Министерство образования Республики Беларусь
Учреждение образования
Белорусский государственный университет
информатики и радиоэлектроники

УДК 004.89

Воронова
Вероника Валерьевна

Модели нейросетевой генерации продолжения музыкальной композиции с сохранением связности и стиля

АВТОРЕФЕРАТ

на соискание степени магистра
по специальности 1-40 80 04 – Информатика и технологии программирования

Научный руководитель
Сиротко С.И.
к.ф.-м.н., доцент

Минск 2022

ВВЕДЕНИЕ

Музыка – это звук, организованный специальным образом для выражения различных идей и эмоций. Например, порядок организации музыкальных нот разных частот (от низких до высоких) влияет на мелодию, гармонию и текстуру музыки. Чередование сильных и слабых долей дает основу восприятию ритма. Повторения и структура также являются важными факторами, позволяющими сделать музыкальное произведение связным и понятным.

Творчество всегда считалось прерогативой человечества. И если в когнитивных задачах, таких как вычисления и обработка информации, люди уже признали превосходство искусственного интеллекта и активно пользуемся плодами автоматизации, то в таких «человеческих» видах деятельности как живопись, поэзия или сочинение музыки алгоритмы им уступают.

Создание машин, способных сочинять музыку подобно людям, является одной из интересных задач в области мультимедиа и искусственного интеллекта. Поэтому задача генерации музыки была предметом пристального интереса в течении двух последних десятилетий.

Работа с музыкальными композициями отличается от задач работы с текстом или изображениями благодаря следующим ключевым особенностям композиций: во-первых, музыка имеет иерархическую структуру и зависимости, разворачивающиеся во времени; во-вторых, музыка может состоять из множества инструментов, которые взаимосвязаны и раскрываются во времени; в-третьих, музыка сгруппирована в аккорды, арпеджио и мелодии, следовательно, каждый временной шаг может иметь несколько выходов.

У аудиоданных есть ряд свойств, которые, в некотором роде, делают их похожими на то, что обычно изучается в рамках глубокого обучения: последовательный характер музыки напоминает NLP; композиция может иметь несколько «каналов» звука с точки зрения тонов и инструментов, которые напоминают изображения.

Нейронные сети представляют собой набор алгоритмов, которые пытаются распознать отношения и связи между входными данными, имитируя работу человеческого мозга. Они интерпретируют данные посредством своего рода машинного восприятия, маркируя или группируя исходные данные. Образцы, которые они распознают, являются числовыми, содержащимися в векторах, в которые должны быть переведены все данные реального мира, будь то изображения, звук, текст или временные ряды. Чтобы иметь возможность использовать данные алгоритмы, мы должны думать о мелодии как о последовательности числовых токенов, или векторов, которые содержат

некоторую информацию о ноте, ритме, тембре, а также о любых других характеристиках, которые могут влиять не полученные результаты в каждом конкретном случае.

Таким образом, генерация музыки в общем смысле зависит главным образом от двух факторов. Первым из них является представление данных, то есть кодирование входных данных в форму, наиболее удобную и эффективную для обучения, а также интерпретация полученных результатов. Вторым фактором является выбор алгоритма нейросетевого моделирования, которые может эффективно работать с предоставленными данными, и таким образом производить благозвучные, связные композиции.

В данной работе необходимо изучить, реализовать и проанализировать различные представления данных, а также алгоритмы нейросетевого моделирования, которые позволят решить поставленную задачу.

Магистерская диссертация проверена в системе «Атиплагиат». Процент оригинальности соответствует норме, установленной кафедрой информатики. Цитирования обозначены ссылками на публикации, указанные в «Списке использованных источников».

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Постановка задачи

В рамках магистерской работы поставлена цель изучить и реализовать модели нейросетевого моделирования, которые позволяют по заданным начальным нотам генерировать благозвучные связные композиции в том же стиле.

Для достижения поставленной цели необходимо решить следующие задачи:

– изучить актуальные технологии нейросетевого моделирования;

– проанализировать структуру и особенности природы музыкальных данных, выделить свойства, которые могут быть полезны;

– найти данные для обучения;

– подобрать библиотеки языка python для наиболее эффективной работы с данными;

– продумать подходы к представлению данных для обучения, реализовать их, проанализировать сильные и слабые стороны подходов в задаче генерации музыки;

– выбрать наиболее подходящие нейросетевые модели, адаптировать их для решения поставленной задачи;

– настроить и обучить выбранные модели;

– сгенерировать выходные данные для каждого подхода;

– проанализировать полученные результаты;

– визуализировать полученные результаты;

– сделать вывод о результатах проделанной работы.

Выбор технологий для разработки

Для разработки был выбран язык python. Его главными достоинствами является:

– кроссплатформенность – python имеет способность работать на разных платформах без необходимости внесения изменений;

– простота и лаконичность языка – код лаконичен и удобочитаем, что позволяет легко писать код, презентовать его, получать и делиться информацией с другими разработчиками в сообществе;

– разнообразие библиотек и фреймворков для текущей задачи - они предлагают среду, которая сокращает время разработки ПО.

В качестве среды разработки был выбран IPython, который представляет собой мощную интерактивную оболочку, поддерживает интерактивную визуализацию данных, просто в использовании, является высокопроизводительным.

Обучение нейронных сетей требует вычислительных ресурсов, поэтому было решено развернуть кластер на личном свободном стационарном компьютере на восьми ядерном процессоре AMD с 64 ГБ оперативной памяти и видео картой Nvidia GeForce GTX 1080 посредством kubernetes.

Для работы с нейросетевыми алгоритмами была использована библиотека keras. Keras — это высокоуровневая библиотека нейронных сетей, написанная на Python и способная работать поверх TensorFlow или Theano.

Она был разработан с акцентом на возможность быстрого экспериментирования. Способность перейти от идеи к результату с минимально возможной задержкой является ключом к проведению хорошего исследования.

Keras имеет следующие особенности:

– позволяет легко и быстро создавать прототипы (благодаря полной модульности, минимализму и расширяемости);

– поддерживает как сверточные сети, так и рекуррентные сети, а также их комбинации;

– поддерживает произвольные схемы подключения (включая обучение с несколькими входами и выходами).

Без проблем работает на CPU и GPU.

Область применения и сравнение с аналогами

Еще в 1957 году учёные из Университета Иллинойса выявили общие принципы в создании музыки. Затем они перенесли эти данные в компьютер и написали программу, которая могла генерировать музыку. Работало это так: алгоритм в случайном порядке генерировал ноты и ритмы. Представлены они были в виде чисел в диапазоне от 0 до 15. За каждым числом была привязана нота из двух октав диатонической гаммы До. Сгенерированные таким образом песни действительно звучали странно, но это было лишь начало пути. С развитием искусственного интеллекта энтузиасты начали развивать более футуристичные проекты.

Генерация музыки сейчас является актуальной темой. Если рассматривать практическое применение генераторов, то можно выделить несколько самых трендовых способов использования сгенерированной музыке:

– Помощь композиторам. Еще Моцарт с современниками изобрел игру в «музыкальные кости», когда броском кубика они выбирали из большой таблицы такты и составляли из них менуэты. В двадцать первом веке музыка уже не является признаком исключительно благородных людей, и не ограничивается только тем, что сейчас мы относим к классическому направлению музыки. Интуитивно мы понимаем под музыкой комбинацию звуков, организованных ритмически и интонационно. Такая музыка востребована практически по всем направлениям – используется в фильмах, звучит фоном в видеоиграх, привлекает внимание к цифровым рекламным бигбордам, способствует развитию детей посредством музыкальных игрушек, отдает сигналы в микроволновых печах и холодильнике. И этот список можно продолжать очень долго. Таким образом, спрос на музыку различных категорий огромный, а ресурс композиторов ограниченный.

– Создание персонализированной музыки. Различные платформы для прослушивания музыки, к примеру, такие как Яндекс.Музыка или Spotify, отлично умеют подбирать композиции в так называемый «плейлист дня», то есть треки, которые пользователь вероятнее всего захочет прослушать сегодня, и, скорее всего, станет от этого более счастливым. Следующей ступенью развития данного подхода является переход от персонализированных плейлистов к персонализированным трекам. То есть тем композициям, которые были сгенерированы, либо изменены по исходной композиции, на то, что пользователю понравится. На текущий момент есть первые прецеденты использования таких композиций: в 2020 году российский стартап Endel и певица Граймс (бывшая жена Илона Маска) объявили о коллаборации. Endel — это приложение генеративной музыки. Алгоритмы адаптируются к погоде, времени суток, пульсу, циркадному ритму и местоположению пользователя. По словам Граймс, на это сотрудничество её подтолкнуло то, что «музыка для детей в целом всегда плохо написана». И эту музыку она не хотела включать своему сыну.

– Использование фоновых композиций, не нарушающих авторских прав. В настоящее время популярные платформы, такие как Instagram и Youtube, активно борются с проблемой нарушения авторских прав, и зачастую можно стать свидетелем публикации видео, у которого звук заблокирован из-за нарушения прав правообладателя, при чем это не обязательно происходит на момент публикации, ведь авторские права могут быть изменены. Генерация музыки искусственным интеллектом может решить данную проблему. То есть, при необходимости получить мелодию для фона, свободную для использования, достаточно будет сгенерировать ее по некоторым критериям, вместо того, чтобы заниматься поиском доступной из существующих.

Рассмотрим некоторые популярные нейронные сети для генерации музыки.

Цифровые представления звука бывают разных форм. Для воспроизведения звук обычно сохраняется путем кодирования формы волны по мере ее изменения во времени. Для анализа часто используются спектрограммы как для вычислительных методов, так и для визуального контроля. Спектрограмму можно получить из сигнала, вычислив преобразование Фурье. Спектрограммы являются комплексными: они представляют как амплитуду, так и фазу различных частотных составляющих в каждый момент времени. При извлечении информации из звуковых сигналов оказывается, что часто можно просто отбросить фазовую составляющую, потому что она неинформативна для большинства интересующих нас вещей. Собственно, поэтому амплитудную спектрограмму часто называют просто как «спектрограмма». Однако при генерации звука фаза очень важна, потому что она существенно влияет на наше восприятие.

WaveNet и SampleRNN — это авторегрессионные модели необработанных сигналов. В то время как WaveNet является сверточной нейронной сетью, SampleRNN использует стек рекуррентных нейронных сетей. До изобретения данных сетей такая идея серьезно не рассматривалась, поскольку моделирование долгосрочных корреляций в последовательностях на тысячах временных шагов казалось невозможным с помощью доступных инструментов.

У данных моделей есть свои недостатки: в том числе медленная выборка из-за авторегрессии и отсутствие возможности интерпретации того, что на самом деле происходит внутри сети.

Стратегия WaveNet для работы с долгосрочными корреляциями заключается в использовании расширенных сверток: последовательные сверточные слои используют фильтры с промежутками между входными данными, так что шаблон связанности на многих слоях образует древовидную структуру. Это обеспечивает быстрый рост рецептивного поля, а это означает, что WaveNet с несколькими слоями может изучать зависимости на многих временных отрезках.

Стратегия SampleRNN немного отличается: несколько RNN накладываются друг на друга, и каждый из них работает на разной частоте. RNN более высокого уровня обновляются реже, что означает, что они могут легче фиксировать долгосрочные корреляции и изучать высокоуровневые функции.

Обе модели также применялись к генерированию фортепианной музыки, что представляло собой хорошую демонстрацию перспективности генерирования музыки в области волновых форм, но они были явно

ограничены в своей способности улавливать долговременную музыкальную структуру.

Кроме моделирования музыки в форме волновых форм существует множество альтернативных подходов.

Модель машинного обучения включает в себя не только сами «обучающие» слои, но еще и подходы к тому, как будет организован входной и выходной слой, то есть как: музыка конвертируется во входное представление и обратную расшифровку сгенерированных данных в мелодию. То есть для получения благозвучных мелодий критическими являются два элемента – способ подготовки данных и алгоритм обучения. В последние два года значительный прогресс был сделан в улучшении «обучающей» части моделей, были придуманы принципиально улучшенные архитектуры. Однако, многие из них используют сходный формат данных – основанный на нотах и аккордах, поскольку самый популярный формат музыкальных композиций для обучения – это MIDI, и из его представления выходит представление данных. В рамках же этой работы делается попытка представить данные иначе, наиболее подходящим к конкретным моделям образом, исходя из их самых удачных применений.

ОСНОВНОЕ СОДЕРЖАНИЕ

Во **введении** диссертационной работы определена предметная область и указано основное направление исследования, показана актуальность темы работы, рассмотрены основные свойства музыкальных композиций, которые могут стать основой для решения поставленной задачи.

В **первой** главе сформулированы цели, которые планируется достичь и задачи, которые необходимо для этого решить. Обосновывается выбор языка программирования python, редактора ipython и библиотеки keras, которые были использованы. Произведена оценка текущего состояния проблемы, а также обзор и анализ существующих аналогов.

Генерация музыки является актуальной задачей. Важной частью моделей нейронных сетей является представление данных на первом, входном слое. Другой важной частью – устройство самих «обучающих» слоев моделей. Эти оба фактора и являются полем для исследователей в сфере генерации музыкальных композиций.

Вторая глава посвящена описанию технологий нейросетевого моделирования, которые были использованы.

LSTM – компонент, способный изучать долгосрочные зависимости, запоминать информацию в течении длительного периода, что положительно повлияет на генерацию ритма и гармонии композиции; умеет бороться с проблемой исчезающего градиента, хотя и не решает ее полностью. Недостатками, существенными для нас, является то, что подход требует достаточно много ресурсов и времени для обучения, требуется много памяти из-за линейных слоев. Еще одним недостатком является склонность к переобучению, но существует ряд подходов, которые помогут бороться с этим, главное, заметить, что переобучение происходит.

Сети-трансформеры показывают себя очень хорошо в задачах обработки естественного языка, а наш план – это использовать представление композиции в том числе на манер текста. Данные сети ловко используют механизм внимания, сосредотачиваясь на важных токенах. В случае генерации музыки улучшение запоминания контекста позволит улучшить чистоту, ритм, акценты в генерируемой мелодии, а так даст возможность генерировать более продолжительные связные мелодии.

Третья глава посвящена выбору и реализации моделей представления данных и нейросетевых моделей для решения поставленной задачи.

Одним из выбранных подходов является разбиение входных данных по спискам нот, сыгранным в каждый момент времени с некоторым шагом. Достоинством этого метода является простота реализации, возможность выбрать частоту разбиения мелодии. Недостатками является то, что

случайные неблагозвучные ноты могут попасть в один «аккорд» при неудачном разбиение; длительность, с которой нота была сыграна, не отображается в явном виде; содержится информация об относительном времени звучания ноты, а не об абсолютном; плохо читается ритм композиции; не учитывается сила нажатия нот. Была настроена, обучена модель, основанная на рекуррентных нейронных сетях.

Иной подход представления имитирует словесные предложения. Каждая нота описывается абсолютным моментом времени, длинной, силой нажатия. Данное представление позволяет лучше генерировать ритм; делает эффективным использование подходов из сферы обработки естественного языка. С другой стороны, более сложная структура данных требует более сложных моделей и\или увеличивает время обучения. Были рассмотрены такие модели, как сети-трансформеры, сети gpt-2. Был добавлен параметр стиля.

ЗАКЛЮЧЕНИЕ

В данной диссертационной работе рассматриваются возможные виды представления данных, а также ряд моделей нейросетевого моделирования, которые позволяют по заданному началу музыкальной композиции генерировать продолжение, при этом это сохраняя стилистическую окраску, связность композиции.

Был рассмотрен такой подход к представлению входных данных как разбиение по нотам, сыгранным в каждый момент времени. Его достоинством является простата в генерации входных данных и полученных результатов. Недостатком является то, что не он не учитывает силу нажатия нот, а также не представляет явно информацию длительности ноты относительно других.

Вторым рассмотренным подходом стало представление музыки на манер текста. Недостатком такого подхода является более громоздкая форма входных данных, необходимость иметь достаточно сложную сеть, чтобы она могла моделировать структуру предложений. Достоинствами является то, что становится эффективным использование моделей обучения из области NLP, учет силы нажатия нот, длительности относительно других нот, что способствует лучшей генерации ритма.

Были рассмотрены, применены, обучены и сравнены в полученных результатах ряд моделей машинного обучения, основанных на рекуррентных нейронных сетях, а также на сетях-трансформерах и gpt-2, которые эффективно применяются в генерации текста.

При должной настройке, применение данных моделей позволяет генерировать благозвучные, связные музыкальные композиции. Таким образом, цель работы была достигнута.

СПИСОК ОПУБЛИКОВАННЫХ РАБОТ

1-А Воронова, В. В. Обзор методов генерации текста на естественном языке / Воронова В. В., Удовин И. А. // Информационные технологии и системы 2021, Минск, 2021. – С. 72–73.