Supplementary Materials for Fine-Grained is Too Coarse: A Novel Data-Centric Approach for Efficient Scene Graph Generation

Maëlic Neau[1,2] Paulo E. Santos[1] Anne-Gwenn Bosser[2] Cédric Buche[2]
[1]College of Science and Engineering, Flinders University, Australia
[2]Ecole Nationale d'Ingénieurs de Brest, France
{neau, buche, bosser}@enib.fr, paulo.santos@flinders.edu.au

A. Imbalance of Predicates

After removing invariant relations, we noticed a slightly better balance in the predicate distribution of VG150-cur. In Table 1, we compare the Imbalance Ratio (IR) [10] and Imbalance Degree (ID) [9] of the different splits of Visual Genome. IR compares the distribution between head and tail classes only while ID compares the normalized distance between the actual distribution and a perfect distribution across all classes. However, as highlighted in [12], this metric is highly dependent on the type of distance chosen as well as the number of minority classes. As neither of the Imbalance Ratio nor Imbalance Degree gives a full picture of the imbalance in multi-class distribution, we also measure the likelihood-ratio imbalance degree (LRID) [12] as follows:

$$LRID = -2\sum_{c=1}^{C} n_c \log \frac{N}{Cn_c} \tag{1}$$

Where C is the number of unique classes, n_c is the frequency of each class and N is the perfect distribution. This metric tests the actual distribution against a complete balance distribution of the data, and is reported to be more accurate for multi-class problems [12]. From Table 1 we observe that the difference between head and tail classes is lower by a strong margin when looking at IR between VG150 and VG150-cur. However, when looking at ID and LRID, we see that the imbalance over the global distribution is slightly similar, with a small advantage for VG150-cur. As unbiased SGG models are very sensitive to the imbalance of the dataset, this small difference could be a factor of the global improvement observed by training baseline models on VG150-cur.

Dataset	IR ↓	ID ↓	LRID ↓
VG80K	600,210	29,278	13.75
VG150	6,549	40.7	2.99
VrR-VG	619	95.61	2.50
VG150-con	1,697	40.69	2.98
VG150-cur	1,319	39.68	2.79

Table 1: Imbalance distribution measurement for different splits of the Visual Genome dataset.

B. Training Strategy

For training on the task of Scene Graph Generation, we kept the same ratio of Training and Test images as VG150 with 0.7 and 0.3, respectively. In contrast to VG150 which set a fixed amount of 5000 validation images, we decided to split the Training set between Train and Validation sets such as the validation set will have a ratio of 0.05 of the total amount of annotated images. Furthermore, the training strategy of VG150 is to keep the same split for both the training of Faster-RCNN and the relations training. However, a consequent amount of images are skipped when training the relation head due to the fact that they only have bounding boxes annotations. To preserve the train/val/test ratio, we then created a second split with images that contains at least one relation per image. This split is used exclusively when training the relation head while the first split is used for the training of the object detection backbone, see Table 2.

	Object	Relationship
Dataset	Train/Val/Test	Train/Val/Test
VG150	68,538/5,000/31,876	68,538/5,000/31,876
VG150-con	68,375/5,386/32,033	61,982/5,021/28,398
VG150-cur	68,380/5,385/32,049	59,948/4,875/26,980

Table 2: Statistics of the different training splits.

C. Advanced Statistics

C.1. Class Similarity

To give a fair comparison between VG150 and our approach, we also computed the similarities between predicate and object class distribution. VG150-cur possesses 88%

of similar predicate classes to VG150 while VG150-con possesses 92% similarity. Regarding object classes, both VG150-cur and VG150-con possess a 77% similarity with VG150. These statistics show that a comparison between VG150 and our approach makes sense as the distribution of classes is very similar.

C.2. Scene Graph Generation Datasets

Even if Visual Genome is the largest and most widely adopted dataset in SGG, other datasets have been used such as VRD [11], Open Image [6] or GQA [1]. Table 3 shows a comparison between these datasets and Visual Genome. The Scene Graph [3] and VRD datasets [8] are too small and do not contains enough samples for efficient learning. On the other hand, OpenImage is a large-scale dataset but with very sparse annotations (e.g. average graph size of 2.7558) and a small number of predicate classes. Finally, the GQA dataset [2] is a subset of Visual Genome with the addition of Question-Answer pairs. Regarding scene graph annotations, GQA sees the addition of *to the left* and *to the right* relations and some manual refinement compared to Visual Genome. In the end, GQA possesses only 112,148 relationships other than *to the left* and *to the right*, making it less diverse and more sparse than Visual Genome or VG150. The manual refinement also did not cover the removal of any irrelevant relation. Because GQA annotations are based on Visual Genome, we did not create a GQA-con and GQA-cur splits. However, our method could be easily applied to the GQA dataset to remove irrelevant relations.

Dataset	Images	Obj.	Pred.	# Rel.	Irr.
Visual Genome [4]	108,073	95,394	33,121	2,316,063	Yes
Scene Graph [3]	5,000	6,745	1,310	112,707	No
VRD [8]	5,000	100	70	37,993	No
OpenImage [5]	133,503	601	30	367,914	No
GQA [2]	85,638	1,703	310	471,614	Yes

Table 3: Comparison of different Scene Graph datasets, # Rel. displays the total number of relationships and Irr. outlines the presence of irrelevant relations.

C.3. Connectivity

Apart from average node degree or average graph size, looking at the average density of graphs across the different datasets gives some insight about the connectivity and interdependencies of samples. In table 4 we compare the average graph density to the number of annotated images. We see that VG150-cur possesses a higher density than VG150 while having more images.

The average density is a good metric however it does not provide a clear overview of inter-dependencies. In fact, all the relations on an image are not always dependent on each other (for instance between foreground/background regions). Thus, the average node degree metric used in the paper is more reliable. Table 4 also highlights the strong sparsity of VrR-VG [7] discussed in the paper when looking at both the average graph size and average node degree. This makes the VrR-VG unsuitable to evaluate the performance of context-aware models.

D. Annotations Comparison

Figure 1 gives an overview of the difference in annotations between the original dataset, VG150, VG150-connected, and VG150-curated.

References

[1] Xingning Dong et al. Stacked Hybrid-Attention and Group Collaborative Learning for Unbiased Scene Graph Generation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19405–19414, New Orleans, LA, USA, June 2022. IEEE. 2

[2] 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. 2

[3] Justin Johnson, Ranjay Krishna, Michael Stark, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. Image retrieval using scene graphs. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3668–3678, Boston, MA, USA, June 2015. IEEE. 2

[4] Ranjay Krishna et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73, 2017. 2

[5] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, Tom Duerig, and Vittorio Ferrari. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *IJCV*, 2020. 2

[6] Rongjie Li, Songyang Zhang, and Xuming He. Sgtr: End-to-end scene graph generation with transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19486–19496, 2022. 2

[7] Yuanzhi Liang, Yalong Bai, Wei Zhang, Xueming Qian, Li Zhu, and Tao Mei. Vrr-vg: Refocusing visually-relevant relationships. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10403–10412, 2019. 2

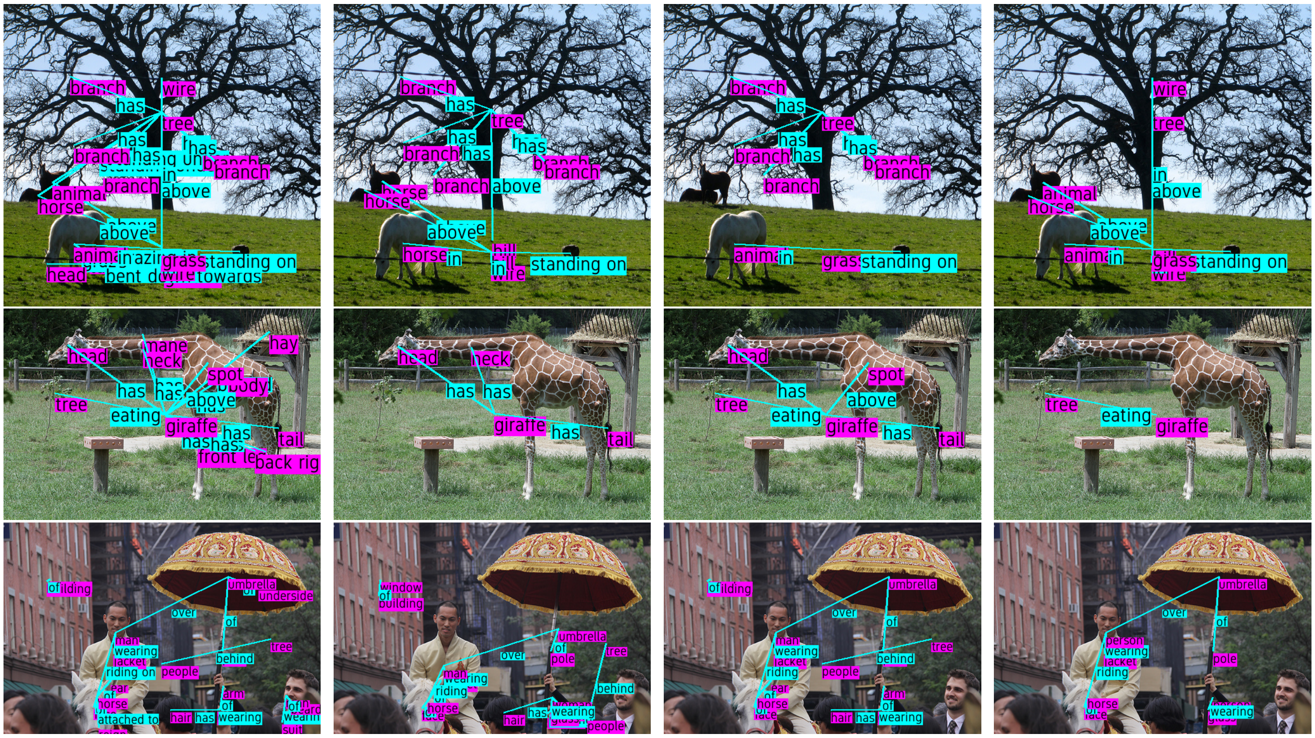
[8] Cewu Lu, Ranjay Krishna, Michael Bernstein, and Li Fei-Fei. Visual relationship detection with language priors. In *European Conference on Computer Vision*, 2016. 2

[9] Jonathan Ortigosa-Hernández, Inaki Inza, and Jose A Lozano. Measuring the class-imbalance extent of multi-class problems. *Pattern Recognition Letters*, 98:32–38, 2017. 1

[10] Ronaldo C Prati, Gustavo EAPA Batista, and Diego F Silva. Class imbalance revisited: a new experimental setup to assess the performance of treatment methods. *Knowledge and Information Systems*, 45:247–270, 2015. 1

Dataset	Graph Size	Node Degree	Subgraphs	Subgraph Size	Density	# Images
VG80K	19.0267	2.3474	3.4352	6.0786	0.1348	104,832
VG150	6.9834	2.0262	1.9719	3.8775	0.2965	89,168
VrR-VG	3.1372	1.5547	1.5368	2.6451	0.3873	56,254
VG150-con	8.3794	2.1909	2.1018	4.1092	0.2752	95,400
VG150-cur	7.1453	2.1248	1.9487	3.6624	0.3219	91,803

Table 4: Comparison of connectivity and number of images in the different datasets. # Images represent the number of annotated images with at least one relation, different than the global number of annotated images as some images have only bounding boxes annotations.



(a) VG80K (b) VG150 (c) VG150-connected (d) VG150-curated

Figure 1: A few examples of the difference between annotations in the original dataset VG80K, VG150, VG150-connected, and VG150-curated. We can easily see that annotation from VG150-curated (right) only detail informative relations, while irrelevant annotations are heavily present in the other data splits.

[11] Ruichi Yu, Ang Li, Vlad I Morariu, and Larry S Davis. Visual relationship detection with internal and external linguistic knowledge distillation. In *Proceedings of the IEEE international conference on computer vision*, pages 1974–1982, 2017. 2

[12] Rui Zhu, Ziyu Wang, Zhanyu Ma, Guijin Wang, and Jing-Hao Xue. Lrid: A new metric of multi-class imbalance degree based on likelihood-ratio test. *Pattern Recognition Letters*, 116:36–42, 2018. 1