It’s Raw! Audio Generation with State-Space Models

Рецензия

Авторы - Stanford University:

- Albert Gu & Karan Goel
  - Efficiently modeling long sequences with structured state spaces [2021] — S4 block
  - Combining recurrent, convolutional, and continuous-time models with linear state space layers [2021] — Linear State Space Layers (Не представлены в текущей статье)
- Chris Donahue
  - Adversarial Audio Synthesis [2019] — Первая попытка применить ГАНы в синтезе аудио
  - GANSynth: Adversarial neural audio synthesis [2019] — Супер эффективно по скорости и памяти генерируем аудио. Побили WaveNet
  - Expediting TTS Synthesis with Adversarial Vocoding [2019] — TTS в сотни раз быстрее чем умели до этого, путем устранения ботлнека предыдущих подходов.
- Christopher Ré

Что было дальше?

- ISTFTNET: Fast and Lightweight Mel-Spectrogram Vocoder Incorporating Inverse Short-Time Fourier Transform [4 march 2022]
- AudioLM: a Language Modeling Approach to Audio Generation [7 September 2022]
- BigVGAN: A Universal Neural Vocoder with Large-Scale Training [9 June 2022]

Что ещё?

Авторегрессионные ранние подходы, на которые опирались авторы:

- Generating long sequences with sparse transformers
- An unconditional end-to-end neural audio generation model
- Wavenet: A generative model for raw audio

Самые важные цитирования:

1. **A Survey on Artificial Intelligence for Music Generation: Agents, Domains and Perspectives**

   Обзор текущих моделей генерации музыки.

2. **GoodBye WaveNet - A Language Model for Raw Audio with Context of 1/2 Million Samples**

   SotA в генерации аудио. Побили даже SaShiMi. С помощью CNN и Трансформера решили проблему сохранения долгосрочных связей и зависимостей в аудио.

3. **GAN You Hear Me? Reclaiming Unconditional Speech Synthesis from Diffusion Models**

   AudioStyleGAN на базе диффузионных моделей.

Вопросы:

Почему именно YouTubeMusic и Beethoven?

Почему лишь SC09 ?

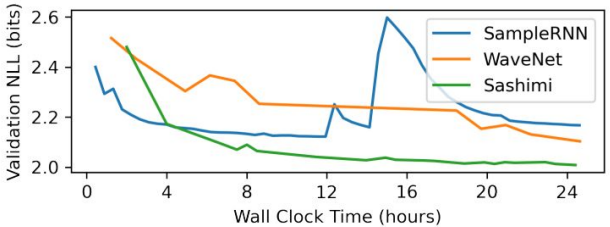
Почему не LJSpeech ?

Хотелось бы оценки по дополнительным музыкальным бенчмаркам.

Сильные стороны

Очень подробное и обширное сравнение с предыдущими моделями

Model	NLL	Time/epoch	Throughput	Params
SampleRNN–2 tier	1.762	800s	112K samples/s	51.85M
SampleRNN–3 tier	1.723	850s	116K samples/s	35.03M
WaveNet–512	1.467	1000s	185K samples/s	2.67M
WaveNet–1024	1.449	1435s	182K samples/s	4.24M
SaShiMi–2 layers	1.446	205s	596K samples/s	1.29M
SaShiMi–4 layers	1.341	340s	316K samples/s	2.21M
SaShiMi–6 layers	1.315	675s	218K samples/s	3.13M
SaShiMi–8 layers	1.294	875s	129K samples/s	4.05M
Isotropic S4–4 layers	1.429	1900s	144K samples/s	2.83M
Isotropic S4–8 layers	1.524	3700s	72K samples/s	5.53M



Model	Test NLL	MOS (fidelity)	MOS (musicality)
SampleRNN	1.723	**2.98 ± 0.08**	1.82 ± 0.08
WaveNet	1.449	2.91 ± 0.08	2.71 ± 0.08
SaShiMi	**1.294**	2.84 ± 0.09	**3.11 ± 0.09**
Dataset	-	3.76 ± 0.08	4.59 ± 0.07

Слабые стороны

Как можно продолжить исследование?

Раскрыть тему TTS.

Тяжело обучать TTS модель на длинных последовательностях

Было бы круто если применить нашу модель к тому чтобы учитывать контекст любой ширины имея эффективность и скорость выполнения как в этой статье.