CIDEr: Consensus-based Image Description Evaluation

Ramakrishna Vedantam[1], C. Lawrence Zitnick[2] Devi Parikh[1]
[1]Virginia Tech [2]Microsoft Research

Automatically describing an image with a sentence is a long-standing challenge at the intersection of computer vision and natural language processing. Recently, there has been much progress in the algorithms for image captioning, based on Convolutional Neural Networks and Recurrent Neural Networks. However, the choice of appropriate evaluation protocols has been unclear.

Previous works have evaluated captioning via. measurement of properties such as grammaticality, saliency, truthfulness etc. Using human studies, these properties may be measured, e.g on separate one to five or pairwise scales. Unfortunately, combining these various results into a single measure of sentence quality is difficult. Alternatively, other works ask subjects to judge the overall quality of a sentence. However, what is liked by humans does not often correspond to what is "human-like".

In this paper, we propose an evaluation protocol for image captioning based on measuring "human-likeness". That is, does an automatically generated sentence sound like a sentence that was written by a human? We propose a novel consensus based protocol, which measures the similarity of a candidate or test sentence to the majority, or consensus of how most people describe the image. Our protocol consists of a new annotation modality for human judgment of consensus, a new automated metric named CIDEr (Consensus-based Image Description Evaluation) that measures consensus and two new datasets named PASCAL-50S and ABSTRACT-50S. Our new datasets contain 50 descriptions per image which enables us to reliably estimate consensus. Our results demonstrate that the proposed CIDEr metric matches human consensus better than existing metrics used for evaluating image captions such as BLEU [3], ROUGE$_L$ [4], and METEOR [1].

CIDEr is available on the MS COCO caption evaluation server [2]. We now describe our contributions in detail.

Human Annotation Modality Our proposed annotation modality for measuring human judgment of consensus uses triplets. That is given three sentences (A,B,C), the question "Which of the two sentences B or C is more similar to sentence A?" is asked to subjects. Sentences B and C are two candidate sentences, while sentence A is a reference sentence provided by humans for the image. For each choice of B and C, we form such triplets using all available reference sentences for the image. Ultimately, we record which sentence B or C was picked as being more similar to a majority of reference sentences.

CIDEr Metric Intuitively, a measure for consensus should encode how often n-grams in the candidate sentence are present in the reference sentences. Similarly, n-grams not present in the reference sentences should not be in the candidate sentence. Finally, n-grams that commonly occur across all images in the dataset should be given lower weight, since they are likely to be less informative. To encode this intuition, we perform a Term Frequency Inverse Document Frequency (TF-IDF) weighting for each n-gram, and collect n-grams into a vector representing the sentence.

Given such TF-IDF vectors for each candidate and reference description, our CIDEr$_n$ score for n-grams of length n is computed using the average cosine similarity between the TF-IDF vectors for candidate and reference sentences, which accounts for both precision and recall. Our CIDEr$_n$ score between a candidate sentence c_i and a set of reference image descriptions $S_i = \{s_{i1}, \ldots, s_{im}\}$ is given as follows:

$$\text{CIDEr}_n(c_i, S_i) = \frac{1}{m} \sum_j \frac{\boldsymbol{g}^n(c_i) \cdot \boldsymbol{g}^n(s_{ij})}{\|\boldsymbol{g}^n(c_i)\| \|\boldsymbol{g}^n(s_{ij})\|}, \quad (1)$$

where $\boldsymbol{g}^n(c_i)$ is the TF-IDF vector for a candidate description and $\boldsymbol{g}^n(s_{ij})$ is a TF-IDF vector for reference sentence j of image i.

This is an extended abstract. The full paper is available at the Computer Vision Foundation webpage.

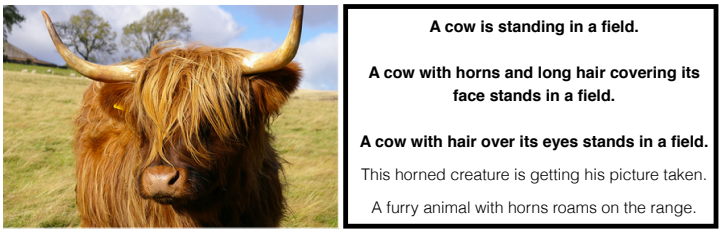

Figure 1: Captions representative of the consensus are shown in bold, for an image from our PASCAL-50S dataset. We propose to capture such captions with our evaluation protocol.

Usage of longer n-grams helps capture grammatical properties as well as richer semantics. Scores from different n-grams (upto 4) are averaged to give the final CIDEr metric:

$$\text{CIDEr}(c, S_i) = \frac{1}{4} \sum_{n=1}^{4} \text{CIDEr}_n(c_i, S_i), \quad (2)$$

Datasets Our proposed datasets PASCAL-50S and ABSTRACT-50S have 1000 and 500 images respectively with 50 descriptions per image. Previous datasets only had a maximum of five. We demonstrate that using more reference sentences it is possible to attain better match to human judgment.

Results We evaluate the performance of automatic metrics at capturing human judgment of consensus using accuracy. Accuracy is defined as the proportion of times a metric and humans agree on which sentence out of B or C matches consensus better.

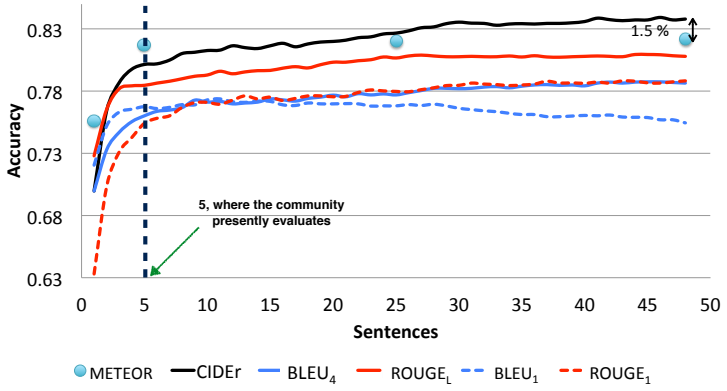

Figure 2: Accuracy of metrics on PASCAL-50S. Note that our proposed metric CIDEr performs the best at 48 descriptions followed by METEOR. We see that the performance of most metrics improves with more reference sentences.

[1] Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. pages 65–72, 2005.

[2] X. Chen, H. Fang, T.-Y. Lin, R. Vedantam, S. Gupta, P. Dollar, and C. L. Zitnick. Microsoft COCO Captions: Data Collection and Evaluation Server. *ArXiv e-prints*, April 2015.

[3] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, 2002.

[4] Chin yew Lin. Rouge: a package for automatic evaluation of summaries. pages 25–26, 2004.