DILATED CONVOLUTIONS FOR MUSIC GENERATION

Kaleb Morris & Hannah Leou
Stanford Computer Science Department · CS 230: Deep Learning

Motivation

Art and science are often considered to be polar opposite disciplines. We would like to challenge this notion and investigate the intersection of these two fields through computer generated music.

Recurrent neural network architectures like LSTMS and GRUs are often plagued by slow training, and unnatural results. Recent research from leading suggests that CNNs are just as capable, if not more capable, at sequence modeling than RNNs if tweaked to increase their receptive fields. Therefore, we decided to apply the idea of dilated convolutional networks to the creative task of music generation.

Past & Related Work

Google DeepMind has made incredible strides with WaveNet, a deep generative model of raw audio waveforms. WaveNet, originally developed for applications in speech generation is now being expanded to applications in music generation through the DeepSound project.

Currently, the WaveNet model is trained on raw audio input and generates raw audio output. In our project, we decided to modify and restructure the WaveNet architecture such that the model is trained on MIDI files as input and generates MIDI files as output.

We believe this will work because MIDI files contain much more focused info related to music than raw audio, meaning that the CNN we are constructing will have to do significantly less work to generate equally natural music to that of a CNN operating on raw audio.

Training Data

Training data consisted of melodies extracted from classical artists such as Beethoven, Bach, Haydn, and Mozart.

Encoding Data

We encoded these melodies into tensors of dimensions (1,128). Each of the elements in the tensor correspond to any of 128 notes. 1's in the tensor denote that the note corresponding to that index is being played and 0's denote the opposite.

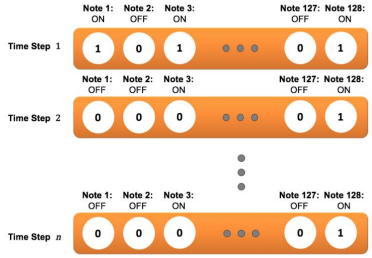

Figure 1: Visualization of encoding scheme for 128 note classes

Training Results

Figure 2: Visualization of training loss over epochs

The Model

We modeled our own architecture after the Google DeepMind's WaveNet, a dilated convolution (convolution where the filter is applied over an area larger than its length by skipping a constant number of inputs)

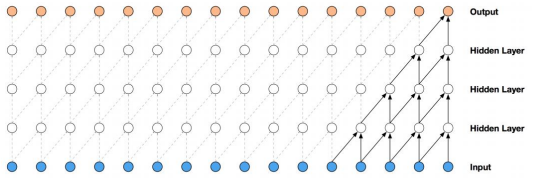

Figure 3: Visualization of a stack of causal convolutional layers.

The model we created takes as input MIDI files and generates as output MIDI files. In order to generate MIDI files rather than raw audio output, we replaced the final activation function in the original WaveNet with a sigmoid function in order to map the values for 128 note classes into a range between 0 and 1. We fine tuned the threshold such that note classes with values above a threshold would be switched on for a defined time-step and note classes with values below a threshold would be switched off for a defined time-step.

Results & Conclusion

Project Stage	Flow	Rhythm	Classical	Repetitive	Correct
Training on ONLY Mozart	3.2	2.0	3.3	1.0	1.8
Training on ONLY Haydn	3.1	2.1	3.0	1.1	2.1
Training on ONLY Beethoven	3.0	1.9	3.2	1.3	2.3
Training on ALL composers	2.6	2.2	2.9	.9	1.7

The music generated through our model did not sound as natural as we had initially thought it would, but certain aspects of the music definitely improved upon altering our model. Notably, the model's capability to generate more complex pieces consisting of notes, chords, and silences came about after altering the processing and feeding of data into our model. In future iterations, we would train on a much larger dataset, and we would test with different architectures, possibly adding residual and skip connections to improve memory capability.