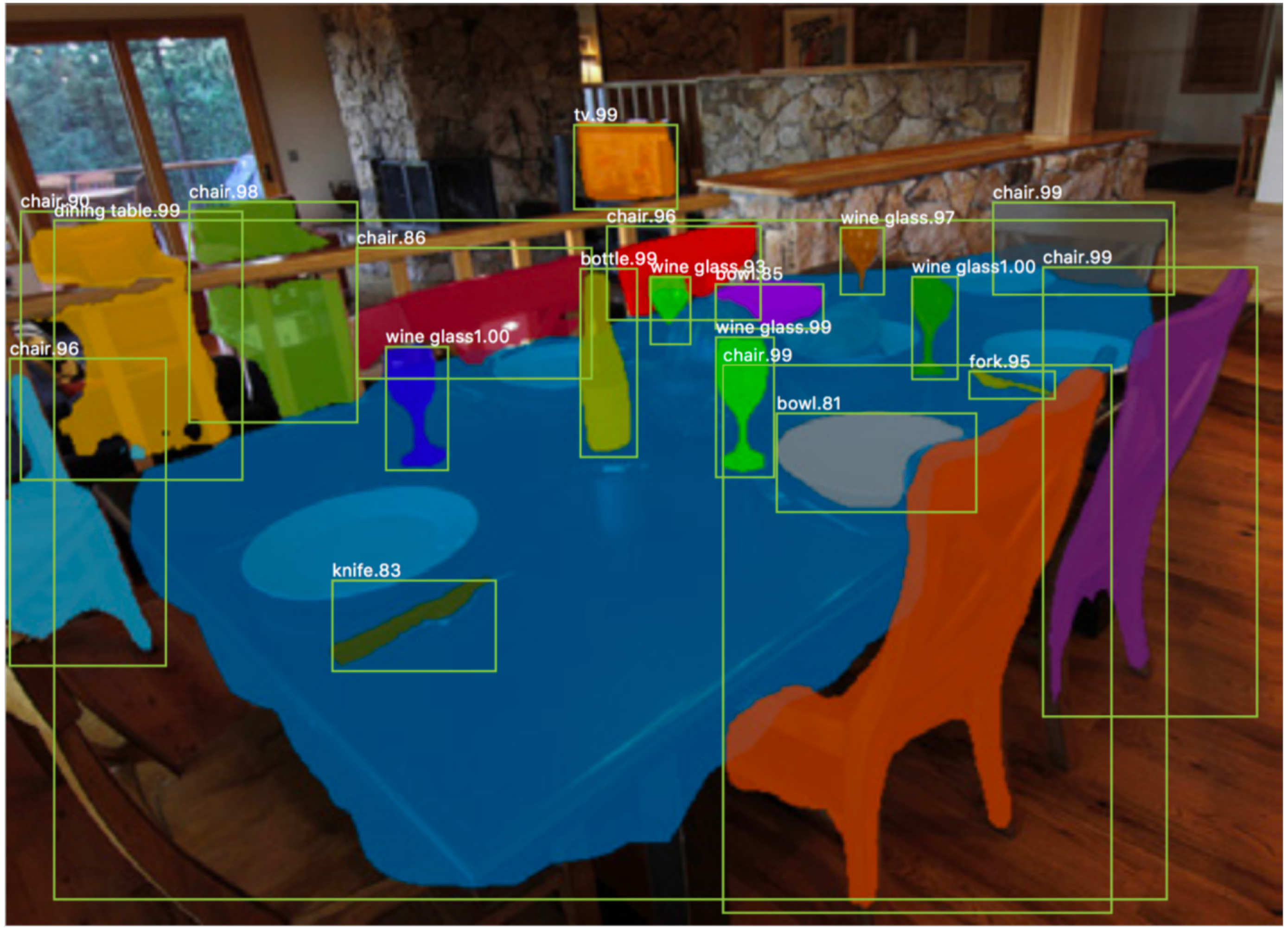
Visual Relationship Reasoning with Scene Graph

Meng-Jiun Chiou

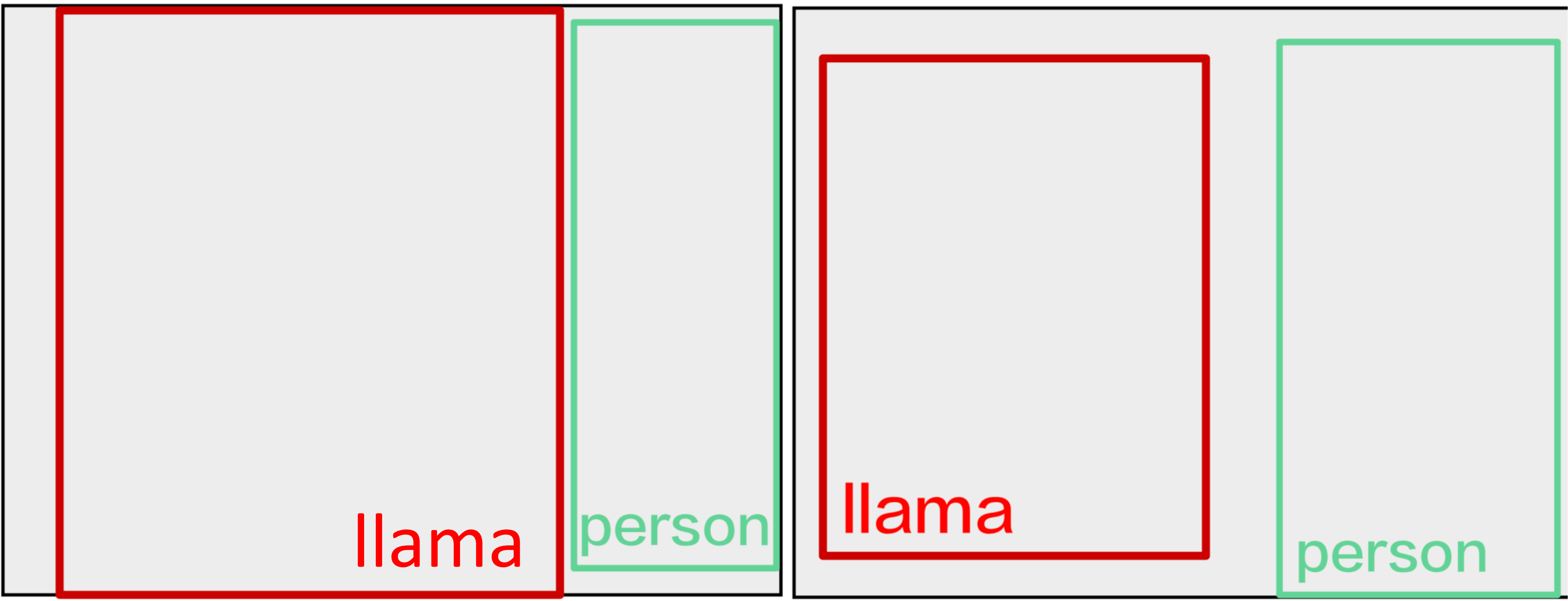
National University of Singapore

Feburary 2, 2019

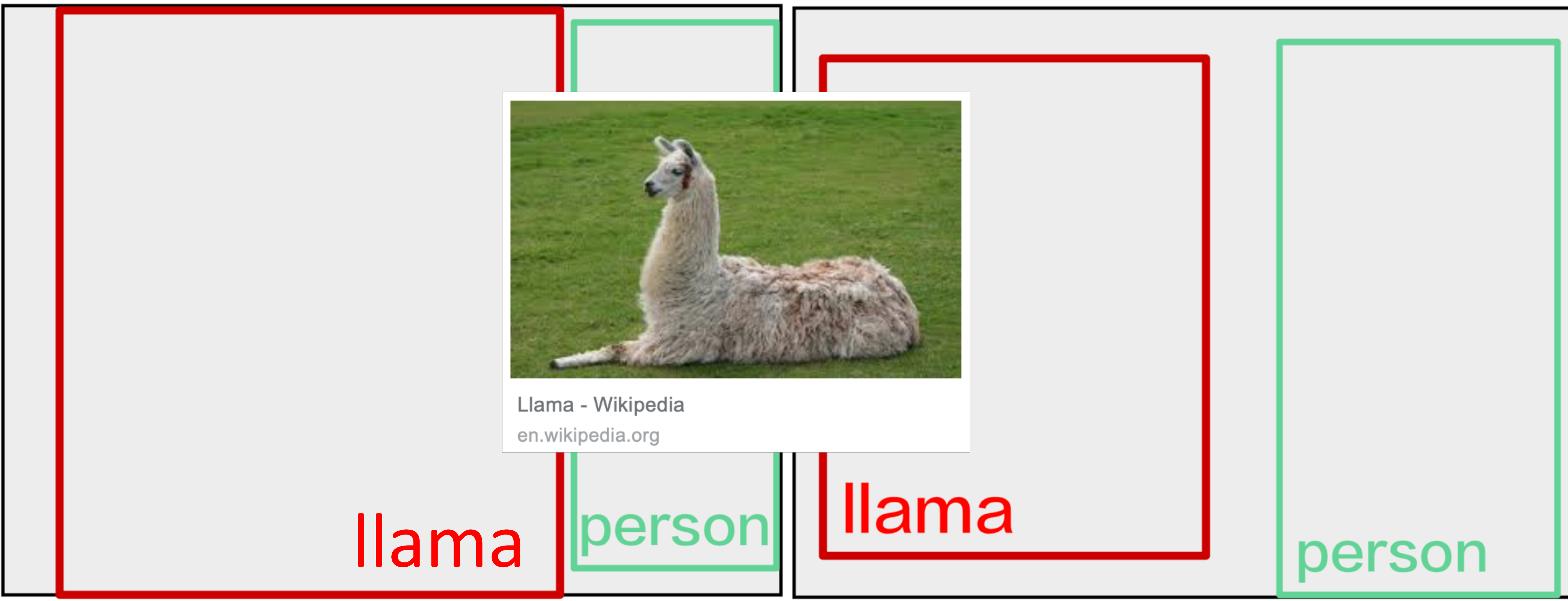
Object detection



He et al. “Mask R-CNN”, ICCV 2017



Slide credit: Ranjay Krishna



Slide credit: Ranjay Krishna

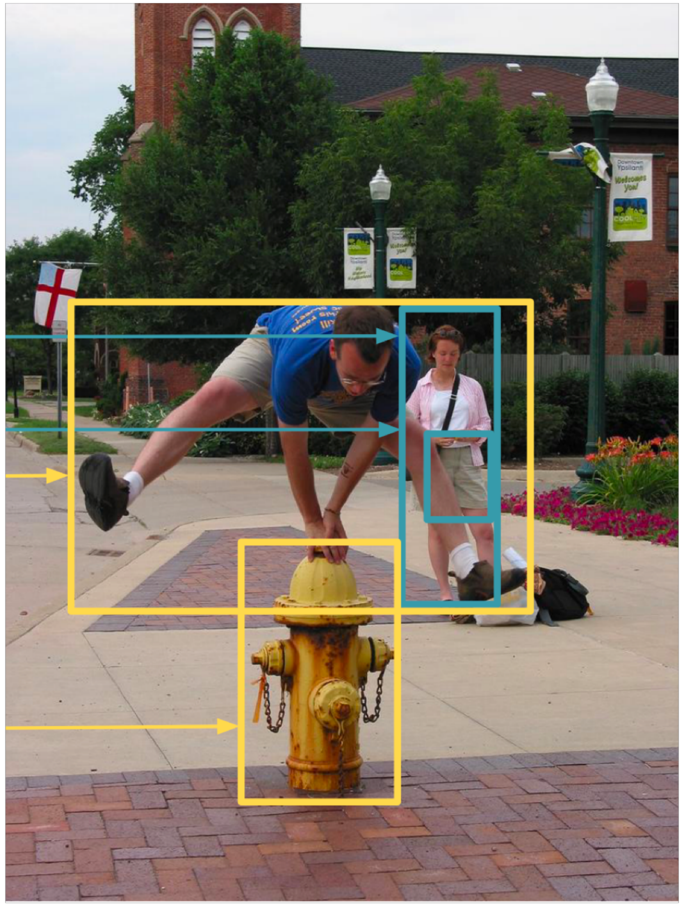
llama next to person

llama chasing person

Slide credit: Ranjay Krishna

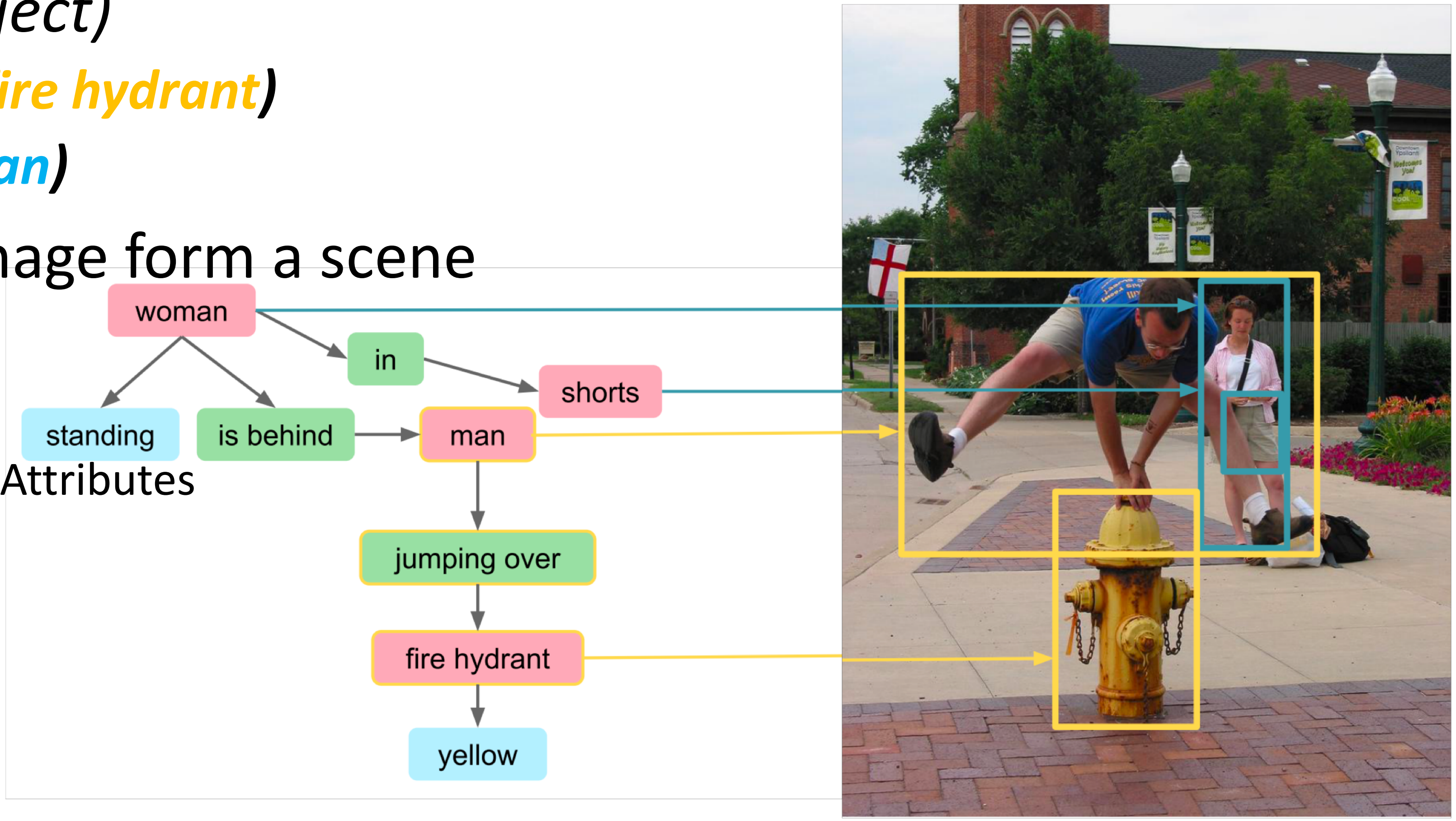
Visual Relationship Detection (VRD)

- Usually represented by visual phrases:
  (subject, predicate, object)
  - ***(man , jumping over, fire hydrant)***
  - ***(woman, is behind, man)***



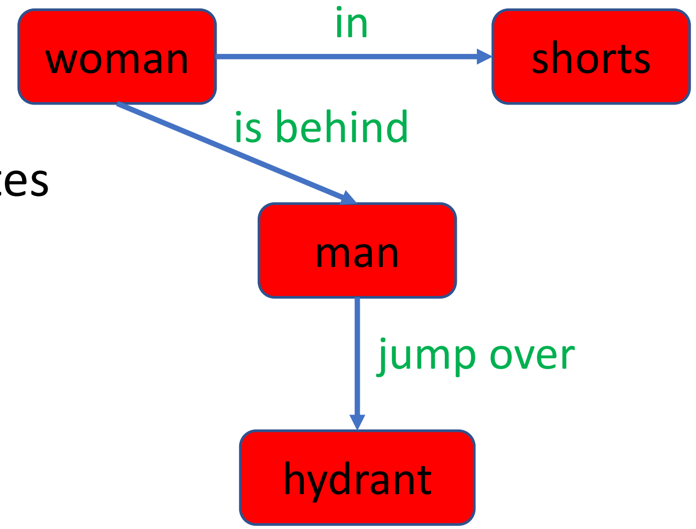
Visual Relationship Detection (VRD)

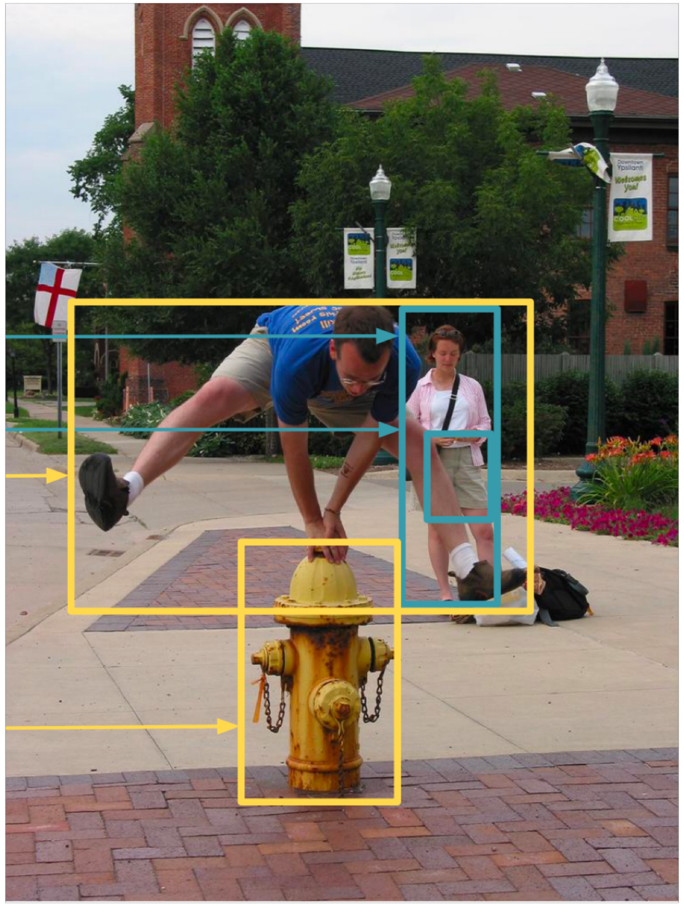
- Usually represented by visual phrases: *(subject, predicate, object)*
  - ***(man , jumping over, fire hydrant)***
  - ***(woman, is behind, man)***
- Visual phrases in an image form a scene graph:
  - Vertices:
    - Objects, Predicates or Attributes



Visual Relationship Detection (VRD)

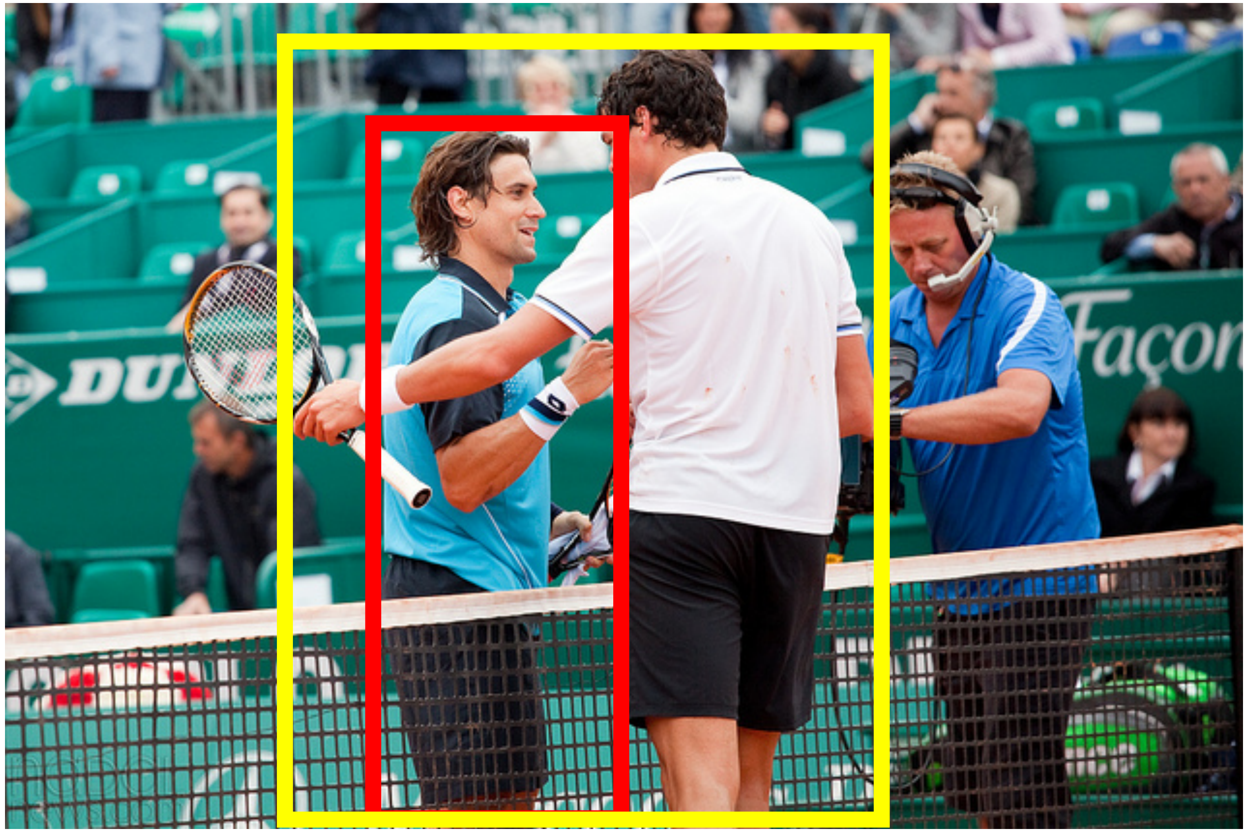
- Usually represented by visual phrases: *(subject, predicate, object)*
  - ***(man , jumping over, fire hydrant)***
  - ***(woman, is behind, man)***
- Visual phrases in an image form a scene graph:
  - Vertices:
    - Objects, Predicates or Attributes
  - Another (simple) definition:
    - Vertices: Objects
    - Edge: Predicates





Applications Benefit from VRD: Image Caption

- Example visual relationships:
  - *(man_1, handshakes, man_2)*
  - *(man_1, talks to, man_2)*
- Ground-truth captions:
  - a **man** giving **another man** a **hand shake** on a tennis court.
  - two tennis players **talk to each other** near the net.

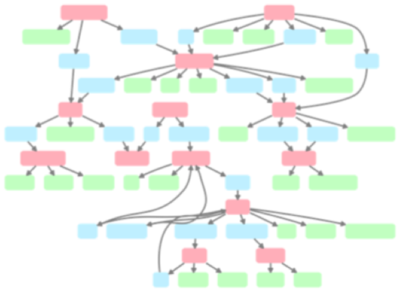


man_1 man_2

Datasets

Scene Graphs 5K

Johnson et al, CVPR 2015



- 5000 images
- 6745 object categories
- 1310 relationship types
- Long-tailed

Visual Relationships

Lu et al, ECCV 2016

- 5000 images
- 100 object categories
- 70 relationship types
- Fully-annotated

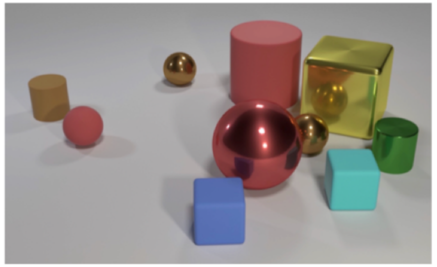
Visual Genome

Krishna et al, IJCV 2017

- 108K images
- 33K object categories
- 42K relationship types
- Long-tailed

CLEVR

Johnson et al, CVPR 2017



- 100K images
- 3 object categories
- 8 relationship types
- Fully-annotated

Outline

- *Visual Relationship Detection with Language Priors* (ECCV 2016)
- *Scene Graph Generation by Iterative Message Passing* (CVPR 2017)
- *Neural Motifs: Scene Graph Parsing with Global Context* (CVPR 2018)

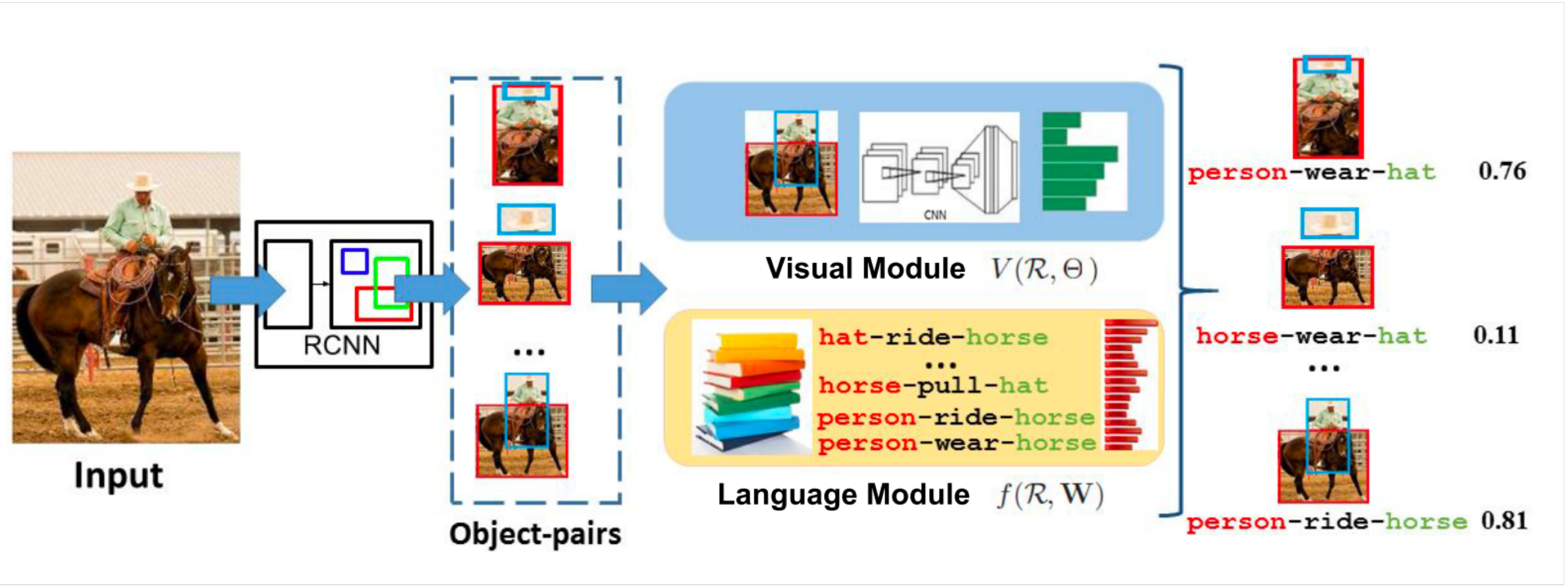
- Experiments Result

Visual Relationship Detection with Language Priors

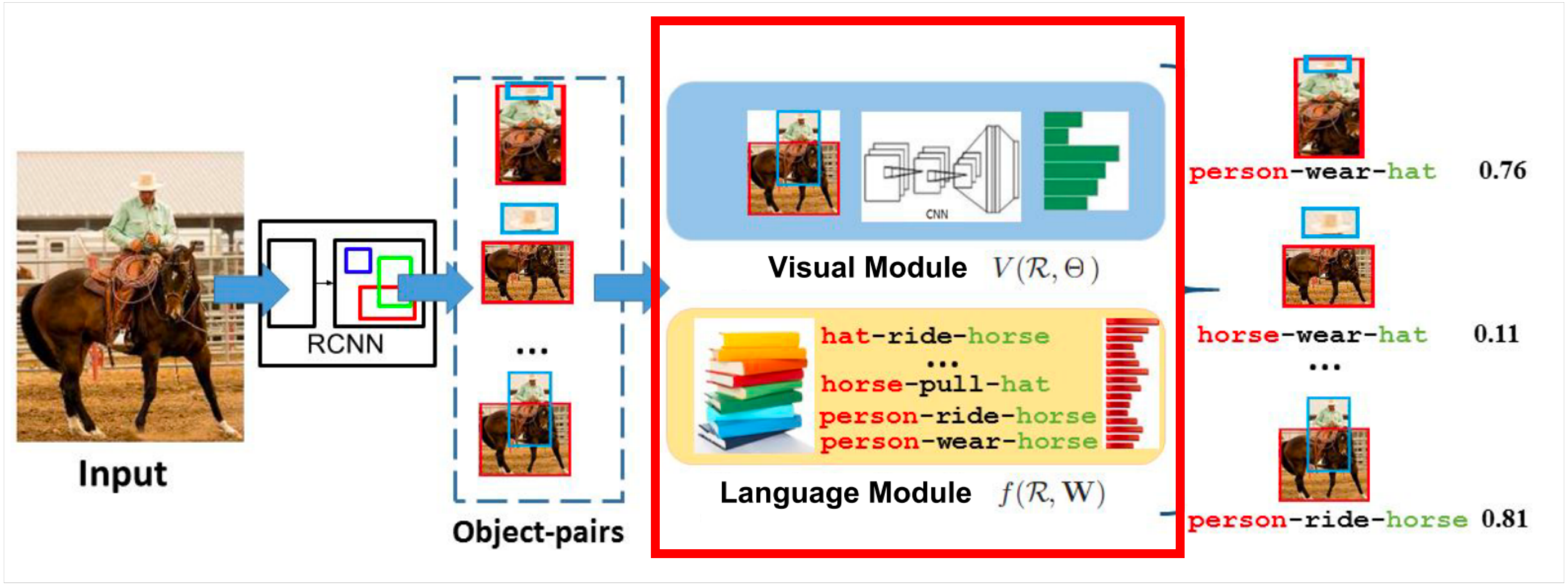
Cewu Lu*, Ranjay Krishna*, Michael Bernstein, Li Fei-Fei
{cwlu, ranjaykrishna, msb, feifeili}@cs.stanford.edu

Stanford University

VRD with Language Prior: Architecture



VRD with Language Prior: Architecture



Visual Appearance Module

- Prior to this work, visual relationship detection is generally based on *visual phrase* classification [1]
  - $O(N^2K)$ unique detectors where we have N objects and K predicates classes
- They propose a **visual appearance module** to predict objects and predicate individually and fuse them together to form a phrase
  - Reduce to $O(N+K)$
- Train two CNNs for classification with *N* classes and *K* predicates respectively and model V as

$$V(R_{\langle i,k,j \rangle}, \Theta | \langle O_1, O_2 \rangle) = P_i(O_1)(\mathbf{z}_k^T \mathrm{CNN}(O_1, O_2) + s_k) P_j(O_2)$$

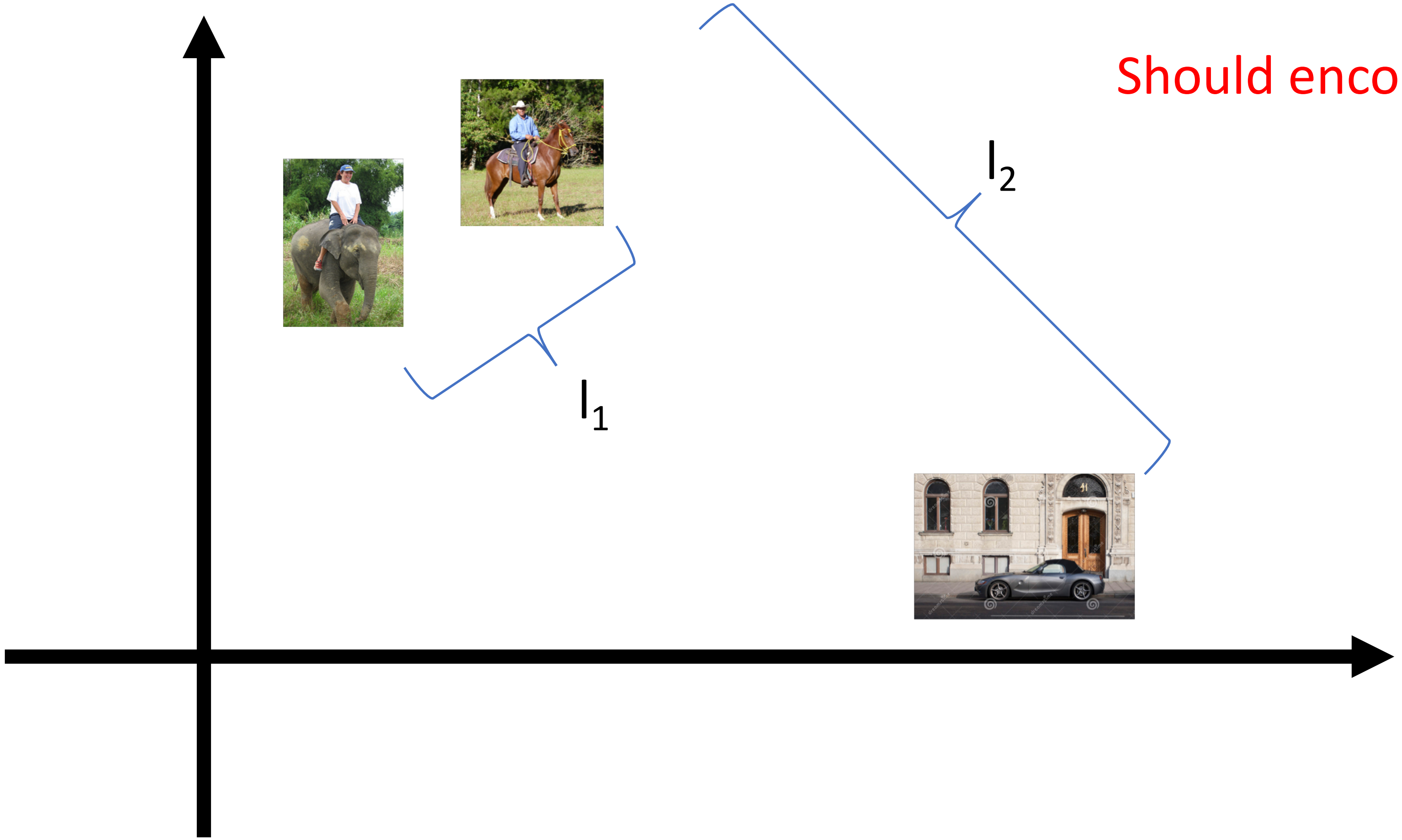
Language Module – Intuition 1

(person, ride, horse)

(person, ride, elephant)

(car, near, house)

Visual Relationship Space



Should encode the idea $l_1 < l_2$

Language Module: Minimize dist. of relationship

- Convert object class labels to 300-dim Word2Vec vectors:

$$f(\mathcal{R}_{\langle i,k,j \rangle}, \mathbf{W}) = \mathbf{w}_k^T [word2vec(t_i), word2vec(t_j)] + b_k$$

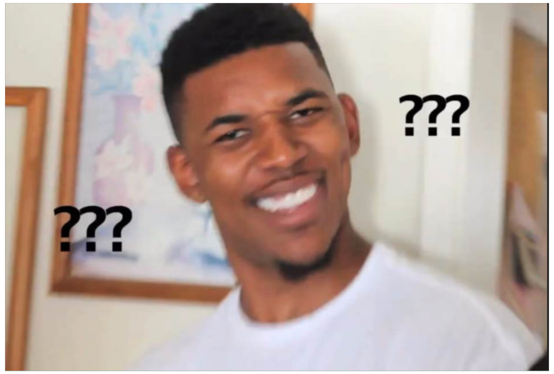
- Under assumption of the distance of visual relationship is proportional to the sum of Word2Vec distance of objects and predicates, randomly sample pairs of $(\langle \mathcal{R}, \mathcal{R}' \rangle)$ and minimize the variance to fulfill the assumption:

$$K(\mathbf{W}) = var(\{\frac{\boxed{[f(\mathcal{R}, \mathbf{W}) - f(\mathcal{R}', \mathbf{W})]^2}}{\boxed{d(\mathcal{R}, \mathcal{R}')}} \quad \forall \mathcal{R}, \mathcal{R}'\})$$

Language Module: Likelihood of Relationship

- Project function f should represent the occurrence likelihood of a relationship: such as ***(monkey, drive, car)*** **should have low likelihood**. We minimize **rank loss function** as follows:

$$L(\mathbf{W}) = \sum_{\{\mathcal{R},\mathcal{R}'\}} \max\{f(\mathcal{R}',\mathbf{W}) - f(\mathcal{R},\mathbf{W}) + 1, 0\}$$



Final Objective

- Maximize the rank of the ground truth relationship R with bounding boxes O_1 and O_2 using **rank loss**:

$$
\begin{aligned}
C(\Theta, \mathbf{W}) = \sum_{\langle O_1 O_2 \rangle, \mathcal{R}} & \max\{\overbrace{1 - V(\mathcal{R}, \Theta | \langle O_1, O_2 \rangle) f(\mathcal{R}, \mathbf{W})}^{\text{Maximize correct labels' likelihood}} \\
& + \max_{\langle O_1', O_2' \rangle \neq \langle O_1, O_2 \rangle, \mathcal{R}' \neq \mathcal{R}} \underbrace{V(\mathcal{R}', \Theta | \langle O_1', O_2' \rangle) f(\mathcal{R}', \mathbf{W})}_{\text{Minimize incorrect labels' likeihood}}, 0\}
\end{aligned}
$$

- Integrating language module, the **final objective** is then

$$
\min_{\Theta, \mathbf{W}} \{C(\Theta, \mathbf{W}) + \lambda_1 L(\mathbf{W}) + \lambda_2 K(\mathbf{W})\}
$$

Strength and Weakness

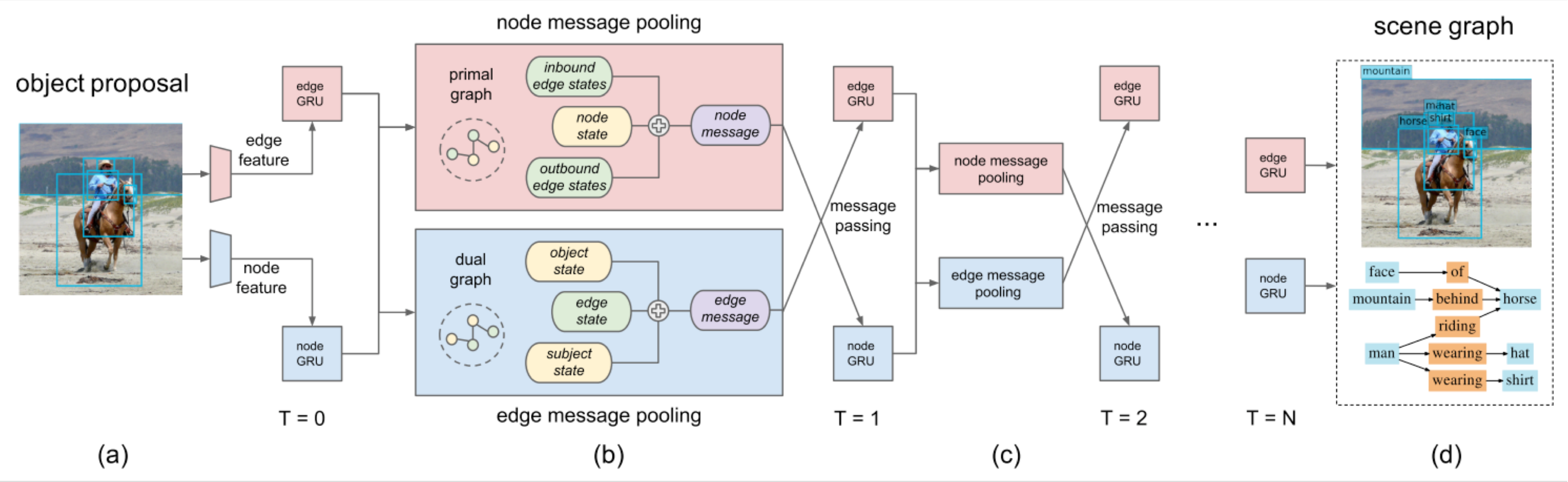
- First to formulate the visual relationship detection as object & predicate prediction respectively, reducing the complexity
- Mapping a relationship into the vector space and exploiting language prior makes the model learn some good dataset bias

- Fails to exploit the **context** of objects and relationships
    - It focuses on *pairwise* relationships

Scene Graph Generation by Iterative Message Passing

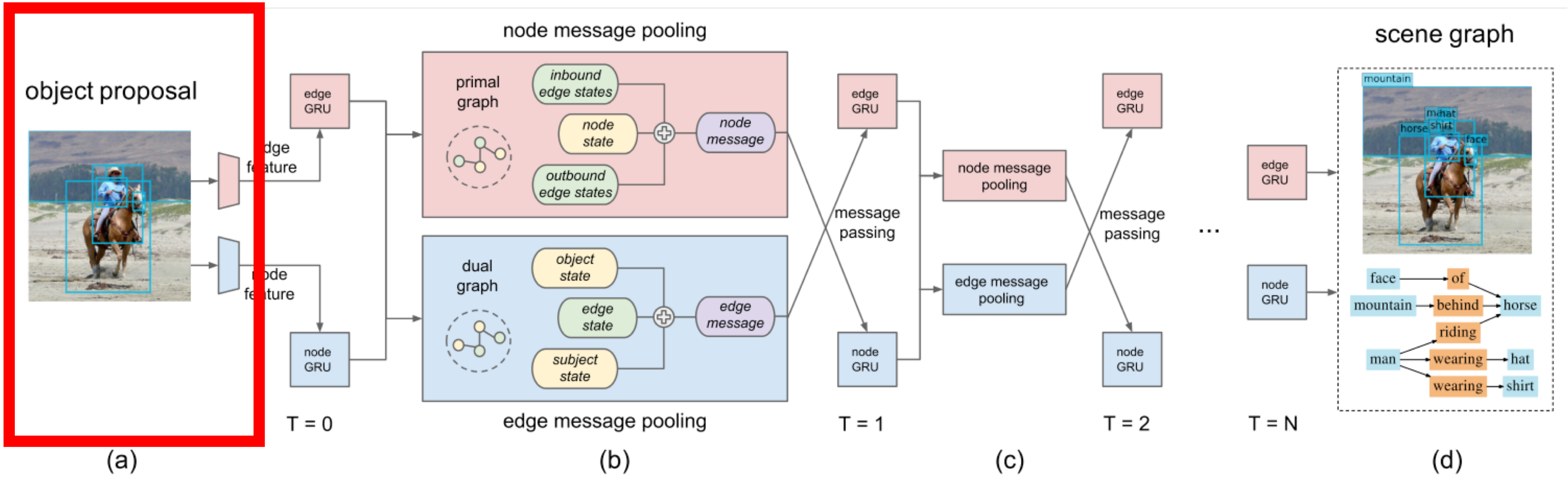
Danfei Xu[1] Yuke Zhu[1] Christopher B. Choy[2] Li Fei-Fei[1]
[1]Department of Computer Science, Stanford University
[2]Department of Electrical Engineering, Stanford University
{danfei, yukez, chrischoy, feifeili}@cs.stanford.edu

Scene Graph Generation by IMP



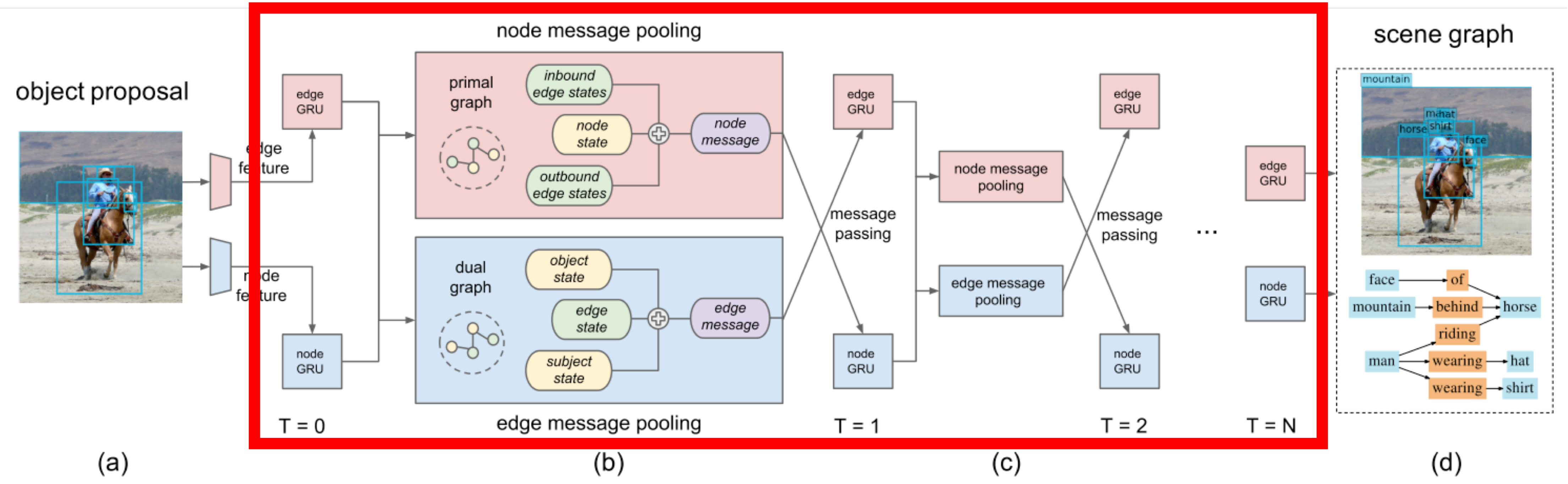
Scene Graph Generation by IMP

CNN + RPN



Scene Graph Generation by IMP

Iterative Message Passing



Graph Inference Problem Setting

- Each node in the graph is associated with a random variable x_i
- We denote the set of all variables to be

$$\mathbf{x} = \{x_i^{cls}, x_i^{bbox}, x_{i\to j} | i = 1 \ldots n, j = 1 \ldots n, i \neq j\}$$

- We want to find

$$\mathbf{x}^* = \arg\max_{\mathbf{x}} \Pr(\mathbf{x}|I, B_I)$$

that maximize the conditional probability (under *Naïve Bayes assumption*)

$$\Pr(\mathbf{x}|I, B_I) = \prod_{i \in V} \prod_{j \neq i} \Pr(x_i^{cls}, x_i^{bbox}, x_{i\to j} | I, B_I)$$

- We need to do **Bayesian inference** to obtain the conditional probability!

Inference with Mean Field Approximation

- Exact inference on densely connected graph can be very expensive, thus we choose **variational inference** to approximate the true distribution *p(x)* with a simpler distribution *q(x)*.

- *Mean field variational inference* factorizes distribution as product of local variational approximation:

$$q(x) = \prod_i q_i(x_i)$$

Mean Field Approximation using GRU

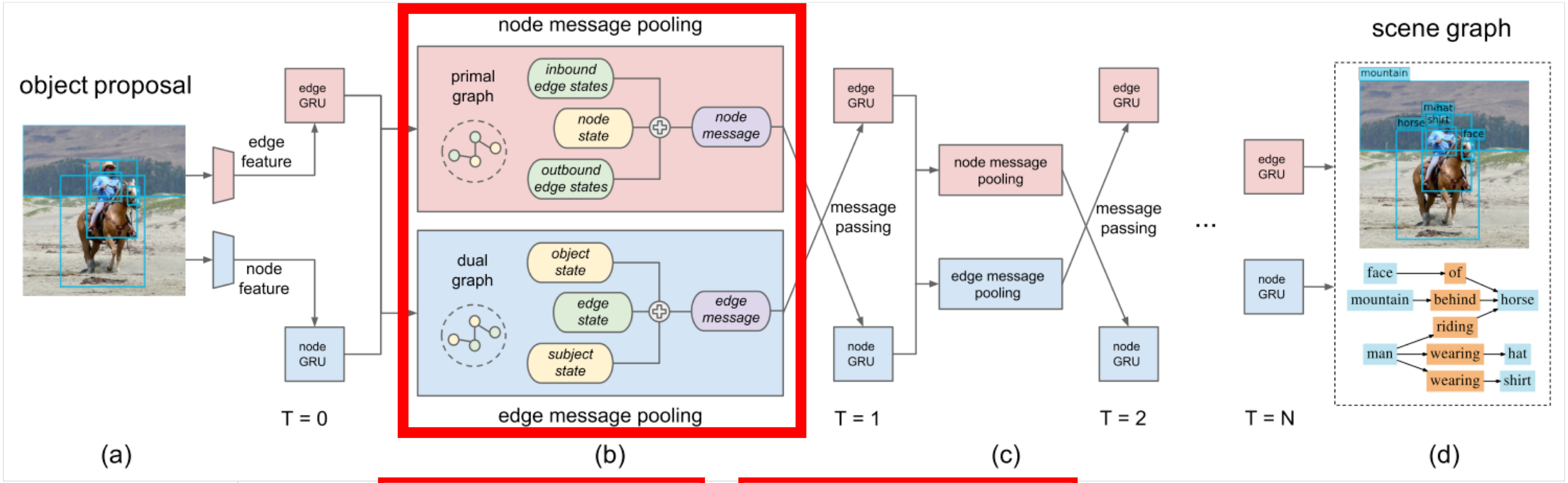
- For our setting, we denote the probability of each variable x as $Q(x|\cdot)$
- Mean field distribution for this setting is then:

$$Q(\mathbf{x}|I, B_I) = \prod_{i=1}^{n} \underbrace{Q(x_i^{cls}, x_i^{bbox}|h_i)Q(h_i|f_i^v)}_{\text{Approximation for nodes (obj)}} \prod_{j \neq i} \underbrace{Q(x_{i \to j}|h_{i \to j})Q(h_{i \to j}|f_{i \to j}^e)}_{\text{Approximation for edges (rel)}}$$

Node/Edge Message Pooling

Outbound edge msg　　　inbound edge msg

$$m_i = \sum_{j:i\to j} \boxed{\sigma(\mathbf{v}_1^T[h_i, h_{i\to j}])h_{i\to j}} + \sum_{j:j\to i} \boxed{\sigma(\mathbf{v}_2^T[h_i, h_{j\to i}])h_{j\to i}}$$

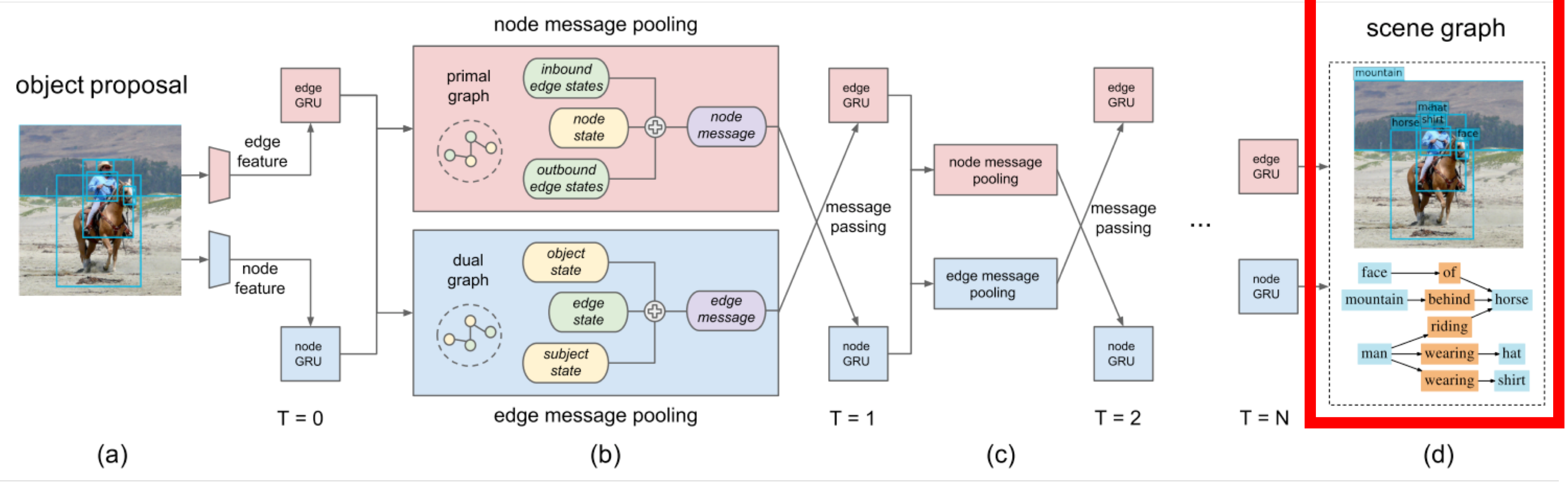


$$m_{i\to j} = \boxed{\sigma(\mathbf{w}_1^T[h_i, h_{i\to j}])h_i} + \boxed{\sigma(\mathbf{w}_2^T[h_j, h_{i\to j}])h_j}$$

Subject node msg　　　Object node msg

Scene Graph Generation by IMP

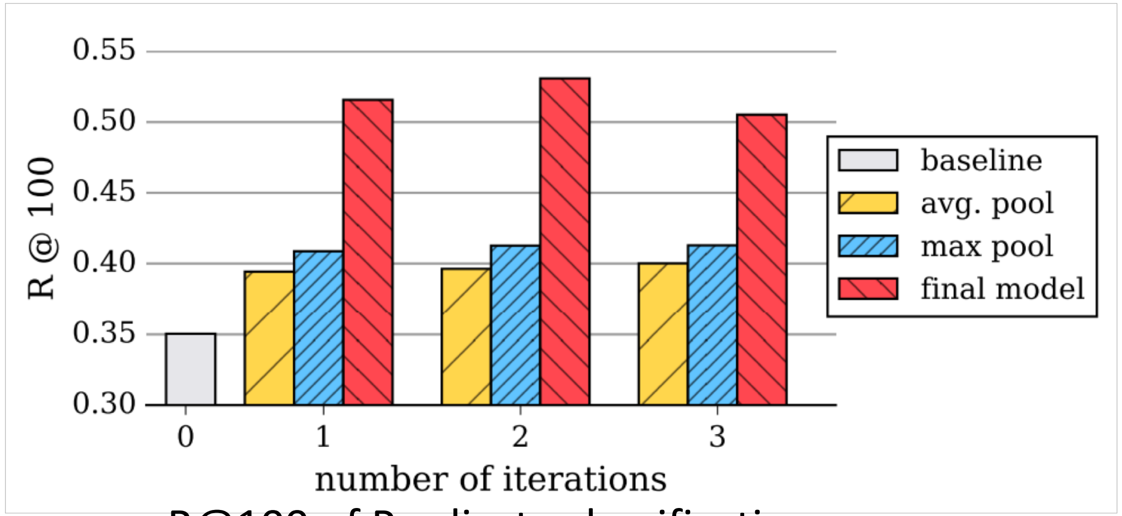
Decoding with
- softmax (labels)
- fc layer (bbox offsets)



Strength and Weakness

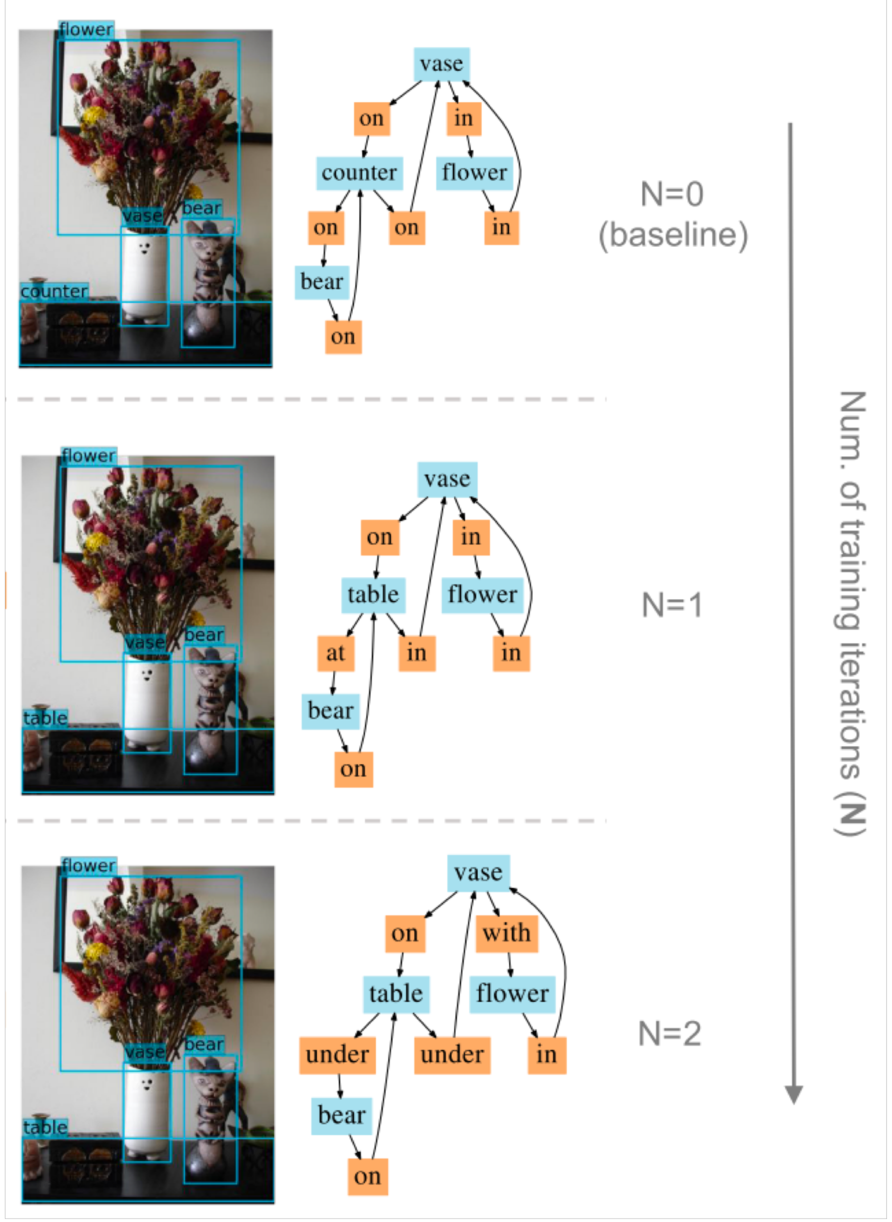
- Exploit the context with graph topology using iterative message passing

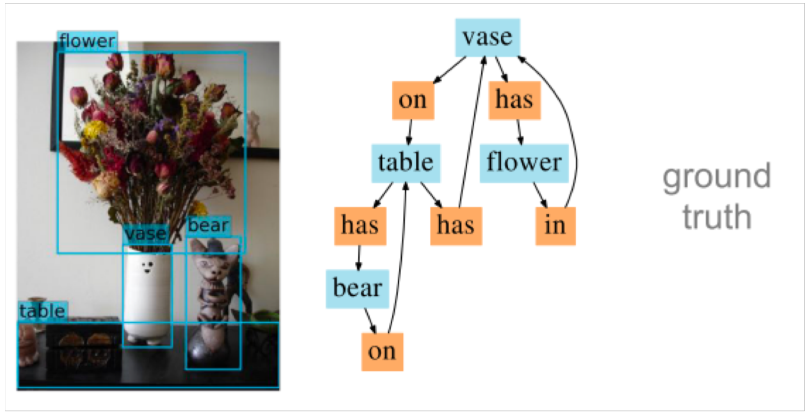
- Model degrades when iterates more than **two round** (noisy message start to permeate through the graph)



R@100 of Predicate classification

Qualitative Result



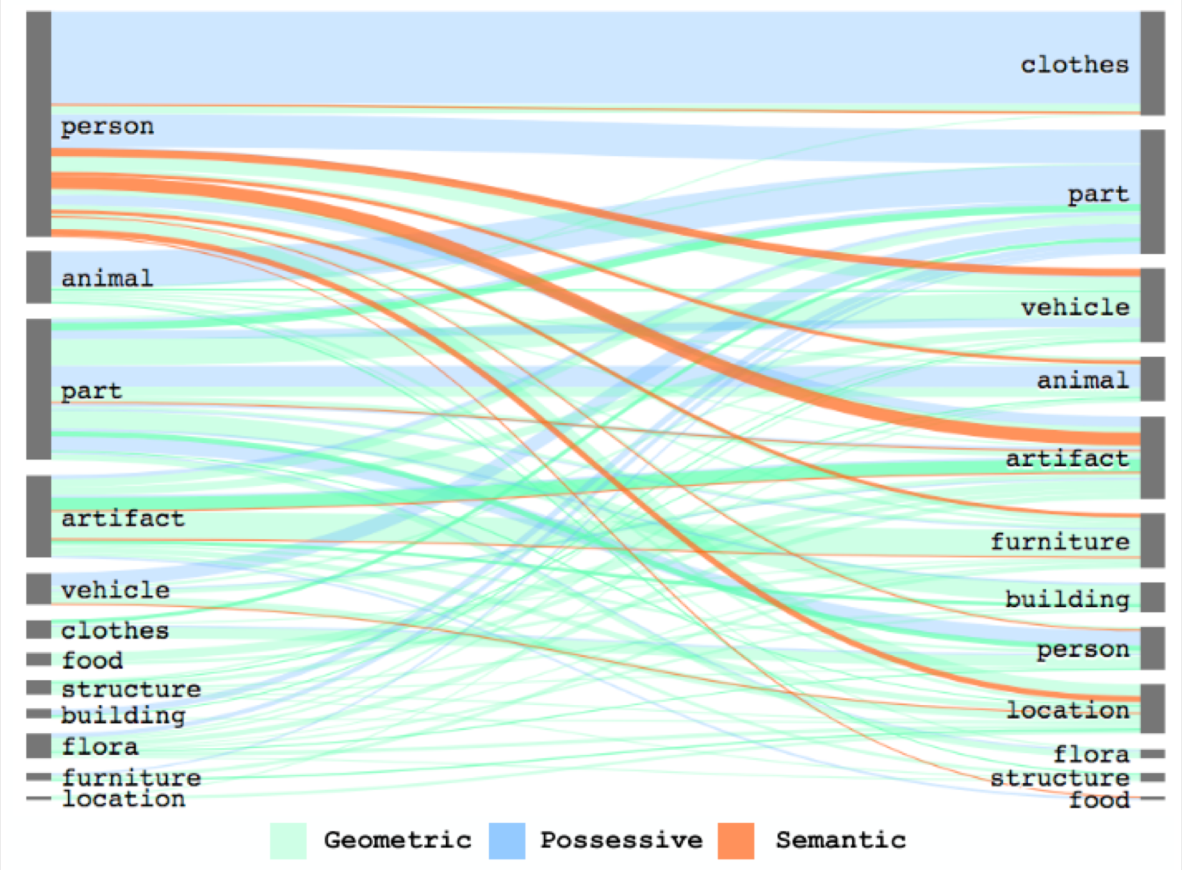


Neural Motifs: Scene Graph Parsing with Global Context

Rowan Zellers[1] Mark Yatskar[1,2] Sam Thomson[3] Yejin Choi[1,2]
[1]Paul G. Allen School of Computer Science & Engineering, University of Washington
[2]Allen Institute for Artificial Intelligence
[3]School of Computer Science, Carnegie Mellon University
{rowanz, my89, yejin}@cs.washington.edu, sthomson@cs.cmu.edu

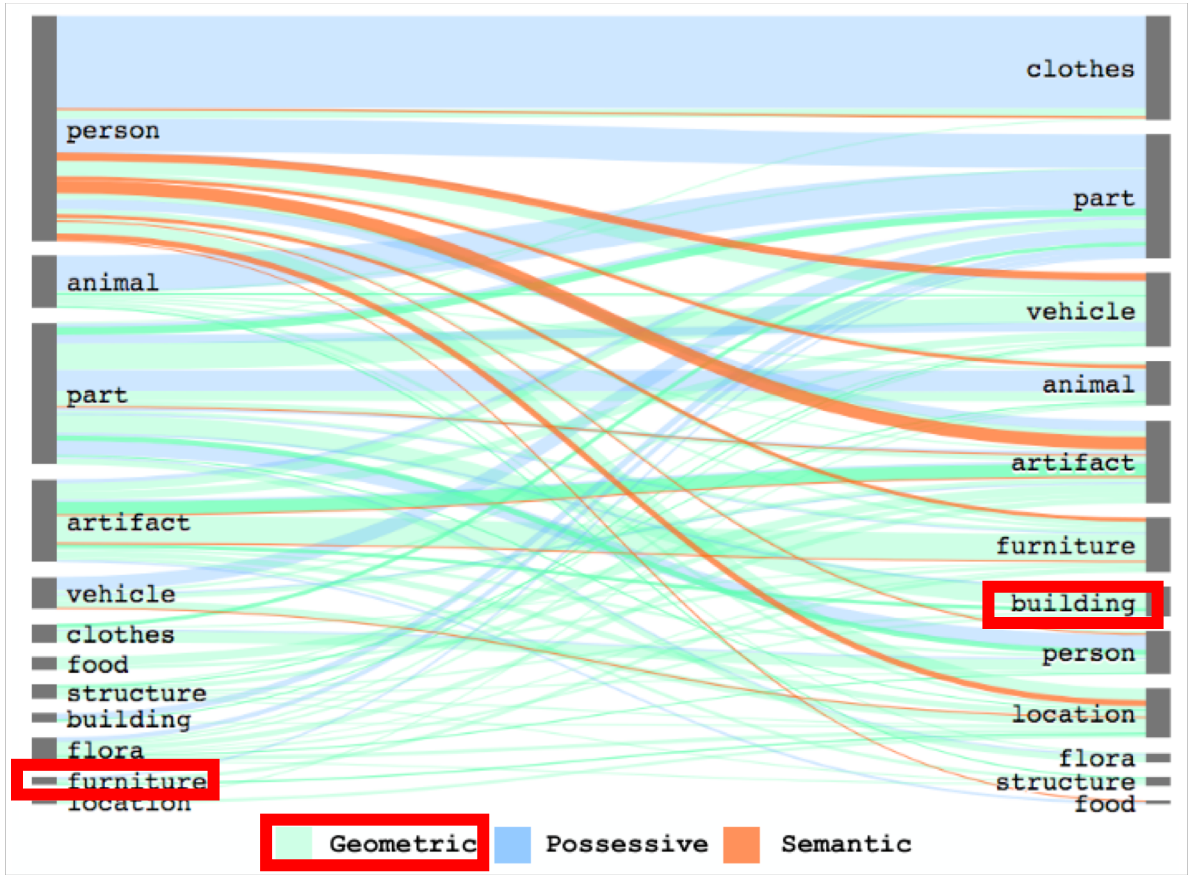
https://rowanzellers.com/neuralmotifs

Visual Genome Dataset Analysis



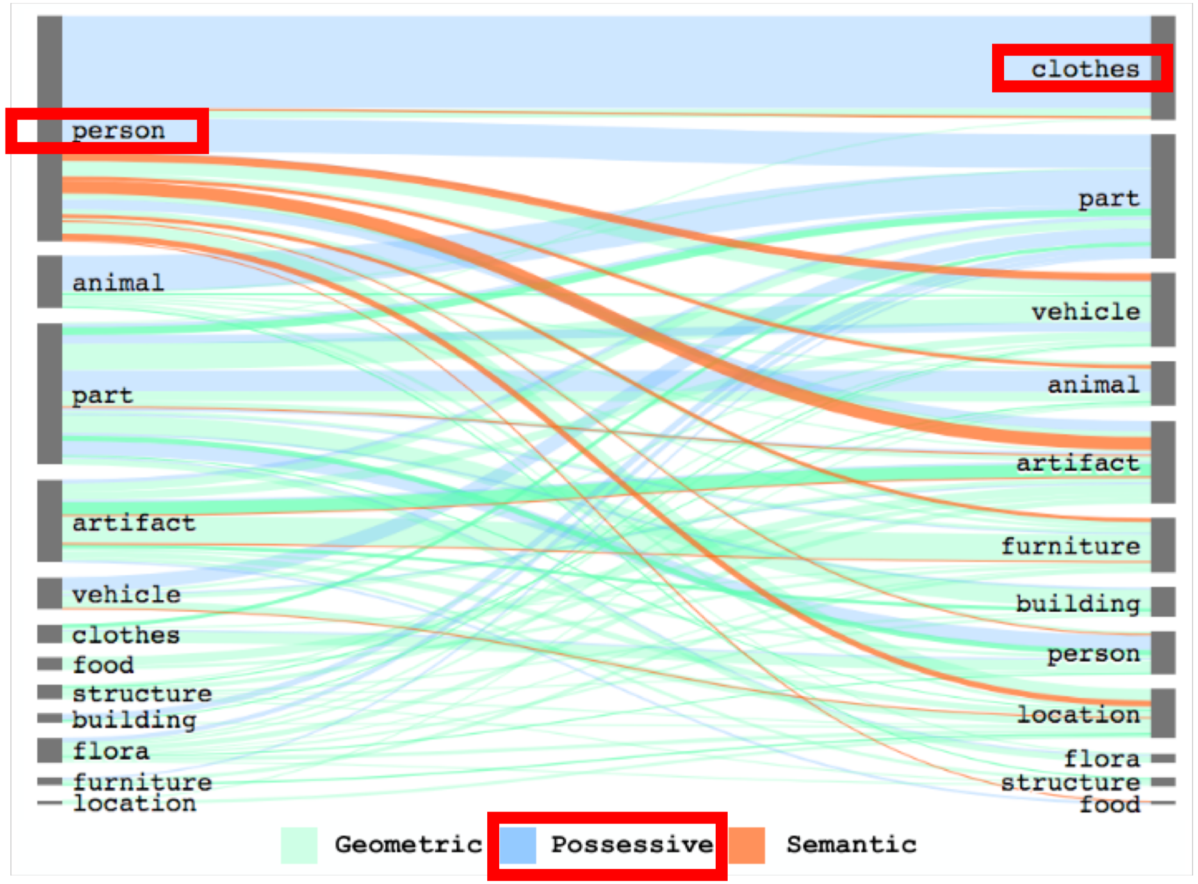
Type	Examples	Classes	Instances
	Entities		
Part	arm, tail, wheel	32	200k (25.2%)
Artifact	basket, fork, towel	34	126k (16.0%)
Person	boy, kid, woman	13	113k (14.3%)
Clothes	cap, jean, sneaker	16	91k (11.5%)
Vehicle	airplane, bike, truck,	12	44k (5.6%)
Flora	flower, plant, tree	3	44k (5.5%)
Location	beach, room, sidewalk	11	39k (4.9%)
Furniture	bed, desk, table	9	37k (4.7%)
Animal	bear, giraffe, zebra	11	30k (3.8%)
Structure	fence, post, sign	3	30k (3.8%)
Building	building, house	2	24k (3.1%)
Food	banana, orange, pizza	6	13k (1.6%)
	Relations		
Geometric	above, behind, under	15	228k (50.0%)
Possessive	has, part of, wearing	8	186k (40.9%)
Semantic	carrying, eating, using	24	39k (8.7%)
Misc	for, from, made of	3	2k (0.3%)

Visual Genome Dataset Analysis



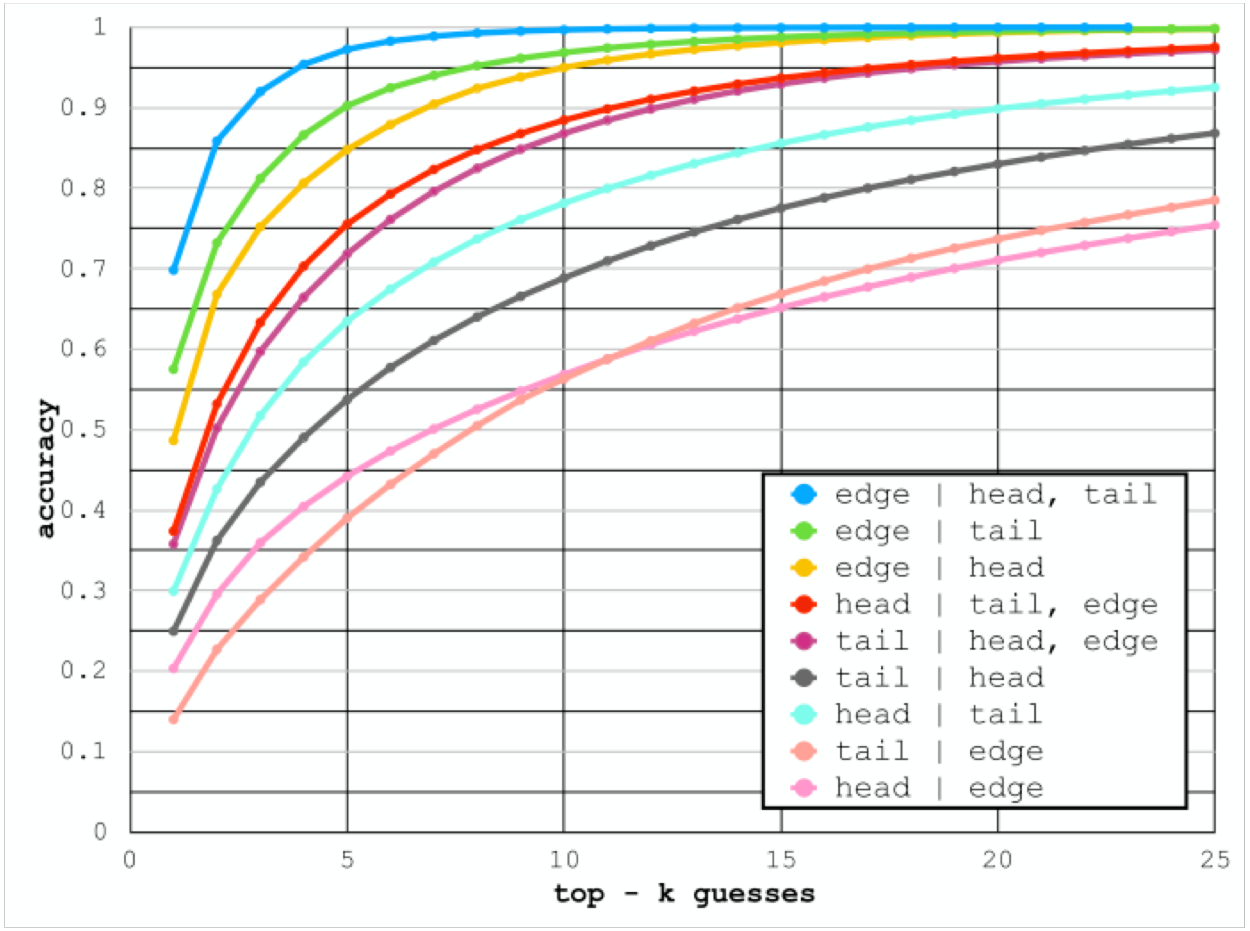
Type	Examples	Classes	Instances
	Entities		
Part	arm, tail, wheel	32	200k (25.2%)
Artifact	basket, fork, towel	34	126k (16.0%)
Person	boy, kid, woman	13	113k (14.3%)
Clothes	cap, jean, sneaker	16	91k (11.5%)
Vehicle	airplane, bike, truck,	12	44k (5.6%)
Flora	flower, plant, tree	3	44k (5.5%)
Location	beach, room, sidewalk	11	39k (4.9%)
Furniture	bed, desk, table	9	37k (4.7%)
Animal	bear, giraffe, zebra	11	30k (3.8%)
Structure	fence, post, sign	3	30k (3.8%)
Building	building, house	2	24k (3.1%)
Food	banana, orange, pizza	6	13k (1.6%)
	Relations		
Geometric	above, behind, under	15	228k (50.0%)
Possessive	has, part of, wearing	8	186k (40.9%)
Semantic	carrying, eating, using	24	39k (8.7%)
Misc	for, from, made of	3	2k (0.3%)

Visual Genome Dataset Analysis



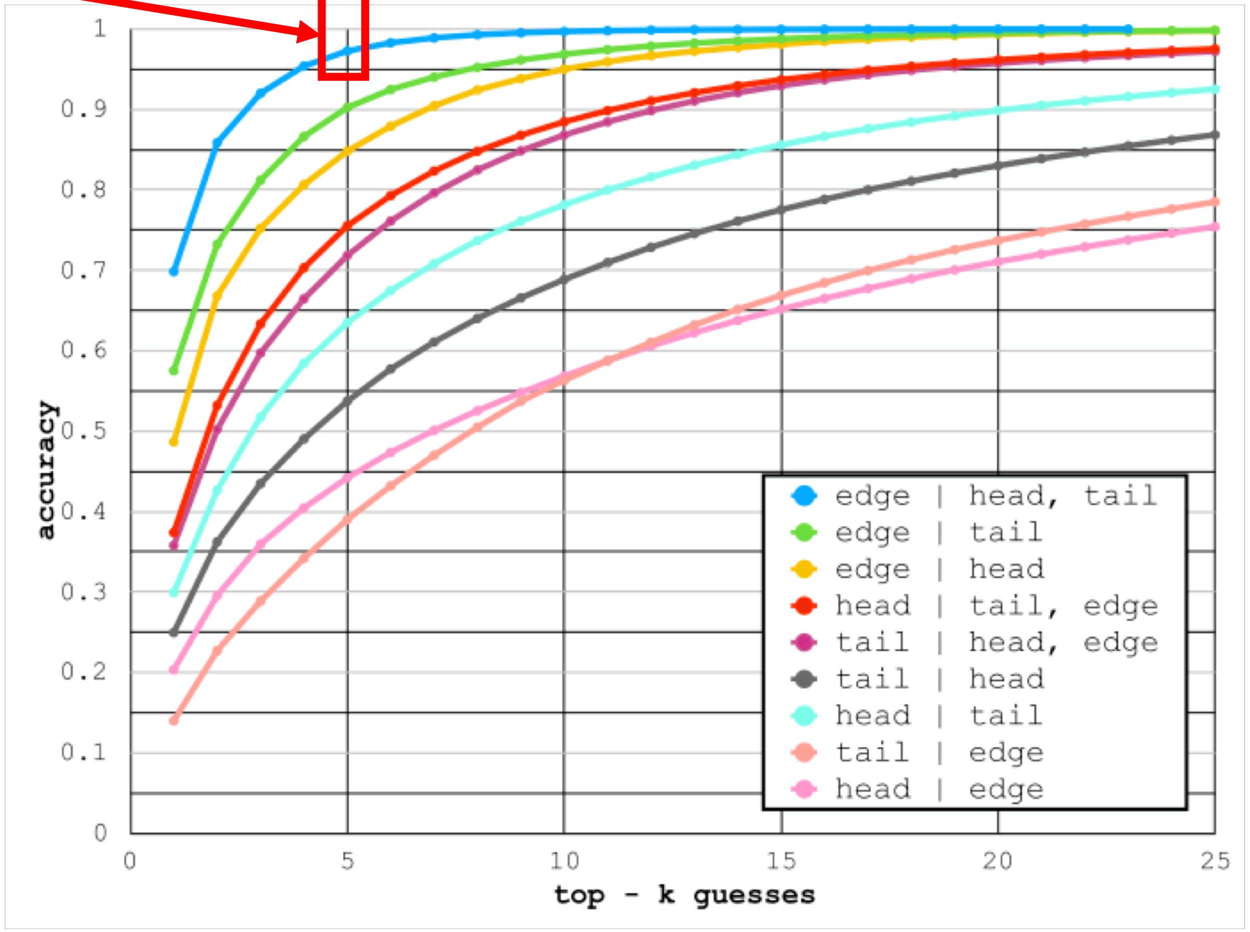
Type	Examples	Classes	Instances
	Entities		
Part	arm, tail, wheel	32	200k (25.2%)
Artifact	basket, fork, towel	34	126k (16.0%)
Person	boy, kid, woman	13	113k (14.3%)
Clothes	cap, jean, sneaker	16	91k (11.5%)
Vehicle	airplane, bike, truck,	12	44k (5.6%)
Flora	flower, plant, tree	3	44k (5.5%)
Location	beach, room, sidewalk	11	39k (4.9%)
Furniture	bed, desk, table	9	37k (4.7%)
Animal	bear, giraffe, zebra	11	30k (3.8%)
Structure	fence, post, sign	3	30k (3.8%)
Building	building, house	2	24k (3.1%)
Food	banana, orange, pizza	6	13k (1.6%)
	Relations		
Geometric	above, behind, under	15	228k (50.0%)
Possessive	has, part of, wearing	8	186k (40.9%)
Semantic	carrying, eating, using	24	39k (8.7%)
Misc	for, from, made of	3	2k (0.3%)

Visual Genome Dataset Analysis



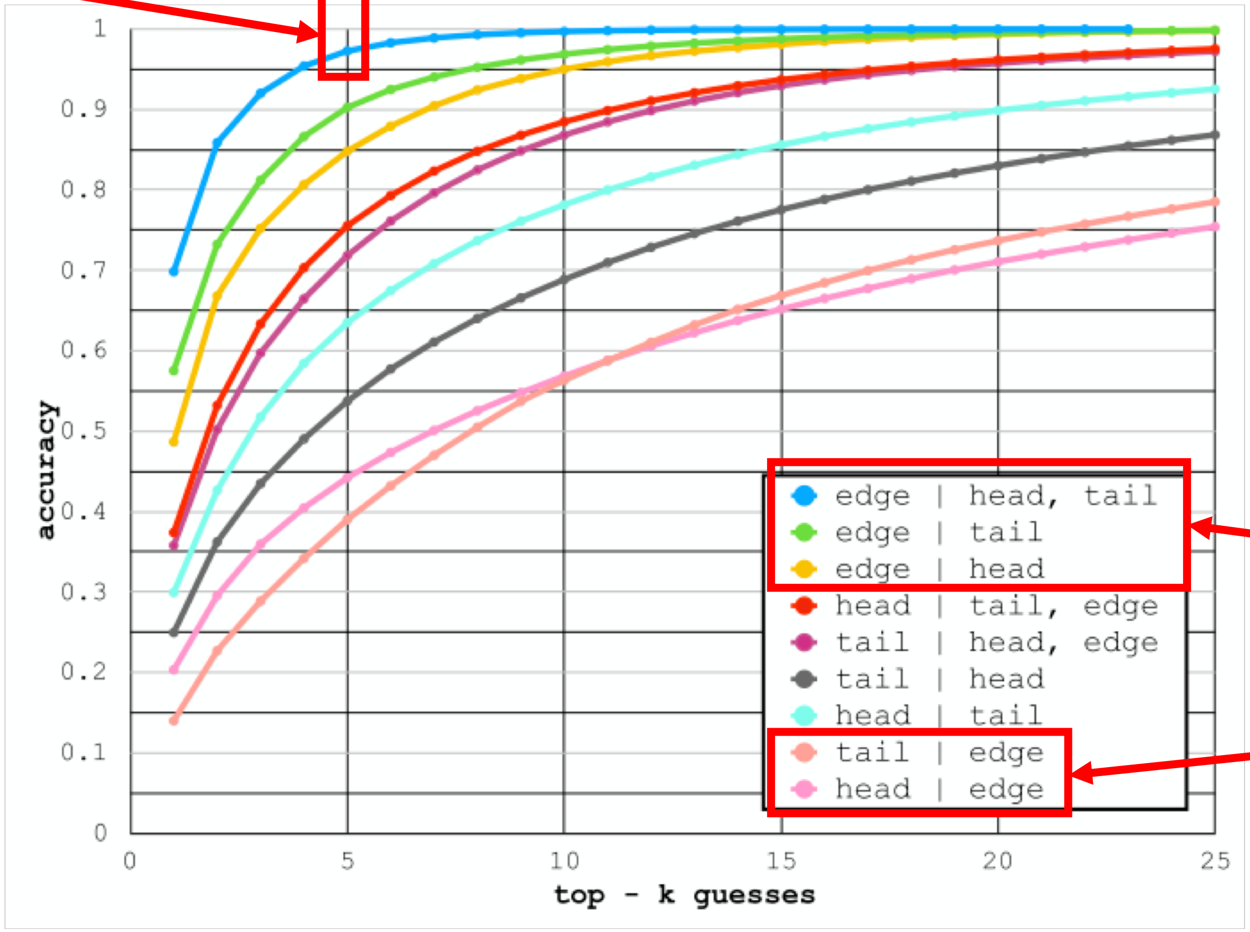
Visual Genome Dataset Analysis

Given head and tail labels, true predicate lies in top-5 guesses **97%** of the time.



Visual Genome Dataset Analysis

Given head and tail labels, true predicate lies in top-5 guesses **97%** of the time.



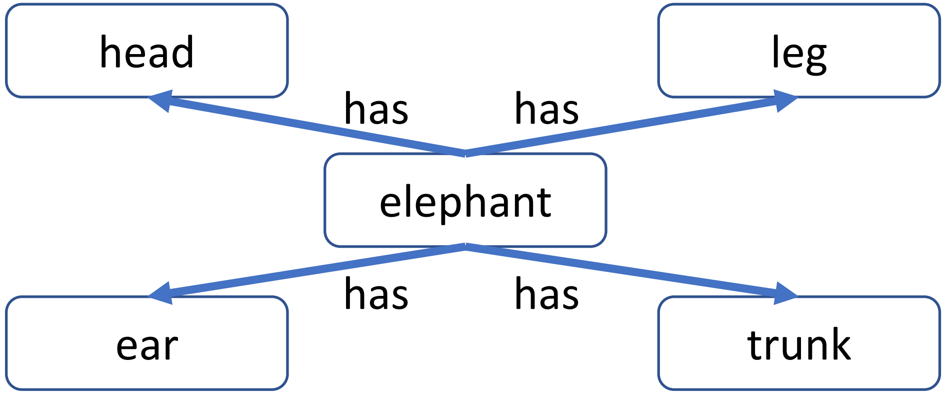
Given head and tails, can infer edges accurately but not vice versa

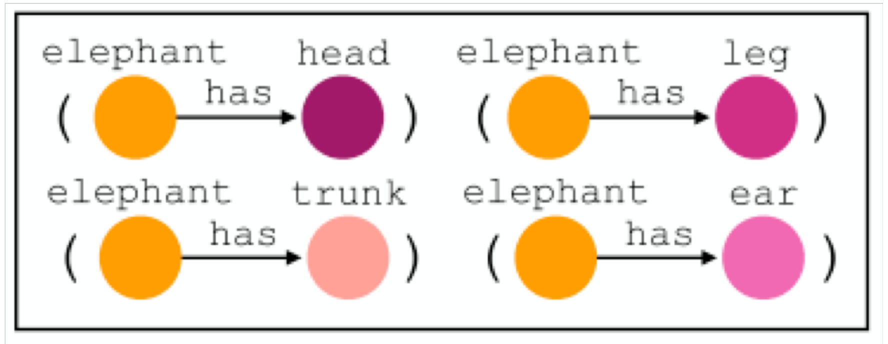
What is Neural Motif?

- *Motif* : (noun [c]) a pattern or design.

What is Neural Motif?

- *Motif* : (noun [c]) a pattern or design.
- Neural motif: repeating higher-order structure in scene graph.





Model

<u>Conditional Probability Chain Rule</u>

- Given Image I and we model graph G = {R, B, O} where R is labeld relations, B is bounding boxes and O is object labels
- Prob of graph $Pr(G|I) = Pr(R, B, O|I)$

$$= Pr(R, O|B, I)\ Pr(B|I)$$

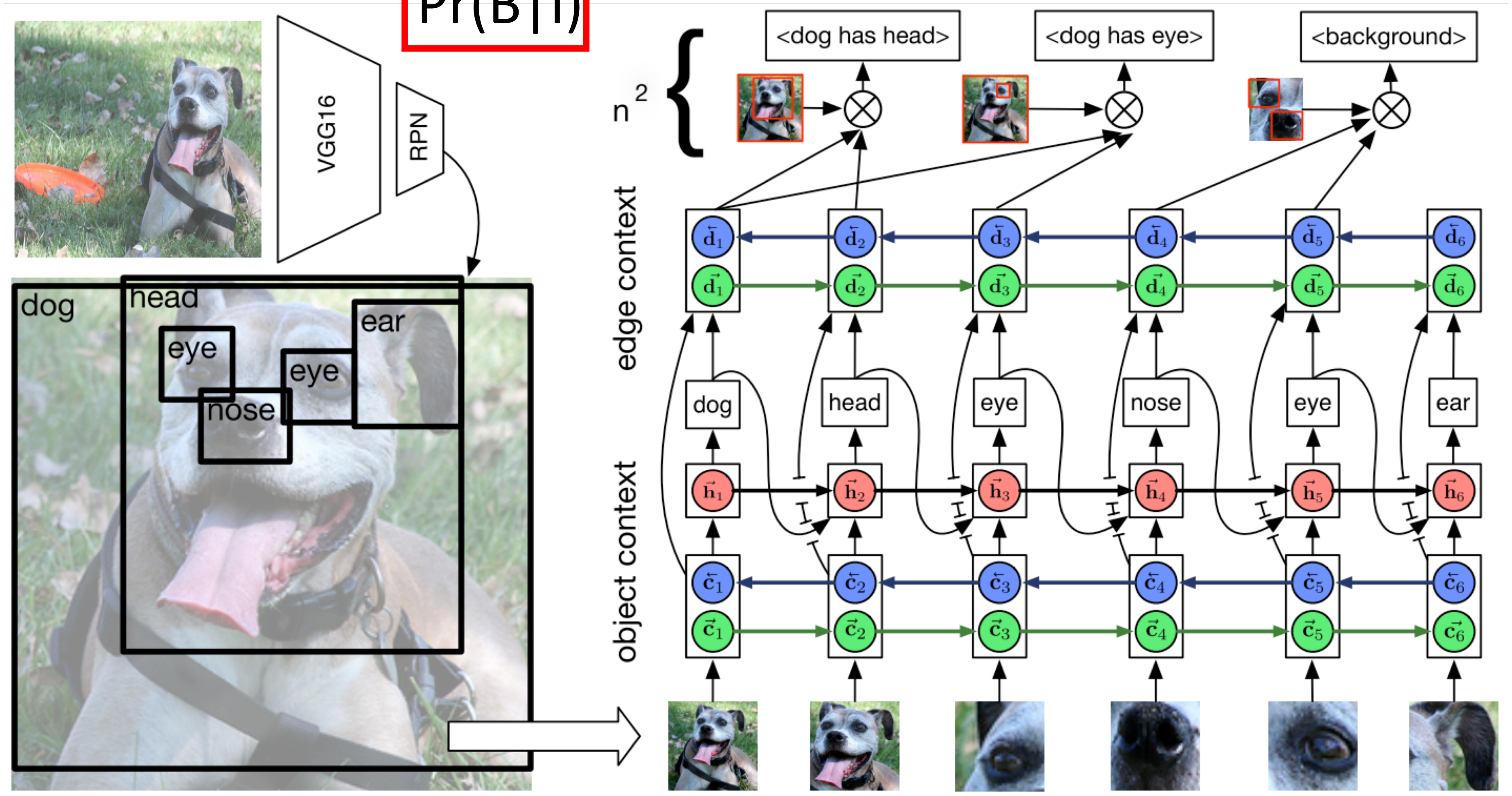
$$= \underbrace{\boxed{Pr(R|B, O, I)}}_{\text{Relation model}}\ \underbrace{\boxed{Pr(O|B, I)}}_{\text{Object model}}\ \underbrace{\boxed{Pr(B|I)}}_{\text{Bounding box model}}$$

Stacked Motif Network

$$\Pr(G|I) = \boxed{\Pr(R|B, O, I)}\,\boxed{\Pr(O|B, I)}\,\boxed{\Pr(B|I)}$$

Bounding box model

$\boxed{\Pr(B|I)}$



Relation model

$\boxed{\Pr(R|B, O, I)}$

$\boxed{\Pr(O|B, I)}$

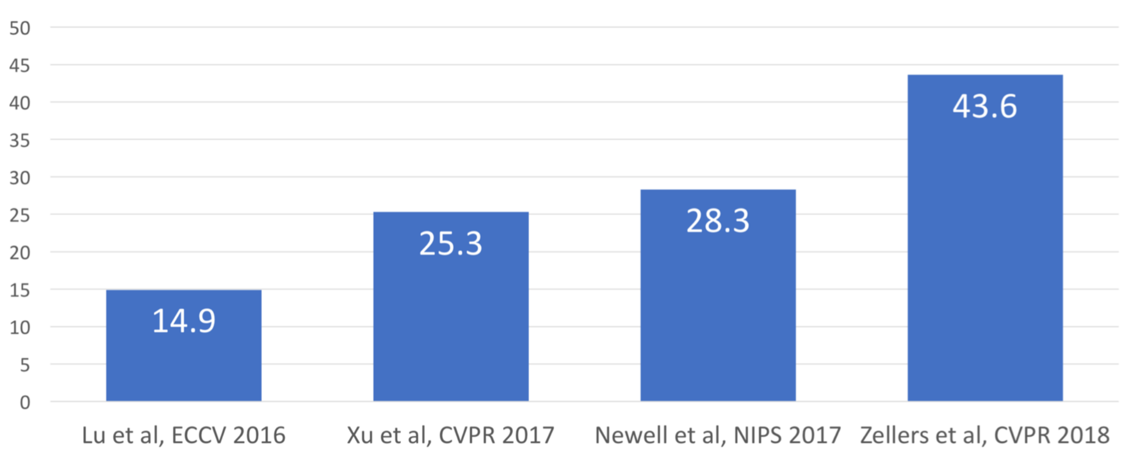
Object model

Strength and Weakness

- This work claims that the current works (and the previous) are only exploiting dataset bias, thus it demonstrates a full power of that bias

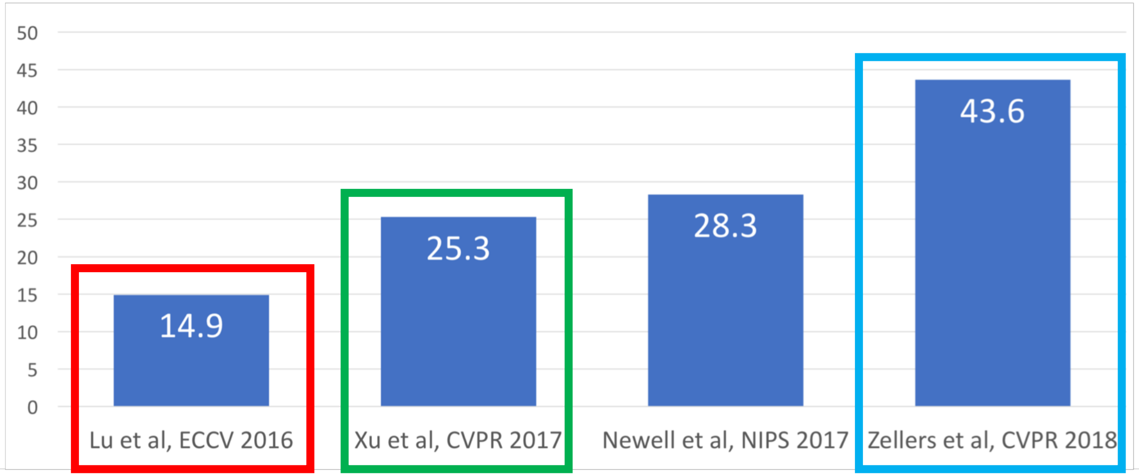
- However cannot see how conditioning on previously decoded object labels help on decoding next label (later in next slide)

Results



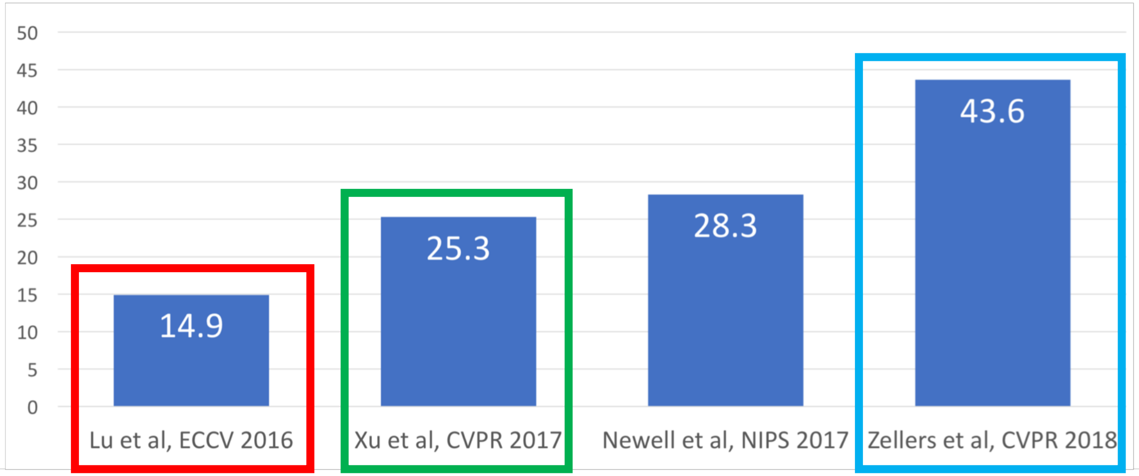
		Scene Graph Detection			Scene Graph Classification			Predicate Classification			Mean
	Model	R@20	R@50	R@100	R@20	R@50	R@100	R@20	R@50	R@100	
models	VRD [29]		0.3	0.5		11.8	14.1		27.9	35.0	14.9
	MESSAGE PASSING [47]		3.4	4.2		21.7	24.4		44.8	53.0	25.3
	MESSAGE PASSING+	14.6	20.7	24.5	31.7	34.6	35.4	52.7	59.3	61.3	39.3
	ASSOC EMBED [31]⋆	6.5	8.1	8.2	18.2	21.8	22.6	47.9	54.1	55.4	28.3
	FREQ	17.7	23.5	27.6	27.7	32.4	34.0	49.4	59.9	64.1	40.2
	FREQ+OVERLAP	20.1	26.2	30.1	29.3	32.3	32.9	53.6	60.6	62.2	40.7
	MOTIFNET-LEFTRIGHT	21.4	27.2	30.3	**32.9**	**35.8**	**36.5**	**58.5**	**65.2**	**67.1**	**43.6**
ablations	MOTIFNET-NOCONTEXT	21.0	26.2	29.0	31.9	34.8	35.5	57.0	63.7	65.6	42.4
	MOTIFNET-CONFIDENCE	**21.7**	**27.3**	**30.5**	32.6	35.4	36.1	58.2	65.1	67.0	43.5
	MOTIFNET-SIZE	21.6	**27.3**	30.4	32.2	35.0	35.7	58.0	64.9	66.8	43.3
	MOTIFNET-RANDOM	21.6	**27.3**	30.4	32.5	35.5	36.2	58.1	65.1	66.9	43.5

Results



		Scene Graph Detection			Scene Graph Classification			Predicate Classification			Mean
	Model	R@20	R@50	R@100	R@20	R@50	R@100	R@20	R@50	R@100	
models	VRD [29]		0.3	0.5		11.8	14.1		27.9	35.0	14.9
	MESSAGE PASSING [47]		3.4	4.2		21.7	24.4		44.8	53.0	25.3
	MESSAGE PASSING+	14.6	20.7	24.5	31.7	34.6	35.4	52.7	59.3	61.3	39.3
	ASSOC EMBED [31]⋆	6.5	8.1	8.2	18.2	21.8	22.6	47.9	54.1	55.4	28.3
	FREQ	17.7	23.5	27.6	27.7	32.4	34.0	49.4	59.9	64.1	40.2
	FREQ+OVERLAP	20.1	26.2	30.1	29.3	32.3	32.9	53.6	60.6	62.2	40.7
	MOTIFNET-LEFTRIGHT	21.4	27.2	30.3	**32.9**	**35.8**	**36.5**	**58.5**	**65.2**	**67.1**	**43.6**
ablations	MOTIFNET-NOCONTEXT	21.0	26.2	29.0	31.9	34.8	35.5	57.0	63.7	65.6	42.4
	MOTIFNET-CONFIDENCE	**21.7**	**27.3**	**30.5**	32.6	35.4	36.1	58.2	65.1	67.0	43.5
	MOTIFNET-SIZE	21.6	**27.3**	30.4	32.2	35.0	35.7	58.0	64.9	66.8	43.3
	MOTIFNET-RANDOM	21.6	**27.3**	30.4	32.5	35.5	36.2	58.1	65.1	66.9	43.5

Results



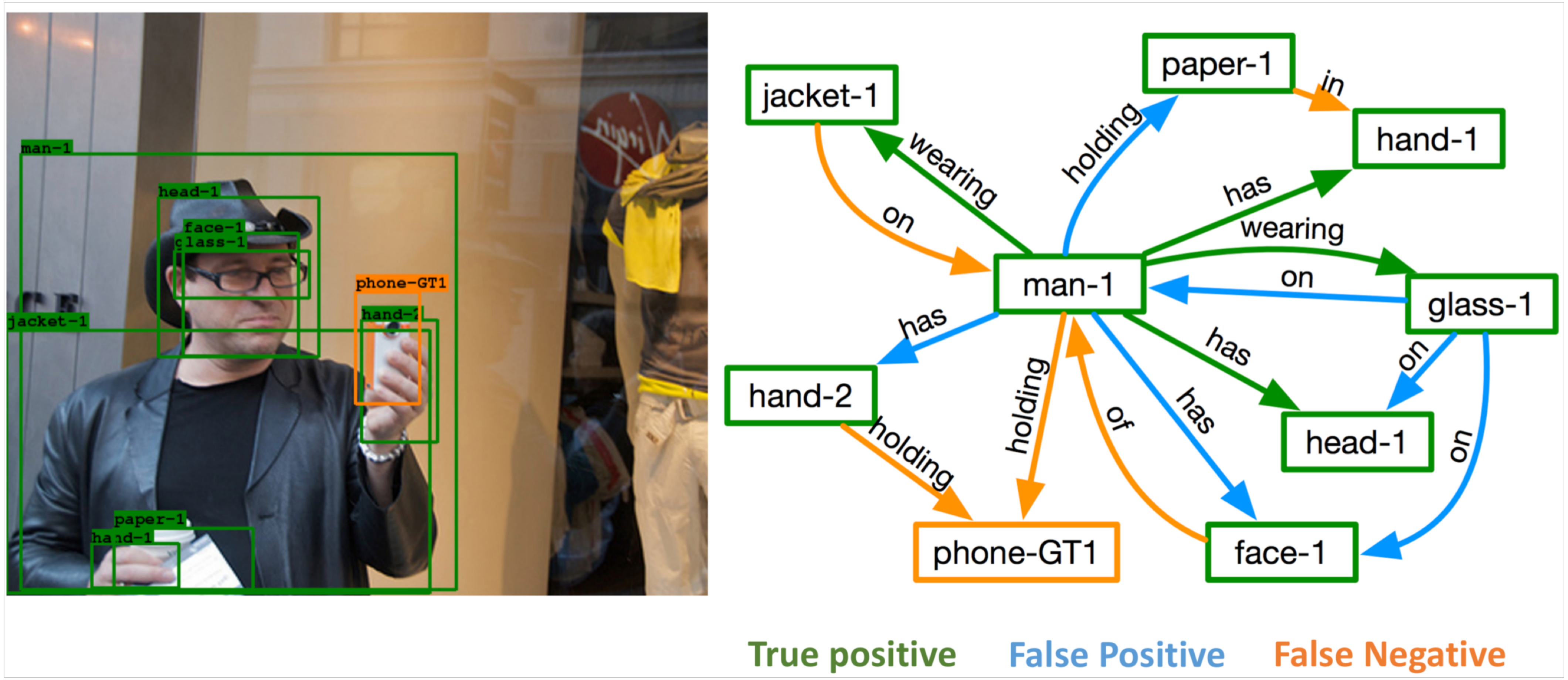
	Model	Scene Graph Detection			Scene Graph Classification			Predicate Classification			Mean
		[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	R@100	R@20	R@50	R@100	
models	VRD [29]		0.3	0.5		11.8	14.1		27.9	35.0	14.9
models	Message Passing [47]		3.4	4.2		21.7	24.4		44.8	53.0	25.3
models	Message Passing+	14.6	20.7	24.5	31.7	34.6	35.4	52.7	59.3	61.3	39.3
models	Assoc Embed [31]⋆	[illegible]	[illegible]	[illegible]	[illegible]	21.8	22.6	47.9	54.1	55.4	28.3
models	Freq	[illegible]	[illegible]	[illegible]	[illegible]	32.4	34.0	49.4	59.9	64.1	40.2
models	Freq+Overlap	20.1	26.2	30.1	29.3	32.3	32.9	53.6	60.6	62.2	40.7
models	MotifNet-LeftRight	21.4	27.2	30.3	**32.9**	**35.8**	**36.5**	**58.5**	**65.2**	**67.1**	**43.6**
ablations	MotifNet-NoCont[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	57.0	63.7	65.6	42.4
ablations	MotifNet-Confide[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	58.2	65.1	67.0	43.5
ablations	MotifNet-Size	21.6	**27.3**	30.4	32.2	35.0	35.7	58.0	64.9	66.8	43.3
ablations	MotifNet-Random	21.6	**27.3**	30.4	32.5	35.5	36.2	58.1	65.1	66.9	43.5

Independent relationship prediction

Jointly predict entire graph

Fully exploit dataset bias with “neural motifs”

Qualitative result (Neural Motifs)



References and acknowledgement

- [1] Sadeghi, M.A., Farhadi, A.: Recognition using visual phrases. In: Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on, IEEE (2011) 1745– 1752

- Several slides credit Justin Johnson's talk in *CVPR 2018 Tutorial on Visual Recognition and Beyond*. https://drive.google.com/open?id=1dG3F6OObF8-ppAlrlE3KWZ0i4YAQ5Uka
- Some pictures come from Google Image search are only for illustration.

Thank you for the attention! ☺

Any questions?