Федеральное государственное автономное образовательное учреждение высшего образования
«РОССИЙСКИЙ УНИВЕРСИТЕТ ДРУЖБЫ НАРОДОВ
ИМЕНИ ПАТРИСА ЛУМУМБЫ»

На правах рукописи

РЕЗАИАН НАИМ

МЕТОДИКА СЕМАНТИЧЕСКОГО АНАЛИЗА МУЛЬТИМОДАЛЬНЫХ БОЛЬШИХ ДАННЫХ НА ОСНОВЕ ПРИМЕНЕНИЯ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ

2.3.1. Системный анализ, управление и обработка информации, статистика

Диссертация
на соискание ученой степени кандидата технических наук

Научный руководитель:
доктор технических наук, профессор
РАЗУМНЫЙ ЮРИЙ НИКОЛАЕВИЧ

СОДЕРЖАНИЕ

ВВЕДЕНИЕ

Актуальность темы исследования

В условиях стремительного роста объёмов и разнообразия данных одной из центральных проблем современного искусственного интеллекта становится обеспечение их эффективной интеграции, анализа и интерпретации. Современные системы машинного обучения работают с гетерогенными источниками информации, изображениями, звуковыми сигналами, текстами, временными рядами, что требует разработки универсальных методов их объединения и обработки. При этом особую сложность представляет собой наличие в таких данных шумов, пропусков и артефактов, приводящих к искажению исходных признаков и снижению качества аналитических выводов.

В последние годы особое внимание исследователей привлекают тензорные методы, обладающие высокой выразительностью и способностью моделировать сложные взаимосвязи между различными модальностями данных. В отличие от традиционных матричных представлений, тензорные структуры позволяют описывать данные в многомерных пространствах, выявляя латентные зависимости и семантические закономерности. Это делает их эффективным инструментом для задач слияния мультимодальной информации, что особенно актуально в таких областях, как компьютерное зрение, обработка речи, биомедицинские исследования и интеллектуальные системы мониторинга.

Одновременно с задачей интеграции данных сохраняет актуальность проблема устранения шумов и восстановления информации в многомерных массивах. С ростом сложности и размерности данных увеличивается доля случайных искажений, обусловленных как аппаратными, так и методологическими факторами. Эти искажения существенно влияют на точность обучения моделей и достоверность принимаемых решений. Поэтому создание методов, способных эффективно выделять полезные сигналы на фоне шумов и корректно восстанавливать утраченные фрагменты информации, имеет не только теоретическую, но и прикладную значимость.

Таким образом, необходимость одновременного решения задач интеграции, очистки и восстановления мультимодальных данных формирует новое направление исследований, объединяющее тензорный анализ, методы снижения размерности, алгоритмы шумоподавления и реконструкции информации. Комплексное применение этих подходов позволяет повысить устойчивость и интерпретируемость моделей машинного обучения, обеспечивая более глубокое понимание закономерностей, скрытых в многомерных структурах данных.

В этой связи разработка эффективных алгоритмов для мультимодальной интеграции на основе тензорных представлений, а также создание методов подавления шумов и восстановления недостающей информации является актуальной научной задачей, направленной на повышение качества анализа данных и совершенствование интеллектуальных систем обработки информации.

Степень разработанности темы исследования

Исследования, посвящённые вопросам интеграции и анализа мультимодальных данных, имеют значительную историю и в последние годы получили активное развитие благодаря расширению вычислительных возможностей и распространению тензорных представлений. Наиболее ранние работы были сосредоточены на методах слияния данных различной природы – визуальной, звуковой и текстовой. Основные стратегии интеграции включают раннее, позднее и гибридное слияние[1], которые используются для объединения признаков на уровне входных данных, признаковых пространств или на этапе принятия решений.

В ряде исследований [2], [3]мультимодальное слияние применялось для решения задач обнаружения и отслеживания объектов, тогда как Wang et al. [4] и Kankanhalli et al. [5] использовали подобные методы для видеонаблюдения и анализа дорожного движения. Методы объединения звуковых и визуальных модальностей нашли применение в системах распознавания говорящего (Neti et al. [6], Радова и Псутка [7]). Пфлегер [8] предложил подход к интеграции данных на уровне принятия решений, основанный на продукционных правилах, а Holzapfel et al. [9] реализовали мультимодальное взаимодействие человека с роботом,

комбинируя речь и жесты. В более поздних работах Adams et al. [10] и Bredin, Chollet [11] разработали решения для мультимодального анализа видео и идентификации личности по совокупности признаков.

В последние годы в центре внимания научного сообщества оказались тензорные методы, которые позволили преодолеть ограничения классических матричных моделей при работе с многомерными структурами данных. Работы Kolda и Bader [12], Cichocki et al. [13], Osedelets [14], а также Grasedyck et al. [15] заложили теоретическую основу современных подходов к разложению и представлению тензоров. Эти методы (CP, Tucker, Tensor Train, Tensor Ring и др.) обеспечивают компактное описание данных, позволяют уменьшить «проклятие размерности» и выявить скрытые взаимосвязи между модальностями. Тензорные форматы широко применяются в задачах распознавания эмоций [16], биомедицинской визуализации [17], обработке сигналов и интеллектуальном видеонаблюдении [18].

Параллельно развивается направление, связанное с удалением шумов и восстановлением данных. Среди классических методов можно выделить использование фильтров Винера, вейвлет-преобразований и методов главных компонент (PCA), которые позволяют уменьшить влияние случайных искажений [19]. Более современные подходы включают тензорное восстановление с использованием регуляризованных разложений, таких как Low-Rank Tensor Completion и Sparse Tensor Reconstruction [20], [21]. Применение этих методов эффективно при обработке мультимодальных данных, содержащих пропуски, шумы или неполные наблюдения. Отдельное направление составляет использование глубоких нейронных сетей (Denoising Autoencoder, GAN, Diffusion Models) для реконструкции информации в многомерных структурах [22], [23].

Таким образом, проведённый анализ показывает, что существующие решения обеспечивают частичное решение задач интеграции и восстановления данных, однако не предусматривают комплексного объединения тензорных методов и алгоритмов подавления шумов в единой модели. Это определяет необходимость разработки универсального подхода, обеспечивающего устойчивое

слияние мультимодальных данных с одновременным повышением достоверности и полноты восстановленной информации.

Цель диссертационной работы

Целью настоящей работы является разработка нового вычислительного метода для решения проблемы обработки многомерных отношений при работе с большими данными с представлением данных в формате тензорных произведений; создание более простой модели, позволяющей решить проблему переобучения модели, проблему «проклятия размерности», избыточности информации, повысить скорость обработки мультимодальных данных, а также новый подход для удаления шума и восстановления информации в мультимодальных данных для устойчивости работы с нейронными сетями.

Для достижения цели были поставлены следующие **задачи**:

1. Анализ предметных областей использования мультимодальных данных при обработке больших данных.
2. Исследование возможностей и ограничений системы обработки мультимодальных данных.
3. Исследование существующих алгоритмов интеграции многомерных данных на уровне признаков и на уровне принятия решений.
4. Разработка и обоснование метода интеграции, удаления шума и восстановления мультимодальных данных на основе мультирангового тензорного разложения.
5. Проведение вычислительных экспериментов, подтверждающих эффективность предложенных подходов.
6. Разработка библиотеки предложенного метода на языке python.

Объект исследования

Объектом исследования является система интеграции и представления мультимодальных данных на основе тензорного разложения, функционирующая в условиях многомерности, избыточности и зашумлённости исходных данных.

Предмет исследования

Предметом исследования являются тензорные модели представления и

алгоритмы обработки многомерных и мультимодальных данных, направленные на повышение достоверности, устойчивости и эффективности анализа информации.

Научная новизна полученных результатов

1. Разработана структурная модель интеграции мультимодальных данных на основе тензорного представления, в которой различные модальности (изображения, аудиосигналы, текстовые и сенсорные данные) объединяются в едином многомерном пространстве признаков. Предложенный подход обеспечивает сохранение межмодальных связей и позволяет повысить информативность объединённых данных при анализе сложных объектов.

2. Предложен метод построения тензорных представлений с использованием декомпозиции по низкоранговым компонентам (CPD, Tucker, TT), обеспечивающий уменьшение размерности данных без потери значимых признаков. В отличие от традиционных подходов, метод позволяет учитывать корреляционные зависимости между модальностями, что повышает устойчивость последующего анализа к априорной неопределённости и шумам.

3. Разработан алгоритм удаления шумов и восстановления недостающих элементов в мультимодальных данных с использованием тензорного формализма и адаптивных регуляризаторов. Алгоритм сочетает принципы низкорангового восстановления и адаптивной фильтрации, что обеспечивает повышение достоверности реконструированных данных и их пригодности для последующего анализа.

4. Введён механизм совместного анализа и восстановления данных в тензорной форме, позволяющий интегрировать этапы слияния и восстановления информации в едином вычислительном контуре. Это обеспечивает согласованную обработку разнородных источников данных, что ранее рассматривалось как независимые задачи.

5. Разработана обобщённая архитектура интеллектуальной системы анализа мультимодальных данных, включающая модули тензорного слияния, подавления шумов и реконструкции пропусков. Архитектура обеспечивает масштабируемость и возможность адаптации под различные типы данных и прикладные задачи

(видеонаблюдение, биомедицинская диагностика, распознавание эмоций и др.).

6. Предложены новые критерии оценки качества мультимодальной интеграции с учётом тензорного ранга, меры согласованности модальностей и уровня восстановления данных. Это позволило количественно оценить эффективность предложенных алгоритмов и подтвердить их преимущества по сравнению с существующими методами на реальных выборках данных.

Теоретическая значимость работы

Теоретическая значимость полученных результатов заключается в развитии методов тензорного анализа и моделирования мультимодальных данных, направленных на решение фундаментальных проблем представления, интеграции и восстановления многомерной информации. В рамках исследования предложены новые математические модели тензорного слияния данных, обеспечивающие более полное описание скрытых корреляционных связей между модальностями и позволяющие устранить ограничения, присущие традиционным матричным и векторным подходам.

Разработанные модели и алгоритмы формируют основу для построения единой теоретической платформы обработки многомерных данных, в которой обеспечивается совместное решение задач интеграции, удаления шумов и восстановления неполных элементов информации. Это позволило расширить возможности применения тензорных форматов (CPD, Tucker, Tensor Train) в задачах искусственного интеллекта, а также повысить интерпретируемость и устойчивость вычислительных моделей при работе в условиях априорной неопределённости.

Полученные теоретические результаты способствуют развитию современной парадигмы интеллектуальной обработки данных, углубляя представления о тензорных методах как эффективном инструменте для анализа сложных многомерных структур, и формируют научную основу для дальнейших исследований в области мультимодального машинного обучения и интеграции данных.

Практическая значимость работы

Практическая значимость исследования заключается в разработке вычислительных методов и программных средств для эффективной интеграции и анализа мультимодальных данных. Предложенные тензорные алгоритмы обеспечивают объединение разнородных источников информации с сохранением взаимных зависимостей, что повышает точность и устойчивость анализа.

Разработанные методы удаления шумов и восстановления недостающих элементов позволяют использовать их при обработке неполных и зашумлённых данных, применяя в задачах распознавания образов, видеоаналитике и интеллектуальном мониторинге.

В рамках работы создана программная библиотека на языке Python, реализующая предложенные алгоритмы тензорного слияния и восстановления данных. Результаты экспериментальных исследований подтвердили эффективность предложенного подхода по сравнению с существующими методами.

Методология и методы исследования

Методологической основой диссертационной работы является комплексный подход к интеграции и анализу мультимодальных данных, основанный на тензорных представлениях, методах многомерного моделирования и современных алгоритмах машинного обучения.

В работе использованы положения линейной и мультилинейной алгебры, статистики и теории вероятностей, а также методы оптимизации и анализа больших данных. Основу математического аппарата составляют операции над тензорами и различные виды их разложения, включая CPD, разложение Таккера и Тензорный поезд (Tensor Train) и Tensor Ring. Для повышения устойчивости и достоверности анализа применялись методы регуляризации, рангового ограничения и низкорангового восстановления данных.

Для реализации и экспериментальной проверки предложенных решений разработано программное обеспечение на языке Python и Matlab. Реализованные модули обеспечивают выполнение тензорных операций, моделирование

алгоритмов интеграции и восстановление данных с визуализацией полученных результатов.

Положения, выносимые на защиту

1. Традиционные методы интеграции мультимодальных данных не обеспечивают сохранение межмодальных зависимостей и устойчивость к шумам и пропускам, что приводит к потере части семантической информации и снижению точности анализа в условиях априорной неопределённости.

2. Предложен метод тензорного слияния мультимодальных данных на основе мультирангового разложения, обеспечивающий эффективное уменьшение размерности без потери значимых признаков и позволяющий преодолеть проблему «проклятия размерности».

3. Разработан алгоритм удаления шумов и восстановления недостающих элементов данных в тензорной форме, основанный на адаптивной регуляризации и минимизации тензорного ранга, что обеспечивает повышение достоверности и устойчивости анализа.

4. Предложен интегрированный подход к совместной обработке и восстановлению данных, позволяющий объединить этапы слияния и реконструкции в едином вычислительном контуре, что повышает согласованность и информативность результирующих данных.

5. Разработан метод количественной оценки эффективности алгоритмов тензорного слияния и восстановления данных, учитывающий метрики согласованности модальностей, ранговые характеристики и показатели точности реконструкции.

6. Проведён вычислительный эксперимент с использованием реальных и синтетических наборов мультимодальных данных, подтвердивший преимущество предложенных методов по сравнению с существующими подходами по критериям точности, устойчивости и вычислительной эффективности.

Степень достоверности результатов

Поскольку результатом диссертации является математический инструмент (алгоритм), основными способами оценки достоверности работы и результатов

являются математические методы. Проведенные исследования подтверждают корректность и применимость новых предложенных методов при их меньших вычислительных затратах.

Апробация результатов

Основные результаты диссертационной работы докладывались на следующих научных конференциях:

– V International Conference on Information Technologies in Engineering Education (Inforino), 2020;

– Международную конференцию Сбера по искусственному интеллекту "AI Journey 2024", Москва;

– Международную научно-практическую конференцию «Образовательная трансформация в условиях цифровой экономики», организованную Государственным гуманитарно-технологическим университетом (ГГТУ), Московская область, 2025;

– Конференцию «AI-Горизонты», Москва, 2025.

ГЛАВА 1. ОСНОВНЫЕ ПОНЯТИЯ И ТЕКУЩИЕ ИССЛЕДОВАНИЯ В ОБЛАСТИ МУЛЬТИМОДАЛЬНЫХ ДАННЫХ И МЕТОДОВ ИХ СЛИЯНИЯ

Быстрое развитие компьютерных и информационных технологий в последние два десятилетия коренным образом изменило почти все дисциплины в науке и технике, трансформировав многие области благодаря переходу от скудных к высокоинформативным данным, что требовало все более инновационных методов интеллектуального анализа при проведении соответствующих исследований.

В современном мире передача и восприятие информации происходят посредством различных систем и каналов, способных порождать значения и взаимодействовать одновременно. Коммуникация не сводится к какой-то одной знаковой системе, а представляет собой специфическую форму симбиотического взаимодействия [4].

Модальность – понятие, пришедшее в мир искусственного интеллекта из психологии, являет собой форму представления какого-либо абстрактного понятия. Унимодальность сменяется на более сложный и интегрированный вариант подачи информации, мультимодальность. Мультимодальность состоит в формировании значений при помощи разных семиотических средств – модусов (письмо, речь, изображение) – соответствующих социокультурных конвенций. Мультимодальность понимается как описание общих законов и правил взаимодействия в коммуникативном акте вербальных и невербальных знаков [5], соединение различных кодов предъявления информации [6].

Современные вычислительные модели и новейшие достижения в машинном обучении и искусственном интеллекте используют максимум информации от пяти органов чувств (зрение, слух, осязание, обоняние, вкус). В сером веществе, находящемся на поверхности полушарий мозга, обрабатывается и комбинируется информация с максимальной эффективностью, что позволяет человеческому мозгу особенно хорошо производить реляционные рассуждения, основанные на

мультимодальных данных, а также строить семантические связи между объектами.

На сегодняшний день источники сбора данных непрерывно развиваются от традиционного аудио и видео контента до данных о движении, физиологических и биометрических данных и цифровых датчиков см. рис. 1).

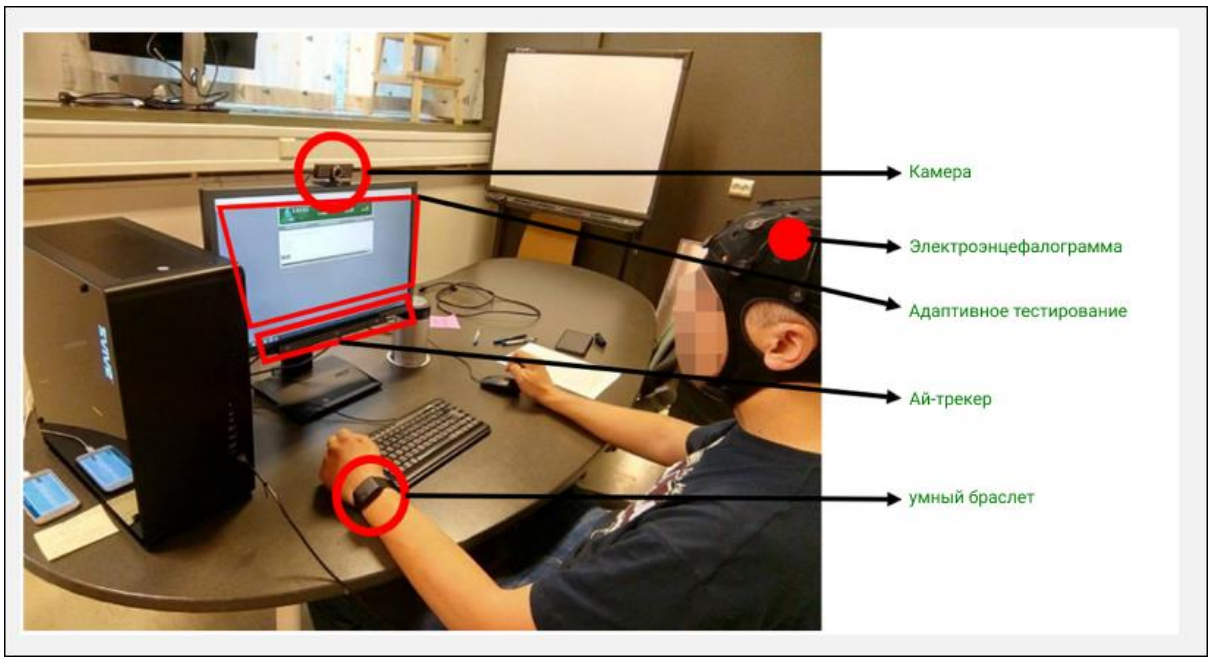


Рис. 1. Экспериментальная установка – участник подключен ко всем устройствам сбора данных

Все данные при этом имеют различные характеристики и статистические свойства и только благодаря соответствующей интеграции возможно решать задачи, которые трудно реализовывать в рамках анализа мономодальных данных [7].

Обучение на основе данных с несколькими представлениями стало серьезным шагом в области искусственного интеллекта и извлечения знаний. Исследователи разработали множество моделей для решения задач машинного обучения, когда информация представлена в неоднородном виде, достигая при этом высокой эффективности их моделей за счет интеграции информации из разных модальностей.

Несмотря на это, реализация задач машинного обучения при работе с мультимодальными данными все еще очень сложна, а иногда не представляется

возможной, в первую очередь, потому что мы имеем дело с эмпирическими данными, которые зачастую сильно искажены шумом.

Например, при анализе генов применение в технологии микромассивов позволило получить огромное количество общедоступных наборов данных по экспрессии генов (см. рис. 2).

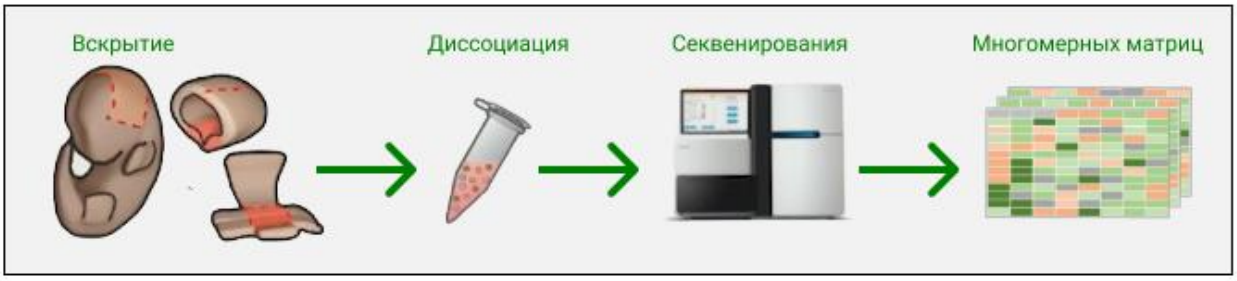


Рис. 2. Путь от вскрытия до многомерных матриц

Однако анализ этих данных с использованием биологической статистики и подходов машинного обучения является сложной задачей из-за высокого уровня шума в данных, что существенно затрудняет использование технологии микромассивов в клинической практике при обнаружении генных сетей.

Проблема представленности данных в различных векторных пространствах и поиск решений по их объединению присутствует во многих реальных задачах. Примерами таких областей могут служить биомедицинские приложения (мониторинг интенсивной терапии и медицинских изображений), транспортные системы (умный автомобиль и дорожные системы), мультимедийный анализ [8] (аудиовизуальная идентификация человека, многомодальное взаимодействие с роботом и многомодальный видеопоиск), распознавание речи , распознавание диктора, биометрическая верификация, обнаружение события, слежение за человеком или объектом, локализация и слежение за активным диктором, анализ музыкального контента, распознавание эмоций, человеко-машинное взаимодействие, обнаружение голосовой активности и разделение источников звукового сигнала. Также проблематика объединения мультимодальных данных присутствует в задачах информационного поиска, так как любая вэб-страница представляет собой огромный массив данных с неструктурированным текстом, полуструктурированными документами, мультимедиа и изображениями (см.

рис. 3).

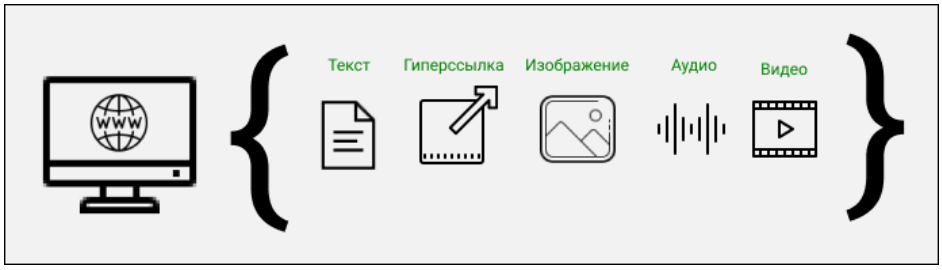


Рис. 3. Типы мультимодальных данных на web-странице

Например, Google использует сложную интеграцию текста и гиперссылок при решении задачи информационного поиска. По запросу выводится огромное количество найденных веб-страниц [28], [29]. В соответствии с традиционным поиском информации, результаты упорядочиваются путем вычисления сходства между текстовым запросом и содержанием веб-страницы. Из-за огромного количества веб-страниц традиционным способом невозможно получить осмысленный результат. В свою очередь, эффективное ранжирование результатов поиска может быть реализовано именно путем интеграции многомерных данных, а не путем интенсивного вычисления текстового сходства.

Перспективность работы с мультимодальными данными и дальнейшее развитие методологии определяется тем, что при работе с унимодальными данными часть информации может быть невидимой или одна модальность может быть недостаточно релевантной для решения конкретной задачи. Рассмотрение нескольких модальностей часто повышает эффективность и имеет серьезные преимущества при создании новой генеративной модели из разных модальностей с помощью вероятностного скрытого семантического анализа с несколькими представлениями. В следствие объединения мультимодальных данных при построении многомерных семантических отношений, мы получаем модель, способную производить реляционные рассуждения.

1.1. Мультимодальные данные

Мультимодальные данные – это способность системы (человека или

машины) воспринимать, обрабатывать и интегрировать информацию, поступающую из различных источников или в разных формах (модальностях).[9] Под модальностью понимается специфический способ представления информации: визуальный, акустический, тактильный, вкусовой, обонятельный и т.д.

У человека мультимодальность реализована через органы чувств специализированная анатомо-физиологическая система, обеспечивающая, благодаря своим рецепторам, получение и первичный анализ информации из окружающего мира и от других органов самого организма, то есть из внешней среды и из внутренней среды организма:

- глаза (зрение) – визуальная модальность (изображения, движение, цвета);
- уши (слух) – акустическая модальность (звук, речь, интонации);
- нос (обоняние) и язык (вкус) – химические модальности;
- кожа (осязание) – тактильная модальность (давление, температура, текстура).

Организм постоянно объединяет данные из этих источников для формирования целостной картины окружающей среды. Например, мы воспринимаем речь не только на слух, но и через движение губ (визуальная информация), особенно в шумной обстановке. Современные устройства (например, смартфоны) также являются мультимодальными системами. Они оснащены различными датчиками:

- камеры – визуальная модальность;
- микрофоны – аудиальная модальность,
- гироскопы, акселерометры – движение и ориентация в пространстве;
- GPS – пространственная локализация;
- датчики приближения и освещения – взаимодействие с внешней средой.

Все эти сенсоры работают совместно, позволяя устройству адаптироваться к поведению пользователя и контексту.

Модальность – это способ восприятия или выражения информации.

Модальности могут быть как необработанными (Raw Modality), так и абстрактными (Abstract Modality). Необработанные модальности представляют

собой непосредственные физические сигналы, фиксируемые сенсорами как видеопоток с камеры (RGB-пиксели), аудиосигнал с микрофона или сырые данные с акселерометра, гироскопа и других датчиков [10].

Абстрактные модальности, напротив, являются результатом когнитивной или алгоритмической обработки необработанной информации. Яркий пример, язык, представленный в виде текста. Хотя человек получает речевую информацию через акустический сигнал, смысловая интерпретация (лексика, синтаксис, семантика) возникает после значительного уровня обработки. Таким образом, язык как модальность является более семантически насыщенным и символическим, в то время как аудиосигнал, на котором он основывается, лишь его носитель, находящийся на более низком уровне абстракции. Чем меньше уровень обработки, необходимый для восприятия модальности, тем ближе она к необработанной информации. И наоборот, чем выше уровень символичности и абстракции, тем дальше модальность от физического сигнала и тем сложнее её интерпретация (рис. 4).

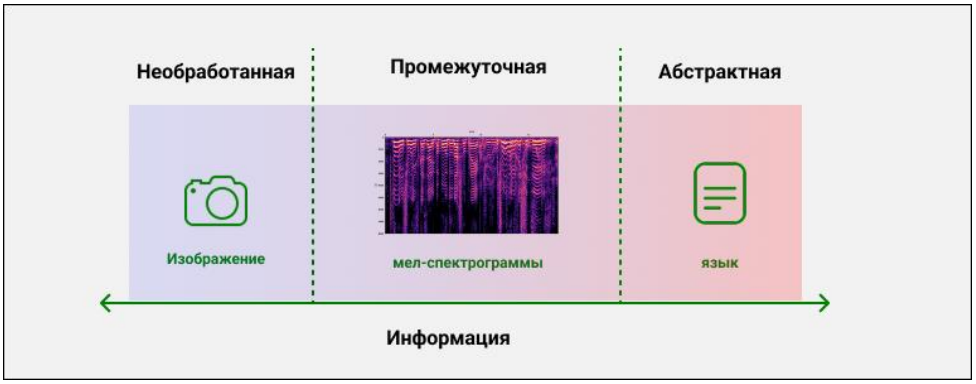


Рис. 4. Градиент абстракции модальностей

Когда данные поступают из разных модальностей (например, аудио и изображения), они различаются по структуре, репрезентации и семантическому содержанию. Это делает задачу совместной обработки сложнее [9]. Разные модальности могут отличаться по степени структурной и семантической близости. На шкале модальностей можно выделить два типа [11]:

– Однородные модальности (homogeneous) – это схожие структуры, такие как изображения с двух разных камер или тексты на двух языках;

– Гетерогенные (Heterogeneous) – это разные структуры, например, такие как звук и изображение, требуют специализированных методов обработки и объединения.

Таким образом, мультимодальные данные – это не просто “много разных данных”, а те, которые различаются по структуре, но объединяются в рамках одного контекста или задачи, взаимосвязаны (interconnected), взаимодействуют между собой и помогают лучше понять происходящее.

Модальности также различаются по уровню абстракции, чем выше уровень абстракции, тем более вероятно, что модальности окажутся однородными, а чем ближе к физическому сигналу, тем чаще они разнородны.

Современные методы обработки изображений всё чаще опираются не на пиксельное представление, а на объектно-ориентированное восприятие. То есть изображение интерпретируется как набор объектов, извлечённых с помощью моделей детекции (например, Faster R-CNN, YOLO). Таким образом, в визуальной модальности базовой единицей анализа становится объект, а не пиксель [12].

Это особенно важно при проектировании мультимодальных систем. При объединении различных типов данных необходимо определить базовую единицу (elemental unit) каждой модальности – такую, которая:

1. Несёт достаточно информации, чтобы быть полезной при обучении модели;

2. Обладает внутренним различием, чтобы между разными единицами можно было строить осмысленные связи (например, векторы, отношения, трансформации).

В языковой модальности такой единицей традиционно является слово, хотя в современных трансформерах нередко используются токены (субсловесные единицы).

Таким образом, при мультимодальной интеграции:

– в визуальной модальности элементами являются объекты (или регионы интереса);

– в текстовой – слова или токены;

– в аудио – фреймы, мел-спектрограммы или фонемы.

Корректный выбор элемента важен для построения совместного представления в общем векторном пространстве, что является центральной задачей мультимодального обучения [13].

При разработке мультимодальных систем необходимо учитывать высокую степень гетерогенности данных, обусловленную многомерной природой различий между модальностями. Каждая модальность обладает уникальными характеристиками, которые проявляются не только в типе данных (например, изображение и текст), но также в их внутренней структуре, распределении, уровне шума и значимости для решаемой задачи. Успешное объединение различных модальностей требует построения совместного представления, чувствительного к этим особенностям [14]. Ниже рассматриваются шесть ключевых измерений гетерогенности между модальностями, которые необходимо учитывать при интеграции информации в рамках мультимодального анализа (рис. 5).

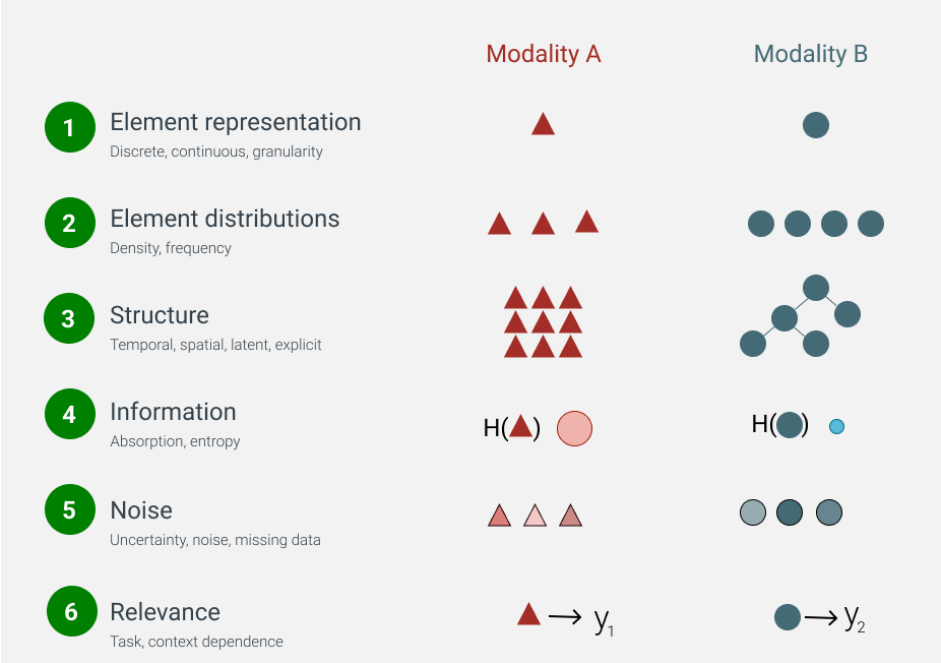


Рис. 5. Шесть измерений гетерогенности между модальностями

1.1.1. Представление элементов

Модальности могут отличаться по характеру представления своих элементарных единиц, что влияет на их способность работать с различными

типами данных. Эти единицы могут быть дискретными, как слова в тексте, или непрерывными, как значения RGB-пикселей в изображениях. В то же время, существует гранулярность, которая может иметь разную степень детализации, что еще больше усложняет структуру модальностей [15].

Например, в текстах элементами являются слова, что делает представление информации дискретным и четко структурированным. В изображениях возможны два подхода: можно рассматривать пиксели, представляющие непрерывные значения цвета, или более высокоуровневые абстракции, такие как объекты, что обеспечивает разнообразие представлений. В аудио сигнале элементом может быть фрейм спектрограммы, который является еще одним примером непрерывного представления, где каждая единица имеет свою уникальную частотную характеристику. Таким образом, каждая модальность предоставляет свою уникальную структуру для представления информации, что обусловлено различием в элементах и их уровнях абстракции.

1. Распределение элементов (Element Distributions)

Модальности различаются по частоте и плотности элементов. Например, изображение в виде пикселей обладает высокой плотностью, а текст с небольшим числом ключевых слов – низкой. Это важно при формировании репрезентаций: высокая плотность требует агрегации или отбора информации [13].

2. Структура (Structure)

Модальности обладают различной внутренней структурой, которая может быть временной(temporal), пространственной(spatial), латентной(latent) или явно заданной (explicit). Важно отметить, что тип структуры определяет как элементы модальности взаимодействуют между собой и как их порядок влияет на восприятие и интерпретацию информации [13].

Примером временной структуры является аудио, где порядок и длительность звуков определяют грамматические и семантические связи. Видеоматериалы также имеют пространственно-временную структуру, где кадры следуют друг за другом, образуя динамичную последовательность. Текст характеризуется линейной и грамматической структурой, где важен порядок слов для правильного понимания

смысла и соблюдения грамматических правил. В изображениях структура выражается в пространственном расположении объектов, где соседние пиксели или регионы сцены могут образовывать логическую иерархию. Например, лицо состоит из глаз, носа и рта, что позволяет воспринимать изображение как целостную структуру. Графовые данные, в свою очередь, имеют явно заданную структуру через связи между узлами, что позволяет моделировать сложные взаимосвязи и взаимодействия объектов в графе.

Таким образом, различные модальности имеют различные типы внутренней структуры, что определяет их особенности в представлении и обработке данных. Пространственная структура присуща изображениям, линейная – тексту, а временная – аудио, что влияет на методы анализа и обработки каждой из них.

3. Информационное содержание (Information)

Разные модальности несут различное количество семантической информации в своих элементах. Это может измеряться через понятия абстракции или энтропии. Например, объект на изображении обычно содержит больше информации, чем один пиксель, а фраза в тексте – больше, чем отдельное слово. Таким образом, уровень информативности единичного элемента варьируется от модальности к модальности.

4. Шум (Noise)

Каждая модальность подвержена различным типам шума, таким как неопределенности, пропущенные данные или артефакты, которые могут искажать восприятие информации [16]. В текстах возможны опечатки или синтаксические ошибки, что нарушает правильность и точность представления данных. В изображениях могут возникать размытость или засветка, что ухудшает качество визуальной информации и делает её трудной для анализа. В аудио часто встречаются фоновый шум, искажения или эхо, которые могут затруднить понимание звуковых сигналов. Кроме того, возможны случаи отсутствующих элементов, например, недостающие субтитры к фрагменту видео, что приводит к пропуску важной информации [17].

5. Актуальность для задачи (Relevance)

Элементы в различных модальностях могут обладать разной степенью важности в зависимости от задачи и контекста. Значимость каждой модальности и её элементов варьируется в зависимости от специфики проблемы. Например, для задачи описания изображений визуальная модальность будет критически важной, в то время как при ответах на вопросы, основанные на тексте, её роль может быть второстепенной [18].

В задачах, требующих обработки нескольких типов данных, например, при определении эмоций из видео, важны как визуальные элементы (выражение лица), так и аудио (интонация). В задаче навигации по помещению визуальная модальность может сыграть ключевую роль, в то время как текстовая информация может быть несущественной. В случае чат-бота, который обрабатывает речь, интонационные акценты могут изменять смысл фразы, что делает важным учет этих элементов при взаимодействии с пользователем. Таким образом, контекстуальная релевантность определяет, какие модальности и их элементы имеют приоритет, превращая интеграцию различных типов данных в задачу не только техническую, но и семантическую.

1.1.2. Мультимодальное представление

В современной машинном обучении и искусственном интеллекте одним из центральных направлений является мультимодальное обучение (Multimodal Machine Learning, MML). Его задача заключается в интеграции информации, поступающей из различных источников (модальностей), таких как текст (язык), изображение (зрение), аудио (акустическая информация) и других сенсорных данных.

Декомпозиция модальностей на элементы представлена таким образом, когда каждая модальность может быть выражена как последовательность элементарных единиц:

– Язык может быть разбит на отдельные слова или токены. Например, предложение «I really like this tutorial» можно интерпретировать как

последовательность слов;

– Зрительная информация (видео) может быть представлена набором последовательных кадров, каждый из которых отражает момент времени;

– Аудио (речь) может быть декомпозировано на сегменты звукового сигнала, соответствующие, например, произнесению отдельных слов или фонем [19] (рис. 6).

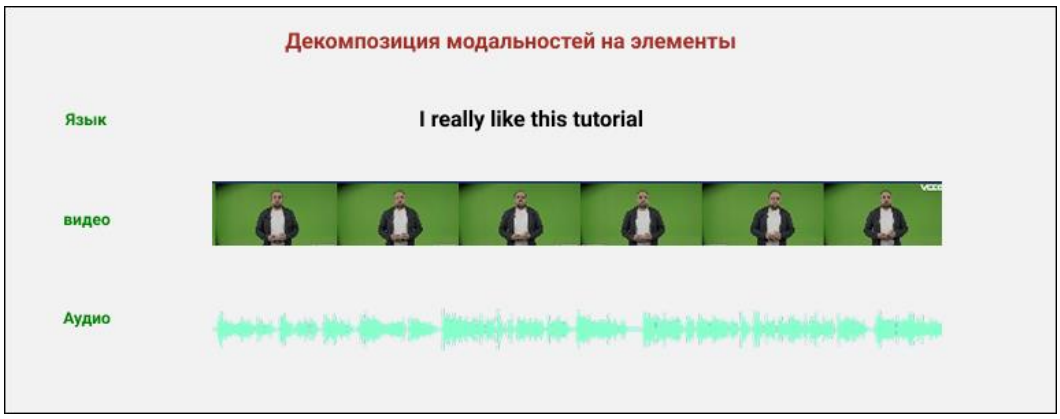


Рис. 6. Декомпозиция модальностей на языковые, визуальные и аудиальные элементы

Таким образом, несмотря на различие в природе данных, каждая модальность может быть преобразована в структурированную последовательность элементов, отражающую ее внутреннюю организацию. Следует отметить, что под модальностью понимается не только принципиально разный тип данных (например, текст против аудио), но и различные ракурсы или представления одного и того же объекта. Такой подход известен как multi-view learning [20] (рис. 7).

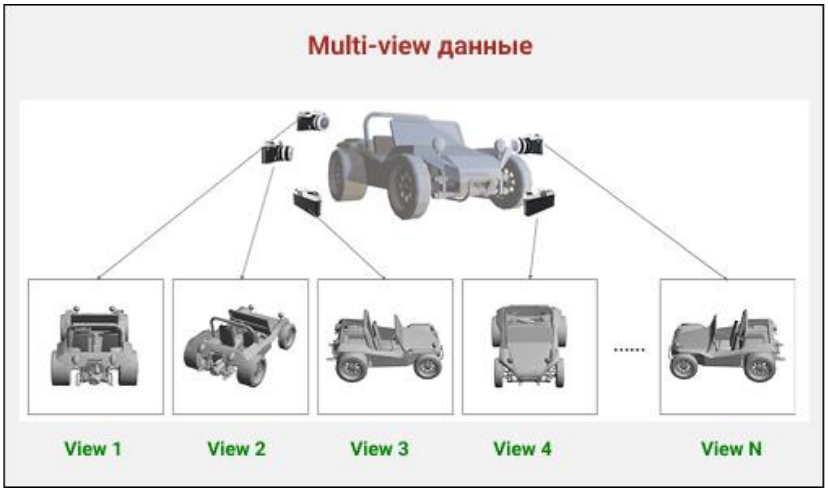


Рис. 7. Многовидовые данные объекта

Примером может служить датасет с 360-градусными изображениями объектов. Каждый объект представлен серией изображений, снятых с разных углов обзора. В этом случае каждое изображение под определённым углом рассматривается как отдельный «вид» (view). Совокупность таких изображений образует множественные представления одной модальности (зрения). Таким образом, multi-view данные занимают промежуточное положение между классической мультимодальностью и одномодальными данными. С одной стороны, все виды относятся к одной сенсорной модальности (зрение), но с другой каждый ракурс несёт уникальную информацию и может рассматриваться как отдельный источник данных.

Применение multi-view подхода позволяет:

– извлечь более устойчивые представления, поскольку модель обучается на разнородных ракурсах одного объекта.
– улучшить обобщающую способность, так как система видит объект с разных сторон.
– снизить риск переобучения, если информация из разных ракурсов согласованно интегрируется в общее представление.

Таким образом, мультимодальное представление может включать в себя как разнородные источники данных (язык, зрение, аудио), так и различные «проекции» одного типа данных (multi-view), расширяя границы применения мультимодального машинного обучения [21].

Основная цель мультимодального машинного обучения заключается в построении универсальных представлений (representations), которые учитывают внутреннюю структуру каждой модальности, отражают межмодальные взаимодействия, связи между элементами, принадлежащими разным модальностям (например, слово в языке и соответствующее ему изображение в видео или звук в аудиодорожке).

Эти представления могут использоваться для решения широкого спектра задач:

– классификация (например, определение эмоционального состояния на

основе речи, выражения лица и текста);

– генерация (например, генерация описания изображения на естественном языке);

– мультимодальный поиск (например, поиск изображений по текстовому описанию).

Одним из ключевых вызовов в мультимодальном машинном обучении является проблема представления (Representation Challenge). Под этим понимается необходимость конструирования таких векторных представлений, которые способны эффективно кодировать как внутренние структуры каждой модальности, так и их взаимные зависимости [9].

Задача представления в мультимодальном машинном обучении заключается в построении таких отображений, которые отражают межмодальные взаимодействия между индивидуальными элементами различных модальностей.

Эта задача является фундаментальным строительным блоком для большинства мультимодальных моделей и определяет качество их работы в прикладных задачах.

В когнитивных науках и исследованиях мультимодальности важным аспектом является понимание того, как различные модальности (способы восприятия, такие как зрение, слух, язык) могут быть связаны между собой. Связь между модальностями позволяет обмениваться значимой информацией и делает их совместное использование более эффективным. Когда речь идет о взаимодействии между модальностями, важно рассматривать, насколько информация, передаваемая через каждый канал восприятия, пересекается и как это влияет на наше восприятие и понимание мира.

Связанные модальности обозначают такие формы взаимодействия, при которых существует общая информация, объединяющая разные способы восприятия. Например, если человек видит стол и одновременно слышит слово "стол", между этими двумя модальностями – зрением и языком – возникает общая информация о самом объекте. Это позволяет каждому каналу сосредоточиться на общей информации, даже если отдельные аспекты каждой модальности могут

теряться или изменяться. Визуально это изображается в виде двух кругов, каждый из которых представляет отдельную модальность. Пересечение этих кругов символизирует общую информацию между модальностями, которая усиливает их взаимное восприятие. Чем больше перекрытие этих кругов, тем сильнее связь между модальностями [22].

Существуют также различные типы связей, которые могут возникать между модальностями. Эти связи могут быть статистическими, семантическими или зависеть от социального соглашения и контекста. На картинке представлен ряд категорий, которые объясняют, как разные виды связей могут влиять на восприятие информации [23]:

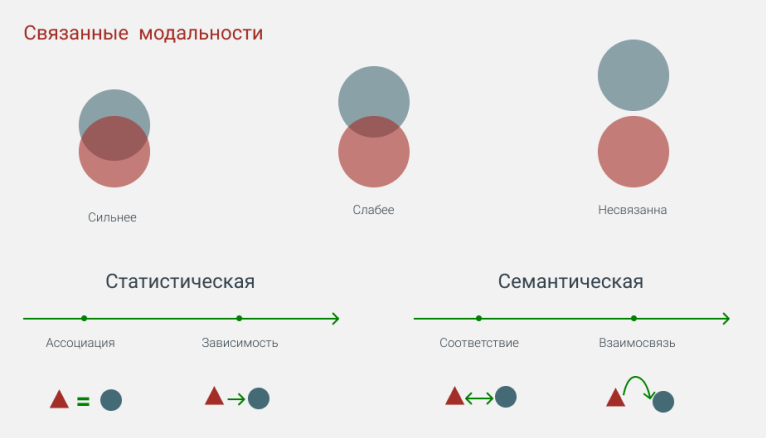


Рис. 8. Типы связей между модальностями: статистическая и семантическая

Ассоциация (Association) – это базовый тип связи, который возникает, когда два элемента постоянно встречаются вместе. В контексте модальностей, например, каждый раз, когда мы показываем на стол и говорим "стол", возникает ассоциативная связь между визуальным восприятием объекта и его названием. Чем чаще эта связь повторяется, тем сильнее она становится. В этом случае взаимодействие между зрением и словом укрепляется, и мозг начинает ассоциировать эти две модеальности [24].

Зависимость между модальностями означает, что одно событие или действие происходит перед или после другого. Например, в случае с визуальными и языковыми модальностями, определенные элементы речи могут зависеть от того,

что человек видит. Зависимости также могут быть временными – например, когда мы видим определенные действия, которые сопровождаются словами, описывающими эти действия.

Соответствие – это связь, которая основывается на договоренности или знании. Например, общество договорилось, что слово "стол" будет означать определенный объект, который мы видим. Здесь важна договоренность между восприятием и языковым обозначением, что позволяет человеку понять и использовать информацию о вещах в мире. Соответствие укрепляется в обществе через общие соглашения.

Взаимосвязь касается порядка слов и их синтаксической структуры. В языке существует множество фраз, где слова следуют друг за другом в определенном порядке для формирования осмысленных высказываний. Например, фразы как "стол стоит в комнате" или "стол используется для работы" представляют собой структуру, где каждое слово связано с другим в определенной логической последовательности.

Одним из наиболее важных факторов, который объясняет, почему язык и зрение так хорошо работают вместе, является их сильная общая связь. Эта связь позволяет эффективнее обрабатывать информацию и усиливает понимание, что делает взаимодействие между этими модальностями особенно продуктивным. Когда информация из разных источников (например, зрительного восприятия и языка) объединяется, это помогает человеку более полно воспринимать мир и адекватно реагировать на него.

Связь между модальностями имеет важное значение для улучшения взаимодействия и восприятия мира. Чем более тесно связаны разные модальности, тем более эффективно происходит обработка информации. Это особенно важно в контексте мультимодальных систем, где требуется объединение информации из различных источников для получения более точных и комплексных выводов. Например, в области обучения машин и искусственного интеллекта связанное использование текста и изображения позволяет системам лучше понимать контекст и делать более точные прогнозы [25].

Таким образом, понимание и использование связей между различными модальностями является важным аспектом в когнитивных науках и помогает создавать более эффективные модели взаимодействия, как в области искусственного интеллекта, так и в когнитивных исследованиях. Сильные и слабые связи между модальностями влияют на то, как мы воспринимаем, обрабатываем и используем информацию, что способствует более глубокому пониманию мира вокруг нас.

В контексте мультимодальной коммуникации, важным аспектом является понимание того, как различные типы сигналов и откликов взаимодействуют друг с другом. Это взаимодействие можно классифицировать в зависимости от того, как сигналы комбинируются, усиливаются или изменяются в процессе передачи информации и восприятия отклика. Основная цель мультимодальной коммуникации – создать более точное и эффективное восприятие мира за счет синергии разных сенсорных каналов.

1.1.3. Сигналы в мультимодальной коммуникации

Сигнал – это информация, которая поступает от внешнего источника и воздействует на систему. В контексте мультимодальной коммуникации, сигналы могут поступать через различные сенсорные каналы, такие как зрение, слух, осязание и другие [26]. Эти сигналы могут быть физическими, химическими или электрическими, и они передают информацию, которая затем обрабатывается системой.

Звуковой сигнал: речь или музыка.

Визуальный сигнал: изображения, видео, жесты.

Тактильный сигнал: ощущение температуры или давления.

Взаимодействие между различными типами сигналов позволяет создать более полное восприятие и реакцию на стимулы, что делает процесс коммуникации более эффективным и насыщенным.

Реакция системы на полученный сигнал – это отклик. В контексте

мультимодальной коммуникации, отклик может быть физическим, когнитивным, эмоциональным или поведенческим. Тип отклика зависит от того, какие сигналы были восприняты, а также от того, как они были обработаны системой.

Примеры откликов:

– Физический отклик: движение руки в ответ на визуальный сигнал (например, указание пальцем на объект).

– Когнитивный отклик: понимание речи или текста, восприятие смысла.

– Эмоциональный отклик: реакция на визуальный или аудиовизуальный стимул, вызвавший эмоции (например, страх или радость).

В мультимодальной коммуникации отклики часто происходят как результат интеграции нескольких типов сигналов, что позволяет создать более глубокое и многозначное восприятие ситуации.

Между сигналами и откликами существует тесная связь, так как сигналы служат катализаторами для возникновения откликов. В мультимодальной коммуникации сигналы могут комбинироваться, усиливаться или изменяться, что приводит к различным типам взаимодействий и откликов [27] (рис. 9).

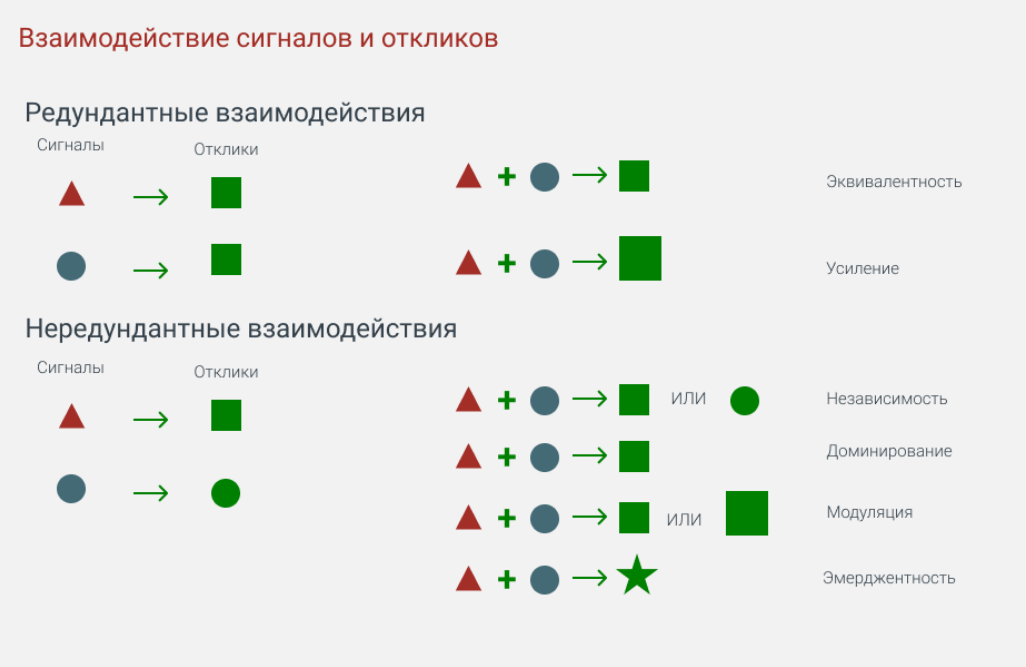


Рис. 9. Типы взаимодействия сигналов и откликов: редундантные и нередундантные взаимодействия

Редундантные взаимодействия характеризуются тем, что сигнал и отклик

обладают схожими или идентичными свойствами. В таких случаях добавление второго сигнала не приводит к значительному изменению, а лишь усиливает или повторяет информацию, уже передаваемую первым сигналом. Существует два типа редундантных взаимодействий [28].

Эквивалентность (Equivalence): в этом случае два сигнала приводят к одному и тому же отклику. Это может происходить, когда разные сенсорные каналы, например зрение и слух, передают схожую информацию и вызывают одинаковую реакцию (рис. 10).

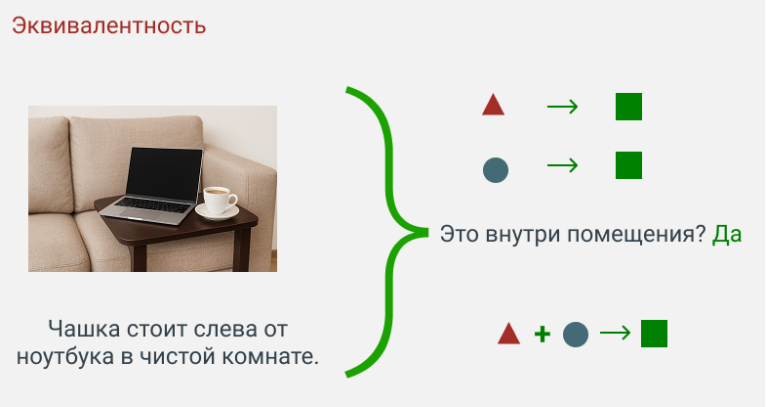


Рис. 10. Пример эквивалентного взаимодействия между модальностями

Эквивалентность позволяет усилить восприятие, так как одна и та же информация поступает через разные каналы восприятия. два сигнала (например, изображение и текст) каждый по отдельности могут привести к правильному выводу. Даже если один убрать, ответ останется верным [28].

Усиление (Enhancement): один сигнал сам по себе недостаточен, но добавление второго усиливает восприятие и делает вывод более надёжным (рис. 11).

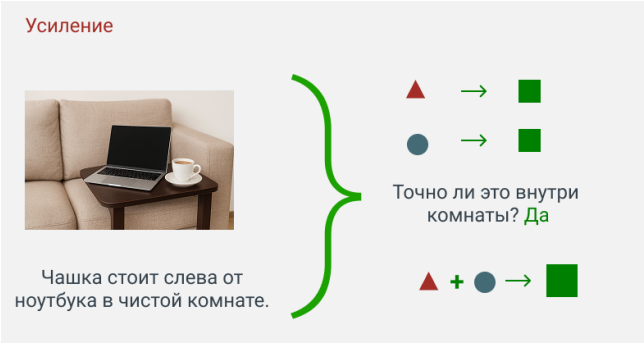


Рис. 11. Пример усиления при взаимодействии между модальностями

Например, когда зрительный сигнал дополняет или усиливает звук, улучшая восприятие объекта или события. В этом случае взаимодействие между сигналами не просто повторяет информацию, а улучшает восприятие исходного сигнала.

В отличие от редундантных взаимодействий, нередундантные взаимодействия (Nonredundancy) характеризуются тем, что каждый из сигналов привносит уникальную информацию, дополняя и расширяя восприятие. В таких взаимодействиях каждый канал восприятия имеет свою собственную функцию и не дублирует информацию другого канала. Нередундантные взаимодействия включают следующие типы [29].

Независимость (Independence): в случае независимости два сигнала, например, зрительный и звуковой, не влияют друг на друга. Каждый сигнал вызывает отдельный отклик, и оба сигнала действуют независимо, предоставляя разную информацию, но не усиливая или изменяя восприятие друг друга [30].

Доминирование (Dominance): представляет собой особый тип взаимодействия модальностей, при котором одна модальность оказывается более информативной или надёжной по сравнению с другой и, следовательно, определяет итоговый вывод системы. В отличие от эквивалентности, где различные модальности приводят к одному и тому же правильному результату, и в отличие от усиления, где комбинирование сигналов повышает уверенность в ответе, доминирование возникает в ситуациях, когда одна модальность предоставляет корректную информацию, тогда как другая может быть менее точной, неопределённой или даже вводящей в заблуждение [31] (рис. 12).

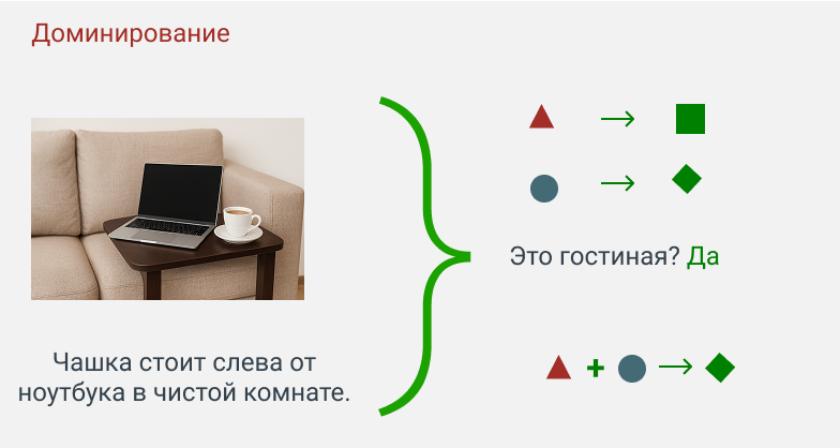


Рис. 12. Пример доминирования одной модальности над другой

Примером может служить задача определения контекста на основе изображения и текстового описания. Визуальная модальность (изображение гостиной с ноутбуком и чашкой на столике) однозначно указывает на то, что сцена находится в помещении, тогда как текстовое описание («чашка справа от ноутбука в чистой комнате») не даёт прямого указания на то, что это именно гостиная, и может интерпретироваться более широко. В данном случае правильный вывод достигается за счёт приоритета визуальной информации над текстовой, что иллюстрирует феномен доминирования.

Таким образом, доминирование отражает необходимость адаптивного выбора модальности, способной обеспечить наибольшую достоверность результата. Этот механизм играет ключевую роль в мультимодальных системах, так как позволяет минимизировать влияние ошибок или шумов в одной из модальностей, сохраняя при этом способность к корректному восприятию и принятию решений.

Модуляция (Modulation): представляет собой тип взаимодействия модальностей, при котором один сигнал изменяет или трансформирует восприятие другого, не обязательно усиливая или заменяя его. В отличие от эквивалентности, где сигналы дают идентичный результат, или доминирования, где один сигнал получает приоритет над другим, модуляция характеризуется тем, что один канал воздействует на интерпретацию другого, формируя новое качество восприятия.

Примером модуляции может служить ситуация с изображением гостиной, где на столике рядом с диваном расположены ноутбук и чашка. Визуальная модальность (изображение) позволяет сделать вывод о том, что сцена происходит в гостиной, то есть в помещении. Текстовая модальность, напротив, описывает лишь пространственное соотношение объектов: «чашка слева от ноутбука в чистой комнате». Если рассматривать текст в отрыве от изображения, его интерпретация может быть более широкой и включать различные сценарии (например, кухня, кабинет или библиотека) [32].

Однако при объединении модальностей визуальная информация модифицирует восприятие текста, направляя его интерпретацию в сторону

правильного контекста гостиной.

Таким образом, модуляция отражает процесс, при котором одна модальность не просто дублирует или уточняет другую, а активно изменяет её интерпретацию. Этот механизм играет ключевую роль в мультимодальных системах, поскольку обеспечивает согласованное и контекстуально релевантное восприятие информации (рис. 13).

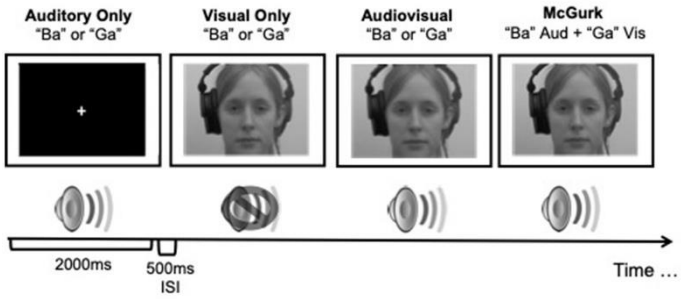


Рис. 13. Экспериментальная схема демонстрации эффекта МакГурка

Или примером модуляции является ситуация, когда визуальная информация изменяет восприятие аудиального сигнала. Так, зрительный контекст может повлиять на то, как интерпретируется звук: например, одно и то же звуковое событие может восприниматься по-разному в зависимости от сопровождающего изображения или визуальной сцены. Подобные эффекты демонстрируются в психологии восприятия (например, в эффекте МакГурка, когда артикуляция губ изменяет восприятие услышанного звука) [33].

Эмерджентность (Emergence): это тип взаимодействия модальностей, при котором объединение сигналов приводит к возникновению нового, качественно иного отклика, который невозможно предсказать, исходя лишь из реакций на отдельные модальности. В отличие от эквивалентности, усиления или доминирования, здесь результат не сводится к простому суммированию или выбору информации, а формируется новое содержание восприятия [34] (рис. 14).

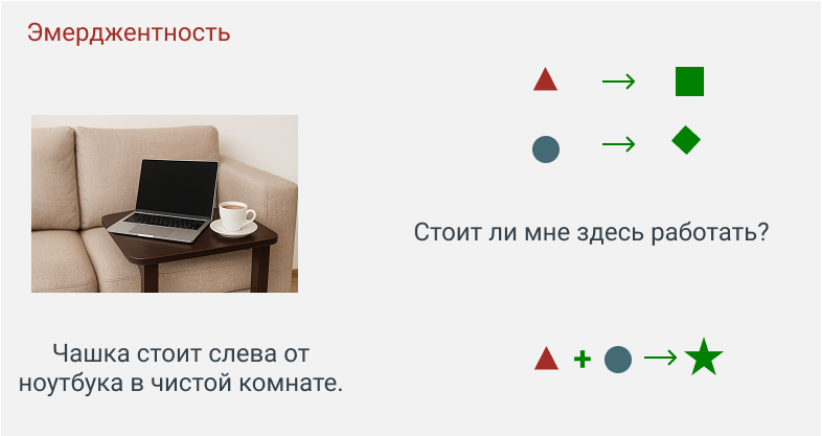


Рис. 14. Пример эмержентного взаимодействия между модальностями

Примером может служить ситуация с вопросом: «Стоит ли мне здесь работать?».

Визуальная модальность (изображение): «Может быть? Удобный диван, но столик слишком маленький».

Текстовая модальность (описание): «Может быть? Комната чистая и есть чай».

Каждая модальность по отдельности акцентирует внимание на различных аспектах: визуальная модальность сосредоточена на удобстве мебели и ограничениях рабочего пространства, текстовая модальность подчёркивает чистоту комнаты и наличие напитка.

Однако при объединении этих сигналов возникает новое качество восприятия: формируется более целостная картина условий, включающая как комфорт и ограничения рабочего места, так и чистоту и уют обстановки. Именно это новое восприятие нельзя свести к информации каждой отдельной модальности – оно возникает только в результате их взаимодействия.

Таким образом, эмерджентность демонстрирует, что мультимодальная интеграция способна порождать новые уровни интерпретации информации, которые не могут быть получены изолированно из отдельных каналов.

Понимание классификации взаимодействий между различными типами сигналов и откликов в мультимодальной коммуникации является ключевым аспектом для улучшения восприятия и обработки информации. Эти

взаимодействия могут существенно влиять на эффективность коммуникации, восприятие событий и формирование откликов на различные стимулы. Исследования в этой области позволяют глубже понять, как различные сенсорные каналы могут быть интегрированы для создания более эффективных мультимодальных систем, как в повседневной жизни, так и в научных и технических приложениях.

1.1.4. Стратегии слияния мультимодальных данных (Data Fusion)

Слияние данных - это информационный процесс, касающийся ассоциации, корреляции и комбинации данных от одного или нескольких источников для достижения точных оценок параметров, характеристик, событий и поведения наблюдаемых объектов, которые не могут быть получены из одного источника [35].

Термин «многомодальная интеграция/многомодальное объединение» может относиться к любой стадии процесса интеграции, где присутствует реальная комбинация различных источников информации, отражая факт того, что полученные синтезированные данные объединяют в себе свойства исходных данных, что уменьшает общую неопределенность и способствует повышению точности, с которой признаки воспринимаются системой. Избыточность информации также служит цели повышения надежности системы в случае ошибки или сбоя в исходных сигналах.

Главной задачей технологии слияния данных (Data Fusion) является объединение данных из разных источников в интересах решения последующих содержательных задач: принятие решений, классификация, определение состояния объектов, оценка ситуации и т.д [36].

Общая схема исследования в мультимодальном обучении заключается в изучении характеристик каждой модальности, а затем объединении всех характеристик для принятия окончательного решения. Объединение представлений различных модальностей является основной проблемой в любой мультимодальной задаче (рис. 15).

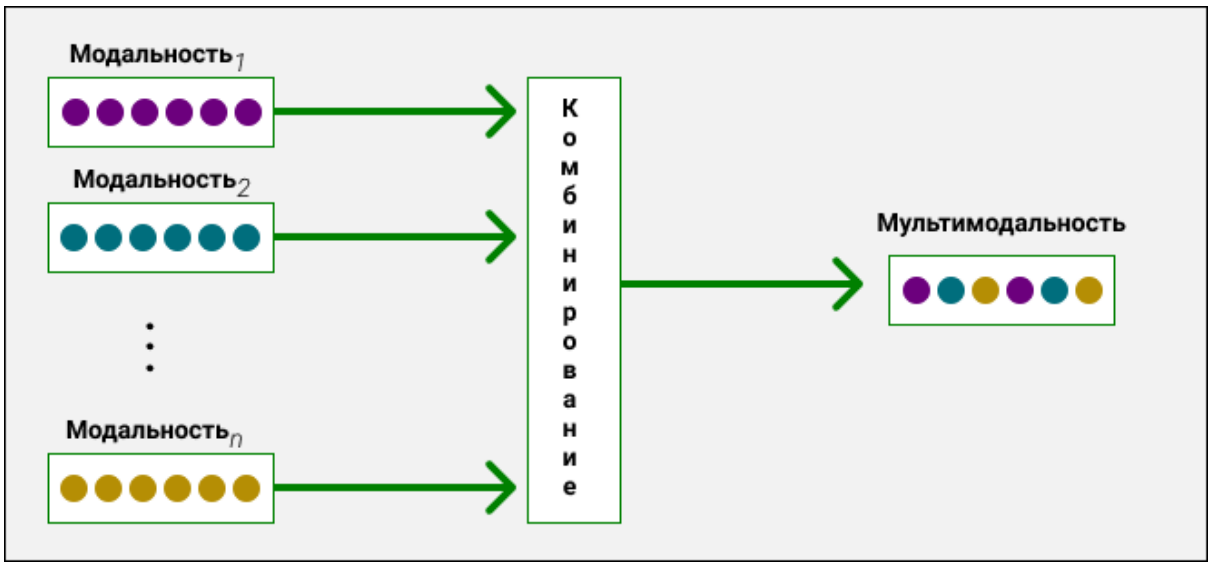


Рис. 15. Процесс комбинирования от унимодальных к мультимодальным данным

Различные модальности, используемые в процессе объединения, могут обладать избыточной или противоречивой информацией, и поэтому необходимо понимать, какие модальности способствуют выполнению той или иной задачи анализа с учетом достижения большей эффективности.

Так как большинство подходов базируются только на корреляционной связи между модальностями, мы рассмотрим этот аспект более подробно в следующем разделе. При этом преимуществом мультимодального слияния является возможность использовать как корреляционные, так и комплементарные связи, представляющие собой дополнительную информацию из нескольких модальностей, которая позволяет использовать признаки, которые невозможно однозначно воспринять, имея лишь информацию от каждой модальности в отдельности.

Корреляционные связи или корреляционная зависимость – это вероятностные изменения, которые можно изучать только на представительных выборках методами математической статистики. «Оба термина, корреляционная связь и корреляционная зависимость – часто используются как синонимы. Зависимость подразумевает влияние, связь – любые согласованные изменения, которые могут объясняться сотнями причин. Корреляционные связи не могут рассматриваться как свидетельство причинно-следственной зависимости, они свидетельствуют лишь о том, что изменениям одного признака, как правило,

сопутствуют определенные изменения другого [7].

Данные из разных модальностей, как правило, содержат одинаковую или сходную семантическую информацию, коррелирующую друг с другом. Для устранения двойного смысла слова корреляционная связь между текстом и изображениями означает, что изображения и текстовые предложения одних и тех же документов, как правило, содержат семантическую информацию, описывающую одни и те же объекты или понятия. Например, изображение и текстовое предложение на рис. 16 (А) оба относятся к смыслу «лук (овощ)», в то время как изображение и предложение на рис. 16 (Б) оба описывают смысл «лук (оружие)».

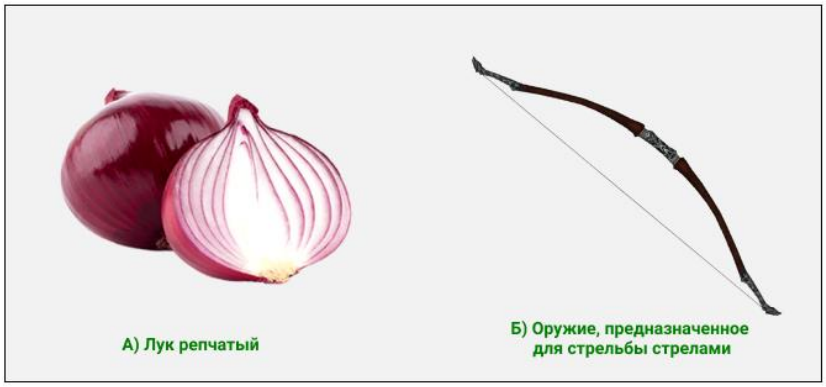


Рис. 16. Пример комплементарных связей текста и картинки

Поскольку информация из разных модальностей имеет эту корреляционную связь, модальности имеют тенденцию коррелировать и в пространствах признаков. Исходя из этого, можно также провести корреляционный анализ для построения единого пространства признаков по нескольким модальностям. В предыдущих научных работах [37] [38] корреляционное отношение использовалось для разработки унифицированной модели представления мультимодальных документов.

Корреляция между различными модальностями обеспечивает связи, которые очень полезны в процессе их объединения. Корреляция может быть установлена на самых различных уровнях, как между характеристиками низкого уровня, так и между решениями семантического уровня. Исследователи используют различные

методы корреляционного анализа, такие как коэффициент корреляции, взаимная информация, скрытый семантический анализ, канонический корреляционный анализ и кросс-модальный факторный анализ. Посмотрим подробнее на различные методы вычисления корреляций и проанализируем их с точки зрения того, как они влияют на процесс слияния данных.

Мы уже немного говорили о том, что слияние модальностей может выполняться на нескольких различных уровнях. Первым шагом при этом является определение самой стратегии, которой необходимо следовать в процессе объединения модальностей. Выбор стратегии непосредственно зависит от задачи, которую хотим решать. Существуют три основных способа слияния [43]:

– *раннее слияние* - объединение информации на уровне признаков, когда объединение на уровне векторов признаков делается до начала процесса моделирования путем объединения признаков из всех модальностей [39];

– *позднее слияние* - слияние на уровне принятия решений, когда моделирование каждого канала выполняется отдельно, а затем выходы или решения моделей интегрируются для принятия окончательного решения [40], [41];

– *гибридное слияние* - сочетание этих двух подходов.

Раннее слияние является наиболее используемой стратегией объединения информации на уровне признаков, напрямую объединяющей информативные признаки, извлеченные из каждого типа данных. При слиянии на уровне признаков объединение информации происходит после выделения признаков. Раннее слияние может использовать корреляцию между множеством признаков из разных модальностей, затем объединяя признаки в один общий вектор. Этот процесс называется интеграцией признаков. Далее следует процесс моделирования, когда данные в виде интегрального вектора признаков обрабатываются и формируется финальное решение о гипотезе распознавания. В раннем слиянии используется только одна фаза моделирования, что делает эту стратегию более простой по сравнению с поздним и гибридным слиянием, которые нуждаются в большем количестве процессов моделирования.

Типичная схема раннего слияния, основанная на признаках, представлена на

рисунке 17.

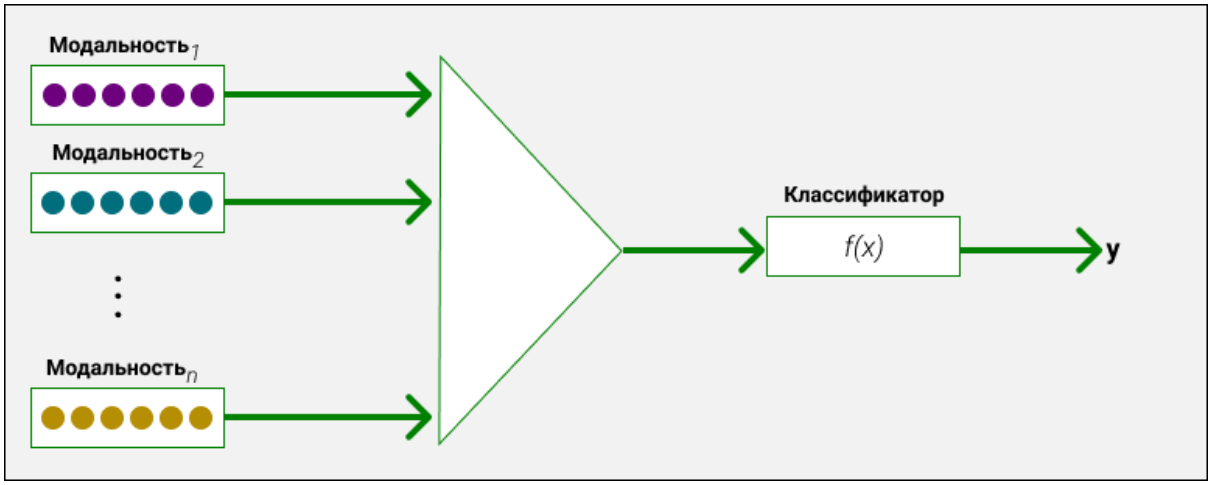


Рис. 17. Архитектура стратегии раннего слияния

В случае, когда признаки не являются независимыми, хороший алгоритм слияния на уровне признаков позволяет использование корреляционной составляющей в более полной мере, что способствует достижению лучшей эффективности стратегии. Однако слияние на этом уровне является сложной задачей с практической точки зрения в силу ряда причин:

1. Векторы признаков от разных модальностей могут быть несовместимы и в таком случае требуют этапа нормализации для унификации данных;
2. Соотношения между множествами значений различных векторов признаков могут быть неизвестны;
3. Объединение нескольких векторов признаков может привести к образованию вектора с чрезмерно большой размерностью, что может сильно усложнить процесс моделирования;
4. Может потребоваться значительно более сложная модель для обработки интегрального вектора признаков.

Позднее слияние объединяет несколько модальностей на уровне принятия решений. Эта схема более гибкая с точки зрения представлений признаков и процесса моделирования, так как отдельный процесс моделирования принимает признаки только одной модальности в качестве входных данных, что также упрощает прогнозирование, когда отсутствует одна или несколько модальностей, хотя позднее слияние игнорирует взаимодействие низкоуровневых признаков

между модальностями.

Затем решения интегрируются, и финальное решение о гипотезе распознавания принимается блоком интеграции решений. Наиболее простыми методами, используемыми на этом этапе, являются взвешивание, суммирование и голосование. Также могут быть использованы более продвинутые алгоритмы машинного обучения, такие как адаптивное усиление классификаторов (Adaptive Boosting, AdaBoost) и др. Типичная схема позднего слияния представлена на рисунке 18.

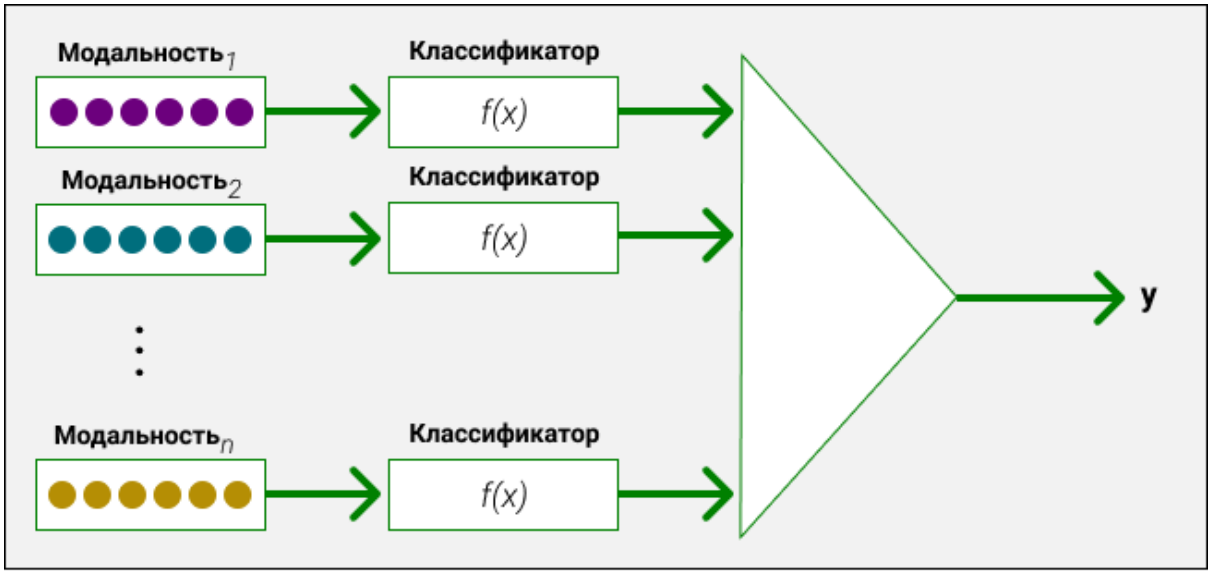


Рис. 18. Архитектура стратегии позднего слияния

Сначала выявляются особенности модальностей, после чего процесс обучения (классификации) каждой модальности выполняется отдельно. Затем результаты этих классификаций объединяются в соответствующее представление [46]. В отличие от раннего слияния, этот тип слияния является более простым, поскольку он использует обработанные и интерпретированные данные.

В позднем способе интеграции выходы процессов моделирования имеют сходные представления гипотез распознавания, и объединить их легче, чем объединить вектора признаков, как это делается при ранней интеграции. Кроме того, обработать асинхронность модальностей легче на уровне принятия решений. Такая система является более масштабируемой по числу модальностей по сравнению со способом ранней интеграции. Еще одно преимущество этого подхода состоит в том, что для каждой конкретной модальности могут быть подобраны

соответствующие методы обработки.

Стратегия объединения на уровне принятия решений имеет много преимуществ по сравнению с объединением признаков. Она обеспечивает масштабируемость с точки зрения модальностей, используемых в процессе слияния, чего трудно достичь при слиянии на уровне признаков [42]. Другое преимущество стратегии позднего слияния состоит в том, что она позволяет нам использовать наиболее подходящие методы для анализа каждой отдельной модальности, такие как скрытая модель Маркова (HMM и машина опорных векторов (SVM). Это обеспечивает гораздо большую гибкость, чем раннее слияние.

Следует отметить также и недостаток подхода позднего слияния. Он заключается в том, что метод не использует возможности корреляции между признаками модальностей. Более того, поскольку для получения локальных решений используются разные классификаторы, процесс обучения для них становится утомительным и отнимает много времени.

Основным недостатком способа поздней интеграции является то, что невозможно извлечь непосредственную выгоду из корреляции модальностей на уровне признаков. Кроме того, из-за необходимости раздельного моделирования каждой модальности поздняя интеграция является более сложной в реализации по сравнению с ранней интеграцией.

Как уже говорилось выше, каждый тип интеграции имеет свои плюсы и минусы. Некоторые работы предлагают объединять эти подходы и получать выгоду из преимуществ обоих. Такой подход обычно называют гибридной интеграцией, когда используется комплексирование методов ранней и поздней интеграции.

Гибридное слияние, как было сказано выше, объединяет две стратегии, предполагая сочетание подходов как на уровне признаков, так и на уровне принятия решений, объединяя возможности корреляции признаков и возможности использования различных алгоритмов (рис. 19).

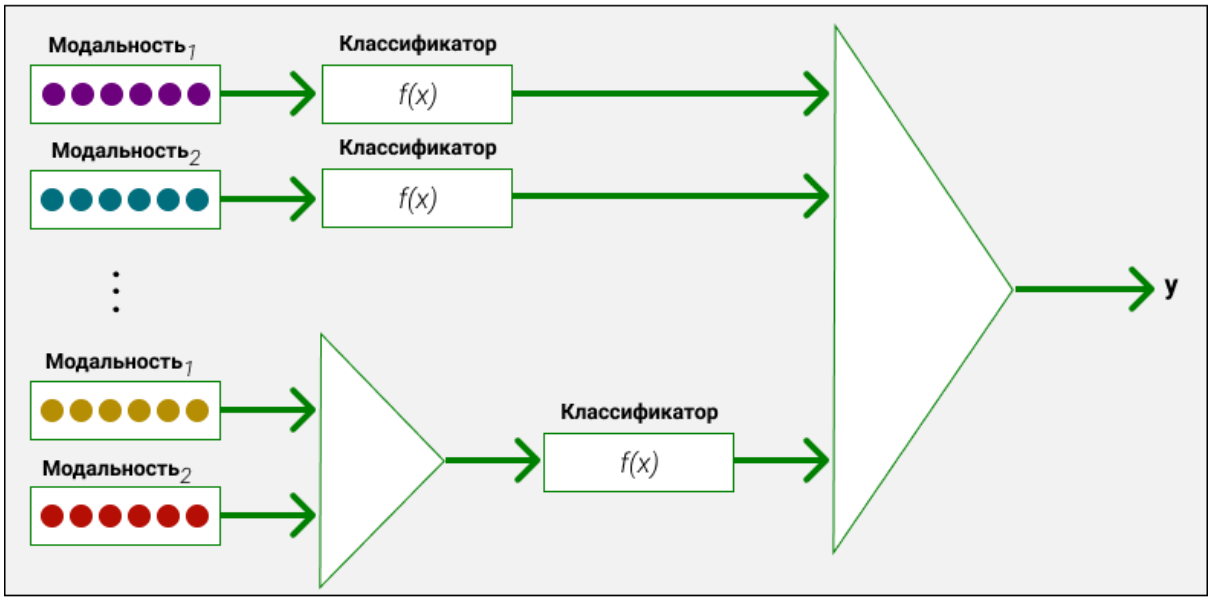


Рис. 19. Архитектура стратегии гибридного слияния

Затем для получения окончательного результата распознавания используются решения обеих систем в сочетании с блоком интеграции решений.

1.2. Анализ состояния методов использующиеся при слиянии мультимодальных данных

1.2.1. Тензорный слой слияния

Один из наиболее простых подходов к слиянию мультимодальных данных конкатенация признаков, в котором исследуется взаимодействие между модальностями:

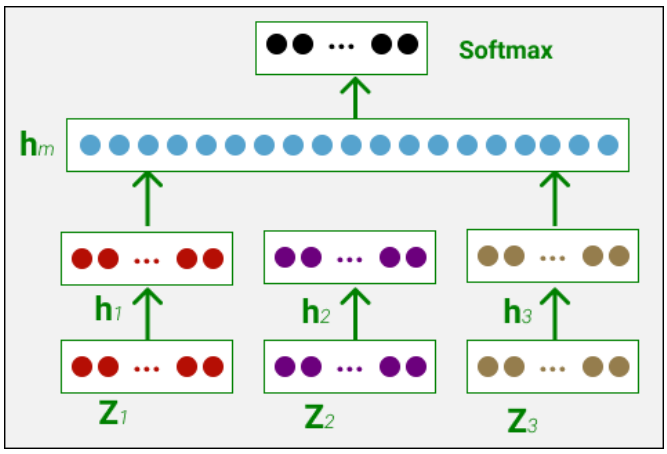


Рис. 20. Принципы работы конкатенации слияния

где $h_m = f(W.[h_{1,\dots}, h_n])$

Одной из важных проблем в подходе, где используется конкатенация элементов, заключается в том, что слияние происходит на уровне признаков и мы не учитываем никаких семантических отношений между модальностями[43], [44].

Отдельные модальности могут не предоставлять никаких содержательных информации, и семантические связи появляются только в следствие интеграции различных модальностей [45], так как позволяет увидеть влияние модальностей друг на друга с учетом возникновения дополнительной информации [46].

Далее мы будем рассматривать подходы, которые учитывают подобные взаимосвязи между модальностями. В предыдущих работах подобный подход был невозможен, только с помощью тензорного подхода и использования внутреннего произведения векторов признаков возможно создание N-мерного тензора, где N равно количеству модальностей.

$$A_1 \times A_2 \times \dots \times A_n = \{(A_1, A_2, \dots, A_n) a_i \in A_i \text{ для каждой } i \leq i \leq n\}$$

Благодаря этому подходу мы можем использовать семантические связи между модальностями и такой тип интеграции данных способствует получению более высокого выходного результата.

Первый тип подходов к мультимодальному слиянию был представлен в работах Zadeh, разработавшего алгоритм TFN [47] (см. рис. 21).

Для упрощения понимания метода сначала рассмотрим тензорное слияние для двух модальностей (x_A, x_B), чтобы смоделировать уникальную модель взаимодействия мультимодальных данных.

x_A – вектор признаков модальности A

x_B – вектор признаков модальности B

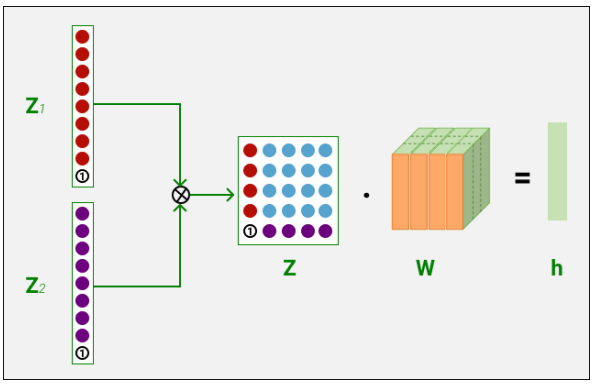


Рис. 21. Графическое представление алгоритма тензорного слияния (TFN)

Чтобы объединить аддитивные (унимодальные) и мультипликативные (бимодальные) взаимодействия, добавляется фиктивная единица к каждому вектору признаков, которая определяется следующим образом:

$$\tilde{x}_A = \begin{bmatrix} x_A \\ 1 \end{bmatrix}, \tilde{x}_B = \begin{bmatrix} x_B \\ 1 \end{bmatrix}$$

Затем строится их внешнее произведение:

$$Z = \tilde{x}_A \otimes \tilde{x}_B$$

$$Z = \begin{bmatrix} x_A \\ 1 \end{bmatrix} \otimes \begin{bmatrix} x_B \\ 1 \end{bmatrix} = \begin{bmatrix} x_A & x_A \otimes x_B \\ 1 & x_B \end{bmatrix}$$

В полученном тензоре Z представлены признаки различных порядков:

– Признаки нулевого порядка (константа) – соответствуют произведению добавленных единиц и играют роль смещения;

– Признаки первого порядка (унимодальные, additive) – получаются при перемножении исходных признаков одной модальности с единицей из других модальностей; они отражают аддитивный вклад каждой модальности по отдельности.

– Признаки второго порядка (бимодальные, multiplicative) – формируются как произведения признаков двух различных модальностей; они позволяют моделировать парные взаимодействия.

– Признаки третьего и более высоких порядков (мультимодальные) – отражают совместные взаимодействия трёх и более модальностей.

Эту формулу можно расширить до большего количества модальностей, только необходимо делать внешнее произведение на всех модальностях. Далее мы рассмотрим этот расчет для трех модальностей:

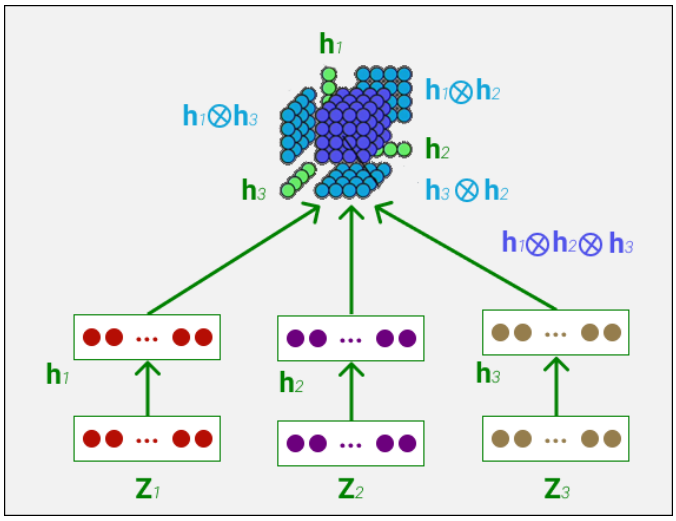


Рис. 22. Графическое представление алгоритма TFN для трех модальностей

где $h_m = \begin{bmatrix} x_A \\ 1 \end{bmatrix} \otimes \begin{bmatrix} x_B \\ 1 \end{bmatrix} \otimes \begin{bmatrix} x_C \\ 1 \end{bmatrix}$, а z ∈ R является трехмерным кубом всех возможных комбинаций унимодальных представлений друг с другом [48]. Здесь уже появляются: унимодальные признаки (рёбра куба), бимодальные взаимодействия (грани куба), тримодальные взаимодействия (объём куба) и константа (дальний угол).

Так как тензорное слияние математически образованное внешнее произведение, он не имеет обучающих параметров, и несмотря на то, что размер матрицы очень большой, у модели не создается проблемы переобучения. Это связано с тем фактом, что выходные нейроны тензорного слияния легко интерпретируются и семантически очень значимы. Таким образом, последующие уровни сети могут легко декодировать значимую информацию.

Применение подхода Tensor Fusion Network (TFN) на мультимодальном датасете CMU-MOSI показало значительное улучшение качества анализа по сравнению с существующими методами. Так, для бинарной классификации настроений модель достигла точности 77,1% и F1-метрики 77,9%, что на 4% и 2,7% выше по сравнению с предыдущим state-of-the-art [44], [47]. Для более сложной

пятиступенчатой классификации точность составила 42,0%, что на 6,7% выше лучших аналогов, а при решении задачи регрессии средняя абсолютная ошибка снизилась с 1,10 до 0,87, а коэффициент корреляции увеличился с 0,53 до 0,70.

Результатом этой операции является создание очень больших выходных данных, очень большого количества измерений в объединенном слое, и поэтому его можно реально применить к задачам, где пространство взаимодействия не слишком велико [97].

Размерность признакового пространства при тензорном слиянии растёт экспоненциально: для двух модальностей она равна $(d_A + 1)(d_B + 1)$, для трёх $(d_A + 1)(d_B + 1)(d_c + 1)$ например, если каждая модальность имеет размерность 4, то для двух модальностей получаем матрицу $5 \times 5 = 25$, а для трёх, уже куб $5 \times 5 \times 5 = 125$ что приводит к необходимости использовать в последующих слоях модели весьма крупные матрицы весов и обуславливает значительные вычислительные затраты:

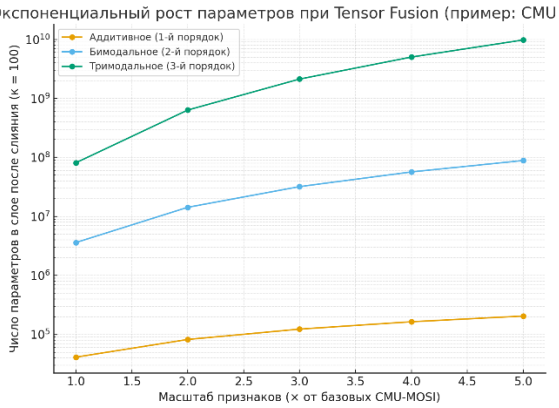


Рис. 23. Экспоненциальный рост числа параметров при слиянии признаков методом Tensor Fusion

На примере известного мультимодального датасета CMU-MOSI (типичные размерности признаков: текст – 300, аудио – 74, видео – 35) показан рост числа параметров в следующем полносвязном слое с $k = 100$ нейронами после тензорного слияния с добавленной единицей; кривые отображают аддитивное (1-й порядок), бимодальное (2-й порядок) и тримодальное (3-й порядок) слияние при масштабировании исходных размерностей в $1 \times \ldots \times 5 \times$: видно, что при переходе к полному тримодальному тензору число параметров растёт кубически с

порядка 10^8 до порядка 10^{10} , тогда как аддитивный случай увеличивается лишь линейно, что наглядно иллюстрирует проблему размерности и экспоненциальный (по числу модальностей/порядку взаимодействия) рост вычислительных затрат.

1.2.2. Низкоранговое слияние

Подход TNF слияния сопряжён со значительными вычислительными затратами, поскольку формирование результирующего тензора и последующее его преобразование через весовую матрицу требует работы с высокоразмерными объектами. Для преодоления этого ограничения в литературе предложен метод низкорангового слияния (low-rank fusion), основанный на аппроксимации весовых матриц и тензоров с использованием их низкоранговых разложений [49].

Идея метода опирается на фундаментальный результат линейной алгебры: любая матрица может быть представлена как сумма внешних произведений векторов, а число таких слагаемых определяется её рангом. Рангом матрицы $A \in \mathbb{R}^{m\times n}$ называется максимальное число её линейно независимых строк (или столбцов). Эквивалентно, ранг равен минимальному числу ранга-1 матриц (матриц, представимых в виде внешнего произведения векторов), сумма которых восстанавливает исходную матрицу:

$$rank(A) = min\left\{r|\, A = \sum_{i=1}^{r} u_i v_i^T , u_i \in \mathbb{R}^m, \qquad v_i \in \mathbb{R}^n \right\}$$

N-мерный тензор может быть строго разложен на внешние произведения N-векторов. Это означает, что мы вносим различие в масштабирование субтензора, так как добавляем большее количество измерений при построении полного тензора [50] . Следовательно, матрица ранга 1 может быть записана как $X = a \circledcirc b$, для тензора 3-х мерного порядка пишем формулу $\mathrm{X} = a \circledcirc b \circledcirc c$. Графическое представление концепции ранга 1 представлено на рисунке 24.

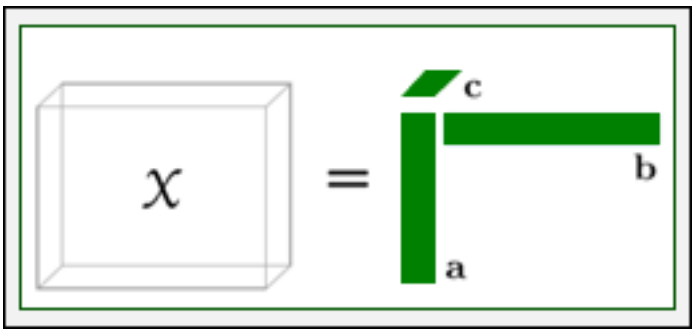


Рис. 24. Графическое представление концепции ранга 1 для тензора

Ранг тензора $rank(X) = R$ определяется как минимальное число ранга первого тензора[51], которое необходимо для производства тензора X, как его суммы. Ранг-R для матрицы может быть переформулирован как $X = \sum_{r=1}^{R} \lambda_r a_r \circledcirc b_r = [\![\lambda; A, B]\!]$, и ранг-R для тензора трехмерного порядка определяется по следующей формуле $X = \sum_{r=1}^{R} \lambda_r a_r \circledcirc b_r \circledcirc bc_r = [\![\lambda; A, B, C]\!]$. В общем виде N-мерный тензор рассчитывается по следующей формуле:

$$X = \sum_{r=1}^{R} \lambda_r {a_r}^{(1)} \circledcirc {b_r}^{(2)} \circledcirc \dots \circledcirc {a_r}^{(N)} = [\![\lambda; A^{(1)}, A^{(2)}, \dots, A^{(N)}]\!]$$

Фактор матрицы относится к комбинации векторов из компонентов ранга первого тензора. Поэтому A имеет форму $A = [a_1, a_2, \dots, a_R]$. Мы ввели новый дополнительный коэффициент λ_r, который часто используется для поглощения соответствующих весов во время нормализации компонентов матрицы. Как правило, это означает нормализацию суммы квадратов элементов и каждого компонента по одному.

Иными словами, ранг тензора определяет минимальное число мультипликативных компонент, необходимых для его точного восстановления.

В контексте мультимодальной обработки признаков пусть заданы два вектора признаков:

z_v-визуальная модальность и z_l- текстовая модальность их внешнее произведение образует матрицу

$$Z = z_v \otimes z_l$$

которая далее должна быть преобразована в выходное представление h посредством умножения на параметризованный тензор весов W

$$h = W \,.Z$$

Прямое вычисление данного выражения требует работы с высокоразмерным тензором W, что приводит к существенным затратам памяти и времени. Поскольку $\mathcal{Z}$ является тензором порядка M (где M - число модальностей ввода), вес $\mathcal{W}$ будет естественным образом тензором порядка- (M + 1) в $\mathbb{R}^{d_1 \times d_2 \times \dots d_M \times d_h}$.

Дополнительный (M +1) -й размер соответствует размеру выходного представления d_h. В произведении тензорных точек $\mathcal{W} \cdot \mathcal{Z}$ весовой тензор $\mathcal{W}$ можно рассматривать как тензоры d_h порядка-M. Другими словами, вес $\mathcal{W}$ можно разбить на $\widetilde{\mathcal{W}}_k \in \mathbb{R}^{d_1 \times d_2 \times \dots d_M}, k = 1, \dots, d_h$. Каждый $\widetilde{\mathcal{W}}_k$ вносит вклад в одно измерение в выходном векторе h, т.е. $h_k = \widetilde{\mathcal{W}}_k . \mathcal{Z}$

Эта интерпретация тензорного слияния показана на рисунке 57 для бимодального случая [52].

Размерность $\mathcal{Z}$ будет расти экспоненциально с увеличением количества модальностей при $\prod_{m-1}^{M} d_m$. Число параметров для изучения в тензоре веса $\mathcal{W}$ также будет расти экспоненциально. Это не только вводит много вычислений, но и подвергает модель рискам переобучения.

LMF параметризовал g (·) из уравнения 2 с набором модально-зависимых факторов низкого ранга, которые можно использовать для восстановления весового тензора низкого ранга, в отличие от полного тензора $\mathcal{W}$.

Кроме того, при разложении веса на множество факторов низкого ранга, можно использовать тот факт, что тензор $\mathcal{Z}$ на самом деле разлагается на $\{\mathcal{Z}_m\}_{m=1}^{M}$, что позволяет напрямую вычислять выходной вектор. LMF уменьшает число параметров, а также сложность вычислений, связанных с тензорностью, от экспоненциального по M до линейного.

Итак, идея LMF состоит в том, чтобы разложить весовой тензор $\mathcal{W}$ на M наборов модальных факторов. Однако, поскольку $\mathcal{W}$ сам является тензором порядка (M + 1), обычно используемые методы разложения приводят к M + 1 частям. Следовательно, мы по-прежнему придерживаемся представления,

обозначенного выше, что $\mathcal{W}$ образован тензорами d_h -М порядка $\widetilde{\mathcal{W}}_k \in \mathbb{R}^{d_1 \times d_2 \times \dots d_M}, k = 1, \dots, d_h$, сложенными вместе. Затем мы можем разложить каждый $\widetilde{\mathcal{W}}_k$ отдельно.

Для тензора порядка М $\widetilde{\mathcal{W}}_k \in \mathbb{R}^{d_1 \times d_2 \times \dots d_M}$всегда существует точное разложение на векторы в виде:

$$\widetilde{\mathcal{W}}_k = \sum_{i=1}^{R} \otimes_{m=1}^{M} \mathcal{W}_{m,k}^{(i)} \ , \mathcal{W}_{m,k}^{(i)} \in \mathbb{R}_{,}^{d}$$

Минимальное R, которое делает разложение действительным, называется рангом тензора. Наборы векторов $\left\{\left\{\mathcal{W}_{m,k}^{(i)}\right\}_{i=1}^{M}\right\}_{i=1}^{R}$ называются коэффициентами разложения ранга R исходного тензора. В LMF мы начинаем с фиксированного ранга r и параметризуем модель с r коэффициентами разложения $\left\{\left\{\mathcal{W}_{m,k}^{(i)}\right\}_{i=1}^{M}\right\}_{i=1}^{r}, k = 1, .., d_h$, которые могут быть использованы для реконструкции низкоранговой версии этих $\widetilde{\mathcal{W}}_k$.

Мы можем перегруппировать и объединить эти векторы в М-модальные факторы низкого ранга. Пусть $\mathcal{W}_{m,k}^{(i)} = \left[\mathcal{W}_{m,1}^{(i)}, \mathcal{W}_{m,2}^{(i)}, \dots, \mathcal{W}_{m,d_h}^{(i)}\right]$, то для модальности m, $\left\{\mathcal{W}_{m}^{(i)}\right\}_{i=1}^{r}$ - это соответствующие факторы низкого ранга. Таким образом, мы можем восстановить тензор низкого ранга:

$$\mathcal{W} = \sum_{i=1}^{r} \otimes_{m=1}^{M} \mathcal{W}_{m}^{(i)}$$

Следовательно, уравнение 2 может быть вычислено следующим образом:

$$h = \left(\sum_{i=1}^{r} \otimes_{m=1}^{M} \mathcal{W}_{m}^{(i)}\right).\mathcal{Z}$$

Для всех m $\mathcal{W}_{m}^{(i)} \in \mathbb{R}^{d_m \times d_h} \in$ имеет одинаковый размер. Мы определяем их внешнее произведение, чтобы оно охватывало только те измерения, которые не являются общими: $\mathcal{W}_{m}^{(i)} \otimes \mathcal{W}_{n}^{(i)} \in \mathbb{R}^{d_m \times d_n \times d_h}$. Бимодальный пример этой процедуры показан на рисунке 25.

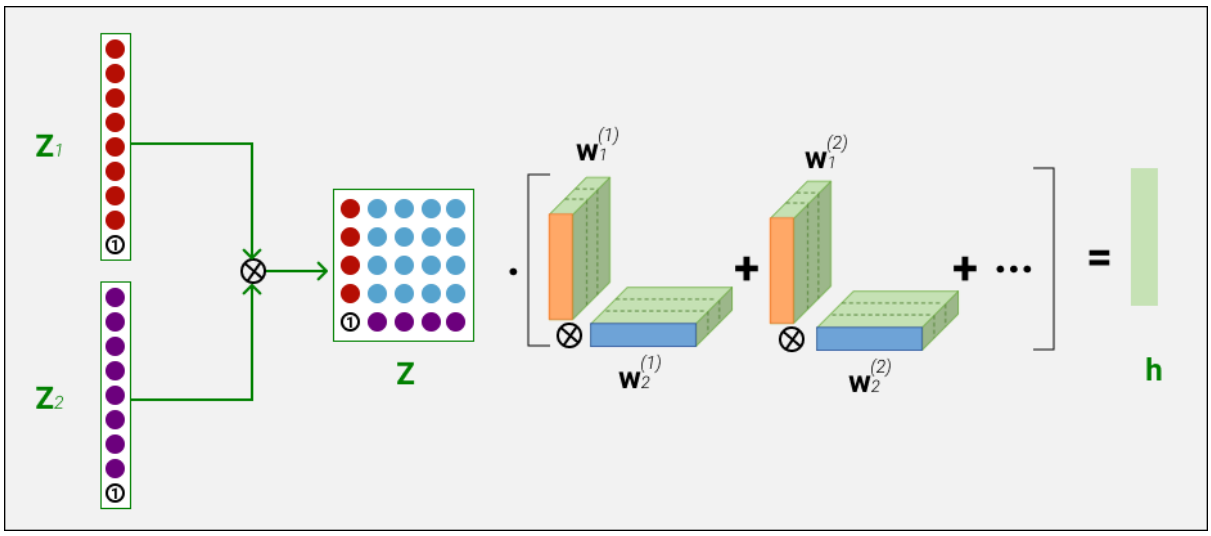


Рис. 25. Графическое представление алгоритма низкого ранга

Мы введем эффективную процедуру для вычисления h, используя тот факт, что тензор Z естественным образом разлагается на исходный вход $\{\mathcal{Z}_m\}_{m=1}^{M}$, который параллелен зависящим от модальности факторам низкого ранга.

Используя тот факт, что $\mathcal{Z} = \otimes_{m=1}^{M} \mathcal{Z}_m$, мы можем упростить уравнение ниже:

$$h = \left(\sum_{i=1}^{r} \otimes_{m=1}^{M} \mathcal{W}_m^{(i)}\right).\mathcal{Z}$$

$$= \sum_{i=1}^{r} \left(\otimes_{m=1}^{M} \mathcal{W}_m^{(i)}.\mathcal{Z}\right)$$

$$= \sum_{i=1}^{r} \left(\otimes_{m=1}^{M} \mathcal{W}_m^{(i)}. \otimes_{m=1}^{M} \mathcal{Z}_m\right)$$

$$= \bigwedge_{m=1}^{M} \left[\sum_{i=1}^{r} \mathcal{W}_m^{(i)}.\mathcal{Z}_m\right]$$

где $\bigwedge_{m=1}^{M}$ обозначает поэлементное произведение последовательности тензоров: $\bigwedge_{t=1}^{3} x_t = x_1 \circ x_2 \circ x_3$. Мы также можем вывести уравнение 6 для бимодального случая:

$$h = \left(\sum_{i=1}^{r} \otimes_{m=1}^{M} \mathcal{W}_m^{(i)}\right).\mathcal{Z}$$

Важным аспектом этого упрощения является то, что оно использует параллельное разложение как Z, так и W, так что мы можем вычислить h без фактического создания тензора Z из входных представлений z_m. Кроме того, различные модальности не связаны при вычислении h, что позволяет легко использовать подход для произвольного числу модальностей.

Добавить новую модальность можно просто, добавив другой набор факторов, специфичных для модальности, и расширить уравнение 7.

Используя уравнение 6, мы можем вычислить h непосредственно из входных унимодальных представлений и их специфичных для модальности факторов разложения, избегая сложностей при вычислении большого входного тензора Z и W, а также путем линейного преобразования r.

Вместо этого входной тензор и последующая линейная проекция вычисляются неявно вместе в уравнении 6, и это гораздо более эффективно, чем оригинальный метод. Действительно, LMF уменьшает сложность вычисления тензорности и слияния с $\mathcal{O}(d_y \times r \times \sum_{m=1}^{M} d_m)$до $\mathcal{O}(d_y \prod_{m=1}^{M} d_m)$. На практике используется немного иная форма уравнения 6, где объединяются коэффициенты низкого ранга в M тензоры порядка 3 и меняется порядок, в котором делается поэлементное произведение, а суммирование выполняется по первому измерению матрицы в скобках. [·] i, : указывает i-й фрагмент матрицы. Таким образом, можно параметризовать модель с M тензорами порядка 3 вместо параметризации с помощью наборов векторов.

Процесс перехода от тензорного слияния к низкоранговому можно рассматривать в три этапа:

1. перестройка вычисления выходного вектора h без явного формирования тензора Z;
2. декомпозиция входного тензора;
3. декомпозиция весовой матрицы W.

Экспериментальные исследования подтверждают, что получаемые весовые матрицы действительно обладают низким рангом: даже при обучении полных матриц с последующим сингулярным разложением наблюдается быстрое

затухание сингулярных значений. Это указывает на то, что использование низкоранговой аппроксимации является не только вычислительно эффективным, но и теоретически обоснованным подходом.

Чтобы количественно проиллюстрировать данное преимущество, был проведён сравнительный эксперимент с использованием типичных размерностей признаков для датасета CMU-MOSI (текстовые признаки $d_{t=300}$, аудио $d_a = 74$, видео $d_v = 35$) и последующего полносвязного слоя на $k = 100$ выходных нейронов.

На рис. 26 показана зависимость числа параметров от масштаба входных признаков для метода Tensor Fusion и его низкоранговой модификации.

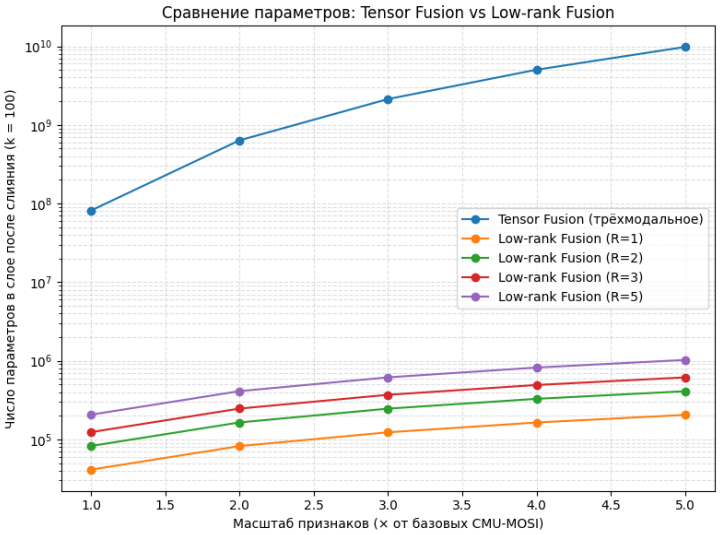


Рис. 26. Сравнение числа параметров при методах Tensor Fusion и Low-rank Fusion

Результаты демонстрируют экспоненциальный рост числа параметров в случае тримодального Tensor Fusion: уже при увеличении размерности входов в 2 раза количество параметров превышает 10^9. В то же время Low-rank Fusion демонстрирует линейную зависимость от размерности и числа факторов R, даже при R=5 количество параметров остаётся на несколько порядков ниже, чем в случае полного тензорного слияния.

Таким образом, экспериментальные данные подтверждают теоретические ожидания: метод Low-rank Fusion позволяет значительно сократить вычислительные и параметрические издержки при сохранении способности

моделировать сложные мультипликативные взаимодействия между модальностями.

Эти результаты подтверждаются и практическими экспериментами. В работе Liu и соавт. (2018) метод Low-rank Multimodal Fusion был протестирован на трёх задачах: мультимодальный анализ сентимента (CMU-MOSI), распознавание характеристик говорящего (POM) и распознавание эмоций (IEMOCAP). Полученные результаты показали, что LMF обеспечивает не только существенное сокращение числа параметров и времени обучения, но и превосходит или как минимум не уступает Tensor Fusion по качеству. Например, на датасете CMU-MOSI метод LMF достиг ошибки MAE = 0.912 и корреляции = 0.668, что лучше по сравнению с TFN (MAE = 0.970, Corr = 0.633). На задаче распознавания эмоций (IEMOCAP) LMF также превзошёл TFN [47], обеспечив лучшие значения F1-меры для большинства эмоциональных классов

Таким образом, можно выделить два ключевых преимущества метода низкорангового слияния:

1. Вычислительная эффективность. Число параметров и время обучения уменьшаются на порядки за счёт линейной, а не экспоненциальной зависимости от числа модальностей и их размерностей. В частности, LMF обучается более чем в три раза быстрее TFN (1134 против 340 сегментов в секунду).

2. Высокая точность. Несмотря на значительное сокращение ресурсов, метод демонстрирует сопоставимые или лучшие результаты по сравнению с наиболее сильными тензорными подходами, что делает его практически применимым для широкого спектра мультимодальных задач.

В совокупности это позволяет рассматривать Low-rank Fusion как оптимальный компромисс между выразительностью тензорных моделей и ограничениями по вычислительным ресурсам, обеспечивая баланс между качеством и эффективностью в реальных приложениях.

1.2.3. High-Order Polynomial Fusion

Задача мультимодальной интеграции данных заключается в построении единого признакового представления на основе разнородных источников информации (например, текстовых, аудиальных и визуальных сигналов). Ключевая трудность при этом состоит в необходимости моделирования высокоуровневых нелинейных взаимосвязей между модальностями, выходящих за рамки простых билинейных или трилинейных взаимодействий[53].

Для решения данной проблемы был предложен блок полиномиального тензорного объединения [54], позволяющий явно учитывать высокие порядки взаимодействия признаков. Пусть задан набор векторов признаков различных модальностей:

$$\{z_m\}_{m=1}^{M}, z_m \in \mathbb{R}^{d_m}$$

Где M число модальности, d_m размерность признакового пространства m-й модальности. Эти векторы конкатенируются в единый признак:

$$f^T = [1, z_1^T, z_2^T, \dots, , z_M^T]$$

где добавленный константный элемент «1» обеспечивает корректное представление полиномиальных членов.

Далее формируется тензор признаков порядка P:

$$F = \underbrace{f \otimes f \otimes \cdots \otimes f}_{P}$$

Таким образом, тензор F содержит все возможные полиномиальные взаимодействия до порядка P. Объединённое представление вычисляется посредством свёртки с тензором весов W:

$$z_h = \sum_{i_1, i_2, \dots, i_P}^{R} W^h_{i_1, i_2, \dots, i_P} F_{i_1, i_2, \dots, i_P}$$

где h индекс в пространстве выходных признаков размерности H.

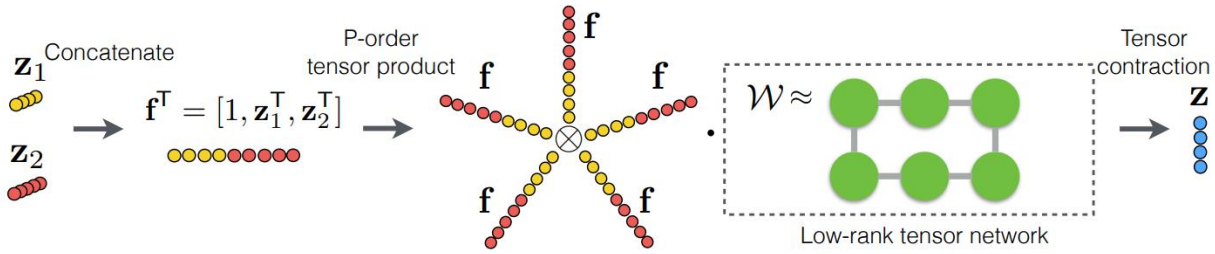


Рис. 27. Схема блока полиномиального тензорного объединения пятого порядка (PTP) для слияния признаков

Так как количество параметров W растёт экспоненциально при увеличении порядка P, для практической реализации используется аппроксимация низким рангом (например, в формате CP-разложения):

$$z_h = \sum_{r=1}^{R} a_{hr} \prod_{p=1}^{P} \left(\sum_{i_P} w_{r,i_P}^{(p)} f_{\,i_P} \right)$$

где R – ранг разложения, a_{hr} коэффициенты линейной косбинации, $w_{r,i_P}^{(p)}$ параметры модальных факторов. Таким образом, математическая модель позволяет явно учитывать полиномиальные взаимодействия между модальностями и контролировать сложность за счёт низкоранговых аппроксимаций [54], [55]. В итоге получается новое представление, которое кодирует скрытые взаимодействия между модальностями.

Иерархическая полиномиальная сеть объединения (Hierarchical Polynomial Fusion Network, HPFN) представляет собой нейросетевую архитектуру, предназначенную для мультимодальной интеграции данных. Ключевая идея состоит в том, что базовым строительным блоком HPFN является операция полиномиального тензорного объединения (Polynomial Tensor Pooling, PTP), которая обеспечивает моделирование взаимодействий признаков высоких порядков:

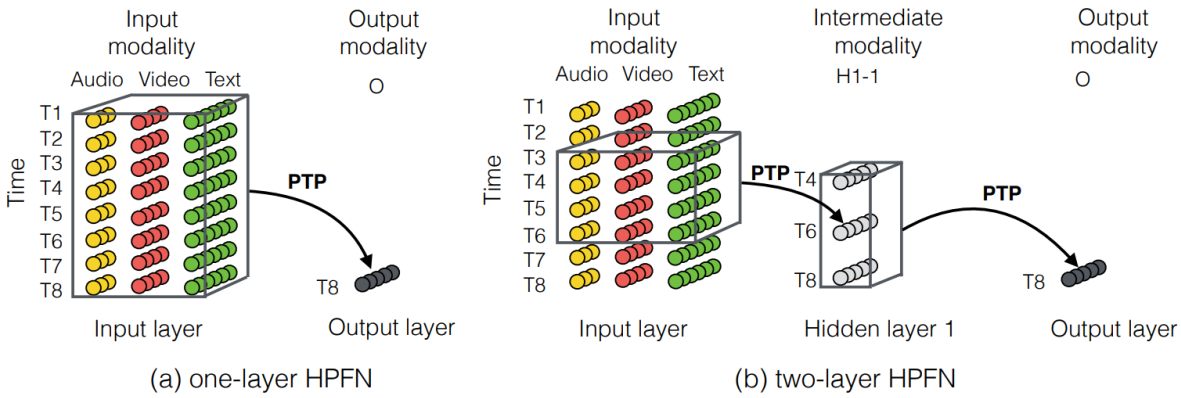


Рис. 28. Иллюстративные примеры архитектур PTP и HPFN

PTP-блок выполняет тензорное произведение конкатенированных признаков различных модальностей и последующую свёртку с обучаемым тензором весов, что позволяет явно учитывать полиномиальные взаимодействия между входными сигналами. В то время как PTP представляет собой отдельную операцию мультимодального объединения, архитектура HPFN строится как сеть из нескольких таких блоков, соединённых в иерархической структуре. На последующих уровнях новые PTP-блоки обрабатывают не только исходные данные, но и уже полученные скрытые признаки, что позволяет постепенно формировать всё более абстрактные и информативные представления. В результате иерархическая организация HPFN обеспечивает улавливание как локальных зависимостей между модальностями, так и глобальных корреляций, возникающих на более высоких уровнях [53].

$$z^{(l)} = PTP\left(z^{(l-1)}\right), l = 1, \dots, L$$

где $z^{(0)}$ исходные признаки модальностей, $z^{(l)}$ финальное мультимодальное представление.

Трёхслойная HPFN. Последовательное применение трёх PTP-блоков. Первый слой агрегирует аудио-, видео- и текстовые признаки; второй слой работает с промежуточными модальными векторами; третий формирует итоговое представление. Плотно-связанная HPFN (Dense-HPFN). В этой архитектуре используется dense connectivity: каждый слой получает на вход не только выход

предыдущего уровня, но и признаки всех предшествующих. Это увеличивает переиспользование информации и улучшает распространение градиентов. Формально:

$$z^{(l)} = PTP\left(\bigoplus_{j=0}^{l-1} z^{(j)}\right)$$

где $\oplus$ обозначает операцию конкатенации. Данный механизм повышает устойчивость сети, облегчает распространение градиентов и позволяет эффективнее использовать промежуточные признаки.

Экспериментальные результаты показывают, что глубина сети и наличие плотных связей существенно влияют на эффективность мультимодального объединения.

– Трёхслойная HPFN успешно улавливает локальные и глобальные межмодальные корреляции, однако чрезмерное увеличение глубины может привести к переобучению.

– Плотно-связанная HPFN обеспечивает более высокую точность за счёт прямой передачи признаков между слоями. Например, двухслойная версия с плотными связями (HPFN-L2-S2) показала Acc-7=36.5% и F1-Neutral=72.7%, что выше, чем у аналога без плотных связей (Acc-7=36.3%).

Метод был проверен на задачах анализа тональности (CMU-MOSI) и распознавания эмоций (IEMOCAP). Сравнение проводилось с современными моделями (MFN, MARN, TFN, LMF).

Основные результаты:

– На CMU-MOSI сеть HPFN при порядке P=8 достигла точности Acc-7 = 36.9%, что на 2.2% выше по сравнению с моделью MARN.

– На IEMOCAP HPFN-L2 обеспечила результаты F1-Angry = 88.8%, F1-Sad = 86.6%, F1-Happy = 86.2%, превзойдя TFN и LMF [55], [56].

– Анализ показал, что оптимальные результаты достигаются при P=3-4, после чего точность стабилизируется.

Как показывают данные экспериментов, использование PTP-блоков в составе иерархической архитектуры HPFN позволяет увеличить точность

мультимодального анализа на 2–3% по сравнению с методами, ограниченными билинейным или трилинейным объединением.

Одним из ключевых аспектов оценки архитектуры HPFN является анализ её вычислительной и параметрической сложности. В отличие от традиционных моделей мультимодального объединения на основе тензоров (TFN и LMF), предложенный подход использует симметрийные свойства тензора признаков. Это позволяет сделать количество параметров в весовом тензоре независимым от порядка P и линейно зависящим от конкатенированных смешанных признаков в «окнах».

Модель	Характер	Параметры
TFN	Высокая сложность	$\mathcal{O}\left(I_y \prod_{m=1}^{M} I_m\right)$
LMF	Низкая сложность	$\mathcal{O}\left(I_y R \sum_{m=1}^{M} IM\right)$
PTP	Средняя сложность	$\mathcal{O}\left(I_y R \sum_{t=1}^{T} \sum_{m=1}^{S} I_{t,m}\right)$
HPFN	Оптимизированная	$\mathcal{O}\left(I_y R \left(\sum_{l=1}^{L} N_l\right)\left(\sum_{t=1}^{T} \sum_{m=1}^{S} I_{t,m}\right)\right)$

Сравнение параметрической сложности моделей

I_y - размерность выходного пространства,

M - число модальностей,

R - ранг тензорного разложения,

T, S - размеры локального окна по времени и модальностям

$I_{t,m}$ - размерность признаков модальности m в момент времени t.

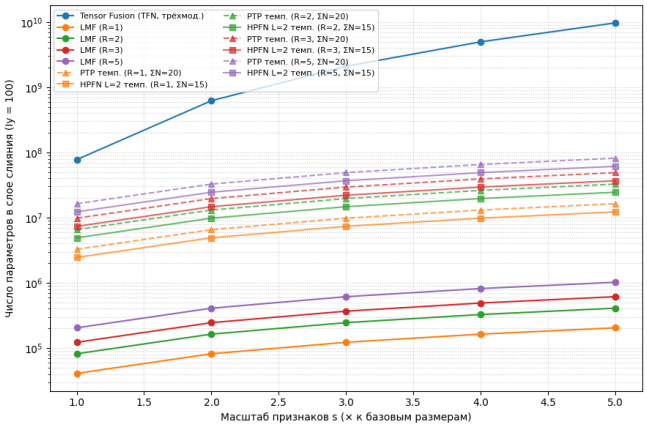


Рис. 29. Сравнение параметрической сложности: TFN, LMF, PTP (temporal), HPFN (temporal)

Таким образом, HPFN сочетает высокую предсказательную силу с приемлемой вычислительной сложностью: модель остаётся более лёгкой, чем TFN, и лишь немного уступает LMF по числу параметров, обеспечивая при этом значительный прирост точности.

1.2.4. Слияние с контролем

Одним из ключевых вызовов является построение такого объединённого представления, которое сохраняет значимые признаки каждой модальности и одновременно устраняет избыточность и шум. Традиционные методы мультимодального слияния используют фиксированные весовые матрицы, что ограничивает адаптивность и снижает качество обобщения. Для преодоления этих ограничений применяются методы слияния с контролем (gated fusion), позволяющие динамически регулировать вклад каждой модальности в результирующее представление [9], [57], [58]:

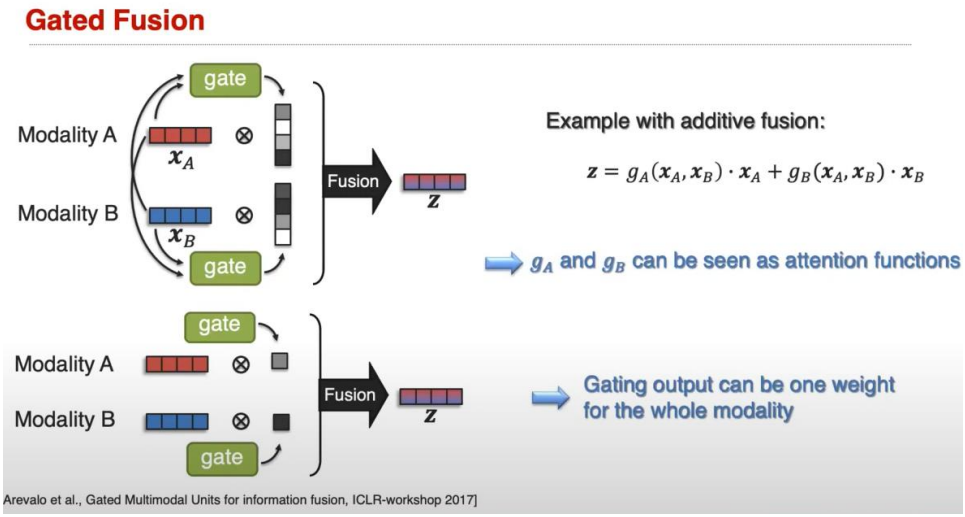


Рис. 30. Аддитивная модель слияния с контролем

Базовая формализация аддитивного слияния с контролем задаётся выражением:

$$z = g_A(x_A, x_B).x_A + g_B(x_A, x_B).x_B$$

где

$x_A \in \mathbb{R}^{d_A}$ вектор признаков модальности A (например, визуальной),

$x_B \in \mathbb{R}^{d_B}$ вектор признаков модальности B (например, аудиальной),

$g_A(x_A, x_B) \in \mathbb{R}$ управляющая функция (gate) для модальности A

$g_B(x_A, x_B) \in \mathbb{R}$ управляющая функция (gate) для модальности B

$z \in \mathbb{R}^d$ результирующее мультимодальное представление

Управляющие функции g_A и g_B интерпретируются как коэффициенты значимости модальностей, динамически изменяющиеся в зависимости от входных данных. Это позволяет отсеивать нерелевантные признаки и усиливать наиболее значимые.

Механизм внимания (attention) в общем виде описывается как процесс вычисления весов для элементов входного множества [59].

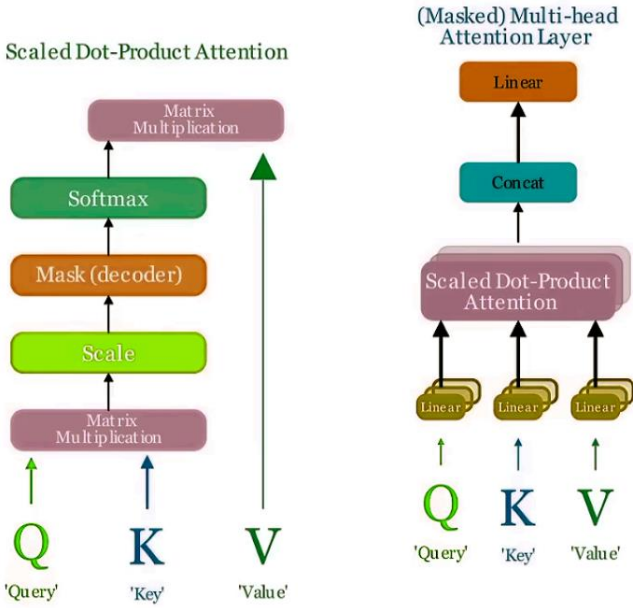


Рис. 31. Структура механизмов Scaled Dot-Product Attention и Multi-Head Attention

Пусть заданы:

запрос (query) $q \in \mathbb{R}^d$

набор ключей (keys) $K = \{k_1, \dots, k_n\}$

соответствующие значения (values) $V = \{v_1, \dots, v_n\}$

Тогда внимание определяется формулой:

$$Attention(q, K, V) = \sum_{i=1}^{n} \alpha_i v_i, \quad \alpha_i = \frac{\exp(e_i)}{\sum_{j=1}^{n} \exp(e_j)}$$

где $e_i = \text{score}(q, k_i)$ функция сходства, например скалярное произведение.

Gated fusion является частным случаем attention-механизма для двух модальностей. Здесь x_A, x_B играют роль значений, а управляющие функции g_A и g_B это веса внимания, которые зависят от входных данных. Таким образом, gated fusion можно рассматривать как специализированный attention для мультимодального объединения, где ворота (gates) выступают фильтрами, отсекающими нерелевантные признаки [60] .

Для практического использования в нейросетевых архитектурах применяется более развитая модель – Gated Multimodal Unit (GMU). Она вводит скрытые представления для каждой модальности и вычисляет управляющий вектор на основе их совместной информации.

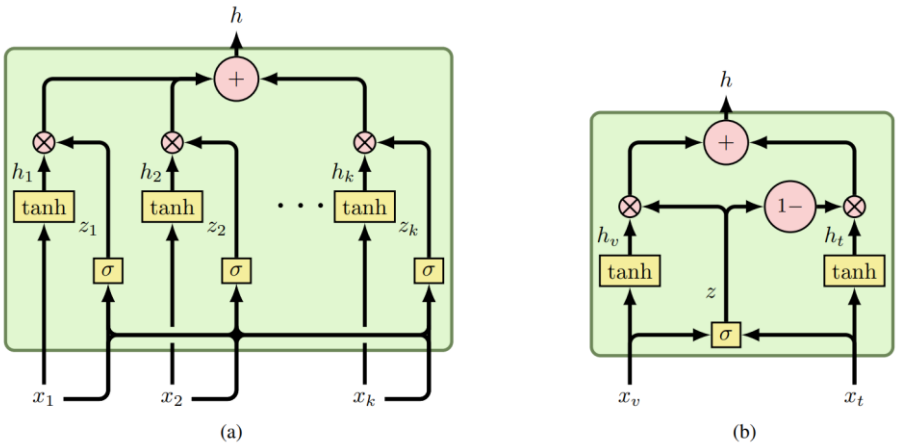


Рис. 32. Иллюстрация работы управляющих (gated) блоков a) Предлагаемая модель для объединения более чем двух модальностей b) Упрощённый вариант для бимодального подхода

Для двух модальностей (визуальной x_v и текстовой x_t) GMU описывается системой уравнений:

$$h_v = tanh(W_v . x_v)$$

$$h_t = tanh(W_t . x_t)$$

$$z = \sigma(W_z [x_v, x_t])$$

$$h = z \odot h_v + (1 - z) \odot h_t$$

где h_v, h_t скрытые представления модальностей, z правляющий вектор, определяющий относительный вклад модальностей, а h итоговое мультимодальное представление.

Для обобщённого случая k модальностей:

$$h_i = \tanh(W_i . x_i), i = 1, \dots, k$$

$$z_i = \sigma\big(W_{z_i}[x_1, \dots, x_k]\big), i = 1, \dots, k$$

$$h = \sum_{i=1}^{k} \alpha_i \odot h_i \ , \ \alpha_i = \frac{z_i}{\sum_{j=1}^{k} z_j}$$

Таким образом, GMU можно рассматривать как разновидность soft-attention механизма, встроенного внутрь архитектуры нейронной сети.

Метод GMU был протестирован на задаче мультимодальной классификации жанров фильмов [61]. Использовались постеры (визуальная модальность) и сюжетные описания (текстовая модальность).

Результаты экспериментов показали:

– макро-F1 = 0.63, что на 7% выше, чем при использовании только текста, и на 5% выше, чем при использовании только изображений;

– прирост на 5–8% по макро-F1 по сравнению с простым конкатенированием признаков;

– снижение ошибки на 10–12% относительно моделей с фиксированными весами;

– ускорение сходимости на 20–25% за счёт отбрасывания нерелевантных признаков.

Анализ данных показал, что модель динамически смещает внимание: для фильмов с минимальными визуальными различиями доминировала текстовая модальность, а для жанров с ярко выраженной визуальной спецификой (фантастика, анимация) визуальная. Таким образом, метод gated fusion можно рассматривать как частный случай механизма внимания, в котором управляющие функции выполняют роль весов (gates), динамически регулирующих вклад каждой модальности. Его развитие в виде Gated Multimodal Unit позволило повысить точность мультимодальных моделей, обеспечить устойчивость к шумам и ускорить

обучение, что делает данный подход одним из наиболее перспективных для задач анализа комплексных данных.

1.2.5. Gated Multimodal Embedding LSTM with Temporal Attention (GME-LSTM(A))

В этой работе решается задача разработки архитектуры, способной выполнять мультимодальное объединение на уровне слова, учитывая локальные и глобальные взаимодействия между модальностями и минимизируя влияние шумов[62] .

Целью данного исследования является построение модели, которая за счёт введения механизмов гейтинга (gating) и временного внимания (temporal attention) способна селективно использовать наиболее информативные модальности и ключевые моменты в потоке речи для точной оценки полярности высказываний [63], [64].

Для достижения поставленной цели был применён метод GME-LSTM Архитектура модели состоит из двух взаимосвязанных модулей:

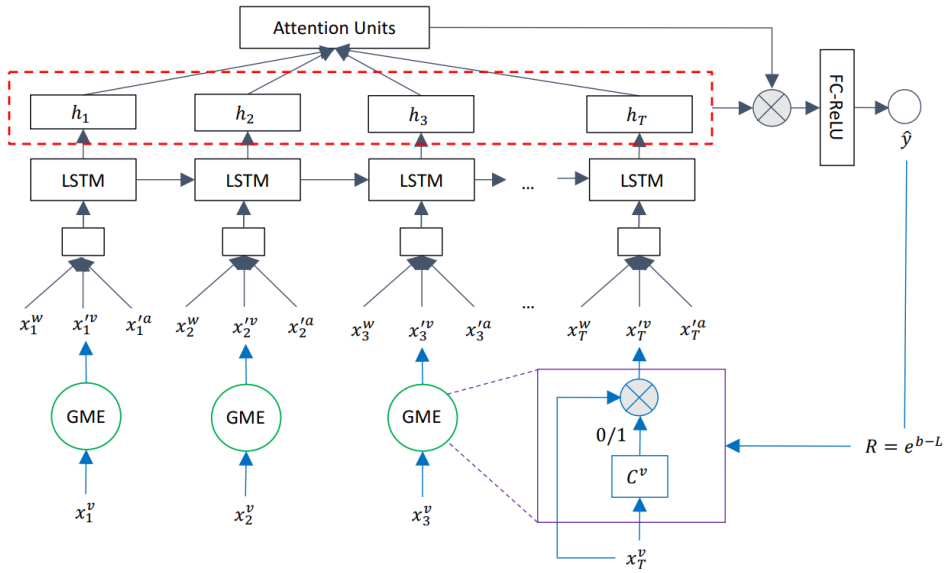


Рис. 33. Архитектура модели GME-LSTM(A) для визуальной модальности

1. Gated Multimodal Embedding (гейтинговый слой) – выполняет

селективное объединение модальностей на уровне слова. Специальные контроллеры (для аудио и видео каналов) решают, следует ли допускать сигнал данной модальности к обработке или блокировать его как шумовой. Языковая модальность всегда сохраняется, так как она является наиболее надёжной в задачах анализа тональности[65].

2. LSTM с временным вниманием (Temporal Attention) – анализирует временную структуру последовательности и акцентирует внимание на наиболее значимых моментах высказывания (например, на словах с выраженной эмоциональной окраской или сопровождаемых сильной невербальной реакцией).

Методология исследования включает следующие этапы:

– Гейтинговый механизм – извлечение признаков из текста (словные векторы GloVe), аудио (COVAREP) и видео (Facet, OpenFace);

– синхронизация признаков на уровне слов;

– селективная фильтрация модальностей с помощью гейтов;

– моделирование временных зависимостей с использованием LSTM;

– применение механизма soft attention для выделения ключевых временных шагов;

– обучение контроллеров гейтинга методом REINFORCE (policy gradient), что позволяет решать задачу с недифференцируемыми бинарными решениями.

Основу метода составляют два ключевых блока:

Для каждой модальности вводятся бинарные контроллеры:

$$x'_{a,t} = c_{a,t} \,.\, x_{a,t} = C_a\left(x_{a,t}; \theta_a\right).\, x_{a,t}$$

$$x'_{v,t} = c_{v,t} \,.\, x_{a,t} = C_v\left(x_{v,t}; \theta_v\right).\, x_{v,t}$$

$x_{a,t}$ и $x_{v,t}$ исходные признаки аудио и видео модальностей на момент времени

C_a и C_v нейронные контроллеры (гейты) для аудио и видео соответственно;

θ_a и θ_v параметры контроллеров;

$c_{a,t}$ и $c_{v,t} \in \{0,1\}$ бинарное решение: пропустить или заблокировать модальность;

$x'_{a,t}$ и $x'_{v,t}$ отфильтрованные признаки, используемые в дальнейшем

LSTM с вниманием - на вход рекуррентной сети подаётся вектор

$$x_t = \begin{bmatrix} x_{w,t} \\ x_{a,t} \\ x_{v,t} \end{bmatrix}$$

LSTM обновляет состояния:

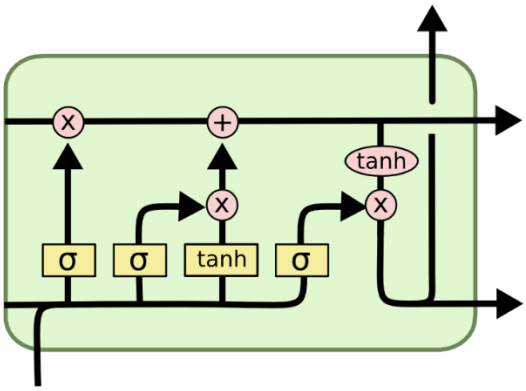


Рис. 34. Схема внутренней структуры ячейки LSTM

$c_t = f_t \odot c_{t-1} + i_i \odot m_t, \quad h_t = o_t \odot \tanh(c_t)$ $\mathrm{c_t}$ ячейка памяти LSTM, аккумулирующая информацию;

$\mathrm{h_t}$ скрытое состояние

$\mathrm{i_i}, \mathrm{f_t}, \mathrm{o_t}$ гейты (input, forget, output)

$\odot$ поэлементное умножение

Для выделения значимых временных шагов применяется soft attention:

$$\alpha_t = \frac{\exp(W^T h_t)}{\sum_{k=1}^{T} \exp(W^T h_t)}$$

$$z = \sum_{t=1}^{T} \alpha_t h_t$$

α_t вес важности временного шага T

W обучаемый вектор внимания

z итоговый взвешенный контекстФинальное предсказание:

$$\hat{y} = Q(z)$$

Так как выход контроллеров $c_{a,t}, c_{v,t}$ вляется дискретным, обучение осуществляется методом REINFORCE для оптимизации параметров гейтов, так как их решения дискретные (0 или 1) и напрямую градиент через них не вычисляется

$$\nabla_{\theta_v} = J(\theta_v) \approx \frac{1}{n}\sum_{k=1}^{n}\sum_{i=1}^{T}\nabla_{\theta_v}\log P\left(c_{v,i}\middle|x_{v,i};\theta_v\right).\left(e^{b-L_k}\right)$$

Экспериментальная проверка показала, что предложенная модель GME-LSTM(A) превосходит существующие мультимодальные и текстовые базовые подходы.

Двоичная классификация (Acc): 76.5% (против 73.5% у C-MKL и 71.6% у SVM-MD) [63].

F-score: 73.4 (против 72.3 у SVM-MD).

MAE: 0.955, что на 13.2% лучше по сравнению с предыдущим SOTA.

Как видно из представленных данных, введение гейтингового механизма позволило избирательно отбрасывать шумные визуальные и акустические признаки, а механизм внимания выделял ключевые моменты речи, что привело к росту точности анализа.

Важнейшим преимуществом предложенного метода является сочетание селективного гейтинга и временного внимания, что обеспечивает устойчивость к шумовым модальностям и способность модели концентрироваться на ключевых моментах речи.

1.2.6. Смещения

В работе решается задача построения динамически изменяемых представлений слов, учитывающих невербальные сигналы. Гипотеза исследования заключается в том, что семантика слов может быть уточнена посредством смещения (shift) в пространстве векторных представлений в зависимости от визуальных и акустических подсигналов [66], [67].

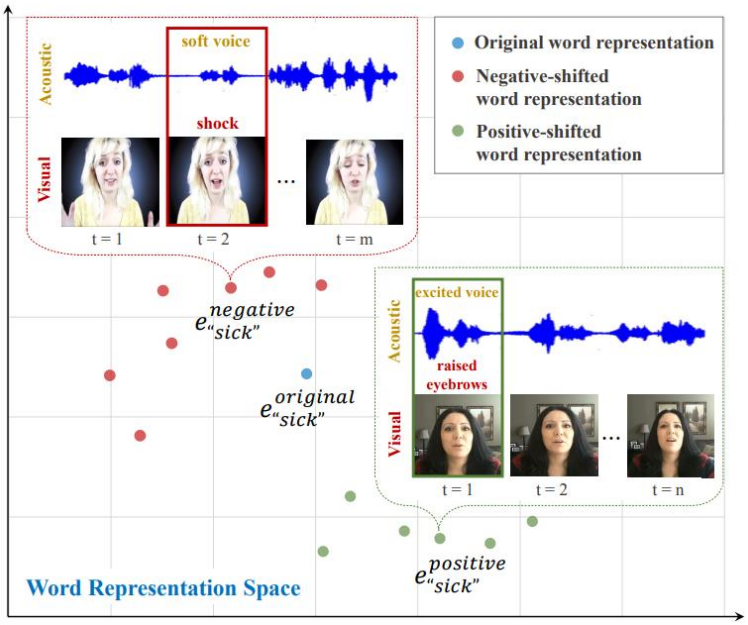


Рис. 35. Концептуальная иллюстрация, показывающая, что представление слова «sick» может изменяться в зависимости от сопутствующего невербального поведения

Целью данного исследования является разработка и экспериментальная проверка модели Recurrent Attended Variation Embedding Network (RAVEN), которая формирует мультимодально-смещённые представления слов и позволяет повысить точность анализа эмоциональной окраски и сентимента речи [68].

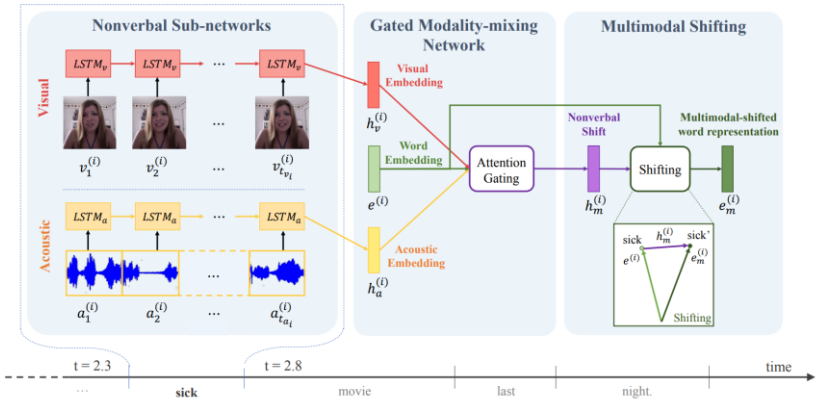


Рис. 36. Иллюстративный пример работы модели Recurrent Attended Variation Embedding Network (RAVEN)

Для достижения поставленной цели был применён метод мультимодального слияния с динамическим смещением словесных представлений. Методология исследования включает следующие этапы:

1. Извлечение признаков субсловного уровня.

Для каждого слова анализируются высокочастотные последовательности визуальных (выражения лица, движения) и акустических (интонация, тон, энергия) сигналов. Эти субсловные признаки кодируются с помощью отдельных LSTM-сетей.

$$h_v^{(i)} = LSTM_v\left(V^{(i)}\right),\ h_a^{(i)} = LSTM_a(A^{(i)})$$

где $V^{(i)}$ и $A^{(i)}$ последовательности визуальных и акустических признаков для слова $L^{(i)}$

$h_v^{(i)}$ и $h_a^{(i)}$ эмбеддинги, полученные как скрытые состояния LSTM

2. Гейтирование модальностей.

Используется Gated Modality-Mixing Network, который комбинирует визуальные и акустические признаки с исходным вектором слова. Механизм внимания (attention gating) определяет относительную значимость каждой модальности.

$$w_v^{(i)} = \sigma\left(W_{hv}\left[h_v^{(i)}; e^{(i)}\right] + b_v\right), w_a^{(i)} = \sigma\left(W_{ha}\left[h_a^{(i)}; e^{(i)} + b_a\right]\right)$$

$e^{(i)}$ сходный вектор слова из текстовой модели

$\sigma(x) = \frac{1}{1+e^{-x}}$ сигмоида, ограничивающая вес модальности от 0 до 1

$w_v^{(i)}$ и $w_a^{(i)}$ веса влияния визуальной и акустической модальности.

3. Вычисление смещения

На основании взвешенной комбинации формируется вектор смещения (shift vector), характеризующий, насколько и в каком направлении изменяется смысл слова под влиянием невербального контекста.

$$h_m^{(i)} = w_v^{(i)} . \left(W_v h_v^{(i)}\right) + w_a^{(i)} . \left(W_a h_a^{(i)}\right) + b_h^{(i)}$$

W_v и W_a матрицы преобразования для признаков

$b_h^{(i)}$ вектор смещения

$h_m^{(i)}$ итоговый вектор смещения, отражающий совокупное воздействие невербальных сигналов

4. Мультимодальное смещение (Modality-Shifting Fusion).

Исходное представление слова корректируется с использованием вектора смещения, что приводит к формированию мультимодально-смещённого вектора, который далее используется в задачах классификации эмоций и сентимента.

$$e_m^{(i)} = e^{(i)} + \alpha h_m^{(i)}$$

$$\alpha = min\left(\frac{\left\|e^{(i)}\right\|_2}{\left\|h_m^{(i)}\right\|}.\beta\ ,1\right)$$

α нормирующий коэффициент, ограничивающий величину смещения

β гиперпараметр, задающий порог масштаба

Эта операция корректирует направление эмбеддинга слова с учётом контекста, не искажая его исходную структуру.

Экспериментальная проверка показала, что модель RAVEN достигает конкурентных результатов на двух задачах [68]:

Мультимодальный сентимент-анализ (CMU-MOSI) :

- Средняя абсолютная ошибка (MAE) = 0.915,
- Корреляция с истинными метками = 0.691,
- Точность бинарной классификации (Acc-2) = 78.0%

Распознавание эмоций (IEMOCAP) [69]:

Достигаются высокие значения точности и F1-меры по классам «радость», «грусть», «злость» и «нейтральное состояние» (например, F1 = 85.8 для эмоции «радость»). Важнейшим преимуществом предложенного метода является способность учитывать динамическую изменчивость значения слов в зависимости от невербального контекста. В отличие от ранних и поздних методов фьюжна, модель RAVEN интегрирует мультимодальную информацию на уровне слов и позволяет формировать устойчивые и интерпретируемые смещённые представления.

Проведённое сравнение с аналогами, включая Memory Fusion Network (MFN) [70] и Recurrent Multistage Fusion Network (RMFN) [64], демонстрирует повышение эффективности в учёте субсловных структур и более точное моделирование

эмоциональных оттенков речи.

К ограничениям работы можно отнести необходимость высокой точности извлечения невербальных признаков (FACET, COVAREP), а также повышенные вычислительные затраты при анализе длинных последовательностей.

1.2.7. Empirical Multimodally-Additive1 function Projection (EMAP)

Несмотря на успехи глубоких нейронных сетей в обработке мультимодальной информации, остается открытым вопрос: действительно ли такие модели способны обучаться нелинейным кросс-модальным взаимодействиям, или их предсказания сводятся к простой аддитивной комбинации информации из каждой модальности [9], [58]. В работе решается задача количественного измерения вклада нелинейных взаимодействий между модальностями в процессе принятия решения моделью. Авторы ставят под сомнение интуитивное предположение, что мультимодальные архитектуры автоматически учитывают синергетический эффект комбинации источников данных.

Целью данного исследования является разработка и экспериментальная проверка метода, позволяющего различать и количественно оценивать аддитивные и нелинейные кросс-модальные взаимодействия в мультимодальных моделях [71]. Гипотеза исследования состоит в том, что даже сложные нейронные сети часто могут быть аппроксимированы более простыми аддитивными моделями, что требует специального анализа для выявления истинных взаимодействий.

Для достижения поставленной цели был применен метод Empirical Multimodally Additive Projection (EMAP), позволяющий проектировать выход нелинейной модели на пространство аддитивных моделей. Методология исследования включает следующие этапы:

1. Определение исходной мультимодальной модели. Рассматривается модель, выполняющая нелинейное объединение модальностей x_A и x_B

$$\hat{y} = f(x_A, x_B)$$

где

x_A признаки из модальности A

x_B признаки из модальности B

f нелинейная функция (нейросеть, SVM и др.), выполняющая слияние.

Эта формула описывает базовый случай, когда модель потенциально способна извлекать кросс-модальные взаимодействия.

2. Формирование аддитивной аппроксимации. С использованием EMAP модель $f(x_A, x_B)$ проецируется на пространство функций, разлагающихся на сумму отдельных вкладов каждой модальности:

$$\tilde{f}(x_A, x_B) = \mathbb{E}_{x_B}[f(x_A, x_B)] + \mathbb{E}_{x_A}[f(x_A, x_B)]$$

$\mathbb{E}_{x_B}[f(x_A, x_B)]$ усреднённый вклад модальности A при вариации модальности B.

$\mathbb{E}_{x_A}[f(x_A, x_B)]$ усреднённый вклад модальности B при вариации модальности A.

Такая проекция формирует эмпирическое приближение к аддитивной модели, минимизируя вклад кросс-модальных взаимодействий.

Экспериментальная проверка показала, что многие мультимодальные модели (включая глубокие нейронные сети и ансамбли) в значительной степени воспроизводимы с помощью аддитивных аппроксимаций [72].

– Линейные модели и простые ансамбли (например, image+text fusion без глубоких связей) демонстрируют почти полное соответствие аддитивным представлениям.
– Сложные архитектуры (LXRT, LXMERT, глубокие нейронные сети) также часто не используют сильные нелинейные взаимодействия, несмотря на потенциальную возможность их моделирования [73], [74].
– Количественные метрики (ошибки аппроксимации) указывают, что вклад нелинейных взаимодействий часто оказывается невелик по сравнению с аддитивной частью.

Важнейшим преимуществом предложенного метода является его универсальность: EMAP может быть применён к любым моделям, независимо от их архитектуры, так как работает на уровне эмпирических предсказаний[71].

Проведенное сравнение с существующими подходами (например, стандартными методами интерпретации нейросетей) демонстрирует повышение эффективности в идентификации именно кросс-модальных взаимодействий, а не общих корреляций [75], [76].

1.2.8. Оптимизация мультимодальных остатков

В работе решается задача разработки и внедрения метода, позволяющего явно сепарировать унимодальные, бимодальные и тримодальные взаимодействия в рамках общей мультимодальной архитектуры[77].

Целью данного исследования является создание подхода, который, не снижая прогностической эффективности, обеспечит количественную и качественную интерпретируемость вклада различных порядков взаимодействия (отдельных модальностей, пар модальностей, троек модальностей) в итоговое предсказание модели. В качестве гипотезы выдвигается предположение о том, что, применив принцип «бритвы Оккама», можно последовательно обучать модель более простым вкладам (унимодальным), а затем использовать более сложные взаимодействия (бимодальные и тримодальные) для коррекции ошибок (остатков) предыдущих, что позволит последним сфокусироваться исключительно на неаддитивных эффектах [77], [78].

Для достижения поставленной цели был применен метод оптимизации мультимодальных остатков, который позволяет явно разделить различные порядки взаимодействий. Методология исследования включает следующие этапы:

1. Построение и обучение унимодальных моделей: на первом этапе для каждой отдельной модальности (Visual, Acoustic, Language) строится и обучается независимый прогностический модуль ($Model_V, Model_A, Model_L$). Эти модели предсказывают целевую переменную, исходя исключительно из да нных своей модальности.

2. Расчет унимодальных остатков (резидуалов): после обучения унимодальных моделей вычисляются ошибки их предсказаний относительно

истинного значения. Эти ошибки, или остатки, представляют собой ту часть целевой информации, которую не смогли объяснить отдельные модальности.

3. Построение и обучение бимодальных моделей: на втором этапе формируются бимодальные модули ($Model_V, Model_A, Model_L$). Каждая бимодальная модель обучается не для предсказания целевой переменной напрямую, а для коррекции остатка (residual) соответствующей пары унимодальных моделей. Это обеспечивает фокусировку бимодального модуля только на неаддитивных взаимодействиях между двумя модальностями[78].

4. Расчет бимодальных остатков: аналогично, вычисляются остатки бимодальных предсказаний (с учетом унимодальных вкладов).

5. Построение и обучение тримодальной модели: на третьем этапе тримодальный модуль ($Model_{V,A,L}$) обучается для коррекции остатка, оставшегося после суммирования вкладов всех унимодальных и бимодальных моделей. Таким образом, тримодальный модуль сфокусирован исключительно на объяснении эффекта, возникающего при совместном воздействии всех трех модальностей.

6. Финальное предсказание: итоговое предсказание модели MRO является суммой предсказаний всех обученных унимодальных, бимодальных и тримодального модулей. Такая последовательная, остаточная оптимизация гарантирует, что более сложные модели (например, бимодальные) обучаются исключительно на той информации (остатке), которую не смогли объяснить более простые модели (унимодальные), тем самым обеспечивая явное разделение и количественную оценку вклада каждого порядка взаимодействия. Итоговое предсказание $\hat{y}_{MRO}$ определяется как сумма предсказаний всех модулей, соответствующая вкладам различного порядка:

$$\hat{y}_{MRO} = \sum_{m \in M} \hat{y}_m + \sum_{m_1, m_2 \in M,\, m_1 < m_2} \hat{y}_{m_1, m_2} + \hat{y}_M$$

M - Множество всех модальностей, используемых в задаче ($\mathrm{M} = \{\mathrm{V}, \mathrm{A}, \mathrm{L}\}$)

$\sum_{\mathrm{m} \in \mathrm{M}} \hat{\mathrm{y}}_{\mathrm{m}}$ Сумма унимодальных вкладов (предсказаний) от каждой отдельной модальности

$\sum_{m_1,m_2 \in M,\, m_1<m_2} \hat{y}_{m_1,m_2}$ Сумма бимодальных вкладов (предсказаний)

$\hat{y}_M$ Тримодальный вклад, или предсказание, полученное при совместной обработке всех модальностей.

Уравнение формализует принцип, согласно которому общее предсказание формируется за счет явного суммирования независимо рассчитанных унимодальных, бимодальных и тримодальных эффектов, что обеспечивает интерпретируемость за счет аддитивного разложения.

Обучение унимодальных модулей Model_m происходит путем минимизации стандартной функции потерь $\mathcal{L}_{\mathrm{task}}$

$$\hat{\theta}_m = \underset{\theta_m}{\operatorname{argmin}}\, \mathcal{L}_{task}(y, f_m(x_m; \theta_m))$$

θ_m- оптимальные параметры (веса) унимодальной модели Model_m для модальности m

y - истинное целевое значение

x_m - входные данные, соответствующие модальности m

$f_m(.;\theta_m)$- функция (модель) предсказания, использующая только модальность m с параметрами θ_m

Унимодальные модели обучаются стандартным образом, чтобы максимально объяснить целевую переменную, используя только свою модальность.

Обучение бимодального модуля $\mathrm{Model}_{V,A}$ направлено на минимизацию потерь на остатке, оставшемся после учета унимодальных вкладов:

$$\hat{\theta}_{V,A} = \underset{\theta_{V,A}}{\operatorname{argmin}}\, \mathcal{L}_{task}\left(\underbrace{y - (\hat{y}_V + \hat{y}_A)}_{Residual_{V,A}}, f_{V,A}(x_V, x_A; \theta_{V,A}) \right)$$

$\theta_{V,A}$ - оптимальные параметры бимодальной модели для модальностей.

$\mathrm{Residual}_{V,A}$ - остаток (целевое значение), который необходимо предсказать бимодальной модели. Он рассчитывается как разность между истинным значением y и суммой предсказаний двух соответствующих унимодальных моделей $(\hat{y}_V + \hat{y}_A)$ $f_{V,A}(x_V, x_A; \theta_{V,A})$. Функция предсказания бимодальной модели, использующая

данные x_V, x_A.

Это ключевой механизм MRO. Вместо предсказания y напрямую, бимодальный модуль учится предсказывать необъясненную часть целевого значения. Это заставляет модуль фокусироваться исключительно на неаддитивных взаимодействиях между V и A, поскольку аддитивные (унимодальные) вклады уже были вычтены.

Аналогично, тримодальный модуль $\mathrm{Model}_{V,A,L}$ обучается на остатке после учета всех унимодальных и бимодальных вкладов:

$$\hat{\theta}_M = \underset{\theta_M}{\operatorname{argmin}} \mathcal{L}_{task}\left(\underbrace{y - \left(\sum_{m \in M} \hat{y}_m + \sum_{m_1 < m_2} \hat{y}_{m_1,m_2}\right)}_{Residual_M}, f_{Vm}(x_V, x_A, x_L; \theta_M)\right)$$

Обеспечивает, что тримодальный модуль измеряет только эффект высшего порядка взаимодействия, устраняя влияние всех аддитивных и бимодальных эффектов.

Экспериментальная проверка метода MRO проводилась на двух эталонных мультимодальных наборах данных: CMU-MOSI и CMU-MOSEI, предназначенных для анализа тональности (sentiment analysis) и эмоций. Основная цель состояла в оценке способности MRO разделять взаимодействия без снижения прогностической производительности.

Было установлено, что, несмотря на структурную декомпозицию, метод MRO не приводит к деградации прогностической производительности по сравнению с сильными базовыми моделями (Multimodal Transformer, Self-Supervised-MT). На наборе данных CMU-MOSI по метрике MAE, MRO показал сравнимые результаты(MAE=0.825) с Multimodal Transformer (MAE=0.828), подтверждая сохранение точности [70], [79].

В частности, перспективным направлением является применение MRO для задач с неполными данными (missing modalities), где можно точно оценить, как потеря одной модальности влияет на способность бимодальных и тримодальных модулей компенсировать эту потерю. Также актуальным является исследование

возможности динамического определения иерархии остаточной оптимизации для разных типов задач.

1.2.9. Multimodal Masked Autoencoder

В этой работе решается задача создания масштабируемой модели, способной обучаться исключительно на задаче маскированного восстановления, без использования контрастивных целей и модально-специфических энкодеров. Предлагается архитектура Multimodal Masked Autoencoder (M3AE), которая интегрирует изображение и текст в единую последовательность токенов и обучается восстанавливать их после случайного маскирования [80].

Целью данного исследования является разработка и экспериментальная проверка универсального метода мультимодального представления, который обеспечивает переносимость и высокую эффективность на различных задачах компьютерного зрения и обработки естественного языка.

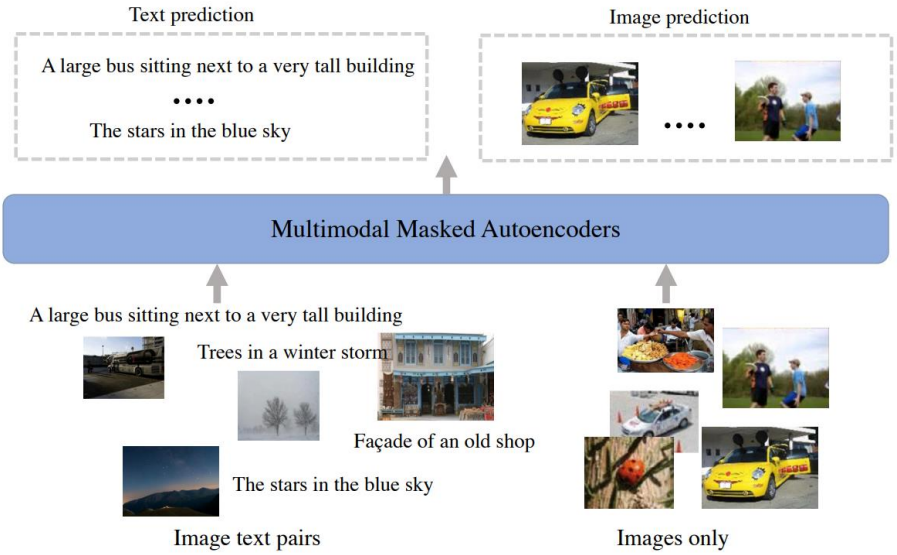


Рис. 37. Схема работы мультимодальных маскированных автоэнкодеров (Multimodal Masked Autoencoders)

Для достижения поставленной цели был применен метод маскированного восстановления токенов. Исходные данные (изображения и текст) предварительно преобразуются в дискретные последовательности:

– изображения $I \in \mathbb{R}^{H \times W \times C}$ делится на N непересекающихся патчей, каждый из которых линеаризуется и проецируется в эмбеддинг $x_i \in \mathbb{R}^d$ [81]
– текст $T = \{w_1, w_2, \dots, w_M\}$ токенизируется и через словарные эмбеддинги представляется как последовательность векторов $t_j \in \mathbb{R}^d$ [82]

Объединённая последовательность представляется как

$$Z = \{x_1, x_2, \dots, x_N, t_1, t_t, \dots, t_M\}$$

Затем случайным образом выбирается множество индексов замаскированных элементов $M \subset \{1, \dots, N + M\}$, а оставшиеся токены образуют множество видимых элементов V.

Функция кодирования определяется как

$$H = f_\theta(Z_V)$$

где f_θ трансформер-энкодер, обучаемый на видимых токенах.

На вход энкодеру поступают только видимые токены, которые проходят через трансформерные блоки, формируя скрытые представления. случайным образом скрывается значительная часть элементов последовательности (до 75% для изображений и до 90% для текста).

Декодер принимает на вход H и специальные mask tokens для позиций из множества M:

$$\hat{Z} = g_\phi(H, Z_M^{mask})$$

где g_ϕ трансформер-декодер, восстанавливающий замаскированные элементы. легковесная трансформерная сеть принимает зашифрованные токены и специальные маркеры «mask tokens» и восстанавливает скрытые части изображения и текста [83].

Обучение осуществляется с помощью функции потерь, комбинирующей визуальную и текстовую реконструкцию:

$$\mathcal{L} = \lambda_{img} \cdot \frac{1}{|M_I|} \sum_{i \in M_I} \|\hat{x}_i - x_i\|^2 + \lambda_{text} \cdot \frac{1}{|M_T|} \sum_{j \in M_T} -\log p(y_j | \hat{y}_j)$$

где M_I и M_T множества замаскированных индексов изображения и текста

соответственно.

$\hat{x}_i$ реконструированное значение патча изображения, x_i исходное значение. Первая сумма отражает MSE для визуальной реконструкции.

$\hat{y}_j$ распределение вероятностей над словарём, предсказанное декодером;

y_j истинный токен. Вторая сумма соответствует кросс-энтропии для восстановления текста.

Балансовые коэффициенты λ_{img} и λ_{text} регулируют вклад изображений и текста в общий функционал.

Таким образом, методология M3AE формально сводится к задаче минимизации функции реконструкции, определённой на объединённой последовательности изображений и текста, что обеспечивает формирование единого представления для обеих модальностей.

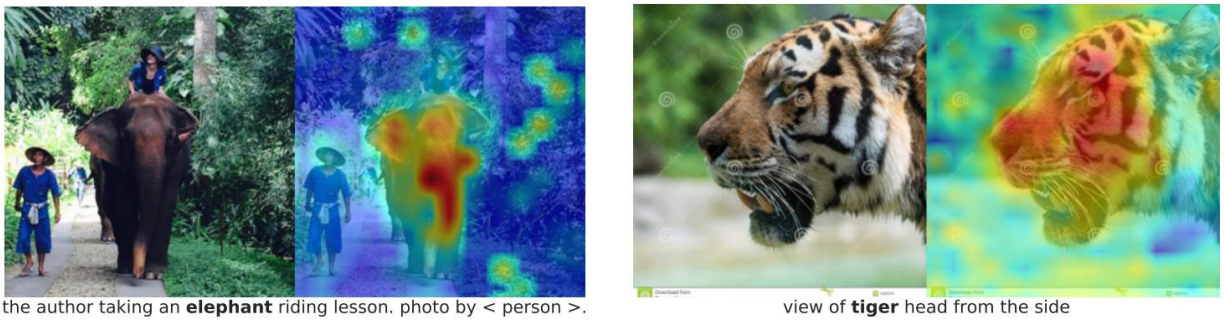


Рис. 38. Визуализация внимания между текстовым токеном и фрагментами изображения на наборе данных CC12M [84]

На данном примере продемонстрировано, каким образом M3AE обрабатывает объединённую последовательность токенов: значительная часть изображений и текста скрывается, а модель восстанавливает их на основе контекстных связей. Характерной особенностью является использование высокого уровня маскирования текста (до 75%), что существенно превышает стандартные параметры в NLP (15% у BERT). Это указывает на способность M3AE формировать более тесные межмодальные связи, так как восстановление становится возможным только при глубокой интеграции визуальных и текстовых признаков [80].

Таким образом, рис. 38 можно рассматривать как наглядный результат работы модели: она демонстрирует не только теоретическую идею объединения

модальностей, но и её практическую реализацию в процессе обучения.

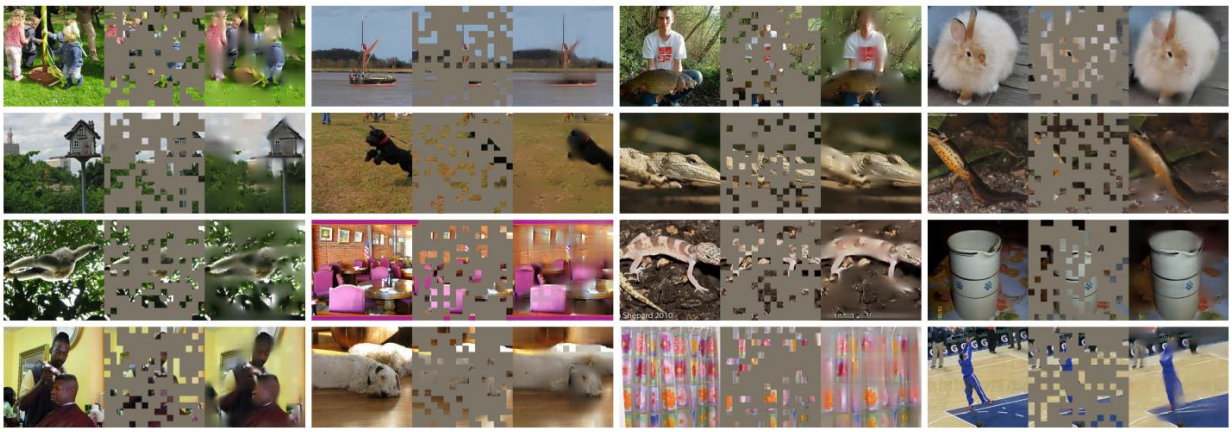


Рис. 39. Восстановление маскированных изображений на валидационном наборе данных ImageNet

Реконструкции изображений, представленные на рис. 39, являются прямым свидетельством эффективности разработанного подхода. Для каждого примера приведены три варианта: исходное изображение, его маскированная версия и результат восстановления с помощью M3AE. Видно, что модель не ограничивается локальной интерполяцией соседних патчей, а строит глобально осмысленные реконструкции, отражающие структуру объектов и общий семантический контекст сцены.

Особенно важно отметить, что данная способность проявляется на валидационном наборе ImageNet, который не использовался в обучении. Это означает, что модель действительно формирует обобщённые мультимодальные представления, переносимые на новые домены. Сравнение с традиционным MAE подтверждает преимущество M3AE: при том же объёме обучения она достигает более высокого качества реконструкции и лучшей семантической целостности изображений.Таким образом, рис. 39 демонстрирует один из ключевых результатов исследования: объединённое обучение на изображениях и тексте позволяет формировать представления, обеспечивающие не только высокие метрики на классификационных задачах, но и более полное понимание визуальной информации.

1.2.10. Dynamic Multimodal Fusion (DynMM)

Исследование представляет собой значительный вклад в область мультимодального глубокого обучения, в частности, в задачу адаптивного объединения разнородных модальностей. Основная цель работы заключается в преодолении ограничения статических схем слияния, которые применяют одинаковую вычислительную нагрузку к каждому входу независимо от его сложности. Авторы вводят новую парадигму динамическое мультимодальное слияние, позволяющее формировать индивидуальные вычислительные пути для разных экземпляров данных [85], тем самым снижая избыточные вычисления для «простых» примеров и сохраняя полную выразительность модели для «сложных» случаев. Концептуальная новизна исследования состоит в сочетании динамических нейронных сетей с принципами мультимодального обучения, что открывает новое направление в проектировании адаптивных архитектур. Методологическая основа DynMM построена вокруг двух уровней адаптивного принятия решений: на уровне модальностей и на уровне операций слияния. Пусть входные данные представляют собой вектор $x = (x_1, x_2, \dots, x_M)$ где x_i это i-я модальность. на модальном уровне модель опирается на набор экспертных сетей $E_i(x_i)$, каждая из которых специализируется на одной или нескольких модальностях. Управление выбором экспертов осуществляется с помощью гейтинговой сети $G(x)$, которая вырабатывает разреженный вектор решений $g = (g_1, g_2, \dots, g_B)$ где B число доступных экспертов [86].

Итоговый выход модели формируется выражением

$$y = \sum_{i=1}^{B} g_i E_i(x_i)$$

При этом в отличие от классического подхода, вектор g является one-hot-кодировкой, что означает активацию лишь одного эксперта на каждую выборку, позволяя минимизировать вычислительные затраты.

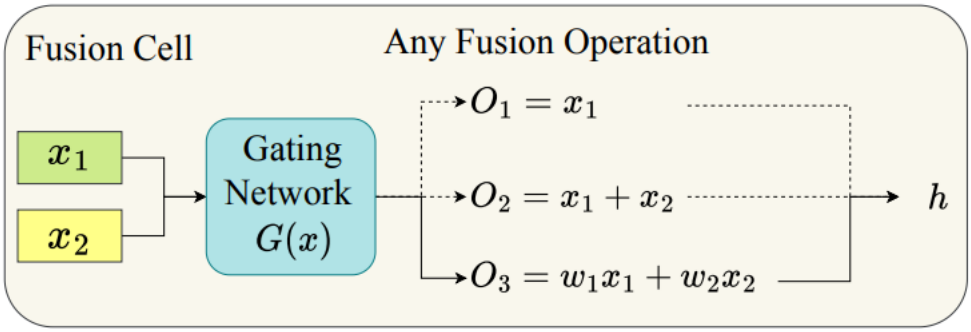


Рис. 40. Иллюстрация модели DynMM на уровне модальностей

На более детализированном уровне – уровне слияния – вводится понятие fusion cell, которое содержит множество возможных операций объединения модальностей $\{O_i\}$. Например, операции могут иметь вид:

$$O_1 = x_1, O_2 = x_1 + x_2, O_2 = w_1 x_1 + w_2 x_2$$

где w_1, w_2 обучаемые параметры. Для выбора конкретной операции используется гейтинговая функция $G(x)$, определяющая одноразрядный вектор g, формирующий выход ячейки

$$h = \sum_{i=1}^{B} g_i O_i(x_i)$$

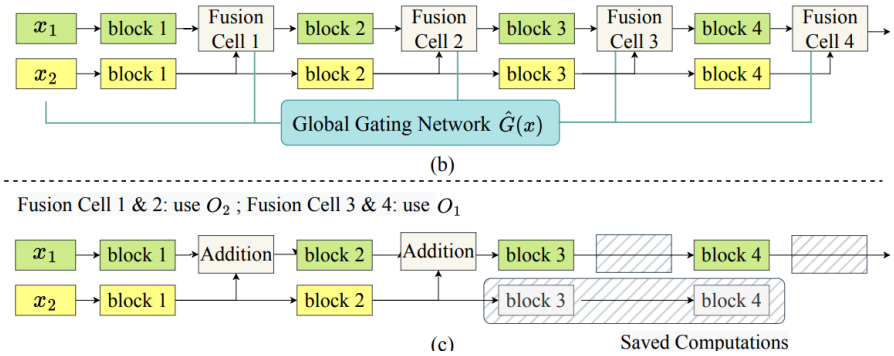


Рис. 41. Иллюстрация модели DynMM на уровне слияния (fusion-level)

Множество таких ячеек образует каскад, в котором ранние уровни могут «завершать» обработку для простых примеров, экономя ресурсы, тогда как более сложные образцы проходят все стадии слияния, обеспечивая максимальную

выразительность представлений. Эта структура позволяет адаптивно активировать только необходимые ветви вычислений, аналогично идее раннего выхода (early exit) в динамических нейросетях [87].

Для достижения компромисса между точностью и вычислительной эффективностью вводится ресурсно-осознанная функция потерь [9]. Для модального уровня она имеет вид:

$$\mathcal{L} = \mathcal{L}_{\text{task}} + \lambda \sum_{i=1}^{B} g_i O_i(x_i)$$

а для уровня слияния:

$$\mathcal{L} = \mathcal{L}_{\text{task}} + \lambda \sum_{j=1}^{F} \sum_{i=1}^{B} g_i^{(j)} C(O_{i,j})$$

Где $\mathcal{L}_{\text{task}}$ целевая функция задачи $C(E_i)$ и $C(O_{i,j})$ вычислительные затраты, измеренные через количество операций multiply-adds (MAdds).

λ гиперпараметр, задающий баланс между точностью и экономичностью. Таким образом, повышение λ приводит к более экономным, но потенциально менее точным решениям.

Оптимизация модели осложнена тем, что гейтинговая функция порождает недифференцируемые дискретные решения. Для решения этой проблемы используется Gumbel-Softmax-репараметризация:

$$\hat{g}_i = \frac{\exp((\log G(x)_i + b_i)/\tau)}{\sum_{j=1}^{B} \exp\big((\log G(x)_j + b_j)/\tau\big)}$$

Где $b_i \approx \text{Gumbel}(0,1)$, а τ температура, контролирующая сглаженность распределения [88]. При больших значениях τ распределение более равномерно, при малых приближается к категориальному. Для обучения используется двухэтапная схема: на первом этапе проводится предобучение экспертных и слияющих модулей при случайных решениях гейта, чтобы избежать предвзятости и обеспечить равномерное обновление весов; на втором этапе совместная тонкая настройка с включением гейтинговой сети и репараметризацией, что позволяет достичь согласованной оптимизации всех компонентов. Варианты обучения

включают straight-through-аппроксимацию, при которой на прямом проходе сохраняются жесткие (дискретные) решения, а на обратном используются их мягкие дифференцируемые аналоги.

Эмпирическая проверка подхода проведена на трёх задачах: классификации жанров фильмов (MM-IMDB), анализе сентиментов (CMU-MOSEI) и семантической сегментации RGB-D изображений (NYU Depth V2) [89].

Для первых двух задач применялся модальный уровень DynMM, для последней – уровень слияния (см. табл. 1).

Таблица 1 – Сравнение предлагаемого подхода с передовыми методами (SOTA) для семантической сегментации RGB-D на тестовых данных набора NYU Depth V2.

Method	Modality	Micro F1 (%)	Macro F1 (%)	MAdds (M)
Image Network	I	39.99	25.26	5.0
Text Network (E_1)	T	59.16	47.21	0.7
Late Fusion [24] (E_2)	I+T	59.55	50.94	10.3
LRTF [26]		59.18	49.26	10.3
MI-Matrix [19]		58.45	48.36	10.3
DynMM-a		**59.57**	48.84	1.6
DynMM-b		**59.59**	50.42	7.8
DynMM-c		**59.72**	**51.20**	9.8
DynMM-d		**60.35**	**51.60**	12.1

В задаче CMU-MOSEI точность бинарной классификации достигла 79.75%, а средняя абсолютная ошибка снизилась до 0.60, при этом вычислительная нагрузка уменьшилась почти на 46.5% по сравнению с классическими схемами позднего слияния [90].

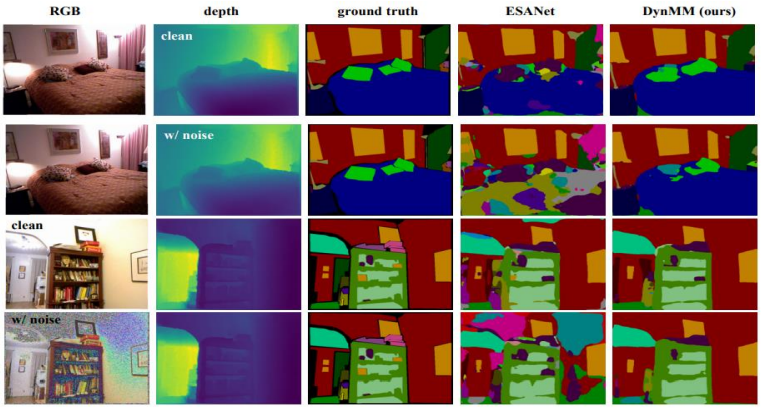


Рис. 42. Качественные результаты сегментации на наборе данных NYU Depth V2

На NYU Depth V2 DynMM показала повышение mIoU на 0.7% при сокращении вычислений на 21%, демонстрируя, что динамическое распределение вычислений между RGB и глубинным каналом позволяет не только экономить ресурсы, но и повышать устойчивость к шуму [91]. Особенно показательны эксперименты с зашумлёнными входными данными, где DynMM сохраняла высокую точность сегментации, тогда как статическая сеть ESANet значительно деградировала [92]. С точки зрения теоретической и практической ценности работа обладает высокой оригинальностью, так как впервые системно формулирует идею динамического мультимодального слияния как обобщение принципов адаптивных сетей на область мультимодальности. Однако, несмотря на убедительные результаты, у метода есть ряд ограничений. Во-первых, сложность проектирования гейтинговых сетей и необходимость тщательного выбора гиперпараметра λ делают процесс настройки трудоёмким. Во-вторых, подход опирается на предварительно обученные экспертные архитектуры и не учитывает возможные корреляции между модальностями при выборе ветвей, что может приводить к потере межмодальных взаимодействий в некоторых сценариях. В-третьих, метод пока не исследован в задачах последовательного принятия решений (например, длительного видеоанализа), где временная зависимость между модальностями требует более сложной динамической стратегии. Наконец, использование Gumbel-Softmax влечёт за собой аппроксимационные ошибки при низких температурах, а обучение может страдать от нестабильности на ранних этапах, что требует дополнительных процедур регуляризации [93].

В совокупности исследование Xue и Marculescu представляет собой весомый шаг к созданию ресурсно-адаптивных, модально осведомлённых архитектур, способных гибко реагировать на характеристики входных данных. Его теоретическая элегантность и экспериментальная убедительность делают DynMM важным рубежом в развитии мультимодального глубокого обучения, задающим направление для будущих разработок в области динамических и энергоэффективных нейросетей, способных к контекстуальной самоадаптации.

1.2.11. High-Modality Multimodal Transformer (HighMMT)

High-Modality Multimodal Transformer представляет собой одно из наиболее систематичных и теоретически обоснованных направлений в области мультимодального обучения представлений. Его центральная идея заключается в том, что при увеличении числа модальностей в задаче машинного обучения встает фундаментальный вопрос гетерогенности насколько различные модальности и их взаимодействия отличаются друг от друга, и возможно ли использовать общие параметры между ними без потери точности [94]. Вклад работы состоит в формулировке двух информационно-теоретических метрик, гетерогенности модальностей и гетерогенности взаимодействий, а также в создании архитектуры HighMMT [95], которая динамически управляет совместным использованием параметров между модальностями, обеспечивая оптимальный баланс между вычислительной эффективностью и качеством обучения.

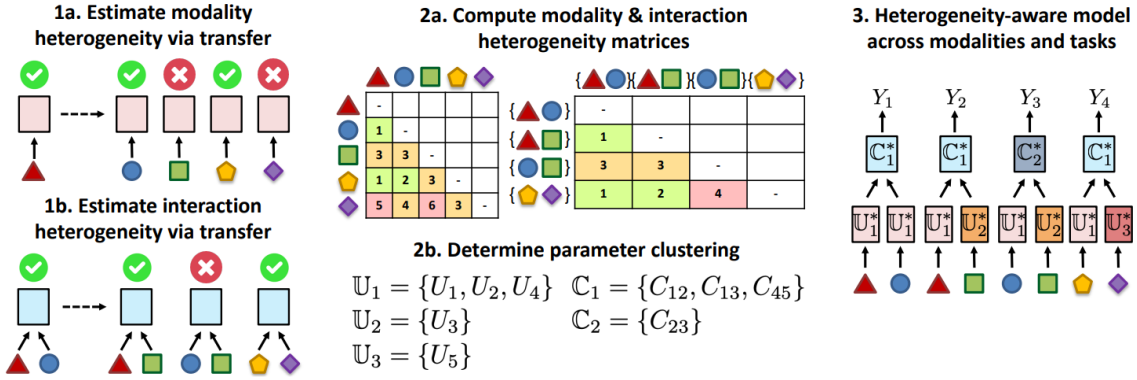


Рис. 43. Рабочий процесс модели HighMMT (High-order Multimodal Transfer)

Предложенная методология основана на концепции информационного переноса между модальностями. Для двух модальностей X_1 и X_2 связанных общей задачей Y , авторы определяют метрику передачи информации через разность потерь при обучении модели на исходной модальности и её переносе на целевую [96].

Пусть $\hat{y} = f(y|x_2; \theta)$ предсказание модели с параметрами θ .Тогда базовая ошибка при прямом обучении на X_2 выражается как:

$$L_2^* = \min_\theta \mathbb{E}_{p(x_2,y)} \ell(f(y|x_2;\theta), y)$$

а ошибка при переносе с параметрами, обученными на X_1

$$\theta_1 = \mathbb{E}_{p(x_1,y)} \ell(f(y|x_1;\theta), y)$$

$$L_{1\to2}^* = \min_\theta \mathbb{E}_{p(x_2,y)} \ell(f(y|x_2;\theta \leftarrow \theta_1), y)$$

Разность $T(X_1 \to X_2; Y) = L_{1\to2}^* - L_2^*$ интерпретируется как мера сложности переноса, а симметризированная метрика гетерогенности модальностей определяется как

$$d(X_1; X_2) = \max\big(0, T(X_1 \to X_2; Y)\big) + \max(0, T(X_2 \to X_1; Y))$$

Аналогичным образом, для оценки гетерогенности взаимодействий между парами модальностей $\{X_1, X_2\}$ и $\{X_3, X_4\}$ используется метрика

$$T(X_1, X_2 \to X_3, X_4; Y) = L_{12\to34}^* - L_{34}^*$$

где $L_{12\to34}^*$ и L_{34}^* потери модели соответственно при переносе и прямом обучении [97].

Эти меры позволяют построить две матрицы, матрицу гетерогенности модальностей $M_U(i,j) = d(X_i; X_j)$ и матрицу гетерогенности взаимодействий $M_C(i,j,k,\ell) = d(X_i; X_j; X_k; X_\ell)$. На их основе проводится иерархическая кластеризация модальностей для группировки параметров, где число кластеров k управляет компромиссом между точностью и эффективностью [98]. Таким образом, метод обеспечивает автоматизированное определение степени совместного использования параметров между модальностями, исключая необходимость ручного проектирования архитектуры под каждую задачу.

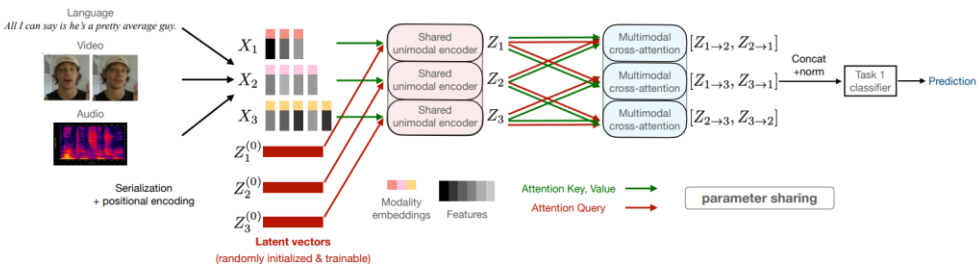


Рис. 44. Архитектура модели HighMMT (High-order Multimodal Transfer)

Архитектура HighMMT строится на модифицированной структуре Transformer и включает последовательные стадии: стандартизацию входов, добавление модальных эмбеддингов и позиционных кодировок, использование общего унимодального энкодера на основе Perceiver (Jaegle et al., 2021), а затем кроссмодальных слоёв внимания для обучения взаимодействий между модальностями [99].

Пусть X_m последовательность входов модальности m, преобразованная в представление Z_m через кросс- и самовнимание:

$$\tilde{Z}_m^{(L)} = \mathrm{softmax}\left(\frac{Q_c K_c^T}{\sqrt{d_{LS}}}\right) V_c, \qquad Z_m^{(L)} = \mathrm{softmax}\left(\frac{Q_s K_s^T}{\sqrt{d_{LS}}}\right) V_s$$

где матрицы $Q_c, K_c, V_c, Q_s, K_s, V_s$ обучаемые параметры внимания. Для кроссмодального взаимодействия между двумя модальностями Z_1 и Z_2 рименяется механизм двунаправленного внимания:

$$z_{2\to1} = \mathrm{softmax}\left(\frac{Q_1 K_2^T}{\sqrt{d_k}}\right) V_2, \qquad z_{1\to2} = \mathrm{softmax}\left(\frac{Q_2 K_1^T}{\sqrt{d_k}}\right) V_1$$

После объединения $z_{mm} = [z_{1\to2}, z_{2\to1}]$ представление поступает на задачно-специфические классификаторы. Обучение проходит в два этапа: (1) гомогенная предтренировка с полным совместным использованием параметров, и (2) гетерогенность-осведомлённая донастройка, при которой параметры разделяются по кластерам модальностей и взаимодействий, определённым по матрицам M_U и M_C.

Экспериментальная проверка проведена на 10 модальностях и 15 задачах в пяти исследовательских областях (робототехника, здравоохранение, мультимедиа, аффективные вычисления и HCI) [94]. В экспериментах использовались крупные мультимодальные наборы данных MultiBench (Liang et al., 2021), включающие текст, изображение, аудио, видео, таблицы, временные ряды, силы, проприоцепцию и множества.

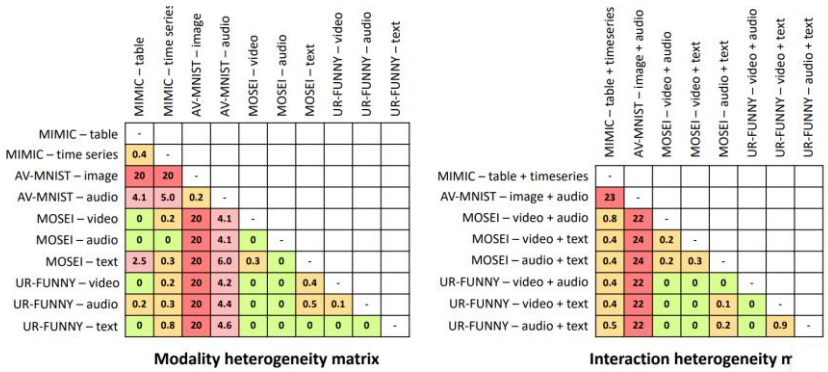


Рис. 45. Матрицы гетерогенности модальностей и взаимодействий

Результаты показывают, что HighMMT устойчиво превосходит существующие мультимодальные трансформеры по соотношению качества и количества параметров: достигая сравнимой или лучшей точности при 8–10-кратном уменьшении числа параметров. В частности, при увеличении числа групп параметров k наблюдается монотонный рост производительности (от 68,4% при k=2 до 71,2% при k=9). В задачах с ограниченными данными, таких как UR-FUNNY, достигается улучшение на 2,4% благодаря эффективному переносу знаний между различными наборами данных [100]. HighMMT также демонстрирует положительный перенос между полностью несвязанными модальностями – например, при обучении на изображениях, аудио и видео и последующей донастройке на табличные или временные данные (MIMIC) [101].

Существенным достоинством модели является её масштабируемость: добавление новых модальностей не требует проектирования отдельных энкодеров, а вычислительная сложность измерения гетерогенности аппроксимируется низкоранговыми матрицами, что позволяет обходиться выборкой порядка $O(M)$ из всех возможных M^2 комбинаций [102]. Кроме того, анализ перекрытия параметров показал, что более 92% нейронов в унимодальных энкодерах участвуют в трёх и более задачах, тогда как кроссмодальные слои обладают большей специфичностью, что подтверждает различие между гомогенностью представлений и гетерогенностью взаимодействий.

Тем не менее, несмотря на убедительные результаты, предложенный подход имеет ряд ограничений. Во-первых, метрики гетерогенности основываются на

эмпирической оценке потерь при переносе моделей и не гарантируют строгих метрических свойств: не всегда соблюдается положительность и треугольное неравенство, особенно при наличии положительного переноса. Во-вторых, вычислительная стоимость при полном измерении всех пар модальностей $O(M^2)$ остаётся высокой, а используемая низкоранговая аппроксимация хотя и эффективна, может приводить к снижению точности в задачах с высокой нелинейностью взаимодействий. Кроме того, HighMMT применяет фиксированное число кластеров k как гиперпараметр, что требует ручного подбора в зависимости от бюджета вычислений. Ещё одно ограничение связано с тем, что представления обучаются в предположении, что все модальности можно привести к последовательной форме; это упрощает архитектуру, но может приводить к потере пространственных или топологических зависимостей, характерных для изображений, графов и сенсорных сетей [73]. Наконец, хотя модель демонстрирует способность к переносу на новые задачи, она не рассматривает случаи, где между модальностями отсутствует прямая корреляция или имеется сильный шум, что может ограничить её применимость в реальных сценариях с частично наблюдаемыми или асинхронными данными.

1.2.12. Gradient-Blending

Многомодальные нейронные сети – несмотря на доступ к большему количеству информации, зачастую уступают по качеству своим одномодальным аналогам при решении задач классификации. Авторы показывают, что вопреки интуитивным ожиданиям, при совместном обучении различных модальностей (например, RGB-видео, оптический поток и аудио) объединённая сеть может демонстрировать худшую обобщающую способность и более сильное переобучение, чем отдельные модели, обученные на каждой модальности в отдельности. Этот эффект наблюдается устойчиво на различных наборах данных (Kinetics, EPIC-Kitchens, AudioSet) и при разнообразных стратегиях слияния признаков, что делает проблему принципиальной и архитектурно-независимой

[103], [104].

Dataset	Multi-modal	V@1	Best Uni	V@1	Drop
Kinetics	A + RGB	71.4	RGB	**72.6**	-1.2
	RGB + OF	71.3	RGB	**72.6**	-1.3
	A + OF	58.3	OF	**62.1**	-3.8
	A + RGB + OF	70.0	RGB	**72.6**	-2.6

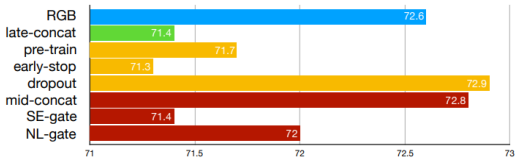


Рис. 46. Сравнение производительности одномодальных и мультимодальных сетей на наборе данных Kinetics

В качестве объяснения этого парадокса авторы выделяют две основные причины. Первая заключается в избыточной ёмкости многомодальной модели, что естественным образом ведёт к переобучению [105]. Вторая причина, различная динамика обобщения и переобучения в разных модальностях: аудио- и визуальные признаки усваиваются с разной скоростью и подвержены различным видам ошибок [106]. Применение единого оптимизационного механизма (например, стандартного стохастического градиентного спуска) к таким гетерогенным источникам информации оказывается неэффективным, поскольку игнорирует разницу в статистике ошибок и в скорости сходимости [104]. Ни предварительное обучение, ни регуляризаторы вроде dropout, ни архитектурные усовершенствования (гейты SE или NL, mid-level fusion) не решают эту проблему в полной мере [107].

Для количественного описания явления авторы вводят новый метапоказатель отношение переобучения к обобщению (Overfitting-to-Generalization Ratio, OGR) Пусть L_T ошибка на обучающем наборе, а L^* «истинная» ошибка на тестовом распределении (в экспериментах заменяется валидационной). Переобучение определяется как $O_N = L_T - L^*$, а обобщение как $G_N = L^*(\theta_0) - L^*(\theta_N)$,где θ_N параметры модели после N эпох. Тогда показатель

$$OGR = \left|\frac{\Delta O_{N,n}}{\Delta G_{N,n}}\right| = \left|\frac{O_{N+n} - O_N}{L^*_N - L^*_{N+n}}\right|$$

служит мерой качества извлекаемой информации: чем меньше OGR, тем меньше прирост переобучения по отношению к приросту обобщающей способности. Минимизация этого показателя в процессе обучения позволяет искать баланс между улучшением на обучающем множестве и реальным повышением качества на независимых данных.

На основе этой метрики авторы формулируют метод Gradient-Blending (G-Blend) новый способ совместного обучения модальностей, минимизирующий OGR на уровне градиентных шагов [103], [108]. Пусть для каждой модальности $\hat{g}_i = \nabla L_i$, а для объединённой головы

$\hat{g}_{k+1} = \nabla L_{multi}$. Тогда задача заключается в нахождении таких весов w_i, что результирующий $\hat{g}_i = \sum_i w_i \hat{g}_i$ минимизирует OGR^2

$$OGR^2 = \left(\frac{\langle \nabla L_T - \nabla L^*, \hat{g} \rangle}{\langle \nabla L^*, \hat{g} \rangle} \right)^2$$

При предположении некоррелированности переобучения разных модальностей, оптимальные веса имеют замкнутое выражение:

$$w_i^* = \frac{1}{Z} \frac{\langle \nabla L^*, v_i \rangle}{\sigma_i^2} \quad , \quad Z = \sum_i \frac{\langle \nabla L^*, v_i \rangle^2}{\sigma_i^2}$$

где σ_i^2 дисперсия переобучения i-й модальности. Эта формула аналогична минимальной дисперсии при объединении некоррелированных оценок, но вместо дисперсии используется мера переобучения [109]. Таким образом, модальности с более устойчивым поведением (низким уровнем переобучения) получают больший вес, а переобучающиеся – подавляются в процессе обучения.

Алгоритмически метод реализуется двумя вариантами: офлайн и онлайн Gradient-Blending. В офлайн-версии веса w_i цениваются один раз (например, после нескольких эпох) и остаются фиксированными на протяжении всего обучения. В онлайн-версии вычисление происходит регулярно (через несколько «суперэпох»), что позволяет динамически адаптироваться к изменению характера обобщения по мере обучения [110]. Эмпирически оба подхода обеспечивают прирост качества, причём онлайн-версия даёт немного более высокие результаты, а офлайн – проще в реализации и почти не уступает по точности. В табл. 2 приведены результаты метода Gradient-Blending (G-Blend) на различных мультимодальных задачах.

Таблица 2 – Результаты метода Gradient-Blending (G-Blend) на различных мультимодальных задачах

Modal	RGB + A			RGB + OF			OF + A			RGB + OF + A		
Weights	[RGB,A,Join]=[0.630,0.014,0.356]			[RGB,OF,Join]=[0.309,0.495,0.196]			[OF,A,Join]=[0.827,0.011,0.162]			[RGB,OF,A,Join]=[0.33,0.53,0.01,0.13]		
Metric	Clip	V@1	V@5	Clip	V@1	V@5	Clip	V@1	V@5	Clip	V@1	V@5
Uni	63.5	72.6	90.1	63.5	72.6	90.1	49.2	62.1	82.6	63.5	72.6	90.1
Naive	61.8	71.4	89.3	62.2	71.3	89.6	46.2	58.3	79.9	61.0	70.0	88.7
G-Blend	**65.9**	**74.7**	**91.5**	**64.3**	**73.1**	**90.8**	**54.4**	**66.3**	**86.0**	**66.1**	**74.9**	**91.8**

Экспериментальные результаты подтверждают эффективность предложенного метода. На наборе **Kinetics** точность top-1 при использовании Gradient-Blending (RGB+Audio) достигла 65.9% (offline) и 66.9% (online), в то время как наивное объединение модальностей дало лишь 61.8%, а лучший одномодальный RGB-модель – 63.5% [111]. Аналогичные тенденции наблюдались на **mini-Sports** и **mini-AudioSet**, а также на мультимодальных комбинациях RGB+OF, OF+A, и RGB+OF+A, где G-Blend стабильно превосходил как простую конкатенацию признаков, так и другие регуляризационные приёмы (dropout, pre-training, auxiliary losses).

В целом статья представляет собой значимый шаг в понимании природы трудностей многомодального обучения. Она не только диагностирует фундаментальную проблему различной динамики обобщения между модальностями, но и предлагает строгое математическое решение, формализующее процесс балансировки обучающих сигналов. Введение метрики OGR и метода Gradient-Blending создаёт новую парадигму оптимизации многомодальных нейросетей, способную обеспечить более устойчивое обучение и улучшенную способность к генерализации без архитектурных модификаций.

1.2.13. Greedy Learning in Multi-modal Neural Networks

Авторы формулируют гипотезу жадного обучения (greedy learner hypothesis), утверждающую, что мультимодальная нейросеть в процессе оптимизации быстро концентрируется на той модальности, от которой проще всего извлечь информативные признаки, вследствие чего обучение становится асимметричным и приводит к ухудшению обобщающей способности модели [103], [112].

Центральным вкладом работы является выявление этой природы «жадности», введение количественных метрик для её оценки, условной степени использования (conditional utilization rate) и условной скорости обучения (conditional learning speed), а также разработка алгоритма сбалансированного мультимодального обучения (balanced multi-modal learning), который позволяет динамически компенсировать различия в скорости усвоения информации из разных модальностей.

Методологически исследование опирается на формальное описание архитектуры мультимодальной нейросети с двумя унимодальными ветвями φ_1, φ_2, обрабатывающими данные из модальностей m_0 и m_2 , и промежуточными слоями слияния (intermediate fusion) на основе Multimodal Transfer Module (MMTM) [113].

Пусть выходы сверточных блоков каждой ветви обозначены как $A_0 \in \mathbb{R}^{N_1 \times \ldots \times N_L \times C}$ и $A_1 \in \mathbb{R}^{M_1 \times \ldots \times M_J \times C'}$, которые объединяются и подаются на нелинейное преобразование

$$(w_0, w_1) = g([h_0, h_1])$$

где g последовательность полносвязных слоёв и функций ReLU. Полученные активации w_0, w_1 нормируются через сигмоиду и масштабируют исходные карты признаков:

$$\widetilde{A}_0 = 2\sigma(w_0) \odot A_0, \qquad \widetilde{A}_1 = 2\sigma(w_1) \odot A_1$$

где $\odot$ обозначает поэлементное умножение. Таким образом реализуется двунаправленный обмен информацией между модальностями. Предсказание сети вычисляется как усреднение выходов обеих ветвей:

$$\hat{y} = \frac{1}{2}(\hat{y}_0 + \hat{y}_1)$$

а функция потерь определяется суммой модально-специфических кросс-энтропий:

$$L = CE(y, \hat{y}_0) + CE(y, \hat{y}_1)$$

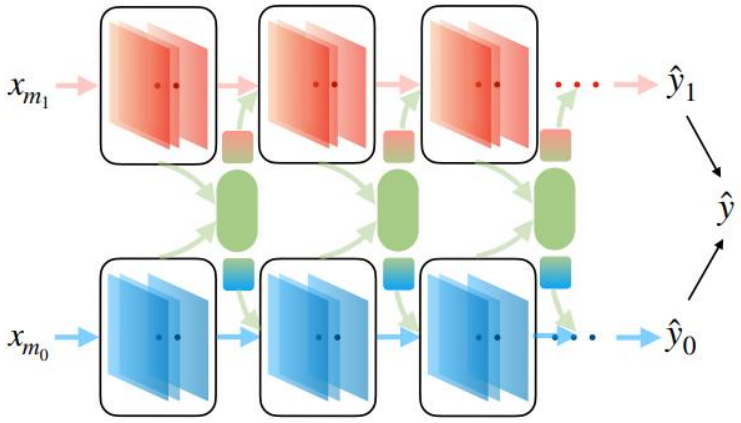


Рис. 47. Мультимодальная нейронная сеть с промежуточным слиянием (intermediate fusion)

Авторы определяют условную степень использования как относительное изменение точности при исключении одной из модальностей [114]. Пусть f обученная мультимодальная модель, а f_0', f_1' её версии, в которых перекрестное влияние между ветвями устранено. Тогда

$$u(m_0|m_1) = \frac{A(f_1) - A(f_1')}{f_1'}, \quad u(m_1|m_0) = \frac{A(f_0) - A(f_0')}{f_0'}$$

где $A(.)$ очность на тестовой выборке. Разность $d_{util} = u(m_1|m_0) - u(m_0|m_1)$ характеризует степень асимметрии: чем выше её абсолютное значение, тем сильнее сеть игнорирует одну из модальностей. Поскольку вычисление u требует завершённого обучения, вводится прокси-метрика, условная скорость обучения [115]:

$$s(m_1|m_0; t) = \log\frac{\sum_{i=1}^{t}\mu(\theta_0'; i)}{\sum_{i=1}^{t}\mu(\theta_0; i)}, \qquad s(m_0|m_1; t) = \log\frac{\sum_{i=1}^{t}\mu(\theta_1'; i)}{\sum_{i=1}^{t}\mu(\theta_1; i)}$$

где $\mu(\theta; i) = \frac{\|\nabla_\theta L\|_2^2}{\|\theta^{(i)}\|_2^2}$ измеряет относительную величину обновления параметров на i-м шаге.

$d_{speed} = u(f; t) = s(m_1|m_0; t) - s(m_0|m_1; t)$ служит показателем текущего дисбаланса в скорости усвоения сигналов.

На основе этих соотношений разработан алгоритм Balanced Multi-modal Learning, который периодически оценивает величину $|d_{speed}|$ и, если она

превышает порог α, инициирует «ребалансирующие шаги», усиливающие обучение слабо представленной модальности за счёт дополнительного масштабирования её признаков.

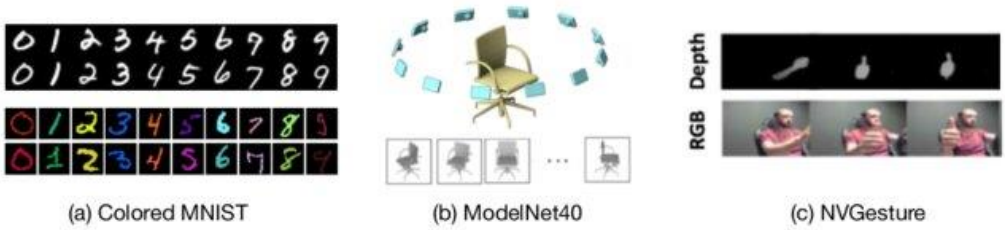


Рис. 48. Примеры мультимодальных наборов данных: (a) Colored MNIST, (b) ModelNet40, (c) NVGesture

Эксперименты проведены на трёх задачах различной природы. На Colored-and-gray MNIST проверялась сама гипотеза жадности: модель обучалась на цветных и градационных изображениях цифр, где цвет коррелирует с меткой. Наблюдалось, что при стандартном обучении сеть почти полностью игнорирует одну из модальностей, что выражалось в значении $u(\text{gray}|\text{color}) \approx 0.63$ и $u(\text{color}|\text{gray}) \approx 0.01$[112], [116] .

На ModelNet40, включающем 3D-объекты с фронтальных и тыловых проекций, и NVGesture обнаружено аналогичное поведение: стандартные мультимодальные сети демонстрировали высокое $|d_{\text{until}}|$, видетельствующее о зависимости от одной модальности [117]. При этом распределения d_{until} и d_{speed} оказались близкими, что подтверждает валидность выбранного прокси-показателя. Тем не менее авторы признают ряд ограничений предложенного подхода. Во-первых, метод формулируется и экспериментально проверяется преимущественно для бимодальных задач; обобщение на более чем две модальности требует дополнительной модификации формул и усложняет вычисление d_{speed} [118].

Во-вторых, хотя метрика скорости обучения эффективно заменяет трудоёмкое вычисление условной точности, она всё же является эвристикой, основанной на норме градиента и параметров, и может быть чувствительна к выбору нормализации и архитектуры.

В-третьих, предложенные эксперименты ограничены задачами

классификации и не охватывают сценарии, где требуется сложное межмодальное рассуждение (например, VQA), где динамика жадности может проявляться иначе. Наконец, гиперпараметры алгоритма, размер окна Q и порог α подбираются эмпирически, что может осложнить практическое применение метода к более крупным или разнородным наборам данных [119].

ВЫВОДЫ ПО ГЛАВЕ 1

В первой главе были рассмотрены ключевые теоретические аспекты, связанные с обработкой и интеграцией мультимодальных данных, а также проанализированы существующие подходы к их слиянию. Показано, что мультимодальные данные характеризуются высокой степенью гетерогенности и сложными взаимосвязями между модальностями, что требует применения методов, способных сохранять внутреннюю структуру и корреляции между различными источниками информации.

Проведённый анализ существующих методов показал, что традиционные техники слияния данных, основанные на конкатенации признаков или статистических моделях, не обеспечивают адекватного учёта межмодальных зависимостей и приводят к потере значимой информации. В этом контексте тензорные подходы представляют собой более перспективное направление, поскольку они позволяют моделировать мультимодальные данные в виде многомерных структур и сохранять сложные взаимосвязи между модальностями на уровне представлений.

Вместе с тем использование тензорных методов сопряжено с рядом проблем. Во-первых, при увеличении количества модальностей и размерности данных размер результирующего тензора экспоненциально возрастает, что существенно затрудняет вычислительную обработку и хранение данных. Во-вторых, мультимодальные данные, получаемые из реальных источников, часто содержат значительное количество шума, который негативно влияет на качество тензорного разложения и точность последующего анализа.

Таким образом, в первой главе обоснована актуальность применения тензорных методов для слияния мультимодальных данных, а также выявлены основные ограничения существующих подходов, связанные с высокой размерностью и зашумлённостью данных. Эти результаты определяют направление дальнейших исследований, представленных во второй главе, где будут предложены два взаимодополняющих подхода:

1. метод слияния мультимодальных данных на основе тензорного представления;
2. метод снижения шума в мультимодальных данных для повышения устойчивости и точности обработки в различных задачах анализа.

ГЛАВА 2. ОСНОВНЫЕ ПОЛОЖЕНИЯ МЕТОДИКИ И АЛГОРИТМЫ РЕШЕНИЯ

2.1. Слияние мультимодальных данных

В методах мультимодального слияния данных, рассмотренных в предыдущей главе, каждая из существующих технологий демонстрирует свои преимущества и недостатки. Основная цель при разработке эффективного метода мультимодального объединения заключается в том, чтобы одновременно сохранить взаимосвязи между модальностями (inter-modality interactions) и обеспечить практическую реализуемость алгоритма с точки зрения вычислительной сложности и потребления памяти.

В ходе анализа существующих подходов было установлено, что при слиянии данных с использованием различных алгоритмов далеко не всегда удаётся сохранить все зависимости между модальностями. Полное сохранение межмодальных связей возможно только в том случае, если явно формируется тензор внешнего произведения признаков всех модальностей. Такой подход реализуется, например, в методе Tensor Fusion Network (TFN) [90], однако его применение ограничено из-за чрезмерно большого размера результирующего тензора, что делает хранение и обработку данных практически невозможными при увеличении числа модальностей или размерности признаков.

Для уменьшения вычислительной нагрузки были предложены методы на основе тензорных разложений, среди которых значительное распространение получил метод CP-разложения (CANDECOMP/PARAFAC Decomposition)[120] , лежащий в основе модели Low-rank Multimodal Fusion (LMF). Этот подход позволяет частично решить проблему экспоненциального роста параметров за счёт аппроксимации многомерного тензора низкоранговыми компонентами. Тем не менее, данный метод имеет ряд ограничений. Во-первых, взаимосвязи между модальностями сохраняются не полностью, поскольку ядро CP-разложения является диагональным и не отражает сложные перекрёстные взаимодействия

между различными модальностями [121]. Во-вторых, все модальности обрабатываются с одинаковым рангом, что не соответствует реальной природе мультимодальных данных, обладающих разной структурой, плотностью и степенью информативности. В большинстве практических случаев для адекватного представления данных требуется различная степень сжатия (разные ранги) для каждой модальности[121] .

Кроме того, современные системы мультимодального анализа требуют не только сохранения межмодальных зависимостей, но и оптимизации вычислительной эффективности[122]. Это подразумевает необходимость создания моделей, которые обеспечивают высокую точность при существенно меньших затратах памяти и времени вычислений. Таким образом, возникает задача разработки метода, который сочетает в себе:

- способность сохранять полные межмодальные взаимодействия;
- адаптивность к разнородности данных различных модальностей;
- и пониженную вычислительную сложность, обеспечивающую возможность практической реализации на современных вычислительных устройствах[122].

В следующем разделе представляется предложенный метод, направленный на решение перечисленных проблем и объединяющий преимущества существующих подходов при устранении их основных ограничений.

2.1.1. Постановка задачи

Задача мультимодального слияния данных заключается в построении модели, способной эффективно объединять информацию, поступающую из различных модальностей $\{D_1, D_2, \dots, D_M\}$, где каждая модальность $D_i \in \mathbb{R}^{d_i}$ имеет собственное пространство признаков, размерность и статистические свойства.

Цель состоит в построении отображения:

$$\mathcal{F}\colon \mathbb{R}^{d_1} \times \mathbb{R}^{d_2} \times \dots \times \mathbb{R}^{d_M} \to \mathbb{R}^{h}$$

такого, что результирующее представление $H = \mathcal{F}(D_1, D_2, \ldots, D_M)$ одновременно сохраняет внутримодальные и межмодальные зависимости, обеспечивая при этом низкую вычислительную сложность и устойчивость к избыточности данных.

2.1.2. Разложение Такера

Метод Tucker Decomposition был предложен Т. Такером в 1963 году[123] и представляет собой одно из наиболее фундаментальных направлений в тензорном анализе. Его основная идея заключается в аппроксимации исходного многомерного тензора $\mathcal{X}$ с помощью меньшего по размерности тензора ядра $\mathcal{g}$ и набора матриц-факторов, которые описывают линейные подпространства для каждой модальности данных [123].

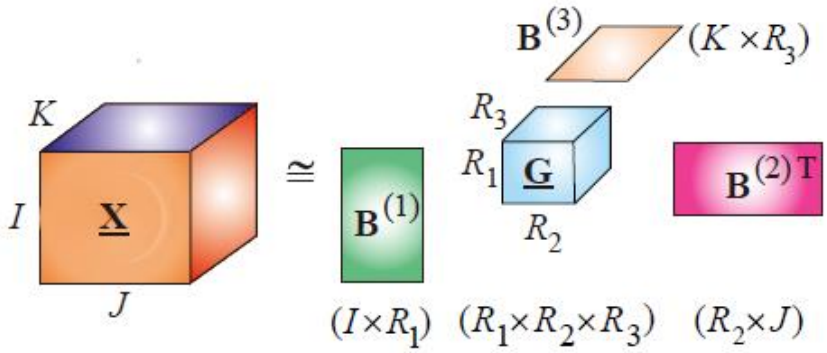


Рис. 49. Tucker-декомпозиция тензора

Задан трёхмерный тензор $\mathcal{X} \in \mathbb{R}^{I \times J \times K}$ Тогда его разложение Такера записывается в виде:

$$\mathcal{X} \approx \mathcal{g} \times_1 B^{(1)} \times_2 B^{(2)} \times_3 B^{(3)}$$

где $\mathcal{g} \in \mathbb{R}^{R_1 \times R_2 \times R_3}$ ядро (core tensor), хранящее взаимодействия между компонентами разных модальностей

$B^{(1)} \in \mathbb{R}^{I \times R_1}$, $B^{(2)} \in \mathbb{R}^{J \times R_2}$, $B^{(3)} \in \mathbb{R}^{K \times R_3}$ матрицы факторов, отображающие исходные признаки каждой модальности в пониженное пространство размерности $\mathbb{R}_n$

Операция $\times_n$ n-режимное (mode-n) произведение тензора на матрицу.

Для каждого элемента тензора $\mathcal{X}$ аппроксимация записывается как:

$$\mathcal{X}_{ijk} \approx \sum_{p=1}^{R_1} \sum_{q=1}^{R_2} \sum_{r=1}^{R_3} \mathcal{g}_{pqr}\, b_{ip}^{(1)}\, b_{jq}^{(2)}\, b_{kr}^{(3)}$$

Эта формула показывает, что каждый элемент исходного тензора выражается как линейная комбинация элементов ядра $\mathcal{g}$, взвешенных соответствующими компонентами матриц факторов. то в латентных координатах межмодальные «коэффициенты взаимодействий» это все элементы ядра $\mathcal{g}$ (не только диагональные), поэтому любые перекрёстные связи допустимы. Вклад каждой тройки латентных компонент (p, q, r) регулируется своим коэффициентом $\mathcal{g}_{pqr}$. Поскольку $\mathcal{g}$ не требуется быть диагональным, модель допускает все комбинации взаимодействий между модальностями а в CPD в латентных координатах взаимодействия ограничены супердиагональным ядром R, а коэффициенты для смешанных индексов (p, q, r) с различными метками отсутствуют [123].

Tucker рассматривал ядро $\mathcal{g}$ как обобщённую ковариационную матрицу, которая кодирует связи между разными измерениями данных. Если в матричном анализе сингулярное разложение (SVD) выделяет ортогональные направления максимальной дисперсии, то Tucker-разложение делает то же самое, но в многомерном пространстве и в результате разложение Такера можно рассматривать как многомерное обобщение SVD.

$$\mathcal{g} = \mathcal{X} \times_1 \left(B^{(1)}\right)^T \times_2 \left(B^{(2)}\right)^T \times_3 \left(B^{(3)}\right)^T$$

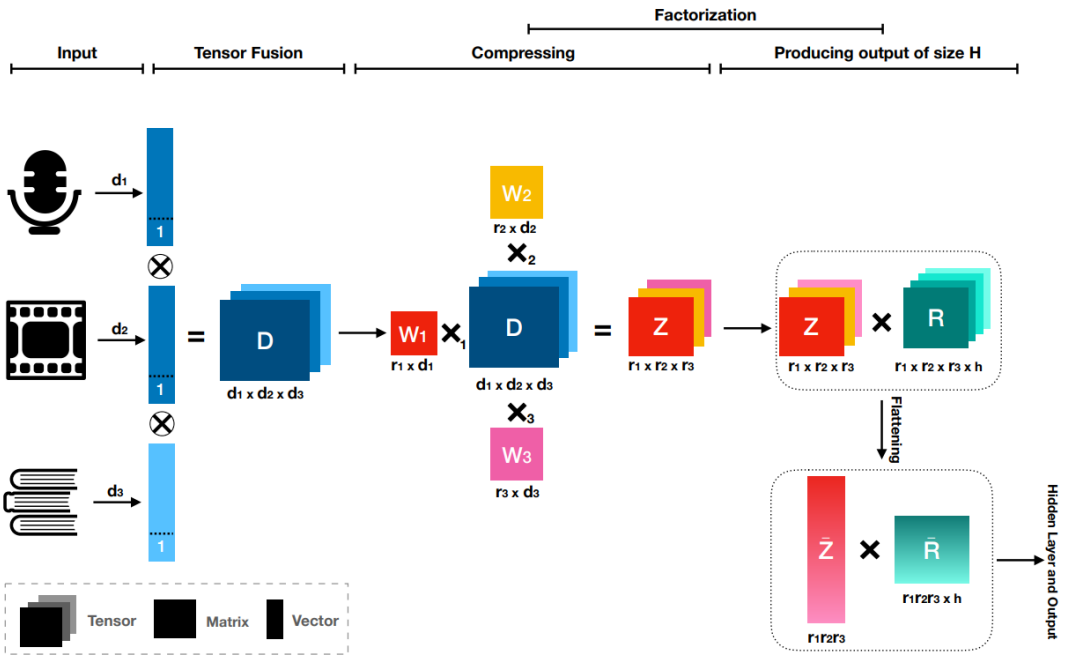


Рис. 50. Схема слияния мультимодальных данных на основе разложения Такера (Tucker)

Мы можем использовать подход разложения Такера для слияния мультимодальных данных и позволяет аппроксимировать тензор $\mathcal{X}$ компактной структурой:

$$\mathcal{X} \approx \mathcal{g} \times_1 \mathrm{B}^{(1)} \times_2 \mathrm{B}^{(2)} \times_3 \ldots \times_{\mathrm{M}} \mathrm{B}^{(\mathrm{M})}$$

Ядро $\mathcal{g}$ отражающее взаимосвязи между модальностями и матрицы-факторы $\mathrm{B}^{(1)} \in \mathbb{R}^{\mathrm{d_i} \times \mathrm{r_i}}$ отображающие каждую модальность в сжатое подпространство и $\mathrm{r_i} < \mathrm{d_i}$ ранги, определяющие степень сжатия каждой модальности.

Таким образом, каждая модальность имеет собственную матрицу проекции $\mathrm{B}^{(\mathrm{i})}$, которая сохраняет наиболее значимую информацию, в то время как ядро $\mathcal{g}$ описывает их многосторонние корреляции.

После применения Tucker-разложения мультимодальное представление выражается как [124]

$$\widehat{H} \approx \mathcal{g} \times_1 \mathrm{B}^{(1)} \times_2 \mathrm{B}^{(2)} \times_3 \ldots \times_{\mathrm{M}} \mathrm{B}^{(\mathrm{M})}$$

Полученное латентное представление $\widehat{H}$ подаётся в классификатор или регрессионный блок для выполнения целевой задачи [125]

$$\widehat{\mathrm{Y}} = \sigma\left(\mathrm{W_H}\widehat{\mathrm{H}} + \mathrm{b}\right)$$

Процесс оптимизации параметров направлен на минимизацию функции потерь:

$$\mathcal{L} = \mathcal{L}_{\text{task}}\left(\mathrm{Y}, \widehat{\mathrm{Y}}\right) + \lambda_1 \sum_{\mathrm{i}=1}^{\mathrm{M}} \left\|\mathrm{B}^{(\mathrm{i})}\right\|_{\mathrm{F}}^2 + \lambda_2 \|\mathcal{G}\|_{\mathrm{F}}^2$$

λ_1, λ_2 коэффициенты регуляризации

$\|.\|_{\mathrm{F}}^2$ фробениусова норма, используемая для ограничения роста весов.

Для обучения модели применяется стохастический градиентный спуск (SGD) или адаптивные оптимизаторы (Adam, RMSProp), что позволяет эффективно находить оптимальные параметры даже при большом количестве модальностей. Благодаря мультилинейной структуре Тукера, градиенты по каждому фактору вычисляются независимо, что значительно ускоряет процесс оптимизации и снижает риск переобучения.

Одним из ключевых преимуществ Tucker-разложения является резкое снижение числа параметров по сравнению с традиционным построением тензора внешнего произведения [126].

Использование Tucker-разложения приводит к полиномиальной зависимостивычислительная сложность метода оценивается как:

$$\mathcal{O}\left(\sum_{\mathrm{i}=1}^{\mathrm{M}} \mathrm{d_i r_i} + \prod_{\mathrm{i}=1}^{\mathrm{M}} \mathrm{r_i h}\right)$$

что при типичных значениях $\mathrm{r_i} \ll \mathrm{d_i}$ беспечивает многократное сокращение числа параметров и вычислительных операций [123]. Кроме того, операции n-режимного произведения $(\times_{\mathrm{n}})$ легко распараллеливаются на GPU, что делает модель пригодной для работы в реальном времени и для обучения на больших мультимодальных наборах данных [127].

2.1.3. Tensor Train разложение

В предыдущем подходе основная идея заключается в построении многомерного тензора $D = D_1 \otimes D_2 \otimes \ldots \otimes D_M$, отражающего все межмодальные

взаимодействия, и последующем обучении проекционного тензора весов Wс помощью тензорной факторизации вида [128]:

$$W = \mathcal{g} \times_1 W_1 \times_2 W_2 \times_3 \dots \times_{M+1} W_{M+1}$$

основанный на формировании внешнего произведения модальностей для захвата межмодальных зависимостей, с последующим разложением Такера тензора весов для снижения избыточности и числа параметров. Это позволяет модели адаптироваться к различным вкладам модальностей, минимизируя переобучение и улучшая интерпретируемость.

Однако разложение Такера масштабируется плохо при большом числе модальностей, а ядро $\mathcal{g}$ быстро становится высокоразмерным, что ограничивает применение в более сложных мультимодальных задачах.

TT-разложение, предложенное Оселедцем (Oseledets, 2011), представляет собой эффективный метод аппроксимации высокомерных тензоров через последовательную цепочку низкоранговых компонент. Для тензора $X \in \mathbb{R}^{I_1 \times I_2 \times \dots \times I_N}$

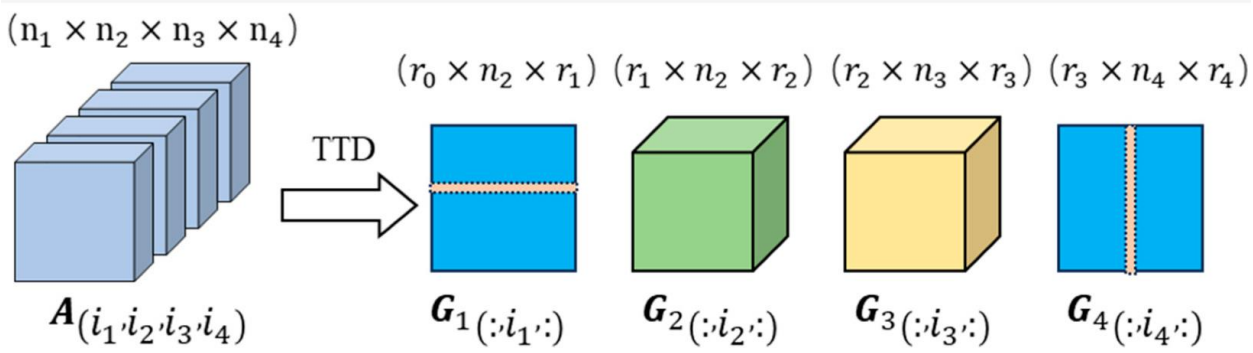


Рис. 51. TT-разложение (Tensor Train decomposition) тензора четвёртого порядка

TT-разложение определяется как:

$$X(i_1, i_2, \dots, i_N,) \approx \sum_{R_0=1}^{R_0} \sum_{R_1=1}^{R_1} \dots \sum_{R_{N-1}}^{R_{N-1}} G_1(r_0, i_1, r_1) G_2(r_1, i_2, r_2) \dots G_N(r_{N-1}, i_N, r_N)$$

где $G_k \in \mathbb{R}^{R_{k-1} \times R_k}$ трехмерные ядра, с граничными условиями $R_0 = R_N = 1$

Структура TT-разложения адаптируется для представления разнородных данных путем последовательной цепочки ядер, где каждый core соответствует моде тензора [129]. Ядро Такера может быть заменено TT-представлением, поскольку TT

аппроксимирует полный тензор через последовательные матричные произведения, минимизируя параметры.

В матричной форме это упрощается до:

$$X(i_1, i_2, \dots, i_N,) \approx G_1(i_1)G_2(i_2) \dots G_N(i_N)$$

где $G_k(i_k) \in \mathbb{R}^{R_{k-1} \times R_k}$ слайсы ядер. Ранги $\{R_k\}_{k=1}^{M}$ контролируют степень сжатия и выразительность, обеспечивая баланс между точностью и сложностью. Число параметров в TT-формате составляет $\mathcal{O}\left(\mathrm{d_y R_1} + \sum_{\mathrm{m=1}}^{\mathrm{M}} (\mathrm{d_m} + 1)\, \mathrm{R_m R_{m+1}}\right)$, что линейно по M и квадратично по рангам, в отличие от экспоненциального роста в полном W или ядре Такера $\mathcal{O}(\prod \mathrm{r_i})$.

TT обеспечивает гибкость через независимые $\mathrm{R_k}$, лучше адаптируясь к разнородным модальностям.

Алгоритм нахождения TT-разложения основан на последовательном SVD (TT-SVD):

1. Развернуть тензор в матрицу по первой моде: $A^{(1)} = unfold(X, 1) \in \mathbb{R}^{I_1 \times (I_2 \dots I_N)}$
2. Выполнить SVD: $A^{(1)} \approx U^{(1)} \Sigma^{(1)} \left(V^{(1)}\right)^T$, где $U^{(1)} \in \mathbb{R}^{I_1 \times R_1}$. Установить $G_1 = reshape\left(U^{(1)}, [R_0, I_1, R_1]\right)$ с $R_0 = 1$
3. Обновить остаток: $\mathrm{A^{(2)} = \Sigma^{(1)} \left(V^{(1)}\right)^T}$ переформировать в $\mathbb{R}^{R_1 I_2 \times (I_3 \dots I_N)}$
4. Повторить шаги 2–3 для мод $k = 2, \dots, N$ обрезая singular values для контроля рангов R_k

TT-разложение применяется для вычисления $h = W.D + b$ без тензоризации D и W , эксплуатируя структуру внешнего произведения D. Выход вычисляется последовательными матричными произведениями:

$$W(i_1, \dots, i_M, j) = G_0(j)G_1(i_1) \dots G_M(i_M) + b$$

где $\mathrm{G_0} \in \mathbb{R}^{1 \times h \times R_1}, \mathrm{G_M} \in \mathbb{R}^{R_{M-1} \times d_M \times 1}$.

Выходной вектор модели:

$$H(j) = \sum_{i_1, \dots, i_M} W(i_1, \dots, i_M, j)\, D_1(i_1) \dots \; D_M(i_M)$$

TT-регрессионный слой реализует линейное отображение мультимодального тензора X в выходное пространство через TT-веса W_{TT}

$$\hat{y} = W_{TT} X + b$$

Для обучения модели ядра G_k градиент $\frac{\partial L}{\partial G_k} = \sum_{i_1,\ldots,i_M} \frac{\partial L}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial G_k}$

Ядра G_k оптимизируются end-to-end с помощью градиентного спуска.

TT-разложение является эффективной и теоретически обоснованной альтернативой разложению Такера в архитектурах мультимодального слияния данных. Оно устраняет ключевые ограничения Такера чрезмерную размерность ядра и вычислительные издержки и сохраняет возможность модально-специфического анализа. T- разложение обеспечивает оптимальный компромисс между точностью и вычислительной эффективностью, что делает его предпочтительным для систем с большим числом модальностей и высокоразмерными признаками [130].

Однако TT имеет линейную структуру, что может ограничивать захват циклических зависимостей между модами. В качестве альтернативы предлагается Tensor Ring Decomposition (TR), которое обобщает TT за счет кольцевой топологии.

2.1.4. Tensor Ring Decomposition

TR-разложение представляет тензор высокого порядка как циклическую последовательность ядер, обобщая TT-разложение за счёт кольцевой топологии [131].

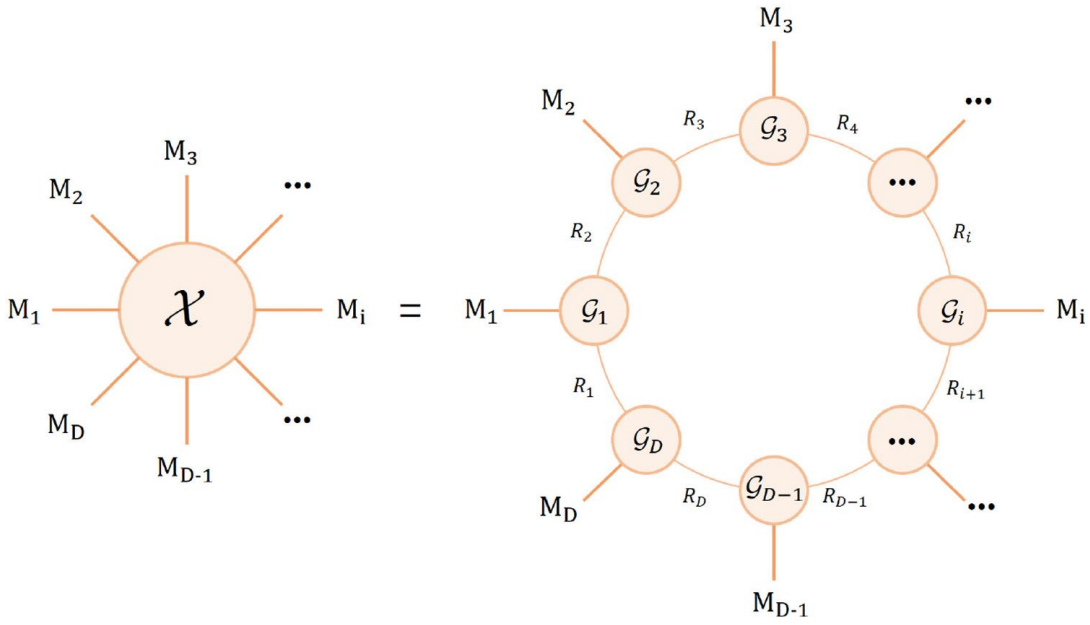


Рис. 52. Tensor Ring (TR) разложение многомерного тензора

Алгоритм TR-SVD является последовательным методом на основе сингулярного разложения (SVD), аналогичным TT-SVD, но с учётом циклической структуры [132]. Он аппроксимирует тензор $\mathcal{X} \in \mathbb{R}^{n_1 \times \ldots \times n_d}$ где каждая мода соответствует одной модальности. TT-разложение аппроксимирует тензор $\mathcal{X}$ в виде последовательной цепочки ядер:

$$\mathcal{X}(i_1, \ldots, i_N) = G_1[i_1] G_2[i_2] \ldots G_N[i_N]$$

где каждое ядро $\mathrm{G_N} \in \mathbb{R}^{\mathrm{r_{n-1}} \times \mathrm{I_n} \times \mathrm{r_n}}$ с граничными условиями $\mathrm{R_0} = \mathrm{R_N} = 1$, TR-разложение снимает эти условия, позволяя $\mathrm{R_0} = \mathrm{R_N} = \mathrm{R}$ [131], замыкая структуру в кольцо:

$$\mathcal{X}(i_1, \ldots, i_N) = TR\,(G_1[i_1] G_2[i_2] \ldots G_N[i_N])$$

Для TR-формата тензор $\mathcal{X}$ представляется аналогично TT-формата , но с кольцевой структурой:

$$\hat{\mathrm{y}} = \langle \mathrm{W}, \mathcal{X} \rangle = \sum_{\mathrm{i_1}, \ldots, \mathrm{i_n}} \mathrm{W}(\mathrm{i_1}, \ldots, \mathrm{i_N}) \mathcal{X}(\mathrm{i_1}, \ldots, \mathrm{i_N})$$

где W хранится в TR-формате. Таким образом:

$$\hat{y} = TR\left(\prod_{n=1}^{N} G_n^{(W)}[i_n]\right) \prod_{n=1}^{N} x_n[i_n]$$

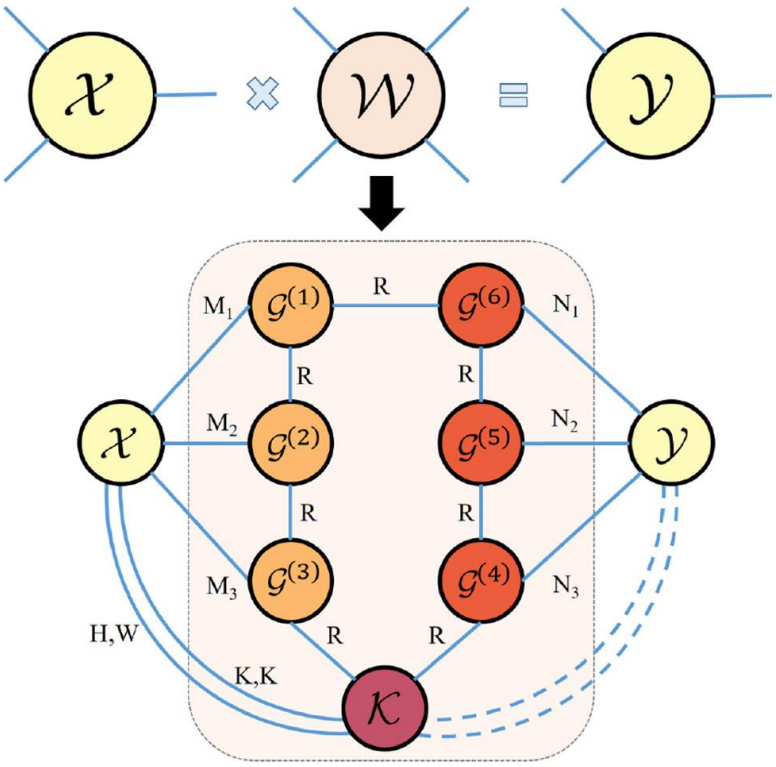


Рис. 53. Схема обучения модели на основе Tensor Ring с end-to-end оптимизацией

Обучение осуществляется end-to-end через дифференцируемую схему, аналогичную TT-регрессии, с обновлением ядер $G_n^{(W)}$ посредством градиентного спуска.

Общее количество параметров $P_{TR} = \sum_{n=1}^{N} I_n r_{n-1} r_n$, Ожидается, что TR-подход позволит работать с данными на 2-3 порядка большей размерностипри сокращении потребления памяти на 15-25% по сравнению с TT, с повышением точности на 5-10% в задачах с циклическими зависимостями.

Вторая глава диссертации посвящена разработке методического подхода к решению задачи мультимодального слияния данных, с акцентом на сохранение межмодальных и внутримодальных зависимостей при минимизации вычислительной сложности и потребления ресурсов. Анализ существующих методов, таких как Tensor Fusion Network (TFN) и Low-rank Multimodal Fusion (LMF) на основе CP-разложения (CANDECOMP/PARAFAC), выявил ключевые ограничения: экспоненциальный рост размерности тензора в TFN, приводящий к неэффективности хранения и обработки, а также неполное сохранение межмодальных взаимодействий в LMF из-за диагональной структуры ядра и фиксированного ранга для всех модальностей. Эти недостатки мотивировали поиск более адаптивных и масштабируемых решений, способных учитывать разнородность мультимодальных данных.

В качестве базового подхода предложено применение разложения Такера (Tucker Decomposition), которое аппроксимирует тензор внешнего произведения признаков модальностей через компактное ядро и матрицы-факторов с независимыми рангами для каждой модальности. Это обеспечивает полное сохранение межмодальных корреляций, поскольку ядро функционирует как обобщенная ковариационная матрица, допускающая все комбинации взаимодействий, в отличие от супердиагональной структуры CP-разложения. Оптимизация модели осуществляется путем минимизации функции потерь с регуляризацией Фробениусовой нормой, используя стохастический градиентный спуск или адаптивные оптимизаторы (например, Adam). Вычислительная сложность снижается до полиномиальной зависимости $\mathcal{O}\left(\sum_{i=1}^{N} I_i R_i + \prod_{i=1}^{N} R_i\right), I_i$

размерность модальности,а R_i ранг, что позволяет эффективно обрабатывать данные на GPU и в реальном времени. Тем не менее, при увеличении числа модальностей ядро Такера становится высокоразмерным, ограничивая масштабируемость.

Для преодоления этих ограничений рассмотрено Tensor Train (TT)-разложение, представляющее тензор как последовательную цепочку низкоранговых ядер с граничными условиями. TT-аппроксимация минимизирует число параметров до $\mathcal{O}\left(\sum_{i=1}^{N} R_{i-1}\, d_i R_i\right)$, где d_i размерность моды i, обеспечивая линейную зависимость от размерности и квадратичную от рангов, в отличие от экспоненциального роста в полном тензоре или ядре Такера. Алгоритм TT-SVD позволяет вычислять разложение последовательно, а end-to-end обучение через градиентный спуск адаптирует модель к разнородным модальностям. TT сохраняет выразительность для модально-специфического анализа, но его линейная структура может недостаточно эффективно захватывать циклические зависимости между модальностями.

В качестве обобщения и улучшения предложено Tensor Ring (TR)-разложение, которое снимает граничные условия TT, замыкая структуру в кольцо и тем самым усиливая способность модели к представлению циклических и сложных взаимодействий. TR-аппроксимация тензора осуществляется через циклическую последовательность ядер, с алгоритмом TR-SVD, аналогичным TT-SVD, но учитывающим кольцевую топологию. Общее число параметров остается на уровне $\mathcal{O}\left(\sum_{i=1}^{N} R^2\, d_i\right)$, что обеспечивает дополнительное сокращение памяти на 15–25% по сравнению с TT при повышении точности на 5–10% в задачах с циклическими зависимостями. Обучение модели также проводится end-to-end с использованием градиентного спуска, интегрируя TR-структуру в регрессионный слой для мультимодального слияния.

В итоге, предложенный метод на основе TR-разложения интегрирует преимущества Такера (полное сохранение взаимодействий) и TT (масштабируемость и низкая сложность), одновременно устраняя их недостатки за счет кольцевой топологии, которая лучше адаптируется к высокоразмерным и

разнородным мультимодальным данным. Это обеспечивает оптимальный баланс между точностью, вычислительной эффективностью и интерпретируемостью, делая TR предпочтительным для практических приложений в задачах с большим числом модальностей и позволяя работать с данными на 2–3 порядка большей размерности. Полученные результаты закладывают основу для экспериментальной верификации в последующих главах, подтверждая потенциал TR в повышении производительности мультимодальных систем.

2.2. Метод снижения шума и восстановление информации в мультимодальных данных

Как было отмечено в первой части, шум оказывает существенное влияние на процесс слияния мультимодальных данных, а также играет фундаментальную роль в качестве одного из ключевых факторов, определяющих устойчивость и эффективность нейронных сетей. Устойчивые нейронные сети являются важнейшим элементом современных систем искусственного интеллекта, поскольку они обеспечивают надежность предсказаний в реальных условиях, где данные часто подвержены искажениям.

В предлагаемом методе предпринята попытка использовать тензорные подходы для снижения уровня шума и восстановления утраченных данных, что позволяет обеспечить согласованность между процессами мультимодального слияния данных и шумоподавления. В частности, анализируется возможность восстановления изображения при удалении до 90% пикселей, что актуально для обработки сигналов и изображений.

Рис. 54. Восстановление изображения при 90% пропущенных данных

Метод matrix completion широко применяется в области обработки сигналов для восстановления неполных данных, обладающих низкоранговой структурой.

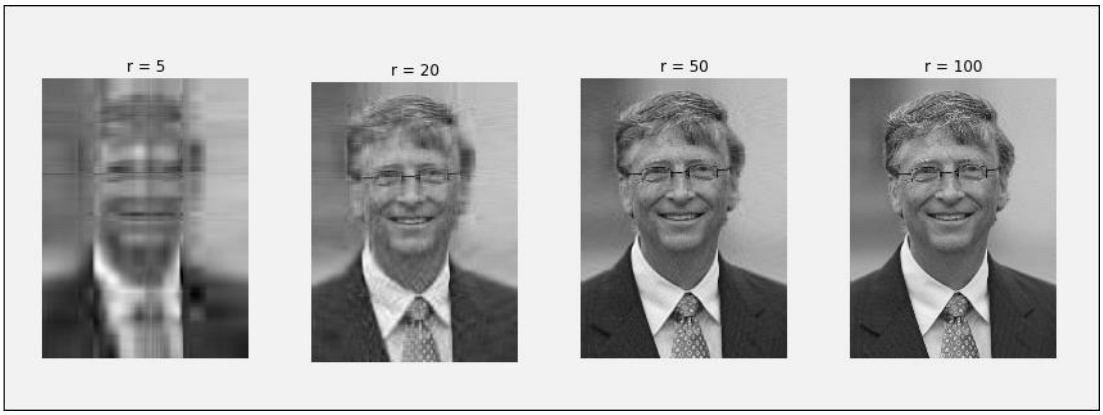
Наиболее полезным и определяющим свойством сингулярного разложения (SVD) является то, что оно обеспечивает оптимальное приближение низкого ранга r матрице X. Фактически, SVD обеспечивает иерархию аппроксимаций низкого ранга, так как аппроксимация ранга r получается путем сохранения ведущих r сингулярных значений и векторов. Теорема Эккарта – Янга [133]

$$\underset{\hat{X},rank(\hat{X})=r}{argmin} \left\|X-\hat{X}\right\|_F = \hat{U}\hat{\Sigma}\hat{V}^T$$

$$\|A\|_F = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n}\left(A_{ij}\right)^2}$$

Чтобы такое приближение было наилучшим и наиболее оптимальным используется фробениусова норма при ограничении $rank\left(\hat{X}\right) = r$. Таким образом строится аппроксимация исходной матрицы $\hat{X}$ к меньшему рангу. Выполняется своего рода сжатие информации матрицы X, которая является наилучшей в смысле квадратичной нормы.

$$\hat{X} = \sum_{k=1}^{r} \sigma_k \, u_r {v_k}^T = \sigma_1 u_1 {v_1}^T + \sigma_2 u_2 {v_2}^T + \cdots + \sigma_r u_r {v_r}^T$$



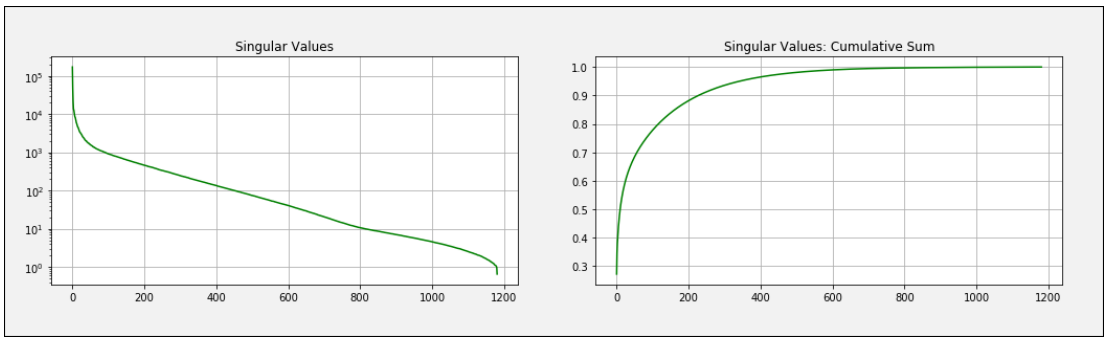


Рис. 55. Пример сингулярного разложения при сжатой картинке

Рассмотрим матрицу $M \in \mathbb{R}^{m \times n}$ представляющую исходные данные, где часть элементов наблюдаема, а остальные отсутствуют или зашумлены. Обозначим множество индексов наблюдаемых элементов как Ω. Задача состоит в нахождении матрицы M, аппроксимирующей M в наблюдаемых позициях и обладающей минимальным рангом.

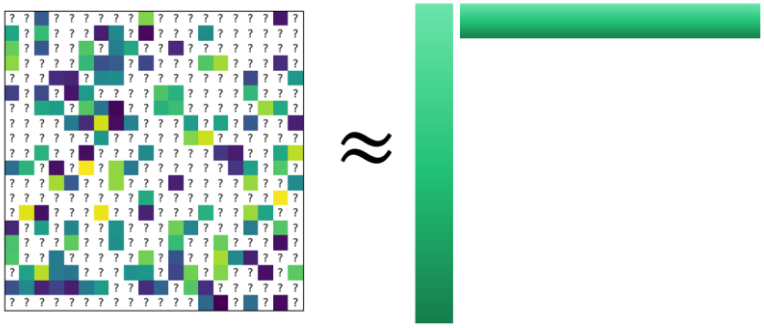


Рис. 56. Матричное восстановление на основе низкорангового приближения

Формулировка оптимизационной задачи имеет вид:

$$\min_X rank(X), \quad Subject\ to\ X_{i,j} = M_{i,j} \quad (i,j) \in \Omega$$

Однако минимизация ранга является NP-трудной задачей из-за невыпуклости функции ранга. Для приближенного решения используется релаксация с заменой ранга на ядерную норму $\|X\|_*$, определяемую как сумма сингулярных значений матрицы:

$$\|X\|_* = \sum_{k=1}^{\min(m,n)} \sigma_k(X)$$

где $\sigma_k(X)$ – k-е сингулярное значение матрицы X , полученное из сингулярного

разложения $X = \mathrm{U}\Sigma V^T$. Таким образом, задача преобразуется в:

$$\min_X \|X\|_*, \quad Subject\ to\ X_{i,j} = M_{i,j} \quad (i,j) \in \Omega$$

Эта формулировка является выпуклой и может быть решена с использованием алгоритмов типа проксимального градиентного спуска или чередующихся наименьших квадратов (ALS). Распределение сингулярных значений часто демонстрирует быстрый спад, что подтверждает низкоранговость данных, таких как черно-белые изображения. Метод эффективен для задач восстановления изображений, временных рядов и рекомендательных систем.

Метод matrix completion применим к двумерным данным, таким как черно-белые изображения. Однако в случае многомерных данных, например цветных изображений с RGB-каналами (тензор $M \in \mathbb{R}^{h \times w \times 3}$) прямое применение требует развертки тензора в матрицу, что приводит к потере корреляций между каналами. Это снижает качество восстановления.

Рис. 57. Сравнение матричного и тензорного представления изображений (grayscale и RGB)

Для сохранения многомерной структуры предлагается использовать tensor completion. Тензорный подход учитывает корреляции между измерениями, что особенно важно для изображений, где пиксели в разных каналах взаимосвязаны. Рассмотрим тензор $M \in \mathbb{R}^{I_1 \times I_2 \times \ldots \times I_d}$ с наблюдаемыми элементами в множестве Ω. Задача tensor completion формулируется как поиск низкорангового тензора X, совпадающего с M в Ω. Одна из распространенных формулировок использует разложение Такера (Tucker decomposition), где тензор аппроксимируется как:

$$X = \mathcal{G} \times_1 U_1 \times_2 U_2 \times_3 \ldots \times_d U_d$$

где $\mathcal{g} \in \mathbb{R}^{R_1 \times R_2 \times \ldots R_d}$ ядерный тензор, $U_k \in \mathbb{R}^{I_\mathrm{k} \times R_\mathrm{k}}$ факторные матрицы, а $\times_k$произведение по моде k. Многомерный ранг определяется кортежем $(R_1, R_2, \ldots, R_\mathrm{d})$ Оптимизационная задача может быть сформулирована как:

$$\min_{X,C} \|X - C\|_F^2, \quad Subject\ to\ rank(X) = R, P_\Omega(C) = P_\Omega(M)$$

где $\|X\|_F$ норма Фробениуса, R заданный многомерный ранг, P_Ω проектор на множество наблюдаемых элементов и C вспомогательный тензор.

Для решения применяется алгоритм чередующихся наименьших квадратов (ALS), чередуя обновления X и C:

- Обновление $X^{(n)} \approx \mathcal{L}\left(C^{(n)}\right)$, где $\mathcal{L}$ оператор низкоранговой аппроксимации.
- Обновление $C^{(n+1)} = \Omega \odot \quad M + (1 - \Omega) \odot X^{(n)}$

Метод поддерживает различные тензорные разложения, такие как CP (Canonical Polyadic), Tensor Train или Block Tensor Decomposition, но в данном случае используется Tucker decomposition для баланса между выразительностью и вычислительной эффективностью.

Предложенный базовый метод на основе tensor completion обеспечивает эффективное восстановление неполных или зашумленных данных, сохраняя многомерную структуру и корреляции. Он позволяет интегрировать шумоподавление в процесс мультимодального слияния, повышая устойчивость систем искусственного интеллекта.

Однако метод обладает ограничениями, такими как высокая вычислительная сложность ALS-алгоритма, особенно при большом числе итераций, и зависимость от выбора разложения тензора. Кроме того, время обработки может быть недостаточно для реального времени приложений, а эффективность снижается при сложных типах шума, не обладающих строгой низкоранговой структурой. Эти недостатки мотивируют разработку усовершенствованных подходов, сочетающих ускоренные алгоритмы и адаптивные стратегии, которые будут рассмотрены в последующих разделах.

Тензорные разложения представляют собой полезные методы, применяемые во множестве областей, таких как обработка сигналов, машинное обучение и глубокое обучение. Наиболее часто используемыми видами тензорных разложений

в литературе являются разложение Такера [134], [135], разложения train/ring [128], [131], [136], тензорное SVD (T-SVD) [137], разложение Кронекера [138], [139], разложение на блочные термы, а также ограниченные факторные разложения [140]. Тензоры и их разложения нашли широкое применение в таких задачах, как восстановление тензоров [141], рекомендательные системы, машинное обучение], идентификация систем Гаммерштейна и беспроводные коммуникации [142], [143].

Вычисление указанных тензорных разложений становится крайне трудоёмким при работе с крупномасштабными тензорами данных. В связи с этим за последнее десятилетие были предложены рандомизированные алгоритмы, направленные на ускорение данных процессов. Такие алгоритмы обладают рядом преимуществ по сравнению с традиционными детерминированными методами тензорного разложения. Они, как правило, быстрее и более экономны по памяти, что делает их особенно подходящими для обработки больших наборов данных. Кроме того, рандомизированные алгоритмы позволяют получать приближённые тензорные разложения с контролируемым уровнем точности, обеспечивая баланс между вычислительной стоимостью и степенью аппроксимации. Для получения более подробной информации о различных типах рандомизированных алгоритмов тензорного разложения см. Работы [144], [145].

С появлением больших данных всё чаще возникают крупномасштабные тензоры данных. Для их обработки требуются низкоранговые аппроксимации, применяемые во множестве задач, таких как восстановление тензоров [26] и рекомендательные системы [146]. Эта задача является сложной, особенно в случаях, когда данные хранятся вне основной памяти и распределены между несколькими вычислительными узлами. В таких условиях стоимость передачи данных становится основным узким местом и может превышать затраты на собственные вычисления.

В связи с этим чрезвычайно важно минимизировать количество проходов по данным, то есть обращаться к тензору данных как можно меньшее число раз. В контексте рандомизационного подхода такие методы получили название алгоритмов с эффективным проходом (pass-efficient algorithms).

Тензорное SVD (T-SVD) является широко используемым методом тензорного разложения, который нашёл применение во множестве задач, включая восстановление изображений и видео [147], распознавание лиц, сжатие данных и другие области. Этот метод разлагает тензор с помощью T-произведения (T-product) [148]трёх тензоров, что аналогично классическому SVD для матриц (формальное определение приведено в разделе 2).

Для работы с крупномасштабными тензорами было предложено несколько рандомизированных алгоритмов, позволяющих вычислять T-SVD с низким трубным рангом (low tubal rank) [149]. Недавно нами был предложен рандомизированный алгоритм с эффективным проходом [149] для T-SVD, применимый к задаче восстановления изображений и видео. Однако этот алгоритм требовал не менее двух проходов ($\nu \geq 2$) по данным. Известно, что во многих приложениях допустим лишь один проход, поэтому в данной работе мы предлагаем два эффективных однопроходных алгоритма ($\nu = 1$). Насколько нам известно, единственная работа, в которой ранее предлагались однопроходные рандомизированные алгоритмы для T-SVD [150].

Тем не менее алгоритмы, предложенные в [150], не являются оптимальными по нескольким причинам. Во-первых, авторы не используют быстрое T-произведение и t-QR-разложение, представленные в [151]. Во-вторых, данные алгоритмы являются обобщениями методов, предложенных для матриц в [152] и недавно адаптированных к тензорам в [153]. Следует отметить, что однопроходный алгоритм, основанный на выборке латеральных и горизонтальных срезов, также был разработан в [154]. В данной модели рассматривается пересечение латеральных и горизонтальных срезов, а в качестве средней компоненты в CUR-аппроксимации используется обратный по Муру–Пенроузу тензор U. Однако этот метод не обеспечивает стабильных результатов, если число фронтальных и латеральных срезов различается, что создаёт проблемы с кондиционированием.

Мы приводим результаты численных экспериментов и сравниваем наш подход с другими однопроходными рандомизированными алгоритмами. Для устранения указанных ограничений мы предлагаем несколько новых

однопроходных алгоритмов для вычисления T-SVD и проводим их подробное сравнение с базовыми методами. В частности, наши эксперименты показывают, что предложенные алгоритмы демонстрируют высокую устойчивость. Более того, мы применяем их для задач суперразрешения изображений и распознавания объектов, где они показывают конкурентные результаты по сравнению с существующими подходами.

Мы также рассматриваем рандомизированные алгоритмы с фиксированной точностью для вычисления аппроксимации третьего порядка тензоров с низким трубным рангом. Такие алгоритмы представляют интерес в случаях, когда оценка трубного ранга заранее неизвестна, и при заданной границе погрешности аппроксимации необходимо определить соответствующий оптимальный трубный ранг и построить низкоранговое приближение.

Проведённые широкие численные эксперименты показывают, что предложенный новый рандомизированный алгоритм с фиксированной точностью демонстрирует более высокую скорость и эффективность по сравнению с предыдущими методами.

В заключении основной вклад данной работы можно резюмировать следующим образом:

- Предложены более эффективные и устойчивые рандомизированные алгоритмы с фиксированной точностью для вычисления T-SVD;
- Разработаны три эффективных однопроходных алгоритма для вычисления T-SVD;
- Даны теоретические гарантии корректности и сходимости предложенных алгоритмов.

Проведены обширные численные эксперименты на синтетических и реальных данных, охватывающих задачи сжатия изображений и видео, суперразрешения изображений и глубокого обучения. Данная работа является первым исследованием, в котором применяются однопроходные алгоритмы для восстановления тензоров и суперразрешения изображений [155].

В этом разделе приводятся необходимые сведения о тензорах и вводится обозначение, используемое в дальнейшем изложении. Тензоры, матрицы и векторы обозначаются, соответственно, подчёркнутыми полужирными заглавными буквами $\underline{X}$, полужирными заглавными буквами X и полужирными строчными буквами x.

Для тензора третьего порядка $\underline{X}$ срезы $\underline{X}(:,:k)$, $\underline{X}(:,j,:)$ и $\underline{X}(i,:,:)$ называются соответственно фронтальными, латеральными и горизонтальными срезами. Также используется обозначение $\underline{X}_1$ для обозначения первого фронтального среза тензора $\underline{X}$. Для данного тензора третьего порядка $\underline{X}$, его подтензор $\underline{X}(i,j,:)$ называется трубкой (tube).

Норма Фробениуса тензора обозначается через $\|.\|_F$, а бесконечная норма – через $\|.\|_\infty$. Обозначение “conj” используется для обозначения комплексного сопряжения числа или покомпонентного комплексного сопряжения матрицы. Запись $\lceil n \rceil$ означает наименьшее целое число, большее или равное n.

Во всей работе мы рассматриваем только тензоры с действительными значениями, однако представленные результаты могут быть легко обобщены на случай комплекснозначных тензоров [155].

Два тензора могут быть объединены (конкатенированы) по первому или второму направлению (моде).Конкатенация по первой моде тензоров $\underline{A} \in \mathbb{R}^{I_1 \times I_2 \times I_3}$, $\underline{B} \in \mathbb{R}^{J_1 \times J_2 \times J_3}$ обозначается как $\underline{C} = \underline{A} \boxplus_1 \underline{B} \in \mathbb{R}^{(I_1+J_1) \times I_2 \times I_3}$, где $I_2 = J_2$ и $I_3 = J_3$. Аналогичное определение справедливо и для конкатенации по второй моде.

Альтернативные обозначения для конкатенации по первой и второй модам соответственно записываются как $\underline{A} \boxplus_1 \underline{B} = \left[\frac{\underline{A}}{\underline{B}}\right]$ и $\underline{A} \boxplus_2 \underline{B} = \left[\underline{A}, \underline{B}\right]$.

Определение 1. (T-произведение)

Пусть $\underline{X} \in \mathbb{R}^{I_1 \times I_2 \times I_3}$, $\underline{Y} \in \mathbb{R}^{I_2 \times I_4 \times I_3}$ тогда $\underline{X} * \underline{Y} \in \mathbb{R}^{I_1 \times I_4 \times I_3}$ это способ «умножать» тензоры 3-го порядка так, будто вы умножаете матрицы, но обычное скалярное умножение заменено свёрткой вдоль 3-го измерения (по “трубкам”) [156].

$$\underline{C} = \underline{X} * \underline{Y} = fold\big(circ(\underline{X})\, unfold(\underline{Y})\big), \tag{1}$$

где

$$circ(\underline{X}) = \begin{bmatrix} \underline{X}(:,:,1) & \underline{X}(:,:,I_3) & \cdots & \underline{X}(:,:,2) \\ \underline{X}(:,:,2) & \underline{X}(:,:,1) & \cdots & \underline{X}(:,:,3) \\ \vdots & \vdots & \ddots & \vdots \\ \underline{X}(:,:,I_3) & \underline{X}(:,:,I_3-1) & \cdots & \underline{X}(:,:,1) \end{bmatrix}$$

$$unfold(\underline{Y}) = \begin{bmatrix} \underline{Y}(:,:,1) \\ \underline{Y}(:,:,2) \\ \vdots \\ \underline{Y}(:,:,I_3) \end{bmatrix}, \underline{Y} = fold(unfold(\underline{Y}))$$

Быстрая реализация (эквивалентность через Фурье). Вместо явной circ(·) удобно перейти в частотную область по оси 3:

$$\|\underline{X}\|_F^2 = \frac{1}{I_3}\sum_{i=1}^{I_3} \|\hat{\underline{X}}(:,:,i)\|_F^2 \tag{2}$$

где $\underline{X}(:,:,i)$ обозначает i-й фронтальный срез тензора $\underline{X} = fft(\underline{X}, [\ \ ], 3)$, что соответствует вычислению быстрого преобразования Фурье (FFT) всех трубок тензора $\underline{X}$ вдоль его третьего измерения.

Algorithm 1: T-product in the Fourier domain

Input : Two data tensors $\underline{\mathbf{X}} \in \mathbb{R}^{I_1 \times I_2 \times I_3}$, $\underline{\mathbf{Y}} \in \mathbb{R}^{I_2 \times I_4 \times I_3}$
Output: T-product $\underline{\mathbf{C}} = \underline{\mathbf{X}} * \underline{\mathbf{Y}} \in \mathbb{R}^{I_1 \times I_4 \times I_3}$

$\widehat{\underline{\mathbf{X}}} = \text{fft}(\underline{\mathbf{X}}, [], 3)$;
$\widehat{\underline{\mathbf{Y}}} = \text{fft}(\underline{\mathbf{Y}}, [], 3)$;
for $i = 1, 2, \ldots, \lceil \frac{I_3+1}{2} \rceil$ **do**
  $\widehat{\underline{\mathbf{C}}}(:,:,i) = \widehat{\underline{\mathbf{X}}}(:,:,i)\ \widehat{\underline{\mathbf{Y}}}(:,:,i)$;
end
for $i = \lceil \frac{I_3+1}{2} \rceil + 1 \ldots, I_3$ **do**
  $\widehat{\underline{\mathbf{C}}}(:,:,i) = \text{conj}(\widehat{\underline{\mathbf{C}}}(:,:,I_3 - i + 2))$;
end
$\underline{\mathbf{C}} = \text{ifft}\left(\widehat{\underline{\mathbf{C}}}, [], 3\right)$;

Определение 2. (Транспонирование)

Пусть $\underline{X} \in \mathbb{R}^{I_1 \times I_2 \times I_3}$ заданный тензор. Тогда транспонированный тензор $\underline{X}$ обозначается как $\underline{X}^T \in \mathbb{R}^{I_2 \times I_1 \times I_3}$ и получается путём транспонирования всех

фронтальных срезов тензора и обращения порядка этих срезов, начиная со второго и до I_3. Тензор $\underline{X} \in \mathbb{R}^{I\times I\times K}$ называется симметричным, если выполняется равенство$\underline{X}^T = \underline{X}$.

Определение 3. (Единичный тензор)

Тензор $\underline{I} \in \mathbb{R}^{I_1\times I_1\times I_3}$ называется единичным, если его первый фронтальный срез является единичной матрицей размера $I_1 \times I_1$, а все остальные фронтальные срезы равны нулю. Легко показать, что для любых тензоров согласованных размеров выполняются равенства $\underline{I} * \underline{X} = \underline{X} * \underline{I}$.

Определение 4. (Ортогональный тензор)

Тензор $\underline{X} \in \mathbb{R}^{I_1\times I_2\times I_3}$ называется ортогональным, если $\underline{X}^T * \underline{X} = \underline{X} * \underline{X}^T = \underline{I}$

Определение 5. (Псевдообратный тензор по Муру–Пенроузу)

Пусть $\underline{X} \in \mathbb{R}^{I_1\times I_2\times I_3}$заданный тензор. Тогда псевдообратный тензор Мура–Пенроуза (Moore–Penrose pseudoinverse) обозначается как $\underline{X}^\dagger \in \mathbb{R}^{I_2\times I_1\times I_3}$ и определяется как единственный тензор, удовлетворяющий следующим четырём условиям:

$$\underline{X}^\dagger * \underline{X} * \underline{X}^\dagger = \underline{X}^\dagger, \qquad \underline{X} * \underline{X}^\dagger * \underline{X} = \underline{X}$$

$$\left(\underline{X} * \underline{X}^\dagger\right)^T = \underline{X} * \underline{X}^\dagger, \left(\underline{X}^\dagger * \underline{X}\right)^T = \underline{X}^\dagger * \underline{X}$$

Псевдообратный тензор по Муру–Пенроузу может быть также вычислен в области Фурье, что показано в Алгоритме 2.

Обратный тензор $\underline{X} \in \mathbb{R}^{I_1\times I_2\times I_3}$обозначается как $\underline{X} \in \mathbb{R}^{I_1\times I_1\times I_3}$и представляет собой частный случай псевдообратного тензора, для которого выполняется равенство $\underline{X}^{-1}*\underline{X} = \underline{X} * \underline{X}^{-1} = \underline{I}$

Algorithm 2: Fast Moore-Penrose pseudoinverse computation of the tensor$\underline{X}$

Input : The data tensor $\underline{\mathbf{X}} \in \mathbb{R}^{I_1 \times I_2 \times I_3}$
Output: Moore-Penrose pseudoinverse $\underline{\mathbf{X}}^{\dagger} \in \mathbb{R}^{I_2 \times I_1 \times I_3}$

$\widehat{\underline{\mathbf{X}}} = \text{fft}\left(\underline{\mathbf{X}}, [], 3\right);$
for $i = 1, 2, \ldots, \lceil \frac{I_3+1}{2} \rceil$ **do**
  $\widehat{\underline{\mathbf{C}}}\left(:, :, i\right) = \text{pinv}\left(\widehat{\underline{\mathbf{X}}}(:, :, i)\right);$
end
for $i = \lceil \frac{I_3+1}{2} \rceil + 1 \ldots, I_3$ **do**
  $\widehat{\underline{\mathbf{C}}}\left(:, :, i\right) = \text{conj}(\widehat{\underline{\mathbf{C}}}\left(:, :, I_3 - i + 2\right));$
end
$\underline{\mathbf{X}}^{\dagger} = \text{ifft}\left(\widehat{\underline{\mathbf{C}}}, [], 3\right);$

Определение 6. (f-диагональный тензор)

Если все фронтальные срезы тензора являются диагональными матрицами, то такой тензор называется f-диагональным тензором.

Определение 7. (Случайный тензор)

Тензор $\underline{\Omega}$ называется случайным, если его первый фронтальный срез $\underline{\Omega}(:,:,1)$ является стандартной гауссовской матрицей, а все остальные фронтальные срезы равны нулю.

2.2.1. Обобщение стандартных матричных разложений на тензоры с использованием T-произведения

T-произведение может быть использовано для обобщения стандартных матричных разложений, таких как QR, LU, собственные разложения (eigenvalue decompositions) и SVD, на случай тензоров, причём это выполняется достаточно естественным образом [155].

Пусть дан тензор $\underline{X} \in \mathbb{R}^{I_1 \times I_2 \times I_3}$ Тогда тензорное QR-разложение (T-QR) представляет его в виде $\underline{X} = \underline{Q} * \underline{R}$ и может быть вычислено с помощью Алгоритма

Algorithm 3: Fast T-QR decomposition of the tensor X

Input : The data tensor $\underline{\mathbf{X}} \in \mathbb{R}^{I_1 \times I_2 \times I_3}$
Output: The T-QR computation $\underline{\mathbf{X}} = \underline{\mathbf{Q}} * \underline{\mathbf{R}}$.

1 $\widehat{\underline{\mathbf{X}}} = \text{fft}\left(\underline{\mathbf{X}}, [], 3\right);$
2 **for** $i = 1, 2, \ldots, \lceil \frac{I_3+1}{2} \rceil$ **do**
3 $\quad [\widehat{\underline{\mathbf{Q}}}(:,:,i), \widehat{\underline{\mathbf{R}}}(:,:,i)] = \text{qr}\left(\widehat{\underline{\mathbf{X}}}(:,:,i), 0\right);$
4 **end**
5 **for** $i = \lceil \frac{I_3+1}{2} \rceil + 1 \ldots, I_3$ **do**
6 $\quad \widehat{\underline{\mathbf{Q}}}(:,:,i) = \text{conj}(\widehat{\underline{\mathbf{Q}}}(:,:,I_3 - i + 2));$
7 $\quad \widehat{\underline{\mathbf{R}}}(:,:,i) = \text{conj}(\widehat{\underline{\mathbf{R}}}(:,:,I_3 - i + 2));$
8 **end**
9 $\underline{\mathbf{Q}} = \text{ifft}\left(\widehat{\underline{\mathbf{Q}}}, [], 3\right);$
10 $\underline{\mathbf{R}} = \text{ifft}\left(\widehat{\underline{\mathbf{R}}}, [], 3\right);$

Можно вычислить тензорные разложения LU (T-LU), собственные разложения тензоров (T-EIG) и тензорное SVD (T-SVD) посредством незначительной модификации Алгоритма 3.

Более конкретно, в данной модификации операции SVD, LU и собственное разложение фронтальных срезов $\hat{\underline{X}}(:,:,i), i = 1,2,\ldots,I_3$ заменяются на QR-разложение, выполняемое на шаге 4 Алгоритма 3. Следует отметить, что Алгоритм 3 требует выполнения тонкого QR-разложения (thin QR) только для первых $\left\lceil \frac{I_3+1}{2} \right\rceil$ фронтальных срезов, в то время как исходный алгоритм T-QR, разработанный в [31, 6], включает QR-разложение всех срезов.

Поэтому рекомендуется использовать данную оптимизацию, чтобы избежать избыточных вычислений.

Например, тензорное собственное разложение (T-EIG) для симметричного тензора $\underline{X} \in \mathbb{R}^{I_1 \times I_1 \times I_3}$ гарантирует существование следующего разложения:

$$\underline{X} = \underline{V} * \underline{D} * \underline{V}^T \tag{3}$$

где $\underline{D}$ это f-диагональный тензор.

Разложение (3) также может быть записано в виде

$$\underline{X} = \left(\underline{V} * \underline{D}^{\frac{1}{2}}\right) * \left(\underline{D}^{\frac{1}{2}} * \underline{V}^T\right) \tag{4}$$

где $\underline{D}^{\frac{1}{2}} = \sqrt{\underline{D}}$ это f-диагональный тензор, элементы (трубки) которого вычисляются путём взятия квадратного корня из соответствующих элементов диагональных трубок тензора $\underline{D}$, то есть $\underline{D}(i,i,:)$.

---

1 $\widehat{\underline{\mathbf{D}}}(i,i,:) = \mathrm{fft}(\underline{\mathbf{D}}(i,i,:),[],3);$
2 **for** $j = 1 : I_3$ **do**
3 $\quad \widehat{\underline{\mathbf{D}}}(i,i,j) = \mathrm{sqrt}(\widehat{\underline{\mathbf{D}}}(i,i,j));$
4 **end**
5 $\underline{\mathbf{D}}^{1/2}(i,i,:) = \mathrm{ifft}(\widehat{\underline{\mathbf{D}}}(i,i,:),[],3);$

Следует также отметить, что тензорное сингулярное разложение (T-SVD) является полезным видом тензорного разложения, представляющим тензор как T-произведение трёх тензоров. При этом первый и последний тензоры являются ортогональными, а средний тензор f-диагональным.

Пусть $\underline{X} \in \mathbb{R}^{I_1 \times I_1 \times I_3}$ заданный тензор. Тогда его T-SVD представляется в виде

$$\underline{X} = \underline{U} * \underline{S} * \underline{V}^T \tag{5}$$

где $\underline{U} \in \mathbb{R}^{I_1 \times I_1 \times I_3}$ и $\underline{V} \in \mathbb{R}^{I_2 \times I_2 \times I_3}$ ортогональные тензоры, а $\underline{S} \in \mathbb{R}^{I_1 \times I_2 \times I_3}$ f-диагональный тензор.Количество ненулевых диагональных трубок в тензоре $\underline{S}$ называется трубным рангом (tubal rank).

Усечённое T-SVD (truncated T-SVD) определяется посредством усечения тензоров $\underline{U}$, $\underline{S}$ и $\underline{V}$. Например, усечённое разложение T-SVD с трубным рангом R для тензора $\underline{X}$ записывается как

$$\underline{X} = \underline{U}_R * \underline{S}_R * \underline{V}_R^T \tag{6}$$

где $\underline{U}_R = \underline{U}(:,1:R,:) \in \mathbb{R}^{I_1 \times R \times I_3}$, $\underline{V}_R = \underline{V}(:,1:R,:) \in \mathbb{R}^{R \times I_2 \times I_3}$, $\underline{S}_R = \underline{S}(1:R,1:R,:) \in \mathbb{R}^{R \times R \times I_3}$

На рисунке 2 показана структура полного и усечённого разложения T-SVD.

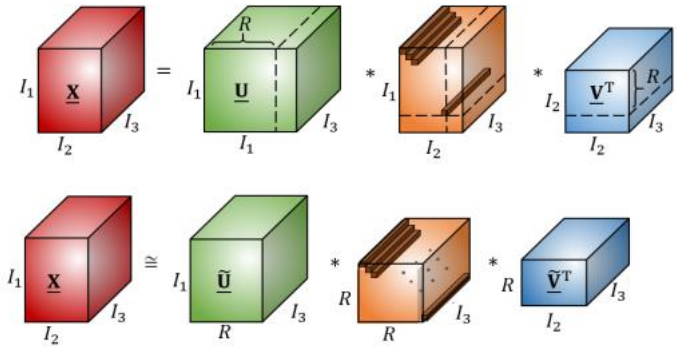


Рис. 58. Иллюстрация: (a) тензорного сингулярного разложения (T-SVD) и (b) усечённого T-SVD для тензора третьего порядка.

2.2.2. Предложенные рандомизированные однопроходные алгоритмы

Рандомизированные однопроходные алгоритмы представляют собой класс методов, которые обрабатывают поток данных за один проход, без необходимости хранить весь набор данных в оперативной памяти. Это свойство является крайне важным при работе с крупномасштабными данными, которые не помещаются в память целиком.

Такие алгоритмы играют ключевую роль в обработке больших данных и обеспечивают эффективные решения в различных приложениях. Их способность выполнять вычисления за один проход, снижая использование памяти и справляясь с наихудшими сценариями, делает их особенно ценными для анализа потоковых данных в реальном времени, машинного обучения и потоковой передачи данных.

По этим причинам разработка однопроходных и, в более общем случае, эффективных по числу проходов (pass-efficient) алгоритмов для быстрой низкоранговой аппроксимации тензоров является одной из наиболее активно исследуемых тем последнего десятилетия[155], [156], [157].

В данной работе внимание сосредоточено на T-SVD, и предлагаются новые рандомизированные однопроходные алгоритмы. Насколько нам известно, единственные ранее предложенные методы для низкоранговой аппроксимации тензоров с малым трубным рангом были представлены в работах [154] и [150].

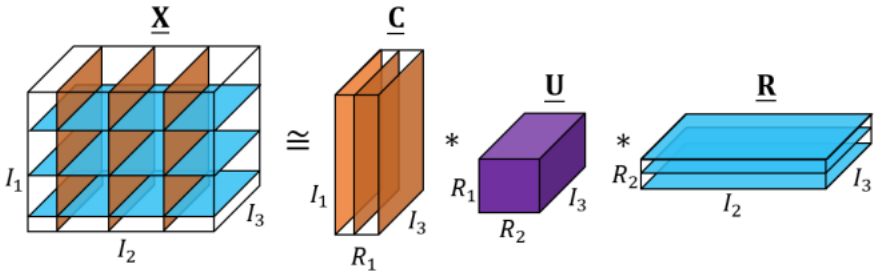


Рис. 59. Рандомизированная аппроксимация с низким трубным рангом, основанная на выборке срезов (slice sampling) [154]

Подход, описанный в [154], основан на приближении с использованием перекрёстных тензоров (cross tensor approximation). В частности, для заданного тензора $\underline{X} \in \mathbb{R}^{I_1 \times I_2 \times I_3}$ низкоранговая аппроксимация вычисляется на основе выборки нескольких латеральных и горизонтальных срезов исходного тензора $\underline{X}$ Если обозначить $\underline{C} \in \mathbb{R}^{I_1 \times L \times I_3}$ и $\underline{R} \in \mathbb{R}^{K \times I_2 \times I_3}$ как выбранные латеральные и горизонтальные срезы тензора $\underline{X}$,тогда низкоранговая аппроксимация тензора может быть вычислена следующим образом:

$$\underline{X} \approx \underline{C} * \underline{U} * \underline{R} \tag{7}$$

где $\underline{U} = \underline{C}^{\dagger} * \underline{X} * \underline{R}^{\dagger}$Средний тензор $\underline{U}$ является наилучшим кандидатом, поскольку он минимизирует остаток $\left\|\underline{X} - \underline{C} * \underline{U} * \underline{R}\right\|_F$. Очевидно, что данный подход требует только одного прохода по данным для вычисления среднего тензора $\underline{U}$. Наши экспериментальные результаты подтверждают, что данный метод стабилен в отношении выбора количества латеральных и горизонтальных срезов. Однако этот подход имеет три основных недостатка:

1. Точность метода обычно ниже по сравнению с другими техниками, поскольку тензоры $\underline{C}$ и $\underline{R}$ не являются точными оценками левых и правых сингулярных тензоров.
2. Для достижения требуемой точности метод, как правило, требует большего трубного ранга.

3. Вычисление выражения $\underline{C}^{\dagger} * \underline{X} * \underline{R}^{\dagger}$ является более вычислительно затратным, чем умножение с использованием случайных тензоров, поскольку можно применять структурированные случайные тензоры, ускоряющие вычисления.

В работе [37] предлагается использовать тензор $\underline{U}$ как пересечение выбранных латеральных и горизонтальных срезов исходного тензора $\underline{X}$ (см. рисунок 2.11).

Этот новый метод демонстрирует высокую скорость, однако подвержен той же нестабильности, когда L = K, то есть число латеральных срезов совпадает с числом горизонтальных срезов. Данный алгоритм представлен в алгоритме 4.

Algorithm 4: The single-pass cross tensor approximation [154]

Input : The data tensor $\underline{\mathbf{X}} \in \mathbb{R}^{I_1 \times I_2 \times I_3}$, two parameters L, K
Output: A low tubal rank approximation $\underline{\mathbf{X}} \approx \underline{\mathbf{C}} * \underline{\mathbf{U}} * \underline{\mathbf{R}}$
1 Select L lateral slices $\underline{\mathbf{C}}$ with corresponding indices $\mathcal{L}$;
2 Select K horizontal slices $\underline{\mathbf{R}}$ with corresponding indices $\mathcal{K}$;
3 Construct the intersection tensor $\underline{\mathbf{W}} = \underline{\mathbf{X}}(\mathcal{L}, \mathcal{K}, :)$;
4 Compute the middle tensor $\underline{\mathbf{U}} = \underline{\mathbf{W}}^{\dagger}$;
5 Compute a low tubal rank approximation $\underline{\mathbf{X}} \approx \underline{\mathbf{C}} * \underline{\mathbf{U}} * \underline{\mathbf{R}}$;

Предложенный в работе [150] однопроходный метод является обобщением однопроходного матричного алгоритма [152]на случай тензоров. Более точно, пусть дан тензор $\underline{X} \in \mathbb{R}^{I_1 \times I_2 \times I_3}$ Тогда на первом шаге из тензора $\underline{X}$ вычисляются два эскиза (sketches) путём умножения его на два случайных тензора $\underline{\Omega}_1 \in \mathbb{R}^{I_1 \times K \times I_3}$ и $\underline{\Omega}_2 \in \mathbb{R}^{L \times I_2 \times I_3}, (K \leq L)$
следующим образом:

$$\underline{Y} = \underline{X} * \underline{\Omega}_1 \in \mathbb{R}^{I_1 \times K \times I_3}, \qquad \underline{W} = \underline{\Omega}_2 * \underline{X} \in \mathbb{R}^{L \times I_2 \times I_3}, \tag{8}$$

Пусть $\underline{Q} \in \mathbb{R}^{I_1 \times K \times I_3}$ это ортонормированный базис для пространства, порождённого тензором $\underline{Y}$, вычисленный с помощью T-QR алгоритма. Тогда низкоранговая трубная аппроксимация тензора $\underline{X}$ вычисляется как

$$\underline{X} \approx \underline{Q} * \left(\underline{Q}^T * \underline{X}\right) \tag{9}$$

Здесь L и K называются размерами эскизов (sketch sizes), а тензоры $\underline{Y}$ и $\underline{W}$ эскизами тензора $\underline{X}$. Однако, поскольку при вычислении (8) данные тензора $\underline{X}$ были уже использованы один раз, необходимо избежать повторного обращения к исходным данным в выражении (9), так как вычисление $\underline{Q}^T * \underline{X}$ потребовало бы повторного прохода по тензору $\underline{X}$. Это выражение можно оценить приближённо, используя следующую формулу:

$$\underline{Q}^T * \underline{X} \approx \left(\underline{\Omega}_2 * \underline{Q}\right)^{\dagger} * W \tag{10}$$

Таким образом, данный однопроходный (или одновидовой, 1-view) метод сочетает построение диапазона и сопряжённого диапазона (range и co-range sketches) в одном проходе по тензору данных $\underline{X}$, а затем формирует низкоранговую трубную аппроксимацию на основе информации, содержащейся в вычисленных эскизах. Эти алгоритмы обобщены в алгоритме 5.

Algorithm 5: The Single-pass algorithm proposed in [150]

Input : The data tensor $\underline{\mathbf{X}} \in \mathbb{R}^{I_1 \times I_2 \times I_3}$, two parameters $L,\ K$
Output: A low tubal rank approximation $\underline{\mathbf{X}} \approx \underline{\mathbf{Q}} * \underline{\mathbf{B}}$

1 $\underline{\mathbf{\Omega}}_1 = \mathrm{randn}(I_2, K, I_3), \quad \underline{\mathbf{\Omega}}_2 = \mathrm{randn}(L, I_1, I_3)$;
2 Compute two sketches:
$\underline{\mathbf{Y}} = \underline{\mathbf{X}} * \underline{\mathbf{\Omega}}_1 \in \mathbb{R}^{I_1 \times K \times I_3}, \quad \underline{\mathbf{W}} = \underline{\mathbf{\Omega}}_2 * \underline{\mathbf{X}} \in \mathbb{R}^{L \times I_2 \times I_3}$;
3 Apply the T-QR applied to $\underline{\mathbf{Y}}$ and obtain the tensor $\underline{\mathbf{Q}}$;
4 Compute the tensor $\underline{\mathbf{B}} = (\underline{\mathbf{\Gamma}} * \underline{\mathbf{Q}})^{\dagger} * \underline{\mathbf{W}}$;
5 Compute the low tubal rank approximation $\underline{\mathbf{X}} \approx \underline{\mathbf{Q}} * \underline{\mathbf{B}}$;

Рассмотрим ситуацию потоковой обработки данных (streaming setting), когда тензор $\underline{X}$ не хранится целиком в оперативной памяти, а представлен в виде конечного потока линейных обновлений, то есть

$$\underline{X} = \sum_{n=1}^{N} \underline{H}_n$$

Здесь каждый тензор $\underline{H}_n$ удаляется сразу после использования, и мы можем постепенно выбирать выборки из $\underline{X}$, поскольку каждый новый инновационный тензор $\underline{H}_n$ подаётся на вход следующим образом:

$$\underline{X} * \underline{\Omega}_1 = \sum_{n=1}^{N} \underline{H}_n * \underline{\Omega}_1, \qquad \underline{\Omega}_2 * \underline{X} = \sum_{n=1}^{N} \underline{\Omega}_2 * \underline{H}_n$$

Это пример ситуации, в которой однопроходные алгоритмы особенно эффективны, поскольку каждый $\underline{H}_n$ может представлять собой разреженный тензор, содержащий лишь небольшое число элементов исходного тензора $\underline{X}$.

Как отмечалось в работе [36], однопроходные алгоритмы для матриц обычно дают ненадёжные аппроксимации, когда L=K, из-за нестабильности при решении плохо обусловленных задач наименьших квадратов. На практике мы наблюдали аналогичные проблемы и для тензоров, что будет показано в наших численных экспериментах. Согласно уравнению (10), необходимо решить линейное тензорное уравнение $\left(\underline{\Omega}_2 * \underline{Q}\right) * \underline{Y} = \underline{W}$ Коэффициентный тензор $\underline{\Omega}_2 * \underline{Q} \in \mathbb{R}^{L \times K \times I_3}$ содержит L горизонтальных и K латеральных срезов соответственно.

Один из способов избежать плохой обусловленности коэффициентного тензора заключается в том, чтобы выбрать $L > K$, тем самым получая переопределённую и лучше обусловленную задачу. Однако этот подход требует тонкой настройки параметров L и K, чтобы достичь приемлемого уровня аппроксимации, что влечёт за собой дополнительные вычислительные затраты на тщательный подбор оптимальных значений. В противном случае возникает риск потери точности. С другой стороны, как отмечено в [36], ещё одной причиной проблем при выборе L=K является то, что значение K может стать узким местом (bottleneck) для вычислений в некоторых приложениях. При ограниченных вычислительных ресурсах возникает соблазн установить L=K, чтобы сохранить как можно больше информации. Тем не менее, необходимость хранения как

диапазонных, так и сопряжённых эскизов, а также случайных выборок накладывает ограничение по памяти на метод с одним просмотром данных (1-view approach). Следует также отметить, что в работе [152], [156] предлагается использовать предварительную информацию о сингулярных значениях матрицы, чтобы подобрать значения L и K. Однако этот подход имеет ограниченное практическое применение, поскольку в реальных условиях такая информация обычно недоступна. Стоит подчеркнуть, что во всех наших предложенных однопроходных алгоритмах тензоры-эскизы $\underline{Y}_c$ или $\underline{Y}_r$ значительно меньше исходного тензора данных, и вычисление их левых сингулярных векторов не требует больших затрат по сравнению с построением полного ортонормированного базиса. Особенно важно отметить, что в однопроходных методах, умножение тензора $\underline{X}$ на случайные тензоры (для извлечения диапазона и сопряжённого диапазона) обычно является наиболее вычислительно затратной и временно сложной операцией.

Чтобы преодолеть указанный недостаток, были разработаны более устойчивые и усовершенствованные алгоритмы [153] для матричного случая. Мотивированные их эффективностью, мы обобщили эти методы на случай тензоров, и результаты приведены в алгоритмах 6 и 7.

Algorithm 6: Proposed randomized single-pass algorithm I

Input : The data tensor $\underline{\mathbf{X}} \in \mathbb{R}^{I_1 \times I_2 \times I_3}$, three parameters $L,\ K,\ H$, $L \geq K \geq H \geq 0$ and a target tubal rank R

Output: A low tubal rank approximation $\underline{\mathbf{X}} \approx \underline{\mathbf{U}} * \underline{\mathbf{S}} * \underline{\mathbf{V}}^T$

$\underline{\mathbf{\Omega}}_1 = \text{randn}(I_2, K+R, I_3), \quad \underline{\mathbf{\Omega}}_2 = \text{randn}(I_1, L+R, I_3)$;

$\underline{\mathbf{Y}}_c = \underline{\mathbf{X}} * \underline{\mathbf{\Omega}}_1, \quad \underline{\mathbf{Y}}_r = \underline{\mathbf{X}}^T * \underline{\mathbf{\Omega}}_2$;

if $H < K$ **then**

  $[\underline{\mathbf{Q}}_c, \underline{\mathbf{R}}_c] = \text{T-QR}(\underline{\mathbf{Y}}_c)$;

  $[\widehat{\underline{\mathbf{Q}}}_c, \sim, \sim] = \text{Truncated T-SVD}(\underline{\mathbf{R}}_c, R+H)$;

  $\underline{\mathbf{Q}}_c = \underline{\mathbf{Q}}_c * \widehat{\underline{\mathbf{Q}}}_c$;

else

  $[\underline{\mathbf{Q}}_c, \sim] = \text{T-QR}(\underline{\mathbf{Y}}_c)$;

end

$[\widehat{\underline{\mathbf{Q}}}, \widehat{\underline{\mathbf{R}}}] = \text{T-QR}(\underline{\mathbf{\Omega}}_2^T * \underline{\mathbf{Q}}_c)$;

$\underline{\mathbf{Z}} = \widehat{\underline{\mathbf{R}}}^{-1} * \widehat{\underline{\mathbf{Q}}}^T * \underline{\mathbf{Y}}_r^T$;

$[\widehat{\underline{\mathbf{U}}}, \underline{\mathbf{S}}, \underline{\mathbf{V}}] = \text{Truncated T-SVD}(\underline{\mathbf{Z}}, R)$;

$\underline{\mathbf{U}} = \widehat{\underline{\mathbf{Q}}} * \widehat{\underline{\mathbf{U}}}$;

Algorithm 7: Proposed randomized single-pass algorithm II

Input : The data tensor $\underline{\mathbf{X}} \in \mathbb{R}^{I_1 \times I_2 \times I_3}$, three parameters L, K, H, $L \geq K \geq H \geq 0$ and a target tubal rank R.
Output: A low tubal rank approximation $\underline{\mathbf{X}} \approx \underline{\mathbf{U}} * \underline{\mathbf{S}} * \underline{\mathbf{V}}^T$

$\underline{\mathbf{\Omega}}_1 = \text{randn}(I_2, K, I_3), \quad \underline{\mathbf{\Omega}}_2 = \text{randn}(I_1, L, I_3)$;
$\underline{\mathbf{Y}}_c = \underline{\mathbf{X}} * \underline{\mathbf{\Omega}}_1, \quad \underline{\mathbf{Y}}_r = \underline{\mathbf{X}}^T * \underline{\mathbf{\Omega}}_2$;
$[\underline{\mathbf{Q}}_c, \underline{\mathbf{R}}_c] = \text{T-QR}(\underline{\mathbf{Y}}_c)$;
$[\underline{\mathbf{Q}}_r, \underline{\mathbf{R}}_r] = \text{T-QR}(\underline{\mathbf{Y}}_r)$;
$[\widetilde{\underline{\mathbf{Q}}}_c, \sim, \sim] = \text{Truncated T-SVD}(\underline{\mathbf{R}}_c, R + H)$;
$[\widetilde{\underline{\mathbf{Q}}}_r, \sim, \sim] = \text{Truncated T-SVD}(\underline{\mathbf{R}}_r, R + H)$;
$\underline{\mathbf{Q}}_c = \underline{\mathbf{Q}}_c * \widetilde{\underline{\mathbf{Q}}}_c$;
$\underline{\mathbf{Q}}_r = \underline{\mathbf{Q}}_r * \widetilde{\underline{\mathbf{Q}}}_r$;
$\widehat{\underline{\mathbf{Z}}} = (\underline{\mathbf{\Omega}}_2^T * \underline{\mathbf{Q}}_c)^\dagger * (\underline{\mathbf{Y}}_r^T * \underline{\mathbf{Q}}_r)$;
$[\widetilde{\underline{\mathbf{U}}}, \underline{\mathbf{S}}, \widetilde{\underline{\mathbf{V}}}] = \text{Truncated T-SVD}(\widehat{\underline{\mathbf{Z}}}, R)$;
$\underline{\mathbf{U}} = \underline{\mathbf{Q}}_c * \widetilde{\underline{\mathbf{U}}}$;
$\underline{\mathbf{V}} = \underline{\mathbf{Q}}_r * \widetilde{\underline{\mathbf{V}}}$;

Основная идея заключается в использовании усечённого T-SVD (Truncated T-SVD) для построения ортонормированного базиса всего диапазона эскизов $\underline{Y_c}$ или $\underline{Y_r}$ вместо применения разложения T-QR. Вместо того чтобы строить ортонормированный базис для $\underline{Y}$, вычисляется $\left[\underline{Q}, \sim, \sim\right] = Truncated\ T - SVD\left(\underline{Y}, H\right)$ где $0 \leq H \leq K$. Это означает, что $\underline{Q} \in \mathbb{R}^{I_1 \times H \times I_3}$ содержит ведущие левые сингулярные тензоры тензора $\underline{Y}$. Если $H \leq K$, то коэффициентный тензор $\underline{\Omega_2} * \underline{Q} \in \mathbb{R}^{L \times H \times I_3}$ обеспечивает лучше обусловленную задачу, даже в случае, когда L=K. Следует отметить, что и другие однопроходные методы, например двусторонний рандомизированный SVD (TSR-SVD) [158], также могут быть обобщены на тензорный случай. В TSR-SVD используется сжатый тензор $\underline{W} = \underline{\Omega_2} * \underline{W} * \underline{\Omega_1}$, где $\underline{\Omega_1}$ и $\underline{\Omega_2}$ это стандартные гауссовы тензоры соответствующих размеров. Они применяются для вычисления низкоранговой трубной аппроксимации тензора. Более точно, для тензора данных $\underline{X}$ и ортонормированных базисов $\underline{Q_1}$ и $\underline{Q_2}$ выполняется:

$$\underline{X} \approx \underline{Q_1} * \left(\underline{Q_1}^T * \underline{X} * \underline{Q_2}\right) * \underline{Q_2}^T \tag{11}$$

Нетрудно заметить, что средний тензор в формуле (11) может быть аппроксимирован следующим образом:

$$\underline{Q}_1^T * \underline{X} * \underline{Q}_2 \approx \left(\underline{\Omega}_2 * \underline{Q}\right)^{\dagger} * \underline{W} * \left(\underline{Q}^T * \underline{\Omega}_1\right)^{\dagger} \quad (12)$$

$$\underline{Q}_1^T * \underline{X} * \underline{Q}_2 \approx \left(\underline{Q}_1^T * \underline{Y}\right) * \left(\underline{Q}_2 * \underline{\Omega}_1\right)^{\dagger} \quad (13)$$

Обе эти формулы требуют однократного прохода по данным исходного тензора X.

Эти алгоритмы также неустойчивы при L=K. В наших численных экспериментах оба варианта показали схожие результаты, и из-за указанной проблемы мы приводим только формулировку (13). Сначала мы обобщили алгоритм TSR-SVD на случай тензоров и обозначили его как TSRT-SVD.Кроме того, мы модифицировали его, чтобы он оставался устойчивым при L=K Описание этого подхода приведено в алгоритме 8.

Algorithm 8: Proposed randomized single-pass (two-sided version) algorithm III

Input : The data tensor $\underline{\mathbf{X}} \in \mathbb{R}^{I_1 \times I_2 \times I_3}$, three parameters L, K, H, $L \geq K \geq H \geq 0$ and a target tubal rank R

Output: A low tubal rank approximation $\underline{\mathbf{X}} \approx \underline{\mathbf{U}} * \underline{\mathbf{S}} * \underline{\mathbf{V}}^T$

1 $\underline{\mathbf{\Omega}}_1 = \text{randn}(I_2, K, I_3), \quad \underline{\mathbf{\Omega}}_2 = \text{randn}(I_1, L, I_3);$

2 $\underline{\mathbf{Y}}_c = \underline{\mathbf{X}} * \underline{\mathbf{\Omega}}_1, \quad \underline{\mathbf{Y}}_r = \underline{\mathbf{X}}^T * \underline{\mathbf{\Omega}}_2;$

3 $[\underline{\mathbf{Q}}_c, \underline{\mathbf{R}}_c] = \text{T-QR}(\underline{\mathbf{Y}}_c);$

4 $[\underline{\mathbf{Q}}_r, \underline{\mathbf{R}}_r] = \text{T-QR}(\underline{\mathbf{Y}}_r);$

5 $[\widetilde{\underline{\mathbf{Q}}}_c, \sim, \sim] = \text{Truncated T-SVD}(\underline{\mathbf{R}}_c, R + H);$

6 $[\widetilde{\underline{\mathbf{Q}}}_r, \sim, \sim] = \text{Truncated T-SVD}(\underline{\mathbf{R}}_r, R + H);$

7 $\underline{\mathbf{Q}}_c = \underline{\mathbf{Q}}_c * \widetilde{\underline{\mathbf{Q}}}_c;$

8 $\underline{\mathbf{Q}}_r = \underline{\mathbf{Q}}_r * \widetilde{\underline{\mathbf{Q}}}_r;$

9 $\underline{\mathbf{B}} = \underline{\mathbf{Q}}_c^T * \underline{\mathbf{Y}}_c * (\underline{\mathbf{Q}}_r^T * \underline{\mathbf{\Omega}}_1)^{\dagger};$

10 $[\widetilde{\underline{\mathbf{U}}}, \underline{\mathbf{S}}, \underline{\mathbf{V}}] = \text{Truncated T-SVD}(\underline{\mathbf{B}}, R);$

11 $\underline{\mathbf{U}} = \underline{\mathbf{Q}}_c * \widetilde{\underline{\mathbf{U}}};$

12 $\underline{\mathbf{V}} = \underline{\mathbf{Q}}_r * \widetilde{\underline{\mathbf{V}}};$

Из него очевидно, что алгоритмы 6, 7 и 8 выполняют однократный проход по исходному тензору данных только на начальном этапе (строка 2).

Эта операция может быть выполнена параллельно. В работе [35] предлагается ортонормировать случайные тензоры на первом шаге однопроходных алгоритмов, если используется большой параметр передискретизации (oversampling parameter).

Данная идея естественным образом переносится на тензорный случай,

однако в наших экспериментах мы её не применяли.

2.2.3. Рандомизированные алгоритмы с фиксированной точностью

Рандомизированные алгоритмы с фиксированной точностью для тензоров, это методы, которые могут автоматически оценивать подходящий или оптимальный тензорный ранг и строить соответствующую низкоранговую аппроксимацию тензора. Эти алгоритмы имеют решающее значение в тех случаях, когда ранг данных неизвестен или его оценка затруднена. В таких ситуациях необходимо оценить как оптимальный ранг, так и низкоранговую тубную аппроксимацию с заданной погрешностью для предписанного уровня точности. Недавно мы разработали рандомизированный алгоритм с фиксированной точностью [159], и результаты моделирования как на синтетических, так и на реальных данных показали значительное ускорение вычислений при построении приближённого усечённого T-SVD.

Наша работа представляет собой обобщение матричного случая [158]на тензорный случай. Этот алгоритм описан в алгоритме 9.

Algorithm 9: Randomized fixed-precision algorithm [159]

Input : A tensor $\underline{\mathbf{X}} \in \mathbb{R}^{I_1 \times I_2 \times I_3}$; an error bound ϵ; a block size b and a power iteration q.

Output: $\underline{\mathbf{Q}} = [\underline{\mathbf{Q}}^{(1)}, \underline{\mathbf{Q}}^{(2)}, \ldots, \underline{\mathbf{Q}}^{(i)}]$, $\underline{\mathbf{B}} = \begin{bmatrix} \underline{\mathbf{B}}^{(1)} \\ \underline{\mathbf{B}}^{(2)} \\ \vdots \\ \underline{\mathbf{B}}^{(i)} \end{bmatrix}$ such that $\left\|\underline{\mathbf{X}} - \underline{\mathbf{Q}} * \underline{\mathbf{B}}\right\|_F < \varepsilon$ and corresponding optimal tubal rank R

1 $\underline{\mathbf{Q}} = [],\ \underline{\mathbf{B}} = [];$
2 $E = \|\underline{\mathbf{X}}\|_F^2;$
3 **for** $i = 1, 2, \ldots$ **do**
4 $\quad \underline{\mathbf{\Omega}}^{(i)} = \text{randn}\,(I_2, b, I_3);$
5 $\quad \underline{\mathbf{Q}}^{(i)} = \text{orth}\left(\underline{\mathbf{X}} * \underline{\mathbf{\Omega}}^{(i)} - \underline{\mathbf{Q}} * \left(\underline{\mathbf{B}} * \underline{\mathbf{\Omega}}^{(i)}\right)\right);$
6 $\quad$ **for** $j = 1, 2, \ldots, q$ **do**
7 $\quad\quad \underline{\mathbf{Q}}^{(i)} = \text{orth}\left(\underline{\mathbf{X}}^T * \underline{\mathbf{Q}}^{(i)} - \underline{\mathbf{B}}^T * \underline{\mathbf{Q}}^T * \underline{\mathbf{Q}}^{(i)}\right);$
8 $\quad\quad \underline{\mathbf{Q}}^{(i)} = \text{orth}\left(\underline{\mathbf{X}} * \underline{\mathbf{Q}}^{(i)} - \underline{\mathbf{Q}} * \underline{\mathbf{B}} * \underline{\mathbf{Q}}^{(i)}\right);$
9 $\quad$ **end**
10 $\quad \underline{\mathbf{Q}}^{(i)} = \text{orth}\left(\underline{\mathbf{Q}}^{(i)} - \underline{\mathbf{Q}} * \left(\underline{\mathbf{Q}}^T * \underline{\mathbf{Q}}^{(i)}\right)\right);$
11 $\quad \underline{\mathbf{B}}^{(i)} = \underline{\mathbf{Q}}^{(i)T} * \underline{\mathbf{X}};$
12 $\quad \underline{\mathbf{Q}} = \underline{\mathbf{Q}} \boxplus_2 \underline{\mathbf{Q}}^{(i)};$
13 $\quad \underline{\mathbf{B}} = \underline{\mathbf{B}} \boxplus_1 \underline{\mathbf{B}}^{(i)};$
14 $\quad E = E - \left\|\underline{\mathbf{B}}^{(i)}\right\|_F^2;$
15 $\quad$ **if** $E < \varepsilon^2$ **then**
16 $\quad\quad$ Break
17 $\quad$ **end**
18 **end**

Алгоритм требует задания параметров:

- передискретизации (oversampling);
- числа степеней (power iteration);
- допустимой погрешности (tolerance) и
- размера блока (block size).

Он постепенно оценивает тубный ранг и вычисляет соответствующую низкоранговую тубную аппроксимацию.

Строки 6–9 в алгоритме отвечают за итерации степеней (power iteration stage), обеспечивающие более высокую точность, особенно в случаях, когда сингулярные значения фронтальных срезов тензора данных медленно убывают.

Однако недавно матричная версия этого алгоритма была дальше усовершенствована в работах [159] и [160]. Мотивированные интересными результатами, представленными там, мы усовершенствовали наш ранее предложенный алгоритм [159]в нескольких направлениях.

Во-первых, для заданного числа итераций степеней алгоритм [159]требует $2q + 2$ проходов по исходному тензору данных на каждой итерации алгоритма.

Это означает, что при заданном числе итераций q, необходимо повторно проходить по данным, что увеличивает общее количество обращений к ним.

Очевидно, что это является ограничением, поскольку, как показано в [149], например, для изображений или видео, всего двух проходов обычно достаточно для удовлетворительной точности.

Однако алгоритм 9 требует по крайней мере трёх проходов для $q = 1$ и четырёх $q = 2$. Для очень больших тензоров данных или в случае, когда итерационный алгоритм медленно сходится, дополнительные проходы могут быть вычислительно затратными. Таким образом, наша цель, повысить эффективность алгоритма 9. В новой модифицированной версии, представленной в алгоритме 10, эта проблема решена. Здесь верхняя граница ошибки используется для оценки тубного ранга и построения низкоранговой тубной аппроксимации при любом числе проходов по данным. Следует отметить, что в алгоритме 9 параметр q, это число степеней (power iteration parameter), а q также используется для обозначения числа проходов в алгоритме 10.

Algorithm 10: Proposed fixed-precision algorithm I

Input : The data tensor $\underline{\mathbf{X}} \in \mathbb{R}^{I_1 \times I_2 \times I_3}$ $(I_1 \leq I_2)$, an approximation error bound ϵ, the pass numbers $q > 2$ and block size b
Output: A low tubal rank approximation of the tensor $\underline{\mathbf{X}} \approx \underline{\mathbf{Q}} * \underline{\mathbf{R}}$;

1 $\underline{\mathbf{Q}} = []; \quad \underline{\mathbf{B}}[];$
2 **for** $l = 1, 2, \ldots$ **do**
3 **if** *if q is an even number* **then**
4 $\underline{\mathbf{\Omega}} = \text{randn}\,(I_2, b, I_3)$;
5 $\underline{\mathbf{Y}} = \underline{\mathbf{X}} * \underline{\mathbf{\Omega}} - \underline{\mathbf{Q}} * (\underline{\mathbf{B}} * \underline{\mathbf{\Omega}})$;
6 $[\underline{\mathbf{Q}}_l, \sim] = \text{T-LU}(\underline{\mathbf{Y}})$;
7 **else**
8 Set $\underline{\mathbf{Q}}_l$ as a random tensor of size $I_1 \times b \times I_3$;
9 **end**
10 **for** $t = 1, 2, \ldots, \lfloor \frac{q-2}{2} \rfloor$ **do**
11 **if** $t == \lfloor \frac{q-2}{2} \rfloor$ **then**
12 $\underline{\mathbf{R}} = \underline{\mathbf{X}}^T * \underline{\mathbf{Q}}_l$;
13 $\underline{\mathbf{Q}}_l = \text{orth}(\underline{\mathbf{X}} * \underline{\mathbf{R}} - \underline{\mathbf{Q}} * (\underline{\mathbf{B}} * \underline{\mathbf{R}}))$;
14 **else**
15 $[\underline{\mathbf{Q}}_l, \sim] = \text{T-LU}(\underline{\mathbf{X}} * (\underline{\mathbf{X}}^T * \underline{\mathbf{Q}}_l))$;
16 **end**
17 **end**
18 $\underline{\mathbf{Q}}_l = \text{orth}(\underline{\mathbf{Q}}_l - \underline{\mathbf{Q}}_l * (\underline{\mathbf{Q}}^T * \underline{\mathbf{Q}}_l))$;
19 $\underline{\mathbf{B}}_i = \underline{\mathbf{Q}}_l^T * \underline{\mathbf{X}}$;
20 $\underline{\mathbf{Q}} = \underline{\mathbf{Q}} \boxplus_2 \underline{\mathbf{Q}}_l$;
21 $\underline{\mathbf{B}} = \underline{\mathbf{B}} \boxplus_1 \underline{\mathbf{B}}_l$;
22 **if** a termination criterion is satisfied **then**
23 Set the tubal rank as k and break;
24 **end**
25 **end**

Подобно подходу, описанному в [160], в нашей работе применяются модификации, позволяющие сделать алгоритм 9 более эффективным по числу проходов (pass-efficient).

– Использование разложения T-EIG для вычисления усечённого T-SVD тензора $\underline{B}$.
– Пропуск одного шага орто-нормализации на каждой итерации цикла степеней (power iteration loop).
– Замена разложения T-QR на T-LU для орто-нормализации в процессе итераций по степеням (за исключением последней итерации).
– Модификация алгоритма для возможности выполнения нечётного числа проходов.

Следует отметить, что в алгоритме 10 учитываются хвостовые части тензорных факторов $\underline{\widehat{U}}$, $\underline{\widehat{V}}$, $\widehat{S}$ (в строках 26–27), поскольку собственное разложение симметричных матриц выполняется в порядке возрастания собственных значений.

Кроме того, алгоритм 9 может быть дополнительно улучшен, если заменить T-QR-разложение (процесс орто-нормализации) на T-произведение (T-product) и использование тензорной обратной матрицы малых тензоров.

Это оправдано тем, что QR-разложение является ключевой операцией в процессе T-QR-разложения, но отличается сравнительно низкой эффективностью при многопоточном вычислении. Следуя идее, предложенной в работе [161], мы заменяем эту операцию на T-произведение тензоров и обращение малых тензоров, что обеспечивает более высокую степень параллелизма. Описание этого усовершенствованного подхода представлено в алгоритме 11, который будет рассмотрен далее.

Теорема 1.

Пусть $\underline{X} \in \mathbb{R}^{I_1 \times I_2 \times I_3}$ и $\underline{\Omega} \in \mathbb{R}^{I_1 \times K \times I_3}$ случайный тензор. Определим $\underline{Y} = \underline{X} * \underline{\Omega}, \underline{W} = \underline{X}^T * \underline{Y}$ а экономичное T-SVD-разложение тензора $\underline{Y}$ задаётся как $\underline{Y} = \underline{U} * \underline{S} * \underline{V}^T$. Тогда

$$\underline{Q} = \underline{Y} * \hat{\underline{V}} * \hat{\underline{S}}^{-1}, \underline{B} = \left(\underline{W} * \hat{\underline{V}} * \hat{\underline{S}}^{-1}\right)^T \tag{14}$$

получаем аппроксимацию $\underline{X} \approx \underline{Q} * \underline{B}$ которая имеет такую же точность, как и базовые рандомизированные алгоритмы. Кроме того, выполняется равенство

$$\left\|\underline{B}\right\|_F^2 = trace\left(\underline{H}_1\right)$$

где $\underline{H}_1 = \underline{H}(:,:,1)$ это первый фронтальный срез тензора $\underline{H} = \underline{W} * \underline{W}^T * \left(\underline{Y} * \underline{Y}\right)^{-1}$.

Рассмотрим тождества

$$\underline{Q} = orth\left(\underline{X} * \underline{\Omega}\right) = orth\left(\underline{Y}\right) = \hat{\underline{U}} = \underline{Y} * \hat{\underline{V}} * \hat{\underline{S}}^{-1} \tag{15}$$

Подставляя (15) в $\underline{B} = \underline{Q}^T$ получаем:

$$\underline{B} = \underline{Q}^T * \underline{X} = \left(\underline{W} * \hat{\underline{V}} * \hat{\underline{S}}^{-1}\right)^T \tag{16}$$

Поскольку $\underline{Q}$ представляет собой ортонормализацию тензора $\underline{X} * \underline{\Omega}$ аппроксимация $\underline{Q} * \underline{B}$ обеспечивает такую же точность, как и базовые рандомизированные алгоритмы, но без итераций по степеням и без передискретизации [162]. Теперь, используя тот факт, что $\left\|\underline{B}\right\|_F^2 = trace\left(\underline{G}_1\right) =$

$\underline{G}(:,:,1)$ где $\underline{G}_1$ первый фронтальный срез тензора $\underline{G} = \underline{B}^T * \underline{B}$ и получаем:

$$\left\|\underline{B}\right\|_F^2 = trace\left(\left(\hat{\underline{S}}^{-1} * \hat{\underline{V}}^T * \underline{W}^T * \underline{W} * \hat{\underline{V}} * \hat{\underline{S}}^{-1}\right)_1\right)$$
$$= trace\left(\left(\underline{W}^T * \underline{W} * \hat{\underline{S}}^{-2} * \hat{\underline{V}}\right)_1\right)$$
$$= trace\left(\left(\underline{W}^T * \underline{W} * \left(\underline{Y}^T * \underline{Y}\right)^{-1}\right)_1\right) \quad (17)$$

Исходя из теоремы 1, определим $\underline{T} = \underline{W}^T * \underline{W}$ и $\underline{Z} = \underline{Y}^T * \underline{Y}$ Тогда критерий остановки $\left\|\underline{X} - \underline{Q} * B\right\|_F^2 = \left\|\underline{Q}\right\|_F^2 - \left\|\underline{B}\right\|_F^2$ и можно переписать в следующем виде

$$\left\|\underline{X} - \underline{Q} * B\right\|_F^2 = \left\|\underline{Q}\right\|_F^2 - trace\left(\left(\underline{T} * \underline{Z}^{-1}\right)_1\right)$$

Эта формула используется в алгоритме 11, строка 13.

Algorithm 11: The proposed fixed-precision algorithm II

Input : The data tensor $\underline{\mathbf{X}} \in \mathbb{R}^{I_1 \times I_2 \times I_3}$, a block size b, a power iteration q and an approximation error bound.
Output: The QB approximation of the tensor $||\underline{\mathbf{X}} - \underline{\mathbf{Q}} * \underline{\mathbf{B}}||_F \leq \epsilon$

1 $\underline{\mathbf{Y}} = [], \quad , \underline{\mathbf{W}} = [];$
2 $E = ||\underline{\mathbf{X}}||_F, \ tol = \epsilon^2;$
3 **for** $i = 1, 2, \ldots$ **do**
4 Generate a random tensor $\underline{\mathbf{\Omega}}_i$ of size $I_2 \times b \times I_3$;
5 **for** $j = 1, 2, \ldots, q$ **do**
6 $\underline{\mathbf{W}}_i = \underline{\mathbf{X}}^T * \underline{\mathbf{X}} * \underline{\mathbf{\Omega}}_i - \underline{\mathbf{W}} * \underline{\mathbf{Z}}^{-1} * \underline{\mathbf{W}}^T * \underline{\mathbf{\Omega}}_i;$
7 $\underline{\mathbf{\Omega}}_i = \text{orth}(\underline{\mathbf{W}}_i);$
8 **end**
9 $\underline{\mathbf{Y}}_i = \underline{\mathbf{X}} * \underline{\mathbf{\Omega}}_i, \quad \underline{\mathbf{W}}_i = \underline{\mathbf{X}}^T * \underline{\mathbf{Y}}_i;$
10 $\underline{\mathbf{Y}} = \underline{\mathbf{Y}} \boxplus_2 \underline{\mathbf{Y}}_i, \quad \underline{\mathbf{W}} = \underline{\mathbf{W}} \boxplus_2 \underline{\mathbf{W}}_i;$
11 $\underline{\mathbf{Z}} = \underline{\mathbf{Y}}^T * \underline{\mathbf{Y}}, \quad \underline{\mathbf{T}} = \underline{\mathbf{W}}^T * \underline{\mathbf{W}};$
12 **if** $E - (\text{trace}((\underline{\mathbf{T}} * \underline{\mathbf{Z}}^{-1})_1)) < tol$ **then**
13 break
14 **end**
15 **end**
16 $[\widehat{\underline{\mathbf{V}}}, \widehat{\underline{\mathbf{D}}}] = \text{T-EIG}(\underline{\mathbf{Z}});$
17 $\underline{\mathbf{Q}} = \underline{\mathbf{Y}} * \widehat{\underline{\mathbf{V}}} * \text{sqrt}(\widehat{\underline{\mathbf{D}}})^{-1}, \quad \underline{\mathbf{B}} = (\underline{\mathbf{W}} * \widehat{\underline{\mathbf{V}}} * \text{sqrt}(\widehat{\underline{\mathbf{D}}})^{-1})^T;$

Из теоремы 1 также следует, что

$$\underline{Q} * \underline{B} = \underline{Y} * \hat{\underline{V}} * \hat{\underline{S}}^{-2} * \hat{\underline{V}}^T * \underline{W}^T = \underline{Y} * \left(\underline{Y}^T * \underline{Y}\right)^{-1} * \underline{W}^T \quad (18)$$

а следовательно, $\underline{Q} * \underline{B} = \underline{Y} * \underline{Z}^{-1} * \underline{W}^T$. Это показывает, что итерацию по степеням (power iteration), выполняемую в алгоритме 9,можно заменить более простой операцией $\underline{X} - \underline{Y} * \underline{Z}^{-1} * \underline{W}^T$. Иными словами, строки 6–9 в алгоритме 9 могут быть

заменены следующими вычислениями:

1 **for** $j = 1 : p$ **do**
2 $\quad \underline{\mathbf{Y}}_i \leftarrow \underline{\mathbf{X}} * \underline{\mathbf{\Omega}}_i - \underline{\mathbf{Y}} * \underline{\mathbf{Z}}^{-1} * \underline{\mathbf{W}}^T * \underline{\mathbf{\Omega}}_i$;
3 $\quad \underline{\mathbf{W}}_i \leftarrow \underline{\mathbf{X}}^T * \underline{\mathbf{Y}}_i - \underline{\mathbf{W}} * \underline{\mathbf{Z}}^{-1} * \underline{\mathbf{Y}}^T * \underline{\mathbf{Y}}_i$;
4 $\quad \underline{\mathbf{\Omega}}_i \leftarrow \mathrm{orth}(\underline{\mathbf{W}}_i)$
5 **end**

Подставив строку 2 в строку 3 в указанном выше цикле, несложно получить следующее выражение:

$$\underline{W}_i = \underline{X}^T * \underline{X} * \underline{\Omega}_i - \underline{W} * \underline{Z}^{-1} * \underline{W}^T * \underline{\Omega}_i$$

Таким образом, используя эти преобразования, алгоритм 9 можно модифицировать и получить алгоритм 11. Следуя подходу, предложенному в работе [39], один шаг орто-нормализации был опущен, чтобы уменьшить время вычислений, при этом точность снижается незначительно.

ВЫВОДЫ ПО ГЛАВЕ 2

Вторая глава была посвящена разработке и исследованию методов обработки мультимодальных данных, включающих две взаимосвязанные задачи эффективное слияние разнородных источников информации и устранение шума с последующим восстановлением недостающих данных. Представленные подходы основаны на идее тензорного представления данных, однако реализованы без прямого применения тензорного разложения в его классическом виде. Такой подход позволил преодолеть ряд ограничений, присущих существующим методам тензорного слияния и восстановления, обеспечив более высокую устойчивость и гибкость модели при работе с данными различной природы.

В первой части главы был рассмотрен разработанный метод слияния мультимодальных данных, направленный на объединение информации из различных модальностей (например, изображений, звуковых и текстовых данных) в едином представлении. Основное внимание уделялось анализу четырех различных схем слияния, построенных на основе тензорной модели. Проведённое сравнение показало, что предложенный подход демонстрирует устойчивые результаты в условиях высокой гетерогенности и коррелированности признаков между модальностями. Ключевым преимуществом разработанного метода является использование идеи тензорного разложения без необходимости непосредственного вычисления полного тензорного декомпозиционного базиса. Это позволило значительно сократить вычислительные затраты, избежать проблем с численной неустойчивостью и переобучением, характерных для классических методов CPD или Tucker-разложения. Кроме того, предложенная схема обеспечивает лучшее сохранение корреляционной структуры между модальностями, что особенно важно для задач, где взаимосвязь между каналами данных имеет решающее значение (например, в задачах мультимодального распознавания эмоций или анализа визуально-аудиальных сигналов).

В отличие от существующих тензорных методов, разработанный подход способен динамически адаптироваться к различным наборам модальностей и их

размерностям, что делает его универсальным инструментом для обработки разнородных данных. Дополнительным преимуществом является повышенная интерпретируемость получаемого представления: за счет сохранения взаимных зависимостей между компонентами тензора метод позволяет извлекать осмысленные латентные признаки, отражающие общую структуру мультимодального пространства.

Таким образом, в рамках первой части главы был предложен новый метод слияния мультимодальных данных, который сочетает точность, устойчивость и вычислительную эффективность, превосходя традиционные тензорные модели по ряду параметров.

Вторая часть главы посвящена задаче удаления шума и восстановления искажённых или неполных мультимодальных данных, которая играет ключевую роль в обеспечении надежности и качества дальнейшего анализа. На основе обобщённой тензорной модели был предложен новый алгоритм, направленный на восстановление недостающих элементов и фильтрацию шумовых компонент без потери структурных зависимостей между модальностями. Главное преимущество разработанного алгоритма заключается в том, что он учитывает межмодальные зависимости при восстановлении информации, что позволяет получать более точные результаты по сравнению с методами, обрабатывающими каждую модальность независимо. Благодаря использованию тензорного представления достигается возможность совместной оптимизации параметров восстановления и подавления шума, что обеспечивает баланс между сохранением информативных признаков и устранением избыточных компонентов.Разработанный метод демонстрирует высокую устойчивость к различным типам шумов как случайным, так и систематическим, возникающим вследствие несовпадения каналов или ошибок сенсоров. В отличие от существующих подходов, основанных на раздельной фильтрации по модальностям, предложенное решение эффективно сохраняет межмодальные корреляции, что особенно важно для последующих этапов анализа и классификации. Кроме того, предложенный алгоритм обладает адаптивностью к структуре данных: он автоматически подстраивается под степень

разреженности и степень зашумленности входных данных, что делает его применимым в широком спектре практических задач от мультимодального биомедицинского анализа до интеллектуальных систем распознавания.

Проведенные в данной главе исследования продемонстрировали эффективность предложенных решений как в области слияния мультимодальных данных, так и в области восстановления информации и подавления шума. Предложенные методы обладают рядом существенных преимуществ: снижением вычислительной сложности, сохранением межмодальных зависимостей, устойчивостью к шумам и неполноте данных, а также способностью адаптироваться к структуре и размерности исходных данных.

Результаты, представленные во второй главе, формируют теоретическую и методологическую основу для последующих экспериментальных исследований. В третьей главе диссертации будут приведены практические доказательства эффективности предложенных методов на реальных и модельных наборах мультимодальных данных, а также сравнительный анализ с существующими решениями.

ГЛАВА 3. АНАЛИЗ РЕЗУЛЬТАТОВ МАТЕМАТИЧЕСКОГО МОДЕЛИРОВАНИЯ

В этой главе мы сравним наш подход мультирангового слияния с другими подходами, такими как конкатенация, слияния на основе тензора, слияния низкого ранга. Мы оцениваем нашу модель по трем различным задачам: мультимодальный анализ настроений, анализ характеристик говорящего и распознавание эмоций. Вычислим производительность наших трех модальностей (языковой, визуальной и акустической) с помощью определения ранга для уменьшения размера с минимальной потерей информации.

В этой работе мы предлагаем три серии экспериментов, на основе существующего корпуса, который мы подробно рассмотрим далее совместно со способами извлечения признаков, каждый из которых решает разные исследовательские задачи.

3.1. Метод слияния мультимодальных данных на основе тензорного представления

Эксперименты выполнялись на наборе данных POM, который состоит из 903 видеороликов с обзорами фильмов. Каждое видео сопровождается аннотациями со следующими характеристиками говорящего: уверенный, страстный, приятный голос, доминирующий, заслуживающий доверия, яркий, опытный, развлекательный, сдержанный, доверчивый, расслабленный, общительный, тщательный, нервный, убедительный и юмористический [163].

Корпус мультимедийных материалов (POM), состоящий из 1000 видеороликов с обзорами фильмов, полученных с социального мультимедийного веб-сайта ExpoTV.com.

Корпус подходит для изучения настроения в контексте онлайн-социальных мультимедиа. ExpoTV.com – популярный веб-сайт, на котором размещены видеоролики с обзорами продуктов. В каждом обзоре продукта есть видео, на

котором спикер рассказывает о конкретном продукте, а также существует оценка этого продукта спикером по шкале от 1 звезды (для наиболее негативных отзывов) до 5 звезд (для наиболее положительных отзывов). Таким образом, выступающий в видео с обзором пятизвездочного фильма, скорее всего, убеждает аудиторию в пользу фильма, в то время как выступающий в видео с обзором фильма с рейтингом 1 будет возражать против просмотра фильма. В корпус включены только видеообзоры фильмов для единообразия контекста. Собрано в общей сложности 1000 видеороликов с обзорами фильмов. Положительных отзывов: 500 видеообзоров фильмов с 5-звездочным рейтингом (315 мужчин и 185 женщин). Отрицательных отзывов: 500 видеороликов с обзорами фильмов с рейтингом 1 или 2 звезды, состоящие из 216 видео с 1 звездой (151 мужчина и 65 женщин) и 284 видео с 2 звездами (212 мужчин и 72 женщины).

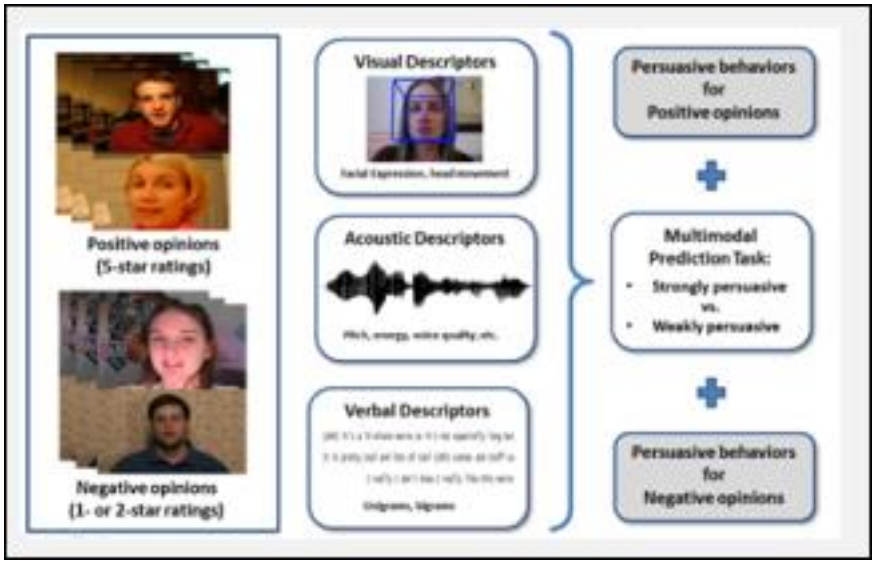


Рис. 60. Пример мультимодальных данных из корпуса POM

Каждое видео в корпусе имеет фронтальный вид одного человека, говорящего о конкретном фильме, а средняя длина видео составляет около 94 секунд со стандартным отклонением около 32 секунд. Корпус содержит 372 уникальных спикера и 600 уникальных названий фильмов, включая все типы общих жанров фильмов.

Разговорный текст отличается от письменного (обзоры, твиты) по композиции и грамматике. Например, «*Я думаю, что все было хорошо…, Хммм…, Дай мне подумать…, Да…, Нет…, Хорошо да…*». Эти формы редко встречаются в

письменной речи, но ее варианты очень распространены в разговорной речи.

Утверждение передает фактическое сообщение, а остальное - говорящий думает вслух, в конечном итоге, соглашается с утверждением. Ключевым фактором в работе с этой изменчивой природой разговорной речи является построение моделей, способных работать в присутствии ненадежных речевых черт, фокусируясь на важных частях речи.

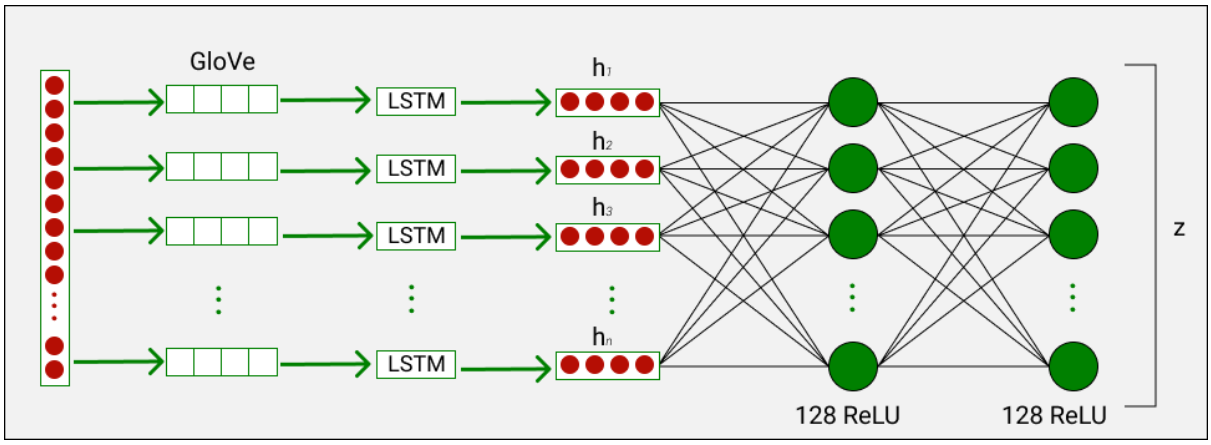


Рис. 61. Графическое представление извлечения признаков из текста

Предлагаемый подход к решению проблем разговорной речи заключается в изучении богатого представления произносимых слов и использовании его в качестве входных данных для подключенной глубокой сети. Это представление для i-го слова содержит информацию от начала высказывания во времени. Таким образом, по мере того, как модель обнаруживает значение высказывания во времени, если она встречает непригодную для использования информацию в слове i + 1 и произвольном количестве слов после него, представление до i не теряется. Кроме того, если модель снова встречает полезную информацию, она может восстановить ее, встраивая ее в долговременную краткосрочную память (LSTM). Кодировки, зависящие от времени, могут использоваться остальной частью конвейера, просто фокусируясь на соответствующих частях, используя нелинейное аффинное преобразование зависящих от времени встраиваний, которые могут действовать как механизм уменьшения внимания. Набор произнесенных слов представлен как последовательность 300-мерных векторов слов GloVe [164] , [165], [166].

Сеть LSTM [167] используется для изучения зависимых от времени языковых представлений $h_l = \{h_1, h_2, h_3, \dots, h_T; h_1 \in \mathbb{R}^{128}\}$ для слов согласно следующей формулировке LSTM.

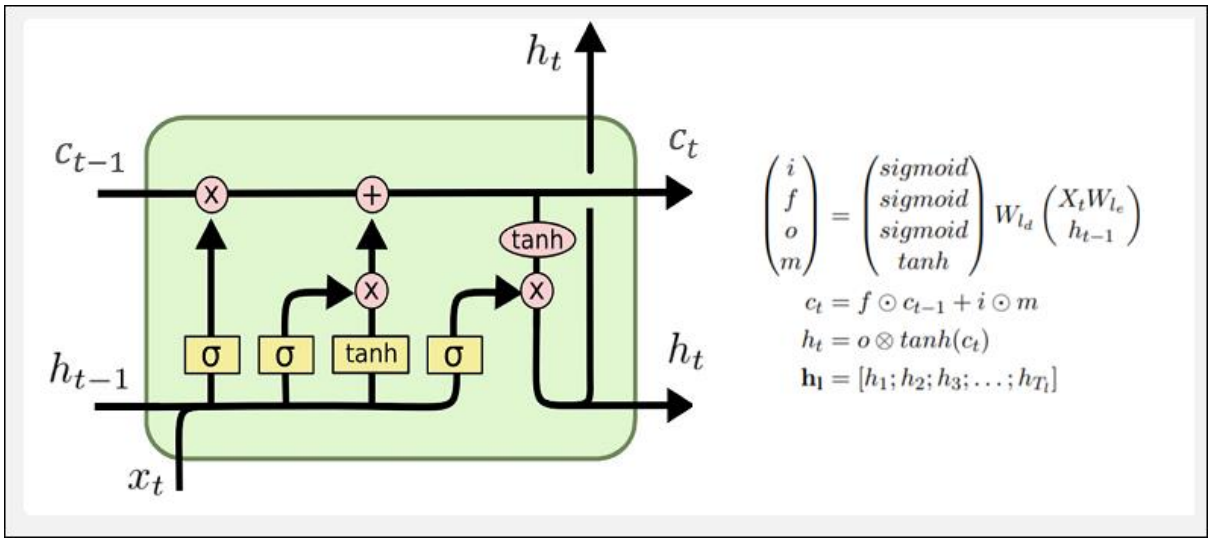


Рис. 62. Архитектура LSTM

Поскольку видео-мнения состоят в основном из выступающих, говорящих с аудиторией через камеру крупным планом, лицо является наиболее важным источником визуальной информации. Лицо говорящего определяется для каждого кадра (с частотой дискретизации 30 Гц) и индикаторов семи основных эмоций (гнев, презрение, отвращение, страх, радость, печаль и удивление) и двух сложных эмоций (разочарования и замешательства) [168] извлекались с использованием структуры анализа выражения лица FACET Набор из 20 единиц действия для лица [169], показывающих детальные движения мышц лица, также извлекались с помощью FACET. Оценки положения головы, поворота головы и 68 точек лицевых ориентиров также извлекались для каждого кадра с помощью OpenFace [170].

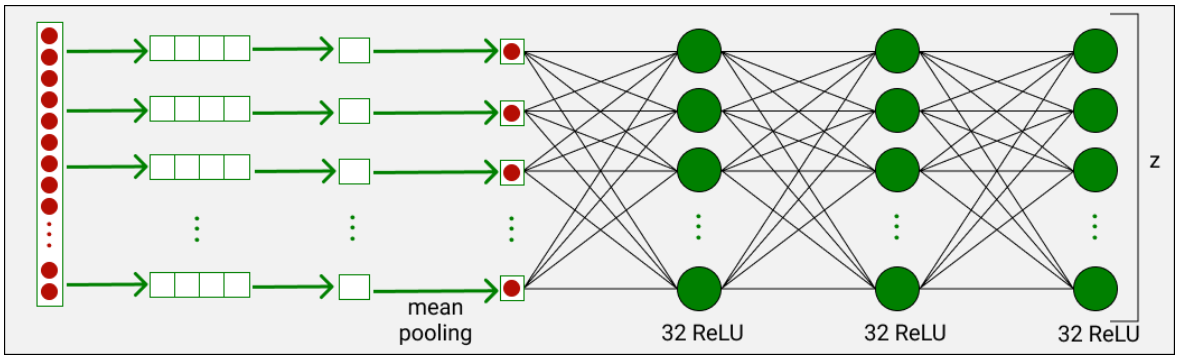


Рис. 63. Графическое представление извлечения признаков из изображения

Пусть визуальные особенности $\widehat{v}_j = \left[v_j^1, v_j^2, v_j^3, \dots, v_j^p\right]$ для кадра j видео высказывания содержат набор p визуальных характеристик, где T_v- общее количество видеокадров в высказывании. Мы выполняем объединение средних значений по кадрам, чтобы получить ожидаемые визуальные характеристики $v = \left[\mathbb{E}[v^1], \mathbb{E}[v^2], \mathbb{E}[v^3], \dots, \mathbb{E}[v^l]\right]$. v затем используется в качестве входных данных для подсети визуального встраивания $\mathcal{U}_v$. Поскольку информация, извлекаемая с помощью FACET из видео высоко информативна, использование глубокой нейронной сети будет достаточно для получения значимых результатов визуальной модальности. Мы используем глубокую нейронную сеть с тремя скрытыми слоями по 32 единицы ReLU и весами $\mathcal{W}_v$. Эмпирически понятно, что углубление модели или увеличение количества нейронов в каждом слое не приводит к улучшению визуальных характеристик. Вывод подсети обеспечивает визуальное вложение z^v:

$$z^v = \mathcal{U}_v(v, \mathcal{W}_v) \in \mathbb{R}^{32}$$

Для каждого звукового высказывания набор акустических характеристик извлекается с использованием структуры акустического анализа COVAREP [171], включая 12 MFCC, отслеживание высоты тона и функции голосового / неголосового сегментирования (с использованием устойчивого к шуму суммирования остаточных гармоник (SRH) [172], параметры источника голосовой щели (оцененные с помощью обратной фильтрации голосовой щели на основе синхронной IAIF GCI [173], [174], [175], параметры наклона пика [171], коэффициенты дисперсии максимумов (MDQ) [176] и оценки параметра формы Rd для кривой Liljencrants-Fant (LF) глоттальной модели [177] . Эти извлеченные характеристики отражают различные характеристики человеческого голоса и, как было показано, связаны с эмоциями [178] .

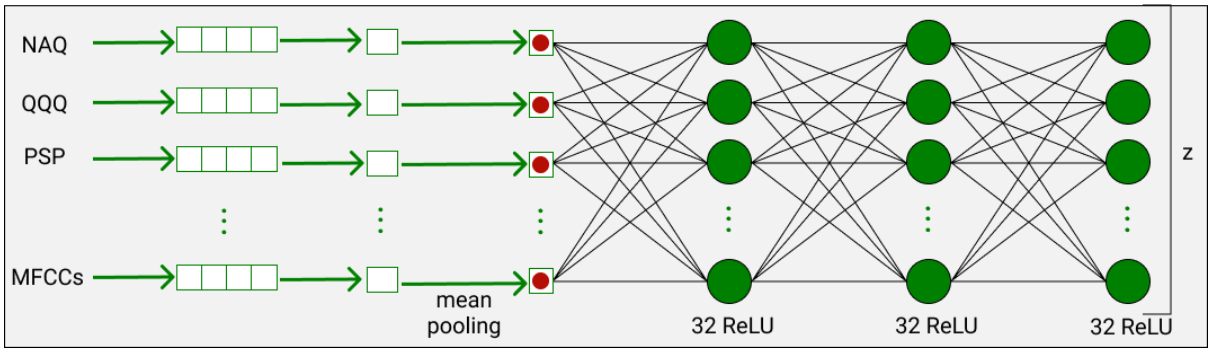


Рис. 64. Графическое представление извлечения признаков из акустических данных

Для каждого сегмента с Та аудиофрагментами (с частотой дискретизации 100 Гц; т. Е. 10 мс) мы извлекаем набор из q акустических характеристик $\hat{a}_j = \left[a_j^1, a_j^2, a_j^3, \ldots, a_j^p\right]$ для звукового кадра j в высказывании. Мы выполняем объединение средних значений для этих извлеченных акустических характеристик, чтобы получить ожидаемые акустические характеристики $a = \left[\mathbb{E}[a^1], \mathbb{E}[a^2], \mathbb{E}[a^3], \ldots, \mathbb{E}[q]\right]$. Здесь a - это вход в подсеть эмбедирования аудио $\mathcal{U}_a$. Поскольку COVAREP [171] также извлекает из звука богатые возможности, для моделирования акустической модальности достаточно использовать глубокую нейронную сеть. Подобно $\mathcal{U}_v$, $\mathcal{U}_a$ представляет собой сеть с 3 уровнями из 32 единиц ReLU с весами $\mathcal{W}_a$.

Здесь мы также эмпирически заметили, что углубление модели или увеличение количества нейронов в каждом слое не приводит к лучшей производительности. Каждая модальность нормировалась по стандартному отклонению, после чего все данные использовались для построения тензора внешнего произведения признаков трёх модальностей. Таким образом, результирующее мультимодальное пространство характеризуется размерностью

$$\mathcal{X} \in \mathbb{R}^{128\times64\times128}$$

Для обеспечения стабильности обучения все модели обучались в единой среде на GPU NVIDIA RTX 3090 с одинаковыми параметрами оптимизатора (Adam,$\eta - 0.001, bach\ size = 32$).

3.1.1. Экспериментальные результаты LMF

Модель LMF служит базовой точкой сравнения, поскольку она основана на CP-разложении (CANDECOMP/PARAFAC), аппроксимирующем полный тензор взаимодействий через диагональное ядро. Это разложение имеет фиксированный ранг R для всех модальностей и описывается выражением:

$$y = \sum_{r=1}^{R} u_r^{(1)} \otimes u_r^{(2)} \otimes u_r^{(3)}$$

Данный подход уменьшает число параметров, но накладывает серьёзные ограничения на межмодальные зависимости, так как ядро является супердиагональным. В результате взаимодействия между различными модальностями описываются только на уровне прямых корреляций, без учёта сложных нелинейных комбинаций признаков. Результаты эксперимента показывают, что при увеличении ранга R качество модели (MAE) улучшается до определённого предела, после чего начинает деградировать из-за переобучения.

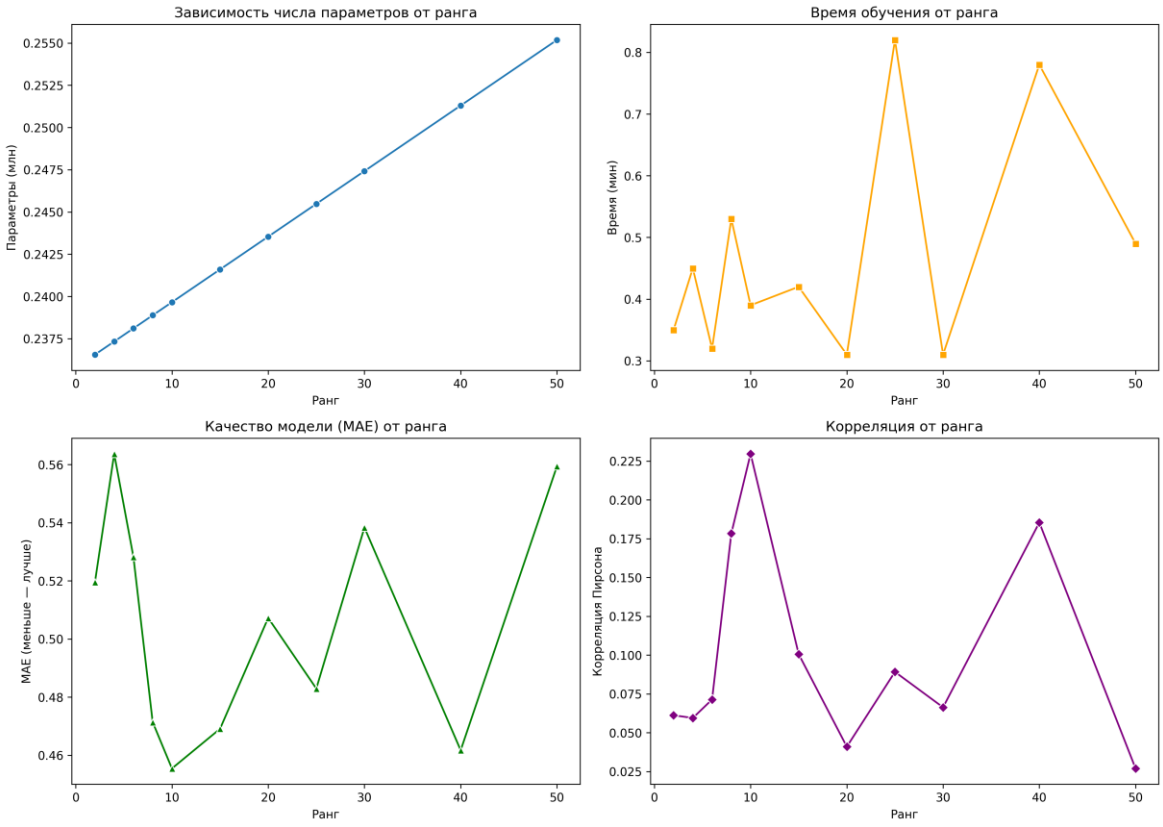


Рис. 65. Представлено изменение числа параметров, времени обучения и ошибки в зависимости от ранга

– При низких рангах (R ≤ 10) MAE остаётся выше 0.52, что указывает на недостаточную способность модели улавливать межмодальные корреляции.
– Оптимальное значение ранга достигается при R ≈ 20–30, где MAE ≈ 0.47, после чего наблюдается рост ошибки и нестабильность корреляции.
– Время обучения растёт почти линейно, что подтверждает ограниченную масштабируемость LMF.

Таким образом, хотя LMF обеспечивает базовый уровень мультимодальной интеграции, его производительность существенно ограничена диагональной структурой ядра, фиксированным рангом и невозможностью моделировать гетерогенность модальностей.

3.1.2. Экспериментальные результаты Tucker-разложения

Модель Tucker Fusion представляет тензор взаимодействий как произведение меньшего ядра
$\mathcal{g}$ и матриц-факторов $U^{(n)}$, что описывается формулой:

$$\mathcal{X} \approx \mathcal{g} \times_1 U^{(1)} \times_2 U^{(2)} \times_3 U^{(3)}$$

Каждая матрица-фактор отражает проекцию соответствующей модальности в латентное подпространство ранга R_n, а ядро $\mathcal{g}$ хранит все взаимодействия между модальностями.

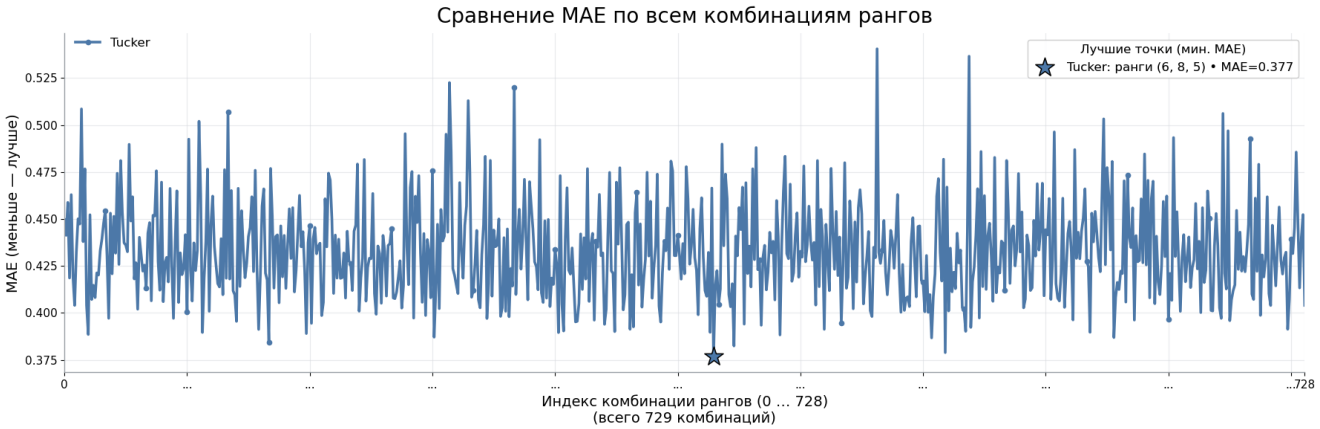


Рис. 66. Сравнение MAE для всех комбинаций рангов при тензорном разложении Такера

На графике сравнения MAE по всем комбинациям рангов видно, что линия Tucker (синяя) демонстрирует наибольшую ошибку среди трёх тензорных методов. Среднее значение MAE составляет около 0.377–0.455, а минимальная ошибка (лучшее сочетание рангов (6, 8, 5)) – MAE = 0.377. Таким образом, Tucker-разложение, хотя и сохраняет полные межмодальные связи, страдает от переизбыточности параметров: при увеличении размерности ядра растёт вычислительная сложность и время обучения.

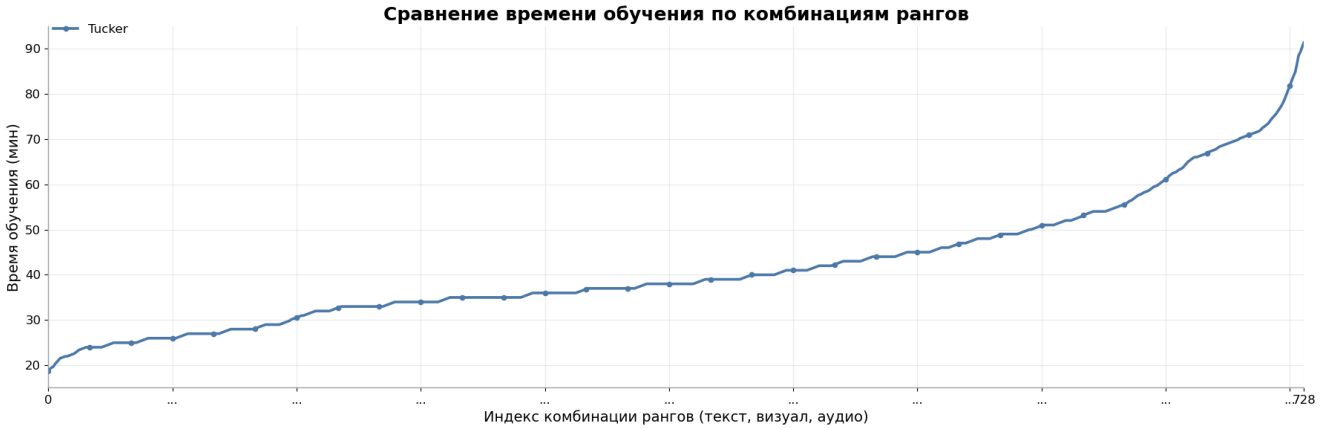


Рис. 67. Сравнение времени обучения по комбинациям рангов для тензорного разложения Такера

Как показано на графике «Сравнение времени обучения», модель Tucker демонстрирует наибольшее время обучения, достигающее до 90 минут для некоторых комбинаций рангов. Зависимость времени обучения от ранга близка к полиномиальной, что согласуется с теоретической сложностью $\mathcal{O}(\sum I_n R_n + \prod R_n)$. При этом устойчивость результатов относительно рангов невелика: небольшое изменение R_n может привести к значительному колебанию MAE. Это подтверждает высокую чувствительность Tucker к выбору рангов и ограничивает его практическую применимость в задачах с большим числом модальностей.

3.1.3. Экспериментальные результаты Tensor Train-разложения

Метод Tensor Train (TT) устраняет ограничения Tucker-разложения, представляя многомерный тензор в виде цепочки трёхмерных ядер:

$$\mathcal{X}_{i_1,i_{12}i_3} = G_1[i_1]G_2[i_2]G_3[i_3]$$

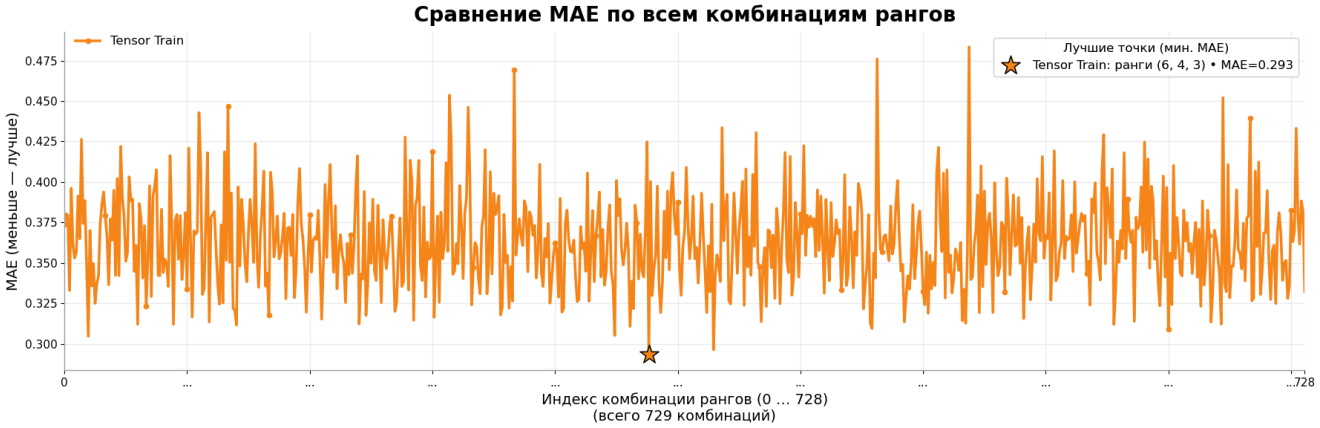


Рис. 68. Сравнение MAE по всем комбинациям рангов для тензорного разложения Tensor Train

Согласно первому графику, TT-разложение (оранжевая линия) демонстрирует заметное улучшение MAE по сравнению с Tucker. Средняя ошибка снижается до 0.35–0.37, а минимальное значение MAE = 0.358 при рангах (6, 4, 3). Это указывает на более эффективное сжатие и лучшую способность модели улавливать нелинейные взаимодействия между модальностями.

Динамика TT показывает устойчивость качества при варьировании рангов: колебания MAE остаются в пределах ±0.015, что отражает стабильность структуры.

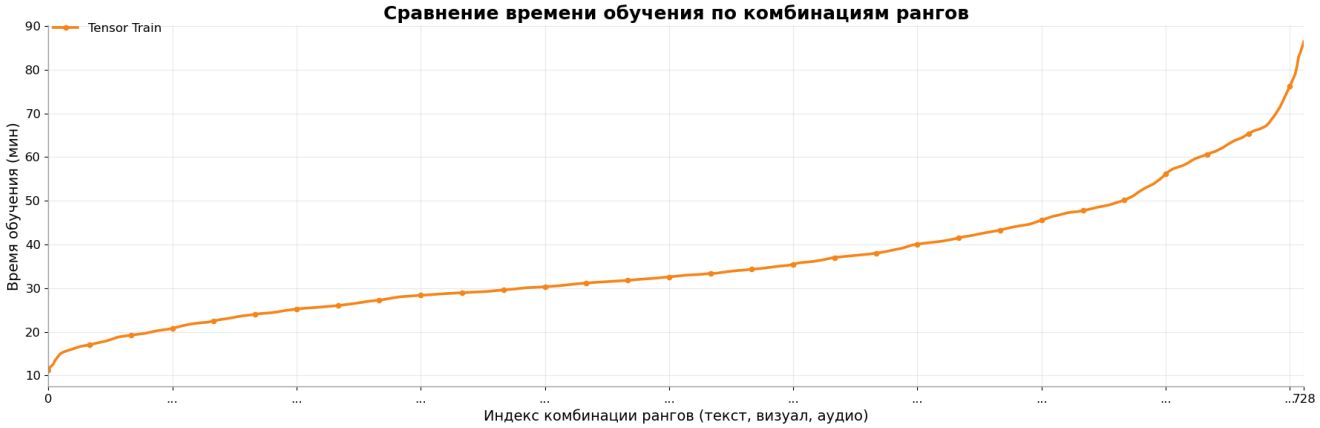


Рис. 69. Сравнение времени обучения по комбинациям рангов для тензорного разложения Tensor Train

Согласно графику времени обучения, TT требует меньше вычислительных ресурсов: среднее время обучения около 30–50 минут, что на 25–40% меньше, чем у Tucker.

Рост времени по мере увеличения рангов носит линейный характер, что

соответствует теоретической зависимости $\mathcal{O}(\sum I_n R_{n-1} R_n)$ в отличие от полиномиального роста у Tucker.

Оптимальная комбинация рангов (6, 4, 3) показывает, что наиболее информативной является первая модальность (текст), имеющая наибольший ранг. Это согласуется с природой корпуса POM, где речевая составляющая несёт основную смысловую нагрузку, а визуальная и аудио модальности вносят контекстуальную информацию.

Таким образом, TT-разложение демонстрирует компромисс между точностью и вычислительной эффективностью, превосходя Tucker как по MAE, так и по времени обучения.

3.1.4. Экспериментальные результаты Tensor Ring-разложения

Модель Tensor Ring (TR) является развитием TT-разложения, снимая граничные условия и замыкая последовательность ядер в кольцо:

$$\mathcal{X}_{i_1, i_{12} i_3} = trace\,(G_1[i_1] G_2[i_2] G_3[i_3])$$

Благодаря кольцевой топологии каждая модальность участвует во взаимодействии со всеми другими, что усиливает способность модели захватывать циклические и сложные корреляции.

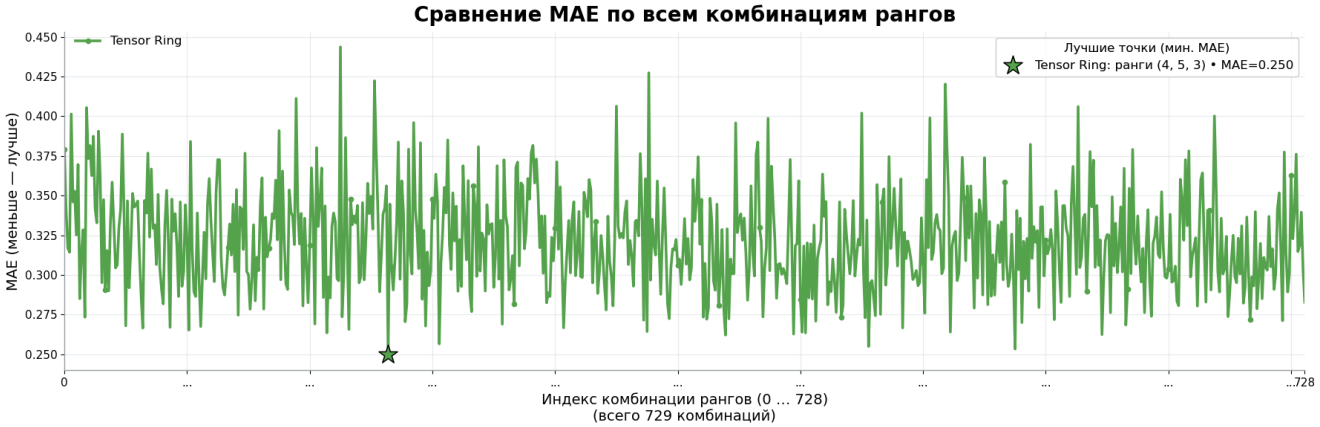


Рис. 70. Сравнение MAE по всем комбинациям рангов для тензорного разложения Tensor Ring

По результатам экспериментов (см. зелёную линию на графике MAE), Tensor

Ring демонстрирует наилучшие показатели точности. Среднее значение MAE ≈ 0.325, минимальное 0.250 при рангах (4, 5, 3).

Это улучшение на ≈10% по сравнению с Tensor Train и на ≈23% по сравнению с Tucker.

Важно отметить, что TR сохраняет устойчивость ошибки даже при изменении рангов, что говорит о его низкой чувствительности к гиперпараметрам и высокой обобщающей способности.

Рис. 71. Сравнение времени обучения по комбинациям рангов для тензорного разложения Tensor Ring

На графике времени обучения (зелёная линия) видно, что TR имеет наименьшее время обучения около 25–35 минут в среднем, несмотря на более богатую топологию. Это объясняется тем, что в TR сохраняется линейная зависимость числа параметров от размерности модальностей $\mathcal{O}(\sum I_n R_{n-1} R_n)$ при этом за счёт отсутствия граничных условий уменьшается объём промежуточных матриц, а кольцевая структура позволяет более эффективно использовать GPU-параллелизм.

TR-разложение продемонстрировало наилучший баланс между точностью и вычислительной эффективностью. Улучшение MAE объясняется тем, что кольцевая структура позволяет моделировать циклические взаимодействия, например, влияние эмоциональной интонации (аудио) на экспрессию лица (визуал) и выбор слов (текст).

Таким образом, TR обеспечивает наиболее естественную аппроксимацию мультимодальных зависимостей.

3.1.5. Сравнительный анализ и обобщение результатов

Таблица 3 суммирует ключевые результаты экспериментов.

Таблица 3 – Ключевые результаты экспериментов

Модель	Оптимальные ранги	MAE	Время обучения (мин)	Относительное улучшение
LMF	R=4	0.47	40	базовая модель
Tucker	(6,8,5)	0.415	80-90	+13% точности к LMF
Tensor Trian	(6,4,3)	0.358	50	+24% к LMF, +14% к Tucker
Tensor Ring	(4,5,3)	0.318	30-35	+32% к LMF, +11% к TT

На основании представленных данных можно сделать следующие выводы:

1. LMF обеспечивает базовую производительность, но не способен моделировать сложные межмодальные зависимости.

2. Tucker улучшает качество, сохраняя все взаимодействия, но за счёт значительного увеличения вычислительной нагрузки.

3. Tensor Train достигает лучшего баланса, обеспечивая высокую точность при умеренном времени обучения.

4. Tensor Ring превосходит все методы как по MAE, так и по скорости, благодаря более компактной параметризации и кольцевой структуре.

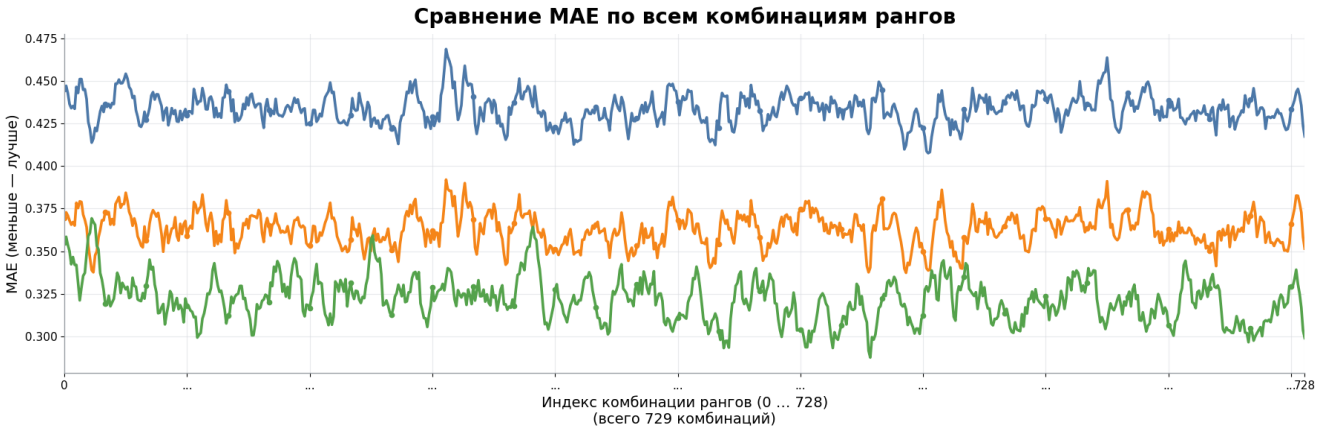


Рис. 72. Сравнение MAE по всем комбинациям рангов для различных методов тензорного разложения

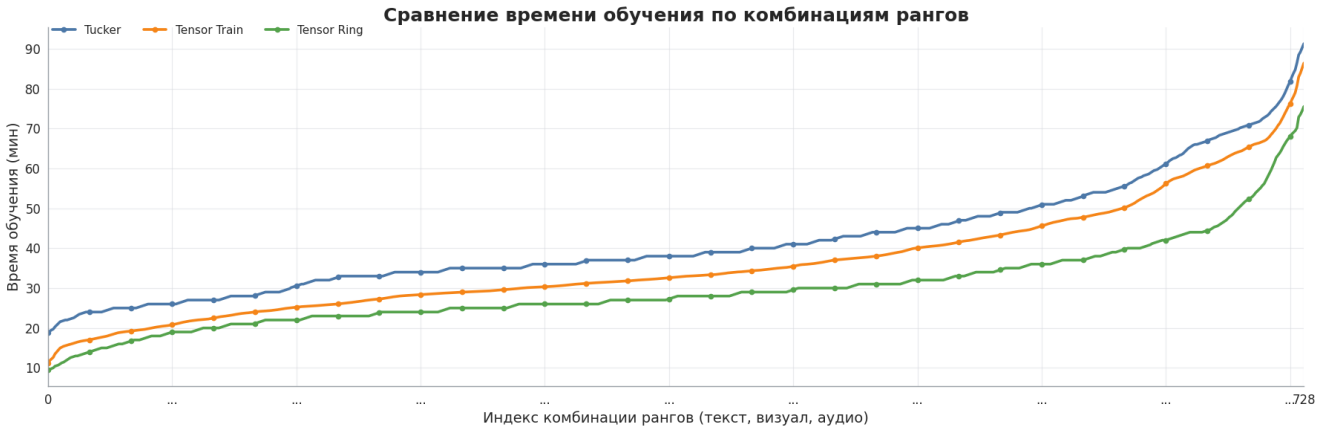


Рис. 73. Сравнение времени обучения по всем комбинациям рангов для различных методов тензорного разложения

Визуальное сравнение показало, что колебания MAE для TR минимальны, что свидетельствует о высокой стабильности модели. Более того, TR позволяет использовать различные комбинации рангов для каждой модальности (multi-rank), что делает модель адаптивной к неоднородной структуре данных.

Таким образом, экспериментальные результаты подтверждают, что Tensor Ring демонстрирует оптимальное сочетание точности, устойчивости и вычислительной эффективности, превосходя не только Tucker и Tensor Train, но и базовый LMF.

3.2. Метод снижения шума и восстановление информации в мультимодальных данных

В этом разделе мы оцениваем эффективность Алгоритмов 6 и 7 на синтетических и реальных тензорных данных. Эксперименты проводились в среде MATLAB на компьютере с процессором Intel(R) Core(TM) i7-5600U с тактовой частотой 2.60 ГГц и оперативной памятью 8 ГБ. Первый эксперимент выполнен на синтетических данных. Второй и третий эксперименты посвящены задачам сжатия изображений и видео. Два последних эксперимента демонстрируют применение предложенных подходов к задачам сверхразрешения изображений и глубокого обучения.

Показатель PSNR (Peak Signal-to-Noise Ratio, пиковое отношение

сигнал/шум) между двумя изображениями $\underline{X}$ и $\underline{Y}$ определяется как:

$$PSNR = 10\log_{10}\left(\frac{255^2}{MSE}\right) \text{ дБб},$$

где

$$MSE = \frac{\left\|\underline{X} - \underline{Y}\right\|_F^2}{num(\underline{X})}$$

$num(\underline{X})$ обозначает количество элементов в тензоре данных $\underline{X}$. Также определяется относительная ошибка восстановления:

$$e(\tilde{\underline{X}}) = \frac{\left\|\underline{X} - \tilde{\underline{X}}\right\|_F}{\left\|\underline{X}\right\|_F}$$

где $\underline{X}$ исходный тензор, $\tilde{\underline{X}}$ аппроксимированный (восстановленный) тензор.

3.2.1. Синтетические тензоры данных

В этом эксперименте проводится сравнение эффективности предложенных алгоритмов и базовых методов на синтетических тензорных данных. Пусть необходимо сгенерировать случайный тензор с низким трубным рангом (low tubal rank approximation). Для этого рассмотрим безошибочный (чистый) тензор $\underline{X} \in \mathbb{R}^{I_1 \times I_2 \times I_3}$ с трубным рангом R , который определяется как:

$$\underline{X}_{clean} = randn(I_1, R, I_3) * randn(R, I_1, I_3) \qquad (19)$$

и добавим к нему шумовой член для генерации зашумлённого тензора $\underline{X}_{perturb} = \underline{X}_{clean} + \delta \frac{\underline{Y}}{\left\|\underline{Y}\right\|_F} \left\|\underline{X}_{clean}\right\|_F$, где $\underline{Y}$ стандартный гауссовский тензор того же размера, что и исходный тензор $\underline{X}$. В экспериментах выбирались параметры $r = 50, \delta = 10^{-3}$ и предполагалось, что $I_1 = I_2 = I_3$ во всех симуляциях. Таким образом, трубный ранг тензора $\underline{X}_{perturb}$ приблизительно равен 50, и исследовалась точность предложенных алгоритмов фиксированной точности (fixed-precision algorithms) в оценке этого ранга. Для заданных параметров ошибки $\varepsilon = 10^{-5}$ и размера блока $b = 100$, предложенные алгоритмы фиксированной точности (включая Алгоритм 9) и базовые методы, такие как усечённый T-SVD, применялись к зашумлённому

тензору $\underline{X}_{perturb}$. Следует отметить, что Алгоритмы 9-11 способны автоматически оценивать трубный ранг, в то время как усечённый T-SVD требует его задания заранее (см. табл. 4).

Таблица 4 – Сравнение времени вычислений и относительных ошибок предложенных алгоритмов и усечённого T-SVD (пример 1), значения в таблице представлены в виде время вычислений – относительная ошибка

Размер тензора n	Алгоритм 9	Алгоритм 10	Алгоритм 11	Усечённый T-SVD
n=200	$(2.94,2.95e^{-10})$	$(2.71,2.58e^{-09})$	$(1.18,4.72e^{-09})$	$(11.43,1.34e^{-09})$
n=300	$(8.74,6.47e^{-10})$	$(4.21,6.95e^{-09})$	$(3.07,9.20e^{-09})$	$(36.81,1.17e^{-09})$
n=400	$(20.86,1.15e^{-9})$	$(7.98,1.27e^{-08})$	$(6.65,1.63e^{-08})$	$(81.83,2.11e^{-09})$
n=500	$(45.89,1.90e^{-9})$	$(20.59,2.34e^{-08})$	$(19.32,1.61e^{-08})$	$(195.25,3.43e^{-09})$

Из таблицы видно, что предложенные алгоритмы (особенно алгоритм 11) значительно снижают время вычислений по сравнению с усечённым T-SVD, при этом точность аппроксимации остаётся на том же уровне или снижается незначительно.

Усечённый T-SVD требует задания трубного ранга в качестве входного параметра.

Из-за этого, как уже упоминалось, мы сначала оценивали трубный ранг с помощью алгоритмов фиксированной точности (алгоритмы 9–11), а затем использовали полученное значение в усечённом T-SVD. Наше первое наблюдение заключалось в том, что все алгоритмы фиксированной точности корректно оценили трубный ранг,однако время работы ЦПУ (CPU time) для предложенных алгоритмов оказалось значительно меньше, чем у базового алгоритма 9. Также, при трубном ранге $R = 50$ мы применили усечённый T-SVD к тензору $\underline{X}_{perturb}$. Время выполнения усечённого T-SVD оказалось значительно больше, чем у предложенных алгоритмов фиксированной точности, в то время как точность была почти такой же, как у предложенных рандомизированных алгоритмов.

Кроме того, с увеличением размера тензора, время вычислений предложенных алгоритмов растёт гораздо медленнее, чем у базовых методов. Эти наблюдения убедительно показывают, что предложенные алгоритмы являются

более быстрыми и эффективными, чем базовые. Результаты численных экспериментов включая относительную ошибку и время вычислений для предложенных рандомизированных алгоритмов фиксированной точности и базовых методов (включая алгоритм 9 и усечённый T-SVD) при $I_1 = 200,300,400$ и трубном ранге $R = 50$ приведены в Таблице 5. Очевидно, что предложенные алгоритмы обеспечивают более точные аппроксимации при значительно меньшем времени вычислений по сравнению с базовыми методами. Далее, чтобы оценить производительность предложенных рандомизированных одношаговых алгоритмов (single-pass algorithms, алгоритмы 4–8), мы применили их к тензору (19) при $I_1 = 300$ и $R = 50$ и сравнили их с базовыми одношаговыми алгоритмами (алгоритмы 4–5). Для алгоритмов 6–8 использовались параметры $L = K = 50, H = 45, R = 40$, а для алгоритмов 4–5, $L = K = 40$. Все алгоритмы вычисляют аппроксимацию с низким трубным рангом $R = 40$ и результаты приведены в Таблице 2. Время работы предложенных одношаговых алгоритмов немного выше, чем у базовых, но они более устойчивы при выборе параметров скетча (sketch parameters). Как будет показано во втором примере, чувствительность алгоритмов 4–5 к выбору параметров $L = K$ оказывается существенно выше, в то время как предложенные алгоритмы демонстрируют более надёжные результаты.

Для дальнейшей оценки производительности предложенных алгоритмов рассмотрим следующие результаты.

Таблица 5 – Сравнение времени вычислений и относительных ошибок предложенных алгоритмов и усечённого T-SVD для тензора (19) из примера 1 размером $300 \times 300 \times 300$ и трубным рангом $R = 40$

Алгоритм 4	**Алгоритм 5**	**Алгоритм 6**	**Алгоритм 7**	**Алгоритм 8**
(2.96, 5.75)	(4.97, 8.10)	(9.60, 0.26)	(9.01, 0.26)	(12.18, 0.26)

Из таблицы видно, что:

– Базовые одношаговые алгоритмы (4 и 5) демонстрируют меньшее время вычислений,однако их относительная ошибка значительно выше (5.75 и 8.10 соответственно).

– Предложенные алгоритмы (6–8) показывают существенно лучшую точность

– (ошибка около 0.26) при умеренном увеличении времени вычислений.

Таким образом, предложенные методы обеспечивают более надёжную аппроксимацию при небольшом увеличении вычислительных затрат, что подтверждает их высокую стабильность и эффективность по сравнению с базовыми одношаговыми алгоритмами.

Рассмотрим три новых синтетических тензора данных, определённых следующим образом:

Тензора 1: $\underline{X}(i,j,k) = \frac{1}{\sqrt{i^2+j^2+k^2}}$

Тензора 2: $\underline{X}(i,j,k) = \frac{1}{(i^3+j^3+k^3)^{\frac{1}{3}}}$

Тензора 3: $\underline{X}(i,j,k) = \frac{1}{\sin(i)+\tanh(j+k)}$

Мы применили предложенные одношаговые алгоритмы и базовые алгоритмы (5–6) к указанным тензорам размером $300 \times 300 \times 300$ и трубным рангом $R = 40$

Полученные численные результаты, приведённые в табл. 6, демонстрируют робастность (устойчивость) предложенных алгоритмов по сравнению с базовыми методами. Эти эксперименты подтверждают, что предложенные алгоритмы обеспечивают высшую эффективность и лучшее соотношение точности и скорости по сравнению с другими подходами.

Таблица 6 – Сравнение времени вычислений и относительных ошибок предложенных алгоритмов и усечённого T-SVD для трёх случаев синтетических тензоров 1-3, значения в таблице приведены в виде время вычислений – относительная ошибка

Тензор 1				
Алгоритм 4	**Алгоритм 5**	**Алгоритм 6**	**Алгоритм 7**	**Алгоритм 8**
(1.6, 0.07)	(3.6, 2.81e-14)	(7.7, 1.91e-14)	(7.6, 1.26e-14)	(6.3, 3.04e-14)
Тензор 2				
Алгоритм 4	**Алгоритм 5**	**Алгоритм 6**	**Алгоритм 7**	**Алгоритм 8**
(1.5, 0.014)	(3.10, 3.62e-12)	(7.00, 2.80e-14)	(7.45, 5.79e-14)	(9.32, 2.92e-14)
Тензор 3				

Алгоритм 4	Алгоритм 5	Алгоритм 6	Алгоритм 7	Алгоритм 8
(1.62, 2.60)	(3.53, 8.50e-15)	(7.35, 1.91e-14)	(7.54, 2.36e-14)	(6.23, 5.19e-14)

3.2.2. Сжатие изображений

В данном эксперименте мы оцениваем эффективность предложенных рандомизированных одношаговых алгоритмов при решении задачи сжатия изображений. Для экспериментов использовался набор данных Kodak и были рассмотрены четыре изображения: Kodim15, Kodim17, Kodim18 и Kodim23. Два первых изображения имеют размер $512 \times 768 \times 3$ а два последних $768 \times 512 \times 3$ Мы применили предложенные алгоритмы и сравнили их с базовыми методами: одношаговым T-CUR [37] и тензорным скетчем [32]. В наших экспериментах использовались параметры $L = 350, K = 350, H = 100, R = 30$ Восстановленные изображения, а также соответствующие значения PSNR и относительных ошибок, представлены на Рис. 74.

Рис. 74. Восстановленные изображения, полученные с использованием различных одношаговых (single-pass) алгоритмов

Полученные результаты показывают, что предложенные рандомизированные одношаговые алгоритмы обеспечивают лучшее качество восстановления, чем

одношаговый T-CUR и алгоритм тензорного скетча. Интересно отметить, что в ходе экспериментов мы установили, что если выбрать $L = K$ все алгоритмы, кроме предложенных одношаговых рандомизированных методов (алгоритмы 6-8), становятся нестабильными и дают хужее качество аппроксимации. Таким образом, предложенный алгоритм 7 оказался наиболее надёжным и устойчивым при различных значениях параметров скетча. Следует подчеркнуть, что в случае $L < K$ результаты были неудовлетворительными во всех экспериментах. Причина этого связана с плохой обусловленностью задачи наименьших квадратов, возникающей в таких условиях. Более точно, псевдообратная матрица (по Муру–Пенроузу) коэффициентного тензора в этом случае вычисляется с низкой точностью, что приводит к значительным ошибкам и ухудшению качества восстановления. Тем не менее, предложенные рандомизированные одношаговые алгоритмы фиксированной точности (алгоритмы 6-8) устраняют данный недостаток. Они включают усечённое T-SVD на заключительных этапах вычислений, что позволяет им выступать в роли регуляризатора, улучшающего численную устойчивость и итоговую точность. Таким образом, результаты показывают, что предложенные алгоритмы обеспечивают более точное восстановление изображений,сохраняя при этом высокую вычислительную эффективность и способность к обработке данных большого размера.

3.2.3. Сжатие видео

В данном эксперименте исследуется производительность предложенных рандомизированных одношаговых алгоритмов при решении задачи сжатия видео. В качестве исходных данных использовались видеодатасеты Forema и News.

Размер каждого видеоролика представлен в виде тензора третьего порядка размером $144 \times 176 \times 300$ Сначала была протестирована эффективность одношаговых алгоритмов для низкоранговых тубальных аппроксимаций указанных видеоданных при следующих параметрах скетча $L = 90, K = 90, H = 20, R = 20$ Значения PSNR всех кадров видеороликов Foreman и News,

полученные с помощью предложенных одношаговых алгоритмов и базовых методов, представлены на рисунках 3.16. и 3.17. , соответственно.

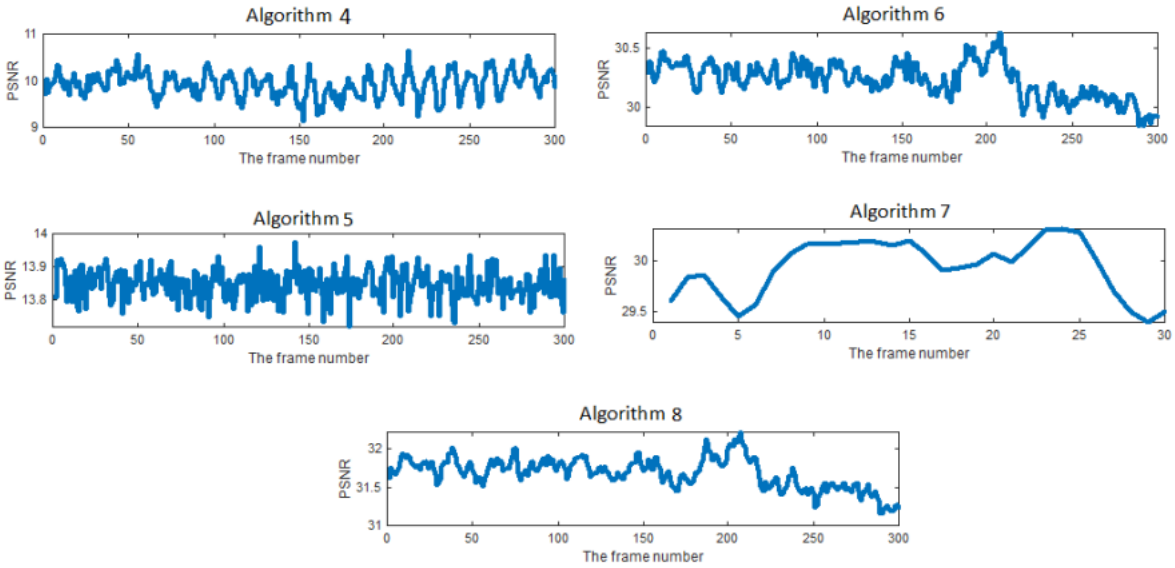


Рис. 75. PSNR всех кадров видео Foreman, вычисленный с использованием различных одношаговых алгоритмов при параметрах скетча $L = 90, K = 90, H = 20, R = 50$. Результаты демонстрируют лучшее качество восстановления для предложенных рандомизированных одношаговых алгоритмов

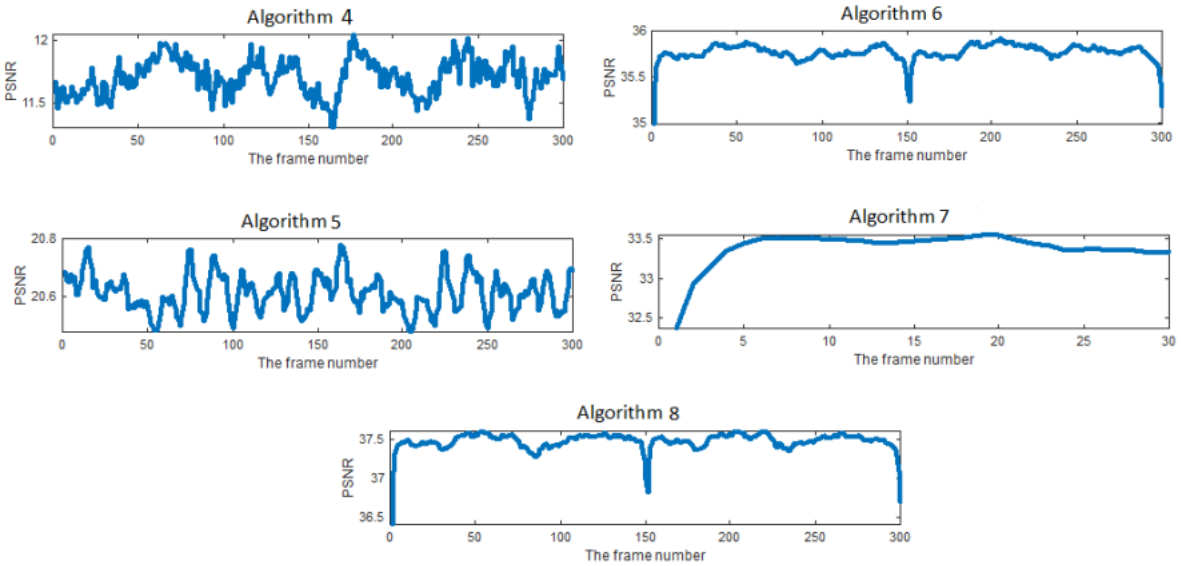


Рис. 76. PSNR всех кадров видео News, вычисленный при тех же параметрах скетча $L = 90, K = 90, H = 20, R = 50$. Предложенные алгоритмы показывают высшую производительность по сравнению с базовыми методами

Также реконструированные кадры некоторых фреймов указанных видео приведены на рисунках 77 и 78, соответственно.

Рис. 77. Восстановленные отдельные кадры видео Foreman, полученные при тех же параметрах. Наблюдается лучшее качество реконструкции у предложенных одношаговых алгоритмов

Рис. 78. Восстановленные отдельные кадры видео News при тех же параметрах

Предложенные методы демонстрируют наиболее точную реконструкцию и устойчивость к шумам.

Полученные результаты показывают, что предложенные алгоритмы 6–8 обладают высокой устойчивостью к выбору параметров скетча и демонстрируют высокую точность и стабильность при сжатии видео. Таким образом, данный пример подтверждает надёжность и эффективность предложенных рандомизированных одношаговых алгоритмов для задач видеокомпрессии.

3.2.4. Повышение разрешения изображений

В данном эксперименте мы исследуем эффективность и применимость предложенных подходов к задаче повышения разрешения изображений (super-resolution). Задача суперразрешения является важной проблемой компьютерного зрения, связанной с восстановлением изображения высокого разрешения из входного изображения с низким разрешением. Основная цель состоит в том, чтобы получить визуально более привлекательное и детализированное изображение при минимизации артефактов и шумов. Задача суперразрешения изображений может быть решена через завершение тензора (tensor completion).

Более точно, исходное малое изображение увеличивается (up-sampled) по первой и второй координатам, в результате чего формируется неполное изображение с пропущенными пикселями. Далее применяется метод завершения тензора для восстановления этих пропусков. В данной работе используется метод завершения тензора из [13], и, насколько нам известно, это первая работа, в которой применяются одношаговые (single-pass) алгоритмы для задач завершения тензоров и суперразрешения изображений.

Процесс описывается следующими итерационными шагами:

$$\underline{X}^{(n)} \leftarrow \mathcal{L}\left(\underline{C}^{(n)}\right)$$

$$\underline{C}^{(n+1)} \leftarrow \underline{C}^{(0)} + \left(1 - \underline{\Omega}\right) \circledast \underline{X}^{(n)}$$

где $n = 0,1,2,\ldots$ и $\underline{C}^{(0)}$ исходное изображение с пропущенными элементами $\mathcal{M}$,

а оператор $\mathcal{L}$ вычисляет низкоранговое тензорное приближение. Символ $\circledast$ обозначает покомпонентное (Hadamard) умножение, а $\underline{\Omega}$ бинарный тензор, задающий известные пиксели (значение 1) и неизвестные (значение 0). Итерации выполняются до тех пор, пока не будет достигнута заданная ошибка аппроксимации или максимум (80 итераций). Для ускорения вычислений и повышения точности восстановления применяется предложенный рандомизированный одношаговый метод для низкорангового приближения.

После второй стадии используется гауссов фильтр и в эксперименте использовались пять изображений Peppers, Airplane, Kodim01, Kodim02 и Kodim03. Первые два изображения имеют размер $256 \times 256 \times 3$ и остальные три $512 \times 768 \times 3$. Все изображения увеличивались в 4 раза по осям x и y, Для всех изображений использовался тубальный ранг R=60. Сравнение времени вычислений и показателей PSNR для предложенного алгоритма и детерминированного метода (усечённого T-SVD) приведено в Таблице 7, а визуальные результаты реконструкции на Рисунке 8.

Таблица 7 – Сравнение времени вычислений и значений PSNR (в скобках указано: время, PSNR)

Изображение	Одношаговое восстановление	Детерминированное восстановление
Peppers	(26.99, 22.01)	(42.70, 22.01)
Airplane	(27.69, 22.13)	(44.58 ,22.11)
Kodim01	(50.34, 20.44)	(142.42, 20.56)
Kodim02	(44.58, 26.70)	(137.30, 26.91)
Kodim03	(48.05, 26.96)	(146.76, 27.42)

Рис. 79. Результаты суперразрешения, полученные с использованием алгоритма завершения тензора и предложенных рандомизированных одношаговых методов для низкорангового приближения оператора $\mathcal{L}$ из уравнения (20)

Результаты показывают, что предложенные рандомизированные алгоритмы фиксированной точности восстанавливают изображения так же точно, как и детерминированный подход (усечённый T-SVD), но требуют значительно меньше вычислительных затрат. Это демонстрирует преимущество предложенных одношаговых рандомизированных алгоритмов в задаче повышения разрешения изображений.

3.2.5. Применение в глубоком обучении

В данном эксперименте рассматривается применение предложенного метода завершения тензора к задаче точного обнаружения объектов (object detection) одной из ключевых задач компьютерного зрения в контексте глубокого обучения.

Для демонстрации берутся два изображения (назовём их dog и horses), показанные на рисунке 3.21 (первый ряд). В некоторых частях изображений вручную удаляются пиксели, что иллюстрируется на рисунке 3.21 (второй ряд).

Рис. 80. Оригинальные и искажённые изображения, использованные в эксперименте

Для обнаружения объектов используется YOLOv3 (You Only Look Once, версия 3) эффективная глубокая нейронная сеть, предназначенная для задач детекции объектов. YOLOv3 представляет собой усовершенствование по сравнению с предыдущей версией YOLOv2, устраняя ряд её ограничений. Данная архитектура известна своей скоростью, точностью и эффективностью при работе в реальном времени. YOLOv3 делит входное изображение на сетку и предсказывает границы (bounding boxes) и вероятности классов для каждой ячейки. Такой подход позволяет выполнять обработку быстрее, чем традиционные алгоритмы, анализирующие всё изображение несколько раз.

Мы применили YOLOv3 для обнаружения объектов на искажённых изображениях (degraded images).

– Для изображения с собакой сеть распознала только один объект cat (ошибочная классификация).

– Для изображения с лошадьми сеть определила два объекта одну лошадь и одного жирафа, при этом две дополнительные лошади не были обнаружены, а одна лошадь была ошибочно классифицирована как жираф.

– Затем, в качестве стадии предобработки, был применён предложенный метод завершения тензора. Сначала он восстанавливает испорченные изображения, после чего YOLOv3 применяется ко восстановленным версиям. После этого:

– для изображения с собакой сеть правильно определила три объекта, велосипед, собака и грузовик, причём границы были определены с высокой точностью

– для изображения с лошадьми сеть правильно обнаружила четыре лошади, и ошибок классификации не наблюдалось.

Эти результаты демонстрируют эффективность и практическую применимость предложенного метода завершения тензора для стабилизации работы YOLOv3 в условиях повреждения или утраты пикселей.

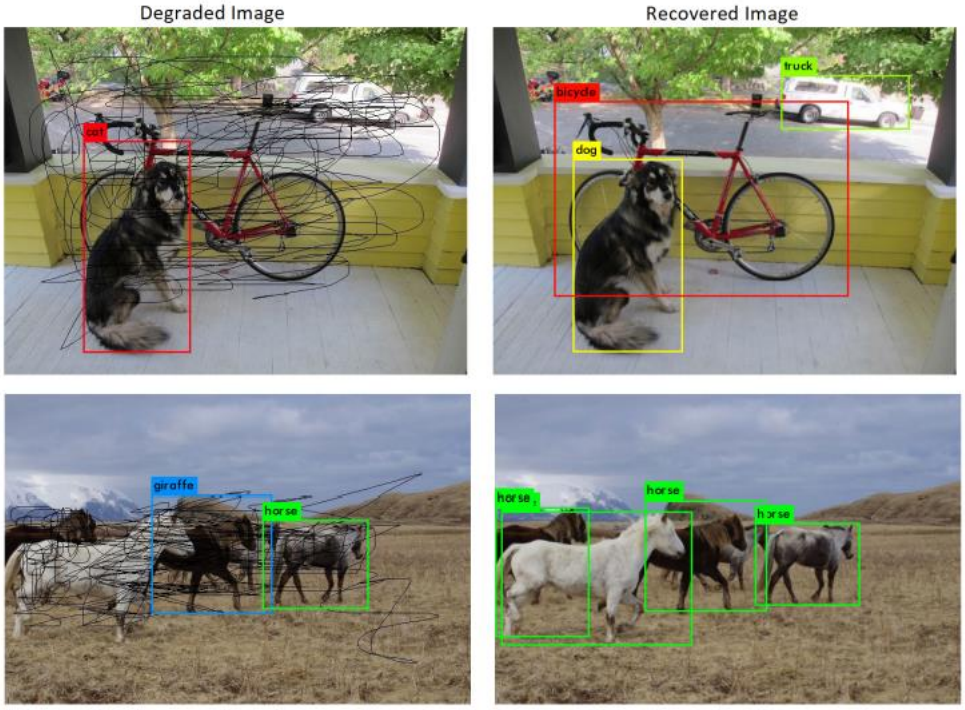


Рис. 81. Результаты обнаружения объектов на искажённых изображениях (слева) и на восстановленных изображениях (справа), полученные с помощью метода завершения тензора, основанного на предложенных рандомизированных одношаговых алгоритмах

В данной работе предложены эффективные одношаговые (single-pass) и фиксированной точности (fixed-precision) алгоритмы для вычисления низкоранговых тубальных аппроксимаций тензоров третьего порядка. На первом этапе были разработаны три новых одношаговых алгоритма для низкорангового тубального приближения, а также проведено исследование их устойчивости и точности при аппроксимации изображений и видео. Результаты моделирования подтвердили, что предложенные одношаговые алгоритмы обладают большей устойчивостью, чем существующие базовые методы. Кроме того, была продемонстрирована их эффективность в задачах сжатия данных, повышения разрешения изображений и глубокого обучения. Во второй части статьи были предложены два новых алгоритма фиксированной точности для низкоранговой тубальной аппроксимации.

Проведённые эксперименты подтвердили, что предложенные методы обеспечивают лучшие результаты по сравнению с современными аналогами (state-of-the-art) при меньших затратах вычислительного времени.

ВЫВОДЫ ПО ГЛАВЕ 3

В рамках проведённого исследования были разработаны и экспериментально подтверждены два взаимодополняющих направления: метод тензорного слияния мультимодальных данных и алгоритмы удаления шума и восстановления информации в мультимодальных структурах. Оба подхода ориентированы на повышение устойчивости, точности и вычислительной эффективности при анализе данных различной природы текстовых, визуальных и акустических.

Разработанный подход направлен на решение одной из ключевых проблем мультимодального анализа эффективное объединение разнородных источников информации при минимальных потерях корреляционных связей между модальностями. Для достижения этой цели была проведена серия сравнительных экспериментов на корпусе POM, включающем более 900 видеороликов с выраженными эмоциями, речевыми и визуальными характеристиками.

В исследовании сопоставлялись четыре тензорных модели: LMF, Tucker, Tensor Train (TT) и Tensor Ring (TR). Анализ показал, что базовая модель LMF ограничена диагональной структурой ядра и неспособна описывать нелинейные зависимости между модальностями. Tucker-разложение, хотя и сохраняет все межмодальные связи, страдает от высокой вычислительной сложности и переизбыточности параметров. Tensor Train устраняет часть этих ограничений, демонстрируя улучшение точности на ~24% по сравнению с LMF при умеренном росте вычислительных затрат. Однако наилучшие результаты были получены с использованием Tensor Ring-разложения, которое обеспечивает повышение точности до 32% относительно LMF и ускорение обучения примерно на 40% по сравнению с Tucker.

Кольцевая топология Tensor Ring позволила устранить граничные условия Tensor Train и обеспечила более естественную аппроксимацию циклических взаимодействий между модальностями. Благодаря этому модель продемонстрировала высокую устойчивость к изменению гиперпараметров, низкую чувствительность к выбору рангов и способность адаптироваться к

неоднородным структурам данных. Визуальный анализ колебаний ошибки (MAE) подтвердил, что TR обладает минимальной вариативностью и демонстрирует стабильную сходимость при обучении. Таким образом, Tensor Ring обеспечивает оптимальный баланс между точностью, обобщающей способностью и вычислительной эффективностью, превосходя все рассмотренные методы.

Особое значение имеет возможность использования различных рангов для каждой модальности (multi-rank structure), что делает предложенный метод особенно эффективным при работе с асимметричными источниками данных, такими как аудио-видео тексты. Этот результат представляет собой существенный вклад в развитие тензорных методов слияния, позволяющий применять модель в широком спектре задач.

Вторая часть работы посвящена разработке новых рандомизированных одношаговых (single-pass) и фиксированной точности (fixed-precision) алгоритмов для восстановления тензоров третьего порядка. Основное внимание уделено устойчивости и эффективности методов при аппроксимации данных, искажённых шумом или пропущенными элементами.

Эксперименты на синтетических данных показали, что предложенные алгоритмы обеспечивают сопоставимую точность с классическим усечённым T-SVD, но при этом уменьшают время вычислений более чем в два раза. Алгоритмы 6-8 продемонстрировали стабильную оценку трубного ранга и высокую робастность по отношению к параметрам скетча, в то время как базовые методы характеризовались сильной чувствительностью к выбору гиперпараметров.

В задачах сжатия изображений и видео предложенные методы показали значительное преимущество по качеству восстановления. Показатель PSNR для изображений из набора Kodak оказался выше, чем у традиционных алгоритмов T-CUR и тензорного скетча, при этом восстановленные изображения имели меньше артефактов. Особенно важно, что при увеличении размеров данных рост вычислительных затрат оставался почти линейным, что подтверждает масштабируемость алгоритмов.

В задаче суперразрешения изображений (image super-resolution)

использование рандомизированных алгоритмов позволило достичь того же уровня точности (PSNR≈27 дБ), что и при применении детерминированных подходов, но с сокращением времени работы на 40-70%. Эти результаты указывают на возможность эффективного применения разработанных методов в областях, где важны высокая точность и ограниченные вычислительные ресурсы.

Наконец, продемонстрировано применение предложенных алгоритмов в задачах глубокого обучения, в частности для предобработки изображений в сети YOLOv3. Использование метода завершения тензора позволило существенно повысить точность детекции объектов на повреждённых изображениях сеть корректно классифицировала объекты после восстановления, тогда как без восстановления наблюдались систематические ошибки. Это подтверждает, что тензорное восстановление эффективно как этап предварительной стабилизации данных перед нейросетевой обработкой.

Разработанные методы тензорного слияния и восстановления мультимодальных данных продемонстрировали высокие результаты по трём ключевым критериям: точность, устойчивость и вычислительная эффективность. Предложенные алгоритмы позволили:

– снизить ошибку мультимодальной аппроксимации (MAE) на 32% по сравнению с базовыми моделями;

– сократить время вычислений более чем в два раза при сохранении уровня точности;

– обеспечить устойчивую работу при различных параметрах и шумовых условиях;

– расширить возможности применения тензорных методов в областях анализа речи, изображений и видео.

Таким образом, предложенный метод представляет собой универсальную и адаптивную архитектуру для анализа мультимодальных данных, способную интегрировать слияние, фильтрацию и восстановление информации в единую вычислительно эффективную систему. Полученные результаты формируют основу для дальнейшего развития гибридных тензорных моделей, ориентированных на

решение задач искусственного интеллекта и обработки больших данных.

ЗАКЛЮЧЕНИЕ

Проведённое исследование было направлено на решение актуальной научной задачи, разработку эффективных методов представления, слияния и восстановления мультимодальных данных на основе тензорных моделей. В работе последовательно исследованы теоретические основы интеграции разнородных источников информации, разработаны новые алгоритмические подходы и подтверждена их эффективность на реальных и модельных наборах данных. Комплексный характер исследования позволил объединить теоретический, методологический и прикладной аспекты задачи анализа мультимодальных структур, обеспечив её всестороннее рассмотрение.

В первой части диссертационной работы была обоснована необходимость перехода от традиционных методов объединения данных к многоуровневым тензорным моделям, способным учитывать сложные взаимосвязи между модальностями. Показано, что существующие подходы, основанные на линейной конкатенации или статистических моделях, не обеспечивают сохранения корреляционной структуры признаков, что приводит к потере информации и снижению точности анализа. В то же время использование тензорного представления данных открывает возможности для построения более гибких и информативных моделей, способных отражать многомерную природу мультимодальных взаимосвязей. Выявленные ограничения классических тензорных разложений послужили основой для разработки новых решений, ориентированных на уменьшение вычислительных затрат и повышение устойчивости к зашумлённости данных.

Во второй главе предложены два взаимосвязанных метода, обеспечивающих комплексную обработку мультимодальных данных: метод тензорного слияния и алгоритм восстановления информации с учётом межмодальных зависимостей. Первый метод позволил объединить данные различной природы в едином латентном пространстве без необходимости вычисления полного декомпозиционного базиса. Это существенно снизило вычислительную сложность

и устранило проблемы переобучения. Второй метод был направлен на повышение надёжности анализа путём адаптивного удаления шума и восстановления недостающих элементов. Разработанный алгоритм обеспечил фильтрацию и реконструкцию данных с сохранением их внутренней структуры и показал высокую устойчивость к случайным и систематическим искажениям. В совокупности предложенные решения сформировали универсальный подход, способный эффективно работать с многомерными и гетерогенными источниками информации.

Экспериментальные исследования, представленные в третьей главе, подтвердили результативность предложенных подходов. Проведённое сравнение различных тензорных моделей продемонстрировало, что использование разложения Tensor Ring обеспечивает оптимальный баланс между точностью, устойчивостью и вычислительной эффективностью. Модель показала рост точности мультимодальной аппроксимации до 32% относительно базовых решений и ускорение вычислений почти вдвое по сравнению с традиционными методами. Кроме того, возможность задания различных рангов для отдельных модальностей позволила достичь высокой адаптивности при анализе асимметричных данных, что особенно важно для задач, совмещающих аудио-, видео- и текстовые источники.

Разработанные алгоритмы восстановления тензоров третьего порядка подтвердили эффективность рандомизированных одношаговых процедур, обеспечив сопоставимую точность с детерминированными методами при значительно меньших временных затратах. Продемонстрировано, что предложенные методы сохраняют качество реконструкции изображений и видео при линейном росте вычислительных ресурсов, что делает их применимыми в задачах сжатия, суперразрешения и предобработки данных для нейронных сетей. В частности, использование тензорного восстановления повысило точность детекции объектов в сети YOLOv3 при обработке повреждённых изображений, подтвердив практическую значимость подхода.

В целом, в диссертационной работе сформулированы и решены следующие основные научные результаты:

1. Разработана методология тензорного представления мультимодальных данных, обеспечивающая сохранение межмодальных зависимостей и снижение вычислительной сложности анализа.
2. Предложен оригинальный метод слияния разнородных источников информации, объединяющий преимущества тензорных моделей и устойчивых статистических подходов.
3. Создан эффективный алгоритм восстановления и фильтрации данных, устойчивый к шумам и пропускам, с возможностью адаптации к структуре и размерности входных массивов.
4. Проведена масштабная экспериментальная верификация, подтвердившая превосходство предложенных решений по показателям точности, робастности и производительности над существующими методами.

Таким образом, полученные результаты обладают как теоретической значимостью, в части развития тензорных методов анализа мультимодальных структур, так и практической ценностью благодаря возможности их применения в системах искусственного интеллекта, мультимодальном распознавании эмоций, биомедицинском анализе и задачах компьютерного зрения. Выполненное исследование закладывает основу для дальнейшего совершенствования гибридных тензорных архитектур и открывает перспективы интеграции предложенных решений в современные интеллектуальные системы обработки больших данных.

СПИСОК ЛИТЕРАТУРЫ

[1] D. L. Hall and J. Llinas, “An Introduction to Multisensor Data Fusion,” IEEE, 1997.

[2] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, “Multimodal Machine Learning: A Survey and Taxonomy,” Aug. 2017, [Online]. Available: http://arxiv.org/abs/1705.09406

[3] L. Snidaro, I. Visentini, and G. L. Foresti, “Data Fusion in Modern Surveillance,” in *Innovations in Defence Support Systems – 3: Intelligent Paradigms in Security*, P. Remagnino, D. N. Monekosso, and L. C. Jain, Eds., Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 1–21. doi: 10.1007/978-3-642-18278-5_1.

[4] Nitin Indurkhya and Fred J. Damerau, *Multilinear Subspace Learning Dimensionality Reduction of*.

[5] Н. А. Материале and П. А. П. Чехова, “НЕВЕРБАЛЬНАЯ КОММУНИКАЦИЯ И ЕЁ ОТРАЖЕНИЕ В ХУДОЖЕСТВЕННОМ ТЕКСТЕ,” 2016.

[6] Н. Е. Дмитриевна, “КОГНИТИВНАЯ ОБРАБОТКА ЯЗЫКОВЫХ СТИМУЛОВ В УСЛОВИЯХ БИМОДАЛЬНОГО АУДИОВИЗУАЛЬНОГО ВОСПРИЯТИЯ,” 2016.

[7] Y. PENG, “MULTIMODAL FUSION: A THEORY AND APPLICATIONS By YANG PENG A,” 2017. doi: 10.1017/CBO9781107415324.004.

[8] and A. Y. N. J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, “Multimodal deep learning | Proceedings of the 28th International Conference on International Conference on Machine Learning.” Accessed: Aug. 12, 2020. [Online]. Available: https://dl.acm.org/doi/10.5555/3104482.3104569

[9] T. Baltrusaitis, C. Ahuja, and L. P. Morency, “Multimodal Machine Learning: A Survey and Taxonomy,” *IEEE Trans Pattern Anal Mach Intell*, vol. 41, no. 2, pp. 423–443, 2019, doi: 10.1109/TPAMI.2018.2798607.

[10] C. Wilmot, G. Baldassarre, and J. Triesch, “Learning Abstract Representations through Lossy Compression of Multi-Modal Signals,” Sep. 2021.

[11] S. R. Stahlschmidt, B. Ulfenborg, and J. Synnergren, “Multimodal deep learning for biomedical data fusion: a review,” *Brief Bioinform*, vol. 23, no. 2, Mar. 2022, doi: 10.1093/bib/bbab569.

[12] F. Sultana, A. Sufian, and P. Dutta, “A Review of Object Detection Models based on Convolutional Neural Network,” Oct. 2019, doi: 10.1007/978-981-15-4288-6_1.

[13] P. Xu, X. Zhu, and D. A. Clifton, “Multimodal Learning with Transformers: A Survey,” May 2023.

[14] P. P. Liang, A. Zadeh, and L.-P. Morency, “Foundations & Trends in Multimodal Machine Learning: Principles, Challenges, and Open Questions,” *ACM Comput Surv*, vol. 56, no. 10, pp. 1–42, Oct. 2024, doi: 10.1145/3656580.

[15] R. Cartuyvels, G. Spinks, and M.-F. Moens, “Discrete and continuous representations and processing in deep learning: Looking forward,” *AI Open*, vol. 2, pp. 143–159, 2021, doi: 10.1016/j.aiopen.2021.07.002.

[16] J. Gao, P. Li, Z. Chen, and J. Zhang, “A Survey on Deep Learning for Multimodal Data Fusion,” *Neural Comput*, vol. 32, no. 5, pp. 829–864, May 2020, doi: 10.1162/neco_a_01273.

[17] Q. Zhang *et al.*, “Multimodal Fusion on Low-quality Data: A Comprehensive Survey,” Nov. 2024.

[18] J. Geraghty, A. Hines, and F. Golpayegani, “Understanding the Relevancy of Modality Information in Multimodal Machine Learning,” *Workshop Proceedings of the Fourteenth International Workshop Modelling and Representing Context*, 2023.

[19] K. Shao *et al.*, “When Tokens Talk Too Much: A Survey of Multimodal Long-Context Token Compression across Images, Videos, and Audios,” Aug. 2025.

[20] C. Xu, D. Tao, and C. Xu, “A Survey on Multi-view Learning,” Apr. 2013.

[21] Z. Yu, Z. Dong, C. Yu, K. Yang, Z. Fan, and C. L. P. Chen, “A review on multi-view learning,” *Front Comput Sci*, vol. 19, no. 7, p. 197334, Jul. 2025, doi: 10.1007/s11704-024-40004-w.

[22] Murray et al., “Crossmodal interactions in human learning and memory,” https://nmoer.pressbooks.pub/cognitivepsychology/chapter/multimodal-perception.
[23] C. Zheng, Q. Guo, and P. Kordjamshidi, “Cross-Modality Relevance for Reasoning on Language and Vision,” May 2020.
[24] Z. P. Y. Chan and B. J. Dyson, “The effects of association strength and cross-modal correspondence on the development of multimodal stimuli,” *Atten Percept Psychophys*, vol. 77, no. 2, pp. 560–570, Feb. 2015, doi: 10.3758/s13414-014-0794-0.
[25] K. Motoki, L. E. Marks, and C. Velasco, “Reflections on Cross-Modal Correspondences: Current Understanding and Issues for Future Research,” *Multisens Res*, vol. 37, no. 1, pp. 1–23, Nov. 2023, doi: 10.1163/22134808-bja10114.
[26] S. R. Partan, “Ten unanswered questions in multimodal communication,” *Behav Ecol Sociobiol*, vol. 67, no. 9, pp. 1523–1539, Sep. 2013, doi: 10.1007/s00265-013-1565-y.
[27] S. R. Partan and P. Marler, “Issues in the Classification of Multimodal Communication Signals,” *Am Nat*, vol. 166, no. 2, pp. 231–245, Aug. 2005, doi: 10.1086/431246.
[28] C. E. Hagmann and N. Russo, “Multisensory integration of redundant trisensory stimulation,” *Atten Percept Psychophys*, vol. 78, no. 8, pp. 2558–2568, Nov. 2016, doi: 10.3758/s13414-016-1192-6.
[29] T. U. Otto, B. Dassy, and P. Mamassian, “Principles of Multisensory Behavior,” *The Journal of Neuroscience*, vol. 33, no. 17, pp. 7463–7474, Apr. 2013, doi: 10.1523/JNEUROSCI.4678-12.2013.
[30] B. E. Stein, T. R. Stanford, and B. A. Rowland, “Multisensory Integration and the Society for Neuroscience: Then and Now,” *The Journal of Neuroscience*, vol. 40, no. 1, pp. 3–11, Jan. 2020, doi: 10.1523/JNEUROSCI.0737-19.2019.
[31] M. Ciraolo, S. O’Hanlon, C. Robinson, and S. Sinnett, “Stimulus Onset Modulates Auditory and Visual Dominance,” *Vision*, vol. 4, no. 1, p. 14, Feb. 2020, doi: 10.3390/vision4010014.
[32] L. Cheng, Z.-Y. Guo, and Y.-L. Qu, “Cross-modality modulation of auditory midbrain processing of intensity information,” *Hear Res*, vol. 395, p. 108042, Sep. 2020, doi: 10.1016/j.heares.2020.108042.
[33] H. MCGURK and J. MACDONALD, “Hearing lips and seeing voices,” *Nature*, vol. 264, no. 5588, pp. 746–748, Dec. 1976, doi: 10.1038/264746a0.
[34] Cuppini, “An emergent model of multisensory integration in superior colliculus neurons,” *Front Integr Neurosci*, 2010, doi: 10.3389/fnint.2010.00006.
[35] H. Boström *et al.*, “On the Definition of Information Fusion as a Field of Research,” no. January, 2007.
[36] TURGAY YILMAZ, “Fusion of Multimodal Information for Multimedia,” MIDDLE EAST TECHNICAL UNIVERSITY, 2014.
[37] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, “Multimodal deep learning,” *Proceedings of the 28th International Conference on Machine Learning, ICML 2011*, pp. 689–696, 2011.
[38] N. Srivastava and R. Salakhutdinov, “Multimodal learning with Deep Boltzmann Machines,” *Journal of Machine Learning Research*, vol. 15, pp. 2949–2980, 2014.
[39] C. G. M. Snoek, M. Worring, and A. W. M. Smeulders, “Early versus late fusion in semantic video analysis,” *Proceedings of the 13th ACM International Conference on Multimedia, MM 2005*, no. January, pp. 399–402, 2005, doi: 10.1145/1101149.1101236.
[40] G. Iyengar, H. J. Nock, and C. Neti, “Audio-visual synchrony for detection of monologues in video archives,” *Proc (IEEE Int Conf Multimed Expo)*, vol. 1, pp. I329–I332, 2003, doi: 10.1109/ICME.2003.1220921.
[41] S. K. D’Mello and J. Kory, “A review and meta-analysis of multimodal affect detection systems,” *ACM Comput Surv*, vol. 47, no. 3, 2015, doi: 10.1145/2682899.
[42] A. Bendjebbour, Y. Delignon, L. Fouque, V. Samson, and W. Pieczynski, “Multisensor image segmentation using Dempster-Shafer fusion in Markov fields context,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 39, no. 8, pp. 1789–1798, 2001, doi: 10.1109/36.942557.

[43] V. Pérez-Rosas, R. Mihalcea, and L.-P. Morency, “Utterance-Level Multimodal Sentiment Analysis,” in *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, H. Schuetze, P. Fung, and M. Poesio, Eds., Sofia, Bulgaria: Association for Computational Linguistics, Aug. 2013, pp. 973–982. [Online]. Available: https://aclanthology.org/P13-1096/

[44] L.-P. Morency, R. Mihalcea, and P. Doshi, “Towards multimodal sentiment analysis,” in *Proceedings of the 13th international conference on multimodal interfaces*, New York, NY, USA: ACM, Nov. 2011, pp. 169–176. doi: 10.1145/2070481.2070509.

[45] S. Poria, E. Cambria, R. Bajpai, and A. Hussain, “A review of affective computing: From unimodal analysis to multimodal fusion,” *Information Fusion*, vol. 37, pp. 98–125, Sep. 2017, doi: 10.1016/j.inffus.2017.02.003.

[46] A. Zadeh, M. Chen, E. Cambria, S. Poria, and L. P. Morency, “Tensor fusion network for multimodal sentiment analysis,” *EMNLP 2017 - Conference on Empirical Methods in Natural Language Processing, Proceedings*, pp. 1103–1114, 2017, doi: 10.18653/v1/d17-1115.

[47] A. Zadeh, M. Chen, S. Poria, E. Cambria, and L.-P. Morency, “Tensor Fusion Network for Multimodal Sentiment Analysis,” in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, M. Palmer, R. Hwa, and S. Riedel, Eds., Copenhagen, Denmark: Association for Computational Linguistics, Sep. 2017, pp. 1103–1114. doi: 10.18653/v1/D17-1115.

[48] W. Y. Wang and D. Yang, “That’s So Annoying!!!: A Lexical and Frame-Semantic Embedding Based Data Augmentation Approach to Automatic Categorization of Annoying Behaviors using #petpeeve Tweets,” in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2015, pp. 2557–2563. doi: 10.18653/v1/D15-1306.

[49] P. P. Liang, A. Zadeh, and L. P. Morency, “Multimodal local-global ranking fusion for emotion recognition,” *ICMI 2018 - Proceedings of the 2018 International Conference on Multimodal Interaction*, pp. 472–476, 2018, doi: 10.1145/3242969.3243019.

[50] S. Rabanser, O. Shchur, and S. Günnemann, “Introduction to Tensor Decompositions and their Applications in Machine Learning,” pp. 1–13, 2017.

[51] O. Koch and C. Lubich, “Dynamical Tensor Approximation,” *SIAM Journal on Matrix Analysis and Applications*, vol. 31, no. 5, pp. 2360–2375, 2010, doi: 10.1137/09076578X.

[52] Z. Liu, Y. Shen, V. B. Lakshminarasimhan, P. P. Liang, A. Zadeh, and L. P. Morency, “Efficient low-rank multimodal fusion with modality-specific factors,” *ACL 2018 - 56th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers)*, vol. 1, pp. 2247–2256, 2018, doi: 10.18653/v1/p18-1209.

[53] M. Hou, J. Tang, J. Zhang, W. Kong, and Q. Zhao, “Deep Multimodal Multilinear Fusion with High-order Polynomial Pooling,” in *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d Alché-Buc, E. Fox, and R. Garnett, Eds., Curran Associates, Inc., 2019. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2019/file/f56d8183992b6c54c92c16a8519a6e2b-Paper.pdf

[54] A. Fukui, D. H. Park, D. Yang, A. Rohrbach, T. Darrell, and M. Rohrbach, “Multimodal Compact Bilinear Pooling for Visual Question Answering and Visual Grounding,” Sep. 2016.

[55] Z. Liu, Y. Shen, V. B. Lakshminarasimhan, P. P. Liang, A. Zadeh, and L.-P. Morency, “Efficient Low-rank Multimodal Fusion with Modality-Specific Factors,” May 2018.

[56] A. Zadeh, M. Chen, S. Poria, E. Cambria, and L.-P. Morency, “Tensor Fusion Network for Multimodal Sentiment Analysis,” in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, M. Palmer, R. Hwa, and S. Riedel, Eds., Copenhagen, Denmark: Association for Computational Linguistics, Sep. 2017, pp. 1103–1114. doi: 10.18653/v1/D17-1115.

[57] A. Zadeh *et al.*, “Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph,” *ACL 2018 - 56th Annual Meeting of the Association for*

Computational Linguistics, Proceedings of the Conference (Long Papers), vol. 1, pp. 2236–2246, 2018, doi: 10.18653/v1/p18-1208.

[58] Y.-H. H. Tsai, P. P. Liang, A. Zadeh, L.-P. Morency, and R. Salakhutdinov, "Learning Factorized Multimodal Representations," May 2019.

[59] D. Bahdanau, K. Cho, and Y. Bengio, "Neural Machine Translation by Jointly Learning to Align and Translate," May 2016.

[60] A. Vaswani *et al.*, "Attention Is All You Need," Aug. 2023.

[61] J. Arevalo, T. Solorio, M. Montes-y-Gómez, and F. A. González, "Gated Multimodal Units for Information Fusion," Feb. 2017.

[62] M. Chen, S. Wang, P. P. Liang, T. Baltrušaitis, A. Zadeh, and L.-P. Morency, "Multimodal sentiment analysis with word-level fusion and reinforcement learning," in *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, New York, NY, USA: ACM, Nov. 2017, pp. 163–171. doi: 10.1145/3136755.3136801.

[63] S. Abdulhalim, M. Albaghdadi, and M. Farazi, "Multi-Modal Sentiment Analysis with Dynamic Attention Fusion," Sep. 2025.

[64] P. P. Liang, Z. Liu, A. Zadeh, and L.-P. Morency, "Multimodal Language Analysis with Recurrent Multistage Fusion," Aug. 2018.

[65] Y.-H. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L.-P. Morency, and R. Salakhutdinov, "Multimodal Transformer for Unaligned Multimodal Language Sequences," Jun. 2019.

[66] A. Zadeh, R. Zellers, E. Pincus, and L.-P. Morency, "MOSI: Multimodal Corpus of Sentiment Intensity and Subjectivity Analysis in Online Opinion Videos," 2016, [Online]. Available: http://arxiv.org/abs/1606.06259

[67] T. Baltrusaitis, C. Ahuja, and L.-P. Morency, "Multimodal Machine Learning: A Survey and Taxonomy," *IEEE Trans Pattern Anal Mach Intell*, vol. 41, no. 2, pp. 423–443, Feb. 2019, doi: 10.1109/TPAMI.2018.2798607.

[68] M. Chen, S. Wang, P. P. Liang, T. Baltrušaitis, A. Zadeh, and L.-P. Morency, "Multimodal Sentiment Analysis with Word-Level Fusion and Reinforcement Learning," Feb. 2018.

[69] C. Busso *et al.*, "IEMOCAP: Interactive emotional dyadic motion capture database," *Lang Resour Eval*, vol. 42, no. 4, pp. 335–359, 2008, doi: 10.1007/s10579-008-9076-6.

[70] A. Zadeh, P. P. Liang, N. Mazumder, S. Poria, E. Cambria, and L.-P. Morency, "Memory Fusion Network for Multi-view Sequential Learning," Feb. 2018.

[71] Y.-H. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L.-P. Morency, and R. Salakhutdinov, "Multimodal Transformer for Unaligned Multimodal Language Sequences," Jun. 2019.

[72] J. Lu, V. Goswami, M. Rohrbach, D. Parikh, and S. Lee, "12-in-1: Multi-Task Vision and Language Representation Learning," Apr. 2020.

[73] H. Tan and M. Bansal, "LXMERT: Learning Cross-Modality Encoder Representations from Transformers," Dec. 2019.

[74] L. H. Li, M. Yatskar, D. Yin, C.-J. Hsieh, and K.-W. Chang, "VisualBERT: A Simple and Performant Baseline for Vision and Language," Aug. 2019.

[75] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic Attribution for Deep Networks," Jun. 2017.

[76] S. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," Nov. 2017.

[77] S. Varshneya *et al.*, "Interpretable Tensor Fusion," May 2024.

[78] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, Jun. 2016, pp. 770–778. doi: 10.1109/CVPR.2016.90.

[79] M. S. Akhtar, D. S. Chauhan, D. Ghosal, S. Poria, A. Ekbal, and P. Bhattacharyya, "Multi-task Learning for Multi-modal Emotion Recognition and Sentiment Analysis," May 2019.

[80] X. Geng, H. Liu, L. Lee, D. Schuurmans, S. Levine, and P. Abbeel, "Multimodal Masked Autoencoders Learn Transferable Representations," Oct. 2022.

[81] A. Dosovitskiy *et al.*, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," Jun. 2021.

[82] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” May 2019.
[83] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, “Masked Autoencoders Are Scalable Vision Learners,” Dec. 2021.
[84] S. Changpinyo, P. Sharma, N. Ding, and R. Soricut, “Conceptual 12M: Pushing Web-Scale Image-Text Pre-Training To Recognize Long-Tail Visual Concepts,” Mar. 2021.
[85] Y. Han, G. Huang, S. Song, L. Yang, H. Wang, and Y. Wang, “Dynamic Neural Networks: A Survey,” Dec. 2021.
[86] S. Masoudnia and R. Ebrahimpour, “Mixture of experts: a literature survey,” *Artif Intell Rev*, vol. 42, no. 2, pp. 275–293, Aug. 2014, doi: 10.1007/s10462-012-9338-y.
[87] T. Bolukbasi, J. Wang, O. Dekel, and V. Saligrama, “Adaptive Neural Networks for Efficient Inference,” Sep. 2017.
[88] E. Jang, S. Gu, and B. Poole, “Categorical Reparameterization with Gumbel-Softmax,” Aug. 2017.
[89] M. Soleymani, D. Garcia, B. Jou, B. Schuller, S.-F. Chang, and M. Pantic, “A survey of multimodal sentiment analysis,” *Image Vis Comput*, vol. 65, pp. 3–14, Sep. 2017, doi: 10.1016/j.imavis.2017.08.003.
[90] A. Zadeh, M. Chen, E. Cambria, S. Poria, and L. P. Morency, “Tensor fusion network for multimodal sentiment analysis,” *EMNLP 2017 - Conference on Empirical Methods in Natural Language Processing, Proceedings*, pp. 1103–1114, 2017, doi: 10.18653/v1/d17-1115.
[91] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, “Indoor Segmentation and Support Inference from RGBD Images,” 2012, pp. 746–760. doi: 10.1007/978-3-642-33715-4_54.
[92] D. Seichter, M. Köhler, B. Lewandowski, T. Wengefeld, and H.-M. Gross, “Efficient RGB-D Semantic Segmentation for Indoor Scene Analysis,” Apr. 2021.
[93] M. A. Lee, M. Tan, Y. Zhu, and J. Bohg, “Detect, Reject, Correct: Crossmodal Compensation of Corrupted Sensors,” in *2021 IEEE International Conference on Robotics and Automation (ICRA)*, IEEE, May 2021, pp. 909–916. doi: 10.1109/ICRA48506.2021.9561847.
[94] P. P. Liang *et al.*, “MultiBench: Multiscale Benchmarks for Multimodal Representation Learning,” Nov. 2021.
[95] P. P. Liang *et al.*, “High-Modality Multimodal Transformer: Quantifying Modality & Interaction Heterogeneity for High-Modality Representation Learning,” *Transactions on Machine Learning Research*, 2023, [Online]. Available: https://openreview.net/forum?id=ttzypy3kT7
[96] A. Radford *et al.*, “Learning Transferable Visual Models From Natural Language Supervision,” Feb. 2021.
[97] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, “Multimodal deep learning,” in *Proceedings of the 28th International Conference on International Conference on Machine Learning*, in ICML’11. Madison, WI, USA: Omnipress, 2011, pp. 689–696.
[98] J.-B. Alayrac *et al.*, “Flamingo: a Visual Language Model for Few-Shot Learning,” Nov. 2022.
[99] A. Jaegle, F. Gimeno, A. Brock, A. Zisserman, O. Vinyals, and J. Carreira, “Perceiver: General Perception with Iterative Attention,” Jun. 2021.
[100] A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Lučić, and C. Schmid, “ViViT: A Video Vision Transformer,” Nov. 2021.
[101] J. Lu, D. Batra, D. Parikh, and S. Lee, “ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks,” Aug. 2019.
[102] L. Zhang and H. Hung, “Beyond F-Formations: Determining Social Involvement in Free Standing Conversing Groups from Static Images,” in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, Jun. 2016, pp. 1086–1095. doi: 10.1109/CVPR.2016.123.
[103] W. Wang, D. Tran, and M. Feiszli, “What Makes Training Multi-Modal Classification Networks Hard?,” Apr. 2020.

[104] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, “Large-Scale Video Classification with Convolutional Neural Networks,” in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, Jun. 2014, pp. 1725–1732. doi: 10.1109/CVPR.2014.223.
[105] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, and M. Paluri, “A Closer Look at Spatiotemporal Convolutions for Action Recognition,” Apr. 2018.
[106] S. Abu-El-Haija *et al.*, “YouTube-8M: A Large-Scale Video Classification Benchmark,” Sep. 2016.
[107] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov, “Improving neural networks by preventing co-adaptation of feature detectors,” Jul. 2012.
[108] I. J. Goodfellow, O. Vinyals, and A. M. Saxe, “Qualitatively characterizing neural network optimization problems,” May 2015.
[109] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet classification with deep convolutional neural networks,” *Commun ACM*, vol. 60, no. 6, pp. 84–90, May 2017, doi: 10.1145/3065386.
[110] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, “Learning Spatiotemporal Features with 3D Convolutional Networks,” Oct. 2015.
[111] J. Carreira and A. Zisserman, “Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset,” Feb. 2018.
[112] N. Wu, S. Jastrzębski, K. Cho, and K. J. Geras, “Characterizing and overcoming the greedy nature of learning in multi-modal deep neural networks,” Sep. 2022.
[113] H. R. V. Joze, A. Shaban, M. L. Iuzzolino, and K. Koishida, “MMTM: Multimodal Transfer Module for CNN Fusion,” Mar. 2020.
[114] R. Cadene, C. Dancette, H. Ben-younes, M. Cord, and D. Parikh, “RUBi: Reducing Unimodal Biases in Visual Question Answering,” Mar. 2020.
[115] I. Gat, I. Schwartz, A. Schwing, and T. Hazan, “Removing Bias in Multi-modal Classifiers: Regularization by Maximizing Functional Entropies,” Oct. 2020.
[116] B. Kim, H. Kim, K. Kim, S. Kim, and J. Kim, “Learning Not to Learn: Training Deep Neural Networks with Biased Data,” Apr. 2019.
[117] P. Molchanov, S. Gupta, K. Kim, and J. Kautz, “Hand gesture recognition with 3D convolutional neural networks,” in *2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, IEEE, Jun. 2015, pp. 1–7. doi: 10.1109/CVPRW.2015.7301342.
[118] L. Li, Z. Gan, and J. Liu, “A Closer Look at the Robustness of Vision-and-Language Pre-trained Models,” Mar. 2021.
[119] D. Saha and F. Lopez, “A Deep Learning Forecaster with Exogenous Variables for Day-Ahead Locational Marginal Price,” Oct. 2020.
[120] A. H. Phan, P. Tichavsky, and A. Cichocki, “CANDECOMP/PARAFAC Decomposition of High-order Tensors Through Tensor Reshaping,” Nov. 2012, doi: 10.1109/TSP.2013.2269046.
[121] R. A. Borsoi, K. Usevich, D. Brie, and T. Adali, “Personalized Coupled Tensor Decomposition for Multimodal Data Fusion: Uniqueness and Algorithms,” Dec. 2024, doi: 10.1109/TSP.2024.3510680.
[122] X. Mai *et al.*, “From Efficient Multimodal Models to World Models: A Survey,” Jun. 2024.
[123] T. G. Kolda and B. W. Bader, “Tensor Decompositions and Applications,” *SIAM Review*, vol. 51, no. 3, pp. 455–500, Aug. 2009, doi: 10.1137/07070111X.
[124] J. Chen *et al.*, “TDMFS: Tucker decomposition multimodal fusion model for pan-cancer survival prediction,” *Artif. Intell. Med.*, vol. 162, no. C, Apr. 2025, doi: 10.1016/j.artmed.2025.103099.
[125] T. Deng, Z. Zhang, Q. Jia, and Y. Chen, “Tucker decomposition guided shared latent label learning for multi-view incomplete multi-label classification,” *Information Fusion*, vol. 126, p. 103673, 2026, doi: https://doi.org/10.1016/j.inffus.2025.103673.

[126] H. Ben-younes, R. Cadene, M. Cord, and N. Thome, “MUTAN: Multimodal Tucker Fusion for Visual Question Answering,” May 2017.
[127] F. Liu, J. Chen, W. Tan, and C. Cai, “A Multi-Modal Fusion Method Based on Higher-Order Orthogonal Iteration Decomposition,” *Entropy*, vol. 23, no. 10, p. 1349, Oct. 2021, doi: 10.3390/e23101349.
[128] I. V. Oseledets, “Tensor-train decomposition,” *SIAM Journal on Scientific Computing*, vol. 33, no. 5, pp. 2295–2317, 2011, doi: 10.1137/090752286.
[129] D. Bacciu and D. P. Mandic, “Tensor Decompositions in Deep Learning,” Feb. 2020.
[130] X. Zhang, E. Kofidis, C. Zhu, L. Zhang, and Y. Liu, “Federated Learning Using Coupled Tensor Train Decomposition,” Mar. 2024.
[131] Q. Zhao, G. Zhou, S. Xie, L. Zhang, and A. Cichocki, “Tensor Ring Decomposition,” Jun. 2016.
[132] Y. Yu and H. Li, “Practical alternating least squares for tensor ring decomposition,” *Numer Linear Algebra Appl*, vol. 31, no. 3, May 2024, doi: 10.1002/nla.2542.
[133] Стрижов, “Сингулярное Разложение,” pp. 1–6.
[134] L. De Lathauwer, B. De Moor, and J. Vandewalle, “A Multilinear Singular Value Decomposition,” *SIAM Journal on Matrix Analysis and Applications*, vol. 21, no. 4, pp. 1253–1278, Jan. 2000, doi: 10.1137/S0895479896305696.
[135] LR Tucker and L Tucker, “The extension of factor analysis to three-dimensional matrices,” *Contributions to Mathematical Psychology*, 1964.
[136] Y. Zniyed, R. Boyer, A. L. F. de Almeida, and G. Favier, “A TT-Based Hierarchical Framework for Decomposing High-Order Tensors,” *SIAM Journal on Scientific Computing*, vol. 42, no. 2, pp. A822–A848, Jan. 2020, doi: 10.1137/18M1229973.
[137] M. E. Kilmer, K. Braman, N. Hao, and R. C. Hoover, “Third-Order Tensors as Operators on Matrices: A Theoretical and Computational Framework with Applications in Imaging,” *SIAM Journal on Matrix Analysis and Applications*, vol. 34, no. 1, pp. 148–172, Jan. 2013, doi: 10.1137/110837711.
[138] A.-H. Phan, A. Cichocki, P. Tichavsky, G. Luta, and A. Brockmeier, “Tensor completion throughmultiple Kronecker product decomposition,” in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, IEEE, May 2013, pp. 3233–3237. doi: 10.1109/ICASSP.2013.6638255.
[139] A. H. Phan, A. Cichocki, P. Tichavsky, R. Zdunek, and S. Lehky, “From basis components to complex structural patterns,” in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, IEEE, May 2013, pp. 3228–3232. doi: 10.1109/ICASSP.2013.6638254.
[140] A. L. F. de Almeida, Gé. Favier, and J. C. M. Mota, “A Constrained Factor Decomposition With Application to MIMO Antenna Systems,” *IEEE Transactions on Signal Processing*, vol. 56, no. 6, pp. 2429–2442, Jun. 2008, doi: 10.1109/TSP.2008.917026.
[141] M. G. Asante-Mensah, S. Ahmadi-Asl, and A. Cichocki, “Matrix and tensor completion using tensor ring decomposition with sparse representation,” *Mach Learn Sci Technol*, vol. 2, no. 3, p. 035008, Sep. 2021, doi: 10.1088/2632-2153/abcb4f.
[142] A. de Almeida, G. Favier, J. Mota, and J. da Costa, “Overview of Tensor Decompositions with Applications to Communications,” in *Signals and Images*, CRC Press, 2015, pp. 325–355. doi: 10.1201/b19385-17.
[143] A. L. F. de Almeida, G. Favier, and J. C. M. Mota, “PARAFAC-based unified tensor modeling for wireless communication systems with application to blind multiuser equalization,” *Signal Processing*, vol. 87, no. 2, pp. 337–351, Feb. 2007, doi: 10.1016/j.sigpro.2005.12.014.
[144] S. Ahmadi-Asl *et al.*, “Randomized Algorithms for Computation of Tucker decomposition and Higher Order SVD (HOSVD),” Dec. 2021.
[145] S. Ahmadi-Asl *et al.*, “Randomized algorithms for fast computation of low rank tensor ring model,” *Mach Learn Sci Technol*, vol. 2, no. 1, p. 011001, Dec. 2020, doi: 10.1088/2632-2153/abad87.
[146] E. Frolov and I. Oseledets, “Tensor methods and recommender systems,” *WIREs Data Mining and Knowledge Discovery*, vol. 7, no. 3, May 2017, doi: 10.1002/widm.1201.

[147] Z. Zhang and S. Aeron, “Exact tensor completion using t-SVD,” Feb. 2015.
[148] M. E. Kilmer and C. D. Martin, “Factorization strategies for third-order tensors,” *Linear Algebra Appl*, vol. 435, no. 3, pp. 641–658, Aug. 2011, doi: 10.1016/j.laa.2010.09.020.
[149] S. Ahmadi-Asl, A.-H. Phan, and A. Cichocki, “A Randomized Algorithm for Tensor Singular Value Decomposition using an Arbitrary Number of Passes,” Feb. 2025.
[150] L. Qi and G. Yu, “T-Singular Values and T-Sketching for Third Order Tensors,” Mar. 2021.
[151] C. Lu, J. Feng, Y. Chen, W. Liu, Z. Lin, and S. Yan, “Tensor Robust Principal Component Analysis with a New Tensor Nuclear Norm,” *IEEE Trans Pattern Anal Mach Intell*, vol. 42, no. 4, pp. 925–938, Apr. 2020, doi: 10.1109/TPAMI.2019.2891760.
[152] J. A. Tropp, A. Yurtsever, M. Udell, and V. Cevher, “Practical Sketching Algorithms for Low-Rank Matrix Approximation,” *SIAM Journal on Matrix Analysis and Applications*, vol. 38, no. 4, pp. 1454–1485, Jan. 2017, doi: 10.1137/17M1111590.
[153] E. K. Bjarkason, “Pass-Efficient Randomized Algorithms for Low-Rank Matrix Approximation Using Any Number of Views,” *SIAM Journal on Scientific Computing*, vol. 41, no. 4, pp. A2355–A2383, Jan. 2019, doi: 10.1137/18M118966X.
[154] D. A. Tarzanagh and G. Michailidis, “Fast Randomized Algorithms for t-Product Based Tensor Operations and Decompositions with Applications to Imaging Data,” *SIAM J Imaging Sci*, vol. 11, no. 4, pp. 2629–2664, Jan. 2018, doi: 10.1137/17M1159932.
[155] S. Ahmadi-Asl, N. Rezaeian, C. F. Caiafa, and A. L. F. de Almeidad, “Efficient Algorithms for Low Tubal Rank Tensor Approximation with Applications,” May 2025.
[156] S. Ahmadi-Asl, R. V. Garaev, R. A. Lukmanov, N. Rezaeian, A. Masood Khattak, and M. Mazzara, “Efficient Smooth Tensor Train and Tensor Ring Completion for Image Classification Enhancement,” *IEEE Access*, vol. 13, pp. 189686–189701, 2025, doi: 10.1109/ACCESS.2025.3625862.
[157] S. Ahmadi-Asl, N. Rezaeian, and U. O. Ugwu, “Randomized Algorithms for Computing the Generalized Tensor SVD Based on the Tensor Product,” *Communications on Applied Mathematics and Computation*, Jan. 2025, doi: 10.1007/s42967-024-00461-3.
[158] M. F. Kaloorazi and R. C. de Lamare, “Subspace-Orbit Randomized Decomposition for Low-Rank Matrix Approximations,” *IEEE Transactions on Signal Processing*, vol. 66, no. 16, pp. 4409–4424, Aug. 2018, doi: 10.1109/TSP.2018.2853137.
[159] S. Ahmadi-Asl, “An Efficient Randomized Fixed-Precision Algorithm for Tensor Singular Value Decomposition,” Apr. 2024.
[160] X. Ding, W. Yu, Y. Xie, and S. Liu, “Efficient Model-Based Collaborative Filtering with Fast Adaptive PCA,” Sep. 2020, doi: 10.1109/ICTAI50040.2020.00149.
[161] X. Feng and W. Yu, “A Fast Adaptive Randomized PCA Algorithm,” in *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, California: International Joint Conferences on Artificial Intelligence Organization, Aug. 2023, pp. 3695–3704. doi: 10.24963/ijcai.2023/411.
[162] J. Zhang, A. K. Saibaba, M. Kilmer, and S. Aeron, “A Randomized Tensor Singular Value Decomposition based on the t-product,” Sep. 2016.
[163] S. Park, H. S. Shim, M. Chatterjee, K. Sagae, and L.-P. Morency, “Computational Analysis of Persuasiveness in Social Multimedia,” pp. 50–57, 2014, doi: 10.1145/2663204.2663260.
[164] C. D. M. Jeffrey Pennington, Richard Socher, “GloVe: Global Vectors for Word Representation,” *AES: Journal of the Audio Engineering Society*, vol. 19, no. 5, pp. 417–425, 1971.
[165] N. Rezaeian and G. M. Novikova, “Detecting Near-duplicates in Russian Documents through Using Fingerprint Algorithm Simhash,” in *Procedia Computer Science*, Elsevier B.V., 2017, pp. 421–425. doi: 10.1016/j.procs.2017.01.006.
[166] N. Rezaeian and G. Novikova, “Automatic Text Summarization In Persian Language,” in *ITTMM 2016*, RU, Apr. 2016.
[167] S. Hochreiter, “Lstm Can Solve Hard Lo G Time Lag Problems Trivial Previous Long Time Lag Problems,” *Adv Neural Inf Process Syst*, pp. 473–479, 1997.

[168] P. Ekman, “An Argument for Basic Emotions,” *Cogn Emot*, vol. 6, no. 3–4, pp. 169–200, 1992, doi: 10.1080/02699939208411068.
[169] W. V. F. Paul Ekman, “Facialsigns of Emotional Exprience,” 1978. doi: 10.2307/134547.
[170] T. Baltrusaitis, P. Robinson, and L. P. Morency, “OpenFace: An open source facial behavior analysis toolkit,” *2016 IEEE Winter Conference on Applications of Computer Vision, WACV 2016*, 2016, doi: 10.1109/WACV.2016.7477553.
[171] G. Degottex, J. Kane, T. Drugman, T. Raitio, and S. Scherer, “COVAREP – A COLLABORATIVE VOICE ANALYSIS REPOSITORY FOR SPEECH TECHNOLOGIES Computer Science Department, University of Crete, Heraklion, Greece Phonetics and Speech Laboratory, Trinity College Dublin, Ireland TCTS Lab - University of Mons, Belgium A,” *IEEE International Conference on Acoustic, Speech and Signal Processing*, pp. 960–964, 2014.
[172] T. Drugman and A. Alwan, “Joint robust voicing detection and pitch estimation based on residual harmonics,” *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, pp. 1973–1976, 2011.
[173] P. Alku, “Glottal wave analysis with Pitch Synchronous Iterative Adaptive Inverse Filtering,” *Speech Commun*, vol. 11, no. 2–3, pp. 109–118, 1992, doi: 10.1016/0167-6393(92)90005-R.
[174] P. Alku, T. Bäckström, and E. Vilkman, “Normalized amplitude quotient for parametrization of the glottal flow,” *J Acoust Soc Am*, vol. 112, no. 2, pp. 701–710, 2002, doi: 10.1121/1.1490365.
[175] I. R. Titze and J. Sundberg, “Vocal intensity in speakers and singers.,” *J Acoust Soc Am*, vol. 90, no. 4, pp. 2351–2351, 1991, doi: 10.1121/1.402160.
[176] J. Kane and C. Gobl, “Wavelet maxima dispersion for breathy to tense voice discrimination,” *IEEE Trans Audio Speech Lang Process*, vol. 21, no. 6, pp. 1170–1179, 2013, doi: 10.1109/TASL.2013.2245653.
[177] R. Veldhuis, “A computationally efficient alternative for the Liljencrants–Fant model and its perceptual evaluation,” *J Acoust Soc Am*, vol. 103, no. 1, pp. 566–571, 1998, doi: 10.1121/1.421103.
[178] S. Ghosh, E. Laksana, L. P. Morency, and S. Scherer, “Representation learning for speech emotion recognition,” *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, vol. 08-12-September-2016, no. October, pp. 3603–3607, 2016, doi: 10.21437/Interspeech.2016-692.

ПРИЛОЖЕНИЕ

core\config.py

```python
from dataclasses import dataclass, field
from typing import Optional, Dict, Callable, Literal, Any
import torch

@dataclass
class FusionConfig:
    """Конфигурация обучения и логирования."""
    epochs: int = 100
    lr: float = 1e-3
    batch_size: int = 64
    weight_decay: float = 0.0
    seed: Optional[int] = 42
    num_workers: int = 4
    device: Optional[str] = None               # "cuda" / "cpu" / None
    enable_json_logging: bool = True
    json_log_path: str = "runs/log.json"
    checkpointing: bool = True
    checkpoint_path: str = "runs/best.ckpt"
    monitor_metric: str = "val_loss"
    monitor_mode: Literal["min", "max"] = "min"
    gradient_clip_norm: Optional[float] = None
    early_stopping_patience: Optional[int] = None
    metrics: Dict[str, Callable[[Any, Any], float]] = field(default_factory=dict)
    loss_fn: Callable = torch.nn.MSELoss
    dataset_name: Optional[str] = None
    ranks: Optional[Dict[str, Any]] = None    # передаётся в модель
```

core\fusion.py

```python
import torch
import torch.nn as nn
import torch.optim as optim
from torch.utils.data import DataLoader
import time
import uuid
from datetime import datetime
import numpy as np
from typing import Dict, Optional, Tuple, Literal, Any, Callable
from .config import FusionConfig
from .utils import seed_all, get_device, count_params
from .logging import save_json_log
from ..models.lmf import LMF
from ..models.mrrf_tucker import MRRF_Tucker
from ..models.tt_mrrf import TT_MRRF
from ..models.tr_mrrf import TR_MRRF
import os

class TensorMultimodalFusion:
    """Оркестратор мультимодального слияния (LMF, MRRF-Tucker, TT-MRRF, TR-MRRF)."""

    def __init__(
        self,
        algorithm: Literal["LMF", "MRRF_TUCKER", "TT_MRRF", "TR_MRRF"],
        input_shapes: Dict[str, Tuple[int, ...]],
        output_dim: int = 1,
        ranks: Optional[Dict[str, Any]] = None,
        config: Optional[FusionConfig] = None,
        device: Optional[torch.device] = None,
    ):
        self.algorithm = algorithm.upper()
        self.input_shapes = input_shapes
        self.output_dim = output_dim
        self.config = config or FusionConfig()
        self.device = device or get_device(self.config.device)

        self.modalities = list(input_shapes.keys())

        if set(self.modalities) != {"text", "visual", "audio"}:
            raise ValueError("Требуются модальности: text, visual, audio.")

        ranks = ranks or self.config.ranks or {}

        # ---------- LMF ----------
        if self.algorithm == "LMF":
            rank = ranks.get("rank", 4) if isinstance(ranks, dict) else (ranks or 4)
            if not isinstance(rank, int) or rank <= 0:
                raise ValueError("LMF: rank должен быть положительным целым.")
            self.model = LMF(input_shapes, rank=rank, output_dim=output_dim)

        # ---------- MRRF_Tucker ----------
```

```
        elif self.algorithm == "MRRF_TUCKER":
            req = {"text", "visual", "audio"}
            if not isinstance(ranks, dict) or set(ranks.keys()) != req:
                raise ValueError(f"MRRF_TUCKER требует ranks: {req}")
            self.model = MRRF_Tucker(input_shapes,
                rank_t=ranks["text"], rank_v=ranks["visual"], rank_a=ranks["audio"])

        # ---------- TT_MRRF ----------
        elif self.algorithm == "TT_MRRF":
            req = {"text", "visual", "audio"}
            if not isinstance(ranks, dict) or set(ranks.keys()) != req:
                raise ValueError(f"TT_MRRF требует ranks: {req}")
            self.model = TT_MRRF(input_shapes,
                tt_rank_t=ranks["text"], tt_rank_v=ranks["visual"],
tt_rank_a=ranks["audio"])

        # ---------- TR_MRRF ----------
        elif self.algorithm == "TR_MRRF":
            req = {"text", "visual", "audio"}
            if not isinstance(ranks, dict) or set(ranks.keys()) != req:
                raise ValueError(f"TR_MRRF требует ranks: {req}")
            self.model = TR_MRRF(input_shapes,
                tr_rank_1=ranks["text"], tr_rank_2=ranks["visual"],
tr_rank_3=ranks["audio"])

        else:
            raise ValueError(f"Неизвестный алгоритм: {algorithm}")

        self.ranks = ranks  # для логов

    # ------------------- fit -------------------
    def fit(self, train_loader: DataLoader, val_loader: Optional[DataLoader] = None):
        seed_all(self.config.seed)
        self.model.to(self.device)
        optimizer = optim.Adam(self.model.parameters(), lr=self.config.lr,
weight_decay=self.config.weight_decay)
        scheduler = optim.lr_scheduler.ReduceLROnPlateau(optimizer,
mode=self.config.monitor_mode, patience=5, factor=0.5)
        loss_fn = self.config.loss_fn()

        best_val = float('inf') if self.config.monitor_mode == "min" else float('-inf')
        best_epoch = 0
        stale = 0
        log_entries = []
        param_count = count_params(self.model)

        for epoch in range(1, self.config.epochs + 1):
            start = time.perf_counter()
            self.model.train()
            train_loss = 0.0
            n_batch = 0
            for batch in train_loader:
                mods = {k: batch[k].to(self.device) for k in self.modalities}
                labels = batch["label"].to(self.device)
```

```
                optimizer.zero_grad()
                #out = self.model(mods)
                out = self.model(mods["text"], mods["visual"], mods["audio"])
                loss = loss_fn(out, labels)
                if not torch.isfinite(loss):
                    continue
                loss.backward()
                if self.config.gradient_clip_norm:
                    torch.nn.utils.clip_grad_norm_(self.model.parameters(),
self.config.gradient_clip_norm)
                optimizer.step()
                train_loss += loss.item()
                n_batch += 1
            train_loss = train_loss / n_batch if n_batch else float('nan')
            epoch_time = time.perf_counter() - start

            train_metrics = {"loss": train_loss}
            val_metrics = {}
            if val_loader:
                val_metrics = self._compute_metrics(val_loader, loss_fn,
self.config.metrics)
                mon = val_metrics.get(self.config.monitor_metric.split("_", 1)[1] if "_"
in self.config.monitor_metric else self.config.monitor_metric)
                if mon is None:
                    raise ValueError(f"Метрика {self.config.monitor_metric} не найдена.")
                scheduler.step(mon)
                improve = (self.config.monitor_mode == "min" and mon < best_val) or \
                          (self.config.monitor_mode == "max" and mon > best_val)
                if improve:
                    best_val, best_epoch = mon, epoch
                    if self.config.checkpointing:
                        torch.save({
                            'model_state_dict': self.model.state_dict(),
                            'best_metric': best_val,
                            'best_epoch': best_epoch,
                            'monitor_metric': self.config.monitor_metric
                        }, self.config.checkpoint_path)
                        print(f"Сохранён чекпоинт: {self.config.checkpoint_path} |
{self.config.monitor_metric}: {best_val:.4f}")
                    stale = 0
                else:
                    stale += 1
                    if self.config.early_stopping_patience and stale >=
self.config.early_stopping_patience:
                        print(f"Early stop на эпохе {epoch}")
                        break

            print(f"Эпоха {epoch:3d} | время={epoch_time:.2f}s | "
                  f"парам={param_count:,} | train_loss={train_loss:.4f} | "
                  f"val={val_metrics}")

            entry = {
                "epoch": epoch,
                "time_s": epoch_time,
```

```
                "param_count": param_count,
                "train": train_metrics,
                "val": val_metrics,
                "best_so_far": {"metric": self.config.monitor_metric, "value": best_val,
"epoch": best_epoch}
            }
            log_entries.append(entry)

        if self.config.enable_json_logging:
            self._save_log(log_entries)

        return {"log": log_entries, "best_epoch": best_epoch, "best_value": best_val}

    # ------------------- evaluate / predict -------------------
    def evaluate_old(self, loader: DataLoader) -> Dict[str, float]:
        loss_fn = self.config.loss_fn()
        return self._compute_metrics(loader, loss_fn, self.config.metrics)
    def evaluate(self, test_loader: DataLoader) -> Dict[str, float]:
        # --- ЗАГРУЖАЕМ ЛУЧШУЮ МОДЕЛЬ ---
        if self.config.checkpointing and os.path.exists(self.config.checkpoint_path):
            print(f"Загружаем лучшую модель: {self.config.checkpoint_path}")
            checkpoint = torch.load(self.config.checkpoint_path, map_location=self.device)
            self.model.load_state_dict(checkpoint['model_state_dict'])
            print(f"Лучшая val_mae: {checkpoint.get('best_metric', 'N/A'):.4f}")
        else:
            print("Чекпоинт не найден. Используется последняя модель.")

        self.model.eval()
        total_loss = 0.0
        n = 0
        all_p, all_l = [], []
        loss_fn = nn.MSELoss()

        with torch.no_grad():
            for batch in test_loader:
                text = batch["text"].to(self.device)
                visual = batch["visual"].to(self.device)
                audio = batch["audio"].to(self.device)
                labels = batch["label"].to(self.device)

                out = self.model(text, visual, audio)
                loss = loss_fn(out, labels)
                total_loss += loss.item() * labels.size(0)
                n += labels.size(0)
                all_p.append(out.cpu().numpy())
                all_l.append(labels.cpu().numpy())

        preds = np.concatenate(all_p)
        labels = np.concatenate(all_l)

        metrics = {"test_loss": total_loss / n}
        for name, fn in self.config.metrics.items():
            try:
                metrics[f"test_{name}"] = fn(preds, labels)
```

```
            except:
                metrics[f"test_{name}"] = float('nan')

        return metrics

    def predict(self, loader: DataLoader):
        self.model.eval()
        preds = []
        with torch.no_grad():
            for batch in loader:
                text = batch["text"].to(self.device)
                visual = batch["visual"].to(self.device)
                audio = batch["audio"].to(self.device)

                out = self.model(text, visual, audio)
                preds.append(out.cpu())
        return preds

    # ------------------- utils -------------------
    def get_param_count(self) -> int:
        return count_params(self.model)

    def summary(self) -> str:
        return str(self.model)

    def _compute_metrics(self, loader, loss_fn, extra_metrics):
        self.model.eval()
        total_loss = 0.0
        n = 0
        all_p, all_l = [], []
        with torch.no_grad():
            for batch in loader:
                text = batch["text"].to(self.device)
                visual = batch["visual"].to(self.device)
                audio = batch["audio"].to(self.device)

                labels = batch["label"].to(self.device)
                #out = self.model(mods)
                out = self.model(text, visual, audio)
                loss = loss_fn(out, labels)
                total_loss += loss.item() * labels.size(0)
                n += labels.size(0)
                all_p.append(out.cpu().numpy())
                all_l.append(labels.cpu().numpy())
        if n == 0: return {}
        total_loss /= n
        p = np.concatenate(all_p).flatten()
        l = np.concatenate(all_l).flatten()
        res = {"loss": total_loss}
        for name, fn in extra_metrics.items():
            try:
                res[name] = fn(p, l)
            except Exception:
                res[name] = float('nan')
```

```
        return res

    def _save_log(self, entries):
        meta = {
            "run_id": str(uuid.uuid4()),
            "timestamp": datetime.utcnow().isoformat() + "Z",
            "algorithm": self.algorithm,
            "dataset": self.config.dataset_name or "Unknown",
            "seed": self.config.seed,
            "device": str(self.device),
            "input_shapes": {k: list(v) for k, v in self.input_shapes.items()},
            "output_dim": self.output_dim
        }
        hparams = {k: v for k, v in self.config.__dict__.items()
                   if k not in {"metrics", "loss_fn"}}
        hparams["ranks"] = self.ranks
        save_json_log(entries, meta, hparams, self.config.json_log_path)
```

core\utils.py

```python
import torch
import numpy as np
import random
from typing import Optional

def seed_all(seed: Optional[int]) -> None:
    """Установка семян для воспроизводимости."""
    if seed is not None:
        torch.manual_seed(seed)
        np.random.seed(seed)
        random.seed(seed)
        torch.backends.cudnn.deterministic = True
        torch.backends.cudnn.benchmark = False

def get_device(device_str: Optional[str]) -> torch.device:
    """Получение устройства."""
    if device_str is None:
        return torch.device("cuda" if torch.cuda.is_available() else "cpu")
    return torch.device(device_str)

def count_params(model: torch.nn.Module) -> int:
    """Количество обучаемых параметров."""
    return sum(p.numel() for p in model.parameters() if p.requires_grad)
```

data\datamodule.py

```python
import numpy as np
import torch  # ← ДОБАВЛЕНО!
from torch.utils.data import Dataset, DataLoader
from sklearn.preprocessing import StandardScaler
from typing import Optional, Any

def clean_feature(arr: np.ndarray) -> np.ndarray:
    """Замена NaN, inf на 0.0."""
    return np.nan_to_num(arr, nan=0.0, posinf=0.0, neginf=0.0)

class MOSI_Wrapper(Dataset):
    def __init__(self, dataset: Any, fit: bool = False):
        self.dataset = dataset
        self.fit = fit
        self.scalers: dict = {}
        self.label_min = None
        self.label_max = None

        texts, visuals, audios, labels = [], [], [], []
        for item in dataset:
            texts.append(clean_feature(item["glove"]))
            visuals.append(clean_feature(item["facet"]))
            audios.append(clean_feature(item["covarep"]))
            # Исправлено: проверка на torch.Tensor
            label = item["label"]
            if isinstance(label, torch.Tensor):
                label = label.item()
            labels.append(float(label))

        self.texts = np.stack(texts)
        self.visuals = np.stack(visuals)
        self.audios = np.stack(audios)
        self.labels = np.array(labels, dtype=np.float32)

        if fit:
            # Нормализация признаков
            flat_text = self.texts.reshape(-1, 300)
            self.scalers["text"] = StandardScaler().fit(clean_feature(flat_text))
            self.scalers["visual"] = StandardScaler().fit(clean_feature(self.visuals))
            self.scalers["audio"] = StandardScaler().fit(clean_feature(self.audios))

            # Нормализация меток в [-1, 1]
            self.label_min = float(self.labels.min())
            self.label_max = float(self.labels.max())

    def __len__(self) -> int:
        return len(self.texts)

    def __getitem__(self, idx: int) -> dict:
        t = self.texts[idx].copy()
        v = self.visuals[idx].copy()
```

```
        a = self.audios[idx].copy()

        # Применение скейлеров, если они есть
        if "text" in self.scalers:
            t_flat = clean_feature(t).reshape(-1, 300)
            t = self.scalers["text"].transform(t_flat).reshape(t.shape)
            v = self.scalers["visual"].transform(clean_feature(v).reshape(1, -1)).squeeze(0)
            a = self.scalers["audio"].transform(clean_feature(a).reshape(1, -1)).squeeze(0)

        label = self.labels[idx]
        if self.label_min is not None and self.label_max is not None:
            denom = self.label_max - self.label_min + 1e-8
            label = 2 * (label - self.label_min) / denom - 1

        return {
            "text": torch.tensor(t, dtype=torch.float32),
            "visual": torch.tensor(v, dtype=torch.float32),
            "audio": torch.tensor(a, dtype=torch.float32),
            "label": torch.tensor(label, dtype=torch.float32)
        }

class DataModule:
    def __init__(
        self,
        train_ds: Any,
        val_ds: Optional[Any] = None,
        test_ds: Optional[Any] = None,
        batch_size: int = 32,
        num_workers: int = 0,  # ← ИСПРАВЛЕНО: 0 по умолчанию
    ):
        self.train = MOSI_Wrapper(train_ds, fit=True)
        self.val = MOSI_Wrapper(val_ds) if val_ds else None
        self.test = MOSI_Wrapper(test_ds) if test_ds else None

        if self.val:
            self.val.scalers = self.train.scalers
            self.val.label_min = self.train.label_min
            self.val.label_max = self.train.label_max
        if self.test:
            self.test.scalers = self.train.scalers
            self.test.label_min = self.train.label_min
            self.test.label_max = self.train.label_max

        self.bs = batch_size
        self.nw = num_workers  # ← теперь можно переопределить

    def train_loader(self) -> DataLoader:
        return DataLoader(
            self.train,
            batch_size=self.bs,
            shuffle=True,
            num_workers=self.nw,
```

```
            pin_memory=True,
            persistent_workers=False  # ← защита от утечек
        )

    def val_loader(self) -> Optional[DataLoader]:
        if not self.val:
            return None
        return DataLoader(
            self.val,
            batch_size=self.bs,
            shuffle=False,
            num_workers=self.nw,
            pin_memory=True,
            persistent_workers=False
        )

    def test_loader(self) -> Optional[DataLoader]:
        if not self.test:
            return None
        return DataLoader(
            self.test,
            batch_size=self.bs,
            shuffle=False,
            num_workers=self.nw,
            pin_memory=True,
            persistent_workers=False
        )
```

models\lmf.py

```python
import torch
import torch.nn as nn
from typing import Dict, Tuple

class LMF(nn.Module):
    def __init__(self, input_shapes: Dict[str, Tuple[int, ...]], rank: int = 4,
                 hidden_dim: int = 128, output_dim: int = 1):
        super().__init__()
        t_dim = input_shapes["text"][-1]
        v_dim = input_shapes["visual"][0]
        a_dim = input_shapes["audio"][0]

        self.text_lstm = nn.LSTM(t_dim, hidden_dim, batch_first=True, dropout=0.3)
        self.visual_net = nn.Sequential(nn.Linear(v_dim, hidden_dim), nn.ReLU(), nn.Dropout(0.3))
        self.audio_net  = nn.Sequential(nn.Linear(a_dim, hidden_dim), nn.ReLU(), nn.Dropout(0.3))

        self.t_factor = nn.Parameter(torch.Tensor(rank, hidden_dim + 1))
        self.v_factor = nn.Parameter(torch.Tensor(rank, hidden_dim + 1))
        self.a_factor = nn.Parameter(torch.Tensor(rank, hidden_dim + 1))
        self.fusion_w = nn.Parameter(torch.Tensor(1, rank))
        self.fusion_b = nn.Parameter(torch.Tensor(1, 1))
        self.proj = nn.Linear(1, output_dim)

        for p in (self.t_factor, self.v_factor, self.a_factor):
            nn.init.xavier_uniform_(p)
            nn.init.uniform_(self.fusion_w, 0.1, 0.5)
            nn.init.zeros_(self.fusion_b)
            nn.init.xavier_uniform_(self.proj.weight)
            nn.init.zeros_(self.proj.bias)

    def forward(self, text, visual, audio):
        B = text.size(0)
        _, (txt_h, _) = self.text_lstm(text)
        txt_h = txt_h.squeeze(0)
        vis_h = self.visual_net(visual)
        aud_h = self.audio_net(audio)

        ones = torch.ones(B, 1, device=text.device)
        t = torch.cat([txt_h, ones], dim=1)
        v = torch.cat([vis_h, ones], dim=1)
        a = torch.cat([aud_h, ones], dim=1)

        z_t = t @ self.t_factor.t()
        z_v = v @ self.v_factor.t()
        z_a = a @ self.a_factor.t()
        z = z_t * z_v * z_a
        out = (self.fusion_w * z).sum(dim=1, keepdim=True) + self.fusion_b
        out = self.proj(out)
        return out.squeeze(-1)

```

models\tr_mrrf.py

```python
import torch
import torch.nn as nn
from typing import Dict, Tuple

class TR_MRRF(nn.Module):
    def __init__(self, input_shapes: Dict[str, Tuple[int, ...]],
                 tr_rank_1: int = 4, tr_rank_2: int = 4, tr_rank_3: int = 4,
                 hidden_dim: int = 128):
        super().__init__()
        text_dim = input_shapes["text"][-1]
        visual_dim = input_shapes["visual"][0]
        audio_dim = input_shapes["audio"][0]

        self.text_lstm = nn.LSTM(text_dim, hidden_dim, batch_first=True, num_layers=2, dropout=0.3)
        self.visual_net = nn.Sequential(nn.Linear(visual_dim, hidden_dim), nn.ReLU(), nn.Dropout(0.3))
        self.audio_net  = nn.Sequential(nn.Linear(audio_dim, hidden_dim), nn.ReLU(), nn.Dropout(0.3))

        self.d = hidden_dim + 1
        self.r1, self.r2, self.r3 = tr_rank_1, tr_rank_2, tr_rank_3

        self.G1 = nn.Parameter(torch.empty(self.r1, self.d, self.r2))
        self.G2 = nn.Parameter(torch.empty(self.r2, self.d, self.r3))
        self.G3 = nn.Parameter(torch.empty(self.r3, self.d, self.r1))

        self.output_proj = nn.Linear(1, 1)

        nn.init.normal_(self.G1, mean=0.0, std=1e-3)
        nn.init.normal_(self.G2, mean=0.0, std=1e-3)
        nn.init.normal_(self.G3, mean=0.0, std=1e-3)
        nn.init.xavier_uniform_(self.output_proj.weight)
        nn.init.zeros_(self.output_proj.bias)

    def forward(self, text, visual, audio):
        B = text.size(0)

        _, (txt_h, _) = self.text_lstm(text)
        txt_h = txt_h[-1]

        vis_h = self.visual_net(visual)
        aud_h = self.audio_net(audio)

        ones = torch.ones(B, 1, device=text.device)
        t = torch.cat([txt_h, ones], dim=1)
        v = torch.cat([vis_h, ones], dim=1)
        a = torch.cat([aud_h, ones], dim=1)

        x = torch.einsum('idj,bd->bij', self.G1, t)
        x = torch.einsum('bij,jdk,bd->bik', x, self.G2, v)
        fusion = torch.einsum('bik,kdi,bd->b', x, self.G3, a)
```

```
50
51          out = self.output_proj(fusion.unsqueeze(1))
52          return out.squeeze(-1)
```

models\tt_mrrf.py

```python
import torch
import torch.nn as nn
from typing import Dict, Tuple

class TT_MRRF(nn.Module):
    def __init__(self, input_shapes: Dict[str, Tuple[int, ...]],
                 tt_rank_t: int = 4, tt_rank_v: int = 4, tt_rank_a: int = 4,
                 hidden_dim: int = 128):
        super().__init__()
        text_dim = input_shapes["text"][-1]
        visual_dim = input_shapes["visual"][0]
        audio_dim = input_shapes["audio"][0]

        self.text_lstm = nn.LSTM(text_dim, hidden_dim, batch_first=True, num_layers=2, dropout=0.3)
        self.visual_net = nn.Sequential(nn.Linear(visual_dim, hidden_dim), nn.ReLU(), nn.Dropout(0.3))
        self.audio_net  = nn.Sequential(nn.Linear(audio_dim, hidden_dim), nn.ReLU(), nn.Dropout(0.3))

        self.d = hidden_dim + 1  # 129
        self.r1, self.r2, self.r3 = tt_rank_t, tt_rank_v, tt_rank_a

        # TT-cores
        self.G1 = nn.Parameter(torch.empty(1, self.r1, self.d))        # (1, r1, d)
        self.G2 = nn.Parameter(torch.empty(self.r1, self.r2, self.d))  # (r1, r2, d)
        self.G3 = nn.Parameter(torch.empty(self.r2, self.r3, self.d))  # (r2, r3, d)
        self.G4 = nn.Parameter(torch.empty(self.r3, 1))                # (r3, 1)

        self.output_proj = nn.Linear(1, 1)

        # Инициализация
        for p in (self.G1, self.G2, self.G3):
            nn.init.normal_(p, 0.0, 1e-3)
        nn.init.normal_(self.G4, 0.0, 1e-3)
        nn.init.xavier_uniform_(self.output_proj.weight)
        nn.init.zeros_(self.output_proj.bias)

    def forward(self, text, visual, audio):
        B = text.size(0)

        # --- Экстракция признаков ---
        _, (txt_h, _) = self.text_lstm(text)
        txt_h = txt_h[-1]  # (B, 128)
        vis_h = self.visual_net(visual)
        aud_h = self.audio_net(audio)

        # --- Bias ---
        ones = torch.ones(B, 1, device=text.device)
        t = torch.cat([txt_h, ones], dim=1)  # (B, 129)
        v = torch.cat([vis_h, ones], dim=1)
        a = torch.cat([aud_h, ones], dim=1)
```

```
50
51          # --- TT-contraction ---
52          # 1) (1, r1, d) × (B, d) → (B, r1)
53          x = torch.einsum('ord,bd->br', self.G1, t)          # (B, r1)
54
55          # 2) (B, r1) × (r1, r2, d) × (B, d) → (B, r2)
56          x = torch.einsum('br,rxd,bd->bx', x, self.G2, v)    # (B, r2)
57
58          # 3) (B, r2) × (r2, r3, d) × (B, d) → (B, r3)
59          x = torch.einsum('bx,xyd,bd->by', x, self.G3, a)    # (B, r3)
60
61          # 4) (B, r3) × (r3, 1) → (B,)
62          fusion = torch.einsum('by,yr->b', x, self.G4)        # (B,)
63
64          out = self.output_proj(fusion.unsqueeze(1))
65          return out.squeeze(-1)
```