Adam Zucker
EE/CS '17
CPSC 490
02/09/17

Web Application for Generating Image Descriptions Using Natural Language Processing and Computer Vision Techniques

Primary Adviser: Dragomir Radev - Computer Science Department

Secondary Adviser: Lawrence Staib - Electrical Engineering Department

Background:

Throughout history, as society has become more technologically advanced, people have been constantly seeking to automate more and more tasks to improve general quality of life. The advent of modern computers, in particular, has opened endless possibilities, as such complex analytical machines can perform far more nuanced tasks than their "naive" purely electro-mechanical machinery counterparts can. It thus comes as no surprise that fields of artificial intelligence and machine learning have seen unbelievable investment and advancement over the past few decades, some standouts being IBM's Deep Blue and Watson, Apple's Siri, and Google's AlphaGo.

One interesting, more recent area of research in the field of AI is computer vision, in which a computer can extract information and understanding from visual data, namely images. One approach to get a computer to make sense of what is going on in an image is to first identify the individual objects in the image, then extract the semantic data from the individual words or phrases describing said objects using natural language processing techniques, and finally to

construct a description in the form of sentence using these words in context. This is precisely the sort of computer science problem I would like to investigate.

Proposal:

For my final project, I would like to explore the usage of deep-learning natural language processing and computer vision techniques to generate descriptions for images. Needless to say, this is a complex problem, but a fascinating one with plenty of ongoing research. But rather than writing an academic paper examining these techniques, I would like to instead build a web application that builds upon existing techniques to generate descriptions for images a user uploads.

The main reason I prefer to program a web app over writing a paper is that while there exists plenty of research into the idea of automatically generating image descriptions, there does not seem to be much in the way of publicly available applications that actually attempt to do this. Of course, this is probably mostly due to the immense technical challenges involved in creating accurate descriptions; nonetheless, regardless of the so-called "quality" of these descriptions, I still think it would be interesting to create such an application, at the very least for experiments' sake. One major possible use case for the technology behind such an app would be as an aid for the visually impaired: the computer receives an image (potentially from a live camera feed), turns it into a description, and reads it aloud to the user.

Another novel component I hope to incorporate into my project is the idea of using image metadata as an additional contextual input for description generation. For example, many image files today have geo-tags encoded into them (e.g. EXIF data in JPGs) that could additionally be

used to try to create interesting descriptions (e.g. incorporate location name and associated data into NLP component by using place-of-interest returned from passing GPS coordinates to a places-of-interest API). There are probably numerous other sources of context I could also end up using, but this is the first that came to mind.

Approach:

The implementation of this project can be broken down into two parts: the computational CV/NLP engine that handles the image description generation and the actual web application that hosts and interfaces with this service.

For the CV/NLP part, I plan to use an existing data set of images and descriptions, such as MSCoco or ImageNet, to query against. Both of these large datasets have Python APIs that will allow me to access the data directly from the backend of my web application. Additionally, a New York start-up called Clarifai has a Python API that given an image, will return all the tagged objects in the image. Clarifai's API would be particularly useful as it has a built-in training feature that would allow my web app to tag custom objects not in their image recognition service (e.g. it could be trained to tag a photo of me with the object "Adam"). The objects returned by these dataset APIs can then be processed together to form sentence descriptions using the NLTK Python library and returned back to the user.

For the actual web application, I will use the Flask Python web framework for the backend and React.js for the front-end. I plan to use Flask as, while it provides a pretty lightweight framework, it is fully customizable, which will allow me to pick and choose my own

plug-ins. Flask also integrates pretty well with React, which is one of the fastest front-end JS libraries that also mostly eliminates the need for writing plain HTML code.

List of Deliverables:

- Website interactions mocks/sketches
- Github repository containing code
- Report describing project implementation in depth (will probably mainly focus on CV/NLP component)