MULTI-SCALE NETWORK BASED ON SPLIT ATTENTION FOR SEMI-SUPERVISED SOUND EVENT DETECTION

DCASE 2021 Workshop

Xiujuan Zhu , Ying Hu , Xinghao Sun, Liang He

School of Information Science and Engineering, Xinjiang University, Urumqi, Chinana

Background

Sound event detection (SED) : determine both the category and time boundaries of a event

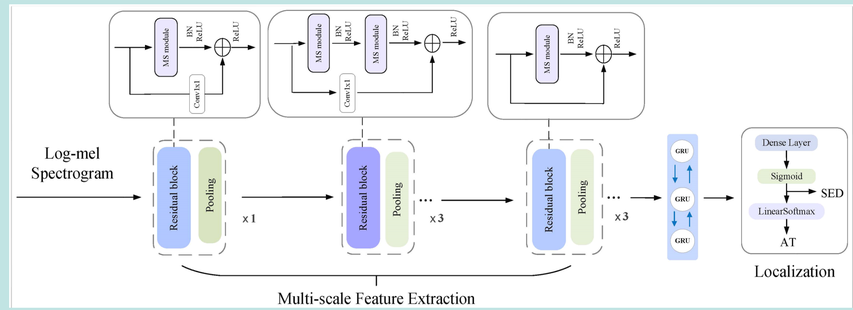
Audio tagging (AT) : only needs to predict event category within an audio

Proposed method

1.Multi-scale mechanism

2.Channel shuffle operation

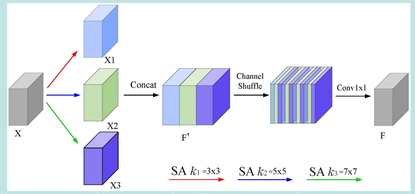
3.Split Attention (SA) module



Motivation

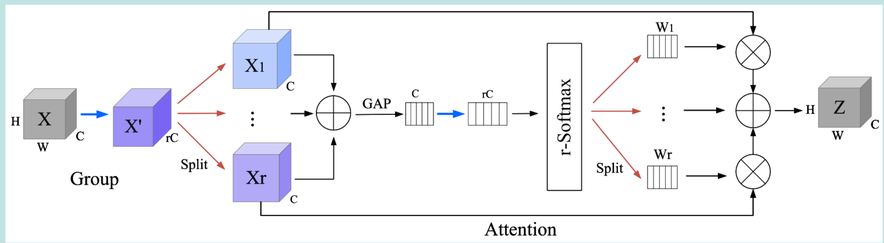
- Short duration events & Long duration events exhibit different time-frequency properties.
  - how to obtain multi-scale features is an issue.
- Features of different channels are often treated equally.
  - Group convolution is used to separately learn sub-features.

Multi-scale module



- Obtain multi-scale features by using different convolution kernel.
- Enhance the cross-channel information flowing among the features with different scales

Split Attention (SA) module



- Group convolution is used to learning sub-features
- Split r branches sub-features
- Attention mechanism

Experiment results

Table 1: Ablation experiments on multi-scale (MS) mechanism with different kernel sizes. We adopt vanilla convolution instead of SA module in MS module of all residual block in this experiment.

Network	PSDS1	PSDS2	IB-F1(%)	CB-F1(%)	Parameter
Base-2021	0.342	0.527	76.60	40.10	1.1M
MS-K=[3]	**0.358**	0.599	81.88	**44.48**	1.2M
MS-K=[3,5]	0.349	**0.602**	83.24	44.13	3.0M
MS-K=[3,5,7]	0.336	0.601	**83.50**	42.21	5.7M

Table 2: Ablation experiments on channel shuffle (CS) operation based on MS-K=[3,5] system. CS-g denotes the channel shuffle operation with g groups. g controls the fusion degree of features.

Network	PSDS1	PSDS2	IB-F1 (%)	CB-F1 (%)
MS-K=[3, 5]	0.349	0.602	**83.20**	44.13
+ CS-g=2	0.349	0.594	82.83	43.58
+ CS-g=4	**0.358**	**0.606**	82.98	**45.36**

Table 3: Ablation experiments on split attention (SA) module based on MS-K=[3,5] system. SA(g, r) means the number of group is g, splitted sub-feaatures r in shuffle attention module.

Network	PSDS1	PSDS2	IB-F1(%)	CB-F1(%)	Parameter
MS-K=[3, 5]	0.349	0.602	83.24	44.13	3.0M
+ SA(1, 1)	0.354	0.598	**84.59**	47.99	3.2M
+ SA(1, 2)	0.350	0.602	84.40	46.64	5.5M
+ SA(2, 1)	0.367	**0.606**	83.80	48.59	1.9M
+ SA(2, 2)	**0.376**	0.599	83.63	49.02	3.3M
+ CS-g=4	0.373	0.602	83.98	**50.28**	3.3M

Conclusion

- A multi-scale SED network based on split attention is proposed.
- Multi-scale module is designed to learn features in parallel.
- Split attention module is used to learndifferent sub-features separately.
- Future work: how to deal with the features with different scales in SED.