Semi-Supervised Sound Event Detection Using Multiscale Channel Attention and Multiple Consistency Training

Yih-Wen Wang, Chia-Ping Chen, Chung-Li Lu, Bo-Cheng Chan

{National Sun Yat–sen University, Kaohsiung |Chunghwa Telecom Laboratories, Taoyuan}, Taiwan

Abstract

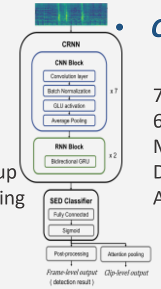
We present a neural network-based sound event detection system that outputs sound events and their time boundaries in audio signals. The network can be trained efficiently with an amount of strongly labeled synthetic data and weakly labeled or unlabeled real data. Based on the mean-teacher framework of semi-supervised learning with RNNs and Transformer, the proposed system employs multi-scale CNNs with efficient channel attention, which can capture the various features and pay more attention to the important area of features. The model parameters are learned with multiple consistency criteria, including interpolation consistency, shift consistency, and clip-level consistency, to improve the generalization and representation power. For different evaluation scenarios, we explore different pooling functions and search for the best layer. To further improve the performance, we use data augmentation and posterior-level score fusion. We demonstrate the performance of our proposed method through experimental evaluation using the DCASE2021 Task4 dataset. On the validation set, our ensemble system achieves the PSDS-scenario1 of 40.72% and PSDS-scenario2 of 80.80%, significantly outperforming that of the baseline score of 34.2% and 52.7%, respectively. On the DCASE2021 challenge's evaluation set, our ensemble system is ranking 7 among the 28 teams and ranking 14 among the 80 submissions.

INTRODUCTION

SED is a useful technique for helping us what is happening in an environment by identifying sounds, which predicts the sound event types with timestamps in audio recording. We employ the RNNs-based and Transformer-based neural networks for SED system. Then, we apply the multi-scale CNNs with ECA-Net to capture the various and important features of sound events. We extend the consistency criteria for model training in mean-teacher framework to include interpolation consistency (ICT), shift consistency (SCT), and clip-level consistency (CCT). We apply data augmentation and posterior-level score fusion to further improve the performance. Finally, on the validation set and public evaluation set of DCASE 2021 Task4, our proposed system both outperforms the baseline system.

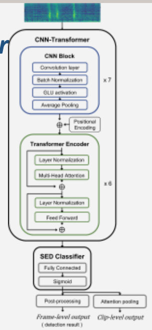
PREVIOUS WORK

- ***CRNN***



7 layers CNN
2 layers GRU
Mean-Teacher
Data Aug: Mixup
Attention pooling

- ***CNN-Transformer***



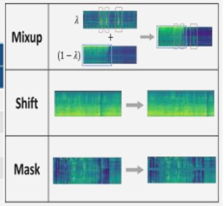
7 layers CNN
6 layers TransEncoder
Mean-Teacher
Data Aug: Mixup, Shift
Attention pooling

DATASET

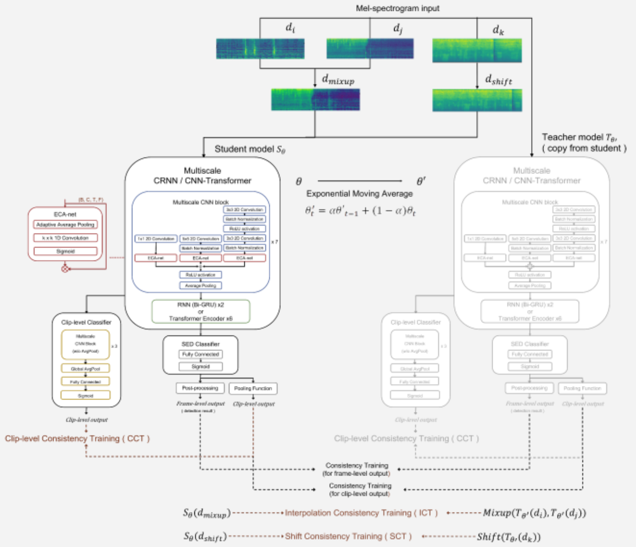
- ***DESED dataset***
  - Strong : class, onset, offset
  - Weak : class
  - Unlabel : None

DESED	Training			Validation	Public eval
	Weak	Unlabel	Strong	Strong	Strong
# Audio	1,578	14,412	10,000	1,168	692
Domain	Real	Synthetic	Real	Real	Real
Length	~10s	~10s	~10s	~10s	~10s
Sample rate	44.1kHz	16kHz	44.1kHz	44.1kHz	44.1kHz
Channel	stereo	mono	stereo	stereo	stereo

- ***Data Augmentation***



PROPOSED METHODS



A. Multiple Consistency Training
- Interpolation Consistency Training (ICT)
- Shift Consistency Training (SCT)
- Clip-level Consistency Training (CCT)

B. Multiscale CNN Blocks
- Using kernel size of 1x1, 3x3, and 5x5
- Integrating features of different scales

C. Efficient Channel Attention
- Using 1D CNN to compute channel attention
- Paying attention to important areas of features

D. Different Pooling Function

Attention	$y = \frac{\sum_i y_i w_i}{\sum_i w_i}$	Linear Softmax	$y = \frac{\sum_i y_i^2}{\sum_i y_i}$
Max pooling	$y = \max_i y_i$	Exponential Softmax	$y = \frac{\sum_i y_i \exp(y_i)}{\sum_i \exp(y_i)}$
Average pooling	$y = \frac{1}{n}\sum_i y_i$		

E. Score Fusion
- Using different data augmentation to build single systems
- Averages the raw posterior outputs of the multiple model

EVALUATION RESULTS

- ***Results of A. B. C.***

Scheme	Model	PSDS 1	PSDS 1
-	CRNN	34.04%	53.30%
	CNN-Transformer	33.46%	48.77%
+ICT	CRNN	36.38%	55.87%
	CNN-Transformer	33.39%	50.07%
+SCT	CRNN	**37.86%**	59.47%
	CNN-Transformer	35.61%	52.01%
+CCT	CRNN	37.64%	60.87%
	CNN-Transformer	**37.33%**	55.87%
+Multiscale	CRNN	37.51%	62.63%
	CNN-Transformer	34.75%	**61.10%**
+ECA-Net	CRNN	34.71%	**65.54%**
	CNN-Transformer	35.13%	60.27%

- ***Results of D.***

Pooling function	Model	PSDS 1	PSDS 1
Attention	CRNN	**37.51%**	62.63%
	CNN-Transformer	**34.75%**	61.10%
Max	CRNN	36.10%	64.59%
	CNN-Transformer	31.73%	59.77%
Average	CRNN	5.34%	73.95%
	CNN-Transformer	4.53%	60.41%
Linear Softmax	CRNN	26.75%	60.17%
	CNN-Transformer	24.21%	60.57%
Exponential Softmax	CRNN	5.82%	**75.35%**
	CNN-Transformer	4.13%	**61.31%**

- ***Results of E.***

# system	Model	Schemes	Validation		Public eval	
			PSDS 1	PSDS 2	PSDS 1	PSDS 2
10	CRNN	ICT, SCT, CCT, Multiscale	**40.72%**	70.25%	37.22%	69.47%
8	CRNN	ICT, SCT, CCT, Multiscale, ECA, Exponential Softmax	6.08%	**80.80%**	8.30%	65.39%
16	CRNN CNN-Transformer	ICT, SCT, CCT, Multiscale	38.79%	67.18%	**37.45%**	68.42%
24	CRNN CNN-Transformer	ICT, SCT, CCT, Multiscale, ECA, Exponential Softmax	37.02%	72.42%	33.56%	**69.73%**

- ***Results of DCASE 2021 Challenge Task 4***
  - Our ranking is **7 among 28 teams**, **14 among 80 submissions**
  - **2.4%** higher than Baseline(SED) on PSDS scenario1
  - **11.5%** higher than Baseline(SED) on PSDS scenario2
  - **8.2%** higher than Baseline(SS+SED) on PSDS scenario2

CONCLUSION

- **ICT** helps models discriminate the ambiguous samples to enhance the generalization ability.
- **SCT** assists models to learn better temporal information.
- **CCT** promotes the model feature representation power.
- **Multiscale CNN blocks** capture various features of sound events.
- **ECA-Net** pays more attention to important area of features.
- Appropriate **pooling function** is applied to the specific scenario.
- **Data augmentation** enhances the data diversity.
- **Posterior-level score fusion** further improves the performance.