A DOMAIN ADAPTIVE ACOUSTIC SCENE CLASSIFICATION MODEL IN SEMI-SUPERVISED TRANSFER LEARNING.

Yizhou Tan, Haojun Ai

Wuhan University

ABSTRACT

Acoustic scene classification (ASC) is a crucial task in audio signal processing with applications ranging from surveillance to augmented reality. In this technical report, we propose an approach to ASC that combines domain adaptation with semi-supervising methods for improved system performance. We first pre-train our model on the TAU Urban Acoustic Scenes 2020 Mobile development dataset to learn robust representations of acoustic scenes. Then, we fine-tune the pre-trained model by combining the Maximum Classification Discrepancy (MCD), adversarial domain classifier and Fixmatch methods on a combination of the above TAU dataset and CAS 2023 dataset to enhance the robustness of the model.

***Index Terms*—** Domain adaptation, Semi-supervised Learning

1. INTRODUCTION

Acoustic scene classification (ASC) is crucial in various applications, including environmental monitoring, urban planning, and surveillance. The task involves classifying audio recordings into predefined categories based on the sound characteristics of the scene. ASC faces challenges such as the variability of acoustic environments, the presence of background noise, and the need for models that can generalize well across different environments.

Traditional ASC approaches have relied on handcrafted features and shallow machine learning models [1, 2, 3, 4]. Recent advancements in deep learning have led to significant improvements in ASC performance. However, existing deep learning models often struggle with model generalization, making them less effective when applied to different acoustic environments, such as in different cities.

The model generalization problem raises a potential catastrophic forgetting [5, 6] in semi-supervised transfer learning, which encourages the pre-trained model to overfit the limited labelled data and forget the knowledge of the pre-training dataset. The emergence of catastrophic forgetting will reduce the model generalization to unlabeled data thus affecting the effectiveness of semi-supervised learning. This motivates us to utilize the pre-training dataset in the fine-tuning process to overcome the catastrophic forgetting, by aligning all datasets into a consistent feature space.

In this work, we propose an ASC model combining the domain adaptation and semi-supervised methods to transfer a pre-trained model to a new dataset. The Maximum Classification Discrepancy (MCD) [7] and adversarial domain classifiers techniques [8] for domain adaptation. aiming to eliminate the domain biases among pre-training, training, and unlabelled datasets. We also introduce the Fixmatch [**?**] method for semi-supervising learning in the fine-tuning process, which can enhance the model generalization by the inference alignment in the unlabelled dataset.

The experiments are conducted on the TAU Urban Acoustic Scenes 2020 Mobile development dataset [9] (pre-trained dataset) and CAS 2023 dataset [10] (training and unlabelled dataset). It is worth noting that the experiment results are based on a random splitted validation set as there is not a data-sufficient validation set.

The key contribution of this research is combining the domain adaptation and semi-supervised methods to overcome the catastrophic forgetting problem in transfer learning.

2. METHOD

With the extraction of Mel spectrograms acoustic features (640x256) from audio recordings, the data of pre-training, training, and unlabelled datasets are represented as $(x_p, y_p), (x_k, y_k), x_u$. For fine-tuning a model $G(\cdot|theta_g)$ and a series classifier, the total Loss consists of pre-training loss, MCD loss, domain classification loss, Cross-entropy loss and Consistency Loss.

Pre-training Loss: It is used to keep the knowledge of pre-training data set, as follows:

$$L_p = CE(F(G(x_p|\theta)|theta_f), y_p) \tag{1}$$

where CE is the cross-entropy loss.

MCD Loss: It is used to make sure that the inference of the unknown dataset is consistent, as follows:

$$L_{mcd} = ||F_1(G(x_u)|\theta_{f1})) - F_2(G((x_u)|\theta_{f2})||_1 \tag{2}$$

where $F_1(\cdot|\theta_{f1}))$ and $F_2(\cdot|\theta_{f2}))$ are two different classifier.

Domain classification Loss: The three datasets x_p, x_k, x_u are considered as three different data domains with the domain label d_p, d_k, d_u. The domain classification loss will introduce a domain classifier with a gradient reversed layer to classify the data domain while encouraging the feature extractor to confuse the domain classifier, as follows:

$$L_d = CE(F_d(R(G(x))|\theta_d), d) \tag{3}$$

where the $R(\cdot)$ is the gradient reversed layer, $F_d(\cdot|\theta_d)$ is the domain classifier, x is all of three dataset and d is device label.

Cross-entropy loss: It is the basic classification loss for the training set, as follows:

$$L_{ce} = CE(F_1(G(x_k)|\theta_{f1}), y_k) + CE(F_2(G(x_k)|\theta_{f2}), y_k) \tag{4}$$

where CE is the cross-entropy loss.

Consistency Loss: Following with the Fixmatch [**?**], two different data augmentation (A_w, A_s) are used:

$$p_w = F_1(G(A_w(x_u)) + F_2(G(A_w(x_u)) \tag{5}$$
$$p_s = F_1(G(A_s(x_u)) + F_2(G(A_s(x_u)) \tag{6}$$
$$L_{con} = CE(p_s, pseudo(p_s)) \tag{7}$$

where $pseudo$ is a pseudo-label generator controlled by a threshold.

The model will be updated with an adversarial process, as follows:

$$\text{Step} \quad 1: \quad \min_{\theta_{f_1}, \theta_{f_2}} L = L_{ce} - \lambda L_{mcd} \tag{8}$$
$$\text{Step} \quad 2: \quad \min_{\theta_g, \theta_f, \theta_d} L = L_{ce} + L_{con} + \lambda L_{mcd} + \beta L_p \tag{9}$$

where $\lambda = 0.1$ and $\lambda = 0.1$ in our work.

3. EXPERIMENTS

Unfortunately, as there is not an official validation set, we have to adopt two highly limited and even flawed methods to evaluate our model.

Firstly, our model performs about 94% in the random validation split (20% data) of the training set.

Secondly, when we use all of the training set in the training process, we use the raw data (without data augmentation) of the training set as a test set, which is flawed but a compromise to pick the best model. The performance is also about 94% accuracy.

4. REFERENCES

[1] Antti Eronen, Juha Tuomi, Anssi Klapuri, Seppo Fagerlund, Timo Sorsa, Gaëtan Lorho, and Jyri Huopaniemi, "Audio-based context awareness-acoustic modeling and perceptual evaluation," in *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03).* IEEE, 2003, vol. 5, pp. 529–532.

[2] Alain Rakotomamonjy and Gilles Gasso, "Histogram of gradients of time–frequency representations for audio scene classification," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 1, pp. 142–153, 2014.

[3] Wenjun Yang and Sridhar Krishnan, "Combining temporal features by local binary pattern for acoustic scene classification," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 6, pp. 1315–1321, 2017.

[4] Shamsiah Abidin, Roberto Togneri, and Ferdous Sohel, "Spectrotemporal analysis using local binary pattern variants for acoustic scene classification," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 11, pp. 2112–2121, 2018.

[5] Robert M French, "Catastrophic forgetting in connectionist networks," *Trends in cognitive sciences*, vol. 3, no. 4, pp. 128–135, 1999.

[6] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al., "Overcoming catastrophic forgetting in neural networks," *Proceedings of the national academy of sciences*, vol. 114, no. 13, pp. 3521–3526, 2017.

[7] Kuniaki Saito, Kohei Watanabe, Yoshitaka Ushiku, and Tatsuya Harada, "Maximum classifier discrepancy for unsupervised domain adaptation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*. IEEE, 2018, pp. 3723–3732.

[8] Yizhou Tan, Haojun Ai, Shengchen Li, and Mark D Plumbley, "Acoustic scene classification across cities and devices via feature disentanglement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2024.

[9] Toni Heittola, Annamaria Mesaros, and Tuomas Virtanen, "Acoustic scene classification in dcase 2020 challenge: generalization across devices and low complexity solutions," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2020 Workshop (DCASE2020)*, 2020, pp. 56–60.

[10] Jisheng Bai, Mou Wang, Haohe Liu, Han Yin, Yafei Jia, Siwei Huang, Yutong Du, Dongzhe Zhang, Mark D Plumbley, Dongyuan Shi, et al., "Description on ieee icme 2024 grand challenge: Semi-supervised acoustic scene classification under domain shift," *arXiv preprint arXiv:2402.02694*, 2024.