INTEGRATING ADVANTAGES OF RECURRENT AND TRANSFORMER STRUCTURES FOR SOUND EVENT DETECTION IN MULTIPLE SCENARIOS

Technical Report

Rui Lu[1]*, Wenzheng Hu*[2]*, Zhiyao Duan*[1]*, Ji Liu*[1]

[1] Beijing Kuaishou Technology Co., Ltd, {lurui, zhiyaoduan}@kuaishou.com, jiliu@kwai.com
[2] The State Key Laboratory of Automotive Satefy and Energy, Tsinghua University, Beijing, China, hwz@mail.tsinghua.edu.cn

ABSTRACT

In this technical report, we detail our submitted systems for task4 of DCASE2021: Sound Event Detection and Separation in Domestic Environments. Our systems exploit both recurrent structure and transformer structure to model the complicated dynamics in real life domestic audio data. In addition to prevalent tricks such as semi-supervised mean-teacher learning, data augmentation and ensemble, we find that different models exhibit differently under the two scenarios, which emphasize different system properties. By integrating advantages of both the recurrent and transformer structures, our proposed systems achieve an overall poly-phonic sound event detection scores (PSDS-scores) of **1.171** (PSDS-scenario1 + PSDS-scenario2) on the hold-out test set of the development dataset, outperforming the baseline system by **34.8%**.

***Index Terms*—** Sound event detection, convolutional recurrent neural network, transformer, semi-supervised learning

1. INTRODUCTION

In this technical report, we describe our submitted systems for task 4 of the DCASE2021 challenge: Sound Event Detection and Separation in Domestic Environments [1]. Different from previous challenges, this year's evaluation criteria takes different real-life scenarios into consideration, thus leading to two different settings: scenario1 emphasizes fast reaction of systems upon sound events while scenario2 imposes more penalty on the confusion between sound event classes [2, 3]. Based on the experimental observations, we propose to deal with each scenario by specific model:

- Scenario1: we exploit the popular convolutional recurrent neural network (CRNN) [4] to capture the temporal variations of audio signal, providing fast reaction upon event detection.
- Scenario2: we make use of the newly proposed convolution-augmented transformer (Conformer) [5, 6] which shows favor for avoiding confusion between sound event classes.

Along with the above models, we implement following tricks to improve the model performance:

- Mean-teacher training framework [7] to fully exploit the unlabelled in-domain data for better semi-supervised learning.
- Mixup and frame-shift [8] data-augmentation strategies to improve the generalization ability of detection systems.
- Ensemble [9] and class-wise median-filter to reduce model variance and smooth probability predictions.

We carry out sufficient experiments on the development dataset of DCASE2021 task4 to verify the effectiveness of our proposed SED system. Evaluation results on the test dataset show that the proposed scenario-specific detection strategy can significantly improve the overall performance. We achieve an overall PSDS-score of **1.171**, which outperforms the baseline by **34.8%**.

This technical report is organized as follows: Section. 2 details the models and tricks we use to train the SED systems; In Section 3, we demonstrate the effectiveness of our proposed scenario-specific detection strategy through adequate experiments; Finally, we conclude in Section 4.

2. METHOD

2.1. Data

We train and evaluate the proposed models on the development dataset of DCASE2021 task4, summarized as following:

- **Weak-Train**: 1578 weakly annotated samples
- **Unlabeled-Train**: 14412 unlabeled in domain samples
- **Synthetic-Train**: 10000 synthetic strongly labeled samples
- **Strong-Valid**: 1168 strongly labeled samples
- **Synthetic-Valid**: 2500 synthetic strongly labeled samples

We take all the **Unlabeled-Train**, **Synthetic-Train** and part of the **Weak-Train** for training; All of the **Synthetic-Valid** and part of the **Weak-Train** constitutes the validation set; We evaluate model performance on the **Strong-Valid**, which is held-out during the whole training process.

2.2. Feature

We follow the baseline system to extract log-mel features with hop size of 256 and window length of 2048, on the resampled 16kHz audio data. 128 mel-filters are applied to obtain the final frame-wise features. For normalization, we calculate the mean and standard deviation of log-mel features across all the samples in **Weak-Train** and **Unlabeled-Train**, these statistics are used during both training and inference time.

Besides the log-mel features, we also experiment with mfcc and pitch features [9], which nevertheless brings no improvements. Neither do the data augmentation strategies such as time-stretch and pitch-shift [10] work in the current task.

2.3. Model

As stated above, this year's task4 takes two scenarios into consideration: scenario1 requires fast reaction upon the occurrences of sound events while scenario2 is more sensitive to confusion between different event classes. We propose a scenario specific strategy to deal with these two scenarios separately: CRNN exhibits higher PSDS-scenario1 score and Conformer performs better under scenario2. Structures of the proposed CRNN and Conformer are detailed as following:

CRNN: The convolutional feature extractor of the CRNN is a stack of 7 convolution layers, each with kernel size of $(3, 3)$ and stride size of $(1, 1)$. Each convolution block is followed by ReLU and batch-normalization. Moreover, average-pooling with kernel size of 2 is applied along the frequency-axis after each block. On the temporal dimension, we only apply average-pooling after the first two convolution blocks, in order to maintain adequate resolution. Consequently, the output resolution of the CRNN model is reduced by 4-times. A two-layer bidirectional-gru with hidden size of 128 is stacked upon the feature extractor, the outputs of which are feed into a fully-connected layer and sigmoid non-linearity to predict the probabilities of the 10 sound event classes.

Conformer: We exploit similar feature extractor with that of the CRNN, except that we use glu as the activation functions and apply one more average-pooling layer along the temporal dimension, resulting in resolution reduction of 8 times. We substitute gru with a 7-block conformer structure [5, 6] to better capture the distinctions between sound event classes.

2.4. Model Training

Despite the BCE loss on strongly-labelled data and weakly-labelled data, we follow [7] to apply mean-teacher semi-supervised learning method to impose consistency constraint (MSE-loss) on the teacher model and student model, thus fully exploiting the unlabelled data. All models are trained with 200 epochs using Adam.

For the synthetic validation data, we compute the PSDS values using 10 operating points, linearly distributed from 0.1 to 0.9. By this means, we can better align with the final criterion on test dataset. For the weakly labelled validation data, we compute the class-averaged F1-score (macro-F1). Above two metrics are further accumulated for best model checkpoint. Mixup and frame-shift [8] are applied on the data separately, each with a probability of 0.5.

2.5. Ensemble and Post-processing

We randomly split the **Weak-Train** dataset into 5-fold for ensemble. On each division of the dataset, models (both CRNN and Conformer) are trained from scratch and the probabilities of which are averaged during the inference process. Median filter is applied to smooth the prediction results. For each sound event, we search for the optimal median filter length from 1 to 49 with increment of 2.

3. EXPERIMENTS

As shown in Table. 3, we exhibit the evaluation metrics of our proposed CRNN and Conformer model, togher with that of the baseline model. For clarity, we omit experiments of model structure search and feature selection, only show the best results with and without data augmentation.

Method	PSDS1	PSDS2	Intersection-F1
baseline	0.342	0.527	76.6%
CRNN	0.408	0.622	79.4%
Conformer	0.188	0.728	74.5%
CRNN + DA	**0.419**	0.638	**80.5%**
Conformer + DA	0.172	**0.752**	74.3%

Table 1: Performance of our proposed CRNN and Conformer model. "DA" indicates data augmentation (mixup and frame-shift).

As can be seen, for scenario1, the best performance is achieved by CRNN with data augmentation; For scenario2, the best performance is achieved by Conformer with data augmentation. Experimental results show that the CRNN model has more advantages when accurate time boundary detection of sound events is required, while Conformer shows more favor for circumstances when confusion between different sound events is given more concern. Finally, by considering the best performance in each specific scenario, we achieve an overall PSDS-score (PSDS1-score + PSDS2-score) of **1.171**, outperforming the baseline by **34.8%**.

4. CONCLUSION

In this technical report, we detail our approach for DCASE2021-task4. Through thorough experiments, we find that when we consider different scenarios that emphasize different system properties, a more favorable strategy is to exploit scenario-specific network structure for sound event detection. By taking the advantages of scenario-specific models, we can achieve significant better performance in terms of the overall metric.

5. REFERENCES

[1] N. Turpault, R. Serizel, A. Parag Shah, and J. Salamon, "Sound event detection in domestic environments with weakly labeled data and soundscape synthesis," in *Workshop on Detection and Classification of Acoustic Scenes and Events*, New York City, United States, October 2019. [Online]. Available: https://hal.inria.fr/hal-02160855

[2] A. Mesaros, T. Heittola, and T. Virtanen, "Metrics for polyphonic sound event detection," *Applied Sciences*, vol. 6, no. 6, p. 162, 2016. [Online]. Available: http://www.mdpi.com/2076-3417/6/6/162

[3] C. Bilen, G. Ferroni, F. Tuveri, J. Azcarreta, and S. Krstulovic, "A framework for the robust evaluation of sound event detection," *arXiv preprint arXiv:1910.08440*, 2019. [Online]. Available: https://arxiv.org/abs/1910.08440

[4] E. Cakır, G. Parascandolo, T. Heittola, H. Huttunen, and T. Virtanen, "Convolutional recurrent neural networks for polyphonic sound event detection," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 6, pp. 1291–1303, 2017.

[5] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, *et al.*, "Conformer: Convolution-augmented transformer for speech recognition," *arXiv preprint arXiv:2005.08100*, 2020.

[6] K. Miyazaki, T. Komatsu, T. Hayashi, S. Watanabe, T. Toda, and K. Takeda, "Convolution-augmented transformer for

semi-supervised sound event detection," DCASE2020 Challenge, Tech. Rep., June 2020.

[7] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," *arXiv preprint arXiv:1703.01780*, 2017.

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

[9] R. Lu and Z. Duan, "Bidirectional gru for sound event detection," *Detection and Classification of Acoustic Scenes and Events*, 2017.

[10] J. Salamon and J. P. Bello, "Deep convolutional neural networks and data augmentation for environmental sound classification," *IEEE Signal Processing Letters*, vol. 24, no. 3, pp. 279–283, 2017.