Transferable Adversarial Attacks on Vision Transformers with Token Gradient Regularization

WED-PM-387

Jianping Zhang,Yizhan Huang, Weibin Wu, Michael R. Lyu

The Chinese University of Hong Kong

Sun Yat-sen University

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

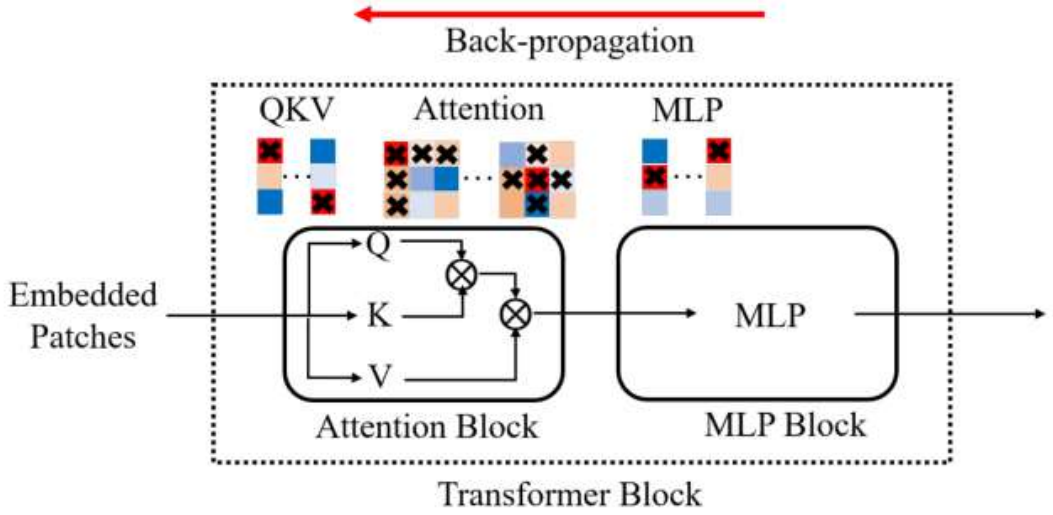
ARISE Automated Reliable Intelligent Software Engineering

香港中文大學 The Chinese University of Hong Kong

Transferable Adversarial Attacks on Vision Transformers with Token Gradient Regularization

Contribution

- Observe the large variance of the back-propagated gradients in intermediate blocks of ViTs leads to a low adversarial transferability
- TGR regularizes the back-propagated gradient in each internal block of ViTs in a token-wise manner.
- Extensive experiments validate the advantages of our method over state-of-the-art baselines.



Method

- Token is the building-block of ViTs
- Extreme token gradient
  - Model specific
  - Unstable features
- Token Gradient Regularization
  - Eliminate extreme tokens
- Implementation
  - QKV
  - Attention
  - MLP block

Introduction

- Vision Transformers (ViTs): successor for CNNs
- ViTs are vulnerable to adversarial attacks
- Identity deficiencies of ViTs: adversarial examples
- Adversarial attacks (model information)
  - White-box
  - Black-box
    - Transfer-based Attacks

Introduction

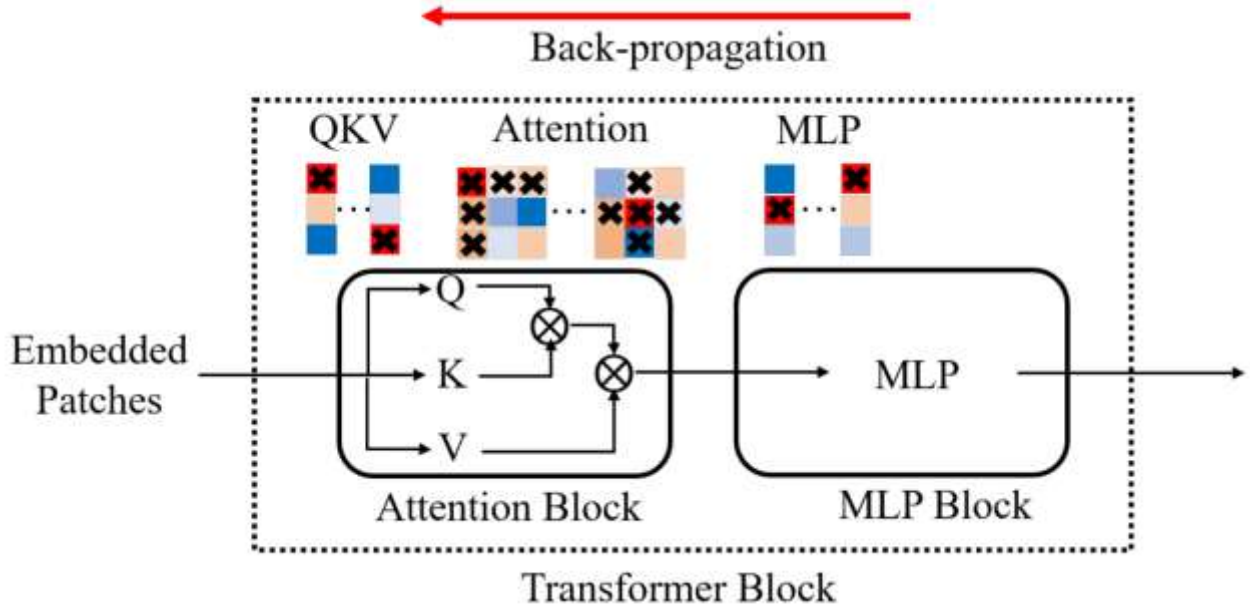
- Transfer-based Attacks
  - Input transformation
  - Gradient regularization
    - VMI: reduce input gradient variance
- Limitations
  - Large internal variance
  - Focus on model-specific features
- Token Gradient Regularization (TGR)
  - Reduce the internal variance of ViTs

Method

- Token Gradient Regularization
  - Tokens are important in ViTs
- Regularization in the intermediate block
  - Regularize the token gradients
- Extreme token gradient
  - Model specific
  - Unstable features
- Eliminate extreme tokens

Method

- Implementation
  - Attention block
    - QKV
    - Attention
  - MLP block



Method

- TGR Algorithm

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
```

Experiment

- Attack success rates
  - Transferability between ViTs

Model	Attack	ViT-B/16	PiT-B	CaiT-S/24	Visformer-S	DeiT-B	TNT-S	LeViT-256	ConViT-B
ViT-B/16	MIM	100.0	34.5	64.1	36.5	64.3	50.2	33.8	66.0
	VMI	99.6	48.8	74.4	49.5	73.0	64.8	50.3	75.9
	SGM	100.0	36.9	77.1	40.1	77.9	61.6	40.2	78.4
	PNA	100.0	45.2	78.6	47.7	78.6	62.8	47.1	79.5
	TGR	**100.0**	**49.5**	**85.0**	**53.8**	**85.6**	**73.1**	**56.5**	**85.4**
PiT-B	MIM	24.7	100.0	34.7	44.5	33.9	43.0	38.3	37.8
	VMI	38.9	99.7	51.0	56.6	50.1	57.0	52.6	51.7
	SGM	41.8	100.0	57.3	73.9	57.9	72.6	68.1	59.9
	PNA	47.9	100.0	62.6	74.6	62.4	70.6	67.3	61.7
	TGR	**60.3**	**100.0**	**80.2**	**87.3**	**78.0**	**87.1**	**81.6**	**76.5**
CaiT-S/24	MIM	70.9	54.8	99.8	55.1	90.2	76.4	54.8	88.5
	VMI	76.3	63.6	98.8	67.3	88.5	82.3	67.0	88.1
	SGM	86.0	55.8	100.0	68.2	97.7	91.1	74.9	96.7
	PNA	82.4	60.7	99.7	67.7	95.7	86.9	67.1	94.0
	TGR	**88.2**	**66.1**	**100.0**	**75.4**	**98.8**	**92.8**	**74.7**	**97.9**
Visformer-S	MIM	28.1	50.4	41.0	99.9	36.9	51.9	49.4	39.6
	VMI	39.2	60.0	56.6	100.0	54.1	62.8	59.1	54.4
	SGM	18.8	41.8	34.9	100.0	31.2	52.1	52.7	29.5
	PNA	35.4	61.5	54.7	100.0	51.0	66.3	64.5	50.7
	TGR	**41.2**	**70.3**	**62.0**	**100.0**	**59.5**	**74.7**	**74.8**	**56.2**

Table 1. The attack success rates (%) against eight models by various transfer-based attacks. The best results are marked in bold.

Experiment

- Attack success rates
  - Transferability between ViTs and CNNs

Model	Attack	Inc-v3	Inc-v4	IncRes-v2	Res-v2	Inc-v3$_{ens3}$	Inc-v3$_{ens4}$	IncRes-v2$_{adv}$
ViT-B/16	MIM	31.7	28.6	26.1	29.4	22.3	19.8	16.5
	VMI	43.1	41.6	37.9	42.6	31.4	30.6	25.0
	SGM	31.5	27.7	23.8	28.2	20.8	18.0	14.3
	PNA	42.7	37.5	35.3	39.5	29.0	27.3	22.6
	TGR	**47.5**	**42.3**	**37.6**	**43.3**	**31.5**	**30.8**	**25.6**
PiT-B	MIM	36.3	34.8	27.4	29.6	19.0	18.3	14.1
	VMI	47.3	45.4	40.7	43.4	35.9	34.4	29.7
	SGM	50.6	45.4	38.4	41.9	25.6	20.8	16.7
	PNA	59.3	56.3	49.8	53.0	33.3	32.0	25.5
	TGR	**72.1**	**69.8**	**65.1**	**64.8**	**43.6**	**41.5**	**32.8**
CaiT-S/24	MIM	48.4	42.9	39.5	43.8	30.8	27.6	23.3
	VMI	58.5	50.9	48.2	52.0	38.1	36.1	30.1
	SGM	53.5	45.9	40.2	45.9	30.8	28.5	21.0
	PNA	57.2	51.8	47.7	51.6	38.4	36.2	30.1
	TGR	**60.3**	**52.9**	**49.3**	**53.4**	**39.6**	**37.0**	**31.8**
Visformer-S	MIM	44.5	42.5	36.6	39.6	24.4	20.5	16.6
	VMI	54.6	53.2	48.5	52.2	33.0	32.0	22.2
	SGM	43.2	41.1	29.6	35.7	16.1	13.0	8.2
	PNA	55.9	54.6	46.0	51.7	29.3	26.2	21.1
	TGR	**65.9**	**66.8**	**55.3**	**60.9**	**36.0**	**32.5**	**23.3**

Table 2. The attack success rates (%) against seven models by various transfer-based attacks. The best results are marked in bold.

Experiment

- Ablation Study – gradient variance

Methods	Deep	Middle	Shallow	Average
MIM	7.5	37.6	70.9	38.7
VMI	4.0	19.1	34.0	19.1
TGR	**1.7**	**5.1**	**6.6**	**4.4**

Table 4. The average gradient variance of ViT-B/16 by different attacking methods. The best results are marked in bold.

Experiment

- Ablation Study - components

Attention	QKV	MLP	ViTs	CNNs	CNNs-adv
-	-	-	56.2	29.0	19.5
✓	-	-	67.4	38.1	25.4
-	✓	-	64.1	33.7	23.1
-	-	✓	57.3	30.0	19.9
✓	✓	-	69.7	40.0	27.3
✓	-	✓	69.3	39.4	26.6
-	✓	✓	66.0	35.5	23.7
✓	✓	✓	73.6	42.7	29.3

Table 6. The average attack success rates (%) against ViTs, CNNs, and adversarially trained CNNs by various setting of components.

Conclusion

- TGR regularizes the back-propagated gradient in each internal block of ViTs in a token-wise manner.
- TGR is extendable to combine with other attack methods and attacks on CNN models.
- Extensive experiments validate the effectiveness of our methods over state-of-the-art baselines.

Thank you!

香港中文大學
The Chinese University of Hong Kong