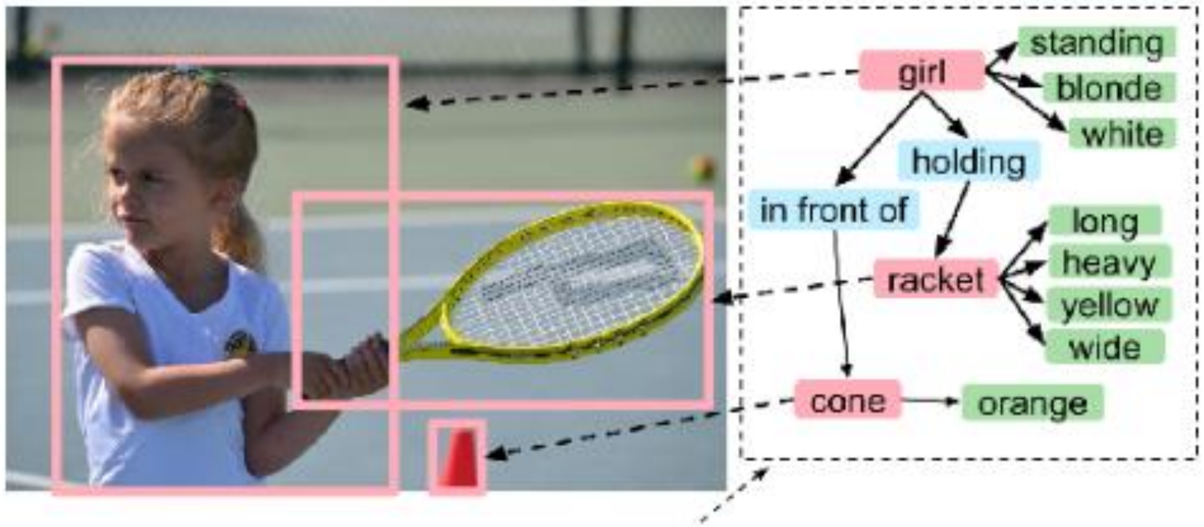
Efficient Graph Features Refinement

张玉杰

2021年4月14日

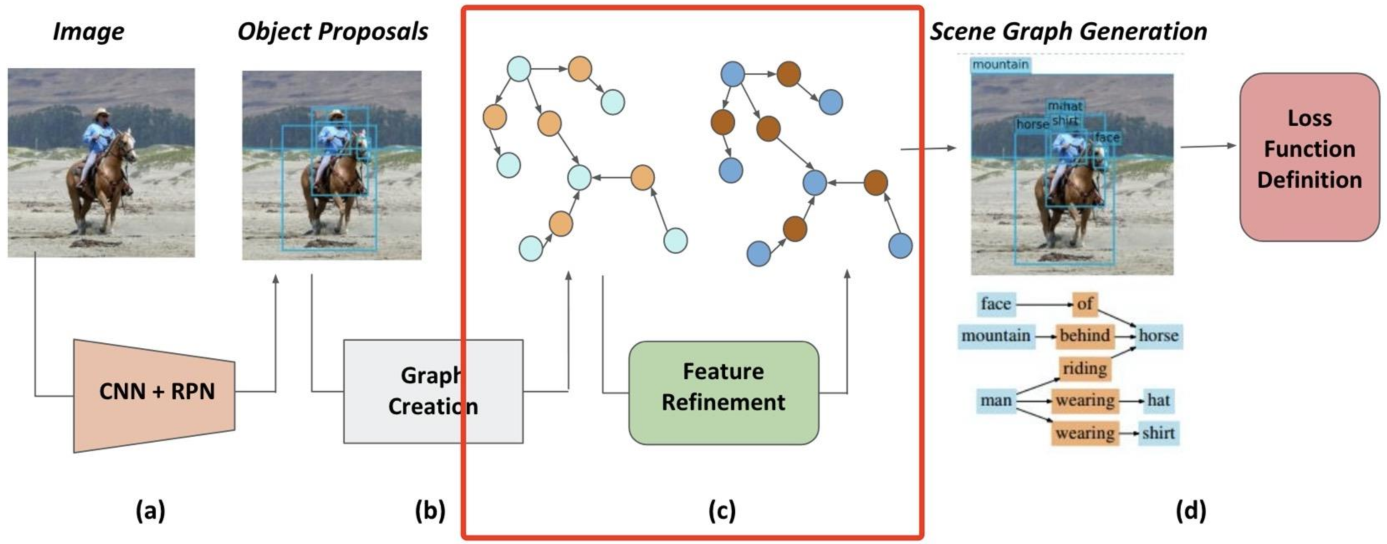
background



- Scene Graph

A Scene Graph is a graphical data structure that describes the contents of a scene. A scene graph encodes object instances, attributes of objects, and relationships between objects.

background



- Feature Refining

The idea is to incorporate contextual information either explicitly or implicitly so that the detection process for objects and relations becomes more context-dependent. The intuition behind feature refining is the superior dependencies among <subject-predicate-object> triplet.

eg. If one object is "boy" and the other is "shirt", there is a high chance of "wear" as a predicate.

- Deep Relational Networks
- Iterative Message Passing
- MSDN
- Graph R-CNN
- Neural Motifs
- VCTree
- UVTransE
- External Knowledge and Image Reconstruction
- Global-Local Attention Transformer(GLAT)
- GB-Net

the spatial and statistical features in a scene

global context

structure and hierarchy

external knowledge and commonsense

Deep Relational Networks

Task

The task is to locate all visual relationships from a given image, and infer the triplets(subject, predicate, object).

Motivation

1. different combinations of objects and relationship predicates as different classes → an extremely large number of imbalanced classes

2. consider each type of relationship predicates as a class → the substantially increased diversity within each class.

Contribution

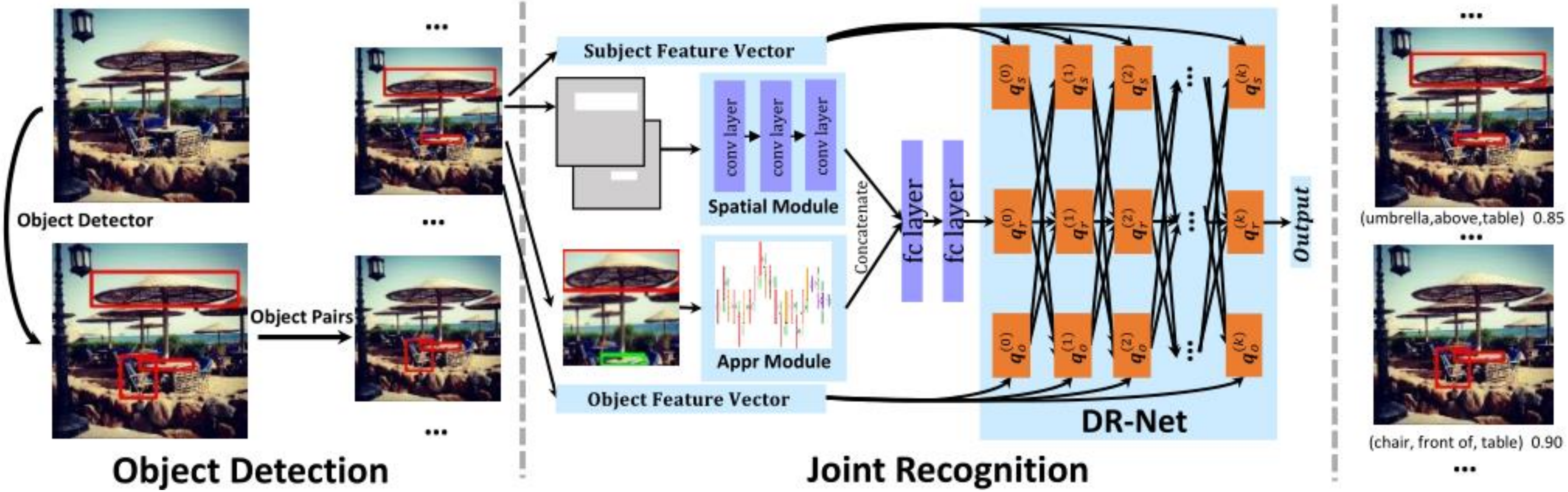
(1) DR-Net, a novel formulation that combines the strengths of statistical models and deep learning

(2) an effective framework for visual relationship detection

Bo Dai, Yuqi Zhang, and Dahua Lin. “Detecting visual relationships with deep relational networks”. In:Proceedings of the IEEE conference on computer vision and Pattern recognition.2017, pp. 3076–3086.

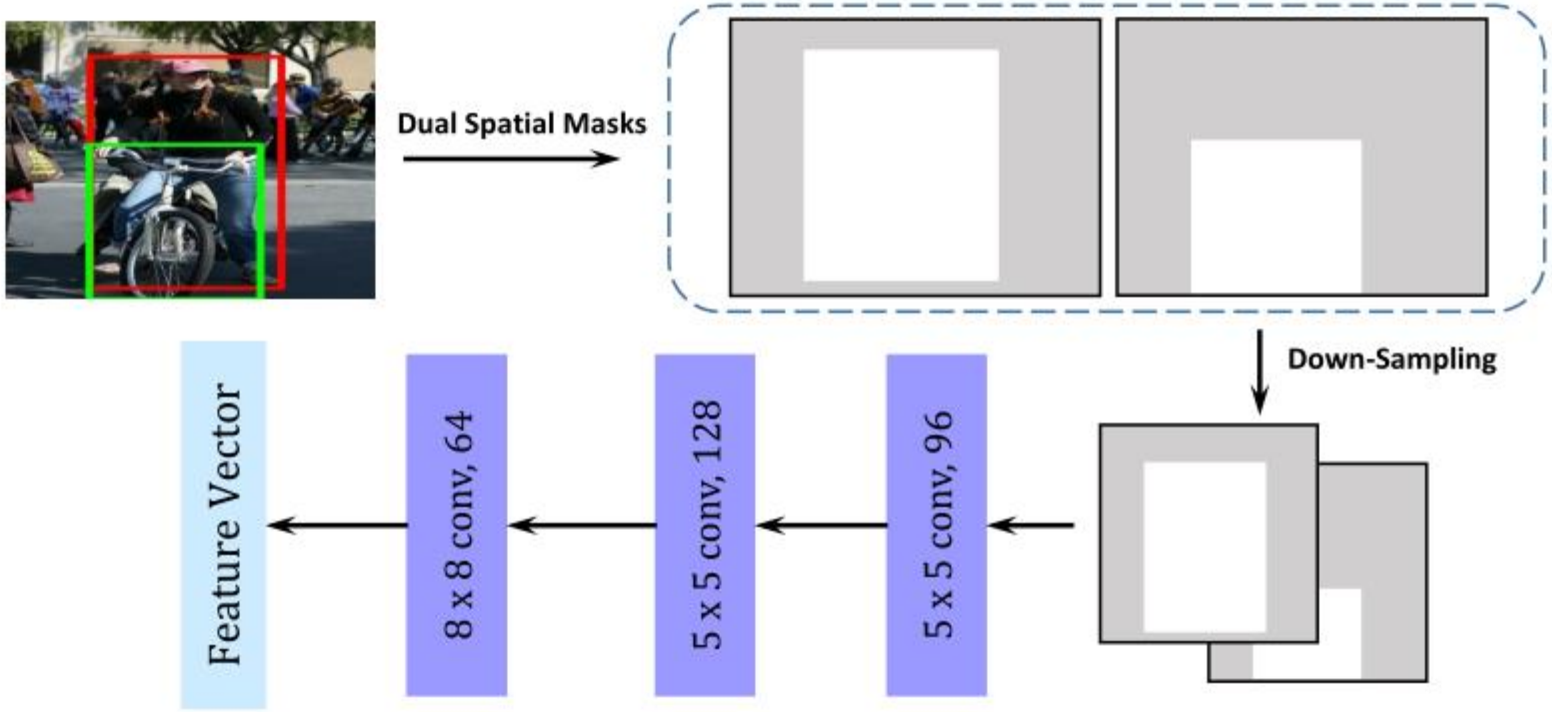
Deep Relational Networks

- The framework for visual relationship detection



Bo Dai, Yuqi Zhang, and Dahua Lin. "Detecting visual relationships with deep relational networks". In:Proceedings of the IEEE conference on computer vision and Pattern recognition.2017, pp. 3076–3086.

Deep Relational Networks



This figure illustrates the process of spatial feature vector generation.

Bo Dai, Yuqi Zhang, and Dahua Lin. "Detecting visual relationships with deep relational networks". In:Proceedings of the IEEE conference on computer vision and Pattern recognition.2017, pp. 3076–3086.

Deep Relational Networks

$$q'_s = \sigma(W_a x_s + W_{sr} q_r + W_{so} q_o),$$
$$q'_r = \sigma(W_r x_r + W_{rs} q_s + W_{ro} q_o),$$
$$q'_o = \sigma(W_a x_o + W_{os} q_s + W_{or} q_r).$$

s : subject

r : relationship

o : object

q : the probability distribution over the classes for features

q' : the updated probabilities

x : the initial feature vectors

σ : the softmax activation

W : weight vectors

Bo Dai, Yuqi Zhang, and Dahua Lin. "Detecting visual relationships with deep relational networks". In:Proceedings of the IEEE conference on computer vision and Pattern recognition.2017, pp. 3076–3086.

Deep Relational Networks

datasets

VRD： 5,000 images | 37,993 visual relationship instances | 6,672triplet types

sVG： 108K images | 998K relationship instances | 74,361 triplet types

		Predicate Recognition		Union Box Detection		Two Boxes Detection	
		Recall@50	Recall@100	Recall@50	Recall@100	Recall@50	Recall@100
VRD	VP [6]	0.97	1.91	0.04	0.07	-	-
	Joint-CNN [49]	1.47	2.03	0.07	0.09	0.07	0.09
	VR [1]	47.87	47.87	16.17	17.03	13.86	14.70
	DR-Net	**80.78**	**81.90**	19.02	22.85	16.94	20.20
	DR-Net + pair filter	-	-	**19.93**	**23.45**	**17.73**	**20.88**
sVG	VP [6]	0.63	0.87	0.01	0.01	-	-
	Joint-CNN [49]	3.06	3.99	1.24	1.60	1.21	1.58
	VR [1]	53.49	54.05	13.80	17.39	11.79	14.84
	DR-Net	**88.26**	**91.26**	20.28	25.74	17.51	22.23
	DR-Net + pair filter	-	-	**23.95**	**27.57**	**20.79**	**23.76**

		A_1	A_2	S	A_1S	A_1SC	A_1SD	A_2SD	A_2SDF
VRD	Predicate Recognition	63.39	65.93	64.72	71.81	72.77	80.66	**80.78**	-
	Union Box Detection	12.01	12.56	13.76	16.04	16.37	18.15	**19.02**	**19.93**
	Two Boxes Detection	10.71	11.22	12.16	14.38	14.66	16.12	**16.94**	**17.73**
sVG	Predicate Recognition	72.13	72.54	75.18	79.10	79.18	88.00	**88.26**	-
	Union Box Detection	13.24	13.84	14.01	16.04	16.08	20.21	**20.28**	**23.95**
	Two Boxes Detection	11.35	11.98	12.07	13.77	13.81	17.42	**17.51**	**20.79**

F： Pair Filter

A： Appearance Module

A1： VGG16， A2： ResNet101

S： Spatial Module

C： CRF

D： DR-Net

Bo Dai, Yuqi Zhang, and Dahua Lin. “Detecting visual relationships with deep relational networks”. In:Proceedings of the IEEE conference on computer vision and Pattern recognition.2017, pp. 3076–3086.

Iterative Message Passing

Task

It takes an image as input, and generates a visually-grounded scene graph.

Motivation

1. tackled detecting and recognizing individual objects in isolation → would struggle to perceive the subtle difference between a man feeding a horse and a man standing by a horse

2. These models that use scene graphs either rely on ground-truth annotations, synthetic images, or extract a scene graph from text domain.

Contribution

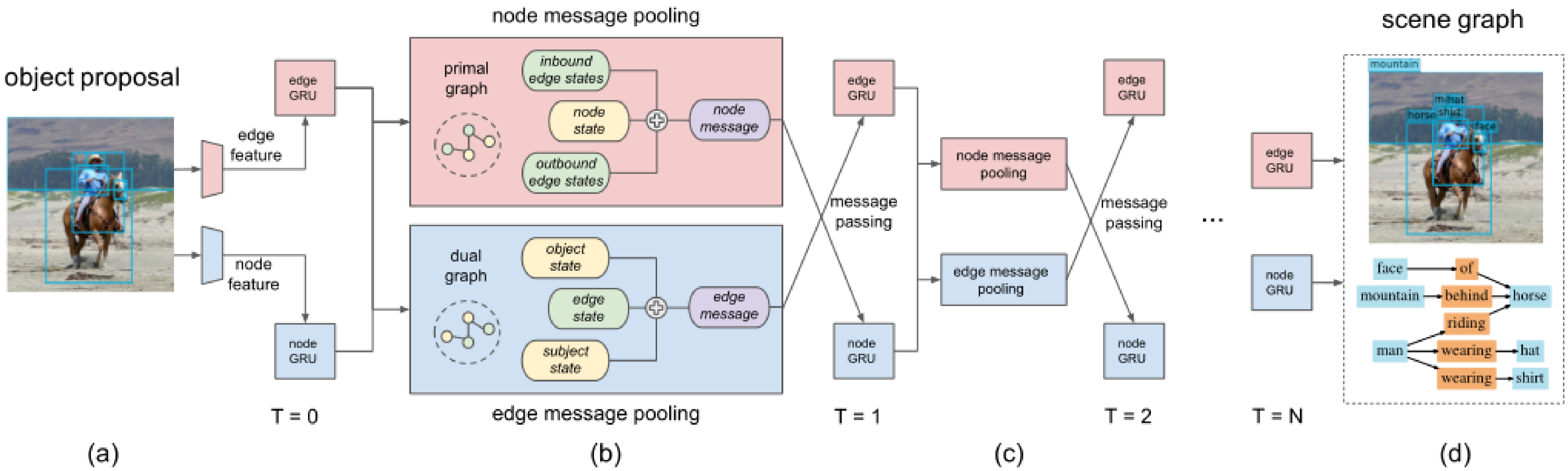
(1) an end-to-end model that generates visually-grounded scene graphs from images.

(2) a novel inference formulation that iteratively refines its prediction by passing contextual messages along the topological structure of a scene graph.

Danfei Xu et al. “Scene graph generation by iterative message passing”. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017, pp. 5410–5419.

Iterative Message Passing

- model architecture



Danfei Xu et al. “Scene graph generation by iterative message passing”. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017, pp. 5410–5419.

Iterative Message Passing

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

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

m_i , $m_{i\to j}$: node and edge message that are to be passed for optimization

h_i , h_j : hidden state of subject and object respectively

$h_{i\to j}$, $h_{j\to i}$: hidden states of outbound edge GRUs and inbound edge GRUs respectively for the i-th object

Danfei Xu et al. “Scene graph generation by iterative message passing”. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017, pp. 5410–5419.

Iterative Message Passing

Result

Dataset: Visual Genome

[26]C. Lu, R. Krishna, M. Bernstein, and L. Fei-Fei. Visual relationship detection with language priors. In European Conference on Computer Vision, 2016.

		[26]	avg. pool	max pool	final
PREDCLS	R@50	27.88	32.39	34.33	**44.75**
	R@100	35.04	39.63	41.99	**53.08**
SGCLS	R@50	11.79	15.65	16.31	**21.72**
	R@100	14.11	18.27	18.70	**24.38**
SGGEN	R@50	0.32	2.70	3.03	**3.44**
	R@100	0.47	3.42	3.71	**4.24**

Top 20 most frequent types (sorted by frequency) are shown. The evaluation metric is recall@5.

predicate	[26]	ours	predicate	[26]	ours
on	**99.71**	99.25	under	28.64	**52.73**
has	**98.03**	97.25	sitting on	31.74	**50.17**
in	80.38	**88.30**	standing on	44.44	**61.90**
of	82.47	**96.75**	in front of	26.09	**59.63**
wearing	**98.47**	98.23	attached to	8.45	**29.58**
near	85.16	**96.81**	at	54.08	**70.41**
with	31.85	**88.10**	hanging from	0.00	0.00
above	49.19	**79.73**	over	**9.26**	0.00
holding	61.50	**80.67**	for	12.20	**31.71**
behind	79.35	**92.32**	riding	72.43	**89.72**

Danfei Xu et al. “Scene graph generation by iterative message passing”. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017, pp. 5410–5419.

MSDN

Task

It simultaneously detect objects, recognize their relationships and predict captions at salient image regions.

Motivation

three vision tasks: object detection, scene graph generation, and image/region captioning.

Though there are connections among the three tasks, the weak alignment across different tasks makes it difficult to learn a model jointly.

Contribution

(1) a novel model to learn features of different semantic levels

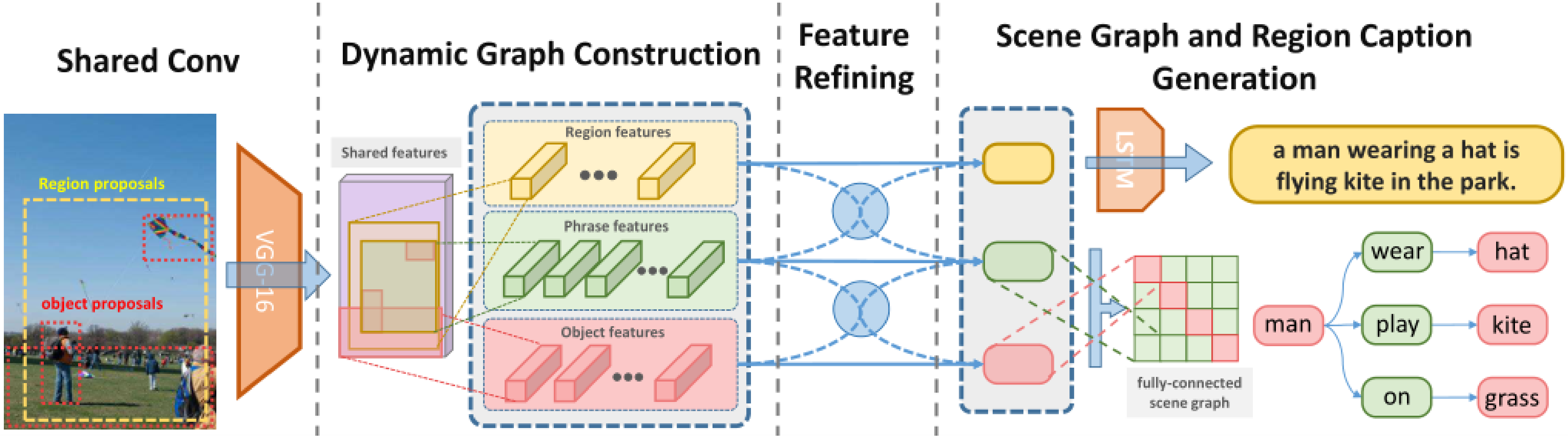
(2) a dynamic graph construction layer in the CNN to construct such a graph.

(3) a feature refining structure to pass message from different semantic levels through the graph

Yikang Li et al. “Scene graph generation from objects, phrases and region captions”. In:Proceedings of the IEEE International Conference on Computer Vision. 2017, pp. 1261–1270.

MSDN

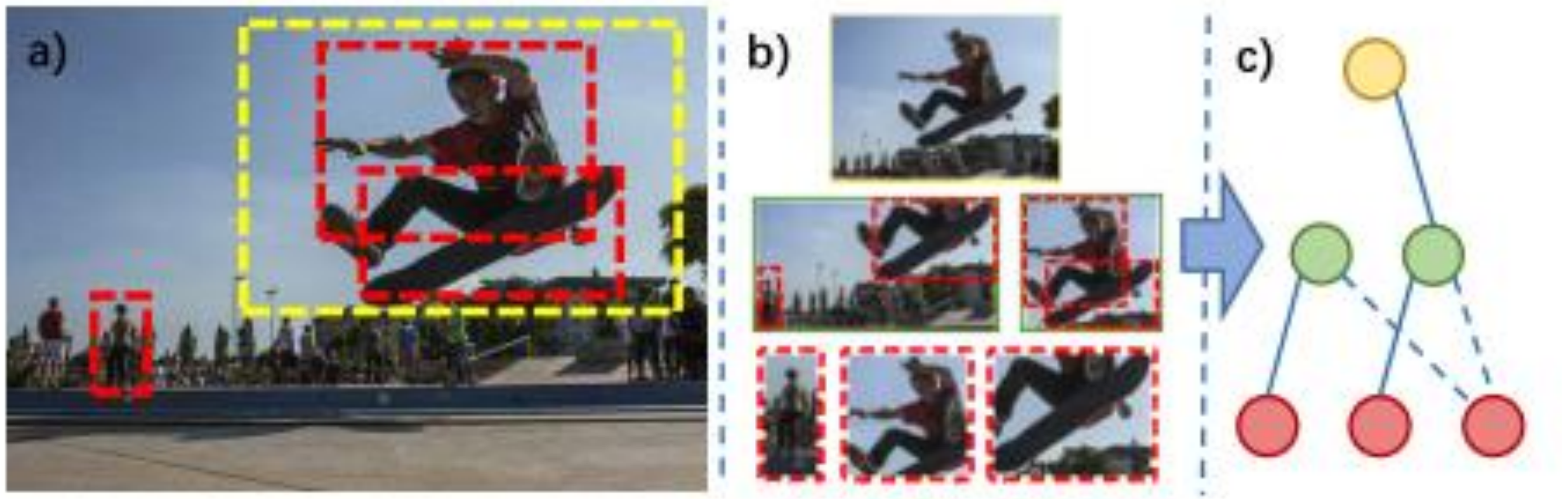
- Overview of MSDN



Yikang Li et al. “Scene graph generation from objects, phrases and region captions”. In:Proceedings of the IEEE International Conference on Computer Vision. 2017, pp. 1261–1270.

MSDN

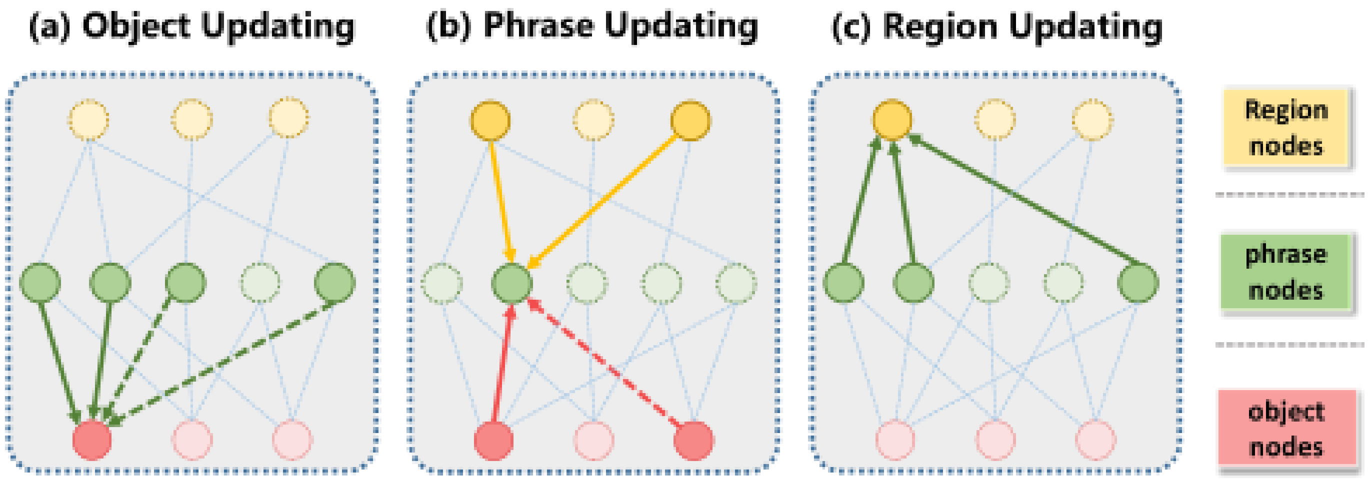
- Dynamical graph construction



Yikang Li et al. “Scene graph generation from objects, phrases and region captions”. In:Proceedings of the IEEE International Conference on Computer Vision. 2017, pp. 1261–1270.

MSDN

- Feature Refining



$$\tilde{\boldsymbol{x}}_i^{(p\to s)} = \frac{1}{\|\boldsymbol{E}_{i,p}\|} \sum_{(i,j)\in \boldsymbol{E}_{s,p}} \sigma_{\langle o,p\rangle}\left(\boldsymbol{x}_i^{(o)}, \boldsymbol{x}_j^{(p)}\right) \boldsymbol{x}_j^{(p)}$$

$$\boldsymbol{x}_{i,t+1}^{(o)} = \boldsymbol{x}_{i,t}^{(o)} + \boldsymbol{F}^{(p\to s)}\left(\tilde{\boldsymbol{x}}_i^{(p\to s)}\right) + \boldsymbol{F}^{(p\to o)}\left(\tilde{\boldsymbol{x}}_i^{(p\to o)}\right)$$

$$x_{j,t+1}^{(p)} = x_{j,t}^{(p)} + F^{(s\to p)}\left(\tilde{x}_j^{(s\to p)}\right) + F^{(o\to p)}\left(\tilde{x}_j^{(o\to p)}\right) + F^{(r\to p)}\left(\tilde{x}_j^{(r\to p)}\right)$$

$$\boldsymbol{x}_{k,t+1}^{(r)} = \boldsymbol{x}_{k,t}^{(r)} + \boldsymbol{F}^{(p\to r)}\left(\tilde{\boldsymbol{x}}_k^{(p\to r)}\right)$$

Yikang Li et al. “Scene graph generation from objects, phrases and region captions”. In:Proceedings of the IEEE International Conference on Computer Vision. 2017, pp. 1261–1270.

MSDN

Result

Dataset: Visual Genome

Task		LP [31]	ISGG [40]	Ours
PredCls	R@50	26.67	58.17	**67.03**
	R@100	33.32	62.74	**71.01**
PhrCls	R@50	10.11	18.77	**24.34**
	R@100	12.64	20.23	**26.50**
SGGen	R@50	0.08	7.09	**10.72**
	R@100	0.14	9.91	**14.22**

ID	Message Passing	Cap. branch	Cap. Supervision	FR-iters	PredCls		PhrCls		SGGen	
					Rec@50	Rec@100	Rec@50	Rec@100	Rec@50	Rec@100
1	-	-	-	0	49.28	52.69	7.31	10.48	2.39	3.82
2	✓	-	-	1	63.12	66.41	19.30	21.82	7.73	10.51
3	✓	✓	-	1	63.82	67.23	20.91	23.09	8.20	11.35
4	✓	✓	✓	1	66.70	**71.02**	23.42	25.68	10.23	13.89
5	✓	✓	✓	2	**67.03**	71.01	**24.22**	**26.50**	**10.72**	**14.22**
6	✓	✓	✓	3	66.23	70.43	23.16	25.28	10.01	13.62

Yikang Li et al. “Scene graph generation from objects, phrases and region captions”. In:Proceedings of the IEEE International Conference on Computer Vision. 2017, pp. 1261–1270.

Graph R-CNN

Task

The task is to detect objects and their relations in images.

Motivation

1. Representing scenes as collections of objects fails to capture relationships which may be essential for scene understanding.
2. Extracting scene graphs from images – efficiently and accurately – is challenging.

Contribution

(1) a novel scene graph generation model called Graph R-CNN

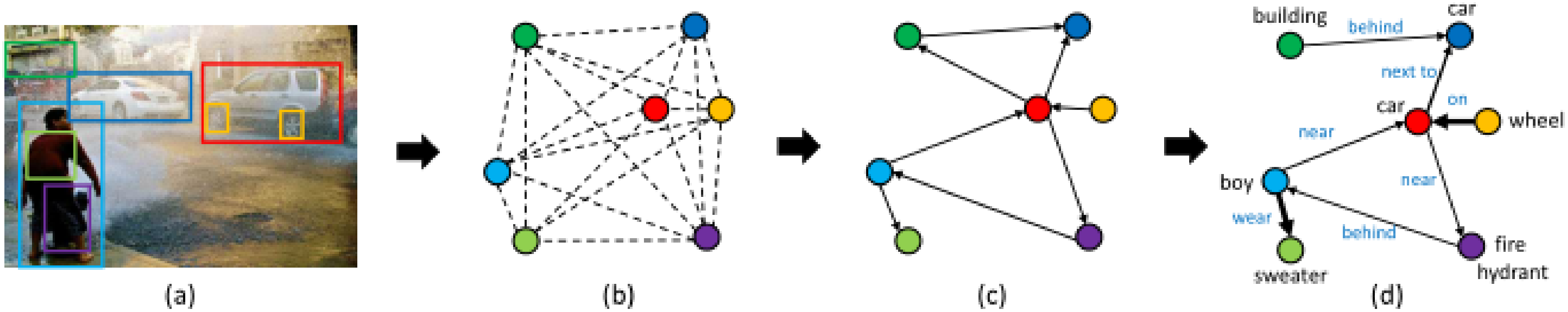
(2) a Relation Proposal Network (RePN)

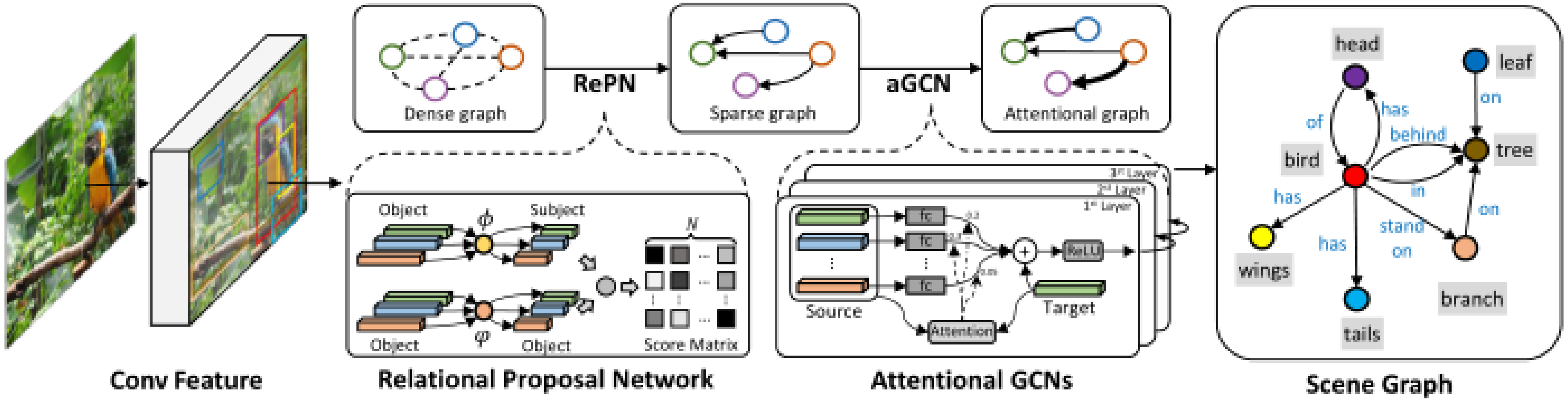
(3) an attentional Graph Convolutional Network (aGCN)

(4) a new evaluation metric

Jianwei Yang et al. “Graph r-cnn for scene graph generation”. In:Proceedings of the Europeanconference on computer vision (ECCV). 2018, pp. 670–685.

Graph R-CNN





Jianwei Yang et al. "Graph r-cnn for scene graph generation". In:Proceedings of the Europeanconference on computer vision (ECCV). 2018, pp. 670–685.

Graph R-CNN

$$z_i^o = \sigma(\overbrace{W^{\texttt{skip}} Z^o \alpha^{\texttt{skip}}}^{\substack{\text{Message from}\\ \text{Other Objects}}} + \overbrace{W^{sr} Z^r \alpha^{sr} + W^{or} Z^r \alpha^{or}}^{\substack{\text{Messages from}\\ \text{Neighboring Relationships}}})$$

$$z_i^r = \sigma(z_i^r + \underbrace{W^{rs} Z^o \alpha^{rs} + W^{ro} Z^o \alpha^{ro}}_{\text{Messages from Neighboring Objects}}).$$

Jianwei Yang et al. “Graph r-cnn for scene graph generation”. In:Proceedings of the Europeanconference on computer vision (ECCV). 2018, pp. 670–685.

Graph R-CNN

dataset： Visual Genome

	SGGen+		SGGen		PhrCls		PredCls	
Method	R@50	R@100	R@50	R@100	R@50	R@100	R@50	R@100
IMP [40]	-	-	3.4	4.2	21.7	24.4	44.8	53.0
MSDN [18]	-	-	7.7	10.5	19.3	21.8	63.1	66.4
Pixel2Graph [26]	-	-	9.7	11.3	26.5	30.0	68.0	75.2
IMP† [40]	25.6	27.7	6.4	8.0	20.6	22.4	40.8	45.2
MSDN† [18]	25.8	28.2	7.0	9.1	27.6	29.9	53.2	57.9
NM-Freq† [42]	26.4	27.8	6.9	9.1	23.8	27.2	41.8	48.8
Graph R-CNN (Us)	**28.5**	**35.9**	**11.4**	**13.7**	**29.6**	**31.6**	**54.2**	**59.1**

Jianwei Yang et al. “Graph r-cnn for scene graph generation”. In:Proceedings of the Europeanconference on computer vision (ECCV). 2018, pp. 670–685.

Neural Motifs

Task

The task is to give object detections, predict the most frequent relation between object pairs with the given labels.

Motivation

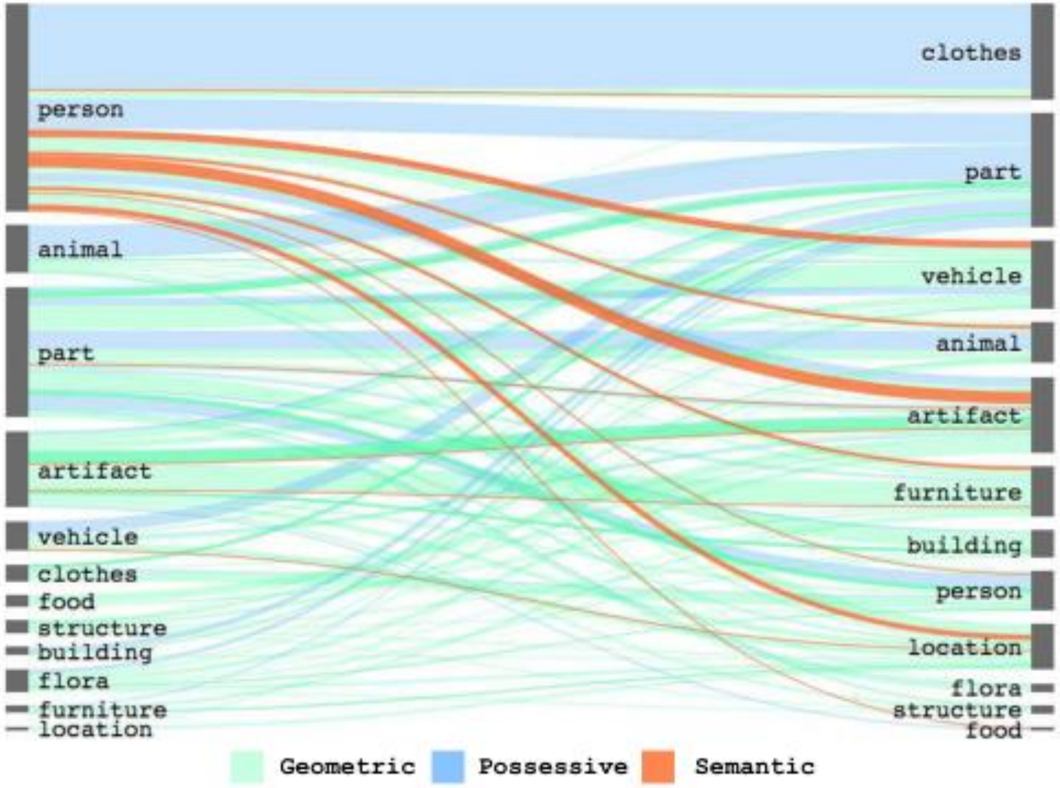
1. There are strong regularities in the local graph structure such that the distribution of the relations is highly skewed once the corresponding object categories are given, but not vice versa.
2. Structural patterns exist even in larger subgraphs; we find that over half of images contain previously occurring graph motifs.

Contribution

a new neural network architecture, the Stacked Motif Network (MOTIFNET), that complements existing approaches to scene graph parsing.

Rowan Zellers et al. “Neural motifs: Scene graph parsing with global context”. In:Proceedingsof the IEEE Conference on Computer Vision and Pattern Recognition. 2018, pp. 5831–5840.

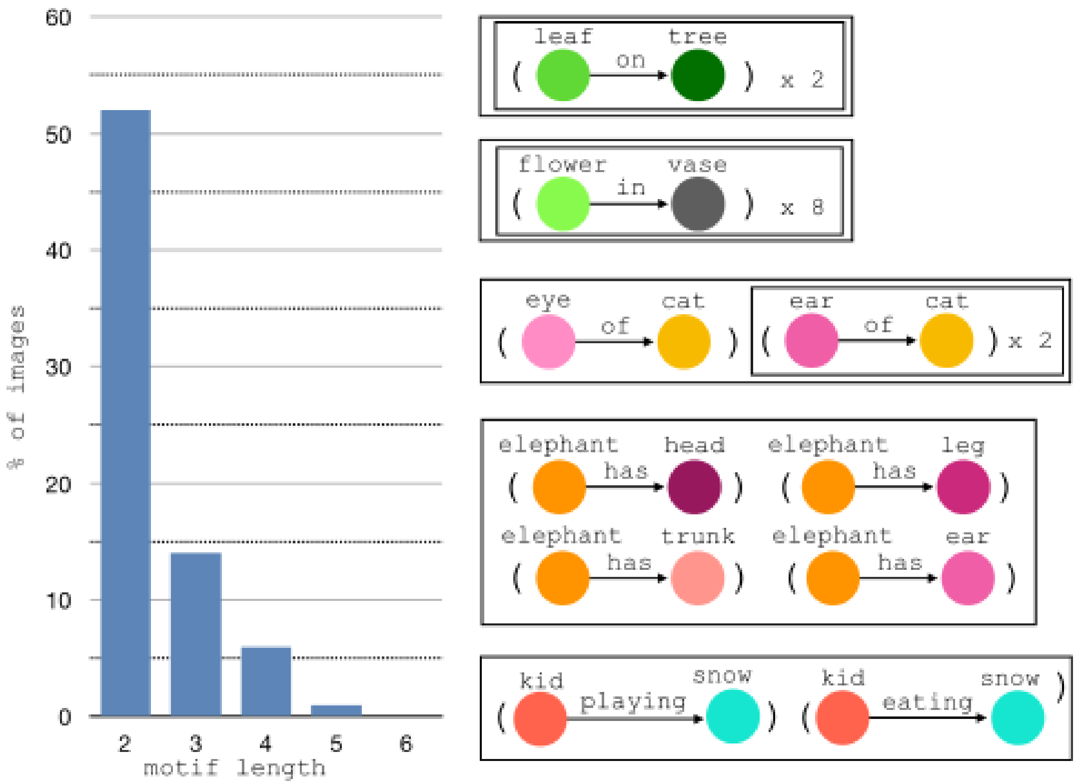
Neural Motifs



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%)

Rowan Zellers et al. “Neural motifs: Scene graph parsing with global context”. In:Proceedingsof the IEEE Conference on Computer Vision and Pattern Recognition. 2018, pp. 5831–5840.

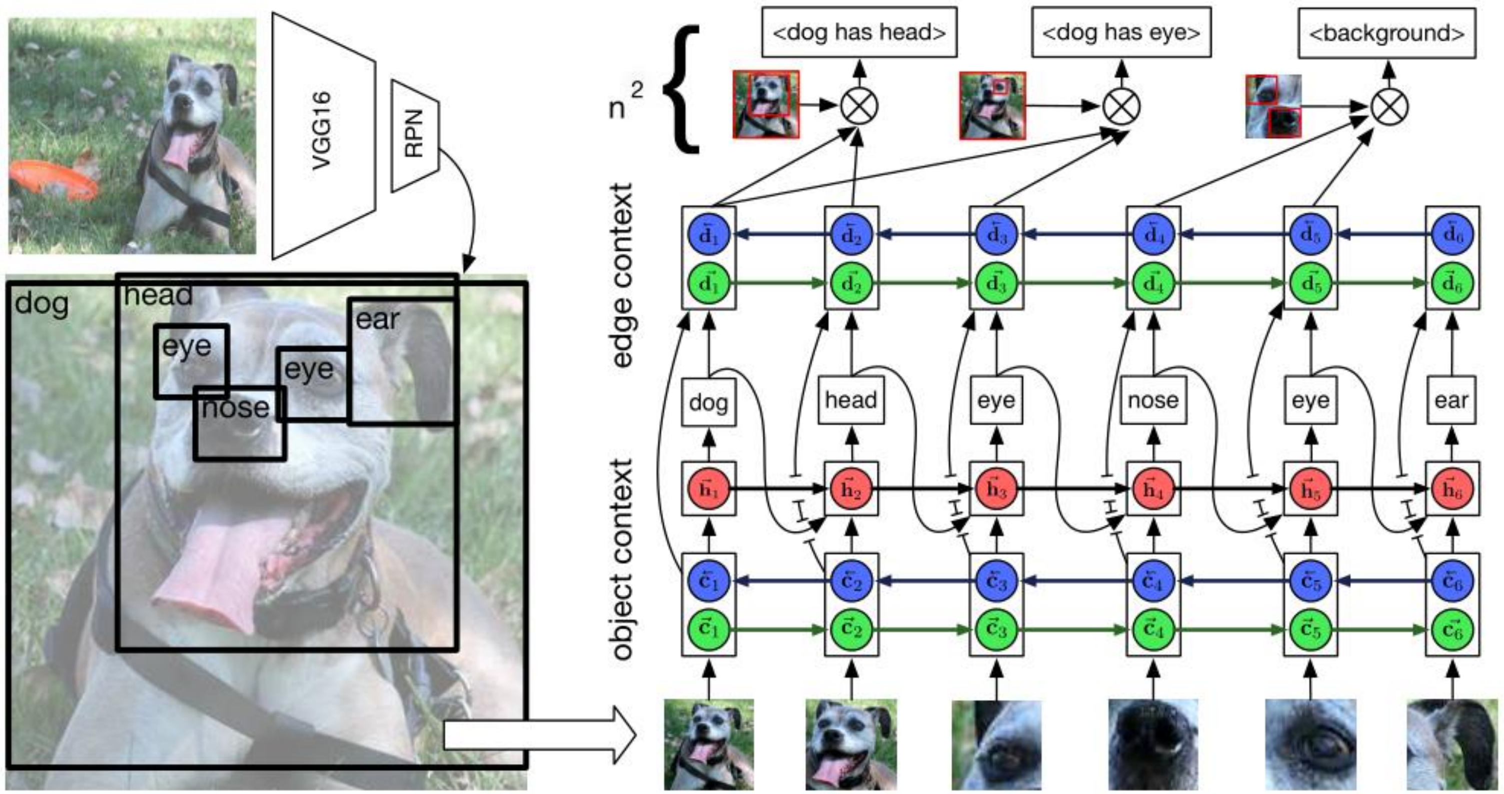
Neural Motifs



Rowan Zellers et al. “Neural motifs: Scene graph parsing with global context”. In:Proceedingsof the IEEE Conference on Computer Vision and Pattern Recognition. 2018, pp. 5831–5840.

Neural Motifs

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



BiLSTM

LSTMs

BiLSTM

graph G

image I

bounding regions B

object labels O

labeled relations R

Rowan Zellers et al. "Neural motifs: Scene graph parsing with global context". In:Proceedingsof the IEEE Conference on Computer Vision and Pattern Recognition. 2018, pp. 5831–5840.

Neural Motifs

	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

Rowan Zellers et al. "Neural motifs: Scene graph parsing with global context". In:Proceedingsof the IEEE Conference on Computer Vision and Pattern Recognition. 2018, pp. 5831–5840.

VCTree

Task

The task is scene graph generation and visual Q&A.

Motivation

1. Bidirectional LSTM for chains are oversimplified and may only capture simple spatial information or cooccurrence bias.
2. CRF-RNN for graphs lack the discrimination between hierarchical relations.

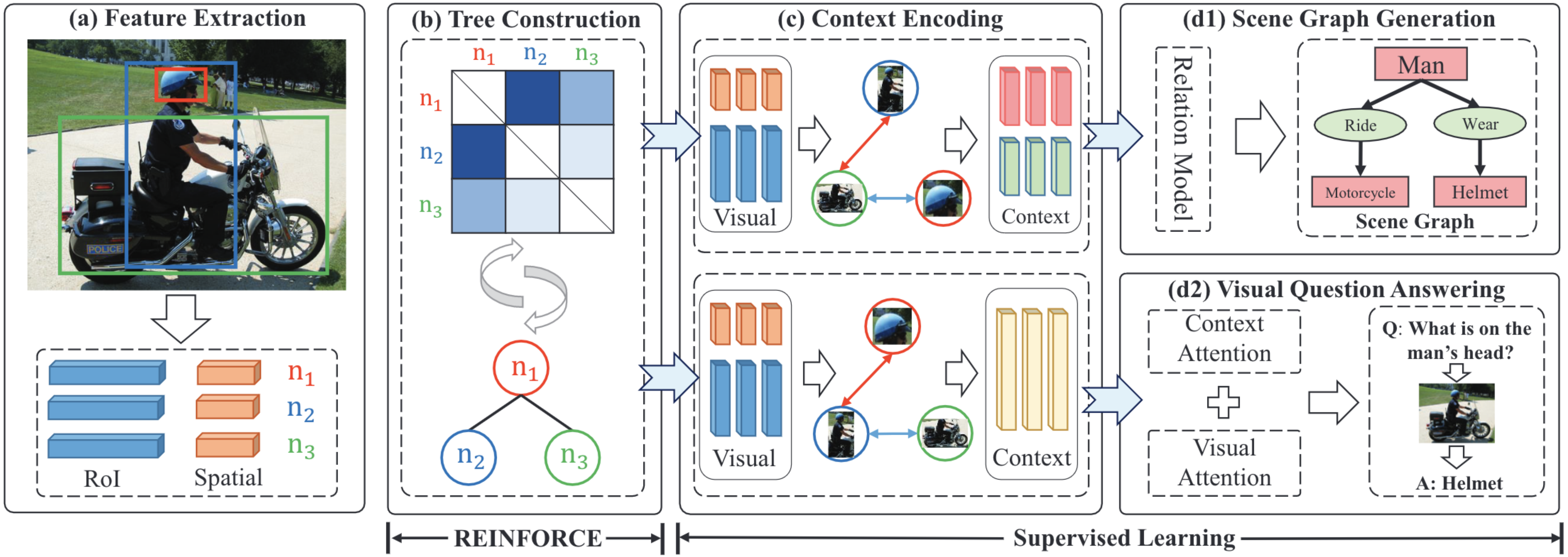
Contribution

(1) a visual context tree model, dubbed VCTREE

(2) a hybrid learning strategy using REINFORCE

Kaihua Tang et al. “Learning to compose dynamic tree structures for visual contexts”. In:Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019,pp. 6619–6628.

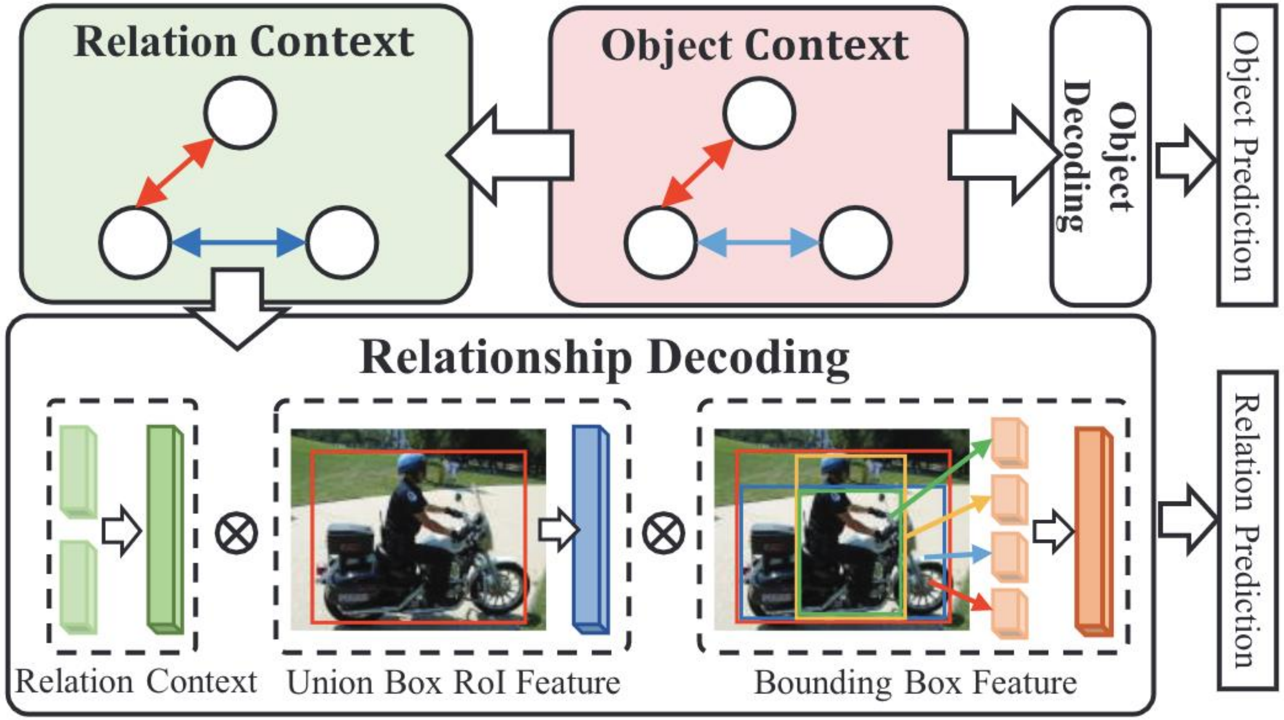
VCTree



$$
\begin{cases}
\boldsymbol{S}_{ij} = f(\boldsymbol{x}_i, \boldsymbol{x}_j) \cdot g(\boldsymbol{x}_i, \boldsymbol{x}_j, \boldsymbol{q}), \\
f(\boldsymbol{x}_i, \boldsymbol{x}_j) = \sigma\left(\mathrm{MLP}(\boldsymbol{x}_i, \boldsymbol{x}_j)\right), \\
g(\boldsymbol{x}_i, \boldsymbol{x}_j, \boldsymbol{q}) = \sigma(h(\boldsymbol{x}_i, \boldsymbol{q})) \cdot \sigma(h(\boldsymbol{x}_j, \boldsymbol{q})),
\end{cases}
$$

Kaihua Tang et al. “Learning to compose dynamic tree structures for visual contexts”. In:Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019,pp. 6619–6628.

VCTree



$$D = \mathrm{BiTreeLSTM}(\{\boldsymbol{z}_i\}_{i=1,2,\ldots,n}),$$

where zi is the input node feature, which will be specified in each task, and D = [d1,d2, …,dn] is the encoded object-level visual context

$$\vec{\boldsymbol{h}}_i = \mathrm{TreeLSTM}(\boldsymbol{z}_i, \vec{\boldsymbol{h}}_p),$$
$$\overleftarrow{\boldsymbol{h}}_i = \mathrm{TreeLSTM}(\boldsymbol{z}_i, [\overleftarrow{\boldsymbol{h}}_l; \overleftarrow{\boldsymbol{h}}_r]),$$

Kaihua Tang et al. “Learning to compose dynamic tree structures for visual contexts”. In:Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019,pp. 6619–6628.

VCTree

	Scene Graph Generation			Scene Graph Classification			Predicate Classification		
Model	R@20	R@50	R@100	R@20	R@50	R@100	R@20	R@50	R@100
VRD [31]	-	0.3	0.5	-	11.8	14.1	-	27.9	35.0
AsscEmbed [34]	6.5	8.1	8.2	18.2	21.8	22.6	47.9	54.1	55.4
IMP$^\diamond$ [50]	14.6	20.7	24.5	31.7	34.6	35.4	52.7	59.3	61.3
TFR [21]	3.4	4.8	6.0	19.6	24.3	26.6	40.1	51.9	58.3
FREQ$^\diamond$ [57]	20.1	26.2	30.1	29.3	32.3	32.9	53.6	60.6	62.2
MOTIFS$^\diamond$ [57]	21.4	27.2	30.3	32.9	35.8	36.5	58.5	65.2	67.1
Graph-RCNN [51]	-	11.4	13.7	-	29.6	31.6	-	54.2	59.1
Chain	21.2	27.1	30.3	33.3	36.1	36.8	59.4	66.0	67.7
Overlap	21.4	27.3	30.4	33.7	36.5	37.1	59.5	66.0	67.8
Multi-Branch	21.5	27.3	30.6	34.3	37.1	37.8	59.5	66.1	67.8
VCTREE-SL	21.7	27.7	31.1	35.0	37.9	38.6	59.8	66.2	67.9
VCTREE-HL	**22.0**	**27.9**	**31.3**	**35.2**	**38.1**	**38.8**	**60.1**	**66.4**	**68.1**

Kaihua Tang et al. “Learning to compose dynamic tree structures for visual contexts”. In:Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019,pp. 6619–6628.

UVTransE

Task

The task is to give an image and get the visual relationship triplets (s, p, o).

Motivation

Consider the Stanford VRD dataset, which has 100 classes of objects, 70 classes of predicates, and a total of 30k training relationship annotations. The number of possible interaction triplets, including unusual cases such as (dog, ride, horse), is 100 ˆ 100 ˆ 70 = 700k, meaning that most relationships do not even have a training example.

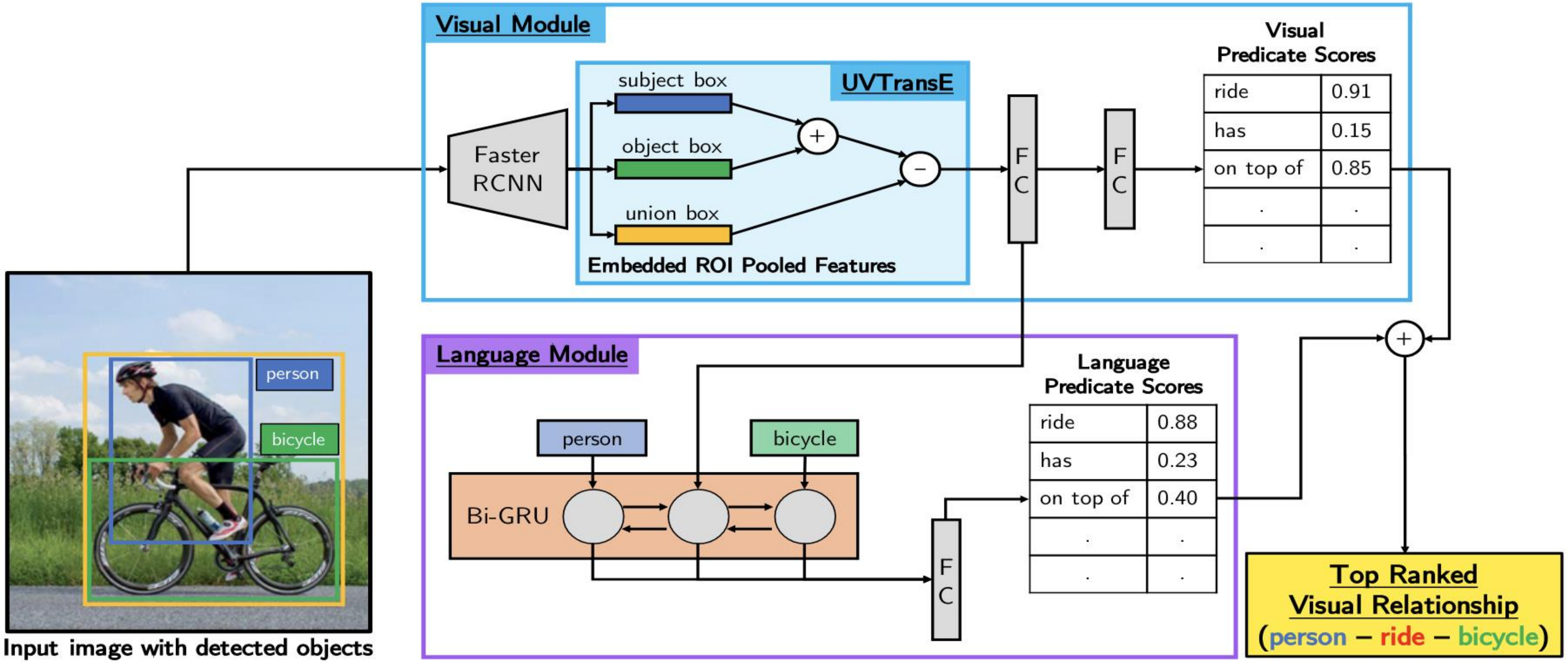
Contribution

（1） a novel framework called Union Visual Translation Embedding, or UVTransE

（2） a language module that benefits the overall detection task

Zih-Siou Hung, Arun Mallya, and Svetlana Lazebnik. Contextual translation embedding for visual relationship detection and scene graph generation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020. 6, 7

UVTransE



Zih-Siou Hung, Arun Mallya, and Svetlana Lazebnik. Contextual translation embedding for visual relationship detection and scene graph generation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020. 6, 7

UVTransE

	Predicate Det.		Phrase Det.				Relationship Det.			
	All	Zero-shot	All		Zero-shot		All		Zero-shot	
	R@50	R@50	R@50	R@100	R@50	R@100	R@50	R@100	R@50	R@100
Appearance	18.17	7.44	8.59	10.68	5.34	10.11	7.52	9.11	4.82	8.97
Appearance + spatial	38.89	14.35	20.06	24.70	7.98	11.84	17.02	20.54	6.90	10.02
Summation	49.01	18.52	21.93	27.80	10.25	14.94	17.78	21.37	9.47	13.33
VTransE [V] (our impl.)	45.12	12.84	19.74	25.62	7.27	10.61	16.21	20.48	6.31	9.55
VTransE [V+L] (our impl.)	50.11	15.31	26.13	31.40	8.73	12.05	22.23	26.14	7.67	10.99
UVTransE [V]	49.98	22.92	23.92	29.57	11.77	17.41	20.22	24.13	10.21	15.92
UVTransE [V+L]	**55.46**	**26.49**	**30.01**	**36.18**	**13.07**	**18.44**	**25.66**	**29.71**	**11.00**	**16.78**

Zih-Siou Hung, Arun Mallya, and Svetlana Lazebnik. Contextual translation embedding for visual relationship detection and scene graph generation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020. 6, 7

External Knowledge and Image Reconstruction

Task

The task is scene graph generation and image reconstruction.

Motivation

1. Training on such a dataset with long-tail distributions will cause the prediction model bias towards those most-frequent relationships.
2. Predicate labels are highly determined by the identification of object pairs.
3. Due to the difficulty of exhaustively labeling bounding boxes of all instances of each object.

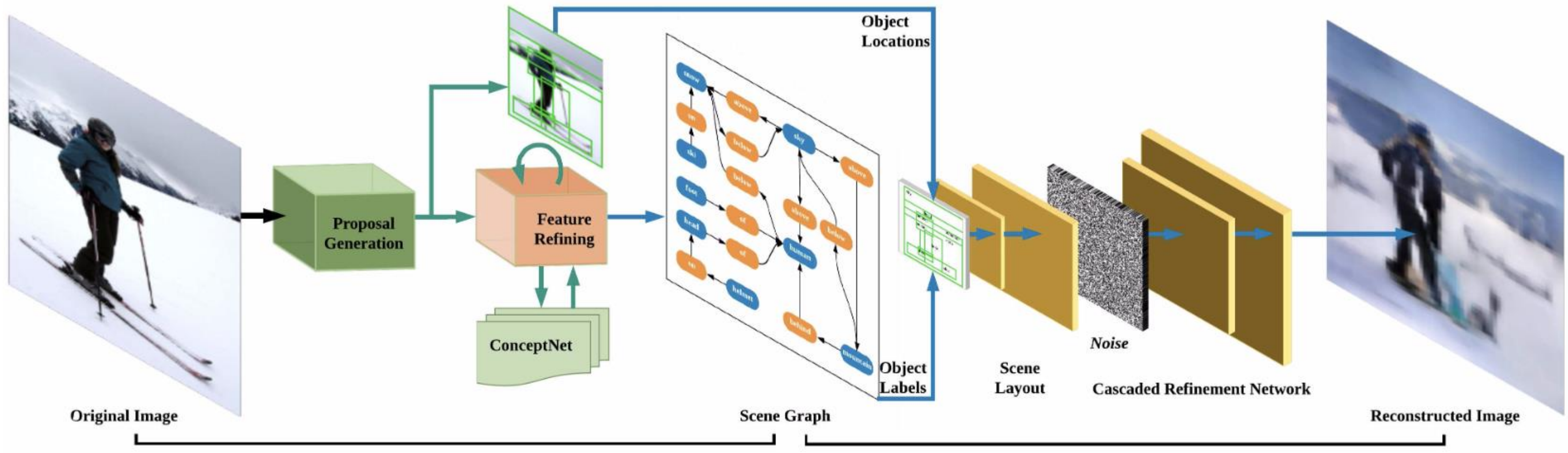
Contribution

(1) a knowledge-based feature refinement module to incorporate commonsense knowledge from an external knowledge base

(2) image-level supervision module by reconstructing the image

Jiuxiang Gu et al. “Scene graph generation with external knowledge and image reconstruction”.In:Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019,pp. 1969–1978.

External Knowledge and Image Reconstruction



Jiuxiang Gu et al. “Scene graph generation with external knowledge and image reconstruction”.In:Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019,pp. 1969–1978.

External Knowledge and Image Reconstruction

- Feature Refinement with External Knowledge
  - Object and Subgraph Inter-refinement

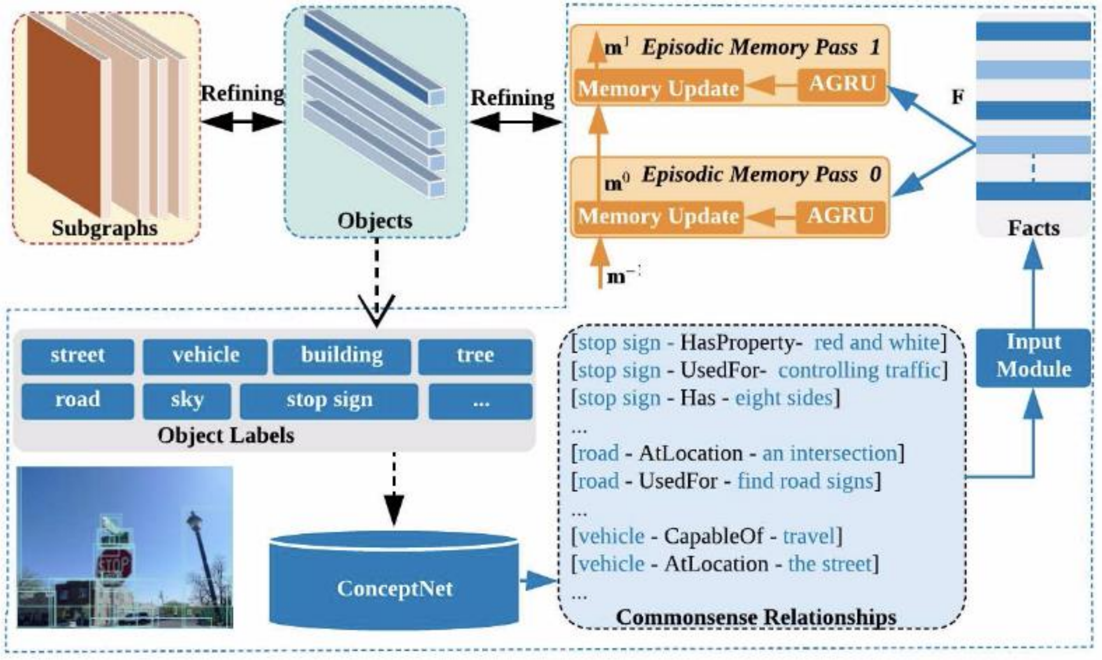
$$\bar{\mathbf{o}}_i = \mathbf{o}_i + f_{s\to o}\left(\sum_{\mathbf{s}_k^i \in \mathbf{S}^i} \alpha_k^{s\to o} \cdot \mathbf{s}_k^i\right)$$

$$\bar{\mathbf{s}}_k = \mathbf{s}_k + f_{o\to s}\left(\sum_{\mathbf{o}_i^k \in \mathbf{O}^k} \alpha_i^{o\to s} \cdot \mathbf{o}_i^k\right)$$

  - Knowledge Retrieval and Embedding

$$a_i \xrightarrow{\text{retrieve}} \langle a_i, a_{i,j}^r, a_j^o, w_{i,j}\rangle, j \in [0, K-1]$$

dynamic memory network



Jiuxiang Gu et al. “Scene graph generation with external knowledge and image reconstruction”.In:Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019,pp. 1969–1978.

External Knowledge and Image Reconstruction

Dataset	Training Set		Testing Set		#Obj	#Pred
	#Img	#Rel	#Img	#Rel		
VRD [29]	4,000	30,355	1,000	7,638	100	70
VG-MSDN [26]	46,164	507,296	10,000	111,396	150	50

Dataset	**Model**	**PhrDet**		**SGGen**	
		Rec@50	Rec@100	Rec@50	Rec@100
VRD [29]	ViP-CNN [27]	22.78	27.91	17.32	20.01
	DR-Net [5]	19.93	23.45	17.73	20.88
	U+W+SF+LK: T+S [45]	26.32	29.43	19.17	21.34
	Factorizable Net [25]	26.03	30.77	18.32	21.20
	KB-GAN	**27.39**	**34.38**	**20.31**	**25.01**
VG-MSDN [26]	ISGG [41]	15.87	19.45	8.23	10.88
	MSDN [26]	19.95	24.93	10.72	14.22
	Graph R-CNN [42]	–	–	11.40	13.70
	Factorizable Net [25]	22.84	28.57	13.06	16.47
	KB-GAN	**23.51**	**30.04**	**13.65**	**17.57**

Jiuxiang Gu et al. "Scene graph generation with external knowledge and image reconstruction".In:Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019,pp. 1969–1978.

Global-Local Attention Transformer(GLAT)

Task

The task is scene graph generation.

Motivation

1. Existing methods rely on an external source of commonsense or statistics directly gathered from training data.
2. Most existing methods are strongly vulnerable to data bias.

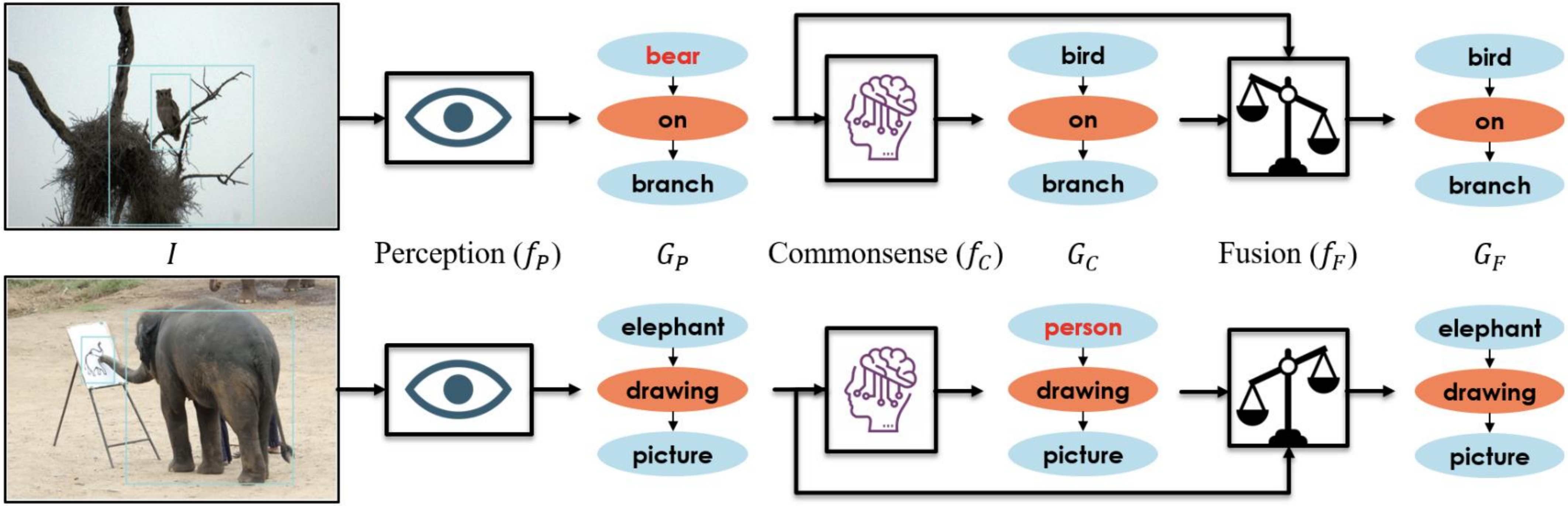
Contribution

(1) the first method for learning structured visual commonsense, Global-Local Attention Transformer (GLAT), which does not require any external knowledge

(2) a cascaded fusion architecture for Scene Graph Generation, which disentangles commonsense reasoning from visual perception

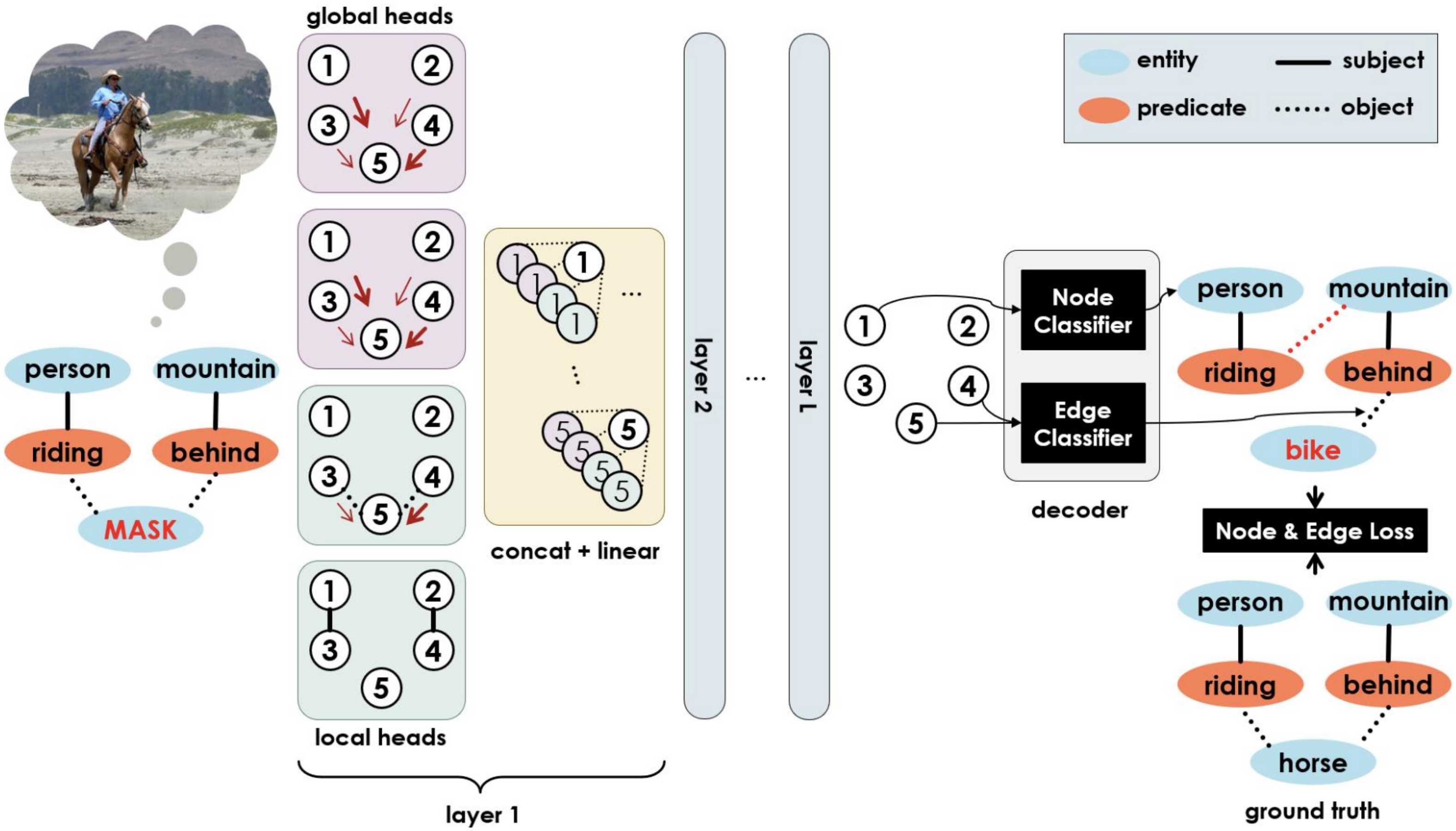
Zareian A., Wang Z., You H., Chang SF. (2020) Learning Visual Commonsense for Robust Scene Graph Generation. In: Vedaldi A., Bischof H., Brox T., Frahm JM. (eds) Computer Vision – ECCV 2020.

Global-Local Attention Transformer(GLAT)



Zareian A., Wang Z., You H., Chang SF. (2020) Learning Visual Commonsense for Robust Scene Graph Generation. In: Vedaldi A., Bischof H., Brox T., Frahm JM. (eds) Computer Vision – ECCV 2020.

Global-Local Attention Transformer(GLAT)



Zareian A., Wang Z., You H., Chang SF. (2020) Learning Visual Commonsense for Robust Scene Graph Generation. In: Vedaldi A., Bischof H., Brox T., Frahm JM. (eds) Computer Vision – ECCV 2020.

Global-Local Attention Transformer(GLAT)

Method	Entity	Predicate	Both
Triplet Frequency [34]	-	44.4	-
Graph Convolutional Nets [11] (local-only, fixed attention)	8.7	43.4	19.7
Graph Attention Nets [24] (local-only)	12.0	45.0	22.3
Transformers [5] (global-only)	14.0	42.3	22.9
Global-Local Attention Transformers (ours)	**22.3**	**60.7**	**34.4**

Method	PredCls		SGCls	
	mR@50	mR@100	mR@50	mR@100
IMP [29]	9.8	10.5	5.8	6.0
IMP + GLAT	11.1	11.9	6.2	6.5
IMP + GLAT + Fusion	**12.1**	**12.9**	**6.6**	**7.0**
SNM [34]	13.3	14.4	7.1	7.5
SNM + GLAT	13.6	14.6	7.3	7.8
SNM + GLAT + Fusion	**14.1**	**15.3**	**7.5**	**7.9**
KERN [2]	17.7	19.2	9.4	10.0
KERN + GLAT	17.6	19.1	9.3	10.0
KERN + GLAT + Fusion	**17.8**	**19.3**	**9.9**	**10.4**

Zareian A., Wang Z., You H., Chang SF. (2020) Learning Visual Commonsense for Robust Scene Graph Generation. In: Vedaldi A., Bischof H., Brox T., Frahm JM. (eds) Computer Vision – ECCV 2020.

GB-Net

Task

The task is scene graph generation.

Motivation

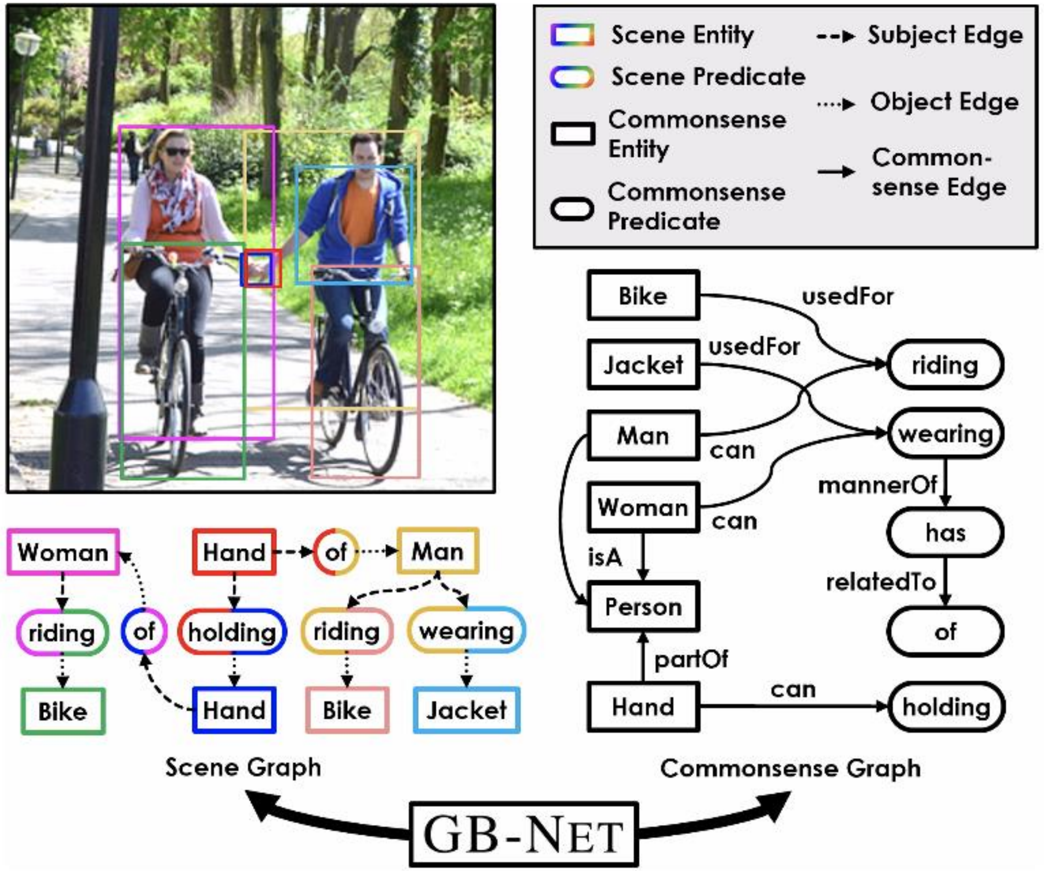
Recent methods either use ad-hoc heuristics to integrate limited types of commonsense into the scene graph generation process, or fail to exploit the rich, graphical structure of commonsense knowledge.

Contribution

(1) Connecting each node to its corresponding class node in the commonsense graph, through an edge we call abridge image-level supervision module by reconstructing the image.
(2) a novel graphical neural network, that iteratively propagates messages between the scene and commonsense graphs, as well as within each of them

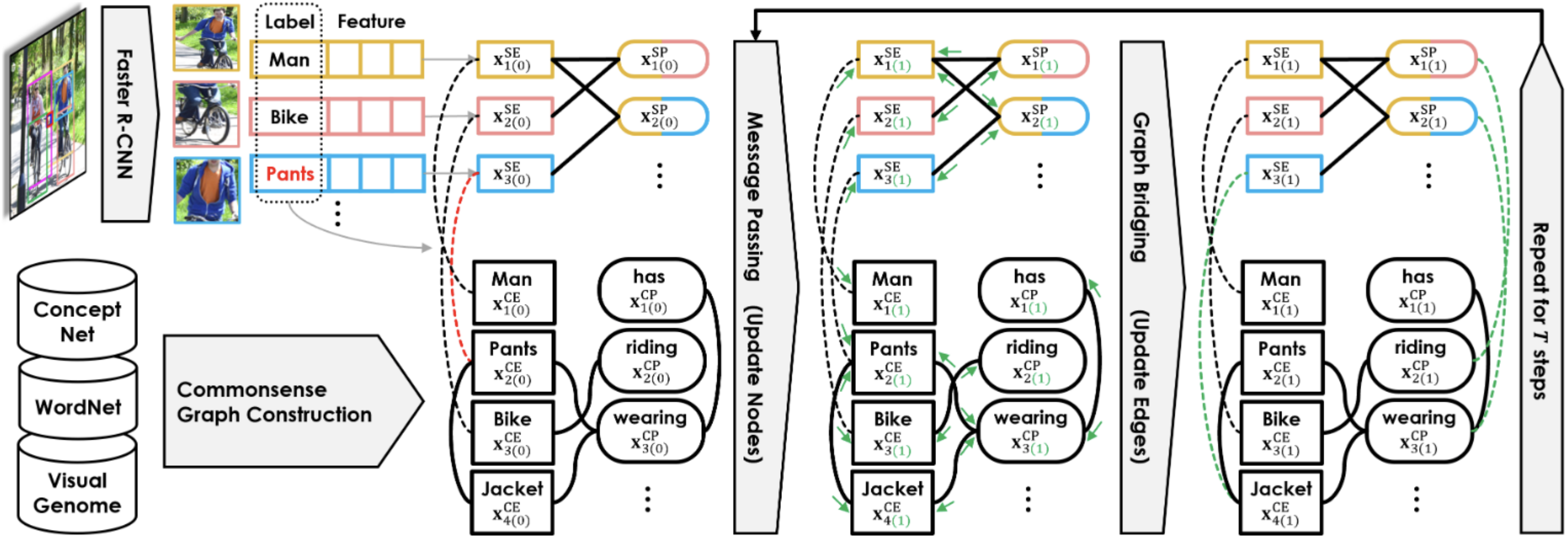
Zareian A., Karaman S., Chang SF. (2020) Bridging Knowledge Graphs to Generate Scene Graphs. In: Vedaldi A., Bischof H., Brox T., Frahm JM. (eds) Computer Vision – ECCV 2020.

GB-Net



Zareian A., Karaman S., Chang SF. (2020) Bridging Knowledge Graphs to Generate Scene Graphs. In: Vedaldi A., Bischof H., Brox T., Frahm JM. (eds) Computer Vision – ECCV 2020.

GB-Net



Zareian A., Karaman S., Chang SF. (2020) Bridging Knowledge Graphs to Generate Scene Graphs. In: Vedaldi A., Bischof H., Brox T., Frahm JM. (eds) Computer Vision – ECCV 2020.

summary

- graph features(local and global context)
- dataset
- structure and hierarchy
- external knowledge and commonsense

Future work

- spatio-temporal scene graph
- solve excessive dependence on background information and external knowledge
- uncommon relationships