Alternate Approach:

In this approach, we divided our task into phases: i) create a clean model from clean images ii) fine-tune the clean model on perturbed images. The provided 100k dataset was divided into two parts: 95k for training, 5k for validation.

i) Create a clean model from clean images:
We first train our model on 95K clean images without perturbation and save the clean model based on accuracy on the validation set, i.e, 5k data. We train for 100 epochs. This is implemented in '**other_approach/model_for_clean_image.py'** script.

ii) fine-tuning the clean model on perturbed images:
Then all 100k images are perturbed using the *torchattack* library's *fgsm* and *pgd* classes. Therefore, from 100k clean images, we now have 100k FGSM adversarial images, 100k PGD adversarial images, along with 100k clean images total of 300k images. Now, using a batch size of 128, we perform batch mixing. We prepare every batch with clean images, fgsm, and pgd perturbed images in the following way:

- 50% clean images
- 25% FGSM perturbed images
- 25% PGD perturbed images

Now the clean model is again loaded and fine-tuned using this batch-mixing strategy. In order to avoid bias, within each batch, we perform shuffling so that the model does not see the same type of image for a long time. In this way, we train for 45 epochs and save the model on accuracy for the validation set, which is 15k, i.e, 5k clean, 5k FGSM perturbed, 5k PGD perturbed. This is implemented in '**other_approach/model_for_perturbed_image.py**' script.
Models Used: For the clean model, we used ResNet50, pre-trained on ImageNet. We replace the last FC layer with a 10-class fully connected layer.

Results: We have fine-tuned the clena model in different combination of FGSM **ε** (epsilon), PGD **ε**, ***α***(alpha), and number of iteration, i.e, fgsm **ε** = 0.01 to 0.05, PGD **ε** = 0.03 to 0.1. Only significant results that we obtained from server are provided in the table below.

	Clean Accuracy	FGSM	PGD
FGSM **ε** = 0.03	**0.6276**	**0.3926**	**0.0006**
PGD **ε** = 0.03, ***α*** = 0.00214, # of iteration = 14			
FGSM **ε** = 0.05	**0.6326**	**0.4046**	**0.001**
PGD **ε** = 0.01, ***α*** = 0.007, # of iteration = 14			

Observations:

We have above 60% clean accuracy, 39%-40% FGSM accuracy for both settings mentioned in the table. However, the PGD accuracy is quite low. One reason can be the PGD attack with the value we chose for $\boldsymbol{\varepsilon}$, $\boldsymbol{\alpha}$ might be quite far from the PGD attack that the submision system is using. On the contrary our chosen parameter for FGSM might be closer to the submission system.