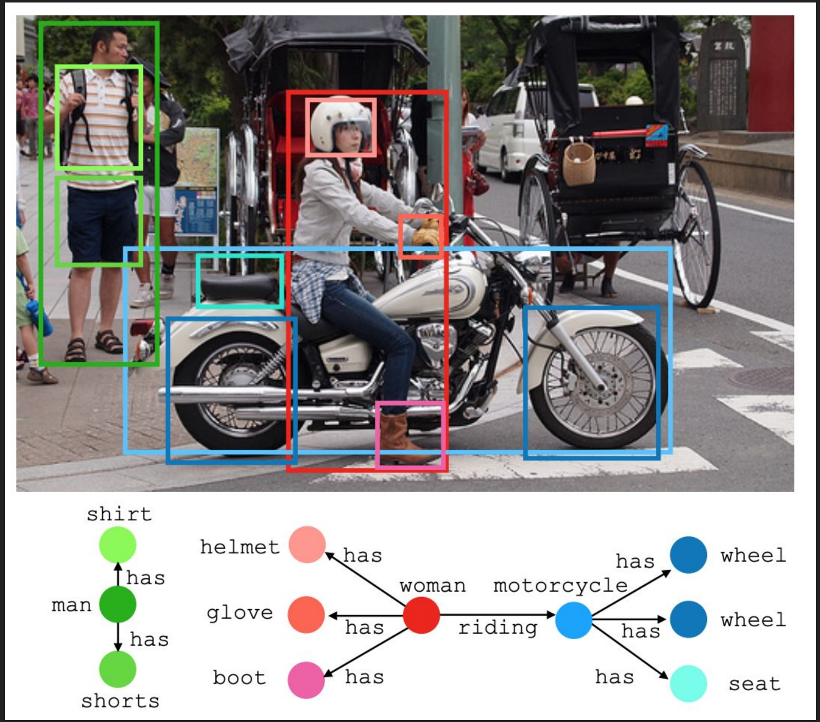
Neural Motifs: Scene Graph Parsing with Global Context

Rowan Zellers, Mark Yatskar, Sam Thomson, Yejin Choi
Presented by: Ying Shen

Scene Graph

A structured representation of the semantic content of an image

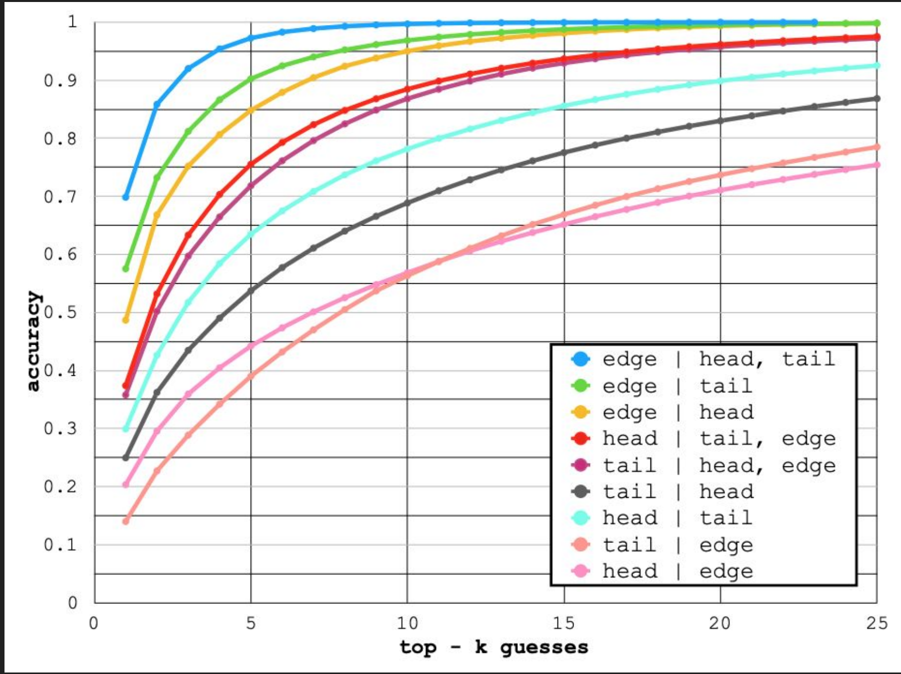
- A set of **bounding boxes**
- A corresponding set of **objects**
- A set of of **binary relationships** between those objects



Zellers, Rowan, et al. "Neural motifs: Scene graph parsing with global context." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018.

Scene Graph Analysis

Prevalent Relations in Visual Genome

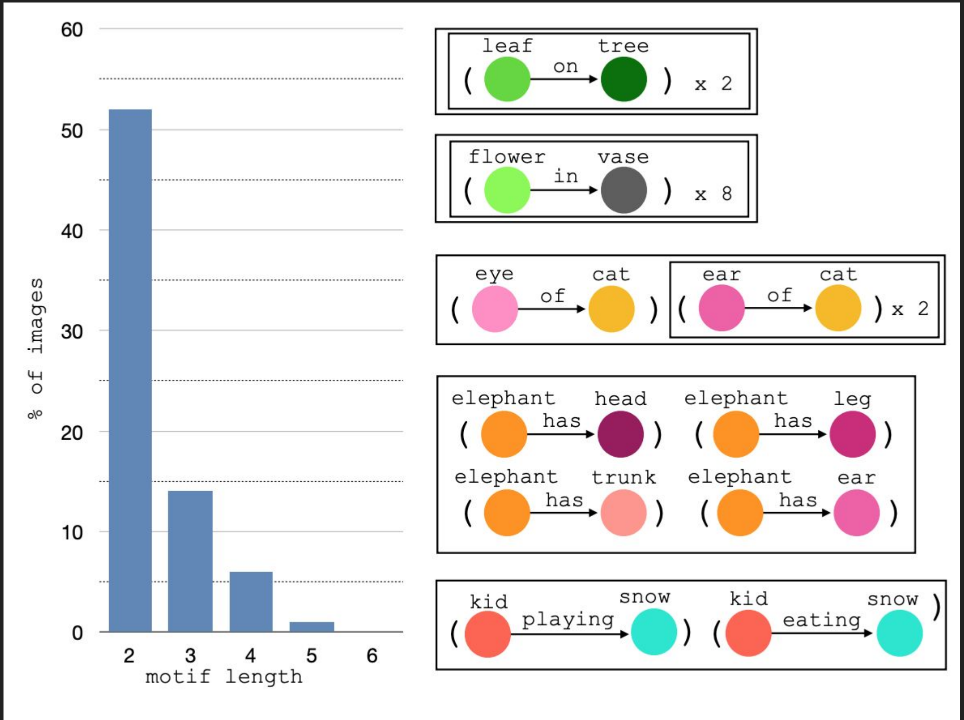


Object labels are highly predictive of relation labels BUT not vice-versa

- (edge | head, tail) is correct **70%** of the time in top-1 guess.
- (edge | head, tail) can be determined with **97%** accuracy in under 5 guesses.

Zellers, Rowan, et al. "Neural motifs: Scene graph parsing with global context." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018.

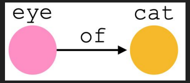
Lager Motifs



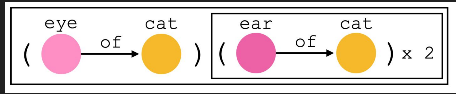
Structure patterns exist in larger subgraphs

- Over 50% graphs contain motifs involving at least two relations

O-R-O:



Motif:



Regularly appearing substructures

Zellers, Rowan, et al. "Neural motifs: Scene graph parsing with global context." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018.

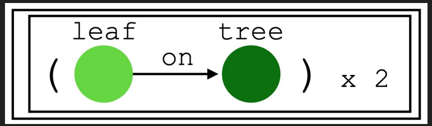
Stacked Motif Networks

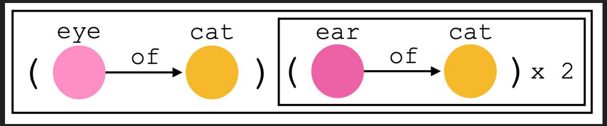
Stacked Motif Networks

- Breaks scene graph parsing into **stages**:
  A. predicting bounding regions Pr(B | I)
  B. predicting labels for regions Pr(O | B, I)
  C. predicting relationships Pr(R | B, O, I)

$$\Pr(G \mid I) = \Pr(B \mid I)\Pr(O \mid B, I)\Pr(R \mid B, O, I).$$

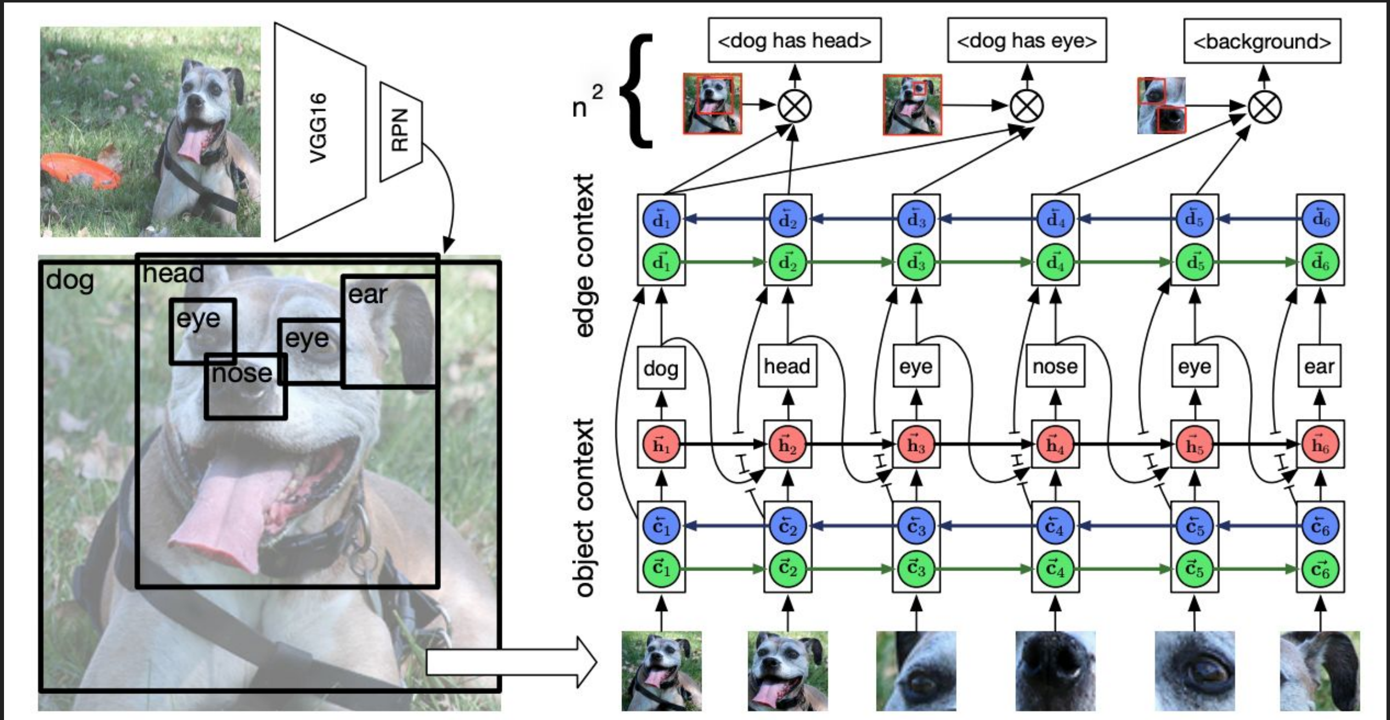
- Encode the **global context** that can directly inform the local predictors





Zellers, Rowan, et al. "Neural motifs: Scene graph parsing with global context." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018.

Stacked Motif Networks



Zellers, Rowan, et al. "Neural motifs: Scene graph parsing with global context." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018.

Frequency Baselines

FREQ:

- Given object detections with labels, predict the most frequent relation between object pairs **without** visual cues.

FREQ-OVERLAP

- Only predict the relationships where there are overlap between the two boxes

Zellers, Rowan, et al. "Neural motifs: Scene graph parsing with global context." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018.

Results

	Model	Scene Graph Detection			Scene Graph Classification			Predicate Classification			Mean
		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

Zellers, Rowan, et al. "Neural motifs: Scene graph parsing with global context." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018.

Discussion Points

- Long tail distribution of relationships [1]
  - dog-ride-skateboard (common) v.s. dog-ride-sufboard (rare)
- Helps from other modalities? (E.g. captions?)
- What makes a good baseline?

[1] Lu, Cewu, et al. "Visual relationship detection with language priors." European Conference on Computer Vision. Springer, Cham, 2016.

Ording of the boxing box

Thank you!