Computer Science Tripos – Part II – Progress Report

Music Style Transfer

Charles J.Y. Yoon

jyy24@cam.ac.uk

1 Feburary 2019

Originator: Prof. Alan Blackwell

Project Supervisor: Dr. Andrea Franceschini

Director of Studies: Dr. Timothy Griffin

Project Overseers: Prof. Lawrence Paulson & Prof. Frank Stajano

Progress

Currently, the project is behind schedule, of approximately 3 weeks. This is due to the lack of work during Michaelmas; since I had 2 Paper 10 units of assessment assigned for Michaelmas, they left no time for dissertation work during the term. Moreover, research and understanding of Deep neural network structures and Generative Adversarial Networks took longer than expected, with additional unforeseen work of understanding audio signal processing. However, data collection is now mostly over; playlists of similar timbre, which is to be used by CycleGAN[1] methods, has been compiled by the Free Music Association[2]. Discography of Boyce Avenue and their original songs are purchased for the methods using `pix2pix` type transforms. Lastly, I am currently liaising with an Youtube music channel, for their “chill & study” type music. Moreover, image style transfers have been implemented, and will be used for my first method. Alike Timbretron[3], I am currently implementing a workflow of converting songs into images using SQT, training the image transfer using CycleGAN[1], and converting the resulting spectrum image using WaveNet[4].

Difficulties

Starting this project, I quickly realised the necessity of qualifying the meaning of “Musical Style”. Although the effectiveness can be interpreted using human perception experiment, the aim of the project needs to be established more specifically; therefore I am currently looking towards defining “style” as Timbre (instrument), and audio processor (EQ, Compressor) style. Moreover, the type of songs that will be used had to be classified separately, specifically to:

- Monotimbral monophony (e.g. Instrument Solo)
- Monotimbral polyphony (e.g. Piano covers)

- Polytimbral polyphony (i.e. Fully mixed songs)

The general aim is for a classifier that works for all three types of songs, but for evaluation the accuracies may be tested for each type separately. There also has been a difficulty in computing resources; GANs and RNNs tend to take significant time to train (up to 8 days). If needed, I am looking to using Google Cloud or AWS GPU instances in order to mitigate this issue.

Going Forward

There are still a number of different neural network structures I intend to experiment. Instead of using image style transfer, it seems possible to use causal convolutional neural networks, as used for WaveNet[4], for audio style transfer. RNNs and bidirectional RNNs work best for temporal data, so they are an option in constructing a GAN structure. These methods should generate a classifier with better performance and reduced complexity. I intend to implement different versions of working classifiers this month (Feburary) and conduct experiments in March to get back on track.

References

[1] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, “Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks,” *CoRR*, vol. abs/1703.1, 2017.

[2] M. Defferrard, K. Benzi, P. Vandergheynst, and X. Bresson, “FMA: A dataset for music analysis,” in *18th international society for music information retrieval conference*, 2017.

[3] S. Huang, Q. Li, C. Anil, X. Bao, S. Oore, and R. B. Grosse, “TimbreTron: A wavenet(CycleGAN(CQT(Audio))) pipeline for musical timbre transfer,” *CoRR*, vol. abs/1811.0, 2018.

[4] A. van den Oord *et al.*, “WaveNet: A Generative Model for Raw Audio,” *CoRR*, vol. abs/1609.0, 2016.