A Hybrid Attention Mechanism for Weakly-Supervised Temporal Action Localization

Ashraful Islam\(^1\), Chengjiang Long\(^2\), Richard J. Radke\(^3\)

\(^1\)Dept. Of Electrical, Computer, and Sytems Engineering, Rensselaer Polytechnic Institute;  \(^2\)JD Digits AI Lab

**Problem & Background**

- Most current action localization methods depend on untrimmed videos with full temporal annotations.
- Full temporal annotations are expensive and time-consuming.
- We propose a weakly supervised temporal action localization method that only requires video-level action instances as supervision during training.
- During evaluation, the model predicts both action instances and temporal localizations.

**Contributions**

- We propose a hybrid attention mechanism to:
  - Detect complete actions
  - Model background activity

- We achieve state-of-the-art performance on both the THUMOS14 and ActivityNet-1.2 datasets.
- We describe extensive experiments to demonstrate the efficacy of different components of our approach.

**Proposed Approach**

**Attention Modules**

- Soft Attention: Remove background frames
- Hard Attention: Remove more confident foreground frames
- Semi-Soft Attention: Remove both confident foreground frames and background frames

**Ablation Studies on THUMOS14**

**Quantitative Results**

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Method</th>
<th>AP(_{50})</th>
<th>AP(_{75})</th>
<th>AOS(_{50})</th>
<th>AOS(_{75})</th>
</tr>
</thead>
<tbody>
<tr>
<td>THUMOS14</td>
<td>Base</td>
<td>42.1</td>
<td>26.4</td>
<td>32.4</td>
<td>20.8</td>
</tr>
<tr>
<td></td>
<td>Soft Attention</td>
<td>43.2</td>
<td>27.2</td>
<td>33.6</td>
<td>21.2</td>
</tr>
<tr>
<td></td>
<td>Hard Attention</td>
<td>44.1</td>
<td>28.3</td>
<td>34.5</td>
<td>22.1</td>
</tr>
<tr>
<td></td>
<td>Semi-Soft Attention</td>
<td>45.2</td>
<td>29.1</td>
<td>35.4</td>
<td>22.4</td>
</tr>
</tbody>
</table>

**Limitations:**

- Cannot detect multiple overlapping actions
- Cannot always solve action ambiguity

Both are partly due to the inherent limitations of the weakly supervised action localization paradigm. Additional supervision is required to solve these issues.

**Qualitative Results**

**Failure Case**

A person starts to do the high jump activity but stops short without completing the full action.

**Key References**

