Convolutional Neural Network with Multi-Task Learning Scheme for Acoustic Scene Classification

Tin Lay Nwe, Tran Huy Dat and Bin Ma
Institute for Infocomm Research, I2R, Singapore
tlnma, hdtran, mabin@i2r.a-star.edu.sg

***Abstract*—Deep Neural Network (DNN) with Multi-Task Learning (MTL) methods have recently demonstrated significant performance gains on a number of classification, detection, recognition tasks compared to conventional DNN. DNN with MTL framework involves cross-task and within-task knowledge sharing layers. MTL methods have benefit for regularization effect from the cross-task knowledge sharing layers. And, within-task knowledge sharing layers allow MTL based DNN to learn information to optimize the performance for individual task. We formulate our acoustic scene classification in MTL framework using Convolutional Neural Network to learn information specific to different types of environment. We conduct experiments using DCASE2016 dataset. Proposed approach achieves 83.8% accuracy to classify 15 acoustic scene classes.**

I. Introduction

Acoustic scene classification is to classify among different environmental sounds that occur for a period of time. If a machine can automatically identify the type of its environment from audio, e.g. through acoustic events inside the recording, it will be useful for many applications such as security surveillance and context-aware services [1]. Acoustic scene classification is a challenging problem that has been studied for several years. Identification of features which is relevant to scene classification is a challenging task due to heterogeneity in nature of scene audio [2] and shared acoustic characteristics between different scene types. For example, in scene classes between forest path and park, sounds such as bird singing and rustling leaves are common to both scene types.

Several studies have been carried out in the area of acoustic scene classification for the past few decades. And, many approaches have been proposed for acoustic scene classification. These include developing various feature extraction algorithms specific to the acoustic scenes and employing machine learning algorithms for modeling acoustic scenes. Some of the explored features are spectrum image features [3], Mel frequency cepstral coefficients (MFCC) [4], log-frequency filter banks [3], time dependent temporal features, frequency dependent spectral features and combined Time-Frequency features [5], [6]. As for machine learning algorithms, several studies employ Gaussian Mixture Model (GMM) [2] and Support Vector Machine (SVM) [7] for modeling environmental sounds. In addition, deep learning methods have gained attention recently, and researchers have started to apply deep learning methods such as DNN [8] and CNN [1] for acoustic scene classification task. Convolution Neural Network (CNN) is one of the most successful modeling method in several acoustic scene classification tasks.

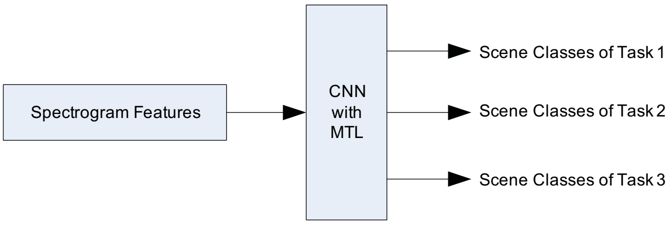

Fig. 1. System block diagram for classification of acoustic scenes.

Recently, Deep Neural Network (DNN) with Multi-Task Learning (MTL) has attracted much attention on classification, detection, prediction and recognition tasks. Multi-Task Learning (MTL) is to learn multiple tasks which are related to each other simultaneously. MTL method provides regularization and optimization effect that leads to achieve better system performance. There are various applications that benefit from MTL approach. These includes multilingual character recognition [9], semantic classification [10] and several others. In this paper, we propose to employ a multi-task representation learning framework based on CNN for acoustic scene classification. In our proposed approach, we formulate the acoustic scene classification in MTL framework by grouping scene classes to form individual task which corresponds to a particular environment. The block diagram of the proposed system is presented in Figure 1.

The rest of the paper is organized as follows. Section II presents the database used in our acoustic scene classification experiments. Section III presents the conventional and multi-task learning based convolutional neural networks. Section IV mentions feature used as input to the networks. Section V presents experimental results and discussions. Finally, section VI concludes the paper.

II. TUT Acoustic Scenes 2016 Data Set

TUT Acoustic Scenes 2016 dataset consists of 15 different acoustic scenes: lakeside beach, bus, cafe/restaurant, car, city center, forest path, grocery store, home, library, metro station, office, urban park, residential area, train and tram. The data set includes 13 hours of audio. Database is divided into training and test sets. Each scene class has total duration of 39 minutes for training set and 13 minutes for test set. The length of the each audio segment is 10 seconds. We notice that acoustic scenes are related to some types of environments. For example, the scenes such as library, office, cafe/restaurant are indoor acoustic scenes. And, the scenes such as car, metro station, bus are transport related outdoor acoustic scenes.

III. Convolutional Neural Networks (CNN)

Convolutional Neural Network (CNN) has shown compelling and efficiency for many classification tasks such as image classification[11], multivariate time series classification [12], acoustic scene classification [1] and several others. CNN is in fact feature extractor as its final output comes from softmax layer. The basic block is a convolutional layer and uses ReLu or Rectilinear units as activation as mentioned in equation (1) [13].

$$\begin{aligned} y &= W \otimes x + b \\ h &= ReLU(y) \end{aligned} \tag{1}$$

$\otimes$ is the convolution operator. We build the baseline conventional CNN and CNN with MTL framework in the following sections.

A. Conventional CNN

Conventional CNN is built by stacking three convolution blocks with the filter sizes $\{128, 256, 512\}$ respectively in each block. Each convolution layer is followed by a maxpooling layer. Finally, the scene classification output is produced by a softmax layer. In the following section, we discuss about CNN with MTL framework.

B. CNN with Multi-Task Learning

Multi-task learning is simultaneously learning multiple classification/predition tasks that are related to one another [14]. In other words, in the framework of multi-task learning, multiple related classification tasks are learned jointly and information is shared across the tasks. MTL framework involves two parts. In the first part, knowledge is shared or learned jointly among many tasks. In the second part, information or knowledge for the specific task is learned separately. In other words, common knowledge learned in the first part is adapted to a specific task in the second part. As common information is learned jointly among different tasks, MTL framework results in better generalization effect than independently learning individual tasks. When generalized knowledge is further adapted to the specific task, modeling process results in better optimization effect. To take advantage on generalization and optimization effect in classifier modeling, we formulate our scene classification problem in MTL framework as discussed in the following section.

1) Environment based Scene Grouping for CNN with Multi-Task Learning: As we observe in Section II, some numbers of scene classes belong to a particular type of environment. Here, we observe that acoustic scenes can be categorized into 3 types of environments. The first environment is outdoor recreational environment. We include the scenes such as beach, park, forest path, city center and residential area in the first environment and refer to as 'Task 1' (outdoor1). The second one is related to transport related outdoor environment. We include the scenes such as car, metro station, bus, train and tram in this group. We will refer the second group as Task 2 (outdoor2). The third group includes audio recorded in indoor environments such as library, office, cafe/restaurant, grocery store and home. We will refer the third group as Task 3 (indoor). All the task groups and respective acoustic scenes are listed in Table I.

As mentioned in Section I, acoustic scenes (example, forest path and park) which belong to outdoor recreational environment share acoustic characteristics such as bird singing and rustling leaves. Network needs to learn knowledge to discriminate these similar acoustic classes correctly. In the above, we divide 15 acoustic scenes into 3 groups based on types of environment. We consider each environment based scene group as an individual task and formulate multi-task learning framework for acoustic scene classification. Grouping based on environment is useful as acoustic scenes have similar acoustic characteristics within a particular environment type. Network can learn information on how to optimize the discrimination ability on scene classes which is related by environment type.

The architecture of our multi-task CNN model is shown in Figure 2. Knowledge from 3 different tasks or environments is shared in lower layers of the network. This allows the network to learn the knowledge which is common to all 3 types of environment. In upper layers, the network learns the knowledge which is specific to individual environment. We would like to compare the environment based scene grouping with random scene grouping for MTL based CNN. In the following section, we present grouping scenes randomly for MTL based CNN.

2) Grouping Scenes Randomly for CNN with Multi-Task Learning: To examine the effectiveness of grouping scenes based on environment types, we compare it with random scene grouping. Hence, we divide 15 scene classes into 3 groups to form 3 tasks. Each task includes 5 acoustic scenes which are randomly selected. In lower layers of MTL based CNN, knowledge sharing effect is the same as in environment based scene grouping. However, in upper layers, knowledge sharing within specific task is among the scenes which correspond to several types of environment. In the following, we present input to the CNN networks.

IV. Generation of Spectrogram

Each scene audio is characterized by their sound intensity by a set of spectral parameters. And, we generate a spec-

TABLE I
Grouping of the 15 acoustic scenes into 3 different environment based tasks

Task group1 (outdoor1)	Task group2 (outdoor2)	Task group3 (indoor)
beach	car	library
park	metro station	office
forest path	bus	cafe restaurant
city center	train	grocery store
residential area	tram	home

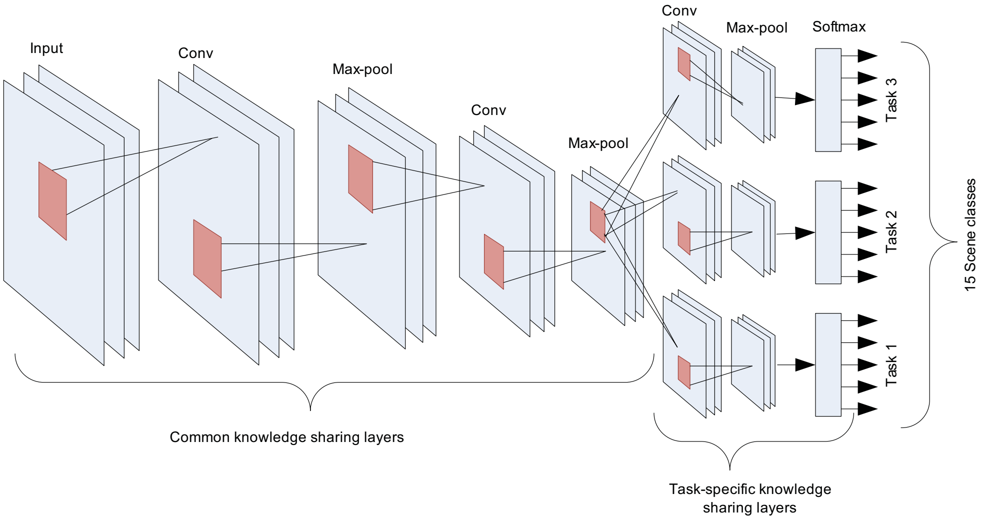

Fig. 2. *Architecture of the multi-task CNN for acoustic scene classification: The lower layers are shared across all tasks while upper layers are for task-specific knowledge learning.*

trogram for each acoustic scene clip. A spectrogram is a visual representation of the frequency spectrum over time, and the spectrograms of most sounds have several distinguishing features. In addition, the spectrogram contains frequency and amplitude information over time, and it is important acoustic information to distinguish different types of sounds.

For each 10 second audio clip which is stereo format, we cut into 3 seconds non-overlapped segments. Then, we compute the log power spectrogram from each of 4 channels from each 3 second segment. Four channels are left, right, sum and difference channels. We compute 12 log power spectrograms from each of 10 second audio clip. Window size and frame rate to compute log power spectrogram are 40 ms and 20 ms respectively. The input into the neural networks was normalized to remove any variance.

V. Experiment

Experiments are conducted to evaluate the effectiveness of the formulating scene classification in MTL framework based on CNN. We use the corpus mentioned in Section II in our experiments. There are two experiments for MTL based CNN. One is for environment based scene grouping and another one is for randomly selected scene grouping. We also conduct experiment with conventional CNN to compare our proposed MTL based CNN with conventional CNN.

As we have mentioned in Section IV, we have 12 log power spectrograms for each 10 second clip. We obtain a set of 12 scores for each of 15 scenes from 12 log power spectrogram inputs to the network. Then, we compute the mean of the 12 scores to make the classification decision for each 10 second clip. We calculate Unweighted Average Recall (UAR) [15] to measure the performance of our scene classification systems. In UAR performance matrix, we calculate accuracy for each scene and then, we obtain mean UAR by taking average over accuracies of all 15 scene classes. In the following sections, we present the experiments and results. In all experiments, we use log power spectrogram mentioned in Section IV as input to the networks.

A. *Effect of Using Conventional CNN*

We conduct scene classification experiments using conventional CNN. We employ 2 Dimensional CNN (2D-CNN) in our experiments. The following are the description of the conventional CNN used in our experiment. CNN has 3 convolutional layers, 3 maxpooling layers and 1 softmax layer. Convolution layers use Relu or Rectilinear units as activation function. Each convolution layer is followed by a max pooling layer. Output from third convolution layer is flattened and fed to fully connected softmax layer. Numbers of output neurons of the softmax layer is 15 which is total numbers of acoustic scene classes to be classified. Numbers of filters for 3 convolutional layers are 128, 256 and 512 respectively. The sizes of the 3 maxpooling layers are $5X5$, $2X2$ and $15X3$ respectively. We use first-order gradient-based optimization of stochastic objective function as the optimizer. Learning rate is 0.0007. Drop out rate is 0.75 and batch size is 128. We implement our CNN model using tensorflow libraries in Python. We conduct experiments to classify 15 acoustic scene types with the above mentioned conventional CNN. We achieve UAR of 80.9% as mention in 2^{nd} row of Table II. The results demonstrate that CNN with log power spectrogram is helpful for acoustic scene classification.

TABLE II
Unweighted Average Recall (UAR [%]) of Scene Classification Systems

CNN Architecture	Mean UAR
Baseline CNN	80.9
CNN with MTL (scene grouping by Environment)	83.8
CNN with MTL (scene grouping randomly)	82.2

B. *Effect of MTL based CNN With Grouping Scenes Based on Environment*

As we have mentioned in Section III-B, MTL based deep learning methods have advantages on generalization and optimization effect. In addition, its framework allows the network to achieve discrimination capability among scenes which have similar acoustic characteristics within a specific environment. Hence, we formulate acoustic scene classification problem into CNN based MTL framework by grouping scenes based on environment types as mentioned in Section III-B. We conduct the experiments to observe the advantages of using MTL based CNN with scene grouping by environment. The two lower convolutional layers are the same as in conventional CNN of previous Section. However, the third convolutional layer and softmax layer are designed for individual task separately as mentioned in Section III-B and illustrated in Figure 2. Network parameters are the same as in conventional CNN. Acoustic scene classification results in terms of UAR using the above CNN set-up is shown in 3^{rd} row of Table II. The results show that proposed approach improves UAR 2.9% absolute over conventional CNN. And, we can observe that formulating scene classification problem in MTL framework is really effective. We also present UAR of individual scene classes in Table III. We notice that we achieve high UAR for all the classes except park which has very similar acoustic characteristics with forest path.

C. *Effect of MTL based CNN With Grouping Scenes Randomly*

In this section, we observe the effect of grouping scenes randomly and compare it with environment based scene grouping of previous section. As mentioned in section III-B2, knowledge sharing mechanism is the same between random task grouping and environment based task grouping in lower layers of the network. However, in upper layers of the network, knowledge sharing mechanism within specific-task is different. For environment based grouping, scenes within a specific task are related to a particular environment. But, in random grouping, scenes within a task belong to different types of environments. We would like to observe the effect of knowledge sharing mechanism on scene classification performance. Hence, we conduct experiments using the tasks with randomly selected scenes on CNN with MTL method. Average classification accuracy over 15 scene classes is presented in 4^{th} row of Table II. Random scene grouping can not perform better than the environment based scene grouping. However, it is better than the conventional CNN. The results further confirm that formulating scene classification problem in MTL framework is effective.

TABLE III
UAR OF INDIVIDUAL SCENE CLASSES FOR ENVIRONMENT BASED SCENE GROUPING AND CNN WITH MTL METHOD

Task 1 (outdoor1)	UAR	Task 2 (outdoor2)	UAR	Task 3 (indoor)	UAR
beach	79.5	car	100	library	73.1
park	55.1	metro station	88.5	office	98.7
forest path	100	bus	94.9	cafe restaurant	75.6
city center	94.9	train	76.9	grocery store	93.6
residential area	70.5	tram	88.5	home	66.7
avg UAR	80		89.76		81.54

VI. Conclusions

We have presented a Convolutional Neural Network with Multi-Task Learning architecture to perform classification for 15 acoustic scenes. The acoustic scenes are grouped based on types of environment to formulate the scene classification problem into multi-task learning framework. The proposed framework benefit for regularization effect from the cross-task knowledge sharing layers and system performance optimization effect from within-task knowledge sharing layers. The experiments show that CNN with MTL framework outperforms baseline CNN and advantage is more significant when scene classes are grouped based on environment types.

References

[1] L. Hertel, H. Phan and A. Mertins, ”Classifying Variable-Length Audio Files with All-Convolutional Networks and Masked Global Pooling”, in *Workshop on Detection and Classification of Acoustic Scenes and Events 2016*, Budapest, Hungary,2016

[2] M. Mulimani and S. G. Koolagudi, Acoustic scene classification using MFCC and MP features in *Detection and Classification of Acoustic Scenes and Events 2016*, Budapest, Hungary,2016.

[3] H. Zhang, I. McLoughlin, and Y. Song, Robust sound event recognition using convolutional neural networks, in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 559563.

[4] E. Cakir, T. Heittola, H. Huttunen, and T. Virtanen, Polyphonic sound event detection using multi label deep neural networks, in *Int. Joint Conf. Neural Networks IJCNN*, 2015.

[5] S. Chu, S. Narayanan, and C. -C. J. Kuo, Environmental sound recognition with timefrequency audio features, *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 6, pp. 11421158, 2009.

[6] B. Ghoraani and S. Krishnan, Timefrequency matrix feature extraction and classification of environmental audio signals, *IEEE transactions on audio, speech, and language processing*, vol. 19, no. 7, pp. 21972209, 2011.

[7] L. Vincent and A. Joakim, ”Binaural Scene Classification with Wavelet Scattering”, in *Workshop on Detection and Classification of Acoustic Scenes and Events 2016*, Budapest, Hungary,2016

[8] R. Patiyal and P. Rajan, ”Acoustic Scene Classification Using Deep Learning”, in *Detection and Classification of Acoustic Scenes and Events 2016*, Budapest, Hungary,2016

[9] Y. Yang and T. M. Hospedales, ”Deep Multi-task Representation Learning: A Tensor Factorisation Approach,” *CoRR* abs/1605.06391 (2016)

[10] X. Liu, J. Gao, X. He, L. Deng, K. Duh and Y. Y. Wang, ”Representation Learning Using Multi-Task Deep Neural Networks for Semantic Classification and Information Retrieval,” *HLT-NAACL* 2015: 912-921

[11] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman, ”Return of the devil in the details: Delving deep into convolutional nets” In BMVC, 2014. 1, 3

[12] Y. Zheng, Q. Liu, E. Chen, Y. Ge, and J. L. Zhao, Exploiting multi-channels deep convolutional neural networks for multivariate time series classification, Frontiers of Computer Science, vol. 10, no. 1, pp. 96 112, 2016.

[13] Z. Wang, W. Yan, and T. Oates, ”Time Series Classification from Scratch with Deep Neural Networks”, A Strong Baseline. CoRR abs/1611.06455 (2016)

[14] C. Rich. Multitask Learning. Machine Learning, 28, 1997.

[15] B. Schuller, S. Steidl, A. Batliner, E. Bergelson, J. Krajewski, C. Janott, A. Amatuni, M. Casillas, A. Seidl, M. Soderstrom et al., The interspeech 2017 computational paralinguistics challenge: Addressee, cold & snoring.