AML Assignment 1

Paarth Iyer - MCS202218

Method and Layout

The notebook is in 2 parts, some helper functions (training, testing, plotting etc.) and results (which consists of training the network with some results, performance of network on scrambled images (a permutation has been fixed beforehand) and FGSM attack on the network, all of which are done each time for a fully connected network and convolution network, with the datasets FashionMNIST and SignLanguageMNIST.)
The functions are described in the notebook.

Results

FashionMNIST - FCNN

We get an accuracy of about 87% on the test set. The performance of the model is similar (around 87%) when it is trained on scrambled images, this is to be expected.
FGSM attack - The attack works pretty well on this model, and the ideal value of ϵ seems to be around 0.08, as that preserves readability and the model accuracy drops to around 20%

FashionMNIST - CNN

The accuracy for the CNN is an improvement over the FCNN with this network giving an accuracy of around 91.3%. When trained on scrambled images, the accuracy takes a hit of around 4%, making it 87-88%. This tells us that the CNN does use some sort of surrounding spatial information.
FGSM attack - Similar to above, a value of 0.04 to 0.08 seems optimal given the recognizability-accuracy trade off. The attack does not work as well as on FCNN for higher values of epsilon. The accuracy seems to be around 0.2 for ϵ 0.1. This may indicate that the CNN is a bit more robust than the FCNN when images are very noisy.

An observation for the above two FGSM attacks, for lower values of ϵ, the label make by the noisy image is usually "closer" to the actual label in possible meaning.

SignLanguageMNIST - FCNN

This model gives an accuracy of about 74%, and a similar accuracy (about 73%) on scrambled images as expected.
FGSM attack - The attack works gives low accuracy pretty fast (¡10% for ϵ=0.04) but the image starts to look considerably tampered for values above 0.02.

SignLanguageMNIST - CNN

This is a major improvement over the FCNN, giving an accuracy of 95.9%. The model, when trained on scrambled images, gives an accuracy of only about 78.8% suggesting that the CNN makes heavy use of local spatial patterns.
FGSM attack - The attack does work, but the accuracy drop-off is not as bad as that of the FCNN. The models managed to maintain an accuracy of more than 20% for ϵ values less than 0.04. Similar to the FCNN, the image looks very different for values more than 0.02.