Supplementary Material – Not All Relations are Equal: Mining Informative Labels for Scene Graph Generation

Arushi Goel[1], Basura Fernando[2], Frank Keller[1], and Hakan Bilen[1]
[1]School of Informatics, University of Edinburgh, UK
[2]CFAR, IHPC, A*STAR, Singapore.

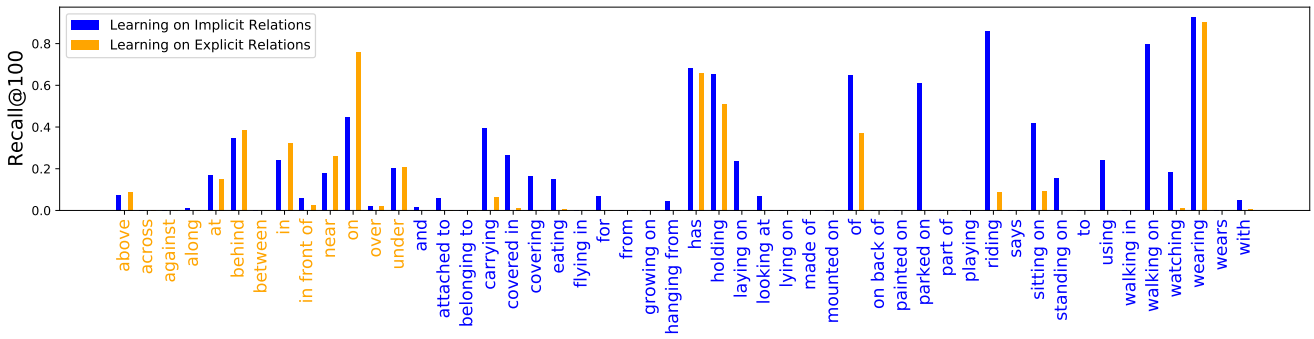


Figure 1. Class Wise Recall@100 for the set of Implicit and Explicit Relations by training on a subset of relations. All models are trained with Motif-TDE-Sum [3, 6].

1. Dataset Statistics

For the Visual Genome dataset [1], Xu *et al*. [5] released a version of the dataset with 50 relations and 150 object categories. These 50 relations are: *above, across, against, along, and, at, attached to, behind, belonging to, between, carrying, covered in, covering, eating, flying in, for, from, growing on, hanging from, has, holding, in, in front of, laying on, looking at, lying on, made of, mounted on, near, of, on, on back of, over, painted on, parked on, part of, playing, riding, says, sitting on, standing on, to, under, using, walking in, walking on, watching, wearing, wears, with.*

In Table 1, we present the explicit set of relations with the number of training instances for each relation in the Visual Genome dataset. Similarly, in Table 2 and Table 3, we define the implicit set of relations[1] and their frequency in the Visual Genome dataset.

2. Additional Results

Quantitative Studies. In Figure 1, we present the class wise recall for all the relation classes in the Visual Genome dataset [1] for training on a subset of relations *i.e*. either learning only on *explicit* relations or only on *implicit* relations. All the models are trained with the MOTIF-TDE-Sum SGG model [3, 6]. The class-wise performances clearly indicates the generalizability of training only on implicit relations as it achieves at-par/similar performances on the explicit relations, whereas the model only trained on explicit relations performs poorly on implicit relations.

Figure 2 compares the class-wise performances for the VCTree model trained with only Energy Based Modeling (EBM) [2] and also with our proposed method. The performance gains of our model over the baseline in the implicit relations such as "carrying", "eating", "covering", walking" *etc*. shows the importance of mining these informative relations from less informative samples while still maintaining recall on the explicit relations hence, improving generalization.

Regular Recall Results. In Table 4, we show the Regular Recall@k results for different SGG backbone architectures when trained with our proposed method compared to the baseline. Although, there is no significant improvement in Regular Recall (when compared to the improvements obtained from mean recall), the at-par performance with the

[1]We break down the implicit relations into two tables for better visualization.

Explicit Relations	above	across	against	along	at	behind	between	in	in front of	near	on	over	under
# of Instances	47341	1996	3092	3624	9903	41356	3411	251756	13715	96589	712409	9317	22596

Table 1. Explicit Relations for the Visual Genome Dataset [1].

Implicit Relations	attached to	and	belonging to	carrying	covered in	covering	eating	flying in	for	from	growing on	hanging from	has	holding	laying on	looking at
# of Instances	10190	3477	3288	5213	2312	3806	4688	1973	9145	2945	1853	9894	277936	42722	3739	3083

Table 2. Implicit Relations for the Visual Genome Dataset [1].

Implicit Relations	lying on	made of	mounted on	of	on back of	painted on	parked on	part of	playing	riding	says	sitting on	standing on	to	using	walking in	walking on	watching	wearing	wears	with
# of Instances	1869	2380	2253	146339.	1914	3095	2721	2065	3810	8856	2241	18643	14185	2517	1925	1740	4613	3490	136099	15457	66425

Table 3. Implicit Relations for the Visual Genome Dataset [1].

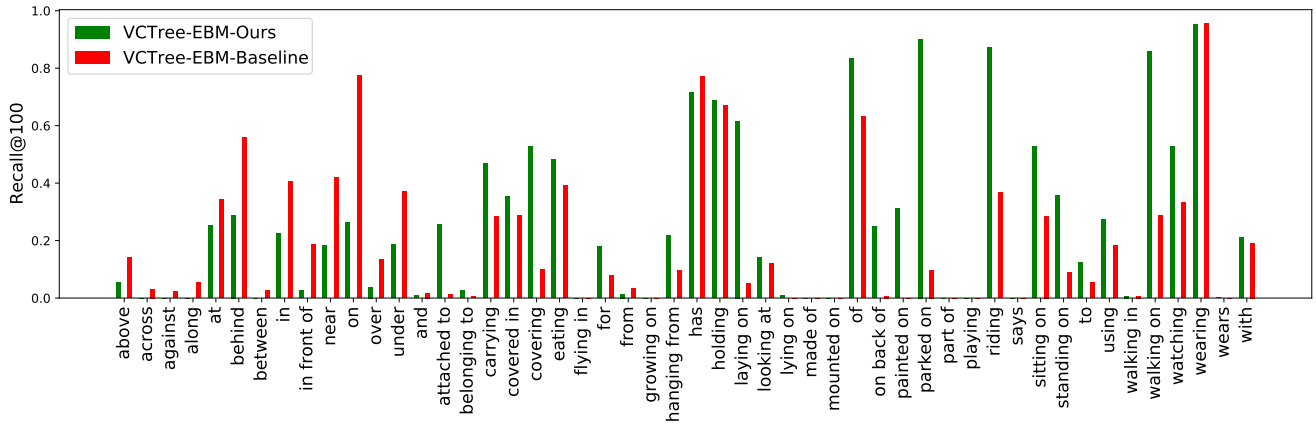


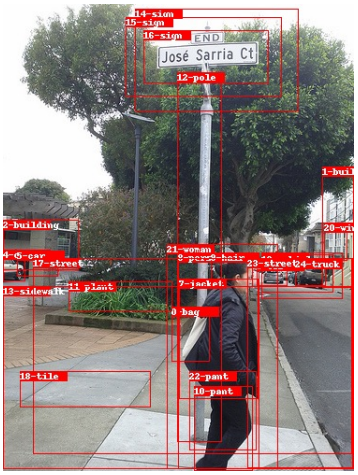
Figure 2. Relation-wise Recall using the VCTree-EBM [2] as the backbone SGG model trained with our proposed method (Ours) vs. the method proposed in [2] (Baseline).

baseline shows that our method maintains the performance on frequent relations while improving significantly on the more informative/infrequent relation classes (as measured by mean recall).

Qualitative Studies. We present additional qualitative visualizations in Figure 3 and Figure 4. Our proposed method predicts informative relations for both the set of pairs present in the ground truth and new set of object pairs that further helps to define a scene comprehensively. In the quantitative evaluation we only reward object pairs that have corresponding ground truth relations, hence, the relations for the remaining set of object pairs can only be visualized qualitatively.

Models	Method	Predicate Classification			Scene Graph Classification			Scene Graph Detection		
		R@20	R@50	R@100	R@20	R@50	R@100	R@20	R@50	R@100
Motif-TDE-Sum [3, 6]	Baseline	33.38	45.88	51.25	20.47	26.31	28.79	11.92	16.56	20.15
	Ours	33.36	43.53	47.44	24.31	29.91	31.75	14.59	17.96	19.70
VCTree [4]	Baseline	59.82	65.93	67.57	41.49	45.16	46.10	24.90	32.02	36.30
	Ours	58.66	64.69	67.05	35.49	38.71	39.51	24.63	31.52	36.42
VCTree-EBM [2]	Baseline	57.31	63.99	65.84	40.31	44.72	45.84	24.21	31.36	35.87
	Ours	57.42	64.37	66.43	35.42	38.79	39.66	23.70	30.74	35.62
VCTree-TDE [3]	Baseline	40.12	50.83	54.91	26.00	33.03	35.97	13.97	19.43	23.34
	Ours	36.90	47.62	52.03	25.67	32.83	35.76	15.20	19.00	20.98

Table 4. Scene Graph Generation performance comparison on Regular Recall@K [3] under all three experimental settings. We compare the results of our proposed framework (Ours) with the original model (Baseline) using different SGG architectures.



Ground Truth Triplets

21-woman **wearing** 7-jacket
21-woman **wearing** 22-pant
23-street **in front of** 21-woman
14-sign **on** 12-pole
8-person **wearing** 7-jacket
8-person **walking on** 13-sidewalk
20-window **on** 1-building
8-person **carrying** 0-bag
7-jacket **on** 8-person
10-pant **on** 21-woman

Predicted Triplets

6-car **parked on** 3-street
16-sign **attached to** 12-pole
8-person **carrying** 0-bag
8-person **walking on** 13-sidewalk
3-car **parked on** 23-street
21-woman **carrying** 0-bag
19-vehicle **parked on** 23-street
20-window **of** 1-building
21-woman **walking on** 13-sidewalk
14-sign **attached to** 12-pole
18-tile **on** 13-sidewalk

Ground Truth Triplets

2-man **on** 3-sign
5-window **on** 0-building
5-window **on** 7-window

Predicted Triplets

5-window **of** 0-building
4-sign **attached to** 6-building
4-sign **attached to** 0-building
0-building **has** 7-window
1-fence **sitting on** 0-building
2-man **painted on** 3-sign
1-fence **sitting on** 4-sign
7-window **on** 6-building
0-building **behind** 4-sign
3-sign **under** 4-sign
0-building **has** 5-window

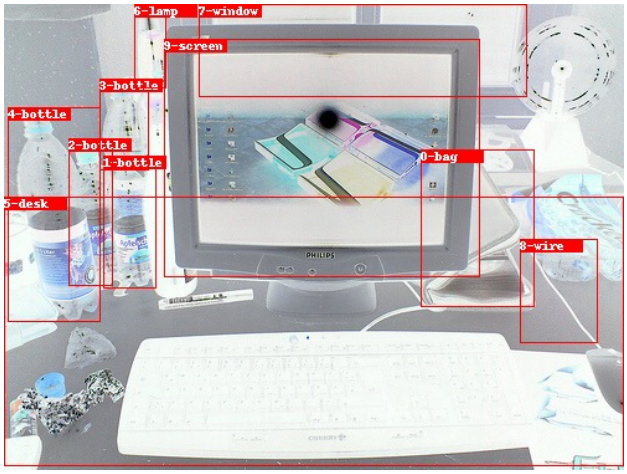
Figure 3. Additional Qualitative Results with the ground truth triplets and the predicted triplets from the VCTree-EBM model trained with our proposed training framework. The predicted triplets are from the SGCls setting.

Ground Truth Triplets

10-woman **with** 9-phone
10-woman **holding** 9-phone
0-bag **with** 1-bottle
3-girl **with** 7-glass
5-girl **with** 8-jean
3-girl **with** 7-glass
10-woman **with** 9-phone
6-girl **with** 1-bottle
6-girl **with** 0-bag
3-girl **with** 7-glass
5-girl **with** 9-phone

Predicted Triplets

3-girl **wearing** 7-glass
10-woman **holding** 9-phone
5-girl **looking at** 9-phone
2-boy **looking at** 9-phone
1-bottle **in** 0-bag
4-girl **holding** 9-phone
3-girl **looking at** 9-phone
0-bag **has** 1-bottle
10-woman **with** 4-girl
7-glass **on** 5-girl
6-girl **wearing** 8-jean
4-girl **wearing** 8-jean



Ground Truth Triplets

4-bottle **near** 9-screen

Predicted Triplets

8-wire l**aying on** 5-desk
8-wire **holding** 6-lamp
1-bottle **behind** 9-screen
7-window **on** 9-screen
4-bottle **sitting on** 5-desk
0-bag **laying on** 5-desk
1-bottle **sitting on** 5-desk
2-bottle **on** 5-desk
6-lamp **on** 5-desk
9-screen **laying on** 5-desk
3-bottle **on** 5-desk
2-bottle **in** 7-window
0-bag **laying on** 9-screen
5-desk **near** 7-window
4-bottle **near** 8-wire

Figure 4. Additional Qualitative Results.

References

[1] 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. *International journal of computer vision*, 123(1):32–73, 2017. 1, 2

[2] 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. 1, 2, 3

[3] 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, 3

[4] Kaihua Tang, Hanwang Zhang, Baoyuan Wu, Wenhan Luo, and Wei Liu. Learning to compose dynamic tree structures for visual contexts. In *Conference on Computer Vision and Pattern Recognition*, 2019. 3

[5] Danfei Xu, Yuke Zhu, Christopher B Choy, and Li Fei-Fei. Scene graph generation by iterative message passing. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5410–5419, 2017. 1

[6] Rowan Zellers, Mark Yatskar, Sam Thomson, and Yejin Choi. Neural motifs: Scene graph parsing with global context. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5831–5840, 2018. 1, 3