ENSEMBLE OF ADACOS BASED SYSTEMS FOR UNSUPERVISED ANOMALOUS SOUND DETECTION FOR MACHINE CONDITION MONITORING APPLYING DOMAIN GENERALIZATION TECHNIQUES

Technical Report

Chengliang Jiang , Yan Wang

PFU SHANGHAI Co., LTD

46 Building 4~5 Floors, 555 GuiPing Road

XuHui District, Shanghai 200233, CHINA

jiang_chengliang.pfu@fujitsu.com, yan.wang.alex@gmail.com

ABSTRACT

In this report, we outline our ensemble of models solution for the DCASE 2022 challenge's Task 2 (Unsupervised Anomalous Sound Detection for Machine Condition Monitoring Applying Domain Generalization Techniques). The basic approach follows our DCASE2021 Task 2 system [1]. In 2022 we diversify our CNN backbone architecture and input size. The final submissions are the ensemble of 2 models for each machine type. The model is trained on a mixture of source and target domains, achieving the same performance on both source and target domains.

Index Terms— DCASE2022, anomalous sounds detection, metric learning, Adacos

1. INTRODUCTION

DCASE2022 Task2 has three main challenges:

1. Only normal sound clips are provided as training data (i.e., unsupervised learning scenario). In real-world factories, anomalies rarely occur and are highly diverse. Therefore, exhaustive patterns of anomalous sounds are impossible to create or collect and unknown anomalous sounds that were not observed in the given training data must be detected. This condition is the same as in DCASE 2020 Task 2 and DCASE 2021 Task 2.

2. Factors other than anomalies change the acoustic characteristics between training and test data (i.e., domain shift). In real-world cases, operational conditions of machines or environmental noise often differ between the training and testing phases. For example, the operation speed of a conveyor can change due to seasonal demand, or environmental noise can fluctuate depending on the states of surrounding machines. This condition is the same as in DCASE 2021 Task 2.

3. In test data, samples unaffected by domain shifts (source domain data) and those affected by domain shifts (target domain data) are mixed, and the source/target domain of each sample is not specified. Therefore, the model must detect anomalies with the same threshold value regardless of the domain (i.e., domain generalization).

For challenge 1, we train three section classification network for each machine type, which tries to identify each section under a certain machine type. Two model uses EfficientNet_B0 [2] as the backbone architecture, and another model uses Swim transformer [3] as the backbone architecture. In test phase, the last classification layer of the network is removed. Each input spectrogram is mapped into a 1280-dim or 1024-dim embedding vector, which is used for measuring cosine similarity in angular space. For challenge 2 and challenge 3, the models trained on source domain data are further fine-tuned on the target domain data. We use serval finetuning strategies to improve performance.

1.1. DCASE 2021 Task2 Dataset

The data used for this task consists of the normal/anomalous operating sounds of seven types of machines. Each recording is a single-channel 10-sec length audio clip that includes both the sounds of the target machine and environmental sounds.

Table 1：Architecture of EfficientNet-B0 based network

Stage i	Operator F	Resolution HxW	#Channels C	#Layers L
1	Conv3x3	224 × 224	32	1
2	MBConv1， k3x3	112 × 112	16	1
3	MBConv6, k3x3	112 × 112	24	2
4	MBConv6, k5x5	56 × 56	40	2
5	MBConv6, k3x3	28 × 28	80	3
6	MBConv6, k5x5	14 × 14	112	3
7	MBConv6, k5x5	14 × 14	192	4
8	MBConv6, k3x3	7 × 7	320	1
9	Conv1x1 & Pooling & FC	7 × 7	1280	1

1.2. Audio preprocessing

Follow [4], we load the audio clips with their raw sampling rate (16,000 Hz), and the spectrogram is adopted through a Short Time Fourier Transform (STFT). We use librosa package [5] to apply STFT and mel spectrogram, the length of the window (nFFT) is 2046, the hop length is 512, so the height of the spectrogram is 1024 (1 + nFFT/2). Then spectrogram is split into 32 columns piece and each piece is normalized by subtracting the mean and dividing by the standard deviation. We use these 1024×32 shape data to train our EfficientNet_B0 network, and the 512×48 reshaped version is used to train Swim transformer, and the 128x128 mel spectrogram is used to train another EfficientNet_B0, we ensemble all the models' predictions scores finally.

2. SOLUTIONS

2.1. Training Loss

AdaCos loss [6] is employed to train our machine section classification network. AdaCos is cosine-based softmax loss. No hyperparameters are required, and the adaptive scale parameter is used to automatically strengthen training supervision during the training process. The cosine similarities between training samples and the corresponding class center vectors (fully connected vectors before softmax) can be dynamically scaled such that their predicted class probabilities satisfy the semantic meaning of these cosine similarities. We also tried ArcFace loss [7], but the score was not good enough in this challenge, we left it for future work.

2.2. Models

Two backbone architectures are incorporated: Swim Transformer based network and EfficientNet-B0.Another EfficientNet-B0 use softmax to do the classification. Figure 1 shows the architecture of our Swim Transformer based network. Table 1 shows the architecture of our EfficientNet-B0 based network. For each backbone architecture, we trained servel input shape version. Finally , we find (512,48) suitable for the Swim Transformer and (1024,32) suitable for EfficientNet-B0.

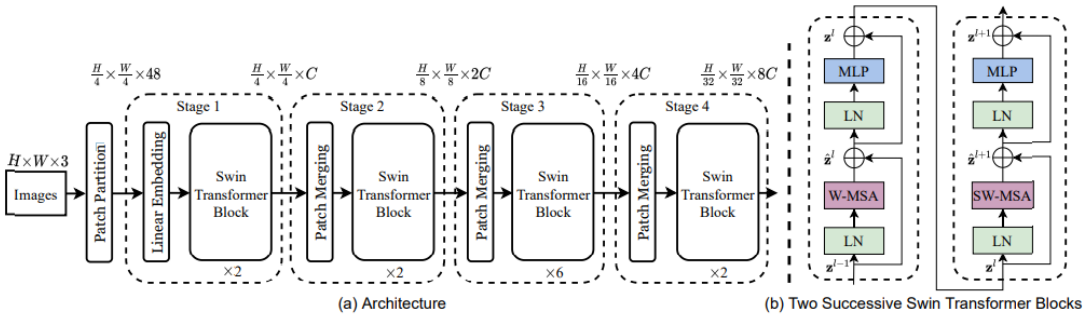

Figure 1 ： (a) The architecture of a Swin Transformer (Swin-T); (b) two successive Swin Transformer Blocks (notation presented with Eq. (3)). W-MSA and SW-MSA are multi-head self attention modules with regular and shifted windowing configurations, respectively.

2.3 Training Details

All models are trained from scratch without using any pre-trained model or external data resources. We use Adam to optimize models. The learning rate is set to 0.0001. Train models on source domain data. Spectrograms are extracted from the 10-second audio and the 1024 ×32 or 512× 48 pieces are cropped to feed to the network. Mel Spectrograms are extracted from the 10-second audio and the 128×128 pieces are cropped to feed to the network. We train each model for 150 epochs. And Table 2 shows hyper-parameters of our models.

Table 2：Summarization of hyper-parameters

Parameters for signal processing	
Sampling rate	16,000 Hz
FFT length	2046 pts
FFT hop length	512 pts
Learning strategy	
learning rate	0.0001
Other parameters	
Batch size	48
K (for embedding's similarity calculation)	10

2.4. Submissions

In Table 3, we present harmonic mean of the partial AUC and arithmetic mean of the partial AUC for 2 baseline systems and our 3 submissions. The 3 submissions are implemented as follows:

Submission1: For each machine type, we use efficientnet-B0 based model's prediction results.

Submission2: We use swim transformer based model's prediction results.

Submission3: We use efficientnet-B0 and softmax based model's prediction results.

Table 3: Evaluation results: Harmonic mean of the partial AUC[%]/Arithmetic mean of the partial AUC[%] on Development Dataset

Algorithm	Toy Car	Toy Train	bearing	fan	gearbox	slider	valve
Baseline(AE-based)	52.78/52.76	50.56/50.50	52.17/52.03	57.98/57.53	58.73/58.50	56.05/55.78	50.39/50.36
Baseline(MobileNetV2-based)	52.54/52.39	51.58/51.56	57.66/57.35	57.61/57.10	56.54/56.18	56.62/54.77	65.44/62.70
Submission1(EfficientNet-B0)	51.42/51.26	50.44/50.41	53.82/53.44	51.68/51.60	57.54/57.14	**70.43/66.56**	61.47/59.06
Submission2(SwimTransformer)	53.54/53.41	52.07/52.05	63.98/63.62	50.33/50.31	55.19/54.27	57.08/56.89	53.59/53.29
Submission3(EfficientNet-B0)	**55.18/56.89**	**52.93/54.65**	**65.14/66.34**	**59.82/61.90**	**61.97/64.53**	66.58/68.45	**66.02/79.66**

3. CONCLUSIONS

This technique report briefly presents our ensemble of AdaCos based systems for the task 2 of DCASE2022 challenge. On the basis of our 2021 Task 2 approach, we diversify CNN backbone to detect anomalies with the same threshold value regardless of the domain.

4. REFERENCES

[1] Q. Zhou, "ENSEMBLE OF ARCFACE BASED SYSTEMS FOR UNSUPERVISED ANOMALOUS SOUND DETECTION UNDER DOMAIN SHIFT CONDITIONS" Tech. report in DCASE2021 Challenge Task 2, 2021.

[2] M. Tan, Quoc V. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks" in arXiv preprint arXiv:1905.11946, 2020.

[3] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, B. Guo," Swin Transformer: Hierarchical Vision Transformer using Shifted Windows" in arXiv preprint arXiv:2103.14030,2021.

[4] O. Dong and I. Yun, "Residual Error Based Anomaly Detection Using Auto-Encoder in SMD Machine Sound," Sensors, 2018, 18(5), pp. 1308–.

[5] B. McFee, C. Raffel, D. Liang, D. P.W. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "librosa: Audio and Music Signal Analysis in Python," in Proceedings of the 14th Python in Science Conference, Kathryn Huff and James Bergstra, Eds., 2015, pp. 18 – 24.

[6] X. Zhang, R. Zhao, Y. Qiao, X. Wang, H. Li," AdaCos: Adaptively Scaling Cosine Logits for Effectively Learning Deep Face Representations"in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019,pp. 10823–10832.

[7] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "ArcFace: Additive Angular Margin Loss for Deep Face Recognition," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4690–4699, 2019.