Transferable and imperceptible Adversarial Attacks on Vision Transformers

Advisor: Ren-Hung Hwang

Reporter: 313834009 陳煒函

Introduction

Introduction

- Vision Transformers (ViTs) have achieved remarkable success in computer vision but remain vulnerable to adversarial attacks.
- Adversarial attacks subtly perturb images to fool models. Two main categories:
  - **White-box attacks:** attacker knows model details.
  - **Black-box attacks:** attacker does not know model details.
- Transfer-based attacks use white-box attacks on a local model to generate perturbations that also fool other (black-box) models.
- Existing transfer-based attacks:
  - **Input Transformation:** uses transformed images to build transferable perturbations.
  - **Gradient Regularization:** stabilizes gradient updates but struggles with high variance in ViT intermediate blocks.

Contributions

- Proposed a **ViT-based** adversarial attack method that **integrates attack techniques** from multiple existing studies.
- Propose a **ViT-based** adversarial attack method that achieves both high attack success and imperceptible perturbations.
- Compare the **robustness of mainstream ViT models** against adversarial attacks
- Evaluate whether adversarial examples generated from ViT models can effectively attack CNN models.

Method

ViT-based adversarial attack overview

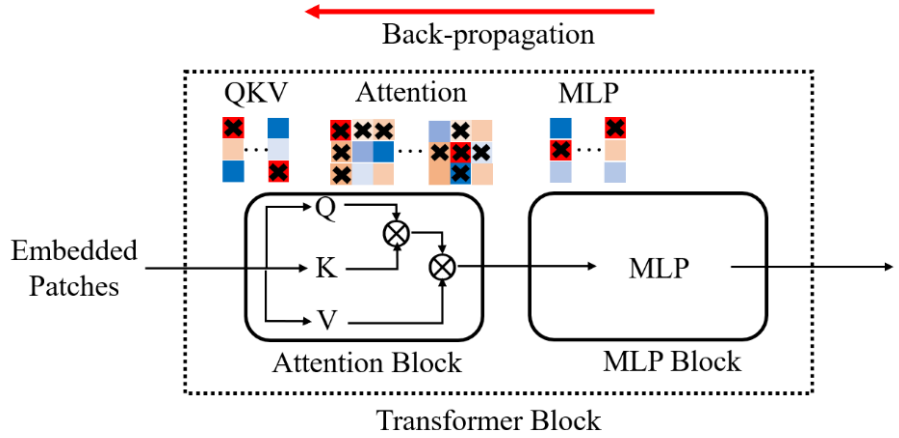
- Adversarial Attack Objective
  - Input image $x \in \mathbb{R}^{H \times W \times C}$
  - ViT splits x into $N = \frac{H \cdot W}{P^2}$ patches of size P×P×C
  - Model prediction: $f(x)$; Loss: $J(x, y; f)$
  - Goal: Find x_{adv} s.t. $f(x_{adv}) \neq f(x)$ and $\|x - x_{adv}\|_\infty < \epsilon$ (imperceptible)

Transferable Adversarial Attacks on Vision Transformers with Token Gradient Regularization (TGR) overview

- Token Gradient Regularization (TGR)
  - Problem: Large gradient variance → unstable updates → weak attacks
  - Limitation: Traditional methods only regularize input gradients
  - TGR Insight:
    - Focus on token gradients inside ViT
    - Identify extreme tokens (top-k / bottom-k gradients)
    - Zero-out those gradients to reduce variance
  - Effect: More stable and effective attack directions

Transferable Adversarial Attacks on Vision Transformers with Token Gradient Regularization (TGR)

$$g' = TGR(Grads, modules, k, s)$$
$$x_{t+1}^{adv} = x_t^{adv} + \alpha \cdot sgn\{g'\}, \quad (1)$$



Algorithm 1 Token Gradient Regularization

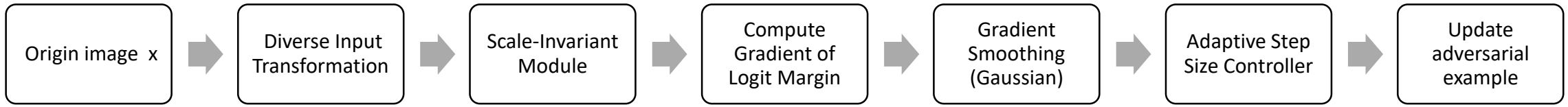
Require: network structure $modules$ and gradients $Grads$
Require: scaling factor s and extreme token number k
Ensure: the gradient on the input g'

```
for m in modules do
    if m is MLP or KQV then
        Grads[m] ← Grads[m] * s
        token ← extreme(Grads[m], k)          ▷ Extreme
    Tokens on MLP or KQV component
        for i = 0 ← 2k − 1 do
            Grads[m][token[i], :] = 0
        end for
    else if m is Attention then
        Grads[m] ← Grads[m] * s
        tokens ← extreme(Grads[m], k)         ▷ Extreme
    Token Pairs on the Attention Map
        for i = 0 ← 2k − 1 do
            Grads[m][tokens[i, 0], :, :] = 0
            Grads[m][:, tokens[i, 1], :] = 0
        end for
    end if
end for
```

Enhance the attack success rate of TGR and the imperceptibility of adversarial samples (TGRv1)

- Diverse Input
  - Add cropping, flipping, translation, simulated rotation
- Scale Invariance
  - Add multi-scale zoom
- Logit Margin Loss
  - Using margin-based cross entropy
- Adaptive Step Size
  - Dynamically adjust the step size based on the logits success rate
- Gradient Smoothing (Gaussian)
  - Smoothing the gradient using a Gaussian kernel

Enhance the attack success rate of TGR and the imperceptibility of adversarial samples (TGRv1)



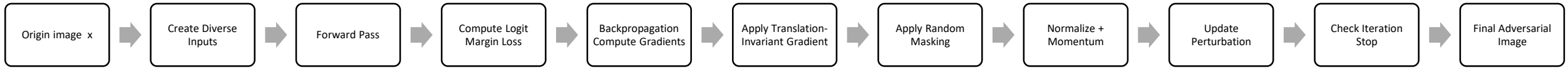
Enhancing the imperceptibility of TGR-generated adversarial examples(TGRv2)

- Diverse Input Transformations
  - Randomly resizes each input image to a size between 224 and 256, then resizes it back to 224×224.
  - This disrupts the spatial structure of the input, making the adversarial perturbation harder to detect visually and preventing patterns from becoming too obvious to the human eye.
- Translation-Invariant Gradient
  - Applies a Gaussian smoothing kernel to the gradient, making it more robust to spatial shifts.
  - This restricts the perturbation to only part of the image in each iteration, limiting the extent of visible noise and making the attack more subtle.
- Random Masking
  - Applies a random binary mask (rate = 0.7 by default) to the gradient before updating.
  - This restricts the perturbation to only part of the image in each iteration, limiting the extent of visible noise and making the attack more subtle.

Enhancing the imperceptibility of TGR-generated adversarial examples (TGRv2)

- Logit Margin Loss
  - Uses Logit Margin Loss instead of standard cross-entropy loss, calculating the difference between the correct class and the highest incorrect class.
  - This stabilizes the gradient updates, reducing local oscillations, and helps the attack push the sample away from the correct class more effectively, while maintaining imperceptibility.
- Momentum Iterative Optimization
  - Incorporates a momentum term into the gradient update process, smoothing the update direction and preventing sharp local perturbations that could cause visible noise.
  - Normalizing the gradient before applying momentum further enhances attack stability and robustness.

Enhancing the imperceptibility of TGR-generated adversarial examples



Experiments

Experiment Setup

- Dataset: 1,000 random images from ImageNet (ILSVRC 2012 val set).
- Models:
  - Source models (ViTs): ViT-B/16, PiT-B, CaiT-S/24, Visformer-S
  - Target models (ViTs): DeiT-B, TNT-S, LeViT-256, ConViT-B
  - Target models (CNNs):
    - Undefended: Inc-v3, Inc-v4, IncRes-v2, ResNet-v2-152
    - Defended: Inc-v3ens3, Inc-v3ens4, IncRes-v2adv

Attack Settings

- Baseline attacks: MIM, VMI, TGR
- Metric: Attack Success Rate (ASR)、FID
- Parameters:
  - Perturbation: $\varepsilon = 16$
  - Iterations: $T = 10$
  - Decay factor: $\mu = 1.0$
  - PatchOut: 130 patches
  - Image size: 224×224
  - Patch size: 16

The attack success rates (%) against VIT models by various transfer-based attacks.

Model	Attack	ViT-B/16	PiT-B	CaiT-S/24	Visformer-S	DeiT-B	TNT-S	LeViT-256	ConViT-B	Avg
ViT-B/16	MIM	**100.0%**	34.5%	64.1%	36.5%	64.3%	50.2%	33.8%	66.0%	56.1%
	VMI	99.6%	48.8%	74.4%	49.5%	73.0%	64.8%	50.3%	75.9%	67.0%
	PNA	**100.0%**	45.2%	78.6%	47.7%	78.6%	62.8%	47.1%	79.5%	67.4%
	TGR	99.6%	50.7%	**83.3%**	55.5%	**83.7%**	74.0%	58.8%	**84.5%**	73.8%
	TGRv1	98.7%	**80.2%**	81.5%	**78.0%**	80.5%	**83.0%**	**77.2%**	82.4%	**82.7%**
	TGRv2	97.1%	31.1%	41.9%	37.4%	41.1%	41.9%	34.5%	44.5%	46.2%
PiT-B	MIM	24.7%	100.0%	34.7%	44.5%	33.9%	43.0%	38.3%	37.8%	44.6%
	VMI	38.9%	99.7%	51.0%	56.6%	50.1%	57.0%	52.6%	51.7%	57.2%
	PNA	47.9%	**100.0%**	62.6%	74.6%	62.4%	70.6%	67.3%	61.7%	68.3%
	TGR	59.4%	**100.0%**	77.9%	87.0%	78.7%	86.8%	81.7%	78.1%	81.2%
	TGRv1	**76.7%**	99.5%	**87.8%**	**92.2%**	**90.1%**	**92.3%**	**87.5%**	**90.8%**	**89.6%**
	TGRv2	39.8%	95.0%	43.4%	57.0%	44.8%	51.7%	41.6%	48.3%	52.7%
CaiT-S/24	MIM	70.9%	54.8%	99.8%	55.1%	90.2%	76.4%	54.8%	88.5%	73.8%
	VMI	76.3%	63.6%	98.8%	67.3%	88.5%	82.3%	67.0%	88.1%	78.9%
	PNA	82.4%	60.7%	99.7%	67.7%	95.7%	86.9%	67.1%	94.0%	81.7%
	TGR	**93.5%**	75.8%	**100%**	85.5%	**99.6%**	**97.6%**	86.4%	**99.3%**	92.2%
	TGRv1	89.9%	**94.6%**	99.3%	**95.3%**	96.4%	97.1%	**94.7%**	95.7%	**95.4%**
	TGRv2	59.6%	57.2%	97.1%	65.0%	73.4%	73.5%	57.1%	74.5%	69.7%
Visformer-S	MIM	28.1%	50.4%	41.0%	99.9%	36.9%	51.9%	49.4%	39.6%	49.6%
	VMI	39.2%	60.0%	56.6%	**100.0%**	54.1%	62.8%	59.1%	54.4%	60.7%
	PNA	35.4%	61.5%	54.7%	**100.0%**	51.0%	66.3%	64.5%	50.7%	60.5%
	TGR	44.8%	72.3%	66.6%	**100.0%**	64.5%	78.2%	77.9%	58.4%	70.3%
	TGRv1	**72.5%**	**94.1%**	**89.9%**	**100.0%**	**88.6%**	**94.4%**	**90.2%**	**88.2%**	**89.7%**
	TGRv2	37.6%	50.7%	49.8%	99.1%	49.8%	56.2%	50.6%	48.8%	55.3%

The attack success rates (%) against CNN models by various transfer-based attacks.

Model	Attack	Inc-v3	Inc-v4	IncRes-v2	Res-v2	Inc-v3$_{ens3}$	Inc-v3$_{ens4}$	IncRes-v2$_{adv}$	Avg
ViT-B/16	MIM	31.7%	28.6%	26.1%	29.4%	22.3%	19.8%	16.5%	24.9%
	VMI	43.1%	41.6%	37.9%	42.6%	31.4%	30.6%	25.0%	36.0%
	PNA	42.7%	37.5%	35.3%	39.5%	29.0%	27.3%	22.6%	33.4%
	TGR	49.3%	45.7%	40.0%	43.6%	35.5%	32.6%	58.8%	43.6%
	TGRv1	**69.4%**	**69.0%**	**64.6%**	**62.2%**	**60.1%**	**59.9%**	**56.2%**	**63.1%**
	TGRv2	33.8%	34.7%	27.9%	31.0%	34.0%	38.6%	31.1%	33.0%
PiT-B	MIM	36.3%	34.8%	27.4%	29.6%	19.0%	18.3%	14.1%	25.6%
	VMI	47.3%	45.4%	40.7%	43.4%	35.9%	34.4%	29.7%	39.5%
	PNA	59.3%	56.3%	49.8%	53.0%	33.3%	32.0%	25.5%	44.1%
	TGR	69.2%	65.7 %	60.9%	61.0%	42.9%	40.7%	32.9%	53.3%
	TGRv1	**78.7%**	**76.4%**	**70.8%**	**68.3%**	**58.5%**	**58.6%**	**52.5%**	**66.3%**
	TGRv2	38.3%	38.1%	29.8%	32.9%	33.0%	36.0%	30.8%	34.1%
CaiT-S/24	MIM	48.4%	42.9%	39.5%	43.8%	30.8%	27.6%	23.3%	36.6%
	VMI	58.5%	50.9%	48.2%	52.0%	38.1%	36.1%	30.1%	44.8%
	PNA	57.2%	51.8%	47.7%	51.6%	38.4%	36.2%	30.1%	44.7%
	TGR	73.5%	67.2%	67.5%	68.0%	56.5%	52.5%	44.2%	61.34%
	TGRv1	**89.0%**	**87.3%**	**87.0%**	**84.8%**	**80.6%**	**78.8%**	**76.5%**	**88.43%**
	TGRv2	50.9%	50.4%	43.7%	45.0%	50.7%	53.3%	46.6%	48.7%
Visformer-S	MIM	44.5%	42.5%	36.6%	39.6%	24.4%	20.5%	16.6%	32.1%
	VMI	54.6%	53.2%	48.5%	52.2%	33.0%	32.0%	22.2%	42.2%
	PNA	55.9%	54.6%	46.0%	51.7%	29.3%	26.2%	21.1%	40.6%
	TGR	72.2%	71.7%	62.2%	67.2%	40.9%	35.4%	28.3%	54.0%
	TGRv1	**87.6%**	**88.3%**	**83.3%**	**80.0%**	**70.5%**	**68.9%**	**63.3%**	**77.4%**
	TGRv2	48.9%	50.1%	41.5%	44.1%	44.9%	47.8%	39.7%	45.3%

The adversarial exapmples FID score models by various transfer-based attacks.

Model	Attack	FID
ViT-B/16	TGR	72.96
	TGRv1	67.91
	TGRv2	**53.70**
PiT-B	TGR	96.77
	TGRv1	78.70
	TGRv2	**61.02**
CaiT-S/24	TGR	99.50
	TGRv1	86.37
	TGRv2	**65.92**
Visformer-S	TGR	94.54
	TGRv1	73.14
	TGRv2	**67.14**

Ablation Study (TGRv1)

Attention	QKV	MLP	ViTs	CNNs	CNNs-adv
-	-	-	73.26	55.9	47.2
✓	-	-	79.18	64.0	54.4
-	✓	-	73.56	56.3	47.0
-	-	✓	78.83	61	52.6
✓	✓	-	78.56	63.3	54.5
✓	-	✓	82.64	67.0	59.4
-	✓	✓	78.76	60.7	52.7
✓	✓	✓	82.7	66.3	58.7

Conclusion

Conclusion

- Analyzed why gradient regularization-based methods have low transferability:
  - They reduce gradient variance only at the input level, but variance in intermediate blocks remains high.
  - This causes models to get stuck in local optima.
- Proposed Token Gradient Regularization (TGRv1):
  - Regularizes gradient variance at each internal block of ViTs.
  - Combined with attack methods proposed in other papers.
  - Uses these regularized gradients to generate transferable adversarial samples.
- Extensive experiments on ViTs and CNNs confirm TGRv1's effectiveness.

Reference

Reference

- Zhang, Jianping, et al. "Transferable adversarial attacks on vision transformers with token gradient regularization." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2023.
- Dong, Yinpeng, et al. "Boosting adversarial attacks with momentum." Proceedings of the IEEE conference on computer vision and pattern recognition. 2018
- Wang, Xiaosen, and Kun He. "Enhancing the transferability of adversarial attacks through variance tuning." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021.
- Wei, Zhipeng, et al. "Towards transferable adversarial attacks on vision transformers." Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 36. No. 3. 2022.
- Weng, Juanjuan, et al. "Logit margin matters: Improving transferable targeted adversarial attack by logit calibration." IEEE Transactions on Information Forensics and Security 18 (2023): 3561-3574.

Thanks