Improving reasoning with contrastive visual information for visual question answering

Yu Long,1 Pengjie Tang,2 Hanli Wang,1,3,4,5 and Jian Yu3
1Department of Computer Science and Technology, Tongji University, Shanghai, P. R. China
2College of Electronics and Information Engineering, Jinguangshan University, Ji’an, P. R. China
3Key Laboratory of Embedded System and Service Computing, Ministry of Education, Tongji University, Shanghai, P. R. China
4Shanghai Research Institute for Intelligent Autonomous System, Shanghai, P. R. China
5Email: hanliwang@tongji.edu.cn

Visual Question Answering (VQA) aims to output a correct answer based on cross-modality inputs including question and visual content. In general pipeline, information reasoning plays the key role for a reasonable answer. However, visual information is commonly not fully employed in many popular models nowadays. Facing this challenge, a new strategy is proposed in this work to make the best of visual information during reasoning. In detail, visual information is divided into two subsets: (1) question-relevant visual set, and (2) question-irrelevant visual set. Then, both of these two sets are employed by reasoning to generate reasonable outputs. Experiments are conducted on the benchmark VQAv2 dataset, which demonstrate the effectiveness of the proposed strategy. The project page can be found in https://mic.tongji.edu.cn/e6/8d/c9778a190093/page.htm.

Introduction: Visual Question Answering (VQA) is a cross-modality task which aims to output a reasonable answer by handling and fusing a target image and the corresponding question. In most of the popular works [1–4], there are three parts for an integrated framework: (1) question feature extraction module, (2) visual feature extraction module, and (3) cross-modality feature reasoning module to fuse and output an answer. In general, the reasoning ability of the third part is of the most significance to produce a correct and decent answer.

Nowadays, there are a few effective models for efficient reasoning. However, according to the evaluation of effectiveness of each input, it is found that only the question text is fully used in most of the current models, and the comprehensive key information of visual content is not sufficiently and effectively mined and employed. For instance, the prediction accuracy of ‘BLOCK [3]’ is still high (i.e. 44% on the VQAv2 dataset [5]) under the condition of only question being used but without any visual information.

Additionally, attention mechanism which is used to align and select feature in the reasoning module is widely employed in a variety of models [2–4]. During this procedure, only a few visual objects in the image are positive for reasoning meanwhile other visual objects are discarded. The discarded visual contents also contain structured information that could not be aligned with the question in the attention mechanism, and they should be classified as the question-irrelevant visual information. Inspired by the fact that contrastive visual information is able to help improve the ability of reasoning, a new strategy is proposed in this work to fully use all these visual semantics, with an overview of the proposed framework shown in Figure 1.

As shown in the feature extraction stage, given an image V and a question Q, the visual and question representations are extracted by Faster R-CNN and RNN/BERT, denoted as VR and QR, respectively. Both the representations are then sent to the following coarse reasoning stage, and the visual information will be split into question-relevant visual set and question-irrelevant visual set, which are denoted as FRre and VRre, respectively. Next, the two subsets are used in the fine-grained reasoning stage with the question feature. Clearly, besides a new coarse reasoning module for vision splitting, a new cross-modality reasoning branch that shares the parameters with the regular reasoning branch is required, as well as the corresponding new prediction target and loss function. The rest of this paper is organized as follows. At first, the related works are reviewed, and then the details of the proposed strategy are presented. Next, experiments are conducted to demonstrate the effectiveness of the proposed strategy. At last, the conclusion and future works are discussed.

Related works: The attention mechanism is usually employed to search semantic image regions in the VQA task. Particularly, a stacked attention module is employed to iteratively discover the available semantic visual regions by the query of question [2]. Besides visual attention for reasoning, the fusion strategy commonly plays the most important role to fuse the cross-modality inputs, and a series of bilinear pooling strategies are developed to replace the simple non-linear projection [6]. Recently, BERT, a powerful framework derived from natural language processing (NLP), is also introduced for better cross-modality feature fusion [4, 7, 8]. In addition, the language bias draws more and more attentions of researchers. In the work of GVQA [9], the training and testing sets of VQAv2 are re-split, clearly revealing that many models are driven by the superficial linguistic correlations. And then a few intuitive models are designed to relieve this bias [10]. However, all these works don’t pay attention to the unaligned visual information.

In fact, irrelevant or negative information is also used in other fields. For metric learning, the Siamese network [11] is a classic model, in which the loss function is established by the correlation measurement of inputs. The training strategy [11] is similar to the positive and negative strategy used in video-text retrieval [12], where the model is optimized by minimizing both the intra-sample loss and the inter-sample loss. However, the reasoning procedure in the VQA task concerns not only the aligning of the cross-modality information but also the reasoning on these aligned visual semantic information.

Algorithm: Attention mechanism, which is inspired by human visual psychological system, is frequently employed in vision-language tasks owing to its good intuition and excellent performance. In general, an attention mechanism pushes the model to select valuable features, which is consistent with the casual reasoning of human. In this procedure, the unnoticed information may be discarded, but it is also regarded to be the question-irrelevant information. From this motivation, both of the noticed and unnoticed visual information are employed to enhance the learning ability of model in this work. In practice, the models of BLOCK [3] and LXMERT [4] are employed as the backbone to represent the UpDn based models [3, 6, 13] and the BERT based models [4, 7, 8, respectively, to fully reveal the effectiveness of the proposed strategy.

Coarse reasoning: One of the important functions in the proposed framework is to align cross-modality features. To this end, the question-relevant visual features are well selected, weighted and then fused with the question feature. As shown in Figure 1, the proposed Coarse Reasoning Module (CRM) focuses on the question-relevant visual feature screening, which acts as a sub-task of the original reasoning procedure. The structure of CRM implemented with each backbone model could be the same as the cross-modality feature reasoning module in the employed backbone model. The details of CRM used in BLOCK and LXMERT are presented in Figure 2.

For the CRM module of BLOCK (the top part in Figure 2), it consists of a fusion unit (i.e., Block Fusion Unit), two non-linear layers for mapping the feature into a suitable size and a sigmoid layer to suppress the value into (0, 1). On the other hand, the CRM structure of LXMERT is similar to that of BLOCK, except that they both share the same architecture of fusion unit in their own respective backbone. In the BERT based model, the output size is equal to that of the input, and only the outputs corresponding to the visual inputs are used in the following.
procedure. As shown in the coarse reasoning stage in Figure 1, the outputs of CRM can be formatted as $\text{Gate} \in \mathbb{R}^N$, where $N$ is the number of visual objects. And then the visual representation $\mathbb{V}R$ is split into two subsets: (1) the question-relevant visual set $\mathbb{V}R_{\text{re}} = \mathbb{V}R \times \text{Gate}$, and (2) the opposite question-irrelevant visual set $\mathbb{V}R_{\text{ir}} = \mathbb{V}R \times (1 - \text{Gate})$, which will be fed to the following reasoning module.

Fine-grained reasoning: In order to enhance the reasoning ability, the question-irrelevant visual information is employed at the training phase besides the question-relevant visual information. In particular, a new reasoning branch is designed and optimized in the training procedure to make full use of the visual semantic information. It indicates that the corresponding loss functions are also required during training. The structure of the new reasoning branch is the same as the regular reasoning branch in the backbone, and two branches share the same parameters.

Regarding the regular reasoning branch in the backbone, it is used to generate the reasonable answer. And as usual, the binary cross entropy loss is used to optimize the branch by minimizing the distance between the output and the target. The loss function is written as:

$$
\mathcal{L}_{\text{ocr}} = \mathcal{T}_{\text{ocr}} \cdot \log (\sigma (P_{\text{ocr}})) + (1 - \mathcal{T}_{\text{ocr}}) \cdot \log (1 - \sigma (P_{\text{ocr}})).
$$

where $P_{\text{ocr}} \in \mathbb{R}^{K \times 1}$ and $\mathcal{T}_{\text{ocr}} \in \mathbb{R}^{K \times 1}$ are the output vector and the target vector of the regular reasoning branch while $K$ is the number of predicted answers, and $\sigma$ is the sigmoid function.

The new additional reasoning branch is designed to make the model to further use the visual content during the procedure of cross-modality information reasoning. And the state should be identified when to further understand the visual content during the procedure of cross-predicted answers, and $\mathcal{L}_{\text{ocr}}$ should be less than 1, so $K$ is adopted as the target label for this branch. The output vector and the target vector are written as $P_{\text{r}} \in \mathbb{R}^{K \times 1}$ and $\mathcal{T}_{\text{r}} \in \mathbb{R}^{1}$, respectively. For the loss function, the cross entropy is also employed as formatted as:

$$
\mathcal{L}_{\text{r}} = -\log \left( \frac{\exp (P_{\text{r}}[K])}{\sum_{j=1}^{K} \exp (P_{\text{r}}[j])} \right) = -P_{\text{r}}[K] + \log \left( \sum_{j=1}^{K} \exp (P_{\text{r}}[j]) \right).
$$

Compared to the regular branch, the convergence of this reasoning branch is unstable. Thus a new flat loss function is built to solve the problem, which is formatted as:

$$
\mathcal{L}_{\text{flat}} = \mathcal{T}_{\text{flat}} \cdot \log (\sigma (P_{\text{r}})) + (1 - \mathcal{T}_{\text{flat}}) \cdot \log (1 - \sigma (P_{\text{r}})).
$$

where $\mathcal{T}_{\text{flat}}$ is the corresponding prediction target of the loss function for this branch, which is a vector and could be formatted as $\mathcal{T}_{\text{flat}} \in \mathbb{R}^{K \times 1}$. To stabilize convergence, $\mathcal{T}_{\text{flat}}$ should keep relatively consistent with the original branch output, which is currently computed as:

$$
\mathcal{T}_{\text{flat}} = \max \left( \frac{\exp (P_{\text{r}})}{\sum \exp (P_{\text{r}})} - \mathcal{T}_{\text{ocr}}, 0 \right).
$$

As a consequence, the loss function of the whole model is

$$
\mathcal{L}_{\text{total}} = \lambda_0 \mathcal{L}_{\text{ocr}} + \lambda_1 \mathcal{L}_{\text{r}} + \lambda_2 \mathcal{L}_{\text{flat}},
$$

in which $\lambda_0$, $\lambda_1$ and $\lambda_2$ are the super-parameters which are set to be 1.0, 1/batchsize and 1.0, respectively.

**Table 1. Performance (%) of the proposed strategy on the VQAv2 dataset, ‘ACC (val set)’ means the accuracy on the val set and ‘ACC (test-dev set)’ means the accuracy on the test-dev set**

| Model       | ACC (val set) | ACC (test-dev set) |
|-------------|---------------|--------------------|
| SAN         | 55.39         | 57.12              |
| SAN + Ours  | 56.69         | 58.16              |
| BAN         | 64.51         | 69.52              |
| BAN + Ours  | 65.32         | 70.16              |
| BLOCK       | 63.03         | 66.00              |
| BLOCK + Ours| 63.61         | 66.31              |
| LXMERT      | 69.97         | 72.42              |
| LXMERT + Ours| 70.32        | 72.56              |

Experiments: In this section, the effectiveness of the proposed strategy is evaluated by several experiments with popular backbones on the public dataset VQA v2. VQA v2 is the improved version of VQA v1 which is the first large dataset annotated by human for the VQA task. The sample format used for open-ended purpose is [Image, Question, Answer]. In practice, 443757, 214354 and 447797 sample pairs are used for training, validation and testing, respectively. As for the questions in the dataset, they are categorized into 3 types: yes/no, number and other. The 3129 most frequent answers are usually used as the labels. All the experiments in this work are implemented on the PyTorch v1.7.1 framework.

Results: Besides the aforementioned models, another two classic UpDn based models including SAN and BAN are also employed as the backbone models to comprehensively reveal the effectiveness of the proposed strategy. The fusion unit of CRM in SAN is the solo attention used in the backbone model, while for simplification the fusion unit of CRM in BAN is the same as the fusion unit used in BLOCK. The optimization methods used for all these UpDn based models are Adamax, but the BertAdam method is employed for the BERT based model.

As shown in Table 1, it is obviously that the performances of the proposed strategy with all these models are better than the baseline backbone models. Particularly, the improvements with the LXMERT backbone based model is more exciting. In addition, the BERT based models are usually pre-trained on other larger datasets, but the parameters of CRM and backbone are initialized as the original model, and all the other additional parameters (i.e., the parameters in the appended fully connected layer) are randomly initialized in our experiments. The better performance shows that even without any pre-trained procedure, the proposed strategy could still gain an obvious performance improvement.

This suggests that the reasoning mode in the original architecture is not changed with the proposed reasoning method, but the reasoning ability of the regular branch is enhanced by the appended new branch. Besides the results listed above, more experiments that concern the sensitivity of the parameters $\lambda_0$, $\lambda_1$, and $\lambda_2$ are conducted. To maintain the effectiveness of the proposed method, the value of $\lambda_1$ should be less than 1/batchsize and the value of $\lambda_2$ should be larger than or equal to 1. By contrast, the model using the proposed strategy performs worse than the backbone without using proper visual input, as seen from the results in Table 2. Concretely, the visual information is changed in the validation stage, and ‘VZ’ means each image in the dataset is changed by another image, while ‘ZV’ means each image is replaced by a zero matrix. The performance trend is consistent with the intuition that the new designed branch can make the model to grab more correlations between visual information and question and ignore the correlation between the
Table 2. Performance (%) of the proposed strategy with different visual inputs on the VQA2 dataset; ACCO (val set), ACCOR (val set) and ACCZV (val set) denote the accuracy with the original image, the wrong image and the zero-value image on the val set, respectively

| Model          | ACCO (val set) | ACCOR (val set) | ACCZV (val set) |
|----------------|----------------|-----------------|-----------------|
| BLOCK          | 63.06          | 34.23           | 32.38           |
| BLOCK + Ours   | 63.61          | 33.87           | 31.94           |

Table 3. Performance (%) comparison of several backbone models with/without CRM on the VQA2 val set

| Model         | ACC (val set) |
|---------------|---------------|
| BLOCK         | 63.06         |
| BLOCK + CRM   | 63.12         |
| BLOCK + Ours  | 63.61         |
| LXMERT        | 69.97         |
| LXMERT + CRM  | 70.06         |
| LXMERT + Ours | 70.32         |

Table 4. Performance (%) comparison of several backbone models with different loss functions on the VQA2 val set

| Model      | ACC (val set) |
|------------|---------------|
| BLOCK      | 63.06         |
| BLOCK + LFL | 63.10        |
| BLOCK + LFIr + LFL | 63.26     |
| BLOCK + LFIr + LFIr | 63.61     |
| LXMERT     | 69.97         |
| LXMERT + LFL | 70.10       |
| LXMERT + LFIr | 70.21      |
| LXMERT + LFIr + LFIr | 70.32    |

input question and the output label. This shows that the reasoning ability can be further enhanced by the new designed reasoning branch.

Ablation study: Ablation experiments are implemented to further reveal the effectiveness of each component of the proposed strategy. The models BLOCK and LXMERT are used as the backbones for our implementation.

Compared to the backbone models, more parameters are required by the newly introduced CRM module. This comes from the new module CRM actually. The effectiveness of CRM is evaluated. Specifically, the CRM module is appended on the backbone models, meanwhile the new reasoning branch and the corresponding loss function are removed from the model, and all the super-parameters are the same as that used in the backbone. The results are presented in Table 3, where it can be seen that both the models could slightly benefit from using CRM, and the main improvement of employing the proposed strategy comes from the usage of the new reasoning branch.

The effectiveness of the new reasoning branch is derived from the usage of the loss functions LFL and LFIr together. To evaluate the effect of these two loss functions, each of them is appended individually. Note that the model without LFIr could not stably converge unless LFL is small enough (i.e., 1e-7). The experimental results are shown in Table 4 (where +LFL means only LFL is used). On the baseline backbone BLOCK, the performance by only using LFL drops compared to the model using CRM, and the improvement on the backbone LXMERT is subtle. This may be caused by the magnitude of λ1 and the effectiveness of the proposed method could be nearly ignored with a small value. As far as LFIr is concerned, it is designed to stabilize the convergence of the new reasoning branch, both of the two models could benefit from using LFIr.

Conclusion: To further exploit the visual information in the reasoning procedure for the VQA task, a novel reasoning strategy is proposed in this work, which makes the model to align not only the question with the question-relevant visual feature, but also the question with the question-irrelevant visual feature. As a consequence, a more accurate correlation between the question and the visual image is obtained. The experimental results demonstrate the effectiveness of the proposed strategy. However, there are still problems remained to be addressed. For instance, the objective function of the question-irrelevant reasoning branch needs to be further investigated and the coarse reasoning module is required to be further explored in future works.

Acknowledgements: This work was supported in part by National Natural Science Foundation of China under Grants 61976159 and 62062041, Shanghai Innovation Action Project of Science and Technology under Grant 2051100700, and Shanghai Municipal Science and Technology Major Project (No. 2021SHZDZX0100).

© 2021 The Authors. Electronics Letters published by John Wiley & Sons Ltd on behalf of The Institution of Engineering and TechnologyThis is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

Received: 9 April 2021 Accepted: 26 May 2021
doi: 10.1049/ell2.12255

References
1. Antol, S., et al.: VQA: visual question answering. In: International Conference on Computer Vision (ICCV), Santiago (2015)
2. Yang, Z., et al.: Stacke d attention networks for image question answering. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas (2016)
3. Ben-Younes, H., et al.: BLOCK: Bilinear superdiagonal fusion for visual question answering and visual relationship detection. In: 33rd AAAI Conference on Artificial Intelligence (AAAI), Honolulu (2019)
4. Tan, H., Bansal, M.: LXMERT: learning cross-modality encoder representations from transformers. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Hong Kong (2019)
5. Goyal, Y., et al.: Making the V in VQA matter: elevating the role of image understanding in visual question answering. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu (2017)
6. Yu, Z., et al.: Beyond bilinear: generalized multimodal factorized high-order pooling for visual question answering. IEEE Trans. Neural Networks and Learning Systems (TNNSL) 29(12), 5947–5959 (2018)
7. Lu, J., et al.: VL-BERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In: Proceedings of the 33rd Conference on Neural Information Processing Systems (NeurIPS), Vancouver (2019)
8. Su, W., et al.: VL-BERT: Pre-training of generic visual-linguistic representations. In: Proceedings of International Conference on Learning Representations (ICLR), New Orleans, LA, USA (2019)
9. Agrawal, A., et al.: Don’t just assume; look and answer: overcoming priors for visual question answering. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Seattle (2020)
10. Chopra, S., Hadsell, R., LeCun, Y.: Learning a similarity metric discriminatively, with application to face verification. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), San Diego (2005)
11. Zhang, Z., et al.: Regularized two-branch proposal networks for weakly-supervised moment retrieval in videos. In: Proceedings of the 28th ACM International Conference on Multimedia (ACMMM), Seattle (2020)
12. Yu, Z., et al.: Deep modular co-attention networks for visual question answering. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach (2019)