# Attention-based Convolutional Neural Network for Music Genre Classification

Tingyi Li(tingyi@kth.se)

## Abstract

In this project, I use convolutional neural networks (CNNs) and try different parameter settings to build and train a model for music genre classification. After that, I focus on developing an attention-based CNN model derived from my baseline models to see the effect of attention mechanism. For evaluation, I provide confusion matrices to measure the performance of my models.

## Model

The baseline models we chose are from Choi et al [1] and Zhang et al [2]. For the first part, I focus on improving the CNN models composed of four convolutional layers, four max pooling layers.

In the second part, I add an additional attention layer to the previously improved CNN. The attention layer computes the weighted sum of all the information extracted from different parts of the input. The output from the max pooling layer and the attention vector are then fed into a fully connected softmax layer.

## Data

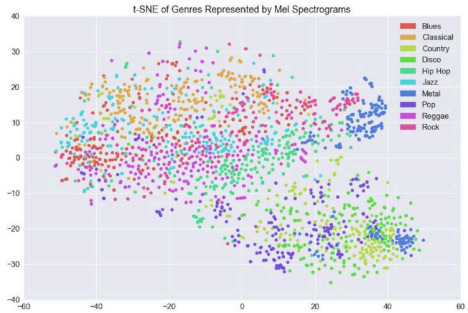
With audio and genres from FreeMusicArchive, I obtained 75 songs (30sec samples) each for the 10 genres: **Blues, Classical, Country, Disco, Hip-hop, Jazz, Metal, Pop, Reggae, Rock.**

I broke them down to three labels: **Sad, Neutral and Happy**.

## Preprocessing

For each song, I converted the audio to Mel-spectrograms and broke the 30 seconds samples down to 10 slices of 3seconds each. I chose **Mel-spectrograms** because it is optimized for human auditory perception and more efficient in size while preserving the most perceptually important information.

## Results & Conclusion

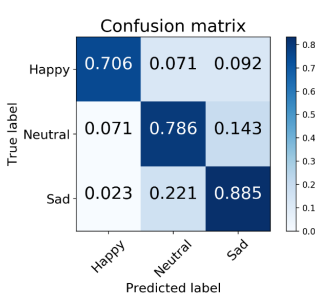


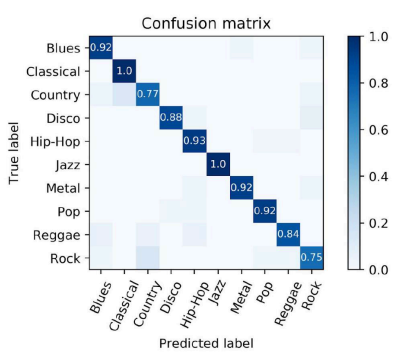

Before training the basic CNN, I primarily run t-SNE on my datasets in order to reduce dimensionality and preserve the similarities of my data samples.

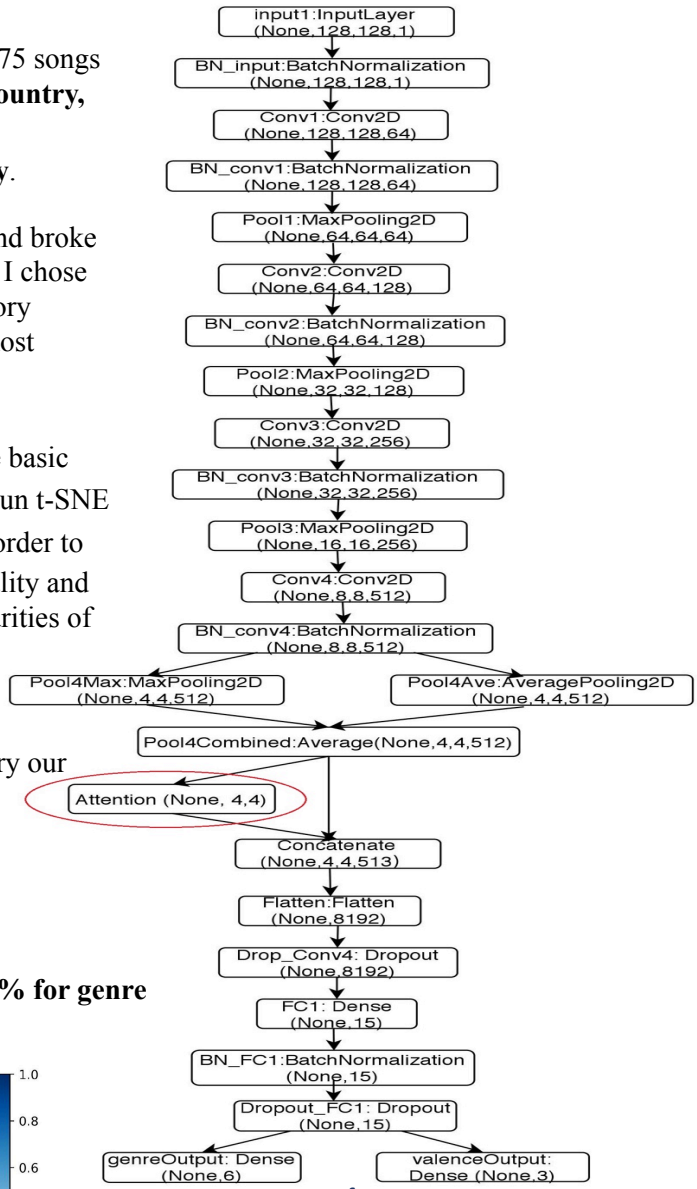
I did a grid search to optimize the performance and then carry our comparison with baseline models:

| Experiment | Genre | | Valence | |
|---|---|---|---|---|
| Model | TrainAcc | TestAcc | TrainAcc | TestAcc |
| 3 layer | 0.473 | 0.384 | 0.482 | 0.402 |
| 5 layer | 0.968 | 0.820 | 0.953 | 0.712 |
| 4 layer(this project) | 0.908 | 0.889 | 0.907 | 0.808 |

For the attention-based CNN, **I reach an accuracy of 89.78% for genre classification and 81.2% for mood classfication**.




## Reference

**[1]** Choi, K. & Fazekas, G. & Cho, K. & Sandler, M.B. (2017). A Tutorial on Deep Learning for Music Information Retrieval. CoRR, abs/1709.04396.

**[2]** Zhang, W. & Lei, W. & Xu, X. & Xing, X. (2016). mproved Music Genre Classification with Convolutional Neural Networks. INTERSPEECH.