Implementing Audio-Based Music Generation and Enhancing the Model through Quantization

FatemehSadat Mireshghallah, Hao Liu, Rithesh Ramapura

January 2019

Speech recognition systems are broadly used in different domains, ranging from voice assistants to music classification and generation. These systems take in waveforms, extract feature vectors from them, model the sequences of these vectors and align them with phones and phonemes, then with the aid of a language model, turn the speech to words and sentences. Algorithms can be applied to detect the subject of speech, and also generate speech based on a training corpus.

What we want to achieve in this project is to come up with a music generation network. Existing music generation models that use deep learning can be broadly classified into two approaches: raw audio models and symbolic models. Although Symbolic models are currently more prevalent and can capture long-range dependencies of melodic structure. However, it fails to grasp the subtle features and richness of raw audio information. In our model, which we want to base on WaveNet, which is trained directly on sampled audio waveforms and can produce some unstructured but realistic music. We want to use The MAESTRO Dataset, which is composed of over 172 hours of virtuosic piano performances captured with fine alignment (3 ms) between note labels and audio waveforms. Here our goal is to explore ways to produce more fluent and realistic music.

In the perspective of enhancing the network performance, we want to have it quantized to lower bitwidths so we can achieve a smaller network that could be easily deployed to edge devices and could generate a response faster which is a recent trend in neural networks. Quantizing neural networks to lower bitwidths (8 bits and lower fixed-point presentations for weights and also activations) helps reduce their memory footprint, and also alleviates their high computational costs (compared to 32-bit floating point operations). If this quantization is done diligently, there would not be noticeable loss in accuracy. Most modern CPUs and GPUs support 8bit computation, which can be employed for our projects. We would first like to investigate the accuracy loss due to quantization by emulating it, then, time permitting we plan to implement our network using 8 or 16 bit instructions.