SKATTN team's submission for DCASE 2022 Task 4 - Sound Event Detection in Domestic Environments

Myeonghoon Ryu
Seoul, South Korea
Deeply Inc.
myeonghoon.ryu@deeply.co.kr

Jeunghyun Byun
Seoul, South Korea
Deargen Inc.
jhbyun@deargen.me

Hongseok Oh
Seoul, South Korea
Deeply Inc.
hongseok.oh@deeply.co.kr

Suji Lee
Seoul, South Korea
Deeply Inc.
suji@deeply.co.kr

Han Park
Seoul, South Korea
Deeply Inc.
han@deeply.co.kr

Abstract

In this technical report, we present our submitted system for DCASE 2022 Task4: Sound Event Detection in Domestic Environments . There are two main aspects we considered to improve the performance of the official baseline system: (1) use of external datasets (2) designing a novel model SKATTN. Our newly proposed SKATNN model combines Selective Kernel Network (SKNet) with the self-attention blocks from the Transformer model. Motivated from the SKNet's successful applications in Computer Vision and Audio domains, we adopted SKNet as a feature extractor for processing the input mel-spectrogram. We used self-attention blocks to process the spectro-temporal features since they are flexible in modeling short and long-range dependencies while being less susceptible to vanishing gradients which commonly occur in RNNs. Experiments on DCASE2022 task 4 validation dataset demonstrate that our system achieves PSDS1 + PSDS2 = 1.372 on the validation dataset, outperforming 0.872 of the baseline system.

1 Introduction

Compared to the last year's challenge, DCASE 2022 Task 4 focuses on the polyphonic sound event detection in domestic environments and the impact of incorporating external data into the development set. The sound separation track was discontinued, and the allowed external datasets include SINS (Aho and Ullman, 2017), AudioSet (Gemmeke et al., 2017), FSD50K (Fonseca et al., 2020), MUSAN (Snyder et al., 2015), and ImageNet (Russakovsky et al., 2015). Since internal and external datasets are consisted of different label types (strong label, weak label, and no label) and drawn from various distributions, the resulting dataset is highly heterogenous.

In the previous series of DCASE task 4, CRNN and the mean-teacher model (Tarvainen and Valpola, 2017) have been popular choices among some top-ranked systems for sound event detection and self-supervised learning approach, respectively. In addition, diverse data augmentation strategies were employed for the generalization ability of the systems. According to the result of (Zheng et al., 2021) in DCASE 2021 Task 4, applying selective kernel units (Li et al., 2019) improved the localization ability of the CRNN and PSDS score in both scenarios 1 and 2 significantly.

In this technical report, we propose a SED system based on our novel SKATTN model trained under mean-teacher semi-supervsed training scheme with use of external datasets: AudioSet, SINS, FSD50K. Our newly proposed SKATTN model consists of a SK network as a feature extractor backbone and self-attention block (Vaswani et al., 2017) to handle spectro-temporal features. The mean-teacher model was trained using a strongly-labeled internal dataset. Then the trained mean-teacher model predicts pseudo labels for weakly-labeled and un-labeled datasets for the self-training of the student model.

2 Proposed Method

In this section, we introduce our proposed method. Our method consists of following components (1) our novel SKATTN model (2) use of the external datasets: SINS (Aho and Ullman, 2017), AudioSet (Gemmeke et al., 2017), FSD50K (Fonseca et al., 2020) (3) the Mean-teacher training scheme.

2.1 SKATTN

2.1.1 SKNet as a feature extractor

Our SKNet consists of 7 SK layers. Each SK layer consists of SK Unit, dropout (with $p = 0.3$) and a 2D Average Pooling layer. To be specific, each SK Unit is composed of a SK Convolution layer with a residual connection and GeLU activation function. In SK Convolution layer, we have two 3x3 Convolution layers stacked in parallel (one with dilation=1 and the other with dilation=2).

In the first two SK layers, we set their 2D Average Pooling's stride along the temporal axis to 2. Thus, the first two SK layers downsamples the mel-spectrogram temporal resolution by 1/4.

The output channel dimensions of each SK layer are as follows: (16, 32, 64, 128, 128, 128, 128).

2.1.2 Self-attention layers

To handle spectro-temporal feature outputted by our SKNet feature extractor, we use stack of two self-attention layers. Each self-attention layer (Vaswani et al., 2017) consists of 4 multi-heads with hidden feature dimension of 128 (to be consistent with the SKNet's output). In addition, we add position embeddings element-wise to the first self-attention layer's output to encode the temporal information. The output of the self-attention layers are then fed into dropout layer ($p = 0.3$) followed by a fully connected feed forward layer which returns predicted class logits for each temporal frame.

2.2 External Datasets

To train our model, we use three external datasets: SINS (Aho and Ullman, 2017), AudioSet (Gemmeke et al., 2017), FSD50K (Fonseca et al., 2020) alongside the internal datasets provided by the competition hosts. One challenge in utilizing the external datasets was that the the class labels of the external datasets (external class labels) do not exactly match with the class labels of the internal datasets (internal class labels). Thus, we identified external class labels that are similar or that belong to an identical category as the internal class labels, see Figure 1. For example, 'cooking' label from the SINS dataset is mapped to 'Frying' label from the internal dataset. Hence, the external class label are re-assigned to its corresponding internal class label. In this way, we could train our model on the external datasets without resorting to having a separate classifier head for each task. Training with the external datasets improved our SKATTN model's performance on sum of PSDS (Bilen et al., 2019) scores (PSDS1 + PSDS2) from 0.83 to 1.37.

2.3 Mean-Teacher scheme

Lastly, we used baseline system's mean-teacher self-supervised training scheme (Tarvainen and Valpola, 2017) to train our model on the unlabelled dataset. We have tuned the hyperparameter of the ema factor, considering ema factor of [0.9, 0.99, 0.999]. We found that optimal ema factor to be 0.99 since 0.9 was too small for the teacher network to 'catch up' with the student network while 0.999 led to noisy pseudo label which led to unstable training.

DESED (Internal) Class Name	SINS (External)	AudioSet (External)	FSD50K (External)
Alarm_bell_ringing		Alarm • Telephone • Telephone bell ringing • Alarm clock • Smoke detector, smoke alarm • Fire alarm	Alarm • Bicycle_bell • Doorbelll Telephone • Ringtone
Blender		Blender	
Cat		Cat • Purr • Meow • Hiss • Caterwaul	Cat • Growling • Hiss • Meow • Purr
Dishes		Dishes, pots, and pans	Dishes_and_pots_and_pans
Dog		Dog • Bark • Yip • Howl • Bow-wow • Growling • Whimper (dog)	Dog • Bark • Growling
Electric_shaver_toothbrush		Toothbrush Electric shaver, electric razor Electric toothbrush	
Frying	cooking	Frying (food)	Frying_(food)
Running_water		Bathtub (filling or washing) Water tap, faucet Sink (filling or washing) Fill (with liquid)	Bathtub_(filling_or_washing) Water_tap_and_faucet Sink_(filling_or_washing) Fill_(with_liquid)
Speech	social_activity	Speech • Male speech, man speaking • Female speech, woman speaking • Child speech, kid speaking • Conversation • Narration, monologue	Speech • Child_speech_and_kid_speaking • Conversation • Female_speech_and_woman_speaking • Male_speech_and_man_speaking
Vacuum_cleaner	vacuum_cleaner	Vacuum cleaner	

Figure 1: Mapping of the class labels of the external datasets to the class labels of the internal datasets.

2.4 References

References

Alfred V. Aho and Jeffrey D. Ullman. 2017. The sins database for detection of daily activities in a home environment using an acoustic sensor network. In *Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE)*.

Cagdas Bilen, Giacomo Ferroni Francesco Tuveri, Juan Azcarreta, and Sacha Krstulovic. 2019. A framework for the robust evaluation of sound event detection. In *ArXiv*.

E. Fonseca, X. Favory, J. Pons, F. Font, and X. Serra. 2020. Fsd50k: an open dataset of human-labeled sound events. In *arXiv*.

J. F. Gemmeke, D. P. W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter. 2017. Audioset: An ontology and human-labeled dartaset for audio events. In *IEEE ICASSP*.

X. Li, W. Wang, X. Hu, and J. Yang. 2019. Selective kernel networks. In *IEEE conference on computer vision and pattern recognition*.

O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, and et al. 2015. Imagenet large scale visual recognition challenge. In *IJCV*.

D. Snyder, G Chen, and D. Povey. 2015. Musan: A music, speech, and noise corpus. In *arXiv*.

Antti Tarvainen and Harri Valpola. 2017. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *In Advances in neural information processing systems*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, , and Illia Polosukhin. 2017. Attention is all you need. In *In Advances in neural information processing systems*.

X. Zheng, H. Chen, and Y. Song. 2021. Zheng ustc team's submission for dcase2021 task4 – semi-supervised sound event detection. In *Tech. Rep., DCASE Challenge*.