Improving Transferable Targeted Attacks with Feature Tuning Mixup

Kaisheng Liang, Xuelong Dai, Yanjie Li, Dong Wang, Bin Xiao

Presenter: Kaisheng Liang

March 2025

CVPR

THE HONG KONG POLYTECHNIC UNIVERSITY
香港理工大學

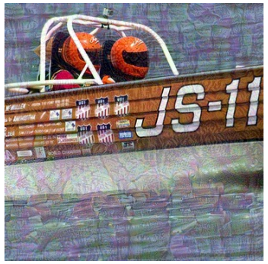
Overview

- Background
- Motivation
- Methodology
- Experiment
- Conclusion

Background

- Adversarial Examples
  - Small perturbations that fool DNNs
- Non-targeted Attacks
  - Push input to any incorrect class
- Targeted Attacks
  - Force input into a specific target class

Benign Image
True label: speedboat
Prediction: speedboat



Non-targeted attack
True label: speedboat
Prediction: candy store

Targeted attack
True label: speedboat
Target label: unicycle
Prediction: unicycle

Transferable Targeted Attack

- Adversarial Transferability
  - Adversarial examples generated on one DNN can fool other DNNs
  - No access to other target models' outputs, architectures, parameters
- Optimization Objective
  - Targeted transfer-based attack

$$\arg\min_{x^{\mathrm{adv}}} \mathcal{L}(F(x^{\mathrm{adv}}), y_t), \quad s.t. \quad \|x - x^{\mathrm{adv}}\|_\infty \leq \epsilon$$

Adversarial Transferability

Targeted Attack on ResNet-50

True Label: balloon
Target Label: pillow
Prediction of ResNet-50: pillow
Prediction of VGG-16: pillow
Prediction of Inception-v3: pillow
Prediction of DenseNet-121: pillow
Prediction of ViT: pillow

Targeted Attack on ResNet-50

True Label: speedboat
Target Label: unicycle
Prediction of ResNet-50: unicycle
Prediction of VGG-16: unicycle
Prediction of Inception-v3: unicycle
Prediction of DenseNet-121: unicycle
Prediction of ViT: unicycle

Attack Scenario

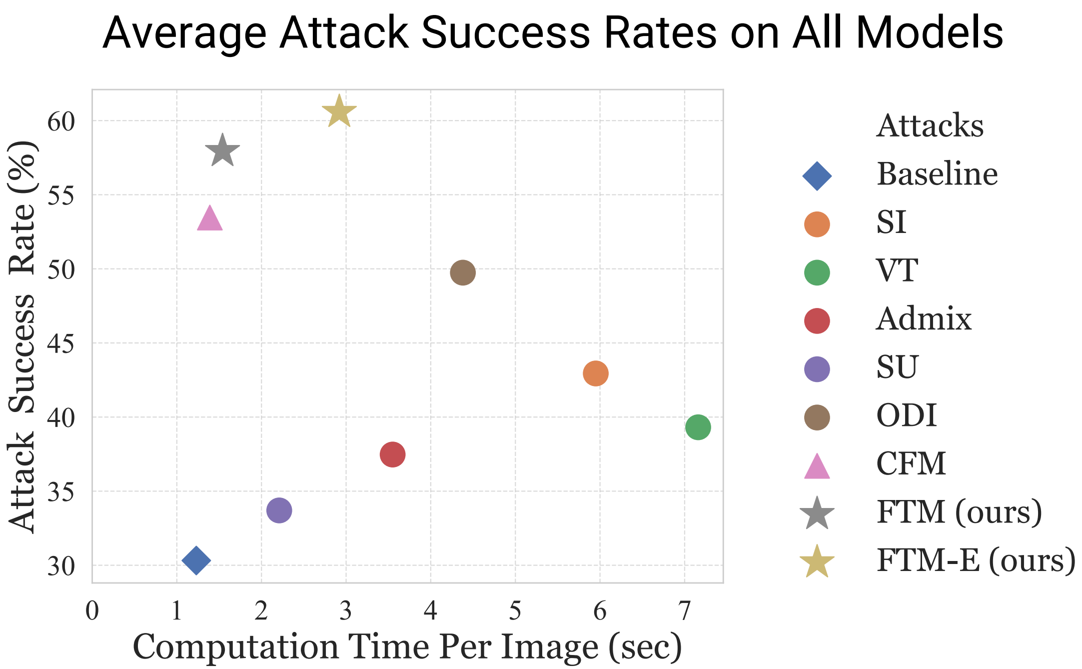
- Given
  - A surrogate model F, a benign image x, a target label y_t
- Goal
  - Generate x^{adv} that is misclassified by the model F as the target label y_t
  - Transferability: other models also classify x^{adv} as the target label y_t
- Constraint
  - Perturbation follows the ℓ_∞-norm constraint
  - No additional training dataset

Existing Methods

- Gradient-based Attack
  - Momentum Iterative FGSM (MI-FGSM)
- Input Transformation
  - Scale Invariant (SI)
  - Admix
  - Object-based Diverse Input (ODI)
- Feature Augmentation
  - Clean Feature Mixup (CFM)

Challenges

- Limited Transferability
- Computational Cost



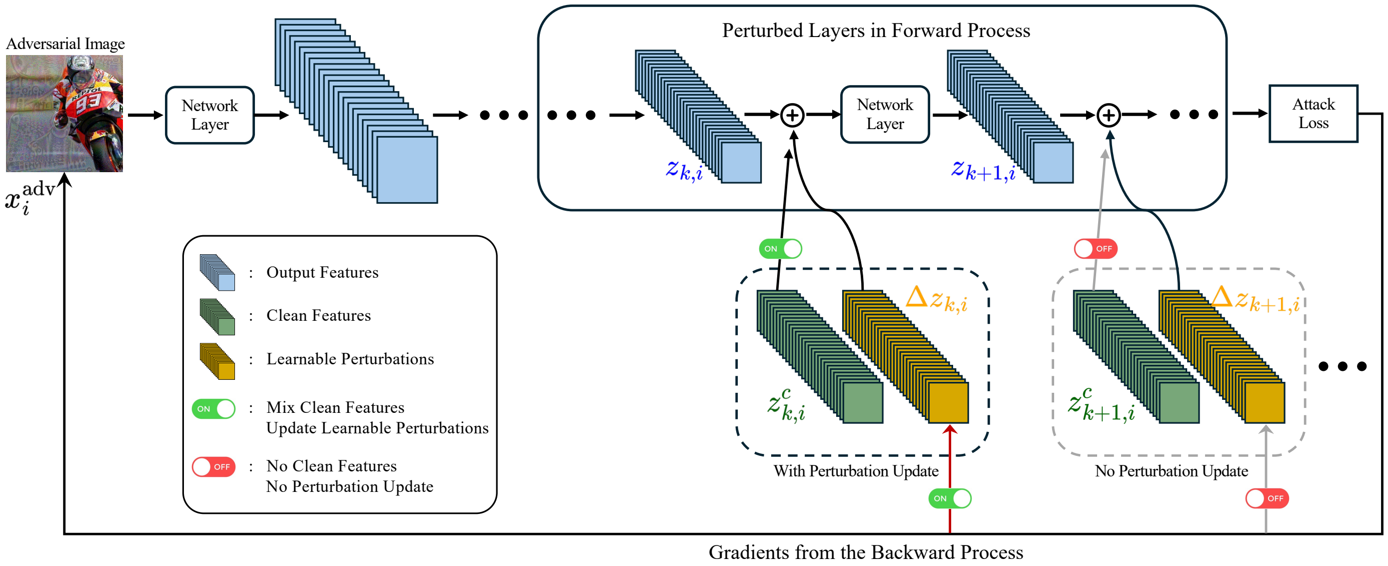
Motivation

- Feature Space Instead of Image Space
- Disrupting the Surrogate Model
  - Improves the transferability of adversarial examples
  - CFM is attack-agnostic in nature
  - Explicitly optimizing feature perturbations remains to be explored

$$\mathcal{L}\big(F(x_i^{\mathrm{adv}}; \Delta z_i), y_t\big) > \mathcal{L}\big(F(x_i^{\mathrm{adv}}), y_t\big)$$

Methodology

- Introducing Learnable Feature Perturbation
  - Perturbation Injection
  - Perturbation Update



x_i^{adv} : i-th iteration of x^{adv}
$\Delta z_{k,i}$: k-th layer of Δz_i

Momentum Stochastic Update

- Momentum
  - Using perturbations in previous iterations
- Stochastic
  - Randomly select a small portion of layers for update

Feature Tuning Mixup (FTM)

Optimization Objective

$$\min_{x_i^{\mathrm{adv}}} \max_{\Delta z_i} \mathcal{L}\left(F(x_i^{\mathrm{adv}}; \Delta z_i), y_t\right)$$

Perturbation Injection

$$\bar{z}_{k,i} = z_{k,i} + \beta \|z_{k,i}\| \cdot \frac{\Delta z_{k,i}}{\|\Delta z_{k,i}\| + \bar{\epsilon}},$$

$$z'_{k,i} = \begin{cases} (1 - \alpha_{k,i}) \odot \bar{z}_{k,i} + \alpha_{k,i} \odot z^{c}_{k,i}, & \tau_k < p, \\ \bar{z}_{k,i}, & \text{else}, \end{cases}$$

$$\Delta z'_i = \{(\Delta z_{k,i}, \tau_k) \mid k = 1, \ldots, n\},$$

Perturbation Update

$$g'_{i+1}, g^{\Delta z'} = \nabla_{x_i^{\mathrm{adv}}, \Delta z'_i} \mathcal{L}\left(F(x_i^{\mathrm{adv}}; \Delta z'_i), y_t\right),$$

$$\Delta z_{k,i+1} = \begin{cases} \Delta z_{k,i} + g_k^{\Delta z'}, & \tau_k < p, \\ \Delta z_{k,i}, & \text{else}, \end{cases}$$

MI-FGSM Framework

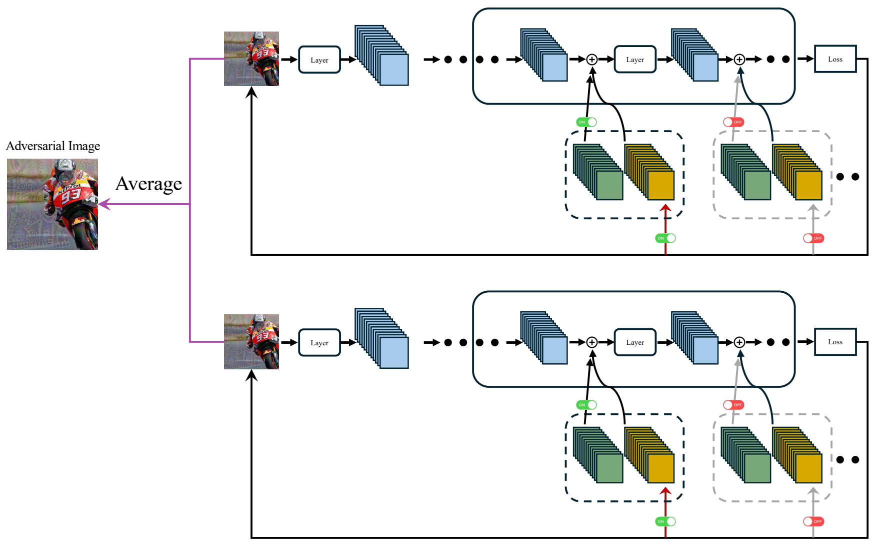
$$\tilde{g}_{i+1} = \mu \cdot g'_i + g'_{i+1} / \|g'_{i+1}\|_1$$

$$x_{i+1}^{\mathrm{adv}} = x_i^{\mathrm{adv}} - \eta \cdot \mathrm{sign}\left(\tilde{g}_{i+1}\right),$$

$$x_{i+1}^{\mathrm{adv}} = \mathrm{Clip}_{x,\epsilon}\left(x_{i+1}^{\mathrm{adv}}\right)$$

Ensemble of Surrogate Variants (FTM-E)

- FTM
  - Efficiently perturb the surrogate model
- FTM-E
  - Multiple copies of the given surrogate model
  - Independently apply FTM to each copy



Efficiency Analysis

- No time-consuming transformation
- No need for extra backpropagation
  - Maintain the one-step optimization of FGSM
- Our method works for different models
  - CNNs, ViTs, etc.

Experimental Setup

- Dataset
  - ImageNet-compatible dataset
  - 1000 images, from NeurIPS 2017
- Iteration Number
  - 300
- Logit-based Loss Function
- Baseline Framework
  - MI-TI-FGSM
- Models
  - 15 pretrained DNNs, consisting of 10 CNNs and 5 ViTs

Experimental Results

- Adversarial Transferability
  - Our method significantly improves the attack success rates

Attack	LeViT ⇒											
	VGG-16	RN-50	Inc-v3	DN-121	IR-v2	Inc-v4	Xcep	ViT	LeViT*	ConViT	Twins	PiT
DI	1.6	2.5	4.2	2.7	1.8	2.6	2.3	0.6	**100**	4.2	9.3	10.3
RDI	3.0	3.5	5.4	4.1	3.6	4.6	2.3	1.1	**100**	7.9	13.8	22.1
RDI-Admix	6.5	8.3	9.9	8.8	5.8	7.2	5.5	3.3	**100**	10.7	23.3	30.9
RDI-Admix$_5$	5.3	8.1	13.9	12.4	9.9	8.4	7.2	8.0	99.9	20.6	30.0	47.2
RDI-SI	3.4	6.3	10.6	10.0	6.4	5.4	5.1	4.3	**100**	18.1	24.1	38.4
RDI-VT	5.3	7.2	12.0	10.1	8.3	8.8	8.9	6.3	99.9	18.7	27.2	40.5
RDI-ODI	21.0	25.0	40.9	38.3	25.7	31.2	26.3	15.3	98.7	34.1	43.8	66.1
RDI-CFM	27.3	30.3	39.8	39.0	23.6	30.1	27.2	18.4	**100**	45.5	63.8	75.7
RDI-FTM	41.3	41.9	56.9	54.2	39.8	46.5	42.2	31.1	99.9	63.2	77.5	86.7
RDI-FTM-E	**50.1**	**52.4**	**62.8**	**63.1**	**48.4**	**53.2**	**49.2**	**40.3**	99.9	**72.2**	**86.2**	**93.0**

Effectiveness and Efficiency

- For Different Surrogates
- Low Computational Cost

Source	Attack	ViT	LeViT	ConViT	Twins	PiT	Avg.	Computation time per image (sec)
RN-50	DI	0.2	3.5	0.4	1.5	1.7	1.5	1.37
	RDI	0.7	13.2	1.7	6.1	7.0	5.7	1.23
	RDI-SI	2.9	29.4	6.3	15.5	17.9	14.4	5.95
	RDI-VT	2.9	28.1	5.2	15.0	14.0	13.0	7.16
	RDI-Admix	1.3	22.5	2.5	8.5	8.4	8.6	3.55
	RDI-SU	0.8	16.9	2.3	6.9	7.8	6.9	2.21
	ODI	5.1	37.0	10.7	20.1	29.1	20.4	4.38
	RDI-CFM	4.3	46.1	8.9	25.2	24.7	21.8	1.39
	RDI-FTM	5.9	52.9	10.8	32.4	31.5	26.7	1.54
	RDI-FTM-E	**6.8**	**58.6**	**13.6**	**35.2**	**34.9**	**29.8**	2.92
Inc-v3	DI	0.1	0.3	0.0	0.0	0.1	0.1	2.01
	RDI	0.2	1.8	0.2	0.4	0.7	0.7	1.76
	RDI-SI	0.3	4.1	0.9	0.7	3.2	1.8	8.11
	RDI-VT	0.4	5.2	0.8	1.6	1.8	2.0	10.5
	RDI-Admix	0.1	4.1	0.6	1.4	1.4	1.5	4.92
	RDI-SU	0.2	2.0	0.3	1.3	1.0	1.0	2.36
	ODI	0.8	12.4	1.7	3.5	6.7	5.0	6.42
	RDI-CFM	2.1	21.9	3.2	6.1	11.6	8.9	2.13
	RDI-FTM	2.4	25.0	4.5	10.0	15.3	11.5	2.35
	RDI-FTM-E	**3.8**	**32.7**	**6.8**	**12.7**	**20.4**	**15.3**	4.37

Evaluation on LLMs

- Surrogate Model
  - ViT
- Target Models
  - Qwen2-VL
  - Llama-3.2
  - Claude-3.5
  - GPT-4o

Response	Qwen2-VL	Llama-3.2	Claude-3.5	GPT-4o	Avg
Total	100	100	100	100	100 %
Refuse to Answer	0	10	0	0	2.50 %
Uncertain	1	5	1	0	1.75 %
Attack Failed	52	42	54	73	55.25 %
Attack Succeeded	47	43	45	27	40.50 %

- We randomly select 100 images for evaluation.
- Each adversarial image is generated using FTM-E on ViT.
- We use the prompt “Is this image a photo of { target label } ? Yes or No?” to obtain predictions of LLMs.

Conclusion

- Incorporating attack-specific feature perturbations can efficiently enhance transferable targeted attacks.
- FTM uses a momentum stochastic update, maintaining computational efficiency while improving attack transferability.
- FTM significantly outperforms state-of-the-art attacks across various source and target models.