VIFIDEL: Evaluating the Visual Fidelity of Image Descriptions
Pranava Madhyastha, Josiah Wang, Lucia Specia

Context: Evaluating Image Descriptions

- Existing automatic metrics conflate various criteria implicitly into a single score
- Our contribution: An image-aware metric named VIFIDEL
  1. Measures faithfulness of description w.r.t. the image
  2. Explicitly takes image content into account
  3. Also works in the absence of reference descriptions!

Claim: VIFIDEL is useful for fine-grained measurement of descriptions

Formally

For image I and description S:

\[ \text{VIFIDEL}(I, S) = \exp \left( - \min_{T \in \mathcal{T}} \sum_{i,j} T_{ij} \cos(x_i, x_j) \right) \]

where transport matrix \( T_{ij} \) contains information about the proportion of semantic content from the image to the description; \( \cos = \text{weighted Euclidean distance} \).

- To weight importance of word \( k \) with a penalty according to human references:
  \[ p_k = \frac{1}{M} \sum_{i} \left( 1 - \max_j \{ \rho(x_i, x_j) \} \right) \]

where \( \{ x_j \} \) is the set of content words in the rth reference for image I; \( x_i \) is the word embedding for word t;

- The cost (weighted according to human references) is:
  \[ \text{cost}(r, j) = \| x_i - p_k x_j \|^2 \]

Properties

- Semantic matching instead of string matching
- Scores even in the absence of references
- Highly scalable compared to SPICE (dependent on linguistic resources)
- Complements fluency-based metrics
- References are only used to weigh the importance of objects and words
- Extendable with other attributes including the environment
- Language agnostic
- Implementation is open source (QR code below)

VIFIDEL in Brief

- Extension of Wasserstein distance with weighted Euclidean distance
- Uses information from the images in the form of detected objects
- Consensus-based scores for multiple references

At a Glance

- Weights are computed for \textit{encyclopedias}, \textit{cat} and \textit{books}
- The word \textit{cat} has a low penalty score
- The penalty scores are then used as weights to compute the cost.

Evaluation

Accuracy as a function of # References

- VIFIDEL is more stable and consistently outperforms other metrics for all numbers of references.

Comparison

<table>
<thead>
<tr>
<th>References</th>
<th>( 0 )</th>
<th>1</th>
<th>5</th>
<th>48</th>
</tr>
</thead>
<tbody>
<tr>
<td>BLEU(_1)</td>
<td>0.58</td>
<td>0.61</td>
<td>0.62</td>
<td></td>
</tr>
<tr>
<td>BLEU(_4)</td>
<td>0.56</td>
<td>0.59</td>
<td>0.60</td>
<td></td>
</tr>
<tr>
<td>ROUGE(_L)</td>
<td>0.58</td>
<td>0.61</td>
<td>0.64</td>
<td></td>
</tr>
<tr>
<td>METEOR</td>
<td>0.62</td>
<td>0.66</td>
<td>0.68</td>
<td></td>
</tr>
<tr>
<td>CIDEr</td>
<td>0.60</td>
<td>0.66</td>
<td>0.68</td>
<td></td>
</tr>
<tr>
<td>SPICE</td>
<td>0.65</td>
<td>0.69</td>
<td>0.69</td>
<td></td>
</tr>
<tr>
<td>WMD(_\text{ent})</td>
<td>0.66</td>
<td>0.70</td>
<td>0.70</td>
<td></td>
</tr>
<tr>
<td>WMD(_\text{ent})</td>
<td>0.66</td>
<td>0.66</td>
<td>0.66</td>
<td></td>
</tr>
<tr>
<td>LM</td>
<td>0.54</td>
<td>0.54</td>
<td>0.54</td>
<td></td>
</tr>
<tr>
<td>VIFIDEL(_\text{gold})</td>
<td>0.68</td>
<td>0.69</td>
<td>0.70</td>
<td>0.71</td>
</tr>
<tr>
<td>VIFIDEL(_\text{gold})</td>
<td>0.69</td>
<td>0.71</td>
<td>0.72</td>
<td>0.71</td>
</tr>
<tr>
<td>VIFIDEL(_\text{gold})+LM</td>
<td>0.69</td>
<td>0.70</td>
<td>0.71</td>
<td>0.71</td>
</tr>
<tr>
<td>VIFIDEL(_\text{gold})+CIDEr</td>
<td>0.69</td>
<td>0.71</td>
<td>0.72</td>
<td>0.72</td>
</tr>
</tbody>
</table>