Music Genre Classification with Neural Networks

Neural networks:

are the core of this project. The history of neural networks in artificial intelligence can be traced back to the 1940s, but their performance only became significant in the past twenty years. Among many models of neural networks, convolutional neural networks (CNN) and recurrent neural networks (RNN) which is used in this project. Any kind of neural network consists of neurons (nodes) and edges. During the learning stage, neurons of the first layer take the inputs, put them through an activation function, and output the results into the next layer.

-Convolutional Neural Networks:

Diagram

Description automatically generatedTraditional multilayer perceptron models are fully connected and work fairly well with image recognition tasks. However, they do not scale well with high-resolution images due to the restriction of computing power. In addition, multilayer perceptrons do not take into account the spatial structure of visual patterns, and thus distant pixels can have the same impact in recognition of an area as a closer pixel. CNN’s overcome this problem by implementing 3D layers that are only connected to a small region of the previous one and filters in the same layer share the weights and biases. Therefore, the number of parameters in one convolutional layer is given by (n2\*x)\*2, where n is the side length for one small region and x is the number of filters in this layer.

-Recurrent Neural Networks:

Diagram

Description automatically generatedRecurrent Neural Networks are often used for sequential data analysis (e.g. text prediction and speech recognition). Units within an RNN are connected along a time sequence, where each unit represents a new time step. There is one input and one output for each time step. A simplified graph for a general RNN model is presented below. The sequential structure is for visualization purposes. In practice, the model is usually circular. There is only one copy of the state, which keeps updating itself with a combination of the previous input and hidden state.

-Aggregation:

After finishing training and comparing different architectures, we saved the best neural net model with its structure and weights. The holdout group (not used during training or testing) from the same dataset was then fed into the model to predict labels. We aggregated the predictions for every track, using the majority rule. In the end, we compare the track-based predictions to the ground truth to compute the final accuracy Dh ll accuracy part

Pre-Processing:

It will be using the Fast-Fourier Transform (FFT), Mel-Spectrograms, And Mel-frequency cepstral coefficients (MFCCs) using librosa

-Librosa:

It is a Python module to generally analyze audio signals but geared more towards music. It includes the nuts and bolts to build a MIR(Music information retrieval) system. It has been very well documented along with a lot of examples and tutorials it’s a library that we will be using in this project.

We also might be using many other libraries as we proceed in the project.

Dataset:

As for the dataset, we will be using the famous [GITZAN](http://marsyasweb.appspot.com/download/data_sets/) dataset for our case study. This dataset was used for the well-known paper in genre classification “ [Musical genre classification of audio signals](https://ieeexplore.ieee.org/document/1021072)“ by G. Tzanetakis and P. Cook in IEEE Transactions on Audio and Speech Processing 2002. The dataset consists of 1000 audio tracks each 30 seconds long. It contains 10 genres namely, **blues, classical, country, disco, hip-hop, jazz, reggae, rock, metal, and pop.**Each genre consists of 100 sound clips.

Conclusion:

After testing several choices of datasets, pre-processing methods, neural network structures, and other factors, we found the optimal combination to be a convolutional neural network using Mel-spectrograms of three-second samples of audio. A bigger dataset reduces the overfitting problem but has very little impact on validation accuracy. Our final best validation accuracy turned out to be 59%. Although it was inferior to the state-of-art accuracy for music genre classification, it outperformed other attempts to solve this challenge with convolutional neural networks. We also discovered that the classification accuracy was highly genre-dependent, which could have impeded the overall performance. It also showed genre selection’s great impact on the classification difficulty.