Аудиоанализ, распознавание и генерация речи

Пояснения к словам иностранного происхождения

Здесь собраны пояснения к словам иностранного происхождения, которые используются в программе курса, но пока не включены в словарную норму русского языка.

01 Кому подойдёт курс

Инженерам машинного обучения (ML)
Освоите работу с аудиоданными и речевыми моделями. Разберётесь в спектрограммах, аудиопризнаках и архитектурах для распознавания речи, научитесь применять современные модели для задач аудиоанализа.

Инженерам по глубокому обучению (DL)
Изучите современные архитектуры для распознавания и синтеза речи, разберётесь в аудиотрансформерах и научитесь строить полноценные аудиосервисы — от обработки сигнала до оптимизированного инференса.

NLP-специалистам
Расширите экспертизу в сторону речевых технологий. Поймёте, как работают системы распознавания и синтеза речи и как интегрировать голосовые интерфейсы и речевые модели с языковыми моделями.

02 Какие знания и навыки освоите

Вы научитесь

- Преобразовывать сырой аудиосигнал в числовые признаки, пригодные для анализа и машинного обучения, и готовить данные для последующих этапов (классификация, распознавание, синтез).
- Переводить аудиоречь в текстовый формат, используя современные нейросетевые архитектуры и алгоритмы декодирования.
- Генерировать естественное аудио из текста, управляя интонацией, тембром и скоростью речи.
- Разворачивать ASR/TTS-системы в продакшене: ускорять вычисления, снижать потребление памяти и объединять компоненты в единый рабочий конвейер.

03 Как проходит курс

- Персональный разбор кейсов с наставником
- Теория и практика на платформе Практикума
- Доступ из любой точки мира в удобное время
- Практические задания

Аудиоанализ, распознавание и генерация речи

03 Как проходит курс

- Персональный разбор кейсов с наставником
- Теория и практика на платформе Практикума
- Доступ из любой точки мира в удобное время
- Практические задания

Что вас ждёт на обучении

Знания и инструменты, актуальные в профессии

Команда сопровождения будет помогать вам на протяжении всего обучения

Удостоверение о повышении квалификации

Опыт экспертов из Яндекса и других крупных компаний

Аудиоанализ, распознавание и генерация речи

Сравнение тарифов

	Аудиоанализ, распознавание и генерация речи	Инженер по глубокому обучению нейросетей: аудиоречевые технологии
Описание	**Освоите современные технологии анализа звука и обработки речи**	**Освоите глубокие нейросети и примените их к задачам анализа звука**
Период обучения	3 месяца	5 месяцев
Количество проектов с ревью	4 проекта для портфолио	8 проектов для портфолио
Ресурсы	Доступ к виртуальным машинам с GPU и хранилищу S3 для выполнения практики	Доступ к виртуальным машинам с GPU и хранилищу S3 для выполнения практики
Образовательные темы	✕ Навыки для уверенной работы на DL-фреймворке PyTorch: сможете самостоятельно строить нейросети, правильно их обучать, подготавливать данные, объяснять принципы и элементы, на которых строятся нейросетевые решения Цифровая обработка аудио, распознавание речи, большие речевые модели, синтез речи, инференс и оптимизация аудио-моделей	Навыки для уверенной работы на DL-фреймворке PyTorch: сможете самостоятельно строить нейросети, правильно их обучать, подготавливать данные, объяснять принципы и элементы, на которых строятся нейросетевые решения Цифровая обработка аудио, распознавание речи, большие речевые модели, синтез речи, инференс и оптимизация аудио-моделей
Документ о завершении обучения	Удостоверение о повышении квалификации	Диплом о профессиональной переподготовке

Аудиоанализ, распознавание и генерация речи

3 месяца

продолжительность курса

4 проекта

в портфолио (опционально)

2 НЕДЕЛИ

01

Введение
Цифровая обработка аудио и классические модели классификации

2 НЕДЕЛИ

02

Распознавание речи: классические CTC и трансформерные модели

2 НЕДЕЛИ

03

SSL и большие речевые модели

2 НЕДЕЛИ

04

Синтез речи и управление характеристиками голоса

2 НЕДЕЛИ

05

Инференс и оптимизация аудио-моделей

1 НЕДЕЛЯ

Итоговый проект

Цифровая обработка аудио и классические модели классификации

01

2 недели
10-15 часов

Выберете задачу под автоматизацию с помощью ИИ-агента из предложенного списка. Изучите архитектуру ИИ-агентов. Освоите подготовку системных промптов для корректной работы ИИ-агента.

Содержание

Тема 1. Аудиосигнал как числовое представление: что такое дискретизация и амплитуда.	• Понимание физической природы звука. • Понимание устройства аналоговых носителей звука. • Знание, что такое амплитуда и глубина кодирования. • Понимание отличий LPCM от float. • Знание, что такое частота дискретизации. • Понимание разницы между популярными кодеками и что такое кодирование с потерями. • Умение загружать, сохранять и визуализировать аудиосигнал (librosa, torchaudio).
Тема 2. Частотные признаки: как получить STFT, мел-спектрограммы и MFCC.	• Понимание основ частотного анализа (спектр, ряд/преобразование Фурье). • Знание принципов STFT (окно, шаг, разрешение) и умение строить спектрограмму. • Понимание мотивации использования мел-шкалы и умение строить мел-спектрограмму. • Понимание назначения MFCC и умение их вычислять. • Навык визуализации частотных признаков.
Тема 3. Подготовка аудиодатасетов: нормализация, выравнивание длины, шумы и фоновые эффекты	• Понимание необходимости единой частоты дискретизации и ресемплинга. • Умение применять пиковую и RMS-нормализацию. • Навык центрирования и случайного кропа для выравнивания длины. • Понимание влияния шумов и умение подмешивать их с заданным SNR. • Понимание природы реверберации и умение применять свёртку с RIR. • Умение симулировать искажения микрофонов и применять аудиоэффекты как аугментацию.
Тема 4. Классические модели: CNN, RNN и CRNN для аудиоклассификации.	• Понимание архитектур CNN для анализа спектрограмм (извлечение локальных паттернов). • Понимание архитектур RNN для моделирования временных зависимостей в аудио. • Понимание комбинированной архитектуры CRNN (CNN для признаков, RNN для времени). • Умение сравнивать эффективность разных архитектур на задаче классификации. • Навык построения конвейера обучения
Проект	• Создание полного конвейера обработки аудио (звук -> спектрограммы -> модель). • Построение аудиоклассификатора на основе свёрточных или рекуррентных сетей. • Проведение экспериментов по сравнению архитектур (CNN vs RNN vs CRNN). • Анализ результатов и выбор оптимальной архитектуры для задачи.

Распознавание речи: классические CTC и трансформерные модели 02

2 недели
10-15 часов

Освоите базовые и трансформерные архитектуры распознавания речи: от подготовки данных и CTC-моделей до seq2seq-подходов и современных encoder-decoder систем. Научитесь обучать модели, выполнять декодирование и оценивать качество распознавания.

Содержание

Тема 1. Подготовка данных для ASR: сегментация аудио, нормализация сигнала, токенизация текста, выбор единиц распознавания и формирование обучающих пар звук-текст.	• Понимание процесса сегментации аудио на фрагменты для обучения. • Умение нормализовать аудиосигнал. • Понимание токенизации текста и выбора единиц распознавания (фонемы, BPE, слова). • Умение формировать обучающие пары "звук — текст".
Тема 2. CTC-модели и выравнивание последовательностей: принцип обучения без временной разметки, роль blank-токена и механизм сопоставления аудио и текстовой последовательности.	• Понимание принципа обучения ASR без временной разметки (CTC). • Знание роли blank-токена в CTC. • Понимание механизма сопоставления аудиокадров и текстовой последовательности. • Умение реализовывать CTC-loss.
Тема 3. Seq2seq и трансформеры в ASR: архитектуры энкодер-декодер, механизм внимания и различия между авторегрессионными и неавторегрессионными моделями.	• Понимание архитектуры энкодер-декодер для распознавания речи. • Знание механизма внимания (attention) в ASR. • Понимание различий между авторегрессионными и неавторегрессионными моделями. • Умение использовать трансформерные модели для ASR (например, Whisper).
Тема 4. Декодирование и языковые модели: изучите жадное декодирование и beam search, разберёте влияние внешней языковой модели на итоговый текст и научитесь оценивать качество распознавания с помощью метрик WER и CER.	• Понимание жадного декодирования и Beam Search. • Понимание влияния внешней языковой модели на качество распознавания. • Умение оценивать качество ASR с помощью метрик WER и CER. • Навык декодирования выходных последовательностей модели в текст.
Проект	Разработка системы распознавания речи: подготовка датасета, реализация и обучение базовой модели, настройка декодирования и оценка качества по метрике WER и CER.

SSL и большие речевые модели 03

2 недели
10-15 часов

Изучите современные подходы к обучению речевых моделей без разметки и разберётесь, как устроены foundation-модели и Speech LLM. Научитесь дообучать self-supervised модели и сравнивать их качество с классическими архитектурами.

Содержание

Тема 1. Self-Supervised Learning в речи: предобучение на неразмеченном аудио и последующее дообучение на задаче распознавания, принципы работы wav2vec и HuBERT.	• Понимание концепции обучения без разметки (SSL) для аудио. • Знание принципов работы wav2vec 2.0 и HuBERT. • Понимание этапов: предобучение на неразмеченном аудио → дообучение на размеченном. • Умение сравнивать SSL-подходы с классическими архитектурами.
Тема 2. Архитектура wav2vec и HuBERT: контрастивное обучение, masked prediction, дискретизация представлений и отличие от обучения с нуля.	• Понимание контрастивного обучения (contrastive learning) в wav2vec. • Знание механизма masked prediction (предсказание замаскированных участков). • Понимание дискретизации представлений (quantization) в HuBERT. • Понимание отличий обучения SSL от обучения с нуля.
Тема 3. Fine-tuning и перенос на новые домены: дообучение на размеченных данных, влияние объёма корпуса и адаптация к новым задачам.	• Умение дообучать SSL-модель на размеченных данных для ASR. • Понимание влияния объёма тренировочного корпуса на качество. • Навык адаптации модели к новым доменам (акценты, шумы, тематики). • Умение сравнивать качество дообученной модели с базовой.
Тема 4. Speech LLM и аудио-conditioned модели: объединение речевого энкодера и языковой модели, обучение на инструкциях, мультимодальные архитектуры и диалоговые системы на основе речи.	• Понимание архитектуры объединения речевого энкодера с языковой моделью (LLM). • Знание принципов обучения на инструкциях (instruction tuning) для речи. • Понимание мультимодальных архитектур (аудио + текст). • Навык создания диалоговой системы на основе речи.
Проект	Дообучение self-supervised модели, сравнение качества с базовой архитектурой, анализ различий и подготовка итогового ASR-скрипт, принимающего аудио и возвращающего текст.

Синтез речи и управление характеристиками голоса

04

2 недели
10-15 часов

Изучите современные архитектуры синтеза речи и поймёте, как формируется тембр, интонация и стиль голоса. Научитесь управлять характеристиками речи и разберётесь, как работают системы voice cloning и какие риски связаны с deepfake-аудио.

Содержание

Тема 1. Архитектура TTS-системы: полный конвейер от текста к аудиосигналу, преобразование текста в фонемы, предсказание спектрограммы и генерация звука.	• Понимание полного конвейера TTS: текст, фонемы, спектрограмма, аудио. • Знание преобразования текста в фонемы (графемно-фонемное преобразование). • Понимание этапов предсказания спектрограммы и генерации звука. • Умение различать компоненты TTS-системы.
Тема 2. Предсказание спектрограммы в FastSpeech: неавторегрессионный синтез, моделирование длительности фонем и формирование акустического представления речи.	• Понимание неавторегрессионного синтеза речи (FastSpeech). • Знание принципа моделирования длительности фонем (duration predictor). • Понимание формирования акустического представления речи из текста. • Умение использовать FastSpeech для синтеза.
Тема 3. Нейросетевые вокодеры и HiFi-GAN: преобразование спектрограммы в аудиосигнал и принципы работы современных вокодеров	• Понимание задачи вокодера: преобразование спектрограммы в аудиосигнал. • Знание принципов работы современных вокодеров (HiFi-GAN, WaveGlow). • Умение применять предобученный вокодер для генерации звука. • Понимание отличий вокодеров от классических методов (Griffin-Lim).
Тема 4. Управление голосом и стилем: использование speaker embeddings, работа с многоспикерными моделями и контроль темпа, высоты и тембра.	• Понимание использования speaker embeddings для управления тембром. • Умение работать с многоспикерными TTS-моделями. • Навык контроля темпа, высоты тона и интонации. • Понимание, как формируется стиль голоса.
Тема 5. Voice cloning и deepfake-аудио: zero-shot и few-shot клонирование голоса, ограничения технологии, риски использования и методы детекции синтетической речи.	• Понимание zero-shot и few-shot клонирования голоса. • Знание ограничений технологий клонирования голоса. • Понимание рисков использования deepfake-аудио. • Знание методов детекции синтетической речи.
Проект	Создание системы синтеза речи с управлением параметрами голоса и оцените качество полученного аудио.

Инференс и оптимизация аудио-моделей 05

2 недели
10-15 часов

Научитесь собирать модели в рабочий инференс контур, оптимизировать скорость их работы и объединять распознавание и синтез в единую цепочку.

Содержание

Тема 1. Экспорт модели в компактный формат: преобразование модели в ONNX и подготовка к оптимизированному запуску.	• Понимание необходимости экспорта модели для инференса. • Умение преобразовывать модель в формат ONNX. • Навык подготовки модели к оптимизированному запуску. • Понимание преимуществ и ограничений ONNX.
Тема 2. Ускорение работы модели на видеокарте: применение оптимизированной точности.	• Понимание применения оптимизированной точности (FP16, mixed precision) для ускорения. • Умение настраивать инференс модели с использованием GPU. • Навык сравнения скорости работы в FP32 и FP16. • Понимание компромисса между скоростью и точностью.
Тема 3. Сборка инференс контура ASR и TTS: объединение шагов обработки.	• Умение объединять ASR и TTS в единую цепочку обработки. • Навык интеграции всех шагов: загрузка аудио, распознавание , обработка текста, синтез ответа. • Понимание узких мест производительности в контуре. • Умение настраивать пайплайн для работы в реальном времени.
Тема 4. Тестирование и валидация готового контура: проверка стабильности.	• Понимание методов проверки стабильности инференс-контура. • Умение тестировать систему на различных входных данных. • Навык выявления и устранения ошибок в пайплайне. • Понимание метрик оценки производительности (latency, throughput).

Итоговый проект 06

1 неделя
10-15 часов

В рамках итоговой работы вы разработаете полноценную систему обработки речи, объединяющую распознавание, языковую обработку и синтез ответа. Проект включает сборку единого инференс-контура, оптимизацию модели и демонстрацию работы сервиса в формате прототипа.