A Comparison of Five Multiple Instance Learning Pooling Functions for Sound Event Detection with Weak Labeling

Yun Wang, Juncheng Li, Florian Metze
May 14, 2019
Sound Event Detection

Detection = audio tagging + localization

Strong labeling is expensive to obtain
Sound Event Detection

- Train with weak labeling

- But still, we want both tagging and localization output
Multiple Instance Learning

- SED with weak labeling is a Multiple Instance Learning (MIL) problem
  - Bag is positive $\iff$ any instance is positive
  - Recording = bag, frames = instances
Multiple Instance Learning

Ground truth

Loss function

Poolung

Instance-level Predictions

Instance-level Classifier

Instances in a bag

Convolutional / Recurrent Network
Pooling Functions

1. **Max pooling**
   \[ y = \max_i y_i \]

2. **Linear softmax**
   \[ y = \frac{\sum_i y_i^2}{\sum_i y_i} \]

3. **Exp. softmax**
   \[ y = \frac{\sum_i y_i \exp(y_i)}{\sum_i \exp(y_i)} \]

4. **Average pooling**
   \[ y = \frac{1}{n} \sum_i y_i \]

5. **Weighted Average**
   \[ y = \frac{\sum_i y_i w_i}{\sum_i w_i} \]

**Attention:**
Learn the weights!
Pooling Functions

- We found **linear softmax** best for localization!
  \[ y = \frac{\sum_i y_i^2}{\sum_i y_i} \]
  \[ \frac{\partial y}{\partial y_i} = \frac{2y_i - y}{\sum_j y_j} \]
  Positive when \( y_i > y/2 \)

- When bag is positive:
  - \( y_i \) gets away from \( y/2 \)
  - Only boosts frames with \( y_i > y/2 \) – nice localization!

- When bag is negative:
  - \( y_i \) approaches \( y/2 \) – finally converges to zero 😊
Pooling Functions

What’s wrong with attention?

\[ y = \frac{\sum_i y_i w_i}{\sum_i w_i} \]

\[ \frac{\partial y}{\partial y_i} = \frac{w_i}{\sum_j w_j} \]

\[ \frac{\partial y}{\partial w_i} = \frac{y_i - y}{\sum_j w_j} \]

Always positive
Positive when \( y_i > y \)

When bag is positive:
- All \( y_i \) increase 😊, attention focuses where \( y_i > y \) 😊

When bag is negative:
- All \( y_i \) decrease 😕, attention focuses where \( y_i < y \) 😞
- Smaller probs get larger weight!
Failure Mode of Attention

- Too many frame-level false positives
- Inconsistent recording-level and frame-level predictions
EVALUATION I:
DCASE 2017 Challenge, Task 4
DCASE 2017: Task

- 17 event types
  - Vehicles, warnings
- Training data:
  - ~50k recordings * 10 seconds each = ~140 hours
  - Weakly labeled
- Test data:
  - 488 recordings * 10 seconds each = ~1.4 h
  - Strongly labeled
- Evaluation metrics:
  - Tagging: F1
  - Localization: error rate & F1 on 1s segments
**DCASE 2017: Model**

- **Input:**
  - Logmel features @ 40 Hz

- **Structure:**
  - 3 conv layers + 1 GRU layer

- **Output:**
  - Frame-level event probs at 10 Hz
  - **For tagging:** pooled globally into recording-level event probs
  - **For localization:** pooled over 1s segments
## DCASE 2017: Results

<table>
<thead>
<tr>
<th>Pooling Func</th>
<th>Tag F1</th>
<th>Loc ER</th>
<th>Loc F1</th>
<th>Loc #FN</th>
<th>Loc #FP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Max</td>
<td>45.3</td>
<td>84.7</td>
<td>35.4</td>
<td>3,154</td>
<td>1,253</td>
</tr>
<tr>
<td>Linear softmax</td>
<td>49.5</td>
<td>84.3</td>
<td>43.7</td>
<td>2,528</td>
<td>2,187</td>
</tr>
<tr>
<td>Attention</td>
<td>49.2</td>
<td>102.5</td>
<td>40.1</td>
<td>2,434</td>
<td>3,309</td>
</tr>
</tbody>
</table>

- **Max**: too many false negatives (FNs) hurt F1
- **Attention**: too many false positives (FPs) hurt ER
- **Linear softmax**: balanced FNs and FPs
EVALUATION II:
Google Audio Set
Audio Set: Task

Data:
- 527 event types (include the 17 events of DCASE)
- Weakly labeled
- Training: ~2M recordings * 10s = 8 months
- Test: ~20k recordings * 10s = 56 hours

Evaluation metrics:
- Audio Set only measures tagging
  - MAP, MAUC, d’
- Reuse DCASE data & metrics for tagging & localization
  - Tag F1, Loc ER, Loc F1 over 1s segments
Audio Set: Model

- **TALNet:**
  - Tagging and Localization Network
  - 10 conv layers, 1 GRU layer
  - Same input & output as before

- No fine-tuning when applied to DCASE data
Audio Set: Result 1/3

- TALNet works out of the box on DCASE
- Linear softmax is best for localization
  - And good enough for tagging
Audio Set: Result 2/3

<table>
<thead>
<tr>
<th>Group</th>
<th>System</th>
<th>No. of Training Recs.</th>
<th>Audio Set</th>
<th>DCASE 2017</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Map pooling</td>
<td></td>
<td>MAP</td>
<td>MAUC</td>
</tr>
<tr>
<td>TALNet</td>
<td>Average pooling</td>
<td>2M</td>
<td>0.351</td>
<td>0.961</td>
</tr>
<tr>
<td>(Sec. 3.3)</td>
<td>Linear softmax</td>
<td></td>
<td>0.361</td>
<td>0.966</td>
</tr>
<tr>
<td></td>
<td>Exp. softmax</td>
<td>2M</td>
<td>0.359</td>
<td>0.966</td>
</tr>
<tr>
<td></td>
<td>Attention</td>
<td></td>
<td>0.362</td>
<td>0.965</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2M</td>
<td>0.354</td>
<td>0.963</td>
</tr>
<tr>
<td>Literature</td>
<td>Hershey [71, 15]</td>
<td>1M</td>
<td>0.314</td>
<td>0.959</td>
</tr>
<tr>
<td></td>
<td>Kumar [128]</td>
<td>22k</td>
<td>0.213</td>
<td>0.927</td>
</tr>
<tr>
<td></td>
<td>Shah [48]</td>
<td>22k</td>
<td>0.229</td>
<td>0.927</td>
</tr>
<tr>
<td></td>
<td>Wu [131]</td>
<td>22k</td>
<td></td>
<td>0.927</td>
</tr>
<tr>
<td></td>
<td>Kong [54]</td>
<td>2M</td>
<td>0.327</td>
<td>0.965</td>
</tr>
<tr>
<td></td>
<td>Yu [55]</td>
<td>2M</td>
<td>0.360</td>
<td>0.970</td>
</tr>
<tr>
<td></td>
<td>Chen [56]</td>
<td>600k</td>
<td>0.316</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Chou [57]</td>
<td>1M</td>
<td>0.327</td>
<td>0.951</td>
</tr>
</tbody>
</table>

- TALNet closely matches state of the art on tagging
  - Yu’s system uses multi-level attention and can’t do localization!
- Amount of training data matters!
## Audio Set: Result 3/3

<table>
<thead>
<tr>
<th>Group</th>
<th>System</th>
<th>No. of Training Recs.</th>
<th>Audio Set</th>
<th>DCASE 2017</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>MAP</td>
<td>MAUC</td>
</tr>
<tr>
<td>TALNet (Sec. 3.3)</td>
<td>Max pooling</td>
<td>2M</td>
<td>0.351</td>
<td>0.961</td>
</tr>
<tr>
<td></td>
<td>Average pooling</td>
<td></td>
<td>0.361</td>
<td>0.966</td>
</tr>
<tr>
<td></td>
<td>Linear softmax</td>
<td></td>
<td>0.359</td>
<td>0.966</td>
</tr>
<tr>
<td></td>
<td>Exp. softmax</td>
<td></td>
<td><strong>0.362</strong></td>
<td><strong>0.965</strong></td>
</tr>
<tr>
<td></td>
<td>Attention</td>
<td></td>
<td>0.354</td>
<td>0.963</td>
</tr>
<tr>
<td>DCASE only (Sec. 3.2.3)</td>
<td>Max pooling</td>
<td>50k</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Average pooling</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Linear softmax</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Exp. softmax</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Attention</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- Adding more data helps the 17 DCASE events
  - Even though most of it belongs to 510 other events
Summary

- **Linear softmax** is the best for localization
  - Better than max: unobstructed gradient flow
  - Better than attention:
    - Balanced false negatives and false positives
    - Consistent frame-level & recording-level predictions

- **We built TALNet**
  - First simultaneous audio tagging and localization
  - Closely matches state of the art on Audio Set
  - Good performance on DCASE 2017 out of the box

- **Future work**
  - Attention pooling with monotonicity constraint?
The first two authors were supported by a graduate research fellowship award from Robert Bosch LLC; the first author was also supported by a faculty research award from Google. This work used the “comet” and “bridges” clusters of the XSEDE environment, supported by NSF grant number ACI-1548562.