Jukebox: A Generative Model for Music

Prafulla Dhariwal, Heewoo Jun, Christine Payne, Jong Wook Kim, Alec Radford, Ilya Sutskever

В статье представлена новая модель генерации музыки с текстом на основе VQ-VAE. Особенностью модели является ее способность генерировать музыкальные треки вместе с вокалом. Генерации модели – это дорожки в “чистом” звуковом домене. Такой подход позволяет увеличить разнообразие треков и создавать музыкальные произведения по образу более 9000 различных исполнителей.

Статья впервые была опубликована на arxiv 30 апреля 2020 года. Она не публиковалась на больших конференциях. Идеи, представленные в статье, заимствованы из нескольких предыдущих работ по генерации музыки. Так, идея использовать VAE в этой задаче впервые появилась в статье MusicVAE https://arxiv.org/abs/1803.05428, а архитектура трансформера схожа с предложенной в статье https://arxiv.org/abs/1809.04281.

Авторы статьи:

Prafulla Dhariwal – исследователь OpenAI, специализируется на генеративных моделях и обучении без учителя. Ушел из MIT, живет в Сан-Франциско, есть видео с объяснением статьи от него на ютубе: https://youtu.be/Jlb1IQ9ooxw. Работает в разных сферах, статьи от Point-E до Improved DDPM, много статей по генерации картинок.

HeeWoo Jun – исследователь OpenAI, так же много публикаций в разных областях, работает вместе с Prafulla. Закончила университет Торонто, также училась в Стэнфорде.

Christine McLeavy Payne – исследователь OpenAI. Закончила Princeton, специализировалась в физике, после этого ушла в медицину и нейронауку. Пианистка, увлекается RL, много публикаций в сферах, не связанных с компьютерными науками.

Наибольшее влияние на статью оказали следующие статьи

https://arxiv.org/abs/1809.04281 – впервые используют self-attention в генерации музыки

https://arxiv.org/abs/1803.05428 – идея применить VAE

https://arxiv.org/abs/2301.11325 – главный конкурент статьи

Цитируется в ней же.

Статья примечательна размерами модели – 5B параметров. Такие большие модели кроме них никто не обучал. Та же Music LM имеет всего 1.6B параметров, хотя она вышла на 3 года позже.

Плюсы работы:

- Генерация музыки с вокалом
- Можно получить спрессованные коды для определенного жанра музыки
- После декомпрессии коды будут звучать как оригинал
- Неплохое качество
- Можно продолжить уже имеющуюся песню с помощью VQ-кодов

Минусы работы:

- Размер модели (5B parameters)
- Много артефактов
- Может не хватить контекстного окна
- Этические проблемы https://transactions.ismir.net/articles/10.5334/tismir.86

Что можно улучшить:

- Разнообразить языки и исполнителей
- Улучшить качество вокала
- Сделать модель эффективнее по памяти и времени обучения

Несколько мыслей по поводу работы:

VQ-VAE показала себя очень хорошо. Однако возможно такое, что структура модели слишком сложна для такой задачи. Я бы предложил для генерации мелодии сделать что-то похожее на MIDINet, то есть обычную сверточную архитектуру. Потому что такие большие модели с миллиардами параметров вряд ли найдут применение в жизни. Вокал можно генерировать отдельно – здесь VQ-VAE, скорее всего, отличный выбор. Получаем 2 модели, их результаты просто наложим друг на друга.

Вообще, применение музыкальных генеративных моделей я вижу только одно – генерация непрерывного потока звука в реальном времени. В случае Jukebox такой вариант, конечно, невозможен.