UNSUPERVISED ANOMALOUS SOUND DETECTION UNDER DOMAIN SHIFT CONDITIONS BASED ON MOBILEFACENETS AND MASKED AUTOREGRESSIVE FLOW

Technical Report

Gang Liu, Yi Liu, Shifang Cai and Minghang Chen
Beijing University of Posts and Telecommunications
School of Artificial Intelligence. No.10, Xitucheng Road
Haidian District, BeiJing 100876, China
liugang@bupt.edu.cn

ABSTRACT

We present our submission to the DCASE2022 Challenge Task 2, which aims to promote research in unsupervised anomalous sound detection under domain shift condition. We propose two architectures to solve this problem, one is a self-supervised model adopting MobileFaceNets, and the other is one density estimation probability distribution model based on Masked Autoregressive Flow.

Index Terms— DCASE2022, Anomalous Sound Detection, Domain Shift, Machine Condition Monitoring, MobileFaceNets, Masked Autoregressive Flow

1. INTRODUCTION

The DCASE2022 Challenge Task 2 is concerned with identifying anomalous behavior from a target machine using sound recordings [1]. The main difference between this task and other DCASE tasks is that it is unsupervised. Consequently, there are only samples from the normal-state distributions in the available training data. Another complicating factor for this challenge is that the acoustic characteristics of the training data and of the test data are different – a condition known as domain shift and there are some known results for reducing the performance gap between the training and test data [2],[3], [4], [5], [6], [7], [8], [9], [10]. Recognizing the potential of these techniques, our experiments combine their advantages and hope to achieve better performance.

In our submission, we use two self-supervised classifiers based on previous work [11], [12], [13], [14]. In the first classifier we adopt dual features of audio waveform and logMel spectrum.Meanwhile, different anomaly score calculation methods are adopted for different types of machines. In the second model, a classifier is utilized that relies on masked normalizing flows to estimate the conditional density of input logMel spectrum and outputs are used to produce an anomaly score [15], [16], [17], [18], [19].

In the following sections, we will demonstrate each submission, how it was trained, its hyperparameters, and their respective results. To ensure our experiment more convincing, the baseline scores are showed in Tables 1 and 2. The data used in this challenge is 16 KHz, single-channel audio. For more details, please see [1], [20], [21].

2. ARCHITECTURES

The first model builds on the work [11]. The input to the model is the raw waveform and logMel spectrum of the audio. The second model is based on the well-known Masked Autoregressive Flow architecture, using Masked Autoencoder for Distribution Estimation(MADE).MADE is an auto-encoder that can capture distribution density. The overall model input is divided into two parts, one is the section ID as a conditional input, and the other is the logMel spectrum of normal audio.

Table 1: Baseline Autoencoder Scores

	ToyCar	ToyTrain	bearing	fan	gearbox	slider	valve
h_mean_auc_source	0.9041	0.7632	0.5442	0.7859	0.6893	0.7795	0.5201
h_mean_auc_target	0.3481	0.2335	0.5838	0.4718	0.6264	0.4767	0.4946
h_mean_pauc	0.5274	0.5048	0.5198	0.5752	0.5849	0.5578	0.5036

Table 2: Baseline MobileNetV2 Scores

	ToyCar	ToyTrain	bearing	fan	gearbox	slider	valve
h_mean_auc_source	0.5912	0.5726	0.6058	0.7075	0.6921	0.6515	0.6709
h_mean_auc_target	0.5196	0.4590	0.5994	0.4822	0.5619	0.3823	0.5722
h_mean_pauc	0.5227	0.5152	0.5714	0.5690	0.5603	0.5467	0.6242

2.1. MobileFaceNets

The architecture of the first model is shown in Figure 1. Furthermore, we adopt different methods to calculate abnormal scores In order to capture the most obvious features of anomalous audio.

2.1.1. Preprocessing

Since there are only a small number of labeled samples of target domain in the provided dataset, we employ a data augmentation algorithm[14]. At the same time, this algorithm can construct the neighborhood value of training samples according to the dataset distribution through the prior knowledge of the training set, and enhance the generalization ability of the model.

2.1.2. Training & Results

The methods of calculating anomaly scores are shown in Table 3. ToyTrain adopts the algorithm of Local Outlier Factor. Several machine types of Valve, Slider, Bearing and Gearbox are determined by experiments. The loss measurement method of cosine similarity

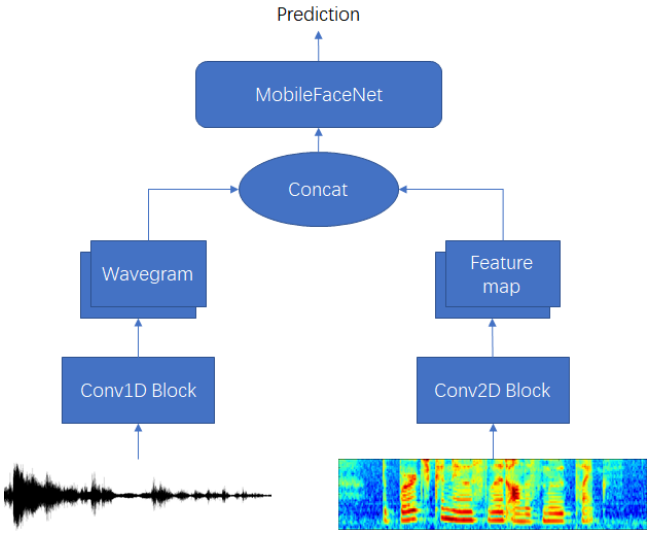

Figure 1: The Overview Architecture Of MobileFaceNets.

Table 3: Methods of Calculating Anomaly Scores

	ToyTrain	bearing	fan	gearbox	slider	valve
anomaly_scores	LOF	COS	GMM	COS	COS	COS

is adopted. At the same time, the model introduces two labels of source domain and target domain for domain judgment. When the audio point is closer to the source domain, the calculation of cosine similarity is performed with the center point of the source domain data. Training. When Fan's abnormal score is calculated, the GMM algorithm is used. The training is usually carried out for 30 epochs, using the training sets in the development and evaluation datasets, and finally calculating the audio embedding and in order to measure the audio from different angles. Loss, we use ArcFace loss during training. Embedding is used to calculate cosine similarity during testing. Table 3 shows our experimental results, which can be seen to be much improved than the baseline indicators.

Table 4: MobileFaceNets Scoring Results

	ToyCar	ToyTrain	bearing	fan	gearbox	slider	valve
h_mean_auc_source	0.6659	0.5650	0.6628	0.6386	0.7951	0.9331	0.8456
h_mean_auc_target	0.5610	0.4714	0.7034	0.5880	0.7394	0.8324	0.8337
h_mean_pauc	0.5258	0.5137	0.5665	0.5672	0.5961	0.7378	0.7011

2.2. MASKED AUTOREGRESSIVE FLOW

The second model architecture we submitted is shown in the figure, the model is a classifier that classifies sections with conditional input.

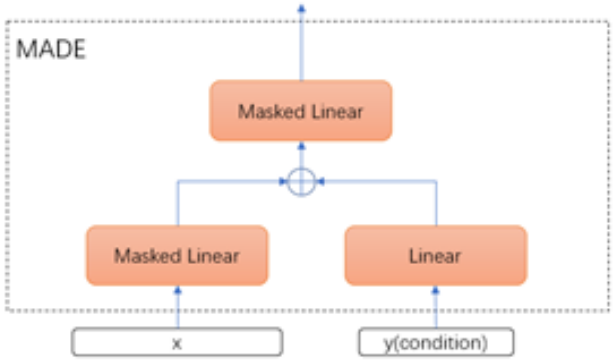

Figure 2: The Architecture Of Masked Autogressive Flow.

2.2.1. Preprocessing

Since there are only a small number of labeled target domain samples in the dataset, we employ a data augmentation algorithm of mixup [ref]. At the same time, this algorithm can construct the neighborhood value of training samples on the distribution of the training set through the prior knowledge of the training set, and enhance the generalization ability of the model.

The model did not use any special preprocessing or augmentation. Logarithmic analyses were performed on both STFT and Mel spectra. All spectrograms are calculated with frequency min and max set to 100 Hz and 8000 Hz, respectively.

Table 5: MAF Scoring Results

	ToyCar	ToyTrain	bearing	fan	gearbox	slider	valve
h_mean_auc_source	0.8936	0.7512	0.6628	0.7487	0.7951	0.9331	0.8456
h_mean_auc_target	0.6761	0.4968	0.7034	0.6780	0.7394	0.8324	0.8337
h_mean_pauc	0.5666	0.5465	0.5665	0.5811	0.5961	0.7378	0.7011

2.2.2. Training & Results

The abnormal score calculation method of each type in the training process is shown in Table 3. ToyTrain adopts the algorithm of Local Outlier Factor. Several machine types of Valve, Slider, Bearing and Gearbox are determined by experiments. The loss measurement method of cosine similarity is adopted. At the same time, the model introduces two labels of source domain and target domain for domain judgment. When the audio point is closer to the source domain, the calculation of cosine similarity is performed with the center point of the source domain data. Training. When Fan's abnormal score is calculated, the GMM algorithm is used. The training is usually carried out for 30 epochs, using the training sets in the dev and eva data sets, and finally calculating the audio embedding, and in order to measure the audio from different angles. Loss, we use ArcFace loss during training. Embedding is used to calculate cosine similarity during testing. Table 3 shows our experimental results, which can be seen to be much improved than the baseline indicators.

3. CONCLUSIONS

We have elaborated on our submission to the DCASE2022 challenge Task 2, the notable difficulty of which is the domain shift between the training and testing datasets. We find that domain adaptative methods that perform well in other modalities (especially visual) do not seem to work well with audio (at least in our experiments). This difference brings a greater need for adaptation methods in the audio domain to the DCASE2022 Challenge. Develop more audio domain adaptative techniques to solve domain shift has become a problem to be solved.

In the model we developed, we found that using Masked Autogressive Flow has some benefit for domain generalization as it is unsupervised and captures the distribution inside the audio. In a real-world setting, it is very difficult for us to collect more data on the target domain. It is even more impossible to obtain anomalous audio, so it is crucial to be able to use Masked Autogressive Flow to predict the distribution inside normal data. Going forward, we plan to work with more variants of Masked Autogressive Flow.

4. REFERENCES

[1] K. Dohi, K. Imoto, N. Harada, D. Niizumi, Y. Koizumi, T. Nishida, H. Purohit, T. Endo, M. Yamamoto, and Y. Kawaguchi, "Description and discussion on dcase 2022 challenge task 2: Unsupervised anomalous sound detection for machine condition monitoring applying domain generalization techniques," *arXiv preprint arXiv:2206.05876*, 2022.

[2] G. Wilson and D. J. Cook, "A survey of unsupervised deep domain adaptation," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 11, no. 5, pp. 1–46, 2020.

[3] Y. Li, N. Wang, J. Shi, X. Hou, and J. Liu, "Adaptive batch normalization for practical domain adaptation," *Pattern Recognition*, vol. 80, pp. 109–117, 2018.

[4] F. Maria Carlucci, L. Porzi, B. Caputo, E. Ricci, and S. Rota Bulo, "Autodial: Automatic domain alignment layers," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 5067–5075.

[5] F. M. Carlucci, L. Porzi, B. Caputo, E. Ricci, and S. R. Bulo, "Just dial: Domain alignment layers for unsupervised domain adaptation," in *International conference on image analysis and processing*. Springer, 2017, pp. 357–369.

[6] M. Mancini, L. Porzi, S. R. Bulo, B. Caputo, and E. Ricci, "Boosting domain adaptation by discovering latent domains," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 3771–3780.

[7] J. Shen, Y. Qu, W. Zhang, and Y. Yu, "Wasserstein distance guided representation learning for domain adaptation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, 2018.

[8] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," in *International conference on machine learning*. PMLR, 2015, pp. 1180–1189.

[9] Y. Asai, "Sub-cluster adacos based unsupervised anomalous sound detection for machine condition monitoring under domain shift conditions."

[10] J. A. Lopez, G. Stemmer, P. Lopez-Meyer, P. Singh, J. A. del Hoyo Ontiveros, and H. A. Cordourier, "Ensemble of complementary anomaly detectors under domain shifted conditions." in *DCASE*, 2021, pp. 11–15.

[11] Q. Kong, Y. Cao, T. Iqbal, Y. Wang, W. Wang, and M. D. Plumbley, "Panns: Large-scale pretrained audio neural networks for audio pattern recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2880–2894, 2020.

[12] M. Germain, K. Gregor, I. Murray, and H. Larochelle, "Made: Masked autoencoder for distribution estimation," in *International conference on machine learning*. PMLR, 2015, pp. 881–889.

[13] G. Papamakarios, T. Pavlakou, and I. Murray, "Masked autoregressive flow for density estimation," *Advances in neural information processing systems*, vol. 30, 2017.

[14] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," *arXiv preprint arXiv:1710.09412*, 2017.

[15] I. Kobyzev, S. J. Prince, and M. A. Brubaker, "Normalizing flows: An introduction and review of current methods," *IEEE transactions on pattern analysis and machine intelligence*, vol. 43, no. 11, pp. 3964–3979, 2020.

[16] D. Rezende and S. Mohamed, "Variational inference with normalizing flows," in *International conference on machine learning*. PMLR, 2015, pp. 1530–1538.

[17] G. Papamakarios, E. T. Nalisnick, D. J. Rezende, S. Mohamed, and B. Lakshminarayanan, "Normalizing flows for probabilistic modeling and inference." *J. Mach. Learn. Res.*, vol. 22, no. 57, pp. 1–64, 2021.

[18] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 4690–4699.

[19] L. Dinh, J. Sohl-Dickstein, and S. Bengio, "Density estimation using real nvp," *arXiv preprint arXiv:1605.08803*, 2016.

[20] K. Dohi, T. Nishida, H. Purohit, R. Tanabe, T. Endo, M. Yamamoto, Y. Nikaido, and Y. Kawaguchi, "Mimii dg: Sound dataset for malfunctioning industrial machine investigation and inspection for domain generalization task," *arXiv preprint arXiv:2205.13879*, 2022.

[21] N. Harada, D. Niizumi, D. Takeuchi, Y. Ohishi, M. Yasuda, and S. Saito, "Toyadmos2: Another dataset of miniature-machine operating sounds for anomalous sound detection under domain shift conditions," *arXiv preprint arXiv:2106.02369*, 2021.