# Appendix for Paper 4531

## 1 Statistical Significance Tests

In this experiment, we also perform the statistical significance tests. The experimental results are summarized in Table A.1, which show that our model achieves statistically significantly better results than the compared baseline methods on MSCOCO dataset (t-test, p-value $< 0.05$).

| Methods | **BLEU-1** | **BLEU-2** | **BLEU-3** | **BLEU-4** | **METEOR** | **ROUGE** | **CIDEr** |
|---|---|---|---|---|---|---|---|
| Soft-Attention | 70.7 | 49.2 | 34.4 | 24.3 | 23.9 | - | - |
| Hard-Attention | 71.8 | 50.4 | 35.7 | 25.0 | 23.0 | - | - |
| VAE | 72.0 | 52.0 | 37.0 | 28.0 | 24.0 | - | 90.0 |
| Google NICv2 | - | - | - | 32.1 | 25.7 | - | 99.8 |
| Attributes-CNN | 74.0 | 56.0 | 42.0 | 31.0 | 26.0 | - | 94.0 |
| $\mathrm{CNN}_{\mathcal{L}}$+RNN | 72.3 | 55.3 | 41.3 | 30.6 | 26.0 | - | 94.0 |
| PG-SPIDEr-TAG | 75.4 | 59.1 | 44.5 | 33.2 | 25.7 | 55.0 | 101.3 |
| Adaptive | 74.2 | 58.0 | 43.9 | 33.2 | 26.6 | 54.9 | 108.5 |
| SCST:Att2in | 76.9 | 60.2 | 45.1 | 33.3 | 26.3 | 55.3 | 111.4 |
| SCST:Att2all | 77.4 | 60.9 | 46.0 | 34.1 | 26.7 | 55.7 | 114.0 |
| TopDown | 79.8 | 63.4 | 48.4 | 36.3 | 27.7 | 56.9 | 120.1 |
| StackCap | 78.4 | 62.5 | 47.9 | 36.1 | 27.4 | 56.9 | 120.4 |
| TextAtt+ResNet | 74.9 | 58.1 | 43.7 | 32.6 | 25.7 | - | 102.4 |
| CNN+Att | 71.1 | 53.8 | 39.4 | 28.7 | 24.4 | 52.2 | 91.2 |
| GroupCap | 74.4 | 58.1 | 44.3 | 33.8 | 26.2 | - | - |
| NBT | 75.5 | - | - | 34.7 | 27.1 | - | 107.2 |
| ICKC (ours) | **80.9*** | **64.6*** | **49.5*** | **37.8*** | **28.6*** | **58.1*** | **121.3** |

Table A.1: Comparisons of ICKC and baseline methods on MSCOCO Karpathy test split. Scores with * mean that improvement of our model is statistically significant over the baseline methods (t-test, p-value $< 0.05$).

## 2 Error Analysis

To examine the limitations of the proposed model, we additionally carry out an analysis of the errors made by ICKC model. Specifically, we randomly choose 100 images from test set whose captions generated by our model have low evaluation scores. We reveal several reasons of the low evaluation scores, which can be divided into two primary categories. **First**, ICKC fails to identify the difference between visually similar images, thus generates general captions that are not tailored to the given images. For example, as shown in Table A.2, the proposed ICKC model generates the same caption for two different images. This may be because that we employ guidance captions in both encoding and decoding. One possible solution is to employ the generative adversarial network (GAN) framework in ICKC and use the discriminative model in GAN to distinguish the correct and incorrect image-caption pairs. **Second**, ICKC fails to detect some objects in the images that have no high-quality guidance captions. As shown in Table A.3, ICKC cannot correctly identify the "umbrella" object in image, thus generates object-irrelevant captions. It suggests that certain object detection strategy needs to be devised in the future so as to generate better captions for specific images.

## 3 Ablation Study of External Knowledge in Encoding and Decoding

For the purpose of analyzing the effectiveness of external knowledge in encoding and decoding phases, we report the ablation test of our model by replacing the knowledge embeddings with Glove (Pennington et al., 2014) embeddings in encoding and decoding respectively, denoted as w/o KB in encoder and w/o KB in decoder. The ablation results are demonstrated in Table A.4. We can observe that the commonsense knowledge has larger impact in encoding phase than in decoding phase. The reason may be that the commonsense knowledge help the encoder to learn better image features, which is the basis of the decoder.

| | | |
|---|---|---|
| |  |  |
| ground truth | "a elephant drinks from a stream with several other elephants walking in background" | "a group of elephants bathing and playing in the water' |
| ICKC | "a herd of elephants walking in a watering hole" | "a herd of elephants walking in the water" |

Table A.2: Two images with ground-truth captions and generated captions by ICKC.

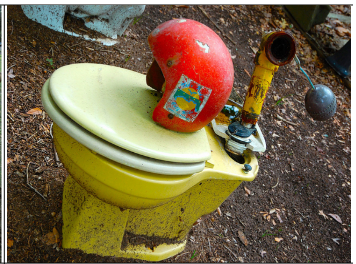
| | | |
|---|---|---|
| |  |  |
| ground truth | "a child holding a flowered umbrella and petting a yak" | "a yellow toilet with a red helmet on top of it" |
| ICKC | "a group of people with a herd of cows" | "a yellow toy sitting on top of a banana" |

Table A.3: Two images with ground-truth captions and generated captions by ICKC.

| | Cross-entropy | | | | | | | CIDEr-optimization | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | B-1 | B-2 | B-3 | B-4 | METEOR | ROUGE | CIDEr | B-1 | B-2 | B-3 | B-4 | METEOR | ROUGE | CIDEr |
| ICKC (ours) | 78.1 | 61.9 | 47.3 | 36.5 | 27.2 | 56.4 | 114.7 | 80.9 | 64.6 | 49.5 | 37.8 | 28.6 | 58.1 | 121.3 |
| w/o Knowledge | 76.2 | 60.5 | 46.4 | 35.3 | 26.7 | 55.9 | 111.4 | 78.7 | 62.5 | 47.9 | 35.6 | 27.5 | 57.3 | 116.5 |
| w/o KB in encoder | 76.9 | 60.9 | 46.8 | 35.7 | 26.8 | 56.1 | 112.9 | 79.6 | 63.4 | 48.7 | 36.5 | 28.1 | 57.8 | 118.5 |
| w/o KB in decoder | 77.3 | 61.2 | 46.9 | 36.1 | 27.1 | 56.3 | 113.6 | 80.2 | 63.8 | 49.0 | 37.2 | 28.3 | 57.9 | 120.3 |

Table A.4: Ablation study of external knowledge in KB on MSCOCO Karpathy test split. Here, B-n is short for BLEU-n.

## 4 Experimental Results on Flickr30k Dataset

We additionally evaluate our model on Flickr30k which is a widely used benchmark in image captioning. In particular, Flickr30k contains 31,000 images, including 29,000 images for training, 1,000 images for validation, and 1,000 images for testing. The experimental results are reported in Table A.5 for Flickr30k. Our model achieves statistically significantly better performance than the state-of-the-art competitors on several evaluation metrics (t-test, p-value $< 0.05$). It is noteworthy that our results can be further improved by tweaking the hyper-parameters.

## References


Long Chen, Hanwang Zhang, Jun Xiao, Liqiang Nie, Jian Shao, Wei Liu, and Tat-Seng Chua. 2017. Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. pages 5659–5667.

Zhe Gan, Chuang Gan, Xiaodong He, Yunchen Pu, Kenneth Tran, Jianfeng Gao, Lawrence Carin, and Li Deng. 2017. Semantic compositional networks for visual captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. pages 5630–5639.

| Methods | BLEU-1 | BLEU-2 | BLEU-3 | BLEU-4 | METEOR | ROUGE | CIDEr |
|---|---|---|---|---|---|---|---|
| Soft-Attention (Xu et al., 2015) | 66.9 | 43.4 | 28.8 | 19.1 | 18.5 | - | - |
| Hard-Attention (Xu et al., 2015) | 66.7 | 43.9 | 29.6 | 19.9 | 18.5 | - | - |
| VAE (Pu et al., 2016) | 72.0 | 53.0 | 38.0 | 25.0 | - | - | - |
| Google NIC (Vinyals et al., 2017) | 63.0 | 41.0 | 27.0 | - | - | - | - |
| ATT-FCN (You et al., 2016) | 64.7 | 46.0 | 32.4 | 23.0 | 18.9 - | - | - |
| Att-CNN+RNN (Wu et al., 2016) | 73.0 | 55.0 | 40.0 | 28.0 | - | - | - |
| SCN-LSTM (Gan et al., 2017) | 73.5 | 53.0 | 37.7 | 25.7 | 21.0 - | - | - |
| SCA-CNN-ResNet (Chen et al., 2017) | 68.2 | 49.6 | 35.9 | 25.8 | 22.4 | 50.9 | 66.5 |
| Adaptive (Lu et al., 2017) | 67.7 | 49.4 | 35.4 | 25.1 | 20.4 | - | 53.1 |
| CNNL+RHN (Gu et al., 2017) | 73.8 | 56.3 | 41.9 | 30.7 | 21.6 | - | 61.8 |
| Self-retrieval-SR-PL (Liu et al., 2018) | 72.9 | 54.5 | 40.1 | 29.3 | 21.8 | 49.9 | 65.0 |
| ICKC (ours) | **74.3** | **57.6*** | **42.5** | **31.3** | **23.4*** | **53.1*** | **67.8*** |

Table A.5: Single-model performance by our proposed method and state-of-the-art methods on Flickr30k dataset. Scores with * mean that improvement of our model is statistically significant over the baseline methods (t-test, p-value $< 0.05$).


Jiuxiang Gu, Gang Wang, Jianfei Cai, and Tsuhan Chen. 2017. An empirical study of language cnn for image captioning. In *Proceedings of the IEEE International Conference on Computer Vision*. pages 1222–1231.

Xihui Liu, Hongsheng Li, Jing Shao, Dapeng Chen, and Xiaogang Wang. 2018. Show, tell and discriminate: Image captioning by self-retrieval with partially labeled data. In *Proceedings of the European Conference on Computer Vision (ECCV)*. pages 338–354.

Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. 2017. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In *The Conference on Computer Vision and Pattern Recognition*.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. pages 1532–1543.

Yunchen Pu, Zhe Gan, Ricardo Henao, Xin Yuan, Chunyuan Li, Andrew Stevens, and Lawrence Carin. 2016. Variational autoencoder for deep learning of images, labels and captions. In *Advances in neural information processing systems*. pages 2352–2360.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2017. Show and tell: Lessons learned from the 2015 mscoco image captioning challenge. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39(4):652–663.

Qi Wu, Chunhua Shen, Lingqiao Liu, Anthony Dick, and Anton Van Den Hengel. 2016. What value do explicit high level concepts have in vision to language problems? In *Proceedings of the IEEE conference on computer vision and pattern recognition*. pages 203–212.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*. pages 2048–2057.

Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. 2016. Image captioning with semantic attention. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. pages 4651–4659.