Towards Transferable Adversarial Attacks on Vision Transformers

Zhipeng Wei, Jingjing Chen, Micah Goldblum, Zuxuan Wu, Tom Goldstein, Yu-Gang Jiang

Shanghai Key Lab of Intelligent Information Processing, School of Computer Science, Fudan University
Shanghai Collaborative Innovation Center on Intelligent Visual Computing
Department of Computer Science, University of Maryland

Introduction

Background

Vision transformers (ViTs) have demonstrated impressive performance on a series of computer vision tasks, yet they still suffer from adversarial examples. In this paper, we posit that adversarial attacks on transformers should be specially tailored for their architecture, jointly considering both patches and self-attention, in order to achieve high transferability.

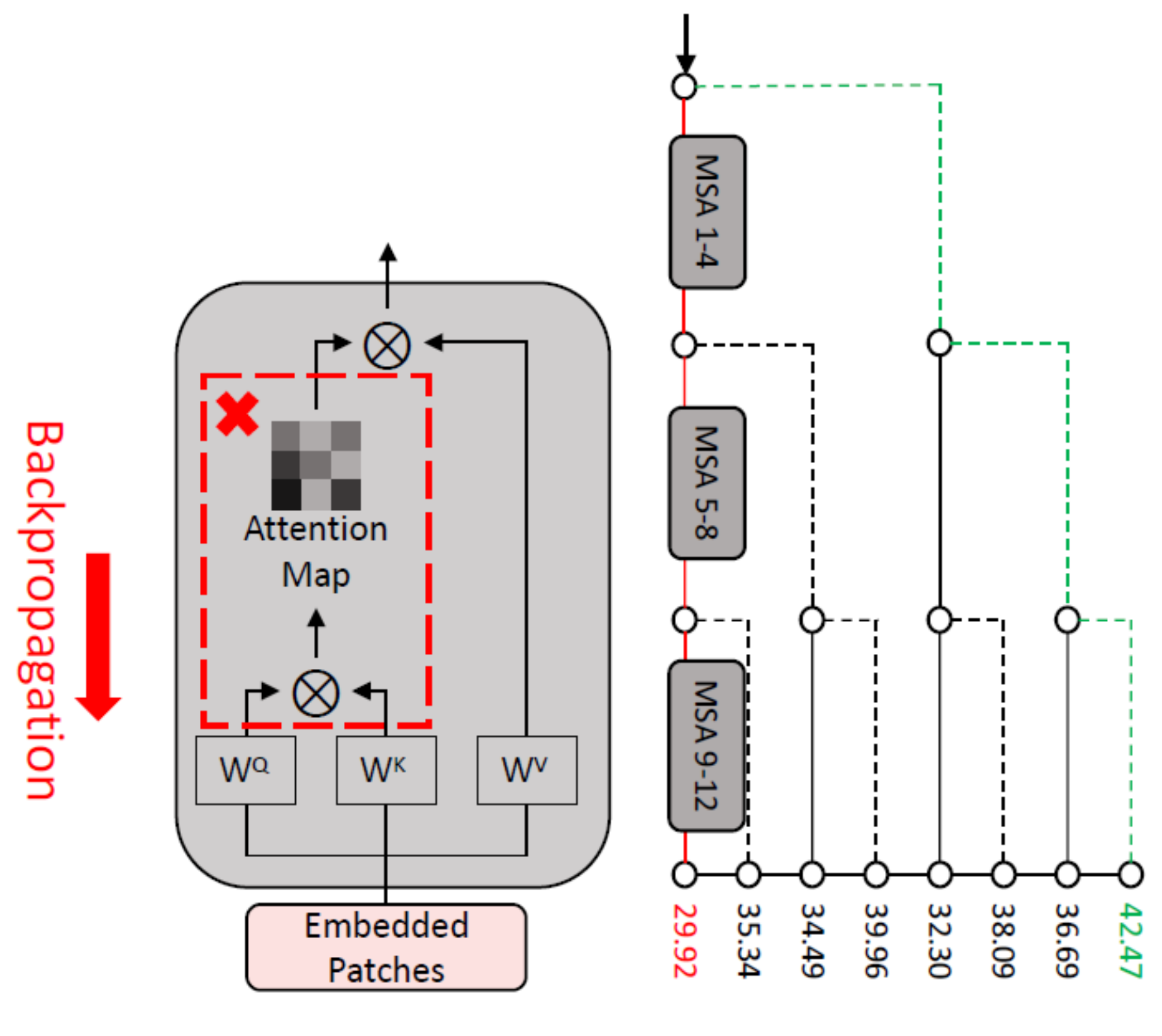
Our Method

We introduce a dual attack framework, which contains a Pay No Attention (PNA) attack and a PatchOut attack, to improve the transferability of adversarial samples across different ViTs. We show that skipping the gradients of attention during backpropagation can generate adversarial examples with high transferability. In addition, adversarial perturbations generated by optimizing randomly sampled subsets of patches at each iteration achieve higher attack success rates than attacks using all patches.

Framework

Pay No Attention Attack

The Pay No Attention (PNA) attack improves adversarial transferability by treating the attention weights computed on the forward pass as constants. In other words, it does not propagate through the branch of the computation graph that produces the attention weights, as illustrated below.

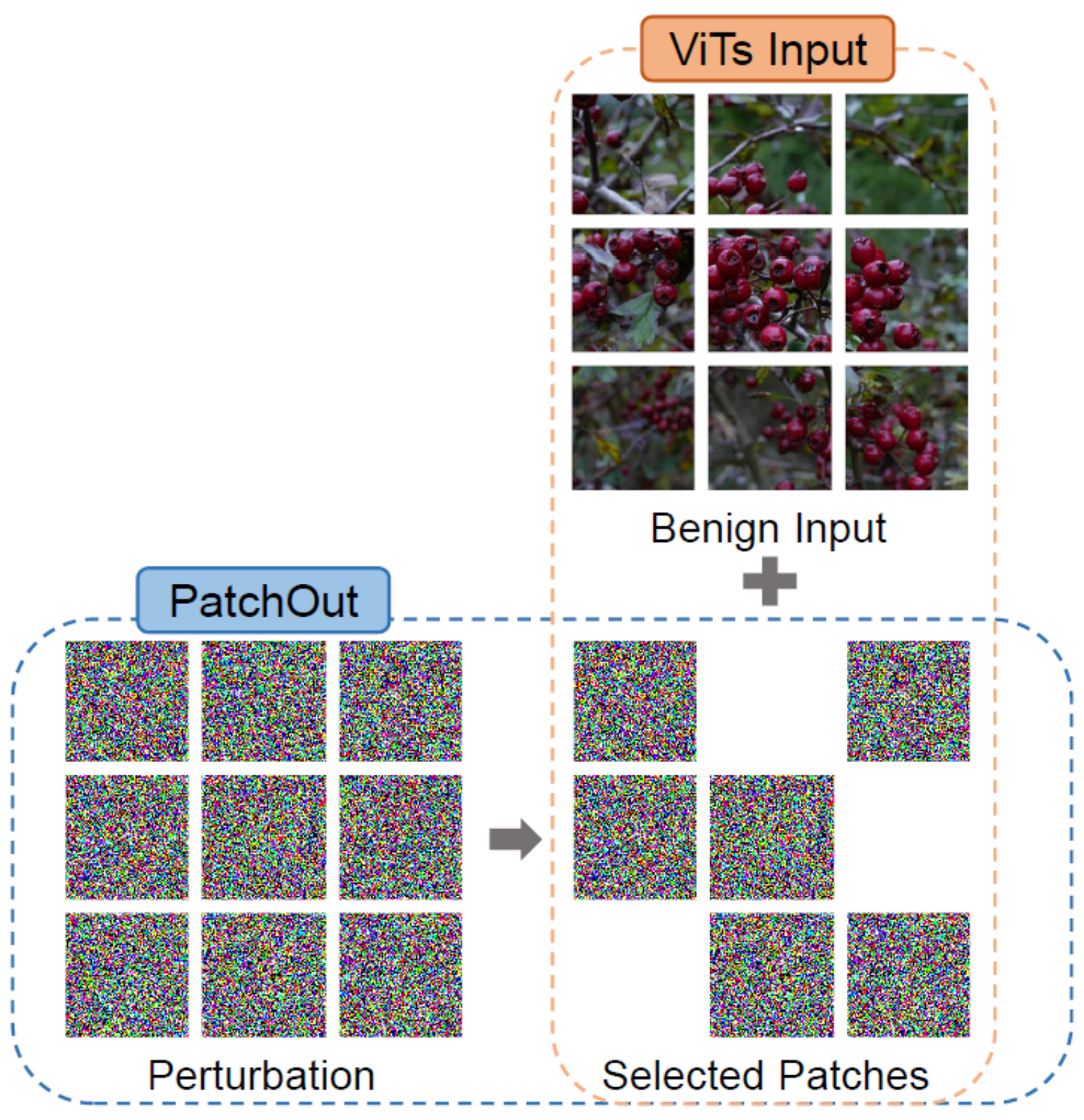


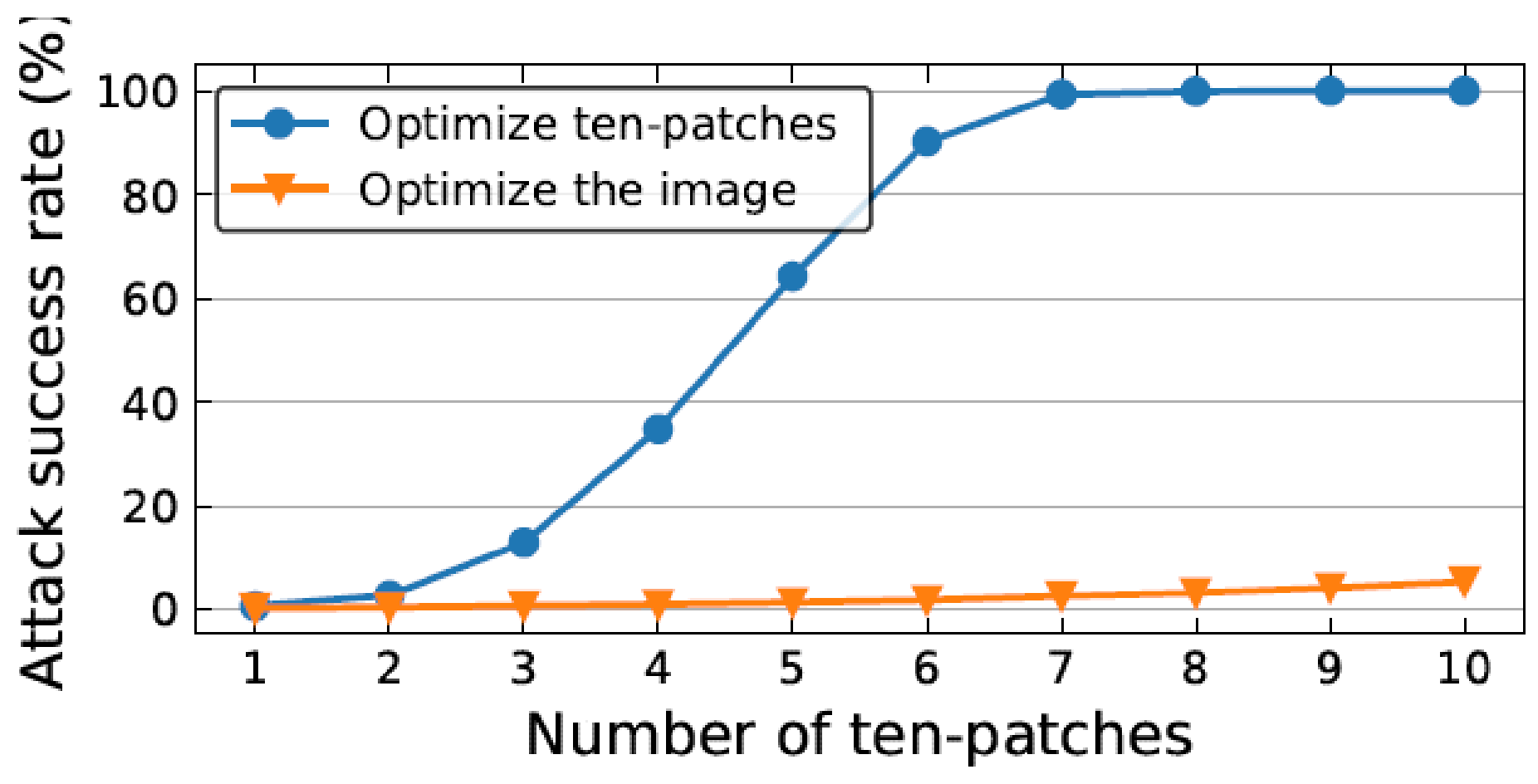
(a) Self-attention Block (b) Bypass Attention

To illustrate how the gradients of attention weights impair adversarial transferability, we conduct a toy experiments using the BIM attack on the white-box model ViT-B/16 using the ImageNet validation dataset. The right figure shows the results of Pay No Attention attack. We observe that the attack success rate (ASR) decreases as more attention gradients are used during backpropagation. Bypassing all gradients of attention (the green path) improves attack success rate from 29% to 42%.

PatchOut Attack

The PatchOut attack randomly samples a subset of patches to receive updates on each iteration of the attack crafting process. This is akin to using dropout on perturbation patches, and helps to combat over-fitting.





This figure shows the results of PatchOut, where we randomly select ten patches as one input pattern. We call such a sparse perturbation a “ten-patch.” We see that stacking multiple ten-patches achieves a higher attack success rate than using perturbations produced by optimizing on the whole image at once. This observation demonstrates that stacking perturbations from diverse input patterns can help alleviate the over-fitting problem.

The Proposed Dual Attack

Algorithm

Algorithm 1: The dual attack on ViTs

Input: The loss function J of Equation 7, a white-box model f, a clean image x with its ground-truth class y.
Parameter: The perturbation budget ϵ, iteration number I, used patch number T.
Output: The adversarial example.

```
$\delta_0 \leftarrow \mathbf{0}$
$\alpha \leftarrow \frac{\epsilon}{I}$
for $i = 0$ to $I - 1$ do
    $x_s \leftarrow PatchOut(x_p, T)$
    $M \leftarrow$ Equation 6
    $g \leftarrow PNA(\nabla_\delta J$ with the $L_2$ norm$)$
    $\delta_i \leftarrow clip_\epsilon(\delta_{i-1} + \alpha \cdot g)$
end for
$x_{adv} = x + \delta_I$
return $x_{adv}$
```

Objective function

$$\arg\max_{\delta} J(f(x + M \odot \delta), y) + \lambda||\delta||_2, s.t. \ ||\delta||_\infty < \epsilon$$

Where M is the mask of selected patches, The added second term encourages perturbations to have a large L2 norm, preferring a large distance from x. λ controls the balance between the loss function and the regularization term.

Experiments

Performance Comparison on ViTs

Method	ViT-B/16	PiT-B	CaiT-S-24	Visformer-S	DeiT-B	TNT-S	LeViT-256	ConViT-B
FGSM	15.57	19.80	20.43	19.37	22.08	22.78	18.80	25.58
BIM	20.77	22.17	22.63	22.70	33.53	32.13	20.45	35.30
MI	41.23	45.23	47.13	45.97	56.03	55.23	43.75	58.25
DI	32.57	45.13	43.07	47.77	48.08	55.18	43.25	49.35
TI	19.33	17.67	16.50	19.00	25.13	28.18	13.70	27.53
SIM	34.97	32.73	35.17	31.13	44.13	46.73	36.43	45.68
SGM	38.87	41.60	52.30	48.80	60.53	64.33	51.13	60.68
IR	21.33	22.70	24.00	23.43	34.00	33.43	21.30	36.38
TAP	25.27	24.73	33.40	32.20	43.20	39.78	30.03	42.20
ATA	3.47	1.13	0.97	2.67	3.68	3.37	2.02	3.72
SE	29.05	21.25	31.40	24.90	45.23	37.87	21.73	46.03
Ours	**46.10**	**52.40**	**59.87**	**58.60**	**63.85**	**67.25**	**57.62**	**63.70**

Performance Comparison on CNNs

Method	Inc-v3	Inc-v4	IncRes-v2	Res-v2	Inc-v3$_{ens3}$	Inc-v3$_{ens4}$	IncRes-v2$_{ens}$
FGSM	20.78	18.80	16.38	19.05	14.62	13.98	10.38
BIM	17.88	14.77	12.40	11.82	8.02	6.20	4.45
MI	39.65	37.43	32.17	33.28	24.80	22.10	17.68
DI	32.78	31.75	26.40	25.00	17.73	15.22	11.12
TI	23.27	23.60	15.28	20.88	18.80	20.80	13.92
SIM	30.55	27.63	24.17	24.48	19.45	17.67	13.40
SGM	38.42	34.00	27.25	27.25	18.70	16.53	11.68
IR	17.65	15.83	12.08	12.48	8.05	6.50	4.78
TAP	29.58	26.10	20.67	19.23	12.58	10.62	6.85
ATA	3.25	2.53	2.03	2.07	1.20	0.85	0.92
SE	18.40	16.47	12.33	12.63	9.13	6.73	5.10
Ours	**47.95**	**45.12**	**38.45**	**38.93**	**26.20**	**22.85**	**18.10**

Combining with Existing Methods

Method	ViTs	Normal CNNs	Robust CNNs
MI	49.10	35.63	21.53
MI + Ours	59.70	48.79	30.13
SGM	52.28	31.73	15.64
SGM + Ours	63.48	52.31	31.37

Qualitative Results

Summary

We identify several properties of ViTs. Specifically, we find that ignoring the gradients of attention units and only perturbing a subset of the patches at each iteration prevents overfitting and creates diverse input patterns, thus increasing transferability. After verifying our intuitions, we propose the dual attack for ViTs consisting of the Pay No Attention (PNA) attack and the PatchOut attack to craft adversarial examples with high transferability. We conduct a series of experiments with 8 ViTs, 4 normally trained CNNs, and 3 robustly trained CNNs to show that the proposed method can greatly improve adversarial transferability.