TASK-AWARE MEAN TEACHER METHOD FOR LARGE SCALE WEAKLY LABELED SEMI-SUPERVISED SOUND EVENT DETECTION

Jie Yan[1] Yan Song[1] Li-Rong Dai[1] Ian McLoughlin[2]
[1] National Engineering Laboratory for Speech and Language Information Processing, University of Science and Technology of China, Hefei, China.
[2] School of Computing, University of Kent, Medway, UK.

中国科学技术大学
University of Science and Technology of China

Contents

Background

- Sound event detection (SED) : determine both the category and occurrence time of a sound event
- Audio tagging (AT) : only needs to predict the category
- Mean teacher (A semi-supervised learning method):
  - It is composed of two networks that both have the same structure.
  - One network is a student model which is trained by back propagation.
  - The other is a teacher model which is updated, much more slowly, by the exponential moving average of the student parameters.

Motivation

- The performance of NN based SED methods depends heavily on the size and quality of the training dataset.
  - Datasets with strong labels are expensive and time-consuming to collect.
  - By contrast, unlabeled or weakly labeled SED recordings are far more easily available.
- SED needs fine-level information, whereas AT tends to provide coarse-level information.
  - Systems for SED are often designed to perform both SED and AT simultaneously.
  - This scale mismatch indicates that systems jointly optimised to perform both tasks may be disadvantaged.

Our Approach

- Mean teacher learning method with data augmentation is used to exploit unlabeled data in an effective way to learn additional structure from the input distribution.
- Multi-branch CRNN structure is proposed to solve the SED and AT tasks differently
    - Specifically, a branch with coarse temporal resolution is designed for the AT task, while a branch with a finer level of temporal resolution is designed for the SED task.

Our Approach

Data Augmentation

- Data augmentation is often used to generate the perturbation of training data to improve the generalization capability of the model.
  - Spec-augment is first applied to the feature inputs.
    - In our implementation, only frequency masking is applied, which means that entire mel frequency bands are consecutively masked.

Our Approach

Data Augmentation

- A method of mixing up labeled and unlabeled data is proposed for the system.
  - Given data x_i , x_j, the mixture method is implemented as below;

$$\hat{x}_{mix} = \lambda * x_i + (1-\lambda)x_j$$

$$\hat{y}_{mix} = \lambda * \hat{y}_i + (1-\lambda)\hat{y}_j$$

$$\hat{y} = \begin{cases} y & (x, y) \in D_L \\ f_{\theta'}(x) & (x) \in D_{UL} \end{cases}$$

where y is the label of data x and $f_{\theta'}$ is the teacher model. D_L and D_{UL} are the labeled and unlabeled dataset respectively.

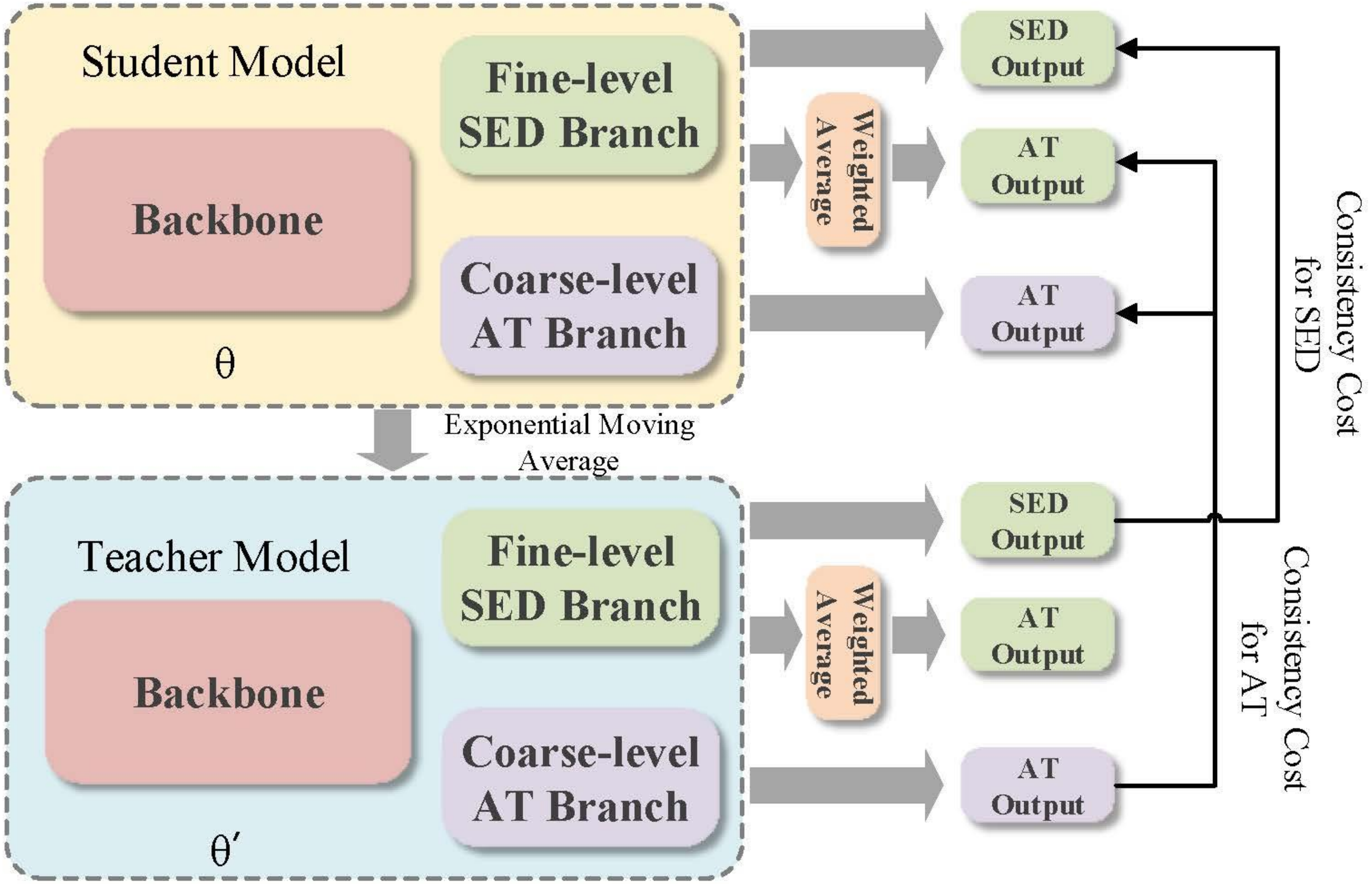
Task-aware Teacher-student Learning

- We designed the proposed system to incorporate two branches with the same backbone but dedicated to fine-level and coarse-level information respectively.

- Prediction of the coarse-level AT branch in the teacher model is used to teach the AT ability of the student model, while prediction of the fine-level SED branch in the teacher model is used to teach the SED ability of student model.
  - Given data x , the consistency loss is organized as below;

$$L_{consistency} = \sum_{n=1}^{N} L_{AT}(S_{\theta_{F\_AT}}(x_n), T_{\theta'_{C\_AT}}(x_n)) + L_{SED}(S_{\theta_{F\_SED}}(x_n), T_{\theta'_{F\_SED}}(x_n)) + L_{AT_{aux}}(S_{\theta_{C\_AT}}(x_n), T_{\theta'_{C\_AT}}(x_n))$$

  - where S_θ and $T_{\theta'}$ are student model and teacher model. F_SED , F_AT and C_AT are the fine-level SED output, fine-level AT output and coarse-level AT output respectively



Task-aware teacher-student learning method

Proposed System

- The architecture used for our experiments is a CRNN structure.
- Context Gating
  - The context gating (CG) module in the CNN block is applied for learning of gated units.
  - Given the input feature X , an output Y the CG module can be represented as

$$Y = \sigma(W * X + b) \cdot X$$

where $*$ denotes the convolutional operator, W and b are filter kernel and bias. σ is the sigmoid function and $\cdot$ is the element-wise product.

Proposed System

- Multi-branch Structure
  - The network has a shared backbone, followed by two branches with fine- and coarse-level information respectively.
  - In each branch, the pooling module following the convolution operator is applied to control the receptive field of the feature representation.
- Multi-resolution Feature
  - Features with a variety of receptive field sizes can be suitable for SED.
  - In our system, we aggregate the last few layer outputs of the CNN part to obtain multi-resolution features.

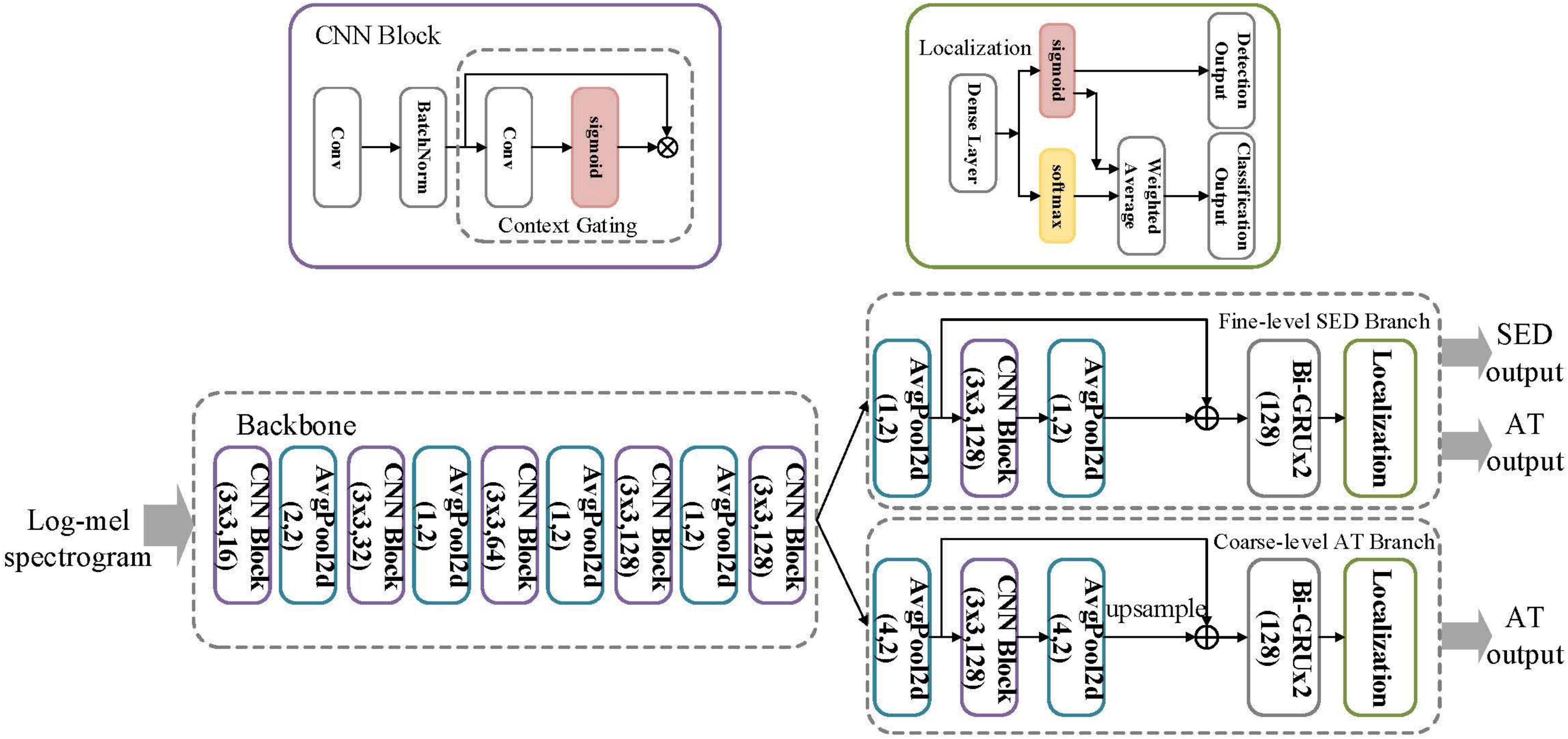


Illustration of the proposed model architecture

Experiments and Results

- Dataset
- The dataset is from Task 4 of the DCASE 2018 Challenge
  - 1,578 weakly labeled training clips
  - 14,412 unlabeled in-domain training clips
  - 39,999 unlabeled out-of-domain training clips
  - 288 development clips
  - 880 evaluation clips
- The dataset has 10 classes of sound events selected from domestic environments.

Sound event class durations occurring in the development dataset.

Event label	Count	Length (s)	
		Total	Average
Alarm_bell_ringing	112	171.87	1.53
Blender	40	214.19	5.35
Cat	97	78.90	0.81
Dishes	122	68.27	0.56
Dog	127	130.33	1.03
Electric_shaver_toothbrush	28	207.63	7.42
Frying	24	224.07	9.34
Running_water	76	426.44	5.61
Speech	261	395.41	1.51
Vacuum_cleaner	36	311.60	8.66

Model	F1
CRNN-ML	71.8
CRNN-MULS	72.4
CRNN-MULT	72.6
CRNN-MULT-MF-F	73.6
CRNN-MULT-MF-C	**73.9**

Audio tagging (AT) results for the proposed methods.

Model	Event-based F1
CRNN-MULT	30.6
CRNN-MULT-MF-F	33.6
CRNN-MULT-MF-C	32.7
CRNN-MULT-MF-CF	35.5
CRNN-MULT-MF-final	**37.7**

Sound event detection (SED) results for the proposed methods.

- “-ML”, “-MULS” and “-MULT” mean mixing up labeled data only, mixing up unlabeled and labeled data separately, and mixing up unlabeled and labeled data together.
- “-MF” refers to a system using the concatenation operation to obtain multi-resolution features.
- “-F” and “-C” refer to systems with only a fine-level branch or a coarse-level branch respectively.

Results

Analysis

- Data augmentation for unlabeled data improves the performance of AT.
- The usage of multiresolution features is found to be beneficial for both SED and AT.
- Compared to the system with fine-level information, the system with coarse-level information has better AT performance and worse SED performance.
- It is evident that the systems with two branches for SED and AT respectively outperform systems having just one.

Conclusion

- This paper proposed a method to mitigate the problem of making predictions for SED and AT through the same network structure when using unlabeled data.
- A multi-branch system was designed to enable detection using fine-level information, and classification using coarse-level information.
- Data augmentation was applied for unlabeled data and multi-resolution features in order to improve system performance.

Thank you!

中国科学技术大学
University of Science and Technology of China