International Journal of Engineering Applied Sciences and Technology, 2020
Vol. 5, Issue 1, ISSN No. 2455-2143, Pages 674-678
Published Online May 2020 in IJEAST (http://www.ijeast.com)

SOUND CLASSIFICATION SYSTEM USING MACHINE LEARNING TECHNIQUES

Dr. S. Veena
Professor,
Department of CSE
S.A. Engineering College
Thiruvekadu, Chennai,
Tamil Nadu, India.

Nerisai. M. V
UG Student,
Department of CSE
S.A. Engineering College
Thiruvekadu, Chennai,
Tamil Nadu, India.

Remya. J. V
UG Student,
Department of CSE
S.A. Engineering College
Thiruvekadu, Chennai,
Tamil Nadu, India.

Sai Tejah.S
UG Student,
Department of CSE
S.A. Engineering College
Thiruvekadu, Chennai,
Tamil Nadu, India.

***Abstract*- Sound is the most important communicative tool among all living organisms but day to day life of people added more type of sounds to the natural environment which may be useful or may not be. So identifying and understanding the sounds from the regular communicative one is high time need of the day. This paper deals with the type of sounds in the urban areas. So the classification of sounds may help the machine to understand the type of sound. This paper analysis talks about various techniques that are used to classify the sounds and to make the machine to learn and analyse the data so as to give the output accordingly. These kinds of analysis also will help to identify criminal activities. In addition, this paper also viewed the different forms of input and other parameters that can be used for classification. Advantages and disadvantages of the methods were also considered.**

Keywords- Convolutional Neural Networks, Machine Learning, Mel-Frequency Cepstral Coefficients.

I. INTRODUCTION

Urban sounds can be easily predicted from a human perspective. But making a machine to learn that sound and classify it to its respective class is still a tedious process. Be it any kind of sound archive, the pre-defined one or the defined one, the system should be able to grasp them, classify them and produce the output to the user. To do so, it needs to learn about the sounds and the variations it possess. Allowing the machine to learn from the sound archive is what we call as Machine Learning (ML). ML is as such a vast subject where it holds many techniques to solve different kinds of complex problems. One such technique mainly used for classifying the urban sounds is the “Feature Extraction Techniques”. This feature extraction technique holds responsible for extracting the sound, compressing it, segregating it and then evaluating it. There are special sub-categories on this technique where all of the above named processes are individually carried out. Various papers presented in this paper use these sub-categories either separately or in combination to train its system. So every paper has provided unique accuracy which proves that all these techniques have both advantage and disadvantage. Further to enhance the output, these techniques make use of Convolutional Neural Networks (CNN) and Deep Neural Networks (DNN) in unmasking the sound.

II. OVERVIEW ON EXISTING TECHNIQUES

David Li *et al*.,(2013) [1] proves the accuracy with the help of Tree Bagger method. First, it aggregates all its classes and clubs them to form a collection tree. These tree sets were used as predictors which were then compared with the actual archives to produce the output. This type of classification process provides less error thus giving higher precision rates. Once the trees are classified individually, it is then sent to a voting based system which tells what classifier has got the highest accuracy rate and in this way the output is provided to the user with stability.

Ayu and Karyono (2014) [2] conducted trails for this technique to prove that it is unapproachable. Hence, they have disproved the theory of “four condition classifier” in building this system.

Jiaxing Ye *et al*., (2016) [3] mentioned that a better understanding of the urban sound can lead to a better environment. This paper has proved to be an application oriented one where the sensor is seem to be deployed in traffic signals so that it could detect the ambulance sound and modulate the signals of this junction making sure that no traffic occurs. Yet, another application would be to deploy it in smart city’s lamp posts where it can detect the gunshots and notify the police with location. Hence, it gives surveillance and safety in both cases.

Karol and Piczak (2015) [4] has written a paper where the experimental model was evaluated with a 5 fold and 2 fold architecture of neural networks and the validation was done. When the machine was manually engineered to classify the sound, it gave poor performance and when it was automated with convolutional neural networks, it was performing at its best. They have used the probability-voting scheme than the traditional majority-voting setup as it proved to be more favourable than the latter.

Yuji Tokozume and Tatsuya Harada (2017) [5] say that the detailed parameter of the system has two features and the set is trained accordingly. The two features are: the static log-mel as one-channel input and the static and delta log-mel [17] as second-channel input. Though there are two different features they possess the same architecture thus providing an accuracy of 71.0% which is capable of contributing in the improvement of the classification performance. They also proved that the system is capable of extracting a discriminative feature that compliments the log-mel features.

Pooja *et al*., (2018) [6] used a system where the input is given to the feature extractor using software which then becomes a feature set. This set is sent to the classification model that is designed using neural networks and machine learning that process the set and categorizes them. The above happened during the training. While the system is tested, the input is processed as it was processed in the training and then they will be classified and provides the output.

Minkyu Lim *et al.*, (2018) [7] have created a system that uses audio event classifier based on Convolutional Neural Networks. This system uses a feature extractor where the audio signal (input in the form of image) that gets transformed into a PCM format. The event classifier used here is made up of three layers: convolutional, pooling and fully connected. The convolutional layer identifies the concurrences that have occurred in the previous layer [25]. Pooling layer helps to reduce the dimension and combines the features that are alike into one. The fully connected layer identifies the sound of the image input.

Mendoza *et al*., (2019) [8] have done the experiments by using three different forms of input for sound classification. They are Constant-Q transform (CQT) [13], Spectrogram features [14] and Spectrogram images. Out of these three, it is proved that CQT is the most suitable feature technique. Here, a three second audio which is given as input using probability voting and the output delay is based on the number of windows processed by the system. The accuracy of the classification techniques is determined by testing the dropout [18] and Batch Normalization [19] in CNN architecture.

Bruno da Silva *et al*., (2019) [9] says that the accuracy obtained for the embedded devices decreased rapidly with increasing versions. The reason stated is connected to the Librosa library [16] packages that are responsible for audio feature extraction. 50-60% of accuracy was obtained in recognizing different sound categories. Google's Edge Tensor processing Unit (TPU) [15] was used to facilitate the edge computation process for embedded devices. It was found through analysis that each and every class did not give the same precision throughout the classification process because every dataset had its own variations of segregating the sound clips.

Aditya Khamparia *et al*., (2019) [10] have designed a model using Convolutional Neural Network with 2 layer architecture: fully connected and prediction layer to classify the environmental sounds [23]. The first layer has 32 filters whereas the second one has 64 both with ReLU activation [22]. The first layer processes the features and the prediction layer predicts the final output (i.e. class). This provides up to 77% accuracy. Here, Tensor Deep Stacking Network Toolkit [24] is used as it gives the functions used for testing and training purposes.

Marc Green and Damian Thomas Murphy (2019) [11] have created an iOS app for sound classification using Augmented Reality (AR) and Machine Learning. The AR part [20] is used in order to display the virtual objects. This app also uses a feature extraction technique called MFCC that extracts the audio and sends it to the Core ML [21]. Core ML converts the model into iOS-supporting format and it also creates an object in the app. For classification of sounds, Gaussian Mixture model and Support Vector Classifiers are used.

Afshan kaleem and Santi Prabha (2019) [12] have analysed different feature extraction techniques and classification models in this paper. It then tells which feature extraction technique provides good results when used in different classification models as a result. In this analysis, the datasets (audio clips) are arranged into ten classes to make the process easier. Librosa library is used to acquire various features of audio data that are considered to be useful. After the experiments, they have concluded by saying that Mel-Frequency Cepstral Coefficients (MFCC) provides the highest accuracy when used in the models.

III. COMPARISONS

S.NO	TITLE	TECHNIQUES/DOMAIN	ADVANTAGES	DISADVANTAGES
1	Auditory Scene classification using Machine Learning Technique	Feature Extraction techniques: Short term feature extraction Wavelet decomposition MFCC Voting based System	Classification takes place independently on different features. The one that gives highest accuracy from the voting based system is chosen.	The scenes that occur in tubes and tube station were found to be difficult in classifying due to the variations in its occurrence.
2	Android based sound detection application for hearing impaired using ADABoostM1 classifier with RepTree Weak Learner	AdaBoostM1 acts as classifier Rep Tree acts as the Weak Learner	Proved to be a non-approachable technique.	The classifiers used here provide only 40%-50% accuracy.
3	Urban Sound Event Classification based on local and global features Aggregation	Feature Learning Descriptive statistics Aggregation pattern to combine both local and global acoustic features	Can be deployed in a real-life scale environment that presented enormous potential to facilitate people’s urban life with various aspects.	Noise reduction was found to be the major problem faced.
4	Environmental Sound Classification with Convolutional Neural Networks	Mel-Frequency Cepstral Coefficients(MFCC) 2-layer Convolutional Neural Network Dropout Learning	With the limited number of datasets, it is proved that higher accuracy can be achieved.	It gives poor performance for short-scale temporal sounds like engine, drilling and Jack hammer.
5	Learning Environmental sounds with end-to-end Convolutional Neural Network	Novel end-to-end ESC system and CNN Delta log-mel + logmel-CNN System	The system has a frequency response similar to that of human perception.	Accuracy is achieved only when there is multi-convolutional layer with small filter size.
6	Sound classification using Machine Learning and Neural Networks	Machine Learning: Support Vector Machine Random Forest Neural Networks MFCC	When the input is sent to classifier model for processing, it will be divided and comparison is done with the labels which is stored in a file.	If the classification of sound is done with Support Vector Machine techniques, the accuracy is reduced.
7	Convolutional Neural Network based audio event classification	Convolutional Neural Networks Gaussian Mixture Models Deep Neural Networks Short time Fourier Transform Mel scale filter bank	In comparison with baseline classification model, the CNN classification model provides higher performance that	Some classes that contain sounds like crowd and wind shows less performance than the other classes.

			leads to high accuracy.	
8	Audio Event detection using wireless sensor networks based on deep learning	Convolutional Neural Networks Constant-Q transform Wireless Sensor Network Ambient Assisted Living STFT	Using batch normalization and dropout, the parallel architecture of CNN is proved to be more stable and also achieves high accuracies.	In CNN, the configuration of the model fails if the input is a spectrogram (it is a time frequency data).
9	Evaluation of classical Machine Learning towards urban Sound recognition on Embedded Systems	Wireless Sensor Networks Machine Learning Techniques embedded with devices	It provides more flexibility to prioritize precision and timing.	When the Convolutional Neural Network classifier was used, the execution time increased by a factor of 100.
10	Sound Classification using Convolutional Neural Network and Tensor Deep Stacking network.	Convolutional Neural Networks Tensor Deep Stacking Networks	77% of accuracy is achieved while using ESC-10 dataset sound archive for the CNN technique.	Only 49% of accuracy was obtained for ESC-50 sound archive while using TDSN.
11	Environmental sound monitoring using machine learning on mobile devices.	MFCC Support Vector Classifier Gaussian Mixture Model Augmented Reality	The results shows that machine learning techniques can be used to calculate soundscapes.	The classifier is not trained enough to detect human ratings.
12	Enhancement of Urban Sound Classification ditUsing various Feature extraction Techniques	The following are the feature extraction techniques used here: Tonnetz Chroma-short time fourier transform Mel-spectrogram MFCC Spectral Contrast Classification Techniques: Naïve Bayes Random Forest J48 Decision Tree Support Vector Machine	Predicting a test dataset class is fast and easy here.	Zero Frequency is one of the main disadvantages in this system. Also, Naïve bayes techniques is said to be bat at estimation.

IV. CONCLUSION

It is clearly seen from the paper that though there exists many techniques to classify the urban sounds, the right combination of Machine Learning Technique with the sound archive system alone produces good results. To make the machine understand the sound that has occurred in the environment will prove to be useful for both research and surveillance purposes. It is also to be noted that training a machine to classify a particular sound and prove that it is equally capable as humans in predicting the environment.

ACKNOWLEDGEMENT

We would like to thank our Management, Advisor and Principal of S.A. Engineering College for constantly motivating us to do research in our work environment and encourages us to publish more research papers.

V. REFERENCES

[1] Li David, Tam Jason and Toub Derek (2013): Auditory Scene Classification using Machine Learning Techniques in IEEE AASP Challenge on Detection and Classification of Acoustic Scenes and Events.

[2] Shekar Melati Ayu Indah (Ayu) and Karyono Kanisius (Karyono) (2014): (AudiTion)Android based Sound Detection application for Hearing Impaired using AdaBoostM1 Classifier with RepTree WeakLearner in APCASE- Asia Pacific Conference on Computer Aided System Engineering .

[3] Ye Jiaxing, Kobayashi Takumi and Murakawa Masahiro (2016): Urban Sound Event Classification based on Local and Global features Aggregation in Applied Acoustics.

[4] Piczak Karol J. (2015): Environmental Sound Classification with Convolutinal Neural Networks in IEEE 25th International Workshop on Machine Learning for Signal Processing (MLSP). (pp. 1–6).

[5] Tokozume Yuji and Harada Tatsuya (2017): Learning Environmental Sounds with end-to-end Convolutional Neural Network in ICASSP- IEEE International Conference on Acoustics, Speech and Signal Processing.

[6] R.K Pooja, Shetty Srishti, M Suhani and Mr. D.R. Janardhana (2014) Sound Classification using Machine Learning and Neural Networks in IJIRT-International Journal of Innovative Research in Technology.

[7] Minkyu Lim, Donghyun Lee, Hosung Park, Yoseb Kang, Junseok Oh, Jeong-Sik Park, Gil-Jin Jang and Ji Hwan Kim (2018): Convolutional Neural Network based audio Event Classification in KSII (Korean Society for Internet Information) Transactions on Internet and Information System.

[8] Mendoza Jose Marie, Tan Vanessa, Fuentes Jr. Vivencio, Perez Gabriel and Tiglao Nestor Michael (2019): Audio Event Detection using Wireless Sensor Network based on Deep Learning in International Wireless Internet Conference.

[9] Silva Bruno da, Happi Axel W., Bracken An and Touhafi Abdellah (2019): Evaluation of classical Machine learning Techniques towards Urban Sound Recognition on Embedded Systems in Applied Sciences.

[10] Khamparia Aditya, Gupta Deepak, GiaNhu Nguyen and Krishna Ashish (2019):Sound Classification using Convolutional Neural Network and Tensor Deep Stacking Network in IEEE Access.

[11] Green Marc and Murphy Damian Thomas (2019): Environmental Sound Monitoring using Machine Learning on Mobile devices in Applied Acoustics 159.

[12] Afshankaleem and Prabha I. Santi (2019): Enhancement of Urban Sound Classification using various Feature Extraction Techniques in IJRTE-International Journal of Recent Technology and Engineering.

[13] Lidy T., Schindler, A (2016): CQT-based convolutional neural networks for audio scene classification and domestic audio tagging in Proceedings of the Detection and Classification of Acoustic Scenes and Events 2016 Workshop (DCASE2016), DCASE2016 Challenge, vol. 90.

[14] Nanni, L., Costa, Y.M.G., Lucio, D.R., Silla, C.N. Jrand and Brahnam, S (2017): Combining visual and acoustic features for audio classification tasks in: Pattern Recognition Letters, vol. 88 (pp. 49–56).

[15] Cass, S (2019): Taking AI to the edge: Google’s TPU now comes in a maker-friendly package in IEEE Spectr.

[16] LibROSA 0.6.3. 2019., Available online: https://librosa.github.io/librosa/.

[17] Sainath Tara N, Weiss Ron J, Senior Andrew, Wilson Kevin W, and Vinyals Oriol (2015): Learning the speechfront-end with raw waveform cldnns in Proc. Interspeech.

[18] Srivastava Nitish, Hinton Geoffrey, Krizhevsky Alex, Sutskever Ilya, and Salakhutdinov Ruslan (2014): Dropout: A simple way to prevent neural networks from overfitting in The Journal of Machine Learning Research, vol. 15, no. 1 (pp. 1929- 1958).

[19] Ioffe Sergey and Szegedy Christian (2015): Batch normalization: Accelerating deep network training by reducing internal covariate shift in Proc. ICML.

[20] Apple ARKit (2019) in online: https://developer.apple.com/arkit.

[21] CoreML (2019) in Online: https://developer.apple.com/documentation/coreml.

[22] Gupta D., Ahlawat A. (2018): Taxonomy of GUM and Usability Prediction using GUM Multistage Fuzzy Expert System, in International Arab Journal of Information Technology.

[23] Chu S., Narayanan S., C and Kuo C, J. (2009): Environmental sound recognition with time–frequency audio features, IEEE Transactions on Audio, Speech, and Language Processing 17.

[24] Palzer D. and Hutchinson B. (2015): The Tensor Deep Stacking Network Toolkit, in IJCNN- International Joint Conference on Neural Networks.

[25] LeCun Y., Bengio Y., and Hinton G. (2015): Deep learning in Nature, vol. 521, (pp.436-444).