Multiple Transformer Mining for VizWiz Image Caption

Xuchao Gong[1], Hongji Zhu[1], Yongliang Wang[1], Aixi Zhang[1],

Biaolong Chen[1], Fangxun Shu[1], Si Liu[2]

[1]Alibaba Group, [2]Beihang University

2021-06

Participant

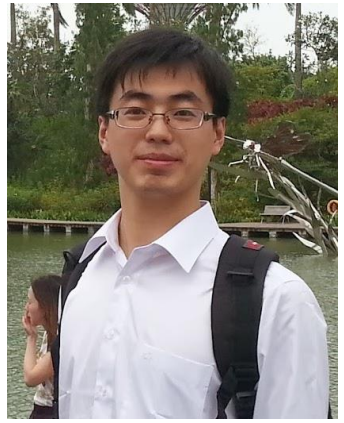
Alibaba Group

Xuchao Gong

Hongji Zhu

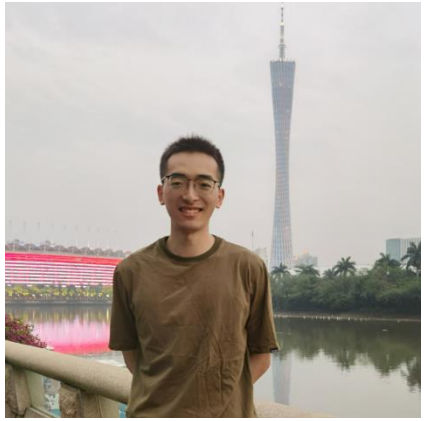
Yongliang Wang

Aixi Zhang

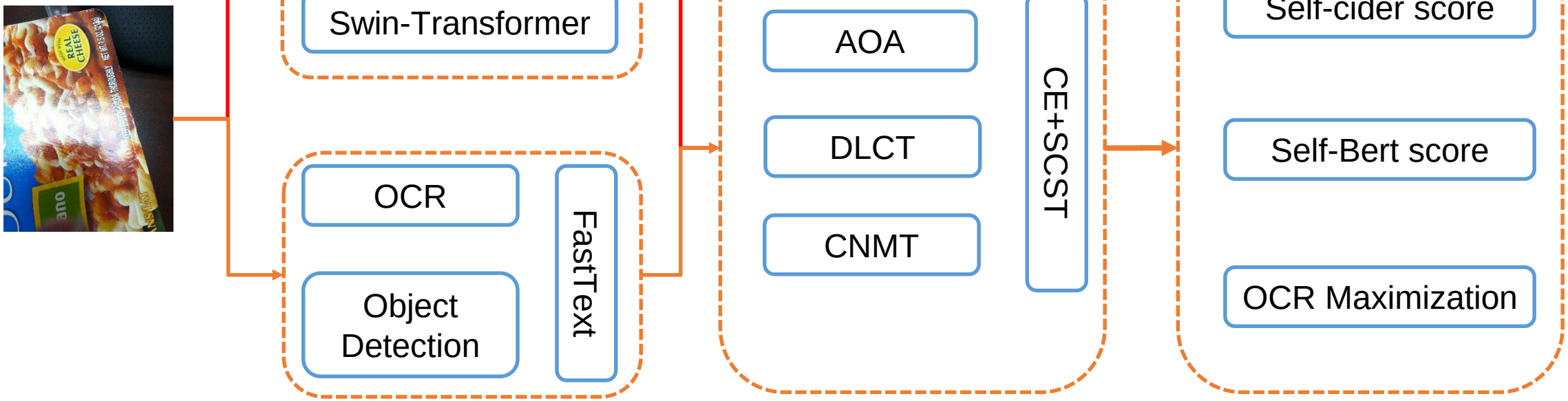


Biaolong Chen

Fangxun Shu



Si Liu


Feature
Multimodal Modeling
Post Processing
ResNext
Swin-Transformer
OCR
Object Detection
FastText
AOA
DLCT
CNMT
CE+SCST
Self-cider score
Self-Bert score
OCR Maximization

Feature

ResNext

Swin-Transformer

Object Detection

OCR

FastText

D. K. Mahajan, R. B. Girshick, V. Ramanathan, K. He, M. Paluri, Y. Li, A. Bharambe, and L. van der Maaten. Exploring the limits of weakly supervised pretraining. In *ECCV*, 2018.

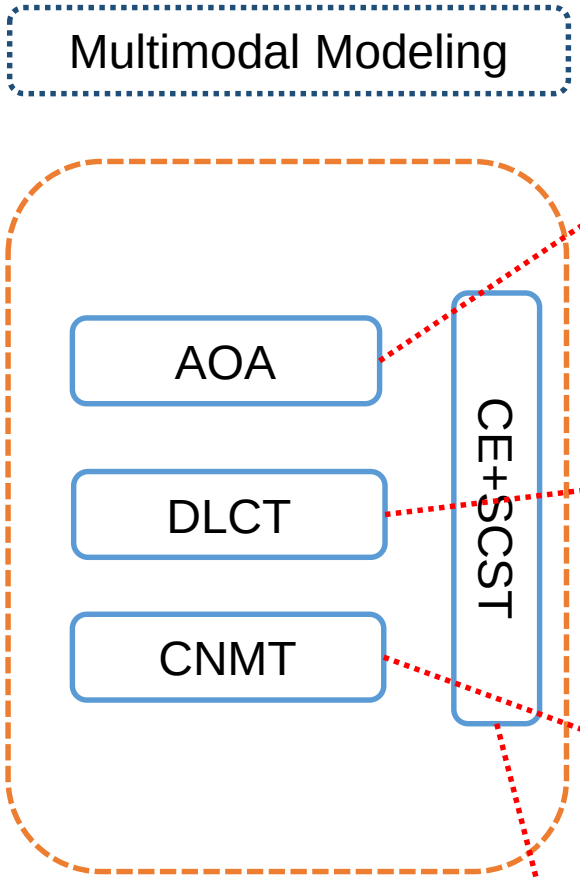
Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo. Swin transformer: Hierarchical vision transformer using shifted windows. *arXiv preprint arXiv:2103.14030*, 2021.

S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *arXiv preprint arXiv:1506.01497*, 2015.

https://github.com/facebookresearch/fastText#text-classification

Y. Baek, B. Lee, D. Han, S. Yun, and H. Lee. Character region awareness for text detection. In *Proceedings of CVPR*, pages 9365 - 9374, 2019.

J. Baek, G. Kim, J. Lee, S. Park, D. Han, S. Yun, S. J. Oh, and H. Lee. What is wrong with scene text recognition model comparisons? dataset and model analysis. In *ICCV*, pages 4715 - 4723, 2019.



L. Huang, W. Wang, J. Chen, and X.-Y. Wei. Attention on attention for image captioning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4634–4643, 2019.

Y. Luo, J. Ji, X. Sun, L. Cao, Y. Wu, F. Huang, C.-W. Lin, and R. Ji. Dual-level collaborative transformer for image captioning. *arXiv preprint arXiv:2101.06462*, 2021.

Z. Wang, R. Bao, Q. Wu, and S. Liu. Confidence-aware non-repetitive multimodal transformers for textcaps. arXiv preprint arXiv:2012.03662, 2020.

S. J. Rennie, E. Marcheret, Y. Mroueh, J. Ross, and V. Goel. Self-critical sequence training for image captioning. InPro-ceedings of CVPR, pages 7008–7024

Alibaba Group

Post Processing

Self-cider score

Qingzhong Wang and Antoni Chan. Describing like humans: on diversity in image captioning. CVPR, 2019

Self-Bert score

T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi.Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*, 2019.

OCR Maximization

We find that selecting a caption whose tokens has most overlap with the OCR tokens helps increase the CIDEr score.

Experimental Results

Alibaba Group

Method	Object Feature	Grid Feature	OCR	SCST	B@1	B@4	M	R	CIDEr	SPICE
AoANet	FRCNN101	-		✓	66.3	23.0	20.1	47.0	61.1	15.4
AoANet†	-	Swin-L		✓	70.3	27.2	22.1	49.6	**77.1**	18.6
DLCT	FRCNN101	ResNeXt101		✓	68.3	24.3	20.4	46.9	63.6	15.7
DLCT†	FRCNN101	Swin-L		✓	70.6	27.1	22.2	49.7	**78.3**	18.6
CNMT	FRCNN101	-	✓		63.7	19.3	18.7	43.7	52.7	13.3
CNMT†	-	ResNeXt101	✓		66.0	21.0	19.6	45.3	60.4	14.9
CNMT†	-	Swin-L	✓		67.9	23.6	21.1	47.1	**71.0**	16.8

Table 1. Comparison of different features. All the results are evaluated on the test-dev. Method names with † are our implement using other feature.

Eval	Model Set	Ensamble	Scorer	B@1	B@4	M	R	CIDEr	SPICE
test-dev	CNMT†	30	bert	71.5	25.6	21.8	49.1	75.8	17.7
	DLCT†	30	bert	72.9	29.3	23.1	51.3	84.1	19.4
	AoANet†	20	bert	74.1	29.0	23.1	51.3	84.5	19.3
	CNMT†+DLCT†+AoANet†	80	bert	75.0	29.6	23.4	51.9	85.7	19.6
	CNMT†+DLCT†+AoANet†	80	cider	74.9	30.7	23.6	51.8	93.7	19.9
	CNMT†+DLCT†+AoANet†	150	cider	74.8	**30.8**	**23.7**	**51.9**	**94.9**	19.9
test-challenge	CNMT†+DLCT†+AoANet†	150	cider	**75.0**	30.7	23.6	51.6	94.1	**19.9**

Table 2. Ensemble results of different models with two type of scorers, bert indicates BERT-score and cider indicates self-CIDEr.

Summary

- Swin-transformer is used to extract the grid features of the image
- we extract OCR and object detection information as feature supplement
- we integrate the results by self cider, and OCR maximization

Thanks!