### Problem Statement

Multimodal documents on the web contain images placed at meaningful locations within the textual narrative. The image captions are commonly conditioned on the surrounding text. We study the generation of such **Contextual Captions**, distinct from conventional image captioning.

### Model Formulation

- **Text Encoder**: BiLSTM
- **Image Encoder**: ResNet152
- **Decoder**: LSTM, attention-weighted sum of encoder states concatenated to current state to incorporate contextual information.

Mathematically:

\[
\mathcal{L}(\theta) = \sum_{t=1}^{N} - \log p(w^{c}_{t} | w^{c}_{1}, \ldots, w^{c}_{t-1}, I, P; \theta)
\]

- **Output of BiLSTM**: \(\tilde{g}_t = \sum_{i=1}^{M} \alpha^t_i g_i\)
- **Attention weights**: \(\alpha^t_i = \frac{v^\top(W_g g_i + W_h h_t + b)}{\sum_{i=1}^{M} v^\top(W_g g^i + W_h h_t + b)}\)

### Dataset

- **Novel Contextual Captioning dataset**:
  - Data scraping from subreddit /r/pics
  - Domain-agnostic posts
  - 250,000 posts spanning one year
  - Post: 1 image, caption, 1-10 comments
  - Captions contain 10.6 words on avg.
  - Concatenated comments serve as image context or associated paragraph
  - Paragraphs contain 59.2 words on avg.

- Data splits based on Named Entities in image captions:
  - 137,732 samples with NE
  - 104,653 samples without NE
  - Additional splits ensuring overlap between context and caption

**SpaCy** is leveraged to detect 14 types of NE in image captions.

### Results

**Table: Quantitative evaluation of baselines and Contextual Captioning on standard text similarity measures.**

<table>
<thead>
<tr>
<th></th>
<th>BLEU-1</th>
<th>ROUGE-L</th>
<th>CIDEr</th>
<th>SPICE</th>
<th>SemSim</th>
</tr>
</thead>
<tbody>
<tr>
<td>Image-only</td>
<td>7.80</td>
<td>7.50</td>
<td>0.38</td>
<td>0.16</td>
<td>0.76</td>
</tr>
<tr>
<td>Text-only</td>
<td>6.87</td>
<td>6.54</td>
<td>0.61</td>
<td>0.36</td>
<td>0.72</td>
</tr>
<tr>
<td>Contextual</td>
<td>9.30</td>
<td>9.68</td>
<td>0.78</td>
<td>0.50</td>
<td>0.77</td>
</tr>
</tbody>
</table>

- Contextual Captions capture information from both visual and textual modalities.
- They are linguistically rich compared to text-only and image-only captions.

### References


[https://github.com/Sreyasi/contextual_captions](https://github.com/Sreyasi/contextual_captions)