Компьютер и человек – кто кого?

«Компьютер делает не то, что вы хотите, а то, что вы ему приказываете»
Автор неизвестен

Продолжение.
Начало в №№ 7, 8, 9, 10/2023, №№ 1, 2, 4, 5, 6, 7/2024

Арсений Ворошилов, по материалам Meta и Tech Xplore

Новые AI-модели

Пока в профессиональной среде продолжаются горячие дискуссии относительно искусственного интеллекта в целом и правил его применения в частности, IT-компании продолжают совершенствовать свои разработки. Недавно появились две новые AI-модели, определенно заслуживающие внимания.

Movie Gen от Meta

Meta, владеющая Facebook и Instagram, представила общественности свою новую AI-модель Movie Gen, способную генерировать реалистичные видео и аудиоклипы в соответствии с текстовым описанием, сформированным пользователем. В компании утверждают, что эта модель является прямым конкурентом аналогичным разработкам компаний OpenAI и ElevenLabs.

Примеры плодов «творчества» Movie Gen, которые представила Meta, содержат видео, на которых животные плавают и занимаются серфингом. Есть также образцы клипов, созданных на основе фото реальных людей, которых «оживили» и заставили выполнять те или иные действия, например, рисовать на холсте.

Movie Gen также способна генерировать фоновую музыку и звуковые эффекты, синхронизированные с содержимым видео. Во всяком случае, так утверждают в Meta. К тому же пользователи получают возможность использовать новую AI-модель для модификации уже имеющихся видеоматериалов.

В одном их таких видео Meta показала, как вложила помпоны в руки человека, бегущего в одиночестве по пустыне, а в другом – модифицировала парковочное пространство, в котором человек катался на скейте, заменив сухое грунтовое покрытие на покрытую брызгами лужу.

Пока что возможности Movie Gen по созданию видео ограничено хронометражем 16 с, что в целом не так уж и мало. Что касается аудио, то его максимальная длительность может достигать 45 с. Слепые тесты показали, что модель вполне способна конкурировать с аналогичными инструментами других разработчиков, таких как Runway, OpenAI, ElevenLabs и Kling.

Марк Цукерберг анонсирует Movie Gen

Кадры видео, сгенерированные с помощью Movie Gen

Любопытно, что презентация AI-модели состоялась на фоне попыток Голливуда осознать, как использовать технологию создания видео с помощью искусственного интеллекта, вскоре после того, как поддерживаемый корпорацией Microsoft стартап OpenAI в феврале нынешнего года впервые показал, как AI-модель Sora способна создавать выглядящее по-кинематографически видео на основе текстового описания.

Сторонникам инновационных технологий в индустрии развлечений не терпится использовать подобные средства для повышения эффективности и ускорения процесса кинопроизводства, тогда как другие испытывают тревогу в связи с применением систем, которые, как считается, были натренированы на защищенных авторским правом работах без разрешения правообладателей.

Законодатели также обращают внимание на проблемы, связанные с тем, как созданные с помощью искусственного интеллекта фейки, проще говоря, очень правдоподобно выглядящие фальшивки, используются по всему миру при проведении избирательных компаний, в том числе в США, Пакистане, Индии и Индонезии.

Представители Meta утверждают, что компания вряд ли сделает Movie Gen доступной для свободного применения разработчиками, как это было сделано для серии больших языковых моделей Llama, поскольку в компании принимают в расчет риски, связанные с использованием каждой из моделей. В Meta отказались комментировать, какие риски в компании видят в связи с Movie Gen. Но отметили, что Meta взаимодействовала непосредственно с профессиональным сообществом индустрии развлечений и другими создателям контента, обсуждая варианты применения Movie Gen и собирается учесть их в собственных разработках, запланированных на следующий год.

Основываясь на данных, обнародованных Meta, и исследовании выпущенной компанией модели, можно сделать вывод, что Meta использовала сочетание лицензированных и широкодоступных наборов данных при создании Movie Gen.

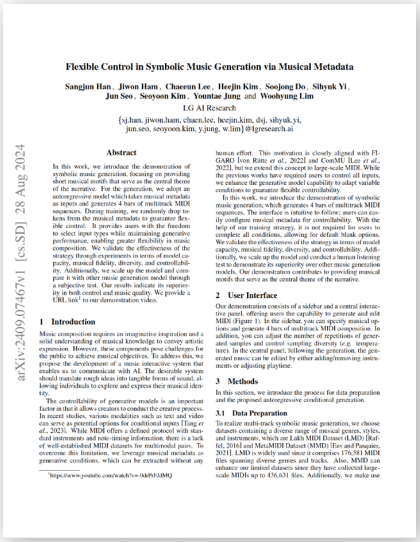
Известно также, что представители OpenAI в нынешнем году встречались с руководством Голливуда и сотрудниками агентств для обсуждения потенциального партнерства с использованием Sora, однако никаких известий о достигнутых результатах пока обнародовано не было. Более того, опасения, связанные с подходом компании, увеличились в мае, когда актриса Скарлет Йоханссон обвинила создателя ChatGPT в том, что соответствующий чат-бот имитирует ее голос без разрешения актрисы.

Тем временем компания Lions Gate Entertainment, создавшая «Голодные игры» и «Сумерки», объявила в сентябре, что предоставила AI-стартапу Runway доступ к своей библиотеке кинематографического и телевизионного контента для обучения AI-модели. За это студия и сотрудничающие с ней кинематографисты получат возможность использовать модель в качестве дополнительного инструмента в своей работе.

Аудиомодель от LG AI Research

Активная деятельность развернулась не только на ниве генеративного искусственного интеллекта для создания видео. В сфере аудио не менее жарко. Недавно была представлена новая AI-модель для так называемого символьного создания музыки на основе применения музыкальных метаданных. Надо отметить, что создание музыкального контента ничуть не проще, чем генерирование видео. Причем в отличие от видео, где все объекты более-менее формализованы, особенно когда речь идет о том, что существует в реальности, сгенерировать звуковой ряд, который был бы именно музыкой, да еще не повторяющей то, что было создано ранее, задача не из простых.

Искусственный интеллект открыл новые интересные возможности для музыкальной индустрии, например, позволив разработать инструменты, которые способны автоматически генерировать музыкальные композиции или отдельные инструментальные треки. Правда, пока что существующие средства были предназначены для использования музыкантами, композиторами и музыкальными продюсерами, а не широкой неискушенной аудиторией.



Flexible Control in Symbolic Music Generation via Musical Metadata

Sangjun Han, Jiwon Ham, Chaeeun Lee, Heejin Kim, Soojong Do, Sihyuk Yi, Jun Seo, Seoyoon Kim, Yountae Jung and Woohyung Lim

LG AI Research

{sj.han, jiwon.ham, chaeun.lee, heejin.kim, dsj, sihyuk.yi, jun.seo, seoyoon.kim, y.jung, w.lim}@lgresearch.ai

Abstract

In this work, we introduce the demonstration of symbolic music generation, focusing on providing short musical motifs that serve as the central theme of the narrative. For the generation, we adopt an autoregressive model which takes musical metadata as inputs and generates 4 bars of multitrack MIDI sequences. During training, we randomly drop tokens from the musical metadata to guarantee flexible control. It provides users with the freedom to select input types while maintaining generative performance, enabling greater flexibility in music composition. We validate the effectiveness of the strategy through experiments in terms of model capacity, musical fidelity, diversity, and controllability. Additionally, we scale up the model and compare it with other music generation model through a subjective test. Our results indicate its superiority in both control and music quality. We provide a URL link[1] to our demonstration video.

1 Introduction

Music composition requires an imaginative inspiration and a solid understanding of musical knowledge to convey artistic expression. However, these components pose challenges for the public to achieve musical objectives. To address this, we propose the development of a music interactive system that enables us to communicate with AI. The desirable system should translate rough ideas into tangible forms of sound, allowing individuals to explore and express their musical identity.

The controllability of generative models is an important factor in that it allows creators to conduct the creative process. In recent studies, various modalities such as text and video can serve as potential options for conditional inputs [Ung et al., 2023]. While MIDI offers a defined protocol with standard instruments and note-timing information, there is a lack of well-established MIDI datasets for multimodal pairs. To overcome this limitation, we leverage musical metadata as generative conditions, which can be extracted without any human effort. This motivation is closely aligned with FIGARO [von Rütte et al., 2022] and ComMU [Lee et al., 2022], but we extend this concept to large-scale MIDI. While the previous works have required users to control all inputs, we enhance the generative model capability to adapt variable conditions to guarantee flexible controllability.

In this work, we introduce the demonstration of symbolic music generation, which generates 4 bars of multitrack MIDI sequences. The interface is intuitive to follow; users can easily configure musical metadata for controllability. With the help of our training strategy, it is not required for users to complete all conditions, allowing for default blank options. We validate the effectiveness of the strategy in terms of model capacity, musical fidelity, diversity, and controllability. Additionally, we scale up the model and conduct a human listening test to demonstrate its superiority over other music generation models. Our demonstration contributes to providing musical motifs that serve as the central theme of the narrative.

2 User Interface

Our demonstration consists of a sidebar and a central interactive panel, offering users the capability to generate and edit MIDI (Figure 1). In the sidebar, you can specify musical options and generate 4 bars of multitrack MIDI composition. In addition, you can adjust the number of repetitions of generated samples and control sampling diversity (e.g. temperature). In the central panel, following the generation, the generated music can be edited by either adding/removing instruments or adjusting playtime.

3 Methods

In this section, we introduce the process for data preparation and the proposed autoregressive conditional generation.

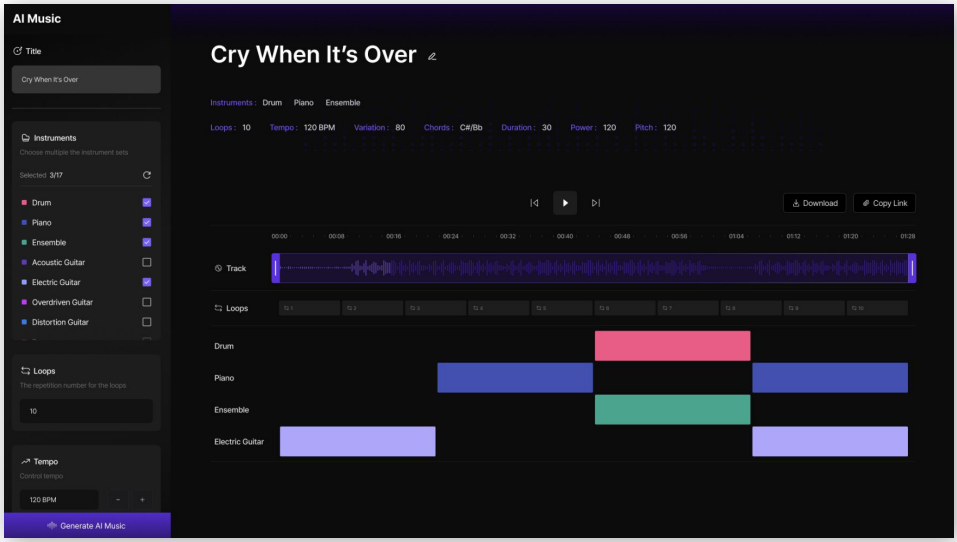
3.1 Data Preparation

To realize multi-track symbolic music generation, we choose datasets containing a diverse range of musical genres, styles, and instruments, which are Lakh MIDI Dataset (LMD) [Raffel, 2016] and MetaMIDI Dataset (MMD) [Ens and Pasquier, 2021]. LMD is widely used since it comprises 176,581 MIDI files spanning diverse genres and tracks. Also, MMD can enhance our limited datasets since they have collected large-scale MIDIs up to 436,631 files. Additionally, we make use

[1]https://www.youtube.com/watch?v=-0dtWFJJMQ

arXiv:2409.07467v1 [cs.SD] 28 Aug 2024

Титульная страница работы, опубликованной на сервере arXiv



Пользовательский интерфейс генератора музыки

Исследователи из LG AI Research недавно разработали новую интерактивную систему, позволяющую любому пользователю легко воплощать свои идеи в музыку. Система, описанная в работе, которая опубликована на сервере препринтов arXiv, содержит только декодирующий авторегрессивный преобразователь, натренированный на музыкальных наборах данных, в сочетании с интуитивно понятным пользовательским интерфейсом.

«*Мы представили демонстрацию символьной генерации музыки, сосредоточившись на обеспечении коротких музыкальных мотивов, что служит основной темой мелодии*, – пишут в работе Сангджун Хан, Дживон Хэм и их соавторы. – *Для генерации музыки мы применили авторегрессивную модель, которая берет музыкальные метаданные в качестве входных и генерирует четыре трека многодорожечных последовательностей MIDI*».

Базирующаяся на преобразователе модель, положенная в основу разработанной группой исследователей системы символьного генерирования музыки, была обучена на двух наборах музыкальных данных, а именно, на Lakh MIDI и MetaMIDI. В сумме эти наборы данных насчитывают более 400 тыс. файлов MIDI, которые представляют собой файлы, содержащие различную информацию о музыкальных треках, такую как воспроизводимые ноты, длительность нот, темп их воспроизведения.

Для обучения своей модели исследователи конвертировали каждый файл MIDI в файл представления музыкального события (REMI). Этот специфический формат предполагает кодирование данных MIDI в токены, представляющие различные музыкальные функции, например, высоту тона и скорость. Файлы REMI фиксируют динамику музыки так, что они предпочтительны для обучения AI-моделей для генерации музыки.

«В процессе обучения мы случайным образом добавляем токены из музыкальных метаданных, чтобы обеспечить гибкое управление, – пишут исследователи. – *Это дает пользователям свободу выбора выходных типов при сохранении эффективности генерирования, что позволяет гибче генерировать музыкальную композицию»*.

В дополнение к разработке модели на основе преобразователя для символьного генерирования музыки, Хан, Хэм и их коллеги создали простой интерфейс, который сделал систему доступной как для опытных пользователей, так и для новичков. Этот интерфейс в настоящее время состоит из боковой панели и центральной интерактивной панели.

С помощью средств на боковой панели пользователи могут указать аспекты музыки, которую должна генерировать модель, например, какие инструменты задействовать и каким должен быть темп композиции. После того, как композиция сгенерирована моделью, можно редактировать трек в центральной панели, в частности, путем добавления/удаления инструментов либо коррекции момента времени, когда они должны вступать.

«Мы проверяем эффективность стратегии путем экспериментов, оценивая такие свойства системы, как ее емкость, музыкальная точность, разнообразие и управляемость, – пишут авторы модели. – *К тому же мы расширяем модель и сравниваем ее с другими моделями генерирования музыки, проводя субъективные тесты. Наши результаты показывают превосходство над другими моделями как по управлению, так и по качеству музыки»*.

Исследователи выяснили, что их модель работает очень хорошо и может надежно генерировать максимум четыре дорожки музыки на основе заданных пользователем параметров. В своих последующих исследованиях они собираются улучшить систему за счет увеличения длительности музыкальных треков, которые способна генерировать модель, расширения спектра параметров, которые могут задавать пользователи, а также путем совершенствования пользовательского интерфейса системы.

«Наша модель, обученная генерировать под общим управлением четыре дорожки музыкальной композиции, имеет ограничения в плане увеличения длительности композиции и управления локальными элементами каждой дорожки, – признают исследователи. – *Тем не менее наши усилия имеют большое значение в смысле генерирования высококачественных музыкальных композиций, которые можно использовать в цикличном режиме»*.

Продолжение следует