Supplementary Material for "LAVT: Language-Aware Vision Transformer for Referring Image Segmentation"

Zhao Yang[1]*, Jiaqi Wang[2]*, Yansong Tang[5,1]†, Kai Chen[2,4], Hengshuang Zhao[3,1], Philip H.S. Torr[1]
[1]University of Oxford, [2]Shanghai AI Laboratory, [3]The University of Hong Kong,
[4]SenseTime Research, [5]Tsinghua-Berkeley Shenzhen Institute, Tsinghua University

This supplementary document provides additional discussions, experiment results, and visualizations, which complement those presented in the main paper.

1. Potential biases of the language model

We note that the pre-trained language model BERT [2] (which we employ) has been reported as containing ethnic biases of potential societal concern in some studies. We refer interested readers to the recent work of Ahn *et al*. [1] for more details, in which different kinds (including racial, gender, geological, *etc*.) of ethnic biases are analyzed and mitigation methods are proposed.

2. The language pathway

For the design of our language pathway, we wanted to find a way to allow the vision Transformer layers to embed multi-modal information effectively. As a result, we built the language pathway as a residual connection [3, 6], which has been shown effective for combining features containing different types of information in a deep neural network. And the design of the language gate is inspired by previous work that featured learnable gates for regulating information flow in deep neural networks, such as the LSTM [4], the SENet [5], and CFBI [7].

Method	P@0.5	P@0.7	P@0.9	oIoU	mIoU
*Replacement (w/o LG)	–	–	–	–	–
*Concatenation (w/o LG)	72.89	58.15	20.02	60.52	63.41
Sum (w/o LG)	84.00	74.96	33.47	72.24	73.94
Sum (with LG; default)	**84.46**	**75.28**	**34.30**	**72.73**	**74.46**

Table 1. Design alternatives for the language pathway (annotated with the asterisk). 'LG' is short for language gate. '–' indicates that training suffered extremely slow convergence.

3. Precision-recall analysis

To understand the precision-recall trade-off of LAVT and two of its ablated models, in Fig. 1 we compute and plot the average precision and the average recall of all test samples in the validation set of RefCOCO at 100 thresholds evenly spaced out from 0 to 1 (where the prediction for a pixel is positive if the softmax-normalized score map of the object class exceeds the threshold and is negative otherwise).

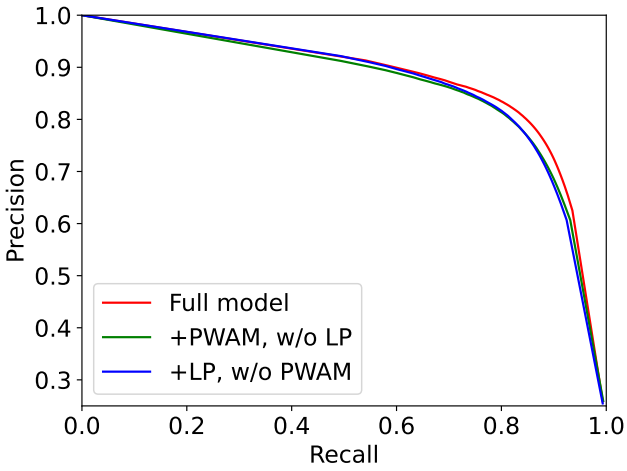

Figure 1. Precision-recall (PR) curves on the RefCOCO validation set. The full model obtains the best PR trade-off compared to the ablated models. Between the "+LP" model (blue) and the "+PWAM" model (green), a close observation will show that LP maintains a slight advantage in precision over PWAM up until around 0.8 recall.

*Equal contribution. †Corresponding author.

4. Mean IoU

Method	Language Model	RefCOCO			RefCOCO+			G-Ref		
		val	test A	test B	val	test A	test B	val (U)	test (U)	val (G)
LAVT (Ours)	BERT	74.46	76.89	70.94	65.81	70.97	59.23	63.34	63.62	63.66

Table 2. Mean IoU of LAVT on the three benchmark datasets. These results complement the overall IoU reported in Table 1 of the main paper. Since mean IoU treats each object equally and does not favor large objects (as overall IoU does), we consider it a fairer metric and recommend more of its use for evaluating this task in the future.

5. Visualizations

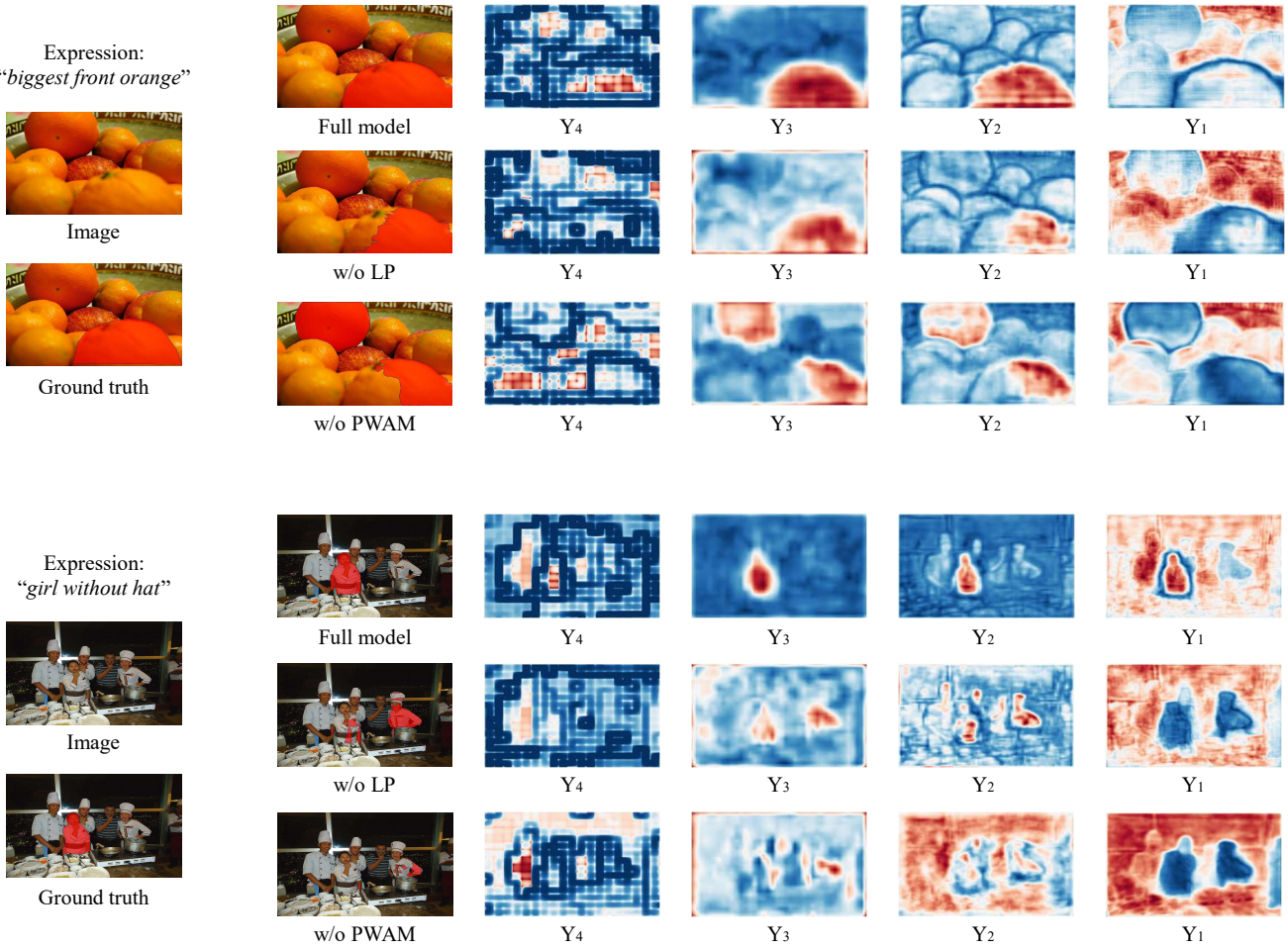

Figure 2. Additional visualizations of predictions and feature maps from the RefCOCO validation set. For each example, the left-most column illustrates the input expression, the input image, and the ground-truth mask overlaid on the input image. In each row, we visualize the predicted mask and the feature maps used for final classification (*i.e.*, Y_4, Y_3, Y_2, and Y_1) from left to right. LP represents the language pathway and PWAM represents the pixel-word attention module.

Figure 3. Visualizations of our predicted masks and the ground-truth masks on examples from the RefCOCO validation set. Examples enclosed with green lines are successful cases, and those enclosed with red lines are failed cases. In the successful cases, our predictions are nearly identical to the ground truth and are sometimes more accurate than the ground truth (see the second example from the right column, where part of the body of the man behind the chair is missing in the annotation). Among the two demonstrated failure cases, the first one is caused by ambiguity in the given expression (there are two boys that are on a skateboard and our model segments out both) and the second one is caused by our model's lack of knowledge of what a "pac man" is (obviously having not played the game Pac-Man, our model fails to associate the shape of the pizza to the shape of a Pac-Man).

References

[1] Jaimeen Ahn and Alice Oh. Mitigating language-dependent ethnic bias in bert. In *EMNLP*, 2021. 1

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In *NAACL*, 2019. 1

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 1

[4] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. In *Neural Computation*, 1997. 1

[5] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *CVPR*, 2018. 1

[6] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *CVPR*, 2017. 1

[7] Zongxin Yang, Yunchao Wei, and Yi Yang. Collaborative video object segmentation by foreground-background integration. In *ECCV*, 2020. 1