Two-stream Fast R-CNN for detecting actions in videos

Qiong Cao

VGG reading group, 15/09/2016
Goal: to detect actions in videos

Input: video
Step 1: action detection in each frame

Architecture comparison

BMVC:

ECCV:
Step 1: action detection in each frame

Multi-region:
Step 2: link the action detections

- **idea**

  Action detections in video \(\rightarrow\) Define similarity between two detections for each action

- **formulations**

  BMVC:
  \[
  E(p_c) = \sum_{t=1}^{T} s_c^*(b_t) + \lambda_o \sum_{t=2}^{T} \psi_o(b_t, b_{t-1}),
  \]

  ECCV:
  \[
  s_c(R_t, R_{t+1}) = \{s_c(R_t) + s_c(R_{t+1}) + \beta \ ov(R_t, R_{t+1})\} \cdot \psi(ov),
  \]

  \[
  s_c(\mathcal{R}^*_c, e) = \arg\max_{(s,e)} \frac{1}{L(s,e)} \sum_{i=s}^{e} s_c(R_i) - \lambda \frac{|L(s,e) - L_c|}{L_c}
  \]

  Action path generalisation through the video

  Temporal trimming
Comparison results in video-level

Video-mAP (%)

<table>
<thead>
<tr>
<th></th>
<th>UCF-101 (delta = .2)</th>
<th>J-HMDB (delta = .5)</th>
</tr>
</thead>
<tbody>
<tr>
<td>BMVC</td>
<td>66.75</td>
<td>71.5</td>
</tr>
<tr>
<td>ECCV</td>
<td>42.27</td>
<td>73.09</td>
</tr>
</tbody>
</table>

Classification accuracy

<table>
<thead>
<tr>
<th></th>
<th>J-HMDB</th>
</tr>
</thead>
<tbody>
<tr>
<td>BMVC</td>
<td>70.0</td>
</tr>
<tr>
<td>ECCV</td>
<td>71.1</td>
</tr>
</tbody>
</table>