A Closer Look at Temporal Ordering in the Segmentation of Instructional Videos

Anil Batra  Shreyank N Gowda  Frank Keller  Laura Sevilla-Lara

Overview

Task
Learning temporal representation of instructional videos.
Temporally identify the key steps and generate temporal summary.
Temporal segmentation is critical for generating correct textual summary.
Closely related to Dense Video Captioning Task[3].

Challenges
Proxy Evaluation Metrics
☑ Do not include 1-to-1 mapping between Ground Truth and Predicted segment
☑ Recursive search to find the best match with highest overlap
☑ Proposal detection metrics are used and overestimate

Proposal Detection Metric [3, 5]

1. P \_\text{det} = \{ p \in P | IoU (g, p) > \gamma \}
2. Recursive Overlap Metric [1]

Example

Dynamic Programming Matching

\[ C_{ij} = IoU(g_i, p_j) \quad S[i][j] = \max \left\{ \frac{S[i-1][j] + C_{ij}}{S[i][j-1] + C_{ij}} \right\} \]

Example

<table>
<thead>
<tr>
<th>Example</th>
<th>Precision</th>
<th>Recall</th>
<th>D</th>
<th>Y</th>
<th>L</th>
<th>L</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.26</td>
<td>0.32</td>
<td>0.32</td>
<td>D</td>
<td>O</td>
<td>L</td>
<td>L</td>
</tr>
<tr>
<td>0.26</td>
<td>0.75</td>
<td>0.75</td>
<td>T</td>
<td>O</td>
<td>L</td>
<td>L</td>
</tr>
<tr>
<td>0.26</td>
<td>0.75</td>
<td>0.75</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>D</td>
</tr>
</tbody>
</table>

Quantitative Results

Existing and Proposed Evaluation Metrics Comparison

<table>
<thead>
<tr>
<th>Method</th>
<th>Video Features</th>
<th>Matcher</th>
</tr>
</thead>
<tbody>
<tr>
<td>PDVC</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Ours</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

Procedure Segmentation and Summarization Comparison

<table>
<thead>
<tr>
<th>Method</th>
<th>Video Features</th>
<th>Matcher</th>
</tr>
</thead>
<tbody>
<tr>
<td>PDVC</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Ours</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

Sequential Matching Optimization

☑ Generate proposal based on Hungarian Matcher - NonDifferentiable
☑ Temporal structure of segments is not incorporated
2. SODA Matching (Ours)
☑ Differentiable matching algorithm.
☑ Plug into training pipeline to improve temporal segmentation performance.

Differentiable SODA Matching

\[ C_{ij} = -IoU(g_i, p_j) \quad S[i][j] = \min \left\{ \frac{S[i-1][j]}{S[i][j-1]} + C_{ij}, \right\} \]

Example

<table>
<thead>
<tr>
<th>Example</th>
<th>Precision</th>
<th>Recall</th>
<th>D</th>
<th>Y</th>
<th>L</th>
<th>L</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.26</td>
<td>0.32</td>
<td>0.32</td>
<td>D</td>
<td>O</td>
<td>L</td>
<td>L</td>
</tr>
<tr>
<td>0.26</td>
<td>0.5</td>
<td>0.5</td>
<td>D</td>
<td>Y</td>
<td>L</td>
<td>L</td>
</tr>
<tr>
<td>0.26</td>
<td>0.25</td>
<td>0.2</td>
<td>T</td>
<td>O</td>
<td>L</td>
<td>L</td>
</tr>
<tr>
<td>0.26</td>
<td>0.25</td>
<td>0.2</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>D</td>
</tr>
</tbody>
</table>

References:
1. Luowei Zhou et al. Towards Automatic Learning of Procedures from Web Instructional Videos, AAAI 2018
2. Anil Batra et al. AI4Life: Automatic KPI Prediction and Segmentation from Videos, CVPR 2019
4. Sener, Fadime et al. Pose Estimation with Video-structured Deep Control, ECCV 2018
6. Song Wang et al. PDVC: Pixel-End-to-End Video Capturing with Parallel Decoding, CVPR 2021

Acknowledgement:
This work was supported in part by the UKRI Centre for Doctoral Training in Natural Language Processing, funded by the UKRI (grant D0521481) and the University of Edinburgh, School of Informatics and School of Philosophy, Psychology & Language Science.