ИМШ: ответы на вопросы

Каменев Матвей

Февраль 2025 г.

1. Что такое обучение с учителем и без учителя, можете ли вы привести пример, связанный с обработкой аудио?
Обучение с учителем – тренировка модели на размеченном датасете, то есть наборе данных который уже включает target-значения. Например, я использовал обучение с учителем для классификации музыки по жанрам.
Обучение без учителя, соответственно, – тренировка модели на неразмеченном датасете, содержащем лишь сами дата-поинты, без target-значений. Так, например, в этой статье используют обучение без учителя для feature learning для последующей классификации аудио.

2. Что означает переобучение в машинном обучении и как это может повлиять на аудиомодели? Как с этим бороться?
Переобучение – явление в машинном обучении, при котором модель настолько "запоминает" тренировочные данные, что плохо перформит на данных реального мира.
Представим, что для задачи классификации пола диктора по голосу, мужчин записывали в одном шумном окружении, а женщин – в другом. Тогда при переобучении такая модель-классификатор может воспринимать фоновый шум тренировочных данных (разнящийся в записях двух разных классов) за реальные фичи сигнала, что, конечно, приведет к плохой генерализации.
Бороться с переобучением можно с помощью ранней остановки, аугментации данных, добавления случайного шума, регулязации входных данных.

3. Что такое batchnorm, layernorm? Когда лучше одно, когда другое?
Batchnorm – нормализация входных данных каждого слоя, т.е. нормализуем для всех семплов каждую фичу независимо относительно батча.
Layernorm – нормализация по фичам в каждом семпле.
То есть, layernorm используем когда размер батча маленький и важно уловить контекстуальные отношения фичей, хорошей областью применения могут послужить архитектуры RNN и Transformer в сфере NLP.
Batchnorm принято использовать когда размер батча достаточно большой, а данные приходят из одной выборки – то есть, отлично подходит для архитектуры CNN.

4. Как можно бороться с проблемой far-field?
Фильтрацией сигнала для удаления нежелательных частот, экстракцией из него значимых признаков. Также помочь может техника beamforming.

5. В чем минус LSTM, GRU? Почему такие сетки учатся сложно?
Требовательны с вычислительной точки зрения. Их обучение завязано на обратном распределении градиента через время, в каждый момент которого у нас есть несколько гейтов, да и кроме того, само их использование подразумевает большие последовательности входных данных, всё это усложняет процесс подбора гиперпараметров и замедляет процесс тренировки.

6. Как можно решать задачу шумоочистки в потоковом аудио? Как формулируется математически, в чем простота по сравнению с деревербераций?

Задача шумоочистки решается с помощью фильтрации звукового сигнала или методами ML.
Согласно этой статье, звуковой сигнал может быть представлен как сумма композиции ревербации звука и его источника и шума: $x = h \bigotimes y + y^{(n)}$. Кроме того, шум чаще всего не имеет временной зависимости в отличие от реверберации.

7. Как можно решать задачу деревербераций в потоковом аудио? В чем сложность? Является ли в принципе задача разрешимой?
Методами машинного обучения. Основная сложность – в отделении реверберации и сухового источника в смешанном сигнале. Задача является разрешимой, но с некоторыми ограничениями – есть много факторов, влияющих на эффект искажения сигнала реверберацией, и для них сложно спроектировать универсальное решение.

8. Как фоновый шум может повлиять на системы распознавания речи и как его можно уменьшить?
Уменьшает точность инференса, если система распознавания речи не проводит процесс шумоочистки, так как вносит изменения в фичи значимой части сигнала. Уменьшить влияние шума можно с помощью фильтров и методов машинного обучения.

9. Какие фичи вы бы выбрали для модели анализа аудио?
В зависимости от задачи наиболее полезными можно считать фичи MFCC, спекрального центроида, ширины спектра, zero-crossing rate и другие.

10. Как можно реализовать обработку аудио в реальном времени в машинном обучении? Какие возникают ограничения? На каком размере окна можно предложить проводить анализ?
Думаю, для обработки аудио в реальном времени больше всего подходит архитектура LSTM, т.к. она справляется с задачей поиска временных зависимостей, и, кроме того, разделяет долговременно важные фичи от кратковременных. Ограничения возникают в скорости инференса, качества выходного сигнала. Анализ обычно проводится на размере окна в 20-50 мс.

11. Как бы вы предобработали аудиофайлы для модели машинного обучения? Какие известны подходы для аугментаций?
Процесс предобработки часто включает в себя применение фильтров, конвертацию сигнала в какой-либо вид спектограммы (если используем архитектуру CNN). Аугментировать имеющиеся данные можно с помощью добавления случайного шума, смещения звука, изменения частот сигнала, использования алгоритмов симуляции реверберации.

12. В чем разница между CNN и RNN, и какой из них вы бы использовали для анализа аудиосигналов? Какие есть ограничения у каждого из подходов?
CNN обрабатывают входные данные через слои свертки и пулинга, постепенно уменьшая их размерность и выделяя фичи; RNN обрабатывают последовательности данных, позволяя использовать предыдущие аутпуты в качестве инпутов будущих моментов времени. И та, и другая архитектура часто применяется в анализе аудио, например, CNN – для классификации, а RNN – для распознавания речи.
CNN не справляется с выделением временных зависимостей и часто требует большей предобработки данных, а RNN часто страдает от проблемы исчезающего градиента и долго обучается. В обработке аудио эти архитектуры часто используют вместе, как, например, в этой статье: архитектура представленной модели включает в себя как слои свертки, так и GRU.

13. Опишите назначение быстрого преобразования Фурье (FFT) в обработке аудио.
Быстрое преобразование Фурье – алгоритм, позволяющий вычислительно эффективно применять функцию преобразования Фурье к сигналу. Преобразование Фурье используется для декомпозиции сложного звукового сигнала на множество волн, которые его образуют, что помогает при экстракции

его характеристик.

14. Объясните, что такое мел-частотные кепстральные коэффициенты (MFCC) и почему они используются в обработке аудио.
MFCC – набор характеристик звукового сигнала, получаемых через его обработку с помощью преобразования Фурье, фильтрации по мел-шкале и кепстрального преобразования. Они используются в обработке аудио из-за того, что основаны на принципах восприятия звука человеческим ухом, т.е. помогают выделить наиболее значимые фичи сигнала.