arXiv:2601.00577v1 [cs.LG] 2 Jan 2026

# 对抗样本并非生而平等

Jennifer Crawford Amol Khanna Fred Lu Amy R.Wagoner Scale AI CrowdStrike
BoozAllenHamiltion BoozAllenHamiltion

Stella Biderman
EleutherAI

Andre T. Nguyen
Booz Allen Hamiltion

Edward Raff
CrowdStrike

## 摘要

过去十年间，人们提出了许多理论来解释深度神经网络普遍易受对抗规避攻击影响的现象。其中，*Ilyas* 等人提出的非鲁棒特征理论*[1]*已广受认可。该理论表明，数据分布中脆弱但具有预测能力的特征可能会被攻击者直接利用。然而，该理论忽视了并未直接利用这些特征的对抗样本。在本研究中，我们认为这两类样本——一类利用脆弱但具有预测能力的特征，另一类则不利用此类特征——代表了两种对抗性弱点，在评估对抗鲁棒性时应加以区分。为此，我们提出一种基于集成的度量方法，用以衡量对抗扰动对非鲁棒特征的操纵程度，并利用该指标分析攻击者生成的对抗样本的构成。这一新视角也使我们能够重新审视多种现象，包括尖锐度感知最小化对对抗鲁棒性的影响，以及在鲁棒数据集上，对抗训练与标准训练之间观察到的鲁棒性差距。

## 1. 引言

人们早已注意到，深度神经网络对输入数据的操纵表现出一种奇特的脆弱性 [2–10]。在所谓的对抗规避攻击中，恶意用户可以向原本正常的输入添加经过精心设计却难以察觉的扰动，诱使模型自信地给出错误答案。这些对抗样本不仅构成显著的安全风险，也揭示了神经网络的运作方式与人类存在根本差异，尽管它们在许多任务上的表现能够与人类相媲美。

发现对抗样本后，人们进行了许多尝试，以理解并刻画这种现象。例如，先前的研究指出，输入数据的高维性 [11], 、过拟合 [12],以及使用局部线性激活函数 [13] 都可能是这种脆弱性的成因。相比之下， [1] 的开创性研究提出，模型用来解决任务的数据中那些脆弱但具有预测性的特征，在对抗易感性中起着重要作用。也就是说，对抗样本是神经网络以不同于人类的方式解决任务的自然结果：神经网络倾向于为任务学习最简单的输入特征 [14–16]。对抗者可以直接操纵这些非鲁棒特征，从模型的角度改变输入的含义，却不改变其在人类看来所表达的含义。

尽管这一理论直观地解释了对抗样本为何存在，但它并未直接探讨这样一种可能性：对抗样本并不涉及明确添加或移除非鲁棒特征。事实上，已有多项研究质疑，数据分布中的脆弱特征是否足以全面解释对抗脆弱性 [17, 18]。在本研究中，我们旨在区分这些不同的弱点，并考察它们在对抗攻击中的普遍程度。我们的贡献如下：

- 我们提出一种基于集成的度量指标，用于判断对抗样本是否直接利用了非鲁棒特征。我们使用该度量指标来区分对抗性错误，并将其定义为不依赖于添加或移除非鲁棒特征的对抗样本。
- 我们考察了非鲁棒特征在对抗攻击中的普遍程度，并发现，当扰动强度足以操纵输入的预测特征时，对抗训练模型的鲁棒性便会崩溃。然而，它们对对抗性错误仍表现出惊人的韧性。

•我们揭示了对抗性错误与

损失景观中的尖锐度之间的联系，并证明尖锐度感知最小化（SAM）能够针对对抗性错误提供保护。这与近期研究的观点形成对比；那些研究认为，SAM 在鲁棒性方面的作用源于它鼓励模型学习鲁棒特征。

- 使用我们的度量方法，我们重新审视了鲁棒数据集，并发现，与以往的假设相反，这些数据集仍然包含非鲁棒特征。这为对抗训练模型与基于鲁棒数据集训练的模型之间普遍存在的鲁棒性差距提供了更清晰的解释。

本文的结构如下。在第 2 节中，我们首先介绍非鲁棒特征理论的相关背景。随后，在第 3 节中，我们定义并论证用于衡量非鲁棒特征操纵程度的度量方法。我们使用该度量方法，研究在标准模型和对抗训练模型上针对 CIFAR10 和 SVHN 发起攻击所生成的对抗样本的构成（第 4 节）、使用 SAM 训练的模型（第 5 节），以及基于鲁棒数据集训练的模型（第 6 节）。

## 2. 预备知识

本节将介绍我们的符号表示，并定义关键术语。

**非鲁棒特征** 在本文中，我们关注多标签分类任务 $\mathcal{X} \to \mathcal{Y}$，其中 $\mathcal{X} \in \mathbb{R}^{m \times m}$和 $\mathcal{Y} = \{1, .., k\}$。我们采用 [1]，提出的特征定义，该定义将特征定义为形如$\phi$的函数：$\mathcal{X} \to \mathbb{R}^k$。因此，特征可以被视为量化输入样本 $\boldsymbol{x}$ 中特定任务相关信号存在程度的指标。在本文中，我们进一步明确，相关特征是数据的特征，即 数据特征，因为它们与数据分布 $\mathcal{D}$ 在统计上相关。每个特征都可以根据其有用性（即预测能力）以及对输入扰动的鲁棒性进行刻画。根据 [1]，，我们将分类中的 $\rho$-有用特征定义为具有以下性质的特征：

$$\mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}}[\mathbb{I}_y \cdot \phi(\boldsymbol{x})] \geq \rho. \tag{1}$$

其中，$\mathbb{I}_y$ 是对$y$真实标签进行编码的独热向量。此外，如果一个特征在一组有效扰动 $\Delta$下仍能保持$\gamma$有用性，那么称该特征具有 $\gamma$鲁棒有用性：

$$\mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}}[\inf_{\boldsymbol{\delta}\in\Delta}(\mathbb{I}_y \cdot \phi(\boldsymbol{x}+\boldsymbol{\delta}))] \geq \gamma. \tag{2}$$

[1] 通过创建非鲁棒数据集和鲁棒数据集（此处分别记为 $D_{NR}$ 和 $D_R$），证明了非鲁棒特征和鲁棒特征对对抗易感性的影响；这两种数据集都能泛化到原始数据分布$\mathcal{D}$。为了创建 $D_{NR}$ 和 $D_R$，[1] 将原始训练集中的每个输入与另一个“目标”输入（同样来自原始训练集）配对，并分别在标准模型或对抗训练模型上执行梯度下降，以便优化

$$\min_{\boldsymbol{x}'} \|g(\boldsymbol{x}_{target}) - g(\boldsymbol{x}')\|_2 \tag{3}$$

其中，$\boldsymbol{x}_{target}$ 是目标输入，$\boldsymbol{x}'$ 初始化为$\boldsymbol{x}'_0 = \boldsymbol{x}$。这里，$g$ 表示模型倒数第二层的隐藏表示。随后，将该优化的输出与目标输入的标签 $y_{target}$配对，从而在 $D_{NR}$ 或 $D_R$中形成一个新样本。作者表明，只有使用 $D_R$时，标准训练才能对对抗样本实现非平凡的鲁棒性。

为完整起见，我们在附录A中复现了这一实验。与 [1]，的结果一致，我们发现，在 $D_{NR}$ 上进行标准训练，尽管其输入看起来被错误标注，仍能取得良好的测试性能。此外，其鲁棒准确率与原始模型相当。相比之下，我们发现，在 $D_R$ 上进行标准训练，可以同时获得非平凡的测试准确率和鲁棒准确率。然而，与 [17, 18], 类似，我们注意到其鲁棒准确率有所下降（61.6% $\to$ 2.0%，对于 $\epsilon = 8/255$），相较于对抗训练而言。我们将在第6节探究这一鲁棒性差距。

此外，[1]构建了两个非鲁棒数据集，$D_{rand}$以及 $D_{det}$，其中包含通过投影梯度下降（PGD）生成的对抗样本 [19] 分别使用随机或确定性目标标签，$y_{target}$。为构建 $D_{det}$，我们遵循 [1] 并设定 $y_{target} = (y + 1) \mod C$，其中 $C$ 表示类别数。每个对抗样本都与其目标标签配对，因此在 $D_{rand}$中，只有非鲁棒特征能够预测输入标签；而在 $D_{det}$中，非鲁棒特征和鲁棒特征都可供学习。[1] 然而，$D_{det}$中的鲁棒特征与解决原始测试任务直接相悖。正如附录A中重现的结果所示，在 $D_{rand}$ 和$D_{det}$ 上训练，能够在测试分布上实现非平凡的泛化。这表明，即使像 $D_{det}$中那样，非鲁棒特征与错位的鲁棒特征配对，非鲁棒特征仍然能够被学习。

**对抗性错误** 由 [1] 提出的非鲁棒特征理论，并未涉及这样一种可能性：某些对抗样本并未直接利用这些特征 [11–13, 20]。本文将这类样本称为对抗性错误，因为它们并非由添加或移除数据特征的对抗扰动直接导致。相反，它们似乎是特定模型权重实例中的错误或盲点。图1展示了对抗性错误与利用非鲁棒特征构造的对抗样本之间的区别。本文旨在通过实证研究考察这两类对抗样本及其在对抗攻击中的出现频率。

[1]此处的特征鲁棒性是指生成 $D_{rand}$ 和 $D_{det}$的 PGD 攻击中所施加扰动的强度。

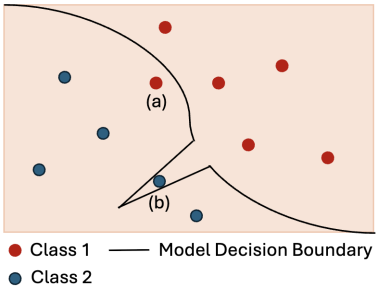



图 1. 深度神经网络决策边界示意图。该网络经过二元分类任务的训练，允许存在 (a) 利用非鲁棒特征的对抗样本，以及 (b) 对抗性错误。前者之所以存在，是因为分类器使用了数据分布中不符合人类认知的预测性特征；而后者则表现为特定模型实例中不规则的盲点。

## 3. 区分对抗性错误

### 3.1. 数据特征的属性

为了识别利用统计数据特征的对抗样本，我们首先描述理想化模型针对这类输入所应具备的两个理想属性。

**泛化属性** 使用能充分利用数据特征的对抗样本进行训练，并为其配上对抗目标标签，应能在原始测试集上取得非平凡的性能。这一现象最早在[1]中被观察到。我们注意到，无法保证完美泛化，因为可能存在分布偏移，而且对抗扰动不一定会使用足够多样的特征。

**不变性属性** 在训练期间采用随机数据增强时，模型会受到强烈引导，使用对所选增强方式保持不变的特征，$T$(即 $f(\boldsymbol{x}) = f(T(\boldsymbol{x}))$)。因此，我们预期利用数据特征的对抗样本在很大程度上对这些变换保持不变（即其不变性与自然测试样本相当）。

### 3.2. 数据特征的度量指标

我们希望有一种指标，能够衡量在创建对抗样本 $\boldsymbol{x}_{adv}$ 时，对其源输入 $\boldsymbol{x}_{src}$ 中数据特征的利用情况。为此，我们注意到，已有观察发现，对于许多神经网络架构和任务，同一模型采用不同随机初始化时，最终会收敛到非常相似的决策边界 [21–23]。这表明它们使用了相同的预测性输入特征，也说明不同模型实例之间对抗样本的可迁移性可能与对抗扰动对数据特征的操纵有关——因此，我们提出从由不同随机种子初始化的模型实例集成的角度来分析对抗样本。

**假设 3.2** 设 $f_{\theta_0}: \mathcal{X} \to \mathbb{R}^k$ 表示一个模型实例，其中 $\theta_0$ 是神经网络的参数，服从分布 $p(\theta)$。为近似 $\mathbb{E}_\theta[f_\theta]$，我们使用由 $N$ 个模型实例组成的集成 $\tilde{f}(\boldsymbol{x}) = \frac{1}{N}\sum_{i=0}^{N-1} f_i(\boldsymbol{x})$，其中 $f_i$ 表示集成中的第 $i$ 个成员。对于 $f_{\theta_0}$ 的给定源输入-对抗输入对 $(\boldsymbol{x}_{src}, \boldsymbol{x}_{adv})$，若 $f_{ens}(\boldsymbol{x}_{src})$ 与 $f_{ens}(\boldsymbol{x}_{adv})$ 存在显著差异，则表明对抗扰动显著操纵了数据特征。反之，这些输出一致则说明，攻击者在生成 $\boldsymbol{x}_{adv}$ 时主要利用了其他对抗弱点。

对于 $f_{\theta_0}$ 的每个源输入-对抗输入对，我们提出计算 $\tilde{}$ $\tilde{}$ 集成输出 $JS(f(\boldsymbol{x}_{adv}) \| f(\boldsymbol{x}_{src}))$ 之间的詹森–香农距离。对于改变数据特征的对抗样本，我们预期该值较大，因为这表明从 $f$ 的角度来看，输入的含义已 $\tilde{}$ 发生改变。然而，该值没有考虑模型对不同标签输入之间詹森–香农距离的尺度判断，因为模型的校准会影响这一距离。[2] 因此，我们提出使用以下归一化指标：

$$JS_\Delta(\boldsymbol{x}_{src}, y_{src}, \boldsymbol{x}_{adv}, y_{adv}) \equiv \frac{JS(\tilde{f}(\boldsymbol{x}_{src}) \| \tilde{f}(\boldsymbol{x}_{adv}))}{\rho_{y_{src}, y_{adv}}} \tag{4}$$

$\tilde{}$ $\tilde{}$ 其中 $\rho_{y_{src}, y_{adv}} = \mathbb{E}[JS(f(\boldsymbol{x}_1) \| f(\boldsymbol{x}_2))]$ 对于 $\{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2) \sim \mathcal{D} | (y_1, y_2) = (y_{src}, y_{adv})\}$。例如，$JS_\Delta = 1$ 的值表示，$f$ 认为对抗样本及其源输入在语义上存在显著差异——这类似于一对标签为 $(y_{src}, y_{adv})$ 的干净输入。反之，$JS_\Delta \sim 0$ 的值表示，对抗攻击基本未改变输入的数据特征，因此该对抗样本是受攻击的模型实例所特有的错误。因此，对于任意源输入-对抗输入对，我们使用公式 4 来检查数据特征在创建 $\boldsymbol{x}_{adv}$ 时的使用情况。

现在，我们通过确认对抗样本向 $f$ 的可迁移性 $\tilde{}$ 与数据特征的泛化性和不变性相关，来证明 $JS_\Delta$。

**$JS_\Delta$ 以及泛化属性** 为了建立 $JS_\Delta$ 与泛化属性之间的联系，我们注意到 [20] 进行了一个实验，在该实验中，攻击者明确创建了

[2] 例如，对于预测结果更有信心的模型，会自然地为标签不同的两个输入的模型输出给出更大的 Jensen-Shannon 距离。

不会迁移到由相同方式训练的独立模型集成中的对抗样本。当使用这些高度不可迁移样本构建类似于 $D_{det}$ 的数据集时，他们观察到，在这个新数据集上进行训练并不会对原始测试分布实现泛化， $D$。我们在附录 B 中复现了这一实验，并确认~对抗样本向 $f$ 的可迁移性与泛化属性相关。

**$JS_\Delta$ 以及不变性属性** 最后，我们考察 $JS_\Delta$ 与不变性属性之间的关系。在此实验中，我们使用针对 $\ell_\infty$ 约束扰动的定向PGD-100攻击，并结合 $\epsilon = 16/255$，生成一组对抗测试样本 $X_{adv}$。我们随机选择目标标签 $Y_{target}$，使其不同于源输入的标签 $Y_{src}$。此外，我们还使用 [20] 提出的方法（详见附录 B）生成了一组高度不可迁移的对抗测试样本 $X^{\perp}_{adv}$。在表 1 中，我们考察 $f_{\theta_0}(\boldsymbol{x})$ 对于用于训练网络的随机数据增强 $T$ 的不变性，并将 $\boldsymbol{x} \sim X_{adv}$ 与 $\boldsymbol{x} \sim X^{\perp}_{adv}$ 进行对比。在此实验中，我们在训练期间采用标准的增强策略，包括随机水平翻转、随机裁剪、小幅随机旋转和随机颜色抖动。请注意，在生成 $X_{adv}$ 和 $X^{\perp}_{adv}$ 时，这些随机增强并未作为防御手段应用。我们推进了 [20] 的工作，确定由原始 PGD 生成的样本在很大程度上对 $T$ 具有不变性，而那些对 $f$ 高度不可迁移（因此具有较小 $JS_\Delta$）的样本中，~绝大多数不具备这种不变性。事实上，当我们将训练增强应用于这些样本时，发现它们通常不再具有对抗性；在大多数（$\sim 86\%$）情况下，受攻击模型能够正确输出其真实标签。[3] 这表明，这些样本的对抗成功很大一部分源于不同于直接操纵数据特征的机制。

**备注** 我们注意到，对抗性错误与利用数据特征的错误很可能存在于一个连续谱上。即使使用公式 6直接针对对抗性错误，非鲁棒特征泄漏仍可能发生在这些样本的生成过程中，而且在实践中可能难以彻底消除。此外，某些架构和数据集可能不满足假设 3.2，尤其是在不同训练轮次间性能差异显著的情况下。

[3]这一结果表明，可以采用另一种指标来帮助区分对抗性错误与利用非鲁棒特征的错误：对训练期间所应用随机数据增强的不变性。然而，这不仅要求在训练期间使用数据增强，而且高度依赖于其强度。

表 1. 对抗测试样本对于训练期间使用的随机数据增强的不变性。我们注意到，原始 PGD 生成的大多数样本 $T$（$X_{adv}$）表现出较高的不变性，在 $T$ 下有 $\sim 95\%$ 的时间保持其目标标签。对于高度不可迁移的样本 $X^{\perp}_{adv}$，情况恰恰相反：其中有 $\sim 86\%$ 的样本在 $T$ 下恢复为其源图像的标签。

| 输入 | 准确率 $Y = Y_{target}$ | 准确率 $Y = Y_{src}$ |
|---|---|---|
| $X_{adv}$ | 100.0% | 0.0% |
| $T(X_{adv})$ | 95.1% | 4.5% |
| $X^{\perp}_{adv}$ | 98.1% | 2.4% |
| $T(X^{\perp}_{adv})$ | 3.9% | 85.7% |

## 4. 对抗样本的构成

**实验设置** 与 [1], 类似，我们使用 ResNet50 和 ResNet18 架构 [24] 分别在 CIFAR10 [25] 和 SVHN [26], 上进行训练。[4] 模型的训练详情~见附录C。为了估计 $f$ 每个被攻击模型的 $f_0$，我们训练了四个相同的模型实例，并使用不同的随机种子进行初始化。需要指出的是，在~每次实验中， $f$ 的各个模型实例都采用与 $f_0$相同的优化过程进行训练。对于攻击，我们使用 PGD [19] 进行 100 步 $\ell_\infty$约束扰动，采用定向交叉熵攻击损失，步长设为 $\epsilon/10$。受篇幅限制，SVHN 的结果见附录 D。请注意，在所有考察对抗样本组成的实验中，我们仅考虑会导致被攻击模型标签改变的成功的 对抗样本。

首先，我们检查使用随机梯度下降和交叉熵训练的非鲁棒模型所产生的对抗样本的组成。在图2 和 8中，我们记录了 CIFAR10 和 SVHN 中每个对抗测试样本的直方图 $JS_\Delta$。我们分别在表2 和 10中对这些直方图进行定量分析，计算具有 $JS_\Delta < \beta$ 的对抗样本占 $\beta = [0.01, 0.05, 0.10]$的百分比。我们发现，对于较小幅度的攻击，相当大比例的对抗样本是对~抗性错误，且很难迁移到 $f$。当攻击扰动达到一定阈值后，攻击者便转而主要利用非鲁棒特征。这表明：1）产生对抗性错误所需的扰动幅度通常小于操纵非鲁棒特征所需的幅度；2）基于梯度的攻击倾向于选择操纵数据特征的对抗样本。后一观点不难理解，因为我们预计，数据的预测特征是训练过程中塑造

[4]卷积架构被发现在 [21] 不同随机种子之间具有较高的决策边界相似性，这为应用假设3.2提供了支持。

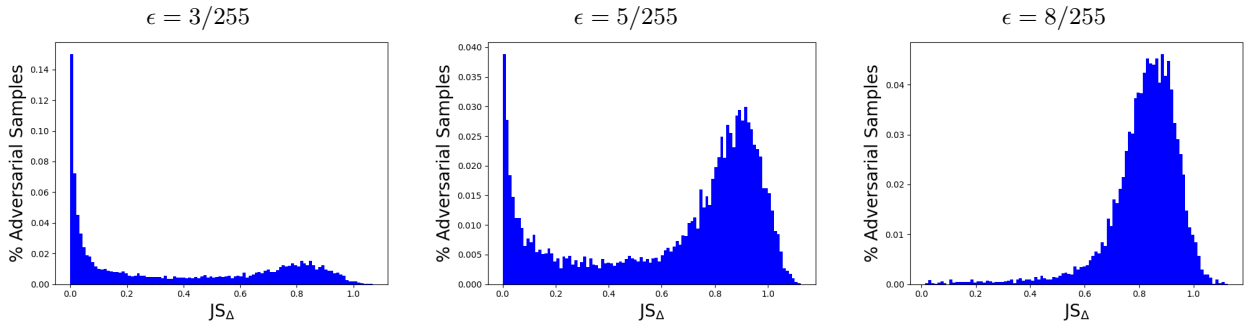



图 2. 在 CIFAR10 上训练的非鲁棒 ResNet50 模型的对抗样本 $JS_\Delta$ 归一化直方图。所有样本均通过定向 PGD-100 攻击生成。我们观察到，在扰动幅度较小时，很大比例的对抗样本可被识别为对抗性错误。扰动幅度较大时，攻击者能够在大多数对抗样本中操纵非鲁棒特征。

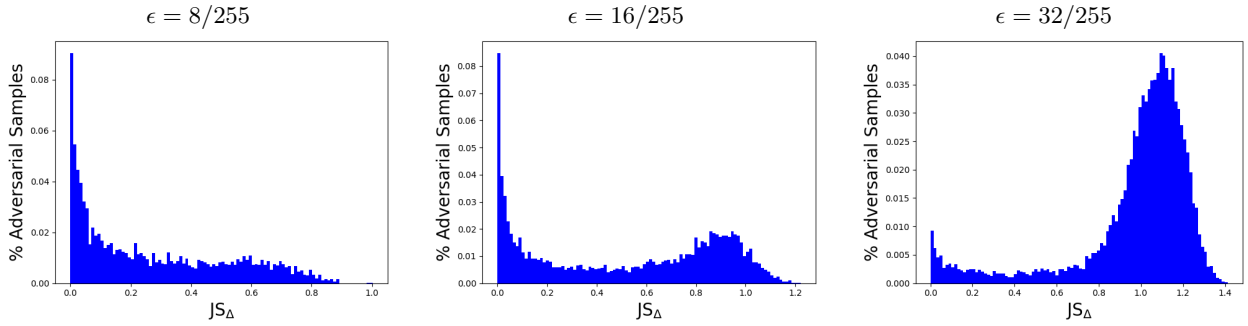



图 3. 在 CIFAR10 上经过对抗训练的鲁棒 ResNet50 模型的成功对抗样本 $JS_\Delta$ 归一化直方图。所有样本均通过定向 PGD-100 攻击生成。我们观察到，要生成对抗性错误以及利用非鲁棒特征的对抗样本，都需要更大的攻击扰动。

损失景观的主要信号，因此这些特征会对基于梯度的攻击产生显著影响。

我们在图 3 和 9 以及表 3和 11 中，对 CIFAR10 和 SVHN 上经过对抗训练的模型重复了这些实验。对于这两个数据集，我们观察到相似的模式：当扰动幅度超过某个阈值后，非鲁棒特征便成为对抗攻击更倾向采用的方式。我们发现，要充分操纵数据特征所需的扰动幅度明显更大（即 $32/255$ 和 $16/255$ 分别对应 CIFAR10 和 SVHN），同时还注意到，在扰动幅度较小时，模型整体上能够抵御对抗性错误。因此，尽管对抗训练会促使模型学习更鲁棒的特征，但似乎也能使模型免受对抗性错误的影响。

## 5. 与尖锐度感知最小化（SAM）的联系

在第 3节中，我们确认了对抗样本的可迁移性与其对随机数据增强的不变性相关。因此，有理由推测

表2。对于在CIFAR10上训练的非鲁棒ResNet50模型，具有 $JS_\Delta < \beta$ 且针对 $\beta = [0.01, 0.05, 0.10]$ 的对抗样本所占百分比。所有样本均通过定向PGD-100攻击生成。我们发现，弱攻击倾向于利用对抗性错误，而当攻击扰动超过某一阈值后，相当一部分对抗样本会使用非鲁棒特征。

| 攻击强度（$\epsilon$） | 鲁棒准确率 | 带有 $JS_\Delta$ <的样本百分比 | | |
|---|---|---|---|---|
| | | 0.01 | 0.05 | 0.10 |
| 1 / 255 | 66.5% | 23.1% | 51.4% | 64.3% |
| 3 / 255 | 3.7% | 14.3% | 32.3% | 39.6% |
| 5 / 255 | 0.7% | 0.3% | 10.4% | 14.5% |
| 8 / 255 | 0.7% | 0.0% | 0.1% | 0.2% |

对抗性错误对随机扰动的稳定性明显低于干净输入或利用数据特征的对抗样本。对抗性错误的这一特性意味着它们可能位于曲率较低的区域，而这一特征此前也被单独发现与对抗可迁移性有关 [27, 28]。为验证我们的假设，我们检查

表3. 在 CIFAR10 上经过对抗训练的鲁棒ResNet50模型中，具有 $JS_\Delta < \beta$ for $\beta = [0.01, 0.05, 0.10]$ for 的对抗样本所占百分比。所有样本均通过定向PGD-100攻击生成。我们观察到，要生成对抗性错误以及利用非鲁棒特征的对抗样本，需要更大的攻击扰动。

| 攻击强度 ($\epsilon$) | 鲁棒准确率 | % 具有 $JS_\Delta <$ 的样本 | | |
|---|---|---|---|---|
| | | 0.01 | 0.05 | 0.10 |
| 5 / 255 | 78.8% | 9.0% | 29.8% | 42.8% |
| 8 / 255 | 61.6% | 9.0% | 26.1% | 36.6% |
| 16 / 255 | 14.2% | 7.5% | 18.1% | 24.5% |
| 32 / 255 | 1.4% | 0.7% | 2.3% | 3.5% |

对抗样本周围的损失景观，其 $JS_\Delta$ 和自然测试样本。具体来说，我们绘制 $L_{CE}(\boldsymbol{x} + \alpha_1 \boldsymbol{\epsilon_1} + \alpha_2 \boldsymbol{\epsilon_2},\ y)$，其中 $\alpha_1$, $\alpha_2 \in$ [-1,1] 表示两个随机方向 $\boldsymbol{\epsilon_1}$, $\boldsymbol{\epsilon_2}$。对于自然测试样本，我们考察其相对于真实标签的损失；对于被攻击者修改的样本，则使用对抗目标标签。在图4中，我们展示了损失景观的曲面图示例，并发现对抗性错误对应的极小值明显比与数据特征相关的极小值更尖锐。

对抗性错误存在于高曲率区域这一发现，可能会让人推测，这些错误的尖锐性与其产生机制直接相关。最近，[29]比较了尖锐度感知最小化（SAM）与对抗训练。SAM 可以表述为以下极小极大优化问题：

$$\min_{\boldsymbol{w}} \mathbb{E}_{(\boldsymbol{x},y)\sim D}[\max_{||\boldsymbol{\epsilon}||<\rho} L(\boldsymbol{w}+\boldsymbol{\epsilon}, \boldsymbol{x}, y)]. \quad (5)$$

其中，扰动被添加到模型权重 $\boldsymbol{w}$ 上，而非像对抗训练那样添加到输入 $\boldsymbol{x}$ 上。这里，$\rho$ 是一个用于限制扰动幅度的超参数。[29] 的作者注意到 SAM 与对抗训练的表述相似，并推测最小化曲率可能会促使模型学习更鲁棒的特征。在这项研究中，研究者发现，使用 SAM 训练的模型在攻击扰动较小时具有对抗鲁棒性，同时干净准确率并未下降。这被用作实证依据，支持这样一种假设：SAM 的作用类似于对抗训练，并抑制非鲁棒特征学习。我们使用针对对抗样本的可迁移性度量来检验这一主张。

在图 5 和 10 以及表 4 和 12中，我们探究了 SAM 对 CI-FAR10 和 SVHN 的对抗样本构成的影响。我们观察到 $JS_\Delta$ 的分布明显向右偏移，而且高度不可迁移的对抗样本所占比例大幅下降。因此，SAM 似乎能够防御对抗性漏洞

。与 [29]，类似，我们还注意到，在攻击扰动较小时，模型具有不可忽视的鲁棒性。

为了衡量 SAM 对所学数据特征鲁棒性的影响，我们希望了解在 SAM 训练的模型中，诱发数据特征所需的最小扰动$^{\epsilon}f$，并将其与标准训练模型和对抗训练模型进行比较。为此，我们重新设计了快速最小范数（FMN）攻击 [30]，使其梯度信息来自˜集成 $f$，并在目标模型上评估攻击是否成功 $f_0$。如表5所示，我们发现，对于使用 SAM 训练的模型，诱发数据特征所需的最小扰动与标准模型的相当，而非与其鲁棒对应模型相当。因此，与 [29], 的假设相反，这种正则化似乎并未促使深度神经网络使用更多鲁棒特征。相反，它不对称地针对对抗性错误背后的弱点，同时基本不改变特征的鲁棒性。这一发现解释了 [29], 指出的 SAM 与对抗训练之间的病理性差异，例如它们对干净准确率的影响。它也表明，模型决策边界的尖锐程度对于对抗性错误的存在起着根本性作用。

表4. 对抗样本的百分比，其 $JS_\Delta < \beta$ 对于 $\beta = [0.01, 0.05, 0.10]$ 对于通过 SAM（$\rho = 0.3$）在 CIFAR10 上训练的 ResNet50 模型。所有样本均通过定向PGD-100攻击生成。与通过随机梯度下降训练的对应模型（表2）相比，我们观察到对抗性错误显著减少。

| 攻击强度 ($\epsilon$) | 鲁棒准确率 | % 个样本具有 $JS_\Delta <$ | | |
|---|---|---|---|---|
| | | 0.01 | 0.05 | 0.10 |
| 1 / 255 | 79.1% | 0.5% | 9.2% | 27.8% |
| 3 / 255 | 7.5% | 0.5% | 8.0% | 16.2% |
| 5 / 255 | 0.7% | 0.1% | 1.8% | 3.8% |
| 8 / 255 | 0.6% | 0.0% | 0.0% | 0.2% |

表5. 对于在 CIFAR10（左）和 SVHN（右）上训练的模型，充分操纵数据特征所需的对抗扰动平均最小幅度，$\|\epsilon_f\|_p$ 其中 $\|\cdot\|_p$ 表示 $L_p$ 范数。我们将对抗训练模型表示为 $M_{adv}$，将使用 SAM 训练的模型表示为 $M_\rho$，其中 $\rho$ 表示 SAM 的强度。我们的结果表明，SAM 并未提高模型所学习的数据特征的鲁棒性。

| | CIFAR10 ($\alpha = 0.3$) | | SVHN ($\alpha = 0.1$) | |
|---|---|---|---|---|
| 模型 | $\|\epsilon_f\|_2$ | $\|\epsilon_f\|_\infty$ | $\|\epsilon_f\|_2$ | $\|\epsilon_f\|_\infty$ |
| $M_{\rho=0}$ | $0.38 \pm 0.21$ | $0.013 \pm 0.006$ | $0.65 \pm 0.44$ | $0.021 \pm 0.013$ |
| $M_{\rho=\alpha}$ | $0.38 \pm 0.19$ | $0.013 \pm 0.006$ | $0.59 \pm 0.41$ | $0.019 \pm 0.011$ |
| $M_{adv}$ | $1.61 \pm 0.81$ | $0.051 \pm 0.024$ | $1.03 \pm 0.64$ | $0.045 \pm 0.024$ |

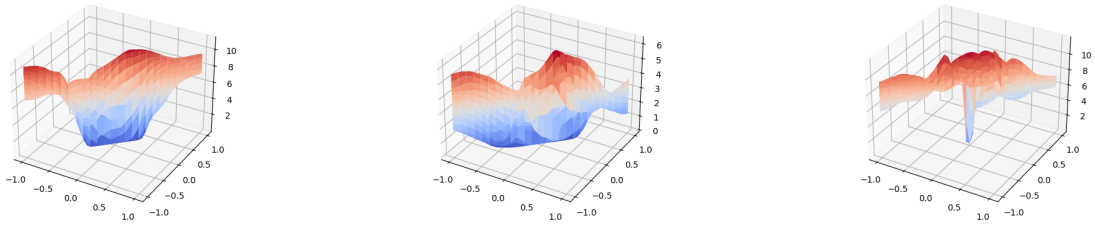


图4. CIFAR10 上训练的 ResNet50 模型中，自然测试样本（左）以及具有较大（中）和较小 $JS_\Delta$ 的对抗测试样本周围的损失景观。我们发现，对抗性错误（右）的损失呈现出尖锐的极小值，而操纵数据特征的对抗样本（中）的损失曲率则与自然样本相近。

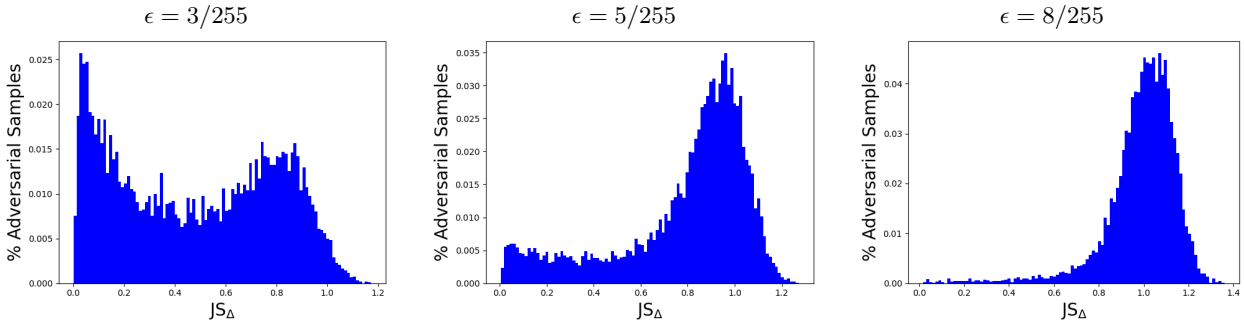



图5. $JS_\Delta$ 使用 SAM（$\rho = 0.3$）训练的 ResNet50 模型在 CIFAR10 上生成的对抗样本的归一化直方图。所有样本均通过定向 PGD-100攻击生成。与通过 SGD 训练的对应模型相比，对抗性错误显著减少（图2）。

## 6. 重新审视鲁棒数据集

既然已经确认非鲁棒特征和对抗性错误都会导致对抗脆弱性，我们或许会想知道，后者的存在能否解释对抗训练模型与在鲁棒数据集上训练的模型之间观察到的鲁棒性差距。也就是说，在鲁棒数据集上训练是否无法防范对抗性错误。事实上，我们注意到，[17] 利用鲁棒性差距提出假设：仅有鲁棒特征不足以实现真正的模型鲁棒性。对抗性错误或许可以解释这一现象。

在图7 和 11中，我们记录了针对在鲁棒 CIFAR10 和 SVHN 数据集上训练的模型发起攻击后得到的 $JS_\Delta$ 直方图 $M_R$。这些直方图的定量分析见表6 和 13。令人惊讶的是，我们观察到在存在鲁棒性差距的扰动幅度下，针对 $M_R$的大多数成功的对抗样本似乎利用的是非鲁棒特征，而非表现为对抗性错误（例如 $\epsilon = 8/255$）。因此，相较于其对抗训练版本， $M_R$ 的特征鲁棒性似乎有所下降。这表明，这些模型鲁棒性的下降并非由于缺乏对对抗性错误的防护，而是非鲁棒特征重新出现的结果。为确认这一可能性，我们通过对 $D_{R'}$的倒数第二层表示应用方程3 中描述的优化过程，创建了一个二阶鲁棒数据集 $M_R$。在图6中，我们直观检查其中的图像，并将其与 $D_{NR}$ 和 $D_R$中的图像进行比较。我们发现， $D_{R'}$ 中的图像明显更像 $D_{NR}$ 中的图像，而非 $D_R$。这表明， $M_R$ 学习到的特征与标准训练模型的特征更为接近，而非对抗训练模型的特征——也就是说， $M_R$ 似乎使用了不连贯的非鲁棒特征。

我们的发现得到了 [31, 32]，的支持；该研究表明，对抗训练模型的特征与非鲁棒模型的特征高度纠缠，而且鲁棒特征可以通过许多有用的非鲁棒特征构建。然而，据我们所知，其对于在鲁棒数据集上训练的启示尚未得到重视。由于鲁棒特征包含有用的非鲁棒特征，因此通常无法通过在鲁棒数据集上进行标准训练来强制使用这些特征。这是因为特征并非由任务（即数据集）唯一确定，而是由输入的处理方式决定的




图 6。（CIFAR10）来自 $D_{NR}$（顶行）、 $D_R$（中间行）以及由 $M_R$ 训练的二阶鲁棒数据集$D'_R$（底行）的输入样本。我们观察到，$D'_R$ 中的图像与其源图像相似，如 $D_{NR}$中所示；而只有 $D_R$ 中的输入在视觉上似乎与目标标签相关。

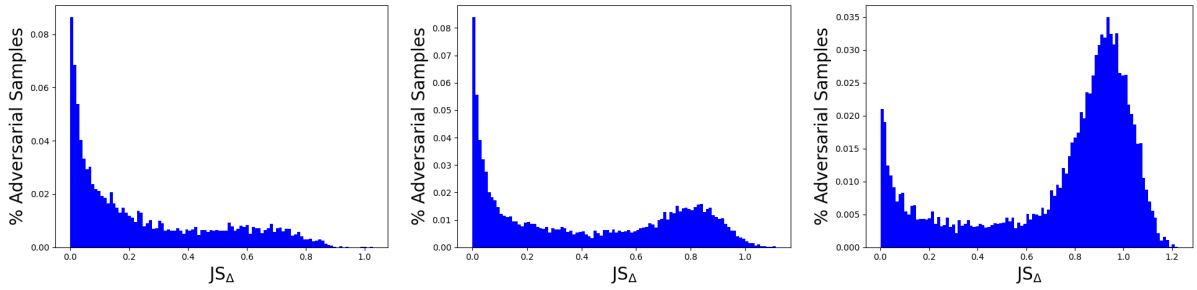



图7。ResNet50 模型在鲁棒CIFAR10数据集上训练后，其对抗样本的 $JS_\Delta$ 归一化直方图（从左至右， $\epsilon = 3/255,\ 5/255,\ 8/255$）。所有样本均通过定向PGD-100攻击生成。我们观察到，攻击者能够以与标准训练模型相当的幅度操纵数据特征，这表明非鲁棒特征再次出现。

表6. 在鲁棒CIFAR10数据集上训练的 ResNet50 模型中，具有 $JS_\Delta < \beta$for $\beta = [0.01, 0.05, 0.10]$ for 的对抗样本百分比。所有样本均通过定向 PGD-100 攻击生成。我们观察到，攻击者能够以与标准训练模型相当的幅度操纵数据特征，这表明非鲁棒特征再次出现。

| 攻击强度 ($\epsilon$) | 鲁棒准确率 | 具有 $JS_\Delta <$的样本百分比 | | |
|---|---|---|---|---|
| | | 0.01 | 0.05 | 0.10 |
| 3 / 255 | 42.2% | 8.3% | 27.8% | 40.3% |
| 5 / 255 | 9.3% | 7.7% | 22.6% | 30.9% |
| 8 / 255 | 2.0% | 1.6% | 6.4% | 9.6% |

以解决任务。因此，我们主张，不应将鲁棒性差距视为使用鲁棒特征时仍未得到防御的对抗性弱点的证据，正如 [17]所提出的那样。相反，这一差距是试图将它们直接提炼到数据集中的简单做法所导致的。这一洞见也自然地解释了为何此前弥合这一差距的尝试大多未能成功[18]。

## 7. 结论

本文强调，对抗样本至少可以通过两种不同的弱点生成：模型对非鲁棒特征的关注，以及与损失景观中的尖锐程度相对应的另一种独立脆弱性。我们提出了一种度量方法，根据对抗样本向具有不同初始化的其他模型实例的可迁移性，衡量对抗样本对数据特征的利用程度。利用这一度量，我们首次分析了非鲁棒模型和鲁棒模型的对抗样本构成，以及它们对数据特征的依赖程度。此外，我们通过阐明文献中此前提及的两种现象——SAM 对鲁棒性的影响，以及鲁棒数据集对鲁棒学习的不充分性——展示了这一视角的实用价值。我们认为，我们的贡献推动了对神经网络普遍存在却又错综复杂的对抗攻击脆弱性的理解。

## 参考文献

[1] Andrew Ilyas、Shibani Santurkar、Dimitris Tsipras、Bran- don Tran Logan Engstrom 和 Aleksander Madry。对抗

样本并非漏洞，而是特征。载于神经信息处理系统大会，2019年。1, 2, 3, 4, 11[2] Christian Szegedy、Wojciech Zaremba、Ilya Sutskever、Joan Bruna、Dumitru Erhan、Ian Goodfellow 和 Rob Fergus。神经网络的奇特属性，2013年。1[3] Edward Raff、Michel Benaroch 和 Andrew L. Farris。管理对抗攻击风险，无需鲁棒机器学习。《AAAI/ACM人工智能、伦理与社会会议论文集》，8(3):2094–2106，2025年10月。ISSN 3065-8365。doi: 10.1609/aies.v8i3.36698。URL https://ojs.aaai.org/index.php/AIES/ article/view/36698.[4] Philip Doldo、Derek Everett、Amol Khanna、Andre T. Nguyen 和 Edward Raff。别再兜圈子！在投影梯度下降中尽早退出。第6373–6382页，2025年。URL https://openaccess.thecvf. com / content / CVPR2025 / html / Doldo _Stop_Walking_in_Circles_Bailing_Out_Early_in_Projected_Gradient_IEEE计算机视觉与模式识别会议_2025_paper.html.[5] Edward Raff、Karen Kukla、Michel Benaroch 和 Joseph Comprix。通过最大违反多目标攻击，对财务报告发起对抗机器学习攻击。载于《2025年应用机器学习与信息安全会议论文集》，第1–27页。机器学习研究论文集，2025年12月。URL https : / / proceedings.mlr.press/v299/raff25a.html。ISSN: 2640-3498.[6] Luke E. Richards、Andr´e Nguyen、Ryan Capps、Steven Forsythe、Cynthia Matuszek 和 Edward Raff。利用未知数据与类别重叠的对抗迁移攻击。载于第14届ACM人工智能与安全研讨会（AISec '21）论文集。计算机协会，2021年。doi: 10.1145/3474369.3486862。URL http://arxiv.org/abs/2109.11125。arXiv: 2109.11125.[7] Arash Rahnama、Andre T. Nguyen 和 Edward Raff。基于李雅普诺夫理论的对抗攻击防御型深度神经网络设计。《IEEE/CVF计算机视觉与模式识别会议》（IEEE计算机视觉与模式识别会议），第8178–8187页，2020年。URL http://arxiv.org/ abs/1911.04636。arXiv: 1911.04636.[8] Edward Raff、Jared Sylvester、Steven Forsyth 和 Mark McLean。通过随机变换提升对抗鲁棒防御。载于《IEEE计算机视觉与模式识别会议》，第6528–6537页，加利福尼亚州长滩，2019年。URL http:// openaccess.thecvf.com/content_IEEE计算机视觉与模式识别会议_2019/ html/Raff_Barrage_of_Random_Transforms_for _ _ _ 2019_

Adversarially Robust Defense CVPR paper.html.[9] Andre T Nguyen 和 Edward Raff。对抗攻击、回归与数值稳定性正则化。载于AAAI-19可靠与安全机器学习系统工程研讨会，2019年。[10] William Fleshman、Edward Raff、Jared Sylvester、Steven Forsyth 和 Mark McLean。抵御对抗攻击的非负网络。AAAI-2019人工智能与网络安全研讨会，2019年。URL http: //arxiv.org/abs/1806.06108。arXiv：1806.06108。1[11] Ali Shafahi、Ronny Huang、Christoph Studer、Soheil Feizi 和 Tom Goldstein。对抗样本是不可避免的吗？载于国际学习表征会议，2019年。1、2[12] Thomas Tanay 和 Lewis Griffin。从边界倾斜视角看对抗样本现象，2016年。1[13] Ian J. Goodfellow、Jonathon Shlens 和 Christian Szegedy。解释并利用对抗样本。载于国际学习表征会议，2015年。1、2[14] Harshay Shah、Kaustav Tamuly、Aditi Raghunathan、Prateek Jain 和 Praneeth Netrapalli。神经网络中简单性偏差的陷阱。载于神经信息处理系统大会，2020年。1[15] Mohammad Pezeshki、S´ekou-Oumar Kaba、Yoshua Bengio、Aaron Courville、Doina Precup 和 Guillaume Lajoie。梯度饥饿：神经网络的一种学习倾向。载于神经信息处理系统大会，2021年。[16] Robert Geirhos、J¨orn-Henrik Jacobsen、Claudio Michaelis、Wieland Brendel、Richard Zemel、Matthias Bethge 和 Fe- lix A. Wichmann。深度神经网络中的捷径学习。《自然·机器智能》，2020年。1[17] Ang Li、Yifei Wang、Yiwen Guo 和 Yisen Wang。对抗样本并非真正的特征。载于神经信息处理系统大会，2023年。1、2、7、8、11[18] Nikolaos Tsilivis、Jingtong Su 和 Julia Kempe。我们能否仅凭数据实现鲁棒性？，2023年。1、2、8、11[19] Aleksander Madry、Aleksandar Makelov、Ludwig Schmidt、Dimitris Tsipras 和 Adrian Vladu。迈向能够抵御对抗攻击的深度学习模型。载于国际学习表征会议，2018年。2、4[20] Preetum Nakkiran。关于《对抗样本不是漏洞，而是特征》的讨论：对抗样本也只是漏洞，2019年。2、3、4、11[21] Gowthami Somepalli、Liam Fowl、Arpit Bansal、Ping Yeh-Chiang、Yehuda Dar、Richard Baraniuk、Micah Goldblum 和 Tom Goldstein。神经网络能否两次学到同一个模型？从决策边界视角研究可复现性与双重下降。载于IEEE计算机视觉与模式识别会议，2022年。3、4[22] Keller Jordan。神经网络训练相对于测试集和分布的方差。载于国际学习表征会议，2024年。[23] Oskar van der Wal、Pietro Lesci、Max Muller-Eberstein、Naomi Saphra、Hailey Schoelkopf、Willem Zuidema 和 Stella Biderman。Polypythias：五十次语言模型预训练运行中的稳定性与离群值。载于国际学习表征会议，2025年。3[24] Kaiming He、Xiangyu Zhang、Shaoqing Ren 和 Jian Sun。用于图像识别的深度残差学习。载于IEEE计算机视觉与模式识别会议，2016年。4[25] Alex Krizhevsky 和 Geoffrey Hinton。从微型图像中学习多层特征。技术报告，多伦多大学，2009年。4[26] Yuval Netzer、Tao Wang、Adam Coates、Alessandro Bissacco、Bo Wu 和 Andrew Y. Ng。通过无监督特征学习识别自然图像中的数字。载于神经信息处理系统大会深度学习与无监督特征学习研讨会，2011年。4[27] Zeyu Qin、Yanbo Fan、Yi Liu、Li Shen、Yong Zhang、Jue Wang 和 Baoyuan Wu。提升对抗攻击的可迁移性，采用对抗

攻击中的反向对抗扰动。载于神经信息处理系统大会，2022年。5[28] Zhijin Ge、Hongying Liu、Xiaosen Wang、Fanhua Shang 和 Yuanyuan Liu。通过实现平坦局部极大值提升对抗可迁移性。载于神经信息处理系统大会，2023年。5[29] Yihao Zhang、Hangzhou He、Huanran Chen、Jingyu Zhu、Yifei Wang 和 Zeming Wei。尖锐度感知最小化与对抗训练之间的对偶关系。载于国际机器学习大会，2024年。6[30] Maura Pintor、Fabio Roli、Wieland Brendel 和 Battista Biggio。通过自适应范数约束实现快速的最小范数对抗攻击。载于神经信息处理系统大会，2021年。6[31] Haydn T. Jones、Jacob M. Springer、Garrett T. Kenyon 和 Juston S. Moore。训练一个就等于训练了全部：跨架构相似性随鲁棒性增强而提高。载于人工智能不确定性会议，2022年。7[32] Jacob M. Springer、Melanie Mitchell 和 Garrett T. Kenyon。对抗扰动并没有那么奇怪：神经网络分类器中鲁棒与非鲁棒特征的纠缠，2021年。7

## A. 复现 Ilyas 等人的实验 [1]

本节给出我们在 CIFAR10 上复现 [1] 中实验的结果。在表 7 中，我们确认，在原始测试分布上，使用 $D_{NR}$ 和 $D_R$ 训练的模型都能取得非平凡的性能。我们的符号约定如下： $M$ 表示经常规训练的模型（即通过交叉熵损失进行随机梯度下降）， $M_{adv}$ 表示其对抗训练版本，而$M_{NR}$ ($M_R$) 表示在 $D_{NR}$ ($D_R$上经常规训练的模型。与 $M$相比，只有使用鲁棒数据集进行标准训练，才能取得非平凡的准确率。不过，与 [17, 18], 类似，我们也注意到 $M_R$ 和$M_{adv}$ 之间存在鲁棒性差距（例如，61.6% $\rightarrow$ 2.0% 在 $\epsilon = 8/255$）。

在表 8中，我们使用非鲁棒数据集 [1] 和 $D_{rand}$ 复现了 $D_{det}$的实验。我们证实了他们的发现：即使存在错位的鲁棒特征（如 $D_{det}$），仅凭非鲁棒特征也足以进行学习。

| 模型 | 干净准确率 | 鲁棒准确率 | | |
|---|---|---|---|---|
| | | $\epsilon = 3/255$ | $\epsilon = 8/255$ | $\epsilon = 16/255$ |
| $M$ | 94.1% | 3.7% | 0.7% | 0.7% |
| $M_{adv}$ | 87.8% | 83.2% | 61.6% | 14.2% |
| $M_{NR}$ | 81.5% | 3.4% | 2.0% | 2.0% |
| $M_R$ | 84.5% | 42.2% | 2.0% | 1.7% |

表7. 在 CIFAR10 上复现 [1] 中的实验：在非鲁棒（鲁棒）数据集上进行常规训练的模型表现出平凡（非平凡）的对抗鲁棒性。鲁棒准确率采用目标 $l_\infty$ 约束的 PGD-100 攻击进行评估。

| 训练集 | $D$上的干净准确率 |
|---|---|
| $D_{rand}$ | 62.0% |
| $D_{det}$ | 33.4% |

表 8. 重现了 [1] 中的实验，该实验表明，在非鲁棒 CIFAR10 数据集 $D_{rand}$ 和 $D_{det}$ 上训练的模型，在 $D$ 上具有不可忽略的测试准确率。这表明，即使存在未对齐的鲁棒特征，模型也能够学习非鲁棒特征，如 $D_{det}$ 中所示。

## B. JS$\Delta$ 与泛化属性之间的经验关联

为了直接创建对抗性错误， [20]提出了一种集成调整的定向攻击损失：

$$\nabla_{\boldsymbol{x}} L_{CE}(f_{\theta_0}(\boldsymbol{x}), y_{target}) \rightarrow \nabla_{\boldsymbol{x}} L_{CE}(f_{\theta_0}(\boldsymbol{x}), y_{target}) + \nabla_{\boldsymbol{x}} L_{CE}(\tilde{f}(\boldsymbol{x}), y_{src}). \tag{6}$$

请注意，公式 6 中的第二项鼓励 $JS_\Delta$取较小值，因为它旨在最小化 $JS(f(\boldsymbol{x}_{src})||f(\boldsymbol{x}))$。我们使用这种集成调整损失创建了非鲁棒 CIFAR10 数据集的第二个版本， $D_{det}$，并将其记为 $D^{\perp}_{det}$。为求完整，我们还构建了 CIFAR10 上 $D_{rand}$ 的类似版本，记为$D^{\perp}_{rand}$。表 9 记录了在这些数据集上进行训练后得到的干净测试准确率；我们发现，其测试准确率明显低于 $D_{rand}$ 和 $D_{det}$，这与表 8 中的观察结果一致。因此，我们确认了 [20] 的发现：对抗样本向 $f$ （因而也向 $JS_\Delta$）的可迁移性与泛化属性相关。

| 训练集 | D 上的测试准确率 |
|---|---|
| $D^{\perp}_{rand}$ | 11.3% |
| $D^{\perp}_{det}$ | 7.8% |

表 9。（CIFAR10）使用由公式 $D_{det}$和 $D_{rand}$构造的集成调整版本训练的模型的干净测试准确率 6(此处记为 $D^{\perp}_{det}$和 $D^{\perp}_{rand}$）。我们证实了 [20]的发现：在这些由高度不可迁移的对抗样本组成的数据集上进行训练，并不能使模型泛化到 $D$。这与表 8中使用 $D_{det}$和 $D_{rand}$进行训练所得的结果形成对比。

## C. 训练详情

本节将介绍我们模型的训练详情。

**CIFAR10** 对于 CIFAR10 使用的 ResNet50 架构，我们采用随机梯度下降训练 200 个 epoch，动量为 0.9，初始学习率为 0.1。学习率每 50 个 epoch 衰减为原来的 0.1 倍，并设置 0.0005 的权重衰减系数。我们采用标准增强策略，包括随机水平翻转、随机裁剪、小幅随机旋转（$\leq 2°$）以及随机颜色抖动。所有主要结果中的 CIFAR10 模型均采用相同设置，包括经过对抗训练的模型、使用 SAM 训练的模型（基础优化器仍为随机梯度下降），以及在鲁棒CIFAR10数据集上训练的模型。

**SVHN** 对于 SVHN，我们使用 ResNet18 架构，并采用随机梯度下降训练 20 个 epoch，动量为 0.9，初始学习率为 0.01。学习率在第 10 个 epoch 衰减为原来的 0.1 倍，并设置权重衰减系数 $0.0005$。增强策略采用随机裁剪和小幅随机旋转（$\leq 8°$）。附录 D中的所有 SVHN 模型均采用相同设置，包括经过对抗训练的模型、使用 SAM 训练的模型（基础优化器仍为随机梯度下降），以及在鲁棒 SVHN 数据集上训练的模型。

**对抗训练** 为每个数据集生成对抗训练模型时，我们使用受 $\ell_\infty$约束的PGD攻击，最大扰动为 $\epsilon = 8/255$。我们的攻击分7步进行，步长为 $\epsilon/5$。

## D. SVHN结果

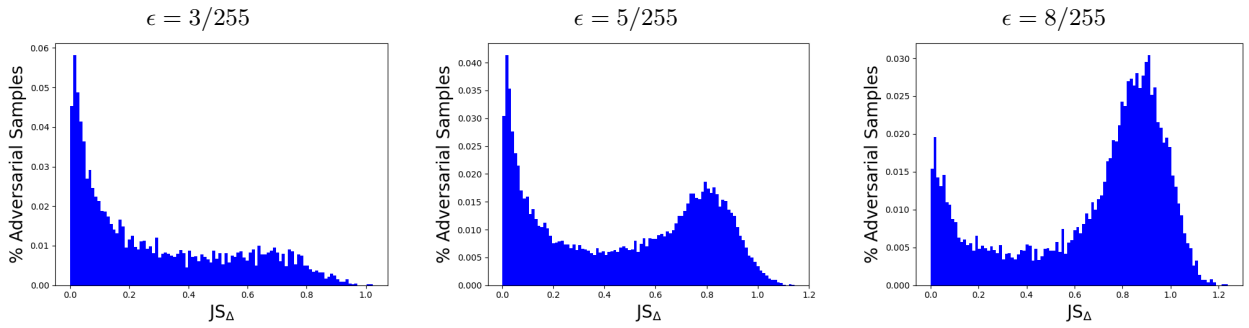



图8。在SVHN上训练的非鲁棒ResNet18模型的成功对抗样本的 $JS_\Delta$ 归一化直方图。我们观察到，扰动幅度较小时，很大比例的对抗样本可被识别为对抗性错误。扰动幅度较大时，攻击者能够操纵大多数对抗样本中的非鲁棒特征。

| 攻击强度（$\epsilon$） | 鲁棒准确率 | % 样本，其中 $JS_\Delta <$ | | |
|---|---|---|---|---|
| | | 0.01 | 0.05 | 0.10 |
| 3 / 255 | 47.7% | 0.4% | 22.3% | 34.7% |
| 5 / 255 | 17.5% | 0.2% | 13.6% | 21.5% |
| 8 / 255 | 4.0% | 0.1% | 0.6% | 10.3% |
| 16 / 255 | 0.7% | 0.0% | 0.7% | 1.2% |

表 10。在 SVHN 上训练的非鲁棒ResNet18模型中，$JS_\Delta < \beta$for $\beta = [0.01, 0.05, 0.10]$ for 的对抗样本百分比。所有样本均通过定向PGD-100攻击生成。我们发现，弱攻击往往会利用对抗性错误，而当攻击扰动超过某个阈值后，相当一部分对抗样本会利用非鲁棒特征。

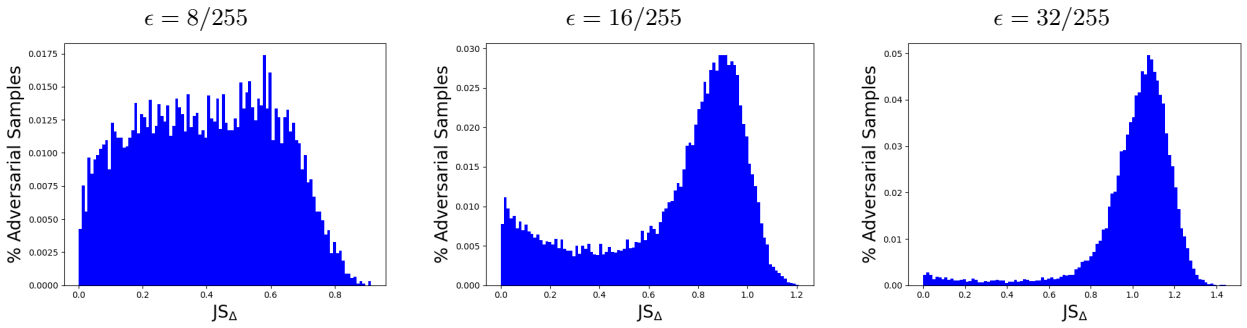



图 9. 在 SVHN 上经过对抗训练的鲁棒 ResNet18 模型的成功对抗样本的 $JS_\Delta$ 归一化直方图。样本由定向PGD-100攻击生成。我们观察到，要生成利用非鲁棒特征的对抗样本，需要更大的攻击扰动，并且对抗性错误显著减少。

| 攻击强度（$\epsilon$） | 鲁棒准确率 | 具有 $JS_\Delta$ <的样本百分比 | | |
|---|---|---|---|---|
| | | 0.01 | 0.05 | 0.10 |
| 5 / 255 | 82.1% | 0.6% | 5.2% | 14.2% |
| 8 / 255 | 63.2% | 0.4% | 3.9% | 9.4% |
| 16 / 255 | 15.9% | 0.6% | 3.8% | 7.0% |
| 32 / 255 | 1.3% | 0.1% | 0.8% | 1.3% |

表11。对于在 SVHN 上经过对抗训练的鲁棒 ResNet18 模型， $JS_\Delta < \beta$ 的 $\beta = [0.01, 0.05, 0.10]$ 百分比。所有样本均通过定向PGD-100攻击生成。我们观察到，要生成利用非鲁棒特征的对抗样本，需要更大的攻击扰动，同时对抗性错误也显著减少。

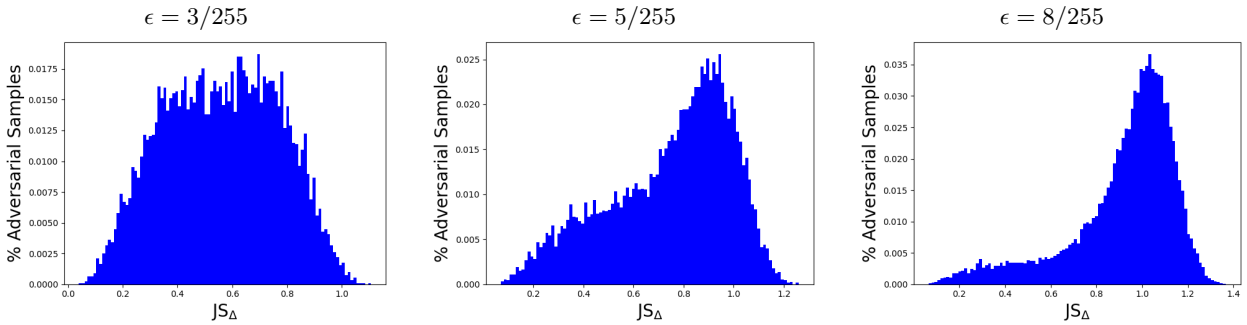



图10。成功的对抗样本的 $JS_\Delta$ 归一化直方图。使用 SAM（$\rho = 0.1$）训练的 SVHN ResNet18 模型的成功对抗样本。样本通过定向PGD-100攻击生成。与使用随机梯度下降训练的对应模型（图 8）相比，我们观察到对抗性错误明显减少。

| 攻击强度（$\epsilon$） | 鲁棒准确率 | % 个样本带有 $JS_\Delta <$ | | |
|---|---|---|---|---|
| | | 0.01 | 0.05 | 0.10 |
| 3 / 255 | 64.0% | 0.0% | 0.0% | 0.2% |
| 5 / 255 | 29.4% | 0.0% | 0.0% | 0.1% |
| 8 / 255 | 8.3% | 0.0% | 0.0% | 0.0% |

表 12。对抗样本的百分比，其中 $JS_\Delta < \beta$ 针对 $\beta = [0.01, 0.05, 0.10]$ 针对通过 SAM 训练的 ResNet18 模型（$\rho = 0.1$）在 SVHN 上。样本使用定向PGD-100攻击生成。与通过随机梯度下降训练的对应模型（表 10）相比，我们观察到对抗性错误显著减少。

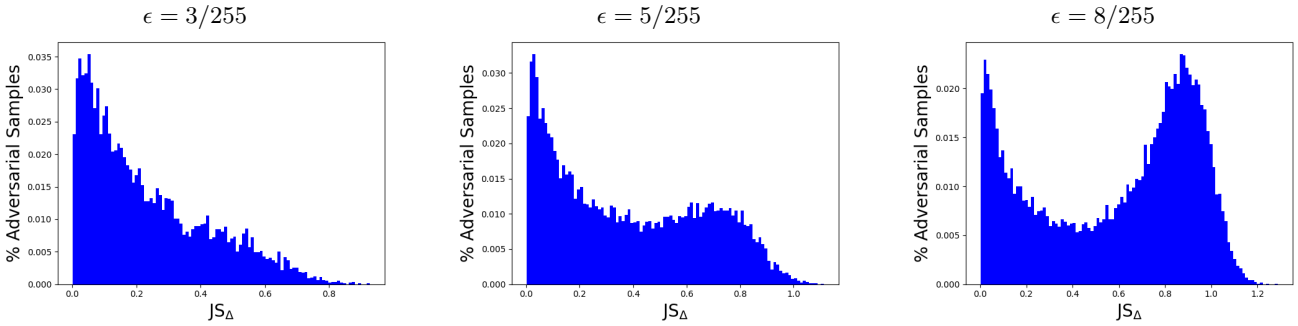



图 11. 在鲁棒 SVHN 数据集上训练的 ResNet18 模型的成功对抗样本（通过定向PGD-100攻击生成）的 $JS_\Delta$ 归一化直方图。我们观察到，攻击者能够以与标准训练模型相当的幅度操纵数据特征，这表明非鲁棒特征再次出现。

| 攻击强度（$\epsilon$） | 鲁棒准确率 | 具有 $JS_\Delta$ <的样本百分比 | | |
|---|---|---|---|---|
| | | 0.01 | 0.05 | 0.10 |
| 3 / 255 | 66.3% | 2.2% | 16.4% | 31.9% |
| 5 / 255 | 37.5% | 1.5% | 12.6% | 22.9% |
| 8 / 255 | 12.7% | 1.0% | 8.0% | 14.0% |
| 16 / 255 | 1.6% | 0.1% | 1.2% | 2.2% |

表 13。在鲁棒 SVHN 数据集上训练的 ResNet18 模型中， $JS_\Delta < \beta$的对抗样本占比为 $\beta = [0.01, 0.05, 0.10]$。所有样本均通过定向PGD-100攻击生成。我们观察到，攻击者能够以与标准训练模型相当的幅度操纵数据特征，这表明非鲁棒特征再次出现。

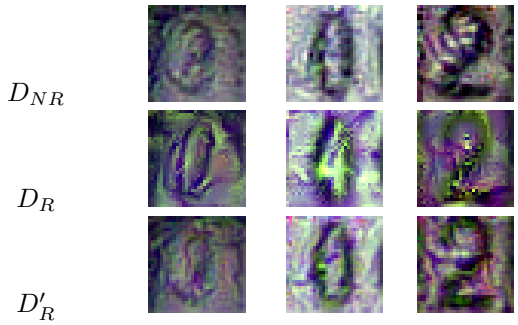



图 12。（SVHN）从 $D_{NR}$（第一行）、 $D_R$（中间行）和二阶鲁棒数据集 $D'_R$ 中的输入样本，这些数据集均由 $M_R$ 训练（最后一行）。我们观察到， $D'_R$ 中的图像与其源图像相似，如 $D_{NR}$ 所示；而只有 $D_R$ 中的输入在视觉上似乎与目标标签相关。