

# Write Up for 百度 AI 安全对抗赛

Daquan Lin\* Zhe Zhao†

ShanghaiTech University

## Abstract

本文为 2019-2020 年百度 AI 安全对抗赛 subianwanwan 队的 Write Up，该队伍成员来自上海科技大学计算机系统与安全中心。文章将分为比赛介绍，初赛和复赛三个部分，后两部分中详细阐述了在对应阶段中使用的方法。

## 1. 比赛介绍

目前人工智能和机器学习技术被广泛应用于人机交互、推荐系统、安全防护等各个领域，其受攻击的可能性以及是否具备强抗打击能力备受业界关注，因此图像识别的准确性对人工智能产业至关重要。这一环节也是最容易被攻击者利用，通过对数据源的细微修改，在用户感知不到的情况下，使机器做出了错误的操作。这种方法会导致 AI 系统被入侵、错误命令被执行，执行后的连锁反应会造成的严重后果。

在本次比赛中，选手必须使用飞桨作为攻击方，对图片进行轻微扰动生成对抗样本，使已有的用于执行图像分类任务的深度学习模型识别错误。

飞桨 (PaddlePaddle)<sup>1</sup> 以百度多年的深度学习技术研究和业务应用为基础，集深度学习核心框架、基础模型库、端到端开发套件、工具组件和服务平台于一体，2016 年正式开源，是全面开源开放、技术领先、功能完备的产业级深度学习平台。飞桨源于产业实践，始终致力于与产业深入融合。目前飞桨已广泛应用于工业、农业、服务业等，服务 150 多万开发者，与合作伙伴一

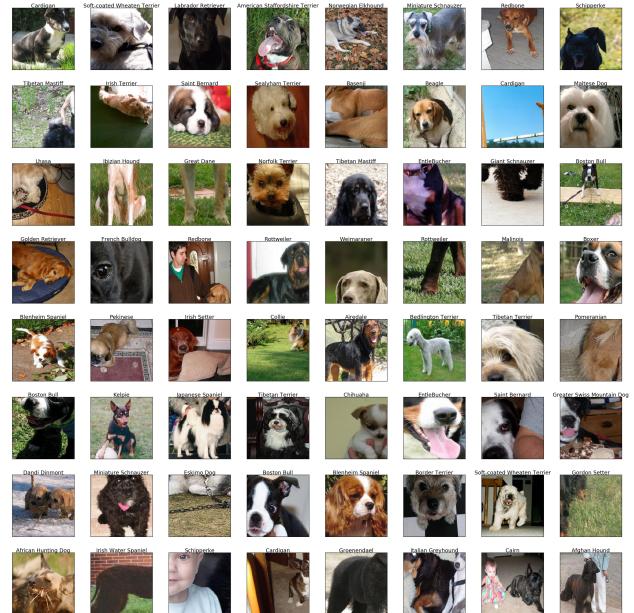


图 1. 本次比赛使用的 Standford Dogs 数据集展示

起帮助越来越多的行业完成 AI 能力赋能。<sup>2</sup>

## 2. 初赛

### 2.1. 初赛介绍

初赛中，主办方提供了三个选手供参赛者攻击，其中包括两个白盒模型和一个黑盒模型，这就要求选手的对抗样本可以同时在两个模型上生效，并且还要拥有不错的迁移性，以便让对抗样本在黑盒上依旧有效。在计算得分时，基于 MSE 和迁移性两个角度考虑。

\*lindq@shanghaitech.edu.cn, equal contribution

†zhaozhe1@shanghaitech.edu.cn

<sup>1</sup><https://www.paddlepaddle.org.cn/>

<sup>2</sup><https://aistudio.baidu.com/aistudio/competition/detail/15>

## 2.2. 方案介绍

在本阶段，队伍主要使用的攻击代码由官方提供的 FGSM 和 PGD demo 演化而来，尝试了以下微调方法：

- Random init
- M-PGD
- Integer domain attack
- L2-norm PGD
- EOT
- C&W

接下来我们将会逐一讨论相关方法的利弊，并给出我们最终使用的方案。

### 2.2.1 BIM or PGD or M-PGD

BIM (basic iteration method) [8]是较为简单高效的对抗样本生成方法，设置好迭代步数及步长  $\epsilon$ ，每次迭代对图片添加对应步长扰动即可。

PGD [9] 主要区别于 BIM 的部分为 Random init，在 PGD 中，第一步需要在图像随机生成 noise，基于这张带有部分 noise 的图片进行后续攻击，这样更有助于找到全局最优。但该步骤会大量增加 MSE，因此没有使用在后续攻击中。

M-PGD 方法是 TSAIL 团队在参与 2017 NIPS Adversarial Competition 时提出的，相关文章随后发表于 2018 CVPR [4]。其主要思路是在 PGD 中，每一步前进的方向为让梯度上升 (Untarget attack) 或者下降 (Target attack) 的方向，但加入动量 (Momentum) 之后，计算方向时除了当前梯度，还要考虑残余动量。直观的理解，这可以帮助对抗样本“冲出”局部最优区域，收敛于全局最优。但相应的，也会导致攻击时长和 MSE 增加。

初赛阶段，我们编写代码实现了以上所有攻击形式，最终考虑到评分标准中对 MSE 的要求，我们使用了 BIM 作为基本的攻击方法。

### 2.2.2 Float domain or Integer domain attack

整数域攻击的思想来自于我们实验室的论文 [5]，该文中详细分析了不同攻击由连续域的 float 转存为整数域的 0-255 图片后，原有攻击的强度衰减情况。其

主要思想是再经过预处理后，png 格式的图片转为了 float numpy array 存储于电脑中，随后我们在 float 的数组上添加扰动并检测其对抗性质，最终，当我们把图片 round 取整并再次存储为 png 时，这会带来一定程度的精度丢失。例如，假设攻击者在 BIM 中攻击一张从 0-255 预处理到 0-1 上的图片，并且使用的步长  $\epsilon < 1/255$ ，那么当转存为图片后，该图片上叠加的所有扰动，都会因为取整操作而丢失。

初赛阶段，考虑到这种情况，我们尽可能的选择了合理的步长和迭代次数，并在取整过程中考虑了梯度方向。

### 2.2.3 Target or Untarget attack

这是两种不同的攻击方式，target 攻击方式的目标是使得正常图片被分类为指定标签，untarget 攻击只需要让正常图片被分类错误即可。

初赛阶段我们考虑了 least-likely target 和 untarget 攻击，前者指的是将图片误分类为正常状况下概率最小的分类，即最不可能的分类。这种分类看起来相比 untarget 强度更高，但在迁移攻击过程中会遇到的问题为，当执行 target 攻击时，原有 label 一般只会被降级为 2nd 或者 3rd 的分类，这就导致了 target 攻击下对抗样本的迁移性较差。因此我们最终使用了 untarget 攻击。

### 2.2.4 增强迁移性的操作

对抗样本的迁移性是本次比赛评分的重要因素。想要增强迁移性，最直观的想法是，生成 confidence 较高的样本，由于 confidence 一般通过 logits 计算得来，考虑到 C&W [2] 攻击中对 loss 函数和 confidence 的分析，我们使用了 logits difference 作为对对抗样本强度的约束，每次攻击，必须确保原有 label 与当前的 1st label 差距大于 30。

Expectation Over Transformation (EOT) [1] 也是增加样本迁移性的操作之一，其做法是，当网络读入一张图片后，首先对图片进行旋转，裁剪，放缩等操作，这些操作的影响会在梯度计算时得到体现，这使得使用 EOT 后得到的梯度更鲁棒，在面对不同模型的卷积操作时，都能保持较为稳定的攻击效果。根据结果来看，该方法在我们的攻击中起到了明显的效果。



图 2. 使用 BIM 生成的对抗样本（左）使用 EOT 生成的对抗样本（右）

### 2.2.5 $L_0$ , $L_2$ or $L_\infty$

基本的 BIM 攻击采用的为  $L_\infty$  norm, 即每次迭代在每个像素上改变相同的步长, 但显然会造成图片的 MSE 较大, 因此我们使用了基于  $L_2$  norm 的 BIM, 在每次迭代中, 约束整个图片扰动的大小, 并根据梯度合理分配步长, 这样可以根据 Jacobian Matrix, 让更重要的像素点更改更大的值。当然攻击者也可以选择  $L_0$  norm, 仅仅修改认为重要的几个像素点。

## 2.3. 结果展示

初赛结果中, 我们本地 MSE 约为 9, 提交后线上 MSE 为 11.21995。生成的图片如图2所示。在以上描述过的策略中, 最重要的为 EOT 的使用。使用 EOT 攻击后的图像明显生成了更为规整的扰动, 直观上看, 类似于在对抗样本中编码了其他分类所具有的特征, 并且针对关键区域的修改更为明显。

## 3. 复赛

### 3.1. 复赛介绍

复赛阶段一共需要攻击五个模型, 其中三个模型与初赛的模型一致。另外两个, 一个是由 AutoDL 技术训练的模型, 一个是人工加固的灰盒模型, 其中灰盒模型网络结构为 ResNeXt50. 初赛的三个模型在最后总分的权重为 0.2, 另外两个模型占最后总分的权重都为 0.4. 复赛的目标与初赛相同, 利用主办方提供的数据和模型, 将指定的 120 张图片样本生成为攻击样本, 主办方根据选手提供的攻击样本在后台使用上述 5 个 Target Model 进行评估, 只要使 Target Model 分类结果与 Label 不一致, 则判定为攻击成功。样本攻击成功

数越多、扰动越小, 得分越高。

### 3.2. 方案介绍

根据最近的一些工作来看, 在图像分类场景上, 通常我们训练的卷积神经网络模型会偏向学到很多 weak feature[7]或者说是低频特征 [10]、纹理特征 [6], 而这些特征和图片本身的属性有很大的联系 [3]。根据这个观点, adversarial image 通常是把其他类的 weak feature 加在干净的图片上面, 导致模型分类失败, 而这些 weak feature 对于人眼来说经常不容易察觉。因此只要模型没做 Adversarial training, 他们都很容易会被 adversarial image 上的其他类的 weak feature 所欺骗, 对于 AutoDL 训练出的模型同样适用。对于人工加固的灰盒模型, 我们根据 EOT 的思想, 在随机噪声、缩放两种变换里面采样五到六个样本, 在多模型里利用。我们先给输入加上 13.5% 左右的均匀分布的噪声, 然后让图片长宽缩小到原来的  $\frac{1}{k}$ ,  $k = 0, \dots, 4$ , 再还原到原本输入的尺寸。因此我们的攻击的可以看成是优化如下问题:

$$\begin{aligned} & \underset{x'}{\operatorname{argmax}} \mathbb{E}_{t \sim T} [L(x', y_{gt})] \\ & \text{s.t. } \mathbb{D}(x', x) < \epsilon, x \in [0, 1]^d \end{aligned}$$

这里  $x'$  是从  $x$  生成的对抗样本,  $T$  为变换集合,  $L$  为损失函数,  $y_{gt}$  为正确的标签,  $\mathbb{D}$  为对抗样本和干净样本之间的距离, 这里我们用 L2 范数,  $\epsilon$  为我们所使用的 L2 范数的上限。通过在一定的扰动范围内优化损失函数找到对抗样本  $x'$ 。

#### 3.2.1 增加新模型

不同的网络结构本质上也是一种 transformation。在复赛中, 为了提高样本的迁移性, 我们训练了新的模型, 并确保了对抗样本能够在本地所有模型上都可以攻击成功。为了更直观的展示该方法的有效性, 我们在表 1 中展示了依次攻击不同模型后, 生成对抗样本的提交结果。从 1 中可以看出, 增加新模型并攻击成功, 可以显著增加对抗样本的迁移性。

#### 3.2.2 困难样本处理

对于一些迁移性较差的图片, 我们采用提高 confidence 和增强扰动, 如图3, 我们使用 untargeted attack

Model Name	Offline MSE	Online MSE	Online Score	Num. of Transfer Failed
MoblibeNet	1.95	NA	NA	NA
+Resnext50	2.54	29.27	64.41	$\approx 133$
+Eff-B0	2.93	20.38	74.75	$\approx 87$
+ResNet2	4.47	15.27	82.53	$\approx 54$
+SE-ResNet34	4.71	14.88	83.32	$\approx 51$
+Shuf-v2	5.02	12.53	85.40	$\approx 38$
+MobileNetV2	5.33	10.19	89.03	$\approx 24$
+SE-ResNet50-vd	6.10	9.31	90.49	$\approx 16$
+SQU	6.19	9.18	90.75	$\approx 15$
+ShuffleNetV2	7.51	8.51	92.58	$\approx 5$
+ResNet34	7.79	8.58	92.67	$\approx 4$

表 1. 攻击模型数量与分数之间的关系



图 3. 第 120 张图片，迁移性较差，增加扰动，MSE 约为 17



图 4. 第 48 张图片，有两只狗，untargeted attack 迁移性较差，换用 targeted attack，MSE 约为 7

使 confidence 提高到 45，同时增加 MSE 到 17。又如图4，我们发现使用 untargeted attack 会出现不同模型输出不同标签，这样为了攻击新的模型而添加的扰动会影响攻击之前模型的效果，导致在新的模型上迁移性较差。因此，这里我们利用 Least Likely Targeted Attack，使为了攻击不同模型而增加的扰动朝相同方向增长，提高迁移性。

**致谢.** 感谢 Paddlepaddle 提供了本次比赛的代码环境，感谢 AIStudio 提供算力供选手使用。

## 参考文献

- [1] A. Athalye, L. Engstrom, A. Ilyas, and K. Kwok. Synthesizing robust adversarial examples. arXiv preprint arXiv:1707.07397, 2017. 2
- [2] N. Carlini and D. Wagner. Towards evaluating the robustness of neural networks. In 2017 IEEE Symposium on Security and Privacy (SP), pages 39–57. IEEE, 2017. 2
- [3] G. W. Ding, K. Y. C. Lui, X. Jin, L. Wang, and R. Huang. On the sensitivity of adversarial robustness to input data distributions. CoRR, abs/1902.08336, 2019. 3
- [4] Y. Dong, F. Liao, T. Pang, H. Su, J. Zhu, X. Hu, and J. Li. Boosting adversarial attacks with momentum. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018. 2
- [5] Y. Duan, Z. Zhao, L. Bu, and F. Song. Things you may not know about adversarial example: A black-box adversarial image attack. arXiv preprint arXiv:1905.07672, 2019. 2
- [6] R. Geirhos, P. Rubisch, C. Michaelis, M. Bethge, F. A. Wichmann, and W. Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. CoRR, abs/1811.12231, 2018. 3
- [7] A. Ilyas, S. Santurkar, D. Tsipras, L. Engstrom, B. Tran, and A. Madry. Adversarial examples are not bugs, they are features. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, Advances in Neural Information Processing Systems 32, pages 125–136. Curran Associates, Inc., 2019. 3
- [8] A. Kurakin, I. Goodfellow, and S. Bengio. Adversarial examples in the physical world. In Proceedings of International Conference on Learning Representations, 2017. 2
- [9] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu. Towards deep learning models resistant to adversarial attacks. In Proceedings of International Conference on Learning Representations, 2018. 2
- [10] N. Rahaman, A. Baratin, D. Arpit, F. Draxler, M. Lin, F. Hamprecht, Y. Bengio, and A. Courville. On the spectral bias of neural networks. In K. Chaudhuri and R. Salakhutdinov, editors, Proceedings of the 36th International Conference on Machine Learning, volume 97 of Proceedings of Machine Learning Research, pages 5301–5310, Long Beach, California, USA, 09–15 Jun 2019. PMLR. 3