SEMI-SUPERVISED SOUND EVENT DETECTION USING PRETRAINED MODEL

Technical Report

Rong Xie, Chuang Shi, Le Zhang and Huiyong Li

School of Information and Communication Engineering,
University of Electronic Science and Technology of China, Chengdu, China.
xierong@std.uestc.edu.cn, shichuang@uestc.edu.cn

ABSTRACT

In this technical report, submitted systems for DCASE 2022 Task4 are described. Early output embeddings of CNN14 in PANNs with a CRNN is designed to achieve a good performance on PSDS-scenario1. The fully connected (FC) layer of CNN14 is replaced by output 10 categories for PSDS-scenario 2. Submitted systms achieve an overall PSDS-scores of 1.31 (0.460 for PSDS scenario 1 and 0.856 for PSDS scenario 2) on test set.

Index Terms— Sound event detection, CNN14, CRNN, pretrained models, CBAM-T.

1. INTRODUCTION

In DCASE 2022 Task 4, a polyphonic sound event detection (SED) system is desired to recognize the categories and localizations of events in an audio signal. SED in domestic environments has great potential for surveillance[1] and healthcare monitoring[2].

Task 4 this year is more flexible compared to previous years. It allows participants to use external datasets and pretrained models. In observation, PSDS-scores[3] in scenario 1 and scenario 2 of Task 4 is proportional to even-based F1-score and audio tagging F1-score[4], respectively. Therefore, a CRNN using early output embeddings from CNN14 is designed to have a good performance on even-based metrics. For audio tagging metrics, the network is built by replacing the FC layer of CNN14. At last, a model is trained from scratch without external data and pretrained models.

For the rest of this report, in Section 2, network architectures are described for submitted systems. In Section 3 implementation details and experiment results are presented. In Section 4, conclusions are drawn.

2. PROPOSED METHOD

In this section, network architectures of submitted systems are described.

2.1. Residual Block with CBAM-T

CBAM-T is slightly changed from CBAM[5]. In our opinion, the temporal dimension for time-frequency features is more like the channel dimension in computer vision. Thus, the input features of CBAM-T are transposed on time and channel dimension, and they are transposed back at the output of CBAM-T. The CBAM-T is embedded into a Residual Block. The architectures of CBAM-T and Residual Block are shown in Fig. 1(a) and Fig. 1(b), respectively.

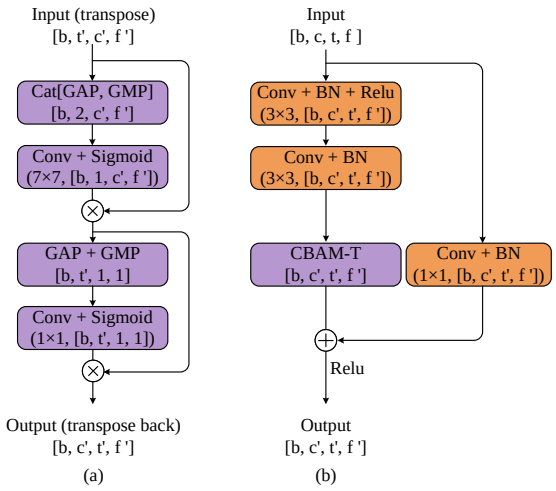

Figure 1: Residual Block and CBAM-T

2.2. Model using early output embeddings from CNN14

The architecture of model using early output embeddings from CNN14 is shown in Fig. 2. Part of CNN14 in PANNs[6] is used to output pretrained embeddings as shown in Fig. 2(a). The parameters of model in Fig. 2(a) are fixed, but dropout is reserved during the training stage. The rest of the model is composed by Residual Blocks in Fig. 1(a), Bi-GRUs, FC layers and linear softmax pooling[7].

2.3. Model for weakly predictions

In order to meet the number of classes in Task 4, the FC layer in CNN14 is replaced to output 10 categories, which is built as a weak label model.

2.4. Model trained from scratch

A model with same network architecture in Fig. 1 is trained from scratch without external data.

3. EXPERIMENT

In this section, the experimental details are firstly decribed. Then experiment results are presented.

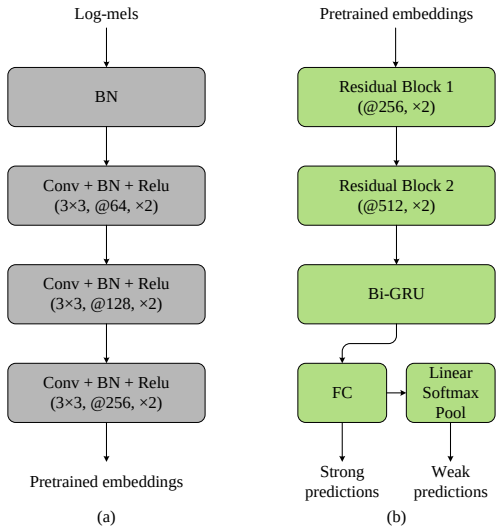

Figure 2: Model using early output embeddings of CNN14

3.1. Dataset and experimental details

In the traning sets of Task 4, there are 10000 synthetic audio clips with strong labels (synthetic training set), 3470 real audio clips with strong labels (strong set), 1578 real audio clips with weak labels (weak set) and 14412 unlabel-in-domain real audio clips (unlabel set). Validation and test set is comprises of 2500 synthetic audio clips (synthetic validation set) and 1168 real audio clips (test set) . Audio clips are downsampled from 44.1kHz to 16kHz. Log-mel spectrograms configured with 512 window size, 160 hop size and 64mel bins are used as input features.

Three systems are presented in this report,

System 1: Ensembled by 9 models with the network described in section 2.2, and trained on strong set, weak set, unlabel set. The models are trained for 200 epochs with Adam optimizer. The learing rate warmups from epoch 1 to 25, and keeps at 0.0008 from epoch 26 to 200.

System 2: Ensembled by 2 models with the network described in section 2.3, and trained on strong set, weak set. The labels of strong set are aggregated into weak labels. The mdoels are trained for 100 epochs with Adam optimizer. And the learing rate is 0.0001.

System 3: A single model with the network described in section 2.4, and trained on synthetic set, weak set and unlabel set. The model is trained for 200 epochs with Adam optimizer. The learing rate warmups from epoch 1 to 25, and keeps at 0.001 from epoch 26 to 200. The output of System 3 is the product of strong predictions and weak predictions.

Mean-teacher based semi-supervised learning methods is applied to train System 1 and 3. Furthermore, Mixup[8] and SpecAugment[9] are applied to generalize the models. And median filtering with 560ms duration is used as a post-processing method.

3.2. Results and discussions

The performances for submitted systems are shown in Table 1. System 1 has best performance on PSDS scenario 1. For PSDS scenario 1, DTC and GTC are 0.7. Thus, PSDS score on scenario 1 and event-based F1 are approximately proportional. System 2 has best performance on PSDS scenario 2. For PSDS scenario 2, DTC and GTC are 0.1. If weak predictions are right and events last longer than 1sec in 10sec audio clips, then they will be counted as true positives. Therefore, PSDS scenario 2 is close to the audio tagging task in Task4. Moreover, System 2 has the best performance on intersection-based F1 (IB F1) and System 1 has the best performance on event-based F1 (EB F1).

Table 1: Performances of submitted systems on real test set

System	PSDS1	PSDS2	IB F1	EB F1
Baseline	0.336	0.536	64.1%	40.1%
System 1	**0.460**	0.768	86.7%	**56.0**%
System 2	0.072	**0.856**	**90.9**%	22.8%
System 3	0.360	0.674	85.6%	39.7%

The performance on test set and synthetic validation set of a model from System 3 is compared in Table 2, which could explain why synthetic set is not involved in the training procedure for System 1 and 2. Differences in the distribution of synthetic audios and real audios may lead to poor performance on real audios.

Table 2: Performance on test set and synthetic validation set for System 3

Training set	Test set name	PSDS1	PSDS2
Synthetic set, weak set, unlabel set	Test set	0.329	0.613
	Synthetic validation set	0.467	0.696
Strong set, weak set, unlabel set	Test set	0.410	0.663
	Synthetic validation set	0.328	0.496

4. CONCLUSION

In conclusion, pretrained model with more training data is helpful in improving the performance of SED task. Moreover, using synthetic audios as training data may degrade the performance of SED systems on real audios due to the difference in data distribution between synthetic audios and real audios.

5. REFERENCES

[1] P. Foggia, N. Petkov, A. Saggese, N. Strisciuglio, and M. Vento, "Reliable detection of audio events in highly noisy environments," *Pattern Recognition Letters.*, vol. 65, pp. 22–28, 2015.

[2] S. Goetze, J. Schroder, S. Gerlach, D. Hollosi, J.-E. Appell, and F. Wallhoff, "Acoustic monitoring and localization for social care," *Journal of Computing Science and Engineering.*, vol. 6, no. 1, pp. 40–50, 2012.

[3] Ç. Bilen, G. Ferroni, F. Tuveri, J. Azcarreta and S. Krstulović, "A framework for the robust evaluation of sound event detection," in *Proc. of IEEE ICASSP*, 2020, pp. 61–65.

[4] M. Annamaria, T. Heittola, and T. Virtanen, "Metrics for polyphonic sound event detection," *Applied Sciences.*, 6.6: 162, 2016.

[5] S. Woo , J. Park , J. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in *Proc. of ECCV*, 2018, pp. 3–19

[6] 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.

[7] Y. Wang, J. Li, and F. Metze, "A comparison of five multi-ple instance learning pooling functions for sound event detec-tion with weak labeling," in *Proc. of IEEE ICASSP*, 2019, pp. 31–35.

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

[9] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "SpecAugment: A simple data augmentation method for automatic speech recognition," *arXiv preprint arXiv:1904.08779*, 2019.