CLEVR-Ref+: Diagnosing Visual Reasoning with Referring Expressions

Runtao Liu[1] Chenxi Liu[2] Yutong Bai[3] Alan Yuille[2]
[1]Peking University [2]Johns Hopkins University [3]Northwestern Polytechnical University

Motivation

Problem Description:

- Current referring expression datasets suffer from bias
- Current state-of-the-art models cannot be easily evaluated on intermediate reasoning process

Our Goal:

- **Building CLEVR-Ref+, a synthetic, diagnostic dataset**
  - Bias can be minimized
  - Ground truth visual reasoning process is available
- **Diagnosing state-of-the-art referring expression models**
- **Simple but effective step-by-step inspection of reasoning**

The CLEVR-Ref+ Dataset

Adaptation from CLEVR

- *From Question to Referring Expression*

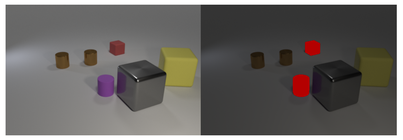
Question (CLEVR)	Referring Expression (CLEVR-Ref+)
How many cyan cubes are there?	The cyan cubes.
Are there any green cylinders to the left of the brown sphere?	The green cylinders to the left of the brown sphere.
How many green spheres are both in front of the red cylinder and left to the yellow cube?	The green spheres that are both in front of the red cylinder and left to the yellow cube.
Are there any other things that have the same size as the red sphere?	The things/objects that have the same size as the red sphere.

- *From Answer to Referred Objects*
  Output is no longer a textual answer; the bounding box or segmentation mask annotations are computed automatically.

Module Addition

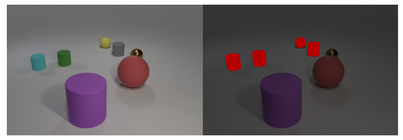
We add two new modules according to our investigation into the real-world referring expression dataset RefCOCO+.

Module - Ordinal



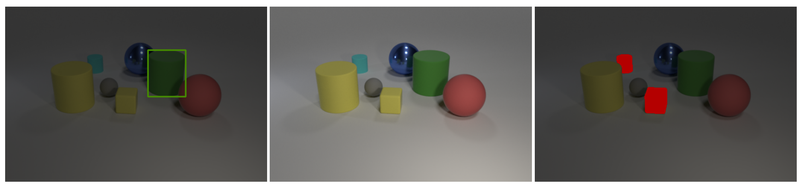
Any other tiny object(s) made of the same material as the second one of the cube(s) from front

Module - Visible



The fully visible small thing(s)

Example



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

Revealing Intermediate Reasoning

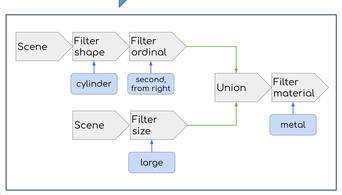
IEP-Ref Workflow

1. During training, the LSTM will first parse the input referring expression into the form of a program.
2. Then each module is parameterized with a small CNN; the `Segment` module performs prediction from the final module's output. *Note that Segment module is always at the end.*
3. During testing, we simply insert the trained `Segment` module after each intermediate module.

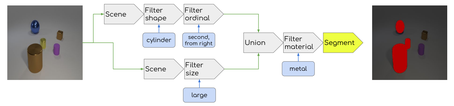
Step1. Parsing (Training)

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

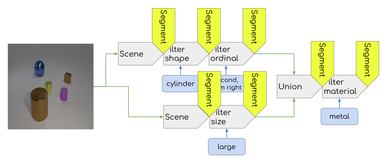
LSTM Parser



Step2. Segmentation (Training)



Step3. Diagnosis (Testing)



Experiments & Results

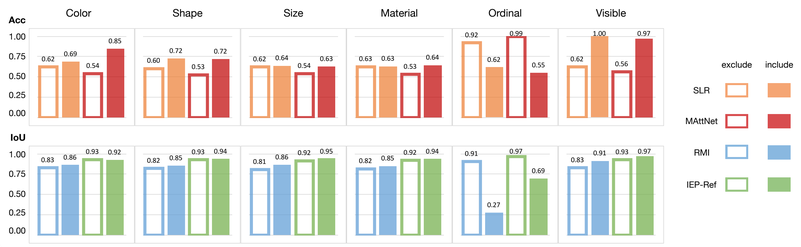
Overall Result on CLEVR-Ref+

The overall result shows that MAttNet and IEP-Ref performs much better, which suggests the importance to model compositionality within the referring expression.

	Basic	Spatial	Logic			
	0-Relate	3-Relate	AND	OR	Same	Acc/IoU
SLR	0.627	0.584	0.594	0.701	0.444	0.577
MAttNet	0.566	0.624	0.723	0.737	0.454	0.609
RMI	0.822	0.715	0.585	0.679	0.251	0.561
IEP-Ref-GT	0.928	0.908	0.879	0.881	0.647	0.816
IEP-Ref-700K	0.920	0.898	0.860	0.869	0.636	0.806
IEP-Ref-18K	0.907	0.862	0.829	0.847	0.605	0.782
IEP-Ref-9K	0.910	0.811	0.778	0.791	0.626	0.760

Results on Different Types of Module

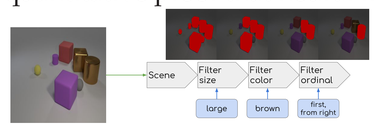
This shows basic referring ability of each model. "Include" means a module is involved. "Exclude" means otherwise. It seems that ordinality is the hardest concept to learn.



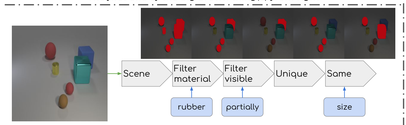
Step-By-Step Inspection

Examples of step-by-step inspection of IEP-Ref visual reasoning

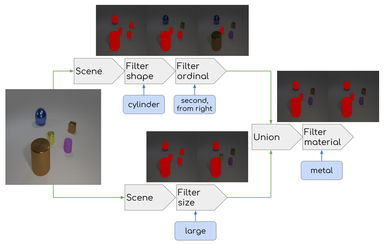
Here are several examples showing intermediate reasoning steps of IEP-Ref. To the best of our knowledge, we give the first direct and quantitative proof that neural modules behave as intended.



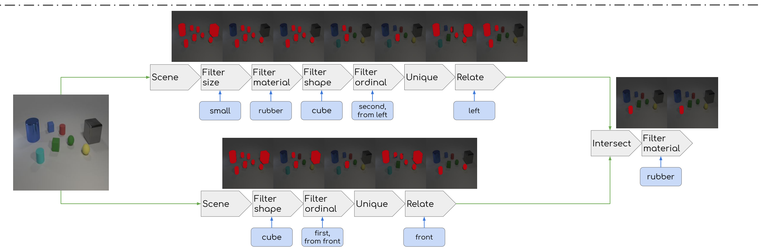
The first one of the big brown thing(s) from right



Any other things that are the same size as the partially visible rubber thing(s)



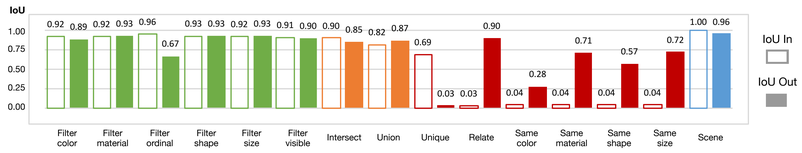
The metallic things that are the second one of the cylinder(s) from right or large object(s)



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

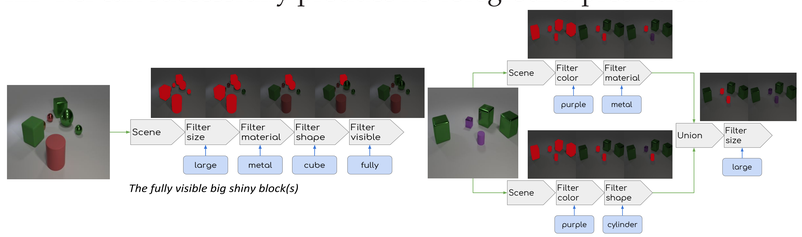
Average IoU going into/out of each IEP-Ref module

Quantitative evaluation shows a high IoU at intermediate steps, proving that the modules learned the job they are supposed to do.



False-Premise Referring Expressions

IEP-Ref can successfully produce no-foreground prediction.



The fully visible big shiny block(s)

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

Conclusion

- We build the CLEVR-Ref+ dataset which complements existing ones for referring expressions.
- We evaluate state-of-the-art referring expression models.
- We propose IEP-Ref, which uses a module network approach and outperforms competing methods by a large margin.
- Our qualitative and quantitative evaluation results shows that the neural modules work as expected.