SOUND EVENT DETECTION BASED ON SELF-SUPERVISED LEARNING OF WAV2VEC 2.0

Technical Report

Hyejin Koo, Hyung-Min Park

Dept. of Electronic Engineering,
Sogang University
Seoul 04107, South Korea
{a9a3710, hpark}@sogang.ac.kr

Jonghyeon Park, Myungwoo Oh

NAVER Corp.
Gyeonggi-do 13561, South Korea
{myungwoo.oh, jong-hyeon.park}
@navercorp.com

ABSTRACT

In this report, we present our system for DCASE2021 Task4: Sound Event Detection (SED) and Separation in Domain Environments. This task evaluates how to capture information of SED with a relatively small amount of labeled data in addition to lots of unlabeled data. We apply wav2vec 2.0 on the SED for the first time. Even though wav2vec 2.0 pre-training using the DCASE2021 Taksk4 dataset spends long time to train audio representations, the presented model achieved higher intersection F1 and PSDS2. The baseline's mean-teacher model and dataset was used to compare wav2vec 2.0 and log-mel features. Under the same conditions, we present how wav2vec 2.0 features work on the SED task.

***Index Terms*—** Sound event detection, wav2vec 2.0, self-supervised learning, DCASE2021 Task4

1. INTRODUCTION

Nowadays, deep learning has provided impressive or promising results on various tasks. Plenty of data and increased computational power may lead us to use complex deep neural networks, but labeling data requires that much cost. Meanwhile, self-supervised learning makes us not to prepare lots of labeled data. We can pre-train the model with unlabeled data, and only a few labeled data is needed to fine-tune the model for its task. Among several self-supervised learning (SSL) model, we use wav2vec 2.0 [1] to train SED. The external data is prohibited, so the submitted model's performance not dramatically outperforms the baseline. However, it is the first trial to apply wav2vec 2.0 to an SED task, and intersection F1 score and PSDS2 are higher than the baseline system [2]. We propose SED model and a pre-training method. To compare wav2vec 2.0 and log-mel features, the baseline's mean teacher system [3] and dataset was used.

2. SELF-SUPERVISED LEARNING AND SOUND EVENT DETECTION NETWORK

2.1. Self-Supervised Learning

Wav2vec 2.0 [1] is a self-supervised framework for speech representation learning. For pre-training of wav2vec 2.0, only audio data are required, without any labeling. After pre-training speech representation, fine-tune the model to learn speech recognition with few labeled data, as proposed by Facebook AI research. Not only speech recognition also other tasks such as speaker identification explored wav2vec 2.0 [4]. However, we introduce this to an SED task, to figure out wav2vec 2.0 could capture sound information, much more comprehensive than speech.

One of the challenge's task rules was not to use external data, so we just used the provided data about 72 hours for pre-training. Among several options to pre-train the wav2vec 2.0, we made difference on latent-groups and latent-variables, which is also known as codebook groups and variances. For speech recognition, previous works used {groups, variables} = {2, 320}, and we experimented with {groups, variables} = {2, 16}, {4, 320}, and {8, 320}. The self-supervised learning loss was minimized with 8 groups and 320 variables. Each pre-training took around 20 days with an RTX 3090 GPU card. For pre-training wav2vec 2.0, fairseq [5] code is referenced.

2.2. Sound Event Detection Network

Since the object of this research is to figure out wav2vec 2.0 can exploit sound representations for event detection, we tried to modify the baseline system to a minimum. For that reason, we applied the basic 3 networks, convolutional recurrent neural network (CRNN, the same as the baseline), recurrent neural network (RNN), and fully connected network (FCN) after the pre-trained wav2vec 2.0 model so that we named each of them 'WCRNN', 'WRNN', and 'WFCN' which are abbreviation for 'wav2vec 2.0 CRNN', 'wav2vec 2.0 RNN', and 'wav2vec FCN'. These models are shown on Fig. 1. Other set-up was the same as the baseline.

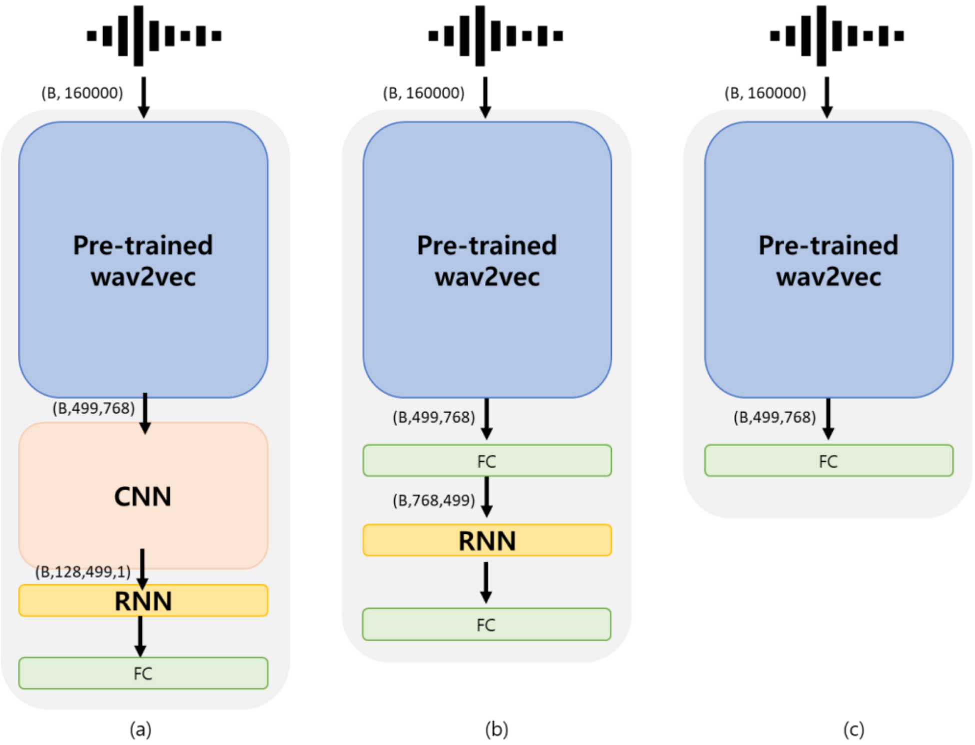

Figure 1: SED networks: (a) WCRNN, (b) WRNN, and (c) WFCN.

Table 1: Experimental results

Methods	***PSDS1***	***PSDS2***
Baseline	0.332	0.526
WCRNN model	0.056	0.206
WRNN model 1	0.295	0.503
WRNN model 2	**0.316**	0.337
WFCN model	0.068	0.629
Ensemble model	0.249	**0.711**

2.2.1. WCRNN

After the wav2vec 2.0, the output was fed into 7 CNN blocks and 2 bidirectional gated recurrent unit (BiGRU) blocks as the baseline CRNN model. We used 128 GRU cells in each BiGRU layer. To match the weight size, we just made difference on pooling [[1,4] *2, [1,2] *5] from [[2,2] *2, [1,2] *5].

2.2.2. WRNN

After the wav2vec 2.0, the output was fed into BiGRU blocks. We experimented on 1, 2, and 3 BiGRU layers, and the 1 layer provided the best performance. To use 128 GRU cells or 256 GRU cells, the wav2vec 2.0 output had to pass through a linear layer whereas the linear layer was not necessary when using 768 GRU cells. Figure 1(b) shows the WRNN model with 768 cells which provided better results than those with the other numbers of cells.

2.2.3. WFCN

After the wav2vec 2.0, the output was fed into a fully connected layer to make a prediction. Before passing through the fully connected layer, every output of the wav2vec 2.0 was fed into an activation function. In most cases, the ReLU function was better than the sigmoid function.

3. EXPERIMENTAL RESULTS

For the DCASE2021 Challenge Task4, the submitted system was evaluated in terms of polyphonic SED scores (PSDS) [5]. Two types of them, PSDS1 and PSDS2 are measures in this challenge. Several selected experimental results are shown in Table 1.

As the wav2vec 2.0 extracted features for audio sound, convolutional layers might not be required after the wav2vec. Rather, the redundancy of the wav2vec 2.0 and convolutional layers might cause degraded performance in the WCRNN model. Instead, direct feeding into an RNN layer made WRNN models achieve higher scores on both the PSDS1 and PSDS2. The WRNN models 1 and 2 used one BiGRU layer with 768 cells and a reversed GRU layer with 768 cells, respectively. Removing the RNN layer, the WFCN model showed a low PSDS1 but a high PSDS2. This implies that the event detection task in a small segment is similar to multi-class classification.

We made an ensemble model using the WRNN and WFCN models to obtain the best score on PSDS2. To fuse the two models, predictions were added and divided by two. Also, this ensemble model achieved significantly improved PSDS1 than a single WFCN model.

4. CONCLUSION

As the first case to apply wav2vec 2.0 to an SED task, our presented system demonstrated that the wav2vec 2.0 could learn representations of audio information much more comprehensive than speech.

From the experimental results, some limitations were shown. There were no better scores on PSDS1 than that of the baseline, which meant that our models could not detect on/offsets effectively. Also, when PSDS2 increases, PSDS1 tends to decrease. Therefore, it was not easy to develop a system that showed high scores on both the measures.

On the other hand, the wav2vec 2.0 is a huge pre-training model, and we could not use enough data in this challenge. Although around 72-hour audio might not be sufficient for pre-training, the model outperformed the baseline in PSDS2 with slightly lower scores in PSDS1. Our object was to figure out whether the wav2vec 2.0 could capture general sound representations, and we believe the outperforming results demonstrated important evidence. For the next work, we will use more data for wav2vec 2.0 pre-training and obtain results on SED tasks. Regardless of labeling, pre-training with more data is expected to get higher performance.

5. REFERENCES

[1] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, “wav2vec 2.0: A framework for self-supervised learning of speech representations,” *Advances in Neural Information Processing Systems*, vol. 33, 2020.

[2] https://github.com/DCASE_REPO/DESED_task/tree/master/recipes/dcase2021_task4_baseline.

[3] C. Liang, H. Ying, Y. Chen, and Z. Wang, "Mean teacher with sound source separation and data augmentation for DCASE 2020 Task 4," in *Proc. DCASE Workshop*, 2020.

[4] F., Zhiyun, M. Li, S. Zhou, and B. Xu, "Exploring wav2vec 2.0 on speaker verification and language identification," *arXiv preprint arXiv:2012.06185* 2020.

[5] M. Ott, S. Edunoy, A. Basvisk, A. Fan, S. Gross, N. Ng, D. Grangier and M. Auli, “fairseq: A Fast, Extensible Toolkit for Sequence Modeling”, Proceedings of NAACL-HLT 2019: Demonstrations, 2019

[6] N. Turpault and R. Serizel. “Training sound event detection on a heterogeneous dataset,” in *Proc. DCASE Workshop*, 2020.