GUIDED MULTI-BRANCH LEARNING SYSTEMS FOR DCASE 2020 TASK 4

Technical Report

Yuxin Huang[1,2], *Liwei Lin*[1,2], *Shuo Ma*[1], *Xiangdong Wang*[1*], *Hong Liu*[1], *Yueliang Qian*[1], *Min Liu*[3], *Kazushige Ouchi*[3]

[1] Bejing Key Laboratory of Mobile Computing and Pervasive Device,
Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China,
{huangyuxin18g, linliwei17g}@ict.ac.cn, ms_hebut@163.com, {xdwang, hliu, ylqian}@ict.ac.cn
[2] University of Chinese Academy of Sciences, Beijing, China
[3] Toshiba China R&D Center, Beijing, China,
liumin@toshiba.com.cn, kazushige.ouchi@toshiba.co.jp
*Corresponding author

ABSTRACT

In this paper, we describe in detail our systems for DCASE2020 task 4. The systems are based on the first-place system of DCASE 2019 task 4, which adopts the multiple instance learning (MIL) framework with embedding-level attention pooling and a semi-supervised learning approach called guided learning. The multi-branch learning method is then incorporated into the system to further improve the performance. Multiple branches with different pooling strategies (embedding-level or instance-level) and different pooling modules (attention pooling, global max pooling or global average pooling) are used and shares the same feature encoder. To better exploit the synthetic data with strong labels, inspired by multi-task learning, a sound event detection (SED) branch is also added. Therefore, multiple branches pursuing different purposes and focusing on different characteristics of the data can help the feature encoder model the feature space better and avoid over-fitting. To combine sound separation with sound event detection, we train models using the output of the baseline system of sound separation and fuse the event detection results of models with of without sound separation.

***Index Terms*—** Multi-branch learning, multi-task learning, guided learning, specialised decision surface, attention

1. INTRODUCTION

DCASE 2020 task 4 [1] is the follow-up to DCASE 2019 task 4 [2]. While DCASE 2019 task 4 targets on exploring the usage of weakly labeled data, unlabeled data and synthetic data in sound event detection (SED), DCASE 2020 task 4 encourages participants to combine sound separation with SED in addition to the same task in DCASE 2019. There are three subtasks in DCASE 2020 task 4: SED without sound separation, SED with sound separation and sound separation (using the SED baseline system). We participated in the first two subtasks. However, for the second subtask, we just use the baseline system for sound separation provided by the challenge organizer and focus on combination of sound separation and SED.

In this paper, we describe in detail our systems for the two subtasks we participated in DCASE2020 task 4. The systems are based on the first-place system of DCASE 2019 task 4 developed by Institute of Computing Technology (ICT), Chinese Academy of Sciences (CAS) [3], which adopts the multiple instance learning (MIL) framework with embedding-level attention pooling [4] and a semi-supervise learning approach called guided learning [5]. The multi-branch learning approach [6] is then incorporated into the system to further improve the performance. Multiple branches with different pooling strategies (embedding-level or instance-level) and different pooling modules (attention pooling, global max pooling or global average pooling) are used and shares the same feature encoder. To better exploit the synthetic data with strong labels, inspired by multi-task learning [7], a sound event detection (SED) branch is also added. Therefore, multiple branches pursuing different purposes and focusing on different characteristics of the data can help the feature encoder model the feature space better and avoid over-fitting. To incorporate sound separation into sound event detection, we train models using output of the baseline system of sound separation and fuse the event detection results of models with of without sound separation.

2. THE DCASE 2019 TASK 4 SYSTEM BY ICT

Our systems for DCASE 2020 task 4 follows the framework of the DCASE 2019 task 4 system by ICT [3], which won the 1st place and the Reproducible system award in the DCASE 2019 task 4 challenge. The system utilizes convolutional neural network (CNN) with embedding-level attention pooling module for weakly-supervised SED and uses disentangled features to solve the problem of unbalanced data with co-occurrences of sound events [4]. To better use the unlabeled data jointly with weakly-labeled data, the system adopts a semi-supervised learning method named Guided Learning [5], which uses different models for the teacher model and student model to achieve different purposes implied in weakly-supervised SED. For the synthetic data, the system regards them as weakly annotated training set and the time stamps of sound events in the strong labels are not used. The system is trained by the DCASE 2019 training data, including weakly-labeled data, synthetic data and unlabeled data without data augmentation. The system won the 1st place in DCASE 2019 task 4 were the fused system of 6 systems with the same model architecture, and the method for fusion is averaging all the probabilities output by the systems.

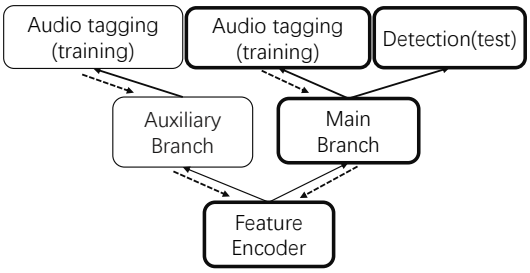


Figure 1: The multi-branch learning method

3. METHOD

3.1. Multi-branch learning for semi-supervised SED

To further improve the performance, the multi-branch learning approach[6] is incorporated into the system. Multiple branches with different pooling strategies (embedding-level or instance-level) and different pooling modules (attention pooling, global max pooling or global average pooling) are used and shares the same feature encoder. One branch is set as the main branch which takes part in training and detection and another one branch is set as the auxiliary branch which is only used for training. In our system, we choose the embedding-level ATP as the main branch and instance-level GMP or instance-level GAP as the auxiliary branch.

3.2. The detection branch for synthetic data

To better exploit the synthetic data with strong labels, inspired by multi-task learning, a sound event detection (SED) branch is also added. For all training data, only the synthetic data is used for training the SED branch and the output of the SED branch is probabilities of each instance.

3.3. Data augmentation

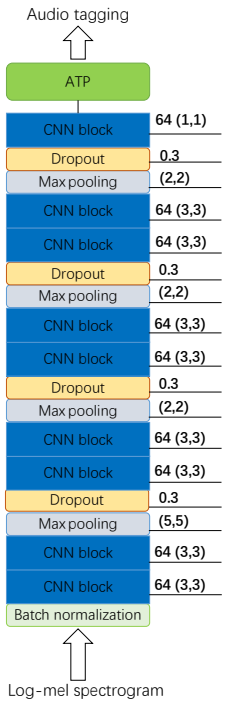
This year, data augmentation is used for training of the models. For all training data, including weakly-labeled data, unlabeled data and synthetic data, we use time-shifting and frequency-shifting to generate augmented data. We set the ratio between the original data and the augmentation data to be 8:1.

3.4. System fusion

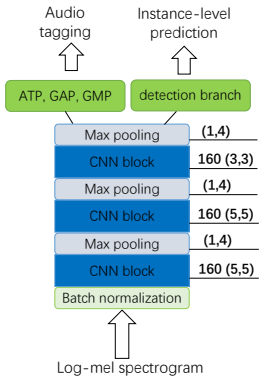
For the systems of which the auxiliary branch is instance-level GAP (global average pooling), we take the average value of the main branch and the GAP branch as the output of the system. For system ensemble, we take the weighted sum of all the system outputs as the final results.

3.5. Combination of sound separation and SED

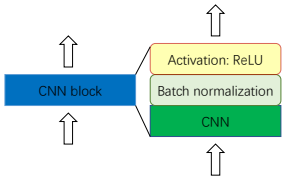
To incorporate sound separation into sound event detection, we train some SED models by the separated data output from the baseline system of sound separation. Then, we fuse the event detection results of models trained by real data and separated data to get the final SS-SED ensemble system result.



(a) PT



(b) PS



(c) CNN Block

Figure 2: The model architectures

4. SYSTEM

4.1. System overview

We use the guided learning (GL) framework which is composed of a professional teacher model (PT-model) and a promising student model (PS-model) as our basic model architecture. The results of all models are combined to get the final system results.

4.2. Model architecture

As shown in Figure 2, for an individual system in our system, we use the guided learning architecture. We mainly modified the PS model, in which multiple branches are added. To make each model be more different from others, We also tried to add the detection branch which only trained with the synthetic data in some systems.

4.3. Model training

Different from the guided learning (GL) system, we add auxiliary branch to the PS model. As a result, the loss function of the PS model is:

$$L_{PS-total} = \alpha L_{PS-main} + \sum \beta L_{PS-auxiliary} \quad (1)$$

For the detection branch, we only use the strong labeled synthetic data to train this branch. The loss function of the detection branch is cross entropy.

$$L = -\sum_{c}\sum_{t}(y_{ct}log(\hat{\mathbf{P}}(y_{ct}|\mathbf{x_t}))+(1-y_{ct})log(1-\hat{\mathbf{P}}(y_{ct}|\mathbf{x_t}))) \quad (2)$$

Table 1: The event-based F1 score on validation set

Model	Average F1	Best F1
E-ATP	0.421 ± 0.0115	0.444
E-ATP + I-GMP	0.430 ± 0.0088	0.445
E-ATP + I-GAP	0.431 ± 0.0156	0.451
SED-Ensemble	-	0.467
SS-SED Ensemble	-	0.472

Table 2: The event-based F1 score on public test set

Model	Average F1	Best F1
E-ATP	0.449 ± 0.0124	0.47
E-ATP + I-GMP	0.458 ± 0.0125	0.478
E-ATP + I-GAP	0.450 ± 0.0130	0.470
SED-Ensemble	-	0.497
SS-SED Ensemble	-	0.495

5. EXPERIMENT

5.1. Experimental setup

The training set of our SED system contains a weakly-labeled training set (1578 clips), an unlabeled training set (14412 clips), and a synthetic strongly labeled set (2584 clips). The validation set contains 1168 strongly-labeled clips. The public test set contains 692 strong-labeled clips. For the SS-SED system, we use the baseline system of source separation to separate the training set of the SED. The separated data is used to train the SS-SED system. We report the event-based marco F1 score [8]. All the experiments are repeated 20 times with random initiation and we report both the average result and the best result of each model.

5.2. Experimental results

Experimental results are shown in Table 1 and 2. For each type of experiment, we do not change the PT model and only change the PS model. In the table, E-* denotes the embedding-level approach and I-* denotes the instance-level approach. We find that adding multi-branch such as I-GMP or I-GAP can have a beneficial effect. For the ensemble system, we use 3 E-ATP + I-GMP and 3 E-ATP + I-GAP to construct it. Besides, to make the difference between models larger, 2 of 3 E-ATP + I-GMP models with detection branches. The ensemble system achieves F1 score of 0.497 on public test set and 0.467 on the validation set. For the SS-SED ensemble system, besides the 6 models used in SED-Ensemble, 2 E-ATP + I-GMP models which are trained by separated data are used and the SS-SED ensemble system achieves F1 score of 0.495 on the public test set and 0.472 on the validation set.

6. CONCLUSIONS

This paper presents the details of our systems for DCASE2020 task 4. The systems are based on the first-place system of DCASE 2019 task 4, which adopts the multiple instance learning (MIL) framework with embedding-level attention pooling and a semi-supervise learning approach called guided learning. The multi-branch learning approach is then incorporated into the system to further improve the performance. Multiple branches with different pooling strategies (embedding-level or instance-level) and different pooling modules (attention pooling, global max pooling or global average pooling) are used and shares the same feature encoder. To better exploit the synthetic data with strong labels, inspired by multi-task learning, a sound event detection (SED) branch is also added. Therefore, multiple branches pursuing different purposes and focusing on different characteristics of the data can help the feature encoder model the feature space better and avoid over-fitting.

7. REFERENCES

[1] http://dcase.community/challenge2020/task-sound-event-detection-and-separation-in-domestic-environments.

[2] http://http://dcase.community/challenge2019/task-sound-event-detection-in-domestic-environments.

[3] L. Lin, X. Wang, H. Liu, and Y. Qian, "Guided learning convolution system for dcase 2019 task 4," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2019 Workshop (DCASE2019)*, 2019, pp. 134–138.

[4] L. Lin, X. Wang, H. Liu, and Y. Qian, "Specialized decision surface and disentangled feature for weakly-supervised polyphonic sound event detection," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 1466–1478, 2020.

[5] L. Lin, X. Wang, H. Liu, and Y. Qian, "Guided learning for weakly-labeled semi-supervised sound event detection," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 626–630.

[6] Y. Huang, X. Wang, L. Lin, H. Liu, and Y. Qian, "Multi-branch learning for weakly-labeled sound event detection," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 641–645.

[7] X. Xia, R. Togneri, F. Sohel, Y. Zhao, and D. Huang, "Multi-task learning for acoustic event detection using event and frame position information," *IEEE Transactions on Multimedia*, vol. 22, no. 3, pp. 569–578, 2020.

[8] A. Mesaros, T. Heittola, and T. Virtanen, "Metrics for polyphonic sound event detection," *Applied Sciences*, vol. 6, no. 6, p. 162, 2016.