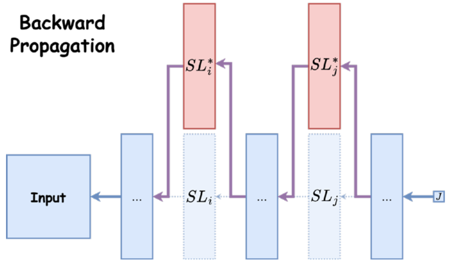
Rethinking the Backward Propagation for Adversarial Transferability

Xiaosen Wang[1], Kangheng Tong[2], Kun He[2]

[1]Huawei Singular Security Lab, [2]School of Computer Science, Huazhong University of Science and Technology

Introduction of transfer-based adversarial attacks

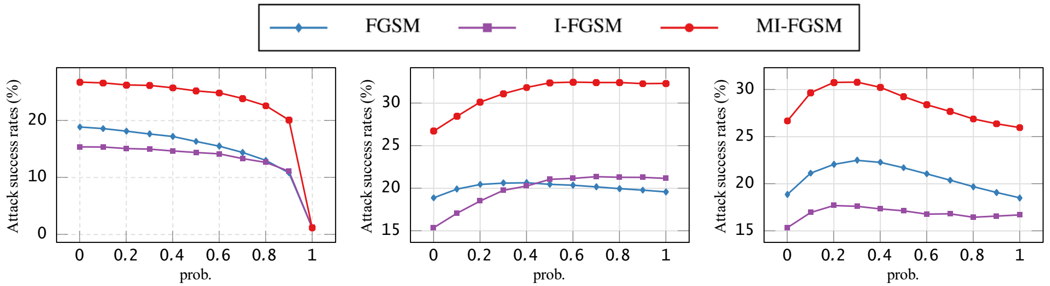
- Dnns' susceptibility to adversarial examples, which are carefully crafted by adding imperceptible perturbations to natural examples, has raised significant concerns regarding their security.
- Transfer-based attacks generate adversarial examples on the surrogate model to fool the target models.
- We find that the gradient truncation introduced by non-linear layers undermines the transferability and modify the backward propagation so as to generate more transferable adversarial examples.



Assumption & Verification

Assumption: The truncation of gradient introduced by non-linear layers in the backward propagation process decays the adversarial transferability.

Verification: Randomly masking the gradient decays the transferability while recovering the gradient of ReLU or max-pooling layers improves the transferability.



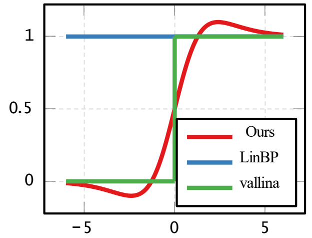
(a) Randomly mask the gradient (b) Recover the gradient of ReLU (c) Recover the gradient of max-pooling

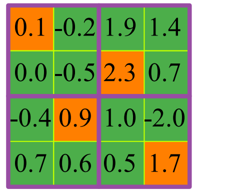
Methodology

To diminish the probability of gradient truncation, we modify the gradient calculation for the ReLU activation function and max-pooling in the backward propagation procedure as follows:

- Use the derivative of SiLU to calculate the gradient of ReLU during the backward propagation process, i.e., $\frac{\partial z_{i+1}}{\partial z_i} = \sigma(z_i) \cdot \left(1 + z_i \cdot \left(1 - \sigma(z_i)\right)\right)$.
- Use the softmax function to calculate the gradient within each window w of the max-pooling operation:

$$\left[\frac{\partial z_{k+1}}{\partial z_k}\right]_{i,j,w} = \frac{e^{t \cdot z_{k,i,j}}}{\sum_{v \in w} e^{t \cdot v}}$$





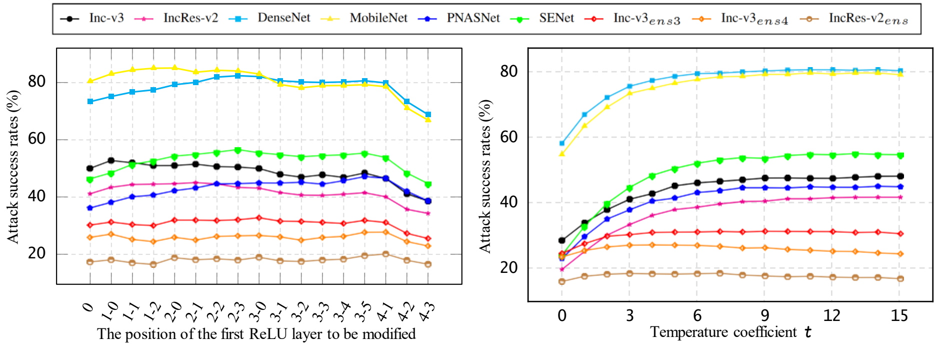
Experiment results

Untargeted attack success rates (%) of various adversarial attacks on nine models when generating the adversarial examples on ResNet-50 w/wo various model-related methods.

Attacker	Method	Inc-v3	IncRes-v2	DenseNet	MobileNet	PNASNet	SENet	Inc-v3$_{ens3}$	Inc-v3$_{ens4}$	IncRes-v2$_{ens}$
PGD	N/A	12.52	9.70	25.82	32.20	13.18	13.82	7.64	7.60	4.14
	LinBP	13.52	10.28	27.60	34.36	14.16	15.12	8.32	7.88	4.20
	Ghost	13.18	9.72	25.78	32.50	12.80	13.68	8.12	7.90	4.48
	BPA	**26.24**	**27.06**	**47.98**	**58.22**	**34.08**	**31.42**	**15.52**	**14.06**	**8.78**
MI-FGSM	N/A	19.74	15.32	37.02	43.42	21.16	23.02	11.46	10.08	5.96
	LinBP	20.28	15.24	36.84	44.44	20.66	23.28	10.92	9.52	5.48
	Ghost	19.88	15.34	36.44	43.20	21.84	24.06	11.54	10.30	6.00
	BPA	**36.88**	**29.98**	**61.10**	**68.58**	**45.98**	**43.06**	**21.44**	**17.68**	**11.94**
VMI-FGSM	N/A	37.20	29.58	58.20	62.20	40.88	38.86	21.14	17.62	11.10
	LinBP	36.18	28.86	55.40	62.46	38.38	39.14	19.20	17.18	10.92
	Ghost	36.94	29.75	58.32	62.16	41.32	38.96	21.18	17.58	11.20
	BPA	**51.60**	**43.00**	**74.08**	**78.74**	**59.54**	**54.74**	**32.88**	**30.04**	**20.18**
ILA	N/A	16.08	13.8	31.28	42.62	19.72	25.16	8.76	7.70	4.62
	LinBP	17.08	14.54	32.74	44.40	20.16	27.08	8.44	7.92	4.54
	Ghost	16.56	14.08	31.80	41.90	20.12	25.98	8.84	7.84	4.76
	BPA	**29.70**	**25.06**	**50.84**	**61.52**	**38.84**	**41.20**	**15.30**	**12.36**	**8.30**
SSA	N/A	33.52	26.38	50.86	60.26	30.94	30.78	17.06	14.52	8.78
	LinBP	35.70	28.08	53.76	63.52	32.32	34.18	18.64	16.10	9.36
	Ghost	33.52	25.92	51.31	60.50	30.96	30.02	17.16	14.74	8.74
	BPA	**50.16**	**40.68**	**70.90**	**78.86**	**51.64**	**47.86**	**29.52**	**26.50**	**18.30**

Ablation Study

We perform parameter studies on two crucial aspects: the position of the first ReLU layer to be modified and the temperature coefficient t for max-pooling.



(a) Attack success rate (%) of BPA using MI-FGSM by modifying the ReLU layers starting from the i-th layer. Here 3-0 indicates the first ReLU layer in the third stage

(b) Attack success rate (%) of BPA using MI-FGSM with various temperature coefficients ($0 \leq t \leq 15$) in Eq. (3) for the max-pooling layer

Conclusion & Limitation

- It is the first work that proposes and empirically validates the detrimental effect of gradient truncation on adversarial transferability. This finding sheds new light on improving adversarial transferability and provides new directions to boost model robustness.
- We propose a model-related attack called BPA to mitigate the negative impact of gradient truncation and enhance the relevance of gradient between the loss function and the input.
- Extensive experiments on ImageNet dataset demonstrate that BPA could significantly boost various untargeted and targeted transfer-based attacks.

Trustworthy-AI-Group