Project 5 Proposal

Automatic Image Captioning via Deep Learning

Charlie Lew

The year-to-year improvement of mobile phone cameras is slowly eating away the use of DSLR cameras for regular consumers. A perfect example are the cameras on the iPhone 11 Max and the Google Pixel. Both are top in the industry offering near DSLR type image quality minus the bulk and weight. It's easy for consumers to simply take their mobile phones from their pockets and snap a picture to capture an event. The ease itself poses somewhat of a quandary for consumers whereby the non-conservative use of cameras generates a ton of images which are left un-captioned. It would be nice if the image capturing device, i.e. cell phone is able to at least generate a caption albeit a summary instead of the user typing it in themselves. The applications this method is not restricted to just consumer photography but to computer vision applied to self-driving cars, autonomous flying drones, crime forensics, .etc.

A Minimum Viable Product (MVP) is to at least detect one object in a photograph or image, correctly classify them and provide a simple narrative of the object or gesture. This project aims to use Convolutional Neural Networks (CNN) in conjunction with packages such as You Only Look Once (YOLO) to properly identify objects in the photograph or image. Source images will be obtained from Kaggle via the Flickr crowdsource photo database containing 8,000 images along with text for captioning.