Motivation & Objective

**Motivation**: large amount of pictures are produced and stored every day, from medical images to personal photos; in this context, automated caption-image retrieval is becoming an increasingly attractive feature to search inside these enormous databases of images.

**Goal**: accurate retrieval of images given an input description.

Dataset

We train the models using the Microsoft COCO dataset [1]:
- 123,287 images: 113,287 for training, 5,000 each for validation and test;
- 5 human-annotated captions per image

Example:
- Three teddy bears laying in bed under the covers.
- A group of stuffed animals sitting next to each other in bed.
- A white beige and brown baby bear under a beige white comforter.
- A trio of teddy bears bundled up on a bed.
- Three stuffed animals lay in a bed cuddled together.

Evaluation metric on the test set:

\[ R@K = \frac{\text{#correct retrievals among top } K \text{ images}}{\text{#text queries}} \]

Results

<table>
<thead>
<tr>
<th>Model</th>
<th>R@10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>65.5%</td>
</tr>
<tr>
<td>Order embedding</td>
<td>75.7%</td>
</tr>
<tr>
<td>Order embedding + GloVe</td>
<td>78.0%</td>
</tr>
</tbody>
</table>

Conclusions and future work

- By using pre-trained word vectors and fine-tuning on the present problem, we were able to achieve a R@10 of 78.0%.
- Future work will focus on finishing the dynamic attention mechanism and context-aware retrieval.