Adversarial Attacks on Vision Transformers

Project Overview

Effectiveness of Adversarial Attacks on ViTs(Vision Transformers)

- Focused on using the ViT-B_16 model
- Convolutional Neural Networks
    - Gold standard in computer vision applications
    - Successful but very vulnerable to adversarial attacks
        - Concerns for security applications

- Different architectures
    - defend against adversarial attacks
- Purpose
    - study the robustness of the ViT architecture or model on different adversarial attack setups

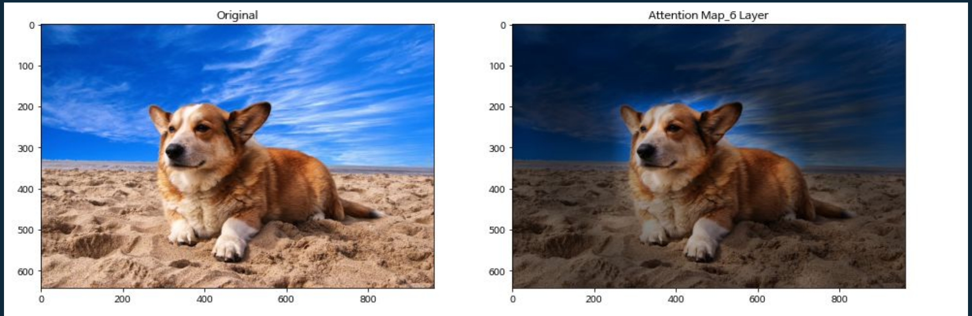
Motivation behind Project

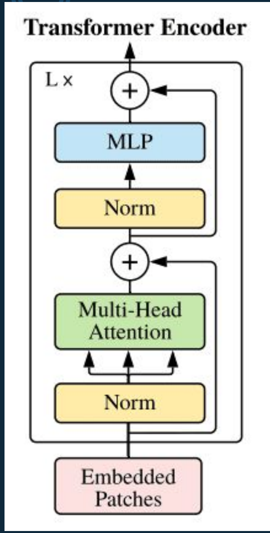
Tricking Machine Learning Models

- AI Systems becoming more integrated with society presents security risks
- Ex:
  - AI Cars
    - Stop sign = speed sign
  - Medical
    - Malignant = Benign

- WAYS TO TRICK
- **Poisoning Attacks-** a dataset or another way to describe this is introducing noise which will slowly make the classifier misidentify by learning certain features
- **Evasion attack**- Intentionally introduce data to deceive an already trained model to make errors
- Studying how adversarial attacks work on AI systems is crucial as AI systems become more integrated with our lives

Vision Transformer Model





ViTs(Vision Transformers) vs CNNs (Convolutional Neural Networks)

- CNNs
  - Uses Pixel Arrays
- ViTs
  - learns by measuring the relationship between input token pairs
- Uses an Attention Map making it more robust
  - patches of images are organized as a token
  - the relationship can be learned by providing attention in the network
- Architecture Model in Image Classification

Types of Attacks

White-Box Setting vs Black Box Setting Attacks

- A whitebox attack
  - complete access to the target model- architecture and its parameters
  - Query-based
  - Transfer based
- A blackbox attack
  - no access to the model- only observe the outputs of the targeted model
- White-box settings introduces one of the strongest possible adversaries that can test the effectiveness of Vision Transformers

ATTACKS TESTED on ViT

- Fast Gradient Sign Method
- Carlini and Wagner
- Deep Fool

Hyperparameters

ImageNet Dataset

ViT-B/16- B-base 16-patch size

- Image: 224
- patch size: 7*7 2
- # patches: 16
- epochs- 25
- Self Attention Layers (Depth)- 6
- Loss- Categorical Cross entropy
    - (the target should be [0,0,0,0,1,0] if the 5 class)
- Learning rate- 0.003
- Epsilon = 0, 0.1, 0.2, 0.3 and 0.4

FSGM/Results

Attack	Epsilon	Vision Transformer Accuracy
No Attack (clean)	0	88.2%
FGSM	0.1	81.7%
FGSM	0.2	78.2%
FGSM	0.3	53.2%
FGSM	0.4	34.2%

- Transformer has overfitted over imagenet
  - self attention layers are fewer than what is required
    - train accuracy is around- 90%
    - test accuracy is around (7-12)%

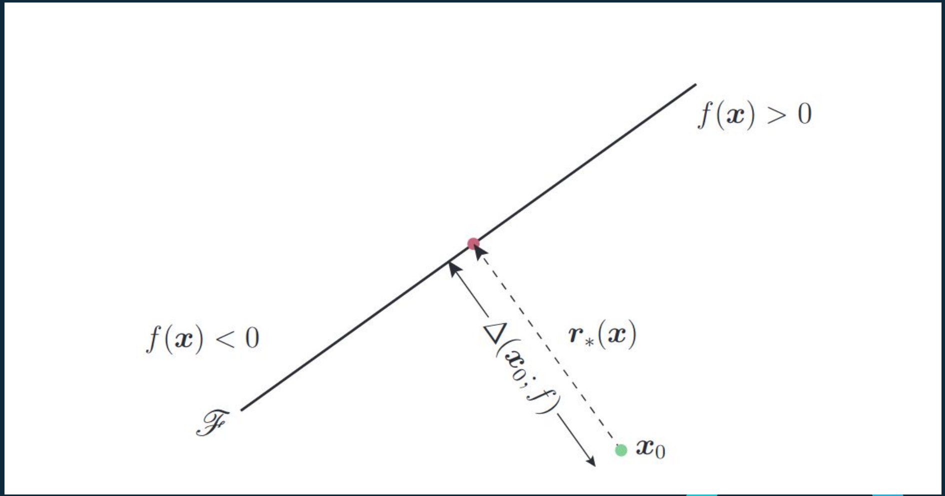
Carlini and Wagner/Results

Clean Accuracy: 76.4
C&W Attack Success Rate: 46.8%

- A model with a lower ASR or a higher l2-distance metric is consider more robust
- ViT-B/16 was able to divert the attack almost 50%
- This attack does not need an epsilon variable

Deep Fool Attack/Results

Linear Binary Classifier for Adversarial Examples



Clean Accuracy: 72.9
DeepFool Attack Success Rate: 54.8%

- A model with a lower ASR or a higher l2-distance metric is consider more robust
- ViT-B/16 was able to divert the attack almost 50%
- This attack was a very simple, yet very effective attack in bypassing a lot models

Clean vs Adversarial

Results

Conclusion

- Vision Transformers are robust at small epsilons
  - But not as epsilon increases
- Patch number
  - Determined how are used in each iteration
  - Small = did not produce strong adversarial examples
- Feature perspective
  - ViTs more reliant on robust features
- ViTs vs CNNs

Further Implementation

- ViTs
  - Still recently new
- More testing on different models
- Different attacks
- Different settings
  - Black-box easier to implement
    - Less or no training time and
    - Less knowledge needed to know of models used and more practical

Works Cited

- https://towardsdatascience.com/deepfool-a-simple-and-accurate-method-to-fool-deep-neural-networks-17e0d0910ac0
- https://blog.floydhub.com/introduction-to-adversarial-machine-learning/#deepfool
- https://portswigger.net/daily-swig/adversarial-attacks-against-machine-learning-systems-everything-you-need-to-know
- https://towardsdatascience.com/adversarial-machine-learning-attacks-and-possible-defense-strategies-c00eac0b395a
- https://viso.ai/deep-learning/vision-transformer-vit/#:~:text=Moreover%2C%20ViT%20models%20outperform%20CNNs,globally%20across%20the%20overall%20image.