EMPIRICAL STUDY ON ENSEMBLE METHOD OF DEEP NEURAL NETWORKS FOR ACOUSTIC SCENE CLASSIFICATION

Jaehun Kim

Music and Audio Research Group
Seoul National University
Seoul, Republic of Korea
eldrin@snu.ac.kr

Kyogu Lee

Music and Audio Research Group
Seoul National University
Seoul, Republic of Korea
kglee@snu.ac.kr

ABSTRACT

The deep neural network has shown superior classification or regression performances in wide range of applications. In particular, the ensemble of deep machines was reported to effectively decrease test errors in many studies. In this work, we extend the scale of deep machines to include hundreds of networks, and apply it to acoustic scene classification. In so doing, several recent learning techniques are employed to accelerate the training process, and a novel stochastic feature diversification method is proposed to allow different contributions from each constituent network. Experimental results with the DCASE2016 dataset indicate that an ensemble of deep machines leads to better performances on the acoustic scene classification.

Index Terms— neural network, ensemble model, ZCA whitening, bagging,

1. INTRODUCTION

Acoustic Scene Classification (ASC) is a task defined as automatically identifying where the environmental acoustic signal is recorded. Traditionally, many existing approaches adopted hand-engineered features such as variants of Mel-frequency cepstral coefficients (MFCCs), Linear prediction coefficients (LPC), or other spectral low-level features, which are mainly derived based on the domain knowledge of acoustics [1, 2]. Some approaches tried to learn the inherent features of data in an unsupervised manner, including principal component analysis (PCA), non-negative matrix factorization (NMF), or sparse feature learning [3, 4]. These features are usually pre-processed and fed into classification models such as Gaussian mixture models (GMMs), support vector machine(SVMs), hidden Markov models (HMM), or artificial neural networks (ANNs) [2]. Recently, more deep structured models or ensemble of a number of models have been reported to be effective on the ASC problem [5, 1].

The ensemble method is a well developed meta-training framework for classification or regression task, and also known as generally more accurate than its constituent predictor[6]. There are a number of strategies for constructing more optimal and effective ensemble estimator, including bagging[7] and boosting[8]. The most well-known ensemble models exploiting bagging and boosting, such as random forest or Adaboost use the decision trees as its base estimator. However, ensemble models constituted with ANNs also reported as successfully reducing test error significantly. Furthermore, simple ensemble strategies, which combining terminal output activation of each constituent ANN, reduce estimation error even when the number of base estimators reaches at 100[6].

In this paper, we introduce a deep-ANN based ensemble method which is coupled with unsupervised feature extraction. The recent researches of deep-ANN models, also known as the deep neural network(DNN), increasingly introduce a very deep and large structure. Even there is a huge enhancement of the computational capability by the enormous advance of parallel computing and GPGPU, the computational complexity of training DNNs still one of the big hurdle for the constituting ensemble of this model. For overcoming the problem, we exploit unsupervised feature learning and other recent techniques not only accelerating the training procedure but also decrease the test error of the DNN. Using the data provided for DCASE challenge[9], we will show how this approach successfully achieve improvement of model accuracy.

2. PROPOSED METHOD

We constructed an ensemble model consist of multiple convolutional neural networks(CNN) models for categorizing input recordings into a certain acoustic environment. We exploited unsupervised feature extraction stage as an initial layer of the constituent CNN, as a pre-trained feature extractor.

2.1. Stochastic Feature Diversification

In order to give diversity, we applied both feature-wise and framework-wise approaches. As a feature-wise strategy, we applied zero-phase component analysis(ZCA) whitening as a preprocessing and also performed stochastic dropping on the transformation weight matrix. The procedure is described as follows:

- Compute time-frequency representation of given recorded audio signal
- Sample sufficient amount of frames of entire data set and compute covariance matrix
- Compute eigenbases with Eigen decomposition algorithm such as singular value decomposition(SVD)
- Select arbitrary eigenbases in a stochastic manner, then whiten the input data with selected eigenbases.

Specifically, first, 60 eigenbases, which covers 99.9% of the variance of given dataset is forced to be preserved and randomly dropped higher than 60th eigenbasis. By the stochastic dropping, we could add 'mild' difference between the whitened features. We assumed this small variance of feature make each constituent model

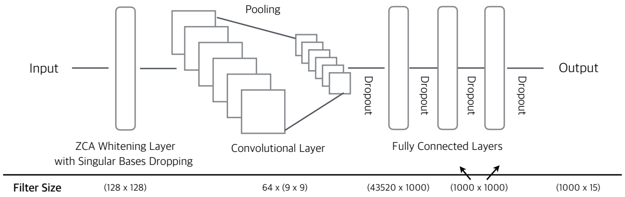


Figure 1: Structure of proposed constituent network.

learning a diverse aspect of the data, which results in the meaningful diversity among the machines. Additionally, we applied bagging to construct ensemble model.

2.2. Supervised Training and Construct Ensemble Model

On top of the ZCA whitening layer, we stacked simple CNN structure for supervised feature extraction and classification. One 2-dimensional convolutional-pooling layer and 3 fully connected layers followed, then one softmax layer plugged in order to terminal classification. We employed this identical structure for all constituent models. Output activation of constituent models aggregated by taking median value.

3. EXPERIMENT

3.1. Dataset

We executed an evaluation of proposed model with the development dataset provided by DCASE challenge. The provided 4 fold partition setup is used for evaluation. All audio signals recorded with the sampling rate of 44.1kHz. To augment the number of training examples, we split given recordings into 1-second length segments. Resultingly, we have total 35,100 samples of audio segments, which contains 2,340 examples per each acoustic environment.

3.2. Preprocessing and Feature Extraction

Each recording files' amplitude is normalized, multiplied by the fraction of the maximum absolute amplitude value before segmentation. We used mono channel source obtained by averaging left and right channel of the signal. We computed Mel-frequency spectrogram with 23ms of Hamming window and half length of overlap and obtained 128 Mel filter bank coefficients for each frame.

We applied ZCA whitening with stochastic masking to extract feature in an unsupervised manner. For computational efficiency, we sampled 5K frames from about 900K of entire training frames to compute covariance matrix and eigenbases. We computed binary masking vector indicating which eigenbases is not masked, by sampling each binary element from the binomial distribution. We set a number of trials as 1 and success probability as 0.9.

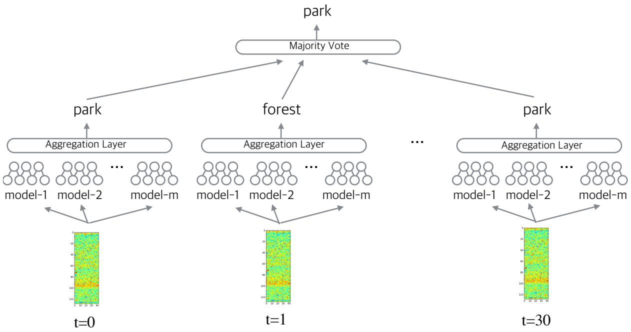


Figure 2: Schema of ensemble classification model.

Table 1: The result of cross-validation. The accuracy of 10 machines is average over 200 trials of randomly choosing 10 machines. The accuracy of 1 machines is average over each machines' test results, and the accuracy of 100 machines is ensemble result of entire machine pool. The values presented this table is the average test accuracy of the 4 folds.

# of Machines	Mean Accuracy(%)				
	Fold1	Fold2	Fold3	Fold4	Avg.
Baseline	-	-	-	-	72.50
1	83.21	78.05	82.08	78.75	80.52
10	**84.62**	78.32	82.78	80.42	81.54
100	84.14	**78.62**	**82.89**	**80.82**	**81.62**

3.3. Base Model Training

The entire structure of the constituent model is following general form of basic CNN, augmented with the unsupervised feature extraction layer as its initial stage. The first layer of the model is linear transformation layer, employing whitening transformation matrix as its weight matrix. It is computed by multiplication of randomly masked eigensystems. The output of this layer is fed into a convolutional layer. We used 64 of 9 × 9 size filters on this layer and compressed its output information with following 2 × 3 size max-pooling layer. We added 3 fully connected layers, each of layers has 1000 of hidden units, then the activation of final fully connected layer propagates to the output layer. We chose all the non-linearity functions as rectified linear unit(ReLU) except the final softmax layer. Except the first feature extraction layer, we applied dropout[10] at every end of the layers. All settings are chosen through cross-validation.

We chose categorical cross entropy between target label, which is encoded with the one-hot coding method, and model softmax output vector. Also weighted penalty terms for the L_2-norm of each weight matrice are added to the loss function for regularization. We weighed 10^{-5} to penalty term except convolution filter weights, which is weighted by 10^{-2}. The batch normalization[11] is applied to each layer, to accelerate training procedure. We chose Adagrad[12] as optimization algorithm for fast fitting, and set the learning rate to 0.01, and the batch size was set as 128. The early stopping strategy is also applied to avoiding overfitting, and patience is set as 10.

3.4. Ensemble Setting

For each fold, we trained 100 constituent networks by bagging and aggregated their output activation by taking median among them.

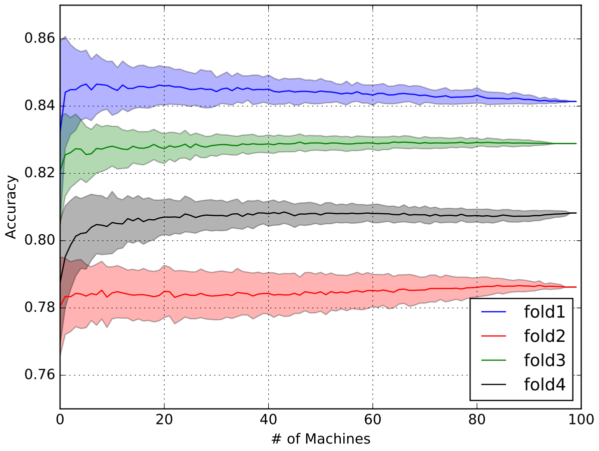


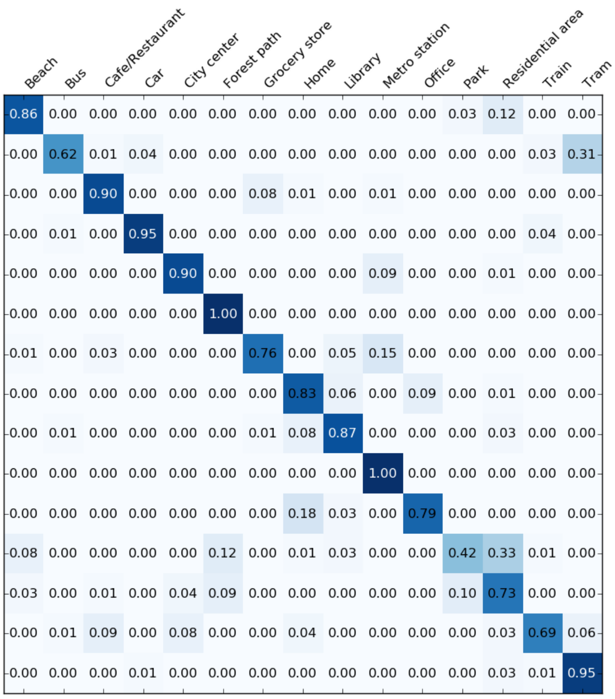
Figure 3: Average accuracy over the number of machines used in ensemble. The solid curve indicates the mean accuracy of ensemble result of randomly sampled machines. The filled area with transparent colors indicates the standard deviation of accuracy.

To verify the effect of the number of constituent models, we conducted the experiment as follows: we repeat 200 trials of randomly choosing N machines from machine pool with replacement, then compute the average of the accuracy of ensemble result. We tried this procedure from $N = 1$ through $N = 100$.

4. RESULT

The result of the experiment is summarized in Table 1 and 2. The proposed method shows about 9% of improvement over the baseline system, which exploited MFCC and GMM in relatively canonical manner. Even with a single constituent model, it showed better performance against the baseline. As shown in Figure 4, the most confused class was Park, and it showed high confusion with Residential Area and Forest Path. Since we did not consider time-domain dependencies between 1-second segmentation, this result might occur because samples of those classes largely sharing the spectral characteristic, but discriminated by acoustic events which are very sparse, such as birds chirp, footstep sounds.

Also, the cross-validation indicates that on most folds bagging method and feature diversity increases classification accuracy of ensemble machine even when the number of constituent growing to 100. As shown in Figure 3, however, the improvement is rapidly saturated when the number of constituents over about 10. These results indicate that the proposed model is not only accurately classifying input acoustic context, but also empirically showing a large-scale ensemble of relatively deep and large neural network model is effective for stabilizing or even boosting the model accuracy. On the evaluation set, the proposed ensemble model achieved 85.4% of overall accuracy, and this result indicates that the cross-validation setting and the ensemble approach have not led the model to be overfit to the given training set.



Beach	Bus	Cafe/Restaurant	Car	City center	Forest path	Grocery store	Home	Library	Metro station	Office	Park	Residential area	Train	Tram
0.86	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.03	0.12	0.00	0.00
0.00	0.62	0.01	0.04	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.03	0.31
0.00	0.00	0.90	0.00	0.00	0.00	0.08	0.01	0.00	0.01	0.00	0.00	0.00	0.00	0.00
0.00	0.01	0.00	0.95	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.04	0.00
0.00	0.00	0.00	0.00	0.90	0.00	0.00	0.00	0.00	0.09	0.00	0.00	0.01	0.00	0.00
0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.01	0.00	0.03	0.00	0.00	0.00	0.76	0.00	0.05	0.15	0.00	0.00	0.00	0.00	0.00
0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.83	0.06	0.00	0.09	0.00	0.01	0.00	0.00
0.00	0.01	0.00	0.00	0.00	0.00	0.01	0.08	0.87	0.00	0.00	0.00	0.03	0.00	0.00
0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00
0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.18	0.03	0.00	0.79	0.00	0.00	0.00	0.00
0.08	0.00	0.00	0.00	0.00	0.12	0.00	0.01	0.03	0.00	0.00	0.42	0.33	0.01	0.00
0.03	0.00	0.01	0.00	0.04	0.09	0.00	0.00	0.00	0.00	0.00	0.10	0.73	0.00	0.00
0.00	0.01	0.09	0.00	0.08	0.00	0.00	0.04	0.00	0.00	0.00	0.00	0.03	0.69	0.06
0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.03	0.01	0.95

Figure 4: Confusion matrix of the result of cross-validation.

5. CONCLUSION

We proposed a DNN based large scale ensemble model, and also proposed stochastic feature whitening for diversifying constituent networks within ensemble model. By the experiment conducted with DCASE 2016 challenge dataset, we showed proposed model gives a significant improvement of classification accuracy compared to the baseline algorithm.

Also, we derived a number of further objectives and potential improvements for the future work. 1) if the selection of principal component actually affects classification accuracy of ensemble model, one might be able to search an optimal subset of feature for improving ensemble performance. 2) Also, the diversity of constituent model structure can affect the performance of ensemble model.

6. REFERENCES

[1] D. Barchiesi, D. D. Giannoulis, D. Stowell, and M. D. Plumbley, “Acoustic Scene Classification: Classifying environments from the sounds they produce,” *IEEE Signal Processing Magazine*, 2015.

[2] D. Stowell, D. Giannoulis, E. Benetos, M. Lagrange, and M. D. Plumbley, “Detection and Classification of Acoustic Scenes and Events,” *IEEE Transactions on Multimedia*, 2015.

[3] J. Nam, Z. Hyung, and K. Lee, “Acoustic scene classification using sparse feature learning and selective max-pooling by event detection.”

[4] V. Bisot, R. Serizel, S. Essid, and G. Richard, “Acoustic scene

classification with matrix factorization for unsupervised feature learning," *Icassp*, pp. 6445–6449, 2016.

[5] Y. Petetin, C. Laroche, and A. Mayoue, "Deep neural networks for audio scene recognition," in *Signal Processing Conference (EUSIPCO), 2015 23rd European*. IEEE, 2015, pp. 125–129.

[6] D. Opitz and R. Maclin, "Popular Ensemble Methods: An Empirical Study," *Journal of Artificial Intelligence Research*, vol. 11, pp. 169–198, 1999.

[7] B. Leo Breiman and L. B. Eiman, "Bagging Predictors," 1994.

[8] R. E. Schapire, "The Boosting Approach to Machine Learning An Overview," *Nonlinear Estimation and Classification*, 2003. [Online]. Available: www.research.att.com/

[9] A. Mesaros, T. Heittola, and T. Virtanen, "TUT Database for Acoustic Scene Classification and Sound Event Detection."

[10] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A Simple Way to Prevent Neural Networks from Overfitting," *Journal of Machine Learning Research*, vol. 15, pp. 1929–1958, 2014.

[11] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *arXiv preprint arXiv:1502.03167*, 2015.

[12] J. Duchi, J. B. Edu, E. Hazan, and Y. Singer, "Adaptive Subgradient Methods for Online Learning and Stochastic Optimization *," *Journal of Machine Learning Research*, vol. 12, pp. 2121–2159, 2011.