

对抗攻击的无数据替代训练

周 1, 2,, 京物 1, 2, j, 柳 1, 柳 2, 策珠 1

中国电子科技大学

旷视科技技术 2

{周明毅, 吴京}@std.uestc.edu.cn, {一品六, 六花一城, eczhu}@uestc.edu.cn。

抽象的

机器学习模型容易受到敌对例子的影响。对于黑盒设置, 当前的替代攻击需要预先训练的模型来生成对抗的例子。然而, 在现实世界的任务中很难获得预先训练的模型。在本文中, 我们提出了一种无数据替代训练方法来获得对抗黑盒攻击的替代模型, 而不需要任何真实数据。为了实现这一点, DaST 利用专门设计的生成对抗网络(GANs)来训练替代模型。特别地, 我们为生成模型设计了多分支结构和标签控制损失, 以处理合成样本的不均匀分布。然后, 由生成模型生成的合成样本训练替代模型, 这些样本随后被攻击模型标记。实验表明, 与由相同训练集训练的基线模型和被攻击模型相比, 由 DaST 产生的替代模型可以获得更好的性能。此外, 为了评估该方法在现实任务中的实用性, 我们在微软 Azure 平台上开发了一个在线机器学习模型。远程模型错误地分类了 98.35% 由我们的方法制作的对抗性例子。据我们所知, 我们是第一个在没有任何真实数据的情况下训练对抗攻击替代模型的人。我们的代码是公开的。

1. 正式介绍

深层神经网络已被证明易受带有不易察觉的扰动的例子的影响[38]。这使得研究人员对研究攻击和防御以评估和提高网络的健壮性非常感兴趣。敌对攻击方法可以分为两种主要的攻击, 白盒攻击

具有部分模型信息的攻击模型和黑盒攻击。

与白盒攻击相比, 黑盒攻击在现实世界系统中更实用。在这些攻击中, 基于分数的攻击[8, 19, 20, 16]和基于决策的攻击[3, 9, 7]利用被攻击模型返回的类概率或硬标签直接攻击被攻击模型。这些攻击方法不需要预先训练的替代模型, 但是, 作为代价, 它们需要对被攻击模型的大量查询来生成每个攻击。

相反, 基于梯度的攻击方法[14, 22, 35, 30]需要被攻击模型的体系结构和权重的知识。Goodfellow 等人。[14]表明, 对抗实例具有可转移性, 这意味着通过白盒攻击方法为一个模型生成的对抗实例也可以攻击其他模型。因此, 为了在黑盒子环境中实施攻击方法, 他们使用替代模型来找到对抗实例, 然后基于这些对抗实例的可转移性来攻击机器学习模型。与当前基于分数和基于决策的攻击相比, 替代攻击不需要查询来生成对抗实例。然而, 他们需要一个预训练的模型来产生对抗性攻击。Papernot 等人。[34]开发了一种方法, 使用大量图像来模拟受攻击模型的输出, 以获得替代网络。预测应用程序接口也被开发来窃取机器学习模型[39]。Orekondu 等人。[32]提出了“山寨”来窃取机器学习模型的功能。这些方法不需要预先训练的模型, 而是需要许多被攻击模型标记的真实数据来训练替代模型。然而, 在一些现实世界的任务中, 很难获得真实的图像。因此, 开发一种无数据替代攻击非常重要, 这样可以更全面地评估当前机器学习模型面临的风险。

在本研究中, 我们提出了一种无数据替代训练方法来训练对抗攻击的替代模型。我们利用生成性对抗网络创建合成样本来训练替代模型。替代模型使用这些样本进行训练, 其中样本的标签由被攻击的模型产生。为了性能, 合成样本应该在输入空间中均匀分布。样品的标签应涵盖所有类别。然而, 没有真实数据的常规遗传神经网络可能产生分布极不均匀且只包含少量类别的样本, 这意味着替代模型不能全面地学习被攻击模型的分类特征。

为了解决这个问题, 我们为生成模型设计了一个多分支结构和一个标签控制损失来处理合成样本的不均匀分布。生成模型可以产生带有被攻击模型给出的随机标签的合成样本。因此, 替代模型可以在对抗训练中学习被攻击模型的分类特征, 并生成对被攻击模型具有很强可移植性的对抗实例。本研究的主要贡献总结如下:

- 我们是第一个在没有任何真实数据的情况下训练对抗攻击替代模型的人。攻击者可以使用这种方法来训练对抗攻击的替代模型, 而无需收集任何真实数据。

- 我们在本地深度学习模型和在线机器学习系统上评估了 DaST 的有效性，这揭示了一个事实，即当前的机器学习模型具有被攻击的显著风险。
- 我们评估了我们的方法在两种攻击场景中的性能，包括攻击者可以访问被攻击模型的输出概率的纯概率场景和攻击者只能访问被攻击模型的输出标签的纯标签场景。我们的方法在两种情况下都能有效地生成对抗性的例子。

此外，我们使用不同的模型架构来替代模型，以测试模型容量对攻击成功率的影响。

论文的其余部分组织如下：在第二部分，我们介绍了相关的工作。第 3 节描述了所提出的方法。我们在第 4 节中评估了 DaST 的性能。

2. 相关作品

敌对攻击是在白盒或黑盒环境下进行的。在白盒设置中，攻击者可以访问被攻击模型的结构和权重。相反，在黑盒设置中，攻击者只有替代模型(基于梯度的攻击)或访问被攻击模型返回的输出(基于查询的攻击)。黑盒攻击方法在现实任务中更实用。

对抗性攻击基于梯度的攻击，如 FGSM [14] 和 BIM [22] 对模型有完全的访问权限，所以它们通常使用预先训练的替代模型来生成对抗性例子，然后利用对抗性例子的可转移性来攻击被攻击的模型。FGSM 的目的是通过直接增加模型的损失来寻找对立的例子，BIM 是 FGSM 的迭代版本。同样地，DeepHull[30] 发现了可能跨越决策边界的对抗性例子。为了找到具有最小 p 范数的扰动，尼古拉斯·卡利尼和戴维·瓦格纳[6] 介绍了一种通过同时最小化扰动来构造这些扰动的方法。类似于这个方法，Rony 等人。[36] 也约束扰动的 2 范数，它们解耦扰动的值和方向。在黑盒环境中，这些攻击依赖于对抗性例子的可转移性。不过，刘等人。[25] 表明这些例子对攻击几乎没有可转移性。反而是程等人。[8] 提出了一种基于分数的零阶攻击方法(ZOO)，使用梯度估计，和易勒雅斯等人。[20] 改进梯度估计的方法。代替梯度估计，郭等人。[16] 介绍了一种简单的黑盒攻击(SimBA)，它根据输出概率的变化来决定扰动的方向。Brendel 等人。[3] 首先提出了基于决策的攻击。基于这种方法，程等人。[9] 和程等人。[7] 提高了查询效率，这是黑盒攻击的一个重要指标。

对抗防御已经提出了几种增强模型鲁棒性的防御方法。对抗训练[38, 27, 23, 40] 修改了模型的训练方案，他们直接用对抗的例子进行训练。另一种方法旨在修改对抗性例子本身，如随机变换[22, 28, 41]。巴克曼等人。[4] 提出了一种基于单热编码的非线性变换模型输入。梯度掩蔽方法[40, 10] 破坏梯度信息，因此它们不能通过基于优化的攻击。然而，这些基于梯度掩蔽的防御方法已经被证明是不可靠的[1]，并且具有上述防御的模型对于一些攻击仍

然是不安全的[5, 17]。此外，检测敌对的例子提高了研究人员的兴趣。他们中的一些人通过一个辅助网络来检测他们是敌对的还是干净的例子[13, 15, 29]，而一些人通过他们的统计属性来找出敌对的例子[2, 18, 12, 26, 33]。

3. 方法

在这一部分，我们描述了本研究中的攻击场景，然后介绍了替代攻击，并提出了一种无数据的方法来训练替代模型。

3.1. 攻击场景

仅标签场景假设在线使用被攻击的机器学习模型，攻击者可以自由地探测被攻击模型的输出标签。攻击者很难获得被攻击模型的输入空间中的任何数据。我们在仅标注场景中将建议的 DaST 命名为 DaST-L。

仅概率场景此场景的其他设置与仅标签场景相同，但攻击者可以访问被攻击模型的输出概率。我们在概率唯一的情况下将提议的 DaST 命名为 DaST-P。

3.2. 敌对攻击

在这一小节中，我们介绍了对抗性替代攻击的定义。

\mathbf{x} 表示来自被攻击模型 t 的输入空间的样本。 y 和 y_0 分别指样品 X 的真实标签和目标标签。 $T(y|X, \theta)$ 是 θ 参数化的被攻击模型。对于非目标攻击，对抗性攻击的目标可以表述为：

$$\min_{\epsilon} \|\epsilon\| \text{ 服从 } \arg \max \text{ 和}$$

对于有针对性的攻击，目标是：

$$\min_{\epsilon} \|\epsilon\| \text{ 服从 } \arg \max \text{ 和}$$

其中 ϵ and r 分别是样本的扰动和扰动的上限。为了攻击难以检测的机器学习系统，设置了 r

攻击方法的小数值。 $\mathbf{X} = \mathbf{X} + \epsilon$ 是

可能导致被攻击的模型 T 输出错误标签(非目标设置)和特定错误标签(目标设置)的对抗性例子。

对于白盒攻击，他们可以充分获取 T 的梯度信息，然后利用它生成对抗的例子来攻击 T 。对于黑盒替代攻击，他们训练一个 T_b 模型来替代被攻击的模型，生成对抗性的例子，然后将这些例子转移到攻击 t 。这些黑盒攻击的成功率在很大程度上依赖于对抗性例子的可转移性。因此，开发一个高效的替代攻击的关键点是训练一个替代模型，该模型的属性尽可能与被攻击的模型相似。当前的攻击方法利用被攻击模型的相同训

练集或者收集被攻击模型标记的大量其他图像来训练替代模型。在接下来的两小节中，我们将介绍一种可以在没有任何图像的情况下训练替代模型的方法。整个过程如图 1 所示。

3.3. 对抗性生成分类器训练

在这一小节中，我们介绍了基本的对抗性训练方法，并讨论了其局限性。

为了训练没有任何图像的替代模型，我们使用生成模型 G 来产生替代模型 D 的训练数据。生成器从输入空间随机采样噪声向量 z 并产生数据

$\hat{X} = G(z)$ 。然后，使用生成的数据探测被攻击模型 T 的输出 $T(Xb)$ 。替代模型由图像输出对 $(\hat{X}, T(\hat{X}))$ 。如图 1 所示，

G 的目标是创建新的样本，探索 T 和 D 的区别， D 的作用是模仿 T 的输出。这是一个特殊的两人游戏，这个游戏中涉及的被攻击模型是一个裁判。为了简化表达式但不失一般性，我们利用二进制分类作为案例进行分析(输出概率在二进制分类中可以看作一个标量，输出标签也可以)。游戏的价值函数表示为：

$$\text{最大最小 } VG, D = d(T(Xb), D(Xb)) \quad (3)$$

其中 $d(T(\hat{X}), D(\hat{X}))$ 是度量 T 和 d 之间输出距离的指标。

对于仅标签攻击场景，该度量可以表述为：

$$d(T, D) = CE(D(\hat{X}), T(\hat{X})), \quad (4)$$

其中 $D(Xb)$ 和 $T(Xb)$ 分别表示替代模型和被攻击模型的输出标签。 $CE(D(Xb), T(Xb))$ 表示交叉熵损失， T 的输出标签被用作该损失的标签。交叉熵损失的作用是约束 T 和 d 之间的差异。对于概率攻击场景，该度量公式如下：

$$d(T, D) = \|D(\hat{X}), T(\hat{X})\|_F \quad (5)$$

其中 $D(X)$ 和 $T(X)$ 分别表示替代模型和被攻击模型的输出概率。

因此，替代模型 D 通过这种对抗训练复制了被攻击模型 T 的信息。在训练中， D 的损失函数设置为 $LD = VG$ ， D 。为了保持训练的稳定性，将 G 的损失函数设计为 $LG = e - D(T, D)$ 。因此，当且仅当 $\forall \hat{X}, T(\hat{X}) = D(\hat{X})$ 时，得到全局最优替代网络 D 。此时， $LD = 0$ 且

$$LG = e - 0 = 1。$$

/

图 1。提议的对抗性无数据模仿。 G 的架构如蓝色虚线框所示。 n 表示类别数。在训练阶段， G 的目标是生成样本 $Xb = G(X)$ ，让 $yD(Xb) = yT(Xb)$ 。 D 的目标是保证。在测试阶段，利用替代模型 D 生成对抗实例来攻击 t 。

我们假设 $\forall \hat{X} = G(z), \hat{X} \in \mathbb{R}, \mathbb{R}$ 是 t 的输入空间。

如果 D 能达到 $D(X) = T(X)$ ，那么我们的替代模型所进行的对抗攻击将会有和没有 T 的梯度信息的白盒攻击一样的成功率。因此，对于一个训练有素的替代网络来说， D 生成的对抗性例子对 t 具有很强的可转移性。

但是，无法保证 $D(\hat{X}) =$

$T(Xb)$ 在有限的时间内。如果我们不约束 G 的输出， T 的合成训练数据很可能只分布在 R 的小范围内，因此这种训练无法工作。为了解决这个问题，我们为 G 设计了一个标签可控的架构，可以控制合成数据的分布，加快训练的收敛速度。

3.4. 标签可控的数据生成

在这一小节中，我们介绍了生成模型 G 的标签可控架构。

为了获得均匀分布的合成数据来训练替代模型 D ，我们考虑开发一种可以控制 Xb 分布的方法。为了训练 T 的复制，合成数据被用来探测被攻击模型的信息。被攻击模型产生的样本标签应该跨越所有类别。因此，如图 1 的蓝色虚线框所示，我们设计了一个包含 N 个上采样去卷积成分的生成网络， N 是类别数。所有上采样组件共享一个后处理卷积网络。模型 G 从输入空间和可变标签值 n 中随机采样噪声向量 z 。然后将 z 输入第 n 个上采样去卷积网络和共享卷积网络，以产生数据 $\hat{X} = G(z, n)$ 。生殖模型 G 的附加标签控制损失公式如下：

$$LC = CE(T(G(z, n)), n). \quad (6)$$

以上方法生成带有随机标签的数据，由 t 产生。然而，这种标签控制损失的反向传播需要被攻击模型 T 的梯度信息，这违反了黑盒攻击的规则。我们需要训练一个没有 t 的梯度信息的标签可控的生成模型。对于模拟过程，可以近似为以下目标函数：

$$\min_D d(T(\hat{X}), D(\hat{X})). \quad (7)$$

在训练过程中，在相同的输入下， D 的输出将逐渐接近 T 的输出。所以我们用 D 来代替等式中的 T 。(6)，其表述为：

$$LC = CE(D(G(z, n)), n). \quad (8)$$

替代 D 的训练可以避免获取 t 的信息。然后我们把 G 的损失更新为:

$$LG = e D(T, D) + \alpha LC, \quad (9)$$

其中 α 控制标签控制损失的重量(我们在实验中将其设置为 0.2)。

在训练阶段,随着 D 的模仿能力增加,由 T 标注的合成样本的多样性会增强。因此, D 可以学习被攻击模型 T 的信息,可以提高 D 生成的对抗性例子的可转移性。我们

算法 1 小批量随机梯度下降训练的提出方法 DaST.

acc 表示 d 的精度。att 表示 d 产生的攻击的攻击成功率。
1:当迭代 $<\delta$ 或 acc 时, att 不增加
2,生成 m 个示例{ $Xb(1)$ }, ..。 $XB(m)$ }由 g 。 3:更新替代模型:
4: $LD = d(T(Xb), D(Xb))$ 。
5: 更新创成式模型:
6: $LG = e D(T, D) + \alpha LC$ 。
7:结束时间

将此方法命名为无数据替代训练,如算法 1 所示。
与目前的替代攻击方法一样,我们的方法训练的替代模型被用来生成攻击 t 的对抗性例子。

4. 实验

4.1. 实验设置

在这一小节中,我们介绍了我们的实验设置,包括数据集、模型架构、攻击方法和评估标准。

数据集:我们在 MNIST [24]和 CIFAR-10 [21]上评估了我们提出的方法。这两个数据集的测试集分别有 10k 个图像。

场景:我们在仅标签攻击和仅概率攻击场景中评估我们的方法。DaST-L 和 DaSTP 分别表示仅标签场景中的 DaST 和仅概率场景中的 DaST。本研究场景中的攻击者可以自由访问被攻击模型的输出。因此,当算法收敛时,我们得到由 DaST 训练的替代模型。

模型架构和攻击方法:替代网络没有被攻击模型的先验知识,这意味着它在实验中不加载任何预先训练好的模型。在 MNIST 的实验中,我们设计了 3 种不同的网络结构,包括一个小型网络(3 个卷积层)、一个中型网络(4 个卷积层)和一个大型网络(5 个卷积层),用于评估我们的具有不同容量模型的分布式存储测试的性能。我们利用预先训练的

中型网络和 VGG-16 [37]分别作为 MNIST 和 CIFAR-10 上的攻击模型。此外,在 CIFAR-10 实验中,我们对替代模型和攻击模型使用不同的体系结构来评估模型结构对我们的方法的影响。为了比较由 DaST 产生的替代模型和预先训练的模型,我们利用 4 种攻击方法来产生对抗的例子,包括 FGSM [14], BIM [22], 投影梯度表 1。拟在 MNIST 安装的分布式存储系统的性能。“预训练”、“DaST-L”和“DaST-P”:由预训练的大型网络和 DaST-L、DaST-P 分别生成的对抗实例的攻击成功率(%). ()表示每幅图像的平均低频扰动距离。

Attack	Non-targeted		
	Pre-trained	DaST-P	DaST-L
FGSM [14]	59.72 (5.40)	69.76 (5.41)	35.74 (5.40)
BIM [22]	85.70 (4.80)	96.36 (4.81)	64.61 (4.82)
PGD [27]	37.93 (3.98)	53.99 (3.99)	23.22 (3.98)
C&W [6]	23.34 (2.91)	27.35 (2.74)	18.16 (2.75)

Attack	Targeted		
	Pre-trained	DaST-P	DaST-L
FGSM [14]	12.10 (5.46)	20.45 (4.49)	13.10 (5.46)
BIM [22]	37.83 (4.90)	57.22 (4.87)	29.18 (4.87)
PGD [27]	28.95 (4.60)	47.57 (4.63)	19.25 (4.63)
C&W [6]	10.32 (2.57)	23.80 (2.99)	12.31 (2.98)

血统(PGD) [27], C&W [6]。为了测试,我们使用[图书馆的广告来生成对抗性的例子。为了评估该方法在现实世界任务中的性能,我们将攻击应用于微软 Azure 的在线 MNIST 模型。该在线模型使用的训练技巧和机器学习方法无法访问。

评估标准:为了评估我们的 DaST 的性能,我们将由其他预先训练的网络生成的敌对示例的攻击成功率设置为基线。非目标攻击和目标攻击的目标分别是导致被攻击模型输出错误标签和特定错误标签。在非目标攻击场景中,我们只在被攻击模型正确分类的图像上生成对抗示例。在有针对性的攻击中,我们只在没有归类到特定错误标签的图像上生成敌对的例子。对抗攻击的成功率用 n/m 来计算,其中 n 和 m 分别是可以欺骗被攻击模型的对抗实例数和对抗实例总数。

4.2. MNIST 实验

在这一小节中,我们使用所提出的 DaST 来训练 MNIST 数据集上对抗攻击的替代模型,并根据仅标记和仅概率场景中的攻击成功率来评估性能。

首先,我们进行实验来评估在仅概率和仅标签攻击场景中的性能。我们使用中型网络作为对 MNIST 的攻击模型,使用大型网络作为 DaST 的替代模型。我们在被攻击模型的同一列车组上训练一个预先训练好的大型网络。我们利用攻击的成功

表 2。MNIST 三种不同替代体系结构的拟建分布式存储系统的性能。“小”、“中”、“大”:DaST 分别与小、中、大替代网络生成的对抗实例的攻击成功率(%)。()表示每幅图像的平均低频扰动距离。

Attack	Non-targeted		
	Small	Medium	Large
FGSM [14]	62.61 (4.38)	56.21 (4.45)	69.76 (5.41)
BIM [22]	94.86 (4.85)	92.47 (4.84)	96.36 (4.81)
PGD [27]	45.31 (3.99)	43.62 (3.99)	53.99 (3.99)
C&W [6]	30.61 (2.89)	24.34 (2.75)	23.80 (2.99)
Attack	Targeted		
	Small	Medium	Large
FGSM [14]	19.92 (4.43))	20.45 (4.49)	23.93 (5.45)
BIM [22]	56.73 (4.89)	53.50 (4.84)	57.22 (4.87)
PGD [27]	39.42 (4.64)	40.76 (4.60)	47.57 (4.63)
C&W [6]	24.86 (3.09)	16.25 (3.13)	23.80 (2.99)

由预先训练的模型作为基线生成的敌对示例的比率。表 1 显示了我们的数据采集终端的性能.用 DaST-P 和 DaST-L 训练的替代模型在测试集上的准确率分别达到 97.82%和 83.95%。我们的 DaST 生成的替代模型的攻击成功率高于非目标的预训练模型(在 FGSM、BIM、PGD 和上分别为 10.04%、10.66%、16.06%和 4.01%)

C&W)和有针对性的攻击(分别比 FGSM、BIM、PGD 和 C&W 高 11.83%、19.39%、18.62%和 13.48%)。结果表明，由 DaST-P 生成的替代模型的性能优于由相同训练集(60000 幅图像)和被攻击模型训练的模型。就连他用 DaST-L 训练的替代模型在 FGSM 和 C&W 攻击上的表现也比基线模型好(有针对性)。

然后我们评估了我们的 DaST 在不同替代架构下的性能。我们还使用中型网络作为 MNIST 的攻击模型，并使用三种不同的替代体系结构(包括大型、中型和小型网络)来应用我们的 DaST。这三种替代体系结构的攻击成功率如表 2 所示.与其他模型相比，大型替代模型在 FGSM、BIM、PGD 攻击上取得了最佳效果。与其他模型相比，小替换模型在 C&W 攻击上获得了最好的结果。结果表明，替代模型的两种体系结构在对抗攻击上都取得了良好的效果。一般来说，结构更复杂的替代模型可以获得更好的对抗攻击性能。

4.3. 在 CIFAR-10 上的实验

在本小节中，我们使用建议的 DaST 为 CIFAR 上的对抗性攻击训练一个替代模型-

表 3 .提议的 DaST 在 CIFAR-10 上的性能。“预训练”、“DaST-P”、“DaST-L”:由预训练的大型网络、DaST-P 和 DaST-L 分别生成的对抗实例的攻击成功率(%)。()表示每幅图像的平均低频扰动距离。

Attack	Non-targeted		
--------	--------------	--	--

	Pre-trained	DaST-P	DaST-L
FGSM [14]	39.10 (1.54)	39.63 (1.54)	22.65 (1.54)
BIM [22]	59.18 (1.01)	59.71 (1.18)	28.42 (1.19)
PGD [27]	35.40 (1.02)	29.10 (1.10)	17.80 (1.10)
C&W [6]	9.76 (0.77)	13.52 (0.74)	10.34 (0.74)
Attack	Targeted		
	Pre-trained	DaST-P	DaST-L
FGSM [14]	9.62 (1.54)	6.69 (1.54)	7.32 (1.54)
BIM [22]	17.43 (1.00)	20.22 (1.18)	15.26 (1.16)
PGD [27]	10.46 (1.05)	14.09 (1.12)	8.32 (1.10)
C&W [6]	23.15 (2.05)	26.53 (1.98)	19.78 (2.04)

10 数据集，并根据仅标记和仅概率场景中的攻击成功率来评估性能。

我们进行了实验，以评估在概率和标签攻击场景中的性能，并使用 VGG-16 网络作为攻击模型。我们在被攻击模型的同一个训练集上训练一个预训练的 ResNet-50 网络。表 3 显示了我们的数据采集终端的性能.用 DaST-P 和 DaST-L 训练的替代模型在测试集上的准确率分别达到 25.15%和 20.35%。我们的 DaST 还通过预先培训的模型实现了有竞争力的性能。在大多数只考虑概率的情况下(FGSM、BIM、C&W 针对非目标攻击、BIM、PGD、C&W 针对目标攻击)，DaST-P 生成的替代模型优于基线模型。由 DaST-L 训练的替代模型在 C&W 攻击(非目标)上的表现优于基线模型。

我们还评估了我们的 DaST 在不同替代架构下的性能。VGG-16 网络被用作攻击模型。我们使用 3 种不同的替代体系结构来应用我们的分布式存储系统，包括 VGG-13、ResNet-18 和 ResNet50.这三种替代体系结构的攻击成功率如表 4 所示.这表明替代模型的两种体系结构在对抗攻击上都取得了良好的效果。在大多数情况下(针对非目标攻击，PGD、C&W；针对目标攻击，新加坡、新加坡、PGD、C&W)，VGG-13 在对抗攻击方面优于其他型号。ResNet-50 在 FGSM 攻击上获得最佳效果(有针对性)。与在 MNIST 的实验不同，该简单模型在 CIFAR-10 上取得了最好的结果.我们分别在图 2 和图 3 中可视化了由 DaST-P 和 DaST-L 生成的对抗性例子。这两种情况的攻击扰动都很小。

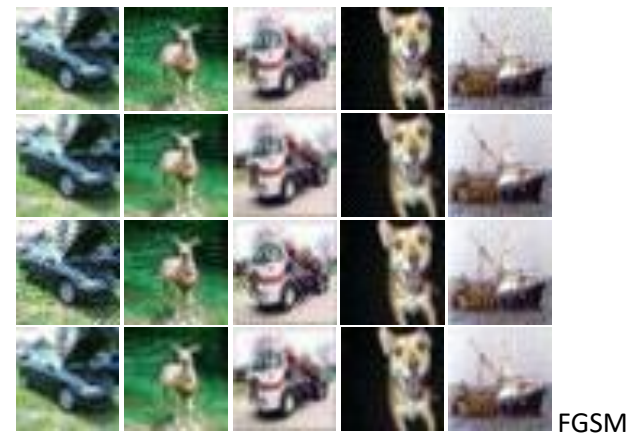
表 4 .在 CIFAR-10 上采用三种不同替代体系结构的建议 DaST 的性能。“VGG-13”、“雷斯网-18”、“雷斯网-50”:分别用 VGG-13、雷斯网-18、雷斯网-50 替代模型，由 DaST 生成的对抗实例的攻击成功率(越高越好)。()中的数字表示每幅图像的平均低频扰动距离。

	非目标(%)。
--	---------

	VGG-13	ResNet-18	ResNet-50
FGSM [14]	6.87 (1.54)	17.97 (1.54)	39.63 (1.54)
BIM [22]	93.13 (1.18)	31.70 (1.54)	59.71 (1.18)
PGD [27]	56.14 (1.08)	10.04 (1.11)	29.10 (1.10)
C&W [6]	56.80 (1.64)	11.54 (1.64)	13.52 (0.74)

	Targeted (%)		
Attack	VGG-13	ResNet-18	ResNet-50
FGSM [14]	18.27 (1.54)	2.07 (1.54)	6.69 (1.54)
BIM [22]	62.23 (1.24)	8.00 (1.52)	20.22 (1.18)
PGD [27]	41.48 (1.17)	3.72 (1.26)	14.09 (1.12)
C&W [6]	33.65 (2.42)	7.31 (1.46)	26.53 (1.98)

攻击



漂亮但不聪明的姑娘

顺时针方向的行波(Continuous Wave)

羟基前列腺素脱氢酶

图 2.CIFAR-10 上由 DaST-L 生成的敌对示例的可视化。我们为每次攻击生成 5 个样本。

4.4. 微软 Azure 上的实验

在这一小节中，我们在两个场景中进行了攻击微软 Azure 在线模型的实验。

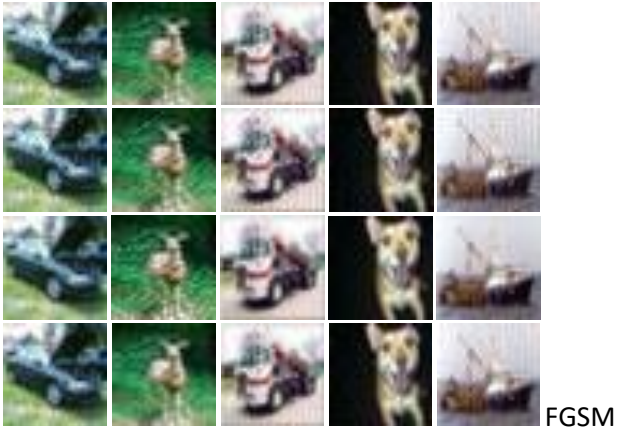
我们使用 Azure 上的机器学习教程的示例 MNIST 模型作为被攻击的模型，并将其用作网络服务。我们不知道这种模型的机器学习方法和体系结构。我们唯一能获得的信息就是这个模型的输出。我们将基于概率的 DaST 和基于标签的 DaST 攻击应用于该模型，以评估该方法在实际应用中的性能。本实验中的替代模型有 5 个卷积层。用 DaST-P 和 DaST-L 训练的替代模型达到 79.35%

表 5。所提出的攻击微软 Azure 示例模型的 DaST 的性能。“预训练”、“DaST-P”、“DaST-L”:由预训练的大型网络、仅概率场景下的 DaST 和仅标签场景下的 DaST 分别生成的对抗实例的攻击成功率(越高越好)。()中的数字表示每幅图像的平均低频扰动距离。因为很难为 C&W [6]上的所有方法生成对抗性的例子，所以我们省略了这种攻击方法。

	非目标(%)。		
	Pre-trained	DaST-P	DaST-L
FGSM [14]	77.96 (5.41)	96.83 (5.25)	98.21 (5.36)
BIM [22]	66.25 (4.81)	96.42 (4.79)	98.35 (4.72)
PGD [27]	59.23 (3.99)	90.63 (3.88)	96.97 (3.96)

	Targeted (%)		
Attack	Pre-trained	DaST-P	DaST-L
FGSM [14]	13.52 (5.46)	32.00 (5.21)	43.99 (5.37)
BIM [22]	19.31 (4.88)	50.21 (4.90)	71.15 (4.56)
PGD [27]	19.31 (4.60)	45.66 (4.46)	65.91 (4.32)

攻击



漂亮但不聪明的姑娘

顺时针方向的行波(Continuous Wave)

羟基前列腺素脱氢酶

图 3.CIFAR-10 上 DaST-P 生成的敌对示例的可视化。我们为每次攻击生成 5 个样本。

和 90.75%的准确度。表 5 显示了所提出的方法在对抗攻击上的性能。

在这个在线模型上，DaST-L 的性能优于 DaST-P。由于被攻击的 Azure 模型过于简单，在 MNIST 上的准确率只有 91.93%。图 6 显示了 DaST-P 的训练，它比 DaST-L 能访问更多的被攻击模型的信息，但存在过度拟合。DaST-L 替代品在 FGSM (98.21%)、BIM (98.35%)、PGD (96.97%)攻击上实现了非常高的攻击成功率。此外，我们的 DaST 方法即使在有针对性的攻击中也能获得很高的攻击成功率。与由 MNIST 训练集训练的模型相比，由 DaST 训练的替代模型仅在标签上表现更好

表 6 。 DaST 与其他攻击的比较。” “ASR”:攻击成功率。“查询”:评估阶段的查询次数。“边界”:基于决策的攻击[3]。“GLS”:基于贪婪本地搜索的基于分数的黑盒攻击[31]。“-”表示我们的 DaST 在评估阶段不需要查询。本实验中的 DaST 使用 BIM 产生攻击。

攻击	自动语音识别 (Automatic Speech Recognition)	距离	问题
DaST-P	96.83%	4.79	-

GLS [31]	40.51%	4.27	297.07
DaST-L	98.35%	4.72	-
边界[3]	100%	4.69	670.54

(非目标 FGSM、BIM、PGD 攻击分别高出 20.25%、32.10%、37.74%。在有针对性的 FGSM、BIM、PGD 攻击上分别高出 30.47%、51.84%、46.60%)和仅概率场景。结果表明，我们的方法在攻击实际的在线模型方面更好，即使所提出的方法不需要任何真实数据。因为 DaST 在评估阶段不需要任何查询，而在训练阶段需要查询，所以我们的 DaST 需要的信息与基于分数的攻击和基于决策的攻击不同(它们在评估阶段需要查询)。我们显示了基于分数和基于决策的攻击的查询数量，它们在非目标攻击中与 DaST 具有相似的扰动距离。结果如表 6 所示.我们的 DaST 在训练阶段针对被攻击的模型进行了 2000 万次查询的训练。与基于决策和基于分数的攻击相比，DaST 每次访问被攻击模型的输入在训练阶段是不同的(当前的基于查询的攻击需要使用一个原始数据多次访问被攻击模型来生成每次攻击)。因此，DaST 的查询比其他攻击更难跟踪。

可视化:我们在 Azure 实验的 DaST 中可视化由生成模型生成的合成样本，如图 4 所示.我们还在图 5 中可视化了由 DaST-P 和 DaST-L 生成的对抗性例子。DaST 的攻击扰动很小。

训练收敛:我们展示了 Azure 实验训练阶段 DaST 生成的 BIM 攻击的攻击成功率曲线，如图 6 所示。DaST-L 和 DaST-P 的攻击成功率分别在 2000 万次和 200 万次查询后收敛。

5. 结论

我们提出了一种无数据的方法来训练对抗攻击的替代模型。DaST 通过利用 GANs 生成合成样本来降低对抗性替代攻击的先决条件。这是第一种不需要任何真实数据就能训练替代模型的方法。实验表明效果良好

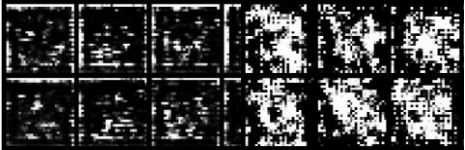
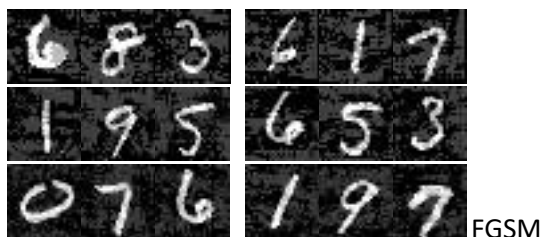


图 4。在 DaST 的训练中由发生器产生的合成样本的可视化。左:由 DaST-L 生成的样本。右:由 DaST-P 生成的样本。



漂亮但不聪明的姑娘

羟基前列腺素脱氢酶

图 5. 由 DaST 生成的攻击 Azure 模型的敌对例子的可视化。
左:DaST-P 生成的例子。右:DaST-L 生成的例子。

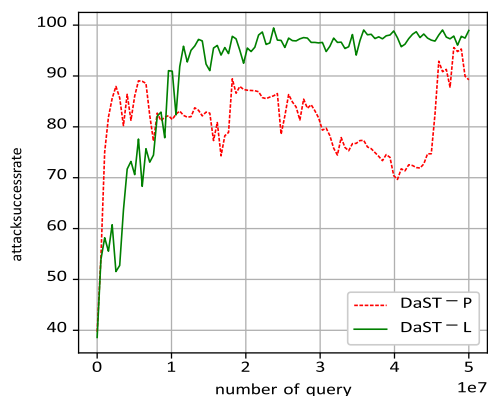


图 6. Azure 实验训练阶段 DaST 生成的 BIM 攻击的攻击成功率。

我们方法的有效性。研究表明，机器学习系统存在很大的风险，即使在真实输入数据难以收集的情况下，攻击者也可以训练替代模型。

提出的 DaST 不能单独产生敌对的例子，它应该与其他基于梯度的攻击方法一起使用。在未来的工作中，我们将设计一种新的替代训练方法，可以直接产生攻击。此外，我们将探索 DaST 的防御。

6. 承认

本研究得到了国家自然科学基金(自然科学基金，编号。61602091，没有。61571102，编号 61872067)和四川科技计划(编号。2019YFH0008，编号 2018JY0035，编号 2019YFH0016)。

引用

[1] 安妮斯·阿萨勒、尼古拉斯·卡利尼和戴维·瓦格纳。模糊的渐变给人一种错误的安全感:回避对抗的例子。《第 35 届国际机器学习会议论文集》，ICML，2018 年，2018 年 7 月。

- [2] 阿尔琼·尼廷·巴格吉、丹尼尔·库里纳和普拉泰克·米塔尔。降维作为一种防御机器学习分类器规避攻击的方法。*arXiv 预印本 arXiv:1704.02654*, 2017。
- [3] 维兰德·布兰德尔、乔纳斯·劳伯和马提亚斯·贝希。基于决策的对抗攻击:针对黑盒机器学习模型的可靠攻击。*arXiv 预印本 arXiv:1712.04248*, 2017.1, 2, 8
- [4] 雅各布·巴克曼、奥科·罗伊、科林·拉斐尔和伊恩·古德费勒。温度计编码:抵制敌对例子的一个热门方法。在学习表征国际会议(ICLR)上, 2018 年。
- [5] 尼古拉斯·卡利尼和戴维·瓦格纳。敌对的例子不容易被发现:绕过十种检测方法。《第十届美国计算机学会人工智能与安全研讨会论文集》, 第 3-14 页.ACM, 2017。
- [6] 尼古拉斯·卡利尼和戴维·瓦格纳。评价神经网络的鲁棒性。在 2017 年电气和电子工程师协会安全和隐私研讨会上, 第 39-57 页.IEEE, 2017.2, 5, 6, 7
- [7] 迈克尔·陈建波。乔丹和马丁·j. 温赖特。Hopskipjumpattack:一种基于查询高效决策的攻击, 2019.1, 2
- [8] 平·陈愉、张欢、雅什·夏尔马、金凤仪和赵瑞希。Zoo:基于零阶优化的黑盒子攻击深度神经网络, 无需训练替代模型。《第十届美国计算机学会人工智能与安全研讨会论文集》, 第 15-26 页.ACM, 2017.1, 2
- [9] 程, 通乐, 平, 金凤仪, 赵瑞雪。查询高效的硬标签黑盒子攻击:一种基于优化的方法。*arXiv 预印本 arXiv:1807.04457*, 2018.1, 2
- [10] Guneet S. Dhillon, Kamyar Azizzadenesheli, Jeremy D. 伯恩斯坦, 让·科斯赛菲, 阿兰·卡纳, 扎卡里·c. 利普顿, 阿尼玛什里·阿南德库马尔。鲁棒对抗防御的随机激活剪枝。2018 年国际学习表征会议。
- [11] 丁伟光、王鲁豫、金。AdverTorch v0.1:一个基于 pytorch 的对抗性健壮性工具箱。*arXiv 预印本 arXiv:1902.07623*, 2019。
- [12] 鲁本·费曼、瑞安·柯廷、索拉博·辛特和安德鲁·加德纳。从人工制品中检测对抗性样本。*arXiv 预印本 arXiv:1703.00410*, 2017。
- [13] 巩,, 魏世南。敌对和干净的数据不是双胞胎。*arXiv 预印本 arXiv:1704.04960*, 2017。
- [14] 伊恩·古德费勒、黄邦贤·史伦斯和克里斯蒂安·塞格迪。解释和利用敌对的例子。学习表征国际会议(ICLR), 2015. 1, 2, 5, 6, 7
- [15] 凯瑟琳·格罗斯、普拉文·马诺哈兰、尼古拉斯·帕佩诺特、迈克尔·巴克斯和帕特里克·麦克丹尼尔。对抗性例子的(统计)检测。*arXiv 预印本 arXiv:1702.06280*, 2017。
- [16] 郭川、雅各布·加德纳、尤荣·尤、安德鲁·戈登·威尔逊和基利安·温伯格。简单的黑盒对抗攻击。在国际机器学习会议上, 第 2484-2493 页, 2019 年.1, 2
- [17] 何华伦、陈、尼古拉斯·卡利尼和黎明之歌。对抗式范例防御:弱防御的集合并不强大。在 2017 年第 11 届进攻性技术研讨会(WOOT, 2017 年)上。

- [18] 丹·亨德里克斯和凯文·金佩尔。检测敌对图像的早期方法。《学习表征国际会议(ICLR)》，2017 年。
- [19] 安德鲁·易勒雅斯、洛根·恩斯特罗姆、安妮斯·阿萨莱和鞠波·林。查询和信息有限的黑盒对抗攻击。机器学习国际会议，2142–2151 页，2018 年。
- [20] 安德鲁·易勒雅斯、洛根·恩斯特罗姆和亚历山大·马达里。前科:与土匪和前科犯进行暗箱对抗攻击。arXiv 预印本 arXiv:1807.07978, 2018.1, 2
- [21] 亚历克斯·克里哲夫斯基和杰弗里·辛顿。从微小图像中学习多层特征。技术报告, Citeseer, 2009。
- [22] 阿列克谢·库拉金、伊恩·古德费勒和萨米·本吉奥。物理世界中的敌对例子。学习表征国际会议(ICLR), 2017 年.1, 2, 5, 6, 7
- [23] 阿列克谢·库拉金、伊恩·古德费勒和萨米·本吉奥。大规模对抗性机器学习。学习表征国际会议(ICLR), 2017 年。
- [24] Yann LeCun, Leon Bottou, Yoshua Bengio, Patrick Haffner, et al. 基于梯度的学习在文档识别中的应用。《电气和电子工程师协会学报》, 86(11):2278–2324, 1998。
- [25] 刘燕佩、陈、宋黎明。探究可转移的对抗性例子和黑盒子攻击。学习表征国际会议(ICLR), 2017 年。
- [26] 马、王。Erfani, Sudanthi Wijewickrema, Grant Schoenebeck, Michael E. 侯乐、黎明歌和詹姆斯·贝利。利用局部内在维度刻画敌对子空间。在学习表征国际会议(ICLR)上, 2018 年。
- [27] 亚历山大·马德里、亚历山大·马克洛夫、路德维希·施密特、迪米特里斯·齐普拉斯和阿德里安·弗拉杜。对抗攻击的深度学习模式。在学习表征国际会议(ICLR)上, 2018 年.2, 5, 6, 7
- [28] 孟冬雨和陈皓。磁石:对抗敌对例子的双管齐下的防御。《2017 年美国计算机学会计算机与通信安全会议录》, 第 135-147 页.ACM, 2017。
- [29] 简·亨德里克·梅岑、蒂姆·杰奈温、沃尔克·费舍尔和巴斯蒂安·比肖夫。关于探测敌对扰动。学习表征国际会议(ICLR), 2017 年。
- [30] 赛义德·穆赫森·穆萨维·德兹布里、阿尔侯赛因·法齐和帕斯卡尔·佛罗萨德。Deepfool:一个简单而精确的方法来愚弄深层神经网络。《美国电气和电子工程师协会计算机视觉和模式识别会议记录》, 第 2574-2582 页, 2016 年.1, 2
- [31] 尼娜·纳罗德茨卡和希瓦·普拉萨德·卡西维斯瓦纳坦。深度网络的简单黑盒对抗扰动。arXiv 预印本 arXiv:1612.06299, 2016。
- [32] 特里布瓦内什·奥雷孔迪、贝特·席勒和马里奥·弗里茨。山寨网:窃取黑盒模型的功能。在2019年的美国电气和电子工程师协会计算机视觉和模式识别会议(CVPR)上。
- [33] 尼古拉斯·帕佩诺特和帕特里克·麦克丹尼尔。深度 k 近邻:走向自信、可解释、稳健的深度学习。arXiv 预印本 arXiv:1803.04765, 2018。
- [34] 尼古拉斯·帕佩诺特、帕特里克·麦克丹尼尔、伊恩·古德费勒、萨默什·贾、兹·伯凯·切利克和安娜瑟拉姆·斯瓦米。针对机器学习的实用黑盒攻击。《2017 年亚洲计算机与通信安全会议论文集》, 第 506-519 页.ACM, 2017。
- [35] 尼古拉斯·帕佩诺特、帕特里克·麦克丹尼尔、萨默什·贾、马特·弗雷德里克松、兹·伯凯·切利克和阿南瑟拉姆·斯瓦米。敌对环境中深度学习的局限性。在 2016 年 IEEE 欧洲安全和隐私研讨会(EuroPeAn & P), 第 372-387 页.IEEE, 2016。
- [36] 杰罗姆·罗尼、路易斯·哈菲曼、路易斯·奥利维拉、伊斯梅尔·贝亚德、罗伯特·萨博林和埃里克·格兰杰。有效的基于梯度的 l2 对抗攻击和防御的解耦方向和规范。arXiv 预印本 arXiv:1811.09600, 2018。
- [37] 卡伦·西蒙扬和安德鲁·齐泽曼。用于大规模图像识别的超深卷积网络。在学习表征国际会议(ICLR)上, 2015 年。
- [38] 克里斯蒂安·塞格迪、沃伊切赫·扎伦巴、伊利亚·萨特基弗、琼·布鲁纳、杜米特鲁·埃尔汉、伊恩·古德费勒和罗布·弗格斯。神经网络有趣的特性。学习表征国际会议(ICLR), 2014 年.1, 2
- [39] 弗洛里安·特拉梅尔、张帆、阿里·朱尔斯、迈克尔·赖特和托马斯·里斯坦帕尔。通过预测 API 窃取机器学习模型。第 25 届{USENIX}安全研讨会({USENIX}安全 16), 第 601-618 页, 2016 年。
- [40] 弗洛里安·特拉梅尔、阿列克谢·库拉金、尼古拉斯·帕佩诺特、伊恩·古德费勒、丹·博纳和帕特里克·麦克丹尼尔。整体对抗训练:攻击和防御。在学习表征国际会议(ICLR)上, 2018 年。
- [41] 谢慈航、张志帅、周仁。通过随机化减轻对抗效应。在学习表征国际会议(ICLR)上, 2018 年。