Разбор кейса
В поисках интересного

Задача

Разработать концепцию архитектуры сервиса
по выделению наиболее интересного фрагмента трека
для пользователя с целью сокращения
длительности аудио контента.

Бизнес-метрики

- Индекс удовлетворенности CSI
- Количество пользователей за период

Основная гипотеза для MVP

Наиболее популярные жанры музыки имеют структуру треков с повторяющимися частями, в которых сконцентрированы самые запоминающиеся ходы и приемы (хуки) — в припевах. При этом зачастую в музыке непосредственно перед припевом имеются гармонические ходы с нарастанием напряжения перед последующим разрешением и наиболее эмоционально заряженные строчки в песне.

Будем считать часть перед припевом и припев самым интересным местом.

Таким образом, задача сводится к Music Track Structure Segmentation/Classification и поиску повторяющихся частей песен с последующей разметкой сниппета.

Статистика по популярности жанров: 1, 2, 3

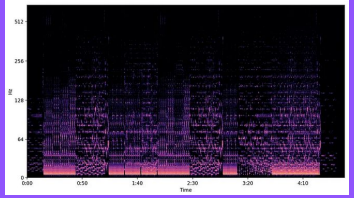
Papers and Technologies Research

- Finding Choruses in Songs with Python на основе статьи
- Audio-Based Music Structure Analysis: Current Trends, Open Challenges, and Applications
- Music Segmentation PhD Defense
- Automatic Structural Segmentation of Music
- Youtube Most Replayed Part Prediction
- и так далее…

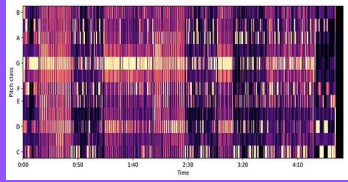
Non-ML Baseline

Идея: Разметка структуры трека на основе анализа повторения нот — находим припев с помощью матриц подобия и размечаем необходимый отрывок. *[источник]*

Трек → Спектрограмма частот



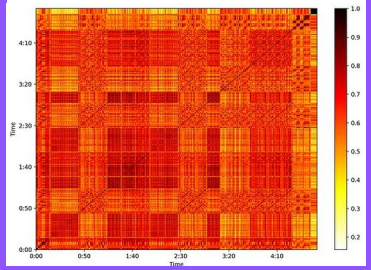
→ Хроматограмма нот, сжатая до одной октавы



→ Считаются матрицы подобия каждого фрейма трека

$$1 - \frac{\|v_1 - v_2\|}{\sqrt{12}}$$

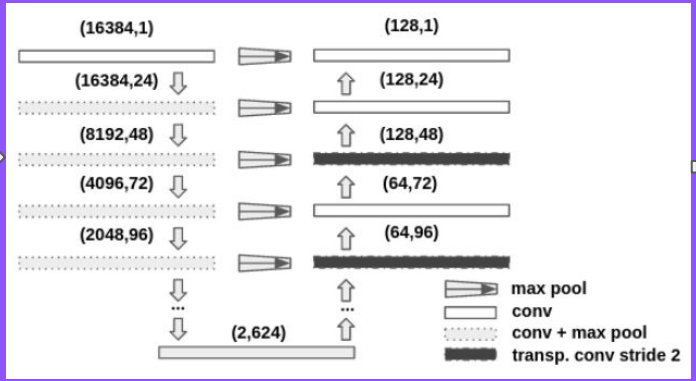
→ Похожие секции выделяются и размечают структуру трека



Pros	Cons
- Простота и дешевизна: не нужны вычисления на видеокартах и разметка	- Возможна низкая точность, т.к. выделяются повторенные один в один части трека - Низкая скорость инференса

CV Solution - Hook-net

Идея: После выделения фичей из аудио, обучаем сверточную нейросеть на основе U-net — Hook-net, которая классифицирует отрезки треков по частям и проводит сегментацию. *[источник]*

Трек → Разбиение на 16384 семплов → 

→ Бинарная классификация 128 регионов в треке

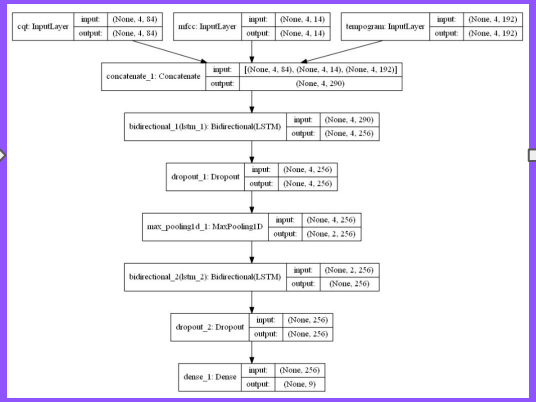
Pros	Cons
- Учитывается структура трека	- Накладные расходы выше Non-ML Baseline

RNN Solution - LSTM

Идея: аудиоинформация и, в частности, музыка может быть обработана не только как изображения, но и как последовательность — возможно применение NLP подходов — Bi-Directional LSTM сеть так же справляется с задачей сегментации и анализа структуры треков. *[источник]*

Трек

Получение признаков методами DSP:
Constant-Q Transform
Mel-Frequency Cepstral Coefficients
Tempogram



Мультиклассовая классификация каждого фрейма последовательности

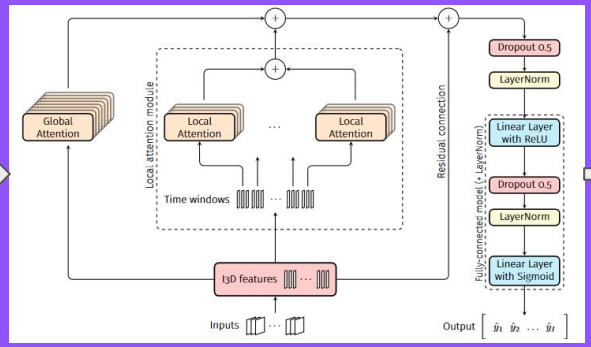
Pros	Cons
- Теоретически может учитывать контекст повторяющихся структур внутри трека при поиске самого интенсивного припева, а не просто размечать структуру	- Более тяжелая архитектура сети

Transformer Solution - Attention-based PGL-SUM

Идея: Задача выделения наиболее просматриваемого фрагмента была решена для видео с помощью архитектуры PGL-SUM, основанной на двух контурах Attention — глобальной и локальной. Возможно данная архитектура справится с аналогичной задачей для аудио. *[источник]*

Трек

Получение признаков методами DSP:
Constant-Q Transform
Mel-Frequency Cepstral Coefficients
Tempogram
…



Определение интересного региона

Pros	Cons
- Двойной контур Attention должен улучшить качество разметки по сравнению с LSTM	- Еще более тяжелая архитектура сети

Метрики

ML-метрики для оценки качества и мониторинга:

1. IoU — по перекрытию ground truth сегмента и предсказания
2. Accuracy
3. Precision, Recall, F1-score

Полезные показатели для мониторинга при A\B-тестировании гипотез:

1. Mean Opinion Score — формировать по опросам пользователей
2. Conversion Rate — конверсия прослушивания трека после прослушивания сниппета (возможно потребуется учитывать предпочтения пользователя для повышения адекватности показателя)

Датасет

1. Открытые датасеты
2. Разметка на Я.Толоке
3. Youtube scraping — возможно получение данных о Most Reviewed Part через API [Stackoverflow thread]
4. Генерация примеров
5. Аугментация для увеличения датасета:
   a. Добавление шумов и глитчирование
   b. Изменение тональности
   c. Перемешивание частей трека (в случае полностью размеченной структуры)

Доступные датасеты

Датасет	Количество треков	Features	Примечание
FMA	106,574	DSP фичи, полученные с помощью librosa	Необходима разметка, 30-секундные треки
SALAMI	1359	Иерархические аннотации структуры треков	Ограничен по исполнителям и жанрам
The Harmonix Set	912	DSP фичи, метадата, структура	Ограничен по исполнителям и жанрам
SPAM	50	DSP фичи, метадата, структура	Ограничен по исполнителям и жанрам
RWC	300	DSP фичи, метадата	Платный, нет разметки структуры
Isophonics	300	DSP фичи, метадата, структура	Ограничен по исполнителям и жанрам

Яндекс.Толока Toloka

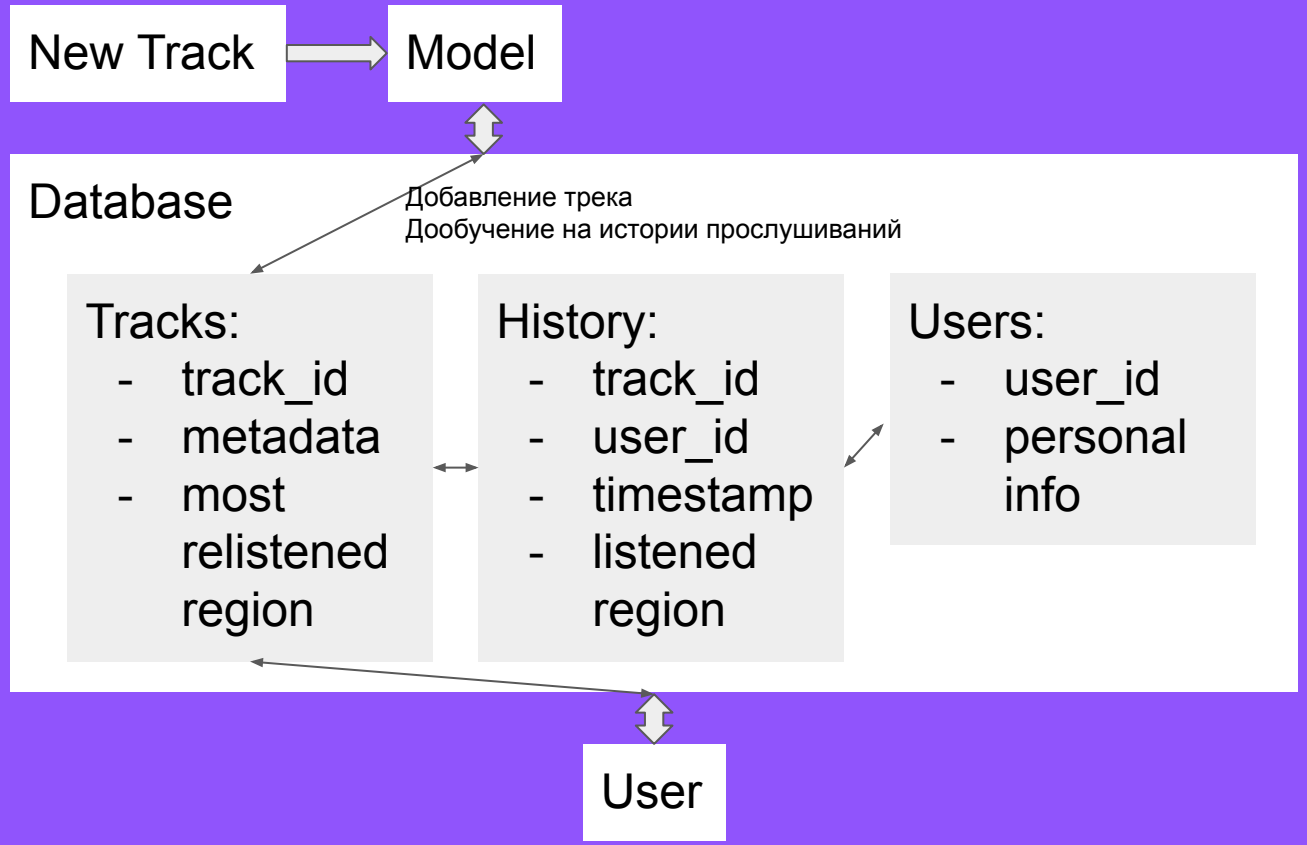
1. Разметка структуры всего трека скорее всего будет требовать толокеров высокого уровня и займет не только больше времени, но и будет больше стоить. Возможно рациональней будет упростить задачу и просить толокеров определить только наиболее интересный регион трека по их мнению.
2. При этом необходимо оставить возможность для выбора у толокера — какие-то жанры могут быть нелюбимыми и разметка интересного региона в таком случае будет необъективной. Вопрос объективности при опросах возможно решить методами статистики [Пример].
3. Пулы следует формировать с большим количеством перекрытий, чтобы агрегация региона после разметки была наиболее объективна.
4. Набор треков должен быть сбалансирован по жанрам и возможно составлен из жанровых топ-листов для упрощения задачи толокерам.

Я.Толока и разметка аудио

Возможные трудности

Описание	Вариант решения
Жанры со сложной структурой	Техно, джаз, академические жанры имеют структуру отличную от популярных и выделение фрагмента на основе повторяющихся сегментов может вызвать трудности. Для подобных жанров можно выделять фрагмент на основе простых эвристик — начало\середина трека, например, пока не будет статистики их реального прослушивания для дообучения модели.
Различная длина треков	Для треков короче выделяемого сниппета — выделяется весь трек. В остальных случаях для подачи в сеть требуется последовательность фиксированной длины. Тогда принимается величина продолжительности трека, к которой будут нормализоваться все отправляемые нейросети треки. При этом треки короче этой величины будут иметь частоту высокую семплирования, а длинней, соответственно, низкую. Для очень длинных треков, возможно, проще будет разбивать трек на части и искать по ним, или же использовать иной подход.

Архитектура системы



Технологии

librosa
Pytorch
PostgreSQL
Airflow
MLflow
DVC

Оценка сроков реализации

1. Сбор, подготовка и очистка датасета: ~1 месяц
2. Реализация и эксперименты с архитектурами моделей: ~2 месяца
3. Создание MVP для Demo Review: ~1 месяц

Итого: ~4 месяца на реализацию MVP

Оценка ресурсов

1. Аренда Яндекс.Облако для обучения\инференса модели: ~75к/мес
2. ФОТ: ~100-150к/мес на сотрудника + отчисления
3. Оборудование (ноутбук для удалённой работы): ~100-150к на сотрудника

Спасибо!