Generation of Captions from Images

Tianyuan Cai

This project aims to use neural network to generate sensible captions for images. By describing objects of images with text, an image captioning model can translate complex images into more compact data formats. This type of model can be applied in a wide range of fields such as voice over technologies, product recommendation, etc.

The training and validation data come from Common Objects in Context (COCO) dataset[1] produced by Microsoft. COCO dataset provides a set of images that identifies common objects in context. The data include both images and captions, which describes the objects and their context.

I make use of the Inception V3[2] model and GloVe[3] embeddings to create image and caption vectors as inputs. I then use LSTM model combined with fully connected layers to perform word-by-word prediction to generate sensible next words based on the training images and the portion of the caption that has already been generated. The model is run on the input sentence sequence and image until a full sentence is created, before continuing to create captions for the next image. The final model is able to produce sensible results, as shown in the example below:

A young man is jumping to catch a tennis ball.

A plain white restroom toilet appointment in corner.

Overall, the final model can produce sensible results after tuning the batch size hyperparameter. However, there are also several areas of improvements to address issues in the areas of prediction bias, performance measure, and hyperparameter tuning.

- The model suffers from bias due to the biased distribution of the data set. For instance, the model tends to predict all people as “man” possibly due to the frequent usage of the word “man” in the caption and the frequent presence of male subjects in the images. Data augmentations and more gender-aware captions can help reduce this type of bias in training and prediction.
- Model performance is measured by the categorical loss, BLEU score, as well as human judgment on how sensible the captions are. However, a systematic way to identify semantic similarities between the predicted and actual captions is needed to more accurately measure the overall model performance.
- Given the large amount of data and the limited computation resources, it was difficult to use the grid search function in *sklearn* to perform hyperparameter tuning. I used the *for* loop to tune the batch size parameter, but more computation resources are needed to perform grid search in combination with other parameters. If I had more time, I will write out a custom function for grid search.

Link to YouTube Video: https://youtu.be/_IFIf4Gn4i0

[1] http://cocodataset.org
[2] https://keras.io/applications/#inceptionv3
[3] https://nlp.stanford.edu/projects/glove/