Adversarial Token Attacks on Vision Transformers

Ameya Joshi
Dept. of ECE
New York University
ameya.joshi@nyu.edu

Gauri Jagatap
Dept. of ECE
New York University
gauri.jagatap@nyu.edu

Chinmay Hegde
Dept. of CSE
New York University
chinmay.h@nyu.edu

Abstract

Vision transformers rely on a patch token based self attention mechanism, in contrast to convolutional networks. We investigate fundamental differences between these two families of models, by designing a block sparsity based adversarial token attack. We probe and analyze transformer as well as convolutional models with token attacks of varying patch sizes. We infer that transformer models are more sensitive to token attacks than convolutional models, with ResNets outperforming Transformer models by up to $\sim 30\%$ *in robust accuracy for single token attacks.*

1. Introduction

Motivation: Convolutional networks (CNNs) have shown near human performance in image classification [1] over non-structured dense networks. However, CNNs are vulnerable to specifically designed adversarial attacks [2]. Several papers in adversarial machine learning literature reveal the brittleness of convolutional networks to adversarial examples. For example, gradient based methods [3, 4] design a perturbation by taking steps proportional to the gradient of the loss of the input image x in a given ℓ_p neighborhood. This has led to refined robust training approaches, or defenses, which train the network to see adversarial examples during the training stage and produce the unaltered label corresponding to it [5, 6].

Vision transformers (ViT) were introduced [7], as a network architecture inspired by transformers [8] which have been successfully used for modeling language data. ViTs rely on self attention [8], a mechanism that allows the network to find correlations between spatially separated parts of the input data. In the context of vision, these are small non-overlapping *patches* which serve as *tokens* to the transformer. ViTs and more recently distillation based Data Efficient Image Transformers (DeIT) [9] have shown to have competitive performance on classification tasks and rely on pre-training on very large datasets. It is of imminent interest to therefore study the robustness of self-attention networks.

There has been some work on adversarial robustness of vision transformers. [10] show that under certain regimes, vision transformers are at least as robust to ℓ_2 and ℓ_∞ PGD attacks as ResNets. While ℓ_2 and ℓ_∞ threat models are useful in understanding fundamental properties of deep networks, they are not realizable in the real world and do not capture actual threats. Transformer based networks also introduce the need for tokenizing the image, leading to an encoded bias in the input. We, therefore propose to analyse the sensitivity of the architecture to token level changes rather than to the full image.

Specifically, we attempt to answer: *Are transformers robust to perturbations to a subset of the input tokens?* We present a systemic approach to answer this query by constructing token level attacks by leveraging block sparsity.

For this analysis, we point out some important vulnerabilities in vision transformers by constructing a token level attack. As ViT models use patches as tokens, this leads us to a natural "block-sparse" attack construction.

Our contributions: We propose a patch based block sparse attack where the attack budget is defined by the number of tokens the attacker is allowed to perturb. We identify top salient pixels using the magnitude of their loss gradients and perturb them to create attacks. We extend a similar idea to block sparsity by constraining salient pixels to lie in non-overlapping patches. We probe three families of neural architectures using our token attack; self-attention (ViT [7], DeIT [9]), convolutional (Resnets [11] and WideResNet [12]) and MLP based (MLP Mixer [13]).

We make the following contributions and observations:

1. We propose a new attack which imposes block sparsity constraints, allowing for *token attacks* for Transformers.
2. We show classification performance of all architectures on token attacks of varying patch sizes and number of patches.
3. We demonstrate that for token attacks accounting for the architecture and token size, vision transformers are less resilient to token attacks as compared to MLP Mixers and ResNets.
4. For token attacks smaller than architecture token size,

vision transformers are comparably robust to ResNets.

5. We also specifically note the shortcomings of previous studies on robustness of transformers [10], where ViTs are shown to be more robust than ResNets.
6. With our token attacks we can break Vision transformers using only 1% of pixels as opposed to ℓ_2 or ℓ_∞ attacks which rely on perturbing all image pixels.

Related work: ***Threat models:*** Deep networks are vulnerable to imperceptible changes to input images as defined by the ℓ_∞ distance [14]. There exist several test-time attack algorithms with various threat models: ℓ_p constrained [2, 4, 15], black-box [16, 17], geometric attacks [18, 19], semantic and meaningful attacks [20–22] and data poisoning based [23].

Defenses: Due to the vast variety of attacks, adversarial defense is a non-trivial problem. Empirical defenses as proposed by [5], [6], and [24] rely on adversarial data augmentation and modified loss functions to improve robustness. [25, 26] propose preprocessing operations as defenses. However, such defenses fail to counter adaptive attacks [27]. [28], [29] and [30] provide methods that guarantee robustness theoretically.

Patch attacks: Patch attacks [31] are practically realizable threat model. [32–34] have successfully attacked detectors and classifiers with physically printed patches. In addition, [35, 35] 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: [10, 36] 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. [37] show that adversarial examples do not transfer well between CNNs and transformers, and build an ensemble based approach towards adversarial defense. [38] claims that Transformers are robust to a large variety of corruptions due to attention mechanism. However, these works consider global perturbations only. Vision transformers on the other hand, have a natural inductive bias with patches. [39] show that ViTs are specifically vulnerable to patch-level transformations, leading to good in-distribution accuracies but poor out-of-distribution performance. [40] present a certified defense for patch attacks, where in ViTs outperform Resnets. This points to a clear correlation between robustness and patch (token) perturbations. We study this in greater detail, specifically restricting ourselves to token-level attacks in order to analyse this phenomenon in greater detail.

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. $i \leftarrow 0$

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

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

  $\mathbf{x}_{b_k} = \mathbf{x}_{b_k} + \nabla_{\mathbf{x}_{b_k}} L; \ \forall \ b_k \in \{b_1, \dots, b_K\}$

  $\mathbf{x}_{b_k} = Project_{\epsilon_\infty}(\mathbf{x}_{b_k})$ (optional)

end while

2. Token Attacks on Vision transformers

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 $[\boldsymbol{x}_1, \dots \boldsymbol{x}_b, \dots, \boldsymbol{x}_B]$, 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 two varying attack budgets. In both cases we consider a block sparse token budget, where we restrict the attacker to modifying K patches or "tokens" (1) with an unconstrained perturbation allowed per patch (2) a "mixed norm" block sparse budget, where the pixelwise perturbation for each token is restricted to an ℓ_∞ ball with radius ϵ defined as K, ϵ-attack.

Sparse attack: To begin, 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 [41, 42]. 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 $b \in \{1, \dots B\}$. 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.

Mixed-norm attacks: Most approaches [35, 43] additionally rely on a mixed ℓ_2-norm based sparse attack in order to generate imperceptible perturbations. Motivated by this setting, we propose a mixed-norm version of our modified attack as well. In order to ensure that our block sparse attacks are imperceptible, we enforce an additional ℓ_∞ projection step post the gradient ascent step. This is enforced via Step 4 in Alg. 1.

3. Experiments and Results

Setup: To ensure a fair comparison, we choose the best models for the Imagenet dataset [44] reported in [7], [9] and [12]. 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` [45] repository. We restrict our analysis to a fixed subset of 300 randomly chosen images from the Imagenet validation dataset.

Models: In order to compare the robustness of transformer models to standard CNNs, we consider three different families of architectures:(1) Vision Transformer (ViT) [7], Distilled Vision Transformers (DeIT) [9], (2) Resnets [11, 12] and (3) MLP Mixer [13]. For transformers, [7] show that best performing Imagenet models have a fixed input token size of 16×16. To ensure that the attacks are fair, we scale the norm or patch budgets appropriatelyas per the pre-processing used [1]. 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 ViT-384 by allowing for more patches (3 to be precise) to be perturbed. However, we do not enforce any imperceptibility constraints. We run the attack on the fixed subset of ImageNet for the network architectures defined above. Fig. 1(a) shows the result of our analysis. Notice that Transformer architectures are more vulnerable to token attacks as compared to ResNets and MLP-Mixer. Further, ViT-384 proves to be the most vulnerable, and ResNet-101 is the most robust model. DeiT which uses a teacher-student network is more robust than ViTs. We therefore conclude that distillation improves robustness to single token attacks.

Varying the Token budget: For this experiment, we start with a block-budget of 1 patch, and iterate upto 40 patches to find the minimum number of tokens required to break an image. We then measure the robust accuracy for each constraint and for each model. For this case, we only study attacks for a fixed patch (token) size of 16×16 and represent our findings in Fig. 1(a). We clearly observe a difference in the behavior of ViT versus ResNets here. In general, for a given token budget, ResNets outperform all other token based models. In addition, the robust accuracies for Transformers fall to zero for as few as *two* patches. The advantage offered by distillation for single token attacks is also lost once the token budget increases.

Varying patch sizes: We also study our attacks for varying patch sizes. Smaller patch sizes are equivalent to partial token manipulation. We fix the token budget to be 5 or 15 tokens as dictated by the input size. Here, this corresponds to allowing the attacker to perturb 5 $p \times p$ patches. Note that a smaller partial token attack is weaker than a full token attack. Surprisingly, the Transformer networks are comparable or better than ResNets for attacks smaller than a single token. This leads us to conclude that Transformers compensate for adversarial perturbations within tokens. However, as the patch size approaches the token size, Resnets achieve better robustness. We also see that MLP-Mixers, while also using the token based input scheme, perform better than Transformers as the patch attack size increases.

However, this approach allows for unrestricted changes to the tokens. Another approach would be to study the effect of "mixed norm" attacks which further constrain the patches to be *imperceptibly* perturbed.

Mixed Norm Attacks: For the mixed norm attacks, we analyse the robustness of all networks for a fixed ϵ ℓ_∞ budget of one gray level. We vary the token budgets from 1 to 5. Here, almost all the networks show similar robustness for a small token budget (K=1,2); refer Table 1. However, as the token budget increases, Transformer and MLP Mixer networks are far more vulnerable. Note that this behavior contradicts [10], where ViTs outperform ResNets. Since our threat model leverages the token based architecture of the Transformers, our attacks are more successful at breaking ViTs over Resnets.

Ablation Study: Saliency v/s Random Selection. We also analyse the efficacy of using the saliency metric to select vulnerable patches. To compare, we randomly select 1, 2 or 5 tokens and run steps 2-4 from Alg. 1 with an ℓ_∞ constraint of $8/255$. Fig. 2 shows the results of the experiment. Our saliency based block-sparse attacks outperforms random sampling and is able to reduce accuracies of all vision transformer models for lower token budgets. This demonstrates the necessity of using a saliency based metric to select tokens for attack.

[1] In case of varying image sizes due to pre-processing, we calculate the scaling factor in terms of the number of pixels and appropriately increase or decrease the maximum number of patches.

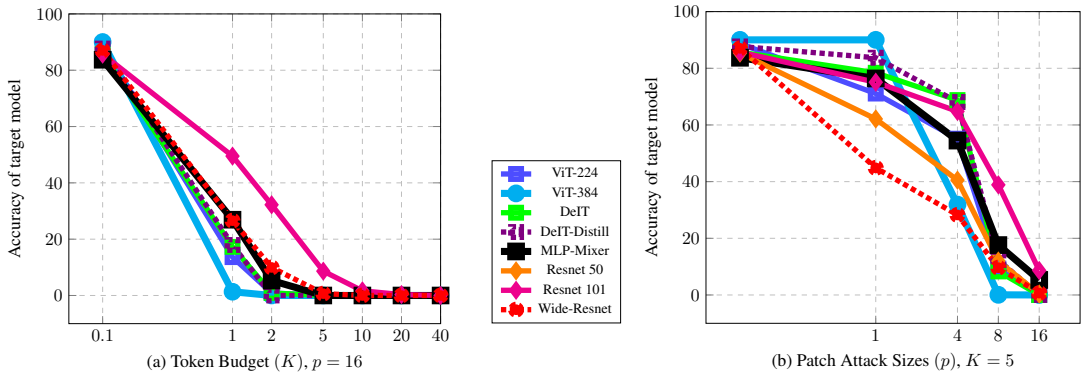


Figure 1. (a) ***Robustness to Token Attacks with varying budgets*** *($p = 16$). Vision transformers are less robust than MLP Mixer and ResNets against patch attacks with patch size matching token size of transformer architecture, (b)* ***Token attacks with varying patch sizes.*** *$K = 5$ When the attack patch size is smaller than token size of architecture, vision transformers are comparably robust against patch attacks, to MLP and ResNets.*

Table 1. *Robust Accuracy for Mixed Norm Attacks:* The models are attacked with a $K, (1/255)$ Patch Attack. Note that for smaller token budgets, the models perform nearly the same. However, as the token budget increases, Resnets are more robust than Transformers.

Model	Clean	Token Budget		
		1	2	5
ViT-224	88.70	68.77	50.83	15.28
ViT-384	**90.03**	*53.48*	*28.57*	*4.98*
DeIT	85.71	**72.42**	46.84	6.31
DeIT-Distilled	87.70	68.77	54.15	16.61
Resnet-101	85.71	69.10	55.14	**32.89**
Resnet-50	85.38	67.44	**55.81**	31.22
Wide Resnet	87.04	54.81	32.89	11.62
MLP-Mixer	*83.78*	63.78	37.87	5.98

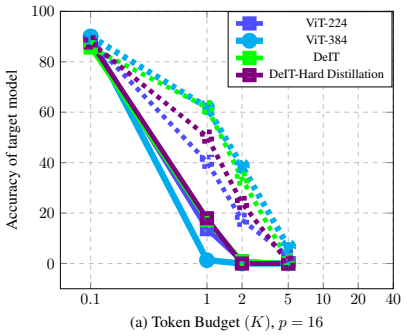


Figure 2. ***Saliency based Token sampling v/s Random Sampling****: The solid lines represent robust accuracies for our token attack whereas dotted lines show the same for randomly sampled tokens. Notice that our saliency based token attack is more successful at constructing attacks with fewer tokens compared to random sampling.*

Sparse Attacks: 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 2). 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, 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.

Table 2. ***Robust accuracies,*** *$s = 256$* ***sparse and*** *$K = 1$,* *16×16* ***patch attack .***

Model	Norm constraint		
	Clean	Sparse	Patch
ViT 224	88.70	5.98	13.62
ViT 384	**90.03**	3.32	*1.33*
DeIT	85.71	4.65	17.27
DeIT (Distilled)	87.70	14.95	17.94
MLP Mixer	*83.72*	5.98	26.91
ResNet 50	85.38	13.95	19.90
ResNet 101	85.71	**23.59**	**49.50**
Wide Resnet	87.04	*1.33*	26.57

4. 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. Further, ResNets and MLP-Mixer outperform Trans-

formers for token attacks consistently.

We aim to further propose strong, certifiable defenses for token attacks. Further directions of research also include analysis of the effect of distillation and semi-supervised pre-training.

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. Kolesnikov, L. Beyer, X. Zhai, J. Puigcerver, J. Yung, S. Gelly, and N. Houlsby, "Big transfer (bit): General visual representation learning," in *ECCV*, 2020. 1, 7

[2] I. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in *ICLR*, 2015. 1, 2

[3] I. Goodfellow, "Defense against the dark arts: An overview of adversarial example security research and future research directions," *arxiv preprint*, vol. 1806.04169, 2018. 1

[4] A. Kurakin, I. Goodfellow, and S. Bengio, "Adversarial examples in the physical world," *arxiv preprint*, vol. 1607.02533, 2017. 1, 2

[5] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," in *ICLR*, 2018. 1, 2, 7

[6] 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. 1, 2, 7

[7] A. Dosovitskiy, L. Beyer, et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in *ICLR*, 2020. 1, 3, 7

[8] 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. 1, 7

[9] 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, 3, 7

[10] 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, 3, 8

[11] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *CVPR*, pp. 770–778, 2016. 1, 3, 8

[12] S. Zagoruyko and N. Komodakis, "Wide residual networks," *ArXiv*, vol. 1605.07146, 2016. 1, 3

[13] I. Tolstikhin, N. Houlsby, et al., "Mlp-mixer: An all-mlp architecture for vision," *ArXiv*, vol. 2105.01601, 2021. 1, 3, 7

[14] 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. 2

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

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

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

[18] 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

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

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

[21] 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

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

[23] 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

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

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

[26] 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

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

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

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

[30] 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

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

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

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

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

[35] F. Croce and M. Hein, "Sparse and imperceivable adversarial attacks," in *CVPR*, 2019, pp. 4724–4732. 2, 3

[36] 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

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

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

[39] 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

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

[41] 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. 2, 8

[42] R. Wiyatno and A. Xu, "Maximal jacobian-based saliency map attack," *ArXiv*, vol. 1808.07945, 2018. 2, 8

[43] 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. 3

[44] 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

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

[46] L. Rice, E. Wong, and Z. Kolter, "Overfitting in adversarially robust deep learning," in *ICML*. PMLR, 2020, pp. 8093–8104. 7

[47] Y. Wang, D. Zou, J. Yi, J. Bailey, X. Ma, and Q. Gu, "Improving adversarial robustness requires revisiting misclassified examples," in *ICLR*, 2019. 7

[48] S. Gowal, C. Qin, J. Uesato, T. Mann, and P. Kohli, "Uncovering the limits of adversarial training against norm-bounded adversarial examples," *ArXiv*, vol. 2010.03593, 2020. 7

[49] T. Pang, X. Yang, Y. Dong, H. Su, and J. Zhu, "Bag of tricks for adversarial training," in *ICLR*, 2021. 7

[50] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *ICML*, 2019. 7

[51] Qizhe Xie, Eduard H. Hovy, Minh-Thang Luong, and Quoc V. Le, "Self-training with noisy student improves imagenet classification," *CVPR*, 2020. 7

[52] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *NAACL*, 2019. 7

[53] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf, "Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter," *ArXiv*, vol. 1910.01108, 2019. 7

[54] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. J. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A.Radford, I. Sutskever, and D. Amodei, "Language models are few-shot learners," *ArXiv*, vol. 2005.14165, 2020. 7

[55] N. Parmar, A. Vaswani, J. Uszkoreit, L. Kaiser, N. Shazeer, A. Ku, and D. Tran, "Image transformer," in *ICML*, 2018, pp. 4055–4064. 7

[56] Z. Dai, H. Liu, Q. Le, and M. Tan, "Coatnet: Marrying convolution and attention for all data sizes," *ArXiv*, vol. 2106.04803, 2021. 7

[57] H. Wu, B. Xiao, N. Codella, M. Liu, X. Dai, L. Yuan, and L. Zhang, "Cvt: Introducing convolutions to vision transformers," *ArXiv*, vol. 2103.15808, 2021. 7

[58] H. Touvron, M. Cord, A. Sablayrolles, G. Synnaeve, and H. J'egou, "Going deeper with image transformers," *ArXiv*, vol. 2103.17239, 2021. 7

[59] C. Yun, S. Sra, and A. Jadbabaie, "Are deep resnets provably better than linear predictors?," in *NeurIPS*, 2019. 8

[60] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," *arXiv preprint arXiv:1312.6034*, 2013. 8

A. Related work

A.1. Defense models

The state of art defense approaches include solving a saddle point min-max optimization via PGD [5] with early stopping [46], TRADES [6] which designs a robust loss consisting of natural loss and boundary loss and MART [47] which leverages a modified loss that considers misclassified examples. Apart from algorithmic approaches, newer papers discuss optimal hyper-parameter tuning as well as combination of regularizers from aformentioned techniques, choice of activation functions, choice of architecture and data augmentation to extract best possible robust accuracies using pre-existing algorithms [48, 49].

A.2. Vision transformers

Vision transformers were recently introduced [7], as a new network architecture inspired by transformers [8] which have been successfully used for modeling language data. Transformers rely on self attention [8], a mechanism that allows the network to find correlations between potentially spatially different parts of the input data. In the context of language, this has to do with different tokens from the input text data. For images, vision transformer breaks down images into smaller patches. Each patch therefore serves as a token to the vision transformer. The position of each patch is also fed to the vision transformer via a positional embedding. Vision transformers have been shown to have competitive performance on classification tasks, at par with the state of art Neural Architecture Search based EfficientNet [50] and rely on pertaining to very large datasets.

A.3. Transformers and Vision Transformers

While convolutional networks have successfully achieved near human accuracy on massive datasets [1, 51], there has been a surge of interest in leveraging self-attention as an alternative approach. Transformers [8] have been shown to be extremely successful at language tasks [52–54]. [55] extend this for image data, where in they use pixels as tokens. While they some success in generative tasks, the models had a large number of parameters and did not scale well. [7] improve upon this by instead using non-overlapping patches as tokens and show state of the art classification performance on the ImageNet dataset. [9] further leverage knowledge distillation to improve efficiency and performance. Further improvements have been suggested by [56], [57] and [58] to improve performance using architectural modifications, deeper networks and better training methods. In parallel, [13] instead propose a pure MLP based architecture that achieves nearly equivalent results with faster training time. However, studies on generalization and robust performance of such networks is still limited. We discuss a few recent works below.

A.4. Transformers

The Transformer block was introduced by [8], for text input. The basic idea of the Transformer model is to leverage an efficient form of "self-attention". A standard attention block is formally defined as,

$$\mathbf{x}_{out} = \text{Softmax}\left(\frac{\mathbf{x}\mathbf{W}_Q\mathbf{W}_k\mathbf{x}^T}{\sqrt{d}}\right)\mathbf{x}\mathbf{W}_V, \tag{1}$$

where $\mathbf{x} \in \mathbb{R}^{d\times n}$ is an input string, $\mathbf{x}_{out} \in \mathbb{R}^{d\times n}$ is the output of the self-attention block, $\mathbf{W}_Q$, $\mathbf{W}_K$ and$\mathbf{W}_V$ are the learnable *query*, *key* and the *value* matrices. Note that $\mathbf{x}$ is actually a concatenation of n "tokens" of size d, which each represent some part of the input. *Multi-headed self attention* stacks multiple such blocks in a single layer. The Transformer model has multiple such layers followed by a final output attention layer with a *classification token*. This architecture makes perfect sense for text where-in tokens are word or sentence embeddings, and each token therefore holds some semantic meaning. These models are trained in an auto-regressive fashion with additional losses for downstream tasks.

However, extending the same architecture for images is non-trivial; primarily as the atomic components of an image are pixels which hold little to no meaning by themselves. [55] propose a solution where they use pixels as tokens and train generative models to solve problems such as image generation and super-resolution. However, the large dimensionality of images forces the Attention blocks to be massively parameterized, leading to issues of scale. In order to remedy this, [7] suggest using local image patches as tokens. This instantly reduces the number of tokens while also leveraging the local consistency property of images. They find that in most cases, it is enough to use non-overlapping patches of 16×16 as tokens to ensure near state of the art accuracies. One disadvantage of such massive models however is the requirement of very large training datasets. [9] propose a data-efficient distillation based method to train Transformers. Their architecture (DeIT) leverages a custom transformer based distillation token as well as standard student-teacher training approaches to improve both the sample complexity and the performance over Vision Transformers.

A standard Resnet model, on the other hand, uses residual blocks:

$$\mathbf{x}_{out} = \text{ReLU}\left(\mathbf{x} + \text{ReLU}(\mathbf{W}\mathbf{x})\right). \tag{2}$$

A Resnet stacks several such residual blocks in succession followed by a classifier. The residual connection allows for easy gradient flow and improves training. There have been several works that prove the generalization and efficacy of Resnets, both empirically [11] and theoretically [59].

A.5. How resnets differ from transformers

In comparison with Resnets, which were the best performing image classifiers previously, we see that there are two major structural differences. The first is that most Resnets downsample activations as we go deeper. This is supposed to help reduce redundancies and propagate discriminative features. However, Vision Transformers with self-attention blocks appear to preserve activation sizes throughout their depth. The second major difference is the structure of the Resnet block in comparison with the Attention block. As is evident, any interaction between non-local pixel groups in Resnets is happens in deeper layers. The initial layers tend to just focus on neighbourhood pixel interactions. However, the Attention mechanism forces each layer of the transformer to consider both local and non-local interactions. There exist additional differences in terms of the non-linearities involved and the number of parameters in each model.

The specific difference in the treatment of local and non-local pixel groups informs our choice of attack. While several papers have previously studied the robustness of vision transformers in the standard adversarial setting, we specifically consider the case where the attacker is only allowed to modify an image locally; for example a set number of tokens.

A.6. Saliency attacks

Such 'salient' pixels are often identified using the magnitudes of gradients. This idea, while not particularly new [60], lends itself naturally to constructing adversarial attacks. Specifically, the idea is to only perturb a subset of the salient pixels thus implicitly satisfying the sparsity constraint. JSMA [41] and Maximal-JSMA [42] leverage this observation to construct k-sparse attacks by maximally perturbing k salient pixels. In maximal-JSMA, the authors calculate saliency of each pixel usign the following equation;

$$S^{+}(x_{i,c}) = \begin{cases} 0 \text{ if } \frac{\partial f(\mathbf{x})_c}{\partial x_i} < 0 \text{ or } \sum_{c' \neq c} \frac{\partial f(\mathbf{x})'_c}{\partial x_i} \\ -\frac{\partial f(\mathbf{x})_c}{\partial x_i} \cdot \sum_{c' \neq c} \frac{\partial f(\mathbf{x})'_c}{\partial x_i} \text{ otherwise,} \end{cases} \tag{3}$$

where x_i is the pixel in question, c is the true class, and f_i is a logit value specific to class i.

In this paper, we propose a patch based block sparse attack where the attack budget is defined by the number of patches (blocks) the attacker is allowed to perturb. Our approach builds on JSMA [41] Maximal-JSMA [42], wherein the attacker identifies top salient pixels using gradients and perturb them to create attacks. We extend a similar idea to block sparsity. The main differences between JSMA and our approach lie in two places: (1) We use a simplified construction for the saliency map that relies on the magnitude of the gradients with respect to each pixel, (2) instead of considering salient pixels, we instead identify the most informative pixel blocks and further rely on gradient updates to generate an attack.

B. Experiments

For all experiments, we use SGD for optimization with a learning rate of 0.1 for a maximum of 100 steps for both variants.

B.1. Mixed norm attacks

For mixed norm block sparse attacks, we impose an additional ℓ_∞ bound (ϵ) on each pixel to enforce imperceptibility. We run our experiments with a constraint of one gray level similar to [10]. Since each of these models scales the input images to varying input ranges, we further scale each ϵ appropriately. We then use a projection step in Alg. 1 using clipping to enforce the constraint.

C. Detailed Results

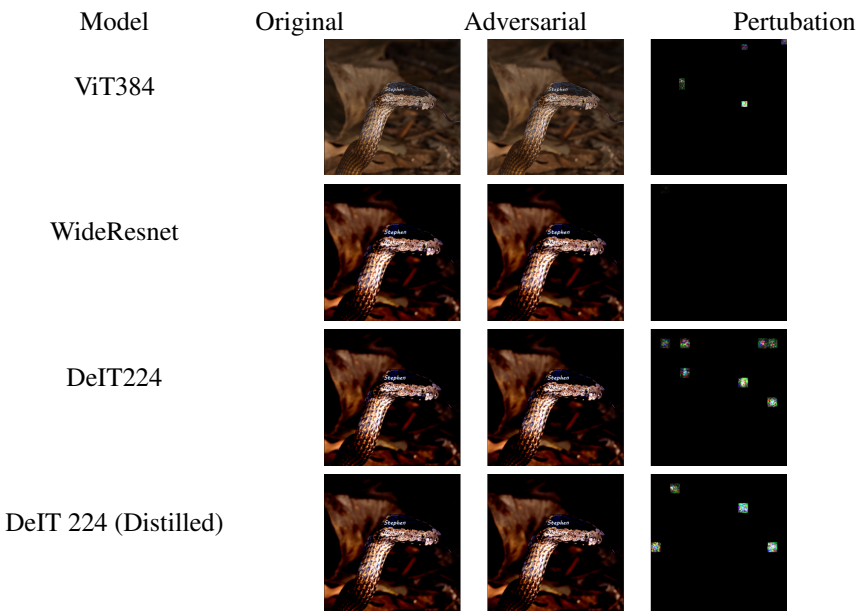


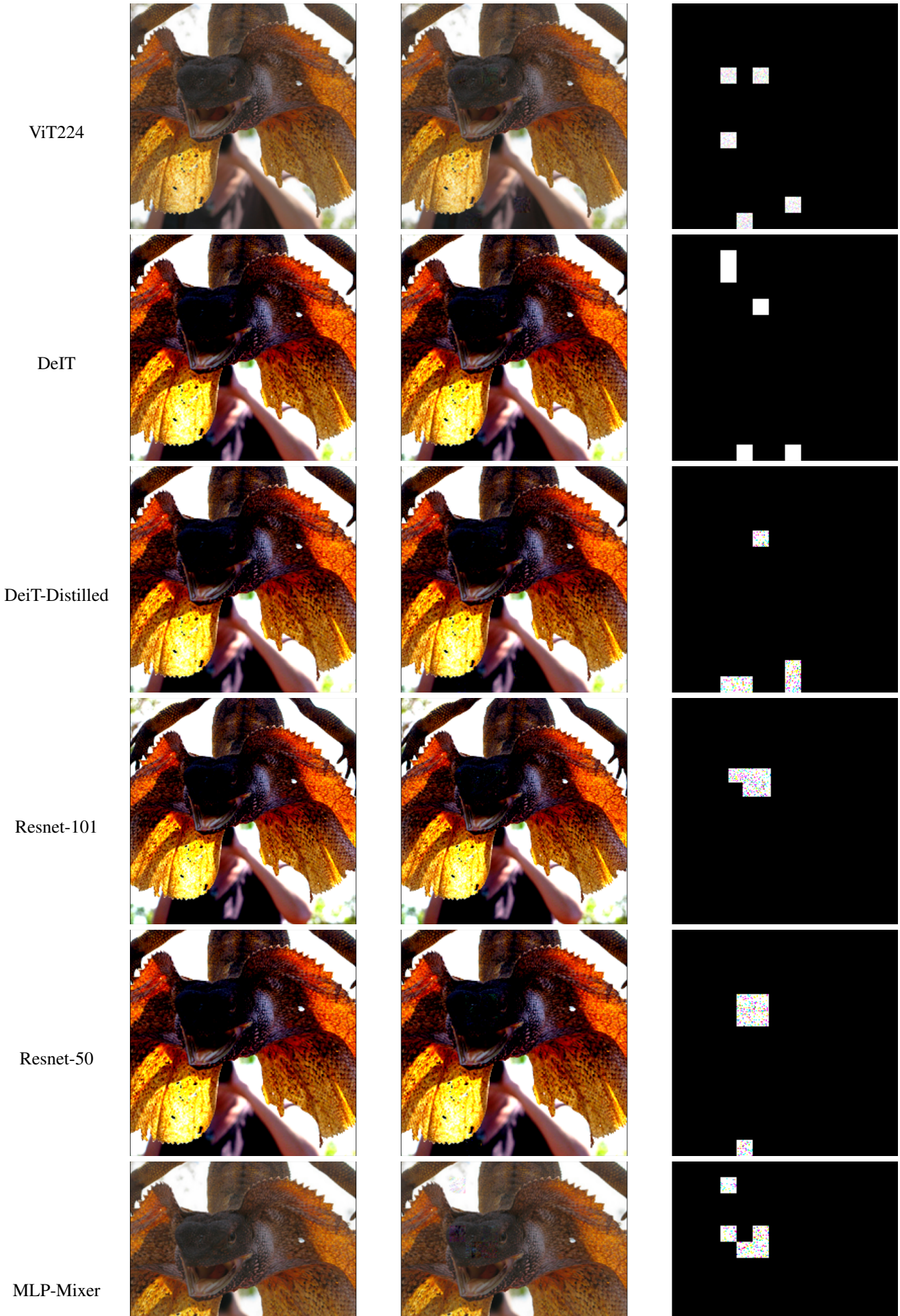
Figure 3. ***Patch attacks on Transformers****: The attack images are generated with a fixed budget of* 20 *patches. Note that the perturbations are imperceptible. The third column shows the perturbation brightened* 10 *times.*

Table 3. Robustness v/s Token Budget

Model	Token Budget					
	1	2	5	10	20	40
ViT-224	13.62	0.9	0.0	0.0	0.0	0.0
ViT-384	1.33	0.0	0.0	0.0	0.0	0.0
DeIT	17.27	0.9	0.0	0.0	0.0	0.0
DeIT (Distilled)	17.94	0.0	0.0	0.0	0.0	0.0
Resnet-101	49.50	32.22	8.64	1.66	0.33	0.0
Resnet-50	19.9	4.65	0.33	0.0	0.0	0.0
Wide-Resnet	26.57	9.96	0.66	0.0	0.0	0.0
MLP-Mixer	26.91	5.31	0.0	0.0	0.0	0.0

Table 4. **Robustness v/s varying patch sizes**

Model	Attack patch sizes			
	1	4	8	16
ViT-224	71.09	55.15	9.30	0.0
ViT-384	68.77	31.89	0.06	0.0
DeIT	78.40	68.77	8.31	0.0
DeIT-Distilled	83.72	68.10	12.29	0.0
Resnet-101d	75.08	64.78	38.87	8.64
Resnet-50	62.12	40.53	11.96	0.33
Wide Resnet	44.85	28.24	9.63	0.66
MLP-Mixer	76.41	54.49	17.61	5.31


ViT224
DeIT
DeiT-Distilled
Resnet-101
Resnet-50
MLP-Mixer