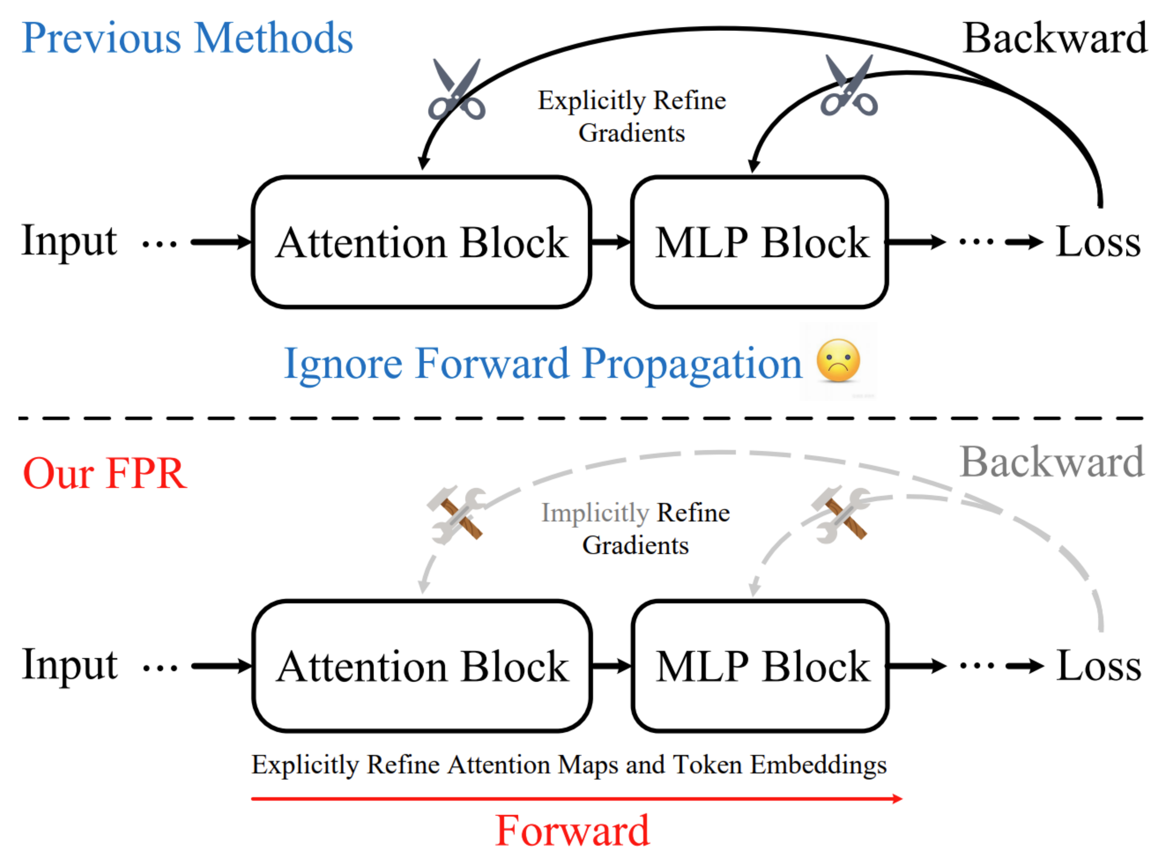
Improving Adversarial Transferability on Vision Transformers via Forward Propagation Refinement

Yuchen Ren[1], Zhengyu Zhao[1*], Chenhao Lin[1], Bo Yang[2], Lu Zhou[3], Zhe Liu[4], Chao Shen[1]
[1]Xi'an Jiaotong University, China; [2]Information Engineering University, China;
[3]Nanjing University of Aeronautics and Astronautics, China; [4]Zhejiang Lab, China

CVPR25
Paper ID: 10733

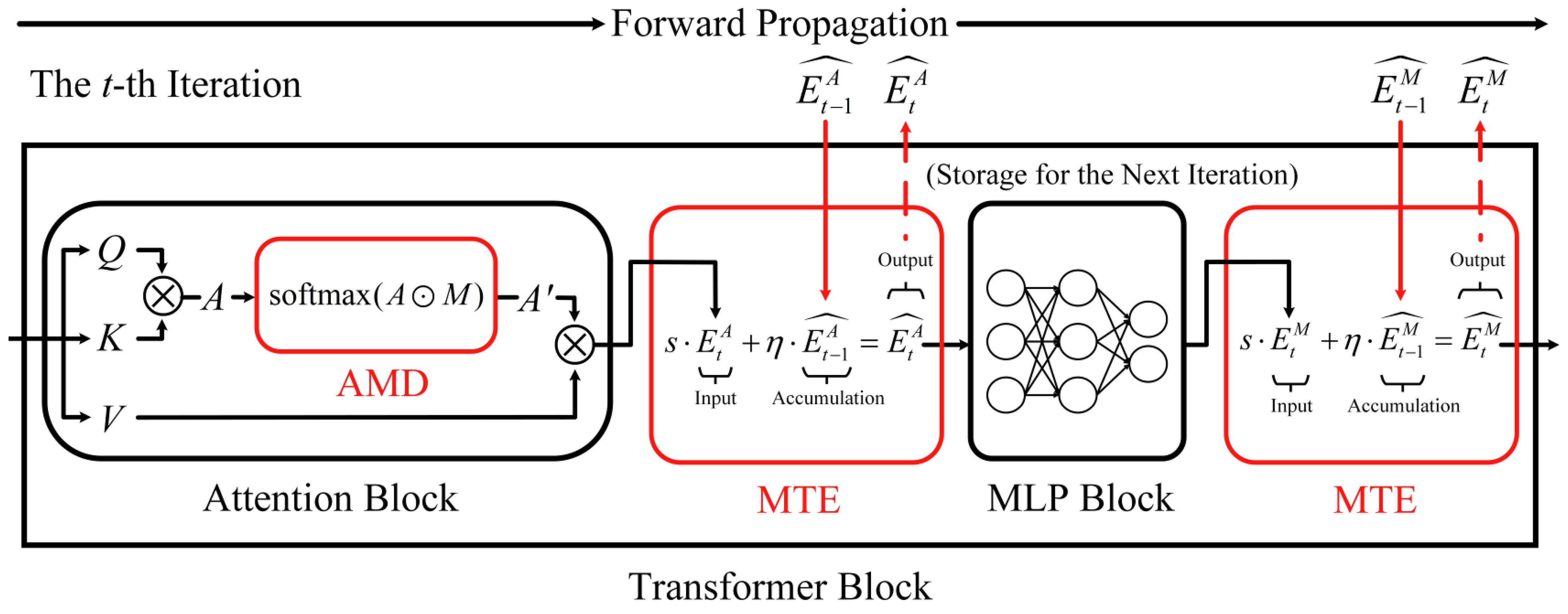
Introduction



(1) Refine ViT surrogate models from the forward propagation

(2) Propose the FPR, consisting of AMD and MTE

Forward Propagation Refinement (FPR)



E_t^A / E_t^M : Output Token Embedding of Attention / MLP Block at t-th Iteration

$\widehat{E_t^A}$ / $\widehat{E_t^M}$: Accumulated Token Embedding of Attention / MLP Block at t-th Iteration

A : Attention Map A' : Diversified Attention Map M : Diversity Matrix $Q/K/V$: Query/Key/Value Component

AMD and MTE

Attention Map Diversification (AMD)

$$A' = AMD(A) = \text{softmax}(A \odot M)$$

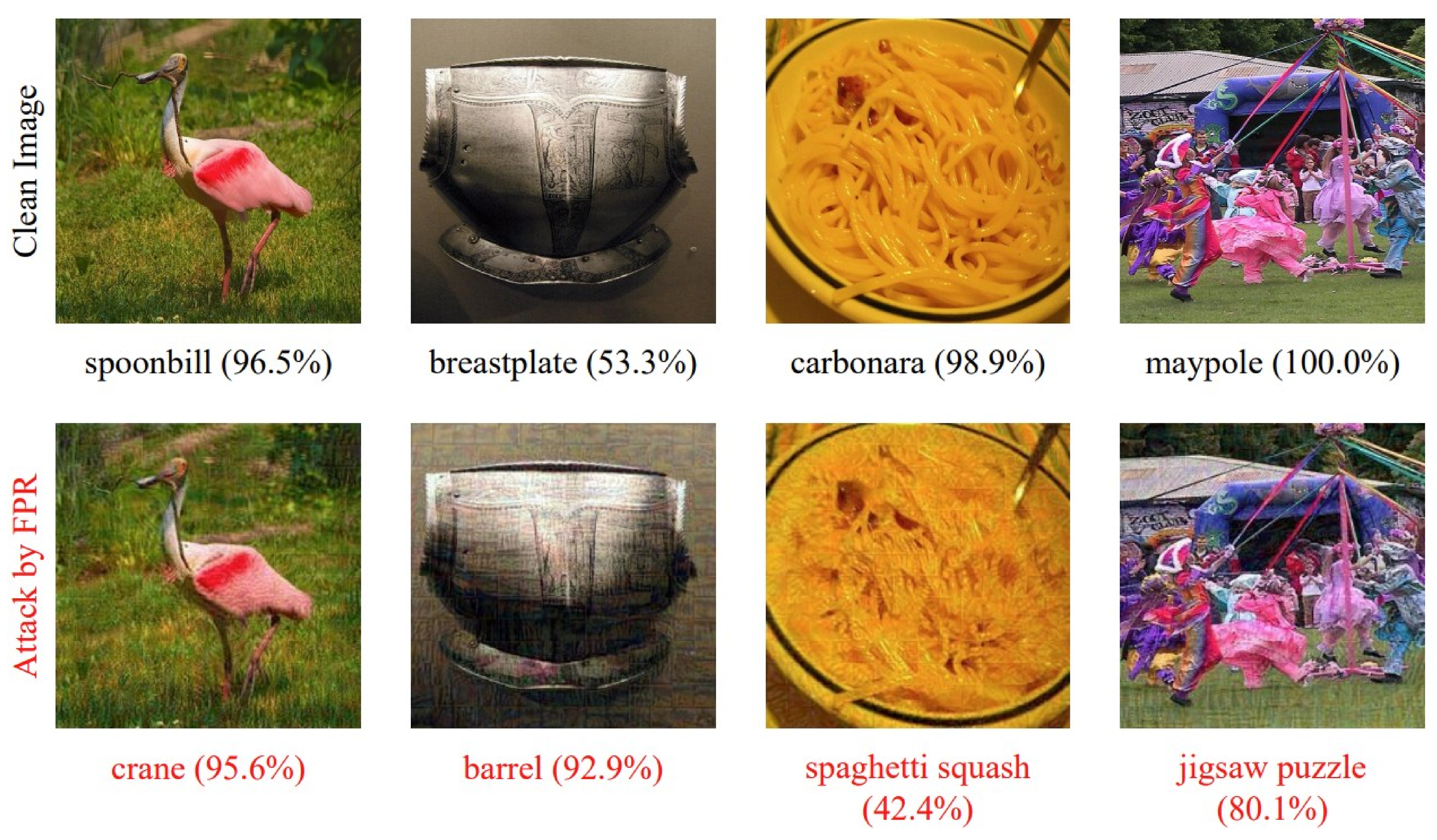
Momentum Token Embedding (MTE)

$$\widehat{E_t} = MTE(E_t) = \eta \cdot \widehat{E_{t-1}} + s \cdot E_t$$

(1) AMD and MTE improve the transferability from different angles

(2) Their complementary effects jointly enhance the adversarial transferability

Visualization



Main Experiment

Model	Attack	ViT-B	CaiT-S	PiT-B	Vis-S	Swin-T	DeiT-T	CoaT-T	RN-18	VGG-16	DN-121	EN-b0	MN-v3	RNX-50	Average
ViT-B	MIM [4]	97.2	61.8	40.4	42.3	54.9	65.8	44.4	51.7	57.2	51.3	50.8	51.7	38.2	50.9
	PNAPO [38]	99.1	83.2	62.1	65.8	74.7	83.0	64.0	67.4	70.0	67.6	68.5	63.0	56.3	68.8
	TGR [45]	99.1	86.2	63.8	69.9	81.1	94.4	70.4	79.1	79.8	75.3	83.1	79.9	60.0	76.9
	GNS [51]	99.9	89.5	68.1	74.0	82.4	93.6	72.2	77.8	76.7	76.0	79.1	76.0	62.0	77.3
	FPR	99.2	**92.5**	**73.0**	**78.3**	**88.5**	**98.2**	**77.7**	**87.3**	**87.6**	**82.0**	**90.3**	**87.4**	**68.3**	**84.3**
CaiT-S	MIM [4]	68.5	98.8	50.0	54.9	68.2	75.3	54.8	61.0	67.8	59.8	61.0	55.7	47.1	60.3
	PNAPO [38]	69.8	88.2	58.1	61.0	70.3	73.8	60.1	64.8	69.7	64.3	65.5	61.7	53.2	64.4
	TGR [45]	90.1	100.0	76.0	81.3	91.2	97.8	80.6	86.9	86.6	84.3	87.7	82.1	69.1	84.5
	GNS [51]	91.2	100.0	78.6	82.2	91.6	97.6	81.1	86.6	87.1	84.9	89.6	81.9	70.8	85.3
	FPR	**93.8**	100.0	**81.6**	**85.6**	**95.0**	**98.8**	**86.8**	**87.4**	**88.7**	**88.6**	**90.6**	**82.9**	**76.8**	**88.1**
PiT-T	MIM [4]	30.0	36.1	38.5	43.3	51.0	75.0	55.5	69.5	68.7	60.8	64.5	71.6	40.4	54.2
	PNAPO [38]	39.5	50.9	49.6	54.9	61.6	85.4	65.0	78.7	76.7	72.2	77.9	81.0	50.8	64.9
	TGR [45]	37.9	47.7	46.9	54.1	62.5	87.4	65.1	82.8	82.0	75.5	81.8	88.8	53.7	66.6
	GNS [51]	37.1	44.1	46.4	51.9	59.4	84.4	63.3	79.9	78.0	71.5	76.3	84.3	48.6	63.5
	FPR	**44.3**	**56.1**	**57.1**	**62.4**	**69.5**	**91.3**	**72.2**	**88.7**	**86.4**	**80.4**	**87.6**	**91.1**	**60.3**	**72.9**
DeiT-B	MIM [4]	86.0	82.4	67.0	69.5	79.7	87.7	65.3	69.0	70.5	70.1	71.4	61.4	58.2	72.2
	PNAPO [38]	83.4	87.7	72.4	74.1	82.2	86.0	71.9	74.9	74.7	74.9	76.1	68.0	66.3	76.4
	TGR [45]	92.4	97.1	81.2	84.3	92.6	97.9	83.5	86.0	85.2	85.0	89.9	84.5	74.6	87.2
	GNS [51]	92.1	98.0	80.5	84.2	91.0	98.0	79.8	83.3	80.7	83.5	86.7	81.3	70.8	85.4
	FPR	**94.6**	**98.4**	**83.8**	**86.7**	**95.5**	**99.5**	**87.0**	**89.8**	**89.2**	**88.1**	**93.9**	**86.8**	**79.8**	**90.2**

Compatibility Experiment

Method	ViTs	CNNs
SIA [36]	73.4	69.1
SIA+FPR	**86.6**	**90.4**
BSR [32]	69.4	66.9
BSR+FPR	**84.8**	**89.9**
VTM [34]	63.1	57.3
VTM+FPR	**84.0**	**81.7**
GRA [49]	73.6	68.3
GRA+FPR	**86.1**	**85.3**