Interaction Compass: Multi-Label Zero-Shot Learning of Human-Object Interactions via Spatial Relations

Dat Huynh
Ehsan Elhamifar
Khoury College of Computer Sciences
Northeastern University
Motivation

• Human-Object Interaction:
  • Classify/Localize *what actions* are performed on *what objects*

• Applications:
  - Assistive Robot
  - Human Computer Interaction
  - Security System
Contributions

• Combinatorically larger number of possible human-object interactions

Peel  Cut  Eat

Difficult to collect samples for all interactions!

• Zero-Shot Human-Object Interaction
  • Recognize unseen interactions as novel combination of seen actions and objects
Prior Works

• (Multi-Label) Zero-Shot Learning [Y. Zhang et al.'16, C. W. Lee et al. '18, D. Huynh et al. '20]

![Multi-Label Zero-Shot Learning Diagram](image)

Cannot capture relationships between action & object!

- **Person Shoe**
- **Skateboard**
- **Unseen labels**

• Human-Object Interaction Detection [Y. W. Chao et al. '18, Y. Li et al. 19', C. Gao et al. 20']

![Human-Object Interaction Detection Diagram](image)

Costly bbox annotations!

- **Action**
- **Object bbox**
Proposed Method

- Relational Direction:
  - *predictable displacement between action and object*

- Overview:
Visual Attention

Max cosine similarity between region feature and query

Visual Attention (Action)

Action Location
Action Feature

Action query:
- Jump
- Step
- Fly
- Unseen
Cross Attention

- **Action Location**
  - Action feature

- **Cross Attention**
  - Action $\rightarrow$ Object
  - Mean
  - Variance

- **Estimate location of object as Gaussian distribution**

- **Action Classifier**
  - Compute action score

- **Interaction score via action**

- **Object Classifier**
  - Compute object score
Experimental Results

- **Recognition**

<table>
<thead>
<tr>
<th>Method</th>
<th>HICO</th>
<th>Visual Genome</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Unseen</td>
<td>All</td>
</tr>
<tr>
<td></td>
<td>R@10</td>
<td>F1@10</td>
</tr>
<tr>
<td>DEVISE</td>
<td>54.4</td>
<td>8.8</td>
</tr>
<tr>
<td>Fast0Tag</td>
<td>76.8</td>
<td>12.5</td>
</tr>
<tr>
<td>LESA</td>
<td>71.1</td>
<td>11.5</td>
</tr>
<tr>
<td>Dual Attention</td>
<td>71.2</td>
<td>11.6</td>
</tr>
<tr>
<td>Combined Attention</td>
<td>71.1</td>
<td>11.5</td>
</tr>
<tr>
<td>ICompass (Ours)</td>
<td>82.7</td>
<td>13.4</td>
</tr>
</tbody>
</table>

- **Localization**

<table>
<thead>
<tr>
<th>Method</th>
<th>Recognition</th>
<th>Localization</th>
</tr>
</thead>
<tbody>
<tr>
<td>DEVISE</td>
<td>2.6%</td>
<td>5.8%</td>
</tr>
<tr>
<td>LESA</td>
<td>2.6%</td>
<td>5.8%</td>
</tr>
<tr>
<td>Dual Attention</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Combined Attention</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ICompass (Ours)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Qualitative Results

• Unseen object-action combination:

• Unseen action:
Code is available at:
https://github.com/hbdat/iccv21_relational_direction