Improved Image Caption Rating – Datasets, Game, and Model

Andrew Taylor Scott
ats@sfsu.edu
Department of Computer Science,
San Francisco State University
San Francisco, CA, USA

Lothar D Narins
lnarins@mail.sfsu.edu
Department of Computer Science,
San Francisco State University
San Francisco, CA, USA

Anagha Kulkarni
ak@sfsu.edu
Department of Computer Science,
San Francisco State University
San Francisco, CA, USA

Mar Castanon
marcastanon4@gmail.com
Department of Computer Science,
San Francisco State University
San Francisco, CA, USA

Benjamin Kao
bkao1@mail.sfsu.edu
Department of Computer Science,
San Francisco State University
San Francisco, CA, USA

Shasta Ihorn
sihorn@sfsu.edu
Department of Psychology,
San Francisco State University
San Francisco, CA, USA

Yue-Ting Siu
ting.siu@wssb.wa.gov
Department of Special Education,
San Francisco State University
San Francisco, CA, USA

Ilmi Yoon
ilmi@sfsu.edu
Department of Computer Science,
San Francisco State University
San Francisco, CA, USA

ABSTRACT

How well a caption fits an image can be difficult to assess due to the subjective nature of caption quality. What is a *good* caption? We investigate this problem by focusing on image-caption ratings and by generating high quality datasets from human feedback with gamification. We validate the datasets by showing a higher level of inter-rater agreement, and by using them to train custom machine learning models to predict new ratings. Our approach outperforms previous metrics – the resulting datasets are more easily learned and are of higher quality than other currently available datasets for image-caption rating.

CCS CONCEPTS

• **Computing methodologies** → **Supervised learning by regression**; *Computer vision*; *Natural language processing*.

KEYWORDS

human-in-the-loop, image captioning, visually-impaired, multimodal learning

ACM Reference Format:
Andrew Taylor Scott, Lothar D Narins, Anagha Kulkarni, Mar Castanon, Benjamin Kao, Shasta Ihorn, Yue-Ting Siu, and Ilmi Yoon. 2023. Improved Image Caption Rating – Datasets, Game, and Model. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems (CHI EA '23), April 23–28, 2023, Hamburg, Germany*. ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3544549.3585632

CHI EA '23, April 23–28, 2023, Hamburg, Germany

ACM ISBN 978-1-4503-9422-2/23/04.
https://doi.org/10.1145/3544549.3585632

1 INTRODUCTION

Image-caption rating (ICR) is the task of estimating the quality of a caption for a given image. It is a growing area of research in the computer vision, natural language processing (NLP), and human computer interaction communities [10, 13, 18, 19, 32], and is becoming especially relevant as computer generated captions are used more and more for closed-captioning and descriptive annotations. Identifying how good a caption is can be difficult because of the multiple aspects involved. At minimum, a good caption should correctly identify all relevant objects, where things are in space, the setting in which the objects are shown, and the interpretation of events depicted in the image. Prior work has contributed multiple ICR datasets where human annotators were asked to assign quality ratings to image-caption pairs, most often with an ad hoc rating scale [12, 20, 22, 29]. While these datasets have been tremendously valuable in advancing the field and have been used extensively [1, 2, 19, 27, 30, 33], several of them suffer from high skew in the ratings with predominantly poor quality captions. Our work seeks to improve the rigor, quality, and scalability of ICR datasets by providing a better rating scale, dataset generation process, dataset validation process, and rating predictor. Our datasets, code, and pretrained models are available through our project website.[1]

A distinguishing characteristic of caption quality metrics is whether or not they depend on reference captions. Traditional NLP approaches such as BLEU, CIDEr, ROUGE, METEOR, and SPICE are monomodal, reference-based, and are unable to measure the nuance found in rich image captions [1, 7, 21, 25, 29]. The success of reference-free approaches has opened new possibilities for ICR estimation in applications without access to reference captions [4, 10, 11, 28]. However, existing reference-free metrics lack rating granularity due to the use of oversimplified scales. Our work seeks to lift this restriction by retaining the benefits of a reference-free rating approach while employing a 5-level rating scale that

[1] https://ai.youdescribe.org

can capture subtle aspects of caption quality, such as correctness, completeness, and local and global context.

Although a more detailed scale can offer higher rating capacity, it can also increase the complexity of the rating task, potentially making the task more subjective and tedious. To tackle this downside, we propose a novel human-in-the-loop solution that employs a rigorous human rater training procedure and a gamified data generation process with built-in quality control. The gamification aspect keeps the raters engaged and the built-in quality control steers raters toward higher quality answers. Our contributions include a web-based image-caption rating game, novel datasets, inter-rater agreement validation, and baseline models that can be used in place of other metrics. We define our research objective to be: *Design and develop a reliable and scalable data generation approach for the task of image-caption rating.*

2 RELATED WORK

2.1 ICR Scale and Datasets

There are only a handful of image-caption datasets with ratings: Google Image Caption (GIC) [20], Conceptual Captions Challenge[2], CapEval1k [18], PASCAL50s [29], and Flickr8k-Expert [12]. Among those previous datasets, we identify two sources of error. One is the rating scale itself which is generally not defined well enough to capture significant meaning. The other source of error is the human rater's inconsistent application of that scale.

The GIC dataset is large with 140k image-caption pairs but only has ratings on a binary scale ("good" or "bad"). One common problem with a binary scale is that it does not have the capacity to handle incomplete or partially correct captions. Figure 1 includes two image and caption examples which were considered "good" by raters but lack contextual information. The Conceptual Caption Challenge dataset contains 5,000 image-caption pairs with binary ratings that were collected in the same manner as GIC and has the same problem with incomplete or partially correct captions. CapEval1k is good because it has a five point scale but the scale is not specific and the dataset is very small with only 1,000 captions and 250 images. PASCAL50s only has ratings in a free-form, non-numeric scale. These datasets are all either too small or have rating scales that are not defined well enough to capture significant meaning.

Table 1: Flickr8k-Expert rating scale.

r	**Meaning**
4	Describes the image without any errors.
3	Describes the image with minor errors.
2	Is somewhat related to the image.
1	Is unrelated to the image.

Flickr8k-Expert is the only one with ratings on a graduated, numeric scale with specific meanings, and represents the current state of the art in ICR datasets. Flickr8k-Expert has 5,822 captions across 1,000 images rated by 21 college students such that each caption has received 3 ratings in a range of 1 to 4 (Table 1). The

[2]https://www.conceptualcaptions.com

Caption: "audience members at the conference hall".

Caption: "street art on the sidewalk".

Figure 1: Two examples from Google Image Caption Dataset illustrating the limitation of a binary scale. For both images all 10 raters chose "good" rating even though salient aspects of the image are not captured by the caption (lecturer at the front of a sparsely attended lecture hall, with a scale model bridge on the table; artist wearing a sunhat, crouching on the ground, holding a piece of paper).

complexity of the ICR task combined with an underspecified rating scale, however, lead to fairly low inter-rater agreement which we show in Section 4 and Table 3. In addition, the rating distribution in Flickr8k-Expert skews heavily toward ratings of 1 and 2, indicating overall lower caption quality (Figure 4a).

2.2 Reference-free ICR Estimators

VSEPP [8] and CLIPScore [10] are multimodal models that use cosine similarity to measure the alignment between an image embedding and text embedding vector representation in a shared visual-linguistic feature space to provide reference-free metrics. While cosine similarity is useful for measuring the alignment between these modes, fine-tuning or manipulation of the similarity of the image and text embeddings for domain-specific applications remains difficult.

Cui and colleagues [6] created a deep learning method for determining if a caption for an image was human-written or machine generated. This binary classifier is not sufficient for diverse use cases since it only detects the source of a caption rather than its quality. Levinboim and colleagues [20] trained a deep learning, image-caption Quality Estimation (QE) model on the GIC dataset. This model inherits the same limitations from the GIC dataset because of its binary classification system discussed in Figure 1.

Lee and colleagues [18] developed Unreferenced Metric for Image Captioning (UMIC) using UNITER [5] via contrastive learning, a process where the model is trained to compare and discriminate the ground-truth captions and diverse synthetic negative samples. Jiang and colleagues [13] developed TIGEr (Text-to-Image Grounding for Image Caption Evaluation) by refining the mapping of the image and the caption pair into carefully grounded vector spaces. These estimators improve over prior metrics but still do not surpass state of the art methods on Flickr8k-Expert.

3 METHODS

To make the task of image-caption rating easier, we developed a rating scale that captures all essential aspects of image-caption quality, and we have developed an engaging tool to facilitate high-quality data generation from human raters. We have evaluated the

efficacy of our approach by analyzing inter-rater agreement and by conducting experiments with custom machine learning models for rating prediction. These components provide a robust, high-quality and scalable image-caption rating methodology.

3.1 Rating Scale

As shown in Table 2, our 5-level scale is designed to capture the extent of four essential aspects of image captions: (1) accuracy (e.g. 'objects are partially correctly identified'), (2) completeness (e.g. 'identifies most of the objects'), (3) local context (e.g. 'where things are in space'), and (4) global context and inferential information (e.g. 'interpretation of overall setting and/or event'). The rating scale was designed through consulting with measurement experts, accessibility experts, and informed by the Integrative Inferential Reasoning (IIR) framework, which is concerned with how humans integrate context for image and text narratives [3]. Our scale differs from the Flickr8k-Expert scale in the ratings of 2 and above. Our rating scale captures more nuance with less ambiguity.

Table 2: Improved image caption rating scale.

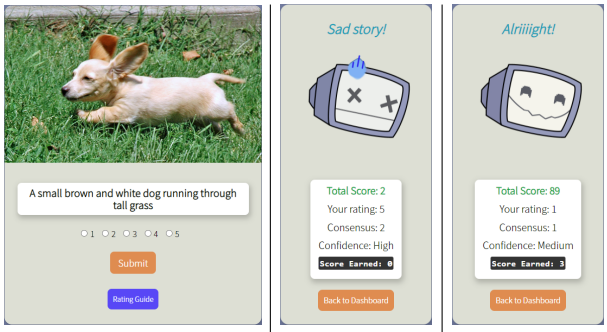
r	Meaning
5	Objects, a general scene, and actions are correctly identified if present in the image. The caption describes what is seen and where things are in space. Interpretation of overall setting and/or event is included.
4	Objects and/or a general scene and/or an action are correctly identified but not every element is completely identified. The caption describes what is seen and where things are in space. There is no interpretation of an event.
3	Relevant objects are correctly identified. The caption describes what is seen but not where objects are in space. There is no description of the overall setting and no interpretation of an event.
2	Objects are partially correctly identified with some errors, but the caption is accurate enough to give an idea of what is happening in the image. The caption identifies most of the objects but might not identify everything. There is no interpretation of what anything means.
1	Objects are incorrectly identified. The caption gives the wrong idea about what is happening in the image.

3.2 Gamified Data Generation

To promote human rater engagement we frame the image-caption rating task as a single-player, asynchronous, point-based game that is played in a web browser. We draw inspiration from the image labeling "ESP Game" [31]. In the ESP Game, two players play against one another, trying to guess the word their opponent is thinking for a given image. In our Rating Game, each player tries to predict the community consensus rating for each image-caption pair in order to maximize their point score.

Human Rater Training: Before their ratings are accepted into our datasets, human raters must go through a tutorial to familiarize themselves with the Rating Game and scale. In the tutorial, the rater is guided through 10 image-caption pairs and asked to rate them. Their answer is compared with the consensus and they are given a score and feedback along with an explanation for why that rating was chosen as the consensus. After the tutorial, they move on to the Rating Game. The rater must score at least 25 points in the first 20 image-caption pairs or else they are put on "probation": their scores are reset and they must complete 20 image-caption pairs from a curated set with a score of at least 25 before they can move on. Point scores do not propagate into the game from the tutorial or probation period. The tutorial and probation image-caption pairs were not included in the final dataset.

Rating Game: The Rating Game is designed to provide a user experience that is similar to that of the Human Rater Training: an image-caption pair is displayed, and the player selects a rating from the 5-level scale (Figure 2a). The player has to wait at least 3 seconds before submitting their answer. After the player submits their rating they receive feedback and a score based on how their rating compares to the consensus rating of the other players so far (Figures 2b and 2c). The scoring algorithm, probation status, and 3 second delay are intended to discourage guessing and promote higher-quality ratings.



(a) Image, caption and 5-level scale. The meanings of the rating scale can be consulted anytime through the 'Rating Guide' button.

(b) Player feedback when the rating does not match the consensus.

(c) Player feedback when the rating matches the consensus.

Figure 2: Image caption rating game.

Scoring Algorithm: The consensus rating, r, is computed by rounding the average of all the previous ratings for that image-caption pair. The score, s, assigned to the player models two intuitions: the first is that if the rating, x, is near the consensus the player should get a high score, and if the rating is far the player should get a low score. The second intuition is that if there is a high variance in the previous ratings the penalty for being far from the consensus should be lower, and if there is a low variance the penalty should be higher. These two intuitions are captured in Algorithm 1, which computes the player score, s, where n is the total number of ratings available for the current image-caption pair (including the player's rating), σ^2 is the variance of the previous ratings, and $V_{max} = 4$ is the largest possible value of the variance.

Since this scoring mechanism requires previous ratings, we have a "cold start" problem for the initial ratings. We deal with this problem by using computer generated initial ratings – from VSEPP at first and from our models once we trained them. We include two

computer generated initial ratings to make the scoring algorithm smoother. As the human ratings are collected, the initial ratings are replaced with human ratings.

Algorithm 1: Player score, s.

$v \leftarrow 1 + \frac{1}{n}(1 + \frac{(n-1)\sigma^2}{V_{max}})$
$d \leftarrow \frac{|x-r|}{v}$
if $d \leq 0.25$ **then** $s \leftarrow 2$
else if $d \leq 0.5$ **then** $s \leftarrow 1$
else if $d \leq 1$ **then** $s \leftarrow 0$
else if $d \leq 1.75$ **then** $s \leftarrow -1$
else $s \leftarrow -2$

3.3 New Datasets

Our work contributes three new ICR datasets: CocoG, FlickrG, and a combination of these two datasets, IICR16k. To generate the datasets, human raters were invited to play the Rating Game through a departmental mailing list and incentivized with monetary rewards that depended on their scores. Participants were required to fill out a consent form.

CocoG: For this dataset, 8,990 distinct images were chosen at random from the MS-COCO 2014 validation set [22]. The captions were selected from four sources: (1) the original MS-COCO captions, (2) captions generated using the Pythia framework [14], (3) captions generated using the GLACNet model [17], and (4) mismatched MS-COCO captions from other images. Most images received one caption but some received two captions, leading to 9,982 image-caption pairs. The goal of sampling from these four sources was to create a balanced and wide variety of caption quality examples. The original captions (1) are known to be good quality, the Pythia captions (2) and GLACNet captions (3) are medium quality, and mismatched captions (4) are poor quality.

There were 72 college students who generated the ratings for the CocoG dataset, earning about $15 per hour on average, depending on their score, and averaging 102 minutes of rating time. The participants took about 10 seconds on average to rate an image-caption pair. Each image-caption pair received between 4 and 7 ratings for a total of 44,252 ratings.

FlickrG: The intent of the FlickrG dataset was to be able to compare our raters directly with Flickr8k-Expert raters. To generate this dataset we used a subset of image-caption pairs from Flickr8k-Expert. We held out 21 images and 2,350 image-caption pairs that received all 1's in the Flickr8k-Expert ratings. We held these out to save time and money because they are known to be poor quality captions. The remaining 979 images and 3,472 image-caption pairs were rated in the game by our raters.

There were 41 college students who generated the ratings for FlickrG, earning about $7.20 per hour in Amazon™ gift cards, on average, depending on their score, and averaging 95 minutes of rating time. The raters took about 13 seconds on average to rate an image-caption pair. Each image-caption pair received between 5 and 7 ratings for a total of 17,721 ratings from the game. We combined these with the known poor quality captions that were held out, keeping the original Flickr8k-Expert ratings that were all 1's, and we removed the 158 pairs that are commonly excluded from previous studies because the captions appear in references for that image [10, 29]. This resulted in 5,664 image-caption pairs with 23,965 total ratings where each image-caption pair received between 3 and 7 ratings.

IICR16k: We have created a third dataset from the combination of CocoG and FlickrG, which resulted in a final IICR16k dataset with 9,990 images, 15,646 image-caption pairs, and a total of 68,217 ratings where each image-caption pair received between 3 and 7 ratings.

3.4 Model Design

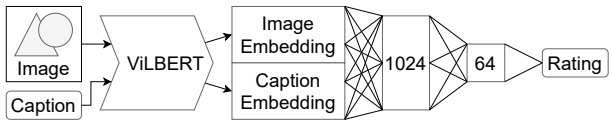


Figure 3: Schematic diagram of our model architecture.

To evaluate and study our datasets, we develop a baseline reference-free image-caption rating predictor (Figure 3). For the representation layer, the predictor takes a novel approach by employing ViLBERT [23, 24] co-attention embeddings, which were trained on the Conceptual Captions dataset [27]. Specifically, the input to the model (image-caption pairs) is represented by a 2048-dimensional vector created by concatenating the image and text embeddings from the final hidden layer of a pretrained ViLBERT model.

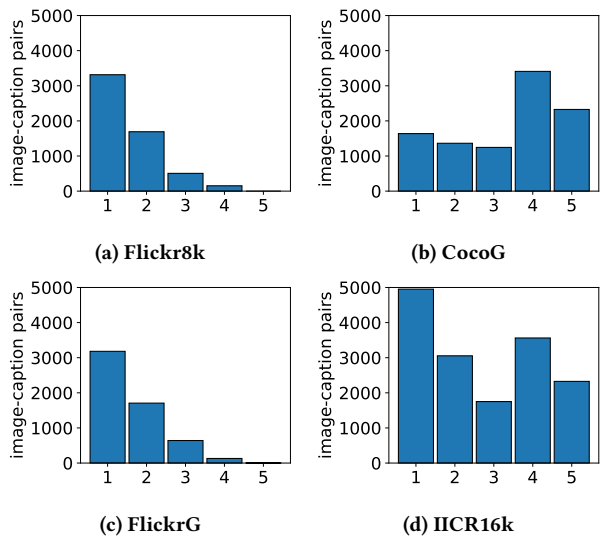
For the rating prediction, our model uses a simple feedforward neural network with two hidden layers. The first hidden layer has 1,024 neurons with a ReLU activation, and the second hidden layer has 64 neurons with a ReLU activation. The output layer consists of a single neuron. We use 80% dropout on both hidden layers. We use mean squared error (MSE) for the loss. We train for 4,000 epochs, with a batch size of 256, and with a learning rate of 10^{-5}, decayed by 1% every 15 epochs. All of our models in Table 4 were trained with these hyperparameters.

4 RESULTS AND ANALYSIS

The standard benchmark in previous work for evaluating ICR estimators is Kendall's τ coefficient [15]. It measures the rank-correlation between pairs of ratings. We take a new approach and use Kendall's τ to analyze inter-rater agreement, along with Kendall's W and Fleiss' κ, in Section 4.1, and in Section 4.2, we use Kendall's τ in the more traditional way to compare our work with other state of the art methods.

4.1 Dataset Analysis

Rating Distributions: For comparative analysis, the rating distributions of the Flickr8k, CocoG, FlickrG, and IICR16k datasets are illustrated in Figure 4. For each image-caption pair, the rounded average of all available ratings for that pair is used as the single value rating for the pair. In Figure 4b, the CocoG dataset has more ratings of 4 and 5, showing a bias toward higher quality captions. In Figure 4c, the FlickrG dataset has a bias toward lower ratings



(a) Flickr8k (b) CocoG

(c) FlickrG (d) IICR16k

Figure 4: Dataset rating distributions.

because it is based on Flickr8k-Expert, which has a known skew toward lower quality captions (Figure 4a). Figure 4d shows the IICR16k dataset is more balanced in its ratings distribution.

Inter-Rater Agreement: To determine inter-rater agreement within each dataset we compute Kendall's τ [15], Kendall's W [16], and Fleiss' κ [9], on virtual raters, scaled by 100. We define a "virtual rater" to be a selection of one rating for each image-caption pair. Flickr8k-Expert only has 3 ratings per image-caption pair, which are sorted from lowest to highest. In order to compare our datasets with Flickr8k-Expert, we construct 3 virtual raters by randomly sampling 3 ratings for each image-caption pair and sorting those ratings such that virtual rater X has the lowest rating, rater Y has the middle rating, and Z has the highest rating. We do this 20 times and take the average for each metric.

Table 3: Inter-rater agreement metrics.

	Flickr8k	FlickrG	CocoG	IICR16k
τ_{X-YZ}	47.7	48.9	70.9	**75.6**
τ_{Y-XZ}	54.8	54.4	70.8	**77.8**
τ_{Z-XY}	54.0	54.9	70.6	**78.1**
W	84.0	87.0	90.8	**93.8**
κ	48.8	**48.9**	34.7	44.5
$\kappa_{4=5}$	48.8	49.5	51.9	**58.1**

In Table 3, we use Kendall's τ to see how well one virtual rater agrees with the other virtual raters individually and Kendall's W to see how well the virtual raters agree with each other overall. For example, the first row of Table 3, τ_{X-YZ}, shows taking the Kendall's τ measure of rater X with respect to raters Y and Z. The fourth row of Table 3 shows how well the raters, X, Y, and Z, all agree with one another. Kendall's τ is on a scale of [-1, 1] where closer to 1 is higher agreement, and Kendall's W is on a scale of [0, 1] where closer to 1 is higher inter-rater reliability. For Kendall's τ we are using "method A" for aggregation and variant τ_C to handle ties [10]. The resulting values show the highest correlations from the IICR16k dataset, demonstrating a high amount of inter-rater agreement.

Fleiss' κ measures how well the virtual raters categorically agree with one another above random chance. In Table 3 there are two rows, κ and $\kappa_{4=5}$, because it is sensitive to the number of categories, and there are fewer categories in Flickr8k-Expert than there are in the other datasets. We collapse categories for ratings 4 and 5 to all be categorical rating 4 in the last row of Table 3. Without collapsing the ratings to the same number of categories, the FlickrG dataset has the highest categorical correlation and Flickr8k-Expert the second highest. This is because those ratings distributions skew heavily (the majority are 1's and 2's, Figures 4a and 4c). When the values are collapsed, the highest categorical correlation is the IICR16k dataset, demonstrating a high amount of categorical agreement.

In order to compare with Flickr8k-Expert, as previsouly explained, we randomly select 3 virtual raters, sort them, and average the metric of 20 runs. The standard deviations were all in the range of 0.04 to 0.5 with the exception of the Kendall τ_{X-YZ} for FlickrG, where it was 1.6. We believe this was because of the imbalanced nature of that dataset and the larger range of rating values. Another observation can be seen in the first two columns of Table 3 where τ_{X-YZ} shows a lower correlation with the other virtual raters compared to τ_{Y-XZ} and τ_{Z-XY}. This is a consequence of the ratings being sorted from lowest to highest and those datasets being skewed toward lower ratings.

4.2 Experimental Results

We summarize the results of our experiments in Table 4 where we compare reference-based and reference-free metrics against our trained models. Each dataset was split into 64% training, 16% validation, and 20% test. All measurements are of Kendall's τ scaled by 100 and were computed on the corresponding test sets. Table 4 is organized in two sections, one for reference-based and one for reference-free metrics. The reference-based metrics are common NLP metrics and include RefCLIPScore [10] as well as ViLBERTScore [19]. RefCLIPScore is the reference-based version of CLIPScore and ViLBERTScore extends BERTScore [34] to the visual-linguistic domain by using ViLBERT embeddings. For ViLBERTScore, we used the fine-tuned model and reported the F1 metric. This corresponds to ViLBERTScore$^*_F$ in [19]. During experimental validation on the entire Flickr8k-Expert dataset, the reference-based metrics and CLIPScore were exactly the same as reported in the CLIPScore paper [10], and ViLBERTScore was exactly the same as reported in the ViLBERTScore paper [19], which gave us confidence in the numbers we received on the results on the test sets.

Besides our models, the other reference-free metrics are CLIPScore, VSEPP and VBAlignment. CLIPScore takes the weighted and scaled cosine similarity of image and text feature embeddings from pretrained CLIP [26] models. VSEPP [8] takes the cosine similarity of image and text feature embeddings from CNN and RNN models. VBAlignment refers to the visual-linguistic alignment prediction task that is used in ViLBERT [23] pretraining. The outputs of this task are two logits representing "alignment" and "non-alignment."

We apply softmax to the logits and use the "alignment" probability as the VBAlignment value for the image-caption pair. We use the pretrained model before fine-tuning on the 12 tasks from [24]. VBAlignment is a stand-in for a ViLBERT metric because ViLBERT is not a caption rater.

We used publicly available code for reference-based metrics and CLIPScore[3]; VSEPP[4]; ViLBERTScore[5]; and pretrained models and VBAlignment code from ViLBERT[6]. Our model code is written from scratch, based on Keras and Tensorflow in Python.

All experiments were run in-house on an Nvidia RTX A6000 GPU server with AMD EPYC 7302 CPU and 256 Gigabytes of system memory. All of our models took approximately 20 to 45 minutes to train. Extracting ViLBERT embeddings was the slowest part (approximately 930 milliseconds per image-caption embedding) due to the serial nature of the ViLBERT model. Our models were trained from scratch 5 times with different seeds and the average Kendall's τ value is reported. The standard deviations ranged from 0.03 to 0.3.

Our $Model_{IICR16k}$ shows the best performance across the board, according to Kendall's τ correlation with ground-truth ratings, with the exception of $Model_{Flickr8k}$. We believe that $Model_{IICR16k}$ generalizes the best because it was trained on a dataset with high-fidelity, the largest number of samples, and has the most balanced ratings distribution. However, $Model_{Flickr8k}$ did the best on the Flickr8k-Expert dataset. We believe this is because the Flickr8k-Expert dataset has fewer categories and they skew toward lower quality, making it more difficult for $Model_{IICR16k}$ to predict.

A final observation from Table 4 is that our best machine learning model achieved an average Kendall's τ of 53.7 on Flickr8k-Expert, whereas in Table 3 the best human rater achieved 54.8 on Flickr8k-Expert. It seems unreasonable to expect anything much higher from a regression model, given the difficulty humans have in correlating the ratings themselves, and it also seems satisfying that the model did so well when compared with human raters.

Table 4: Kendall's τ correlation with ground-truth ratings on a test subset of each dataset for various metrics and predictors. We used "method A" in aggregation [10] and τ_C to be consistent with prior work.

Reference-based	Flickr8k	FlickrG	CocoG	IICR16k
BLEU-1	33.7	32.4	43.7	55.8
BLEU-4	31.6	28.7	42.9	51.8
METEOR	40.4	39.0	51.9	60.2
ROUGE	33.4	32.1	42.8	53.2
CIDEr	44.1	40.9	58.1	66.6
SPICE	41.2	39.1	54.4	60.4
RefCLIPScore	51.9	51.7	65.8	71.7
ViLBERTScore	50.1	52.2	59.3	66.9
Reference-free				
CLIPScore	50.7	49.9	63.4	67.3
VSEPP	48.6	49.5	60.8	65.1
VBAlignment	49.9	50.8	60.7	65.8
$\text{Model}_{\text{Flickr8k}}$	**53.7**	54.1	61.2	71.8
$\text{Model}_{\text{FlickrG}}$	53.0	54.4	61.9	71.7
$\text{Model}_{\text{CocoG}}$	50.6	52.1	66.0	73.0
$\text{Model}_{\text{IICR16k}}$	53.1	**54.7**	**66.1**	**75.8**

5 CONCLUSIONS

The big discovery from the results of our work is that, given a high-quality dataset, it is better to model a rating predictor than to use previously available metrics for caption quality. We also showed that higher quality ICR datasets may be generated by using a human-in-the-loop game. Our datasets have greater inter-rater agreement than previous work and are more easily modeled with machine learning techniques. A reference-free, modeled approach allows us to automate the caption quality rating process, which opens up new possibilities, not only for more accurate, rich, and descriptive image annotations, but also for searching through visual content such as for image and video retrieval. The only downside to our reference-free model is the dependence on a pretrained ViLBERT model and the time it takes to extract image and text embeddings.

In this work we presented several novel contributions. We introduced a gamified human-in-the-loop dataset generation methodology to create new datasets, demonstrated a novel inter-rater agreement validation approach on those datasets, and trained custom machine learning models to be a reference-free metric. Another discovery we made in the process of this work is the use of VBAlignment as a metric, which has, as far as we know, not been done before. It is essentially a reference-free ViLBERTScore, whereas the actual ViLBERTScore [19] is reference-based.

Future work could extend the ViLBERT model to include a regression head similar to our model and fine-tune from a pretrained ViLBERT model like with the 12 tasks from [24]. It would also be useful to pretrain ViLBERT on larger datasets to create richer image and text embeddings. We would like to conduct more user studies and collect larger datasets. Additionally, we would like to experiment with different scale definitions. We had considered using a multi-dimensional scale at first but decided on a single-dimensional scale for this work. We would like to conduct a study with unpaid volunteers to see if the monetary reward had undue influence. Future work may also include experimenting with different multimodal embeddings as new algorithms are developed.

ACKNOWLEDGMENTS

This work is part of the YouDescribe Project[7] in collaboration with Smith-Kettlewell Eye Research Institute[8], and made possible by funding from Ability Central[9].

Thanks to Charity Pitcher-Cooper, Brenna Gibson Tirumalashetty, Raya Farshad, Aditya Bodi, Abhishek Das, Yash Kant, Alexander Blum, and James Mason for their invaluable contributions to this project.

REFERENCES

[1] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2016. Spice: Semantic propositional image caption evaluation. In *European conference*

[3] https://github.com/jmhessel/clipscore
[4] https://github.com/fartashf/vsepp
[5] https://github.com/hwanheelee1993/ViLBERTScore
[6] https://github.com/facebookresearch/vilbert-multi-task
[7] https://youdescribe.org
[8] https://www.ski.org/
[9] https://abilitycentral.org

on computer vision. Springer, 382–398.
[2] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*. 2425–2433.
[3] Alexander Mario Blum, James M Mason, Jinho Kim, and P David Pearson. 2020. Modeling question-answer relations: the development of the integrative inferential reasoning comic assessment. *Reading and Writing* 33, 8 (2020), 1971–2000.
[4] Eva Cetinic. 2021. Towards generating and evaluating iconographic image captions of artworks. *Journal of Imaging* 7, 8 (2021), 123.
[5] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. Uniter: Universal image-text representation learning. In *European conference on computer vision*. Springer, 104–120.
[6] Yin Cui, Guandao Yang, Andreas Veit, Xun Huang, and Serge Belongie. 2018. Learning to evaluate image captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 5804–5812.
[7] Michael Denkowski and Alon Lavie. 2014. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the ninth workshop on statistical machine translation*. 376–380.
[8] Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler. 2017. Vse++: Improving visual-semantic embeddings with hard negatives. *arXiv preprint arXiv:1707.05612* (2017).
[9] Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin* 76, 5 (1971), 378.
[10] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2021. CLIPScore: A Reference-free Evaluation Metric for Image Captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 7514–7528. https://doi.org/10.18653/v1/2021.emnlp-main.595
[11] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. 2022. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303* (2022).
[12] 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 (2013), 853–899.
[13] Ming Jiang, Qiuyuan Huang, Lei Zhang, Xin Wang, Pengchuan Zhang, Zhe Gan, Jana Diesner, and Jianfeng Gao. 2019. Tiger: Text-to-image grounding for image caption evaluation. *arXiv preprint arXiv:1909.02050* (2019).
[14] Yu Jiang, Vivek Natarajan, Xinlei Chen, Marcus Rohrbach, Dhruv Batra, and Devi Parikh. 2018. Pythia v0. 1: the winning entry to the vqa challenge 2018. *arXiv preprint arXiv:1807.09956* (2018).
[15] Maurice G Kendall. 1938. A new measure of rank correlation. *Biometrika* 30, 1/2 (1938), 81–93.
[16] Maurice G Kendall and B Babington Smith. 1939. The problem of m rankings. *The annals of mathematical statistics* 10, 3 (1939), 275–287.
[17] Taehyeong Kim, Min-Oh Heo, Seonil Son, Kyoung-Wha Park, and Byoung-Tak Zhang. 2018. Glac net: Glocal attention cascading networks for multi-image cued story generation. *arXiv preprint arXiv:1805.10973* (2018).
[18] Hwanhee Lee, Seunghyun Yoon, Franck Dernoncourt, Trung Bui, and Kyomin Jung. 2021. UMIC: An Unreferenced Metric for Image Captioning via Contrastive Learning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*. Association for Computational Linguistics, Online, 220–226. https://doi.org/10.18653/v1/2021.acl-short.29
[19] Hwanhee Lee, Seunghyun Yoon, Franck Dernoncourt, Doo Soon Kim, Trung Bui, and Kyomin Jung. 2020. Vilbertscore: Evaluating image caption using vision-and-language bert. In *Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems*. 34–39.
[20] Tomer Levinboim, Ashish V. Thapliyal, Piyush Sharma, and Radu Soricut. 2021. Quality Estimation for Image Captions Based on Large-scale Human Evaluations. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Online, 3157–3166. https://doi.org/10.18653/v1/2021.naacl-main.253
[21] Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*. 74–81.
[22] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*. Springer, 740–755.
[23] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems* 32 (2019).
[24] Jiasen Lu, Vedanuj Goswami, Marcus Rohrbach, Devi Parikh, and Stefan Lee. 2020. 12-in-1: Multi-task vision and language representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10437–10446.
[25] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*. 311–318.
[26] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*. PMLR, 8748–8763.
[27] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2556–2565.
[28] Matteo Stefanini, Marcella Cornia, Lorenzo Baraldi, Silvia Cascianelli, Giuseppe Fiameni, and Rita Cucchiara. 2021. From show to tell: A survey on image captioning. *arXiv preprint arXiv:2107.06912* (2021).
[29] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 4566–4575.
[30] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3156–3164.
[31] Luis Von Ahn and Laura Dabbish. 2004. Labeling images with a computer game. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 319–326.
[32] Xin Wang, Wenhu Chen, Yuan-Fang Wang, and William Yang Wang. 2018. No metrics are perfect: Adversarial reward learning for visual storytelling. *arXiv preprint arXiv:1804.09160* (2018).
[33] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*. PMLR, 2048–2057.
[34] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675* (2019).