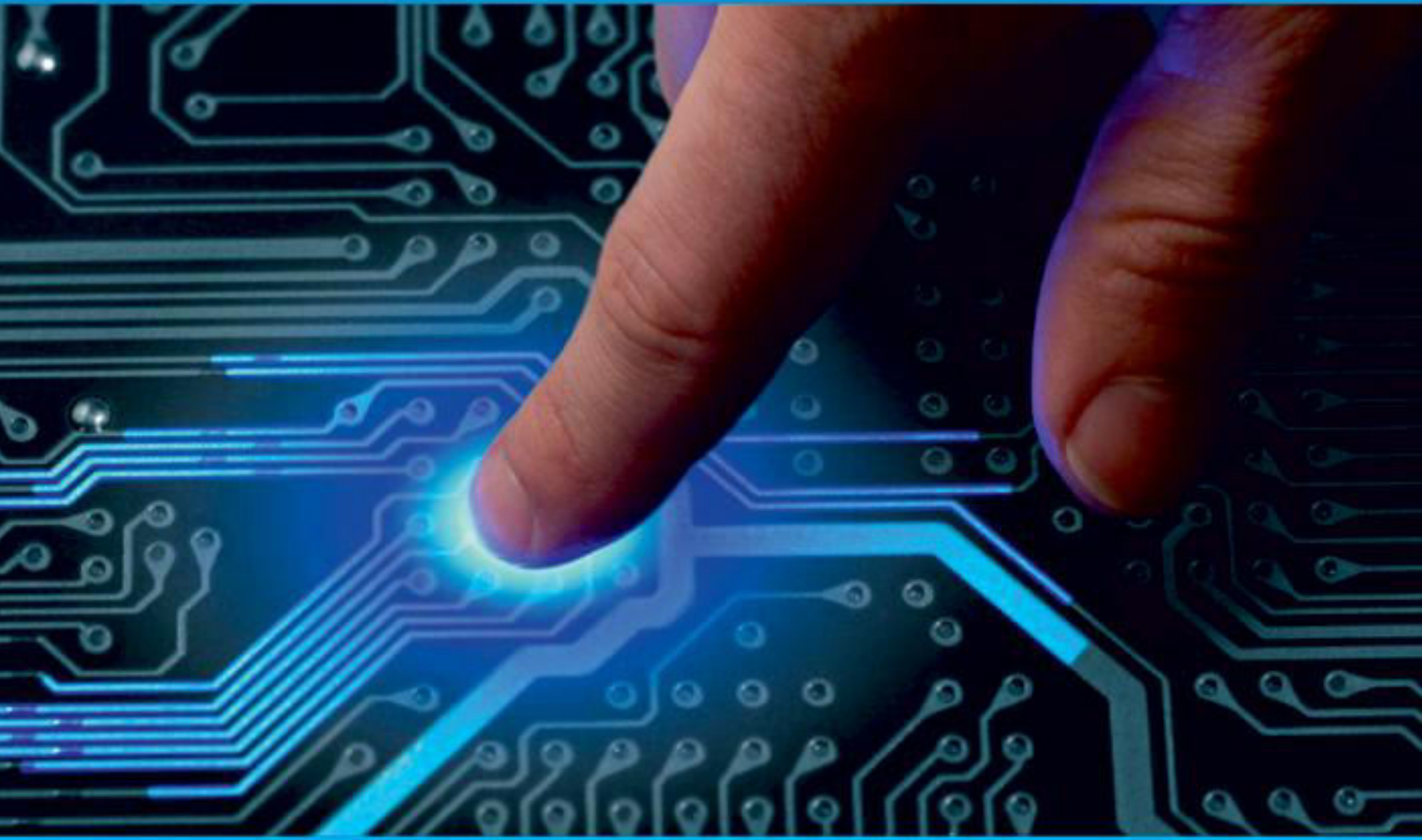
IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798



INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 9, Issue 7, July 2021

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 7.542

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.542 |

|| Volume 9, Issue 7, July 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0907058 |

Survey on Music Genre and Instrument Classification using Machine Learning

Adithi H Koushik[1], Gautam Naik[2], Pranav Prakash K J[3], Suhas N[4], Madhusudan G K[5]

Students, Dept. of Computer Science and Engineering, Vidya Vikas Institute of Engineering and Technology, Mysuru, Karnataka, India[1,2,3,4]

Assistant Professor, Dept. of Computer Science and Engineering, Vidya Vikas Institute of Engineering and Technology, Mysuru, Karnataka, India[5]

ABSTRACT: With the wealth of music available at the fingertips of users around the world, there is an ever-increasing need for automatic classification of music for cataloging of music for organization and quicker retrieval which is often done manually by experts in the field. To further complicate the issue, there is no standard definition on what determines a song's genre, which can be a culmination of various themes and moods that the song generates in listeners. This work aims to examine one of the cornerstone problems of Musical Instrument Retrieval (MIR), in particular, music genre recognition and instrument classification. In particular, this research involves adapting the natural taxonomy of musical genres to generate a machine learning model in an attempt to capture some of the natural hierarchy in music.

I. INTRODUCTION

The aim of this project was to compare machine learning algorithms in their ability to automatically classify song excerpts into the correct musical genre and find out the predominant instrument used. For humans who are familiar with the genres and instruments in question, it is not an especially difficult task. Most people well acquainted with music are able to identify the genre of a song and the instruments used just by listening to the audio. Importantly, they are usually able to perform this type of broad genre classification even if they are unfamiliar with the song or artist in question: 'heard' qualities of the sound are normally enough. As Gjerdingen (ibid.) describes, only a short sample of audio is usually needed, sometimes less than a second. A more fine-grained classification into musical subgenre usually requires some level of expertise on the part of the listener, and the greater the level of expertise, the more accurate the classification that can be made. The question is raised as to whether software can be written to perform such classifications as well as humans, and what type of approaches work best.

II. BACKGROUND

In the era of the internet, music has proven to be one of the most popular classes of information that is viewed and downloaded. There are hundreds of music streaming sites that host tens of millions of songs that deliver content to millions of users worldwide. One of the largest content providers in this space is Spotify, which boasts over 70 million tracks and 356 million users as of December 31, 2019 . Such large collections of music pose an incredible challenge for the organization and retrieval of music.

Extracting information from music to be used in music information retrieval can be described as Content-Based Music Information Retrieval (CB-MIR) . The intent of CB-MIR is to enable automatic cataloguing and retrieving of music. There are several use cases that CB-MIR aims to solve. The use cases span a broad spectrum of potential audiences, such as industries involved in recording, cataloguing, and distribution of music, the end-users who are consuming content from aforementioned industries, and professionals involved in the research and creation of music such as artists, producers, professors, and teachers Music genre has always been a widely used classification for grouping similar songs together. Genre is a fundamental way of describing music and is often how people describe their preferences and tastes when discussing music. With the wealth of available music streaming services that have access to millions of songs in recent years, there is a need in Music Information Retrieval to be able to quickly and automatically classify songs according to classes such as genre so that they can be distributed.

Instrument recognition is also widely studied problem from various perspectives. Essid et al. studied the classification of five different woodwind instruments. Mel frequency cepstral coefficient(MFCC) features were extracted from the training tracks as they were found helpful for classification based on tremolo, vibrato and sound attack. PCA was performed on the MFCC features for dimension reduction before feeding the transformed features to Gaussian Mixture model (GMM) and support vector machine (SVM) classification. GMM with 16, 32 Gaussian components were used, which resulted in better classification accuracy for the later. SVM was also performed with linear and polynomial kernels where the former was found to be efficient.

As part of collecting its Audioset data, Google has also created a simple benchmark, albeit with a much larger set of categories. The benchmark utilized a shallow, fully-connected neural network in conjunction with a pre-trained CNN [Hershey, Gemmeke] to achieve an average precision of 0.314. This project attempts to classify wide variety of sounds, into a variety of categories. Within the realm of music, it includes categories for both the instrument(s) making the sound and the music genre. Music genres are highly abstract categories, over which humans often disagree, whereas it is relatively easy to assign ground truth for instruments.

Furthermore, music recommendation systems are vital to the success of such services as they allow for the tailoring of music towards the customer based on past use and content similarity. Traditional musical classification is a manual process. It is typically performed by domain experts such as artists or curators that have authority and knowledge in the field. With the ever-growing number of songs out in the wild, there is an increasing need to be able to classify these songs automatically and in a timely fashion.

Our work includes:

1. Extraction of information from the audio files.
2. Classifying the music based on the spectrogram.
3. Finding the best method for classification of these audio files.

III. METHODS

- **Datasets**

GTZAN

The first used dataset is GTZAN (Tzanetakis and Cook, 2002). It consists of 10 different genres: blues, classical, country, disco, hiphop, jazz, metal, pop, reggae and rock. Each of the genres consists of 100 tracks with 30 second snippets each, stored as 22050 Hz, 16-bit mono *.au audio files.The 10 genres are as follows: Blues, Classical, Country, Disco, Hip-hop, Jazz, Metal, Pop, Reggae and Rock. It is used as the de facto standard dataset in the domain with more than 100 applications (Sturm, 2013). This allows for comparisons with other researchers operating at the same conditions.

IRMAS

The IRMAS dataset for instrument classification includes musical audio excerpts with annotations of the predominant instruments present and is intended to be used for the automatic identification of the predominant instruments in the music. This dataset was used in the paper on predominant instrument classification by Bosch et al.and includes music from various decades from the past century, hence differing in audio quality to a great extent. In addition, the dataset covers a wide variability in musical instrument types, articulations, recording and production styles, and performers. The dataset is divided into training and testing data, and all audio files are in 16-bit stereo wave with 44,100 Hz of sampling rate. The training data consisted of 6705 audio files with excerpts of 3 s from more than 2000 distinct recordings.

- **Audio Feature**

Mel-Frequency Cepstral Coefficients (MFCCs) are one of the most common feature types used in audio classification of all kinds, having been used in speech and ambient noise recognition as well as music genre classification. Unlike the estimated tempo of a given audio signal, which consists of a single value, ordinarily between 13 and 30 MFCCs are extracted, with the exact number of coefficients.

- **Classification**

Artificial Neural Networks

An Artificial Neural Network (ANN) is an information processing structure that is composed of a large number of highly interconnected processing elements—called neurons or units—working in unison tosolve specific problems. Neurons are grouped into layers (usually called input, output, and hidden) thatcan be interconnected through different connectivity patterns. An ANN learns complex mappingsbetween input and output vectors by changing the weights that interconnect neurons. These changes mayproceed either supervised or unsupervised. In the supervised case, a teaching instance is presented to theANN, it is asked to generate an output, this out is then compared with an expected "correct" output, andthe weights are consequently changed in order to minimize future errors. In the unsupervised case, theweights "settle" into a pattern that represents the collection of input stimulus.

Some ANN architectures are capable of approximating any function. This attribute makes neuralnetworks a good choice when the function to be learned is not known in advance, or it is suspected to benonlinear. ANN's do have some important drawbacks, however, that must be considered before they areimplemented: the computation time for the learning phase is very long, adjustment of parameters can betedious and prohibitively time consuming, and data over-fitting can degrade their generalizationcapabilities. It is still an open question whether ANN's can outperform simpler classification approaches.They do, however, exhibit one strong attribute that recommends their use: once the learning phase iscompleted, the classification decision is very fast when compared to other popular methods such as k-NN.

Support Vector Machines

SVMs are based on statistical learning theory (Vapnik, 1998). The basic training principle underlyingSVMs is finding the optimal linear hyperplane such that the expected classification error for unseen testsamples is minimized (i.e., they look for good generalization performance). According to the structuralrisk minimization inductive principle, a function that classifies the training data accurately, and whichbelongs to a set of functions with the lowest complexity, will generalize best regardless of thedimensionality of the input space. Based on this principle, a SVM uses a systematic approach to find alinear function with the lowest complexity. For linearly non-separable data, SVMs can (non-linearly) mapthe input to a high dimensional feature space where a linear hyperplane can be found. This mapping is done by means of a so-called kernel function.

Although there is no guarantee that a linear solution will always exist in the high dimensional space, inpractice it is quite feasible to construct a working solution. In other words, it can be said that training aSVM is equivalent to solving a quadratic programming with linear constraints and as many variables asdata points. Anyway, SVM present also some drawbacks: first, there is a risk of selecting a non-optimalkernel function; second, when there are more than two categories to classify, the usual way to proceed isto perform a concatenation of two-class learning procedures; and third, the procedure is computationallyintensive.

Random Forests

Random Forests are built from ensembles of decision trees. Each tree is built using a random selection of examples, and each node in a tree splits based on a random selection of the features. The algorithm we used selects √N features out of the total N features for each split. The algorithm trains an ensemble of these decision trees, with each tree being trained differently due to the random selections. To classify an input vector, this algorithm sends the data to each tree in the forest, and determines the best output based on the consensus of the trees.

Since random forest classifiers take only single-dimensional data, we collapsed the 128x10 matrix into a single 1280-element feature vector. We then used a randomized parameter search for hyperparameter optimization, optimizing for the maximum depth, the minimum samples per leaf, the minimum number of samples needed to split a node, the number of estimators, and the method used to determine the number of features used to determining splits in the tree. We also tested whether to use bootstrap samples for tree building, and found that this configuration was clearly disadvantageous. We ran the random search with 60 candidate sets of hyperparameter configurations, with three runs for each configuration for a total of 180 training runs. The results of random forest modelling are discussed in the following section.

Recurrent Neural Networks

Recurrent Neural Network (RNN) is a collection of a man made neural networks where the connections between nodes form a directed graph along a temporal sequence. This permits it to exhibit a temporal dynamic behaviour. Derived from feed forward neural networks, RNNs can use their internal state (memory) to process the variable length sequences of given inputs. This makes it applicable to tasks like unsegmented, connected handwriting recognition or speech recognition.

Recurrent nets are a type of artificial neural network designed to recognize patterns in sequences of data, such as textual data, genomes, audio, video, or numerical times series data emanating from sensors, stock markets and government agencies. These algorithms take time and sequence into account, they have a temporal dimension. RNNs are applicable even to images, which can be decomposed into a series of patches and treated as a sequence.

Recurrent networks are distinguished from feed forward networks by that feedback loop connected to their past decisions, ingesting their own outputs moment after moment as input. It is often said that recurrent networks have memory. Adding memory to neural networks has a purpose: There is information in the sequence itself, and recurrent nets use it to perform tasks that feedforward networks can't. Long Short-Term Memory networks – usually just called "LSTMs" – are a special kind of RNN, capable of learning long-term dependencies. They were introduced by Hochreiter & Schmidhuber (1997).

Convolutional Neural Networks

The model of CNN has variety of layers in hierarchical order in a particular sequence. A typical model sometimes comprises the layers of convolutional wherever the contents of visual (i.e., spectrograms) are diagrammatic a group of features obtained when the input involved with a spread of extractors that are learned throughout the part of training. The layer of Pooling is also presented when to accumulate the convolutional layer most function of activation from Convolutional model. As results of pooling, spatial resolution of those maps is reduced. Moreover, CNNs may additionally contain absolutely connected (FC) layers wherever every somatic cell of the input layer is linked with each alternative neuron within the layer. The Convolutional, pooling, and FC layers are caring the removal pipeline that designs the input data in abstract form. At last, a softmax layer involves the ultimate recognizing processes supported this illustration.

IV. POSSIBLE OUTCOMES

The possible outcomes of this project are as follows:

1. Retrieving information of music files using their spectrogram.
2. Classifying music files into their respective genre.
3. Recognizing the predominant instrument used in the audio files.
4. Discovering the optimal model for classification of music files based on their accuracy and data loss.

V. CONCLUSION

In conclusion, we train six models, one model or retrieving spectrograms of the audio files and five more for classification of the audio files into their respective genres and predominant instruments, namely, Artificial Neural Networks (ANNs), Recurrent Neural Networks (RNNs), Support Vector Machines (SVMs), Random Forests, Convolutional Neural Networks (CNNs). We found that CNNs surpassed the other models in terms of accuracy and data loss. We described how to apply CNN to classify music files based on their genre and predominant instrument in the real-world music. We train the network using fixed-length single-labeled data, and identify an arbitrary number of the genre and the predominant instrument in a music clip with a variable length. Our results show that CNN is capable of achieving good performance by learning the appropriate feature automatically from the input data. Mel-spectrogram was used as an input to the models, and we do not use any source separation in the preprocessing unlike in existing works.

REFERENCES

[1] G. Tzanetakis and P. Cook "Musical genre classification of audio signals", IEEE Transactions on speech and audio processing, vol. 10, pp. 293-302, 2002.

[2]Jia Dai,Wenju Liu, Hao Zheng,WeiXue and Chongjia Ni, "Semi-supervised learning of bottleneck feature for music genre classification", Springer Nature, T. Tan et al. (Eds.): CCPR 2016, Part II, CCIS 663, pp. 552562, Singapore 2016.

[3] Juan J Bosch, Jordi Janer, Ferdinand Fuhrmann, and Perfecto Herrera. A comparison of sound segregation techniques for predominant instrument recognition in musical audio signals. In ISMIR, pages 559–564, 2012.

[4]Beth Logan et al. Mel frequency cepstral coefficients for music modeling.

[5]Karpathy. "Convolutional Neural Networks for Visual Recognition." CS231n Convolutional Neural Networks for Visual Recognition. N.p., n.d. Web. 14 June 2019.

[6] https://valeriovelardo.com/the-sound-of-ai/

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details