Robustness and Adversarial Properties of Vision Transformers

Songlin Liu, Bingzhao Shan, Zihan Wang
{slili,shanbz,yurihime}@umich.edu
Depertment of Computer Science and Engineering
University of Michigan

Abstract

Transformer-based approaches have been an active area of research. Recently, the success of state-of-the-art transformer-based model Vision Transformer (ViT) [3] in computer vision related tasks has attracted many research interest. In this work, we study the robustness ViT in the following aspects: (1) Robustness under white-box attacks and the transferability between different attack algorithms (such as FGSM, PGD, etc) and ResNet50. (2) Frequency analysis in identifying ViT adversarial examples (3) Training a detector attempts to distinguish between clean images and ViT adversarial images.

1 Introduction

Transformer-based models have been a great success in the field of natural language processing since the attention mechanism was first proposed in 2017 by Vaswani et al [10]. As a brief summary, transformers are models with a stack of encoders and decoders each consisting of a self-attention layer and a feed-forward layer. The intuition behind such architecture is helping models to pay attention to the spatial relation between each individual word more effectively than existing approaches such as LSTM.

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. Recently, several novel works have presented significant improvements in vision tasks including DETR (Detection Transformer) for object detection and segmentation [1], Image Transformer for image generation and super-resolution [8], and most popular of them all: ViT (Vision Transformer) for image classification [3], etc. Out of the interest in the unique architecture of the Vision Transformer, we will focus on ViT in our work for all experiments and analysis.

Although transformer-based models achieved great success in some computer vision tasks, the robustness of transformer-based models against adversarial examples is still a new research topic that remains to be explored. Existing white-box adversarial attacks mainly focus on the gradients used in back-propogation. ViT also uses back-propagation to update its weights, thus we expect that these attack might very likely succeed in ViT as well. Traditionally, adversarial examples in CNN-based models are very difficult to detect. However, since the architecture of ViT is significantly different from CNN-based models, it could leave a specific trace that made its adversarial examples easier to be identified. To prove our theory, we conducted a series of experiments to compare the robustness and transferability of adversarial examples between ViT and CNN-based ResNet50. After observing the grid square pattern in the adversarial samples generated on ViT, we also performed frequency analysis and attempted to build a ViT adversarial example detector using a simple binary classifier. Our experiments results made the following contributions:

Preprint. Under review.

- We found that existing white-box attacks have similar performance on ViT and ResNet50, however, ViT seems to be more robust in our experiments.
- The transferability between ViT and ResNet50's adversarial examples is quite low.
- Adversarial examples of ViT have distinctive checkerboard patterns.
- Detector for ViT adversarial examples

2 Related Works

Vision Transformer Inspired by Transformers' success in NLP, researchers started to apply Transformers as hybrids or alternatives of CNNs in computer vision. In this work, we focus on a state-of-the-art transformer-based image classification model, the Vision Transformer (ViT). As shown in Figure1, ViT divides an input image into 16x16 grid of squares patches, similar to dividing words in a sentence, and unrolls them into sequences. These patches along with their positions would be fetched into alternating layers of multi-head attention layers. As opposed to CNN, ViT can learn integrated spatial information across the entire images using positional embeddings and global attention. Also, it outperforms the state-of-the-art CNNs on image classification with significant less model complexity.

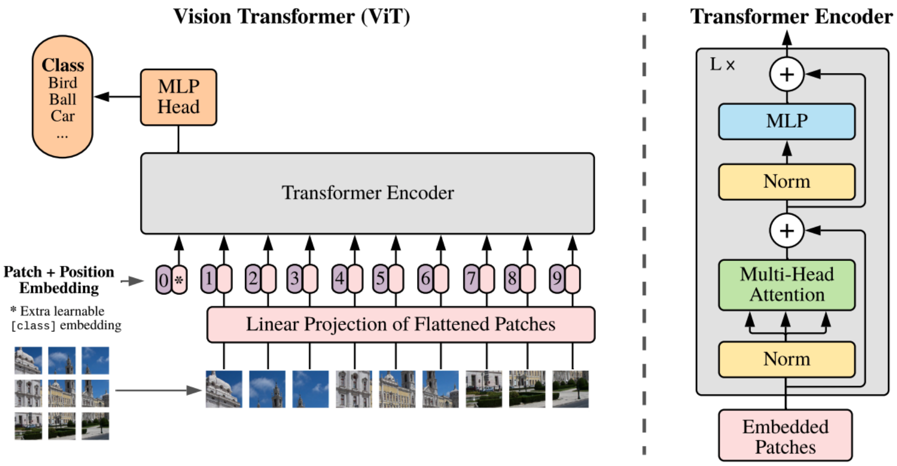


Figure 1: The architecture of Vision Transformer.

White-box attack White-box attacks have achieved great success in confusing neural networks by adding small perturbations on the input images. Based on the type of norm these algorithms adopts, bounded white-box attacks can be further divided into several sub-categories: L_0 attacks, L_1 attacks, L_2 attacks, and L_∞ attacks. In this work, we measured the robustness of ViT under two L_2 attacks and two L_∞ attacks with a range of different clipping thresholds (epsilons):

Two L_∞ attacks we used in this work:

- Fast Gradient Sign Method (FGSM) [5].
- Projected Gradient Descent (PGD, with L_∞ norm) [6].

Two L_2 attacks we used in this work:

- Projected Gradient Descent (PGD, with L_2 norm) [6].
- Carlini and Wagner attack (C&W, with L_2 norm) [2].

Detecting adversarial samples using binary classifier Previous work has shown that binary classifiers trained on large-scale datasets can identify real and CNN-generated images effectively [11]. Moreover, leveraging frequency analysis also provides a powerful approach to identify fake images from real ones by catching specific fingerprints of fake images [4]. However, identifying adversarial images from real ones is a much more challenging task due to the following reasons: (1) Perturbations added onto images can be very small for classifiers to be captured. (2) No predictable artifacts such as the de-convolution checkerboard pattern can be observed in the adversarial samples. Metzen et.al [7] proposed a method for training a binary classifier by extracting features at specific layers of the original model as input to the adversarial classifier. Although the classification accuracy was good in their paper, the classifier failed to generalize to other attacking methods very well. In this work, we show that the unique fingerprints on adversarial samples generated on ViT help the binary classifier to generalize better to a specific set of unseen attacking methods. In this work, we design a similar attack-specific detector that uses a binary classifier to distinguish between clean images and adversarial images generated by attack on ViT.

3 Methodology

Adversarial examples from white-box attacks on ViT have shown checkerboard pattern (as shown in Figure2) and our hypothesis is that such pattern is caused by ViT's patch embedding. With such hypothesis, we perform frequency analysis and train of CNN-based detector on adversarial examples from attacking ViT.

3.1 White-box Attack & Transferability & Adversarial Training

We attack both ViT and ResNet50 using L_∞FGSM, L_2PGD, L_∞PGD, and L_2C&W using various epsilon values. Other than this, we also evaluate the performance of ViT on adversarial samples generated from ResNet50 and vice versa. From 2(c) we can see a clear 14x14 checkerboard pattern from L_{inf}, $\epsilon = 0.1$ attack on ViT. Though less obvious, same pattern also exists in 2(d) L_2PGD, $\epsilon = 5$ attack on ViT. This checkerboard pattern does not exhibits in the 2 (a) original images and 2 (b) ResNet50. Thus, we make the hypothesis that ViT unique architecture of grid of square patches leads to the checkerboard signatures in ViT adversarial examples.

To further measure the robustness of ViT, we do adversarial training on our fine-tuned ViT for another 4 epochs adversarial samples generated using different attacking methods. The results show that ViT's performance improves a lot on the adversarial samples after adversarial trainning.

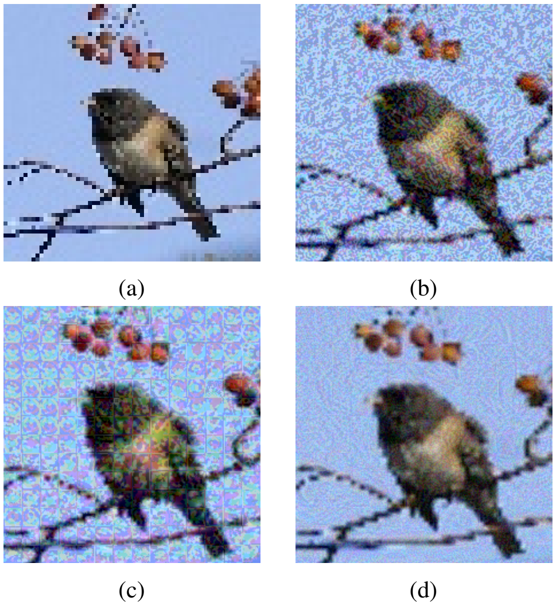
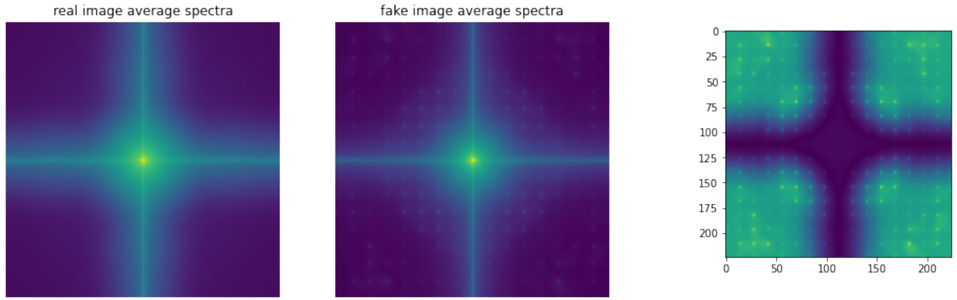


Figure 2: (a) Clean Image (b) An adversarial image generated from ResNet50 with L_∞FGSM, ϵ=0.1 (c) An adversarial image generated from ViT with L_∞FGSM, $\epsilon = 0.1$ (d) An adversarial image generated from ViT with L_2PGD-50, $\epsilon = 5$. It is obvious that adversarial samples generated from ViT have unique checkerboard pattern that those generated from ResNet50 do not have.

3.2 Frequency Analysis

We find that in all the white-box attacks tested on ViT, the adversarial examples with larger perturbations have visible gird-like signatures. Such signatures resemble 14x14 grids, in which each cell consists of 16x16 pixels. The cell size is the same as the ViT model's patch size. To check if such signature in the spatial domain is also presented in the frequency domain, we perform frequency analysis on clean examples and adversarial examples generated by L_2PGD, L_∞FGSM, and L_2C&W attacks. As shown in Figure3, the dot-pattern in frequency domain is obvious in adversarial samples generated from ViT by FGSMLinf ($\epsilon = 0.1$).



(a) clean vs adversarial avg spectrum

(b) spectrum diff

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

3.3 ViT Adversarial Examples Detector

We hypothesize that ViT's patch embedding might be the cause of the checkerboard patterns and they can be used to distinguish between clean images and adversarial images in advance. Even though these patterns are easily observable by human when purtabations are large, such as FGSM in Figure2, as perturbations decrease in more efficient attack methods, such as PGD and C&W, we need to rely on the classifier models to tell them apart.

Since adversarial samples generated from ViT also have unique fingerprints which might be informative for the adversarial detection, we employ the same method as [11] to distinguish clean images and adversarial images - a simple binary classifier with clean images as 1 and adversarial images as 0 [11]. Frank et al. has shown that these some artifacts can be more easily identified in frequency space instead of image space[4]. Therefore, in some of our detector's training configurations, we also adopt some pre-processing steps such as discrete Fourier transform (DFT) and discrete Cosine transform (DCT) to exploit the promising features of frequency domain images.

Frequency Spectra of real training images and adversarial images from L_2PGD attack with $\epsilon = 0.5$ are selected as our training set. PGD with $\epsilon = 0.5$ is selected because its attack efficiency is relative in the middle among all attacks tested, so it allows us to test the generalizability of the detector against unseen attack with both larger and smaller perturbations.

4 Experiment Setups

4.1 Dataset

We used two datasets in our experiments: a customized ImageNet subset with 16 classes (we call it ImageNet-16) and CIFAR10. ImaegNet-16 consists of 16k training images and 1.6k validation images. Since the image resolution(32x32) of CIFAR10 is far below the requirement of ViT(224x224), we focus on ImageNet-16 in most of our experiments.

4.2 White-box attacks & Transferability & Adversarial Training

We first fine-tuned both ViT and ResNet50 on ImageNet-16. After fine-tuning, we used the foolbox adversarial attack library [9] to perform the attacks. For each attack methods, we set several different epsilons , attacked on the validation set, and calculate the accuracies.

For transferability experiments, we only used three configurations: L$_2$PGD-50 ($\epsilon = 0.3$), ResNet50 L$_2$C&W-40 ($\epsilon = 0.3$), and ResNet50 L$_2$C&W-40 ($\epsilon = 0.3$) since they gave the most strong attack results without adding too much artifacts in the original images. We first evaluate ViT on adversarial samples generated by ResNet50 and then the other direction. Results can be found in Section 5.1.

To further measure the robustness of ViT, we do adversarial training on our fine-tuned ViT for another 4 epochs adversarial samples generated by L_2C&W (epsilon=0.3) and L_∞FGSM (epsilon=0.01).

4.3 Frequency Analysis

We attack the ViT model fine-tuned on our customized tinyImageNet dataset (16 classes), and perform various pre-processings on gray-scaled adversarial images from the validation set. In the first set of experiment, we apply high pass filter (subtract the median blurred version). In the second set, we apply low pass filter (median blurred version). In the third set, no filter is applied. After the pre-processings, we perform discrete Fourier transform (DFT) on the result images, and convert them to averaged spectrum to check if there is visible signature. We perform the same process for the clean examples, just to form side by side comparison of the spectra.

4.4 Adversarial Examples Detector

We select a relative simple CNN classifier - ResNet18 as our detector base model. Training set consists of real image from our customized tiny ImageNet training set and adversarial examples from L_2 norm PGD attack with $\epsilon = 0.5$. Testing set consists of adversarial images from previous white-box attack experiment. More specifically, to test the generalization of our detector against different levels of perturbation and white-box attacks, we selected adversarial examples from FGSM($= 0.1, 0.01$), L_2norm PGD-50($\epsilon = 0.1, 0.3, 0.5, 2$), L_∞norm PGD-50($= 0.01, 0.005$), L_2norm C&W-40 ($\epsilon = 0.5$).

To further investigate which factor could influence the performance of classifier, we train 5 detectors using different step count, features extraction, pre-processing and normalization techniques. First detector is the base line model that uses raw adversarial examples from L_2 norm PGD-50 attack ($\epsilon = 0.5$, step size = 0.0125). On top of first detector, second detector adds a pre-processing step that uses DFT to convert raw image into frequency spectra. On top of the second model, the third detector replace PGD-50 with L_2 norm PGD-5 attack ($\epsilon = 0.5$, step size = 0.15). This change increases PGD attack maximum perturbation from 125% into 150%. On top of the third detector, we remove the raw images normalization in spatial domain and add frequency spectra normalization in frequency domain instead. One top of detector 4, we add have both spatial and frequency domain normalization and replace DFT with DCT.

5 Experiment Results

5.1 White-box attacks & Transferability & Adversarial Training

Table 1 shows the evaluation results on different white-box attacks on ViT and ResNet50 respectively. Although the clean accuracy of ResNet50 is higher than that of ViT, ViT's classification accuracy drops slower as epsilon goes up compared with ResNet50 in most cases other than L_∞FGSM. We conclude that ViT is more robust since L_∞FGSM with epsilon 0.01 and 0.1 added too much noise into the image while the perturbation generated by other attacking methods are invisible to naked eye.

As for transferability, according to the result shown in Table 2, the adversarial samples generated from these two model don not transfer well. The models' accuracy only drop less than 2% in both cases.

We also tried adversarial training on the ViT model. After normal training, we trained our model on adversarial samples generated by by L$_2$C&W ϵ = 0.3 and L$_\infty$FGSM, ϵ = 0.01 respectively for another 4 epochs. As shown in Table3, the robustness for ViT increases a little but it harms the accuracy on the original dataset as well.

	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$_2$PGD-50 ($\epsilon = 0.1$)	77.6%	77.0%	79.2%	82.1%
L$_2$PGD-50 ($\epsilon = 0.3$)	53.3%	35.8%	48.0%	41.7%
L$_2$PGD-50 ($\epsilon = 0.5$)	33.9%	15.2%	24.1%	14.9%
L$_2$PGD-50 ($\epsilon = 2$)	1.6%	0.2%	0.2%	0.0%
L$_2$C&W-40 ($\epsilon = 0.1$)	77.8%	78.4%	83.1%	85.9%
L$_2$C&W-40 ($\epsilon = 0.3$)	35.9%	13.8%	29.6%	26.1%
L$_2$C&W-40 ($\epsilon = 0.5$)	15.0%	2.6%	6.8%	4.9%
L$_2$C&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.

ViT accuracy on ResNet50 Adversarial Samples

Attack	Accuracy on ViT
Clean Accuracy	87.2%
ResNet50 L$_2$PGD-50 ($\epsilon = 0.3$)	86.6%
ResNet50 L$_2$C&W-40 ($\epsilon = 0.3$)	86.8%
ResNet50 L$_\infty$PGD-50 ($\epsilon = 0.005$)	85.7%

ResNet50 accuracy on ViT Adversarial Samples

Attack	Accuracy on ResNet50
Clean Accuracy	92.8%
ViT L$_2$PGD-50 ($\epsilon = 0.3$)	91.9%
ViT L$_2$C&W-40 ($\epsilon = 0.3$)	92.1%
ViT L$_\infty$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.

Attack(on Original ViT)	Accuracy (adversarially trained on L$_2$C&W $\epsilon = 0.3$)	Accuracy (adversarially trained on L$_\infty$FGSM, $\epsilon = 0.01$)
Clean Images	77.1% (87.2%)	59.3% (87.2%)
L$_\infty$FGSM ($\epsilon = 0.01$)	46.7% (25.3%)	54.1% (25.3%)
L$_\infty$PGD-50 ($\epsilon = 0.01$)	10.3% (0.2%)	18.4% (0.2%)
L$_2$PGD-50 ($\epsilon = 0.3$)	59.8% (53.3%)	53.8% (53.3%)
L$_2$C&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.

5.2 Frequency Analysis

When applying high pass filter, for L_2PGD and $L_2C\&W$ attacks, there is no visible difference between the clean images average spectrum and the adversarial images average spectrum. However, for $L_{inf}FGSM$ attacks, the difference between clean images average spectrum and the adversarial images average spectrum is quite noticeable.

When applying low pass filter, again, no visible clean vs adversarial average spectrum difference is found for L_2PGD and $L_2C\&W$ attacks. For $L_\infty FGSM$ attacks, although not quite obvious, there is visible difference when comparing clean and adversarial average spectrum. As one can see in Figure 13a, the adversarial average spectrum's center light right is larger, and the clean average spectrum's center light right has clearer edge.

When not applying any filter, there is also no visible difference between clean and adversarial spectrum for L_2PGD and $L_2C\&W$ attacks. However, the clean and adversarial contrast is quite obvious for $L_{inf}FGSM$ attacks.

Attack	Accuracy on clean ViT	Max abs diff vs. clean spectrum
L$_2$PGD-50 ($\epsilon = 0.5$)	18.4%	0.033
L$_2$PGD-50 ($\epsilon = 1$)	12.8%	1.03
L$_2$PGD-50 ($\epsilon = 2$)	1.3%	2.72
L$_\infty$FGSM ($\epsilon = 0.01$)	26.1%	4.71
L$_\infty$FGSM ($\epsilon = 0.1$)	14.2%	23.33
L$_2$C&W-40 ($\epsilon = 1$)	2.2%	0.18
L$_2$C&WL2 ($\epsilon = 2$)	0.1%	0.18
L$_2$C&WL2 ($\epsilon = 5$)	0.1%	0.18

Table 4: Result of high pass filter frequency analysis. The attack examples are generated from validation set, maximum absolute value differences of attack examples' average spectrum vs clean examples' average spectrum are evaluated on 0-255 scale

Attack	Accuracy on clean ViT	Max abs diff vs. clean spectrum
L$_2$PGD-50 ($\epsilon = 2$)	1.1%	1.14
L$_\infty$FGSM ($\epsilon = 0.1$)	14.2%	17.39
L$_2$C&W-40 ($\epsilon = 1$)	2.1%	0.15

Table 5: Result of low pass filter frequency analysis. The attack examples are newly generated from validation set.

Attack	Accuracy on clean ViT	Max abs diff vs. clean spectrum
L$_2$PGD-50 ($\epsilon = 0.5$)	18.4%	0.04
L$_\infty$FGSM ($\epsilon = 0.1$)	2.5%	53.46
L$_2$C&W-40 ($\epsilon = 1$)	18.4%	0.06

Table 6: Result of no filter frequency analysis. The attack examples are newly generated from validation set.

5.3 Adversarial Examples Detector

From 7, all detectors correctly classify clean images from validation set and all detectors except Detector 3 can perfectly classify adversarial examples from the attack - L_2PGD-50($\epsilon = 0.5$, they are trained on. Our baseline model Detector 1 misclassifies all adversarial examples from attack with large perturbation, including FGSM and L_∞PGD-50($\epsilon = 0.01$), as clean images. For attacks with smaller perturbation, including L_2C&W-40 and L_∞PGD-50($\epsilon = 0.005$), Detector 1 can identify some of the adversarial images. Specifically, the smaller the perturbation, the better Detector 1 generalizes. From 7, we can also see Detector 2 behavior is similar to Detector 1. However, it completely misclassifies all attacks that it is not trained on. Similar to Detector 1 & 2, Detector 4 perfectly classifies adversarial examples from all L_2PGD attacks we tested. Opposite to Detector 1, Detector 4 only generalize well to attack with larger perturbation (with the exception of L_∞FGSM). The almost zero accuracy in L_2 PGD-50 ($\epsilon = 2$) in Detector 5 shows that it only generalize well to smaller perturbed images regardless which attack it is trained on.

	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.

6 Discussion & Conclusion

6.1 Robustness Analysis

In this work, we showed that ViT is more robust under adversarial attacks compared to ResNet50 when perturbations are relatively small. From our white-box attack experiment, ViT's accuracy drops slower as ϵ increases gradually. We also showed that the transferability between ViT and ResNet50 are very poor on ImageNet-16. As for adversarial training, ViT's robustness increases when we do adversarial training but the accuracy on clean samples is harmed heavily.

6.2 Frequency Analysis

For frequency analysis, we observed a special dot patterns in the spectrum generated by adversarial samples from L_∞FGSM. However, for other attacks (such as L_2PGD-50 and L_2C&W-40) with smaller perturbations, such pattern cannot be captured. One reason for such observation is that the signature of L_∞FGSM attacks on ViT is much more visible to naked eye compared to the other two type of attacks, which needs careful inspection or magnification. Our hypothesis is that the perturbation in frequency domain for other attacks is relatively small to visually change the spectrum. Nevertheless, some of our detector configurations successfully picked up the unnoticeable signatures of these attacks in the frequency domain.

6.3 Adversarial Example Detector

The only difference between the training setup of detector 1 and 2 is the domain on which the data is presented. For detector 1, we feed it images on spatial domain, for detector 2, we feed it images on frequency domain. They have similar performance across all attacks tested. Our hypothesis is that the detector model, which is ResNet, is able to capture some frequency domain features directly from the spatial domain input during feature extraction in the convolutional layers. For detector 3, 4 and 5, they used L_2PGD-5 as training set. the difference lies between when the zero mean normalization is performed, and the spatial to frequency domain conversion applied. We found that for L_2PGD with smaller number of steps, frequency domain normalization is quite necessary to ensure the detector even works. There is also difference between using DFT and DCT for spatial to frequency domain conversion, as we can see although both detector 4 and 5 generalize to attacks they are not trained on, the direction of generalization is quite different.

1. features extraction ahead of time is necessary 2. In general, our detector generalizes well to the attack it is trained on 3.

7 Future Works

Due to limitation in time and computational resource, there are several experiments we would like to leave as our future work:

- Moving our experiment to the complete ImageNet dataset.
- We used ResNet18 as our adversarial detector in both image and frequency domain. However, doing convolution in the frequency domain may average out some important features. One future work can be replacing ResNet with ridge regression in the frequency domain with more pre-processings to get rid of unimportant distracts.
- Attacks with large epsilon values may leave strong checkerboard pattern in the adversarial images due to the patch embedding step in ViT. However, adversarial images generated from a ensemble model that consists of ViT and ResNet may offset the checkerboad pattern while preserving high attack accuracy on ViT.

References

[1] Nicolas Carion et al. *End-to-End Object Detection with Transformers*. 2020. arXiv: 2005.12872 [cs.CV].

[2] Nicholas Carlini and David Wagner. *Towards Evaluating the Robustness of Neural Networks*. 2017. arXiv: 1608.04644 [cs.CR].

[3] Alexey Dosovitskiy et al. "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale". In: *International Conference on Learning Representations*. 2021. URL: https://openreview.net/forum?id=YicbFdNTTy.

[4] Joel Frank et al. *Leveraging Frequency Analysis for Deep Fake Image Recognition*. 2020. arXiv: 2003.08685 [cs.CV].

[5] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. *Explaining and Harnessing Adversarial Examples*. 2015. arXiv: 1412.6572 [stat.ML].

[6] Aleksander Madry et al. *Towards Deep Learning Models Resistant to Adversarial Attacks*. 2019. arXiv: 1706.06083 [stat.ML].

[7] Jan Hendrik Metzen et al. *On Detecting Adversarial Perturbations*. 2017. arXiv: 1702.04267 [stat.ML].

[8] Niki Parmar et al. "Image Transformer". In: *CoRR* abs/1802.05751 (2018). arXiv: 1802.05751. URL: http://arxiv.org/abs/1802.05751.

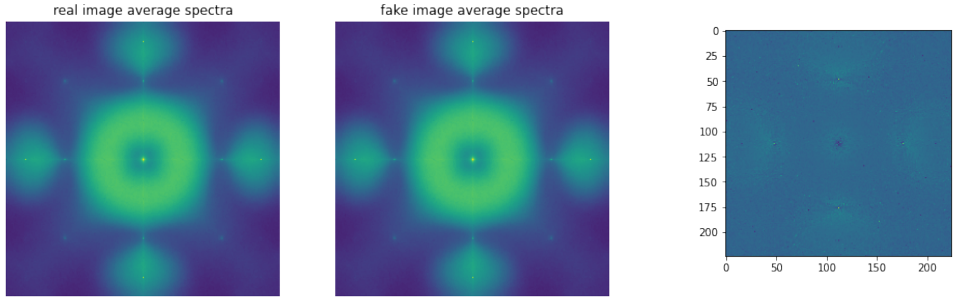
[9] Jonas Rauber et al. "Foolbox Native: Fast adversarial attacks to benchmark the robustness of machine learning models in PyTorch, TensorFlow, and JAX". In: *Journal of Open Source Software* 5.53 (2020), p. 2607. DOI: 10.21105/joss.02607. URL: https://doi.org/10.21105/joss.02607.

[10] Ashish Vaswani et al. *Attention Is All You Need*. 2017. arXiv: 1706.03762 [cs.CL].

[11] Sheng-Yu Wang et al. *CNN-generated images are surprisingly easy to spot... for now*. 2020. arXiv: 1912.11035 [cs.CV].

8 Appendix

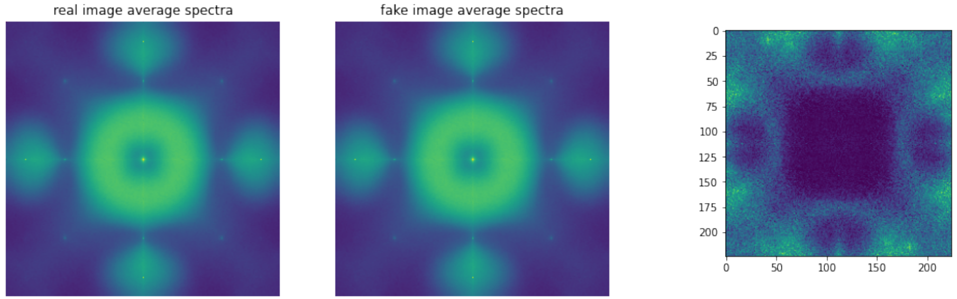
8.1 Additional Frequency Analysis Spectrum



(b) spectrum diff

(a) clean vs adversarial avg spectrum

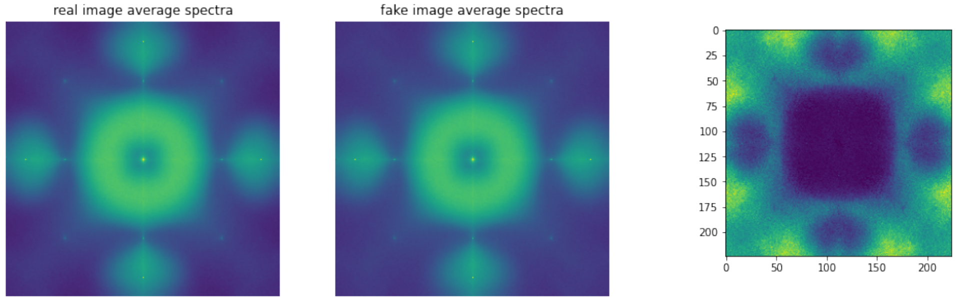
Figure 4: Spectra after applying high pass filter on PGDL2 ($\epsilon = 0.5$) generated adversarial examples. The spectra absolute difference is calculated by abs(adv avg spectrum - clean avg spectrum)



(b) spectrum diff

(a) clean vs adversarial avg spectrum

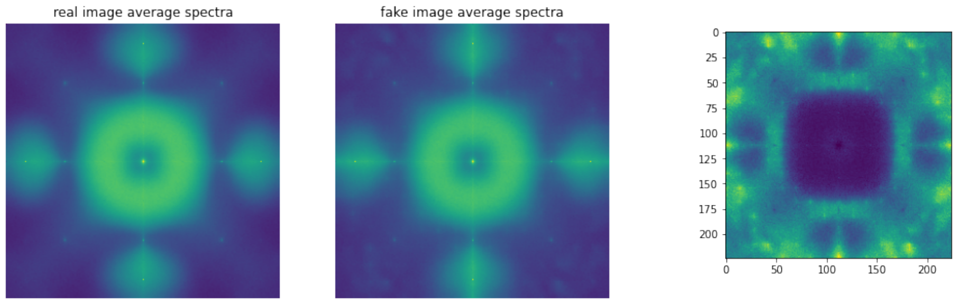
Figure 5: Spectra after applying high pass filter on PGDL2 ($\epsilon = 1$) generated adversarial examples.



(b) spectrum diff

(a) clean vs adversarial avg spectrum

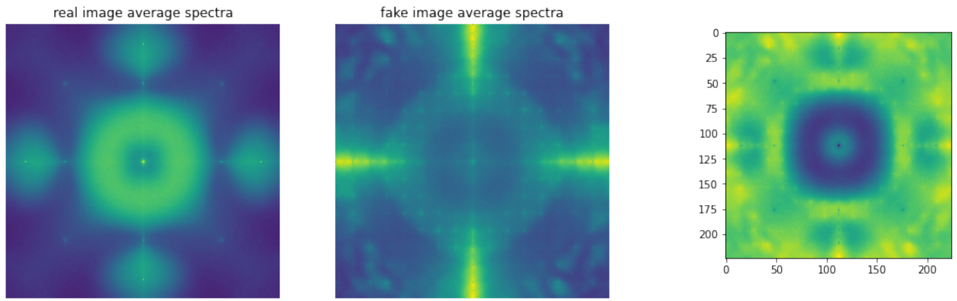
Figure 6: Spectra after applying high pass filter on PGDL2 ($\epsilon = 2$) generated adversarial examples.



(b) spectrum diff

(a) clean vs adversarial avg spectrum

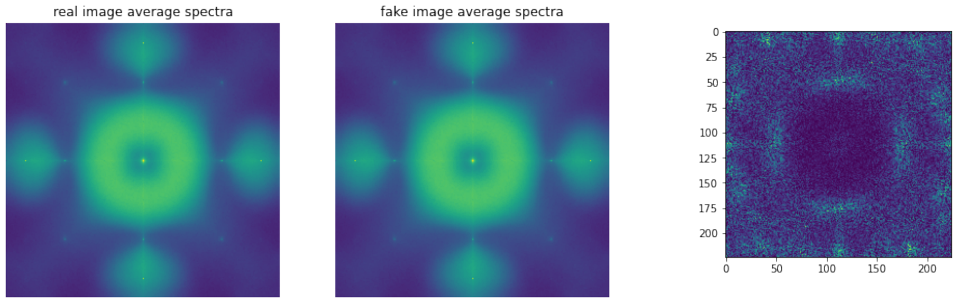
Figure 7: Spectra after applying high pass filter on FGSMLinf ($\epsilon = 0.01$) generated adversarial examples.



(b) spectrum diff

(a) clean vs adversarial avg spectrum

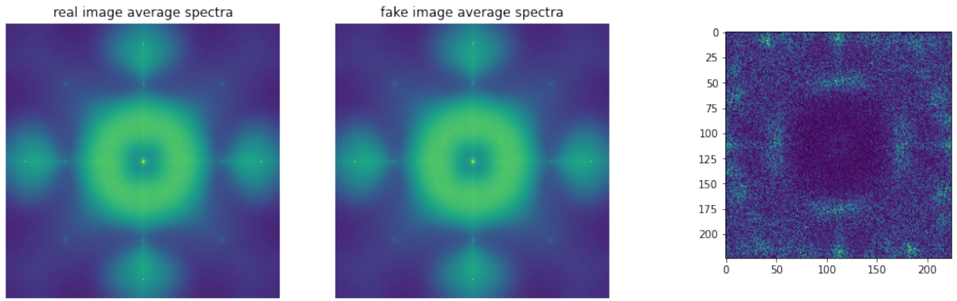
Figure 8: Spectra after applying high pass filter on FGSMLinf ($\epsilon = 0.1$) generated adversarial examples.



(b) spectrum diff

(a) clean vs adversarial avg spectrum

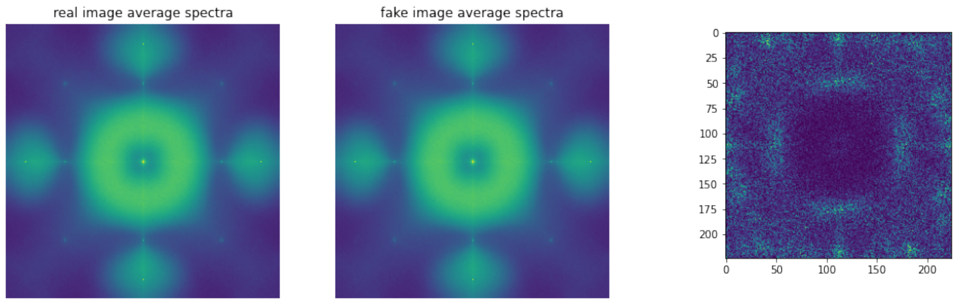
Figure 9: Spectra after applying high pass filter on C&WL2 ($\epsilon = 1$) generated adversarial examples.



(b) spectrum diff

(a) clean vs adversarial avg spectrum

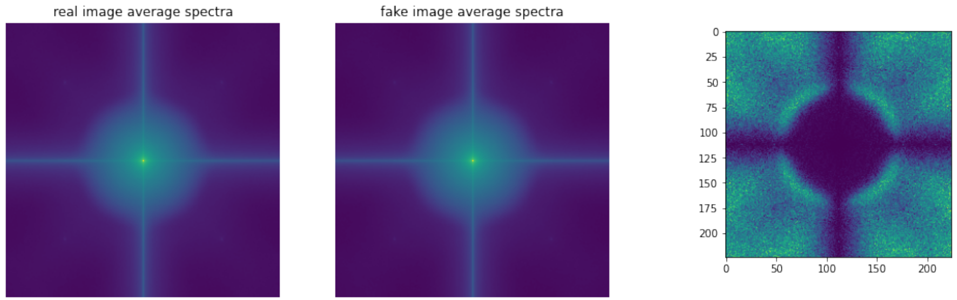
Figure 10: Spectra after applying high pass filter on C&WL2 ($\epsilon = 2$) generated adversarial examples.



(b) spectrum diff

(a) clean vs adversarial avg spectrum

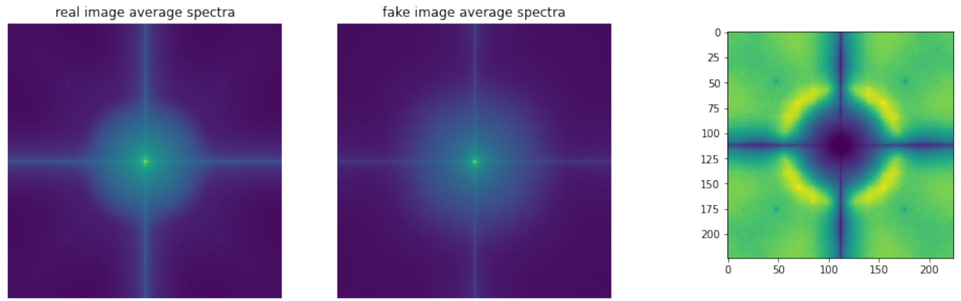
Figure 11: Spectra after applying high pass filter on C&WL2 ($\epsilon = 5$) generated adversarial examples.



(b) spectrum diff

(a) clean vs adversarial avg spectrum

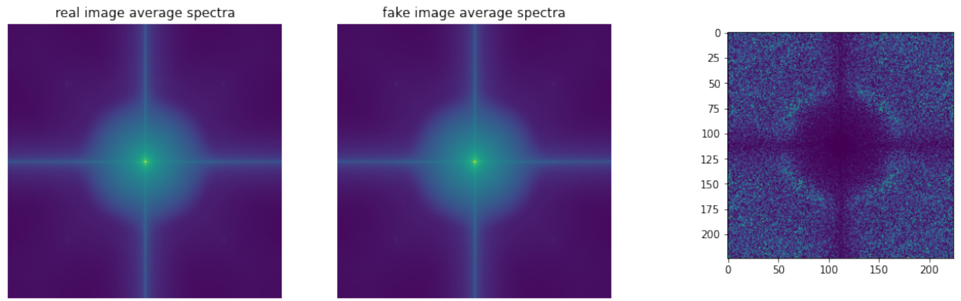
Figure 12: Spectra after applying low pass filter on PGDL2 ($\epsilon = 2$) generated adversarial examples.



(b) spectrum diff

(a) clean vs adversarial avg spectrum

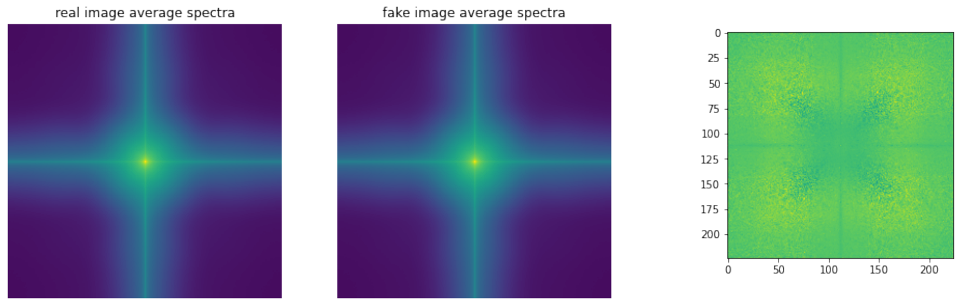
Figure 13: Spectra after applying low pass filter on FGSMLinf ($\epsilon = 0.1$) generated adversarial examples.



(b) spectrum diff

(a) clean vs adversarial avg spectrum

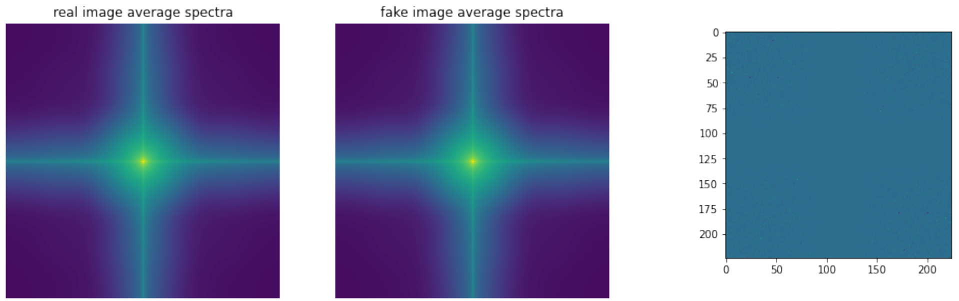
Figure 14: Spectra after applying low pass filter on C&WL2 ($\epsilon = 1$) generated adversarial examples.



(b) spectrum diff

(a) clean vs adversarial avg spectrum

Figure 15: Spectra after applying no filter on PGDL2 ($\epsilon = 0.5$) generated adversarial examples.



(a) clean vs adversarial avg spectrum

(b) spectrum diff

Figure 16: Spectra after applying no filter on C&WL2 ($\epsilon = 1$) generated adversarial examples.

8.2 Adversarial Example Detector Trained on ResNet50

Attack	ResNet50(FFT)	ResNet50(raw)
clean accuracy	1	1
L_2PGD-50 ($\epsilon = 0.1$)	1	0
L_2PGD-50 ($\epsilon = 0.3$)	1	0
L_2PGD-50 ($\epsilon = 0.5$)	1	0
L_2PGD-50 ($\epsilon = 2$)	0.472	0
L_∞FGSM ($\epsilon = 0.01$)	0	0
L_∞FGSM ($\epsilon = 0.1$)	0	0
L_2 C&W-40($\epsilon = 0.5$)	0.926	0
L_∞PGD-50($\epsilon = 0.01$)	0.001	0
L_∞PGD-50($\epsilon = 0.005$)	1	0

Table 8: ResNet50 trained detectors classification accuracy. Column 1 is detector trained and tested on RestNet50 with (DFT extracted frequency spectra), Column 2 is detector trained and tested on ResNet50 raw images. Both detectors clean images and adversarial images generated by L_2PGD-50 ($\epsilon = 0.5$)