Convolutional Neural Network Achieves Human-level Accuracy in Music Genre Classification

Mingwen Dong (mingwen.dong@rutgers.edu)
Psychology, Rutgers University
Piscataway, New Jersey 08854 USA

Abstract

Music genre classification is one example of content-based analysis of music signals. Traditionally, human-engineered features were used to automatize this task and 61% accuracy has been achieved in the 10-genre classification. Here, we propose a method that achieves human-level accuracy (70%) in the same classification task. The method is inspired by knowledge of human perception study in music genre classification and the neurophysiology of the auditory system. It works by training a simple convolutional neural network (CNN) to classify short segments of music waveforms. During prediction, the genre of an unknown music is determined as the majority vote of all classified segments from a music waveform. The filters learned in the CNN qualitatively resemble the spectro-temporal receptive fields (STRF) in the auditory system and potentially provide insights about how human auditory system classifies music genre.

Keywords: music genre classification; STRF, CNN.

Introductions

With the rapid development of digital technology, the amount of digital music content increases dramatically everyday. To give better music recommendations for the users, it's essential to have an algorithm that could automatically characterize the music. This process is called Musical Information Retrieval (MIR) and one specific example is music genre classification.

However, music genre classification is a very difficult problem because the boundaries between different genres could be fuzzy in nature. For example, testing with a 10-way forced choices task, college students could achieve 70% classification accuracy after hearing 3-seconds segment of the music and the accuracy does not improve with longer segment. (Tzanetakis & Cook, 2002). Also, the number of labeled data often is much smaller than the dimension of the data. For example, GTZAN dataset [1] used in the current work contains only 1000 audio tracks, but each audio track is 30s long with a sampling rate 22,050 Hz.

Traditionally, using human-engineered features like MFCC (Mel-frequency cepstral coefficients), texture, beat and so on, 61% accuracy has been achieved in the 10-genre classification task (Tzanetakis & Cook, 2002). More recently, using PCA-whitened spectrogram as input, convolutional deep belief network has achieved 70% accuracy in a 5-genre classification task. These results are reasonable but still not as good as humans, suggesting there's still space to improve.

Psychophysics and physiology study show that human auditory system works in a hierarchical way (Schnupp, Nelken, & King, 2011). First, the ear decomposes the continuous sound waveform into different frequencies with higher precision on low frequencies. Then, neurons from lower to higher auditory structures gradually extract more complex acoustic features with more complex spectro-temporal receptive field (STRF) (Theunissen & Elie, 2014). The features used by human auditory system for music genre classification may rely on STRFs but at different time scales. By having the spectrogram as input and the corresponding genre as label, CNN will learn filters that extract features in the frequency and time domain. These learned filters can be seen as STRFs for music classification. Because music signal often is high-dimension in the time domain, having a CNN that fits the complete spectrogram of the music signal is not feasible. To solve this problem, we split the spectrogram of the music signal into consecutive 3-second segments, make predictions for each segment, and finally combine the predictions from all segments using the majority vote. The main rational for this method is that humans' classification accuracy plateaus at 3 seconds and good results were obtained using 3-second segments to train convolutional deep belief network (Tzanetakis & Cook, 2002) (Lee, Pham, Largman, & Ng, 2009). It also intuitively makes sense because different parts of the same music probably should belong to the same genre.

To further reduce the dimension on the spectrogram, we used mel-spectrogram as the input to the CNN. Mel-spectrogram approximates how human auditory system works and can be seen as the spectrogram smoothed in the frequency domain, with high precision in the low frequencies and low precision in the high frequencies (O'shaughnessy, 1987) (Picone, 1993).

Data Processing & Models

Data pre-processing

Each music signal is first converted from waveform into mel-spectrogram z_i using Librosa library with 23ms time window and 50% overlap (Figure 1). Then, the mel-spectrogram is log transformed to bring values at different mel-scale to the same range ($f(z_i) = ln(z_i + 1)$). Because mel-spectrogram is a biological-inspired representation (Picone, 1993), it has a simpler interpretation than the PCA-whitening method used in (Lee et al., 2009).

[1] Available at: `http://marsyasweb.appspot.com/download/data_sets/`

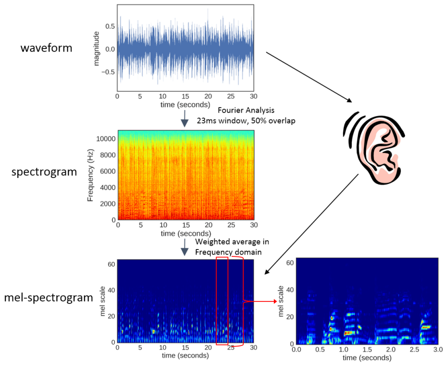


Figure 1: Convert waveform into mel-spectrogram and an example 3-second segment. Mel-spectrogram mimics how human ear works, with high precision in low frequency band and low precision in high frequency band. Note, the mel-spectrogram shown in the figures is already log transformed.

Network Architecture

1. Input layer: 64 * 256 neurons, corresponds to 64 mel scales and 256 time windows(23ms, 50% overlap).
2. Convolution layer: 64 different 3 * 3 filters with a stride of 1.
3. Max pooling layer: 2 * 4.
4. Convolution layer: 64 different 3 * 5 filters with a stride of 1.
5. Max pooling layer: 2 * 4.
6. Fully connected layer: 32 neurons that are fully connected to the neurons in the previous layer.
7. Output layer: 10 neurons that are fully connected to neurons in the previous layer.

For 2D layers/filters, the first dimension corresponds to the mel-scale and the second dimension corresponds to the time. All hidden layers use RELU activation functions, the output layer use softmax function, and the loss is calculated using cross-entropy function. Dropout and L2 regularization were used to prevent extreme weights. The model is implemented using Keras (2.0.1) with tensorflow as backend and trained on a single GTX-1070 using stochastic gradient descent.

Training & Prediction

1000 music tracks (converted into mel-spectrogram) are split into 50% training, 20% validation, and 30% testing. The training procedure is as following:

1. Select a subset of tracks from the training set.
2. Randomly sample a starting point and take the 3-second continuous segments from all selected tracks.
3. Calculate the gradients using back-propagation algorithm using the segments as input and the labels of the original music as target genres.
4. Update the weights using the gradients.

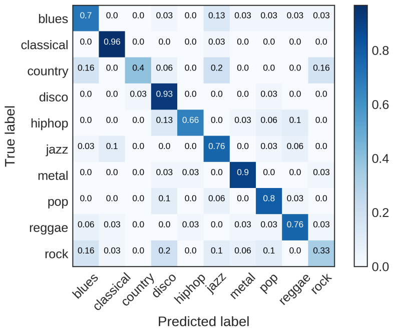


Figure 2: Confusion matrix of the CNN classification on testing set.

5. Repeat the procedure until classification accuracy on the cross-validation data set doesn't improve anymore.

During testing, all music (mel-spectrogram) are split into consecutive 3-second segments with 50% overlap. Then, for each segment, the trained neural network predicts the probabilities of each genre. The predicted genre for each music is the genre with most votes.

Calculate the filters learned by the CNN

After training, all musics are split into 3-second segments with 10% overlap. All the segments are then fed into the trained CNN and intermediate outputs are calculated and stored. Then, we estimated the learned filters using the following method:

1. Identify the range of input neurons (specific section of the input mel-spectrogram) that could activate the target neuron at a specific layer. E.g., $c_{i,j}^{(l)}$ indicates the neuron at location (i, j) from the l^{th} layer.
2. Perform Lasso regression with the specific section of the mel-spectrogram (reshaped as a vector) as the regressors and the corresponding activations of the neuron $c_{i,j}^{(l)}$ as the target values.
3. The fitted Lasso coefficients were reshaped to estimate the learned filters.

Results

The current method achieves human-level (70%) accuracy in the 10-genre classification task (Figure 2). It's 10% higher than in (Tzanetakis & Cook, 2002) and classifies 5 more different genres than (Lee et al., 2009) with similar accuracy.

Classification accuracies varies by different genres.

Confusion matrix (Figure 2) shows that the classification accuracy varies a lot across different genres. Especially, the accuracies for country and rock genre are not only lower than the current average but also lower than those from (Tzanetakis & Cook, 2002). Because some important human-engineered features used in (Tzanetakis & Cook, 2002) are the long-term

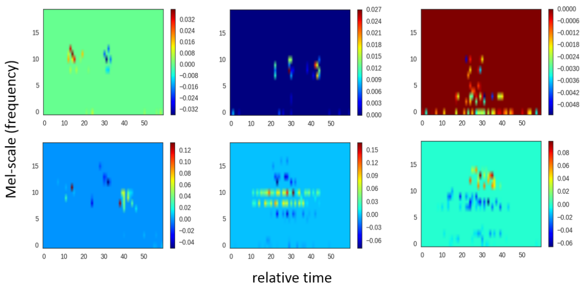


Figure 3: Filters learned by the CNN are similar to the STRF from physiological experiments. Mel scale corresponds to frequency and relative time corresponds to latency in figure 4. Note that 1 unit of time is 60ms, which is different from figure 4.

feature like beat and rhythm, this suggests country and rock music may have characteristic features (e.g., beat) that require longer time (> 3 seconds) to capture and 3s segments used in our CNN are not long enough. One future direction is to explore how to use CNN to extract long-term features for classification and one possibility is to use another down-sampled mel-spectrogram of the whole audio as input.

CNN learns filters like spectro-temporal receptive field.

Figure 2 shows some filters learned by the CNN's 2nd max pooling layer and they are qualitatively similar to the STRF obtained from physiological experiments (Figure 4), even though at different time scales. To visualize how these filters help classify the audios, we feed all the 3s segments from the testing set into the CNN and calculated the activations of the last hidden layer. After this non-linear transformation, most testing data points become linearly separable (Figure 5). In contrast, the testing data points are much less separable when raw mel-spectrogram is used.

These results together show that human auditory system may use filters like those learned in the CNN to classify music genre. The STRF-like filters transform the original mel-spectrogram into a representation where the data is linearly separable. But to test this hypothesis, physiological experiments are needed.

Discussion

By combining the knowledge from human psychophysics study and neurophysiology, we used a simple CNN which successfully classified the audio waveforms into different genres with human-level accuracy. It may not be the methods that give the highest classification accuracy ((Van Den Oord et al., 2016); (Stokowiec, 2016)), but it is simple and potentially provides insights about how humans perform music genre classification. The same technique may be used to solve problems that share similar characteristics, for example, music tagging and artist identification using raw audio waveform.

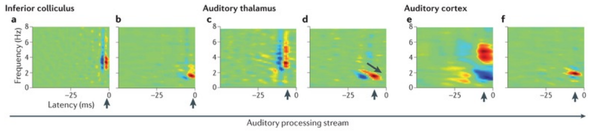
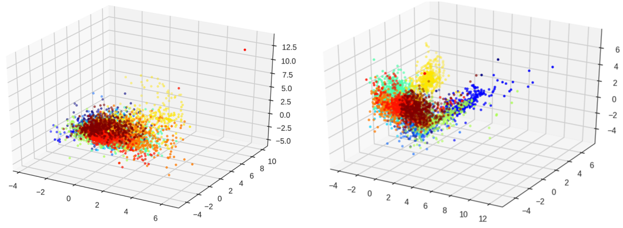


Figure 4: STRF obtained from physiological experiments. From left to right are the STRFs obtained from lower to higher auditory structures. Adapted from (Theunissen & Elie, 2014) with permission. Note that 1 unit of time is 1ms.



a) LDA representation from raw mel-spectrogram

b) LDA representation from last hidden layer

Figure 5: Comparison between the separability of the raw representation and last layer representation of the CNN of the testing data. The axes are the first three components when data is projected onto the directions obtained from linear discriminant analysis (LDA). using training data.

References

Lee, H., Pham, P., Largman, Y., & Ng, A. Y. (2009). Unsupervised feature learning for audio classification using convolutional deep belief networks. In *Advances in neural information processing systems* (pp. 1096–1104).

O'shaughnessy, D. (1987). *Speech communication: human and machine*. Universities press.

Picone, J. W. (1993). Signal modeling techniques in speech recognition. *Proceedings of the IEEE*, *81*(9), 1215–1247.

Schnupp, J., Nelken, I., & King, A. (2011). *Auditory neuroscience: Making sense of sound*. MIT press.

Stokowiec, W. (2016). A comparative study on music genre classification algorithms. In *Machine intelligence and big data in industry* (pp. 123–132). Springer.

Theunissen, F. E., & Elie, J. E. (2014). Neural processing of natural sounds. *Nature Reviews Neuroscience*, *15*(6), 355–366.

Tzanetakis, G., & Cook, P. (2002). Musical genre classification of audio signals. *IEEE Transactions on speech and audio processing*, *10*(5), 293–302.

Van Den Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., ... Kavukcuoglu, K. (2016). Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*.