Guess The Genre

Atalay Gürel [1]
Muhammed Fidan [1]

Abstract

We could not even imagine a life without music. Throughout history wherever there were people, there was music too. Empires fell and rose, revolutions were made and humankind has come to these days.The music was with them for all these moments. Humankind took shape with music but also the music took shape with humankind. Music has evolved for generations and generations and found its form in many ways. Today we use the term "genre" for different categories of music. The goal of this paper is to give machines a chance to predict music genres given input features from music tracks. To do that, we apply various techniques based on machine learning on the dataset called FMA which consists of 161 subgenres among 106,574 tracks.

1. Introduction

As a part of being human, we all have different tastes for any kind of hobby. There is even a popular expression for this: "Tastes and colors are indisputable". This also includes our music tastes; the difference may come from the person's upbringing, reliefs as well as the environment and era in which the person grew up.

What we aim in this project is to successfully make a Music Genre Recognition when we have input as an audio file by utilizing Machine Learning algorithms. Popular music streaming services and platforms are using these algorithms to understand the preferences of their audience and provide them better recommendations so that the audience will enjoy their music more and spend time on these mediums more.

In this project we tried to implement and discuss baseline classification models to solve the problem of music genre classification. These models are:

- K-Nearest Neighbor Classifier
- Logistic Regression
- Support Vector Machines
- Deep Neural Network
- Random Forest
- Multi-Layer Perceptron
- Convolutional Neural Network

To represent the audio tracks we planned to use Mel Frequency Cepstral Coefficients(MFCC) and Spectral Contrast features to begin with, which have been shown to be effective in the task of predicting genres.

2. Related Work

For the music genre classification one of the most common datasets are FMA: A Dataset For Music Analysis (Defferrard & Bresson, 2017), and GTZAN (Tzanetakis & Cook, July 2002.). However FMA is the one more up to date and fresh dataset.

Kim et al (Jaehun Kim & Liem, 2018) used the same FMA Medium dataset but took a different approach to the genre recognition problem. This work claimed that genre labels are noisy, subjective, and not clearly separable from one another. Genre labels may be inter-correlated, and a generally accepted taxonomy of musical genres does not exist. Given this claim, the authors asserted that artist labelings do not suffer from the same faults that genre labelings do. Furthermore, they contended that since artists tend to stay within one or several related genres, musical characteristics of the artists may be indicative of certain musical genres

Automatic classification of audio has also a long history originating from speech recognition. Mel-frequency cepstral coefficients (Davis & Mermelstein, Aug. 1980), are a set of perceptually motivated features that have been widely used in speech recognition. They provide a compact representation of the spectral envelope, such that most of the signal energy is concentrated in the first coefficients.

[1]Department of Computer Engineering, University of Hacettepe, Ankara, Turkey. Correspondence to: Muhammed Fidan <muhammed.fidan77@gmail.com>, Atalay Gürel <atagrl98@gmail.com>.

BBM406 Fundamentals of Machine Learning, Spring 2021. *Copyright 2021 by the author(s), .*

In the academic study titled "A comparative study on content-based music genre classification", DWCH(Daubechies Wavelet Coefficients Histogram) (T. Li & Li, 2003) which is a new method of extracting attributes on music genre classification, is mentioned. This method is based on feature extraction by calculating the Daubechies Wavelength Coefficients of the signals. Using the Support Vector Machine, K-NN, Gaussian Mixing Model and Linear Discriminant Analysis algorithms, the success percentage of this new method compared to the previous methods was compared.

Michael I. Mandel and Daniel P.W. Ellis (Mandel & Ellis, 2005) used Support Vector Machines to classify music in his work. He thought that SVM might be more successful than Knn. These advantages become evident when comparing four combinations of classifiers and features. SVM using the Mahalanobis distance performed the best, achieving a classification accuracy of 69%. When the songs were randomly distributed between cross validation sets, an SVM using the KL divergence between single Gaussians was able to classify 84% of songs correctly.

The first work that uses time decomposition using regular classifiers applied to complete feature vectors was proposed by Costa, Valle-Jr and Koerich. (C.H.L. Costa & Koerich, 2004) This work presents experiments based on ensemble of classifiers that uses three time segments of the music audio signal, and where the final decision is given by the majority vote rule. They employ a MLP neural net and the k-NN classifiers. Experiments were conducted on a database of 414 music pieces of 2 genres. However, final results regarding the quality of the method for the classification task were inconclusive.

3. Dataset Analyze

The FMA dataset, a dump of the Free Music Archieve, includes 106,574 tracks with 161 sub-genres. In this project, we used 8000 of the tracks with 8 genres sampled considering their metadata and popularity for computational efficiency. Our data includes clips of 30s and a balanced distribution among genres.

In FMA dataset, the features are generated using librosa, and stored as statics, including kurtosis, max, min, mean, median, std and skew, for each feature. For pitch feature, chroma representations are a preferred way to encode harmony, and suppressing perturbations in octave height, loudness, or timbre. Chroma features could be extracted in different ways, e.g. by convolving a fixed size window Short Time Fourier Transform (chroma stft), or a variable sized window, constant-Q transform (chroma cqt), over the audio signal to extract the time-varying frequency spectrum.

We split our data preserving the percentage of tracks per genre as a reflection of population (stratified sampling) into training and test by 80/20%. Thus, our training data turned into a matrix of 6400 rows and 140 columns.

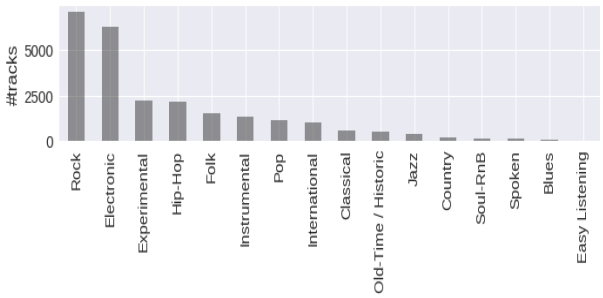

Figure 1. Top Genres

4. The Approach

4.1. NEAREST NEIGHBORS CLASSIFIER

KNN is a non parametric algorithm (meaning, it does not make any underlying assumptions about the distribution of data)belonging to supervised learning community. KNN algorithm can also be used for regression problems.The only difference will be using averages of nearest neighbors rather than voting from nearest neighbors. Although kNN is an easy implemented algorithm and performs well in a large number of classification problems, it suffers from the curse of dimensionality. Our model has a dimension space of 140 features which makes kNN a bit vulnerable.

4.2. LOGISTIC REGRESSION

Logistic regression is a statistical machine learning algorithm that classifies the data by considering outcome variables on extreme ends and tries makes a logarithmic line that distinguishes between them. It is a go-to method for binary classification problems.

We used multinomial logistic regression. It uses for three or more categories without ordering. Linear regression uses mean squared error as its cost function. But in multiclass classification problems, it would be more logical to use the maximum likelihood cost function. Maximum likelihood estimation (MLE) is a method of estimating the parameters of a probability distribution by maximizing a likelihood function, so that under the assumed statistical model the observed data is most probable.

4.3. SUPPORT VECTOR MACHINE

A Support Vector Machine (SVM) is a discriminative classifier formally defined by a separating hyperplane. In other words, given labeled training data (supervised learning), the

algorithm outputs an optimal hyperplane which categorizes new examples. In two dimentional space this hyperplane is a line dividing a plane in two parts where in each class lay in either side.

We tried to preprocess the dataset before applying Linear SVC using standard scaler to represent standard normally distributed data. Beside from Logistic Regression, SVM uses one-against-one approach for multi-class classification. This method is consistent, which is not true for one-vs-rest classification.

4.4. DEEP NEURAL NETWORK

Deep Neural Networks (DNNs) are typically Feed Forward Networks in which data flows from the input layer to the output layer without going backward . DNN is an artificial neural network (ANN) with multiple layers between the input and output layers. There are different types of neural networks but they always consist of the same components: neurons, synapses, weights, biases, and functions. These components functioning similar to the human brains and can be trained like any other ML algorithm. DNNs can model complex non-linear relationships. DNN architectures generate compositional models where the object is expressed as a layered composition of primitives. The extra layers enable composition of features from lower layers, potentially modeling complex data with fewer units than a similarly performing shallow network. For instance, it was proved that sparse multivariate polynomials are exponentially easier to approximate with DNNs than with shallow networks. A DNN model comprises three layers: Input, hidden, and output. The input and output layers are, in general, single layers, whereas the hidden layer may comprise two or more layers. Data features are fed to the input layer, and prediction values are derived from the output layer after processing them in the hidden layers

4.5. RANDOM FOREST

The Random Forest algorithm is an extension of the bagging method as it utilizes both bagging and feature randomness to create an uncorrelated forest of decision trees.Feature randomness, also known as feature bagging or the ”random subspace method”, generates a random subset of features that provides low correlation between decision trees. This is the key difference between decision trees and random forests. Decision trees consider all possible feature segments, whereas random forests select only a subset of those features. Random forest has mainly three hyper-parameters. Node size, The number of trees and the number of features sampled. Overfitting is common in decision trees. However, when there's a robust number of decision trees in a random forest, the classifier won't overfit the model since the averaging of uncorrelated trees lowers the overall variance and prediction error. And also Random Forest provides flexibility, it is easy to determine feature importance.

4.6. MULTI-LAYER PERCEPTRON

Multi layer perceptron (MLP) is a supplement of feed forward neural network. It consists of three types of layers—the input layer, output layer and hidden layer. The input layer receives the input signal to be processed. The required task such as prediction and classification is performed by the output layer. An arbitrary number of hidden layers that are placed in between the input and output layer are the true computational engine of the MLP. Similar to a feed forward network in a MLP the data flows in the forward direction from input to output layer. The neurons in the MLP are trained with the back propagation learning algorithm. MLPs are designed to approximate any continuous function and can solve problems which are not linearly separable. The major use cases of MLP are pattern classification, recognition, prediction and approximation.

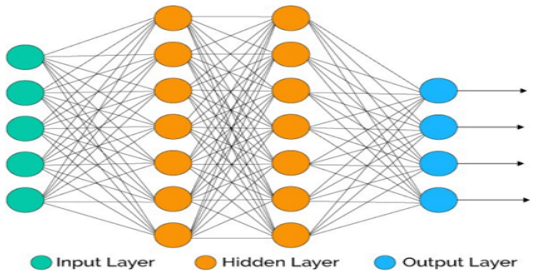

Figure 2. MLP model example

Every hidden layer has inputs and outputs. Inputs and outputs have their own weights that pass through the activation function and its derivative computation. For the hidden layer, the input values to the layer are multiplied by the synaptic weights of the layer. And this result is given as input to other hidden layers. This cycle is repeated until it reaches the output layer. Softmax is mostly used for multiclass classification in the output layer and the node value with the highest ratio becomes the predicted class.

4.7. CONVOLUTIONAL NEURAL NETWORK

A Convolutional Neural Network (ConvNet/CNN) is a Deep Learning algorithm that can take an input image, assign importance (learnable weights and biases) to various aspects in the image, and distinguish one from the other. The preprocessing required in a ConvNet is much lower compared to other classification algorithms. While filters are designed by hand in primitive methods, with sufficient training ConvNets are capable of learning these features. Convolutional Neural Networks take advantage of the input being composed of

images and constrain the architecture in a more logical way. ConvNet's layers have neurons arranged in 3 dimensions: width, height, depth.

5. Experiment and Conclusion

In our models, we tried to approach our results with six classification method. These are; Nearest Neighbors(kNN), Logistic Regression and Support Vector Machine (SVM), Deep Neural Network(DNN), Random Forest(RF) and Multilayer Perceptron(MLP) . We split our dataset into train and test subsets. In our train set there are 6400 clips meanwhile test set included 800 clips.

In the first observations we made with 8 music genres, the success rate was low. We examined the confusion matrices in each model and looked at which types were found to be true and false. In general, models made mistakes in folk and pop music. While the models predicted music genres such as Hip-hop and rock in the range of 60%-70%, the correct prediction of Folk genre was around 30%.

In the classifications, the rate of predicting folk music was found to be low. And the features of this genre caused confusion with other music genres during model training. That's why we decided to remove the Folk genre from our dataset. We observed an increase in accuracy of nearly 10% in models that we trained with the remaining 7 music genres.

The following tables show the accuracy performance obtained with all the features and Mel Frequency Cepstral Coefficients (MFCC) alone.

Table 1. Methods accuracy with 8 genres

MODELS	TRAIN ACC.(%)	TEST ACC(%)
KNN	57.8	36.1
LOGISTIC REG.	56.0	42.1
SVM	75.8	46.3
DNN	78.0	43.4
RANDOM FOR.	99.9	43.2
MLP	96.3	46.8

Table 2. Methods accuracy after removing Folk genre

MODELS	TRAIN ACC.(%)	TEST ACC(%)
KNN	58.6	46.8
LOGISTIC REG.	57.3	48.7
SVM	76.7	55.2
DNN	89.2	49.3
RANDOM FOR.	99.9	51.4
MLP	79.4	54.3

Our kNN and Logistic Regression models are outperformed by SVM regardless of the chosen features. The reason behind of the issue is that Nearest Neighbors algorithm treats vectors as an input which makes this method work uneffi-ciently in complex dimensionals while also logistic regression does not try to find the best margin (distance between the line and the support vectors) that separates the classes and this reduces the risk of error on the data. Instead it can have different decision boundaries with different weights that are near the optimal point.

Support vector machine with linear kernel performing a classification with a linear decision boundary. So when facing an input set of high dimensions, we can reach to a conclusion that we can improve our decision boundary by using a radial basis kernel in order to perform a well non-linear decision boundary.

As an addition to the MFCC feature set of 140 dimensions used in Table 1, Despite the increase in dimension, all of our methods except for kNN, we achieved better results in both train and test accuracy. Therefore, with better feature and model selection, improved accuracy results can be achieved.

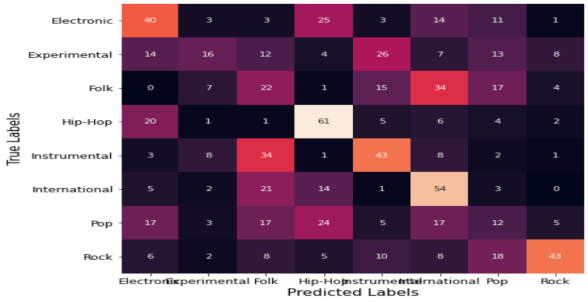

Figure 3. kNN Confusion Matrix

We run the model with different k values to find the best k value of KNN. After trials between 1-25, we observed that the most optimal k value is 13. We have seen that the most wrongly predicted music genres in KNN are experimental and pop on the confusion matrix. Meanwhile International and Hip hop music genres were predicted more accurately than other music genres.

Our SVM model found clips in hiphop, rock, and electronic genres highly accurate. However, as in KNN and logistic regression, SVM also made unsuccessful predictions in pop and folk genres. Music such as hip hop, rock tend to have very different sounds such as auto-tuning and therefore the model is easier to distinguish. The same style of instruments is commonly used in pop and experimental music, mostly guitar. This causes the model to mix them easily, resulting in worse performance with both types.

Our DNN model is a 6 layer sequential network, with first layer has 512 neurons, second layer has 256 neurons, third

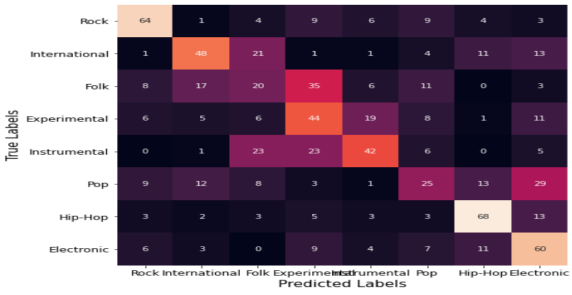

Figure 4. SVM Confusion Matrix

layer has 128, fourth and fifth layer has 64, 32 neurons and the output layer has 7 neurons.

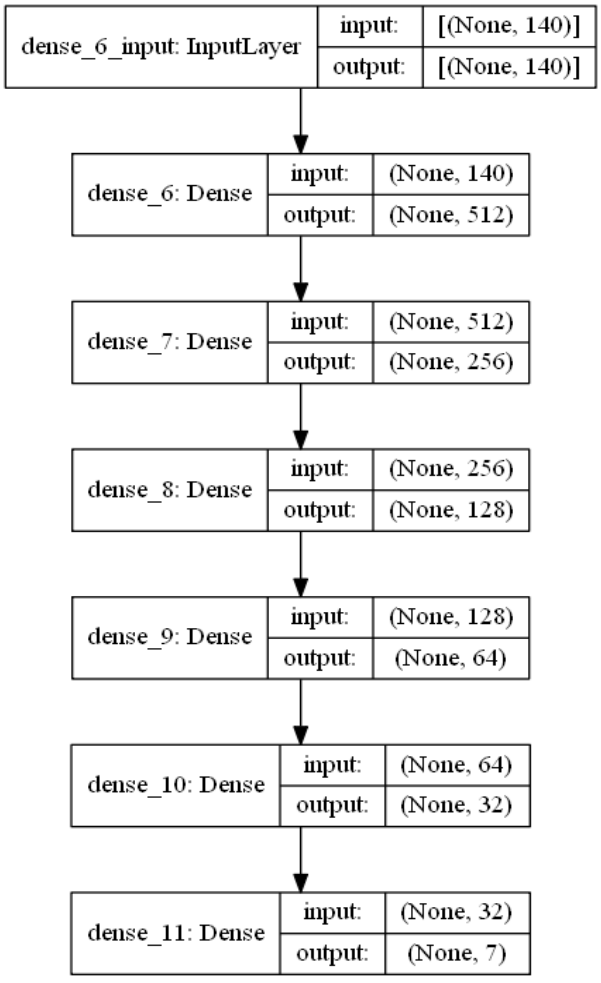

Figure 5. DNN Layers

Sparse categorical crossentropy was the most suitable for loss in the DNN model. We tried changing the optimizer hyperparameter. As an optimizer, we achieved 48.6% accuracy with 'adam' as a result of 20 epoch and batch size 32 When we trained the model with the same values, but this time by 'sgd' the optimizer, the training accuracy was 89% and the test accuracy was 49.3%. We assumed that the stochastic gradient descent optimizer was better than 'adam' and increased the number of epochs when training with sgd. While training the model with 100 epochs, the training accuracy of the model was around 99% after the 35th epoch, and after this epoch, our model's training slowed down a lot. The best value we can reach is 50.3% accuracy, which we found in the test set after 100 epochs for DNN.

The random forest consists of a large number of individual decision trees that operate as a community. Each tree in the random forest gives a class prediction, and the class with the most votes will be the class predicted by model. The biggest advantage of Random Forest is that the model consists of many separate trees. If there are a few more trees that make wrong guesses, the selection is made according to the decision of the majority. The predictions made by individual trees should have low correlations with each other. The model we trained with Random Forest gave the higher accuracy rate than Logistic Regression, kNN and DNN. This model achieved 50% success on test data. Our Random Forest model made nearly 65% accurate predictions on Rock and International genres. The correct estimation rate in Experimental and Pop genres remained at around 30%.

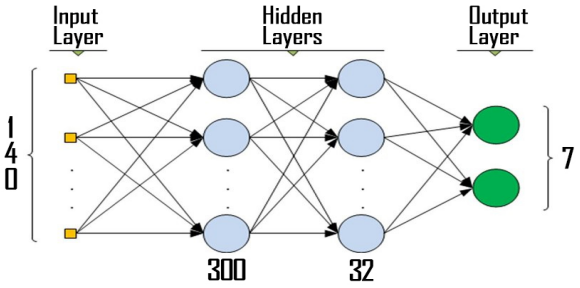

Figure 6. MLP Layers

Multilayer Perceptron classifier which is the subset of DNN. MLP is a feed forward neural network that produces a series of outputs from a series of inputs. It uses backpropagation to train the network. We got better results in ReLu activation function in MLP model training. The model consists of two hidden layers. There are 300 nodes in the first hidden layer and 32 nodes in the second hidden layer. We tried 'sgd' and 'adam' as optimizer. Both gave accuracy at close ratios, but the 'adam' was slightly better. So we trained the model with 'adam' optimizer and activation function 'ReLU'. We observed 54% accuracy in the test data after 30 epochs. We compared the outputs of the models by changing the number of hidden layers and the number of nodes in the hidden layers. We expected the success rate to increase as the number of hidden layers increased. However,

the accuracy values of the models containing more than 2 hidden layers were lower than our 2 hidden layers model. That's why we decided that the most optimal MLP model for the dataset we have is with 2 hidden layers as in the picture above.

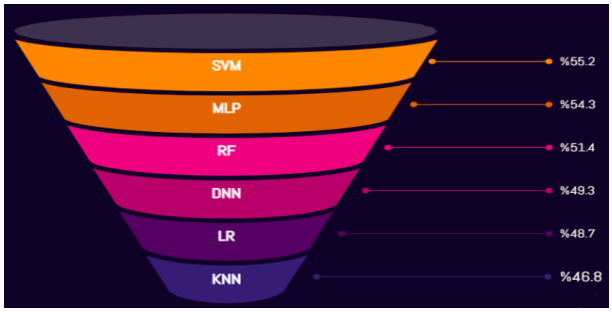

Figure 7. Compare All Methods

In this project, we implemented and discussed various machine learning models such as Nearest Neighbors Classifier, Support Vector Machines, Logistic Regression, Random Forest, Deep Neural Network and Multilayer Perceptron to recognize music genres using FMA dataset. We have carefully chosen hyperparameters to avoid overfitting, especially in Deep Neural Network and Multilayer Perceptron models. We observed the results by reducing the complexity of all models to give the most optimal accuracy.

Our attempts to increase the results were not successful at first. However at some point we've discovered that if we drop some of the genres from the dataset, the models seem to work better. The same style of instruments is commonly used in some music genres, mostly guitar. This caused the models to confuse while learning, so the results were in bad condition with every model. That's why we decided to remove the folk music genre from the dataset to reduce confusion during model training. After removing the folk genre from the dataset, we observed a nearly 10% improvement in success in each model. We found an 8% difference between our highest and lowest models. The results in general were close to each other. The lowest was kNN, even though the process was fast with it since we've used small data, however we got low accuracy because kNN is sensitive to the scale of the data and irrelevant features. Logistic Regression and Random Forest results were similar. We can also say that the best models were SVM and MLP.

To train the CNN model, we need an image as an input. For this, we extracted mel spectrograms from audio files and saved these spectrograms as an image file. We converted 30 second mp3 music clips to wav files. This process took a long time. We extracted the melspectroms from the wav files we obtained with the help of librosa. We saved these spectrogram images we extracted in folders according to 8 music genres in the dataset.

In CNN there are several convolution kernels between the Layers. Each feature map of the previous layer and each convolution kernel perform the convolution calculations, which will generate a feature map of the next layer. The feature map is a helpful tool for the CNN to understand images, so the performance of the CNN will certainly improve if weights are given to a feature map so that effective feature maps have large weights, and feature maps with little effect have small weights.

We used Keras, a deep learning library for Python, which provides a convenient way to define and train our Convolutional Neural Network model. We tried 'sgd' and 'adam' as optimizers during the training phase and observed that the 'adam' optimizer gave better results. In the model trainings we made with 'sgd', the training accuracy showed fluctuating graphics, while the 'adam' optimizer showed a linear increase and a linear decrease in the loss value. Therefore, we used the 'adam' optimizer in the model training phase.

It took very long time to convert mp3 audio files to wav files and then save them as images. At the same time, wav files took up too much space on our computer. Therefore, the images we obtained do not cover the whole data set. The images we obtained were stored on our computer as unbalanced data. For example, while we were able to create 600 images of the Pop type, 250 images of the Instrumental type were created. Because of the unbalanced data, our trained CNN model did not find success at high values.

```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
conv2d_55 (Conv2D)           (None, 62, 62, 32)        896
_________________________________________________________________
max_pooling2d_42 (MaxPooling (None, 20, 20, 32)        0
_________________________________________________________________
batch_normalization_26 (Batc (None, 20, 20, 32)        128
_________________________________________________________________
conv2d_56 (Conv2D)           (None, 18, 18, 64)        18496
_________________________________________________________________
max_pooling2d_43 (MaxPooling (None, 6, 6, 64)          0
_________________________________________________________________
batch_normalization_27 (Batc (None, 6, 6, 64)          256
_________________________________________________________________
conv2d_57 (Conv2D)           (None, 5, 5, 32)          8224
_________________________________________________________________
max_pooling2d_44 (MaxPooling (None, 2, 2, 32)          0
_________________________________________________________________
batch_normalization_28 (Batc (None, 2, 2, 32)          128
_________________________________________________________________
flatten_28 (Flatten)         (None, 128)               0
_________________________________________________________________
dense_55 (Dense)             (None, 64)                8256
_________________________________________________________________
dropout_28 (Dropout)         (None, 64)                0
_________________________________________________________________
dense_56 (Dense)             (None, 8)                 520
=================================================================
Total params: 36,904
Trainable params: 36,648
Non-trainable params: 256
_________________________________________________________________
```

Figure 8. CNN Model Summary

During this process we also deleted wav files after getting

images to open space in our disk but this also resulted with error while converting. So with strong hardware setup we would be able to do this even with larger datasets as we already write the codes for converting wav and getting images.

We observed the results by constructing convolutional neural network models with different numbers of layers. First, we trained a one-layer CNN model. We used kernel size (3,3) , 32 filters and activation function ReLU in the Conv2D layer. We increased the number of Conv2D layers in the model when the accuracy we received was low. We achieved an accuracy rate of around 40% with the three-layer CNN model. As we mentioned before, this rate came to be so low because we used unbalanced data.

6. Future Work and Improvements

In future studies, we plan to produce models with higher accuracy by using FMA medium or large datasets. By working on FMA's medium or larger data set, music genre classification can have more music genres. The fact that the sample numbers of these datasets are more than the small dataset can lead to high accuracy values by ease. We've already tried larger datasets and got higher results but we didn't include in this report as we aimed small dataset at the beginning.

As we mentioned earlier in our experiment section, we can extract spectogram images from the audio files. By adding this with better hardware setup and computers with high processing speed we can work on Convolutional Neural Network further. We can also do this with larger datasets (Small - 7.2GB, Medium - 22GB, Large - 93GB). Do keep in mind these are mp3 files. To get images we need wav files. In example after converting small dataset to wav it become over 50GB.

As an additional feature, we can use the lyrics of the music in the audio files. It is possible to train the lyrics with one model and at the same time train the spectrogram images in another model. A more successful and beautiful application can be created by combining the results with two different models.

References

C.H.L. Costa, J. D. V.-J. and Koerich, A. Automatic classification of audio data. *in IEEE International Conference on Systems, Man and Cybernetics*, pp. 562–567, 2004.

Davis, S. and Mermelstein, P. Experiments insyllable-based recognition of continuous speech. 28:357–366, Aug. 1980.

Defferrard, M., B. K. V.-P. and Bresson. Fma: A dataset for music analysis. *18th International Society for Music Information Retrieval Conference*, 2017. URL https://arxiv.org/abs/1612.01840.

Jaehun Kim, Minz Won, X. S. and Liem, C. Transfer learning of artist group factors to musical genre classification. *Companion Proceedings of the The Web Conference 2018*, pp. 1929–1934, 2018.

Mandel, M. I. and Ellis, D. P. Song-level features and support vector machines for music classification. 2005. URL http://www.ee.columbia.edu/~dpwe/pubs/ismir05-svm.pdf.

T. Li, M. O. and Li, Q. A comparative study on content-based music genre classification. *SIGIR '03 Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pp. 282–289, 2003.

Tzanetakis, G. and Cook. P. musical genre classification of audio signals. volume 10:293–302, July 2002. doi: 10.1109/TSA.2002.800560.