ICLR 2023

ILA-DA: Improving Transferability of Intermediate Level Attack with Data Augmentation

Chiu Wai Yan
cwyan@connect.ust.hk

Tsz-Him Cheung
thcheungae@connect.ust.hk

Dit-Yan Yeung
dyyeung @cse.ust.hk

From top to bottom:
Clean example, I-FGSM, ILA-DA

Introduction

- DL models are vulnerable to adversarial attacks /examples
- Transfer white-box adversarial attacks to form black-box attacks
  - I-FGSM, PGD, etc.

Transfer-based Attacks

- DIM (Xie et al., 2019)
  - Random resizing and zero-padding
- TIM (Dong et al., 2019)
  - Image translation
- SIM (Lin et al., 2020)
  - Scaling the pixel values
- Admix (Wang et al., 2021)
  - Mixing gradients from different label classes

Predefined and fixed image transformations

We propose ILA-DA, which

- consists of 3 novel augmentation techniques
- outperforms SOTA attacks on 9 undefended models and 6 defended models
- can be incorporated into other transfer-based attacks to further strengthen its attack transferability

Intermediate-Level Attack (ILA)

ILA (Intermediate Level Attack) (Huang et al., 2019)

- Given 3 Inputs:
  - A clean example $\boldsymbol{x}$
  - An existing adversarial example $\boldsymbol{x}'$
  - An example to be fine-tuned $\boldsymbol{x}''$
- F_l: The output of a DNN model F up to the l^{th} layer
- ILA (Projection) Loss: Maximize the intermediate feature discrepancy

$$L(\mathbf{x}, \mathbf{x}', \mathbf{x}'') = -\Delta \mathbf{y}''_l \cdot \Delta \mathbf{y}'_l$$

, where

$$\Delta \mathbf{y}'_l = F_l(\mathbf{x}') - F_l(\mathbf{x})$$
$$\Delta \mathbf{y}''_l = F_l(\mathbf{x}'') - F_l(\mathbf{x})$$

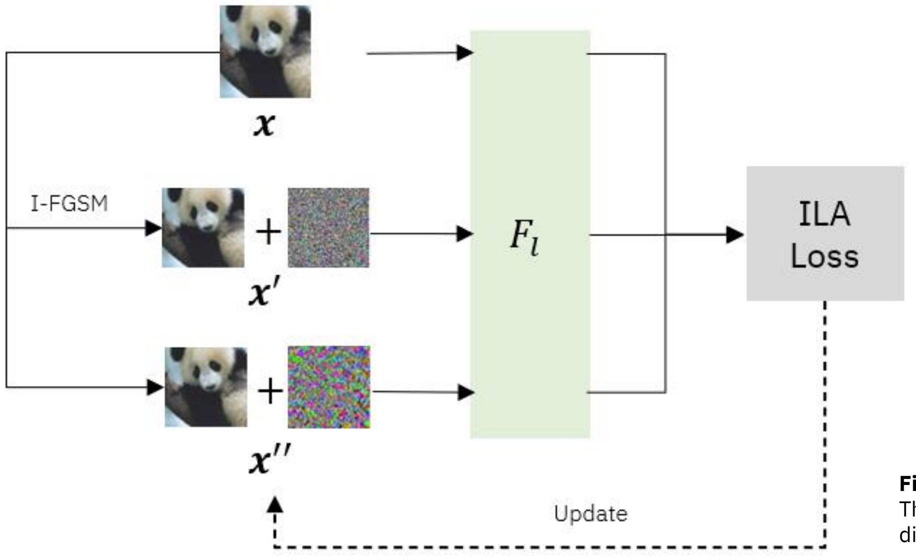

Fig. 1
The schematic diagram of ILA.

Our Method

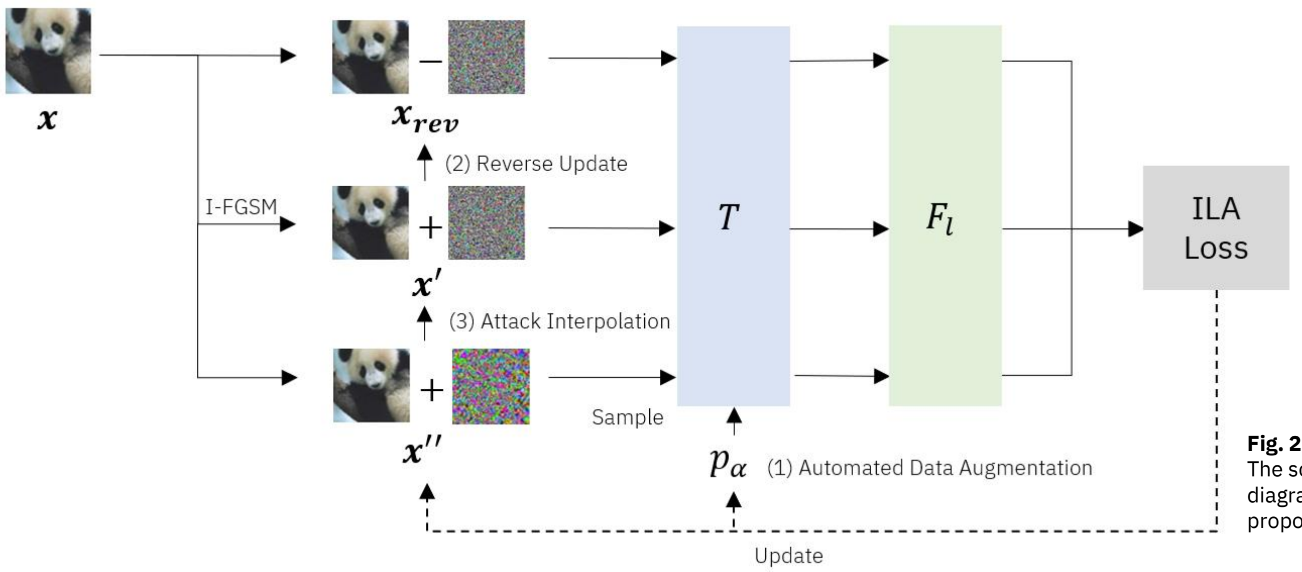

Fig. 2
The schematic diagram of the proposed ILA-DA.

Automated Data Augmentation

- Learn the most effective augmentation from a set of candidates
- Sample a transformation function T from a learnable distribution p_α
- The probability parameter α is updated using the Gumbel Softmax reparameterization trick

Reverse Adversarial Update

- $x_{rev} = T_{adv}(x) = 2x - x'$
- Increase the confidence of x being recognized as the right class
- Minus the adversarial example to boost the confidence

Attack Interpolation

- $x'_{t+1} \leftarrow \lambda x''_t + (1-\lambda)x'_t$
- Strengthen the reference attack with the finetuned attack
- Adaptive λ based on the norm of the feature maps discrepancy

$$\lambda = \frac{\|\Delta \mathbf{y}''_l\|_2}{\|\Delta \mathbf{y}''_l\|_2 + \|\Delta \mathbf{y}'_l\|_2}$$

Results

Undefended Models

- ImageNet, CIFAR-10, CIFAR-100
- 9 pretrained models
- $\epsilon = 8/255$ and $13/255$
- I-FGSM$_{10}$ → ILA-DA$_{50}$

Defended Models

- ImageNet
- 6 defences from NIPS-2018 Competition
- $\epsilon = 16/255$
- I-FGSM$_{10}$ → ILA-DA$_{500}$

Table 1: Attack success rates of ImageNet adversarial examples on nine undefended models, generated from ResNet50 with $\epsilon = 8/255$ (0.03). The column 'Average' is the average of all models except the source model.

Method	ResNet50*	Inc-v3	WRN	VGG19	PNASNet
I-FGSM	99.9%	14.9%	41.3%	26.4%	17.4%
I-FGSM + ILA	99.9%	34.6%	79.9%	66.7%	43.3%
I-FGSM + ILA++	99.9%	41.5%	87.1%	75.2%	49.2%
I-FGSM + LinBP + SGM	**100.0%**	35.3%	88.7%	78.7%	45.0%
MI-CT-FGSM	97.9%	65.0%	77.5%	69.7%	67.6%
NI-CT-FGSM	99.4%	59.6%	78.1%	67.1%	64.9%
VMI-CT-FGSM	99.4%	66.6%	84.8%	73.0%	70.2%
VNI-CT-FGSM	99.8%	**67.3%**	87.4%	76.1%	71.8%
I-FGSM + ILA-DA (Ours)	99.0%	64.1%	**92.6%**	**91.6%**	**72.4%**

Method (cont.)	DenseNet	ResNeXt	MobileNet	SENet	Average
I-FGSM	31.4%	41.8%	31.9%	44.0%	31.1%
I-FGSM + ILA	69.2%	78.5%	67.6%	80.2%	65.0%
I-FGSM + ILA++	79.1%	78.5%	75.3%	87.6%	71.7%
I-FGSM + LinBP + SGM	81.4%	77.1%	75.1%	91.0%	74.7%
MI-CT-FGSM	77.2%	71.6%	74.0%	80.8%	72.9%
NI-CT-FGSM	76.0%	67.0%	73.6%	81.1%	70.9%
VMI-CT-FGSM	82.9%	77.3%	80.5%	86.7%	77.8%
VNI-CT-FGSM	85.1%	80.3%	81.5%	88.9%	79.8%
I-FGSM + ILA-DA (Ours)	**89.8%**	**86.2%**	**87.9%**	**91.3%**	**84.5%**

* The source model used to generate the attack.

Table 2: Attack success rates of ImageNet adversarial examples on six defended methods, generated from Inception V3 with $\epsilon = 16/255$ (0.063).

Attack	Inc-v3$_{ens3}$	Inc-v3$_{ens4}$	IncRes-v2$_{ens}$	HGD	R&P	MMD	Average
I-FGSM	12.1%	10.9%	5.8%	2.7%	4.0%	8.3%	7.3%
+ ILA-DA	**86.3%**	**81.8%**	**66.4%**	**82.2%**	**68.3%**	**70.9%**	**75.9%**
MI-FGSM	14.1%	13.0%	6.6%	4.6%	5.0%	8.3%	8.6%
+ ILA-DA	**83.6%**	**79.2%**	**64.6%**	**79.5%**	**65.9%**	**70.2%**	**73.8%**
MI-CT-FGSM	65.5%	62.1%	45.5%	56.6%	44.5%	52.5%	54.5%
+ ILA-DA	**88.1%**	**84.4%**	**72.3%**	**84.4%**	**73.3%**	**76.2%**	**79.8%**
NI-CT-FGSM	58.8%	54.4%	40.0%	49.2%	38.0%	46.1%	47.8%
+ ILA-DA	**87.3%**	**83.9%**	**68.5%**	**81.0%**	**71.0%**	**74.7%**	**77.7%**
MI-Admix-TI-DIM	73.4%	70.7%	53.9%	65.4%	53.7%	58.0%	62.5%
+ ILA-DA	**88.9%**	**86.3%**	**74.6%**	**85.2%**	**77.1%**	**79.7%**	**81.9%**
VMI-CT-FGSM	77.6%	75.2%	63.6%	72.1%	63.0%	69.7%	70.2%
+ ILA-DA	**89.1%**	**85.7%**	**74.5%**	**84.1%**	**75.6%**	**78.7%**	**81.3%**
VNI-CT-FGSM	79.1%	77.4%	65.3%	72.7%	63.5%	70.8%	71.5%
+ ILA-DA	**88.0%**	**86.1%**	**74.5%**	**84.2%**	**75.8%**	**78.0%**	**81.1%**

Results (2)

Table 3: Comparison of the attack success rates of ImageNet adversarial examples on various models using ILA-DA with different augmentation configurations.

Method	ResNet50*	Inc-v3	WRN	VGG19	PNASNet
I-FGSM + ILA-DA	99.0%	**64.1%**	**92.6%**	91.6%	**72.4%**
w/o Augmentation	99.7%	39.7%	86.6%	79.8%	49.7%
w/ Random Augmentation	98.4%	62.6%	90.4%	90.9%	71.3%
w/ All Augmentation	93.4%	59.4%	78.0%	79.9%	64.3%
w/o Reverse adversarial update	97.3%	50.6%	88.0%	**92.1%**	65.9%
w/o Attack interpolation	**99.9%**	54.0%	91.8%	83.8%	61.6%

Method (cont.)	DenseNet	ResNeXt	MobileNet	SENet	Average
I-FGSM + ILA-DA	**89.8%**	**86.2%**	87.9%	91.3%	**84.5%**
w/o Augmentation	77.0%	76.0%	75.6%	85.9%	71.3%
w/ Random Augmentation	87.3%	83.8%	88.0%	89.3%	82.9%
w/ All Augmentation	74.3%	70.2%	78.0%	78.7%	72.8%
w/o Reverse adversarial update	85.7%	82.1%	**88.3%**	88.6%	79.9%
w/o Attack interpolation	87.2%	86.0%	83.9%	**91.8%**	80.0%

* The source model used to generate the attack.

Ablation Study

+ We study each of the three proposed techniques
+ Data augmentation is the most effective among the three
+ Sampling based on the learned distribution is more effective than sampling randomly
+ Applying all three techniques together gives the highest average success rate

Summary

ILA-DA applies 3 novel augmentation techniques:

- automated data augmentation
- reverse adversarial update
- attack interpolation

ILA-DA outperforms SOTA on 9 undefended models and 6 defended models.

ILA-DA highlights the effectiveness of data augmentation in transfer-based attacks.

ICLR2023

ILA-DA: Improving Transferability Of Intermediate Level Attack with Data Augmentation