SPICE: Semantic Propositional Image Caption Evaluation

Peter Anderson[1(✉)], Basura Fernando[1], Mark Johnson[2], and Stephen Gould[1]

[1] The Australian National University, Canberra, Australia
{peter.anderson,basura.fernando,stephen.gould}@anu.edu.au
[2] Macquarie University, Sydney, Australia
mark.johnson@mq.edu.au

Abstract. There is considerable interest in the task of automatically generating image captions. However, evaluation is challenging. Existing automatic evaluation metrics are primarily sensitive to n-gram overlap, which is neither necessary nor sufficient for the task of simulating human judgment. We hypothesize that semantic propositional content is an important component of human caption evaluation, and propose a new automated caption evaluation metric defined over scene graphs coined *SPICE*. Extensive evaluations across a range of models and datasets indicate that SPICE captures human judgments over model-generated captions better than other automatic metrics (e.g., system-level correlation of 0.88 with human judgments on the MS COCO dataset, versus 0.43 for CIDEr and 0.53 for METEOR). Furthermore, SPICE can answer questions such as *which caption-generator best understands colors?* and *can caption-generators count?*

1 Introduction

Recently there has been considerable interest in joint visual and linguistic problems, such as the task of automatically generating image captions [1,2]. Interest has been driven in part by the development of new and larger benchmark datasets such as Flickr 8K [3], Flickr 30K [4] and MS COCO [5]. However, while new datasets often spur considerable innovation—as has been the case with the MS COCO Captioning Challenge [6]—benchmark datasets also require fast, accurate and inexpensive evaluation metrics to encourage rapid progress. Unfortunately, existing metrics have proven to be inadequate substitutes for human judgment in the task of evaluating image captions [3,7,8]. As such, there is an urgent need to develop new automated evaluation metrics for this task [8,9]. In this paper, we present a novel automatic image caption evaluation metric that measures the quality of generated captions by analyzing their semantic content. Our method closely resembles human judgment while offering the additional advantage that the performance of any model can be analyzed in greater detail than with other automated metrics.

One of the problems with using metrics such as Bleu [10], ROUGE [11], CIDEr [12] or METEOR [13] to evaluate captions, is that these metrics are

B. Leibe et al. (Eds.): ECCV 2016, Part V, LNCS 9909, pp. 382–398, 2016.
DOI: 10.1007/978-3-319-46454-1_24

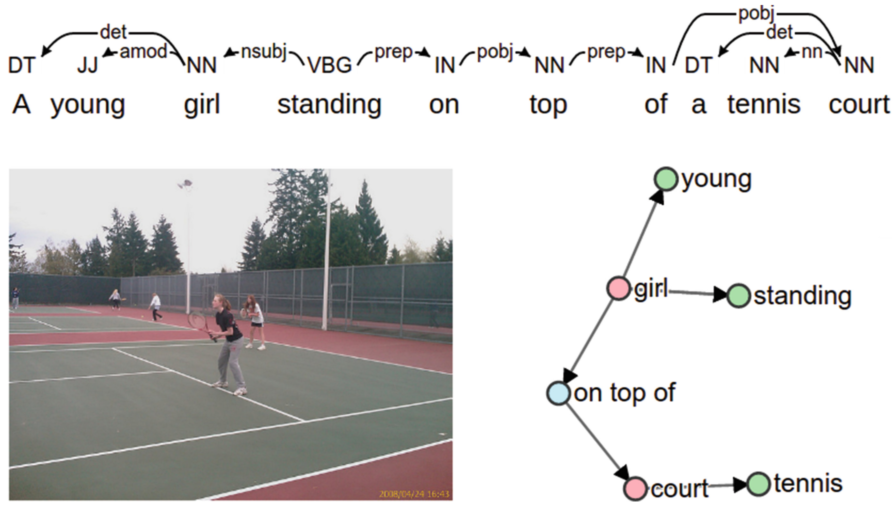


Fig. 1. Illustrates our method's main principle which uses semantic propositional content to assess the quality of image captions. Reference and candidate captions are mapped through dependency parse trees (top) to semantic *scene graphs* (right)—encoding the objects (red), attributes (green), and relations (blue) present. Caption quality is determined using an F-score calculated over tuples in the candidate and reference scene graphs (Color figure online)

primarily sensitive to n-gram overlap. However, *n-gram overlap is neither necessary nor sufficient for two sentences to convey the same meaning* [14].

To illustrate the limitations of n-gram comparisons, consider the following two captions (a,b) from the MS COCO dataset:

(a) A young girl *standing on top of a* tennis court.
(b) A giraffe *standing on top of a* green field.

The captions describe two very different images. However, comparing these captions using any of the previously mentioned n-gram metrics produces a high similarity score due to the presence of the long 5-gram phrase *'standing on top of a'* in both captions. Now consider the captions (c,d) obtained from the same image:

(c) A shiny metal pot filled with some diced veggies.
(d) The pan on the stove has chopped vegetables in it.

These captions convey almost the same meaning, but exhibit low n-gram similarity as they have no words in common.

To overcome the limitations of existing n-gram based automatic evaluation metrics, in this work we hypothesize that *semantic propositional content is an important component of human caption evaluation.* That is, given an image with the caption 'A young girl standing on top of a tennis court', we expect that a human evaluator might consider the truth value of each of the semantic propositions contained therein—such as (1) there is a girl, (2) girl is young, (3) girl is standing, (4) there is a court, (5) court is tennis, and (6) girl is on top of court. If each of these propositions is clearly and obviously supported by the image, we would expect the caption to be considered acceptable, and scored accordingly.

Taking this main idea as motivation, we estimate caption quality by transforming both candidate and reference captions into a graph-based semantic representation called a *scene graph*. The scene graph explicitly encodes the objects, attributes and relationships found in image captions, abstracting away most of the lexical and syntactic idiosyncrasies of natural language in the process. Recent work has demonstrated scene graphs to be a highly effective representation for performing complex image retrieval queries [15,16], and we demonstrate similar advantages when using this representation for caption evaluation.

To parse an image caption into a scene graph, we use a two-stage approach similar to previous works [16–18]. In the first stage, syntactic dependencies between words in the caption are established using a dependency parser [19] pre-trained on a large dataset. An example of the resulting dependency syntax tree, using Universal Dependency relations [20], is shown in Fig. 1 top. In the second stage, we map from dependency trees to scene graphs using a rule-based system [16]. Given candidate and reference scene graphs, our metric computes an F-score defined over the conjunction of logical tuples representing semantic propositions in the scene graph (e.g., Fig. 1 right). We dub this approach SPICE for *Semantic Propositional Image Caption Evaluation*.

Using a range of datasets and human evaluations, we show that SPICE outperforms existing n-gram metrics in terms of agreement with human evaluations of model-generated captions, while offering scope for further improvements to the extent that semantic parsing techniques continue to improve. We make code available from the project page[1]. Our main contributions are:

1. We propose SPICE, a principled metric for automatic image caption evaluation that compares semantic propositional content;
2. We show that SPICE outperforms metrics Bleu, METEOR, ROUGE-L and CIDEr in terms of agreement with human evaluations; and
3. We demonstrate that SPICE performance can be decomposed to answer questions such as ‘which caption-generator best understands colors?’ and ‘can caption generators count?’

2 Background and Related Work

2.1 Caption Evaluation Metrics

There is a considerable amount of work dedicated to the development of metrics that can be used for automatic evaluation of image captions. Typically, these metrics are posed as similarity measures that compare a candidate sentence to

[1] http://panderson.me/spice.

a set of reference or ground-truth sentences. Most of the metrics in common use for caption evaluation are based on n-gram matching. Bleu [10] is a modified precision metric with a sentence-brevity penalty, calculated as a weighted geometric mean over different length n-grams. METEOR [13] uses exact, stem, synonym and paraphrase matches between n-grams to align sentences, before computing a weighted F-score with an alignment fragmentation penalty. ROUGE [11] is a package of a measures for automatic evaluation of text summaries using F-measures. CIDEr [12] applies term frequency-inverse document frequency (tf-idf) weights to n-grams in the candidate and reference sentences, which are then compared by summing their cosine similarity across n-grams. With the exception of CIDEr, these methods were originally developed for the evaluation of text summaries or machine translations (MT), and were subsequently adopted for image caption evaluation.

Several studies have analyzed the performance of n-gram metrics when used for image caption evaluation, by measuring correlation with human judgments of caption quality. On the PASCAL 1K dataset, Bleu-1 was found to exhibit weak or no correlation (Pearson's r of -0.17 and 0.05) [7]. Using the Flickr 8K [3] dataset, METEOR exhibited moderate correlation (Spearman's ρ of 0.524) outperforming ROUGE SU-4 (0.435), Bleu-smoothed (0.429) and Bleu-1 (0.345) [8]. Using the PASCAL-50S and ABSTRACT-50S datasets, CIDEr and METEOR were found to have greater agreement with human consensus than Bleu and ROUGE [12].

Within the context of automatic MT evaluation, a number of papers have proposed the use of shallow-semantic information such as semantic role labels (SRLs) [14]. In the MEANT metric [21], SRLs are used to try to capture the basic event structure of sentences – '*who* did *what* to *whom*, *when*, *where* and *why*' [22]. Using this approach, sentence similarity is calculated by first matching semantic frames across sentences by starting with the verbs at their head. However, this approach does not easily transfer to image caption evaluation, as verbs are frequently absent from image captions or not meaningful – e.g. 'a very tall building with a train *sitting* next to it' – and this can de-rail the matching process. Our work differs from these approaches as we represent sentences using scene graphs, which allow for noun / object matching between captions. Conceptually, the closest work to ours is probably the bag of aggregated semantic tuples (BAST) metric [23] for image captions. However, this work required the collection of a purpose-built dataset in order to learn to identify Semantic Tuples, and the proposed metric was not evaluated against human judgments or existing metrics.

2.2 Semantic Graphs

Scene graphs, or similar semantic structures, have been used in a number of recent works within the context of image and video retrieval systems to improve performance on complex queries [15,16,18]. Several of these papers have demonstrated that semantic graphs can be parsed from natural language descriptions [16,18]. The task of transforming a sentence into its meaning representation has also received considerable attention within the computational linguistics community. Recent work has proposed a common framework for semantic graphs called an abstract meaning representation (AMR) [24], for which a number of parsers [17,25,26] and the Smatch evaluation metric [27] have been developed. However, in initial experiments, we found that AMR representations using Smatch similarity performed poorly as image caption representations. Regardless of the representation used, the use of dependency trees as the starting point for parsing semantic graphs appears to be a common theme [16–18].

3 SPICE Metric

Given a candidate caption c and a set of reference captions $S = \{s_1, \ldots, s_m\}$ associated with an image, our goal is to compute a score that captures the similarity between c and S. For the purposes of caption evaluation the image is disregarded, posing caption evaluation as a purely linguistic task similar to machine translation (MT) evaluation. However, because we exploit the semantic structure of scene descriptions and give primacy to nouns, our approach is better suited to evaluating computer generated image captions.

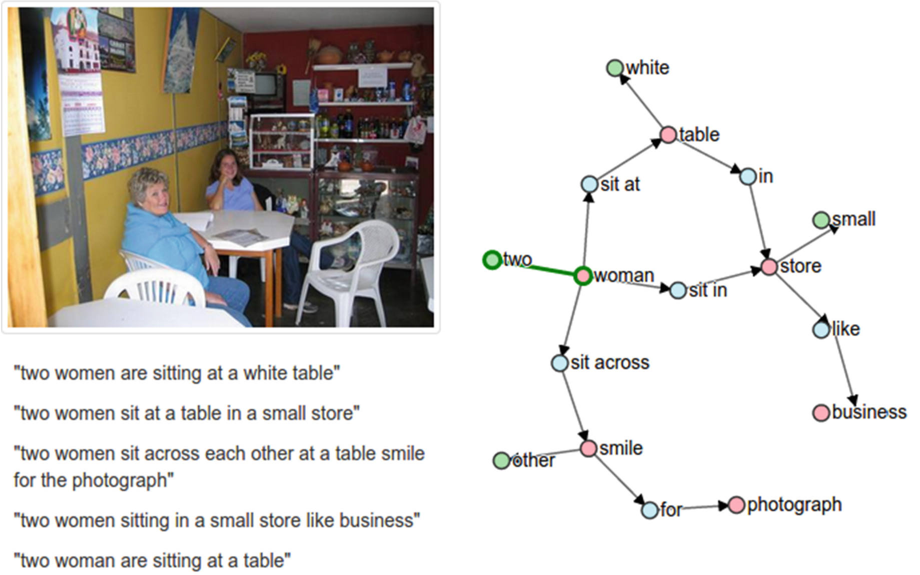
First, we transform both candidate caption and reference captions into an intermediate representation that encodes semantic propositional content. While we are aware that there are other components of linguistic meaning—such as figure-ground relationships—that are almost certainly relevant to caption quality, in this work we focus exclusively on *semantic meaning*. Our choice of semantic representation is the *scene graph*, a general structure consistent with several existing vision datasets [15,16,28] and the recently released Visual Genome dataset [29]. The scene graph of candidate caption c is denoted by $G(c)$, and the scene graph for the reference captions S is denoted by $G(S)$, formed as the union of scene graphs $G(s_i)$ for each $s_i \in S$ and combining synonymous object nodes. Next we present the semantic parsing step to generate scene graphs from captions.

3.1 Semantic Parsing—Captions to Scene Graphs

We define the subtask of parsing captions to scene graphs as follows. Given a set of object classes C, a set of relation types R, a set of attribute types A, and a caption c, we parse c to a scene graph:

$$G(c) = \langle O(c), E(c), K(c) \rangle \tag{1}$$

where $O(c) \subseteq C$ is the set of object mentions in c, $E(c) \subseteq O(c) \times R \times O(c)$ is the set of hyper-edges representing relations between objects, and $K(c) \subseteq O(c) \times A$ is the set of attributes associated with objects. Note that in practice, C, R and A are *open-world* sets that are expanded as new object, relation and attribute types are identified, placing no restriction on the types of objects, relation and attributes that can be represented, including 'stuff' nouns such as grass, sky, etc. An example of a parsed scene graph is illustrated in Fig. 2.



"two women are sitting at a white table"

"two women sit at a table in a small store"

"two women sit across each other at a table smile for the photograph"

"two women sitting in a small store like business"

"two woman are sitting at a table"

Fig. 2. A typical example of a *scene graph* (right) parsed from a set of reference image captions (left)

Our scene graph implementation departs slightly from previous work in image retrieval [15,16], in that we do not represent multiple instances of a single class of object separately in the graph. In previous work, duplication of object instances was necessary to enable scene graphs to be grounded to image regions. In our work, we simply represent object counts as attributes of objects. While this approach does not distinguish collective and distributive readings [16], it simplifies scene graph alignment and ensures that each incorrect numeric modifier is only counted as a single error.

To complete this subtask, we adopt a variant of the rule-based version of the Stanford Scene Graph Parser [16]. A Probabilistic Context-Free Grammar (PCFG) dependency parser [19] is followed by three post-processing steps that simplify quantificational modifiers, resolve pronouns and handle plural nouns. The resulting tree structure is then parsed according to nine simple linguistic rules to extract lemmatized objects, relations and attributes, which together comprise the scene graph. As an example, one of the linguistic rules captures adjectival modifiers, such as the $\textit{young} \xleftarrow{\text{amod}} \textit{girl}$ example from Fig. 1, which results in the object mention 'girl' with attribute 'young'. Full details of the pipeline can be found in the original paper.

SPICE slightly modifies the original parser [16] to better evaluate image captions. First, we drop the plural nouns transformation that duplicates individual nodes of the graph according to the value of their numeric modifier. Instead, numeric modifiers are encoded as object attributes. Second, we add an additional linguistic rule that ensures that nouns will always appear as objects in the scene graph—even if no associated relations can identified—as disconnected graph nodes are easily handled by our semantic proposition F-score calculation.

Notwithstanding the use of the Stanford Scene Graph Parser, our proposed SPICE metric is not tied to this particular parsing pipeline. In fact, it is our hope that ongoing advances in syntactic and semantic parsing will allow SPICE to be further improved in future releases. We also note that since SPICE operates on scene graphs, in principle it could be used to evaluate captions on scene graph datasets [15,16,28] that have no reference captions at all. Evaluation of SPICE under these circumstances is left to future work.

3.2 F-Score Calculation

To evaluate the similarity of candidate and reference scene graphs, we view the semantic relations in the scene graph as a conjunction of logical propositions, or tuples. We define the function T that returns logical tuples from a scene graph as:

$$T(G(c)) \triangleq O(c) \cup E(c) \cup K(c) \tag{2}$$

Each tuple contains either one, two or three elements, representing objects, attributes and relations, respectively. For example, the scene graph in Fig. 1 would be represented with the following tuples:

$$\begin{gathered}\{(\text{girl}), (\text{court}), (\text{girl}, \text{young}), (\text{girl}, \text{standing}) \\ (\text{court}, \text{tennis}), (\text{girl}, \text{on-top-of}, \text{court})\}\end{gathered}$$

Viewing the semantic propositions in the scene graph as a set of tuples, we define the binary matching operator $\otimes$ as the function that returns matching tuples in two scene graphs. We then define precision P, recall R, and $SPICE$ as:

$$P(c,S) = \frac{|T(G(c)) \otimes T(G(S))|}{|T(G(c))|} \tag{3}$$

$$R(c,S) = \frac{|T(G(c)) \otimes T(G(S))|}{|T(G(S))|} \tag{4}$$

$$SPICE(c,S) = F_1(c,S) = \frac{2 \cdot P(c,S) \cdot R(c,S)}{P(c,S) + R(c,S)} \tag{5}$$

where for matching tuples, we reuse the wordnet synonym matching approach of METEOR [13], such that tuples are considered to be matched if their lemmatized word forms are equal—allowing terms with different inflectional forms to match—or if they are found in the same wordnet sysnet.

Unlike *Smatch* [27], a recently proposed metric for evaluating AMR parsers that considers multiple alignments of AMR graphs, we make no allowance for partial credit when only one element of a tuple is incorrect. In the domain of image captions, many relations (such as *in* and *on*) are so common they arguably deserve no credit when applied to the wrong objects.

Being an F-score, SPICE is simple to understand, and easily interpretable as it is naturally bounded between 0 and 1. Unlike CIDEr, SPICE does not use cross-dataset statistics—such as corpus word frequencies—and is therefore equally applicable to both small and large datasets.

3.3 Gameability

Whenever the focus of research is reduced to a single benchmark number, there are risks of unintended side-effects [30]. For example, algorithms optimized for performance against a certain metric may produce high scores, while losing sight of the human judgement that the metric was supposed to represent.

SPICE measures how well caption generators recover objects, attributes and the relations between them. A potential concern then, is that the metric could be 'gamed' by generating captions that represent only objects, attributes and relations, while ignoring other important aspects of grammar and syntax. Because SPICE neglects fluency, as with n-gram metrics, it implicitly assuming that captions are well-formed. If this assumption is untrue in a particular application, a fluency metric, such as *surprisal* [31,32], could be included in the evaluation. However, by default we have not included any fluency adjustments as conceptually we favor simpler, more easily interpretable metrics. To model human judgement in a particular task as closely as possible, a carefully tuned ensemble of metrics including SPICE capturing various dimensions of correctness would most likely be the best.

4 Experiments

In this section, we compare SPICE to existing caption evaluation metrics. We study both system-level and caption-level correlation with human judgments. Data for the evaluation is drawn from four datasets collected in previous studies, representing a variety of captioning models. Depending on the dataset, human judgments may consist of either pairwise rankings or graded scores, as described further below.

Our choice of correlation coefficients is consistent with an emerging consensus from the WMT Metrics Shared Task [33,34] for scoring machine translation metrics. To evaluate system-level correlation, we use the Pearson correlation coefficient. Although Pearson's ρ measures linear association, it is smoother than rank-based correlation coefficients when the number of data points is small and systems have scores that are very close together. For caption-level correlation, we evaluate using Kendall's τ rank correlation coefficient, which evaluates the similarity of pairwise rankings. Where human judgments consist of graded scores rather than pairwise rankings, we generate pairwise rankings by comparing scores over all pairs in the dataset. In datasets containing multiple independent judgments over the same caption pairs, we also report inter-human correlation. We include further analysis, including additional results, examples and failure cases on our project page[2].

[2] http://panderson.me/spice.

Table 1. System-level Pearson's ρ correlation between evaluation metrics and human judgments for the 15 competition entries plus human captions in the 2015 COCO Captioning Challenge [6]. SPICE more accurately reflects human judgment overall (M1–M2), and across each dimension of quality (M3–M5, representing correctness, detailedness and saliency)

	M1		M2		M3		M4		M5	
	ρ	p-value	ρ	p-value	ρ	p-value	ρ	p-value	ρ	p-value
Bleu-1	0.24	(0.369)	0.29	(0.271)	0.72	(0.002)	−0.54	(0.030)	0.44	(0.091)
Bleu-4	0.05	(0.862)	0.10	(0.703)	0.58	(0.018)	−0.63	(0.010)	0.30	(0.265)
ROUGE-L	0.15	(0.590)	0.20	(0.469)	0.65	(0.006)	−0.55	(0.030)	0.38	(0.142)
METEOR	0.53	(0.036)	0.57	(0.022)	0.86	(0.000)	−0.10	(0.710)	0.74	(0.001)
CIDEr	0.43	(0.097)	0.47	(0.070)	0.81	(0.000)	−0.21	(0.430)	0.65	(0.007)
SPICE-exact	0.84	(0.000)	0.86	(0.000)	0.90	(0.000)	0.39	(0.000)	0.95	(0.000)
SPICE	**0.88**	(0.000)	**0.89**	(0.000)	**0.89**	(0.000)	**0.46**	(0.070)	**0.97**	(0.000)
M1	Percentage of captions evaluated as better or equal to human caption.									
M2	Percentage of captions that pass the Turing Test.									
M3	Average correctness of the captions on a scale 1–5 (incorrect - correct).									
M4	Average detail of the captions from 1–5 (lacking details - very detailed).									
M5	Percentage of captions that are similar to human description.									

4.1 Datasets

Microsoft COCO 2014. The COCO dataset [6] consists of 123,293 images, split into an 82,783 image training set and a 40,504 image validation set. An additional 40,775 images are held out for testing. Images are annotated with five human-generated captions (C5 data), although 5,000 randomly selected test images have 40 captions each (C40 data).

COCO human judgements were collected using Amazon Mechanical Turk (AMT) for the purpose of evaluating submissions to the 2015 COCO Captioning Challenge [6]. A total of 255,000 human judgments were collected, representing three independent answers to five different questions that were posed in relation to the 15 competition entries, plus human and random entries (17 total). The questions capture the dimensions of overall caption quality (M1 - M2), correctness (M3), detailedness (M4), and saliency (M5), as detailed in Table 1. For pairwise rankings (M1, M2 and M5), each entry was evaluated using the same subset of 1000 images from the C40 test set. All AMT evaluators consisted of US located native speakers, white-listed from previous work. Metric scores for competition entries were obtained from the COCO organizers, using our code to calculate SPICE. The SPICE methodology was fixed before evaluating on COCO. At no stage were we given access to the COCO test captions.

Flickr 8K. The Flickr 8K dataset [3] contains 8,092 images annotated with five human-generated reference captions each. The images were manually selected to focus mainly on people and animals performing actions. The dataset also contains graded human quality scores for 5,822 captions, with scores ranging from 1 ('the selected caption is unrelated to the image') to 4 ('the selected caption describes the image without any errors'). Each caption was scored by three expert human evaluators sourced from a pool of native speakers. All evaluated captions were sourced from the dataset, but association to images was performed using an image retrieval system. In our evaluation we exclude 158 correct image-caption pairs where the candidate caption appears in the reference set. This reduces all correlation scores but does not disproportionately impact any metric.

Composite Dataset. We refer to an additional dataset of 11,985 human judgments over Flickr 8K, Flickr 30K [4] and COCO captions as the composite dataset [35]. In this dataset, captions were scored using AMT on a graded correctness scale from 1 ('The description has no relevance to the image') to 5 ('The description relates perfectly to the image'). Candidate captions were sourced from the human reference captions and two recent captioning models [35,36].

PASCAL-50S. To create the PASCAL-50S dataset [12], 1,000 images from the UIUC PASCAL Sentence Dataset [37]—originally containing five captions per image—were annotated with 50 captions each using AMT. The selected images represent 20 classes including people, animals, vehicles and household objects.

The dataset also includes human judgments over 4,000 candidate sentence pairs. However, unlike in previous studies, AMT workers were not asked to evaluate captions against images. Instead, they were asked to evaluate caption triples by identifying 'Which of the sentences, B or C, is more similar to sentence A?', where sentence A is a reference caption, and B and C are candidates. If reference captions vary in quality, this approach may inject more noise into the evaluation process, however the differences between this approach and the previous approaches to human evaluations have not been studied. For each candidate sentence pair (B,C) evaluations were collected against 48 of the 50 possible reference captions. Candidate sentence pairs were generated from both human and model captions, paired in four ways: human-correct (HC), human-incorrect (HI), human-model (HM), and model-model (MM).

4.2 System-Level Correlation

In Table 1 we report system-level correlations between metrics and human judgments over entries in the 2015 COCO Captioning Challenge [6]. Each entry is evaluated using the same 1000 image subset of the COCO C40 test set. SPICE significantly outperforms existing metrics, reaching a correlation coefficient of 0.88 with human quality judgments (M1), compared to 0.43 for CIDEr and 0.53 for METEOR. As illustrated in Table 1, SPICE more accurately reflects human judgment overall (M1 - M2), and across each dimension of quality (M3 - M5, representing correctness, detailedness and saliency). Interestingly, only SPICE

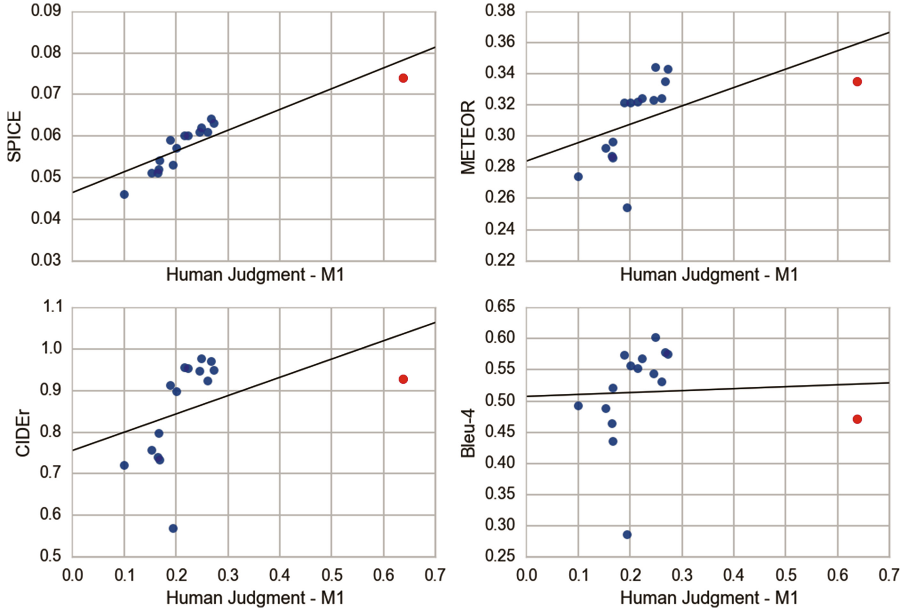


Fig. 3. Evaluation metrics vs. human judgments for the 15 entries in the 2015 COCO Captioning Challenge. Each data point represents a single model with human-generated captions marked in red. Only SPICE scores human-generated captions significantly higher than challenge entries, which is consistent with human judgment (Color figure online)

rewards caption detail (M4). Bleu and ROUGE-L appear to penalize detailedness, while the results for CIDEr and METEOR are not statistically significant.

As illustrated in Fig. 3, SPICE is the only metric to correctly rank human-generated captions first—CIDEr and METEOR rank human captions 7th and 4th, respectively. SPICE is also the only metric to correctly select the top-5 non-human entries. To help understand the importance of synonym-matching, we also evaluated SPICE using exact-matching only (SPICE-exact in Table 1). Performance degraded only marginally, although we expect synonym-matching to become more important when fewer reference captions are available.

4.3 Color Perception, Counting and Other Questions

Existing n-gram evaluation metrics have little to offer in terms of understanding the relative strengths and weaknesses, or error modes, of various models. However, SPICE has the useful property that it is defined over tuples that are easy to subdivide into meaningful categories. For example, precision, recall and F-scores can be quantified separately for objects, attributes and relations, or analyzed to any arbitrary level of detail by subdividing tuples even further.

To demonstrate this capability, in Table 2 we review the performance of 2015 COCO Captioning Challenge submissions in terms of *color perception*,

Table 2. F-scores by semantic proposition subcategory. SPICE is comprised of object, relation and attribute tuples. Color, count and size are attribute subcategories. Although the best models outperform the human baseline in their use of object color attributes, none of the models exhibits a convincing ability to count

	SPICE	Object	Relation	Attribute	Color	Count	Size
Human [6]	**0.074**	**0.190**	**0.023**	**0.054**	0.055	**0.095**	**0.026**
MSR [38]	0.064	0.176	0.018	0.039	**0.063**	0.033	0.019
Google [39]	0.063	0.173	0.018	0.039	0.060	0.005	0.009
MSR Captivator [40]	0.062	0.174	0.019	0.032	0.054	0.008	0.009
Berkeley LRCN [1]	0.061	0.170	**0.023**	0.026	0.030	0.015	0.010
Montreal/Toronto [2]	0.061	0.171	**0.023**	0.026	0.023	0.002	0.010
m-RNN [41]	0.060	0.170	0.021	0.026	0.038	0.007	0.004
Nearest Neighbor [42]	0.060	0.168	0.022	0.026	0.027	0.014	0.013
m-RNN [43]	0.059	0.170	0.022	0.022	0.031	0.002	0.005
PicSOM	0.057	0.162	0.018	0.027	0.025	0.000	0.012
MIL	0.054	0.157	0.017	0.023	0.036	0.007	0.009
Brno University [44]	0.053	0.144	0.012	0.036	0.055	0.029	0.025
MLBL [45]	0.052	0.152	0.017	0.021	0.015	0.000	0.004
NeuralTalk [36]	0.051	0.153	0.018	0.016	0.013	0.000	0.007
ACVT	0.051	0.152	0.015	0.021	0.019	0.001	0.008
Tsinghua Bigeye	0.046	0.138	0.013	0.017	0.017	0.000	0.009
Random	0.008	0.029	0.000	0.000	0.000	0.004	0.000

counting ability, and understanding of *size attributes* by using word lists to isolate attribute tuples that contain colors, the numbers from one to ten, and size-related adjectives, respectively. This affords us some insight, for example, into whether caption generators actually understand color, and how good they are at counting.

As shown in Table 2, the MSR entry [38] —incorporating specifically trained visual detectors for nouns, verbs and adjectives—exceeds the human F-score baseline for tuples containing color attributes. However, there is less evidence that any of these models have learned to count objects.

4.4 Caption-Level Correlation

In Table 3 we report caption-level correlations between automated metrics and human judgments on Flickr 8K [3] and the composite dataset [35]. At the caption level, SPICE achieves a rank correlation coefficient of 0.45 with Flickr 8K human scores, compared to 0.44 for CIDEr and 0.42 for METEOR. Relative to the correlation between human scores of 0.73, this represents only a modest improvement over existing metrics. However, as reported in Sect. 4.2, SPICE more closely

Table 3. Caption-level Kendall's τ correlation between evaluation metrics and graded human quality scores. At the caption-level SPICE modestly outperforms existing metrics. All p-values (not shown) are less than 0.001

	Flickr 8K [3]	Composite [35]
Bleu-1	0.32	0.26
Bleu-4	0.14	0.18
ROUGE-L	0.32	0.28
METEOR	0.42	0.35
CIDEr	0.44	0.36
SPICE	**0.45**	**0.39**
Inter-human	0.73	-

Table 4. Caption-level classification accuracy of evaluation metrics at matching human judgment on PASCAL-50S with 5 reference captions. SPICE is best at matching human judgments on pairs of model-generated captions (MM). METEOR is best at differentiating human and model captions (HM) and human captions where one is incorrect (HI). Bleu-1 performs best given two correct human captions (HC)

	HC	HI	HM	MM	All
Bleu-1	**64.9**	95.2	90.7	60.1	77.7
Bleu-2	56.6	93.0	87.2	58.0	73.7
ROUGE-L	61.7	95.3	91.7	60.3	77.3
METEOR	64.0	**98.1**	**94.2**	66.8	**80.8**
CIDEr	61.9	98.0	91.0	64.6	78.9
SPICE	63.3	96.3	87.5	**68.2**	78.8

approximates human judgment when aggregated over more captions. Results are similar on the composite dataset, with SPICE achieving a rank correlation coefficient of 0.39, compared to 0.36 for CIDEr and 0.35 for METEOR. As this dataset only includes one score per image-caption pair, inter-human agreement cannot be established.

For consistency with previous evaluations on the PASCAL-50S dataset [12], instead of reporting rank correlations we evaluate on this dataset using accuracy. A metric is considered accurate if it gives an equal or higher score to the caption in each candidate pair most commonly preferred by human evaluators. To help quantify the impact of reference captions on performance, the number of reference captions available to the metrics is varied from 1 to 48. This approach follows the original work on this dataset [12], although our results differ slightly which may be due to randomness in the choice of reference caption subsets, or differences in metric implementations (we use the MS COCO evaluation code).

On PASCAL-50S, there is little difference in overall performance between SPICE, METEOR and CIDEr, as shown in Fig. 4 left. However, of the four

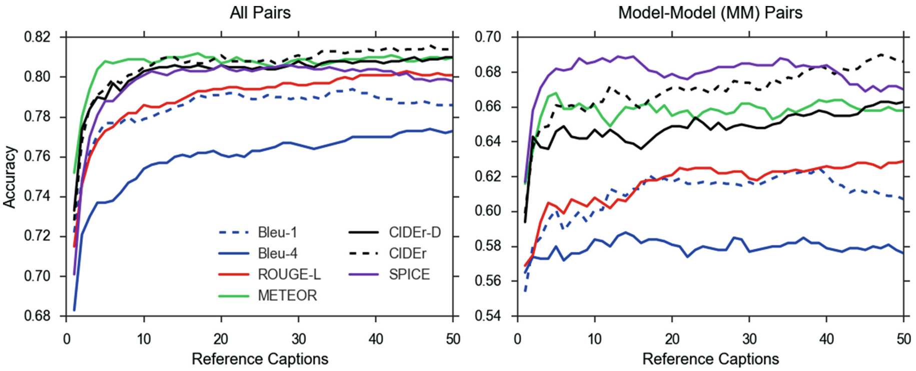


Fig. 4. Pairwise classification accuracy of automated metrics at matching human judgment with 1–50 reference captions

kinds of captions pairs, SPICE performs best in terms of distinguishing between two model-generated captions (MM pairs) as illustrated in Table 4 and Fig. 4 right. This is important as distinguishing better performing algorithms is the primary motivation for this work.

5 Conclusion and Future Work

We introduce SPICE, a novel semantic evaluation metric that measures how effectively image captions recover objects, attributes and the relations between them. Our experiments demonstrate that, on natural image captioning datasets, SPICE captures human judgment over model-generated captions better than existing n-gram metrics such as Bleu, METEOR, ROUGE-L and CIDEr. Nevertheless, we are aware that significant challenges still remain in semantic parsing, and hope that the development of more powerful parsers will underpin further improvements to the metric. In future work we hope to use human annotators to establish an upper bound for how closely SPICE approximates human judgments given perfect semantic parsing. We release our code and hope that our work will help in the development of better captioning models.

Acknowledgements. We are grateful to the COCO Consortium (in particular, Matteo R. Ronchi, Tsung-Yi Lin, Yin Cui and Piotr Dollár) for agreeing to run our SPICE code against entries in the 2015 COCO Captioning Challenge. We would also like to thank Sebastian Schuster for sharing the Stanford Scene Graph Parser code in advance of public release, Ramakrishna Vedantam and Somak Aditya for sharing their human caption judgments, and Kelvin Xu, Jacob Devlin and Qi Wu for providing model-generated captions for evaluation. This work was funded in part by the Australian Centre for Robotic Vision.

References

1. Donahue, J., Hendricks, L.A., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., Darrell, T.: Long-term recurrent convolutional networks for visual recognition and description. In: CVPR (2015)
2. Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A.C., Salakhutdinov, R., Zemel, R.S., Bengio, Y.: Show, attend and tell: Neural image caption generation with visual attention (2015). arXiv preprint arXiv:1502.03044
3. Hodosh, M., Young, P., Hockenmaier, J.: Framing image description as a ranking task: data, models and evaluation metrics. JAIR **47**, 853–899 (2013)
4. Young, P., Lai, A., Hodosh, M., Hockenmaier, J.: From image descriptions to visual denotations: new similarity metrics for semantic inference over event descriptions. TACL **2**, 67–78 (2014)
5. 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: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014, Part V. LNCS, vol. 8693, pp. 740–755. Springer, Heidelberg (2014)
6. Chen, X., Fang, H., Lin, T.Y., Vedantam, R., Gupta, S., Dollar, P., Zitnick, C.L.: Microsoft COCO captions: Data collection and evaluation server (2015). arXiv preprint arXiv:1504.00325
7. 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. PAMI **35**(12), 2891–2903 (2013)
8. Elliott, D., Keller, F.: Comparing automatic evaluation measures for image description. In: ACL, pp. 452–457 (2014)
9. Bernardi, R., Cakici, R., Elliott, D., Erdem, A., Erdem, E., Ikizler-Cinbis, N., Keller, F., Muscat, A., Plank, B.: Automatic description generation from images: a survey of models, datasets, and evaluation measures. JAIR **55**, 409–442 (2016)
10. Papineni, K., Roukos, S., Ward, T., Zhu, W.: Bleu: a method for automatic evaluation of machine translation. In: ACL (2002)
11. Lin, C.Y.: ROUGE: a package for automatic evaluation of summaries. In: ACL Workshop, pp. 25–26 (2004)
12. Vedantam, R., Zitnick, C.L., Parikh, D.: CIDEr: consensus-based image description evaluation. In: CVPR (2015)
13. Denkowski, M., Lavie, A.: Meteor universal: language specific translation evaluation for any target language. In: EACL 2014 Workshop on Statistical Machine Translation (2014)
14. Giménez, J., Màrquez, L.: Linguistic features for automatic evaluation of heterogenous MT systems. In: ACL Second Workshop on Statistical Machine Translation
15. 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)
16. Schuster, S., Krishna, R., Chang, A., Fei-Fei, L., Manning, C.D.: Generating semantically precise scene graphs from textual descriptions for improved image retrieval. In: EMNLP 4th Workshop on Vision and Language (2015)
17. Wang, C., Xue, N., Pradhan, S.: A transition-based algorithm for AMR parsing. In: HLT-NAACL (2015)
18. Lin, D., Fidler, S., Kong, C., Urtasun, R.: Visual semantic search: retrieving videos via complex textual queries. In: CVPR (2014)
19. Klein, D., Manning, C.D.: Accurate unlexicalized parsing. In: ACL (2003)

20. De Marneffe, M.C., Dozat, T., Silveira, N., Haverinen, K., Ginter, F., Nivre, J., Manning, C.D.: Universal stanford dependencies: a cross-linguistic typology. LREC **14**, 4585–4592 (2014)
21. Lo, C.k., Tumuluru, A.K., Wu, D.: Fully automatic semantic MT evaluation. In: ACL Seventh Workshop on Statistical Machine Translation (2012)
22. Pradhan, S.S., Ward, W., Hacioglu, K., Martin, J.H., Jurafsky, D.: Shallow semantic parsing using support vector machines. In: HLT-NAACL, pp. 233–240 (2004)
23. Ellebracht, L., Ramisa, A., Swaroop, P., Cordero, J., Moreno-Noguer, F., Quattoni, A.: Semantic tuples for evaluation of image sentence generation. In: EMNLP 4th Workshop on Vision and Language (2015)
24. Banarescu, L., Bonial, C., Cai, S., Georgescu, M., Griffitt, K., Hermjakob, U., Knight, K., Koehn, P., Palmer, M., Schneider, N.: Abstract meaning representation (AMR) 1.0 specification. In: EMNLP, pp. 1533–1544 (2012)
25. Flanigan, J., Thomson, S., Carbonell, J., Dyer, C., Smith, N.A.: A discriminative graph-based parser for the abstract meaning representation. In: ACL (2014)
26. Werling, K., Angeli, G., Manning, C.: Robust subgraph generation improves abstract meaning representation parsing. In: ACL (2015)
27. Cai, S., Knight, K.: Smatch: an evaluation metric for semantic feature structures. In: ACL (2), pp. 748–752 (2013)
28. Plummer, B.A., Wang, L., Cervantes, C.M., Caicedo, J.C., Hockenmaier, J., Lazebnik, S.: Flickr30k entities: collecting region-to-phrase correspondences for richer image-to-sentence models. In: CVPR, pp. 2641–2649 (2015)
29. Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.J., Shamma, D.A., Bernstein, M., Fei-Fei, L.: Visual genome: Connecting language and vision using crowdsourced dense image annotations (2016). arXiv preprint arXiv:1602.07332
30. Torralba, A., Efros, A.A.: Unbiased look at dataset bias. In: CVPR, June 2011
31. Hale, J.: A probabilistic earley parser as a psycholinguistic model. In: NAACL, pp. 1–8 (2001)
32. Levy, R.: Expectation-based syntactic comprehension. Cognition **106**(3), 1126–1177 (2008)
33. Stanojević, M., Kamran, A., Koehn, P., Bojar, O.: Results of the WMT15 metrics shared task. In: ACL Tenth Workshop on Statistical Machine Translation, pp. 256–273 (2015)
34. Machacek, M., Bojar, O.: Results of the WMT14 metrics shared task. In: ACL Ninth Workshop on Statistical Machine Translation, pp. 293–301 (2014)
35. Aditya, S., Yang, Y., Baral, C., Fermuller, C., Aloimonos, Y.: From images to sentences through scene description graphs using commonsense reasoning and knowledge (2015). arXiv preprint arXiv:1511.03292
36. Karpathy, A., Fei-Fei, L.: Deep visual-semantic alignments for generating image descriptions. In: CVPR (2015)
37. Rashtchian, C., Young, P., Hodosh, M., Hockenmaier, J.: Collecting image annotations using Amazon's Mechanical Turk. In: HLT-NAACL, pp. 139–147 (2010)
38. Fang, H., Gupta, S., Iandola, F.N., Srivastava, R., Deng, L., Dollar, P., Gao, J., He, X., Mitchell, M., Platt, J.C., Zitnick, C.L., Zweig, G.: From captions to visual concepts and back. In: CVPR (2015)
39. Vinyals, O., Toshev, A., Bengio, S., Erhan, D.: Show and tell: a neural image caption generator. In: CVPR (2015)
40. Devlin, J., Cheng, H., Fang, H., Gupta, S., Deng, L., He, X., Zweig, G., Mitchell, M.: Language models for image captioning: The quirks and what works (2015). arXiv preprint arXiv:1505.01809

41. Mao, J., Wei, X., Yang, Y., Wang, J., Huang, Z., Yuille, A.L.: Learning like a child: fast novel visual concept learning from sentence descriptions of images. In: CVPR, pp. 2533–2541 (2015)
42. Devlin, J., Gupta, S., Girshick, R.B., Mitchell, M., Zitnick, C.L.: Exploring nearest neighbor approaches for image captioning (2015). arXiv preprint arXiv:1505.04467
43. Mao, J., Xu, W., Yang, Y., Wang, J., Huang, Z., Yuille, A.: Deep captioning with multimodal recurrent neural networks (m-rnn) (2014). arXiv preprint arXiv:1412.6632
44. Kolár, M., Hradis, M., Zemcík, P.: Technical report: Image captioning with semantically similar images (2015). arXiv preprint arXiv:1506.03995
45. Kiros, R., Salakhutdinov, R., Zemel, R.S.: Multimodal neural language models. ICML **14**, 595–603 (2014)