CONVOLUTIONAL NETWORK WITH CONFORMER FOR SEMI-SUPERVISED SOUND EVENT DETECTION

Technical Report

Tong Na

Beijing, 100191, China
natong19@outlook.com

Qinyi Zhang

Beijing, 100191, China
735207834@qq.com

ABSTRACT

In this technical report, we describe our system submission for DCASE 2021 Task 4. Our model employs a convolutional network in conjunction with conformer blocks and utilizes the Mean-Teacher semi-supervised learning technique for further improvement.

Index Terms— Sound event detection, CNN, conformer, Mean-Teacher, semi-supervised learning.

1. INTRODUCTION

The goal of DCASE 2021 Task 4: sound event detection and separation in domestic environments [1] is to build systems that can detect and identify sound events. Given a sound sequence, such systems are expected to produce strong labels indicating the sound events' categories, onsets and offsets, but are trained using unlabeled, weakly labeled and synthetic strongly labeled data. To address this task, we propose a network composed of a CNN and conformer blocks [2]. By stacking these structures, the model can capture features from both local and global contexts. To further enhance performance, Mean-Teacher semi-supervised learning [3] and median filter post-processing are applied.

2. DATASET

2.1. DESED

DESED [4], a dataset for SED in domestic environments, is composed of 10 second audio clips that are recorded or synthesized. The training/validation subset consists of unlabeled data, weakly labeled data (labels on a clip scale) and synthetic strongly labeled data (labels on a frame scale). The public evaluation subset contains audio that are extracted from YouTube and Vimeo.

2.2. Evaluation Metrics

Systems are evaluated according to PSDS scores that are calculated from two scenarios and then normalized by the baseline PSDS. The first scenario tests the system's ability to react in time upon event detection while the second scenario requires the system to distinguish between event classes. Additionally, event-based F1 scores and intersection-based F1 scores are computed as a contrastive measure.

3. PROPOSED METHOD

3.1. Mel spectrogram

We extract log-mel spectrograms without source separation pre-processing. The audio clips are zero-padded/truncated to 10 seconds, then converted to single channel and resampled at 16 kHz. The window size is 2048, the hop length is 256 and the number of mels is 128. Then, the mel spectrograms for each bin are normalized based on the global minimum and maximum for that bin.

3.2. Network architecture

Our model is inspired by [5] and consists of three modules: a CNN, a conformer module and a dense layer. The architecture of the 7-layer CNN feature extractor follows that of the DCASE 2021 Task 4 baseline system [6]. The conformer module includes 4 conformer blocks, whose structure is detailed in [7] and implemented in [8]. The CNN focuses on local features while the conformer module considers global features. Finally, the dense layer is used for producing the final outputs corresponding to the event types.

3.3. Semi-supervised learning

The mean-teacher semi-supervised learning technique is employed to take advantage of the large amount of unlabeled training data. The model is a combination of two models: a student model and a teacher model. The two models share the same structure, but the weights of the student model are learned through training, and the teacher model's weights are updated through an exponential moving average of the student model's weights, rather than through back propagation. We set the number of ramp-up epochs and consistency cost to 50 and 2, respectively.

3.4. Post-processing

A median filter with fixed length was applied to smooth the output sequence.

4. EXPERIMENTAL RESULTS

Method	PSDS-Scenario1	PSDS-Scenario2
Baseline	0.320	0.504
Conformer	0.315	0.516
Conformer w/ dropout	0.256	0.479
Conformer w/ layernorm	0.313	0.535

We first train a model from scratch using the baseline code. The PSDS scores of the model are lower than expected possibly due to incomplete training and validation data. For the conformer model, we replace the RNN layers and the dropout layer from the baseline model with 4 conformer blocks. For the conformer model with dropout, we replace only the RNN layers and keep the dropout layer. For the conformer model with layernorm, we add layer normalization to each feed-forward module in the conformer blocks. The conformer model with layernorm achieves the best performance, scoring 0.313 and 0.535 in the two PSDS scenarios, respectively.

5. CONCLUSION

In this technical report, we described a system for DCASE 2021 Task 4. The system, composed of a CNN and conformer blocks, exhibits comparable performance compared to the baseline and demonstrates that conformer blocks are a viable substitution for RNNs for the purpose of sound event detection.

6. REFERENCES

[1] http://dcase.community/workshop2021/.

[2] A. Gulati, J. Qin, C.-C. Chiu, et al., “Conformer: Convolution-augmented transformer for speech recognition,” arXiv preprint arXiv:2005.08100, 2020.

[3] A. Tarvainen and H. Valpola, “Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results,” in NIPS, 2017, pp. 1195– 1204.

[4] https://project.inria.fr/desed

[5] Miyazaki, Koichi et al. “CONVOLUTION-AUGMENTED TRANSFORMER FOR SEMI-SUPERVISED SOUND EVENT DETECTION Technical Report.” (2020).

[6]https://github.com/DCASE-REPO/DESED_task/tree/master/recipes/dcase2021_task4_baseline

[7] Gulati, Anmol et al. “Conformer: Convolution-augmented Transformer for Speech Recognition.” ArXiv abs/2005.08100 (2020): n. pag.

[8] https://github.com/lucidrains/conformer