MAIS 202: Progress and Preliminary Results

Nathalie Redick

14 February 2020

I. Problem Statement:

For my final project, I am creating my own dataset of images by web scraping the Smithisonian website for images of rocks, minerals, gemstones, and fossils. I will use train a convolutional neural network on the data and use the model to make predictions. Then, I will create a webapp to display my model.

II. Data Preprocessing

For my database, I selected the Smithsonian Natural Museum of Natural History's Geology Collections Data Portal[1] because it has approximately 47,000 images that I can utilize that are classified in an organized fashion. Creating my own dataset proved extremely difficult and took much more time than I anticipated. I was able to preprocess the data and ended up with a total set of approximately 11,000 images. I resized all of the images, removed duplicates, and removed Gaussian noise. Resizing images is necessary for training and testing the CNN, removing duplicates prevents the model from learning images rather than patterns, and Gaussian noise improves the model's ability to decipher the important aspects of each image.

III. Machine Learning Model

I used a convolutional neural network (CNN) on my image dataset. The data was split by training on 9423 samples and validating on 1663 samples. The split I used 8 different layers, including a convolutional 2D layer, dense, flatten, dropout, and max pooling. Originally, I also implemented a function to remove noise via Gaussian blur. I used standard hyperparameters because I wanted to get the model up and running.

IV. Preliminary Results

The preliminary results report (test data):

- Test loss: 4.099792193162176
- Test accuracy: 0.7456404091146289

For the first few rounds of training, the model performed satisfactorily. Tweaks to hyperparameters and possibly more image preprocessing need to be made.

V. Coming Soon

The goal for the next deliverable is to further fine tune the model to increase the model metrics. Pros of my approach include extensive research and extreme dedication, while ons include my inability to make office hours due to scheduling conflicts. I plan to play around with the convolutional layers as well as other hyperparameters and image preprocessing to increase the model accuracy.