UntrimmedNets for Weakly Supervised Action Recognition and Detection

Limin Wang, Yuanjun Xiong, Dahua Lin, Luc Van Gool

ETH Zurich      The Chinese University of Hong Kong

Motivation: weakly supervised action recognition and detection

- Action recognition: training on trimmed videos
- Temporal annotation: expensive and subjective
- Large numbers of videos are untrimmed in nature.
- Learning directly from untrimmed video without temporal annotations.

UntrimmedNet: learning from untrimmed videos

- Untrimmed Video
- Clip Sampling
  - Clip proposal 1
  - Three stream CNNs
  - Temporal Segment Networks

Feature Extraction

Two stream CNNs

Spatial Stream CNN
Temporal Stream CNN

Classification

Softmax

Selection

Over proposals

Comparisons

Method

THUMO14
ActivityNet (v.1)
ActivityNet (v.2)

UntrimmedNet (w/o)

UntrimmedNet (w/t)


<table>
<thead>
<tr>
<th>Method</th>
<th>THUMO14</th>
<th>ActivityNet (v.1)</th>
<th>ActivityNet (v.2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>TNN (w/o) [5]</td>
<td>61.4%</td>
<td>64.4%</td>
<td>68.0%</td>
</tr>
<tr>
<td>TNN (w/t) [5]</td>
<td>66.8%</td>
<td>71.0%</td>
<td>72.3%</td>
</tr>
<tr>
<td>UntrimmedNet (w/o) [8]</td>
<td>71.7%</td>
<td>70.7%</td>
<td>71.6%</td>
</tr>
<tr>
<td>UntrimmedNet (w/t) [8]</td>
<td>71.7%</td>
<td>71.9%</td>
<td>68.9%</td>
</tr>
</tbody>
</table>

Evaluations on datasets of THUMO14 and ActivityNet 1.2

Examples

References: