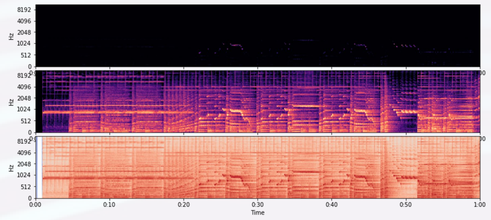
Music Generation with Deep Learning

Swapnil Agrawal*, Tanmay J. Raj*, Bhiksha Raj†
*Department of Chemical Engineering, Carnegie Mellon University
†Language Technologies Institute, School of Computer Science, Carnegie Mellon University

Language Technologies Institute

Abstract

The use of Deep Learning techniques to solve problems in the field of literary arts has gained recent interest among researchers. Generating musical content that is both- pleasant to hear and obeys the rules and structure of music theory is the biggest challenge. The ulterior motive of this task is to enhance the creativity of an artist and produce new compositions that are similar in style and dynamics to an original artist. In this project, we attempt to translate the task of music generation to a language modelling problem. We apply the 'bag-of-frames' approach and vector quantize the log mel-spectrograms of raw audio files into a vector of 'symbols' that can now be interpreted as a language sequence.



Literature Review

There have been several attempts of generating music. The earliest attempt by Chen et al. generated music with only one melody and no harmony.
More recent work where musical features such as notes, chords and notations have been used to generate music using LSTMs.

LSTMs

2002 Eck et al. use 2 LSTM networks to learn chord structure and local note structure

2016 Magenta by Google Brain developed an LSTM model tuned with Reinforcement Learning

CNNs

2016 Wavenet applies CNN with multiple dilation factors and

MuseGAN applies CNN to generate bars of music rather than notes

2017

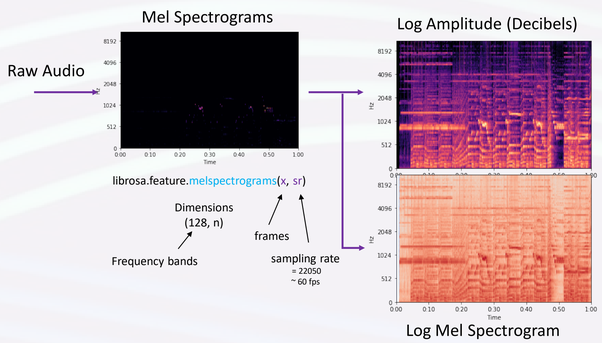
NLP architecture and Transformers

2019 MuseNet by OpenAI applied large-scale transformer model to predict next token in a sequence of encoded music notations

Approach

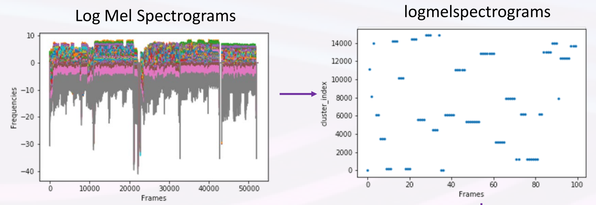
Proof of Concept

Audio Preprocessing



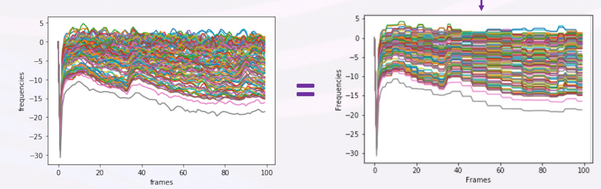
Vector Quantization

K Means Clustering



km = cluster.KMeans(*init*='k-means++', *n_clusters*=15000, *n_init*=3)
km.fit(flatlogmel.T)

Generate Codebooks

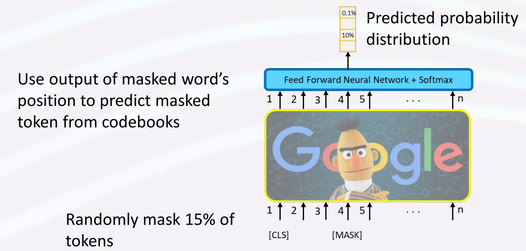


Pre-Training BERT on VQ data

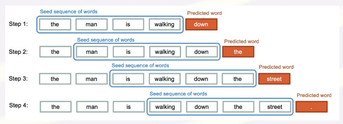
Trained for 10 epochs with Adam Optimizer and minimized Cross Entropy Loss between actual masked token and probability distribution of predicted tokens for masked input

BERT Generation

Bidirectional **E**ncoder **R**epresentations from **T**ransformers



We use NLP generation techniques to predict the masked token based on the previous tokens.



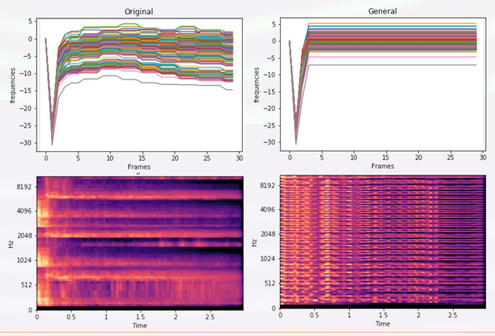
Results & Discussion

The music generation task is carried by providing first two tokens as input. Using the first two tokens, the next token is predicted which is then used recursively along the previous tokens to predict the next token.
The tokens are converted back to log Mel Spectrograms from the codebooks generated during Vector Quantization and the results are below.

Song Sample 1: Low Variation

The generated music approximates the subsequent tokens but due to lower variation, is unable to distinguish nearby frequencies.

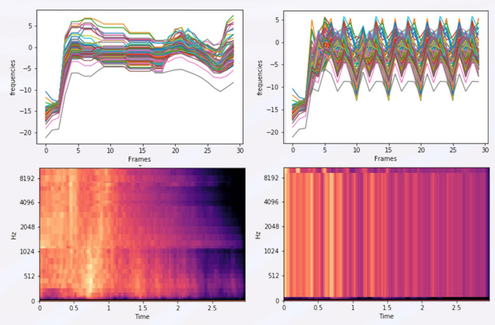
This results in the pattern being able to capture the initial trend and failing to generate any long-term relationships



Song Sample 2: High Variation

The generated music is able to capture the pattern of music over short time.

The input sequence triggers the generation of tokens similar to the original piece and then begins to repeat the sequence.



Future Work

- Train BERT on larger dataset and for more epochs to better learn the structure of music
- Apply models such as GPT-2, XLNet, Albert and compare their performance with BERT for music generation

References

Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova, Google AI Language; BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, 2018

Makhoul, John, Salim Roucos, and Herbert Gish. "Vector quantization in speech coding." Proceedings of the IEEE 73.11 (1985): 1551-1588