ABSTRACT

This thesis deals with the use of state-of-the-art deep learning algorithms in the creative field of generating new content. The aim is to develop a deep learning model that can generate a random song in the electronic music genre without vocals. This is motivated by the objective to expand the existing customer base of the partner company of this thesis by presenting AI expertise on a fair or conference. The resulting model can serve as an exhibit.

There are three mayor ways to present audio data to a computer that are used for generative tasks. The physical analog waveform can be digitized by measuring the change in air pressure over time for a certain location, and then by taking a certain number of samples from these values each second. This resulting vector of values is the raw audio data. This can further be processed to higher level formats. Different authors tackled the task of generating music in various combinations of audio formats and deep learning architectures. The most intriguing is the WaveGAN [12] that uses a Generative Adversarial Network (GAN) to model raw audio data. It can only generate audio of one second. In 2018 an architecture called progressive growing of GANs [31] showed that GANs can be used to generate over one million pixels in a coherent image when trained iteratively. This magnitude would result in audio of over a minute. This thesis implements this algorithm for raw audio.

Raw audio is the most challenging format due to its high dimensional properties. To reduce computational cost, the generated song is kept at eleven seconds, resulting in 262, 144 samples for the final sampling rate. The generator uses a CNN architecture to upsample and analyze a latent vector to a song of eleven seconds. The discriminator is a mirrored version of the generator and downsamples a song to a single value, which indicates whether the discriminator thinks the song is real or fake. Both networks grow iteratively, starting with a low sampling rate and therefore fewer samples for the eleven seconds of audio. The sampling rate along with the output size of the generator and the input size of the discriminator is doubled every iteration. In that way the model first learns lower frequencies and global structure and later local details.

The results of the conducted model development and training showed that working with raw audio data has very high computational requirements. The model can successfully generate songs that can be steered by the user. The audio lacks global coherence and still contains noise. Increasing the kernel size of the convolutional networks could improve global structure but would increase the model size considerably. The occurring noise is mainly due to the small amount of training data and the shorter training time compared to networks of similar size. Some further development and training resources are needed to use the model as an exhibit for a fair or conference.