Milestone

Introduction

Problem:

Describe the content in the picture precisely using sentences.

Application:

Transform pictures into sentences or braillie for the purpose of assisting blinds to comprehend the content in the pictures.

Problem statement

State the full problem :

As technology advances, smartphones and social medias become an indepensable part in human beings' lives. Nevertheless, for social medias such as Facebook and Instagram using images to deliver messages are difficult for blinds to perceive, let along for them to be involved in using these social medias. Therefore, we hope to apply techniques including image processing and natural language processing to describe the context in the images and then transform them in braillie or voice messages. Eventually, we hope to assist blinds to participate more in using social medias and thus be more involved in society.

Dataset:

First stage dataset: **Flickr8k**

Sentence-based image description and search, consisting of 8,000 images, each paired with five different captions which provide clear descriptions. The images were chosen from six different Flickr groups, and tend not to contain any well-known people or locations, but were manually selected to depict a variety of scenes and situations.

Expected results:

We hope that we can use our model to describe images precisely, which means we would try our best to generate a clear statement about an image. However, in reality, image caption can be a great challenge. To successfully capture the keywords of images would be the first goal. The second goal is to combine the keywords to an complete sentence so that we can help visually impaired people.

Evaluatuon: cross entropy loss

In Rachael's book, she had validation loss: 4.39 and 2.99(with transfer learning)

1. Our first evaluation is to generate validation loss lower than Rachael's result.

2. Besides, we will check the statements of our test data. We wish that we can reduce the irrelevant words and descriptions about images.

3. We will check the completeness of our captions.

Approach:

Model Usage:

Image Pre-process Model: **ResNet50**

ResNet-50 is a convolutional neural network that is trained on more than a million images from the ImageNet database. The network is 50 layers deep and can classify images into 1000 object categories. As a result, the network has learned rich feature representations for a wide range of images. The network has an image input size of 224-by-224. It is a variant of ResNet model which has 48 Convolution layers along with 1 MaxPool and 1 Average Pool layer.

RNN Model: **LSTM**

LSTM is a type of RNN method that can be used to learn ordered sequences(Time series data). The details of its algorithm are to input data in chronological order, convert it, and then mix it with past data. The final result will be the last hidden feature.

Use RestNet50 and LSTM to extract features from images and use NLP to extract features from captions.

Preliminary results:

(1) There are 8091 images in Flickr8k, each images have 5 captions(Total 40455 captions).

(2) All captions are stored in a txt file.

(3) The format of the file**:** image_name, caption (each image name repeat 5 times)

(4) The longest caption has 37words, therefore we use 40 sequence_length in LSTM model.