Towards Unique and Informative Captioning of Images

Zeyu Wang[1 [0000−0002−7057−1613]], Berthy Feng[1,2 [0000−0002−1843−2165]], Karthik Narasimhan[1 [0000−0001−9894−9983]], and Olga Russakovsky[1 [0000−0001−5272−3241]]

[1] Princeton University
{zeyuwang,karthikn,olgarus}@cs.princeton.edu
[2] California Institute of Technology
bfeng@caltech.edu

Abstract. Despite considerable progress, state of the art image captioning models produce generic captions, leaving out important image details. Furthermore, these systems may even misrepresent the image in order to produce a simpler caption consisting of common concepts. In this paper, we first analyze both modern captioning systems and evaluation metrics through empirical experiments to quantify these phenomena. We find that modern captioning systems return higher likelihoods for incorrect distractor sentences compared to ground truth captions, and that evaluation metrics like SPICE can be 'topped' using simple captioning systems relying on object detectors. Inspired by these observations, we design a new metric (SPICE-U) by introducing a notion of *uniqueness* over the concepts generated in a caption. We show that SPICE-U is better correlated with human judgements compared to SPICE, and effectively captures notions of diversity and descriptiveness. Finally, we also demonstrate a general technique to improve any existing captioning model – by using mutual information as a re-ranking objective during decoding. Empirically, this results in more unique and informative captions, and improves three different state-of-the-art models on SPICE-U as well as average score over existing metrics.[3]

1 Introduction

Over the last few years, there has been considerable progress in image captioning, with current methods producing fluent captions for a variety of images [41, 45, 48, 26, 42, 2, 8]. However, all these systems tend to produce generic captions, re-using a small set of common concepts to describe vastly different images. Consider the example caption in Figure 1, produced by a state of the art model [8]. Despite obvious differences between the sixteen images, the model produces the same caption, missing several other details specific to certain images and generating incorrect facts about others. A human, on the other hand, would identify unique aspects of each image, such as whether the person is serving, is it a match or a practice, the type of tennis court, the color of the person's shirt, etc. While the

[3] Code is available at https://github.com/princetonvisualai/SPICE-U.

Fig. 1. Diverse images from the COCO validation set for which a trained captioning system [8] generates the same caption: "A man holding a tennis racket on a tennis court". The caption misses important details, such as the action of the person, the type of tennis court, whether there is audience, etc.

inadequacies of the captioning models can be partially attributed to the "mode collapse" problem of current techniques and loss functions like cross-entropy, the issue is more fundamental — defining and benchmarking image captioning adequately remains a challenging task.

To this end, we investigate modern captioning systems in terms of their ability to produce *unique and complete* captions. Specifically, we find that the problem of producing common concepts is deeply ingrained in modern captioning systems. As we demonstrate empirically, one reason for this could be that end-to-end training results in strong language model priors that lead to models preferring more commonly occurring sentences, irrespective of whether they are relevant to the image or not. For instance, we find that state-of-the-art captioning systems [2, 27, 8] incorrectly assign higher likelihoods to irrelevant common captions compared to even ground truth captions paired with a particular image. Furthermore, we also show that this is not just a problem with the captioning models – existing evaluation metrics frequently fail to reward diversity and uniqueness in captions, in fact *preferring* simple automatically generated captions to more descriptive captions produced by human annotators.

In this paper, we take a step towards quantitatively characterizing these deficiencies by proposing a new measure which captures the ability of a caption to uniquely identify an image. We convert a caption into a set of objects, attributes, and relations. For each such concept, we compute its uniqueness as a function of the global number of images containing the concept. This is then aggregated over concepts to compute the uniqueness of the overall caption. This uniqueness metric is orthogonal to standard measures like precision and recall, and allows us to combine them using a harmonic mean to define a new metric, SPICE-U. We empirically demonstrate that this metric correlates better with human judgements than the commonly-used SPICE [1] metric.

Next, we propose techniques to improve current captioning systems at producing unique, more meaningful captions. We employ the strategy of re-ranking

captions during the decoding process by maximizing mutual information between the image and the caption (inspired by a similar line of work in machine translation [18]). Our method achieves an absolute improvement of up to 1.6% in SPICE-U and a relative improvement of up to 2.4% on the average across different metrics. The captions produced are more informative and relevant to the image while not losing out on fluency.

To summarize, the contributions of this paper are:

- quantitatively demonstrating limitations of current captioning systems and metrics.
- proposing a new metric (SPICE-U) that measures the ability of captions to be unique and descriptive.
- investigating new decoding objectives to generate more informative captions.

2 Related work

Discriminative captioning. A number of recent approaches address the task of producing discriminative captions [28, 37, 24, 27, 22]. One such method considers the task of distinguishing a target image from a distractor image using generated captions [37]. The proposed method balances the objectives of maximizing the probability of seeing the predicted caption conditioned on the target image and minimizing the probability of seeing the predicted caption conditioned on the distractor image. Other methods [27, 24] incorporate image retrieval into the training process, encouraging the generation of captions that are more likely to be uniquely aligned to the original image than to other images. While these approaches help generate more discriminative captions, they are approximate versions of maximizing mutual information between the image and caption, which we aim to do explicitly.

Descriptive captioning. Prior work has also focused on improving the amount of information present in a single generated caption. Dense captioning [9] aims to identify all the salient regions in an image and describe each with a caption. Diverse image annotation [44] focuses on describing as much of the image as possible with a limited number of tags. Entity-aware captioning [25] employs hashtags as additional input. Image paragraph captioning [13, 29] aims to produce more than a single sentence for an image. While these papers do capture some notion of expressiveness, they do not explicitly quantify it or examine trade-offs such as the caption length or uniqueness of generated concepts.

Diversity and mutual information in captioning. Another related line of research to this paper is the area of diversity-promoting objectives for captioning [40, 39, 34, 43, 21]. While the similarity lies in aiming to prevent generic, dull captions, these approaches do not explicitly try to make sure that the information content of the caption matches well with the image. In terms of measuring diversity, some papers propose metrics that use corpus-level statistics to provide

coarse judgements [34, 43, 21]. For instance, one can measure how distinct a set of different captions are for a single image, or how many different captions a model generates across the entire test set. In contrast, our metric provides measurements for *each image-caption pair* using aggregated corpus-level information.

Using mutual information to re-rank scores has been explored in speech recognition [3, 31], machine translation [17, 18, 36, 12], conversational agents [50], and multimodal search and retrieval [5, 7, 47]. Maximizing the mutual information as an objective (during either training or inference) has provided reasonable performance gains on all the above tasks. However, to the best of our knowledge, ours is the first work to explore re-ranking via maximizing mutual information specifically to improve the uniqueness of machine-generated captions.

Image captioning metrics. The most commonly used metrics for image captioning evaluation are BLEU [30], METEOR [16], CIDEr [38], and SPICE [1]. BLEU, METEOR, and CIDEr all rely on n-gram matching between the candidate caption and reference captions. BLEU and METEOR are traditionally used in machine translation and thus concerned with syntactical soundness. CIDEr measures the similarity between the candidate caption and "consensus" of the set of reference captions, essentially calculating how often n-grams in the candidate appear in the reference set. SPICE (Semantic Propositional Image Caption Evaluation) is more concerned with the semantics of a caption. It scores a caption based on how closely it matches the scene graph of the target image, where a scene graph consists of a set of object classes, set of relation types, and set of attribute types. While other metrics capture "naturalness" of captions, SPICE correlates better with human judgement by focusing on semantic correctness. Attempts at combining metrics have also been made (e.g. SPIDER [23]). More recent work [33] points out existing models often hallucinate objects when generating captions and proposes the CHAIR score to explicitly evaluate this problem. In contrast to the above rule-base metrics, recent work has also proposed learning statistical models to evaluate captioning systems [4, 6]. While these metrics provide a good measure for the accuracy of a caption, they do not explicitly evaluate how descriptive or informative a caption is. Our metric (SPICE-U) incorporates a new 'uniqueness' measure, while also capturing notions of caption fluency and accuracy through traditional precision and recall components.

3 Analysis: prevalence and causes of common concepts in captions

Current image captioning systems produce captions that are surprisingly fluent and frequently accurate, but generic and uninformative. We begin by demonstrating that the problem of generating common concept is deeply ingrained in both the current captioning systems and in the evaluation metrics. These two factors are closely related as captioning systems are trained to optimize performance on existing metrics. To analyze the captioning systems in the absence of any pre-defined metrics, we take a look directly at the underlying probability

Table 1. Five most common captions in the COCO [20] training set with appearance numbers (based on exact matches over the entire sentence). In Section 3 we demonstrate that captioning models frequently prefer such distractors to ground truth human captions of the images.

a man riding a wave on top of a surfboard (160)
a man flying through the air while riding a skateboard (137)
a man riding skis down a snow covered slope (124)
a man holding a tennis racquet on a tennis court (122)
a large long train on a steel track (116)

distributions learned by the models; to further demonstrate the brittleness of the metrics we design a simple competing baseline that outperforms state-of-the-art captioning systems on standard metrics (this section). Equipped with this analysis, we then go on to propose a new metric (Section 4) along with a potential technical solution (Section 5) to address the problem.

3.1 Captioning systems prefer common concepts

Modern captioning systems are trained to maximize the likelihood of generating the correct caption sentence $\mathbf{s}$ conditioned on the image I, or $P(\mathbf{s}|I)$. Even though the model is learned jointly and does not neatly decompose, intuitively the probability distribution is influenced by two factors: (1) whether a particular concept appears in the image, and (2) the likelihood that the particular concept would appear in a caption. We run a simple experiment to showcase that the latter language prior plays a surprisingly strong role in modern captioning models, helping to partially explain why the systems frequently resort to returning generic image captions corresponding to common concepts.

To do so, we examine the learned probability distribution of the bottom-up top-down attention model [2] trained on the popular Karpathy split [11] of the COCO dataset [20]. On every validation image, we compare the model's likelihood of the human generated ground truth captions for this image with the model's likelihood corresponding to generic distractor sentences applied to this image. For distractor sentences, we use the five captions that appear most frequently in training set (Table 1). During evaluation, to ensure that these distractor sentences are not correct description of the corresponding image, we use the code from [33] to only keep the sentences that contain at least one hallucinated object not present in the image. We observe that in an amazing 73% of the images the model returns a higher likelihood $P(d|I)$ for one of these *wrong* distractor sentences d than its likelihood $P(g|I)$ of one of the ground truth caption, i.e., $\exists d, g : P(d|I) > P(g|I)$. Figure 2 qualitatively illustrates why this is the case: an incorrect caption associated with common concepts may end up with a higher overall $P(\mathbf{s}|I)$ than a correct caption albeit with rare words, which would receive lower language model scores.

Ground truth (log likelihood = -3.03)

a	boy	shows	off	his	arm	cast	on	his	skateboard	<end>
0.84	0.10	0.00038	0.38	0.54	0.019	0.000019	0.13	0.19	0.062	0.82

Common distractor (log likelihood = **-1.09**)

a	man	holding	a	tennis	racquet	on	a	tennis	court	<end>
0.84	0.49	0.035	0.81	0.073	0.42	0.53	0.52	0.08	0.83	0.94

Fig. 2. The ground truth caption "a boy shows off his arm cast on his skateboard" has much lower mean log likelihood (-3.03) according to the captioning model of [2] than a common (on this dataset) but incorrect caption "a man holding a tennis racquet on a tennis court." Numbers under each word w_k correspond to $P(w_k|I, w_{<k})$ of the captioning model, color-coded according to their magnitude. This preference for common captions even at the expense of accuracy is a problem in modern captioning models.

3.2 Captioning metrics prefer common concepts

We now demonstrate that the problem is not just in the captioning models but also in the metrics used to evaluate those models, such as SPICE [1].

Background: SPICE. While older metrics such as BLEU [30], METEOR [16] and CIDEr [38] aim to evaluate both the correctness and the fluency of a caption through n-gram matching, SPICE takes a departure from fluency to focus primarily on caption correctness, i.e., whether the caption reflects visual concepts that are indeed in the image. Here, a *visual concept* is a concrete thing or abstract notion that can be both localized in an image and described using natural language. For the purposes of evaluation, visual concepts are restricted to objects, their attributes, and their relations [10, 1, 14].

Consider an image with a set of visual concepts $\mathbf{G}$ and a set of predicted visual concepts $\mathbf{P}$. The accuracy of this description $\mathbf{P}$ is commonly measured using *precision* and *recall* with regard to the ground truth concepts $\mathbf{G}$ [1], where:

$$\text{Rec}(\mathbf{P}; \mathbf{G}) = \frac{|\mathbf{P} \cap \mathbf{G}|}{|\mathbf{G}|}, \qquad \text{Pr}(\mathbf{P}; \mathbf{G}) = \frac{|\mathbf{P} \cap \mathbf{G}|}{|\mathbf{P}|}$$

The SPICE metric trades off between them using the harmonic mean:

$$\text{SPICE}(\mathbf{P}; \mathbf{G}) = \frac{2}{1/\text{Rec}(\mathbf{P}; \mathbf{G}) + 1/\text{Pr}(\mathbf{P}; \mathbf{G})} \tag{1}$$

We can observe that this metric ignores entirely the uniqueness of concepts and implicitly rewards models which predict common concepts (which are easier to recognize) over rare yet more distinctive concepts.

Findings. We run a simple experiment to show that the SPICE metric can be fooled by very simple baseline models that only recognize the **10** *most common*

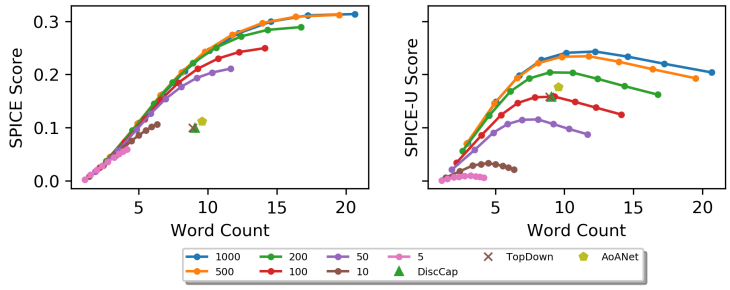

Fig. 3. Comparison of state of the art TopDown model [2], DiscCap model [27], AoANet model [8], and our object detection-based models (best viewed in color). The x-axis is the average caption length in words. The y-axis is the SPICE score [1] (left) and proposed SPICE-U score (right) on 1,076 images (the intersection of the COCO [20] and the Visual Genome dataset [14] which not appear in the training set of both object detection and captioning models). The different curves of the object-based model correspond to running different numbers of object detectors (e.g., detecting only the 1000, 500, etc most common object classes in the image) and producing simplistic captions of the form "There is a tennis ball, court and person". For each curve, performance is shown across varying detection thresholds from 0.1 to 0.9. A simple object-based model that only outputs the 10 most common object classes seen in images (brown) outperforms a state of the art discriminative captioning model (green triangle) on SPICE, but not on SPICE-U.

object classes in images, and nothing else![4] To do so, we design an object-based captioning model consisting of a set of object detectors. The object detectors are trained jointly as a Faster R-CNN model [32], on the Visual Genome training dataset [14].[5] Given a set of detected objects such as "tennis ball," "court" and "person," the final caption is generated following a template as: "There is a tennis ball, court and person".[6] We evaluate the accuracy of this system using the SPICE (Eqn. 1). The evaluation is done on 1,076 images (the intersection of the COCO [20] and the Visual Genome dataset [14] which not appear in the training set of both object detection and captioning models) using their ground truth concept annotations from Visual Genome.

To help interpret the results, we compare this baseline model with three modern captioning systems: the bottom-up and top-down attention model [2], which combines the bottom-up region features generated from object detector with top-down attention mechanism, the model of Luo et al. [27], which includes a "discriminability" loss to encourage unique captions, and the model of Huang et

[4] The objects classes are: man, person, tree, ground, shirt, wall, sky, window, building, and head.

[5] The trained object detectors are taken from the bottom-up part of the captioning model [2].

[6] The resulting model is similar to *Baby Talk* [15], which uses object, attribute, and relationship classifiers to generate image descriptions.

al. [8], which extends conventional models with a stronger attention mechanism. The models are trained on the COCO dataset [20] with the split of [11]. Figure 3 (left) details the results of the experiment. Surprisingly, according to this metric an object detector that only knows **10** object classes rivals a state of the art captioning model: our object-based captioning model achieves a SPICE score of 0.11 versus 0.10 of [27]! This occurs even despite producing fewer words on average per caption: 6.4 versus 9.1. Further, we observe that given access to a (still limited) set of 500 object detectors, our simple baseline produces significantly higher SPICE scores (≥ 0.3).

Conclusions. These surprising findings are likely due to two reasons. First, the SPICE score gives equal weight to different concepts. This means that, for example, a caption that names generic objects like "tree" and "person" scores the same as a caption that identifies the two unique objects in the image, such as "volleyball" or "gazebo", giving a perhaps unfair advantage to our simple baseline. We will address this by proposing a new uniqueness-based metric in Section 4. Second, modern captioning systems are optimized to rely too heavily on the common concepts, failing to fully leverage their image understanding capabilities, and we propose some strategies to mitigate that in Section 5.

4 SPICE-U: A uniqueness-aware metric

Inspired by the observations in Section 3, we introduce the SPICE-U metric ("Semantic Propositional Image Caption Evaluation with Uniqueness") to encourage captions to capture the diversity and uniqueness of real-world images.

Uniqueness. We define the uniqueness of a single visual concept p as:

$$\mathrm{Un}(p) = \frac{\#\text{ images not containing p}}{\#\text{ images total}} \tag{2}$$

This is similar to the notion of *inverse document frequency* (IDF) in text retrieval [35], which allows for weighting down common words in text. While this concept is also used in CIDEr [38], they compute IDF over n-grams, not visual concepts. Note that our definition of uniqueness is complementary to saliency – while saliency measures how prominent a concept is in the image, uniqueness aims to identify parts of the image that make it *interesting*. Future work could involve investigating combinations of these.

For computational tractability, we approximate the denominator using a large set of images (e.g. the training set). For example, if p is *tree*, contained in 28,186 of 113,287 images in the COCO training set [20], $\mathrm{Un}[tree] = 0.75$. We realize that this approximation introduces some dependence on the corpus, but this is similar to calculating IDF using a large text corpus in metrics like CIDEr. Further, even measures like recall implicitly make corpus-specific assumptions,

e.g. by considering the set of ground truth concepts to be those concepts seen in the dataset.

To define the uniqueness of a set of predictions $\mathbf{P}$, we want to consider the uniqueness of its constituent concepts. One natural definition would be:

$$\mathrm{Un}(\mathbf{P}) = \sum_{p \in \mathbf{P}} \mathrm{Un}(p) \tag{3}$$

However, this definition is undesirable for several reasons. First, it's not between 0 and 1, making it difficult to reason about in comparison with precision and recall. Second, and more problematically, it increases with every additional concept (unless the concept is present in 100% of the training images), encouraging long captions. Finally, it encourages the models to make incorrect predictions and detect unusual concepts not present in the image just to increase the uniqueness score.

Instead, we use a definition that measures the uniqueness of a set of predictions compared to the best (most unique) set of predictions which could have been made. To do so, consider alternative predictions $\mathbf{A}$ of the same length as $\mathbf{P}$. As to not encourage a reduction in accuracy through uniqueness, we further assume $\mathbf{A}$ consists only of the concepts that appear either within $\mathbf{P}$ or within the ground truth set $\mathbf{G}$. Concretely:

$$\mathcal{A}(\mathbf{P}; \mathbf{G}) = \{\mathbf{A} : \mathbf{A} \in \mathbf{G} \cup \mathbf{P}, |\mathcal{A}| = |\mathbf{P}|\} \tag{4}$$

For example, if the image contains a *cat* and a *dog*, and the prediction was *cat* and $fish$:

$$\mathcal{A}(\{(cat, fish)\}, \{(cat, dog)\}) = \{(cat, dog), (cat, fish), (dog, fish)\} \tag{5}$$

Given this definition, we then define the *uniqueness* of a prediction as:

$$\mathrm{Uniq}(\mathbf{P}; \mathbf{G}) = \frac{\mathrm{Un}(\mathbf{P}) - \min_{\mathbf{A} \in \mathcal{A}(\mathbf{G};\mathbf{P})} \mathrm{Un}(\mathbf{A})}{\max_{\mathbf{A} \in \mathcal{A}(\mathbf{G};\mathbf{P})} \mathrm{Un}(\mathbf{A}) - \min_{\mathbf{A} \in \mathcal{A}(\mathbf{G};\mathbf{P})} \mathrm{Un}(\mathbf{A})} \tag{6}$$

Intuitively, this measures how unique the caption is compared to others of the same length that could have been conceivably generated. For example, consider an image that contains a *person* (uniqueness score of 0.75), *table* (score of 0.87), and *elephant* (score of 0.98). If the model captions only one of these objects and nothing else, it will be rewarded with a uniqueness score of 1 if the object it chooses is *elephant*, 0 if it outputs *person*, and 0.52 if it outputs *table*. Note that predicting a more unique, yet incorrect, object would not give the model an additional reward. Similarly, if the image did not contain an *elephant*, then the model would receive the full uniqueness score of 1 for predicting the most unique object *table*. This ensures that models are rewarded for noticing unique things in the image but not unfairly penalized on images with only common concepts.

Combined metric. The uniqueness-aware measure of the quality of a caption is then a combination through harmonic mean of SPICE (Eqn. 1), and uniqueness (Eqn. 6):

Table 2. Evaluation of various metrics against human judgements. First five columns show pairwise judgment accuracy with fifty reference captions on the PASCAL-50 dataset (HC: both sentences written by humans for the corresponding image, HI: both sentences written by humans – one for the corresponding image and one for a random image, HM: one caption written by human and another generated by a model, MM: captions generated by two different models.) The last column is Pearson's correlation between human preferences and each metric on images from PASCAL-50.

	HC	HI	HM	MM	ALL	Pearson's
BLEU-4	55.00	97.30	92.60	61.80	76.68	0.581
ROUGE	54.60	98.70	96.00	62.00	77.83	0.732
METEOR	57.50	**99.30**	**96.90**	62.30	79.00	0.710
CIDEr	53.00	**99.30**	92.10	67.10	77.88	0.641
SPICE	**66.80**	98.50	93.80	**71.10**	82.55	0.749
SPICE-U	66.50	98.60	94.40	70.80	**82.58**	**0.767**

$$\text{SPICE-U}(\mathbf{P};\mathbf{G}) = \frac{2}{1/\text{SPICE}(\mathbf{P};\mathbf{G}) + 1/\text{Uniq}(\mathbf{P};\mathbf{G})} \tag{7}$$

Consider the example above of an image that contains a *person*, *table* and *elephant*, and two captions: "There is a table" and "There is an elephant." The original SPICE score of Eqn. 1 would be 0.5 for both captions (recall 1/3, precision 1), failing to recognize that one is a much more useful caption than the other. However, SPICE-U score would be 0.67 for "There is an elephant" and 0.51 for "There is a table," correctly selecting the most informative description.[7]

Advantage of SPICE-U. We follow the setup of [1] to analyze correlation of SPICE-U with human judgements when determining the similarity of sentences. We use the PASCAL-50S dataset [38], which contains 50 ground truth captions for each image. Human annotators were provided with a pair of candidate sentences (b, c) and asked which was more similar to sentence a, which is one of the ground truth captions for an image. Consider an image with a set of ground truth captions $A = \{a_k\}$ and a reference pair of sentences (b, c) as above, where without loss of generality we assume that humans favored b over c for this image (i.e., on average over all a_k, humans found a_k to be more similar to b than c). We say that a metric agrees with humans if $\text{metric}(b, A) \geq \text{metric}(c, A)$. From table 2, we observe that SPICE-U achieves better judgement accuracy than other metrics and comparable accuracy with SPICE, especially outperforming SPICE on HM pairs. This shows that SPICE-U can indeed capture the diverse nature of human written captions and can help separate two captions that are both

[7] For "There is a person" uniqueness is 0, since it's the most common of the objects, and SPICE-U score is 0 by definition.

correct but differ in quality. Despite being a standard test on PASCAL-50S, measuring the accuracy abstracts away detailed human preferences, and causes issue when two candidate captions get similar human votes. To mitigate this, we also evaluate Pearson's correlation between human preferences and each metric[8]. SPICE-U achieves the best correlation score among all metrics.

5 Generating unique and informative captions

SPICE-U aims to capture the uniqueness of a particular caption given an image. Intuitively, any captioning model that maximizes SPICE-U must forge a strong connection between the semantic concepts in the image and the linguistic concepts in the caption it generates. However, in the predominant (current) regime of end-to-end training with loss functions such as cross entropy, there is no explicit objective which enables this connection.

Formally, current captioning models decode using the following objective:

$$\hat{s} = \arg\max_{s} \log P(s|I;\theta) \tag{8}$$

where s is the caption, I is the image and θ are the learned parameters of the model. However, this ignores the dependency from the caption to the image $P(I|s)$, which is critical for ensuring that the caption adequately (and uniquely) describes the image. A similar observation was made in machine translation [17, 18] where the input and output are sentences in two different languages.

One solution to this problem is to maximize mutual information (MMI) instead of cross-entropy:

$$\begin{aligned}\hat{s} &= \arg\max_{s} \log \frac{P(I,s)}{P(I)P(s)^{\lambda}} \\ &= \arg\max_{s} \log P(s|I) - \lambda \log P(s) \\ &= \arg\max_{s} (1-\lambda) \log P(s|I) + \lambda \log P(I|s)\end{aligned} \tag{9}$$

However, since training a model to predict $P(I|s)$ is not trivial [19, 49, 46], we propose to use second line in the MMI objective above to *re-rank captions* produced by a standard beam decoding mechanism. To this end, we train language models to obtain likelihood estimates for captions, $\log P(s) = \sum_i \log P(s_i|s_{<i})$. In particular, we investigate three variants of language models:

1. **Unigram LM:** A simple unigram language model estimated from the train set, $P(s) = \prod_i P(s_i)$
2. **LSTM LM:** An LSTM language model trained on captions in the train set.

[8] We calculate the correlation between the mean value of human votes (+1 if they prefer caption b over caption c, -1 otherwise) and the score $R_m(b) - R_m(c)$, where $R_m(s)$ is the score of sentence s given by metric m.

Algorithm 1 Generating caption with beam decoding and re-ranking

Input: Caption model with parameter θ_c, language model with parameter θ_l, image I, weighting factor λ

Output: Generated caption s

1: Beam decode top-k captions $\{s^{(1)}, ..., s^{(k)}\}$ along with probabilities $\{P(s^{(1)}|I;\theta_c), ..., P(s^{(k)}|I;\theta_c)\}$ with caption model
2: Generate probabilities for entire captions $\{P(s^{(1)};\theta_l), ..., P(s^{(k)};\theta_l)\}$ with language model
3: $s \leftarrow \arg\max_{s^{(i)}} \log P(s^{(i)}|I;\theta_c) - \lambda \log P(s^{(i)};\theta_l)$

3. **Interpolated LM:** A log-linear interpolation[9] between the variants above:

$$P_{int}(s_i|s_{<i}) = P_{uni}(s_i)^{\alpha} P_{LSTM}(s_i|s_{<i})^{1-\alpha} \tag{10}$$

We generate the top-k captions using the baseline model and then re-rank them using their newly computed scores, described in Algorithm 1.

6 Experiments

Data. We conduct experiments on the COCO [20] dataset which contains images of everyday scenes with common objects in their natural context. For captioning task, every image is annotated with five human captions, mostly short sentences summarizing the important parts of the scene. We adopt the popular split of this dataset from Karpathy et al. [11], which contains 113,287 images for training and 5,000 images for validation and test respectively.

Model. We use three recent captioning models as our baselines. The bottom-up and top-down attention model (TopDown) from Anderson et al. [2] utilizes object detector to propose salient image regions as bottom-up features and then uses top-down attention to decide weight for each region. The discriminative captioning model (DiscCap) from Luo et al. [27] is trained explicitly with proposed 'discriminability' loss besides standard cross-entropy loss to encourage unique captions that can distinguish between different images. The attention on attention model (AoANet) from Huang et al. [8] extends conventional attention mechanism with another attention to determine the relevance between attention results and queries. We use off-the-shelf implementations for these models[10]. For language model, we train a one-layer LSTM with hidden size of 512 and embedding size of 300.

[9] We also tried linear interpolation and it works not as good as the log-linear interpolation.

[10] The TopDown model from https://github.com/poojahira/image-captioning-bottom-up-top-down, the DiscCap from https://github.com/ruotianluo/DiscCaptioning and AoANet from https://github.com/husthuaan/AoANet.

Table 3. Comparison of three different state-of-the-art captioning systems [2, 27, 8], along with our proposed re-ranking schemes, evaluated using different metrics on the COCO test split from [11].

	BLEU	METEOR	CIDEr	CHAIRs (↓)	SPICE	SPICE-U	GeoMean
TopDown [2]	**23.03**	28.98	**108.13**	8.68	20.62	23.70	12.63
TopDown+Unigram	22.88	**29.06**	107.04	8.10	**20.82**	25.05	12.89
TopDown+LSTM	22.79	28.48	107.59	8.20	20.52	24.46	12.74
TopDown+Interpolated	22.77	28.84	106.42	**7.80**	20.72	**25.27**	**12.94**
DiscCap [27]	**21.93**	**27.55**	**112.39**	11.92	**20.32**	23.74	11.84
DiscCap+Unigram	21.56	27.38	110.41	10.88	20.28	24.60	12.00
DiscCap+LSTM	21.64	27.40	111.73	11.34	20.17	23.79	11.87
DiscCap+Interpolated	21.58	27.42	110.90	**10.84**	20.27	**24.52**	**12.02**
AoANet [8]	**27.53**	30.37	**129.12**	10.40	22.77	26.04	13.54
AoANet+Unigram	27.30	**30.43**	128.66	9.52	22.79	26.46	13.75
AoANet+LSTM	27.36	30.26	128.79	10.24	22.71	26.12	13.55
AoANet+Interpolated	27.18	30.39	128.15	**9.28**	**22.81**	**26.53**	**13.80**

Re-ranking. We use the captioning model with beam decoding to generate top 10 candidates along with probabilities $P(s|I)$ for re-ranking. The language model is then used to generate the $P(s)$ for each candidate caption and finally the caption with the maximum mutual information is selected according to Eqn. 9 as the predicted caption.

The hyperparameters λ (language model weight in Eqn. 9) and α (coefficient in interpolation model, Eqn. 10) are selected for each model on the validation set using a grid search (0 to 1, step size of 0.1).

We cross-validate with the objective of optimizing the geometric mean[11] across several evaluation metrics (BLEU-4, METEOR, CIDEr, CHAIRs, SPICE and SPICE-U). The resulting hyperparameters are: $\lambda = 0.3$ on TopDown+Unigram, $\lambda = 0.2$ on TopDown+LSTM, $\lambda = 0.4, \alpha = 0.8$ on TopDown+Interpolated, $\lambda = 1.0$ on DiscCap+Unigram, $\lambda = 0.1$ on DiscCap+LSTM, $\lambda = 0.8, \alpha = 0.9$ on DiscCap+Interpolated, and $\lambda = 0.4$ on AoANet+Unigram, $\lambda = 0.1$ on AoANet+LSTM, $\lambda = 0.5, \alpha = 0.9$ on AoANet+Interpolated.

Results. Table 3 summarizes the results. For the TopDown baseline, the TopDown+Interpolated modification improves SPICE-U by an absolute 1.6% over the baseline (from 23.7% to 25.3%) and the geometric mean over all metrics by a relative 2.4% (from 12.6% to 12.9%). For DiscCap model, DiscCap+Interpolated led to an absolute improvement of 0.8% on SPICE-U (from 23.7% to 24.5%) and 1.7% relative on the geometric mean (from 11.8% to 12.0%). For AoANet, AoANet+Interpolated improves SPICE-U by an absolute 0.5% (from 26.0% to 26.5%) and a relative 2.2% improvement on geometric mean.

[11] The captioning metrics measure different aspects of the captions and are largely uncorrelated with each other [33]; we use the geometric mean as a simple summary statistic of the overall performance of the models. For CHAIR lower scores are better so we use $\frac{1}{CHAIR}$ in the geometric mean.

Fig. 4. Captions generated by the AoANet [8] model (in italics) and by our variation AoANet+Interpolated (in regular font). The modification we introduce encourages the model to output more descriptive and accurate captions, such as describing the place ("beach"), the type of the sign ("no parking sign"), the presence of a prominent object ("paper container", "statue") in the first four images. However, there are also some images (like the last one) where despite improvements in SPICE-U the changes are less interesting, such as simply replacing "group" with "herd".

Figure 4 shows qualitative examples: as expected, the updated captions correspond to more detailed descriptions of the image. The improvements demonstrated here are the result of quite simple algorithmic modification yet propose a promising path forward for improving modern image captioning system.

7 Conclusion

State of the art image captioning models produce generic captions, leaving out important image details and misrepresenting facts. In this paper, we quantitatively demonstrated that both modern captioning systems and evaluation metrics tend towards generating and rewarding captions with commonly occurring concepts from the training data. We then introduced a new notion of *uniqueness* and used it to propose a new metric, SPICE-U. Our studies show that SPICE-U correlates better with human judgements compared to SPICE. Finally, we utilized the notion of maximizing mutual information to re-rank captions produced by any captioning system. Our experiments demonstrate that our method results in unique and informative captions, and yields promising improvements over three different state-of-the-art models.

Acknowledgments. This work is partially supported by KAUST under Award No. OSRCRG2017-3405, by Samsung and by the Princeton CSML DataX award. We would like to thank Arjun Mani, Vikram Ramaswamy and Angelina Wang for their helpful feedback on the paper.

References

1. Anderson, P., Fernando, B., Johnson, M., Gould, S.: SPICE: Semantic Propositional Image Caption Evaluation. In: ECCV (2016)
2. Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., Zhang, L.: Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering. In: CVPR (2018)
3. Bahl, L., Brown, P., de Souza, P., Mercer, R.: Maximum mutual information estimation of hidden Markov model parameters for speech recognition. In: ICASSP (1986)
4. Cui, Y., Yang, G., Veit, A., Huang, X., Belongie, S.: Learning to Evaluate Image Captioning. In: CVPR (2018)
5. Datta, D., Varma, S., Chowdary C., R., Singh, S.K.: Multimodal Retrieval using Mutual Information based Textual Query Reformulation. Expert Systems with Applications (2017)
6. Dognin, P., Melnyk, I., Mroueh, Y., Ross, J., Sercu, T.: Adversarial Semantic Alignment for Improved Image Captions. In: CVPR (2019)
7. Henning, C.A., Ewerth, R.: Estimating the Information Gap between Textual and Visual Representations. In: ICMR (2017)
8. Huang, L., Wang, W., Chen, J., Wei, X.Y.: Attention on Attention for Image Captioning. In: ICCV (2019)
9. Johnson, J., Karpathy, A., Fei-Fei, L.: DenseCap: Fully Convolutional Localization Networks for Dense Captioning. In: CVPR (2016)
10. Johnson, J., Krishna, R., Stark, M., Li, L.J., Shamma, D.A., Bernstein, M.S., Fei-Fei, L.: Image retrieval using scene graphs. In: CVPR (2015)
11. Karpathy, A., Fei-Fei, L.: Deep visual-semantic alignments for generating image descriptions. In: CVPR (2015)
12. Kimura, R., Iida, S., Cui, H., Hung, P.H., Utsuro, T., Nagata, M.: Selecting Informative Context Sentence by Forced Back-Translation. In: MT Summit XVII (2019)
13. Krause, J., Johnson, J., Krishna, R., Fei-Fei, L.: A Hierarchical Approach for Generating Descriptive Image Paragraphs. In: CVPR (2017)
14. Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.J., Shamma, D.A., Bernstein, M.S., Fei-Fei, L.: Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations. International Journal of Computer Vision (2017)
15. Kulkarni, G., Premraj, V., Ordonez, V., Dhar, S., Li, S., Choi, Y., Berg, A.C., Berg, T.L.: BabyTalk: Understanding and Generating Simple Image Descriptions. IEEE Transactions on Pattern Analysis and Machine Intelligence (2013)
16. Lavie, A., Agarwal, A.: Meteor: an automatic metric for MT evaluation with high levels of correlation with human judgments. In: StatMT (2007)
17. Li, J., Galley, M., Brockett, C., Gao, J., Dolan, B.: A Diversity-Promoting Objective Function for Neural Conversation Models. In: NAACL HLT (2016)
18. Li, J., Jurafsky, D.: Mutual Information and Diverse Decoding Improve Neural Machine Translation. arXiv:1601.00372 [cs] (2016), arXiv: 1601.00372
19. Li, W., Zhang, P., Zhang, L., Huang, Q., He, X., Lyu, S., Gao, J.: Object-Driven Text-To-Image Synthesis via Adversarial Training. In: CVPR (2019)
20. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: Common Objects in Context. In: ECCV (2014)

21. Lindh, A., Ross, R.J., Mahalunkar, A., Salton, G., Kelleher, J.D.: Generating Diverse and Meaningful Captions. In: ICANN (2018)
22. Liu, L., Tang, J., Wan, X., Guo, Z.: Generating Diverse and Descriptive Image Captions Using Visual Paraphrases. In: ICCV (2019)
23. Liu, S., Zhu, Z., Ye, N., Guadarrama, S., Murphy, K.: Improved Image Captioning via Policy Gradient optimization of SPIDEr. In: ICCV (2017)
24. Liu, X., Li, H., Shao, J., Chen, D., Wang, X.: Show, Tell and Discriminate: Image Captioning by Self-retrieval with Partially Labeled Data. In: ECCV (2018)
25. Lu, D., Whitehead, S., Huang, L., Ji, H., Chang, S.F.: Entity-aware Image Caption Generation. In: EMNLP (2018)
26. Lu, J., Xiong, C., Parikh, D., Socher, R.: Knowing When to Look: Adaptive Attention via a Visual Sentinel for Image Captioning. In: CVPR (2017)
27. Luo, R., Shakhnarovich, G., Cohen, S., Price, B.: Discriminability Objective for Training Descriptive Captions. In: CVPR (2018)
28. Mao, J., Huang, J., Toshev, A., Camburu, O., Yuille, A., Murphy, K.: Generation and Comprehension of Unambiguous Object Descriptions. In: CVPR (2016)
29. Melas-Kyriazi, L., Rush, A., Han, G.: Training for Diversity in Image Paragraph Captioning. In: EMNLP (2018)
30. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: BLEU: a method for automatic evaluation of machine translation. In: ACL (2001)
31. Povey, D., Woodland, P.: Minimum Phone Error and I-smoothing for improved discriminative training. In: ICASSP (2002)
32. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. IEEE Transactions on Pattern Analysis and Machine Intelligence (2017)
33. Rohrbach, A., Hendricks, L.A., Burns, K., Darrell, T., Saenko, K.: Object Hallucination in Image Captioning. In: EMNLP (2018)
34. Shetty, R., Rohrbach, M., Hendricks, L.A., Fritz, M., Schiele, B.: Speaking the Same Language: Matching Machine to Human Captions by Adversarial Training. In: ICCV (2017)
35. Sparck Jones, K.: A Statistical Interpretation of Term Specificity and Its Application in Retrieval. Journal of Documentation (1972)
36. Tu, Z., Liu, Y., Shang, L., Liu, X., Li, H.: Neural Machine Translation with Reconstruction. In: AAAI (2017)
37. Vedantam, R., Bengio, S., Murphy, K., Parikh, D., Chechik, G.: Context-Aware Captions from Context-Agnostic Supervision. In: CVPR (2017)
38. Vedantam, R., Zitnick, C.L., Parikh, D.: CIDEr: Consensus-based image description evaluation. In: CVPR (2015)
39. Vijayakumar, A.K., Cogswell, M., Selvaraju, R.R., Sun, Q., Lee, S., Crandall, D., Batra, D.: Diverse Beam Search for Improved Description of Complex Scenes. In: AAAI (2018)
40. Vijayakumar, A.K., Cogswell, M., Selvaraju, R.R., Sun, Q., Lee, S., Crandall, D., Batra, D.: Diverse Beam Search: Decoding Diverse Solutions from Neural Sequence Models. arXiv:1610.02424 [cs] (2018), arXiv: 1610.02424
41. Vinyals, O., Toshev, A., Bengio, S., Erhan, D.: Show and tell: A neural image caption generator. In: CVPR (2015)
42. Vinyals, O., Toshev, A., Bengio, S., Erhan, D.: Show and Tell: Lessons Learned from the 2015 MSCOCO Image Captioning Challenge (2017)
43. Wang, Q., Chan, A.B.: Describing Like Humans: On Diversity in Image Captioning. In: CVPR (2019)

44. Wu, B., Jia, F., Liu, W., Ghanem, B.: Diverse Image Annotation. In: CVPR (2017)
45. Xu, K., Ba, J.L., Kiros, R., Cho, K., Courville, A., Salakhutdinov, R., Zemel, R.S., Bengio, Y.: Show, attend and tell: neural image caption generation with visual attention. In: ICML (2015)
46. Xu, T., Zhang, P., Huang, Q., Zhang, H., Gan, Z., Huang, X., He, X.: AttnGAN: Fine-Grained Text to Image Generation with Attentional Generative Adversarial Networks. In: CVPR (2018)
47. Yao, T., Mei, T., Ngo, C.W.: Co-reranking by mutual reinforcement for image search. In: CVPR (2010)
48. You, Q., Jin, H., Wang, Z., Fang, C., Luo, J.: Image Captioning with Semantic Attention. In: CVPR (2016)
49. Zhang, H., Xu, T., Li, H., Zhang, S., Wang, X., Huang, X., Metaxas, D.: StackGAN: Text to Photo-Realistic Image Synthesis with Stacked Generative Adversarial Networks. In: ICCV (2017)
50. Zhang, Y., Galley, M., Gao, J., Gan, Z., Li, X., Brockett, C., Dolan, B.: Generating Informative and Diverse Conversational Responses via Adversarial Information Maximization. In: NeurIPS (2018)