Review: Practical No-box Adversarial Attacks Against DNNs

Zhiying Li
lizhiyi3@msu.edu

Abstract

This report reviews the paper ***Practical No-box Adversarial Attacks Against DNNs*** Li et al. [2020]. This report contains two parts: *Part 1*, I reviewed the background/motivation, problem formulation, method, experiments, and results of this work. In *Part 2*, I provided my post-reading insights and discussed my critics of this work

1 Paper Review

1.1 Background, Motivation and Problem Formulation

Evasion attack, also known as ℓ_p evasion attack or adversarial example, is a type of attacks that generate small perturbation to fool a trained ML system (Carlini and Wagner, 2017). The general problem formulation is

$$\begin{aligned} \min_{\delta} \quad & \ell_{\text{atk}}(\mathbf{x}+\delta;\theta) \\ \text{s.t.} \quad & \|\delta\|_p \leq \epsilon, x \end{aligned}$$

where $\mathbf{x}$ is the training data, δ is the perturbation, θ is a model parameter, and ϵ is the perturbation boundary. Generally, at inference-phase (test time), we use input gradient by back-propagation (bp) $\nabla_\delta l_{\text{atk}}(\mathbf{x}+\delta;\theta)$ to find the input perturbation.

Based on knowing or not knowing the victim model's parameters θ, evasion attacks can also be broadly categorized as white-box attacks and black-box attacks. White-box attack directly calculates the gradient, and black-box attack estimates the gradient through queries.

However, it's impractical for many real-world cases when we don't know the victim model's parameters (white-box), or when we are not allowed to query frequently (black-box). Therefore, this paper, *No-Box Attack* is motivated to address these issues: *No-Box Attack* leverages transfer attack to achieve attacks not knowing victim model parameters and not querying the victim model.

The problem formulation of *No-Box Attack* is: "Assume a benign instance x_0 is to be perturbed such that being misclassified into an arbitrary label. We aim to train a substitute discriminative model θ_{sub} on a small and thus easily gathered (and labeled) auxiliary dataset $\mathcal{X} := \{(x_i, y_i)\}_{i=0}^{n-1}$, including the instance x_0 to be perturbed."

1.2 Method

The core method of *No-Box Attack* is transfer attack. A "substitute" model is trained based on a small auxiliary dataset in the same domain as training data. The adversarial perturbation to attack victim model is designed on top of the "substitute" model. In this way, we do not need to have any knowledge about the victim model.

Another requirement for practicability in this problem is to use as small amounts of training data as possible to train the "substitute" model. When people try to train a DNN classifier (with data augmentation and regularization) to distinguish the two classes with a small dataset, the model will overfit because the dataset is too small to generalize.

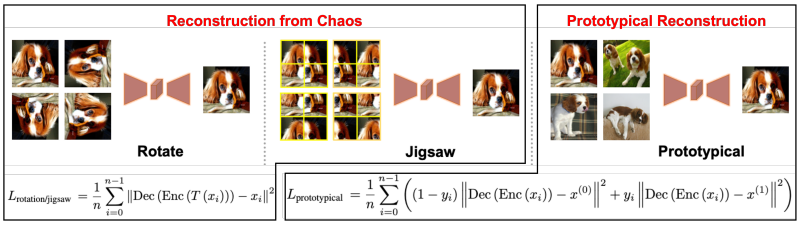


Figure 1: 2 Types and 3 Schemes of Training Mechanism of the Substitute Model.

Instead, the paper used 2 training mechanisms for the substitute model 'Reconstruction from Chaos' and 'Prototypical Reconstruction'.

- **Reconstruction from Chaos** is to train a model to capture the underlying variation of augmented data for recovering the original image. It is self-supervised learning. This paper proposed two schemes: 'Rotation Reconstruction' and 'Jigsaw Puzzle Reconstruction'.
- **Prototypical Reconstruction** is a training mechanism to select a single sample image from the input of the class-specific subset of the dataset. This is a supervised mechanism, and it encourages the model to reconstruct class-specific prototypes.

Attack the "Substitute Model": At decision time, when we launch the attack, the adversaries apply the attack loss (negative cross-entropy loss):

$$L_{\text{adversarial}} = -\log p\left(y_i \mid x_i\right) \quad \text{where} \quad p\left(y_i \mid x_i\right) = \frac{\exp\left(-\lambda \left\|\text{Dec}\left(\text{Enc}\left(x_i\right)\right) - \tilde{x}_i\right\|^2\right)}{\sum_j \exp\left(-\lambda \left\|\text{Dec}\left(\text{Enc}\left(x_i\right)\right) - \tilde{x}_j\right\|^2\right)},$$

in which $\tilde{x}_i$ is "positive prototype"(true case), and $\tilde{x}_i$ (negative case). λ is a ≥ 0 parameter (In this paper, $\lambda = 1$). Thus, minimizing attack loss $\ell_{\text{atk}}(\mathbf{x} + \delta; \theta_{\text{sub}}) = -L_{\text{adversarial}}$ is to maximizing the difference between $\text{Dec}\left(\text{Enc}\left(x_i + \delta\right)\right)$ and positive prototype of $\tilde{x}_i$, thus minimizing the likelihood of correct output under input perturbation. Notice, θ_{sub} is the model parameters of"substitute model", *No-Box Attack* has no knowledge of the victim model.

ILA (intermediate level attack) to Improve Transferability In general, ILA is a method to enhance to transferability of a black-box evasion attack by increasing the perturbation on a pre-specified layer of a model Huang et al. [2019]. In this paper, the attack is performed in conjunction with ILA. ILA maximizes projections on the mid-layer representation. In this way, ILA enlarges intermediate-level perturbation in the direction of guiding examples. The direction of the guiding example is first obtained in the attack step while minimizing ℓ_{atk}, and ILA is then performed on the output of the encoder.

1.3 Experiment and Result

This paper conducted experiments on two applications: Image Classification and Face Verification. The substitute model they used is CycleGAN. The attack type is ℓ_∞ ($\epsilon = 0.1$ or 0.8), and the method to get input gradient is I-FGSM (also PGD in supplementary materials).

Image Classification ImageNet contains 1,200,000 images with 1,000 classes. They choose 9 popular models (VGG-19, Inception, ResNet, DenseNet, SENet, WRN, RNASNet, and MobileNet) as victim models. They trained the substitute model with 5,000 images in 500 classes from the validation set.

Face Verification LFW (Labeled Face from Wild) contains 13,233 images with 5,749 classes. They choose 2 popular models (FaceNet and CosFace) as victim models. They trained the substitute model with 2,110 images in 400 classes.

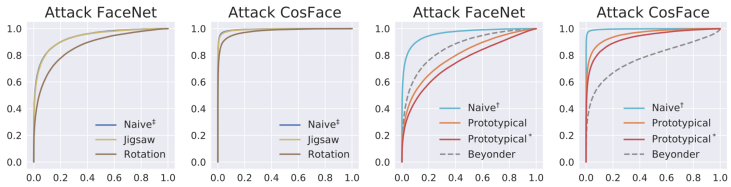
Comparison They compare with the following competitor models to benchmark their performance.

- ***naive***‡ **and** ***naive***‡: 2 Baseline (trained over same small-scale dataset): Transferring Adversarial examples from (***naive***†) Supervised ResNet trained with possible regularization and data augmentation, and (***naive***‡) unsupervised AutoEncoder with the same policy as training rotation/jigsaw/prototype.

(a) Compare the transferability of adversarial examples crafted on different models on ImageNet. The prediction accuracy on adversarial examples under $\epsilon = 0.1$ are shown (lower is better).

Method	Sup.	VGG-19 [42]	Inception v3 [45]	ResNet [15]	DenseNet [17]	SENet [16]	WRN [56]	PNASNet [28]	MobileNet v2 [39]	Average
Naïve‡	✗	45.92%	63.94%	60.64%	56.48%	65.54%	58.80%	73.14%	37.76%	57.78%
Jigsaw	✗	31.54%	50.28%	46.24%	42.38%	59.06%	51.24%	62.32%	25.24%	46.04%
Rotation	✗	31.14%	48.14%	47.40 %	41.26%	58.20%	50.72%	59.94%	26.00%	45.35%
Naïve†	✓	76.20%	80.86%	83.76%	78.94%	87.00%	84.16%	86.96%	72.44%	81.29%
Prototypical	✓	19.78%	36.46%	37.92%	29.16%	44.56%	37.28%	48.58%	17.78%	33.94%
Prototypical*	✓	18.74%	33.68%	34.72%	26.06%	42.36%	33.14%	45.02%	16.34%	**31.26%**
Beyonder	✓	24.96%	51.12%	30.30%	27.12%	43.78%	33.94%	51.80%	27.02%	36.26%

* The prototypical models with multiple decoders. To be more specific, 20 decoders are introduced in each model.



(b) ROC curves of face verification on adversarial examples crafted on different substitute models. The left two sub-figures show *unsupervised* results and the right two show *supervised* results.

Figure 2: (a) Result of ImageNet (b) Result of Face Detection from LFW

- ***Beyonder***: adversarial example transferred from models pre-trained on a large-scale or even the same dataset as training the victim model.

Result: Their result demonstrates their approach outperformed the baseline model and performs similarly or even better than the transfer attack based on pre-trained models using large-scale training sets.

- **Image Classification**: Their result is presented in Figure 2a. Across all 9 victim models, ***naive***‡ and ***naive***‡ methods perform the worst as expected, and all 3 autoencoder training schemes outperforms Naive model and reduce the accuracy below 50%. Prototypical auto-encoding models transfer the best to a variety of image classification, and it even outperforms the ***Beyonder*** model.
- **Face Verification**: Their result is presented in Figure 2b. They present their result with receiver operating characteristic (ROC) curves. Ideally, a successful attack is low true positive and high false positive (under the diagonal line). In an unsupervised autoencoding training mechanism, Jigsaw doesn't attain an effective attack, and Rotation achieves a slightly worse attack. For the supervised part, we can see the prototypical methods achieves the best performance and also outperforms the ***Beyonder***.

2 Insights and Discussion

2.1 Post-reading Insights

This paper provides a novel way to achieve transfer evasion attacks with a small amount of training data. Thus, it achieves "practicability of attack", when the model parameter is infeasible (defense of white-box model) or querying and large-scale training are infeasible (defense of black-box model).

To leverage a small dataset, the core of their method is the 3 auto-encoder's training schemes they designed. Compared with DNN supervised learning as the substitute model, it requires less data, and 2 of the 3 schemes are even unsupervised. Unsupervised schemes, from my perspective, is even more practical than the supervised prototyical scheme when only the data and not labels are available.

In terms of transferability, they used the existing Intermediate Level Attack to improve their model. Besides the existing method, they did not put in special effort to innovate transferability-improving schemes when the training dataset is very small. To the best of my knowledge, ILA is a general transfer method tested with knowledge of entire victim training data. A question that may need further investigation is whether there exists a method tailored made to improve the transferability when the victim model is trained with a small subset of the dataset.

2.2 Discussions

Choice of ϵ. In Image Classification application, the prototypical training schemes only outperform Beyounder example when $\epsilon = 0.1$ (Table 1). In the supplementary material, when $\epsilon = 0.08$, we can see the attack performance of Beyonder is still the best (Table 2). This shows when the maximum perturbation is small, the no-box attack cannot beat the performance transfer attack from the model trained by the whole training data. Although the paper did not explain intuition, I think this is a non-trivial question worth investigating.

Jigsaw Scheme Face Verification. In Face Verification results, we can tell the Jigsaw model fails to learn. In Figure 2(b) *left*, the Jigsaw curve is almost the same as the Naive model. The intuition the paper gives is that Jigsaw model cannot discriminate the distortion of facial structure. Based on the visualization and explanation in the supplementary material, attacks on Jigsaw Model will not provide useful information to divert model attention, given it fails to learn the distinctive discriminate parts at first. However, the failure of Jigsaw model can be an implication that Jigsaw, Rotation, and Prototypial model schemes can be application-specific. There need the further experiments of No-Box attacks on other datasets and applications to test the effectiveness in general.

Number of Training Images. I found the attack performance of Jigsaw and Rotation with the number of training images is counter-intuitive. In the experiment of this paper, they choose 10 images per class. How they decided on the number 10 through a test they conducted. They have tested the cases with 2, 10, 20, and 40 images. However, their result shows, varying the number of images, the attack effectiveness of different training schemes of auto-encoder shows different trends (Figure 7 in paper). The prototypical scheme's accuracy decreases with the number of training images, which means better attack performance. This fits our intuition because more data is likely to lead to better performance. However, the Jigsaw's and Rotation's accuracy increases with the number of training images, which is counter-intuitive. This is an unanswered question throughout this paper. In the end, how they choose 10 is a balance between these 2 trends when all schemes achieve accuracy on the same level.

References

Qian Huang, Isay Katsman, Horace He, Zeqi Gu, Serge Belongie, and Ser-Nam Lim. Enhancing adversarial example transferability with an intermediate level attack. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4733–4742, 2019.

Qizhang Li, Yiwen Guo, and Hao Chen. Practical no-box adversarial attacks against dnns. *Advances in Neural Information Processing Systems*, 33:12849–12860, 2020.