Appendix for Weakly Supervised Referring Image Segmentation with Intra-Chunk and Inter-Chunk Consistency

Anonymous ICCV submission

Paper ID 5477

A1. Detailed Experimental Setup

We utilized images with a resolution of 384×384 pixels for both training and inference. To input data into the visual Transformer, we used a patch size of 16, resulting in a N_w of 24×24. We set batch size to 32, learning rate to 1e-5, and d to 768. We train our model with 5 epochs with adamW optimizer [2]. On the architecture, we follow ALBEF [1]: we use 12-layer visual encoder, 6-layer text encoder, and 6-layer cross-attention transformer. We plan to make our code publicly available upon acceptance.

A2. Additional Analysis

Hyper-parameters: We present the performance using the different values of hyper-parameters. The values of λ_1 and λ_2 control the influence of L_{inter} and L_{intra} respectively (Eq. 2). We first investigate the effectiveness of these hyper-parameters in Table A1. We can see that L_{inter} and L_{intra} can produce satisfactory results with a wide range of values of λ_1 and λ_2. Next, we show the effectiveness of the values of τ (Eq. 4) in Table A2. We can see that τ is even less sensitive than λ_1 and λ_2. Finally, we analyze the effect of the number of MCG proposals in Section 3.3.2. We used 200 MCG proposals for each image. 100 proposals drop the performance by 0.9%p. No performance change is observed with $>$ 200 proposals.

Negative Meaning: As mentioned in Section 4.3 in the main paper, our method does not work properly for the negative terms (*e.g.,* 'no', 'not'). We can partly address this issue with the utilization of box labels. We collect the 'Negative' set by selectively collecting the expressions containing one of the following words: 'no', 'not', and 'without'. Table A3 compares the results on the 'Negative' set with 'All' set. Learning with image-text pairs produces 24.3% drop for the 'Negative' set, while learning with boxes produces 16.5% drop. Negative meanings can be effectively modeled through explicit localization cues provided by the box. However, learning negative expressions is still a challenging problem. We expect that more sophisticated language modeling is required.

A3. Additional Examples

We present additional examples of localization maps produced by GroupViT [3], ALBEF [1], and our method.

References

[1] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705, 2021. 1, 3

[2] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 1

[3] Jiarui Xu, Shalini De Mello, Sifei Liu, Wonmin Byeon, Thomas Breuel, Jan Kautz, and Xiaolong Wang. Groupvit: Semantic segmentation emerges from text supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18134–18144, 2022. 1, 3

Table A1: Performance with various combinations of values of λ_1 and λ_2.

		RefCOCO			RefCOCO+			G-Ref
λ_1	λ_2	val	testA	testB	val	testA	testB	val
0	0	23.12	22.79	23.42	22.44	22.07	22.51	24.18
0.5	0	25.49	25.16	25.93	24.36	24.17	25.06	26.32
0	1.5	25.29	25.89	25.01	24.67	24.93	24.64	25.37
0.5	1.5	26.45	26.93	26.52	25.49	25.53	25.89	26.61
0.5	3	26.52	27.24	26.64	25.33	25.50	25.82	26.82
1	3	26.19	26.92	26.26	25.68	26.11	25.94	26.97
1	5	25.34	25.98	25.71	25.14	25.79	25.43	27.14
1.5	5	25.47	25.64	25.84	24.35	24.04	25.26	27.06

Table A2: Performance with various values of τ on the RefCOCO+ dataset.

τ	*val*	*testA*	*testB*
0.1	25.81	26.63	25.64
0.15	25.82	26.50	25.81
0.2	25.68	26.11	25.94
0.25	25.44	25.79	25.85
0.3	25.20	25.44	25.59

Table A3: Performance for the 'Negative' and 'All' sets on the RefCOCO+ dataset.

Set	*val*	*testA*	*testB*
Supervision: image-text pairs			
Negative	19.44	18.80	18.14
All	25.68	25.11	25.94
Supervision: boxes			
Negative	40.21	41.56	35.96
All	48.19	53.01	42.83

Figure A1: Examples of localization maps obtained by GroupViT [3], our baseline ALBEF [1], and ours method, which are obtained without refinement techniques.