Detecting Off-topic Responses to Visual Prompts
Marek Rei
University of Cambridge

Image Relevance Detection

• Given an image, evaluate the relevance of a text to that image.

  An astronaut is celebrating on Mars. Two signs are sticking out of the sand. The sun is setting behind mountains. The whales are breaching the surface.

• Important for automated essay scoring and high-stakes testing.

Relevance Model

• Text is represented with word embeddings \([w_1, w_2, ..., w_N]\) and composed to a single vector \(u = h_N\) using an LSTM:
  \[ h_n = LSTM(w_n, h_{n-1}) \]

• The image is represented with vector \(x\), which we extract from a pre-trained GoogLeNet image recognition network.

• The model first reads the text, then decides which parts of the image are relevant to the task, using gating:
  \[ z = \sigma(uW_z + b_z) \]

• The image vector is then mapped to a new space which is specialised for relevance scoring:
  \[ v = tanh(x'W_z) \]

• The final relevance score is given as a dot product of the text vector \(u\) and the image vector \(v\):
  \[ score_{rel}(u, v) = uv \]

Optimisation

• The model is optimised using cross-entropy over all examples in the minibatch:
  \[ Loss_{ce} = - \sum_i log\left(\frac{\exp(score_{rel}(u_i, v_i))}{Z}\right) \]

Analysis

• Looking at examples that the system incorrectly classifies as unrelated.

• The uses of rare terms and unusual images are potential sources of confusion for the model.

• Visualising the gating vector for two different sentences, values close to 0 are represented with white.

Evaluation

• Training on 31,014 images from the Flickr30k captioning dataset.

• Evaluating on a dataset of 543 answers written by language learners.

<table>
<thead>
<tr>
<th>Learner texts</th>
<th>Flickr30k</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACC</td>
<td>AP</td>
</tr>
<tr>
<td>LSTM-COS</td>
<td>68.2</td>
</tr>
<tr>
<td>+ gating</td>
<td>69.6</td>
</tr>
<tr>
<td>+ cross-ent</td>
<td>71.1</td>
</tr>
<tr>
<td>+ dropout</td>
<td>75.4</td>
</tr>
</tbody>
</table>

Conclusion

• An automated system can reliably evaluate the relevance of a written text to a given image.

• Texts and images can be mapped into a shared semantic space for comparison.

• Each of the modifications gives a consistent improvement: gating the image based on the text, applying dropout, and jointly optimising all examples in the minibatch.

Examples

In this picture there are lot of people and each one has a different attitude. 0.65
In the foreground, people are waiting for the green light in order to cross the street. 0.81
Generally speaking, the picture is full of bright colours and it conveys the idea of crowded city. 0.63
Looking at this pictures reminds me of the time I went scuba diving in the sea. -2.38
You swim to the surface and you see the sunlight coming nearer and nearer until you get out. -1.70