MAIS 202: Data Selection Proposal

Nathalie Redick

27 January 2020

I. *Database and 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. The data is downloadable in sets of 1,000, but I plan on creating a custom dataset by scraping the website for images and rock classifications.

II. *Machine Learning Model*

Ideally, I'd like to train a model to return rock classifications based on the images given. I'd like to implement a Convolution Neural Network (CNN) to classify the image data. CNN's are popular for image classification and recognition because they are known for producing highly accurate results. I also considered using a Deep Neural Network (DNN), but after some research I ultimately decided that it would be best to use a CNN because DNN models typically have larger losses and longer training times when working with images.[2]

III. *Final Conceptualization*

To conceptualize the model I designed, I'd like to create a simple landing-page web app that allows you to upload an image and have the model assign the most likely classification of the rock type based on the input image. Given enough time, I am excited by the prospect of designing a mobile app to utilize the model.

1. *NMNH Geology Collections Data Portal*, geogallery.si.edu/portal.

2. Pal, Lalit. "Image Classification: A Comparison of DNN, CNN and Transfer Learning Approach." *Medium*, Analytics Vidhya, 13 Sept. 2019, medium.com/analytics-vidhya/image-classification-a-comparison-of-dnn-cnn-and-transfer-learning-approach-704535beca25.