PERFORMANCE EVALUATION OF DEEP LEARNING ARCHITECTURES FOR ACOUSTIC SCENE CLASSIFICATION

Dinesh Vij[*] and *Naveen Aggarwal*[†]

UIET, Panjab University, Chandigarh
[*]vijdinesh@gmail.com
[†]navagg@gmail.com

ABSTRACT

This paper is a submission to the sub-task Acoustic Scene Classification of the IEEE Audio and Acoustic Signal Processing challenge: Detection and Classification of Acoustic Scenes and Events 2017. The aim of the sub-task is to correctly detect 15 different acoustic scenes, which consist of indoor, outdoor, and vehicle categories. This work is based on log mel-filter bank features and deep learning. In this short paper, the impact of different parameters while applying a basic Deep Neural Network (DNN) architecture is first analyzed. The accuracy gains obtained by the different types of deep learning architectures such as Recurrent Neural Network (RNN), Gated Recurrent Unit (GRU), Long Short-Term Memory (LSTM), and Convolutional Neural Network (CNN) are then reported. It has been observed that the overall best scene classification accuracy was obtained with CNN.

Index Terms— Acoustic Scene Classification, Deep Learning, RNN, LSTM, GRU, CNN

1. INTRODUCTION

This short paper describes our submission to the sub-task Acoustic Scene Classification of the Detection and Classification of Acoustic Scenes and Events 2017 (DCASE 2017) challenge. The aim of the sub-task is to correctly identify 15 different acoustic scenes, which consist of indoor, outdoor, and vehicle categories. It is 3rd official IEEE Audio and Acoustic Signal Processing challenge, organized by IEEE Signal Processing Society. This short paper has analyzed the relevance of different parameters while applying a basic DNN for the scene classification task. Then the accuracy gains obtained by using various deep learning architectures such as RNN, GRU, LSTM, and CNN are further presented. An overview of the whole system is shown in Figure 1. The binaural input signal composed of two channels is first combined by taking the mean of the two channels, thereby converting the binaural signal into the mono-channel signal. The combined input audio signal obtained after channel combination is then segmented into smaller manageable chunks called windows. This process is known as windowing. It has been observed that the spectral characteristics of the cumulative acoustic scenes do not significantly change over the short time spans. Therefore, larger window sizes are more appropriate for detecting the cumulative acoustic scenes. But on using larger window sizes, the number of output feature frames becomes less in number, which means fewer data to train the neural network. Therefore larger window sizes are not appropriate while applying deep architectures because neural networks require a large amount of data for training. The length of window size used in our work is 1024 samples. Log mel-filter bank features are extracted from these short windows. These feature frames are then combined to form a longer concatenated feature vector. This concatenated feature vector is then passed as an input to the DNN. After passing through the hidden layers of this DNN, this concatenated feature vector is then given one class label at the output layer. Then, majority voting is done amongst all such feature vectors to give one class label to the full 10-sec recording.

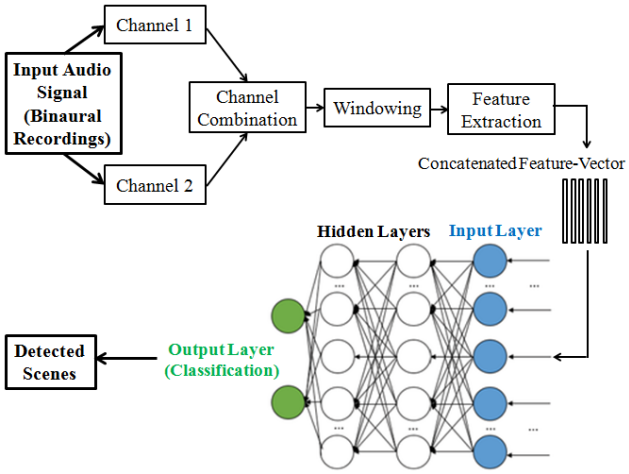

Figure 1: Basic architecture of Acoustic Scene Classification.

The paper is organized as follows: Section 2 details the preprocessing and feature extraction process used for capturing the characteristics of various acoustic scenes. Section 3 describes the basic deep learning architecture used in this work. Section 4-7 describes the four deep learning architectures widely used by researchers in the literature, i.e., RNN, LSTM, GRU, and CNN. The evaluation results obtained by using these four deep archi-

This research is supported by the CARTS Project funded by ITRA, Media Lab Asia; and Design Innovation Center, Panjab University, Chandigarh.

tectures are presented in section 8. Finally, section 9 concludes the paper and gives future directions. In section 10, the classification results achieved on evaluation dataset (after challenge completion) are discussed.

2. PRE-PROCESSING AND FEATURE EXTRACTION

The binaural input audio signal is first combined by taking mean of the two channels. Then the combined signal is segmented into smaller portions using short windows of 1024 samples. For feature extraction, we have used log mel-filter bank features with 40 channels to capture the distinctive acoustic signatures of various scenes. Librosa library [1] is used for extracting mel-filter bank features. The extracted features are then normalized to the same scale using z-score normalization. Ten such feature frames are then concatenated, to form a longer feature vector of 400 units. This feature vector is then passed as an input to the DNN described in the next section.

3. DEEP LEARNING

In the recent years, deep learning has been widely used in various fields such as computer vision, speech recognition, natural language processing, etc. Researchers have also used various types of deep learning architectures to solve the above-mentioned problems. DNNs have the ability to learn hierarchical representations without explicitly providing the system with hand-engineered features. In this section, we try to use a general deep learning architecture as provided by Kong *et al.* [2], and in the subsequent sections, we experiment with the various popular deep architectures used in the recent past. Our aim is to evaluate that how the different architectures perform on the scene classification task as compared to the baseline method.

3.1. Neural Network Structure

We have used a fully connected neural network with two hidden layers in our work. The input layer consists of 400 input nodes corresponding to 10 concatenated log mel-filter bank feature frames (10 frames * 40 features = 400 values). In the next two hidden layers, 400 hidden units per layer are used. A dropout value of 0.5 is used between layers to prevent over-fitting. At each node, Leaky-Relu activation function is used. Rmsprop optimizer with categorical cross-entropy loss function (objective function) is used to minimize the loss. For training the neural network, the batch size is set to 256. The learning rate is set to 0.001, and the maximum number of training epochs is set to 20.

3.2. Scene Classification

For the scene classification task, our system used a softmax output layer. Further, the last layer consisted of 15 units corresponding to 15 acoustic scenes to be predicted. The implementation is based on deep learning framework built on top of Theano [3] by Qiuqiang Kong. We achieved an overall scene classification accuracy of 72.56 % by using this setup.

4. RNN

RNN is a type of DNN learning model, which is used to learn sequence data. DNN has a disadvantage that after each data point is processed; the entire state of the network is lost. RNN removes this problem by passing information across sequence steps. The use of RNN improved the overall scene classification accuracy to 77.26 % over the basic DNN architecture.

5. LSTM

The training of an RNN model is very complex. LSTM is a specific type of RNN architecture, which is easier to train and can learn long-term dependencies. The basis of LSTM is the use of cell states, and the ability to add or remove information from cell states by using structures called gates. The overall scene classification accuracy obtained with LSTM is 78.97%.

6. GRU

GRU is a simpler variation of LSTM, which has gained popularity since 2014. Many researchers have reported accuracy gains by using GRU over the standard LSTM architecture. Therefore, in our work, we also tried to use GRU for the acoustic scene classification task. The overall scene classification accuracy obtained with GRU is 78.03%.

7. CNN

A CNN is basically a DNN in which stacks of convolutions are used instead of stacks of matrix multiplication layers. In a DNN, unique weights are learned for each point. However, in a CNN, the weights are shared across space. CNN has been mainly used for visual applications. But in the recent years, researchers have started exploring its' application in the audio domain as well. In this work, the input layer shape for CNN is set to (batch number, 1, 10, 40) corresponding to (batch number, number of audio channels i.e. mono channel, concatenated frames i.e. height, length of single frame i.e. width). The convolution size is set to (3, 3). The number of output feature maps is set to 32 in both the layers of the 2-layered architecture. Further, pooling operation is not performed. The overall scene classification accuracy obtained by using CNN is 82.74%., which is highest amongst the four architectures used in this work.

8. EVALUATION

We evaluated our approach using fold1 of the cross-validation setup provided by the DCASE 2017 organizers. Development dataset is used for this purpose. It consists of 78 segments (10-sec each, totaling 13 min of audio) for each acoustic scene [4]. Final classification accuracy and scene-wise classification results obtained on fold1 using the provided development dataset are shown in Table 1. The deep learning models are then applied on Evaluation dataset, which consists of a total of 1620 segments (10-sec each, totaling 270 min of audio) for all the acoustic scenes and the results are submitted to the challenge. As a comparison with machine learning approaches, we have compared our results with our last year's approach based on machine

learning using Support Vector Machines (SVM) [5]. The results obtained by using SVM classifier are listed under the column SVM in Table 1.

Table 1: Scene-wise classification accuracy in %age

Acoustic Scene	Baseline System	SVM	RNN	LSTM	GRU	CNN
Beach (outdoor)	76.6	81.73	94.87	93.59	94.87	93.59
Bus (vehicle)	74.7	73.72	47.44	43.59	41.03	61.54
Café (indoor)	64.1	67.95	78.21	97.44	93.59	98.72
Car (vehicle)	94.9	90.71	88.46	78.21	83.33	91.03
City center (outdoor)	90.4	80.77	79.49	93.59	91.03	97.44
Forest path (outdoor)	84.0	74.68	82.05	83.33	88.46	87.18
Grocery store (indoor)	68.9	84.62	78.21	79.49	76.92	84.62
Home (indoor)	66.4	64.94	75.64	80.77	74.36	74.36
Library (indoor)	51.9	63.14	51.28	46.15	46.15	46.15
Metro station (indoor)	92.6	93.91	83.33	84.62	85.90	98.72
Office (indoor)	99.4	90.07	100	97.44	100	100
Park (outdoor)	60.3	56.73	69.23	73.08	73.08	76.92
Residential (outdoor)	63.5	63.78	44.87	46.15	48.72	44.87
Train (vehicle)	34.0	59.30	100	97.44	96.15	96.15
Tram (vehicle)	78.5	73.08	85.90	89.74	76.92	89.74
Overall accuracy	**73.3**	**74.61**	**77.26**	**78.97**	**78.03**	**82.74**

9. CONCLUSION AND FUTURE DIRECTIONS

Overall accuracy reported by the baseline system is 73.3%. Compared to the baseline system, all the four deep architectures used in this work provide an improvement in the classification accuracy. Also, all the four deep architectures outperformed our last year's machine learning approach based on SVM. The least accurate results correspond to basic deep neural network architecture, and the overall best scene classification accuracy has been obtained with CNN. Future directions include representing acoustic scenes in terms of component acoustic events, and inferring important properties of these events which would aid in detecting the acoustic scenes associated with them. To differentiate between two scenes having common type of events, the repetition frequency of individual events within each scene can be useful. Also, two scenes can be matched for some pre-defined sequence of events, as this sequence might be different for different scenes. Wavelet packet features could also be explored as an alternative input to the neural network architecture, for extracting important information from various sub-bands of the input acoustic signal.

10. RESULTS ON EVALUATION DATASET

Evaluation dataset consists of a total of 1620 segments (10-sec each, totaling 270 min of audio) for all the acoustic scenes. On this dataset, it has been observed that best results are achieved by using convolutional neural networks. An overall scene classification accuracy of 65.0% is obtained with CNN, which is better than the baseline system's accuracy of 61.0%. The results obtained on the evaluation dataset by using various deep architectures are shown in Table 2.

Table 2: Results on Evaluation Dataset

System	Baseline	RNN	LSTM	GRU	CNN
Classification Accuracy (%)	61.0	61.2	57.5	59.6	**65.0**

11. REFERENCES

[1] "Librosa: A python library for audio signal processing and music analysis." Available: https://github.com/librosa.

[2] Q. Kong, I. Sobieraj, W. Wang, and M. D. Plumbley. "Deep Neural Network Baseline for DCASE Challenge 2016." in Proceedings of the Detection and Classification of Acoustic Scenes and Events 2016 Workshop (DCASE2016), pp. 50-54, September 2016.

[3] Theano Development Team. "Theano: A Python framework for fast computation of mathematical expressions." Available: http://deeplearning.net/software/theano/.

[4] A. Mesaros, T. Heittola, A. Diment, B. Elizalde, A. Shah, E. Vincent, B. Raj, and T. Virtanen. "DCASE 2017 challenge setup: tasks, datasets and baseline system." in Proceedings of the Detection and Classification of Acoustic Scenes and Events 2017 Workshop (DCASE2017), November 2017. submitted.

[5] D. Vij, N. Aggarwal, B. Raman, K.K. Ramakrishnan, and D. Bansal. "Acoustic Scene Classification Based On Spectral Analysis And Feature-Level Channel Combination." in Proceedings of the Detection and Classification of Acoustic Scenes and Events 2016 Challenge (DCASE2016), September 2016.