Dual Visual Attention Network for Visual Dialog

Dan Guo, Hui Wang and Meng Wang
School of Computer Science and Information Engineering,
Hefei University of Technology
guodan@hfut.edu.cn, wanghui.hfut@gmail.edu.cn, eric.mengwang@gmail.com

Introduction

Visual dialog is a challenging task, which involves multi-round semantic transformations between vision and language. This paper aims to address cross-modal semantic correlation for visual dialog. Motivated by that V_g (global vision), V_l (local vision), Q (question) and H (history) have inseparable relevance, the paper proposes a novel Dual Visual Attention Network (DVAN) to realize (V_g, V_l, Q, H) $\Rightarrow$ A. DVAN is a three-stage query-adaptive attention model. In order to acquire accurate A (answer), it first explores the textual attention, which imposes the question on history to pick out related context H'. Then, based on Q and H', it implements respective visual attentions to discover related global image visual hints V'_g and local object-based visual hints V'_l. Next, a dual crossing visual attention is proposed. V'_g and V'_l are mutually embedded to learn the complementary of visual semantics. Finally, the attended textual and visual features are combined to infer the answer. Experimental results on the VisDial v0.9 and v1.0 datasets validate the effectiveness of the proposed approach.

Methods

Framework

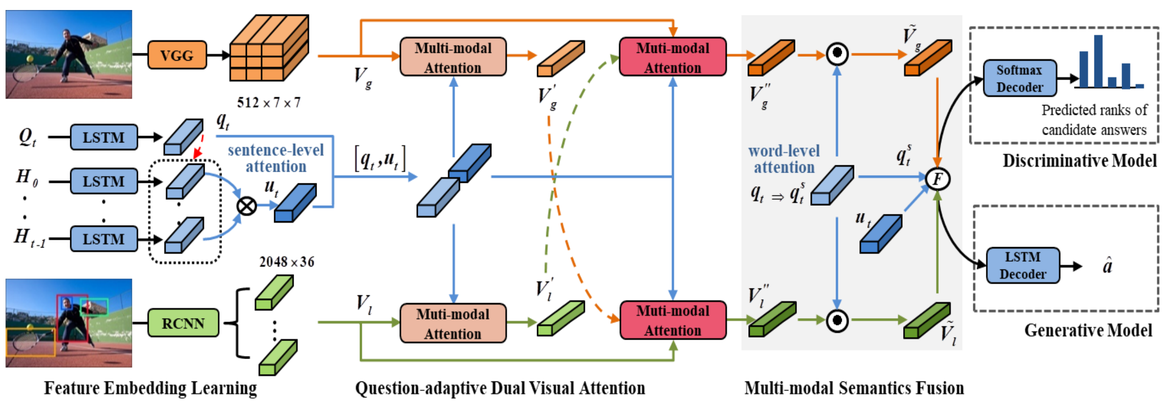

Figure 1: The overall framework of Dual Visual Attention Network (DVAN) for visual dialog.

DVAN consists of three modules:

(1) Feature Embedding Learning. To better learn the correlation among multi-modal features, we embed each feature to the same feature dimension.

(2) Question-adaptive Dual Visual Attention. A multi-stage attention mechanism is proposed to progressively enhance the question-adaptive cues on history, global and local visions. It first tackles the textual co-reference problem. Then, the visual semantic is explored by independent visual reasoning and dual crossing visual reasoning. Learning the intra- and inter-visual relation is beneficial in performing visual grounding in a more accurate way.

(3) Multi-modal Semantics Fusion, in this module, the attended textual and visual features are fused to infer the answer.

Independent Visual Attention

To gradually implement visual reasoning, this reasoning step focuses on discovering the independent semantic responses on visual contents. The question q_t and attended history u_t are jointly used to learn the intra-visual relation in both types of visions. Taking global visual features as an example, the whole calculation process is as follows:

$$z_t^{g1} = tanh\left(W_{q1}q_t + W_{h1}u_t + W_{g1}V_g\right)$$

$$\alpha_t^{g1} = softmax\left(P_{g1}^T z_t^{g1}\right)$$

$$V'_g = \alpha_t^{g1} \cdot V_g^T$$

where V_g is the original global visual feature, and V'_g captures the intra-visual relation in its own feature space.

Dual Crossing Visual Attention

As global and local visions are different types of features, the next reasoning step aims to explore the inter-visual relation between them, which helps to make their semantics more consistent and also achieves visual complementarity, the whole calculation process is as follows:

$$z_t^{g2} = tanh\left(W_{q2}q_t + W_{h2}u_t + W_{l2}V'_l + W'_{g2}V_g\right)$$

$$\alpha_t^{g2} = softmax\left(P_{g2}^T z_t^{g2}\right)$$

$$V''_g = \alpha_t^{g2} \cdot V_g^T$$

where V''_g learns the correlation between global and local visions, and the intra-relation aware visual semantics complement each other.

Experiments

Our methods are evaluated on VisDial v0.9 & v1.0 datasets. The experimental performance is evaluated by retrieving the ground-truth answer from a list of 100 candidate answers.

On VisDial v0.9 dataset, we evaluate **DVAN** in both generative and discriminative models. As shown in table1, our basic model **DVAN w/o OF**, which only consider global visions, can already outperform most of the compared methods. Then, by integrating global and local visions, **DVAN** has significantly improved the results compared with the state-of-the-art methods.

On VisDial v1.0 dataset, compared with the previous methods, **DAVN** still achieves the best performance.

Model	Generative Models					Discriminative Models				
	MRR	R@1	R@5	R@10	Mean	MRR	R@1	R@5	R@10	Mean
LF [Das *et al.*, 2017]	0.5199	41.83	61.78	67.59	17.07	0.5807	43.82	74.68	84.07	5.78
HRE [Das *et al.*, 2017]	0.5237	42.29	62.18	67.92	17.07	0.5846	44.67	74.50	84.22	5.72
HREA [Das *et al.*, 2017]	0.5242	42.28	62.33	68.17	16.79	0.5868	44.82	74.81	84.36	5.66
MN [Das *et al.*, 2017]	0.5259	42.29	62.85	68.88	17.06	0.5965	45.55	76.22	85.37	5.46
HCIAE [Lu *et al.*, 2017a]	0.5386	44.06	63.55	69.24	16.01	0.6222	48.48	78.75	87.59	4.81
AMEM [Seo *et al.*, 2017]	-	-	-	-	-	0.6227	48.53	78.66	87.43	4.86
CoAtt [Wu *et al.*, 2018]	0.5411	44.32	63.82	69.75	16.47	0.6398	50.29	80.71	88.81	4.47
CorefNMN [Kottur *et al.*, 2018]	-	-	-	-	-	0.6410	50.92	80.18	88.81	4.45
DVAN w/o OF	0.5443	44.58	64.30	70.11	15.27	0.6381	50.09	80.58	89.03	4.38
DVAN w/o GF	0.5538	46.01	65.06	70.68	15.11	0.6522	51.86	81.64	89.96	4.22
DVAN w/o Att3	0.5573	46.32	65.28	70.92	14.91	0.6601	52.78	82.22	90.21	4.09
DVAN w/o SQ	0.5579	46.40	65.33	71.02	14.95	0.6604	52.83	82.41	90.37	4.03
DVAN	**0.5594**	**46.58**	**65.50**	**71.25**	**14.79**	**0.6667**	**53.62**	**82.85**	**90.72**	**3.93**

Table 1: Retrieval performance of both generative and discriminative models on VisDial v0.9.

Model	MRR	R@1	R@5	R@10	Mean	NDCG
LF	0.5542	40.95	72.45	82.83	5.95	45.31
HRE	0.5416	39.93	70.45	81.50	6.41	45.46
MN	0.5549	40.98	72.30	83.30	5.92	47.50
CorefNMN‡	0.6150	47.55	78.10	88.80	4.40	**54.70**
DVAN	**0.6258**	**48.90**	**79.35**	**89.03**	**4.36**	**54.70**

Table 2: Retrieval performance of discriminative models on VisDial v1.0 test-std.

Qualitative Results

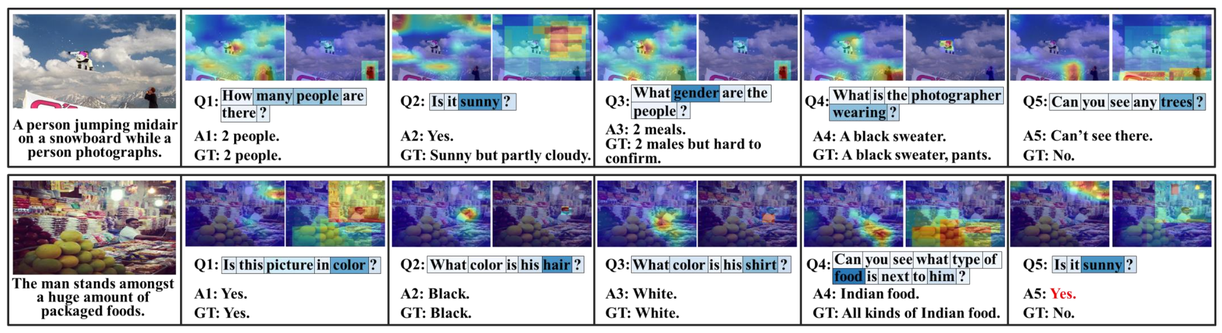

Figure 2: Visualization results on VisDial v0.9. In the blank box at each round, the first column is the global attention map, and the second column is the local object-based attention map.

Visualization results validate the effectiveness of our proposed model. There are some conclusions:

(1) The textual attention distribution at word-level relative to each round answering is interpretable, which captures the keywords of each question.

(2) We discuss the two types of visual attention maps, i.e., global and local visual attention maps. In each round dialogue, if both two maps focus on similar regions, the model has high confidence in visual reasoning; otherwise, their visual cues can complement each other for better visual grounding.

The proposed DVAN effectively represents multi-modal entities and infers their rich semantic relationships, which is beneficial in constructing a good visual dialogue system.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (NSFC) under grants 61725203, 61732008, and 61876058.