Knowing Where to Look? Analysis on Attention of Visual Question Answering System

Wei Li[1,2], Zehuan Yuan[2(✉)], Xiangzhong Fang[1], and Changhu Wang[2]

[1] Shanghai Jiao Tong University, Shanghai, China
{liweihfyz,xzfang}@sjtu.edu.cn
[2] ByteDance AI Lab, Beijing, China
{levi.li,yuanzehuan,wangchanghu}@bytedance.com

Abstract. Attention mechanisms have been widely used in Visual Question Answering (VQA) solutions due to their capacity to model deep cross-domain interactions. Analyzing attention maps offers us a perspective to find out limitations of current VQA systems and an opportunity to further improve them. In this paper, we select two state-of-the-art VQA approaches with attention mechanisms to study their robustness and disadvantages by visualizing and analyzing their estimated attention maps. We find that both methods are sensitive to features, and simultaneously, they perform badly for counting and multi-object related questions. We believe that the findings and analytical method will help researchers identify crucial challenges on the way to improve their own VQA systems.

Keywords: Attention · Visual question answering

1 Introduction

Visual question answering (VQA) attracts increasing attentions in both computer vision and natural language processing community. The goal of VQA is to answer questions based on the information of any given image. As deep learning witnessed a series of remarkable success in artificial intelligence, VQA also made tremendous progress [1,6,15] over past few years such as several benchmark datasets, e.g., VQA 2.0 [2], CLEVR [4] and Visual Genome [7], and tons of approaches, e.g., MFB [15] and BAN [5].

VQA is usually formulated as a classification task with different answers as candidate categories. The current mainstream pipeline is to firstly extract image and question representations with Convolutional Neural Network and Recurrent Neural Network, respectively. Then, a lot of fusion methods such as early fusion [18] and bilinear pooling [1,5,6,15] are adopted to combine two-stream features.

In addition, attention is playing an increasingly important role as the mechanism encourages deep cross-domain interactions without introducing substantial parameters. There are two main branches to add attention to VQA system: *uni-attention* and *co-attention*. *Uni-attention* merely considers question-guided visual attentions. In contrast, *co-attention* additionally takes image-guided question attentions into account to jointly model the multimodal correlations [5,9,10].

Although much progress has been made, few works lie on deep analysis on the influence of different attention mechanisms. In this paper, we dive into two state-of-the-art methods: multi-model factorized pooling (MFB) [15] and bilinear attention network (BAN) [5] to discover their inherent limitations. Both methods adopt the popular bilinear pooling to perform multimodal fusion. However, MFB only performs question-guided visual attention (*uni-attention*) while BAN extends *co-attention* into bilinear attention to enable more image and language interactions. We conduct all our experiments on VQA 2.0 dataset with a more balanced answer distribution than VQA 1.0 [16] and Visual Genome dataset. In addition, it covers more relations of real-world objects compared with CLEVR dataset full of synthetic images. In order to make a deeper understanding of both methods, we propose to directly delve into their attention maps. Observing whether estimated attentions relate to real answers could reflect the robustness and limitations of corresponding approaches.

To summarize, we present three key observations after thorough experiments on both approaches:

– The performance is sensitive to selected features. Representations based on object proposals are better than image-level features.
– Attention distribution becomes much more inaccurate for questions related to multiple objects.
– Counting problem is not well solved by soft attention mechanism.

In terms of each observation, we also analyze main reasons behind these phenomenons and claim that similar limitations probably exist in most of methods with attention mechanisms. We believe that these findings will inspire researchers to design more effective methods. Furthermore, our analytical method is hopeful to offer researchers an opportunity to identify potential roadblocks when debugging their VQA systems.

2   Multimodal Factorized Bilinear Pooling Revisited

Since bilinear pooling [12] allows abundant multimodal cross-channel interactions, the fusion method has been widely used in VQA systems compared to simple summation and concatenation operators. To further reduce the number of parameters in bilinear pooling, multimodal factorized bilinear Pooling (MFB) [15] decomposes the weight matrix as two low-rank matrices.

Specifically, given a question vector $x \in \mathbb{R}^m$ and an image feature vector $y \in \mathbb{R}^n$, each output channel of MFB pooling is formulated as:

$$\text{pool}(x,y)_i = x^T \mathbf{W}_i y + b_i = x^T \mathbf{U}_i \mathbf{V}_i^T y + b_i = \mathbb{I}(\mathbf{U}_i^T x \circ \mathbf{V}_i^T y) + b_i \qquad (1)$$

where $\mathbb{I} \in \mathbb{R}^k$ is a vector of all elements ones, $\mathbf{W}_i \in \mathbb{R}^{m \times n}$ is the weight matrix and $\mathbf{U}_i \in \mathbb{R}^{m \times k}$ and $V_i \in \mathbb{R}^{n \times k}$ are two factorized matrices.

The whole pipeline of MFB for VQA can be summarized as follows. First, an overall question representation $\hat{x} \in \mathbb{R}^m$ is obtained by a *self-attention* manner with weights α^x. Then, the weighted question feature guilds the visual attention on the image as follows:

$$\alpha^y = softmax(\{\mathbf{W}_p^T \mathrm{pool}(\hat{x}, y_j)\}), \hat{y} = \sum_j \alpha_j^y y_j \tag{2}$$

where y_j is an image feature vector and $\mathbf{W}_p \in \mathbb{R}^{m \times 1}$. Finally, attention weighted language feature $\hat{x}$ and visual feature $\hat{y}$ are fused together as $f = \mathrm{pool}(\hat{x}, \hat{y})$ for further prediction.

3   Bilinear Attention Revisited

Co-attention based model jointly integrates question-guided visual attention and visual-guided question attention together. To further consider every pair of multimodal features, BAN [5] extends co-attention into bilinear attention. The fused feature can be defineds as:

$$f_i = (\mathbf{X}^T \tilde{\mathbf{U}})_i^T A (\mathbf{Y}^T \tilde{\mathbf{V}})_i \tag{3}$$

where $\tilde{\mathbf{U}} \in \mathbb{R}^{m \times k}$, $\tilde{\mathbf{V}} \in \mathbb{R}^{n \times k}$, $\mathbf{X} \in \mathbb{R}^{m \times \theta}$, $\mathbf{Y} \in \mathbb{R}^{n \times \gamma}$, and $A \in \mathbb{R}^{\theta \times \gamma}$ is the bilinear attention map that sums to 1 as follows:

$$A = softmax((\mathbb{I} \cdot p^T) \circ X^T \mathbf{U}) \mathbf{V}^T Y) \tag{4}$$

where $\mathbb{I} \in \mathbb{R}^\theta$ is a vector with all elements ones, $p \in \mathbb{R}^k$, and $softmax$ is applied element-wisely. Then the fused feature f can be used for further classification.

MFB and BAN represent popular attempts in *uni-attention* and *co-attention* directions, respectively. A thorough analysis for both methods is also expected to shed light on similar limitations of other approaches with attention mechanisms.

4   Deep Study

In this section, we will present detailed analysis for our key observation results. As shown above, we investigate MFB [15] and BAN [5] to make a thorough study. All experiments are conducted on VQA2.0 benchmark, where we train on *train* split with 82,783 images and 443,757 questions, and evaluate on *val* split with 40,503 images and 214,354 questions totally. Each question is annotated with 10 answers by crowdsourcing. In order to give an intuitive demonstration, we report visualizations of image attention vectors α^y in MFB and the bilinear attention maps A in BAN.
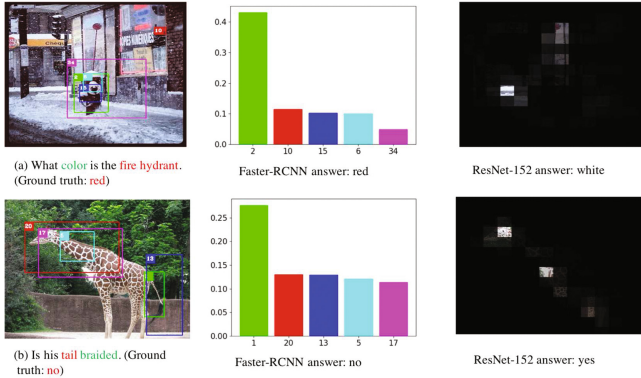
Fig. 1. Visualization of MFB with different visual features. From left to right are the original images, the MFB attention weights of Faster-RCNN proposals and the MFB attention map of the ResNet-152 feature map. The most salient boxes (numbered in the top-left corner of each bounding box and x-axis of the grids) are visualized in both images.

4.1   Object Feature and Image Feature

Visual object features have been proven effective in VQA task [5,13] compared with image-level features. However, the reason behind the performance gain has not been well investigated. In this work, we delve deeper into this from the attention perspective.

In our experiments, we select top-36 Faster-RCNN proposals [11] and ResNet-152 last feature map before *pool5* [3] as object features ($36 \times 2,048$) and image features ($196 \times 2,048$), respectively. We set the batch size to 64 and the dimension of hidden states to 1024 in BAN. To simplify experiments, we do not integrate counting module [17]. Unlike the original implementation, we augment 300-dimensional random initialized word embedding instead of 300-dimensional computed word embedding to each 300-dimensional Glove word embedding. The performance comparison on the VQA 2.0 validation set is shown in Table 1. Unsurprisingly, we achieve better performance with object features for both methods compared with image-level features. In addition, we found that a more accurate attention distribution can be obtained for object features compared with image features. For example in Fig. 1, given a question about fire hydrant, we can see that MFB with object proposals focuses on the correct entity while image-level representation directs attentions to snow regions. Due to the inaccurate attention distribution, the model with image features predicts a wrong answer, *white*. Similarly when "**Is his tail braided?**" is asked, the tail proposal is highlighted for the method with object-level representations as opposed to arbitrary emphasis with a single feature map.

Although it is difficult to measure the negative effect of features quantitatively on attention maps over the entire dataset, we hypothesize that inaccurate attention maps take a large amount of responsibility for decline in performance.

Table 1. Detailed performance comparison on VQA 2.0 validation set

Feature type	Methods	Overall	Other	Number	Yes/No
ResNet-152 feature map	MFB [15]	60.94	52.93	38.48	79.28
	BAN [5]	59.52	51.19	38.92	77.64
Faster-RCNN proposals	MFB [15]	65.19	57.17	44.37	82.98
	BAN [5]	64.3	55.7	45.45	82.16

We analyse that object proposals have much more specific semantic meanings compared with feature maps and thus the corresponding relations between words and visual features are easier to learn, which leads to a more accurate attention distribution and further performance boost.

4.2   Single Object and Multiple Objects

Based on how many objects are necessary to infer final answers, questions in VQA2.0 can be roughly divided into single object, e.g., "what is the color of the dog?" and multiple objects, e.g.,"what color is the book on the desk?". In our experiment, we conduct the comparison for both kinds of questions. The observation shows that the attention distribution is much more inaccurate for questions related to multiple objects. For example in Fig. 2, both models incorrectly focus on the laptop used by the woman in (a), which implies that the relation between the woman and the laptop are not well captured and modeled. Additionally, relative positions are not well integrated by both models. We can see in Fig. 2, both models make predictions (*white* and *yellow*) based on the person on the left and the person on the middle respectively in (b). In a word, the estimated attention maps cannot learn relative positions. Moreover, spatial locations are crucial to infer the *what* question in (c). Both models concentrate on the wrong objects in other positions, e.g., sink and toilet.

It is worth noting that current attention mechanisms learn attention distributions by only comparing visual and question representations and object features ignore their own locations in images.

However, without well-captured object relations or position information, models are unable to set these visually or semantically similar objects apart when the questions are related to multiple objects or multiple instances exist in an image. The confusion causes an inaccurate attention distribution which leads to a significant accuracy drop between single-object questions and those with multiple objects, which constitutes the main hurdle for current VQA systems.

In order to reduce the performance gap, it could be a crucial step to explicitly consider object relation and position. In particular, graph-based neural networks might be an effective way to handle unstructured object correlations [8,14]. Object relations modeling is still an open question and worth further explorations.
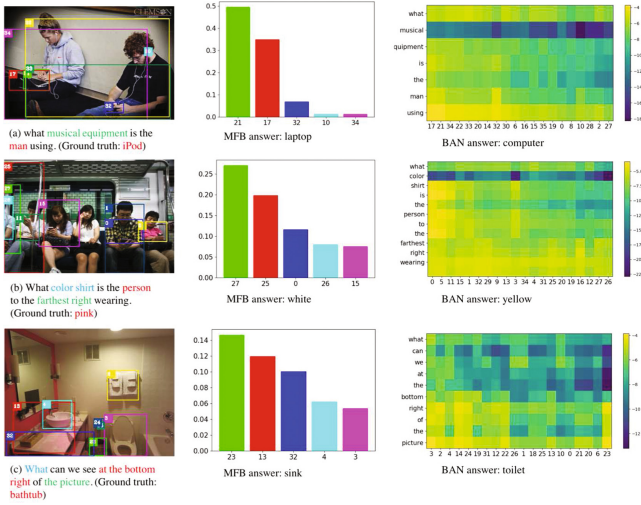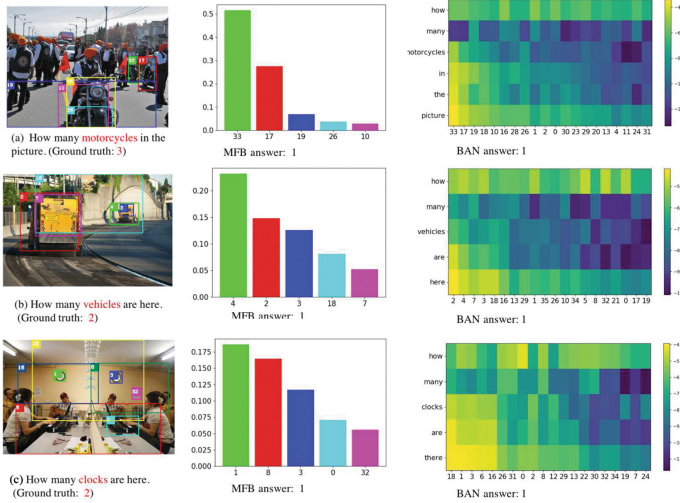
Fig. 2. Visualization of MFB and BAN on questions related to multiple objects. From left to right are the original images, MFB attention vectors and BAN bilinear attention maps. The most salient boxes (numbered in the top-left corner of each bounding box and x-axis of the grids) are visualized in both images. (Color figure online)

4.3   Counting Problem

Counting problem is a special case of questions related to multiple objects. As mentioned in [17], due to that soft-attention mechanism normalizes the attention weights, which leads to the loss of counting-related information. Soft attention is replaced by the gate strategy in [17] and then overlapping object proposals are processed in a differentiable manner.

In this work, we show that poor results can also be obtained even with an accurate attention distribution. For example in Fig. 3, both models focus their attention on multiple detected objects, namely, motorcycles in (a), vehicles in (b) and clocks in (c). However, detected objects are obviously visually similar and thus the weighted average of these visual features is probably similar to one of them, which means cues for counting are lost during soft attention process regardless of attention distributions. The limitations probably exist in a large amount of VQA systems. Therefore, in order to improve the counting performance essentially, additional structures or more flexible attention mechanisms might be needed.

Fig. 3. Visualization of MFB and BAN on counting problems. From left to right are the original images, MFB attention vectors and BAN bilinear attention maps. The most salient boxes (numbered in the top-left corner of each bounding box and x-axis of the grids) are visualized in both images. Both models give the wrong answer, 1.

5   Conclusions

To facilitate further research on the VQA task, we delve into two state-of-the-art methods MFB [15] and BAN [5] on VQA 2.0 dataset by visualizing and analysing their estimated attention maps. We form three main observations. Firstly, the performance improvement with Faster-RCNN proposals is probably related to a more accurate attention distribution. Second, the attention distribution is much more inaccurate for questions related to multiple objects. Finally, counting problem is not well solved by soft attention mechanism due to the attention weight normalization. We believe that these observation results can help future VQA research and analysing attention maps will also assist researchers to debug their own VQA systems.

References

1. Fukui, A., Park, D.H., Yang, D., Rohrbach, A., Darrell, T., Rohrbach, M.: Multimodal compact bilinear pooling for visual question answering and visual grounding. In: Conference on Empirical Methods in Natural Language Processing (EMNLP) (2016)
2. Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., Parikh, D.: Making the V in VQA matter: elevating the role of image understanding in visual question answering. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2017)

3. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
4. Johnson, J., Hariharan, B., van der Maaten, L., Fei-Fei, L., Zitnick, C.L., Girshick, R.: CLEVR: a diagnostic dataset for compositional language and elementary visual reasoning. arXiv preprint arXiv:1612.06890 (2016)
5. Kim, J.H., Jun, J., Zhang, B.T.: Bilinear attention networks. arXiv preprint arXiv:1805.07932 (2018)
6. Kim, J.H., On, K.W., Lim, W., Kim, J., Ha, J.W., Zhang, B.T.: Hadamard product for low-rank bilinear pooling. In: International Conference on Learning Representations (ICLR) (2017)
7. Krishna, R., et al.: Visual genome: connecting language and vision using crowd-sourced dense image annotations. Int. J. Comput. Vis. (IJCV) **123**(1), 32–73 (2017)
8. Liu, Y., Wang, R., Shan, S., Chen, X.: Structure inference net: object detection using scene-level context and instance-level relationships. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2018)
9. Lu, J., Yang, J., Batra, D., Parikh, D.: Hierarchical question-image co-attention for visual question answering. arXiv preprint arXiv:1606.00061 (2016)
10. Nguyen, D.K., Okatani, T.: Improved fusion of visual and language representations by dense symmetric co-attention for visual question answering. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2018)
11. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: towards real-time object detection with region proposal networks. In: Advances in Neural Information Processing Systems (NIPS) (2015)
12. Tenenbaum, J.B., Freeman, W.T.: Separating style and content. In: Advances in Neural Information Processing Systems (NIPS) (1997)
13. Teney, D., Anderson, P., He, X., van den Hengel, A.: Tips and tricks for visual question answering: learnings from the 2017 challenge. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2018)
14. Xu, D., Zhu, Y., Choy, C., Fei-Fei, L.: Scene graph generation by iterative message passing. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
15. Yu, Z., Yu, J., Xiang, C., Fan, J., Tao, D.: Beyond bilinear: generalized multimodal factorized high-order pooling for visual question answering. IEEE Trans. Neural Netw. Learn. Syst. **99**, 1–13 (2018)
16. Zhang, P., Goyal, Y., Summers-Stay, D., Batra, D., Parikh, D.: Yin and Yang: balancing and answering binary visual questions. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
17. Zhang, Y., Hare, J., Prügel-Bennett, A.: Learning to count objects in natural images for visual question answering. In: International Conference on Learning Representations (ICLR) (2018)
18. Zhou, B., Tian, Y., Sukhbaatar, S., Szlam, A., Fergus, R.: Simple baseline for visual question answering. arXiv preprint arXiv:1512.02167 (2015)