Extracting Action Hierarchies from Action Labels and their Use in Deep Action Recognition

Konstantinos Bacharidis, Antonis Argyros
Computer Science Department, University of Crete, Greece
and
Institute of Computer Science, FORTH
Problem Statement

• **HAR datasets** smaller and less diverse compared to image recognition datasets
  -> *training data bottleneck in deep neural network learning*

• **Data limitation**: Multi-modal model designs utilizing language, audio or other sensory information

• **Language-infused designs**:
  - Script data introduced to speech recognition models
  - Constrained to a limited action set -> Action datasets with script data are scarce

• **Action/activity labels**: source of linguistic information *present in every dataset*
  - Contain motion motif(s), object presence, visual relationships
  - Motion motif commonalities -> common verbs or verbs with high semantic similarity
  - Object commonalities -> common nouns or nouns with high semantic similarity
Proposed method: Define Action Granularity Tree from Action Labels

- **Part-of-speech detection**: use a part-of-speech (POS) tagger from the Natural Language ToolKit (NLTK), to classify words into lexical categories.

- **Tag refinement**: refine with syntax rules to account for words with multiple semantic interpretations (e.g. *screw*: noun/verb, *take off* & *take out*).
  
  - **Verbs**: discriminate between cases of the same verb when followed by an ad-position or a particle (at, on, out, over, etc.).
  
  - **From noun to verb**: acceptable action description format
    
    \[ \text{verb} + \text{adposition/particle} + \text{noun} \]
  
- **Cluster** label sentences based on POS commonalities or high semantic content similarity.
  
  - **Similarity**: defined with a form of distance between the verb word embeddings in WordNet.
Incorporate Action Hierarchy in DNN designs

- **DNN design directions:**
  - **Modify** the temporal modelling sub-network.
  - **Mimic** the N-level action granularity with a set of subnets, one for each level (coarse-to-fine).
  - **Introduce** the learned representations of the coarser ones to finer sub-nets, using
    - Skip-connections & feature vector concatenations

- **Shallow action hierarchy** cost function:

  \[
  C = -\frac{1}{N} \sum_{n=1}^{N} \left[ \sum_{k=1}^{K} T_{n,k}^{gn} \log \left( Y_{n,k}^{gn} \right) + \sum_{l=1}^{L} w_l T_{n,l}^{fn} \log \left( Y_{n,l}^{fn} \right) \right]
  \]

  with \( w_l \) : vector of label associations of the fine-grained action classes,
  \((T_{n}^{gn}, T_{n}^{fn})\) : ground-truth labels for coarse- and fine-grained actions sets,
  \((Y_{n}^{gn}, Y_{n}^{fn})\) : the estimated action classes
Experimental Results

- **Datasets:** MHAD (11-actions), J-HMDB (21-actions), MPII Cooking Activities (64-actions)

<table>
<thead>
<tr>
<th>Datasets</th>
<th>MHAD</th>
<th>J-HMDB</th>
<th>MPII Cooking</th>
</tr>
</thead>
<tbody>
<tr>
<td>Num unique verbs</td>
<td>9 verbs</td>
<td>19 verbs</td>
<td>42 verbs</td>
</tr>
<tr>
<td>Avg num verbs/lbl</td>
<td>1.128 verb/lbl</td>
<td>1.0 verb/lbl</td>
<td>1.188 verbs/lbl</td>
</tr>
<tr>
<td>Avg asc via verb</td>
<td>0.545 asc/lbl</td>
<td>0.286 asc/lbl</td>
<td>1.656 asc/lbl</td>
</tr>
<tr>
<td>Max/min asc verb</td>
<td>1/0 asc</td>
<td>2/0 asc</td>
<td>5/0 asc</td>
</tr>
<tr>
<td>Num finer labels</td>
<td>11</td>
<td>21</td>
<td>64</td>
</tr>
<tr>
<td>Num Gen labels</td>
<td>8</td>
<td>18</td>
<td>36</td>
</tr>
</tbody>
</table>

- **Learning speed difference:**

- **Accuracy:**

<table>
<thead>
<tr>
<th>Architecture Design</th>
<th>Datasets (mAcc. (Coarse, Fine)%):</th>
<th>MHAD</th>
<th>J-HMDB</th>
</tr>
</thead>
<tbody>
<tr>
<td>NH-BiLSTM</td>
<td>(82.50, 70.25)%</td>
<td>(82.50, 70.25)%</td>
<td>(82.50, 70.25)%</td>
</tr>
<tr>
<td>H-BiLSTM</td>
<td>(89.61)%</td>
<td>(89.61)%</td>
<td>(89.61)%</td>
</tr>
<tr>
<td>NH-I3D</td>
<td>(98.75, 96.38)%</td>
<td>(98.75, 96.38)%</td>
<td>(98.75, 96.38)%</td>
</tr>
<tr>
<td>H-I3D</td>
<td>(98.75, 96.38)%</td>
<td>(98.75, 96.38)%</td>
<td>(98.75, 96.38)%</td>
</tr>
</tbody>
</table>

**TABLE:** Action recognition performance for MHAD, JHMDB and MPII datasets between hierarchical (H) and non-hierarchical (NH) deep architecture designs.

- **Observations on the impact of hierarchical design:**
  - 4-6% score improvement in every deep model & dataset case
  - Increases learning speed in the earlier epochs of learning
Experimental Results

- Visualization of learned representations with PCA:

Classification layer, MHAD dataset (a) Non-Hierarchical, (b) Proposed Hierarchical model design
Conclusions & Future Work

Conclusions:

- **Action labels** contain a considerable amount of action-related information.
- Identifying action class similarities a priori using the linguistic description provides useful insights regarding action complexity and hierarchy.
- Mimicking this hierarchical structure in a deep model design, leads to learning speed and accuracy improvement (up to 6%), compared to a non-hierarchical design.
- Despite the increase in the number of hyperparameters (+20-24%), learning at early stages is faster compared to a non-hierarchical design.

Future Work:

- Datasets with complex actions require elaborate linguistic analysis to capture the semantics contained in the action labels.
- Investigate the effect of increasing the levels of action granularity in
  - accuracy and learning speed
  - layer level fusion selection
Thank you!