Co-attention Spatial Reasoning for Visual Question Answering

Ye Qin*, Zhiping Zhou, Chen Biao, Wenjie Li
Department of Computer science and technology, University of Nanchang Hangkong, Faculty of Information Engineering, room E518, Nanchang, China
*Email: 1299511673@qq.com

Abstract. Visual question answering (VQA) requires both inferring and providing correct answers based on the semantics of the question and the content of the image. We propose a co-attention spatial reasoning model, it models the co-attention, and spatial reasoning attention for the VQA task. The model maps the question text information to image region to reason the fusion features, which contain fine-grained semantic information of two models. Meanwhile, it can obtain the spatial relationship between question words and image regions. The model was trained and verified on the VQA-v2 dataset, the best single model achieves an accuracy rate of 69.55% on the test-dev set.

1. Introduction
Multi-modal learning across vision and language has attracted wide interest in the field of deep learning, such as image-text matching [1], image captioning [2], visual dialog [3-5] and visual question answering (VQA) [6-9] have yielded significant progress. The input of the VQA system is an image and free form natural language questions, and the output is a natural language answer of the question.

Up to now, many VQA studies have used attention mechanisms [10-12]. There are two attention mechanisms: top-down and bottom-up. The former focuses on the parts that are closely related to the task, the latter emphasizes the concept of part and focus attention on a specific region of the image or a valuable word in question, which is more consistent with the requirements of the VQA research.

Based on the bottom-up attention mechanism [13], we propose a co-attention spatial reasoning model to handle the VQA task, it can improve the representation ability of image and question features and the reason for the spatial relationships of multiple models. Experiments on the VQA-v2 dataset verify the effectiveness of the model.

2. Related Work

2.1. Visual Question Answering
Early methods used global features to solve the VQA task [6,14]. Firstly, feature extractors are used to obtain the global feature of two models, and then simple methods are used for feature fusion. Some methods adopt the LSTM to learn better question representation [7], or a more complex multi-modal
feature fusion approach [15]. The global feature concluded background and other irrelevant noise information. The attention mechanism can be helpful to focus on the specific regions or words related to the task [16], increasing the accuracy of the model. Ref [17] introduced a stacked attention network to simulated a multi-step reasoning process. Reference [18] proposed the "where to look (WTL)" attention model used text query to find the position of the image. References [19], [20], and [21][22] used different multimodal bilinear pooling methods to fuse images and question features to predict the attention.

2.2. Co-attention Models
The attention mechanism limited to the image may make the model unable to capture the key information related to answer prediction. Therefore, the text attention mechanism is integrated to consider the importance of words in the question, which helps the model to select important words in the question and eliminate some interference information. Reference [23] method modeled the self-attention of the question and the question-condition attention of the image. Reference [24] exploited a multi-level attention network model, which combines image attention and text attention into an end-to-end framework to complete the VQA task. Reference [25] proposed a Deep Modular Co-Attention Networks to reason the dense multi-modal interactions.

3. Co-attention Spatial Reasoning Mechanism

![Figure 1. Overall flowchart of co-attention spatial reasoning mechanism.](image)

3.1. Attention Mechanism
The attention mechanism is used in both images and questions to obtain important feature both of image and words. The self-attention model of the question and the guided attention model of the image adopt the transform model mechanism, through multi-head attention with h parallel ‘head’ to calculate the attended features [26]. The self-attention of the image exploited the channel attention mechanism to enhance the expression ability of image features [27].

3.2. Co-Attention Spatial Reasoning Networks

![Diagram](image)
Figure 2. Overall flowchart of the co-attention spatial reasoning networks.

The input of the spatial reasoning network is the output of two attention models and a fusion feature of them. Using bilinear transformation to fuse the multi-modal features.

Firstly, to reduce the computation time of the model, the two inputs are transformed into \( \tilde{I}_{\text{att}} \in \mathbb{R}^{K \times d_0} \) and \( \tilde{O}_{\text{aur}} \in \mathbb{R}^{K \times d_0} \) respectively. The specific implementation process is as follows:

- The multi-modal fusion features are obtained by matrix multiplication.
  \[
  J = \tilde{O}_{\text{aur}} \times \tilde{O}_{\text{aur}}^{T}
  \]

- Using softmax to encoding J to get the spatial attention probability distribution of each region.

- To enhance the reasoning of the model on the spatial relationship between the image and the question, each component \( (A \in \mathbb{R}^{K \times 1}, 1 \leq i \leq t) \) of A is fused with the attention feature according to equation (3) to obtain the corresponding spatial correlation feature \( f_t \in \mathbb{R}^{d_t} \). Where \( U \in \mathbb{R}^{K \times d}, V \in \mathbb{R}^{d_t \times d} \) are the model parameter, \( \odot \) indicates dot-product.
  \[
  f_t = (I_{\text{att}} \odot U) A (Q_{\text{aur}}^{T} \odot V)
  \]

- After obtaining spatial attention features \( f_t \), referring to the multi-modal residual proposed by [18], multiple spatial feature components are obtained spatial reasoning features.
  \[
  f = \sum_t f_t
  \]

- The spatial reasoning features \( f \) are through nonlinear layers to predict to answers, then the prediction results \( p \in \mathbb{R}^N \) are obtained. According to the prediction output of the network, binary-cross entropy with Logits (BCEWithLogits) is selected as the loss function to optimize the classifier.

4. Experiments

4.1. Datasets

VQAv2 dataset: VQAv2 is the most used public large-scale visual VQA dataset [7]. It’s improved based on VQAv1 and greatly reduces the language bias in VQAv1. It comprised of question-answer pairs relating to the images from the MS-COCO dataset, with 3 questions per image and 10 answers per question. The dataset is split into three: train (80k images and 444k QA pairs); Val (40 images and 214k QA pairs); and test (80k images and 448k QA pairs). VQAv2 was used to train and evaluate the visual question answering task model in the experiment. We use tools provided by Antol et al to evaluate the performance of our model.

4.2. Implementation Details

The dimension of image features is \( d_t = 2048 \), the number of regions per image is fixed at \( K=100 \). We use the weight of Glove-300d [28] as the pre-training weight of the word embedding, select the feature of question text extracted by LSTM. The question dimension \( d_q \) and the hidden layer dimension are 1024,1024. In the spatial reasoning model, a total of \( t \) bilinear attention maps are generated. The strategy proposed by [13] is used to set the number of \( N \) candidate answers. In the process of training, the Adam method is used to optimize the network model. The parameters are set to \( \beta = (0.9, 0.98) \), the base learning rate is set to 0.001, batch_size is set to 20, all the models are trained up to 20 epochs.

4.3. Ablation Analysis

We conduct many ablations to test model performance. Firstly, we experiments are run on different models to investigate the effect of each model.
As can be seen from table 1, the SRA with the self-attention model which based on the “Bottom-up and top-down” attention model benefits accuracy. The accuracy of Yes/No, number, other, and all four types of answer is improved by 1.73%, 1.91%, 1.7%, and 1.62% respectively. The results show that the self-attention module plays an important role in improving the accuracy of the model, which makes the module capture the details of different features. Combined with guided-attention, a co-attention model (SGA-SRA) is constructed, which enables the model to obtain the interactive information between multi-modal, understand the relationship between images and questions, all of this is conducive to the process of reasonable answers.

From table 2, we can see that the accuracy rate of the SGR+SRA model is higher than some current methods of VQA. Although the overall accuracy rate has not been improved much, it has increased by 1.03%, 2.23% respectively in the Yes/No, and number question types, indicating that the co-attention spatial reasoning model can effectively learn the corresponding relationship between multi-modal. It also means that our model has a good expression ability for VQA task.

| Y/N (%) | Num (%) | Other (%) | All (%) |
|--------|---------|-----------|---------|
| SRA    | 81.82   | 44.21     | 56.05   | 65.32   |
| SRA+SA | 83.63   | 45.01     |         | 64.8    |
| SGA+SRA| 83.71   | 45.40     | 57.74   | 65.63   |

Table 2. Contrast with the state-of-the-arts on the test-dev set of the VQA2.0 dataset. The best results are bolded.

| Y/N (%) | Num (%) | Other (%) | All (%) |
|---------|---------|-----------|---------|
| Bottom-Up and Top-Down | 81.82 | 44.21 | 56.05% | 65.32% |
| MFH      | 84.27   | 49.56     | 56.89   | 68.76   |
| BAN      | 85.31   | 50.93     | 60.26   | 69.52   |
| SGA+SRA(OURS) | **86.34** | **53.16** | 60.08 | **69.55** |

4.4. Result and Analysis
In figure 3, we visualize part of the model prediction. According to the prediction results, we can find that the number of targets in the top images are simple, our network can well predict the correct answer. However, in the bottom pictures, the target regions, and background information overlap, or the target is small, the performance of the model is not satisfactory. This is also a bottleneck in the current research of the VQA tasks.
5. Conclusion
In this paper, we develop a co-attention spatial reasoning model. Based on the attention mechanism, we can strengthen the expression ability of both image and question features, meanwhile spatial reasoning module reason the spatial relationship between two modalities, which can help to learn the rich interactions of multi-modal instances. Experiments on the VQAv2 dataset shows that the model is effective in Yes/No, number types of questions.

References
[1] Donahue J, Hendricks L A, Guadarrama S, Rohrbach M, and Darrell T. 2014 Long-term recurrent convolutional networks for visual recognition and description. *IEEE Transactions on Software Engineering* 2014 99
[2] Nam H, Ha J and Kim J. 2017 Dual Attention Networks for Multimodal Reasoning and matching. *in 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp 2156-2164
[3] Guo D, Wang H, Wang S, and Wangb M. 2020 Textual-visual reference-aware attention network for visual dialog. *IEEE Transactions on Image Processing* 2020 99 1-1
[4] Guo D, Xu C and Tao S. 2019 Image-Question-Answer Synergistic Network for Visual dialog. *in 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* pp 10426-35
[5] Liu F, Xiang T, Hospedales T M et al. 2018 Inverse visual question answering: a new benchmark and vqa diagnosis tool *IEEE Transactions on Pattern Analysis and Machine Intelligence* 42(2) 460-474
[6] Malinowski M and Fritz M 2014 A multi-world approach to question answering about real-world scenes based on uncertain input. *Advances in Neural Information Processing Systems (NIPS)* 2014 1682–90
[7] Antol S, Agrawal A, Lu J, Mitchell M, Batra D, and Zitnick C L, et al. 2016 VQA: Visual Question Answering 2015 *IEEE International Conference on Computer Vision (ICCV).* IEEE.
[8] Li H, Wang P, Shen C, and Hengel A V D 2020 Visual Question answering as reading comprehension. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).* IEEE
[9] Gupta A, Garg M, Deshmane S, Singh P K, and Agarwal B 2020 Object-based classification for Visual Question Answering 2020 3rd International Conference on Emerging Technologies in Computer Engineering: Machine Learning and Internet of Things (ICETCE)
[10] Xu K, Ba J, Kiros R, Cho K, Courville A, and Salakhutdinov R, et al. 2015 Show, attend and tell: neural image caption generation with visual attention. *Computer Science* 20 48-57

![Figure 3. Typical examples of the model.](image-url)
[11] Lu J, Yang J, and Batra D 2016 Hierarchical question-image co-attention for visual question answering C Advances In Neural Information Processing Systems pp 289-297

[12] Chen K, Wang J, Chen L C, Gao H, Xu W, and Nevatia R 2015 Abc-cnn: an attention based convolutional neural network for visual question answering. arXiv preprint arXiv:1511.05960

[13] Peter A, Chris B, Damien T, and Lei Z 2018 Bottom-up and top-down attention for image captioning and visual question answering. IEEE Conference on Computer Vision and Pattern Recognition 2018 6077–86

[14] Zhou B, Tian Y, Sukhbaatar S, Szlam A, and Fergus R 2015 Simple baseline for visual question answering. arXiv preprint arXiv:1512.02167

[15] Kim J H, Lee S W, Kwak D H, Heo M O, Kim J, and Ha J W et al. 2016 Multimodal residual learning for visual qa. Advances in Neural Information Processing Systems 2016 361-369

[16] Peng L, Yang Y, Bin Y, Xie N, Shen F, and Ji Y, et al. 2019 Word-to-region attention network for visual question answering Multimedia Tools and Applications 78(3) 3843-58

[17] Yang Z, He X, Gao J, Deng L, and Smola A 2016 Stacked attention networks for image question answering IEEE Conference on Computer Vision and Pattern Recognition IEE.

[18] Shih K J, Singh S, and Hoiem D 2016 Where to look: focus regions for visual question answering 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) IEE.

[19] Ben-Younes H, Cadene R, Cord M, and Thome N 2017 Mutan: Multimodal Tucker fusion for visual question answering. In international Conference on Computer Vision (ICCV) pp 2612-2620

[20] Kim J H, On K W, Lim W, Kim J, Ha J W, and Zhang B T 2017 Hadamard product for low-rank bilinear pooling In International Conference on Learning Representation (ICLR)

[21] Fukui A, Park D H, Yang D, Rohrbach A, and Rohrbach M 2016 Multimodal compact bilinear pooling for visual question answering and visual grounding. in Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing pp 457-468

[22] Yu Z, Yu J, Fan J, and Tao D 2017 Multi-modal factorized bilinear pooling with co-attention learning for visual question answering. IEEE International Conference on Computer Vision (ICCV) 2017 1839-48

[23] Yu Z, Yu J, Xiang C, Fan J, and Tao D 2018 Beyond bilinear: generalized multimodal factorized high-order pooling for visual question answering. IEEE Transactions on Neural Networks and Learning Systems 29(12) 5947-59

[24] Yu D, Fu J, Mei T, and Rui Y 2017 Multi-level attention networks for visual question answering Computer Vision and Pattern Recognition. IEEE.

[25] Yu Z, Yu J, Cui Y, Tao D, and Tian Q 2019 Deep modular co-attention networks for visual question answering. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2019 6281-90

[26] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, and Gomez A N, et al. 2017 Attention is all you need Advances in Neural Information Processing Systems pp 5998-6010

[27] Hu J, Shen L, Albanie S, Sun G, and Wu E 2017 Squeeze-and-excitation networks. IEEE Transactions on Pattern Analysis and Machine Intelligence 42(8) 2011-23

[28] Pennington J, Socher R, and Manning C 2014 Glove: Global vectors for word representation Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP) pp 1532-1543