# 深度学习为何不鲁棒？万字长文综述对抗鲁棒性

极市平台 2022-11-06 22:33:31 发表于广东 手机阅读

以下文章来源于王晋东不在家，作者陶略

**王晋东不在家**

分享科研与研究生活的点点滴滴，包括但不限于：机器学习、迁移学习、元学习等，以...





作者丨陶略

来源丨王晋东不在家

编辑丨极市平台

**极市导读**

本文从对抗样本的发现过程、研究意义、寻找对抗样本、增强模型对抗鲁棒性、对抗鲁棒性问题导致的攻防竞赛现象方法等方面出发，对深度神经网络的鲁棒性问题进行了深入探讨。加入极市CV技术交流群，走在计算机视觉的最前沿

**作者介绍：**陶略，PARNEC Group，南京航空航天大学

**个人主页：**https://www.zhihu.com/people/michael-63-42

本文介绍**对抗鲁棒性**，主要针对的是深度学习时代下深度神经网络的鲁棒性问题。大部分内容探讨**最基本的机器学习分类任务**。

囿于有限的认知水平，本文是个人偏好（bias）的产物，不代表所有人的观点。

## 目录

1. 对抗样本的发现

2. 我们为什么要研究对抗样本现象?

3. 寻找对抗样本：一个约束优化问题

- 替代损失
- 优化算法
- 度量函数

4. 增强模型对抗鲁棒性

5. 对抗鲁棒性问题导致的攻防竞赛现象

6. 深度学习为何不鲁棒?

- 人类是鲁棒的吗?

7. 对抗鲁棒性带来的额外益处

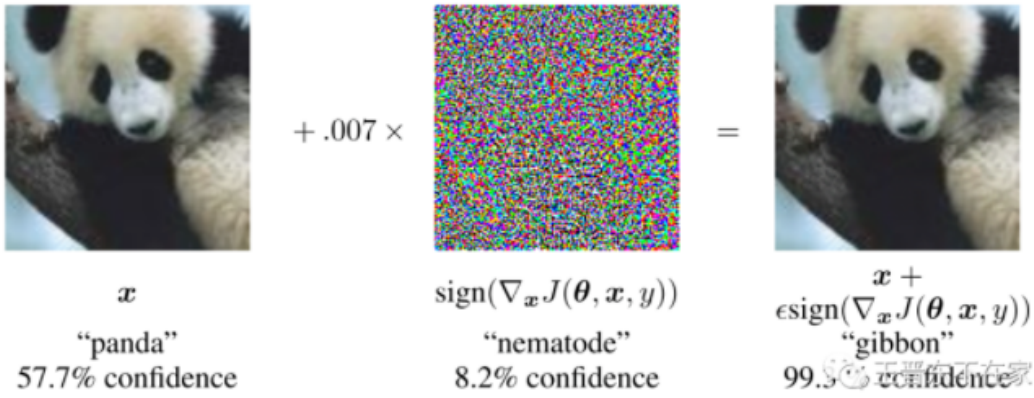
8. 大厦已然建成?

## 1 对抗样本的发现

What has been is what will be, and what has been done is what will be done; there is nothing new under the sun.[1]

日光之下并无新事。

下图是用FGSM方法[2]生成的一个对抗样本。

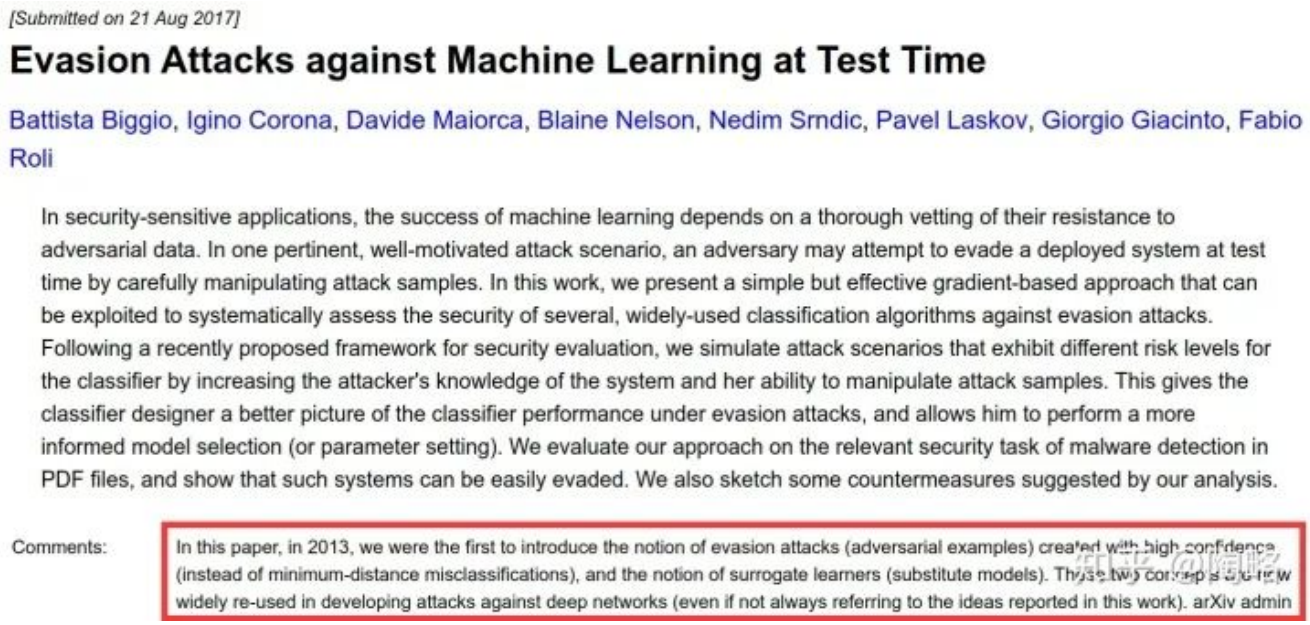



一个对抗样本的例子。左：一张干净图片，被模型正确分类为熊猫。中：扰动。右：扰动后的熊猫图片，被模型识别为长臂猿。

**对抗样本（Adversarial Examples）**一词首次出现在Szegedy等人在ICLR 2014的一篇论文里[3]。通过对测试集的图片添加微小的扰动，导致神经网络对其误分类（而人类依然能够正确分类），可以使得经过良好训练的神经网络的准确率降为0。

但早在本世纪初，就有许多工作在设计传统机器学习模型的敌手（Adversary），它能够操纵输入数据从而欺骗分类器（比如，通过向垃圾邮件里添加特殊词汇以欺骗垃圾邮件检测器，而检测器也可以随敌手的出现而改进自己的词库，从而获得更好的性能[4]）。当时就在许多领域里掀起了攻防角逐（arms races），比如欺骗计算机入侵检测系统、躲避空中监视系统、操纵搜索引擎排序系统等[5]。

事实上，第一个针对神经网络分类器的敌手是Biggio等人在2013年设计的，他们称之为规避攻击（Evasion Attacks）[6]。Biggio等人显然有些气愤，因为后来大部分研究对抗鲁棒性的论文只引用Szegedy等人的那篇论文[3]，而不引用他们这篇；大部分论文都使用[3]里发明的术语，而不使用[6]里的术语（evasion attacks = adversarial examples，surrogate learners = substitute models）。Biggio等人在2017年终于把他们在2013年的那篇论文挂到了arxiv上面，并附了一段评语：



*[Submitted on 21 Aug 2017]*

**Evasion Attacks against Machine Learning at Test Time**

Battista Biggio, Igino Corona, Davide Maiorca, Blaine Nelson, Nedim Srndic, Pavel Laskov, Giorgio Giacinto, Fabio Roli

In security-sensitive applications, the success of machine learning depends on a thorough vetting of their resistance to adversarial data. In one pertinent, well-motivated attack scenario, an adversary may attempt to evade a deployed system at test time by carefully manipulating attack samples. In this work, we present a simple but effective gradient-based approach that can be exploited to systematically assess the security of several, widely-used classification algorithms against evasion attacks. Following a recently proposed framework for security evaluation, we simulate attack scenarios that exhibit different risk levels for the classifier by increasing the attacker's knowledge of the system and her ability to manipulate attack samples. This gives the classifier designer a better picture of the classifier performance under evasion attacks, and allows him to perform a more informed model selection (or parameter setting). We evaluate our approach on the relevant security task of malware detection in PDF files, and show that such systems can be easily evaded. We also sketch some countermeasures suggested by our analysis.

Comments: In this paper, in 2013, we were the first to introduce the notion of evasion attacks (adversarial examples) created with high confidence (instead of minimum-distance misclassifications), and the notion of surrogate learners (substitute models). Those two concepts are now widely re-used in developing attacks against deep networks (even if not always referring to the ideas reported in this work). arXiv admin

所以日光之下真的并无新事吗？当时的明月换拨人看？[7]

1. 我们新发现的对抗样本与十多年前的没有区别吗？
2. 我们深度学习时代的攻防竞赛是在重蹈覆辙吗？
3. 我们要因为[3]没有引用[6]而把[3]批判一番吗？

我个人认为，答案都是否定的。


虽然从抽象的形式上看，十多年前的对抗样本与现在的并无二致，都是“攻击者精心设计的导致机器学习模型出错的输入”[8]。但时代变了，模型变了。能力越大，责任越大。深度学习复兴给大家带来了成功的喜悦，在许多任务上，以前模型做不到或做不好的事情，现在能出人意料地做得很好，需要的似乎只有数据和算力，深度学习一时间成了人工智能的代名词，学术界工业界一片欣欣向荣的景象。而深度网络的优异性能和深度网络在对抗攻击下的灾难性脆弱似乎是同一枚硬币的两面[9]。深度网络中的对抗样本现象给狂热的人群敲墙了警钟：我们的深度网络似乎什么都知道，但其实什么都不知道（You know nothing!）。

在传统机器学习的攻防竞赛中，攻击和防御都围绕着手工特征（handcrafted features）做文章，且主要是在浅层模型应用过的领域（如垃圾邮件过滤器，入侵检测，生物认证，欺诈检测）。而近些年来针对深度网络鲁棒性的攻防竞赛主要对原始数据（raw data）做扰动，且集中在深度学习最擅长的领域（视觉、听觉、自然语言处理等），同时更加强调对抗扰动对人类的影响，即人类察觉不到扰动，而模型却过分地敏感，以突出模型与人类的差距。更多讨论将推迟到 本文第五节“对抗鲁棒性问题导致的攻防竞赛现象”。

任何研究者都不希望自己的研究被埋没在书山书海中。但我倾向于相信[6]和[3]是两个独立的且同期的研究成果（ICLR 2014是在2013年12月截稿），[3]没有引用[6]的原因可能很简单：

作者并不知道同行的工作。但这并不代表[3]和[6]的工作是等价的！实际上，[6]是延续传统机器学习里的那一套思路，在二分类的任务下对 Adversary 进行了设定，通过强调自己发现的机器学习系统存在的安全风险来讲故事。而[3]没有理会安全语境下那些繁琐的设置，从新的角度出发，旨在寻找模型的盲点（Blind Spots in Neural Networks）：虽然深度学习的泛化性能非常好，但在高维数据流形附近存在着许多低概率的洞洞（pockets），在这些洞洞上模型表现很差，而我们人类在这些洞洞上的表现依然非常好。深度学习的成功让我们看到了人工智能的晴朗天空，但[3]告诉我们，晴朗天空中的远处漂浮着一朵乌云：你们可以暂时假装看不见，但它迟早会飘到你头顶。

## 2 我们为什么要研究对抗样本现象？

First of all，对抗样本突出了**机器智能与人类智能的差距**。我们希望我们的模型是“人工智能”而不是“人工智障”。我们最远大的理想就是训练出像人类一样强大的分类器帮我们执行分类任务，这样我们就可以躺着喝咖啡了。因此像把熊猫识别成长臂猿这样的现状是完全无法忍受的，我们会气得跺脚：虽然你每个神经元连接、每个参数我都清清楚楚，但我还是无法理解你的脑子里究竟在想些什么，所以我要骂你是个智障黑盒非线性深度神经。

Last but not least，**安全性**。因为你看起来让别人觉得你很能干，许多人就请你去帮他们做事。但我知道你外强中干，一旦有个敌手在旁边干扰你，你就崩溃了。外面有很多敌手在恶意设计扰动，让自动驾驶汽车直线拐弯[10]，让目标检测失灵[11]，让人脸识别系统失效[12]。我不放心你做事啊，尤其是人命关天的任务。

## 3 寻找对抗样本：一个约束优化问题

要消除盲点，我们首先要找到盲点。

（盲生，你发现了华点[13].

我们有个分类任务：data $(x, y) \in \mathbb{R}^d \times \{1, \ldots, C\}$ from a distribution $\mathcal{D}$ .我们有个标注员（labeling oracle） $\mathcal{O}$ ： $\mathbb{R}^d \to \{1, \ldots, C\}$ .注意到，对任何 $(x, y) \sim \mathcal{D}$ ，我们有 $y = \mathcal{O}(x)$ .

常规的机器学习目标是学习一个分类器 $f$ ，使得在分布 $\mathcal{D}$ 下期望风险最小。而终极的人工智能目标是学习一个分类器 $f$ ，使得对任何 $x \in \mathbb{R}^d$ ，有 $f(x) = \mathcal{O}(x)$ 。

所以广义上来说，分类器的任何一个能够被发现的失败点都是对抗样本[14]。

**对抗样本（广义）**：一个对抗样本是任何满足 $f(x^*) \neq \mathcal{O}(x^*)$ 的输入 $x^*$ 。

但上面这个定义由于过于依赖标注员 $\mathcal{O}$ 而太难处理，我们通常习惯于研究它的一个放松版本，这个版本限制敌手只能在一个干净样本 $x$ 的邻域附近寻找对抗样本[15]。

**对抗样本**：给定一个分类器 $f$ 和一个被正确分类的输入 $(x, y) \sim \mathcal{D}$ （即 $f(x) = y$ ），一个$\epsilon$-bounded 对抗样本 是一个输入 $x^*$ ，它满足：$f(x^*) \neq y$ 且 $\|x^* - x\| \leq \epsilon$ 。

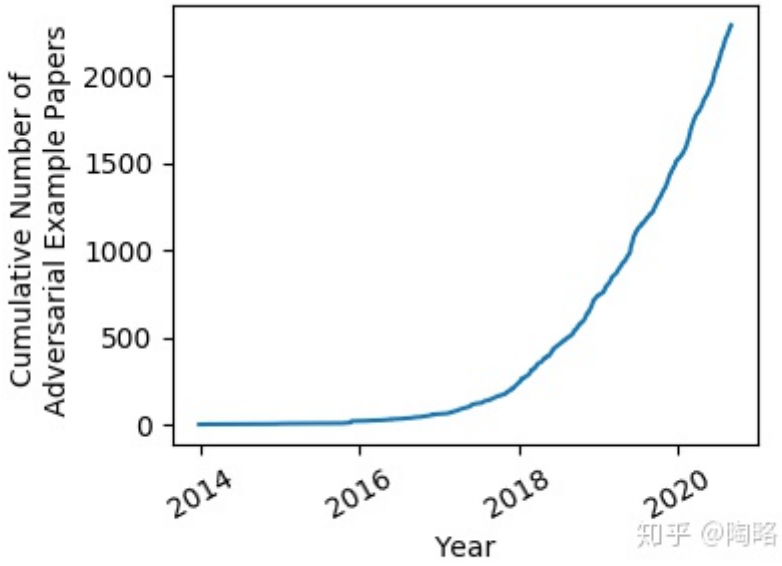
这里的假设是所有满足 $\|x^* - x\| \leq \epsilon$ 的扰动都不会影响标注员的标注（ $\mathcal{O}(x^*) = \mathcal{O}(x)$ ）。这里的度量函数 $\|\cdot\|$ 通常是 $l_p$ 范数。

那么在这个定义下，我们寻找对抗样本这件事，就变成了一个带约束的优化问题：

$$x_{\text{adv}} = \underset{\|x'-x\| \leq \epsilon}{\arg\max}\, \mathbf{1}(f(x') \neq y)$$

其中 $\mathbf{1}(\cdot)$ 代表指示函数（aka 0-1损失）。

下图显示了近年来对抗样本相关论文的累计数目[16]：




对抗样本论文的累计数量随年份呈现指数级增长趋势

对抗样本论文的累计数量随年份呈现指数级增长趋势

就这样一个简短的对抗样本问题，每年就有成百上千篇论文在研究。这有啥好研究的呢？

我粗略地从三个方面扯扯：替代损失、优化方法、度量函数。

### 替代损失

0-1损失不便于用梯度方法优化，故许多人在寻找替代损失。这里 just to name a few.

最常见的是使用交叉熵损失 $\mathcal{L}_{\text{CE}}$ 。Goodfellow等人为了突出问题的严重性，使用 $l_\infty$ 范数做度量，对 $f(x)$ 做一阶近似后直接获得了 $x_{\text{adv}}$ 的闭式解（实际上就是一步梯度上升）：$x_{\text{adv}} = x + \epsilon\,\text{sign}(\nabla_x \mathcal{L}_{\text{CE}}(f(x), y))$ ，该方法称作 FGSM[17]。

而后 Kurakin等人为了产生攻击性更强的对抗样本，设计了迭代方法 BIM[18]：$x^{n+1}=\mathrm{Clip}_{x,\epsilon}(x^n+\alpha\,\mathrm{sign}(\nabla_x\mathcal{L}_{\mathrm{CE}}(f(x),y)))$ 。这实际上就是后来Madry等人那篇非常seminal的文章里做inner maximization的时候，从优化的角度使用的投影梯度下降方法（PGD）在 $l_\infty$ 下的特例[19]。

同期Carlini和Wagner在信息安全顶会 S&P 上发表的论文里针对该问题设计了一个专门的替代损失[20]，类似于hinge loss：$\mathcal{L}_{\mathrm{C\&W}}=\max\big(\max_{i\neq l'}(\log[f(x)]_i)-\log[f(x)]_{l'},-\kappa\big)$ 。在当时的论文里C&W没有用PGD处理约束，而是使用Lagrangian relaxation，并一举攻破了当时许多看起来有效的防御方法。在之后的追求最小扰动幅度攻击的论文里，大都使用Lagrangian relaxation；但在之后的追求对抗鲁棒性的文章里，大都使用PGD。

Gowal等人认为在这个高度非凸非光滑的优化问题里随机初始化的PGD并不高效，于是设计了一个新的替代损失，称作 MultiTargeted[21]。Tashiro等人则提出一种更高效的初始化方法[22]。Li等人则认为我们应该去寻找一个最坏的对抗扰动分布，而不是一个最坏的对抗扰动点（有点贝叶斯的味道了），并且运用NES技巧，优化时可以无需梯度信息[23]。

还有些论文在原问题上又加了些其它的目标。比如让一个对抗扰动同时欺骗多个模型[24]。比如让一个扰动同时使多张图片变成对抗样本[25]。比如让一个对抗扰动更容易去欺骗一个没见过的模型（即对抗样本的迁移性）[26]。

**优化算法**

我们有了合适的替代损失了，就需要用高效的优化算法去优化这个问题。上面提到的PGD和Lagrangian relaxation通常需要一阶的梯度信息，因而被称为 First-order 算法。还有一种在优化领域常见的算法叫做 Frank-Wolfe（aka conditional gradient method or projection free method）被搬出来解决上面那个约束优化问题了[27]，它用Linear Minimization Oracle来替代Projection，与PGD相比的优势是，在某些约束下计算复杂度更低，比如 $l_1$ 范数、核范数。

有些论文认为敌手不是什么时候都能获取梯度信息的，比如Google Cloud Vision API，比如要攻击一个未见过的目标检测器。这些模型对于敌手来说都是黑盒。敌手只能知道模型的输入和输出，而不知道梯度。这时候 Zeroth-order 优化算法就被拿出来玩了。在对抗样本领域又被称为黑盒攻击（Black-box attacks），与之对应的是前文的白盒攻击（White-box attacks）。

黑盒攻击的花样就多了。有的人直接使用一个白盒模型来作为黑盒的替代。也有人把白盒替代和zeroth-order结合做的[28]。更多论文在直接做zeroth-order。最开始利用最简单的 coordinate梯度估计器去揣摩模型的梯度，然后用在高维数据下更高效的Gaussion和Uniform梯度估计器。之后一大批论文针对梯度估计器的查询代价过高的问题，认为实际系统的API不会允许你查询太多次，于是大家都限制自己在有限的查询次数下做性能的比较然后追求SOTA，这方面的论文我看得不太多，印象最深刻的是这篇[29]。

之后就有许多搞优化的朋友进场表演了。上面提到的所有替代损失，不管是用 first-order算法 还是用 zeroth-order算法 做优化，都存在一个可以改进的地方，就是加速。当然许多搞优化的文章只是把对抗样本问题作为一个小应用以证明自己存在的意义。同时也有些优化文章专门做对抗攻击，比如[30]和[31]。这方面的文章我看得也不太多。

**度量函数**

这其实是一个最本质的问题，也是一个最难搞的东西。

早期为了去更方便地且深刻地认识对抗样本现象，从而推动社区发展，不管是从理论上还是从实验上，我们都习惯于选择使用 $l_p$ 范数作为 toy 来玩。但很明显， $l_p$ 范数并不适合许多结构性很强的高维数据，比如图像。而且在 $l_p$ 范数下对所有样本使用同一个 $\epsilon$ 也是不合理的，有些样本天然离分类面近，一个很小的扰动就改变它的label了；而有些样本天然离分类面远，可以忍受较大的扰动；少量文章对这件事做了初步探索[32][33][34]。

因此有些论文在探索其它的样本邻域。比如旋转和平移诱导出的邻域[35]，Wasserstein邻域[36]，Trace-Norm邻域[37]，函数化扰动诱导出的邻域[38]，形变扰动[39][40]等。以上手工构造的扰动邻域容易被人诟病的地方就是，我们依旧认为真正的符合人类感知的距离度量是无法被简单的数学表达式刻画的。

于是有许多论文利用各种生成模型的潜空间来诱导出扰动邻域[41][42][43][44][45][46]。同时也有人在深度网络分类器的特征空间做距离度量（这类距离被称为Perceptual Distance）来诱导出扰动邻域[47]。

虽然这件事有不少人在做，但都无法真正做好。构造一个与人类感知一致的度量有多难呢？这件事的难度被证明等价于构造一个与标注员 $\mathcal{O}$ 一致的分类器 $f$ （即 $f(x)=\mathcal{O}(x)$ ）[15]，i.e.， $f$ 完美地解决了$\mathcal{O}$的分类任务，我们可以躺着喝咖啡了。

革命尚未成功，同志仍需努力！

## 4 增强模型对抗鲁棒性

黑客寻找对抗样本的目的是去攻击模型；而白客的目的是修补漏洞。

（红客的目的是啥？

白客通常有两种防御方法：一种是发现有人攻击漏洞了就阻止这个人进入系统；另一种是尽量修补已知的漏洞。

对应到对抗鲁棒性这里，就是两种策略：1、设计检测方法，检测到对抗样本后拒绝分类；2、

真正增加模型鲁棒性，使模型正确分类对抗样本。

这里先讲讲第二个策略——增强模型鲁棒性。

这里本质上最有效的方法就是 Madry 等人在 ICLR 2018 上的那篇非常 seminal 的论文《Towards Deep Learning Models Resistant to Adversarial Attacks》。近三年大浪淘沙后剩下的有效的鲁棒化方法大都是它的变种。

它在经验风险最小化问题的基础上，增加了一个内部的敌手，变成了一个 min-max 问题：

$$\min_{\theta} \mathbb{E}_{(x,y)\sim\mathcal{D}} \left[ \max_{\|x'-x\|\leq\epsilon} \mathcal{L}(f_\theta(x'), y) \right]$$

上式实际上是一个经典的鲁棒优化问题（Robust Optimization），同时在机器学习里也被称作对抗训练（Adversarial Training）。上面这个需要被最小化的期望值也被称为对抗风险（adversarial risk），与之对应的传统的风险被称为自然风险（natural risk）。

把对抗样本加入到训练集中以提高模型鲁棒性在2017年9月ICLR 2018投稿时已经不是什么新鲜玩意了，但当时大部分人还是在用FGSM生成的对抗样本来做对抗训练，Madry等人最大的贡献是正本清源，从鲁棒优化的角度告诉大家，我们应该用PGD，而且对于这个非凸的问题，我们应该多次随机初始化以寻找更好的局部极值点。

在Madry等人那篇文章受到重视之前，大量论文在使用FGSM做对抗训练，这种方式只能使模型对FGSM攻击鲁棒，而C&W和PGD之类的迭代攻击依旧能够成功找到许多对抗样本。对此，最先用FGSM做对抗训练的Goodfellow也很头疼，他曾在twitter上表示："When I invented adversarial training as a defense against adversarial examples, I focused on making it as cheap and scalable as possible."[48]

鲁棒优化是个min-max问题，如果我们用10-steps的PGD做对抗训练，那么我们的训练代价就是普通训练的10倍以上。对于MNIST这种数据集比较无所谓，对于CIFAR-10就有点吃力（e.g. 在V100上用Wide-ResNet-28-10搞一个PGD-10的对抗训练需要24~40个小时），对于ImageNet一般实验室就受不了了。Kaiming He等人曾经用128块V100在ImageNet上做对抗训练，一个ResNet-101需要38 hours，一个ResNet-152需要52 hours[49]。

所以后来就有些论文在思考怎样减少对抗训练的计算代价了。NeurIPS 2019同时收了两篇通过玩弄做PGD时的梯度流来减少计算量：一篇是《Adversarial Training for Free!》[50]，另一篇是Yiping Lu大佬等人的《You Only Propagate Once》[51]。ICLR 2020出了篇改进 FGSM-based adversarial training 的[52]，深得Goodfellow欢心[48]。后面还有篇利用对抗样本在epoch之间的迁移性质来减少计算量[53]。但这些加速对抗训练的方法或多或少都会对模型性能造成负面影响，且在推广到较大规模数据集和较大扰动幅度上可能存在困难。

## 5 对抗鲁棒性问题导致的攻防竞赛现象

上文提到的原始的基于FGSM的对抗训练作为一种防御方法，之后被更强大的迭代式的攻击方法击破，后来基于PGD的对抗训练的防御方法又进一步增强了模型的鲁棒性。这是最淳朴的攻防竞赛。

然而，由于敌手的攻击大部分都严重依赖模型的梯度，许多花里胡哨的被称为混淆梯度（Obfuscated Gradients）的防御方法被提出。这些防御方法的作者有意或无意地使得分类器的梯度信息无效，从而使得基于梯度的攻击方法失效。

比如在模型中添加不可微结构，引入数值不稳定性质，随机化模型中的部分组件，把输入先进行随机变换后再送入网络，将网络的输出作为输入多次循环迭代以造成梯度爆炸或消失等。

从安全性的角度来看，这样的技巧是有一定意义的。但从鲁棒性上看，这样的技巧并不能真正解决漏洞。它们只是让基于梯度的敌手找不到对抗样本，但漏洞本身还是存在，可能会有别的敌手能够找到。

2017年，Carlini和Wagner宣布大名鼎鼎的混淆梯度方法 defensive distillation 无效[20]。

2018年，Athalye、Carlini和Wagner在ICML 2018上深刻总结并抨击了混淆梯度，并且一举攻破了7个发表在ICLR 2018上的防御方法，只有Madry那篇基于PGD的对抗训练无法击破，该论文直接拿下ICML 2018 Best Paper Awards，名声大噪[54]。

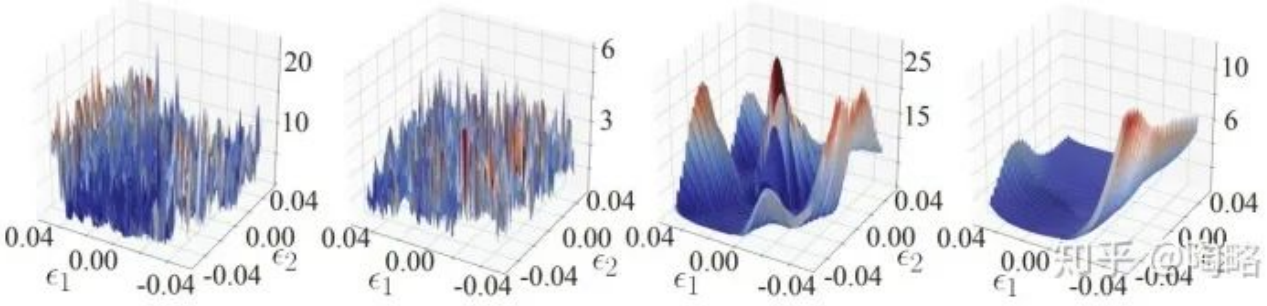
在这之前，许多防御方法提出防御策略，而并没有设计对应的适应性的攻击方法来评估防御策略的有效性。这给社区带来的是虚假的安全感。Athalye、Carlini和Wagner击破了这种虚假的安全感，树立了自己的地位。

在2018年之后，又有许多防御方法被提出，这些论文大都吸取了教训，在方法评估部分设计了适应性的攻击方法来测试模型鲁棒性。但有时，有意或无意，这些自我评估还是不可靠。

2020年1月，Carlini等人放了篇《On Adaptive Attacks to Adversarial Example Defenses》在arxiv上，他们以示范性和教育性的目的，在文中列举了13个最近发表在ICLR、ICML和NeurIPS上的防御方法，并教人如何正确地评估这些方法。他们展示了如何规避这13篇论文里提出的防御方法，并且证明这些模型的鲁棒精度远远低于他们原始论文里所宣称的精度。

这是一篇爽文，怼天怼地怼空气。比方说，一篇发表在ICLR 2020上名为 k-Winners Take All 的防御方法[55]，被Carlini等人评价为“the existence of such a defense would go against common wisdom”[56]，因为大佬们在ICML 2018就说了，这种混淆梯度形式的方法有根本性的缺陷，而这篇论文还在故意设计混淆梯度。

但有一说一，这篇论文本身还挺好玩的，你看它做了个啥事就知道了：




基于梯度的攻击过程的损失曲面（Loss Landscape）。第一张和第二张图使用了 k-Winners Take All技巧，第三张和第四张是正常模型的结果。

可以看出，k-Winners Take All技巧把 loss 对 input space 的 loss landscape 从比较光滑变成了极度扭曲的形状。从而造成了非常严重的混淆梯度现象。

好玩但是漏洞依然存在。

另外，有一系列防御工作是去尽可能检测对抗样本。最开始是直接训练一个二分类器去判断是干净样本还是对抗样本。而这个二分类器本身就是不鲁棒的，所以没啥用。也有些工作提出某些统计量来做检测。Carlini和Wagner也有篇工作专门攻破这些基于检测的防御方法[57]。后来可能有的检测器把握了问题本质，似乎有效，至少目前还没人来锤它[58]。

但无论如何，对抗样本和干净样本对人来说区别不大，甚至有时人眼都无法区分出哪个是对抗图片。而基于检测的防御方法要去做这件对人来说反直觉的事，我个人不太偏好。而且就算检测出来了，只能做拒绝，还是无法做分类识别，鲁棒的分类器还是无法实现。

## 6 深度学习为何不鲁棒?

传统的基于经验风险最小化的神经网络为何不鲁棒？为何需要对抗训练才能使其鲁棒？为什么越鲁棒的模型，在干净测试集上的精度反而越低[59]?

有许多论文对此进行了解释，其中我认为最接近本质的是 Madry等人的这篇《Adversarial Examples Are Not Bugs, They Are Features》[60]。

对抗样本不是缺陷，而是特征。对抗脆弱性是我们的模型对数据里易于泛化的特征过于敏感的直接结果。这里的易于泛化的特征，我们可以把它理解为隐藏在图片像素里的某些模式（pattern）。

它把特征分成两类：有用且鲁棒的特征$F_r$（useful, robust features），有用但不鲁棒的特征$F_{nr}$（useful, non-robust features）。这里的有用指的是能帮助分类器做分类任务的特征。这里的$F_{nr}$可能是某些高频特征，但也不完全是[61]。

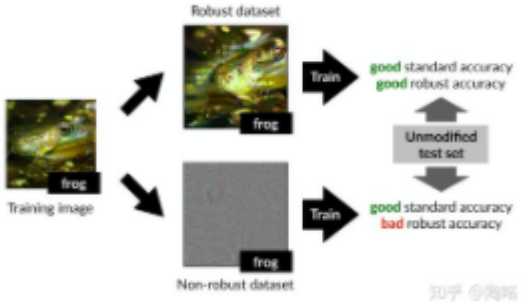
它认为这两类特征天然存在与我们的图像数据集中。虽然我们人类主要是依赖于 $F_r$ 做分类任务，我们人类对 $F_{nr}$ 天然无感。但是我们的深度网络在训练过程中，是竭尽全力地在学习所有

有利于减小目标函数的特征，因此我们的模型会过于依赖这部分有用但是不鲁棒的特征 $F_{nr}$ ，而这些不鲁棒的特征正是导致模型对抗脆弱性的关键：在通常情况下，这些 $F_{nr}$ 是很有用的，它让模型获得了令人诧异的泛化性能；然而一旦有敌手通过操纵 $F_{nr}$ 来攻击模型，它就让模型的精度直接降为0，而人类的精度并不会受到影响。

正所谓，成也萧何败萧何[9]。

那么如何证明我们的数据集里存在这两类特征呢？如果我们的数据集里仅仅含有$F_r$而不存在$F_{nr}$，是不是普通训练的模型就会具备对抗鲁棒性了？

这篇文章里设计了一个实验。它利用一个现存的鲁棒分类器和一个原始数据集构造了两个新的数据集：一个数据集仅含有 $F_r$ ，另一个数据集仅含有 $F_{nr}$ ：




实验表明，在一个鲁棒的数据集上普通地训练一个模型，能得到良好的标准精度和鲁棒精度；而在一个不鲁棒的数据集上普通训练的模型，能有良好的标准精度但是不鲁棒。

这篇文章很好地回答了前面三个问题。模型不鲁棒是因为数据集里存在不鲁棒的特征 $F_{nr}$ ，对抗训练能够使模型避免依赖 $F_{nr}$ ，但 $F_{nr}$ 其实对标准精度有益，所以对抗训练会使得模型的标准精度下降。

所以，from this perspective，模型的对抗脆弱性是一个纯粹以人为中心的现象，因为从标准的监督学习的视角看来，$F_r$和$F_{nr}$其实一样重要。

对抗鲁棒性揭示了深度学习与人类的差异之处。我们接下来要做的，是把更多的人类先验（human priors）编码到模型的训练过程之中。我们需要的不是对某一种特定的扰动邻域过分鲁棒的分类器，除非这个扰动邻域与人类的距离感知一致[15]。我们需要的是像人一样的分类器，以实现我们最远大的理想——训练分类器帮我们做事然后安心地躺着喝咖啡。

**人类是鲁棒的吗?**

前面批判了这么久深度神经网络，现在来批判一下我们人类自己。

就图像分类任务来说，我们人类当然是最鲁棒的，因为在这个任务下，我们人类就是 Oracle 啊，所有的模型都是学习我们人类的标注，模型如果与人类的表现不一致，那一定是模型错

了，人不会错。

人真的不会错吗？这里有两个例子，都来自纪录片《深入大脑》[62]。

第一个例子是一个心理视觉实验，人受到参照物的影响，对色块的颜色产生了误判。

**Color Illusion**

https://www.zhihu.com/zvideo/1284242038373048320

第二个例子是一个简单实验，利用穿颅磁刺激技术（TMS）干扰人的决策：举起左手或举起右手。TMS是在无需手术的情况下，在脑皮质以电流刺激特定区域，它可以用来引发非自主动作。

**穿颅磁刺激技术（TMS）干扰人脑的决策**

https://www.zhihu.com/zvideo/1284242655015063552

这两个例子都表明，即使是我们人类自身，也存在某些扰动，能够干扰甚至操纵我们的决策。

事实上，在 NeurIPS 2018 就有一篇论文分析了人类的对抗鲁棒性——《Adversarial Examples that Fool both Computer Vision and Time-Limited Humans》[63]。
它通过限制人类做分类任务时决策的可用时长（60ms ~ 70ms），把人类变成一个严格受时间限制的分类器。然后让受时间限制的人类来分类对抗样本，这里的对抗样本是针对深度神经网络分类器生成的，对抗扰动幅度很小，故样本的标签并没有改变。在实验中，这种情形下的人类也可能会被对抗样本成功攻击，即人类的分类准确率降低了。

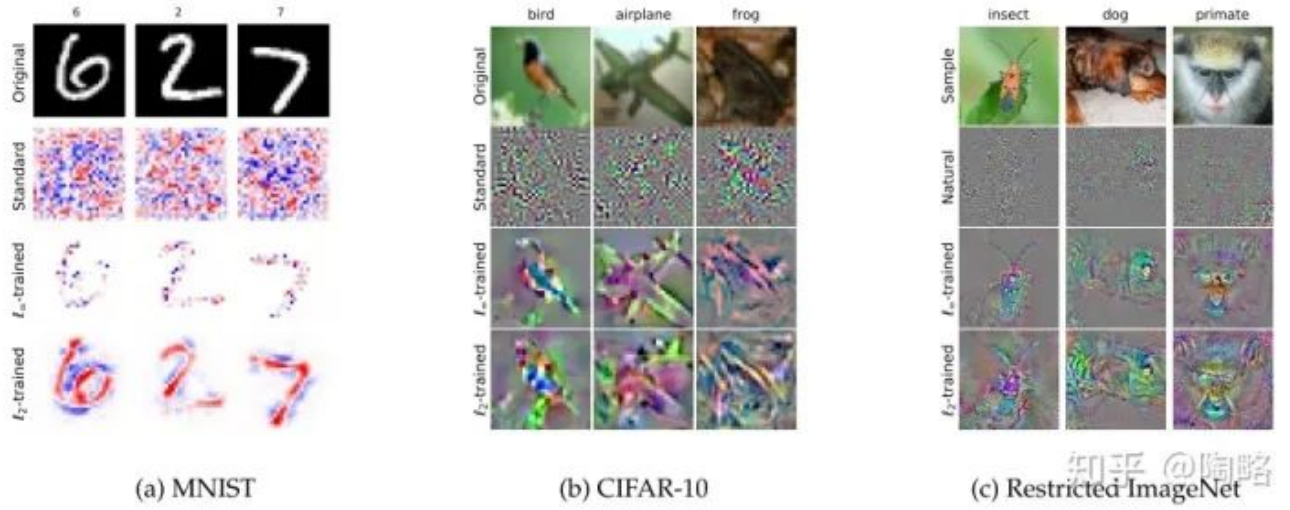
该论文对此现象的解释是，人脑通常依靠横向的和自上而下的脑回路做决策。如果人脑的视觉输入刺激的时间过短，人脑分类器就变成了一个前向传播的神经网络，这种脑子的鲁棒性不太好。这能够一定程度上启发未来我们对深度网络架构的设计。

## 7 对抗鲁棒性带来的额外益处

对抗训练可以增加模型在测试集上的鲁棒精度（即分类对抗样本的准确率）。通过让模型对人类不敏感的信号不敏感，鲁棒的训练目标可能会使得模型学习到更类似我们人脑使用的特征表示。

这种与人类感知更相符的特征表示，会有一些额外的益处。

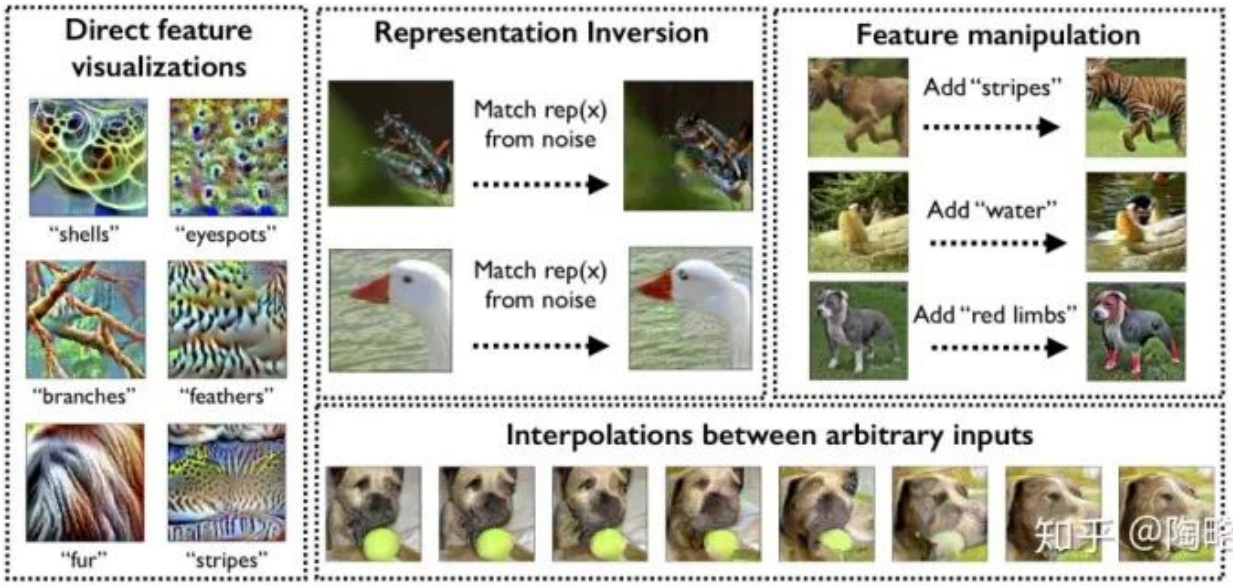
1. 可解释性：鲁棒模型对输入的梯度比普通模型更加有意义[64][65][66]。




(a) MNIST (b) CIFAR-10 (c) Restricted ImageNet



2. 图像合成：利用一个鲁棒的分类器可以做许多从前只有生成模型可以做的图像合成任务[67][68][69]。




3. 异常检测：对抗鲁棒性可以增加异常检测的性能[70][71]。

4. 迁移学习：对抗鲁棒的预训练模型迁移到新任务时具有更好的性能[72][73][74]。

5. What's next? Continual learning? Zero-shot learning? Open-set learning? Noisy label?

## 8 大厦已然建成?

近几年随着社区不断的努力，模型鲁棒性问题得到了深入的理解。我们刻画了对抗攻击的max问题，对抗训练的min-max问题，得到了许多toy任务在理论上的结论，对许多大规模数据集也做了大量的实验分析，我们的对抗攻击和对抗训练也被扩展到了其它机器学习模型和任务上：目标检测、分割、跟踪，自然语言处理，语音识别，图数据，医疗影像分析，行人重识别，度量学习，最近邻模型，决策树等等。

似乎对抗鲁棒性的大厦已经建成，剩下的只有修修补补的工作。

但实际上，剩下的是难摘的果实。

前文提到的距离度量问题是一个，对抗鲁棒性的泛化性能问题也是一个。

事实上，在鲁棒优化框架下，MNIST的性能还说得过去，但到CIFAR-10上就只有可怜的50%~70%的鲁棒分类精度，ImageNet就更别提了。非常多的工作在想方设法地提高泛化能力，比如加正则化项、修改模型结构、更改损失函数、利用未标注的数据等等。效果都差强人意，真正有效的改进能把CIFAR-10的鲁棒精度稳定提升一两个点已经是不错的工作了。

有人可能期望更多的训练数据和更大的深度网络能够解决问题。但问题是，究竟需要多大的数据集、多大的模型才能让我们对模型鲁棒性满意？如果现有的深度学习范式下模型的鲁棒性能对样本数量和模型复杂度的依赖是类似于指数爆炸的呢？解决鲁棒性问题需要新的思想，我们依然任重而道远[75]。

**参考资料**


1. 出自《圣经 旧约 传道书》第一章第九节
2. Explaining and harnessing adversarial examples, ICLR 2015
3. abcdefghij Intriguing properties of neural network,s ICLR 2014
4. "In vivo" spam filtering: a challenge problem for KDD, ACM SIGKDD Explorations Newsletter, 2003
5. Adversarial classification, KDD 2004
6. abcdefgEvasion attacks against machine learning at test time, ECML-PKDD 2013
7. 出自《拓在墙》次崇




## 极市干货

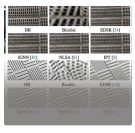
**算法竞赛：往届获奖方案总结以及经验详解 | ACCV2022国际细粒度图像分析挑战赛**

**技术综述：BEV 学术界和工业界方案、优化方法与tricks综述 | PyTorch下的可视化工具（网络结构/训练过程可视化）**

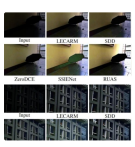
**极视角动态：极视角与华为联合发布基于昇腾AI的「AICE赋能行业解决方案」 | 算法误报怎么办？自训练工具使得算法迭代效率提升50%！**

喜欢此内容的人还喜欢

ICCV 2023｜南开程明明团队提出适用于SR任务的新颖注意力机制（已开源）
极市平台



ICCV23｜将隐式神经表征用于低光增强，北大张健团队提出NeRCo
极市平台



YOLOv5帮助母猪产仔？南京农业大学研发母猪产仔检测模型并部署到Jetson Nano开发板

极市平台