UMIC: An Unreferenced Metric for Image Captioning via Contrastive Learning

Hwanhee Lee¹, Seunghyun Yoon², Franck Dernoncourt², Trung Bui², Kyomin Jung¹

¹Seoul National University, Seoul, Korea
²Adobe Research, San Jose, CA, USA
The metric score for a given candidate caption varies significantly depending on the reference type due to the diverse nature of image captions.

Reference-based metrics usually require multiple references, which are difficult to obtain, to get meaningful score.

### Difficulty of Image Caption Evaluation

**Ref 1:** A dog standing in the snow with a stick in its mouth.
**Ref 2:** A little dog holding sticks in its mouth.
**Candidate:** A dog standing on the snow with a dog

*CIDEr with Ref 1:* 3.166  
*CIDEr with Ref 2:* 0.281

**Human Judgments:** 1.875 out of 5 (Average of 5 people)
Reference-less Metrics

- Humans do not require reference captions when evaluating the captions.

- We can simply argue that the candidate caption is wrong *(only one dog in the picture)*

**Ref 1:** A dog standing in the snow with a stick in its mouth.

**Ref 2:** A little dog holding sticks in its mouth.

**Candidate:** A dog standing on the snow with a dog

CIDEr with Ref 1: 3.166

CIDEr with Ref 2: 0.281

**Human Judgments:** 1.875 out of 5 *(Average of 5 people)*
Overall Training Procedure of UMIC

- We fine-tune UNITER via contrastive learning, where the model is trained to compare and discriminate the ground-truth captions and diverse synthetic negative samples like the example.
Overall Training Procedure of UMIC

1) \([CLS], i_1, ..., i_N, x_1, ..., x_T = \text{UNITER}(I, X)\)

2) \(S(I, X) = \text{sigmoid}(W_i[CLS] + b)\)

3) \(\text{Loss} = \max(0, M - (S(I, X) - S(I, \hat{X})))\)

---

\(I\) : Image  
\(x\) : Positive caption  
\(\hat{x}\) : Negative caption  
\(S_x\) : Score of positive caption  
\(S_{\hat{x}}\) : Score of negative caption
Generating Negative Captions

- We prepare the negative captions that can represent most of the undesirable cases in captioning, such as relevant but have wrong keyword, irrelevant to the image, grammatically incorrect.

1) Substituting Keywords: substitute 30% of the words (verb, adjective, noun) in the reference captions

2) Random Captions: sample captions from other images utilize the captions of the images similar to the given images

3) Repetition & Removal: repeat or remove some words in the reference captions with a probability of 30%

4) Word Order Permutation: changing the word order of the reference captions
Generating Negative Captions

• We prepare the negative captions that can represent most of the undesirable cases in captioning, such as relevant but have wrong keyword, irrelevant to the image, grammatically incorrect.

Original: a woman hugging a girl who is holding a suitcase
Substitution: a boy hugging a girl who is holding a suitcase
Random(Hard Negative): a very small cute child by a suitcase
Repetition & Removal: a woman hugging a girl is holding a suitcase
Problems in Previous Benchmark Datasets

When evaluating the metric’s performance, it is required to compare the correlations between human judgments and the metric’s evaluation score for given datasets.

We investigate the human judgments in Flickr8k and Composite, and visualize the distributions of judgment scores for two popular datasets, Flickr8k and Composite, and find several problems.
Problems in Previous Benchmark Datasets

Flickr8k
- Most of the scores are less than 0.2 since the candidate captions were sampled by an image retrieval system from a reference caption pool, not model-generated captions. In other words, most captions are not related to images and differ significantly from the model-generated captions.

Composite
- Most of the scores are placed near 0 or 1.
- Captions for this dataset were generated by the old model.
CapEval1k: Introducing New Benchmark Dataset

• We introduce a new dataset CapEval1k, which is composed of human judgments for the model-generated captions from four recently proposed models.

• We ask each annotator to evaluate the captions by considering three dimensions: fluency, relevance, descriptiveness and assign overall score.

• CapEval1k contains the annotators’ comprehensive judgment across multiple dimensions in evaluating the quality of the generated captions, so we can see that the score distribution score is not concentrated in a particular area.
CapEval1k: Instructions to Workers

Read the instructions and examples below and evaluate candidate captions (Click to collapse)

Evaluate the captions comparing them with reference captions and considering "fluency", "relevance" and "descriptiveness".

[Image]

Caption 1: a couple of ducks swimming in the water

Caption 2: two ducks swimming in the water in a body of water

Caption 3: three ducks are swimming in the water

Caption 4: three ducks swimming in the water

[Reference Captions]
Ref1: two ducks floating together on a body of water.
Ref2: two ducks are swimming in the green colored pond.
Ref3: two canadian geese swim in a green pond.
Ref4: two ducks swim in a pond with green water.
Ref5: two swam swimming next to each other on a lake.

[Overall]
In this task, you are supposed to evaluate the quality of the caption for the given image.
Please read the image and the captions carefully and assign the score for each caption considering three criterias.

[Instructions]
1. Read the candidate captions, reference captions and see the given image.
2. Evaluate the four candidate captions considering three criterias (refer to the negative examples below) and comparing them to the reference captions
- Note that reference captions are not always perfect.

Criteria & Common negative examples in the captions
Please consider 3 things comprehensively and rate the overall score for the capture.

(1) Fluency
Whether the caption is fluent, natural and grammatically correct
Ex: Grammatically correct but strange a plate of food and food

(2) Relevance
Whether the sentence correctly describes the visual content and be closely relevant to the image.
Ex: Relevant/Minor Mistake: relevant but tiny parts are wrong a plate of fruits and a crepe on a grey dish

(3) Descriptiveness
Whether the sentence is a precise, informative caption that describes important details of the image.
Ex: Too General Captin a plate of fruits

<Annotation Interface>

<Full Guideline>
## Experimental Results

<table>
<thead>
<tr>
<th>Metric</th>
<th>Flickr8k</th>
<th>Composite</th>
<th>CapEval1k</th>
<th>PASCAL50s</th>
</tr>
</thead>
<tbody>
<tr>
<td>BLEU-1</td>
<td>0.274</td>
<td>0.406</td>
<td>0.233</td>
<td>74.3</td>
</tr>
<tr>
<td>BLEU-4</td>
<td>0.286</td>
<td>0.439</td>
<td>0.238</td>
<td>73.4</td>
</tr>
<tr>
<td>ROUGE-L</td>
<td>0.300</td>
<td>0.417</td>
<td>0.220</td>
<td>74.9</td>
</tr>
<tr>
<td>METEOR</td>
<td>0.403</td>
<td>0.466</td>
<td>0.288</td>
<td>78.5</td>
</tr>
<tr>
<td>CIDEr</td>
<td>0.419</td>
<td>0.473</td>
<td>0.307</td>
<td>76.1</td>
</tr>
<tr>
<td>SPICE</td>
<td>0.457</td>
<td>0.486</td>
<td>0.279</td>
<td>73.6</td>
</tr>
<tr>
<td>BERTScore</td>
<td>0.396</td>
<td>0.456</td>
<td>0.273</td>
<td>79.5</td>
</tr>
<tr>
<td>BERT-TBR</td>
<td>0.467</td>
<td>0.439</td>
<td>0.257</td>
<td>80.1</td>
</tr>
<tr>
<td>VBTScore</td>
<td>0.525</td>
<td>0.514</td>
<td>0.352</td>
<td>79.6</td>
</tr>
<tr>
<td>VIFIDEL</td>
<td>0.336</td>
<td>0.191</td>
<td>0.143</td>
<td>70.0</td>
</tr>
<tr>
<td>UMIC</td>
<td>0.468</td>
<td>0.561</td>
<td>0.328</td>
<td>85.1</td>
</tr>
<tr>
<td>UMIC_c</td>
<td>0.431</td>
<td>0.554</td>
<td>0.299</td>
<td>84.7</td>
</tr>
</tbody>
</table>

- We show that although UMIC does not utilize any reference captions, UMIC outperforms most of the baseline metrics in all of the datasets that depend on multiple references.

**Flickr8k, Composite, CapEval1k:** Kendall Correlation Coefficient

**PASCAL50s:**
Accuracy of matches between human judgments for comparing two candidate captions

*UMIC\_c*: UMIC without contrastive learning (i.e. UNITER)
Example

References
- two giraffes standing next to each other in a field.
- two giraffes are climbing a hill with mountains in the background.

Candidate
- three giraffes standing in a field of grass

| BLEU1: 0.324 | ROUGE-L: 0.320 | METEOR: 0.173 | CIDER: 0.866 |
| SPICE: 0.289 | UMIC: 0.352 | UMIC_{-C}: 0.770 | Human: 0.200 |

References
- a person breaking a bottle with a baseball bat
- a boy in yellow shirt swinging a baseball bat

Candidate
- a man swinging a baseball bat at a ball

| BLEU1: 0.360 | ROUGE-L: 0.354 | METEOR: 0.176 | CIDER: 1.205 |
| SPICE: 0.192 | UMIC: 0.094 | UMIC_{-C}: 0.062 | Human: 0.450 |

• **Case1 – Good Case**: UMIC detects the wrong keyword “three” in the candidate caption and give lower score.

• **Case2 - Error Case**: UMIC could not recognize the important object like the “baseball bat” and UMIC outputs very low score compared to human judgment.
Closing Remarks

• We propose UMIC, an unreferenced metric that does not require any reference captions for image captioning task through contrastive learning to UNITER.

• We propose a new benchmark dataset for image captioning that relieve the issues (e.g. biased distributions) in previous datasets.

• Experimental results on four benchmark datasets, including our new dataset, show that UMIC outperforms most of the previous metrics that require multiple references.

Code: https://github.com/hwanheeelee1993/UMIC