Comparative Analysis of Random Forests and Inception-v3 for Broadcast Audio Classification

1Kamatchy B, 2P. Dhanalakshmi
1Research Scholar, 2Professor,Dept of Computer Science and Engineering
Annamalai university

***Abstract* - This paper focuses on the comparative analysis of two different audio pattern classifiers for classifying TV broadcast audio data into one of five categories namely Advertisement, Cartoon, News, Songs and Sports. For Classifying TV broadcast audio data, machine learning algorithm Random Forests and deep learning pretrained model Inception -v3 are implemented and the results are compared. In this comparative analysis, the pretrained model Inception-v3 exhibits increased efficiency in classifying the broadcast audio data.**

***keywords* - Audio Segmentation, Mel Frequency Cepstral Coefficients (MFCC), Audio Classification, Inception-v3, Random Forests, Spectrogram, Transfer learning.**

1. **INTRODUCTION**

Rapid increase in the number of audio data calls for powerful and quick processing methods that allow the automatic classification of broadcast audio data which are very useful in many multidisciplinary domains. Several modules for segmentation, and classification are proposed in this paper. The key move in the processing of audio data is to identify automatically or separate the content of the audio into homogeneous segments. This separation criterion is called as segmentation. This is normally carried out along the process of the feature extraction.
The main effort in this content-based audio analysis tends to five types of events separation, and classification namely Advertisement, Cartoon, News, Songs and Sports.

After the acoustic segmentation stage each segment is passed through the classifiers random forests and inceptionv3 and the final output will be one of the predefined categories.
This paper investigates machine learning technique random forests and deep learning pretrained model Inceptionv3 that will support multilingual semantic analysis of broadcasted content, derived from data set created from different tv channels and downloaded from you tube channels.

2. LITERATURE SURVEY

A. J. Vavrek, E. Vozáriková, M. Pleva and J. Juhár, in the paper "Broadcast news audio classification using SVM binary trees," addresses the problem of broadcast news audio classification, by support vector machine - binary tree (SVM-BT) architecture, into the five classes: pure speech, speech with music, speech with environment sound, pure music and environment sound. One of the most substantial steps in creating such classification architecture is selection of an optimal feature set for each binary SVM classifier. They incorporate the F-score feature selection algorithm as an efficient search algorithm across a set of features that are often used for speech/non-speech discrimination.

B. In the paper entitled “Robust Sports Image Classification using Inceptionv3 and Neural Networks”, authors Ketan Joshi, Vikas Tripathi, Chitransh Bose, Chaitanya Bhardwaj, proposes an method focused on the use of Inception V3 for the extraction of features and Neural Networks for the classification of different types of sport. Six types of sports football, tennis, hockey, basketball, volleyball and badminton were used for review and classification. To validate the efficacy of the framework and the Neural Networks, comparisons have been made with other classifiers such as Random Forest, K-Nearest Neighbors (KNN) and Support Vector Machine (SVM). The framework has consistently attained an overall accuracy of 96.64 per cent across six categories, which illustrate the efficacy of the framework and can be used to detect and classify diverse sporting activities in an efficient manner.

C. L. Grama and C. Rusu, in the paper "Audio signal classification using Linear Predictive Coding and Random Forests," present an audio signal classification system based on Linear Predictive Coding and Random Forests. The problem of multiclass classification with imbalanced datasets is taken. The signals under classification belong to the class of sounds from wildlife intruder detection applications: birds, gunshots, chainsaws, human voice and tractors. The proposed system achieves an overall correct classification rate of 99.25%. There is no probability of false alarms in the case of birds or human voices. For the other three classes the probability is low, around 0.3%. The false omission rate is also low: around 0.2% for birds and tractors, a little bit higher for chainsaws (0.4%), lower for gunshots (0.14%) and zero for human voices.

D. Fatemeh Noroozi, Tomasz Sapiński, Dorota Kamińska & Gholamreza Anbarjafari, in the paper "Vocal-based emotion recognition using random forests and decision tree", proposes a new vocal-based emotion-recognition approach using random Forests, where pairs of features on the entire speech signal, including pitch, intensity, the first four formants, the first four formants bandwidths, mean autocorrelation, mean noise-to-harmonic ratio and standard deviation, are used to identify the emotional state of the speaker. The suggested methodology adopts random forests to represent speech signals, along with a decision-tree approach, in order to assign them into various groups. Emotions are divided into six categories, which are happiness, fear, sadness, neutral, surprise, and disgust. The Surrey Audio-Visual Expressed Emotion database is included. According to preliminary findings using leave-one-out cross-validation, by integrating the most important prosodic characteristics, the suggested system has an average recognition rate of 66.28 per cent and at the highest level, a recognition rate of 78 per cent has been reached, which belongs to the happiness voice signals. The proposed method has 13.78%higher average recognition rate and 28.1% higher best recognition rate compared to the linear discriminant analysis as well as 6.58% higher average recognition rate than the deep neural networks results, both of which have been implemented on the same database.

3. METHEDOLOGY

3.1 AUDIO SEGMENTATION

The ultimate aim of segmentation is to create a series of discreet utterances with specific characteristics that remain consistent within each of them. In this work segmentation is composed of Autoassociative Neural Network (AANN) with MFCC features. Types of acoustic classes include Advertisement, Cartoon, News, Songs and Sports. Category change points in audio signal such as news to advertisement, advertisement to song are some examples of segmentation boundaries. In this work, the category change point detection is made using MFCC features extracted from the broadcast audio data. To capture the distribution of the acoustic feature vectors, a five-layer Autoassociative Neural Network model is used. Autoassociative neural network models are feed forward neural networks that perform an identity mapping of input space and are used to capture the distribution of input data. The second and fourth layers of the network have more units than the input layer. The third layer has fewer units than the first or fifth Figure.3.1 shows the five-layer Autoassociative Neural Network.

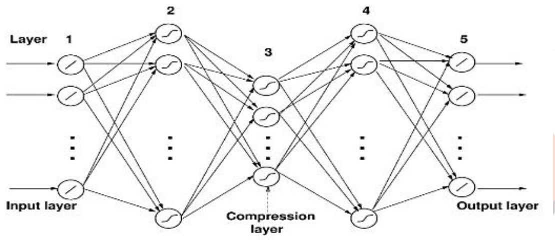


Fig. 3.1 Five-layer Autoassociative Neural Networks

3.2 AUDIO CLASSIFICATION USING RANDOM FORESTS

Audio content classification is basically a pattern recognition problem that can be divided into two components: feature extraction and classification based on the extracted feature. The feature extraction is executed based on MFCC and given to the classifier Random Forests. The Random Forests method can effectively improve the performance of television broadcast audio data classification even with fewer training examples.

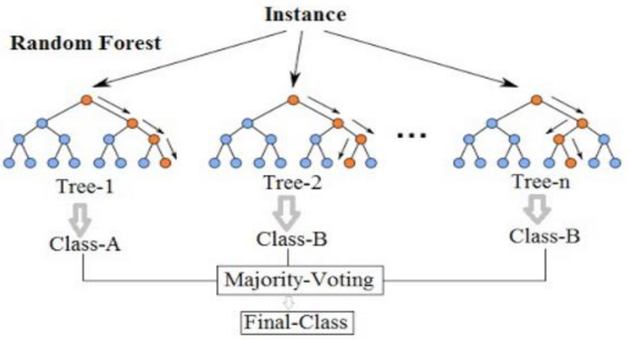
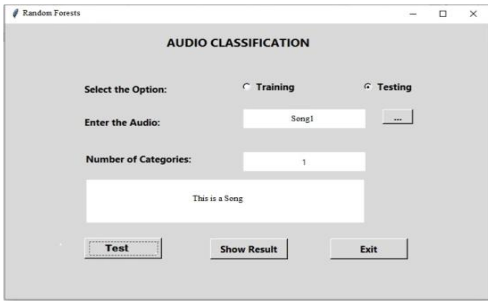


Fig. 3.2 Random Forests

Random Forests or Random Decision Forests are an ensemble learning method for classification, regression and other tasks that works by constructing a variety of decision trees at the time of training, and the system outputs the category which is the mode of the classes (for classification) or mean prediction (for regression) of the individual trees. Random decision forests correct the decision trees' habit of over fitting to their training set. Figure. 3.3 shows the snapshot of Audio Classification System using Random Forests.



Random Forests

AUDIO CLASSIFICATION

Select the Option: ○ Training ● Testing

Enter the Audio: Song1

Number of Categories: 1

This is a Song

Fig 3.3 Snapshot of Audio Classification System using Random Forests.

3.3 AUDIO CLASSIFICATION USING INCEPTON-V3

In this work, deep Convolutional Neural Networks (CNNs) pretrained model Inception-v3 is proposed to extract features and classify the audio segments into one of the five predefined categories. The pretrained model is then trained with the spectrogram images.

3.3.1 Spectrogram

A spectrogram is a visual representation of sound. It displays the amplitude of the frequency components of the signal over time. In a spectrogram representation plot - one axis indicates the time; the second axis indicates frequencies and the colors indicate magnitude (amplitude) of the noticed frequency at a specific time. To plot the spectrogram, the audio signal is broken into millisecond chunks and Short-Time Fourier Transform (STFT) for each chunk is computed. Then this time chunk is plotted as a colored vertical line in the spectrogram. Figure 3.4 shows the spectrogram of an audio sample.

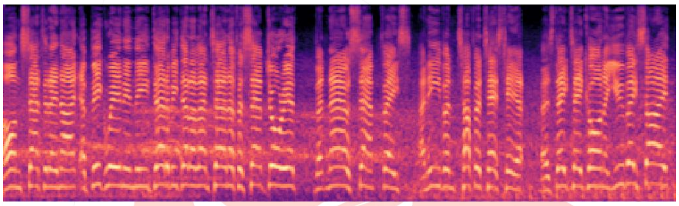


Fig 3.4 Spectrogram of an audio sample

3.3.2 Transfer Learning

Transfer learning utilizes the acquired knowledge when solving one problem and applying it to another but similar problem. In deep learning, transfer learning is a technique whereby a neural network model is initially trained on a problem which resembles the problem which is being solved. One or many more layers from the trained model are made use of in a new model trained on the problem of interest. Transfer learning is generally explained by using the pretrained models. A pre-trained model is the one which was trained on a large benchmark dataset for solving a problem which is nearly the same that we need to be solved.

Inception-v3

The inception deep convolutional architecture was introduced as GoogLeNet, and named as Inception-v1. Then the Inception architecture was refined in several ways, first by the introduction of batch normalization which is named as Inception-v2.

Later by additional factorization ideas in the third iteration is referred to as Inception-v3. Figure 3.5 shows the architecture of Inception-v3.

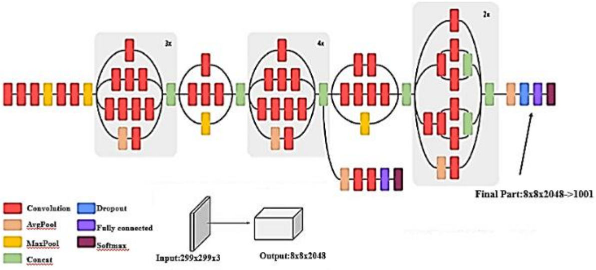


Fig. 3.5 Architecture of Inception-v3

Repurposing the pretrained models

The spectrogram images are given as input data to the pretrained model Inception-v3. Five hidden layers are added to the end of the model Inception-v3 and the parameters are fine-tuned for television broadcast audio classification task. Finally, an output layer with 5 neurons with sigmoid activation function is added as a classifier, which outputs a probability for each class, and the one with the highest probability was chosen as the predicted class. Fig.3.6 shows the snapshot of Audio Classification System using Inception-v3.

4. 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.

5. RESULTS & DISCUSSIONS

5.1 Dataset

The dataset for the television broadcast audio classification is collected from different channels using television tuner card and downloaded from you tube. 200 clips of advertisements, 200 cartoon clips, 200 news clips, 200 clips of songs and 200 sports clips are collected, each are about 10 seconds duration for each category.

5.2 Performance Measures

MFCC Features are extracted from the TV broadcast audio and segmented using AANN. The segmented Audio is given to the classifiers Random Forests and Inception-v3. The performance measures Precision, Recall, and F1 Score are used.

Random Forests and the pretrained model Inception-v3 implementations are executed and the results are tabulated below.

Table 5.1 Performance of Random Forests

Category	Precision (%)	Recall (%)	F-score (%)
Advertisemer	97.0	95.0	96.0
News	95.0	93.0	94.0
Cartoon	90.0	88.0	89.0
Songs	89.0	85.0	87.0
Sports	87.0	97.0	92.0
Accuracy			92.0

Table 5.2 1 Performance of Inception-v3

Category	Precision (%)	Recall (%)	F-Score (%)
Advertisement	97.0	97.0	97.0
Cartoon	97.0	95.0	96.0
News	95.0	90.0	92.0
Songs	90.0	88.0	89.0
Sports	89.0	97.0	93.0
Accuracy			94.0

6. CONCLUSION

In this work, two models Random Forests and Inception-v3 for classifying the TV broadcast audio are implemented and analysed. The comparative analysis shows that the Inception-v3 gives the best accuracy of 94.0% than Random Forests in classifying the TV broadcast audio data.

REFERENCES

[1] J. Vavrek, E. Vozáriková, M. Pleva and J. Juhár, "Broadcast news audio classification using SVM binary trees," 2012 35th International Conference on Telecommunications and Signal Processing (TSP), Prague, 2012, pp. 469-473.

[2] Ketan Joshi, Vikas Tripathi, Chitransh Bose, Chaitanya Bhardwaj**, "**Robust Sports Image Classification Using InceptionV3 and Neural Networks" Procedia Computer Science, Volume 167,2020, Pages 2374-2381.

[3] L. Grama and C. Rusu, "Audio signal classification using Linear Predictive Coding and Random Forests," *2017 International Conference on Speech Technology and Human-Computer Dialogue (SpeD)*, Bucharest, 2017, pp. 1-9,

[4] Fatemeh Noroozi, Tomasz Sapiński, Dorota Kamińska & Gholamreza Anbarjafari, Vocal-based emotion recognition using random forests and decision tree. *International Journal of Speech Technology*,2020**,** pp 239–246.

[5] Zhao, B., Huang, B., & Zhong, Y. (2017). Transfer Learning With Fully Pretrained Deep Convolution Networks for Land-Use Classification. *IEEE Geoscience and Remote Sensing Letters*, *14*(9), 1436–1440.

[6] Zhang, Yan, & Lv, D. (2015). Selected Features for Classifying Environmental Audio Data with Random Forest. *The Open Automation and Control Systems Journal*, *7*(1).

[7] Yanai, K., & Kawano, Y. (2015). Food image recognition using deep convolutional network with pre-training and fine-tuning. *2015 IEEE International Conference on Multimedia Expo Workshops (ICMEW)*, 1–6
[8] Stowell, D., Giannoulis, D., Benetos, E., Lagrange, M., & Plumbley, M. D. (2015). Detection and Classification of Acoustic Scenes and Events. *IEEE Transactions on Multimedia*, *17*(10), 1733–1746.
[9] Singh, J., & Joshi, R. (2019). Background Sound Classification in Speech Audio Segments. *2019 International Conference on Speech Technology and Human-Computer Dialogue (SpeD)*, 1–6.
[10] Pannirselvam, S., & Balakrishnan, G. (2015). Comparative Study on Preprocessing Techniques on Automatic Speech Recognition for Tamil Language. *IJCA Proceedings on National Conference on Research Issues in Image Analysis and Mining Intelligence*, *NCRIIAMI 2015*(2), 25–28.