УДК 004.9

ИССЛЕДОВАНИЕ СОВРЕМЕННЫХ ТЕХНОЛОГИЙ АНАЛИЗА МУЗЫКАЛЬНЫХ ПРОИЗВЕДЕНИЙ

Игнатов К. О. (Университет ИТМО)
Научный руководитель - кандидат технических наук Жданов А. Д. (Университет ИТМО)

Введение. В современном мире, в эпоху, когда музыкальные стриминговые сервисы преобладают над остальными подходами к прослушиванию музыки, необходимо использовать подходы на основе современных технологий для создания качественных систем рекомендаций и других утилит для улучшения пользовательского опыта в контексте прослушивания музыки. Например, системы распознавания музыки, такие, как Shazam, а также системы для работы с музыкой не с позиции обывателя, а с точки зрения музыканта, - подобные системы могут помочь определить тональность, темп и т.д., иными словами – разбирать композиции на составляющие, чтобы создавать свои аранжировки, кавера или проводить исследования. Алгоритмы рекомендаций же должны быть построены так, чтобы рекомендованные музыкальные композиции имели схожий жанр, темп, стиль, атмосферу и даже лирику, если таковая есть. Для эффективной работы перечисленных систем необходимо задействовать весь спектр современных технологий, который включает как новые подходы, в лице машинного обучения и анализа данных, так и технологии, и подходы, которые укоренились в индустрии уже давно, например, различные методы анализа спектрограмм.

Основная часть. В настоящее время существует два основных направления исследования алгоритмов анализа музыкальных треков – это традиционные алгоритмы и алгоритмы, основанные на нейронных сетях. Традиционные методы как правило основываются на алгоритмах аудио-отпечатков при помощи которых каждой композиции сопоставляется “цифровая подпись” для дальнейшего поиска в базе данных. Алгоритмы, основанные на нейронных сетях, базируются на анализе спектрограмм, сырых аудио и символьных данных, например, MIDI. Для анализа спектрограмм используются архитектуры глубокого обучения [1], одной из самых популярных является ResNet [2], а также, методы машинного обучения без учителя, основывающиеся на моделях wav2vec2, позволяющие извлекать эмбеддинги для последующего анализа.

Для начала рассмотрим один из методов на основе аудио-отпечатков, - это метод 2D-пиков спектрограммы, используемый приложением Shazam для поиска песни по ее звучанию. Перед тем, как перейти к самому алгоритму, необходимо из аудиозаписи получить спектрограмму, это делается при помощи быстрого преобразования Фурье [3]. После этого можно переходить к самому алгоритму, суть которого заключается в поиске 2D-пиков на спектрограмме [4]. Данный метод состоит из нескольких этапов – сначала спектрограмма делится на ячейки по времени и частоте для облегчения анализа, после разбиения идет поиск локальных максимумов, максимумы ищутся по величине амплитуды в окрестности. Важным этапом в данном подходе является исключение незначительных колебаний за счет определения порога, значения ниже которого будут исключаться. Данный этап является важным, так как делает аудио-отпечатки устойчивыми к шумам и прочим помехам. Заключительными этапами алгоритма 2D-пиков спектрограммы является формирование пар пиков, т.е. каждому пику сопоставляется пара в определенном частотном и временном диапазоне, в результате чего и получаются уникальные комбинации. После этого создаются хэши аудио-отпечатков для дальнейшего поиска в базе данных, хэш-код включает в себя частоту первого пика в паре, частоту второго и временной интервал между пиками. Среди преимуществ данного подхода можно выделить устойчивость к шумам, эффективность поиска благодаря хэшам и индексам и уникальность получаемых отпечатков.

Среди алгоритмов, основывающихся на нейронных сетях, можно выделить один из алгоритмов на основе эмбеддингов [5]. Суть этого алгоритма заключается в извлечении признаков аудиофайла путем обработки предпоследнего слоя предобученной модели

Wav2Vec2, а после поиска похожих композиций при помощи алгоритма поиска k ближайших соседей. Он также делится на несколько этапов. Первый – .wav файл подается на вход модели Wav2Vec2 для последующего извлечения предпоследнего слоя. После извлечения предпоследний слой усредняется по временной оси и нормируется, размерность полученных векторов можно изменять, как гиперпараметр. Как только эмбеддинги получены для каждой композиции из интересующей библиотеки, похожие песни можно искать при помощи алгоритма поиска k ближайших соседей, для поиска релевантных песен достаточно просто посмотреть на расстояния между полученными векторами. Одними из главных преимуществ данного подхода можно считать эффективность за счет быстрой работы алгоритма поиска ближайших соседей и гибкость, благодаря возможности расширять систему, добавляя метаданные и слои обработки.

Выводы. Системы, направленные на анализ музыкальных композиций, могут быть построены на основе самых разных алгоритмов, исходя из своих нужд. Можно задействовать традиционные алгоритмы для определения конкретного трека по входному сигналу, можно применять машинное обучение для этих задач, например, для поиска похожих песен. Стоит отметить, что существующие подходы не являются совершенными. Так, например, метод 2D-пиков спектрограммы подвержен коллизии хэшей, если у каких-то песен будут схожие частотные характеристики, а метод эмбеддингов может быть не точен из-за того, что в качестве модели Wav2Vec2 может быть задействована модель, заточенная под задачи далекие от музыки, хотя по-прежнему работающая со звуком. Именно поэтому технологии анализа музыкальных произведений продолжают совершенствоваться по сей день.

Список использованных источников:

1. Girshick R., Donahue J., Darrell T., Malik J. Rich feature hierarchies for accurate object detection and semantic segmentation [Электронный ресурс] / R. Girshick, J. Donahue, T. Darrell, J. Malik. – Tech report (v5). – Berkeley: UC Berkeley, 2013. – 14 с. – Режим доступа: https://arxiv.org/pdf/1311.2524 (Дата обращения 15.11.24)
2. Zhang J. Music Genre Classification with ResNet and Bi-GRU Using Visual Spectrograms [Электронный ресурс] / J. Zhang. – arXiv preprint arXiv:2307.10773. – 2023. – Режим доступа: https://arxiv.org/pdf/2307.10773 (Дата обращения 23.11.24)
3. Neammalai P., Phimoltares S., Lursinsap C. Speech and Music Classification using Hybrid Form of Spectrogram and Fourier Transformation [Электронный ресурс] // Proceedings of the 10th International Conference on Signal Processing and Communication Systems (ICSPCS), Opatija, Croatia, 2015. – 1–6 p. – DOI: 10.1109/ICSPCS.2015.7041658. – Режим доступа: https://ieeexplore.ieee.org/abstract/document/7041658 (Дата обращения 10.11.24)
4. Froitzheim S. A Short Introduction to Audio Fingerprinting with a Focus on Shazam [Электронный ресурс] // MUS-17. – 5 июля 2017 года. – Режим доступа: https://hpac.cs.umu.se/teaching/sem-mus-17/Reports/Froitzheim.pdf (Дата обращения 13.12.24)
5. Koh E., Dubnov S. Comparison and Analysis of Deep Audio Embeddings for Music Emotion Recognition [Электронный ресурс] // arXiv preprint arXiv:2104.06517. – 13 апреля 2021 года. – Режим доступа: https://arxiv.org/abs/2104.06517 (Дата обращения 13.12.24)