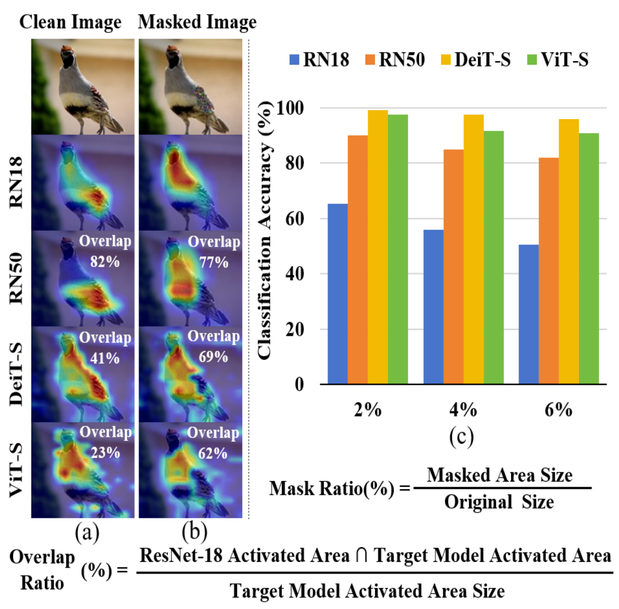
Boosting Adversarial Transferability via Ensemble Non-Attention

Yipeng Zou[1], Qin Liu[1]*, Jie Wu[2,3], Yu Peng[4] , Guo Chen[1], Hui Zhou[1], Guanghui Ye[1]
[1]Hunan University, [2]China Telecom Cloud Computing Research Institute, [3]Temple University,
[4]University of Electronic Science and Technology of China

AAAI-26 / IAAI-26 / EAAI-26

MOTIVATION

- The non-attention areas of CNNs are probably to be the focus of ViTs, and vice versa.
- The masked image induced high ratios of attention overlaps across both homogeneous and heterogeneous models



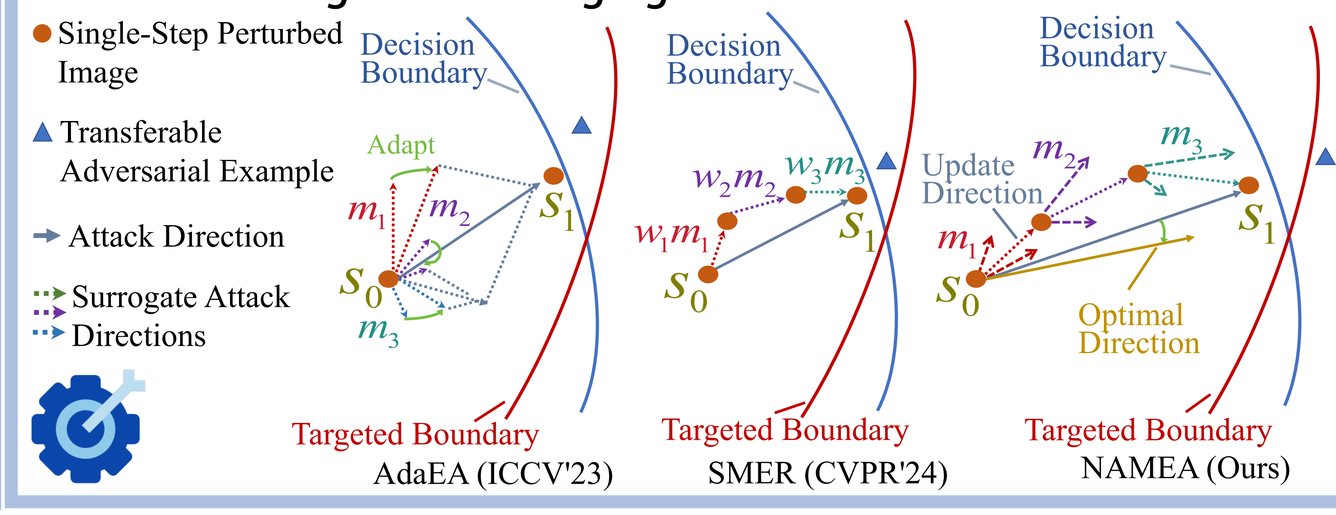
$$\text{Mask Ratio}(\%) = \frac{\text{Masked Area Size}}{\text{Original Size}}$$

$$\text{Overlap Ratio}\ (\%) = \frac{\text{ResNet-18 Activated Area} \cap \text{Target Model Activated Area}}{\text{Target Model Activated Area Size}}$$

Challenge: How to make the best of individual model while stabilizing update direction among ensemble models?

Solution & Contribution

- We propose a novel ensemble attack, NAMEA, which ensures stable update direction and model diversity at once, exhibiting superior cross-architecture transferability.
- NAMEA decouples gradients from non-attention and attention regions and integrates meta-learning into iterative optimization for efficient gradient merging.

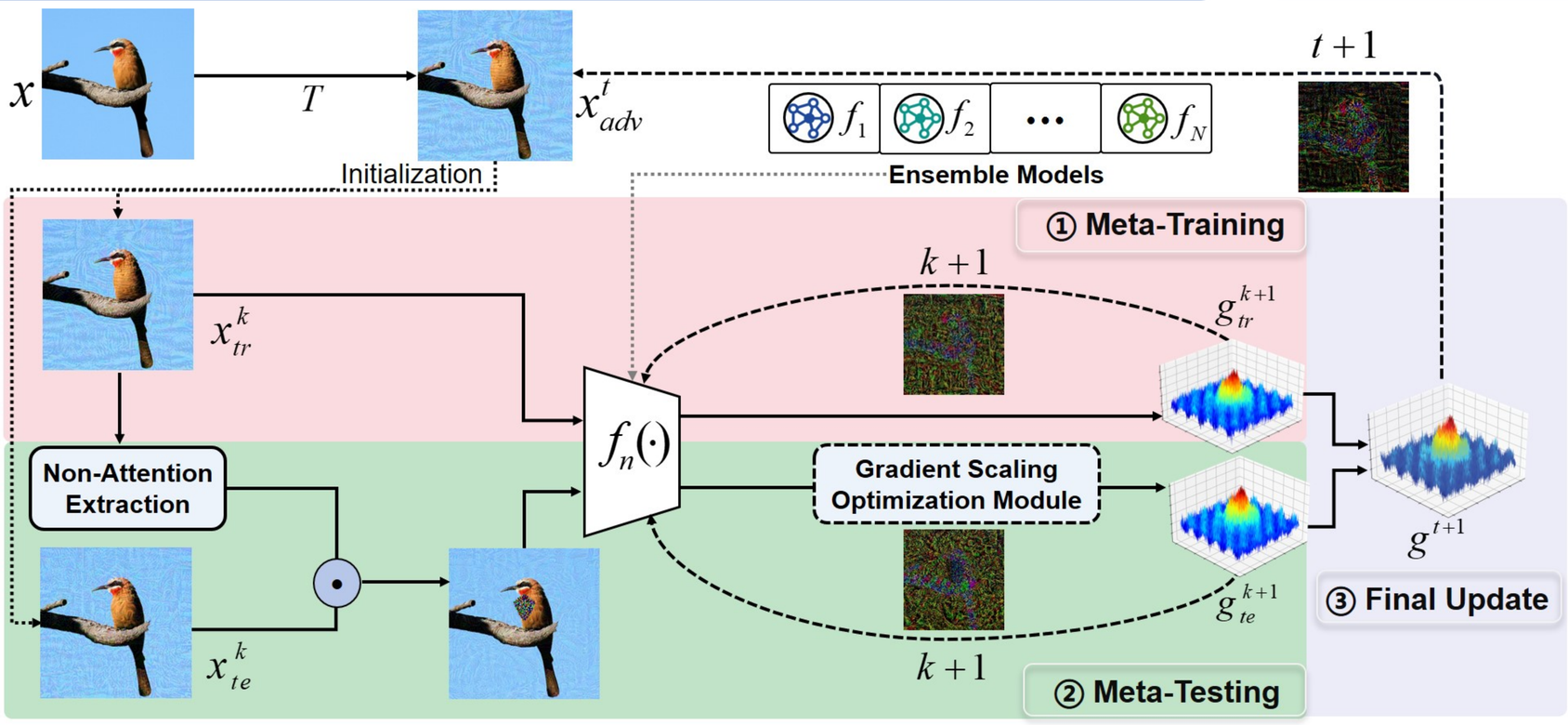


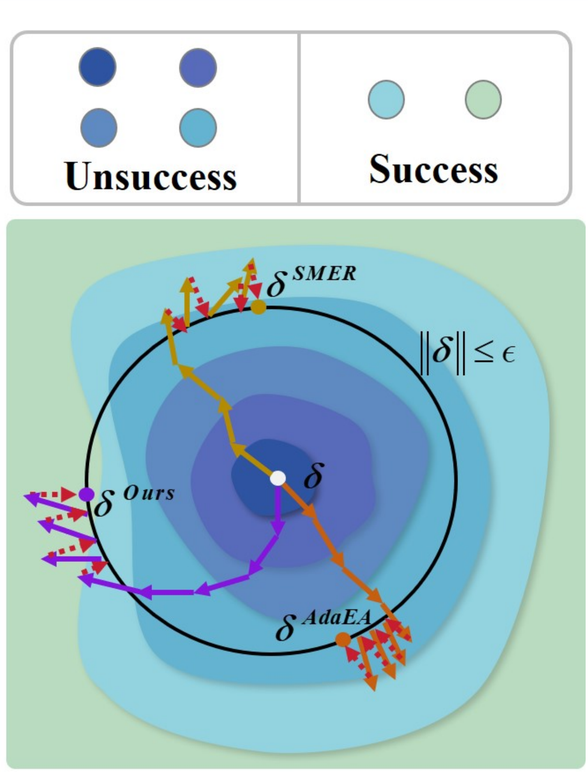
METHOD: Overview of NAMEA

① Attention meta-training updates the gradient g_{tr}^{k+1} based on model's attention areas.
② Non-attention meta-testing updates the gradient g_{te}^{k+1} based on model's non-attention areas.
③ Final update merges the gradients from meta-training and meta-testing steps to obtain the final gradient g^{t+1}.

Non-Attention Extraction
- Generate attention maps (Grad-CAM) & Create non-attention masks
- Mask attention areas & Obtain gradients from non-attention regions

Gradient Scaling Optimization Module
- CNN: emphasize intermediate-layer gradients
- ViT: suppress low-magnitude gradients





RESULTS

Base	Attack	ViTs										CNNs								
		ViT-B	PiT-B	CaiT-S	ViS	DeiT-B	TNT-S	LeViT	ConV	Swin-B	**Avg.**	RN50	RN152	DN201	DN169	VGG16	VGG19	WRN101	BiT50	**Avg.**
I-FGSM	Ens	16.0	10.7	25.0	17.2	26.8	28.4	17.9	30.8	9.9	20.3	22.7	13.0	30.3	34.7	35.5	33.6	22.6	28.2	27.6
	SVRE	13.1	11.5	21.9	19.2	23.2	28.2	19.3	23.9	10.1	18.9	29.0	16.2	34.8	39.5	42.1	28.9	26.0	32.5	32.4
	AdaEA	25.1	17.6	39.2	27.5	40.4	40.2	28.8	42.7	15.6	30.8	38.7	21.1	47.0	50.1	53.0	48.4	34.5	39.6	41.6
	CWA	27.8	10.6	41.5	16.7	49.9	46.7	21.1	48.8	11.7	30.5	12.9	6.9	20.8	22.6	34.3	32.1	15.2	25.5	21.3
	SMER	27.4	16.4	42.6	26.0	43.9	44.7	27.7	48.9	15.4	32.6	33.2	18.4	43.1	45.7	50.0	48.4	31.4	39.6	38.7
	CSA	27.5	17.8	42.1	27.3	43.0	48.6	30.4	43.7	16.0	32.9	36.6	20.4	49.7	50.2	51.9	51.0	36.2	42.3	42.3
	Ours	**43.0**	**25.5**	**61.2**	**38.0**	**63.0**	**61.2**	**42.9**	**63.6**	**21.8**	**46.7**	**46.2**	**26.4**	**55.8**	**58.5**	**64.4**	**60.7**	**43.8**	**52.1**	**51.0**
MI-FGSM	Ens	34.0	24.9	48.5	34.7	51.7	49.8	38.7	51.2	20.6	39.3	43.4	26.5	52.8	53.6	55.2	52.9	39.6	46.4	46.3
	SVRE	31.3	24.2	43.2	35.1	44.6	50.5	38.9	46.5	19.3	37.1	49.6	30.5	58.1	60.5	59.2	58.0	45.3	50.6	51.5
	AdaEA	41.2	25.5	56.3	38.8	59.4	55.8	41.4	58.7	21.7	44.3	49.0	29.2	56.2	59.9	59.5	57.8	43.7	52.2	47.6
	CWA	35.1	18.4	53.5	28.6	55.4	56.7	38.9	58.2	18.0	40.3	37.7	22.1	48.7	51.4	58.8	53.6	37.5	44.9	44.3
	SMER	45.4	26.8	61.2	40.2	63.0	61.8	47.5	64.9	25.1	48.4	51.0	31.5	59.8	61.0	66.0	61.7	47.9	55.1	54.3
	CSA	48.5	29.8	61.3	45.4	63.2	66.2	49.0	64.2	27.1	50.5	52.0	32.0	60.4	62.3	66.1	63.6	49.6	53.8	55.0
	Ours	**56.6**	**34.9**	**72.6**	**51.1**	**74.5**	**72.5**	**59.0**	**74.5**	**32.8**	**58.7**	**59.7**	**39.7**	**69.9**	**69.9**	**73.3**	**72.2**	**57.1**	**63.4**	**63.2**
DI-MI-FGSM	Ens	42.5	38.3	56.6	50.5	56.1	62.0	53.7	59.3	31.4	50.0	59.5	41.9	70.1	71.5	71.4	70.0	60.4	63.5	63.5
	SVRE	45.2	43.1	65.4	57.0	62.5	70.5	63.0	63.3	32.2	55.8	66.8	49.1	76.7	77.8	78.2	75.4	67.7	71.7	70.4
	AdaEA	47.7	36.6	67.2	52.6	66.2	69.3	56.0	66.4	30.8	54.8	60.5	42.1	69.3	72.4	72.8	70.9	58.5	64.9	63.9
	CWA	53.6	44.4	73.6	57.9	71.1	79.4	66.1	73.7	33.1	61.4	64.3	47.8	76.7	77.9	79.6	78.5	65.0	72.9	70.3
	SMER	66.9	57.2	81.9	70.6	82.0	85.4	75.7	83.2	46.0	72.1	75.3	59.2	85.0	85.7	84.0	82.7	75.5	80.2	78.5
	CSA	54.8	46.2	68.1	60.1	69.2	73.4	63.2	68.8	38.8	60.3	63.2	48.2	76.7	75.2	76.5	73.4	66.2	72.7	69.0
	Ours	**72.7**	**63.6**	**85.9**	**77.8**	**86.5**	**89.2**	**80.8**	**86.6**	**54.1**	**77.5**	**80.9**	**68.4**	**88.6**	**89.4**	**87.7**	**87.6**	**81.1**	**86.0**	**83.7**