Modular Graph Attention Network for Complex Visual Relational Reasoning

Yihan Zheng[1,2*], Zhiquan Wen[1*], Mingkui Tan[1], Runhao Zeng[1,2], Qi Chen[1], Yaowei Wang[2†], and Qi Wu[3]

[1] South China University of Technology, Guangzhou, China
yihanzheng7@gmail.com
sewenzhiquan@mail.scut.edu.cn
mingkuitan@scut.edu.cn
[2] PengCheng Laboratory, Shenzhen, China
wangyw@pcl.ac.cn
[3] University of Adelaide
qi.wu01@adelaide.edu.au

We organise our supplementary materials as follows. In Section 1, we provide more details about our MGA-Net. In Section 2, we draw a discussion on simple REF tasks and conduct experiments on RefCOCO [1], RefCOCO+ [1] and RefCOCOg [2] datasets. In Section 3, we provide implementation details of our MGA-Net. In Section 4, we show more visualisation examples on CLEVR-Ref+ [3] and GQA [4] to demonstrate the effectiveness of our method.

Algorithm 1 Training details of MGA-Net.

Require: Training data $\{I_k, r_k, \boldsymbol{y}_k\}_{k=1}^{K}$, the number of updating steps T in GGNNs, the number of training iterations D

1: **for** $d = 1, \ldots, D$ **do**
2:     *// Language Attention Network, type* $\in$ {*att, loc, rel_vis, rel_loc*}
3:     Calculate the word attention values $\left\{a_l^{type}\right\}_{l=1}^{L}$ using Eq. (1).
4:     Obtain the query representations $\boldsymbol{s}^{type}$ using Eq. (2).
5:     *// Object Attention Network, obj* $\in$ {*att, loc*}
6:     Calculate the object attention values $\left\{a_i^{o,obj}\right\}_{i=1}^{N}$ using Eq. (3).
7:     Obtain the object representation $\hat{\boldsymbol{x}}_i^{obj}$ using Eq. (4).
8:     *// Relational Inference Network, rel* $\in$ {*rel_vis, rel_loc*}
9:     Construct a relational graph $\mathcal{G}=\{\mathcal{V}, \mathcal{E}\}$ using Eqs. (5) and (6).
10:     **for** $t = 1, \ldots, T$ **do**
11:         Obtain the object representation $\hat{\boldsymbol{x}}_i^{rel}$ through GGNNs using Eq. (7).
12:     **end for**
13:     Calculate the matching scores p_i^{type} between $\boldsymbol{s}^{type}$ and $\hat{\boldsymbol{x}}_i^{type}$ using Eq. (10).
14:     Obtain the final scores $\boldsymbol{p}$ using Eq. (12).
15:     Update MGA-Net by minimising the loss in Eq. (13).
16: **end for**

* Authors contributed equally.
† Corresponding author.

1 More details about MGA-Net

The details of our proposed MGA-Net are shown in Algorithm 1. We first decompose the input query r_k into four types ($type \in \{att, loc, rel\_vis, rel\_loc\}$) and obtain the language representations $\boldsymbol{s}^{type}$. For each object in the input image I_k, we calculate the object attention value and obtain the object representation $\hat{\boldsymbol{x}}_i^{obj}$ under the guidance of $\boldsymbol{s}^{att}$ and $\boldsymbol{s}^{loc}$. Then, we construct a relational graph and obtain the object representation $\hat{\boldsymbol{x}}_i^{rel}$ via Gated Graph Neural Networks (GGNNs) [5] guided by $\boldsymbol{s}^{rel\_vis}$ and $\boldsymbol{s}^{rel\_loc}$. Last, the final score $\boldsymbol{p}$ is obtained by matching the object representations with the language representations. To train the model, we use the Adam optimiser to minimise the loss.

2 Discussions on Simple REF Tasks

In this paper, we focus on complex visual relational reasoning tasks. In particular, we consider complex queries which require multi-steps reasoning over a chain of visual attributes and relationships.

In some cases, however, the data may contain only short and simple queries. For example, for RefCOCO [1], more than 90% of queries contain fewer than 6 words. In this case, the complex relational reasoning is not necessary and may not significantly boost the performance [9]. Nevertheless, it is still interesting to investigate whether our method could help with simple tasks. To this end, we conduct experiments on RefCOCO, RefCOCO+ and RefCOCOg datasets and report the results in Table 1.

From these results in Table 1, compared with the state-of-the-art methods, our MGA-Net performs worse. As discussed in [9], many of the queries in RefCOCO [1,2] datasets do not require resolving relations. However, our MGA-Net performs relational reasoning step-by-step, and thus introducing some useless even noisy information, which affects the performance of our MGA-Net on RefCOCO datasets. It is worth mentioning that in real-world applications, we are facing more complex and challenging reasoning tasks rather than simple and toy ones. Note that our MGA-Net is orthogonal with previous referring expression models and we leave it as further works to combine them together to solve both short and long reasoning expressions.

		RefCOCO			RefCOCO+			RefCOCOg	
	feature	val	testA	testB	val	testA	testB	val	test
MAttNet [6]	resnet101	85.65	85.26	84.57	71.01	75.13	66.17	78.10	78.12
DGA [7]	resnet101	86.34	86.64	**84.79**	73.56	78.31	68.15	80.21	80.26
CMRIN [8]	resnet101	**86.99**	**87.63**	84.73	**75.52**	**80.93**	**68.99**	**80.45**	**80.66**
MGA-Net	resnet101	84.59	84.60	83.81	72.68	74.89	68.73	76.84	77.07

Table 1. Comparison with state-of-the-art methods on RefCOCO, RefCOCO+ and RefCOCOg when ground-truth bounding boxes are used. The best performing method is marked in bold.

3 Implementation Details

On CLEVR-Ref+ and CLEVR-CoGenT, we follow the settings in [9] and obtain the feature of each object by using ResNet101 [10] pre-trained on ImageNet [11]. To train our model, we use Adam [12] with a learning rate 1e-4. We evaluate our methods in two settings: (i) bounding boxes detected by Mask-RCNN [13];[4] (ii) ground truth bounding boxes. On GQA, we use the 2048-dimensional visual features of objects provided by the dataset and encode the queries using GloVe word embedding [14]. During training, we use Adam with the learning rate 1e-3. For all datasets, the batch size is 30, which means that we feed 30 images and all the queries associated with these images to the network for each training iteration. The updating step of GGNNs is set to 3. Following [15], we set the dimensions of the final language representations and object representations to 512. We implement our method based on PyTorch [16].

4 More visualisation examples

To demonstrate the effectiveness of our MGA-Net, we show more visualisation examples on GQA and CLEVR-Ref+. Specifically, given a query and an image with candidate bounding boxes, our MGA-Net selects the most relevant object from the candidates. As shown in Figs. A and B, our method obtains the object related to the answer (on GQA) or the referent (on CLEVR-Ref+) correctly.

In addition, we also show some failure cases of our MGA-Net on CLEVR-Ref+. As shown in Fig. C, our method makes wrong predictions. The main reason may lie in that the complex relational reasoning is not necessary for the short and simple referring expressions.

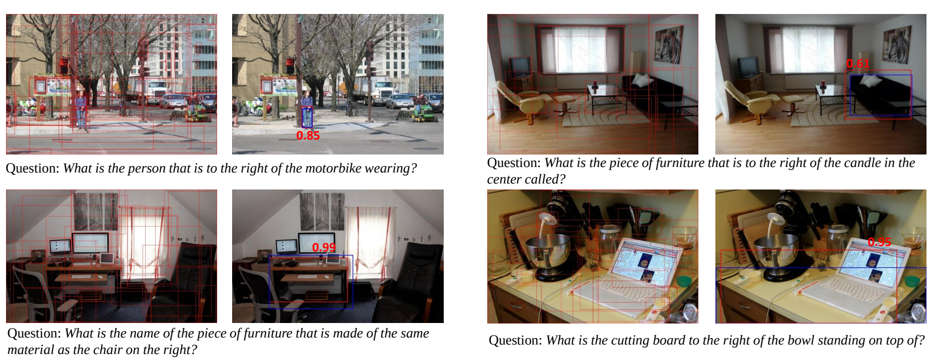

Fig. A. Visualisation examples of our MGA-Net on GQA. With the inputs of a question and an image (on the left), we obtain the object with the highest matching score (on the right). The object in the red bounding box is the prediction, while the blue bounding box corresponds to the ground truth.

[4] We take the pre-trained Mask-RCNN from https://github.com/kexinyi/ns-vqa

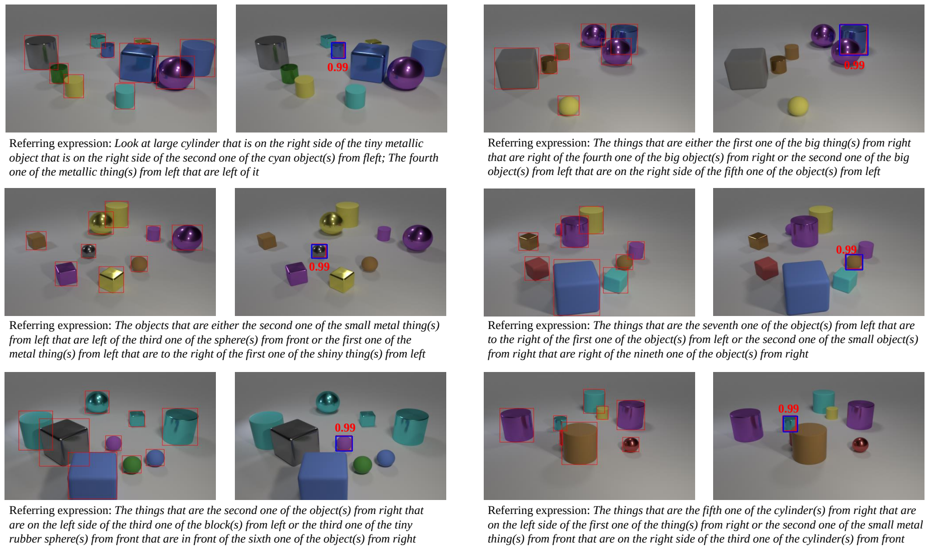

Fig. B. Visualisation examples of our MGA-Net on CLEVR-Ref+. With the inputs of a referring expression and an image (on the left), we obtain the object with the highest matching score (on the right). The object in the red bounding box is the prediction, while the blue bounding box corresponds to the ground truth.

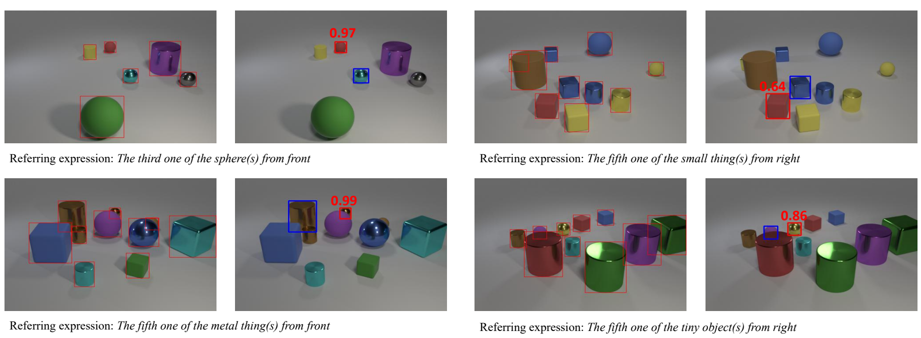

Fig. C. Some failure cases of MGA-Net on CLEVR-Ref+. With the inputs of a referring expression and an image (on the left), we obtain the object with the highest matching score (on the right). The object in the red bounding box is the prediction, while the blue bounding box corresponds to the ground truth.

References

1. Kazemzadeh, S., Ordonez, V., Matten, M., Berg, T.L.: Referitgame: Referring to objects in photographs of natural scenes. In: Proceedings of the Conference on

Empirical Methods in Natural Language Processing (EMNLP). (2014) 787–798
2. Mao, J., Huang, J., Toshev, A., Camburu, O., Yuille, A.L., Murphy, K.: Generation and comprehension of unambiguous object descriptions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2016) 11–20
3. Liu, R., Liu, C., Bai, Y., Yuille, A.L.: Clevr-ref+: Diagnosing visual reasoning with referring expressions. In: Proceedings of the of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2019) 4185–4194
4. Hudson, D.A., Manning, C.D.: GQA: A new dataset for real-world visual reasoning and compositional question answering. In: Proceedings of the of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2019) 6700–6709
5. Li, Y., Tarlow, D., Brockschmidt, M., Zemel, R.S.: Gated graph sequence neural networks. In: International Conference on Learning Representations (ICLR). (2016)
6. Yu, L., Lin, Z., Shen, X., Yang, J., Lu, X., Bansal, M., Berg, T.L.: Mattnet: Modular attention network for referring expression comprehension. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2018) 1307–1315
7. Yang, S., Li, G., Yu, Y.: Dynamic graph attention for referring expression comprehension. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). (2019) 4644–4653
8. Yang, S., Li, G., Yu, Y.: Cross-modal relationship inference for grounding referring expressions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2019) 4145–4154
9. Hu, R., Rohrbach, A., Darrell, T., Saenko, K.: Language-conditioned graph networks for relational reasoning. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). (2019) 10294–10303
10. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2016) 770–778
11. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M.S., Berg, A.C., Li, F.: Imagenet large scale visual recognition challenge. International Journal of Computer Vision **115** (2015) 211–252
12. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: International Conference on Learning Representations (ICLR). (2015)
13. He, K., Gkioxari, G., Dollár, P., Girshick, R.B.: Mask R-CNN. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). (2017) 2980–2988
14. Pennington, J., Socher, R., Manning, C.D.: Glove: Global vectors for word representation. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP). (2014) 1532–1543
15. Wang, P., Wu, Q., Cao, J., Shen, C., Gao, L., Hengel, A.v.d.: Neighbourhood watch: Referring expression comprehension via language-guided graph attention networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2019) 1960–1968
16. 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., Chintala, S.: Pytorch: An imperative style, high-performance deep learning library. In: Advances in Neural Information Processing Systems 32 (NeurIPS). (2019) 8024–8035