(CS 4262): Gang of Four - Josh Scherer, Jackson Shelby, Michelle Wang, & Alex Zhang
March 24, 2023

Project Interim Report

1) Data exploration:

We are using a dataset that provides images of MRI brain scans of patients with Alzheimer's at various stages of the disease. The dataset contains both original data and augmented data.
It is important for us to understand both types of data, as this is what we will be training on.
We began by investigating whether the images within the sets are of a consistent size. To do so, we iterated over all images and used Image.open().size through the PIL.Image library.

We came to the following conclusions:
- Within the augmented dataset, we found the following:
  - Mildly Demented: 8,064 190x200 images and 896 180x180 images.
  - Moderately Demented: 6,400 190x200 images and 64 180x180 images.
  - Healthy: 6,400 190x200 images and 3,200 180x180 images.
  - Very Mildly Demented: 6,720 190x200 images and 2,240 180x180 images.
- Within the original dataset, we found the following:
  - Mildly Demented: 896 208x176 images.
  - Moderately Demented: 64 208x176 images.
  - Healthy: 3,200 208x176 images.
 - Very Mildly Demented: 2,240 208x176 images.

The original dataset has a standard image size (but is a much smaller set), while the augmented dataset does not. However, the images that have a different size within the augmented dataset are almost certainly derived from those within the original dataset, as there are exactly the same number of differently sized images in the augmented set as the total within the original dataset. Further the original dataset has colored images, while the augmented images are normalized to black and white. Below is an example of a 190x200 image from the Mildly Demented dataset converted to a size of 185x190 (the significance of this number is described in part [4]). The final image is the same image converted to a 57x60 image to observe a noticeable change in quality.
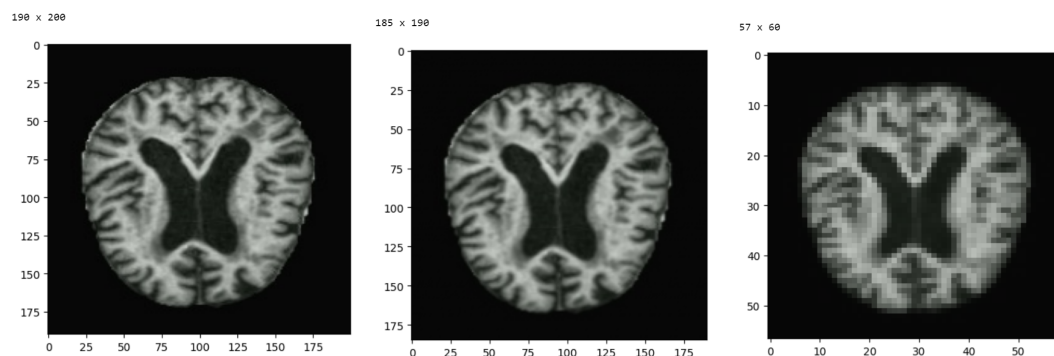
2) Preliminary results:

We have not learned about Neural Networks in class yet, so we have been unable to begin working directly on the model. However, we have done some out-of-class research to learn typical processes for ML processing with images.
We followed a video (https://www.youtube.com/watch?v=iGWbqhdjf2s), which gave us some key insights into how images can be stored in numpy, how to load images from a directory, how to check the size of an image, how to resize an image using existing libraries, and how a convolutional neural network interacts with training images, testing images, and new images that are fetched from external sources. Below are some key plots from the guided practice:
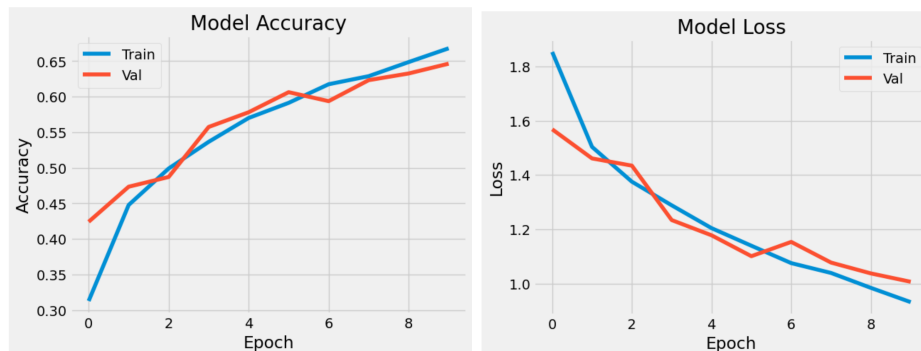
3) Teamwork: if you are working as part of a team, how have you been collaborating and dividing the work?

We created a GitHub repository so that all members can collaborate on the same files simultaneously. We made sure to use branching and versioning to ensure that the collaboration does not result in anyone's work being overwritten by accident.
We plan to continue to use this repository moving forward, as it has proven beneficial throughout the early stages of this project.

4) Next steps: based on your progress and any challenges you have encountered, what are the next steps you are considering?

Since we found our data to be of varying size, our current plan is to use the average size of the images in the augmented dataset and convert all images to be that size for training. We also anticipate that the image quality may be too granular, so we may need to use a common factor (or an approximate common factor) for the image dimensions to reduce training time. Of course, this is dependent on us beginning to implement the neural network, which we plan to do immediately after gaining some exposure in class.

NOTE: We are happy to provide any notebooks that we used for exploring our data. They are located in our private repository, but we can easily provide access.

References

[Computer Science]. (2020, June 5). *Classify Images Using Python & Machine Learning*

[Video]. Youtube. https://www.youtube.com/watch?v=iGWbqhdjf2s

Hunter, J., Dale, D., Firing, E., Droettboom, M., & Matplotlib Development Team. (2012).

matplotlib.pyplot.imread. matplotlib.

https://matplotlib.org/stable/api/_as_gen/matplotlib.pyplot.imread.html

uraninjo. (2022, September). Augmented Alzheimer MRI Dataset, Version 1. Retrieved February

3, 2023 from

https://www.kaggle.com/datasets/uraninjo/augmented-alzheimer-mri-dataset