Learning of Visual Relations: The Devil is in the Tails – Supplementary Material

Alakh Desai*[1], Tz-Ying Wu*[1], Subarna Tripathi[2], and Nuno Vasconcelos[1]

[1]University of California San Diego, USA
[2]Intel Labs, USA

1. Model complexity

The model complexity is quite low for DT2, which has **10**× fewer trainable parameters than most of the recent approaches in the literature. For example, the SGCls model sizes of DT2, MOTIFS [5], VCTree [3] and TDE-MOTIFS [2] are **224 MB**, 1.68 GB, 1.65 GB and 2.1 GB respectively. This is by design, since our goal is to emphasize the importance of accounting for long tails during training.

2. Additional results

In this section, we provide additional results of the proposed DT2-ACBS.

2.1. More visual examples

In addition to section **??**, Figure 1 presents more visual examples of PredCls (left column) and SGCls (right column) generated by DT2-ACBS. These examples show that DT2-ACBS can predict predicates ranging from head classes such as *has* and *wearing* to less populated classes like *walking on*. While some examples are counted as incorrect predictions under the metric, as discussed in the main paper, they are still reasonable predictions such as a subclass or a synonym of the ground truth. For example, *belonging to/of* and *wears/wearing*. In the SGCls task, DT2-ACBS correctly predicts head entities (*boy*, *horse*, and *head*) and tail entities (*racket* and *sock*).

2.2. Ablations on appearance branch

As discussed in section **??**, the goal of appearance branch is to convey the image information *not* encoded in the entity labels but *relevant to predicate predictions.* We tested the effectiveness of the appearance branch F^a by removing it and training the network with ACBS. Table 1 shows that entity classification accuracy remains similar, but PredCls and SGCls performance drops dramatically, i.e.

*Authors have equal contributions.

Table 1. Ablations of appearance branch in SGCls. (subj, obj) Acc. denotes the accuracy of a pair of subject and object class.

Method	PredCls mR@ 20 / 50 / 100	SGCls mR@ 20 / 50 / 100	(subj, obj) Acc.
w/o F^a	18.1 / 24.5 / 26.8	11.0 / 14.7 / 16.3	25.77
w/ F^a (ours)	**27.4 / 35.9 / 39.7**	**18.7 / 24.8 / 27.5**	26.26

Table 2. Ablations of ACBS with different teachers in SGCls.

Teacher	mR@ 20 / 50 / 100
E-step	15.2 / 20.2 / 22.0
P-step (ours)	**18.7 / 24.8 / 27.5**

the appearance branch contributes substantially to predicate classification. Note that the gains hold even when the ground truth entity labels are used (PredCls), confirming the argument that simply knowing entity classes is not enough for predicate prediction.

2.3. E-step as teacher

Note that the predicate $\mathbf{W}^p$ and entity $\mathbf{W}^e$ weight matrices are interdependent. Using E-step as the teacher would negatively affect $\mathbf{W}^p$. In ACBS, $\mathbf{W}^e$ receives class-balanced *entity supervision,* so there is no risk of overfitting. The role of the teacher is to guarantee that the E-step update of $\mathbf{W}^e$ is not incompatible with the P-step update of $\mathbf{W}^p$. This distillation is exactly how ACBS fuses the knowledge learnt with different distributions. Using E-step as the teacher has weaker results, as shown in Table 2.

2.4. Recall values

The metric of Recall@K is highly biased toward dominated classes (such as "on"), and thus it is not suitable for long-tailed visual relations, as discussed in the main paper. However, we provide the numbers in Table 3 for reference.

3. Implementation Details

DT2-ACBS is a two-stage training process. While SRS is adopted in the first stage when training the parameter of θ, ϕ and ψ, the proposed ACBS is adopted in the second stage to learn the classifiers. Apart from the differences in sampling strategies, both stages share a similar optimization

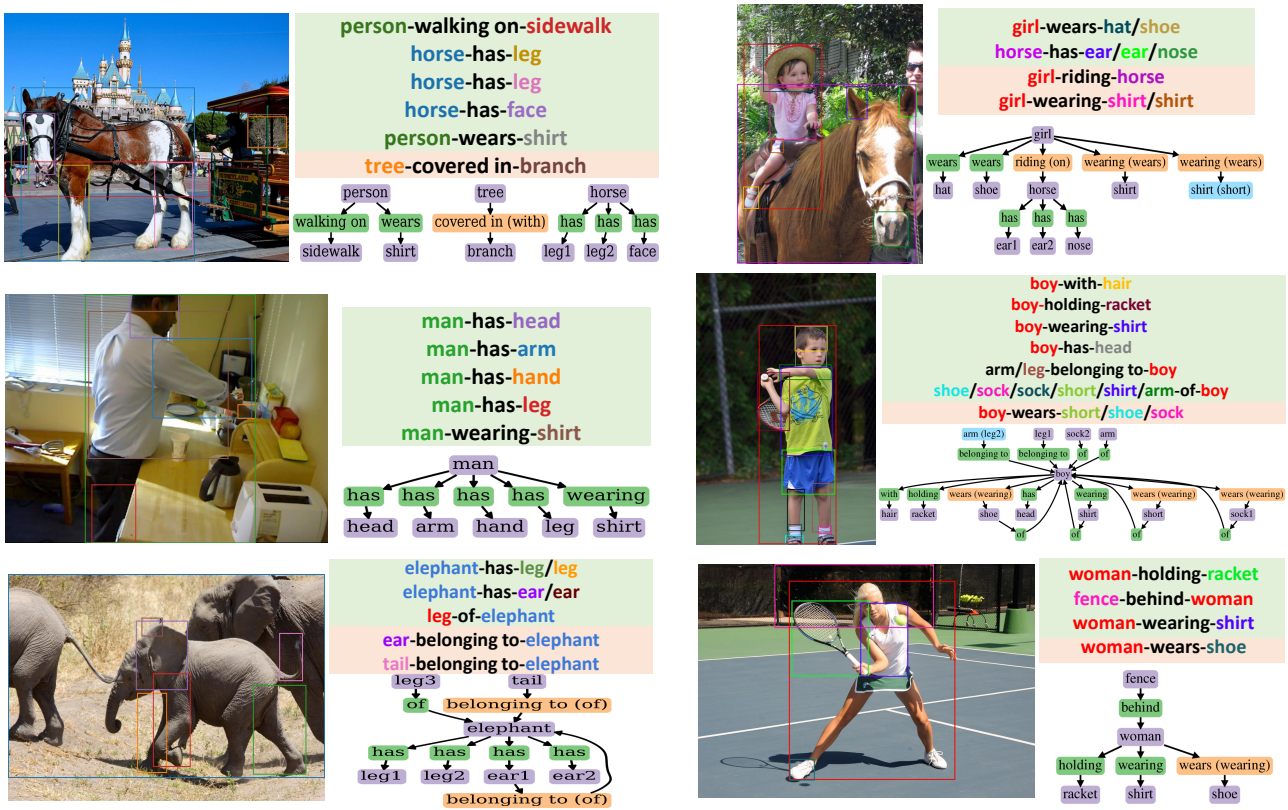

Figure 1. **Additional visual examples of PredCls (left column) and SGCls (right column).** In each sub-figure, colors of bounding boxes in the image (left) are corresponding to the entities in the triplets (upper-right) with the background color green/orange for correct/incorrect predicate predictions. In the generated graphs (lower-right), correct/incorrect predictions of entities and predicates are shown in purple/blue and green/orange respectively, with the ground truth noted in the bracket (best viewed in color).

Table 3. Recall and mRecall values for SGG tasks.

Method	Predicate Classification		Scene Graph Classification		Scene Graph Detection	
	R@50 / 100	mR@50 / 100	R@50 / 100	mR@50 / 100	R@50 / 100	mR@50 / 100
KERN [1]	65.8 / 67.6	17.7 / 19.2	36.7 / 37.4	9.4 / 10.0	27.1 / 29.8	6.4 / 7.3
TDE-MOTIFS-SUM [2]	46.2 / 51.4	25.5 / 29.1	27.7 / 29.9	13.1 / 14.9	16.9 / 20.3	8.2 / 9.8
TDE-VCTree-SUM [2]	47.2 / 51.6	25.4 / 28.7	25.4 / 27.9	12.2 / 14.0	19.4 / 23.2	9.3 / 11.1
PCPL [4]	50.8 / 52.6	35.2 / 37.8	27.6 / 28.4	18.6 / 19.6	14.6 / 18.6	9.5 / 11.7
DT2-ACBS (ours)	23.3 / 25.6	35.9 / 39.7	16.2 / 17.6	24.8 / 27.5	15.0 / 16.3	22.0 / 24.4

scheme, where the Adam optimizer with initial learning rate 10^{-3} is adopted, with the learning rate decay of 0.5 for every 5 epochs. The batch size in the first stage is 256, while in the second stage, objects and predicates are sampled with 2 and 5 samples per class respectively. The hyperparameters α, β and τ_s are set to 0.2, 1 and 10 respectively using the validation set. For evaluation on SGG tasks, we adopt the protocol of [5, 3] to filter out the subject-object pairs that do not have a relationship.

References

[1] Tianshui Chen, Weihao Yu, Riquan Chen, and Liang Lin. Knowledge-embedded routing network for scene graph generation. In *Conference on Computer Vision and Pattern Recognition*, 2019. 2

[2] Kaihua Tang, Yulei Niu, Jianqiang Huang, Jiaxin Shi, and Hanwang Zhang. Unbiased scene graph generation from biased training. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 3713–3722. IEEE, 2020. 1, 2

[3] Kaihua Tang, Hanwang Zhang, Baoyuan Wu, Wenhan Luo, and Wei Liu. Learning to compose dynamic tree structures for visual contexts. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 1, 2

[4] Shaotian Yan, Chen Shen, Zhongming Jin, Jianqiang Huang, Rongxin Jiang, Yaowu Chen, and Xian-Sheng Hua. PCPL: predicate-correlation perception learning for unbiased scene graph generation. In *MM '20: The 28th ACM International Conference on Multimedia, Virtual Event / Seattle, WA, USA, October 12-16, 2020*, pages 265–273, 2020. 2

[5] Rowan Zellers, Mark Yatskar, Sam Thomson, and Yejin Choi. Neural motifs: Scene graph parsing with global context. *CoRR*, abs/1711.06640, 2017. 1, 2