Evaluating Model Robustness to Patch Perturbations

Jindong Gu [1] **Volker Tresp** [1] **Yao Qin** [2]

Abstract

Recent advances in Vision Transformer (ViT) have demonstrated its impressive performance in image classification, which makes it a promising alternative to Convolutional Neural Network (CNN). Unlike CNNs, ViT represents an input image as a sequence of image patches. The patch-based input image representation makes the following question interesting: How does ViT perform when individual input image patches are perturbed with natural corruptions or adversarial perturbations, compared to CNNs? In this submission, we propose to evaluate model robustness to patch-wise perturbations. Two types of patch perturbations are considered to model robustness. One is natural corruptions, which is to test models' robustness under distributional shifts. The other is adversarial perturbations, which are created by an adversary to specifically fool a model to make a wrong prediction. The experimental results on the popular CNNs and ViTs are surprising. We find that ViTs are more robust to naturally corrupted patches than CNNs, whereas they are more vulnerable to adversarial patches. Given the architectural traits of state-of-the-art ViTs and the interesting results above, we propose to add the robustness to natural patch corruption and adversarial patch attack into the robustness benchmark.

1. Motivation

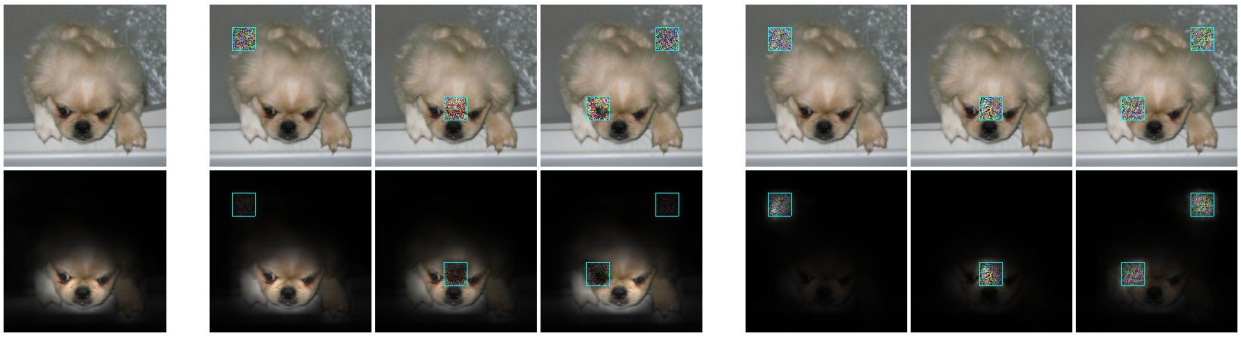
Recently, Vision Transformer (ViT) has demonstrated impressive performance (Dosovitskiy et al., 2020; Touvron et al., 2021; Wu et al., 2020; Xiao et al., 2021; Graham et al., 2021; Chen et al., 2021b; Han et al., 2021; Chen et al., 2021a; Liu et al., 2021), which makes it become a potential alternative to convolutional neural networks (CNNs). Meanwhile, the robustness of ViT has also received great attention (Bhojanapalli et al., 2021; Joshi et al., 2021; Salman et al., 2021; Shao et al., 2021; Shi & Han, 2021; Tang et al., 2021). On the one hand, it is important to improve its robustness for safe deployment in the real world. On the other hand, diagnosing the vulnerability of ViT can also give us a deeper understanding of its underlying working mechanisms. Existing works have intensively studied the robustness of ViT and CNNs when the whole input image is perturbed with natural corruptions or adversarial perturbations (Bhojanapalli et al., 2021; Shao et al., 2021; Mahmood et al., 2021; Bai et al., 2021; Aldahdooh et al., 2021). Unlike CNNs, ViT processes the input image as a sequence of image patches. In this work, instead, we propose to study the robustness of ViT to patch-wise perturbations based on its special patch-based architecture.

In this work, two typical types of perturbations are considered to compare the robustness between ViTs and CNN (e.g., ResNets (He et al., 2016)). One is natural corruptions (Hendrycks & Dietterich, 2019), which is to test models' robustness under distributional shift. The other is adversarial perturbations (Szegedy et al., 2014; Goodfellow et al., 2014), which are created by an adversary to specifically fool a model to make a wrong prediction. Surprisingly, we find ViT does *not always* perform more robustly than ResNet. When individual image patches are naturally corrupted, ViT is more robust compared to ResNet. However, when input image patch(s) are adversarially attacked, ViT shows a higher vulnerability than ResNet.

To better understand the model robustness to patch perturbation, we revealed that ViT's stronger robustness to natural corrupted patches and higher vulnerability against adversarial patches are both caused by the attention mechanism. Specifically, the self-attention mechanism of ViT can effectively ignore the natural patch corruption, while it's also easy to manipulate the self-attention mechanism to focus on an adversarial patch. This is well supported by rollout attention visualization (Abnar & Zuidema, 2020) on ViT. As shown in Fig. 1 (a), ViT successfully attends to the class-relevant features on the clean image, *i.e.*, the head of the dog. When one or more patches are perturbed with natural corruptions, shown in Fig. 1 (b), ViT can effectively ignore the corrupted patches and still focus on the main foreground to make a correct prediction. In Fig. 1 (b), the attention weights on the positions of naturally corrupted patches are much smaller even when the patches appear on the fore-

[1]University of Munich [2]Google Research. Correspondence to: Yao Qin <yaoqin@google.com>.

Shift Happens Workshop : Crowdsourcing metrics and test datasets beyond ImageNet in ICML 2022, Copyright 2022 by the author(s).



(a) Clean Image (b) with Naturally Corrupted Patch (c) with Adversarial Patch

Figure 1. Images with patch-wise perturbations (top) and their corresponding attention maps (bottom). The attention mechanism in ViT can effectively ignore the naturally corrupted patches to maintain a correct prediction in Fig. b, whereas it is forced to focus on the adversarial patches to make a mistake in Fig. c. The images with corrupted patches (Fig. b) are all correctly classified. The images with adversary patches (Fig. c) are misclassified as *dragonfly*, *axolotl*, and *lampshade*, respectively.

ground. In contrast, when the patches are perturbed with adversarial perturbations by an adversary, ViT is successfully fooled to make a wrong prediction, as shown in Fig. 1 (c). This is because the attention of ViT is misled to focus on the adversarial patch instead.

In summary, given the architectural traits of state-of-the-art ViTs and the interesting results we found, we propose to add the robustness to corrupted patches and adversarial patches into the robustness benchmark.

2. Data Generation

Natural Corrupted Patches. For each image, we first select n image patch $\boldsymbol{x}_i$ from the input image and perturb them with natural corruptions. As in (Hendrycks & Dietterich, 2019), 15 types of natural corruptions are applied to the selected patches, respectively. 5 severity levels of perturbation are considered. The final performance is averaged over different corruption types, all severity levels, and selected patches. The positions and sizes of patches as well as the number of patches are hyper-parameters to be specified.

Adversarial Patches. We now introduce adversarial patch attack (Karmon et al., 2018) used in our study. The first step is to specify a patch position and replace the original pixel values of the patch with random initialized noise δ. The second step is to update the noise to minimize the probability of ground-truth class, *i.e.* maximize the cross-entropy loss via multi-step gradient ascent (Madry et al., 2017). Similarly, the positions and sizes of patches as well as the number of patches are hyper-parameters to be specified.

For both types of perturbations, the bound (255/255) is applied to the perturbed patches. The resulting images are kept in the valid image space [0, 255]. Our code will be integrated into the ShiftHappens robustness benchmark.

3. Evaluation Metric

We use the standard metric **Fooling Rate (FR)** to evaluate the model robustness. First, we collect a set of images that are correctly classified by both models that we compare. The number of these collected images is denoted as P. When these images are perturbed with natural patch corruption or adversarial patch attack, we use Q to denoted the number of images that are misclassified by the model. The Fooling Rate is then defined as FR = $\frac{Q}{P}$. The lower the FR is, the more robust the model is.

4. Special Requirements on the Models to Be Evaluated

The evaluation of model robustness to patch perturbation is model-agnostic. All the classification models can be evaluated on the proposed task, including the traditional machine learning classifiers.

There is a special setting for evaluation of robustness of Vision Transformers. ViT represent an input images as a list image patches. The setting of our patch perturbation can be perfect (un)aligned with the patch representations. Given the popularity of current ViT family models, it is indeed an interesting setting.

5. Evaluation Procedure

In the experiment, 10k test images are randomly selected from ImageNet-1k validation dataset (Deng et al., 2009) that are correctly classified by models to be evaluated. Then for each image, we randomly sample n input image patches $\boldsymbol{x}_i$ from 196 patches and perturb them with natural corruptions. As in (Hendrycks & Dietterich, 2019), 15 types of natural corruptions with the highest level (or 5 severity levels) are

Table 1. Fooling Rates (in %) are reported. DeiT is more robust to naturally corrupted patches than ResNet, while it is significantly more vulnerable than ResNet against adversarial patches. Bold font is used to mark the lower fooling rate, which indicates the higher robustness.

Model	No. of Naturally Corrupted Patches				No. of Adversarial Patches			
	32	96	160	196	1	2	3	4
ResNet50	3.7	18.2	43.4	49.8	**30.6**	**59.3**	**77.1**	**87.2**
DeiT-small	**1.8**	**7.4**	**22.1**	**38.9**	61.5	95.4	99.9	100
ResNet18	6.8	31.6	56.4	61.3	**39.4**	**73.8**	**90.0**	**96.1**
DeiT-tiny	**6.4**	**14.6**	**35.8**	**55.9**	63.3	95.8	99.9	100

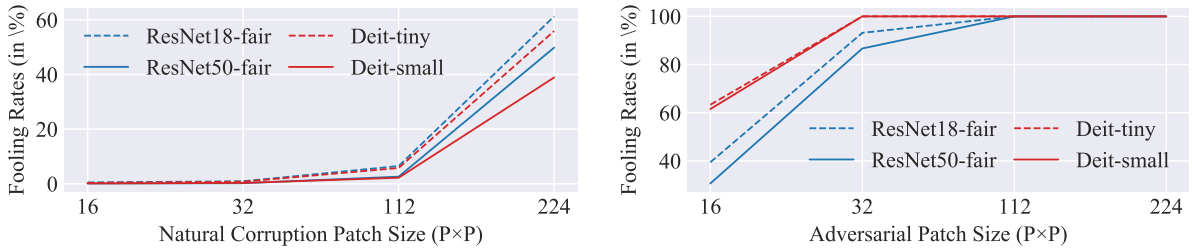


Figure 2. DeiT with red lines shows a smaller FR to natural patch corruption and a larger FR to adversarial patch of different sizes than counter-part ResNet.

applied to the selected patches. The fooling rate of the patch-based natural corruption is averaged over all the test images and corruption types. The same image selection is also applied in case of adversarial patches. The ℓ_∞-norm bound of 255/255, the step size of 2/255, and the attack iterations of 10K is applied to create adversarial patches.

6. Evaluation Output and Analysis

Patch-wise Natural Corruption. First, we investigate the robustness of DeiT and ResNet to patch-based natural corruptions. We find that both DeiT and ResNet hardly degrade their performance when a small number of patches are corrupted (*e.g.*, 4). When we increase the number of patches, the difference between two architectures emerges: DeiT achieves a lower FR compared to its counter-part ResNet (See Tab. 1). This indicates that DeiT is more robust against naturally corrupted patches than ResNet. The same conclusion holds under the extreme case when the number of patches $n = 196$ where the whole image is perturbed with natural corruptions. This is aligned with the observation in the existing work (Bhojanapalli et al., 2021) that ViTs are more robust to ResNet under distributional shifts.

In addition, we also increase the patch size of the perturbed patches, *e.g.*, if the patch size of the corrupted patch is 32×32, it means that it covers 4 continuous and independent input patches as the input patch size is 16×16. As shown in Fig. 2 (Left), even when the patch size of the perturbed patches becomes larger, DeiT (marked with red lines) is still more robust than its counter-part ResNet (marked with blue lines) to natural patch corruption.

Patch-wise Adversarial Attack As shown in Tab. 1, DeiT achieves much higher fooling rate than ResNet when one of the input image patches is perturbed with adversarial perturbation. This consistently holds even when we increase the number of adversarial patches, sufficiently supports that DeiT is more vunerable than ResNet against patch-wise adversarial perturbation. When more than 4 patches ($\sim$2% area of the input image) are attacked, both DeiT and ResNet can be successfully fooled with almost 100% FR.

When we attack a large continuous area of the input image by increasing the patch size of adversarial patches, the FR on DeiT is still much larger than counter-part ResNet until both models are fully fooled with 100% FR. As shown in Fig. 2 (Right), DeiT (marked with red lines) has higher FR than ResNet under different adversarial patch sizes.

Taking all the results above together, we discover that DeiT is more robust to natural patch corruption than ResNet, whereas it is significantly more vulnerable to adversarial patch perturbation.

We also report different versions of ViT (Dosovitskiy et al., 2020; Touvron et al., 2021; Liu et al., 2021), CNN (He et al., 2016; Huang et al., 2017) as well as Hybrid architectures (Graham et al., 2021). We train all the models in the same setting as in (Touvron et al., 2021) and report fooling rate on each model in Fig. 3. Four main conclusions can be drawn from the figure. 1). CNN variants are more robust than ViT models. 2). The robustness of LeViT model (Graham et al., 2021) with hybrid architecture (*i.e.*, Conv Layers + Self-Attention Blocks) lives somewhere between ViT and CNNs, as expected. 3). Swin Transformers (Liu et al., 2021)

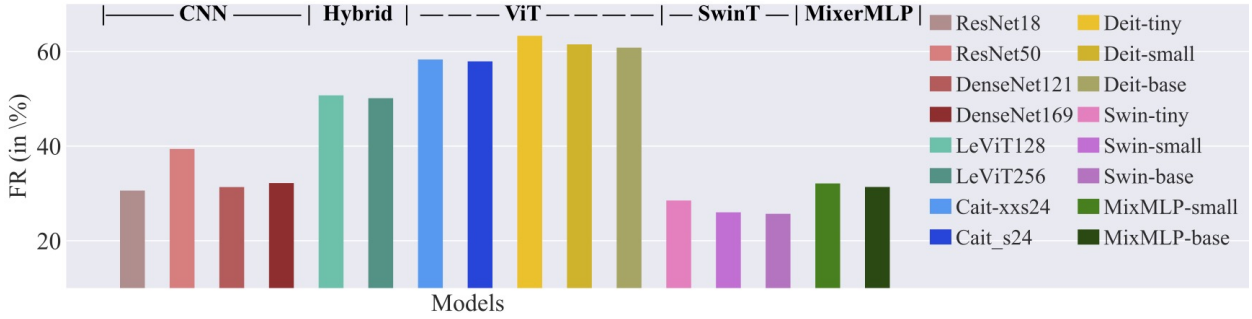


Figure 3. We report Fooling Rates on different versions of ViT, CNN as well as Hybrid architectures under Adversarial Patch Attack.

are as robust as CNNs since attention cannot be manipulated by a single patch due to hierarchical attention and the shifted windows therein. The self-attention in Swin Transformers is only conducted on patches within a local region. With shifted windows, a single patch will interact with patches from different groups in different layers. Both designs make effective adversarial patches challenging. That's the reason why Swin Transformer performs more robustly than popular ViTs. 4). Mixer-MLP (Tolstikhin et al., 2021) uses the same patch-based architecture as ViTs and has no attention module. Mixer-base with FR (31.36) is comparable to ResNet and more robust than ViTs. The results confirm that the vulnerability of ViT can be attributed to self-attention mechanism.

7. Related Works

The robustness of ViT have achieved great attention due to its great success (Bhojanapalli et al., 2021; Naseer et al., 2021a; Shao et al., 2021; Benz et al., 2021; Mahmood et al., 2021; Bai et al., 2021; Naseer et al., 2021b; Aldahdooh et al., 2021; Salman et al., 2021; Yu et al., 2021; Hu et al., 2021; Mao et al., 2021b;a; Naseer et al., 2021a; Tang et al., 2021). On the one hand, (Bhojanapalli et al., 2021; Paul & Chen, 2021) show that vision transformers are more robust to natural corruptions (Hendrycks & Dietterich, 2019) compared to CNNs. On the other hand, (Bhojanapalli et al., 2021; Shao et al., 2021; Paul & Chen, 2021) demonstrate that ViT achieves higher adversarial robustness than CNNs under adversarial attacks. These existing works, however, mainly focus on investigating the robustness of ViT when a whole image is naturally corrupted or adversarially perturbed. Instead, our work focuses on patch perturbation, given the patch-based architecture trait of ViT. The patch-based attack (Joshi et al., 2021; Fu et al., 2021) and defense (Mu & Wagner, 2021; Shi & Han, 2021) methods have also been proposed recently. Different from their work, we aim to understand the robustness of patch-based architectures under patch-based natural corruption and adversarial patch perturbation.

8. Conclusion

This work first shows our motivation to evaluating model robustness to patch robustness. With experimental results. our work shows an interesting observation on the robustness of ViT to patch perturbations. Namely, vision transformer (e.g., DeiT) is more robust to natural patch corruption than ResNet, whereas it is significantly more vulnerable against adversarial patches. A deep understanding of the observation is then provided. We reveal that the self-attention mechanism of ViT can effectively ignore natural corrupted patches but be easily misled to adversarial patches to make mistakes. We hope this study can help the community better understand the robustness of ViT to patch perturbations.

References

Abnar, S. and Zuidema, W. Quantifying attention flow in transformers. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2020.

Aldahdooh, A., Hamidouche, W., and Deforges, O. Reveal of vision transformers robustness against adversarial attacks. *arXiv:2106.03734*, 2021.

Bai, Y., Mei, J., Yuille, A., and Xie, C. Are transformers more robust than cnns? *arXiv:2111.05464*, 2021.

Benz, P., Ham, S., Zhang, C., Karjauv, A., and Kweon, I. S. Adversarial robustness comparison of vision transformer and mlp-mixer to cnns. *arXiv preprint arXiv:2110.02797*, 2021.

Bhojanapalli, S., Chakrabarti, A., Glasner, D., Li, D., Unterthiner, T., and Veit, A. Understanding robustness of transformers for image classification. *arXiv:2103.14586*, 2021.

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

Chen, Z., Xie, L., Niu, J., Liu, X., Wei, L., and Tian, Q. Visformer: The vision-friendly transformer. *arXiv:2104.12533*, 2021b.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv:2010.11929*, 2020.

Fu, Y., Zhang, S., Wu, S., Wan, C., and Lin, Y. Patch-fool: Are vision transformers always robust against adversarial perturbations? In *International Conference on Learning Representations*, 2021.

Goodfellow, I. J., Shlens, J., and Szegedy, C. Explaining and harnessing adversarial examples. *arXiv:1412.6572*, 2014.

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

Han, K., Xiao, A., Wu, E., Guo, J., Xu, C., and Wang, Y. Transformer in transformer. *arXiv:2103.00112*, 2021.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

Hendrycks, D. and Dietterich, T. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations (ICLR)*, 2019.

Hu, H., Lu, X., Zhang, X., Zhang, T., and Sun, G. Inheritance attention matrix-based universal adversarial perturbations on vision transformers. *IEEE Signal Processing Letters*, 28:1923–1927, 2021.

Huang, G., Liu, Z., Van Der Maaten, L., and Weinberger, K. Q. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4700–4708, 2017.

Joshi, A., Jagatap, G., and Hegde, C. Adversarial token attacks on vision transformers. *arXiv:2110.04337*, 2021.

Karmon, D., Zoran, D., and Goldberg, Y. Lavan: Localized and visible adversarial noise. In *International Conference on Machine Learning (ICML)*, 2018.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., and Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. *arXiv:2103.14030*, 2021.

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

Mahmood, K., Mahmood, R., and Van Dijk, M. On the robustness of vision transformers to adversarial examples. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7838–7847, 2021.

Mao, X., Qi, G., Chen, Y., Li, X., Duan, R., Ye, S., He, Y., and Xue, H. Towards robust vision transformer. *arXiv:2105.07926*, 2021a.

Mao, X., Qi, G., Chen, Y., Li, X., Ye, S., He, Y., and Xue, H. Rethinking the design principles of robust vision transformer. *arXiv:2105.07926*, 2021b.

Mu, N. and Wagner, D. Defending against adversarial patches with robust self-attention. In *ICML 2021 Workshop on Uncertainty and Robustness in Deep Learning*, 2021.

Naseer, M., Ranasinghe, K., Khan, S., Hayat, M., Khan, F. S., and Yang, M.-H. Intriguing properties of vision transformers. *arXiv:2105.10497*, 2021a.

Naseer, M., Ranasinghe, K., Khan, S., Khan, F. S., and Porikli, F. On improving adversarial transferability of vision transformers. *arXiv:2106.04169*, 2021b.

Paul, S. and Chen, P.-Y. Vision transformers are robust learners. *arXiv:2105.07581*, 2021.

Salman, H., Jain, S., Wong, E., and Madry, A. Certified patch robustness via smoothed vision transformers. *arXiv:2110.07719*, 2021.

Shao, R., Shi, Z., Yi, J., Chen, P.-Y., and Hsieh, C.-J. On the adversarial robustness of visual transformers. *arXiv:2103.15670*, 2021.

Shi, Y. and Han, Y. Decision-based black-box attack against vision transformers via patch-wise adversarial removal. *arXiv preprint arXiv:2112.03492*, 2021.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., and Fergus, R. Intriguing properties of neural networks. *International Conference on Learning Representations (ICLR)*, 2014.

Tang, S., Gong, R., Wang, Y., Liu, A., Wang, J., Chen, X., Yu, F., Liu, X., Song, D., Yuille, A., et al. Robustart: Benchmarking robustness on architecture design and training techniques. *arXiv preprint arXiv:2109.05211*, 2021.

Tolstikhin, I., Houlsby, N., Kolesnikov, A., Beyer, L., Zhai, X., Unterthiner, T., Yung, J., Keysers, D., Uszkoreit, J., Lucic, M., et al. Mlp-mixer: An all-mlp architecture for vision. In *arXiv:2105.01601*, 2021.

Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., and Jégou, H. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning (ICML)*, 2021.

Wu, B., Xu, C., Dai, X., Wan, A., Zhang, P., Yan, Z., Tomizuka, M., Gonzalez, J., Keutzer, K., and Vajda, P. Visual transformers: Token-based image representation and processing for computer vision. *arXiv:2006.03677*, 2020.

Xiao, T., Singh, M., Mintun, E., Darrell, T., Dollár, P., and Girshick, R. Early convolutions help transformers see better. *arXiv:2106.14881*, 2021.

Yu, Z., Fu, Y., Li, S., Li, C., and Lin, Y. Mia-former: Efficient and robust vision transformers via multi-grained input-adaptation. *arXiv preprint arXiv:2112.11542*, 2021.