Learning to Evaluate Image Captioning - Supplementary Material

Yin Cui[1,2] Guandao Yang[1] Andreas Veit[1,2] Xun Huang[1,2] Serge Belongie[1,2]
[1]Department of Computer Science, Cornell University [2]Cornell Tech

1. Summary

In this supplementary material, we present implementation details and more experiment results with different model hyper-parameters. Further, we compare correlation to human score on all 5 metrics used in 2015 COCO Captioning Challenge [1]. In addition, Figure 1 provides examples captions of both success and failure cases.

2. Implementation Details

2.1. Image Representations

To extract image features, we use a 152-layer Residual Network (ResNet-152) [2] pretrained on ImageNet, which achieved state-of-the-art performance on the large-scale image classification task [4]. Instead of the standard feature extraction procedure of extracting features from a resized and cropped 224×224 image, we extract the features from the original image without any resizing and cropping. The feature map from the last convolution layer is average-pooled, resulting in a 2048-dimensional feature vector as our image feature representation. We do not fine-tune the weights of the residual network during training; thus the image features remain fixed.

2.2. Caption Representations

We construct a vocabulary list by taking the 10,000 most frequent words that appear at least 5 times in the human annotated captions from the training set. A special token is added to the vocabulary to represent any word that is not among the top 10,000 words. Suppose the length of the vocabulary list is n. Each word in the vocabulary can be represented by a one-hot vector $\mathbf{w} \in \{0,1\}^n$, where for word i, $w_i = 1$ and for all $j \neq i$, $w_j = 0$. Then, a word embedding matrix $E \in \mathbb{R}^{n \times d}$ is used to encode each word as a d-dimensional vector $\mathbf{x} = \mathbf{w}E \in \mathbb{R}^d$ as the input to the LSTM. The word embedding is initialized from GloVe [3]. We use a word embedding dimension of $d = 300$ for all of our experiments. We fix the step size of the LSTM to be 15. That is, shorter sentences are padded with a special token and longer captions are cut at 15 words. During training, a mask is applied to remove the padded part of a caption when we compute the classification loss.

2.3. Training

During training, we sample equal number of positive and negative examples. To generate positive examples, we first randomly choose an image from the database, and such image should correspond to several reference captions. We use one reference caption as the context, and a different one as the candidate caption. To compose a negative example, we first choose with equal probability one of the following types of negative examples: 1) using a caption generator; 2) sample a caption from a pathologically transformed dataset; or 3) generate a caption using Monte Carlo Sampling. If we are using a pathologically transformed dataset, we will choose in equal probability among three transformations: $\mathcal{T}_{RC}$ (human caption for a different image), $\mathcal{T}_{WP}$ (reference caption with word permutation), and $\mathcal{T}_{RW}$ (reference caption with random word replacement).

2.4. Evaluation

To evaluate how good a candidate caption is, we iterate through all the reference captions for the image and compute a score using each reference caption as context for the candidate caption. The average of these scores is the final score for the candidate caption.

To evaluate a caption generator, we train our model for 10 epochs using only this generator to produce the first type of negative examples. We use pathological transformation and Monte Carlo Sampling for all model evaluation. Finally, we use our model to score all candidate captions this generator produces on a held-out set of data. The average of these score is used as the final indicator for how good the caption generator is.

While computing the caption level correlation with human, we first use a candidate metric to compute a score for each pair of image and candidate caption (i, c), where i indicates the image and c indicates the candidate caption. Suppose a (i, c) pair has corresponding human annotations $\mathcal{A}_{i,c}$ and our computed scores $\mathcal{S}_{i,c}$, we create all pairs between human annotations and computed scores $[(h, s) | h \in \mathcal{A}_{i,c}, s \in \mathcal{S}_{i,c}]$. Finally, we compute the Kendalls τ Rank Correlation for all score pairs we could generate, *i.e.*,

$$\tau([(h, s) | h \in \mathcal{A}_{i,c}, s \in \mathcal{S}_{i,c}, \forall (i, c)]) \qquad (1)$$

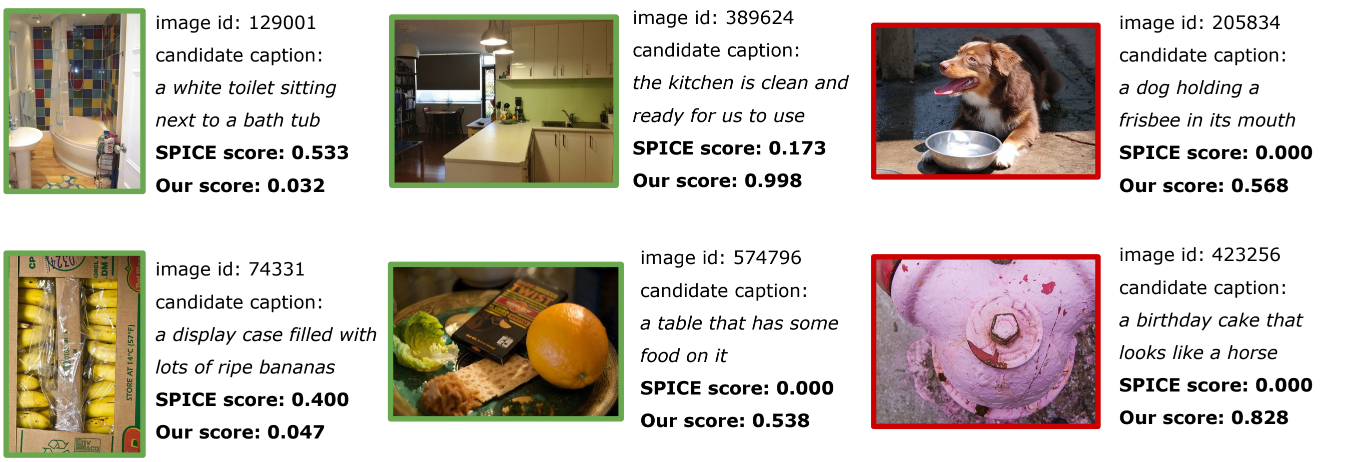
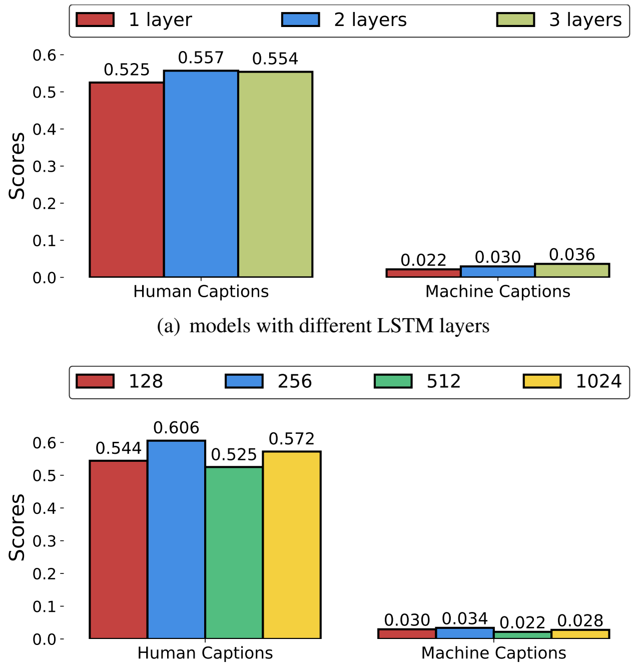


Figure 1. Exemplar candidate captions and their evaluation scores using our metric and SPICE on the COCO validation set. Examples where our metric performs better than SPICE are marked with green bounding boxes, while examples where our metric is worse are marked with red ones. By utilizing the image as context, our metric is able to recognize some captions that are referring to wrong objects (left), and give high scores to captions that are semantically relevant to the image (center). Typical failure cases of our metric are due to misleading visual information (right).

3. The Choice of Hyper-parameters

Fig. 2 compares capability performance of models with different LSTM layers and hidden feature sizes. The proposed model is robust with respect to variant LSTM parameters. Using models with higher capacity, *i.e.*, more layers, higher dimensional hidden features, have no obvious benefit in terms of capability performance. Considering the trade-off between performance gain and efficiency, we therefore use 1 LSTM layer and make the hidden feature of the LSTM to be 512 dimensional in our paper.

Fig. 3 shows models trained without data augmentation. Models trained with or without data augmentation are capable of learning to give higher scores to human captions than machine generated captions. Interestingly, a critique trained without data augmentation can achieve even higher discrimination performance than models with data augmentation. However, as shown in Sec. 4.3 and Fig. 6 in the paper, models trained without data augmentation are actually learning to perform a much simpler task, focusing only on discriminating human generated captions from the machine generated ones without considering the context (*i.e.*, image and ground truth captions). Therefore, models that merely perform well in discrimination task might be easily gamed with pathological transformations. Training with appropriate data augmentation and architecture (non-linearity) is essential to force critiques to pay attention to contexts.



(a) models with different LSTM layers

(b) models with different LSTM hidden feature size

Figure 2. Top: models with variant LSTM layers (512 hidden size). Bottom: models with variant LSTM hidden feature size (1 layer). All the models are trained with both image and reference ground truth captions as contexts, using concatenation of context information and candidate caption followed by a linear classifier and with data augmentation.

4. System Level Human Correlation on COCO

In the original paper, we didn't compare to metrics M3, M4 and M5 because they were not used to rank image captioning models, but were intended for an ablation study to understand which aspects make captions good [1]. Since

	M1		**M2**		**M3**		**M4**		**M5**	
	ρ	p-value	ρ	p-value	ρ	p-value	ρ	p-value	ρ	p-value
BLEU-1	0.124	(0.687)	0.135	(0.660)	0.549	(0.052)	-0.517	(0.070)	0.241	(0.428)
BLEU-2	0.037	(0.903)	0.048	(0.877)	0.483	(0.094)	-0.572	(0.041)	0.162	(0.598)
BLEU-3	0.004	(0.990)	0.016	(0.959)	0.471	(0.105)	-0.588	(0.035)	0.143	(0.641)
BLEU-4	-0.019	(0.951)	-0.005	(0.987)	0.459	(0.114)	-0.577	(0.039)	0.139	(0.650)
METEOR	0.606	(0.028)	0.594	(0.032)	0.808	(0.001)	0.085	(0.784)	0.685	(0.010)
ROUGE-L	0.090	(0.769)	0.096	(0.754)	0.529	(0.063)	-0.526	(0.065)	0.208	(0.494)
CIDEr	0.438	(0.134)	0.440	(0.133)	0.763	(0.002)	-0.149	(0.628)	0.559	(0.047)
SPICE	0.759	(0.003)	0.750	(0.003)	**0.871**	(0.000)	0.250	(0.411)	0.809	(0.001)
Ours (no DA)	**0.821**	(0.000)	**0.807**	(0.000)	0.430	(0.143)	**0.844**	(0.000)	0.704	(0.007)
Ours	**0.939**	(0.000)	**0.949**	(0.000)	0.720	(0.006)	**0.626**	(0.026)	**0.867**	(0.000)

M1: Percentage of captions that are evaluated as better or equal to human caption.

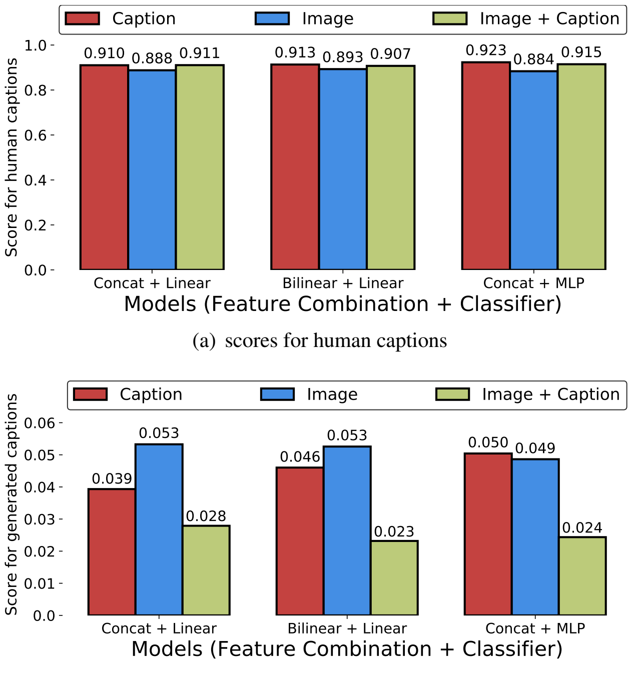
M2: Percentage of captions that pass the Turing Test.

M3 (Correctness): Average correctness of the captions on a scale 1-5 (incorrect - correct).

M4 (Detailness): Average amount of detail of the captions on a scale 1-5 (lack of details - very detailed).

M5 (Salience): Percentage of captions that are similar to human description.

Table 1. Pearson's ρ correlation between human judgements and evaluation metrics. We use the 12 available entries to the 2015 MS-COCO captioning challenge that submitted results on the validation set. "Ours (no DA)" means our metric trained without data augmentation.



(a) scores for human captions

(b) scores for generated captions by **ST**, **SAT** and **NT**

Figure 3. This figure is same as Fig. 5 in the paper except all models are trained without data augmentation.

our metric was designed to evaluate the overall quality of an image caption, we only compared M1 and M2. For better understanding of our metric from different perspectives, in Table 1, we calculate the Pearson's ρ correlation between human judgements on all 5 metrics (M1-M5) used in 2015 COCO Captioning Challenge [1].

From the results, we can see that the human correlation of our proposed evaluation metrics surpasses all other metrics by large margins on M1, M2, M4 and M5. On M3, our metric achieves comparable correlation scores with other commonly-used metrics. It is worth noticing that all other metrics fail to capture the human correlation on the detailness of captions (M4), whereas our metric correlates reasonably well with humans on M4.

References

[1] The coco 2015 captioning challenge. `http://mscoco.org/dataset/#captions-challenge2015`. 1, 2, 3

[2] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 1

[3] J. Pennington, R. Socher, and C. D. Manning. Glove: Global vectors for word representation. In *EMNLP*, 2014. 1

[4] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. Imagenet large scale visual recognition challenge. *IJCV*, 2015. 1