A Experimental Details

A.1 Reproducibility Checklist

Source Code We provide the source code for both training UMIC and computing UMIC as supplementary material. We will publicly release the full source with the pre-trained model to easily compute UMIC.

Computing Infrastructure We use AMD Ryzen Threadripper 2950X (3.50 GHz) with GeForce GTX 2080 Ti for the experiments. The software environments are Python 3.6.8 and PyTorch 1.1.0.

Average runtime for each approach Each epoch of our training UMIC on average takes 20 minutes using a single GPU. For evaluation, it takes a minute.

Number of Model Parameters The number of parameters in UMIC is about 109.9M.

A.2 Correlation Coefficient

We compute Kendall-C for Flickr8k (Hodosh et al., 2013), since we could produce the similar results for most of the previous papers. And we compute Kendall-B for Composite (Aditya et al., 2015) and CapEval1k. For Composite, we use five references and some of the candidate captions are exact same with one of the references.

A.3 Significance Test

For all of the correlation coefficients we computed in this paper, we conduct a standard way to test the significance of the correlation coefficient. We use a t-test using a null hypothesis that is an absence of association to report the p-value for each coefficient.

B Data Collection

B.1 Generating Captions

We generate the captions from the images in Karphathy's test split that do not have any overlaps in the training set and validation set of UMIC. We use four models, Att2in (Rennie et al., 2017), Transformer (Vaswani et al., 2017), BUTD (Anderson et al., 2018), and AoANet (Huang et al., 2019) to generate captions. We use the pre-trained model that uses self-critical loss (Luo et al., 2018) in the public repository [1]. We set beam size 2 for all of the models during the inference. We sample 1,000 captions for a total of 250 images for each model, where each caption does not have a single equivalent as shown in Figure 1.

[1] https://github.com/ruotianluo/self-critical.pytorch

B.2 Instructions to Annotators

The interface and instructions to annotators in MTurk are shown in Figure 1 and Figure 2. We request the worker to evaluate four captions at once in a single assignment so that the worker can consider the difference among the captions.

B.3 Inter-annotator Agreement

We compute the annotator agreement using Krippendorff's α (Krippendorff, 1970). We observe that Krippendorff's α is 0.37 that indicates a "fair" agreement according to one of the general guidelines (Landis and Koch, 1977) for kappa-like measures.

B.4 Worker Pool & Pay

We hire the annotators whose locations in one of the US, UK, CA, NZ, AU. We restrict the workers whose HIT approval rates are higher than 96%, and minimum hits are over 5000. We pay workers more than USD $10 in an hour through several preliminary experiments on the compensation.

References

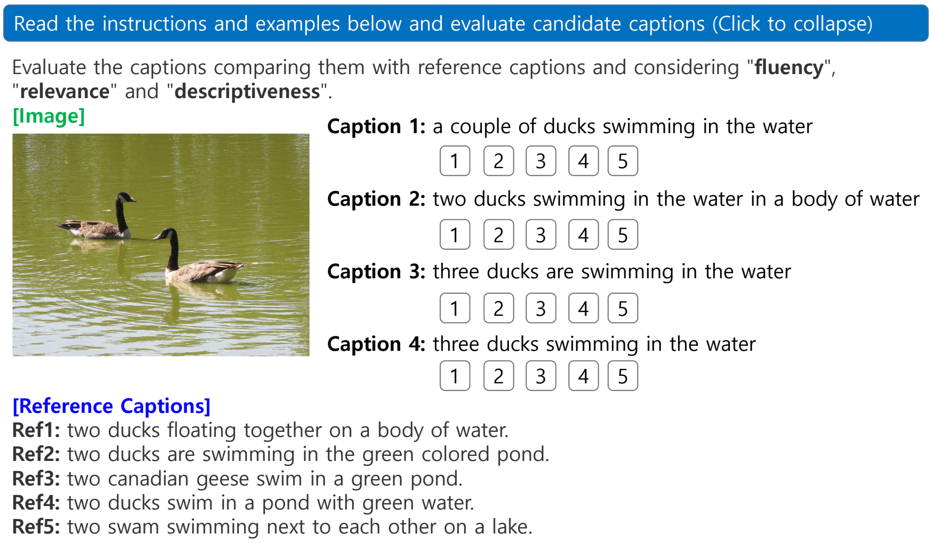
Somak Aditya, Yezhou Yang, Chitta Baral, Cornelia Fermuller, and Yiannis Aloimonos. 2015. From images to sentences through scene description graphs using commonsense reasoning and knowledge. *arXiv preprint arXiv:1511.03292*.

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6077–6086.

Micah Hodosh, Peter Young, and Julia Hockenmaier. 2013. Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research*, 47:853–899.

Lun Huang, Wenmin Wang, Jie Chen, and Xiao-Yong Wei. 2019. Attention on attention for image captioning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4634–4643.

Klaus Krippendorff. 1970. Estimating the reliability, systematic error and random error of interval data. *Educational and Psychological Measurement*, 30(1):61–70.



Read the instructions and examples below and evaluate candidate captions (Click to collapse)

Evaluate the captions comparing them with reference captions and considering "**fluency**", "**relevance**" and "**descriptiveness**".

[Image]

Caption 1: a couple of ducks swimming in the water

1 2 3 4 5

Caption 2: two ducks swimming in the water in a body of water

1 2 3 4 5

Caption 3: three ducks are swimming in the water

1 2 3 4 5

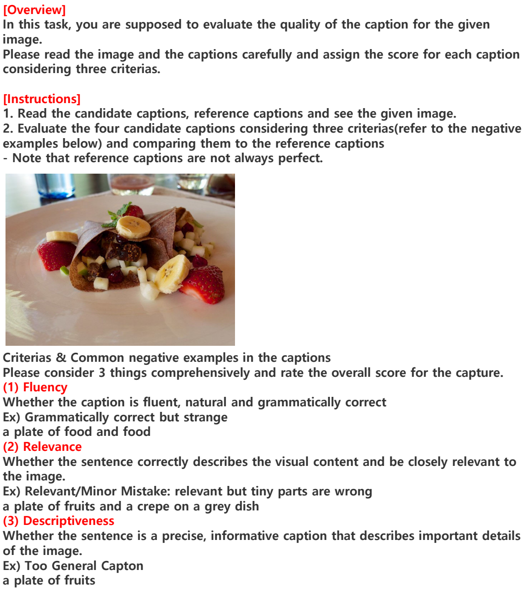
Caption 4: three ducks swimming in the water

1 2 3 4 5

[Reference Captions]

Ref1: two ducks floating together on a body of water.
Ref2: two ducks are swimming in the green colored pond.
Ref3: two canadian geese swim in a green pond.
Ref4: two ducks swim in a pond with green water.
Ref5: two swam swimming next to each other on a lake.

Figure 1: Annotation interface and short instructions for captioning evaluation task.



[Overview]
In this task, you are supposed to evaluate the quality of the caption for the given image.
Please read the image and the captions carefully and assign the score for each caption considering three criterias.

[Instructions]
1. Read the candidate captions, reference captions and see the given image.
2. Evaluate the four candidate captions considering three criterias(refer to the negative examples below) and comparing them to the reference captions
- Note that reference captions are not always perfect.

Criterias & Common negative examples in the captions
Please consider 3 things comprehensively and rate the overall score for the capture.
(1) Fluency
Whether the caption is fluent, natural and grammatically correct
Ex) Grammatically correct but strange
a plate of food and food
(2) Relevance
Whether the sentence correctly describes the visual content and be closely relevant to the image.
Ex) Relevant/Minor Mistake: relevant but tiny parts are wrong
a plate of fruits and a crepe on a grey dish
(3) Descriptiveness
Whether the sentence is a precise, informative caption that describes important details of the image.
Ex) Too General Capton
a plate of fruits

Figure 2: Full instructions for the captioning evaluation task. We provide an image and five reference captions to the workers and request them to evaluate four captions.

J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174.

Ruotian Luo, Brian Price, Scott Cohen, and Gregory Shakhnarovich. 2018. Discriminability objective for training descriptive captions. *arXiv preprint arXiv:1803.04376*.

Steven J Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. 2017. Self-critical sequence training for image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7008–7024.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.