Comparing raw audio generation algorithms on the task of music generation

Fabian Stahl

Hochschule **RheinMain**
University of Applied Sciences
Wiesbaden Rüsselsheim

Introduction

Music can be found in almost every movie, video game or public location. Its targeted use can change people's mood, encourage buying decisions or add context to accompanying content. However, composing, recording and mixing music is a creative process, that takes a lot of time and skill to master. This poses the question, if pleasing music can be generated autonomously.
Deep neural networks have been known to solve a wide range of problems. During this semester's project I will compare different network architectures to generate sample based music, show how they can be trained and rate results. Some of the most promising candidates can be found in [1].

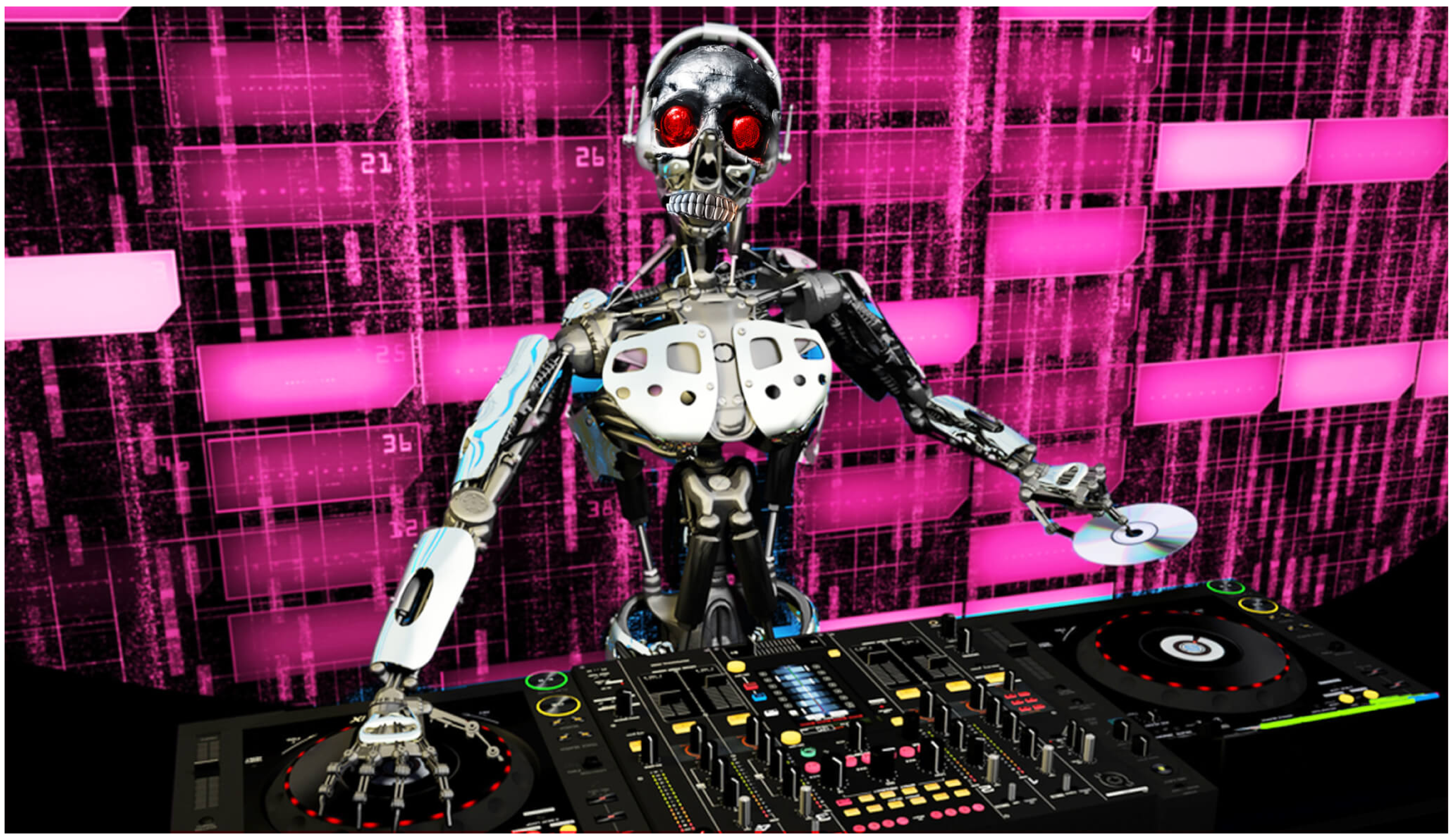


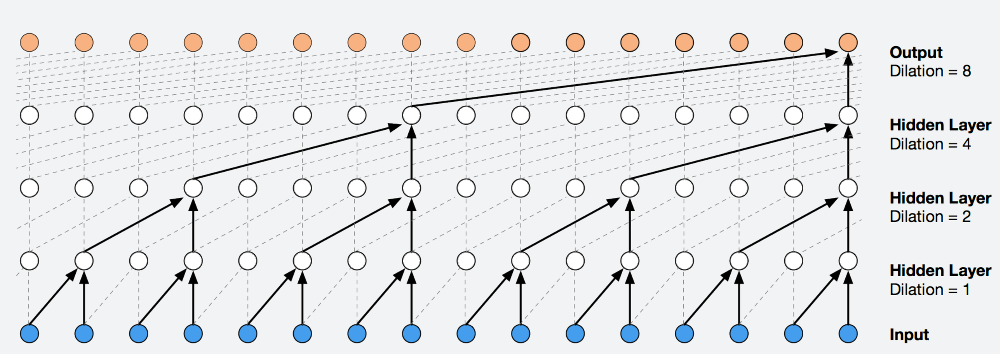
image source: `https://www.electronicbeats.net/app/uploads/2018/03/robotdj_spotify.jpg`

Related Work

The field of audio generation offers a lot of different research fields. For example, it was found, that deep neural networks perform excellent in the text-to-speech (TTS) domain [2, 3, 4, 5]. However, when it comes to music synthesis, these architectures seem to lack the ability to capture long term time structures which results, if not in noise, in rapidly changing motives.
Most approaches to generate music use MIDI data, a binary format specifying musical notes [6, 7, 8, 9]. This symbolic representation makes the format very slim, however the musical interpretation, (the raw audio sent to the sound port), is up to an external MIDI synthesizer.
Only few sample based approaches to generate music are known, that generate pleasing results. The most promising candidates are presented in the following.

CNN-based Approach

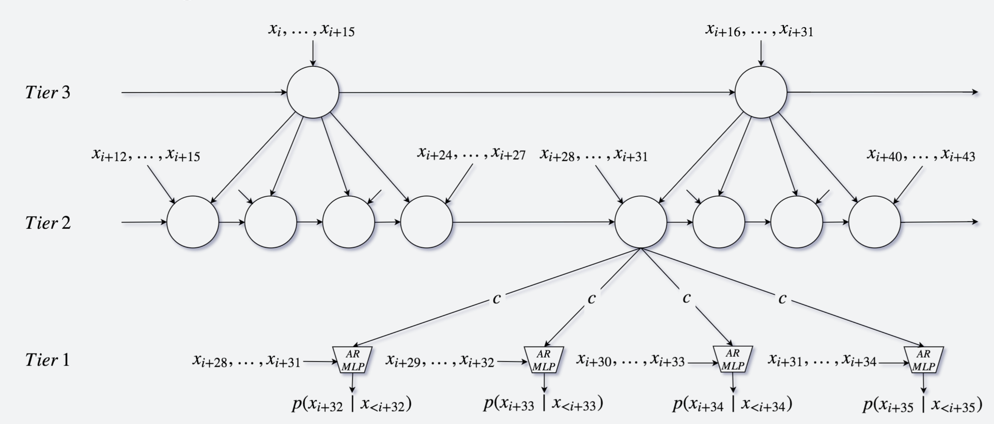
Convolutional Neural Networks (CNNs) are very popular in the field of computer vision because they ~~preserve~~ the spatial structure of input data. But They are rather uncommon when it comes to sequential data. take into account!
However, in 2016 a Microsoft research team presented State-of-the-Art TTS results using a fully convolutional neural network, called **WaveNet** [2]. Dilated convolutions were used to increase the field of perception (see Figure below). Experiments also included piano music synthesis. Unfortunately choosing the right hyperparameters for training is quite difficult, generation is slow and training takes several days on a cluster of GPUs.



simplified WaveNet architecture

RNN-based Approach

Recurrent Neural Networks (RNNs) are used to process sequences of data. Output data is not only based upon input data, but also on a hidden state vector, that encodes previous input.
A promising candidate for music generation is the **SampleRNN** [10] (see Figure below). Stacked *Frame-Level Modules* with different fields of perception are used to capture temporal context over different time periods (tier 2 and 3). A *Sample-Level Module* puts their state vector as well as the current input sample into consideration to generate a probability distribution for the next sample (tier 1). Originally tested only for piano music, [11] and [12] showed, that SampleRNN is especially good to generate loud music genres, like Metal and Dark Ambient. Its biggest drawback is, that sample quality varies strongly even after days of training.



simplified architecture of the SampleRNN network

GAN-based Approach

Generative Adversarial Neural Networks (GANs) are a new kind of architecture. While a Generator network tries to produce realistic data, a Discriminator network aims to separate real from generated data.
Donahue et al. proposes **SpecGAN**, which is based on convolutions of spectrogram images and **WaveGAN**, which is based on 1D convolutions on raw audio data [13]. Engel et al. suggests to use progressively growing GANs to improve training stability [14]. This **Gansynth** network also works on spectrogram images.
However, all of these approaches only work on fixed length audio input and output, which can be a drawback due to limited memory.

Future Work

- Implement and train more architectures
- Directly compare generated music from different networks that were trained on the same dataset
- Human evaluation

es fehlen ein paar aktuelle Ergebnisse!

References

[1] Deepsound.
`http://www.deepsound.io.`
Accessed: 2019-06-05.

[2] 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.

[3] Sercan Ö Arik, Mike Chrzanowski, Adam Coates, Gregory Diamos, Andrew Gibiansky, Yongguo Kang, Xian Li, John Miller, Andrew Ng, Jonathan Raiman, et al.
Deep voice: Real-time neural text-to-speech.
In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 195–204. JMLR. org, 2017.

[4] Junyoung Chung, Kyle Kastner, Laurent Dinh, Kratarth Goel, Aaron C Courville, and Yoshua Bengio.
A recurrent latent variable model for sequential data.
In *Advances in neural information processing systems*, pages 2980–2988, 2015.

[5] Nal Kalchbrenner, Erich Elsen, Karen Simonyan, Seb Noury, Norman Casagrande, Edward Lockhart, Florian Stimberg, Aaron van den Oord, Sander Dieleman, and Koray Kavukcuoglu.
Efficient neural audio synthesis.
arXiv preprint arXiv:1802.08435, 2018.

[6] 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.

[7] Olof Mogren.
C-rnn-gan: Continuous recurrent neural networks with adversarial training.
arXiv preprint arXiv:1611.09904, 2016.

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

[9] 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.

[10] Soroush Mehri, Kundan Kumar, Ishaan Gulrajani, Rithesh Kumar, Shubham Jain, Jose Sotelo, Aaron Courville, and Yoshua Bengio.
Samplernn: An unconditional end-to-end neural audio generation model.
arXiv preprint arXiv:1612.07837, 2016.

[11] Zack Zukowski and CJ Carr.
Generating black metal and math rock: Beyond bach, beethoven, and beatles.
arXiv preprint arXiv:1811.06639, 2018.

[12] CJ Carr and Zack Zukowski.
Generating albums with samplernn to imitate metal, rock, and punk bands.
arXiv preprint arXiv:1811.06633, 2018.

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

[14] 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.

fabian.stahl@student.hs-rm.de