International Journal of Technical Innovation in Modern Engineering & Science (IJTIMES)
Impact Factor: 5.22 (SJIF-2017), e-ISSN: 2455-2585
Volume 5, Issue 04, April-2019

REVIEW OF CONVOLUTIONAL NEURAL NETWORK FOR SOUND CLASSIFICATION

Anam Bansal

Assistant Professor, Central University of Punjab, Bathinda

Abstract— ***Audio event detection and audio event classification are emerging field. The detection and classification of audio and sounds are prominently accomplished using the machine learning models like Support Vector Machines, K- Nearest Neighbour, Artificial Neural Networks, etc. These models give considerable accuracy. But deep learning models have surpassed the machine learning models in detecting and classifying the audio events, sources of sound and audio scenes. Convolutional Neural Network(CNN) is the deep learning model that has been used extensively in the field of audio. It has already made astonishing achievements in the field of image classification and other computer vision applications. This paper introduces the usage of CNN in the field of audio classification, describes the CNN model, layers of CNN, feature extraction in CNN. Then, the use of CNN in the various areas related to sound are described.***

Keywords— Deep Learning, Convolutional Neural Network, Audio classification, Audio detection.

I. INTRODUCTION

Audio detection and classification is the study of detecting the audio scenes and events and further classifying them. Audio scenes are the environment surrounding the sounds such as home, restaurant, office, meeting rooms, etc. Audio events are audio signals at the particular point of time like birds singing, the car passing by, , etc. Audio scene classification assigns semantic labels to temporal regions of sound recordings. Audio event classification systems aim to classify sound events in the audio recognizing start and stop times of events.

At present, audio recognition technology has a great commercial market and several benefits. Speech Processing finds a number of applications in speech recognition [1][2], speech synthesis [3]. In the field of music, various musical instruments [4], genres of music [5] are classified using machine learning models. In the field of medicines and hospitals, the normal and pathological data are discriminated using audio [6]. The invisible damage in the panels of aircraft which is difficult to detect by staff inspectors can be detected using audio processing[7]. Low flying aircraft can be detected through sound processing [8]. Various events such as sounds of closing or opening doors, dropping or breaking objects, gunshot, dog barking and screams can help in identification of various abnormal activities [9]. Wang et al. [10] presented a robust environmental sound recognition system for home automation. In machine learning models, features are extracted manually and fed to machine learning classifier. But CNN is the black box, features are extracted and classification is performed internally. CNN has been used massively for sound related applications.

II. CONVOLUTIONAL NEURAL NETWORK

CNN is an artificial neural network and a deep learning model. It is a variation of Multilayer Perceptron (MLP) Networks and requires very less preprocessing as compared to MLP. Instead of handcrafted features as in the case of machine learning algorithms, CNN has the ability to learn features on its own. It was developed primarily for tasks of object recognition. E.g. handwritten digit recognition [11]. CNN uses a mathematical operation called convolution instead of matrix multiplication in at least one of its layers. The CNN architectures are usually built with four main types of layers: Convolutional layer, Detector Layer, Pooling layer and Fully Connected layer (Fig. 1).

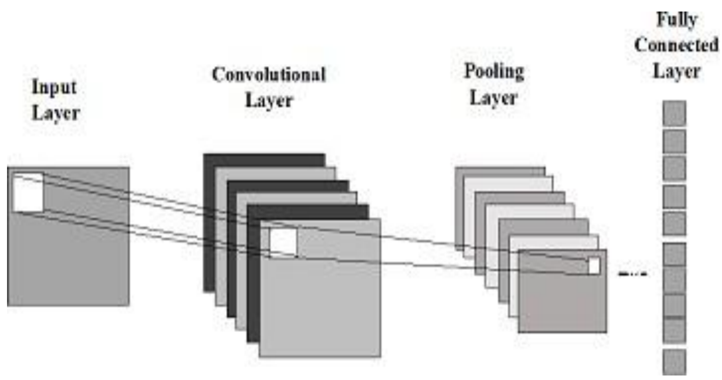

Fig. 1: Layers of Convolutional Neural Network

A. Convolutional Layer

The convolutional layer has a convolution operation between two signals or functions. One dimensional convolution can be depicted as below:

$$F(t) = (f * g)(t) = \sum_{a=-\infty}^{\infty} f(a)\,g(t-a) \tag{1}$$

Where *f* is the input and *g* is the filter. *t* is the time index and *a* is the value of time shift. Feature Map, also called activation map is the output of the convolution. This layer consists of a set of learnable filters which convolve across the input and produces a feature map. Different filters lead to different feature maps. It implies that filters are activated when certain features are detected. In the case of sounds, features can be a certain frequency component or pitch.

B. Detector Layer

A non-linear activation function called ReLU is used which makes the linear output from the previous layer to be non-linear. The ReLU activation function is preferred over tanh and sigmoid activation functions because ReLU removes vanishing gradient problem [12]. ReLU makes all the negative values as zero.

$$f(x) = max(0, x)$$

C. Pooling Layer

The spatial size of the output from the previous layer is reduced by the pooling layer. The value of the output is replaced by the aggregation operation on its neighborhood values as the relative location of the features is more important than the actual location. The aggregation or pooling operations used are max-pooling, min-pooling, average and L2 norm pooling, out of which max-pooling is the most common one. Pooling layer reduces the chances of overfitting and the number of parameters to be learned.

D. Fully connected Layer

Fully connected layer takes as input the output of the previous layer(convolutional or detector or pooling layer). The output of this layer is the vector whose length is equal to the number of classes into which the data is to be classified. Different activation functions can be used in this layer. The most commonly used activation function is softmax function which outputs the vector whose each value is the probability of the input value belonging to the class.

III. CNN IN SOUND RELATED APPLICATIONS

CNN has been widely employed for various sound related applications. In CNN, hierarchical feature learning takes place.

In the field of biodiversity, CNN has achieved remarkable performance through audio sensing. It is used to classify various bird species in large scale dataset and has achieved a considerable accuracy [13] [14]. Bat calls are detected and classified using CNN which are ultrasonic in nature [15].

In the field of medicines and hospitals, CNN is used to classify the heart sounds as pathological, non- pathological or uncertain [16]. In this approach, human-driven features, Mel Frequency Cepstral Coefficients(MFCCs) are given as input to CNN. CNN is employed for recognizing S1 and S2 heart sounds when duration and interval information between S1 and S2 are not available [17]. The MFCCs are clustered using the K-means algorithm and further classified using CNN. CNN is employed for feature extraction only and classification is performed using a machine learning model to classify heart sounds and it performed well [18].

Audio scenes such as beach, car/bus, cafe, park, etc. are classified using CNN [19]. Vehicles are classified using sounds. CNN is used for extracting features from the sounds of vehicles and further classification is performed using a machine learning model i.e. Support Vector Machine(SVM)[20]. CNN classifies short audio clips of environmental sounds and achieves the accuracy more than state of art methods[21][22] [23]. Large scale audio classification in videos is performed using deep Convolutional Neural Network [24] [25] CNN is used for recognizing and classifying various sound events using spatial features[26]. In the case of the weakly labeled dataset, CNN achieves good performance in classifying various sound events [26]. CNN is applied for Polyphonic sound event detection[27] and other speech event and audio event recognition tasks [28] [29]. The audio can be converted to images and further CNN can be used for classifying the audio because CNN became immensely popular through image classification tasks only[30].

Dividing the audio recordings into frames can give good results as audio recordings are continuous signals and using frames, even the small events can be captured and the signals can be considered quasi-stationary. CNN works well when applied on framed audio signals [31]. Traditional speech recognition algorithms such as Hidden Markov Models(HMM) are replaced by CNN to recognize speech [32] [33]

Deep learning models need large scale datasets and a huge amount of data to train efficiently. If the data available is less, then techniques of data augmentation can be employed to increase the data. CNN has given good results after data augmentation in case of audio events [34] and environmental sound classification [35]. In some cases, injecting noise can lead to better results for backpropagation networks like CNN [36].

IV. CONCLUSION

Convolutional Neural Network is used widely in various classification tasks such as image and sound classification. In this paper, the usage of CNN in sound-related applications is described. CNN is extensively used in environmental sounds, heart sounds, bird sounds classification, and speech, etc. Transfer learning can help in audio classification using CNN [37] [38] [39].

REFERENCES

[1] Bhupinder Singh, Neha Kapur, and Puneet Kaur. "Speech recognition with a hidden Markov model: a review". In: *International Journal of Advanced Research in Computer Science and Software Engineering* 2.3 (2012).
[2] Mathias De Wachter et al. "Template-based continuous speech recognition". In: *IEEE Transactions on Audio, Speech, and Language Processing* 15.4 (2007), pp. 1377–1390.
[3] Andrew J Hunt and Alan W Black. "Unit selection in a concatenative speech synthesis system using a large speech database". In: *1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings*. Vol. 1. IEEE. 1996, pp. 373–376.
[4] Jing Liu and Lingyun Xie. "SVM-based automatic classification of musical instruments". In: *2010 International Conference on Intelligent Computation Technology and Automation*. Vol. 3. IEEE. 2010, pp. 669–673.
[5] George Tzanetakis and Perry Cook. "Musical genre classification of audio signals". In: *IEEE Transactions on speech and audio processing* 10.5 (2002), pp. 293– 302.
[6] Zulfiqar Ali et al. "Clinical informatics: mining of pathological data by acoustic analysis". In: *2017 International Conference on Informatics, Health & Technology (ICIHT)*. IEEE. 2017, pp. 1–8.
[7] LP Dickinson and Neville H Fletcher. "Acoustic detection of invisible damage in aircraft composite panels". In: *Applied Acoustics* 70.1 (2009), pp. 110–119.
[8] Richard O Nielsen. "Acoustic detection of low flying aircraft". In: *2009 IEEE Conference on Technologies for Homeland Security*. IEEE. 2009, pp. 101–106.
[9] Aki Harma, Martin F McKinney, and Janto Skowronek. "Automatic surveillance of the acoustic activity in our living environment". In: *2005 IEEE International Conference on Multimedia and Expo*. IEEE. 2005, 4–pp.
[10] Jia-Ching Wang et al. "Robust environmental sound recognition with fast noise suppression for home automation". In: *IEEE Transactions on Automation Science and Engineering* 12.4 (2015), pp. 1235–1242.
[11] Deep Learning. *Ian Goodfellow, Yoshua Bengio and Aaron Courville*. 2016.
[12] Sepp Hochreiter. "The vanishing gradient problem during learning recurrent neural nets and problem solutions". In: *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 6.02 (1998), pp. 107–116.
[13] Stefan Kahl et al. "Large-Scale Bird Sound Classification using Convolutional Neural Networks." In: *CLEF (Working Notes)*. 2017.
[14] Sharath Adavanne et al. "Stacked convolutional and recurrent neural networks for bird audio detection". In: *2017 25th European Signal Processing Conference (EUSIPCO)*. IEEE. 2017, pp. 1729–1733.
[15] Oisin Mac Aodha et al. "Bat detective-Deep learning tools for bat acoustic signal detection". In: *PLoS computational biology* 14.3 (2018), e1005995.
[16] Jonathan Rubin et al. "Classifying heart sound recordings using deep convolutional neural networks and mel-frequency cepstral coefficients". In: *2016 Computing in Cardiology Conference (CinC)*. IEEE. 2016, pp. 813–816.
[17] Tien-En Chen et al. "S1 and S2 heart sound recognition using deep neural networks". In: *IEEE Transactions on Biomedical Engineering* 64.2 (2017), pp. 372–380.
[18] Michael Tschannen et al. "Heart sound classification using deep structured features". In:2016 *Computing in Cardiology Conference (CinC)*. IEEE. 2016, pp. 565– 568.
[19] Michele Valenti et al. "A convolutional neural network approach for acoustic scene classification". In: *2017 International Joint Conference on Neural Networks (IJCNN)*. IEEE. 2017, pp. 1547–1554.
[20] Anam Bansal et al. "An Off the Shelf CNN Features Based Approach for Vehicle Classification Using Acoustics". In: *International Conference on ISMAC in Computational Vision and Bio-Engineering*. Springer. 2018, pp. 1163–1170.
[21] Karol J Piczak. "Environmental sound classification with convolutional neural networks". In: *2015 IEEE 25th International Workshop on Machine Learning for Signal Processing (MLSP)*. IEEE. 2015, pp. 1–6.
[22] Inyoung Hwang, Hyung-Min Park, and Joon-Hyuk Chang. "Ensemble of deep neural networks using acoustic environment classification for statistical model-based voice activity detection". In: *Computer Speech & Language* 38 (2016), pp. 1–12.
[23] Yong Xu et al. "Unsupervised feature learning based on deep models for environmental audio tagging". In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 25.6 (2017), pp. 1230–1241.
[24] Shawn Hershey et al. "CNN architectures for large- scale audio classification". In: *2017 IEEE international conference on acoustics, speech and signal processing (icassp)*. IEEE. 2017, pp. 131–135.
[25] Yong Xu et al. "Large-scale weakly supervised audio classification using gated convolutional neural network". In: *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2018, pp. 121–

125.

[26] Sharath Adavanne, Pasi Pertilä, and Tuomas Virtanen. “Sound event detection using spatial features and convolutional recurrent neural network”. In: *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2017, pp. 771– 775.

[27] Emre Cakır et al. “Convolutional recurrent neural networks for polyphonic sound event detection”. In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 25.6 (2017), pp. 1291–1303.

[28] Haomin Zhang, Ian McLoughlin, and Yan Song. "Robust sound event recognition using convolutional neural networks”. In: *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE. 2015, pp. 559–563.

[29] Huy Phan et al. “Robust audio event recognition with 1-max pooling convolutional neural networks”. In: *arXiv preprint arXiv:1604.06338* (2016).

[30] Venkatesh Boddapati et al. “Classifying environmental sounds using image recognition networks”. In: *Price- dia computer science* 112 (2017), pp. 2048–2056.

[31] Szu-Yu Chou, Jyh-Shing Roger Jang, and Yi-Hsuan Yang. “FrameCNN: A weakly-supervised learning framework for frame-wise acoustic event detection and classification”. In: *Recall* 14 (2017), pp. 55–4.

[32] Ossama Abdel-Hamid et al. “Applying convolutional neural networks concepts to hybrid NN-HMM model for speech recognition”. In: *2012 IEEE international conference on Acoustics, speech and signal processing (ICASSP)*. IEEE. 2012, pp. 4277–4280.

[33] Ossama Abdel-Hamid et al. “Convolutional neural networks for speech recognition”. In: *IEEE/ACM Transactions on audio, speech, and language processing* 22.10(2014), pp. 1533-1545.

[34] Naoya Takahashi et al. “Deep convolutional neural networks and data augmentation for acoustic event detection”. In: *arXiv preprint arXiv:1604.07160* (2016).

[35] Justin Salamon and Juan Pablo Bello. “Deep convolutional neural networks and data augmentation for environmental sound classification”. In: *IEEE Signal Processing Letters* 24.3 (2017), pp. 279–283.

[36] Kartik Audhkhasi, Osonde Osoba, and Bart Kosko. “Noise-enhanced convolutional neural networks”. In: *Neural Networks* 78 (2016), pp. 15–23.

[37] Andreas Fritzler, Sven Koitka, and Christoph M Friedrich. “Recognizing Bird Species in Audio Files Using Transfer Learning.” In: *CLEF (Working Notes)*. 2017.

[38] Keunwoo Choi et al. “Transfer learning for music classification and regression tasks”. In: *arXiv preprint arXiv:1703.09179* (2017).

[39] Aäron Van Den Oord, Sander Dieleman, and Benjamin Schrauwen. “Transfer learning by supervised pre-training for audio-based music classification”. In: *Conference of the International Society for Music Information Retrieval (ISMIR 2014)*. 2014.

