A Fast and Accurate One-Stage Approach to Visual Grounding

Presenter: Tianlang Chen
Visual grounding

- Grounding a language query onto a region of the image

Query: bottom right grass
Existing framework

- Two-stage framework

Query: center building
Existing framework

- Performance is capped by the region candidates
- Slow in speed
One-stage visual grounding

(a). Two-stage visual grounding

Query: center building

Query: bottom right grass

(b). The proposed one-stage method

Query: bottom right grass

- One-stage approach
- Generally applicable for sub-tasks in grounding
Why one-stage visual grounding

- No region candidates -> 7~20% higher in accuracy
- One-stage -> 10x faster
Architecture overview

- Encoder
- Fusion module
- Grounding module
Architecture

- Encoder
- Fusion module
- Grounding module

- Visual encoder: DarkNet53+FPN
- Language encoder: Bert, LSTM, FV
- Spatial encoder: location related queries
Architecture

- Encoder
- Fusion module
- Grounding module

- Image-level fusion
Architecture

- Encoder
- Fusion module
- Grounding module

- Output format: box + confidence
Datasets

- Phrase localization: Flickr 30K Entities
- Referring expression comprehension: ReferItGame

An older man in a white jacket works at a stand featuring a wide variety of colorful food.

Flickr 30K Entities

the black backpack on the bottom right

ReferItGame

HAJIM SCHOOL OF ENGINEERING & APPLIED SCIENCES UNIVERSITY OF ROCHESTER
Comparison to other methods

ReferItGame

Flickr30K Entities

Inference Speed

<table>
<thead>
<tr>
<th>Method</th>
<th>Accu@0.5</th>
<th>Time (ms)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sim. Net-Resnet</td>
<td>34.54</td>
<td>184</td>
</tr>
<tr>
<td>CITE-Resnet</td>
<td>35.07</td>
<td>196</td>
</tr>
<tr>
<td>Ours</td>
<td>59.18</td>
<td>16</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Method</th>
<th>Accu@0.5</th>
<th>Time (ms)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sim. Net-Resnet</td>
<td>60.89</td>
<td></td>
</tr>
<tr>
<td>CITE-Resnet</td>
<td>61.33</td>
<td></td>
</tr>
<tr>
<td>Ours</td>
<td>68.38</td>
<td></td>
</tr>
</tbody>
</table>
Qualitative results

- Union of multiple objects
- Stuff as opposed to things
- Challenging regions
A Fast and Accurate One-Stage Approach to Visual Grounding

Code & models: https://github.com/zyang-ur/onestage_grounding

Poster: #26

Contact: zyang39@cs.rochester.edu