Boosting Adversarial Attacks with Momentum

Yinpeng Dong[1], Fangzhou Liao[1], Tianyu Pang[1], Hang Su[1], Jun Zhu[1], Xiaolin Hu[1], Jianguo Li[2]

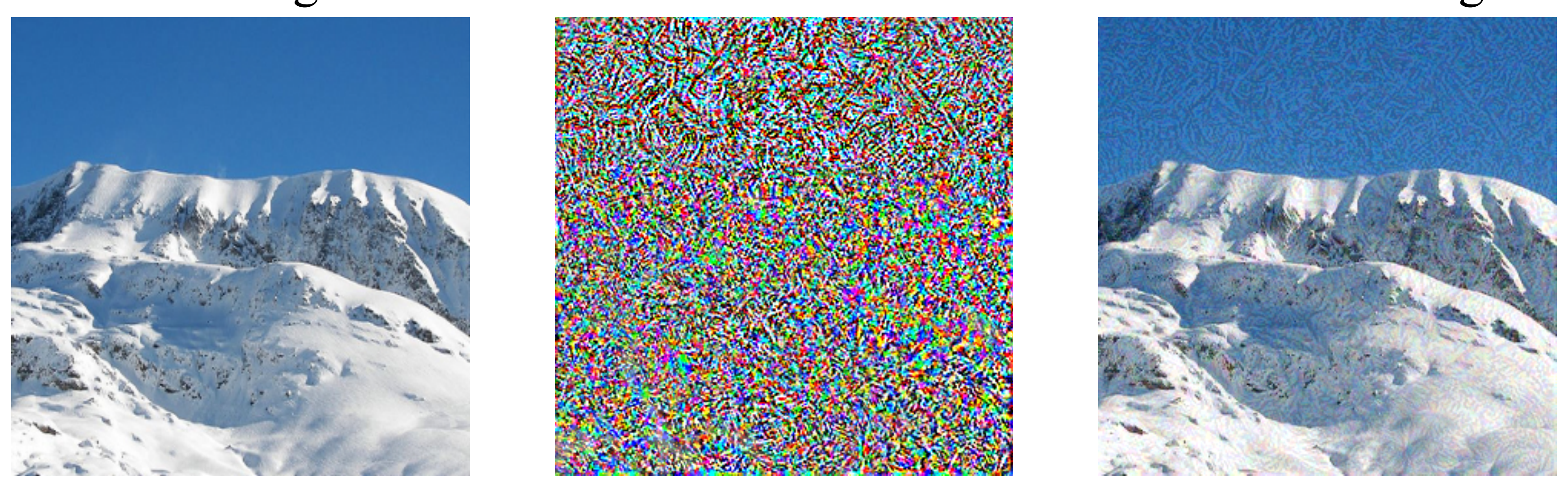
[1]Department of Computer Science and Technology, Tsinghua University, [2]Intel Labs China

SALT LAKE CITY • JUNE 18-22

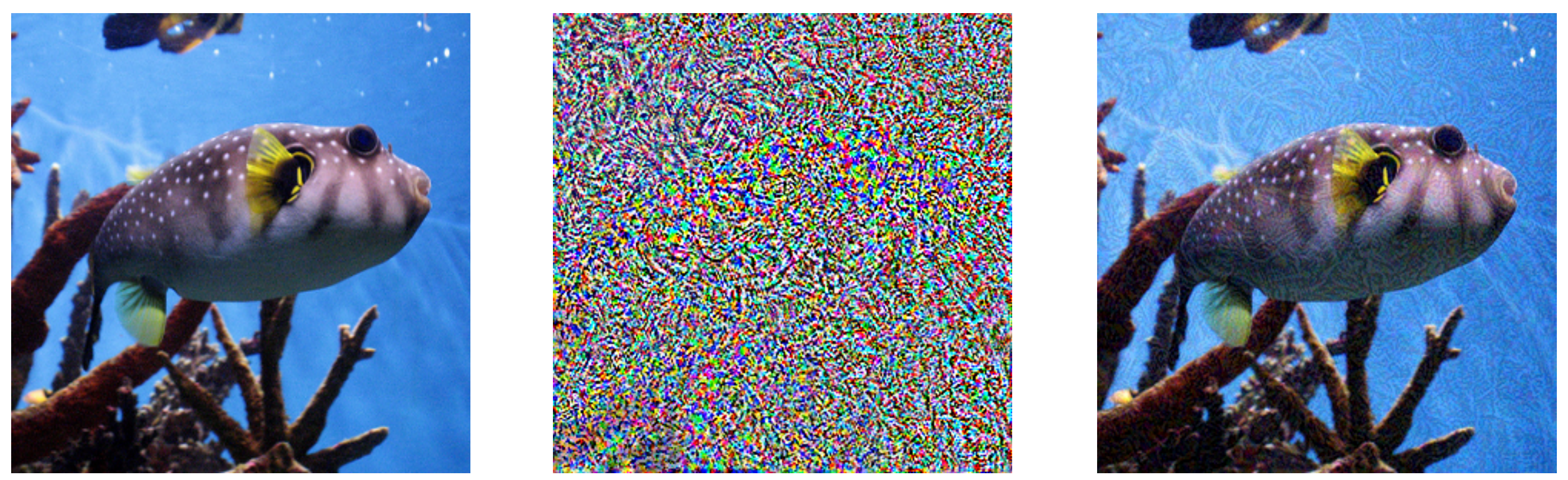
Introduction

- **Adversarial examples** are crafted by adding small, human-imperceptible noises to legitimate examples, but make a model output attacker-desired inaccurate predictions.
- **Adversarial attacks:**
  - Identify the robustness of deep learning models.
  - Provide more varied training data (i.e., adversarial training).

Real Images | Perturbations | Adversarial Images



Alps: 94.39% | Dog: 99.99%



Puffer: 97.99% | Crab: 100.00%

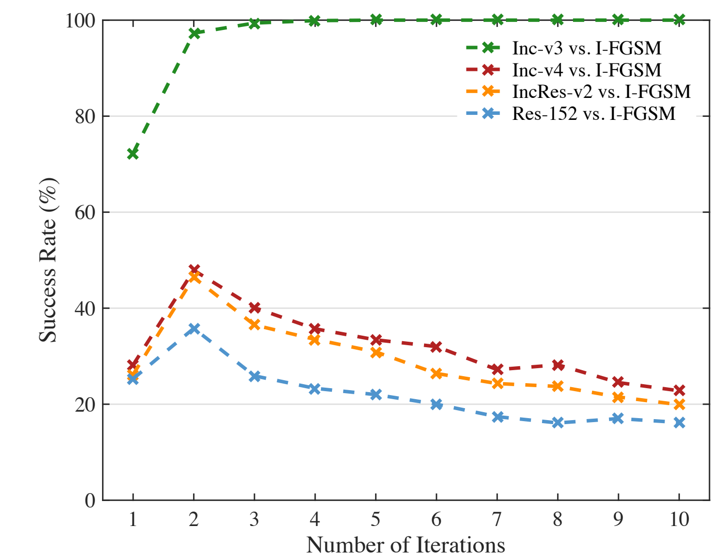
- **Generating adversarial examples:**
  - Constrained optimization problem:
    $$\operatorname*{argmax}_{x^*} J(x^*, y) \quad s.t. \|x^* - x\|_\infty \le \epsilon$$
  - Fast gradient sign method (FGSM, Goodfellow et al., 2015):
    $$x^* = x + \epsilon \cdot \mathrm{sign}(\nabla_x J(x, y))$$
  - Iterative fast gradient sign method (I-FGSM, Kurakin et al., 2016):
    $$x_0^* = x, \quad x_{t+1}^* = \mathrm{clip}(x_t^* + \alpha \cdot \mathrm{sign}(\nabla_x J(x_t^*, y)))$$
  - Optimization-based method (Carlini and Wagner, 2017):
    $$\operatorname*{argmin}_{x^*} \lambda \cdot d(x^*, x) - J(x^*, y)$$
- **Transferability**
  - The adversarial examples generated for one model can also fool another model (Liu et al., 2017).
  - Black-box attacks: how to generate more efficient adversarial examples for a black-box model (challenge).

Motivation

- **The trade-off between the attack ability and transferability**

1. FGSM: more transferable adversarial examples; low success rates for the white-box models. (**Reason: linear assumption may not hold for large distortion; "underfit" the model.**)
2. I-FGSM: high success rates for white-box models; poor transferability. (**Reason: drop into poor local maxima; "overfit" the model.**)

The Success rates when attacking Inc-v3, Inc-v4, IncRes-v2 and Res-152 by I-FGSM with different number of iterations. The adversarial examples are generated for Inc-v3.



- **Optimization with Momentum (Polyak, 1964)**
  - Accelerate gradient descent
  - Escape from poor local minima and maxima
  - Stabilize update directions of stochastic gradient descent

Methodology

- **Momentum Iterative Fast Gradient Sign Method (MI-FGSM)**

$$x_0^* = x, \quad x_{t+1}^* = \mathrm{clip}(x_t^* + \alpha \cdot \mathrm{sign}(\nabla_x J(x_t^*, y)))$$

Momentum

$$x_0^* = x, g_0 = 0$$
$$g_{t+1} = \mu \cdot g_t + \frac{\nabla_x J(x_t^*, y)}{\|\nabla_x J(x_t^*, y)\|_1}$$
$$x_{t+1}^* = \mathrm{clip}(x_t^* + \alpha \cdot \mathrm{sign}(g_{t+1}))$$

where g_t gathers the gradients of the first t iterations.

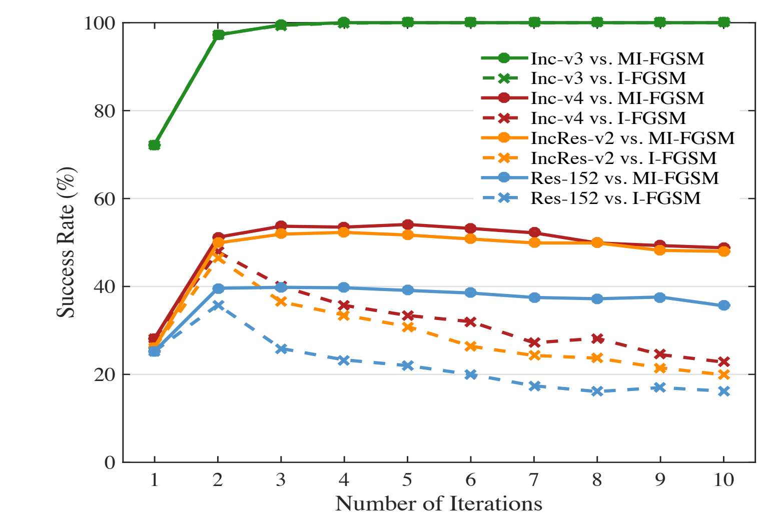
- **Attacking an ensemble of models**
  - The adversarial examples generated for multiple models are more transferable (Liu et al., 2017).
  - We propose to attack multiple models whose logits are fused together and then use MI-FGSM to attack the ensemble model.
    $$l(x) = \sum_{i=1}^{K} w_i l_i(x)$$
- **Extension: MI-FGSM can be extended to targeted attacks and L_2 norm bound attacks**

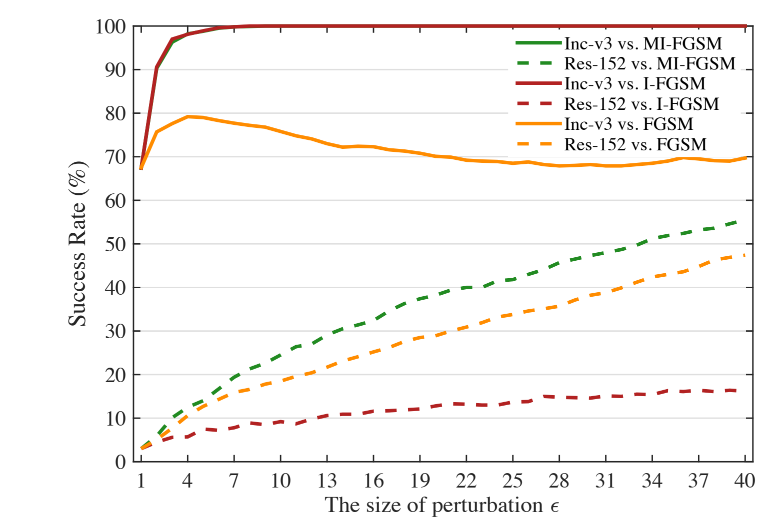
Experiments

- **Attacking a single model**

	Attack	Inc-v3	Inc-v4	IncRes-v2	Res-152	Inc-v3$_{ens3}$	Inc-v3$_{ens4}$	IncRes-v2$_{ens}$
Inc-v3	FGSM	72.3*	28.2	26.2	25.3	11.3	10.9	4.8
	I-FGSM	**100.0***	22.8	19.9	16.2	7.5	6.4	4.1
	MI-FGSM	**100.0***	**48.8**	**48.0**	**35.6**	**15.1**	**15.2**	**7.8**
Inc-v4	FGSM	32.7	61.0*	26.6	27.2	13.7	11.9	6.2
	I-FGSM	35.8	**99.9***	24.7	19.3	7.8	6.8	4.9
	MI-FGSM	**65.6**	**99.9***	**54.9**	**46.3**	**19.8**	**17.4**	**9.6**
IncRes-v2	FGSM	32.6	28.1	55.3*	25.8	13.1	12.1	7.5
	I-FGSM	37.8	20.8	**99.6***	22.8	8.9	7.8	5.8
	MI-FGSM	**69.8**	**62.1**	99.5*	**50.6**	**26.1**	**20.9**	**15.7**
Res-152	FGSM	35.0	28.2	27.5	72.9*	14.6	13.2	7.5
	I-FGSM	26.7	22.7	21.2	**98.6***	9.3	8.9	6.2
	MI-FGSM	**53.6**	**48.9**	**44.7**	98.5*	**22.1**	**21.7**	**12.9**

- **Ablation studies**





- **Attacking an ensemble of models**

	Ensemble method	FGSM		I-FGSM		MI-FGSM	
		Ensemble	Hold-out	Ensemble	Hold-out	Ensemble	Hold-out
-Inc-v3	Logits	**55.7**	**45.7**	**99.7**	**72.1**	**99.6**	**87.9**
	Predictions	52.3	42.7	95.1	62.7	97.1	83.3
	Loss	50.5	42.2	93.8	63.1	97.0	81.9
-Inc-v4	Logits	**56.1**	**39.9**	**99.8**	**61.0**	**99.5**	**81.2**
	Predictions	50.9	36.5	95.5	52.4	97.1	77.4
	Loss	49.3	36.2	93.9	50.2	96.1	72.5
-IncRes-v2	Logits	**57.2**	**38.8**	**99.5**	**54.4**	**99.5**	**76.5**
	Predictions	52.1	35.8	97.1	46.9	98.0	73.9
	Loss	50.7	35.2	96.2	45.9	97.4	70.8
-Res-152	Logits	**53.5**	**35.9**	99.6	**43.5**	99.6	**69.6**
	Predictions	51.9	34.6	**99.9**	41.0	**99.8**	67.0
	Loss	50.4	34.1	98.2	40.1	98.8	65.2

Conclusion

- **We propose a broad class of momentum-based iterative methods for generating more transferable adversarial examples.**
- **We propose to attack an ensemble of models whose logits are fused.**
- **Our method won the first places in both of the NIPS 2017 Non-target Adversarial Attack and Targeted Adversarial Attack competitions.**
- **Code available at:**