Московский государственный технический университет им. Н.Э. Баумана
Факультет «Информатика и системы управления»
Кафедра «Системы обработки информации и управления»

Домашнее задание

на тему:

«Задача классификации музыкальных композиций по исполнителю с помощью методов машинного обучения с учителем»

по дисциплине «Методы машинного обучения»

ИСПОЛНИТЕЛЬ:

Попов М.А.

Группа ИУ5-24М

_____

"__"_____2022 г.

ПРОВЕРИЛ:

Гапанюк Ю.Е.

_____

"__"_____2022 г.

Москва, 2022

Введение

В настоящее время всё большую популярность приобретает задача классификации музыкальных композиций по жанру, настроению, наличию музыкальных инструментов определённого типа, а также по сходству композиций для выявления музыкальных предпочтений и рекомендаций. Онлайн-сервисы по прослушиванию музыки, такие как Spotify, Яндекс.Музыка, Apple Music и другие, активно внедряют технологии классификации музыки в свои приложения, предоставляя пользователям большой спектр возможностей по поиску музыки в соответствии с их настроением и вкусовыми предпочтениями.

Наилучших результатов в сфере распознавания и классификации аудио информации удалось достичь путём применения нейросетевых технологий, в частности методов глубокого машинного обучения.

В данной работе подробнее рассмотрена задача классификации музыкальных композиций по исполнителю, которая может быть использована в музыкальных библиотеках для автоматического распределения по разделам и поиска нужной композиции по названию исполнителя. Эта задача была рассмотрена в других работах по данной теме, при этом были использованы различные модели, подходы и параметры. Рассмотрим некоторый срез текущих знаний по этой теме, чтобы составить общую картину понимания данной задачи в научном сообществе.

Классификация музыкальных композиций

Классификация музыки — это задача поиска музыкальной информации (MIR) для классификации музыкальных элементов по определённому признаку или набору признаков [1]. Принято рассматривать музыкальную информацию как систему, состоящую из разнородных элементов: звука и его характеристик, текста, ритма и т.д. Каждый элемент характеризуется собственными специфическими параметрами, которые используются в качестве отличительных черт одной композиции от другой. Для успешного решения задачи классификации необходимо иметь возможность представить музыкальную композицию в формате, пригодном для извлечения необходимых классификационных признаков, и далее использовать полученные данные на входе некоторого классификатора, представляющего из себя многослойную нейронную сеть, способную на выходе предложить один из классов в рассматриваемом пространстве, как наиболее подходящий.

Модель классификации музыкальных композиций состоит из модуля предварительной обработки, внешнего блока и внутреннего блока. На этапе предварительной обработки модель извлекает различные входные представления. Внешний блок фиксирует локальные акустические характеристики, такие как тембр, высота звука или наличие конкретного инструмента. Затем внутренний блок суммирует последовательность извлеченных данных.

Способы представления аудиоданных

Для использования методов глубокого машинного обучения с целью классификации музыкальных композиций необходимо представить аудио информацию, содержащуюся в них, одним из подходящих способов,

позволяющих добиться максимально эффективного результата работы нейронной сети [2].

Большинство подходов к глубокому обучению при решении задачи поиска музыкальной информации используют двухмерное представление аудио информации вместо исходного одномерного представления, которое несёт в себе изначальный дискретный аудиосигнал. Часто этими двумя измерениями являются частотная и временная оси. При применении методов глубокого обучения к задачам MIR особенно важно понимать свойства представления аудиоданных. Обучение глубоких нейронных сетей требует больших вычислительных затрат, поэтому оптимизация необходима на каждом этапе. Одним из методов оптимизации является предварительная обработка входных данных.

Рассмотрим основные способы представления аудиоданных, представленных на рисунке 1:

- **Аудиосигнал**. Исходный звук, неподверженный обработке, остальные представления основаны на нём и получены путём его преобразования. Цифровой аудиосигнал состоит из звуковых выборок, которые задают амплитуды с временными шагами. Предполагается, что музыкальный контент передается в виде цифрового аудиосигнала без воздействия акустических каналов. Для обучения нейронной сети с использованием исходного аудиосигнала требуется очень большой набор данных.
- **Кратковременное преобразование Фурье** (STFT) обеспечивает частотно-временное представление с линейно разнесенными центральными частотами. Вычисление STFT происходит быстрее, чем других частотно-временных представлений, благодаря быстрому преобразованию Фурье (FFT), которое снижает стоимость относительно количества точек FFT. Линейные центральные частоты не всегда

желательны при анализе музыки. Они не соответствуют частотному разрешению слуховой системы человека и музыкально не обоснованы. По этой причине STFT не является самым популярным выбором представления при использовании глубокого обучения. Это не самый эффективный способ представления. Одним из достоинств STFT является то, что он инвертируем к аудиосигналу.

- **Mel-спектрограмма** — это 2D-представление, оптимизированное для слухового восприятия человека. Оно сжимает STFT по частотной оси и, следовательно, может быть более эффективным по размеру, сохраняя при этом наиболее важную для восприятия информацию. Mel-спектрограмма показывает только величину (или энергию) частотно-временных интервалов, что означает, что она не обратима к звуковым сигналам.
- **Преобразование постоянной добротности** (CQT) обеспечивает 2D-представление с центральными частотами логарифмического масштаба. Это хорошо согласуется с частотным распределением высоты тона, поэтому CQT преимущественно используется там, где необходимо точно определить основные частоты нот. Вычисление CQT сложнее, чем вычисление спектрограммы STFT или Mel.
- **Хромаграмма**, также часто называемая профилем класса высоты тона, обеспечивает распределение энергии по набору классов высоты тона. Можно рассматривать хромаграмму как представление CQT, откладываемое по частотной оси. Хромаграмма более "обработана", чем другие представления, и может использоваться как функция сама по себе.

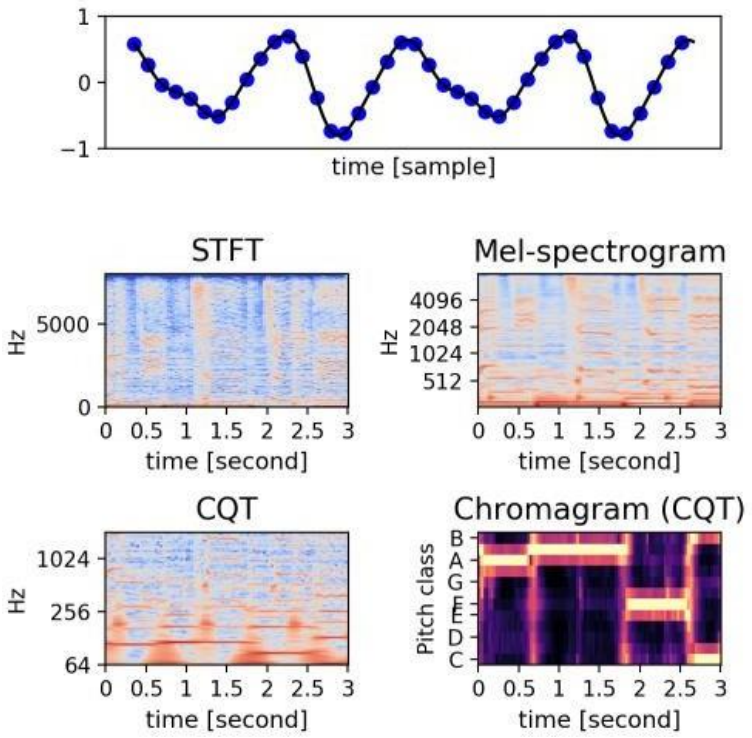

Рисунок 1 — Способы представления аудиосигнала.

Архитектуры нейронных сетей

Рассмотрим архитектуры нейронных сетей для обучения с учителем, наиболее часто использующиеся и показавшие максимальную продуктивность при решении задачи классификации музыкальных композиций [1]. Парадигма обучения с учителем полностью полагается на существование базовой истиной выборки, используемой для обучения и решения задач классификации. Нам нужна исходная информация для обучения модели нейронной сети, называемая размеченным набором данных. Чтобы достичь адекватной производительности, глубоким нейронным сетям часто требуется большое количество многообразных размеченных данных. Необходимо учитывать, что создание большого набора данных — это дорогостоящий и сложный процесс.

- **Полностью сверточные сети (FCN)**

Благодаря значительным успехам сверточных нейронных сетей (CNN) в области компьютерного зрения исследователи MIR использовали успешные архитектуры для решения проблем автоматической классификации музыки. Полностью сверточная сеть (FCN) является одним из ранних подходов глубокого обучения для анализа музыки, который включает в себя четыре сверточных слоя. За каждым слоем следует нормализация (batch normalization), нелинейность исправленной линейной единицы (ReLU) и слой пуллинга (max-pooling). Сверточные фильтры 3x3 используются для захвата спектрально-временных акустических характеристик входной спектрограммы Mel.

- **Сети типа VGG / Short-chunk CNN**

Эти архитектуры очень похожи на FCN, за исключением входных данных. Вместо изучения представления на уровне песни они используют обучение на уровне экземпляра (на уровне фрагмента). Поскольку при этом длина ввода короче, чем у FCN, такие модели не требуют увеличения размера восприимчивых полей с разреженными шагами. Вместо этого, например, Short-chunk CNN (сеть короткого фрагмента) состоит из 7 сверточных слоев с dense max-pooling (2, 2), что соответствует звуковому фрагменту длиной 3,69 с. Когда входные данные становятся длиннее, модель суммирует функции с использованием max pooling.

- **Гармонические CNN**

Сверточные модули гармонических CNN идентичны модулям CNN с коротким фрагментом, но используют несколько другие входные сигналы. Гармонические CNN используют преимущества обучаемых полосовых фильтров и гармонически сложенных входных сигналов частотно-временного

представления. В отличие от фиксированных блоков фильтров Mel, обучаемые фильтры обеспечивают большую гибкость модели. Гармонически сложенное представление также обеспечивает спектрально-временную локальность, сохраняя гармонические структуры через канал входного тензора в первом сверточном слое.

- **MusiCNN**

Вместо использования фильтров 3x3 авторы MusiCNN предложили использовать созданные вручную формы фильтров для разметки музыки. Предположим, что оси x и y соответствуют времени и частоте. Вертикально удлиненные фильтры предназначены для улавливания тембральных характеристик, в то время как горизонтально удлиненные фильтры - для улавливания временного потока энергии, который, вероятно, связан с ритмическими паттернами и темпом.

- **CNN на уровне сэмплов**

CNN на уровне сэмплов и его варианты обеспечивают сквозную автоматическую разметку музыки, напрямую используя необработанные звуковые сигналы в качестве входных данных. В этой архитектуре используются фильтры свертки 1x2 или 1x3 (1D). Каждый слой состоит из 1D свертки, пакетной нормализации и нелинейности ReLU. Ступенчатая свертка используется для увеличения размера воспринимающего поля.

- **Сверточные рекуррентные нейронные сети (CRNN)**

В отличие от ранее представленных моделей уровня экземпляра, сверточные рекуррентные нейронные сети предназначены для представления музыки в виде длинной последовательности из нескольких экземпляров. Эту архитектуру можно описать как комбинацию CNN и RNN. Интерфейс CNN фиксирует локальные акустические характеристики (на уровне экземпляра), а внутренний модуль RNN суммирует последовательность функций на уровне

экземпляра. CRNN можно описать как модифицированный CNN, заменив последние сверточные слои на RNN [3].

- **Трансформатор музыкальных тегов (Music tagging transformer)**

Идея архитектуры идентична модели CRN. Внешний модуль фиксирует локальные акустические характеристики, а внутренний суммирует последовательность. В области обработки естественного языка трансформатор показал свою пригодность для моделирования длинных последовательностей с использованием слоев самоконтроля. Music tagging transformer использует внешний модуль CNN и внутренний модуль Transformer, который эффективно обобщает функции на уровне экземпляра.

Классификация по исполнителю с помощью CRNN

В работе [4] авторы отмечают, что сверточные рекуррентные нейронные сети обеспечивают наилучшую производительность в жанровой классификации среди известных архитектур классификации звука. Авторы предлагают адаптировать модель CRNN для создания основы глубокого обучения для классификации композиций по исполнителю.

Набор данных для идентификации музыкального исполнителя artist20 используется для оценки эффективности классификации. Он содержит шесть альбомов двадцати исполнителей, охватывающих широкий спектр музыкальных стилей.

Разделение на обучающую и тестовую выборки в классификации по исполнителю является важным этапом. Необходимо, чтобы фрагменты тестовых песен не использовались при обучении, а также обязательно нужно учитывать эффект продюсера, при котором наблюдается завышенная

эффективность классификации в наборах данных, разделенных по песням, из-за того, что детали производства могут сильнее влиять на классификацию по сравнению с музыкальным стилем. Стандартный подход для борьбы с этой проблемой заключается в разделении набора данных по альбомам таким образом, чтобы тестовый набор состоял исключительно из песен из альбомов, не используемых в обучении. Однако любая модель, обученная в соответствии с этой парадигмой, не будет устойчива к изменениям в музыкальных стилях на разных альбомах. Кроме того, детали производственного уровня, связанные с альбомом, также могут рассматриваться как часть уникального стиля исполнителя. По этим причинам авторы данной работу используют сплит-тесты как песен, так и альбомов и сравнивают результаты экспериментов.

В этой работе спектрограммы создаются для всей длины каждой песни, чтобы сформировать начальный набор данных. Этот набор данных подвергается разделению на 90/10 с разбивкой песен по исполнителям для создания обучающих и тестовых данных соответственно. Затем набор данных для обучения разделяется с использованием того же стратифицированного разделения 90/10 для создания подмножеств для обучения и валидации. Стратификация гарантирует, что каждый набор содержит эквивалентное количество песен от каждого исполнителя. Для разделения альбомов два альбома от каждого исполнителя случайным образом удаляются из исходного набора данных — один добавляется к тестовому набору, а другой - к валидационному. Оставшиеся четыре альбома каждого исполнителя объединяются, образуя обучающую выборку.

Кратковременное преобразование Фурье применяется к необработанному аудиосигналу для каждой песни с целью создания спектрограмм. Эти операции считаются стандартными методами обработки звука с целью повышения производительности в задачах классификации. Частота

дискретизации устанавливается на указанное значение для звуковых дорожек в наборе данных artist20.

Спектрограммы каждой песни разделяются на обучающие, тестовые и валидационные наборы данных. После разделения каждая песня разбивается на аудиоклипы длиной t, которая варьируется в ходе исследования. Также вместо того, чтобы использовать один клип на песню, во время обучения и валидации используется вся песня целиком. Преимущество данного подхода заключается в том, что он дает большее количество обучающих сэмплов и позволяет экспериментировать с прогнозами уровня песен. Кроме того, существует компромисс между размером обучающего набора и длиной каждого аудиоклипа. Более длинные клипы содержат больше временной структуры в каждой обучающем экземпляре, в то время как более короткие клипы могут быть перетасованы и интерпретированы как больший набор независимых обучающих примеров.

В работе [4] была адаптирована архитектура CRNN из работы [3] по классификации жанров. Авторы предположили, что эта архитектура также будет хорошо работать для классификации исполнителей, поскольку понимание музыкального стиля включает в себя характеристику того, как частотный контент меняется с течением времени. Учитывая, что эта информация содержится в спектрограмме, идеальная архитектура сети должна быть способна суммировать шаблоны по частоте (где первенствуют сверточные слои), а затем также рассматривать любые результирующие временные последовательности в этих шаблонах (где первенствуют повторяющиеся слои). CRNN содержит оба этих компонента. Архитектуру в целом можно разделить на три части: сверточные, рекуррентные и полностью связанные слои. CNN использует 2-слойную RNN с закрытыми повторяющимися блоками (GRU) для суммирования временных паттернов поверх двумерных 4-слойных CNN. В подструктуре CNN размеры сверточных

слоев и слоев с максимальным объединением составляют 3×3 и (2×2)-(3×3)-(4×4)-(4×4). Это приводит к размеру карты объектов N×1×15 (количество карт объектов×частота×время). Они подаются в 2-слойный RNN, из которого последнее скрытое состояние подключается к выходу сети [3].

Идентификацию исполнителя можно рассматривать как задачу многоклассовой классификации. В своих экспериментах авторы выбрали в качестве функции потерь категориальная кросс-энтропия. Adam был выбран в качестве оптимизатора, потому что он показал хорошую производительность в задачах классификации на основе свертки с ограниченной настройкой гиперпараметров. Однако установленная по умолчанию скорость обучения (0,001) снижена на порядок для повышения стабильности обучения. Ранняя остановка также используется с параметром 10, чтобы смягчить переобучение.

Модель CRNN, используемая в этом исследовании, обучается на аудиоклипах длиной {1 секунда, 3 секунды, 5 секунд, 10 секунд, 20 секунд, 30 секунд} в различных условиях, таких как тип разделения {песня, альбом} и уровень функции {фрагмент, песня}.

Для разделения по альбомам результаты проведённых тестов на одной секунде звука изначально разочаровывают, но улучшаются по мере добавления временной структуры к представлению объектов. Через тридцать секунд средние и лучшие показатели, составляющие 0,603 и 0,612 соответственно, демонстрируют преимущества спектрограммы звуковое представление за счет улучшения базовой линии. По мере добавления большей временной структуры улучшается производительность прогнозирования. Эту закономерность также можно увидеть в результатах с разделением на основе песни, за исключением того, что производительность прогнозирования лучше, чем базовые показатели уровня кадра при любой длине клипа. Однако мы наблюдаем, что средняя производительность
начинает снижаться через десять секунд, в отличие от разделения по

альбомам. Это говорит о том, что, хотя дополнительные временные данные приносят пользу, модель может быть переоснащена при разделении на песни или что она выигрывает от наличия больший обучающей выборки с большим количеством коротких независимых фрагментов. Эти выводы также подразумевают, что избыточная временная информация теряется, вероятно, на ранних уровнях объединения, когда выборка проходит через сеть.

Несоответствие между разделением по песням и альбомам также подтверждает эффект продюсера: результаты тестирования намного лучше при оценке неиспользованных для обучения песен по сравнению с неиспользованными альбомами. Важно отметить, что способ созданияальбома также может рассматриваться как часть стиля исполнителя и имеет значение в определенных контекстах. Следовательно, для общего назначения,желательна высокая производительность при обоих типах разделения.

Наиболее результативный трехсекундный (3 секунды) кейс достигает средних и лучших результатов в 0,937 и 0,966 метрики точности соответственно. Такая модель может быть полезна для реальных приложений. Однако, когда продолжительность звука увеличивается более чем на три секунды, производительность начинает уменьшаться, и это, вероятно, связано с тем, что подавление шума при голосовании с использованием большего числатестовых выборок перевешивает дополнительную временную выгоду. Однако использование слишком короткого аудио фрагмента ограничивает способность модели распознавать исполнителей, в то время как более длинные фрагменты снижают вероятность смягчения ошибочных прогнозов путем голосования. На практике мы рекомендуем использовать дискографию исполнителя для обучения и оценки уровня песни с аудио сэмплами продолжительностью от трех до десяти секунд, если это возможно.

На рисунке 2 разделение классов по исполнителю, полученное с помощью метода t-распределенного стохастического вложения соседей (t-SNE),

представлено визуально. В данном случае модель обучена ной с продолжительностью звука в десять секунд. Видим, что свертка и рекуррентные слои в модели CRNN создают эффективное представление данных для того, чтобы отличить исполнителей друг от друга. Даже в двух измерениях большинство аудио сэмплов хорошо разделены и образуют кластеры, которые однозначно описывают конкретного исполнителя. Существование шумов может быть связано с тем фактом, что у исполнителей скорее всего будут песни или, по крайней мере, звуковые сегменты в песнях, которые похожи друг на друга.

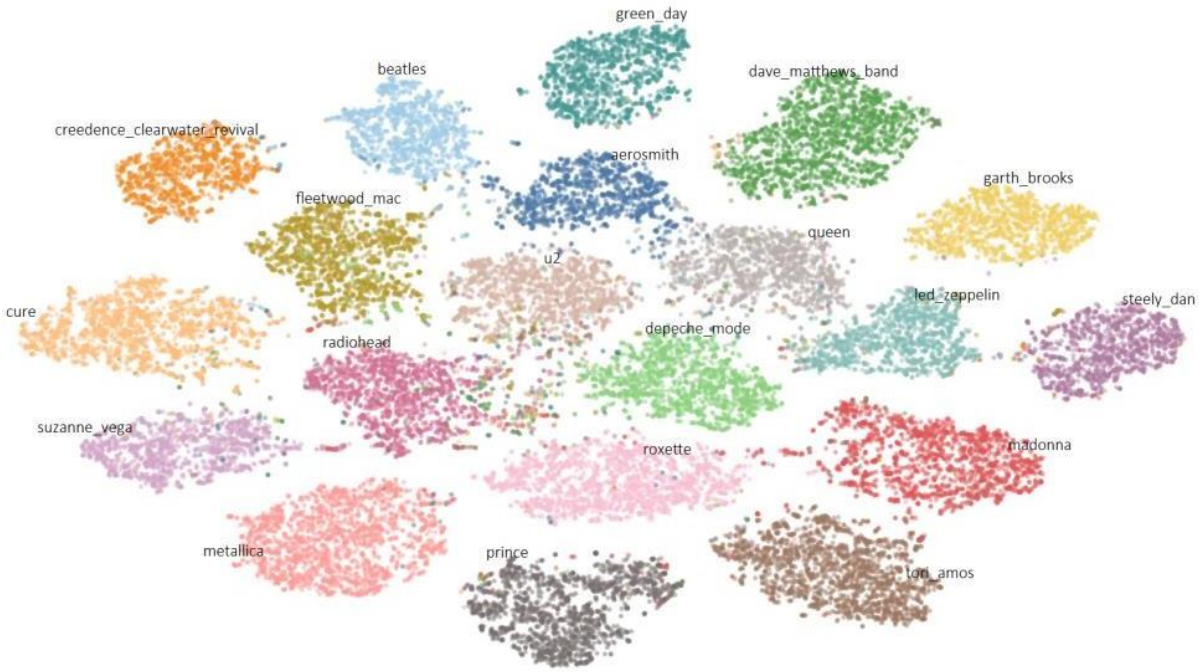

Рисунок 2 – Результат разделения классов по исполнителю при длине аудио фрагмента в 10 секунд

Выводы

В данной работе была рассмотрена задача классификации аудио информации в формате музыкальных композиций и более подробно изучен частный случай классификации по исполнителю. Решение задач обработки и выделения информации из аудио файлов позволяет создавать автоматические системы распознавания и классификации музыки, которые находят широкое применения в настоящее время в связи с распространением онлайн-сервисов хранения, прослушивания и поиска музыки. Задача классификации музыкальных композиций может быть решена большим количеством разных способов, но наиболее продуктивные результаты удалось получить, используя модели глубокого машинного обучения. Для обучения таких моделей необходимо представить входную аудио информацию в определённом виде, пригодном для обработки в нейронной сети. Одной из наиболее эффективных архитектур нейронной сети для классификации музыки по исполнителю можно назвать RCNN, с помощью которой при оптимальных условиях достижима средняя точность распознавания до 93,7 %.

Список использованных источников

1. Music Classification: Beyond Supervised Learning, Towards Real-world Applications / Won Minz, Spijkervet Janne, Choi Keunwoo, 2021 — 146 с.
2. A Tutorial on Deep Learning for Music Information Retrieval / Choi Keunwoo, Fazekas György, Cho Kyunghyun, Sandler Mark, 2017 — 16 с.
3. Convolutional recurrent neural networks for music classification / Keunwoo Choi, Gyorgy Fazekas, Mark Sandler, Kyunghyun Cho, 2016 — 5 с.
4. Music Artist Classification with Convolutional Recurrent Neural Networks / Zain Nasrullah, Yue Zhao, 2019 — 8 с.