# Attack on Image Recognition



# 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.

# 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)**：

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

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：

$$
\arg\max_{x'} \mathbb{E}_{t\sim T}\Big[\log P(y_t|t(x')) \\
-\lambda||LAB(t(x')) - LAB(t(x))||_2\Big]
$$

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 | | $\ell_2$ |
|---|---|---|---|---|---|
| | mean | stdev | mean | stdev | mean |
| Original | 70.0% | 36.4% | 0.01% | 0.3% | 0 |
| Adversarial | 0.9% | 2.0% | 96.4% | 4.4% | $5.6\times10^{-5}$ |

(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 | | $\ell_2$ |
|---|---|---|---|---|---|
| | mean | stdev | mean | stdev | mean |
| Original | 68.8% | 31.2% | 0.01% | 0.1% | 0 |
| Adversarial | 1.1% | 3.1% | 83.4% | 21.7% | $5.9 \times 10^{-3}$ |

(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)]