# 摘要

我们建议使用降维来防御针对 ML 分类器的规避攻击。我们研究了一种通过主成分分析来降维的策略，以增强机器学习的适应能力，既可以应用于分类又可应用于训练阶段。我们使用多个真实世界的数据集证明了数据降维在防御逃避攻击方面的可行性。我们的主要研究结果是：（1）有效对抗文献中的策略性的逃避攻击，将对方成功攻击所需的资源增加约 2 倍，（2）适用于一系列 ML 分类器包括支持向量机和深度神经网络，（3）可推广到多个应用领域，包括图像分类和人类活动分类。


# 1 介绍

我们生活在一个到处充满着机器学习（ML）和人工智能的时代。机器学习被用于诸如图像识别，自然语言处理，垃圾邮件检测，车辆自动驾驶甚至恶意软件检测等多种基础应用中。

此外，最近在深度学习方面取得的进展表明分类的准确性可以接近于人类操作的准确性，这使得 ML 系统的广泛应用成为可能。鉴于 ML 应用程序的无处不在，它越来越多地应用于敌对情景中，在这种情况下，攻击者可以从 ML 系统的失败中对输入进行正确的分类。那么问题就出现了：ML 系统在对抗环境中安全吗?

对抗性机器学习：从 21 世纪初开始，已经有大量工作将机器学习算法的脆弱性暴露给战略对手。例如，中毒攻击在训练阶段系统地引入敌对数据，从而在测试阶段导致数据分类错误。另一方面，规避攻击的目的是通过向测试数据中添加策略性的干扰数据来欺骗现有的 ML 分类器。

规避攻击：在本文中，我们重点关注规避攻击，其中攻击者的目标是干扰 ML 分类器的测试输入以引起错误分类。针对各种机器学习分类器都提出过规避攻击，如支持向量机，基于树的分类器，随机森林和增强树，以及最近的神经网络。使用机器学习的应用程序（例如人脸检测，语音命令识别和 PDF 恶意软件检测）的脆弱性也已得到证明，这也突出了防御的必

要性。令人惊讶的是，这也表明，敌对方修改后的数据（针对特定分类器）的规避属性持续存在于不同的 ML 分类器中，这使得即使对 ML 系统了解很有限的对手都可以攻击它。因此，在敌对情境下使用 ML 系统时考虑敌对数据和躲避攻击的可能性至关重要。然而，针对这些攻击的防御措施极少，并且每种攻击的适用性仅限于某些已知的攻击和特定类型的 ML 分类器（请参见第 7 节获得详细描述）。

## 1.1 贡献

通过广泛的评估，我们发现我们的防御机制明显降低了逃避攻击的成功率。就我们所知，这是针对具有以下属性的规避攻击的唯一防御措施：（1）适用于多个 ML 分类器（如 SVM，DNN），（2）适用于多个应用领域（图像和活动分类），（3）减轻多种攻击类型，包括战略攻击类型。此外，我们的防御可调性允许系统设计人员根据应用选择公共安防权衡曲线上适当的操作点。

### 1.1.1 防御

在本文中，我们提出使用数据的降维来防御针对 ML 系统的规避攻击。降维技术（如主成分分析）旨在将高维数据投影到较低维度的空间，同时满足特定的条件。我们研究了一种降维的策略，以增强机器学习的适应能力，既可以应用于分类又可应用于训练阶段。我们考虑一种方法，将降维应用于训练数据和测试数据，以增强训练分类器的可靠性。

### 1.1.2 实证评估

我们证明了我们的防御措施的可行性和有效性：

- 多重分类器，例如支持向量机（SVM）和深度神经网络（DNN）
- 几种不同类型的规避攻击，例如 Moosavi-Dezfooli 等人对线性 SVMs 的攻击、Goodfel-low 等人的深层神经网络攻击以及针对我们的防御

措施的策略性攻击

- 各种现实世界的数据集/应用程序：MNIST 图像数据集和 UCI 人类活动识别（HAR）数据集。

我们的主要发现是，即使面对一个几乎完全了解 ML 系统的强大对手，(1) 我们的防御措施使得成功攻击所需的修改程度有着高达 5 倍的显著提高，同样的，以固定的修改程度攻击的成功率降低约 2-50 倍，(2)防御措施可以用于不同的 ML 分类器，对原始分类器进行最小限度的修改，同时仍然有效地防御攻击，（3）在大多数情况下良性样品的分类成功率有约 1-4 % 的适度变化。我们还提供了公共安防权衡曲线的分析以及我们的防御措施产生的计算开销。我们的结果开源在https://github.com/inspire-group/ml_defense上。

然而，我们的防御措施并没有完全解决规避攻击的问题，因为它可以降低固定预算下的敌对成功率，但这并不是在所有情况下都忽略不计。在第 4 节中，我们讨论了对手在不同应用场景下可用的预算范围，并明确了防御有效的场景。我们希望我们的工作能够激发进一步的研究，以解决规避攻击来保证机器学习的系统的安全性。

本文的其余部分安排如下：首先，在第 2 节中，我们介绍了对抗机器学习的必要背景。然后，在第 3 节中，我们描述了我们的防守措施。接下来，我们分别在第 4 节和第 5 节中设置并提出我们的实证评估。我们在第 6 节讨论我们的结果。最后，我们在第 7 节中详细介绍相关工作，并在第 8 节中做出结论。

## 2 对抗性机器学习

在本节中，我们提出了对抗性机器学习所需的背景，重点关注（a）ML 分类器，如 SVM 和 DNN，以及（b）通过干扰测试输入引发错误分类的规避攻击。

动机和运行示例：我们的运行示例使用来自 MNIST 数据集的图像数

据（详见第 4 节）。图 1（a）描绘了来自 MNIST 数据集的正确测试图像，这些图像被 SVM 分类器正确分类；而图 1（b）描绘了对手制作的测试图像（使用 Papernot 的规避攻击的扰动图像），它们被 SVM 分类器错误分类。

(a) Typical test images from the MNIST dataset. Correctly classifed as 0, 4, 5 and 6 respectively.

(b) Corresponding adversarial images obtained using the evasion attack on Linear SVMs [29]. Now, **misclassified as 9, 9, 3, 2 and 0.** respectively.

Figure 1: **Comparison of benign and adversarial images taken from the MNIST dataset.**

## 2.1 使用机器学习分类

在本文中，我们关注有监督的机器学习，其中分类器通过预先存在的标签对数据进行训练。一个训练完成的监督机器学习分类器是一个函数，通过输入点 $\boldsymbol{x} \in \mathbb{R}^d$（二进制时为 $\{0,1\}d$），会输出 $\hat{y} \in C$，其中 C 是所有可能分类的集合。例如，在 MNIST 数据集的情况下，$\boldsymbol{x}$ 将是 28×28 像素的手写数字的灰度图像，而 C 将是有限集合 {0,1,2,3,4,5,6,7,8,9}。

## 2.2 攻击机器学习系统

在本小节中，我们首先讨论对抗模型，之后我们会讲述一般的规避攻击，最后讲述对特定的 ML 分类器的规避攻击。

注: 我们把完整的训练集表示为 $S_{train}$，完整的训练数据表示为 $S_{test}$，将 ML 分类器表示为 $f$，并且针对 ML 分类器的特定参数表示为 $\boldsymbol{\theta}$。数据的原始维度表示为 $d$。接下来，我们把攻击者的攻击算法表示为 $A(\boldsymbol{x}_{in}|K)$，其中，$\boldsymbol{x}_{in}$ 表示对手开始时的输入，$K$ 代表对手的已知信息，可能是 $\{S_{train}, f, \boldsymbol{\theta}\}$ 的任一子集。$\widetilde{\boldsymbol{x}}$ 表示 A 生成的敌对样本。

### 2.2.1 敌对方的目标和能力

在本文中，我们关注的情景是，攻击者的目标是通过修改一个正确的输入，以便使它被误分为其他的任何分类，或者使其被归类为与原始类不同的目标分类。请注意，这些目标分类在二元分类器的情况下是等价的。

我们的基本假设是对手具有以下能力。

- 对手完全了解原始分类器已经训练过的训练集，即她知道分类器作为输入所采用的特征向量的类型。
- 对手知道分类器结构，超参数和训练过程。
- 错误数据是由对手离线创建的，在测试阶段提交给 ML 分类器。

简而言之，$\widetilde{\boldsymbol{x}} = A(\boldsymbol{x}_{in}|S_{train}, f, \boldsymbol{\theta}, K_{add})$，其中 $K_{add}$ 表示关于对手可能拥有的系统的任何其他知识。

我们对对手的能力的假设是保守的，因为从安全角度来讲，系统在完全了解系统安全的对手的强力的攻击下，依旧是健壮的。而且，一个有着 ML 系统知识的攻击者，即使有着有限权限的访问（如黑盒访问），也可以很好的对分类器进行推断来进行规避攻击。这和一个拥有完全访问权的对手攻击的效果集合一样，这证明了我们的假设是合理的。

### 2.2.2 规避攻击

在正常操作，即没有攻击者时，当输入 $\boldsymbol{x}_i \in S$，$f$ 会输出 $\hat{y}$，其中 $S$ 是输入集合。输出的分类中正确匹配的比例为 $\alpha$，即，

$$\alpha(S) = \frac{\#\{(\boldsymbol{x}, y) \in S : f(\boldsymbol{x}) = y\}}{\#S} \tag{1}$$

其中　给出了一组的基数。攻击者的目标是设计一个作用在 $x \in S$ 上的算法 A 来生成敌对数据，即，$A(\boldsymbol{x}) = \widetilde{\boldsymbol{x}}$，令

$$S^{adv} = \{(A(\boldsymbol{x}), y) : (\boldsymbol{x}, y) \in S\}$$

这是一组对比修改的例子，其中修改之处应满足：

- 与分类器的正常操作相比，增加错误分类的占比，即 $\alpha(S^{adv}) < \alpha(S)$，
- 在诸如图像和文本等人类可解释的数据的情况下，不被人类察觉到异常；在诸如恶意软件样本，网络和系统日志等数据的情况下，可被基于规则的检测系统通过。例如：在恶意软件的情况下，攻击者受到这样的限制，即她的修改必须确保最终的样本仍然是恶意的。

  我们接下来讨论敌对干扰，以及在图片数据的情况下，他们对人类的感知力。

### 2.2.3 敌对干扰

模拟人类对图像扰动的感知是一个难题。作为人类可感知性的代理，我们将对某个范数 $||\cdot||$ 的修正程度定义为 $||A(\boldsymbol{x}) - \boldsymbol{x}||$。需要强调的是，我们将考虑受 $\ell_2$ 约束的限制，即 $||\widetilde{\boldsymbol{x}} - \boldsymbol{x}||_2 \leqslant \xi$，其中 $\xi$ 决定了干扰的强度。[35] 中给出了用于约束敌对干扰的各种规范与其感知之间关系的详细描述。

现在，我们定义了实现不同对抗目标所需的最小扰动。为了在特定的类 $z$ 中导致错误分类，必须添加一个输入数据 $(\boldsymbol{x}, y)$ 作为最小的扰动，其中 $z \neq y$，

$$\Delta(\boldsymbol{x}, z) = \inf_{\widetilde{\boldsymbol{x}}} \{||\widetilde{\boldsymbol{x}} - \boldsymbol{x}|| : f(\widetilde{\boldsymbol{x}} = z)\}$$

这是导致 $\boldsymbol{x}$ 被归类为 $z$ 所需的最小失真。导致 $\boldsymbol{x}$ 在任何类中被错误分类所需的最小失真是，

$$\Delta(x) = \min_{z \in C \setminus \{y\}} \Delta(\boldsymbol{x}, z)$$

对于图像数据，这些量与最小可检测失真之间的关系决定了分类器 $f$ 对敌对扰动的鲁棒性。在图 1 中，图像中的干扰值导致线性 SVM 几乎将所有输入都错误错误，但干扰对于人眼几乎不可见。这表明线性标准形式的 SVM 对抗扰动是不稳健的。在第 4.4 节中进一步讨论了用于约束对手的指标。

## 2.3 针对特定分类器的规避攻击

我们现在描述现有文献记载的针对特定 ML 分类器的攻击，并展示来自 MNIST 数据集的一些对抗性例子。表 1 给出了各种攻击的总结。

### 2.3.1 对线性 SVM 的最佳攻击

在线性支持向量机的多类分类设置中，分类器 $g_i$ 针对每个类别 $i \in C$ 进行训练，其中

$$g_i : \boldsymbol{x} \mapsto \boldsymbol{w}_i^T \boldsymbol{x} + b_i \tag{2}$$

$\boldsymbol{x}$ 被分配给类 $f(\boldsymbol{x}) = \arg\max_{i \in C} g_i(\boldsymbol{x})$。假定真正的类别是 $t \in C$，攻击的目标是找到最接近的点 $\widetilde{\boldsymbol{x}}$，使得 $f(\widetilde{\boldsymbol{x}}) \neq t$。

从 [29] 我们知道，对于多类分类器的最优无目标的攻击，即如果我们只关心 $f(\widetilde{\boldsymbol{x}})$ 使得 $||\widetilde{\boldsymbol{x}} - \boldsymbol{x}||$ 尽可能小，令 $\widetilde{\boldsymbol{x}}$ 的最优选择是 $\widetilde{\boldsymbol{x}}_{\boldsymbol{k}}$，则，

$$k = \arg\min_j \frac{g_t(\boldsymbol{x}) - g_j(\boldsymbol{x})}{||\boldsymbol{w}_t - \boldsymbol{w}_j||} \tag{3}$$

进而得到，

$$\widetilde{\boldsymbol{x}}(\xi) = \boldsymbol{x} + \xi \frac{\boldsymbol{w}_t - \boldsymbol{w}_k}{||\boldsymbol{w}_t - \boldsymbol{w}_k||} \tag{4}$$

这里的 $\xi$ 代表干扰的程度。导致误分类的 $\xi$ 的最小值是 $\xi^* = \frac{|g_t(\boldsymbol{x}-g_k(\boldsymbol{x}))|}{||\boldsymbol{w}_t-\boldsymbol{w}_k||}$。很容易可以证明 $f(\widetilde{\boldsymbol{x}}(\xi^*)) = k$。请注意，由于 $||\xi\frac{\boldsymbol{w}_t-\boldsymbol{w}_k}{||\boldsymbol{w}_t-\boldsymbol{w}_k||}|| = \xi$，这个攻击受到 $\ell_2$ 约束的限制。

### 2.3.2 基于梯度的神经网络攻击

FGS 攻击是 [19] 中引入的针对神经网络的高效攻击。在这种情况下，通过添加与损失函数的梯度（$\nabla J_f(\boldsymbol{x}, y, \theta)$）成正比的对立噪声来生成对抗示例，其中 $J_f(\cdot)$ 表示损失函数，$\theta$ 表示用于训练的超参数。可以使用反向传播有效地计算梯度。具体为，

$$\widetilde{\boldsymbol{x}} = \boldsymbol{x} + \eta sign(\nabla J_f(\boldsymbol{x}, y, \theta)) \tag{5}$$

其中 $\eta$ 是对手可以改变的参数，以控制对抗性例子的有效性。随着 $\eta$ 的增加，攻击的成功率一般也在增长。然而，较大的干扰可能会使人们难以辨识图像（请参阅附录中的图像，其变化范围为 $\eta$）。FGS 攻击者受到 $\ell_2$ 约束的限制，因为 $||\eta sign(\nabla J_f(\boldsymbol{x}, y, \theta))||_\infty = \max_i |\eta sign(\nabla J_f(\boldsymbol{x}, y, \theta))_i| = \eta$，这控制着干扰的程度。

FGS 攻击和对线性 SVM 的攻击根据不同的标准受到限制。为了便于比较各种分类器的鲁棒性以及我们对它们的防御效果，我们提出了一种修改 FGS 攻击的方法，该攻击受到 $\ell_2$ 约束的限制。我们将这称为快速梯度（FG）攻击，我们将敌对示例定义为，

$$\widetilde{\boldsymbol{x}} = \boldsymbol{x} + \eta\frac{\nabla J_f(\boldsymbol{x}, y, \theta)}{||\nabla J_f(\boldsymbol{x}, y, \theta)||} \tag{6}$$

对于 FG 攻击而言，$\eta$ 是对干扰标准 $\ell_2$ 的约束。

## 2.4 降低机器学习的维度

在处理高维数据（意味着每个样本具有大量特征）的同时，很难弄清哪些特征很重要。应用程序约束也可能会使原始高维空间中的数据执行学习任务变得不切实际。因此降维是对高维数据有效的预处理步骤。它还可以

表 1: 对线性 SVM 和神经网络的攻击总结

| 攻击 | 分类器 | 约束 | 直观结果 |
|---|---|---|---|
| 线性 SVM 最优攻击 | 线性 SVM | $\ell_2$ | 趋向分类器边界 |
| 快速梯度 | 神经网络 | $\ell_2$ | 最小扰动方向的一阶近似 |
| 快速渐变标志 | 神经网络 | $\ell_\infty$ | 建议不断缩放每个像素模型 |

帮助解决与“维度诅咒”相关的问题。在这种情况下，数据首先投影到较低维空间，然后将其作为 ML 系统的输入。常见的降维算法是 PCA [27]，随机投影 [36] 和核 PCA [37]。在本文中，我们使用 PCA 等降维方法不仅有助于解释性和提高效率，而且还可以提高 ML 系统对敌对实例的鲁棒性。

# 3 基于降维的防御

# 4 进行实验

# 5 实验结果

在本节中，我们将概述我们的实验结果。我们试图回答的主要问题有：

i) 我们的防御对策略性攻击是否有效?

ii) 我们的防御能够对抗 vanilla 攻击?

iii) 我们的防御能否作用在不同分类器上?

iv) 我们的防御是否推广到不同的数据集?

我们的评估结果证实了我们的防御在各种场景下的有效性，每种场景都有数据集、机器学习算法、攻击和降维算法的不同组合。对于每一组评估，我

们改变分类管道的特定步骤并修复其他步骤。基本配置：我们首先考虑将 MNIST 数据集作为输入数据的分类流水线，将线性 SVM 作为我们的分类算法，将 PCA 作为我们的防御中使用的降维算法。由于我们将线性 SVM 作为分类器，因此我们评估其对使用 2.2 节中描述的线性 SVM 攻击生成的敌对样本的敏感性。下面我们对每个数据集来评估我们针对从测试集开始创建的对抗样本的防御。除非另有说明，否则所有防御结果均适用于完整的测试集。为了证明我们的防御不仅在这个基线和各种配置下是鲁棒的，所以我们系统地研究了它的影响，因为管道的每个组成部分以及攻击都被改变了。

（a）数字'9' 的良性和干扰图像（针对不具有防御的线性 SVM）：左边的第一张图像是原始图像，而其他图像是利用线性 SVM 的攻击（从左到右）修改的，$\xi = 0.5, 1.0, 1.5, 2.0$。干扰爬在 $\xi = 1.5$ 时开始可见，在 $\xi = 1.5$ 的图像中非常明显。攻击是在没有任何降维的分类器 $f$ 上进行的。

（b）数字'7' 的干扰图像（针对没有防御的神经网络）：图像通过对神经网络的快速梯度攻击（从左到右）进行修改，$\eta \approx 0.5, 1.0, 1.5, 2.0, 2.5$。在 $\eta = 1.5$ 时，再次开始可见，在 $\eta > 2.0$ 的图像中非常明显。攻击是在没有任何降维的分类器上进行的。

（c）数字'7' 的干扰图像（针对具有 $k = 70$ 的基于 PCA 的防御的神经网络）：图像已经通过神经网络上的快速梯度攻击进行修改，采用降维输入（从左到右），$\eta \approx 0.5, 1.0, 1.5, 2.0, 2.5$。将降维矢量投影回图像空间进行可视化。在这种情况下，干扰在 $\eta = 0.5$ 时开始可见，并且在 $\eta > 1.5$ 的图像中非常明显。这表明我们的防守也会使敌对扰动更容易被察觉。

图 3：为避开线性 SVM 和神经网络而生成的敌对图像

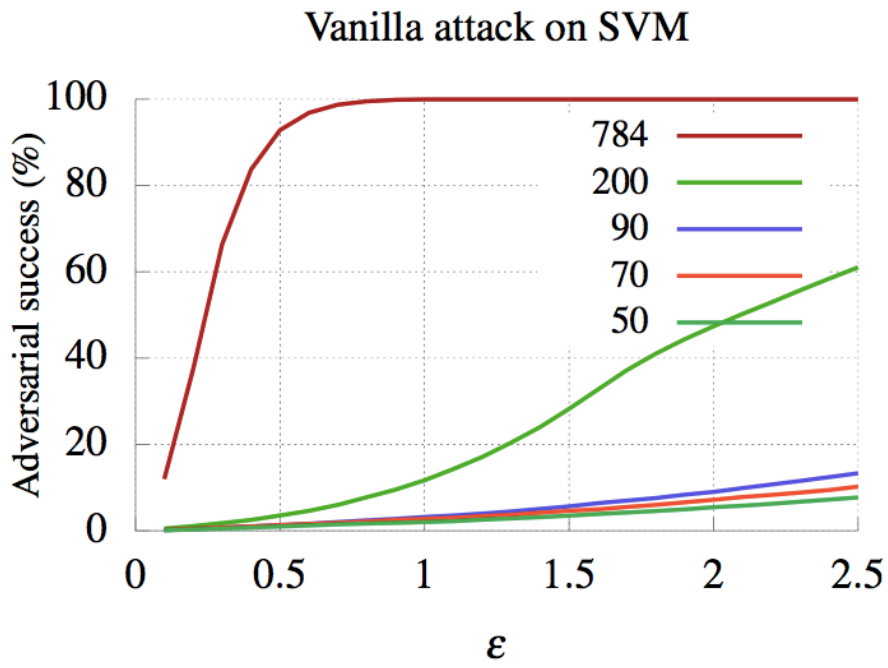


图 4：针对 MNIST 数据集的防御对线性 SVM 的 vanilla 攻击的有效性。MNIST 数据集上的敌对示例成功率与干扰程度 $\xi = ||\widetilde{\boldsymbol{x}} - \boldsymbol{x}||$ 的关系图。针对原始分类器进行攻击，并针对每个减小的维度 k 绘制防御的效果。

## 5.1 防御对支持向量机的影响

在标准情况下，我们首先回答问题 ii），即"防御能否降低 vanilla 攻击的有效性?"和 i），即针对线性 SVMs 的"防御能否降低策略攻击的有效性?"。

### 5.1.1 防御 vanilla 攻击

图 4 显示了成功防御对 SVMs 的防御攻击的变化。防御大大降低了敌对的成功率。例如，在 $\xi = 1.0$ 时，使用 $k = 50$ 的 PCA 的防御方法，使对手的成功率从 99.97% 降低到 1.85%。在 $\xi = 0.5$ 的情况下，敌对的成功率是 92.77% ,$k = 50$ 的防御将敌对成功率降低到 0.9% 这是两个数量级的下降。使用降维的数据训练会使得线性 SVMs 更健壮，这可以从防御的附加效果中看到。

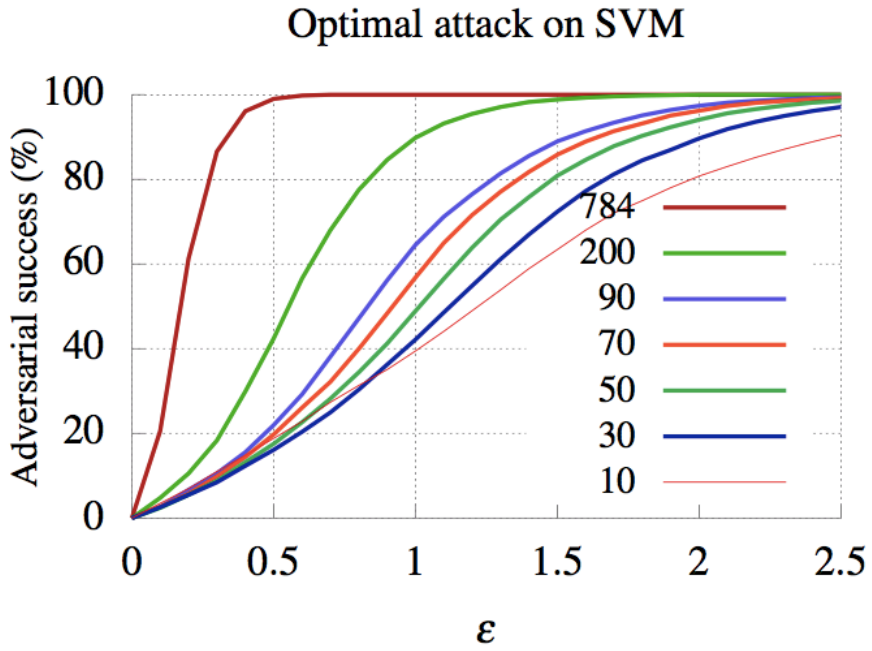


图 5：针对 MNIST 数据集的防御对线性 SVM 的最优攻击的有效性。将 MNIST 数据集上的干扰示例成功率与干扰幅度 $\xi = ||\widetilde{\boldsymbol{x}} - \boldsymbol{x}||$ 作图。针对每个降维分类器执行攻击并绘制防御效果。

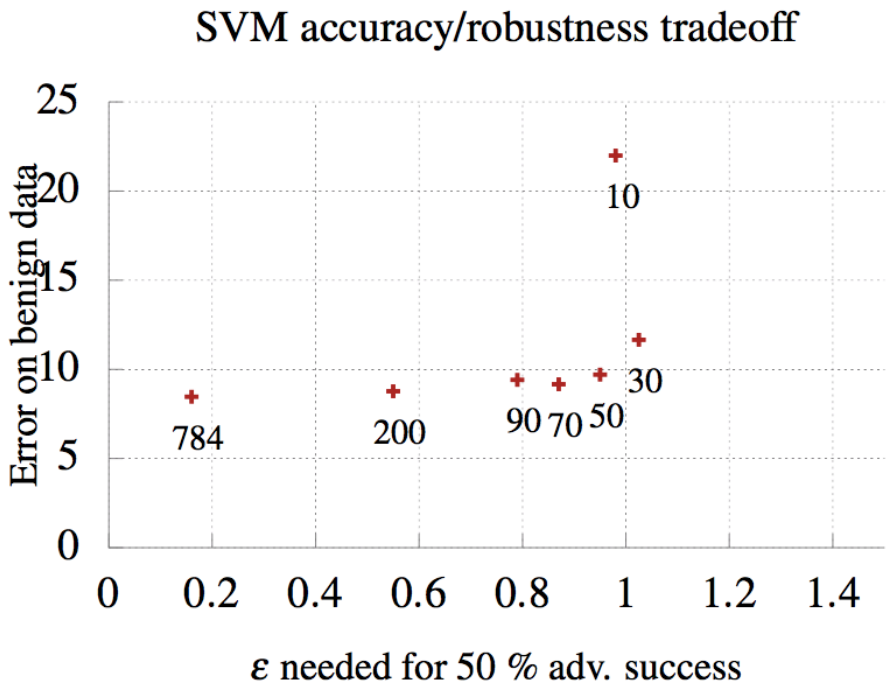


图 6：在良性测试数据和敌对性能之间的 SVM 分类性能之间的权衡。对抗性的健壮性是 $||\widetilde{\boldsymbol{x}} - \boldsymbol{x}||$ 的值，它允许对手达到 50% 的错误率。

同样，我们也注意到当我们减少在防御的投影步骤中使用的减少的维度 $k$ 时，对抗的成功率也在降低。在 $\xi = 1.0, k = 331$ 时，对抗成功率是 48.75%，当 $k = 100$ 时下降到 5.53%。在 $k = 30$ 时，对抗性的成功率下降到 2.63%，在 $k = 10$ 时减少到 2.52%。

在 vanilla 攻击下，防御就像一个噪音移除过程，消除了敌对的干扰并留下了干净的输入数据。与策略攻击相比，我们看到的是防御的鲁棒性。

### 5.1.2 防御对最佳攻击的影响

图 5 显示了针对线性 SVMs 的最优策略攻击的防御成功的变化。这个图对应的是对手意识到维度减少防御并将样本输入到管道中的情景，它的设计是为了最有效地避开降维分类器。在 0.5 的扰动程度下，没有防御的

分类器的分类错误率是 99.04%，$k = 70$ 的降维分类器的分类错误率只有 19.75%，即攻击成功率分别为 80.25% 和 5.01%。然而，由于 0.5 是一个小的干扰参数，即使当缩小的维度图像被投射回像素空间时，微扰也将是不可见的。在 1.3 的干扰参数中，开始清晰可见（见第 4.4 节），没有防御的分类器的错误分类率是 100%，大概是 77.11% 的分类错误率对于的 $k = 70$ 的降维分类器，几乎是降低了 23% 的攻击成功率。回想一下，避开低维度分类器所需要的扰动对人眼来说更清晰可见，使这些数字变得保守。

我们也可以研究我们的防御对达到一定的敌对成功率所需要的对抗预算的效果。为了达到 86.6%，需要一个 0.3 的预算，在没有防御的情况下进行分类，而对于一个 $k = 70$ 的分类器的所需预算是 1.6。对应的数字达到 90% 误分类率的 $k = 0.4$。因此，我们的防御很明显的降低一个非常强大的对手所进行的攻击的有效程度，它完全了解防御和分类器，并拥有最有效攻击的能力。

### 5.1.3 对防御的效用-安全权衡

图 6 显示了在普通和敌对条件下的性能之间的权衡。这个数据集的最佳维数显然在 50 到 30 之间，其中的扭结发生在这里。通过使用更多的维度，在分类性能方面几乎没有什么好处，而且使用更少的性能对健壮性没有任何好处。在 k=50 时，我们看到没有任何防御时测试集上的分类成功率下降了 91.5%, 而在防御下为 90.29%，因此，大约有 1.2% 的效用。

有了这些结果，我们就可以得出结论，我们的防御至少在基线情况下是有效的，对于线性 SVMs 的普通和最优攻击都是有效的。现在，为了证实我们关于我们的防御在机器学习分类器中的适用性的主张，我们研究了我们在神经网络上的防御表现。

## 5.2 防御对神经网络的影响

## 5.3 对不同数据集的适用性

接下来，我们通过更改所使用的数据集来修改基线配置。我们用线性 SVMs 作为分类器和 PCA 作为维度还原算法来显示结果。我们为人类活动识别数据集提供结果。

### 5.3.1 对 HAR 数据集的保护

在图 9 中，显示了由于防御而导致的攻击成功率降低。在 $\xi = 1.0$ 时，防御成功率从没有防御时的 99.56% 下降到 91.75%（$k = 70$）和 76.21%（$k = 30$）。为了达到分类错误率 90%，没有防御时需要的干扰程度是 0.65，在 $k = 70$ 时它增加到 $0 \cdot 876$，在 $k = 30$ 时为 1.26。因此，攻击方的成本增加了两倍以达到同样的成功率。这对效用的影响是适用的，在 $k = 70$ 时为下降了 2.3%，在 $k = 30$ 时为 5.4%，与安全获得的收益相比，这是微不足道的。

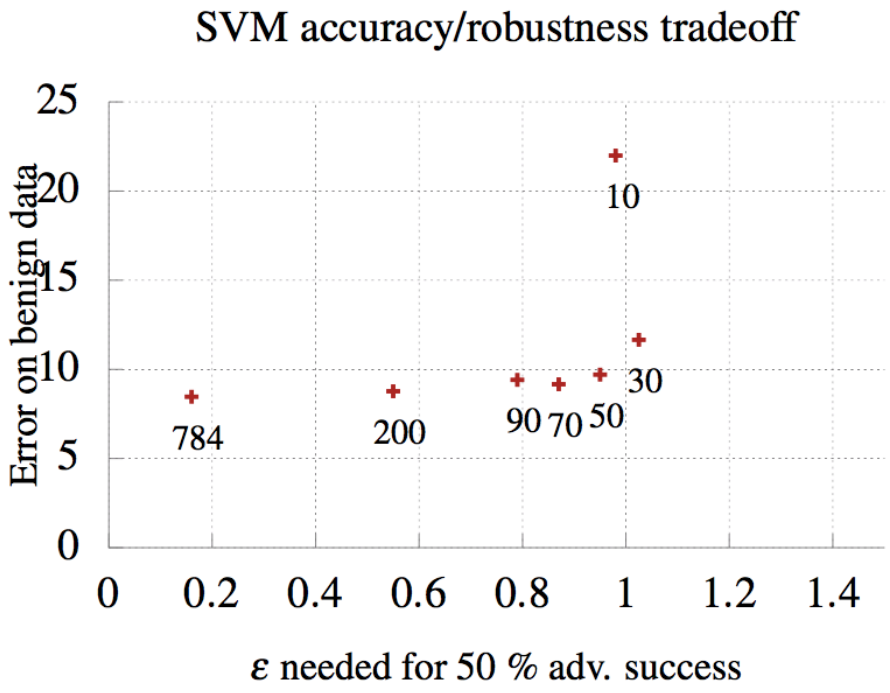


图 9：对于线性 SVM 攻击（针对原始分类器），HAR 数据集上的敌对示例与摄动幅度 e。针对防御中使用的每个减少的维度 k 绘制。

| | MNIST data | | | HAR data |
|---|---|---|---|---|
| | FC100-100-10 | Linear SVM | | Linear SVM |
| $k$ (MNIST) | | | $k$ (HAR) | |
| No D.R. | 97.47 | 91.52 | No D.R. | 96.67 |
| 784 | 97.32 | 91.54 | 561 | 96.57 |
| 331 | 97.35 | 91.37 | 200 | 96.61 |
| 200 | 97.04 | 91.28 | 100 | 92.43 |
| 100 | 97.36 | 90.89 | 90 | 94.60 |
| 90 | 97.14 | 90.58 | 80 | 94.54 |
| 80 | 97.25 | 90.64 | 70 | 94.37 |
| 70 | 97.52 | 90.76 | 60 | 93.72 |
| 60 | 97.38 | 90.47 | 50 | 92.47 |
| 50 | 97.26 | 90.18 | 40 | 92.06 |
| 40 | 96.71 | 89.03 | 30 | 91.11 |
| 30 | 96.56 | 88.37 | 20 | 88.63 |
| 20 | 96.67 | 86.69 | 10 | 86.67 |
| 10 | 93.22 | 77.79 | | |

表 2：降维防御的效用值。对于 MNIST 和 HAR 数据集，良性测试集的分类准确性针对用于基于 PCA 的防御的降维 k 的各种值以及没有防御的准确性提供。

## 5.4 对效用的影响

表 2 显示了我们的防御对良性数据的分类精度的影响。关键的结论是，神经网络和线性 SVMs 的精度降低到 k=50 的程度是最多 4%，此外，我们注意到，使用 PCA 的维数减少实际上可以提高分类精度，当 $k = 70$ 时，MNIST 数据集的准确性从 97.47% 降低到 97.52%，然而，更重要的维度减少，这将导致分类精度的急剧下降，这是意料之中的，因为用于分类的大部分信息都丢失，这些结果突出了我们在应用领域的防御的广泛适用性。很明显，我们防御的有效性并不是来自于 MNIST 数据集的特定结构的产物，他们对不同的数据都有影响。

# 6 讨论和限制

## 6.1 满足设计目标

在第 3.2 节中，我们列出了任何辩护都应该具备的理想目标。首先，防御应该保持较高的分类精度。从表 2 中可以看出，对于数据集和分类器来说，有一系列缩小的维度对分类精度影响最小，在一些特定的情况下可以证明这一点。其次，基于 PCA 的防御版增加了样本数量 n 和维度 d 的多项式级别的代价。训练降维分类器所需的时间和空间高于高维空间中的分类器，因此，我们的防御在训练和测试阶段保持高效率。我们的辩护所带来的附加安全已经在前一节的各种设置中得到了说明。从图 6 中可以清楚地看出，改变尺寸允许 ML 系统所有者在实用安全空间中导航不同的点。然而，当一个系统可能受到攻击时，我们的防御系统并不有效，这将导致我们在下面讨论的限制。

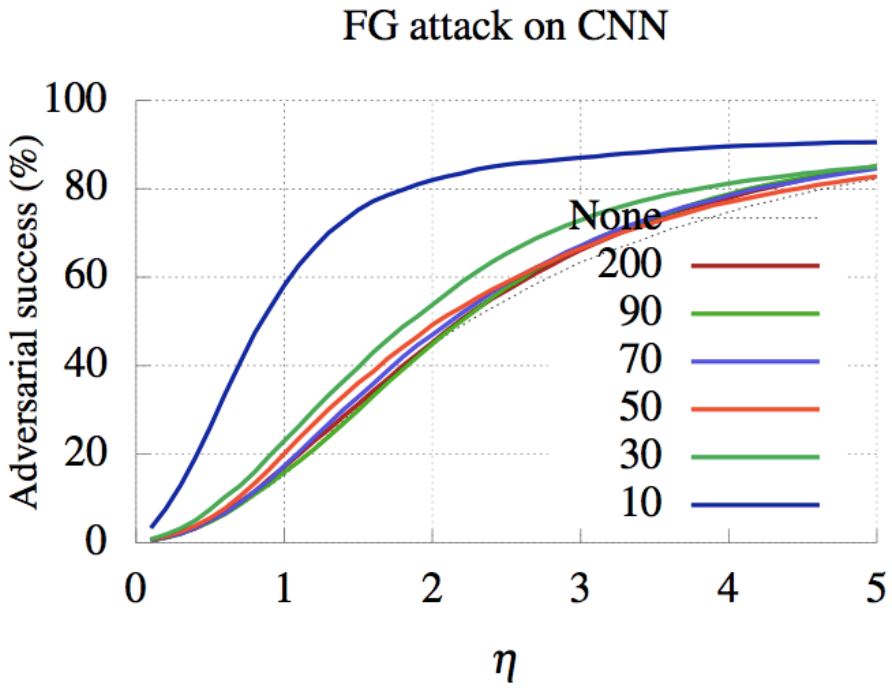


图 10：MNIST 数据集的防御效果与针对 Papernot-CNN 的战略性 FG 攻

击相抵触。在 MNIST 数据集上的对抗成功示例是相对于扰动幅度 $\eta = ||\widetilde{\boldsymbol{x}} - \boldsymbol{x}||$ 绘制的。针对每个子分类器（从算法 2 获得）进行攻击并绘制防御效果。

## 6.2 限制

尽管我们的防御在许多情况下降低了对抗性的成功率，但有两个主要的方面，它没有成为一种针对逃避攻击的全面防御机制：

1 在自己的不足：虽然我们的防守在各种情况下都能显著降低对手的成功率，但在某些情况下，对手的成功率仍然太高。在这种情况下，我们的防御系统很可能会被合并有其他的防御措施，如对抗训练 [19] 和整体方法 [43] 以建立一个针对逃避攻击的 ML 系统。我们的防御有一个优势，它可以与各种各样的 ML 分类器一起使用，它不会干扰其他防御机制的操作。此外，正如在第四部分示范的那样，我们的辩护导致了一种具有更大视觉感知能力的“意即性”的混乱。这可能有助于防御的防御，目的是探测敌对的扰动。

2 缺乏普遍性: 在某些情况下，我们的防御能力有限。例如, 在图 10 中, 我们看到，基于 PCA 的防御系统几乎没有为 Papernot-CNN 提供安全改进（详情见第 9.3 条）。这一效应很可能源于这样一个事实，即 CNNs 已经在其卷积层中已经处于企业领域特定的知识，另外，使用 PCA 进行预处理的附加层不会带来任何额外的健壮性。此外，PCA 可能会减少 CNN 的卷积层用于分类目的的本地信息的数量。

解决我们防御的局限性的一个关键步骤是使用其他维度减少技术这可以将敌对的成功降低到可以忽略的水平，并与诸如 CNNs 这样的分类器结合在一起。在未来的工作中，我们计划探索减少维度的技术，例如自动编码器，内核 PCA 和各种压缩方案，以更好地理解维度减少与分类器的鲁棒性之间的关系。

# 7 相关工作

# 8 结论

# 致谢

我们要感谢 Chawin Sitawarin 在实验和讨论方面提供帮助。Arjun Nitin Bhagoji 由 NSF 和 DARPA 提供支持。Daniel Cullina 由 DARPA 支持。

# 参考

1 Y. LeCun, Y. Bengio, and G. Hinton, "Deep learn- ing,"Nature, vol. 521, no. 7553, pp. 436–444, 2015.

2 R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa, "Natural language processing (almost) from scratch,"Journal of Ma- chine Learning Research, vol. 12, no. Aug, pp. 2493–2537, 2011.

3 G. V. Cormack, "Email spam filtering: A system- atic review,"Foundations and Trends in Informa- tion Retrieval, vol. 1, no. 4, pp. 335–455, 2007.

4 D. Cires An, U. Meier, J. Masci, and J. Schmidhu- ber, "Multi-column deep neural network for traffic sign classification,"Neural Networks, vol. 32, pp. 333–338, 2012.

5 NVIDIA, "Self driving vehicles development plat- form."

6 N. S˘rndic´ and P. Laskov, "Hidost: a static machine-learning-based detector of malicious files,"EURASIP Journal on Information Security, vol. 2016, no. 1, p. 22, 2016.

7 G. E. Dahl, J. W. Stokes, L. Deng, and D. Yu, “Large-scale malware classification using random projections and neural networks,”in 2013 IEEE In- ternational Conference on Acoustics, Speech and Signal Processing. IEEE, 2013, pp. 3422–3426.

8 A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Im- agenet classification with deep convolutional neural networks,”in Advances in neural information pro- cessing systems, 2012, pp. 1097–1105.

9 Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, “Deepface: Closing the gap to human-level perfor- mance in face verification,”in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 1701–1708.

10 L. Huang, A. D. Joseph, B. Nelson, B. I. Rubin- stein, and J. Tygar, “Adversarial machine learning,”in Proceedings of the 4th ACM workshop on Secu- rity and Artificial Intelligence. ACM, 2011, pp. 43–58.

11 M. Barreno, B. Nelson, R. Sears, A. D. Joseph, and J. D. Tygar, “Can machine learning be secure?”in Proceedings of the 2006 ACM Symposium on In- formation, computer and communications security. ACM, 2006, pp. 16–25.

12 P. Laskov and M. Kloft, “A framework for quan- titative security analysis of machine learning,”in Proceedings of the 2nd ACM workshop on Security and artificial intelligence. ACM, 2009, pp. 1–4.

13 B.Biggio,B.Nelson,andP.Laskov,“Poisoningat- tacks against support vector machines,”in Proceed- ings of the 29th International Conference on Ma- chine Learning (ICML-12), 2012, pp. 1807–1814.

14 B. Biggio, I. Corona, D. Maiorca, B. Nelson, N. S˘rndic´, P. Laskov, G. Giacinto, and F. Roli, "Evasion attacks against machine learning at test time,"in Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Springer, 2013, pp. 387–402.

15 N. Papernot, P. McDaniel, S. Jha, M. Fredrikson, Z. B. Celik, and A. Swami, "The limitations of deep learning in adversarial settings," in 2016 IEEE Eu- ropean Symposium on Security and Privacy (EuroS&P). IEEE, 2016, pp. 372–387.

16 C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intrigu- ing properties of neural networks,"in International Conference on Learning Representations, 2014.

17 N. Papernot, P. McDaniel, and I. Goodfellow, "Transferability in machine learning: from phe- nomena to black-box attacks using adversarial sam- ples,"arXiv preprint arXiv:1605.07277, 2016. 17

18 A. Kantchelian, J. Tygar, and A. D. Joseph, "Eva- sion and hardening of tree ensemble classifiers,"in Proceedings of the 33rd International Conference on Machine Learning (ICML-16), 2016.

19 I. J. Goodfellow, J. Shlens, and C. Szegedy, "Ex- plaining and harnessing adversarial examples,"in International Conference on Learning Representa- tions, 2015.

20 A. Kurakin, I. Goodfellow, and S. Bengio, "Ad- versarial examples in the physical world,"arXiv preprint arXiv:1607.02533, 2016.

21 N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks,"in IEEE Sympo- sium on Security and Privacy, 2017.

22 A. Nguyen, J. Yosinski, and J. Clune, “Deep neu- ral networks are easily fooled: High confidence predictions for unrecognizable images,” in 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2015, pp. 427–436.

23 M. McCoyd and D. Wagner, “Spoofing 2d face de- tection: Machines see people who aren’t there,”arXiv preprint arXiv:1608.02128, 2016.

24 N. Carlini, P. Mishra, T. Vaidya, Y. Zhang, M. Sherr, C. Shields, D. Wagner, and W. Zhou, “Hidden voice commands,”in 25th USENIX Secu- rity Symposium (USENIX Security 16), Austin, TX, 2016.

25 W. Xu, Y. Qi, and D. Evans, “Automatically evad- ing classifiers,”in Proceedings of the 2016 Network and Distributed Systems Symposium, 2016.

26 P. Russu, A. Demontis, B. Biggio, G. Fumera, and F. Roli, “Secure kernel machines against evasion attacks,”in Proceedings of the 2016 ACM Workshop on Artificial Intelligence and Security, ser. AISec ’16. New York, NY, USA: ACM, 2016, pp. 59–69.

Online . Available: http://doi.acm.org/10.1145/2996758.2996771

27 J. Shlens, “A tutorial on principal component anal- ysis,”arXiv preprint arXiv:1404.1100, 2014.

28 L. Van Der Maaten, E. Postma, and J. Van den Herik, “Dimensionality reduction: a comparative review,”J Mach Learn Res, vol. 10, pp. 66–71, 2009.

29 S.-M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard, “Deepfool: a simple and accurate method to fool deep neural networks,”arXiv preprint arXiv:1511.04599, 2015.

30 Y. LeCun and C. Cortes, “The mnist database of handwritten digits,” 1998.

31 D. Anguita, A. Ghio, L. Oneto, X. Parra, and J. L. Reyes-Ortiz, “A public domain dataset for human activity recognition using smartphones.”in ESANN, 2013.

32 N. Papernot, P. McDaniel, I. Goodfellow, S. Jha, Z. Berkay Celik, and A. Swami, “Practical black- box attacks against deep learning systems using adversarial examples,”in Proceedings of the 2017 ACM Asia Conference on Computer and Commu- nications Security.

33 B. Nelson, B. I. Rubinstein, L. Huang, A. D. Joseph, S.-h. Lau, S. J. Lee, S. Rao, A. Tran, and J. D. Tygar, “Near-optimal evasion of convex- inducing classifiers.”in AISTATS, 2010, pp. 549–556.

34 B. Biggio, I. Corona, B. Nelson, B. I. Rubinstein, D. Maiorca, G. Fumera, G. Giacinto, and F. Roli, “Security evaluation of support vector machines in adversarial environments,”in Support Vector Machines Applications. Springer, 2014, pp. 105–153.

35 A. Fawzi, O. Fawzi, and P. Frossard, “Analysis of classifiers’robustness to adversarial perturbations,”arXiv preprint arXiv:1502.02590, 2015.

36 E. Bingham and H. Mannila, “Random projection in dimensionality reduction: Applications to image and text data,”in Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.

37 B. Scholkopf and A. J. Smola, Learning with Ker- nels: Support Vector Machines, Regularization, Optimization,andBeyond. Cambridge,MA,USA: MIT Press, 2001.

38 I. Goodfellow, Y. Bengio, and A. Courville, Deep learning. MIT Press, 2016.

39 D. Arp, M. Spreitzenbarth, M. Hubner, H. Gascon, and K. Rieck, "Drebin: Effective and explainable detection of android malware in your pocket."in NDSS, 2014.

40 F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Ma- chine learning in Python,"Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011. 18

41 Theano Development Team, "Theano: A Python framework for fast computation of mathematical expressions,"arXiv e-print arXiv:1605.02688.

42 S. Dieleman and J. S. et.al., "Lasagne: First release."Aug. 2015.

Online . Available: http: //dx.doi.org/10.5281/zenodo.27878

43 C. Smutz and A. Stavrou, "When a tree falls: Us- ing diversity in ensemble classifiers to identify eva- sion in malware detectors,"in 23nd Annual Net- work and Distributed System Security Symposium, NDSS 2016.

44 D. Lowd and C. Meek, "Adversarial learning,"in Proceedings of the eleventh ACM SIGKDD interna- tional conference on Knowledge discovery in data mining. ACM, 2005, pp. 641–647.

45 N.Dalvi,P.Domingos,S.Sanghai,D.Vermaetal., "Adversarial classification," in Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2004, pp. 99–108.

46 B. I. Rubinstein, B. Nelson, L. Huang, A. D. Joseph, S.-h. Lau, S. Rao, N. Taft, and J. Tygar, "Antidote: understanding and defending against poisoning of anomaly detectors,"in Proceedings of the 9th ACM SIGCOMM conference on Internet measurement conference. ACM, 2009, pp. 1–14.

47 ——, "Stealthy poisoning attacks on pca-based anomaly detectors," ACM SIGMETRICS Perfor- mance Evaluation Review, vol. 37, no. 2, pp. 73–74, 2009.

48 M. Kloft and P. Laskov, "Online anomaly detection under adversarial impact."in AISTATS, 2010, pp. 405–412.

49 A. Fawzi, S.-M. Moosavi-Dezfooli, and P. Frossard, "Robustness of classifiers: from adversarial to random noise,"arXiv preprint arXiv:1608.08967, 2016.

50 T. Tanay and L. Griffin, "A boundary tilting persepective on the phenomenon of adversarial ex- amples,"arXiv preprint arXiv:1608.07690, 2016.

51 N. Papernot, P. D. McDaniel, X. Wu, S. Jha, and A. Swami, "Distillation as a defense to adversar- ial perturbations against deep neural networks,"in IEEE Symposium on Security and Privacy, SP 2016, 2016, pp. 582–597.

52 S. Gu and L. Rigazio, "Towards deep neural net- work architectures robust to adversarial examples,"arXiv preprint arXiv:1412.5068, 2014.

53 U. Shaham, Y. Yamada, and S. Negahban, "Under- standing adversarial training: Increasing local sta- bility of neural nets through robust optimization,"arXiv preprint arXiv:1511.05432, 2015.

54 Q. Zhao and L. D. Griffin, “Suppressing the un- usual: towards robust cnns using symmetric activa- tion functions,”arXiv preprint arXiv:1603.05145, 2016.

55 Y. Luo, X. Boix, G. Roig, T. Poggio, and Q. Zhao, “Foveation-based mechanisms alleviate adversar- ial examples,”arXiv preprint arXiv:1511.06292, 2015.

56 R. Huang, B. Xu, D. Schuurmans, and C. Szepesva´ri, “Learning with a strong adver- sary,”CoRR, abs/1511.03034, 2015.

57 D. Hendrycks and K. Gimpel, “Visible progress on adversarial images and a new saliency map,”arXiv preprint arXiv:1608.00530, 2016.

58 G. K. Dziugaite, Z. Ghahramani, and D. M. Roy, “A study of the effect of jpg compression on adver- sarial images,”arXiv preprint arXiv:1608.00853, 2016.

59 Q. Wang, W. Guo, K. Zhang, X. Xing, C. L. Giles, and X. Liu, “Random feature nullification for ad- versary resistant deep architecture,” arXiv preprint arXiv:1610.01239, 2016.

60 F. Zhang, P. P. Chan, B. Biggio, D. S. Yeung, and F. Roli, “Adversarial feature selection against eva- sion attacks,”IEEE Transactions on Cybernetics, vol. 46, no. 3, pp. 766–777, 2016.

# 9 附录

## 9.1 测量对抗成功

回想一下，我们在第 4 部分中使用了一种特殊的对抗性的成功。还有两个相关的概念可能被使用：

- 对于每一个 $\boldsymbol{x}$，我们检查 $y_a dv(= f(\boldsymbol{x}_a dv)) = f(\boldsymbol{x})$ 是否成立。这计算出了对抗样本的总数，其中的扰动会导致由分类器为干净的样本 $\boldsymbol{x}$ 分配的类别发生变化。可能会出现这样的情况，无论是干净的还是敌对的样本都没有被分配到正确的类别，因为分类器在测试装置上没有百分之百的精度（也可能在训练集上），然而，也可能是添加微扰导致分类器正确地对先前错误的输入进行分类。我们可能有 $f(\boldsymbol{x})not \in y$，但是 $y_a dv = y$，这是不太可能但可能发生的情况。

- 对于每一个 $\boldsymbol{x}$，我们检查是否 $y_a dv = y$。这计算了在摄动后的类不等于真正的类的敌对样本的总数。然而，这一数字还将包括那些具有对抗性的样本，而这些样本已经被错误地分类了。在这种情况下，不管攻击或防御的有效性如何，错误分类的对抗样本的百分比不能低于基线的分类器在良性测试样本上的不准确，这代表了对任何攻击的有效性的下限。

目前还不清楚这三种统计中哪一项在之前的工作中被认为是对方成功。我们在实验中计算了所有 3 项，发现它们是相似的。

## 9.2 用于防御评估的测试数据的直觉

使用的数据：规避攻击通常涉及到现有样本的修改化，使它们被错误地分类。如果一种攻击可以被认为是有效的，如果它导致了对来自训练集的敌对样本的高误分类率。由于各种原因，分类器在测试集上的准确性可能不高，而将敌对修改隔离为错误分类的原因可能是有问题的。此外，由于分类器通常是经过训练的，直到它们在训练集上有非常高的准确性，他们的决策界限反映了培训数据的分布，一个涉及到最少修改训练数据的攻击是成功的。

我们对来自测试集的反式修改样本进行了评估，主要原因是在训练集上的过度拟合可能是防御效果的一个可能的原因，因此，对防御的准确评估应该包括从测试集中制作的对抗性样本。因此，一种防御机制使一个分

类器更加安全，如果从测试集中的、经过修改的样本中，管道的分类精度高于原始分类器。精确度越高，防守就越精确。

## 9.3 CNNs

我们还在一个卷积神经网络 [38] 上进行实验我们从纸上获得的架构。这个 CNNs 的架构如下：它有两个卷积层，每个层有 32 个过滤器，后面是一个最大的池层，然后是另一个两个卷积层是 64 的过滤器, 然后是一个最大的池层。最后，我们有两个完全连接的层，每个层都有 200 个神经，然后是一个回归函数输出，有 10 个神经元（在 MNIST 的 10 个类中）隐藏层的所有神经元都是 ReLUs。我们把这个网络称为“网络报纸”。它的学习速度是 0.1（在过去的 10 年里调整到 0.01），并且 50 个时期的动量为 0.9。批大小是 MNIST 的 500 个样本，在 MNIST 测试数据上我们在 Papernot-CNN 网络得到了一个分类精度是 98.91%。