CLEVR-Ref+: Diagnosing Visual Reasoning with Referring Expressions

Runtao Liu, Chenxi Liu, Yutong Bai, Alan Yuille
02/12/2019
What is referring expression

giraffe on the right
What is wrong with referring expression

- Dataset bias
  - [1] showed that even if the referring expression is discarded and prediction is made solely from the image, accuracy is much higher than random.

What is wrong with referring expression

- Dataset bias
  - [1] showed that even if the referring expression is discarded and prediction is made solely from the image, accuracy is much higher than random

- Cannot inspect the entire visual reasoning process
  - Getting the final result right does not imply true understanding
  - Getting all the intermediate results right is a much stronger evidence

What is wrong with referring expression

- **Dataset bias**
  - [1] showed that even if the referring expression is discarded and prediction is made solely from the image, accuracy is much higher than random
- **Solution: Synthetic Dataset**

- **Cannot inspect the entire visual reasoning process**
  - Getting the final result right does not imply true understanding
  - Getting all the intermediate results right is a much stronger evidence

What is wrong with referring expression

- **Dataset bias**
  - [1] showed that even if the referring expression is discarded and prediction is made solely from the image, accuracy is much higher than random

- **Solution: Synthetic Dataset**

- **Cannot inspect the entire visual reasoning process**
  - Getting the final result right does not imply true understanding
  - Getting all the intermediate results right is a much stronger evidence

- **Solution: Modular Approach**

The CLEVR-Ref+ Dataset
CLEVR

- Simple, synthetic scene with full knowledge
- Synthetic sentences generated by templates
- Designed for visual question answering

Questions in CLEVR test various aspects of visual reasoning including attribute identification, counting, comparison, spatial relationships, and logical operations.

Q: Are there an equal number of large things and metal spheres?
Q: What size is the cylinder that is left of the brown metal thing that is left of the big sphere?
Q: There is a sphere with the same size as the metal cube; is it made of the same material as the small red sphere?
Q: How many objects are either small cylinders or red things?

## Step 1: Question -> Referring Expression

<table>
<thead>
<tr>
<th>Category</th>
<th>Question (CLEVR)</th>
<th>Referring Expression (CLEVR-Ref+)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Basic</td>
<td>How many cyan cubes are there?</td>
<td>The cyan cubes.</td>
</tr>
<tr>
<td>Spatial Relation</td>
<td>Are there any green cylinders to the left of the brown sphere?</td>
<td>The green cylinders to the left of the brown sphere.</td>
</tr>
<tr>
<td>AND Logic</td>
<td>How many green spheres are both in front of the red cylinder and left to the yellow cube?</td>
<td>The green spheres that are both in front of the red cylinder and left to the yellow cube.</td>
</tr>
<tr>
<td>OR Logic</td>
<td>Are there any cylinders that are either purple metal objects or small red matte things?</td>
<td>Cylinders that are either purple metal objects or small red matte things.</td>
</tr>
<tr>
<td>Same Relation</td>
<td>Are there any other things that have the same size as the red sphere?</td>
<td>The things/objects that have the same size as the red sphere.</td>
</tr>
<tr>
<td>Compare Integer</td>
<td>Are there more brown shiny objects behind the large rubber cylinder than gray blocks?</td>
<td>-</td>
</tr>
<tr>
<td>Comparison</td>
<td>Does the small ball have the same color as the small cylinder in front of the big sphere?</td>
<td>-</td>
</tr>
</tbody>
</table>
Step 2: Answer -> Bounding Box/Segmentation Mask

The big thing(s) that are behind the second one of the big thing(s) from front and to the right of the first one of the large sphere(s) from left

Any other things that are the same size as the fifth one of the thing(s) from right
Step 3: Module Additions

- We listed the frequent words in RefCOCO+ [1] and manually categorized them.
- We found that ordinal and visible are not covered by referring expression templates, so we added them.

<table>
<thead>
<tr>
<th>Category</th>
<th>Example words</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>object</td>
<td>shirt, head, chair, hat, pizza</td>
<td>63.66%</td>
</tr>
<tr>
<td>human</td>
<td>man, woman, guy, girl, person</td>
<td>42.54%</td>
</tr>
<tr>
<td>color</td>
<td>white, black, blue, red, green</td>
<td>38.76%</td>
</tr>
<tr>
<td>spatial</td>
<td>back, next, behind, near, up</td>
<td>23.86%</td>
</tr>
<tr>
<td>animal</td>
<td>zebra, elephant, horse, bear</td>
<td>15.36%</td>
</tr>
<tr>
<td>attribute</td>
<td>big, striped, small, plaid, long</td>
<td>10.55%</td>
</tr>
<tr>
<td>action</td>
<td>standing, holding, looking</td>
<td>10.34%</td>
</tr>
<tr>
<td>ordinal</td>
<td>closest, furthest, first, third</td>
<td>5.797%</td>
</tr>
<tr>
<td>compare</td>
<td>smaller, tallest, shorter, older</td>
<td>5.247%</td>
</tr>
<tr>
<td>visible</td>
<td>fully visible, barely seen</td>
<td>4.639%</td>
</tr>
</tbody>
</table>

Step 4: Changes to Generation Procedure

- Better balance between templates
- Remove referring expressions that are too peculiar and rare
- Better prevention of the referring expression from being degenerate
- Refer to at least one object at the end
Summary

- Same scenes as CLEVR
  - 70K images in train
  - 15K images in val
  - 15K images in test
- Every image is associated with 10 referring expressions
The IEP-Ref Model: Unifying Segmentation and Diagnosis
IEP

- Stands for “Inferring and Executing Programs”
- Modular approach
- Designed for visual question answering

Johnson, Justin, Bharath Hariharan, Laurens van der Maaten, Judy Hoffman, Li Fei-Fei, C. Lawrence Zitnick, and Ross Girshick. "Inferring and executing programs for visual reasoning." In ICCV, 2017.
Step 1: Parsing

The metallic things that are the second one of the cylinder(s) from right or large object(s)

LSTM Parser

- Scene
- Filter shape: cylinder
- Filter ordinal: second, from right
- Union
- Filter material: metal
- Scene
- Filter size: large
Step 2: Segmentation
Step 3: Diagnosis
Diagnosing Visual Reasoning: REF vs VQA

- The “logic” part is very similar. Both have to understand spatial relationship, logical operations etc.
Diagnosing Visual Reasoning: REF vs VQA

- The “logic” part is very similar. Both have to understand spatial relationship, logical operations etc.
- Recall that we are interested in visualizing intermediate steps. The output space of **visualization is basically the same as segmentation** (as in the REF setting), as opposed to textual answers (as in the VQA setting).
- In other words, we would have very little insight by plugging the “Exist” module (VQA setting) at intermediate steps.
Experimental Results:
Accuracy
Models

- In addition to IEP-Ref (seg), we also evaluated three existing SOTA referring expression models on CLEVR-Ref+, to see their strengths and weaknesses:
  - Speaker-Listener-Reinforcer (det) [1]
  - MAttNet (det) [2]
  - Recurrent Multimodal Interaction (seg) [3]

## Overall Evaluation

<table>
<thead>
<tr>
<th></th>
<th>Basic 0-Relate</th>
<th>Spatial Relation</th>
<th>Logic AND</th>
<th>Logic OR</th>
<th>Same</th>
<th>Accuracy</th>
<th>IoU</th>
</tr>
</thead>
<tbody>
<tr>
<td>SLR [35]</td>
<td>0.627</td>
<td>0.569</td>
<td>0.570</td>
<td>0.584</td>
<td>0.594</td>
<td>0.701</td>
<td>0.444</td>
</tr>
<tr>
<td>MAAttNet [33]</td>
<td>0.566</td>
<td>0.623</td>
<td>0.634</td>
<td>0.624</td>
<td>0.723</td>
<td>0.737</td>
<td>0.454</td>
</tr>
<tr>
<td>RMI [21]</td>
<td>0.822</td>
<td>0.713</td>
<td>0.736</td>
<td>0.715</td>
<td>0.585</td>
<td>0.679</td>
<td>0.251</td>
</tr>
<tr>
<td>IEP-Ref (GT)</td>
<td>0.928</td>
<td>0.895</td>
<td>0.908</td>
<td>0.908</td>
<td>0.879</td>
<td>0.881</td>
<td>0.647</td>
</tr>
<tr>
<td>IEP-Ref (700K prog.)</td>
<td>0.920</td>
<td>0.884</td>
<td>0.902</td>
<td>0.898</td>
<td>0.860</td>
<td>0.869</td>
<td>0.636</td>
</tr>
<tr>
<td>IEP-Ref (18K prog.)</td>
<td>0.907</td>
<td>0.858</td>
<td>0.874</td>
<td>0.862</td>
<td>0.829</td>
<td>0.847</td>
<td>0.605</td>
</tr>
<tr>
<td>IEP-Ref (9K prog.)</td>
<td>0.910</td>
<td>0.858</td>
<td>0.847</td>
<td>0.811</td>
<td>0.778</td>
<td>0.791</td>
<td>0.626</td>
</tr>
</tbody>
</table>
Basic Referring Ability

- Easy: color, shape, visibility
- Hard: ordinality
Spatial Reasoning Ability

- Seems that once the model has grasped spatial reasoning, there is little trouble in successfully applying it multiple times.
Different Reasoning Topologies

- Trees are generally harder, though not consistent.
Different Relation Types

- “Same” is harder than “Spatial”
- Presumably because “Same” requires global context
Number of Objects in a Scene

- More objects -> Harder
Schedule of Acquiring Reasoning Abilities

- No obvious schedule; all modules are learned at same pace
Novel Compositions

- Small drop in performance; generalize well in general
Experimental Results: IEP-Ref Interpretability
Figure 3. Visualizations of the norm of the gradient of the sum of the predicted answer scores with respect to the final feature map. From left to right, each question adds a module to the program; the new module is underlined in the question. The visualizations illustrate which objects the model attends to when performing the reasoning steps for question answering. Images are from the validation set.
Step-by-step inspection: chain structure

The first one of the big brown thing(s) from right
Step-by-step inspection: chain structure

Any other things that are the same size as the partially visible rubber thing(s)
Step-by-step inspection: tree structure

The metallic things that are the second one of the cylinder(s) from right or large object(s)
Step-by-step inspection: tree structure

The rubber object(s) that are in front of the first one of the cube(s) from front and left of the second one of the tiny rubber cube(s) from left
Step-by-step inspection: quantitative

- Every module is performing its intended job pretty well, except the red group.
Step-by-step inspection: quantitative

- Every module is performing its intended job pretty well, except the red group.
- “Same” and “Relate” are the only modules that may come after “Unique”.
- IoU after “Unique” is very low; yet after one more module, the segmentation mask becomes normal again.
Step-by-step inspection: quantitative

- Every module is performing its intended job pretty well, except the red group.
- “Same” and “Relate” are the only modules that may come after “Unique”.
- IoU after “Unique” is very low; yet after one more module, the segmentation mask becomes normal again.
- Perfect disentanglement for other modules, except learning some mechanism to treat “Unique” as the preprocessing step of “Same” and “Relate”.
False-premise referring expressions: quantitative

- We tested 10,000 referring expressions that refer to zero object at the end.
- Note that in every training example, at least one object is referred.
- IEP-Ref predicts:
  - 0 foreground pixel more than \(\frac{1}{4}\) of the time
  - \(\leq 8\) foreground pixels more than \(\frac{1}{3}\) of the time
False-premise referring expressions: chain structure

The fully visible big shiny block(s)
False-premise referring expressions: tree structure

*The big objects that are either purple metal object(s) or purple cylinder(s)*
Conclusion

1. CLEVR-Ref+: Synthetic dataset
   ○ not-so-novel, but necessary for 3 and 4
2. IEP-Ref: Modular approach
   ○ not-so-novel, but necessary for 3 and 4
3. Detailed analysis of the strengths and weaknesses of existing models
   ○ useful; impossible on real image datasets
4. An easy technique to clearly reveal the entire visual reasoning process
   ○ novel; definitive and quantitative proof that neural modules are doing the intended jobs!
Everything has been released

- **Paper:**

- **Dataset:**
  - https://cs.jhu.edu/~cxliu/data/clevr_ref+_1.0.zip
  - https://cs.jhu.edu/~cxliu/data/clevr_ref+_cogent_1.0.zip

- **Code:**
  - https://github.com/ccvl/clevr-refplus-dataset-gen
  - https://github.com/ccvl/iep-ref
Thank you!