Глубинное обучение в звуке 2024

Программа экзамена

На экзамене студенту случайно выбирается блок тем из 2-6, по которым его будут спрашивать. Времени на подготовку не даётся. В зависимости от выбранной темы выбирается принимающий, который обладает экспертизой в данной области. Перед ответом на выбранный блок обязательно происходит обсуждение блока 1, так как это является обязательной теоретической базой. Проверяющим могут задаваться не все вопросы в силу большого объёма материала в каждом блоке, с другой стороны, некоторые вопросы могут быть более обширными (так, к примеру, в вопросе про модель может потребоваться объяснить пайплайн инференса, обучения, архитектуру модели, используемый лосс и другие важные концептуальные аспекты). В любом случае, объём материала ограничивается лекциями, семинарами и домашними работами.
Статьи, прочитанные студентом самостоятельно для ДЗ, которых не было на лекциях, входить не будут.

Список вопросов

1. Блок Signal Processing. Цифровой и аналоговый сигнал, как происходит преобразование из одного в другой, зачем нужен один и зачем нужен другой. Основные характеристики аудио в цифровом формате: частота семплирования (+примеры), форматы (без деталей). Преобразование Фурье: зачем оно нужно для DL, как оно вычисляется (+формула), какую информацию о сигнале помогает извлечь, что такое амплитуда\магнитуда, фаза и частоты после преобразования Фурье, какое отношение они имеют к исходному сигналу. Спектрограммы, STFT, получение спектрограммы с помощью STFT, Mel-спектрограммы.

2. Блок Automatic Speech Recognition. Задача распознавания речи, примеры датасетов, используемых в речевых задачах. Метрики качества в задаче ASR. CTC Loss, формула его подсчета и оптимизация. Варианты инференса: greedy, beam search. Примеры архитектур с CTC Loss. Модели декодеров LAS, RNN-Transducer, обучение и инференс. Плюсы и минусы каждой из моделей. Использование сторонних языковых моделей для CTC, LAS, RNN-t моделей, как встроить, какую модель взять и чем может быть полезно. Задача Self-Supervised Learning. Модели Wav2Vec2.0 и HuBERT: архитектура, метод претрейна, как использовать в downstream задачах (случай ASR).

3. Блок Source Separation. Задача разделения аудио, классификация задач разделения, некоторые примеры датасетов. Задача денойзинга, основная идея решений, использующих спектрограммы. Модель DEMUCS, HT-DEMUCS, BandSplit-RNN. Использование комплексных сигналов (DCCRN, FullSubNet).
Permutation-Invariant Training. Модель TasNET. Модель DPRNN. Модель ConvTasNET. Модель SpEx+. Модель VoiceFilter. Разделение аудио в аудиопотоке, как меняется пайплайн инференса и обучения по сравнению со

случаем полного аудио.

4. Блок Self-Supervised Models и Audio-Visual systems. SSL может также встречаться как часть ASR или Voice Biometry блоков. Wav2Vec, Audio/Visual/Cross-Modal/Audio-Visual HuBERT (+может быть вопрос о том как и где применять, см. ASR/Voice Biometry). Мотивация Audio-Visual в ASR/SourceSeparation. Audio-Visual Source Separation: подходы к Fusion, CTCNet, RTFSNet (может включать вопросы из раздела Source Separation). Lyp-Sync with Wav2Lip

5. Блок TextToSpeech. Постановка задачи TTS. Из каких частей состоит система. Метрики и методы оценки качества. Механизм внимания, GST, FastSpeech, WaveNet, Parallel WaveGAN, MelGAN, HiFiGAN. Лоссы в синтезе: STFT, Feature Matching. TTS with Diffusion: WaveGrad, DiffWave, GradTTS

6. Блок VoiceBiometry. Постановка задачи, метрики, ASV vs CM vs SASV. Виды атак, артефакты при синтезе, артефакты при перезаписи. Fusion. LCNN, RawNet2. CM для синтезированной речи, CM для перезаписи. Виды SoftMax лоссов. ECAPA-TDNN. RawNet1-3 (+виды time-frequency преобразований). SincNet (+слой, +формула). Методы получения SASV систем.