МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ

Федеральное государственное автономное образовательное учреждение  
высшего образования

**«Южно-Уральский государственный университет**

**(национальный исследовательский университет)»**

**Высшая школа электроники и компьютерных наук**

**Кафедра системного программирования**

|  |  |
| --- | --- |
| РАБОТА ПРОВЕРЕНА  Рецензент  <ученая степень, ученое звание>  \_\_\_\_\_\_\_ <И.О. Фамилия рецензента>  «\_\_\_»\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ 2024 г. | ДОПУСТИТЬ К ЗАЩИТЕ  Заведующий кафедрой, д.ф.-м.н., профессор  \_\_\_\_\_\_\_\_\_\_ Л.Б. Соколинский  «\_\_\_»\_\_\_\_\_\_\_\_\_\_\_ 2024 г. |

**Разработка веб-приложения для генерации музыки**

ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА

ЮУрГУ – 09.04.04.2024.308-1492.ВКР

|  |  |
| --- | --- |
|  | Научный руководитель,  доцент кафедры СП, к.ф.-м.н., доцент \_\_\_\_\_\_\_\_\_\_ Г.И. Радченко  Автор работы, студент группы КЭ-229  \_\_\_\_\_\_\_\_\_\_ Д.В. Кутюшкин  Ученый секретарь  (нормоконтролер)  \_\_\_\_\_\_\_\_\_\_\_\_\_ И.Д. Володченко  «\_\_\_»\_\_\_\_\_\_\_\_\_\_\_ 2024 г. |

Челябинск, 2024 г.

МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ

Федеральное государственное автономное образовательное учреждение  
высшего образования

**«Южно-Уральский государственный университет**

**(национальный исследовательский университет)»**

**Высшая школа электроники и компьютерных наук**

**Кафедра системного программирования**

УТВЕРЖДАЮ

Зав. кафедрой СП

\_\_\_\_\_\_\_\_\_\_ Л.Б. Соколинский

29.01.2024 г.

**ЗАДАНИЕ**

**на выполнение выпускной квалификационной работы магистранта**

студенту группы КЭ-229

Кутюшкину Дмитрию Владимировичу,

обучающемуся по направлению

09.04.04 «Программная инженерия»

(магистерская программа «Искусственный интеллект и инженерия данных»)

1. **Тема работы** (утверждена приказом ректора от \_\_.\_\_.2024 г. № \_\_\_)

Разработка веб-приложения для генерации музыки.

1. **Срок сдачи студентом законченной работы:** 20.05.2024 г.
2. **Исходные данные к работе2**
3. Kailash A. Generative Adversarial Networks Projects: Build next-generation generative models using TensorFlow and Keras. – Packt Publishing Ltd., 2019. – 316 p.
4. Foster D. Generative Deep Learning, 2nd Edition. – O'Reilly Media, Inc., 2023. – 426 p.
5. **Перечень подлежащих разработке вопросов**
6. Провести анализ предметной области.
7. Подготовить набор данных.
8. Выбрать топологию нейронной сети.
9. Реализовать нейросетевую модель и провести ее обучение.
10. Спроектировать приложение для генерации музыки.
11. Реализовать приложение.
12. Провести тестирование приложения.
13. **Дата выдачи задания:** 29.01.2024 г.

**Научный руководитель,**

доцент кафедры СП, к.ф.-м.н., доцент Г.И. Радченко

**Задание принял к исполнению** Д.В. Кутюшкин

ОГЛАВЛЕНИЕ

[ВВЕДЕНИЕ 5](#_Toc159880671)

[1. АНАЛИЗ ПРЕДМЕТНОЙ ОБЛАСТИ 7](#_Toc159880672)

[1.1. Анализ аналогичных проектов 7](#_Toc159880673)

[1.2. Анализ архитектур нейронных сетей 9](#_Toc159880674)

[2. ПОДГОТОВКА НАБОРА ДАННЫХ 19](#_Toc159880675)

[2.1. Формат входных данных 19](#_Toc159880676)

[ЗАКЛЮЧЕНИЕ 20](#_Toc159880677)

[ЛИТЕРАТУРА 21](#_Toc159880678)

# ВВЕДЕНИЕ

**Актуальность**

В настоящее время ежемесячно появляется огромное множество разнообразного программного обеспечения, основанного на технологиях искусственного интеллекта. Эти технологии созданы для решения самых разных задач.

В современном музыкальном мире, где искусственный интеллект и машинное обучение играют все более значимую роль, генерация мелодии становится все более актуальной и инновационной темой исследования, которая может иметь значительное влияние на музыкальную индустрию и творческий процесс.

Тема генерации мелодии упрощает работу с таким важным аспектом музыкальной индустрии, как композиция музыки. Одними из ключевых проблем, с которыми сталкиваются музыканты являются сложность и затраты времени на придумывание интересного и запоминающегося мотива, особенно в ситуациях, когда появляется необходимость развивать идеи уже начатых произведений. Существование инструмента генерации мелодий может открыть больше возможностей для творчества. Написание мелодии – процесс, зачастую зависящий от вдохновения и настроен. Генератор мелодий в свою очередь мог бы помочь в поиске идей, продолжении уже существующих идей, и просто помог бы ускорить и оптимизировать процесс. Подобный инструмент также поможет начинающим музыкантам развивать свои навыки.

Эта технология также имеет потенциал для творческого экспериментирования и инноваций. Благодаря случайности в процессе генерации музыканты могут открывать новые жанры, настроения и стили, что способствует разнообразию и оригинальности в музыкальном творчестве.

Все эти факторы подтверждают актуальность и значимость исследования генерации мелодии. Этот инновационный подход не только упрощает и оптимизирует процесс создания музыки, но и стимулирует разнообразие, творческое экспериментирование и эксклюзивность в музыкальной индустрии.

**Постановка задачи**

Целью выпускной квалификационной работы является разработка веб-приложения для генерации музыки. Для достижения поставленной цели необходимо решить следующие задачи:

1. провести анализ предметной области;
2. сформировать набор данных для обучения;
3. выбрать топологию нейросети;
4. реализовать нейросетевую модель и провести ее обучение;
5. спроектировать приложение для генерации музыки;
6. реализовать приложение;
7. провести тестирование приложения.

**Структура и содержание работы**

Работа состоит из введения, двух глав, заключения и списка литературы. Объем работы составляет 22 страницы, объем списка литературы – 11 источников.

В первой главе описываются существующие аналоги, а также производится анализ архитектур нейронных сетей, применимых в задаче генерации звука.

Вторая глава посвящена формату входных данных и описанию использованного при обучении набора данных.

# 1. АНАЛИЗ ПРЕДМЕТНОЙ ОБЛАСТИ

1.1. Анализ аналогичных проектов

На данный момент тема относительно новая, и задача генерации музыки не очень часто затрагивается в статьях, а какие-либо проекты только начинают появляться.

Рассмотрим несколько наиболее актуальных и современных решений данной задачи.

**Jukebox**

Проект разработан исследователями из OpenAI, он предназначен для генерации полноценных музыкальных композиций, включая мелодии, гармонии, ритмы и тексты. Отличительной особенностью является его способность генерировать как инструментальные, так и вокальные партии.

Jukebox стремится к созданию новых и уникальных произведений, которые звучат аутентично и имеют характеристики определенного жанра.

Система реализована с помощью автоэнкодера, который сжимает музыку в дискретное пространство с помощью модели VQ-VAE. Эта модель основана на квантизации. Иерархическая VQ-VAE способна к генерации коротких инструментальных паттернов аудио из небольшого набора музыкальных инструментов.

Исследователи используют три уровня сжатия, которые сжимают аудиозапись в 8, 32 и 128 раз соответственно. В результате таких преобразований данные теряют большую часть деталей, но сохраняют информацию о тоне, тембре и громкости мелодии.

Модель обучали на собранном из интернета наборе из 1,2 миллионов композиций. Половина из них на английском языке. Для каждой аудиозаписи доступны метаданные и текст песни из LyricWiki [8].

**MuseNet**

Созданная исследователями из OpenAI MuseNet способна генерировать полноценные музыкальные композиции с использованием 10 различных инструментов и сочетать стили. Сеть научилась обнаруживать закономерности гармонии, ритма и стиля и предсказывать следующий токен. MuseNet использует ту же модель, что и GPT-2 – трансформер, обученный предсказывать следующий токен последовательности, будь то аудио или текст [9].

В качестве набора данных использовались коллекции MIDI-файлов, пожертвованные ClassicalArchives и BitMidi, дополнительно использовался набор данных MAESTRO.

**MusicLM**

Это модель, разработанная исследователями из Google. Модель предназначена для генерации музыки по текстовому описанию. Для решения задачи использовался decoder-only Transformer, состоящий из 24 слоев и 16 слоев внимания. Обучение производилось на наборе данных Free Music Archive [11].

**Нейромузыка**

Это модель, разработанная исследователями из Yandex, включенная в сервис Yandex Music для генерации бесконечного аудиопотока определенного настроения и жанра.

Для решения задачи использовались трансформеры. Модель обучалась на триплетах (басовая, гармоническая и мелодическая петли), закодированных в текстовой форме, впоследствии научившись генерировать новые петли и продолжать существующие [10].

1.2. Анализ архитектур нейронных сетей

Если упростить задачу, генерация аудио сводится к предсказанию последовательности по ее уже имеющимся элементам, последовательность может являться как существующим аудио фрагментом, так и случайно сгенерированным шумом. В случае генерации мелодий, задача упрощается, т.к. последовательность нот имеет гораздо меньше возможных значений элементов, чем абстрактный звуковой сигнал.

В задаче обработки последовательностей хорошо себя зарекомендовали рекуррентные нейронные сети (RNN) и различные их архитектуры, например, LSTM (Long-short term memory) и GRU (Gated Recurrent Units), чьи способности улавливать предыдущий контекст сыграли в этом главную роль.

Также для решения задачи подходят более сложные архитектуры, такие как трансформеры, генеративно-состязательные сети (GAN), которые показывают более хорошие результаты.

Достойны упоминания и другие подходы искусственного интеллекта, такие как обучение с подкреплением.

**Рекуррентные нейронные сети**

Рекуррентная нейронная сеть (Recurrent Neural Network, RNN) использует принцип обратной связи, подавая на вход рекуррентного слоя его выход в предыдущий момент времени [1]. Она обрабатывает последовательность, перебирая ее элементы и сохраняя состояние, полученное при обработке предыдущих элементов. Однако простейшая рекуррентная сеть обладает таким недостатком как «забывание» информации через длительное время.

Для решения этой проблемы была придумана архитектура LSTM, ее схема представлена на рисунке 1.

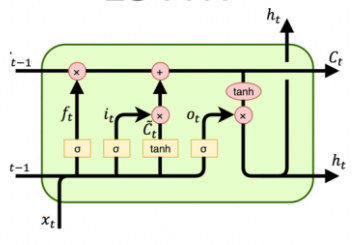


Рисунок 1 – Архитектура LSTM

Основное отличие заключается в реализации механизма долгосрочной памяти, таким образом информация из ранних интервалов может использоваться на поздних. За реализацию этого механизма на данной схеме отвечает выходной сигнал , имитация забывания ненужной информации реализована операцией поэлементного умножения, а запоминание новой информации операцией поэлементного сложения [1].

Существует упрощенная версия архитектуры LSTM – архитектура GRU (Gated Recurrent Unit), которая схожа с сетью LSTM, с тем отличием, что механизм памяти и выход (скрытое состояние) реализованы в одном сигнале, это уменьшает число параметров, тем самым уменьшая вычислительную сложность. Ее схема представлена на рисунке 2.

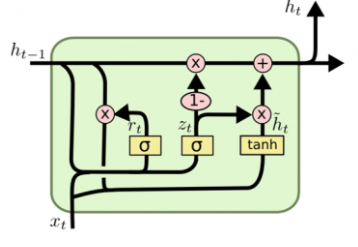


Рисунок 2 – Архитектура GRU

В статье [2] исследователи используют один блок GRU и один блок LSTM. В качестве входных данных в статье используются аудиосэмплы, представленные в виде фаз и амплитуд частот. После генерации скрытого состояния используется афинное преобразование для преобразования его в представление частоты. Затем оценивается ошибка L2 для истинного и сгенерированного значения. В результате получается сеть, обученная предсказывать следующие элементы последовательности.

Для генерации выполняется прямое распространение на исходной последовательности, в результате получается вектор предсказанных следующих элементов последовательности. Затем этот вектор используется как исходный для генерации новых элементов, и снова выполняется прямой проход до тех пор, пока не будет сгенерирована последовательность необходимой длины.

В результате экспериментов с генерацией было выяснено, что более крупные начальные последовательности дают лучшие результаты, число итераций также зависит от длины начальной последовательности. Эмпирически, генерация последовательности, которая примерно в три раза больше исходной имеет тенденцию генерировать связную музыку, которая не приводит к проблеме петель (случаев, когда предыдущее скрытое состояние похоже на текущее, а все последующие сохраняют эту тенденцию).

По генерируемым результатам архитектуры GRU и LSTM сильно отличались: GRU показала неудовлетворительные результаты, генерируя в основном белый шум, в то время как LSTM смогла воспроизводить вполне правдоподобную музыку. Таким образом сеть LSTM намного лучше справляется с задачей.

В статье [3] используется слегка иной подход, сеть LSTM прогнозирует не только саму частоту, но и время, в которое она воспроизводится. На рисунке 3 изображена архитектура модели.

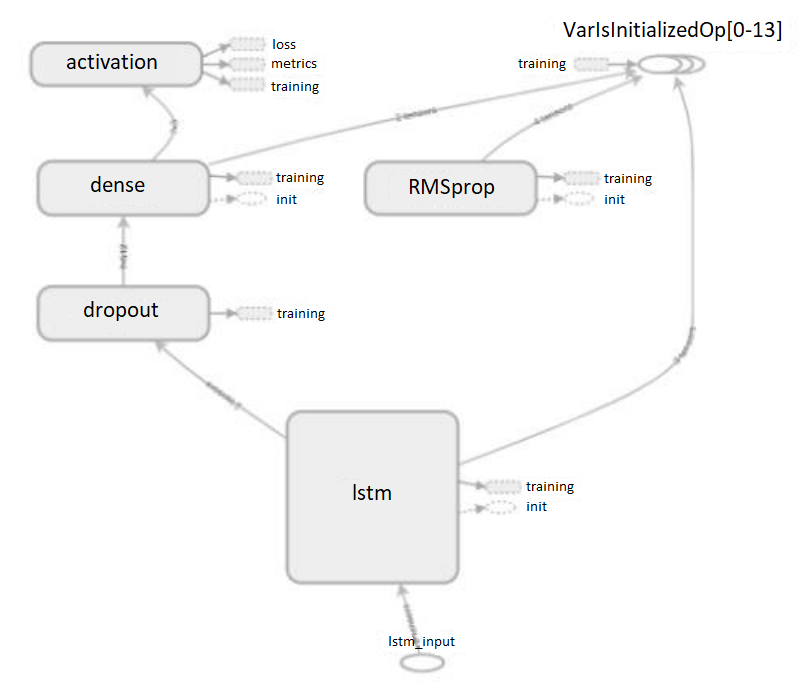


Рисунок 3 – Модель генерации аудио

**Трансформеры**

Трансформер – архитектура нейронных сетей, избегающая рекуррентности и полностью полагающаяся на механизм внимания для построения зависимостей между входом и выходом [4]. Основное преимущество данной архитектуры по сравнению с RNN – высокая эффективность в условиях параллелизации. Архитектура представлена на рисунке 4.

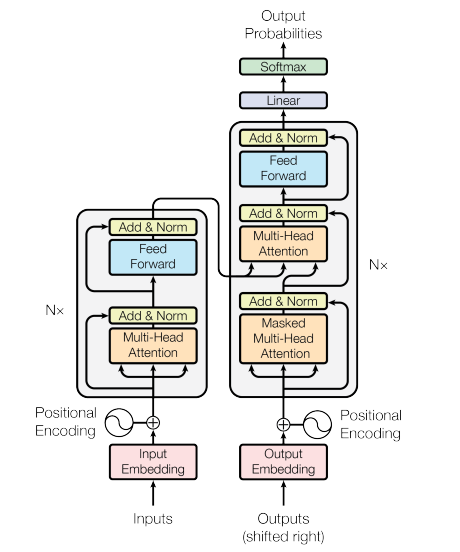


Рисунок 4 – Архитектура трансформера

Устройство трансформера состоит из кодирующего и декодирующего блока, они состоят из нескольких кодеров и энкодеров соответственно. На вход поступает некоторая последовательность, формируется ее векторное представление, прибавляется вектор позиционного кодирования, затем эта последовательность без учета порядка поступает в кодирующий блок, после этого декодирующий блок получает на вход часть последовательности и выход кодирующего блока, в результате получается новая последовательность.

Кодировщики состоят из слоя, реализующего механизм самовнимания, и простого полносвязного слоя, декодировщики также состоят из этих слоев, а также реализуют механизм внимания к выходам кодировщиков.

Механизм внимания основан на присвоении элементам последовательности значений, отражающих их значимость на текущем шаге, то есть модель обращает свое «внимание» на данные элементы.

Функцию внимания можно описать как сопоставление запроса и набора пар ключ-значение с выходными данными, где запрос, ключи, значения и выходные данные являются векторами. Результат рассчитывается как взвешенная сумма значений, где вес, присвоенный каждому значению, вычисляется функцией совместимости запроса с соответствующим ключом. Расчет для матрицы результата представлен в формуле (1).

|  |  |  |
| --- | --- | --- |
|  |  | (1) |

В этом лежит основа работы трансформеров. Также имеются механизмы, улучшающие результаты работы трансформеров, например, множественное внимание, здесь создаются несколько наборов матриц, затем, результаты объединяются, благодаря чему модель может фокусироваться на нескольких аспектах. Можно также выделить позиционное кодирование, позволяющее учитывать порядок во входной последовательности.

На рисунке 5 изображены схемы работы механизмов внимания и множественного внимания.

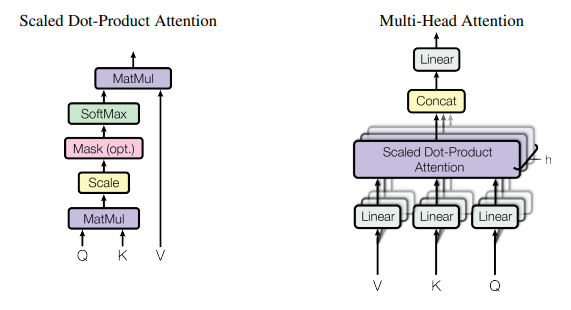


Рисунок 5 – Организация слоев внимания и множественного внимания

Одной из архитектур для задачи генерации музыки является архитектура Wave2Midi2Wave. Эта архитектура представляет собой комбинацию из трех моделей.

1. Первая модель используется для представления аудио в MIDI формате.
2. Вторая сеть является особым типом трансформера, генерирующим новые последовательности музыки с долгосрочной когерентностью. Т.к. данные от первой модели приходят в MIDI формате, задача сводится к работе с символьными данными, что упрощает обучение. В отличии от стандартных моделей трансформеров, данная сеть использует относительное внимание, что позволяет ей генерировать последовательности, выходящие за рамки обучающих данных.
3. Третья сеть, используя набор данных MAESTRO, содержащий записи с конкурса фортепианных исполнителей, и сгенерированные трансформером последовательности в MIDI формате, синтезирует эти последовательности в реалистичный звук [5].

На рисунке 6 представлена схема работы данной архитектуры.

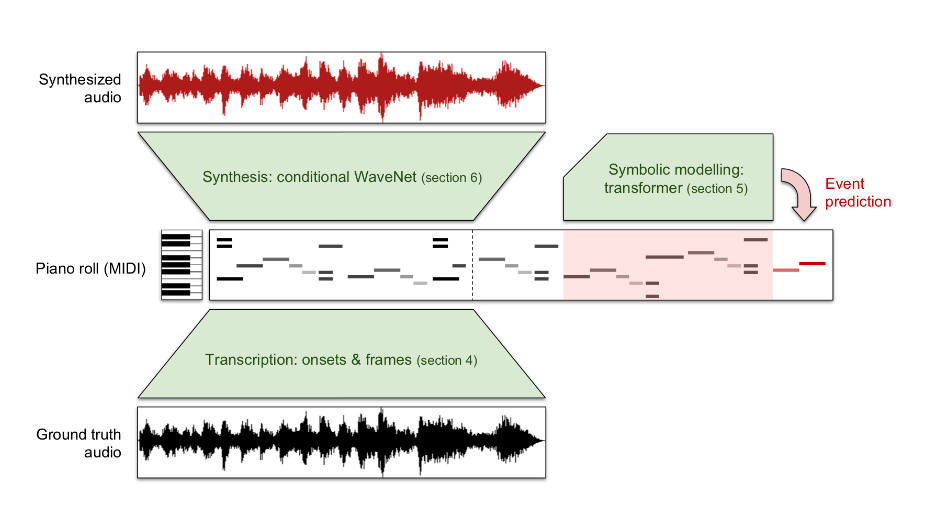


Рисунок 6 – Схема работы архитектуры Wave2Midi2Wave

Также архитектура трансформера используется в задаче генерации аудио по тексту, такая архитектура представлена в статье [6]. Сеть AudioGPT состоит из слоев самовнимания и слоев с обратной связью. Она способна улавливать неочевидные корреляции и производить последовательные выводы. Архитектура представлена на рисунке 7.

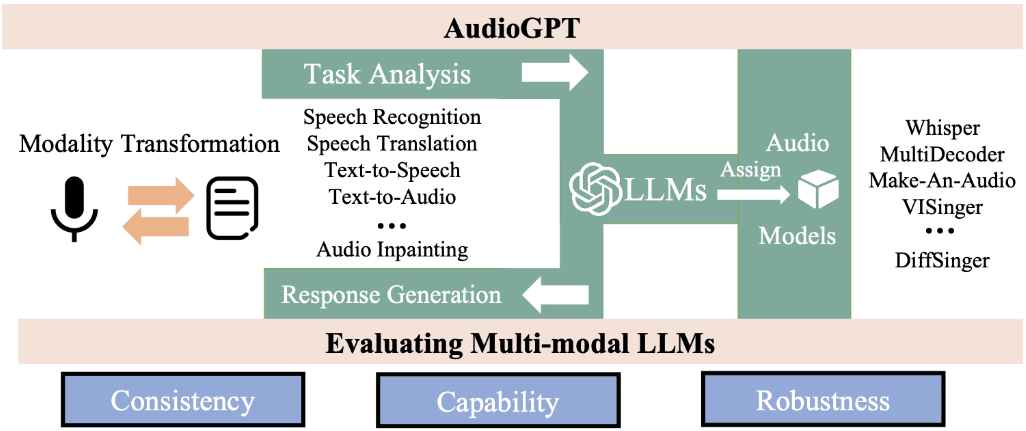


Рисунок 7 – Архитектура AudioGPT

**Генеративно-состязательные сети**

Генеративно-состязательные сети (GAN) получили широкое применение в задачах генерации. Такая архитектура построена на двух нейронных сетях – генераторе и дискриминаторе.

Задачей генератора является генерирование правдоподобных данных, а задача дискриминатора – отличать сгенерированные генератором данные от реальных. В этом кроется идея данной архитектуры – последовательно обучая дискриминатор все лучше и лучше классифицировать реальные и сгенерированные данные, а генератор генерировать все более правдоподобные данные, мы получаем очень качественную генерацию.

В статье [7] представлена архитектура для генерации музыки, состоящей из нескольких партий. Данные представлены как тензор, состоящий из матриц с бинарными значениями, представляющими наличие нот на каждом временном шаге. Архитектура построена на моделировании реальных композиционных подходах и состоит из трех GAN моделей – джемов, композитора и гибридной.

Модель джемов представляет собой несколько независимых генераторов, генерирующих музыку из собственного случайного вектора , где – число генераторов (дорожек), эти генераторы получают «критику», то есть сигналы обратного распространения, от разных дискриминаторов, их число равно числу генераторов.

Модель композитора состоит из одного единственного генератора и одного дискриминатора. Генератор создает аудиозапись, состоящую из дорожек, он требует для этого одного общего для всех дорожек случайного вектора , а дискриминатор исследует все сгенерированные дорожки вместе, чтобы определить является ли входная музыка настоящей или сгенерированной.

Гибридная модель состоит из генераторов и объединяет эти две модели, подавая на вход генераторов две случайные дорожки, одну из сгенерированных моделью джема, а другую из модели композитора, затем результат генераторов объединяется в одну композицию и подается на вход одному единственному дискриминатору. Архитектуры этих трех моделей приведены на рисунке 8, а структура общей модели показана на рисунке 9.

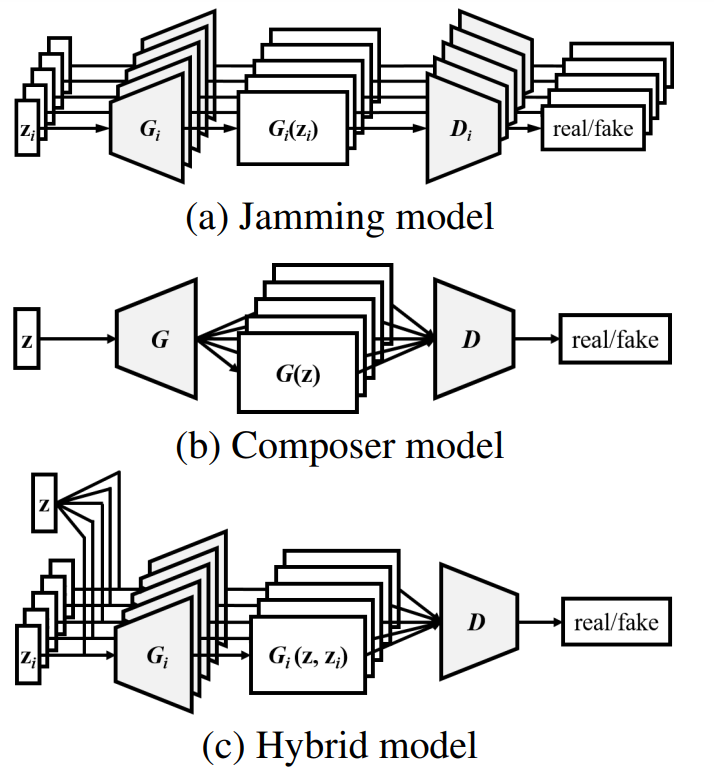


Рисунок 8 – Архитектуры моделей джема, гибридной и композитора

# 2. ПОДГОТОВКА НАБОРА ДАННЫХ

2.1. Формат входных данных

Поговорим немного о формате данных. Для начала, необходимо собирать минусы композиций, т.е. варианты песен без вокала, т.к. вокал в данном случае будет выступать скорее, как шум в данных.

Поскольку нейронная сеть напрямую не может принимать на вход аудиофайлы, то необходимо преобразовать их в векторный вид.

**Представление аудио данных**

Рассмотрим несколько вариантов представления аудиоданных.

1. Временные последовательности. Простой способ представления аудио – это использование временного сигнала, который представляет изменения амплитуды аудиоволны во времени.
2. Спектрограммы. Преобразование аудиосигнала в двумерное представление, где по одной оси временная шкала, а по другой – частотная. Как правило для ее получения используется оконное преобразование Фурье.
3. Мел-спектрограммы. Это модификация спектрограммы, которая имитирует восприятие звука человеческим ухом. Она использует мел-шкалу частот. Переход к мелам осуществляется с помощью применения мел-фильтров к исходной спектрограмме.
4. MFCC (Mel-Frequency Cepstral Coefficients). Спектрограмма аудиосигнала подвергается процедуре выделения мел-частотных кепстральных коэффициентов, которые представляют собой компактное представление.
5. Embedding-подход. Представление аудиосигналов в виде векторов фиксированной длины, что упрощает их использование в нейронных сетях, предназначенных для обработки последовательностей.

# ЗАКЛЮЧЕНИЕ

В рамках данной работы был проведен обзор существующих проектов для генерации музыки.

Был проведен анализ архитектур нейронных сетей, применяемых в задаче генерации аудио.

Были рассмотрены способы представления аудиоданных.

# ЛИТЕРАТУРА

1. Фостер Д. Генеративное глубокое обучение. Творческий потенциал нейронных сетей. – СПб.: Питер, 2020. – 352 с.
2. Nayebi A., Vitelli M. GRUV: Algorithmic Music Generation using Recurrent Neural Networks. [Электронный ресурс] URL: https://cs224d.stanford.edu/reports/NayebiAran.pdf (дата обращения: 15.01.2024 г.).
3. Mangal S., Modak R., Joshi P. LSTM Based Music Generation System. // ArXiv, 2019. – 6 p.
4. Vaswani, A., Shazeer, N.M., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I. Attention is All you Need. // ArXiv, 2017. – 15 p.
5. The MAESTRO Dataset and Wave2Midi2Wave. [Электронный ресурс] URL: https://magenta.tensorflow.org/maestro-wave2midi2wave (дата обращения: 15.01.2024 г.).
6. Huang R., Li M., Yang D., Shi J., Chang X., Ye Z., Wu Y., Hong Z., Huang J., Liu J., Ren Y., Zhao Z., Watanabe S. Audiogpt: Understanding and generating speech, music, sound, and talking head. // ArXiv, 2023. – 14 p.
7. Hao-Wen Dong, Wen-Yi Hsiao, Li-Chia Yang, Yi-Hsuan Yang. MuseGAN: Multi-track Sequential Generative Adversarial Networks for Symbolic Music Generation and Accompaniment. // ArXiv, 2017. – 13 p.
8. Jukebox. [Электронный ресурс] URL: https://openai.com/research/jukebox (дата обращения: 15.01.2024 г.).
9. Musenet. [Электронный ресурс] URL: https://openai.com/research/musenet (дата обращения: 15.01.2024 г.).
10. Glazyrin N. CONTEXT-AWARE GENERATION OF MELODIC MIDI LOOPS. [Электронный ресурс] URL: https://archives.ismir.net/ismir2021/latebreaking/000037.pdf (дата обращения: 15.01.2024 г.).
11. Agostinelli A., Denk T. I., Borsos Z., Engel J., Verzetti M., Caillon A., Huang Q., Jansen A., Roberts A., Tagliasacchi M., Sharifi M., Zeghidour N., Frank C. Musiclm: Generating music from text. // arXiv, 2023. – 15 p.