Multi-Attention Relation Network for Figure Question Answering

Ying Li, Qingfeng Wu(✉), and Bin Chen

School of Informatics, Xiamen University, Xiamen 361005, China
24320191152542@stu.xmu.edu.cn, qfwu@xmu.edu.cn

Abstract. Figure question answering (FQA) is proposed as a new multimodal task for visual question answering (VQA). Given a scientific-style figure and a related question, the machine needs to answer the question based on reasoning. The Relation Network (RN) is the proposed approach for the baseline of FQA, which computes a representation of relations between objects within images to get the answer result. We improve the RN model by using a variety of attention mechanism methods. Here, we propose a novel algorithm called Multi-attention Relation Network (MARN), which consists of a CBAM module, an LSTM module, and an attention relation module. The CBAM module first performs an attention mechanism during the feature extraction of the image to make the feature map more effective. Then in the attention relation module, each object pair contributes differently to reasoning. The experiments show that MARN greatly outperforms the RN model and other state-of-the-art methods on the FigureQA and DVQA datasets.

Keywords: Attention mechanism · Figure question answering · Relation network · Deep Learning

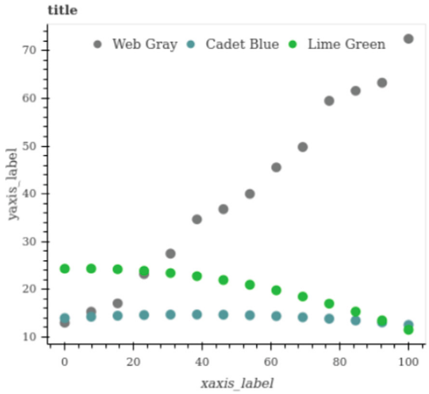
1 Introduction

Charts, such as line plots, bar graphs, pie charts, and so on, are effective and commonly used methods for presenting complex data. They exist in various text files such as academic papers and business reports and are widely used in various fields. After the machine understands the characteristics of the chart, it can help people extract relevant information from a large number of documents. Therefore, the use of computer vision to analyze chart information has high practical significance and application value. This task has only been proposed in recent years, and there are still many challenges.

Figure question answering (FQA) is an independent task of visual question answering (VQA) [1]. VQA is usually regarded as a classification problem about natural images, while FQA is to make inferences and predictions for a given chart and a related question to get the answer. Different from VQA, FQA will completely change the information of the chart even if only minor modifications are made to the image, resulting in different results [2]. In recent years, there have been many good results on FQA tasks. On the FigureQA [3] and DVQA [2] datasets, the model relational network (RN) [4]

G. Memmi et al. (Eds.): KSEM 2022, LNAI 13369, pp. 667–680, 2022.
https://doi.org/10.1007/978-3-031-10986-7_54

performs well, but the RN model still has some shortcomings, such as the limita-tion of the information extracted from images. Figure 1 is an example of a graph type with questions and answers pairs on the FigureQA dataset.



Q: Does Web Gray have the maximum area under the curve?
A: Yes

Q: Does Cadet Blue have the minimum area under the curve?
A: Yes

Q: Is Web Gray the roughest?
A: Yes

Q: Does Lime Green have the lowest value?
A: Yes

Q: Is Lime Green less than Web Gray?
A: No

Fig. 1. Dot-line graph with question answer pairs.

In this paper, we present a novel algorithm called Multi-attention Relation Network (MARN) to improve the performance of the FQA task. In the image encoder, we use Convolutional Block Attention (CBAM) module [5] to calculate the attention map of the feature map from channel and space dimensions and then multiply the attention map with the feature map to carry out adaptive feature learning. And we propose an attention-driven relation module to filter useful object pairs. The experimental results show that our proposed method achieves better performance than the RN model and other state-of-the-art methods.

The key contributions of this paper are:

1. We propose a novel method called MARN to improve the performance of the FQA task. MARN surpasses the RN model and most existing methods on the FigureQA and DVQA datasets.
2. To obtain more useful image features, we use the CBAM to calculate the attention map of the feature map obtained by CNN. In this way, the feature map can be more effective to represent the features of the image information.
3. To make the relation features more concise, we propose a novel attention-driven relation module to give each object pair a different weight. A larger weight indicates a larger impact on reasoning.

2 Related Work

In recent years, many VQA datasets [1, 6, 7] and VQA methods [8, 9] have been proposed. Due to the difference between FQA and VQA, the algorithm proposed for VQA is not

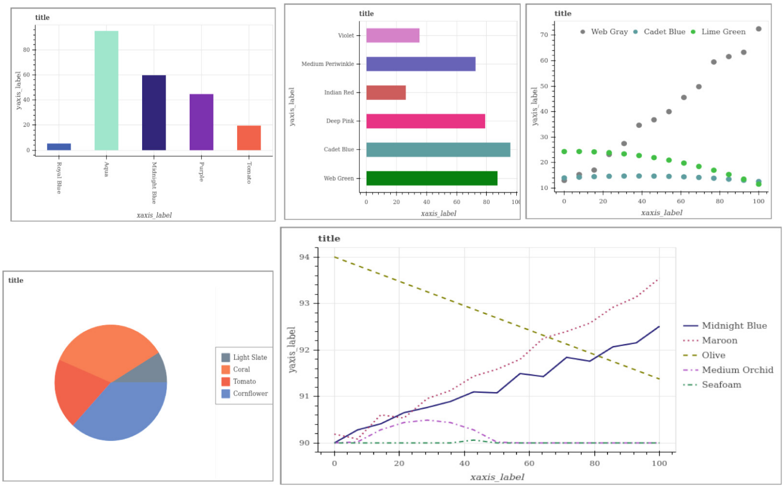


Fig. 2. The example of five types of charts in FigureQA database: vertical and horizontal bar graphs, line plots, dot-line plots, and pie charts.

suitable for figure question answering. For example, a small change in a natural image usually only affects a local area, and because the chart information is concise, even a small change will change the entire image information.

2.1 FQA Dataset

Many datasets have been proposed to study FQA tasks, such as FigureQA, DVQA, PlotQA [10], and LEAF-QA [11]. Because only the data of FigureQA and DVQA datasets is open source, we verify the performance of our model on these two datasets.

Table 1. The detailed information on FigureQA dataset.

Dataset split	# Images	# Questions	Has answers & annotations?	Color scheme
Train	100,000	1,327,368	Yes	Scheme 1
Validation 1	20,000	265,106	Yes	Scheme 1
Validation 2	20,000	265,798	Yes	Scheme 2
Test1	20,000	265,024	No	Scheme 1
Test2	20,000	265,402	No	Scheme 2

The FigureQA dataset is a synthetic corpus, which contains more than one million questions and answer pairs of more than 100,000 images for visual reasoning. It contains 5 forms of charts: line plots, dot-line plots, vertical and horizontal bar graphs, and pie charts, as shown in Fig. 2. The drawing elements of FigureQA are color-coded, with 100 unique colors. These colors are divided into the two-colored scheme: scheme 1 and scheme 2. Each scheme has 50 colors and does not overlap with each other. The dataset is divided into five separate packages, including one training set, two validation sets, and two testing sets. These packages differ by train/validation/test split and the color-to-figure assignment scheme that is used. Each drawing type in the data set has the corresponding question and answer pairs and some bounding boxes. The detailed information on the training set and two validation sets is shown in Table 1.

The DVQA dataset is a large open-source statistical graph question and answer data set proposed by Kushal Kafle and others in cooperation with the Adobe Research Laboratory. It contains a training set and two test sets (Test-Familiar and Test-Novel). The training set consists of 200,000 images and 2325316 questions, the Test-Familiar test set consists of 50,000 images and 580557 questions, and the Test-Novel test set consists of 50,000 images and 581321 questions. There are three types of questions in the DVQA data set: the first type is Structure Understanding, which is about the understanding of the overall structure of an image, the second type is Data Retrieval, and the third is reasoning. The reasoning is a type of reasoning problem about the relationship between elements in the image.

2.2 Existing FQA Algorithms

At present, four types of basic algorithms are proposed based on the FigureQA dataset. They are Text-only baseline, CNN+LSTM, CNN+LSTM on VGG-16 features, and Relation Network (RN), of which RN has the best effect. The Text-only baseline is a text-only model, which is trained with batch size 64. CNN+LSTM uses the MLP classifier to connect the received LSTM to generate the problem code and the learned visual representation of the CNN with five convolutional layers. And CNN+LSTM on VGG-16 features extracts features from the fifth-layer pool of the ImageNet pre-trained VGG-16 network [12]. RN is currently the best baseline method, a simple and powerful neural module for relational reasoning. RN is proven to have the most advanced performance on a challenging data set called CLEVR [13].

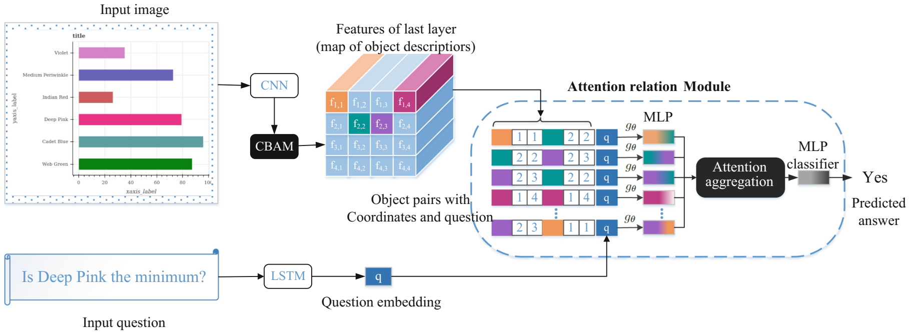

Input image
title
Violet
Medium Periwinkle
Indian Red
Deep Pink
Cadet Blue
Web Green
yaxis_label
0
20
40
60
80
100
xaxis_label
CNN
CBAM
Features of last layer
(map of object descriptors)
$f_{1,1}$ $f_{1,2}$ $f_{1,3}$ $f_{1,4}$
$f_{2,1}$ $f_{2,2}$ $f_{2,3}$ $f_{2,4}$
$f_{3,1}$ $f_{3,2}$ $f_{3,3}$ $f_{3,4}$
$f_{4,1}$ $f_{4,2}$ $f_{4,3}$ $f_{4,4}$
Attention relation Module
MLP
g_θ
q
Attention aggregation
MLP classifier
Yes
Predicted answer
Object pairs with Coordinates and question
Is Deep Pink the minimum?
LSTM
q
Question embedding
Input question

Fig. 3. The framework of our proposed Multi-attention Relation Network (MARN).

Recently, many novel methods have been proposed for FQA tasks. The FigureNet [14] model proposes a multi-module algorithm framework to solve the question and answer of statistical graphs. The LEAF-Net [11] model uses several open-source pre-training models. First, the character information in the image is recognized through OCR, and then it is located in the problem for embedding. At the same time, the image feature map is obtained through the pre-trained ResNet-152. The ARN [15] model is a relation network framework algorithm. It first recognizes the elements, characters, structure, and other information of the image through multiple recognition modules, then constructs it into the form of a table through the obtained information, and finally passes a form question and answer model to get the answer.

2.3 Attention Mechanism

It is well known that attention plays an important role in human perception [16, 17]. An important feature of the human visual system is that people don't try to process the whole scene at the same time. On the contrary, to better capture the visual structure, humans use a series of partial glimpses and selectively focus on the salient [18].

Lately, several attempts incorporate attention processing to improve the performance of CNNs in large-scale classification tasks. Wang et al. [19] propose a Residual Attention Network that uses an encoder-decoder style attention module. Hu et al. [20] introduce a compact module to exploit the inter-channel relationship.

3 Methods

In this section, we give a detailed introduction to the proposed method, namely, Multi-attention Relation Network (MARN), as shown in Fig. 3. In the image representation, we employ a Convolutional Block Attention (CBAM) module on the feature map obtained by CNN to capture the more effective information. The final feature map is denoted as F. In the question representation, we apply an LSTM module to convert the text content into a low-dimension embedding. We regard the hidden state as the question representation, denoted as q. At last, F and q feed into the attention relation module to get the result.

3.1 Image Representation

In the image decoder, the feature map firstly comes to form a CNN with five convolutional layers, each with 64 kernels of size 3 × 3, stride 2, zero-padding of 1 on each side, and batch normalization [21]:

$$F_i = ConvBlock(F_{i-1}), F_0 = I \tag{1}$$

where F_0 is the original image I and $F_i\{1 \leq i \leq 5\}$ denote the feature map after the ith convolutional layer. After the CNN module, the feature map $F_5 \in R^{H*W*C}, H = 8, W = 8, C = 64$ is obtained. H is the height of the feature map, W is the width, and C denotes the number of channels.

To make the feature map more effectively show the information of image I, we apply the CBAM to the feature map F_5 after CNN. The framework of the CBAM module is shown in Fig. 4. It can be seen that the CBAM module is divided into two steps: the channel attention module and the spatial attention module.

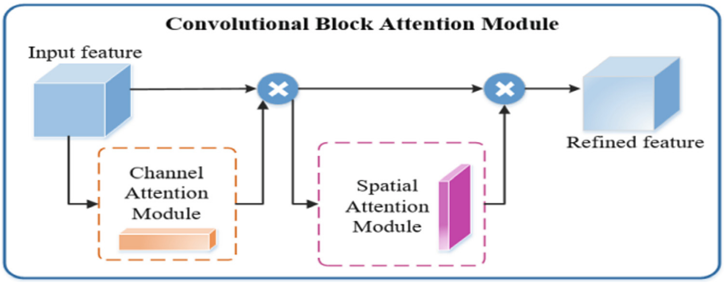


Fig. 4. CBAM module.

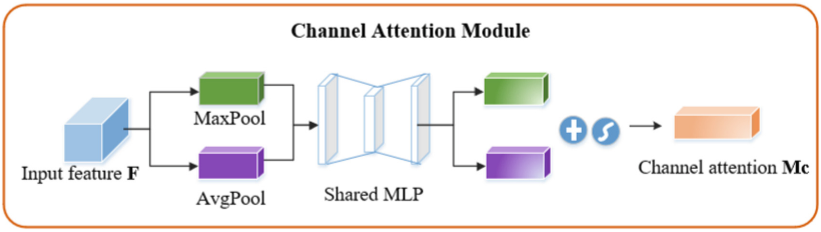


Fig. 5. Channel attention module.

The channel attention module focuses on which channels in the feature map F_5 are more useful. Figure 5 shows the structure of the channel attention module. We use max pooling and average pooling to compress the feature map in the spatial dimension and get two different spatial background description vectors F_{max}^c and $F_{avg}^c \in R^{1*1*C}$:

$$F_{max}^c = MaxPool(F_5) \tag{2}$$

$$F_{avg}^c = AvgPool(F_5) \tag{3}$$

Then for F_{max}^c and F_{avg}^c, shard MLP is used to calculate the channel attention map $M_c \in R^{1*1*C}$:

$$M_c = \sigma(W_1(W_0(F_{max}^c)) + W_1\left(W_0\left(F_{avg}^c\right)\right)) \tag{4}$$

where σ is the sigmoid function, and $W_0 \in R^{\frac{C}{r}*C}$, $W_1 \in R^{C*\frac{C}{r}}$, r is the reduction ratio. Then, F_5 is multiplied by M_c to get $F^{'}$:

$$F^{'} = F_5 * M_c \tag{5}$$

where * denotes the element-wise multiplication.

The spatial attention module focuses on location information (where). Figure 6 shows the structure of the channel attention module. This time, we use max pooling and average pooling to compress the feature map in the channel dimension and get two vectors: F^s_{max} and $F^s_{avg} \in R^{H*W*1}$:

$$F^s_{max} = MaxPool\left(F^{'}\right) \tag{6}$$

$$F^s_{avg} = AvgPool\left(F^{'}\right) \tag{7}$$

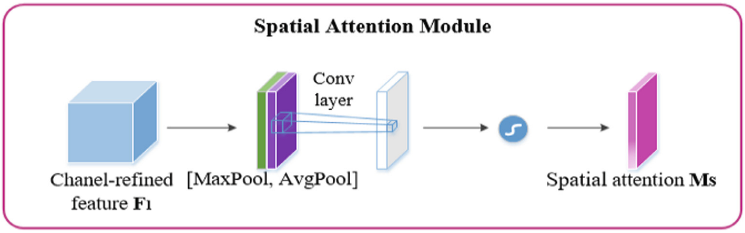


Fig. 6. Spatial attention module.

The two vectors are merged by concatenation and use the convolutional layer to generate the spatial attention map $M_s \in R^{H*w*1}$:

$$M_s = \sigma(ConvBlock2(\left[F^s_{max}; F^s_{avg}\right])) \tag{8}$$

where ConvBlock2 represents the 7 × 7 convolutional layer. The final feature map $F \in R^{H*W*C}$ after the image decoder is obtained by:

$$F = F^{'} * M_s \tag{9}$$

3.2 Question Representation

Like the RN model, we first combine all the existing words into a dictionary. Then each question can be expressed as $Q = [x_1, \ldots, x_T]$, where x_t represents the vector after the one-hot encoding in the dictionary, and T is the length of the question. We apply the simple unidirectional LSTM with 512 hidden units to obtain the hidden state:

$$h_t = LSTM(x_t), 1 \leq t \leq T \tag{10}$$

We regard the hidden state in the last step as the question representation, i.e., $q = h_T \in R^{512}$.

3.3 Attention Relation Module

The core idea of RN is to regard the feature map F of the image as a set of objects and combine two objects into a pair and then connect the question representation vector q after each pair.

For a feature map of size n × n, the object set can be denoted as $F = \{f_{i,j} | 1 \leq i, j \leq n\} \in R^{64*64}$, where $f_{i,j} \in R^{64}$ denotes the ith row and jth column of the feature map F, and n = H = W = 8. Then the set of all object pairs is represented as:

$$P = \{p_{(i,j),(u,v)} | 1 \leq i, j, u, v \leq n\} \tag{11}$$

where $p_{(i,j),(u,v)}$ is the concatenation of the corresponding object vectors, their location information, and the question vector q, i.e., $p_{(i,j),(u,v)} = [f_{i,j}, i, j, f_{u,v}, u, v, q] \in R^{644}$. Each object is paired with all objects including itself, i.e., $p_{(1,1),(1,1)}, p_{(1,1),(1,2)}, \ldots, p_{(n,n),(n,n)}$. Then $P \in R^{4096*644}$ is the matrix containing 4096 object pairs representations.

Then, every object pairs are separately processed by MLPs to produce a feature representation $o_{(i,j),(u,v)}$ of the relation between the corresponding objects:

$$O = \{o_{(i,j),(u,v)} = g_{\theta}(p_{(i,j),(u,v)}) | 1 \leq i, j, u, v \leq n\} \tag{12}$$

where g_{θ} is implemented as MLPs.

Then the RN model is to average all feature O to get the final result. But we think that the contribution of the generated feature $o_{(i,j),(u,v)}$ of each object pairs to the final result is different, so we propose a new aggregation method, namely, attention aggregation. The structure of attention aggregation is shown in Fig. 7.

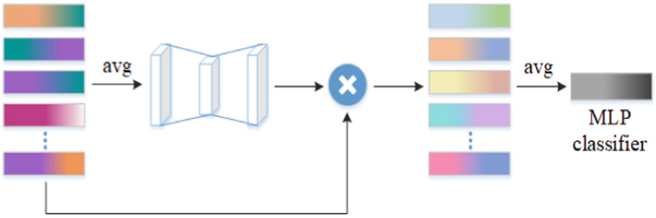


Fig. 7. Attention aggregation.

As shown in Fig. 7, we average each feature $o_{(i,j),(u,v)}$ and splice it into a vector $A \in R^{4096}$. Then, through two layers of MLP, we get an attention map M_a to represent the contribution of each object pair feature to the final result. At last, O is multiplied by M_a to get the attention-based object pair features O_a.This process can be formulated as:

$$M_a = \sigma(W_1^{'}\left(W_0^{'}(O)\right)) \tag{13}$$

$$O_a = O * M_a \tag{14}$$

where $W_0^{'}$ and $W_1^{'}$ are the weight matrixes of two MLPs.

At last, the sum over all relational features O_a is then processed by MLP, yielding the predicted outputs:

$$Label = f_{\varphi}(\frac{1}{N^2}\sum_{i,j,u,v} o_{a(i,j),(u,v)} \tag{15}$$

where f_φ is implemented as MLP, and N = 64 is the number of all objects. The classifier f_φ has two hidden layers with 256 ReLU units and the second layer applies the dropout with a rate of 50%.

4 Experiments

4.1 Experimental Setting

We evaluate our proposed MARN model on the FigureQA and DVQA datasets. For FigureQA, we use the train data as the training set and the validation2 data as the validation set. Then we verify the effectiveness of our model on validation1 and validation2 data because test sets are non-publicly available. For DVQA, we use two versions to verify our model: the method without a dynamic dictionary (No OCR) and the method with Oracle Version (Oracle).

The accuracy is adopted as the evaluation metric. The images in the dataset are resized into a size of 256 × 256. For data augmentation, each image is padded to 264 × 264 and then is randomly cropped back to 256 × 256. And at each training step, we compute the accuracy of one randomly selected batch from the validation set and keep an exponential moving average with a decay of 0.9. Starting from the 100th update, we perform early-stopping using this moving average. The train batch size is 160, and the validation batch size is 64. We trained our model in 350000 steps. Our model is trained using the Adam optimizer with a learning rate of 1.25e-4.

4.2 Experimental Results

In this section, we show the comparison of our model with other methods both on FigureQA and DVQA datasets. Table 2 shows our results and comparison with other methods on the FigureQA dataset, such as QUES [5] (Text only), IMG+QUES [5] (CNN+LSTM), RN [5], FigureNet [14], LEAF-Net [11] and ARN [15]. It should be noted that FigureNet carries out the experiments only on three types of charts, i.e., vBar, hBar, and Pie, in the Validation1 data of the FigureQA dataset. Then Table 3 shows the performance of our model compared with other methods on two versions, such as SANDY (No OCR+ Oracle) [2], ARN (No OCR+Oracle) [15], and LEAF-Net (Oracle) [11].

In FigureQA, our approach significantly outperforms the RN baseline model both on two validation sets. Specifically, our method obtains an accuracy promotion of approximately 8.04%, 10.78%,4.37%, 10.92%, and 11.58% in the five types of charts of validation1 data and 14.06%, 13.49%, 8.47%, 9.45%, and 13.23% in the five types of charts of validation2 data, respectively. Overall, the accuracy of our model is improved by 9.40% on validation1 data and 11.24% on validation2 data compared with the RN model. Then compared with FigureNet, MARN obtains a promotion of approximately 6.39%, 10.78%, and 3.80% in the vBar, hBar, and pie charts of validation1 data. LEAF-Net reaches an accuracy of about 81.15% on validation2 data, while our approach performs slightly better, about 83.78%. At last, we also improve the accuracy by 0.31% and 0.83% on both validation sets compared with ARN. In addition, we also found that the accuracy

Table 2. Results comparisons with other methods on FigureQA dataset.

	Validation 1 -same colors						Validation 2 -alternated colors					
	vBar	hBar	Pie	Line	Dot-line	Overall	vBar	hBar	Pie	Line	Dot-line	Overall
QUES	–	–	–	–	–	–	–	–	–	–	–	50.01
IMG+QUES	61.98	62.44	59.63	57.07	57.35	59.41	58.60	58.05	55.97	56.37	56.97	57.14
RN	85.71	80.60	82.56	69.53	68.51	76.39	77.35	77.00	74.16	67.90	69.40	72.54
FigureNet	87.36	81.57	83.13	–	–	–	–	–	–	–	–	–
LEAF-Net	–	–	–	–	–	–	–	–	–	–	–	81.15
ARN	92.49	91.20	84.25	**81.31**	**81.03**	85.48	90.46	89.56	80.29	**77.60**	78.28	82.95
MARN(Ours)	**93.75**	91.38	**86.93**	80.45	80.09	**85.79**	**91.41**	**90.49**	**82.63**	77.35	**82.63**	**83.78**

Table 3. Results comparisons with other methods on DVQA dataset.

	Test-familiar	Test-novel
IMG+QUES	32.01	32.01
SANDY (No OCR)	36.02	36.14
ARN (No OCR)	44.50	44.51
MARN (ours)	**45.10**	**44.97**
SANDY (Oracle)	56.48	56.62
LEAF-Net (Oracle)	72.72	72.89
ARN (Oracle)	79.43	79.58
MARN (Oracle)	**79.96**	**80.23**

of our model in the bar and pie charts is significantly higher than that in the line charts. After analysis, this may be because line charts often contain more complex information and the questions are quite more difficult to answer, so the accuracy of line charts is relatively low.

On DVQA, first, compare the performance of the No OCR version, it can find that our method achieves the best results, which achieves higher accuracy of 13.09% and 12.86% than the IMG+QUES baseline and 0.6% and 0.46% higher than ARN in the two verification sets, respectively. In the Oracle version, we also achieve the best performance compared with these three methods on both two test sets.

In general, our method greatly improves the performance of the original Relation Network (RN) model and proves the effectiveness of introducing the attention mechanism into the RN model both on FigureQA and DVQA datasets.

4.3 Ablation Study on Modules of Our Approach

We conduct an ablation study on the FigureQA dataset to explore the importance of each module in MARN. The detailed results are shown in Table 4, where RN+CBAM indicates that the RN model adds the CBAM module in the image encoding module, and RN+attention aggregation indicates we use the attention aggregation module in the RN model.

Table 4. Results for ablation studies on our method.

Ablation model	Val 1	Val 2
RN+CBAM	81.45	79.65
RN+attention aggregation	80.34	79.89
MARN (Full model)	85.79	83.78

Table 5. Accuracy comparison per question type on the validation1 and validation2.

Template	Validation 1	Validation 2
Is X the minimum?	92.45	89.38
Is X the maximum?	96.71	94.41
Is X less than Y? (bar,pie)	97.92	95.46
Is X greater than Y? (bar,pie)	97.99	95.53
Is X the low median?	78.42	7573
Is X the high median?	80.99	78.88
Does X have the minimum area under the curve?	86.86	86.00
Does X have the maximum area under the curve?	91.65	90.29
Is X the smoothest?	65.01	64.60
Is X the roughest?	6472	63.63
Does X have the lowest value?	83.38	81.19
Does X have the highest value?	88.24	86.73
Is X less than Y? (line)	81.40	79.43
Is X greater than Y? (line)	81.18	79.67
Does X intersect Y?	80.78	78.79

As shown in Table 4, the model using one of the attention modules alone performs better than the original RN model. And the relation network using multi attention (MARN) can achieve the best performance.

4.4 Comparison Per Question Type

Tables 5 show the performances of the MARN on each question type of the FigureQA dataset. There are 15 question types in the FigureQA dataset.

From Table 5, we can see that the prediction accuracy of different problem types is different. For instance, ‘Is X the minimum?’, ‘Is X the maximum?’, ‘Is X less than Y? (bar, pie)’, ‘Is X greater than Y? (bar, pie)’, and ‘Does X have the maximum area under the curve?’ these five question types all achieves more than 90% accuracy, while ‘Is X the smoothest?’, and ‘Is X the roughest?’ these two question types have an accuracy rate of only over 60%. After analysis, we conclude that this type of question like ‘Is X the smoothest?’ is more complex, and even human beings are difficult to answer.

In addition, we can find that even for the same problem, the accuracy of the model is different due to different chart types. For example, ‘Is X less than Y?’ this type of question can reach the accuracy of 97.92% on bar and pie chart, but only 81.40% in a line chart on the validation1 data. This also proves that the line charts are more difficult to answer than the bar and pie charts because they contain more complex picture information.

4.5 Comparison with Human Annotates

Table 6 shows the accuracy of the overall validation set and shows a comparison with the results of human annotations. We can find that although the accuracy of our MARN model is 11.33% higher than that of the previous RN model, there is still a certain gap compared with the accuracy of human annotation. There are still many big challenges for figure question answering to improve the performance.

Table 6. Performance of our method, other method and human annotates on full validation set.

Model	Accuracy
IMG+QUES	58.27
RN	74.46
MARN (ours)	**85.79**
Human	91.21

5 Conclusion

In this paper, we proposed a multi-attention relation network (MARN) to improve the performance of the original relation network. Our model uses the CBAM module in the image representation to make the feature map more effective. And we propose a novel attention relation module to give the different weights to the object pairs features, which can help the model to find more useful information. Our proposed MARN performs significantly better than the original relation network baseline and most state-of-the-art methods. In future work, we intend to improve the performance of these low-accuracy figure types and question types. And we will try to change the structure of the model and use more complex image and text models to improve the accuracy.

References

1. Zhou, B., Tian, Y., Sukhbaatar, S., Szlam, A., Fergus, R.: VQA: Visual question answering. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2425–2433 (2015)
2. Kafle, K., Price, B., Cohen, S., Kanan, C.: Dvqa: Understanding data visualizations via question answering. In: Proceedings of the 2018 IEEE/ CVF Conference on Computer Vision and Pattern Recognition, pp. 5648–5656. IEEE (2018)
3. Kahou, S.E., Michalski, V., Atkinson, A., Kadar, A., Trischler, A., Bengio, Y.: Figureqa: An annotated figure dataset for visual reasoning (2017). arXiv preprint arXiv:1710.07300
4. Santoro, A., Raposo, D., Barrett, D.G., Malinowski, M., Pascanu, R.: A simple neural network module for relational reasoning (2017). arXiv preprint arXiv:1706.01427
5. Woo, S., Park, J., Lee, J.-Y., Kweon, I.S.: CBAM: Convolutional block attention module. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11211, pp. 3–19. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01234-2_1
6. 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: Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6325–6334 (2017). doi: https://doi.org/10.1109/CVPR.2017.670
7. Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K.: Visual genome: connecting language and vision using crowdsourced dense image annotations. Int. J. Comput. Vis. **123**(1), 32–73 (2017)
8. Kafle, K., Kanan, C.: Answer-type prediction for visual question answering. In: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4976–4984 (2016)
9. Andreas, J., Rohrbach, M., Darrell, T., Klein, D.: Deep compositional question answering with neural module networks. Comput. Sci. **27** (2015)
10. Methani, N., Ganguly, P., Khapra M., Kumar, P.: PlotQA: Reasoning over scientific plots. In: Proceedings of the 2020 IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 1516–1525 (2020)
11. Ritwick, C., Sumit, S., Utkarsh, G., Pranav, M., Prann, B., Ajay, J.: Leaf-qa: Locate, encode and attend for figure question answering. In: Proceedings of the 2020 IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 3501–3510 (2020)
12. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: Proceedings of the International Conference on Learning Representations (2015)
13. Johnson, J., Hariharan, B., Maten, L. Fei-Fei, L.: CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. In: Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1988–1997 (2017)
14. Reddy, R., Ramesh, R.: Figurenet: A deep learning model for question-answering on scientific plots. In: Proceedings of the 2019 International Joint Conference on Neural Networks (IJCNN), pp. 1–8 (2019)
15. Jialong, Z., Guoli, W., Taofeng, X., Qingfeng, W.: An affinity-driven relation network for figure question answering. In: Proceedings of the 2020 IEEE International Conference on Multimedia and Expo (ICME), pp. 1–6 (2020)
16. Itti, L., Koch, C., Niebur, E.: A model of saliency-based visual attention for rapid scene analysis. IEEE Trans. Pattern Anal. Mach. Intell. (TPAMI) **20**, 1254–1259 (1998)
17. Rensink, R.A.: The dynamic representation of scenes. Vis. Cogn. **7**, 17–42 (2000)
18. Larochelle, H., Hinton, G.E.: Learning to combine foveal glimpses with a thirdorder Boltzmann machine. Neural Inf. Process. Syst. (NIPS) (2010)

19. Wang, F., et al.: Residual attention network for image classification. In: Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), arXiv preprint arXiv:1704.06904 (2017)
20. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. IEEE Trans. Pattern Anal. Mach. Intell. arXiv preprint arXiv:1709.01507 (2017)
21. Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: Proceedings of the International Conference on Machine Learning (2015)