DeepMix

Joseph Ballouz, Ashwin Neerabail, William Park
{jballouz, ashwinn, wpark73}@stanford.edu
Computer Science Department

Introduction

In this project, our goal is to mix songs using deep learning. Specifically, the generated audio mix will have content (i.e. tempo, melody) of one audio clip and the style of another audio clip (i.e. instrument types, pitch). The way this is done is by applying a Fourier transform to generate a spectogram that is fed to a CNN and optimized to reduce the loss and we take an Inverse Fourier transform to generate the output audio file.

2D CNN of 1 layer with Random weights and 4096 filters works really well for monphonic sounds. We are able to take an audio with a primary instrument (like Ukelele) and replace it with other instruments like Piano/Guitar/Bird Sound.

Data

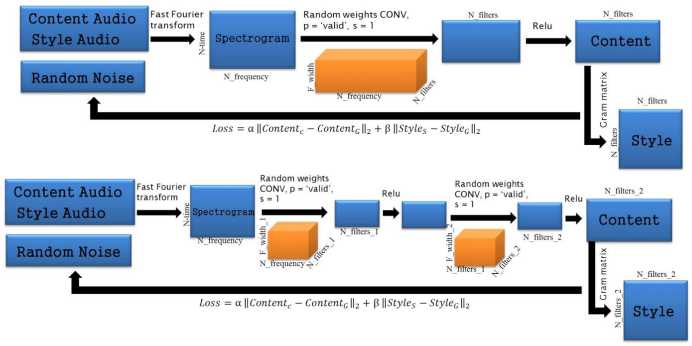
For this project we have used Random weighted CNN and pretrained models. So we did not need any data for training. For mixing , we have collected 10 second style and content clips from the royalty free website www.bensound.com.

Features

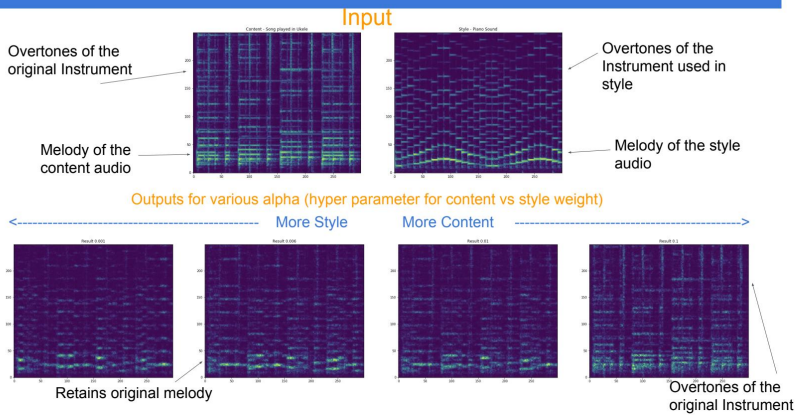
Content Extraction : 2D CNN of 1 layer with Random weights and 4096 filters is used for extracting melody/tempo (content)

Style Extraction : Gram Matrix is used for extracting timber (style).

Models



Results



Discussion

✔ Shallow CNN with Random-Weights performs really well for texture synthesis for audio content extraction.

✔ We get really good results for monophonic sounds like generating original melody played in the style of a different instrument.

✗ Polyphonic sounds and human voice transfer dont work well.

✗ Sound quality of Synthesized audio has room for improvement.

Future Work

Style for audio files is not a well defined paradigm. We have used timbre of the instrument to represent style in our work. This works well for monophonic sounds. Further research is needed to extract human voice (to replace one voice with another) as well as complex polyphonic sounds.

Using a pretrained network for audio like Google Magenta to learn compositional style of an artist to generate music interpretation from a different composer for the same input song is something that needs further exploration.

References

1. Neural Style Transfer for Audio Spectrograms by Prateek Verma, Julius O. Smith
https://arxiv.org/pdf/1801.01589.pdf
2. Audio texture synthesis and style transfer by Dmitry Ulyanov and Vadim Lebedev
https://dmitryulyanov.github.io/audio-texture-synthesis-and-style-transfer/
3. Audio Style Transfer by Eric Grinstein, Ngoc Q. K. Duong, Alexey Ozerov and Patrick Perez
https://arxiv.org/pdf/1710.11385.pdf
4. Image Style Transfer Using Convolutional Neural Networks by Leon A. Gatys, Alexander S. Ecker, Matthias Bethge
https://ieeexplore.ieee.org/document/7780634
5. A Powerful Generative Model Using Random Weights for the Deep Image Representation by Kun He, Yan Wang, John Hopcroft
https://arxiv.org/abs/1606.04801