Final Project Report
CSE-6363 Machine Learning
Fall-2020

Chandra Shekhar Kasturi
UTA-ID: 1001825454
The University of Texas at Arlington
chandrashekhar.kasturi@mavs.uta.edu

Unnathi Reddy Nalla
UTA-ID: 1001828087
The University of Texas at Arlington
unnathireddy.nalla@mavs.uta.edu

Srithisha Katam
UTA-ID: 1001829488
The University of Texas at Arlington
srithisha.katam@mavs.uta.edu

Abstract

Music plays a very important role in people's lives. Music bring like-minded people together and is the glue that holds communities together. Communities can be recognized by the type of songs that they compose, or even listen to. The purpose of our project and research is to find a better machine learning algorithm than the pre-existing models that predicts the genre of songs. In this project, we built multiple classification models and trained them over the GTZAN dataset. We have compared the performances of all these models and logged their results in terms of prediction accuracies. Few of these models are trained on the mel-spectrograms of the songs along with their audio features, and few others are trained solely on the spectrograms of the songs. It is found that one of the models, a convolutional neural network(CNN), which was given just the spectrograms as the dataset, has given the highest accuracy amongst all other models.

1 Introduction

Genre classification is an important task with many real world applications.As the quantity of music being released on a daily basis continues to sky-rocket, especially on internet platforms such as Soundcloud and Spotify.Being able to instantly classify songs in any given playlist or library by genre is an important functionality for any music streaming/purchasing service, and the capacity for statistical analysis that correct and complete labeling of music and audio provides is essentially limitless.We experimented with support vector machine(SVM), k-nearest neighbors(knn), principal component analysis(PCA), logistic regression and convolutional neural network(CNN).We implemented these classification algorithms admitting two different types of input(the raw data and the transformed mel-spectrograms),then predicted an output genre out of 10 common music genres.The advanced convolution neural network stood out the best with the input as the transformed mel-spectrograms.Therefore we decided to focus our efforts on implementing a high-accuracy CNN, with other models used as a baseline.

1.1 Dataset and Features

The dataset we used is a public GTZAN dataset.The data set comes with 1000 30-second audio notes, labelled with one of the 10 possible audio genres as .au files. From this dataset we sampled a 2 second window at random locations.Since the data was sampled at 22050HZ, it leaves us with 44100 features as the raw audio input.After preprocessing the input shape is (8000,44100),we also used unaugumented data of 100 samples for our cross-validation and test-sets.The transformed melspectrograms data set is generated by applying short-time Fourier transforms across sliding windows of audio, most commonly around 20ms wide.Then we take the discrete cosine transform of the result (common in signal processing) in order to get our final output

mel-spectogram.Doing this, we experienced significant performance increases across all models. Mel-spectograms are a commonly used method of featurizing audio because they closely represent how humans perceive audio

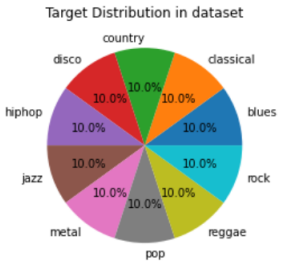

Figure 1: Target distribution in dataset

1.2 Principal Component Analysis

Other than pre-processing our dataset we have used Principal Component Analysis to reduce dimension for our models(Knn ,Logistic regression, SVM). PCA is a linear technique for dimensionality reduction performing a linear mapping of data to a lower-dimensional space by maximizing the variance of the data in the lower-dimension. In order to preserve as much variance as possible with m examples x(i), unit length u, PCA maximizes

$$\frac{1}{m}\sum_{i=1}^{m}(x^{(i)^T}u)^2 = \frac{1}{m}\sum_{i=1}^{m}x^{(i)}x^{(i)^T}.$$

2 Methods

2.1 K-Nearest Neighbors

K-Nearest neighbors is an instance-based learning or lazy-learning algorithm used for classification as well as regression. Given a set of data (x1; x2; x3:::) where each observation is an d-dimensional vector in which it determines the k-closest samples in the feature space.In this project after pre-rpocessing and reducing the dimensionality of the data using PCA with 15 components ,we used k=10 for the classification and found out the accuracy to be highest with distance metric being euclidean distance. Finally we return the label with most nearest to 10 other neighbors when calculated by distance.

2.2 Support Vector Machine

Similar to what we have done in knn by reducing the dimensionality of our data using PCA, we have trained an SVM classifier.Since our data is not linearly separable we used a gaussian kernel , corresponds to an infinite dimensional feature space is related to Euclidean distance. This function as a whole is often minimized using sequential minimization optimization.

$$K(x^{(j)}, x^{(i)}) = \exp\left(-\frac{||x^{(j)} - x^{(i)}||_2^2}{2\sigma^2}\right).$$

2.3 Linear Regressor (Feed Forward Neural Network)

We used a fully connected neural network as well, with ReLU activation and 6 layers, with cross-entropy loss. As the input to our model was 1D, when using mel-spectrograms, we flattened the data. Our model is fully connected, which means each node is connected to every other node in the next layer. At each layer, we applied a ReLU activation function to the output of each node, then we constructed a 10 class softmax function by running each output, to optimize we used cross entropy loss. We used a L2 regularization as a penalty term to the loss function as well.

2.4 Convolution Neural Network

We mainly focused on implementing the convolution neural network using 3 Convolution layers each with its own maxpool and regularization with relu as activation function ,softmax output and cross-entropy loss.This approach involves convolution windows that scan over the input data and output the sum of the elements within the window.An example of how our model will be represented can be seen below:

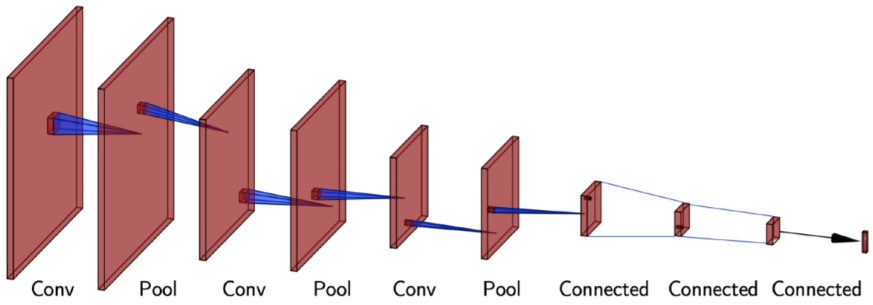

Figure 2: CNN Architecture

3 Results

The results in our project are mainly compared using accuracy metric as a quantitative measure , which measures the probability with which the correct class-label(genre) has been predicted. We have experimented each model with pre-processing data as well without pre-processing the data. The best way to visualise our result is by viewing the confusion matrices as they not only offer a way to visualize our data but also provide additional metrics such as precision and recall to quantitatively measure the performance of our models.

We chose to use Adam optimization for a few reasons. Working with audio time-series data over 2 dimensions cause sparse gradient problems, similar to those often encountered in natural language or computer vision problems. We also felt our data was somewhat noisy and messy.With Adam our models trained more quickly and did not plateau as early.Table 1 for model accuracy both without data processing – i.e. with raw audio as input – and with data processing – i.e. using the scaled mel-spectrograms as input.

The identifiable metric in the below table can be seen when we shifted our data from raw-data to mel-spectrogram.We observed an significant increase in the accuracy of our models.Converting the data to mel-spectrogram is a long process with the number of training samples we have .So we have downloaded an already converted GTZAN data to mel-spectrogram and included in the .zip file of our submission.

Table 1: Accuracy of predictions by model used.

	With data processing			Without data processing		
	Train	CV	Test	Train	CV	Test
Support Vector Machine	.97	.60	.60	.75	.32	.28
K-Nearest Neighbors	1.00	.52	.54	1.00	.21	.21
Feed-forward Neural Network	.96	.55	.54	.64	.26	.25
Convolution Neural Network	.95	.84	**.82**	.85	.59	.53

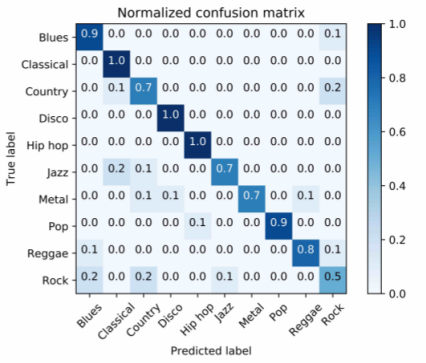

Figure 3: Confusion matrix of CNN

In the implementation of CNN we have faced the issue of over-fitting but tried to minimize it using the L2 Regularization and using the dropout layers.The confusion matrix of the cnn depicts that it misclassified some instances of Blues,Classical and Reggae as Rock. So the classifier had much difficulty in predicting the class-

label(genre) Rock the most as it correctly classified only half of its instances correctly.

4 Conclusion

We have implemented CNN,KNN,SVM and Linear regression using the raw and mel-spectrograms. The longest time it took was to train the CNN,the hardest and the most accurate classifier in our project, however KNN,SVM and linear regressor were baselined due to their expected accuracies.In the future we like to experiment with different classifier such as LSTM,GAN considering different metrics such as classifying based on the artist and the year of release.We would like to experiment the multi-label classification of any song where a song is expected to be classified into more then a single genre.

5 Collaborations and Contributions

The entire project has been a combined byproduct of all the team-members starting with idea of the project with the respective selection on model implementations and preparing the report.Each and every task such as selecting the data set,determining the appropriate architecture for the convolution neural network it has been a collective effort.

6 References

[1] Y. Panagakis, C. Kotropoulos, and G. R. Arce. Music genre classification via sparse representations of auditory temporal modulations. In 2009 17th European Signal Processing Conference, pages 1–5, Aug 2009.

[2] Mingwen Dong. Convolutional neural network achieves human-level accuracy in music genre classification. CoRR, abs/1802.09697, 2018.

[3] G. Tzanetakis and P. Cook. Musical genre classification of audio signals. IEEE Transactions on Speech and Audio Processing, 10(5):293–302, July 2002.

[4] Changsheng Xu, N. C. Maddage, Xi Shao, Fang Cao, and Qi Tian. Musical genre classification using support vector machines. In 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03)., volume 5, pages V–429, April 2003.