On the Adversarial Robustness of Vision Transformers

Rulin Shao
Carnegie Mellon University
rulins@cs.cmu.edu

Zhouxing Shi
University of California, Los Angeles
zshi@cs.ucla.edu

Jinfeng Yi
JD AI Research
yijinfeng@jd.com

Pin-Yu Chen
IBM Research
pin-yu.chen@ibm.com

Cho-Jui Hsieh
University of California, Los Angeles
chohsieh@cs.ucla.edu

Abstract

Following the success in advancing natural language processing and understanding, transformers are expected to bring revolutionary changes to computer vision. This work provides a comprehensive study on both empirical and certified robustness of vision transformers (ViTs), with analysis that casts light on creating models that resist adversarial attacks. We find that ViTs possess better empirical and certified adversarial robustness when compared with various baselines. In our frequency study, we show features learned by ViTs contain less high-frequency patterns which tend to have spurious correlation, and there is a high correlation between how much the model learns high-frequency features and its robustness against different frequency-based perturbations. Moreover, modern CNN designs that borrow techniques from ViTs including activation function, layer norm, larger kernel size to imitate the global attention, and patchify the images as inputs, etc., could help bridge the performance gap between ViTs and CNNs not only in terms of performance, but also certified and empirical adversarial robustness. Introducing convolutional or tokens-to-token blocks for learning high-frequency features in ViTs can improve classification accuracy but at the cost of adversarial robustness.

1 Introduction

Transformers are originally applied in natural language processing (NLP) tasks as a type of deep neural network (DNN) mainly based on the self-attention mechanism (21; 8; 2). (9) applied a pure transformer directly to sequences of image patches (i.e., a vision transformer, ViT) and showed that the Transformer itself can be competitive with convolutional neural networks (CNN) on image classification tasks. Since then transformers have been extended to various vision tasks and show competitive or even better performance compared to CNNs and recurrent neural networks (RNNs) (3; 4; 26). While ViT and its variants hold promise toward a unified machine learning paradigm and architecture applicable to different data modalities, it is critical to study the robustness of ViT against adversarial perturbations for safe and reliable deployment of many real-world applications. In this work, we examine the adversarial robustness of ViTs on image classification tasks and make comparisons with CNN and MLP baselines. As highlighted in Figure 1, our experimental results illustrate the superior robustness of ViTs than CNNs and MLP-Mixer in various settings, and we show more results in the following section. Our work provides a deep understanding of the intrinsic robustness of ViTs and can be used to inform the design of robust vision models based on the transformer structure.

ML Safety Workshop, 36th Conference on Neural Information Processing Systems (NeurIPS 2022).

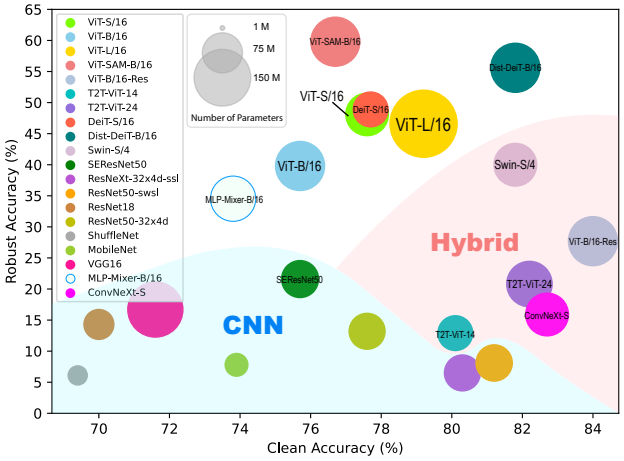

Figure 1: Robust accuracy v.s. clean accuracy. The robust accuracy is evaluated by AutoAttack (6).

2 Empirical Study and Reasoning on the Source of Robustness

Unless otherwise specified, Clean Accuracy (CA) is the accuracy evaluated on the entire ImageNet-1k (7) test set, Robust Accuracy (RA) is the accuracy on the adversarial examples generated with 1,000 test samples.

2.1 Frequency Study Using Filtered-PGD

Table 1: RA (%) of the target models against frequency-filtered PGD attack in our frequency study. In the "Low-pass" column, only low-frequency adversarial perturbations are preserved and added to the input images. In the "High-pass" column, only high-frequency perturbations are preserved. The "Full-pass" mode preserves all frequency and is the same as the traditional PGD attack. We set the attack step fixed to 40 and vary the attack radius to different values.

	Low-pass					High-pass					Full-pass				
Attack Radius	**0.001**	**0.003**	**0.005**	**0.01**	**0.1**	**0.001**	**0.003**	**0.005**	**0.01**	**0.1**	**0.001**	**0.003**	**0.005**	**0.01**	**0.1**
ViT-S/16	74.0	68.1	64.7	59.8	56.2	70.8	60.7	50.6	40.4	23.4	55.4	24.6	10.2	1.0	0.0
ViT-B/16	71.9	64.3	60.3	55.8	49.6	66.3	53.1	44.0	33.4	21.9	48.9	14.6	6.0	0.9	0.0
ViT-L/16	74.9	64.1	58.3	50.2	42.0	72.9	62.3	56.6	47.5	28.9	55.1	23.4	9.9	1.8	0.0
ViT-B/16-Res	83.1	81.4	80.4	79.0	75.1	62.9	29.2	16.0	7.3	3.3	45.5	8.4	2.3	0.1	0.0
T2T-ViT-14	78.0	77.2	76.0	75.8	74.3	49.6	20.5	9.1	3.1	1.4	37.1	7.0	1.8	0.0	0.0
T2T-ViT-24	80.2	79.2	78.4	77.7	74.4	58.3	31.1	17.7	8.2	3.1	47.7	12.3	3.4	0.2	0.0
MLP-Mixer-B/16	69.4	64.7	62.4	60.0	59.8	56.6	32.5	19.5	6.7	1.3	34.5	3.8	0.0	0.0	0.0
ConvNeXt-S	80.3	79.0	77.9	76.4	74.3	53.1	24.4	14.4	6.8	6.2	15.9	0.0	0.0	0.0	0.0
ResNet50-swsl	78.2	74.9	73.7	71.6	72.5	45.3	12.4	5.0	2.2	3.5	24.7	2.9	1.4	0.4	0.0
ResNet50-32x4d	75.0	66.3	62.7	59.0	61.5	47.7	17.1	7.4	3.3	3.5	28.2	3.2	1.2	0.4	0.1

ViTs are more resistant to high-frequency perturbations and have less bias towards high-frequency features that have spurious correlation with labels. We design a frequency study to verify our hypothesis that ViTs are adversarially more robust compared with CNNs and MLP-Mixer because ViTs learn less high-frequency features. For adversarial perturbations generated by PGD attack, we first project them to the frequency domain by DCT. We design three frequency filters where each filter allows only the corresponding frequencies to pass through as shown in Table 1. We then apply these filters to the frequencies of the perturbations, and project them back to the spacial domain with the IDCT. We test the RA under different frequency areas, and show the results in Table 1. The RA of ViTs are much higher in the "High-pass" column when only the high-frequencies of the perturbations are preserved. In contrast, CNNs show significantly lower RA in the "High-pass" column than in the "Low-pass" column. It reflects that CNNs tend to be more sensitive to high-frequency adversarial perturbations compared to ViTs.

Introducing CNN or T2T blocks makes ViTs less robust to high-frequency perturbations. One interesting and perhaps surprising finding is that ViTs have worse robustness when modules that

Table 2: Clean Accuracy (%) and Robust Accuracy (%) of target models against 40-step PGD attack with different radii. More results can be found in Appendix B.

	CA	RA against PGD			
Attack Radius		0.001	0.003	0.005	0.01
ViT-S/16	77.6	**55.4**	**24.6**	**10.2**	1.0
ViT-B/16	75.7	48.9	14.6	6.0	0.9
ViT-L/16	79.2	55.1	23.4	9.9	**1.8**
ViT-SAM-B/16	76.7	**63.4**	**37.0**	**20.1**	**3.8**
ViT-B/16-Res	84.0	45.5	8.4	2.3	0.1
T2T-ViT-14	80.1	37.1	7.0	1.8	0.0
T2T-ViT-24	82.2	47.7	12.3	3.4	0.2
Deit-S/16	77.7	48.9	17.6	7.1	1.1
Dist-Deit-B/16	81.8	55.6	17.7	4.5	0.4
Swin-S/4	81.8	40.0	12.4	3.2	0.2
MLP-Mixer-B/16	73.8	41.9	10.7	4.3	0.4
ConvNeXt-S	82.7	42.4	8.1	2.6	0.0
SEResNet50	75.7	35.4	4.9	0.8	0.1
ResNeXt-32x4d-ssl	80.3	23.0	2.9	1.2	0.5
ResNet50-swsl	81.2	24.7	2.9	1.4	0.4
ResNet18	70.0	24.9	2.0	0.6	0.1
ResNet50-32x4d	77.6	28.2	3.2	1.2	0.4
ShuffleNet	69.4	15.0	0.6	0.2	0.0
MobileNet	71.9	16.7	0.4	0.0	0.0
VGG16	71.6	26.3	3.2	1.3	0.0

Table 3: Clean Accuracy (%) and Robust Accuracy (%) of target models against AutoAttack with different attack radii. More results can be found in Appendix B.

	CA	RA against AutoAttack			
Attack Radius		0.001	0.003	0.005	0.01
ViT-S/16	77.6	**48.1**	6.0	0.5	0.0
ViT-B/16	75.7	39.8	5.4	0.6	0.0
ViT-L/16	79.2	46.6	**8.5**	**1.0**	0.0
ViT-SAM-B/16	76.7	**59.8**	**26.0**	**8.4**	**0.1**
ViT-B/16-Res	84.0	27.7	0.9	0.0	0.0
T2T-ViT-14	80.1	12.9	0.1	0.0	0.0
T2T-ViT-24	82.2	20.8	0.3	0.0	0.0
Dist-Deit-S/16	79.3	43.1	3.7	0.2	0.0
Dist-Deit-B/16	81.8	42.7	3.4	0.2	0.0
Swin-S/4	81.8	7.9	0.1	0.0	0.0
MLP-Mixer-B/16	73.8	34.5	3.8	0.0	0.0
ConvNeXt-S	82.7	15.9	0.0	0.0	0.0
SEResNet50	75.7	21.6	0.6	0.0	0.0
ResNeXt-32x4d-ssl	80.3	6.5	0.0	0.0	0.0
ResNet50-swsl	81.2	8.1	0.0	0.0	0.0
ResNet18	70.0	14.3	0.4	0.0	0.0
ResNet50-32x4d	77.6	13.2	0.2	0.0	0.0
ShuffleNet	69.4	6.1	0.0	0.0	0.0
MobileNet	71.9	7.8	0.0	0.0	0.0
VGG16	71.6	16.7	0.5	0.0	0.0

claimed to help learning local structures are added ahead of the transformer blocks. We observe that ResNet and T2T modules that could help improve the CA of the hybrid ViTs makes the models more sensitive to high-frequency perturbations. One possible explanation is that the introduced modules improve the classification accuracy by remembering the high-frequency patterns that repeatedly appear in the training dataset. These structures such as edges and lines are high-frequency and sensitive to perturbations (22). Learning such features makes the model more vulnerable to adversarial attacks.

2.2 Empirical Study: Adversarial Robustness under Various Adversarial Attacks

Robustness under PGD and AutoAttack We present the results of PGD and AutoAttack in Table 2 and Table 3 respectively. The RA is approximately 0.0% on all the models when $\epsilon = 0.01$ is large,indicating models trained without any adversarial augmentations are vulnerable to large perturbations. However, such vulnerability doesn't hold equally for all models within mild perturbations: For smaller attack radii, **ViT models have higher RA than CNNs under both PGD attack and AutoAttack.** We also find that **introducing ResNet or T2T blocks decreases the RA under both PGD and AutoAttack.** When the features learned by ResNet are introduced, the RA of ViT-B/16 decreases from 48.9% to 45.5% of ViT-B/16-Res under PGD attack, and from 39.8% to 27.7% under AutoAttack, with attack radius $\epsilon = 0.001$. Besides, the results show that **SAM can further improve model's adversarial robustness**. This is because the SAM objective is formulated as a min-max optimization similar to adversarial training: but instead of adding adversarial perturbations to the input space, SAM adds adversarial perturbations to the weights.

Transferability of Adversarial Examples from ViTs to CNNs and Vice Versa We consider attacks with ℓ_∞-norm perturbation no larger than 0.1 and present the results in Figure 2. When the ViTs serve as the target models and CNNs serve as the source models, as shown in the lower left of each subplot, the RER of the transfer attack is quite low. On the other hand, when the ViTs are the source models, the adversarial examples they generate have higher RER when transferred to other target models. As a result, the first three rows and the last seven columns are darker than the others. Besides, for the diagonal lines in the figure where FGSM actually attacks the models in a white-box setting, we can observe that ViTs are less sensitive to attack with smaller radii compared to CNNs, and T2T modules make ViTs more robust to such one-step attack. In addition, adversarial examples transfer well between models with similar structures.

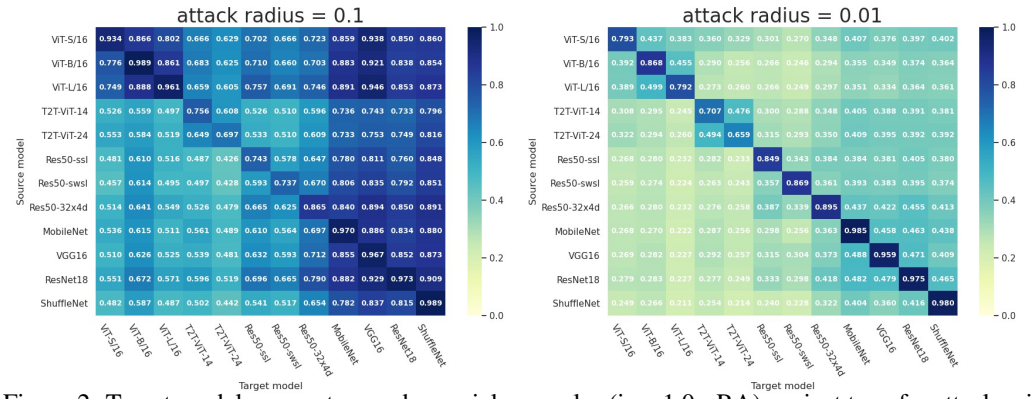

Figure 2: Target model error rate on adversarial examples (i.e., 1.0 - RA) against transfer attack using FGSM. The rows stand for the surrogate models and the columns stand for the target models.

Table 4: Certified robust accuracy w.r.t. different radii using denoised randomized smoothing (19) ($\sigma = 0.25$). Pure ViTs are highlighted with gray shadows.

Radius	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
ViT-S/16	0.5944	0.5452	0.4936	0.4424	0.3972	0.3428	0.2820	0.2044	0.0
DeiT-S/16	0.6352	0.5880	0.5380	0.4948	0.4476	0.4004	0.3408	0.2604	0.0
Dist-DeiT-S/16	0.6072	0.5600	0.5176	0.4716	0.4172	0.3620	0.3108	0.2360	0.0
SAM-ViT	0.6320	0.6040	0.5520	0.5360	0.5040	0.4600	0.4160	0.3560	0.0
T2T-ViT-14	0.4044	0.3816	0.3580	0.3348	0.3044	0.2660	0.2276	0.1816	0.0
VGG16	0.3772	0.3220	0.2796	0.2372	0.1964	0.1580	0.1176	0.0768	0.0
ResNet50	0.4584	0.4096	0.3604	0.3140	0.2676	0.2208	0.1820	0.1268	0.0
SEResNet50	0.4880	0.4440	0.3880	0.3360	0.2920	0.2680	0.2160	0.1760	0.0
ConvNext-S	0.5160	0.4760	0.4320	0.3920	0.3480	0.2880	0.2440	0.1880	0.0

2.3 Provably Certified Robustness Comparison Using Denoised Randomized Smoothing

We train the denoisers using the stability objective for 25 epochs with a noise level of $\sigma = 0.25$, learning rate of 10^{-5} and a batch size of 64. We report the certified accuracy versus different radii as defined in (19) in Table 4. We conclude our findings in two points: **Pure ViTs possess better certified robust accuracy than CNNs.** As shown in Table 4, pure ViTs (ViT-S/16, DeiT-S/16, Dist-DeiT-S/16 and SAM-ViT) have higher certified robust accuracy than CNNs. Introducing T2T blocks to ViTs can cause the model to have inferior certified robust accuracy even than CNNs especially for small radii, e.g., radii smaller than 0.5. While sharpness-aware minimization helps further improve ViTs' certified robust accuracy. **Modern CNN design helps bridge the performance gap between CNNs and ViTs.** The design of modern non-transformer models, e.g. ConvNext, has borrowed many techniques from transformers. For example, using larger kernel size to imitate the global attention mechanism of transformers, following the transformers to change stem to "Patchify", using invertible bottleneck as transformers do, substituting BN with LN, replacing ReLU with GeLU, etc. All these changes are targeted to imitate the transformer's operation without introducing the attention blocks. MLP-Mixer also does similar modifications. Our experiments (Table 2, Table 3 and Table 4) show that such modification helps bridge the performance gap between CNNs and transformers not only in terms of clean accuracy, but also empirical and certified robust accuracy. We also show that CNNs with attention mechanism, i.e. SEResNet50, also has better certified robustness than CNNs.

3 Conclusion

This paper presents a comprehensive study on the robustness of ViTs against adversarial perturbations. Our results indicate that ViTs are more robust than CNNs on the considered adversarial attacks and in certified robustness settings. We show that the features learned by ViTs contain less low-level information, contributing to improved robustness against adversarial perturbations that often contain high-frequency components. Also, introducing convolutional blocks in ViTs can facilitate

learning low-level features but has a negative effect on adversarial robustness and makes the models more sensitive to high-frequency perturbations. Moreover, both the sanity performance and the (certified and empirical) adversarial robustness are improved in the modern CNN designs that leverage techniques from ViTs to imitate the global attention behavior. Our work provides a deep understanding of the intrinsic robustness of ViTs and can be used to inform the design of robust vision models based on the transformer structure. In Appedix A, we also conduct a sanity check to verify that ViT's improvement is not caused by insufficient attack optimization, and an explanation from Hopfield network perspective is provided. Besides, we verify that adversarial training could be directly applied to ViTs in Appendix F.

References

[1] Rima Alaifari, Giovanni S Alberti, and Tandri Gauksson. Adef: an iterative algorithm to construct adversarial deformations. *arXiv preprint arXiv:1804.07729*, 2018.

[2] Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020.

[3] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision*, pages 213–229. Springer, 2020.

[4] Mark Chen, Alec Radford, Rewon Child, Jeffrey Wu, Heewoo Jun, David Luan, and Ilya Sutskever. Generative pretraining from pixels. In *International Conference on Machine Learning*, pages 1691–1703. PMLR, 2020.

[5] Xiangning Chen, Cho-Jui Hsieh, and Boqing Gong. When vision transformers outperform resnets without pretraining or strong data augmentations. *arXiv preprint arXiv:2106.01548*, 2021.

[6] Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *International Conference on Machine Learning*, pages 2206–2216. PMLR, 2020.

[7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[10] Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. *arXiv preprint arXiv:2010.01412*, 2020.

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *European conference on computer vision*, pages 630–645. Springer, 2016.

[12] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[13] Dmitry Krotov and John Hopfield. Dense associative memory is robust to adversarial inputs. *Neural computation*, 30(12):3151–3167, 2018.

[14] Dmitry Krotov and John J Hopfield. Dense associative memory for pattern recognition. *Advances in Neural Information Processing Systems*, 2016.

[15] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. *arXiv preprint arXiv:2103.14030*, 2021.

[16] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

[17] Tianyu Pang, Xiao Yang, Yinpeng Dong, Hang Su, and Jun Zhu. Bag of tricks for adversarial training. *arXiv preprint arXiv:2010.00467*, 2020.

[18] Hubert Ramsauer, Bernhard Schäfl, Johannes Lehner, Philipp Seidl, Michael Widrich, Lukas Gruber, Markus Holzleitner, Milena Pavlović, Geir Kjetil Sandve, Victor Greiff, et al. Hopfield networks is all you need. *arXiv preprint arXiv:2008.02217*, 2020.

[19] Hadi Salman, Mingjie Sun, Greg Yang, Ashish Kapoor, and J Zico Kolter. Denoised smoothing: A provable defense for pretrained classifiers. *arXiv preprint arXiv:2003.01908*, 2020.

[20] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*, pages 10347–10357. PMLR, 2021.

[21] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *arXiv preprint arXiv:1706.03762*, 2017.

[22] Haohan Wang, Xindi Wu, Zeyi Huang, and Eric P Xing. High-frequency component helps explain the generalization of convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8684–8694, 2020.

[23] Eric Wong, Leslie Rice, and J Zico Kolter. Fast is better than free: Revisiting adversarial training. *arXiv preprint arXiv:2001.03994*, 2020.

[24] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016.

[25] Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric Xing, Laurent El Ghaoui, and Michael Jordan. Theoretically principled trade-off between robustness and accuracy. In *International Conference on Machine Learning*, pages 7472–7482. PMLR, 2019.

[26] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020.

Supplemental Material

In this supplemental material, we provide more analysis and results in our experiments.

A The Source of Adversarial Robustness

In this section we examine the source of the adversarial robustness revealed in our experiments.

The improved robustness of ViT is not caused by insufficient attack optimization. We first demonstrate that the better robustness of ViTs in white-box attacks is not caused by the difficult optimization in ViT by plotting the loss landscape with sufficient attack steps.

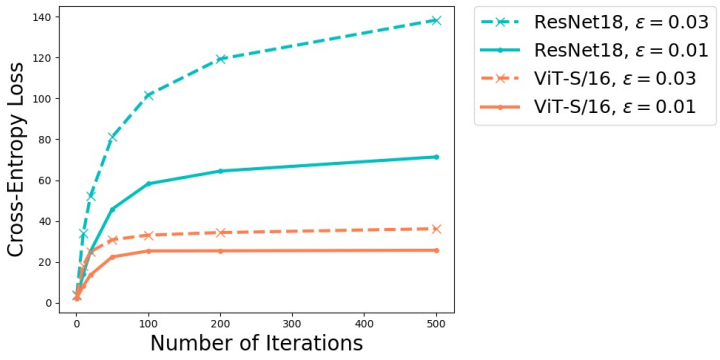

Figure 3: Cross entropy loss versus varying PGD attack steps for ViT-S/16 and RestNet18. The dashed lines corresponds to larger attach radius of 0.03 and the full lines to smaller attack radius of 0.01.

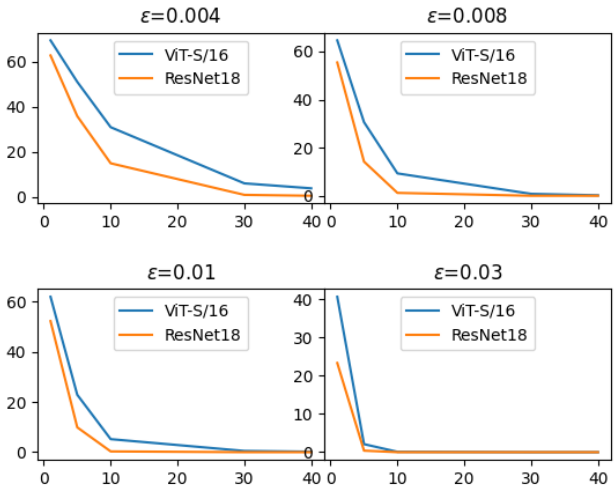

Figure 4: Robust accuracy versus varying PGD attack steps. The attack radii used for evaluation are shown in subtitles.

Figure 3 shows the cross entropy loss versus varing PGD attack steps for ViT-S/16 and ResNet18. Figure 4 shows the robust accuracy versus varing PGD attack steps. As shown in the figures, ViT's loss curves converge at a much lower value than RestNet18, suggesting that the improved robustness of ViT is not caused by insufficient attack optimization.

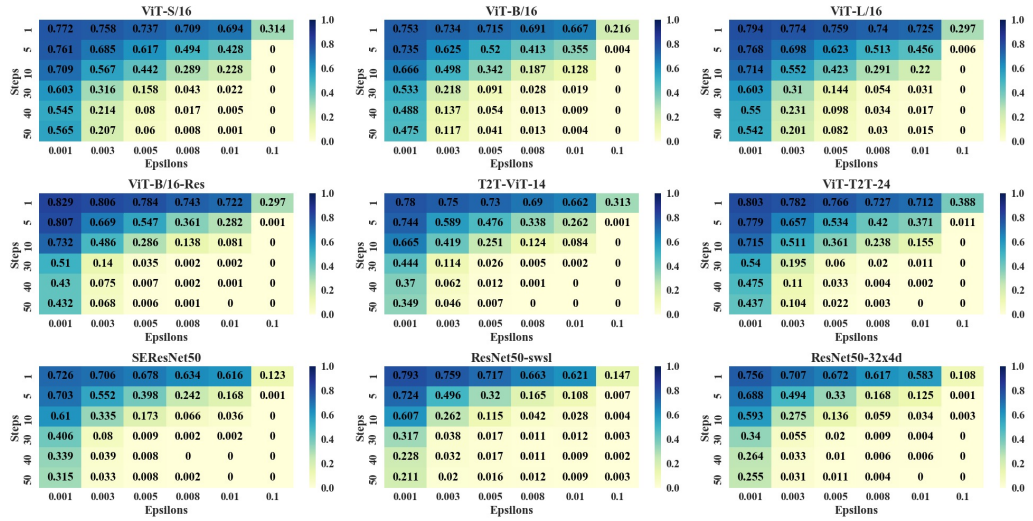

Figure 5: Adversarial accuracy of the target models against PGD attack with different attack radii ("eps") and attack steps ("steps"). When the attack radius and attack steps are increased, the adversarial accuracy of the target model decreases to zero. Darker blocks stand for more robust models against PGD attack.

Figure 5 shows the robust accuracy of more target models against PGD attack with different attack radii ("eps") and attack steps ("steps"). Vision transformers have darker blocks than CNNs', which stands for their superior adversarial robustness against PGD attack.

A Hopfield Network Perspective The equivalence between the attention mechanism in transformers to the modern Hopfield network (14) was recently shown in (18). Furthermore, on simple Hopfield network (one layer of attention-like network) and dataset (MNIST), improved adversarial robustness was shown in (13). Therefore, the connection of attention in transformers to the Hopfield network can be used to explain the improved adversarial robustness for ViTs.

B Experiments on SOTA ViT Structures

In this section, we supplement the experimental results of recently proposed SOTA ViTs.

Swin-Trasnformer (15) computes the representations with shifted windows scheme which brings greater efficiency by limiting self-attention computation to non-overlapping local windows while also allowing for cross-window connection.

DeiT (20) further improves the ViTs' performance using data augmentation or distillation from CNN teachers with an additional distillation token.

SAM-ViT (5) uses sharpness-aware minimization (10) to train ViTs from scratch on ImageNet without large-scale pretraining or strong data augmentations.

Table 5 summarizes the information of models investigated in our experiments. The window size of the swin transformers in Table 5 is 7. The pre-trained weights of these models are available in `timm` package.

Table 6 shows the clean and robust accuracy of ViTs in Table 5 against 40-step PGD attack with different radii. And results for AutoAttack are shown in Table 7. Swin-transformers introduce shifted windows scheme that limit self-attention computation to non-overlapping local windows, which harms the robustness as Tokens-to-Token scheme according to the above results.

Table 5: SOTA ViT models investigated in our experiments.

Model	Layers	Hidden size	Heads	Params
Deit-T/16 (20)	12	192	3	6M
Deit-S/16 (20)	12	384	6	22M
Deit-B/16 (20)	12	768	12	87M
Dist-Deit-T/16 (20)	12	192	3	6M
Dist-Deit-S/16 (20)	12	384	6	22M
Dist-Deit-B/16 (20)	12	768	12	87M
ViT-SAM-B/16 (5)	12	768	12	87M
ViT-SAM-B/32 (5)	12	768	12	88M
Swin-T/4 (15)	(2,2,6,2)	96	(3,6,12,24)	28M
Swin-S/4 (15)	(2,2,18,2)	96	(3,6,12,24)	50M
Swin-B/4 (15)	(2,2,18,2)	128	(4,8,16,32)	88M
Swin-L/4 (15)	(2,2,18,2)	192	(6,12,24,48)	197M

Table 6: Robust accuracy (%) of ViTs described in Table 5 against 40-step PGD attack with different attack radii, and also the clean accuracy ("Clean"). A model is considered to be more robust if the robust accuracy is higher.

Model	Clean	0.001	0.003	0.005	0.01
Deit-T/16	72.3	36.8	8.3	2.6	0.3
Deit-S/16	77.7	48.9	17.6	7.1	1.1
Deit-B/16	81.3	46.6	14.3	6.0	0.9
Dist-Deit-T/16	74.4	40.6	5.7	0.7	0.2
Dist-Deit-S/16	79.3	52.4	15.1	4.3	0.3
Dist-Deit-B/16	81.8	55.6	17.7	4.5	0.4
ViT-SAM-B/16	76.7	63.4	37.0	20.1	3.8
ViT-SAM-B/32	63.8	53.2	32.3	19.7	3.1
Swin-T/4	78.8	33.5	6.0	1.2	0.1
Swin-S/4	81.8	40.0	12.4	3.2	0.2
Swin-B/4	82.3	38.8	11.1	4.1	0.3
Swin-L/4	84.2	38.7	11.1	2.9	0.4

Table 7: Robust accuracy (%) of ViTs described in Table 5 against AutoAttack with different attack radii, and also the clean accuracy ("Clean"). A model is considered to be more robust if the robust accuracy is higher.

Model	Clean	0.001	0.003	0.005	0.01
Deit-T/16	72.3	23.4	0.5	0.0	0.0
Deit-S/16	77.7	30.2	1.2	0.0	0.0
Deit-B/16	81.3	20.4	0.3	0.1	0.0
Dist-Deit-T/16	74.4	31.1	0.8	0.1	0.0
Dist-Deit-S/16	79.3	43.1	3.7	0.2	0.0
Dist-Deit-B/16	81.8	42.7	3.4	0.2	0.0
ViT-SAM-B/16	76.7	59.8	26.0	8.4	0.1
ViT-SAM-B/32	63.8	48.9	23.6	9.7	0.8
Swin-T/4	78.8	6.8	0.1	0.0	0.0
Swin-S/4	81.8	7.9	0.1	0.0	0.0
Swin-B/4	82.3	2.4	0.1	0.0	0.0
Swin-L/4	84.2	4.3	0.1	0.0	0.0

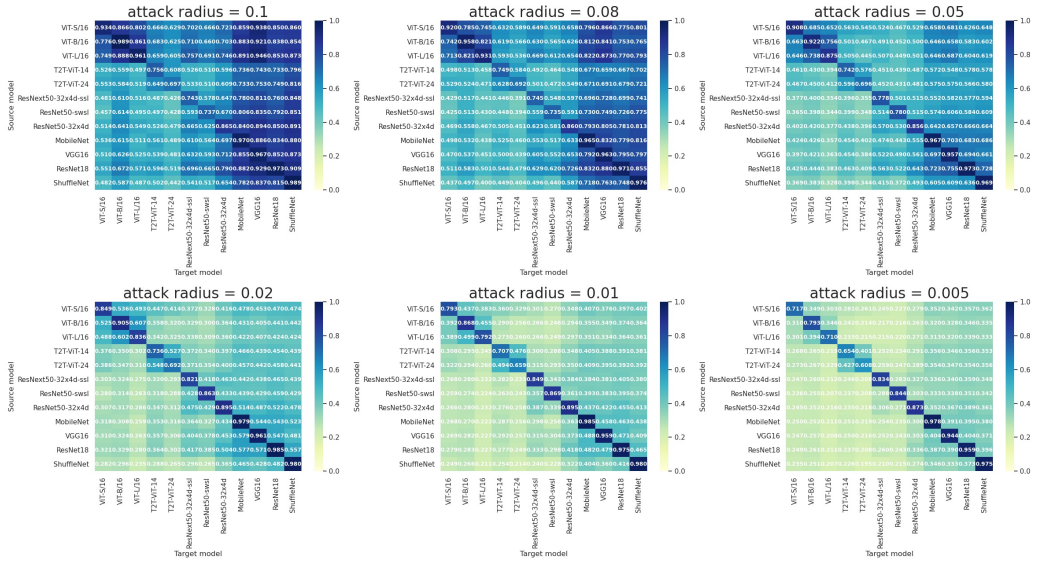

Figure 6: ASR of transfer attack using FGSM with different attack radii. The rows stand for the surrogate models used to generated adversarial examples in the white-box attack approach. The columns stand for the target models. Darker rows correlate to the source models that generate more transferable adversarial examples. While darker columns mean that the target models are more vulnerable to the transfer attack."Res50-ssl" and "Res50-swsl" are in short of "ResNeXt-32x4d-ssl" and "ResNet50-swsl" respectively.

C Experiments on Cifar-10

We choose the ImageNet as the benchmark because ViTs can hardly converge when training directly on small datasets like Cifar. Therefore, we finetune the ViTs instead. As shown in Table 8, ViT-B/4 performs higher robust accuracy than WideResNet, which is consistent with the trend on ImageNet.

Table 8: Robust accuracy of ViT-B/4 and WideResNet against PGD-10 attack with different attack radii.

Model	**0.001**	**0.003**	**0.01**	**0.03**
ViT-B/4	0.9202	0.6242	0.0994	0.0103
WideResNet	0.7744	0.5923	0.0854	0.0000

D Robustness Against Adversarial Deformation

Besides additively perturbing the correctly classified image, ADef (1) iteratively applies small deformations to the clean data. We show the robust accuracy against such perturbations in Table 9, which is in accordance to the results of PGD and AutoAttack.

Table 9: Robust accuracy (%) against AFef under the default setting described in (1).

Model	ViT-S/16	VGG16	DenseNet	MobileNet	ResNet18
Robust Accuracy	**12.4**	10.8	11.1	11.7	11.8

E Transfer Attack Results

Transfer attack results using more attack radii are provided in Figure 6

Table 10: Results of adversarial training for different models using PGD-7 (7-step PGD attack) and TRADES respectively on CIFAR-10. ViT-B/4 is a variant of ViT-B/16 where we downsample the patch embedding kernel from 16×16 to 4×4 to accommodate the smaller image size on CIFAR-10. We report the clean accuracy (%) and robust accuracy (%) evaluated with PGD-10 and AutoAttack respectively. Each model is trained using only 20 epochs to reduce the cost.

Model	Method	Clean	PGD-10	AutoAttack
PreActResNet18	PGD-7	77.3	48.9	44.4
	TRADES	77.6	49.4	44.9
WideResNet-34-10	PGD-7	80.3	52.2	48.4
	TRADES	81.6	53.4	49.3
ViT-B/4	PGD-7	85.9	51.7	47.6
	TRADES	85.0	53.9	49.2

F Adversarial Training

Settings We also conduct a preliminary experiment on adversarial training for ViT. For this experiment we use CIFAR-10 (12) with $\epsilon = 8/255$ and the ViT-B/16 model. Since originally this ViT was pre-trained on ImageNet with image size 224×224 and patch size 16×16 while image size on CIFAR-10 is 32×32, we downsample the weights for patch embeddings and resize patches to 4×4, so that there are still 8×8 patches and we name the new model as ViT-B/4. Though ViT originally enlarged input images on CIFAR-10 for natural fine-tuning and evaluation, we keep the input size as 32×32 to make the attack radius comparable. For training, we use PGD-7 (PGD with 7 iterations) (16) and TRADES (25) methods respectively, with no additional data during adversarial training. We compare ViT with two CNNs, ResNet18 (11) and WideResNet-34-10 (24). To save training cost, we train each model for 20 epochs only, although some prior works used around hundreds of epochs (16; 17) and are very costly for large models. We use a batch size of 128, an initial learning rate of 0.1, an SGD optimizer with momentum 0.9, and the learning rate decays after 15 epochs and 18 epochs respectively with a rate of 0.1. While we use a weight decay of 5×10^{-4} for CNNs as suggested by (17) that 5×10^{-4} is better than 2×10^{-4}, we still use 2×10^{-4} for ViT as we find 5×10^{-4} causes an underfitting for ViT. We evaluate the models with PGD-10 (PGD with 10 iterations) and AutoAttack respectively.

Results We show the results in Table 10. The ViT model achieves higher robust accuracy compared to ResNet18, and comparable robust accuracy compared to WideResNet-34-10, while ViT achieves much better clean accuracy compared to the other two models. Here ViT does not advance the robust accuracy after adversarial training compared to large CNNs such as WideResNet-34-10. We conjecture that ViT may need larger training data or longer training epochs to further improve its robust training performance, inspired by the fact that on natural training ViT is not able to perform well either without large-scale pre-training. And although T2T-ViT improved the performance of natural training when trained from scratch, our previous results in Table 2 and Table 3 show that the T2T-ViT structure may be inherently less robust. We have also tried (23) which was proposed to mitigate the overfitting of FGSM to conduct fast adversarial training with FGSM, but we find that it can still cause catastrophic overfitting for ViT such that the test accuracy on PGD attacks remains almost 0. We conjecture that this fast training method may be not suitable for pre-trained models or require further adjustments. Our experiments in this section demonstrate that the adversarial training framework with PGD or TRADES is applicable for transformers on vision tasks, and we provide baseline results and insights for future exploration and improvement.