Journal of Artificial Intelligence and Capsule Networks (ISSN: 2582-2012)
www.irojournals.com/aicn/

Audio Tagging Using CNN Based Audio Neural Networks for Massive Data Processing

J. Samuel Manoharan

Professor, Department of Electronics & Communication Engineering, Sir Isaac Newton College of Engineering and Technology, Nagapattinam, India

E-mail: drjsm1530@ieee.org

Abstract

Sound event detection, speech emotion classification, music classification, acoustic scene classification, audio tagging and several other audio pattern recognition applications are largely dependent on the growing machine learning technology. The audio pattern recognition issues are also addressed by neural networks in recent days. The existing systems operate within limited durations on specific datasets. Pretrained systems with large datasets in natural language processing and computer vision applications over the recent years perform well in several tasks. However, audio pattern recognition research with large-scale datasets is limited in the current scenario. In this paper, a large-scale audio dataset is used for training a pre-trained audio neural network. Several audio related tasks are performed by transferring this audio neural network. Several convolution neural networks are used for modeling the proposed audio neural network. The computational complexity and performance of this system are analyzed. The waveform and leg-mel spectrogram are used as input features in this architecture. During audio tagging, the proposed system outperforms the existing systems with a mean average of 0.45. The performance of the proposed model is demonstrated by applying the audio neural network to five specific audio pattern recognition tasks.

Keywords: Transfer learning, pretrained audio neural networks, audio pattern recognition, audio tagging, machine learning

1. Introduction

Among the existing machine learning applications, research in audio pattern recognition gains significance and attraction as it plays an important role in the current smart systems. The events happening around us and our location consists of sounds that are rich in

DOI: https://doi.org/10.36548/jaicn.2021.4.008
Received: 03.11.2021, received in revised form: 26.11.2021, accepted: 12.12.2021, published: 24.12.2021

data. Sound event detection, speech emotion classification, music classification, acoustic scene classification and audio tagging are some of the most commonly used audio pattern recognition applications. Individual researchers used their private datasets for audio pattern recognition related work in early days. Classification of sounds such as poured water, dropped metal, opening and shutting of a wooden door are performed using the hidden Markov model (HMM) by the authors in [1]. Several datasets for sound event detection, acoustic scene classification and so on are made publicly available with the challenge series - Detection and Classification of Acoustic Scenes and Events (DCASE) [2]. The research interest towards the field of audio pattern recognition has increased with the DCASE challenges. Across six subtasks, around 394 entries were received at the most recent DCASE 2021 challenge. When large scale datasets are used, the efficiency of the audio pattern recognition system performance is an open question. ImageNet, a large scale dataset is used for building several image classification systems in computer vision. Wikipedia and other large scale text datasets are used for building several language models in natural language processing [3]. However, only a limited number of large-scale audio datasets and trained systems exist.

In this paper, a large scale pre-trained audio neural network dataset consisting of 527 sound classes and 1.9 million audio clips is used [4]. The performances of the audio neural networks were verified on the audio tag accuracy with its computational time. The wavegram CNN and mel spectrogram are combined to achieve a precise and efficient audio tagging system [5]. When compared to the existing state of the art models, the proposed model offers a better audio tagging solution. Various pattern recognition tasks can be performed in an efficient manner with this model.

2. Audio Tagging Systems

Audio clips are analyzed and the presence or absence of audio tags are predicted with the crucial audio pattern recognition task termed as audio tagging [6]. Mel-Frequency Cepstrum Coefficients (MFCCs), zero crossing rate, audio energy and other manually designed features are used as an input to the audio tagging systems available in the existing research [7]. Discriminative Support Vector Machines (SVMs) [8], Hidden Markov Models (HMMs) [9], Gaussian Mixture Models (GMMs) [10] and other generative models are used for the purpose of classification. Audio recording tags are also predicted with Convolutional Neural Networks (CNNs) [11] and other neural network-based models in recent research. In sound event detection, acoustic scene classification and several other DCASE challenge tasks, state of the

art performance are achieved by the CNN based systems [12]. However, limited sized datasets with few sound classes are focused in those works and a multiple sound classes are not recognized. In this work, a large scale pre-trained audio neural network dataset is used for overcoming the general issue of audio tagging.

2.1 Convolution Neural Networks

2.1.1 Conventional CNNs:

Image classification and other related computer vision tasks have successfully made use of CNNs [13]. Several convolution layers are present in a CNN. The input feature maps are convoluted with several kernels at each convolution layer and the local patterns are captured [14]. The log mel spectrograms are commonly used as inputs for audio tagging by the CNN based systems. Spectrograms are calculated by applying time-domain waveforms with Short Time Fourier Transforms (STFTs) [15]. Further, the log mel spectrograms are extracted by applying the logarithmic model with a mel filter bank.

2.1.2 CNN for audio tagging:

The cross-task CNN systems are used by several audio neural network systems on the DCASE challenge model. The representation ability is improved by adding the penultimate CNN layer with an additional fully-connected layer [16]. 14, 10 and 6 layer CNNs are available. AlexNet based 4 convolution layers are present in the 6-layer CNN. The kernel size of 5 x 5 is used at each convolution layer. The 14 and 10 layer CNNs are inspired by the CNNs similar to VGG and make use of 6 and 4 convolution layers respectively [17]. 3 x 3 kernel size in 2 convolutional layers are present in each convolutional block [18]. The training speed is increased and stability is achieved by adding ReLU nonlinearity while applying batch normalization between each convolutional layer [19]. Downsampling is performed by applying each convolutional block with an average pooling size of 2 x 2 as this outperforms the max pooling of size 2 x 2.

2.2 ResNets

2.2.1 Conventional residual networks (ResNets)

Audio classification is performed in an enhanced manner with deep CNN when compared to the conventional CNN models [20]. However, in deep CNN models, propagation

of gradients is not performed efficiently from the bottom layers to the top layers. This is addressed by introducing shortcut connections between convolution layers using ResNets. Direct propagation on either forward or backward direction from one layer to another is enabled by this technique. With some extra computational complexity, a few extra parameters are introduced by the shortcut connections [21]. Shortcut between the input and output connectivity and multiple data blocks with two convolution layers and kernel size of 3 x 3 are available in ResNet. These basic blocks in the ResNet can be replaced with a network-in-network architecture of three convolution layers in each bottleneck block.

2.2.2 ResNets for audio tagging

ResNet is executed such that the first step involves downsampling layer and convolutional layer incorporated in the log mel spectrum to decrease the spectrogram size and input log [22]. In this work three ResNets are incorporated such as: 38 layers with 16 basic blocks, 22 layers with 8 basic blocks and 68 layers with 32 basic blocks.

3. Proposed Hybrid- CNN system

The one-dimensional CNN systems that previously exist do not outperform the systems that are tuned by log mel spectrograms as a tuning parameter. The most important drawback of the existing CNN models in time domain is their inability to capture the frequency pattern of the audio with respect to several pitch shifts as there is an absence of frequency axis in the one-dimensional CNN and the system is not suitable to capture frequency details.

3.1 Wavegram CNN

In this paper, a CNN based wavegram is introduced along with Hybrid-CNN. This is a time-domain oriented audio tagging model same as like log mel spectrogram, the wavegram is a feature that is built using the incorporation of neural networks [23]. This methodology introduces a time-frequency representation in the wavegram which is incorporated with Fourier transform modification. It holds a wavegram with a frequency and time axis. It is also crucial for audio pattern recognition to detect the frequency patterns like various pitch shifts based on the sounds in a particular class. The frequency information is learnt by the wavegram such that there is a lag in CNN one-dimensions [24]. These wavegrams are also used to enhance the log mel spectrograms by way of learning and adapting from the data with time-frequency. Similarly based on the input features fed from wavegrams replacing log mel spectrograms leads

to the Wavegram CNN system. There are 4 steps involved in building the proposed Hybrid CNN:

1. A one-dimensional CNN is applied as the first step in the transformation process. A convolutional layer with 5 strides and length of 11 is used as the filter to decrease the input size. This results in a subsequent reduction of memory usage and length of inputs by almost 6 times.

2. The next step involves the use of two convolutional blocks with two convolution layers that can be used to improve the total reception. A stride of 5 is used to downsample each of the convolutional blocks. With the help of the stride, it is possible to downsample the audio of 32 kHz into a simple frame count of 100/s.

3. The 1D-CNN layer formed thus far has an output size of TxC where C represents the number of channels and T is the number of frames. This is reframed to TxFxC/F such that there are C/F groups and every group is fixed with frequency bins 'F'. This is known as a Wavegram.

4. In the C/F channels, the frequency information is learnt by the Wavegram using the frequency bins, F.

3.2 Wavegram Logmel CNN

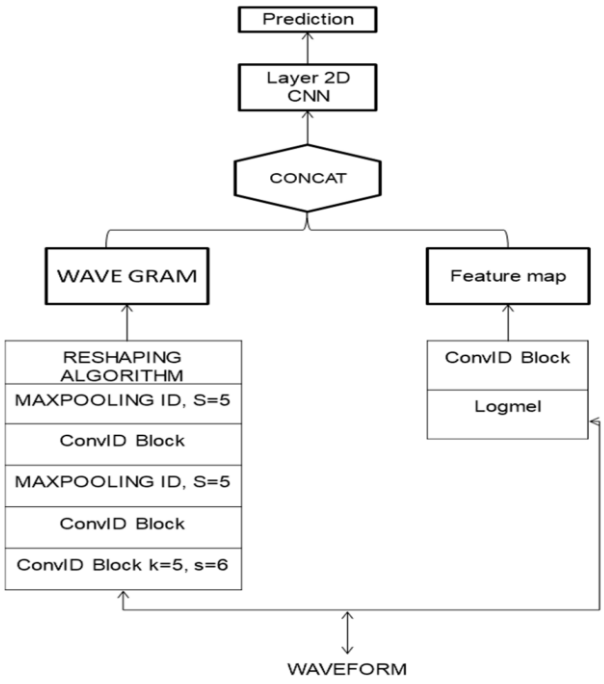

Figure 1. Architecture of Hybrid- CNN

The Architecture of the proposed hybrid CNN is represented in Fig. 1. A combination of log mel spectrogram and wavegram results in the hybrid CNN. This can be used to extract information in log mel spectrograms and time-domain waveforms. The channel dimension is used to guide the combination. This methodology also provides extra details for complementing logmel spectrogram and audio tagging.

4. Results and Discussions

4.1 Data Processing

Data augmentation and data balancing are the two aspects used for audio tagging in which data processing is involved.

- Data Augmentation: Overfitting can be prevented in a system with the help of data augmentation. A small number of training clips are present in the audio datasets and are used to define the pretrained audio neural networks limitation in terms of performance. In this methodology, SpecAugment [24] and mixup [25] are combined which leads to better audio quality.

- Data Balancing: This represents the total audio clips that are available for training between two sound classes. Based on the sound class, they are also distributed. In a pre-trained audio neural network, the inputs are training data and this data needs to be sampled uniformly before being used. Sometimes, data in a mini-batch may be a part one sound class. In such cases, the pretrained audio neural networks are trained with a balanced sampling strategy.

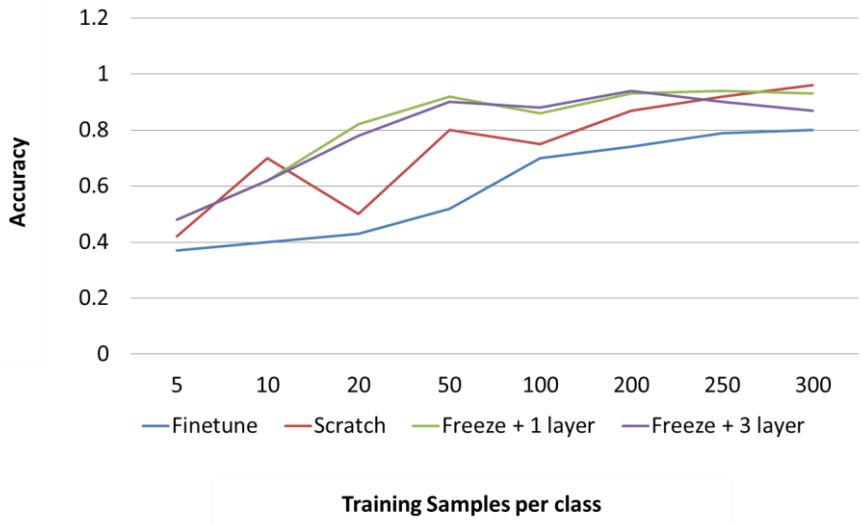

Figure 2. Accuracy of Training Samples for MSoS

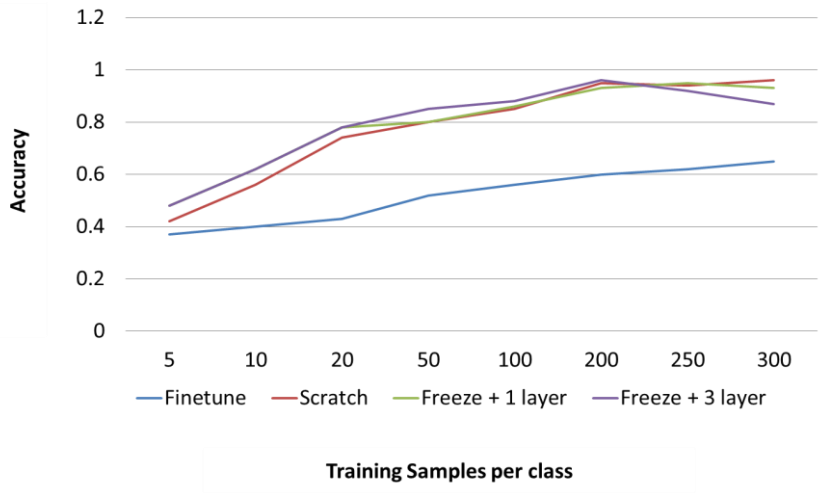

Figure 3. Accuracy of Training Samples for GTZAN

The accuracy of two sets of training samples in MSoS and GTZAN are observed in Fig. 2 and Fig. 3 respectively.

4.2 Evaluation metrics

The evaluation metrics for audio tagging are d-prime, mean Area Under the Curve (mAUC) and mean Average Precision (mAP). The area under the precision and recall curve is the average precision which doesn't depend on count of true negatives. However, AUC comprises both true and false positive rate that is observed in true negative outcomes. Every individual class is used for calculating the metrics and an average is calculated. Fig. 4. shows a comparison of the various audio Tagging mechanisms with respect to the average precision accuracy.

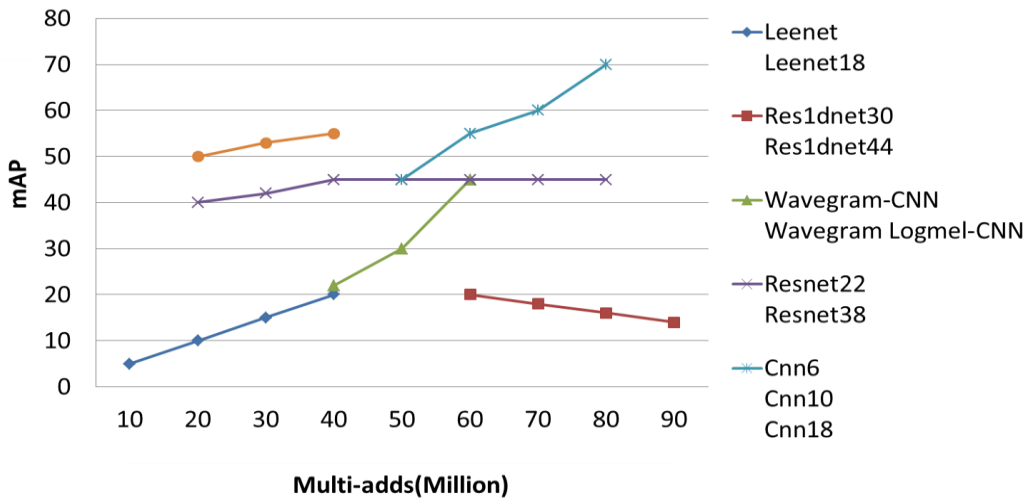

Figure 4. mAP Vs Multi-adds (Million) in audio Tagging

5. Conclusion

In this paper, a novel methodology that uses audio pattern recognition incorporated in an Audio Set is used as a pre-trained Audio Neural Network. To build the neural networks, a large number of neural networks are examined. A Wavegram-Log mel CNN and a wavegram feature are used in this paper which can archive 0.5 mAP. The system computational complexity is also examined indicating the efficiency of the proposed work. Experimental observations and simulations indicated that this work can be utilized in many audio pattern recognition tasks and proves to be more efficient than other traditional methods. The audio neural networks are found to be very useful when used on small data. This methodology can be extended for identifying audio pattern recognition applications in the future.

References

[1] Verbitskiy, S., Berikov, V., & Vyshegorodtsev, V. (2021). Eranns: Efficient residual audio neural networks for audio pattern recognition. arXiv preprint arXiv:2106.01621.

[2] Adam, E. E. B. (2020). Deep Learning based NLP Techniques In Text to Speech Synthesis for Communication Recognition. Journal of Soft Computing Paradigm (JSCP), 2(04), 209-215.

[3] Xu, K., Zhu, B., Kong, Q., Mi, H., Ding, B., Wang, D., & Wang, H. (2019). General audio tagging with ensembling convolutional neural networks and statistical features. The Journal of the Acoustical Society of America, 145(6), EL521-EL527.

[4] Rodrigo, W. U. D., H. U. W. Ratnayake, and I. A. Premaratne. "Identification of Music Instruments from a Music Audio File." In Proceedings of International Conference on Sustainable Expert Systems: ICSES 2020, vol. 176, p. 335. Springer Nature, 2021.

[5] Dhaya, R. "Efficient Two Stage Identification for Face mask detection using Multiclass Deep Learning Approach." Journal of Ubiquitous Computing and Communication Technologies 3, no. 2 (2021): 107-121.

[6] de Benito-Gorron, D., Lozano-Diez, A., Toledano, D. T., & Gonzalez-Rodriguez, J. (2019). Exploring convolutional, recurrent, and hybrid deep neural networks for speech and music detection in a large audio dataset. EURASIP Journal on Audio, Speech, and Music Processing, 2019(1), 1-18.

[7] Manoharan, S. (2019). A smart image processing algorithm for text recognition, information extraction and vocalization for the visually challenged. Journal of Innovative Image Processing (JIIP), 1(01), 31-38.

[8] Sankar, MS Arun, Tharak Sai Bobba, and PS Sathi Devi. "Stage Audio Classifier Using Artificial Neural Network." In International Conference on Communication, Computing and Electronics Systems, pp. 139-147. Springer, Singapore, 2020.

[9] Nanni, L., Maguolo, G., Brahnam, S., & Paci, M. (2021). An ensemble of convolutional neural networks for audio classification. Applied Sciences, 11(13), 5796.

[10] Chandy, A. (2019). A review on iot based medical imaging technology for healthcare applications. Journal of Innovative Image Processing (JIIP), 1(01), 51-60.

[11] Adapa, S. (2019). Urban sound tagging using convolutional neural networks. arXiv preprint arXiv:1909.12699.

[12] Hamdan, Yasir Babiker. "Construction of Statistical SVM based Recognition Model for Handwritten Character Recognition." Journal of Information Technology 3, no. 02 (2021): 92-107.

[13] Zhu, B., Xu, K., Kong, Q., Wang, H., & Peng, Y. (2020). Audio tagging by cross filtering noisy labels. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 28, 2073-2083.

[14] Duraipandian, M. (2020). Adaptive Algorithms for Signature Wavelet recognition in the Musical Sounds. Journal of Soft Computing Paradigm (JSCP), 2(02), 120-129.

[15] Iqbal, T., Kong, Q., Plumbley, M., & Wang, W. (2018). Stacked convolutional neural networks for general-purpose audio tagging. DCASE2018 Challenge.

[16] Vishva, R., P. Harish Annamalai, K. Raja Raman, B. Vijay, J. Rolant Gini, and M. E. Harikumar. "Automated Industrial Sound Power Alert System." In International Conference on Communication, Computing and Electronics Systems: Proceedings of ICCCES 2020, vol. 733, p. 175. Springer Nature, 2021.

[17] Vinothkanna, M. R. (2019). A secure steganography creation algorithm for multiple file formats. Journal of Innovative Image Processing (JIIP), 1(01), 20-30.

[18] Pamina, J., J. Beschi Raja, S. Sam Peter, S. Soundarya, S. Sathya Bama, and M. S. Sruthi. "Inferring Machine Learning Based Parameter Estimation for Telecom Churn Prediction." In International Conference On Computational Vision and Bio Inspired Computing, pp. 257-267. Springer, Cham, 2019.

[19] Koszewski, D., & Kostek, B. (2020). Musical instrument tagging using data augmentation and effective noisy data processing. Journal of the Audio Engineering Society, 68(1/2), 57-65.

[20] REDDY, M. R. (2020). IoT Based Air And Sound Pollution Monitioring System Using Machine Learning Algorithms. Journal of IoT in Social, Mobile, Analytics, and Cloud, 2(1), 13-25.

[21] Lee, J., & Nam, J. (2017). Multi-level and multi-scale feature aggregation using pretrained convolutional neural networks for music auto-tagging. IEEE signal processing letters, 24(8), 1208-1212.

[22] Narmadha, S., and V. Vijayakumar. "An Effective Imputation Model for Vehicle Traffic Data Using Stacked Denoise Autoencoder." In International Conference On Computational Vision and Bio Inspired Computing, pp. 71-78. Springer, Cham, 2019.

[23] Adam, E. E. B., Babikir, E., & Sathesh, P. (2021). Survey on medical imaging of electrical impedance tomography (eit) by variable current pattern methods. Journal of ISMAC, 3(02), 82-95.

[24] Wang, H. C., Syu, S. W., & Wongchaisuwat, P. (2021). A method of music autotagging based on audio and lyrics. Multimedia Tools and Applications, 80(10), 15511-15539.

[25] Ranganathan, G. (2021). A Study to Find Facts Behind Preprocessing on Deep Learning Algorithms. Journal of Innovative Image Processing (JIIP), 3(01), 66-74.

Author's biography

J. Samuel Manoharan is a professor in the Department of Electronics and Communication Engineering at Sir Isaac Newton College of Engineering and Technology, India. His area of research includes Digital Image and Signal Processing, Data Security and Cryptography, Embedded Systems, Biomedical Instrumentation, Artificial Intelligence, Robotics, Deep Learning, Cognitive Science, Ad-hoc Networks, Artificial Neural Network, Evolutionary Computing, Speech Recognition and Autonomous Systems.