Multiple Transformer Mining for VizWiz Image Caption

Xuchao Gong[1], Hongji Zhu[1], Yongliang Wang[1], Biaolong Chen[1], Aixi Zhang[1], Fangxun Shu[1], Si Liu[2]
[1]Alibaba Group
[2]Beihang University
{hongli.gxc, zhj283587, wangyongliang.wyl, biaolong.cbl, aixi.zhax, shufangxun.sfx}@alibaba-inc.com
liusi@buaa.edu.cn

Abstract

This paper proposes a multiple transformer mining algorithm (MTMA) for the VizWiz image captioning task. MTMA consists of grid image feature extraction, OCR and object detectors to effectively describe the image information. Self-Critical Sequence Training (SCST) approach is adopted for image captioning models in the training phase, and semantic similarity aggregation is adopted in the post-processing phase. Meanwhile, ensemble power is leveraged in multi-modal feature fusion and post-caption generation to further enhance the performance. As a result, the proposed algorithm outperforms others with 94.06 CIDEr.

1. Introduction

Now there are many methods based on encoding and decoding in image captioning[8][11][6][3][5]. In order to realize the complementary advantages of the contextual information and fine-grained details, a dual-level collaborative transformer network is proposed in [8]. An attention on attention image captioning method [5] uses attention mechanism to improve image captioning performance. Context-aware auxiliary guidance mechanism is proposed in [11] that can guide the captioning model to perceive global contexts. The global enhanced transformer to enable the extraction of a more comprehensive global representation proposed in [6], and [3] uses a length level embedding to control caption generation.

In this paper, a multiple transformer mining algorithm (MTMA) for the VizWiz caption task is proposed. Our method consists of three parts: (1) multi-modal feature extraction, which contains grid image feature extraction, OCR and object detectors; (2) multi-modal training and the self-critical sequence training (SCST) approach for image captioning models; (3) semantic similarity aggregation post-processing strategy to help boost the CIDEr score. The proposed algorithm achieves 94.06 CIDEr in the VizWiz 2021 image captioning challenge.

2. Method

MTMA method is composed of four parts. Image grid feature extraction is discussed in Section 2.1, OCR and object detectors are detailed in Section 2.2, training models and and post processing of captions ensemble described in 2.3.

2.1. Image Grid Feature Extraction with Swin-Transformer

For image processing, we choose Swin-Transformer [7] as our feature extractor, and as shown in Table 1 we observe a better CIDEr score compared to ResNeXt [9]. We keep the size of the image as long as one of the sides is greater than 384 pixels. Otherwise, we enlarge the size of the image axis to 384 pixels while maintaining the aspect ratio. To extract features, we use the layers before and after the last layernorm respectively [7]. Before being processed by the multi-modal transformer, the image feature vector with 12×12×1536 is reshaped to 144×1536.

2.2. The OCR and Object Detectors

The effectiveness of OCR in image captioning is approved [4], thus several open source OCR solutions have been evaluated, and finally [2] and [1] are leveraged in our method. The open source solution is an effective supplement to our OCR coverage. Like [4] for each image, after the ocr result is obtained, we get OCR embedding features by fastText model.

To get object embedding, we apply pretrained Faster R-CNN [10] as the detector to extract appearance feature. In addition to the OCR component, we also utilize an object detection module to assist in generating accurate image descriptions.

2.3. Model Training and Ensemble

Considering only part of the VizWiz data has OCR and object recognition results, to obtain better captions, we use the strategy of separate training of multiple models to com-

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.

plement each other. For the samples with a large proportion of visual information, the visual model DLCT [8] and AOA [5]is mainly used, and for samples with rich visual and OCR information, the multi-modal model CNMT[12] is mainly used.The experiment is shown in Table 2.

To generate better caption results, we incorporate a series of model ensemble steps. Four strategies are applied: (a) de-tokenization, (b) self-BERT[13] (c) CIDEr ensemble, and (d) OCR maximization.

References

[1] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4715–4723, 2019.

[2] Y. Baek, B. Lee, D. Han, S. Yun, and H. Lee. Character region awareness for text detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9365–9374, 2019.

[3] C. Deng, N. Ding, M. Tan, and Q. Wu. Length-controllable image captioning. *arXiv preprint arXiv:2007.09580*, 2020.

[4] P. Dognin, I. Melnyk, Y. Mroueh, I. Padhi, M. Rigotti, J. Ross, Y. Schiff, R. A. Young, and B. Belgodere. Image captioning as an assistive technology: Lessons learned from vizwiz 2020 challenge. *arXiv preprint arXiv:2012.11696*, 2020.

[5] 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.

[6] J. Ji, Y. Luo, X. Sun, F. Chen, G. Luo, Y. Wu, Y. Gao, and R. Ji. Improving image captioning by leveraging intra- and inter-layer global representation in transformer network. *arXiv preprint arXiv:2012.07061*, 2020.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] Z. Song, X. Zhou, Z. Mao, and J. Tan. Image captioning with context-aware auxiliary guidance. *arXiv preprint arXiv:2012.05545*, 2020.

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

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