Exploiting Knowledge Embedding to Improve the Description for Image Captioning

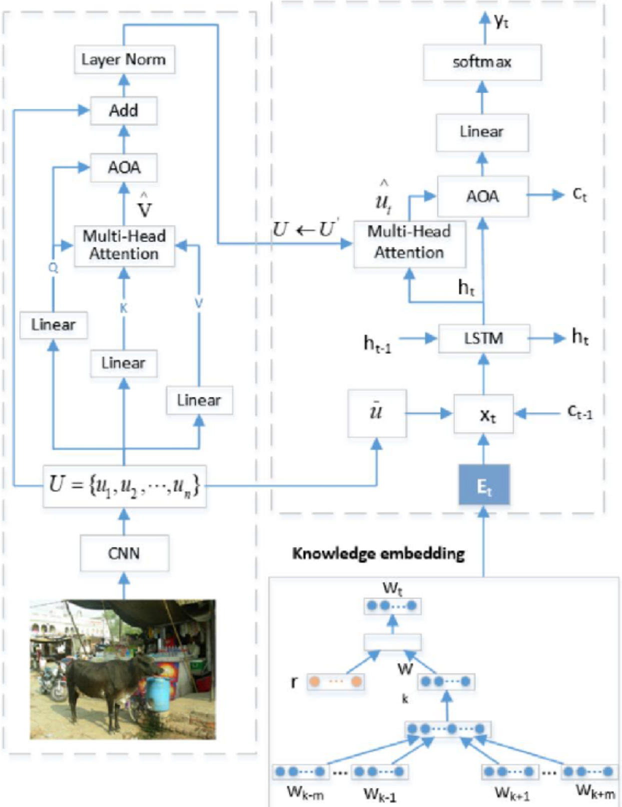
Dandan Song, Cuimei Peng, Huan Yang, and Lejian Liao

Beijing Engineering Research Center of High Volume Language Information Processing & Cloud Computing Applications,
Beijing Key Laboratory of Intelligent Information Technology, School of Computer Science and Technology, Beijing
Institute of Technology, Beijing, China { sdd, pcm, yh,and liaolj,}@ bit.edu.cn

Introduction

In this work, we propose a Knowledge Embedding with Attention on Attention (KE-AoA) method for image captioning, which judges whether or how well the objects are related and augments semantic correlations and constraints between them. The KE-AoA method combines knowledge base method (TransE) and text method (Skip-gram), adding external knowledge graph information (triplets) into the language model to guide the learning of word vectors as the regularization term. Then it employs the AoA module to model the relations among different objects. As more inherent relations and commonsense knowledge are learned, the model can generate better image descriptions.

Network Framework



The KE-AoA model applies the AoA module to the image encoder and combines knowledge embedding with the decoder. By combining semantic features from refining knowledge representations with visual features of the image, our model can better model relations among different objects and generate more accurate captions.

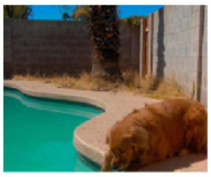
Experimental Results

	Cross-Entropy Loss								CIDEr Optimization							
	B-1	B-2	B-3	B-4	MT	RG	CD	S	B-1	B-2	B-3	B-4	MT	RG	CD	S
DeepVS [26]	62.5	45.0	32.1	23.0	19.5	-	66.0	-	-	-	-	-	-	-	-	-
gLSTM [7]	67.0	49.1	35.8	26.4	22.7	-	81.3	-	-	-	-	-	-	-	-	-
Soft-Attention [11]	70.7	49.2	34.4	24.3	23.9	-	-	-	-	-	-	-	-	-	-	-
Hard-Attention [11]	71.8	50.4	35.7	25.0	23.0	-	-	-	-	-	-	-	-	-	-	-
Adaptive [35]	74.2	58.0	43.9	33.2	26.6	54.9	109.5	-	-	-	-	-	-	-	-	-
LSTM [6]	-	-	-	29.6	25.2	52.6	94.0	-	-	-	-	31.9	25.5	54.3	106.3	-
SCST [36]	-	-	-	30.0	25.9	53.4	99.4	-	-	-	-	34.2	26.7	55.7	114.0	-
Up-Down [12]	77.2	-	-	36.2	27.0	56.4	113.5	20.3	79.8	-	-	36.3	27.7	56.9	120.1	21.4
SGAE [17]	77.6	-	-	36.9	27.7	57.2	116.7	20.9	80.8	-	-	38.4	28.4	58.6	127.8	22.1
Baseline: AoANet	76.8	61.2	47.3	36.4	28.1	57.1	117.3	21.1	80.1	64.7	50.6	38.4	28.5	58.4	126.7	22.3
Ours	**77.9**	**61.8**	**48.6**	**37.7**	**28.6**	**58.0**	**119.9**	**21.6**	**80.9**	**65.5**	**51.2**	**39.2**	**29.4**	**58.9**	**128.9**	**22.6**

B-n is the abbreviation of BLEU-n, M for METEOR, R for ROUGE-l, and C for CIDEr, and S for SPICE.

The experiments on MSCOCO data sets achieve a significant improvement on the existing methods and validate the effectiveness of our prior knowledge-based approach. The KE-AoA model shows significant improvements across all metrics regardless of whether cross-entropy loss or CIDEr optimization is used.

Competitive captions generated by KE-AoA are shown below:



Ours: a brown dog laying on the ground next to a pool

Baseline: a brown dog laying next to a pool

GT1: a dog laying on the ground next to a pool

GT2: a brown dog laying next to a pool in the water

Ours: a man and a woman sitting on a bench with a dog

Baseline: a man and a dog sitting on a bench in the street

GT1: a man and a woman sitting on a bench with a dog

GT2: a man and a woman is sitting on a bench with a dog

Ours: a large airplane is parked on the tarmac at an airport

Baseline: a large airplane sitting on the tarmac at an airport

GT1: a large airplane is parked on the tarmac at an airport

GT2: a large airplane is parked on the runway at an airport

Ours: a bedroom with a white bed and curtains on the wall

Baseline: a bedroom with a bed and a television in the wall

GT1: a bedroom with a bed and curtains in the room

GT2: a bedroom with a bed and a television on the wall

Acknowledgements

This work was supported by National Key Research and Development Program of China (Grant No. 2020AAA0106600) and National Natural Science Foundation of China (Grant Nos. 61976021, 61672100 and U1811262).