Milestone 1: Caption this Pic

Team KKST

Team Members

1) Al-Muataz Khalil (almuatazkhalil@gmail.com)
2) Stephen Knapp (stephen.t.knapp14@gmail.com)
3) Matthew Stewart (HUID: 31309891, matthew_stewart@g.harvard.edu)
4) Shih-Yi Tseng (shihyi_tseng@g.harvard.edu)

Problem Definition

The World Health Organization (WHO) estimated that 314 million people have visual impairment across the world, including 269 million who have low vision, and 45 million who are blind (Ono et al 2010). Many people with visual impairments rely on screen readers in order to access the internet through audio, and thus depend on image captions (Yesilada et al 2004). Therefore, accessibility, as well as automatic indexing and other goals, make accurate image captioning an important priority (Hossain et al 2018).

Proposed Solution

We will design, build, and deploy at-scale an application which receives an image of an everyday activity which then assigns a caption of the image contents, based on state-of-art computer vision and natural language models. Additionally, the app will provide a visualization of the image components reflected in the caption through saliency maps.

Project Scope

The project has the following requirements and limitations:

- The model will be trained and tested only on the images in the data sets listed in the Data Sets section of this report.
- Images of niche activities or objects (e.g., medical imaging, outer space) will not be considered in the model.
- The user may upload images for captioning that are not in the data set.
- Some image pre-processing will be required but manual labelling of the data is not required.
- The model will involve both image processing and NLP. It will utilize a deep learning neural network built in TensorFlow using Google Colab.

- Model accuracy will be measured by the number of objects and activities the model successfully identifies in each photo and the overall coherency of the caption.
- The user interface will be accessible on the internet through Google Cloud Platform utilizing a Docker container and Kubernetes.

Timeline and Components

- **Mid-to-late September:** background reading on SOTA models for image captioning; download and explore the data structure.
- **Early-to-mid October:** implement and train the models on a small subset of image data as proof-of-concept experiments. Apply transfer learning techniques from pre-trained transformer-based language models and CNN.
- **Mid-to-late October:** Build data pipelines with the whole dataset on GCP and train models and run experiments.
- **Early-to-mid November:** Design the web app, identify the containers, frameworks and structure of the codebase.
- **Mid-to-late November:** Implement and scale the app with Kubernetes and deploy it to GCP
- **Early December:** Wrap up and work on final delivery.

Datasets

- Flickr8k — 8k images paired with five different captions to describe entities and events
- Common Objects in Context (COCO) — 330k images paired with five captions per image

Models

- Image to sequence problem: process an image, represent it's context then transform that into a sequence of words maintaining fluency of language.
- CNN Encoder to LSTM decoder architecture.
    - Visual CNN encoder to to represent the image as feature vectors
        - Can Potentially use a prettrained computer vision model
    - Language LSTM Decoder to generate a sequence of words from the image feature vectors.
- CNN to Transformer
    - Create a feature vector from a pretrained CNN and input its pretrained language model as the language decoder.
- Research other SOTA models