Microsoft COCO Captions: Data Collection and Evaluation Server

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam
Saurabh Gupta, Piotr Dollár, C. Lawrence Zitnick

Abstract—In this paper we describe the Microsoft COCO Caption dataset and evaluation server. When completed, the dataset will contain over one and a half million captions describing over 330,000 images. For the training and validation images, five independent human generated captions will be provided. To ensure consistency in evaluation of automatic caption generation algorithms, an evaluation server is used. The evaluation server receives candidate captions and scores them using several popular metrics, including BLEU, METEOR, ROUGE and CIDEr. Instructions for using the evaluation server are provided.

1 Introduction

The automatic generation of captions for images is a long standing and challenging problem in artificial intelligence [1], [2], [3], [4], [5], [6], [7], [8], [9], [10], [11], [12], [13], [14], [15], [16], [17], [18], [19]. Research in this area spans numerous domains, such as computer vision, natural language processing, and machine learning. Recently there has been a surprising resurgence of interest in this area [20], [21], [22], [23], [24], [25], [26], [27], [28], [29], [30], due to the renewed interest in neural network learning techniques [31], [32] and increasingly large datasets [33], [34], [35], [7], [36], [37], [38].

In this paper, we describe our process of collecting captions for the Microsoft COCO Caption dataset, and the evaluation server we have set up to evaluate performance of different algorithms. The MS COCO caption dataset contains human generated captions for images contained in the Microsoft Common Objects in COntext (COCO) dataset [38]. Similar to previous datasets [7], [36], we collect our captions using Amazon's Mechanical Turk (AMT). Upon completion of the dataset it will contain over a million captions.

When evaluating image caption generation algorithms, it is essential that a consistent evaluation protocol is used. Comparing results from different approaches can be difficult since numerous evaluation metrics exist [39], [40], [41], [42]. To further complicate matters the implementations of these metrics often differ. To help alleviate these issues, we have built an evaluation server to enable consistency in evaluation of different caption generation approaches. Using the testing data, our evaluation server evaluates captions output by different approaches using numerous automatic metrics: BLEU [39], METEOR [41], ROUGE [40] and CIDEr [42]. We hope to augment these results with human evaluations on an annual basis.

Fig. 1: Example images and captions from the Microsoft COCO Caption dataset.

This paper is organized as follows: First we describe the data collection process. Next, we describe the caption evaluation server and the various metrics used. Human performance using these metrics are provided. Finally the annotation format and instructions for using the evaluation server are described for those who wish to submit results. We conclude by discussing future directions and known issues.

2 Data Collection

In this section we describe how the data is gathered for the MS COCO captions dataset. For images, we use the dataset collected by Microsoft COCO [38]. These images are split into training, validation and testing sets.

- *Xinlei Chen is with Carnegie Mellon University.*
- *Hao Fang is with the University of Washington.*
- *T.Y. Lin is with Cornell NYC Tech.*
- *Ramakrishna Vedantam is with Virginia Tech.*
- *Saurabh Gupta is with the Univeristy of California, Berkeley.*
- *P. Dollár is with Facebook AI Research.*
- *C. L. Zitnick is with Microsoft Research, Redmond.*

The images were gathered by searching for pairs of 80 object categories and various scene types on Flickr. The goal of the MS COCO image collection process was to gather images containing multiple objects in their natural context. Given the visual complexity of most images in the dataset, they pose an interesting and difficult challenge for image captioning.

For generating a dataset of image captions, the same training, validation and testing sets were used as in the original MS COCO dataset. Two datasets were collected. The first dataset MS COCO c5 contains five reference captions for every image in the MS COCO training, validation and testing datasets. The second dataset MS COCO c40 contains 40 reference sentences for a randomly chosen 5,000 images from the MS COCO testing dataset. MS COCO c40 was created since many automatic evaluation metrics achieve higher correlation with human judgement when given more reference sentences [42]. MS COCO c40 may be expanded to include the MS COCO validation dataset in the future.

Our process for gathering captions received significant inspiration from the work of Young etal. [36] and Hodosh etal. [7] that collected captions on Flickr images using Amazon's Mechanical Turk (AMT). Each of our captions are also generated using human subjects on AMT. Each subject was shown the user interface in Figure 2. The subjects were instructed to:

- Describe all the important parts of the scene.
- Do not start the sentences with "There is.
- Do not describe unimportant details.
- Do not describe things that might have happened in the future or past.
- Do not describe what a person might say.
- Do not give people proper names.
- The sentences should contain at least 8 words.

The number of captions gathered is 413,915 captions for 82,783 images in training, 202,520 captions for 40,504 images in validation and 379,249 captions for 40,775 images in testing including 179,189 for MS COCO c5 and 200,060 for MS COCO c40. For each testing image, we collected one additional caption to compute the scores of human performance for comparing scores of machine generated captions. The total number of collected captions is 1,026,459. We plan to collect captions for the MS COCO 2015 dataset when it is released, which should approximately double the size of the caption dataset. The AMT interface may be obtained from the MS COCO website.

3 Caption Evaluation

In this section we describe the MS COCO caption evaluation server. Instructions for using the evaluation server are provided in Section 5. As input the evaluation server receives candidate captions for both the validation and testing datasets in the format specified in Section 5. The validation and test images are provided to the submitter. However, the human generated reference sentences are only provided for the validation set. The reference sentences for the testing set are kept private to reduce the risk of overfitting.

Fig. 2: Example user interface for the caption gathering task.

Numerous evaluation metrics are computed on both MS COCO c5 and MS COCO c40. These include BLEU-1, BLEU-2, BLEU-3, BLEU-4, ROUGE-L, METEOR and CIDEr-D. The details of the these metrics are described next.

3.1 Tokenization and preprocessing

Both the candidate captions and the reference captions are pre-processed by the evaluation server. To tokenize the captions, we use Stanford PTBTokenizer in Stanford CoreNLP tools (version 3.4.1) [43] which mimics Penn Treebank 3 tokenization. In addition, punctuations[1] are removed from the tokenized captions.

3.2 Evaluation metrics

Our goal is to automatically evaluate for an image I_i the quality of a candidate caption c_i given a set of reference captions $S_i = \{s_{i1}, \ldots, s_{im}\} \in S$. The caption sentences are represented using sets of n-grams, where an n-gram $\omega_k \in \Omega$ is a set of one or more ordered words. In this paper we explore n-grams with one to four words. No stemming is performed on the words. The number of times an n-gram ω_k occurs in a sentence s_{ij} is denoted $h_k(s_{ij})$ or $h_k(c_i)$ for the candidate sentence $c_i \in C$.

3.3 BLEU

BLEU [39] is a popular machine translation metric that analyzes the co-occurrences of n-grams between the candidate and reference sentences. It computes a corpus-level clipped n-gram precision between sentences as follows:

$$CP_n(C,S) = \frac{\sum_i \sum_k \min(h_k(c_i), \max_{j \in m} h_k(s_{ij}))}{\sum_i \sum_k h_k(c_i)}, \quad (1)$$

1. The full list of punctuations: {", ", ', ', -LRB-, -RRB-, -LCB-, -RCB-, ., ?, !, ,, :, -, –, ..., ;}.

where k indexes the set of possible n-grams of length n. The clipped precision metric limits the number of times an n-gram may be counted to the maximum number of times it is observed in a single reference sentence. Note that CP_n is a precision score and it favors short sentences. So a brevity penalty is also used:

$$b(C,S) = \begin{cases} 1 & \text{if } l_C > l_S \\ e^{1-l_S/l_C} & \text{if } l_C \leq l_S \end{cases}, \tag{2}$$

where l_C is the total length of candidate sentences c_i's and l_S is the length of the corpus-level effective reference length. When there are multiple references for a candidate sentence, we choose to use the *closest* reference length for the brevity penalty.

The overall BLEU score is computed using a weighted geometric mean of the individual n-gram precision:

$$BLEU_N(C,S) = b(C,S)\exp\left(\sum_{n=1}^{N} w_n \log CP_n(C,S)\right), \tag{3}$$

where $N = 1, 2, 3, 4$ and w_n is typically held constant for all n.

BLEU has shown good performance for corpus-level comparisons over which a high number of n-gram matches exist. However, at a sentence-level the n-gram matches for higher n rarely occur. As a result, BLEU performs poorly when comparing individual sentences.

3.4 ROUGE

ROUGE [40] is a set of evaluation metrics designed to evaluate text summarization algorithms.

1) ROUGE$_N$: The first ROUGE metric computes a simple n-gram recall over all reference summaries given a candidate sentence:

$$ROUGE_N(c_i,S_i) = \frac{\sum_j \sum_k \min(h_k(c_i), h_k(s_{ij}))}{\sum_j \sum_k h_k(s_{ij})} \tag{4}$$

2) ROUGE$_L$: ROUGE$_L$ uses a measure based on the Longest Common Subsequence (LCS). An LCS is a set words shared by two sentences which occur in the same order. However, unlike n-grams there may be words in between the words that create the LCS. Given the length $l(c_i, s_{ij})$ of the LCS between a pair of sentences, ROUGE$_L$ is found by computing an F-measure:

$$R_l = \max_j \frac{l(c_i,s_{ij})}{|s_{ij}|} \tag{5}$$

$$P_l = \max_j \frac{l(c_i,s_{ij})}{|c_i|} \tag{6}$$

$$ROUGE_L(c_i,S_i) = \frac{(1+\beta^2)R_l P_l}{R_l + \beta^2 P_l} \tag{7}$$

R_l and P_l are recall and precision of LCS. β is usually set to favor *recall* ($\beta = 1.2$). Since n-grams are implicit in this measure due to the use of the LCS, they need not be specified.

3) ROUGE$_S$: The final ROUGE metric uses skip bi-grams instead of the LCS or n-grams. Skip bi-grams are pairs of ordered words in a sentence. However, similar to the LCS, words may be skipped between pairs of words. Thus, a sentence with 4 words would have $C_2^4 = 6$ skip bi-grams. Precision and recall are again incorporated to compute an F-measure score. If $f_k(s_{ij})$ is the skip bi-gram count for sentence s_{ij}, ROUGE$_S$ is computed as:

$$R_s = \max_j \frac{\sum_k \min(f_k(c_i), f_k(s_{ij}))}{\sum_k f_k(s_{ij})} \tag{8}$$

$$P_s = \max_j \frac{\sum_k \min(f_k(c_i), f_k(s_{ij}))}{\sum_k f_k(c_i)} \tag{9}$$

$$ROUGE_S(c_i,S_i) = \frac{(1+\beta^2)R_s P_s}{R_s + \beta^2 P_s} \tag{10}$$

Skip bi-grams are capable of capturing long range sentence structure. In practice, skip bi-grams are computed so that the component words occur at a distance of at most 4 from each other.

3.5 METEOR

METEOR [41] is calculated by generating an alignment between the words in the candidate and reference sentences, with an aim of 1:1 correspondence. This alignment is computed while minimizing the number of chunks, ch, of contiguous and identically ordered tokens in the sentence pair. The alignment is based on exact token matching, followed by WordNet synonyms [44], stemmed tokens and then paraphrases. Given a set of alignments, m, the METEOR score is the harmonic mean of precision P_m and recall R_m between the best scoring reference and candidate:

$$Pen = \gamma\left(\frac{ch}{m}\right)^{\theta} \tag{11}$$

$$F_{mean} = \frac{P_m R_m}{\alpha P_m + (1-\alpha)R_m} \tag{12}$$

$$P_m = \frac{|m|}{\sum_k h_k(c_i)} \tag{13}$$

$$R_m = \frac{|m|}{\sum_k h_k(s_{ij})} \tag{14}$$

$$METEOR = (1 - Pen)F_{mean} \tag{15}$$

Thus, the final METEOR score includes a penalty Pen based on chunkiness of resolved matches and a harmonic mean term that gives the quality of the resolved matches. The default parameters α, γ and θ are used for this evaluation. Note that similar to BLEU, statistics of precision and recall are first aggregated over the entire corpus, which are then combined to give the corpus-level METEOR score.

3.6 CIDEr

The CIDEr metric [42] measures consensus in image captions by performing a Term Frequency Inverse Document Frequency (TF-IDF) weighting for each n-gram. 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. CIDEr computes the TF-IDF weighting $g_k(s_{ij})$ for each n-gram ω_k using:

$$g_k(s_{ij}) = \frac{h_k(s_{ij})}{\sum_{\omega_l \in \Omega} h_l(s_{ij})} \log\left(\frac{|I|}{\sum_{I_p \in I} \min(1, \sum_q h_k(s_{pq}))}\right), \quad (16)$$

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 sentences describing an image, while IDF reduces the weight of n-grams that commonly occur across all descriptions. 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.

The 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:

$$\text{CIDEr}_n(c_i, S_i) = \frac{1}{m} \sum_j \frac{\boldsymbol{g^n}(c_i) \cdot \boldsymbol{g^n}(s_{ij})}{\|\boldsymbol{g^n}(c_i)\| \|\boldsymbol{g^n}(s_{ij})\|}, \quad (17)$$

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 $\boldsymbol{g^n}(s_{ij})$.

Higher order (longer) n-grams to are used to capture grammatical properties as well as richer semantics. Scores from n-grams of varying lengths are combined as follows:

$$\text{CIDEr}(c_i, S_i) = \sum_{n=1}^{N} w_n \text{CIDEr}_n(c_i, S_i), \quad (18)$$

Uniform weights are used $w_n = 1/N$. $N = 4$ is used.

CIDEr-D is a modification to CIDEr to make it more robust to gaming. Gaming refers to the phenomenon where a sentence that is poorly judged by humans tends to score highly with an automated metric. To defend the CIDEr metric against gaming effects, [42] add clipping and a length based gaussian penalty to the CIDEr metric described above. This results in the following equations for CIDEr-D:

TABLE 1: Human Agreement for Image Captioning: Various metrics when benchmarking a human generated caption against ground truth captions.

Metric Name	MS COCO c5	MS COCO c40
BLEU 1	0.663	0.880
BLEU 2	0.469	0.744
BLEU 3	0.321	0.603
BLEU 4	0.217	0.471
METEOR	0.252	0.335
ROUGE_L	0.484	0.626
CIDEr-D	0.854	0.910

$$\text{CIDEr-D}_n(c_i, S_i) = \frac{10}{m} \sum_j e^{\frac{-(l(c_i)-l(s_{ij}))^2}{2\sigma^2}} * \frac{min(\boldsymbol{g^n}(c_i), \boldsymbol{g^n}(s_{ij})) \cdot \boldsymbol{g^n}(s_{ij})}{\|\boldsymbol{g^n}(c_i)\| \|\boldsymbol{g^n}(s_{ij})\|}, \quad (19)$$

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

The final CIDEr-D metric is computed in a similar manner to CIDEr (analogous to eqn. 18):

$$\text{CIDEr-D}(c_i, S_i) = \sum_{n=1}^{N} w_n \text{CIDEr-D}_n(c_i, S_i), \quad (20)$$

Note that just like the BLEU and ROUGE metrics, CIDEr-D does not use stemming. We adopt the CIDEr-D metric for the evaluation server.

4 Human Performance

In this section, we study the human agreement among humans at this task. We start with analyzing the inter-human agreement for image captioning (Section. 4.1) and then analyze human agreement for the word prediction sub-task and provide a simple model which explains human agreement for this sub-task (Section. 4.2).

4.1 Human Agreement for Image Captioning

When examining human agreement on captions, it becomes clear that there are many equivalent ways to say essentially the same thing. We quantify this by conducting the following experiment: We collect one additional human caption for each image in the test set and treat this caption as the prediction. Using the MS COCO caption evaluation server we compute the various metrics. The results are tabulated in Table 1.

4.2 Human Agreement for Word Prediction

We can do a similar analysis for human agreement at the sub-task of word prediction. Consider the task of tagging the image with words that occur in the captions. For this task, we can compute the human precision and recall for

TABLE 2: Model defintions.

o	=	object or visual concept
w	=	word associated with o
n	=	total number of images
k	=	number of captions per image
q	=	$P(o=1)$
p	=	$P(w=1\|o=1)$

a given word w by benchmarking words used in the $k+1$ human caption with respect to words used in the first k reference captions. Note that we use weighted versions of precision and recall, where each negative image has a weight of 1 and each positive image has a weight equal to the number of captions containing the word w. Human precision (H_p) and human recall (H_r) can be computed from the counts of how many subjects out of k use the word w to describe a given image over the whole dataset.

We plot H_p versus H_r for a set of nouns, verbs and adjectives, and all 1000 words considered in Figure 3. Nouns referring to animals like 'elephant' have a high recall, which means that if an 'elephant' exists in the image, a subject is likely to talk about it (which makes intuitive sense, given 'elephant' images are somewhat rare, and there are no alternative words that could be used instead of 'elephant'). On the other hand, an adjective like 'bright' is used inconsistently and hence has low recall. Interestingly, words with high recall also have high precision. Indeed, all the points of human agreement appear to lie on a one-dimensional curve in the two-dimension precision-recall space.

This observation motivates us to propose a simple model for when subjects use a particular word w for describing an image. Let o denote an object or visual concept associated with word w, n be the total number of images, and k be the number of reference captions. Next, let $q = P(o=1)$ be the probability that object o exists in an image. For clarity these definitions are summarized in Table 2. We make two simplifications. First, we ignore *image level saliency* and instead focus on *word level saliency*. Specifically, we only model $p = P(w=1|o=1)$, the probability a subject uses w given that o is in the image, without conditioning on the image itself. Second, we assume that $P(w=1|o=0) = 0$, i.e. that a subject does not use w unless o is in the image. As we will show, even with these simplifications our model suffices to explain the empirical observations in Figure 3 to a reasonable degree of accuracy.

Given these assumptions, we can model human precision $\widetilde{H_p}$ and recall $\widetilde{H_r}$ for a word w given only p and k. First, given k captions per image, we need to compute the expected number of (1) captions containing w (cw), (2) true positives (tp), and (3) false positives (fp). Note that in our definition there can be up to k true positives per image (if $cw = k$, i.e. each of the k captions contains word w) but at most 1 false positive (if none of the k captions contains w). The expectations, in terms of k, p, and q are:

$$\begin{aligned}
E[cw] &= \Sigma_{i=1}^{k} P(w^i = 1) \\
&= \Sigma_i P(w^i = 1|o=1)P(o=1) \\
&\quad +\Sigma_i P(w^i = 1|o=0)P(o=0) \\
&= kpq + 0 = \boxed{kpq} \\
E[tp] &= \Sigma_{i=1}^{k} P(w^i = 1 \wedge w^{k+1} = 1) \\
&= \Sigma_i P(w^i = 1 \wedge w^{k+1} = 1|o=1)P(o=1) \\
&\quad +\Sigma_i P(w^i = 1 \wedge w^{k+1} = 1|o=0)P(o=0) \\
&= kppq + 0 = \boxed{kp^2q} \\
E[fp] &= P(w^1 \ldots w^k = 0 \wedge w^{k+1} = 1) \\
&= P(o=1 \wedge w^1 \ldots w^k = 0 \wedge w^{k+1} = 1) \\
&\quad +P(o=0 \wedge w^1 \ldots w^k = 0 \wedge w^{k+1} = 1) \\
&= q(1-p)^k p + 0 = \boxed{q(1-p)^k p}
\end{aligned}$$

In the above $w^i = 1$ denotes that w appeared in the i^{th} caption. Note that we are also assuming independence between subjects conditioned on o. We can now define model precision and recall as:

$$\begin{aligned}
\widetilde{H_p} &:= \frac{nE[tp]}{nE[tp] + nE[fp]} = \frac{pk}{pk + (1-p)^k} \\
\widetilde{H_r} &:= \frac{nE[tp]}{nE[cw]} = p
\end{aligned}$$

Note that these expressions are independent of q and only depend on p. Interestingly, because of the use of weighted precision and recall, the recall for a category comes out to be exactly equal to p, the probability a subject uses w given that o is in the image.

We set $k = 4$ and vary p to plot $\widetilde{H_p}$ versus $\widetilde{H_r}$, getting the curve as shown in blue in Figure 3 (bottom left). The curve explains the observed data quite well, closely matching the precision-recall tradeoffs of the empirical data (although not perfectly). We can also reduce the number of captions from four, and look at how the empirical and predicted precision and recall change. Figure 3 (bottom right), shows this variation as we reduce the number of reference captions per image from four to one annotations. We see that the points of human agreement remain at the same recall value, but decrease in their precision, which is consistent with what the model predicts. Also, the human precision at infinite subjects will approach one, which is again reasonable given that a subject will only use the word w if the corresponding object is in the image (and in the presence of infinite subjects someone else will also use the word w).

In fact, the fixed recall value can help us recover p, the probability that a subject will use the word w in describing the image given the object is present. Nouns like 'elephant' and 'tennis' have large p, which is reasonable. Verbs and adjectives, on the other hand, have smaller p values, which can be justified from the fact that a) subjects are less likely to describe attributes

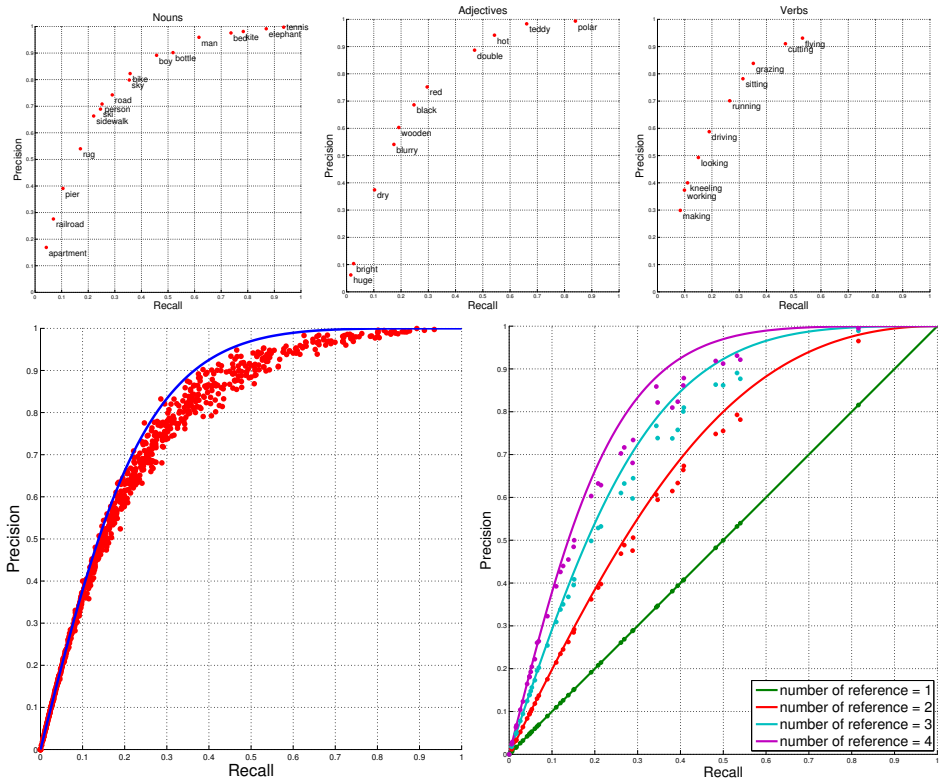

Fig. 3: Precision-recall points for human agreement: we compute precision and recall by treating one human caption as prediction and benchmark it against the others to obtain points on the precision recall curve. We plot these points for example nouns (top left), adjectives (top center), and verbs (top right), and for all words (bottom left). We also plot the fit of our model for human agreement with the empirical data (bottom left) and show how the human agreement changes with different number of captions being used (bottom right). We see that the human agreement point remains at the same recall value but dips in precision when using fewer captions.

of objects and b) subjects might use a different word (synonym) to describe the same attribute.

This analysis of human agreement also motivates using a different metric for measuring performance. We propose Precision at Human Recall (PHR) as a metric for measuring performance of a vision system performing this task. Given that human recall for a particular word is fixed and precision varies with the number of annotations, we can look at system precision at human recall and compare it with human precision to report the performance of the vision system.

5 Evaluation Server Instructions

Directions on how to use the MS COCO caption evaluation server can be found on the MS COCO website. The evaluation server is hosted by CodaLab. To participate, a user account on CodaLab must be created. The participants need to generate results on both the validation and testing datasets. When training for the generation of results on the test dataset, the training and validation dataset may be used as the participant sees fit. That is, the validation dataset may be used for training if desired. However, when generating results on the validation set, we ask participants to only train on the training dataset, and only use the validation dataset for tuning meta-parameters. Two JSON files should be created corresponding to results on each dataset in the following format:

$$
\begin{array}{l}
[\{ \\
\text{“}image\_id\text{”} \quad : \quad int, \\
\text{“}caption\text{”} \quad : \quad str, \\
\}]
\end{array}
$$

The results may then be placed into a zip file and uploaded to the server for evaluation. Code is also provided on GitHub to evaluate results on the validation dataset without having to upload to the server. The number of submissions per user is limited to a fixed amount.

6 Discussion

Many challenges exist when creating an image caption dataset. As stated in [7], [42], [45] the captions generated by human subjects can vary significantly. However even though two captions may be very different, they may be judged equally “good” by human subjects. Designing effective automatic evaluation metrics that are highly correlated with human judgment remains a difficult challenge [7], [42], [45], [46]. We hope that by releasing

results on the validation data, we can help enable future research in this area.

Since automatic evaluation metrics do not always correspond to human judgment, we hope to conduct experiments using human subjects to judge the quality of automatically generated captions, which are most similar to human captions, and whether they are grammatically correct [45], [42], [7], [4], [5]. This is essential to determining whether future algorithms are indeed improving, or whether they are merely over fitting to a specific metric. These human experiments will also allow us to evaluate the automatic evaluation metrics themselves, and see which ones are correlated to human judgment.

REFERENCES

[1] K. Barnard and D. Forsyth, "Learning the semantics of words and pictures," in *ICCV*, vol. 2, 2001, pp. 408–415.

[2] K. Barnard, P. Duygulu, D. Forsyth, N. De Freitas, D. M. Blei, and M. I. Jordan, "Matching words and pictures," *JMLR*, vol. 3, pp. 1107–1135, 2003.

[3] V. Lavrenko, R. Manmatha, and J. Jeon, "A model for learning the semantics of pictures," in *NIPS*, 2003.

[4] G. Kulkarni, V. Premraj, S. Dhar, S. Li, Y. Choi, A. C. Berg, and T. L. Berg, "Baby talk: Understanding and generating simple image descriptions," in *CVPR*, 2011.

[5] M. Mitchell, X. Han, J. Dodge, A. Mensch, A. Goyal, A. Berg, K. Yamaguchi, T. Berg, K. Stratos, and H. Daumé III, "Midge: Generating image descriptions from computer vision detections," in *EACL*, 2012.

[6] 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 *ECCV*, 2010.

[7] M. Hodosh, P. Young, and J. Hockenmaier, "Framing image description as a ranking task: Data, models and evaluation metrics." *JAIR*, vol. 47, pp. 853–899, 2013.

[8] P. Kuznetsova, V. Ordonez, A. C. Berg, T. L. Berg, and Y. Choi, "Collective generation of natural image descriptions," in *ACL*, 2012.

[9] Y. Yang, C. L. Teo, H. Daumé III, and Y. Aloimonos, "Corpus-guided sentence generation of natural images," in *EMNLP*, 2011.

[10] A. Gupta, Y. Verma, and C. Jawahar, "Choosing linguistics over vision to describe images." in *AAAI*, 2012.

[11] E. Bruni, G. Boleda, M. Baroni, and N.-K. Tran, "Distributional semantics in technicolor," in *ACL*, 2012.

[12] Y. Feng and M. Lapata, "Automatic caption generation for news images," *TPAMI*, vol. 35, no. 4, pp. 797–812, 2013.

[13] D. Elliott and F. Keller, "Image description using visual dependency representations." in *EMNLP*, 2013, pp. 1292–1302.

[14] A. Karpathy, A. Joulin, and F.-F. Li, "Deep fragment embeddings for bidirectional image sentence mapping," in *NIPS*, 2014.

[15] Y. Gong, L. Wang, M. Hodosh, J. Hockenmaier, and S. Lazebnik, "Improving image-sentence embeddings using large weakly annotated photo collections," in *ECCV*, 2014, pp. 529–545.

[16] R. Mason and E. Charniak, "Nonparametric method for data-driven image captioning," in *ACL*, 2014.

[17] P. Kuznetsova, V. Ordonez, T. Berg, and Y. Choi, "Treetalk: Composition and compression of trees for image descriptions," *TACL*, vol. 2, pp. 351–362, 2014.

[18] K. Ramnath, S. Baker, L. Vanderwende, M. El-Saban, S. N. Sinha, A. Kannan, N. Hassan, M. Galley, Y. Yang, D. Ramanan, A. Bergamo, and L. Torresani, "Autocaption: Automatic caption generation for personal photos," in *WACV*, 2014.

[19] A. Lazaridou, E. Bruni, and M. Baroni, "Is this a wampimuk? cross-modal mapping between distributional semantics and the visual world," in *ACL*, 2014.

[20] R. Kiros, R. Salakhutdinov, and R. Zemel, "Multimodal neural language models," in *ICML*, 2014.

[21] J. Mao, W. Xu, Y. Yang, J. Wang, and A. L. Yuille, "Explain images with multimodal recurrent neural networks," *arXiv preprint arXiv:1410.1090*, 2014.

[22] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," *arXiv preprint arXiv:1411.4555*, 2014.

[23] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," *arXiv preprint arXiv:1412.2306*, 2014.

[24] R. Kiros, R. Salakhutdinov, and R. S. Zemel, "Unifying visual-semantic embeddings with multimodal neural language models," *arXiv preprint arXiv:1411.2539*, 2014.

[25] 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," *arXiv preprint arXiv:1411.4389*, 2014.

[26] H. Fang, S. Gupta, F. Iandola, R. Srivastava, L. Deng, P. Dollár, J. Gao, X. He, M. Mitchell, J. Platt *et al.*, "From captions to visual concepts and back," *arXiv preprint arXiv:1411.4952*, 2014.

[27] X. Chen and C. L. Zitnick, "Learning a recurrent visual representation for image caption generation," *arXiv preprint arXiv:1411.5654*, 2014.

[28] R. Lebret, P. O. Pinheiro, and R. Collobert, "Phrase-based image captioning," *arXiv preprint arXiv:1502.03671*, 2015.

[29] ——, "Simple image description generator via a linear phrase-based approach," *arXiv preprint arXiv:1412.8419*, 2014.

[30] A. Lazaridou, N. T. Pham, and M. Baroni, "Combining language and vision with a multimodal skip-gram model," *arXiv preprint arXiv:1501.02598*, 2015.

[31] A. Krizhevsky, I. Sutskever, and G. Hinton, "ImageNet classification with deep convolutional neural networks," in *NIPS*, 2012.

[32] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[33] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A Large-Scale Hierarchical Image Database," in *CVPR*, 2009.

[34] M. Grubinger, P. Clough, H. Müller, and T. Deselaers, "The iapr tc-12 benchmark: A new evaluation resource for visual information systems," in *LREC Workshop on Language Resources for Content-based Image Retrieval*, 2006.

[35] V. Ordonez, G. Kulkarni, and T. Berg, "Im2text: Describing images using 1 million captioned photographs." in *NIPS*, 2011.

[36] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier, "From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions," *TACL*, vol. 2, pp. 67–78, 2014.

[37] J. Chen, P. Kuznetsova, D. Warren, and Y. Choi, "Déjá image-captions: A corpus of expressive image descriptions in repetition," in *NAACL*, 2015.

[38] 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.

[39] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *ACL*, 2002.

[40] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," in *ACL Workshop*, 2004.

[41] M. Denkowski and A. Lavie, "Meteor universal: Language specific translation evaluation for any target language," in *EACL Workshop on Statistical Machine Translation*, 2014.

[42] R. Vedantam, C. L. Zitnick, and D. Parikh, "Cider: Consensus-based image description evaluation," *arXiv preprint arXiv:1411.5726*, 2014.

[43] C. D. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. J. Bethard, and D. McClosky, "The Stanford CoreNLP natural language processing toolkit," in *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 2014, pp. 55–60. [Online]. Available: http://www.aclweb.org/anthology/P/P14/P14-5010

[44] G. A. Miller, "Wordnet: a lexical database for english," *Communications of the ACM*, vol. 38, no. 11, pp. 39–41, 1995.

[45] 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*, vol. 2, 2014, pp. 452–457.

[46] C. Callison-Burch, M. Osborne, and P. Koehn, "Re-evaluation the role of bleu in machine translation research." in *EACL*, vol. 6, 2006, pp. 249–256.