Dual Attention Network for Visual Question Answering

Huijuan Xu and Kate Saenko

Computer Science,
Boston University
{hxu,saenko}@bu.edu

Abstract. Visual Question Answering (VQA) is a popular research problem that involves inferring answers to natural language questions about a given visual scene. Recent neural network approaches to VQA use attention to select relevant image features based on the question. In this paper, we propose a novel Dual Attention Network (DAN) that not only attends to image features, but also to question features. The selected linguistic and visual features are combined by a recurrent model to infer the final answer. We experiment with different question representations and do several ablation studies to evaluate the model on the challenging VQA dataset.

Keywords: Visual Question Answering, Attention, Deep Learning

1 Introduction

Visual Question Answering (VQA) is an interdisciplinary research problem, which has the potential applications for the visual impaired and the automatic text querying of large image collections or surveillance videos.

The attention concept has been applied for solving the image caption problem, visual question answering, text question answering, and so on. There are two general attention architectures. One is the LSTM architecture with each sentence/question word as input and attending on the candidate features at each time step [1, 2]. The other is using the whole sentence/question embedding to attend on the candidate features from the beginning, similar to the memory network style [3–6].

Several attention models [4–6] have been proposed for the VQA problem, but these models only attend on image features. This paper explores the attention on both image and question features for the VQA problem. Four different question representations (word embedding, LSTM, Bidirectional LSTM, question convolutional neural network) are experimented in the proposed Dual Attention Network. We also conduct two ablation studies to help us understand the attention models for VQA problem.

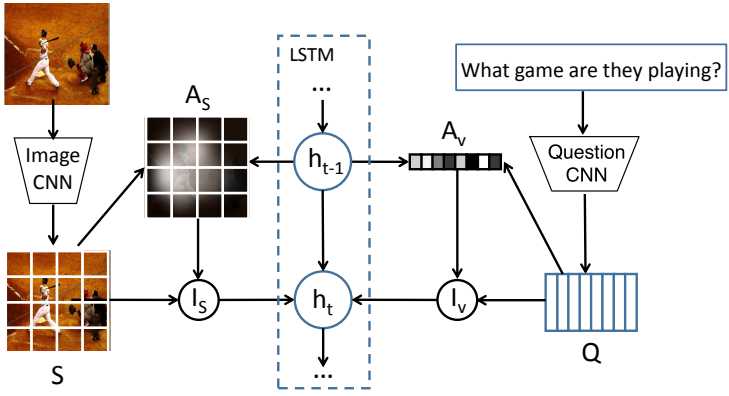

Fig. 1. Our Dual Attention Network (DAN) model applies the same hidden state of the LSTM to attend on both the question features and image convolutional features. The question features are denoted as "Question CNN" while other question features can also be used.

2 Related Work

In one of the early works [7] about visual question answering, VQA is proposed as a Turing test proxy. Their approach is based on handcrafted features using a semantic parse of the question and scene analysis of the image and combined in a latent-world Bayesian framework. Later papers [8, 9] directly adapt the image captioning models to solve the VQA problem by generating the answer using a recurrent LSTM network conditioned on the image CNN output and each question word. These end-to-end deep neural networks learn features directly from data and achieve improved results compared to earlier hand-crafted approaches.

Recently, [10] proposes a large scale VQA data set with two tasks, the Open-Ended task and the Multiple-Choice task. Several recent papers report results on this dataset. The non-attention based models include iBOWIMG [11], ACK [12] and DPPnet [13]. For the attention based models, the D-NMN [14] uses attention operation in the composable modules. SAN [4], SMem-VQA [5] and DMN [6] are similar to memory network [3] in principle, which all use the collected visual evidences and question features to predict the final answer.

3 Dual Attention Network

An overview of the proposed Dual Attention Network is illustrated in Fig. 1. The input to our network is an image of fixed size and a question comprised of a variable-length sequence of words. The model uses a recurrent neural network LSTM as the state update mechanism. The hidden state of the LSTM is used to attend to both the question features and image features.

3.1 Image Attention

The input image is processed by a convolutional neural network (CNN) to extract high-level M-dimensional visual features on a grid of spatial locations. Specifically, we use $S = \{s_i | s_i \in \mathbb{R}^M; i = 1, \cdots, L\}$ to represent the spatial CNN features at each of the L grid locations. In this paper, the outputs of GoogLeNet (inception 5b/output) [15] are used as the image features. The LSTM hidden state h_t is used to attend to the visual features S, as follows:

$$F_s = W_s S + b_s \tag{1}$$

$$h_s = W_{hs} h_{t-1} + b_{hs} \tag{2}$$

$$A_s = softmax(W_{as}(h_s \oplus F_s)) \tag{3}$$

$$I_s = A_s S \tag{4}$$

where F_s is the embedding of image features S, h_s is the embedding of previous LSTM hidden state h_{t-1}, A_s are the attention weights, and I_s are the attended image features that are passed to the next time step of LSTM. $\oplus$ is an operator that adds the vector h_s to the columns of the matrix F_s.

3.2 Question Attention

Each word in the question is first represented as a one-hot vector of the size of the vocabulary, with a value of one only in the corresponding word position and zeros in the other positions. Each one-hot vector is then embedded into a real-valued word vector, $V = \{v_j | v_j \in \mathbb{R}^N; j = 1, \cdots, T\}$, where T is the maximum number of words in the question and N is the dimensionality of the embedding space. Sentences with length less than T are padded with special -1 value, which are embedded to all-zero word vectors.

We explore four kinds of question features, namely direct word embedding, LSTM hidden state, Bidirectional LSTM hidden state and question convolutional neural network (question CNN). The word embedding of each question word does not contain the overlap context information, while the other three features contain certain nearby context information. We use Q to represent the question features. The word embedding for each word in the question is a learned real-valued word vector V. In the LSTM, each word in the question is fed into the LSTM at each time step in sequence. The hidden state of each time step is used for the question features Q. In Bidirectional LSTM, one additional LSTM is used to encode the question words in reverse order. The forward and backward hidden states at each time step are fused to form the hidden state for the Bidirectional LSTM, which is the question features Q. For the question CNN, we modify from a model proposed in [16]. This model is a 3-layer CNN architecture which uses 256/256/128 filters in the first/second/third layer. The question convolutional features Q represents the output of the question CNN applied on word vectors V.

The question side attention is listed as follows:

$$F_v = W_v Q + b_v \tag{5}$$
$$A_v = softmax(W_{av}(h_s \oplus F_v)) \tag{6}$$
$$I_v = A_v Q \tag{7}$$

where F_v is the embedding of question features Q, A_v are the attention weights, and I_v are the attended question features that are fed into the next time step of the LSTM. Here, the model uses the same embedding h_s of the LSTM hidden state to attend on the question features Q as in the image attention layer.

3.3 Recurrent Dual Attention Model

Our Dual Attention Network combines the question and image attention layers described above. At each time step of the LSTM, the same embedding of the hidden state h_{t-1} is used to generate the attended image features I_s and attended question features I_v. These are then used as inputs to the LSTM to update the LSTM hidden state in the next time step. We modify the LSTM units to include the attended features from both the image and question side. The modified LSTM equations are as follows:

$$f_t = \sigma(W_{fs} I_s + W_{fv} I_v + W_f h_{t-1} + b_f) \tag{8}$$
$$i_t = \sigma(W_{is} I_s + W_{iv} I_v + W_i h_{t-1} + b_i) \tag{9}$$
$$o_t = \sigma(W_{os} I_s + W_{ov} I_v + W_o h_{t-1} + b_o) \tag{10}$$
$$\widetilde{c_t} = \phi(W_{cs} I_s + W_{cv} I_v + W_c h_{t-1} + b_c) \tag{11}$$
$$c_t = i_t \widetilde{c_t} + f_t c_{t-1} \tag{12}$$
$$h_t = o_t \phi(c_t) \tag{13}$$

The hidden state from the last time step of the fusion LSTM is used to predict the answer and connect to the cross entropy loss layer. Following the tradition of LSTM attention model, no attention is done at the first time step since the hidden state at the first time step is randomly initialized, and the model just inputs the average of each location's image features and the average of each word vector in the question at the first time step.

4 Experiments

4.1 Question Features

The learning curves of the four question features on sampled validation batches are shown in Fig. 2 (0-1 accuracy). We observe that among these four question features, direct word embedding performs the worst. LSTM is significantly better than word embedding, Bidirectional LSTM is slightly better than LSTM by about 2%, and the question CNN exceeds the Bidirectional LSTM by about 5%. Based on these results, we select the question CNN features for our model in the following experiments.

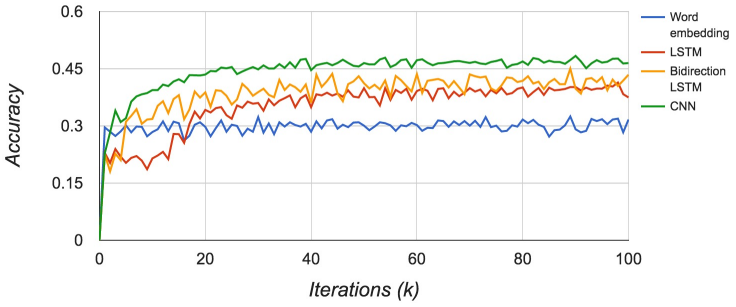

Fig. 2. Learning curves of four different question features in DAN model.

4.2 Ablation Studies

We design two ablation studies to help us understand the DAN model as well as other attention models for solving VQA problem, and present the results in Fig. 3. The two ablation studies are as follows: 1) showing each word in sequence on the question side of the fusion LSTM (First Ablation Model); and 2) removing the question side attention from the second time step and only attending on image convolutional features (Without Question Attention).

The model in the first ablation study is similar to the image captioning attention model [1] and the grounded question answering model Visual7W [2] with image patches as multiple choice answers. If we directly use the model in this ablation study to predict the answer for the VQA problem, the accuracy is about 10% lower compared to our DAN model in Fig. 3. The ablation model accepts each question word in sequence and attends on the convolutional image features at each time step. So the ablation model uses incomplete question information to attend on visual features at each time step, and the model will not see the whole question until the last time step of the fusion LSTM where it predicts the answer. This may be the reason for the low performance of the ablation model on the VQA problem. This ablation study indicates that allowing the model to access whole question information from the beginning is important for solving the VQA problem.

The second ablation study takes away the whole question-side attention and only keeps the image-side attention, and the LSTM hidden state in the last time step is used to predict the final answer. Fig. 3 shows that the sampled batch validation accuracy becomes extremely poor for this ablation model (Without Question Attention). We compare this ablation model to an attention based VQA model SAN [4] to investigate the reason of the poor performance. In SAN model, the final answer is predicted using the addition of the collected visual evidence and one whole question representation. In this ablation model, the hidden state in the last time step of the LSTM contains the collected visual evidence, but the whole question representation component is missed for directly predicting the final answer, compared to SAN. So in these memory network like models

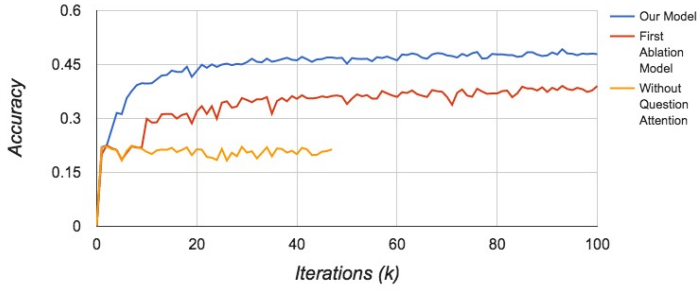

Fig. 3. Ablation studies of DAN model. "Without Question Attention" is early stopped due to the poor accuracy.

	test-dev				test-standard			
methods	**Overall**	yes/no	number	others	**Overall**	yes/no	number	others
iBOWIMG [11]	55.7	76.6	35.0	42.6	55.9	76.8	35.0	42.6
DPPnet [13]	57.2	80.7	37.2	41.7	57.4	80.3	36.9	42.2
ACK [12]	59.2	81.0	38.4	45.2	59.4	81.2	37.1	45.8
deeper LSTM Q + Norm I [10]	57.8	80.5	36.8	43.1	58.2	80.6	36.5	43.7
NMN [17]	54.8	77.7	37.2	39.3	55.1	-	-	-
D-NMN [14]	57.9	80.5	37.4	43.1	58.0	-	-	-
SAN [4]	58.7	79.3	36.6	46.1	58.9	-	-	-
DMN [6]	60.3	80.5	36.8	48.3	60.4	80.4	36.8	48.3
MRN [18]	61.7	82.3	38.8	49.3	61.8	82.4	38.2	49.4
[19]	61.8	79.7	38.7	51.7	62.1	-	-	-
[20]	63.3	81.9	39.0	53.0	63.2	81.7	38.2	52.8
MCB [21]	64.7	82.5	37.6	55.6	-	-	-	-
DAN (ours)	60.1	81.1	35.3	47.7	-	-	-	-

Table 1. Test-dev and Test-standard results on the Open-Ended VQA dataset (in percentage).

for solving VQA problem [4–6], incorporating one direct question representation path for the final answer prediction is very important. In our Dual Attention Network, the attended question is a kind of question representation with focus on certain parts of the question. The attended question features and attended image features are fused to update the LSTM hidden state which contains the collected evidence, and the LSTM hidden state at the last time step which is used to predict the final answer has direct access to one whole question representation.

4.3 Comparison to other Models

We evaluate our proposed Dual Attention Network on the large scale data set for visual question answering, VQA full release (V1.0) [10]. The evaluation results of the Open-Ended task are shown in Tab. 1. The Open-Ended server evaluation uses the evaluation metric introduced by [10], which gives partial credit to certain synonym answers: $Acc(ans) = \min\{(\#\text{ humans that said } ans)/3, 1\}$. The image features take ResNet-50 [22] and the question features take question CNN.

5 Conclusion

In this paper, we explore the idea of attending on both the image and question for the VQA problem. Different question features are experimented and two ablation studies are designed to help understand the VQA models. We also get competitive results on the Open-Ended task of VQA.

References

1. Xu, K., Ba, J., Kiros, R., Courville, A., Salakhutdinov, R., Zemel, R., Bengio, Y.: Show, attend and tell: Neural image caption generation with visual attention. arXiv preprint arXiv:1502.03044 (2015)
2. Zhu, Y., Groth, O., Bernstein, M., Fei-Fei, L.: Visual7w: Grounded question answering in images. arXiv preprint arXiv:1511.03416 (2015)
3. Sukhbaatar, S., Weston, J., Fergus, R., et al.: End-to-end memory networks. In: Advances in Neural Information Processing Systems. (2015) 2431–2439
4. Yang, Z., He, X., Gao, J., Deng, L., Smola, A.: Stacked attention networks for image question answering. arXiv preprint arXiv:1511.02274 (2015)
5. Xu, H., Saenko, K.: Ask, attend and answer: Exploring question-guided spatial attention for visual question answering. arXiv preprint arXiv:1511.05234 (2015)
6. Xiong, C., Merity, S., Socher, R.: Dynamic memory networks for visual and textual question answering. arXiv preprint arXiv:1603.01417 (2016)
7. Malinowski, M., Fritz, M.: A multi-world approach to question answering about real-world scenes based on uncertain input. In: Advances in Neural Information Processing Systems. (2014) 1682–1690
8. Malinowski, M., Rohrbach, M., Fritz, M.: Ask your neurons: A neural-based approach to answering questions about images. In: Proceedings of the IEEE International Conference on Computer Vision. (2015) 1–9
9. Ren, M., Kiros, R., Zemel, R.: Exploring models and data for image question answering. In: Advances in Neural Information Processing Systems. (2015) 2935–2943
10. Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Lawrence Zitnick, C., Parikh, D.: Vqa: Visual question answering. In: Proceedings of the IEEE International Conference on Computer Vision. (2015) 2425–2433
11. Zhou, B., Tian, Y., Sukhbaatar, S., Szlam, A., Fergus, R.: Simple baseline for visual question answering. arXiv preprint arXiv:1512.02167 (2015)
12. Wu, Q., Wang, P., Shen, C., Hengel, A.v.d., Dick, A.: Ask me anything: Free-form visual question answering based on knowledge from external sources. arXiv preprint arXiv:1511.06973 (2015)
13. Noh, H., Seo, P.H., Han, B.: Image question answering using convolutional neural network with dynamic parameter prediction. arXiv preprint arXiv:1511.05756 (2015)
14. Andreas, J., Rohrbach, M., Darrell, T., Klein, D.: Learning to compose neural networks for question answering. arXiv preprint arXiv:1601.01705 (2016)
15. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2015) 1–9
16. Ma, L., Lu, Z., Li, H.: Learning to answer questions from image using convolutional neural network. arXiv preprint arXiv:1506.00333 (2015)

17. Andreas, J., Rohrbach, M., Darrell, T., Klein, D.: Deep compositional question answering with neural module networks. arXiv preprint arXiv:1511.02799 (2015)
18. Kim, J.H., Lee, S.W., Kwak, D.H., Heo, M.O., Kim, J., Ha, J.W., Zhang, B.T.: Multimodal residual learning for visual qa. arXiv preprint arXiv:1606.01455 (2016)
19. Lu, J., Yang, J., Batra, D., Parikh, D.: Hierarchical question-image co-attention for visual question answering. arXiv preprint arXiv:1606.00061 (2016)
20. Noh, H., Han, B.: Training recurrent answering units with joint loss minimization for vqa. arXiv preprint arXiv:1606.03647 (2016)
21. Fukui, A., Park, D.H., Yang, D., Rohrbach, A., Darrell, T., Rohrbach, M.: Multimodal compact bilinear pooling for visual question answering and visual grounding. arXiv preprint arXiv:1606.01847 (2016)
22. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. arXiv preprint arXiv:1512.03385 (2015)