Attack on Image Recognition

Attack on Image Recognition

- Todo List
- Synthesizing Robust Adversarial Examples
  - Contribution
  - Notes
  - Links
- NES: Black-box Adversarial Attacks with Limited Queries and Information
  - Notes
  - Links

Todo List

1. Kurakin, A., Goodfellow, I., and Bengio, S. Adversarial examples in the physical world. 2016.
2. Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., and Fergus, R. Intriguing properties of neural networks. 2013.
3. Goodfellow, I. J., Shlens, J., and Szegedy, C. Explaining and harnessing adversarial examples. In Proceedings ofthe International Conference on Learning Representations (ICLR), 2015.
4. Carlini, N. and Wagner, D. Towards evaluating the robustness of neural networks. In IEEE Symposium on Security & Privacy, 2017c.
5. Evtimov, I., Eykholt, K., Fernandes, E., Kohno, T., Li, B., Prakash, A., Rahmati, A., and Song, D. Robust PhysicalWorld Attacks on Deep Learning Models. 2017.
6. Tom B Brown, Dandelion Man´e, Aurko Roy, Mart´ın Abadi, and Justin Gilmer. Adversarial patch. arXiv preprint arXiv:1712.09665, 2017.
7. Danny Karmon, Daniel Zoran, and Yoav Goldberg. Lavan: Localized and visible adversarial noise. arXiv preprint arXiv:1801.02608, 2018.
8. Zuxuan Wu, Ser-Nam Lim, Larry Davis, and Tom Goldstein. Making an invisibility cloak: Real world adversarial attacks on object detectors. arXiv preprint arXiv:1910.14667, 2019.
9. Cassidy Laidlaw and Soheil Feizi. Functional adversarial attacks, 2019
10. Mahmood Sharif, Sruti Bhagavatula, Lujo Bauer, and Michael K Reiter. Adversarial generative nets: Neural network attacks on state-of-the-art face recognition. arXiv preprint arXiv:1801.00349, 2017.
11. Anish Athalye, Nicholas Carlini, and David Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. arXiv preprint arXiv:1802.00420, 2018.
12. advbox
13. paddlepaddle 对抗工具箱
14. Stealthy Porn: Understanding Real-World Adversarial Images for Illicit Online Promotion
15. Stealing Hyperparameters in Machine Learning
16. Phantom of the ADAS: Securing Advanced Driver-Assistance Systems from Split-Second Phantom Attacks
17. Text Captcha Is Dead? A Large Scale Deployment and Empirical Study
18. A Tale of Evil Twins: Adversarial Inputs versus Poisoned Models

19. Adversarial Sensor Attack on LiDAR-based Perception in Autonomous Driving
20. Privacy Risks of Securing Machine Learning Models against Adversarial Examples
21. Procedural Noise Adversarial Examples for Black-Box Attacks on Deep Convolutional Networks
22. Seeing isn't Believing: Towards More Robust Adversarial Attack Against Real World Object Detectors
23. Model-Reuse Attacks on Learning Systems
24. A. Ilyas, L. Engstrom, A. Athalye, and J. Lin, "Black-box adversarial attacks with limited queries and information," in ICML, 2018.
25. A. Kurakin, I. J. Goodfellow, and S. Bengio, "Adversarial examples in the physical world," in ICLR, 2017.

Synthesizing Robust Adversarial Examples

Contribution

1. 提出了一种增加物理环境下对抗样本鲁棒性的一般化方法 EOT；
2. 不仅在 2D 下测试，而且在 3D 下测试；
3. 模拟物理变换的想法十分具有借鉴意义，已被后续的对抗攻击算法广泛使用；

Notes

1. **白盒**的、**针对物理环境**下的、**有目标**的对抗攻击算法。攻击的算法不仅在 2D 下可行，同时在 **3D** 下也可以生成成功的对抗样本；
2. 已有的对抗攻击算法，训练的目标如下：

$$\begin{aligned} \underset{x'}{\arg\max} \quad & \log P(y_t|x') \\ \text{subject to} \quad & ||x' - x||_p < \epsilon \\ & x' \in [0,1]^d \end{aligned}$$

   但是这样生成的对抗样本，在视角等物理环境发生改变时**无法保持对抗性。**故作者提出改进后的训练目标 **EOT (Expectation Over Transformation)**：

$$\begin{aligned} \underset{x'}{\arg\max} \quad & \mathbb{E}_{t\sim T}[\log P(y_t|t(x'))] \\ \text{subject to} \quad & \mathbb{E}_{t\sim T}[d(t(x'), t(x))] < \epsilon \\ & x \in [0,1]^d \end{aligned}$$

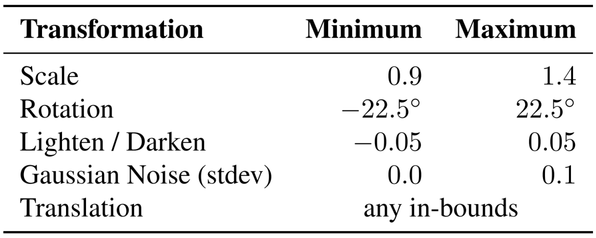
   其含义是，**在保证对抗样本经过物理变换的"感受"修改量在一定范围内时，使得对抗样本（经过物理变换）能够尽可能地被分类为目标类别。**这类物理变换可以是 2D/3D 的变换，包括随机旋转、平移、噪声、视角变化、光照等。作者将公式转换为 Carlini &Wagner (2017c) 的形式，并使用**二级范数**和 **PGD** (Projected Gradient Descent) 优化器进行计算：

$$\begin{aligned} \underset{x'}{\arg\max}\, \mathbb{E}_{t\sim T} \Big[& \log P(y_t|t(x')) \\ & -\lambda ||LAB(t(x')) - LAB(t(x))||_2 \Big] \end{aligned}$$

   其中 *LAB* 代表指的是 LAB 色域 。

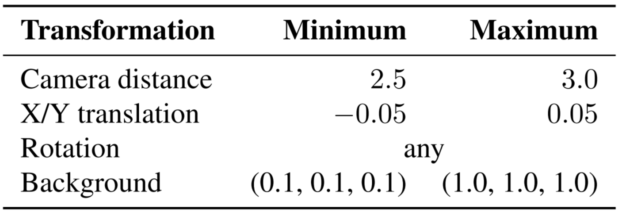
3. Distributions of Transformations:

   (1) **2D Case**



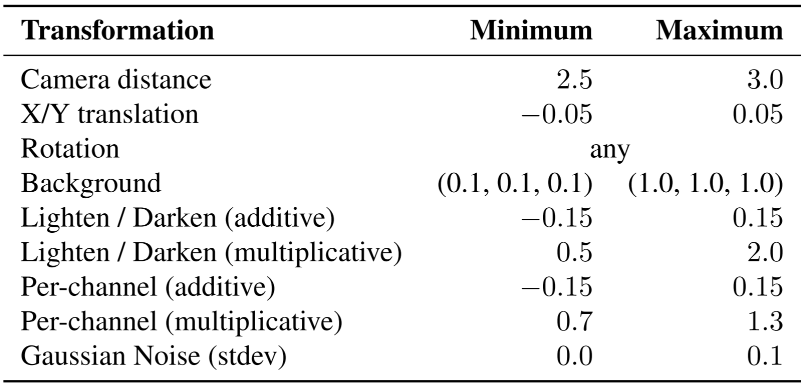
Transformation	Minimum	Maximum
Scale	0.9	1.4
Rotation	−22.5°	22.5°
Lighten / Darken	−0.05	0.05
Gaussian Noise (stdev)	0.0	0.1
Translation	any in-bounds	

(2) **3D Case**



Transformation	Minimum	Maximum
Camera distance	2.5	3.0
X/Y translation	−0.05	0.05
Rotation	any	
Background	(0.1, 0.1, 0.1)	(1.0, 1.0, 1.0)

(3) **Physical Case**

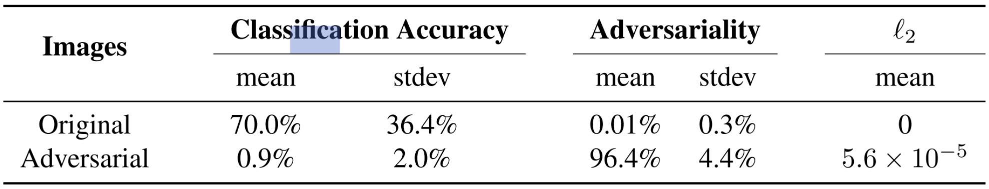


Transformation	Minimum	Maximum
Camera distance	2.5	3.0
X/Y translation	−0.05	0.05
Rotation	any	
Background	(0.1, 0.1, 0.1)	(1.0, 1.0, 1.0)
Lighten / Darken (additive)	−0.15	0.15
Lighten / Darken (multiplicative)	0.5	2.0
Per-channel (additive)	−0.15	0.15
Per-channel (multiplicative)	0.7	1.3
Gaussian Noise (stdev)	0.0	0.1

4. Evaluation：

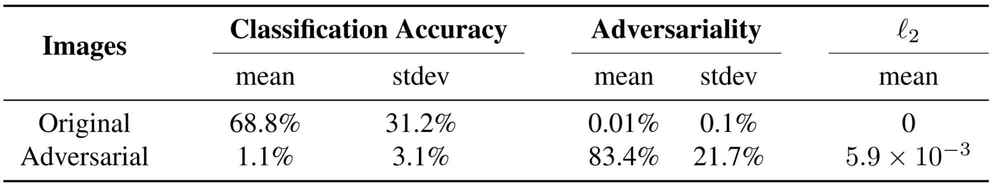
(1) 攻击基于数据集 ImageNet 的 **Inception V3** 模型（`Top-1 Accuracy = 78.0%`），随机选择目标分类；

(2) **Robust 2D adversarial examples**：在 2D 下考虑的物理变换有 **缩放 、旋转、亮度调节、高斯噪声和平移。**每个样本都在 **1000** 个随机的模拟物理变换上进行测试，结果如下：



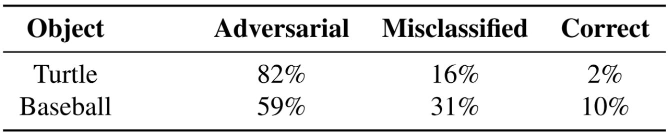
Images	Classification Accuracy		Adversariality		ℓ_2
	mean	stdev	mean	stdev	mean
Original	70.0%	36.4%	0.01%	0.3%	0
Adversarial	0.9%	2.0%	96.4%	4.4%	5.6×10^{-5}

(3) **Robust 3D adversarial examples**：在 3D 下考虑**不同的相机距离、照明条件、对象的平移和旋转以及纯色背景色。**挑选了 10 个 3D 模型 —— 木桶、棒球、够、橘子、海龟、小丑鱼、沙发、泰迪熊、汽车和出租车。每个 3D 模型都挑选 20 个随机的目标分类标签；每个样本都在 100 个随机的模拟物理变换上进行测试，结果如下：



Images	Classification Accuracy		Adversariality		ℓ_2
	mean	stdev	mean	stdev	mean
Original	68.8%	31.2%	0.01%	0.1%	0
Adversarial	1.1%	3.1%	83.4%	21.7%	5.9×10^{-3}

(4) **Physical adversarial examples**：在 3D 的基础上，考虑**摄像机的噪声、照明的影响和颜色的失真。**作者考虑将 ”海龟“ 错误分类成 ”手枪“、 ”棒球“ 错误分类成 ”咖啡“ 两种情况，将对抗样本经过 3D 打印后，拍 100 张照片进行测试，结果如下：



Object	Adversarial	Misclassified	Correct
Turtle	82%	16%	2%
Baseball	59%	31%	10%

(5) **Perturbation budget**：在物理环境下越鲁棒，需要模拟更多的物理变换，添加的噪声也会更多；

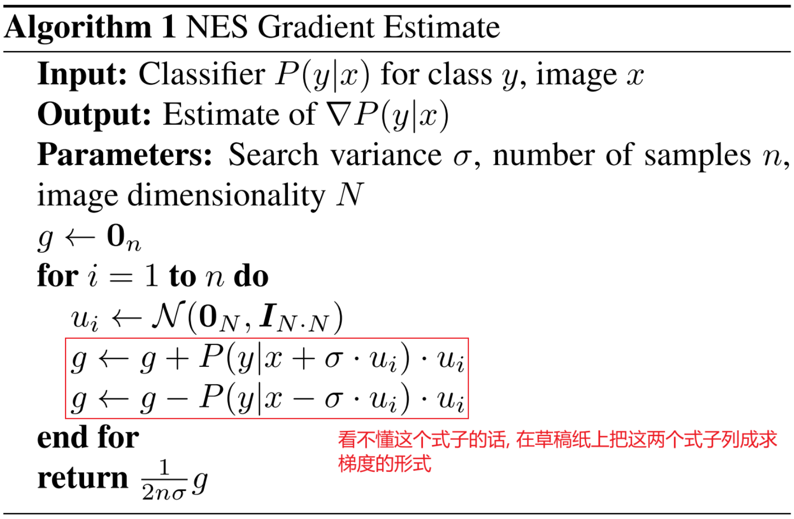
Links

- 论文链接：Athalye, Anish, et al. "Synthesizing robust adversarial examples." *International conference on machine learning*. PMLR, 2018.
- 开源代码：prabhant/synthesizing-robust-adversarial-examples (github.com)

NES: Black-box Adversarial Attacks with Limited Queries and Information

Notes

1. 黑盒设定：
    - Query-limited Setting: 限制访问次数；
    - Partial-information Setting: 只知道 Top-K 的结果 (包括概率);
    - Label-only Setting: 只知道 Top-K 的标签 (不包括概率); (这一项我觉得没必要看)
2. NES (Natural Evolutionary Strategies) 进行梯度估计: 最小化期望的损失大小, 算法伪代码如下 (如何挑选这个参数?)



Algorithm 1 NES Gradient Estimate

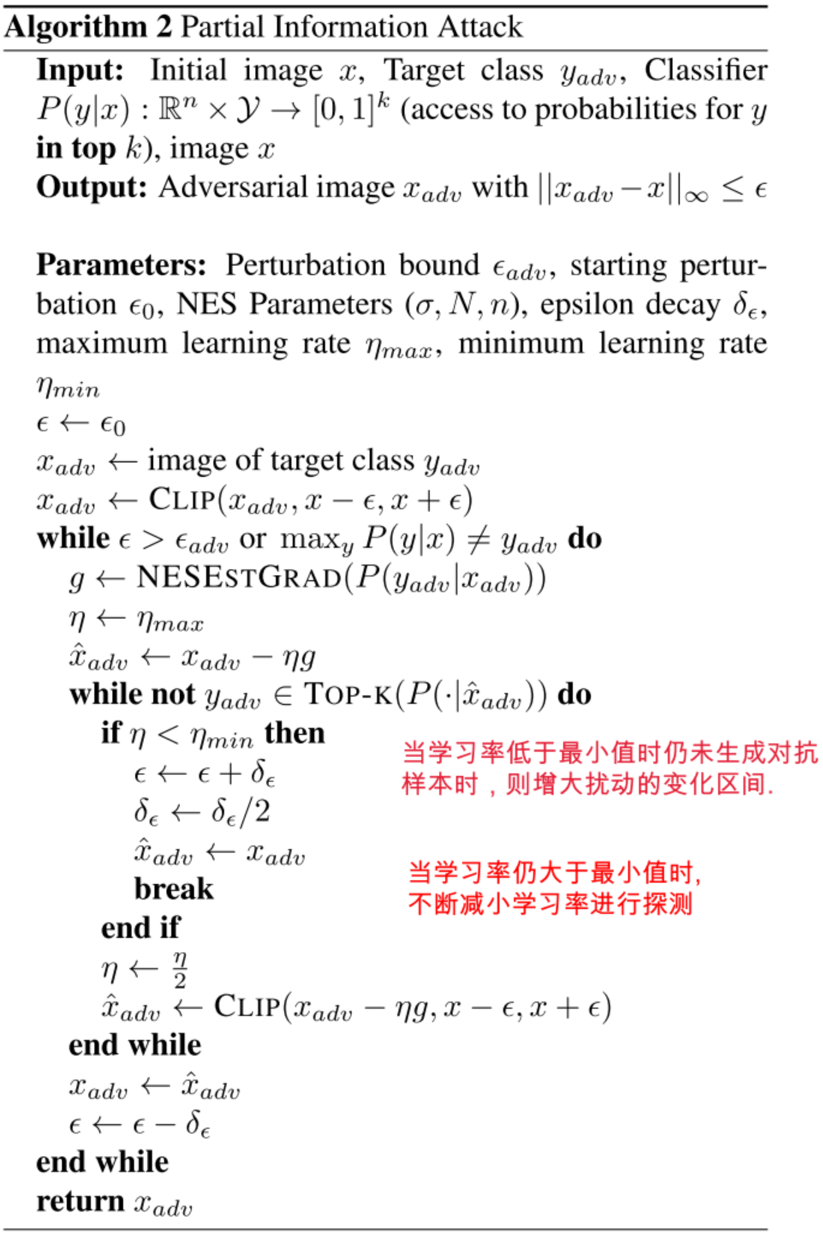
```
Input: Classifier $P(y|x)$ for class $y$, image $x$
Output: Estimate of $\nabla P(y|x)$
Parameters: Search variance $\sigma$, number of samples $n$, image dimensionality $N$
$g \leftarrow \mathbf{0}_n$
for $i = 1$ to $n$ do
    $u_i \leftarrow \mathcal{N}(\mathbf{0}_N, \boldsymbol{I}_{N \cdot N})$
    $g \leftarrow g + P(y|x + \sigma \cdot u_i) \cdot u_i$
    $g \leftarrow g - P(y|x - \sigma \cdot u_i) \cdot u_i$
end for
return $\frac{1}{2n\sigma} g$
```

看不懂这个式子的话, 在草稿纸上把这两个式子列成求梯度的形式

3. PGD (Projected Gradient Descent) 进行梯度更新:

$$x^{(t)} = \Pi_{[x_0-\epsilon, x_0+\epsilon]}(x^{(t-1)} - \eta \cdot \text{sign}(g_t))$$

4. 仅知道 Top-K 的概率:



Algorithm 2 Partial Information Attack

```
Input: Initial image $x$, Target class $y_{adv}$, Classifier $P(y|x) : \mathbb{R}^n \times \mathcal{Y} \to [0,1]^k$ (access to probabilities for $y$ in top $k$), image $x$
Output: Adversarial image $x_{adv}$ with $||x_{adv} - x||_\infty \le \epsilon$

Parameters: Perturbation bound $\epsilon_{adv}$, starting perturbation $\epsilon_0$, NES Parameters $(\sigma, N, n)$, epsilon decay $\delta_\epsilon$, maximum learning rate $\eta_{max}$, minimum learning rate $\eta_{min}$
$\epsilon \leftarrow \epsilon_0$
$x_{adv} \leftarrow$ image of target class $y_{adv}$
$x_{adv} \leftarrow \textsc{Clip}(x_{adv}, x - \epsilon, x + \epsilon)$
while $\epsilon > \epsilon_{adv}$ or $\max_y P(y|x) \neq y_{adv}$ do
  $g \leftarrow \textsc{NESEstGrad}(P(y_{adv}|x_{adv}))$
  $\eta \leftarrow \eta_{max}$
  $\hat{x}_{adv} \leftarrow x_{adv} - \eta g$
  while not $y_{adv} \in \textsc{Top-K}(P(\cdot|\hat{x}_{adv}))$ do
    if $\eta < \eta_{min}$ then
      $\epsilon \leftarrow \epsilon + \delta_\epsilon$
      $\delta_\epsilon \leftarrow \delta_\epsilon / 2$
      $\hat{x}_{adv} \leftarrow x_{adv}$
      break
    end if
    $\eta \leftarrow \frac{\eta}{2}$
    $\hat{x}_{adv} \leftarrow \textsc{Clip}(x_{adv} - \eta g, x - \epsilon, x + \epsilon)$
  end while
  $x_{adv} \leftarrow \hat{x}_{adv}$
  $\epsilon \leftarrow \epsilon - \delta_\epsilon$
end while
return $x_{adv}$
```

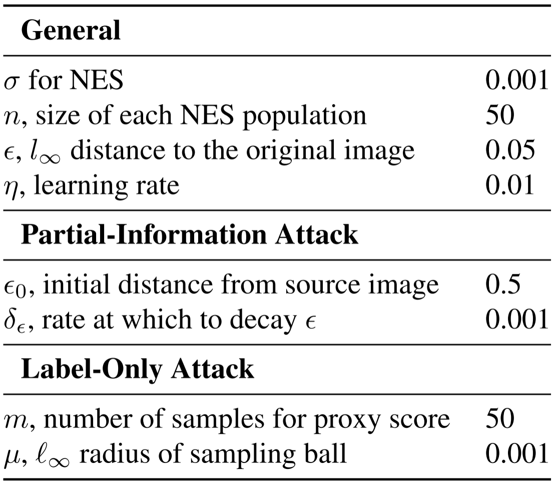
当学习率低于最小值时仍未生成对抗样本时，则增大扰动的变化区间.

当学习率仍大于最小值时，不断减小学习率进行探测

- 使用目标分类的样本来初始化扰动，从而减少 query 的数量；
- 在保证目标分类在 Top-K 中的前提下，不断缩小对抗扰动，直至生成对抗样本且满足修改量的限制；

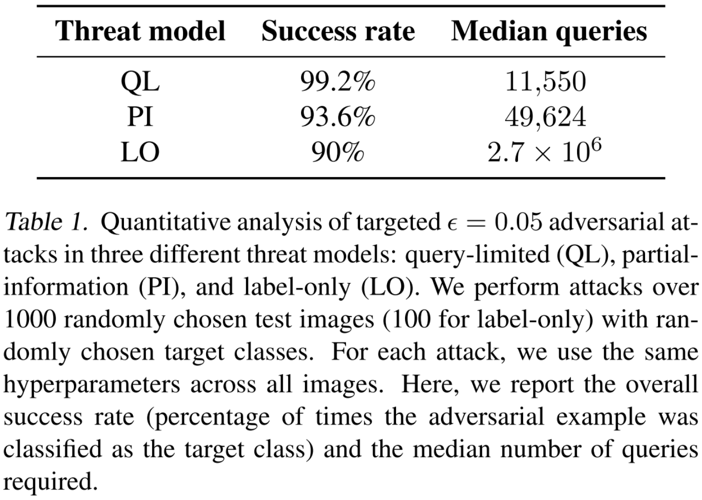
5. Evaluation：

(1) 参数的选择：



General	
σ for NES	0.001
n, size of each NES population	50
ϵ, l_∞ distance to the original image	0.05
η, learning rate	0.01
Partial-Information Attack	
ϵ_0, initial distance from source image	0.5
δ_ϵ, rate at which to decay ϵ	0.001
Label-Only Attack	
m, number of samples for proxy score	50
μ, ℓ_∞ radius of sampling ball	0.001

(2) On ImageNet：这里大概的 query 数量级为上万级别的



Threat model	Success rate	Median queries
QL	99.2%	11,550
PI	93.6%	49,624
LO	90%	2.7×10^6

Table 1. Quantitative analysis of targeted $\epsilon = 0.05$ adversarial attacks in three different threat models: query-limited (QL), partial-information (PI), and label-only (LO). We perform attacks over 1000 randomly chosen test images (100 for label-only) with randomly chosen target classes. For each attack, we use the same hyperparameters across all images. Here, we report the overall success rate (percentage of times the adversarial example was classified as the target class) and the median number of queries required.

Links

- 论文链接: Ilyas, Andrew, et al. "Black-box adversarial attacks with limited queries and information." *PRML* (2018).
- 论文代码: https://github.com/labsix/limited-blackbox-attacks