COMP 790.139 (Fall 2017)
Natural Language Processing

Language+Vision; Guest Research Talks by Ramakanth, Hao

THE UNIVERSITY
of NORTH CAROLINA
at CHAPEL HILL

Mohit Bansal

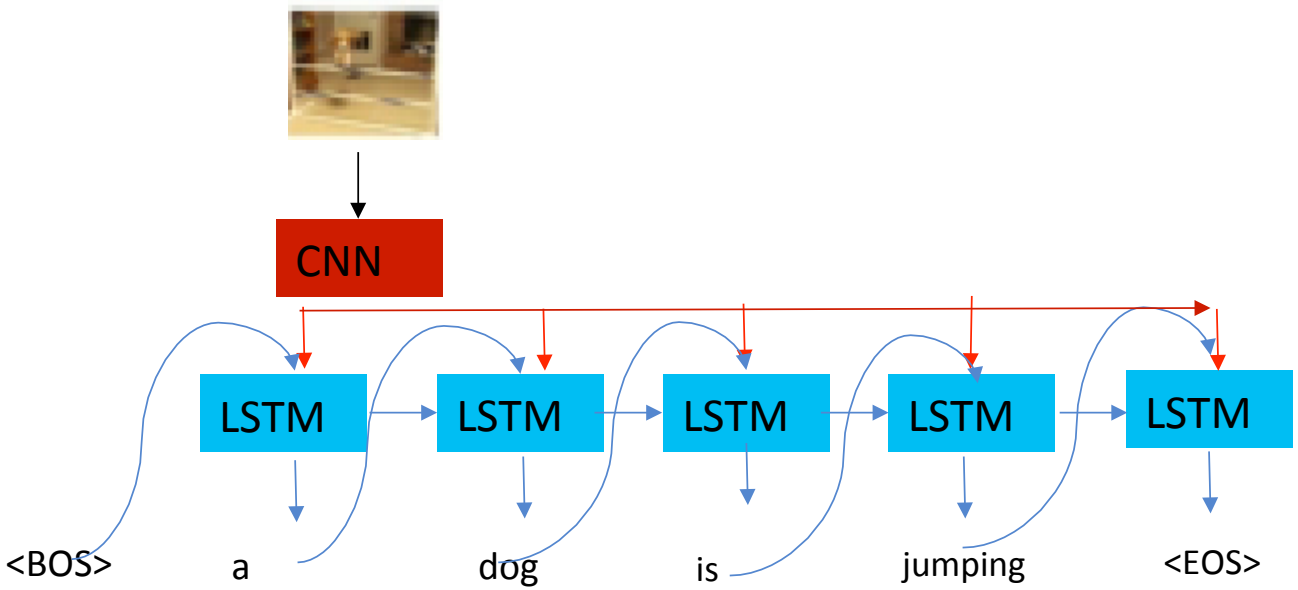
Language+Vision

Major Language+Vision Tasks

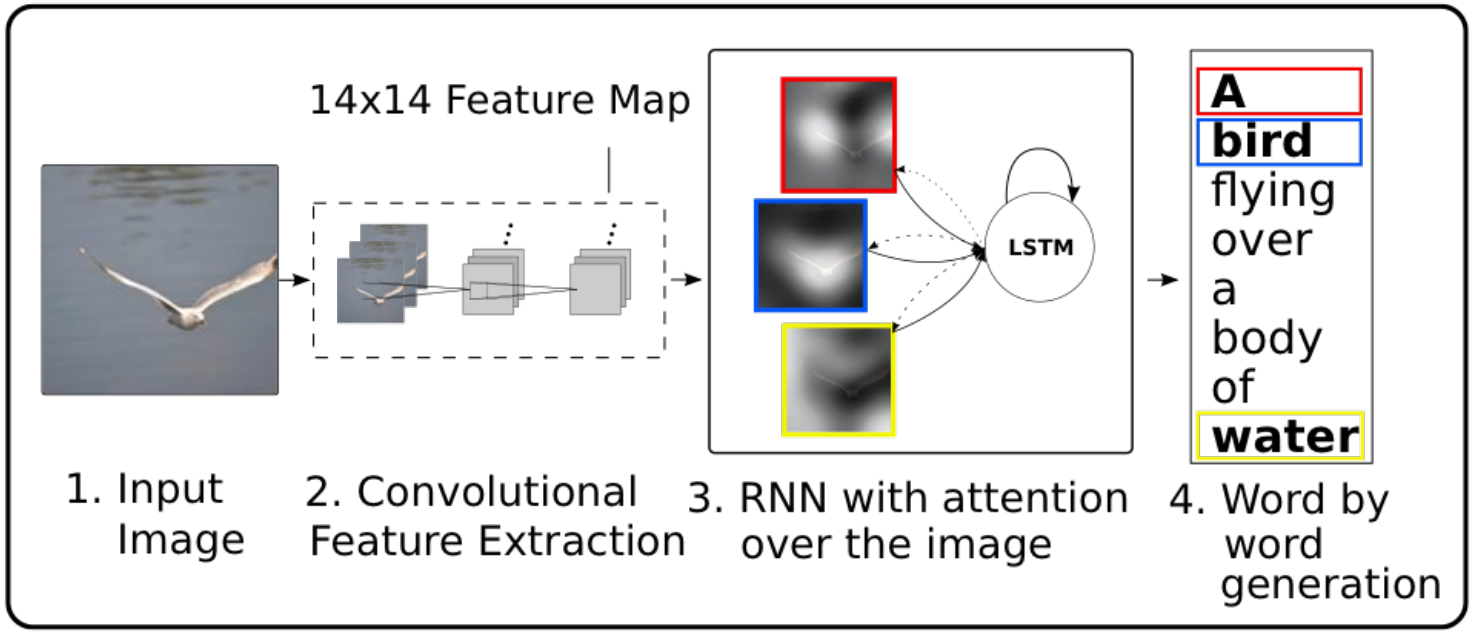
- Image Captioning
- Referring Expressions
- Image/Visual Question Answering
- Visual Dialog
- Video Captioning

Brief Task Definitions and Example Papers/Models

Image Captioning



Show, Attend, and Tell



Attention:

$$e_{ti} = f_{\text{att}}(\mathbf{a}_i, \mathbf{h}_{t-1})$$

$$\alpha_{ti} = \frac{\exp(e_{ti})}{\sum_{k=1}^{L} \exp(e_{tk})}.$$

[Xu et al., 2015]

Show, Attend, and Tell

A woman is throwing a frisbee in a park.

A dog is standing on a hardwood floor.

A stop sign is on a road with a mountain in the background.

A little girl sitting on a bed with a teddy bear.

A group of people sitting on a boat in the water.

A giraffe standing in a forest with trees in the background.

[Xu et al., 2015]

Visual Referring Expressions

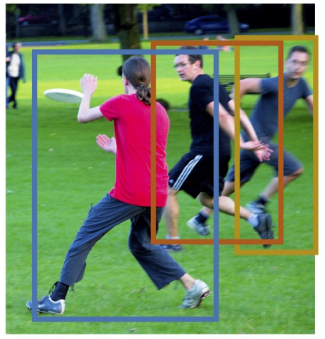
RefCOCO TestB

bottom left banana

second banana from left

top right banana

RefCOCO+ TestA

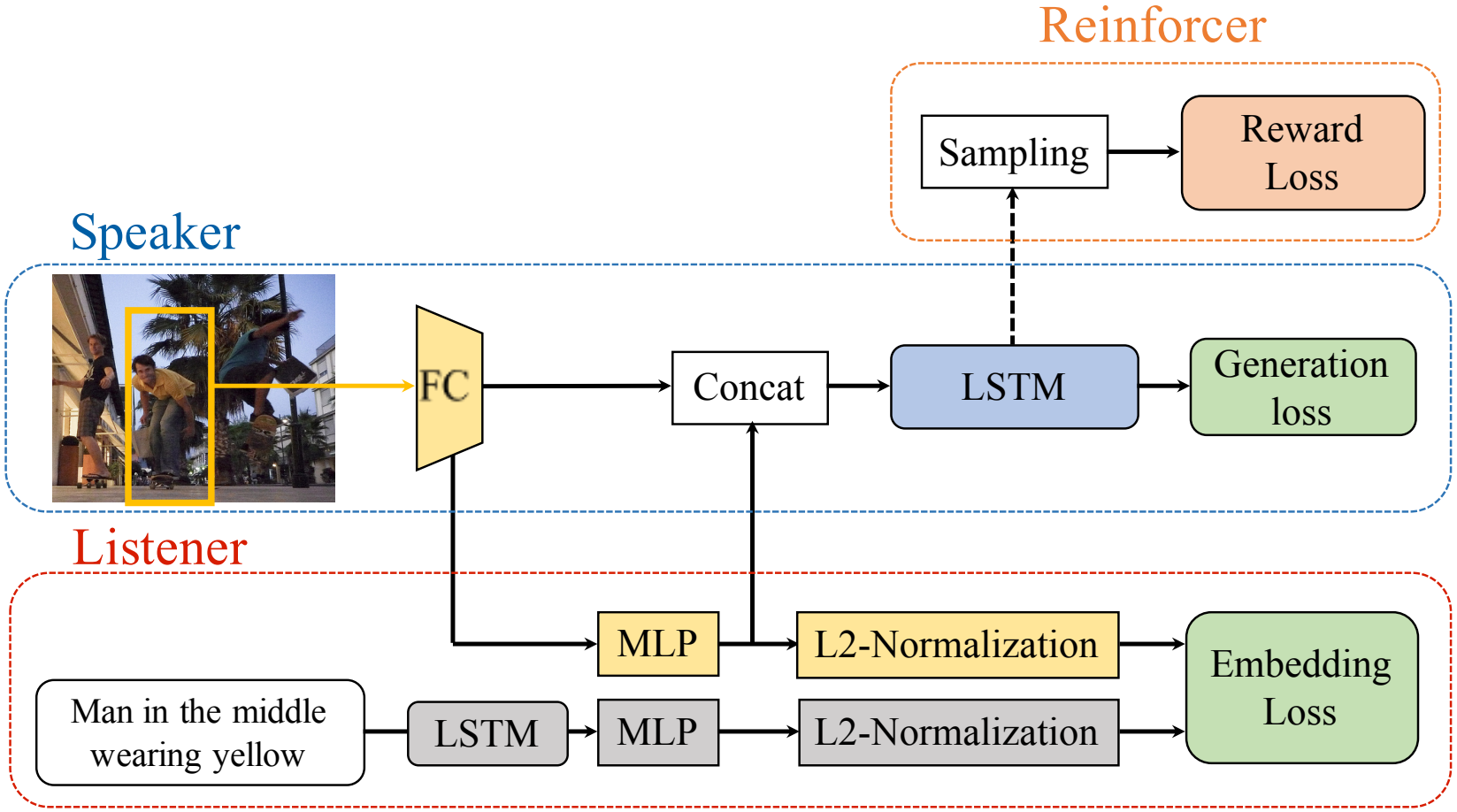


red shirt

man in black

blue shirt

Joint Comprehension+Generation Model



[Yu et al., 2017]

Joint Comprehension+Generation Model

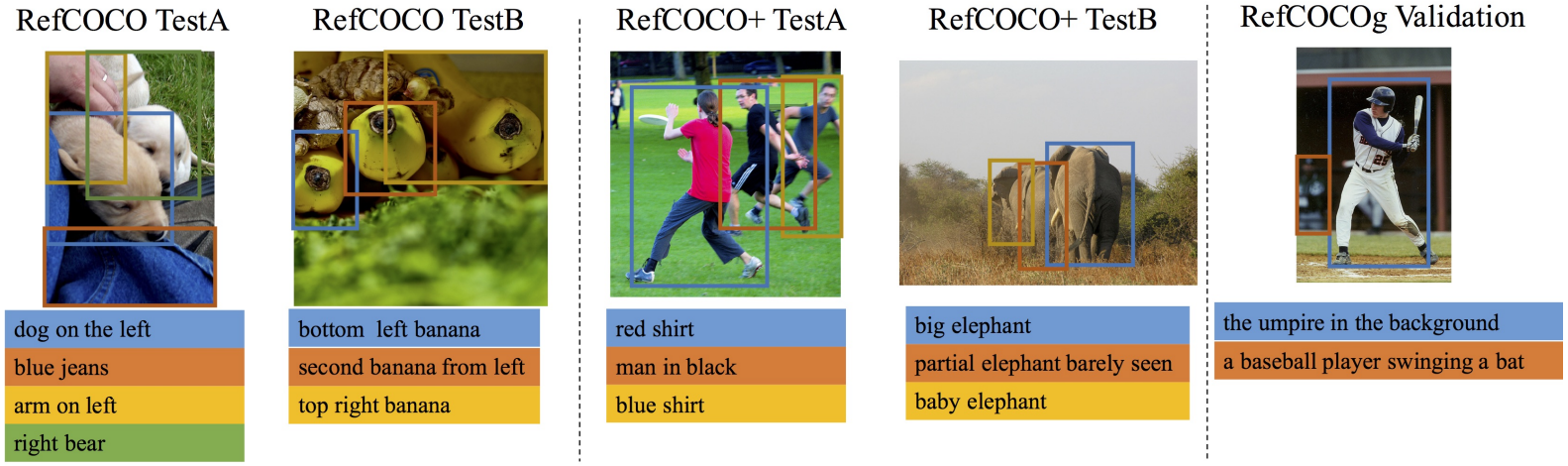


Figure 1: Joint generation examples using our full model with "+rerank" on three datasets. Each sentence shows the generated expression for one of the depicted objects (color coded to indicate correspondence).

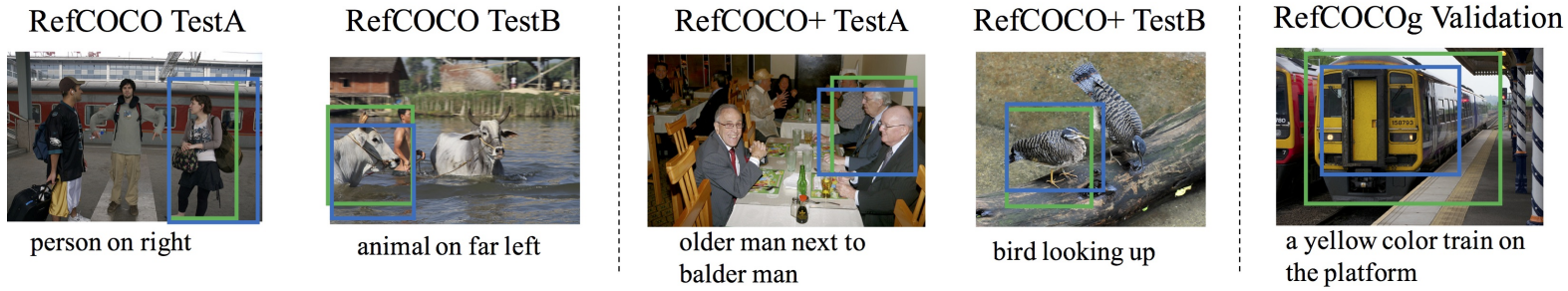


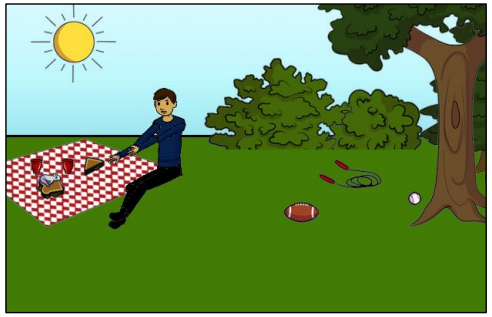
Figure 2: Example comprehension results using our full model on three datasets. Green box shows the ground-truth region and blue box shows our correct comprehension based on the detected regions.

[Yu et al., 2017]

VQA: Visual Question Answering

What color are her eyes?
What is the mustache made of?

How many slices of pizza are there?
Is this a vegetarian pizza?



Is this person expecting company?
What is just under the tree?

Does it appear to be rainy?
Does this person have 20/20 vision?

[Agrawal et al., 2015]

Demo

- http://vqa.cloudcv.org/

how many giraffes are there?

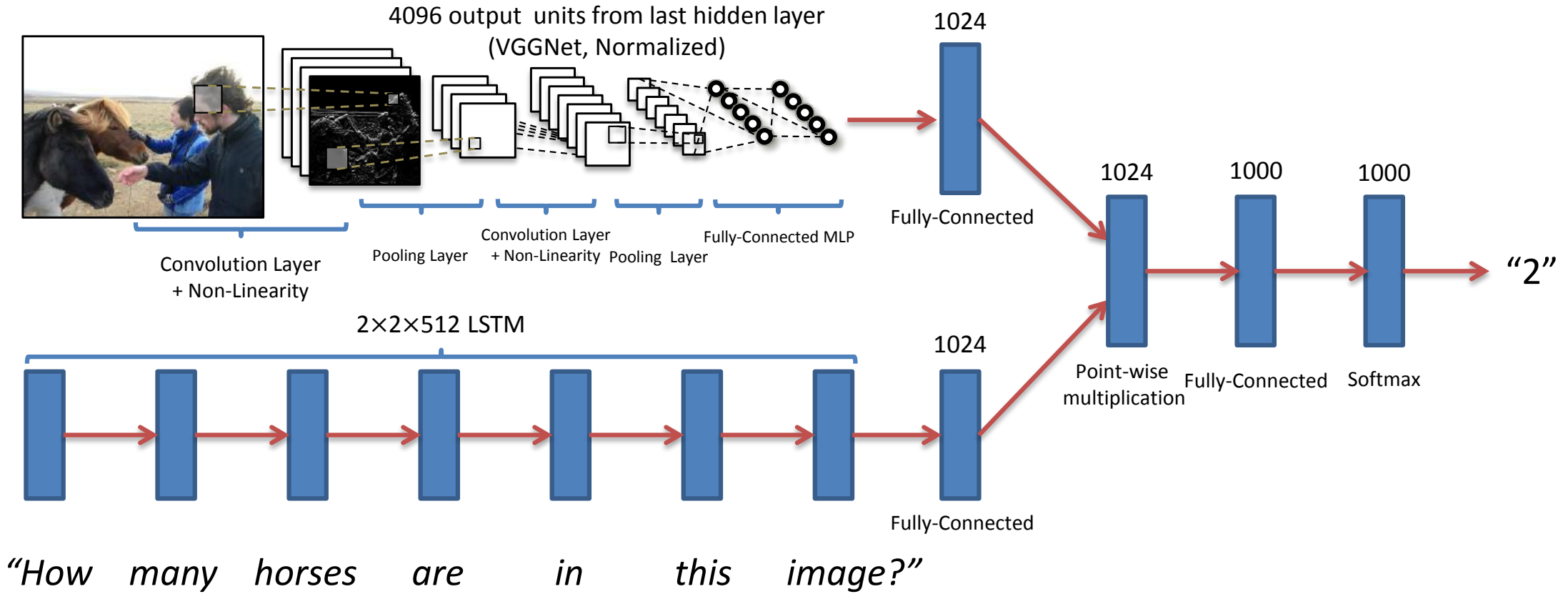
Submit

Predicted top-5 answers with confidence:

2	67.267%
1	22.324%
3	9.115%
4	0.945%
0	0.242%

[Agrawal et al., 2015]

Simple VQA Baseline



[Agrawal et al., 2015]

Hierarchical Co-Attention Model

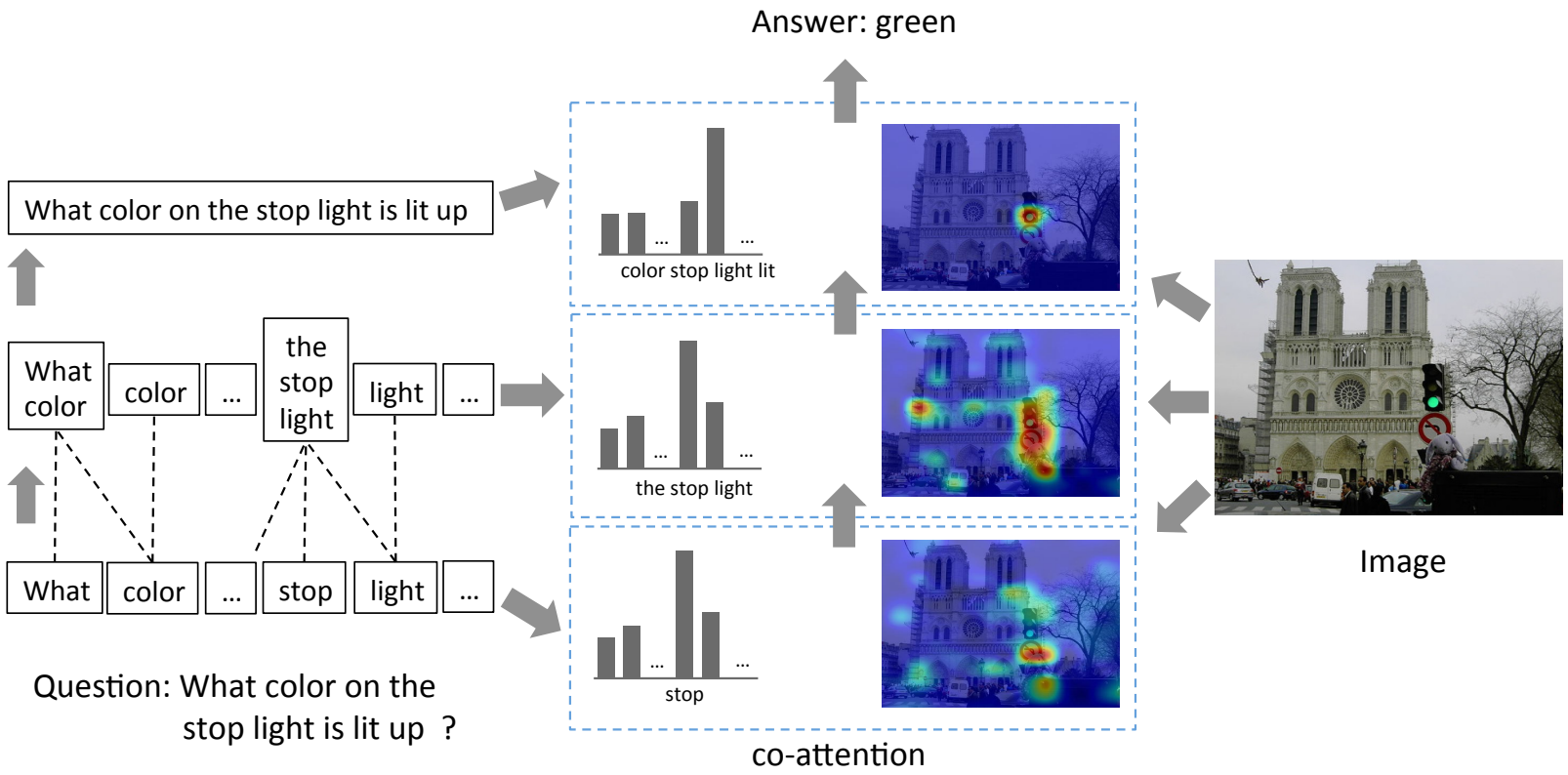


Figure 1: Flowchart of our proposed hierarchical co-attention model. Given a question, we extract its word level, phrase level and question level embeddings. At each level, we apply co-attention on both the image and question. The final answer prediction is based on all the co-attended image and question features.

[Lu et al., 2016]

Hierarchical Co-Attention Model

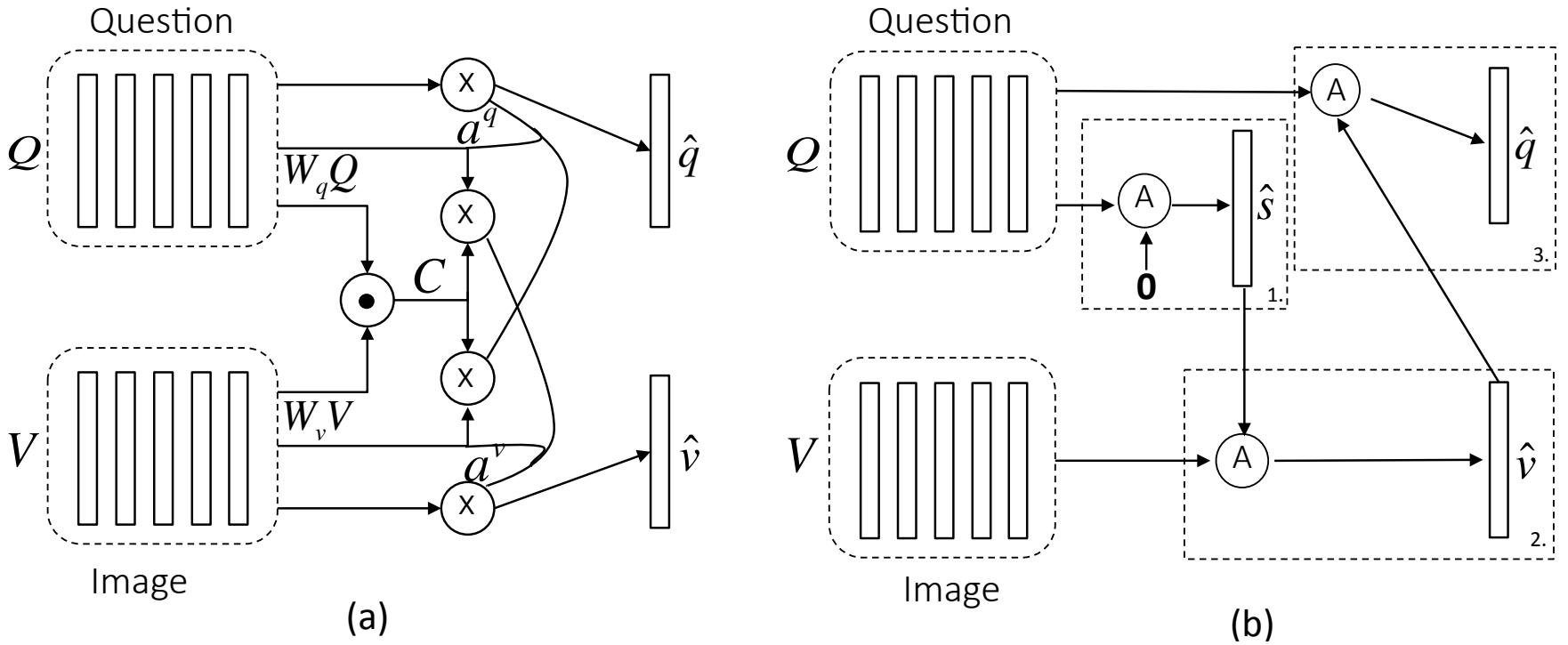


Figure 2: (a) Parallel co-attention mechanism; (b) Alternating co-attention mechanism.

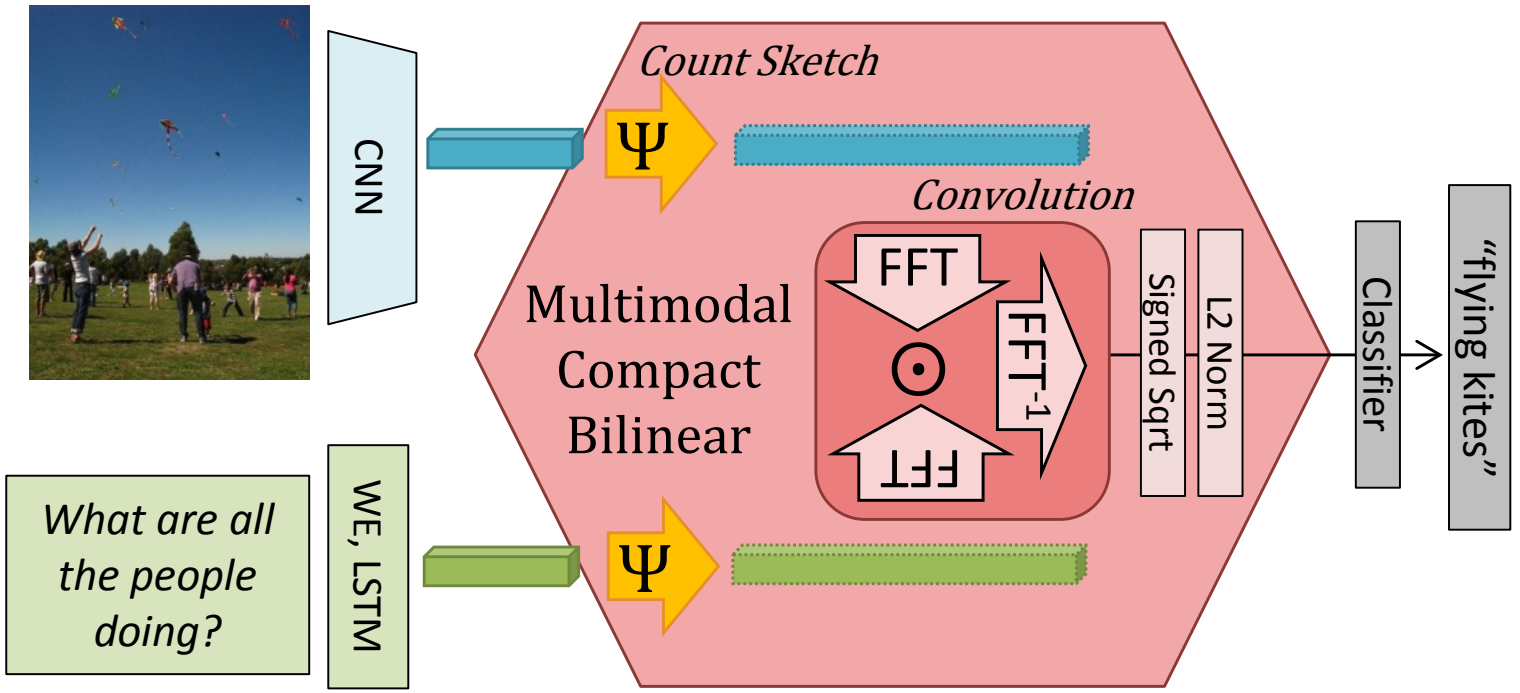
[Lu et al., 2016]

Hierarchical Co-Attention Model

Figure 4: Visualization of image and question co-attention maps on the COCO-QA dataset. From left to right: original image and question pairs, word level co-attention maps, phrase level co-attention maps and question level co-attention maps. For visualization, both image and question attentions are scaled (from red:high to blue:low). Best viewed in color.

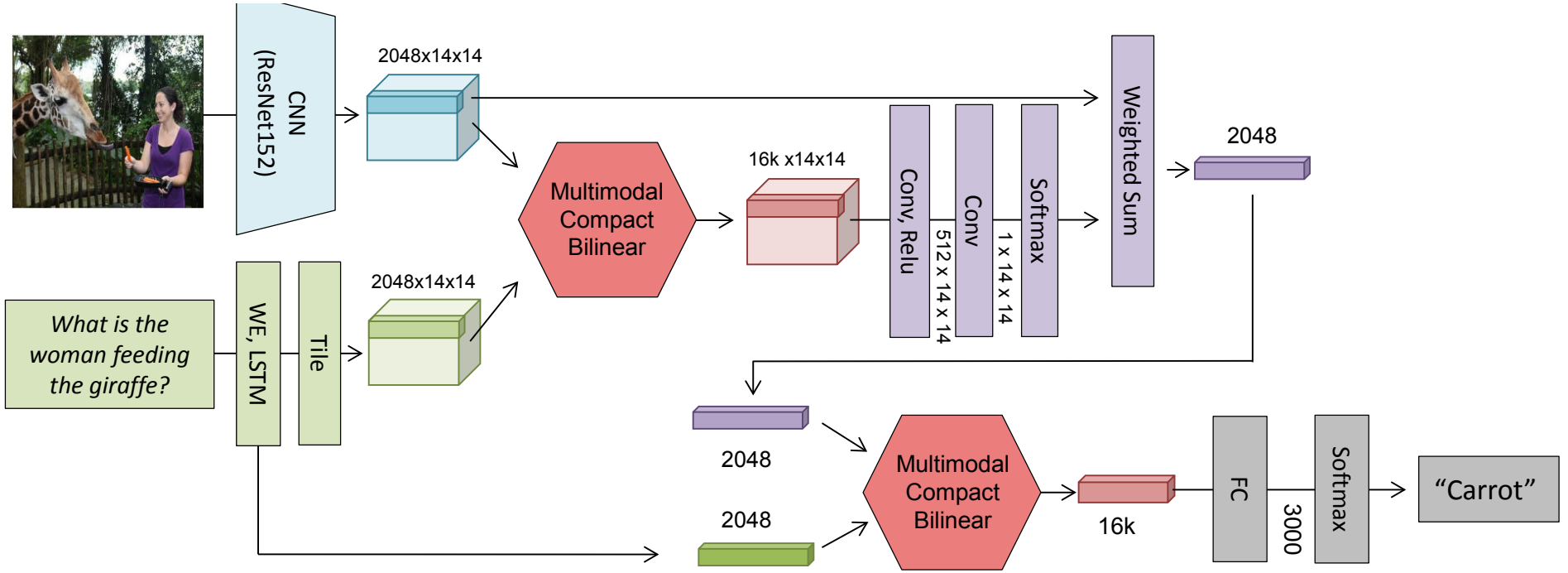
[Lu et al., 2016]

Multimodal Compact Bilinear (MCB) Model



[Fukui et al., 2016]

MCB Model with Attention



[Fukui et al., 2016]

Results

	Test-dev					Test-standard				
	Open Ended				MC	Open Ended				MC
	Y/N	No.	Other	All	All	Y/N	No.	Other	All	All
MCB	81.2	35.1	49.3	60.8	65.4	-	-	-	-	-
MCB + Genome	81.7	36.6	51.5	62.3	66.4	-	-	-	-	-
MCB + Att.	82.2	37.7	54.8	64.2	68.6	-	-	-	-	-
MCB + Att. + GloVe	82.5	37.6	55.6	64.7	69.1	-	-	-	-	-
MCB + Att. + Genome	81.7	38.2	57.0	65.1	69.5	-	-	-	-	-
MCB + Att. + GloVe + Genome	82.3	37.2	57.4	65.4	69.9	-	-	-	-	-
Ensemble of 7 Att. models	**83.4**	**39.8**	**58.5**	**66.7**	**70.2**	**83.2**	**39.5**	**58.0**	**66.5**	**70.1**
Naver Labs (challenge 2nd)	83.5	39.8	54.8	64.9	69.4	83.3	38.7	54.6	64.8	69.3
HieCoAtt (Lu et al., 2016)	79.7	38.7	51.7	61.8	65.8	-	-	-	62.1	66.1
DMN+ (Xiong et al., 2016)	80.5	36.8	48.3	60.3	-	-	-	-	60.4	-
FDA (Ilievski et al., 2016)	81.1	36.2	45.8	59.2	-	-	-	-	59.5	-
D-NMN (Andreas et al., 2016a)	81.1	38.6	45.5	59.4	-	-	-	-	59.4	-
AMA (Wu et al., 2016)	81.0	38.4	45.2	59.2	-	81.1	37.1	45.8	59.4	-
SAN (Yang et al., 2015)	79.3	36.6	46.1	58.7	-	-	-	-	58.9	-
NMN (Andreas et al., 2016b)	81.2	38.0	44.0	58.6	-	81.2	37.7	44.0	58.7	-
AYN (Malinowski et al., 2016)	78.4	36.4	46.3	58.4	-	78.2	36.3	46.3	58.4	-
SMem (Xu and Saenko, 2016)	80.9	37.3	43.1	58.0	-	80.9	37.5	43.5	58.2	-
VQA team (Antol et al., 2015)	80.5	36.8	43.1	57.8	62.7	80.6	36.5	43.7	58.2	63.1
DPPnet (Noh et al., 2015)	80.7	37.2	41.7	57.2	-	80.3	36.9	42.2	57.4	-
iBOWIMG (Zhou et al., 2015)	76.5	35.0	42.6	55.7	-	76.8	35.0	42.6	55.9	62.0

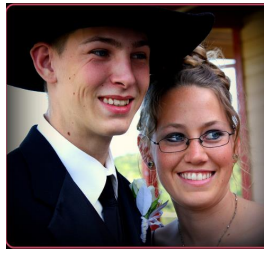
[Fukui et al., 2016]

Making the V in the VQA matter !

Who is wearing glasses?

man

woman



Where is the child sitting?

fridge

arms

Is the umbrella upside down?

yes

no

How many children are in the bed?

2

1

[Goyal et al., 2017]

Results

Approach	Ans Type	UU	UB	$B_{half}B$	BB
MCB [9]	Yes/No	81.20	70.40	74.89	77.37
	Number	34.80	31.61	34.69	36.66
	Other	51.19	47.90	47.43	51.23
	All	60.36	54.22	56.08	59.14
HieCoAtt [25]	Yes/No	79.99	67.62	70.93	71.80
	Number	34.83	32.12	34.07	36.53
	Other	45.55	41.96	42.11	46.25
	All	57.09	50.31	51.88	54.57

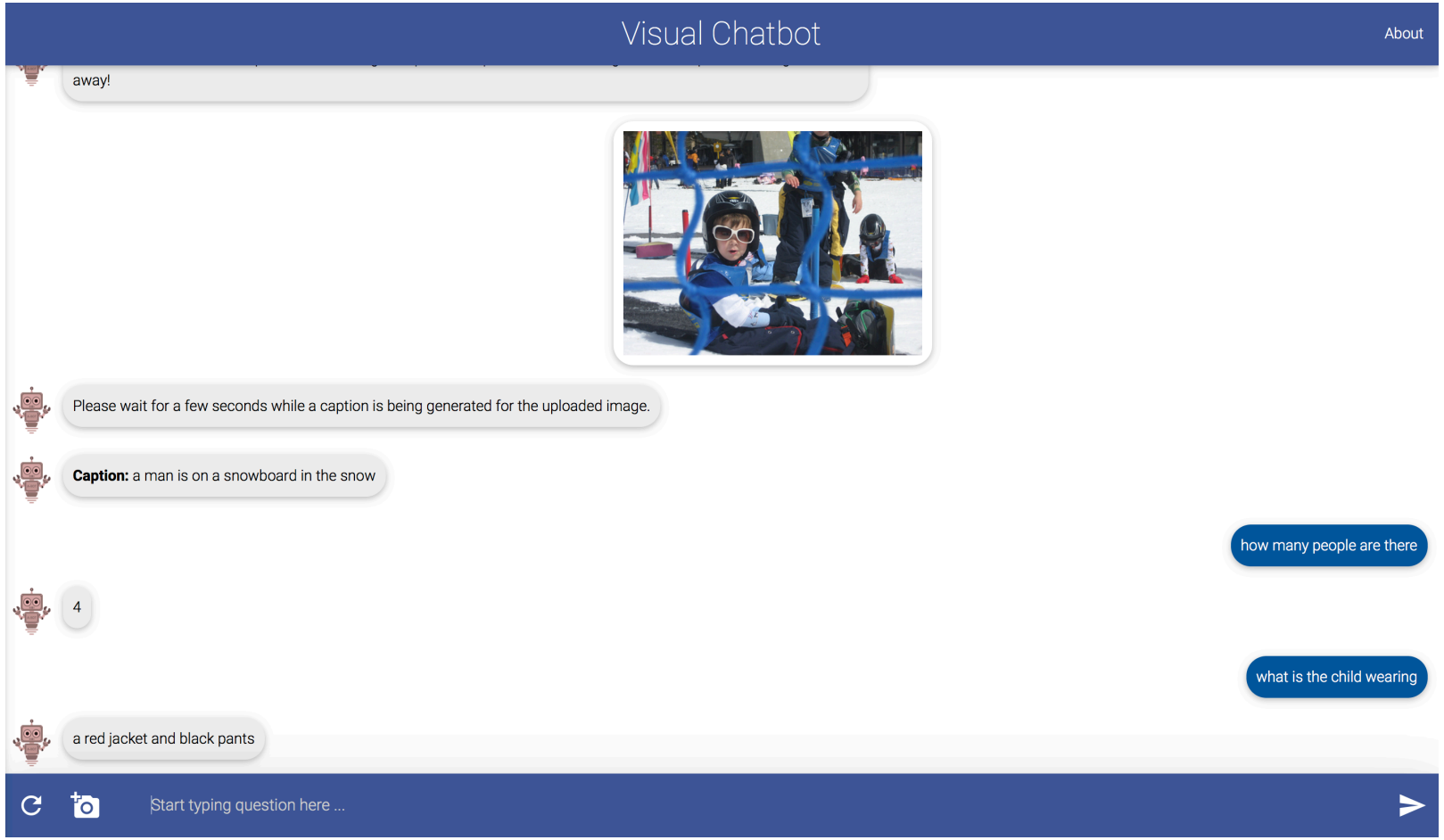
[Goyal et al., 2017]

Visual Dialog

[Das et al., 2017]

Demo

- http://visualchatbot.cloudcv.org/



[Das et al., 2017]

Visual Dialog vs VQA

VQA

Q: How many people on wheelchairs ?

A: Two

Q: How many wheelchairs ?

A: One

Captioning

Two people are in a wheelchair and one is holding a racket.

Visual Dialog

Q: How many people are on wheelchairs ?

A: Two

Q: What are their genders ?

A: One male and one female

Q: Which one is holding a racket ?

A: The woman

Visual Dialog

Q: What is the gender of the one in the white shirt ?

A: She is a woman

Q: What is she doing ?

A: Playing a Wii game

Q: Is that a man to her right

A: No, it's a woman

[Das et al., 2017]

Results

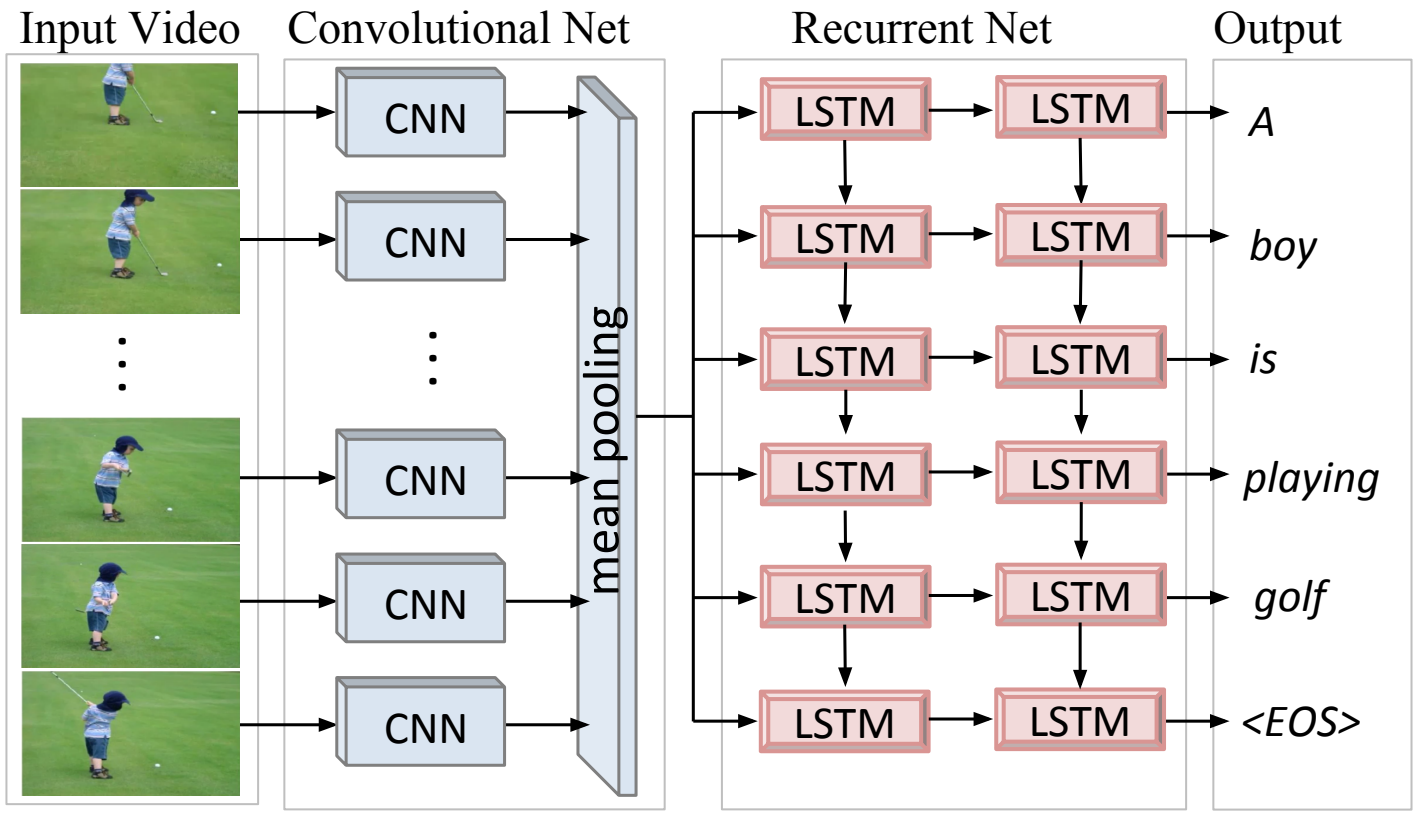
	Model	MRR	R@1	R@5	R@10	Mean
Baseline	Answer prior	0.3735	23.55	48.52	53.23	26.50
	NN-Q	0.4570	35.93	54.07	60.26	18.93
	NN-QI	0.4274	33.13	50.83	58.69	19.62
Generative	LF-Q-G	0.5048	39.78	60.58	66.33	17.89
	LF-QH-G	0.5055	39.73	60.86	66.68	17.78
	LF-QI-G	0.5204	42.04	61.65	67.66	16.84
	LF-QIH-G	0.5199	41.83	61.78	67.59	17.07
	HRE-QH-G	0.5102	40.15	61.59	67.36	17.47
	HRE-QIH-G	0.5237	**42.29**	62.18	67.92	17.07
	HREA-QIH-G	0.5242	42.28	62.33	68.17	**16.79**
	MN-QH-G	0.5115	40.42	61.57	67.44	17.74
	MN-QIH-G	**0.5259**	**42.29**	**62.85**	**68.88**	17.06
Discriminative	LF-Q-D	0.5508	41.24	70.45	79.83	7.08
	LF-QH-D	0.5578	41.75	71.45	80.94	6.74
	LF-QI-D	0.5759	43.33	74.27	83.68	5.87
	LF-QIH-D	0.5807	43.82	74.68	84.07	5.78
	HRE-QH-D	0.5695	42.70	73.25	82.97	6.11
	HRE-QIH-D	0.5846	44.67	74.50	84.22	5.72
	HREA-QIH-D	0.5868	44.82	74.81	84.36	5.66
	MN-QH-D	0.5849	44.03	75.26	84.49	5.68
	MN-QIH-D	**0.5965**	**45.55**	**76.22**	**85.37**	**5.46**
VQA	SAN1-QI-D	0.5764	43.44	74.26	83.72	5.88
	HieCoAtt-QI-D	0.5788	43.51	74.49	83.96	5.84

[Das et al., 2017]

Video Captioning

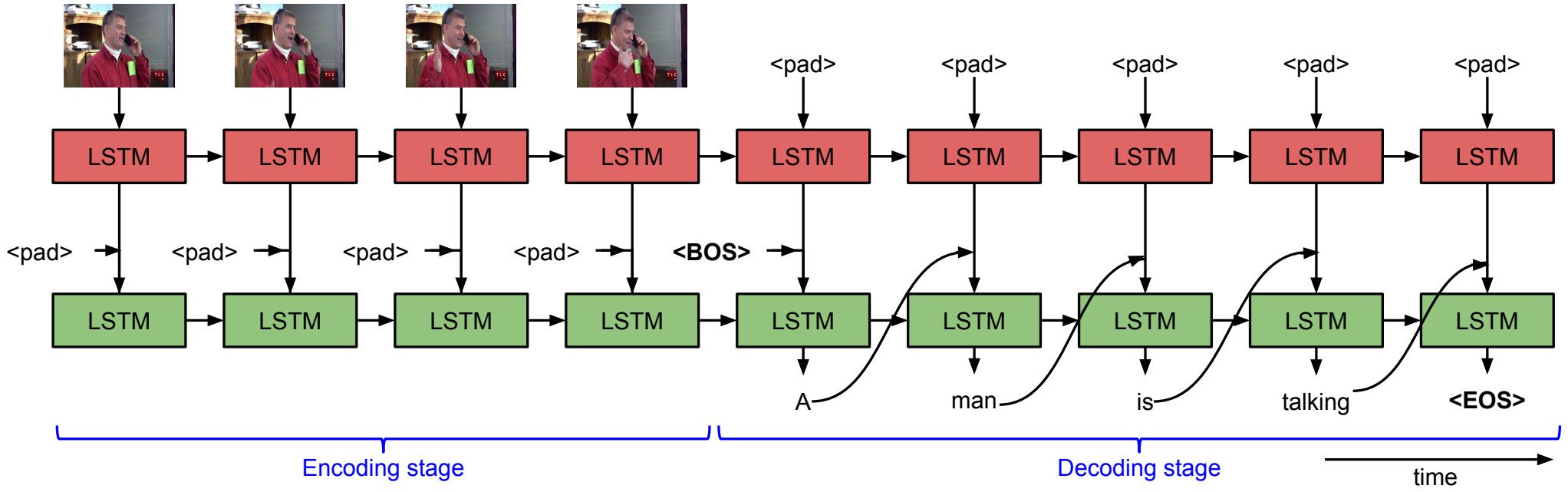
Ground truth: A woman is slicing a red pepper.

Early Video Captioning



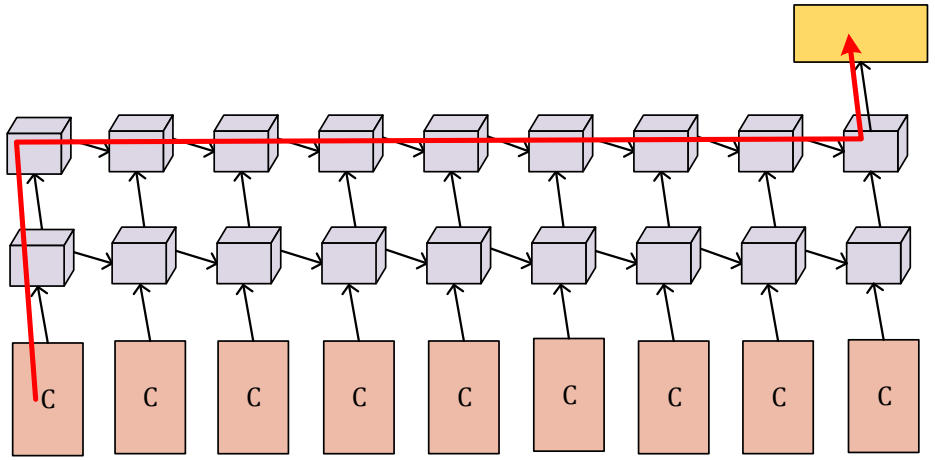
[Venugopalan et al., 2014]

S2VT

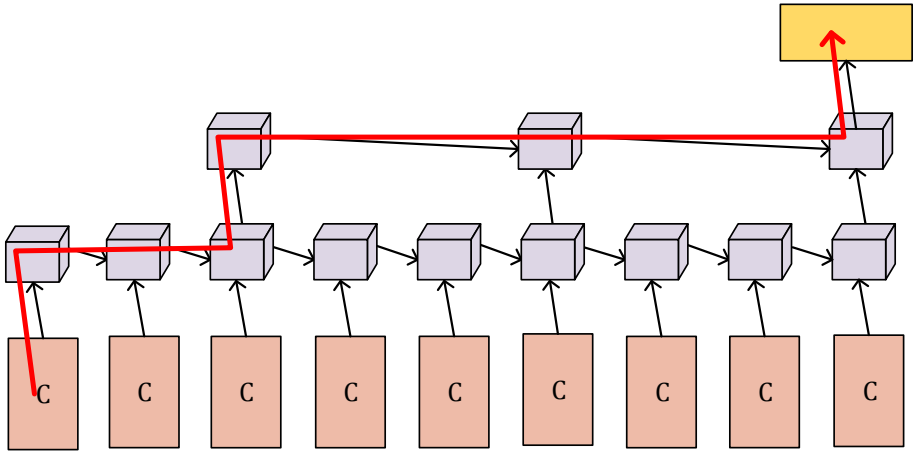


[Venugopalan et al., 2015]

Hierarchical Encoder



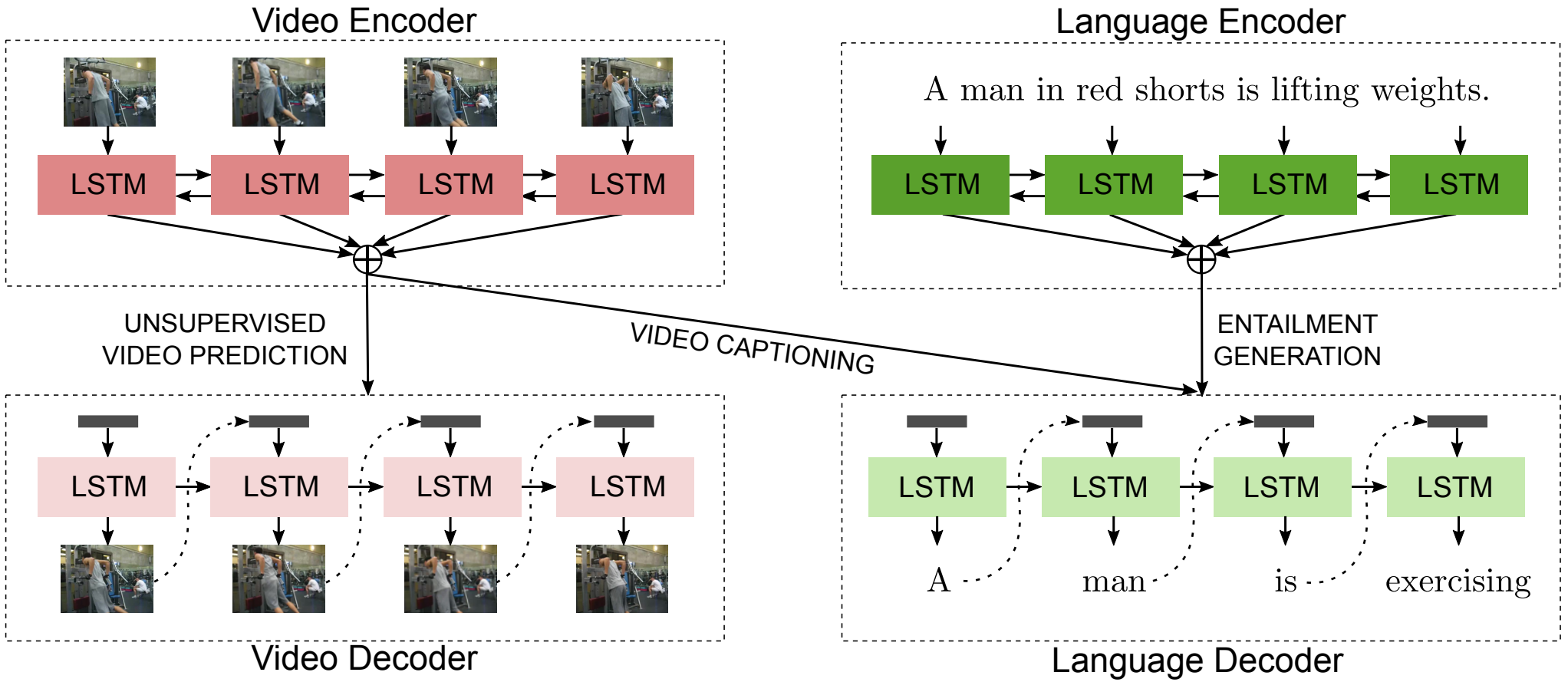
(a) Stacked LSTM video encoder



(b) Hierarchical Recurrent Neural Encoder

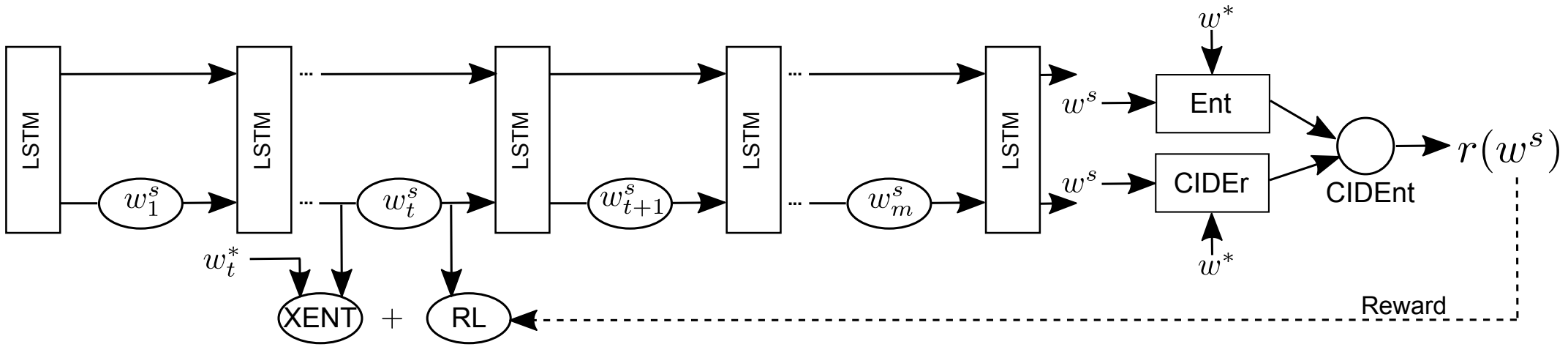
[Pan et al., 2016]

M-to-M Multi-Task for Video Captioning



[Pasunuru and Bansal, 2017a]

Reinforced Video Captioning w/ Entailment



[Pasunuru and Bansal, 2017b]

Guest Research Talks by Ramakanth Pasunuru:

1) Multi-Task Video Captioning with Video and Entailment Generation (ACL 2017)
2) Reinforced Video Captioning with Entailment Rewards (EMNLP 2017)

(45 mins)

Guest Research Talks by Hao Tan:

1) A Joint Speaker-Listener-Reinforcer Model for Referring Expressions (CVPR 2017)
2) Source-Target Inference Models for Spatial Instruction Understanding (AAAI 2018)

(45 mins)