RelTransformer: A Transformer-Based Long-Tail Visual Relationship Recognition Supplementary

Jun Chen[1], Aniket Agarwal[2], Sherif Abdelkarim[1], Deyao Zhu[1], Mohamed Elhoseiny[1]
[1]King Abdullah University of Science and Technology
[2] Indian Institute of Technology
{jun.chen,deyao.zhu,mohamed.elhoseiny}@kaust.edu.sa
aagarwal@ma.iitr.ac.in, sherif.abdelkarim91@gmail.com

A. Training and Implementation Details

To have fair comparisons with all the baseline models, we follow the same experimental setup as [1] for GQA-LT and VG8K-LT evaluation and also the same setup as [10] for VG200 evaluation. We train our model and all its ablations with 8 V100 GPUs. The batch size is 8. We train our models with 12, 8, and 7 epochs on GQA-LT, VG8K-LT, and VG200 datasets, respectively. The hidden size of h is 768. The number of Transformer heads is 12. Relational and global-context encoders both have 2 layers. The memory size is 100×768. We use the Faster R-CNN [6] with VGG-16 [7] backbone to extract the object proposal features. We also apply the pretrained Word2Vec [5] embeddings to represent the relation and object labels.

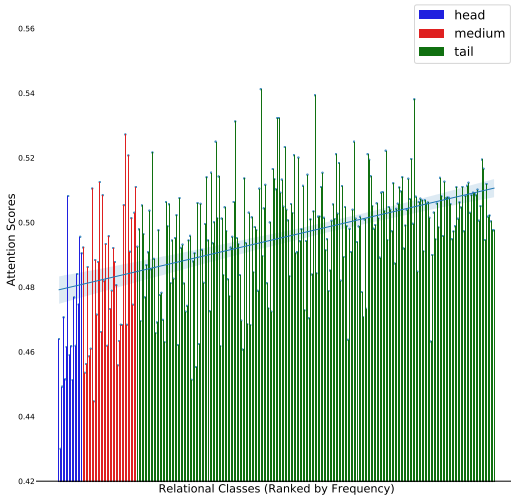


Figure 1. Average memory attention scores per each relation class on evaluation of GQA-LT. The relation classes are ranked by its frequency (from high to low)

B. Subject and Object Per-Class Accuracy

We also provide the per-class accuracy for subjects and objects in Table 1. We can observe that our RelTransformer can also easily outperform many baselines. Combining RelTransformer with CE loss can outperform all the baseline models on each category and it significantly improves over the "many" category. Combining it with WCE can furthermore improve over "medium" and "tail" categories on both datasets. These results demonstrate the effectiveness of our model on the subject and object prediction. We also notice that RelTransformer (WCE or DCPL) drops the performance over "many" categories compared to CE loss function; We hypothesis that it is due to the assigned lower weights on the high-frequent subjects/objects, and lower weights will lead to the lower confidence values during the classification. This phenomenon can explain why RelTansformer (WCE) underperforms on the "many" classes and why it underperforms on the compositional prediction for "many" and "medium" classes.

C. Per-Example Accuracy

We provide the per-example accuracy for the subjects and objects on GQA-LT dataset. It shows that RelTransformer (CE) achieves the best performance on predicting subjects/objects and relations among all the baselines. Per-example accuracy is mainly dominated by the "head" classes since they are very large in example numbers. This indicates that RelTransformer (CE) improves both head and tail classes. Compared to the best baseline LSVRU (CE), RelTransformer (CE) improves it by 10.6 acc on subject/object and 0.3 acc on relation predictions. However, we could also notice that RelTransformer (WCE or DCPL) brings the performance down on per-example accuracy, and this phenomenon is also observed in LSVRU baselines. But the results in Table 1(Supplementary) and Table 1 (main paper) shows that RelTransformer (WCE or DCPL) has a

Architecture	Learning Methods	VG8K-LT many 267	VG8K-LT medium 799	VG8K-LT few 4,264	VG8K-LT all 5,330	GQA-LT many 86	GQA-LT medium 255	GQA-LT few 1,362	GQA-LT all 1,703
LSVRU	VilHub [1]	61.6	20.3	10.1	14.2	68.6	44.0	10.3	18.3
LSVRU	VilHub + RelMix [1]	59.5	15.1	10.4	13.6	68.8	42.1	10.1	18.1
LSVRU	OLTR [4]	56.8	12.0	9.6	12.3	68.2	37.2	7.0	14.6
LSVRU	EQL [8]	56.9	12.1	10.0	12.7	68.9	43.7	10.0	18.0
LSVRU	Counterfactual$^\sharp$ [9]	57.3	11.1	8.5	11.4	68.3	37.0	6.9	14.5
LSVRU	CE	57.3	11.1	8.5	11.4	68.3	37.0	6.9	14.5
RelTransformer	CE	**67.1**	**25.9**	**11.5**	**16.5**	**78.0**	**56.6**	**14.2**	**23.8**
LSVRU	Focal Loss [3]	58.1	13.9	8.9	12.1	68.2	39.2	7.5	15.3
RelTransformer	Focal Loss	**65.6**	**21.7**	**10.8**	**15.2**	**75.0**	**51.4**	**11.9**	**21.0**
LSVRU	DCPL [2]	**53.8**	5.9	7.9	9.9	**64.0**	35.3	6.4	13.7
RelTransformer	DCPL	50.3	**30.9**	**13.4**	**17.8**	51.8	**44.6**	**19.2**	**24.7**
LSVRU	WCE	**52.8**	27.2	10.8	14.5	**53.4**	42.0	14.0	20.2
RelTransformer	WCE	50.1	**31.3**	**13.7**	**18.0**	50.3	**46.2**	**28.7**	**32.4**

Table 1. Average per-class accuracy for subject/object. We separately evaluate the average per-class accuracy for many, medium, few, and all categories. The best performance from each category is underlined. $\sharp$ denotes our reproduction. Our model is denoted in gray

Architecture	Learning Methods	Per-example Accuracy Subject/Object	Per-example Accuracy Relation
LSVRU	CE	51.9	94.8
LSVRU	VilHub [1]	53.9	91.2
LSVRU	VilHub + RelMix [1]	53.5	91.0
LSVRU	EQL [8]	51.1	93.9
LSVRU	WCE	37.6	72.6
RelTransformer	CE	**62.5**	**95.1**
RelTransformer	Focal Loss [3]	59.4	95.0
RelTransformer	DCPL [2]	34.3	80.5
RelTransformer	WCE	34.2	74.2

Table 2. Per-example Accuracy on GQA-LT. The best performance from each category is underlined. Our model is denoted in gray

good-performing result on "medium" and "tail" categories. This indicates that combining with class-imbalance loss functions can benefit low-frequent class predictions with the cost of the performance from a few number of top-frequent classes; RelTransformer improves the results more general.

D. Memory Attention Further Analysis

To further investigate the role of memory attention, we compute the memory attention scores, $J - \alpha$ in our fusion function Eq. 1, for each testing example from well-trained RelTransformer (WCE) on the GQA-LT dataset. We average the attention scores per each relation-class and demonstrate them in Fig. 1. The relation classes are ranked according to their frequency in the training data. We can observe a clear increasing trend for the attention scores from high-frequent relations to low-frequent ones, meaning that the features from memory can contribute to the long-tail relations more than the head-relations, which can reflect why memory can gain more improvement on the "medium" and "tail" classes.

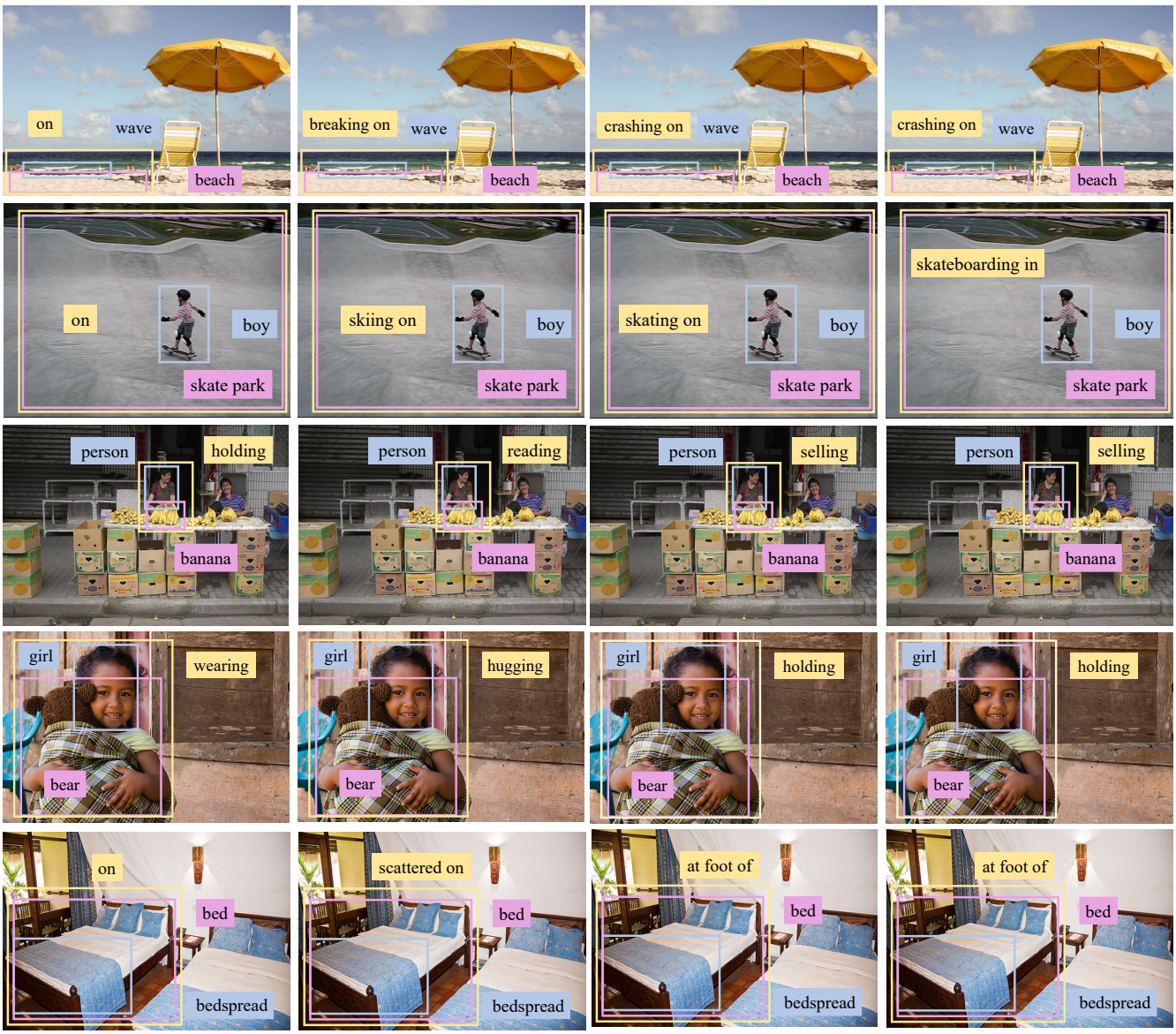
$$\begin{aligned} g(x, y) &= \alpha \odot x + (J - \alpha) \odot y \\ \alpha &= \sigma(W[x; y] + b) \end{aligned} \tag{1}$$

where W is a 2D $\times$ D matrix. b is a bias term. $[;]$ denotes the concatenation. $\odot$ is the Hadamard product. J is an all-one matrix with the same dimensions as α.

E. Additional Qualitative Examples

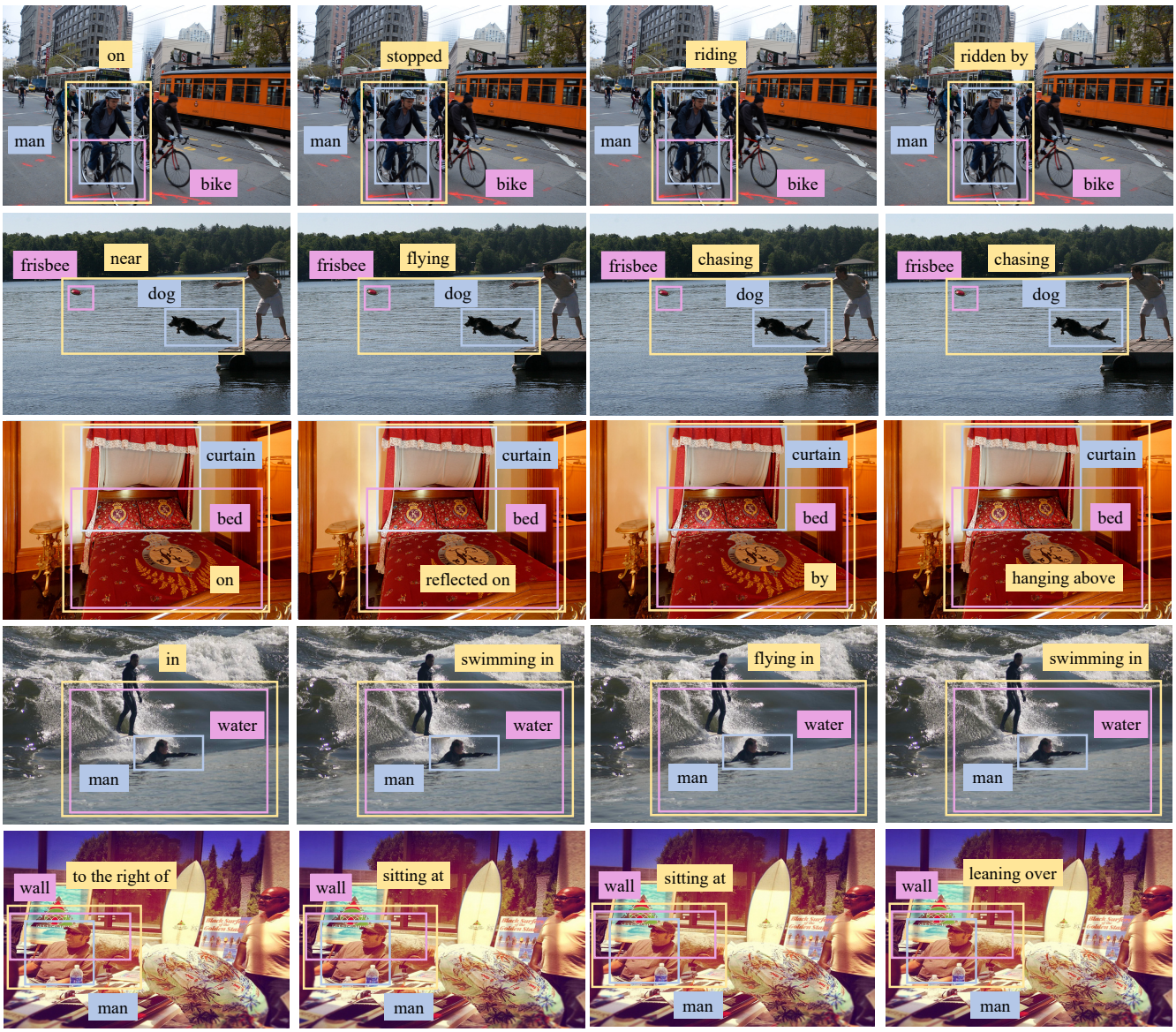
We show the relation prediction results on VG200 dataset in Fig. 2 and provide more long-tail relation prediction results on GQA-LT and VG8K-LT dataset in Fig. 3 and 4.

Figure 2. Qualitative results on VG200 dataset



LSVRU | RelTransformer | LSVRU (WCE) | RelTransformer (WCE)

Figure 3. More long-tail relation prediction examples on GQA-LT dataset



LSVRU | RelTransformer | LSVRU (WCE) | RelTransformer (WCE)

Figure 4. More long-tail relation prediction examples on VG8K-LT dataset

References

[1] Sherif Abdelkarim, Aniket Agarwal, Panos Achlioptas, Jun Chen, Jiaji Huang, Boyang Li, Kenneth Church, and Mohamed Elhoseiny. Exploring long tail visual relationship recognition with large vocabulary. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15921–15930, 2021. 1, 2

[2] Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis. Decoupling representation and classifier for long-tailed recognition. In *International Conference on Learning Representations*, 2019. 2

[3] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017. 2

[4] Ziwei Liu, Zhongqi Miao, Xiaohang Zhan, Jiayun Wang, Boqing Gong, and Stella X Yu. Large-scale long-tailed recognition in an open world. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2537–2546, 2019. 2

[5] Tomas Mikolov, Ilya Sutskever, Kai Chen, Gregory S Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality. In *NIPS*, 2013. 1

[6] Shaoqing Ren, Kaiming He, Ross B Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NIPS*, 2015. 1

[7] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *ICLR*, 2015. 1

[8] Jingru Tan, Changbao Wang, Buyu Li, Quanquan Li, Wanli Ouyang, Changqing Yin, and Junjie Yan. Equalization loss for long-tailed object recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11662–11671, 2020. 2

[9] 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. 2

[10] Ji Zhang, Yannis Kalantidis, Marcus Rohrbach, Manohar Paluri, Ahmed Elgammal, and Mohamed Elhoseiny. Large-scale visual relationship understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 9185–9194, 2019. 1