Peer Collaborative Learning for Polyphonic Sound Event Detection

Hayato Endo and Hiromitsu Nishizaki
endo@alps-lab.org, hnishi@yamanashi.ac.jp

Integrated Graduate School of Medicine, Engineering, and Agricultural Sciences, University of Yamanashi, JAPAN
Polyphonic Sound Event Detection Task

- **Task Definition**: detection of multiple sound event intervals in acoustic data for domestic environments

**Goal**

Improvement of detection accuracy of sound event intervals in practical environment situations

---


# Label Information on the Task

Three sorts of label types are included in the dataset

<table>
<thead>
<tr>
<th>Label image</th>
<th>Hard label</th>
<th>Soft label</th>
<th>Unlabeled</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1.png" alt="Label image" /></td>
<td><img src="image2.png" alt="Label image" /></td>
<td><img src="image3.png" alt="Label image" /></td>
<td><img src="image4.png" alt="Label image" /></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Label</th>
<th>class</th>
<th>interval</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image5.png" alt="Label image" /></td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td><img src="image6.png" alt="Label image" /></td>
<td>○</td>
<td>×</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Amount of data</th>
<th>small</th>
<th>small</th>
<th>large</th>
</tr>
</thead>
<tbody>
<tr>
<td>Difficulty of collection</td>
<td>high</td>
<td>middle</td>
<td>low</td>
</tr>
</tbody>
</table>

Because collecting hard-labeled data is very costly, soft-labeled or unlabeled data should be utilized.

This study proposes a model structure that can utilize soft-labeled and unlabeled data.
Related Work (1/2)

Online Knowledge Distillation \textsuperscript{[3]}

- Considering the output of the ensemble net as a reference, each subnetwork extracts powerful features for classification.

\textbf{Improved performance of each subnetwork} $\Rightarrow$ \textbf{Improved overall performance}

Mean-Teacher model (the baseline model of the DCASE 2019 · 2020 Task4)

- Student model (For training and evaluation): Use the recent weights for classification
- Teacher model (For training only): Use the past to recent weights of the student model

Guiding the student model training, effective use of unlabeled data
Summary of Our Research

■ **Goal** of our research
  - Improvement of accuracy of sound event detection on the DCASE Task 4

■ **Proposed approach**
  - Use Peer Collaborative Learning (PCL)\[[4]\], an integration and development of online knowledge distillation and mean-teacher approaches
  - Propose an effective combination of PCL and acoustic data augmentation

※ F1-score was used as evaluation measure

RESULT: Baseline (31.1%※) ➤ Proposed (44.2%※)

【Proposed】PCL with Data Augmentation

**Pre-processing**

- **Input data:** $X$
  - Hard-labeled data
  - Soft-labeled data
  - Unlabeled

**Data augmentation (DA)**

1. $X$ (w/o DA)
2. $X + $ mixup
3. $X + $ Gaussian noise
4. $X + $ frequency mask
5. $X + (3+4)$

**Peer Collaborative Learning (PCL)**

- **Student model**
  - lower layer
  - upper layers
  - Subnetwork 1
  - Subnetwork 5
  - Ensemble net
  - Output

- **Teacher model**
  - lower layer
  - upper layers
  - Subnetwork 5
  - Output

**Update weight parameters using the exponential moving average**

- $L_{BCE}$: Classification loss
- $L_{MSE}$: Consistency loss
- BCE: Binary cross entropy
- MSE: Mean squared error

**Hard label**
- Soft label

**Soft label**
- Hard label
- Unlabeled

**Data augmentation (DA)**

- $X_1$
- $X_2$
- $X_3$
- $X_4$
- $X_5$

- Hard label
- Soft label
- Unlabeled

- $L_{BCE}$
- $L_{MSE}$

**Input data:** $X$
Data Pre-Processing

Pre-processing

Input data: $X$

- Hard-labeled data
- Soft-labeled data
- Unlabeled

Data augmentation (DA)

1. $X$ (w/o DA)
2. $X +$ mixup
3. $X +$ Gaussian noise
4. $X +$ frequency mask
5. $X + (3+4)$

Student model
- lower layer
- upper layers
- Subnetwork 1
- Subnetwork 5
- Ensemble net

Teacher model
- lower layer
- upper layers
- Subnetwork 1
- Subnetwork 5

Peer Collaborative Learning (PCL)

- Hard label
- Soft label
- Hard label
- Soft label
- Unlabeled

Update weight parameters using the exponential moving average

$X_1, X_2, X_3, X_4, X_5$

Loss functions:
- $L_{BCE}$: Classification loss
- $L_{MSE}$: Consistency loss
- BCE: Binary cross entropy
- MSE: Mean squared error
Peer Collaborative Learning

Pre-processing

Input data: $X$

- Hard-labeled data
- Soft-labeled data
- Unlabeled

Data augmentation (DA)

1. $X$ (w/o DA)
2. $X$ + mixup
3. $X$ + Gaussian noise
4. $X$ + frequency mask
5. $X$ + (3+4)

Peer Collaborative Learning (PCL)

Student model

lower layer

shared layers

Subnetwork 1

... 

Subnetwork 5

upper layers

output

Ensemble net

output

Update weight parameters using the exponential moving average

Hard label
Soft label

Hard label
Soft label
Unlabeled

Teacher model

shared layers

upper layers

Subnetwork 5

output

Subnetwork 1

output

Hard label
Soft label

$L_{BCE}$: Classification loss
$L_{MSE}$: Consistency loss
BCE: Binary cross entropy
MSE: Mean squared error

Input data: $X$

- $X_1$
- $X_2$
- $X_3$
- $X_4$
- $X_5$
PCL Model Details

1. \(X\)
2. \(X + \text{Mix up}\)
3. \(X + \text{Gaussian noise}\)
4. \(X + \text{Frequency mask}\)
5. \(X + 3 \& 4\)

**Augment for subnets**

- Basic augment
  - No change
  - Filter
  - Time shift
  - Time mask
- \(X_1\)
- \(X_2\)
- \(X_3\)
- \(X_4\)
- \(X_5\)

**PCL Model Formulation**

\[L = L_{BCE}^{n,w} + L_{BCE}^{n,stu} + L_{BCE}^{n,te} + L_{MSE}^{n,stu} + L_{MSE}^{n,te} + L_{MSE}^{en,stu} + L_{MSE}^{en,te}\]

- \(n\): Subnet Number
- \(m\): Subnet number, excluding \(n\)
- \(L_{BCE}^{n,w}\): BCE for each subnet's output
- \(L_{BCE}^{n,stu}\): BCE for ensemble output
- \(L_{MSE}^{n,stu}\): MSE of each subnet output of the student model and each subnet output of the teacher model except \(n\)
- \(L_{MSE}^{en,stu}\): MSE with each subnet output and ensemble output

**Input:** \(X\)

**Student Model**

- Low-level layer (Share layers)
- High-level layers

**Teacher Model**

- Shared layers
- High-level layers

**Augment for subnets**

- No change
- Filter
- Time shift
- Time mask

**Augment for subnets**

- No change
- Filter
- Time shift
- Time mask

**Basic augment**

- No change
- Filter
- Time shift
- Time mask

**PCL Model Formulation**

\[L = L_{BCE}^{n,w} + L_{BCE}^{n,stu} + L_{BCE}^{n,te} + L_{MSE}^{n,stu} + L_{MSE}^{n,te} + L_{MSE}^{en,stu} + L_{MSE}^{en,te}\]

- \(n\): Subnet Number
- \(m\): Subnet number, excluding \(n\)
- \(L_{BCE}^{n,w}\): BCE for each subnet's output
- \(L_{BCE}^{n,stu}\): BCE for ensemble output
- \(L_{MSE}^{n,stu}\): MSE of each subnet output of the student model and each subnet output of the teacher model except \(n\)
- \(L_{MSE}^{en,stu}\): MSE with each subnet output and ensemble output

**Wap**

- Weighted average pooling

**BCE**

- Binary Cross Entropy

**MSE**

- Mean Squared Error
Experimental Setup

- **Dataset**
  - DCASE 2019 Task4 [1]
  - Sounds expected to occur in home environment (1 file = 10 seconds duration)

<table>
<thead>
<tr>
<th></th>
<th>Label type</th>
<th># of data [ /file ]</th>
<th>Remarks</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training</td>
<td>Hard label</td>
<td>2,045</td>
<td>Known event intervals</td>
</tr>
<tr>
<td></td>
<td>Soft label</td>
<td>1,578</td>
<td>Unknown event intervals</td>
</tr>
<tr>
<td></td>
<td>Unlabeled</td>
<td>14,412</td>
<td></td>
</tr>
<tr>
<td>Validation</td>
<td>Hard label</td>
<td>1,168/692</td>
<td>Known event intervals</td>
</tr>
<tr>
<td>Evaluation</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Num. of event classes: 10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alarm/bell/ringing</td>
</tr>
<tr>
<td>Blender</td>
</tr>
<tr>
<td>Cat</td>
</tr>
<tr>
<td>Dishes</td>
</tr>
<tr>
<td>Dog</td>
</tr>
</tbody>
</table>

- **Evaluation measure**
  - F1-score [%] based on the interval of sound event occurrence
    - The student model is used for evaluation

[Diagram of sound event classification with Dog label and time intervals 2.20~4.10 [s]]
Experimental Setup

- **Dataset**
  - DCASE 2019 Task4 [1]
  - Sounds expected to occur in home environment (1 file = 10 seconds duration)

<table>
<thead>
<tr>
<th></th>
<th>Label type</th>
<th># of data [ /file ]</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Training</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hard</td>
<td></td>
<td>2,045</td>
</tr>
<tr>
<td>Soft</td>
<td></td>
<td>1,578</td>
</tr>
<tr>
<td>Unlabeled</td>
<td></td>
<td>14,412</td>
</tr>
<tr>
<td><strong>Validation</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hard</td>
<td></td>
<td>1,168</td>
</tr>
<tr>
<td><strong>Evaluation</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hard</td>
<td></td>
<td>692</td>
</tr>
</tbody>
</table>

- **Evaluation measure**
  - F1-score [%] based on the interval of sound event occurrence
    - The student model is used for evaluation

<table>
<thead>
<tr>
<th>Num. of event classes: 10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alarm/bell/ringing</td>
</tr>
<tr>
<td>Blender</td>
</tr>
<tr>
<td>Cat</td>
</tr>
<tr>
<td>Dishes</td>
</tr>
<tr>
<td>Dog</td>
</tr>
</tbody>
</table>

2.20~4.10 [s]
# Four Competitive Approaches

<table>
<thead>
<tr>
<th>Model image</th>
<th>Baseline (mean-teacher)</th>
<th>Online KD(^{[3]}) w/ DA</th>
<th>PCL w/ DA (proposed)</th>
<th>PCL w/o DA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Student model</td>
<td>single network → output</td>
<td>single model → output</td>
<td>Student model → subnetwork 1 → output</td>
<td>Student model → subnetwork 1 → output</td>
</tr>
<tr>
<td>Teacher model</td>
<td>single network → output</td>
<td>single model → output</td>
<td>Teacher model → subnetwork 5 → output</td>
<td>Teacher model → subnetwork 5 → output</td>
</tr>
</tbody>
</table>

**Online KD w/ DA**: Online knowledge distillation with data augmentation

**PCL w/ DA**: Peer collaborative learning with data augmentation

**PCL w/o DA**: Peer collaborative learning *without* data augmentation
### Evaluation Results (F1-score [%])

<table>
<thead>
<tr>
<th></th>
<th>Baseline</th>
<th>Online KD w/ DA</th>
<th>PCL w/ DA</th>
<th>PCL w/o DA</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Validation</strong></td>
<td>25.9</td>
<td>43.1</td>
<td><strong>43.8</strong></td>
<td>41.7</td>
</tr>
<tr>
<td><strong>Evaluation</strong></td>
<td>31.1</td>
<td>43.4</td>
<td><strong>44.2</strong></td>
<td>42.4</td>
</tr>
</tbody>
</table>

- **Experimental findings**
  1. PCL Online KD > Baseline
  2. PCL w/ DA Online KD w/ DA > PCL w/o DA

- Confirmation of the effectiveness of the PCL model, which evolved from the online knowledge distillation and mean-teacher methods
- It is valid to design sub-networks based on the data augmentation process
## Evaluation Results (F1-score [%])

<table>
<thead>
<tr>
<th></th>
<th>Baseline</th>
<th>Online KD w/ DA</th>
<th>PCL w/ DA</th>
<th>PCL w/o DA</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Validation</strong></td>
<td>25.9</td>
<td>43.1</td>
<td><strong>43.8</strong></td>
<td>41.7</td>
</tr>
<tr>
<td><strong>Evaluation</strong></td>
<td>31.1</td>
<td>43.4</td>
<td><strong>44.2</strong></td>
<td>42.4</td>
</tr>
</tbody>
</table>

### Experimental findings

1. **PCL Online KD** > Baseline
2. **PCL w/ DA Online KD w/ DA** > **PCL w/o DA**

- Confirmation of the effectiveness of the PCL model, which evolved from the online knowledge distillation and mean-teacher methods
- It is valid to design sub-networks based on the data augmentation process
Conclusions

- **Motivation (Goal)**
  - Improvement of accuracy of polyphonic sound event detection on the DCASE Task4 task

- **Proposed approach**
  - **Peer collaborative learning** model, which evolved from the online knowledge distillation and mean-teacher methods with **audio data augmentation**

- **Experimental results (F1-score)**
  - Baseline (mean-teacher) 31.1% →⇒ PCL with data augmentation 44.2%

- **Future work**
  - We will implement and experiment with new knowledge distillation methods, such as collaborating with other knowledge distillation methods