Asynchronous Temporal Fields for Action Recognition
Gunnar A. Sigurdsson, Santosh Divvala, Ali Farhadi, Abhinav Gupta

Understanding Sequences of Activities

High-order Temporal Model Trained From Randomly Sampled Individual Images

Results

- Evaluation on the Charades dataset
  157 classes, 66k instances, object, verb, scene, captions, etc

<table>
<thead>
<tr>
<th>Approach</th>
<th>mAP</th>
<th>Loss</th>
<th>mAP Improvement</th>
</tr>
</thead>
<tbody>
<tr>
<td>RGB++</td>
<td>15.6%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Two-Stream+</td>
<td>16.8%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Two-Stream+LSTM</td>
<td>17.8%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Two-Stream Extended</td>
<td>18.8%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>One (RGB Only)</td>
<td>18.3%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>One</td>
<td>22.4%</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- With Mean-Field, training is simple:
  1. Sample any frame
  2. Get stored "messages"
  3. Backprop using just this frame

Learned "Intent"

- We model a latent variable for groups of activities
- Learns higher-level activity concepts:
  Cluster 1: Get Ready to Leave the House
  Cluster 2: Photograph Something

Zooming in...

A Two-Stream network predicts the potentials

Nodes model labels in a frame
Objects, Actions, Scene, etc

"Intent" groups of activities
Latent variable

Fully-Connected over time
Prediction influenced by past/future

Message Server
Input
Messages
Input
Messages
Output
CNN
RGB &
Optical Flow
Video
Single Timepoint
Time

Training Method

- We model activities as a fully-connected CRF over time
- Nodes for each frame in the video, plus "intent"

Potential model labels
Predicted with a CNN

We model activites as a fully-connected CRF over time
Nodes for each frame in the video, plus "intent"