Integrating Historical States and Co-attention Mechanism for Visual Dialog

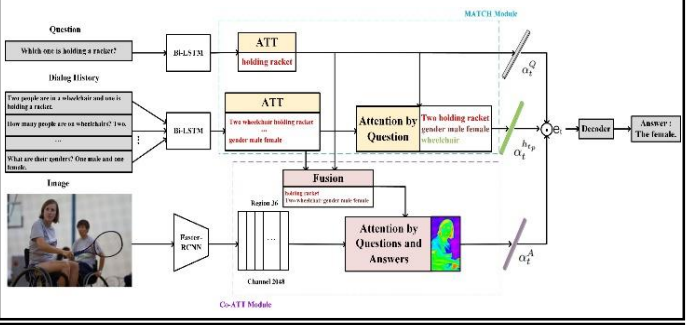
Tianling Jiang ∗ , Yi Ji † , Chunping Liu ‡
School of Computer Science and Technology Soochow University
tljiang@stu.suda.edu.cn,{jiyi, cpliu} @suda.edu.cn

Introduction

Visual dialog is a multi-modal mission, which needs to make a comprehensive understanding and reasonable reasoning based on both visual and textual content. The core issue to be resolved in the visual dialog is visual reference resolution.

Main contributions:

1.The Co-ATT module to locate objects precisely in the given image.
2.The MATCH module retrieves historical content more accurately.
3.we achieve state-of-the-arts on most metrics of VisDial v1.0 compared to other methods.



Experiments

Ablation Studies

In our ablation studies, we give a detailed comparison for the roles of various modules in our proposed model.

B+H(L): refers we integrate the history information in the late phase.
B+H(E):processes the image together with the given question and dialog history in the early stage.
B+M: utilizes similarity calculations to retrieve the most relevant content so that we get the maximum information gain while removing irrelevant information.
B+H(L)+M: is the combination of two modules, but we add the history condition in the later during the training procedure.
B+H(E)+M: is the complete combined architecture we finally proposed.

Table || ABLATION ANALYSIS OF OUR PROPOSED MODELS

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**

Compared Methods

We compare the proposed novel model with the other approaches on VisDial v1.0 dataset, which can be categorized into three groups.

· Fusion-based models: LF[15] and HRE[15]
· Attention-based models: MN[15] and Sync(Guo et al.)[24]
· Visual reference resolution models: CorefNMN(Kottur et al.)[22], RvA(Niu et al.)[23], DAN(Kang et al.)[25] and HACAN(Yang et al.)[26].

Table I RETRIEVAL PERFORMANCE OF DISCRIMINATIVE MODELS ON THE TEST-STANDARD SPLIT OF VISDIAL V 1.0.

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**

Approach

Co-ATT

The Co-ATT module is designed to early combine the textual and visual features in the model.

In Algorithm 1, we firstly embed the question features, history features and visual features into the same space. Then, we fuse the question and the whole dialog history features. Finally, we make use of the textual features to locate more objects in the image.

As illustrated in Figure 3, we extract the textual features (boat and head) from c (the image caption at h_0), then we make use of the textual features to locate objects in the image. At q_1 & a_1 , the noun "drum" has appeared in a_1 , so we can also use the textual feature to guide the image. Thus the object "drum" shown in the image has a high score.

Algorithm 1 Co-ATT Module

1 : **function** Co-ATT :
2 : $\alpha_t^{q'} \leftarrow f_q^A(\alpha_t^q)$
3 : $\alpha^{H'} \leftarrow f_h^A(\alpha_{t_p}^H)$ ($t_p = 0, \cdots, t\text{-}1$)
4 : **for** i← 1,···,K **do**
5 : $e_i^v \leftarrow f_v^A(v_i)$
6 : $z_{t,i}^A \leftarrow L_2\text{Norm}(\alpha_t^{q'} \circ \alpha^{H'} \circ e_i^v)$
7 : **end for**
8 : $\alpha_t^A \leftarrow \text{softmax}(W^A Z_t^A)$
9 : **return** α_t^A
10: **end function**

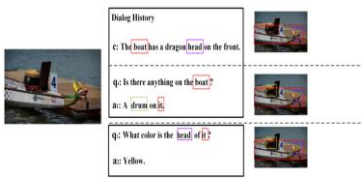


Fig. 3. Visualization of the Co-ATT module.

MATCH

The MATCH module aims to find the history block most relevant to current ambiguous question. In this module, we retrieve the history block based on the given question instead of using the whole history feature. By this operation, we retain more relevant features and filter out many unrelated features.

In Algorithm 1, we firstly takes the attended question features and the attended history features as inputs. Then locates which historical round is most pertinent to current question. Finally, we get the core historical content to retain more relevant features and filter out many unrelated features.

As in figure 4 (a), the historical information at time t_3 guided by question 4 can give us the highest information gain.

Algorithm 2 MATCH Module

1 : **function** MATCH(H, q, t)
2 : $\alpha_t^{q''} \leftarrow f_q^A(\alpha_t^q)$
3 : **for** t_p← 0 ,···,t-1 **do**
4 : $\alpha_t^{h_{t_p}} \leftarrow f_h^A(\alpha_{t_p}^H)$
5 : $z_{t,t_p}^P \leftarrow \text{MLP}([\alpha_t^{q''}, \alpha_t^{h_{t_p}}])$
6 : $\triangle_{t,t_p} \leftarrow t\text{-}t_p$
7 : **end for**
8 : $o_t^P \leftarrow \text{GS\_Sampler}(W^P[Z_t^P, \triangle_t])$
9 : $t_p' \leftarrow \sum_{t_p} o_{t,t_p}^P \cdot t_p$
10: $z_t^{h_{t_p}} \leftarrow L_2\text{Norm}(\alpha_t^{q''} \circ \alpha_t^{h_{t_p'}})$
11: $\alpha_t^{h_{t_p}} \leftarrow \text{softmax}(W^H z_t^{h_{t_p}})$
12: **return** $\alpha_t^{h_{t_p}}$
13: **end function**

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)

Fig. 4. Illustration of the MATCH module.

Conclusion

In this paper, we propose Integrating Historical States and Co-attention (HSCA) for visual reference resolution in visual dialog task. As compared with the previous methods, our Co-ATT module takes into account the importance of joint guidance of questions and answers in the dialog history in the early stage. Then, the MATCH module resolves ambiguous references in the current question by retrieving the history block. On the benchmark real-world dataset VisDial v1.0, HSCA achieves the new state-of-the-art performance. It demonstrates our model is more grounded and effective.