Self-Supervised Relation Alignment for Scene Graph Generation (Supplementary Material)

Bicheng Xu[1,2] Renjie Liao[1,2,3] Leonid Sigal[1,2,3]
[1]University of British Columbia [2]Vector Institute for AI [3]Canada CIFAR AI Chair
bichengx@cs.ubc.ca rjliao@ece.ubc.ca lsigal@cs.ubc.ca

1. Hyperparameters

1.1. Model Performance w.r.t. the Masking Ratio

We conduct ablation experiments with respect to the masking ratio p (in Equation 1 of the main paper) on the Neural Motifs [3] model under the PredCls setting. Table 1 below shows the validation mean-Recall results with different p values, which are obtained from a single run on the same machine with the same random seed. The loss weight λ (in Equation 3 of the main paper) is fixed to 10 for all the experiments. As can be seen, p being 0.1 gives us the best validation results among the values we tried. Based on this ablation, we set p to be 0.1 for other Neural Motif's training settings (SGCls and SGDet), and for SGTR. Notably, even with a sub-optimal p value, over a wide range of p values, we obtain significant improvements over the Motifs baseline.

Method	mR@50	mR@100
Motifs	22.1	24.2
Align-Motifs ($p = 0.05$)	23.7	25.9
Align-Motifs ($p = 0.1$)	**23.9**	**26.2**
Align-Motifs ($p = 0.2$)	23.5	25.8
Align-Motifs ($p = 0.4$)	23.2	25.4
Align-Motifs ($p = 0.6$)	22.8	24.8

Table 1. **Validation results of different masking ratio p values for the Neural Motifs model under the PredCls setting.** Motifs is our trained Neural Motifs [3] model. Align-Motifs is the Neural Motifs model containing our proposed self-supervised relation alignment mechanism during training, where the p value in the bracket is the masking ratio used for the experiment.

1.2. Model Performance w.r.t. the Loss Weight

Again, on the Neural Motifs [3] model under the PredCls setting, we conduct ablation experiments with respect to the loss weight λ (in Equation 3 of the main paper). Table 2 below shows the validation mean-Recall results with different λ values, which are obtained from a single run on the same machine with the same random seed. The masking ratio p (in Equation 1 of the main paper) is fixed to 0.1 for all the experiments. As the results suggest, λ being 10 gives us the most effective validation results among the values we experimented. Based on this ablation, we set λ to be 10 for all other experiment settings.

Method	mR@50	mR@100
Motifs	22.1	24.2
Align-Motifs ($\lambda = 0.1$)	22.0	23.5
Align-Motifs ($\lambda = 1$)	23.8	25.9
Align-Motifs ($\lambda = 10$)	**23.9**	**26.2**
Align-Motifs ($\lambda = 50$)	22.5	24.2
Align-Motifs ($\lambda = 100$)	21.7	23.4

Table 2. **Validation results of different loss weight λ values for the Neural Motifs model under the PredCls setting.** Motifs is our trained Neural Motifs [3] model. Align-Motifs is the Neural Motifs model containing our proposed self-supervised relation alignment mechanism during training, where the λ value in the bracket is the loss weight used for the experiment.

2. Masking Illustration - Align-SGTR*

Figure 1 illustrates the last layer of the mirrored relation predictor (mirrored structural predicate decoder) of Align-SGTR* – the SGTR [1] model equipped with our proposed self-supervised relation alignment mechanism during training. The mirrored structural predicate decoder shares weights with the original one, except for the untied projection heads. Random masking is applied on the attention matrix of the cross-attention Transformer blocks. Mask is generated independently for every cross-attention Transformer block at each Transformer layer, while the alignment losses are only enforced at the last layer.

3. Per-Predicate R@100 Difference - SGTR

Figure 2 shows the per-predicate R@100 difference between Align-SGTR* and SGTR* (the SGTR [1] model

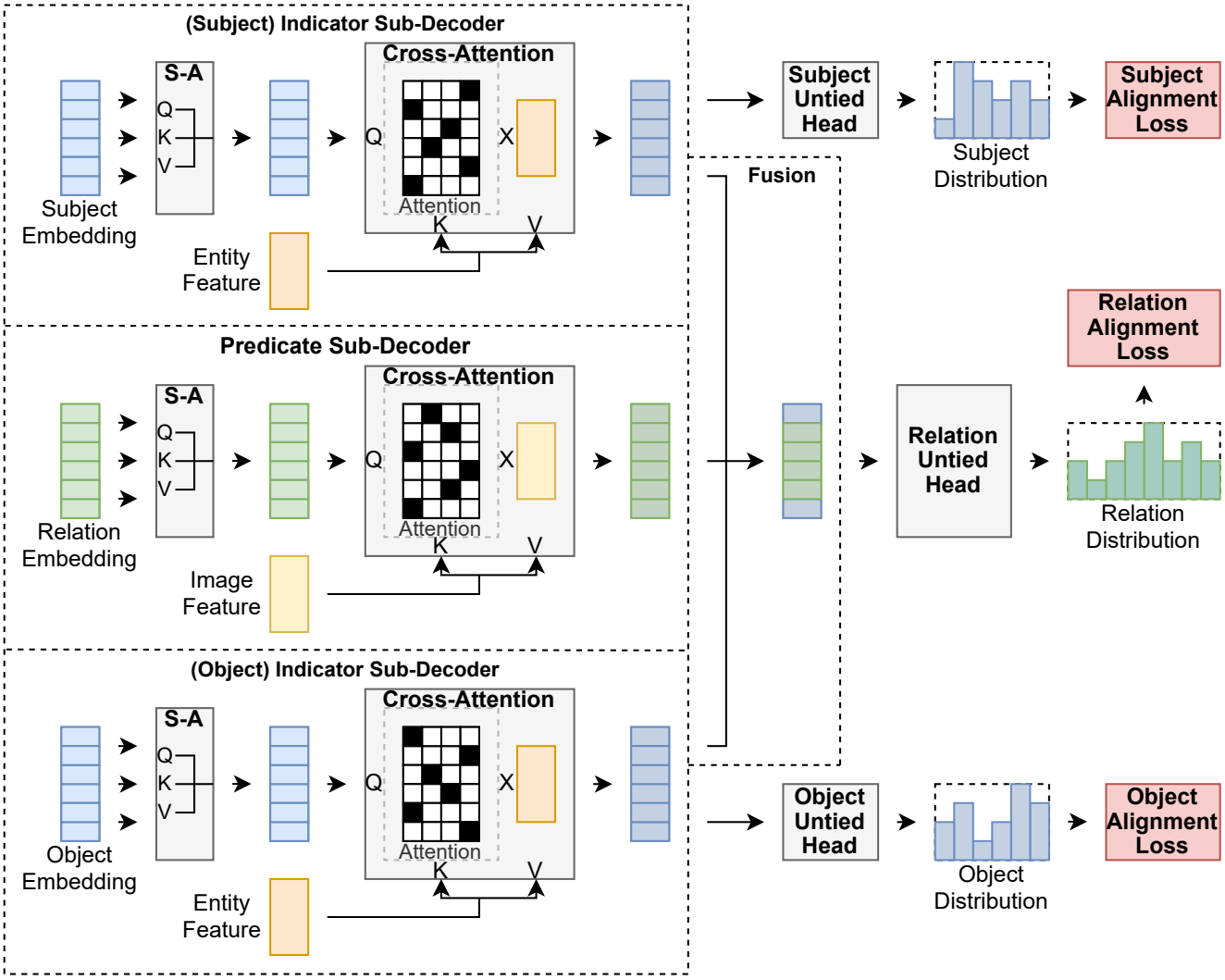

Figure 1. **Instantiation of our proposed self-supervised relation alignment mechanism under SGTR.** This figure illustrates the last Transformer layer of the mirrored structural predicate decoder, which shares weights with the original one, except for the untied projection heads (`Relation/Subject/Object Untied Head`). Random masking is applied on the attention matrix of the cross-attention Transformer blocks. Mask is generated independently for every cross-attention Transformer block at each Transformer layer, while the alignment losses are only enforced at the last layer. In the figure, `S-A` stands for a self-attention Transformer block. `Q`, `K`, and `V` are the `query`, `key`, and `value` of a Transformer block respectively.

trained by us). The predicates are sorted by their frequencies in descending order from left to right. The predicate order, and the head-body-tail partitions are from [2]. Results are averaged across 4 runs. As can be seen, our Align-SGTR* is better than SGTR* on 40 predicate labels (out of a total of 50), in many cases by a sizable margin.

References

[1] 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. 1

[2] 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. 2

[3] 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

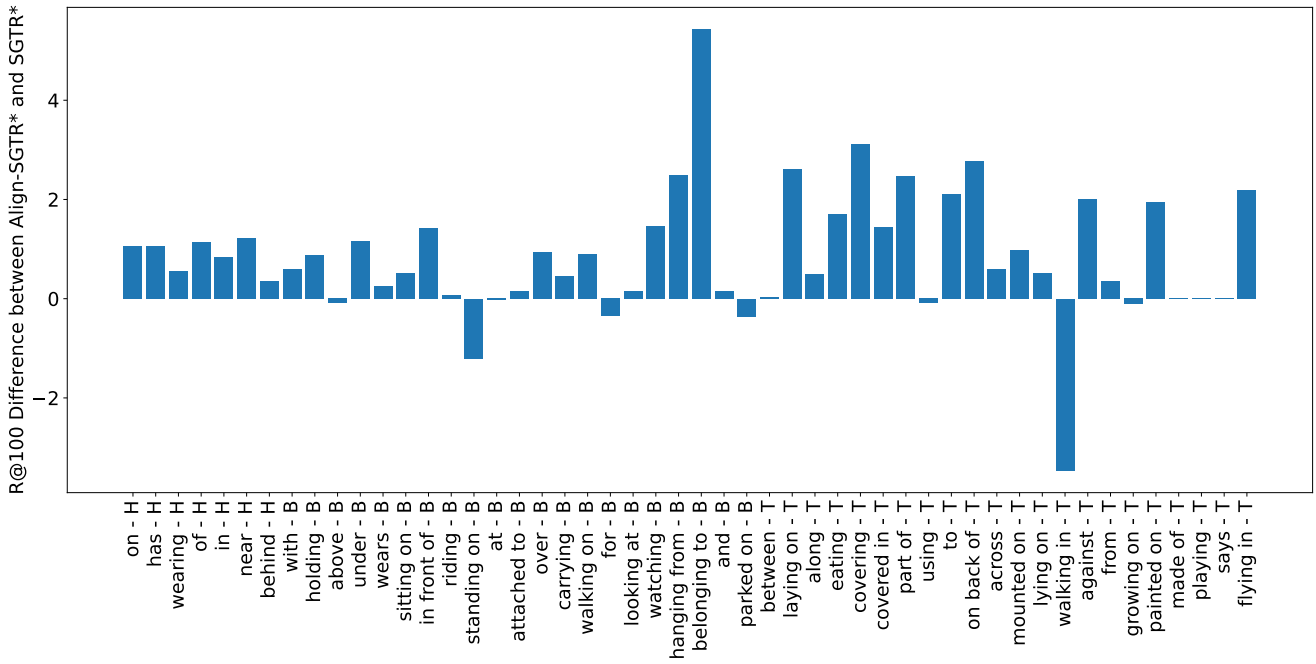

Figure 2. **Per-predicate R@100 difference between Align-SGTR* and SGTR*.** The predicates are sorted by their frequencies in descending order from left to right. `H`, `B`, and `T` indicate the head, body, and tail partitions respectively. Results are averaged across 4 runs.