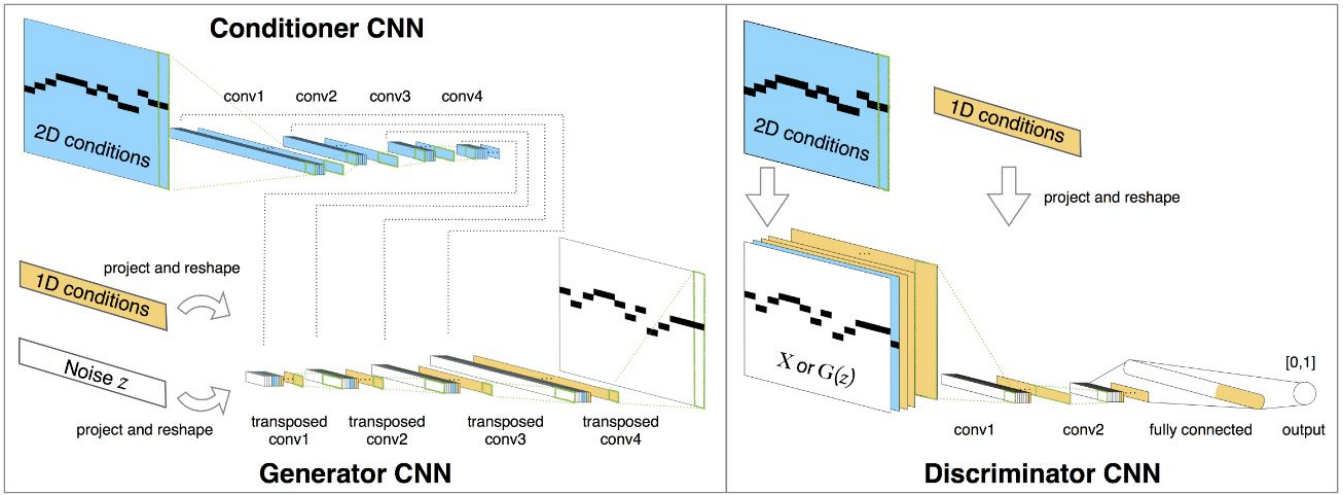
Jukebox: A Generative Model for Music

Daniil Panteleev

Генерация музыки в чистом звуковом домене

Большая часть предыдущих моделей генерации музыки выдают на выходе просто набор нот

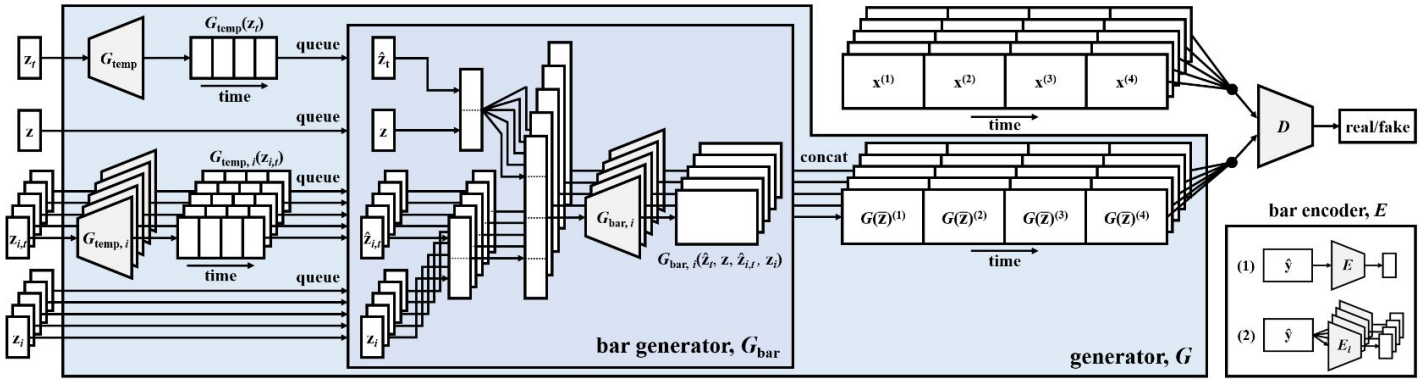
Самый простой вариант: MIDINet



https://arxiv.org/abs/1703.10847

MuseGAN: попытка учесть особенности музыки при генерации

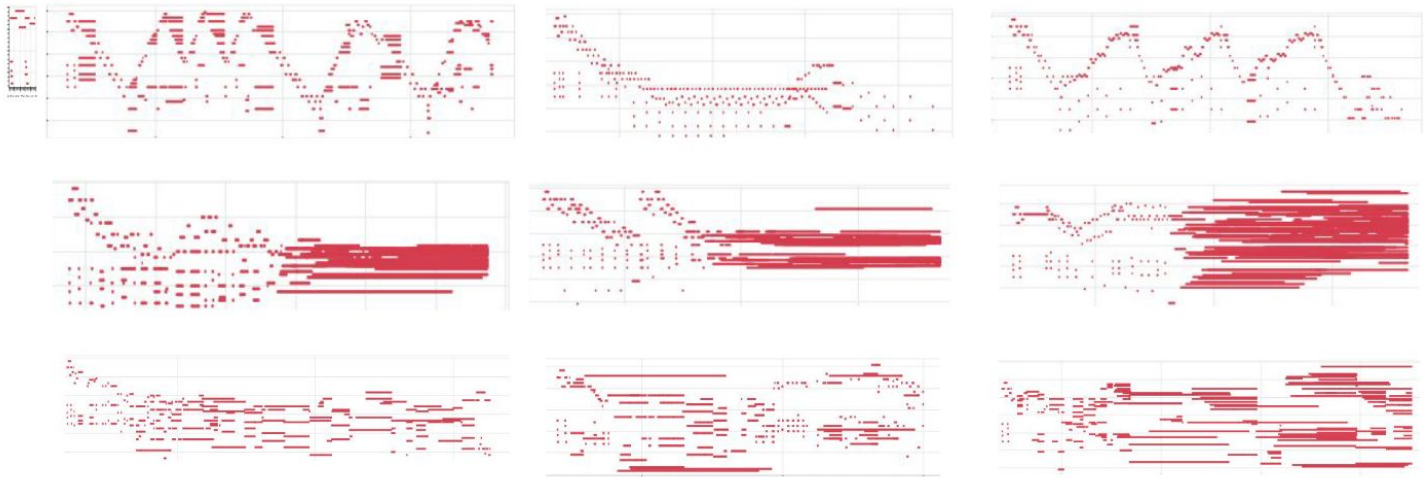
Появилась возможность генерировать треки, похожие на заданный пользователем. Используется очень сложная архитектура, слишком вычислительно затратно.



https://arxiv.org/abs/1709.06298

Music Transformer: по мотивам “Attention is all you need”

Впервые используется self-attention. Структура сгенерированных музыкальных треков выглядит гораздо привычнее, однако все еще генерация по нотам.

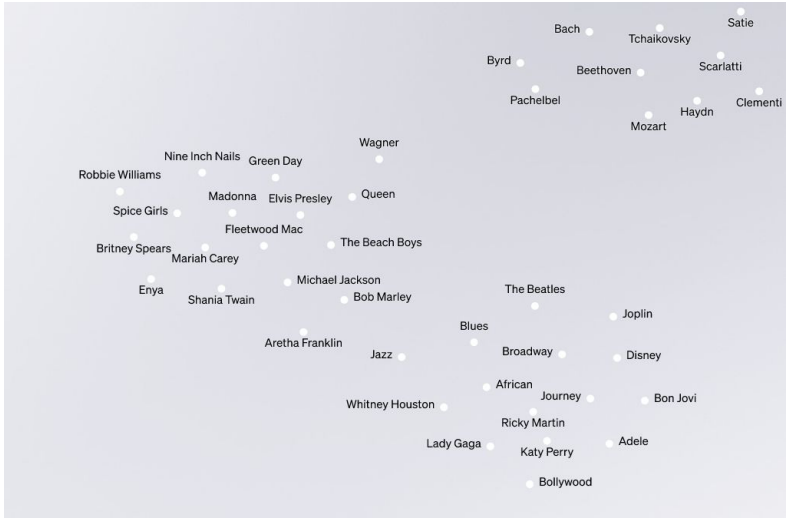


https://arxiv.org/abs/1809.04281

MuseNet

OpenAI обучили SparseTransformer на куче данных, опять же генерация по нотам, но качество в разы лучше, плюс возможность генерировать 10 разных инструментов. Хорошо умеет комбинировать стили различных исполнителей.

В основе GPT-2 для предсказания следующей ноты по предыдущим.



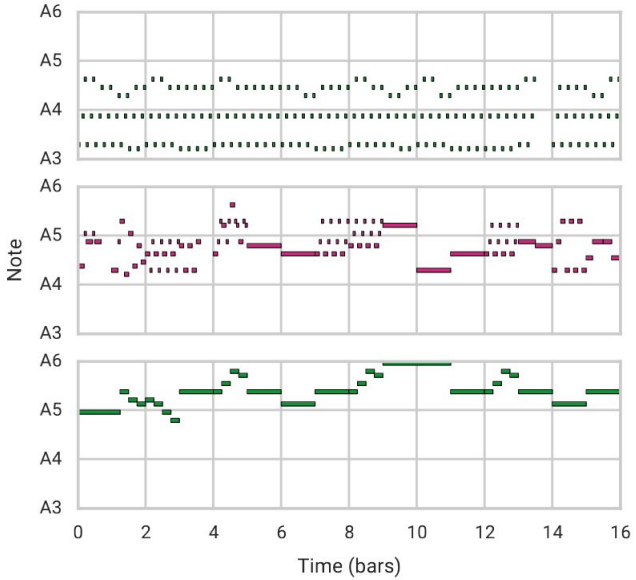
Ссылка по которой можно послушать генерации:
https://openai.com/research/musenet

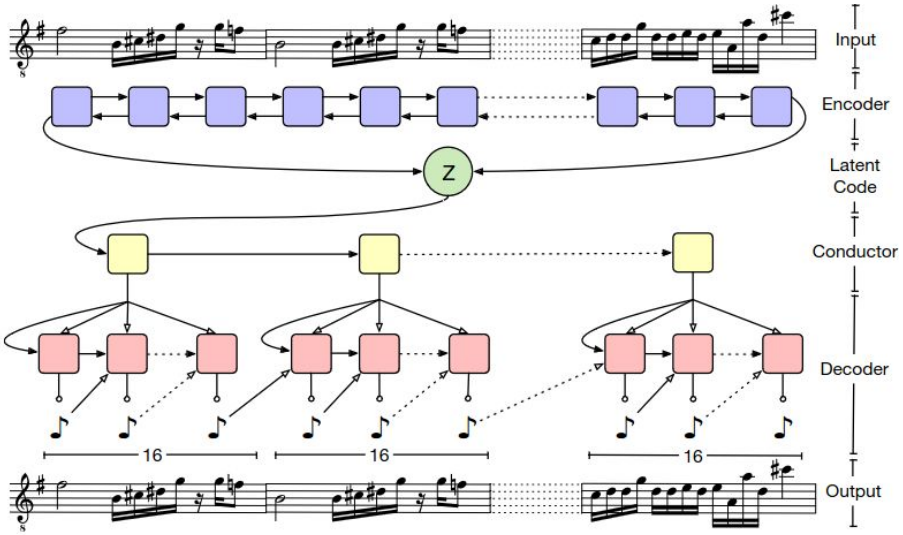
Сама статья:
https://www.ijitee.org/wp-content/uploads/papers/v9i6/F3580049620.pdf

Гитхаб:
https://github.com/AbhilashPal/MuseNet

MusicVAE

Первое использование VAE в генеративных моделях для музыки, один из основных источников статьи про Jukebox.

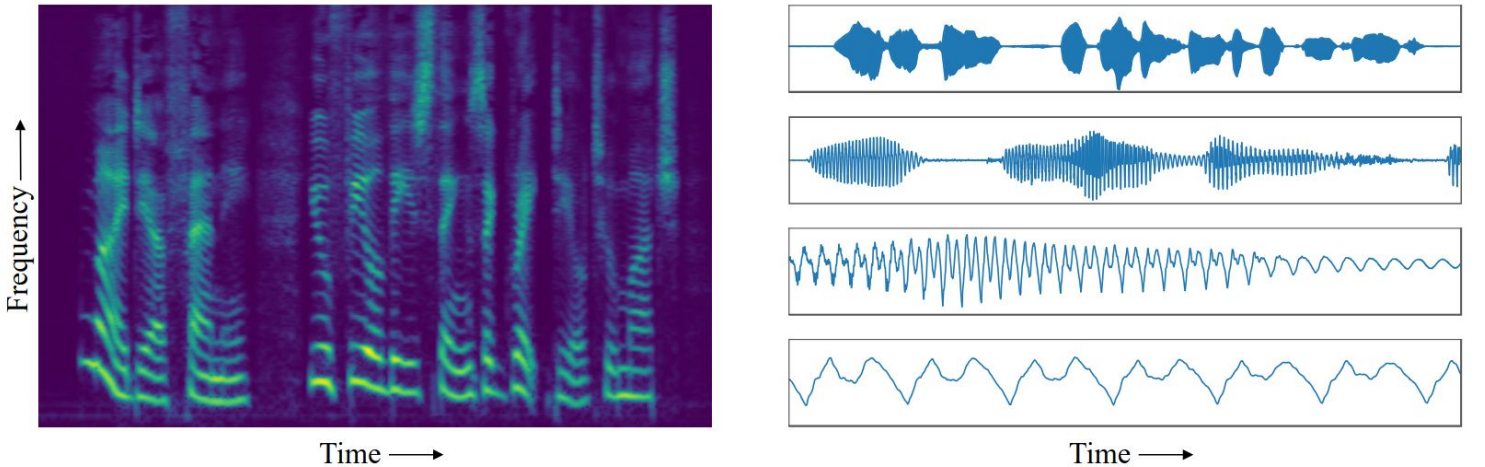




https://arxiv.org/abs/1803.05428

MelNet

Здесь для предсказания далеких по времени связей используются 2d спектрограммы.

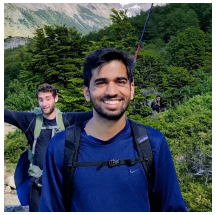


(a) Spectrogram representation

(b) Waveform representation (1x, 5x, 25x, 125x magnifications)

https://arxiv.org/pdf/1906.01083.pdf

Авторы



Prafulla Dhariwal - OpenAI, первая статья по генерации музыки, работает в разных сферах – от диффузионок до Point-E, есть видео объяснялка статьи на ютубе от него: https://youtu.be/Jlb1lQ9ooxw

HeeWoo Jun - Open Al, Fast Spectrograms, speech recognition, Point-E, Image Synthesis

Christine McLeavy Payne - OpenAI, RL, пианистка

Плюсы:

- 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:

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

Модель обучалась 4 недели на 512V100

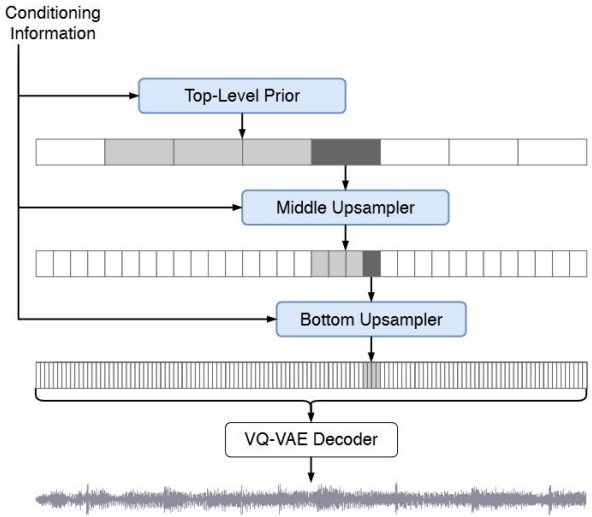
Jukebox vs MusicLM

	Jukebox	**MusicLM**
Число параметров	5B	1.6B
Качество звука	8kHz – 24kHz	24kHz
Разнообразие жанров и стилей	9000 исполнителей, 27 жанров	Более 100 жанров
Генерация вокала	✅	❌
Иерархическое моделирование	✅	✅

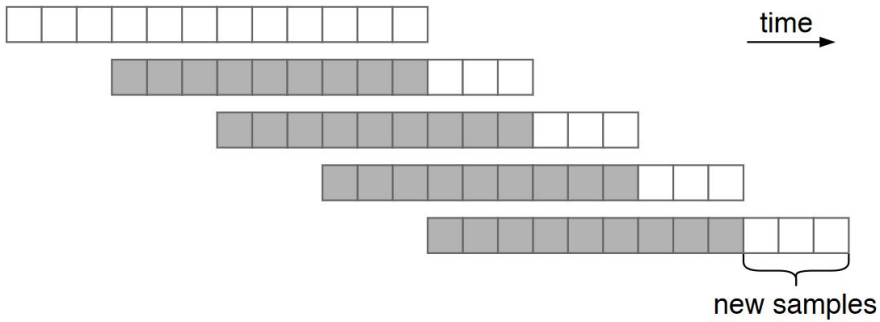
Некоторые детали:

3 способа sampling’a

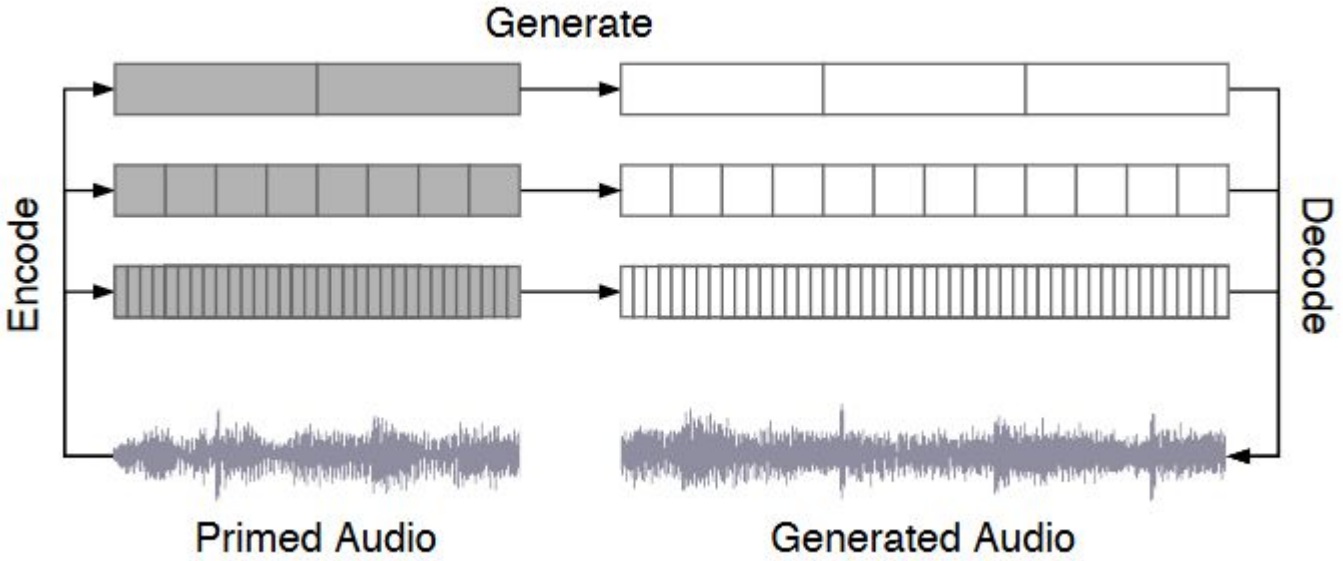
Ancestral sampling



Windowed sampling



Primed sampling



VAE vs VQ-VAE

VAE: представление латентных векторов в виде подпространства, “непрерывное представление”
VQ-VAE: представление латентных векторов в виде дискретных объектов (почти как эмбеддинги), “дискретное представление”

Простой пример дискретного представления:

стиль	темп	басы	высоты	вокал
21	48	55	14	93

https://ml.berkeley.edu/blog/posts/vq-vae/

Future work

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

Идеи: можно разделить модель на 2 отдельные – минус и вокал

Использование: генерация непрерывного потока аудио

Полезные ссылки

https://openai.com/research/jukebox главная страница

https://jukebox.openai.com/ песочница

https://github.com/openai/jukebox гитхаб

https://arxiv.org/abs/2005.00341 статья