Федеральное государственное автономное образовательное учреждение высшего образования
«Российский университет дружбы народов имени Патриса Лумумбы»

Факультет искусственного интеллекта
(наименование основного учебного подразделения (ОУП)-разработчика ОП ВО)

РАБОЧАЯ ПРОГРАММА ДИСЦИПЛИНЫ

ОПТИМИЗАЦИЯ МОДЕЛЕЙ МАШИННОГО ОБУЧЕНИЯ
(наименование дисциплины/модуля)

Рекомендована МССН для направлений подготовки:

02.03.02 ФУНДАМЕНТАЛЬНАЯ ИНФОРМАТИКА И ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ;
09.03.03 ПРИКЛАДНАЯ ИНФОРМАТИКА
(код и наименование направления подготовки/специальности)

Освоение дисциплины ведется в рамках реализации основной профессиональной образовательной программы высшего образования (ОП ВО):

ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ: РАЗРАБОТКА И ОБУЧЕНИЕ ИНТЕЛЛЕКТУАЛЬНЫХ СИСТЕМ
(наименование (профиль/специализация) ОП ВО)

2026 г.

1. ЦЕЛЬ ОСВОЕНИЯ ДИСЦИПЛИНЫ

Дисциплина «Оптимизация моделей машинного обучения» входит в программу бакалавриата «Искусственный интеллект: разработка и обучение интеллектуальных систем» по направлениям подготовки 02.03.02 Фундаментальная информатика и информационные технологии и 09.03.03 Прикладная информатика, и изучается в 7 семестре 4 курса. Дисциплину реализует Кафедра прикладного искусственного интеллекта. Дисциплина состоит из 3 разделов и 39 тем и направлена на изучение математических основ и практических методов оптимизации в контексте машинного обучения: теории выпуклой и невыпуклой оптимизации (градиентный спуск, условия сходимости, стохастическая оптимизация), продвинутых методов обучения нейронных сетей (адаптивные оптимизаторы, learning rate scheduling, нормализация, регуляризация), методов автоматического подбора гиперпараметров (байесовская оптимизация, multi-fidelity, ранняя остановка), архитектурного поиска (NAS), методов сжатия и ускорения моделей (pruning, quantization, knowledge distillation), а также практик системной оптимизации ML-пайплайнов с учётом компромиссов между качеством, вычислительной стоимостью и временем разработки

Целью освоения дисциплины является формирование у студентов системных знаний о теоретических основах оптимизации в машинном обучении и практических навыков применения методов оптимизации на всех уровнях ML-системы — от математической постановки задачи и выбора алгоритма обучения до сжатия моделей и оптимизации инференса, включая способность анализировать ландшафт функции потерь, обосновывать выбор оптимизатора и стратегии обучения, проводить автоматический подбор гиперпараметров, сжимать модели с контролем потерь качества, а также оценивать компромиссы между качеством и вычислительными затратами при проектировании ML-решений.

2. ТРЕБОВАНИЯ К РЕЗУЛЬТАТАМ ОСВОЕНИЯ ДИСЦИПЛИНЫ

Освоение дисциплины «Оптимизация моделей машинного обучения» направлено на формирование у обучающихся следующих компетенций (части компетенций):

Таблица 2.1. Перечень компетенций, формируемых у обучающихся при освоении дисциплины (результаты освоения дисциплины)

Шифр	**Компетенция**	**Индикаторы достижения компетенции** (в рамках данной дисциплины)
УК-10	Способен принимать обоснованные экономические решения в различных областях жизнедеятельности	УК-10.2 Умеет обосновывать и применять основные положения и методы социально-экономических наук для принятия решений в различных областях жизнедеятельности;
УК-2	Способен определять круг задач в рамках поставленной цели и выбирать оптимальные способы их решения, исходя из действующих правовых норм, имеющихся ресурсов и ограничений	УК-2.2 Умеет анализировать альтернативные варианты решений для достижения намеченных результатов; разрабатывать план, определять целевые этапы и основные направления работ;
ОПК-1	Способен применять фундаментальные знания, полученные в области математических и естественных наук, методы математического анализа и моделирования, теоретического и экспериментального	ОПК-1.2 Умеет строить математические модели процессов и явлений, применять методы численного анализа и оптимизации для решения задач машинного обучения и обработки данных; ОПК-1.3 Владеет навыками проведения вычислительных экспериментов, анализа их результатов и обоснования выбора математического аппарата для решения конкретных профессиональных задач в области ИИ;

Шифр	Компетенция	Индикаторы достижения компетенции (в рамках данной дисциплины)
	исследования в профессиональной деятельности	
ОПК-6	Способен анализировать и разрабатывать организационно-технические процессы с применением методов системного анализа, математического моделирования и технологий искусственного интеллекта	ОПК-6.3 Владеет навыками построения онтологий и моделей предметных областей, оценки целесообразности и ограничений применения ИИ для решения конкретных организационно-технических задач;
ПК-1	Способен анализировать требования к программному обеспечению систем ИИ, разрабатывать технические спецификации и техническое задание на систему	ПК-1.1 Анализирует возможности реализации функциональных и нефункциональных требований к ПО систем ИИ, выявляет противоречия и ограничения;
ПК-3	Способен разрабатывать и реализовывать стратегии тестирования и контроля качества программного обеспечения систем ИИ	ПК-3.3 Оценивает результаты тестирования, реализует процесс контроля качества ПО систем ИИ;
FC-1	Способен проводить передовые исследования в области архитектур, алгоритмов МО, оптимизации и математики	FC-1.1 Разрабатывает фундаментальные основы и новые алгоритмы машинного обучения; FC-1.2 Разрабатывает новые архитектуры глубоких нейросетей; FC-1.3 Развивает методы ускорения обучения;
MF-3	Способен применять современные методы оптимизации для обучения моделей машинного обучения, настройки гиперпараметров и решения задач искусственного интеллекта	MF-3.1 Применяет методы оптимизации для разработки и исследования обучающих алгоритмов; MF-3.2 Применяет методы оптимизации для настройки гиперпараметров моделей машинного обучения, включая использование методов поиска (поиск по сетке, случайный поиск) и байесовской оптимизации;
ML-3	Способен применять классические алгоритмы машинного обучения с пониманием их математических основ и областей применения	ML-3.1 Обосновывает способы и варианты применения классических методов и моделей МО в задачах ИИ, включая их математическое (алгоритмическое) преобразование и адаптацию к специфике задачи;
SS-3	Способен к критическому анализу, метарефлексии и переносу знаний при работе с системами ИИ	SS-3.2 Определяет релевантность применения ИИ для решения конкретных задач, анализирует поведение ИИ в техническом, социальном и правовом контекстах, переносит идеи и методы за пределы исходной предметной области;

3. МЕСТО ДИСЦИПЛИНЫ В СТРУКТУРЕ ОП ВО

Дисциплина «Оптимизация моделей машинного обучения» относится к обязательной части блока 1 «Дисциплины (модули)» образовательной программы высшего образования.

В рамках образовательной программы высшего образования обучающиеся также осваивают другие дисциплины и/или практики, способствующие достижению запланированных результатов освоения дисциплины «Оптимизация моделей машинного обучения».

Таблица 3.1. Перечень компонентов ОП ВО, способствующих достижению запланированных результатов освоения дисциплины

Шифр	Наименование компетенции	Предшествующие дисциплины/модули, практики*	Последующие дисциплины/модули, практики*
УК-10	Способен принимать обоснованные экономические решения в различных областях жизнедеятельности	Правоведение;	MLOps и промышленная разработка систем искусственного интеллекта;
УК-2	Способен определять круг задач в рамках поставленной цели и выбирать оптимальные способы их решения, исходя из действующих правовых норм, имеющихся ресурсов и ограничений	Эксплуатационная практика (учебная); Правоведение; Этика и безопасность использования искусственного интеллекта; Методы разработки решений на основе искусственного интеллекта (Git, Docker); Проектирование и разработка систем компьютерного зрения; Практикум по обработке естественного языка (NLP);	Проектирование и разработка систем компьютерного зрения; Практикум по обработке естественного языка (NLP); MLOps и промышленная разработка систем искусственного интеллекта;
ОПК-6	Способен анализировать и разрабатывать организационно-технические процессы с применением методов системного анализа, математического моделирования и технологий искусственного интеллекта	Введение в искусственный интеллект; Онтология и графы знаний; Искусственный интеллект и когнитивная психология; Этика и безопасность использования искусственного интеллекта; Методы машинного обучения;	Методы машинного обучения;
ОПК-1	Способен применять фундаментальные знания, полученные в области математических и естественных наук, методы математического анализа и моделирования, теоретического и экспериментального исследования в профессиональной деятельности	Линейная алгебра; Дискретная математика; Математический анализ; Теория вероятностей и математическая статистика; Дифференциальные уравнения; Численная линейная алгебра; Методы машинного обучения; Основы глубокого обучения; Статистические методы и первичный анализ данных;	Методы машинного обучения;
ПК-1	Способен анализировать требования к программному обеспечению систем ИИ, разрабатывать технические спецификации и техническое задание на систему	Эксплуатационная практика (учебная); Эксплуатационная практика (производственная); Технологическая (проектно-технологическая) практика (учебная); Правоведение; Параллельное и распределенное программирование; Введение в искусственный интеллект; Искусственный интеллект и когнитивная психология; Этика и безопасность	Преддипломная практика; Методы машинного обучения; MLOps и промышленная разработка систем искусственного интеллекта; Проектирование и разработка систем компьютерного зрения; Практикум по обработке естественного языка (NLP);

Шифр	Наименование компетенции	Предшествующие дисциплины/модули, практики*	Последующие дисциплины/модули, практики*
		использования искусственного интеллекта; Методы машинного обучения; Массово-параллельные вычисления в машинном обучении (GPU); Основы глубокого обучения; Большие языковые модели**; История и теория программирования; Программирование на языке C++; Методы разработки решений на основе искусственного интеллекта (Git, Docker); Введение в базы данных; Онтология и графы знаний; Проектирование и разработка систем компьютерного зрения; Практикум по обработке естественного языка (NLP);	
ПК-3	Способен разрабатывать и реализовывать стратегии тестирования и контроля качества программного обеспечения систем ИИ	Технологическая (проектно-технологическая) практика (учебная); Эксплуатационная практика (учебная); Эксплуатационная практика (производственная); Теория вероятностей и математическая статистика; Этика и безопасность использования искусственного интеллекта; Статистические методы и первичный анализ данных; Методы машинного обучения; Обработка и анализ изображений и видео с помощью методов искусственного интеллекта; Анализ естественного языка с помощью методов искусственного интеллекта; Программирование на языке Python; Методы разработки решений на основе искусственного интеллекта (Git, Docker); Проектирование и разработка систем компьютерного зрения; Практикум по обработке естественного языка (NLP);	Преддипломная практика; Методы машинного обучения; MLOps и промышленная разработка систем искусственного интеллекта; Проектирование и разработка систем компьютерного зрения; Практикум по обработке естественного языка (NLP);
SS-3	Способен к критическому анализу, метарефлексии и переносу знаний при работе с системами ИИ	Эксплуатационная практика (учебная); Эксплуатационная практика (производственная); Теория вероятностей и математическая статистика; Искусственный интеллект и когнитивная психология; Этика и безопасность использования искусственного	Преддипломная практика; Методы машинного обучения; Вайб-кодинг**; MLOps и промышленная разработка систем искусственного интеллекта; Проектирование и разработка систем компьютерного зрения;

Шифр	Наименование компетенции	Предшествующие дисциплины/модули, практики*	Последующие дисциплины/модули, практики*
		интеллекта; Статистические методы и первичный анализ данных; Методы машинного обучения; Обработка и анализ изображений и видео с помощью методов искусственного интеллекта; Анализ естественного языка с помощью методов искусственного интеллекта; Правоведение; Введение в искусственный интеллект; Введение в компьютерное зрение; Проектирование и разработка систем компьютерного зрения; Практикум по обработке естественного языка (NLP); Основы программирования HTML - CSS - JavaScript**; Основы программирования на языке NodeJS**; Основы программирования на языке Go**; Основы программирования на языке Julia**; Основы робототехники**; Цифровые двойники**; Философия; Большие языковые модели**;	Практикум по обработке естественного языка (NLP); Обработка сигналов**; Анализ временных рядов**;
MF-3	Способен применять современные методы оптимизации для обучения моделей машинного обучения, настройки гиперпараметров и решения задач искусственного интеллекта	Эксплуатационная практика (учебная); Численная линейная алгебра; Методы машинного обучения;	Методы машинного обучения;
ML-3	Способен применять классические алгоритмы машинного обучения с пониманием их математических основ и областей применения	Эксплуатационная практика (учебная); Методы машинного обучения; Основы глубокого обучения;	Преддипломная практика; Методы машинного обучения;
FC-1	Способен проводить передовые исследования в области архитектур, алгоритмов МО, оптимизации и математики	Эксплуатационная практика (учебная); Линейная алгебра; Математический анализ; Теория вероятностей и математическая статистика; Методы машинного обучения; Введение в искусственный интеллект; Основы глубокого обучения; Численная линейная алгебра; Параллельное и распределенное программирование; Массово-параллельные вычисления в машинном	Методы машинного обучения; Преддипломная практика;

Шифр	Наименование компетенции	Предшествующие дисциплины/модули, практики*	Последующие дисциплины/модули, практики*
		обучении (GPU);	

* - заполняется в соответствии с матрицей компетенций и СУП ОП ВО

** - элективные дисциплины /практики

4. ОБЪЕМ ДИСЦИПЛИНЫ И ВИДЫ УЧЕБНОЙ РАБОТЫ

Общая трудоемкость дисциплины «Оптимизация моделей машинного обучения» составляет «4» зачетные единицы.

Таблица 4.1. Виды учебной работы по периодам освоения образовательной программы высшего образования для очной формы обучения.

Вид учебной работы	**ВСЕГО, ак.ч.**		**Семестр(-ы)**
			7
Контактная работа, ак.ч.	*78*		*78*
Лекции (**ЛК**)	26		26
Лабораторные работы (**ЛР**)	0		0
Практические/семинарские занятия (**СЗ**)	52		52
Самостоятельная работа обучающихся, ак.ч.	*39*		*39*
Контроль (экзамен/зачет с оценкой), ак.ч.	*27*		*27*
Общая трудоемкость дисциплины	**ак.ч.**	**144**	**ак.ч.**
	зач.ед.	**4**	**зач.ед.**

5. СОДЕРЖАНИЕ ДИСЦИПЛИНЫ

Таблица 5.1. Содержание дисциплины (модуля) по видам учебной работы

Номер раздела	Наименование раздела дисциплины	Наименование темы		Содержание темы	Вид учебной работы *	Формируемые индикаторы
Раздел 1	Математические основы оптимизации для ML	1.1	Введение: оптимизация как ядро машинного обучения	Обучение ML-модели как задача оптимизации: минимизация эмпирического риска. Функции потерь: MSE, Cross-Entropy, Hinge, Huber — свойства, гладкость, выпуклость. Ландшафт функции потерь: локальные минимумы, седловые точки, плато, острые и плоские минимумы. Связь ландшафта с обобщающей способностью (sharp vs. flat minima). Обзор курса: от теории к промышленной оптимизации	ЛК	MF-3.1, ОПК-1.2, FC-1.1
		1.2	Выпуклая оптимизация: теория и условия сходимости	Выпуклые множества и функции: определения, примеры. Строгая выпуклость, сильная выпуклость. Условия оптимальности: KKT (Каруша-Куна-Таккера). Градиентный спуск: теорема о сходимости для выпуклых и сильно выпуклых функций. Скорость сходимости: $O(1/t)$ vs. $O(\rho^t)$. Число обусловленности и его влияние. Связь с линейной регрессией (выпуклая задача) и нейросетями (невыпуклая)	ЛК	MF-3.1, ОПК-1.2
		1.3	Стохастическая оптимизация и методы первого порядка	SGD: постановка, сходимость, дисперсия градиента, расписание learning rate ($O(1/\sqrt{t})$). Мини-батч SGD: снижение дисперсии. Моментум: Polyak momentum, Nesterov accelerated gradient. Адаптивные методы: AdaGrad, RMSProp, Adam — вывод, свойства. AdamW: decoupled weight decay. Сравнение теоретических свойств: regret bounds, сходимость. Проблемы Adam: обобщение хуже SGD (дискуссия)	ЛК	MF-3.1, MF-3.2, ОПК-1.2
		1.4	Невыпуклая оптимизация и ландшафт нейросетей	Невыпуклость функции потерь нейросетей. Седловые точки: почему их больше, чем локальных минимумов в высокой размерности. Результаты о ландшафте: все локальные минимумы примерно равноценны (обзор). Overparametrization и implicit regularization. Лотерейная гипотеза (Lottery Ticket Hypothesis, обзор). Двойной спуск (double descent). Связь с практикой: почему SGD находит хорошие решения	ЛК	FC-1.1, MF-3.1, ОПК-1.2
		1.5	Методы второго порядка и ограниченная	Метод Ньютона: гессиан, квадратичная аппроксимация,	ЛК	MF-3.1,

Номер раздела	Наименование раздела дисциплины	Наименование темы		Содержание темы	Вид учебной работы *	Формируемые индикаторы
			оптимизация	суперлинейная сходимость. Проблемы: вычисление и хранение гессиана $O(n^2)$. Квази-ньютоновские методы: L-BFGS (обзор). Natural gradient: метрика Фишера (обзор). K-FAC (обзор). Ограниченная оптимизация: метод проекции, proximal gradient (L1-регуляризация). Связь с Lasso: soft-thresholding. Обзор: когда методы второго порядка полезны в ML		ОПК-1.2, ОПК-1.3
		1.6	Практикум: визуализация ландшафта функции потерь	Визуализация 2D-срезов loss surface нейросети (Li et al., 2018): случайные направления, фильтровая нормализация. Сравнение: ResNet vs. plain network (sharp vs. flat). Визуализация траектории оптимизации: SGD, Adam, SGD+momentum. Практика: построение ландшафта для учебной сети на CIFAR-10	СЗ	MF-3.1, FC-1.1
		1.7	Практикум: реализация градиентного спуска с нуля	Реализация на NumPy: GD, SGD, mini-batch SGD, SGD+momentum, Nesterov. Тестирование на квадратичной функции и функции Розенброка. Визуализация траекторий. Влияние learning rate и batch size. Практика: сравнение скорости сходимости	СЗ	MF-3.2, ОПК-1.3
		1.8	Практикум: реализация адаптивных оптимизаторов с нуля	Реализация AdaGrad, RMSProp, Adam, AdamW на NumPy. Тестирование на задаче с ill-conditioned Hessian. Визуализация: как адаптивные методы справляются с разным масштабом признаков. Сравнение с vanilla SGD. Связь с PyTorch оптимизаторами	СЗ	MF-3.2, ОПК-1.3
		1.9	Практикум: сравнение оптимизаторов на нейросети	Обучение CNN на CIFAR-10 с разными оптимизаторами: SGD, SGD+momentum, Adam, AdamW. Фиксация всех гиперпараметров кроме оптимизатора. Метрики: train loss, val loss, val accuracy по эпохам. Анализ: скорость начальной сходимости vs. финальное качество. Обсуждение: «Adam сходится быстрее, SGD обобщает лучше»	СЗ	MF-3.2, ML-3.1
		1.10	Практикум: learning rate scheduling	Schedulers в PyTorch: StepLR, ExponentialLR, CosineAnnealingLR, CosineAnnealingWarmRestarts, OneCycleLR. Warmup: линейный, константный. Learning rate finder (Smith, 2017): range test. Практика:	СЗ	ML-3.1, MF-3.2

Номер раздела	Наименование раздела дисциплины	Наименование темы		Содержание темы	Вид учебной работы *	Формируемые индикаторы
				визуализация LR по эпохам, обучение с разными schedulers, сравнение кривых обучения		
		1.11	Практикум: L1/L2 регуляризация как ограниченная оптимизация	L2 (weight decay): интерпретация как ограниченная оптимизация (Тихоновская регуляризация). L1 (Lasso): sparsity, soft-thresholding, proximal gradient. Визуализация: контуры функции потерь + ограничение. Практика: сравнение L1 vs. L2 на задаче регрессии, анализ разреженности весов	СЗ	MF-3.1, ОПК-1.2
		1.12	Практикум: инициализация весов и её влияние на оптимизацию	Xavier (Glorot) initialization: вывод для sigmoid/tanh. Kaiming (He) initialization: вывод для ReLU. Нулевая и случайная инициализация: проблемы (мёртвые нейроны, симметрия). Практика: обучение глубокой сети с разными инициализациями, визуализация градиентов по слоям, связь с затухающими/взрывающимися градиентами	СЗ	MF-3.2, FC-1.1
		1.13	Практикум: gradient clipping и стабильность обучения	Проблема взрывающихся градиентов: RNN, глубокие сети. Gradient clipping: по норме (clip_grad_norm_), по значению (clip_grad_value_). Gradient accumulation: эмуляция большого батча. Практика: обучение LSTM с/без gradient clipping, визуализация нормы градиента	СЗ	MF-3.2, ML-3.1
		1.14	Практикум: мини-проект — анализ оптимизации модели	Сквозная задача: выбор модели и датасета → обучение с разными оптимизаторами, schedulers, инициализациями, регуляризациями → систематическое сравнение → анализ ландшафта → выбор лучшей конфигурации с обоснованием.	СЗ	ML-3.1, ПК-3.3, FC-1.1

Номер раздела	Наименование раздела дисциплины	Наименование темы		Содержание темы	Вид учебной работы *	Формируемые индикаторы
				Отчёт: таблица экспериментов, визуализации, выводы		
Раздел 2	Автоматическая настройка гиперпараметров и архитектурный поиск	2.1	Подбор гиперпараметров: формализация и методы	Гиперпараметры: определение, примеры (LR, batch size, архитектура, регуляризация). Формализация: bilevel optimization. Grid Search: полный перебор, проклятие размерности. Random Search: теоретическое обоснование (Bergstra & Bengio, 2012). Successive Halving и Hyperband: ранняя остановка неперспективных конфигураций. Сравнение методов: качество, вычислительная стоимость	ЛК	ML-3.1, MF-3.2
		2.2	Байесовская оптимизация гиперпараметров	Суррогатная модель: Gaussian Process (повторение), функция приобретения (Expected Improvement, UCB, PI). Цикл: fit surrogate → maximize acquisition → evaluate → update. Tree-structured Parzen Estimator (TPE): Optuna. Multi-fidelity: комбинация с Hyperband (BOHB). Практические аспекты: пространство поиска, начальные точки, warm-start	ЛК	FC-1.1, ML-3.1, MF-3.2
		2.3	Neural Architecture Search (NAS)	Пространство поиска: макро (архитектура целиком), микро (ячейка/блок). Стратегии поиска: RL-based (NASNet), evolutionary, differentiable (DARTS). One-shot NAS: weight sharing. Hardware-aware NAS: оптимизация latency + accuracy (EfficientNet, MnasNet). Связь с AutoML. Ограничения: вычислительная стоимость, воспроизводимость	ЛК	FC-1.2, ML-3.1
		2.4	AutoML: интеграция автоматического подбора	AutoML как концепция: автоматизация data preprocessing + feature engineering + model	ЛК	FC-1.2, ML-3.1, УК-2.2

Номер раздела	Наименование раздела дисциплины	Наименование темы		Содержание темы	Вид учебной работы *	Формируемые индикаторы
				selection + HPO. Инструменты: auto-sklearn, AutoGluon, FLAML, H2O AutoML. Automated feature engineering: Featuretools (повторение). Связь с NAS, HPO, model selection. Ограничения AutoML: интерпретируемость, контроль, ресурсы. Когда AutoML полезен, когда нет		
		2.5	Практикум: Optuna — базовый подбор гиперпараметров	Optuna: study, trial, suggest_int/float/categorical. Цель: минимизация val_loss. Pruning: MedianPruner, HyperbandPruner. Visualization: optimization_history, param_importances, parallel_coordinate. Практика: подбор гиперпараметров для CNN на CIFAR-10	СЗ	ML-3.1, MF-3.2
		2.6	Практикум: Optuna — многокритериальная оптимизация	Multi-objective optimization: accuracy + latency. Pareto front. Optuna: multi-objective study. Визуализация Pareto front. Выбор точки на фронте: trade-off для конкретного SLA. Практика: одновременная оптимизация accuracy и inference time	СЗ	ML-3.1, ПК-1.1, УК-10.2
		2.7	Практикум: Optuna — пространство поиска архитектуры	Поиск архитектуры через Optuna: число слоёв, hidden size, dropout rate, тип активации, тип нормализации. Conditional hyperparameters. Практика: совместный поиск архитектуры и гиперпараметров обучения для MLP/CNN	СЗ	FC-1.2, ML-3.1
		2.8	Практикум: Hyperband и ранняя остановка	Hyperband: принцип successive halving. Optuna HyperbandPruner. Интеграция с PyTorch: reporting intermediate values. Практика: сравнение Hyperband vs. MedianPruner vs. без pruning: время поиска, качество найденной конфигурации	СЗ	ML-3.1, FC-1.1
		2.9	Практикум: AutoML — AutoGluon и FLAML	AutoGluon: TabularPredictor, автоматический подбор моделей и ансамбля. FLAML: бюджетная оптимизация. Практика: запуск AutoML на реальном датасете, анализ выбранных моделей, сравнение с ручным подбором. Обсуждение: trade-offs (качество, время, контроль)	СЗ	FC-1.2, ML-3.1
		2.10	Практикум: оценка стоимости обучения и бюджетирование	Оценка стоимости: GPU-часы, облачные тарифы, energy consumption. Формулирование бюджета: сколько экспериментов можно провести за T часов на K GPU. Стратегия: начать с дешёвых экспериментов (маленькая	СЗ	УК-10.2, ПК-1.1

Номер раздела	Наименование раздела дисциплины	Наименование темы		Содержание темы	Вид учебной работы *	Формируемые индикаторы
				модель, подвыборка), масштабировать лучшие. Практика: расчёт стоимости HPO-кампании, обоснование бюджета		
		2.11	Практикум: NAS с DARTS (обзорная реализация)	DARTS: differentiable architecture search, relaxation of discrete choice, bilevel optimization. Упрощённая реализация: поиск архитектуры ячейки для CNN. Практика: запуск DARTS на маленьком датасете, анализ найденной архитектуры. Обсуждение: затраты NAS vs. ручной дизайн	СЗ	FC-1.2, ML-3.1
		2.12	Практикум: transfer learning как оптимизация	Transfer learning: инициализация близко к хорошему решению (warm start optimization). Выбор стратегии: linear probing vs. fine-tuning vs. progressive unfreezing. Layer-wise LR decay. Практика: сравнение стратегий fine-tuning как задача оптимизации, анализ sensitivity к LR	СЗ	FC-1.2, ML-3.1
		2.13	Практикум: мини-проект — оптимальная конфигурация ML-системы	Сквозная задача: реальный датасет → HPO (Optuna с multi-objective: accuracy vs. latency) → подбор архитектуры → подбор стратегии обучения → оценка стоимости → документирование. Формат: отчёт с Pareto front, обоснованием выбора, оценкой затрат	СЗ	ML-3.1, УК-10.2, ПК-3.3
Раздел 3	Сжатие моделей и оптимизация инференса	3.1	Knowledge Distillation: обучение компактных моделей	Идея: обучение «ученика» на soft predictions «учителя». Temperature scaling: softmax с температурой. Distillation loss: KL-divergence между soft predictions + task loss. Варианты: logit matching, feature matching, attention transfer. DistilBERT, TinyBERT. Self-distillation. Born-again networks. Связь с оптимизацией: регуляризация через учителя	ЛК	FC-1.2, ML-3.1
		3.2	Pruning: разреженность моделей	Мотивация: большинство весов можно обнулить без потери качества. Типы: unstructured (отдельные веса), structured (каналы, головы, слои). Критерии: magnitude pruning, movement pruning, gradient-based. Расписание: one-shot, iterative (gradual pruning). Lottery Ticket Hypothesis: sparse subnetwork при инициализации. Инструменты: PyTorch pruning, Neural Magic	ЛК	FC-1.2, FC-1.3
		3.3	Quantization: снижение точности вычислений	: post-training quantization (PTQ), quantization-aware training (QAT). Symmetric vs. asymmetric quantization. Per-tensor vs. per-channel. INT8, INT4, FP16, BF16, FP8. Калибровка: MinMax, Percentile, MSE. QAT: fake quantize + STE (Straight-Through Estimator). Инструменты: PyTorch	ЛК	FC-1.3, FC-1.2

Номер раздела	Наименование раздела дисциплины	Наименование темы		Содержание темы	Вид учебной работы *	Формируемые индикаторы
				quantization, ONNX Runtime, TensorRT. Связь с Tensor Cores (INT8, FP16)		
		3.4	Системная оптимизация: от модели к production	Оптимизация графа вычислений: operator fusion, constant folding, dead code elimination. torch.compile (TorchDynamo + Inductor). ONNX: graph optimization levels. TensorRT: layer fusion, precision selection. Batching: dynamic batching для инференса. Кэширование: KV-cache для LLM. Profiling: bottleneck analysis. Оценка компромиссов: accuracy vs. latency vs. throughput vs. cost vs. memory	ЛК	ПК-1.3, ПК-1.1, ОПК-6.3
		3.5	Практикум: knowledge distillation для классификации	Teacher: ResNet-50 (обученный). Student: MobileNet / shallow CNN. Distillation loss: α × KL(soft_teacher, soft_student) + (1-α) × CE(labels, student). Подбор temperature и α. Практика: обучение student с/без distillation, сравнение accuracy, latency, model size	СЗ	ПК-1.2, ML-3.1
		3.6	Практикум: distillation для NLP (DistilBERT)	Distillation BERT → DistilBERT: архитектура (6 слоёв vs. 12), обучение. Hugging Face: загрузка DistilBERT, сравнение с BERT. Fine-tuning DistilBERT на задаче классификации. Практика: сравнение BERT vs. DistilBERT: accuracy, latency, model size, memory	СЗ	ПК-1.2, ПК-1.3
		3.7	Практикум: pruning — unstructured и structured	PyTorch pruning: L1Unstructured, RandomStructured, LnStructured. Iterative pruning: train → prune → fine-tune → repeat. Анализ: sparsity vs. accuracy curve. Structured pruning: удаление каналов CNN. Практика: pruning CNN до 70–90% sparsity, оценка потери качества	СЗ	ПК-1.2, ПК-1.3
		3.8	Практикум: post-training quantization (PTQ)	PyTorch dynamic quantization: torch.quantization.quantize_dynamic. Static quantization: калибровка на calibration set. ONNX Runtime INT8. Практика: quantization CNN и BERT, замеры: model size, latency (CPU), accuracy. Анализ: какие слои чувствительны к quantization	СЗ	ПК-1.3, ОПК-1.3
		3.9	Практикум: quantization-aware training (QAT)	QAT в PyTorch: torch.quantization.prepare_qat, fake quantize nodes. Обучение с QAT: fine-tuning с имитацией INT8. Сравнение PTQ vs. QAT по accuracy. Практика: QAT для CNN, замеры: восстановление accuracy после quantization	СЗ	ПК-1.3, ПК-1.2
		3.10	Практикум: torch.compile и оптимизация	torch.compile: режимы, принцип работы (TorchDynamo	СЗ	ПК-1.3,

Номер раздела	Наименование раздела дисциплины	Наименование темы		Содержание темы	Вид учебной работы *	Формируемые индикаторы
			графа	capture + Inductor codegen). ONNX optimization: graph optimizers. Практика: компиляция модели, замеры ускорения, анализ graph breaks. Сравнение: eager mode vs. compiled vs. ONNX Runtime vs. TensorRT		ОПК-6.3
		3.11	Практикум: комплексное сжатие — distillation + pruning + quantization	Комбинирование методов: distillation → pruning → quantization. Порядок применения: влияние последовательности. Практика: полный пайплайн сжатия (ResNet-50 → MobileNet-distilled → pruned → INT8), итоговая таблица: accuracy, latency, size, memory	СЗ	ГС-1.2, ГС-1.3, SS-3.2
		3.12	Практикум: итоговый проект — оптимизация ML-модели для production	Финальная интеграция: выбор модели и задачи → HPO (Optuna, multi-objective) → обучение с оптимальной конфигурацией → distillation → pruning → quantization → экспорт (ONNX/TensorRT) → замеры на целевом оборудовании → анализ trade-offs (accuracy vs. latency vs. cost). Документирование: отчёт с Pareto front, обоснование каждого решения, оценка стоимости инференса в production. Презентация	СЗ	ГС-1.2, ГС-1.3, ПК-1.1, ПК-3.3

* - заполняется только по **ОЧНОЙ** форме обучения: *ЛК – лекции; ЛР – лабораторные работы; СЗ – практические/семинарские занятия.*

6. МАТЕРИАЛЬНО-ТЕХНИЧЕСКОЕ ОБЕСПЕЧЕНИЕ ДИСЦИПЛИНЫ

Таблица 6.1. Материально-техническое обеспечение дисциплины

Тип аудитории	**Оснащение аудитории**	**Специализированное учебное/лабораторное оборудование, ПО и материалы для освоения дисциплины (при необходимости)**
Лекционная	Аудитория для проведения занятий лекционного типа, оснащенная комплектом специализированной мебели; доской (экраном) и техническими средствами мультимедиа презентаций.	
Семинарская	Аудитория для проведения занятий семинарского типа, групповых и индивидуальных консультаций, текущего контроля и промежуточной аттестации, оснащенная комплектом специализированной мебели и техническими средствами мультимедиа презентаций.	Персональные компьютеры, необходимое ПО
Для самостоятельной работы	Аудитория для самостоятельной работы обучающихся (может использоваться для проведения семинарских занятий и консультаций), оснащенная комплектом специализированной мебели и компьютерами с доступом в ЭИОС.	Персональные компьютеры, необходимое ПО

* - аудитория для самостоятельной работы обучающихся указывается **ОБЯЗАТЕЛЬНО**!

7. УЧЕБНО-МЕТОДИЧЕСКОЕ И ИНФОРМАЦИОННОЕ ОБЕСПЕЧЕНИЕ ДИСЦИПЛИНЫ

Основная литература:

1. «Золкин, А. Л. Машинно-ориентированные языки программирования в сфере искусственного интеллекта : учебное пособие для вузов / А. Л. Золкин. — 2-е изд., стер. — Санкт-Петербург : Лань, 2026. — 168 с. — ISBN 978-5-507-56208-4. — Текст : электронный // Лань : электронно-библиотечная система. — URL: https://e.lanbook.com/book/514155 » (Золкин, А. Л. Машинно-ориентированные языки программирования в сфере искусственного интеллекта : учебное пособие для вузов / А. Л. Золкин. — 2-е изд., стер. — Санкт-Петербург : Лань, 2026. — ISBN 978-5-507-56208-4. — Текст : электронный // Лань : электронно-библиотечная система. — URL: https://e.lanbook.com/book/514155

2. Баланов, А. Н. Машинное обучение и искусственный интеллект : учебное пособие для вузов / А. Н. Баланов. — 3-е изд., стер. — Санкт-Петербург : Лань, 2026. — 172 с. — ISBN 978-5-507-54962-7. — Текст : электронный // Лань : электронно-библиотечная система. — URL: https://e.lanbook.com/book/513580

Дополнительная литература:

1. Искусственный интеллект. Лабораторный практикум : учебное пособие для вузов / А. И. Галиева, Г. И. Галиева, В. Г. Дмитриев, Ф. А. Баязитов. — Санкт-Петербург : Лань, 2026. — 316 с. — ISBN 978-5-507-54528-5. — Текст : электронный // Лань :

электронно-библиотечная система. — URL: https://e.lanbook.com/book/516483

Ресурсы информационно-телекоммуникационной сети «Интернет»:

1. ЭБС РУДН и сторонние ЭБС, к которым студенты университета имеют доступ на основании заключенных договоров

- Электронно-библиотечная система РУДН – ЭБС РУДН https://mega.rudn.ru/MegaPro/Web
- ЭБС «Университетская библиотека онлайн» http://www.biblioclub.ru
- ЭБС «Юрайт» http://www.biblio-online.ru
- ЭБС «Консультант студента» www.studentlibrary.ru
- ЭБС «Знаниум» https://znanium.ru/

2. Базы данных и поисковые системы

- Sage https://journals.sagepub.com/
- Springer Nature Link https://link.springer.com/
- Wiley Journal Database https://onlinelibrary.wiley.com/
- Наукометрическая база данных Lens.org https://www.lens.org

Учебно-методические материалы для самостоятельной работы обучающихся при освоении дисциплины/модуля:*

1. Курс лекций по дисциплине «Оптимизация моделей машинного обучения».

* - все учебно-методические материалы для самостоятельной работы обучающихся размещаются в соответствии с действующим порядком на странице дисциплины **в ТУИС**!