Integrating Historical States and Co-attention Mechanism for Visual Dialog

Tianling Jiang , Yi Ji , Chunping Liu

School of Computer Science and Technology Soochow University

ICPR 2021

Outline

- Visual Dialog
- Framework
- Experiments
- Conclusion

What is Visual Dialog?

Given

given Image I

dialog History H

A man and a woman are holding umbrellas

What color is his umbrella?

His umbrella is black

What about hers?

Hers is multi-colored

follow-up Question Q

How many other people are in the image?

Task

give an answer A
based on I, H, Q

...

image

history

A man and a woman are holding umbrellas

What color is his umbrella?

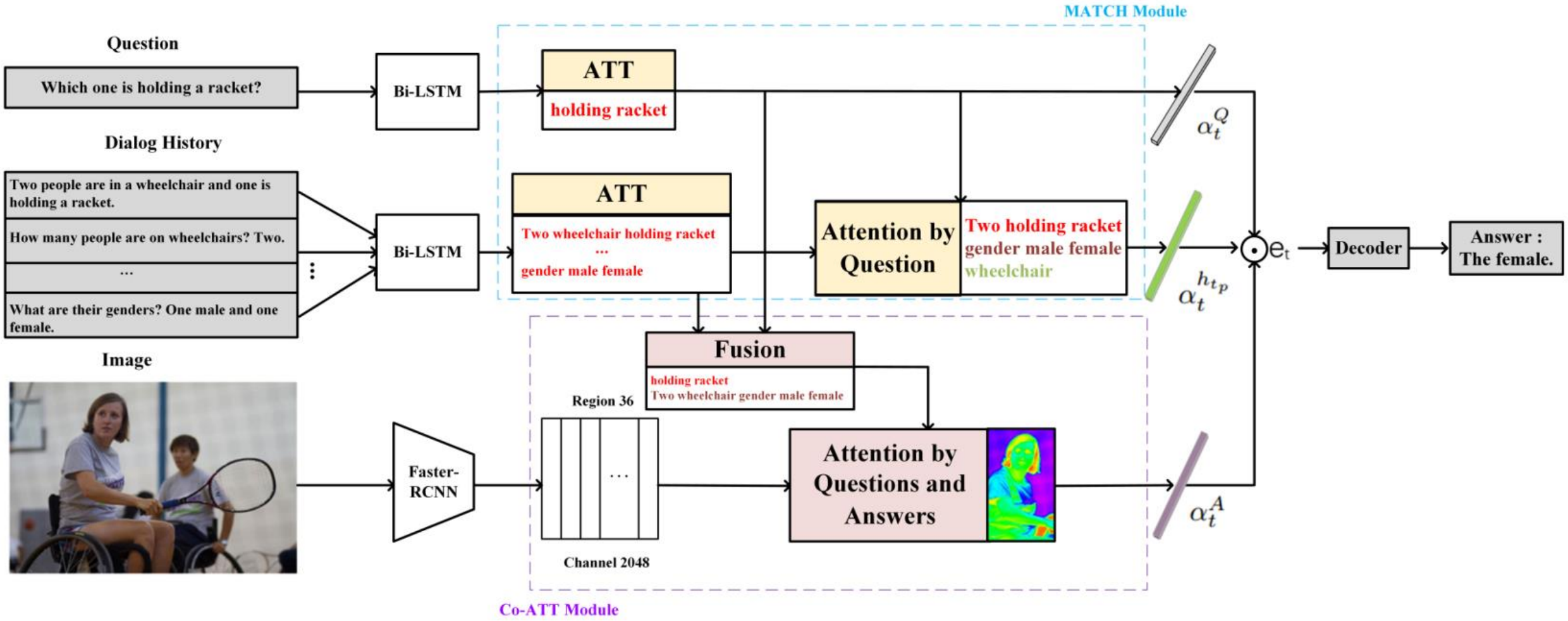
His umbrella is black

question

What about hers?

Answer: Hers is multi-colored

Framework


MATCH Module
Question
Which one is holding a racket?
Bi-LSTM
ATT
holding racket
Dialog History
Two people are in a wheelchair and one is holding a racket.
How many people are on wheelchairs? Two.
...
What are their genders? One male and one female.
Bi-LSTM
ATT
Two wheelchair holding racket
...
gender male female
Attention by Question
Two holding racket
gender male female
wheelchair
α_t^Q
$\alpha_t^{h_{tp}}$
e_t
Decoder
Answer : The female.
Image
Fusion
holding racket
Two wheelchair gender male female
Region 36
Faster-RCNN
...
Channel 2048
Attention by Questions and Answers
α_t^A
Co-ATT Module

Co-ATT:

Dialog History

c: The boat has a dragon head on the front.

q_1: Is there anything on the boat ?

a_1: A drum on it.

q_2: What color is the head of it ?

a_2: Yellow.

Fig. 3. Visualization of the Co-ATT module. The example of current question and previous Q & A pairs to jointly guide the visual information in the early stage. The objects mentioned in the textual information will be traced as early as possible in the given image. For example, as the noun "drum" is mentioned in the a_1 answer, we will also use it to guide the image. In this way, the features located in the final stage will be more comprehensive.

MATCH:

q4:What color are the towels?

t1	h0:A bathroom with a white bath tub, sink and large window.	0.02
t2	h1:What color is the bathroom? Most white.	0.05
t3	h2:Are there towels hanging? No, on the floor.	0.83
t4	h3:Are there any people in there? No.	0.03

(a)

q4:Are they standing next to each other?

t1	h0:Two zebra standing next to each other in a dry grass field.	0.76
t2	h1:Are zebras in zoo? No.	0.25
t3	h2:Are zebras eating? No, they are just standing around.	0.12
t4	h3:Any trees? Yes many trees.	0.05

(b)

q4:Is the sun shining?

t1	h0:A couple of people in the snow on skis.	0.04
t2	h1:Is the person male or female? Male.	0.02
t3	h2:What color is his hair? Brown.	0.01
t4	h3:What kinds of bags does he have? Backpack.	0.02

(c)

q4:Is it clear?

t1	h0:A sink and toilet in a small room.	0.76
t2	h1:Is this a bathroom? Yes.	0.53
t3	h2:What color is the room? It looks cream colored.	0.32
t4	h3:Is anyone in the room? Nobody.	0.05

(d)

Fig. 4. Illustration of the MATCH module. The correlation between current question and the historical information block. At each round, the weight value of the similarity between the historical block and the question will be generated. Finally, we choose the round with the largest weight value. As in figure (a), the historical information at time t_3 guided by question 4 can give us the highest information gain. In figure (b) and (d), the q_4 is more related to the information at t_1 , so the value of the information at t_1 is higher than other historical information. However, as shown in figure (c), current question is not related to the antecedent historical information, so the values are all very low, which causes the selected history block to be of little help to the given question. At this time, this MATCH module has little effect.

Ablation Studies

Model	NDCG ↑	MRR ↑	R@1 ↑	R@5 ↑	R@10 ↑	Mean ↓
B[23]	0.5559	0.6303	49.03	80.40	89.83	4.18
B+H(L)	0.5676	0.6452	50.81	81.52	90.18	4.09
B+H(E)	0.5679	0.6458	50.86	81.50	90.28	4.07
B+M	0.5675	0.6450	50.70	81.49	90.21	4.13
B+H(L)+M	0.5673	0.6451	50.93	81.41	90.10	4.04
B+H(E)+M	**0.5701**	**0.6471**	**51.09**	**81.58**	**90.36**	**4.03**

B： Baseline
H： Co-ATT
M： MATCH

Comparisons

Model	NDCG ↑	MRR ↑	R@1 ↑	R@5 ↑	R@10 ↑	Mean ↓
LF w/o RPN[15]	0.4531	0.5542	40.95	72.45	82.83	5.95
HRE[15]	0.4546	0.5416	39.93	70.45	81.50	6.41
MN[15]	0.4750	0.5549	40.98	72.30	83.30	5.92
CorefNMN[22]	0.5470	0.6150	47.55	78.10	88.80	4.40
RvA w/o RPN[23]	0.5176	0.6060	46.25	77.88	87.83	4.65
HACAN w/o RPN[26]	0.5281	0.6174	47.91	78.59	87.81	4.63
LF[15]	0.5163	0.6041	46.18	77.80	87.30	4.75
RvA[23]	0.5559	0.6303	49.03	80.40	89.83	4.18
Sync[24]	0.5732	0.6220	47.90	80.43	89.95	4.17
DAN[25]	**0.5759**	0.6320	49.63	79.75	89.35	4.30
HACAN[26]	0.5717	0.6422	50.88	80.63	89.45	4.20
Ours	0.5701	**0.6471**	**51.09**	**81.58**	**90.36**	**4.03**

Conclusion

- Co-ATT module takes into account the joint guidance of questions and answers in the dialog history
- MATCH module resolves ambiguous references in the current question by retrieving the history block
- HSCA achieves the new state-of-the-art performance

THANKS!