TRM-UAP: Enhancing the Transferability of Data-Free Universal Adversarial Perturbation via Truncated Ratio Maximization

Yiran Liu, Xin Feng, Yunlong Wang, Wu Yang, Di Ming*

School of Computer Science and Engineering, Chongqing University of Technology

重慶理工大學 CHONGQING UNIVERSITY OF TECHNOLOGY

ICCV23 PARIS

Introduction

➢ Background

❖ Adversarial Example (AE): crafted by adding tiny perturbations deliberately to benign samples.

❖ Universal Attack: try to find the universal adversarial perturbation $\boldsymbol{v}$ which maximizes the classification loss $\mathcal{L}$ over the data distribution $\mathbb{D}$:

$$\max_{\boldsymbol{v}\sim\mathbb{S}} \mathop{\mathbb{E}}_{(\boldsymbol{x},\boldsymbol{y})\sim\mathbb{D}} [\mathcal{L}(f(\boldsymbol{v}+\boldsymbol{x}),\boldsymbol{y})] \quad \text{s.t. } \|\boldsymbol{v}\|_p \le \epsilon$$

➢ Contribution

❖ We propose a novel data-free universal attack method to craft universal adversarial perturbations without utilizing real samples during training.

❖ Our proposed TRM-UAP enhances the transferability of UAPs via ratio maximization, truncation strategy, and curriculum optimization.

Truncated Ratio Maximization

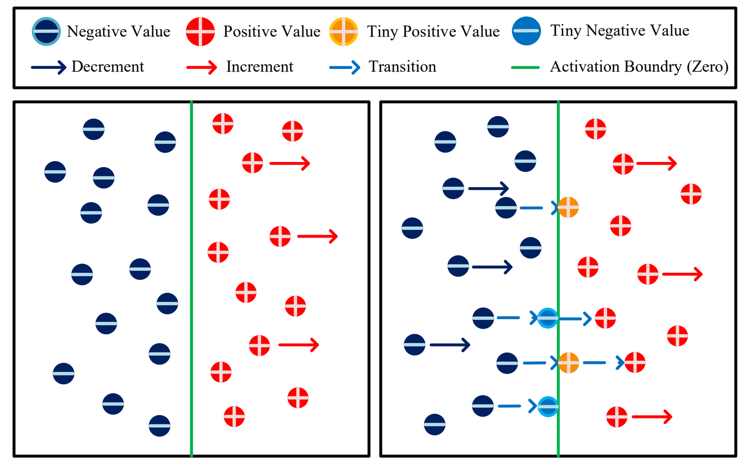
➢ Preliminary

❖ The output of the i-th convolution layer: ❑ $\mathcal{C}^{(i)}(\boldsymbol{v})$

❖ CNN/Pos./Neg. Activations: ❑ $\mathcal{A}^{(i)}(\boldsymbol{v}) = \text{Activation}(\mathcal{C}^{(i)}(\boldsymbol{v}))$

❑ $\mathcal{C}_+^{(i)}(\boldsymbol{v}) = \max(\mathcal{C}^{(i)}(\boldsymbol{v}), 0)$ ❑ $\mathcal{C}_-^{(i)}(\boldsymbol{v}) = \min(\mathcal{C}^{(i)}(\boldsymbol{v}), 0)$

➢ Maximizing the Ratio of Activations



(a) Positive Maximization (b) Ratio Maximization

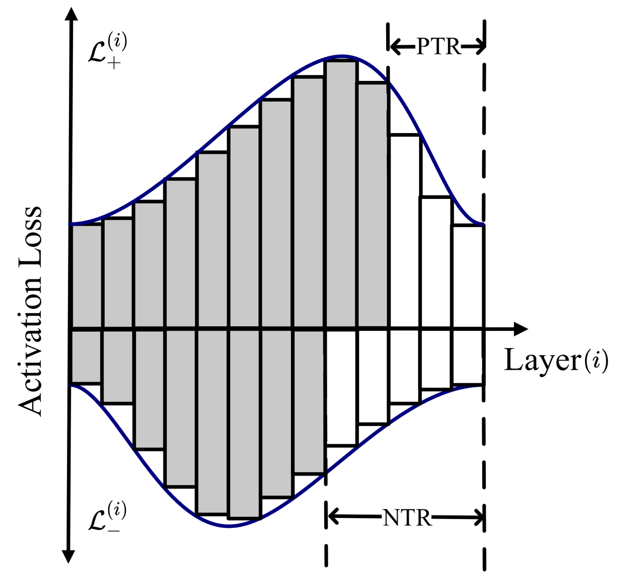
❖ Positive Maximization:

$$\max_{\boldsymbol{v}} \ \|\mathcal{A}^{(i)}(\boldsymbol{v})\|_2, \quad \text{for } i = 1, 2, \cdots, L$$
$$\text{s.t. } \|\boldsymbol{v}\|_\infty \le \epsilon$$

❖ Ratio Maximization:

$$\max_{\boldsymbol{v}} \ \frac{\|\mathcal{C}_+^{(i)}(\boldsymbol{v})\|_2}{\|\mathcal{C}_-^{(i)}(\boldsymbol{v})\|_2}, \quad \text{for } i = 1, 2, \cdots, L$$
$$\text{s.t. } \|\boldsymbol{v}\|_\infty \le \epsilon$$

➢ Truncated Ratio Maximization



❖ Truncated Positive Activation Loss:

❑ $\mathcal{L}_+^{(i)}(\boldsymbol{v}) = \tau \ (i > l')$ ❑ $PTR = \lfloor (L - l')/L \rfloor \%$

❖ Truncated Negative Activation Loss:

❑ $\mathcal{L}_-^{(i)}(\boldsymbol{v}) = \tau \ (i > l'')$ ❑ $NTR = \lfloor (L - l'')/L \rfloor \%$

where $\mathcal{L}_+^{(i)}(\boldsymbol{v}) = \|\mathcal{C}_+^{(i)}(\boldsymbol{v})\|_2$ and $\mathcal{L}_-^{(i)}(\boldsymbol{v}) = \|\mathcal{C}_-^{(i)}(\boldsymbol{v})\|_2$.

❖ Rescaled Ratio Loss: ❑ $\mathcal{L}_\alpha^{(i)}(\boldsymbol{v}) = \dfrac{\mathcal{L}_+^{(i)}(\boldsymbol{v})}{(\mathcal{L}_-^{(i)}(\boldsymbol{v}))^\alpha}$

Curriculum Optimization Algorithm

➢ The Overall Loss Function

❖ To maximize the ratio of activations in convolution layers, the overall loss function of truncated ratio maximization is reformulated as:

$$\mathcal{L}(\boldsymbol{v}) = \sum_{i=1}^{L} \log \mathcal{L}_\alpha^{(i)}(\boldsymbol{v})$$
$$\propto \sum_{i=1}^{l'} \log \mathcal{L}_+^{(i)}(\boldsymbol{v}) - \alpha \cdot \sum_{i=1}^{l''} \log \mathcal{L}_-^{(i)}(\boldsymbol{v})$$

❖ Craft the universal adversarial perturbation $\boldsymbol{v}$ satisfying:

$$\max_{\boldsymbol{v}} \ \sum_{i=1}^{l'} \log \|\mathcal{C}_+^{(i)}(\boldsymbol{v})\|_2 - \alpha \cdot \sum_{i=1}^{l''} \log \|\mathcal{C}_-^{(i)}(\boldsymbol{v})\|_2 \quad \text{s.t.} \quad \|\boldsymbol{v}\|_\infty \le \epsilon$$

➢ Curriculum Optimization

❖ To improve the diversity of inputs, the set of artificial images is generated from simple to difficult pattern with the increase of training iterations:

$$D_1 \prec D_2 \prec \cdots \prec D_n, \ D_t = \{\boldsymbol{x} | \boldsymbol{x} \sim P(\theta_0, t)\}.$$

❖ For t-th iteration, our curriculum optimization algorithm is maximizing:

$$\mathcal{L}_t = \frac{1}{|D_t|} \sum_{\boldsymbol{x}\in D_t} \left(\sum_{i=1}^{l'} \log \mathcal{L}_+^{(i)}(\boldsymbol{v}+\boldsymbol{x}) - \alpha \cdot \sum_{i=1}^{l''} \log \mathcal{L}_-^{(i)}(\boldsymbol{v}+\boldsymbol{x}) \right)$$

Main Result

➢ Experimental Setting

❖ Models: ❑ AlexNet ❑ VGG16 ❑ VGG19 ❑ ResNet152 ❑ GoogleNet

❖ Methods: ❑ FFF ❑ AAA ❑ GD-UAP ❑ PD-UA ❑ Cosine-UAP

➢ Fooling Rates of Data-Free Universal Attacks

Attack	AlexNet	VGG16	VGG19	ResNet152	GoogleNet	Average
FFF	80.92	47.10	43.62	-	56.44	-
AAA	89.04	71.59	72.84	60.72	75.28	73.89
GD-UAP	85.24	90.01	87.34	45.96	45.87	64.65
PD-UA	-	70.69	64.98	46.39	67.12	-
Cosine-UAP	91.07	89.48	86.81	65.35	**87.57**	84.08
TRM-UAP(Ours)	**93.53±0.07**	**94.30±0.15**	**91.35±0.30**	**67.46±0.35**	85.32±0.04	**86.39**

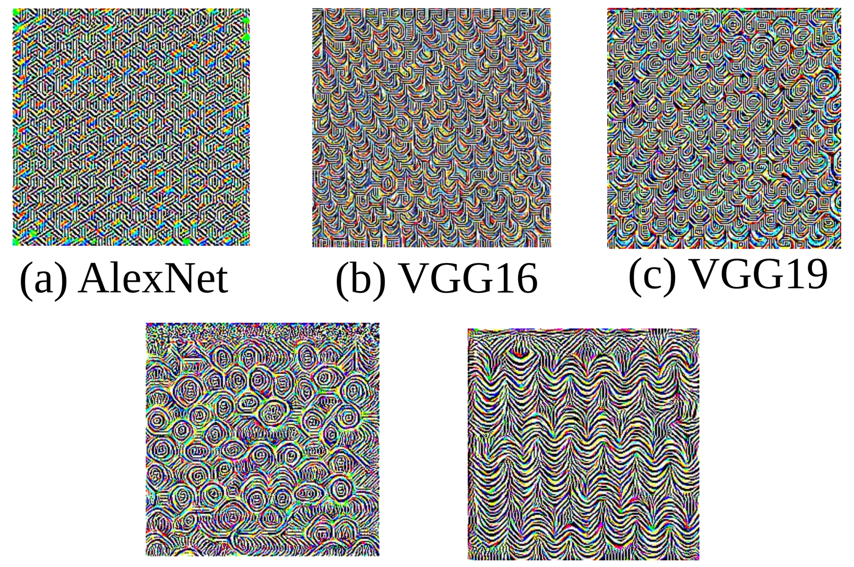
➢ Transferable Results of TRM-UAP (White-Box & Black-Box)

	AlexNet	VGG16	VGG19	ResNet152	GoogleNet
AlexNet	93.53±0.07	60.10±0.24	57.08±0.15	27.31±0.30	32.70±0.22
VGG16	47.53±0.51	94.30±0.12	89.68±0.14	61.43±0.40	53.95±0.59
VGG19	46.01±0.44	89.82±0.15	91.35±0.30	47.19±0.66	46.48±0.78
ResNet152	53.56±0.75	77.20±0.35	73.30±0.41	67.46±0.35	57.54±0.50
GoogleNet	60.10±1.16	79.66±0.95	79.98±1.06	58.85±1.94	85.32±0.04

Visualization & Analysis

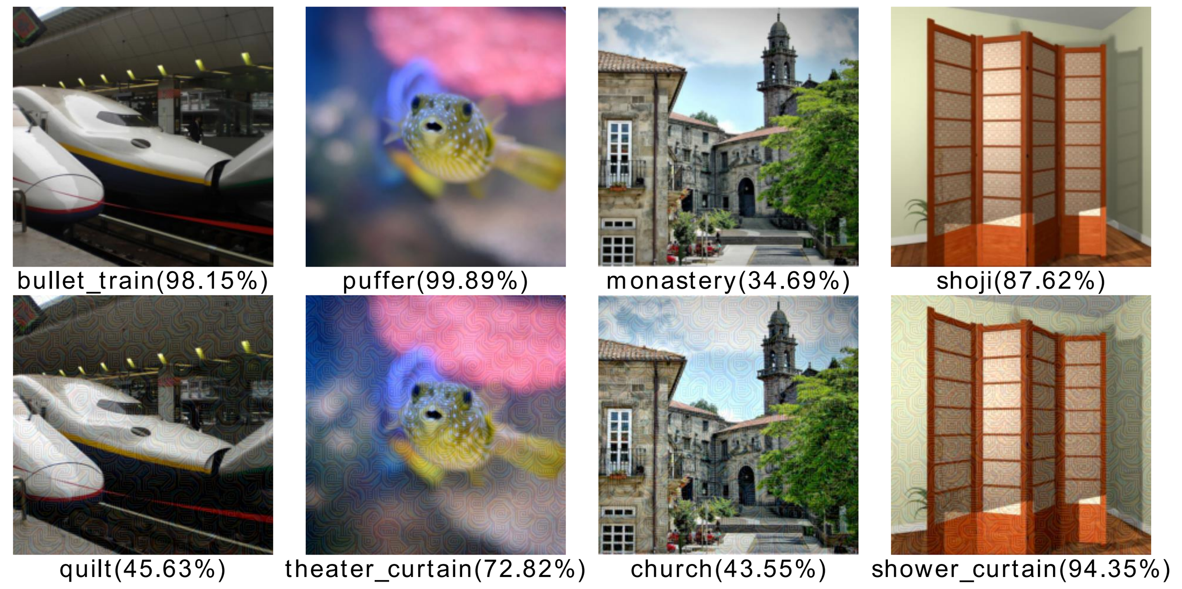
➢ UAPs and AEs Crafted by TRM-UAP

❖ UAPs of different CNN models



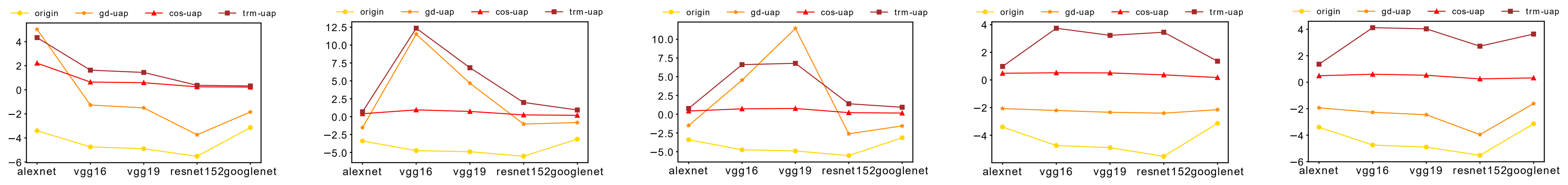
(a) AlexNet (b) VGG16 (c) VGG19 (d) ResNet152 (e) GoogleNet

❖ Original Examples vs Adversarial Examples



➢ Evaluating the Transferability

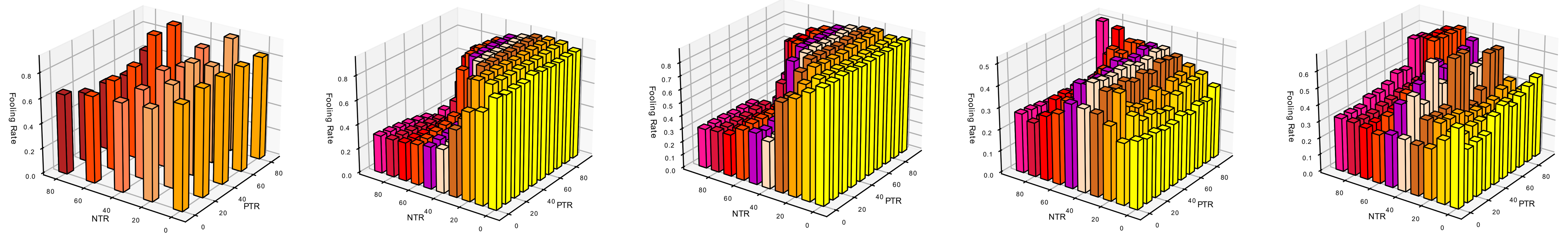
❖ Comparison between GD-UAP, Cosine-UAP and TRM-UAP on the logit loss



(a) AlexNet (b) VGG16 (c) VGG19 (d) ResNet152 (e) GoogleNet

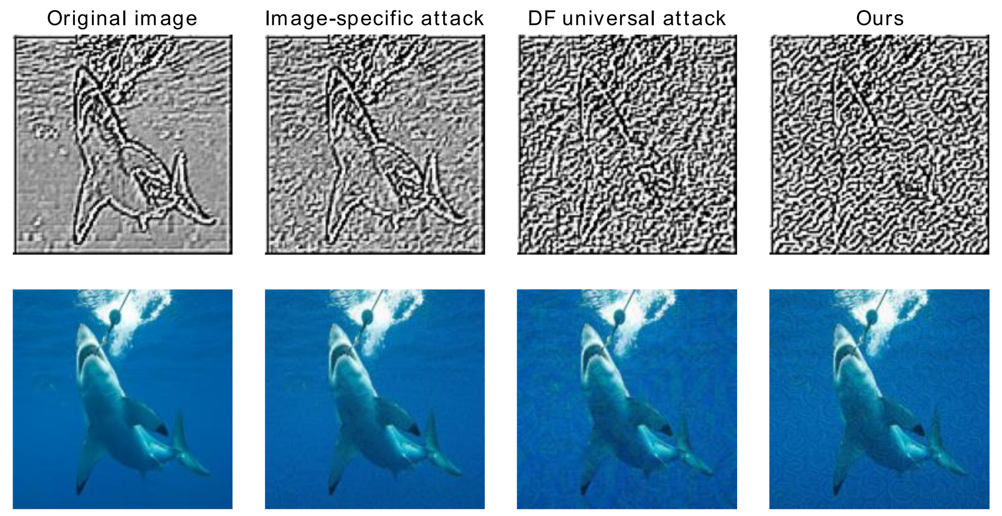
➢ Parameter Study

❖ Fooling rate with respect to positive truncation rate and negative truncation rate

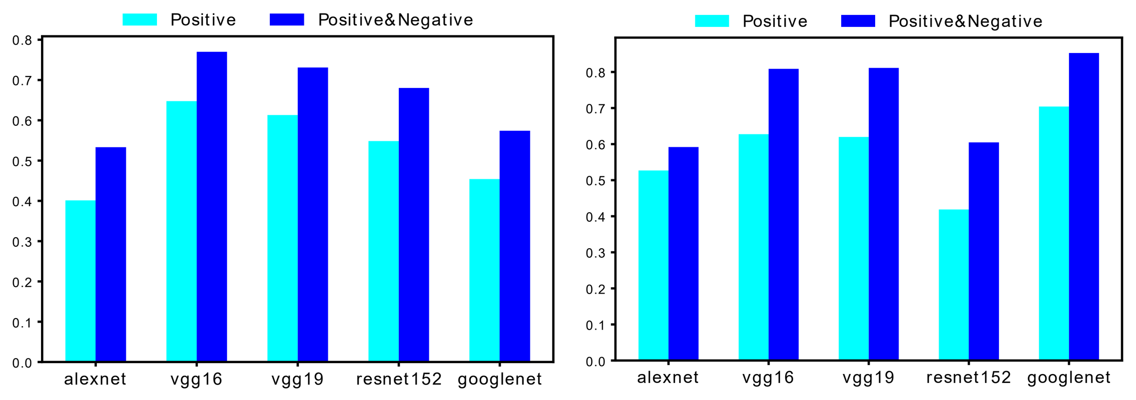


(a) AlexNet (b) VGG16 (c) VGG19 (d) ResNet152 (e) GoogleNet

❖ Feature map of different attacks



❖ Positive Maximization vs Ratio Maximization



(a) ResNet152 (b) GoogleNet