# Modeling Relationships in Referential Expressions with Compositional Modular Networks

Ronghang Hu[1] Marcus Rohrbach[1,2] Jacob Andreas[1] Trevor Darrell[1] Kate Saenko[3]
[1]University of California, Berkeley [2]Facebook AI Research [3]Boston University

Project page and code: http://ronghanghu.com/cmn/

BERKELEY ARTIFICIAL INTELLIGENCE RESEARCH

## Relationships in referential expressions

Task: given a referential expression in natural language, ground (localize) the corresponding visual entities

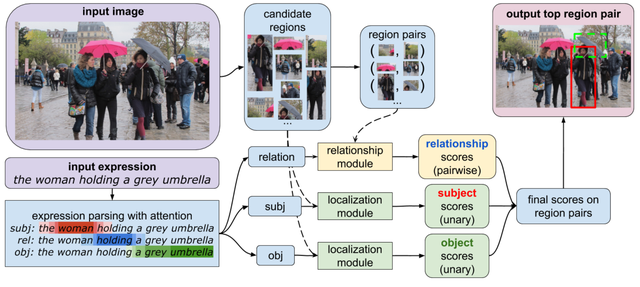
Localize the image regions corresponding to:
*"the woman in a cream-colored wedding dress cutting cake"*

- Expressions involve inter-object relationships
- Exploit the compositionality of natural language

## Our model: Compositional Modular Networks (CMNs)

**Input**: an image and a referential expression

**Output**: bounding box pair of the subject and the object



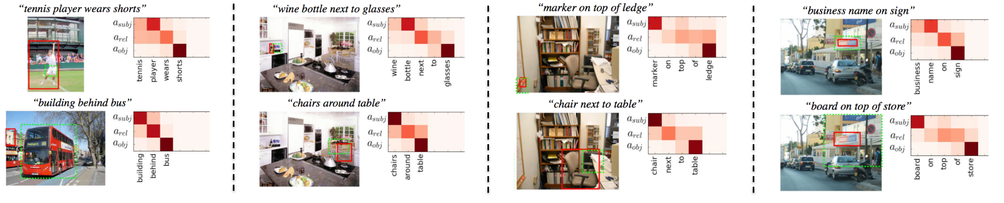

1. Extract a set of candidate regions (object proposals)
2. Soft-parse the expression into (*subject*, *relationship*, *object*) using soft-attention windows over the text
3. Unary subject and object scores with *localization module*
4. Pairwise relationship scores with *relationship module*
5. Sum the three scores into final pairwise scores

## Analysis of CMNs on multiple datasets and tasks

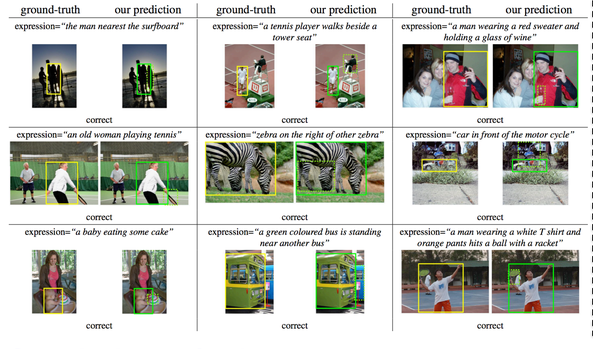
- Compositionality: grounding by localizing the entities and analyzing their relationships
- Modularity: different sub-tasks are handled by different modules

### Visualization

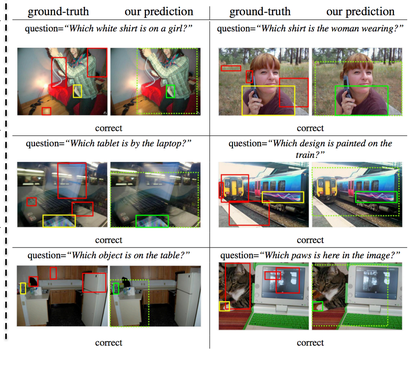
Localizing relationship expressions in Visual Genome [1]




Localizing referential expressions in Google-Ref [2]




"Which" questions in Visual-7W [3]




### Quantitative evaluation

- Evaluation: top-1 precision of localizing the described entities (subject or subj-obj pair)
- End-to-end training with subj-obj bounding box pair or subject bounding box only

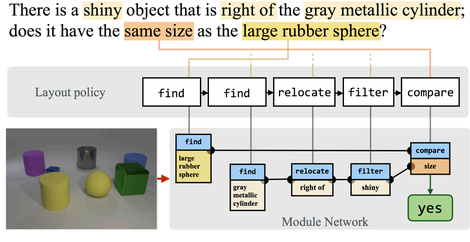
| on Visual Genome [1] | training supervision | P@1-subj | P@1-pair |
|---|---|---|---|
| our baseline (loc module) | subject-GT | 41.20% | - |
| our baseline (loc module) | subject-object-GT | - | 23.37% |
| our full model | subject-GT | 43.81% | 26.56% |
| our full model | subject-object-GT | **44.24%** | **28.52%** |

| on Google-Ref [2] | P@1-subj |
|---|---|
| Mao et al. 2016 [2] | 60.7% |
| Yu et al. 2016 [4] | 64.0% |
| Nagaraja et al. 2016 [5] | 68.4% |
| our baseline (loc module) | 66.5% |
| our full model | **69.3%** |

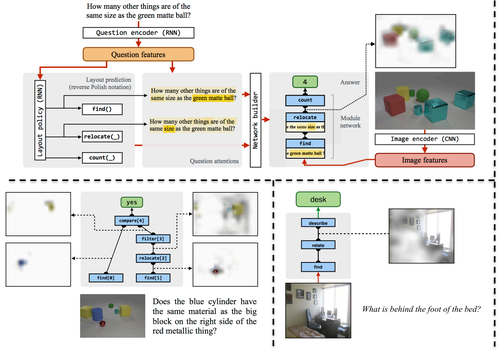
| on Visual-7W "Which" questions [3] | P@1-subj |
|---|---|
| Nagaraja et al. 2016 [5] | 56.10% |
| our baseline (loc module) | 71.61% |
| our full model | **72.53%** |

## Extension – End-to-End Module Networks (N2NMN)

- Learning the network structure and module layout end-to-end
- Free-form layouts beyond (*subject, relationship, object*) triplet




- Details
  - Checkout the paper: Learning to Reason: End-to-End Module Networks for Visual Question Answering *in arXiv preprint arXiv:1704.05526, 2017*




| on CLEVR [6] | Accuracy |
|---|---|
| CNN+BoW [8] | 48.4 |
| CNN+LSTM [7] | 52.3 |
| CNN+LSTM+MCB [9] | 51.4 |
| CNN+LSTM+SA [10] | 68.5 |
| NMN [11] | 72.1 |
| ours (N2NMN) | **83.7** |

| on VQA [7] | Accuracy |
|---|---|
| CNN+BoW [8] | 55.7 |
| CNN+LSTM [7] | 53.7 |
| CNN+LSTM+MCB [9] | 64.7 |
| CNN+LSTM+SA [10] | 57.6 |
| NMN [11] | 58.6 |
| ours (N2NMN) | **64.2** |

- Project page and code of N2NMN: http://ronghanghu.com/n2nmn/

**References**
[1] Krishna, Ranjay, et al. "Visual genome: Connecting language and vision using crowdsourced dense image annotations." *IJCV* 123.1 (2017): 32-73.
[2] Mao, Junhua, et al. "Generation and comprehension of unambiguous object descriptions." *CVPR*. 2016.
[3] Zhu, Yuke, et al. "Visual7w: Grounded question answering in images." *CVPR*. 2016.
[4] Yu, Licheng, et al. "Modeling context in referring expressions." *ECCV*, 2016.
[5] Nagaraja, Varun K., Vlad I. Morariu, and Larry S. Davis. "Modeling context between objects for referring expression understanding." *ECCV*, 2016.
[6] Johnson, Justin, et al. "CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning." *CVPR*. 2017.
[7] Antol, Stanislaw, et al. "Vqa: Visual question answering." *ICCV*. 2015.
[8] Zhou, Bolei, et al. "Simple baseline for visual question answering." *ECCV*, 2016.
[9] Fukui, Akira, et al. "Multimodal compact bilinear pooling for visual question answering and visual grounding." *EMNLP*, 2016.
[10] Yang, Zichao, et al. "Stacked attention networks for image question answering." *CVPR*, 2016.
[11] Andreas, Jacob, et al. "Neural module networks." *CVPR*, 2016.