Multimedia Tools and Applications (2024) 83:7085–7096
https://doi.org/10.1007/s11042-023-15418-6

OECA-Net: A co-attention network for visual question answering based on OCR scene text feature enhancement

Feng Yan[1] · Wushouer Silamu[1,2] · Yachuang Chai[1] · Yanbing Li[1]

Received: 20 July 2022 / Revised: 15 February 2023 / Accepted: 18 April 2023 /
Published online: 5 June 2023

Abstract
Most VQA(visual question answering) models can not understand the scene text in the image. Poor text reading ability is a significant reason for the current VQA model's poor performance. To solve the problems, we designed a co-attention model that incorporates the scene text features in images. We detect and obtain the OCR token in the image through the OCR model, which is conducive to further understanding the image. We design a model based on a co-attention mechanism, including a question self-attention unit, question-guided image visual attention unit and question-guided image OCR token attention unit. The redundant question information is filtered under the question self-attention module. The question-guided attention module is used to obtain the final visual features and OCR token features in the image. The information of question text features, visual image features and OCR token features in the image is fused. We design a classifier which can get an answer from the fixed answer set or directly copy the text detected from the OCR model as the final answer so that the model can answer the questions about the text in the image. The experimental results show that our model is improved.

Keywords Visual question answering · Faster R-CNN · Self-attention · Spatial position relationship

✉ Feng Yan
yanfeng@stu.xju.edu.cn

Wushouer Silamu
wushour@xju.edu.cn

Yachuang Chai
cyc@stu.xju.edu.cn

Yanbing Li
liyb@xju.edu.cn

[1] School of Information Science and Engineering, Xinjiang University, Shengli Road 666, Urumqi, 830046, Xinjiang, China

[2] Xinjiang Key Laboratory of Multilingual Information Technology, Shengli Road 666, Urumqi, 830046, Xinjiang, China

1 Introduction

Deep learning has made significant progress in computer vision [14] and natural language processing [26]. Interdisciplinary disciplines between vision and natural languages, such as image caption [31], VQA [2, 23] and visual dialogue [18], have attracted strong attention in the field of vision and natural language.

VQA is a very challenging research direction, but most VQA models [34] focus on vision processing. For some images with scene text, most models [8, 39] need help to answer the questions about the scene text in the image. These images have actual text semantics and need to understand the image scene text.

The Lorra model [27] can effectively infer and answer the scene text information in the image. It uses the BUTD [1] attention model to infer the visual object. However, this method is limited by simple attention model interaction. The M4C [16] model jointly encodes questions, images, and text by using multimodal transformer Architecture [30] and obtains answers from OCR tokens or some fixed vocabulary iterations. Although this isomorphism processing method is easy to implement and fast to train, it does not distinguish between text and visual objects after isomorphism.

To solve the above problems, after obtaining the text features and the visual features based on BUA(bottom-up attention), and using OCR and FastText [5] to obtain the OCR text features in the image, we use the question self-attention unit to filter the redundant features, and then use the question-guide attention mechanism to process the visual features and the OCR token features. Finally, we fuse the three features, and the features input to the classifier are obtained. Based on the newly introduced question-guided OCR token feature, we add the copy of the OCR token as the extension of candidate answers so that the model can predict the OCR token outside the fixed candidate answers as the answer.

Our innovations and contributions can be summarized as follows:

- To enable the model to have the ability to understand text in the image, we designed a co-attention model that incorporates the scene text features in images. The deep-stacked co-attention mechanism guarantees to answer common questions that do not involve scene texts and enable the model to recognize texts in images and answer questions using scene texts.
- We design a classifier which can get an answer from a fixed answer set or directly copy the text in the image detected from the OCR model as the final answer.
- Ablation experiments show that the methods we proposed are effective. Our model has significantly improved the performance on the VQA 2.0 dataset [13].

2 Related work

2.1 Attention

The attention mechanism has been successfully applied to uni-modal tasks and simple multi-modal tasks. Paper [1] learned the visual attention of the image region from the input question of VQA, embedded the question into the visual space using the attention structure, and constructed a convolution kernel to search the noticed region in the image, which effectively promoted the representation ability of the model; Subsequently, many studies [10, 11, 29, 35, 36] introduced the use of visual attention to extract features and reduce the interference of redundant features in image and text information through an attention mechanism. In addition, papers [4, 20] use different multi-modal bilinear pooling methods to combine

the grid visual features with the text features to predict the answer. The results show that attention to learning vision and text modals helps enhance the fine-grained representation of images and questions to improve the model's accuracy effectively. However, these rough attention models can not infer the correlation between regions and question words.

Therefore, learning the co-attention between two modals can effectively improve the VQA results. Paper [38] simplifies the co-attention method into two steps. Firstly, the question is put into the self-attention mechanism to learn the dependency between question words. Then the most relevant visual region is searched in the question-guided attention module. At the same time, paper [19] proposed a bi-linear attention network to refine attention based on previously noticed features. In addition, we can use better models to encode features [17, 40] and further enhance the models.

2.2 Pre-training model

A visual language pre-training model (VLP) is a type of deep learning model that combines images and text for joint learning to obtain rich visual and language representations.

VLP models typically consist of an image encoder and a text encoder. The image encoder converts the input image into a low-dimensional vector, while the text encoder converts the input text into a low-dimensional vector. These vectors can then be aligned in a shared representation space to capture semantic similarities between the images and text.

Currently, some popular VLP models include UNITER [9], ViLBERT [22], LXMERT [28], BLIP [21], OFA [32], among others. These models have achieved state-of-the-art results in various visual language tasks, such as image captioning, question answering, visual reasoning, image classification, and more.

3 Proposed model

The multi-modal co-attention model is shown in Fig. 1. GloVe [25] and LSTM are used to extract question features, Faster R-CNN is used to extract image features, and FastText is

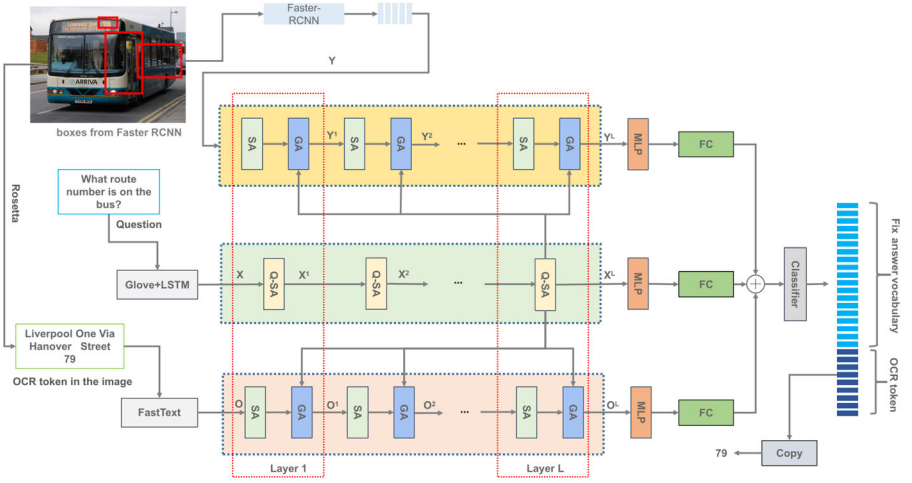


Fig. 1 Overview of the proposed model

used to encode OCR tokens in the image. Among them, SA is a self-attention unit, and GA corresponds to a guided attention unit.

3.1 Text question representation

Each question is encoded as a Glove vector for word representation. If the question length is less than 14, it is extended with a zero vector. The sequence after word embedding is encoded by LSTM [15], the memory mechanism of LSTM can effectively process long text information and prevent network gradient explosion, and the calculation formula is as follows:

$$X = LSTM(Glove(ques)) \tag{1}$$

where $X \in \mathbb{R}^{d_x * M}$ is the question representations, M is the length of question.

3.2 Image OCR tokens extraction and representation

We use the Rosetta [6] OCR system to extract text marks on each image, identify up to 50 OCR tokens in the image, and then use FastText [5] to represent OCR tokens. The calculation formula is as follows:

$$tokens = Rosetta(image) \tag{2}$$

After obtaining the image OCR tokens, we use the FastText to encode:

$$O = FastText(tokens) \tag{3}$$

where tokens is the OCR text of the image, $O \in \mathbb{R}^{d_{ocr} * P}$ is the OCR representations. P is the number of detected OCR token.

3.3 Image feature representation

Currently, the mainstream image feature extraction adopts the BUA [1] method to identify specific objects in the image. Figure 2 is the overview of the BUA model. The calculation formula is as follows:

$$Y = FRCNN(image) \tag{4}$$

where $Y \in \mathbb{R}^{d_y * N}$ is the vision feature, N is the number of the targets.

3.4 Co-attention model

Our model includes a question self-attention unit, a question-guided image OCR token attention unit, and a question-guided image vision attention unit. Learn the interaction

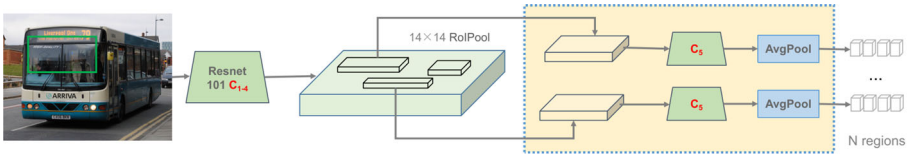


Fig. 2 Overview of the region feature extractor based on Bottom-Up Attention

information between the same modal or different modals through the attention mechanism. The question-guided image attention unit has the same implementation method as the question-guided image OCR token attention unit, except that the image OCR token feature representation replaces the image feature representation.

3.4.1 Multi-head self-attention mechanism

The multi-head self-attention mechanism module includes a multi-head attention layer, layer normalization [3] layer, residual link layer, and forward layer. As shown in Fig. 3a, the input feature X is mapped by three matrices to obtain the corresponding matrices Q, K, and V, which are calculated by scaled dot-product. The calculation formula is as follows:

$$\begin{cases} Q = XW^Q, K = XW^K, V = XW^V \\ Attn(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \end{cases} \tag{5}$$

where d_k is the dimension of the Query, Key and Vaule.

A multi-head attention mechanism can be adopted to improve the presentation ability further. Multi-head attention includes h attention operations, and each attention operation corresponds to a scaling dot product operation. The operation result is connected to the output representation of the multi-head attention layer:

$$\begin{cases} head_i = Attn(Q_i, K_i, V_i) \\ MHA(Q, K, V) = Concat(head_1, head_2, ..., head_h)W_o \end{cases} \tag{6}$$

where Q_i, K_i, V_i is the i-th head's matrix. $W_o \in \mathbb{R}^{h*d_h*d}$ are the projection matrices, d_h is the dimension of the output features.

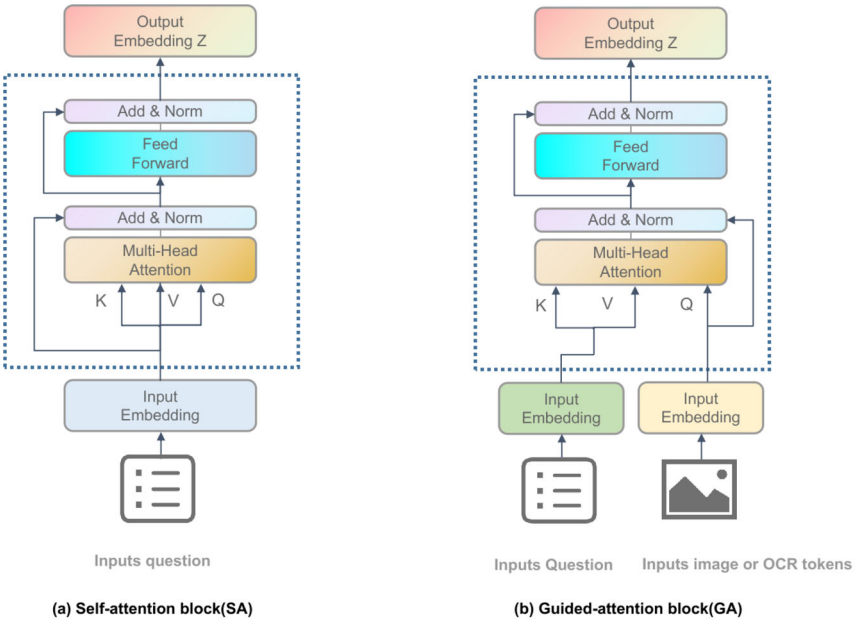


Fig. 3 Architecture of the two basic attention units

The output of the multi-head attention layer is then processed through the residual link layer and layer normalization to prevent the gradient from disappearing and accelerate the convergence of the model:

$$f = LayerNorm(X + MHA(Q, K, V)) \tag{7}$$

where d_k is the dimension of the query and the key.

After passing through the forward layer, the final output of the self-attention module is:

$$FFN(f) = FC(Dropout(ReLU(FC(f)))) \tag{8}$$

$$Z = LayerNorm(f + FFN(f)) \tag{9}$$

where FC is the fully-connected layers.

3.4.2 Self-attention unit and guided-attention unit

As shown in the Fig. 3a, SA is based on MHA. $X = \{x_1; x_2; ...; x_M\} \in \mathbb{R}^{M*d_x}$ is the input features of SA:

$$\begin{cases} Q = XW^Q, K = XW^K, V = XW^V \\ f = LayerNorm(X + MHA(Q, K, V)) \\ SA(X) = LayerNorm(f + FFN(f)) \end{cases} \tag{10}$$

where MHA is the multi-head attention layers.

The difference between the guided-attention mechanism and self-attention lies in the input of two different features. As shown in Fig. 3b, the corresponding matrices of question feature X as guidance features are K and V, and the corresponding matrix of visual feature (or image description feature) is Q:

$$\begin{cases} Q = YW^Q, K = XW^K, V = XW^V \\ Attn(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \\ f = LayerNorm(Y + MHA(Q, K, V)) \\ GA(X, Y) = LayerNorm(f + FFN(f)) \end{cases} \tag{11}$$

The calculation of GA is similar to SA.

3.4.3 Cascade of attention modules

To further improve the representation ability of the feature, we use a cascade approach to combine attention modules. The output of the last layer can be used as the input of the next layer. For the question self-attention mechanism, a total of L layers are stacked as the final question feature:

$$X^k = SA^k\left(X^{k-1}\right) \tag{12}$$

where $SA^1, SA^2, ..., SA^L$, represents the question self-attention of different layers. The final layer X^L is obtained as the final question feature.

Then, the final image features are obtained by using the SA and GA. The calculation formula is as follows:

$$Y^k = GA^k(X^L, SA(Y^{k-1})) \tag{13}$$

where $GA^1, GA^2, ..., GA^L$, represents the question Guided-attention of different layers. The obtained final image feature Y^L is used as the final image region feature.

Similarly, after obtaining the image OCT token feature O processed by LSTM, use the self-attention mechanism unit to process it, then input it to the question-guided attention module GA. The calculation formula is as follows:

$$O^k = GA^k(X^L, SA(O^{k-1})) \tag{14}$$

The last layer O^L obtained is used as the final image OCR token feature.

4 Feature fusion and answer prediction

After obtaining the question text features, image features, and image OCR token feature processed by the attention module, we need to fuse these three features. Before fusion, we use MLP to process these three features to get the final features.

$$\begin{cases} MLP(X) = FC_{2d}^d \circ Relu \circ FC_d^d(X) \\ \alpha = Softmax\left(MLP(X^L)\right) \\ V = \sum_{i=1}^{M} \alpha_i x_i \\ \beta = Softmax\left(MLP(Y^L)\right) \\ Q = \sum_{i=1}^{N} \beta_i y_i \\ \gamma = Softmax\left(MLP(O^L)\right) \\ OCR = \sum_{i=1}^{P} \gamma_i o_i \end{cases} \tag{15}$$

where FC_{2d}^d, $FC_d^d()$ are fully connected layers. Vector $\alpha = [\alpha_1, \alpha_2, ..., \alpha_M] \in \mathbb{R}^M$, vector $\beta = [\beta_1, \beta_2, ..., \beta_N] \in \mathbb{R}^N$ and vector $\gamma = [\gamma_1, \gamma_2, ..., \gamma_N] \in \mathbb{R}^N$.

Finally, we obtain the fusion feature:

$$r = LayerNorm(W_v^T V + W_q^T Q + W_d^T OCR) \tag{16}$$

where vector $W_v^T, W_q^T, W_d^T \in \mathbb{R}^{d*d_z}$, $h \in \mathbb{R}^{d_z}$ is the fusion feature.

After the final fusion feature r is obtained, we use a linear classifier to classify it and then use a sigmoid function to get the predicted probability value:

$$\hat{y} = Sigmoid(W_z^T r) \tag{17}$$

where vector $W_z^T \in \mathbb{R}^{d_z * A}$.

If the model's predicted index is greater than A(the number of fixed candidate answers), we use the corresponding OCR token as the final answer.

5 Experiment

5.1 Datasets

VQA v2.0 [13] dataset is composed of natural images in MSCOCO [47], which is consistent with MSCOCO in the division of training set, verification set, and test set. It is a large-scale data set disclosed in the VQA task. The test dev model is divided into four parts, which the developers can use to test the system more flexibly, and the test dev model can

be used to prevent the developers from passing the test. Each image question pair collects ten answers and selects the answer with the most occurrences as the correct answer. There are two kinds of questions in the dataset: open and multi-topic. This paper focuses on the open task.

5.2 Experimental setup

The basic setup of the experiment follows MCAN. The maximum number of detected OCR token P is 50.

5.3 Ablation experiment

Based on the MCAN [37] model, we conducted the following ablation experiments:

- MCAN: denotes benchmark model.
- MCAN+Q-GA(Rosetta OCR): indicates MCAN with question-guide OCR token attention (Q-GA) unit and the OCR token in the image is extracted by Rosetta modle.
- MCAN+Q-GA(Paddle OCR): refers to that extracts image OCR token based on the Paddle model and introduces question-guided image OCR token attention.
- MCAN+Q-GA(Rosetta Meaningful OCR): refers to that when the OCR token is obtained, some characters without practical significance are filtered.
- MCAN+Q-GA(Rosetta OCR)+V-GA(Rosetta OCR): refers to introduces question-guided OCR token attention and image-guided OCR token attention.

The ablation results are shown in Table 1. The MCAN model is used as the benchmark in the first row. In the second row, after introducing the question-guide OCR token attention (Q-GA) unit, the accuracy is improved by 0.38% compared with the benchmark model, indicating that the Q-GA module is effective.

The Paddle model is used to extract the image OCR token in the third row, and the accuracy is improved by 0.26% compared to the benchmark model. It demonstrates that different OCR models have varying effects on our model.

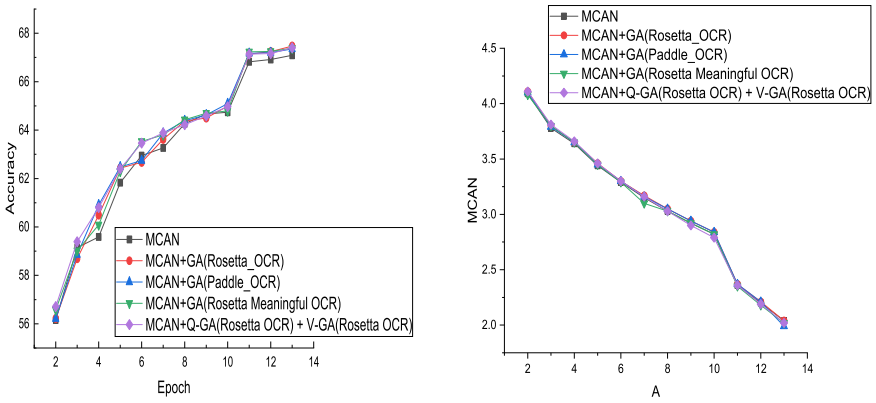
In the fourth line, the Rosetta model is used to extract the image OCR token of the image. Some characters without practical significance are filtered when the OCR token is obtained. But the accuracy still needs to be improved.

In the fifth line, the question-guide OCR token attention (Q-GA) unit and image-guide OCR token attention (V-GA) unit are added based on the MCAN, compared with MCAN, 0.33% improves the accuracy. That means the image-guide OCR token attention (V-GA) unit didn't promote.

The accuracy and loss curves during training are depicted in Fig. 4.

Table 1 The results of ablation experiment on VQA 2.0 val set

Model	Y/N(%)	Num(%)	Other(%)	Overall(%)
MCAN	84.88	48.91	58.37	67.09
MCAN+Q-GA(Rosetta OCR)	84.96	50.74	58.60	67.47
MCAN+Q-GA(Paddle OCR)	50.18	51.23	58.48	67.35
MCAN+Q-GA(Rosetta Meaningful OCR)	84.97	58.60	50.03	67.38
MCAN+Q-GA(Rosetta OCR) + V-GA(Rosetta OCR)	84.84	50.74	58.55	67.40



(a) the accuracy curve of the ablation model (b) the loss curve of the ablation model

Fig. 4 the accuracy and loss curve of the ablation model on VQA 2.0 val set

We randomly selected four demos in Fig. 5 to demonstrate the effects of our model. The MCAN model can not answer the question related to image tokens at all through these four examples. Compared with the MCAN model, our model can identify whether the question is related to the tokens in the image and can copy one from the tokens in the image as the answer. The following two wrong examples show that when more tokens are involved in the image, Our model can not well judge which token to choose as the answer and can not deal with the situation where two tokens need to be answered.

5.4 Comparison with current main models

Table 2 shows the comparison results. Our model is compared with the existing Bottom-Up, MuRel, MFH, MCAN, DFAF, DMBA-NET and MDFNet. It can be concluded from Table 2 that the model proposed in this paper is superior to other models.

BUTD [1] is the champion model of the 2017 VQA challenge. The region features based on bottom-up attention are extracted for the first time, and our model is improved by 5.68%. To improve the reasoning ability of the model, MRA-NET [24] combines the relationship between text and vision, and 1.89% improves our model. MCAN [37] and DFAF [12]

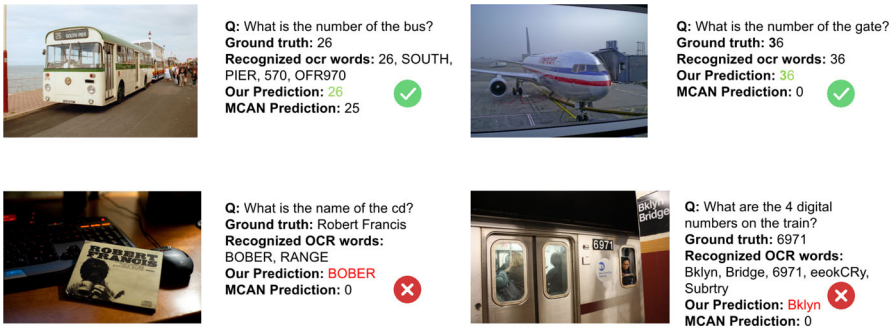


Fig. 5 Some typical examples of our model prediction

Table 2 Accuracy of single model on VQA v2.0 test-dev and test-standard dataset

Method	test-dev(%)				test-std(%)			
	Y/N	Num	Other	Overall	Y/N	Num	Other	Overall
Bottom-Up [1]	81.82	44.21	56.05	65.32	82.20	43.90	56.26	65.67
MuRel [7]	84.77	49.84	57.85	68.03	-	-	-	68.41
MRA-NET [24]	85.58	48.92	59.46	69.02	85.83	49.22	59.86	69.46
MCAN [37]	86.82	53.26	60.72	70.63	-	-	-	70.9
DFAF [12]	86.73	52.92	61.04	70.59	-	-	-	70.81
DMBA-NET [33]	87.55	51.15	60.72	70.69	87.81	50.26	60.79	70.85
MDFNet [41]	86.85	53.73	61.78	71.19	-	-	-	71.32
OECA-Net(Our)	86.92	54.83	61.07	71.01	87.18	54.45	61.44	71.35

explore the attention mechanism within and between modals. Compared with the two models, our model is improved by 0.45% and 0.54%, respectively. MDFNet [41] proposes graphical reasoning and fusion layer (GRFL) to infer the complex spatial and semantic relationships between visual objects and adaptively fuse the two relationships, 0.03% improving our model.

6 Conclusions

In this paper, we designed a deep co-attention model that is able to fuse the scene text information in images, thus, equipping the model with the ability to read and understand the text in images, and compared to other models of similar direction, our model is able to deal with datasets that use general, rather than text-oriented specific VQA datasets. Through validation on the VQA 2.0 dataset, it is found that our model is capable of answering general questions as well as questions involving text in scenes in images. As a result, our model is more generalizable.

Author Contributions For this research, Y.F. and W.S. designed the concept of the research; Y.F.and C.Y. implemented experimental design; Y.F. conducted data analysis; Y.F. wrote the draft paper; W.S. and Y.L. reviewed and edited the whole paper.

Funding This work was supported in part by the National Natural Science Foundation of China under Grant U1911401 and Key Project of Science and Technology Innovation 2030 supported by the Ministry of Science and Technology of China under Grant ZDI135-96.

Data Availability The public dataset VQA 2.0 used in this paper can be found here: https://visualqa.org/download.html (access on 13 Feb 2023). We use VQA challenge website (https://eval.ai/challenge/830/overview) to evaluate the scores on test-dev or test-std split. The link of the experiment results is as follows: https://evalai.s3.amazonaws.com/media/submission_files/submission_202957/2aa0cb55-7cb9-4505-8cd8-37ca0382ff45.json

Code Availability The current version of the code is available at https://github.com/yanfeng918/openvqa-ocr-softcopy

Declarations

Ethics approval This article does not contain any studies with human participants or animals performed by any of the authors.

Consent for Participate Not Applicable.

Consent for Publication All authors have read and agreed to the published version of the manuscript.

Conflict of Interests The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

1. Anderson P, He X, Buehler C, Teney D, Johnson M, Gould S, Zhang L (2018) Bottom-up and top-down attention for image captioning and visual question answering. In: Proceedings of the IEEE Conference on computer vision and pattern recognition, pp 6077–6086
2. Antol S, Agrawal A, Lu J, Mitchell M, Batra D, Zitnick CL, Parikh D (2015) Vqa: Visual question answering. In: Proceedings of the IEEE International conference on computer vision, pp 2425–2433
3. Ba JL, Kiros JR, Hinton GE (2016) Layer normalization. arXiv:1607.06450
4. Ben-Younes H, Cadene R, Cord M, Thome N (2017) Mutan: Multimodal tucker fusion for visual question answering. In: Proceedings of the IEEE International conference on computer vision, pp 2612–2620
5. Bojanowski P, Grave E, Joulin A, Mikolov T (2017) Enriching word vectors with subword information. Trans Assoc Computat Linguistics 5:135–146
6. Borisyuk F, Gordo A, Sivakumar V (2018) Rosetta: Large scale system for text detection and recognition in images. In: Proceedings of the 24th ACM SIGKDD International conference on knowledge discovery & data mining, pp 71–79
7. Cadene R, Ben-Younes H, Cord M, Thome N (2019) Murel: Multimodal relational reasoning for visual question answering. In: Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition, pp 1989–1998
8. Chen C, Han D, Chang C-C (2022) Caan:Context-aware attention network for visual question answering. Pattern Recogn 132:108980
9. Chen Y-C, Li L, Yu L, El Kholy A, Ahmed F, Gan Z, Cheng Y, Liu J (2020) Uniter: Universal image-text representation learning. In: Computer Vision–ECCV 2020: 16th European conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX, pp 104–120. Springer
10. Chen K, Wang J, Chen L-C, Gao H, Xu W, Nevatia R (2015) Abc-cnn:, An attention based convolutional neural network for visual question answering. arXiv:1511.05960
11. Fukui A, Park DH, Yang D, Rohrbach A, Darrell T, 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. Association for computational linguistics, ???. https://doi.org/10.18653/v1/d16-1044
12. Gao P, Jiang Z, You H, Lu P, Hoi SC, Wang X, Li H (2019) Dynamic fusion with intra-and inter-modality attention flow for visual question answering. In: Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition, pp 6639–6648
13. Goyal Y, Khot T, Summers-Stay D, Batra D, Parikh D (2017) Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In: Proceedings of the IEEE Conference on computer vision and pattern recognition, pp 6904–6913
14. Guo M-H, Xu T-X, Liu J-J, Liu Z-N, Jiang P-T, Mu T-J, Zhang S-H, Martin RR, Cheng M-M, Hu S-M (2022) Attention mechanisms in computer vision: a survey. Computational Visual Media, pp 1–38
15. Hochreiter S, Schmidhuber J (1997) Long short-term memory. Neural Comput 9(8):1735–1780
16. Hu R, Singh A, Darrell T, Rohrbach M (2020) Iterative answer prediction with pointer-augmented multimodal transformers for textvqa. In: Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition, pp 9992–10002
17. Jia S, Zhang Y (2018) Saliency-based deep convolutional neural network for no-reference image quality assessment. Multimed Tools Appl 77:14859–14872
18. Jiang X, Yu J, Qin Z, Zhuang Y, Zhang X, Hu Y, Wu Q (2020) Dualvd: an adaptive dual encoding model for deep visual understanding in visual dialogue. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol 34, pp 11125–11132
19. Kim J-H, Jun J, Zhang B-T (2018) Bilinear attention networks. arXiv:1805.07932
20. Kim J-H, On K-W, Lim W, Kim J, Ha J-W, Zhang B-T (2016) Hadamard product for low-rank bilinear pooling. arXiv:1610.04325

21. Li J, Li D, Xiong C, Hoi S (2022) Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: International conference on machine learning, pp 12888–12900. PMLR
22. Lu J, Batra D, Parikh D, Lee S (2019) Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. Advances in neural information processing systems, p 32
23. Manmadhan S, Kovoor BC (2020) Visual question answering: a state-of-the-art review. Artif Intell Rev 53(8):5705–5745
24. Peng L, Yang Y, Wang Z, Huang Z, Shen HT (2020) Mra-net: Improving vqa via multi-modal relation attention network. IEEE Transactions on Pattern Analysis and Machine Intelligence
25. Pennington J, Socher R, Manning CD (2014) Glove: Global vectors for word representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp 1532–1543
26. Shahi TB, Sitaula C (2021) Natural language processing for nepali text: a review. Artif Intell Rev, pp 1–29
27. Singh A, Natarajan V, Shah M, Jiang Y, Chen X, Batra D, Parikh D, Rohrbach M (2019) Towards vqa models that can read. In: Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition, pp 8317–8326
28. Tan H, Bansal M (2019) Lxmert:, Learning cross-modality encoder representations from transformers. arXiv:1908.07490
29. Teney D, Anderson P, He X, Van Den Hengel A (2018) Tips and tricks for visual question answering: Learnings from the 2017 challenge. In: Proceedings of the IEEE Conference on computer vision and pattern recognition, pp 4223–4232
30. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L, Polosukhin I (2017) Attention is all you need. arXiv:1706.03762
31. Vinyals O, Toshev A, Bengio S, Erhan D (2015) Show and tell: a neural image caption generator. In: Proceedings of the IEEE Conference on computer vision and pattern recognition, pp 3156–3164
32. Wang P, Yang A, Men R, Lin J, Bai S, Li Z, Ma J, Zhou C, Zhou J, Yang H (2022) Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In: International conference on machine learning, pp 23318–23340. PMLR
33. Yan F, Silamu W, Li Y (2022) Deep modular bilinear attention network for visual question answering. Sensors 22(3):1045
34. Yan F, Silamu W, Li Y, Chai Y (2022) Spca-net: a based on spatial position relationship co-attention network for visual question answering. The Vis Comput 38(9-10):3097–3108
35. Yang Z, He X, Gao J, Deng L, Smola A (2016) Stacked attention networks for image question answering. In: Proceedings of the IEEE Conference on computer vision and pattern recognition, pp 21–29
36. Yu D, Fu J, Tian X, Mei T (2019) Multi-source multi-level attention networks for visual question answering. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM) 15(2s):1–20
37. Yu Z, Yu J, Cui Y, Tao D, Tian Q (2019) Deep modular co-attention networks for visual question answering. In: Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition, pp 6281–6290
38. Yu Z, Yu J, Xiang C, Fan J, Tao D (2018) Beyond bilinear: Generalized multimodal factorized high-order pooling for visual question answering. IEEE Trans Neural Netw Learn Syst 29(12):5947–5959. https://doi.org/10.1109/tnnls.2018.2817340
39. Zhang S, Chen M, Chen J, Zou F, Li Y-F, Lu P (2021) Multimodal feature-wise co-attention method for visual question answering, vol 73
40. Zhang Y, Hutchinson P, Lieven NA, Nunez-Yanez J (2020) Remaining useful life estimation using long short-term memory neural networks and deep fusion. IEEE Access 8:19033–19045
41. Zhang W, Yu J, Wang Y, Wang W (2021) Multimodal deep fusion for image question answering. Knowl-Based Syst 212:106639