CIDEr: Consensus-based Image Description Evaluation

Ramakrishna Vedantam
Virginia Tech
vrama91@vt.edu

C. Lawrence Zitnick
Microsoft Research
larryz@microsoft.com

Devi Parikh
Virgnia Tech
parikh@vt.edu

Abstract

Automatically describing an image with a sentence is a long-standing challenge in computer vision and natural language processing. Due to recent progress in object detection, attribute classification, action recognition, etc., *there is renewed interest in this area. However, evaluating the quality of descriptions has proven to be challenging. We propose a novel paradigm for evaluating image descriptions that uses* human consensus. *This paradigm consists of three main parts: a new triplet-based method of collecting human annotations to measure consensus, a new automated metric that captures consensus, and two new datasets: PASCAL-50S and ABSTRACT-50S that contain 50 sentences describing each image. Our simple metric captures human judgment of consensus better than existing metrics across sentences generated by various sources. We also evaluate five state-of-the-art image description approaches using this new protocol and provide a benchmark for future comparisons. A version of CIDEr named CIDEr-D is available as a part of MS COCO evaluation server to enable systematic evaluation and benchmarking.*

1. Introduction

Recent advances in object recognition [15], attribute classification [23], action classification [26, 9] and crowdsourcing [40] have increased the interest in solving higher level scene understanding problems. One such problem is generating human-like descriptions of an image. In spite of the growing interest in this area, the evaluation of novel sentences generated by automatic approaches remains challenging. Evaluation is critical for measuring progress and spurring improvements in the state of the art. This has already been shown in various problems in computer vision, such as detection [13, 7], segmentation [13, 28], and stereo [39].

Existing evaluation metrics for image description attempt to measure several desirable properties. These include grammaticality, saliency (covering main aspects), correctness/truthfulness, *etc*. Using human studies, these properties may be measured, *e.g*. on separate *one* to *five* [29, 37, 44, 11] or *pairwise* scales [45]. Unfortunately, combining these various results into one measure of sentence quality is difficult. Alternatively, other works [22, 18] ask subjects to judge the overall quality of a sentence.

An important yet non-obvious property exists when image descriptions are judged by humans: What humans like often does not correspond to what is human-like.[1] We introduce a novel consensus-based evaluation protocol, which measures the similarity of a sentence to the majority, or *consensus* of how most people describe the image (Fig. 1). One realization of this evaluation protocol uses human subjects to judge sentence similarity between a candidate sentence and human-provided ground truth sentences. The question "Which of two sentences is more similar to this other sentence?" is posed to the subjects. The resulting quality score is based on how often a sentence is labeled as being *more* similar to a human-generated sentence. The relative nature of the question helps make the task objective. We encourage the reader to review how a similar protocol has been used in [41] to capture human perception of image similarity. These annotation protocols for similarity may be understood as instantiations of 2AFC (two alternative forced choice) [3], a popular modality in psychophysics.

Since human studies are expensive, hard to reproduce, and slow to evaluate, automatic evaluation measures are commonly desired. To be useful in practice, automated metrics should agree well with human judgment. Some popular metrics used for image description evaluation are BLEU [33] (precision-based) from the machine translation community and ROUGE [46] (recall-based) from the summarization community. Unfortunately, these metrics have been shown to correlate weakly with human judgment [22, 11, 4, 18]. For the task of judging the overall quality of a description, the METEOR [11] metric has shown better correlation with human subjects. Other metrics rely on the ranking of captions [18] and cannot evaluate novel

[1] This is a subtle but important distinction. We show qualitative examples of this in [42]. That is, the sentence that is most similar to a typical human generated description is often not judged to be the "best" description. In this paper, we propose to directly measure the "human-likeness" of automatically generated sentences.

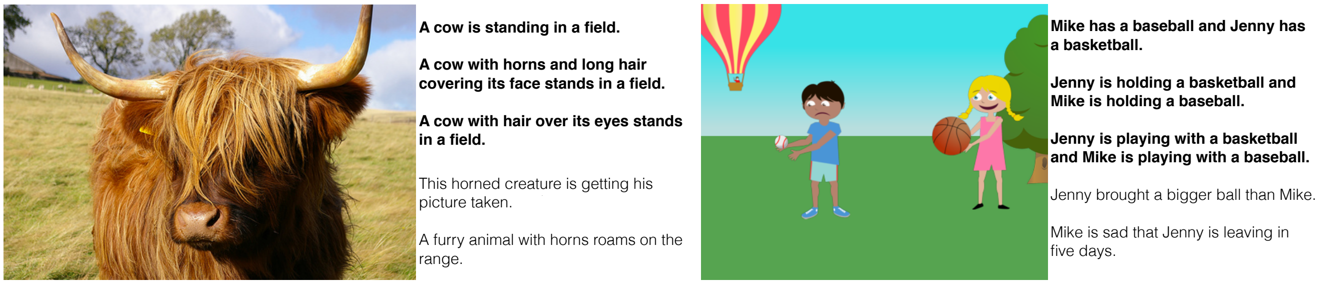

A cow is standing in a field.
A cow with horns and long hair covering its face stands in a field.
A cow with hair over its eyes stands in a field.
This horned creature is getting his picture taken.
A furry animal with horns roams on the range.
Mike has a baseball and Jenny has a basketball.
Jenny is holding a basketball and Mike is holding a baseball.
Jenny is playing with a basketball and Mike is playing with a baseball.
Jenny brought a bigger ball than Mike.
Mike is sad that Jenny is leaving in five days.

Figure 1: Images from our PASCAL-50S (left) and ABSTRACT-50S (right) datasets with a subset of corresponding (human) sentences. Sentences shown in **bold** are representative of the consensus descriptions for these images. We propose to capture such descriptions with our evaluation protocol.

image descriptions.

We propose a new automatic *consensus* metric of image description quality – CIDEr (Consensus-based Image Description Evaluation). Our metric measures the similarity of a generated sentence against a set of ground truth sentences written by humans. Our metric shows high agreement with consensus as assessed by humans. Using sentence similarity, the notions of grammaticality, saliency, importance and accuracy (precision and recall) are inherently captured by our metric.

Existing datasets popularly used to evaluate image description approaches have a maximum of only five descriptions per image [35, 18, 32]. However, we find that five sentences are not sufficient for measuring how a "majority" of humans would describe an image. Thus, to accurately measure consensus, we collect two new evaluation datasets containing 50 descriptions per image – PASCAL-50S and ABSTRACT-50S. The PASCAL-50S dataset is based on the popular UIUC Pascal Sentence Dataset, which has 5 descriptions per image. This dataset has been used for both training and testing in numerous works [29, 22, 14, 37]. The ABSTRACT-50S dataset is based on the dataset of Zitnick and Parikh [47]. While previous methods have only evaluated using 5 sentences, we explore the use of 1 to ~50 reference sentences. Interestingly, we find that most metrics improve in performance with more sentences.[2] Inspired by this finding, the MS COCO testing dataset now contains 5K images with 40 reference sentences to boost the accuracy of automatic measures [5].

Contributions: In this work, we propose a consensus-based evaluation protocol for image descriptions. We introduce a new annotation modality for human judgment, a new automated metric, and two new datasets. We compare the performance of five state-of-the-art machine generation approaches [29, 22, 14, 37]. Our code and datasets are available on the author's webpages. Finally, to facilitate the adoption of this protocol, we have made CIDEr available as a metric on the newly released MS COCO caption evaluation server [5].

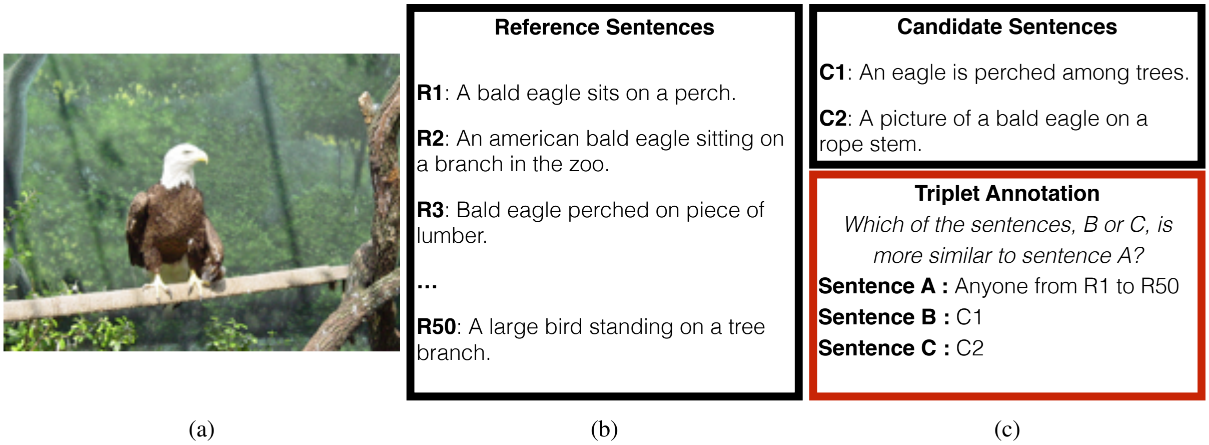
2. Related Work

Vision and Language: Numerous papers have studied the relationship between language constructs and image content. Berg *et al.* [2] characterize the relative importance of objects (nouns). Zitnick and Parikh [47] study relationships between visual and textual features by creating a synthetic Abstract Scenes Dataset. Other works have modeled prepositional relationships [16], attributes (adjectives) [23, 34], and visual phrases (*i.e.* visual elements that co-occur) [38]. Recent works have utilized techniques in deep learning to learn joint embeddings of text and image fragments [20].

Image Description Generation: Various methods have been explored for generating full descriptions for images. Broadly, the techniques are either retrieval- [14, 32, 18] or generation-based [29, 22, 45, 37]. While some retrieval-based approaches use global retrieval [14], others retrieve text phrases and stitch them together in an approach inspired by extractive summarization [32]. Recently, generative approaches based on combination of Convolutional and Recurrent Neural Networks [19, 6, 10, 43, 27, 21] have created a lot of excitement. Other generative approaches have explored creating sentences by inference over image detections and text-based priors [22] or exploiting word co-occurrences using syntactic trees [29]. Rohrbach *et al.* [37] propose a machine translation approach that goes from an intermediate semantic representation to sentences. Some other approaches include [17, 24, 44, 45]. Most of the approaches use the UIUC Pascal Sentence [14, 22, 29, 37, 17] and the MS COCO datasets [19, 6, 10, 43, 27, 21] for evaluation. In this work we focus on the problem of evaluating image captioning approaches.

Automated Evaluation: Automated evaluation metrics have been used in many domains within Artificial Intelligence (AI), such as statistical machine translation and text summarization. Some of the popular metrics in machine translation include those based on precision, such as BLEU [33] and those based on precision as well as recall, such as METEOR [1]. While BLEU (BiLingual Evaluation Understudy) has been the most popular metric, its effective-

[2] Except BLEU computed on unigrams


Reference Sentences

R1: A bald eagle sits on a perch.

R2: An american bald eagle sitting on a branch in the zoo.

R3: Bald eagle perched on piece of lumber.

...

R50: A large bird standing on a tree branch.

Candidate Sentences

C1: An eagle is perched among trees.

C2: A picture of a bald eagle on a rope stem.

Triplet Annotation

Which of the sentences, B or C, is more similar to sentence A?

Sentence A : Anyone from R1 to R50

Sentence B : C1

Sentence C : C2

Figure 2: Illustration of our triplet annotation modality. Given an image (a), with reference sentences (b) and a pair of candidate sentences (c, top), we match them with a reference sentence one by one to form triplets (c, bottom). Subjects are shown these 50 triplets on Amazon Mechanical Turk and asked to pick which sentence (B or C) is more similar to sentence A.

ness has been repeatedly questioned [22, 11, 4, 18]. A popular metric in the summarization community is ROUGE [46] (Recall Oriented Understudy of Gisting Evaluation). This metric is primarily recall-based and thus has a tendency to reward long sentences with high recall. These metrics have been shown to have weak to moderate correlation with human judgment [11]. Recently, METEOR has been used for image description evaluation with more promising results [12]. Another metric proposed by Hodosh *et al.* [18] can only evaluate ranking-based approaches, it cannot evaluate novel sentences. We propose a consensus-based metric that rewards a sentence for being similar to the majority of human written descriptions. Interestingly, similar ideas have been used previously to evaluate text summarization [31].

Datasets: Numerous datasets have been proposed for studying the problem of generating image descriptions. The most popular dataset is the UIUC Pascal Sentence Dataset [35]. This dataset contains 5 human written descriptions for 1,000 images. This dataset has been used by a number of approaches for training and testing. The SBU captioned photo dataset [32] contains one description per image for a million images, mined from the web. These are commonly used for training image description approaches. Approaches are then tested on a query set of 500 images with one sentence each. The Abstract Scenes dataset [47] contains cartoon-like images with two descriptions. The recently released MS COCO dataset [25] contains five sentences for a collection of over 100K images. This dataset is gaining traction with recent image description approaches [19, 6, 10, 43, 27, 21]. Other datasets of images and associated descriptions include ImageClef [30] and Flickr8K [18]. In this work, we introduce two new datasets. First is the PASCAL-50S dataset where we collected 50 sentences per image for the 1,000 images from UIUC Pascal Sentence dataset. The second is the ABSTRACT-50S dataset where we collected 50 sentences for a subset of 500 images from the Abstract Scenes dataset. We demonstrate that more sentences per image are essential for reliable automatic evaluation.

The rest of this paper is organized as follows. We first give details of our triplet human annotation modality (Sec. 3). Then we provide the details of our consensus-based automated metric, CIDEr (Sec. 4). In Sec. 5 we provide the details of our two new image-sentence datasets, PASCAL-50S and ABSTRACT-50S. Our contributions of triplet annotation, metric and dataset make consensus-based image description evaluation feasible. Our results (Sec. 7) demonstrate that our automated metric and our proposed datasets capture consensus better than existing choices.

All our human studies are performed on the Amazon Mechanical Turk (AMT). Subjects are restricted to the United States, and other qualification criteria are imposed based on worker history.[3]

3. Consensus Interface

Given an image and a collection of human generated *reference* sentences describing it, the goal of our consensus-based protocol is to measure the similarity of a *candidate* sentence to a majority of how most people describe the image (*i.e.* the *reference* sentences). In this section, we describe our human study protocol for generating ground truth consensus scores. In Sec. 7, these ground truth scores are used to evaluate several automatic metrics including our proposed CIDEr metric.

An illustration of our human study interface is shown in Fig. 2. Subjects are shown three sentences: A, B and C. They are asked to pick which of two sentences (B or C)

[3] Approval rate greater than 95%, minimum 500 HITs approved

is most similar to sentence A. Sentences B and C are two candidate sentences, while sentence A is a reference sentence. For each choice of B and C, we form triplets using all the reference sentences for an image. We provide no explicit concept of "similarity". Interestingly, even though we do not say that the sentences are image descriptions, some workers commented that they were imagining the scene to make the choice. The relative nature of the task – "Which of the two sentences, B or C, is more similar to A?" – helps make the assessment more objective. That is, it is easier to judge if one sentence is more similar than another to a sentence, than to provide an absolute rating from 1 to 5 of the similarity between two sentences [3].

We collect three human judgments for each triplet. For every triplet, we take the majority vote of the three judgments. For each pair of candidate sentences (B, C), we assign B the winner if it is chosen as more similar by a majority of triplets, and similarly for C. These pairwise relative rankings are used to evaluate the performance of the automated metrics. That is, when automatic metrics give both sentences B and C a score, we check whether B received a higher score or C. Accuracy is computed as the proportion of candidate pairs on which humans and the automatic metric agree on which of the two sentences is the winner.

4. CIDEr Metric

Our goal is to automatically evaluate for image I_i how well a candidate sentence c_i matches the consensus of a set of image descriptions $S_i = \{s_{i1}, \ldots, s_{im}\}$. All words in the sentences (both candidate and references) are first mapped to their stem or root forms. That is, "fishes", "fishing" and "fished" all get reduced to "fish." We represent each sentence using the set of n-grams present in it. An n-gram ω_k is a set of one or more ordered words. In this paper we use n-grams containing one to four words.

Intuitively, a measure of consensus would 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 [36]. The number of times an n-gram ω_k occurs in a reference sentence s_{ij} is denoted by $h_k(s_{ij})$ or $h_k(c_i)$ for the candidate sentence c_i. We compute the TF-IDF weighting $g_k(s_{ij})$ for each n-gram ω_k using:

$$g_k(s_{ij}) = \frac{h_k(s_{ij})}{\sum_{\omega_l \in \Omega} h_l(s_{ij})} \log\left(\frac{|I|}{\sum_{I_p \in I} \min(1, \sum_q h_k(s_{pq}))}\right), \quad (1)$$

where Ω is the vocabulary of all n-grams and I is the set of all images in the dataset. The first term measures the TF of each n-gram ω_k, and the second term measures the rarity of ω_k using its IDF. Intuitively, TF places higher weight on n-grams that frequently occur in the reference sentence describing an image, while IDF reduces the weight of n-grams that commonly occur across all images in the dataset. That is, the IDF provides a measure of word saliency by discounting popular words that are likely to be less visually informative. The IDF is computed using the logarithm of the number of images in the dataset $|I|$ divided by the number of images for which ω_k occurs in any of its reference sentences.

Our CIDEr$_n$ score for n-grams of length n is computed using the average cosine similarity between the candidate sentence and the reference sentences, which accounts for both precision and recall:

$$\text{CIDEr}_n(c_i, S_i) = \frac{1}{m} \sum_j \frac{\boldsymbol{g^n}(c_i) \cdot \boldsymbol{g^n}(s_{ij})}{\|\boldsymbol{g^n}(c_i)\| \|\boldsymbol{g^n}(s_{ij})\|}, \quad (2)$$

where $\boldsymbol{g^n}(c_i)$ is a vector formed by $g_k(c_i)$ corresponding to all n-grams of length n and $\|\boldsymbol{g^n}(c_i)\|$ is the magnitude of the vector $\boldsymbol{g^n}(c_i)$. Similarly for $\boldsymbol{g^n}(s_{ij})$.

We use higher order (longer) n-grams to capture grammatical properties as well as richer semantics. We combine the scores from n-grams of varying lengths as follows:

$$\text{CIDEr}(c_i, S_i) = \sum_{n=1}^{N} w_n \text{CIDEr}_n(c_i, S_i), \quad (3)$$

Empirically, we found that uniform weights $w_n = 1/N$ work the best. We use N = 4.

5. New Datasets

We propose two new datasets – PASCAL-50S and ABSTRACT-50S – for evaluating image caption generation methods. Both the datasets have 50 reference sentences per image for 1,000 and 500 images respectively. These are intended as "testing" datasets, crafted to enable consensus-based evaluation. For a list of training datasets, we encourage the reader to explore [25, 32]. The PASCAL-50S dataset uses all 1,000 images from the UIUC Pascal Sentence Dataset [35] whereas the ABSTRACT-50S dataset uses 500 random images from the Abstract Scenes Dataset [47]. The Abstract Scenes Dataset contains scenes made from clipart objects. Our two new datasets are different from each other both visually and in the type of image descriptions produced.

Our goal was to collect image descriptions that are objective and representative of the image content. Subjects were shown an image and a text box, and were asked to "Describe what is going on in the image". We asked subjects to capture the main aspects of the scene and provide descriptions that others are also likely to provide. This includes writing descriptions rather than "dialogs" or overly descriptive sentences. Workers were told that a good description should help others recognize the image from a collection of similar images. Instructions also mentioned that work with poor grammar would be rejected. Snapshots of our interface can be found in [42]. Overall, we had 465 subjects for ABSTRACT-50S and 683 subjects for PASCAL-50S datasets. We ensure that each sentence for an image is written by a different subject. The average sentence length for the ABSTRACT-50S dataset is 10.59 words compared to 8.8 words for PASCAL-50S.

6. Experimental Setup

The goals of our experiments are two-fold:

- Evaluating how well our proposed metric CIDEr captures human judgement of consensus, as compared to existing metrics.
- Comparing existing state-of-the-art automatic image description approaches in terms of how well the descriptions they produce match human consensus of image descriptions.

We first describe how we select candidate sentences for evaluation and the metrics we use for comparison to CIDEr. Finally, we list the various automatic image description approaches and our experimental set up.

Candidate Sentences: On ABSTRACT-50S, we use 48 of our 50 sentences as reference sentences (sentence A in our triplet annotation). The remaining 2 sentences per image can be used as candidate sentences. We form 400 pairs of candidate sentences (B and C in our triplet annotation). These include two kinds of pairs. The first are 200 human–human correct pairs (HC), where we pick two human sentences describing the same image. The second kind are 200 human–human incorrect pairs (HI), where one of the sentences is a human description for the image and the other is also a human sentence but describing some other image from the dataset picked at random.

For PASCAL-50S, our candidate sentences come from a diverse set of sources: human sentences from the UIUC Pascal Sentence Dataset as well as machine-generated sentences from five automatic image description methods. These span both retrieval-based and generation-based methods: Midge [29], Babytalk [22], Story [14], and two versions of Translating Video Content to Natural Language Descriptions [37] (Video and Video+).[4] We form 4,000 pairs of candidate sentences (again, B and C for our triplet annotation). These include four types of pairs (1,000 each). The first two are human–human correct (HC) and human–human incorrect (HI) similar to ABSTRACT-50S. The third are human–machine (HM) pairs formed by pairing a human sentence describing an image with a machine generated sentence describing the same image. Finally, the fourth are machine–machine (MM) pairs, where we compare two machine generated sentences describing the same image. We pick the machine generated sentences randomly, so that each method participates in roughly equal number of pairs, on a diverse set of images. Ours is the first work to perform a comprehensive evaluation across these different kinds of sentences.

For consistency, we drop two reference sentences for the PASCAL-50S evaluations so that we evaluate on both datasets (ABSTRACT-50S and PASCAL-50S) with a maximum of 48 reference sentences.

Metrics: The existing metrics used in the community for evaluation of image description approaches are BLEU [33], ROUGE [46] and METEOR [1]. BLEU is precision-based and ROUGE is recall-based. More specifically, image description methods have used versions of BLEU called BLEU_1 and BLEU_4, and a version of ROUGE called ROUGE_1. A recent survey paper [12] has used a different version of ROUGE called ROUGE_S, as well as the machine translation metric called METEOR [1]. We now briefly describe these metrics. More details can be found in [42]. **BLEU** (BiLingual Evaluation Understudy) [33] is a popular metric for Machine Translation (MT) evaluation. It computes an n-gram based precision for the candidate sentence with respect to the references. The key idea of BLEU is to compute precision by *clipping*. Clipping computes precision for a word, based on the maximum number of times it occurs in any reference sentence. Thus, a candidate sentence saying "The The The", would get credit for saying only one "The", if the word occurs at most once across individual references. BLEU computes the geometric mean of the n-gram precisions and adds a brevity-penalty to discourage overly short sentences. The most common formulation of BLEU is BLEU4, which uses 1-grams up to 4-grams, though lower-order variations such as BLEU1 (unigram BLEU) and BLEU2 (unigram and bigram BLEU) are also used. Similar to [12, 18] for evaluating image descriptions, we compute BLEU at the sentence level. For machine translation BLEU is most often computed at the corpus level where correlation with human judgment is high; the correlation is poor at the level of individual sentences. In this paper we are specifically interested in the evaluation of accuracies on individual sentences. **ROUGE** stands for Recall Oriented Understudy of Gisting Evaluation [46]. It computes n-gram based recall for the candidate sentence with respect to the references. It is a popular metric for summarization evaluation. Similar to BLEU, versions of ROUGE can be computed by varying the n-gram count. Two other versions

[4] We thank the authors of these approaches for making their outputs available to us.

of ROUGE are ROUGE$_S$ and ROUGE$_L$. These compute an F-measure with a recall bias using *skip-bigrams* and *longest common subsequence* respectively, between the candidate and each reference sentence. Skip-bigrams are all pairs of ordered words in a sentence, sampled non-consecutively. Given these scores, they return the maximum score across the set of references as the judgment of quality. **METEOR** stands for Metric for Evaluation of Translation with Explicit ORdering [1]. Similar to ROUGE$_L$ and ROUGE$_S$, it also computes the F-measure based on matches, and returns the maximum score over a set of references as its judgment of quality. However, it resolves word-level correspondences in a more sophisticated manner, using exact matches, stemming and semantic similarity. It optimizes over matches minimizing *chunkiness*. Minimizing chunkiness implies that matches should be consecutive, wherever possible. It also sets parameters favoring recall over precision in its F-measure computation. We implement all the metrics, except for METEOR, for which we use [8] (version 1.5). Similar to BLEU, we also aggregate METEOR scores at the sentence level.

Machine Approaches: We comprehensively evaluate which machine generation methods are best at matching consensus sentences. For this experiment, we select a subset of 100 images from the UIUC Pascal Sentence Dataset for which we have outputs for all the five machine description methods used in our evaluation: Midge [29], Babytalk [22], Story [14], and two versions of Translating Video Content to Natural Language Descriptions [37] (Video and Video+). For each image, we form all $^5C_2$ pairs of machine–machine sentences. This ensures that each machine approach gets compared to all other machine approaches on each image. This gives us 1,000 pairs. We form triplets by "tripling" each pair with 20 random reference sentences. We collect human judgement of consensus using our triplet annotation modality as well as evaluate our proposed automatic consensus metric CIDEr using the same reference sentences. In both cases, we count the fraction of times a machine description method beats another method in terms of being more similar to the reference sentences. To the best of our knowledge, we are the first work to perform an exhaustive evaluation of automated image captioning, across retrieval- and generation-based methods.

7. Results

In this section we evaluate the effectiveness of our consensus-based metric CIDEr on the PASCAL-50S and ABSTRACT-50S datasets. We begin by exploring how many sentences are sufficient for reliably evaluating our consensus metric. Next, we compare our metric against several other commonly used metrics on the task of matching human consensus. Then, using CIDEr we evaluate several existing automatic image description approaches. Finally, we compare performance of humans and CIDEr at predicting consensus.

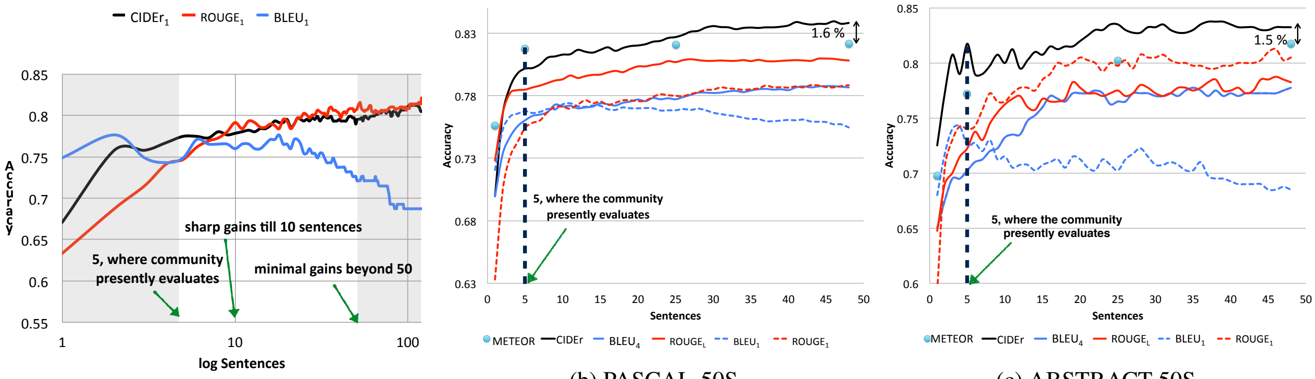
7.1. How many sentences are enough?

We begin by analyzing how the number of reference sentences affects the accuracy of automated metrics. To quantify this, we collect 120 sentences for a subset of 50 randomly sampled images from the UIUC Pascal Sentence Dataset. We then pool human–human correct, human–machine, machine–machine and human–human incorrect sentence pairs (179 in total) and get triplet annotations. This gives us the ground truth consensus score for all pairs. We evaluate BLEU$_1$, ROUGE$_1$ and CIDEr$_1$ with up to 100 reference sentences used to score the candidate sentences. We find that the accuracy improves for the first 10 sentences (Fig. 3a) for all metrics. From 1 to 5 sentences, the agreement for ROUGE$_1$ improves from 0.63 to 0.77. Both ROUGE$_1$ and CIDEr$_1$ continue to improve until reaching 50 sentences, after which the results begin to saturate somewhat. Curiously, BLEU$_1$ shows a decrease in performance with more sentences. BLEU does a max operation over sentence level matches, and thus as more sentences are used, the likelihood of matching a lower quality reference sentence increases. Based on this pilot, we collect 50 sentences per image for our ABSTRACT-50S and PASCAL-50S datasets. For the remaining experiments we report results using 1 to 50 sentences.

7.2. Accuracy of Automated Metrics

We evaluate the performance of CIDEr, BLEU, ROUGE and METEOR at matching the human consensus scores in Fig. 3. That is, for each metric we compute the scores for two candidate sentences. The metric is correct if the sentence with higher score is the same as the sentence chosen by our human studies as being more similar to the reference sentences. The candidate sentences are both human and machine generated. For BLEU and ROUGE we show both their popular versions and the version we found to give best performance. We sample METEOR at fewer points due to high run-time. For a more comprehensive evaluation across different versions of each metric, please see [42].

At 48 sentences, we find that CIDEr is the best performing metric, on both ABSTRACT-50S as well as PASCAL-50S. It is followed by METEOR on each dataset. Even using only 5 sentences, both CIDEr and METEOR perform well in comparison to BLEU and ROUGE. CIDEr beats METEOR at 5 sentences on ABSTRACT-50S, whereas METEOR does better at five sentences on PASCAL-50S. This is because METEOR incorporates soft-similarity, which helps when using fewer sentences. However, METEOR, despite its sophistication does a max across reference scores, which limits its ability to utilize larger numbers of reference sentences. Popular metrics like



(a) Pilot Performance

(b) PASCAL-50S

(c) ABSTRACT-50S

Figure 3: **(a)**: We show accuracy (y-axis) versus *log* number of sentences (x-axis) for our pilot study. We note that the gains saturate after 50 sentences. **(b) and (c)**: Accuracy of automated metrics (y-axis) plotted against number of reference sentences (x-axis) for PASCAL-50S (b) and ABSTRACT-50S (c). Metrics currently used for evaluating image descriptions are shown in *dashed* lines. Other existing metrics and our proposed metric are in **solid** lines. CIDEr is the best performing metric on both datasets followed by METEOR. METEOR is sampled at fewer points, due to high run-time. Note that more reference sentences that we collect clearly help.

ROUGE$_1$ and BLEU$_1$ are not as good at capturing consensus. CIDEr provides consistent performance across both the datasets, giving 84% and 84% accuracy on PASCAL-50S and ABSTRACT-50S respectively.

Considering previous papers only used 5 reference sentences per image for evaluation, the relative boost in performance is substantial. Using BLEU$_1$ or ROUGE$_1$ at 5 sentences, we obtained 76% and 74% accuracy on PASCAL-50S. With CIDEr at 48 sentences, we achieve 84% accuracy. This brings automated evaluation much closer to human performance (90%, details in Sec. 7.4). On the Flickr8K dataset [18] with human judgments on 1-5 ratings, METEOR has a correlation (Spearman's ρ) of 0.56 [12], whereas CIDEr achieves a correlation of 0.58 with human judgments.[5]

We next show the best performing versions of the metrics CIDEr, BLEU, ROUGE and METEOR on PASCAL-50S and ABSTRACT-50S, respectively, for different kinds of candidate pairs (Table 1). As discussed in Sec. 5 we have four kinds of pairs: (human–human correct) HC, (human–human incorrect) HI, (human–machine) HM, and (machine–machine) MM. We find that out of six cases, our proposed automated metric is best in five. We show significant gains on the challenging MM and HC tasks that involve differentiating between fine-grained differences between sentences (two machine generated sentences and two human generated sentences). This result is encouraging because it indicates that the CIDEr metric will continue to perform well as image description methods continue to improve. On the easier tasks of judging consensus on HI and HM pairs, all methods perform well.

Metric	PASCAL-50S				ABSTRACT-50S	
	HC	HI	HM	MM	HC	HI
BLEU$_4$	64.8	97.7	93.8	63.6	65.5	93.0
ROUGE	66.3	98.5	95.8	64.4	**71.5**	91.0
METEOR	65.2	99.3	**96.4**	67.7	69.5	94.0
CIDEr	**71.8**	**99.7**	92.1	**72.2**	**71.5**	**96.0**

Table 1: Results on four kinds of pairs for PASCAL-50S and two kinds of pairs for ABSTRACT-50S. The best performing method is shown in **bold**. Note: we use ROUGE$_L$ for PASCAL-50S and ROUGE$_1$ for ABSTRACT-50S

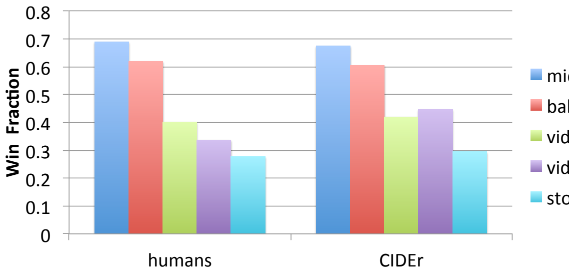


Figure 4: Fraction of times a machine generation approach wins against the other four (y-axis), plotted for human annotations and our automated metric, CIDEr.

7.3. Which automatic image description approaches produce consensus descriptions?

We have shown that CIDEr and our new datasets containing 50 sentences per image provide a more accurate metric over previous approaches. We now use it to evaluate some existing automatic image description approaches. Our methodology for conducting this experiment is described in Sec. 6. Our results are shown in Fig. 4. We show the fraction of times an approach is rated better than other ap-

[5] We thank Desmond Elliot for the result.

proaches on the y-axis. We note that Midge [29] is rated as having the best consensus by both humans and CIDEr, followed by Babytalk [22]. Story [14] is the lowest ranked, by both humans and CIDEr. Humans and CIDEr differ on the ranking of the two video approaches (Video and Video+) [37]. We calcuate the Pearson's correlation between the fraction of wins for a method on human annotations and using CIDEr. We find that humans and CIDEr agree with a high correlation (0.98).

7.4. Human Performance

In our final set of experiments we measure human performance at predicting which of two candidate sentences better matches the consensus. Human performance puts into context how clearly consensus is defined, and provides a loose bound on how well we can expect automated metrics to perform. We evaluate both human and machine performance at predicting consensus on all 4,000 pairs from PASCAL-50S dataset and 400 pairs from the ABSTRACT-50S dataset described in Sec. 6. To create the same experimental set up for both humans and machines, we obtain ground truth consensus for each of the pairs using our triplet annotation on 24 references out of 48. For predicting consensus, humans (via triplet annotations) and machines both use the remaining 24 sentences as reference sentences. We find that the best machine performance is 82% on PASCAL-50S using CIDEr, in contrast to human performance which is at 90%. On the ABSTRACT-50S dataset, CIDEr is at 82% accuracy, whereas human performance is at 83%.

8. Gameability and Evaluation Server

Gameability When optimizing an algorithm for a specific metric undesirable results may be achieved. The "gaming" of a metric may result in sentences with high scores, yet produce poor results when judged by a human. To help defend against the future gaming of the CIDEr metric, we propose several modifications to the basic CIDEr metric called CIDEr-D.

First, we propose the removal of stemming. When performing stemming the singular and plural forms of nouns and different tenses of verbs are mapped to the same token. The removal of stemming ensures the correct forms of words are used. Second, in some cases the basic CIDEr metric produces higher scores when words of higher confidence are repeated over long sentences. To reduce this effect, we introduce a Gaussian penalty based on the difference between candidate and reference sentence lengths. Finally, the sentence length penalty may be gamed by repeating confident words or phrases until the desired sentence length is achieved. We combat this by adding clipping to the n-gram counts in the CIDEr$_n$ numerator. That is, for a specific n-gram we clip the number of candidate occurrences to the number of reference occurrences. This penalizes the repetition of specific n-grams beyond the number of times they occur in the reference sentence. These changes result in the following equation (analogous to Equation 2):

$$\text{CIDEr-D}_n(c_i, S_i) = \frac{10}{m} \sum_j e^{\frac{-(l(c_i)-l(s_{ij}))^2}{2\sigma^2}} * \frac{\min(\boldsymbol{g^n}(c_i), \boldsymbol{g^n}(s_{ij})) \cdot \boldsymbol{g^n}(s_{ij})}{\|\boldsymbol{g^n}(c_i)\| \|\boldsymbol{g^n}(s_{ij})\|}, \quad (4)$$

Where $l(c_i)$ and $l(s_{ij})$ denote the lengths of candidate and reference sentences respectively. We use $\sigma = 6$. A factor of 10 is added to make the CIDEr-D scores numerically similar to other metrics.

The final CIDEr-D metric is computed in a similar manner to CIDEr (analogous to Equation 3):

$$\text{CIDEr-D}(c_i, S_i) = \sum_{n=1}^{N} w_n \text{CIDEr-D}_n(c_i, S_i), \quad (5)$$

Similar to CIDEr, uniform weights are used. We found that this version of the metric has a rank correlation (Spearman's ρ) of 0.94 with the original CIDEr metric while being more robust to gaming. Qualitative Examples of ranking can be found in [42].

Evaluation Server To enable systematic evaluation and benchmarking of image description approaches based on consensus, we have made CIDEr-D available as a metric in the MS COCO caption evaluation server [5].

9. Conclusion

In this work we proposed a consensus-based evaluation protocol for image description evaluation. Our protocol enables an objective comparison of machine generation approaches based on their "human-likeness", without having to make arbitrary calls on weighing content, grammar, saliency, *etc.* with respect to each other. We introduce an annotation modality for measuring consensus, a metric CIDEr for automatically computing consensus, and two datasets, PASCAL-50S and ABSTRACT-50S with 50 sentences per image. We demonstrate CIDEr has improved accuracy over existing metrics for measuring consensus.

Acknowledgements: We thank Chris Quirk, Margaret Mitchell and Michel Galley for helpful discussions in formulating CIDEr-D. This work was supported in part by The Paul G. Allen Family Foundation Allen Distinguished Investigator award to D.P.

References

[1] S. Banerjee and A. Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. pages 65–72, 2005. 2, 5, 6

[2] A. C. Berg, T. L. Berg, H. D. III, J. Dodge, A. Goyal, X. Han, A. Mensch, M. Mitchell, A. Sood, K. Stratos, and K. Yamaguchi. Understanding and predicting importance in images. In *CVPR*. IEEE, 2012. 2

[3] R. Bogacz, E. Brown, J. Moehlis, P. Holmes, and J. D. Cohen. The physics of optimal decision making: a formal analysis of models of performance in two-alternative forced-choice tasks. *Psychol Rev*, 113(4):700–765, Oct. 2006. 1, 4

[4] C. Callison-burch and M. Osborne. Re-evaluating the role of bleu in machine translation research. In *In EACL*, pages 249–256, 2006. 1, 3

[5] 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*, Apr. 2015. 2, 8

[6] X. Chen and C. L. Zitnick. Learning a recurrent visual representation for image caption generation. *CoRR*, abs/1411.5654, 2014. 2, 3

[7] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009. 1

[8] M. Denkowski and A. Lavie. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the EACL 2014 Workshop on Statistical Machine Translation*, 2014. 6

[9] P. K. Dokania, A. Behl, C. V. Jawahar, and P. M. Kumar. Learning to rank using high-order information. *ECCV*, 2014. 1

[10] J. Donahue, L. A. Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell. Long-term recurrent convolutional networks for visual recognition and description. *CoRR*, abs/1411.4389, 2014. 2, 3

[11] D. Elliott and F. Keller. Image description using visual dependency representations. In *EMNLP*, pages 1292–1302. ACL, 2013. 1, 3

[12] D. Elliott and F. Keller. Comparing automatic evaluation measures for image description. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 452–457, Baltimore, Maryland, June 2014. Association for Computational Linguistics. 3, 5, 7

[13] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2010 (VOC2010) Results. http://www.pascal-network.org/challenges/VOC/voc2010/workshop/index.html. 1

[14] A. Farhadi, M. Hejrati, M. A. Sadeghi, P. Young, C. Rashtchian, J. Hockenmaier, and D. Forsyth. Every picture tells a story: Generating sentences from images. In *Proceedings of the 11th European Conference on Computer Vision: Part IV*, ECCV'10, 2010. 2, 5, 6, 8

[15] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9):1627–1645, 2010. 1

[16] A. Gupta and L. S. Davis. Beyond nouns: Exploiting prepositions and comparative adjectives for learning visual classifiers. In D. A. Forsyth, P. H. S. Torr, and A. Zisserman, editors, *ECCV (1)*, volume 5302 of *Lecture Notes in Computer Science*, pages 16–29. Springer, 2008. 2

[17] A. Gupta, Y. Verma, and C. Jawahar. Choosing linguistics over vision to describe images. 2012. 2

[18] M. Hodosh, P. Young, and J. Hockenmaier. Framing image description as a ranking task: Data, models and evaluation metrics. *J. Artif. Intell. Res. (JAIR)*, 47:853–899, 2013. 1, 2, 3, 5, 7

[19] A. Karpathy and L. Fei-Fei. Deep visual-semantic alignments for generating image descriptions. *CoRR*, abs/1412.2306, 2014. 2, 3

[20] A. Karpathy, A. Joulin, and L. Fei-Fei. Deep fragment embeddings for bidirectional image sentence mapping. *CoRR*, 2014. 2

[21] R. Kiros, R. Salakhutdinov, and R. S. Zemel. Unifying visual-semantic embeddings with multimodal neural language models. *CoRR*, abs/1411.2539, 2014. 2, 3

[22] G. Kulkarni, V. Premraj, S. Dhar, S. Li, Y. Choi, A. C. Berg, and T. L. Berg. Baby talk: Understanding and generating image descriptions. In *Proceedings of the 24th CVPR*, 2011. 1, 2, 3, 5, 6, 8

[23] C. H. Lampert, H. Nickisch, and S. Harmeling. Learning to detect unseen object classes by betweenclass attribute transfer. In *In CVPR*, 2009. 1, 2

[24] S. Li, G. Kulkarni, T. L. Berg, A. C. Berg, and Y. Choi. Composing simple image descriptions using web-scale n-grams. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning*, CoNLL '11, pages 220–228, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics. 2

[25] T. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft COCO: Common objects in context. In *ECCV*, 2014. 3, 4

[26] S. Maji, L. Bourdev, and J. Malik. Action recognition from a distributed representation of pose and appearance. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011. 1

[27] J. Mao, W. Xu, Y. Yang, J. Wang, and A. L. Yuille. Explain images with multimodal recurrent neural networks. *CoRR*, abs/1410.1090, 2014. 2, 3

[28] D. Martin, C. Fowlkes, D. Tal, and J. Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *Proc. 8th Int'l Conf. Computer Vision*, volume 2, pages 416–423, July 2001. 1

[29] M. Mitchell, X. Han, and J. Hayes. Midge: Generating descriptions of images. In *Proceedings of the Seventh International Natural Language Generation Conference*, INLG '12, pages 131–133, Stroudsburg, PA, USA, 2012. Association for Computational Linguistics. 1, 2, 5, 6, 8

[30] H. Mller, P. Clough, T. Deselaers, and B. Caputo. *ImageCLEF: Experimental Evaluation in Visual Information Retrieval*. Springer Publishing Company, Incorporated, 1st edition, 2010. 3

[31] A. Nenkova and R. J. Passonneau. Evaluating content selection in summarization: The pyramid method. In *HLT-NAACL*, pages 145–152, 2004. 3

[32] V. Ordonez, G. Kulkarni, and T. L. Berg. Im2text: Describing images using 1 million captioned photographs. In *Neural Information Processing Systems (NIPS)*, 2011. 2, 3, 4

[33] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 311–318, Stroudsburg, PA, USA, 2002. Association for Computational Linguistics. 1, 2, 5

[34] D. Parikh and K. Grauman. Relative Attributes. In *ICCV*, 2011. 2

[35] C. Rashtchian, P. Young, M. Hodosh, and J. Hockenmaier. Collecting image annotations using amazon's mechanical turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, CSLDAMT '10, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics. 2, 3, 4

[36] S. Robertson. Understanding inverse document frequency: On theoretical arguments for idf. *Journal of Documentation*, 60:2004, 2004. 4

[37] M. Rohrbach, W. Qiu, I. Titov, S. Thater, M. Pinkal, and B. Schiele. Translating video content to natural language descriptions. In *IEEE International Conference on Computer Vision (ICCV)*, December 2013. 1, 2, 5, 6, 8

[38] M. A. Sadeghi and A. Farhadi. Recognition using visual phrases. 2011. 2

[39] D. Scharstein and R. Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *Int. J. Comput. Vision*, 2002. 1

[40] A. Sorokin and D. Forsyth. Utility data annotation with amazon mechanical turk. In *Computer Vision and Pattern Recognition Workshops, 2008. CVPRW '08. IEEE Computer Society Conference on*, pages 1–8, June 2008. 1

[41] O. Tamuz, C. Liu, S. Belongie, O. Shamir, and A. T. Kalai. Adaptively learning the crowd kernel. In *In ICML11*, 2011. 1

[42] R. Vedantam, C. L. Zitnick, and D. Parikh. Cider: Consensus-based image description evaluation. *CoRR*, abs/1411.5726, 2014. 1, 5, 6, 8

[43] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell: A neural image caption generator. *CoRR*, abs/1411.4555, 2014. 2, 3

[44] Y. Yang, C. L. Teo, H. D. III, and Y. Aloimonos. Corpus-guided sentence generation of natural images. In *EMNLP*. ACL, 2011. 1, 2

[45] M. Yatskar, M. Galley, L. Vanderwende, and L. Zettlemoyer. See no evil, say no evil: Description generation from densely labeled images. In *Proceedings of the Third Joint Conference on Lexical and Computational Semantics (*SEM 2014)*, page 110120, Dublin, Ireland, August 2014. Association for Computational Linguistics and Dublin City University. 1, 2

[46] C. yew Lin. Rouge: a package for automatic evaluation of summaries. pages 25–26, 2004. 1, 3, 5

[47] C. L. Zitnick and D. Parikh. Bringing semantics into focus using visual abstraction. In *CVPR*, 2013. 2, 3, 4