Music Tagging With Convolutional Neural Network

Xiao Fei Yu (xfy@stanford.edu)

Overview

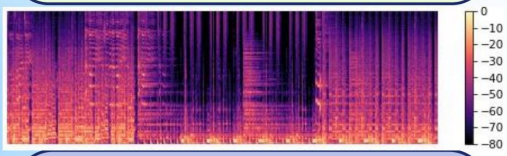
Using Deep Learning for Music Information Retrieval has garnered a lot of attention recently as it is a lot more effective and automated than traditional MIR techniques, which lack universality and are difficult to design. Some MIR techniques have also been made proprietary such as Spotify's audio features. This paper aims to simultaneously classify the genre as well as valence (mood) of the audio by using a multi-output CNN to learn the features of mel-spectrograms generated from the audio.

Data

With audio and genres from FreeMusicArchive, Iobtained 75 songs (30sec samples) each for the genres: Rock, Hip-Hop, Pop, Folk, Instrumental and Electronic. For each of these songs, I then used Spotify API to obtain their valence metric. The valence metrics range from 0 to 1, with 0 being sad and 1 being happy. I broke them down to three labels: Sad, Neutral and Happy.

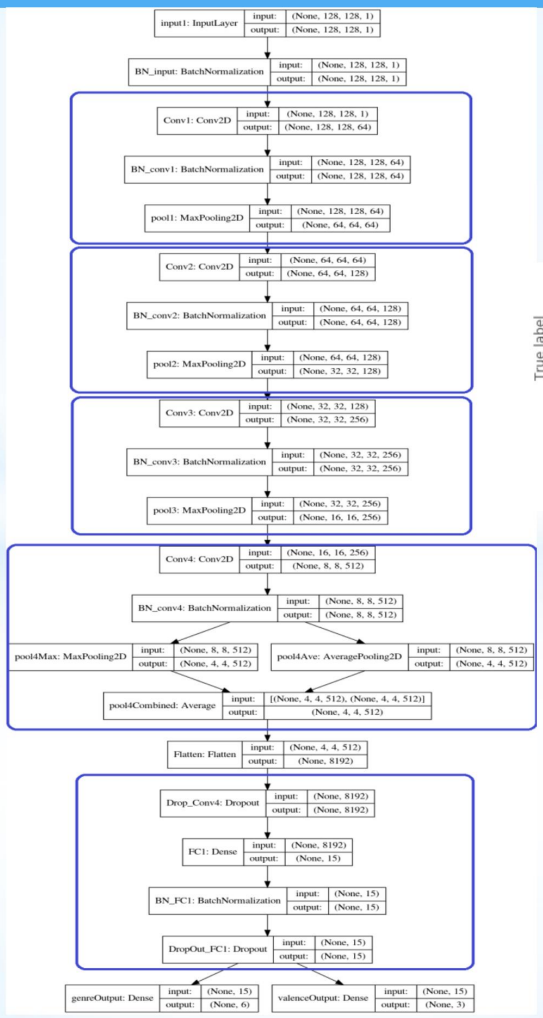
Preprocessing

For each song, I converted the audio to Mel-Spectrograms and broke the 30 seconds samples down to 10 slices of 3 seconds 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. It provides the energy of the frequency bins through time. I used gray scale in the data as color does not add any new information. Example:



Model

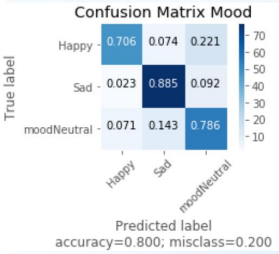
Drawing inspiration from Zhang et al and Choi et al as base models, I tried different number of convolution layers, strides, kernel sizes, FC layers and nodes. The model uses 4 convolution layers and 1 FC layer and 1 decision layer for each output. I use kernel size of 2x2 for all layers and stride of 1 for all except the 4th layer where I use stride of 2. I regularize the flattened and the FC layers and Batch normal every layer. With Learning rate = .001, l2 lambda = .002, Batch size = 75. See image in center:

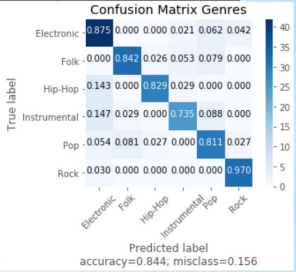


Model Comparisons

Training set has 4074 samples while test and dev sets each have 225 samples. They have been randomly chosen to ensure balanced number of labels. I see that Choi et al's model applied to the data set over fits while Zhang et al's underfits. This project's model achieves a good balance.

	Genre		Mood	
Model	**TrainAccuracy**	**TestAccuracy**	**TrainAccuracy**	**TestAccuracy**
Zhang et al	0.476	0.333	0.478	0.42
Choi et al	0.968	0.79	0.953	0.702
This Project	0.908	0.844	0.907	0.8





Discussion

Accuracy is calculated as number of songs categorized correctly divide by total number of songs. The genre classification overall is on par with state of the art (~85%). Instrumental songs have the worst performance (73.5%) and are often mistaken as electronic. This makes sense intuitively as these two types of songs are similar. Happy songs and mood-neutral songs suffer with 70.6% and 78.6% accuracy respectively. This is because there are less happy songs in the data sets (only 25% are Happy while the optimal number would be 33%), expanding and balancing the dataset in future works can potentially fix this issue.

Future

One can extend the dataset and balance the number of songs in each mood category to increase accuracy. Also, one can extend this model to predict danceability, energy and other Spotify features. It would be interesting to examine what each of the 15 nodes of the FC layer represent, and see whether this vector can be used for music recommendation, where songs with similar vectors are recommended.

References

- Keunwoo Choi, Gyorgy Fazekas, Kyunghyun Cho, and MarkS, "A Tutorial on Deep Learning for Music Information Retrieval," May 2018.
- Iibin Zhang, Ikang Lei, Xiangmin Xu, and Xiaofeng Xing, "Improved Music Genre Classification with Convolutional Neural Networks," September 2016.
- Keunwoo Choi, Gyorgy Fazekas, and MarkS, "Convolutional Recurrent Neural Networks for Music Classification," December 2016.

*For the video, please visit:

* https://youtu.be/1FaVVqhY9e8