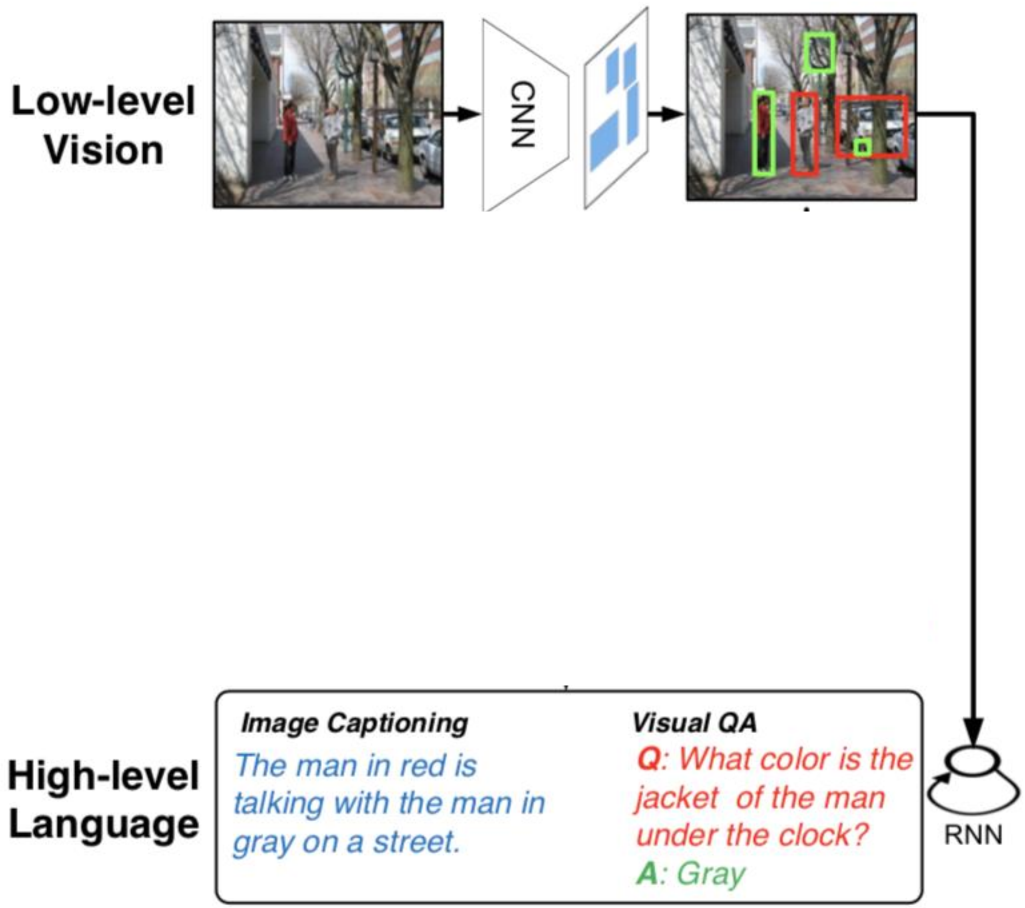
Towards X Visual Reasoning

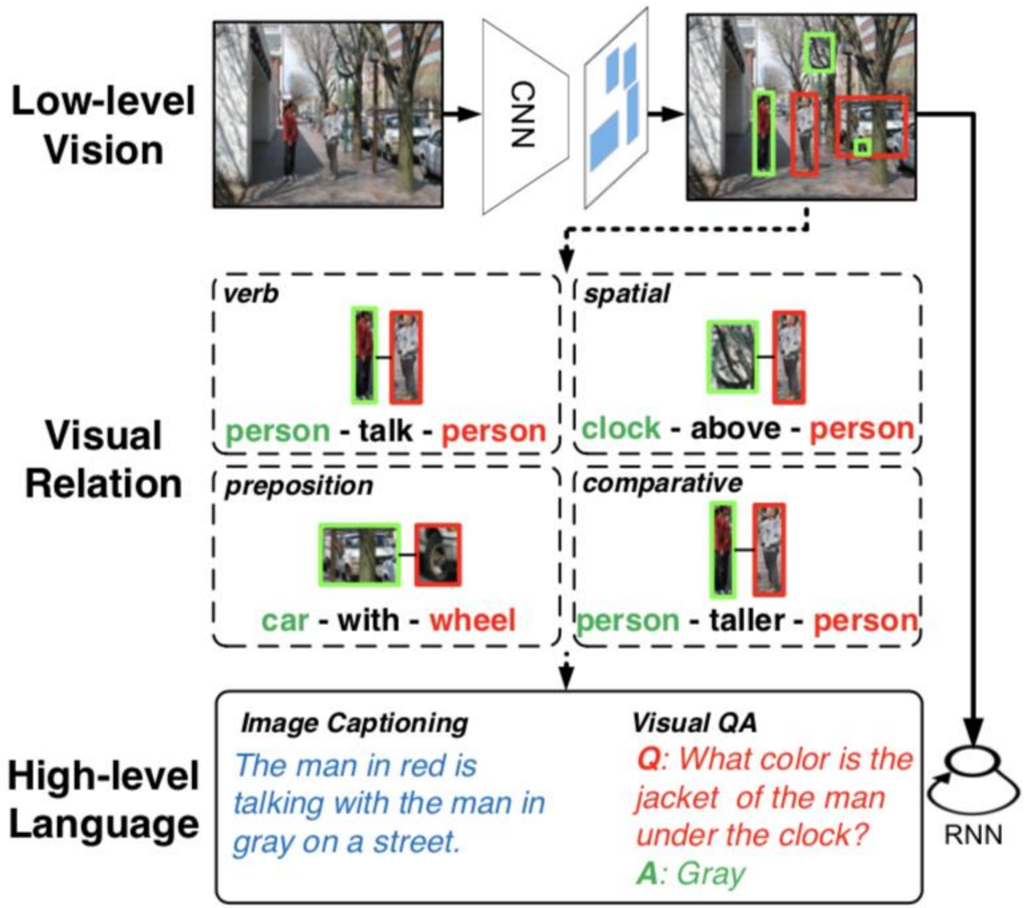
Hanwang Zhang 张含望
hanwangzhang@ntu.edu.sg

NANYANG TECHNOLOGICAL UNIVERSITY
SINGAPORE
School of Computer Science and Engineering

Pattern Recognition v.s. Reasoning



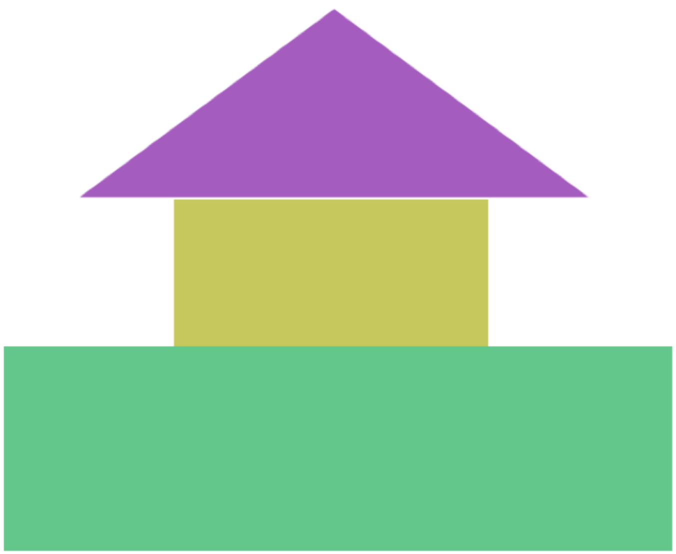
Pattern Recognition v.s. Reasoning



Caption: Lu et al. Neural Baby Talk. CVPR’18
VQA: Teney et al. Graph-Structured Representations for Visual Question Answering. CVPR’17
Cond. Image Generation: Jonson et al. Image Generation from Scene Graphs. CVPR’18

Reasoning: Core Problems

Compositionality



Learning to Reason

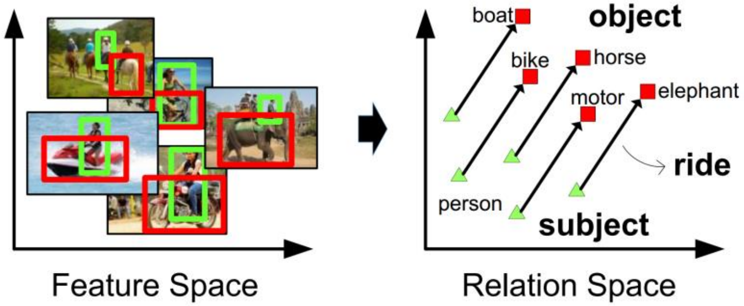
$$1+1=2$$

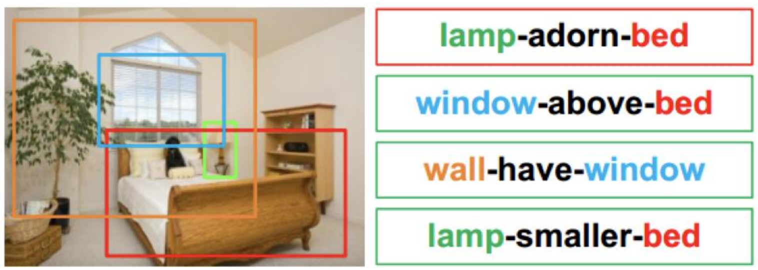
$$a+a=2a$$

NANYANG TECHNOLOGICAL UNIVERSITY

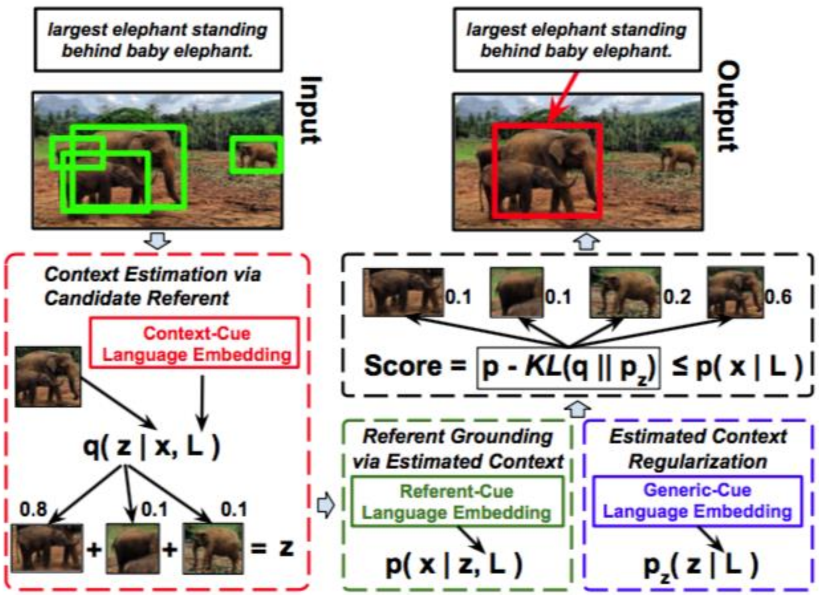
Three Examples

Visual Relation Detection [CVPR'17, ICCV'17]





Referring Expression Grounding [CVPR'1



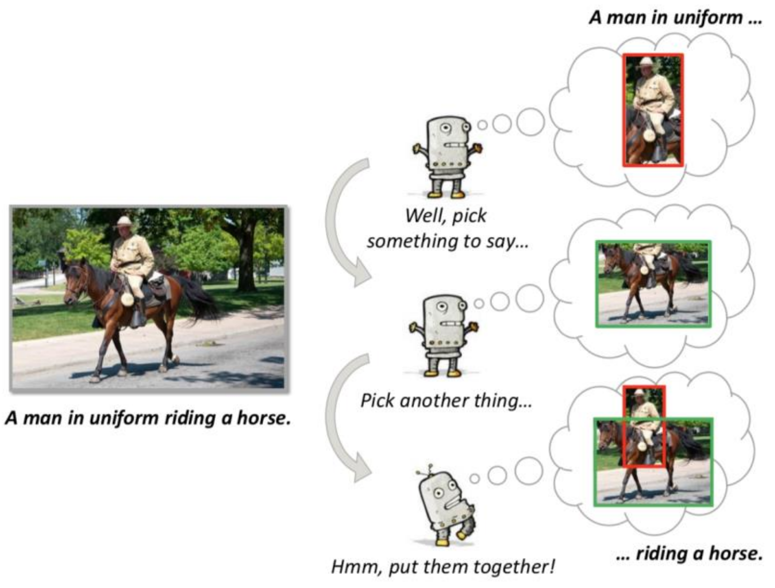
Compositionality

Learning to Reason

NANYANG TECHNOLOGICAL UNIVERSITY

Three Examples

Sequence-level Image Captioning [MM'18 submission]



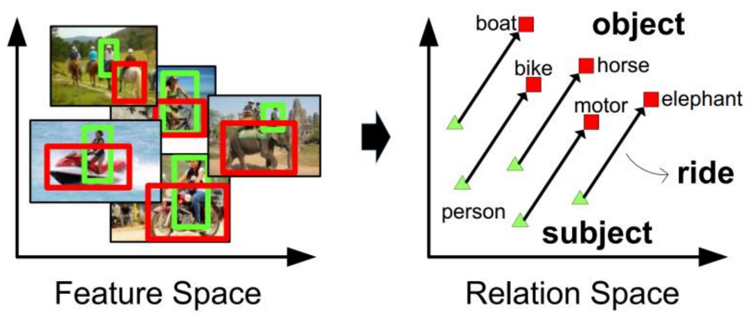
Learning to Reason

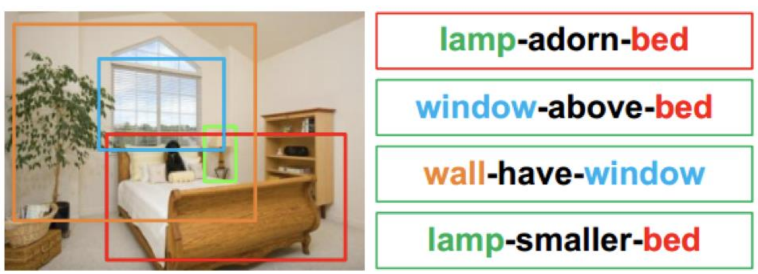
Two Future Works

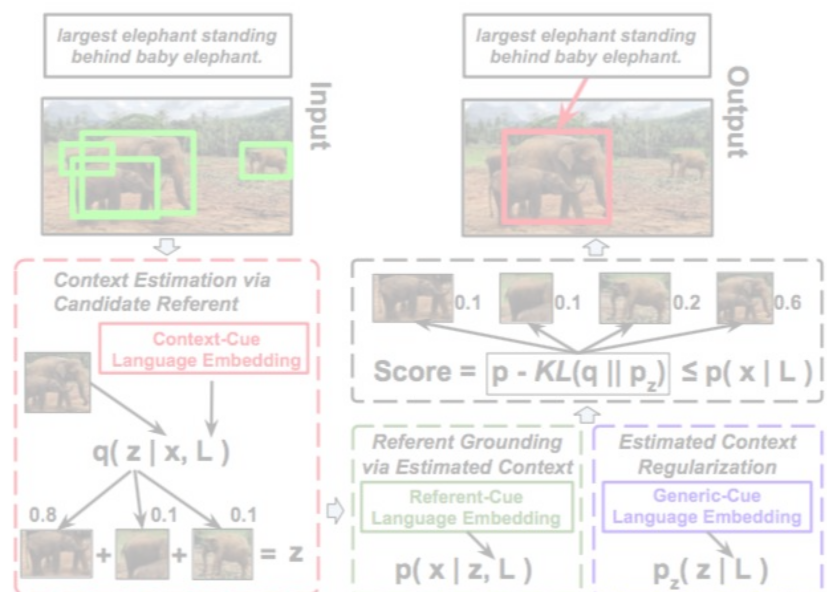
- Scene Dynamics
- Design-free NMN for VQA

Three Examples

Visual Relation Detection [CVPR'17, ICCV'17] **Referring Expression Grounding** [CVPR'







Compositionality **Learning to Reason**

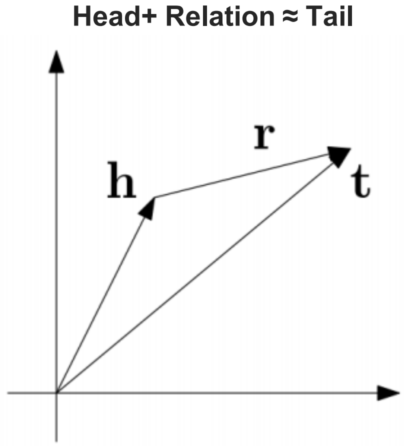
NANYANG TECHNOLOGICAL UNIVERSITY

Challenges in Visual Relation Detection

- Modeling <Subject, Predicate, Object>
  - Joint Model: direct triplet modeling
    - Complexity $O(N^2R)$→hard to scale up
  - Separate Model: separate objects & predicate
    - Complexity $O(N+R)$→visual diversity

TransE: Translation Embedding

[Bordes et al. NIPS'13]



WALL-E

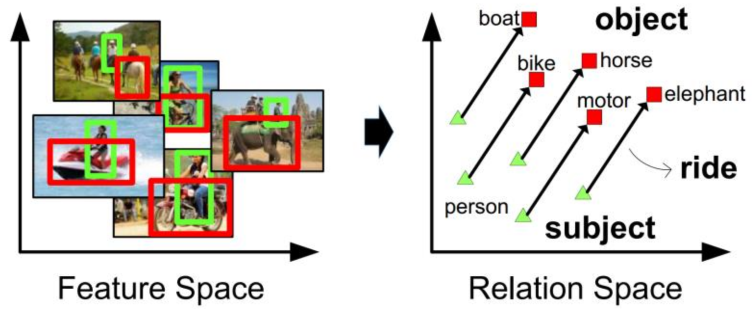
_has_genre

Animation
Computer
Anim.
Comedy film
Adventure film
Science Fiction
Fantasy
Stop motion
Satire
Drama
Connecting

Visual Translation Embedding

[Zhang et al. CVPR'17, ICCV'17]

- VTransE: Visual extension of TransE

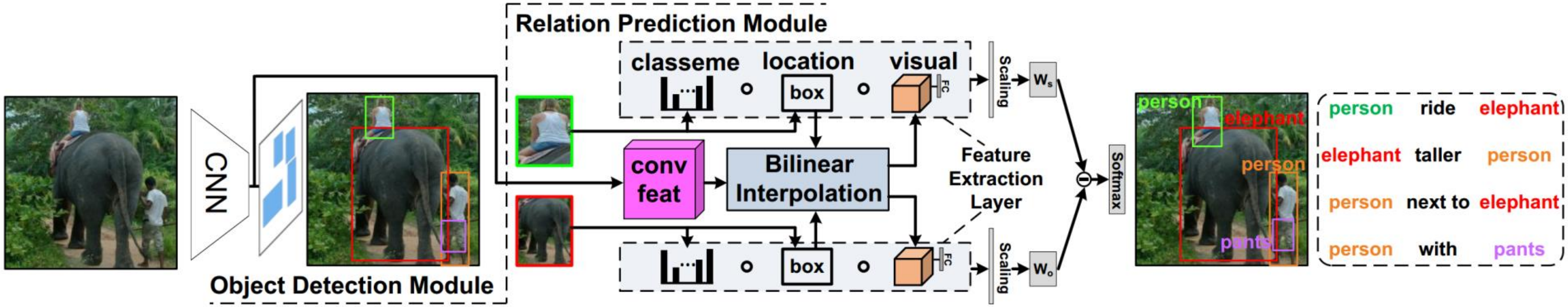


$$\mathbf{W}_s \mathbf{x}_s + \mathbf{t}_p \approx \mathbf{W}_o \mathbf{x}_o$$

$$\mathcal{L}_{rel} = \sum_{(s,p,o)\in\mathcal{R}} \sum_{(s',p,o')\in\mathcal{R}'} [d(\mathbf{W}_s\mathbf{x}_s + \mathbf{t}_p, \mathbf{W}_\mathbf{o}\mathbf{x}_\mathbf{o}) + 1 - d(\mathbf{W}_s\mathbf{x}'_s + \mathbf{t}_p, \mathbf{W}_\mathbf{o}\mathbf{x}'_\mathbf{o})]_+$$

$$\mathcal{L}_{rel} = \sum_{(s,p,o)\in\mathcal{R}} -\log\mathrm{softmax}\left(\mathbf{t}_p^T(\mathbf{W}_o\mathbf{x}_o - \mathbf{W}_s\mathbf{x}_s)\right)$$

VTransE Network

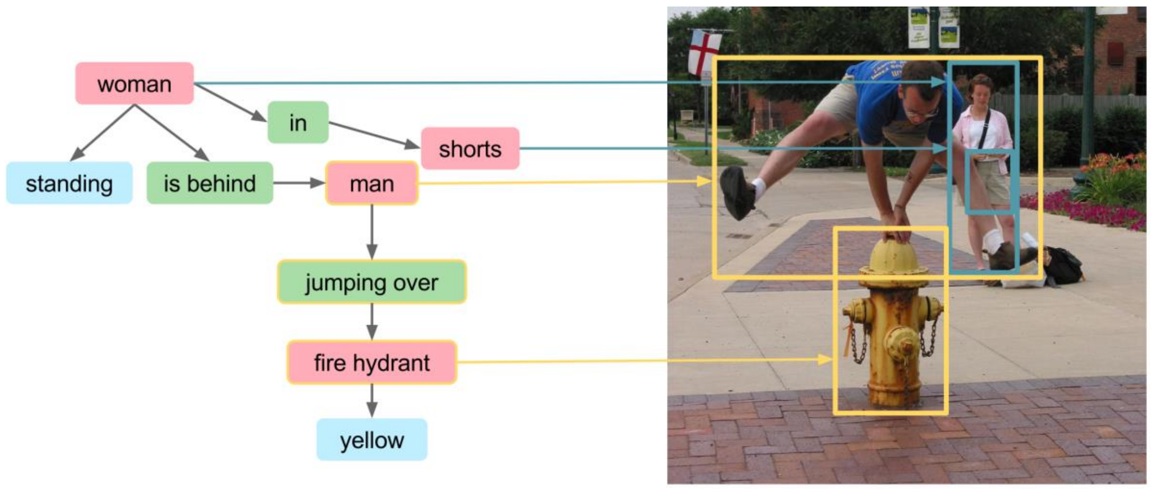


NANYANG TECHNOLOGICAL UNIVERSITY

Evaluation: Relation Datasets

- Visual Relationship Lu et al. ECCV'16
- Visual Genome Krishna et al. IJCV'16

DataSet	Image	Object	Predicate	Unique Relation	Relation/ Object
VRD	5,000	100	70	6,672	24.25
VG	99,658	200	100	19,237	57



Main Deficiency: Incomplete Annotation

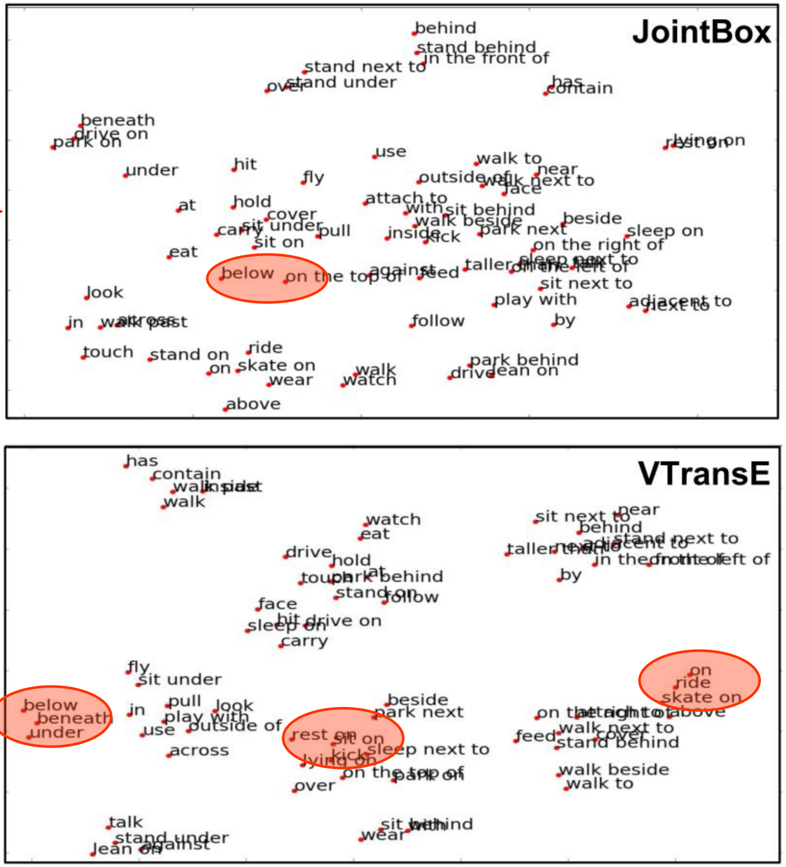
Does TransE work in visual domain?

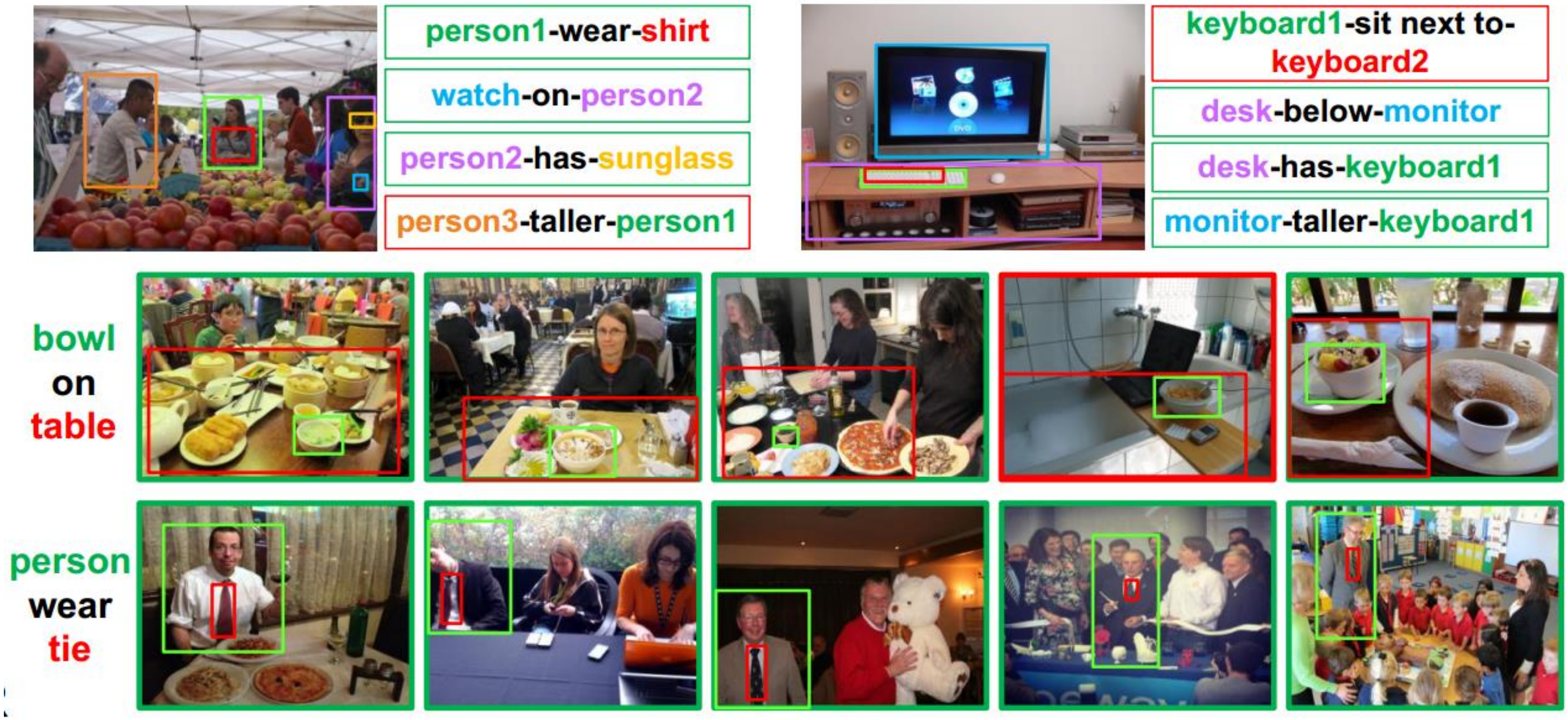
- **Predicate Prediction**

NANYANG TECHNOLOGICAL UNIVERSITY

Does TransE work in visual domain?

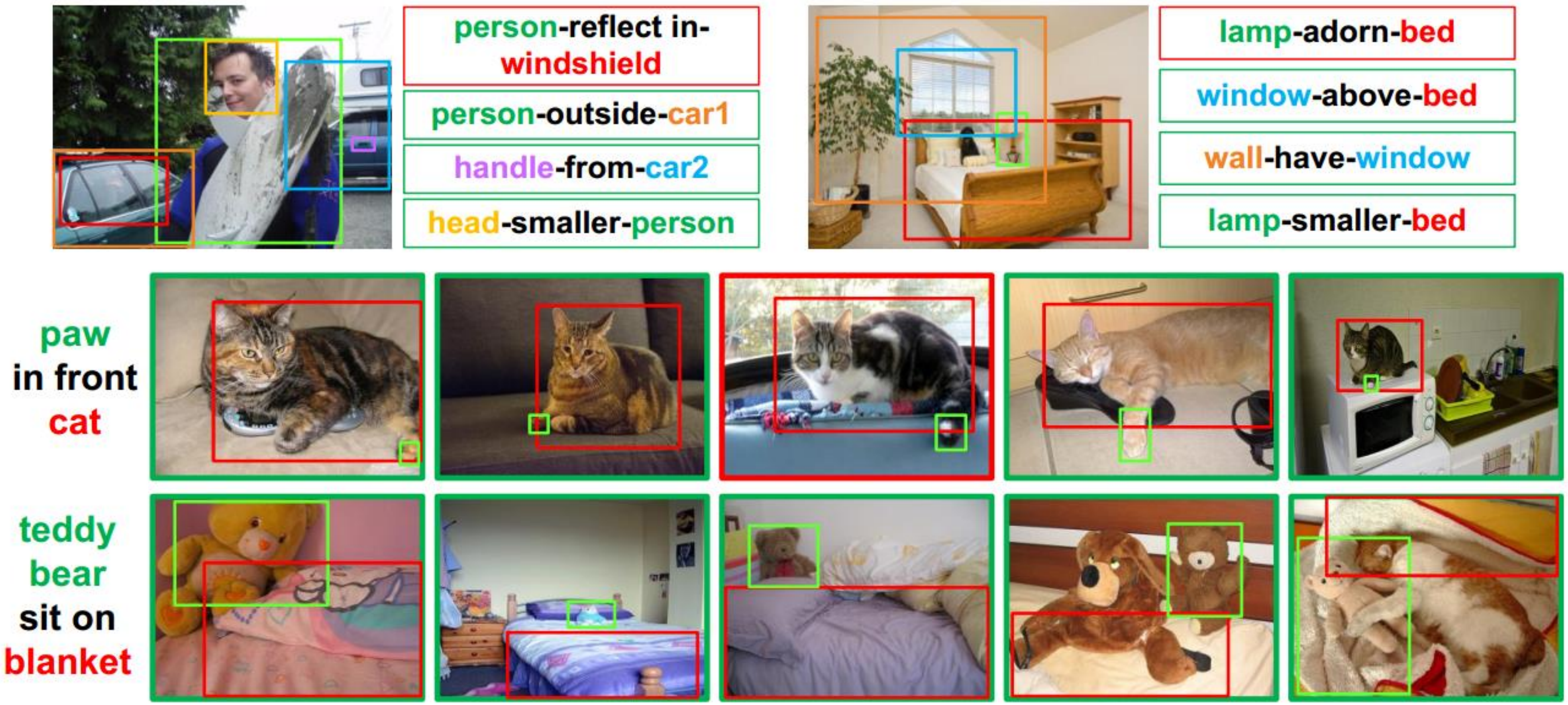
$$\mathcal{L}_{rel} = \sum_{(s,p,o)\in\mathcal{R}} -\log \mathrm{softmax}\left(\boxed{\mathbf{t}_p^T}(\mathbf{W}_o\mathbf{x}_o - \mathbf{W}_s\mathbf{x}_s)\right)$$





Demo link: cvpr.zl.io

NANYANG TECHNOLOGICAL UNIVERSITY



Demo link: cvpr.zl.io

NANYANG TECHNOLOGICAL UNIVERSITY

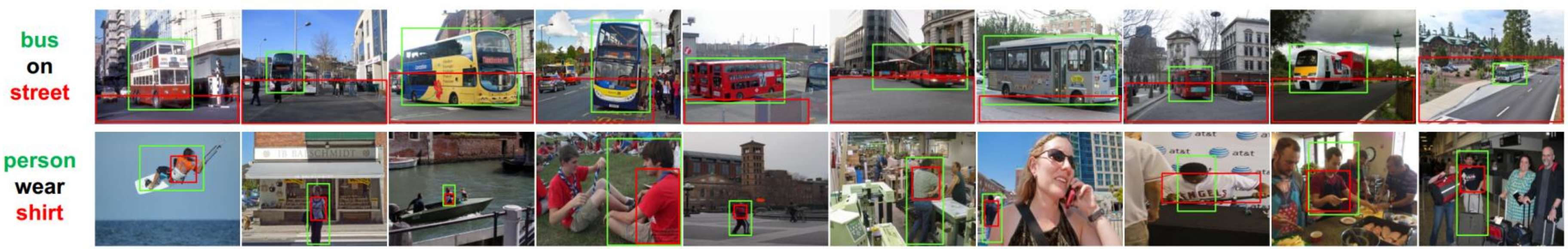
Dataset	VRD [27]						VG [23]					
Task	Phrase Det.		Relation Det.		Retrieval		Phrase Det.		Relation Det.		Retrieval	
Metric	R@50	R@100	R@50	R@100	Rr@5	Med r	R@50	R@100	R@50	R@100	Rr@5	Med r
VisualPhrase [37]	0.54	0.63	–	–	3.51	204	3.41	4.27	–	–	11.42	18
DenseCap [19]	0.62	0.77	–	–	4.16	199	3.85	5.01	–	–	12.95	13
Lu's-V [27]	2.24	2.61	1.58	1.85	2.82	211	–	–	–	–	–	–
Lu's-VLK [27]	16.17	17.03	13.86	14.70	**8.75**	137	–	–	–	–	–	–
VTransE	**19.42**	**22.42**	**14.07**	**15.20**	7.89	**41**	**9.46**	**10.45**	**5.52**	**6.04**	**14.65**	**7**
VTransE-2stage	18.45	21.29	13.30	14.64	7.14	41	8.73	10.11	4.97	5.48	12.82	12
Random	0.06	0.11	7.14×10^{-3}	1.43×10^{-2}	2.95	497	0.04	0.07	1.25×10^{-3}	2.50×10^{-3}	3.45	1.28×10^{4}

Phrase Detection: only need to detect the <subject, object> joint box
Relation Detection: detect both subject and object
Retrieval: given a query relation, return images

VTransE were best separate models in 2017. ([Li et al. and Dai et al. CVPR'17 are (partially joint models)

New state-of-the-art: Neural MOTIF (Zellers et al. CVPR'18, 27.2/30.3 R@50/R@100)

Bad retrieval on VR is due to incomplete annotation



NANYANG TECHNOLOGICAL UNIVERSITY

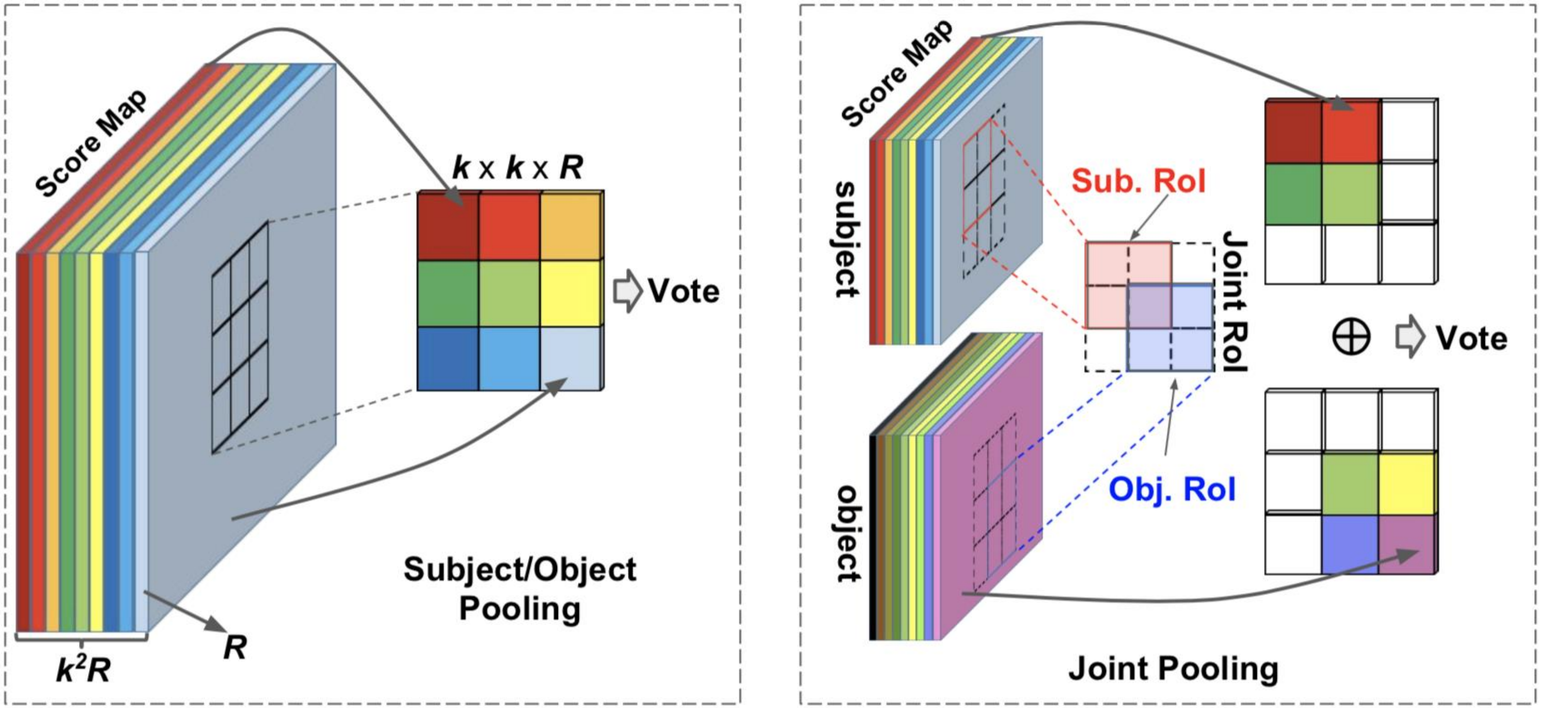
Two follow-up works

- The key: pure visual pair model $f(x_1, x_2)$
- f(x1,x2) underpins almost every VRD
- Evaluation: predicate classification

- 1. Faster pairwise modeling (ICCV'17)
- 2. Object-agnostic modeling (ECCV'18 submission)

NANYANG TECHNOLOGICAL UNIVERSITY

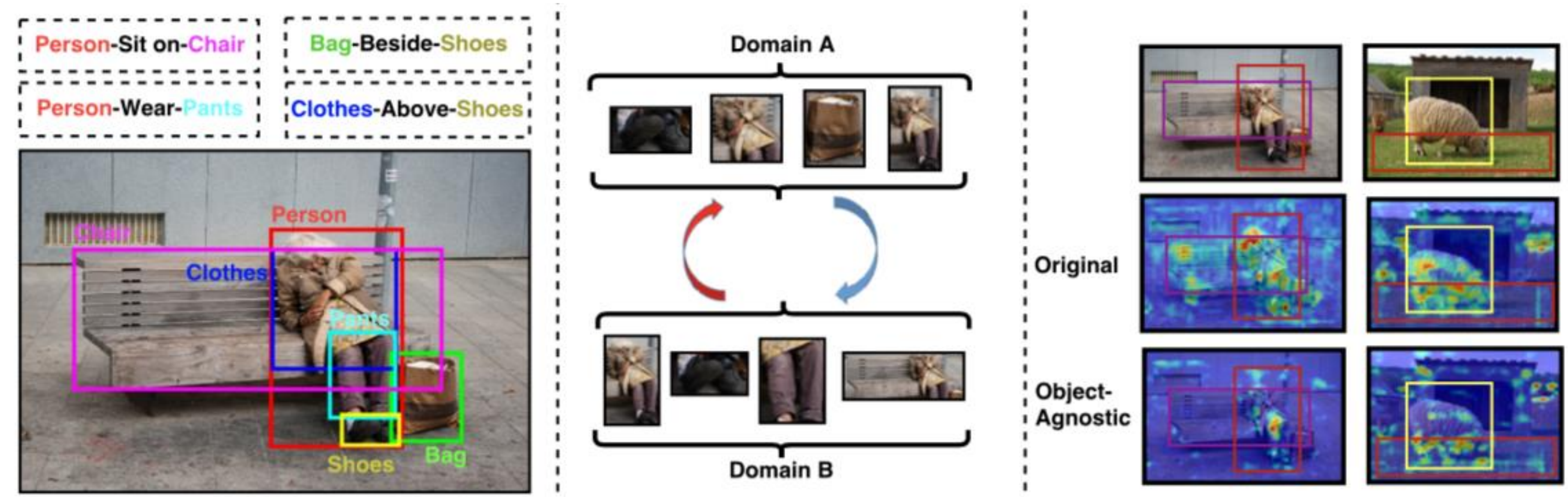
Parallel Pairwise R-FCN (Zhang et al. ICCV'17)



	VRD R@50	VRD R@100	VG R@50	VG R@100
VTransE	44.76	44.76	62.63	62.87
PPR-FCN	**47.43**	**47.43**	**64.17**	**64.86**

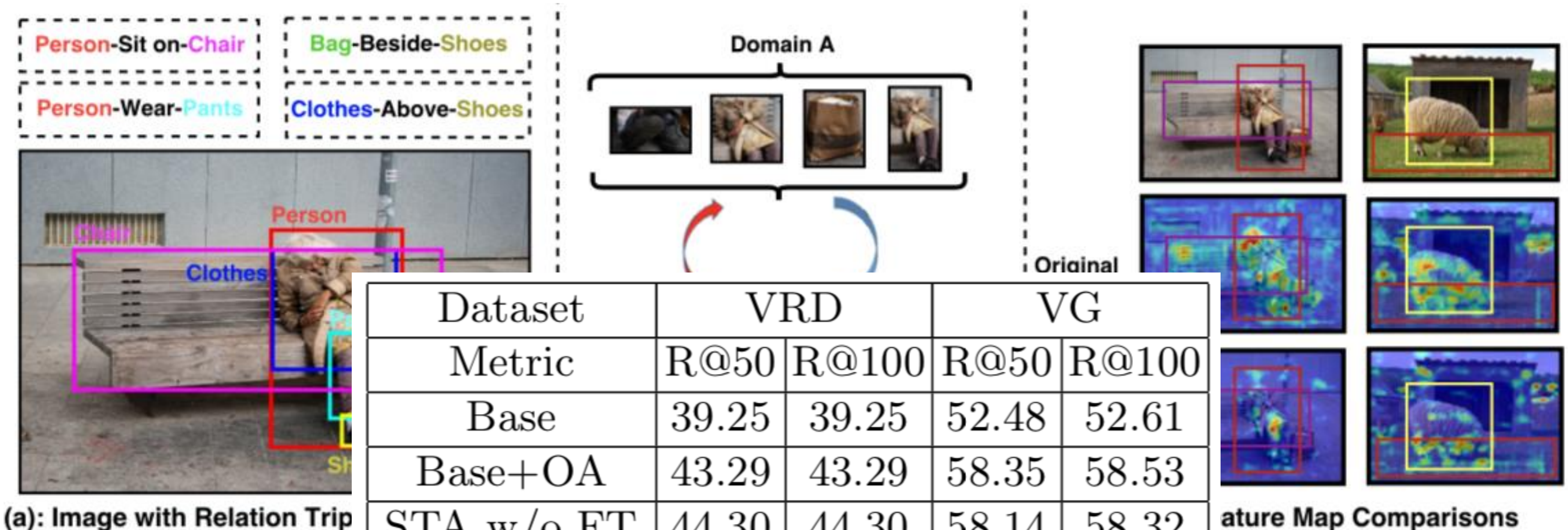
NANYANG TECHNOLOGICAL UNIVERSITY

Shuffle-Then-Assemble (Yang et al. 18')



(a): Image with Relation Triplet Annotations (b): Shuffle-Then-Assemble (c): Feature Map Comparisons

Shuffle-Then-Assemble (Yang et al. 18')

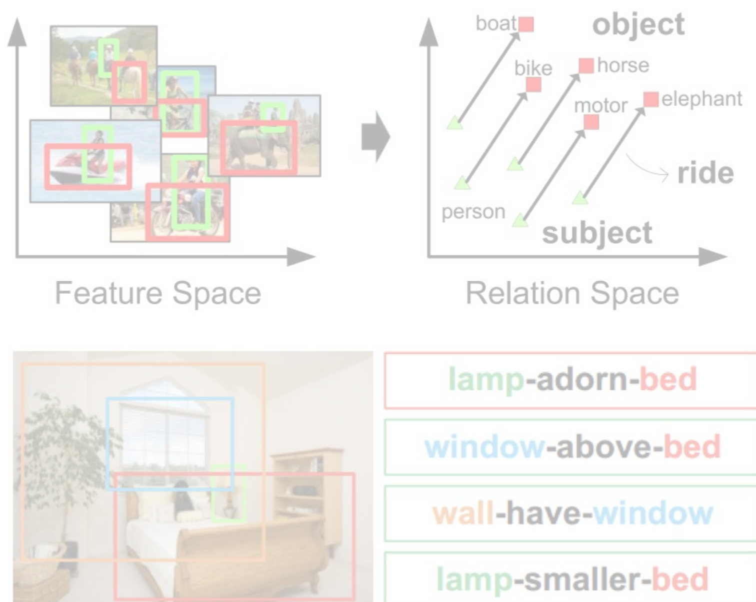


Dataset	VRD		VG	
Metric	R@50	R@100	R@50	R@100
Base	39.25	39.25	52.48	52.61
Base+OA	43.29	43.29	58.35	58.53
STA w/o FT	44.30	44.30	58.14	58.32
STA w/o Res	46.83	46.83	62.08	62.32
STA	**48.03**	**48.03**	**62.71**	**62.94**
Lu's-V [26]	7.11	7.11	–	–
Lu's-VLK [26]	47.87	47.87	–	–
VTransE [48]	44.76	44.76	62.63	62.87
Peyre's-A[31]	46.30	46.30	–	–

NANYANG TECHNOLOGICAL UNIVERSITY

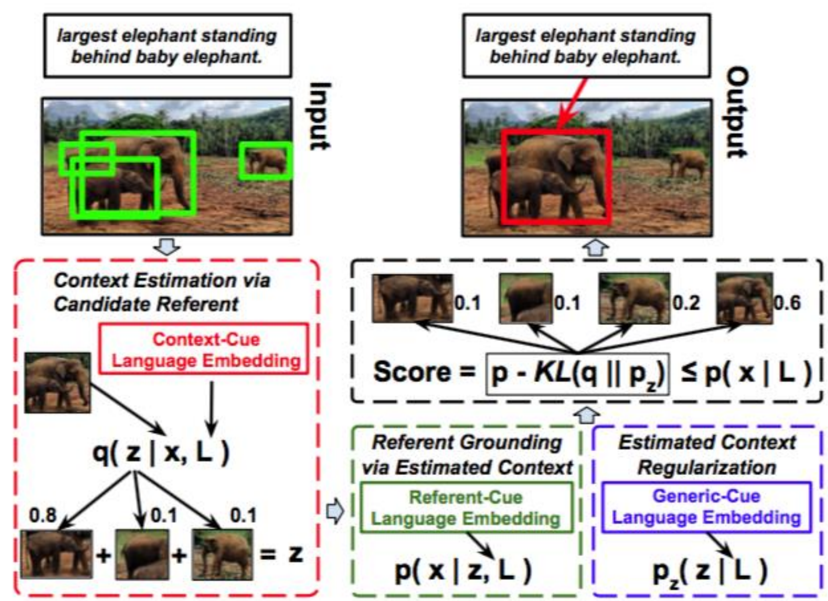
Three Examples

Visual Relation Detection [CVPR'17, ICCV'17]



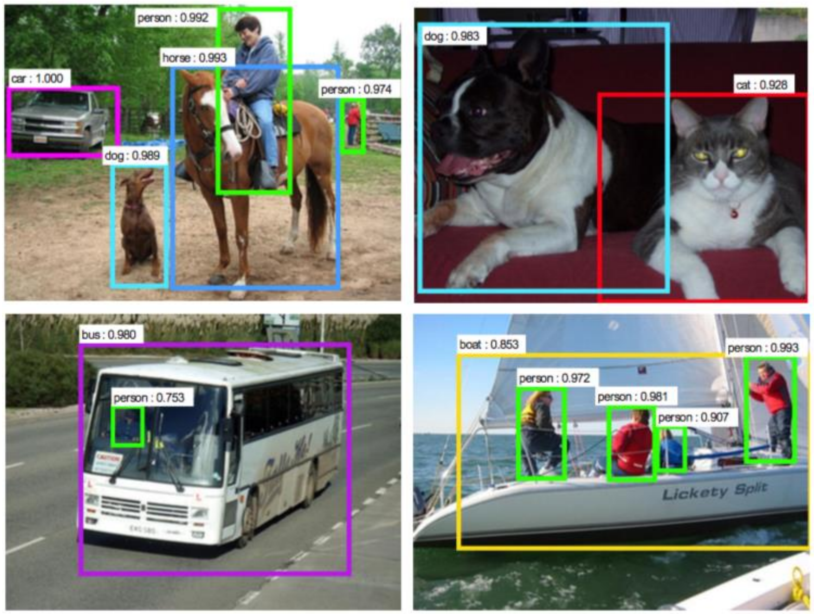
Compositionality

Referring Expression Grounding [CVPR'1



Learning to Reason

What is grounding? Object Detection



R Girshick ICCV'15

Link words (from a fixed vocab.) to visual objects

O(N)

What is grounding? Phrase-to-Region

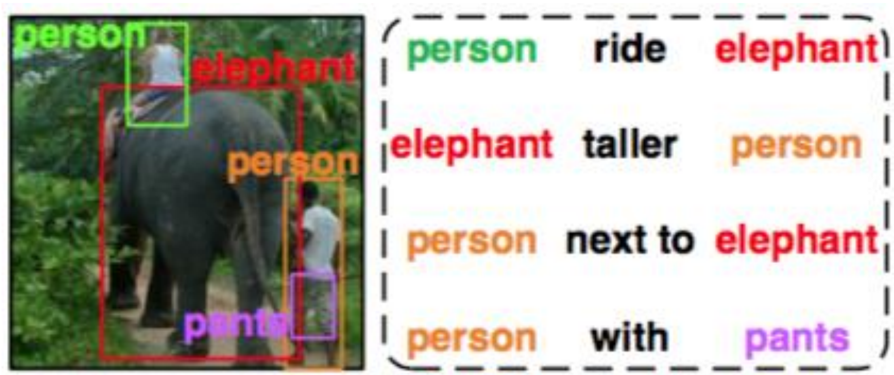
A man in a gray sweater speaks to two women and a man pushing a shopping cart through Walmart.

Plummer et al. ICCV'15

Link phrases to visual objects

O(N)

What is grounding? Visual Relation Detection

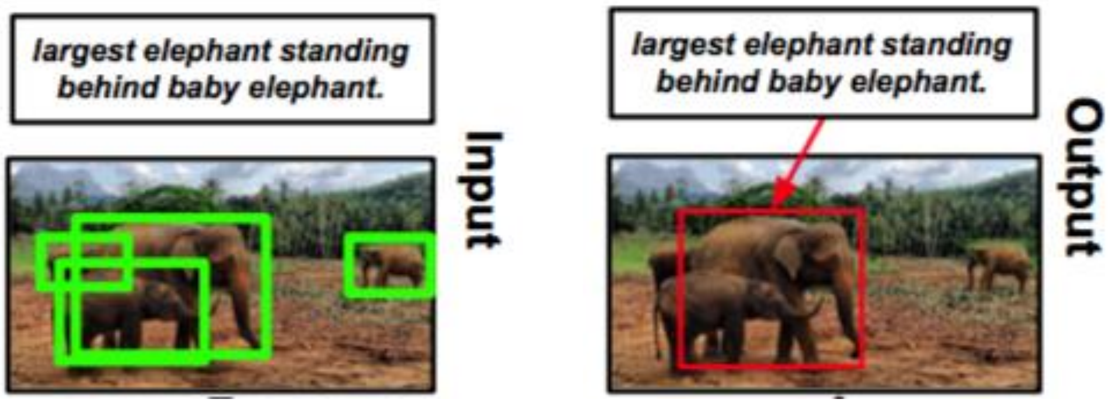


$O(N^2)$

Zhang et al. CVPR'17

NANYANG TECHNOLOGICAL UNIVERSITY

What's referring expression grounding?



$O(2^N)$

Prior Work: Multiple Instance Learning

$$x^* = \arg\max_{x \in \mathcal{X}} \log \sum_z p(x, z|L)$$

Max-Pool [Hu et al. CVPR'17]

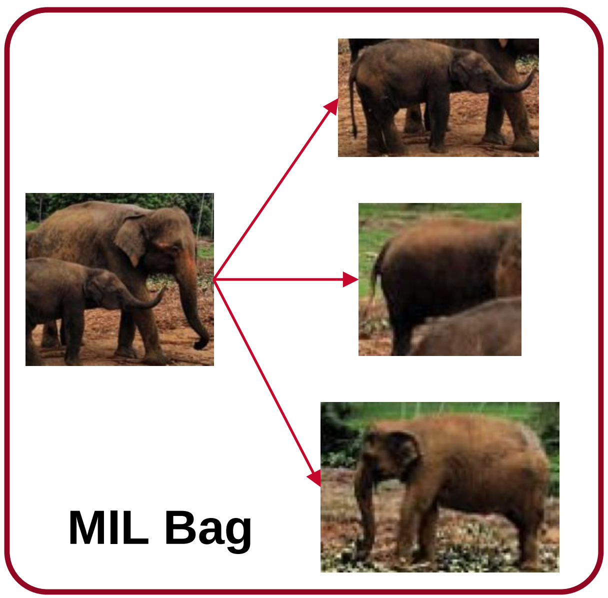
$$\log \max_z p(x, z)$$

Noisy-Or [Nagaraja et al. ECCV'16]

$$\log(1 - \prod_z (1 - p(x, z)))$$

$$\sum_z \log p(x, z)$$

$O(2^N)$ → $O(N^2)$



Bad Approximation:

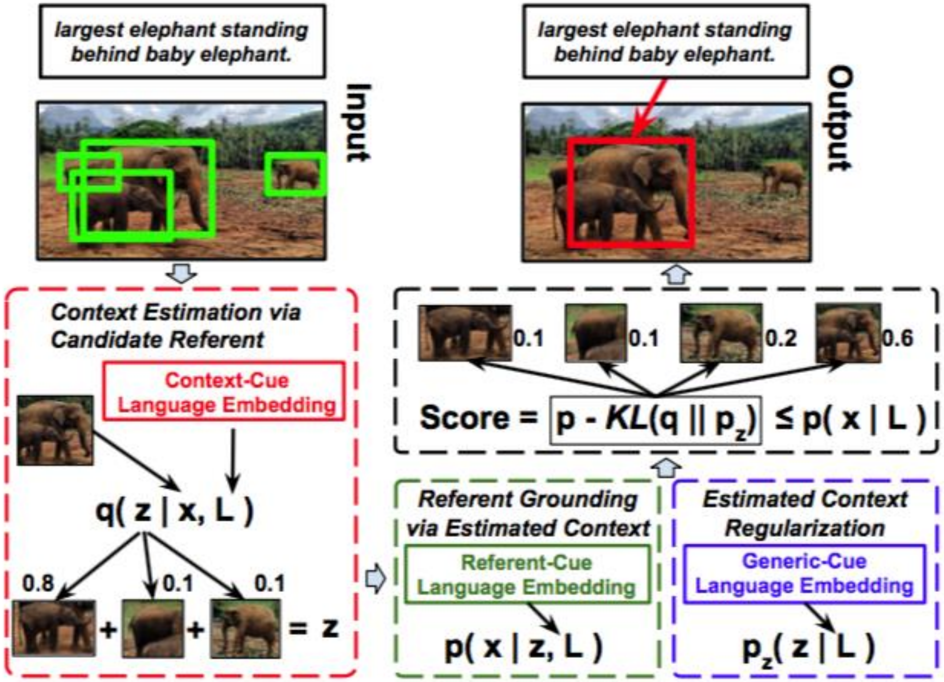
1. Context z is not necessarily to be a single region
2. Log-sum directly to sum-log is too coarse, i.e., forcing every pair to be equally possible

NANYANG TECHNOLOGICAL UNIVERSITY

Our Work: Variational Context [Zhang et al CVPR'18]

$$\log p(x|L) = \log \sum_z p(x, z|L) \geq \boxed{\mathcal{Q}(x, L)} =$$
$$\underbrace{\mathbb{E}_{z\sim q_\phi(z|x,L)} \log p_\theta(x|z, L)}_{\text{Localization}} - \underbrace{KL\left(q_\phi(z|x, L)||p_\omega(z|L)\right)}_{\text{Regularization}}.$$

Variational lower-bound:
Sum-log



SGD Details

$$\mathbb{E}_{z\sim q_\phi(z|x,L)} \log p_\theta(x|z,L) - KL\left(q_\phi(z|x,L)||p_\omega(z|L)\right)$$

z: reasoning over 2^N

Deterministic function
(Soft attention)

REINFORCE with baseline
(MC, hard sampling)

$$z = f(x,L) = \sum_{x'\in\mathcal{X}} x' \cdot q_\phi(x'|x,L)$$

$$z = f(x,L;\epsilon), \epsilon \sim p(\epsilon)$$

$$\mathcal{Q}(x,L) = \log p_\theta(x|z,L) - \log q_\phi(z|x,L) + \log p_\omega(z|L)$$

$$\mathcal{Q}(x,L) \propto \mathcal{S}(x,L) = s_\theta(x,L) - s_\phi(x,L) + s_\omega(x,L)$$

NANYANG TECHNOLOGICAL UNIVERSITY

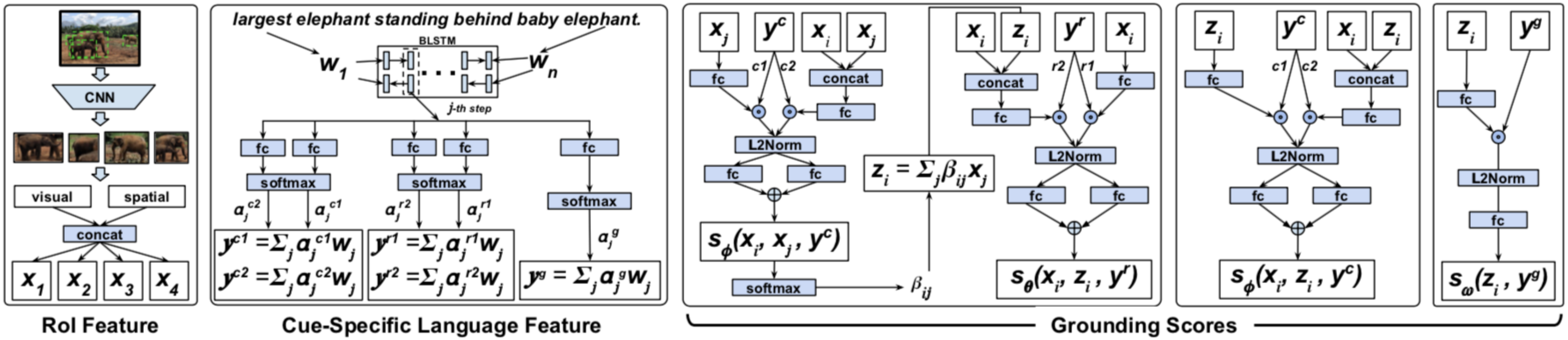
Network Details

$$\mathcal{Q}(x,L) \propto \mathcal{S}(x,L) = s_\theta(x,L) - s_\phi(x,L) + s_\omega(x,L)$$

$$\downarrow$$

$$\mathbf{z}_i = \sum_j \mathrm{softmax}_j \left(s_\phi(\mathbf{x}_i, \mathbf{x}_j, \mathbf{y}^c)\right) \mathbf{x}_j$$

$$s_\theta(x,L) \leftarrow s_\theta(\mathbf{x}_i, \mathbf{z}_i, \mathbf{y}^r) \quad s_\phi(x,L) \leftarrow s_\phi(\mathbf{x}_i, \mathbf{z}_i, \mathbf{y}^c) \quad s_\omega(x,L) \leftarrow s_\omega(\mathbf{z}_i, \mathbf{y}^g)$$

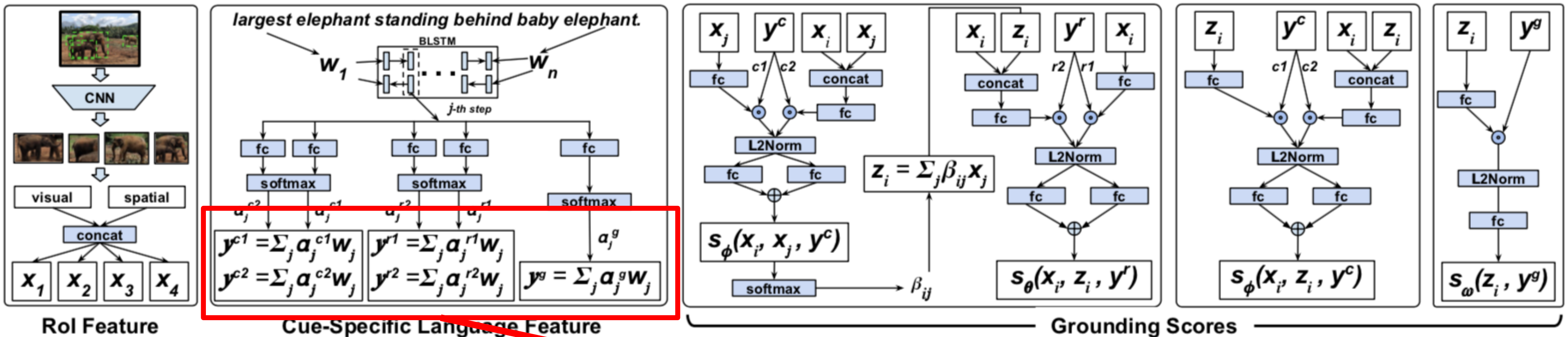


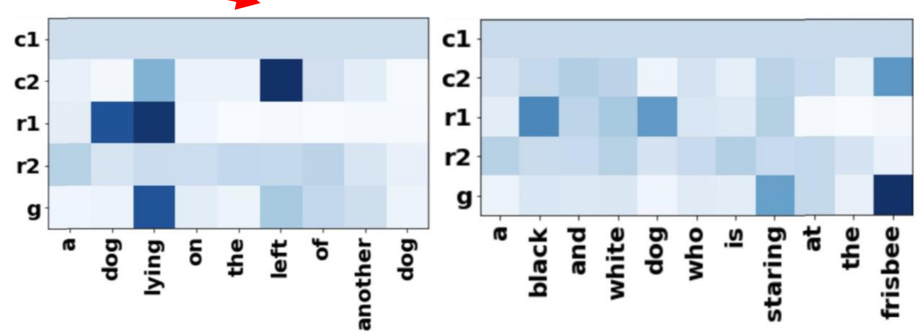
Network Details

$$\mathcal{Q}(x, L) \propto \mathcal{S}(x, L) = s_\theta(x, L) - s_\phi(x, L) + s_\omega(x, L)$$

$$\mathbf{z}_i = \sum_j \mathrm{softmax}_j \left(s_\phi(\mathbf{x}_i, \mathbf{x}_j, \mathbf{y}^c) \right) \mathbf{x}_j$$

$$s_\theta(x, L) \leftarrow s_\theta(\mathbf{x}_i, \mathbf{z}_i, \mathbf{y}^r) \quad s_\phi(x, L) \leftarrow s_\phi(\mathbf{x}_i, \mathbf{z}_i, \mathbf{y}^c) \quad s_\omega(x, L) \leftarrow s_\omega(\mathbf{z}_i, \mathbf{y}^g)$$





Grounding Accuracy

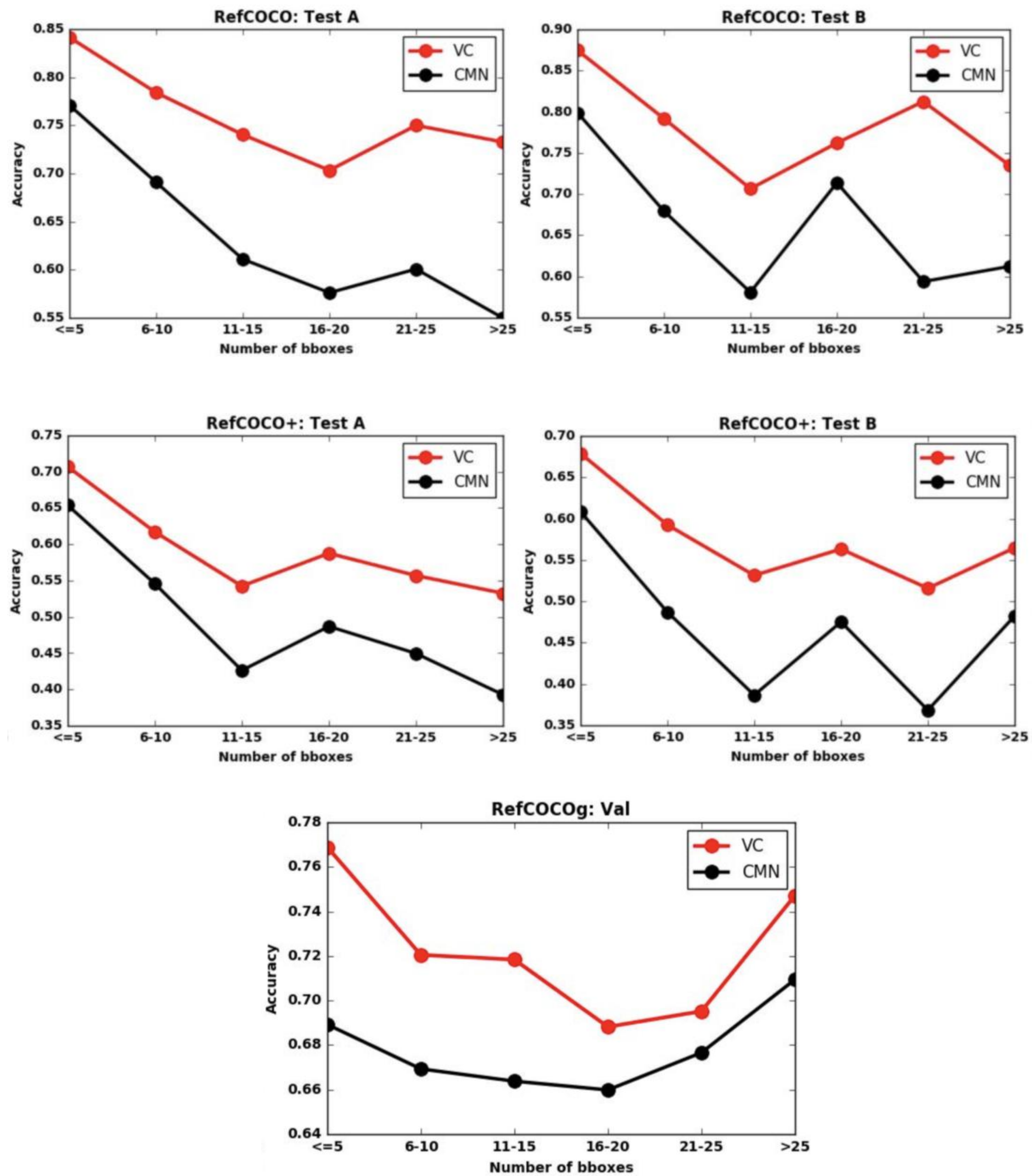
		State-of-The-Arts						Our Baselines		
Dataset	Split	MMI [25]	NegBag [26]	Attr [19]	CMN [11]	Speaker [46]	Listener [46]	VC w/o reg	VC w/o α	VC
RefCOCO	Test A	71.72	75.6	78.85	75.94	78.95	78.45	75.59	74.03	**78.98**
	Test B	71.09	78.0	78.07	79.57	80.22	80.10	79.69	78.27	**82.39**
RefCOCO+	Test A	58.42	—	61.47	59.29	**64.60**	63.34	60.76	57.61	62.56
	Test B	51.23	—	57.22	59.34	59.62	58.91	60.14	54.37	**62.90**
RefCOCOg	Val	62.14	68.4	69.83	69.30	72.63	72.25	71.05	65.13	**73.98**
RefCOCO(det)	Test A	64.90	58.6	72.08	71.03	72.95	72.95	70.78	70.73	**73.33**
	Test B	54.51	56.4	57.29	65.77	63.43	62.98	65.10	64.63	**67.44**
RefCOCO+(det)	Test A	54.03	—	57.97	54.32	**60.43**	59.61	56.82	53.33	58.40
	Test B	42.81	—	46.20	47.76	48.74	48.44	51.30	46.88	**53.18**
RefCOCOg(det)	Val	45.85	39.5	52.35	57.47	59.51	58.32	60.95	55.72	**62.30**

The best VGG SINGLE model to date.

Best ResNet Model: Licheng Yu et al. MAttNet: Modular Attention Network for Referring Expression Comprehension. CVPR'18

NANYANG TECHNOLOGICAL UNIVERSITY

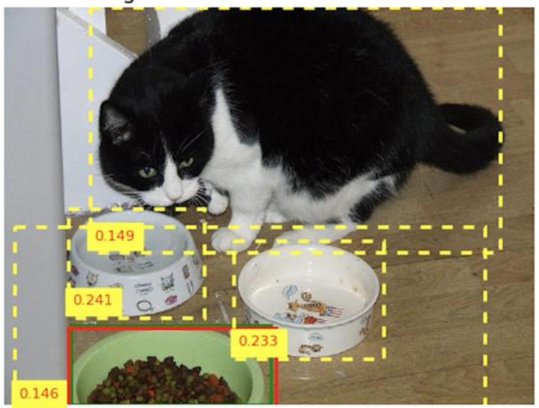
More effective than MIL



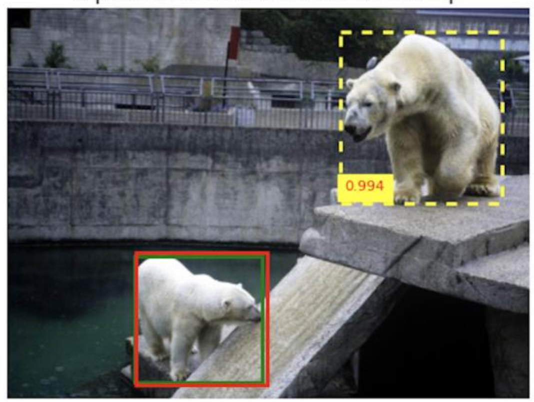
R. Hu et al. Modeling relationships in referential expressions with compositional mod- ular networks. In *CVPR*, 2017

Qualitative Results

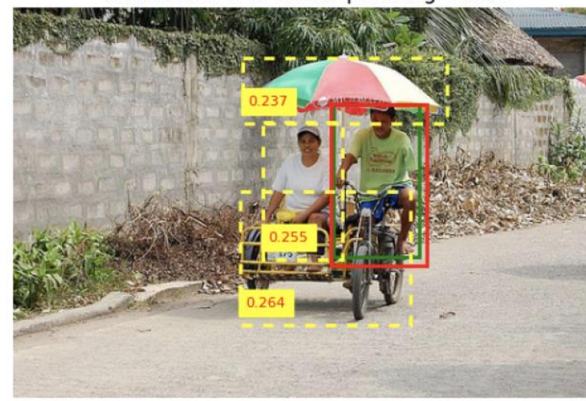
a green food dish with cat food



a polar bear at the bottom of a ramp

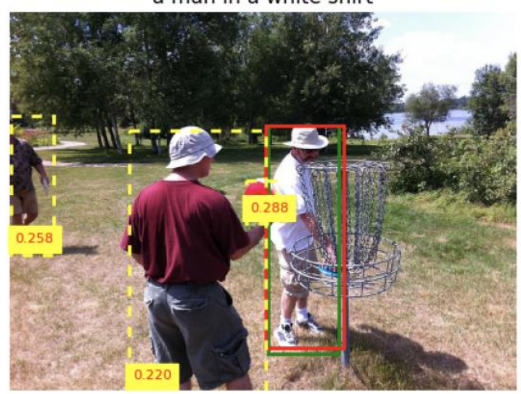


a man with one arm is pedaling a bike



a green boat afloat near a flooded river lined with refugees

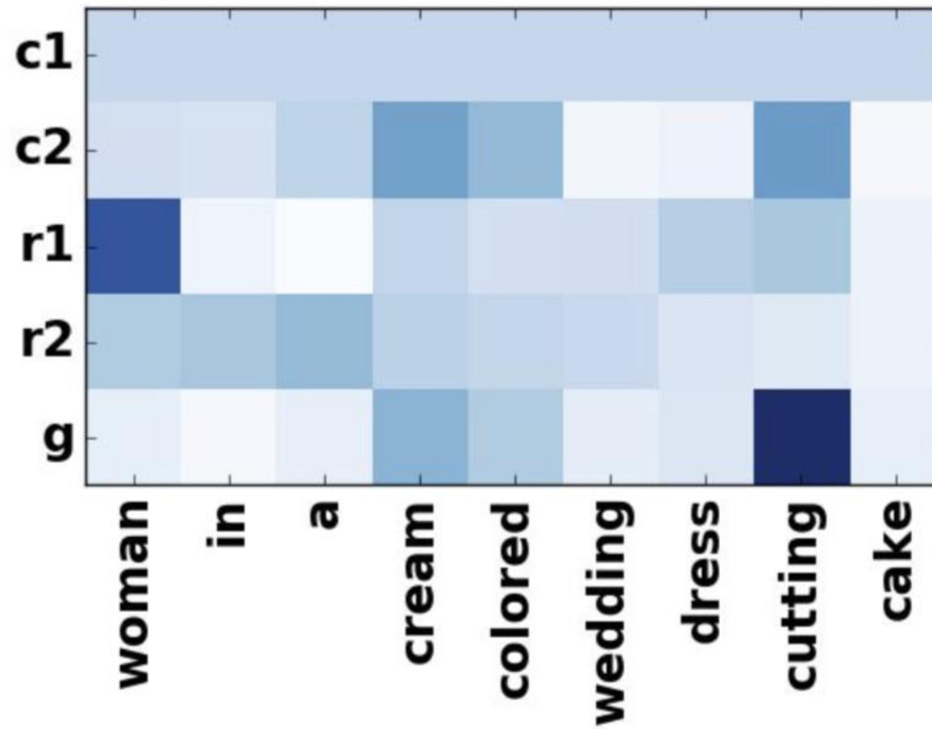
a man in a white shirt



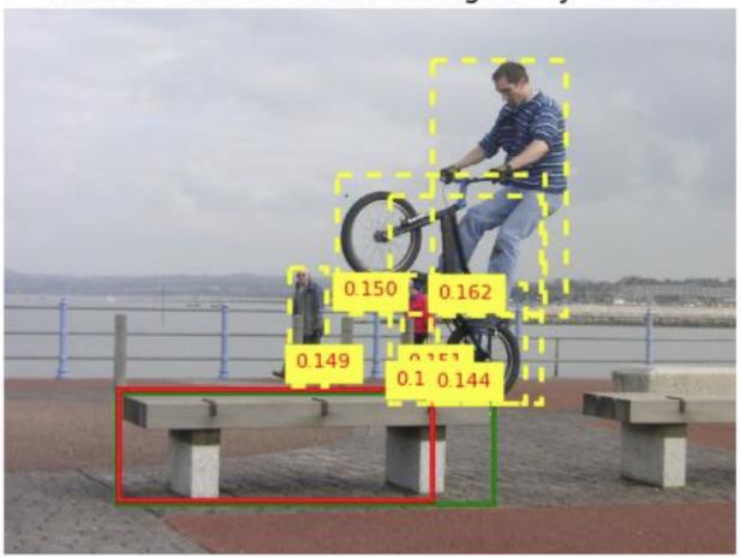
A dark horse between three lighter horses

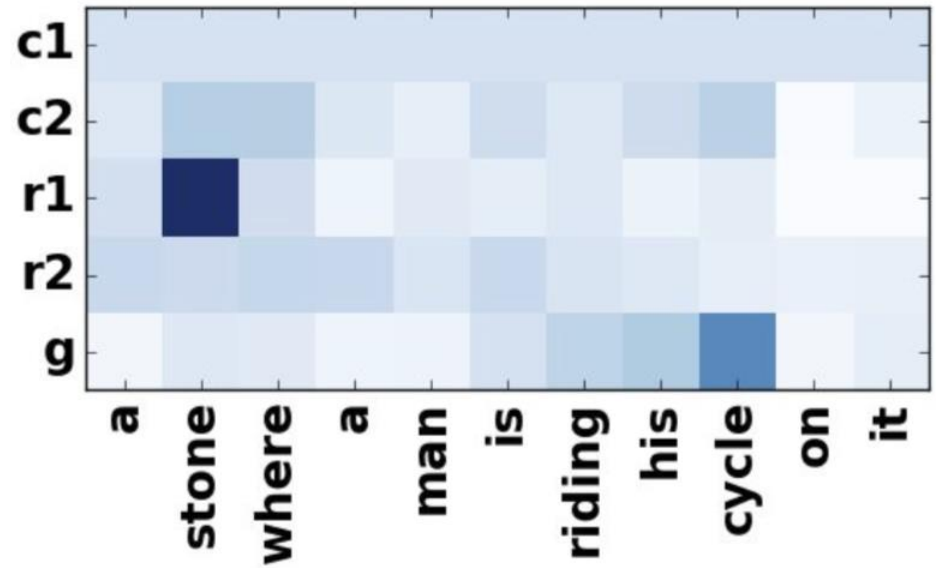
NANYANG TECHNOLOGICAL UNIVERSITY

woman in a cream colored wedding dress cutting cake



a stone where a man is riding his cycle on it

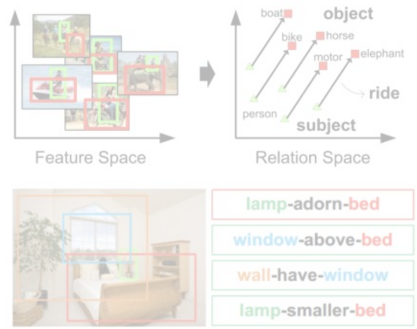




NANYANG TECHNOLOGICAL UNIVERSITY

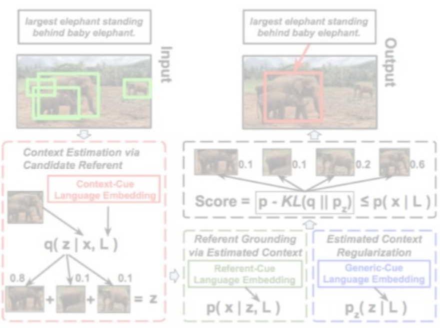
Three Examples

Visual Relation Detection [CVPR'17, ICCV'17]



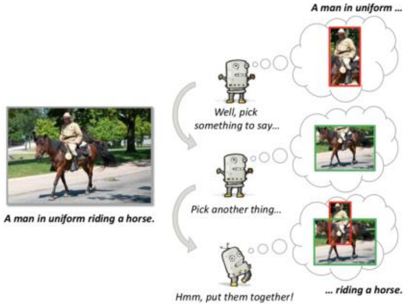
Compositionality

Referring Expression Grounding [CVPR'18]



Learning to Reason

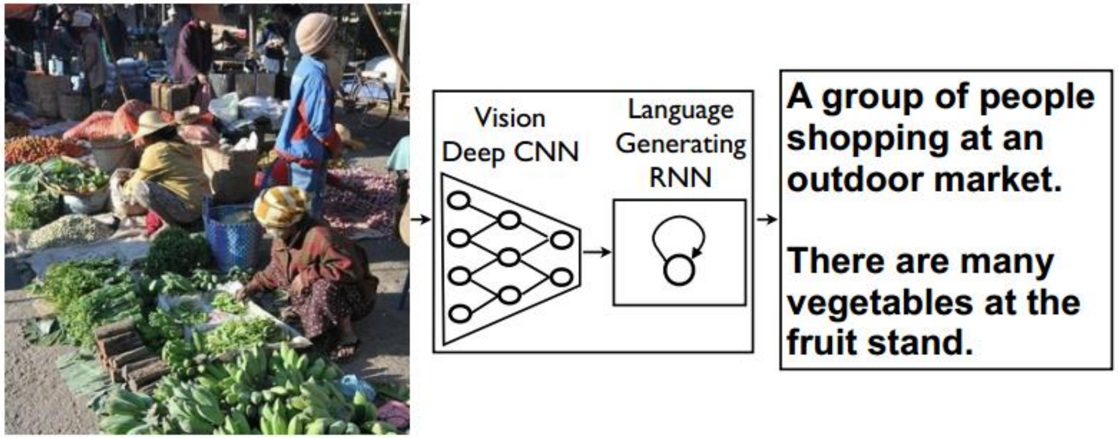
Sequence-level Image Captioning [MM'18 submission]



Learning to Reason

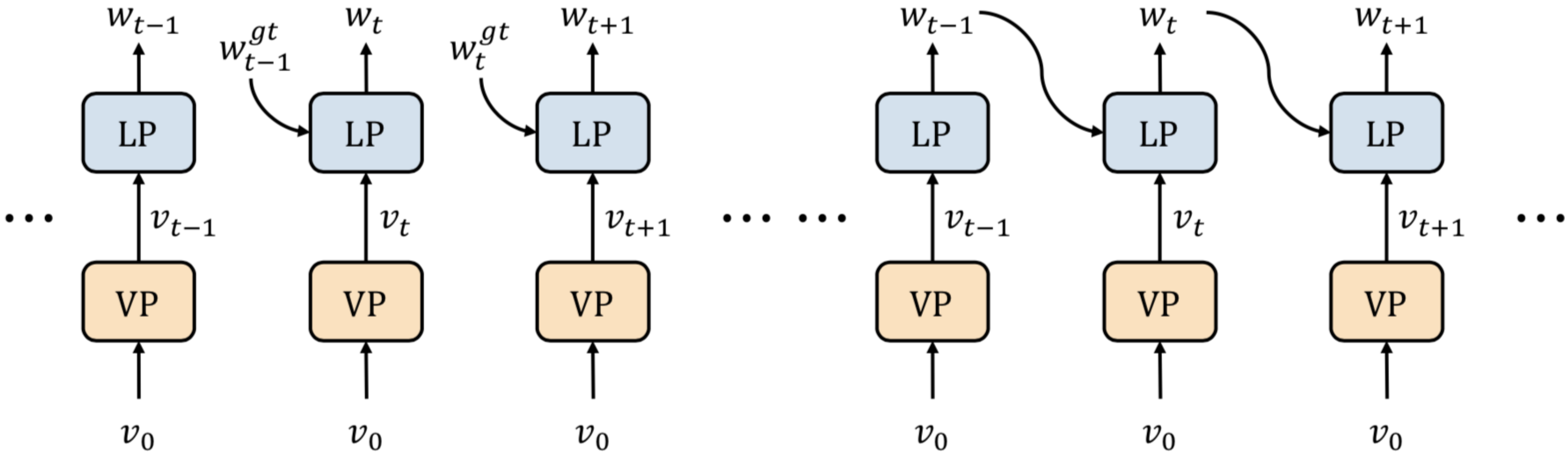
Neural Image Captioning

GoogleNIC (Vinyals et al. 2014)



Encoder (Image→CNN→Vector) → ***Decoder*** (Vector→Word Seq.)

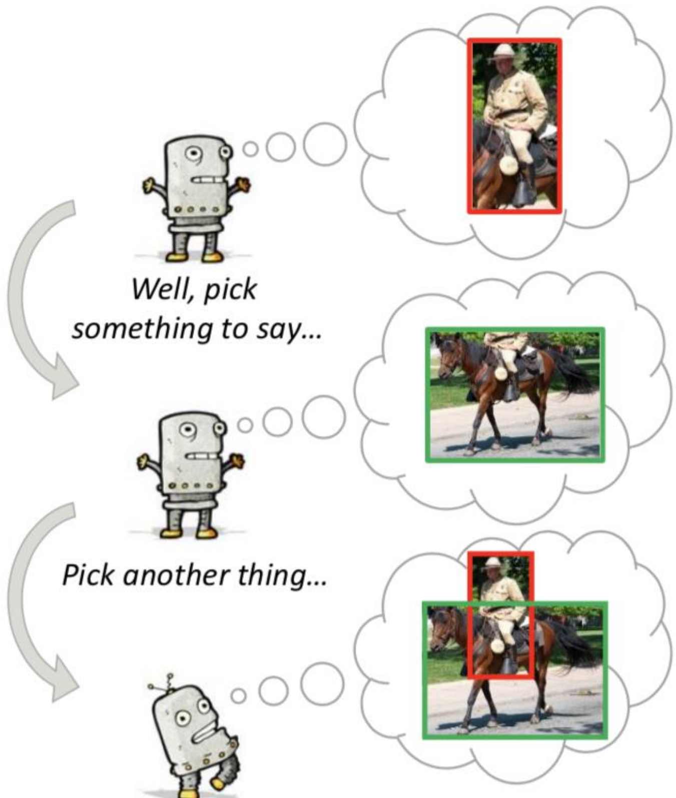
Sequence-level Image Captioning



(a) Traditional Framework

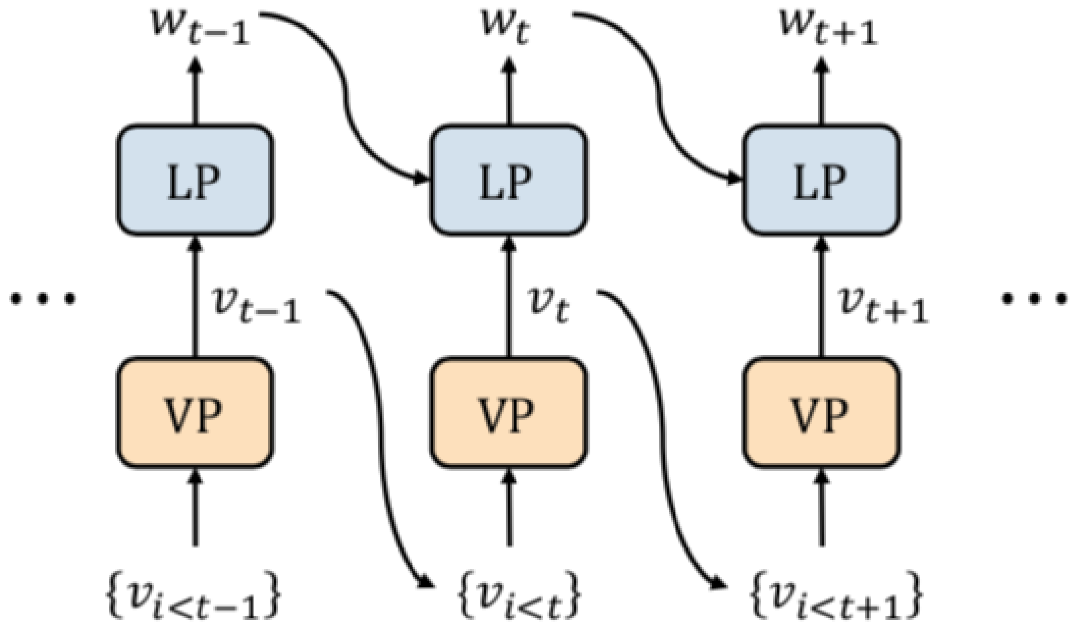
(b) RL-based Framework

Context in Image Captioning



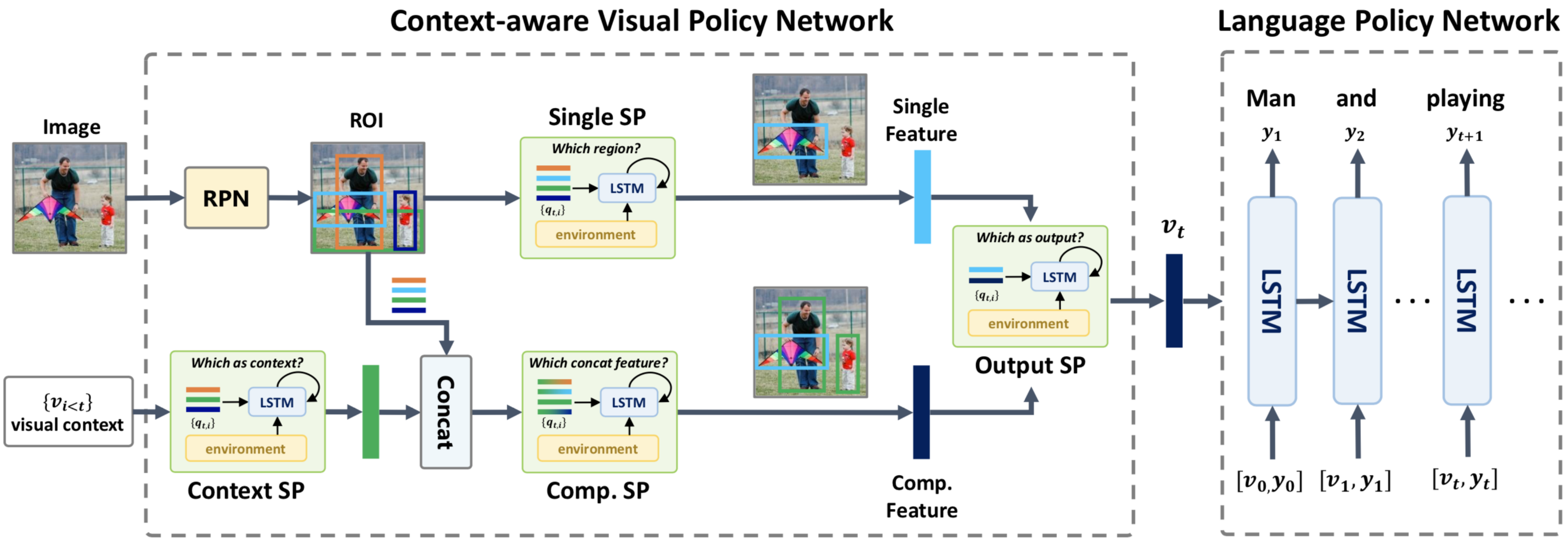
A man in uniform riding a horse.

Context-Aware Visual Policy Network



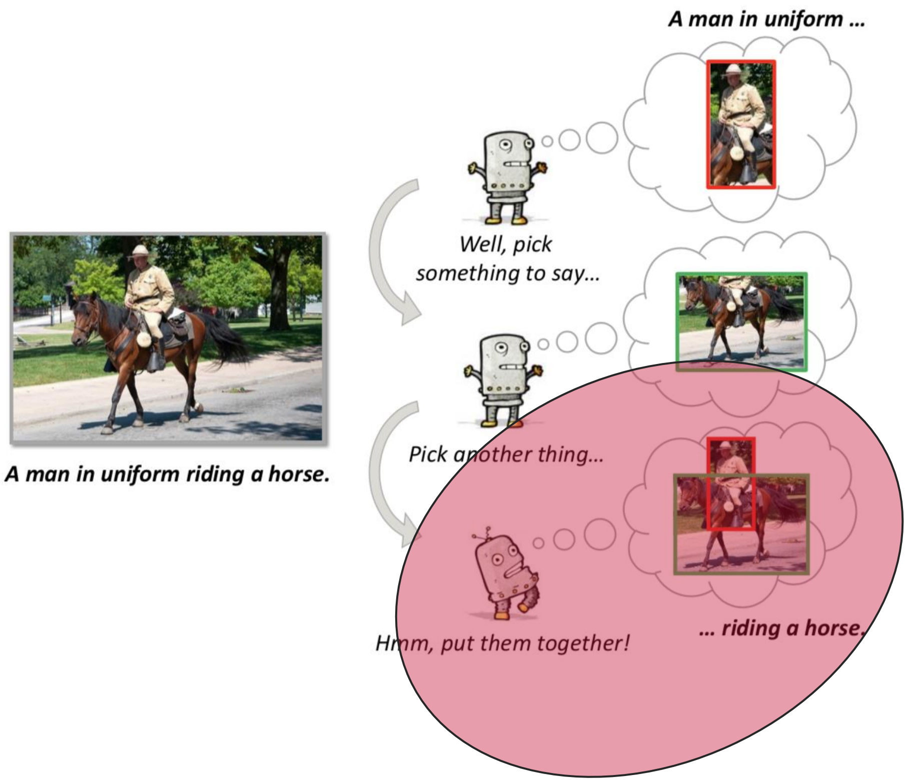
(c) Our Framework

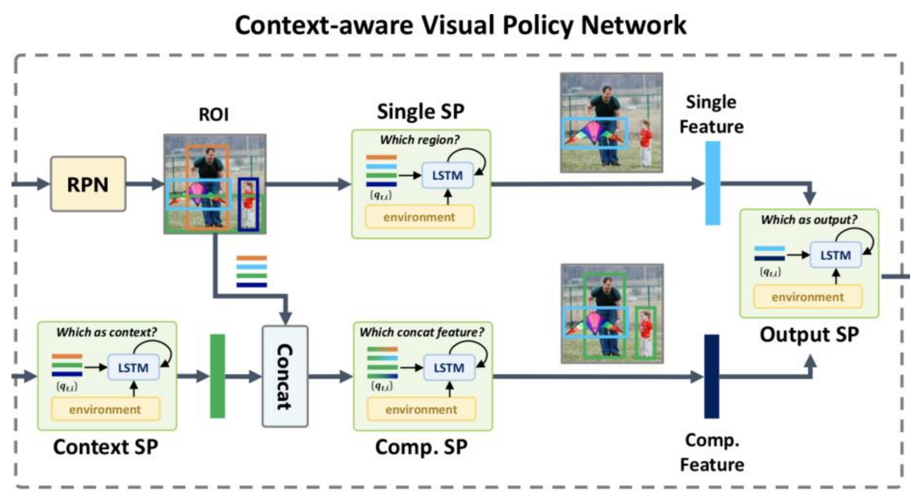
Context-Aware Policy Network

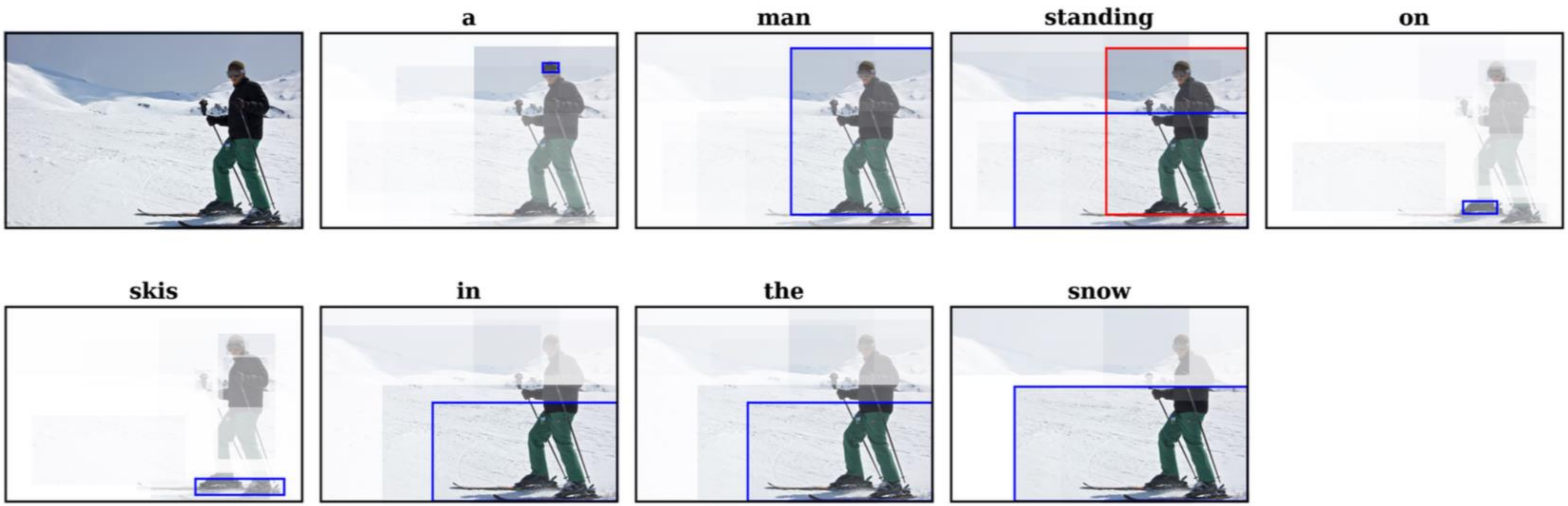


NANYANG TECHNOLOGICAL UNIVERSITY

Context-Aware Policy Network

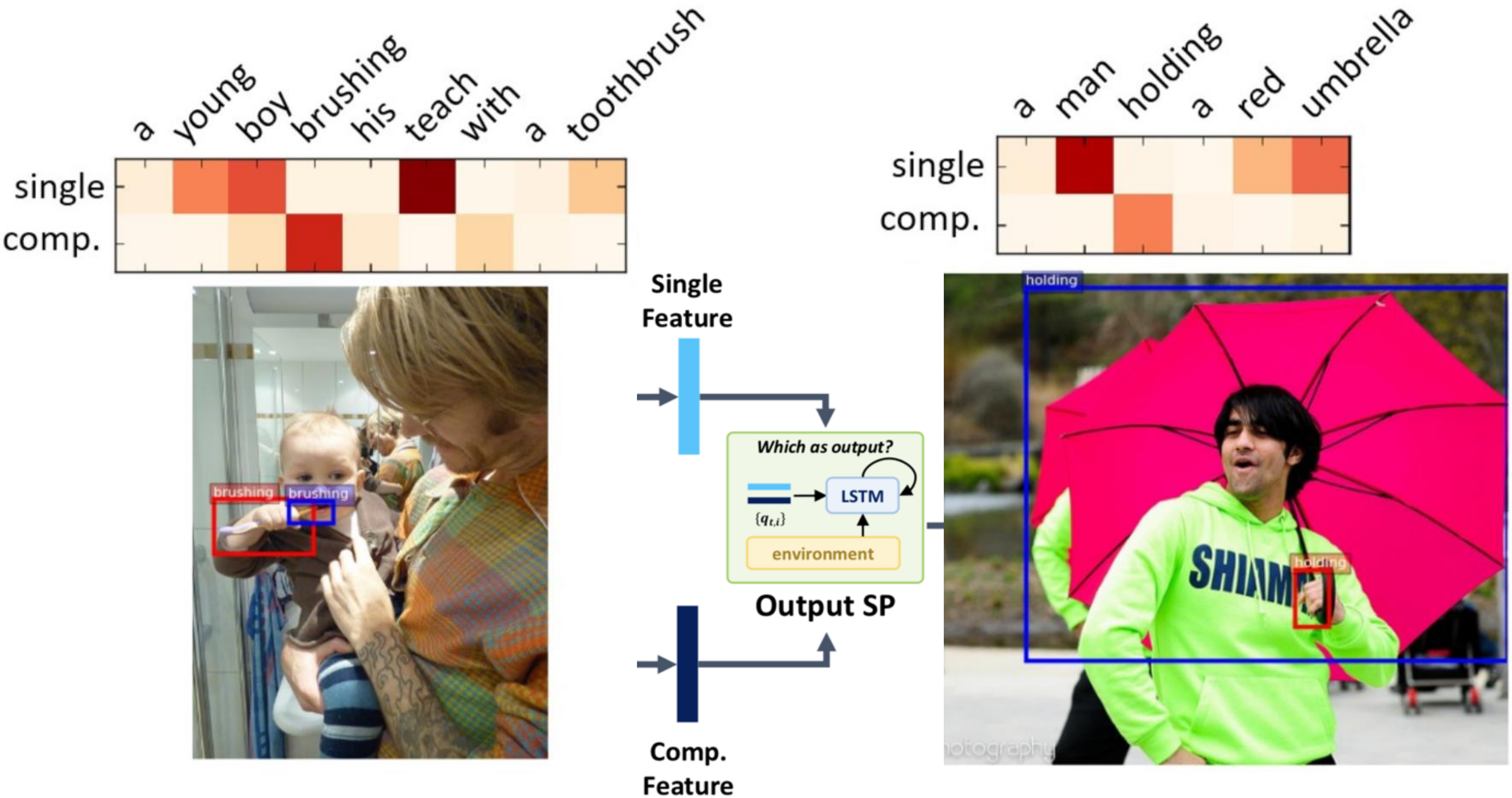




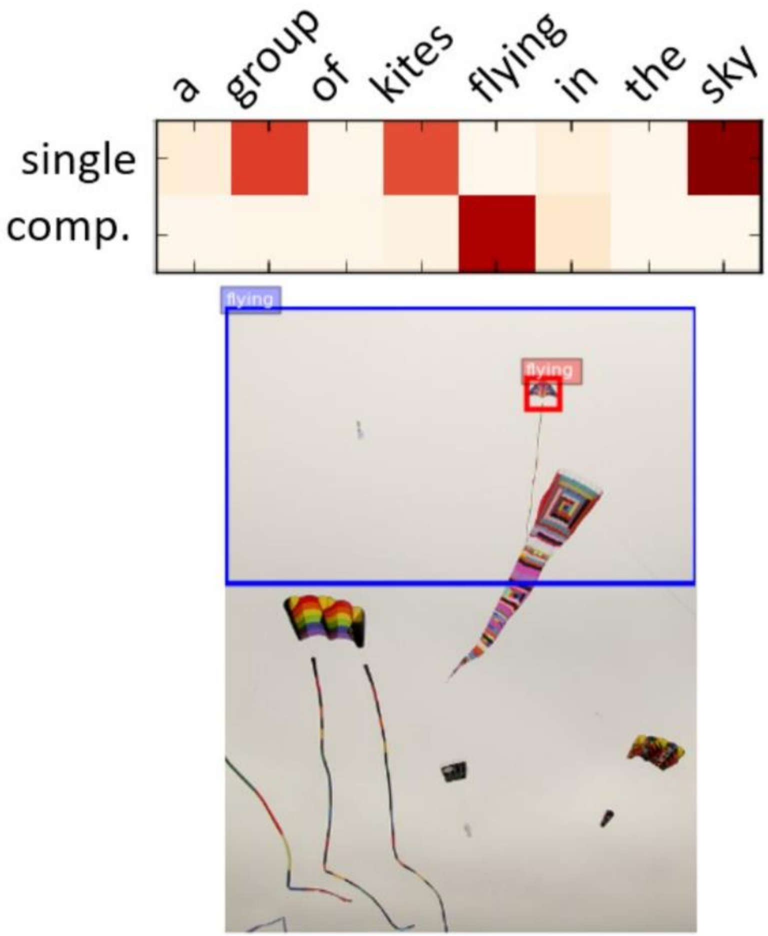


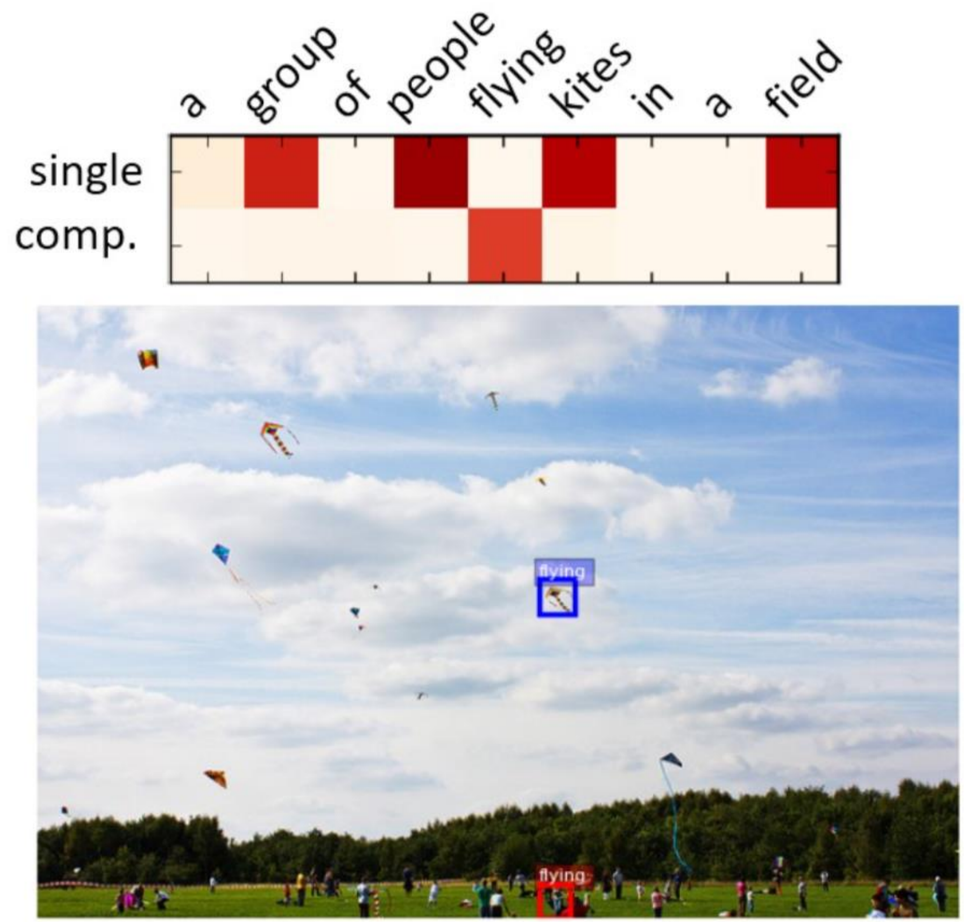
NANYANG
TECHNOLOGICAL
UNIVERSITY

NANYANG TECHNOLOGICAL UNIVERSITY



NANYANG TECHNOLOGICAL UNIVERSITY





NANYANG TECHNOLOGICAL UNIVERSITY

MS-COCO Leaderboard

#	User	Entries	Date of Last Entry	BLEU-4		METEOR		ROUGE-L		CIDEr-D	
				c5 ▲	c40 ▲	c5 ▲	c40 ▲	c5 ▲	c40 ▲	c5 ▲	c40 ▲
1	TencentAI.v2	5	12/15/17	0.386 (1)	0.701 (1)	0.286 (1)	0.377 (1)	0.587 (1)	0.737 (1)	1.254 (1)	1.278 (1)
2	AnonymousTeam	5	11/13/17	0.380 (3)	0.692 (2)	0.282 (3)	0.372 (3)	0.582 (3)	0.731 (2)	1.229 (3)	1.251 (2)
3	TingYao	4	09/03/17	0.382 (2)	0.691 (3)	0.283 (2)	0.373 (2)	0.582 (2)	0.729 (4)	1.232 (2)	1.246 (3)
4	LiuDaqing	3	04/08/18	0.379 (4)	0.690 (4)	0.281 (4)	0.370 (5)	0.582 (4)	0.731 (3)	1.216 (4)	1.238 (4)

We are SINGLE model.

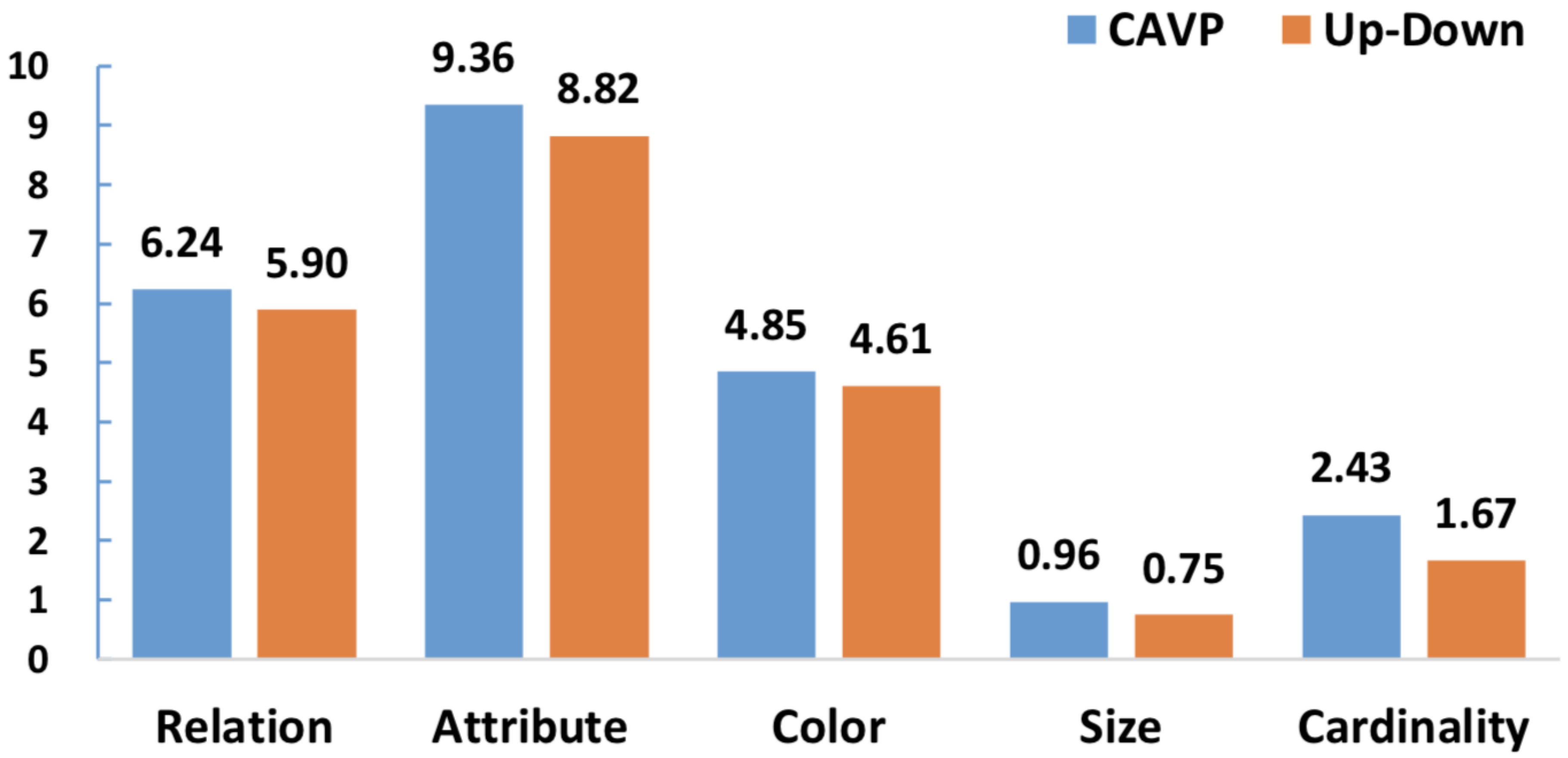
NANYANG TECHNOLOGICAL UNIVERSITY

Compare with Academic Peers

Model	B@4	M	R	C	S
Google NIC[35]	32.1	25.7	-	99.8	-
Hard-Attention[38]	24.3	23.9	-	-	-
Adaptive[23]	33.2	26.6	54.9	108.5	19.4
LSTM-A[39]	32.5	25.1	53.8	98.6	-
PG-SPIDEr[22]	32.2	25.1	54.4	100.0	-
Actor-Critic[43]	34.4	26.7	55.8	116.2	-
EmbeddingReward[28]	30.4	25.1	52.5	93.7	-
SCST[29]	35.4	27.1	56.6	117.5	-
StackCap[9]	36.1	27.4	56.9	120.4	20.9
Up-Down[2]	36.3	27.7	56.9	120.1	21.4
Ours	**38.6**	**28.3**	**58.5**	**126.3**	**21.6**

Table 2: Performance comparisons on MS-COCO "Karpathy" offline split. B@n is short for BLEU-n, M is short for METEOR, and C is short for CIDEr.

Detail Comparison with Up-Down

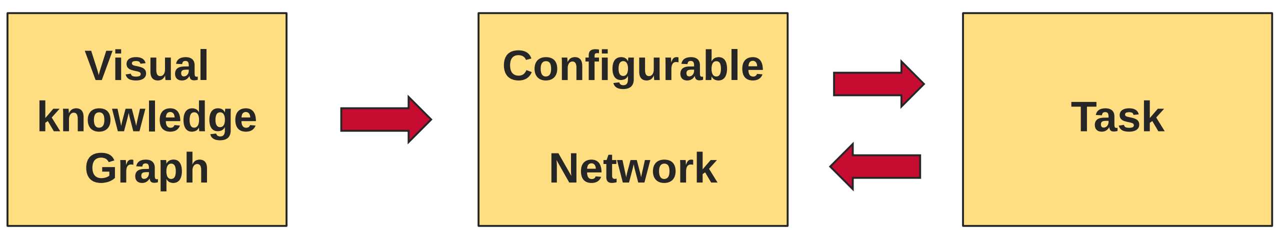


P. Anderson et al. Bottom-up and top-down attention for image captioning and VQA. In *CVPR'18*

NANYANG TECHNOLOGICAL UNIVERSITY

Visual Reasoning: A Desired Pipeline

- Configurable NN for various reasoning applications :

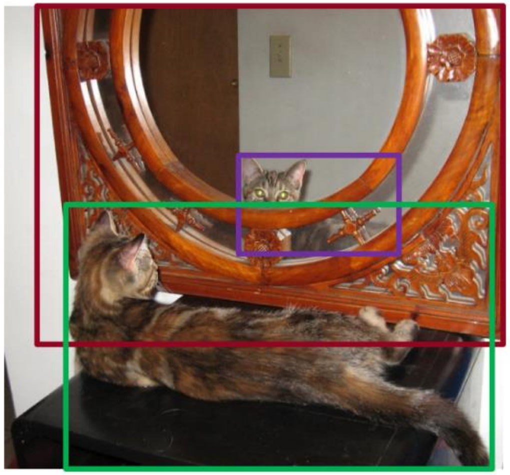


Captioning, VQA, and Visual Dialogue

Visual Reasoning: Future Directions

- Compositionality
  - Good SG generation
  - Robust SG representation
  - Task-specific SG generation
- Learning to reason
  - Task-specific network
  - Good policy-gradient RL for large SG

Scene Dynamics



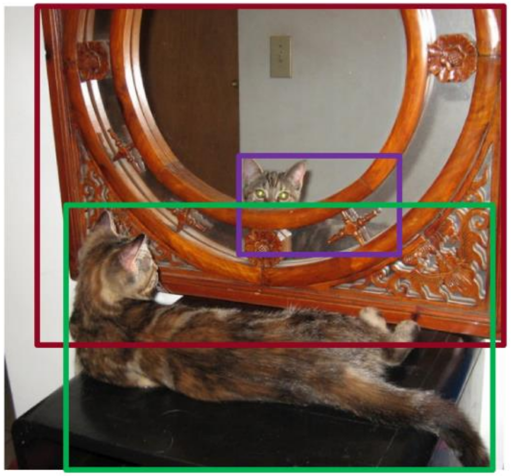
Scene Dynamics

time step: 0
(Initial State)

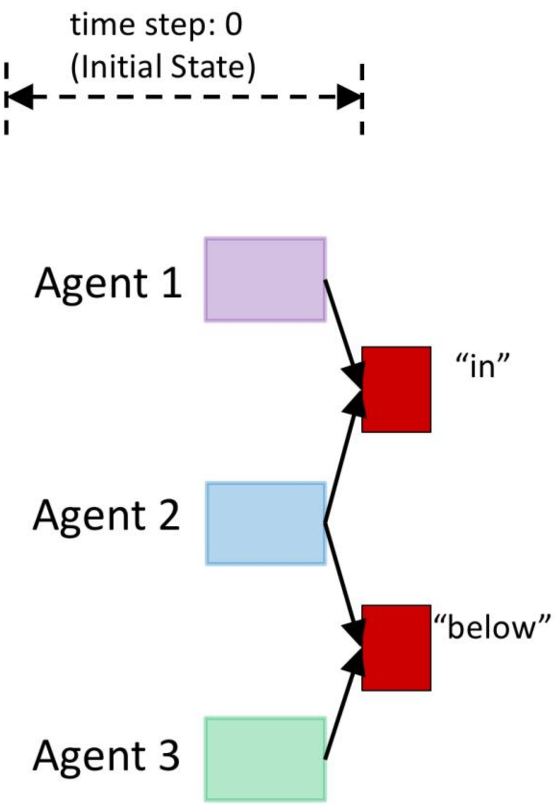
Agent 1

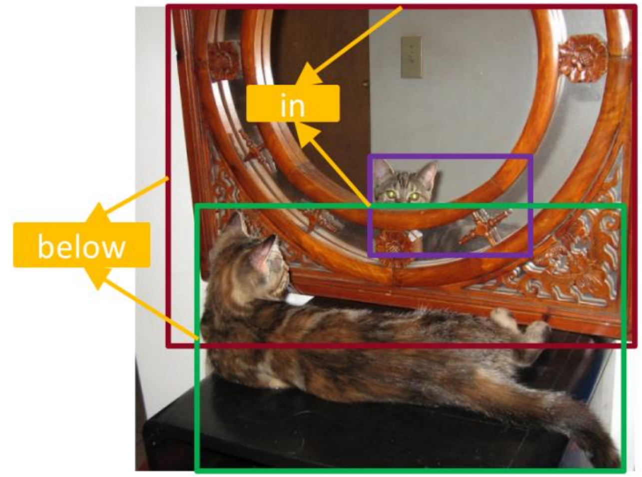
Agent 2

Agent 3

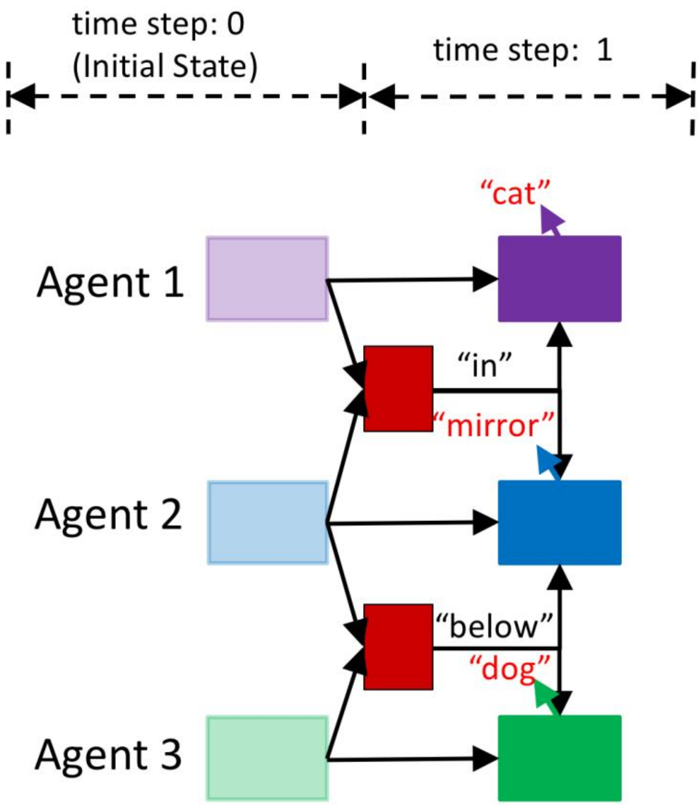


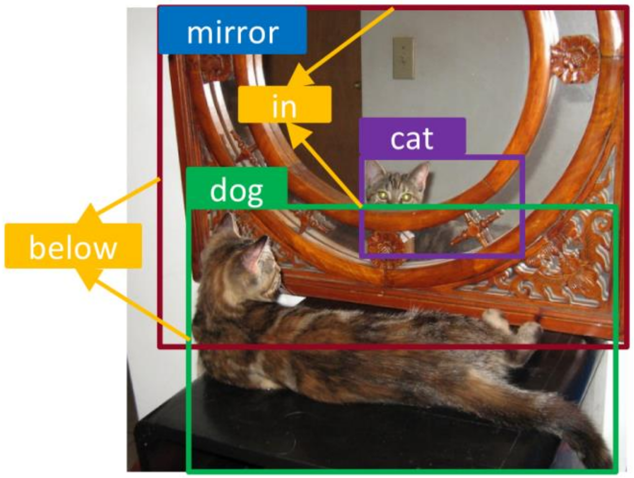
Scene Dynamics



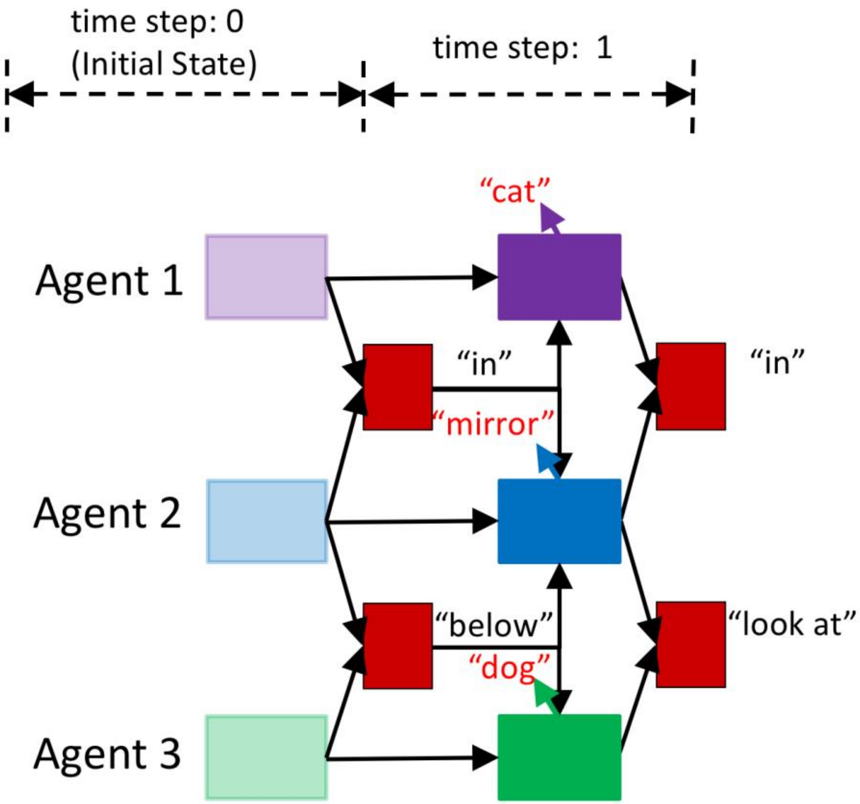


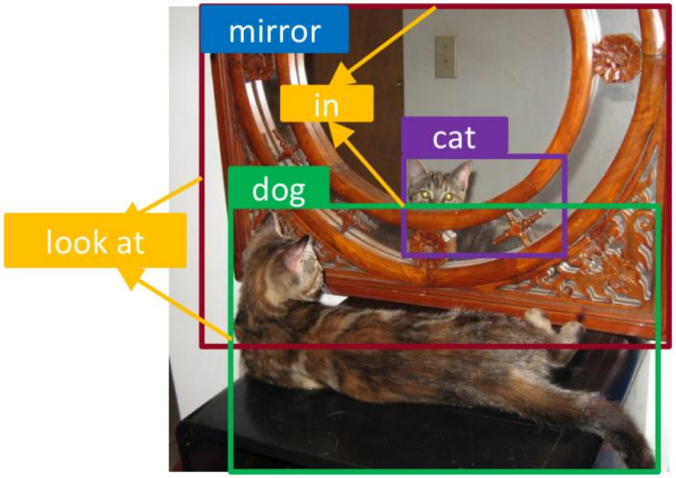
Scene Dynamics



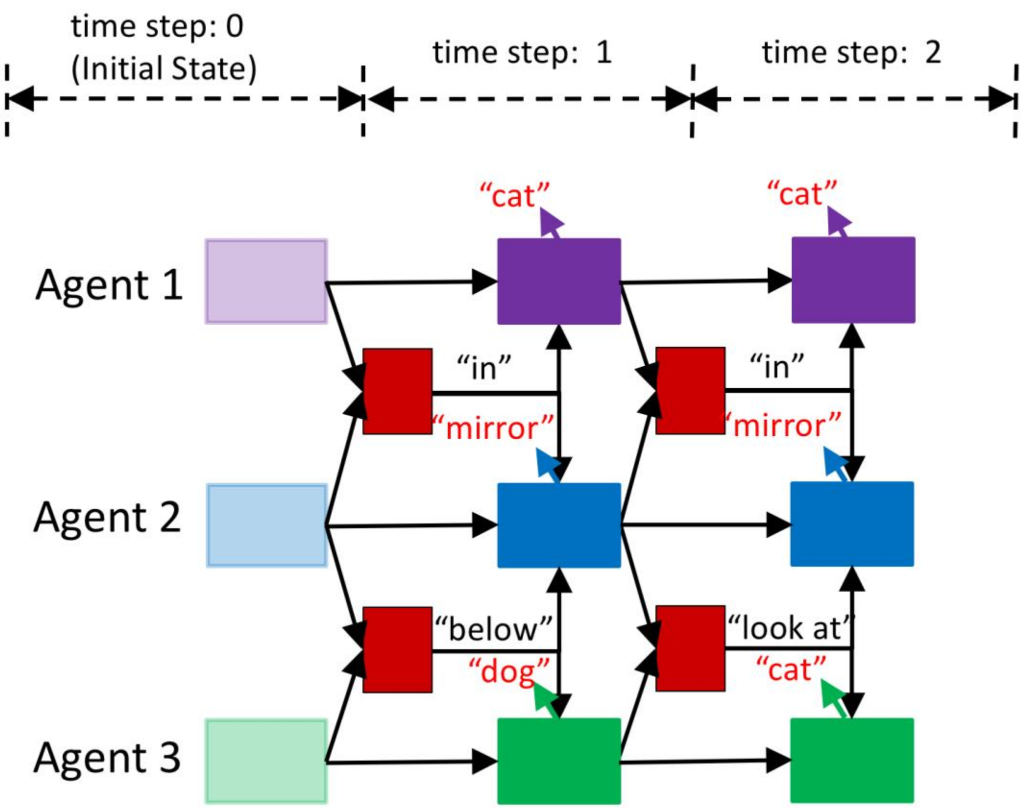


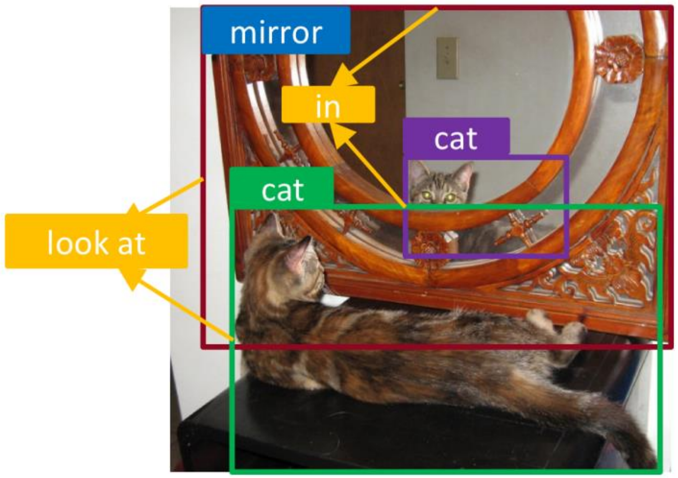
Scene Dynamics





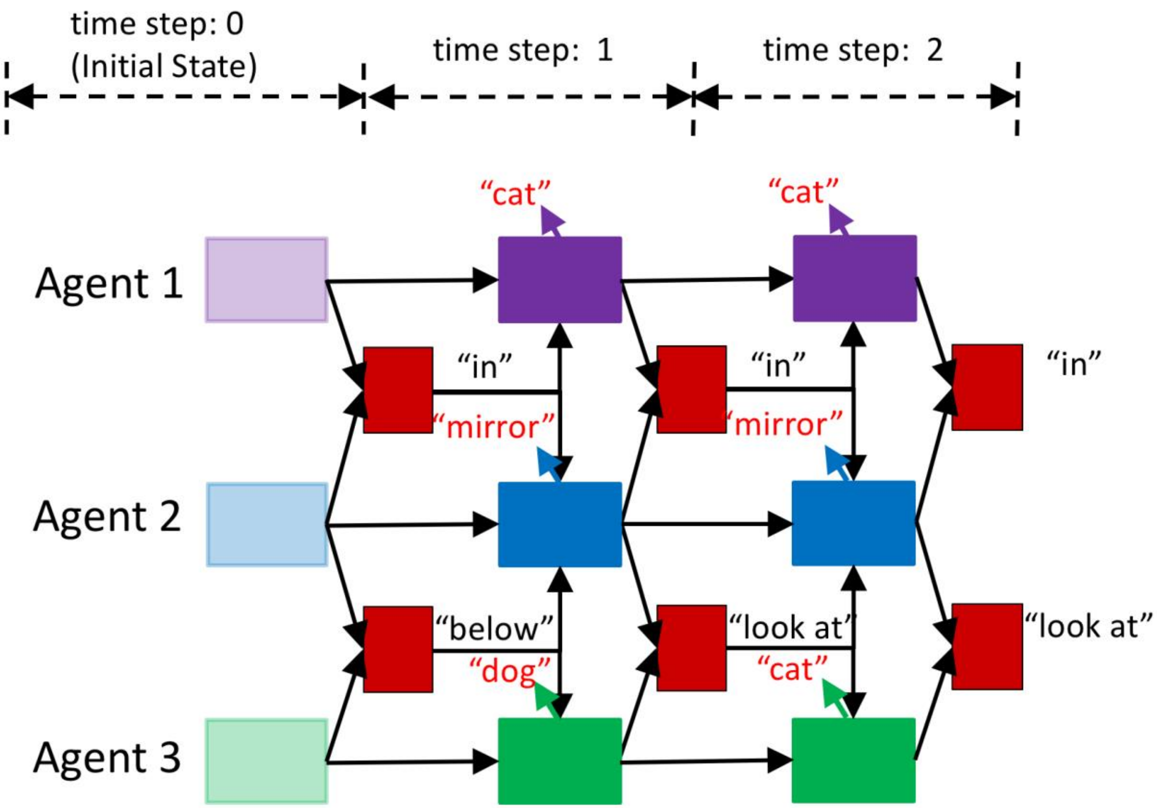
Scene Dynamics

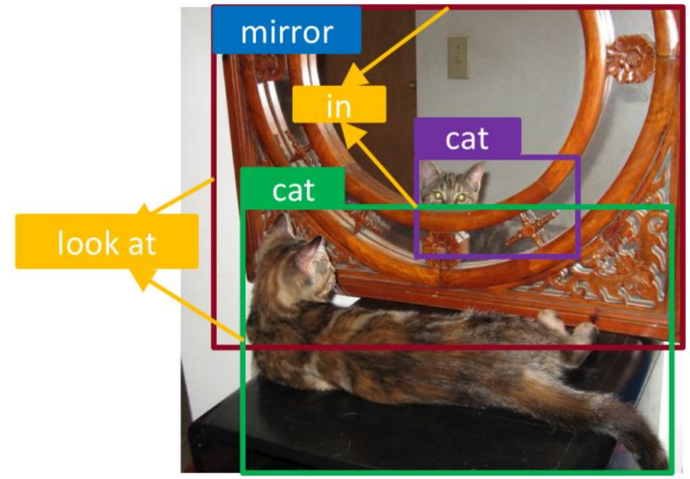




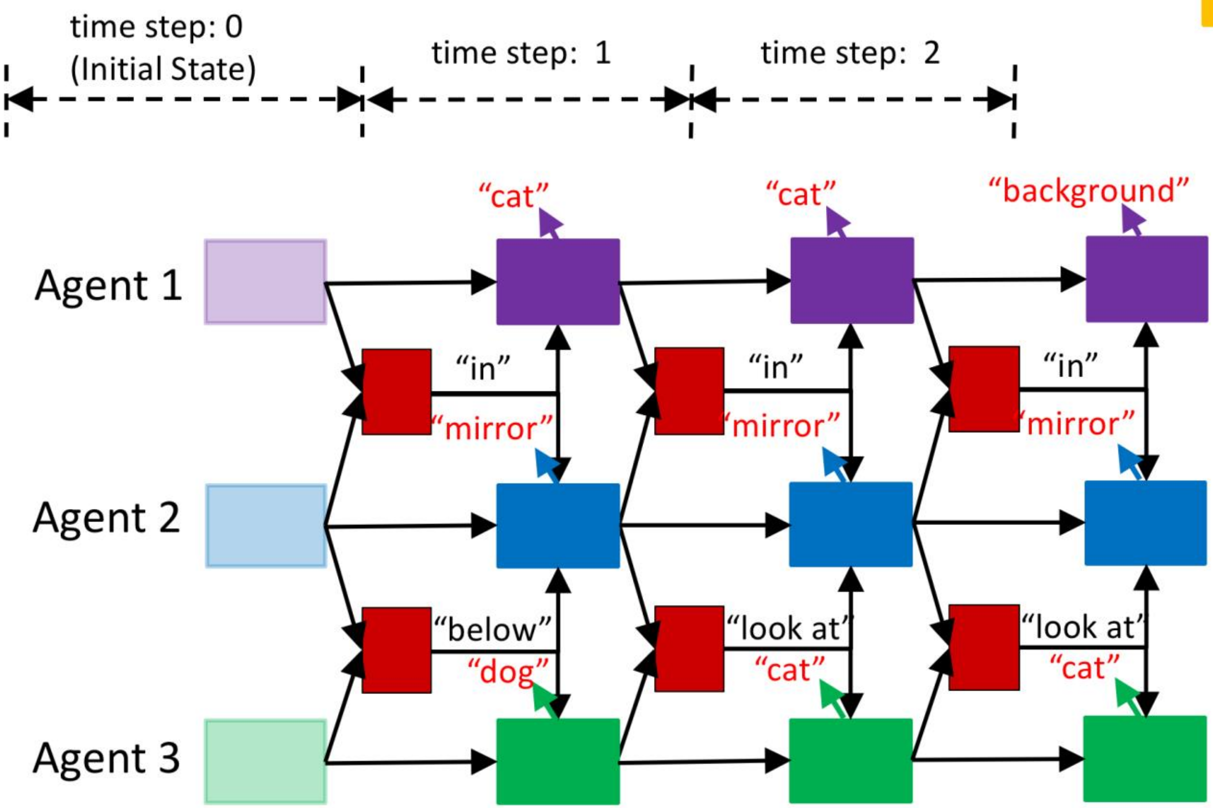
NANYANG TECHNOLOGICAL UNIVERSITY

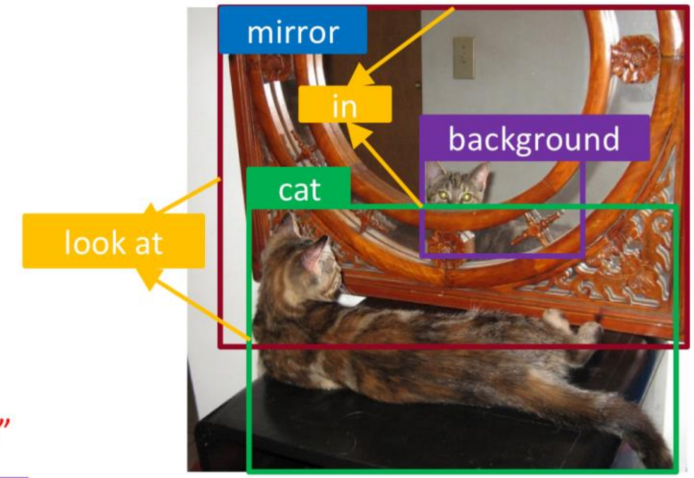
Scene Dynamics



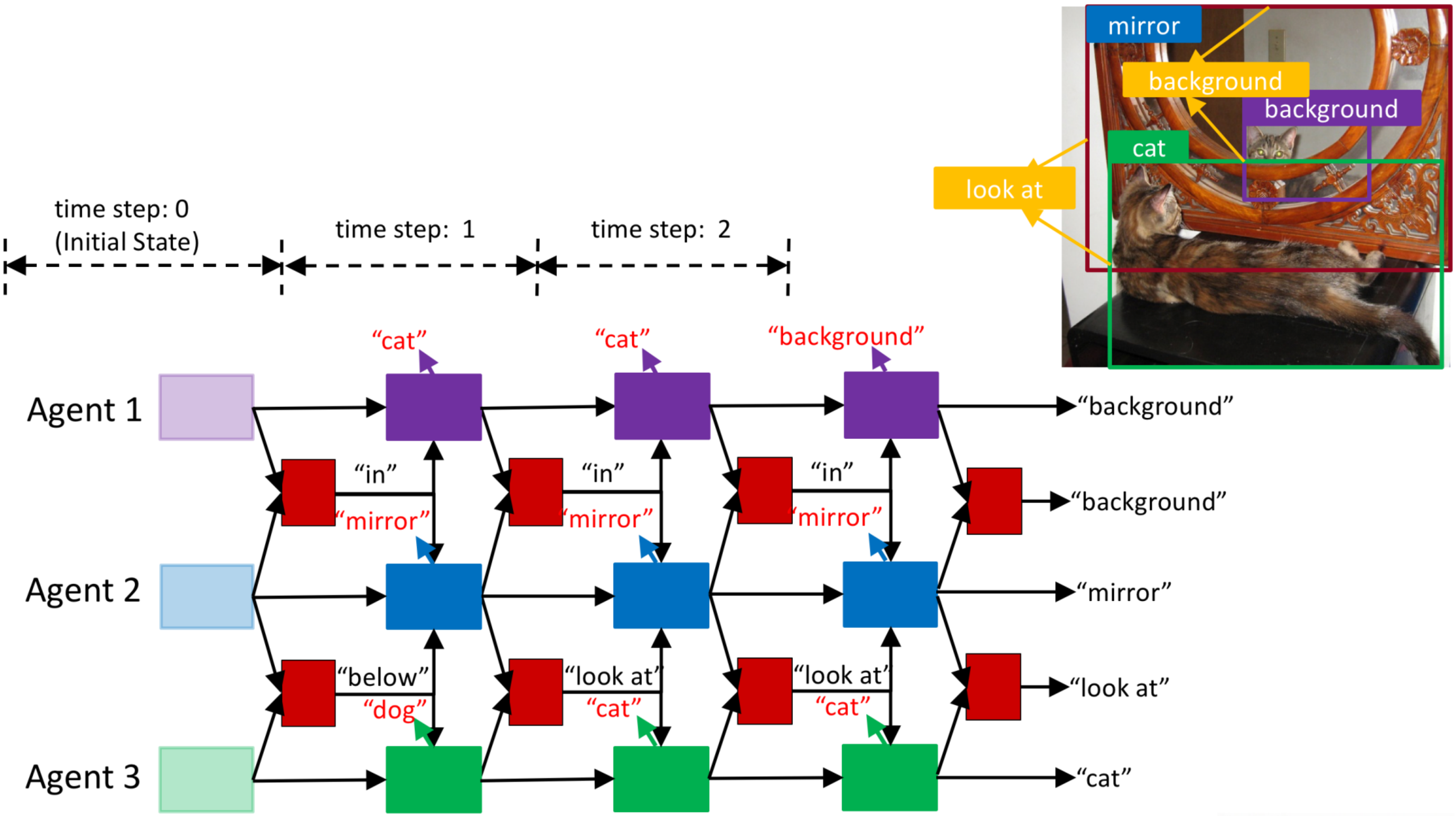


Scene Dynamics



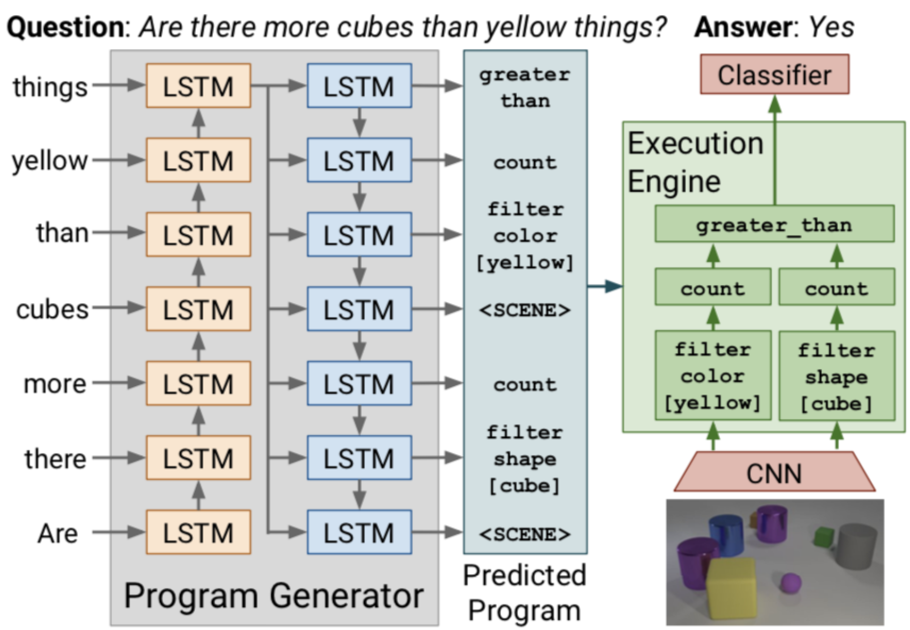


Scene Dynamics



NANYANG TECHNOLOGICAL UNIVERSITY

Hard-design X Module Network



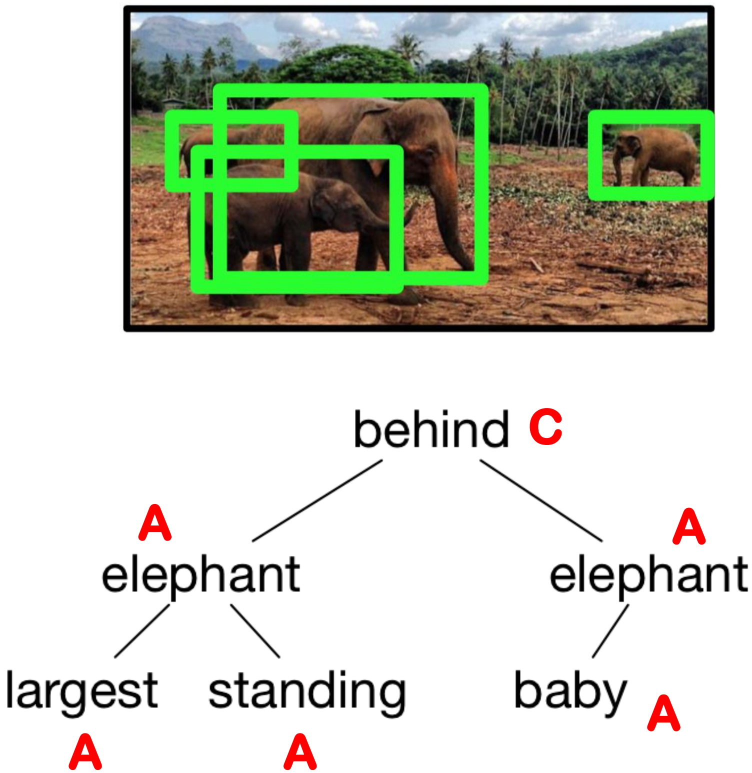
- Q → Program not X
- Module X but hard-design

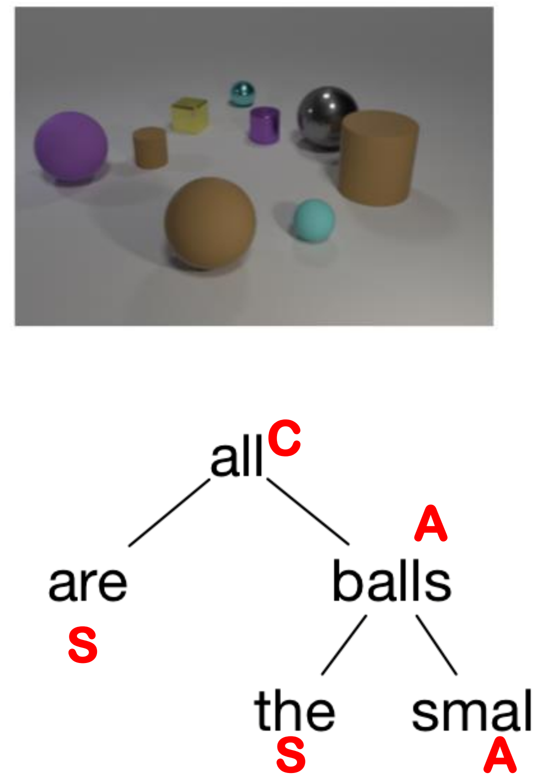
- CLEVER hacker
- Poor generalization to COCO-VQA

Jonson et al. ICCV’17
Hu et al. ICCV’17
Mascharka et al. CVPR’18

NANYANG TECHNOLOGICAL UNIVERSITY

Design-Free Module Network





Choi et al. Learning to compose task-specific tree structures. AAAI'17