Relation-aware Instance Refinement for Weakly Supervised Visual Grounding Supplementary materials

Yongfei Liu[1,5,6*] Bo Wan[1,2*] Lin Ma[3] Xuming He[1,4]
[1]School of Information Science and Technology, ShanghaiTech University
[2]Department of Electrical Engineering (ESAT), KU Leuven
[3]Meituan [4]Shanghai Engineering Research Center of Intelligent Vision and Imaging
[5]Shanghai Institute of Microsystem and Information Technology, Chinese Academy of Sciences
[6]University of Chinese Academy of Sciences
{liuyf3, wanbo, hexm}@shanghaitech.edu.cn forest.linma@gmail.com

In this material, we supplement more implementation details and deep discussions of several components in our model as following.

1. Ranking Loss $\mathcal{L}_{rank}$

In this section, we depict the ranking loss for image-caption pairs similar to [3]. Specifically, for each image sentence pair (I, D), we compute the image representation at coarse-level $\mathbf{x}_I^c$ by taking the average pooling of the visual features for all phrases $\{\mathbf{z}_i^c\}_{i=1}^N$ (c.f Eq. 13), and the sentence representation $\mathbf{x}_D$ as the average pooling of the sentence embedding $\mathbf{H}$ in Eq. 1. The similarity $S(I, D)$ between I and D in a minibatch $\mathcal{B}$ is defined as the cosine distance of $\mathbf{x}_I^c$ and $\mathbf{x}_D$. We compute the ranking loss on the coarse-level $\mathcal{L}_{dis}^c$ as follow,

$$\mathcal{L}_{rank}^c = \sum_{D \in \mathcal{B}} \max_{I' \neq I}(0, \Delta - S(I, D) + S(I', D)) + \sum_{I \in \mathcal{B}} \max_{D' \neq D}(0, \Delta - S(I, D) + S(I, D')) \quad (1)$$

Similarly we compute the ranking loss on the fine-level $\mathcal{L}_{rank}^f$ and the total ranking loss $\mathcal{L}_{rank}$ is define as:

$$\mathcal{L}_{rank} = \mathcal{L}_{rank}^c + \mathcal{L}_{rank}^f \quad (2)$$

2. Effectiveness of coarse network

The coarse net aims to select a small set of relevant proposals, which is beneficial to the visual object graph and fine net. To further investigate the effectiveness of this component, we remove it in our model and observe the accuracy drops from 58.3% to 31.2% dramatically due to severe noise propagation within graph net.

3. More results for visual object graph network & relation constraints

In this paper, we take the visual object graph network (VOGN) and relation constraints (RC) as a whole, because our model is not able to encode visual context cues without VOGN, and cannot suppress noise propagation over VOGN without explicitly relationship supervision. To investigate the capability of each component, we conduct drop-one-out ablation studies on our final model, and observe a significant performance drop without any part, as shown in Tab. 1.

Methods	TSD	STR	VOGN	RC	*Acc%*
ours (w/o VOGN&RC)	✓	✓	-	-	56.88
ours (w/o RC)	✓	✓	✓	-	57.10
ours (w/o VOGN)	✓	✓	-	✓	57.13
ours	✓	✓	✓	✓	58.30

Table 1. Ablation Study on Flickr30K Entities val set.

4. Ablation study for four loss terms

Phrase reconstruction loss is a default supervision in our paper and has been widely used in weakly-supervised grounding [2,28,19,20]. As shown in Tab. 3 below, we report the results of using either phrase reconstruction loss ($\mathcal{L}_{rec}$) or ranking loss ($\mathcal{L}_{rank}$) in the baseline model, and show the performance gain of STR loss ($\mathcal{L}_{reg}$) and RC loss ($\mathcal{L}_{rel}$). We observe that all the loss terms are effective in model learning.

5. Relation types encoded in graph network

Our graph net mainly captures semantic and spatial relations, and encodes spatial cues (box locations) in object

*Both authors contributed equally. This work was done when Yongfei Liu was a research intern at Tencent AI Lab, and Bo Wan was a master student in ShanghaiTech University. This work was supported by Shanghai NSF Grant (No. 18ZR1425100).

Methods	people	clothing	bodyparts	animal	vehicles	instruments	scene	other
GroundR[28]	44.32	9.02	0.96	46.91	46.00	19.14	28.23	16.98
KAC Net[2]	58.42	7.63	2.97	**77.80**	69.00	20.37	45.53	17.05
UTG[39]	58.37	14.87	2.29	68.91	55.00	**22.22**	24.87	20.77
MATN[42]	54.71	13.38	2.87	58.31	45.04	19.48	21.97	17.02
KAC Net*[2]	58.42	46.14	23.90	65.06	56.75	9.87	49.87	26.94
ours	**73.70**	**54.68**	**31.36**	77.41	**69.50**	14.81	**58.65**	**37.91**

Table 2. Comparison of phrases grounding accuracy over coarse categories on Flickr30K Entities test set.

Methods	$\mathcal{L}_{rec}$	$\mathcal{L}_{rank}$	$\mathcal{L}_{reg}$ (STR)	$\mathcal{L}_{rel}$ (VOGN&RC)	$Acc\%$
Baseline	✓	-	-	-	47.5
	-	✓	-	-	43.2
	✓	✓	-	-	48.18
Baseline + TSD	✓	✓	-	-	50.80
	✓	✓	✓	-	56.88
	✓	✓	✓	✓	58.30

Table 3. Ablation Studies of four loss terms on Flickr30K val set.

feature as in [21]. Concretely, we select top-88 frequent relation types on Flickr30K (63% for semantic and 37% for spatial) and top-34 relations on ReferitGame (34% for semantic, 66% for spatial).

We further investigate the efficacy of relation encoding in Flickr30K, and report relations classification accuracy in Tab. 4 below. It shows our relation encoding module can capture semantic and spatial relations indeed.

	# classes	top-1 (%)	top-5 (%)	top-10 (%)
semantic	67	41.2	79.1	88.6
spatial	21	53.1	84.6	91.8
all	88	45.6	81.1	89.8

Table 4. Relation classification results on Flickr30K val set.

6. Coarse Categories Accuracy

As shown in Tab. 2, our method outperforms the previous state-of-the-art in most coarse categories in Flickr30k test set, which validates the effectiveness of our network. In addition, our model performs inferior result in instruments category, which is caused by lower proposal recall when using object detector pretrained on Visual Genome dataset. We find that most instruments phrases are ”guitar”, which is not contained in Visual Genome category space.

7. Comparison with Concurrent Work

We compare our model with concurrent work MAF Net [35], of which feature extractor is pretrained with additional supervision from object attributes on Visual Genome dataset. For a fair comparison and keeping in line with the previous works [2, 7], we re-implement their released code with the same visul features as ours, denoted as MAF*. As shown in Tab. 5, we outperforms MAF Net with 1.01% grounding accuracy, which validates the superiority of our proposed flexible and context-aware object representation for weakly supervised visual grounding.

Methods	$Acc\%$
MAF* [35]	58.26
ours	59.27

Table 5. Comparison with concurrent work on Flickr30K Entities test set.

8. Implementation details for ReferItGame dataset

For the visual feature extraction, we take the same object detector pretrained on Visual Genome to generate M=50 object proposals and compute their visual representation via RoI-Align. We also select K=5 proposals in coarse-level matching network to suppress most of the background distractors. For the semantic relations, we select top C_r=34 relations whose frequency are greater than 10. It worth noting that we explicitly parse the expression in ReferItGame dataset into $\langle$subject, relation, object$\rangle$ pairs following KPRN [20], and regard the subject as target grounding phrase.

For model learning, we keep the same training configuration as in Flickr30k Entities but the initial learning rate is set to 0.005.