Image Classification using Vision Transformers and Enhancing Adversarial Robustness

•••

Shubham Mishra
Group #04

Vision Transformers (ViTs)

Traditionally, Convolutional Neural Networks (CNNs) have been for computer vision tasks like image classification

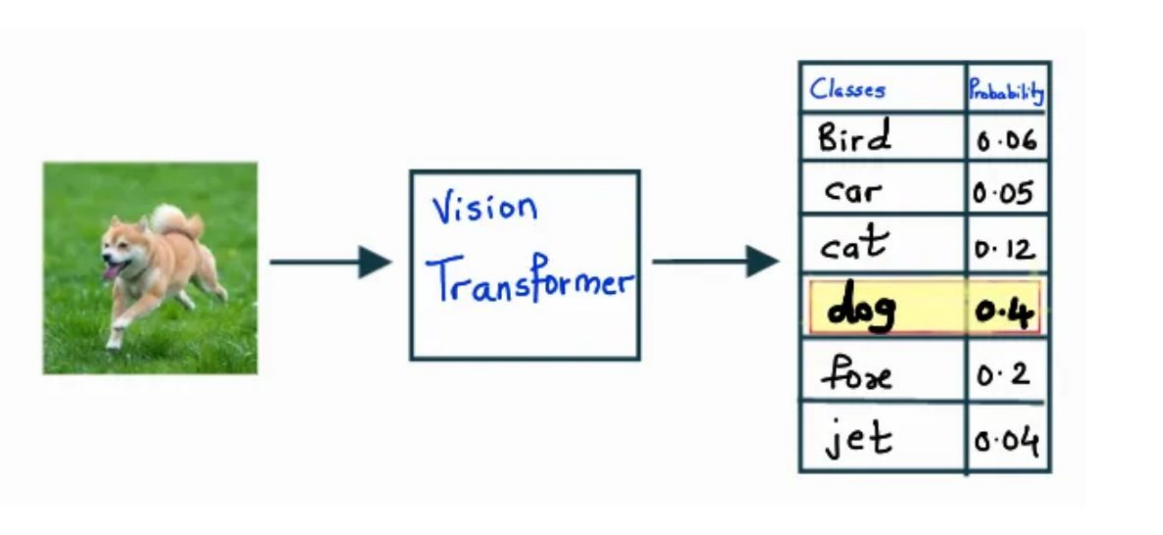
However, CNNs struggle to capture long-range dependencies and global relationships within images effectively

Vision Transformers (ViTs) addresses this limitation

This enables ViTs to capture global context and long-range dependencies more effectively than CNNs

Image Classification

Image classification deals with assigning a class label to the input image. For example, as you can see in the below image, we predict the class as Dog for our input image as it has the highest confidence score after applying softmax.

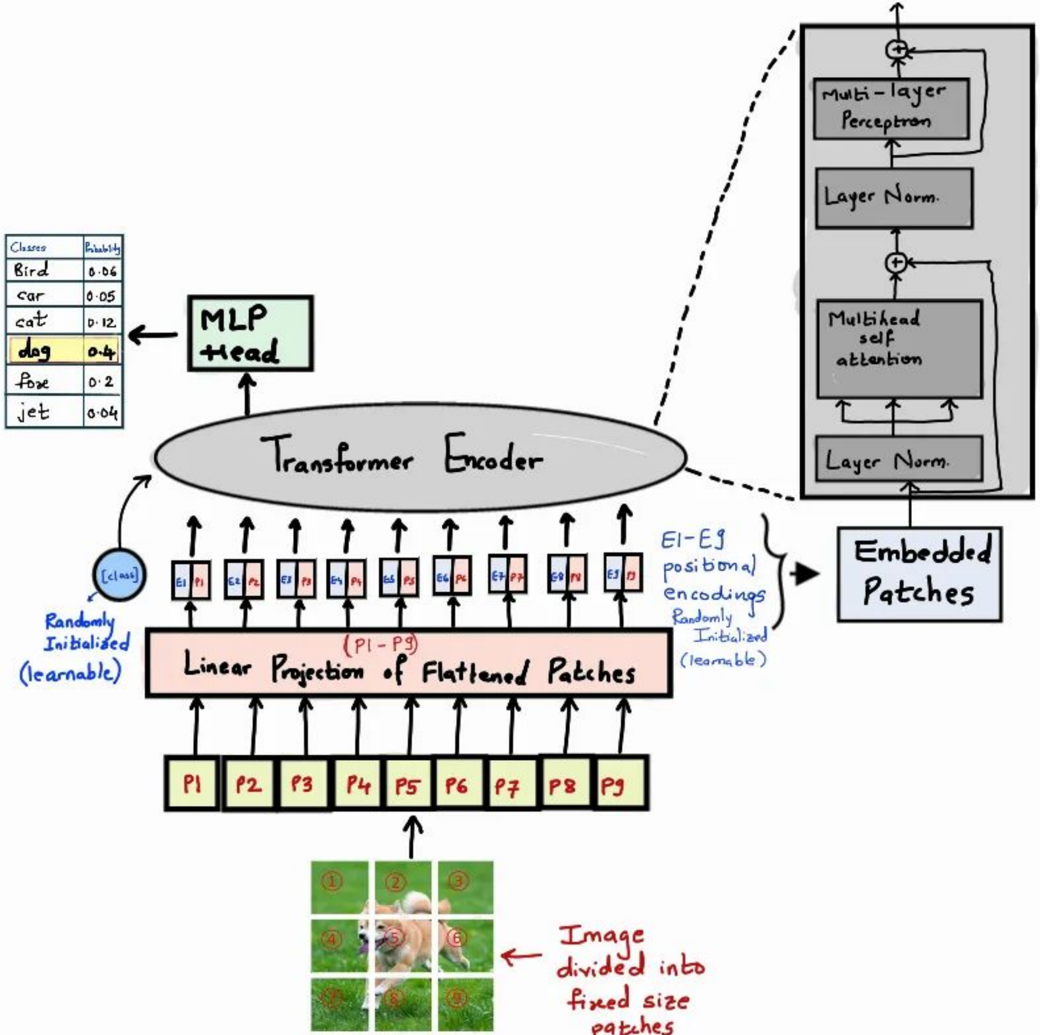


Classes	Probability
Bird	0.06
car	0.05
cat	0.12
dog	0.4
fox	0.2
jet	0.04

The Vision Transformer

Three main Components

- Embedding
- Transformer Encoder
- MLP Head



Classes	Probability
Bird	0.06
car	0.05
cat	0.12
dog	0.4
fox	0.2
jet	0.04

Objectives

Train a ViT model on CIFAR-10 for image classification:

- Utilize the CIFAR-10 dataset to train a Vision Transformer model for the task of image classification.
- This step serves as the foundation for evaluating the model's performance and assessing its robustness.

Evaluate performance on clean and adversarial examples:

- Assess the trained model's performance not only on clean images but also on adversarial examples.
- Adversarial examples are crafted to deceive the model, providing insights into its vulnerability to adversarial attacks.

Improve robustness through adversarial training:

- Incorporate adversarial examples into the training process to improve the model's resilience against adversarial attacks.
- By exposing the model to adversarial perturbations during training, it learns to generalize better and become more robust

Initial Image Classification

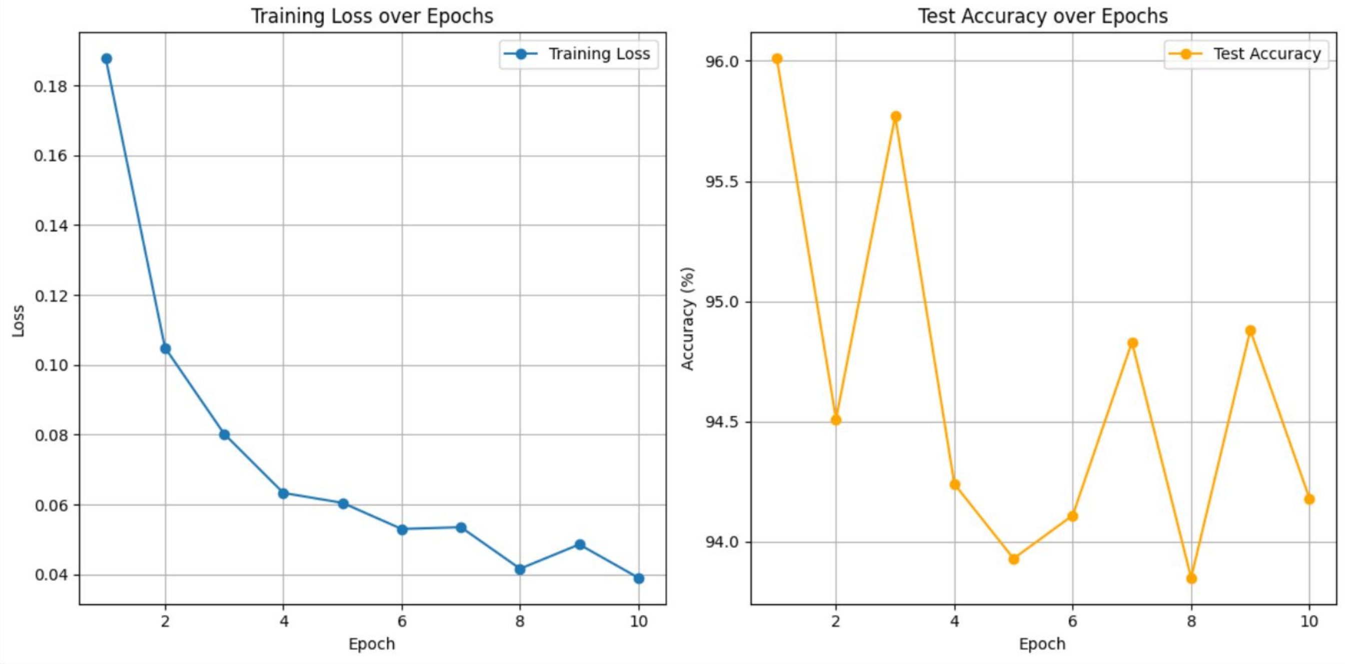
ViT Model Architecture:

- Used vit_base_patch16_224 architecture for the initial image classification task.
- This variant of the Vision Transformer model has demonstrated promising performance on various benchmarks, including image classification tasks.

Training Setup:

- The training setup included the following hyperparameters:
  - Batch size: 64
  - Learning rate: 1e-4
  - Number of epochs: 10

Initial Classification Results

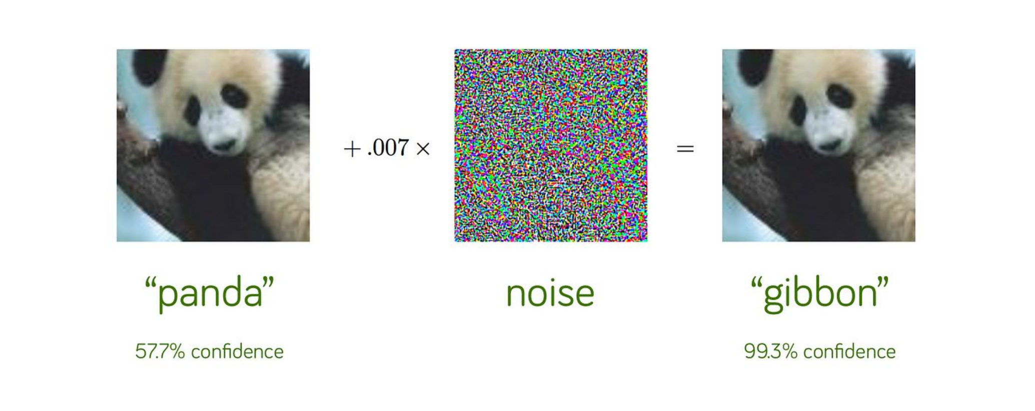


Adversarial Attacks

Adversarial attacks are a significant challenge in deep learning, especially for computer vision tasks.

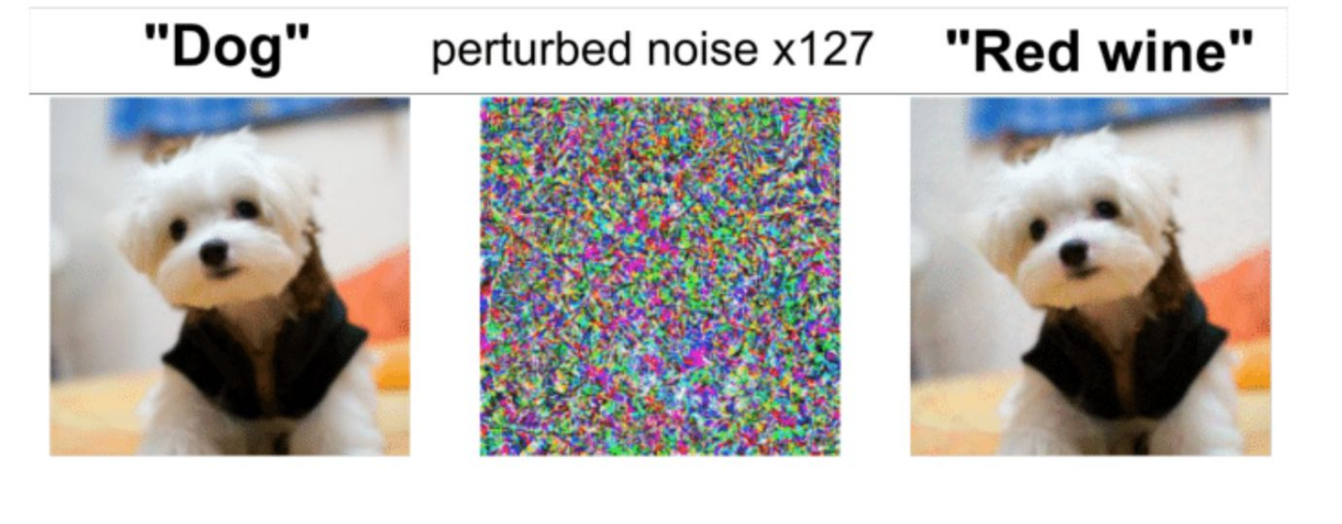
These attacks involve introducing small, carefully crafted perturbations to input images, leading to misclassification by deep learning models.

Despite being imperceptible to human eyes, these perturbations can cause significant changes in the model's predictions.



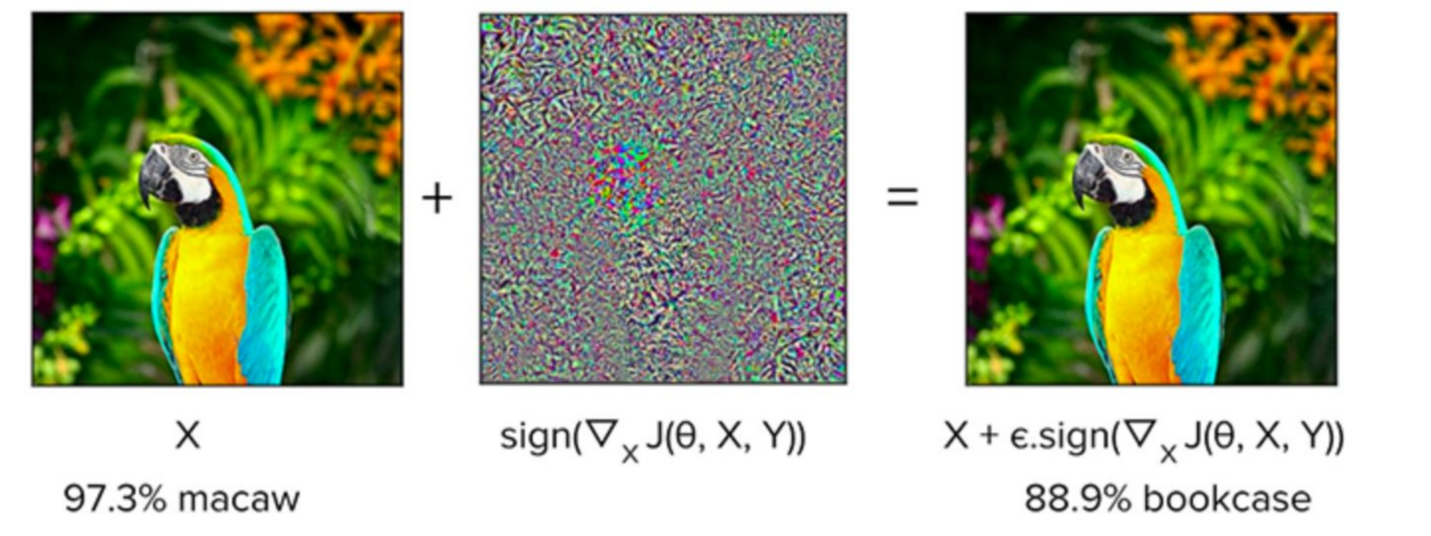
Adversarial Attack Methods

PGD (Projected Gradient Descent): Iteratively perturbs input images to maximize the model's loss, ensuring the perturbations are within a specified epsilon range.

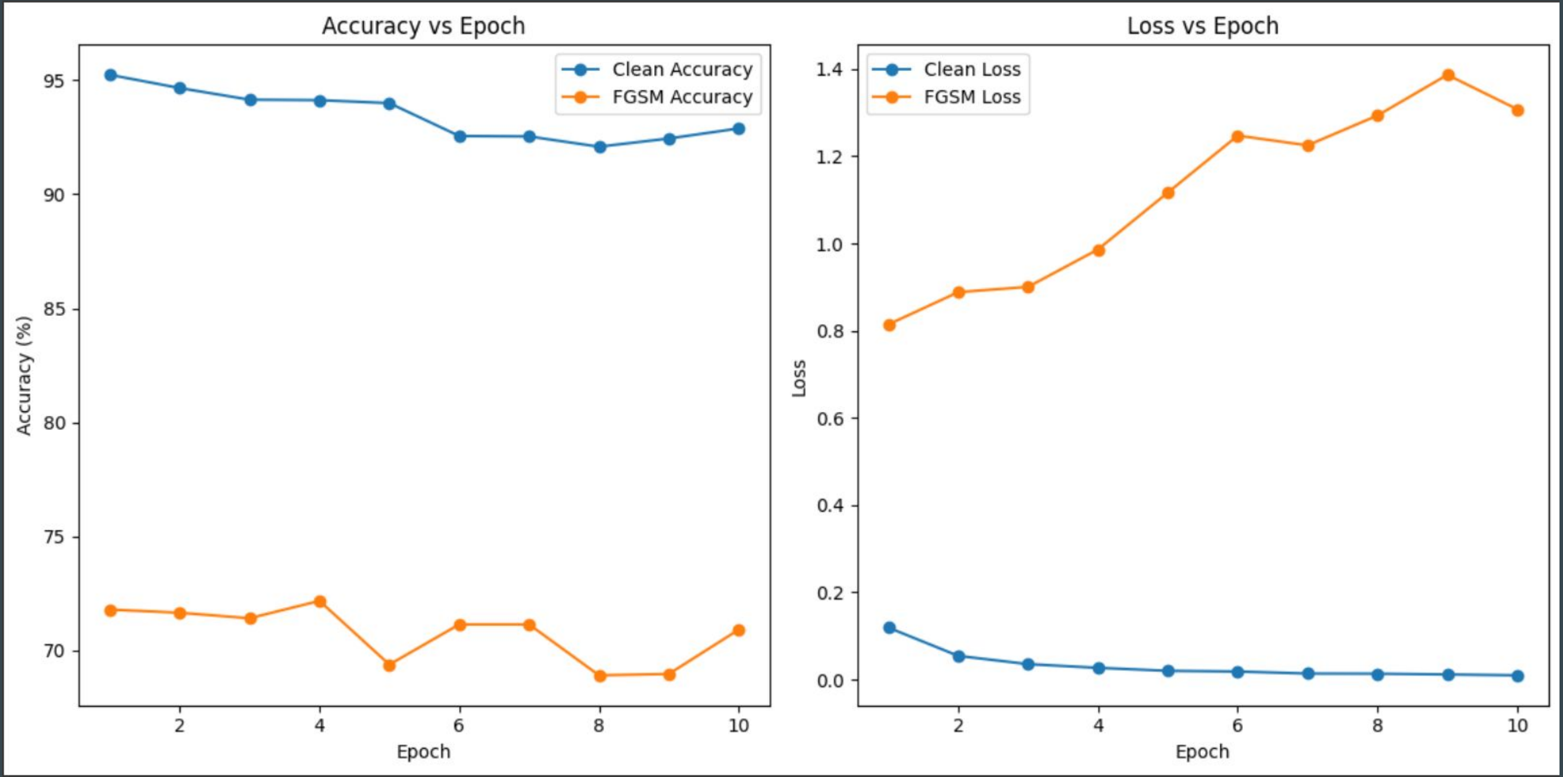


Adversarial Attack Methods

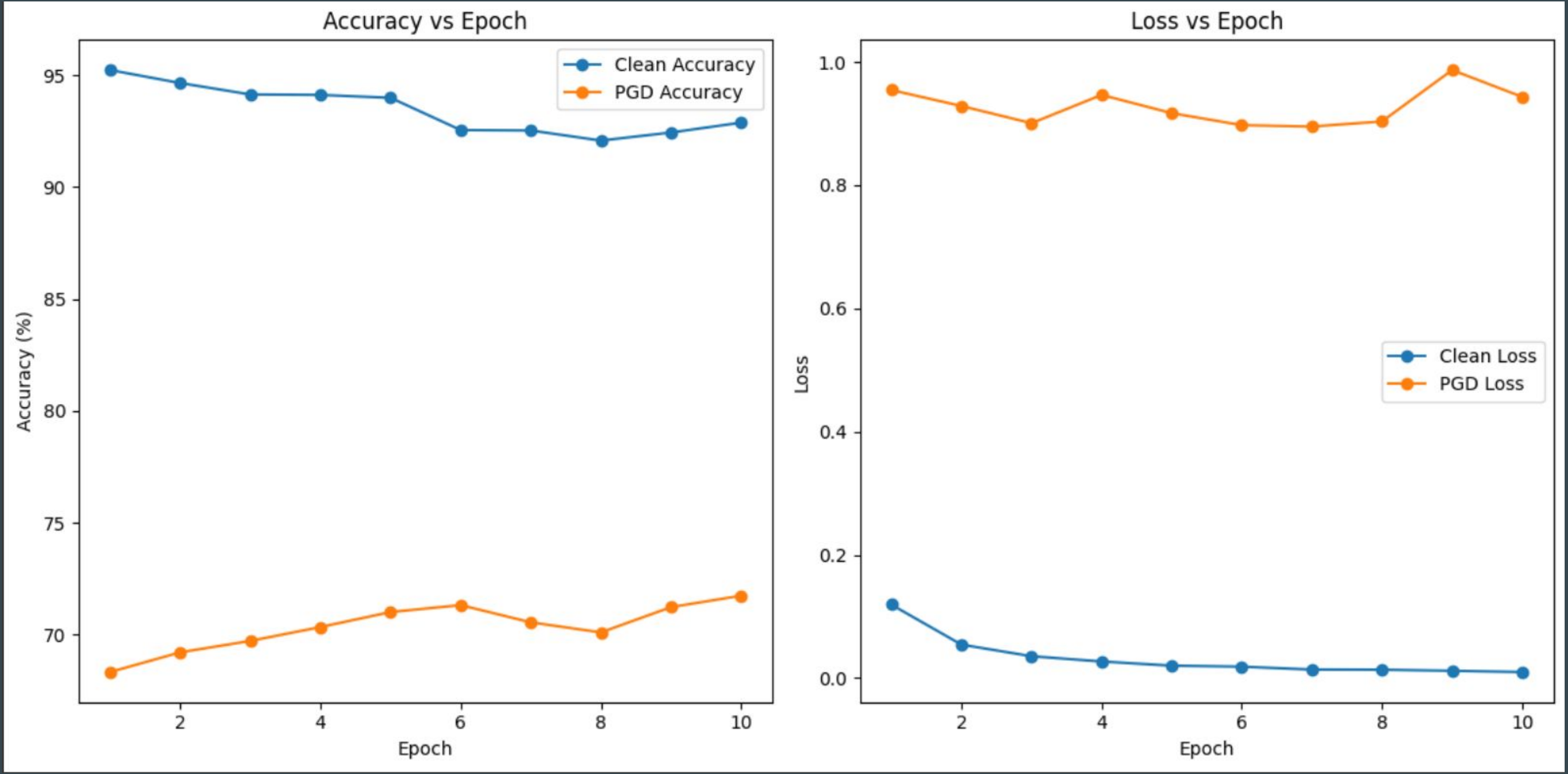
FGSM (Fast Gradient Sign Method): Generates adversarial examples by perturbing input images in the direction of the gradient of the loss function.



Clean vs FGSM



Clean vs PGD


Accuracy vs Epoch
Clean Accuracy
PGD Accuracy
Accuracy (%)
95
90
85
80
75
70
2
4
6
8
10
Epoch
Loss vs Epoch
Clean Loss
PGD Loss
Loss
1.0
0.8
0.6
0.4
0.2
0.0
2
4
6
8
10
Epoch

Conclusion

Demonstrated the vulnerability of Vision Transformers to adversarial attacks and the effectiveness of adversarial training in enhancing their robustness.

While the ViT model initially achieved high accuracy on clean data, its performance was significantly impacted by adversarial examples generated using FGSM and PGD attacks.

THANK YOU