**Paper: Re-evaluating Automatic Metrics for Image Captioning**

Name: Milon Hossain

ID: 1610497042

**Abstract:** In recent years, the challenge of producing natural language explanations from images has gained a lot of interest. Consequently, it is becoming increasingly important to automatically test certain approaches to image captioning. In this article, through a series of deliberately planned tests, this system provides an in-depth review of the current image captioning metrics. In addition, here investigating the use of the newly proposed Word Mover’s Distance (WMD) software metric for image captioning purpose. Via detailed companions, precious and distraction-based assessments, for result outline the distinctions and correlations between measures and their relative robustness. This system also suggests that WMD presents clear advantages over other indicators.

**Introduction:** There has been a growing in research on integrating vision and language in natural language processing and computer vision communities. Image captioning, as one of the main problems in this emerging field, attempts to produce natural explanations of a given image. The recently proposed deep image captioning studies follow this interpretation and model the process via an encoder-decoder based architecture. These approaches have attained considerable success in the recent benchmarks such as FLICKR8K. With the size of the benchmark datasets becoming larger and larger, evaluating image captioning models has become increasingly important.

**Evaluation Metrics**: All the evaluation metrics except SPICE and WMD define the similarity over words of reference and candidate descriptions by considering different formulas. Now I am going to give a short description about different evaluation metrics which are important for our project.

1. **BLEU:** BLEU is one of the first metrics that have been in use for measuring similarity between two sentences.
2. **ROGUE:** ROGUE is initially proposed for evaluation of summarization systems, and this evaluation is done via comparing overlapping word sequence and word pairs.
3. **METEOR:** METEOR is another machine translation metric. It is defined as the harmonic mean of precision and recall of unigram matches between sentences.
4. **CIDEr:** CIDEr is a recent metrics proposed for evaluating the quality image descriptions. It measures the consensus between the candidate image description and the reference sentences.
5. **SPICE:** Another recently proposed metric for evaluating image caption similarity is SPICE. It is based on the agreement of the scene graph tuples of the candidate sentences and all reference sentences.
6. **WMD:** Two captions may not share the same words or any synonyms, yet they can be semantically similar. On the other hand, two captions may include similar objects, attributes or relations yet they may be not semantically similar. Metrics that are currently in use fail to correctly identify and assess the quality of such cases.

**Robustness:** The system evaluates the robustness of automatic image captioning metrics in this section. For this purpose, the system employs the binary forced choice task introduced to compare the existing image captioning models. For a given image, this task involves distinguishing a correct description from its slightly distracted incorrect versions. But in this system, a robust image captioning metrics should always choose the correct caption over the distracted ones.

**Dataset & Analyze:** To analyze statistical significance in the automatic metrics, this approach uses the publicly available FLICKR-8K and COMPOSITE dataset. In this experiment, they first lowercase and tokenize the candidate and reference captions using ptbtokenizer.py script from MSCOCO evaluation tools. FLICKR-8K dataset contains quality judgement for 5822 candidate sentences for the image in its test set. COMPOSITE dataset contains human judgements for 11,985 candidate captions for the sub sets of FLICKR-8K and FLICKR-30K and MSCOCO datasets.

**Discussion:** CPICE, METEOR and WMD give the best performance in terms of the correlation analysis against human judgements. CIDEr and WMD provide the best classification score for accuracy experiments. Moreover, CIDEr, METEOR and WMD are found to be less affected by distractors.

**Conclusion:** A careful evaluation of the automatic image captioning metrics and propose to use WMD, which utilizes word-to-vector embedding of the words to compute semantic similarity of sentences. Also highlight the drawbacks of the existing metrics and empirically show that they are significantly different from each other.

**Reference:**

1. <https://arxiv.org/abs/1612.07600>

<https://arxiv.org/abs/1612.07600>