Transparent Human Evaluation for Image Captioning

Anonymous ACL submission

Abstract

We establish a rubric-based human evaluation protocol for image captioning models. Our scoring rubrics and their definitions are carefully developed based on machine- and human-generated captions on the MSCOCO dataset. Each caption is evaluated along two main dimensions in a tradeoff (*precision* and *recall*) as well as other aspects that measure the text quality (*fluency*, *conciseness*, and *inclusive language*). Our evaluations demonstrate several critical problems of the current evaluation practice. Human-generated captions show substantially higher quality than machine-generated ones, especially in coverage of salient information (i.e., recall), while most automatic metrics say the opposite. Our rubric-based results reveal that CLIPScore, a recent metric that uses image features, better correlates with human judgments than conventional text-only metrics because it is more sensitive to recall. We hope that this work will promote a more transparent evaluation protocol for image captioning and its automatic metrics.

1 Introduction

Recent progress in large-scale training has pushed the state of the art in vision-language tasks (Li et al., 2020; Zhang et al., 2021, *inter alia*). One of these tasks is image captioning, whose objective is to generate a caption that describes the given image. The performance in image captioning has been primarily measured in automatic metrics (e.g., CIDEr, Vedantam et al., 2015; SPICE, Anderson et al., 2016) on popular benchmarks, such as MSCOCO (Lin et al., 2014) and Flickr8k (Hodosh et al., 2013). Use of these metrics is justified based on their correlations with human judgments collected in previous work (Hodosh et al., 2013; Elliott and Keller, 2014; Kilickaya et al., 2017, *inter alia*).

Continuous use of these previous human judgments, however, raises significant concerns for development of both captioning models and automatic metrics because of their lack of transparency.

Machines	P	R	CIDEr
A red fire hydrant spewing water on a street.	5	3	139.2
A red fire hydrant spraying water on a street.	5	3	205.2
Human			
A busted red fire hydrant spewing water all over a street creating a rainbow.	5	5	120.5

Figure 1: These machine captions are *precise* (in the scale of 1–5) but lose points in recall (i.e., coverage of salient information); they both ignore the rainbow in the picture. Automatic metrics, such as CIDEr, do not capture this failure.

In previous work, annotators (crowdworkers, typically) rate image captions directly (Hodosh et al., 2013), pairwise (Vedantam et al., 2015), or along multiple dimensions such as thoroughness (Aditya et al., 2015) and truthfulness (Yatskar et al., 2014). These scoring judgments depend highly on individual annotators' discretion and understanding of the annotation scheme (Freitag et al., 2021; Clark et al., 2021), making it difficult to decompose, interpret, and validate annotations. This lack of transparency also makes it difficult to interpret evaluation results for downstream applications where some aspects are particularly important (e.g., accessibility for people with visual impairments; Gleason et al., 2019, 2020). Further, these annotations were done only on relatively old models (e.g., MSCOCO leaderboard submissions in 2015; Anderson et al., 2016). Correlations of automatic metrics with human judgments can break down especially when model types change (Callison-Burch et al., 2006), or generation models become increasingly powerful (Ma et al., 2019; Edunov et al., 2020). We thus develop an up-to-date, transparent human evaluation protocol to better understand how current models perform and how automatic metrics are correlated when applied to current models.

At the core of our rubrics are two main scores in a tradeoff: *precision* and *recall* (Fig. 1). The former measures accuracy of the information in a caption, and the latter assesses how much of the salient information in the image is covered. We then penalize a caption if we find a problem in *fluency*, *conciseness*, or *inclusive language*. Two or more authors evaluate every instance and collaborate to resolve disagreements, ensuring high quality of the annotations. We assess outputs from four strong models as well as human-generated reference captions from MSCOCO. We call our scores THUMB 1.0 (**T**ransparent **Hum**an **B**enchmark), and release them publicly.[1] Our key findings from the evaluations are:

- Machine-generated captions from recent models have been claimed to achieve superhuman performance using popular automatic metrics (human performance is ranked at the 250th place in the MSCOCO leaderboard),[2] but they still show substantially lower quality than human-generated ones.
- Machines fall short of humans, especially in recall (Fig. 1), but most automatic metrics say the opposite.
- Human performance is underestimated in the current leaderboard paradigm, and there is still much room for improvement on MSCOCO captioning.
- CLIPScore and RefCLIPScore (Hessel et al., 2021), recently proposed metrics that use image features, improve correlations particularly in recall. While they fail to score human generation much higher than machine one, they capture an aspect that is less reflected in text-only metrics.
- Currently available strong captioning models generate highly fluent captions. Fluency evaluation is thus no longer crucial in ranking these models.

2 Evaluation Protocol

We establish a transparent evaluation protocol for image captioning models. Our rubrics and rules are developed through discussions among all annotators (first four authors of this paper).

2.1 Evaluation Setups and Quality Control

We used images from the test data in the standard *Karpathy split* (Karpathy and Fei-Fei, 2015) of the MSCOCO dataset (Lin et al., 2014). The dataset consists of 113K, 5K, and 5K train/dev./test everyday-scene photos sampled from Flickr. We randomly sampled 500 test images and prepared one human- and four machine-generated captions for every image (§2.3). We first performed developmental evaluations of 250 captions for 50 images and created rubrics. We then proceeded with the rest of the captions. For every image, captions were shuffled, and thus annotators did not know which caption corresponded to which model, thereby avoiding a potential bias from knowledge about the models. We conducted two-stage annotations: the first annotator scores all captions for given images, and the second annotator checks and modifies the scores when necessary. After the developmental phase, the κ coefficient (Cohen, 1960) was 0.86 in precision and 0.82 in recall for the rest of the evaluated captions (§2.2.1).[3] The first four authors of this paper conducted all evaluations; none of them are color blind or low vision, two are native English speakers, and one is a graduate student in linguistics. We finally ensured that at least one native speaker evaluated the fluency of every caption (§2.2.2), meaning that if a caption is annotated by the two non-native speakers, one native speaker checks the fluency in an additional round.

2.2 THUMB 1.0

We base our evaluations on two main scores (**precision** and **recall**) and three types of penalty (**fluency**, **conciseness**, and **inclusive language**) The overall score is computed by averaging precision and recall and deducting penalty points.

2.2.1 Main Scores

The two main scores are assessed in the scale of 1–5. They balance information accuracy and coverage. See §A.4 for score distribution histograms.

Precision Precision (P) measures how precise the caption is given the image. For instance, Caption 1-B in Table 1 is perfectly precise, while 1-A (*dog* vs. *otter*, *one* vs. *two frisbees*) and 1-C (*three* vs. *two*

[1] Anonymized.

[2] https://competitions.codalab.org/competitions/3221#results.

[3] Furthermore, we found that a third annotator did not change the scores for all 100 captions randomly sampled for meta-evaluations, confirming the sufficiently high quality of our two-stage annotations.

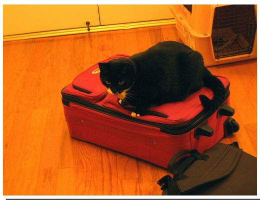
Image	Caption	P	R	Flu.	Total
	**1-A**: Up-Down ***A dog playing with a frisbee on the ground.***	3	4	0	3.5
	1-B: VinVL-base ***A otter is laying on the sand next to two frisbees.***	5	4	0.1	4.4
	1-C: VinVL-large ***A small animal laying on a rock with three frisbees.***	4	3	0	3.5
	**2-A**: Up-Down ***A close up of a plate of broccoli.***	5	3	0	4
	2-B: Unified-VLP, VinVL-base, VinVL-large ***A plate of pasta and broccoli on a table.***	4	4	0	4
	2-C: Human ***A multi colored dish with broccoli and white twisted pasta in it.***	5	5	0.1	4.9
	**3-A**: Unified-VLP ***A little girl holding a video game controller.***	3	4	0	3.5
	3-B: VinVL-large ***A little girl is blow drying her hair on a couch.***	4	5	0	4.5
	3-C: Human ***A little girl holding a blow dryer next to her head.***	5	5	0	5
	**4-A**: Up-Down ***A black cat laying in a red suitcase.***	3	5	0	4
	4-B: Unified-VLP, VinVL-base, VinVL-large ***A black cat sitting on top of a red suitcase.***	5	5	0	5
	4-C: Human ***A large black cat laying on top of a pink piece of luggage.***	4	5	0	4.5
	**5-A**: Up-Down, Unified-VLP ***A man standing in front of a display of donuts.***	3	2	0	2.5
	5-B: VinVL-large ***A woman standing behind a counter at a donut shop.***	5	3	0	4
	5-C: Human ***Woman selling doughnuts with doughnut stock in the background.***	5	5	0.3	4.7

Table 1: Example evaluations of machine- and human-generated captions. None of these captions get penalties in conciseness and inclusive language. Evaluated captioning models are described in §2.3

frisbees) are not precise. Precision guards against hallucinations from the language model (*table* in 2-B) that are known to be common failures of image captioning models (Rohrbach et al., 2018). The score of 4 is reserved for relatively minor issues, such as attributes that are almost correct (e.g., *pink* vs. *red* in 4-C, Table 1) or cases where the caption does not contradict with the image but is not guaranteed to be true (e.g., it is unclear whether the girl is sitting on a couch in 3-B). In addition to objects themselves, precision deals with information like properties, attributes, occasions, locations, and relations between objects (e.g., *in a red suitcase* vs. *on a red suitcase* in 4-A).

Recall Recall (R) measures how much of the salient information (e.g., objects, attributes, and relations) from the image is covered by the caption. This includes color (e.g., color of the frisbees in 1-A, 1-B, and 1-C) and guards against generic, uninformative captions. For instance, an otter is a *small animal*, and thus *small animal* is *precise* (1-C); however, it is much less informative than saying an otter. Similarly, Caption 5-B only says a woman is standing behind a counter at a donut shop, but she is selling donuts, not buying or looking at donuts, which is salient information from the picture. We do not take a point off if missing information is already expected from the caption (e.g., a double-decker bus is typically red). We often find it useful to take a generative approach when evaluating recall: *what image does the caption lead us to imagine?* When the caption entails many potential images that substantially diverge from the given image, the recall score should be low.

2.2.2 Penalties

Fluency Fluency (Flu.) measures the quality of captions as English text regardless of the given image. Initially, we scored fluency in the scale of 1–5, similar to P and R, but we found most captions from modern neural network models were highly fluent. Thus, we instead decided to take points off from the average of P and R if there's a fluency problem to account for minor issues that are much less problematic than losing one P/R point. The four annotators had extensive discussions and developed rubrics for fluency. Similar to recent work on professional evaluations for machine translation (Freitag et al., 2021), we evaluated under the following principle: if a fluency problem is expected to be easily corrected by a text postprocessing algorithm (e.g., grammatical error correction: Yuan and Briscoe, 2016; Sakaguchi et al., 2017), the penalty should be 0.1. This includes obvious misspellings or grammatical errors (e.g., *A otter* in 1-B) and missing determiners/hyphens (*multi colored* in 2-C). 0.5+ points were subtracted for more severe problems, such as duplication (e.g., *A display case of donuts and doughnuts*), ambiguity (e.g., *A cat is on a table with a cloth on **it***), and broken sentences (e.g., *A large concrete sign small buildings behind it.*). See Table 6 in §A.1 for more extensive fluency rubrics. Note that the average fluency penalty was 0.01; this confirms that fluency is no longer crucial in ranking models for MSCOCO captioning and contrasts with human evaluations previously done for older captioning models.

Conciseness The scores so far do not take into account conciseness of captions. Specifically, a model could simply increase all scores by describing every detail in a picture. For instance, the following caption is overly repetitive: *a woman lying on her back with knees bent on a beach towel under a multicolored, striped beach umbrella, surrounded by sand, and with clear blue sky above*. We subtract 0.5 points for these captions. Note that most machine captions were short, and this penalty was only applied to two human-generated captions. It might become more crucial for future models with a more powerful object detection module that catches many objects in the picture.

Inclusive Language We found that some instances substantially diverge from inclusive language, raising a concern for downstream applications. In these cases, we added a penalty: 0.5 points were deducted for a subjective comment about appearance (e.g., *very pretty girl*), and 2 points for more severe problems (e.g., *beautiful breasts*).

2.2.3 Rules of THumb

In our development phase, we established the following additional rules to clarify our annotation scheme.

Avoiding Double Penalties When an error is accounted for in precision, we correct the error before scoring the recall, thereby avoiding penalizing the precision and recall for the same mistake. For example, P=3 is given to Caption 1-A in Table 1 because of its wrong detection (*dog* vs. *otter*; *one* vs. *two frisbees*), but we score the recall assuming that the caption is now *an **otter** playing with two frisbees on the ground*. This ensures that a generic, useless caption, such as *there is something on something* (P=5, R=1), would be ranked considerably lower than *a dog on the beach with two pink and yellow frisbees* (P=3, R=5). Similarly, the wrong detection in 5-A (*man* vs. *woman*) is handled only in precision. Note that such error correction is not applicable to hallucinations because there is no alignment between a part of the image and a hallucinated object (e.g., *table* in 2-B). This rule departs from the definition of recall in SPICE (Anderson et al., 2016), an automatic metric that measures the F_1 score in scene graphs predicted from reference and generated captions; their alignment is limited to WordNet synonyms (Miller, 1995). This means that classifying an otter as a dog or even a small animal would result in cascading errors both in precision and recall, overrating captions that completely overlook the otter or ones that make a more severe classification error (e.g., miscategorize the otter as a car, compared to a dog).

Object Counts as Attributes All counts are considered as object attributes, and wrong counts are handled in precision. This simplifies the distinction between precision and recall. For instance, both *a frisbee* (1-A) and *three frisbees* (1-C) are precision problems, while saying *some frisbees* would be a recall problem when it is clear that there are exactly two frisbees. Note that this is in line with SPICE, which treats object counts as attributes in a scene graph, rather than duplicating a scene graph for every instance of an object (Anderson et al., 2016).

Black and White Photo MSCOCO contains black and white or gray-scale pictures. Some captions explicitly mention that they are black and white, but we disregard this difference in our evaluations. The crowdsource instructions for creating reference captions do not specify such cases (Chen et al., 2015). Further, we can potentially run post-processing to determine whether it is black and white to modify the caption accordingly, depending on the downstream usage.

Text Processing Image captioning models often differ slightly in text preprocessing. As a result, we found that generated captions were sometimes slightly different in format (e.g., tokenized or detokenized; lowercased or not). For better reproducibility, we follow the spirit of SACREBLEU (Post, 2018), which has become the standard package to compute BLEU scores for machine translation: all evaluations, including automatic metrics, should be done on clean, untokenized text, independently of preprocessing design choices. We apply the following minimal postprocessing to the model outputs and human captions.

- Remove unnecessary spaces at the start or end of every caption.
- Uppercase the first letter.
- Add a period at the end if it doesn't exist, and remove a space before a period if any.

We keep the postprocessing minimal for this work and encourage future model developers to follow the standard practice in machine translation: every model has to output clean, truecased, untokenized text that is ready to be used in downstream modules. This also improves the transparency and reproducibility of automated evaluations (Post, 2018).

2.3 Evaluated Captions

We evaluated the following four strong models from the literature as well as human-generated captions. They share similar pipeline structure: object detection followed by crossmodal caption generation. They vary in model architecture, (pre)training data, model size, and (pre)training objective. Evaluating captions from them will enable us to better understand what has been improved and what is still left to future captioning models.

- **Up-Down** (Anderson et al., 2018) trains Faster R-CNN (Ren et al., 2015) on the Visual Genome datset (Krishna et al., 2016) for object detection. It then uses an LSTM-based crossmodal generation model.
- **Unified-VLP** (Zhou et al., 2020) uses the same object detection model as Up-Down. The transformer-based generation model is initialized with base-sized BERT (Devlin et al., 2019) and further pretrained with 3M images from Conceptual Captions (Sharma et al., 2018).
- **VinVL-base** and **VinVL-large** (Zhang et al., 2021) train a larger-scale object detection model with the ResNeXt-152 C4 architecture (Xie et al., 2017) on ImageNet (Deng et al., 2009). The transformer generation model is initialized with BERT and pretrained with 5.7M images.
- **Human** randomly selects one from the five human-generated reference captions in MSCOCO. Those captions were created by crowdworkers on Amazon Mechanical Turk (Chen et al., 2015).

Further details are described in §A.3 of Appendix.

3 Results and Analysis

We present results and analysis from our evaluations. Our transparent evaluations facilitate assessments and analysis of both captioning models (§3.1) and automatic metrics (§3.2).

3.1 Comparing Models

Seen in Table 2 (left section) is the model performance that is averaged over the 500 test images and broken down by the rubric categories. Overall, Human substantially outperforms all machines in the P, R, and total scores. In particular, we see a large gap between Human and the machines in recall (e.g., Human 4.35 vs. VinVL-large 3.97). This contrasts with the automatic metric-based ranking of the MSCOCO leaderboard, where Human is ranked at the 250th place.[4] This result questions claims about human parity or superhuman performance on MSCOCO image captioning. The four machine captioning models are ranked in the expected order, though the small difference between VinVL-large and VinVL-base suggests that simply scaling up models would not lead to a substantial improvement. We see that the three models that are initialized with pretrained BERT (VinVL-large/base, Unified-VLP) are particularly fluent, but the problem is small in the other models as well.

While we compute representative, total scores, our transparent rubrics allow for adjusting weighting of the categories depending on the application of interest. For instance, in the social media domain, recall can be more important than precision to make captions engaging to users (Shuster et al., 2019). To assess the models indepen-

[4] The official leaderboard ranks submissions using CIDEr (Vedantam et al., 2015) with 40 references on the hidden test data. We use the public Karpathy test split instead, but we suspect the same pattern would hold on the hidden data as well, given the large gap between machines and Human.

	THUMB 1.0						Automatic Metrics						
Model	P↑	R↑	Flu.↓	Con.↓	Inc.↓	Total↑	BLEU	ROUGE	BERT-S	SPICE	CIDEr	CLIP-S	RefCLIP-S
Human	4.82	4.35	0.019	0.02	0.00	$\mathbf{4.56}^{+0.03}_{-0.03}$	26.2	50.4	0.938	23.7	111.5	0.791	0.834
VinVL-large	4.54	3.97	0.005	0.00	0.00	$4.25^{+0.04}_{-0.04}$	33.3	56.5	0.946	26.4	141.8	0.784	0.834
VinVL-base	4.47	3.95	0.001	0.00	0.00	$4.21^{+0.04}_{-0.04}$	32.3	55.9	0.945	25.6	138.4	0.779	0.830
Unified-VLP	4.35	3.77	0.004	0.00	0.00	$4.06^{+0.04}_{-0.04}$	31.6	55.8	0.945	24.3	128.5	0.771	0.821
Up-Down	4.29	3.50	0.014	0.00	0.00	$3.88^{+0.05}_{-0.05}$	28.4	52.2	0.939	21.0	110.7	0.746	0.803

Table 2: Performance of image captioning models with respect to THUMB 1.0 (left) and automatic metrics (right). All scores are averaged over 500 images randomly sampled from the Karpathy test split. P: precision; R: recall; Flu.: fluency; Con.: conciseness; Inc.: inclusive language. 90% confidence intervals for total scores are calculated by bootstrapping (Koehn, 2004). All reference-based metrics take as input the same four crowdsourced captions that are not used in Human for fair comparisons.

dently of these aggregation decisions, we count the number of times when each model outperforms/underperforms all the others both in P and R (*strictly* best/worst, Table 3). We see patterns consistent with Table 2. For example, Human is most likely to be strictly best and least likely to be strictly worst. This suggests that machine captioning models would still fall short of crowdworkers in a wide range of downstream scenarios.

Model	Human	Vin-large	Vin-base	U-VLP	Up-Down
# Best ↑	**327**	180	161	112	74
# Worst ↓	**65**	128	150	190	269

Table 3: # times when each captioning model is *strictly* best/worst in the caption set (i.e., best/worst both in precision and recall).

	w/o Human			w/ Human		
Metric	P	R	Total	P	R	Total
RefCLIP-S	0.34	**0.27**	**0.44**	0.31	0.26	$\mathbf{0.41}^{+0.05}_{-0.05}$
RefOnlyC	**0.42**	0.14	0.41	**0.37**	0.11	$0.34^{+0.04}_{-0.05}$
CLIP-S	0.18	**0.27**	0.32	0.17	**0.28**	$0.32^{+0.05}_{-0.05}$
CIDEr	0.27	0.18	0.33	0.21	0.11	$0.23^{+0.04}_{-0.04}$
BERT-S	0.27	0.18	0.33	0.20	0.10	$0.21^{+0.04}_{-0.04}$
SPICE	0.26	0.15	0.30	0.20	0.09	$0.21^{+0.04}_{-0.04}$
ROUGE-L	0.26	0.17	0.31	0.18	0.07	$0.18^{+0.04}_{-0.04}$
BLEU	0.21	0.13	0.25	0.15	0.04	$0.13^{+0.04}_{-0.04}$

Table 4: Instance-level correlations of automatic evaluation scores. RefCLIP-S and CLIP-S use image features unlike the others, and all but CLIP-S require references. All of these reference-based metrics use the same subset of four captions as in Table 2 that exclude Human. All metrics had correlations lower than 0.1 for fluency.

3.2 Comparing Automatic Metrics

While carefully-designed human judgments like ours should be considered more reliable, automatic metrics allow for faster development cycles. Our transparent evaluations can also be used to analyze how these automatic metrics correlate with different aspects of image captioning. Table 2 (right section) shows automatic scores of the captioning models over 7 popular metrics for image captioning. CLIP-S (Hessel et al., 2021) is a *referenceless* metric that uses image features from CLIP (Radford et al., 2021), a crossmodal retrieval model trained on 400M image-caption pairs from the web. RefCLIP-S augments CLIP-S with similarities between the generated and reference captions. All other metrics, such as SPICE (Anderson et al., 2016) and CIDEr (Vedantam et al., 2015), only use reference captions without image features.

These automatic metrics generally agree with our evaluations in ranking the four machines, but completely disagree in the assessment of Human. Most metrics rank Human near the bottom, showing that they are not reliable in evaluating high-quality, human-generated captions. The two metrics with powerful image and text features (CLIP-S and RefCLIP-S) give high scores to Human compared to the other metrics, but they still fail to score Human substantially higher than VinVL-large. This suggests that automatic metrics should be regularly updated as our models become stronger (and perhaps more similar to humans), and raises a significant concern about the current practice that fixes evaluation metrics over time.

Seen in Table 4 are instance-level Pearson correlation scores between automatic scores and our evaluations.[5] We also add an ablation study: RefOnlyC removes image features from RefCLIP-S to

[5]Instance-level Pearson correlations with human judgments were often computed in prior work to compare automatic metrics for image captioning (e.g., Hessel et al., 2021). An alternative is system-level correlations, but they would be uninformative with five systems only.

quantify the effect of image features. We consider two types of scenarios: one *with* Human and one *without*. Correlations drop from the latter to the former for all metrics and aspects except CLIP-S, again showing that the metrics are not reliable in assessing human-generated captions. Interestingly, CLIP-S correlates best in recall (0.28 w/ Human) but suffers in precision (0.17 w/ Human). RefOnlyC, in contrast, achieves the best correlations in P at the expense of R. RefCLIP-S balances the two and achieves the best correlation in total scores. This indicates that the CLIP image features particularly help assess coverage of salient information that can be ignored in some reference captions from crowdworkers.[6] Prior work (Hessel et al., 2021) found that SPICE can still improve correlations when combined with CLIP-S, even though CLIP-S better correlates with human judgments than SPICE. This implies that image-based and reference-only metrics capture different aspects of image captioning. Our analysis indeed agrees with their finding and, further, identifies that recall is one such aspect. For an extensive description of these metrics and their configurations, see §A.2 of Appendix.

3.3 Machine vs. Human Examples

Table 5 provides examples that contrast machine- and human-generated captions. We see that machine-generated captions ignore salient information or make critical errors for these images. These problems often occur in relatively rare cases: a tennis player is showing excitement rather than hitting a ball; a bride and groom are cutting a wedding cake; a boy is wearing a tie without a shirt; a man is putting clothing and a tie on a dummy instead of a person. But these situations are exactly the most important information because of their *atypicality* (Feinglass and Yang, 2021). This illustrates fundamental problems of current image captioning models that are left to future work.

4 Related Work

Human Evaluations for Image Captioning Several prior works conducted human evaluations for image captioning with varying models, datasets, and annotation schemes. Much work used crowdworkers from Amazon Mechanical Turk on Flickr-based datasets, including the PASCAL (Rashtchian et al., 2010), Flickr8k/30k (Hodosh et al., 2013; Young et al., 2014), and MSCOCO datasets. Annotators scored the overall quality directly (Kulkarni et al., 2011; Hodosh et al., 2013), pairwise (Vedantam et al., 2015), or along multiple dimensions, such as truthfulness/correctness (Yatskar et al., 2014; Anderson et al., 2016), thoroughness (Aditya et al., 2015), relevance (Yang et al., 2011; Li et al., 2011), and grammaticality/readability (Mitchell et al., 2012; Elliott and Keller, 2013). There are similarities between our rubrics and previous annotations, but our framework defines every dimension in a decomposable way through discussions among all annotators, while focusing on outputs from strong models currently available. Apart from these conventional Flickr-based datasets, some other work evaluated image captions for social media (engagingness, Shuster et al., 2019; accessibility for Twitter users with vision impairments, Gleason et al., 2019, 2020) and news articles (Biten et al., 2019). Our transparent evaluations would enable us to adjust the aggregation method based on the nature of downstream applications. More specializing categories can be added for these applications in later versions (e.g., THUMB 2.0).

Human Evaluations for Other Generation Tasks Much previous work explored human evaluations for other language generation tasks than image captioning. The WMT shared task (Barrault et al., 2020) conducts human evaluations of state-of-the-art machine translation systems every year; participants or crowdworkers directly rate a translation in a 100-point scale, which is a method developed by Graham et al. (2013, 2014, 2017). GENIE takes a similar approach but hosts human evaluations in leaderboards for machine translation, summarization, and commonsense reasoning (Khashabi et al., 2021). Kryscinski et al. (2019) and Fabbri et al. (2021) assessed many summarization models in a similar annotation scheme to the DUC 2006/2007 evaluations (Dang, 2006). Our transparent evaluation framework is inspired by rubric-based machine translation judgments by professional translators (Freitag et al., 2021), which resulted in different system rankings than the WMT evaluations. As top-performing models and automatic metrics are becoming increasingly similar across various natu-

[6] The low recall correlations of reference-only metrics can be partly because the maximum (as opposed to minimum or average) is typically taken over multiple reference captions (e.g., BERTScore, Zhang et al., 2020). Nevertheless, this alone does not explain the recall gap from image-based metrics because RefCLIP-S also takes the maximum score over all references. Future work can explore the relation between precision/recall and different treatments of multiple references.

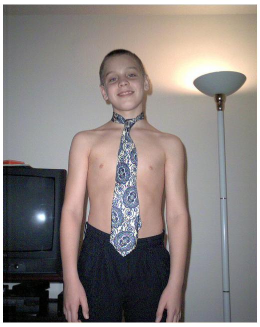
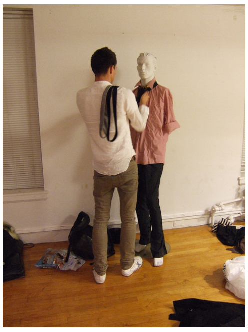
Image	Caption	P	R	Flu.	Total
	**6-A**: Up-Down ***A man holding a tennis racquet on a tennis court.***	5	3	0	4
	6-B: Unified-VLP, VinVL-base, VinVL-large ***A man holding a tennis racket on a tennis court.***	5	3	0	4
	6-C: Human ***A tennis player shows controlled excitement while a crowd watches.***	5	5	0	5
	**7-A**: Up-Down ***A person cutting a cake with a knife.***	3	3	0	3
	7-B: Unified-VLP ***A person cutting a wedding cake with a knife.***	3	5	0	4
	7-C: VinVL-base ***A couple of cakes on a table with a knife.***	5	3	0	4
	7-D: VinVL-large ***A woman cutting a cake with a knife.***	3	3	0	3
	7-E: Human ***Bride and grooms arms cutting the wedding cake with fruit on top.***	5	5	0.1	4.9
	**8-A**: Up-Down ***A young boy wearing a blue shirt and a blue tie.***	3	3	0	3
	8-B: Unified-VLP ***A young boy wearing a shirt and a tie.***	3	3	0	3
	8-C: VinVL-base ***A young boy wearing a tie standing in front of a lamp.***	5	3	0	4
	8-D: VinVL-large ***A young man wearing a tie and a shirt.***	3	3	0	3
	8-E: Human ***A man wearing only a tie standing next to a lamp.***	4	5	0	4.5
	**9-A**: Up-Down ***A couple of men standing next to each other.***	2	2	0	2
	9-B: Unified-VL ***Two men standing in a room.***	2	2	0	2
	9-C: VinVL-base ***A couple of men standing in a room.***	2	2	0	2
	9-D: VinVL-large ***Two men standing next to each other in a room.***	2	2	0	2
	9-E: Human ***A man standing next to a dummy wearing clothes.***	5	3	0	4

Table 5: Examples that contrast machine- and human-generated captions. All machine-generated captions overlook or misinterpret salient information: the excitement the tennis player expresses, the bride and groom cutting a wedding cake, the boy not wearing a shirt, and the man putting a tie on a dummy. None of these captions are penalized for conciseness or inclusive language. See §A.5 in Appendix for more examples.

ral language generation tasks, our findings on image captioning may be useful for other generation tasks as well.

5 Conclusion

We developed THUMB 1.0, transparent evaluations for the MSCOCO image captioning task. We refined our rubrics through extensive discussions among all annotators, and ensured the high quality by two-stage annotations. Our evaluations demonstrated critical limitations of current image captioning models and automatic metrics. While recent image-based metrics show promising improvements, they are still unreliable in assessing high-quality captions from crowdworkers. We hope that our annotation data will help future development of better captioning models and automatic metrics, and this work will become a basis for transparent human evaluations for the image captioning task and beyond.

References

Somak Aditya, Yezhou Yang, Chitta Baral, Cornelia Fermüller, and Yiannis Aloimonos. 2015. From images to sentences through scene description graphs using commonsense reasoning and knowledge.

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2016. SPICE: semantic propositional image caption evaluation. In *Proc. of ECCV*.

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *Proc. of CVPR*.

Loïc Barrault, Magdalena Biesialska, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Matthias Huck, Eric Joanis, Tom Kocmi, Philipp Koehn, Chi-kiu Lo, Nikola Ljubešić, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Santanu Pal, Matt Post, and Marcos Zampieri. 2020. Findings of the 2020 conference on machine translation (WMT20). In *Proc. of WMT*.

Ali Furkan Biten, Lluís Gómez, Marçal Rusiñol, and Dimosthenis Karatzas. 2019. Good news, everyone! context driven entity-aware captioning for news images. In *Proc. of CVPR*.

Chris Callison-Burch, Miles Osborne, and Philipp Koehn. 2006. Re-evaluating the role of BLEU in machine translation research. In *Proc. of EACL*.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C. Lawrence Zitnick. 2015. Microsoft COCO captions: Data collection and evaluation server.

Elizabeth Clark, Tal August, Sofia Serrano, Nikita Haduong, Suchin Gururangan, and Noah A. Smith. 2021. All that's 'human' is not gold: Evaluating human evaluation of generated text. In *Proc. of ACL*.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*.

Hoa Trang Dang. 2006. Overview of DUC 2006. In *Proc. of DUC*.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: A large-scale hierarchical image database. In *Proc. of CVPR*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proc. of NAACL*.

Sergey Edunov, Myle Ott, Marc'Aurelio Ranzato, and Michael Auli. 2020. On the evaluation of machine translation systems trained with back-translation. In *Proc. of ACL*.

Desmond Elliott and Frank Keller. 2013. Image description using visual dependency representations. In *Proc. of EMNLP*.

Desmond Elliott and Frank Keller. 2014. Comparing automatic evaluation measures for image description. In *Proc. of ACL*.

Alexander R Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021. SummEval: Re-evaluating summarization evaluation. *TACL*.

Joshua Feinglass and Yezhou Yang. 2021. SMURF: SeMantic and linguistic UndeRstanding fusion for caption evaluation via typicality analysis. In *Proc. of ACL*.

Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021. Experts, errors, and context: A large-scale study of human evaluation for machine translation.

Cole Gleason, Patrick Carrington, Cameron Tyler Cassidy, Meredith Ringel Morris, Kris M. Kitani, and Jeffrey P. Bigham. 2019. "It's almost like they're trying to hide it": How user-provided image descriptions have failed to make Twitter accessible. In *Proc. of WWW*.

Cole Gleason, Amy Pavel, Emma McCamey, Christina Low, Patrick Carrington, Kris M. Kitani, and Jeffrey P. Bigham. 2020. Twitter a11y: A browser extension to make twitter images accessible. In *Proc. of CHI*.

Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2013. Continuous measurement scales in human evaluation of machine translation. In *Proc. of LAW*.

Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2014. Is machine translation getting better over time? In *Proc. of EACL*.

Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2017. Can machine translation systems be evaluated by the crowd alone. *Natural Language Engineering*.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2021. CLIPScore: A reference-free evaluation metric for image captioning. In *Proc. of EMNLP*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*.

Micah Hodosh, Peter Young, and Julia Hockenmaier. 2013. Framing image description as a ranking task: Data, models and evaluation metrics. *JAIR*.

Andrej Karpathy and Li Fei-Fei. 2015. Deep visual-semantic alignments for generating image descriptions. In *Proc. of CVPR*.

Daniel Khashabi, Gabriel Stanovsky, Jonathan Bragg, Nicholas Lourie, Jungo Kasai, Yejin Choi, Noah A. Smith, and Daniel S. Weld. 2021. GENIE: A leaderboard for human-in-the-loop evaluation of text generation.

Mert Kilickaya, Aykut Erdem, Nazli Ikizler-Cinbis, and Erkut Erdem. 2017. Re-evaluating automatic metrics for image captioning. In *Proc. of EACL*.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proc. of EMNLP*.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalanditis, Li-Jia Li, David A Shamma, Michael Bernstein, and Li Fei-Fei. 2016. Visual Genome: Connecting language and vision using crowdsourced dense image annotations. *IJCV*.

Wojciech Kryscinski, Nitish Shirish Keskar, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Neural text summarization: A critical evaluation. In *Proc. of EMNLP*.

Girish Kulkarni, Visruth Premraj, Sagnik Dhar, Siming Li, Yejin Choi, Alexander C Berg, and Tamara L Berg. 2011. Baby talk: Understanding and generating simple image descriptions. In *Proc. of CVPR*.

Siming Li, Girish Kulkarni, Tamara L Berg, Alexander C Berg, and Yejin Choi. 2011. Composing simple image descriptions using web-scale n-grams. In *Proc. of CoNLL*.

Xiujun Li, Xi Yin, Chunyuan Li, Xiaowei Hu, Pengchuan Zhang, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. 2020. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *Proc. of ECCV*.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Proc. of Text Summarization Branches Out*.

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: common objects in context. In *Proc. of ECCV*.

Qingsong Ma, Johnny Wei, Ondřej Bojar, and Yvette Graham. 2019. Results of the WMT19 metrics shared task: Segment-level and strong MT systems pose big challenges. In *Proc. of WMT*.

George A. Miller. 1995. WordNet: A lexical database for English.

Margaret Mitchell, Jesse Dodge, Amit Goyal, Kota Yamaguchi, Karl Stratos, Xufeng Han, Alyssa Mensch, Alex Berg, Tamara Berg, and Hal Daumé III. 2012. Midge: Generating image descriptions from computer vision detections. In *Proc. of EACL*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proc. of ACL*.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proc. of WMT*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision.

Cyrus Rashtchian, Peter Young, Micah Hodosh, and Julia Hockenmaier. 2010. Collecting image annotations using Amazon's Mechanical Turk. In *Proc. of NAACL Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster R-CNN: Towards real-time object detection with region proposal networks. In *Proc. of NeurIPS*.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. 2018. Object hallucination in image captioning. In *Proc. of EMNLP*.

Keisuke Sakaguchi, Matt Post, and Benjamin Van Durme. 2017. Grammatical error correction with neural reinforcement learning. In *Proc. of IJCNLP*.

Sebastian Schuster, Ranjay Krishna, Angel Chang, Li Fei-Fei, and Christopher D. Manning. 2015. Generating semantically precise scene graphs from textual descriptions for improved image retrieval. In *Proc. of VL*.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual Captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proc. of ACL*.

Kurt Shuster, Samuel Humeau, Hexiang Hu, Antoine Bordes, and Jason Weston. 2019. Engaging image captioning via personality. In *Proc. of CVPR*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proc. of NeurIPS*.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2015. CIDEr: Consensus-based image description evaluation. In *Proc. of CVPR*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. HuggingFace's transformers: State-of-the-art natural language processing. In *Proc. of EMNLP: System Demonstrations*.

Saining Xie, Ross B. Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. 2017. Aggregated residual transformations for deep neural networks. In *Proc. of CVPR*.

Yezhou Yang, Ching Teo, Hal Daumé III, and Yiannis Aloimonos. 2011. Corpus-guided sentence generation of natural images. In *Proc. of EMNLP*.

Mark Yatskar, Michel Galley, Lucy Vanderwende, and Luke Zettlemoyer. 2014. See no evil, say no evil: Description generation from densely labeled images. In *Proc. of *SEM*.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *TACL*.

Zheng Yuan and Ted Briscoe. 2016. Grammatical error correction using neural machine translation. In *Proc. of NAACL*.

Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. 2021. VinVL: Making visual representations matter in vision-language models. In *Proc. of CVPR 2021*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating text generation with BERT. In *Proc. of ICLR*.

Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason J. Corso, and Jianfeng Gao. 2020. Unified vision-language pre-training for image captioning and VQA. In *Proc. of AAAI*.

A Appendix

A.1 Fluency Rubrics

Table 6 presents our fluency rubrics. They were developed by the first four authors (two of whom were native English speakers, and one was a graduate student in linguistics). Generally, if a fluency problem is expected to be easily corrected by a text postprocessing algorithm, the penalty is 0.1. More severe errors (e.g., broken sentence and ambiguity) are penalized more.

A.2 Automatic Metrics

Here we discuss details and configurations of the automatic metrics used in §3.2. CLIPScore and RefCLIPScore use image features from CLIP (Radford et al., 2021), a crossmodal retrieval model trained on 400M image-caption pairs from the web. All the other five metrics only use reference captions.

BLEU BLEU (Papineni et al., 2002) is a precision-oriented metric and measures n-gram overlap between the generated and reference captions. We use the SacreBLEU implementation of BLEU-4 and get sentence-level scores (Post, 2018).[7]

ROUGE ROUGE (Lin, 2004) measures the number of overlapping n-grams between the generated and reference captions. We use the HuggingFace implementation of ROUGE-L (Wolf et al., 2020).

CIDEr CIDEr (Vedantam et al., 2015) measures the cosine similarity between the n-gram counts of the generated and reference captions with TF-IDF weighting. We use the implementation from the pycocoevalcap package.[8]

SPICE SPICE (Anderson et al., 2016) predicts scene graphs from the generated and reference captions using the Stanford scene graph parser (Schuster et al., 2015). It then measures the F_1 score between scene graphs from the generated and reference captions. WordNet Synsets are used to cluster synonyms (Miller, 1995). We again use the implementation from the pycocoevalcap package.

BERTScore BERTScore (Zhang et al., 2020) aligns tokens between the generated and reference captions using contextual word representations from BERT (Devlin et al., 2019). We use

[7] `https://github.com/mjpost/sacreBLEU/blob/v1.2.12/sacrebleu.py#L999`.

[8] `https://github.com/salaniz/pycocoevalcap`.

Fluency Error Type	Penalty	Example
Obvious spelling error, one vs. two words	0.1	*cel phone, surf board*
Grammatical error that can be easily fixed	0.1	*a otter*
Casing issue	0.1	*tv, christmas*
Hyphenation	0.1	*horse drawn carriage*
Interpretable but unnatural wording	0.1	*double decked bus*
Non-trivial punctuation	0.2	*A bird standing in the wooded area with leaves all around.*
Misleading spelling error	0.5	*A good stands in the grass next to the water.* (good→goose)
Duplication	0.5	*A display case of donuts and doughnuts.*
Ambiguity	0.5	*A cat is on a table with a cloth on it.*
Awkward construction	0.1–0.5	*There is a freshly made pizza out of the oven.*
Broken sentence	0.5+	*A large concrete sign small buildings behind it.*

Table 6: Fluency penalty rubrics.

the HuggingFace implementation and compute the F_1 score. As in Zhang et al. (2020), we take the maximum score over all reference captions.

CLIPScore CLIPScore (Hessel et al., 2021) is the only *referenceless* metric out of the 7 metrics. It measures the cosine similarity between the generated caption and given image using the representations from CLIP. It is shown to correlate better with human judgments from prior work, compared to previous reference-based metrics (Hessel et al., 2021). We use the official implementation by the authors.[9]

RefCLIPScore RefCLIPScore augments CLIPScore with the maximum similarity between the generated and reference captions. We again use the official implementation.

A.3 Evaluated Captions

We evaluated the following four strong models from the literature as well as human-generated captions. They share similar pipeline structure but vary in model architecture, (pre)training data, model size, and (pre)training objective. Evaluating captions from them will enable us to better understand what has been improved and what is still left to future captioning models.

Up-Down The bottom-up and top-down attention model (Up-Down, Anderson et al., 2018) performs pipelined image captioning: *object detection* that finds objects and their corresponding image regions and *crossmodal generation* that predicts a caption based on the features from object detection. The bottom-up attention finds salient image regions during object detection, and the top-down one attends to these regions during crossmodal generation. Up-Down uses Faster R-CNN (Ren et al., 2015) and LSTMs (Hochreiter and Schmidhuber, 1997) for object detection and crossmodal generation respectively. Faster R-CNN is trained with the Visual Genome dataset (Krishna et al., 2016), and the crossmodal generation model is trained on the MSCOCO dataset. We generate captions for the test data with a model optimized with crossentropy.[10]

Unified-VLP Unified-VLP (Zhou et al., 2020) also runs a pipeline of object detection and crossmodal generation. Faster R-CNN and the transformer architecture (Vaswani et al., 2017) are used for object detection and crossmodal generation respectively. Similar to Up-Down, the Faster R-CNN object detection model is trained with the Visual Genome dataset. The transformer generation model, on the other hand, is initialized with base-sized BERT (Devlin et al., 2019) and pretrained on the Conceptual Captions dataset (3M images, Sharma et al., 2018) with the masked and left-to-right language modeling objectives for the captions. The crossmodal generation model is then finetuned on the MSCOCO dataset. We apply beam search of size 5 to the model with CIDEr optimization.

VinVL-base, VinVL-large VinVL with Oscar (Li et al., 2020; Zhang et al., 2021) performs a similar pipeline of object detection, followed by crossmodal generation. The crossmodal model is initialized with BERT (Devlin et al., 2019) as in Unified-VLP but uses detected object tags to encourage alignments between image features and word representations. The object detection model

[9]https://github.com/jmhessel/pycocoevalcap.

[10]https://vision-explorer.allenai.org/image_captioning.

with the ResNeXt-152 C4 architecture (Xie et al., 2017) is pretrained with ImageNet (Deng et al., 2009) and trained on 2.5M images from various datasets. The transformer-based crossmodal generator is initialized with BERT, pretrained with 5.7M images, and finetuned for MSCOCO captioning. We use VinVL-base and VinVL-large that are both finetuned with CIDEr optimization[11] and generate captions with beam search of size 5.

Human In addition to machine-generated captions from the four models, we assessed the quality of human-generated reference captions from MSCOCO. This will allow us to understand the performance gap between machines and humans, as well as the quality of crowdsourced captions. Human-generated captions were created using Amazon Mechanical Turk (Chen et al., 2015). Crowdworkers were only given the following instructions (Chen et al., 2015):

- 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.

Every image has five human-generated captions, and we randomly selected one for each to evaluate. We found, however, a non-negligible number of noisy captions in the MSCCOCO dataset from annotation spammers. We often find subjective adjectives (e.g., *very nice/clean/cute*) or words that diverge from *inclusive language* in reference captions, probably because crowdworkers increased the number of words in captions effortlessly (see the last instruction item that says captions have to have 8+ words). To better estimate the performance of a human that invests reasonable effort into the captioning task, we resampled a caption for 13% of the test images, which would have been given a total score lower than 4.0.

A.4 Score Distributions

Seen in Fig. 2 are distributions of precision and recall scores for human and machine-generated captions. We see that the precision distribution looks similar between Human and machines, but not recall. This provides further support for our claim that current machines fall short of humans particularly in recall.

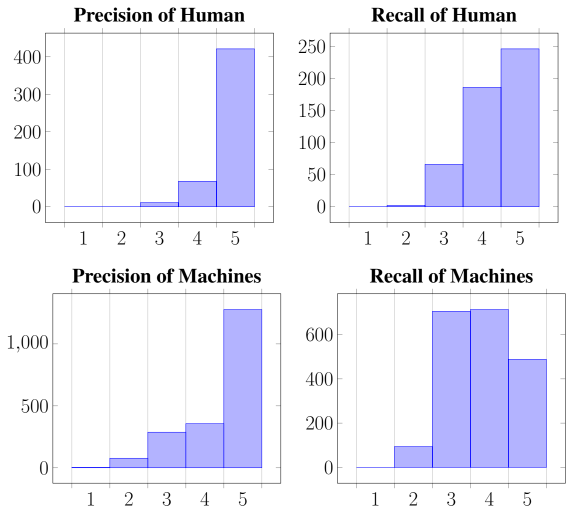


Figure 2: Precision/recall histograms for human- and machine-generated captions.

A.5 Additional Machine vs. Human Examples

Table 7 provides an additional example that contrasts machine- and human-generated captions. All machines generate generic captions and ignore the most important information that a traditional Thanksgiving dinner is being served on the table.

[11] https://github.com/microsoft/Oscar/blob/master/VinVL_MODEL_ZOO.md#Image-Captioning-on-COCO.

Image	Caption	P	R	Total
	**10-A**: Up-Down ***A table that has some food on it.***	5	2	3.5
	10-B: Unified-VLP ***A table with plates of food on a table.***	5	2	3.5
	10-C: VinVL-base ***A red table topped with plates of food and bowls of food.***	4	2	3
	10-D: VinVL-large ***A table with a turkey and other food on it.***	5	3	4
	10-E: Human ***A table set for a traditional Thanksgiving dinner.***	5	5	5

Table 7: Additional example that contrasts machine- and human-generated captions. Similar to Table 5, machine-generated captions ignore the most salient information: Thanksgiving dinner. None of these captions are penalized for fluency, conciseness, or inclusive language.