IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Using audio as a basis for recognition method using CNN

Vimal Kumar, Tanuj Bhatt, Sumit Paul, Pooja Rana

Student

Lovely Professional University

Abstract

In the last decade, **Deep learning (DL)** has emerged as the solution to the problems that are easy for a human to understand but very difficult to put in a way so that computers can solve it. Image recognition is one example of such problems. **Convolutional Neural Networks (CNN)** emerge as the best solution to image recognition problems. CNN consists of several convolutional layers and pooling layers. Over the years, many people have experimented with CNN models to use them for audio classification and results have been encouraging. In this paper, we have reviewed such attempts and try to analyze them.

Keyword

CNN, Deep Learning, Audio Classification, Spectrogram

Introduction

Deep Learning started gaining popularity with its applications in real life. CNN has produced better results in the field of image recognition. The results it achieved in the field of image recognition is an example of how powerful it can be. Although, we have large image datasets, but they are not sufficient for efficiently monitoring the wildlife, especially at night, but this can be achieved with an audio event detection device. Which encouraged people to use CNN for audio classification and it has already shown great results **[1][2]**. Because CNN operates on images rather than the audio, the first step is to transform audio data to image datasets. This is achieved by generating spectrograms from the audio dataset. Spectrograms of audio signals of different species show different patterns and CNN can learn these patterns. In the early days of trying to use CNN for audio classification, computing resources and availability of large enough datasets was a concern but today we have large audio-sets to feed deep learning models. After generation, spectrograms are fed to the

CNN model which then generates output. The audio consists of many types of features and it is important to retain as many of them as possible during the process of spectrogram generation. This has been extensively discussed in **[3][4][5]**. This process is computationally heavy. In **[6]**, the authors show how formatting a dataset in a certain way, can make the process more efficient by eliminating the need for processing the clip for its whole duration without affecting the feature extraction accuracy.

Some researchers have tried to find other ways to classify audio because they felt that spectrograms lost many of the audio's important features. Although it is true, there is no denying that CNN is still very effective and produces good results as shown in **[7]**, where authors take pre-trained models that are trained on normal images and try to use them for audio classification.

Different methods used

1. CNN

Classification and recognition are the major fields in deep learning. Distinguishing images, audio from others. Image recognition, object detection are some areas where CNNs are widely used. CNN image classification takes an image as input, processes it, and classifies it under certain categories.

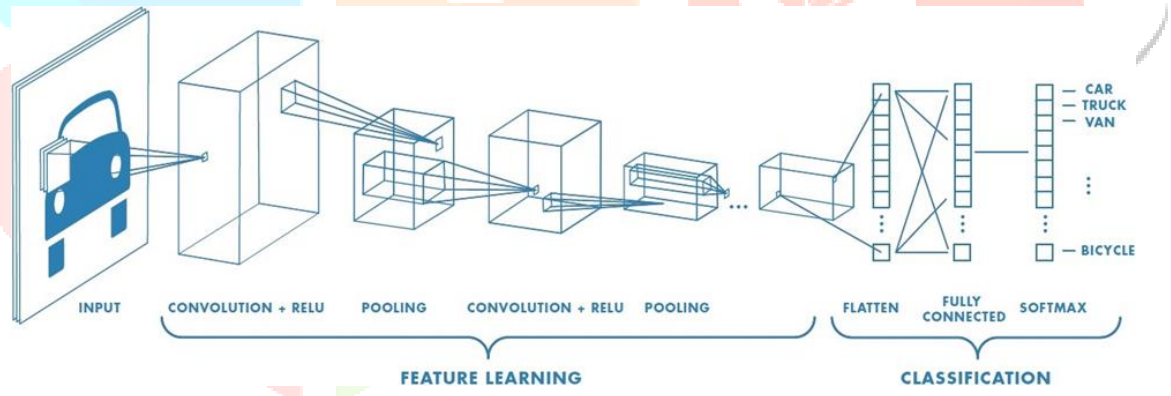

Fig: Neural network with many convolutional layers[https://medium.com/@RaghavPrabhu/understanding-of-convolutional-neural-network-cnn-deep-learning-99760835f148]

It consists of several convolutional layers that serve a different-different purpose from firstly extracting features from the input image as a matrix, pooling layer helps in reducing the number of parameters by reducing the dimensionality of each map, then the matrix is flattened into a vector and fed to a fully connected layer for final output.

Researchers utilized these features to classify the species using audio signals. The idea was to make an automatic bird identification system without physical intervention. This was built as manual detection either physically or with the help of a model is prone to errors as it requires multiple levels of inspections. The data is gathered manually as well as from Xeno-Canto – an online website dedicated to birds sounds from all around the world. Some human and surrounding environment clips are also added as the condition won't be ideal in the testing environment [8].

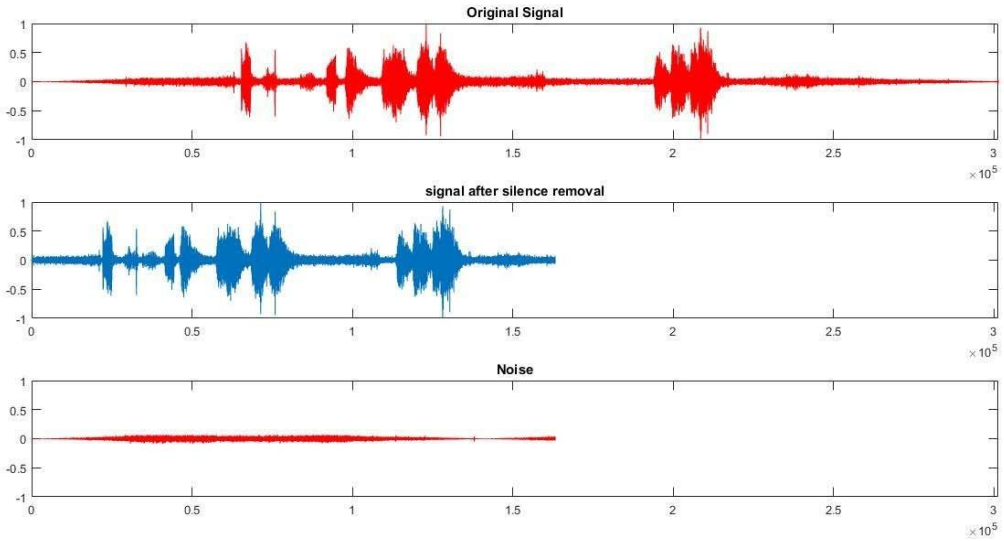

Fig : Silence removal and reconstruction of the audio signal

The data was first passed through an equation as to filter out the relevant frequency. After that, the audio signal was framed for silence removal as the signal was not stationary it has a silent part that needs to be removed. Then the frames are reconstructed by concatenating the frames. The resulted audio signal is free from the silence to some extent. Then the spectrogram was generated from that. As for the modeling, part transfer learning was used in which a pre-trained AlexNet (CNN used in image classification) was used by removing its last fully connected layer and then changing the output according to the dataset.

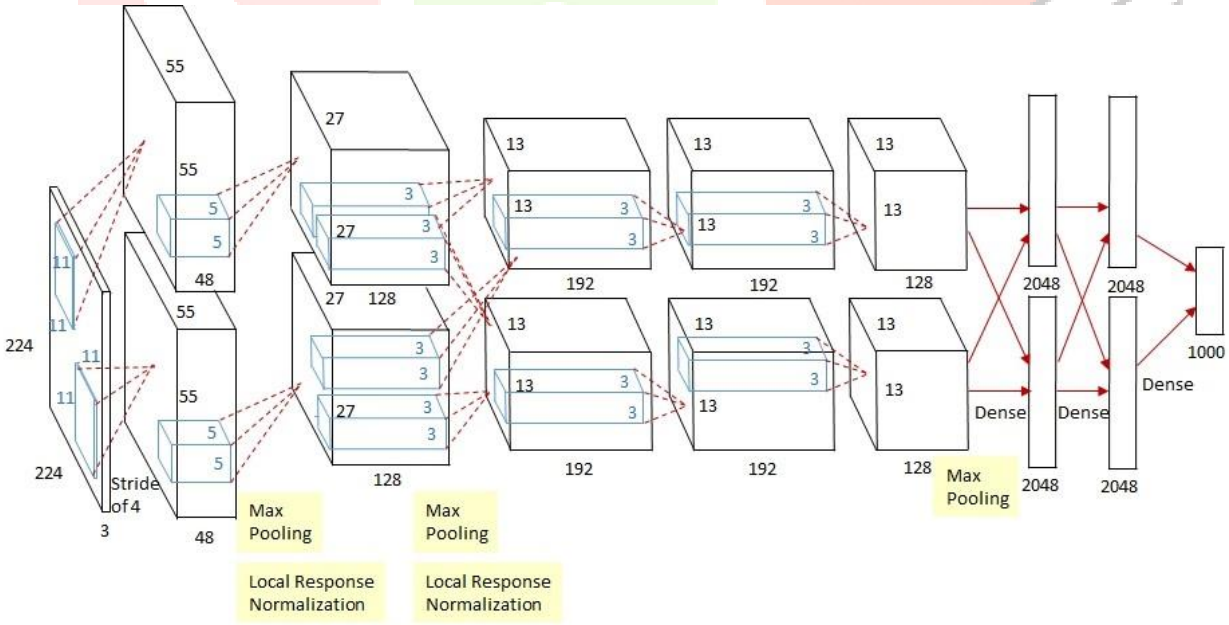

Fig : AlexNet Architecture

The resulting model achieved 97% accuracy but as mentioned by the author these audio signals were from ideal environments and in some cases, noise is removed. To work in real-time and ensure accuracy CNN was retrained with a new dataset having an ideal audio signal as well as collected from the ambient environment and the frequency was chosen randomly as well. After converting to the desired format, the spectrogram is generated and modal was trained and when tested in real-time reached a 91% accuracy.

In this paper, four different species of birds were identified. The model was trained with the audio signals of birds in their natural habitat with other sounds. The outputs were considered in different epochs and data split resulting in the highest of 97% accuracy. That can be further increased by fine-tuning the performance parameters.

Although CNN alone provided very well accuracy researchers still want some best and efficient method so they included different machine learning techniques with CNN.

Different CNN-based models for improved bird sound classification

Automatic bird classification plays an important role in monitoring and helping to take measures in protecting biodiversity. With the help of deep learning, audio classification has advanced very much. In these different previous papers and models are considered and selectively fused for improving the birds' sound classification. Data used was from a public dataset CLO-43DS.

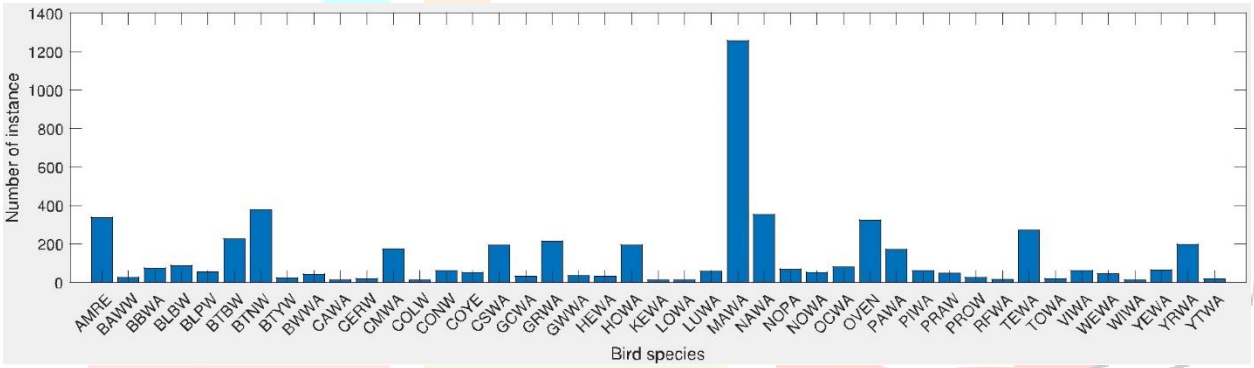

Fig: The number of instance for all bird species in the CLO-43DS dataset. Here, x-axis denotes the abbreviations of common names of those 43 bird species, which can be found in [27].

The clips are recorded in different conditions and with different recording devices. For feature extraction, three different **time-frequency representations** (TFRs) are used [9].

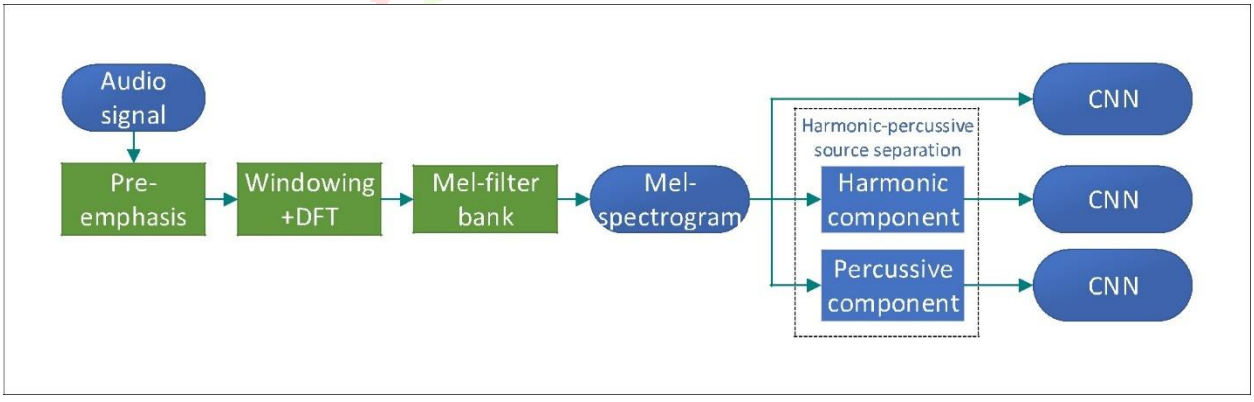

Fig : The time-frequency representation generation procedure. Here, Mel-spectrogram is divided into harmonic and percussive components using Harmonic-percussive source separation. DFT denotes discrete Fourier transform.

For the learning part, VGG style network model is trained with Adam optimizer. Other than this a SubSpectralNet is also used to train features. For further improvement over various CNN-based models, we use class-based late fusion.

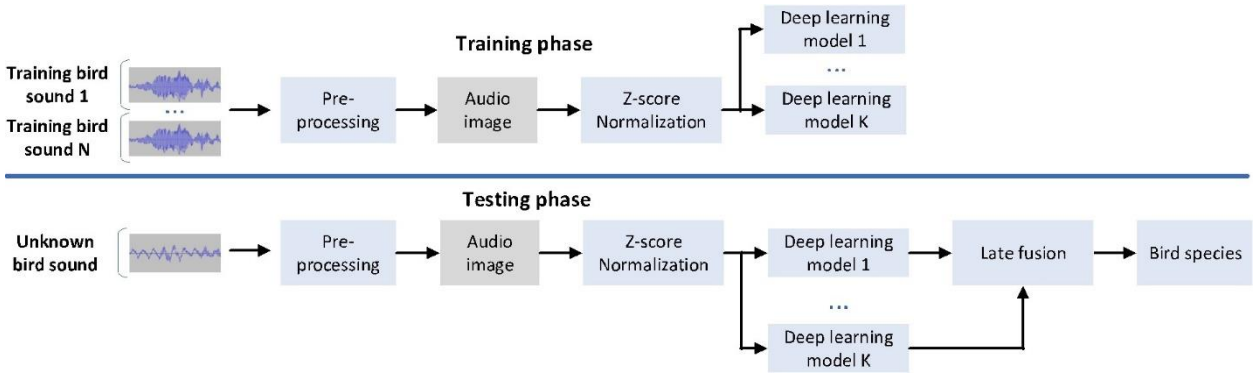

Fig: The flow diagram that is proposed. In here K denotes number of CNN-based models that are used for fusion.

Using the above two architectures multi variate varying length acoustic data, repeating signals achieved higher accuracy (91.2%) than resizing spectrograms (88.5%). For class-based late fusion of different inputs to the same CNNs accuracy is 84.76% which is higher than 83.27%.

In this the fusion of CNN-based models using three types of TFRs for classifying 43 bird species. For those three TFRs, Mel-spectrogram, Harmonic-component-based spectrogram, and Percussive-component-based spectrogram are used to capture different acoustic patterns for the same audio file. From these spectrograms, the VGG style model was trained to classify bird species. Since class-based late fusion for that different CNN were trained that made the whole framework less efficient.

a. CNN with SVM [10]

Pattern classification done using deep networks performs well as compared to other networks. In audio classification, some weakness arises when the dataset that we are training on substantial similarity in different classes. While CNN works well in animal audio classification, CNN shows a weakness if the animal sounds are similar. In this, a novel approach is proposed in which multiple CNNs are trained on fewer levels generating features for a particular class and features are extracted and merged into a combined CNN with SVM for classification.

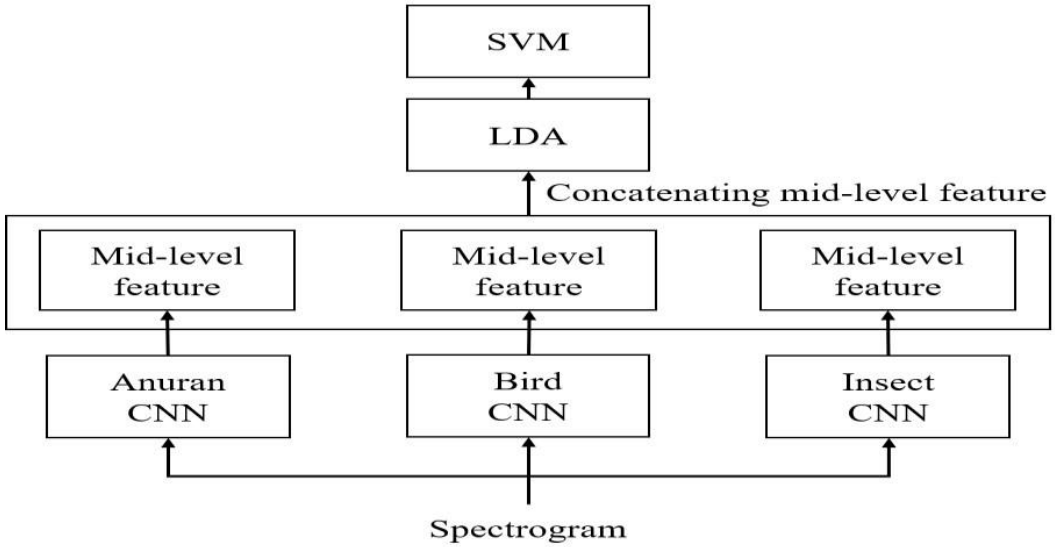

Fig: Proposed approach composed of CNNs and SVM

Data was taken from pretrained classes of anurans, birds, and insects from previous papers. The anuran dataset was manually collected while the other two were taken from **ebird.org** and Korea Wild Animal Sound Dictionary released by the National Institute of Biological Resources respectively. In the proposed method pre-trained CNN are used for the three classes for feature extraction. The final features that are extracted are then fed to LDA for dimension reduction and then to SVM for classification.

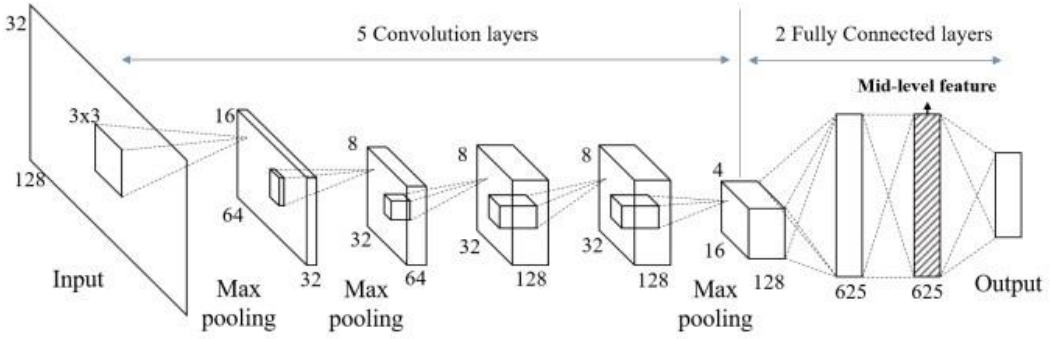

Fig : CNN architecture for extracting mid-level feature; the basic structure is from AlexNet (modified for this purpose)

The overall accuracy that was achieved with other baseline cases is 97.18% and the minimum class accuracy achieved was 70.97%.

A novel approach was proposed for classifying animal sounds. The database was established from real-time and from websites. The method was compared with two different CNN architecture and it outperformed both of them in overall accuracy.

b. CNN with multiple kernel learning [11]

In this the authors instead of relying on a single method for the classification they used deep neural features of both visual and audio datasets based on multiple kernel learning. Dataset was used from two different sources CUB-200-2011 dataset that contains 200 species of birds and an audio dataset from sharing bird audio dataset Xeno-Canto.

For feature extraction, they used a fine-tuned CNN. CaffeNet which is a conventional CNN for large-scale image classification was used. The first few layers of the pre-trained model were used and the latter was discarded to mold the network according to the requirement. After extracting the features are used to create multimodal features dataset by matching the above datasets with their corresponding labels and trained using multiple kernel learning.

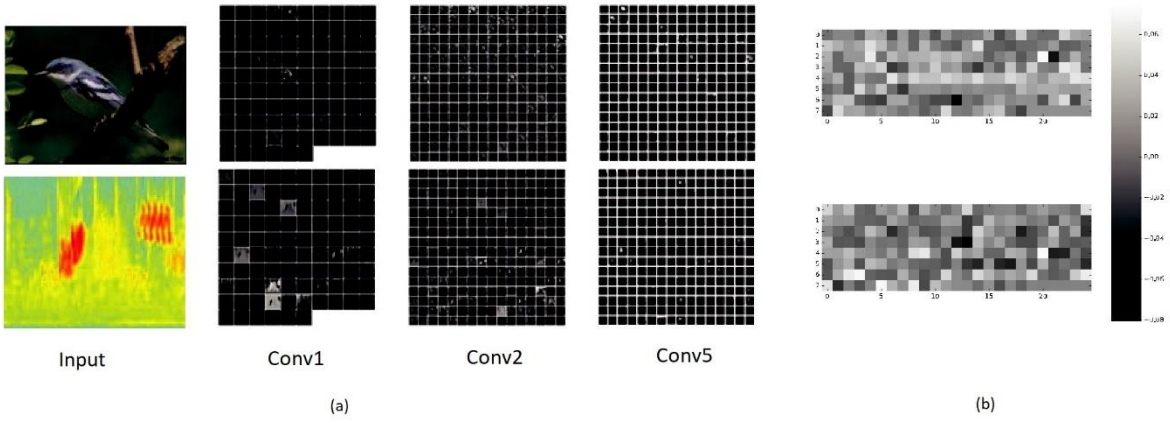

Fig : (a) shows the features from different CNN layers of image (top) and audio (bottom) , and (b) shows the last FC layer that are used in MKL

The result was compared to the other single modality models (trained on a single dataset) and other feature combination methods for birds' classification. The first evaluated single kernel SVM on both the features is compared to a single kernel classifier (CNN+SVM) trained on combined features. The accuracy is increased by a large margin with the second approach.

The combined feature trained on CNN+SVM tends to have higher accuracy as compared to a single feature approach. Also, CNN combined with two kernel functions (RBF and Polynomial) gave higher accuracy from other kernels.

c. CNN-Support vector system [12]

This paper is about the Audio Event Detection (AED) and Acoustic Scene Classification (ASC), intending to understand the environment and detect events and anomalies, are among the growing topics in the research community. There are several challenges to be faced in AED and ASC that include: Lack of a basic set of sound units such as phonemes and words, presence of constant noises in the environment, the rare occurrence of some events, and unexpected events.

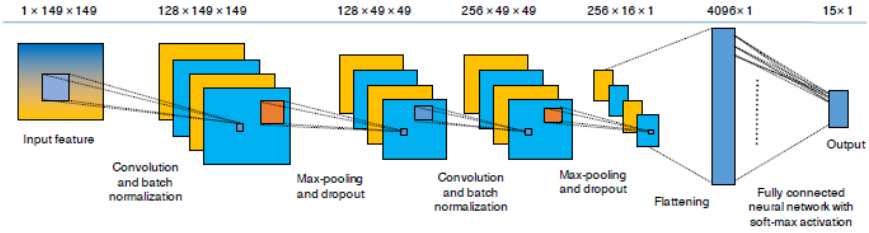

Fig : The flattening layer activations are used to form the purposed CNN-SuperVectors which than sent to softmax layers.

The authors researched various auditory and spectrogram image features for Acoustic Scene Classification using CNN architecture. The individual systems score provides a 7% relative improvement in overall accuracy compared to the CNN system.

The authors have studied the performance of the CNN-based audio classification system using different auditory and spectrogram image features. The authors, noting the benefits of the GMM-SVPLDA system, used the high-dimensional CNN Supervector by the combination of the output of the final layer of CNN functions and used them as features in the PLDA classifier. The evaluation results in the DCASE 2016 Sound Classification Database demonstrate the effectiveness of the CNN-SV approach compared to sole CNN and GMM-SV systems.

2. Deep CNN [13]

In this paper, recognizing speech emotions (SER) refers to the acceptance of the speaker's emotional state by analyzing its speech. The SER can be used to extract useful semantics from speech, as well as to improve the performance of the speech recognition system.

The authors aim to create an SER based on existing emotional speech data sets and later develop an SER model using telephone speech data. The Berlin database was used to evaluate the effectiveness of the proposed SER program. The database contains explicit speech data from four different users. All audio files are defined using one of the seven different senses including neutrality, fear, anger, joy, sadness, disgust, and boredom.

The proposed framework seeks to implement visual learning program schemes from the speech. This mode learning feature paradigm exceeds the traditional feature engineering pipeline. Powerful and discriminatory features are studied in spectrograms that automatically form the basis of SER.

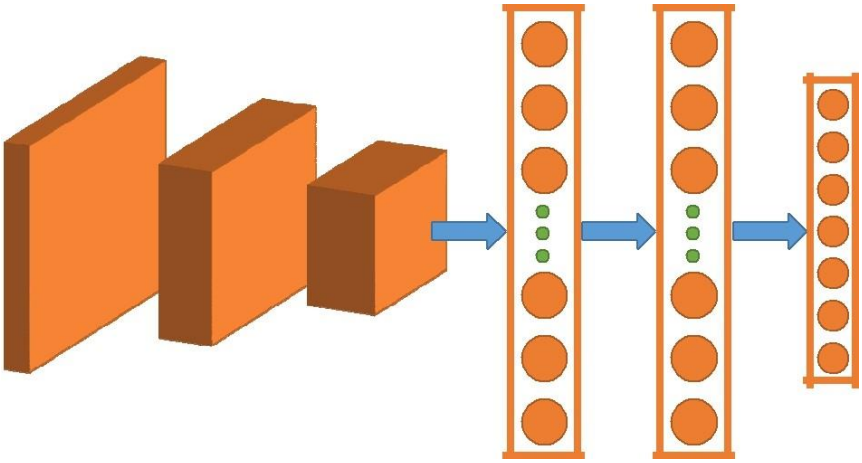

Fig : Proposed CNN architecture for SER using spectrograms

Fresh Trained CNN based SER

The model was able to accurately predict more than 50% of emotions such as anger, boredom, disgust, and sadness. However, predicting the performance of fear, excitement, and neutral feelings was less than 50%. Feelings of fear are often confused with anger, disgust, and excitement. Although the amount of confusion about 19% in each case is lower than the correct predictions that were 25.33%.

Fine-tuned CNN based SER

The confusion matrix was acquired by Alex Net's well-designed model for emotion recognition. This well-designed fine-tuned model improves predictability in the event of anger, neutrality, and feelings of sadness. However, the performance of the forecast in the event of four other emotions decreased. In this case, the authors used a newly trained CNN model because of its less complexity and better performance.

The results reveal that the proposed framework can accurately predict multiple emotions by producing more reliable results in more than 50% of the time. When there are fear and happy emotions, performance is acceptable but not very good because some spectrograms are confused by other emotions. The average prediction for anger, disgust, and feelings of sadness is more than 0.68, while loneliness, fear, happiness, and neutrality are 0.48, 0.33, 0.35, and 0.44, respectively. In all cases, the mean prediction for all emotions was greater than for other emotions. The authors attempt to solve the SER problem using a feature learning program based on deep convolutional neural networks. The speech signal is represented as spectrograms that act as an input to deep CNNs. The CNN model with three flexible and three fully connected layers releases features from these spectrograms and predicts the release of seven sensory categories. The CNN model with three flexible and three fully connected layers releases features from these spectrograms and predicts the release of seven sensory categories. In this regard, two separate test sets were developed. The authors plan to use a lot of data on complex models to improve SER performance continuously.

3. Adaptive multiscale detection of acoustic events [14]

This paper introduced a novel method to detects the acoustic event using the **Adaptive Multi-scale Detection of Acoustic Events (AdaMD)**. This model differs from the above-mentioned models in the sense that they try to classify the audio event while this model tries to detect the time-frame at which the target event occurs which can be useful in events such as monitoring of a specific bird, animal, or any other object. This model builds upon the **Convolutional recurrent neural network (CRNN)**. It introduced a layer called *hourglass* in the CNN part of the CRNN, other concepts

borrowed from the computer vision field, which help to detect a particular event at any position in time. Then in the RNN part of the CRNN, it introduced another unit namely *Gated Recurrent Unit (GRU)*, which helps in processing the temporal information. At every branch, an output is generated and then the branch which produces the worst result is weakened during

optimization. The model was tested on the dataset provided during DCASE competitions and it

outperformed the winner of the DCASE16, DCASE17. Events such as baby cry, gunshot, glass break were used for the evolution of the model. AdaMd with default parameters was competitive with other state-of-the-art models AdaMd with further optimization performed better than those

models.

Method	Baby Cry	Glass Break	Guns Shot	Average
DCASE17 Winner	97.6	99.6	91.6	96.3
R-FCN	97.2	94.6	81.4	90.5
AdaMD-Default	97.1	97.3	91.1	95.5
AdaMD-Balanced	98.2	98.8	92.5	97.6

4. Masked conditional neural networks [15]

Convolutional Neural Networks, originally designed for Image processing, has been used for sound processing for long but because they were not created with sound processing in mind, they often loss invaluable information of acoustic feature and give a simple non-efficient but working model for sound processing. To overcome this problem Fady Medhat et. al. proposed a new neural network architecture which apart from the current state, also consider the output of the previous state which allow it take the whole spectrogram into the account, unlike CNN which at any layer only consider a part of the spectrogram and do care about the previous output. They called it **MASKED CONDITIONAL NEURAL NETWORKS(MCLNN)**. It extends upon the **CONDITIONAL NEURAL NETWORKS(CLNN)**. The CLNN is trained over a window of frames. The CLNN has a hidden layer of e neurons of a vector shape. The input to the CLNN is a window having d frames. The width of the window is $d = 2n + 1$.

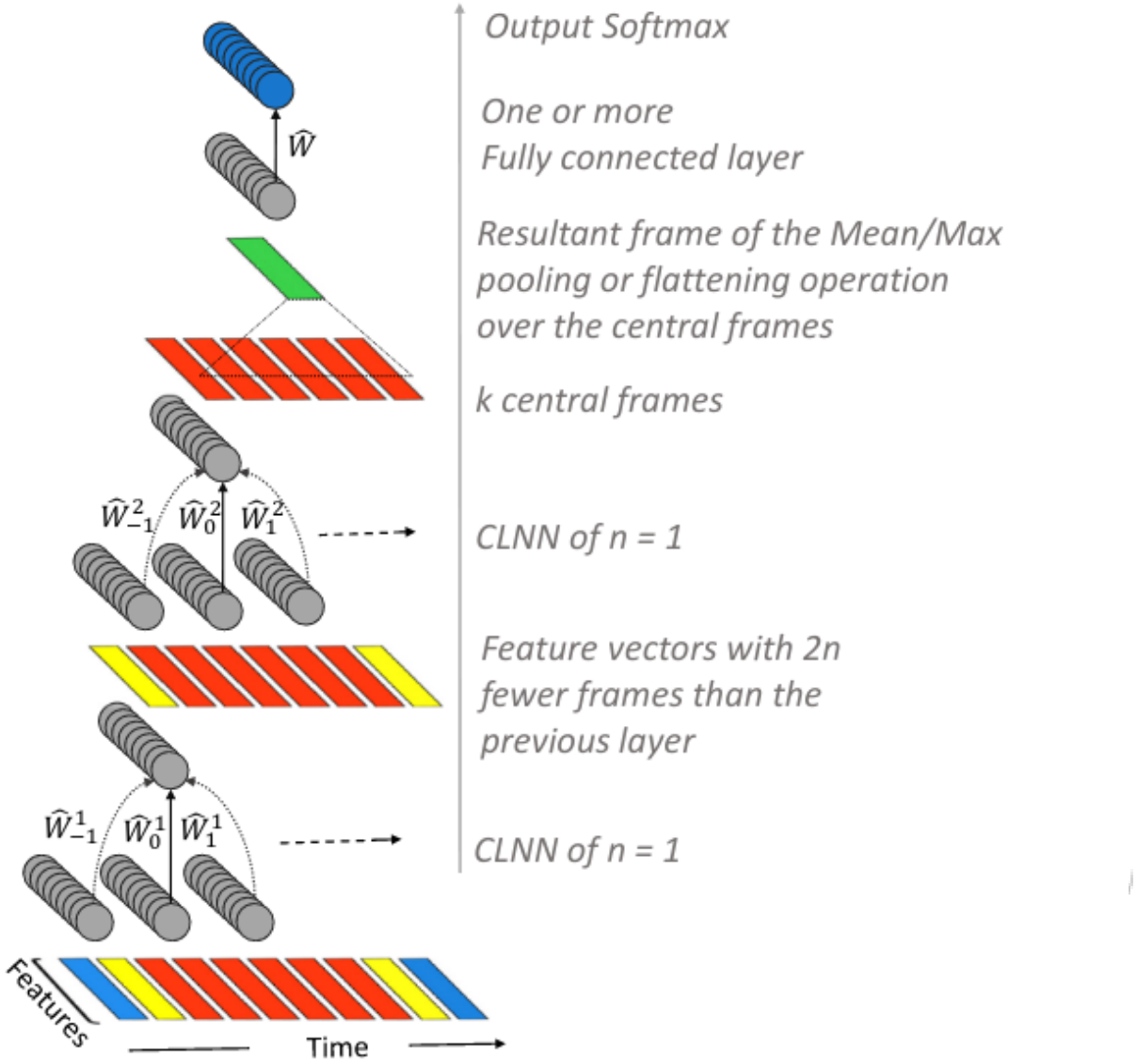

Fig: A CNN with layer n=l [15]

MCLNN adds a filter bank in the CLNN and this filter-bank subdivides the spectrogram into a frequency band. This allows the spectrogram to be frequency shift-invariant by aggregating the energy across several frequency bins and also reduce the complexity of the spectrogram thus making the model lightweight and faster to train. The ballroom dataset was used to evaluate the model which consists of 698 music clips of 30 seconds each. 10-fold cross-validation was used to report accuracies. The model achieved an accuracy of 92.12% without any manual feature selection which was highest in this case. However, if we manual feature selection then the highest accuracy was recorded by a model presented by Peeters et. al. which utilized SVM.

Conclusion

Audio signals has a lot of potential in the field of recognition and it is not bound to the external factors like day and night. There are some constraints like access noise and silence in between but that can be overcome by applying proper pre-processing. CNN has shown a lot of potential in using the extracted features from the audio signals for classification, which can further be increased and also can be branched in different fields like sonography, medical etc.

References

1. N. Takahashi, M. Gygli, B. Pfister, and L. V. Gool, “Deep convolutional neural networks and data augmentation for acoustic event detection,” 2016
2. S. Hershey, S. Chaudhuri, D. P. W. Ellis, J. F. Gemmeke, A. Jansen, C. Moore,M. Plakal, D. Platt, R. A. Saurous, B. Seybold, M. Slaney, R. Weiss, andK. Wilson, “Cnn architectures for large-scale audio classification,” in International Conference on Acoustics, Speech and Signal Processing (ICASSP),2017
3. Huiyu Zhou, A. Sadka, and R. M. Jiang, “Feature extraction for speech and music discrimination,” in 2008 International Workshop on Content-Based Multimedia Indexing, pp. 170–173, 2008
4. F. Alías, J. C. Carrié, and X. Sevillano, “A review of physical and perceptual feature extraction techniques for speech, music and environmental sounds,” Applied Sciences, vol. 6, p. 143, 05 2016
5. N. Patil and M. Nemade, Content-Based Audio Classification and Retrieval Using Segmentation, Feature Extraction and Neural Network Approach, pp. 263–281. 05 2019
6. D. Parikh and S. Sachdev, “Improving the efficiency of spectral features extraction by structuring the audio files,” 2020
7. K. Palanisamy, D. Singhania, and A. Yao, “Rethinking cnn models for audio classification,” 2020
8. B. Chandu, A. Munikoti, K. S. Murthy, G. Murthy V. and C. Nagaraj, "Automated Bird Species Identification using Audio Signal Processing and Neural Networks," 2020 International Conference on Artificial Intelligence and Signal Processing (AISP), Amaravati, India, 2020, pp. 1-5, doi: 10.1109/AISP48273.2020.9073584
9. Xie, Jie & hu, Kai & Zhu, Mingying & Yu, Jinghu & Zhu, Qibing. (2019). Investigation of Different CNN-Based Models for Improved Bird Sound Classification. IEEE Access. 7. 1-1. 10.1109/ACCESS.2019.2957572
10. K. Ko, S. Park and H. Ko, "Convolutional Feature Vectors and Support Vector Machine for Animal Sound Classification," 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Honolulu, HI, 2018, pp. 376-379, doi: 10.1109/EMBC.2018.8512408.
11. N. Bold, C. Zhang and T. Akashi, "Bird Species Classification with Audio-Visual Data using CNN and Multiple Kernel Learning," 2019 International Conference on Cyberworlds (CW), Kyoto, Japan, 2019, pp. 85-88, doi: 10.1109/CW.2019.00022.
12. Hyder, Rakib & Ghaffarzadegan, Shabnam & Feng, Zhe & Hansen, John & Hasan, Taufiq. (2017). Acoustic Scene Classification Using a CNN-SuperVector System Trained with Auditory and Spectrogram Image Features. 3073-3077. 10.21437/Interspeech.2017-431.
13. Badshah, Abdul & Ahmad, Jamil & Rahim, Nasir & Baik, Sung. (2017). Speech Emotion Recognition from Spectrograms with Deep Convolutional Neural Network. 1-5. 10.1109/PlatCon.2017.7883728
14. W. Ding and L. He, "Adaptive Multi-Scale Detection of Acoustic Events," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 28, pp. 294-306, 2020, doi: 10.1109/TASLP.2019.2953350.
15. F. Medhat, D. Chesmore and J. Robinson, "Automatic Classification of Music Genre Using Masked Conditional Neural Networks," 2017 IEEE International Conference on Data Mining (ICDM), New Orleans, LA, 2017, pp. 979-984, doi: 10.1109/ICDM.2017.125.