Music Instrument Recognition from Spectrogram Images Using Convolution Neural Network

G Jawaherlalnehru, S Jothilakshmi

ABSTRACT- The projected system presents a unique approach to instrument Recognition (MIR) supported Convolution Neural Networks (CNNs). Previous MIR strategies are supported planning and extracting spectrogram features from the audio signal so as to explain its characteristics and classify it. In distinction, CNNs learn the features directly from the input file. The model has evidenced successful in solving various advanced multimedia system information Retrieval (MIR) issues, like image classification or voice recognition. The projected system seeks to explore whether or not the success may be imported to MIR also. The results from this work shows 97% accuracy for all instrument classes.

Keywords - Music Information Retrieval, Musical Instrument Recognition (MIR), Spectrogram images, Convolution neural network.

I. INTRODUCTION

Automated instrument recognition has been at the core of sound research for a long time. Taking care of the issue would permit gains in related region, for example, automated translation, source partition and some more. In spite of the fact that it appears to be a clear issue, computer battle in the assignment set for basic sound recordings that comprise just of a only one instrument.

Convolution neural networks (CNNs) are actively used for a variety of music classification functions such as music tagging [1][2], genre classification[3][4], and user-item latent characteristic prediction for recommendation[5]. Most past strategy requires a double pipeline: first, descriptors should be removed utilizing a predefined calculation and parameters; and second, temporal models require an extra tied over the proposed descriptor. In this manner, descriptors and fleeting models are regularly not mutually planned. All through this work, the framework investigates perceiving instrument by deep learning with the info set to be log magnitude spectrogram. CNN characteristics are in different levels of hierarchy and could be extracted by kernels of convolution. Throughout supervised training, stratified characteristics are learned to perform a specific job. As an approach, learned characteristics from a CNN that is trained for low-level choices for genre classification to high-level features [6]. The most common approach to applying CNNs to an audio-recognition task is employing a spectrogram image as the feeding data [7, 8]. On the other hand, there have been a few attempts at feeding a raw time-series audio signal into a neural network for music [9] and speech [10] signals. However, we suggest in this article another technique for feeding the phase information into a neural network.

Revised Manuscript Received on July 05, 2019.

G.Jawaherlalnehru, Department of Computer Science and Engineering, Annamalai University, Annamalainagar. Tamilnadu.

S. Jothilakshmi, Department of Information Technology, Annamalai University, Annamalainagar. Tamilnadu..

Because a time-series is a sequence of one-dimensional data, the filter (kernel) used for the time-series in CNNs should also be one-dimensional. This restriction has given rise to a limitation when applying a filter convolution because information in a certain temporal location can only be convoluted a single time; in contrast, a spectrogram image, which is two-dimensional, can be analyzed using a two dimensional filter multiple times per single temporal location, which provides more dimensions for an analysis.

In this work, the framework has registered spectrogram picture highlights to segregate the instrumental signals and classify the signals (CNN) technique is utilized. The rest of the document is sorted as follows: The outline of linked work is given in Section 2.The proposed technique is depicted in Section 3 where the framework has exhibited the component calculation and classification plan. Section 4 introduces the experimental outcomes and the concluding comments are presented in the section 5.

II. RELATED WORK

Mostly, the automatic instrument recognition in instrument classification of feature observe activities or monophonic phrases and melodies used by single instruments. Classification situations with 10 classes are considered for the best-performance to find instrument class. The best structures acquire focus charges above 90%, as shown for instance in [11, 12].

The deep learning network specifically CNN has been widely used for a variety of image processing [13]. Since, these successful strategies were implemented in music instrument recognition (MIR) such as chord recognition [14] and track transcription [15], where the work extensively improved upon previous results. Likewise, the first successful automated instrument recognition(AIR) method based on the deep learning algorithm has been presently proposed and designed from the convolution layers for learning, and fully-linked layers for recognition [16, 17]. Park et al. uses CNN to recognize instruments the purpose of tone recordings [16].

Furthermore, most CNN designs utilize unique channel shapes in each layer [1] [17]. Ongoing works call attention to that utilizing diverse channel frames in each layer is a proficient method to exploit CNN's ability [18] [19]. For instance, Pons et al. [19] suggested utilizing various musically motivated channel frames in the initial layer to proficiently display a few musically important time-scales for learning fleeting features.

In this job, the model discusses whether the latest achievement of deep learning in other areas can be imported into MIR. The following sections explain what artificial deep neural

networks are, also regarded as deep learning.

The model will aim on a particular variety of architecture of deep learning, namely ConvNets, which is the one used in this study.

III. PROPOSED METHODOLOGY

The proposed system plans to use the procedure of feature learning for spectrograms created from music instrument. The features are extracted from spectrograms automatically to form MIR. The main elements of the proposed system are described in the following areas.

A. Spectrograms

A spectrogram defines signal strength of visual information at various frequencies available in input waveform. The spectrogram represents two dimensional graph contains horizontal and vertical axis for frequency and amplitude. These are basic components specified using by colour in a particular time in the spectrogram. Low amplitudes indicated by dark blue. Strong amplitude indicated by red colour. The amplitudes are computed by Fast Fourier Transform (FFT) in speech signal and it gives time frequency. The frequency is separated into small portions and implemented in wave form of the music. Spectrograms of instruments are showed in Figure 1.

Spectrograms provide a facility to analyze speech includes sound event recognition [20], speaker recognition [8], speech emotion recognition (SER) [21] and speech recognition [22].

B. 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 feature obtained when the input involved with a spread of extractors that are learned throughout the part of training.

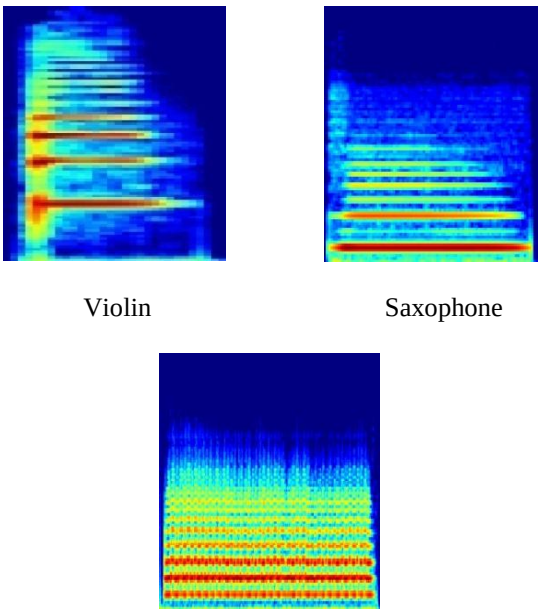

Figure 1: Sample Instrument spectrograms

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.

C. Model Architecture

The suggested model of the CNN, proven in Figure. 2 Consists of three convolutionary layers, three fully connected layers and a Softmax layer. The network input is a 150x 150 spectrogram produced from instrument signals. The preliminary convolutionary layers extract elements of these spectrograms from the use of convolution activities. Layer C1 has 32 (3x 3) kernels that are used at a four-pixel step environment. It is detected by rectified linear gadgets (ReLU) and a 3x 3 max pooling layer with stride 2. ReLU acts as activation functions as a replacement for the usual softmax functions that improve the effectiveness of the training method. Layer C2 has 32 kernels of dimension 3x 3 and they are used to enter with a step 1. Similarly, C3 has 64 kernels of dimension 3x 3 each of these convolution layers are provided by ReLU units. The feature extractor is attached to the classifier by an intermediate globally max pooling layer and flat layer (F). Finally, a totally connected layer (FC), 0.5 dropout, and 10 classes of softmax activation (S) of the final output layer are used. The model was trained with a learning speed of 0.001, a batch size of 32, and the rms prop optimizer. Compared to natural images, it is highly difficult to acquire discriminative features from spectrograms for strong instrument recognition.

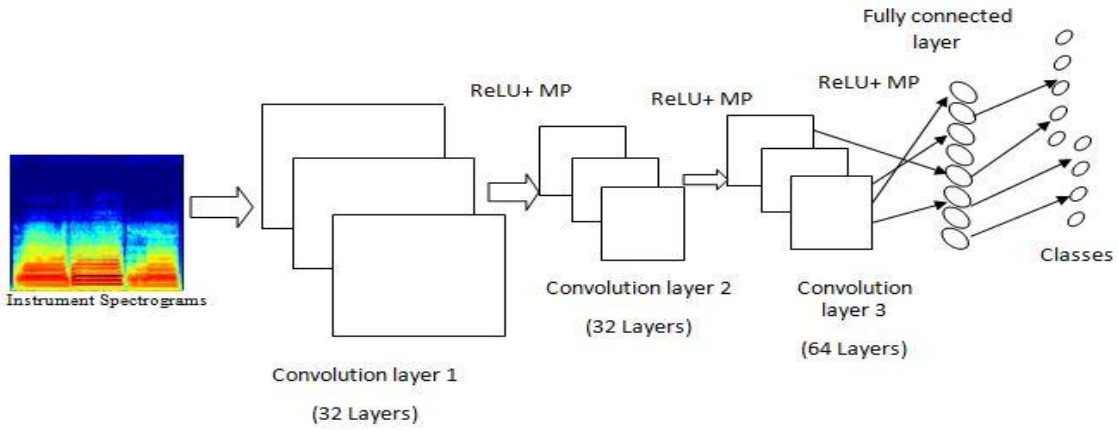

Figure 2: CNN architecture of proposed system for Music Instrument Recognition using spectrogram

IV. EXPERIMENTAL RESULTS AND ANALYSIS

The experimental configuration and evaluation of the results are provided in this section.

A. Dataset

IRMAS (Instrument Recognition in Music Audio Signals) dataset was used to assess the performance of proposed MIR system. Every audio file utilizes one of the ten different instruments such as bassoon, contrabassoon, cello, clarinet, French horn, viola, violin, saxophone, bass clarinet, cor anglais. Each music file contains around 1sec duration. The data sampling frequency is 22050 Hz. Samples are 16-bit and mono-type. Eighty percent of the information was used for training and fifteen percent of the information is used for validation and the remainder was used for testing.

B. Model Training

Monaural audio signals are processed as a first level at a sampling rate of 22050 Hz. It then estimates a 2048 window size Mel spectrogram, a 512 hop length size, and 96 Mel bands. After applying logarithmic magnitude compression, one-second fast spectral patches are used as input to the deep neural network.

The spectrogram pictures had been generated in Anaconda IPython from the IRMAS sample data and size to a 150 x 150. 4500 spectrograms have been produced from all the audio samples in dataset. 80% of the data were once used for training and 15% percentage of the data used to be used for validation and relaxation was used in the test phase. For each instrument, the model has gathered about 400 pictures in sample dataset. The process of training was once executes for 30 epochs with a bunch measurement fixed to 32. The learning charge was once set to 0.001 with a delay of 0.1 after each 100 epochs.

C. Instrument Prediction using CNN

The model CNN presented in Figure 2 trained on the spectrograms produced from the IRMAS dataset. The model used to perform predictions with accuracy above 90% for ten exclusive instrument units which include bassoon, contrabassoon, cello, clarinet, French horn, viola, violin, saxophone, bass clarinet, cor anglais.

The model trained and used to attain predictions for spectrogram of each instrument generated for the instrumental information. The softmax layer in the CNN model predicts for ten various instrumentals in shape of probabilities are used in imply prediction primarily created reasoning technique to get usual prediction ratings for instruments. Prediction described by the CNN model for individual spectrogram performance as signal to update the acceptance values for all instrumentals. The guesses from CNN model are amassed to decide the probabilities for character instrumentals by means of computing suggest predictions from the collected evidence.

D. Prediction Performance

In training process the system has taken round 60 minutes and satisfactory precision was executed after 30 epochs. In the training data set, a loss of 0.24 was attained, whereas 0.23 loss was once logged on the test dataset. An accuracy of 97 % was once achieved per spectrogram.

Prediction outcomes (i.e. possibilities of actual class) for spectrograms produced for number instrumental sound archives from the test dataset are provided in Figure 3. The results of the CNN proposed framework is successfully expecting most of the instrumentals using making excessive outputs extra than 50% of the time. In viola and violin instruments, the overall performance is acceptable but no longer very exact due to the fact some of the spectrograms are stressed with different instruments. Overall, the proposed CNN method assisted to acquire an accuracy of 97% to all the gadgets on the test set.

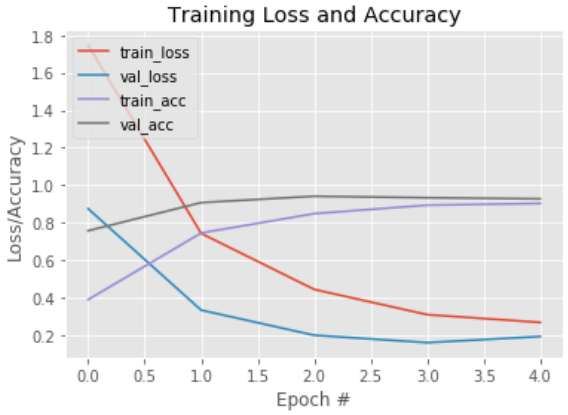

Figure 3: CNN Performance of proposed system with 30 epochs

V. CONCLUSION

In the proposed system, the model try to solve the trouble of MIR the use of feature gaining knowledge of system based on the network of deep convolution. Music signal is represented spectrograms which performance on the input to deep CNNs. The CNN model consisting of 3 convolution and 3 absolutely layers extract features of spectrograms and predictions of ten instrumental classes. In this work, CNN model used to be skilled created on the spectrograms produced for the IRMAS dataset. The results had been performed the use of the model for most of the instruments. The future work is wanted to recover the proposed CNN framework so that all instruments are diagnosed efficiently in robust way. Future idea is to utilize data with comparatively complicated models to expand music instrument recognition performance.

REFERENCES

1. S. D. Man and B. S.wen. “End to end learning for music audio”. International conference on acoustics, speech and signal processing, pages: 6964-6968, 2014.
2. K. Choi, G. Fazekas and M. Sandler, “Automatic tagging using DCNN”, 2016.
3. S. Sigta , S. Dixon . “Improved music features learning with DNN”. International conference on acoustics, speech and signal processing, pages: 6959-6963, 2014.
4. P. C. Guano and G. Fazekas. “Hybrid music recommender using CBSI”, International conference on acoustics, speech and signal processing, pages: 2618-2622. 2016.
5. A.V. D. Oord, S. Dieleman at all. “Deep Content based music recommendation”, Advance in neural information processing systems. Pages: 2643-2651. 2013.
6. K. Choi, G. Fazekas, “Explaining DCNN on music classification”. 2016.
7. O. A. Hamid, A. Mohamed. “CNN for speech recognition” Audio, speech and language processing, IEEE/ACM. Pages: 1533-1545. 2014.
8. H. Lee and P. Pham, “Unsupervised feature learning for audio classification using CDBN”, Advances in neural information processing system. 2012.
9. S. Dieleman and S. Benjamin, “End to end learning for music audio”, International conference on Acoustics, speech and signal processing. 2014.
10. D. Palaz and M. M. Doss, “Estimating phoneme class conditional probabilities from raw speech signal using CNN”, Proceedings of Interspeech. 2013.
11. Jakob, M. Grasis and H. Lukashevich, “ A multiple expert framework for instrument recognition”, International symposium on computer music multidisciplinary research. 2013.
12. S. Tjoa and K. J. Ray, “MIR using biologically inspired filtering of temporal dictionary atoms”, International society for MIR conference. 2010.
13. Y. Bengio and A. Courville, “Deep learning”, MIT press. 2016.
14. G. Widmer and F. Korzeniowski, “A fully convolutional deep auditory model for musical chord recognition”, IEEE International workshop on MLSP. Pages: 1-6. 2016.
15. Peter Li, B. McFee, J.Salamon, “Deep salience representations for f0 estimation in polyphonic music”, International society of MIR. 2017.
16. T. Park ad T. Lee, “Musical instrument sound classification with DCNN using feature fusion approach. 2015.
17. Y. Hab and J. Kim,” DCNN for predominant instrument recognition in polyphonic music”, IEEE/ACM Transactions on audio, speech and language processing. pages: 208-221. 2017.
18. H.Phan, L. Hertel and M. Maass, “Robust audio event recognition with 1-max pooling CNN”. 2016.
19. J. Pons and Xavier, “Designing efficient architectures for modelling temporal features with CNN”, International conference on acoustics, speech and signal processing”. 2017.
20. J. Dennis and H. Li , “Spectrogram image feature for sound event classification in mismatched conditions”, IEEE signal processing letters. Pages: 130-133. 2011.
21. M. Dong, Z. Huang and Y. Zhan, “Learning salient features for SER using CNN”, IEEE Transactions on multimedia, pages: 2203-2213. 2014.
22. M. L. Seltzer, J. Li and F. Seide, “Feature learning in DNN studies on speech recognition tasks”. 2013.

AUTHORS PROFILE

First Author G. Jawaherlalnehru M.E, Ph.D (CSE) received M.E degree from Anna University in 2014, Chennai. India. He is currently pursuing Ph.D in Annamalai University, India. His research interests are the area of speech processing, machine learning and deep learning.

Second Author S. Jothilakshmi received Post-Doctoral degree in Marshall University, U.S.A., Ph.D (CSE) degree in Annamalai University, India and M.E degree in Annamalai University, India. Her interests are the area of Speech Processing and Big Data Analysis. She is Associate Professor in Annamalai University, Tamil Nadu, India. She has 20 years’ experience in teaching.