Attentive Max Feature Map and Joint Training for Acoustic Scene Classification

Hye-jin Shim[1], Jee-weon Jung[2], Ju-ho Kim[1], and Ha-Jin Yu[1]
[1]School of Computer Science, University of Seoul, [2]Naver Corporation

Poster 1160

Overview

❑ Motivation

- ***Attentive Max Feature Map (AMFM)***
  - Analyze the phenomena of performance degradation
  - Aim to alleviate the excessive information loss
- ***Joint Training***
  - Frequently misclassified pairs mostly happens between the same abstract class categories
- Most DCASE related works adopt various ensembles using a number of different models with a high-complexity architecture

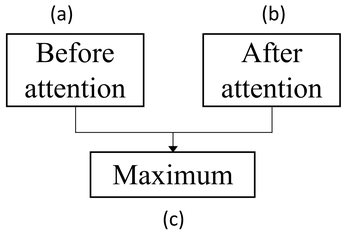
❑Contributions

- ***Attentive Max Feature Map***
  - A new module that combines the attention mechanism and the max feature map
- ***Joint training***
  - Adopt two relevant tasks
  - Improves the performance of both 10-class and 3-class classifications
- Proposal of single system to keep the line with recent challenge of DCASE 2022, which only deals with low-complexity systems

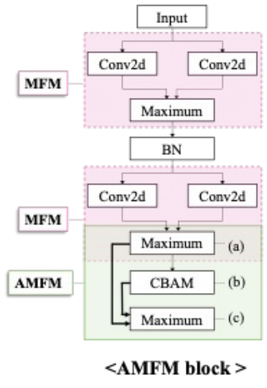
Proposed three integrated architectures & Joint training

1. Attentive Max Feature Map (AMFM)

- Assumption: attention mechanism discards excessive information
- Inspired by max feature map of Light CNN
- Comparison of the feature maps of before and after attention
- ***Seamlessly applied to any other architecture and domains***



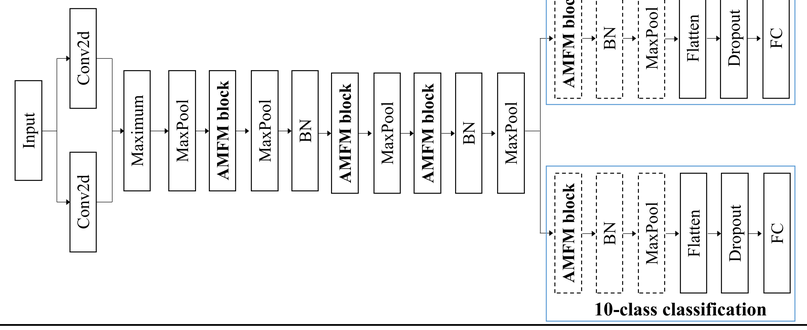
$\rightarrow \max(attention(a), a),$
$where\ a \in \mathbb{R}^{K \times T \times F}$



<AMFM block >

2. Joint Training

- ***Utilizes the Subtask A labels and Subtask B labels***
- Adopts various joint training approaches
  - pre-training
  - multi-task learning(MTL)
  - Sequential MTL
  - ***extended MTL(best performance)***
    - refer to the below figure



Task definition

❑ DCASE 2020 Subtask A

- 10-class classification (scene - specific)

❑ DCASE 2020 Subtask B

- 3-class classification (overall categories - abstract)

❑ Class labels for each tasks → utilized for joint training method

(Subtask A / Subtask B)

- Airport / Indoor
- Indoor shopping mall / Indoor
- Metro station / Indoor
- Pedestrian street / Outdoor
- Public square / Outdoor
- Street with medium level of traffic / Outdoor
- Travelling by a tram / Transportation
- Travelling by a bus / Transportation
- Travelling by an underground metro / Transportation
- Urban park / Outdoor

Experiment results

- **Dataset** : DCASE 2020 Task 1 Subtask A

- **Features**
  - 256 dimensional Mel-Spectrograms
  - Hamming window with a length of 50 ms & 20 ms shift
  - 2048 FFT bins for all kinds

- **Training details**
  - Data augmentation: Mixup, SpecAugment
  - Optimizer: SGD
  - Batch size: 24

- **Experimental results of applying the attention mechanism**

System	Attention	Accuracy (%)
CNN w/ ReLU	✗	70.2
	✓	68.3
CNN w/ LeakyReLU	✗	69.6
	✓	68.2
MFM	✗	69.4
	✓	70.3 ± 0.13
AMFM	✓	**70.7 ± 0.08**

- **Comparison with the state-of-the-art systems**

System	Accuracy (%)	# Params
Proposed Method	**71.3**	**0.6M**
DCASE2020 Baseline	54.1	5M
Suh et al.	73.7	13M
Hu et al.	76.9	-
Gao et al.	71.8	4M
Liu et al.	72.1	3M
Koutini et al.	71.8	225M

- **Application of various joint training strategies**

System	Joint prediction	# Params	Accuracy (%)
w/o joint training	✗	1.5M	70.8
Pre-training	✗	1.5M	69.2
Conventional MTL	✗	1.5M	69.7
Extended MTL	✗	0.6M	**71.3**
	✓	0.6M	70.0
Sequential MTL	✗	0.7M	71.0
	✓	0.7M	69.1
Separated Classifier	✓	1.5M	69.4