Supplementary Material towards Stacked Hybrid-Attention and Group Collaborative Learning for Unbiased Scene Graph Generation

Xingning Dong[1], Tian Gan[1,†], Xuemeng Song[1], Jianlong Wu[1], Yuan Cheng[2,†], Liqiang Nie[1]
[1]Shandong University, [2]Ant Group
dongxingning1998@gmail.com, gantian@sdu.edu.cn, sxmustc@gmail.com
jlwu1992@sdu.edu.cn, chengyuan.c@antgroup.com, nieliqiang@gmail.com

1. Introduction

In this supplementary material, we present more experiments and analyses, as well as discuss the limitations and future work of our method. Specifically, we first provide the visualization results of our proposed SHA+GCL to intuitively show its outstanding performance in generating unbiased scene graphs. We then analyse the number of parameters towards our proposed method, and present additional results of various model-agnostic debiasing approaches towards the regular Recall@K and the unbiased Mean Recall@K. Ultimately, we discuss the limitations of our method, based on which we provide several potential directions to further improve our SHA+GCL network.

2. Visualization Results

To get an intuitive perception of the superior performance in generating unbiased scene graphs of our proposed GCL, we visualize several PredCls examples generated from the biased SHA and the unbiased SHA+GCL. As shown in Figure 1, the model employing the proposed GCL strategy prefers to providing more informative and specific relationship predictions (*e.g.*, lying on and riding) rather than common and trivial ones (*e.g.*, on and has), *e.g.*, "person1-riding-elephant" in the top-right example and "train-pulling-car" in the bottom-left example. Moreover, the model equipped with our model-agnostic GCL could also capture potential reasonable relationships, such as "person1-watching-person2" in the top-right example and "sidewalk-beside-train" in the bottom-left example. In a nutshell, the proposed GCL could enhance the unbiased relationship predictions, thus achieving more informative scene graphs to support various down-stream tasks.

3. Parameter Statistics

We compare the total number of parameters between three baseline methods (*i.e.*, Motifs, VCTree, and SHA) and their enhanced versions that are equipped with our model-agnostic GCL in Table 1. As can be observed, compared with the original methods which possess massive number of total trainable parameters (about 200M), GCL only additionally introduces a limited number of parameters (about 2M), which could hardly influence the overall training procedure.

4. Detailed Performance

We present the complete results of our experiments employing the regular Recall@K [9], the unbiased Mean Recall@K [1, 13], and their mean [8] on all three tasks (*i.e.*, PredCls, SGCls, and SGDet) on VG150 [6] and GQA200 [4] dataset in Table 2, where $K \in \{50, 100\}$. Note that all the methods are implemented with a pre-trained Faster R-CNN [10] with ResNeXt-101-FPN [14] provided by [12] as the object detector, thus we could give a fair comparison to prove the superiority of our method.

From Table 2, we observe that 1) our proposed SHA+GCL achieves the best performance on all three tasks towards the unbiased metric mR@K in both two datasets. In VG150, we breakthrough the 40% precision in both mR@50 and mR@100 on PredCls, and 20% precision in mR@100 on both SGCls and SGDet, thus establishing a new state-of-the-art in the unbiased metric. 2) Our improvement towards the relation decoder, namely GCL strategy, is model-agnostic and could largely enhance the unbiased SGG. In both VG150 or GQA200, the method equipped with GCL nearly doubles the performance compared with the original one, showing the outstanding capability in generating unbiased scene graphs.

5. Limitations and Future Work

In this section, we would like to discuss the limitations of our method, based on which we provide several potential directions to further improve our SHA+GCL.

[†]Corresponding authors.

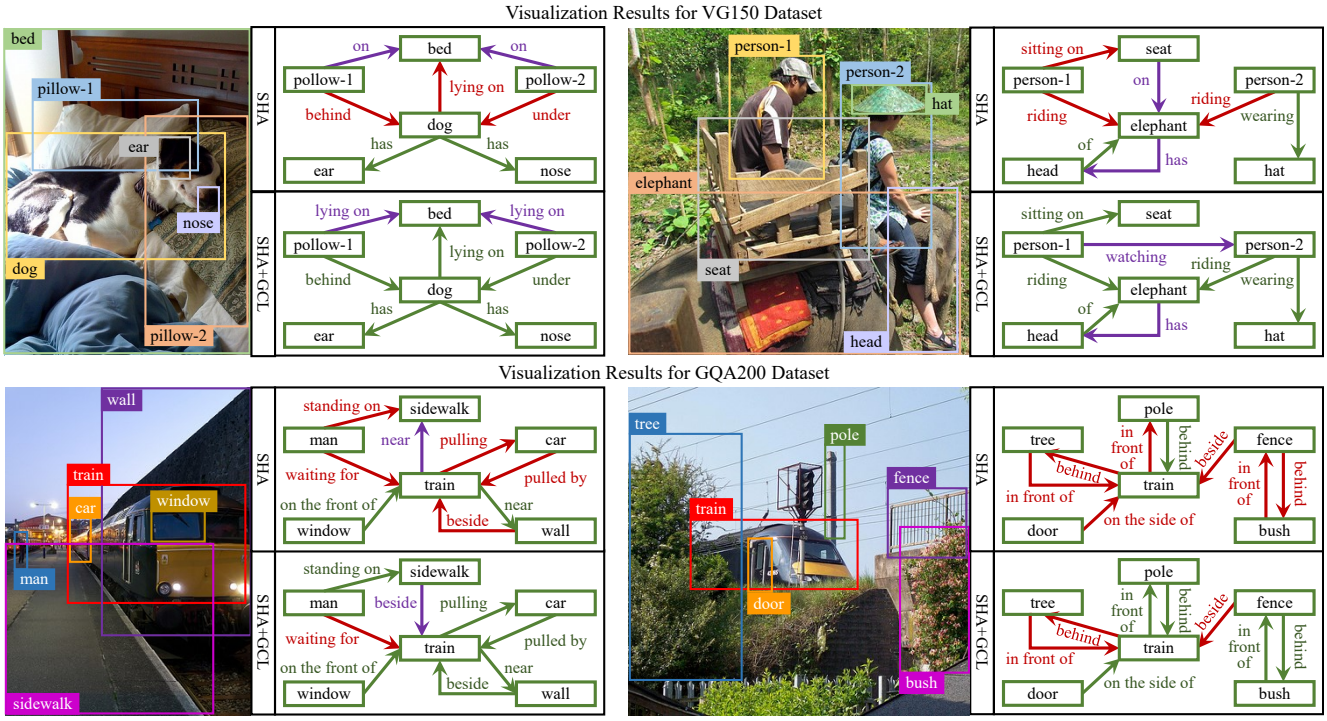

Figure 1. Qualitative comparisons between SHA and SHA+GCL with regard to R@20 on PredCls setting. Green edges represent the ground truth relationships that are correctly predicted, red edges represent the ground truth relationships that are failed to be detected, and purple edges represent the reasonable relationships which are predicted by the model but are not annotated in the ground truth.

Model	Fixed	Trainable	Model	Fixed	Trainable	Model	Fixed	Trainable
Motifs	158.7M	208.5M	VCTree	158.7M	199.8M	SHA	158.7M	228.8M
Motifs + GCL	158.7M	210.5M	VCTree + GCL	158.7M	201.8M	SHA + GCL	158.7M	230.9M

Table 1. Comparison of different methods on number of parameters. "Fixed" counts the number of parameters that belong to the pre-trained object detector, and "Trainable" counts the number of parameters that can be updated during the training procedure.

5.1. More Configurations Could be Further Explored

As aforementioned, we follow the intuition of "divide-conquer-cooperate" to address the biased relationship predictions. In the second step, namely "conquer", we borrow the idea from class-incremental learning [3] and employ the group-incremental configuration. Actually, we employ this configuration mainly due to its simplicity and efficiency, as we could directly leverage the final classifier that covers all the candidate classes to obtain the predictions in the evaluation stage. However, we should argue that it is not the only alternative to fulfill the "conquer" step. Therefore, in the future, we aim to explore more robust group dividing methods as well as classifier configuration strategies to promote the unbiased SGG, *e.g.*, the group-split configuration in Figure 2.

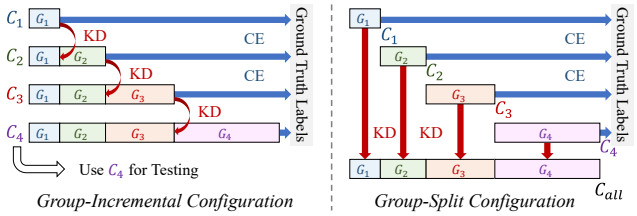

Figure 2. The group-incremental configuration (left) may not be the only alternative to fulfill the "conquer" step in GCL. For example, the group-split configuration (right) is another promising strategy. Therefore, we aim to explore more robust group dividing methods and classifier configuration strategies in the future.

5.2. "Strong Constraint" Could be Further Enhanced

As aforementioned, in the "cooperate" step, we use the collaborative knowledge distillation to establish an effective

knowledge transfer mechanism, where a regular Kullback-Leibler Divergence loss is employed. However, since various novel methods [5] have been proposed in the knowledge distillation area, we could further enhance our GCL by devising more efficient strategies, thus strengthening the "Strong Constraint" and promoting the unbiased SGG.

References

[1] Tianshui Chen, Weihao Yu, Riquan Chen, and Liang Lin. Knowledge-embedded routing network for scene graph generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6163–6171, 2019. 1

[2] Meng-Jiun Chiou, Henghui Ding, Hanshu Yan, Changhu Wang, Roger Zimmermann, and Jiashi Feng. Recovering the unbiased scene graphs from the biased ones. *arXiv preprint arXiv:2107.02112*, 2021. 4

[3] Xinting Hu, Yi Jiang, Kaihua Tang, Jingyuan Chen, Chunyan Miao, and Hanwang Zhang. Learning to segment the tail. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14045–14054, 2020. 2

[4] Drew A Hudson and Christopher 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

[5] Taehyeon Kim, Jaehoon Oh, NakYil Kim, Sangwook Cho, and Se-Young Yun. Comparing kullback-leibler divergence and mean squared error loss in knowledge distillation. *arXiv preprint arXiv:2105.08919*, 2021. 3

[6] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *arXiv preprint arXiv:1602.07332*, 2016. 1

[7] Rongjie Li, Songyang Zhang, Bo Wan, and Xuming He. Bipartite graph network with adaptive message passing for unbiased scene graph generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11109–11119, 2021. 4

[8] Xin Lin, Changxing Ding, Jinquan Zeng, and Dacheng Tao. Gps-net: Graph property sensing network for scene graph generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3746–3753, 2020. 1

[9] Cewu Lu, Ranjay Krishna, Michael Bernstein, and Li Fei-Fei. Visual relationship detection with language priors. In *Proceedings of the European Conference on Computer Vision*, pages 852–869, 2016. 1

[10] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 39(6):1137–1149, 2017. 1

[11] Mohammed Suhail, Abhay Mittal, Behjat Siddiquie, Chris Broaddus, Jayan Eledath, Gerard Medioni, and Leonid Sigal. Energy-based learning for scene graph generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13936–13945, 2021. 4

[12] Kaihua Tang, Yulei Niu, Jianqiang Huang, Jiaxin Shi, and Hanwang Zhang. Unbiased scene graph generation from biased training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3716–3725, 2020. 1, 4

[13] Kaihua Tang, Hanwang Zhang, Baoyuan Wu, Wenhan Luo, and Wei Liu. Learning to compose dynamic tree structures for visual contexts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6619–6628, 2019. 1

[14] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1492–1500, 2017. 1

[15] Jing Yu, Yuan Chai, Yue Hu, and Qi Wu. Cogtree: Cognition tree loss for unbiased scene graph generation. *arXiv preprint arXiv:2009.07526*, 2020. 4

Evaluation on Visual Genome Dataset

Model	PredCls		SGCls		SGDet		MEAN	
	R@50/100	mR@50/100	R@50/100	mR@50/100	R@50/100	mR@50/100	R-M	mR-M
IMP† [11]	61.1 / 63.1	11.0 / 11.8	37.4 / 38.3	6.4 / 6.7	23.6 / 28.7	3.3 / 4.1	42.0	7.2
GPS-Net† [7]	65.2 / 67.1	15.2 / 16.6	37.8 / 39.2	8.5 / 9.1	31.1 / 35.9	6.7 / 8.6	46.1	10.8
SG-CogTree [15]	38.4 / 39.7	28.4 / 31.0	22.9 / 23.4	15.7 / 16.7	19.5 / 21.7	11.1 / 12.7	27.6	19.3
BGNN [7]	59.2 / 61.3	30.4 / 32.9	37.4 / 38.5	14.3 / 16.5	31.0 / 35.8	10.7 / 12.6	43.9	19.6
VTransE [12]	**65.7** / **67.6**	14.7 / 15.8	38.6 / 39.4	8.2 / 8.7	29.7 / 34.3	5.0 / 6.0	45.9	9.7
VTransE + TDE [12]	48.5 / 43.1	24.6 / 28.0	25.7 / 28.5	12.9 / 14.8	18.7 / 22.6	8.6 / 10.5	31.2	16.6
VTransE + GCL	35.4 / 37.3	34.2 / 36.3	25.8 / 26.9	20.5 / 21.2	14.6 / 17.1	13.6 / 15.5	26.2	23.5
Motifs [12]	65.2 / 67.0	14.8 / 16.1	38.9 / 39.8	8.3 / 8.8	**32.8** / **37.2**	6.8 / 7.9	46.8	10.4
Motifs + Reweight [2]	54.7 / 56.5	17.3 / 18.6	29.5 / 31.5	11.2 / 11.7	24.4 / 29.3	9.2 / 10.9	37.7	13.2
Motifs + TDE [12]	46.2 / 51.4	25.5 / 29.1	27.7 / 29.9	13.1 / 14.9	16.9 / 20.3	8.2 / 9.8	32.1	16.8
Motifs + PCPL† [2]	54.7 / 56.5	24.3 / 26.1	35.3 / 36.1	12.0 / 12.7	27.8 / 31.7	10.7 / 12.6	40.4	16.4
Motifs + CogTree [15]	35.6 / 36.8	26.4 / 29.0	21.6 / 22.2	14.9 / 16.1	20.0 / 22.1	10.4 / 11.8	26.4	18.1
Motifs + DLFE [2]	52.5 / 54.2	26.9 / 28.8	32.3 / 33.1	15.2 / 15.9	25.4 / 29.4	11.7 / 13.8	37.8	18.7
Motifs + EMB [11]	65.2 / 67.3	18.0 / 19.5	39.2 / 40.0	10.2 / 11.0	31.7 / 36.3	7.7 / 9.3	46.6	12.6
Motifs + GCL	42.7 / 44.4	36.1 / 38.2	26.1 / 27.1	20.8 / 21.8	18.4 / 22.0	16.8 / 19.3	30.1	25.5
VCTree [12]	65.4 / 67.2	16.7 / 18.2	**46.7** / **47.6**	11.8 / 12.5	31.9 / 36.2	7.4 / 8.7	**49.2**	12.6
VCTree + Reweight [2]	60.7 / 62.6	19.4 / 20.4	42.3 / 43.5	12.5 / 13.1	27.8 / 32.0	8.7 / 10.1	44.8	14.0
VCTree + TDE [12]	47.2 / 51.6	25.4 / 28.7	25.4 / 27.9	12.2 / 14.0	19.4 / 23.2	9.3 / 11.1	32.5	16.8
VCTree + PCPL† [2]	56.9 / 58.7	22.8 / 24.5	40.6 / 41.7	15.2 / 16.1	26.6 / 30.3	10.8 / 12.6	42.5	17.0
VCTree + CogTree [15]	44.0 / 45.4	27.6 / 29.7	30.9 / 31.7	18.8 / 19.9	18.2 / 20.4	10.4 / 12.1	31.8	19.8
VCTree + DLFE [2]	51.8 / 53.5	25.3 / 27.1	33.5 / 34.6	18.9 / 20.0	22.7 / 26.3	11.8 / 13.8	37.1	19.5
VCTree + EMB [11]	64.0 / 65.8	18.2 / 19.7	44.7 / 45.8	12.5 / 13.5	31.4 / 35.9	7.7 / 9.1	47.9	13.5
VCTree + GCL	40.7 / 42.7	37.1 / 39.1	27.7 / 28.7	22.5 / 23.5	17.4 / 20.7	15.2 / 17.5	29.6	25.8
SHA	64.3 / 66.4	18.8 / 20.5	38.0 / 39.0	10.9 / 11.6	30.6 / 34.9	7.8 / 9.1	45.5	13.1
SHA + GCL	35.1 / 37.2	**41.6** / **44.1**	22.8 / 23.9	**23.0** / **24.3**	14.9 / 18.2	**17.9** / **20.9**	25.4	**28.6**

Evaluation on GQA Dataset

Model	PredCls		SGCls		SGDet		MEAN	
	R@50/100	mR@50/100	R@50/100	mR@50/100	R@50/100	mR@50/100	R-M	mR-M
VTransE	55.7 / 57.9	14.0 / 15.0	33.4 / 34.2	8.1 / 8.7	27.2 / 30.7	5.8 / 6.6	39.9	9.6
VTransE + GCL	35.5 / 37.4	30.4 / 32.3	22.9 / 23.6	16.6 / 17.4	15.3 / 18.0	14.7 / 16.4	25.4	21.4
Motifs	**65.3** / **66.8**	16.4 / 17.1	**34.2** / **34.9**	8.2 / 8.6	**28.9** / **33.1**	6.4 / 7.7	**43.9**	10.9
Motifs + GCL	44.5 / 46.2	36.7 / 38.1	23.2 / 24.0	17.3 / 18.1	18.5 / 21.8	16.8 / 18.8	29.7	24.2
VCTree	63.8 / 65.7	16.6 / 17.4	34.1 / 34.8	7.9 / 8.3	28.3 / 31.9	6.5 / 7.4	43.1	10.5
VCTree + GCL	44.8 / 46.6	35.4 / 36.7	23.7 / 24.5	17.3 / 18.0	17.6 / 20.7	15.6 / 17.8	29.6	23.6
SHA	63.3 / 65.2	19.5 / 21.1	32.7 / 33.6	8.5 / 9.0	25.5 / 29.1	6.6 / 7.8	41.6	12.1
SHA + GCL	42.7 / 44.5	**41.0** / **42.7**	21.4 / 22.2	**20.6** / **21.3**	14.8 / 17.9	**17.8** / **20.1**	27.3	**27.3**

Table 2. Detailed performance comparison of different methods on PredCls, SGCls, and SGDet tasks of both VG150 and GQA200 with respect to R@50/100 (%), mR@50/100 (%), and their mean (%). R-M and mR-M denote the mean on all three tasks over R@50/100 and mR@50/100, respectively. The optimal results from the same baseline (*i.e.*, VTransE, Motifs and VCTree) in VG150 are underlined. The global optimal results over all the methods in VG150 and GQA200 are in bold. The superscript † denotes that the method is reproduced. Note that all the methods are implemented on the same object detector, *i.e.*, a pre-trained Faster R-CNN with ResNeXt-101-FPN.