Modulating Bottom-Up and Top-Down Visual Processing via Language-Conditional Filters - Supplementary Material

Ilker Kesen[1,2] Ozan Arkan Can[3] Erkut Erdem[1,4] Aykut Erdem[1,2] Deniz Yuret[1,2]

[1] Koç University, KUIS AI Center [2] Koç University, Computer Engineering Department
[3] Amazon Search [4] Hacettepe University, Computer Engineering Department

A. Supplementary Material

This supplementary material contains the implementation details (Section A.1) and the complete ablation studies (Section A.2) of our work.

A.1. Implementation Details

Referring Expression Segmentation. Following previous work [1, 6, 9, 11], we limit the maximum length of expressions to 20. We set input image size to 512×512 and 640×640 for training and inference phase respectively. We use the first four layers of DeepLabv3+ with ResNet-101 backbone, pre-trained on COCO dataset by excluding images appear on the validation and the test sets of UNC, UNC+ and G-Ref datasets similar to previous work [4, 7, 12]. Thus, our low-level visual feature map I has the size of $64 \times 64 \times 64 \times 1032$ in training, and $80 \times 80 \times 1032$ in inference phase, both including 8-D location features. In all convolutional layers, we set the filter size, stride, and number of filters (ch) as $(5, 5)$, 2, and 512, respectively. The depth is 4 in the multimodal encoder part of the network. We apply dropout regularization [10] to language representation r with 0.2 probability. We use Adam optimizer [5] with default parameters. We freeze the DeepLab-v3+ ResNet-101 weights. There are 32 examples in each minibatch. We train our model for 20 epochs on a Tesla V100 GPU with mixed precision and each epoch takes at most two hours depending on the dataset.

Language-guided Image Colorization. Unless otherwise speficied, we follow the same design choices applied for the referring expression segmentation task. We set the number of language-conditional filters as 512, replace the LSTM encoder with a BiLSTM encoder, and we use the first two layers of ResNet-101 trained on ImageNet as image encoder to have a similar model capacity and make a fair comparison with the previous work [8]. We set input image width and height to 224 in both training and validation. Thus, the low-level visual feature map has the size of $28 \times 28 \times 512$, and we don't use location features. Additionally, in our experimental analysis, we consider the same design choices with previous work [8, 13]. Specifically, we use LAB color space, and our model predicts ab color values for all the pixels of the input image. We perform the class re-balancing procedure to obtain class weights for weighted cross entropy objective. We use 313 ab classes present in ImageNet dataset, and encode ab color values to classes by assigning them to their nearest neighbors. We use input images with a size of 224×224, and output target images with a size of 56×56 which is same with the previous work.

A.2. Ablation Studies

We performed additional ablation experiments on *referring expression segmentation* task in order to understand the contributions of the remaining components of our model. We share results in Table A1. Each row stands for a different architectural setup. Horizontal lines separate the different ablation studies we performed, and first column denotes the ablation study group. Columns on the left determine these architectural setups. ✓ on the *Top-down* column indicates that the corresponding setup modulates top-down visual branch with language, and similarly ✓ on the *Bottom-up* column indicates that the corresponding setup modulates bottom-up visual branch with language. *Depth* indicates how many layers the multi-modal encoder has. *Layer* indicates the type of language-conditional layer used. *Visual* and *Textual* indicates which visual encoder and textual encoder used for the corresponding setup, respectively. The remaining columns stand for results.

Network Depth (2). We performed experiments by varying the depth size of the multi-modal encoder. We originally started with the depth size of 4. Increasing the depth size slightly increased the scores for some metrics, but more importantly, decreasing the depth size caused the model to perform worse than the bottom-up baseline. This happens because decreasing the depth size shrinks the receptive field of the network, and the model becomes less capable of drawing conclusions for the scenes that requires to be seen as a whole in order to fully understand.

FiLM vs. Language-conditional Filters (3). Another

#	Top-down	Bottom-up	Depth	Layer	Visual	Textual	*p@0.5*	*p@0.6*	*p@0.7*	*p@0.8*	*p@0.9*	*IoU*
1	✓		4	Conv	ResNet-50	LSTM	66.40	58.59	49.35	36.01	13.42	58.06
		✓	4	Conv	ResNet-50	LSTM	71.40	65.14	57.36	45.11	19.04	60.74
	✓	✓	4	Conv	ResNet-50	LSTM	75.12	70.08	63.32	50.50	22.29	63.59
2	✓	✓	3	Conv	ResNet-50	LSTM	69.96	63.13	55.04	41.33	15.98	60.23
	✓	✓	4	Conv	ResNet-50	LSTM	75.12	70.08	63.32	50.50	22.29	63.59
	✓	✓	5	Conv	ResNet-50	LSTM	75.56	70.59	63.82	51.68	**22.84**	63.52
3	✓	✓	4	Conv	ResNet-50	LSTM	75.12	70.08	63.32	50.50	22.29	63.59
	✓	✓	4	FiLM	ResNet-50	LSTM	71.18	65.14	57.32	44.66	18.75	61.12
4	✓	✓	4	Conv	ResNet-50	LSTM	75.12	70.08	63.32	50.50	22.29	63.59
	✓	✓	4	Conv	ResNet-50	BERT	75.60	70.39	63.05	49.93	21.16	63.57
5	✓	✓	4	Conv	ResNet-50	LSTM	75.12	70.08	63.32	50.50	22.29	63.59
	✓	✓	4	Conv	ResNet-101	LSTM	**76.67**	**71.77**	**64.76**	**51.69**	22.73	**64.63**

Table A1. The complete ablation studies on the UNC validation set with $p@X$ and overall IoU metrics.

method for modulating language is using conditional batch normalization [2] or its successor, FiLM layers. When we replaced language-conditional filters with FiLM layers in our model, we observed ≈2.5 IoU decrease. This is natural, since the FiLM layer can be thought as grouped convolution with language-conditional filters, where the number of groups is equal to number of channels / filters.

LSTM vs. BERT as language encoder (4). We also experimented with BERT [3] as input language encoder in addition to LSTM network. We update BERT weights simultaneously with the rest of our model, where we use a smaller learning rate for BERT (0.00005). We use the *CLS* output embedding as our language representation r, than split this embedding into pieces to create language-conditional filters. Our model achieved similar quantitative results using BERT as language encoder. This points out a language encoder pre-trained on solely textual data might be sub-optimal for integrating vision and language.

The impact of the visual backbone (5). We first start training our model with DeepLabv3+ ResNet-50 backbone pre-trained on Pascal VOC dataset. Then, we pre-trained a DeepLabv3+ with ResNet-101 backbone on COCO dataset by excluding the images appear on the validation and the test splits of all benchmarks similar to the previous work [4, 7, 12]. We only used 20 object categories present in Pascal VOC. Thus, using a more sophisticated visual backbone resulted with ≈ 1 improvement on the IoU score.

References

[1] Ding-Jie Chen, Songhao Jia, Yi-Chen Lo, Hwann-Tzong Chen, and Tyng-Luh Liu. See-through-text grouping for referring image segmentation. In *ICCV*, pages 7454–7463, 2019. 1

[2] Harm De Vries, Florian Strub, Jérémie Mary, Hugo Larochelle, Olivier Pietquin, and Aaron C Courville. Modulating early visual processing by language. In *NeurIPS*, pages 6594–6604, 2017. 2

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *ACL*, pages 4171–4186, June 2019. 2

[4] Henghui Ding, Chang Liu, Suchen Wang, and Xudong Jiang. Vision-language transformer and query generation for referring segmentation. In *ICCV*, 2021. 1, 2

[5] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 1

[6] Chenxi Liu, Zhe L. Lin, Xiaohui Shen, Jimei Yang, Xin Lu, and Alan L. Yuille. Recurrent Multimodal Interaction for Referring Image Segmentation. *ICCV*, pages 1280–1289, 2017. 1

[7] Gen Luo, Yiyi Zhou, Xiaoshuai Sun, Liujuan Cao, Chenglin Wu, Cheng Deng, and Rongrong Ji. Multi-task collaborative network for joint referring expression comprehension and segmentation. In *CVPR*, pages 10034–10043, 2020. 1, 2

[8] Varun Manjunatha, Mohit Iyyer, Jordan Boyd-Graber, and Larry Davis. Learning to color from language. In *NAACL-HLT, vol. 2*, pages 764–769, June 2018. 1

[9] Edgar Margffoy-Tuay, Juan C Pérez, Emilio Botero, and Pablo Arbeláez. Dynamic multimodal instance segmentation guided by natural language queries. In *ECCV*, pages 630–645, 2018. 1

[10] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *JMLR*, 15(1):1929–1958, 2014. 1

[11] Linwei Ye, Mrigank Rochan, Zhi Liu, and Yang Wang. Cross-Modal Self-Attention Network for Referring Image Segmentation. *CVPR*, June 2019. 1

[12] Licheng Yu, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu, Mohit Bansal, and Tamara L. Berg. MAttNet: Modular Attention Network for Referring Expression Comprehension. *CVPR*, June 2018. 1, 2

[13] Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *ECCV*, pages 649–666, 2016. 1