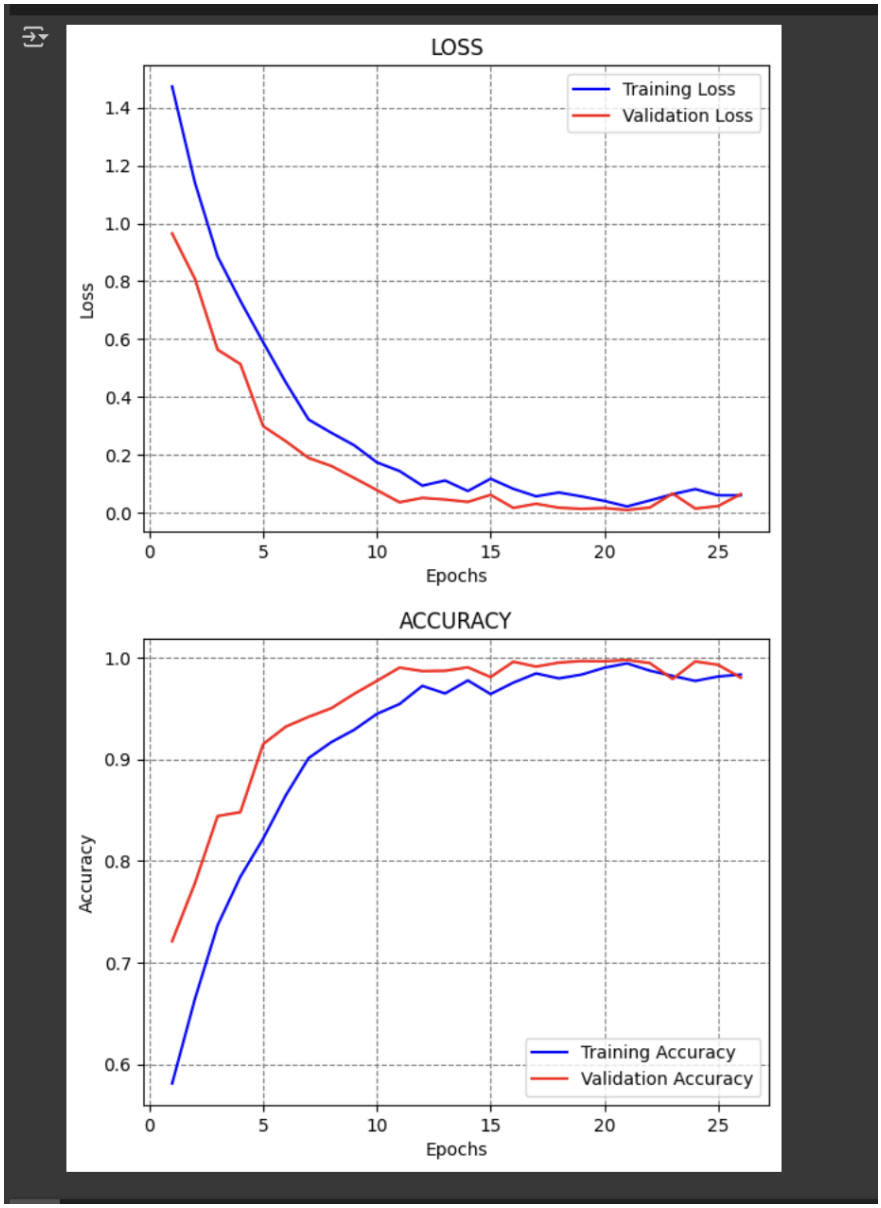
Lalith Aditya Chunduri

Cs-487
Assignment 1
Task-1

Model	Train Set	Validation Set	Test set
Vgg16	98.02%	98.02%	81.12%

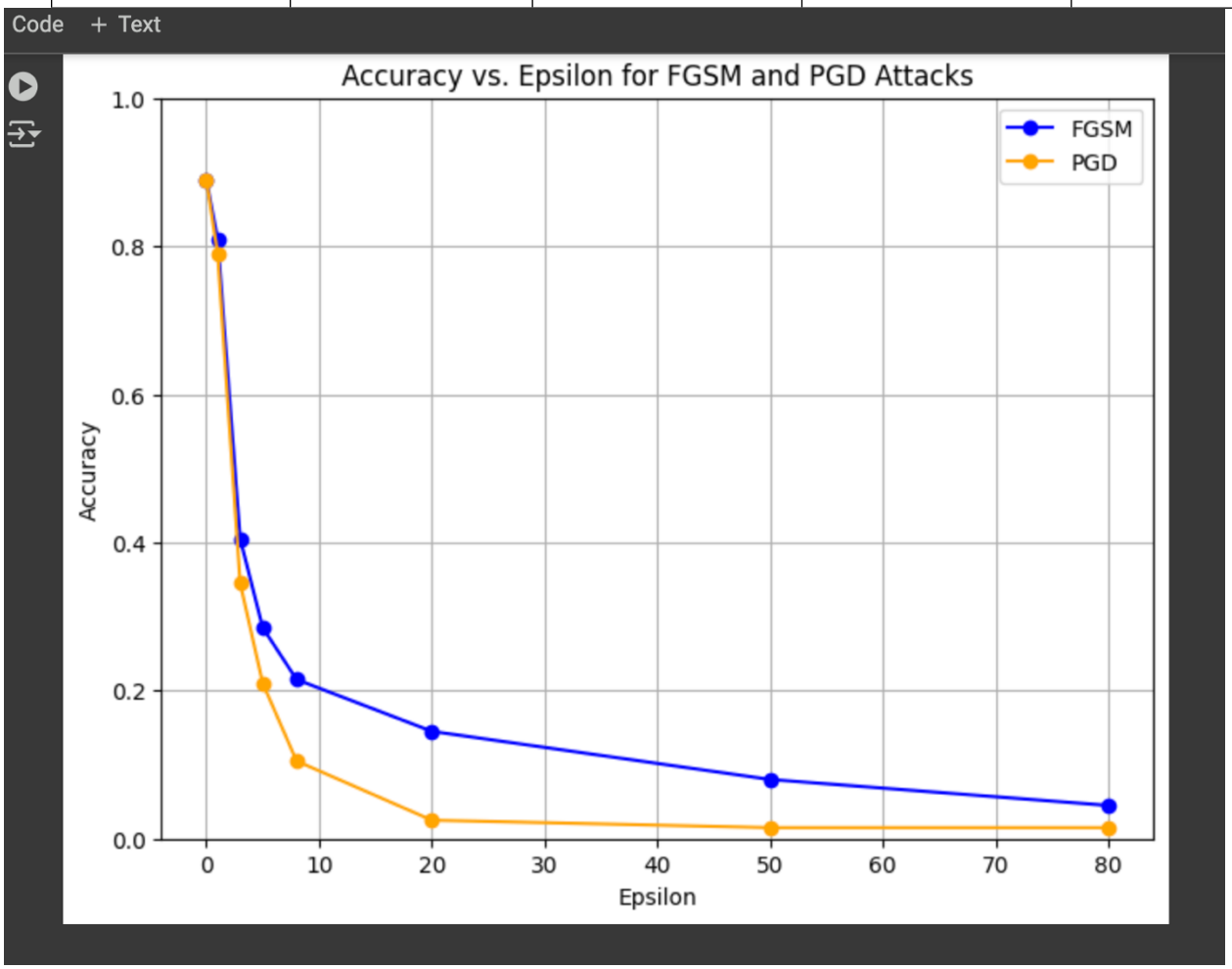
b.



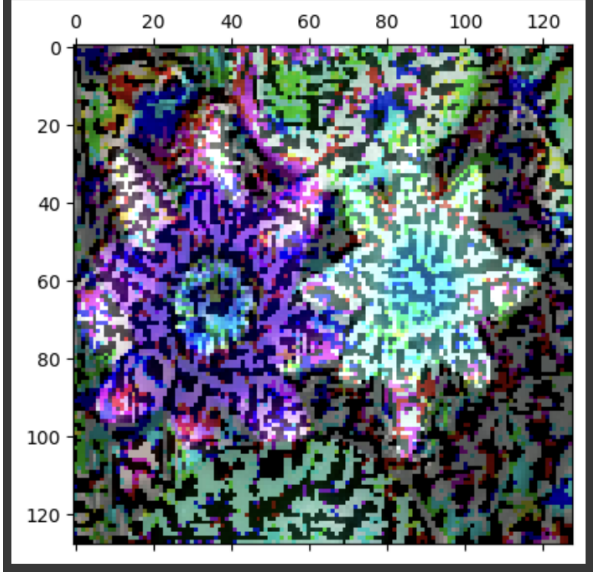
Lalith Aditya Chunduri

Task-2

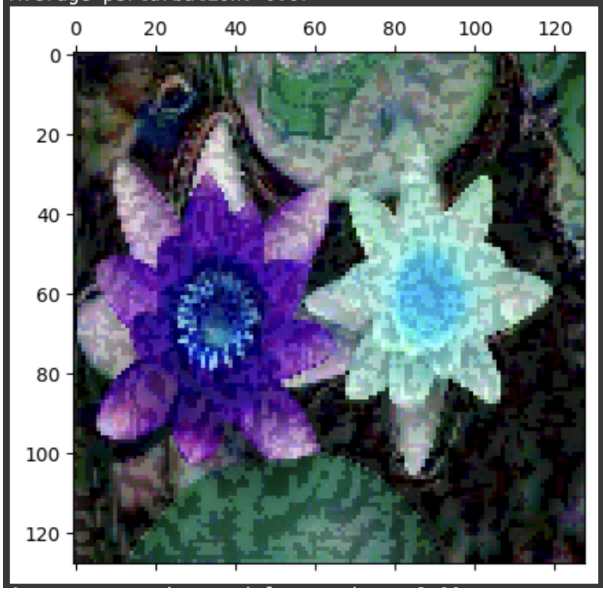
Model	Clean images	Adversarial images ϵ=1/255	Adversarial images ϵ=5/255	Adversarial images ϵ=8/255
FGSM	**89%**	**81%**	**14.50%**	**4.50%**
PGD	**89%**	**79%**	**10.50%**	**1.50%**



```
Accuracy on adversarial test data: 4.50%
Average perturbation: 0.26
```




```
Accuracy on adversarial test data: 14.50%
Average perturbation: 0.07
```



d. The analysis indicates that both FGSM and PGD attacks lower the model's accuracy significantly as epsilon increases, as it is evident with a higher efficiency of attack for PGD around a larger epsilon range. At low epsilon regions, the model's accuracy is still greater than 90%, which means that the model is still robust enough as the perturbations are very minor (for example, $\epsilon = 0/255, 1/255$). But as moderate epsilon values are increased like $\epsilon = 5/255, 8/255$, there is a noticeable drop in accuracy particularly for PGD as stronger iterative perturbations are applied. Around epsilon values of $\epsilon = 20/255, 50/255$, the stability of the model's accuracy for both attacks greatly increases revealing a drop of almost 60%. This shows that both approaches lead to high vulnerability around adversarial noise shifts. This also indicates that the FGSM approach with lesser attacks is less potent when compared to the iterative PGD technique, which leads to more degradation of model performance.