Music Genre Classification

George Herbert
Department of Computer Science
University of Bristol
Bristol, United Kingdom
cj19328@bristol.ac.uk

***Abstract*—**
***Index Terms*—music information retrieval, convolutional neural networks**

I. Introduction

Music information retrieval (MIR) is the interdisciplinary science of retrieving information from music. Genre classification—classifying a sample of music into one or more genres—is a fundamental problem in MIR.

Schindler et al. [1] investigated performance differences of different convolutional neural network (CNN) architectures on the task of genre classification.

II. Related Work

III. Dataset

The GTZAN dataset [2] contains 1000 WAV audio tracks, each 30 seconds in length. There are 100 tracks for each of the 10 genres in the dataset: blues, classical, country, disco, hip-hop, jazz, metal, pop, reggae and rock.

The CNNs were not trained on the raw tracks. Each audio track was divided into chunks, with log-mel spectograms produced for randomly selected chunks. To produce a training and validation set, a stratified split was deemed suitable to prevent imbalance. Spectograms for 75 of the 100 WAV audio tracks for each genre were randomly selected to make up the training set, with the spectograms for the other 25 audio tracks for each genre making up the validation set.

IV. CNN Architecture

I recreated the shallow CNN architecture outlined by Schindler et al. Log-mel spectrograms of shape 80×80 are provided as input to the network.

Since the dimensions of the spectograms correspond to time and frequency, Schindler et al. implemented a parallel architecture. The left pipeline aims to capture frequency relations. It first contains a convolutional layer (with padding) with 16 kernels of shape 10×23 to produce 16 square feature maps of shape 80×80. These are then downsampled using a 1×20 max pooling layer to produce 16 vertical rectangular feature maps of shape 80×4. The kernel and max pooling shapes were specifically selected to capture spectral characteristics. Conversely, the right pipeline aims to capture temporal relations. It too initially contains a convolutional layer (with padding) with 16 kernels, but of approximately square shape 21×20 to produce 16 square feature maps of shape 80×80. These are then downsampled using a 20×1 max pooling layer to produce 16 horizontal rectangular feature maps of shape 4×80, specifically to capture temporal changes in intensity.

The 16 feature maps from each piepline are flattened and concatenated to a shape of 1×10240, which serves as input to a 200 neuron fully connected layer—10% dropout is utilised at this layer to prevent overfitting. These final 200 neurons are then mapped to 10 output neurons, which represent the probabilities of each of the ten genres for a given input.

With the exception of the final layer, each convolutional and fully connected layer is passed through the Leaky ReLU activation function. The final layer uses the softmax activation function.

V. Implementation Details

VI. Replicating Quantitative Results

VII. Training Curves

VIII. Qualitative Results

IX. Improvements

X. Conclusion and Future Work

References

[1] A. Schindler, T. Lidy, and A. Rauber, "Comparing shallow versus deep neural network architectures for automatic music genre classification," Nov. 2016.

[2] G. Tzanetakis and P. Cook, "Musical genre classification of audio signals," *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 5, pp. 293–302, 2002.