Exposé

Comparing raw audio generation algorithms on the task of music generation

Fabian Stahl

April 24, 2019

Deep neural networks have been known to solve a wide range problems. Often a models architecture is chosen based upon the topology of the input data. For example CNN-based networks perform especially good with image tasks, since the spatial structure of neighbouring pixels can be recognized. LSTMs are often used for input streams of text since previous states are important for the output.

In the field of audio generation there is a wide variety of approaches. Most of them work with MIDI data, a binary format specifying musical notes [12, 8, 10, 6]. This makes the format very slim, however the musical interpretation, (the raw audio sent to the sound port), is up to an external MIDI synthesizer.

When it comes to raw-audio synthesizes and moreover to music-generation using deep neural networks, there are only few approaches. Since most approaches have heavy downsides, there are no clear superior network architectures for this kind of data. For example, an already compressed 3 minute mp3 Song is about 3MB. Data of this size cannot be fed into most network architectures due to resource limitations. Microsoft's WaveNet architecture [11] for example, an autoregressive model intended for speech synthesis, is known to produce very realistic sounds, but takes hours to produce a single second of audio.

Another approach is to convert audio data to spectrogram images and use CNN-based networks intended for image data. Conversion back and forth can be done by using public tools such as ARSS [1]. However, converting spectrogram images back to audio data is known to be very lossy which results in poor audio quality, even if generated spectrogram images are realistic.

My aim for this project is to make myself familiar with recent sample-based approaches, and - if possible - implement them. This may include

- GAN-based architectures like WaveGAN / SpecGAN [4] (produces about a second of audio data) or Google's GANSynth [5](produces about 4 seconds of audio data).
- Music Auto-Encoders like [3, 9] or [7]
- Fully convolutional audio synthesis using the ARSS conversion tool.
- More approaches from [2]
- ...

Finally I want to compare different approaches, and evaluate results concerning realism, synthesis speed, musical coherency and creativity.

Sources

[1] The arss. `http://www.arss.sourceforge.net`. Accessed: 2019-03-19.

[2] Jean-Pierre Briot, Gaëtan Hadjeres, and François Pachet. Deep learning techniques for music generation-a survey. *arXiv preprint arXiv:1709.01620*, 2017.

[3] Joseph Colonel, Christopher Curro, and Sam Keene. Improving neural net auto encoders for music synthesis. In *Audio Engineering Society Convention 143*. Audio Engineering Society, 2017.

[4] Chris Donahue, Julian McAuley, and Miller Puckette. Adversarial audio synthesis. *arXiv preprint arXiv:1802.04208*, 2018.

[5] Jesse Engel, Kumar Krishna Agrawal, Shuo Chen, Ishaan Gulrajani, Chris Donahue, and Adam Roberts. Gansynth: Adversarial neural audio synthesis. *arXiv preprint arXiv:1902.08710*, 2019.

[6] Jay A Hennig, Akash Umakantha, and Ryan C Williamson. A classifying variational autoencoder with application to polyphonic music generation. *arXiv preprint arXiv:1711.07050*, 2017.

[7] Adam Roberts, Jesse Engel, and Douglas Eck. Hierarchical variational autoencoders for music. In *NIPS Workshop on Machine Learning for Creativity and Design*, 2017.

[8] Adam Roberts, Jesse Engel, Colin Raffel, Curtis Hawthorne, and Douglas Eck. A hierarchical latent vector model for learning long-term structure in music. *arXiv preprint arXiv:1803.05428*, 2018.

[9] Andy M Sarroff and Michael A Casey. Musical audio synthesis using autoencoding neural nets. In *ICMC*, 2014.

[10] Alexey Tikhonov and Ivan P Yamshchikov. Music generation with variational recurrent autoencoder supported by history. *arXiv preprint arXiv:1705.05458*, 2017.

[11] Aäron Van Den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew W Senior, and Koray Kavukcuoglu. Wavenet: A generative model for raw audio. *SSW*, 125, 2016.

[12] Li-Chia Yang, Szu-Yu Chou, and Yi-Hsuan Yang. Midinet: A convolutional generative adversarial network for symbolic-domain music generation. *arXiv preprint arXiv:1703.10847*, 2017.