Are Vision Transformer Models more Adversarially Robust than Convolutional Neural Networks?

Chenkai Su
University of Michigan
cksu@umich.edu

K. Faryab Haye
University of Michigan
kfaryabh@umich.edu

Zuoyi Li
University of Michigan
zuoyili@umich.edu

Abstract

Starting out primarily as a language model, transformer models have raised up as a state-of-the-art image classification method in recent years. Google's Vision Transformer Model has proven to be one of the best models on classic challenges such as CIFAR-10 and ImageNet (1). More and more researchers are applying transformer models in the computer vision realm (2). However, there hasn't been much study in adversarial robustness of transformer models in image classification. In our work we demonstrate that Transformer models, albeit having stronger representation capabilities, are just as easy to adversarially attack as traditional CNN models. Next we investigate visualizations of how the original vs attacked images look like in the new Vision Transformer models (ViT) and find that adversarial attacks can create saliency patterns that are not commonly seen in clean images for ViT. Our third experiment investigates the transferrability of attacks designed for Convolutional Neural Network (CNN) models to Transformer Based models, we find that these attacks have low transferability. Finally we discuss the future implications of our findings and the potential for creating ensemble or fingerprint based defences.

1 Introduction

The transformer model, which initially became the model of choice in Natural Language Procesing (NLP), is a type of Deep Neural Network (DNN) based on the self attention mechanism first introduced in A. Vaswani et. al. (3). Given its high performance and non-reliance on human-defined inductive biases, the transformer is receiving more and more attention from the computer vision community. This includes tasks such as object detection (4) (5), image segmentation (6), pose estimation (7), image classification (1) (8), and much more (2). In our work we specifically investigate the widely popular Vision Transformer Image Classification Model introduced by Dosovitskiy et. al. (1)

Given the rising interest in the new model architecture, it is natural to investigate security concerns which may be brought up by it. If a company were to switch over their Machine Learning models from a CNN based model to a Transformer one that could be a huge security risk. Suppose this was company which worked on self-driving cars, then it is possible that even if the old model would not have been fooled by a slightly perturbed image of a stop sign, the new model may interpret it as a 35 MPH sign instead, potentially leading to a serious accident. Outside of hypothetical, this is a very real possibility, and has been extensively demonstrated for CNN based models in prior work (9), (10), (11). Thus the general adoption of this new model in the community demands an investigation into its robustness.

We have also observed in prior work H. Zheng et. al. (12) that adversarial examples from CNN models tend to transfer well to one another, this makes it not hard to generalize defences from one model to another. The case of transformer models is more dangerous, however, because in our experiments we learn that adversarial examples from CNN models do not generalize to this new

Preprint. Under review.

model architecture. In addition to investigating the transferability of these examples, in our work we further explore the robustness of these new models.

To the best of our knowledge, our work is the first to investigate the adversarial robustness of transformer models, and compare their robustness to traditional CNN models. In summary, we make the following contributions:

- We are the first to reveal the low transferability between CNN generated adversarial attacks against Vision Transformer models and vice-versa. This is a serious security vulnerability since a defence against an attack generated against one model may not generalize to the other.
- We demonstrate that like CNN models, Vision Transformer models can just as easily be attacked through whitebox attacks such as PGD-k (13) under the L-∞ norm. However, unlike CNN models, Vision Transformer models are considerably robust against PGD-k attack in L-2 norm.
- We demonstrate that like CNN models, Vision Transformer models can also be adversarially trained to be more robust towards attack examples generated by the same architecture.
- In addition, our visualiztion results show that saliency map pattern can potentially be used in adversarial attack detection mechanism for ViT models.

In Section 2 of the paper we cover the datasets and models we use in our experiments and briefly go over the adversarial attacking techniques we investigate. Next, in Section 3 we cover each of our experiments and detail their results. In Section 4 we discuss the implications of our results and discuss possible future directions which may be explored from our findings. Finally in Section 5 we summarize our findings and conclude.

2 Background

2.1 Datasets

We conduct experiments on two main datasets:

- **CIFAR-10.** Following many prior works in the field (13), (12), (14) we conducted the majority of our experiments on the CIFAR-10 dataset. The CIFAR-10 dataset consists of 60000 32x32 colour images in 10 classes, with 6000 images per class. There are 50000 training images and 10000 test images. The test batch contains exactly 1000 randomly-selected images from each class. The training batches contain the remaining images in random order, between them, the training batches contain exactly 5000 images from each class.
- **Tiny ImageNet.** Since the majority of our models were pre-trained on ImageNet (15), we decided to use a smaller dataset with a subset of images from it for some of our experiments. Tiny Imagenet is a dataset of 120,000 labeled images belonging to 200 object categories. The categories are synsets of the WordNet hierarchy, and the images are similar in spirit to the ImageNet images used in the ILSVRC benchmark (16), but with lower resolution. Each of the 200 categories consists of 500 training images, 50 validation images, and 50 test images, all downsampled to a fixed resolution of 64x64.

For our visualization experiments in addition to using images from CIFAR-10 we also full 224x224 sized images from the original ImageNet dataset.

2.2 Transformer For Image Classification

Transformer models utilizes self-attention mechanism to capture positional relationship between different parts of the input (3). Dosovitskiy et. al. (1) segment input images into 16 by 16 patches, and use linear transformation to flatten the sequence of input image patches, in order to feed the sequence into a transformer. Vision Transformer (ViT) was shown to be able to capture the positional relationship between patches. After training with image labels, the ViT model is able to capture achieve state-of-the-art accuracy (1). The structue of ViT is shown in Figure 1. We believe transformer

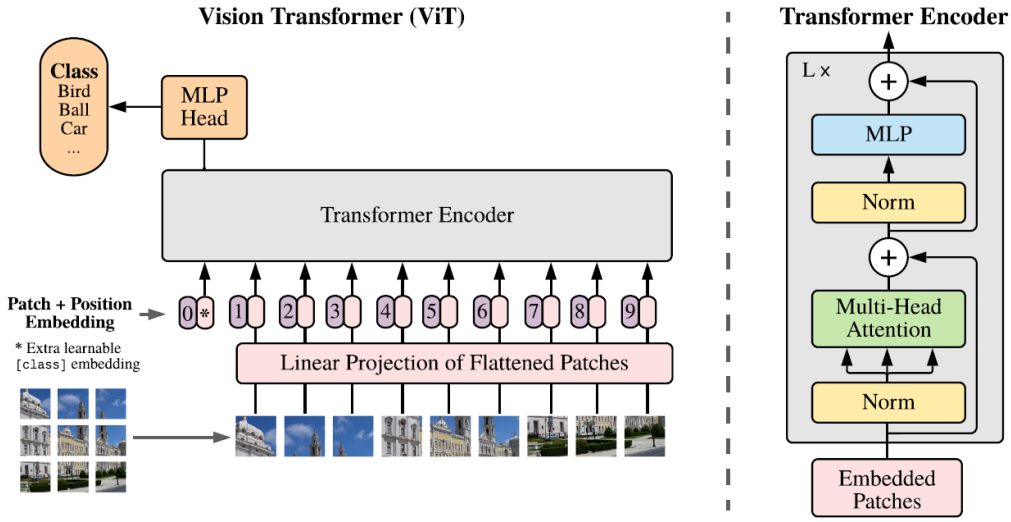

Figure 1: Model overview. ViT splits an image into fixed-size patches, linearly embed each of them, add position embeddings, and feed the resulting sequence of vectors to a standard Transformer encoder. Image labels are added during training. This figure is provided by Dosovitskiy et. al. (1).

based image classification models will become more prevalent in the field, and studying the adversarial characteristics of models such as ViT would be of great importance.

2.3 Common Adversarial Attacks

Gradient based attack methods are common, and generally applicable method of adversarial attacks against deep learning models (17). These attacks takes advantages of the fact that deep learning models use gradients to optimize themselves, and add small perturbations on the input image to create adversarial examples, such that the image looks about the same to human eyes, but can cause significant difference at the deep learning model's output. FGSM and PGD attacks are classic examples of this class of attacks (17). In response, adversarial training has been proposed as a general method to improve model's robustness against such attacks. By using adversarial examples during training, neural networks can learn to generalize to perturbed images. We explore how gradient based attacks work on transformer models, and how adversarial training can mitigate such attacks.

2.4 Transferability of Adversarial Attacks

Transferability of adversarial examples between deep learning models has been shown in Zheng et al.'s work. Adversarial examples generated based on one neural network can be effective on a different model, and this property has been utilized by Zheng et al (12) to propose a more efficient method for adversarial training. Previous works generate new adversarial examples from clean original images in each new epoch. Zheng et al.'s proposed Adversarial Training with Transferable Adversarial examples (ATTA), which succeeds perturbations from previous epochs (12). We want to explore how well adversarial examples generated from CNNs can be transferred to transformer models and vice-versa.

2.5 Prior Work on ViT

There are several papers focused on improving the accuracy and speed of ViT models in 2021. Zhou et. al. (18) proposed Re-Attention mechanism to solve the problem that attention maps will become similar when models goes deeper, which improve the model accuracy. Chen et. al.(19) used multi-scale feature to improve the accuracy. Zhang et. al.(20) proposed a new ViT architecture Multi-scale Vision Longformer which enhances the ViT for encoding high-resolution images. Graham et. al.(21)

Model Name (Input Dimension)	Model Size	Input Dimension	Original Accuracy
ViT fine tuned on cifar-10 (224)	~343 MB	3*224*224	98.52%
ViT fine tuned on cifar-10 (32)	~343 MB	3*32*32	98.74%
VGG16 fine tuned on cifar-10	~58 MB	3*32*32	92.77%

Figure 2: Baseline model summaries.

Attack \ Model	epsilon	alpha	ViT (224*224)	ViT (32*32)	VGG16 (32*32)
None	None	None	98.52 %	98.74%	92.52 %
PGD-1	8/255	2/255	44.84 %	48.26%	40.19 %
PGD-5			1.78 %	2.07%	6.93 %
PGD-10			0.98 %	0.86%	4.7 %
PGD-20			0.96 %	0.77%	4.46 %
PGD-20 L2	8/255	2/255	98.45 %	96.69 %	32.65 %
PGD-20 L2	25/255	5/255	95.94 %	78.15 %	29.51 %
PGD-20 L2	50/255	12/255	86.71 %	37.88 %	23.34 %
PGD-20 L2	100/255	20/255	53.37 %	9.32 %	18.16 %

Figure 3: White-box Attack results on CIFAR-10. The green rows are attacked under the L-∞ norm. The percentages are raw accuracy values on the original test set and adversarial examples generated by the attacks in column 1.

optimize the trade-off between accuracy and efficiency in a high-speed regime. However, to our knowledge there is no work related to the robustness of the Vision Transformer (ViT) Model.

3 Experiment Results

3.1 Baseline Models

As shown in Figure 2, we use three different models in our experiments as baseline. We use ViT models with different input size to observe its effect on adversarial attacks generated on different sizes of images, and on the saliency maps. We retrieved the fine tuned ViT model on CIFAR-10 from works of Nathan Raw (22), and the fine tuned VGG16 model on CIFAR-10 from Yaofo Chen (23). We will refer to ViT model with input dimension 224 as ViT224, and the one with input dimension 32 as ViT32.

Attack \ Model	ViT(32*32) (Acc)	Attack \ Model	VGG16(32*32) (Acc)
None	98.74 %	None	92.77 %
PGD-1-CNN	79.74 %	PGD-1-ViT	56.65 %
PGD-5-CNN	85.65 %	PGD-5-ViT	60.74 %
PGD-10-CNN	83.50 %	PGD-10-ViT	60.89 %
PGD-20-CNN	80.97 %	PGD-20-ViT	63.78 %

Figure 4: Transferrability of Attack Examples. The ViT (224*224) column tabulates the accuracy of the fine tuned ViT (224*224) on attack examples generated against the VGG-16 Model. Similarly the VGG16 (32*32) column tabulates the accuracy of the fine tuned VGG16 (32*32) on attack examples generated against the ViT (224*224) Model.

3.2 Whitebox Attacks

We conduct Projected Gradient Descent Attack (PGD) (17) attack against ViT models and CNN model. We evaluate both of the attack results on the CIFAR-10 dataset. We firstly apply L-∞ norm to the attack. Then we also try L-2 norm to the attack. The maximum perturbation ϵ and the step size (alpha) during attack, along with the model's accuracy under attack, are shown in Figure 3.

3.3 Visualizations

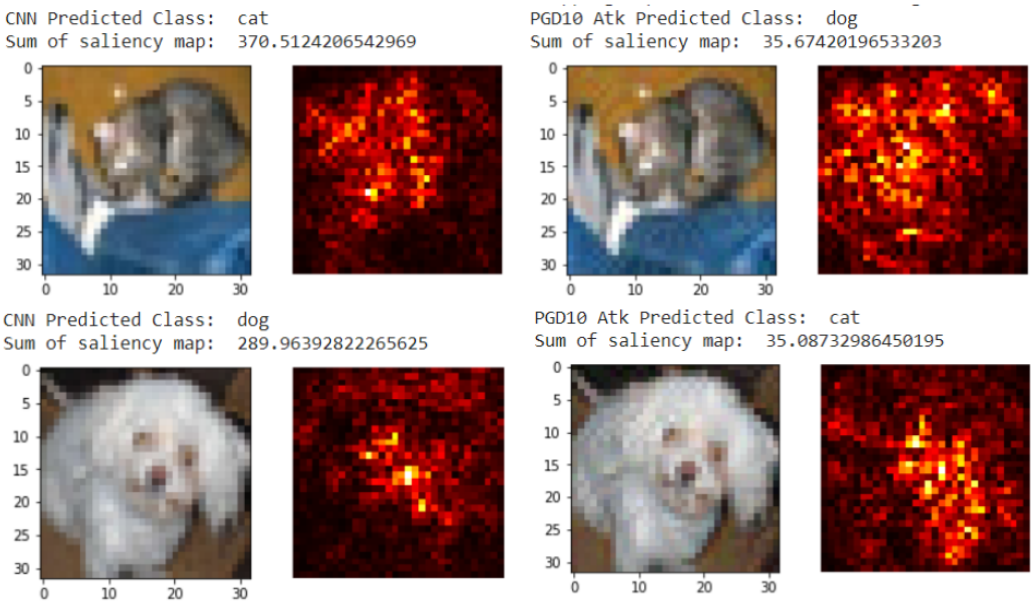

Figure 5: Saliency map for VGG16 on CIFAR-10 clean images (left), and PGD10 L-∞ attacked images (right).

In our first visualization experiment we created saliency maps. Saliency maps were introduced by Simonyan et al (24) to visualize how each pixel in the input image contributes to the final prediction (24). Intuitively, saliency maps highlight the pixels that need to be changed the least to affect the overrall class score of the model. One can expect that such pixels correspond to the object's location in the image. In Figure 6 and Figure 7, we observe the saliency map for ViT224 to get finer grain details compared with using 32*32 images. For comparison, we show in Figure 5 the saliency map for the same images in 32*32 from a CNN model.

Next we created attention maps. Various mechanisms of measuring visual attention have been introduced in the past (25). In our experiments we use the visual attention mechanism used in (1), taking inspiration form the implementation in (1). Attention maps measure the average sum of all weights across all attention heads. They are more 'general' than saliency maps in that sense. Our results are shown in figures 8, 9, 10, 11, 9, and discussed in section 4.

3.4 Transfer Attacks Between ViT and CNN

To study the transferability of attack between the models we conduct several experiments. Our first experiment is illustrated in Figure 4. We generate attack examples of varying degrees of strength against both ViT and CNN based models. It appears that the ViT32 model is comparably robust towards CNN generated adversarial examples than the VGG16 model is towards ViT generated adversarial examples.

We expected that adversarial examples would not transfer as well due to the completely different underlying architectures of the two models, but this asymmetrical transferability was unexpected. Implications of this are discussed in section 4.

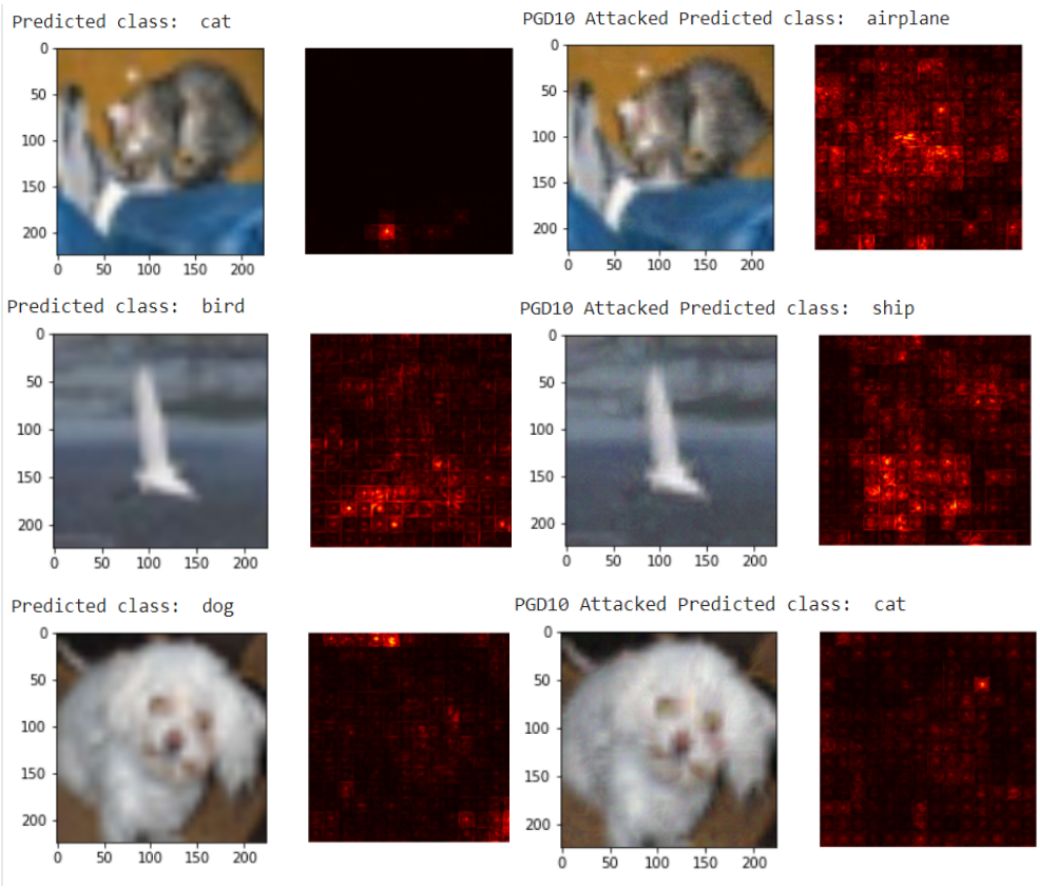

Figure 6: Saliency map for ViT224 on CIFAR-10 clean images (left), and PGD-10 L-∞ attacked images (right).

We also conducted another experiment on TinyImagenet to see whether the same results would generalize. However, we have we ran out of time to finish fine-tuning a ViT model for TinyImagenet and were only able to go halfway through that experiment.

3.5 Adversarial Training

In Figure 13, we show our preliminary experiment results of adversarially training ViT model with ATTA methodology. We use CIFAR-10 as the dataset, and ViT model with 3x32x32 input size. We used 8/255 as ϵ value, and 2/255 for step size for PGD attacks. The PGD attacks use L-∞ norm. We evaluate our models performance on 10 thousand clean CIFAR-10 test images, and adversarial images generated by PGD-10 and PGD-20 from the 10 thousand test images, based on the original model.

In our procedure to get the result shown in Figure 13, all adversarial training data are generated by PGD-5, with the same epsilon and step size mentioned above. At the beginning, we generate 50 thousand adversarial training images based on the clean training set of CIFAR-10 for the first epoch, and generate adversarial examples based on training images from the previous epoch starting from epoch 2. We didn't incorporate random restart mechanism, or mix in clean training image into adversarial training set as mentioned in the original paper (12).

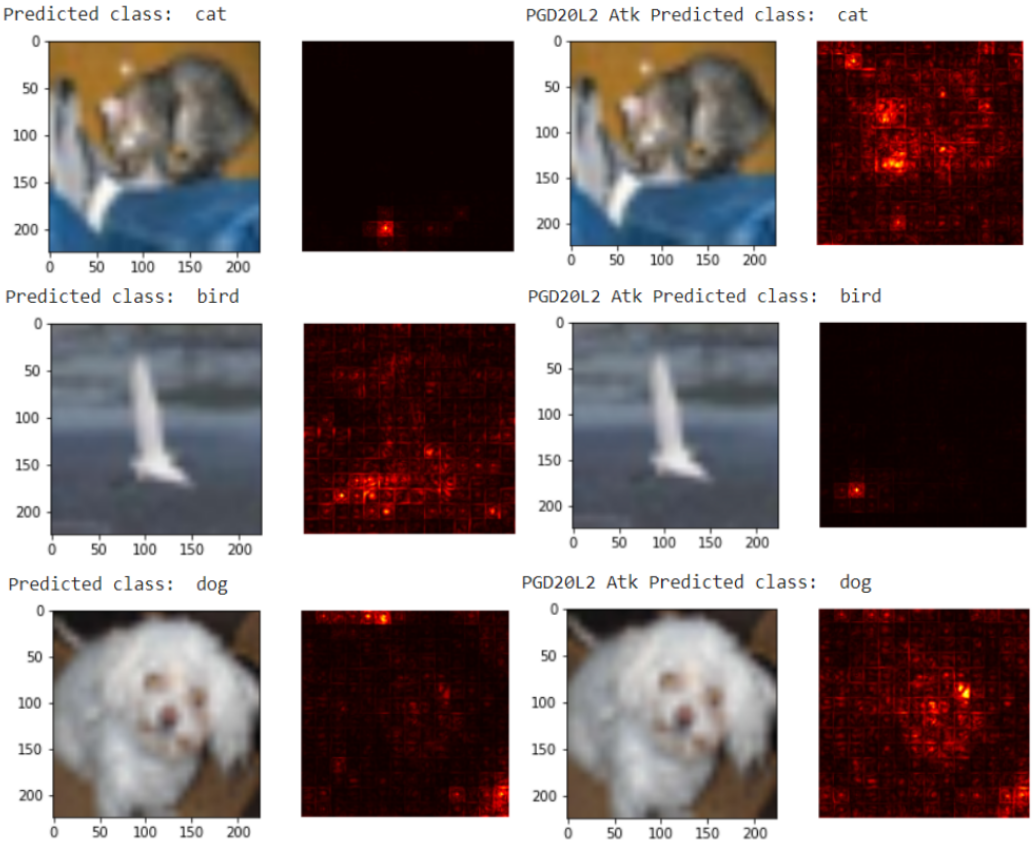

Figure 7: Saliency map for ViT224 on CIFAR-10 clean images (left), and PGD-10 L-2 attacked images (right).

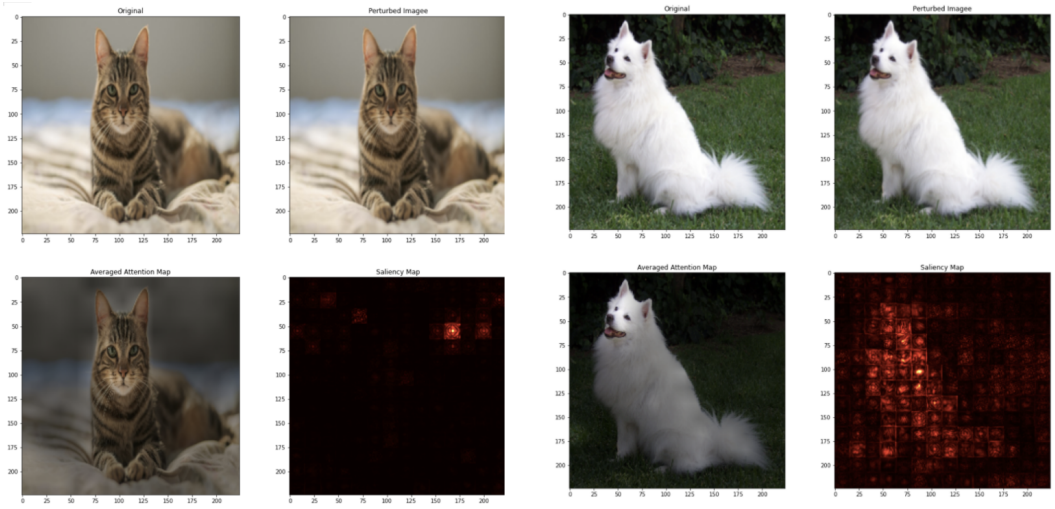

Figure 8: Cat and Dog Originals

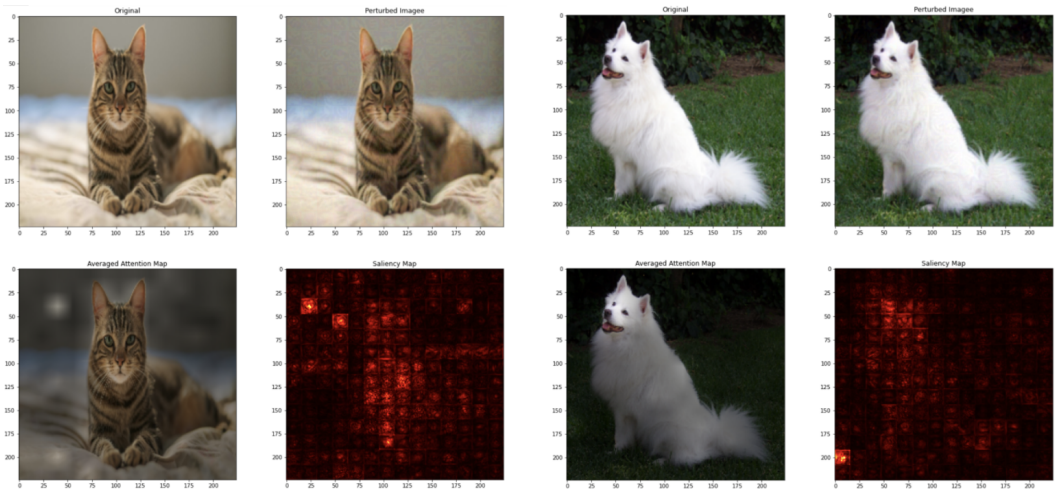

Figure 9: Untargeted, PGD-20, L-Infty, $\epsilon = 8/255$

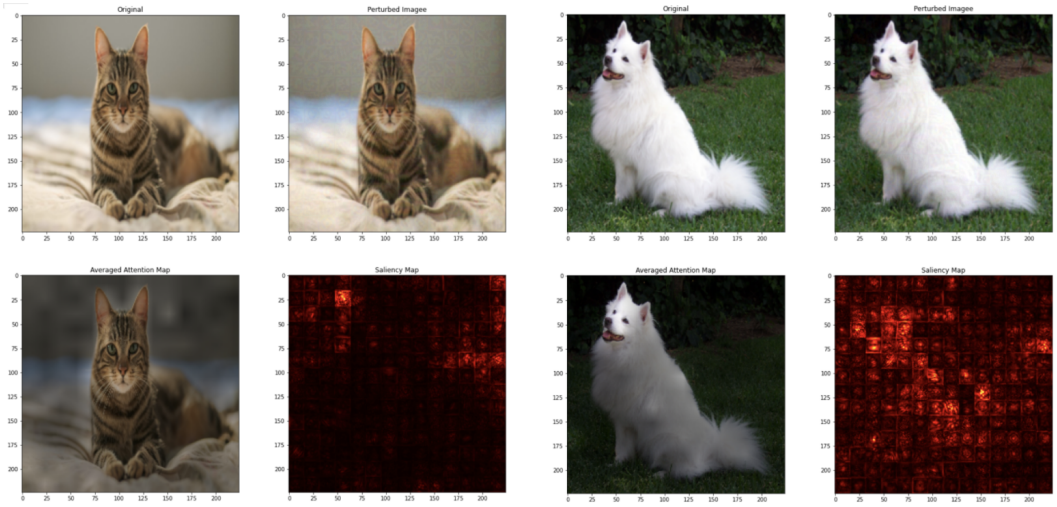

Figure 10: Targeted, PGD-20, L-Infty $\epsilon = 8/255$

4 Discussion and Future Work

4.1 White Box Attacks

The results shows in figure 3 that PGD attack works well on both of the CNN models and ViT models under the L-∞ norm. However, under the L-2 norm, when ϵ value is relatively small, ViT model is much more robust against PGD attack compared with the CNN model. Here we mainly focus on comparing ViT32 model with VGG16 model, since L-2 attack's strength is correlated with input image size given the same set of ϵ and α values. When perturbation becomes larger, with $\epsilon = 100/255$, the ViT32 model's performance degrades quickly, and shows lower accuracy compared with the CNN model. It is also worth noting that the adversarial examples generated with $\epsilon = 100/255$ remains visually similar to clean images. From the comparison between CNN and ViT facing L-2 norm PGD-20 attack, we can conclude that ViT models are significantly more robust than the CNN models when ϵ is small, but with high ϵ value, performance of ViT model is similar to that of a CNN model under L-2 PGD attacks.

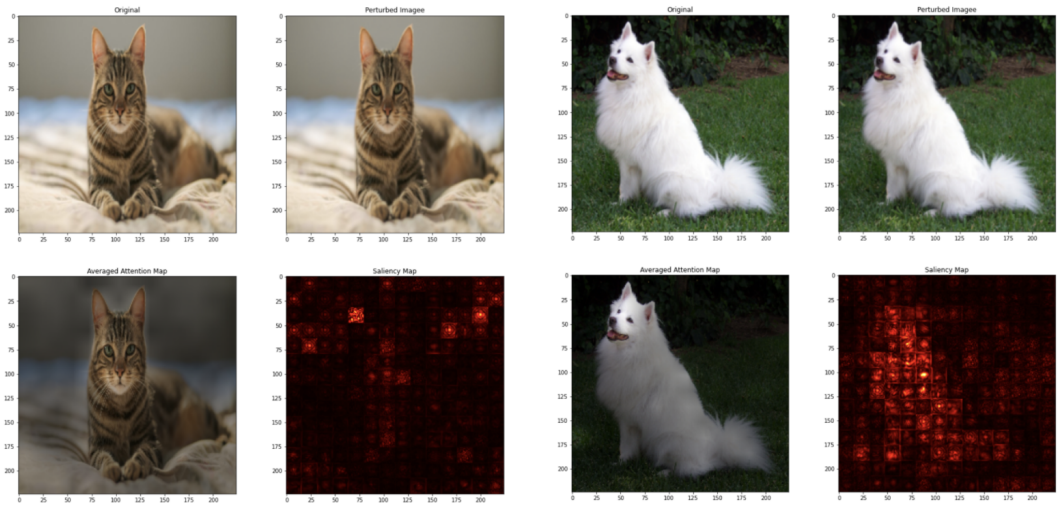

Figure 11: Untargeted, PGD-20, L-2, $\epsilon = 8/255$

Figure 12: Untargeted PGD-20 L-2 $\epsilon = 100/255$

Epoch	Adv Ex Gen Time (sec)	Adv Train Time (sec)	Adv Train Acc (%)	Eval Acc On PGD10 (%)	Eval Acc On PGD20 (%)	Eval Acc on Orig (%)
Orig	NA	NA	NA	0.98	0.96	98.74
1	2458.99	663.34	93.79	91.29	91.07	93.21
2	2457.49	663.52	95.70	81.74	82.49	89.22
3	2453.59	660.72	94.00	82.36	82.49	84.82
4	2446.89	660.52	95.18	69.49	69.91	72.50
5	2446.30	660.50	96.22	76.38	76.52	78.20

Figure 13: Model accuracy before and after ATTA training.

4.2 Visualizations

We can observe from Figure 5, 6, and 7, there is significant difference between the patterns of saliency map for CNN, and for ViT. CNN's saliency maps have a common pattern that generally highlights

the shape of the object in the image. However, the saliency pattern for ViT can vary a lot between images. Some images have very small, concentrated pattern, while some can have a pattern that resembles CNN more. It is unclear why sometimes the saliency maps are concentrated to specific image patches.

Another observation that is immediately different from the CNN maps is that the ViT maps have distinct checker-board Pattern. This is clearly due to the way the input images are processed; each image being cropped multiple into 16x16 patch embeddings that is.

In order to further investigate the peculiar saliency patterns we conducted more experiments on higher resolution ImageNet images. In addition to the saliency maps we also extracted attention maps. Two examples of a cat and dog are highlighted in figures 8.

The dog lights up very similarly to how we expect. It appears that the single patch lighting up is not uncommon as we observe the same behavior for the original cat image. What is surprising is that the attention map of the cat is still focused on the cat, and not at all around the highlighted image patch.

While investigating the whitebox attacks our visualization results agreed with our initial findings in section 3.2. As can be seen in figure 9 it is pretty easy to divert the models attention and successfully conduct an untargeted attack on the model. The L-∞ attacks add some noise to the image to the human eye, but to the model result in a completely different image.

In contrast, targeted L-∞ and untargeted L2-Attacks were not so easy. When L-∞ attacked to predict 'coffee-mug' on both images, the model succeeded for the brown 'tabby_cat' but not for the 'samoyed' (see figure 10). The saliency map for the dog changes a lot on thee perturbed image but the attention map stays focused on its face.

For the PGD-20 L2-norm untargeted attack, with an $\epsilon = 8/255$, both attacks failed (See figure 11). Only after bumping up epsilon to $\epsilon = 100/255$ to both attacks pass (See 12). Since the ViT has global information as soon as the initial hidden layers we hypothesize it is hard to conduct a localized L2 attack against it. Only after changing to a higherr ϵ does the model start failing. This is consistent with our findings in figure 3.

It is still unclear why sometimes the saliency maps are concentrated to a single patch or two. Our hypothesis is that because transformer models can learn the relationship from one location to all the other locations in the input sequence. There could be a patch in the input sequence that concentrated scores for important features from the overall image, and therefore leads to high heat value in a small region.

In the future, we want to explore the possibility to using saliency map property as a weighted part of an adversarial attack detection mechanism for ViT. The saliency maps produced by ViT models have a distinct checkerboard pattern which may be leveraged to create a detection mechanism.

4.3 Transfer Attacks Between ViT and CNN

There is low transferability of ViT generated attack examples to CNN based models and even lower transferability of CNN generated attack examples towards ViT based models.

Though it is unclear to us exactly why this may be the case, we have one hypothesis. Since transformers lack the inductive biases of CNNs, such as translation invariance and a locally restricted receptive field, they are inherently more robust towards adversarial examples generated in such a setting. Of course the reverse does not hold true for the CNN models which are attacked by ViT generated images.

It is interesting that the PGD-1 ViT attack results in the lowest VGG16 accuracy (56.65%) compared to the more powerful attacks. This might make sense because the higher PGD-k examples would be more 'overfit' against the model they are targeting instead of the ViT model. Surprisingly no similar trend appears in the PGD-k CNN attacks, all of which seem to decrease the model accuracy by similar amounts.

In the future, we plan to conduct the same experiment on TinyImagenet and some other datasets to see if the results generalize. We also plan on comparing ViT to other CNN architectures to see whether our findings are stable. Another interesting future experiment might be construct an ensemble adversarial attacks for both of these two models to explore more inner relationship between them.

4.4 Adversarial Training As A Defense Strategy

In experiments shown in Figure 13, we didn't incorporate some training strategies proposed in the original paper, such as random restart (12). But from the results shown so far, we can see that after first epoch of adversarial training, the model is able to improve its accuracy on PGD-10 attacked test set from 0.98% to 91.29%, and its accuracy on PGD-20 attacked test set from 0.96% to 91.07%, while keeping its accuracy on original task at 93.21%. We can qualitatively state that adversarial training remains a valid option to significantly improve ViT model's robustness against evasion attacks.

5 Conclusion

In conclusion, our work is the first to investigate the adversarial robustness of transformer models, and we make the following contributions:

- We are the first to reveal the low transferability between CNN generated adversarial attacks against Vision Transformer models and vice-versa. This is a serious security vulnerability since a defence against an attack generated against one model may not generalize to the other.
- We demonstrate that like CNN models, Vision Transformer models can just as easily be attacked through whitebox attacks such as FGSM and PGD-k (13). But Vision Transformer models are very robust against PGD-k attack in L-2.
- We demonstrate that like CNN models, Vision Transformer models can also be adversarially trained to be more robust towards attack examples generated by the same architecture.
- In addition, our visualiztion results show that saliency map pattern can potentially be used in adversarial attack detection mechanism for ViT models..

6 Acknowledgements

We would like to thank Professor Atul Prakash and Tianji Cong for their continued support throughout our research. In addition we would like to acknowledge the open source research community which enabled us to quickly deploy pre-trained models and play around with them without having to worry about setting up infrastructure and training gigantic models from scratch.

References

[1] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," 2020. [Online]. Available: https://shairozsohail.medium.com/a-survey-of-visual-attention-mechanisms-in-deep-learning-1043eb25f343

[2] K. Han, Y. Wang, H. Chen, X. Chen, J. Guo, Z. Liu, Y. Tang, A. Xiao, C. Xu, Y. Xu, Z. Yang, Y. Zhang, and D. Tao, "A survey on visual transformer," *ArXiv*, vol. abs/2012.12556, 2020.

[3] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *arXiv preprint arXiv:1706.03762*, 2017.

[4] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable detr: Deformable transformers for end-to-end object detection," 2021.

[5] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," 2020.

[6] S. Zheng, J. Lu, H. Zhao, X. Zhu, Z. Luo, Y. Wang, Y. Fu, J. Feng, T. Xiang, P. H. S. Torr, and L. Zhang, "Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers," 2021.

[7] L. Huang, J. Tan, J. Liu, and J. Yuan, "Hand-transformer: Non-autoregressive structured modeling for 3d hand pose estimation," in *ECCV*, 2020.

[8] M. Chen, A. Radford, J. Wu, H. Jun, P. Dhariwal, D. Luan, and I. Sutskever, "Generative pretraining from pixels," in *ICML*, 2020.

[9] A. Kurakin, I. Goodfellow, and S. Bengio, "Adversarial examples in the physical world," 2017.

[10] K. Eykholt, I. Evtimov, E. Fernandes, B. Li, A. Rahmati, C. Xiao, A. Prakash, T. Kohno, and D. Song, "Robust physical-world attacks on deep learning models," 2018.

[11] R. Feng, J. Chen, E. Fernandes, S. Jha, and A. Prakash, "Robust physical hard-label attacks on deep learning visual classification," 2020.

[12] H. Zheng, Z. Zhang, J. Gu, H. Lee, and A. Prakash, "Efficient adversarial training with transferable adversarial examples," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 1181–1190.

[13] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," *arXiv preprint arXiv:1706.06083*, 2017.

[14] P.-Y. Chen, H. Zhang, Y. Sharma, J. Yi, and C.-J. Hsieh, "Zoo," *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*, Nov 2017. [Online]. Available: http://dx.doi.org/10.1145/3128572.3140448

[15] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248–255.

[16] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "Imagenet large scale visual recognition challenge," 2015.

[17] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," 2019.

[18] D. Zhou, B. Kang, X. Jin, L. Yang, X. Lian, Q. Hou, and J. Feng, "Deepvit: Towards deeper vision transformer," *arXiv preprint arXiv:2103.11886*, 2021.

[19] C.-F. Chen, Q. Fan, and R. Panda, "Crossvit: Cross-attention multi-scale vision transformer for image classification," *arXiv preprint arXiv:2103.14899*, 2021.

[20] P. Zhang, X. Dai, J. Yang, B. Xiao, L. Yuan, L. Zhang, and J. Gao, "Multi-scale vision longformer: A new vision transformer for high-resolution image encoding," *arXiv preprint arXiv:2103.15358*, 2021.

[21] B. Graham, A. El-Nouby, H. Touvron, P. Stock, A. Joulin, H. Jégou, and M. Douze, "Levit: a vision transformer in convnet's clothing for faster inference," *arXiv preprint arXiv:2104.01136*, 2021.

[22] N. Raw, "huggingface-vit-finetune," 2021. [Online]. Available: https://github.com/nateraw/huggingface-vit-finetune

[23] Y. Chen, "pytorch-cifar-models," 2021. [Online]. Available: https://github.com/chenyaofo/pytorch-cifar-models

[24] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," 2014.

[25] S. Sohail, "A survey of visual attention mechanisms in deep learning," Dec 2019. [Online]. Available: https://shairozsohail.medium.com/a-survey-of-visual-attention-mechanisms-in-deep-learning-1043eb25f343