YouTube2Text: Recognizing and Describing Arbitrary Activities Using Semantic Hierarchies and Zero-shot Recognition

Sergio Guadarrama¹, Niveda Krishnamoorthy², Girish Mallarkarnerk², Subhashini Venugopalan²,
Raymond Mooney³, Trevor Darrell¹, Kate Saenko⁴
¹UC–Berkeley ²UT Austin ³UMass Lowell

Goals
Given a short YouTube video, output a natural language sentence that describes the main activity in the video.
When the model is not confident enough it should produce a less specific answer, but not over generalize.

Vision Pipeline

The outputs are over the leaves of the Hierarchies

Learning Hierarchies

Qualitative Results

K(x_i, x_j) = \exp \left( - \sum \frac{1}{A} D_i(x_i, x_j) \right)

Binary 0-1 accuracy

Comparison of WUP similarity

Zero-shot activity recognition

Human evaluation

Conclusions

We presented a system that takes a short video clip “in-the-wild” and outputs a brief sentence that sums up the main activity in the video, such as the actor, the action and its object. The semantic hierarchies learned from the data help to choose an appropriate level of generalization, and a prior learned from web-scale natural language corpora penalizes unlikely combinations of actors/actions/objects.

YouTUBE DATASET

We use the YouTube dataset collected by (Chen and Dolan, ACL 2011) consisting of 1970 videos and around 41 sentences on average per video, see (c) below

(a) Hollywood (8 actions)
(b) TRECVID MED (6 actions)
(c) YouTube (218 actions)

This new dataset (c) contains many more actions than the other previously used activity datasets (a-b).

Defining Semantic Accuracy

Given a Hierarchy of labels and a matching function \( \mu_i \), the accuracy \( \phi_H(f) \) over a hierarchy \( H \) with respect to a ground truth set leaf nodes \( L \) is defined by:

\[
\mu_i(c) = \max_{c \in L} \{ n_i(c) \} / |c|
\]

\[
m_{\mu_i}(v, l) = 2 \times \text{depth}(v) / \text{depth}(l)
\]

\[
\phi_H(f) = E[\mu_i(f(X))]
\]

We use Amazon Mechanical Turk to compare the methods by evaluating them on a video retrieval task.

Average Rating

The differences in the ratings of the three systems are statistically significant.

<table>
<thead>
<tr>
<th>FL</th>
<th>HE</th>
<th>OU</th>
<th>Ground Truth</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.81</td>
<td>1.54</td>
<td>1.99</td>
<td>3.90</td>
</tr>
</tbody>
</table>

Conclusions