Graph-Symbolic VQA with Rich Visual Estimators and No Question-Answer Labels

Zhexiong Liu Adriana Kovashka
Department of Computer Science, University of Pittsburgh
{zhexiong, kovashka}@cs.pitt.edu

1. Introduction

Visual question answering (VQA) [1] is a fundamental task towards real-world AI applications, e.g. robots that learn from and respond to human decision making and reasoning. To answer a visual question, a human may first parse it into components, and reason about recognized visual estimators, their attributes and relations. In contrast, most previous approaches for algorithmic VQA primarily rely on deep representation learning [2, 3, 5] to directly enumerate answers from common representation spaces, e.g. by fusing embeddings from CNNs for images and RNNs for questions, or more recently, through multimodal transformers. While effective when sufficient training data is available [1, 9, 8], this mechanism is complex, unexplainable, and entangles multiple procedures from human reasoning.

In addition to lack of interpretability, neural-based models might suffer in-sufficiency when handling innumerable visual estimators and question phrasing variants due to large variability of real-world scenarios. To adapt to any particular scenario, supervision is required, thus constructing domain-robust models is challenging and typically requires extensive annotated data, as [19, 4, 22] show. Some researchers tackle robustness and sample efficiency [10, 14] on synthetic datasets like CLEVR [9], but applicability to real-world VQA questions was not demonstrated.

To bridge the gap between existing work on sample-efficiency in VQA research and real-world VQA data, we move towards processing question answering as a three-stage inference structure that fully disentangles vision and language understanding from reasoning. In particular, we leverage off-to-shelf image recognition modules to generate image scene graphs that capture visual estimators and their spatial and semantic subject-object-predicate relations. We construct graph representations for questions in terms of object attributes and relations, complemented with quantifiers and logical connectives. We incorporate a graph-symbolic answering executor that runs novel algorithms on the question and image scene graphs to obtain answers. Our proposed framework is fully modular, extensible, and efficient, with an ability to handle questions not specific to datasets.

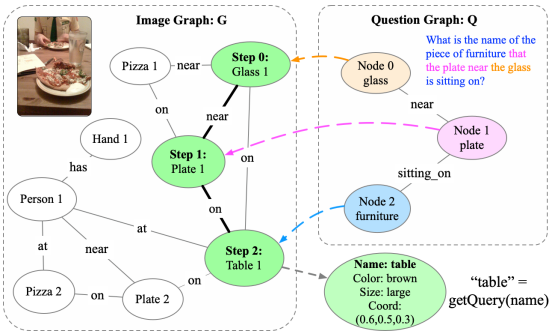

Figure 1. The graph-symbolic framework obtains question answers by traversing nodes on image graph G, guided by algorithm and question graph Q. The step 0 to step 2 are visiting trace to navigate target node that is retrieved to provide question answers

2. Methods

2.1. Image Scene Parser

BASE: we construct image BASE scene graph using off-the-shelf Mask-RCNN [7] pretrained on the COCO dataset [13], which returns detected objects (from a vocabulary of 80 objects). Each object is represented as a node on the image graph and has several attributes (e.g. color [17], size [18], and 3d-coordinates [16]), and graph edges represent object spatial relations. Moreover, we train CNN-based action classifier on Stanford 40 Actions [20] to detects actions portrayed in images. We use the Places-CNNs [23] to detect image scenes places in 205 scene categories.

LARGE: we next construct LARGE-vocabulary image graphs. In particular, we use a Faster-RCNN object detector [15] trained on the large-scale (600) object detection dataset Open Images V4 [12]. The construction process follows BASE. This large-scale detector greatly improves the diversity and extensivility of the image graph by adding more object nodes as well as node labels.

SCENE: In addition to spatial relations, we represent object semantic relations in image graphs. For example, the relation *eating* in the phrase *girl eating cake* is significant for answering questions *what is the girl eating?* To capture such semantic relations, we leverage the Visual Semantic Parsing Network, VSPNet [21] that constructs semantic

graphs by linking object relations with *subjects*, *objects*, and *predicates* (SOP).

2.2. Question Graph Parser

We formulate the question graph parsing as a sequence-to-sequence translation problem: we train a neural translator that can directly translate our textual question to a question graph. Inspired by [18], we define fine-grained special tags to segment question text into several tagged segments, which are automatically produced through executing semantic programs for questions in the GQA dataset [8]. For example, the question *what type of sign is the same color as the fire hydrant to the left of the car?* can be annotated as three segments *<sel> [0] car <suj> [1] fire_hydrant to_the_left_of [0] <obj> [2] sign same_color [1] <que> name* by running semantic programs [8], where numbers represents node ids, <sel>, <suj>, <obj>, <que> are tags. Compared to [18], we defined 15 special tags that are richer, more extensive, and exactly fit for answering GQA questions. The annotated questions with special tags paired with original question string serve as training data. In particular, we train a LSTM-based sequence-to-sequence model based on Open Neural Machine Translation [11] to convert a question (source) to a sequence of question with tags (target). Since the target sequence is ordered, we can easily convert it into a question graph as shown in Fig 1.

2.3. Question Answering

Our novel answering algorithm finds valid assignments between the image graph (Sec. 2.1) and question graph (Sec. 2.2) and outputs a question answer. In particular, we transverse each node in the question graph and call corresponding functions defined in Table 1 to navigate object nodes in the image graph. For example, The input data for function *getQuery()* is a subgraph Q', which queries attribute values in image node shown in Fig. 1, and returns a query result (e.g. the name value of the node).

3. Results

Since we are not aware of previous fully unsupervised work that extensively evaluates on GQA [8] or VQA-v2 [6], we propose a basic and advanced models. All methods use the question parser in Sec. 2.2.

- BASELINE: uses BASE image graph in Sec. 2.1. It is similar to [18] but we use fine-grained question graphs with 15 tags, and richer visual estimators, as well as extensive evaluation on GQA and VQA-v2.
- BASELINE+LARGE: uses the BASE and LARGE image graphs in Sec. 2.1, and merges nodes on two image graphs using object IoU.
- BASELINE+SCENE: merges BASE and SCENE image graphs in Sec. 2.1 using object IoU.
- BASELINE+LARGE+SCENE: merges BASE, LARGE, and SCENE image graphs in Sec. 2.1 using object IoU.

Functions	Args.	Out.	Explanation
getNodes()	G/G', N	G	return image (sub)graph
getQuery()	G/G', N	q	retrieve node attributes (*color*)
compare()	G/G', N	b	compared two nodes (*than*)
logicConn()	b, N	b	run logic connectives (*and/or*)

Table 1. Function arguments and outputs in answering algorithm; N denotes node on question graph Q, G' denotes a sub-graph of image graph G, q is a query result of node M on image graph G/G' (e.g. q=table for M={name:table, color:brown, size:large, coord:(0.6,0.5,0.3)} in Fig. 1), b denotes boolean outputs for node M in image graph G/G' (e.g. yes/no questions).

	GQA			
	overall	yes/no	choose	others
BASELINE	32.34	56.43	18.16	19.52
+LARGE	35.01	57.52	24.38	22.55
+SCENE	40.44	58.41	40.61	28.97
+LARGE+SCENE	**43.66**	**62.96**	**42.17**	**31.65**
	VQA-v2			
	overall	yes/no	number	others
BASELINE	31.88	49.12	**35.43**	12.79
+LARGE	32.53	49.64	23.96	14.20
+SCENE	32.69	49.96	25.45	13.98
+LARGE+SCENE	**33.25**	**50.17**	20.44	**15.31**

Table 2. Performance of baseline and our proposed models on GQA and VQA-v2; best method per column bolded.

The results of our proposed models are shown in Table 2. The BASELINE model achieves strong performance even if we do not have any question-answer training; however, each of our advanced methods bring significant improvements. In GQA, our BASELINE+LARGE+SCENE models show the best performance with respect to *overall, yes/no, choose*, and *other* types of questions, which demonstrates the effectiveness of the integration of LARGE and SCENE image graphs. Concretely, there is a major improvement due to BASELINE+SCENE, e.g. improved from 18.16 to 40.61 for the *choose* category, which suggests that the additional semantic SOP relations extracted from VSPNet are significant for answering real-world questions that involve object relations like (*girl, cake, eating*). Besides, BASELINE+LARGE also has significant contribution over the baseline for all types of questions on GQA.

These observations also found on VQA-v2 dataset, except the *number* questions which require accurate object detection with respect to its bounding box and labels to count objects. More complex model, such as BASELINE+LARGE+SCENE that integrates object image graphs in Sec 2.1 will introduce more variance. This problem can be ameliorated by optimizing our graph merging algorithm in our future work. In general, BASELINE+LARGE and BASELINE+SCENE improve baseline, and combining BASELINE+LARGE+SCENE improves the most in both GQA and VQA-v2 datasets, which suggests that our framework are independent of and extensible to different datasets.

References

[1] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015. 1

[2] H. Ben-Younes, R. Cadene, M. Cord, and N. Thome. Mutan: Multimodal tucker fusion for visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2612–2620, 2017. 1

[3] R. Cadene, H. Ben-Younes, M. Cord, and N. Thome. Murel: Multimodal relational reasoning for visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1989–1998, 2019. 1

[4] W.-L. Chao, H. Hu, and F. Sha. Cross-dataset adaptation for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5716–5725, 2018. 1

[5] P. Gao, Z. Jiang, H. You, P. Lu, S. C. Hoi, X. Wang, and H. Li. 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*, pages 6639–6648, 2019. 1

[6] Y. Goyal, T. Khot, D. Summers-Stay, D. Batra, and D. Parikh. 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*, pages 6904–6913, 2017. 2

[7] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 1

[8] D. A. Hudson and C. D. Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6700–6709, 2019. 1, 2

[9] J. Johnson, B. Hariharan, L. Van Der Maaten, L. Fei-Fei, C. Lawrence Zitnick, and R. Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2901–2910, 2017. 1

[10] J. Johnson, B. Hariharan, L. Van Der Maaten, J. Hoffman, L. Fei-Fei, C. Lawrence Zitnick, and R. Girshick. Inferring and executing programs for visual reasoning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2989–2998, 2017. 1

[11] G. Klein, Y. Kim, Y. Deng, J. Senellart, and A. Rush. OpenNMT: Open-source toolkit for neural machine translation. In *Proceedings of ACL 2017, System Demonstrations*, pages 67–72, Vancouver, Canada, July 2017. Association for Computational Linguistics. 2

[12] A. Kuznetsova, H. Rom, N. Alldrin, J. Uijlings, I. Krasin, J. Pont-Tuset, S. Kamali, S. Popov, M. Malloci, A. Kolesnikov, et al. The open images dataset v4. *International Journal of Computer Vision*, pages 1–26, 2020. 1

[13] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 1

[14] J. Mao, C. Gan, P. Kohli, J. B. Tenenbaum, and J. Wu. The neuro-symbolic concept learner: Interpreting scenes, words, and sentences from natural supervision. In *International Conference on Learning Representations*, 2019. 1

[15] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015. 1

[16] A. Saxena, S. H. Chung, A. Y. Ng, et al. Learning depth from single monocular images. In *NIPS*, volume 18, pages 1–8, 2005. 1

[17] J. Van De Weijer, C. Schmid, and J. Verbeek. Learning color names from real-world images. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2007. 1

[18] B.-Z. Vatashsky and S. Ullman. Vqa with no questions-answers training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 1, 2

[19] Y. Xu, L. Chen, Z. Cheng, L. Duan, and J. Luo. Open-ended visual question answering by multi-modal domain adaptation. *Association for Computational Linguistics*, Nov. 2020. 1

[20] B. Yao, X. Jiang, A. Khosla, A. L. Lin, L. Guibas, and L. Fei-Fei. Human action recognition by learning bases of action attributes and parts. In *2011 International conference on computer vision*, pages 1331–1338. IEEE, 2011. 1

[21] A. Zareian, S. Karaman, and S.-F. Chang. Weakly supervised visual semantic parsing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3736–3745, 2020. 1

[22] M. Zhang, T. Maidment, A. Diab, A. Kovashka, and R. Hwa. Domain-robust vqa with diverse datasets and methods but no target labels. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3520–3530, 2021. 1

[23] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva. Learning deep features for scene recognition using places database. Neural Information Processing Systems Foundation, 2014. 1