Learning to Answer Complex Visual Questions from Multi-View Analysis

Minjun Zhu[1,2⋆], Yixuan Weng[1*], Shizhu He[1,2], Kang Liu[1,2], Jun Zhao[1,2]

[1] National Laboratory of Pattern Recognition,, Institute of Automation, CAS
[2] School of Artificial Intelligence, University of Chinese Academy of Sciences
zhuminjun2020@ia.ac.cn, wengsyx@gmail.com
{shizhu.he, kliu, jzhao}@nlpr.ia.ac.cn

Abstract. Visual Question Answering (VQA) has received increasing attention in NLP research. Most VQA images focus on natural scenes. However, some images widely used in textbooks such as diagrams often contain complicated and abstract information (e.g. constructed graphs with logic and concepts). Therefore, Diagram Question answering (DQA) is a challenging but significant task, which is also helpful for machines to understand human cognitive behaviors and learning habits. On DQA task, we propose a multi-perspective understanding based visual question-answering method, which constructs a variety of different self-monitoring tasks in the form of prompts to help the model learn deeper information. For the first time, we propose a decoding method of "Cross Entropy constraint Decoding", which can effectively constrain the content generated by the text when performing multiple selection tasks. This method has obtained SOTA in the evaluation task of CCKS-2022, which fully proves the effectiveness of the method.

Keywords: Diagram Question Answering · Visual Question answering · Computer science.

1 Introduction

Question Answering (QA) systems have long pursued the ability to understand human cognitive behaviors and learning habits. Diagram Question Answering (DQA) task requires dynamic and complex reasoning of knowledge representation, which helps to improve the understanding of abstract images by computers. Generally, Diagram is manually constructed with abstract meanings and widely used in pre-defined scenarios, such as textbooks and dictionaries, which requires fully understand the abstract schematic information according to questions. DQA is still a challenge task because the complicate expression and the lack of data.

CSDQA is a novel visual question answering dataset which contains 1294 different diagram such as circle, rectangle and triangle. CSDQA is a multiple-choice dataset that is mainly collected from textbooks. Each sample contains the

⋆ These authors contribute equally to this work.

question, diagram, choice, and answer. It divides datasets questions into simple reasoning and complex reasoning. Complex reasoning requires two-step reasoning on the image to get the answer, and the proportion of complex reasoning is 22.98% in all questions.

Pre-trained models have promoted the progress of natural language processing and even multimodality. In previous studies, sequence-to-sequence models have been widely used for a large number of downstream tasks. CLIP uses a large number of text and text pairs in order to narrow them in the same embedding space; Vl-bert takes both visual and textual features as input. VLMo jointly learns a dual encoder and a fusion encoder with a modular Transformer network. However, multi-modal neural network model shows poor performance on diagram question answering task. Because existing multi-modal models are pre-trained on a large number of natural scenes images, which lack the ability to understand abstract information of diagram such as concept and logical relationship within different graphics. Moreover, the scarcity of diagram resources also limits the large corpus training of DQA task models.

We propose a Multi-View Analysis (MVA) method to enhance model's understanding of schematic abstract information. MVA are constructed with five different task forms and unified in Text2Text form. We take OFA model as backbone, and then introduce MVA to enhance the abstract understanding ability. In order to avoid the invalid generation in Decoding phase, we introduce Cross Entropy constraint Decoding, which restricts the results to achieves higher inference accuracy.

2 Main Methods

2.1 Multi-View Training

Existing multi-modal pre-trained models have strong ability to extract information from natural scene images, but it is still difficult for model to directly understand the abstract diagram information in low-resource Diagram Question Answering (DQA) task. Therefore, we introduce multi-View multi-task training to help model deeply understand the diagram. As depicted in figure 1. We constructed five different task forms based on Prompt and unified five different objectives as Seq2Seq form for training. Five tasks are respectively shown as follows:

- 1. Original task: the original data, including images, questions and answers.
- 2. Answer Judge task: ask model to answer whether the statement combined with the question, answer and judgement is correct.
- 3. Image Description task: require model to describe the image with details.
- 4. Ask Question task: require model to ask a question based on the input image and answer.
- 5. Image Text Match task: change some image-question pairs, and ask model to judge whether the description is grounding with the input image.

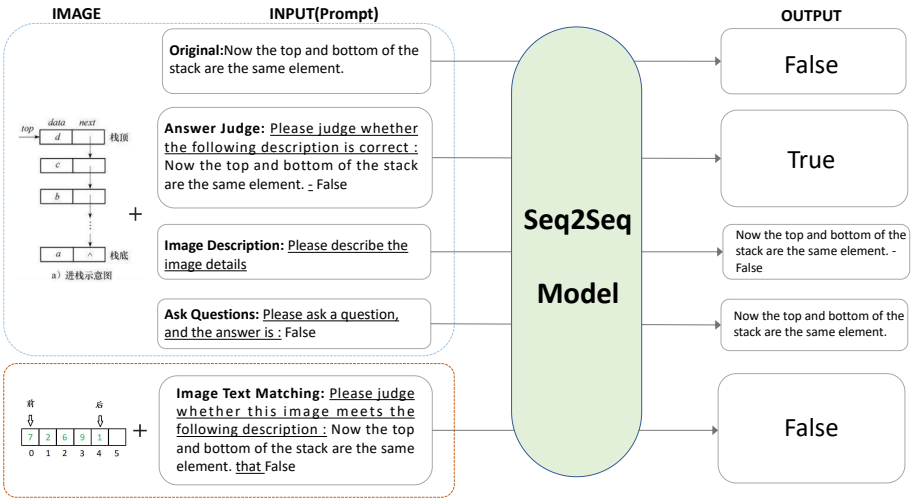

Fig. 1. The Multi-View Training by different tasks. We take origin as the initial sample and realize multi-view training by changing the position of its questions / answers / diagram.

2.2 Step Training

In order to alleviate the catastrophic forgetting problem caused by overtraining of the pre-trained model in Multi-View Training. We implemented the step learning based on Child-tuning [24] method. The Child-tuning method is used to fine-tune the backbone model in our method, where the parameters of the Child network are updated with the gradients mask. For the DQA task, the task-independent algorithm is used for child-tuning. When fine-tuning, the gradient masks are obtained by Bernoulli Distribution [2] sampling from in each step of iterative update, which is equivalent to randomly dividing a part of the network parameters when updating. The equation of the above steps is shown as follows

$$w_{t+1} = w_t - \eta \frac{\partial \mathcal{L}(w_t)}{\partial w_t} \odot B_t B_t \sim Bernoulli(p_F) \tag{1}$$

where the notation $\odot$ represents the dot production, p_F is the partial network parameter.

3 Experiments

In this section, we will introduce the experimental settings and evaluation indicators. Then we compare MVA with the existing VQA technology and ablation experiments to prove the effectiveness of our method.

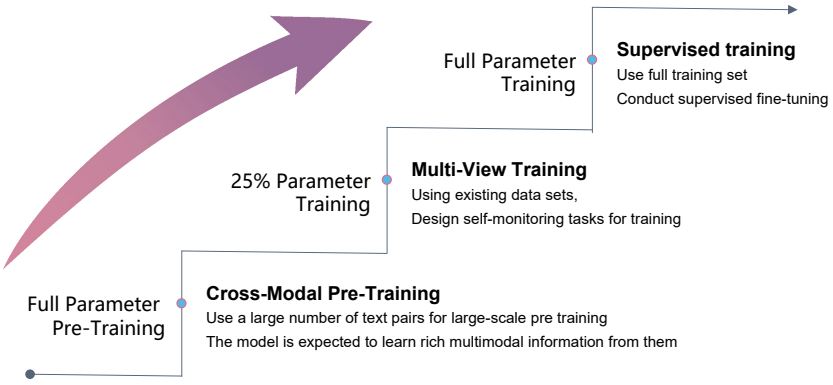

Fig. 2. In the training stage, we use step-by-step training to gradually make the model learn the image features of the schematic diagram.

3.1 Evaluation Metrics

Following prior work [13, 19, 1, 4]. We use the accuracy rate as the evaluation metrics. We assume that the total quantity is n and the predicted correct quantity is C, then the calculation of concurrency is as follows:

$$\textbf{Accuracy} = \frac{\mathbf{C}}{\mathbf{n}} \times 100\% \tag{2}$$

3.2 Implementation Details

In order to compare the functions of the system more fairly. In recent years, natural language processing significant progress has been achieved [5, 17] due to the introduction of Pre-trained Language Model [15, 3, 18]. Therefore, more and more methods begin to introduce the pre-trained language model in the VQA task [11, 22, 16, 8, 10].

For all methods, we use the same size model for finetuning. And we all use the large-size model for testing. We follow the original code for the remaining settings. We train the model using the Pytorch[3] [14] on the NVIDIA RTX3090 GPU and use the hugging-face[4] [23] framework. We use the AdamW [12] as the optimizer and the learning rate is set to 1e-5 with the warm-up [7]. The batch size is 8. We set the maximum length of 512, and delete the excess. We use the linear decay of the learning rate and gradient clipping of 1e-6. The dropout [20]

[3] https://pytorch.org
[4] https://github.com/huggingface/transformers

Hyper-parameter	**Value**
Image Encoder	ResNet152 [6]
Encoder Hidden Size	4096
Encoder Num Layers	12
Encoder Attention Heads	16
Decoder Hidden Size	4096
Decoder Num Layers	12
Decoder Attention Heads	16
Dropout	0.1
Max Token Length	512
Language Model Loss Function	Cross Entropy
Learning Rate	1e-5
Batch size	8
Num Epochs	20
Weight Decay	1e-4
FP16	True
Gradient Accumulation	1
Beam Search	5

Table 1. Hyper-parameter settings.

Metric	**Accuracy**			
Learning Rate	**1e-5**	**2e-5**	**3e-5**	**Avg.**
Random Mode	/	/	/	35.84
LayoutLMv3$_{Base}$[9](2022)	38.25	35.52	36.89	36.89
LayoutLMv3$_{Large}$[9](2022)	40.21	36.89	37.04	38.05
OFA$_{Base}$[21](2022)	52.86	53.31	52.25	52.80
OFA$_{Large}$[21](2022)	<u>54.06</u>	<u>53.37</u>	<u>53.61</u>	<u>53.68</u>
MVA	**58.89(4.83↑)**	**58.58(5.21↑)**	**57.23(3.62↑)**	**58.23 (4.55↑)**

Table 2. Performance comparison of the variants methods on Computer Science Diagrams dataset. We highlight the best score ineach column in **bold**, and the second best score with <u>underline</u>. We will also show the improvement between first place and second place.

of 0.1 is applied to prevent overfitting. The detailed experimental settings are shown in **Table 1**.

All hyperparameters are optimized on the Valid set. In all our experiments, at the end of each training phase, we will test the effective data set and select the highest model (mainly depending on Accuracy) in the test data set for prediction. We report the results in the test data set. We repeated the experiment three times and reported the average score.

3.3 Comparison with State-of-the-Art Methods

In the CSDia dataset, we compared the baseline scheme with the existing dialogue generation.

The $LayoutLM_{v3}$ [9] is based on $LayoutLM_{v2}$ [26] and $LayoutLM_{v1}$ [25], and it uses unified text and image mask modeling objectives to pre train the

Metric	Accuracy			
Learning Rate	**1e-5**	**2e-5**	**3e-5**	**Avg.**
OFA_{Large}[21]	54.06	53.37	53.61	53.68
W/O MV	56.93	56.63	55.57	56.38
W/O Step training	58.13	58.58	56.93	57.88
W/O CE Decode	58.28	57.98	57.23	57.83
MVA	58.89	58.58	57.23	58.23
+Full Data	60.24	59.49	59.19	59.64

Table 3. Performance comparison of the variants methods on Computer Science Diagrams dataset. We conducted some control experiments and tried to train the model using full data and showed its scores.

multimodal model, which simplifies the model design. It requires that the hidden words in the text be restored according to the uncovered text and layout information in the document data set, and the masked image block data be restored at the same time. The $LayoutLM_{v3}$ achieved better results in form tasks than previous work.

The OFA [21] model realizes the unification of modes, tasks and structures, unifies the multi-modal and single-modal understanding and generation tasks into a simple seq2seq generative framework, and performs pre training and fine-tuning using task instructions. The OFA has achieved SOTA in four cross modal tasks: image capture, VQA, visual entailment and referring expression synthesis.

3.4 Experimental Result

We report the performance of the model in Table 2. We compared several models in different forms and sizes and selected $LayoutLM_{v3}$ and OFA respectively. We show the improvement of our method compared with the baseline model.

Among them, the performance of $LayoutLM_{v3}$ is weak. The $LayoutLM_{v3}$ has learned a large number of abstract characters and symbols in the pre-training process, it is difficult to learn the relevant features in the question and answer of complex schematic diagrams due to the lack of understanding of the overall graph and the task of the pre-training phase is mainly mask recovery rather than visual question and answer, which leads to difficulty in finetuning. In addition, we have directly fine-tuned the OFA model. It is not difficult for us to find that the OFA model can have strong performance on the visual question and answer the task, and the performance of the *Base* size and the *Large* size are relatively close, which indicates that the additional knowledge brought by the pre-training for the model has reached the limit. For the abstract schematic question and answer task, the OFA model still has a large room for improvement.

When we use the MVA, the baseline model can further learn more relevant knowledge. In Table 2, MVA exceeds the baseline model in three different learning rate settings, and its average score exceeds the baseline by 4.55%, which fully proves the reliability of our method.

A			B		
Rank	**Team**	**Score**	**Rank**	**Team**	**Score**
1	**灵境_CASIA**	**55.57**	1	**灵境_CASIA**	**60.19**
2	maoada	54.22	2	key7	58.09
3	福气boy	53.61	3	Cube	55.02
4	国足10号	53.01	4	福气boy	54.53
5	qddy	52.56	5	maoada	54.53
6	northsky	51.81	6	国足10号	53.39

Table 4. Online performance of CCKS-2022 in DQA task.

3.5 Ablation Study

In **Table 3**, we can see some performance comparisons. We further carry out care learning in OFA [21], which is the best pre-trained model in Diagram Question Answering task. It can fully show the effect differences brought by different methods.

First, we try to cancel the Multi-View, which means that we no longer require the model to pretrain multi-view task in Diagram Question Answering. This may lead to the lack of understanding of the diagram so that the generated answer lacks the modeling of the diagram.

After canceling the Step Training, we directly train the model in one step. However, the experimental results show that compared with the Step Training, the performance will be reduced. We believe that this is because after the introduction of MV, the training tasks may deviate from DQA, resulting in the focus of the model learning is no longer DQA tasks. Therefore, using two-stage training and fixing parameters in the MV stage can help the model mitigate catastrophic forgetting and bring higher performance. Finally, if the CE decode is cancelled, there will be more than 2% of the answers, and it will be difficult to answer because it cannot be matched.

Since the task provides the target of the verification set, we additionally use full data for training and test in test, which brings us additional performance improvement.

3.6 Online Result

In **Table 4**, We showed the online results of two different lists, and we all got the results of SOTA, which reflects the superior performance of the method.

4 Conclusions

In this paper, we propose a new method to solve complex text questi answering. We provide multimodal and multi-perspective learning for the pre-training language generation model. By constructing a large number of different learning tasks, we can make the pre-training model play a more effective role in the low resource Abstract schematic scenario. Our MVA module is very flexible. We

have built a unified task framework for multitasking learning, which can support almost all multimodal seq2seq models. In addition, we introduce CE decode decoding to constrain the generation results, which enhances the controlonslability of the multi-modal generation model and improves the performance. In the DQA task of CCKS-2022, our method won first place, which provides a powerful solution for the complex visual question and answers task.

References

1. Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C. L., and Parikh, D. Vqa: Visual question answering. *international conference on computer vision* (2015).
2. Chen, S. X., and Liu, J. S. Statistical applications of the poisson-binomial and conditional bernoulli distributions. *Statistica Sinica* (1997), 875–892.
3. Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (Minneapolis, Minnesota, June 2019), Association for Computational Linguistics, pp. 4171–4186.
4. Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., and Parikh, D. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. *International Journal of Computer Vision* (2016).
5. Han, X., Zhang, Z., Ding, N., Gu, Y., Liu, X., Huo, Y., Qiu, J., Yao, Y., Zhang, A., Zhang, L., Han, W., Huang, M., Jin, Q., Lan, Y., Liu, Y., Liu, Z., Lu, Z., Qiu, X., Song, R., Tang, J., Wen, J.-R., Yuan, J., Zhao, W. X., and Zhu, J. Pre-trained models: Past, present and future. *AI Open 2* (2021), 225–250.
6. He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. *arXiv: Computer Vision and Pattern Recognition* (2015).
7. He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2016), pp. 770–778.
8. Hu, R., and Singh, A. Unit: Multimodal multitask learning with a unified transformer. *international conference on computer vision* (2021).
9. Huang, Y., Lv, T., Cui, L., Lu, Y., and Wei, F. Layoutlmv3: Pre-training for document ai with unified text and image masking.
10. Li, B., Weng, Y., Sun, B., and Li, S. Towards visual-prompt temporal answering grounding in medical instructional video. *arXiv preprint arXiv:2203.06667* (2022).
11. Li, W., Gao, C., Niu, G., Xiao, X., Liu, H., Liu, J., Wu, H., and Wang, H. Unimo: Towards unified-modal understanding and generation via cross-modal contrastive learning. *meeting of the association for computational linguistics* (2020).
12. Loshchilov, I., and Hutter, F. Decoupled weight decay regularization. In *International Conference on Learning Representations* (2018).
13. Malinowski, M., and Fritz, M. A multi-world approach to question answering about real-world scenes based on uncertain input. *neural information processing systems* (2014).

14. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems* (2019), H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds., vol. 32, Curran Associates, Inc.
15. Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. Deep contextualized word representations. In *North American Chapter of the Association for Computational Linguistics* (2018).
16. Qi, D., Su, L., Song, J., Cui, E., Bharti, T., and Sacheti, A. Imagebert: Cross-modal pre-training with large-scale weak-supervised image-text data.
17. Qiu, X., Sun, T., Xu, Y., Shao, Y., Dai, N., and Huang, X. Pre-trained models for natural language processing: A survey. *CoRR abs/2003.08271* (2020).
18. Radford, A., and Narasimhan, K. Improving language understanding by generative pre-training.
19. Ren, M., Kiros, R., and Zemel, R. S. Exploring models and data for image question answering. *neural information processing systems* (2015).
20. Srivastava, N., Hinton, G. E., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research 15* (2014), 1929–1958.
21. Wang, P., Yang, A., Men, R., Lin, J., Bai, S., Li, Z., Ma, J., Zhou, C., Zhou, J., Yang, H., and Zhou<ericzhou, C. Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework.
22. Wang, W., Bao, H., Dong, L., and Wei, F. Vlmo: Unified vision-language pre-training with mixture-of-modality-experts. *arXiv: Computer Vision and Pattern Recognition* (2021).
23. Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T. L., Gugger, S., Drame, M., Lhoest, Q., and Rush, A. M. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations* (Online, Oct. 2020), Association for Computational Linguistics, pp. 38–45.
24. Xu, R., Luo, F., Zhang, Z., Tan, C., Chang, B., Huang, S., and Huang, F. Raise a child in large language model: Towards effective and generalizable fine-tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing* (2021), pp. 9514–9528.
25. Xu, Y., Li, M., Cui, L., Huang, S., Wei, F., and Zhou, M. Layoutlm: Pre-training of text and layout for document image understanding. *knowledge discovery and data mining* (2019).
26. Xu, Y., Xu, Y., Lv, T., Cui, L., Wei, F., Wang, G., Lu, Y., Florencio, D., Zhang, C., Che, W., Zhang, M., and Zhou, L. Layoutlmv2: Multi-modal pre-training for visually-rich document understanding. *meeting of the association for computational linguistics* (2020).