Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 11, November-2020

A Deep Learning CNN Model for TV Broadcast Audio Classification

Kamatchy. B
Research Scholar
Department of Computer Science and Engineering,
Annamalai University

Dr. P. Dhanalakshmi
Professor
Department of Computer and Information Science
Annamalai University

***Abstract*— In the media, there are many electronic devices used in our day to day life. Television plays a predominant role. A method using deep learning Convolution Neural Network is introduced here to classify TV programs into one of the five categories namely Advertisement, Cartoon, News, Songs and Sports, based on the analysis of audio content. The objective of this work is to develop a CNN architecture to classify the audio segments significantly. The required dataset is created from different channels of Television using TV tuner card and by downloading from you tube channels. The proposed CNN model gives the accuracy of 95 % for TV broadcast audio classification.**

Index Terms*— Keywords: Audio Classification, Spectrograms, Convolutional Neural Network (CNN), and Board cast audio classification.

I. INTRODUCTION

Television is one of the most widely used electronic devices i n our day to day life. In addition to the simple transmission of multimedia content from broadcast service providers to end users, recent advance in content analysis technologies enable television to assume various functionalities, such as scene browsing and summarization. In this respect, the progressing enthusiasm is found in classifying the types of TV programs and videos to favor viewers in different ways.

In this work, a deep learning CNN model for the classification of TV broadcast audio data into one of the five categories namely advertisements, cartoon, news, songs and sports is proposed.

The Convolutional Neural Network (CNN) is a well-known deep learning architecture influenced by the natural visual perception of living beings. A convolutional neural network is a particular type of artificial neural network that uses perceptrons, a machine learning unit algorithm, for supervised learning, for data analysis.

Basic CNN Architecture

CNN's architecture is inspired by the organisation and functionality of the visual cortex and is designed to mimic the pattern of neuron connectivity within the human brain. The CNN architecture consists of a stack of separate layers that transforms the input volume into the output volume via a differentiable function.

A few distinct types s layers commonly used are:

Convolution layers: They are used to preserve the spatial orientation of features in an image.

Pooling Layers: These are used to down-sample an image (shrink it).

Fully Connected layers: The output from the convolutional and pooling layers of a CNN is the image features vector. The purpose of the fully connected layer is to use these features vector for classifying the input images into several classes based on a labelled training dataset.

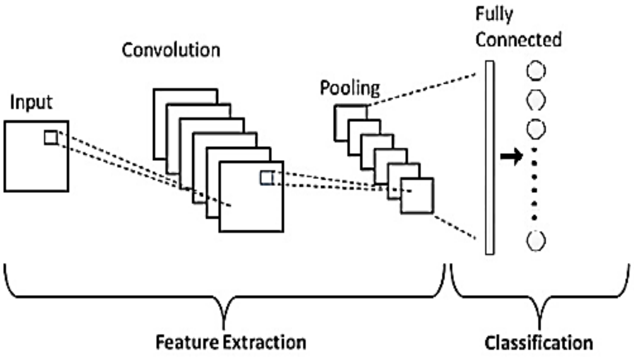


Fig. 1.1 Schematic diagram of Basic Convolution Neural Network

A CNN model is built for classifying the TV broadcast audio. The audio clips are first converted into images using spectrogram which is a visual representation of the spectrum of frequencies of a signal as it varies with time and then used in the model. The spectrogram images of audio clips are then given as input and a series of convolution and pooling operations, followed by a number of fully connected layers are added. The output is one of the predefined five categories: advertisement, cartoon, news, songs, and sports.

II. LITERATURE SURVEY

A. In the year 2018, Boxue Zhang, Qi Zhao ∗, Wenquan Feng, Shuchang Lyu, in their paper, entitled “AlphaMEX: A smarter global pooling method for convolutional neural networks” proposed a novel end-to-end trainable global pooling operator AlphaMEX Global Pool for convolutional neural network. A nonlinear smooth log-mean-exp function is designed, called AlphaMEX, to extract features effectively and make networks smarter. Compared to the original global pooling layer, our proposed method can improve classification accuracy without increasing any layers or too much redundant parameters.The feasibility of the proposed method is demonstrated by experimental results on CIFAR-10/CIFAR100, SVHN and ImageNet. The AlphaMEX-ResNet outperforms original ResNet-110 by 8.3% on CIFAR10+, and the top-1 error rate of AlphaMEX-DenseNet (k = 12) reaches 5.03% which outperforms original DenseNet (k = 12) by 4.0%

B. In the paper "CNN-Based Electronic Camouflage Audio Restoration Mechanism",2018, the authors Zhengyu Shi, Lixian Zheng, Yongquan Wang, Libo Wu proposed a restoration mechanism based on convolution neural network (CNN) for electronic camouflage audio. Since there are certain change rules in the process of converting original audio into electronic camouflage audio and audio is short-time stationary, convolution and nonlinear mapping are performed on the historical sampling acoustic information and restoring factors of the electronic camouflage audio. After companding transformation, the reduction audio is outputted. In the experiment, the voiceprint features comparison, LPC analysis and human ear identity judgement are made between restoring audio and original audio. The results show the validity of the proposed mechanism. It is of great theoretical and practical significance to the restoration of electronic camouflage audio in judicial expertise.

C. In the year 2017, Chien-Yao Wang, Andri Santoso, Seksan Mathulaprangsan, Chin-Chin Chianz, Chung-Hsien Wu and Jia-Ching Wang, in their paper titled "Recognition and retrieval of sound events using sparse coding convolutional neural network" proposed a novel deep convolutional neural network (CNN), called a sparse coding convolutional neural network (SC-CNN), to address the issue of sound event detection and retrieval. Unlike the general framework of CNN, where the feature learning process is conducted hierarchically, the proposed framework models all memorizing processes in the human brain, including encoding, storage, and recollection. Sound data from the RWCP sound scene dataset with additional noise from the NOISEX92 noise dataset are used to equate the output of the proposed system with the state-of-the-art noise baselines. The experimental results showed that the proposed SC-CNN was superior to the state-of-the-art systems for sound event detection and retrieval. In the sound event recognition task, the proposed system obtained an accuracy of 94.6 % 100 % and 100 % under 0db, 10db and clean noise conditions. In the retrieval task, the proposed system improves CNN's overall MAP rate by approximately 6%.

D. In the year 2017, authors Yang Lu, Shujuan Yi, Nianyin Zeng, Yurong Liud, Yong Zhang in their paper titled "Identification of rice diseases using deep convolutional neural networks", proposed an innovative technique to enhance the deep learning ability of CNNs. The proposed CNN model, using a dataset of 500 natural images of diseased and healthy rice leaves and stems taken from experimental rice fields, will effectively identify 10 different rice diseases by image recognition. Under the 10-fold cross-validation strategy, the proposed CNNs-based model achieves 95.48 % accuracy. The application to the rice disease identification shows that the proposed CNNs model can correctly and effectively recognize rice diseases through image recognition. Compared with the other model, the proposed method has a better training performance, faster convergence rate, as well as a better recognition ability than the other model.

E. In the Paper entitled "CNN-based Learnable Gammatone Filterbank and Equal-loudness Normalization for Environmental Sound Classification",2020, the authors Hyunsin Park and Chang D Yoo provides a learnable auditory filterbank with a clear psychophysiological inductive bias in the form of a gammatone filterbank based on a one-dimensional (1D) convolutional neural network and an equivalent loudness prompting normalisation. In the past, a variety of ESC techniques have been proposed based on learnable auditory features obtained by plain 1D convolutions on raw input waveforms to outperform typical handcrafted features such as a mel-frequency filterbank. The large number of parameters involved in convolutions, however, indicates that these models do not generalise better than the model described by a smaller number of parameters which is considered in this paper. For obtaining a time-frequency representation of the raw waveform, a learnable gammatone filterbank layer consisting of 1D kernels represented by a parametric form of the bandpass gammatone filters is suggested here. A normalisation is suggested with learnable parameters that govern the trade-off between energy equalisation and maintenance of structures in the spectro-temporal domain. ESC experiments on the ESC-50 and UrbanSound8K datasets were performed to validate the efficacy of the considered network and the normalisation. Compared to other state-of-the-art networks, the considered network performed better on the two datasets. In addition, an ensemble architecture achieved further performance improvement.

F. "Audio Steganography Based on Iterative Adversarial Attacks Against Convolutional Neural Networks",2020, Junqi Wu, Bolin Chen, Weiqi Luo and Yanmei Fang introduced a novel steganography method based on adversarial examples for digital audio in the time domain The proposed method may start from a flat or even a random embedding cost and then iteratively update the initial costs by leveraging the adversarial attacks until satisfactory security efficiency is obtained, unlike similar methods for image steganography, which are highly dependent on certain current embedding costs. The detailed experimental results show that the system greatly outperforms and produces state-of-the-art outcomes of current non-adaptive and adaptive steganography methods. Moreover, experimental results are provided to investigate why the proposed embedding modifications seem evenly located at all audio segments despite their different content complexities, which is contrary to the content adaptive principle widely employed in modern steganography methods

III. SYSTEM DESIGN

The system design is developed using Python with Keras in Jupyter notebook. Keras is one of the most powerful and easy-to-use Python libraries for developing and evaluating deep learning models; It wraps the efficient numerical computation libraries Theano and TensorFlow.

IV. SYSTEM ARCHITECTURE

The proposed CNN- model for audio classification has the following architecture:

First, a convolution layer is added with 64 filters of size 4x4 and with ReLu activation function to make all negative values to zero. Then Batch normalization is applied. Followed by the Convolution layer, a maxpooling layer with a pool size 2x2 and with a stride of two is added. Furthermore, a second convolutional layer with 64 filters of size 4×4 is added with ReLu activation function. A second 2x2 max-pooling layer is added with a stride of two in both directions. Another convolutional layer with 64 filters of size 4×4 is added with ReLu. Batch normalization is performed. Again, a 2x2 max-pooling layer is added with a stride of two in both directions. A dropout with a probability of 0.2is added to all the convolution layers. Another strategy called global max pooling is used here to down samples the entire feature map to a single value. One of the benefits of applying global max pooling is that in the global max pooling there is no parameter to optimise, so overfitting is prevented at this layer.

Finally, a dense layer consisting of 32 neurons and another dense layer with 5 neurons (one for each class) are added, followed by a sigmoid activation function to classify the audio samples. This model is trained with SGD optimizer. The model iterates over 50 epochs to improve its parameters and to get a highest test accuracy.

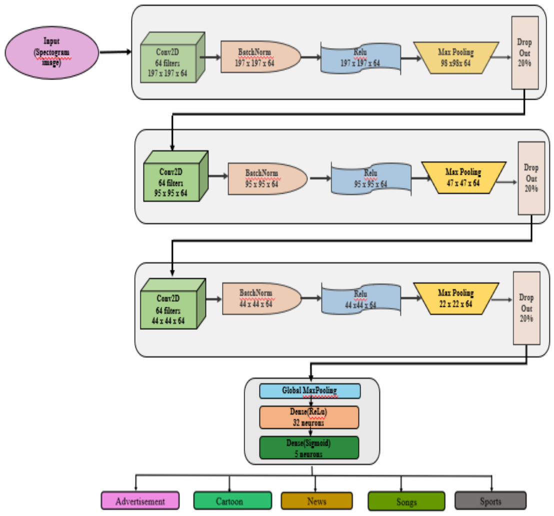


Fig: Block diagram of the proposed CNN model architecture

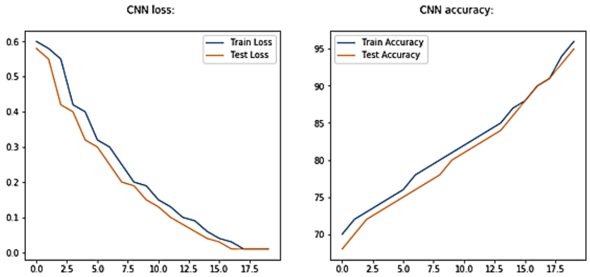
RESULTS & DISCUSSIONS

Dataset

The dataset for the TV broadcast audio classification is collected from different channels using TV tuner card and downloaded from YouTube. Audio clips ranging from 1 to 10 minutes are recorded with a sampling rate of 8kHz for the different categories. 200 clips of advertisement, 200 clips of cartoon, 200 news clips, 200 clips of songs and 200 sports clips are collected. From the dataset ,80% of data are taken for training and 20% of data are taken for testing the model.

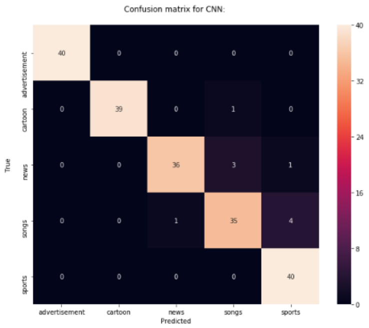
In this model, "accuracy" is taken as classification metrics. Precision, Recall, F-Score are used as performance measures. The loss function -Binary Cross Entropy is used here. Then the Stochastic gradient descent method is used to update the weights of the neural network such that the loss is minimized. The plots of the CNN-model accuracy and loss on the training and validation set is shown below. These are useful to check for overfitting.

A. Performance Measures



The graph showing the learning curves of CNN-Model of Audio Classification

Confusion Matrix: A confusion matrix or an error matrix is used for describing the efficiency of a classifier. It is a specific table that describes the capability of a classifier on a set of data used for testing for which the true values are already known. It contains details of the actual classifications and the predicted classifications performed by the classification system. The confusion matrix displayed below shows how well the CNN-Model can predict against unseen data.



Confusion Matrix of CNN-Model for audio Classification

V.TABULATIONS

Table 5.4 Performance measures of CNN-Model

	Precision (%)	Recall (%)	F-Score (%)
Advertisement	100	100	100
Cartoon	100	97.0	99.0
News	97.0	90.0	94.0
Songs	90.0	88.0	89.0
Sports	89.0	100	94.0
Accuracy			95.0

VI. CONCLUSION

A CNN -Model for audio classification is examined to classify the segmented audio into one of its five predefined categories and the model gives the best accuracy for TV broadcast audio classification.

REFERENCES

[1] S. Veena, Nerisai. M. V Remya. J. V Sai Tejah.S; SOUND Classification System Using Machine Learning Techniques. *International Journal of Engineering Applied Sciences and Technology, 2020 Vol. 5, Issue 1, ISSN No. 2455-2143, Pages 674-678*

[2] Tong He, Zhi Zhang, Hang Zhang, Zhongyue Zhang, Junyuan Xie, Mu Li; Bag of Tricks for Image Classification with Convolutional Neural Networks. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 558-567.*

[3] Agrawal, A., & Mittal, N. (2019). *Using CNN for facial expression recognition: a study of the effects of kernel size and number of filters on accuracy. The Visual Computer. pp* 405–412.

[4] Rehman, A., Naz, S., Razzak, M. I., Akram, F., & Imran, M. (2019). *A Deep Learning-Based Framework for Automatic Brain Tumors Classification Using Transfer Learning. Circuits, Systems, and Signal Processing. pp* 757–775.

[5] Alzubaidi, L., Fadhel, M. A., Oleiwi, S. R., Al-Shamma, O., & Zhang, J. (2019). *DFU_QUTNet: diabetic foot ulcer classification using novel deep convolutional neural network. Multimedia Tools and Applications. Pp* 15655–15677.

[6] Sampath Kumar, A., Erler, R., & Kowerko, D. (2019). *A Real-Time Demo for Acoustic Event Classification in Ambient Assisted Living Contexts. Proceedings of the 27th ACM International Conference on Multimedia - MM '19. pp* 2205–2207.

[7] K. N. Bui, H. Oh and H. Yi, "Traffic Density Classification Using Sound Datasets: An Empirical Study on Traffic Flow at Asymmetric Roads," in *IEEE Access*, vol. 8, pp. 125671-12567.

[8] R. Mars, P. Pratik, S. Nagisetty & C. Lim, "Acoustic Scene Classification From Binaural Signals Using Convolutional Neural Networks", Proceedings of the Detection and Classification of Acoustic Scenes and Events 2019 Workshop (DCASE2019), pages 149–153, New York University, NY, USA, Oct. 2019.

[9] Vafeiadis, A., Votis, K., Giakoumis, D., Tzovaras, D., Chen, L., & Hamzaoui, R. (2020). *Audio content analysis for unobtrusive event detection in smart homes. Engineering Applications of Artificial Intelligence, 89, 103226*.

[10] K. N. Bui, H. Oh and H. Yi, "Traffic Density Classification Using Sound Datasets: An Empirical Study on Traffic Flow at Asymmetric Roads," in *IEEE Access*, vol. 8, pp. 125671-125679, 2020.

[11] T. V. Kumar, R. S. Sundar, T. Purohit and V. Ramasubramanian, "End-to-end audio-scene classification from raw audio: Multi time-frequency resolution CNN architecture for efficient representation learning," *2020 International Conference on Signal Processing and Communications (SPCOM)*, Bangalore, India, 2020, pp. 1-5.

[12] A. Torfi, S. M. Iranmanesh, N. Nasrabadi and J. Dawson, "3D Convolutional Neural Networks for Cross Audio-Visual Matching Recognition," in *IEEE Access, vol. 5, pp. 22081-22091, 2017*.

[13] A. Jansen, J. F. Gemmeke, D. P. W. Ellis, X. Liu, W. Lawrence and D. Freedman, "Large-scale audio event discovery in one million YouTube videos," *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, New Orleans, LA, 2017, pp. 786-790.

[14] J. Wu, B. Chen, W. Luo and Y. Fang, "Audio Steganography Based on Iterative Adversarial Attacks Against Convolutional Neural Networks," in *IEEE Transactions on Information Forensics and Security*, vol. 15, pp. 2282-2294, 2020.

[15] De Freitas, P. V. A., Santos, G. N. P. dos, Busson, A. J. G., Guedes, Á. L. V., & Colcher, S. (2019). *A baseline for NSFW video detection in e-learning environments. Proceedings of the 25th Brazillian Symposium on Multimedia and the Web - WebMedia '19.* Pages 357–360