CS288 Natural Language Processing:

Homework5 Report

Name: Diyun Lu SID: 3039752224

1 Part 1: Image and Text Retrieval

1.1. Analysis on Retrieval Failures

There are 92 failures using captions and 127 failures using descriptions where 39 are in common. I looked into the first 10 examples where they both made mistakes retrieving images. I observed that there were many charts and maps wrongly retrieved. The model failed to correspond well with specific content in the visual representation and words in the description or caption. For the same examples,

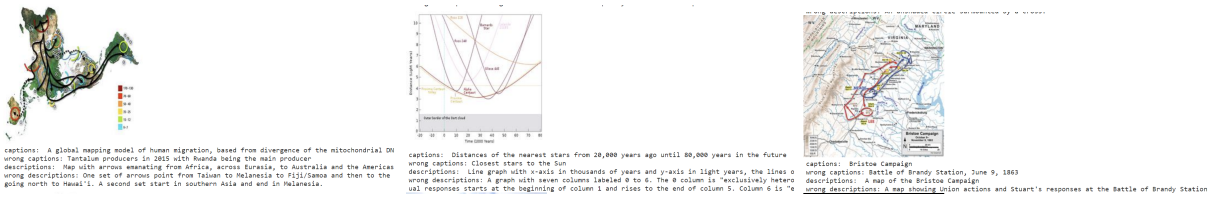

Fig. 1: Wrongly Retrieved Images

I've noticed that errors in captions are more coarse-grained, whereas errors in descriptions are more fine-grained. That means in the wrong descriptions, there are many errors in describing the specific features while the wrong captions got the whole idea wrong. It is somehow aligned with common sense that captions are more general while descriptions provide more details.

1.2. Analysis on Different Metrics

For the 3 different metrics we implemented, they have different focuses. MRR highlights the importance of quickly providing the most relevant information, Top-1 has a strict judgment, and Top-k provides a broader perspective by considering multiple predictions. From the accuracy results of different metrics on the batched outputs of captions and descriptions, we can see the big differences between using Top-1 and Top-k metrics. Though the model can reach 90%+ using Top-k, it does not mean it is perfectly well. MRR is rather more comprehensive than other metrics by not only focusing on how high the ranking is.

1.3. Analysis on Different Scores

I stored the difference between the score of using captions and descriptions every time when captions outperformed descriptions. Then I sorted and got the top 10 images with the highest different scores. I observed many short descriptions that intuitively lead to bad outputs. And what is interesting is that there are many images with a simple description indicating that it is a photo.

Fig. 2: Images with Higher Scores using Captions

2 Part 2: RSA

The fixed number is 26. For specific reasons behind the better performance, I printed out and looked into the probabilities of the first five corrected corresponding text-image pairs. It can be easily observed that the literal listeners returned probabilities that are very sparse and extremely large probability. But the pragmatic listener returned to relatively more concentrated probabilities. And I also plotted a distribution of a specific example. RSA fixes the problem by considering a broader context of given information. It allowed the model to be less biased and understand the dataset more comprehensively.

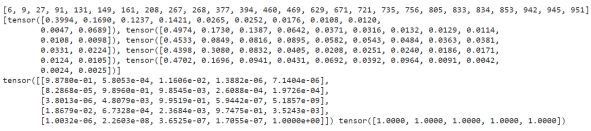

Fig. 3: Probabilities

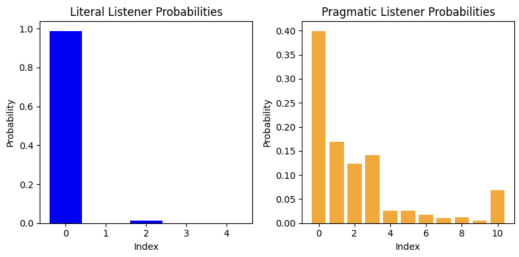

Fig. 4: Probabilities Distribution Comparisons between 2 Methods

3 Part 3: Image Captioning

By observing the image captioning results generated by the two different methods, I found that pragmatic listener returned to more detailed captions. It performed better both in capturing the general frame of the image as well as the detailed features. For a specific example, the literal caption is: a painting of a woman in a red dress, and the pragmatic caption is: an oil painting of a woman wearing red robes and a crown. It has more details thanks to the incorporation of contexts.