Robustness and Adversarial Properties of Vision Transformers

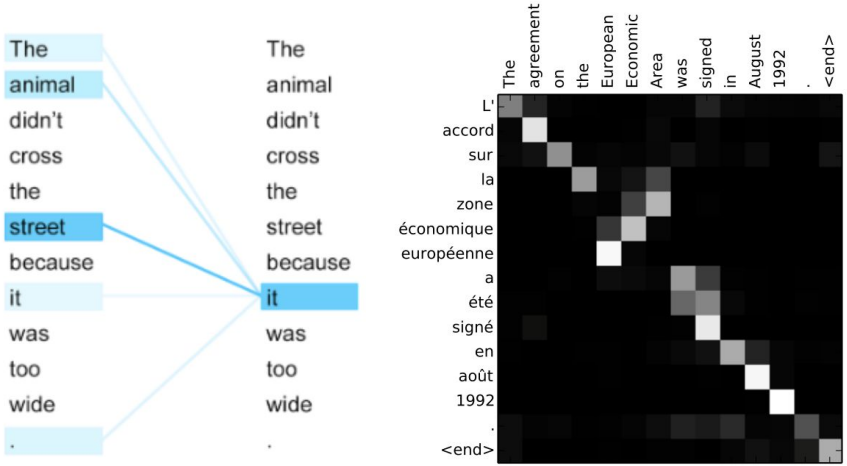
Songlin Liu, Bingzhao Shan, Zihan Wang

Outline

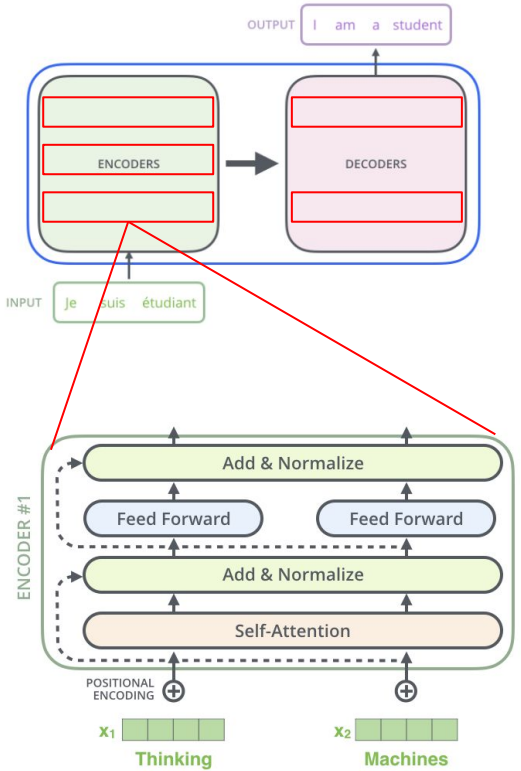
- Introduction
- Related Work
- Methodology
- Experiments Result & Discussion
- Conclusion
- Future work

Introduction

Attention



Transformer



Introduction

Transformers in Computer Vision

- Due to the high dimensional and noisy nature of images, the performance of transformer-based models in the field of computer vision used to be far lower than CNN-based models for a long time…
- Why we need it?
  - Convolution Neural Networks are not perfect!

The pooling operation used in convolutional neural networks is a big mistake and the fact that it works so well is a disaster.

-Geoffrey Hinton

Introduction

Transformers in Computer Vision

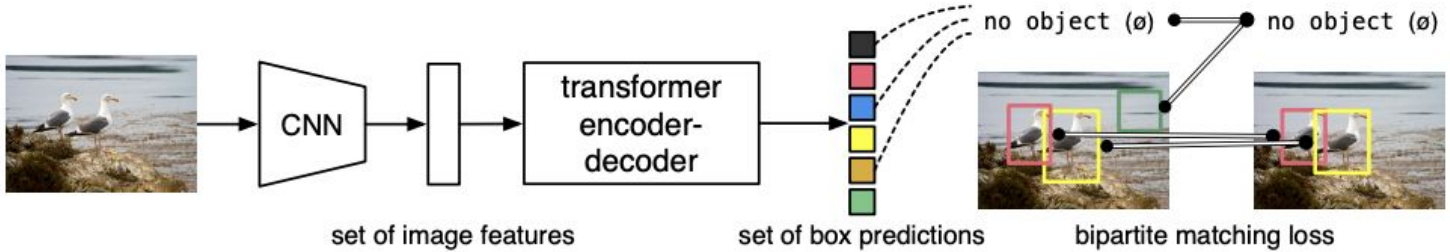
- Why we need it?
  - Convolution Neural Networks are not perfect!

> *The pooling operation used in convolutional neural networks is a big mistake and the fact that it works so well is a disaster.*
>
> *-Geoffrey Hinton*

Introduction

Transformers in Computer Vision

- Transformed Based **Classification**:
  - Vision Transformer (ViT, Dosovitskiy et al., 2020)
- Transformed Based **Detection**:
  - Detection Transformer (DETR, Carion et al., 2020)



Introduction

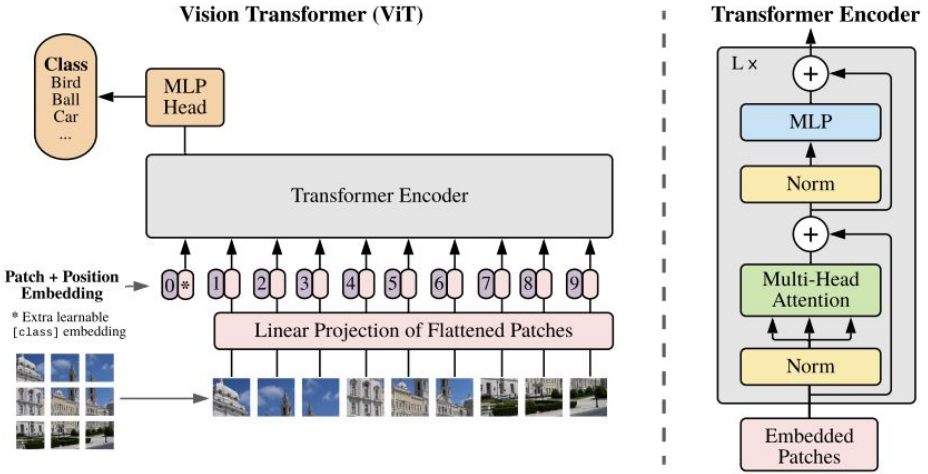
Robustness and Adversarial Property of ViT

- **Robustness of transformer-based models against adversarial examples is still a new research topic that remains to be explored.**
  - Is Vit more robust to attacks we developed for CNN-based models?
  - Is there any fingerprints on adversarial images generated from ViT?
  - If so, how can we utilize such fingerprint to classify clean images and adversarial ones?

Related Work

Vision Transformer (ViT)

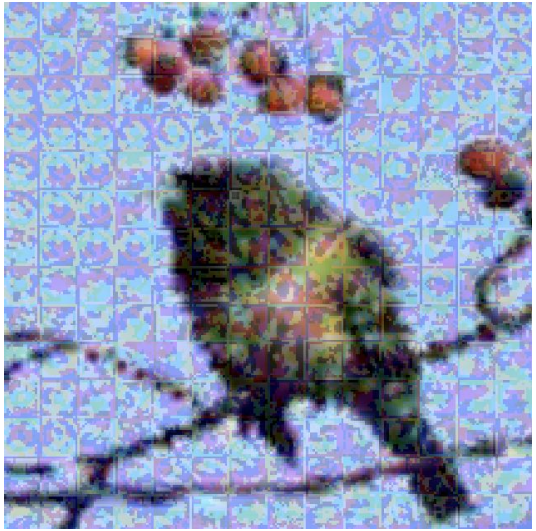
- Patch embedding (16x16)
- Transformer Encoder
- MLP Layer



Related Work

White-box attacks:

- Two L∞ attacks:
  - Fast Gradient Sign Method (FGSM)
  - Projected Gradient Descent (PGD, with L∞ norm)
- Two L2 attacks:
  - Projected Gradient Descent (PGD, with L2 norm)
  - Carlini and Wagner attack (C&W, with L2 norm)



Related Work

Adversarial Detectors

- Detecting real images and CNN-generated fake images:
  - CNN-generated images are surprisingly easy to spot...for now (Image domain, Wang et al., 2020)
  - Leveraging Frequency Analysis for Deep Fake Images (Frequency domain, Frank et al., 2020)
- Detecting clean images and adversarial images
  - On Detecting Adversarial Perturbations (Metzen et al., 2017)
- Why detecting adversarial perturbations is harder?
  - Perturbations added onto images can be very small for classifiers to capture.
  - No predictable artifacts such as the de-convolution checkerboard pattern can be observed in the adversarial samples.

Methodology - Robustness & Transferability

- Various White-box Attack on ViT and ResNet50
- Adversarial training on fine-tuned ViT using various attacking methods.
- Transfer adversarial examples generated from ResNet50 to ViT and vice versa

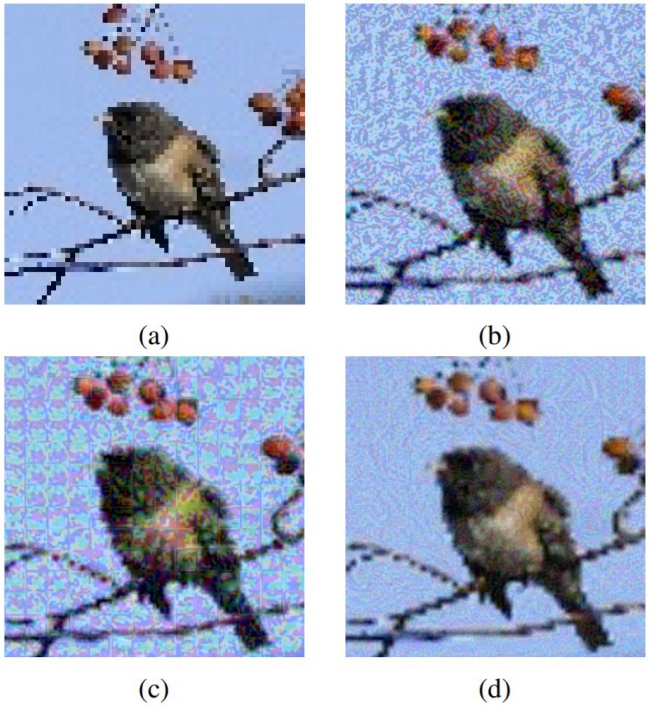
Methodology - Robustness & Transferability

- Observation:

  ViT adversarial examples have unique 14x14 checkerboard pattern (16 x 16 pixels per cell)

- Hypothesis:

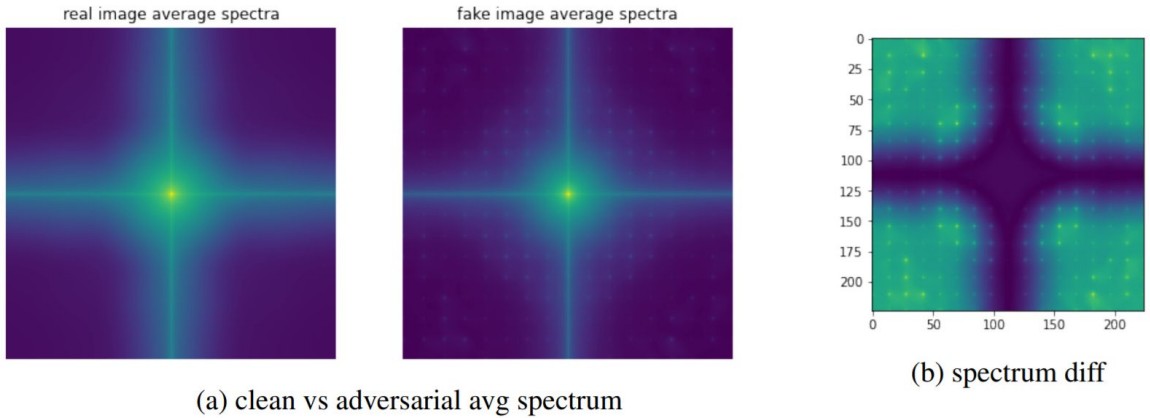
  ViT unique architecture of grid of square patches leads to such checkerboard signatures



Methodology - Frequency Analysis

- Compare average frequency spectra of clean images and adversarial images from various L_2PGD, L_∞FGSM, and L_2C&W attacks.
- Observation:

Dot-pattern in ViT adversarial images in frequency domain



(a) clean vs adversarial avg spectrum

(b) spectrum diff

Figure 3: Spectra after applying no filter on FGSMLinf ($\epsilon = 0.1$) generated adversarial examples.

Methodology - ViT Adversarial Examples Detector

- Preprocessing: convert input image from spatial domain into frequency domain (DCT or DFT)

- A binary classifier with clean images as 1 and adversarial images as 0

Experiment Setups - Dataset

- ImageNet - 16 (our customized ImageNet subset)
    - 16 classes
    - 16,000 training images
    - 1600 validation images
- CIFAR-10

Experiment Setups - Robustness & Transferability

- Model: Fine-tune ViT and ResNet50 on ImageNet-16 and CIFAR-10 respectively

- Selected White-box attack using various ϵ
-

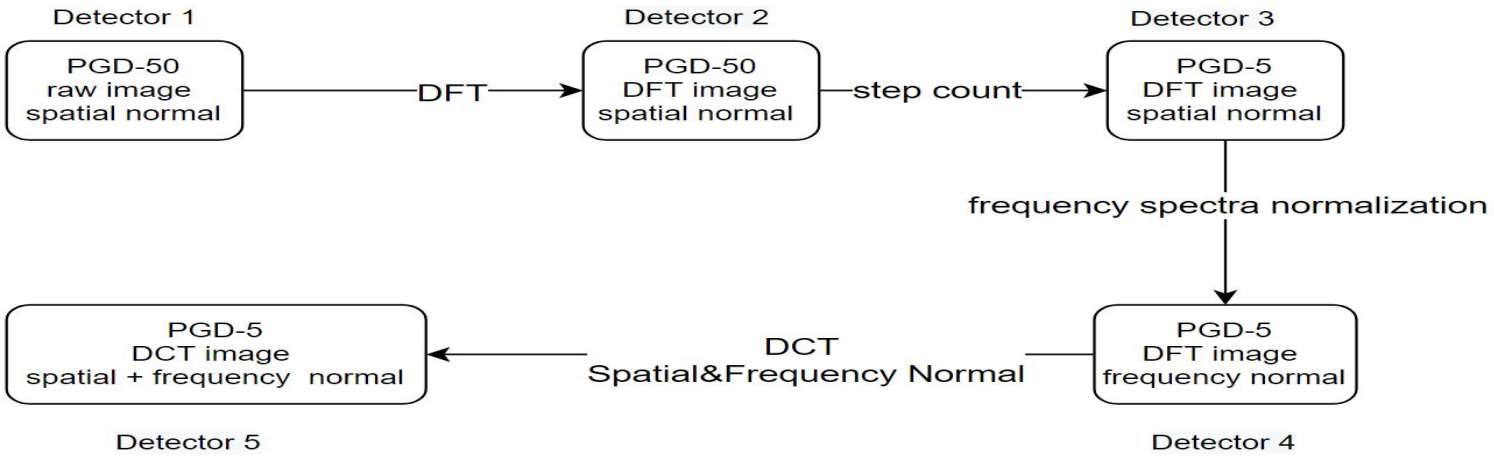
Attack Method	FGSM	PGD-50	PGD-50	C&W
Norm	L_∞	L_∞	L_2	L_2

Experiment Setups - Frequency Analysis

- Obtain average frequency spectra using three different filtering techniques:
    - High pass filter
    - Low pass filter
    - No filter

Experiment Setups - ViT Adversarial Examples Detector

- Baseline Model: ResNet18
- Training set: clean images and adversarial examples from L2 PGD ϵ = 0.5 attack
- Testing set: Adversarial examples from FGSM, L2 PGD-50, L_∞ PGD-50, C&W-40



Experiment Results & Discussion

Whitebox Attacks

- ViT is susceptible to gradient based whitebox attacks
- ViT is still more robust compared to ResNet

	ViT Accuracy		ResNet50 Accuracy	
Attack	ImageNet(tiny)	CIFAR10	ImageNet(tiny)	CIFAR10
clean accuracy	87.2%	93.0%	92.8%	94.4%
L_∞FGSM ($\epsilon = 0.001$)	65.7%	61.0%	67.2%	63.5%
L_∞FGSM ($\epsilon = 0.005$)	33.6%	30.9%	35.1%	21.4%
L_∞FGSM ($\epsilon = 0.01$)	25.3%	25.4%	28.7%	15.0%
L_∞FGSM ($\epsilon = 0.1$)	12.6%	13.3%	27.8%	19.8%
L_∞PGD-50 ($\epsilon = 0.001$)	60.7%	47.8%	59.1%	52.9%
L_∞PGD-50 ($\epsilon = 0.005$)	7.6%	1.3%	1.7%	0.2%
L_∞PGD-50 ($\epsilon = 0.01$)	0.2%	0.0%	0.1%	0.0%
L_∞PGD-50 ($\epsilon = 0.1$)	0.0%	0.0%	0.0%	0.0%
L_2PGD-50 ($\epsilon = 0.1$)	77.6%	77.0%	79.2%	82.1%
L_2PGD-50 ($\epsilon = 0.3$)	53.3%	35.8%	48.0%	41.7%
L_2PGD-50 ($\epsilon = 0.5$)	33.9%	15.2%	24.1%	14.9%
L_2PGD-50 ($\epsilon = 2$)	1.6%	0.2%	0.2%	0.0%
L_2C&W-40 ($\epsilon = 0.1$)	77.8%	78.4%	83.1%	85.9%
L_2C&W-40 ($\epsilon = 0.3$)	35.9%	13.8%	29.6%	26.1%
L_2C&W-40 ($\epsilon = 0.5$)	15.0%	2.6%	6.8%	4.9%
L_2C&W-40 ($\epsilon = 2$)	0.0%	0.0%	0.1%	0.0%

Table 1: White box attacking results of ViT and ResNet50 on ImageNet(tiny) and CIFAR10.

Transferability of Attacks

ViT accuracy on ResNet50 Adversarial Samples

Attack	Accuracy on ViT
Clean Accuracy	87.2%
ResNet50 L_2PGD-50 ($\epsilon = 0.3$)	86.6%
ResNet50 L_2C&W-40 ($\epsilon = 0.3$)	86.8%
ResNet50 L_∞PGD-50 ($\epsilon = 0.005$)	85.7%

ResNet50 accuracy on ViT Adversarial Samples

Attack	Accuracy on ResNet50
Clean Accuracy	92.8%
ViT L_2PGD-50 ($\epsilon = 0.3$)	91.9%
ViT L_2C&W-40 ($\epsilon = 0.3$)	92.1%
ViT L_∞PGD-50 ($\epsilon = 0.005$)	91.6%

Table 2: Result of accuracy of ViT on adversarial samples generated using ResNet50 and accuracy of ResNet50 on adversarial samples generated using ViT on ImageNet tiny.

Adversarial Training

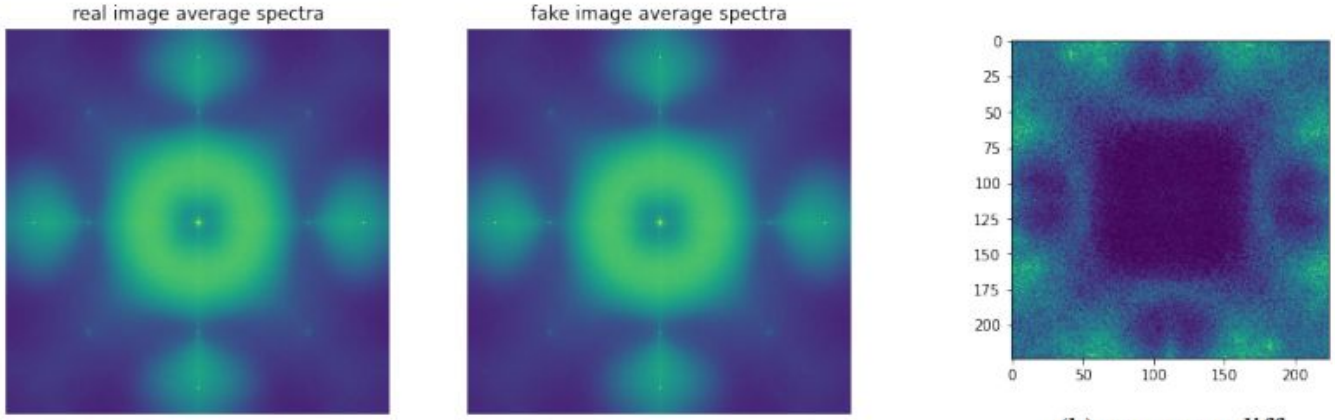
- ViT benefit from adversarial training

Attack(on Original ViT)	Accuracy (adversarially trained on L_2C&W $\epsilon = 0.3$)	Accuracy (adversarially trained on L_∞FGSM, $\epsilon = 0.01$)
Clean Images	77.1% (87.2%)	59.3% (87.2%)
L_∞FGSM ($\epsilon = 0.01$)	46.7% (25.3%)	54.1% (25.3%)
L_∞PGD-50 ($\epsilon = 0.01$)	10.3% (0.2%)	18.4% (0.2%)
L_2PGD-50 ($\epsilon = 0.3$)	59.8% (53.3%)	53.8% (53.3%)
L_2C&W-40 ($\epsilon = 0.3$)	61.4% (35.9%)	54.6% (35.9%)

Table 3: Adversarial training results of ViT. Numbers in the parenthesis indicates the accuracy before adversarial trianing.

Frequency Analysis

- 2PGD and L2C&W's signature not visible, might be due to their perturbation being small to register anything visibly different in frequency domain. The difference still exist, might've been picked up by the detector model



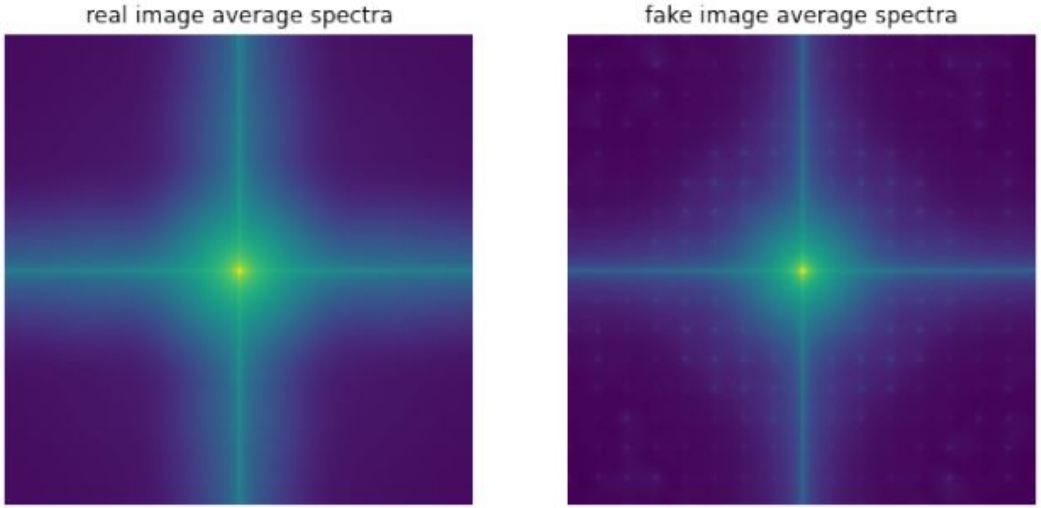
(a) clean vs adversarial avg spectrum

(b) spectrum diff

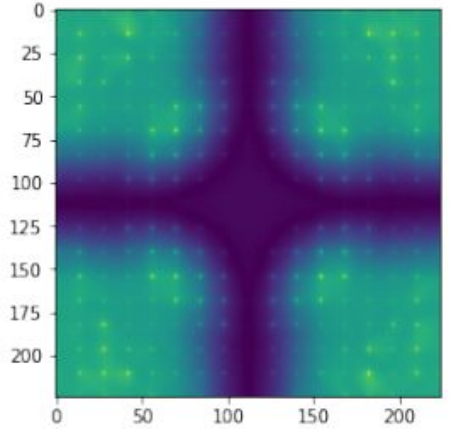
Figure 5: Spectra after applying high pass filter on PGDL2 ($\epsilon = 1$) generated adversarial examples.

Frequency Analysis (Continued)

- L infinity FGSM's signature is visible to naked-eyes in frequency domain as well



(a) clean vs adversarial avg spectrum



(b) spectrum diff

Adversarial Detection

- CNN-based ViT adversarial detectors can be trained with both spatial and frequency domain data, with varied generalizability

	Accuracy				
Validation Set	Detector 1	Detector 2	Detector 3	Detector 4	Detector 5
clean accuracy	100%	100%	100%	100%	100%
L_2PGD-50($\epsilon = 0.1$)	100%	100%	0.0%	100%	100%
L_2PGD-50($\epsilon = 0.5$)	100%	100%	0.0%	100%	100%
L_2PGD-50($\epsilon = 2$)	99.9%	82.6%	0.0%	100%	0.1%
L_∞FGSM($\epsilon = 0.01$)	0.1%	0.0%	0.0%	100%	0.0%
L_∞FGSM($\epsilon = 0.1$)	0.0%	0.0%	0.0%	0.0%	0.0%
L_2C&W-40($\epsilon = 0.5$)	68.1%	86.5%	0.0%	0.1%	86.8%
L_∞PGD-50($\epsilon = 0.005$)	100%	100%	0.0%	0.0%	99.8%
L_∞PGD-50($\epsilon = 0.01$)	33.5%	1.8%	0.0%	100%	0.0%

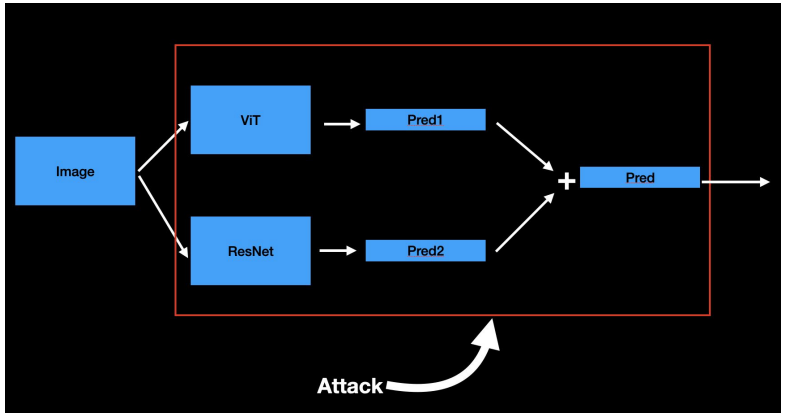
Table 7: The Detectors' accuracy on various examples. Detector 1 is trained on PGD-50 (raw images) with spatial domain normalization, Detector 2 is trained on PGD-50($\epsilon = 0.5$) (DFT images) with spatial domain normalization, Detector 3 is trained on PGD-5($\epsilon = 0.5$) (DFT images) with spatial domain normalization, Detector 4 is trained on PGD-5($\epsilon = 0.5$) (DFT images) with frequency domain normalization, Detector 5 is trained on PGD-5($\epsilon = 0.5$) (DCT images) with both spatial and frequency domain normalization.

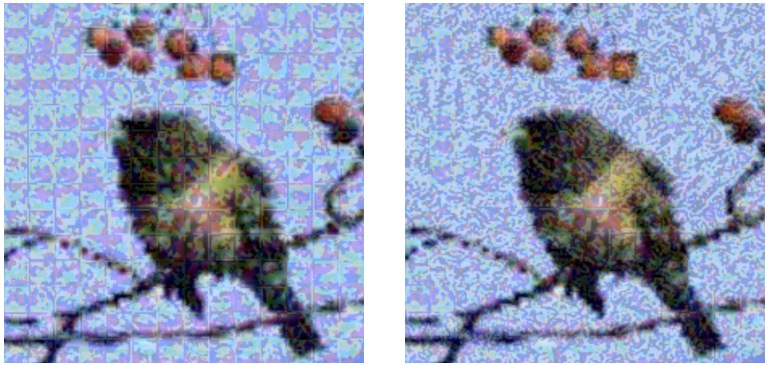
Conclusion

- ViT is susceptible to gradient based whitebox attacks
- ViT is still more robust compared to ResNet
- ViT benefit from adversarial training
- There is distinct grid like signature on whitebox adversarial examples of ViT
- L infinity FGSM’s signature is visible to naked-eyes in frequency domain as well
- L2PGD and L2C&W’s signature not visible, might be due to their perturbation being small to register anything visibly different in frequency domain. The difference still exist, might’ve been picked up by the detector model
- CNN-based ViT adversarial detectors can be trained with both spatial and frequency domain data, with varied generalizability

Future Works

- Test on full version ImageNet
- Better preprocessing on frequency domain
- Explore more on the generalizability of the detectors
- Making the checkerboard fingerprint less obvious without modifying the attacking method or influencing attack success rate too much by utilizing ensemble model of ViT and ResNet (i.e. still FGSM with epsilon=0.1, less checkerboard pattern, similar attacking success rate).





Two adversarial samples that fooled ViT successfully.
Left: FGSM epsilon=0.1 on ViT.
Right: FGSM epsilon=0.1 on Ensemble Model.

Thank you!