

CMPT 983

Grounded Natural Language Understanding

February 16th, 2022

Compositionality and Structure

Today

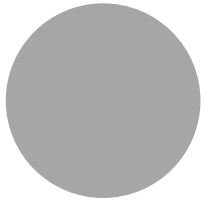
- Compositionality
- Structured representations
- Structured reasoning

Compositionality

Compositional Generalization

Grounding

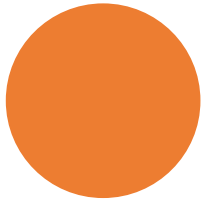
osk



vap



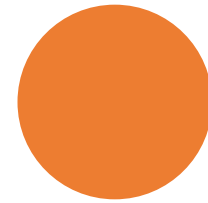
tod



bo



Compositionality



osk tod

Generalization

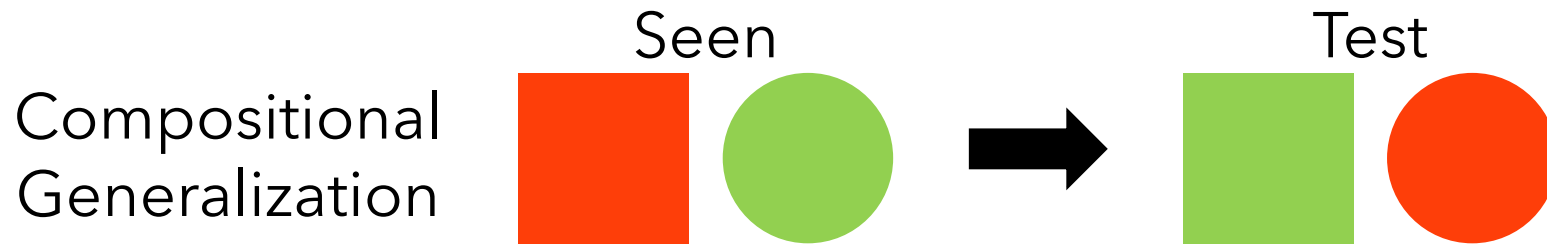
unseen combination



vap bo

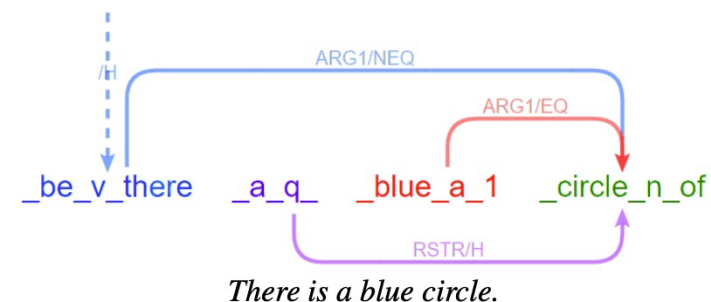
Studying compositionality

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

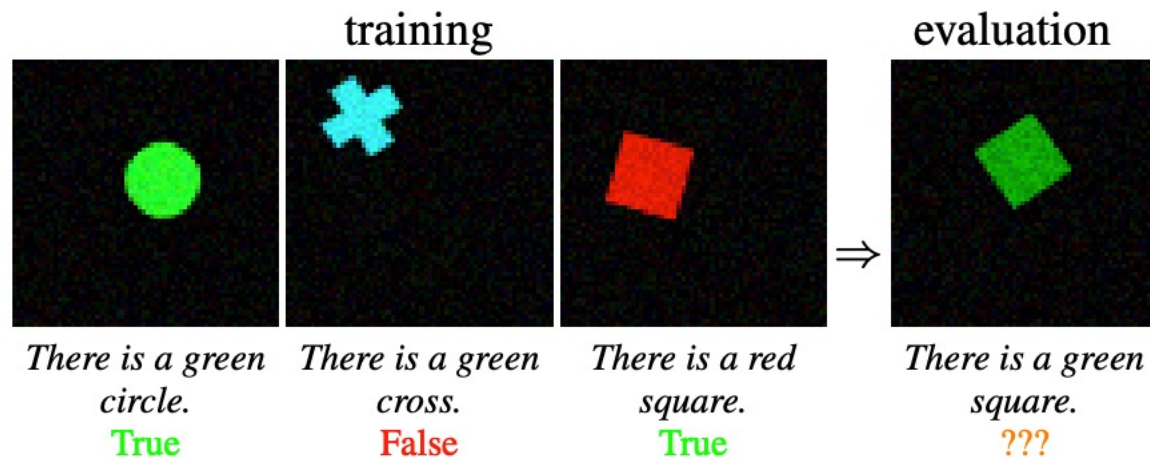


ShapeWorld

- 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



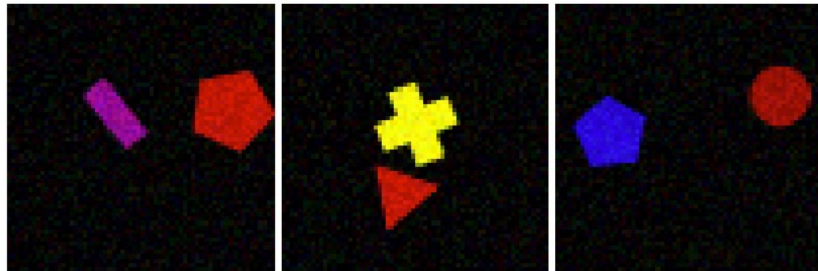
ONESHape



ShapeWorld – 4 datasets

SPATIAL

training



An ellipse is to the left of a red pentagon.

False

A red triangle is below a cross.

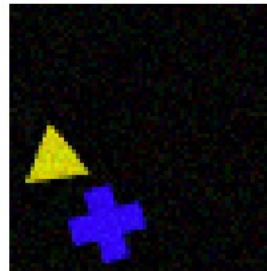
True

A blue shape is to the left of a circle.

True



evaluation

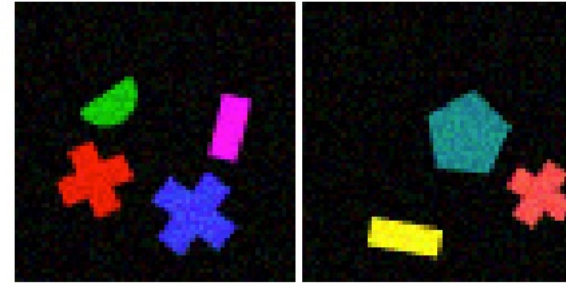


A triangle is below a blue cross.

???

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

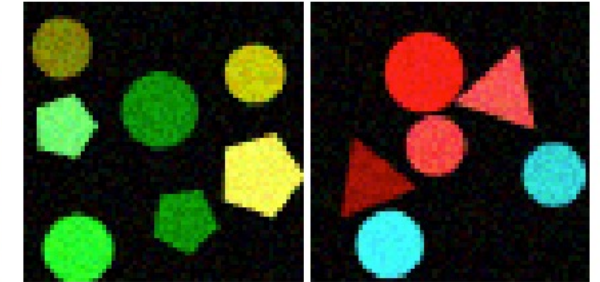
MULTISHAPE



- *There is a magenta semicircle.*
- *There is a pentagon.*
- *There is a cyan shape.*

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

QUANTIFICATION



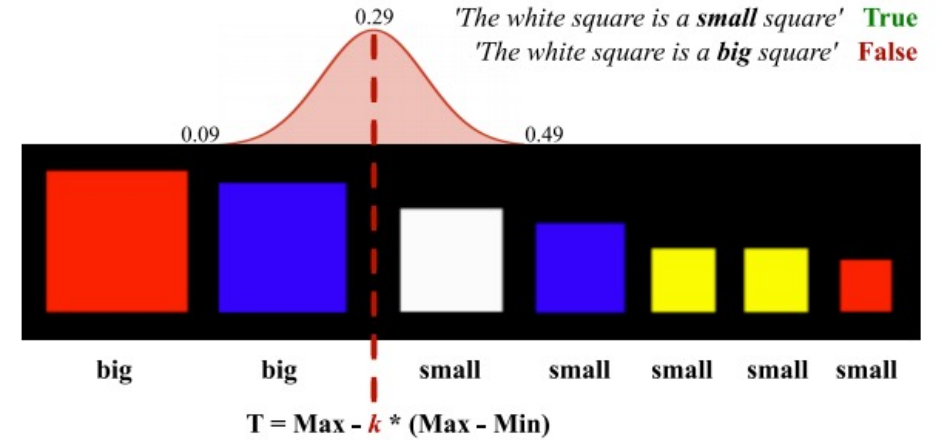
- *The shape is green.*
- *Most shapes are rectangles.*
- *No shape is a red triangle.*
- *All triangles are green.*
- *Two blue shapes are pentagons.*

ShapeWorld Results

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 <i>no</i>	52 / 51 / 48	61 / 60 / 61	56 / 56 / 51	55 / 55 / 58
instances with <i>the</i> (=1)	53 / 58 / 61	55 / 59 / 58	59 / 59 / 55	63 / 63 / 63
instances with <i>a</i> (≥ 1)	34 / 35 / 36	34 / 36 / 37	49 / 50 / 51	48 / 52 / 50
instances with <i>two</i> (≥ 2)	53 / 48 / 48	50 / 50 / 49	70 / 69 / 62	72 / 67 / 58
instances with <i>most</i>	49 / 50 / 49	48 / 48 / 49	69 / 68 / 60	60 / 52 / 51
instances with <i>all</i>	52 / 54 / 50	48 / 50 / 51	47 / 52 / 51	49 / 50 / 51

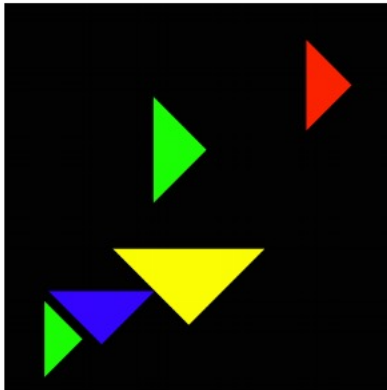
MaleVic

- Size understanding
- Programmatically determine big / small using thresholds



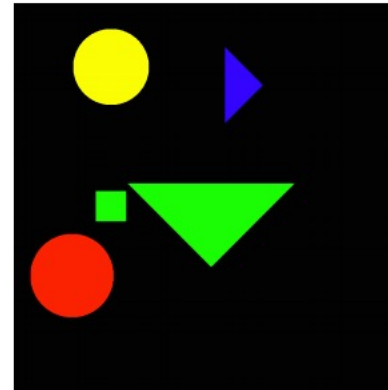
Superlative

The yellow triangle is the **biggest** triangle.



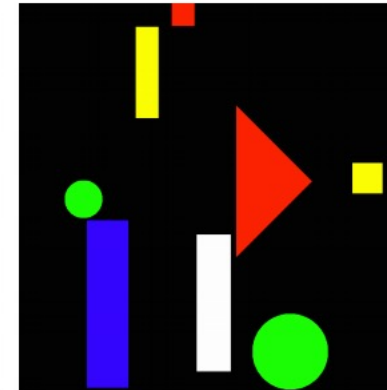
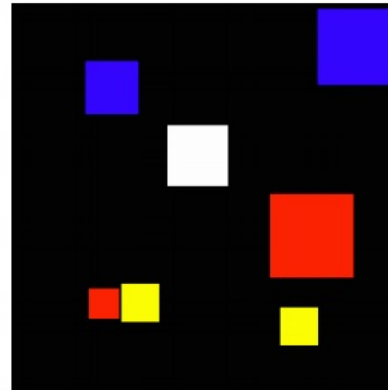
Any shape

The red circle is a **big** object.



Same shape

The white square is a **small** square.

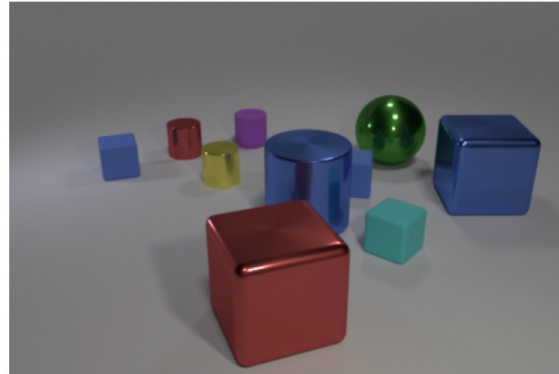


Pick shape from different shapes

The white rectangle is a **big** rectangle.

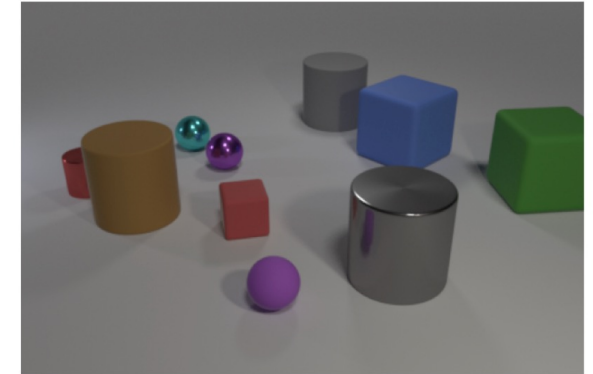
CLEVR: Compositionality and reasoning

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



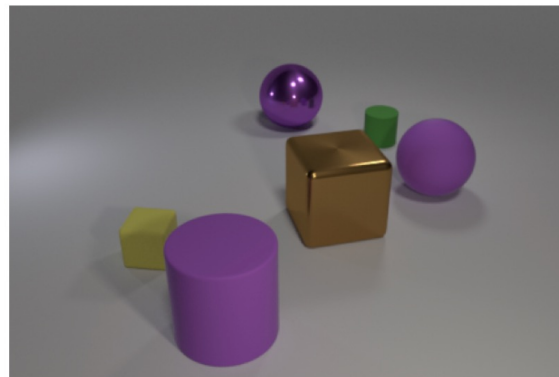
Q: What shape is the object reflected in the blue cylinder?

A: cube



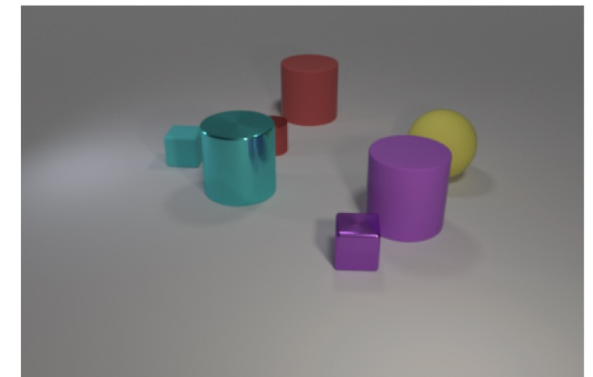
Q: What number of cylinders share the same color?

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

Compositionality with actions

Generate worlds and language

Command: walk to a yellow small cylinder
 Meaning: walk to a yellow small cylinder
 Target: turn left turn left walk walk walk walk turn left walk walk walk

ROOT → VP

VP → VP RB

VP → VV_i 'to' DP

VP → VV_t DP

DP → 'a' NP

NP → JJ NP

NP → NN

VV_i → {walk}

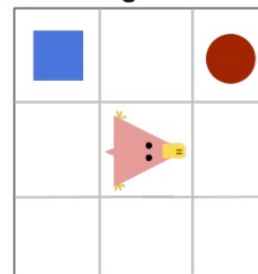
VV_t → {push, pull}

RB → {while spinning, while zigzagging, hesitantly, cautiously}

NN → {circle, square, cylinder}

JJ → {red, green, blue, big, small}

Training

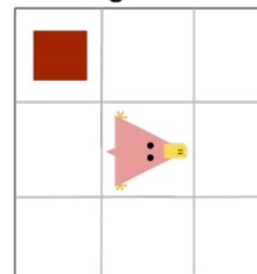


"Walk to the **blue square**."

"Walk to the **red circle**."

"Push the **red circle**."

Testing

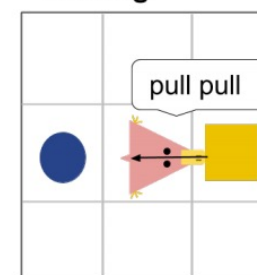


"Walk to the **red square**."

"Push the **red square**."

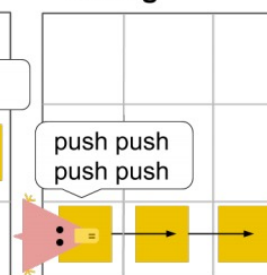
"Pull the **small red square**."

Training



"Pull the square."

Testing



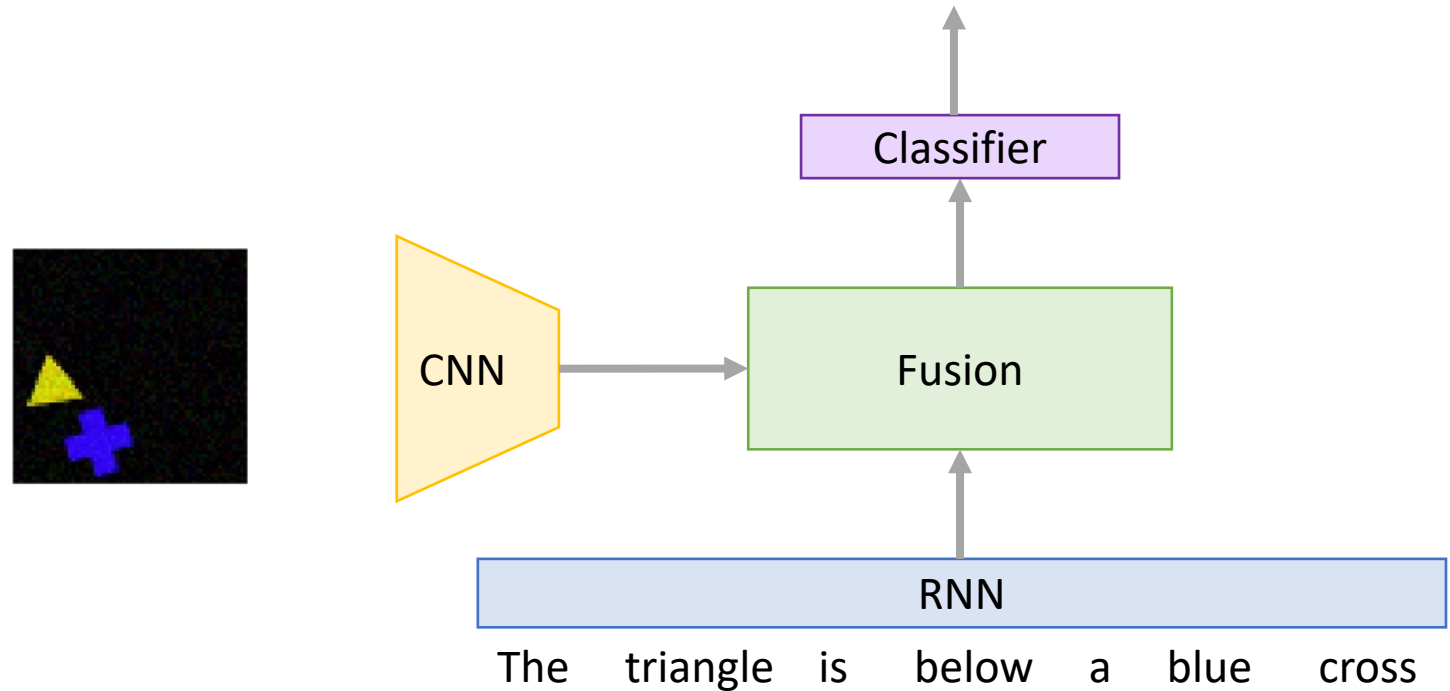
"Push the square."



How to achieve compositionality?

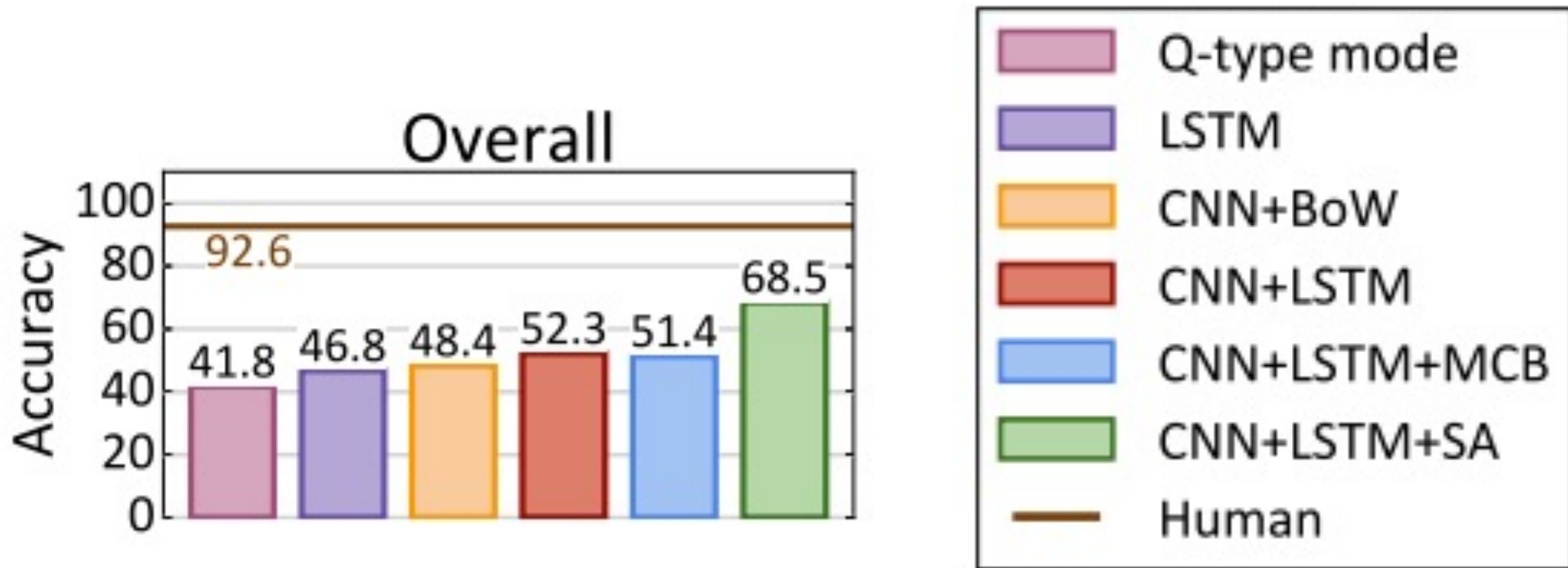
Baseline networks

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



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

CLEVR baseline performance

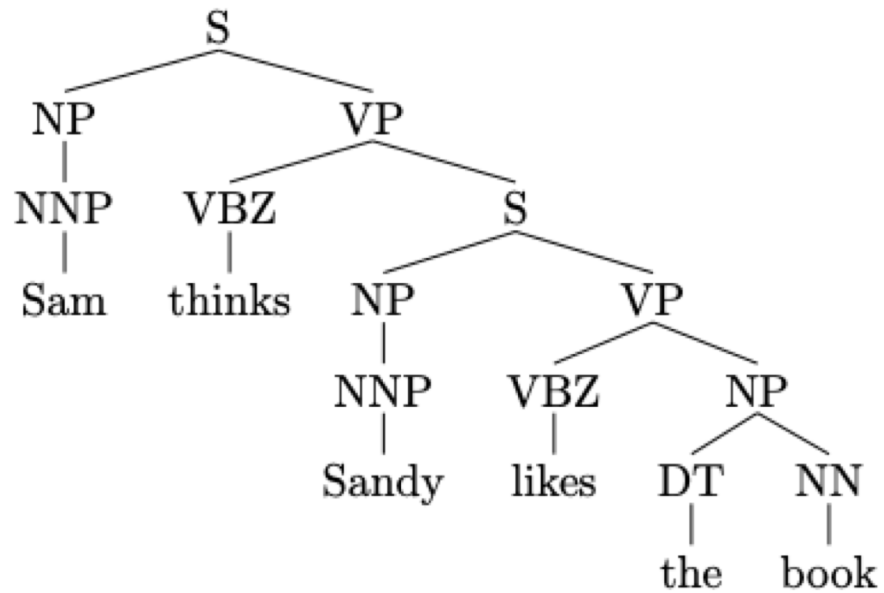


Structured representations

Structured representation of sentences

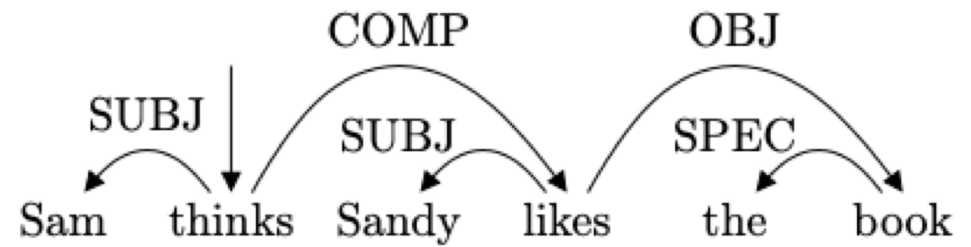
Constituency Parse Tree

Hierarchical



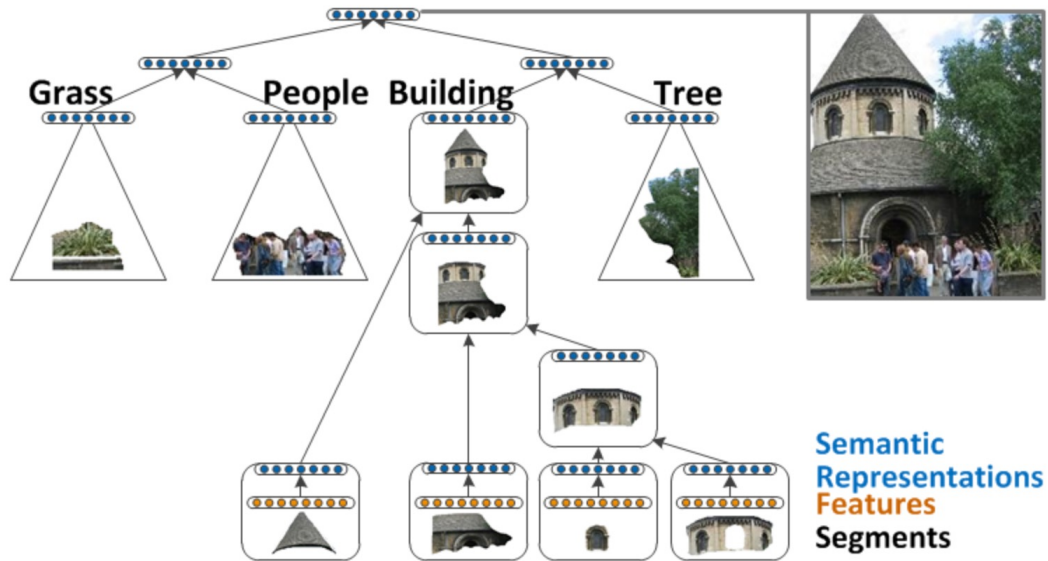
Dependency Parse

Relational

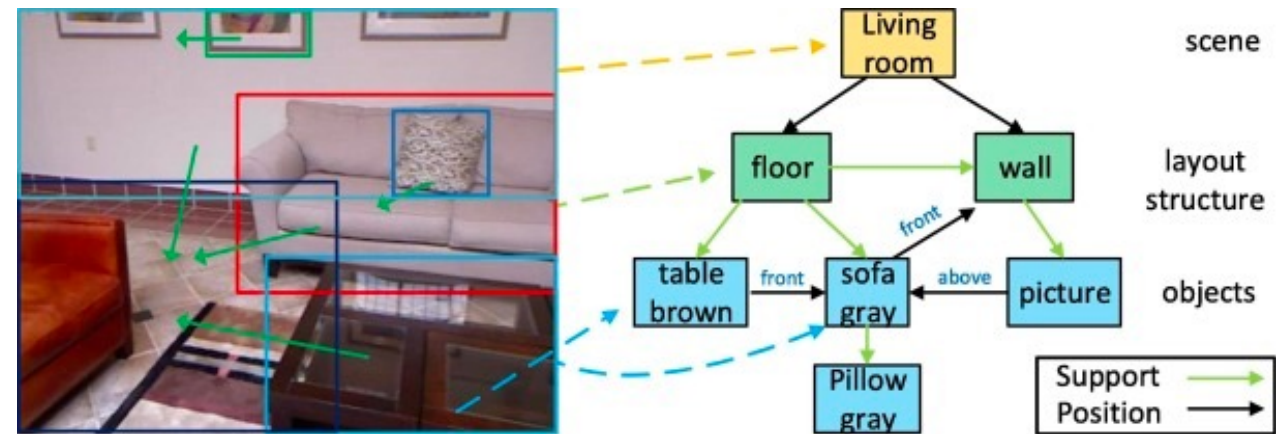


Structured representation of images

Scene Parse Tree Hierarchical



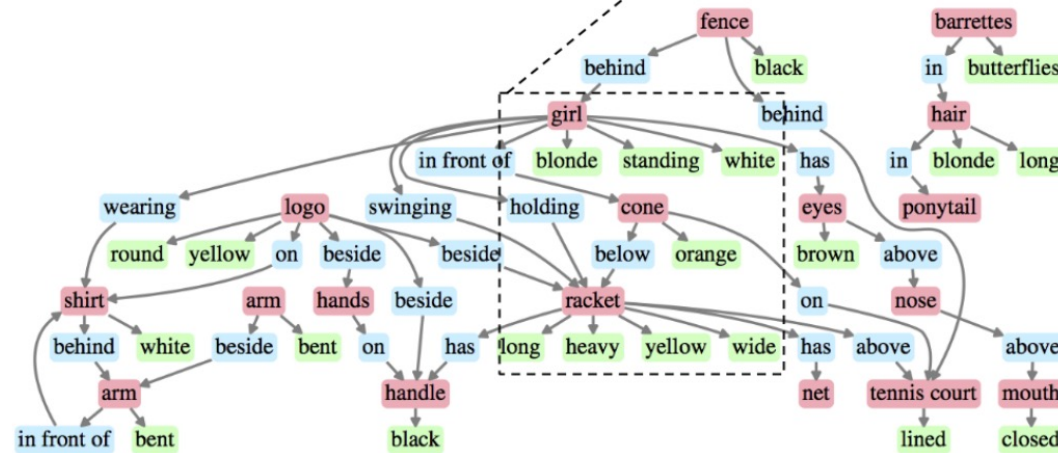
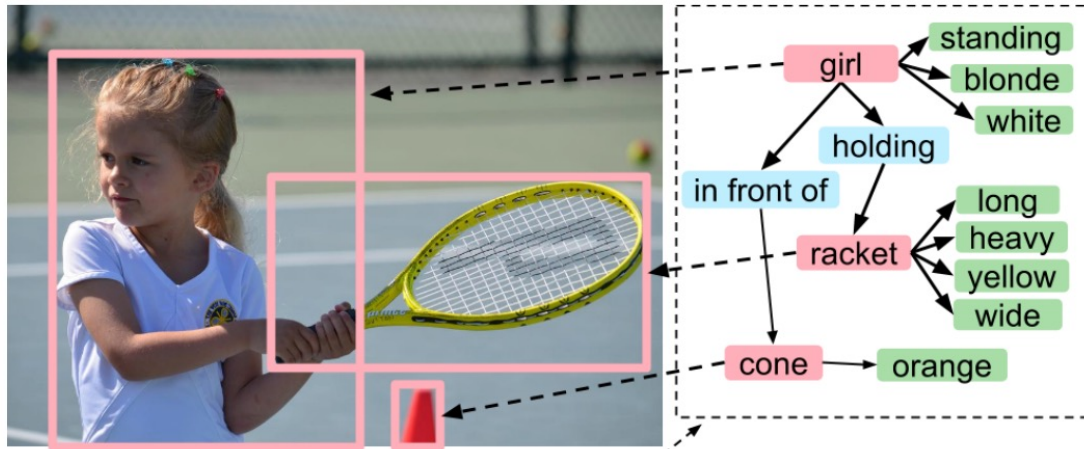
Scene Graph Relational



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

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

Objects + Relationships = Scene Graphs



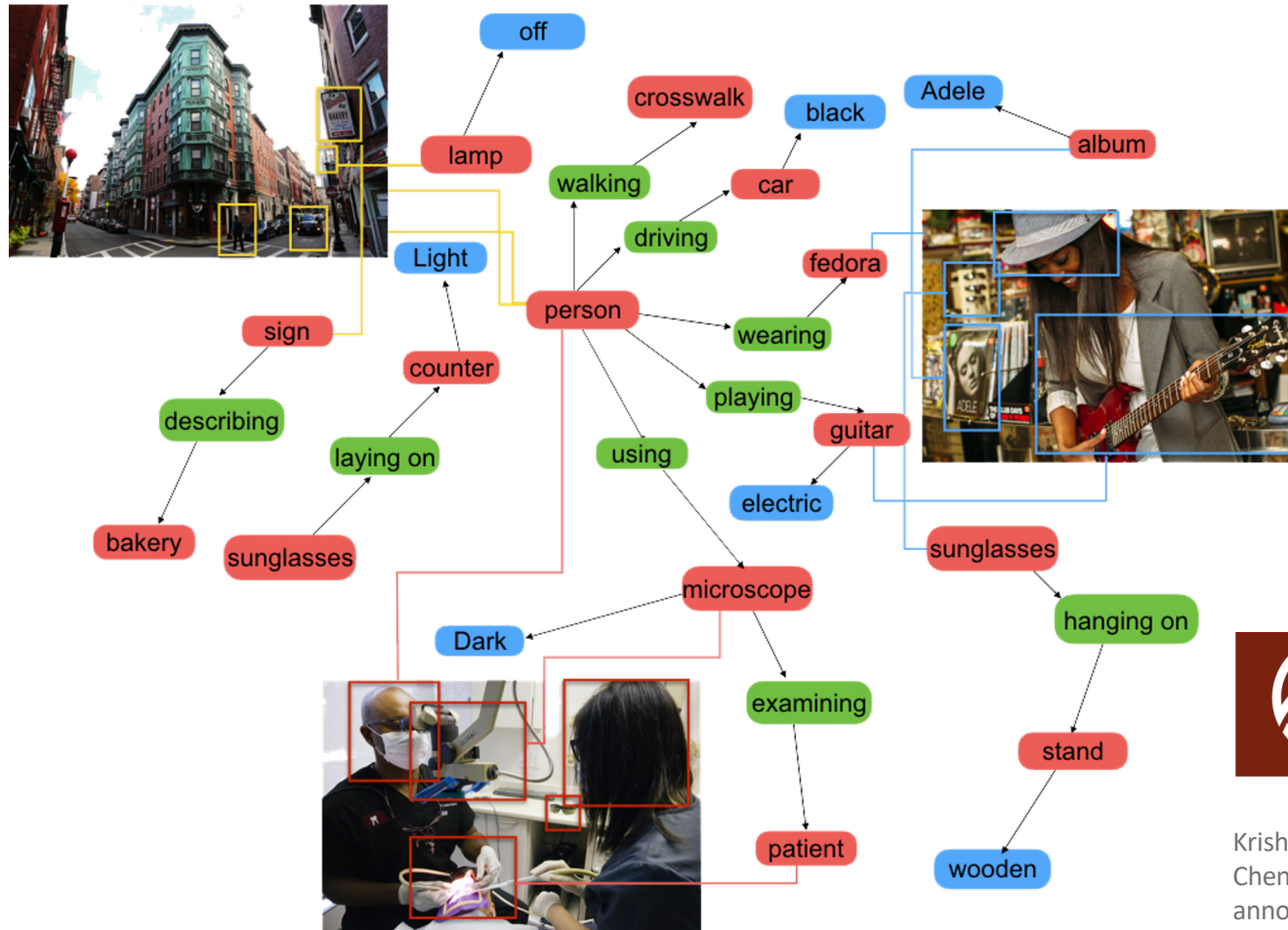
Legend: objects attributes relationships

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



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.

Objects + Relationships = Scene Graphs

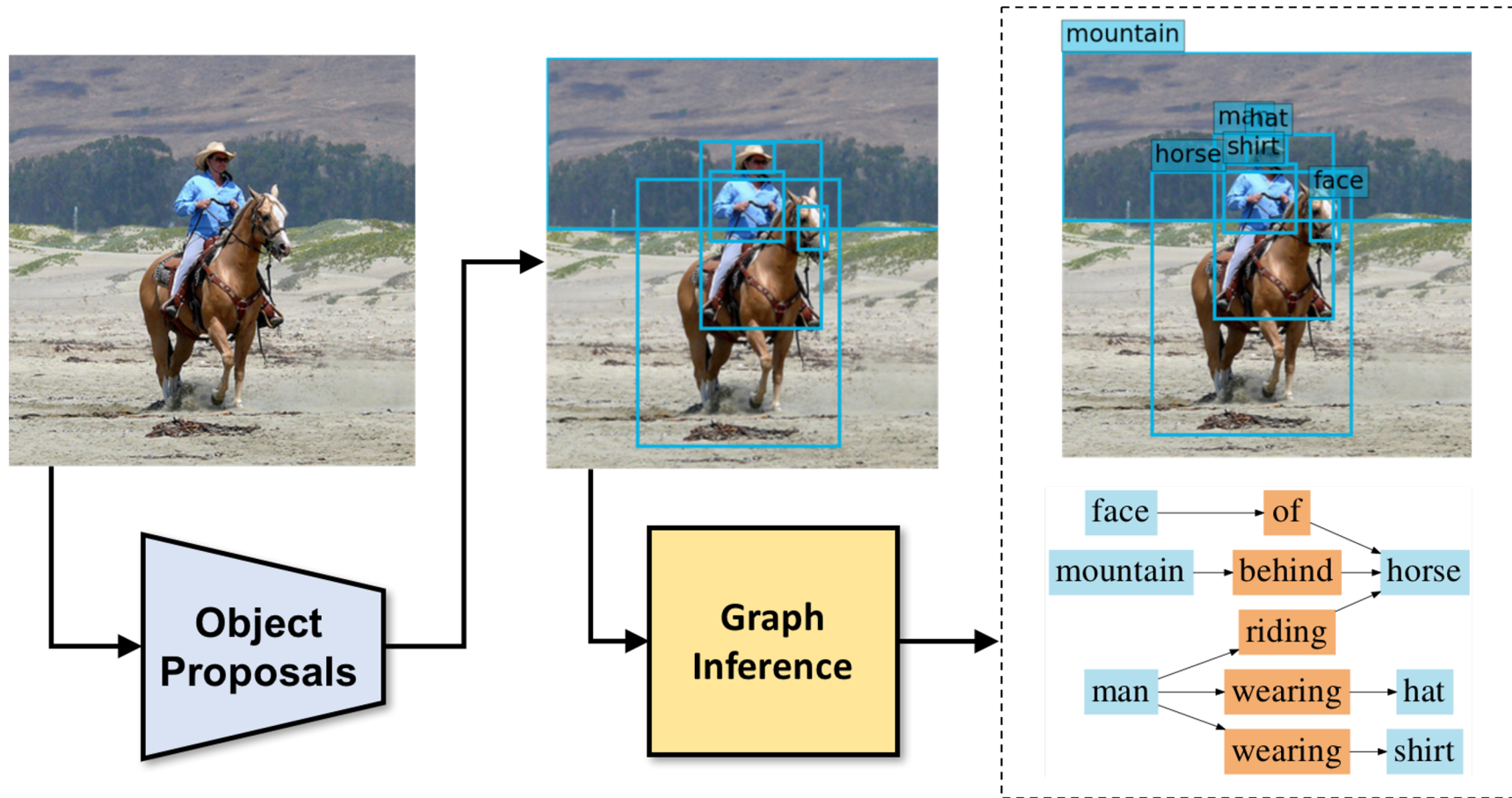


108,077 Images
5.4 Million Region Descriptions
1.7 Million Visual Question Answers
3.8 Million Object Instances
2.8 Million Attributes
2.3 Million Relationships
Everything Mapped to Wordnet Synsets



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
Figure copyright IEEE, 2018. Reproduced for educational purposes.

Neural networks for
structured representations
and for structured reasoning

Structured predictions

- Can use neural networks for structured prediction

- Treat as sequence to sequence problem

John has a dog . → (S (NP NNP)_{NP} (VP VBZ (NP DT NN)_{NP})_{VP} .)_S

- Predict actions to take to build the structure incrementally
 - Use neural models to score local compatibility and then perform search to find the structure with the highest overall score

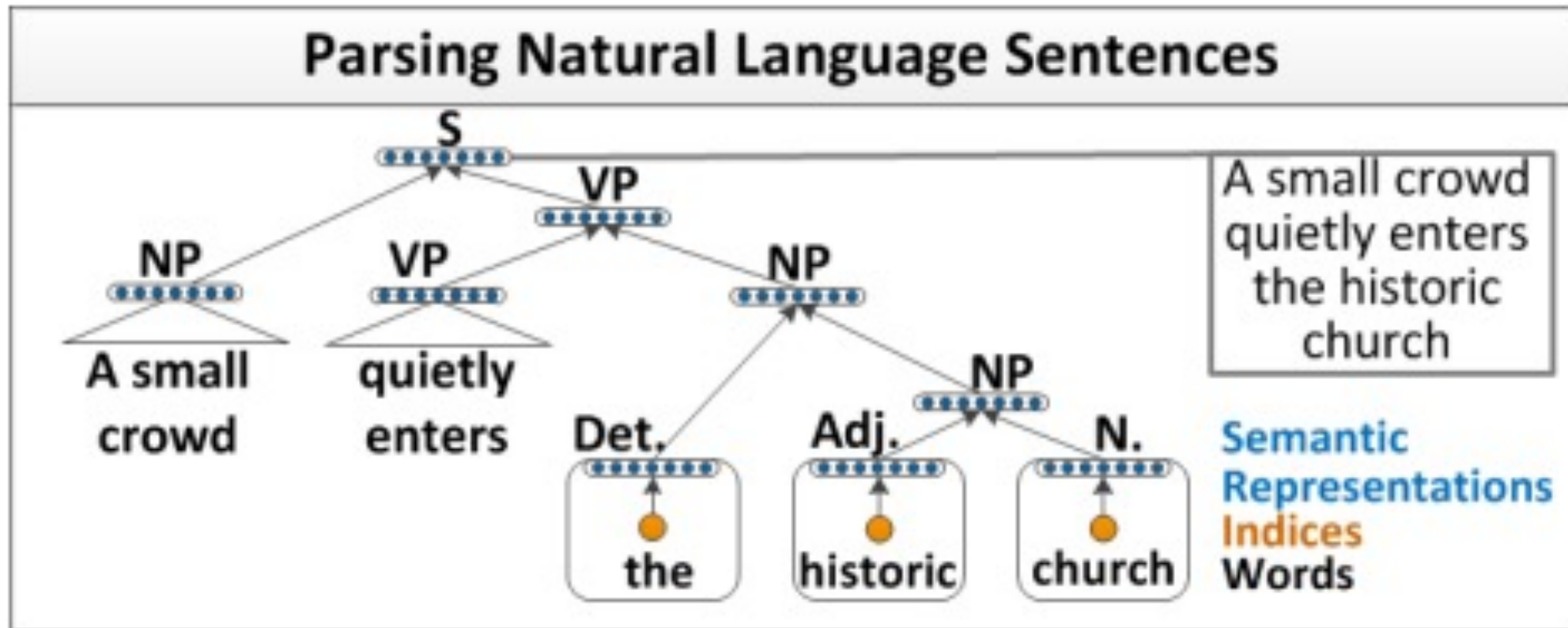
Structured neural models

- Models for working with structured representations
 - Tree structure models
 - Graph neural networks

Tree-structured models

Recursive NNs

- Recursive model for building up compositional representations



Compositional phrase embeddings

Train to increase scores of segment pairs with good

Score of two nodes combining

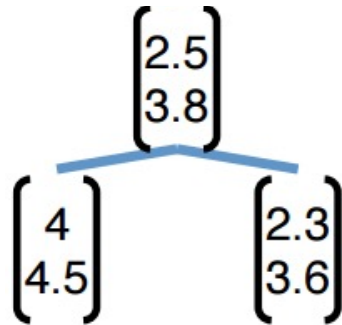
Embedding for parent node



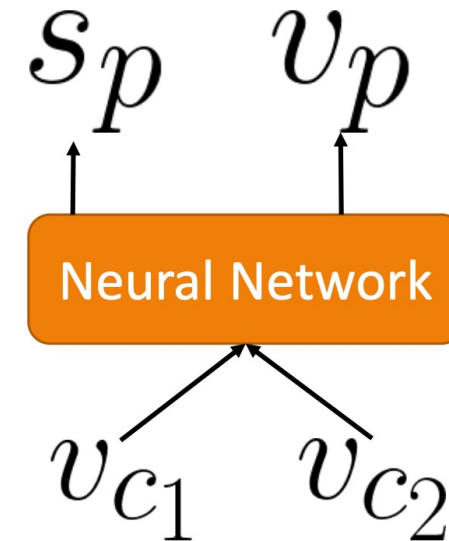
“house teapot”

$$s_p = u^T v_p \quad v_p = \sigma\left(W \begin{bmatrix} v_{c1} \\ v_{c2} \end{bmatrix} + b\right)$$

“house”



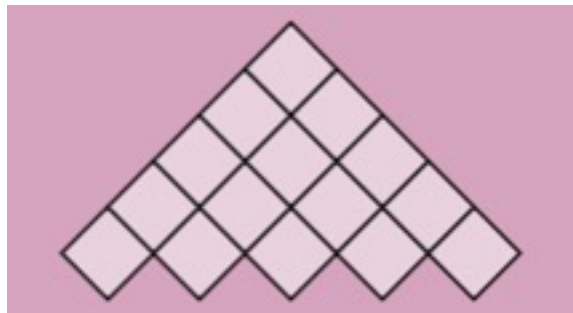
“teapot”



Tied weights

Compositional phrase embeddings

- Train: structured max-margin
- Structure prediction (inference)
 - Given current set of segments to merge, compute scores for all pairs of neighboring segments
 - Greedily merge the highest scoring segment
 - Iterate until only one segment left

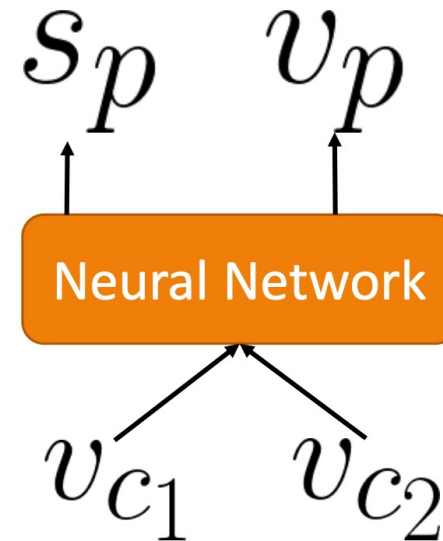


Score of two nodes combining

$$s_p = u^T v_p$$

Embedding for parent node

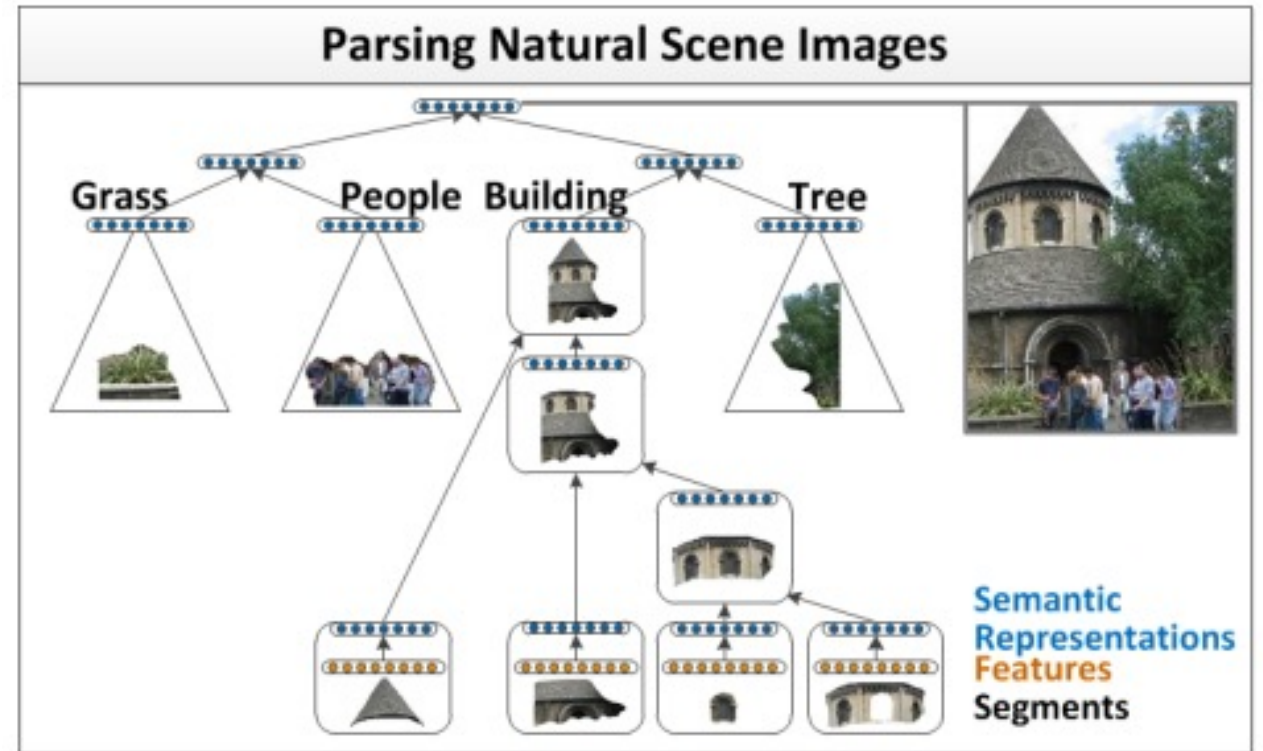
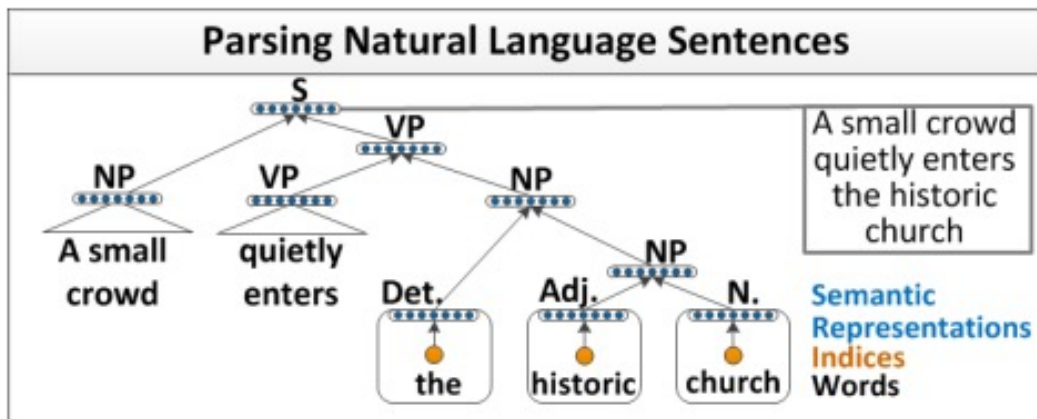
$$v_p = \sigma\left(W \begin{bmatrix} v_{c1} \\ v_{c2} \end{bmatrix} + b\right)$$



Tied weights

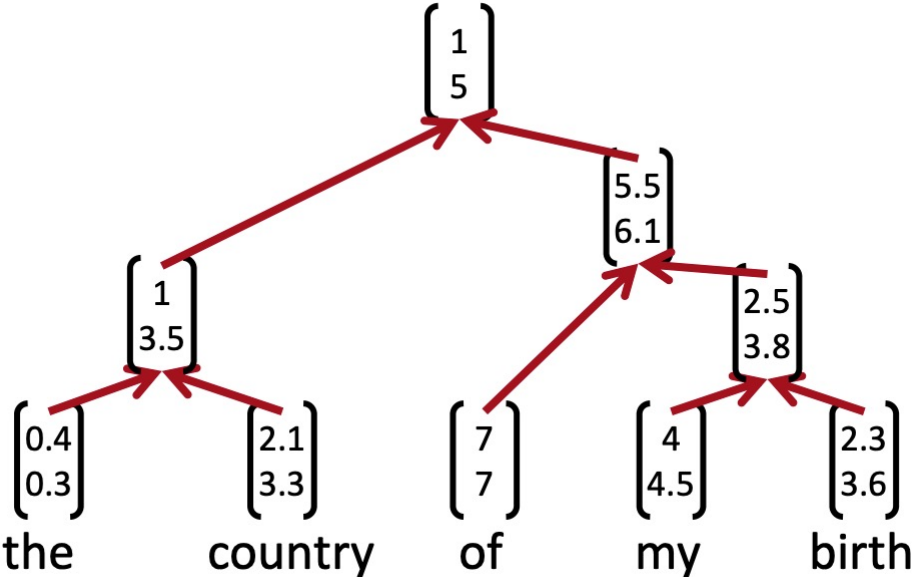
Recursive NNs

- Recursive model for building up compositional representations
- Can be applied to sentences and images



RvNNs vs RNNs

Recursive



Recurrent

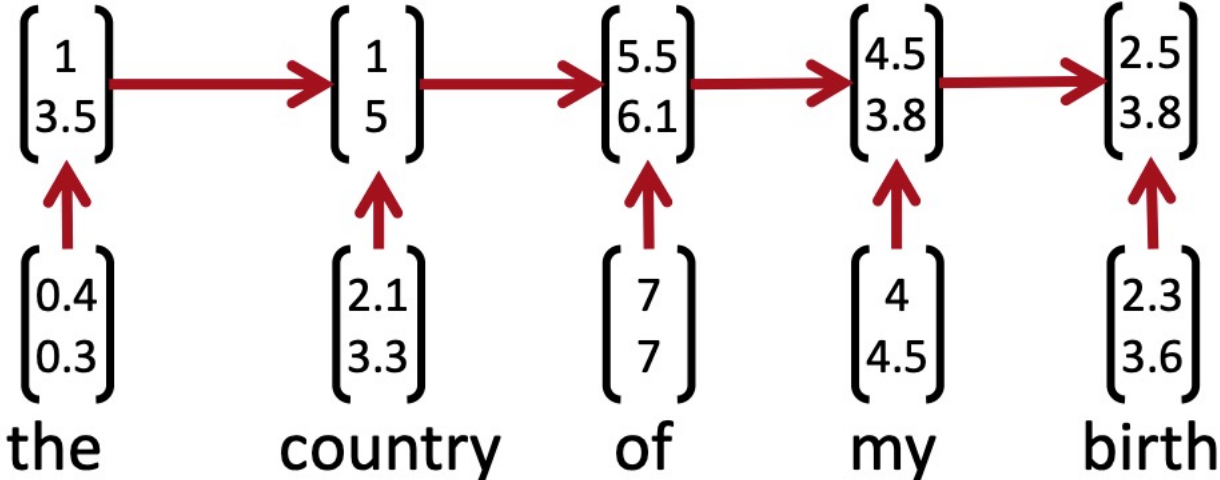
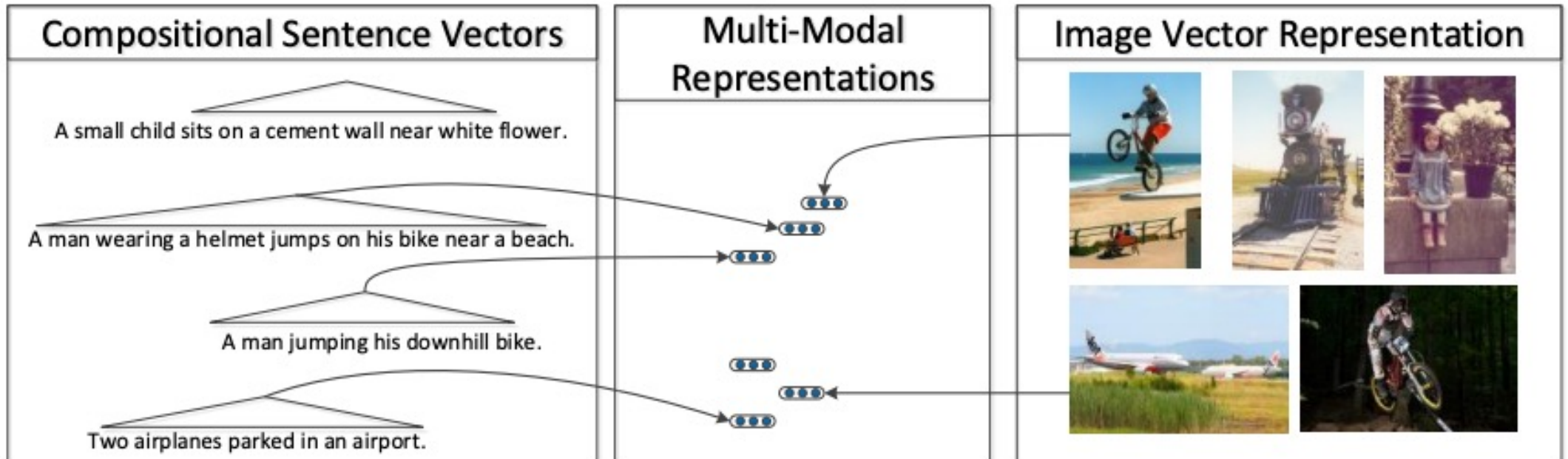
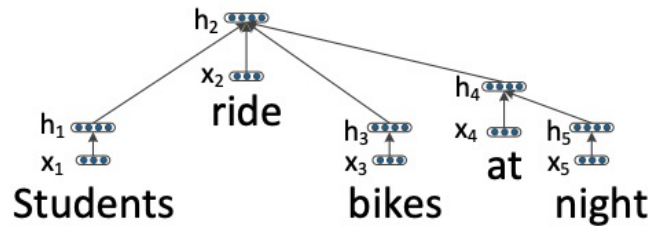


Image credit: Chris Manning

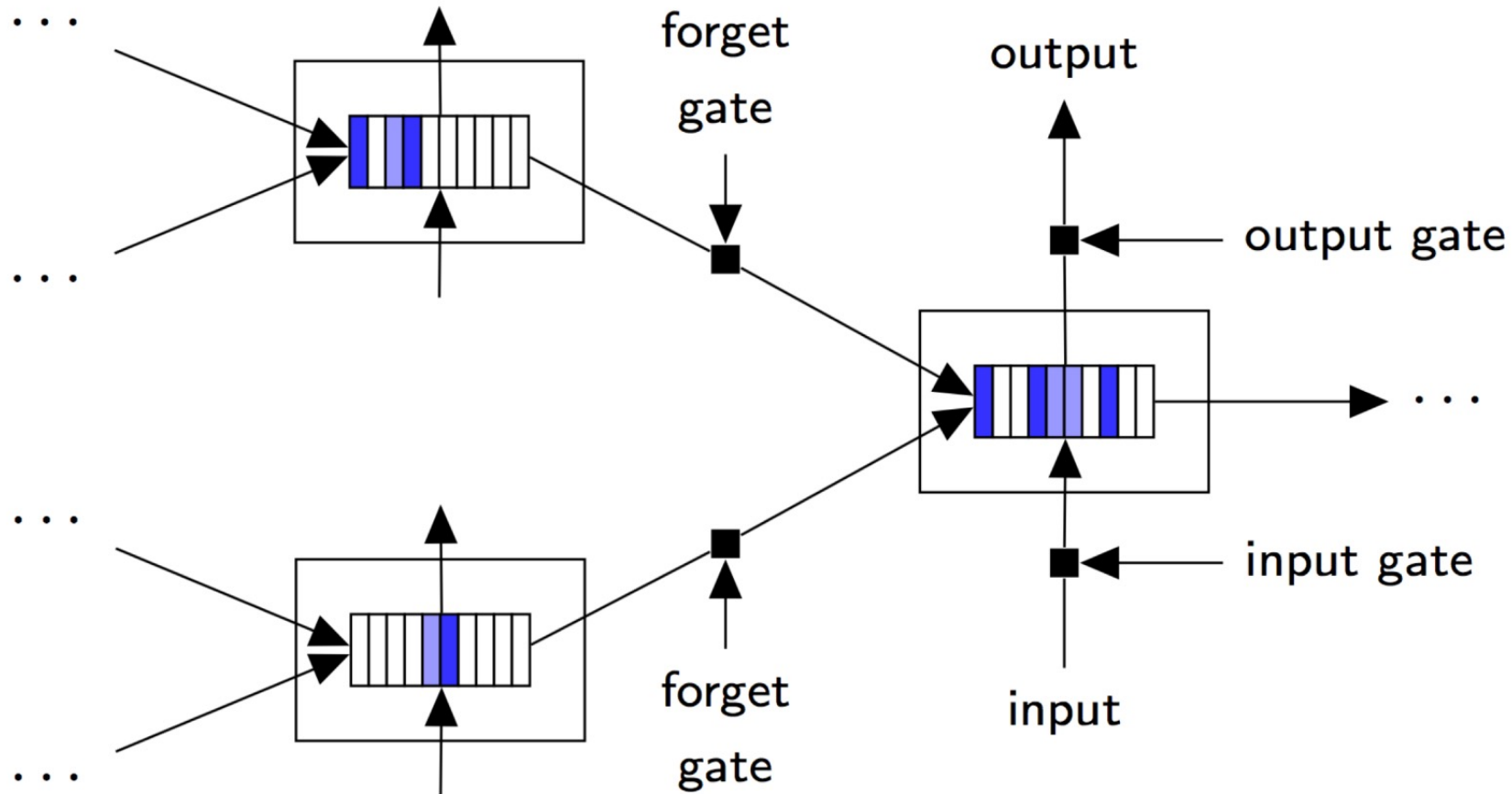
Recursive NNs

- Use dependency trees instead of constituency trees
- Apply to multi-modal search



TreeRNNs

- Extend to a n-ary trees



Inductive biases

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

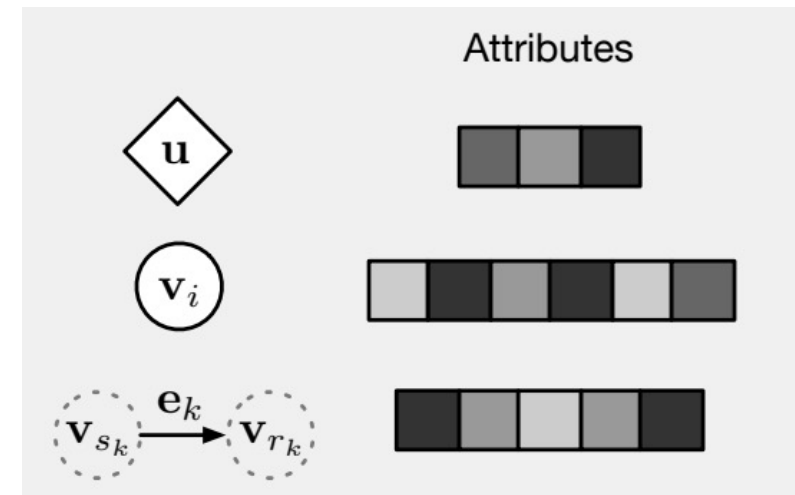
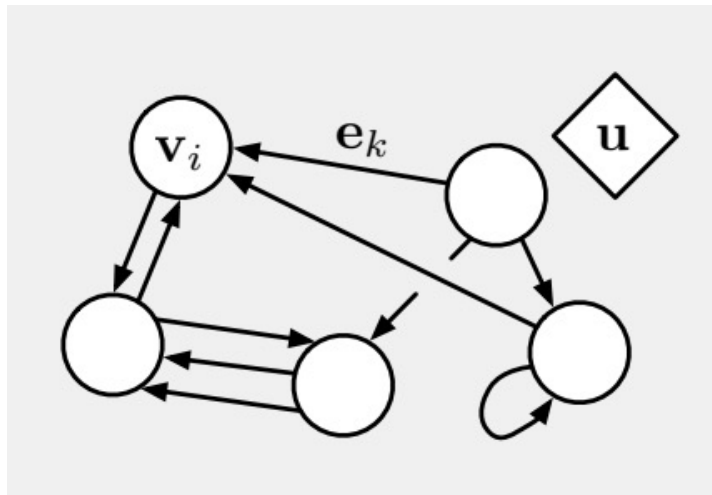
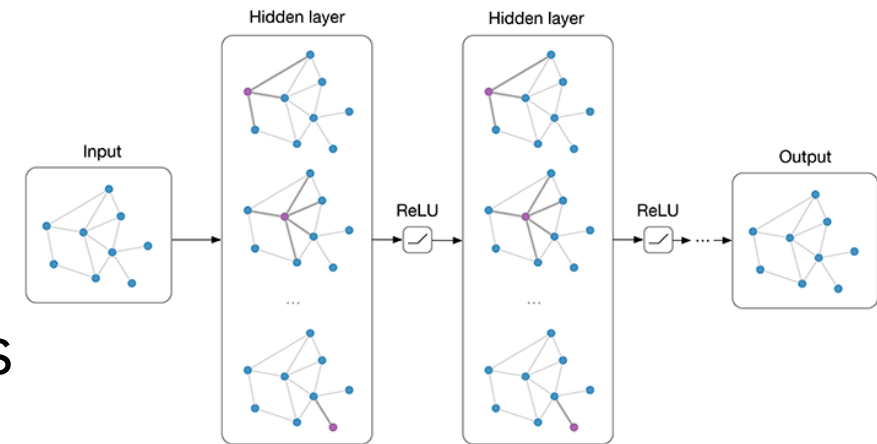
Component	Entities	Relations	Rel. inductive bias	Invariance
Fully connected	Units	All-to-all	Weak	-
Convolutional	Grid elements	Local	Locality	Spatial translation
Recurrent	Timesteps	Sequential	Sequentiality	Time translation
Graph network	Nodes	Edges	Arbitrary	Node, edge permutations

- These architecture constraints can help your network learn faster

Graph neural networks

GraphNNs

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



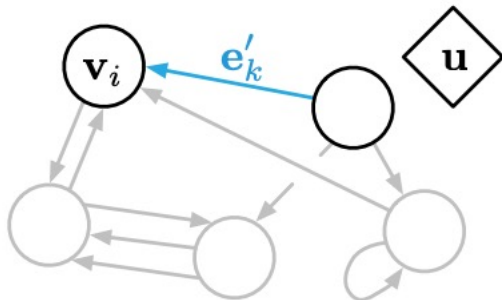
Relational inductive biases, deep learning, and graph networks, Battaglia et al, arXiv 2018

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

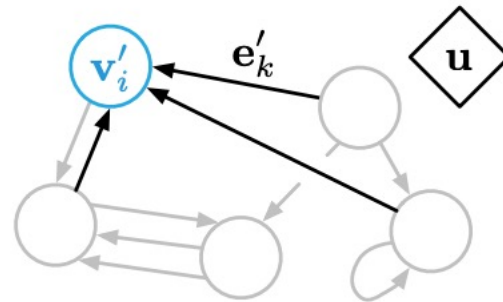
$$\mathbf{e}'_k = \phi^e(\mathbf{e}_k, \mathbf{v}_{r_k}, \mathbf{v}_{s_k}, \mathbf{u})$$



Update each edge e_k

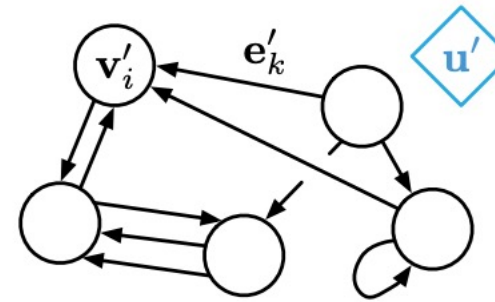
$$\bar{\mathbf{e}}'_i = \rho^{e \rightarrow v}(E'_i)$$

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



Update each node v_i

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



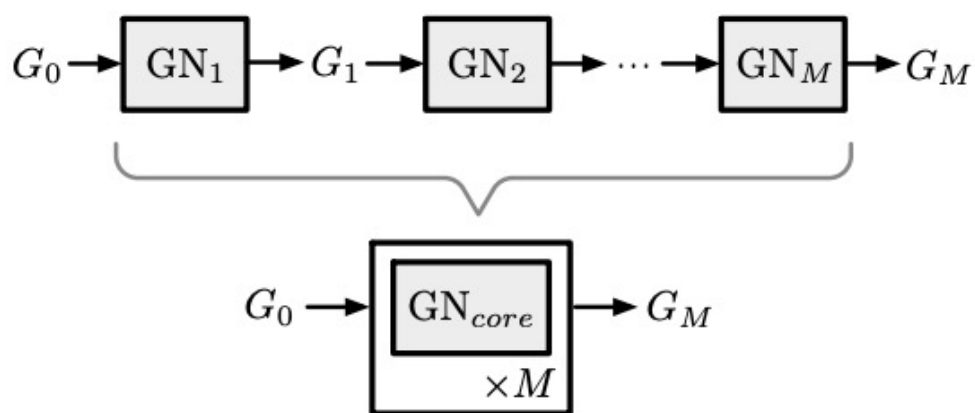
Update global graph u

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

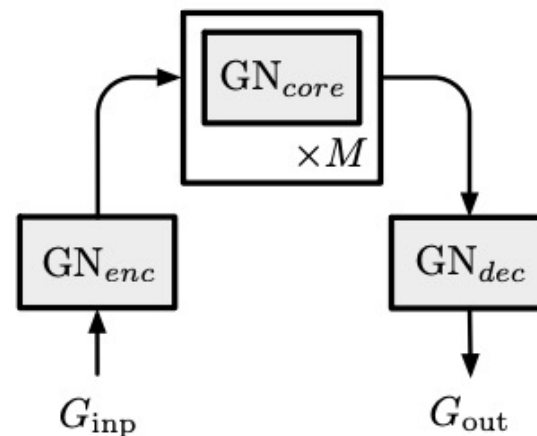
$$\bar{\mathbf{v}}' = \rho^{v \rightarrow u}(V')$$

GraphNNs

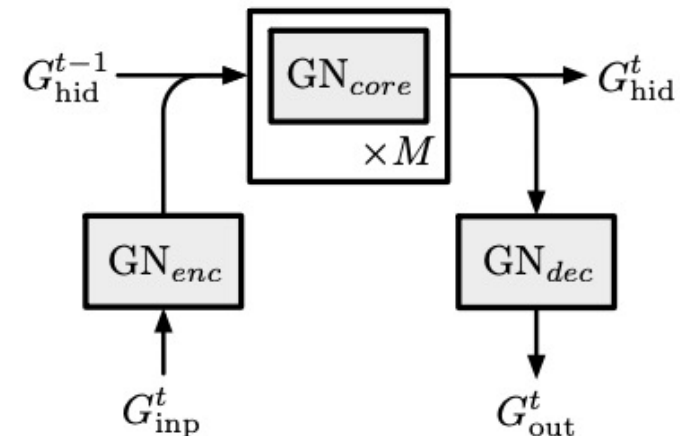
- GN blocks can be composed



(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

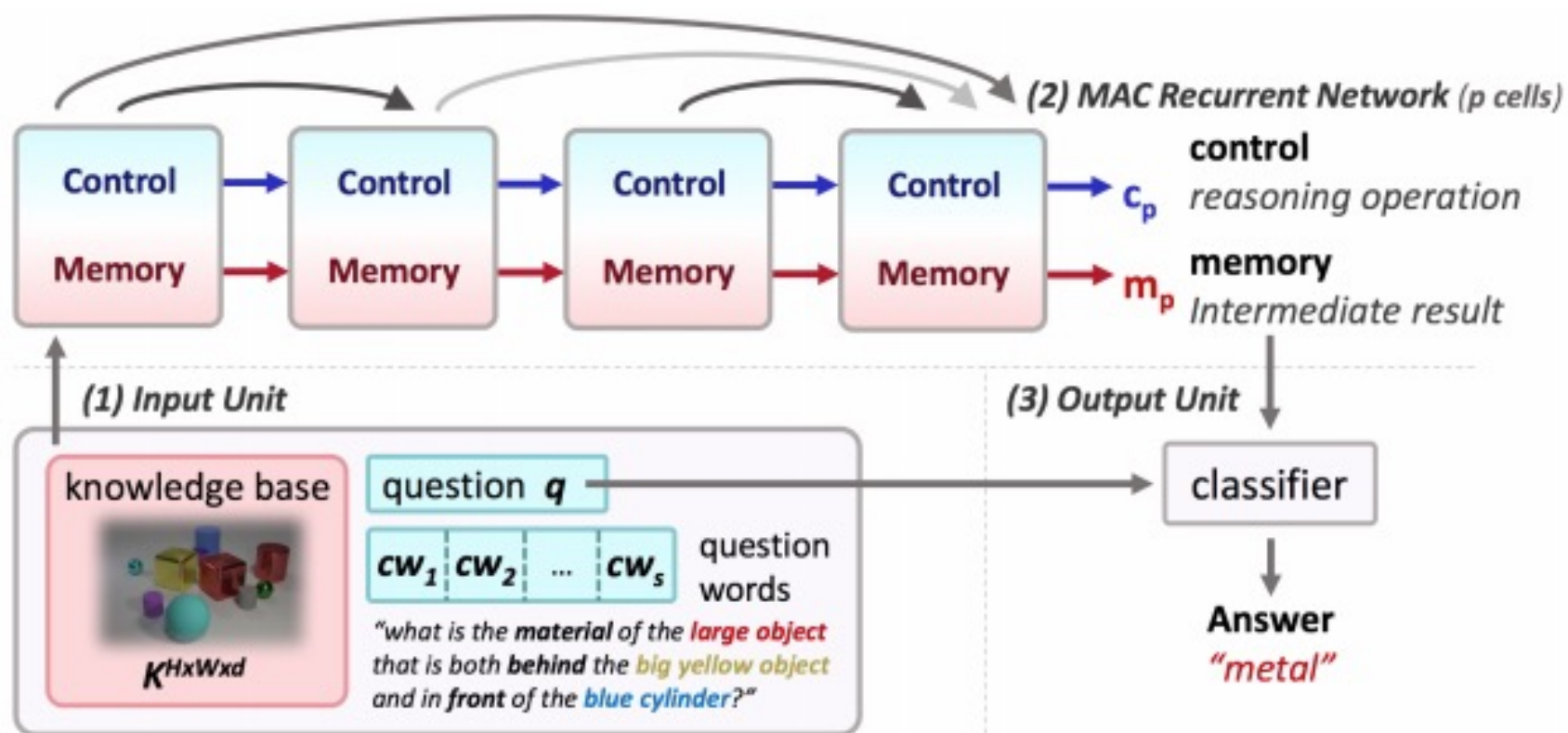
Code for working with GraphNNs

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

Structured reasoning

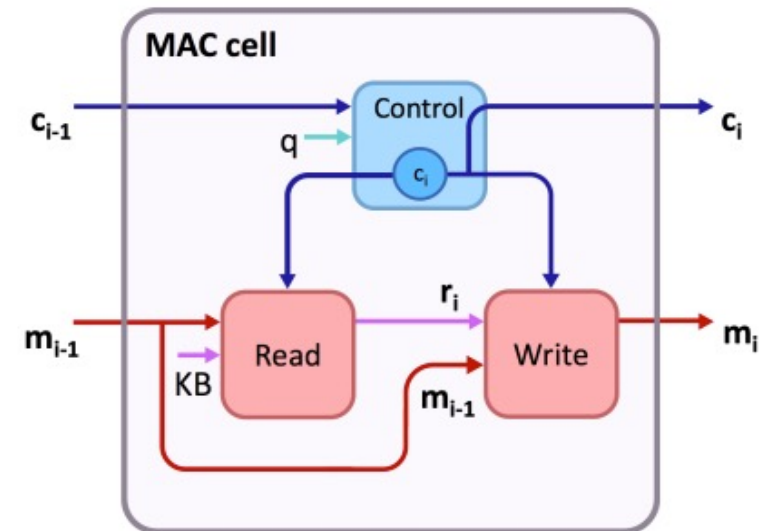
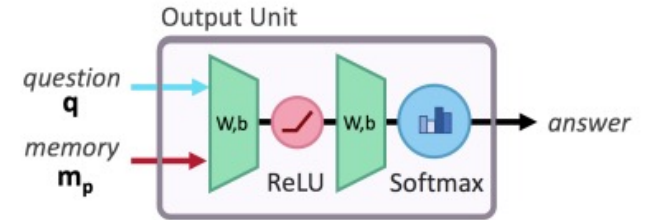
MAC (Memory, Attention, Control)

- Recurrent network with cell with read/write/control



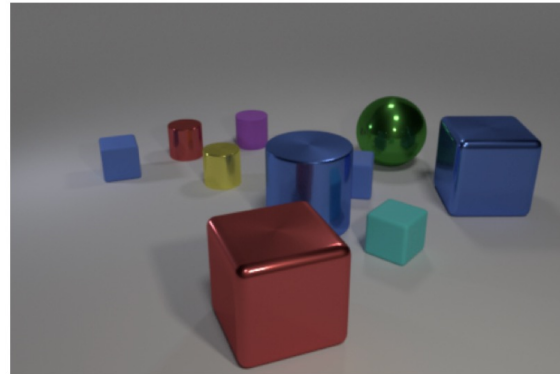
MAC (Memory, Attention, Control)

- 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



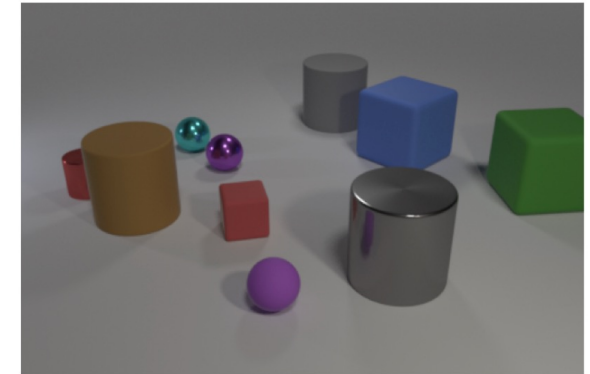
Compositionality and reasoning (CLEVR dataset, Johnson et al, 2017)

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



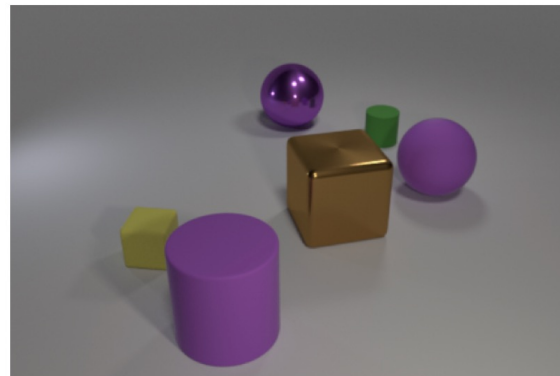
Q: What shape is the object reflected in the blue cylinder?

A: cube



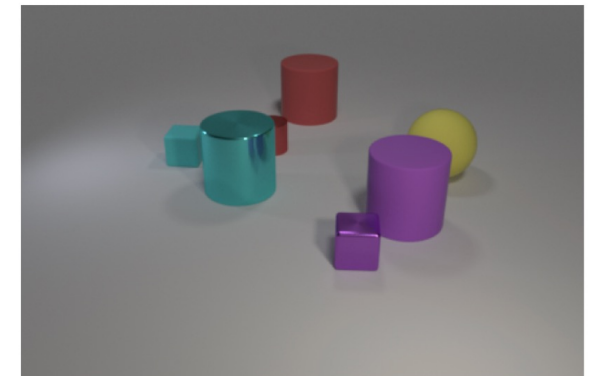
Q: What number of cylinders share the same color?

A: 2



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

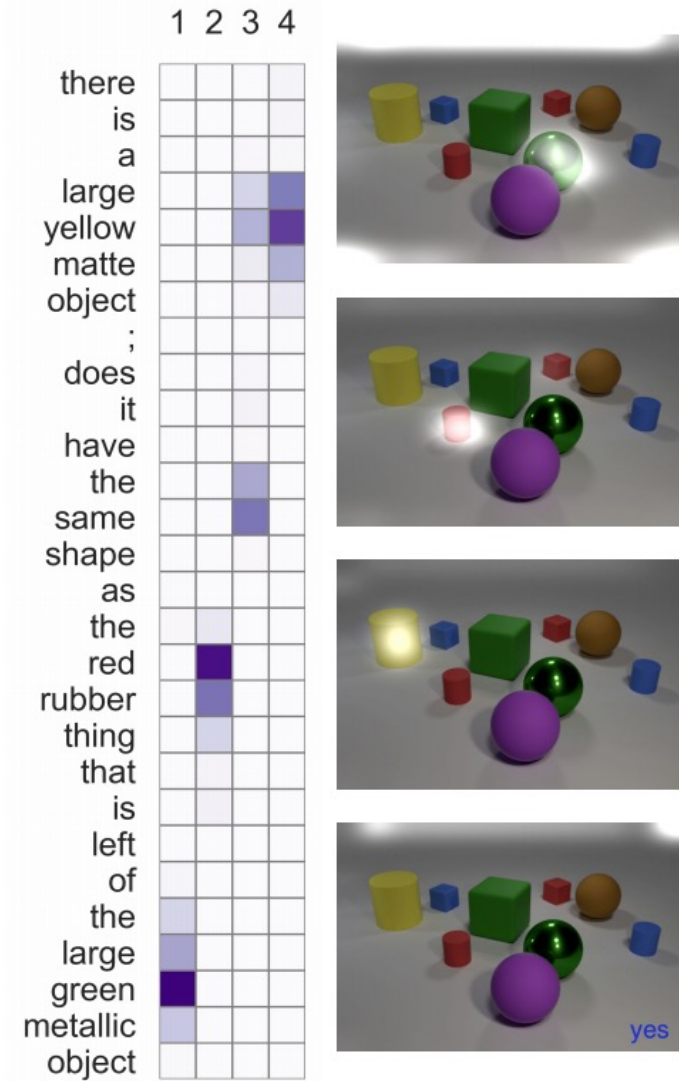
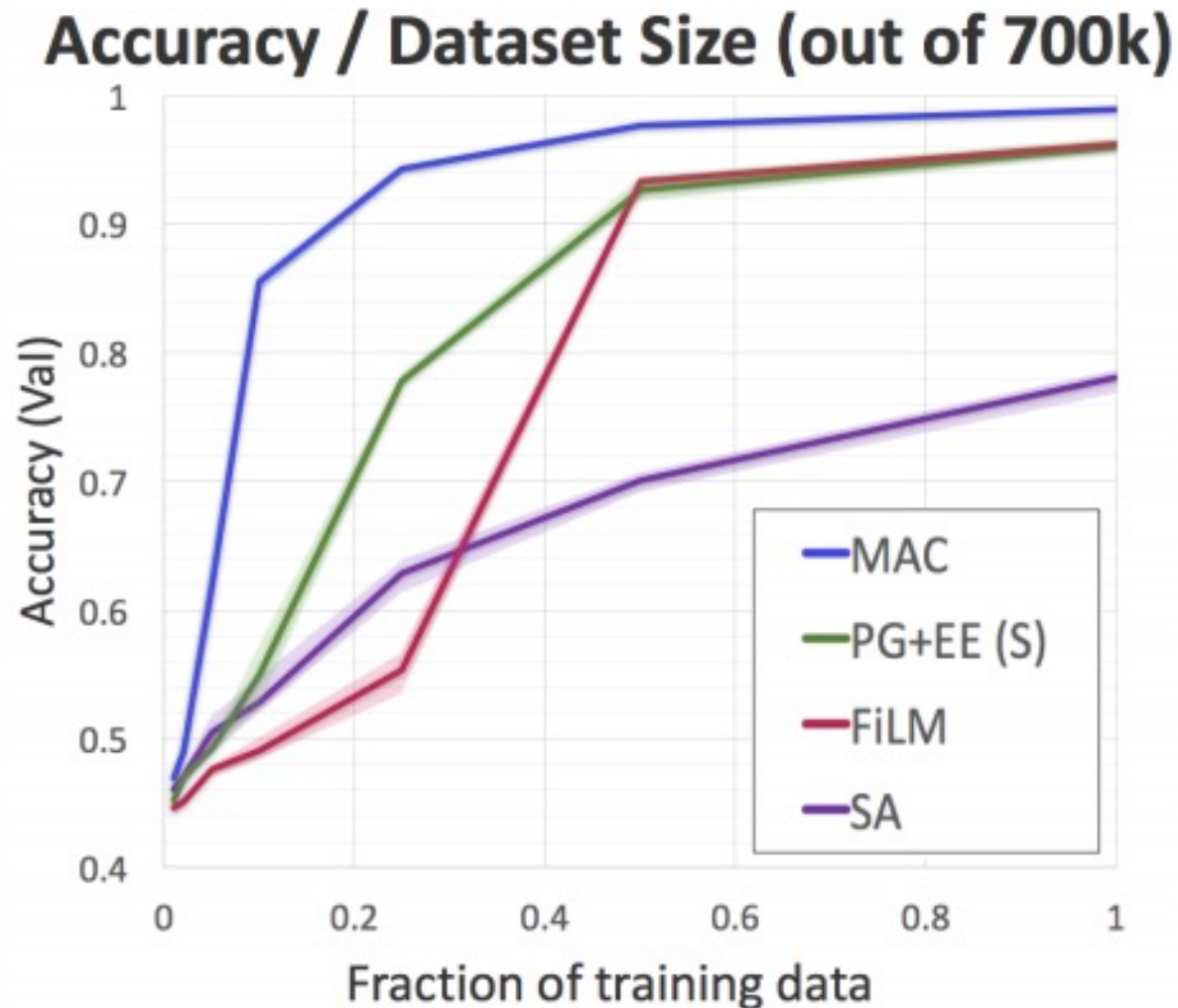
A: 2



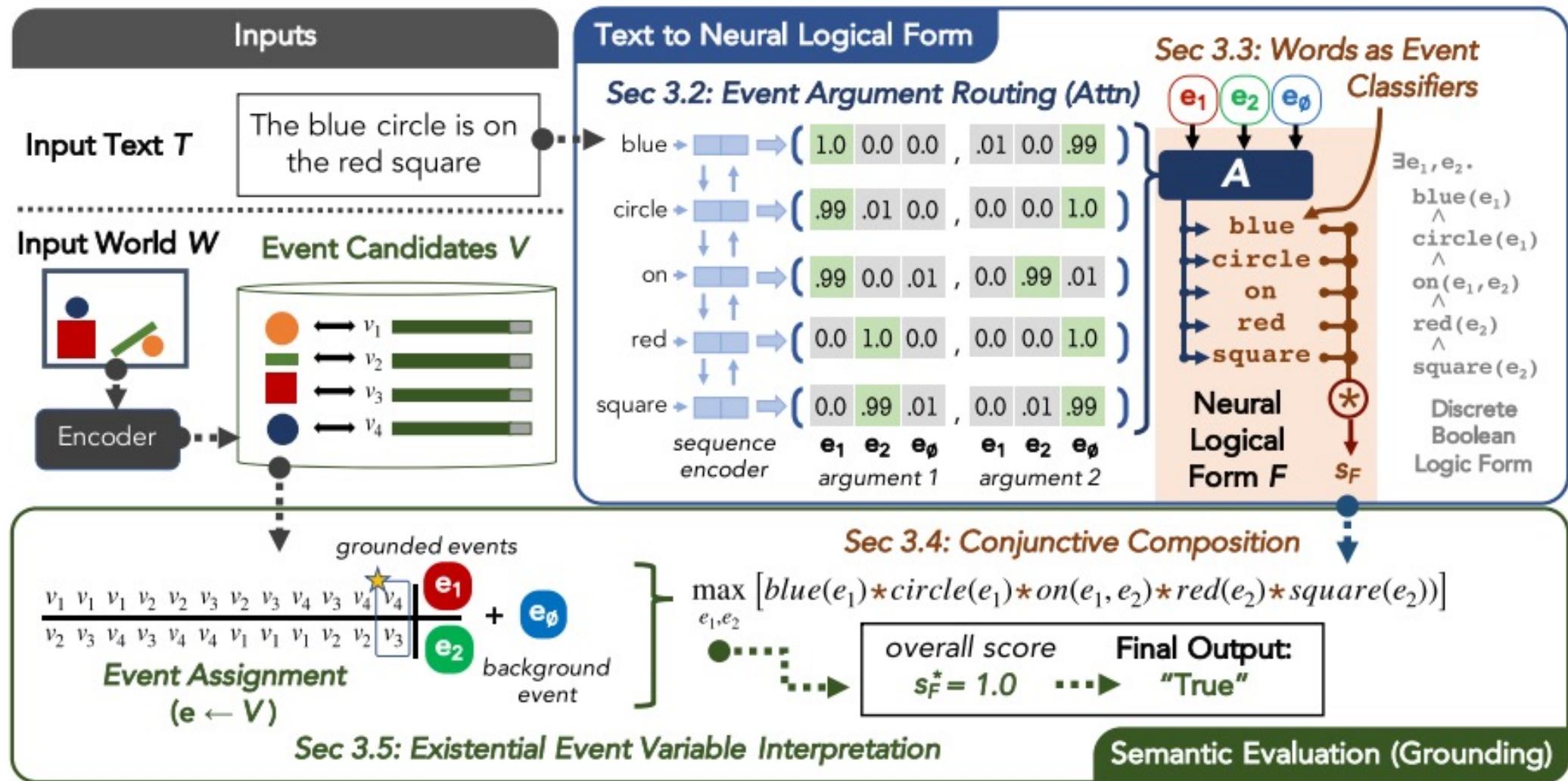
Q: What color is the object partially blocked by the purple cylinder?

A: yellow

MAC can learn with smaller amount of data



Neural Event Semantics



Next time (after the break)

- Paper presentations
 - Grounded Compositional Semantics For Finding and Describing Images with sentences (Yanshu)
 - Neural Event Semantics for Grounded Language Understanding (Sihui)
- Project proposal (2 groups)
- Wednesday (3/2): Semantic Parsing (language to programs)