A Few Adversarial Tokens Can Break Vision Transformers

Ameya Joshi* Sai Charitha Akula* Gauri Jagatap
Chinmay Hegde
New York University
{ameya.joshi, sca321, gbj221, chinmay.h}@nyu.edu

Abstract

Vision transformers rely on self-attention operations between disjoint patches (tokens) of an input image, in contrast with standard convolutional networks. We investigate fundamental differences between the adversarial robustness properties of these two families of models when subjected to adversarial token attacks (i.e., where an adversary can modify a tiny subset of input tokens). We subject various transformer and convolutional models with token attacks of varying patch sizes. Our results show that vision transformer models are much more sensitive to token attacks than the current best convolutional models, with SWIN outperforming transformer models by up to $\sim 20\%$ in robust accuracy for single token attacks. We also show that popular vision-language models such as CLIP are even more vulnerable to token attacks. Finally, we also demonstrate that a simple architectural operation (patch-merging), which is used by transformer variants such as SWIN, can significantly enhance robustness to token attacks.

1. Introduction

Motivation: Vision transformers (or ViTs [1]) are now ubiquitous across the entire spectrum of tasks in computer vision. ViT-based models now rank among the state-of-the-art for a variety of tasks, while also providing additional benefits like zero-shot classification [2] and distributional robustness [3]. At the heart of vision transformers is the *self-attention* operation, a mechanism that allows the network to find and exploit correlations between spatially-disjoint, potentially-far away patches of a given input image image; indeed, an image can now be viewed as a collection of disjoint patches. In the context of vision, small non-overlapping patches serve as input *tokens* to the transformer, and models such as ViT, Data Efficient Image Transformers (DeIT) [4], and many other variants all rely on this token-based mechanism to represent images. In comparison, convolutional networks (CNNs) take raw image pixels as input, and each layer only considers localized correlations as inductive bias.

By now, it is well-known that convolutional networks (CNNs) are vulnerable to adversarial attacks [5–7] under a variety of threat models. We can therefore also ask: how well do vision transformers fare against adversarial attacks? This has previously been addressed by several papers [8, 9] which showed that ViTs are at least as robust as CNNs under norm-bounded adversarial perturbation attacks.

However, since transformers process inputs in the form of tokens, this motivates an unique threat model, where a malicious attacker can modify a tiny number of tokens (imperceptibly or otherwise). In this work, we focus on the "token attack" threat model for transformer-based architectures. Specifically, we attempt to answer the question: *Are transformers robust to malicious perturbations to a small subset of the input tokens?*

Our findings bear both good and bad news. On the negative side, we show that vanilla ViT models are worryingly brittle when subjected to token attacks; *even a single adversarially designed token in an image can dramatically affect performance*, compared to similar attacks on modern convolutional models (such as ConvNextv2). This effect is *even worse* when we consider CLIP vision-language models with transformer backbones. On the positive side, we show that *modern variations of ViTs (such as SWIN or BeIT) that have reduced dependency on singular tokens through overlapping patches or masked pretraining are significantly more robust*. Finally, we demonstrate that other similar architectural variants that perform "patch merging" have a degree of inbuilt robustness to token attacks.

Our contributions: As mentioned above, our focus is on the "token attack" model [10] where an adversary is permitted to modify K tokens (patches) of a given input image. Finding the optimal attack under this threat model is combinatorially hard, but we can employ natural relaxations that can be solved by (projected) gradient descent; see Section 3 below for technical details.

Using this attack, we interrogate vulnerabilities of sev-

*Equal contribution

eral families of neural architectures using our token attack; transformer-based (ViT [1], DeIT [4], BeIT, and others), convolutional (Resnets [11], WideResNet [12], ConvNextv2), and finally vision-language models that perform zero-shot classification (specifically, CLIP [2]) with a transformer-based image backbone.

We make the following contributions:

1. With our token attack algorithm, we can significantly degrade the performance of vision transformers using only a small number of tokens (corresponding to 0.5% of pixels) — as opposed to ℓ_2- or ℓ_∞-attacks which rely on perturbing all image pixels. We show consistent degradation of classification performance of all architectures on token attacks of increasing patch sizes and number of patches.
2. We demonstrate that for token attacks accounting for the architecture and token size, transformer architectures relying on non-overlapping patches are far less robust as compared to convolutional networks. Intriguingly, we also show that CLIP [2] models based on transformer backbones, which have generally been shown to be robust to distribution shifts, are far less robust to token attacks.
3. We also observe that models that have reduced dependency on singular tokens, generally achieved through overlapping patches or masked pretraining (SWIN, ConvNextv2 and BeIT) are more robust than other models. We further analyse this effect through various experiments on SWIN and show that using overlapping patches as tokens leads to robustness.

2. Related Work

Vision Transformers: Transformers, introduced by [13], have led to significant improvements in NLP tasks. Following this success in NLP, [1] propose Vision Transformers (ViT) that leverage non-overlapping patches as tokens input to a similar attention based architecture. ViTs have led to significant developments across vision tasks, including zero-shot classification [2], captioning [14], and image generation [15] among others. Vision transformers have further been improved through the use of distillation [4], masked image pre-training (BeIT) [16] and linear time attention layers [17]. Given the recent ubiquity of vision transformers across computer vision, it is of great importance to quantify and analyse their robustness to adversarial perturbations.

Adversarial attacks: Deep networks are vulnerable to imperceptible changes to input images as defined by the ℓ_∞ distance [5]. There exist several test-time attack algorithms with various threat models: ℓ_p constrained [18–20], black-box [21,22], geometric attacks [23,24], semantic and meaningful attacks [25–27] and data poisoning based [28].

Defenses: Due to the vast variety of attacks, adversarial defense is a non-trivial problem. Empirical defenses as proposed by [29], [30], and [31] rely on adversarial data augmentation and modified loss functions to improve robustness. [32, 33] propose preprocessing operations as defenses. However, such defenses fail to counter adaptive attacks [34]. [35], [36] and [37] provide methods that guarantee robustness theoretically.

Patch attacks: Patch attacks [38] are a more practically realizable threat model. [39–41] have successfully attacked detectors and classifiers with physically printed patches. In addition, [7, 42] also show that spatially limited sparse perturbations suffice to consistently reduce the accuracy of classification model. This motivates our analysis of the robustness of recently invented architectures towards sparse and patch attacks.

Attacks and defenses for vision transformers: The popularity of transformer models in image classification have inspired a number of studies about their robustness. [8, 43] analyse the performance of vision transformers in comparison to massive ResNets under various threat models and concur that vision transformers (ViT) are at least as robust as Resnets when pretrained with massive training datasets. However, [10] show that ViTs are less robust than Resnets for adversarial token attacks.

The transferability of adversarial attacks on ViT has also been examined. [44] show that adversarial examples do not transfer well between CNNs and transformers, and build an ensemble based approach towards adversarial defense. [45,46] suggested that adversarial attacks can be transferred between ViTs and CNNs by specifically tailoring attacks to transformers. We consider a orthogonal setup, where we construct adversarial attacks specifically for transformer models to leverage the special input modality. [47] show that ViTs are specifically vulnerable to patch-level transformations, leading to good in-distribution accuracies but poor out-of-distribution performance. [48] present a certified defense for patch attacks, where in ViTs outperform Resnets.

[9] claims that ViTs are robust to a large variety of corruptions due to the attention mechanism. However, [49] shows that dot-product attention can result in vulnerability to adversarial patch attacks and propose adversarial objectives for crafting patches that target this explicitly. [50] finds that ViTs are more effective in dealing with naturally distorted image patches compared to CNNs but are more susceptible to adversarial patches, where the attention mechanism can be easily fooled to focus more on the adversarially perturbed patches. [51] implements a patch attack by using a set of attention-aware optimization techniques that are specifically designed to deceive the self-attention mechanism of the model.

[52] show that it is possible to improve the robustness of CNNs to changes in natural distribution shifts by patchifying input images without incorporating any attention-related techniques. [53] shows that the patchified stem no-

tably improves the robustness with respect to ℓ_2 attacks while being comparable to ℓ_∞ attacks.

In any case, there seems to exist a strong effect on model robustness when subjected to patch-wise (token) perturbations. In this paper we illuminate this effect in greater detail for several model families. We also show through ablations that techniques like patch-merging architectures and masked patch pretraining, which reduce dependence on single patches, further provide significant robustness towards adversarial patches.

3. Token Attacks on Vision Transformers

We begin by introducing Token attacks [10], which specifically are tailored towards targeting transformer architectures that rely on patch-based inputs.

Threat Model: Let $\mathbf{x} \in \mathbb{R}^d$ be a d-dimensional image, and $f : \mathbb{R}^d \rightarrow [m]$ be a classifier that takes $\mathbf{x}$ as input and outputs one of m class labels. For our attacks, we focus on sparsity as the constraining factor. Specifically, we restrict the number of pixels or blocks of pixels that an attacker is allowed to change. We consider $\mathbf{x}$ as a concatenation of B blocks $[\boldsymbol{x}_1, \ldots \boldsymbol{x}_b, \ldots, \boldsymbol{x}_B]$, where each block is of size p. In order to construct an attack, the attacker is allowed to perturb up to $K \leq B$ such blocks for a K-token attack. We also assume a white-box threat model, that is, the attacker has access to the model including gradients and preprocessing.We consider a block sparse token budget, where we restrict the attacker to modifying K patches or "tokens" with an unconstrained perturbation allowed per patch.

Sparse attack: We first consider the simpler case of a sparse (ℓ_0) attack. This is a special case of the block sparse attack with block size is *one*. Numerous such attacks have been proposed in the past [54, 55]. The general idea behind most such attacks is to analyse which pixels in the input image tend to affect the output the most $S(x_i) := \left|\frac{\partial L(f(\mathbf{x},\mathbf{y}))}{\partial x_i}\right|$, where $L(\cdot)$ is the adversarial loss, and c is the true class predicted by the network. The next step is to perturb the top s most salient pixels for a s-sparse attack by using gradient descent to create the least amount of change in the s pixels to adversarially flip the label.

Patchwise token attacks: Instead of inspecting saliency of single pixel we check the norm of gradients of pixels belonging to non-overlapping patches using patch saliency $S(\mathbf{x}_b) := \sqrt{\sum_{x_i \in \boldsymbol{x}_b} \left|\frac{\partial L(f(\mathbf{x},\mathbf{y}))}{\partial x_i}\right|^2}$, for all $b \in \{1, \ldots B\}$. We pick top K blocks according to patch saliency. The effective sparsity is thus $s = K \cdot p$. These sequence of operations are summarized in Alg. 1.

We use non-overlapping patches to understand the effect of manipulating salient tokens instead of arbitrarily choosing patches. In order to further test the robustness of transformers, we also propose to look at the minimum number of patches that would required to be perturbed by an attacker. For this setup, we modify Alg. 1 by linearly searching over the range of 1 to K patches. Fig. 1 show examples of token attacks on transformers.

Algorithm 1 Adversarial Token Attack

Require: $\mathbf{x}_0$:Input image, $f(.)$: Classifier, $\mathbf{y}$: Original label, K: Number of patches to be perturbed, p: Patch size. $i \leftarrow 0$

1: $[b_1 \ldots b_K]$= Top-K of $S(\mathbf{x}_b) = \sqrt{\sum_{x_i \in \boldsymbol{x}_b} \left|\frac{\partial L(f(\mathbf{x},\mathbf{y}))}{\partial x_i}\right|^2}, \forall b.$

2: **while do** $f(\mathbf{x}) \neq y$ OR MaxIter

3: $\quad \mathbf{x}_{b_k} = \mathbf{x}_{b_k} + \nabla_{\mathbf{x}_{b_k}} L; \ \forall \ b_k \in \{b_1, \ldots, b_K\}$

4: **end while**

4. Experiments and Results

Setup: To ensure a fair comparison, we choose the best models for the ImageNet dataset [56] reported in literature.

The models achieve near state-of-the-art results in terms of classification accuracy. They also are all trained using the best possible hyperparameters for each case. We use these weights and the shared models from the `Pytorch Image models` [57] repository. We have done our analysis on 10000 images from the ImageNet validation dataset.

Models: In order to compare the robustness of transformers to CNNs, we consider multiple families of models: Vision Transformers (ViT) [1, 4, 16], Resnets [11, 12], ConvViTs [58], ConvNexts [59], SWIN [17], and CLIP [2]. We note that the vision transformer architectures except SWIN rely on non-overlapping patches as tokens. SWIN uses a shifted window based approach to construct tokens. Note that [1] show that best performing ImageNet models have a fixed input token size of 16×16. We therefore fix a token size of 16×16 for all our models.

In order to ensure that the attacks are equivalent, we ensure that any norm or patch budgets are appropriately scaled as per the pre-processing used. We also scale the ϵ-norm budget for mixed norm attacks to eight gray levels of the input image post normalization. Additionally, we do a hyper parameter search to find the best attacks for each model analysed.

Patch attacks: We allow the attacker a fixed budget of tokens as per Algorithm 1. We use the robust accuracy as the metric of robustness, where a higher value is better. We start with an attack budget of 1 token for an image size of 224×224 for the attacker where each token is a patch of the size 16×16. In order to compensate for the differences in the size of the input, we scale the attack budget for ConvNextv2-Huge by allowing for bigger patch size (24×24 to be precise) to be perturbed. For this setup, we do not enforce any imperceptibility constraints. We run the attack on the fixed subset of ImageNet for the network architectures defined above. Fig. 2 shows the result of our

Figure 1. **Examples of Token attacks.** Token attacks are successful in creating nearly imperceptible perturbations that fool ViTs. The leftmost image in every triplet is an original image, followed by the adversarial image with a single token, and the token perturbation.

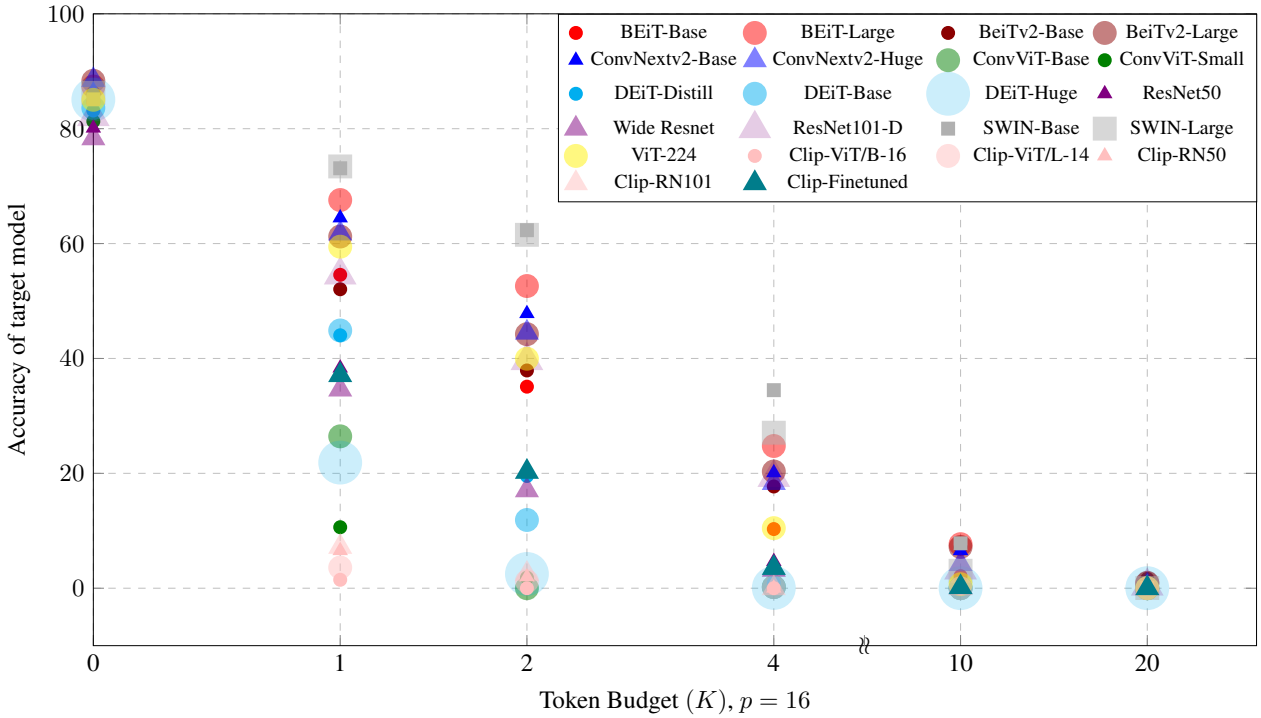

Figure 2. **Robustness to Token Attacks with varying budgets.** $p = 16$**.** Vision transformers are less robust than SWIN, BeIT and convnets for high token budgets, with patch size matching token size of transformer architecture. Detailed results for all models can be found in the appendix.

analysis. Notice that vision transformer architectures are less robust as compared to ResNet-101 and ConvNextv2 models. However, we observe that SWIN, and BeIT reject this trend and are more robust than CNNs for lower token budgets and comparable for higher budgets. We conjecture that this is a consequence of the architectural novelties that SWIN and BeIT use. SWIN, for example, leverages patch-merging - tokens are essentially overlapping patches. BeIT on the other hand, uses a mask-based pretraining approach which inuitively reduces the models dependence on a single patch. We empirically validate this conjecture in the next section by ablating over the amount of patch-overlap between tokens.

Varying the Token budget: We now study the robustness of models by varying the token budget. For this case, we only study attacks for a fixed patch (token) size of 16×16. See Fig. 2 for the results We clearly observe a difference in the behavior of transformer models and convnets here. In

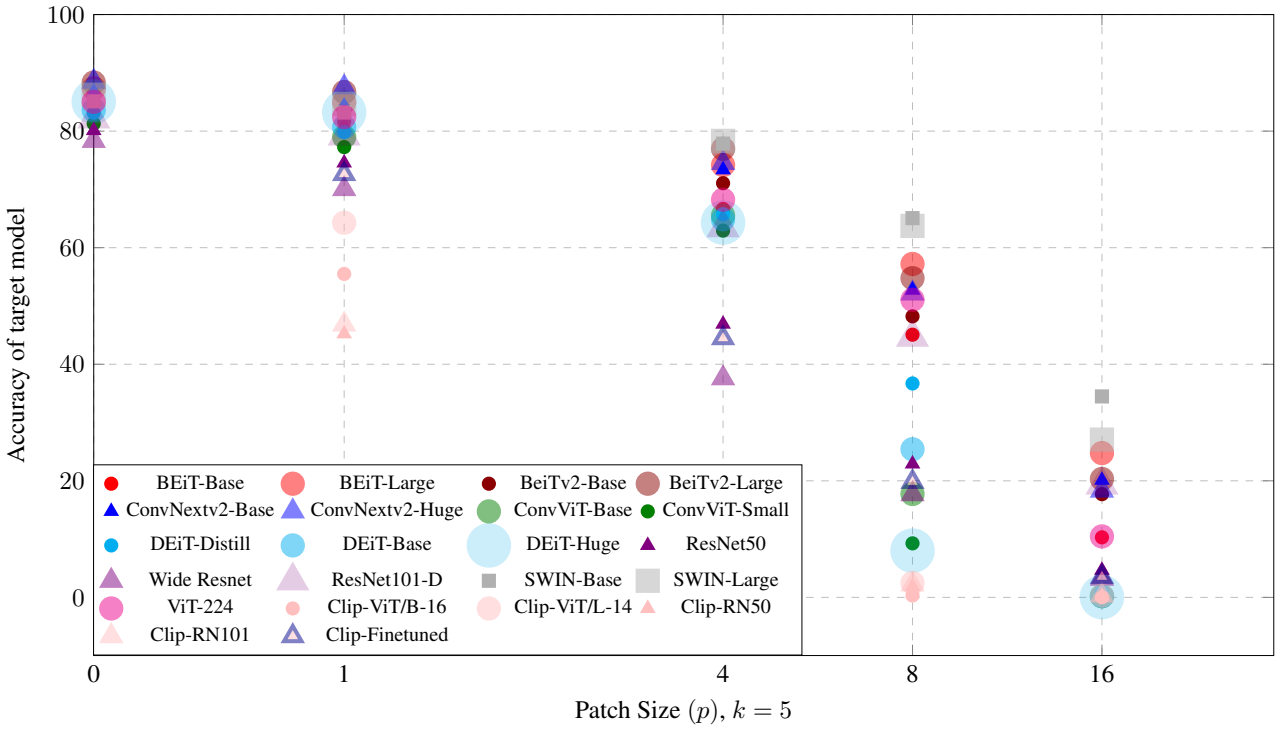

Figure 3. ***Token attacks with varying patch sizes.*** *$K = 5$. When the attack patch size is smaller than token size of architecture, all models except CLIP are comparably robust against patch attacks. However, as the patch attacks appraoch token size, we see significant deterioration in robustness for vision transformers. Detailed results can be found in the Appendix.*

general, for larger token budgets, convnets outperform all other token based models. For smaller token budgets, while transformers are still comparably robust, convnets tend to be more robust that ViT. In addition, the robust accuracies for Transformers fall significantly for as few as *four* tokens. The advantage offered by distillation in DeIT is also lost under token attacks. In addition, a surprising observation is that CLIP models are vulnerable to even single token attacks. This is of particular concern as CLIP embeddings are now used for a variety of downstream tasks. We also analyse fine-tuned CLIP models and observe that while they improve in robustness, the models are still worse than convolutional models.

We also analyse finetuned CLIP models from [60]. We consider two models from their setup: (1) the best performing finetuned model, and (2) the averaged greedy-soup model. We observe that the finetuned models perform better than the zero-shot CLIP models for low token budgets. However, as token budgets increase (> 4 tokens), the robust accuracy drops to nearly zero in both instances.

Varying patch sizes:In order to further analyse if these results hold across stronger and weaker block sparse constraints, we further run attacks for varying patch sizes. Smaller patch sizes are equivalent to partial token manipulation. We fix the token budget to be 5 tokens. Here, this corresponds to allowing the attacker to perturb 5 $p \times p$ patches. See Fig. 3 for the results. As one would expect, a smaller partial token attack is weaker than a full token attack. Surprisingly, the Transformer networks are comparable or better than ResNets and other convnets for attacks smaller than a single token. This leads us to conclude that Transformers can compensate for adversarial perturbations within a tokens. However, as the patch size approaches the token size, SWIN, BeIT and convnets outperform ViTs and ConvViTs. Notice that CLIP follows the same trend as well with CLIP-finetuned models being slightly more robust than the zero-shot classifier.

Ablation Study: Sparse Attacks: We also study the effect of the block-sparsity constraint which forces token level attacks here. The sparse variant of our algorithm restricts the patch size to 1×1. We allow for a sparsity budget of 0.5% of original number of pixels. In case of the standard 224×224 ImageNet image, the attacker is allowed to perturb 256 pixels. We compare the attack success rate of both sparse attack and patch-based token attack at same sparsity budget; to compare we chose $1, 16 \times 16$ patch attack (refer Table 1). Notice that sparse attacks are stronger as compared to token attacks. We see that as is the case with token attacks, even for sparse attacks, vision transformers are less robust as compared to ResNets. With the same sparsity budget,

Table 1. ***Robust accuracies,*** $s = 256$ ***sparse and*** $K = 1$**,** 16×16 ***token attack***

Model	Clean	Sparse	Token
BEiT-Base-224	84.69	29.28	54.46
BEiT-Large-224	87.34	42.60	67.58
BEiTv2-Base-224	86.27	45.16	52.05
BEiTv2-Large-224	88.34	52.03	61.23
ConvNextv2-Base	86.72	44.77	64.47
ConvNextv2-Huge	88.48	59.90	73.28
ConvNextv2-Large	86.89	51.01	65.45
ConvViT-Base	82.18	12.96	26.44
ConvViT-Small	81.28	13.61	10.62
ConvViT-Tiny	73.48	18.35	4.03
DeiT224-Distill	83.16	24.06	44.03
DeiT3-Base-224	83.61	12.51	44.87
DeiT3-Huge-224-14	85.07	6.76	21.87
DeiT3-Large-224	84.62	8.58	55.27
DeiT3-Medium-224	82.86	24.04	46.59
DeiT3-Small-224	81.46	4.14	21.34
ResNet101-D	82.10	33.78	54.53
ResNet50	80.10	9.03	38.33
Wide Resnet	78.33	4.78	34.59
SWIN-224	82.90	48.42	69.11
SWIN-224-Base	85.11	48.65	73.11
SWIN-224-Large	86.24	48.00	73.43
ViT-224	85.03	25.44	59.46

sparse attacks are stronger than token attacks; however we stress that sparse threat model is less practical to implement as the sparse coefficients may be scattered anywhere in the image.

5. Does Patch-Merging help?

Observing that SWIN and ConvNextv2 perform much better, we conjecture that this is because these models reduce the model dependency on single tokens. Primarily, SWIN leverages special attention layers; multi-head self attention modules with regular (W-MSA) and shifted windowing (SW-MSA) configurations. The shifted window self-attention model provides connections across the boundaries of the windows using patch merging, thus reducing dependency on independent tokens. The shift here refers to the number of pixels that overlap between consecutive tokens.

To further analyse this, we trained SWIN transformers with varying shift sizes in the SW-MSA and analysed their robustness to patch attacks.

We find that as the shift size increases from zero, the robustness increases; see Table 2. Further, we note that there is a higher difference between 0-shift to 1-shift compared to others. This clearly shows that reducing the independent, non-overlapping token dependency plays a major role in improving the robustness of the transformers to token attacks.

Table 2. **Robust Accuracy for SWIN models trained with different shift sizes.** Notice that the SWIN model trained with non-overlapping patches is more vulnerable to adversarial token attacks.

Shift (Patch Overlap)	Clean	Patch Size			
		1	4	8	16
0	81.04	78.56	71.62	60.31	33.52
1	82.01	79.75	73.82	64.00	36.18
2	82.02	80.18	75.20	66.80	42.49
3	81.94	79.89	74.37	64.55	38.07

6. Discussion and Conclusion

Analysing the above results, we infer certain interesting properties of transformers.

1. We find that Transformers are generally susceptible to token attacks, even for very low token budgets.
2. However, Transformers appear to compensate for perturbations to patch attacks smaller than the token size.
3. We also observe that pure convolutional models (ResNet, ConvNextv2), SWIN and BeIT are more robust to such token level attacks. Further analysis of SWIN models reveals that using patch-merging helps reduce dependence of the model predictions on a few tokens, and improve robustness through enforcing redundancy.

Acknowledgements

The authors were supported in part by the National Science Foundation under grants CCF-2005804 and CCF-1815101, USDA/NIFA under grant USDA-NIFA:2021-67021-35329, and ARPA-E under grant DE:AR0001215.

References

[1] A. Dosovitskiy, L. Beyer, et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in *ICLR*, 2020. 1, 2, 3

[2] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever, "Learning transferable visual models from natural language supervision," in *ICML*, 2021. 1, 2, 3

[3] Alex Fang, Gabriel Ilharco, Mitchell Wortsman, Yu Wan, Vaishaal Shankar, Achal Dave, and Ludwig Schmidt, "Data

determines distributional robustness in contrastive language image pre-training (clip)," in *ICML*, 2022. 1

[4] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. J'egou, "Training data-efficient image transformers & distillation through attention," in *ICML*, 2021. 1, 2, 3

[5] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," *arXiv preprint arXiv:1312.6199*, 2013. 1, 2

[6] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," in *2017 IEEE Symposium on Security and Privacy*. IEEE, 2017, pp. 39–57. 1

[7] F. Croce and M. Hein, "Sparse and imperceivable adversarial attacks," in *CVPR*, 2019, pp. 4724–4732. 1, 2

[8] S. Bhojanapalli, A. Chakrabarti, D. Glasner, D. Li, T. Unterthiner, and A. Veit, "Understanding robustness of transformers for image classification," *ArXiv*, vol. 2103.14586, 2021. 1, 2

[9] S. Paul and P. Chen, "Vision transformers are robust learners," *arXiv preprint arXiv:2105.07581*, 2021. 1, 2

[10] Ameya Joshi, Gauri Jagatap, and Chinmay Hegde, "Adversarial token attacks on vision transformers," in *CVPR Workshop on Transformers for Vision*, 2022. 1, 2, 3, 9

[11] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *CVPR*, pp. 770–778, 2016. 2, 3

[12] S. Zagoruyko and N. Komodakis, "Wide residual networks," *ArXiv*, vol. 1605.07146, 2016. 2, 3

[13] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *NeurIPS*, 2017. 2

[14] Junnan Li, Dongxu Li, Caiming Xiong, and Steven C. H. Hoi, "Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation," in *International Conference on Machine Learning*, 2022. 2

[15] Robin Rombach, A. Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer, "High-resolution image synthesis with latent diffusion models," *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10674–10685, 2021. 2

[16] Hangbo Bao, Li Dong, and Furu Wei, "Beit: Bert pre-training of image transformers," *ArXiv*, vol. abs/2106.08254, 2021. 2, 3

[17] 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," *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9992–10002, 2021. 2, 3

[18] I. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in *ICLR*, 2015. 2

[19] A. Kurakin, I. Goodfellow, and S. Bengio, "Adversarial examples in the physical world," *arxiv preprint*, vol. 1607.02533, 2017. 2

[20] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," *IEEE (SP)*, 2017. 2

[21] A. Ilyas, L. Engstrom, and A. Madry, "Prior convictions: Black-box adversarial attacks with bandits and priors," *arxiv preprint*, vol. 1807.07978, 2018. 2

[22] A. Ilyas, L. Engstrom, A. Athalye, and J. Lin, "Black-box adversarial attacks with limited queries and information," in *PMLR*, 2018, vol. 80. 2

[23] L. Engstrom, D. Tsipras, L. Schmidt, and A. Madry, "A rotation and a translation suffice: Fooling cnns with simple transformations," *arxiv preprint*, vol. 1712.02779, 2017. 2

[24] C. Xiao, J. Zhu, B. Li, W. He, M. Liu, and D. Song, "Spatially transformed adversarial examples," *arxiv preprint*, vol. 1801.02612, 2018. 2

[25] A. Joshi, A. Mukherjee, S. Sarkar, and C. Hegde, "Semantic adversarial attacks: Parametric transformations that fool deep classifiers," in *ICCV*, 2019. 2

[26] Y. Zhang, H. Foroosh, P. David, and B. Gong, "Camou: Learning physical vehicle camouflages to adversarially attack detectors in the wild," in *ICLR*, 2019. 2

[27] Y. Song, R. Shu, N. Kushman, and S. Ermon, "Constructing unrestricted adversarial examples with generative models," in *NeurIPS*, 2018. 2

[28] A. Shafahi, W R. Huang, M. Najibi, O. Suciu, C. Studer, T. Dumitras, and T. Goldstein, "Poison frogs! targeted clean-label poisoning attacks on neural networks," in *NeurIPS*, 2018. 2

[29] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," in *ICLR*, 2018. 2

[30] H. Zhang, Y. Yu, J. Jiao, E. Xing, L. El Ghaoui, and M. Jordan, "Theoretically principled trade-off between robustness and accuracy," in *ICML*, 2019, pp. 7472–7482. 2

[31] G. Jagatap, A. Joshi, A. Chowdhury, S. Garg, and C. Hegde, "Adversarially robust learning via entropic regularization," *ArXiV*, vol. 2008.12338, 2020. 2

[32] P. Samangouei, M. Kabkab, and R. Chellappa, "Defense-GAN: Protecting classifiers against adversarial attacks using generative models," in *ICLR*, 2018. 2

[33] H. Yin, Z.and Wang, J. Wang, J. Tang, and W. Wang, "Defense against adversarial attacks by low-level image transformations," *International Journal of Intelligent Systems*, vol. 35, no. 10, pp. 1453–1466, 2020. 2

[34] A. Athalye, N. Carlini, and D. Wagner, "Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples," in *ICML*, 2018. 2

[35] E. Wong and Z. Kolter, "Provable defenses against adversarial examples via the convex outer adversarial polytope," in *ICML*. PMLR, 2018. 2

[36] J. Cohen, E. Rosenfeld, and Z. Kolter, "Certified adversarial robustness via randomized smoothing," in *ICML*. PMLR, 2019. 2

[37] H. Salman, G. Yang, J. Li, P. Zhang, H. Zhang, I. Razenshteyn, and S. Bubeck, "Provably robust deep learning via adversarially trained smoothed classifiers," in *NeurIPS*, 2019. 2

[38] T. Brown, D. Mané, A. Roy, M. Abadi, and J. Gilmer, "Adversarial patch," *arXiv preprint arXiv:1712.09665*, 2017. 2

[39] A. Zolfi, M. Kravchik, Y. Elovici, and A. Shabtai, "The translucent patch: A physical and universal attack on object detectors," in *CVPR*, 2021. 2

[40] S. Thys, W. Van Ranst, and T. Goedemé, "Fooling automated surveillance cameras: adversarial patches to attack person detection," in *CVPR Workshops*, 2019. 2

[41] Z. Wu, S. Lim, L. Davis, and T. Goldstein, "Making an invisibility cloak: Real world adversarial attacks on object detectors," in *ECCV*, 2020. 2

[42] F. Croce, M. Andriushchenko, et al., "Sparse-rs: a versatile framework for query-efficient sparse black-box adversarial attacks," *arXiv preprint arXiv:2006.12834*, 2020. 2

[43] D. Hendrycks, X. Liu, E. Wallace, A. Dziedzic, R. Krishnan, and D. Song, "Pretrained transformers improve out-of-distribution robustness," *arXiv preprint arXiv:2004.06100*, 2020. 2

[44] K. Mahmood, R. Mahmood, and M. Van Dijk, "On the robustness of vision transformers to adversarial examples," *arXiv preprint arXiv:2104.02610*, 2021. 2

[45] Muzammal Naseer, Kanchana Ranasinghe, Salman Hameed Khan, Fahad Shahbaz Khan, and Fatih Murat Porikli, "On improving adversarial transferability of vision transformers," *ArXiv*, vol. abs/2106.04169, 2021. 2

[46] Zhipeng Wei, Jingjing Chen, Micah Goldblum, Zuxuan Wu, Tom Goldstein, and Yu-Gang Jiang, "Towards transferable adversarial attacks on vision transformers," *ArXiv*, vol. abs/2109.04176, 2021. 2

[47] Yao Qin, Chiyuan Zhang, Ting Chen, Balaji Lakshminarayanan, Alex Beutel, and Xuezhi Wang, "Understanding and improving robustness of vision transformers through patch-based negative augmentation," *ArXiv*, vol. abs/2110.07858, 2021. 2

[48] Hadi Salman, Saachi Jain, Eric Wong, and Aleksander Mkadry, "Certified patch robustness via smoothed vision transformers," *ArXiv*, vol. abs/2110.07719, 2021. 2

[49] Giulio Lovisotto, Nicole Finnie, Mauricio Munoz, Chaithanya Kumar Mummadi, and Jan Hendrik Metzen, "Give me your attention: Dot-product attention considered harmful for adversarial patch robustness," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022, pp. 15234–15243. 2

[50] Jindong Gu, Volker Tresp, and Yao Qin, "Are vision transformers robust to patch perturbations?," in *European Conference on Computer Vision*, 2021. 2

[51] Y. Fu, Shunyao Zhang, Shan-Hung Wu, Cheng Wan, and Yingyan Lin, "Patch-fool: Are vision transformers always robust against adversarial perturbations?," *ArXiv*, vol. abs/2203.08392, 2022. 2

[52] Zeyu Wang, Yutong Bai, Yuyin Zhou, and Cihang Xie, "Can cnns be more robust than transformers?," *ArXiv*, vol. abs/2206.03452, 2022. 2

[53] Francesco Croce and Matthias Hein, "On the interplay of adversarial robustness and architecture components: patches, convolution and attention," *ArXiv*, vol. abs/2209.06953, 2022. 2

[54] N. Papernot, P. Mcdaniel, S. Jha, M. Fredrikson, Z. B. Celik, and A. Swami, "The limitations of deep learning in adversarial settings," *EuroS&P*, pp. 372–387, 2016. 3, 9

[55] R. Wiyatno and A. Xu, "Maximal jacobian-based saliency map attack," *ArXiv*, vol. 1808.07945, 2018. 3, 9

[56] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. Berg, and F. Li, "ImageNet Large Scale Visual Recognition Challenge," *Intl. J. Comp. Vision*, vol. 115, no. 3, pp. 211–252, 2015. 3

[57] R. Wightman, "Pytorch image models," `https://github.com/rwightman/pytorch-image-models`, 2019. 3

[58] Haiping Wu, Bin Xiao, Noel C. F. Codella, Mengchen Liu, Xiyang Dai, Lu Yuan, and Lei Zhang, "Cvt: Introducing convolutions to vision transformers," in *ICCV*, 2021. 3

[59] Zhuang Liu, Hanzi Mao, Chaozheng Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie, "A convnet for the 2020s," in *CVPR*, 2022. 3

[60] Mitchell Wortsman, Gabriel Ilharco, Samir Yitzhak Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S. Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, and Ludwig Schmidt, "Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time," in *International Conference on Machine Learning*, 2022. 5

[61] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," *arXiv preprint arXiv:1312.6034*, 2013. 9