CMPT 983

Grounded Natural Language Understanding

February 11th, 2021

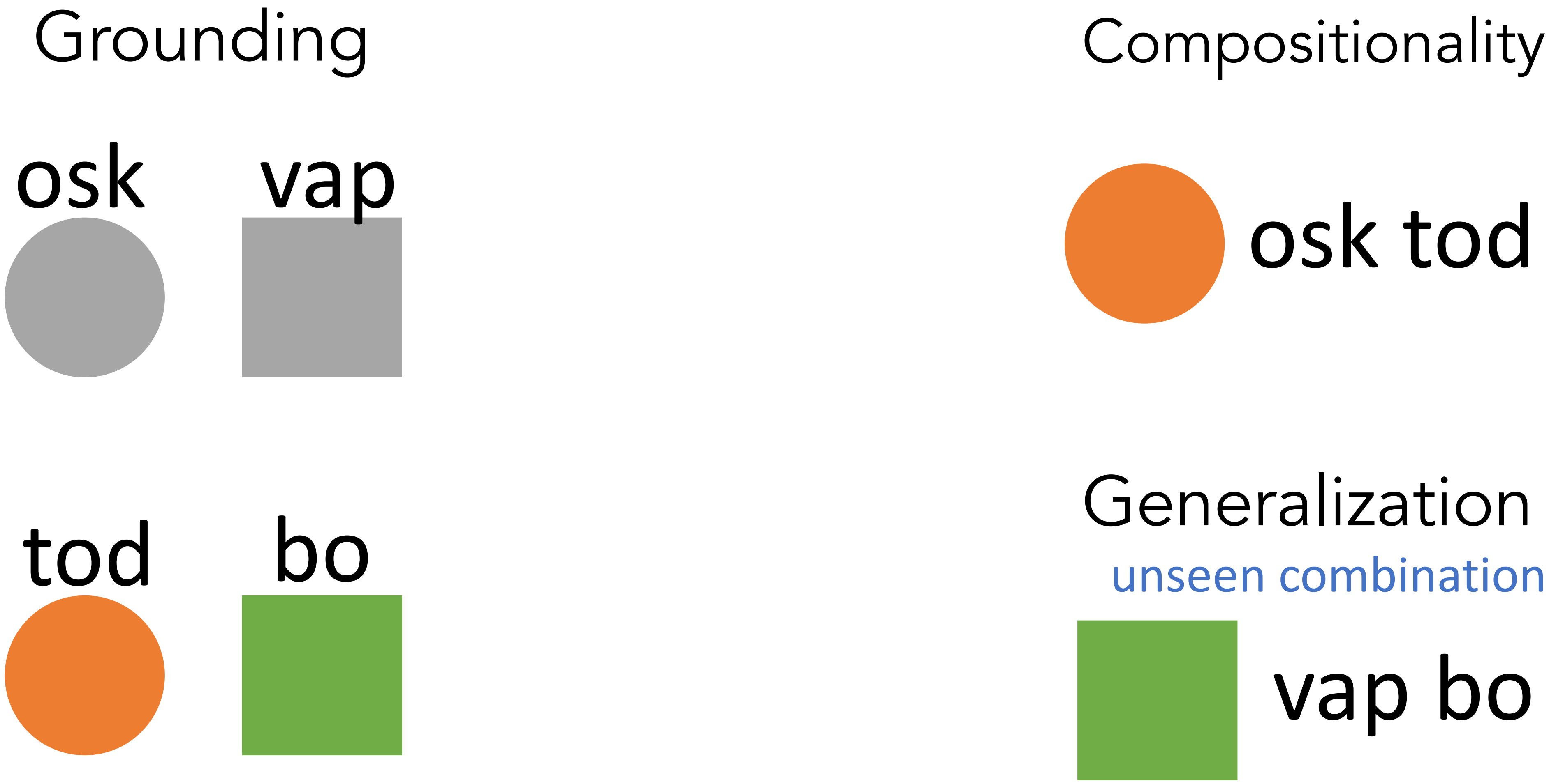
Compositionality and Structure

Today

- Compositionality
- Structured representations
- Structured reasoning

Compositionality

Compositional Generalization



Studying compositionality

- Systematic study

  Controlled settings to study specific aspects of language learning:
- Easier to study in smaller, synthetic generated datasets

Compositional Generalization

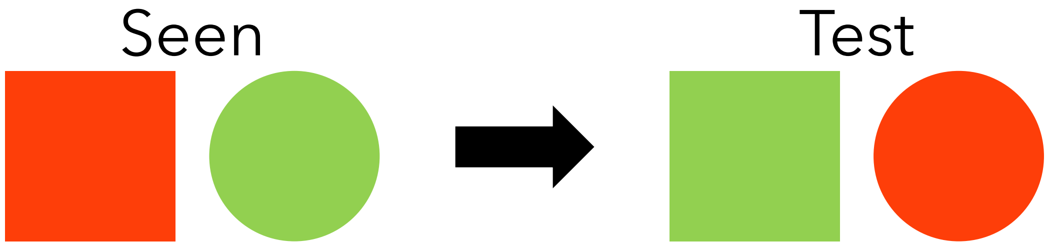
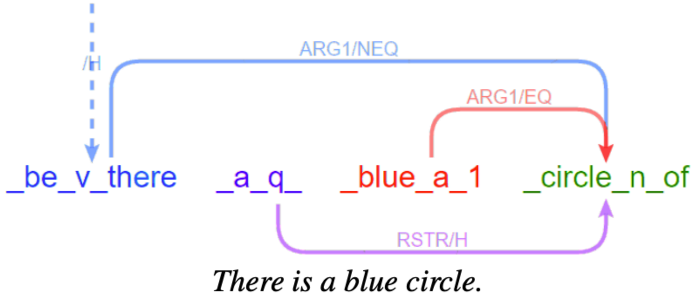


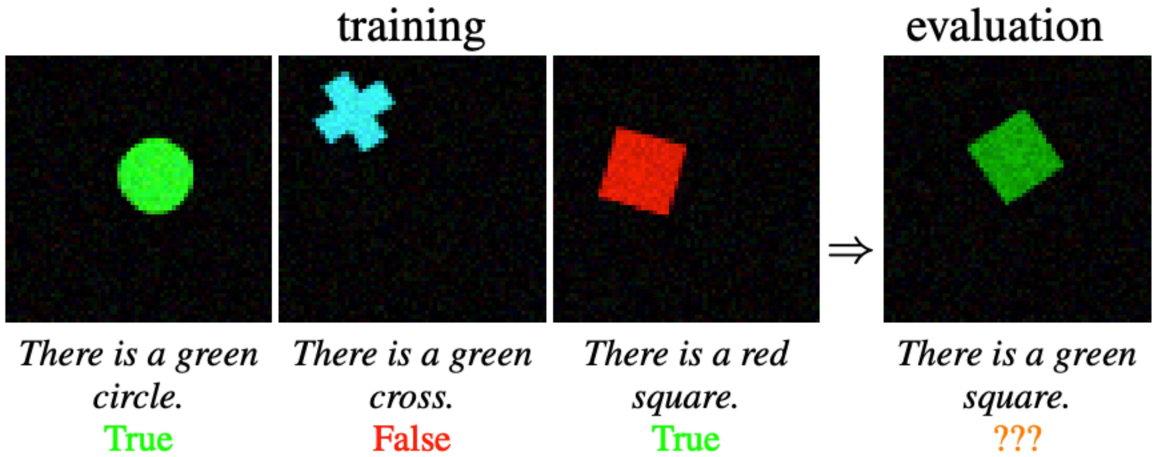
Image credit: Stefan Lee

ShapeWorld

- Framework to generate ``worlds'' and matching captions
- Language generated from semantic graph
- Task: Does the Image-Caption match?
- Training: Simple color + shape combination
- Evaluation: unseen color shape combination

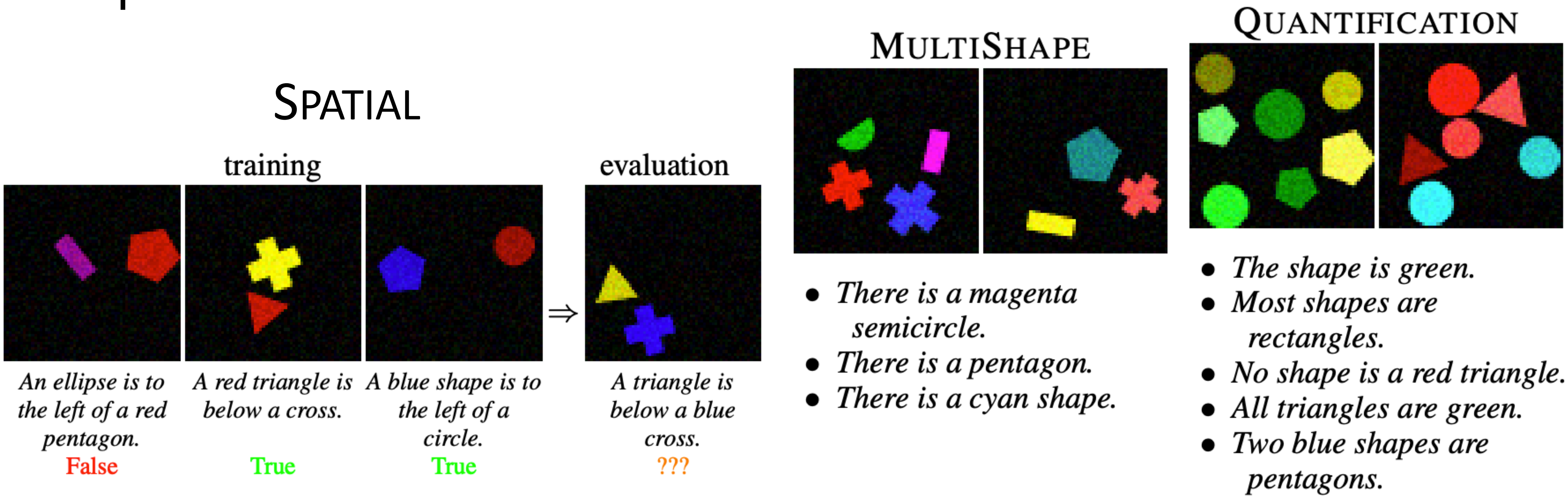


ONESHAPE



SHAPEWORLD: A new test methodology for multimodal language understanding, Kuhnle and Copestake, NeurIPS 2019

ShapeWorld – 4 datasets



Can the model generalize to unseen relation + color + shape combinations?

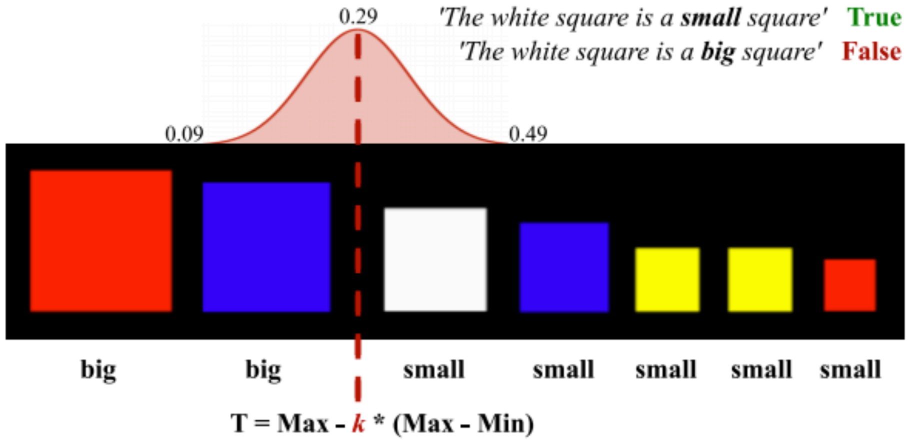
Can the model pick out shape from many, and generalize to unseen number of objects?

SHAPEWORLD: A new test methodology for multimodal language understanding, Kuhnle and Copestake, NeurIPS 2019

Dataset configuration	LSTM-only	CNN+LSTM:Mult	CNN+CNN:HCA-par	CNN+CNN:HCA-alt
ONESHAPE	51 / 46 / 50	81 / 70 / 66	90 / 77 / 78	**92 / 81 / 77**
C: no hypernyms	90 / 70 / 100	95 / 64 / 57	98 / 71 / 73	97 / 68 / 66
C: only hypernyms	100 / 100 / 100	52 / 34 / 30	96 / 78 / 82	95 / 75 / 73
I: changed shape	6 / 5 / 7	70 / 81 / 82	60 / 63 / 58	73 / 78 / 78
I: changed color	8 / 15 / 0	100 / 100 / 99	100 / 92 / 96	100 / 97 / 89
I: changed both	7 / 5 / 6	96 / 97 / 98	87 / 85 / 84	93 / 92 / 89
MULTISHAPE	62 / 67 / 67	**72 / 71 / 72**	**72 / 71 / 69**	71 / 68 / 68
correct instances	48 / 49 / 50	76 / 64 / 54	81 / 68 / 65	71 / 59 / 53
I: random attr.	58 / 63 / 68	67 / 74 / 79	64 / 67 / 68	70 / 73 / 78
I: random existing attr.	100 / 100 / 100	78 / 86 / 95	55 / 71 / 79	72 / 87 / 95
SPATIAL	52 / 51 / 50	57 / 52 / 54	**63 / 65 / 64**	54 / 52 / 55
C: no hypernyms	85 / 85 / 69	45 / 44 / 41	83 / 83 / 86	92 / 62 / 100
C: only hypernyms	95 / 95 / 97	4 / 6 / 4	60 / 59 / 65	49 / 40 / 52
I: swapped direction	11 / 13 / 16	98 / 97 / 98	36 / 39 / 30	50 / 61 / 47
I: object random attr.	15 / 12 / 16	88 / 88 / 91	69 / 68 / 68	63 / 66 / 60
I: subject random attr.	13 / 12 / 17	87 / 88 / 89	69 / 71 / 70	61 / 64 / 56
QUANTIFICATION	57 / 57 / 56	56 / 56 / 58	**76 / 77 / 78**	**74 / 77 / 78**
correct instances	23 / 22 / 18	25 / 30 / 26	74 / 71 / 72	70 / 71 / 75
incorrect instances	94 / 93 / 93	88 / 90 / 88	81 / 83 / 88	78 / 82 / 82
instances with *no*	52 / 51 / 48	61 / 60 / 61	56 / 56 / 51	55 / 55 / 58
instances with *the* (=1)	53 / 58 / 61	55 / 59 / 58	59 / 59 / 55	63 / 63 / 63
instances with *a* (≥1)	34 / 35 / 36	34 / 36 / 37	49 / 50 / 51	48 / 52 / 50
instances with *two* (≥2)	53 / 48 / 48	50 / 50 / 49	70 / 69 / 62	72 / 67 / 58
instances with *most*	49 / 50 / 49	48 / 48 / 49	69 / 68 / 60	60 / 52 / 51
instances with *all*	52 / 54 / 50	48 / 50 / 51	47 / 52 / 51	49 / 50 / 51

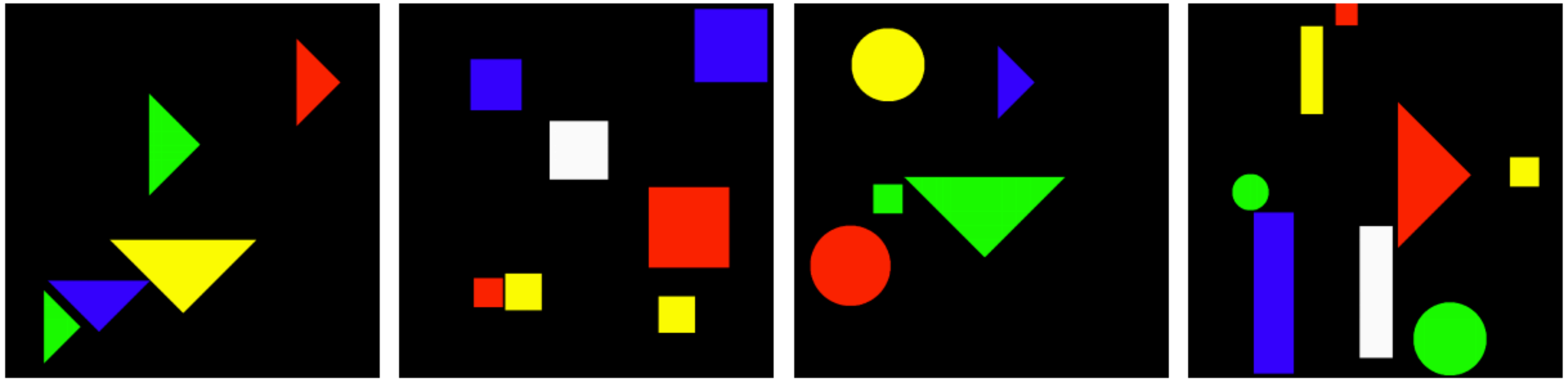
MaleVic

- Size understanding
- Programmatically determine big / small using thresholds



Superlative
The yellow triangle is the biggest triangle.

Any shape
The red circle is a big object.



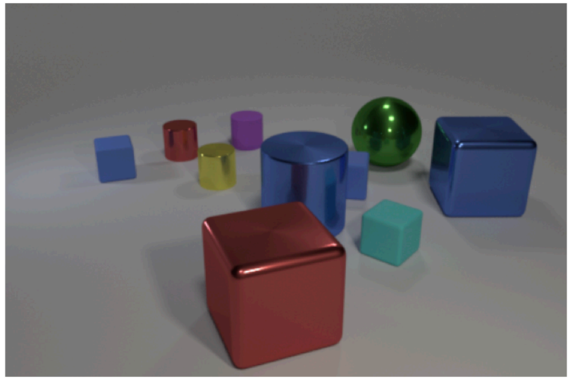
Same shape
The white square is a small square.

Pick shape from different shapes
The white rectangle is a big rectangle.

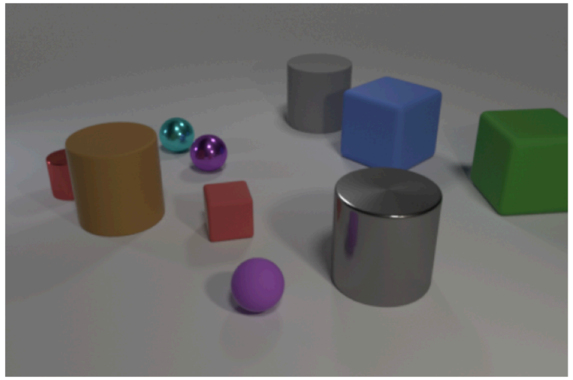
Is the Red Square Big? MALeViC: Modeling Adjectives Leveraging Visual Contexts, Pezzelle and Fernandez, EMNLP-IJCNLP 2019

CLEVR: Compositionality and reasoning

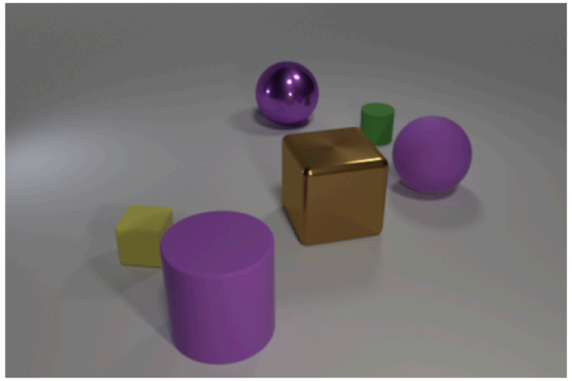
- VQA - Answering questions is a good way to assess understanding
- Diagnostic dataset for probing visual understanding and reasoning



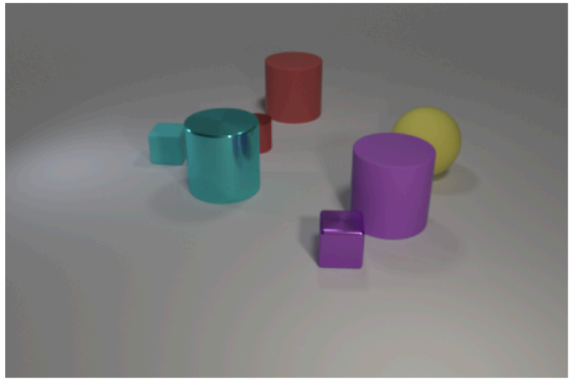
Q: What shape is the object reflected in the blue cylinder?
A: cube



Q: What number of cylinders share the same color?
A: 2



Q: How many objects are not purple and not metallic?
A: 2

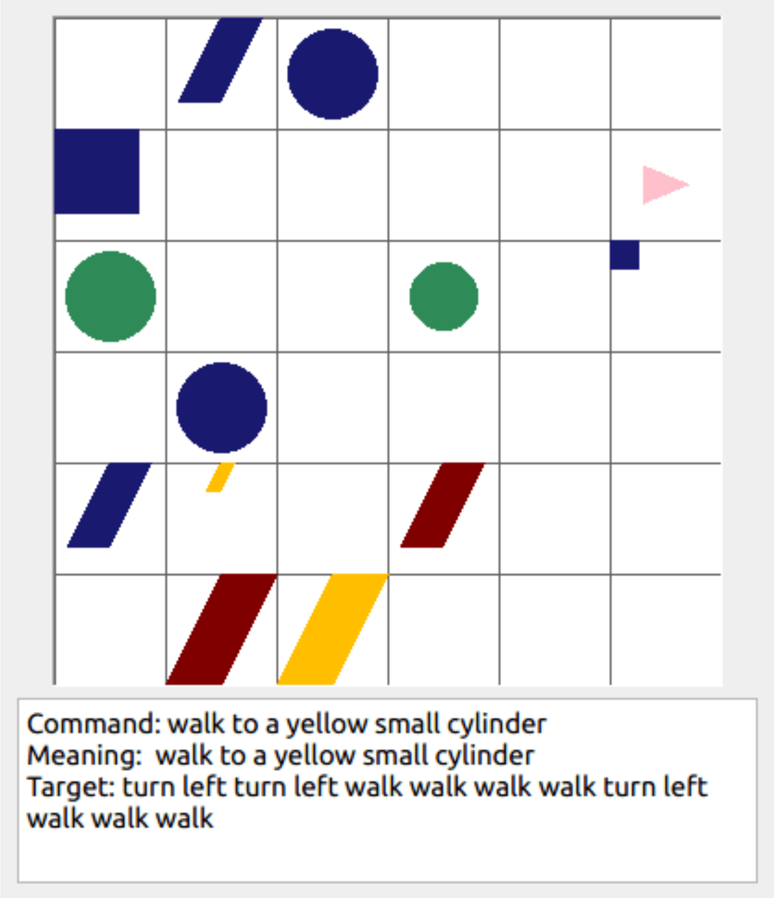


Q: What color is the object partially blocked by the purple cylinder?
A: yellow

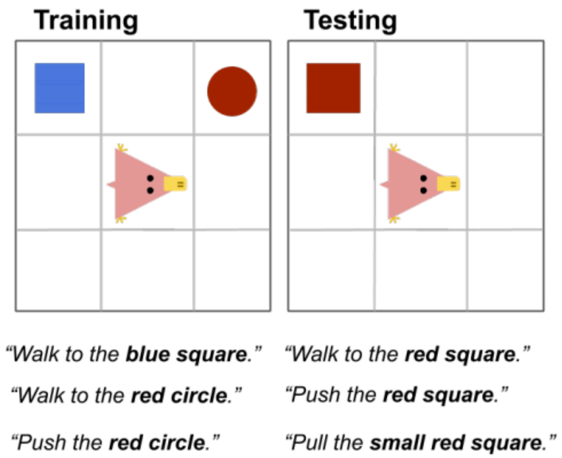
CLEVR: A Diagnostic Dataset for Compositional Language and Elementary Visual Reasoning, Johnson et al, CVPR 2017

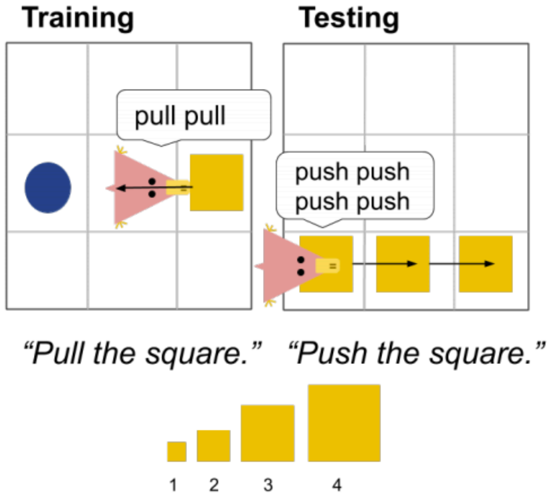
Compositionality with actions

Generate worlds and language



$$
\begin{aligned}
&\text{ROOT} \rightarrow \text{VP} \\
&\text{VP} \rightarrow \text{VP RB} && \text{VV}_i \rightarrow \{\text{walk}\} \\
&\text{VP} \rightarrow \text{VV}_i \text{ `}to\text{' DP} && \text{VV}_t \rightarrow \{\text{push, pull}\} \\
&\text{VP} \rightarrow \text{VV}_t \text{ DP} && \text{RB} \rightarrow \{\text{while spinning, while zigzagging, hesitantly, cautiously}\} \\
&\text{DP} \rightarrow \text{`}a\text{' NP} && \text{NN} \rightarrow \{\text{circle, square, cylinder}\} \\
&\text{NP} \rightarrow \text{JJ NP} && \text{JJ} \rightarrow \{\text{red, green, blue, big, small}\} \\
&\text{NP} \rightarrow \text{NN}
\end{aligned}
$$

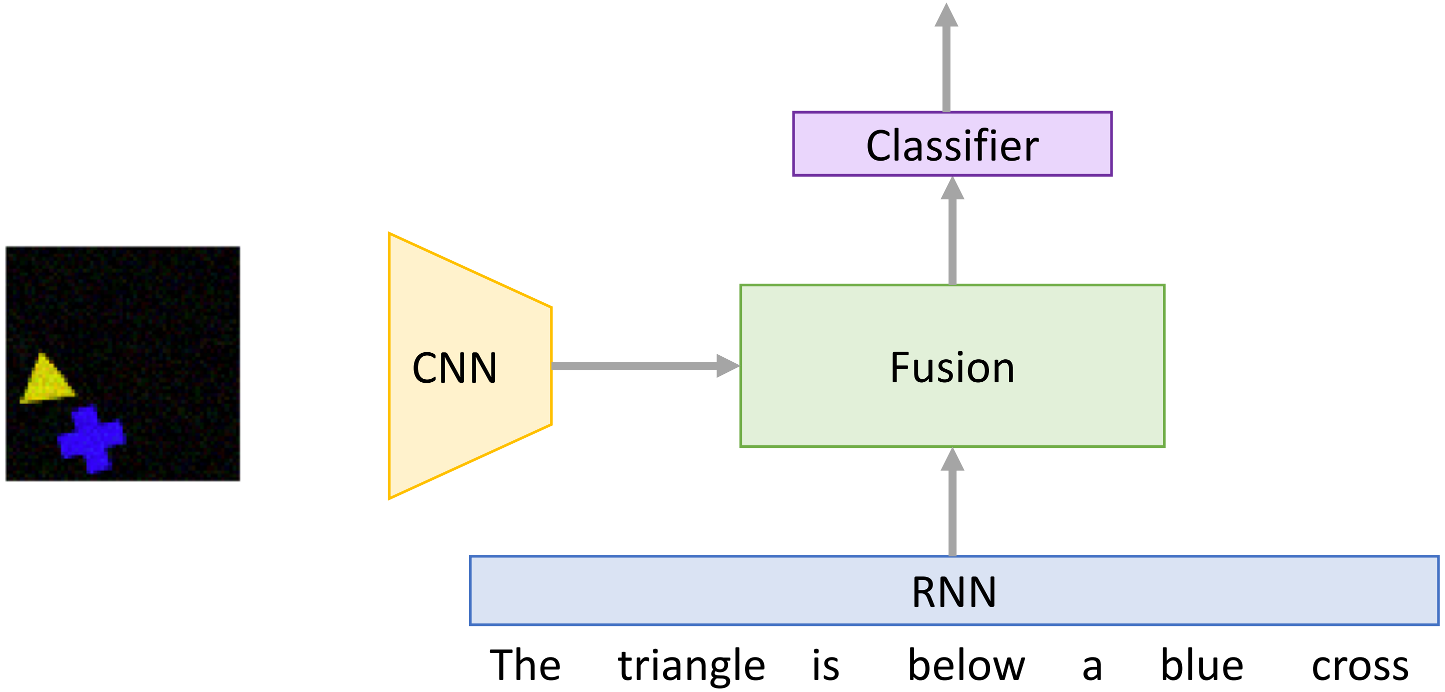




How to achieve compositionality?

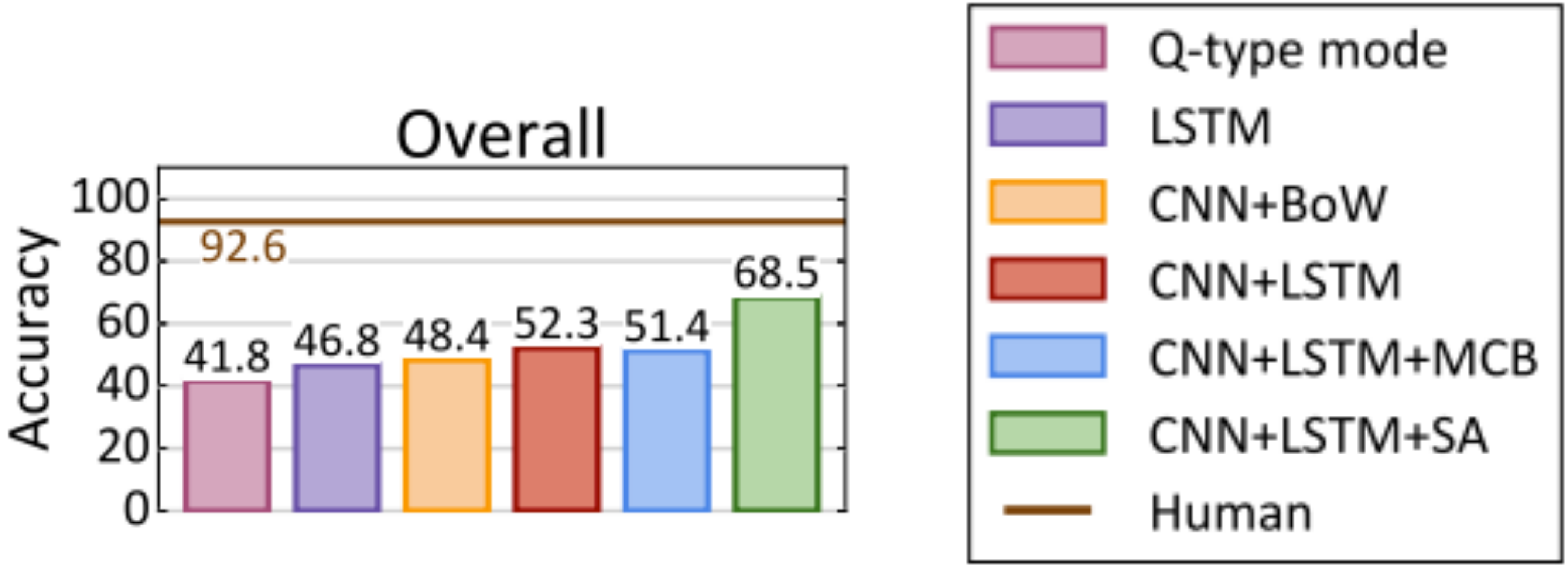
Baseline networks

- Language: RNN
- Vision: CNN
- Fusion
- Classifier



- One way to achieve compositionality is by considering structured representations and reasoning over structures

CLEVR baseline performance

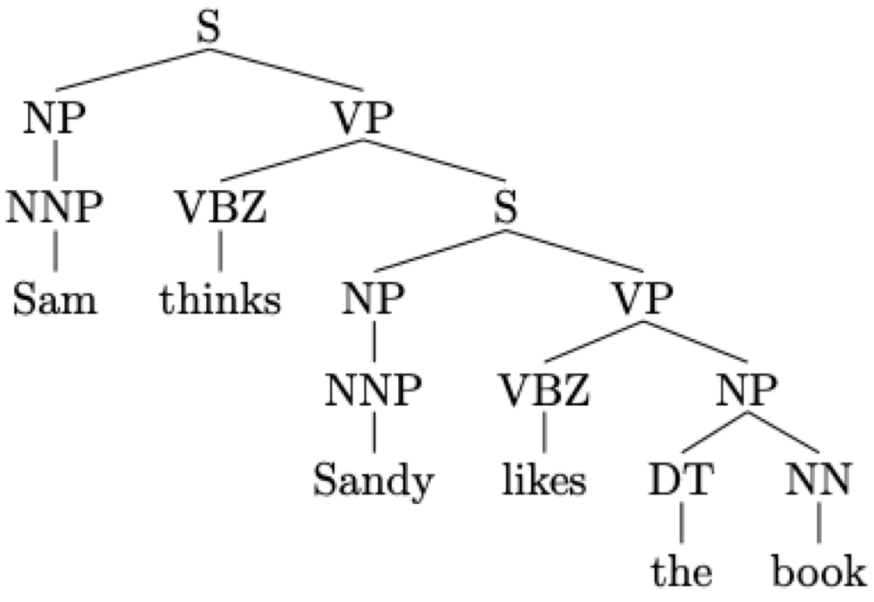


Structured representations

Structured representation of sentences

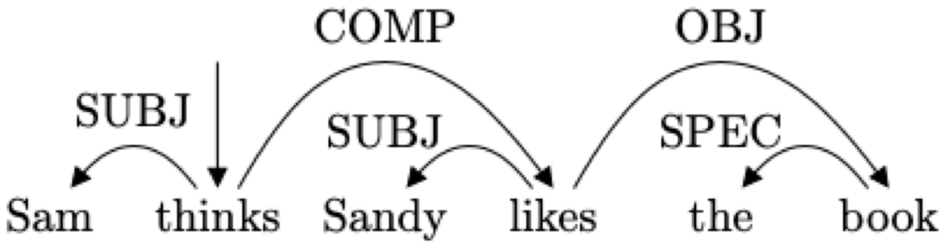
Constituency Parse Tree

Hierarchical



Dependency Parse

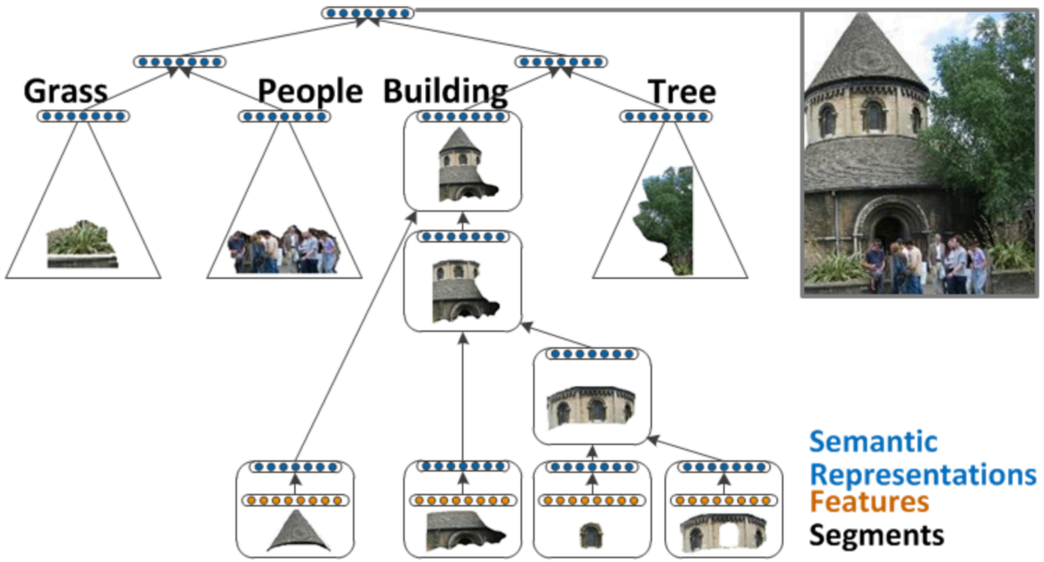
Relational



Structured representation of images

Scene Parse Tree

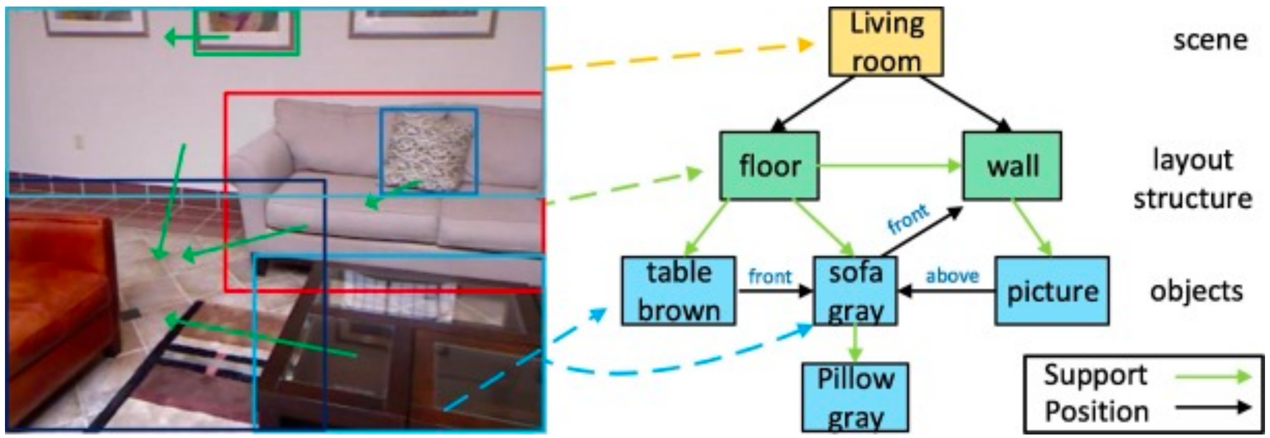
Hierarchical



Socher, Lin, Ng, and Manning, "Parsing Natural Scenes and Natural Language with Recursive Neural Networks", ICML 2011

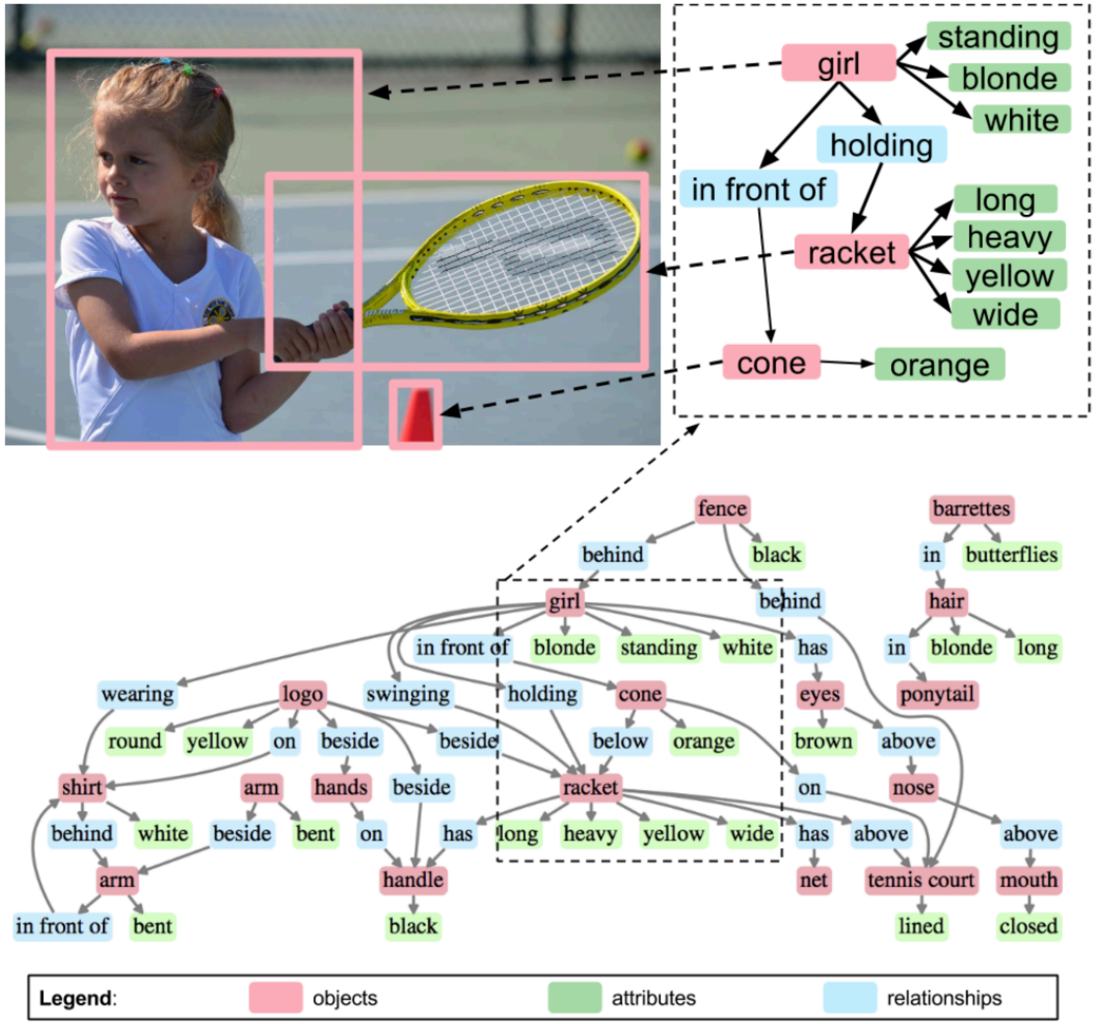
Scene Graph

Relational



Yang, Liao, Ackermann, and Rosenhahn, "On support relations and semantic scene graphs", ISPRS Journal of Photogrammetry and Remote Sensing, 2017

Objects + Relationships = Scene Graphs



108,077 Images

5.4 Million Region Descriptions

1.7 Million Visual Question Answers

3.8 Million Object Instances

2.8 Million Attributes

2.3 Million Relationships

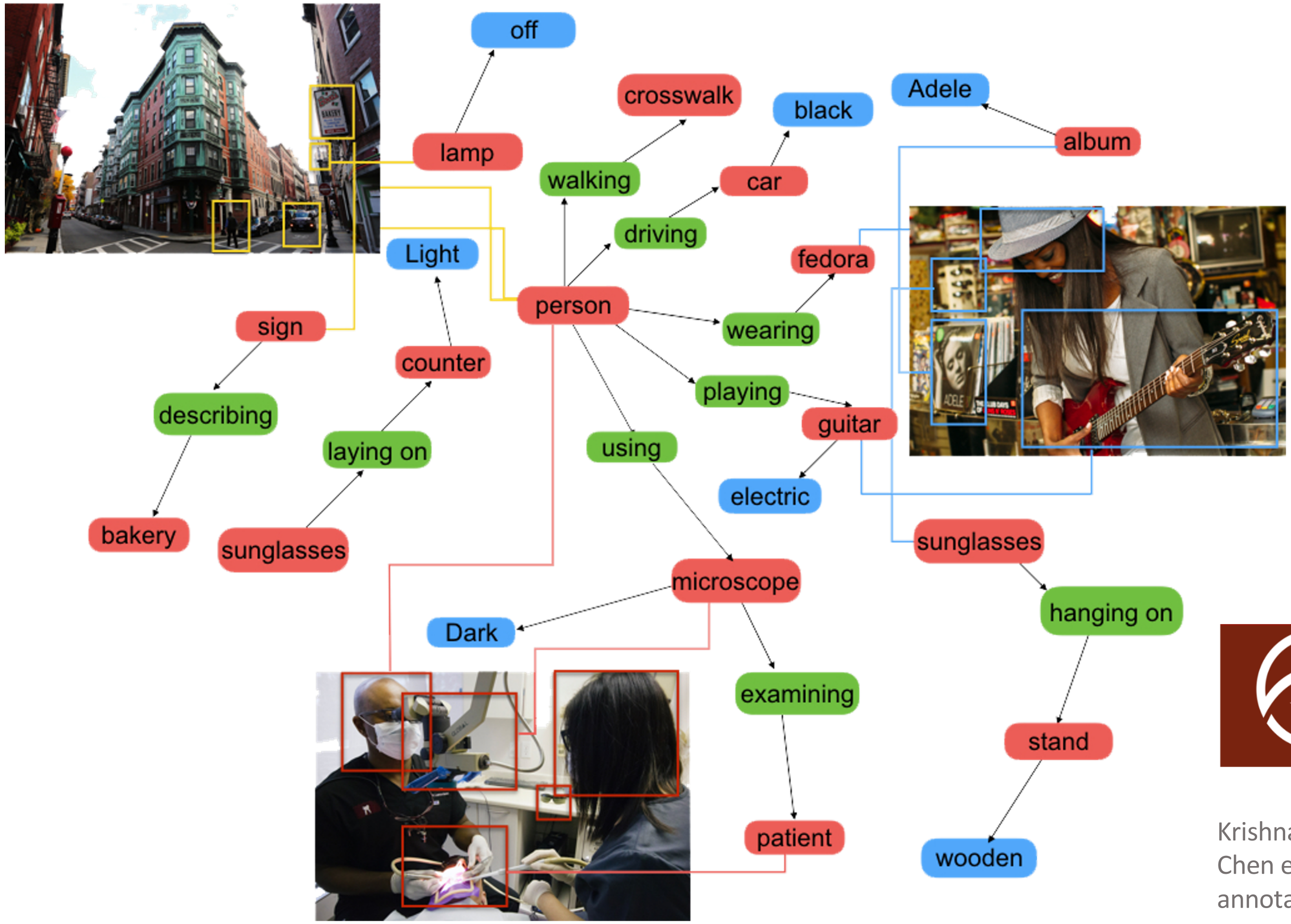
Everything Mapped to Wordnet Synsets

VISUALGENOME

Krishna, Ranjay, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen et al. "Visual genome: Connecting language and vision using crowdsourced dense image annotations." International Journal of Computer Vision 123, no. 1 (2017): 32-73.

Slide from Stanford CS231n [Johnson, Yeung, and Fei-Fei]

Objects + Relationships = Scene Graphs

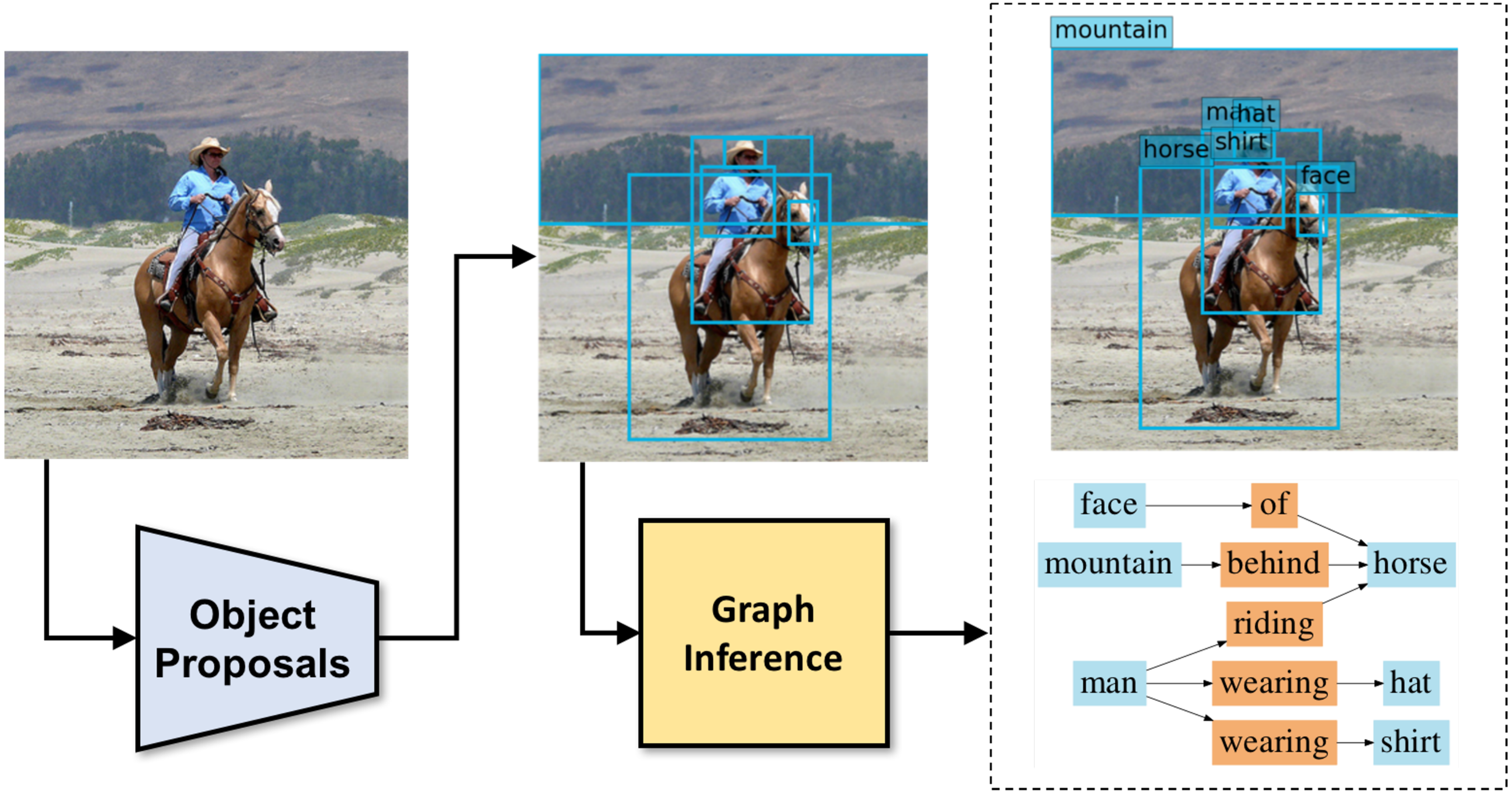


108,077 Images
5.4 Million Region Descriptions
1.7 Million Visual Question Answers
3.8 Million Object Instances
2.8 Million Attributes
2.3 Million Relationships
Everything Mapped to Wordnet Synsets

VISUALGENOME

Krishna, Ranjay, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen et al. "Visual genome: Connecting language and vision using crowdsourced dense image annotations." International Journal of Computer Vision 123, no. 1 (2017): 32-73.

Scene Graph Prediction



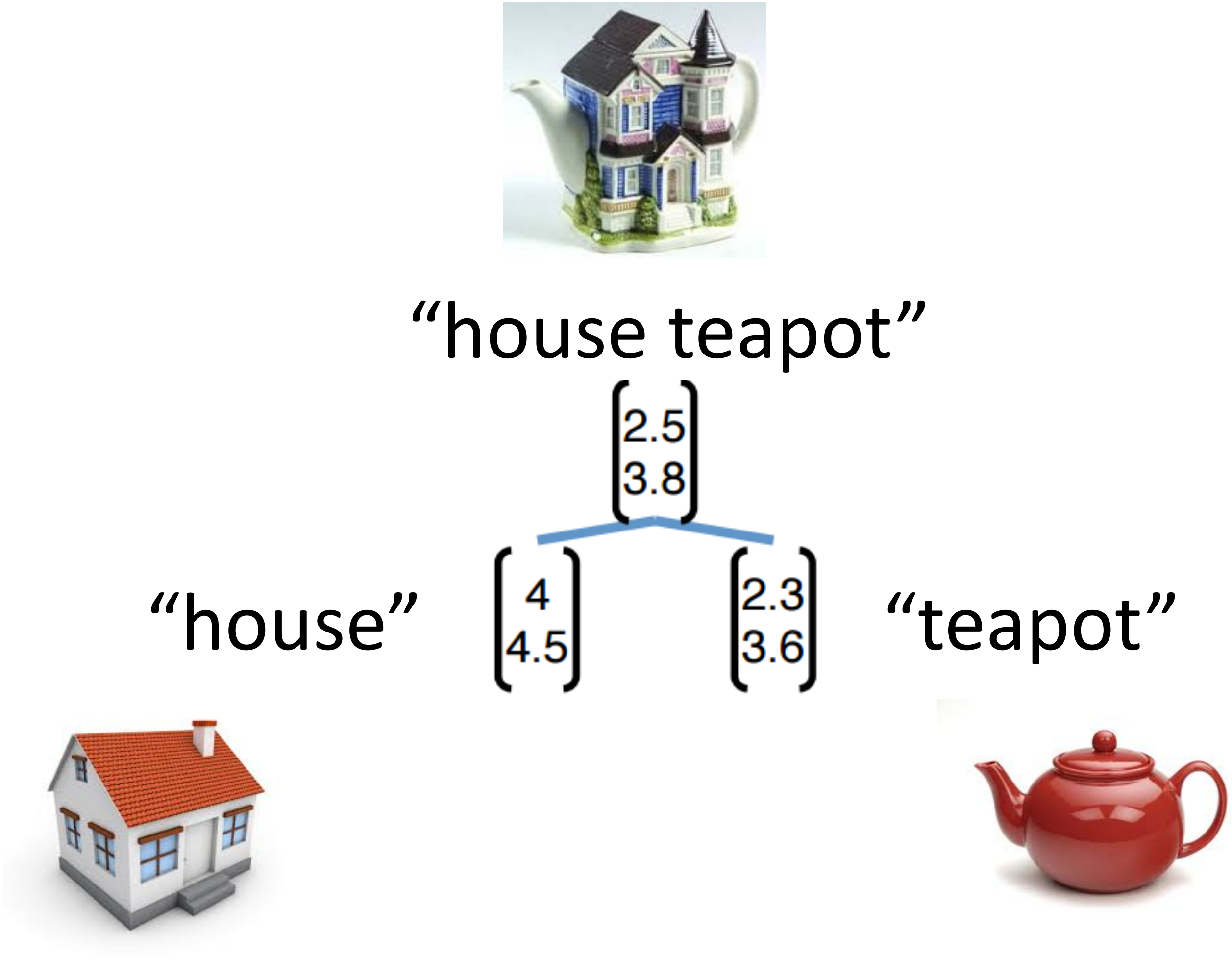
Xu, Zhu, Choy, and Fei-Fei, "Scene Graph Generation by Iterative Message Passing", CVPR 2017

Neural networks for structured representations and for structured reasoning

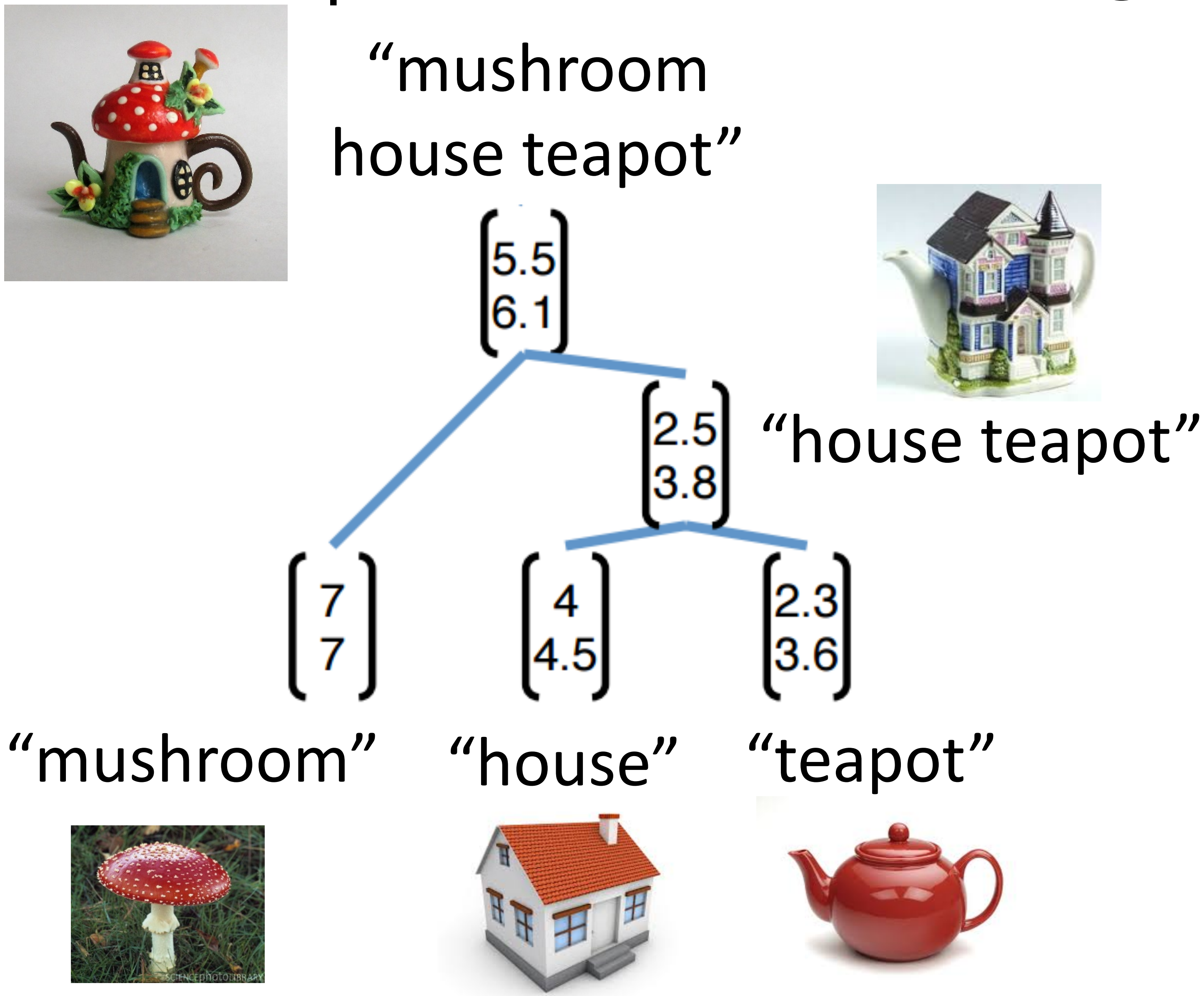
Structured neural models

- Two types of models for working with structured representations
  - Tree structure models
  - Graph neural networks

Compositional phrase embeddings

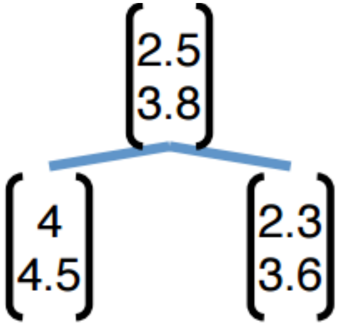


Compositional phrase embeddings



Compositional phrase embeddings

"house teapot"



"house"

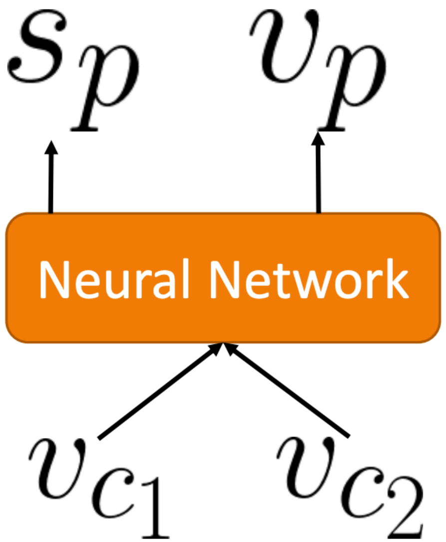
"teapot"

Score of two nodes combining

$$s_p = u^T v_p$$

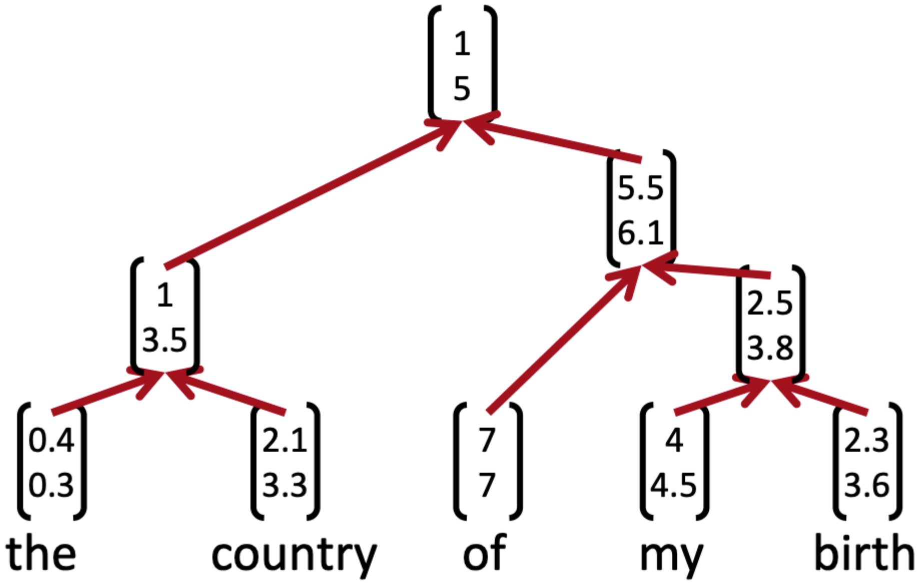
Embedding for parent node

$$v_p = \sigma(W \begin{bmatrix} v_{c1} \\ v_{c2} \end{bmatrix} + b)$$



Tied weights

Recursive



Recurrent

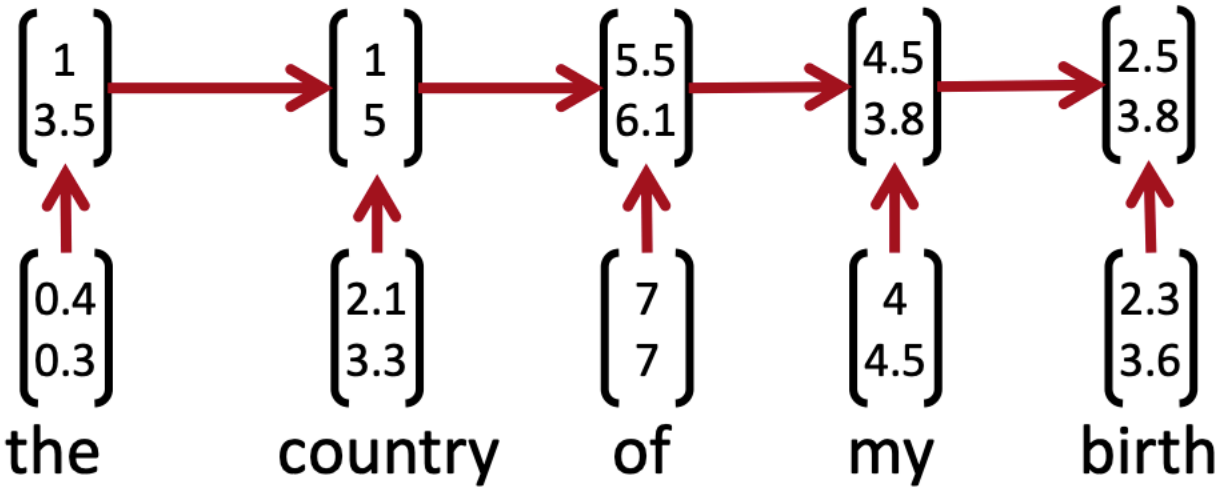
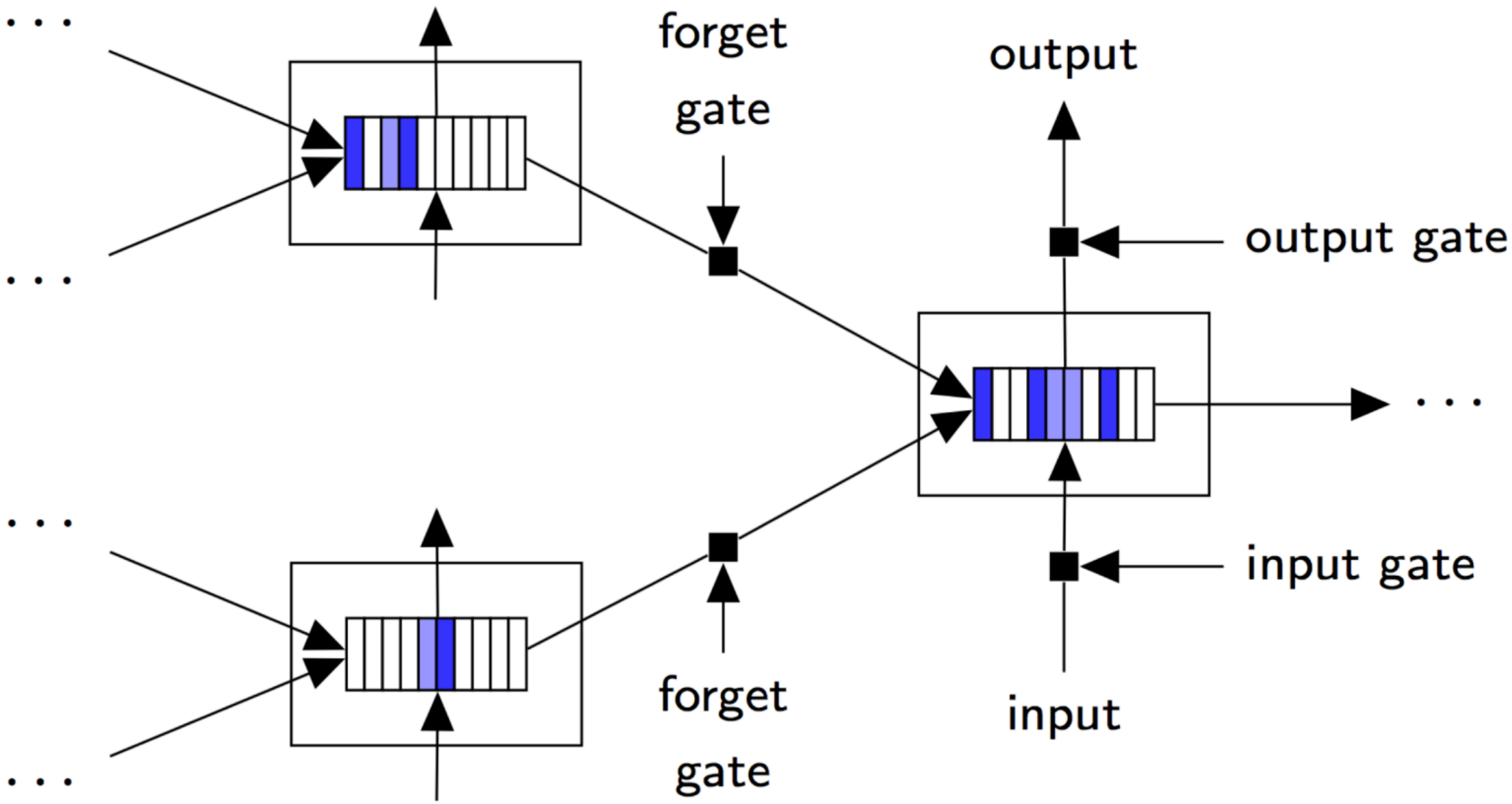


Image credit: Chris Manning

TreeRNNs

- Extend to a n-ary trees



Improved Semantic Representations From Tree-Structured Long Short-Term Memory Networks, Tai et al, ACL 2015

Inductive biases

- Assumptions to favor one set of solutions over another
- Structure priors

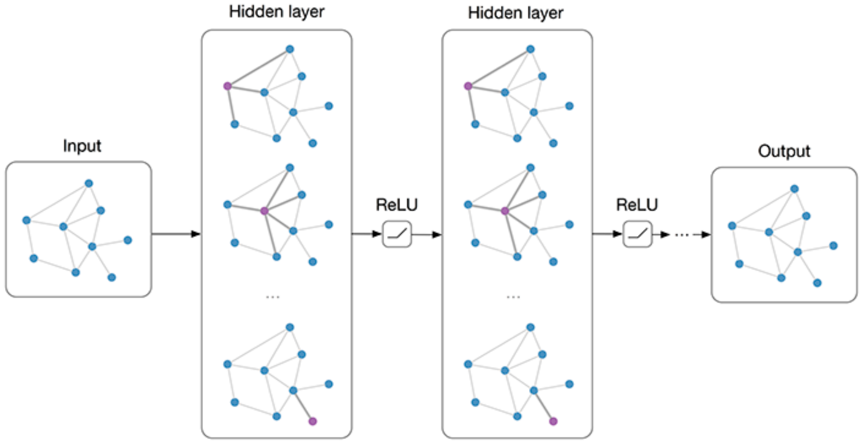
Component	Entities	Relations	Rel. inductive bias	Invariance
Fully connected	Units	All-to-all	Weak	-
Convolutional	Grid elements	Local	Locality	Spatial translation
Recurrent	Timesteps	Sequential	Sequentiality	Time translation
Graph network	Nodes	Edges	Arbitrary	Node, edge permutations

- These architecture constraints can help your network learn faster

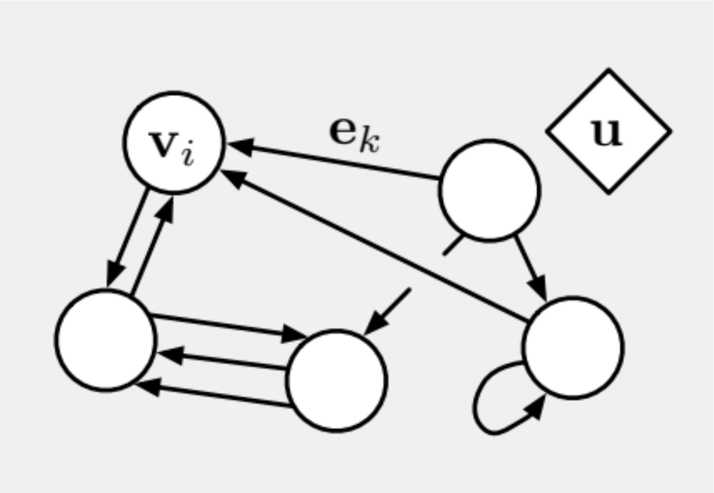
Table from Relational inductive biases, deep learning, and graph networks, Battaglia et al, arXiv 2018

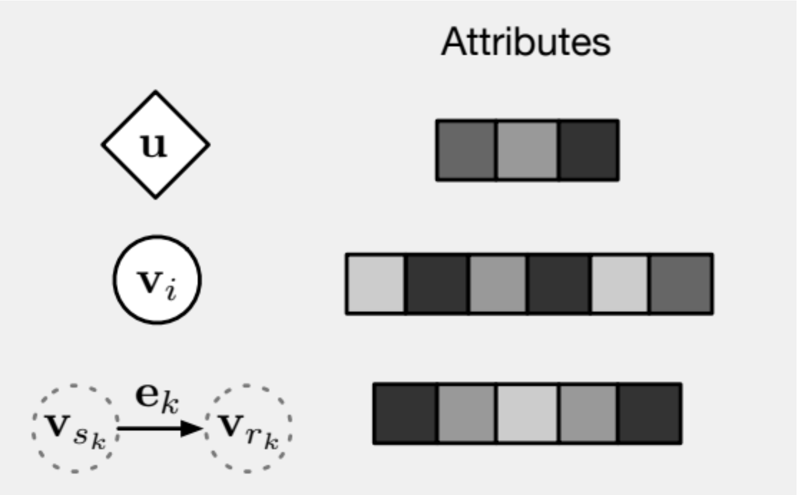
More general graph neural networks

GraphNNs



- Need to decide what will be nodes, edges
- Embeddings (attributes) for
  nodes v_i, edges e_k, entire graph u





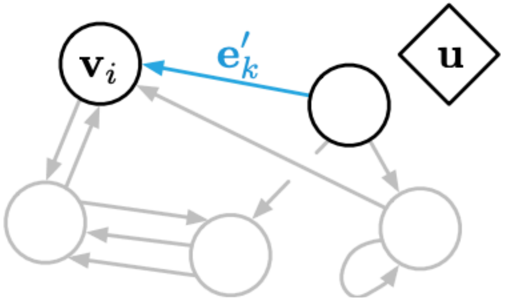
Relational inductive biases, deep learning, and graph networks, Battaglia et al, arXiv 2018

GraphNNs

- Embeddings (attributes) for nodes v_i, edges e_k, entire graph u
- Embeddings are iteratively updated

- Different architecture differ on what functions are used
- Use neural network for ϕ (shared weights) (MLP, CNN, RNN)
- Use sum / weighted average for ρ
- In some architectures, some components or inputs may be ignored

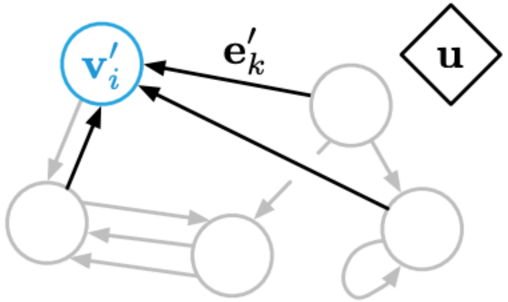
$$\mathbf{e}'_k = \phi^e(\mathbf{e}_k, \mathbf{v}_{r_k}, \mathbf{v}_{s_k}, \mathbf{u})$$



Update each edge e_k

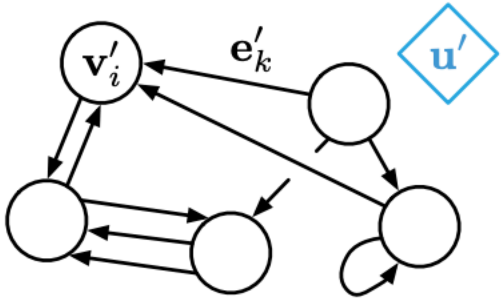
$$\bar{\mathbf{e}}'_i = \rho^{e \to v}(E'_i)$$

$$\mathbf{v}'_i = \phi^v(\bar{\mathbf{e}}'_i, \mathbf{v}_i, \mathbf{u})$$



Update each node v_i

$$\mathbf{u}' = \phi^u(\bar{\mathbf{e}}', \bar{\mathbf{v}}', \mathbf{u})$$



Update global graph u

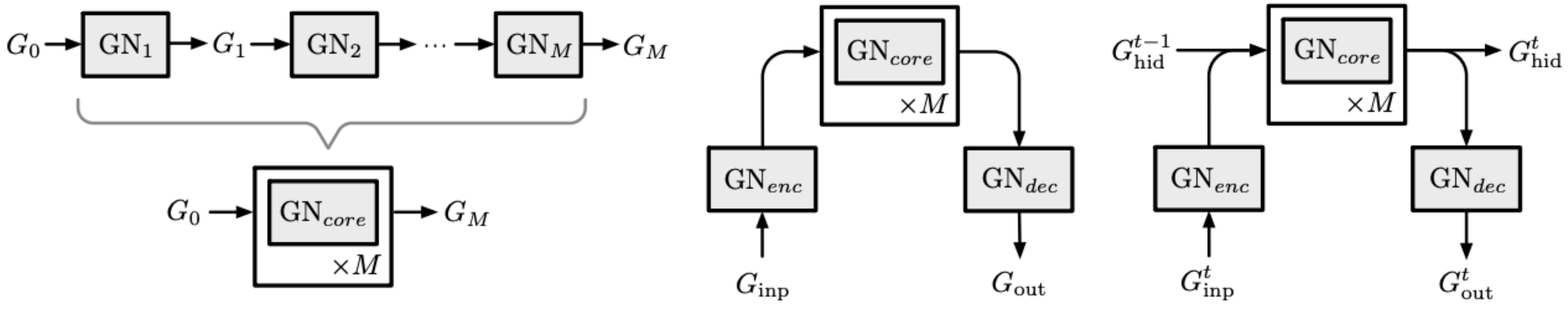
$$\bar{\mathbf{e}}' = \rho^{e \to u}(E')$$

$$\bar{\mathbf{v}}' = \rho^{v \to u}(V')$$

Relational inductive biases, deep learning, and graph networks, Battaglia et al, arXiv 2018

GraphNNs

- GN blocks can be composed



(a) Composition of GN blocks (b) Encode-process-decode (c) Recurrent GN architecture

Relational inductive biases, deep learning, and graph networks, Battaglia et al, arXiv 2018

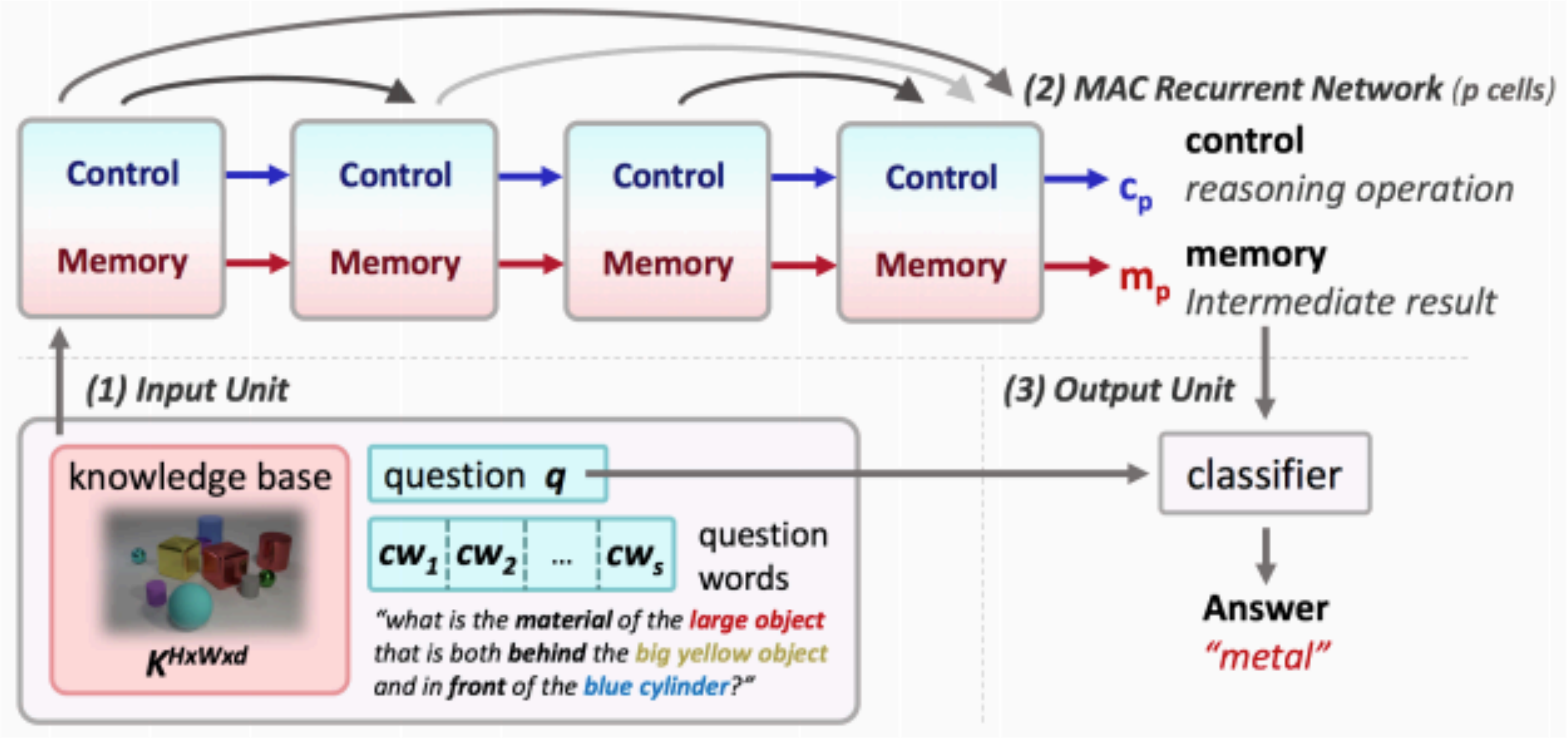
Code for working with GraphNNs

- https://github.com/deepmind/graph_nets
- https://pytorch-geometric.readthedocs.io

Structured reasoning

MAC (Memory, Attention, Control)

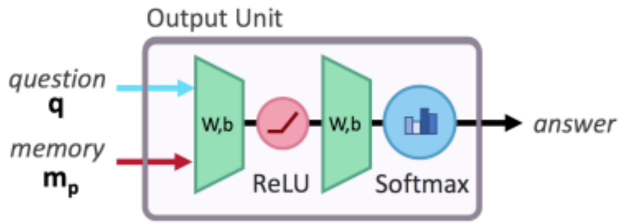
- Recurrent network with cell with read/write/control

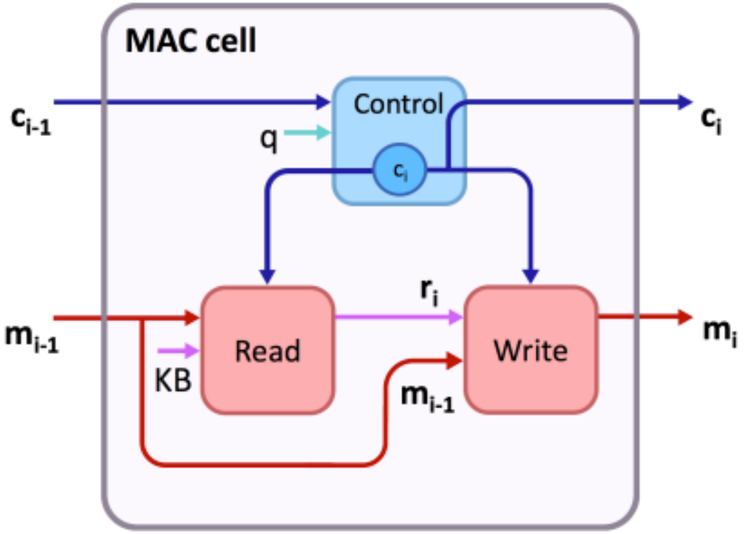


Compositional Attention Networks for Machine Reasoning, Hudson and Manning, ICLR 2018

MAC (Memory, Attention, Control)

- Recurrent network with cell with read/write/control
- Control – extract ``instruction'' from attention over query words
- Read – retrieves information from a knowledge base (image) given **current control** and **previous memory**
- Write – updates memory (combines old + new information)
- Fully differentiable

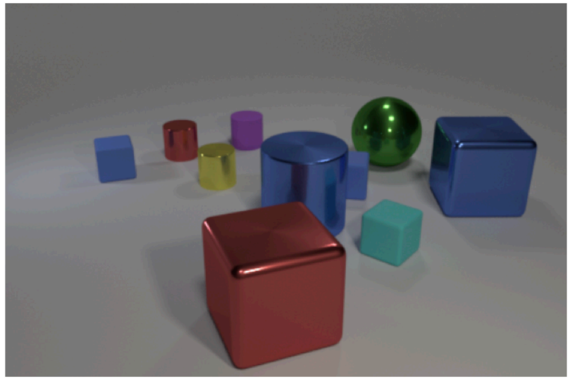




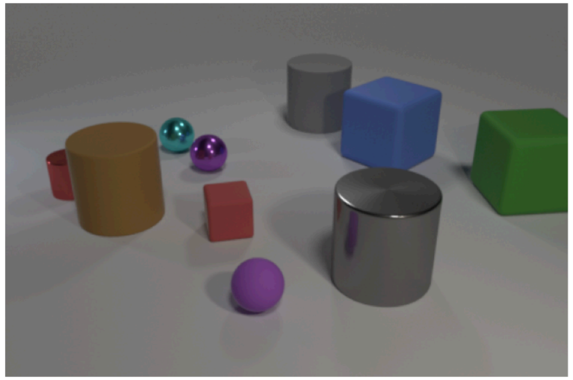
Compositional Attention Networks for Machine Reasoning, Hudson and Manning, ICLR 2018

Compositionality and reasoning (CLEVR dataset, Johnson et al, 2017)

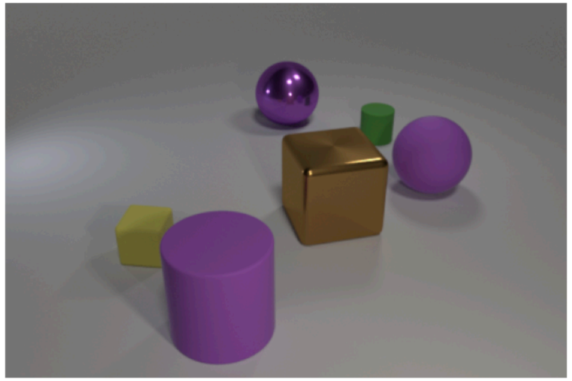
- Constructed by building functional program converted to natural language
- Small space of objects and attributes



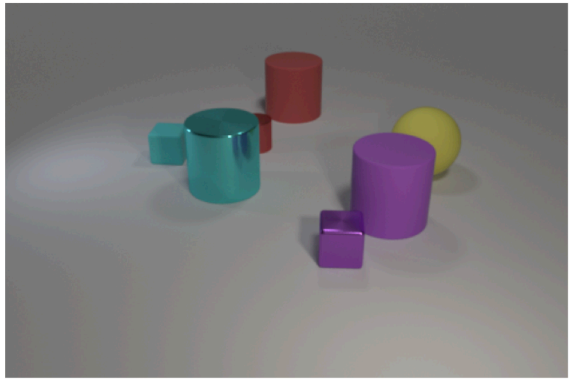
Q: What shape is the object reflected in the blue cylinder?
A: cube



Q: What number of cylinders share the same color?
A: 2

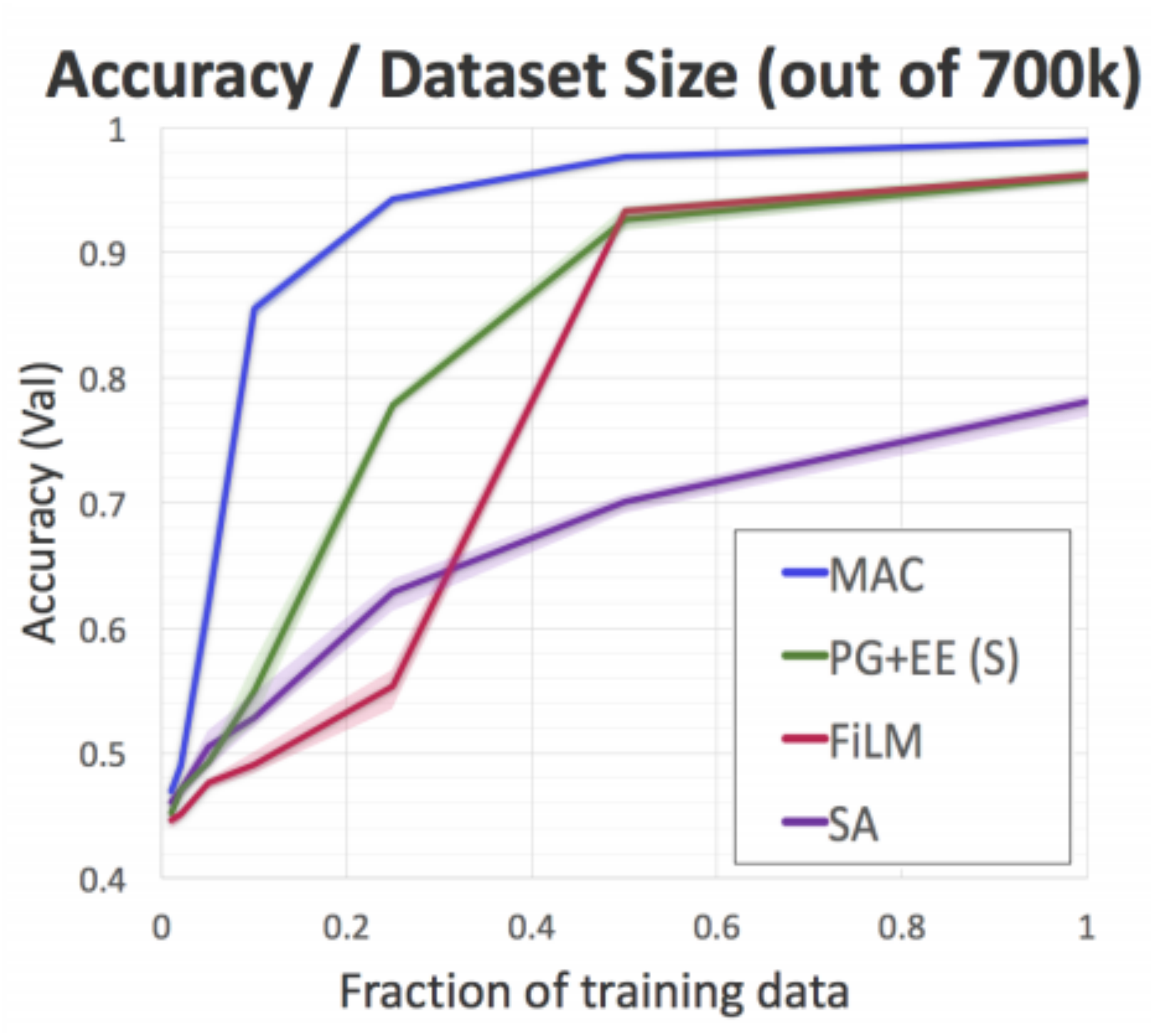


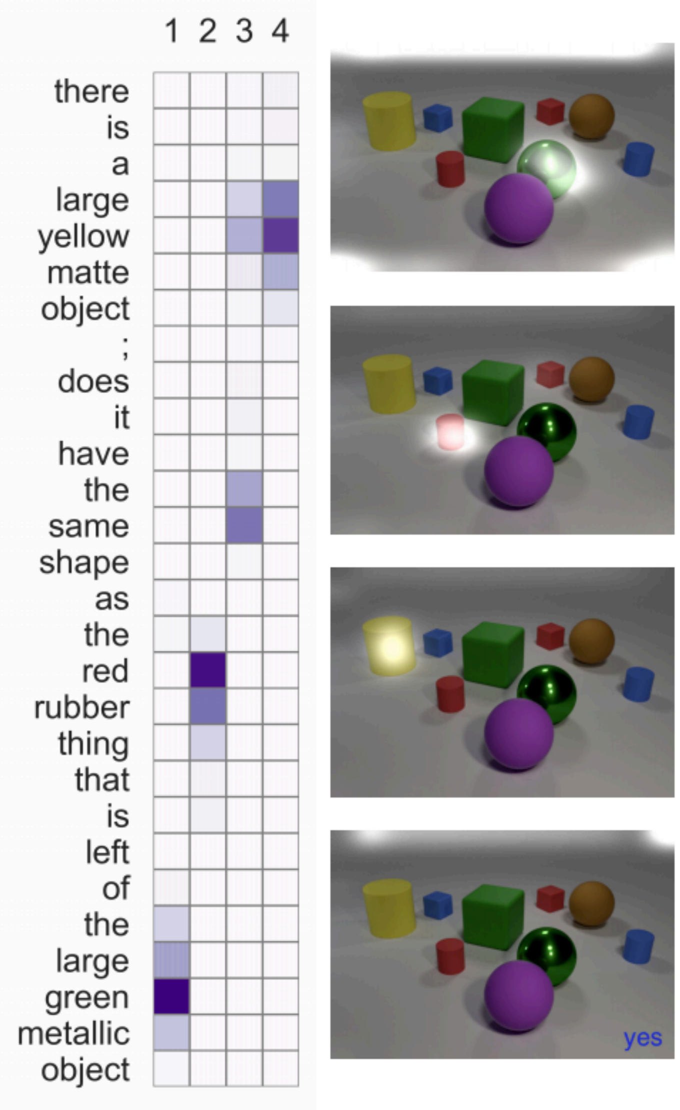
Q: How many objects are not purple and not metallic?
A: 2



Q: What color is the object partially blocked by the purple cylinder?
A: yellow

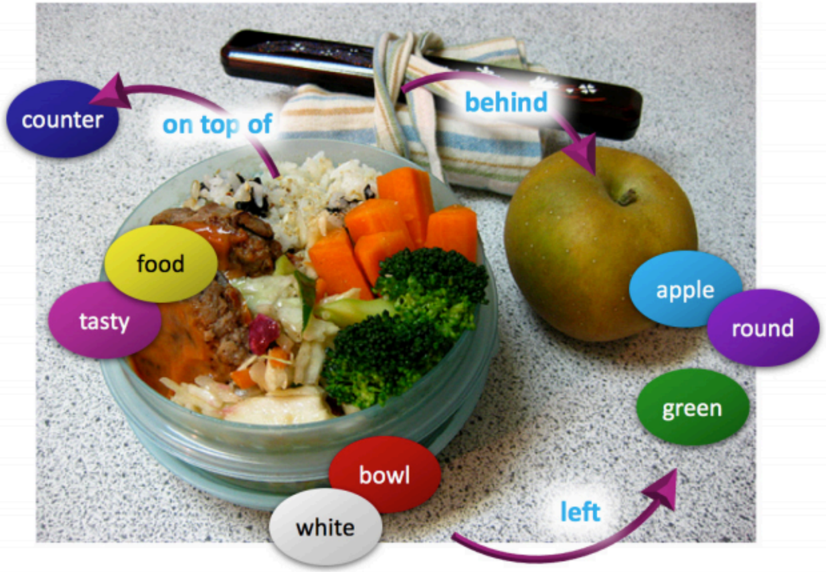
MAC can learn with smaller amount of data





Issues with real world VQA datasets

- Real world visual question benchmarks
- Strong biases
  - Language biases Can guess answer based on looking at picture)
  - Visual biases: focus on salient objects
- Unclear error sources
- Don’t need reasoning/compositionality
- Simple questions

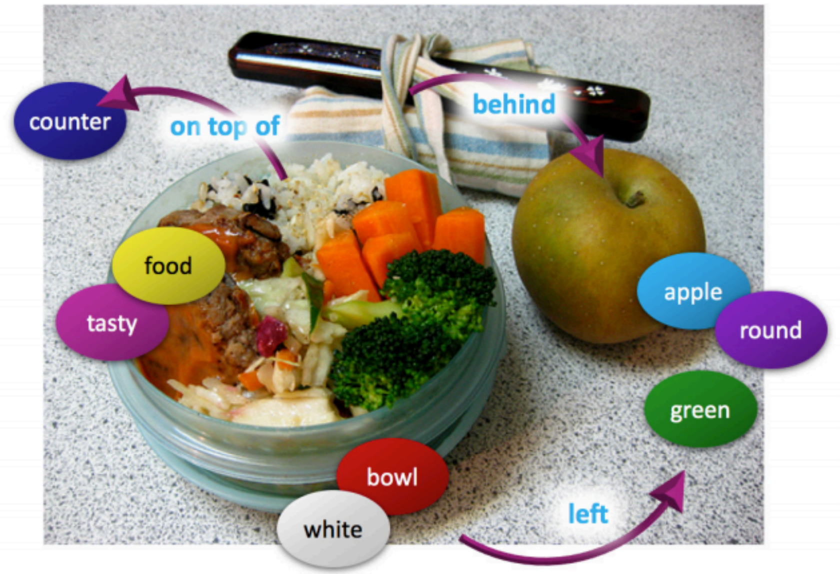


*Is the **bowl** to the right of the **green apple**?*
*What type of **fruit** in the image is **round**?*
*What color is the **fruit** on the right side, red or **green**?*
*Is there any **milk** in the **bowl** to the left of the **apple**?*

GQA: A New Dataset for Real-World Visual Reasoning and Compositional Question Answering
Hudson and Manning, CVPR 2019

GQA

- CLEVR on real images
- Generate questions in a compositional manner
- Start with scene-graph (Visual Genome)
    - Use segmentation
    - Resolve synonyms, use ontology
    - Generate questions in a controlled way
- Closely control answer distribution
- Multi-step question with large linguistic and visual variety
- Metrics that assess the model's ability in different ways

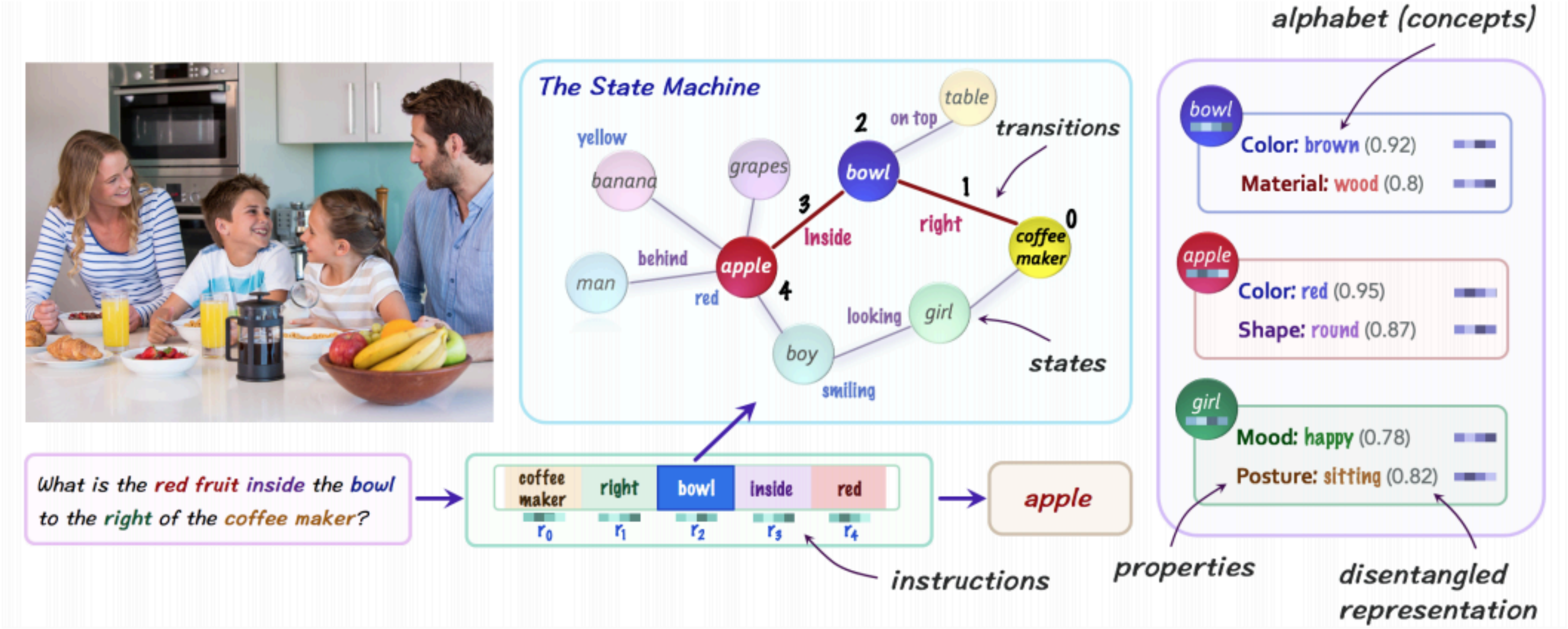


*Is the **bowl** to the right of the **green apple**?*
*What type of **fruit** in the image is **round**?*
*What color is the **fruit** on the right side, red or **green**?*
*Is there any **milk** in the **bowl** to the left of the **apple**?*

GQA: A New Dataset for Real-World Visual Reasoning and Compositional Question Answering
Hudson and Manning, CVPR 2019

Neural state machines

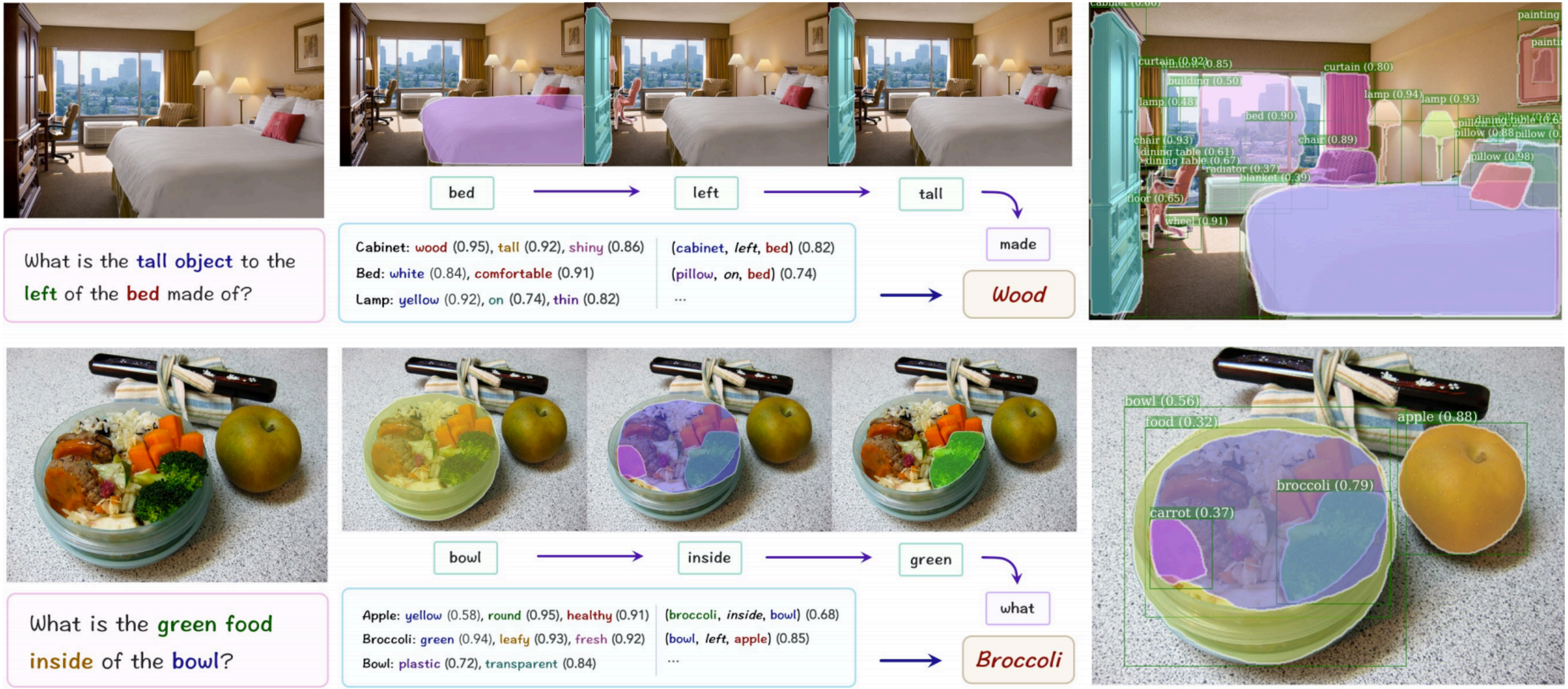
- Extract scene graph using Mask-RCNN + scene graph generation
- End to end differentiable model on graphs (after graph extraction)
- MAC with graphs



Learning by Abstraction: The Neural State Machine, Hudson and Manning, NeurIPS 2019

Neural state machines

- Uses learned concept embeddings for object category, attribute types (shape, color, material, etc), and relations.
- Construct graph with
    - Objects as nodes (states) with probabilities for each of the object category + attributes
      (computed from bounding box + visual features)
      Node embedding is weighted sum of concept embeddings $s^j = \sum_{c_k \in C_j} P_j(k) c_k$
    - Edges between objects capture the probability of each relation
      Edge embeddings is weighted sum of relation embeddings $e' = \sum_{c_k \in C_{L+1}} P_{L+1}(k) c_k$
    - Probability (attention) over states (objects)
- Question is converted into sequence of reasoning instructions
    - Run on the graph for a fixed number of steps
    - Each step will update the probabilities on the states (objects)
- Answer is obtained by putting a two-layer FCN softmax classifier on the question encoding and a vector with aggregated information from final object representations



Learning by Abstraction: The Neural State Machine, Hudson and Manning, NeurIPS 2019

NSM performance on GQA

Model	Binary	Open	Consistency	Validity	Plausibility	Distribution	Accuracy
Human [41]	91.20	87.40	98.40	98.90	97.20	-	89.30
Global Prior [41]	42.94	16.62	51.69	88.86	74.81	93.08	28.90
Local Prior [41]	47.90	16.66	54.04	84.33	84.31	13.98	31.24
Language [41]	61.90	22.69	68.68	96.39	**87.30**	17.93	41.07
Vision [41]	36.05	1.74	62.40	35.78	34.84	19.99	17.82
Lang+Vis [41]	63.26	31.80	74.57	96.02	84.25	7.46	46.55
BottomUp [5]	66.64	34.83	78.71	96.18	84.57	5.98	49.74
MAC [40]	71.23	38.91	81.59	96.16	84.48	5.34	54.06
SK T-Brain*	77.42	43.10	90.78	96.26	85.27	7.54	59.19
PVR*	77.69	43.01	90.35	**96.45**	84.53	5.80	59.27
GRN	77.53	43.35	88.63	96.18	84.71	6.06	59.37
Dream	77.84	43.72	91.71	96.38	85.48	8.40	59.72
LXRT	77.76	44.97	92.84	96.30	85.19	8.31	60.34
NSM	**78.94**	**49.25**	**93.25**	**96.41**	84.28	**3.71**	**63.17**

Next time (after the break)

- Paper presentations
  - Grounded Compositional Semantics for Finding and Describing Images with Sentences (RvNNs with Ali Arab?)
  - Learning to Represent Image and Text with Denotation Graph (Atmika)

- Project proposal

- Thursday (2/25): Semantic Parsing (language to programs)