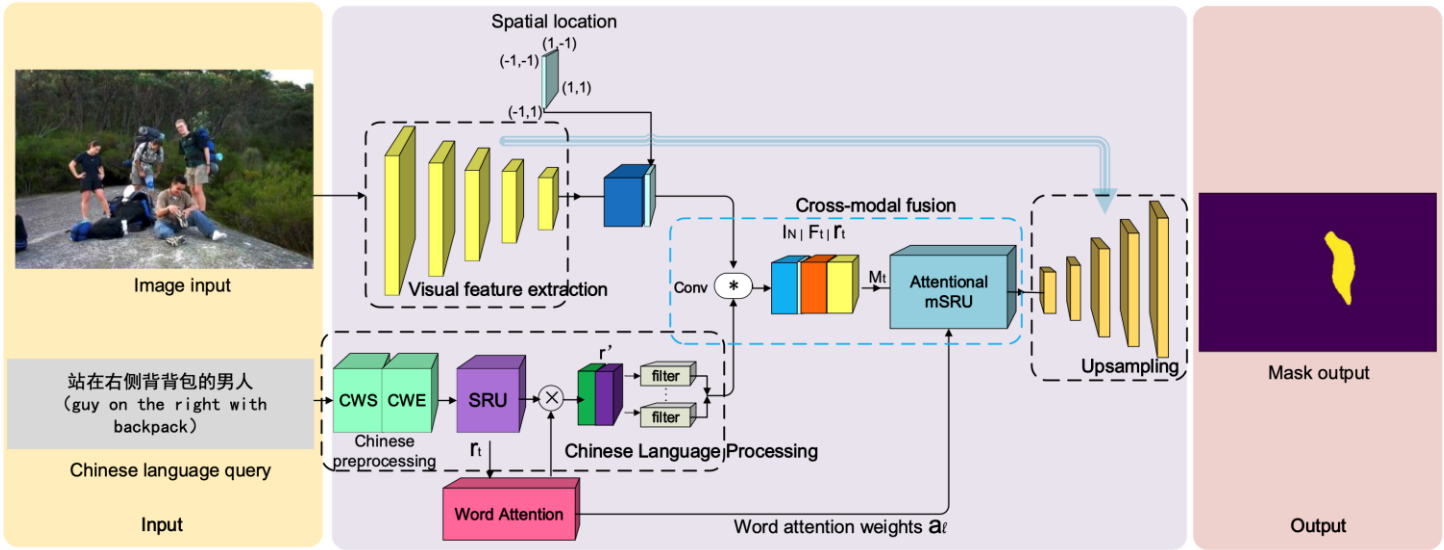
Referring Image Segmentation with Attention Guided Cross Modal Fusion for Semantic Oriented Languages

Qianli ZHOU, Rong WANG, Haimiao HU, Quange TAN, Wenjin ZHANG

Frontiers of Computer Science , DOI: 10.1007/s11704-022-1136-3

Problems & Ideas

- Problems of performing unstably in complicated environment
  - Current methods cannot achieve the similar performance on semantic-oriented language like Chinese due to the different language characteristics.
  - The lack of the collaborative learning of word-level and multimodal-level attention between visual and linguistic modalities for semantic-oriented languages.
- Idea: the word attention and cross-modal fusion collaboration mechanisms are vital to emphasize on the more im_x0002_portant words



Framework of our Proposed model

Main Contributions

- Contributions:
  - We introduce a novel modular network to leverage instance segmentation based on Chinese description.
  - Dual attention-guided learning method has been proposed for both linguistic and cross modal levels to make the network more suitable for processing Chinese expression.
  - Our proposed model achieves competitive performances on four Chinese-based datasets.

Datasets	Splits	Prec@0.5	Prec@0.6	Prec@0.7	Prec@0.8	Prec@0.9	mIoU
Chinese Referit	val	37.83	30.73	24	16.57	8.27	***53.74***
	test	37.16	30.14	22.8	16.11	7.86	***51.26***
Chinese UNC	val	54.71	41.33	25.6	11.05	1.33	***50.29***
	testA	52.87	45.41	34.61	18.21	2.56	***53.74***
	testB	52.38	43.69	32.85	18.45	3.36	***47.45***
Chinese UNC+	val	32.95	27.51	20.43	11.23	2.82	***40.71***
	testA	44.95	35.87	25.45	13.01	1.97	***42.71***
	testB	35.08	28.33	20.93	12.52	2.27	***34.27***
Chinese Gref	val	27.58	20.11	12.41	5.67	0.72	***30.02***

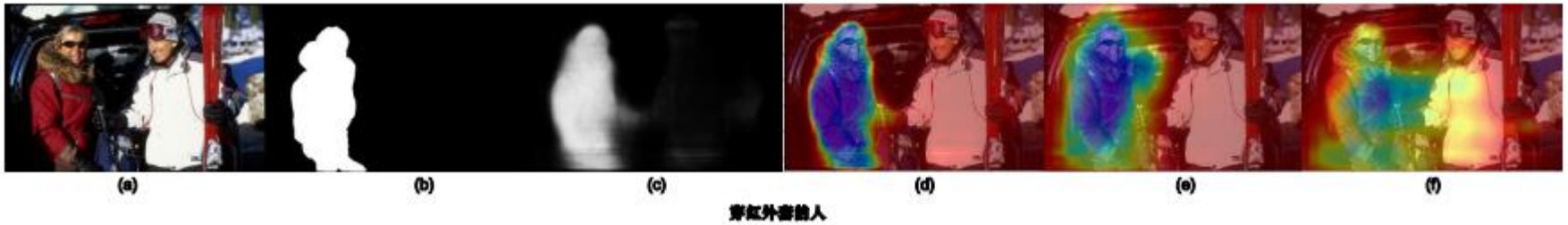
Table 1: The result of our approach. The inference has been carried out on the four datasets and result has been listed in this table.

Main Contributions

- Ablation Study:

Method	Referit Splits	Prec@0.5	Prec@0.6	Prec@0.7	Prec@0.8	Prec@0.9	Overall mIoU
No word attention	val	32.54	26.43	20.2	14.28	6.92	***49.46***
	test	33.6	26.67	19.98	13.9	6.56	***48.8***
No attentional mSRU	val	36.68	30.09	22.56	15.32	7.36	***50.92***
	test	34.4	27.68	21.11	14.3	7.11	***50.37***
Ours	val	37.83	30.73	24	16.57	8.27	***53.74***
	test	37.16	30.14	22.8	16.11	7.86	***51.26***
Ours-EN	val	39.40	30.50	23	15.08	6.92	***53.78***
	test	40.90	31.60	23.12	15.19	7.32	***53.08***

Table 4: Ablation study on the Referit val and test set. No word attention means that we only utilize the Chinese language processing module, and No attentional mSRU means we utilize both language processing and word attention modules but deactivate the attention guidance for mSRU. Our-En represents the results of our model with English support, which we replace the Chinese preprocessing module by normal English process. The results show that our method outperforms the tow variants of the model, as justifying the effectiveness of our attention mechanism.



Visual examples of ablation study. Four examples in rows have been illustrated. (a) is the original image,(b) is the Ground-Truth, (c) is the final result, (d) is the heatmap of our method,(e) is the one without att-mSRU, (f) is the one without both att-mSRU and word attention.