TIGEr: Text-to-Image Grounding for Image Caption Evaluation

Ming Jiang¹, Qiuyuan Huang², Lei Zhang², Xin Wang³, Pengchuan Zhang², Zhe Gan², Jana Diesner¹, Jianfeng Gao²

¹. University of Illinois at Urbana-Champaign  ². Microsoft Research  ³. University of California Santa Barbara

Motivation and Contribution

• Metrics based on pure text-level comparison and face the challenge of language ambiguity.
• Propose a novel automatic evaluation metric called TIGEr.
  - Consider both image and human-generated references.
  - Measure the consistency with human attention distribution among image regions.

TIGEr Framework

Data Encoding
  • Region-level & Word-level embedding vectors

Text-to-Image Grounding
  • Grounding a caption into each image region.

[(Reference vs. Candidate) | Image]
  • RRS: how similar is the order of image regions based on grounding weights?
  • WDS: how similar is the attention distributed by a caption among image regions?

TIGEr
  • Average value of RRS and WDS

TIGEr Workflow

• Encoding images and texts by a pre-trained Bottom-Up Attention and a RNN model.
• Grounding texts and images by a pre-trained SCAN model.
• Calculating RRS based on Normalized Discounted Cumulative Gain (NDCG).
• Measuring WDS based on KL Divergence.

Metric Performance

• TIGEr achieved a noticeable improvement in the assessment of caption quality on three benchmark datasets.
• Identifying irrelevant human-written captions in HI is relatively easy for all metrics, while judging the quality of two correct human-annotated captions in HC is more difficult than other comparison groups.
• Given the change of reference sizes, TIGEr achieves a higher judgment accuracy and more stable performance.

<table>
<thead>
<tr>
<th>Composite</th>
<th>Flickr8K</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\rho)</td>
<td>(\rho)</td>
</tr>
<tr>
<td>BLEU-1</td>
<td>0.395</td>
</tr>
<tr>
<td>ROUGE-L</td>
<td>0.429</td>
</tr>
<tr>
<td>METEOR</td>
<td>0.472</td>
</tr>
<tr>
<td>CIDEr</td>
<td>0.514</td>
</tr>
<tr>
<td>SPICE</td>
<td>0.553</td>
</tr>
<tr>
<td>Ours</td>
<td>0.403</td>
</tr>
<tr>
<td>RRS</td>
<td>0.497</td>
</tr>
<tr>
<td>WDS</td>
<td>0.528</td>
</tr>
<tr>
<td>TIGEr</td>
<td>0.497</td>
</tr>
</tbody>
</table>

Caption-level correlation between metrics and human grading scores in Composite and Flickr8K dataset by using Kendall’s and Spearman’s rank. All \(p\) values < 0.01.

Analysis

• Image region has a higher grounding weight with the corresponding caption than other unrelated regions.
• Text-to-image grounding is more challenging at action-level compared to object-level.
• Reference captions may not fully cover visual information and TIGEr can measure a caption quality by considering the semantic information of image contents.
• Human interpretation inspired by the image is hard to be judged by an automatic evaluation metric.

Related Resource


Github Link: https://github.com/SeleenaJM/CapEval