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

Jie Yan¹ Yan Song¹ Li-Rong Dai¹ Ian McLoughlin²
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.
Contents

- Motivation
- Our Approach
- Proposed System
- Experiments and Results
- Conclusion
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.
Our Approach

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:
    
    $$L_{\text{consistency}} = \sum_{n=1}^{N} L_{\text{AT}}(S_{\theta_{F\_AT}}(x_n), T'_{\theta'_{C\_AT}}(x_n)) + L_{\text{SED}}(S_{\theta_{F\_SED}}(x_n), T'_{\theta'_{F\_SED}}(x_n)) + L_{\text{AT\_aux}}(S_{\theta_{C\_AT}}(x_n), T'_{\theta'_{C\_AT}}(x_n))$$

  - where $S_{\theta}$ and $T'_{\theta'}$ are student model and teacher model. $F\_\text{SED}$, $F\_\text{AT}$ and $C\_\text{AT}$ 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. $\sigma$ 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.

<table>
<thead>
<tr>
<th>Event label</th>
<th>Count</th>
<th>Length (s)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Total</td>
</tr>
<tr>
<td>Alarm.bell.ringing</td>
<td>112</td>
<td>171.87</td>
</tr>
<tr>
<td>Blender</td>
<td>40</td>
<td>214.19</td>
</tr>
<tr>
<td>Cat</td>
<td>97</td>
<td>78.90</td>
</tr>
<tr>
<td>Dishes</td>
<td>122</td>
<td>68.27</td>
</tr>
<tr>
<td>Dog</td>
<td>127</td>
<td>130.33</td>
</tr>
<tr>
<td>Electric.shaver_toothbrush</td>
<td>28</td>
<td>207.63</td>
</tr>
<tr>
<td>Frying</td>
<td>24</td>
<td>224.07</td>
</tr>
<tr>
<td>Running_water</td>
<td>76</td>
<td>426.44</td>
</tr>
<tr>
<td>Speech</td>
<td>261</td>
<td>395.41</td>
</tr>
<tr>
<td>Vacuum_cleaner</td>
<td>36</td>
<td>311.60</td>
</tr>
</tbody>
</table>

Sound event class durations occurring in the development dataset.
Audio tagging (AT) results for the proposed methods.

<table>
<thead>
<tr>
<th>Model</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>CRNN-ML</td>
<td>71.8</td>
</tr>
<tr>
<td>CRNN-MULS</td>
<td>72.4</td>
</tr>
<tr>
<td>CRNN-MULT</td>
<td>72.6</td>
</tr>
<tr>
<td>CRNN-MULT-MF-F</td>
<td>73.6</td>
</tr>
<tr>
<td>CRNN-MULT-MF-C</td>
<td>\textbf{73.9}</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Model</th>
<th>Event-based F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>CRNN-MULT</td>
<td>30.6</td>
</tr>
<tr>
<td>CRNN-MULT-MF-F</td>
<td>33.6</td>
</tr>
<tr>
<td>CRNN-MULT-MF-C</td>
<td>32.7</td>
</tr>
<tr>
<td>CRNN-MULT-MF-CF</td>
<td>35.5</td>
</tr>
<tr>
<td>CRNN-MULT-MF-final</td>
<td>\textbf{37.7}</td>
</tr>
</tbody>
</table>

- "-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.
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!