Capstone Proposal

Machine Learning Engineer Nanodegree

Sana Shah

April 23rd, 2020

IMAGE CAPTONING

Domain Background

Generating a description of an image is called image captioning. Image captioning requires to recognize the important objects, their attributes and their relationships in an image. It also needs to generate syntactically and semantically correct sentences. Deep learning-based techniques and machine learning are capable of handling the complexities and challenges of image captioning. Image captioning can be used in a variety of applications, here are a few examples:

- In web development, it is a good practice to label the images, but it can be a tedious task for a human. Automatically generated captions can be very helpful.
- Captions can also be used to describe a video in real time, and provide subtitles.
- Would be greatly helpful for visually impaired people.

There are a number of scenarios in which we can find great use of automatic image captioning. By designing an algorithm which can automatically generate these captions using computer vision, we can take great advantage.

Problem Statement

I will be tackling two of the problems in this project, first will be recognizing an imagine that is considered a computer vision problem. Secondly, after recognizing an image, we would be required to generate some text to caption that image, which is a natural language procession problem. So, in conclusion we need to design a machine learning model that can firstly detect what is happening in the image, and then provide the appropriate captions.

Dataset and Input

The dataset that I will be using to train my model will be a collection of labeled images from Microsoft's COCO (Common Objects in Context) dataset. Each image in the dataset has a total of five related captions. The COCO dataset is an excellent choice as it comes with 80 classes, 80,000 training images and 40,000 validation images. As for input, an image will be provided to the model and it will produce an output prediction consisting of the captions.

Solution Statement

The solution I will be designing to solve the problem of image captioning will have two main components, the CNN (Convolutional Neural Network) and RNN (Recurrent Neural Network). CNN is great for classifying images and RNN can be used for tasks like sentence generation. By combining these two, a model can be obtained with the powerful attributes of both CNN and RNN.

Benchmark Model

By using a combination of CNN and RNN, my model will generate captions, which will be aimed to be accurate and be better than other deep learning-based (Reinforcement learning and GAN-based methods) using them as a benchmark.

Evaluation Metrics

I will be using Cross Entropy Loss as the loss function and will train the model for long enough to keep the value minimized. As I have learned that during one of my nanodegree (Computer Vision) at Udacity, training such a model is a GPU heavy task and there is no better evaluation metrics than to Tinker with your model - and train it for long enough - to obtain results that are comparable to (or surpass!) recent research articles.

Project Design

The data will first be preprocessed. I will write use a data loader that can be used to load the COCO dataset in batches. The data loaded will have all the necessary features that will convert the images into tensors before using them as an input to the CNN, a batch size that will the number of image-caption pairs used to amend the model weights in each training step and some other parameters for the vocabulary and captions which will be fed to the RNN model. The combination of these two, RNN and CNN will provide a caption for any given input image.

References

- https://arxiv.org/pdf/1810.04020.pdf
- http://cocodataset.org/