Music Generation Using WaveNet

Sally Shin, William Krska, Yuke Li
salshin@bu.edu, wkrska@bu.edu, yukeli@bu.edu

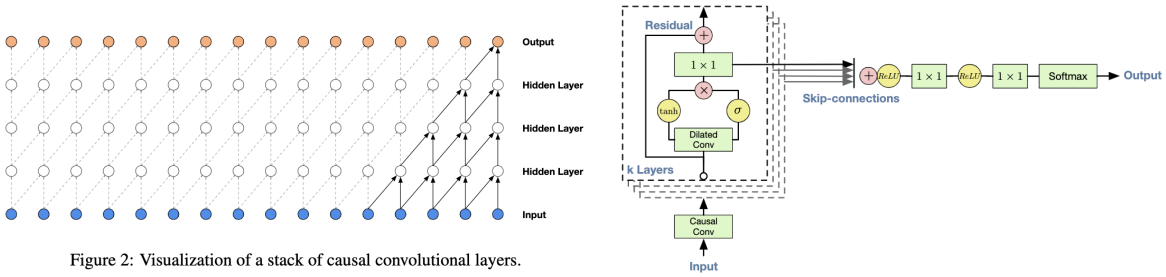

Figure 2: Visualization of a stack of causal convolutional layers.

1. Task

For this project, we'll be generating piano music using Google's WaveNet model. We plan to use MIDI as our input training data from the MAESTRO dataset to generate piano music that sounds natural on inspection and indistinguishable from human-made piano music, in 5-10 second intervals. The goal is to produce performances that are musically plausible and coherent, and that capture the style and characteristics of the training data. We will also evaluate our generated music using several metrics, including note accuracy, rhythmic accuracy, pitch coherence, and human evaluation.

2. Related Work

"Wavenet: A Generative Model For Raw Audio" by Aaron van den Oord et. al.

In this work, the authors introduce the WaveNet, a fully probabilistic and autoregressive model that produces a predictive audio distribution that takes into account all previous audio. The authors describe the general architecture of the model to be a series of convolutional layers without any pooling layers in between. The input length is the same as the output length. The key point for this model is that they utilize causal convolution to reduce the computational load and time in training. To reduce the number of layers needed, the authors also utilize dilated convolutions to increase the receptive field by implementing a coarser scale than regular convolutions.

"Enabling Factorized Piano Music Modeling and Generation with the MAESTRO Dataset" by Christopher Hawthorne et al.

In this work, the authors use the WaveNet model with MAESTRO dataset and design a state-of-the-art architecture to produce longer music sequences that sound natural. The factorized hierarchical model separates the musical structure, timing, and dynamics of piano performances into different factors, which are modeled independently at each level. This allows for greater flexibility in generating new piano performances, as different factors can be combined in different ways. The authors evaluate the music based on musicality, coherence, and expressiveness. A user survey was also conducted.

"DeepJ: Style-Specific Music Generation" by Huanru Henry Mao et. al.

In this work, the authors look to improve on previous efforts to create a deep neural network that can compose music in a specific style. Prior

architectures such as the Biaxial LSTM used a more simplistic method of representing notes, only storing their “on” or “off” values. The authors, still using a biaxial LSTM, wanted to make the music tunable, and include more aspects of music such as dynamics to better mimic human created music. The notes are stored in an NxT matrix, where N is the number of possible notes, and T is that notes value at a given time sample.

3. Approach

We plan to use the codebase from Hawthorne et al. using a modified version of their preprocessing pipeline to separate the MAESTRO dataset into training, validation, and testing datasets, for both audio and midi versions of the audio. We’ll also use their approach in downsampling and vectorizing the music into numpy arrays for input into our model.

We then plan to input either the audio or midi into our WaveNet model and train over the entire training dataset for multiple epochs, and validating for underfitting or overfitting of the data. Time and dataset compatibility permitting, we can apply this method to either different instruments or different genres of music.

We are hoping to generate 5-10 seconds of generated piano music, and aiming for longer musical pieces without lowering our evaluation metrics.

4. Dataset and Metric

We will be using MAESTRO, a dataset that is composed of about 200 hours of piano performances. These audio recordings have been meticulously labeled with the appropriate MIDI note values, which are each composed of pitch, velocity, onset time, and offset time. A training, validation, and testing set are split into 5.66, 0.64, and 0.74 million notes respectively.

Metrics for generated music is still debated between the different resources we read. For now, we plan to evaluate our model's performance using note accuracy, rhythmic accuracy, pitch coherence, and subjective human evaluation. Note accuracy measures the match between generated and ground-truth notes, while rhythmic accuracy measures the alignment of generated notes with the beat and bar structure of the ground-truth. Pitch coherence measures the similarity of pitch distribution in generated and ground-truth performances. Along with these quantitative metrics, we will also perform subjective human evaluation.

5. Preliminary Results

We’ve unfortunately had to switch our topic from our previous one involving piano music transcription task. We found the previous topic to be too difficult and too novel for the field for our level of expertise. For this intermediate report, we’ve defined our new project direction, found base code to work with, and references to base our project direction off of.

Code implementation that has been completed so far is our work on our preprocessing pipeline of working with MAESTRO audio and midi files.

6. Approximate Timeline

Task	**Deadline**
Get dataset to use - COMPLETED	Start/Mid-March
Read more background - Almost Complete	Start of April
Test out/understand original codebase - In Progress	Start of April
Training our model	Start-Mid April
Fine-tune training, get evaluation metrics	Mid-End April
Prepare presentation and report	End of April

7. Preliminary Code

GITHUB:
https://github.com/sassmander/EC523_MusicGeneration

8. References

1) "Enabling factorized piano music modeling and generation with the Maestro dataset" by Hawthorne et al. 2019. https://arxiv.org/pdf/1810.12247.pdf
2) "WaveNet: A Generative Model for Raw Audio" by Oord et al. 16 Sep 2016. https://arxiv.org/pdf/1609.03499.pdf
3) "On the evaluation of generative models in music" by Yang and Lerch. 2018. https://musicinformatics.gatech.edu/wp-content_nondefault/uploads/2018/11/postprint.pdf
4) "DeepJ: Style-Specific Music Generation" by Mao et al. 2018. https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8334500