**Contributions**
- SBU Captured Photo Dataset: A large novel data set containing 1 million images from the web with associated captions written by people, filtered so that the descriptions are likely to refer to visual content. ([http://tamaraberg.com/sducaptions](http://tamaraberg.com/sducaptions))
- A description generation method that utilizes global image representations to retrieve and transfer captions from our data set to a query image.
- A description generation method that utilizes both global representations and direct estimates of image content (objects, actors, stuff, attributes, and scenes) to produce relevant image descriptions.

**Method overview**

**SBU Captured Photo Dataset**

- The bridge over the lake on Suzhou Street.
- The view from the 13th floor of an apartment building in Seattle.
- The Gherkin.
- The old Premium Oil Co. sign in downtown Calais Beach.
- The view from the 13th floor of an apartment building in Seattle.
- The bottom end of a kilometre of water.
- The face of the RCA building.
- The boat ended up a kilometre from the water.

**Im2Text: Describing Images Using 1 Million Captioned Photographs**

**Vicente Ordonez, Girish Kulkarni, Tamara L. Berg**

**Stony Brook University**

**High level information**

- **Objects:** 80 object categories using part-based deformable models and compute distances with objects detected in the query image based on visual attributes and raw visual descriptions.
- **People/Actions:** Detect people and pose using state-of-the-art methods and compute person similarly using an attribute-based representation of pose.

**Dataset size**

- Query Image
- 1,000
- 10,000
- 100,000
- 1,000,000

**Good results**

- Incorrect objects Incorrect context Completely wrong
- The sky is blue over the Breakers.
- The sun was coming through at the time I was taking the shot.
- Stone cross in the middle of beautiful desert.
- A woman sits in a chair by the river.
- A female mallard duck in the lake.
- The old Premium Oil Co. sign in downtown Calais Beach.

**Bad results**

- Incorrect objects Incorrect context Completely wrong
- The sun was coming through at the time I was taking the shot.
- Our dog Zoe in her bed.
- The boat ended up a kilometre from the water.
- The view from the 13th floor of an apartment building in Seattle.

**BLEU score evaluation**

<table>
<thead>
<tr>
<th>Method</th>
<th>BLEU score</th>
<th>1</th>
<th>5</th>
<th>10</th>
<th>100</th>
<th>1,000,000</th>
</tr>
</thead>
<tbody>
<tr>
<td>Global matching (1k)</td>
<td>0.285</td>
<td>0.067</td>
<td>0.040</td>
<td>0.001</td>
<td>0.000</td>
<td></td>
</tr>
<tr>
<td>Global matching (10k)</td>
<td>0.285</td>
<td>0.067</td>
<td>0.040</td>
<td>0.001</td>
<td>0.000</td>
<td></td>
</tr>
<tr>
<td>Global matching (100k)</td>
<td>0.285</td>
<td>0.067</td>
<td>0.040</td>
<td>0.001</td>
<td>0.000</td>
<td></td>
</tr>
<tr>
<td>Global matching (1M)</td>
<td>0.285</td>
<td>0.067</td>
<td>0.040</td>
<td>0.001</td>
<td>0.000</td>
<td></td>
</tr>
</tbody>
</table>

**Human evaluation**

In addition, we propose a new evaluation task where a user is presented with two photographs and one caption. The user must assign the caption to the most relevant image. For evaluation we use a query image, a random image and a generated caption.

**Method**

- Original human caption
- Top captions
- Best from our top-4 captions

**Voted success rate**

- 96.0%
- 65.7%
- 92.7%