**iCAN: Instance-Centric Attention Network**
for Human-Object Interaction Detection

Chen Gao  Yuliang Zou  Jia-Bin Huang
Virginia Tech


---

### Human-Object Interaction Detection

- Input image
- Object detection
- HOI detection

What is where? vs. What is happening?

---

### Instance-centric Attention Network

- Detecting objects with Faster R-CNN
- Predicting action scores with object/human/pairwise streams
- Fusing score to produce final predictions

**Core idea:** Appearance of an instance provides informative cues on where in the image we should pay attention to

---

### Instance-centric Attention Module

- Generating attention map conditioned on instance appearance
- Measuring the similarity in embedding space
- Using instance-centric attentional feature to complement the instance appearance feature

**Attention map visualization**

---

### Experimental Results

**Qualitative results on V-COCO**

- ride elephant
- ride motorcycle
- ride boat
- ride horse
- ride bike
- ride truck
- eat banana
- eat hot dog
- eat carrot
- eat sandwich
- eat donut
- eat pizza
- sit on coach
- sit on elephant
- sit on chair
- sit on horse
- sit on bike
- sit on coach
- lay on coach
- lay on bench
- lay on bed
- drink w/ sunglasses
- drink w/ cup
- drink w/ bottle

**Qualitative results on HICO-DET**

- hold motorcycle
- impact mass object
- jump motorcycle
- once motorcycle
- ren motorcycle
- shuttle-ouze
- lead elephant
- hose elephant
- kiss elephant
- pet elephant
- ride elephant
- walk elephant
- sit on boat
- wash boat
- read on boat
- drive boat
- repair boat
- row boat
- pot dog
- walk dog
- wash dog
- kiss dog
- dry dog
- carry dog
- catch sports ball
- throw sports ball
- kick sports ball
- hold sports ball
- hit sports ball
- dribble sports ball

**Detecting multiple actions**

- work on laptop
- read book
- sit on coach
- hold spoon
- eat with spoon
- sit on chair

**Quantitative results on V-COCO**

<table>
<thead>
<tr>
<th>Method</th>
<th>Feature backbone</th>
<th>AP$_{poly}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model C of [Gupta et al. 2015]</td>
<td>ResNet-50-FPN</td>
<td>31.8</td>
</tr>
<tr>
<td>BAR-CNN [Kolesnikov et al. 2018]</td>
<td>Inception-ResNet</td>
<td>41.1</td>
</tr>
<tr>
<td>iCAN (ours) w/ late fusion</td>
<td>ResNet-50</td>
<td>44.7</td>
</tr>
<tr>
<td>iCAN (ours) w/ early fusion</td>
<td>ResNet-50</td>
<td>45.3</td>
</tr>
</tbody>
</table>

**Quantitative results on HICO-DET**

<table>
<thead>
<tr>
<th>Method</th>
<th>Feature backbone</th>
<th>AP$_{poly}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zero-shot [Shen et al. 2018]</td>
<td>VGG-19</td>
<td>6.46</td>
</tr>
<tr>
<td>HD-BCNN [Chao et al. 2017]</td>
<td>CaffeNet</td>
<td>7.81</td>
</tr>
<tr>
<td>iCAN (ours)</td>
<td>ResNet-50</td>
<td>14.84</td>
</tr>
</tbody>
</table>

**Ablation study: mAP vs. time/model size**

- Baseline
- iCAN (early fusion)
- iCAN (late fusion)