International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-9 Issue-5, March 2020

An Efficient Model for TV broadcast Audio Classification through InceptionV3 and ResNet50

Kamatchy B., P. Dhanalakshmi

Abstract: ***In the recent advancements of applications, one of the challenging task in many gadgets are incorporated, which is based on audio classification and recognition. A set of emotion detection after post-surgical issues, classification of various voice sequence, classification of random voice data, surveillance and speaker detection audio data act as a crucial input. Most of the audio data is inherent with the environmental noise or instrumental noise. Extracting the unique features from the audio data is very important to determine the speaker effectively. Such kind of a novel idea is evaluated here. The research focus is based on classification of TV broadcast audios in which the type of audio is being class separated through a novel approach. The design evaluates, the five different categories of audio data such as advertisement, news, songs, cartoon and sports from the data collected using the TV tuner card. The proposed design associated with python as a Development environment. The audio samples are converted to images using Spectrogram and then transfer learning is applied on the pretrained models ResNet50 and Inceptionv3 to extract the deep features and to classify the audio data. Inception V3 is compared here with the ResNet50 to get greater accuracy in classification. The pre-trained models are models that was trained on the ImageNet data set for a certain task and are used here to quick train the audio classification model on training set with high accuracy. The proposed model produces accuracy of 94% for Inceptionv3 which gives greater accuracy when compared with the ResNet50 which gives 93%. accuracy.***

Keywords: ***Audio Classification, Spectrograms, Inception modeling, and Board cast audio classification.***

I. INTRODUCTION

Human classify audio signals all the time without cognizant effort. Distinguishing a power of speech on the telephone, telling the modification between a receiver ring and a carillon ring, are chores that are not considered very difficult. Difficulties do arise when the wide-ranging sound is weak or there is a nearby noise or it is comparable to another sound.

There are three main areas of motivation for audio classification research. First, it would be instructive to know how the humans do that and what they do. If the general systems used to classify audio is known, it might be easy to better diagnose and treat auditory ailments. The research that would answer these questions tends to be more psychological and physiological than computational, but the methods used in computer audio classification systems might provide a starting point for human audio classification research. Second, it would be efficient to have a machine that could analyze like a human which can implement the same, how the humans do with sound. In many research focus is on classification of random audio signals to find out the emotions of the patients and post-surgical analysis through the audio variations. In many applications audio signals are impacted more with the noise. The environmental noise is modulated with the original signal can be difficult when making classification. In police department keeping the large set of audio data which tell the information of various recordings, statements, and many secret data is being managed with the server which will be analyzed to recognize the unique voice held in different audios. In communication systems, audio data plays a major role. Audio data is converted into spectrogram images in many applications to handle the pattern-based recognition principle.

A. Categories of Audio Broadcast Classification

Audio broadcast data is a combination of variety of different speakers with different voice tones contains the background tone and music etc. The different types of TV broadcast audio can be classified as advertisements, cartoons, news, songs, and sports. Each kind of signal have their own identity through a type of pitch, segmented audio part, the Ceptral componenet etc. The acoustic feature representing the audio information is extracted from the audio data. This unique components act as a factor for classification key. In the proposed research work around 200 data samples are collected from various TV channels. Each category consists of 200 samples. The audio data contains the cartoons, sports, advertisements, news and songs. . The pre-trained model of Convolution Neural Network, RestNet50 and Inceptionv3 are evaluated here for comparisons.

The quality of audio is based on the broadcast bandwidth in which it travels. Normally the fact is Digital audio broadcast is not very much efficient compared with the analogue type of audio transmission in older days but still the recent advancements of digital broadcasting encapsulates the benefits of reduced signal interferences. Because of the large set of massive devices nowadays sending the audio broadcast data in the free space is emulated with improved quality.

Revised Manuscript Received on March 30, 2020.
***** Correspondence Author
Kamatchy B. *, Research Scholar, Department of Computer and Information Science, Annamalai University, Indian. E-mail: kamatchi6282@gmail.com
Dr. P. Dhanalakshmi, Professor, Department of Computer Science and Engineering, Annamalai University, India. E-mail: abidhana01@gmail.com

The advantageous software modules like neural networks, trained googleNet, AlexNet and so many Deep learning machines provides the benefit of implementing the audio recognition systems, Broadcast type detection systems, News speaker identification and so many categories of Broadcast Audio classification Models.

II. LITERATURE SURVEY

A. Irene KMart´ın-Morat´o, Maximo Cobus, and Francesc J. Ferri, in the year 2018 found in his research work that 'an efficient Adaptive MidTerm Symbols for Vigorous Audio Occurrence Classification', in which the use of an substitute event depiction based on non-linear time stabilization prior to the taking out of mid-term statistics. The short-term structures are converted into a new fixed-length depiction that cogitates unbroken distance subsampling over a well-defined feature interstellar in contrast to the typical short-term sequential framing. The grades show that the use of distance-based consistency windows affords an enhanced numerical portrayal of the event full-bodied to errors in the event breakdown stage lower than noisy conditions.

B. "Design of an Argumentative Refinement Method for Audio Classification, *LiangoGoa, *Haibo mi evaluated in the year 2019 recently in his work that, various audio classification task aims to differentiate between unlike audio signal types. In this task, deep neural networks have accomplished better concert than the outmoded shallow architecture-based machine-learning method. Nevertheless, deep neural webs often involve huge computational and stowing necessities that hamper the distribution in surrounded campaigns. In his research paper, he declared a distillation method which transferals acquaintance from well-trained networks to a trivial network, and the process can compress model size while enlightening audio classification precision. The assistances of the proposed method are two folds: a multi-level feature refinement method was anticipated and combative learning policy was engaged to improve the knowledge transfer. The wide-ranging experiments are showed on three audio classification tasks, audio scene classification, all-purpose audio tagging, in accumulation speech expertise appreciation. The investigational results determine that: the small web can provide better concert while achieves the planned amount of floating-point procedures per second solidity ratio of 76:1 and constraints firmness ratio of 3:1.

C. In the year 2016, author named xXiaodan *Lin, a-Jingxian Liu*, and Xiaangui Kang* research work entitled , Audio recapture uncovering with convolutional neural networks and proposes, The process works for slight audio clips of 2 seconds' time, where the state of the art may fail with such insignificant audio clips. Investigational results determine that the anticipated linkage earnings high exposure truthfulness with each ENF choral element denoted as a single-channel input. The routine can be supplementary improved by a united input depiction which integrates both the ultimate ENF and its harmonics. Conjunction goods of the link and the effect of using scrutiny opening with various sizes are also studied. Concert evaluation against the backing tensor contraption demonstrates the specialist of using CNN for the task of auditory recapture recognition. Moreover, conception of the transitional feature maps affords some intuition into what the deep neural webs actually learn and how they kind decisions.

D. In the year 2016, Authors named *Lakshmish Kaushik, Abhijeet ASSangwan and John HL*. Hansen Automatic Soppiness Detection in Natural Audio, in the study, they shown that this baseline organization is sub-optimal for audio sentiment abstraction. Otherwise, new construction using keyword spotting (KWS) is projected for soppiness detection. In the new manner, a text-based sentiment classifier is operated to repeatedly regulate the most useful and discriminative sentiment-bearing keyword terms, which are then used as a span list for KWS. In order to attain a compacted yet discriminative soppiness term list, repeated or recursive feature optimization for determined entropy soppiness model is proposed to condense model density while maintaining operative arrangement precision. A new hybrid ME-KWS joint keep score methodology is developed to model both text and auditory based bounds in a single assimilated formulation. For valuation, two new catalogs are settled for audio constructed gush detection, namely, YouTube sentiment catalog and another lately developed amount called UT-Opinion Judgment audio record. These catalogs grasp lifelike opinionated audio together in real-world surroundings. The anticipated solution is gauged on audial attained from videos in youtube.com and Opinion body. Our investigational domino effect show that the anticipated KWS based system ominously outclasses the out-of-date ASR planning in perceiving soppiness for inspiring useful tasks.

E. H.Annamaria Mesaros*, Toni Heiittola , Emsmanouil Benetos , Peter Foster, Mathieu Lagrangee, Tuomass Virtanen*, Year 2016 , the research work entitles an efficient Detection and Classification of Acoustic Scenes and In this paper, they exploysion on the responsibilities and conclusions of the DCASE-*2016 challenge. The encounter covered four tasks: audile scene sorting, sound incident recognition in manmade audio, wide-ranging event uncovering in real-life audio, and inland audio tagging. We contemporaneous each task in element and consider the give in to structures in terms of proposal and performance. We observe the development of deep scholarship as the most popular arrangement method, exchanging the old styles based on Gaussian concoction models and support vector machines. By disparity, unique representations have not improved greatly during the course of the years, as mel-frequency-based depictions preponderate in all tasks. The datasets fashioned for and used in DCASE*2016 are freely offered and are a treasured source for added research.

F. In the year 2017, author named HuypPhan, Lars Heertel, Marcoo Maass*, Philipp Koch, Radoslaw Mazur, Alfred Mertins ,development of Improved Audio Scene Ordering Constructed on Label-Tree Embedding and Convolutional Neural Networks We show that the recipe of several features is important to acquire good routine. While be in the region of label-tree entrenching images over while yields good routine, we maintain that middling pooling keeps an essential deficiency. We on the other hand propose an enhanced arrangement structure to circumvent this control.

We aim at robotically erudition public models that are advantageous for the ordering task from these imaginings using meek but handmade convolutional neural networks. The accomplished webs are then engaged as a chin extractor that contests the learned prototypes diagonally a label-tree surrounding image and yield the thoroughgoing similar scores as geographies for classification. Since auditory scenes revelation rich contented, template scholarship and alike on low-level landscapes would be ineffective. With label-tree surrounding landscapes, we have quantized and condensed the low-level features into the prospects of the Meta programmers, on which the prototype scholarship and same are competent.

III. SYSTEM DESIGN

The system design is developed using python Anaconda Tool & Jupiter note book. The system incorporates pre-trained models RestNet50 and Inceptionv3 with the input samples of TV broadcast audio which are preprocessed and feature extracted. Python is one of the efficient open source language. The tool is user friendly and associated with easy configuration of libraries for plotting the data, machine learning toolboxes, other benefits include:

1) Python is a tool which can be used to develop real time prototypes, and it is so laid-back to work with and delivered.

2) The field of automation, data mining of large sets, and bigdata handling platforms rely on Python language. It is the superlative linguistic tool to work with for general purpose chore.

IV. SYSTEM ARCHITECTURE

A. Transfer Learning

Define abbreviations and acronyms the first time they are used in the text, even after they have been defined in the abstract. Abbreviations such as IEEE, SI, MKS, CGS, sc, dc, and rms do not have to be defined. Do not use abbreviations in the title or heads unless they are unavoidable.

Transfer learning via feature extraction: When performing feature extraction, the pre-trained network is treated as an arbitrary feature extractor, which allows the input image to propagate forward, stopped at pre-specified layer, and the outputs of that layer is taken as the features.

Transfer learning via fine –tuning Fine-tuning, on the other hand, requires that the model architecture itself is updated by removing the previous fully-connected layer heads, which provides new, freshly initialized ones, and then the new FC layers are trained to predict the input classes.

B. RestNet50

ResNet,(Residual Networks) is a classic neural network that is used for many computer vision tasks as backbone. ResNet's fundamental breakthrough was that it allows successful training of extremely deep neural networks with 150+layers.. ResNets help to eliminate the varnishing and exploding gradient problems.

Vanishing gradients: The gradient becomes very small, that even a very big change in the input will not affect the output as desired.

Exploding gradients: The gradient becomes exponentially big that the algorithm can no longer be used to train the model.

For the tv broadcast audio classification task, 6 hidden (dense) layers are added with 1024, 512, 256, 128, 64, 32 neurons respectively as shown in the fig 1. Here the audio segment should be classified into one of five predefined classes. So the output layer is added with 5 neurons which correspond to the number of categories in which we need to classify the input image.

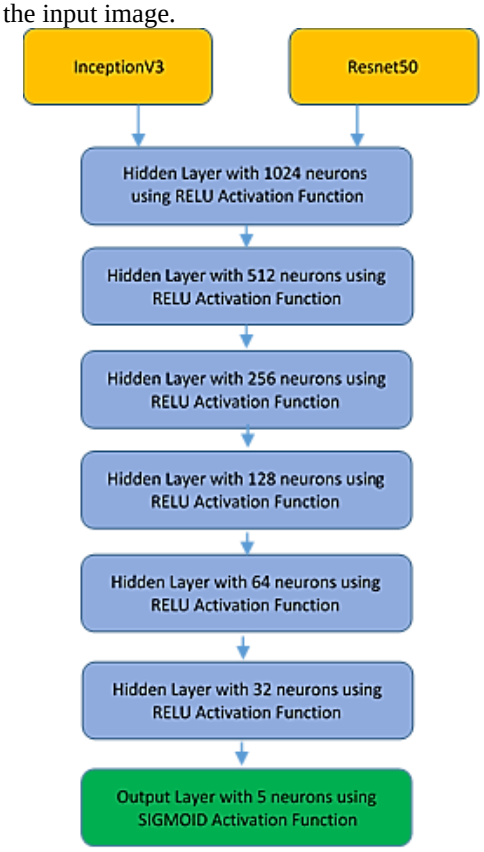

Fig. 1. System architecture of proposed Inception and Resnet Models

C. Inception V3 Model

The proposed novel work is focused on creating an efficient model which is compared with the pre-trained RestNet50 The model Inception V3 is most probably used for, the image recognition model .It consists of two efficient part of design (1) The feature extraction part with a convolutional neural network and another is the (2) Classification part with fully-connected and softmax layers. All the audio clips are transformed into spectrograms firstly and the pretrained model- the InceptionV3, is then re-trained with the spectrograms.

Then, six fully connected layers are added to the end of the Inception modules as shown in the figure to utilize the pretrained model and finetune the parameters for TV broadcast audio classification task. It is only the final layers of the network, the layers that learn to identify classes specific to the task that need training. So finally, an output layer with 5 neurons with sigmoid activation function is added as a classifier, outputs a probability for each class, and the one with the highest probability was chosen as the predicted class.

D. Design Methodology Adapted

(1) In the way of layman point of view, a Residual network is a kind of deep network in which the modules basically contains multiple convolutional filters which able to dig deeper with the same images, same type of input with same pooling strategy henceforth the results are concatenated each other at the end. The design perceptions or the methodology is adopted here with the help of converting the input audio data into spectrogram images.

(2) Spectrogram is a unique method of representing the raw audio data in the form of frequency spectrums. The spectrums varies with time. Each audio component have the unique spectrum peaks which will be helpful for getting the data uniquely. Spectrograms are sometimes called as a name sonograms or voiceprints.

V. RESULTS & DISCUSSIONS

A. Dataset

The TV broadcast audio classification is done with the dataset collected by using TV tuner card from different channels.200 clips of audio samples are taken for each category in the dataset. The research methodology utilizes the spectrogram of images in which 80% of the input is used for testing and 20% of the data is used for training.

B. Experimental Results

With various versions of training aspects it is found that Inception V3 model is benefited in computation time and scaling the dimensionality aspects. To compensate the best perception model for TV broadcast classification, the proposed idea is emulated better which is clearly depicted in the results.

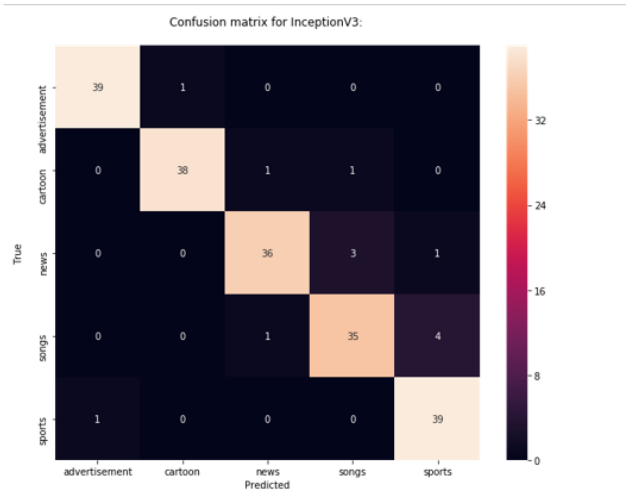

Fig 2 Confusion Matrix showing Inception V3 results on audio classification

It is being clearly depicted from the above image that classification of advertisements, cartoons, news predictions, songs and sports are classified and the relative score is displayed diagonally

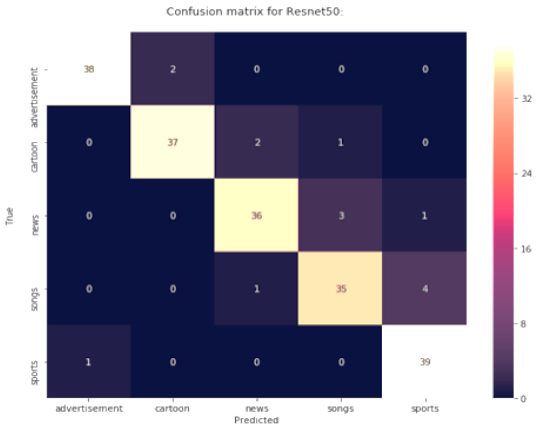

Fig 3. Confusion Matrix showing RestNet50 results on audio classification

In the above Fig 53 the confusion matrix showing RestNet50 classifying the perception of given broadcast audio data and clearly depicts the category which is displayed diagonally.

C. Performance Measures

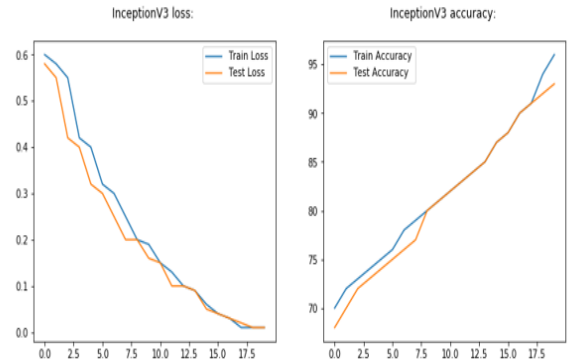

Fig 4. The graph showing Loss and accuracy of Inception-V3 Model of audio classification.

The above graph shows the performance measure in terms of Loss and accuracy of the proposed model. The graph depicts the training as well as the testing Loss and accuracy comparatively.

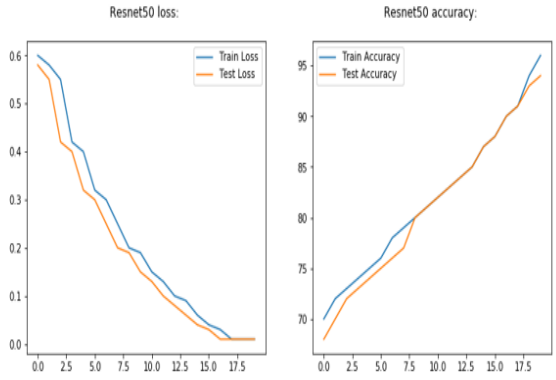

Fig 5. The graph showing Loss and accuracy of RestNet50 model of audio classification

The above graph shows the performance measure of RestNet50 in terms of Loss and accuracy of the training set and testing set.

Table 1.1 Tabulation showing the performance results on Inception V3 model

Classification Results on ResNet50			
CATEGORY	PRECISION	RECALL	F1-SCORE
Advertisements	0.97	0.35	0.96
Cartoon	0.95	0.33	0.94
News	0.92	0.3	0.91
Songs	0.90	0.38	0.89
Sports	0.89	0.37	0.93
Accuracy			0.93

Table 1.2 Tabulation showing the performance results on ResNet50

Classification Results on Inception V3			
CATEGORY	PRECISION	RECALL	F1-SCORE
Advertisements	0.97	0.97	0.97
Cartoon	0.97	0.95	0.96
News	0.95	0.9	0.92
Songs	0.9	0.88	0.89
Sports	0.89	0.97	0.93
Accuracy			0.94

VI. CONCLUSION

The perceptive analysis ended with the classification results showing that the InceptionV3 Model producing best accuracy comparing with the existing pre-trained model of ResNet50. The challenge on getting the deep feature extraction of the audio data is hopefully concluded with the better accuracy and less in loss. The Computational time persists in the proposed process also being achieved. The proposed model is accurate and good in performance in class separating the audio sequences of TV broadcast spectrums.

The research can be further extended to recognize the type of speaker, gender identification of various broad cast data and the theme classification of the particular broadcast event etc

REFERENCES

1. P. Dhanalakshmi, S. Pazhanirajan, "MRI Classification of Parkinson's Disease Using SVM and Texture Features", in Advances in Intelligent Systems and Computing,
2. Springer 380 (2), Springer, August 2015.
3. R Visalakshi, P Dhanalakshmi, S Palanivel , "Analysis of throat microphone using MFCC features for speaker recognition", in Intelligence, Cyber Security and Computational Models, pages 35-41, Springer 2016.
4. R Thiruvengatanadhan, P Dhanalakshmi, PS Kumar, "Speech/music classification using SVM", in International Journal of Computer Applications 65 (6), Published by Foundation of Computer Science, Jan 2013.
5. Zf. Qawaqneh, AA. Malloasuh, B. D. Baarkanaa, "Deep neural network for Speaker age classification", in Ieee knowledge based systems, 2015.
6. CS. Oi, K. PSeng, L.-M. -Ang, LWchhhew, "Design of audio emotion recognition", in IEEE 2014 conference summit.
7. SO. Poria, E. Cambria EP, Ak. Hussain, G.-B. Huanga, IEEE 2015 Neural network summit, Multi-modal intelligent framework on Neural network. 2015
8. TFLi, S.-C. Changd, design of Speech recognition of syllables Ieee linguistic conference 2015
9. MAkagi, X.v Han, Rr. Elbaarougy, Y. Hamaada*, entitled Design of speech-tospeech translation: speech recognition Strategy in IEEE conference summit 2015
10. IEEE journal 2015classifiying the DA/MCI patients Dept of Biomedical. Fc. Lic, L. Tran, K.-H. Thung*, S.D. Shen, JLi
11. Authors named YZhu, Sf. Lucey*, CSC-Convolutional sparse coding for trajectory reconstruction, IEEE Transactions on Pattern Analysis and Machine Intelligence 2014
12. Authors Yy. LeCun, Y. Beengio, G. Hiinto*n, Deep learning, Natural Dataset
13. Kd. Simonyan, AkZisserman, DCNN deep convolutional networks for largescale image recognition, Computer Science. IEEE 2015 Journal
14. KHe, X. Zhaang, S. Reen, J. *Sun, DEEP RL - Deep residual learning for image recognition, Computer Vision , IEEE conference 2016 Sumit
15. Mw. Espi, Mm. Fuujiimoto, K. Kinooshita, T. Naakatani, In IEEE Journal on Audio speech processing in 2015
16. WLim, D. Jaang, T. Leee, Speech emotion recognition using CNN* and RNN* networksin IEEE conference 2016
17. Dharmarajan, K., and M. A. Dorairangaswamy. "Web user navigation pattern behavior prediction using nearest neighbor interchange from weblog data." International Journal of Pure and Applied Mathematics 116.21 (2017): 761-775.