Language-Driven Visual Reasoning for Referring Expression Comprehension

李冠彬

中山大学 数据科学与计算机学院

Outline

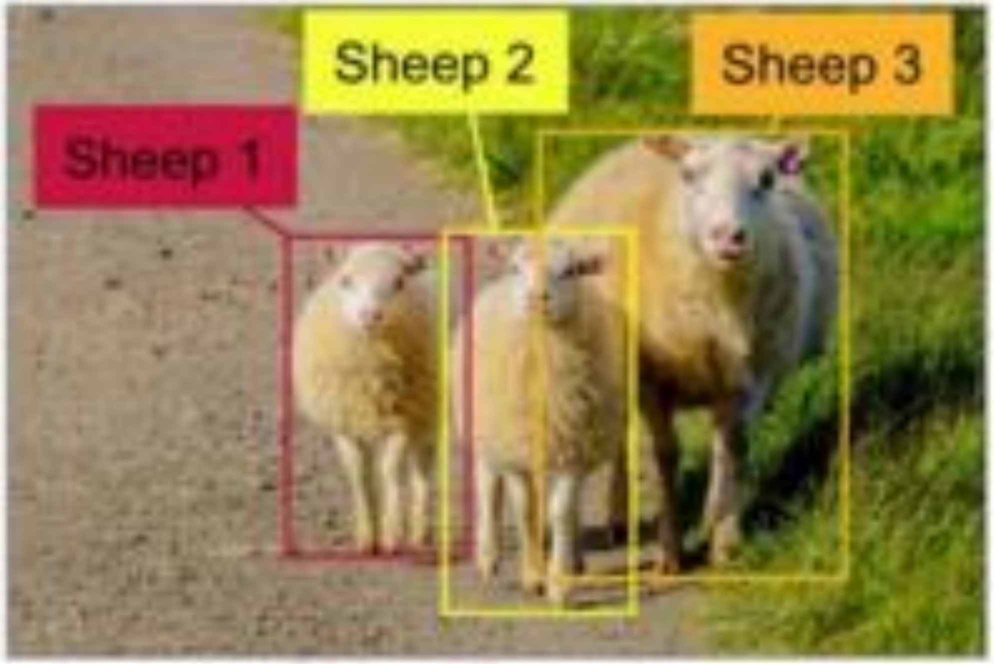
- Introduction and Related Work
- Cross-Modal Relationship Inference Network, CVPR 2019
- Dynamic Graph Attention for Visual Reasoning, ICCV2019
- Scene Graph guided Visual Reasoning, CVPR2020
- Conclusion and Future Work Discussion

Outline

- ☐ Introduction and Related Work
- ☐ Cross-Modal Relationship Inference Network, CVPR 2019
- ☐ Dynamic Graph Attention for Visual Reasoning, ICCV2019
- ☐ Scene Graph guided Visual Reasoning, CVPR2020
- ☐ Conclusion and Future Work Discussion

Introduction

Referring Expression Comprehension

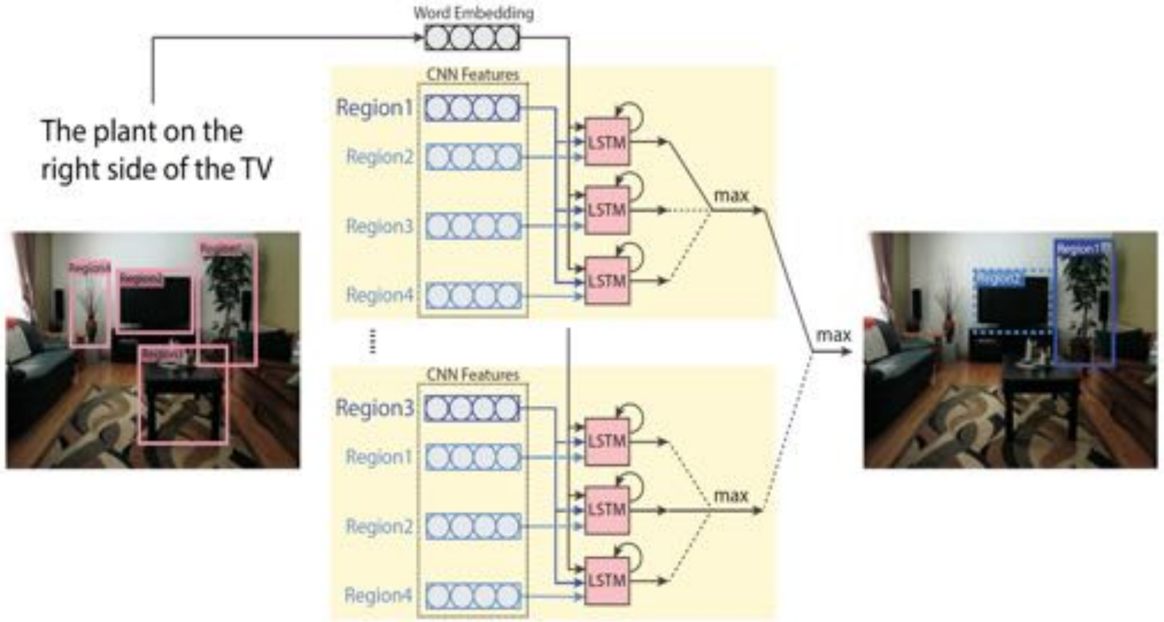


1. The sheep in the middle	Sheep 2
2. The fattest sheep	Sheep 3
3. The sheep farthest from the grass	Sheep 1

Requires Relationship Reasoning

1. The hat worn by the man bending over and stroking the dog
2. The hat on the guy to the left of the man in the yellow shirt

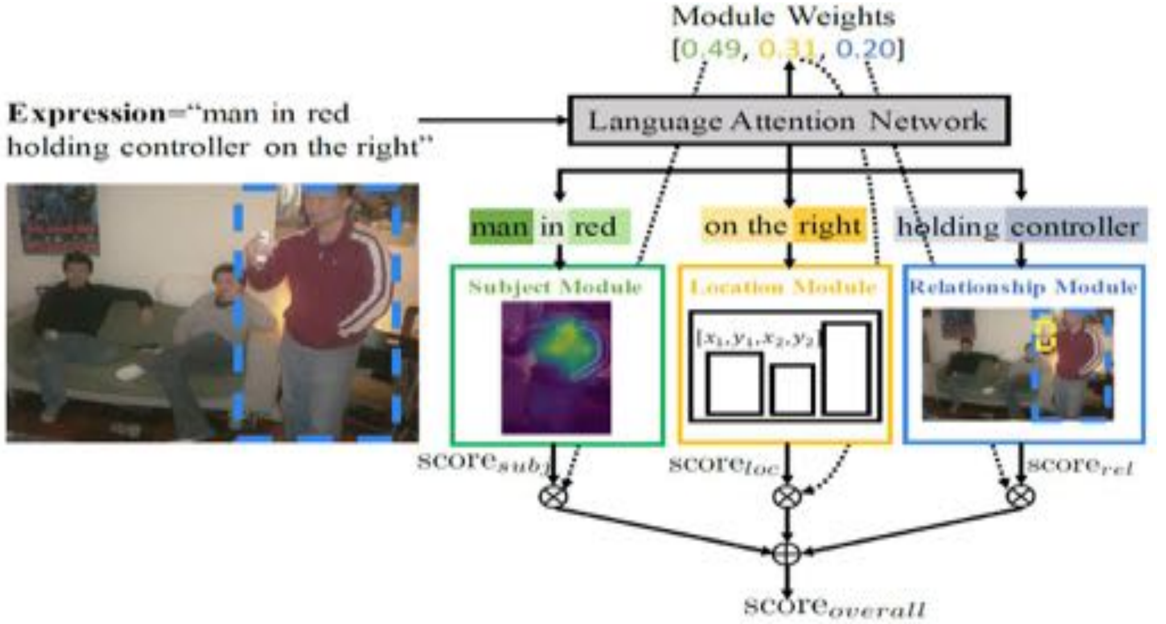
Related Work



$$R^* = \arg\max_{R \in \mathcal{C}} p(R|S, I)$$

$$R^* = \arg\max_{R \in \mathcal{C}} p(S|R, I)$$

(Nagaraja et al. ECCV2016)



$$[w_{subj}, w_{loc}, w_{rel}] = \text{softmax}(W_m^T[h_0, h_T] + b_m)$$

$$S(o_i|r) = w_{subj} S(o_i|q^{subj}) + w_{loc} S(o_i|q^{loc}) + w_{rel} S(o_i|q^{rel})$$

Modular Attention Network (CVPR2018)

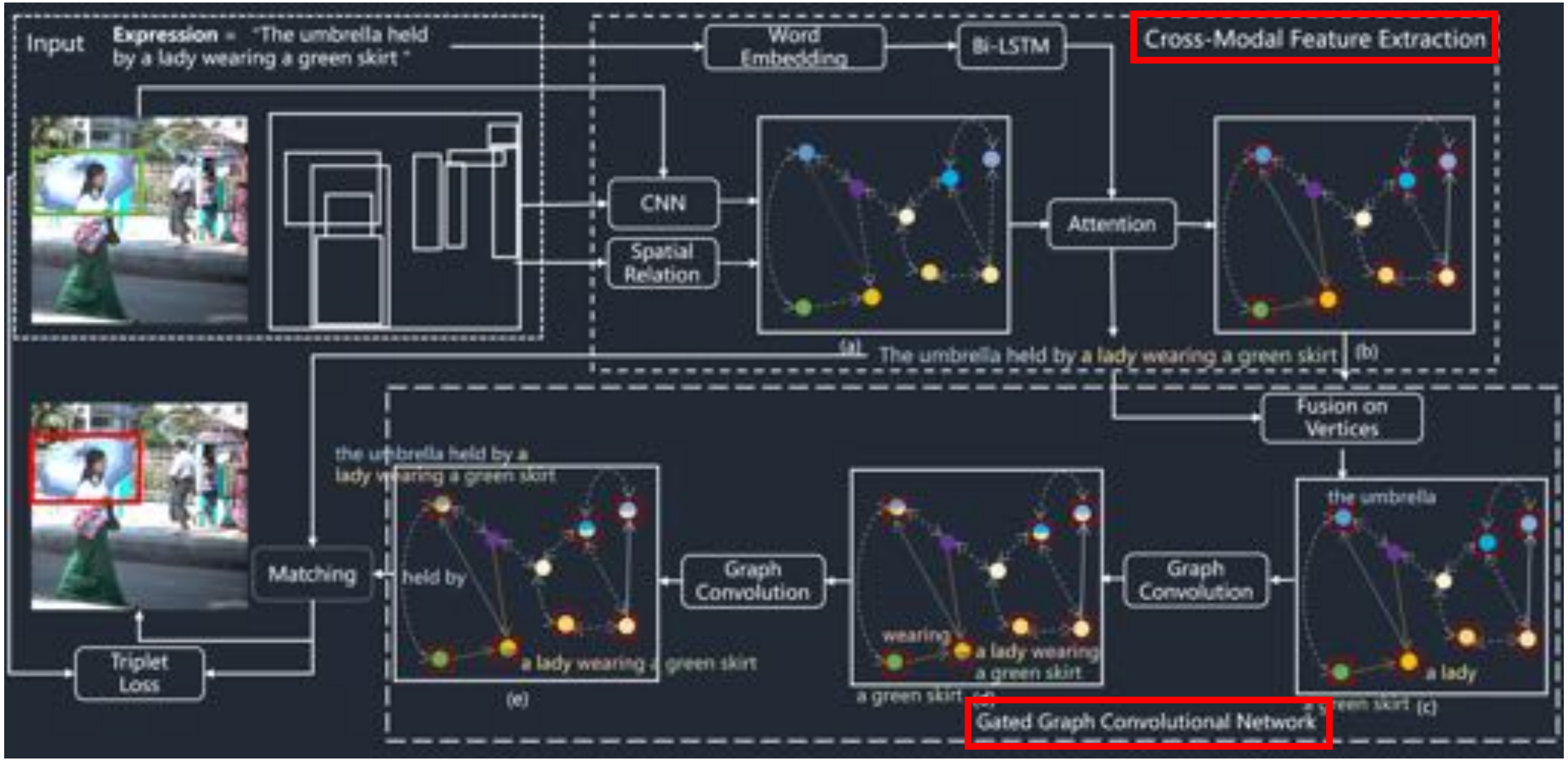
Outline

- ☐Introduction and Related Work
- ☐Cross-Modal Relationship Inference Network, CVPR 2019
- ☐Dynamic Graph Attention for Visual Reasoning, ICCV2019
- ☐Scene Graph guided Visual Reasoning, CVPR2020
- ☐Conclusion and Future Work Discussion

Cross-Modal Relationship Inference (CVPR2019)

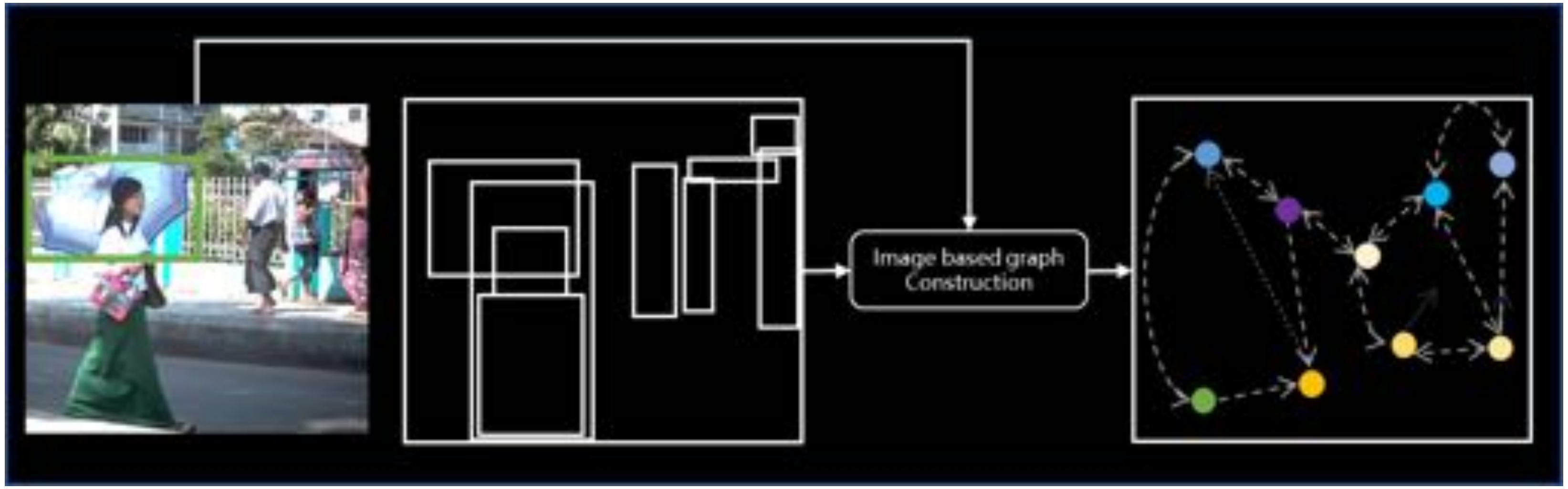
Motivation:

- Relationships (including first-order and multi-order) is essential for visual grounding.
- Graph based information propagation helps to explicitly capture multi-order relationships.

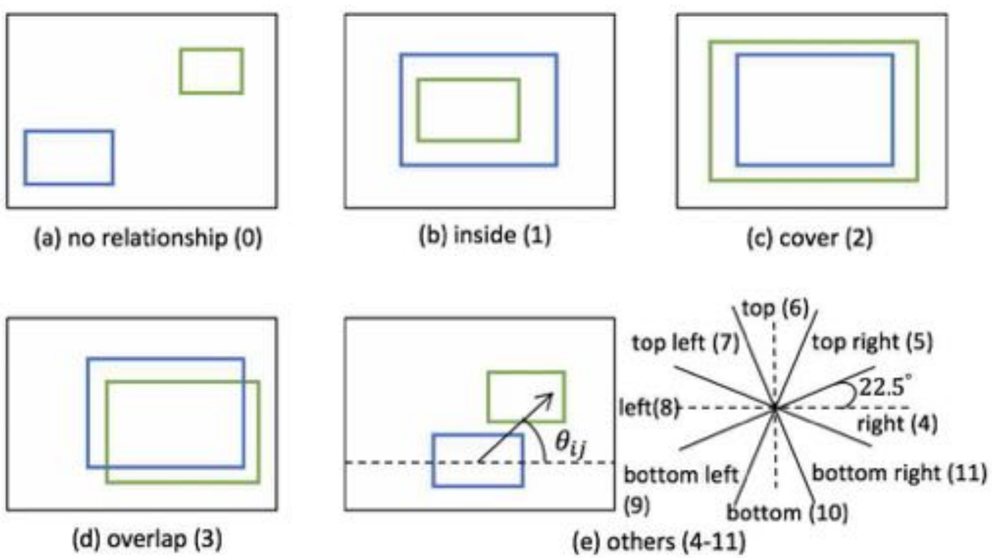


Language-Guided Visual Relation Graph

Spatial Relation Graph Construction



$$G^s = (V, E, \mathbf{X}^s)$$
$$V = \{v_i\}_{i=1}^{K}$$
$$\mathbf{X}^s = \{\mathbf{x}_i^s\}_{i=1}^{K}$$

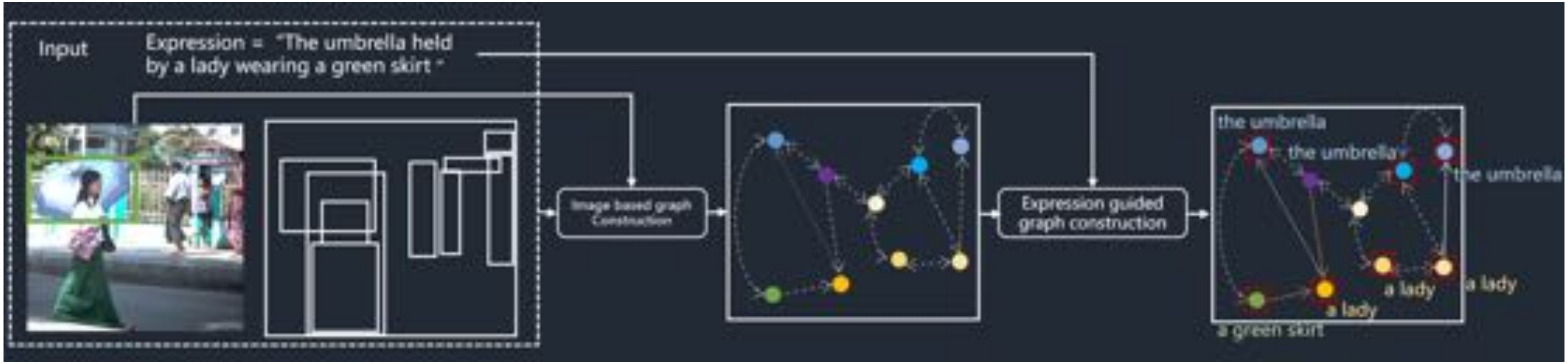


$$E = \{e_{ij}\}_{i,j=1}^{K}$$

e_{ij} is the index label of relationship r_{ij}

Language-Guided Visual Relation Graph

Language-Guided Visual Relation Graph Construction



1. Given expression $L = \{l_t\}_{t=1}^{T}$, Bidirectional LSTM for word feature extraction $\mathbf{h}_t \in \mathbb{R}^{D_h}$

2. The type (i.e. entity, relation, absolute location and unnecessary word) for each word

$$\mathbf{m}_t = \mathrm{softmax}(\mathbf{W}_{l1}\sigma(\mathbf{W}_{l0}\mathbf{h}_t + \mathbf{b}_{l0}) + \mathbf{b}_{l1})$$

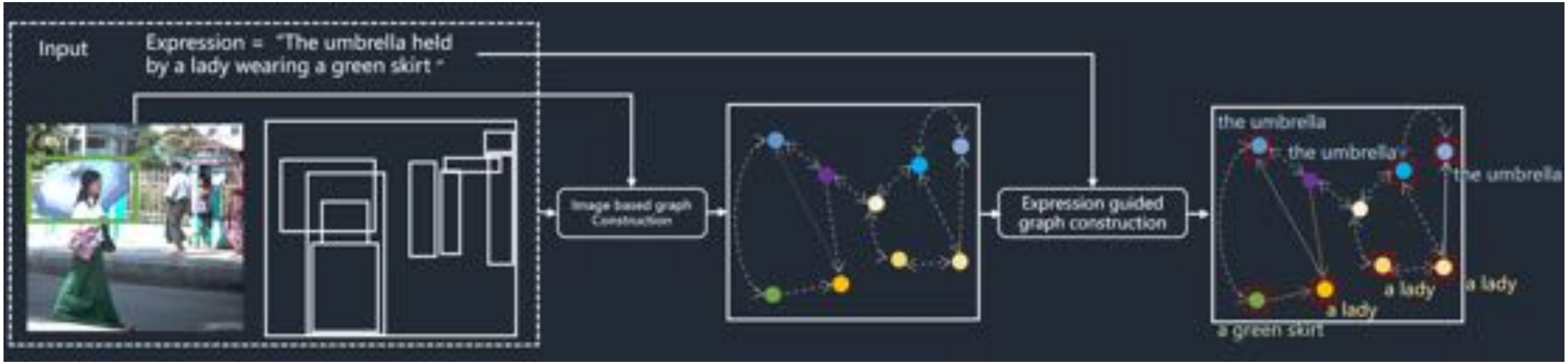
Weighted normalized attention of word l_t refer to vertex v_i ,

$$\begin{cases} \alpha_{t,i} = \mathbf{W}_n[\tanh(\mathbf{W}_v\mathbf{x}_i^s + \mathbf{W}_h\mathbf{h}_t)] \\ \lambda_{t,i} = \mathbf{m}_t^{(0)} \frac{\exp(\alpha_{t,i})}{\sum_i^K \exp(\alpha_{t,i})} \end{cases}$$

The language context $\mathbf{c}_i$ at vertex v_i : $\mathbf{c}_i = \sum_{t=1}^{T} \lambda_{t,i}\mathbf{h}_t$

Language-Guided Visual Relation Graph

Language-Guided Visual Relation Graph Construction



3. The gate p_i^v for vertex v_i is defined as: $p_i^v = \sum_{t=1}^{T} \lambda_{t,i}$

the gate p_j^e for edges with type $j \in \{1, 2, ..N^e\}$ is: $p_j^e = \sum_{t=1}^{T} w_{t,j}^e$

The language-guided multi-modal graph is defined as: $G^m = (V, E, \mathbf{X}^m, P^v, P^e)$

$$\mathbf{X}^m = \{\mathbf{x}_i^m\}_{i=1}^K \qquad \mathbf{x}_i^m = [\mathbf{x}_i^s, \mathbf{c}_i]$$

Language-Guided Visual Relation Graph

Gated graph convolution operation at vertex: v_i

$$\overrightarrow{\mathbf{x}_i}^{(n)} = \sum_{e_{i,j}>0} p^e_{e_{i,j}} (\overrightarrow{\mathbf{W}}^{(n)} \hat{\mathbf{x}}_j^{(n-1)} p^v_j + \mathbf{b}^{(n)}_{e_{i,j}})$$

$$\overleftarrow{\mathbf{x}_i}^{(n)} = \sum_{e_{j,i}>0} p^e_{e_{j,i}} (\overleftarrow{\mathbf{W}}^{(n)} \hat{\mathbf{x}}_j^{(n-1)} p^v_j + \mathbf{b}^{(n)}_{e_{j,i}})$$

$$\tilde{\mathbf{x}}_i^{(n)} = \widetilde{\mathbf{W}}^{(n)} \hat{\mathbf{x}}_i^{(n-1)} + \widetilde{\mathbf{b}}^{(n)}$$

$$\hat{\mathbf{x}}_i^{(n)} = \sigma(\overrightarrow{\mathbf{x}_i}^{(n)} + \overleftarrow{\mathbf{x}_i}^{(n)} + \tilde{\mathbf{x}}_i^{(n)})$$

$$\mathbf{X}^c = \{\mathbf{x}_i^c = \hat{\mathbf{x}}_i^{(N)}\}_{i=1}^K$$

$$\mathbf{x}_i = [\mathbf{W}_p \mathbf{p}_i, \mathbf{x}_i^c]$$

Matching Score and Loss Function:

$$s_i = \mathrm{L2Norm}(\mathbf{W}_{s0}\mathbf{x}_i) \odot \mathrm{L2Norm}(\mathbf{W}_{s1}\mathbf{h}_g)$$

$$loss = \max(s_{neg} + \Delta - s_{gt}, 0)$$

Experiments

Evaluation Datasets: RefCOCO, RefCOCO+ and RefCOCOg
Evaluation Metric: Precision@1 metric (the fraction of correct predictions)

			RefCOCO			RefCOCO+			RefCOCOg	
		feature	val	testA	testB	val	testA	testB	val	test
1	MMI [23]	vgg16	-	63.15	64.21	-	48.73	42.13	-	-
2	Neg Bag [26]	vgg16	76.90	75.60	78.00	-	-	-	-	68.40
3	CG [22]	vgg16	-	74.04	73.43	-	60.26	55.03	-	-
4	Attr [19]	vgg16	-	78.85	78.07	-	61.47	57.22	-	-
5	CMN [7]	vgg16	-	75.94	79.57	-	59.29	59.34	-	-
6	Speaker [36]	vgg16	76.18	74.39	77.30	58.94	61.29	56.24	-	-
7	Listener [37]	vgg16	77.48	76.58	78.94	60.50	61.39	58.11	69.93	69.03
8	Speaker+Listener+Reinforcer [37]	vgg16	79.56	78.95	80.22	62.26	64.60	59.62	71.65	71.92
9	VariContext [41]	vgg16	-	78.98	82.39	-	62.56	62.90	-	-
10	AccumulateAttn [4]	vgg16	81.27	81.17	80.01	65.56	68.76	60.63	-	-
11	ParallelAttn [42]	vgg16	81.67	80.81	81.32	64.18	66.31	61.46	-	-
12	MAttNet [35]	vgg16	80.94	79.99	82.30	63.07	65.04	61.77	73.04	72.79
13	Ours CMRIN	vgg16	84.02	84.51	82.59	71.46	75.38	64.74	76.16	76.25
14	MAttNet [35]	resnet101	85.65	85.26	84.57	71.01	75.13	66.17	78.10	78.12
15	Ours CMRIN	resnet101	**86.99**	**87.63**	**84.73**	**75.52**	**80.93**	**68.99**	**80.45**	**80.66**

Comparison with state-of-the-art approaches on RefCOCO, RefCOCO+ and RefCOCOg

Experiments

global langcxt+vis instance: Visual feature + location feature, last hidden unit of LSTM, matching

global langcxt+global viscxt(2): GCN on the spatial relation graph

weighted langcxt+guided viscxt: Gated GCN on the language-guided visual relation graph

weighted langcxt+guided viscxt+fusion: Gated GCN on cross-modal relation graph

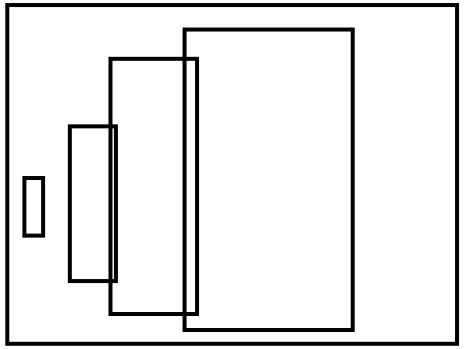
		RefCOCO			RefCOCO+			RefCOCOg	
		val	testA	testB	val	testA	testB	val	test
1	global langcxt+vis instance	79.05	81.47	77.86	63.85	69.82	57.80	70.78	71.26
2	global langcxt+global viscxt(2)	82.61	83.22	82.36	67.75	73.21	63.06	74.29	75.23
3	weighted langcxt+guided viscxt(2)	85.29	86.09	84.12	73.70	79.60	67.52	78.47	79.39
4	weighted langcxt+guided viscxt(1)+fusion	85.80	86.09	83.98	73.95	78.43	67.21	79.37	78.90
5	weighted langcxt+guided viscxt(3)+fusion	86.55	87.50	84.53	75.29	80.46	68.79	80.11	80.45
6	weighted langcxt+guided viscxt(2)+fusion	**86.99**	**87.63**	**84.73**	**75.52**	**80.93**	**68.99**	**80.45**	**80.66**

Ablation study on variances of our proposed CMRIN on RefCOCO, RefCOCO+ and RefCOCOg

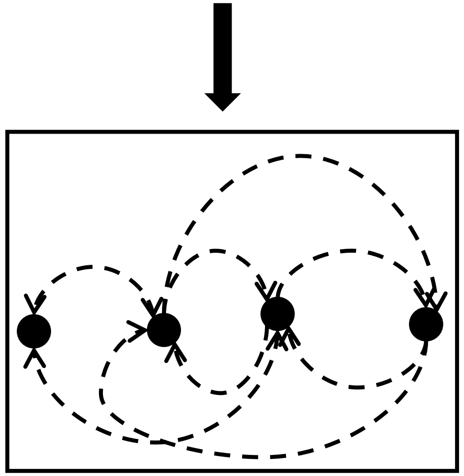
Visualization Results

“an elephant between two other elephants”

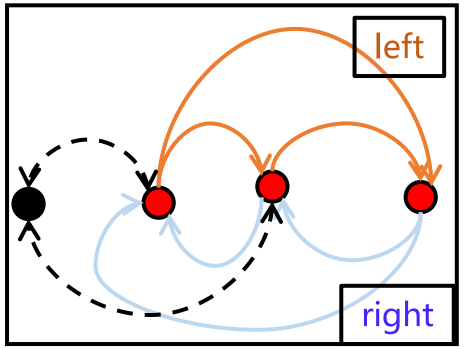
Input Image



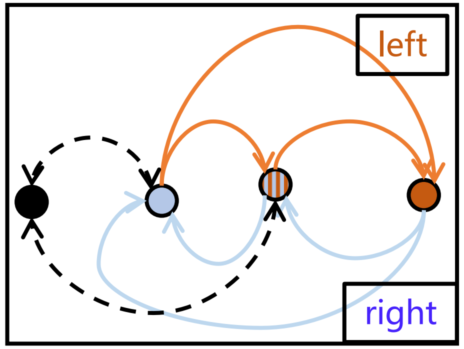
objects



Initial Attention Score



Final matching score

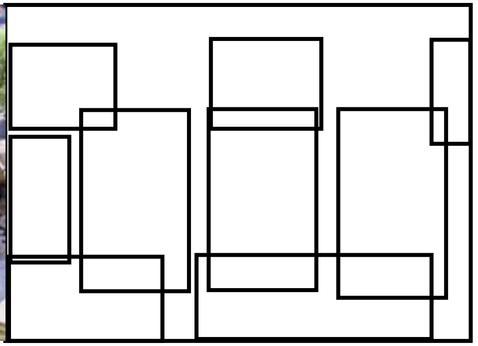


Result

Visualization Results

"green plant behind a table visible behind a lady's head"

Input Image



Objects

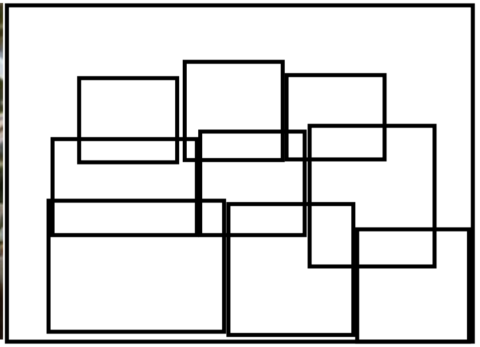
Initial Attention Score

Final matching score

Result

"sandwich in center row all the way on right"

Input Image



Objects

Initial Attention Score

Final matching score

Result

Outline

- Introduction and Related Work
- Cross-Modal Relationship Inference Network, CVPR 2019
- Dynamic Graph Attention for Visual Reasoning, ICCV2019
- Scene Graph guided Visual Reasoning, CVPR2020
- Conclusion and Future Work Discussion

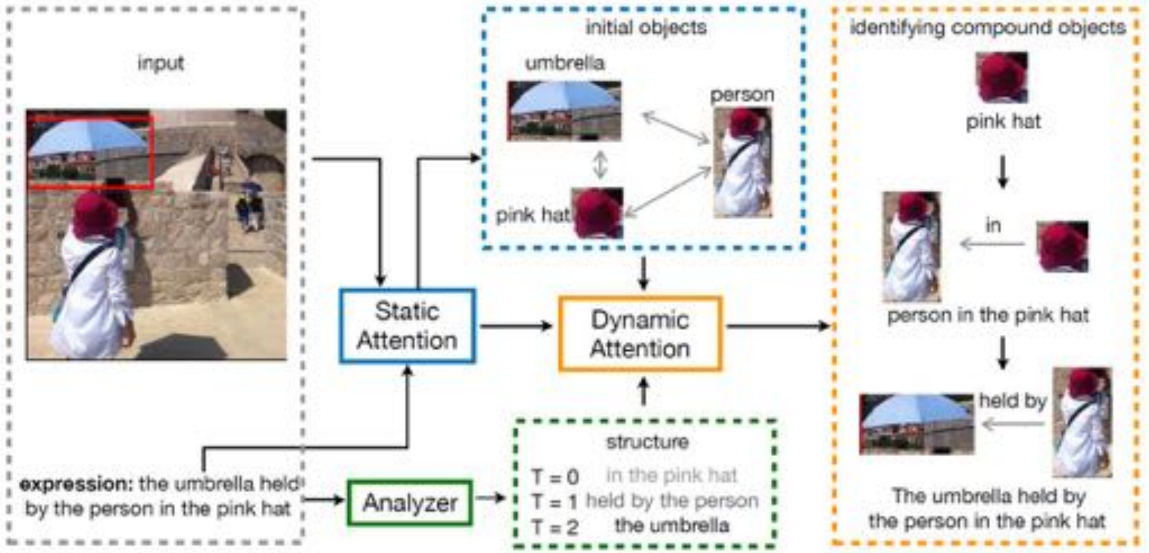
Dynamic Graph Attention (ICCV2019)

Motivation:

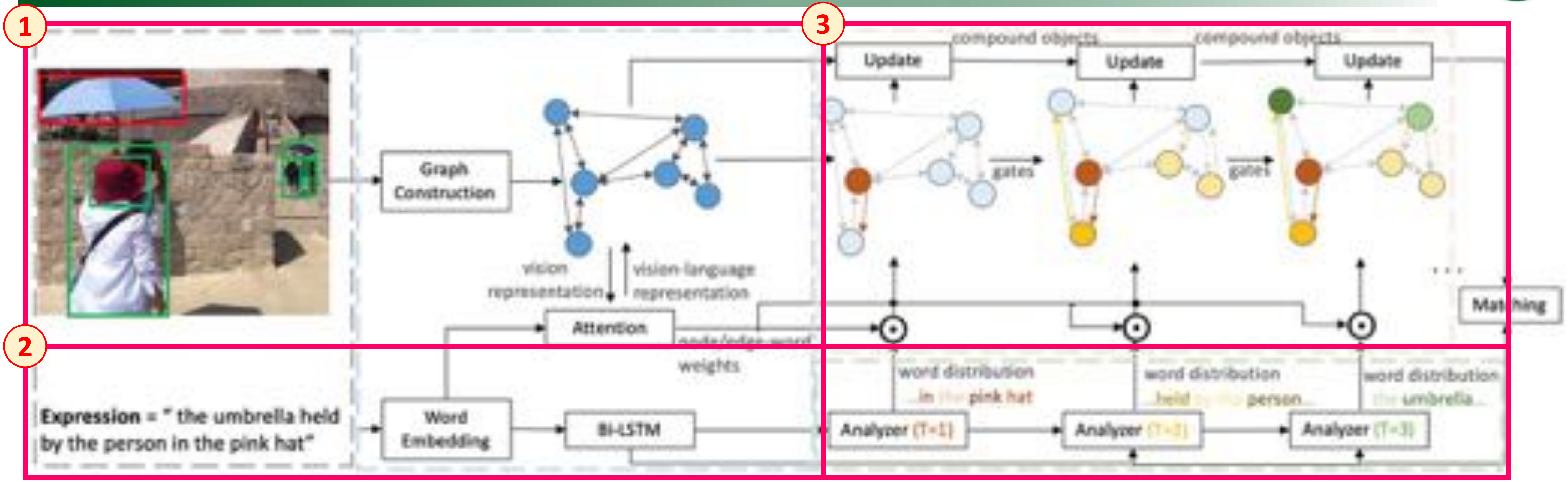
- Referring expression comprehension inherently requires visual reasoning on top of the relationships among the objects in the image. Example "the umbrella held by the person in the pink hat"
- Human visual reasoning of grounding is guided by the linguistic structure of the referring expression.

Our Proposed Method:

- Specify the reasoning process as a sequence of constituent expressions.
- A dynamic graph attention network to perform multi-step visual reasoning to identify compound objects by following the predicted reasoning process.



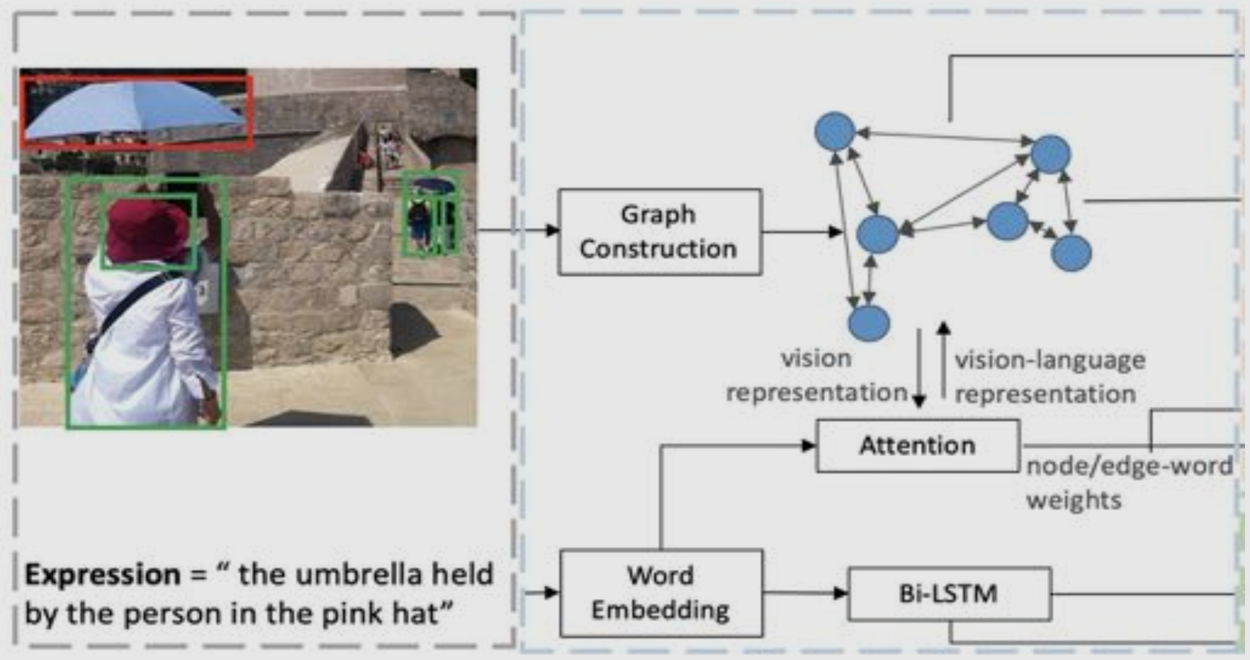
Dynamic Graph Attention Network



1. Graph construction
   - ➢Visual graph → Multi-modal graph
2. Linguistic structure analysis
   - ➢Constituent expressions → Guidance of reasoning

3. Step-wisely dynamic reasoning
   - ➢ performs on the top of the graph under the guidance
   - ➢ highlight edges and nodes → identify compound objects

Graph construction



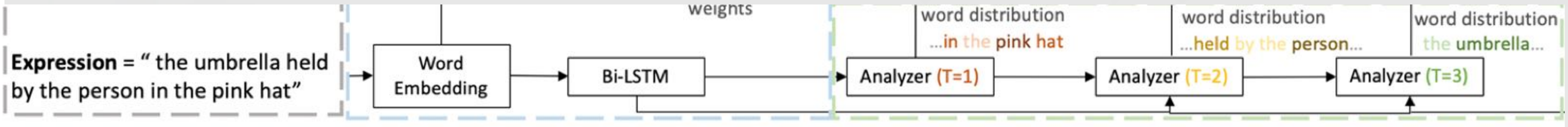
Directed graph: $G^I = (V, E, \boldsymbol{X}^I)$ $\boldsymbol{x}_k^I = [\boldsymbol{x}_k^o; \boldsymbol{p}_k]$ $\boldsymbol{p}_k = \boldsymbol{W}_p[x_{0k}, x_{1k}, w_k, h_k, w_k h_k]$

Multi-modal graph: $G^M = (V, E, \boldsymbol{X}^M)$ $a_{k,l} = \boldsymbol{W}_{\alpha 2}[\tanh(\boldsymbol{W}_{\alpha 1}\boldsymbol{x}_k^I + \boldsymbol{W}_{\alpha 0}\boldsymbol{f}_l)]$

word embedding $\boldsymbol{F} = \{\boldsymbol{f}_l\}_{l=1}^L$ $\alpha_{k,l} = z_{0l}\dfrac{\exp(a_{k,l})}{\sum_{k=1}^K \exp(a_{k,l})}$

language representation $\boldsymbol{c}_k$ at node v_k: $\boldsymbol{c}_k = \sum_{l=1}^L \alpha_{k,l}\boldsymbol{f}_l$ $\boldsymbol{x}_k^M = \boldsymbol{W}_m[\boldsymbol{x}_k^I; \boldsymbol{c}_k] + \boldsymbol{b}_m$

Language Guided Visual Reasoning Process



Model expression as a sequence of constituent expressions (soft distribution over words in the expression)

$$R^{(t)} = \{r_l^{(t)}\}_{l=1}^{L}$$

$$\boldsymbol{F} = \{\boldsymbol{f}_l\}_{l=1}^{L} \xrightarrow{\text{bi-directional LSTM}} \boldsymbol{H} = \{\boldsymbol{h}_l\}_{l=1}^{L} \xrightarrow{\text{overall expression}} \boldsymbol{q}$$

$$\boldsymbol{q}^{(t)} = \boldsymbol{W}^{(t)}\boldsymbol{q} + \boldsymbol{b}^{(t)}$$

$$\boldsymbol{u}^{(t)} = [\boldsymbol{q}^{(t)}; \boldsymbol{y}^{(t-1)}]$$

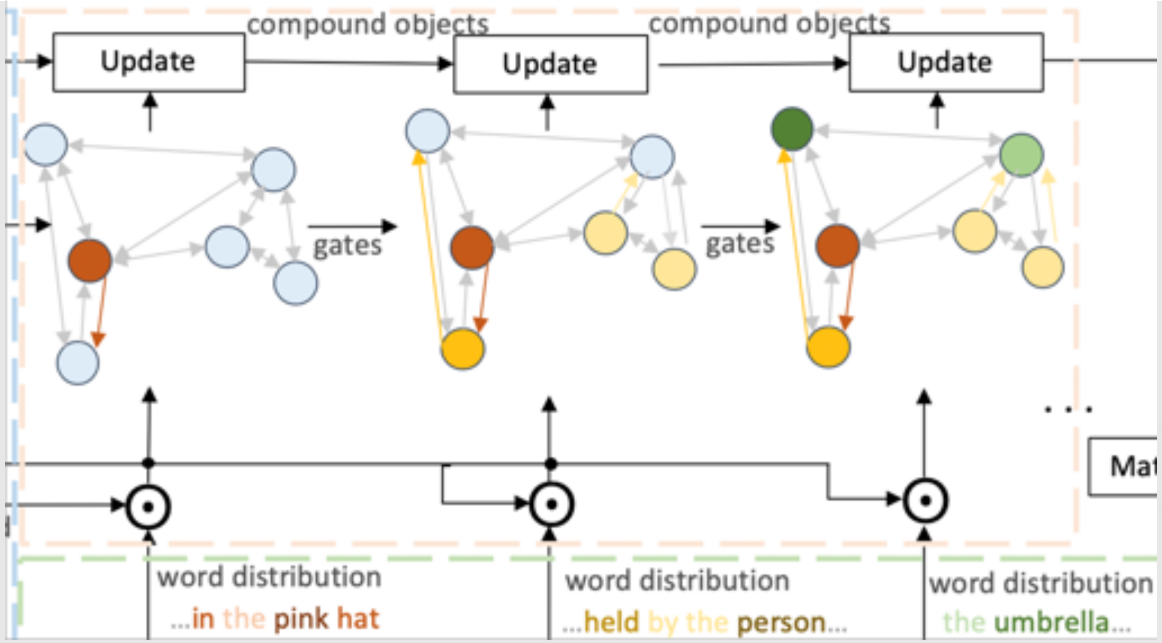
$$\boldsymbol{s}^{(t)} = \mathrm{relu}(\boldsymbol{W}_u \boldsymbol{u}^{(t)} + \boldsymbol{b}_u)$$

$$a_l^{(t)} = \boldsymbol{W}_{s2}[\tanh(\boldsymbol{W}_{s0}\boldsymbol{s}^{(t)} + \boldsymbol{W}_{s1}\boldsymbol{h}_l)]$$

$$r_l^{(t)} = \frac{\exp(a_l^{(t)})}{\sum_{l=1}^{L}\exp(a_l^{(t)})}$$

$$\boldsymbol{y}^{(t)} = \sum_{l=1}^{L} r_l^{(t)} \boldsymbol{h}_l$$

Step-wisely Dynamic Reasoning



The probability of the l-th word referring to each node and type of edge: $\gamma_{k,l}^{(t)} = r_l^{(t)} \alpha_{k,l},\ \delta_{n,l}^{(t)} = r_l^{(t)} \beta_{n,l}$

The weight of each node (or the edge type) being mentioned in time step: $\lambda_k^{(t)} = \sum_{l=1}^{L} \gamma_{k,l}^{(t)},\ \mu_n^{(t)} = \sum_{l=1}^{L} \delta_{n,l}^{(t)}$

Update the gates for every node or the edge type: $p_k^{(t)} = \lambda_k^{(t)} + p_k^{(t-1)},\ \nu_n^{(t)} = \mu_n^{(t)} + \nu_n^{(t-1)}$

Identify the compound object corresponding to each node:

$$\overleftarrow{\boldsymbol{m}}_k^{(t)} = \sum_{e_{j,k}>0} \nu_{e_{j,k}}^{(t)} (\overleftarrow{\boldsymbol{W}} \boldsymbol{m}_j^{(t-1)} p_j^{(t-1)} + \overleftarrow{\boldsymbol{b}}_{e_{j,k}})$$

$$\widetilde{\boldsymbol{m}}_k^{(t)} = \widetilde{\boldsymbol{W}} \boldsymbol{m}_k^{(t-1)} + \widetilde{\boldsymbol{b}},$$

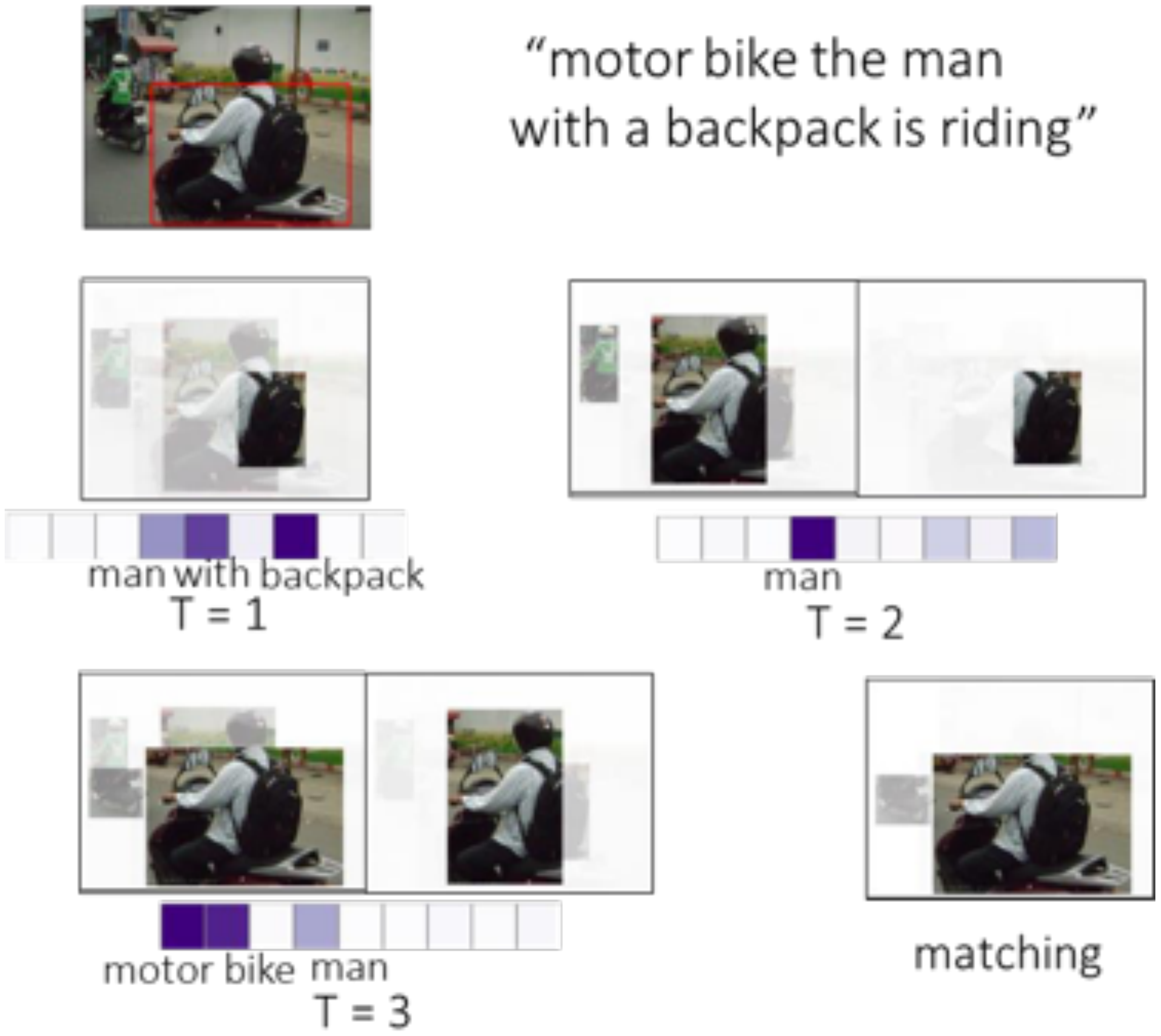
$$\boldsymbol{m}_k^{(t)} = \frac{\lambda_k^{(t)} (\hat{\boldsymbol{W}} (\overleftarrow{\boldsymbol{m}}_k^{(t)} + \widetilde{\boldsymbol{m}}_k^{(t)}) + \hat{\boldsymbol{b}}) + p_k^{(t-1)} \boldsymbol{m}_k^{(t-1)}}{p_k^{(t)}}$$

Experiments

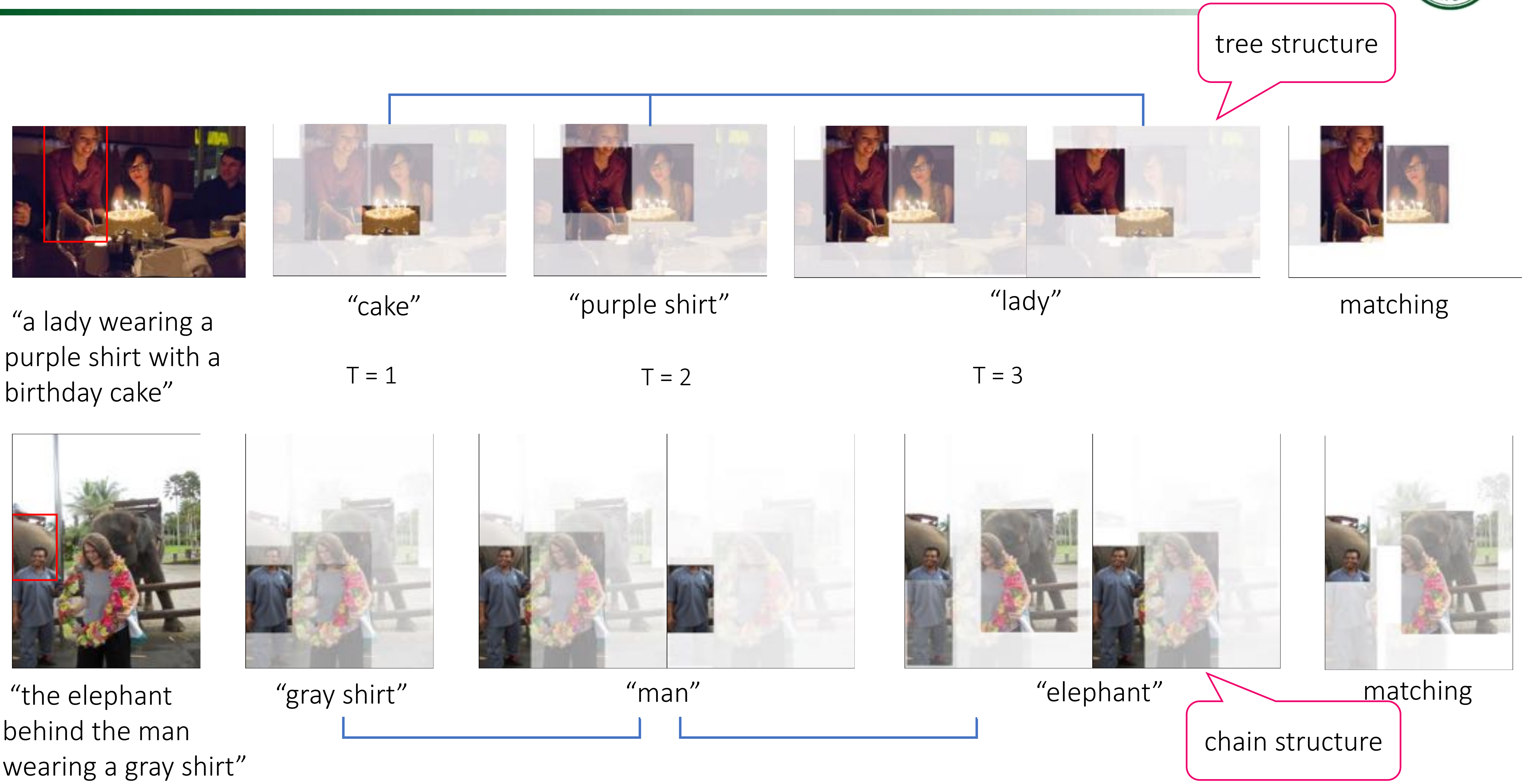
	feature	RefCOCO			RefCOCO+			RefCOCOg	
		val	testA	testB	val	testA	testB	val	test
MMI [18]	vgg16	-	63.15	64.21	-	48.73	42.13	-	-
Neg Bag [19]	vgg16	76.90	75.60	78.00	-	-	-	-	68.40
CG [16]	vgg16	-	74.04	73.43	-	60.26	55.03	-	-
Attr [13]	vgg16	-	78.85	78.07	-	61.47	57.22	-	-
CMN [7]	vgg16	-	75.94	79.57	-	59.29	59.34	-	-
Speaker [31]	vgg16	76.18	74.39	77.30	58.94	61.29	56.24	-	-
Spearker+**Listener**+Reinforcer[32]	vgg16	78.36	77.97	79.86	61.33	63.10	58.19	71.32	71.72
Speaker+Listener+Reinforcer [32]	vgg16	79.56	78.95	80.22	62.26	64.60	59.62	71.65	71.92
AccumulateAttn [4]	vgg16	81.27	81.17	80.01	65.56	68.76	60.63	-	-
ParallelAttn [33]	vgg16	81.67	80.81	81.32	64.18	66.31	61.46	-	-
MAttNet [30]	vgg16	80.94	79.99	82.30	63.07	65.04	61.77	73.04	72.79
Ours DGA	vgg16	**83.73**	**83.56**	**82.51**	**68.99**	**72.72**	**62.98**	**75.76**	**75.79**
MAttNet [30]	resnet101	85.65	85.26	84.57	71.01	75.13	66.17	78.10	78.12
Ours DGA	resnet101	**86.34**	**86.64**	**84.79**	**73.56**	**78.31**	**68.15**	**80.21**	**80.26**

Comparison with state-of-the-art methods on RefCOCO, RefCOCO+ and RefCOCOg when ground-truth bounding boxes are used.

Explainable Visualization



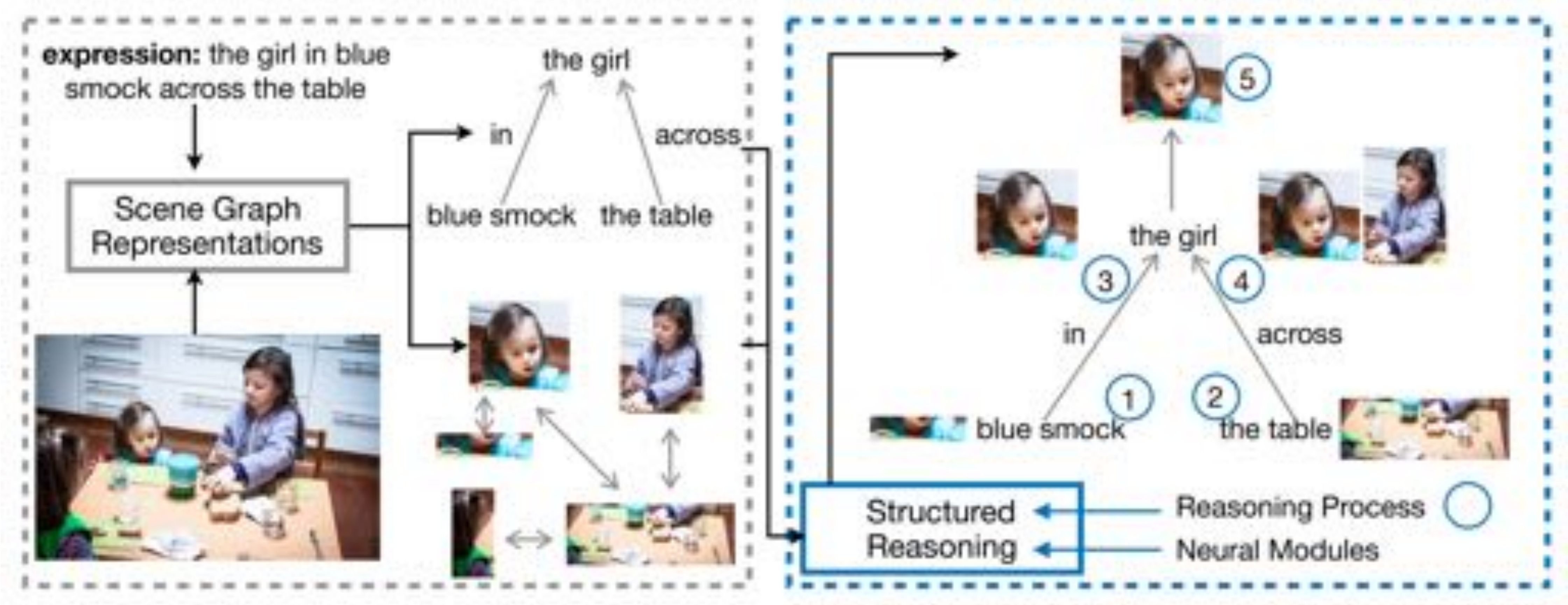
Visualization Results



Outline

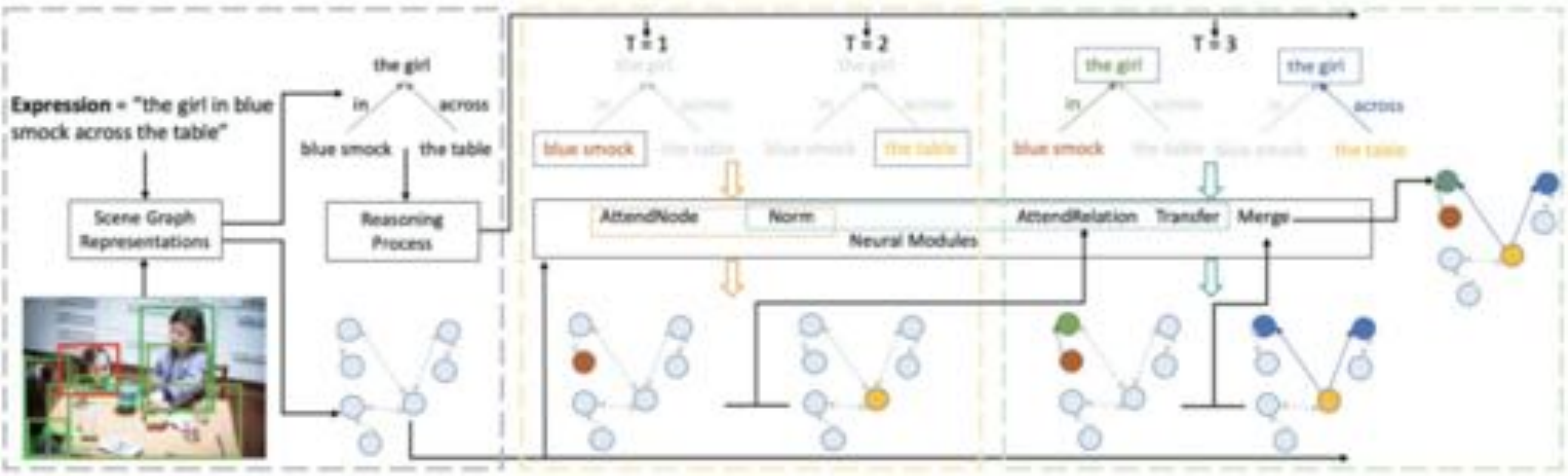
- Introduction and Related Work
- Cross-Modal Relationship Inference Network, CVPR 2019
- Dynamic Graph Attention for Visual Reasoning, ICCV2019
- Scene Graph guided Visual Reasoning, CVPR2020
- Conclusion and Future Work Discussion

Scene Graph guided Modular Network



Performs structured reasoning with neural modules under the guidance of the language scene graph

Scene Graph guided Modular Network



Overview of our Scene Graph guided Modular Network (SGMN)

Scene Graph Representations

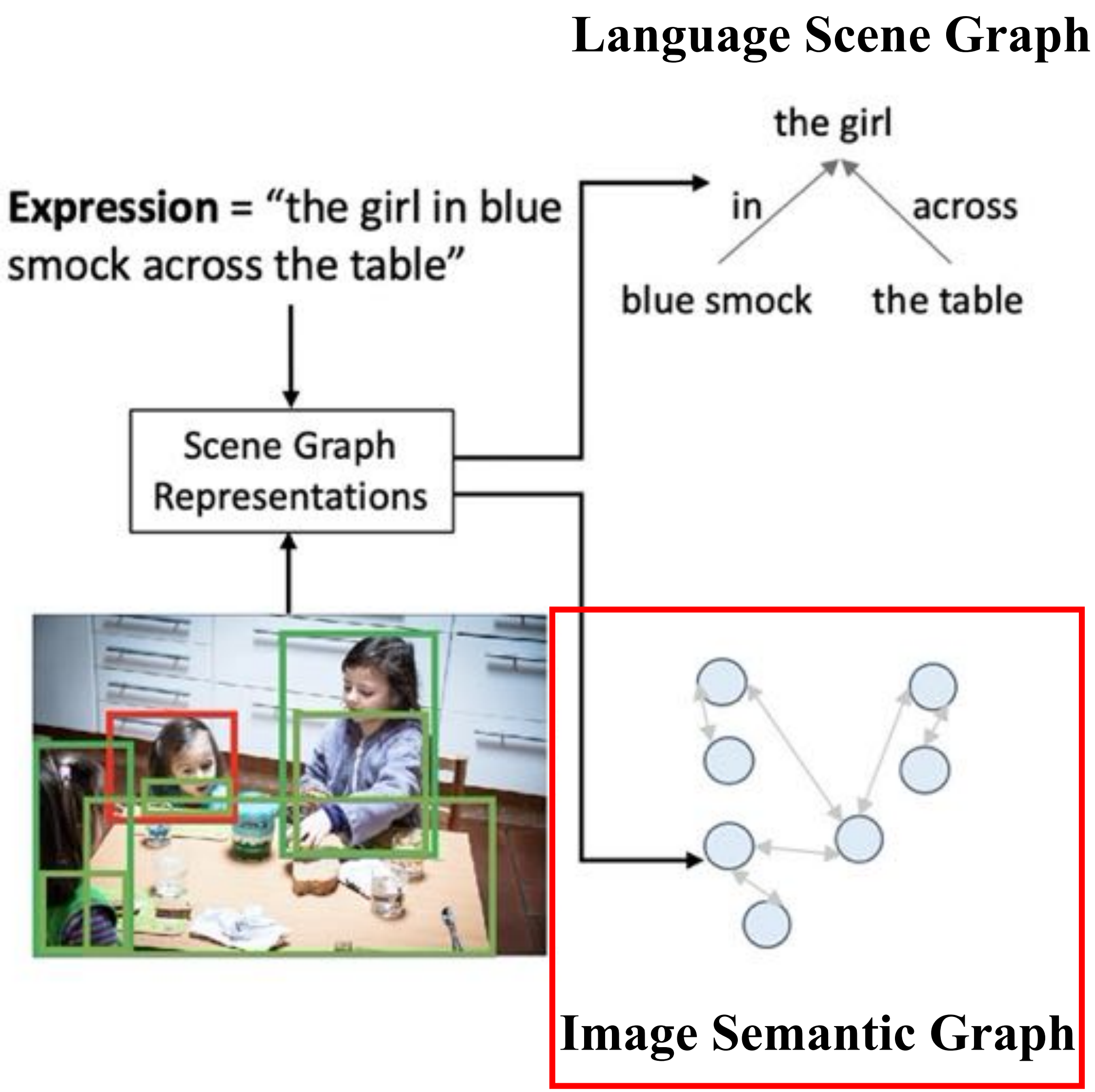


Image Semantic Graph:

$$\mathcal{G}^o = (\mathcal{V}^o, \mathcal{E}^o) \qquad \mathcal{V}^o = \{v_i^o\}_{i=1}^N$$

$$\mathcal{E}^o = \{e_{ij}^o\}_{i,j=1}^N$$

Node:

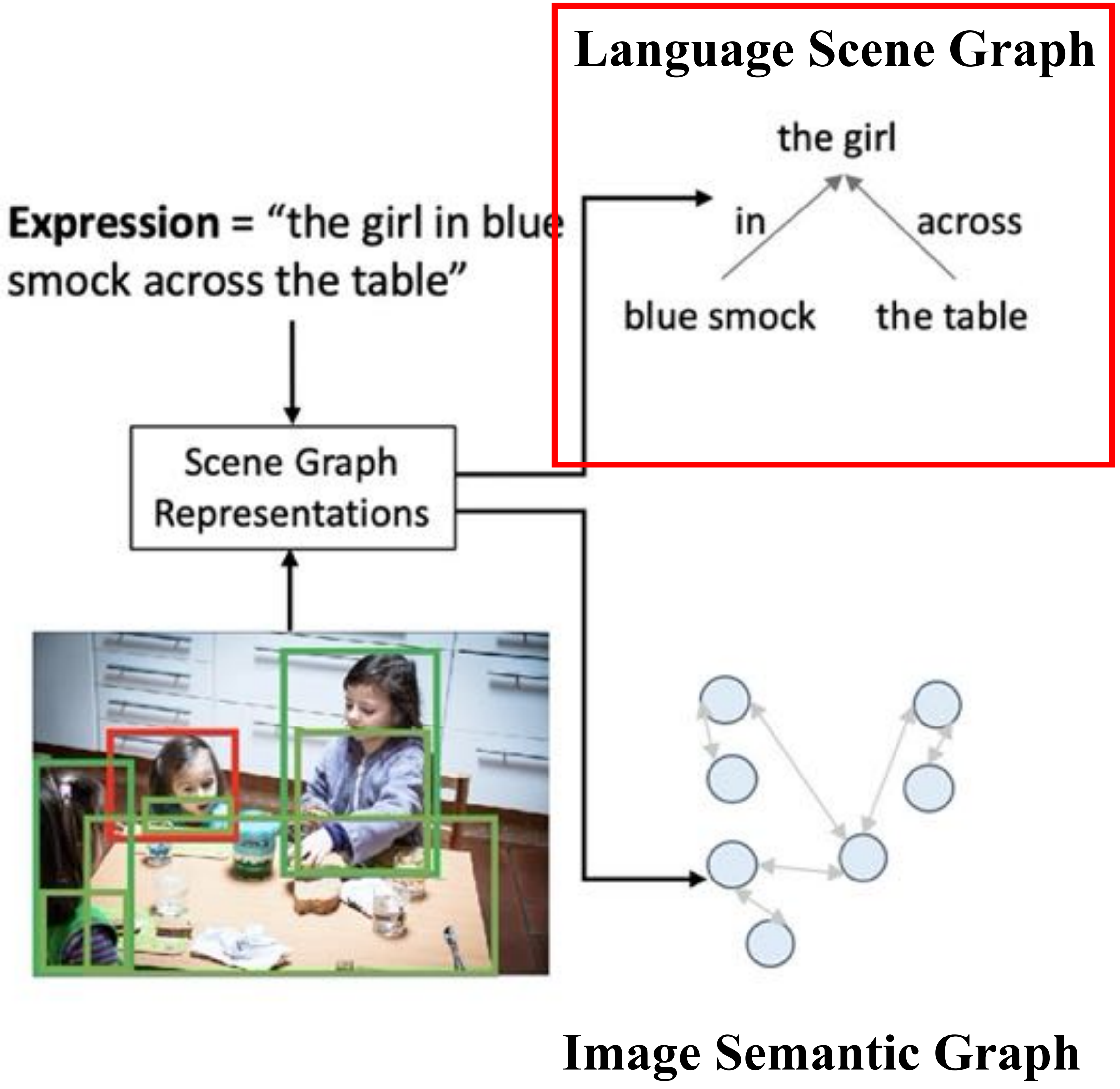
Visual feature: $\mathbf{v}_i^o$

Spatial feature: $\mathbf{p}_i^o = [x_i, y_i, w_i, h_i, w_i h_i]$

Edge feature: $\mathbf{e}_{ij}^o = [\mathbf{W}_o^T \mathbf{l}_{ij}^o, \mathbf{v}_j^o]$

$$\mathbf{l}_{ij}^o = [\frac{x_j - x_{ci}}{w_i}, \frac{y_j - y_{ci}}{h_i}, \frac{x_j + w_j - x_{ci}}{w_i}, \frac{y_j + h_j - y_{ci}}{h_i}, \frac{w_j h_j}{w_i h_i}]$$

Scene Graph Representations



Language Scene Graph

$$\mathcal{G} = (\mathcal{V}, \mathcal{E})$$

$\mathcal{V} = \{v_m\}_{m=1}^{M}$ noun or noun phrase

$\mathcal{E} = \{e_k\}_{k=1}^{K}$ $\quad e_k = (v_{k\,s}, r_k, v_{k\,o})$

Relation r_k a preposition/verb word or phrase

e_k indicates that subject node $v_{k\,s}$ is modified by object node

Structured Reasoning

the girl in blue smock across the table

the girl

in

across

blue smock

the table

breadth-first traversal

stack

stack
the table
blue smock
the girl

a boy who is to the left of a skater and is wearing dark t-shirt, and the skater is on a skateboard

is to the left of

a skater

a boy

on

is wearing

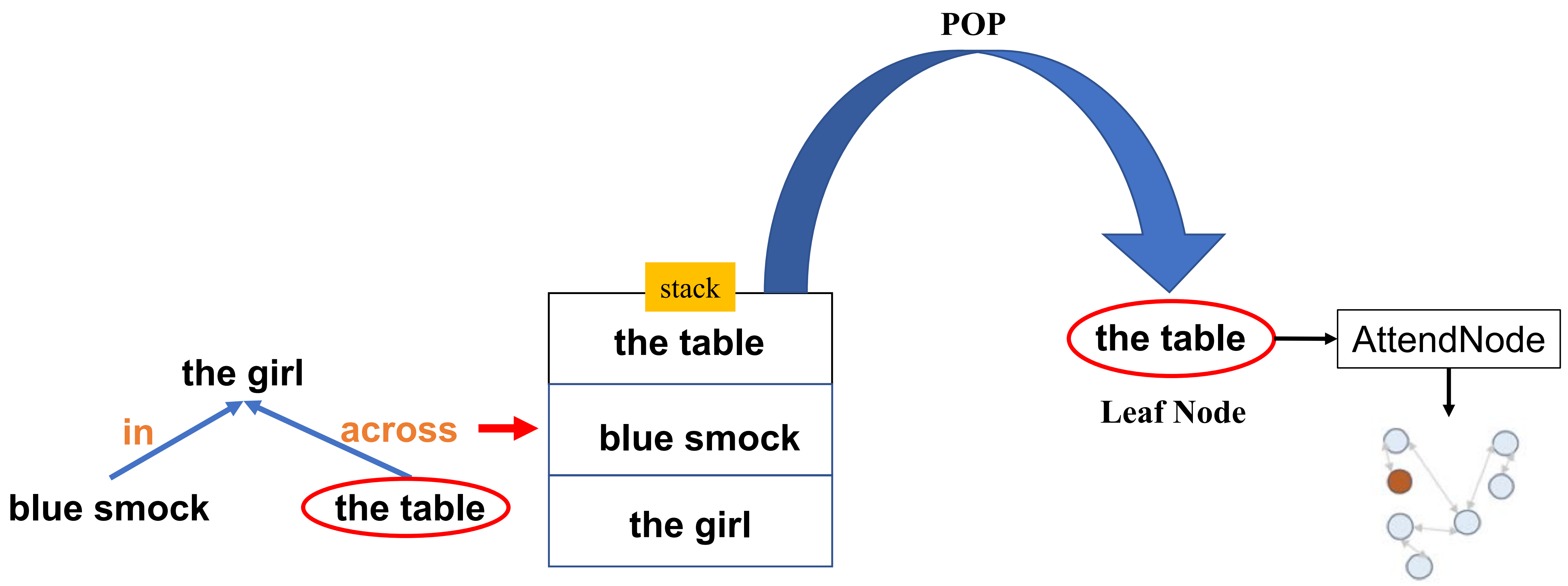
a skateboard

dark t-shirt

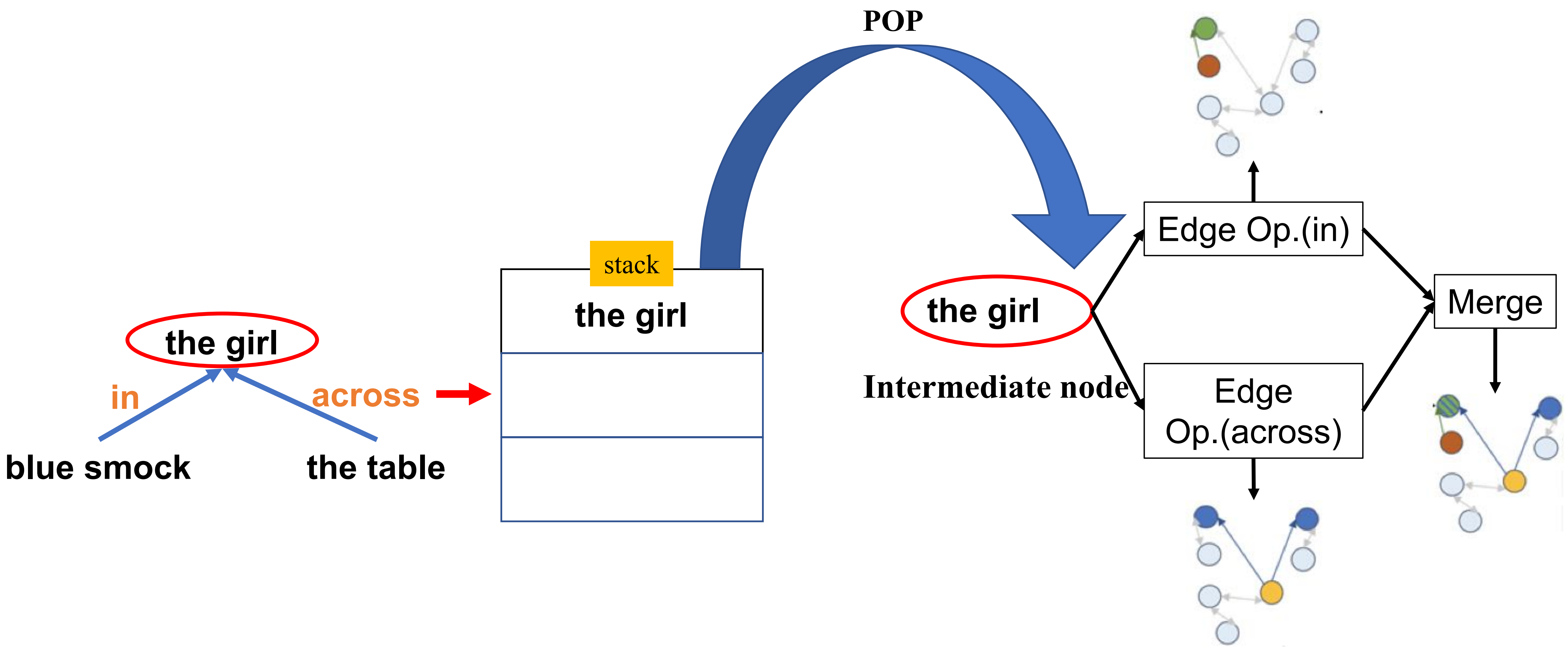
breadth-first traversal

stack
a skateboard
dark t-shirt
a skater
a boy

Structured Reasoning



Structured Reasoning



Leaf node operation

Given node v_m, with its associated phrase consists of words $\{w_t\}_{t=1}^{T}$

Embedded feature vectors: $\{\mathbf{f}_t\}_{t=1}^{T}$

Bi-directional LSTM for context feature representation: $\mathbf{h}_t$

represent the whole phrase feature as: $\mathbf{h}$

An individual entity is often described by its appearance and spatial location. We learn feature representations for node v_m from both appearance and spatial location:

$$\alpha_{t,m}^{look} = \frac{\exp(\mathbf{W}_{look}^T \mathbf{h}_t)}{\sum_{t=1}^{T}\exp(\mathbf{W}_{look}^T \mathbf{h}_t)}, \mathbf{v}_m^{look} = \sum_{t=1}^{T} \alpha_{t,m}^{look}\mathbf{f}_t$$

$$\alpha_{t,m}^{loc} = \frac{\exp(\mathbf{W}_{loc}^T \mathbf{h}_t)}{\sum_{t=1}^{T}\exp(\mathbf{W}_{loc}^T \mathbf{h}_t)}, \mathbf{v}_m^{loc} = \sum_{t=1}^{T} \alpha_{t,m}^{loc}\mathbf{f}_t,$$

$\longrightarrow$ AttendNode

$$\{\lambda_{n,m}^{look}\}_{n=1}^{N}$$

$$\{\lambda_{n,m}^{loc}\}_{n=1}^{N}$$

$$\beta^{look} = \mathrm{sigmoid}(\mathbf{W}_0^T \mathbf{h} + b_0)$$

$$\beta^{loc} = \mathrm{sigmoid}(\mathbf{W}_1^T \mathbf{h} + b_1)$$

$$\boxed{\begin{aligned}\lambda_{n,m} &= \beta^{look}\lambda_{n,m}^{look} + \beta^{loc}\lambda_{n,m}^{loc}\\ \{\lambda_{n,m}\}_{n=1}^{N} &= \mathrm{Norm}(\{\lambda_{n,m}\}_{n=1}^{N})\end{aligned}}$$

Intermediate node operation

Intermediate node v_m is connected to nodes that modify it, denote the connected edge subset as: $\mathcal{E}_m \in \mathcal{E}$

For each edge $e_k = (v_{k\,s}, r_k, v_{k\,o})$, form an associated sentence by concatenating the words or phrases

Obtain embedded feature vectors: $\{\mathbf{f}_t\}_{t=1}^{T}$, Bi-directional LSTM for context feature representation: $\mathbf{h}_t$

Compute the attention map for node $v_{k\,s}$ from both subject description and relation-based transfer

For subject description, compute as Leaf Operation and obtain: $\{\lambda_{n,k_s}^{look}\}_{n=1}^{N}$ and $\{\lambda_{n,k_s}^{loc}\}_{n=1}^{N})$

For relation-based transfer, relational feature representation $\alpha_{t,k}^{rel} = \dfrac{\exp(\mathbf{W}_{rel}^{T}\mathbf{h}_t)}{\sum_{t=1}^{T}\exp(\mathbf{W}_{rel}^{T}\mathbf{h}_t)}, \mathbf{r}_k = \sum_{t=1}^{T}\alpha_{t,k}^{rel}\mathbf{f}_t$

$\boxed{\mathbf{r}_k}$ → AttendRelation → $\boxed{\{\gamma_{ij,k}\}_{i,j=1}^{N}}$

$\boxed{\{\gamma_{ij,k}\}_{i,j=1}^{N}}$, $\boxed{\{\lambda_{n,k_o}\}_{n=1}^{N}}$ → Transfer → Norm → $\boxed{\{\lambda_{n,k_s}^{rel}\}_{n=1}^{N}}$

$$\beta_k^{rel} = \text{sigmoid}(\mathbf{W}_2^T\mathbf{h} + b_2)$$
$$\lambda_{n,k_s} = \beta_{k_s}^{look}\lambda_{n,k_s}^{look} + \beta_{k_s}^{loc}\lambda_{n,k_s}^{loc} + \beta_{n,k_s}^{rel}\lambda_{n,k_s}^{rel}$$
$$\{\lambda_{n,k_s}\}_{n=1}^{N} = \text{Norm}(\{\lambda_{n,k_s}\}_{n=1}^{N})$$

$\{\{\lambda_{n,k_s}\}_{n=1}^{N}\}$ → Merge → Norm → $\boxed{\{\lambda_{n,m}\}_{n=1}^{N}}$

Neural Modules

AttendNode [appearance query, location query]:

$$\lambda_n^{look} = \langle \mathrm{L2Norm}(\mathrm{MLP}_0(\mathbf{v}_n^o)), \mathrm{L2Norm}(\mathrm{MLP}_1(\mathbf{v}^{look})) \rangle$$
$$\lambda_n^{loc} = \langle \mathrm{L2Norm}(\mathrm{MLP}_2(\mathbf{p}_n^o)), \mathrm{L2Norm}(\mathrm{MLP}_3(\mathbf{v}^{loc})) \rangle$$

AttendRelation [relation query]:

$$\gamma_{ij} = \sigma(\langle \mathrm{L2Norm}(\mathrm{MLP}_5(\mathbf{e}_{ij}^o)), \mathrm{L2Norm}(\mathrm{MLP}_1(\mathbf{e})) \rangle)$$

Transfer: $\lambda_n^{new} = \sum_{j=1}^{N} \gamma_{n,j} \lambda_j$ **Merge:** $\lambda_n = \sum_{\{\lambda'_n\}_{n=1}^N \in \Lambda} \lambda'_n$

Norm: Rescale attention maps to [-1, 1].

Loss Function

The final attention map for the referent node is obtained: $\{\lambda_{n,ref}\}_{n=1}^{N}$

Adopt the cross-entropy loss for training: $p_i = \exp(\lambda_{i,ref}) / \sum_{n=1}^{N} \exp(\lambda_{n,ref}), \text{loss} = -\log(p_{gt})$

p_{gt} is the probability of the ground-truth object

During inference, choose the object with the highest probability.

Ref-Reasoning Dataset

Motivation:

- Dataset biases exist
- Samples in existing datasets have unbalanced levels of difficulty
- Evaluation is only conducted on final predictions but not on intermediate reasoning process

Ref-Reasoning Dataset:

Built on the scenes from the GQA dataset.

Generate referring expressions according to the ground-truth image scene graphs.

Design a family of referring expression templates for each reasoning layout.

During expression generation: (the referent node + a sub-graph + a template), check uniqueness.

Define the difficulty level as the shortest sub-expression which can identify the referent in the scene graph.

Experimental Datasets

Dataset	Specification
RefCOCO	➢ 142,210 expression referent pairs in 19,994 images ➢ Average length of expression < 4
RefCOCO+	➢ 141,564 expression-referent pairs in 19,992 images ➢ Forbids describing the absolute locations ➢ Average length of expression < 4
RefCOCOg	➢ 95,010 expression-referent pairs from 25,799 images ➢ Average length of expression 8.43
Ref-Reasoning	**➢ 810,012 referring expressions in 195,288 images** **➢ Semantically rich expressions describing objects, attributes, direct relations and indirect relations with different layouts**

Experiments

Comparison with baselines and state-of-the-art methods on Ref-Reasoning dataset

	Number of Objects				Split	
	one	two	three	>= four	val	test
CNN	10.57	13.11	14.21	11.32	12.36	12.15
CNN+LSTM	75.29	51.85	46.26	32.45	42.38	42.43
DGA	73.14	54.63	48.48	37.63	45.37	45.87
CMRIN	79.20	56.87	50.07	35.29	45.43	45.87
Ours SGMN	**79.71**	**61.77**	**55.57**	**41.89**	**51.04**	**51.39**

- The CNN model 12.15%, much lower than 41.1%[4] for the Ref-COCOg dataset.
- CNN+LSTM 75.29% on one-node split (Not require reasoning).
- DGA and CMRIN achieve higher performance on the two-, three and four-node splits because they learn a language-guided contextual representation.

Experiments

Comparison with state-of-the-art methods on
RefCOCO, RefCOCO+ and RefCOCOg

	RefCOCO		RefCOCO+		RefCOCOg
	testA	testB	testA	testB	test
Holistic Models					
CMN [9]	75.94	79.57	59.29	59.34	-
ParallelAttn [29]	80.81	81.32	66.31	61.46	-
MAttNet* [26]	85.26	84.57	75.13	66.17	78.12
CMRIN* [23]	**87.63**	84.73	**80.93**	68.99	80.66
DGA* [24]	86.64	84.79	78.31	68.15	80.26
Structured Models					
MattNet* + parser [26]	79.71	81.22	68.30	62.94	73.72
RvG-Tree* [8]	82.52	82.90	70.21	65.49	75.20
DGA* + parser [24]	84.69	83.69	74.83	65.43	76.33
NMTree* [15]	85.63	85.08	75.74	67.62	78.21
MSGL* [16]	85.45	85.12	75.31	67.50	78.46
Ours SGMN*	86.67	**85.36**	78.66	**69.77**	**81.42**

- SGMN consistently outperforms existing structured methods across all the datasets.
- Holistic models usually have higher performance.
- However, inference mechanism of holistic methods has poor interpretability.

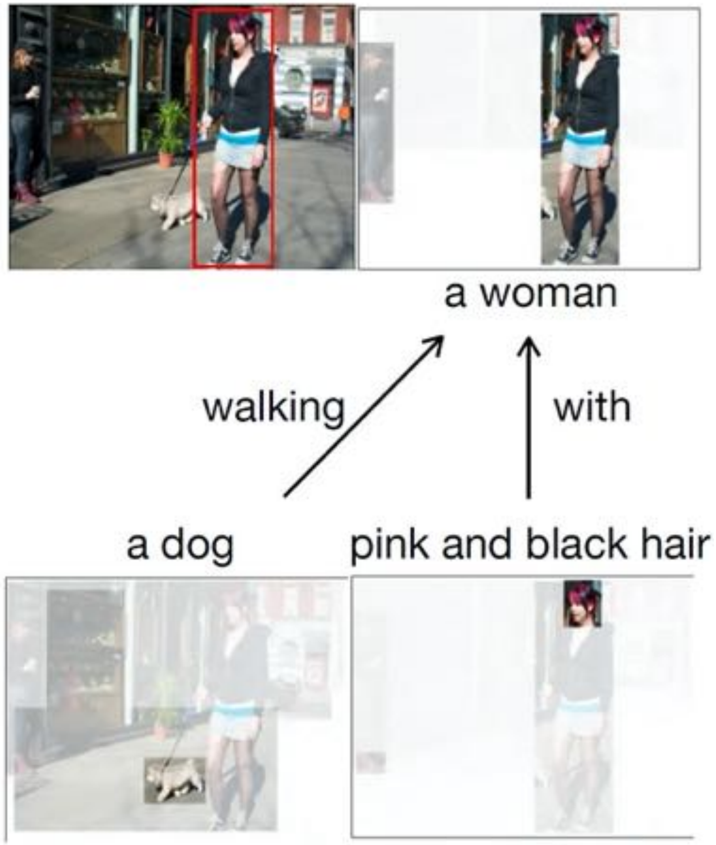
Experiments

Ablation study on the design of neural modules

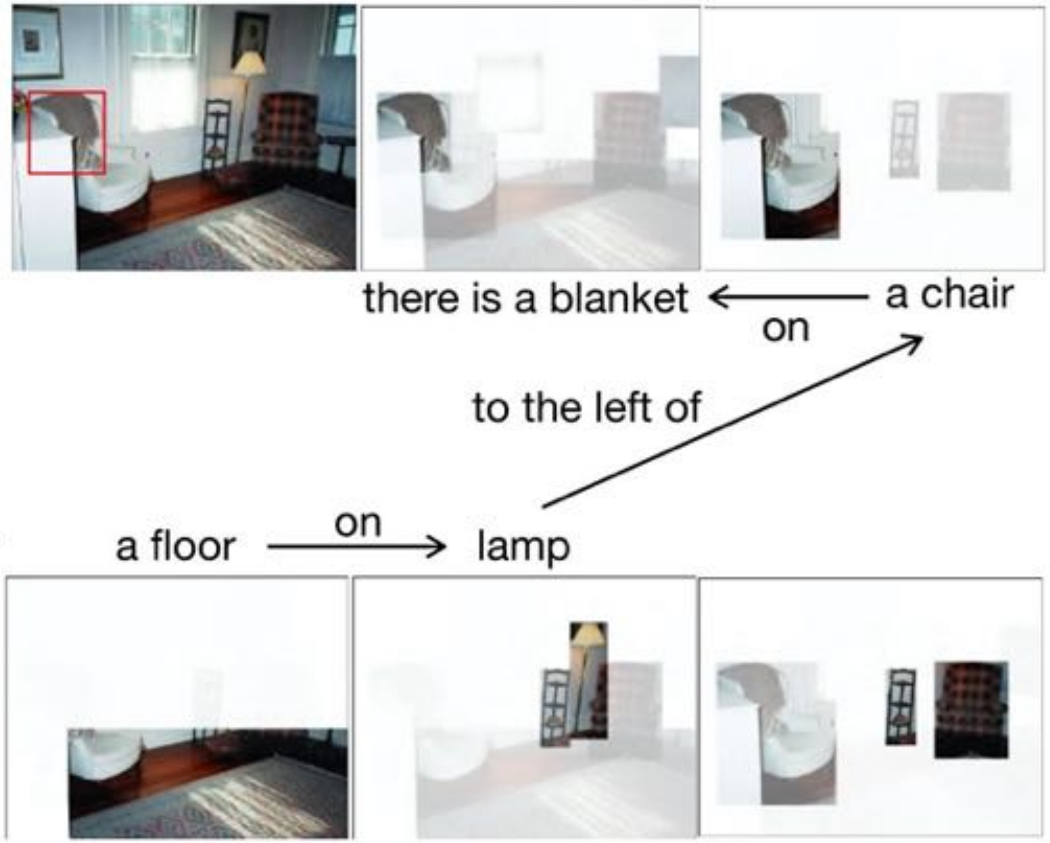
	Number of Objects				Split	
	one	two	three	>= four	val	test
w/o transfer	79.14	48.51	45.97	31.57	40.66	41.88
w/o norm	79.37	49.44	45.61	31.57	40.80	41.93
max merge	78.71	54.00	50.34	34.76	44.50	45.27
min merge	78.83	53.83	51.11	35.79	45.25	46.00
Ours SGMN	**79.71**	**61.77**	**55.57**	**41.89**	**51.04**	**51.39**

- All the models have similar performance on the split of expressions directly describing the referents (one node split).
- SGMN without the Transfer module and without Norm module have much lower performance.
- min-merge and max-merge drops because max-merge only captures the most significant relation and min-merge is sensitive to parsing errors.

Experiments



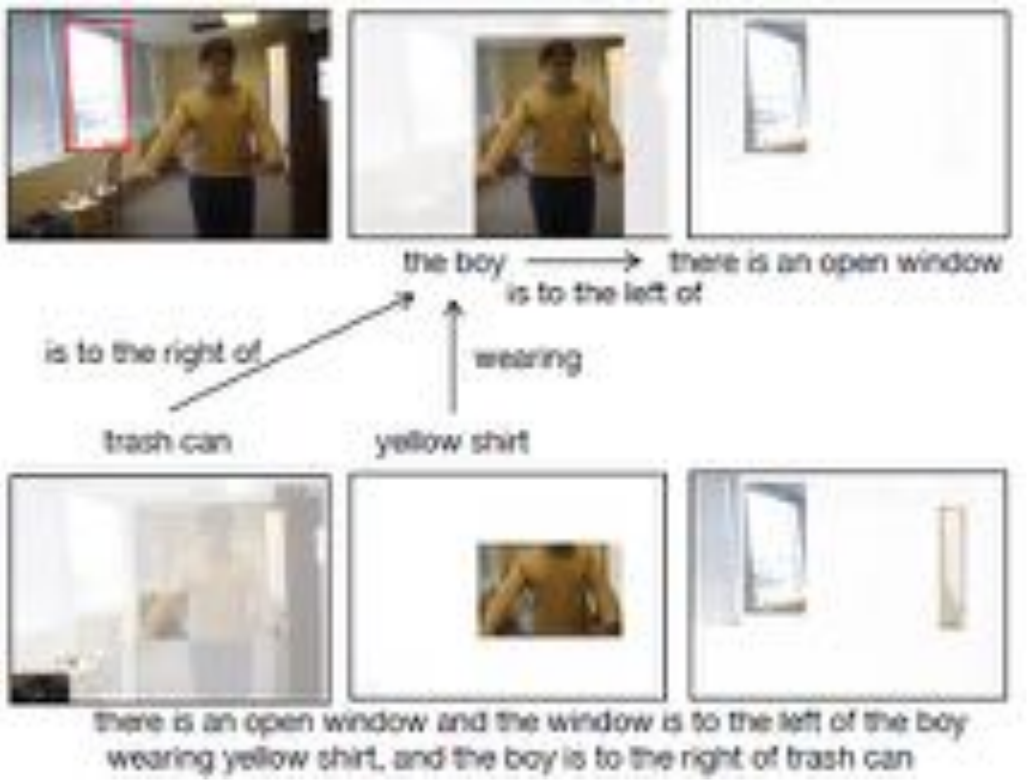
(a) a woman with pink and black hair walking a dog

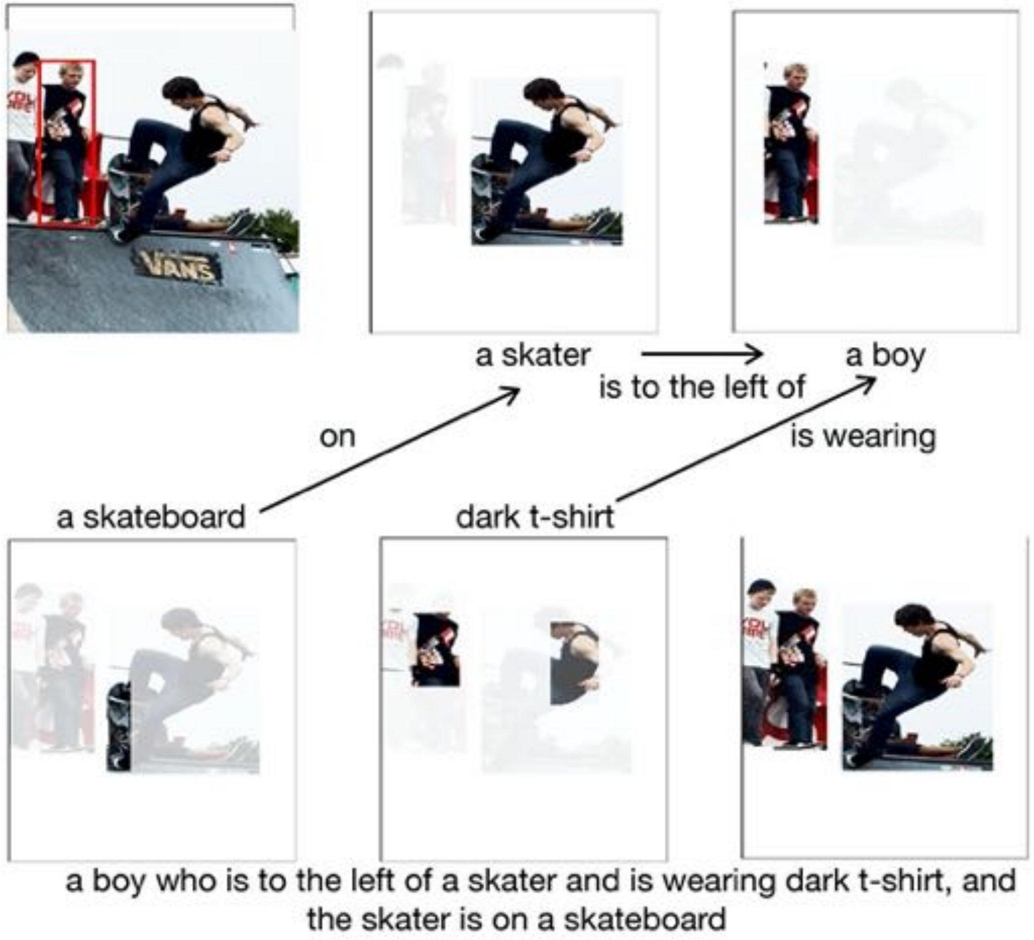


(b) there is a blanket which is on the chair to the left of lamp on a floor

SGMN can generate interpretable visual evidences of intermediate steps in the reasoning process.

Experiments





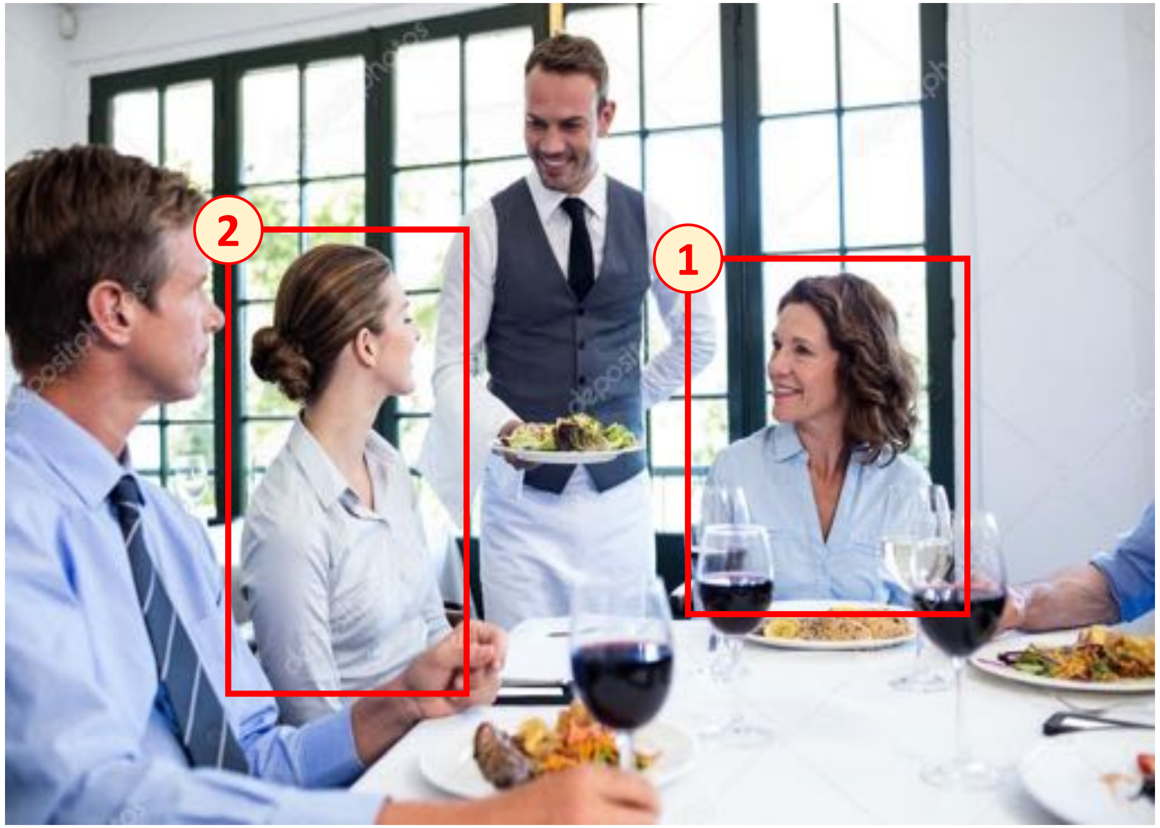
SGMN can generate interpretable visual evidences of intermediate steps in the reasoning process.

Outline

- Introduction and Related Work
- Cross-Modal Relationship Inference Network, CVPR 2019
- Dynamic Graph Attention for Visual Reasoning, ICCV2019
- Scene Graph guided Visual Reasoning, CVPR2020
- Conclusion and Future Work Discussion

Future Work Discussion

Commonsense Reasoning for Visual Grounding



1. The lady to the right of the waiter
2. The person who ordered the dish served by the waiter

From Recognition to Cognition: Visual Commonsense Reasoning, CVPR2019

Future Work Discussion

Task Driven Object Detection

What object in the scene would a human choose to serve wine?

[Sawatzky et al. CVPR2019]

I want to watch the "The Big Bang Theory" now, by the way, the room is too bright.

Thank You!

[1] Sibei Yang, Guanbin Li, Yizhou Yu, "Relationship-Embedded Representation Learning for Grounding Referring Expressions", **T-PAMI**, 2020.

[2] Sibei Yang, Guanbin Li, Yizhou Yu, "Graph Structured Referring Expression Reasoning in The Wild", **CVPR**, **Oral Presentation**, 2020.

[3] Sibei Yang, Guanbin Li, Yizhou Yu, "Dynamic Graph Attention for Referring Expression Comprehension", **ICCV**, **Oral Presentation**, 2019.

[4] Sibei Yang, Guanbin Li, Yizhou Yu, "Cross-Modal Relationship Inference for Grounding Referring Expressions", **CVPR**, 2019.

Source code available at:

https://github.com/sibeiyang/sgmn