DEEP NEURAL NETWORK BASED LEARNING AND TRANSFERRING MID-LEVEL AUDIO FEATURES FOR ACOUSTIC SCENE CLASSIFICATION

Seongkyu Mun[*]*, Suwon Shon*[**]*, Wooil Kim*[***]*, David K. Han*[****]*, Hanseok Ko*[**]

[*] Dept. of Visual Information Processing, Korea University, Seoul, Korea
[**] School of Electrical Engineering, Korea University, Seoul, Korea
[***]Dept. of Computer Science and Engineering, Incheon National University, Incheon, Korea
[****] Office of Naval Research, Arlington, VA, USA
{skmoon,swshon}@ispl.korea.ac.kr, wikim@inu.ac.kr, ctmkhan@gmail.com, hsko@korea.ac.kr

ABSTRACT

Deep Neural Network (DNN) based transfer learning has been shown to be effective in Visual Object Classification (VOC) for complementing the deficit of target domain training samples by adapting classifiers that have been pre-trained for other large-scaled DataBase (DB). Although there exists an abundance of acoustic data, it can also be said that datasets of specific acoustic scenes are sparse for training Acoustic Scene Classification (ASC) models. By exploiting VOC DNN's ability of learning beyond its pre-trained environments, this paper proposes DNN based transfer learning for ASC. Effectiveness of the proposed method is demonstrated on the database of IEEE DCASE Challenge 2016 Task 1 and home surveillance environment via representative experiments. Its improved performance is verified by comparing it to prominent conventional methods.

Index Terms— Transfer learning, deep neural network, acoustic scene classification, mid-level feature

1. INTRODUCTION

Acoustic Scene Classification (ASC) is a field of autonomously recognizing different environments via sounds. It has recently attracted considerable attention due to a variety of new applications and potential uses [1-4]. Most of top 10 ranked approaches in the IEEE Detection and Classification of Acoustic Scenes and Events (DCASE) Challenge 2016 used DNN based approaches such as Recurrent Neural Network (RNN) [5], Convolutional Neural Network (CNN) [6] and DNN based bottleneck features [7].

As shown in the results of DCASE 2016, various structures of DNN have been implemented in ASC applications. Successes of these approaches have been attributed to effectiveness of novel structures proposed by these teams.

However, one of the powerful features in DNN based approaches, which is called 'transfer learning', has not been exploited yet in the ASC researches. The "transfer learning" scheme aims at transferring knowledge between the source domain used for pre-training and the target domain of interest [8]. In computer vision, transfer learning overcomes deficit of target domain training samples by adapting classifiers that are pre-trained for other large-scaled DB [9-10]. In recent VOC fields, CNN based supervised transfer learning methods pre-train lower layers in source domain first and then transfer these lower layer parameters for training target domain categories [11].

Transfer learning can address the issue of ASC DB being significantly smaller compared to that of other audio signal applications such as speech recognition, natural language processing or speaker recognition. Therefore, this paper proposes to pre-train a classifier with large-scaled source domain DB and transfer the parameters for training with target DB. To the best of our knowledge, this is the first use of transfer learning in acoustic scene classification.

2. PROPOSED APPROACH

The process of training ASC system using transfer learning is depicted in Figure 1. Similar to the previous VOC research [11], the internal layers of the DNN can act as a mid-level feature extractor (or refiner), which is pre-trained on source domain task and then re-used on other target domain tasks. Details of the source/target domain tasks are listed on Section 3.

2.1. Network structure

For the source task, the network is composed of four hidden full connected layers which use a 'ReLU' non-linear function and a one output layer with a 'SoftMax' function. For the target task, similar to the transfer learning in VOC, output layer of the pre-trained network is removed and two hidden fully connected layers and a single output layer are added for adaptation. As depicted in Figure 1, output vector Y4 is used as input of target task layer TL#1. Note that Y4 is

978-1-5090-4117-6/17/$31.00 ©2017 IEEE

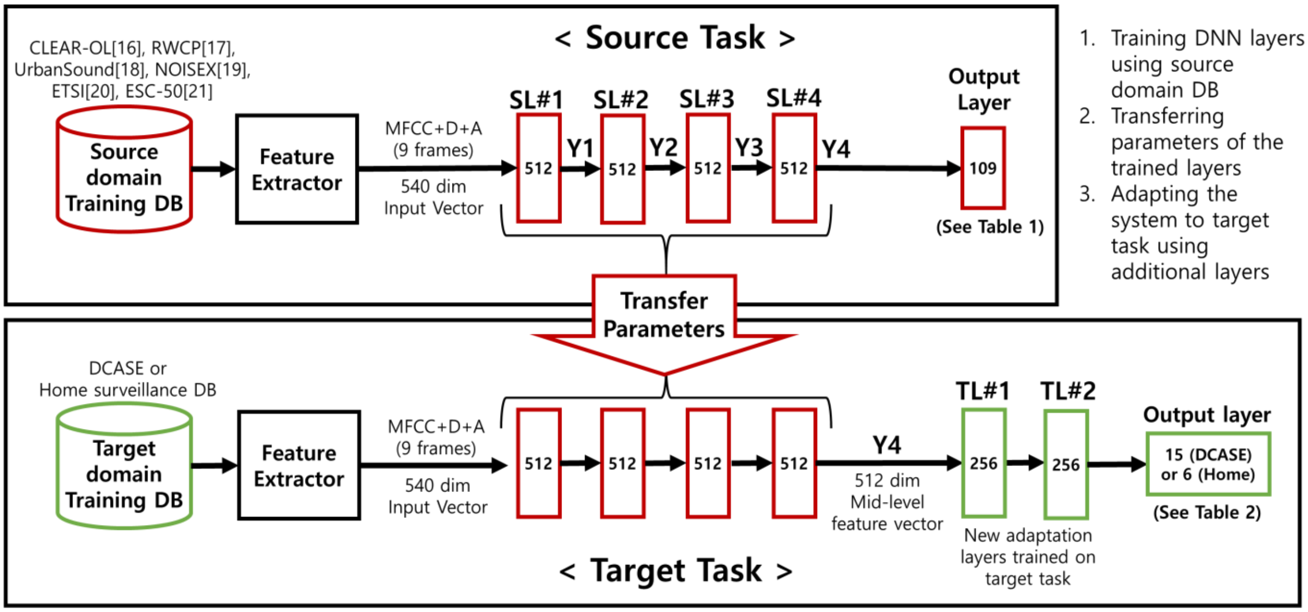
< Source Task >
CLEAR-OL[16], RWCP[17], UrbanSound[18], NOISEX[19], ETSI[20], ESC-50[21]
Source domain Training DB
Feature Extractor
MFCC+D+A (9 frames)
540 dim Input Vector
SL#1
SL#2
SL#3
SL#4
512
Y1
Y2
Y3
Y4
Output Layer
109
(See Table 1)
1. Training DNN layers using source domain DB
2. Transferring parameters of the trained layers
3. Adapting the system to target task using additional layers
Transfer Parameters
DCASE or Home surveillance DB
Target domain Training DB
Feature Extractor
MFCC+D+A (9 frames)
540 dim Input Vector
512 dim Mid-level feature vector
TL#1
TL#2
256
Output layer
15 (DCASE) or 6 (Home)
(See Table 2)
New adaptation layers trained on target task
< Target Task >

Fig. 1. Proposed transfer learning based DNN framework for acoustic scene classification

obtained from the output of non-linear hidden layers in source domain. Since these non-linear layers are pre-trained for classifying various classes of source task, the layer outputs may capture discriminative features of sounds [1, 12-13]. In target task, these mid-level features are adapted to target domain labels by being used as inputs for training additional two hidden layers.

In summary, the parameters of layers SL#1~4 are first trained on the source task, then transferred to the target task and kept fixed. Only the adaptation layers are trained on the target task training data as described next.

3. EXPERIMENTAL SETTINGS

The 'Caffe' based deep learning framework was used for training [14]. The training procedure periodically evaluated the cross-entropy objective function on a subset of the training set and on a validation set (rate for train/validation set was 4:1 in training DB). The initial learning rates are set to 0.01 and the network is trained until the training cross-entropy is stabilized. The learning rates are then divided by 10 and the training procedure repeats. The momentum parameter and weight decay were set to 0.9 and 0.0005 respectively.

The input features were 60-dimensional MFCC features including both delta and acceleration of MFCC coefficients (including the 0th order coefficient). Input layer was composed of a concatenation of 9 input frames (the current frame and the four previous and four next frames) resulting in 540 input units.

3.1. Source domain task

The 'ImageNet' DB which is commonly used for source domain in VOC has 15 million images and 22,000 classes [15]. However, there is no single ASC DB as large-scaled as ImageNet. Therefore, web accessible six different acoustic DB sets are merged for source domain DB in this work [16-21]. Since each DB set has a different wave length for their own classes, wave length of each class is normalized to about 800 seconds. The classes over 800 seconds are randomly cut into 800 seconds and those under 800 seconds are reproduced by filtering room impulse response of office environment [22] as like as re-recording sounds in the office environment. Detail descriptions are shown in Table 1.

3.2. Target domain task

IEEE DCASE 2016 Challenge Task 1 and home surveillance DB of real life recordings are individually used for each target domain task. The DCASE DB [23], also known as TUT acoustic scenes 2016, contains 15 different acoustic scenes (see Table 2). The TUT DB consists of two subsets: training dataset and evaluation dataset. For each acoustic scene, 78 segments were included in the training dataset and 26 segments were kept for evaluation (each segment is 30 seconds of wave file). Training set contains in total 9h 45mins of audio, and evaluation set 3h 15mins.

The other target task using home surveillance DB is for ASC application which has significantly small-scaled target DB. The DB have 6 different acoustic event (see Table 2) and each event consists of 150 segments for training set and 50 segments for evaluation set (each segment is 3 seconds of wave file). Training set contains in total 45mins of audio, and evaluation set 15mins.

Table 1. Source domain database description

DB set	Contents	# of classes
Clear-OL [16]	Alert, cough, door slam, drawer, key, keyboard, knocking, laughing, mouse, page turn, pen drop, phone, printer, speech, switch, clear throat	16
RWCP [17]	Air-cap, bell, break stick, buzzer, castanet, ceramic collision, clap, clock ringing, coin, cymbals, drum, dryer, grinding coffee, kara, maracas, metal collision, article dropping, plastic collision, pump, punch stapler, rubbing, shaver, spray, string, tambourine, toy, whistle, wood collision	28
UrbanSound [18]	Air-conditioner, dog bark, drilling, engine idling, car horn, jackhammer, children playing, siren, street music, shot	10
NOISEX [19]	Voice babble, destroyer noises, F16 noise, Factory noise, tank noise, machine gun, pink noise, Volvo 340, white noise	9
ETSI noise [20]	Living room, kindergarten, playing sports, pub, traffic, shopping, schoolyard	7
ESC-50 [21]	Airplane, breathing, brushing teeth, can opening, cat, chainsaw, chirping birds, church bells, clapping, clock alarm, clock tick, coughing, cow, crackling fire, crickets, crow, door - wood creaks, door knock, drinking – sipping, engine, fireworks, footsteps, frog, hand saw, helicopter, hen, insects (flying), pig, pouring water, rooster, sea waves, sheep, sneezing, snoring, thunderstorm, toilet flush, vacuum cleaner, washing machine, wind	39
The similar classes were merged and target domain related classes were excluded /16KHz, 16bit		**Total : 109**

Table 2. Target domain database description

Experiment #1 : IEEE DCASE 2016 Challenge Task 1	
15 Classes	**Bus** - traveling by bus in the city, **Cafe / Restaurant** - small cafe/restaurant, **Car** - driving or traveling as a passenger, in the city, **City center, Forest path, Grocery store, Home, Lakeside beach, Library, Metro station, Office** - multiple persons, typical work day, **Residential area, Train, Tram, Urban park**
Experiment #2 : Home surveillance environmental DB	
6 Classes	Crying kid, Glass breaking, Water drop(rain), Doorbell, Home appliance beeping sound, Scream

4. EXPERIMENTAL RESULTS

This paper compared the average accuracies over all scenes for the conventional methods and the proposed transfer learning based DNN. Tables 3 and 4 show the segment-based classification accuracy and the proposed method achieved higher accuracy than other approaches in ASC.

In the DCASE 2016 experiment, the baseline accuracy of audio scene classification task in the Challenge [23], which was based on MFCCs and GMMs, was 77.2% and the best Challenge result among full-connected DNN based approaches was 85.6% (The approach with additional GMM classifiers using DNN classification results) [24]. Utilizing the information transferred from source domain task, the accuracy of proposed method was 86.3% without using the additional GMM classifier.

In the home surveillance experiment, due to significantly small size of training DB, utilizing the transfer learning from source domain task outperformed conventional methods in all of classes. Based on the experimental results, the transfer learning can be used for complementing the deficit of target domain training DB in ASC applications.

Table 3. Classification result of Experiment #1

Acc. Rate [%]	GMM (DCASE)	DNN-GMM (DCASE)	DNN	**DNN***
Beach	84.6	92.3	92.3	96.2
Bus	88.5	100.0	84.6	96.2
Café	69.2	61.5	61.5	76.9
Car	96.2	100.0	96.2	88.5
City cent.	80.8	88.5	85.5	88.5
Forest	65.4	88.5	84.6	96.2
Grocery	88.5	96.2	96.2	92.3
Home	92.3	84.6	76.9	82.1
Library	26.9	57.7	53.8	57.7
Metro	100	80.8	65.4	76.9
Office	96.2	100.0	84.6	92.3
Park	53.8	92.3	80.8	92.3
Resid.	88.5	80.8	80.8	84.6
Train	30.8	61.5	53.8	96.2
Tram	96.2	100.0	76.9	76.9
Average	77.2	85.6	78.3	**86.3**

- GMM(DCASE) : Base line system of DCASE, GMM-MFCC
- DNN-GMM(DCASE) : The best result among the full connected DNN based approach in DCASE challenge [24]
- DNN : DNN structure in Figure 1 without transfer learning (Conventional training with target domain only)
- **DNN* : DNN structure in Figure 1 with transfer learning (Proposed method)**

Table 4. Classification result of Experiment #2

Acc. Rate [%]	GMM (DCASE)	DNN	**DNN***
Crying kid	86	90	98
Glass breaking	86	84	100
Water drop	88	88	92
Doorbell	90	88	92
Home appliance	88	90	96
Scream	86	90	92
Average	87.3	88.3	**95.0**

5. CONCLUSION AND FUTURE WORKS

This paper proposed a novel DNN framework with transfer learning. The proposed mid-level features derived from pre-trained by source domain task yielded overall improved the acoustic scene classification performance in DCASE 2016 and home surveillance experiments.

Additional work will investigate effective transfer learning methods for Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN) and refining source domain DB in terms of acoustic labels, volume and format is needed.

6. REFERENCES

[1] S. Mun, S. Shon, W. Kim and H. Ko, “Deep Neural Network Bottleneck Features for Acoustic Event Recognition”, *INTERSPEECH 2016 Proceedings–Annual Conference of the International Speech Communication Association*, San Francisco, September 8–12, San Francisco, USA, 2016.

[2] S. Park, J. Rho, M. Shin, D. K. Han, and H. Ko, “Acoustic feature extraction for robust event recognition on cleaning robot platform”, *IEEE Conference on Consumer Electronics*, pp. 149-150, 2014.

[3] I. McLoughlin, et al., "Robust sound event classification using deep neural networks." *IEEE/ACM Transactions on Audio, Speech, and Language Processing,* vol. 23 no.3, pp. 540-552, 2015

[4] H. Zhang, I. McLoughlin and S, Yan, “Robust sound event recognition using convolutional neural networks”, *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* p. 559-563, 2015.

[5] S. Bae, I. Choi and N. Kim, "Acoustic Scene Classification Using Parallel Combination of LSTM and CNN", *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2016 Workshop (DCASE2016)*, pp.11-15, 2016. ,

[6] M. Valenti, A. Diment, G. Parascandolo, S. Squartini and T. Virtanen, "DCASE 2016 Acoustic Scene Classification Using Convolutional Neural Networks", *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2016 Workshop (DCASE2016)*, pp.95-99, 2016. ,

[7] S. Mun, S. Park, Y. Lee and H. Ko, “Deep Neural Network Bottleneck Feature for Acoustic Scene Classification”, *DCASE2016 challenge technical report,* 2016

[8] S. Pan and Q. Yang, “A survey on transfer learning”, *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 10 pp. 1345–1359, 2010.

[9] Y. Aytar, and Z. Andrew, "Tabula rasa: Model transfer for object category detection." *2011 International Conference on Computer Vision (ICCV),* pp. 2252-2259, 2011.

[10] T. Tommasi, F. Orabona and B. Caputo, “Safety in numbers: Learning categories from few examples with multi model knowledge transfer”, *IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* pp. 3081-3088, 2010.

[11] M. Oquab, et al., "Learning and transferring mid-level image representations using convolutional neural networks." *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1717-1724, 2014.

[12] Q. Le, M. Ranzato, R. Monga, M. Devin, K. Chen, G. Corrado, J. Dean, and A. Ng. “Building high-level features using large scale unsupervised learning”, *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 8595-8598, 2012.

[13] M. Zeiler, G. Taylor, and R. Fergus, “Adaptive deconvolutional networks for mid and high level feature learning”, *2011 International Conference on Computer Vision (ICCV),* pp. 2252-2259, 2011.

[14] Y. Jia et al., “Caffe: Convolutional architecture for fast feature embedding”, *In Proceedings of the 22nd ACM international conference on Multimedia*, pp. 675-678. 2014.

[15] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “ImageNet: A Large-Scale Hierarchical Image Database”, *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1717-1724, 2009.

[16] A. Temko, R. Malkin, C. Zieger, D. Macho, C. Nadeu, and M. Omologo, “CLEAR evaluation of acoustic event detection and classification systems,” *In Proc CLEAR*, pp. 311–322, 2007.

[17] S. Nakamura, K. Hiyane, F. Asano, T. Yamada and T. Endo, “Data collection in real acoustical environments for sound scene understanding and hands-free speech recognition,” *in Proc. EUROSPEECH*, pp. 2255–2258. 1999.

[18] J. Salamon, C. Jacoby and J. P. Bello, “A dataset and taxonomy for urban sound research” *In Proceedings of the 22nd ACM international conference on Multimedia,* pp. 1041-1044, 2014.

[19] A. Varga and H. J. M. Steeneken, “Assessment for automatic speech recognition II: NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems,” *Speech Communication*, vol. 12, no. 3, pp. 247–251, 1993.

[20] European Telecommunications Standards Institute, “ETSI: EG 202 396-1 v1.2.2,” 2008.

[21] K. Piczak, “ESC: Dataset for environmental sound classification”, *In Proceedings of the 23rd ACM international conference on Multimedia*, pp. 1015-1018, 2015.

[22] J. B. Allen, and D. A. Berkley, “Image method for efficiently simulating small-room acoustics”, *The Journal of the Acoustical Society of America,* vol. 65, no.4, pp. 943-950, 1979.

[23] A. Mesaros, T. Heittola and T. Virtanen, “TUT database for acoustic scene classification and sound event detection”, *In 24th European Signal Processing Conference*, 2016.

[24] G. Takahashi, et al., “Acoustic Scene Classification Using Deep Neural Network and Frame-Concatenated Acoustic Feature”, *DCASE2016 challenge technical report,* 2016