Generative Adversarial Partial Domain Adaptation for Acoustic Scene Classification

Ningyu He[1] and Jie Zhu[2]

[1,2]Department of Electronic Engineering, Shanghai Jiao Tong University (SJTU), Shanghai 200240, China
Email: NingyuHe_ruby@163.com [1], zhujie@sjtu.edu.cn [2]

Published: 29 April 2020
Copyright © He et al.

Abstract

Previous domain adaptation methods generally assume that the source and the target data have the same label spaces. However, in the acoustic scene classification (ASC) task, target data collected from unseen conditions may have different label spaces. Therefore, conventional methods forcefully reducing the domain discrepancy in latent spaceignores the relationship between data with different labels. In this paper, a generative adversarial nets-based domain adaptation method is proposed for ASC task to perform a weighted partial transfer. Experiments are carried out on TUT acoustic scenes dataset, which improve the mean classification accuracy on untrained data from 21%to 52%. The network has also been practically applied to perimeter security system, which proves the reliability and universality of our method.

Keywords: acoustic scene classification, partial domain adaptation, generative adversarial training, perimeter security system.

Cite this article: He, N. & Zhu, J. (2020). Generative Adversarial Partial Domain Adaptation for Acoustic Scene Classification. *European International Journal of Science and Technology*, 9(4), 25-36.

I. Introduction

Acoustic scene classification (ASC) is the task of classifying the sounds from different environments. Compared to image recognition task, ASC has more confusing labels (e.g., a "public square" scene may also be labeled as "residential area"). Another problem is that different recording devices lead to differences in audio signal quality, which will result in the degradation of the performance of classification.

Current methods for ASC tasks are mainly based on deep learning [1] [2] [3]. It is generally assumed that the training data and the test data have the same distribution in feature and label spaces. However, this assumption is difficult to implement in practical applications. Therefore, one promising way to solve the problems mentioned above is domain adaptation (DA) [4] [5] [6]. Domain adaptation is a learning algorithm that employs training source data to solve the task in a new target data with few labels.

The purpose of domain adaptation is to minimize the distribution difference between the source domain and the target domain. With the development of deep learning technology, people have found that more domain invariant features can be extracted through deep networks than traditional mathematical methods. And pervious domain adaptation algorithms can be roughly divided into two stages. The first stage is the optimization of the source domain and then domain adaptation for target domain. However, these algorithms ignore the difference in label distribution between the source and target domain and the connection between different categories in the same domain. Forcefully reducing the domain discrepancy in the latent space will result in the destruction of intrinsic data structure.

To address the issues above, a partial domain adaptation method for acoustic scene classification is proposed based on the theory of generative adversarial learning. As mentioned above, our method aims to address the situation where the source and target domains have different label categories. We assume that source domain label category is sufficient to include all possible categories in target domain. In this situation, directly reducing the domain shift cannot achieve good results. Therefore, we proposed a generative adversarial module. Generative module aims to generate the augmented samples on each domain to preserve the class-level structure during domain adaptation process. On the contrast, adversarial module ensures that the closest class part of source domain can be transferred to the target domain. Experiments results on TUT dataset [2] show that our algorithm improves the average recognition accuracy rate by more than 15%. What is more, the algorithm has also been used in practical application of the perimeter security system [7] [8] and has achieved good results.

The rest of the paper is organized as follows. Section II provides a brief description on related work. In Section III, we introduce the proposed method in details. Then experiments are carried out on TUT acoustic scene dataset and the application of the algorithm in the perimeter security system is introduced in Section IV. Finally, we conclude the paper in Section V.

II. Related Work

A. Domain Adaptation

In this section, the basic ideas of domain adaptation are introduced. Source and target domain are noted as $D_S = \langle Z_S, f_S \rangle$ and $D_T = \langle Z_T, f_T \rangle$, where Z represents the data distribution and f represents the label processing. The key to the domain adaptation algorithm is to design a classifier h over z. The expected error of the h over its input z can be expressed as follow.

$$\grave{o}(h, f) = \mathrm{L}\left(h(z), f(z)\right) \tag{1}$$

Where L is the loss function, $\grave{o}(h, f)$ indicates the difference between the output of the classifier h and the label f. The goal of domain adaptation is to adjust h to get the small error $\grave{o}_S(h, f)$ [9] in the source domain and adapt it to the target domain D_T with low value of $\grave{o}_T(h, f)$.

The classifier h with the low $\grave{o}_S(h, f)$ can be obtained from classical training on the source domain D_S. However, we cannot optimize h by retaining on target domain D_T due to the lack of labels. Therefore, our focus has changed to reduce the discrepancy between Z_S and Z_T.

The discrepancy between Z_S and Z_T is usually expressed with $H\Delta H$ distance [10], which is defined in equation 2.

$$d_{H\Delta H} = \sup_{h,h' \in H} \left| P_{z \square Z_S}\left[h(z) \neq h'(z)\right] - P_{z \square Z_T}\left[h(z) \neq h'(z)\right] \right| \tag{2}$$

To sum up, $d_{H\Delta H}$ represents the prediction difference under two distributions Z_S and Z_T. Thus, the problem of domain adaptation reduces to minimizing the discrepancy $d_{H\Delta H}$ between two domains.

B. Generative Adversarial Domain Adaptation

In adversarial adaptation methods [11] [12], a domain discriminator is trained to minimize the domain discrepancy. In generative adversarial learning theory, adversarial losses with auxiliary loss can make sure that the learned function can transfer an individual source sample to the desired domain. However, previous methods only focus on the global transform. Since the discriminator can reduce the domain discrepancy, it destroys the class semantic feature in each category. Differently, we propose a partial domain adaptation method to solve the acoustic scene classification problem.

III. Proposed Method

This section introduces our proposed method in details. Some mathematical notation is set to explain our method. Source and target domain are defined as $D_S = \langle Z_S, f_S \rangle$ and $D_T = \langle Z_T, f_T \rangle$, where Z

represents the data distribution and f represents the label. In acoustic scene classification task, we assume that the feature spaces are the same ($M_S = M_T$) while the target label is a subset of the source label ($Y_T \in Y_S$).

A. Generative Adversarial Network

Fig.1 shows the overall framework of our network. The network can be viewed as two parts. The first part completes the task of generative adversarial training and the second part does the partial domain adaptation.

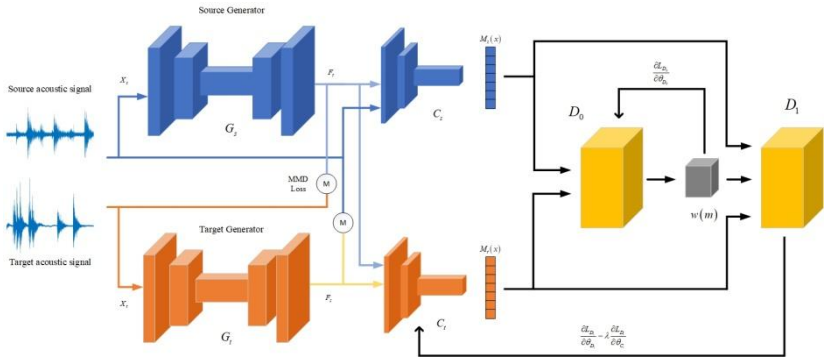

Fig.1. Overview of the proposed generative adversarial network

In first part, we draw on the idea of generative adversarial neural networks (GAN). Two generators G_s and G_t are designed to transfer one domain data to the other domain distribution. In other words, G_s aims to generate F_t through the source data X_s which is similar to the target data X_t. Similarly, G_t does the same job to generate the data F_s. The generator loss in source domain can is defined as the following formulation.

$$
\begin{aligned}
&\mathrm{L}_{GAN}^{s}\left(X_s\right) = \mathrm{L}_{dis_s} + \mathrm{L}_{cls_s} \\
&\mathrm{L}_{dis_s} = \mathrm{E}\left[\log C_s\left(X_s\right)\right] + \mathrm{E}\left[\log\left(1 - C_s\left(G_s\left(X_s\right)\right)\right)\right] \\
&\mathrm{L}_{cls_s} = \mathrm{E}\left[\log C_s\left(X_s, Y_s\right)\right] + \mathrm{E}\left[\log C_s\left(G_s\left(X_s\right), Y_s\right)\right]
\end{aligned}
\tag{3}
$$

Where L_{dis_s} is the discrimination loss and L_{cls_s} is the classification loss. C_s is a classifier aims to classify the real source data X_s and the fake data F_t. In the same way, C_t aims to classify F_s and F_t. So the generator loss in target domain is similarly defined as follows.

$$\begin{aligned} &\mathrm{L}_{GAN}^{t}\left(X_s, X_t\right)=\mathrm{L}_{dis_t}+\mathrm{L}_{cls_t} \\ &\mathrm{L}_{dis_s}=\mathrm{E}\left[\log C_t\left(X_s\right)\right]+\mathrm{E}\left[\log\left(1-C_t\left(G_t\left(X_t\right)\right)\right)\right] \\ &\mathrm{L}_{cls_s}=\mathrm{E}\left[\log C_t\left(F_t, Y_s\right)\right]+\mathrm{E}\left[\log C_t\left(F_s, \hat{Y}_t\right)\right] \end{aligned} \tag{4}$$

Where $\hat{Y}_t$ is the pseudo label of the target domain which comes from the $C_0\left(X_t\right)$. Here C_0 is a pre-trained classifier on the source domain.

The Maximum Mean Discrepancy (MMD) is another indicator that our network needs to be optimized. The MMD loss can compute the domain discrepancy to compare distinct distributions. In our network, we apply two kinds of MMD loss. One is the global MMD that shows the distance between two domains center. The other is class MMD that shows the distance in each class data. So the whole MMD loss is defined in equation 5.

$$\mathrm{L}_{MMD}^{s/t}=\mathrm{L}_{gMMD}^{s/t}+\frac{1}{N}\mathrm{L}_{cMMD}^{s/t} \tag{5}$$

Where N is the class number. We integrate the generator loss and MMD loss to obtain the overall goal of generative adversarial training.

$$\mathrm{L}=\mathrm{L}_{GAN}^{s}+\mathrm{L}_{GAN}^{t}+\lambda\left(\mathrm{L}_{MMD}^{s}+\mathrm{L}_{MMD}^{t}\right) \tag{6}$$

Where λ control the relative weight of two losses. To sum up, our network is to train two classifiers which take inputs from the generators. The classifiers aim to achieve the best classification performance while the generators aim to minimize the overall loss shown in equation 6.

B. Weighted Partial Domain Adaptation

In the above section, we trained generators and classifiers in the network by using the theory of generative adversarial networks. In this section, the trained classifiers are used to complete the subsequent partial domain adaptation [13].

We first extract the parameters of the last convolutional layer in the trained classifier as the input of our partial adaptation algorithm. The last convolution layer in the classifier can be considered as the output of a feature extractor $M\left(x\right)$. Therefore, the discriminator loss in our network is similar to the original GAN with its minimax loss:

$$\begin{aligned} \min_{M_s, M_t}\max_{D}\mathrm{L}\left(D, M_s, M_t\right)=&\mathrm{E}_{x\sim Z_s(x)}\left[\log D\left(M_s\left(x\right)\right)\right] \\ &+\mathrm{E}_{x\sim Z_t(x)}\left[\log\left(1-D\left(M_t\left(x\right)\right)\right)\right] \end{aligned} \tag{7}$$

Where M_s and M_t are the feature extractors and D is the domain discriminator to identify whether the data come from the source or target domain. The loss minimizes the data distribution divergence on the feature space M while produces a stricter bound for the discriminator D to achieve the purpose of adversarial learning.

As mentioned above, in our network, different feature extractors are trained for the source and target domain to get more domain specific features. After completing the classifier training on the source domain data, we turn our focus to training the domain classifier D. For the learned $M_s(x)$, the domain adversarial loss in equation 7 is simplified to equation 8 to reduce the shift between two domains.

$$\begin{aligned}\min_{M_t}\max_{D} \mathrm{L}\left(D, M_s, M_t\right) = & \mathrm{E}_{x\square Z_s(x)}\left[\log D\left(M_s\left(x\right)\right)\right] \\ & + \mathrm{E}_{x\square Z_t(x)}\left[\log\left(1-D\left(M_t\left(x\right)\right)\right)\right]\end{aligned} \tag{8}$$

Therefore, for any $M_s(x)$ and $M_t(x)$, training the domain discriminator D is to maximize the loss:

$$\begin{aligned}\max_{D} \mathrm{L}\left(D, M_s, M_t\right) = & \int_x Z_s\left(x\right)\log D\left(M_s\left(x\right)\right) \\ & + Z_t\left(x\right)\log\left(1-D\left(M_t\left(x\right)\right)\right)dx \\ = & \int_m Z_s\left(m\right)\log D\left(m\right) + Z_t\left(m\right)\log\left(1-D\left(m\right)\right)dm\end{aligned} \tag{9}$$

Where $m = M(x)$ is the feature sample after feature extraction. To solve this optimization problem, we can get the theoretical optimal solution of D through Leibniz's rule.

$$D^*\left(m\right) = \frac{Z_s\left(m\right)}{Z_s\left(m\right) + Z_t\left(m\right)} \tag{10}$$

As mentioned in the introduction, our approach is intended to solve the partial domain adaptation problem where the target label category is smaller than the source domain. So we need to determine whether the features extracted are from the class that is unique to the source domain or is shared by the two domains.

Fortunately, we find that the optimum D^* is a good indicator. It can be found in equation 10 that if the value of $D^*(m)$ is close to 1 which means that the feature is more likely come from the particular classes in the source domain. On the contrary, if $D^*(m)$ is small, the feature is more likely from the shared classes. Hence, we set a weight value for each feature based on the value of $D^*(m)$ and give the larger weight to the features from public classes to reduce the shift on the shared classes. Simply, the relationship between weight and D is defined as follows.

$$\tilde{w}\left(m\right) = 1 - D^*\left(m\right) = \frac{Z_t\left(m\right)}{Z_s\left(m\right) + Z_t\left(m\right)} \tag{11}$$

It is clear from equation 11 that weight can not only complete the weight assignment of features but also can reflect the distribution ratio between two domains. It is reasonable to give larger weights to data with coincident distributions. The weights are normalized for training as follows

$$w(m)=\frac{\tilde{w}(m)}{\mathrm{E}_{m\square Z_s(m)}\tilde{w}(m)} \tag{12}$$

In our proposed method, two domain discriminators D are designed to complete our adversarial partial domain adaptation network. The first discriminator D_0 is designed to achieve the weights on M_s and M_t. Another discriminator D_1 with the weighted data from two domains is trained to reduce the shift on the shared classes.

After adopting weights to the source data, the training goal in equation 8 turns to:

$$\begin{aligned}\min_{M_t}\max_{D}\mathrm{L}\left(D,M_s,M_t\right)&=\mathrm{E}_{x\square Z_s(x)}\left[w(m)\log D\left(M_s\left(x\right)\right)\right]\\&+\mathrm{E}_{x\square Z_t(x)}\left[\log\left(1-D\left(M_t\left(x\right)\right)\right)\right]\end{aligned} \tag{13}$$

Where $w(m)$ is independent of D_1 which can be seen as a constant. Similar to that shown in equation 10, the optimum D_1 is obtained at:

$$D_1^*\left(m\right)=\frac{w(m)Z_s\left(m\right)}{w(m)Z_s\left(m\right)+Z_t\left(m\right)} \tag{14}$$

Therefore, for the given optimum D_1, the minimax goal in equation 13 can be transferred as:

$$\begin{aligned}&\mathrm{L}\left(M_t\right)=\mathrm{E}_{x\square Z_s(x)}\left[w(m)\log D\left(M_s\left(x\right)\right)\right]\\&\quad+\mathrm{E}_{x\square Z_t(x)}\left[\log\left(1-D\left(M_t\left(x\right)\right)\right)\right]\\&=\int_m w(m)Z_s\left(m\right)\log\frac{w(m)Z_s\left(m\right)}{w(m)Z_s\left(m\right)+Z_t\left(m\right)}\\&+Z_t\left(m\right)\log\frac{w(m)Z_s\left(m\right)}{w(m)Z_s\left(m\right)+Z_t\left(m\right)}dm\end{aligned} \tag{15}$$

To sum up, the fundamental goal of our partial adaptation method is to reduce the divergence between the weighted distribution density of the source domain and the target domain, so as to achieve the purpose of partial domain adaptation. Compared to other common domain common domain adaptation algorithms, like Adversarial Discriminative Domain Adaptation (ADDA) and Selective Adversarial Networks (SAN), the contributions of our method are concluded as follows:

· We establish a connection between two generators in the source and target domains. Therefore, generators can not only expand the data samples but also preserve the class-level structure during domain adaptation.

·A weighting scheme based on adversarial nets is proposed to do the selective adaptation from source domain. This solves the problem that the source and target domain have different label distributions.

IV. Experiments And Applications

Several experiments are carried out to evaluate the effectiveness of our methods. Comparisons are performed on TUT Acoustic Scenes dataset between our method and pervious SOTA method to prove the superiority of our algorithm. What is more, our method is also applied to the actual perimeter security system to prove the practicability of our algorithm.

A. TUT Acoustic Scenes Classification

TUT dataset is widely used in Acoustic Scenes Classification task. The dataset includes audio recordings collected under ten different acoustic scenes, such as “airport”, “metro station”, “shopping mall” and so on. Each category of audio data is recorded by three different acquisition devices and marked with A, B and C to distinguish. Device A is a professional recording device that can capture high quantity audio data, while device B and C are common recording devices. Thus we regard the data from device A as the source domain data and data from B, C as the target domain data.

In our method, we take the energy spectrum of the audio as the input to the network and use the deep network as the feature extractor M to extract abstract features instead of traditional speech features. Feature extractor M is designed as a residual convolution neural network with long short term memory (LSTM). The network establishes five layers of convolution layers and three layers of LSTM to extract the time-frequency features of the input. Moreover, residual concepts are introduced in M to address the possible degradation problem. M_s and M_t are designed with the same structure but trained separately for respective domain data. Domain discriminators D are simply designed as four layers of fully connected layers.

The purpose of our method is to complete partial domain adaptation in small target domain category. Therefore, all category data in the source domain is trained. In contrast, only part of the category data is selected for training in the target domain.

We select different types of target data for multiple experiments. Comparisons are carried out between our method and SOTA domain adaptation methods. In this paper, Adversarial Discriminative Domain Adaptation (ADDA) [14] and Selective Adversarial Networks (SAN) [15] are selected for comparative experiments.

Fig.2 records the average classification accuracy without domain adaptation. In contrast, mean accuracy of method is shown in Fig.3. It is clear that target data without domain adaptation is difficult to obtain good recognition results on the classifier trained in the source domain. The average recognition accuracy is only 20.4%. This is because the differences in the acquisition equipment will greatly affect the characteristics of the audio signals. In addition, the shift on label domain also increases the difficulty of classification task. In contrast, results in Fig.3 demonstrate the effectiveness of our algorithm. It is obvious that the recognition accuracy on the diagonal of Fig.3 is significantly improved which proves that our method decreases the discrepancy between Z_s and Z_t . The recognition accuracy of any category has been improved to more than 30% and the average classification accuracy rate has increased to 52%.

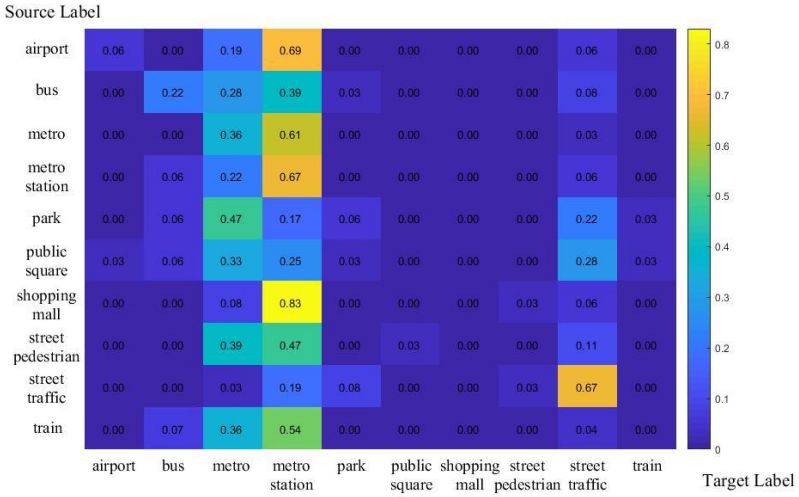

Fig.2. TUT classification results before domain adaptation

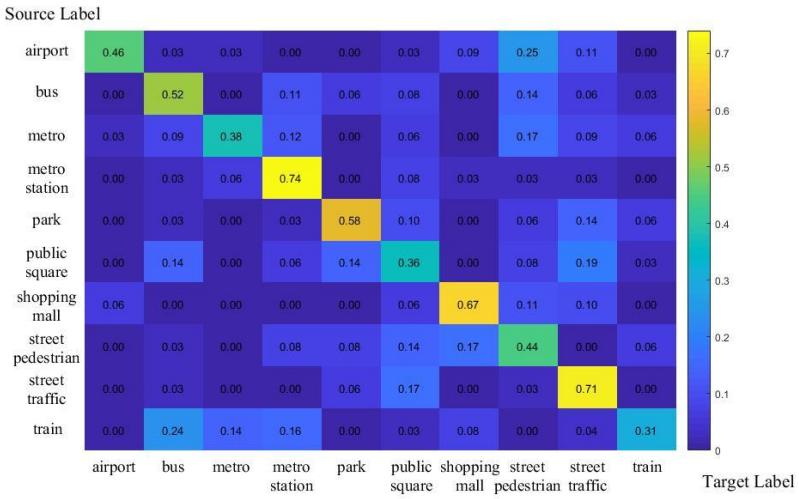

Fig.3. TUT classification results after domain adaptation

Another experiment is used to prove the superiority of our method. Adversarial Discriminative Domain Adaptation (ADDA) and Selective Adversarial Networks (SAN) are selected for comparison experiments. The results of the comparative experiments are recorded in Table I. Observing the results in Table I, we find that our method achieves better recognition accuracy in target domain, indicating that the weight setting of domain-invariant features in our algorithm is reasonable and effective.

Table I. Classification accuracy on different domain adaptation methods

Algorithm	D_s	D_t
ADDA	64.6%	42.4%
SAN	65.7%	47.6%
Our method	65.2%	51.6%

B. Application in Perimeter Security System

With the development of society, the intelligent perimeter security system has been applied in various occasions. The traditional perimeter security system is generally built on the camera monitoring. Processing the video signals requires huge computing power. Therefore, we proposed a new type of perimeter security system based on the optical fiber sensors. The system framework is shown in Fig.4.

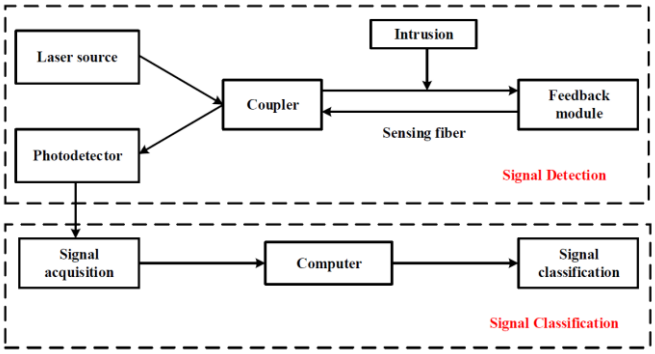

Fig.4. The framework of our proposed perimeter security system

As illustrated in Fig.4, the optical fiber sensor collects signals from the vibration on the optical path, so the collected signal can be regarded as an audio signal. Thus the perimeter security system can be seen as an alternative audio scene classification task. What is more, although the security system has trained various types of intrusion signals, each single intrusion needs to be identified in the security system, which also coincides with the thought of partial adaptation.

Experiments are taken under variant environments to do the comparison between our method and traditional recognition algorithms and the results are shown in Table II.

Table II. Alarm accuracy on different climates

Algorithm	Sunny days	Rainy days
BPNN	86.9%	82.7%
SVM	87.6%	84.5%
DNN	76.3%	74.8%
Our method	90.2%	86.7%

It is clear from Table II that our method achieves the highest accuracy in both sunny and rainy days. This proves that the idea of weighting similar categories of data in our method is effective and universal.

V. Conclusion

In this paper, a generative adversarial partial domain adaptation method is proposed. We draw on the ideas of the Generative Adversarial Networks (GANs) to solve the problem that the test data and the training data have different label spaces in deep learning. We establish a connection between two generators in the source and target domains. Therefore, generators can not only expand the data samples but also preserve the class-level structure during domain adaptation. Then a network with two domain

discriminators is designed to reduce the shift on source and target domain. The first domain discriminator D_0 is used to find the source domain data that are similar to the target and assign the larger weight to these data. This effectively reduces the distribution difference between the source and target domain. Another discriminator D_1 is trained with the weighted data to reduce the shift on the shared classes.

We focus on the Acoustic Scenes Classification (ASC) tasks in this paper. Experiments are taken on TUT dataset among our method and the SOTA domain adaptation algorithms (ADDA, SAN). Results show that our method plays a good domain adaptation effect. The mean classification accuracy increased from 21% to 53% after domain adaptation. What is more, our method also achieves better classification accuracy than ADDA and SAN methods.

In order to further prove the versatility of our method, we apply it to a fiber optical security system. The experimental results also show that our algorithm has achieved better recognition results under various environments than traditional classification recognition algorithms(SVM, BPNNand DNN). Various experiments have proved that our algorithm is effective and flexible. In our further work, we will try to apply the network to more complex datasets with different label spaces.

References

[1] Han Y, Park J, Lee K. Convolutional neural networks with binauralrepresentations and background subtraction for acoustic scene classification[J]. the Detection and Classification of Acoustic Scenes and Events (DCASE), 2017: 1-5.

[2] Gharib S, Drossos K, Cakir E, et al. Unsupervised adversarial domain adaptation for acoustic scene classification[J]. arXiv preprint arXiv:1808.05777, 2018.

[3] Weiping Z, Jiantao Y, Xiaotao X, et al. Acoustic scene classification using deep convolutional neural network and multiple spectrograms fusion[J]. Detection and Classification of Acoustic Scenes and Events (DCASE), 2017.

[4] Ganin Y, Ustinova E, Ajakan H, et al. Domain-adversarial training of neural networks[J]. The Journal of Machine Learning Research, 2016, 17(1): 2096-2030.

[5] Hou C A, Yeh Y R, Wang Y C F. An unsupervised domain adaptation approach for cross-domain visual classification[C]//2015 12th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS). IEEE, 2015: 1-6.

[6] Ganin Y, Ustinova E, Ajakan H, et al. Domain-adversarial training of neural networks[J]. The Journal of Machine Learning Research, 2016, 17(1): 2096-2030.

[7] He N, Zhu J, Li L. An optic-fiber fence intrusion recognition system using the optimized curve fitting model based on the svm method[C]//2018 International Joint Conference on Neural Networks (IJCNN). IEEE, 2018: 1-6.

[8] He N, Zhu J. Deep Learning Approach For Audio Signal Classification And Its Application In Fiber Optic Sensor Security System[C]//2019 9th International Conference on Information Science and Technology (ICIST). IEEE, 2019: 263-267.

[9] Ben-David S, Blitzer J, Crammer K, et al. A theory of learning from different domains[J]. Machine learning, 2010, 79(1-2): 151-175.

[10] Drossos K, Magron P, Virtanen T. Unsupervised adversarial domain adaptation based on the Wasserstein distance for acoustic scene classification[C]//2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA). IEEE, 2019: 259-263.

[11] Goodfellow I. NIPS 2016 tutorial: Generative adversarial networks[J]. arXiv preprint ar Xiv:1701.00160, 2016.

[12] Ganin Y, Ustinova E, Ajakan H, et al. Domain-adversarial training of neural networks[J]. The Journal of Machine Learning Research, 2016, 17(1): 2096-2030.

[13] Zhang J, Ding Z, Li W, et al. Importance weighted adversarial nets for partial domain adaptation[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 8156-8164.

[14] Tzeng E, Hoffman J, Saenko K, et al. Adversarial discriminative domain adaptation[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 7167-7176.

[15] Cao Z, Long M, Wang J, et al. Partial transfer learning with selective adversarial networks[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 2724-2732.