ECCV 2018
European Conference on Computer Vision

THE UNIVERSITY OF SYDNEY

Image-Question-Answer Synergistic Network for Visual Dialog

Dalu Guo Chang Xu Dacheng Tao

UBTECH Sydney AI Centre, School of Computer Science,
FEIT, University of Sydney, Darlington, NSW 2008, Australia

Paper: https://arxiv.org/abs/1902.09774 (accepted by cvpr2019)

From VQA to Visual Dialog

– Beyond Dialog History

What color are their jackets?

VQA

- Black and grey
- Grey and blue
- Dark grey
- Ray and red
- White and grey
- Yellow and grey

Visual Dialog

- Black and grey
- 1 is green and black and 1 is purple and black
- Black
- Red green grey and black
- Grey, black and white
- Dark blue
- Blue, red, black, grey and green

➢ Candidate answers of Visual Dialog are longer than VQA

Blue, red, black, grey and green VS Black and grey

➢ Candidate answers of Visual Dialog are similar

Blue, red, black, grey and green VS Red green grey and black

Motivation

– Transfer Ranking to Matching

What color are their jackets?

- Black and grey
- Red green grey and black
- Grey, black and white
- Dark blue
- Blue, red, black, grey and green

Hard

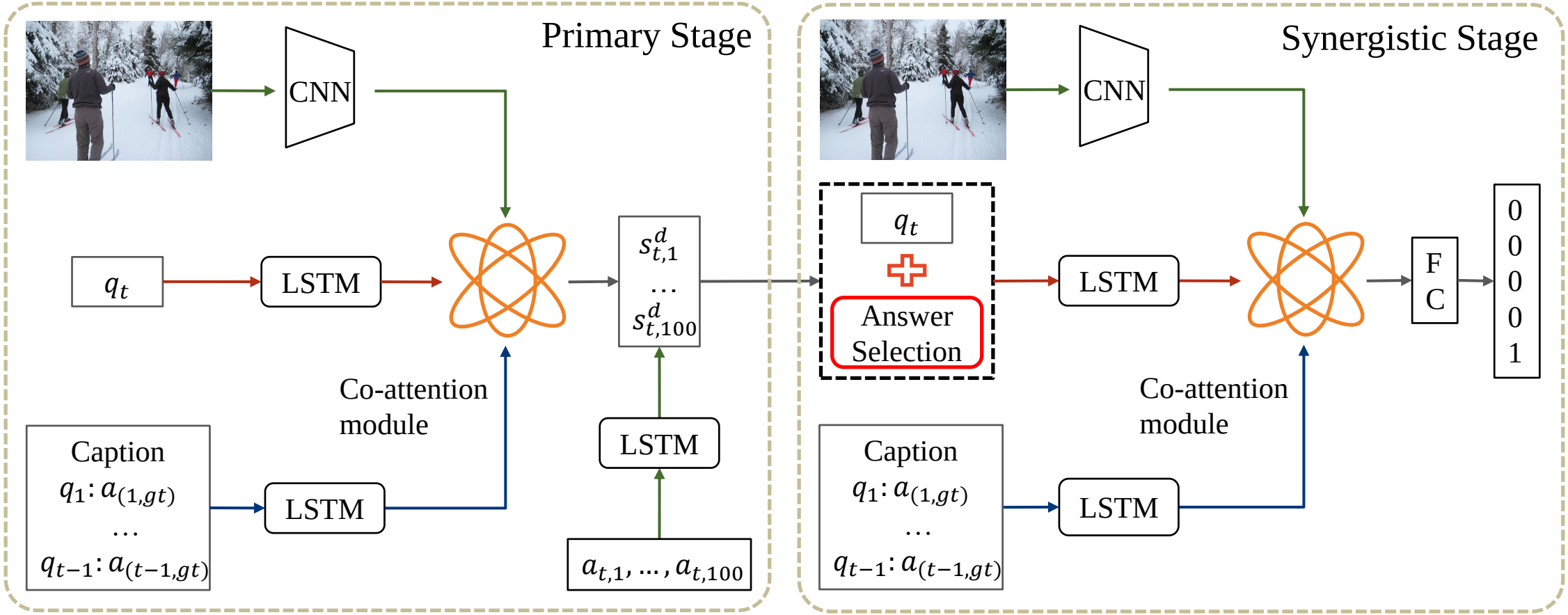
The color of their jackets are black and grey
The color of their jackets are red green grey and black
The color of their jackets are grey, black and white
The color of their jackets are dark blue
The color of their jackets are blue, red, black, grey and green

0
0
0
0
1

Easy

➢ Answers can also attend the image to collect related information for classification

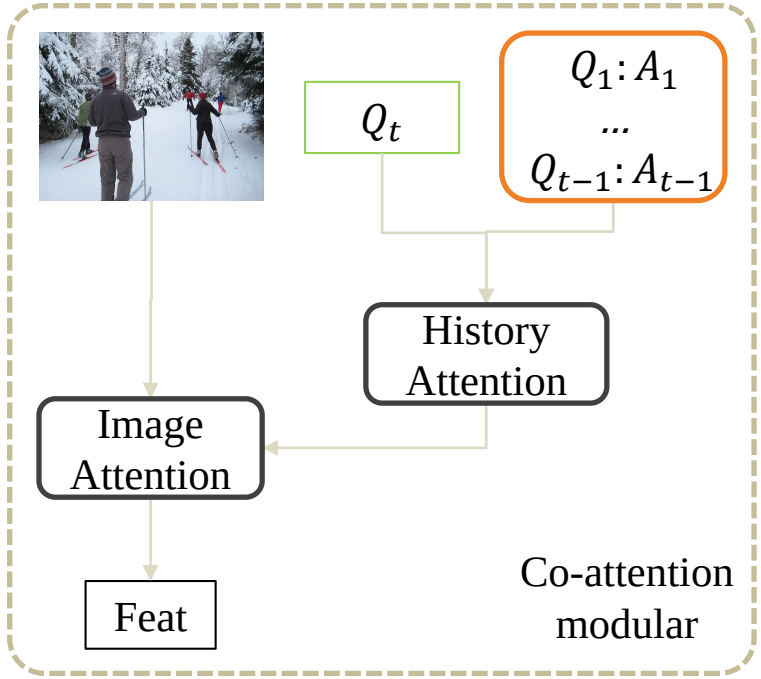
Framework

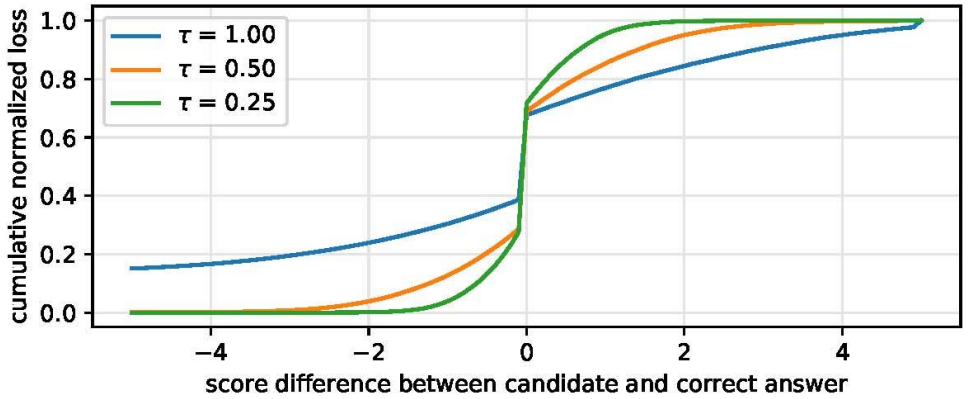


- In primary stage, the fusion of image, question and history scores the given candidate answers.
- In synergistic stage, fusion of the selected answers concatenated with question, image and history are re-ranked.

Implement Details

- Image Feature
  - Bottom-Up and Top-Down
- Attention and Fusion Method
  - Multi-modal Factorized Bilinear Pooling
    - Question and history fusion
    - Question, history and image fusion
- Discriminator Loss
  - In primary stage, N-pair loss with temperature $\tau = 0.25$
  - In synergistic stage, Softmax with cross entropy loss
- Training
  - Only primary stage for 7 epochs
  - Both primary and review stage for 17 epochs
  - Initial learning rate 1e-3, decay every 7 epoch





Experiments

Result on v1.0 test-std dataset

Model	NDCG	MRR	R@1	R@5	R@10	Mean
LF[4]	45.31	55.42	40.95	72.45	82.83	5.95
HRE[4]	45.46	54.16	39.93	70.45	81.50	6.41
MN[4]	47.50	55.49	40.98	72.30	83.30	5.92
MN-att[4]	49.58	56.90	42.43	74.00	84.35	5.59
LF-att[4]	49.76	57.07	42.08	74.83	85.05	5.41
Technion	54.46	67.25	53.40	85.28	92.70	3.55
MS AI	55.35	63.27	49.53	80.40	89.60	4.15
USTC-YTH	56.47	61.44	47.65	78.13	87.88	4.65
Single (ours)	57.32	62.20	47.90	80.43	89.95	4.17
Ensemble (ours)	57.88	63.42	49.30	80.77	90.68	3.97

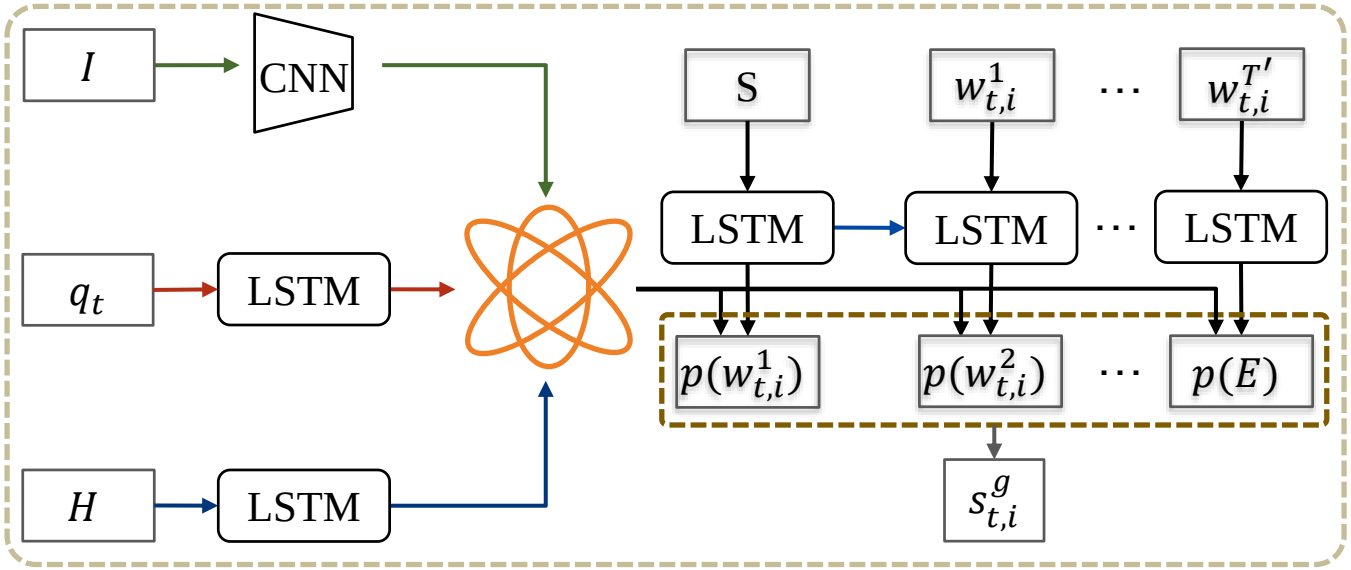
Ablation Study

Result on v1.0 validation dataset

N	M	τ	MRR	R@1	R@5	R@10	Mean
-	-	1.00	61.92	47.53	79.78	89.28	4.42
-	-	0.50	62.51	48.30	80.05	89.57	4.27
-	-	0.25	62.63	48.31	80.39	89.65	4.23
10	10	0.25	62.31	48.18	79.45	90.21	4.22
10	20	0.25	62.83	48.45	80.70	90.28	4.11
10	30	0.25	**63.54**	**49.21**	**81.01**	**90.32**	**4.09**
10	40	0.25	63.14	48.77	80.97	90.02	4.13
15	30	0.25	63.16	48.91	80.75	90.22	4.12

Extension to Generative Model

– Framework for primary stage



Extension to Generative Model

Result on v1.0 validation dataset

Model	N	M	MRR	R@1	R@5	R@10	Mean
MN-att [4]	-	-	47.94	37.48	58.56	65.57	17.61
Primary (ours)	-	-	49.01	38.54	59.82	66.94	16.69
Synergistic (ours)	10	30	51.62	40.77	63.58	67.00	16.51
Synergistic (ours)	20	30	53.23	**41.42**	67.22	72.91	15.87
Synergistic (ours)	30	30	**53.73**	41.28	**69.01**	**75.85**	**15.12**

Qualitative Results for Discriminative Model

Is there any signs around?

Primary	Synergistic
A1:No signs	A1:No, only the giraffe
A2:No signs present	A2:No I do not see any signs
A3:There are no signs	A3:No signs
A4:No	A4:No signs present
A5:No I do not see any signs	A5:There are no signs
A6:No, only the giraffe	A6:No

Is there trees and animals?

Primary	Synergistic
A1:Just rocks and grass	A1:No plants or trees
A2:No grass	A2:I see some grass
A3:No plants or trees	A3:Dried up grass or weeds, yes
A4:Dried up grass or weeds, yes	A4:Just rocks and grass
A5:Just dirt	A5:No grass
A6:I see some grass	A6:Just dirt

Is there anyone near him?

Primary	Synergistic
A1:No one else is visible	A1:Yes, there are another pair of feed
A2:No	A2:Not that I can see
A3:Nobody in the photo	A3:No one else is visible
A4:Not that I can see	A4:No
A5:Yes, there are another pair of feed	A5:Nobody in the photo

Can you tell if it's day or night?

Primary	Synergistic
A1:Day	A1:Evening
A2:Daytime	A2:Night
A3:It is day	A3:Day
A4:Night	A4:Daytime
A5:Evening	A5:It is day

What's the desk made out of?

Primary	Synergistic
A1:Wood	A1:Wood
A2:Laminate	A2:Cheap plastic or wood
A3:Dark brown	A3:Laminate
A4:Cheap plastic or wood	A4:Dark brown

Are they both wearing a hat?

Primary	Synergistic
A1:0 of the 3 guys are wearing hats	A1:Only 1 is wearing a hat
A2:Yes, the man is wearing a hat	A2:2 of them are wearing hats
A3:2 of them are wearing hats	A3:0 of the 3 guys are wearing hats
A4:Only 1 is wearing a hat	A4:Yes, the man is wearing a hat

Qualitative Results for Generative Model (without Candidate Answers)

Q:Is this outdoors?
A:No, a beach

Primary	Synergistic
A1:Yes	A1:Yes
A2:Yes, it is	A2:Yes, on a beach
A3:Yes, it is outdoor	A3:Yes, it is outdoor
A4:Yes, on a beach	A4:Yes, it is outdoors
A5:No	A5,Yes, it is
A6:Yes, in a park	A6:No

Q:What color is the jet?
A:White and blue

Primary	Synergistic
A1:White	A1:White
A2:Gray	A2:White and blue
A3:Grey	A3:White with a blue tail
A4:White and blue	A4:White and black
A5:Silver	A5:Blue and white
A6:Blue and white	A6:Grey and blue

Q:Is she wearing a hat?
A:A visor

Primary	Synergistic
A1:Visor	A1:Yes, a white visor
A2:Yes	A2:Visor, yes
A3:A visor	A3:A visor
A4:Yes, a visor	A4:Yes, a visor
A5:No, a visor	A5:Visor
A6:No	A6:No, a visor

Q:Is the church brick?
A:It looks made of stone

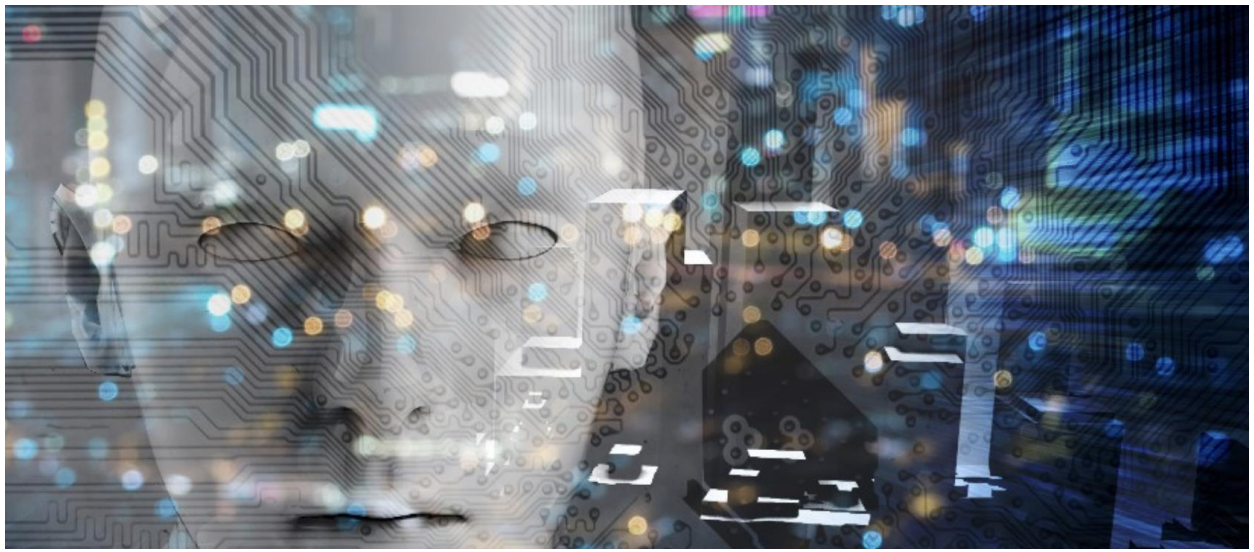
Primary	Synergistic
A1:Yes	A1:No, it looks like its made of stone
A2:Yes, it looks like it	A2:No, it looks like stone
A3:Yes, it is	A3:Yes, it appears to be brick
A4:No, it looks like stone	A4:No, it looks like concrete
A5:No	A5:Yes, it is brick
A6:I can't tell	A6:Yes

Q:Can you see any people?
A:Part of a person

Primary	Synergistic
A1:Yes	A1:Just the legs of someone
A2:Just 1	A2:Just the legs of 1 person
A3:Part of 1	A3:I can see the legs of 1 person
A4:1	A4:Just legs
A5:Yes 1	A5:Just the legs of 1
A6:Just the legs of someone	A6: Yes

Q:Are there any buildings?
A:In the background, I can see 1

Primary	Synergistic
A1:Yes, in the background	A1:Yes, there is 1 behind the truck
A2:Yes, there are	A2:Yes, I see a few
A3:Yes, in the distance	A3:Yes, there is a building behind it
A4:Yes, there 1 in the background	A4:Yes, there is 1 behind the bus
A5:Yes, there are	A5:Yes, I see 1



THE UNIVERSITY OF SYDNEY

UBTECH Sydney Artificial Intelligence Centre