Explanation

The goal is to be able to start with a small dataset and blow it up into a larger one for better training results. I want to come at it from a medical point of view as I told the professor that I wanted to train on one medical issue and then use GANs to generate data to train a network to identify a more rare disease.

I’ve used balanced data for everything, equal number of images for both classes in the training, validation and test data.

I started with the MedMNIST data set. I took the PneumoniaMNIST that had enough images to train a NN to classify Pneumonia Chest X-Rays from Healthy chest X-rays at about 96% accuracy. When I tested it though, it had about a 90% accuracy rate on test data. (Dataset contained around 3,000 + images)

I then took another dataset, the ChestMNIST data set, which has 14 different classes in it. I wrote an algorithm to separate all of the different classes. I then chose to use the cardiomegaly class (enlarged heart – that can indicate many other diseases). That dataset has I think 1950 images in it. I trained a NN using that data vs healthy chest x-rays and was able to get 75% accuracy on test data. I used a GAN algorithm for data augmentation (just finally got this to work today). I’m now going to add the generated images to the dataset and retrain to try and improve the 75% accuracy. I then want to take less than the 1950 and re-run them through the GAN to see if I can beat the 75% with even less training data than the 1950. I really want to keep looping that test with less and less initial training data to see what the lowest amount of starting data I can achieve the same (or similar) results with.

(If this isn’t enough stuff to write a whole paper on, I can expand to include the other classes in the ChestMNIST dataset – Emphysema and some other stuff is in there too. I thought just one class would be good enough to do this on, but I’m not sure)