Audio Event Classification Using Deep Neural Networks

Minkyu Lim, Donghyun Lee, Hosung Park, Unsang Park and Ji-Hwan Kim*
Department of Computer Engineering
Sogang University
Seoul, Korea
{lmkhi, redizard, hosungpark, unsangpark, kimjihwan }@sogang.ac.kr

Jeong-Sik Park
Department of Information and Communication Engineering
Yeungnam University
Gyeongsan, Korea
parkjs@yu.ac.kr

Gil-Jin Jang
School of Electronics Engineering
Kyungpook National University
Daegu, South Korea
gjang@knu.ac.kr

***Abstract*—This paper proposes an audio event classification method using deep neural networks (DNNs). The proposed method applies a feed-forward neural network (FFNN) to generate event probabilities for twenty audio events (e.g., birds, wind, rain) for each frame. For each frame, mel-scale filter-bank features of the adjacent frames are used as the input vector for the FFNN. These event probabilities are accumulated for the events, and the classification result is determined as being the event having the highest accumulated probability. The best accuracy achieved by the proposed method was 72.2%, for the UrbanSound8K and BBC SoundFX dataset when 560 mel-scale filter-bank features from a frame and its left and right three context frames (80 features per frame) were used as the input vector for the FFNN with two hidden layers and 2,000 neurons per hidden layer. In this configuration, the rectified linear unit was suggested as its activation function.**

Keywords—audio event classification; deep neural networks;

I. INTRODUCTION

Deep neural networks (DNNs) are receiving attention as a technology that has shown notable performance enhancement in current machine learning fields. A DNN is a deep artificial neural network composed of many hierarchies and can achieve better performance in classification problems because complex nonlinear learning boundaries can be separated better than with conventional artificial neural networks. It presents the challenge of requiring many calculations for estimating various parameters, but DNN can be successfully applied in a variety of fields because of the recent development of hardware technology and algorithms, and availability of big data.

DNN has shown great enhancement of performance when applied to voice recognition and image classification, but there have not been many cases in which it has been applied to audio event classification. In this study, an audio event classifier using DNN was implemented, and the hyper-parameters that compose the DNN were experimentally estimated.

This paper is structured as follows. Previous works for audio event recognition are described in Section II, and an event classifier using DNN is described in Section III. In Section IV, hyper-parameters of the audio event classifier using DNN are experimentally estimated, and the level of performance enhancement is evaluated by a comparison with existing classifiers. The conclusion is presented in Section V.

*Corresponding author: Ji-Hwan Kim

II. PREVIOUS WORKS

From a study on audio event classifiers based on a hidden Markov model (HMM) and support vector machine (SVM), 15 audio event classifiers were trained by combining various feature vectors such as mel-frequency cepstral coefficients (MFCCs), perceptual linear prediction (PLP), and zero crossing rate (ZCR), showing differences in recognition performance depending on the selection of feature vectors [1]. The rate of event detection was measured by directly labeling audio events that appear in movies, documentaries, talk shows, and news. The experiment showed that performance was best when using PLP as the feature vector.

The difference between conventional SVM and DNN was theoretically analyzed [2]. SVM and DNN were classified into the categories of shallow architecture and deep architecture. SVM reduces dimensionality through the kernel function to classify the class distribution and can be seen as a form of artificial neural network that has one hidden layer. On the other hand, DNN, which has several hidden layers, hierarchically stacks nonlinear boundaries in which decision boundary is more complex than SVM or single hidden layer artificial neural networks.

III. DEEP NEURAL NETWORK–BASED AUDIO EVENT CLASSIFIER

The proposed DNN-based audio event classifier is composed of two modules: the audio feature extractor and the audio event classifier.

A. Audio Feature Extractor

The audio feature extractor generates column vectors that are used in the DNN input vector. Short-time Fourier transform (STFT) is performed using a 20-ms Hamming window on the input audio (2 bytes per sample, mono, 16 kHz) moving in 10-ms units. A triangular-shaped bin that increases by mel scale in every window is applied, and the weighted value is multiplied by the energy of each frequency to extract the feature value. This

is expressed in one vector to generate the mel-scale filter-bank (FBank) feature vector.

The feature vector extracted from one window is the feature value that corresponds to 10 ms, which has only a very temporary sound feature and had a very large feature change for window movements within the equivalent sound range. Therefore, to reflect the contextual information, *N* windows on the left and right of the current window are combined to generate one vector, and this is used as the input of the classifier.

B. Audio Event Classifier

The structure of the audio event classifier is described in this section. The audio event classifier is composed of a feed-forward neural network (FFNN). The FFNN consists of one input layer, one output layer, and one or more hidden layers. Each layer is composed of neurons, and each neuron has weight and bias as parameters. The output of all neurons on the lower layer becomes the input of that neuron. The feature vector generated from the audio feature extractor described in Section III-A becomes the input vector $\bar{x}$ in the FFNN, which makes up the audio event classifier. The input vector $\bar{x}$ passes through the functions given by (1) and (2) until reaching the final output layer in the FFNN. The values of the output layer corresponding to the input vector represent the occurrence probability of each class.

$$a_i^k(\bar{x}) = b_i^k + \sum_{j=1}^{N^{k-1}} w_j^k \, h_j^{k-1}(\bar{x}) (1 \le k \le L+1) \qquad (1)$$

$$h_i^k(\bar{x}) = g\left(a_i^k(\bar{x})\right) (1 \le k \le L+1) \qquad (2)$$

$a_i^k(\bar{x})$ in (1) is the calculated preactivation function in the *i*-th *k*-layer neuron when the vector $\bar{x}$ input. The final output of the relevant neuron becomes the result of the activation function g applied to the result of the preactivation function.

IV. Performance Evaluation of Audio Event Classifier According to Changes in Hyper-parameter Values

The data used for training and evaluation are UrbanSound8K and BBC SoundFX [3]. UrbanSound8K is data labeled by 10 event classes on audio signals collected from www.freesound.org. It consists of 8,732 samples, duration of each is around four seconds. There are about nine hours of data with 10 audio event types. BBC SoundFX is a sound effect corpus extracted from BBC audio sound of the 1990's and before. The dataset is composed of 1,655 sound tracks in total 40 audio CDs. We manually tagged sound events for all of the sound tracks. Then the top ten audio event classes were selected from the longest according to the length of audio events. Theses ten classes are appended to the audio event set used in this research. All data were converted to 16-bit mono 16-kHz format, of which 9/10 were used for training and 1/10 were used for evaluation. The total audio event list is shown in Table I.

Voice recognition toolkit HTK was used for the feature vector extraction, and Theano[4] was used for the DNN-based audio event classifier. The size of the minibatch was 98,569. The learning rate started from 0.01, and the rate was reduced by 10% until it reaches to 0.001, when the validation error was not reduced for consecutive 30 sessions.

TABLE I. Event List for UrbanSound8K and BBC SoundFX Dataset

Air conditioner	Engine idling	Children play	Jack hammer	Street music
Car horn	Gun shot	Dog barks	Siren	Drilling
Baby cry	Road	River	Wind	Bird
Rain	Cow	Horse	Crowd	Ship

The experiment was conducted by changing the number of neurons per hidden layer. The test results are shown in Table II. The model with two hidden layers and 3,000 neurons per hidden layer when the ReLU activation function is used.

TABLE II. Audio Event Classification Accuracy for Various Hyperparameter Values

Number of Neurons per Hidden Layer	FBank Dimension	Number of Input Frames	Recognition Rate (%)
500	80	7	52.3
1,000	80	7	53.6
2,000	80	7	68.5
3,000	80	7	72.2

V. Conclusion

An audio event classifier using DNN was proposed in this study, and its performance was evaluated through experiments. Performance was measured through DNN-based classifiers with a few hyper-parameters. A maximum performance of 72.2% was shown when the number of hidden layers is 2 and the number of neurons per hidden layer is 3,000.

Acknowledgment

This work was supported by the ICT R&D program of MSIP/IITP. [R0126-16-1112, Development of Media Application Framework based on Multi-modality which enables Personal Media Reconstruction]

References

[1] J.P. Elo et al., "Non-speech audio event detection," in Proc. International Conference on Acoustics, Speech and Signal Processing, pp. 1973-1976, 2009.

[2] Y. Bengio and Y. LeCun, "Scaling learning algorithms towards AI," Large-scale Kernel Machines, Vol. 34, No. 5, pp. 321-360, 2007.

[3] J. Salamon, C. Jacoby, and J.P. Bello, "A dataset and taxonomy for urban sound research," in Proc. ACM International Conference on Multimedia, pp. 1041-1044, 2014.

[4] J. Bergstra et al., "Theano: A CPU and GPU math expression compiler," in Proc. Python for Scientific Computing Conference, Vol. 4, p. 3, 2010