Adversarial Token Attacks on Vision Transformers

NYU TANDON SCHOOL OF ENGINEERING

Ameya Joshi, Gauri Jagatap, and Chinmay Hegde

New York University

Introduction

- Vision transformers use a patch token based self-attention mechanism unlike CNNs.
- However, this may lead to specific vulnerabilities to token-level attacks.
- Token level attacks = Block-sparse constraint on attacks.
- We probe and analyze effect of token attacks on Vision transformers and CNNs.

Contributions

- New block-sparsity constraint based token attack
- Token attacks leverage saliency to implement block sparsity
- Effect of token attacks on flavors of ViTs and CNNs
- **ViTs** are **less robust** than CNNs!

Setup

Dataset: Imagenet
Models: Pretrained

- ViT-(224, 384)
- DeIT (hard & soft distillation)
- MLP-Mixer
- Resnets (50, 101, Wide)

Token Attacks

Saliency:

$$S(x_b) := \sqrt{\sum_{x_i \in x_b} \left| \frac{\partial L(f(x, y))}{\partial x_i} \right|^2}$$

Adversarial Token Attack

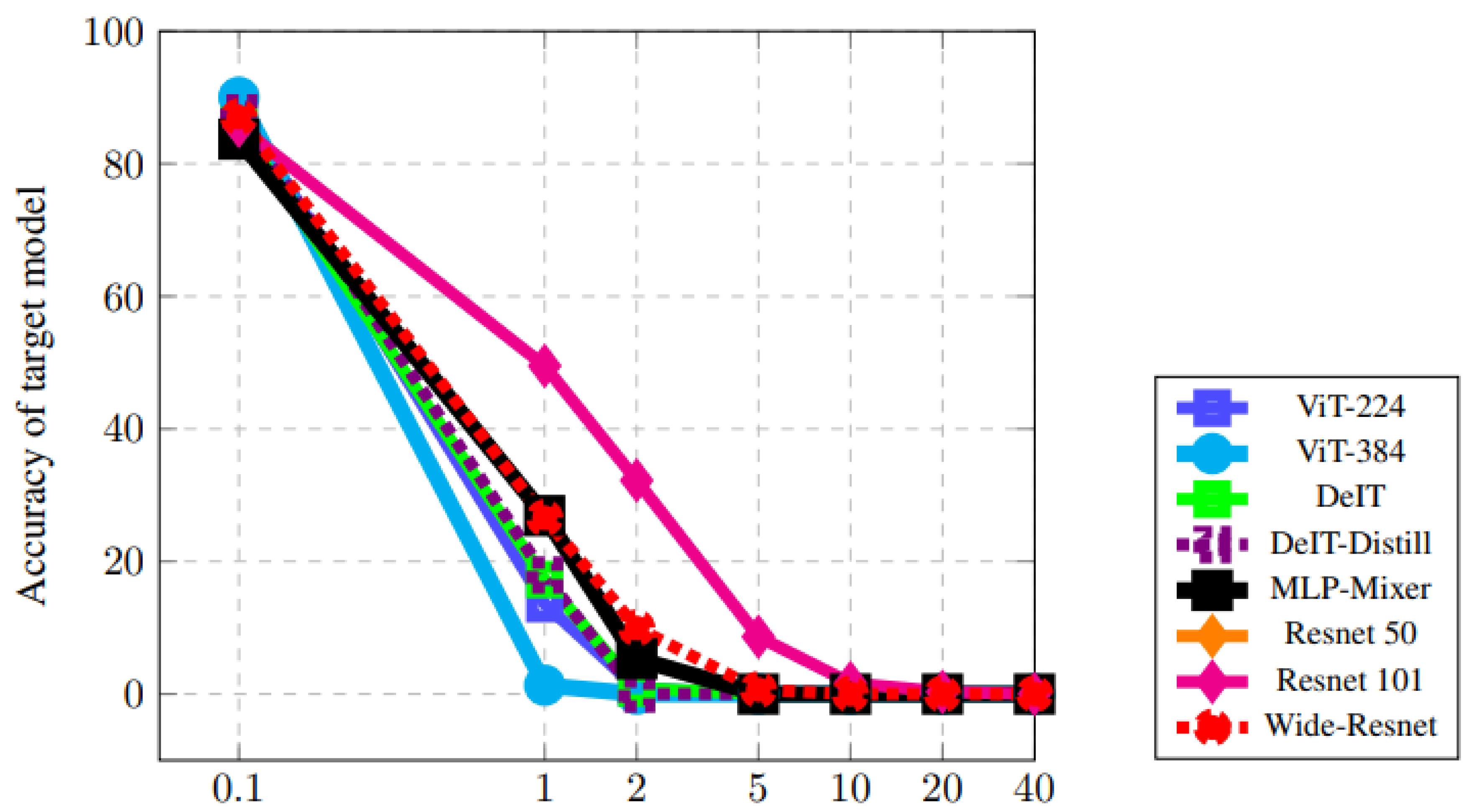
Require: x_0:Input image, $f(.)$: Classifier, y : Original label, K: Number of patches to be perturbed, p: Patch size.

1: $[b_1 \dots b_K]$= Top-K of $S(x_b) = \sqrt{\sum_{x_i \in x_b} \left| \frac{\partial L(f(x,y))}{\partial x_i} \right|^2}$, $\forall b$.
2: **while** **do** $f(x) \neq y$ OR MaxIter
3: $x_{b_k} = x_{b_k} + \nabla_{x_{b_k}} L; \ \forall \ b_k \in \{b_1, \dots, b_K\}$
4: $x_{b_k} = Project_{\epsilon_\infty}(x_{b_k})$ (optional)
5: **end while**

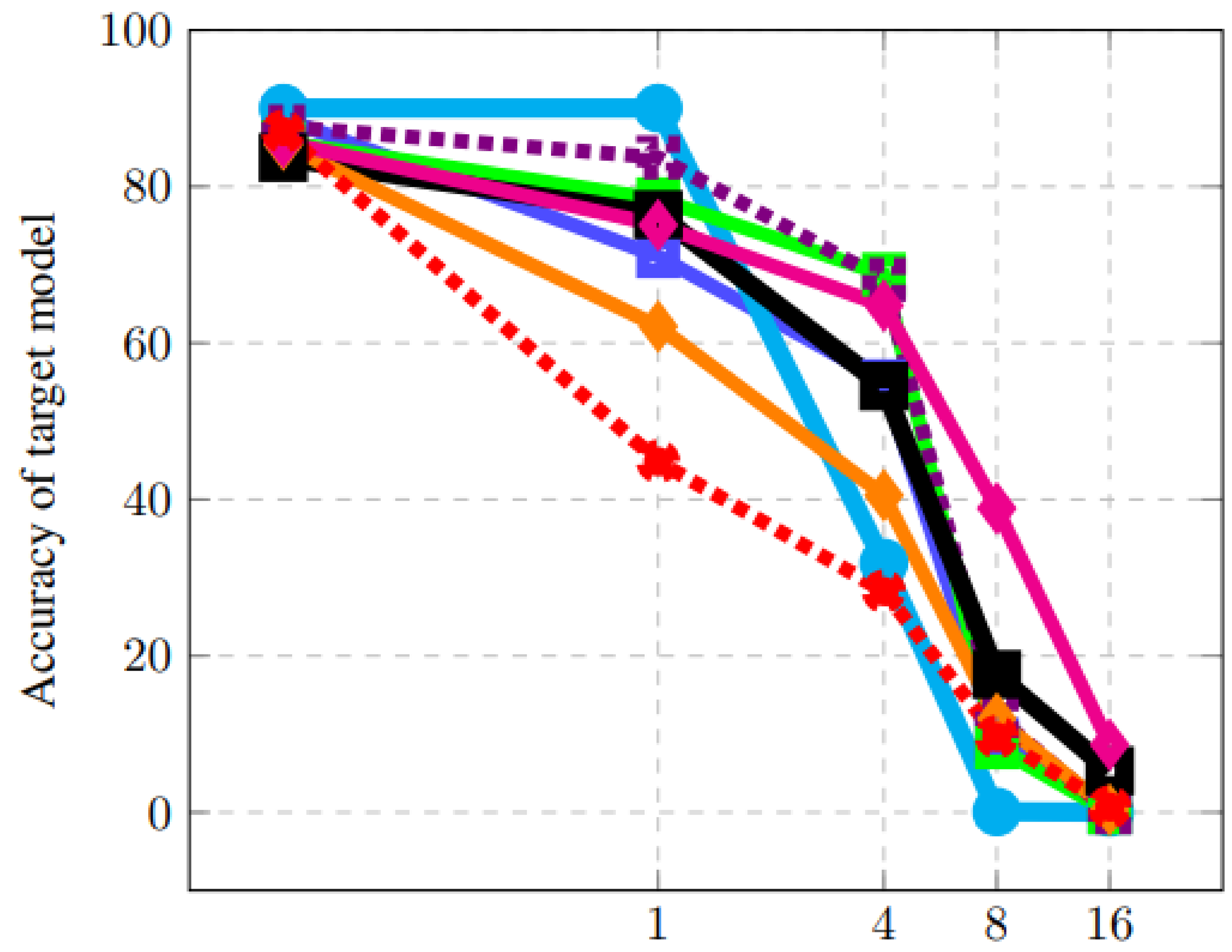
Attacks

- Sparse Attacks
- Token attacks ($p = 16 \times 16$)
- Mixed norm attacks

Results



(a) Token Budget (K), $p = 16$
Number of tokens

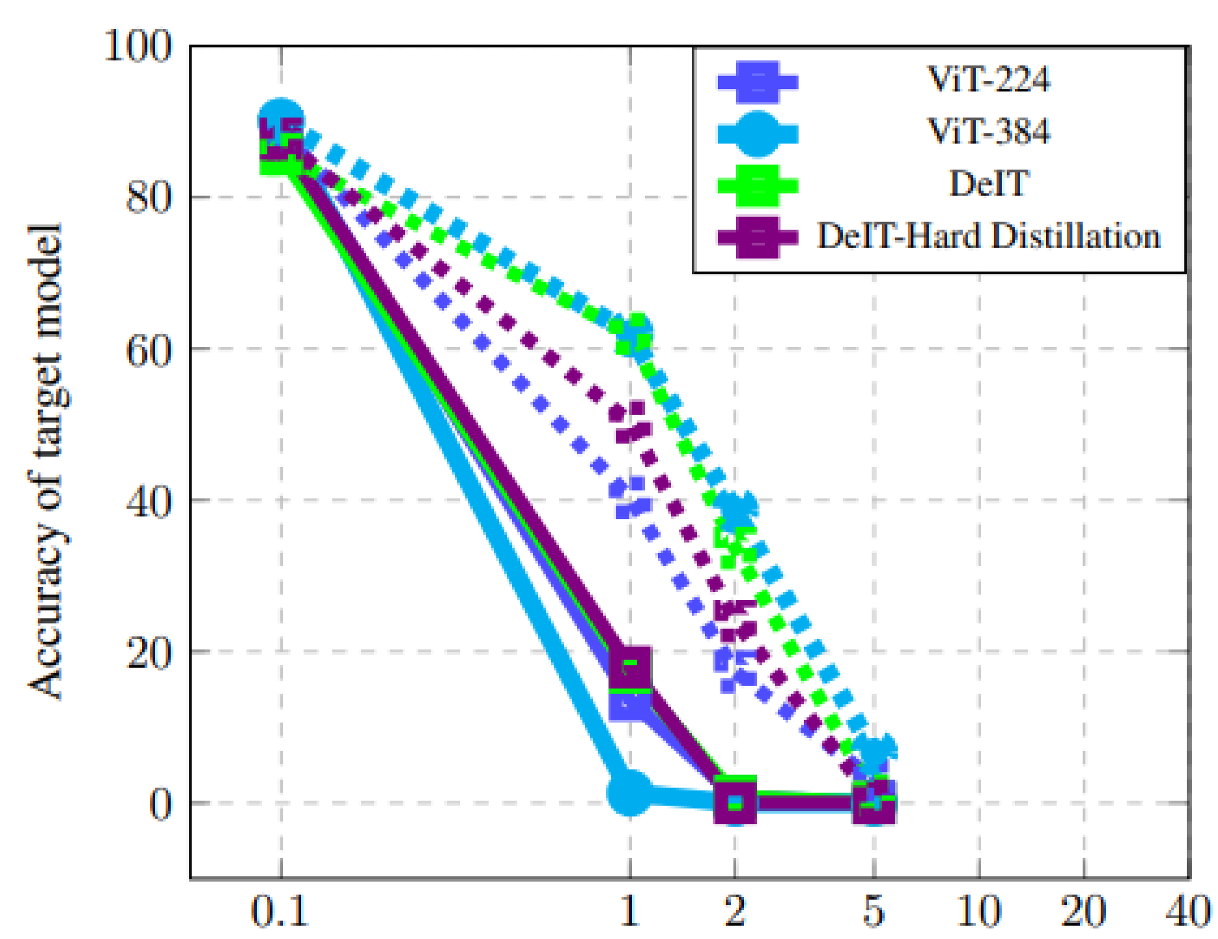


(b) Patch Attack Sizes (p), $K = 5$
Effect of patch size

Mixed Norm attacks

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

Saliency v/s Random Tokens



(a) Token Budget (K), $p = 16$