FROM MUSIC TO SEMANTICS: AUTOMATICALLY GENERATING TIME-VARYING SEMANTIC TAGS FROM MUSIC

Tong Shan

Biomedical Engineering Deparmemt, University of Rochester

Abstract

This project aimed to use Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM) cell to build a model to predict time-varying tags for music.

- **CAL500exp** was used. There are 500 western pop songs and semantic time-varying labels in the dataset. The dataset was perprocessed by extracting Mel-spectrogram and then attached with 18 emotional time-varying labels.
- A Convolutional Neural Network (CNN) and a CNN combining Long Short-term Memory (LSTM) cell was constructed to predict the time-varying labels.
- The work need further improvement, and it paves a way for understanding the semantic meaning of music that related to human brain processing of music.

Introduction

Music tagging is a music information retrieval (MIR) task that gives music descriptive tags based on music content and its metadata.

Since digital music has been more and more popular, music applications like Spotify and Apple Music are now developing music recommendation systems to their users based on content and tags of the mus. With the development of machine learning, music auto-tagging has been developed for several years to meet such a demand. However, all of the methods could generate tags fora whole piece of music, which does not make sense since most music has time-varying semantic representations, especially for a symphony or a movie original sound track.The instrument, emotion, and even vocal artist could change over time. It has been shown that only track-level tagging is not enough since different segment of music tent to have different tags. Thus, time-varying auto-tagging is in need.

Dataset

CAL500exp introduced by Wang et al [1] was used in this project. The data is adapted from **CAL500** dataset [2].

- 500 western pop music.
- Each track was segmented into several segmentation (3-16 secs per segmentation) depends on the similarity. The labels were given for each segmentation.
- Each segment was tagged with 67 binary semantic labels including emotion, genre, instrument, instrument solo, vocal style, song characteristic.

Data Preprocessing

- **Feature Extraction** Mel-spectrogram with 128 bins (with 2048 window size and 1024 hop size).
- **Assigning Labels to Segmentation** For each track, labels were converted to map each frame so that each frame has a corresponding label. An example is shown in Figure 1. Only the 18 emotion labels were used afterwards.
- **Long-term Frame Segmentation** Long-term frame was then aggregated by a window of 128 frames and hop size of 64 frames.
- **Splitting Dataset** The 500 tracks in the **CAL500exp** dataset were split randomly: 400 tracks in train set, 50 in validation set, and 50 in test set.

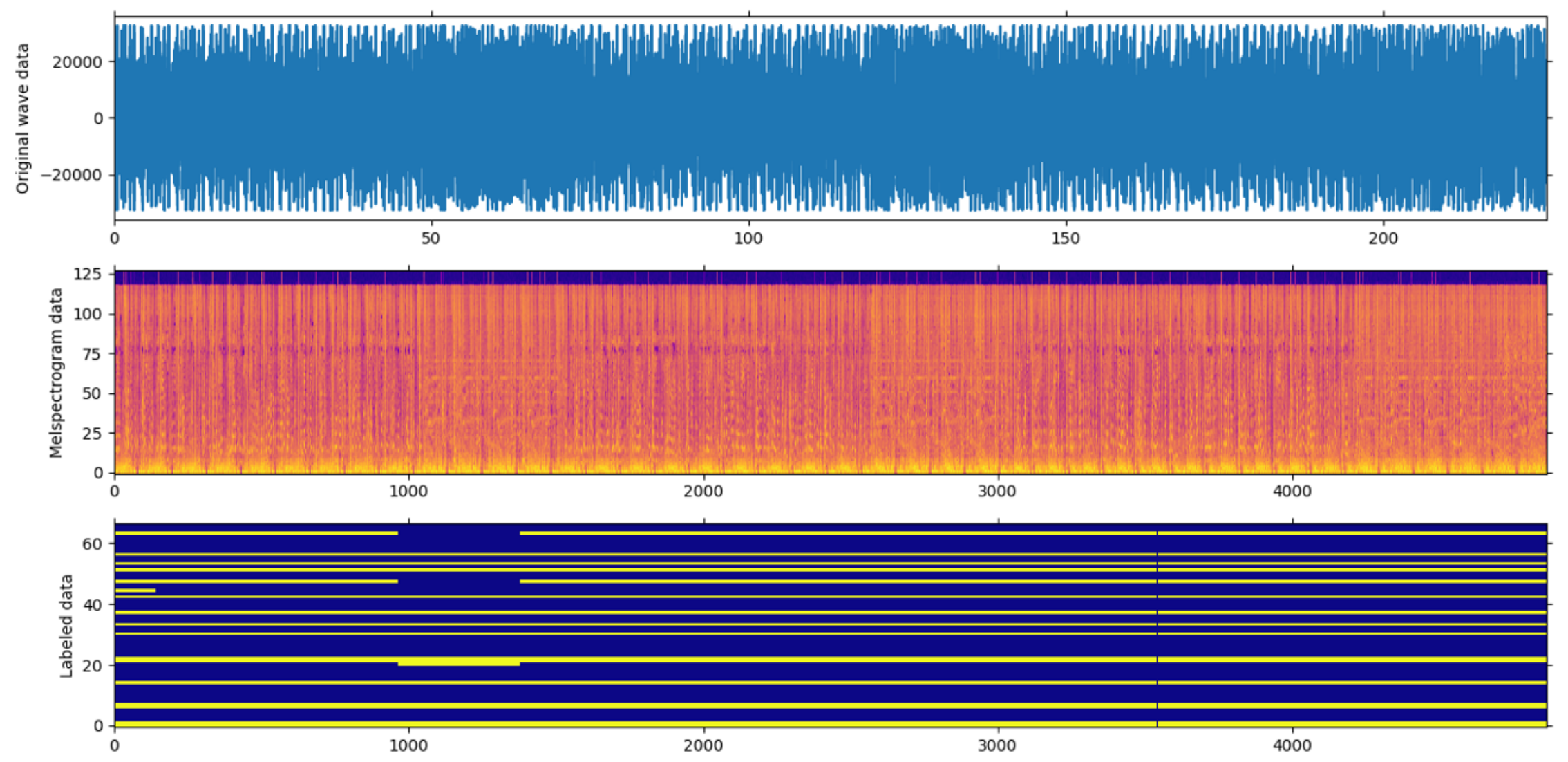

Figure 1:Example of preprocessed track

Neural Network Models

CNN only Model Four convolutional layer with two maxpooling layer were used to construct the model. ReLU activation was used after each convolutional layer. Data were then fed in fully connected layer with 1000 neurons, 800 neurons and 600 neurons. Finally, the output layer had 18 neurons corresponding to 18 labels. Sigmoid activation is used after output layer, *Adam* optimizer with learning rate of $1 * 10^-5$ were used to train. *Multi Label Margin Loss* were used as loss function.

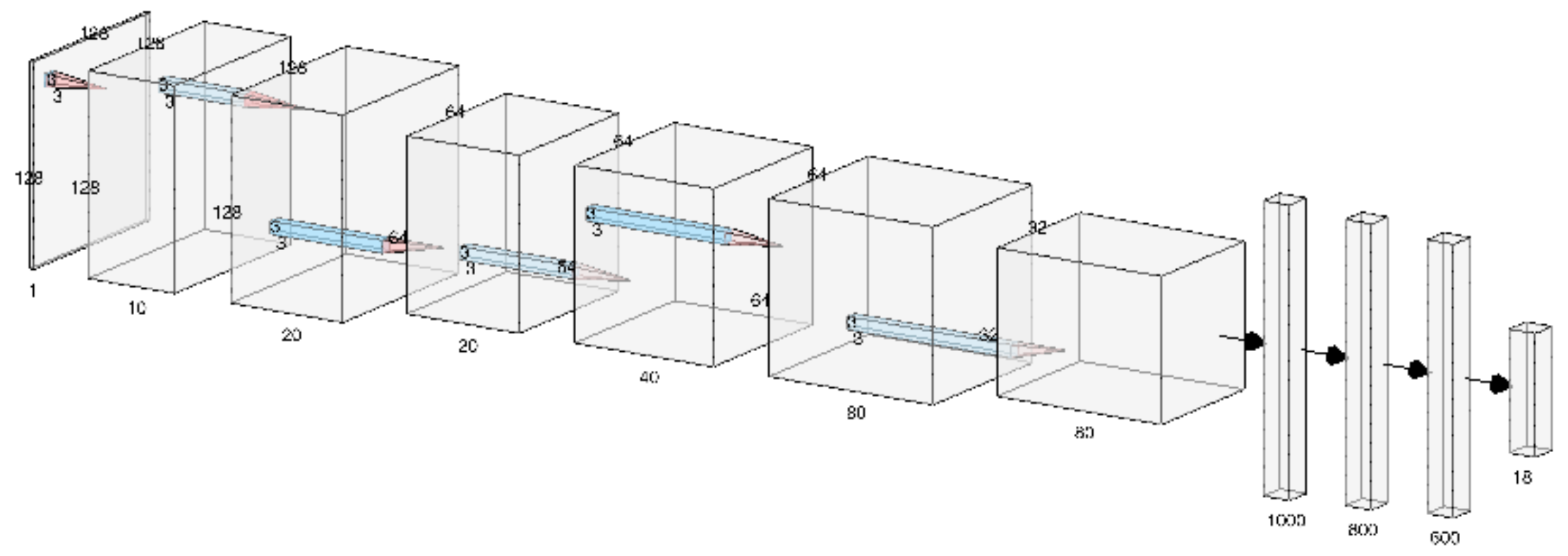

Figure 2:CNN only model.

CNN Combined with LSTM Model The architecture of this model is similar to the CNN only model. The only difference happens after the first fully connected layer. One LSTM cell was implemented after the fully connected layer with an output of 500. Then output from the layer before LSTM was concatenated with the output from LSTM layer and fed into the next fully connected layer.

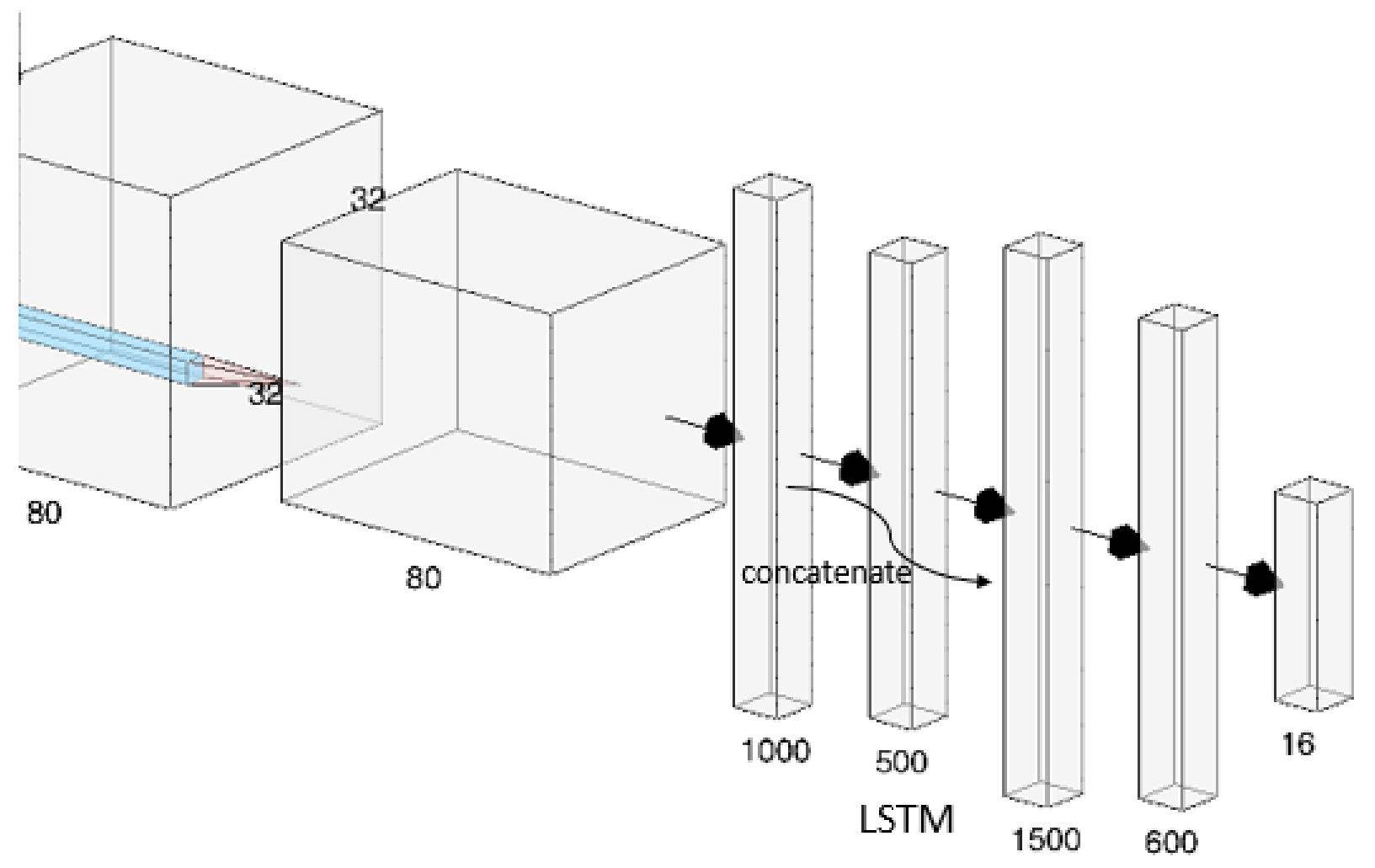

Figure 3:CNN combined with LSTM model.

Evaluation

Mean average accuracy (MAA) is computed for every label. MAA is defined as

$$MAA_l = \frac{1}{N}\sum_N \frac{TP_{nl} + TN_{nl}}{TP_{nl} + TN_{nl} + FP_{nl} + FN_{nl}}$$

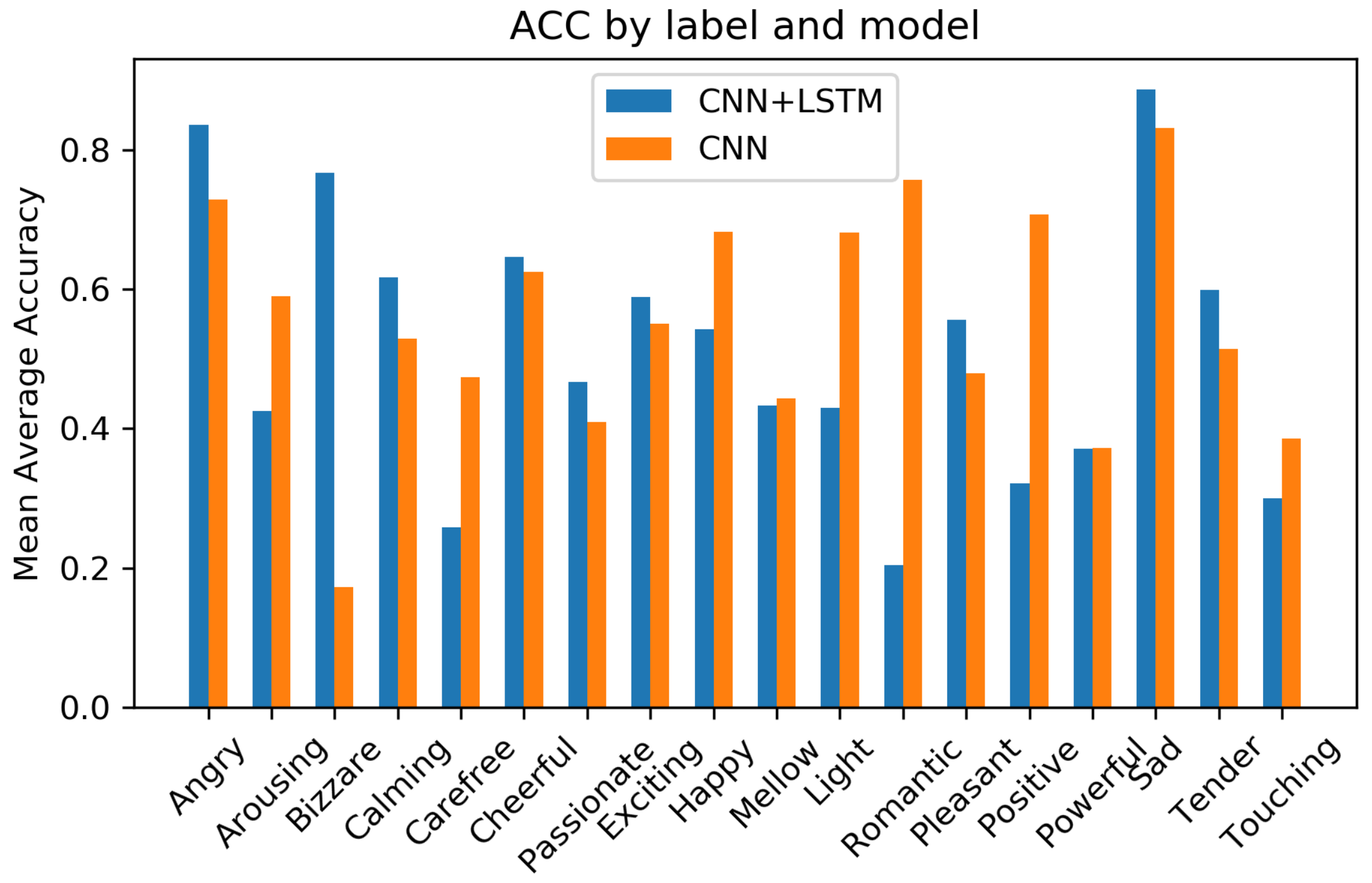

Figure 4:Mean average accuracy for CNN only and CNN combined with LSTM models.

Discussion

Improvement could be done in future

- Log-scale Mel-spectrogram might be a better choice for feature
- Normalization and deeper layers

Motivation for Neuroscience aspect It paves a way for understanding the semantic meaning of music that related to human brain processing of music and the relationship between music and speech.

References

[1] S. Wang, J. Wang, Y. Yang, and H. Wang, "Towards time-varying music auto-tagging based on cal500 expansion," in *2014 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1–6, July 2014.

[2] D. Turnbull, L. Barrington, D. Torres, and G. Lanckriet, "Semantic annotation and retrieval of music and sound effects," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 16, pp. 467–476, February 2008.