# ACOUSTIC SCENE AND EVENT RECOGNITION USING RECURRENT NEURAL NETWORKS

*Toan H. Vu, Jia-Ching Wang*

National Central University
Department of Computer Science and Information Engineering
Taoyuan, Taiwan

## ABSTRACT

The DCASE2016 challenge is designed particularly for research in environmental sound analysis. It consists of four tasks that spread on various problems such as acoustic scene classification and sound event detection. This paper reports our results on all the tasks by using Recurrent Neural Networks (RNNs). Experiments show that our models achieved superior performances compared with the baselines.

***Index Terms***— RNN, GRU, acoustic scene classification, sound event detection, audio tagging

## 1. INTRODUCTION

Environmental sound analysis has attracted a lot of researchers' attention recently. In real, a sound does not often come from a single source but it is frequently a combination of sounds from many sources, which make machines more challenging to perceive. The four tasks in the DCASE2016 challenge cover several aspects of environmental sound analysis: acoustic scene classification, sound event detection, and domestic audio tagging.

In the paper, we introduce our experimental results by employing Recurrent Neural Networks (RNNs) to all the tasks. The remainder of this paper is organized as follows. Section 2 presents general information about RNNs, and describes RNN architectures for classification. In Section 3, we conduct experiments of the four tasks, and report our results. Finally, Section 4 draws our conclusions.

## 2. RECURRENT NEURAL NETWORKS

A recurrent neural network is a computational neural network that has feedback connections, so it works efficiently and flexibly with time-series signals such as audio and video. Figure 1a shows a simple RNN structure where hidden units are self-connected. The corresponding diagram of its recurrent units in Figure 1b illustrates how the hidden state accumulate information from its previous state and an input at a specific time.

In simple RNNs, the hidden state at a time $t$ is computed by

$$h_t = f(W_{ih} i_t + W_{hh} h_{t-1}) \quad (1)$$

where $i_t$ is input at time $t$; $f$ is an activation function; $W_{ih}$ and $W_{hh}$ represent weight matrices of connections between input and hidden and between hidden and hidden layers, respectively. For conciseness, biases are not included.

In fact, due to the exploding and vanishing gradient problem [1], a simple RNN is not easy to train, and not able to deal with long-range dependencies. Alternatively, we employ Gated Recurrent Neural Networks that are RNNs whose hidden units are gate-based. Two well-known types of Gated RNNs are Long Short Term Memory networks (LSTMs) and Gated Recurrent Units (GRUs). While LSTMs have been widely used for a long time, GRUs have just been introduced recently in [2]. However, it was shown that GRUs have comparable performances to LSTMs in various applications especially in sequence modeling, but with lower computational cost [3]. Hence, GRU is our choice in this challenge, RNN models in the experiment section are referred as GRUs from now on. Description about LSTM and GRU can be found in [4, 5] and [2, 3], respectively.

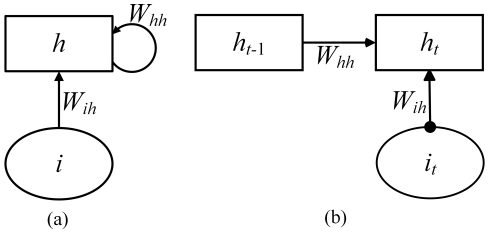


Figure 1: (a) A simple recurrent neural network; (b) The corresponding block diagram of a simple recurrent unit.

RNNs are very flexible in classifying sequential data in both cases of sequence-to-one classification and sequence-to-sequence classification as can be seen in Figure 2a&2b. Instead of using a sequence-to-sequence RNN, a bidirectional RNN (BiRNN) in which there is a second hidden layer that learns input sequence in an inverse direction (Figure 2c) is frequently employed. It is supposed to generate better prediction since information to make a prediction at each time-step comes from both the backward and forward directions.

## 3. EXPERIMENTS

### 3.1. Task 1: Acoustic scene classification

The used features are 13 mel-frequency cepstral coefficients (MFCCs) extracted from frames of duration 0.02 second and 0.01 second overlap, with their first and second temporal derivatives, which forms 39 dimensional features. The features are processed by doing zero mean and unit variance on each dimension over the training set. We divide each audio features into windows, each is 0.5 second long. Totally, around 35.000 windows are segmented from each class. We use them as input sequences to the RNN model.

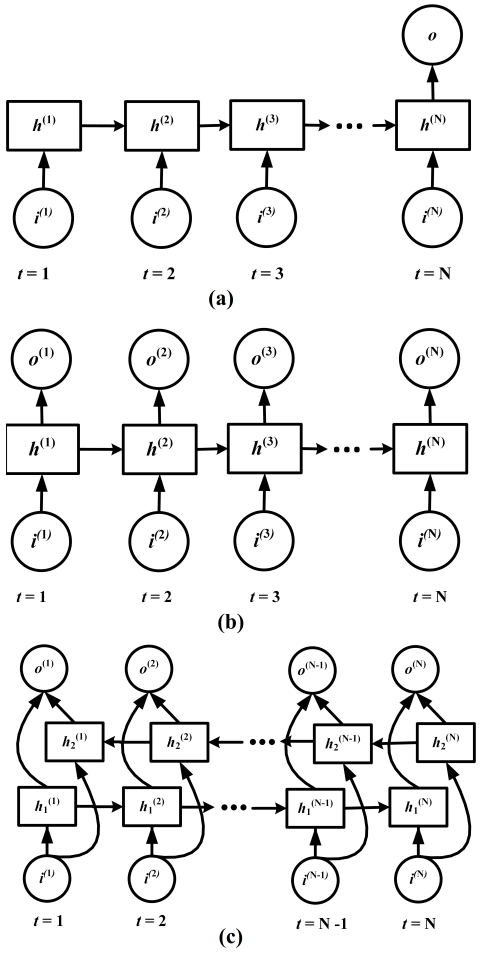


Figure 2: Recurrent neural networks for classification: (a) A sequence-to-one RNN architecture, (b) A sequence-to-sequence RNN architecture, (c) A sequence-to-sequence BiRNN architecture.

The network type is sequence-to-one, the net has 100 hidden units. Output of a window is probabilities of labels computed by softmax connections. Cross-entropy was used as cost function. We train the network by gradient descent: gradient clipping with a threshold of 1, and ADADELTA method with hyper parameters $\rho = 0.95$, $\varepsilon = 1e-6$ as in [6]. The optimal set of parameters is found at the best results on the development set. In testing time, a prediction of a recording is an average of predictions from all its segmented windows. Table 1 shows average results over 4 folds from our system and the baseline [7]. We achieve an overall average accuracy of 82.09% outperforming the baseline with a margin of 9.59%.

### 3.2. Task 2: Sound event detection in synthetic audio

The task 2 is for detecting events in synthetic audio. Besides 11 given categories, we consider an "unknown" category, which would help our model distinguish better. Training data for "unknown" are extracted from a random file in the development set. After that, the file will be removed from the development set, and not be included for evaluation. CQT coefficients are extracted from audio files with the same configuration as the baseline [7]. Then, we apply PCA whitening with 99% of variance retained to reduce dimension size of features. We exploit a BiRNN with size of 100 hidden units that learns on input sequences of 50 frames. Table 2 reports our results in this task. In segment-based overall metrics, the ER and F-score of our model are 0.3412 and 81.15%, respectively, which surpass that of the baseline.

Table 1: Acoustic scene classification results (average over 4 folds)

| Label | Accuracy (%) | |
|---|---|---|
| | **Baseline** | **Ours** |
| Beach | 69.3 | 88.16 |
| Bus | 79.6 | 60.53 |
| Cafe / Restaurant | 83.2 | 73.98 |
| Car | 87.2 | 92.30 |
| City center | 85.5 | 97.55 |
| Forest path | 81.0 | 92.46 |
| Grocery store | 65.0 | 89.60 |
| Home | 82.1 | 79.59 |
| Library | 50.4 | 83.63 |
| Metro station | 94.7 | 96.05 |
| Office | 98.6 | 100.00 |
| Park | 13.9 | 67.50 |
| Residential area | 77.7 | 72.84 |
| Train | 33.6 | 50.96 |
| Tram | 85.4 | 86.14 |
| **Overall accuracy** | **72.5** | **82.09** |

Table 2: Results of sound event detection in synthetic audio

| Model | Segment-based overall metrics | | Event-based overall metrics |
|---|---|---|---|
| | **ER** | **F-score** | **F-score (onset-only)** |
| **Baseline[7]** | 0.7859 | 41.6 % | 30.3 % |
| **Ours** | 0.3412 | 81.15% | 32.32% |

### 3.3. Task 3: Sound event detection in real life audio

Each recording is encoded to extract 40 log mel-filter bank coefficients, then these features are normalized to zero mean and unit variance. A BiRNN with size of 50 hidden units is exploited to work on input sequences that is 50 frames long. We evaluate results by using metrics that are defined in [8]. In table 3, we present values of 0.815 average ER and 49.75% F-score achieved from our model those are superior than the baseline.

### 3.4. Task 4: Domestic audio tagging

In this task, each recording is converted to get 13 MFCCs with frame size of 0.04 second and hop size of 0.02. These features then are normalized to zero mean and unit variance. RNN type is sequence-to-one. Given the setup of five folds, at each training time, 3 folds are used to train, 1 fold is for development set, and the remainder is test set. All folds are used as test data one time. The optimal set of parameters is determined by the best performance on the development set. As can be seen in table 4, our experimental results are comparable to the baseline with 0.20 EER.

Table 3: Results of sound event detection in real life audio

| | Baseline[7] | | Ours | |
|---|---|---|---|---|
| **Acoustic scene** | **ER** | **F-score** | **ER** | **F-score** |
| **Home** | 0.96 | 15.9 % | 1.03 | 39.5% |
| **Residential area** | 0.86 | 31.5 % | 0.60 | 60.0% |
| **Average** | **0.91** | **23.7 %** | **0.815** | **49.75%** |

Table 4: Audio tagging results over evaluation folds

| **Audio tag** | **EER** | |
|---|---|---|
| | **Baseline[7]** | **Ours** |
| adult female speech | 0.29 | 0.26 |
| adult male speech | 0.30 | 0.24 |
| broadband noise | 0.09 | 0.11 |
| child speech | 0.20 | 0.21 |
| other | 0.29 | 0.29 |
| percussive sound | 0.25 | 0.23 |
| video game/tv | 0.07 | 0.06 |
| **Mean error** | **0.21** | **0.20** |

## 4. CONCLUSIONS

In this report, we introduce our experimental results in the DCASE2016 challenge. Recurrent neural networks showed their effectiveness and flexibility in working with various problems in audio analysis.

## 5. REFERENCES

[1] R. Pacanu, T. Mikolov, and Y. Bengio, "On the difficulties of training recurrent neural networks," in *Proceedings of the 30th International Conference on Machine Learning, ICML 2013*, no. 2, 2013, pp. 1310–1318.

[2] K. Cho, B. van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 1724–1734. [Online]. Available: http://arxiv.org/abs/1406.1078

[3] J. Chung, Ç. Gülçehre, K. Cho, and Y. Bengio, "Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling arXiv : 1412 . 3555v1 [ cs . NE ] 11 Dec 2014," pp. 1–9.

[4] S. Hochreiter, S. Hochreiter, J. Schmidhuber, and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–80, 1997. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/9377276

[5] A. Graves, A.-r. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2013, pp. 6645–6649.

[6] M. D. Zeiler, "ADADELTA: An adaptive learning rate method," *CoRR*, vol. abs/1212.5, p. 6, 2012. [Online]. Available: http://arxiv.org/abs/1212.5701

[7] A. Mesaros, T. Heittola, and T. Virtanen, "TUT database for acoustic scene classification and sound event detection," in *24th European Signal Processing Conference 2016 (EUSIPCO 2016)*, Budapest, Hungary, 2016.

[8] A. Mesaros, T. Heittola, and T. Virtanen, "Metrics for polyphonic sound event detection," *Applied Sciences*, vol. 6, no. 6, p. 162, 2016. [Online]. Available: http://www.mdpi.com/2076-3417/6/6/162