HYU SUBMISSION FOR DCASE 2022 TASK 4

PA-NET: PATCH-BASED ATTENTION FOR SOUND EVENT DETECTION

Technical Report

Sojeong Kim

Hanyang University
Artificial Intelligent Dept. 222, Wangsimni-ro
Seongdong-gu, Seoul, Korea
sj2021162470@hanyang.ac.kr

ABSTRACT

In this paper, we describe details about submitted systems for DCASE 2022 challenge task 4: sound event detection in domestic environments. We focus on how to effectively use a spectrogram as input for SED model since it has different time-frequency characteristics. Frequencies have various characteristics for some reasons like recording devices and type of sound event. Specifically, each time frame has different features from each other due to uncertainty on whether any sound event may happen or not in an audio clip and what type of sound event. Therefore, we propose a patch attention(PA) mechanism capturing patch-range dependencies across input sequences so that the model can learn by training with important local information. We use PA with efficient channel attention for learning important channels in feature maps. In addition, we adopt a strategy called subspectral normalization (SSN), which split the input frequencies into multiple sub-groups and normalizes each group to stand out specific features. Experiments result on the DESED 2022 validation dataset show that our proposed model outperforms the baseline system. Particularly, our model demonstrates improvement in performance on PSDS scores of 0.4438 and 0.683 on scenario1 and scenario2 respectively.

Index Terms—patch-based attention, sound event detection, DCASE

1. METHOD

2.1. Patch-based Attention

In this section, we present patch-based attention (PA) for sound event detection. PA is the first spatial attention mechanism considering both frequency and time dimensions for sound event detection. The effectiveness of the attention mechanism in the audio domain has been sufficiently demonstrated by previous studies[10, 11]. But they take differences into account between frequencies in the spectrogram but time frames are not. In SED, only a few time frames of the whole patches have sound events, and they need to be emphasized differently than others. We consider frequency dimension and time dimension by using patch units. Because patch units contain local context information of frequency and time are suitable for sound event detection.

Patch units are typically used to make an image transformer input embeddings in vision tasks such as token embeddings in NLP. Patches act like not only token embeddings but also local information units of frequency and time axis. We focus on the latter of the patch's role.

Figure 1 shows PA architecture. PA consists of two main processes: patch compression and expansion to a patch, such as squeeze and excitation attention mechanism [8]. The spectrogram is split into patches without overlapping. (patch compression) Each patch is changed to a compressed element of the patch's information. Elements use sigmoid to compute their importance. after that, they are no longer local context but their attention scores. (expansion to a patch) Multiply the attention score by the patch pair to highlight important local contexts, such as those involving sound events. The deeper the network, the smaller the functional map size along with the patch size. we use the set of patch sizes [4, 4, 4, 4, 4, 2, 2].

In our experiments, we use PA and ECA [9] together to pay attention to the spatial and channel dimensions of the feature map.

Table 1. Results with network architectures and PA with ECA

Model	PSDS1	PSDS2
Baseline (CRNN)	0.3557	0.5664
double CRNN	0.4007	0.6152
double CRNN + PA,ECA	**0.4175**	**0.6631**

2.2. SubSpectral Normalization

Adopting SSN [12] is another way to align with our goal of utilizing local information in a spectrogram by training. SubSpectral Normalization (SSN) splits the input feature map into several groups along the frequency dimension and normalizes inter-groups. SSN overcomes the drawback of batch normalization, which can lose the unique characteristics of each frequency dimension because it performs equally in frequency and time dimensions. The experimental results are demonstrated in Table2. As shown in the table, SSN makes superior performance for PSDS1 0.3862 and PSDS2 0.5999 than other techniques in the SED task.

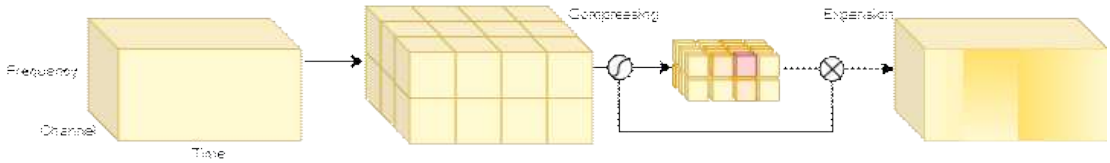

Figure 1. The architecture of the proposed PA

Table 2. Results of normalization techniques with baseline model.

Type	PSDS1	PSDS2
Batch Norm [13]	0.3557	0.5664
Group Norm [14] (G=16)	0.3809	0.5927
Instance Norm [15]	0.357	0.578
Subspectral Norm (G=4)	**0.3862**	**0.5999**

2.3. Network Architecture

We use a double-CRNN architecture that changes the baseline system model, CRNN, to double the width of the model. Simply, increase the channels of CRNN to double and use context gating as activation function instead of ReLU. Using SSN(G=4) and CNN output time axis pool 4x, and to fit the feature map to the patch size, the CNN's padding size is [2, 2, 2, 1, 1, 0, 0] and the CNN's pooling is the size is [[2,2], [2,2], [1,2], [1,2], [1,2], [1,1], [1,4]].

2. EXPERIMENTS

2.1. Dataset and Feature Extraction

All of our experiments are conducted on the DCASE 2022 challenge task4: sound event detection in domestic environments. The dataset is composed of 10-sec audio clips, 10 classes of sound events and consists of four types of datasets: strong-labeled synthetic dataset contained 10000 clips, strong-labeled real dataset contained 3470 clips, weakly-labeled real dataset (only have sound event labels) contained 1578 clips, and unlabeled real dataset contained 14412 clips. Real datasets are from the Audioset and sytnthetic dataset are generated with the Scaper soundscape synthesis and augmentation library. Audio was resampled to 16 kHz and log mel spectrograms were extracted using the same options as the baseline system, a 2048 window with 256 hop lengths, 128 mel bins. As a result, the shape of input feature for the DCASE model is 1x128x625.

2.2. Experimental Settings.

The network is trained by the mean-teacher method of semi-supervised learning to effectively learn unlabeled dataset. We employ three data augmentation methods: mix-up[6], time-shifting, and frequency masking[15] to make the training dataset more diverse which improves model performance. The first two are used commonly, the last one is only used for the ensemble. Unlike the rest, frequency masking applies only student model because it does not change the label. Train the network with the Adam optimizer (max lr= 0.001) and exponential warmup learning rate scheduler. we apply the BCE loss to the supervised loss for strong, weak labels and the MSE loss for the self-supervised loss between teacher and student model's prediction.

3. RESULTS

The performances of the submitted systems are demonstrated in Table 3. All submitted models use the same network architecture, double-CRNN, along with PA and ECA. The results for an ensemble of six models trained only on the training dataset are the best. (PSDS1 0.4438 and PSDS2 0.683) However, the model could not be submitted because an error occurred during evaluation. Model 1 is a three-model ensemble. Different models have different augmentation methods applied and whether or not strong labeled real data is used. Model 2 is a single model trained by only training set and mix-up, time-shifting applied for data augmentation. Models 3 and 4 were trained by integrating the validation dataset into the training dataset.Model 3 is a two-model ensemble trained by different parameters (data augmentation, mixup rate). The last model 4 is a five-ensemble from different two model's top-2, top-3 checkpoints.

Model	Training	Ensemble	PSDS1	PSDS2
double CRNN	Training set	3-model	0.434	0.675
double CRNN	Training set	single	0.422	0.667
double CRNN	Full set	2-model	0.494	0.748
double CRNN	Full set	5-model	0.48	0.726

4. REFERENCES

[1] Hershey, Shawn, et al. "The benefit of temporally-strong labels in audio event classification." ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2021.
[2] Fonseca, Eduardo, et al. "Fsd50k: an open dataset of human-labeled sound events." IEEE/ACM Transactions on Audio, Speech, and Language Processing 30 (2021): 829-852.
[3] Font, Frederic, Gerard Roma, and Xavier Serra. "Freesound technical demo." Proceedings of the 21st ACM international conference on Multimedia. 2013.
[4] Turpault, Nicolas, et al. "Sound event detection in domestic environments with weakly labeled data and soundscape synthesis." (2019).
[5] Serizel, Romain, et al. "Sound event detection in synthetic domestic environments." ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2020.
[6] Zhang, Hongyi, et al. "mixup: Beyond empirical risk minimization." arXiv preprint arXiv:1710.09412 (2017).
[7] Bilen, Çağdaş, et al. "A framework for the robust evaluation of sound event detection." ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2020.
[8] Hu, Jie, Li Shen, and Gang Sun. "Squeeze-and-excitation networks." Proceedings of the IEEE conference on computer vision and pattern recognition. 2018.
[9] Wang, Q., Wu, B., Zhu, P., Li, P., Zuo, W., & Hu, Q. (2020). ECA-Net: Efficient Channel Attention for Deep Convolutional Neural Networks. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)
[10] Liu, Tianchi, et al. "MFA: TDNN with multi-scale frequency-channel attention for text-independent speaker verification with short utterances." ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2022.
[11] Zheng, Xu, et al. "An Effective Mutual Mean Teaching Based Domain Adaptation Method for Sound Event Detection." Interspeech. 2021.
[12] Chang, Simyung, et al. "Subspectral normalization for neural audio data processing." ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2021.
[13] Ioffe, Sergey, and Christian Szegedy. "Batch normalization: Accelerating deep network training by reducing internal covariate shift." International conference on machine learning. PMLR, 2015.
[14] Wu, Yuxin, and Kaiming He. "Group normalization." Proceedings of the European conference on computer vision (ECCV). 2018.
[15] Ulyanov, Dmitry, Andrea Vedaldi, and Victor Lempitsky. "Instance normalization: The missing ingredient for fast stylization." arXiv preprint arXiv:1607.08022 (2016).
[16] Park, Daniel S., et al. "Specaugment: A simple data augmentation method for automatic speech recognition." arXiv preprint arXiv:1904.08779 (2019).