Audio event detection using multiple-input convolutional neural network
Il-Young Jeong, Subin Lee, Yoonchang Han, Kyoung Lee

1 Music and Audio Research Group, Seoul National University, Seoul, Korea 2 Cochlear.ai, Seoul, Korea
{iyjeong, sblee, ychan}@cochlear.ai, kglee@snu.ac.kr

Introduction

This paper describes the model and training framework in our submission for DCASE 2017 task 3: sound event detection in real life audio. Extending the basic convolutional neural network architecture, we use both short- and long-term audio signal simultaneously as input data. In the training stage, we calculated validation error more frequently than one epoch with adaptive thresholds. We also used class-wise early-stopping strategy to find the best model for each class. The proposed model showed meaningful improvements in cross-validation experiments compared to the baseline system.

Architecture

- logMel: log mel-spec. (1024 window and 729 shift)
- logAvgMel: log of frequency-wise average of mel-spec of full music track.
- W: respective convnet. (consists of convolution and pooling)
- merge: combining the outputs of two layers (adding)
- Optimizer: Adam with 8 mini-batch size
- Batch generation: 
  ±88, 573 stereo samples from the random offset (short-term) pre-computed logAvgMel (long-term)
- Augmentation: channel swapping

Learning strategy

- Adaptive threshold
  - 0.5 of detecting threshold may not be optimal. (Due to imbalance data distribution and different error function between training/validation)
  - It is empirically chosen for every class/validation, to minimize the error of validation data.
- Class-wise early stopping
  - Class-wise cost converged with different speed.
  - The optimal early-stopping point for one class may be too early (or too late) for the other classes.
  - Our solution: do early stopping individually.
  - In the evaluation (test) stage, classes are detected by using the respective model.
- Frequent validation
  - Model validation for every fixed number of mini-batch iteration.
  - Allows the proper early-stopping in training procedure whose validation error has fluctuation.
- Cross-validation ensemble
  - 4-fold CV models are combined for DCASE submission
    - submission 1: majority vote (50% voting is considered as ‘active’)
    - submission 2: majority vote (50% voting is considered as ‘inactive’)
    - submission 3: majority vote without fold 1, that shows poor performance.
    - submission 4: weighted vote based on those validation ERL.

Results

Cross-validation results

<table>
<thead>
<tr>
<th>fold</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>average</th>
<th>baseline</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ER</td>
<td>F</td>
<td>ER</td>
<td>F</td>
<td>ER</td>
<td>F</td>
</tr>
<tr>
<td>brakes squeaking car children</td>
<td>1.00</td>
<td>0.0</td>
<td>1.00</td>
<td>0.0</td>
<td>0.97</td>
<td>11.8</td>
</tr>
<tr>
<td>large vehicle</td>
<td>0.78</td>
<td>35.8</td>
<td>0.71</td>
<td>65.7</td>
<td>0.27</td>
<td>85.4</td>
</tr>
<tr>
<td>people speaking</td>
<td>0.91</td>
<td>28.7</td>
<td>0.97</td>
<td>16.8</td>
<td>0.81</td>
<td>32.9</td>
</tr>
<tr>
<td>people walking</td>
<td>0.92</td>
<td>19.3</td>
<td>0.25</td>
<td>86.0</td>
<td>0.46</td>
<td>76.8</td>
</tr>
<tr>
<td>total</td>
<td>0.72</td>
<td>48.7</td>
<td>0.43</td>
<td>75.6</td>
<td>0.42</td>
<td>73.3</td>
</tr>
</tbody>
</table>

DCASE 2017 submission results

Discussion and future work

- The following difficulties were encountered during the research process.
  - Imbalance data distribution: we tried adaptive threshold, but it is unlikely to be a fundamental solution.
  - multi-label classification: It is not easy to use class-wise early stopping for many more classes (>100).
  - Annotation error: we could not handle it in this study.
  - data augmentation, noise-invariance, reducing computational cost for real-time detection…

any ideas or tips?