Supplementary Material for Linguistic Structure Guided Context Modeling for Referring Image Segmentation

Anonymous ECCV submission

Paper ID 986

1 Reproducibility

We provide code of our LSCM model in the folder named *code_of_lscm* which is under the same directory as this PDF file.

2 Mutan Fusion

Our model takes an image and a referring expression with T words as input. We use a CNN backbone and a LSTM [2] to extract visual feature $V_i \in \mathbb{R}^{H\times W\times C_v}, i \in \{2,3,4,5\}$ corresponding to different stages in CNN, and words features $Q \in \mathbb{R}^{T\times C_l}$. Since we conduct the same operations on each level of the visual features, we use V to denote a single level of them for ease of presentation. We also use an 8D spatial coordinate feature [4] denoted as $P \in \mathbb{R}^{H\times W\times 8}$. The simplified Mutan fusion [1] which we adopt to fuse $\{V, Q, P\}$ is conducted as follows:

$$\tilde{V} = Conv([V, P]), \tag{1}$$

$$L = MaxPool(Q), \tag{2}$$

$$M_j = (LW_{1j}) \odot (\tilde{V}W_{2j}), \tag{3}$$

$$M = \sum_{j=1}^{r} M_j, \tag{4}$$

where $[,]$ denotes concatenation operation, $W_{1j} \in \mathbb{R}^{C_l\times C_h}$, $W_{2j} \in \mathbb{R}^{C_v\times C_h}$ are learnable parameters, $\odot$ denotes elementwise multiplication, r is a hyperparameter set as 5 in our experiments, $M \in \mathbb{R}^{H\times W\times C_h}$ is the output multimodal feature.

3 Qualitative Results

Qualitative results of our model on four benchmark datasets including UNC [6], UNC+ [6], G-Ref [5] and ReferIt [3] are illustrated in Fig. 1. Our model can well generalize to multiple datasets with different characteristics such as length of expression and entity categories.

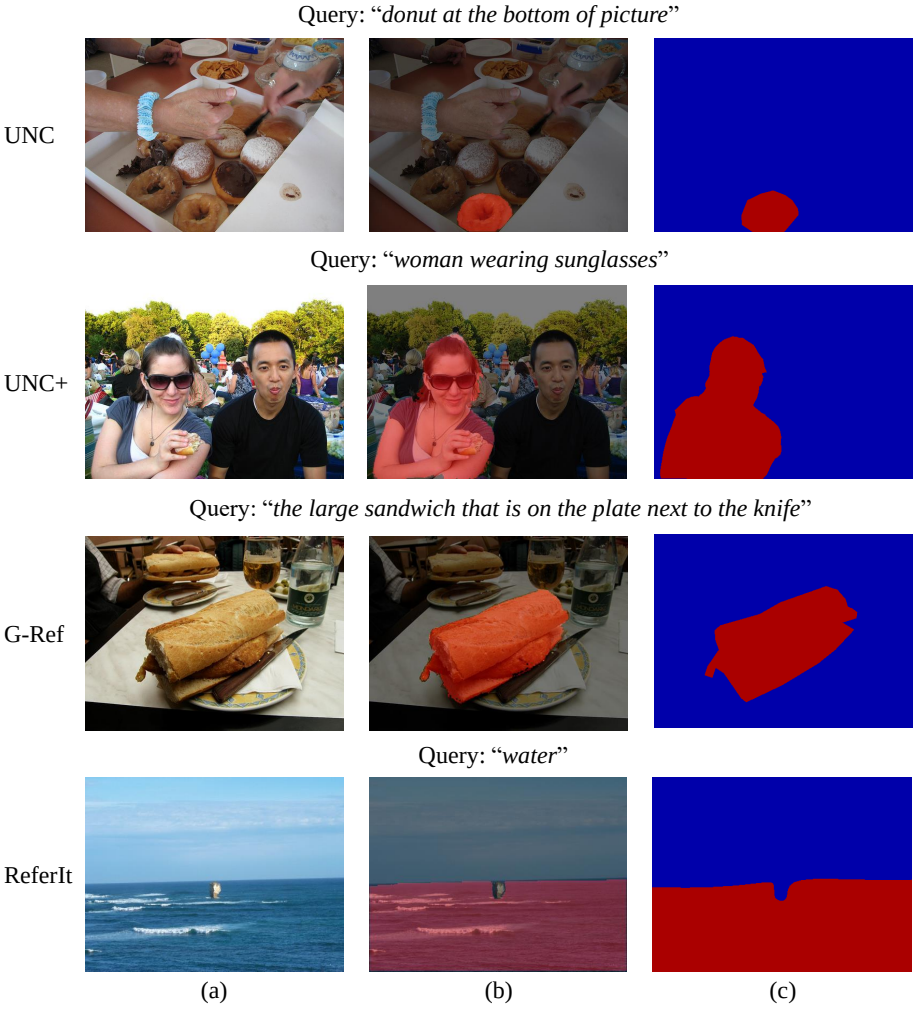

Fig. 1. Qualitative results of our model on four benchmark datasets. (a) Original image. (b) Prediction of our model. (c) Ground-truth.

References

1. Ben-Younes, H., Cadene, R., Cord, M., Thome, N.: Mutan: Multimodal tucker fusion for visual question answering. In: ICCV (2017)
2. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Computation (1997)
3. Kazemzadeh, S., Ordonez, V., Matten, M., Berg, T.: Referitgame: Referring to objects in photographs of natural scenes. In: EMNLP (2014)
4. Liu, C., Lin, Z., Shen, X., Yang, J., Lu, X., Yuille, A.: Recurrent multimodal interaction for referring image segmentation. In: ICCV (2017)
5. Mao, J., Huang, J., Toshev, A., Camburu, O., Yuille, A.L., Murphy, K.: Generation and comprehension of unambiguous object descriptions. In: CVPR (2016)
6. Yu, L., Poirson, P., Yang, S., Berg, A.C., Berg, T.L.: Modeling context in referring expressions. In: ECCV (2016)