Re-evaluating automatic metrics for image captioning

Mert Kilickaya, Aykut Erdem, Nazli Ikizler-Cinbis, and Erkut Erdem
Hacettepe University Computer Vision Lab
Dept. of Computer Engineering, Hacettepe University, Ankara, TURKEY
kilickayamert@gmail.com, {aykut,nazli,erkut}@cs.hacettepe.edu.tr

1. Introduction

Image captioning is the task of generating a human-like description for an image. It is a combination of two challenges: The model needs to understand the visual content of an image and generate a linguistic description of that content. The quality of a generated description can either be evaluated by humans or by the automatic metrics. The automatic metrics are of utmost importance to study the problem at scale. To that end, the researchers either borrowed available metrics from NLP community [7, 6, 2] or proposed problem specific metrics [8, 1]. As none of the existing metrics perform human-like evaluation quality [4], it is critical to compare and contrast the proposed metrics: In which case(s) the one is superior to another?

In this paper, we first briefly review existing metrics. Observing that none of the metrics considers the semantic side of the problem, we propose to fill this gap via the use of a document similarity metric named Word Mover's Distance [5]. Then, we proceed to explore the proposed metrics in terms of correlation with human judgements and whether their improvement is significant, their syntactic robustness, and finally their semantic robustness.

2. Metrics

All metrics compare a candidate image description c_i (generated by a model) with the reference sentences $S_i = \{s_{i1}, \ldots, s_{im}\}$ (provided by the dataset).

Metrics borrowed from other NLP tasks. BLEU [7] is one of the first metrics, proposed for machine translation, and defined as the geometric mean of n-gram precision scores multiplied by a brevity penalty for short sentences. ROUGE [6] is initially proposed for evaluation of summarization systems, and works by comparing overlapping n-grams, word sequences and word pairs. METEOR [2] is another machine translation metric, defined as the harmonic mean of precision and recall of uni-gram matches between sentences. Additionally, it makes use of paraphrase matching to handle synonyms.

Metrics developed for image captioning. CIDEr [8] is a recent metric proposed for evaluating the quality of image descriptions, taking into account the frequency of words and phrases in a specified corpus (*i.e.* highly frequent words contribute more to the score). SPICE [1] deviates from n-gram similarity focus of the previous metrics by defining similarity over scene-graph tuples. Scene-graph is a structured image representation taking into account the objects (*i.e.* man, dog), their modifiers (*i.e.* *white* dog) and their relations (*i.e.* man *with* dog). Then, the similarity between these entities are measured via WordNet, as is done by METEOR.

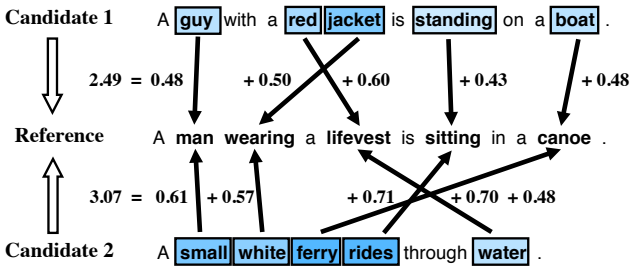

Figure 1. An illustration of the distance calculation of WMD metric comparing two candidate captions with a reference caption.

A new suggestion: Word mover's distance [5]. Two captions may not share the same words or any synonyms; yet they can be semantically similar. On the contrary, two captions may include similar objects, attributes or relations yet they may not be semantically similar. Metrics that are currently in use fail to correctly identify and assess the quality of such cases. To address this issue, we propose to use a recently introduced document distance measure called Word Mover's Distance (WMD) [5] for evaluating image captioning approaches. WMD casts the similarity between documents as an instance of Earth Mover's Distance, where the travel costs are calculated based on *word2vec* embeddings of the words (see Figure 1).

3. Analysis

In this part, we provide an analysis of the metrics.

Distractor Type		
	Gold Caption	a man wearing a life jacket is in a small boat on a lake with a ferry in view
	"Replace Scene"	a man wearing a life jacket is in a small boat on **takeoff** with a ferry in view
	"Replace Person"	**a woman in a blue shirt and headscarf** is in a small boat on a lake with a ferry in view
	"Share Person"	**a man** is selecting a chair from a stack under a shady awning
	"Share Scene"	a black and brown dog is playing on the ice at the edge of **a lake**

Figure 2. Distracted versions of a description for a sample image.

Table 1. Correlation analysis.

	BLEU	METEOR	ROUGE	CIDER	SPICE	WMD
Flickr-8k	0.44	0.58	0.44	0.56	**0.64**	0.60
Composite	0.38	**0.44**	0.39	0.42	0.42	0.43

Correlation and significance testing. A common way of assessing the performance of a new automatic image captioning metric is to analyze how well it correlates with the human judgements of description quality. In this paper, we provide Spearman correlations of each metric on 2 different datasets (Flickr-8k and Composite [1]) in Table 1. However, comparing the corresponding correlations relative to each other does not say much since they are both computed on the same dataset, and thus not independent. To address this issue, we use Williams significance testing previously suggested by [3], and found that improvement of each metric to another is significant.

Syntactic robustness. We show which metrics are sensitive to syntactic changes in a candidate caption c_i in Table 2. To that end, we apply three modifications to the caption (without modifying the meaning of the sentence), namely **synonym** (replace few words by their synonyms), **redundancy** (append a few redundant words the end of a caption) and **order** (change the order of a few words). Ideally the scores assigned by a metric should not change (indicated by $\leftrightarrow$), however many times most of the metrics are affected by these small changes (indicated by $\downarrow$)

Table 2. Syntactic analysis.

	BLEU	METEOR	ROUGE	CIDER	SPICE	WMD
Synonym	$\downarrow$	$\downarrow$	$\leftrightarrow$	$\downarrow$	$\downarrow$	$\downarrow$
Redundancy	$\leftrightarrow$	$\leftrightarrow$	$\leftrightarrow$	$\leftrightarrow$	$\leftrightarrow$	$\leftrightarrow$
Order	$\downarrow$	$\leftrightarrow$	$\downarrow$	$\downarrow$	$\leftrightarrow$	$\leftrightarrow$

Semantic robustness. As stated before, many metrics fail to account for the semantic side of the automatic evaluation. To quantitatively measure that, we designated a task, where in addition to an accurate candidate description $\boldsymbol{c_i}$, we are given a set of modified versions of it $\boldsymbol{m_i}$, by replacing scene (**R-scene**) or person (**R-person**) or sharing only scene (**S-scene**) or person (**S-person**) (see Figure 2). We measure the ability of each metric to distinguish accurate metric $\boldsymbol{c_i}$ from its modified version $\boldsymbol{m_i}$ in Table 3. Evaluation is conducted via accuracy (higher is better).

Table 3. Distraction analysis.

	BLEU	METEOR	ROUGE	CIDER	SPICE	WMD
R-scene	0.62	0.69	0.63	**0.83**	0.54	0.76
R-person	0.73	0.77	0.78	0.78	0.67	**0.80**
S-scene	0.79	0.85	0.79	0.81	0.70	**0.87**
S-person	0.78	0.85	0.78	0.83	0.67	**0.88**
Overall	0.73	0.79	0.75	0.81	0.65	**0.83**

4. Conclusion

This work proposed to consider semantic information in image caption evaluation procedure via WMD, and sketched new testing scenarios like significance testing, syntactic and semantic robustness tests for all the existing and upcoming metrics in the literature.

References

[1] P. Anderson, B. Fernando, M. Johnson, and S. Gould. SPICE: Semantic propositional image caption evaluation. In *ECCV*, 2016.

[2] S. Banerjee and A. Lavie. METEOR: An automatic metric for mt evaluation with improved correlation with human judgments. In *ACL Workshops*, 2005.

[3] Y. Graham and T. Baldwin. Testing for significance of increased correlation with human judgment. In *EMNLP*, 2014.

[4] M. Kilickaya, A. Erdem, N. Ikizler-Cinbis, and E. Erdem. Re-evaluating automatic metrics for image captioning. In *EACL*, 2017.

[5] M. J. Kusner, Y. Sun, N. I. Kolkin, and K. Q. Weinberger. From word embeddings to document distances. In *ICML*, 2015.

[6] C.-Y. Lin. ROUGE: A package for automatic evaluation of summaries. In *ACL-04 workshop*, 2004.

[7] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. BLEU: a method for automatic evaluation of machine translation. In *ACL*, 2002.

[8] R. Vedantam, C. Lawrence Zitnick, and D. Parikh. CIDEr: Consensus-based image description evaluation. In *CVPR*, 2015.