基于特征分布差异的对抗样本检测

韩 蒙[1,5], 俞伟平[2], 周依云[3], 杜文涛[1], 孙彦斌[4], 林昶廷[1*,5]

[1]浙江大学 杭州 中国 310007*
[2]杭州涂鸦科技有限公司 杭州 中国 310010
[3]CareerBuilder 芝加哥 美国 60601
[4]广州大学 广州 中国 510006
[5]浙江君同智能科技有限责任公司 杭州中国 310051

摘要 诸多神经网络模型已被证明极易遭受对抗样本攻击。对抗样本则是攻击者为模型所恶意构建的输入, 通过对原始样本输入添加轻微的扰动, 导致其极易被机器学习模型错误分类。这些对抗样本会对日常生活中的高要求和关键应用的安全构成严重威胁, 如自动驾驶、监控系统和生物识别验证等应用。研究表明在模型的训练期间, 检测对抗样本方式相比通过增强模型来预防对抗样本攻击更为有效, 且训练期间神经网络模型的中间隐层可以捕获并抽象样本信息, 使对抗样本与干净样本更容易被模型所区分。因此, 本文针对神经网络模型中的不同隐藏层, 其对抗样本输入和原始自然输入的隐层表示进行统计特征差异进行研究。本文研究表明, 统计差异可以在不同层之间进行区别。本文通过确定最有效层识别对抗样本和原始自然训练数据集统计特征之间的差异, 并采用异常值检测方法, 设计一种基于特征分布的对抗样本检测框架。该框架可以分为广义对抗样本检测方法和条件对抗样本检测方法, 前者通过在每个隐层中提取学习到的训练数据表示, 得到统计特征后, 计算测试集的异常值分数, 后者则通过深层神经网络模型对测试数据的预测结果比较, 得到对应训练数据的统计特征。本文所计算的统计特征包括到原点的范数距离 L_2 和样本协方差矩阵的顶奇异向量的相关性。实验结果显示了两种检测方法均可以利用隐层信息检测出对抗样本, 且对由不同攻击产生的对抗样本均具有较好的检测效果, 证明了本文所提的检测框架在检测对抗样本中的有效性。

关键词 神经网络, 特征分布差异, 对抗样本检测, 异常值检测
中图法分类号 TP 393.08 **DOI 号** 10.19363/J.cnki.cn10-1380/tn.2023.05.01

Exploiting Feature Space Divergence For Adversarial Example Detection

HAN Meng[1,5], YU Weiping[2], ZHOU Yiyun[3], DU Wentao[1], SUN Yanbin[4], LIN Changting[1*,5]

[1] Zhejiang University, Hangzhou 310007, China*
[2] Tuya Inc., Hangzhou 310010, China
[3]CareerBuilder, IL, 60601, USA
[4]Guangzhou University, Guangzhou 510006, China
[5]Gentel.ai, Hangzhou 310051, China

Abstract Neural network models have been shown to be vulnerable to adversarial examples, which are the maliciously crafted inputs with adding slight perturbation to the original natural inputs, resulting in incorrect classification by the ML model. Such adversarial samples threaten the security of high requirements and key applications in daily life, such as autonomous driving, surveillance systems, and biometric authentication. Recent works have shown detecting adversarial examples can be more effective than preventing them by enhancing models during training time. Moreover, the neural network model is more easier to distinguish adversarial samples from original natural samples as its middle hidden layer capture and extract sample information during training time. Therefore, we investigate the statistical divergence of hidden representations between adversarial inputs and benign inputs on different layers in neural networks in this study. Our results show that this divergence can vary among different layers. By identifying the most effective layers for identifying the divergence and the statistical representation distribution of the benign training datasets, a framework for adversarial samples detection using feature distribution is proposed in this paper. The framework can be divided into generalized adversarial samples detection method and conditional adversarial samples detection method. The former calculates the outlier score of the test set after obtaining statistical features by extracting the learned training data representation from each hidden layer. The latter obtains the statistical features of the corresponding training data by comparing the prediction results of deep neural network models. The

通讯作者: 林昶廷, 博士, 副研究员, Email: linchangting@zju.edu.cn。
本文研究成果得到国家自然科学基金项目(No. 62072404), CCF-蚂蚁科研基金(CCF-AFSG No. RF20220003), 杭州市领军型创新创业团队(No. TD2022011)资助。
收稿日期: 2022-08-09; 修改日期: 2022-12-15; 定稿日期: 2023-03-28

calculated statistical features include the L_2 norm distance from the origin and the correlation with the top eigenvalue of the sample covariance matrix. Our experiment results show that both detection methods can detect adversarial samples using hidden layer information and have good detection effects on adversarial samples generated by different attacks, thereby demonstrating the effectiveness of the proposed detection framework in detecting adversarial samples.

Key words neural network; feature space divergence; adversarial example detection; outlier detection

1 引言

深度学习在多个预测任务中广泛应用, 如图像分类[1]、语言翻译[2]及语音分析[3]等。然而, 其模型的鲁棒性也备受学术界和工业界关注[4-5]。大量深度学习模型应用在高要求的关键领域中, 因此模型在应对对抗样本攻击时所具备的鲁棒性便显得尤为重要。对抗样本通常指由敌手明确设计且不易察觉的扰动输入, 用以误导某个目标深度学习模型。对抗样本会对日常生活中高要求和关键应用的安全构成严重威胁, 比如自动驾驶、监控系统[6]和生物识别验证[7]等领域。

为了降低对抗样本攻击所带来的安全性影响, 诸多解决方案被学术界和工业界所提出, 但这些方案均存在相关缺陷。如对抗性训练[8]和防御性蒸馏法[9], 这两类方法都只能在模型训练期间提高模型的鲁棒性。另一类方案则利用输入转换方法的防御措施减轻恶意扰动[10]。然而, 该类方法并不能有效适应生成对抗样本变化, 同时还需要改变原始模型的训练过程。由于这些局限性, 当前学术界的研究方向主要聚焦于针对对抗样本的检测。现有的检测方法一般利用已知攻击的对抗样本, 训练与原模型分离的对抗样本检测分类器[11], 或在原模型中增加一个额外的对抗样本类别进行分类[12]。另外, 还有相关研究利用原始输入和转换后的输入之间的模型结果差异检测对抗样本。然而, 这些方法在面对不同攻击生成的对抗样本时, 其鲁棒性有限。

本文研究神经网络模型中, 其输入的原始自然样本和对抗样本在不同隐层上, 表现出的特征分布差异。基于实验结果, 确定了一种利用特征分布差异进行对抗样本检测的方法。本文的实验结果表明, 源自同一类别样本的隐层表示往往逼近于同一分布, 而对抗样本的特征表示在某些层上与原始自然样本的分布则有所不同。主要原因为, 攻击者为了达到攻击效果, 对抗样本的特征表示必须至少在输出层前偏离原始自然样本的特征表示。由于每一层的特征表示都是不同层输入特征的抽象提取, 因此对抗样本可能在某一层具有不同于原始自然样本的特征表示。通过构建隐层的统计特征来捕获训练数据集中数据在确定层上的隐藏特征分布, 若隐藏特征表示与统计特征有较大偏差, 基于本文提出的方法则可以识别出对抗样本。本文的实验结果也证明了该检测方法的有效性。

1.1 相关工作

迄今为止, 已有大量工作针对模型推理期间的对抗样本检测进行研究[13-14]。其中, 对抗样本检测方法通常依据以下理论: 给定一个 K 类别的神经网络分类器, 原本的训练集为 $D=\left\{x_i \in R^d\right\}_{i=1}^N$, 构建一个对抗样本集 $D'=\left\{x'_j \in R^d\right\}_{j=1}^N$, 然后设计一个检测方法区分 D' 和 D 。

Gong 等人[9]提出了一种二元分类器, 以高精度分离对抗样本数据和干净数据。同时本文作者还发现, 二进制分类器对对抗攻击算法较为敏感。具体来说, 在 FGSM 样本上训练的二进制分类器对 JSMA 样本不具备鲁棒性, 反之亦然。然而, 如果分类器在 FGSM 和 JSMA 的混合对抗样本上进行训练, 则分类器可以以较好的性能检测出两者。Grosse 等人[15]将“攻击”类添加为模型的异常值类, 用干净和对抗性数据一起训练模型。Metzen 等人[11]通过使用中间层特征作为输入数据来识别 D' 和 D 来构建检测器。检测器可以通过发现在附近类边界的特定方向上稍微偏离干净数据流中心的输入来识别对抗样本, 该类边界泛化为未知数据, 需要某些扰动规律。然而, Carlini 等人[16]指出, 检测方法在识别具有强攻击的自适应敌手时效率将显著降低。为了应对这一挑战, 作者提出了一种范数约束的对抗样本检测器方法, 该方法保证识别潜在威胁。Lu 等人[17]提出使用每个 ReLu 层输出的二进制阈值作为对抗样本检测器的特征, 并通过 RBF-SVM 分类器识别对抗样本。Lu 等人[17]还表明, 即使对抗样本知道检测器, 其方法也很难被对抗样本所攻破。

Feinman[18]则通过研究对抗样本的预测置信度, 提出一种贝叶斯神经网络实现对抗样本和原始干净数据的有效区分。Meng 等人[19]不仅实现了对抗样本检测, 还利用了一个改型自动编码器将对抗样本移向正常样本的流形, 这对于在小扰动下正确分类对抗样本极为有效。Pang 等人[20]提出训练反向交叉熵,

以实现鼓励DNN通过潜在表示将对抗样本与干净数据区分开。反向交叉熵采用 DNN 对目标类具有较高的置信度并且在其他类上具有相似的分布。相较标准交叉熵最小化方法, 该方法实现简单, 额外计算成本更少。

1.2 本文贡献

在实验中, 本文发现当 D'和 D 中的样本映射到神经网络隐层中的学习表示时, D'和 D 的样本统计特征可以被区分。神经网络模型的中间隐层捕获并抽象样本信息, 利用此信息对抗样本与干净样本则更易被区分。本文利用表示特征的 Z-Score 异常值补偿的统计工具, 提出基于特征分布差异的对抗样本检测框架。本文的贡献点可总结为如下几点:

- 提出了一种基于特征分布的对抗样本检测框架, 该框架分为广义对抗样本检测方法和条件对抗样本检测方法, 利用特征分布的信息可以从一个崭新的维度展开检测;
- 构建的对抗样本检测方案具备较好的检测能力, 能有效检测并区分不同攻击产生的对抗样本, 同时还能对不同类型神经网络所产生的对抗样本进行有效识别;
- 除在理论上可对特征分布及特征提取技术领域提供参考, 通过与多个当前先进算法进行对比实验, 其结果表明本文所提出的算法和框架可以利用隐层信息检测出对抗样本。

2 检测框架

如图 1 所示, 本文利用神经网络模型的记忆特性设计了对抗样本检测框架。神经网络模型的记忆特性是指模型的隐藏层可以全面捕获并提取数据的抽象特征信息。虽然原始干净数据 D 和对抗样本数据 D'的统计特征很难在数据集层面上被直接区分, 但本文利用隐层的抽象特征信息区分出干净数据和对抗样本, 这也是隐层表示的统计特征差异可用作对抗样本检测方法并提高模型鲁棒性的最初动机。

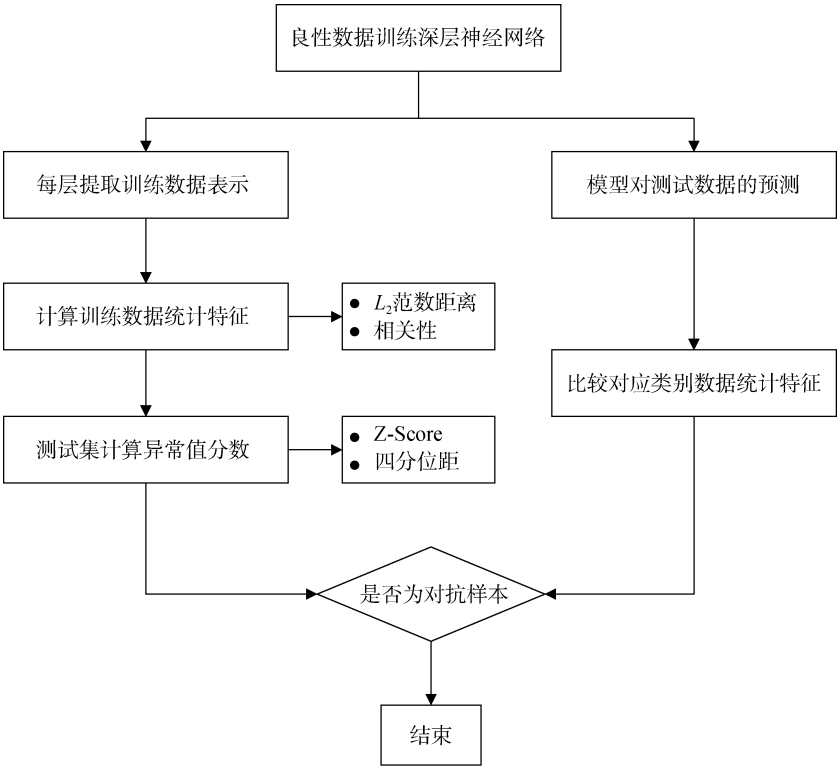

图 1 对抗样本检测框架图

Figure 1 Adversarial samples detection framework

本文所提出的检测框架包含两种检测对抗样本的方法: 广义对抗样本检测方法与条件对抗样本检测方法。对于这两种检测方法, 都需要先在干净的训练数据上训练深层的神经网络模型。具体地, 在广义方法中, 本文在每个隐层中提取学习到的训练数据表示, 然后计算训练数据的统计特征, 并利用它们为可能包含对抗样本的测试集计算出异常值分数; 在条件方法中, 本文首先得到深层神经网络模型对测试数据的预测结果, 然后与对应类别的训练数据比较统计特征表示。另外, 本文在该框架中应用了两

个统计特征, 一个是到原始数据的 L_2 范数距离, 另一个则是与样本协方差矩阵的顶奇异向量的相关性。对于每个测试数据, 本文计算异常值分数, 如果异常分数超过某一阈值, 则判定其为对抗样本。上述两种方法均可检测出隐藏层中的对抗样本。

2.1 隐层表示的统计特征

为了使用统计特征来区分来自 D 的干净训练数据 x_i 以及来自 D_0 的对抗样本 x'_j, 本文假设 D 的隐层表示遵循正态分布, 并且在某些层上与 D_0 隐层表示的统计特征截然不同。

在实验中, 本文发现当 D 和 D_0 映射到神经网络的特征表示时, D 和 D_0 的统计特征相互分离。神经网络模型的中间层会捕获 D_0 中的信号, 从中间层中更容易区分对抗样本。因此, 本文利用相关统计工具, 如 Z-Score 异常值检验, 在隐层表示上检测对抗样本。在第 4 节中, 本文通过实验展示了如何有效地删除 D_0, 并防止对抗样本的潜在威胁。本文提出的框架也可以推广至任何其它深度神经网络, 且不需要过多的计算成本。对于隐层的统计特征, 本文则考虑计算训练样本的 L_2 范数, 其与训练样本协方差矩阵的随机向量和顶奇异向量的相关性。

通过图 2 和图 3 可以发现, ResNet 神经网络不同层下, 对抗样本分布与干净训练数据 L_2 分布之间存在一定的分布差异。图 2 展示了数据点分布的 L_2 范数, 即在数据层级中通过基本迭代方法 BIM(Basic Iterative Method)生成的对抗样本和干净训练数据一般位于相同分布中。然而, 在隐藏层中, 模型将增强或减弱对抗性样本的信号, 使得对抗样本与干净数据分离。同样地, 如图 3 所示, 通过 DeepFool 方法所生成的对抗性样本, 在不同的表示层上具有相似的属性。在 ResNet 神经网络中, 两种对抗样本具有不同的分离级别。由于无法得知哪一隐层表示的统计特征最适合分离对抗样本, 为了有效识别未知攻击, 因此本文在 DNN 的每一层中检测对抗样本。

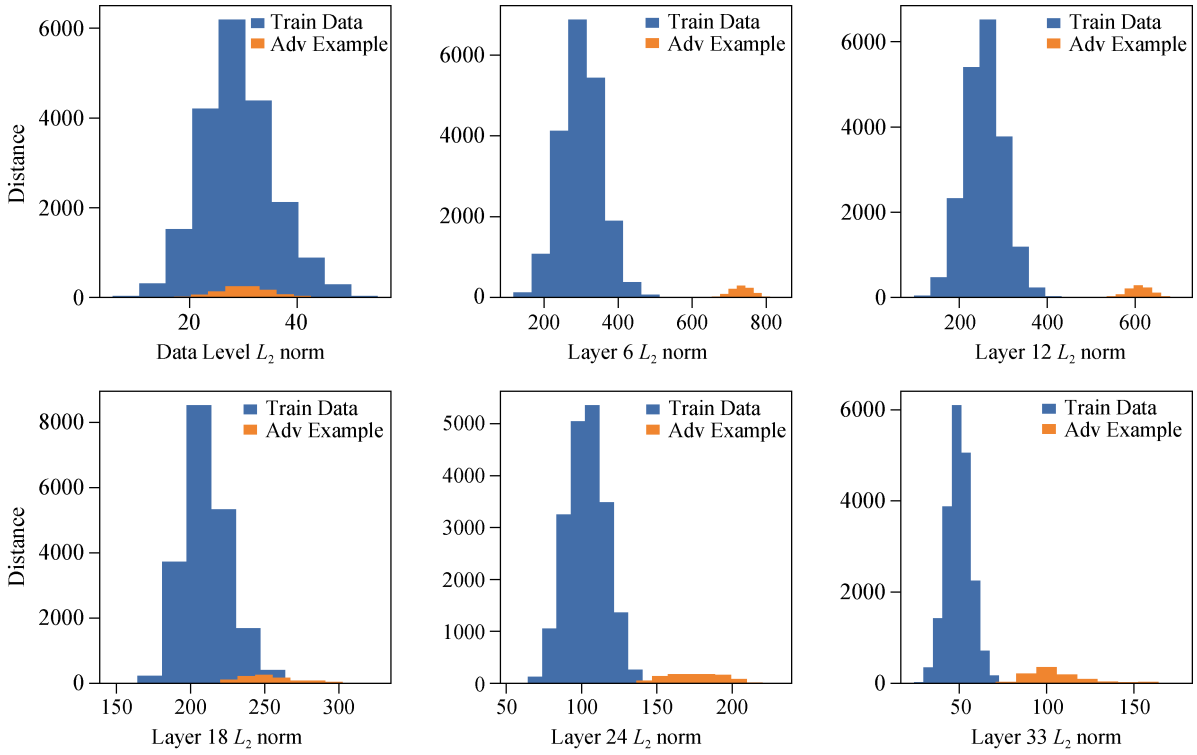

图 2 BIM 对抗样本和训练数据 L_2 分布

Figure 2 L_2 distribution of BIM adversarial samples and training data

传统的 CNN 操作可以形成迁移学习表示, 因此, 本文提出的检测框架能够识别各种神经网络结构中不同类型的对抗样本。在该项工作中, 本文在框架中应用了两个统计特征: 一个是到原点的 L_2 范数距离; 另一个则是样本协方差矩阵的顶奇异向量的相关性, 对于每个测试数据点, 计算异常值分数, 若分数超过阈值, 则认为该测试数据为对抗样本。

2.2 通过统计特征检测异常

首先, 本文需要计算训练数据集的统计特征。本文将训练好的模型应用于测试数据, 比较测试数据

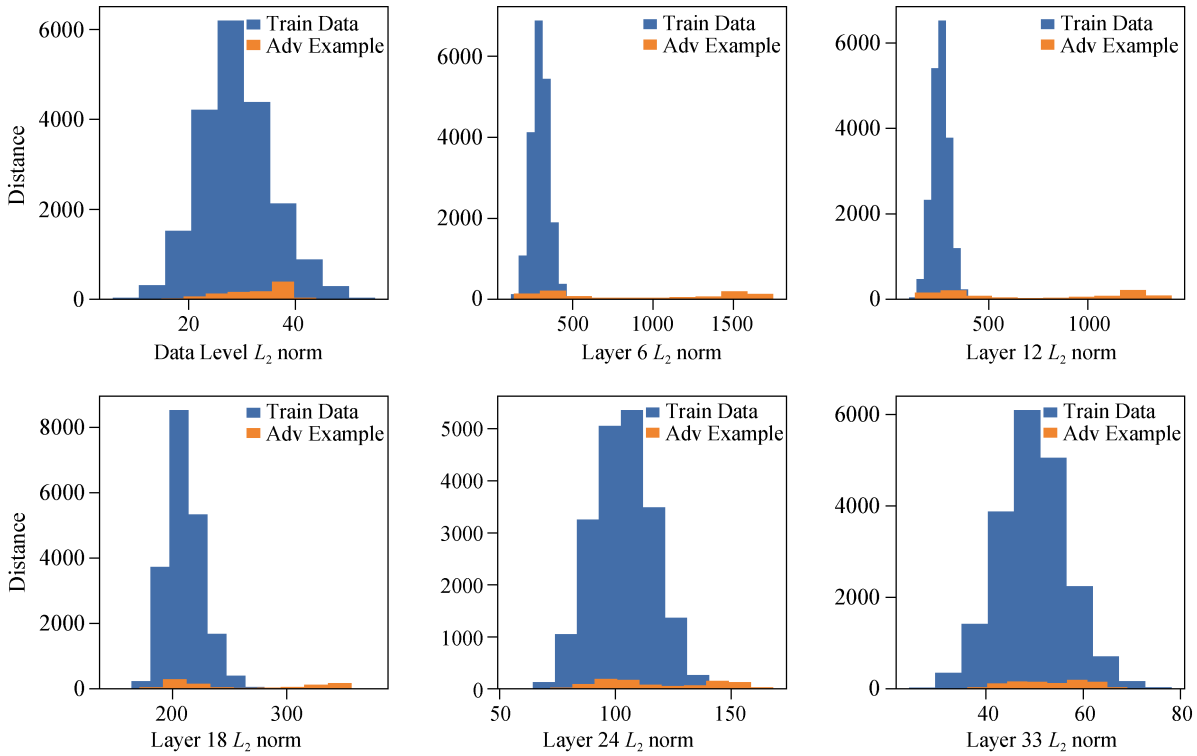

图 3　DeepFool 对抗样本和训练数据 L_2 分布

Figure 3　L_2 distribution of DeepFool adversarial samples and training data

和训练数据之间的统计特征，一些隐层的特征表示可以很好地检测对抗样本，而一些层则无效。数值线性代数中最常用的矩阵范数则是 Frobenius 范数[21]，其 L_2 范数定义为:

$$\|f\|_{L_2(\Omega)} = \sup_{\substack{v\in L_2(\Omega)\\ v\neq 0}} \frac{|(f,v)|}{\|v\|L_2(\Omega)} \tag{1}$$

为了得到另一个统计特征，本文需要在每一层计算训练数据集的奇异值分解(SVD)表示。奇异值分解是一种矩阵分解方法，其提取矩阵的抽象表示，可以删除不太重要的信息并降低矩阵维度。定义一个秩为 r，维度为 $m*n$ 的矩阵 A，本文可以找到两个正交矩阵 U, V 和一个对角矩阵 σ 满足以下公式:

$$\underbrace{A}_{m\times n} = \underbrace{U}_{m\times r} \times \underbrace{\Sigma}_{r\times r} \times \underbrace{V^T}_{r\times n} \tag{2}$$

矩阵 A 的维度是 $m*n$，包含 V 的奇异向量，这里考虑 V 是 A 的 m 个样本的 n 维空间中的最佳拟合线，矩阵 A 第 i 行在 V 上的投影是 $|a_i \cdot v|$，最佳拟合线最大化投影是 $|A\cdot v|^2$ 的长度平方和，并最小化点到该线的距离平方和。A 的第一个奇异向量 v_1，也称为 A 的协方差矩阵的顶奇异向量，定义为:

$$v_1 = \arg\max_{|v|=1} |A\cdot v| \tag{3}$$

2018 年，Tran 等人提出了鲁棒性的统计指标[22]，即 a_i 与 v_1 的相关性，并实现了利用隐层表示识别深度学习模型中的后门攻击。本文参考该做法，利用各隐层中的这些指标来检测对抗样本，可将其定义为:

$$\rho = \frac{\mathrm{cov}(v_1, a_i)}{\sigma_{v_1}\sigma_{a_i}} \tag{4}$$

本文假设干净数据样本 L_2 范数和表示层中第一个奇异向量的相关性服从正态分布。对抗样本 D' 与干净训练数据 D 在数据层级上位于相同的分布中，但是 D' 在某些隐层表示上可能是可分离的。即本文可以应用异常值检测方法，通过学习表示的统计特征来识别 D'。

一种方法是 Z-Score 异常值检测。在该检测过程中，本文计算每个可能包含对抗样本测试数据的 Z-Score，若观察值 Z-Score 的绝对值大于 3，则被视为异常值，在本研究中 Z-Score 被定义为:

$$z = \frac{x^i_{\text{test}} - \mu_{\text{train}}}{\sigma_{\text{train}}} \tag{5}$$

这个方法遵循经验法则，因为几乎所有(99.7%)数据的 Z-Score 绝对值都会在 3 之内。

另一个方法是四分位距 IQR(Interquartile Range)规则，本文需要获得第一四分位数 Q_1 和第三四分位数 Q_3 的值，代表训练数据特定统计特征值的四分之一和四分之三。这里四分位距的值为第三四分位数

减去第一四分位数后的结果, 如下所示:

$$IQR = Q_3 - Q_1 \tag{6}$$

计算出测试数据的统计特征值, 任何大于 $Q_3 + 1.5 * IQR$ 或小于 $Q_1 - 1.5 * IQR$ 可被认为是离群值, 即对抗样本。在本文的实验中, 这两种离群点检测方法具有相似的性能, 能够以极低的误报率识别出对抗样本 D'。

2.3 分布距离

为了评估对抗样本在神经网络不同层中的敏感度, 进而选择出最佳的隐层, 本文测量两个概率分布 D'和 D 之间的距离, 使用两个分布距离来衡量对抗样本与干净数据在不同隐层中的差异性。

本文考虑两个分布距离指标: Wasserstein 距离和能量距离。Wasserstein 距离是一种用来估计特征空间中两个分布之间差异性的方法, 其中单个特征之间的距离测量也称为搬土距离 EMD(Earth Mover's Distance)[23], 更具体地说, 给定两个分布, 一个可以被视为在空间中充分分布的“一团土”, 另一个是需要在同一空间中填充的一组“孔洞”, EMD 则估算了用泥土填充孔洞所需的最小工作量。能量距离评估样本与相同或不相同维度的假设分布之间的差异[24]。

图 4-图 7 显示了隐层表示中数据分布距离的变化。在 BIM 和 DeepFool 对抗样本中, 两个指标的结果是一致的。BIM 样本在 ResNet 结构的开始部分和结束部分是可分离的。DeepFool 样本仅在开始部分可分离。根据实验结果, 本文发现很难确定某一隐层表示最适用于对抗样本检测, 因此本文所提出的对抗样本检测框架则利用了神经网络中每一个隐层。

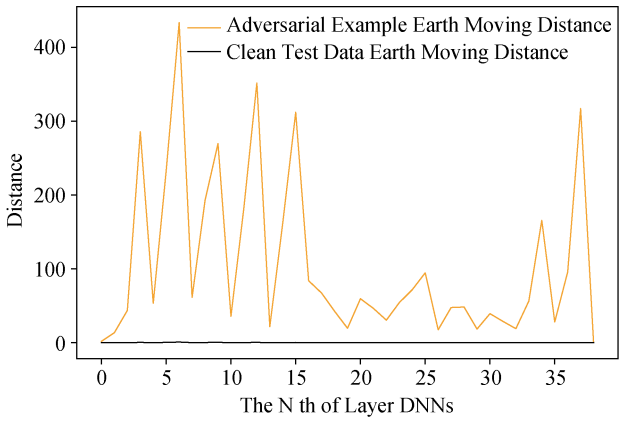

图 4 BIM 样本和测试数据到训练数据的 EMD

Figure 4 EMD from BIM adversarial samples and test data to training data

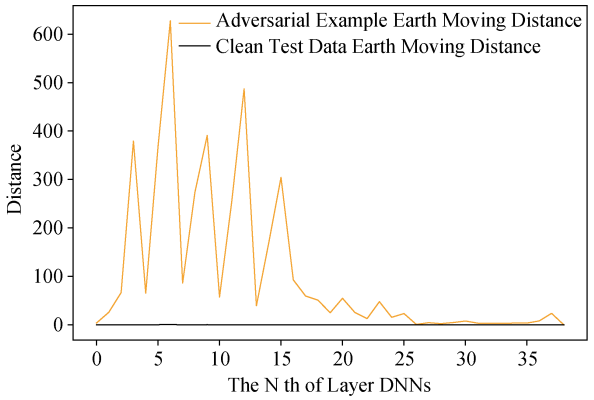

图 5 DeepFool 样本和测试数据到训练数据的 EMD

Figure 5 EMD from DeepFool adversarial samples and test data to training data

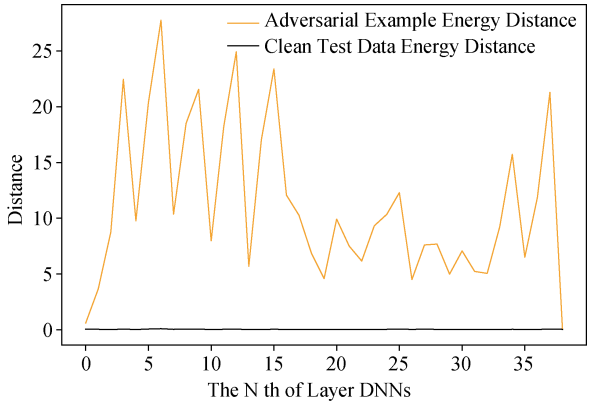

图 6 BIM 样本和测试数据到训练数据的能量距离

Figure 6 Energy distance from BIM adversarial samples and test data to training data

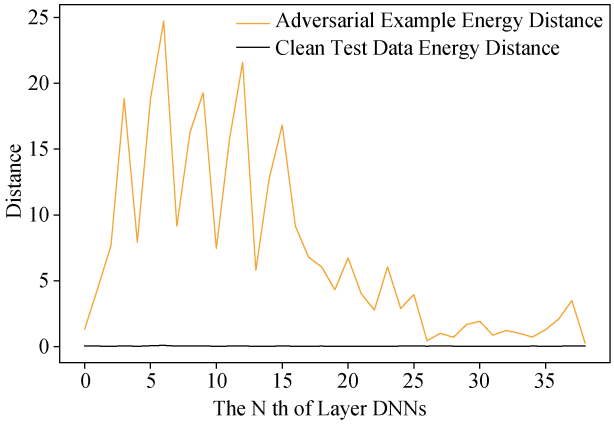

图 7 DeepFool 样本和测试数据到训练数据的能量距离

Figure 7 Energy distance from DeepFool adversarial samples and test data to training data

3 实验

3.1 实验设置

在实验中, 本文研究了 CIFAR10 数据集[25]上对

抗样本的特征, 在 ResNet 深度学习模型中测试本文的检测方法。为了生成对抗样本, 本文使用了 python 3 对抗样本生成工具[26]。神经网络模型中的不同层形成各种隐层表示, 通常顶部隐层被视为“高层”。实验结果表明, 防御框架在“低层”表示中能够更有效地检测对抗样本。

3.2 攻击和防御方法

本文使用三种目前最先进的对抗样本来评估检测框架: 基本迭代法(BIM)[27]、DeepFool[28]和C&W[29]。根据攻击强度级别, 可以将三个对抗样本分为: BIM-低、DeepFool-中、C&W-高。在文献[16]中, 作者表明, 通过修改 C&W 攻击算法中相应的代价函数, C&W 方法可以打败大多数对抗性检测框架。在实验中, 本文选择 32*32color CIFAR10 数据集, 其中包含 10 个类别的 60000 个数据观察值。本文在 3 块大小的 ResNet 神经网络上训练 CIFAR 数据, 分类精度达到 93.42%。然后根据训练后的 ResNet 的梯度信息生成三种类型的对抗样本, 每种类型有 1000 个对抗样本。

3.3 检测框架评估

在本节中, 本文展示了基于各种统计特征和离群点检测指标的防御框架, 实验评估了广义和条件检测框架中的性能。在检测指标中, 本文利用干净训练数据集的统计特征, 然后将测试数据集统计特征与干净训练数据集中进行比较。本文还评估了将干净测试数据识别为对抗样本的假阳性案例, 并在不同的 Z-Score 检测方法设置下测试结果。

图 8-图 13 展示了所提出的防御框架在神经网络每个隐藏层中的性能。一般来说, Z-Score 和 IQR 检测方法具有相似的性能。在默认设置中, Z-Score 检测器识别出偏离平均值的三个标准差的数据观察值, IQR 检测器识别出高于 1.5 倍 IQR 的数据观察结果。基于扰动大小和攻击强度, 高级攻击算法生成的可检测对抗样本较少。与协方差矩阵的顶奇异向量的相关性相比, L_2 范数则是更好的评估统计特征。

3.3.1 广义检测方法

BIM 样本可以在不同的层表示中轻松检测。通过本文提出的模型可以识别所有 BIM 样本。DeepFool 可以在“低级”特征中检测到, 但在“高级”特征中很难找到。除平坦层外, 在所有隐藏层中几乎检测不到“最强”C&W 对抗样本。此外, 一些对抗样本无法在一层表示中识别, 但可以在其他层表示中检测到。聚合所有层检测结果应该是一个好的策略, 并且统计特征不会消耗太多计算能力。

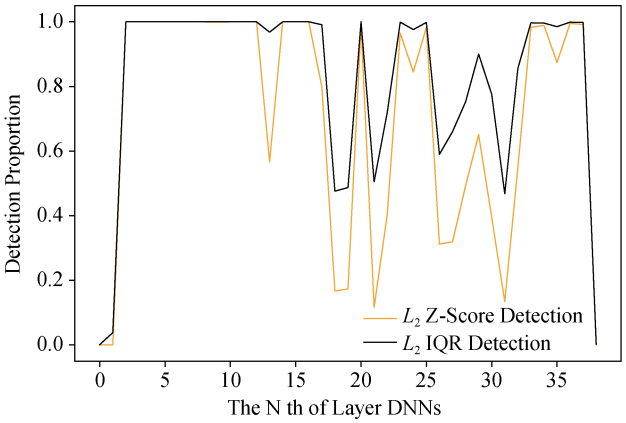

图 8 ResNet 不同层 BIM 对抗样本 L_2 检测

Figure 8 L_2 detection of BIM adversarial samples on different layers of ResNet

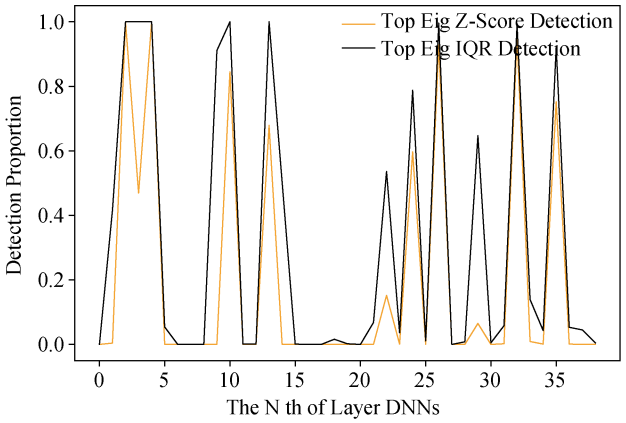

图 9 ResNet 不同层 BIM 对抗样本顶部特征检测

Figure 9 Top Eigen detection of BIM adversarial samples on different layers of ResNet

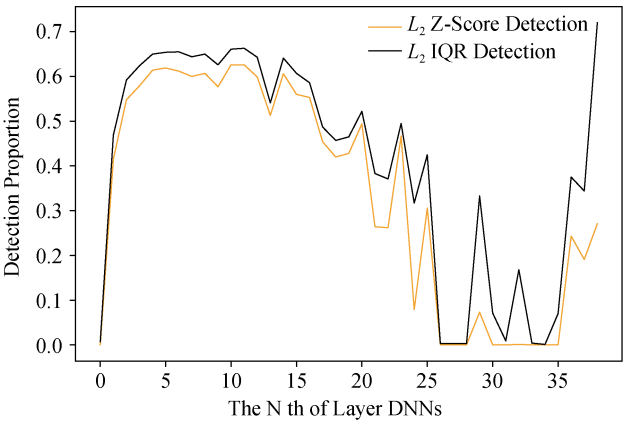

图 10 ResNet 不同层 DeepFool 对抗样本 L_2 检测

Figure 10 L_2 detection of DeepFool adversarial samples on different layers of ResNet

除了对抗样本检测外, 误报率则是针对防御框架的另一个重要评估指标。如图 14 和图 15 所示, 除最后一层的表示外, 其他层的假阳性率极低。这表明

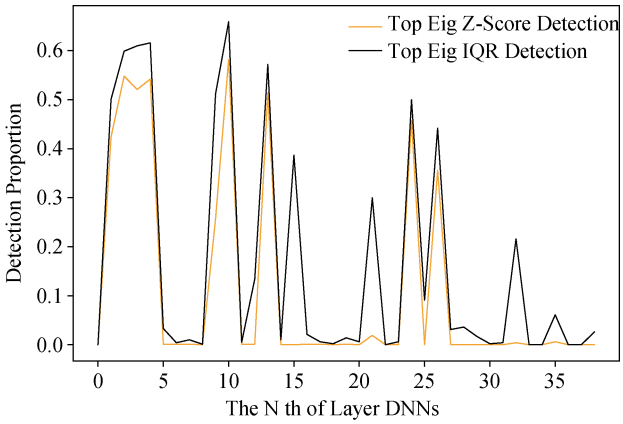

图 11　ResNet 不同层 DeepFool 对抗样本顶部特征检测

Figure 11　Top Eigen detection of DeepFool adversarial samples on different layers of ResNet

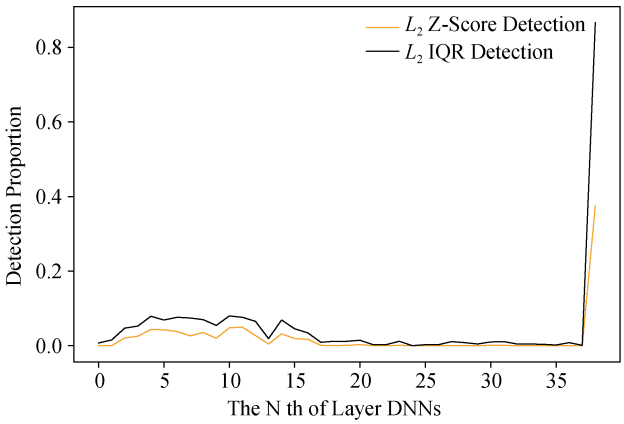

图 12　ResNet 不同层 C&W 对抗样本 L_2 检测

Figure 12　L_2 detection of C&W adversarial samples on different layers of ResNet

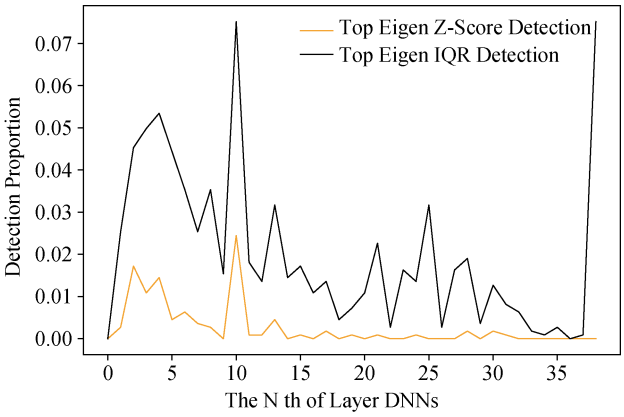

图 13　ResNet 不同层 C&W 对抗样本顶部特征检测

Figure 13　Top Eigen detection of C&W adversarial samples on different layers of ResNet

本文需要调整检测器的参数, 并制定更严格的标准, 以减轻对干净测试数据的负面影响。

表 1 展示了 Z-Score 检测器在 2.5、3 和 4 标准差设置下的不同性能。这意味着统计特征值大于 2.5、3 或 4 可能是对抗样本。因此, 即使有严格的标准设置, 4 个标准差, Z-Score 检测器仍然以非常低的假阳性率过滤出实质性对抗样本, 在两个相应的统计特征中, 从 28.3%下降到 6.5%, 从 33.4%下降到 2.3%。然而检测率并没有显著影响。BIM 检测率在 L_2 中保持不变且与顶部特征检测率一致。DeepFool 的 L_2 检

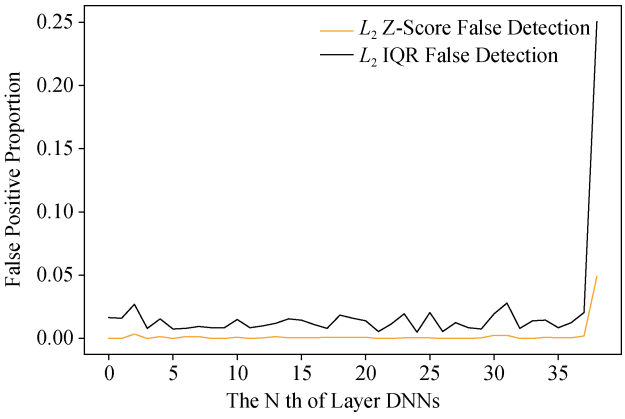

图 14　ResNet 不同层 L_2 检测的假阳性率

Figure 14　L_2 detection of false positive proportion on different layers of ResNet

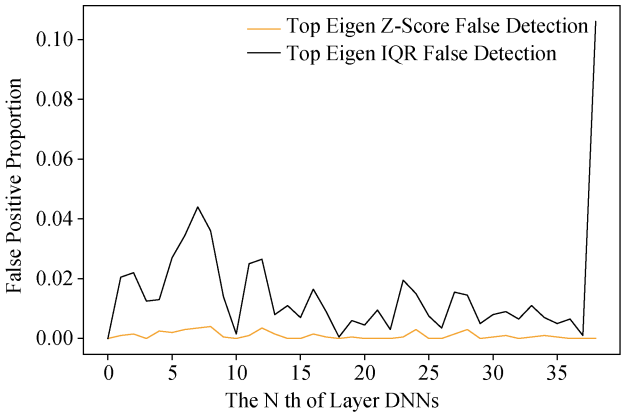

图 15　ResNet 不同层顶部特征检测的假阳性率

Figure 15　Top Eigen detection of false positive proportion on different layers of ResNet

表 1　不同 Std 标量下的 Z-Score 检测精度

Table 1　Z-Score detection accuracy under different Std scales

Z-Score 检测	2.5 Std	3 Std	4Std
BIM L_2	100%	100%	100%
DeepFool L_2	85.2%	81.6%	77.7%
C&W L_2	65.9%	57.5%	42.7%
False Positive L_2	28.3%	16.2%	6.5%
BIM Top Eigen	100%	100%	100%
DeepFool Top Eigen	76.0%	68.9%	59.4%
C&W Top Eigen	35.6%	16.3%	4.9%
False Positive Top Eigen	33.4%	11.3%	2.3%

然而检测率并没有显著影响。BIM 检测率在 L_2 中保持不变且与顶部特征检测率一致。DeepFool 的 L_2 检

测率从 85.2%降到 77.7%, 顶部特征检测率从 76.0%降低到 59.4%。比较两种静态度量的有效性, L_2 统计特征在对抗样本检测中更有用。顶部特征几乎不能反映 C&W 对抗样本的特征。

总的来说, 广义检测框架可以很好地识别未知的对抗样本, 而无需大量额外计算, 并且优于 3.4 中其他最先进的对抗防御模型性能。

3.3.2 条件检测方法

在条件检测框架中, 本文首先得到每个观察的预测结果, 并检测每个类别中的对抗样本。Z-Score 条件检测结果如表 2 所示, 有意思的是, 某些类别的检测性能明显优于其他类别。船舶对抗样本易于识别, 鹿对抗样本难以识别。总体性能略低于一般检测方法。L_2 统计签名的性能优于顶部特征向量统计签名, 这与一般检测方法一致。

如图 16 和图 17 所示, 其展示了 ResNet 不同层的船舶 C&W 对抗样本 L_2 和顶部特征的检测结果。检测模式与一般检测方法一致, 实验结果显示: “低水平”的检测率较差, “高水平”的识别效果较好。

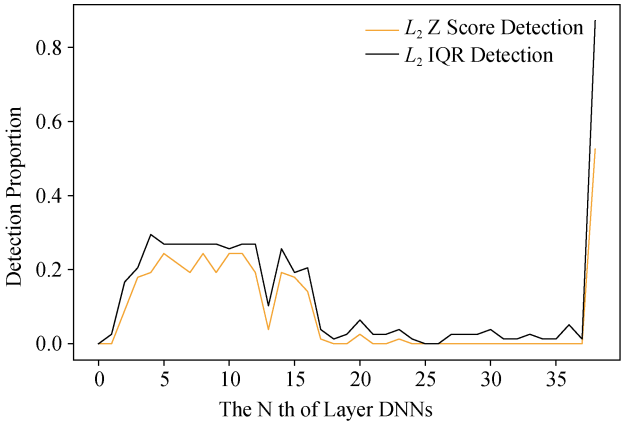

图 16 ResNet 不同层船舶 C&W 对抗样本 L_2 检测

Figure 16 L_2 detection of ship C&W adversarial samples on different layers of ResNet

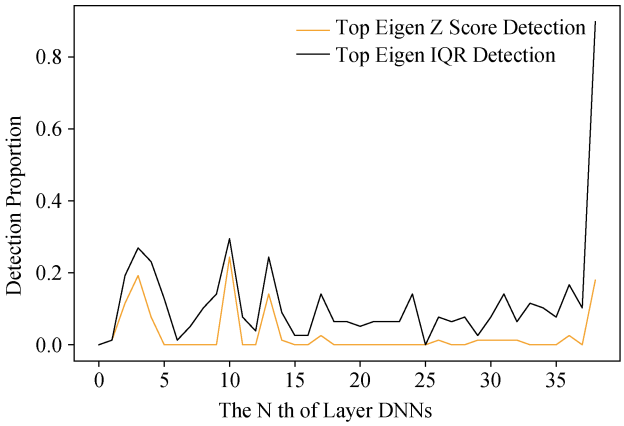

图 17 ResNet 不同层船舶 C&W 对抗样本顶部特征检测

Figure 17 Top Eigen detection of ship C&W adversarial samples on different layers of ResNet

3.4 方法对比

在本小节中, 本文将提出的方法与其他最先进的对抗防御框架进行比较。文献[29]提出了一种自动架构搜索鲁棒神经网络来防御对抗性攻击。作者利用遗传算法提高了网络在每次迭代中的鲁棒性。该架构可以在对抗性样本上发展为固有的精确性。文献[30]提出了高斯数据增强方法, 以提高预测的鲁棒性。文献[31]提出了特征压缩和空间平滑来传输数据, 以增强 DNN 对抗对抗样本的能力。

表 3 显示了对抗数据集上四种防御策略的预测精度。如果本文没有这些防御方法, 那么在干净的测试数据上的分类器预测是 93.4%, 在对抗样本上是 0.00%。注意, 与上述提出的方法相比, 本文重点关注光计算方法。因此, 本文仅利用高斯增强、空间平滑和特征压缩作为去噪变换器来消除对分类器的扰动影响。结果表明, 这些防御处理器的性能比本文提出的方法差。有趣的是, 虽然这些框架防御对于强对抗样本(C&W)更好, 但在“弱”对抗样本中表现相对

表 2 Z-Score 条件检测

Table 2 Z-Score condition detection

Z-Score 检测	飞机	汽车	鸟	猫	鹿	狗	青蛙	马	船	卡车	总计
BIM L_2	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
DeepFool L_2	64.7%	62.5%	73.3%	77.6%	15.2%	42.8%	25.3%	61.8%	96.9%	87.5%	60.8%
C&W L_2	41.3%	63.3%	66.0%	64.2%	9.2%	34.8%	12.4%	34.7%	70.5%	52.7%	44.9%
False Positive L_2	6.5%	6.9%	6. 1%	5.4%	16.5%	10.0%	15.2%	11.8%	2. 1%	8.2%	8.9%
BIM Top Eigen	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
DeepFool Top Eigen	45. 1%	40.6%	26.7%	56.4%	15.2%	11.4%	24.1%	26.3%	95.9%	75.6%	41.7%
C&W Top Eigen	20.6%	25.0%	8.0%	37.9%	4.6%	6.2%	7.3%	4.9%	44.9%	18.2%	17.8%
False Positive Top Eigen	5.6%	5.4%	9.8%	7.4%	3.2%	4.7%	2.8%	10.3%	5.7%	9.7%	6.5%

表 3 其他防御框架性能

Table 3 Performance of other defense frameworks

方法	BIM	DeepFool	C&W
自动架构[29]	10.6%	10.2%	17.0%
高斯增强[30]	10.0%	10.1%	17.2%
空间平滑[31]	0.5%	26.3%	38.2%
特征压缩[31]	0.4%	9.8%	18.8%
集成方法	8.9%	9.2%	23.2%

较差。对于 DeepFool 和 C&W 对抗样本, 集成方法效果不佳, 空间平滑显著提高了预测精度。通过上述实验可得出结论: 各种对抗样本具有不同的属性, 很难确定哪种防御方法最适合目标分类器。

4 讨论与总结

本文解释了统计特征的概念, 以及如何将其用于检测各种对抗样本。防御框架依赖于分类器每一层学习到的特征表示, 该特征表示可以提高原始信息的分类能力。当对抗样本混合在干净数据集中时, 隐藏层的特征表示将统计特征放大, 本文通过离群点检测器在干净数据中分离出对抗样本。

通过实验发现, 在某些层中, 对抗样本表示的统计特征足以改变原本的分布, 从而可以利用检测方法进行检测。此外, 本文还证明在每一层中提取特征表示的必要性, 因为每一层的检测率是不同的, 在某些隐藏层中无法检测到对抗样本, 但可能在其它层中识别。此外, 不同类型的对抗样本在每一层中都不具有相同的增强强度信号。一般性检测方法的性能优于条件检测方法, L_2 范数到原始点的距离是识别对抗样本的良好度量。

参考文献

[1] Chan T H, Jia K, Gao S H, et al. PCANet: A Simple Deep Learning Baseline for Image Classification?[J]. *IEEE Transactions on Image Processing*: *a Publication of the IEEE Signal Processing Society*, 2015, 24(12): 5017-5032.

[2] Vaswani A, Bengio S, Brevdo E, et al. Tensor2Tensor for Neural Machine Translation[EB/OL]. 2018: arXiv: 1803.07416. https://arxiv.org/abs/1803.07416

[3] Zhang S X, Chen Z, Zhao Y, et al. End-to-End Attention Based Text-Dependent Speaker Verification[EB/OL]. 2017: arXiv: 1701.00562. https://arxiv.org/abs/1701.00562

[4] Gan Y Y, Mao Y H, Zhang X H, et al. Is your Explanation Stable?: A Robustness Evaluation Framework for Feature Attribution[EB/OL]. 2022: arXiv: 2209.01782. https://arxiv.org/abs/2209.01782

[5] Zheng H B, Li X H, Chen J Y, et al. One4All: Manipulate one Agent to Poison the Cooperative Multi-Agent Reinforcement Learning[J]. *Computers & Security*, 2023, 124: 103005.

[6] Fang Q, Li H, Luo X C, et al. Detecting Non-Hardhat-Use by a Deep Learning Method from Far-Field Surveillance Videos[J]. *Automation in Construction*, 2018, 85: 1-9.

[7] Mao Z J, Yao W X, Huang Y F. EEG-Based Biometric Identification with Deep Learning[C]. *2017 8th International IEEE/EMBS Conference on Neural Engineering*, 2017: 609-612.

[8] Liao F Z, Liang M, Dong Y P, et al. Defense Against Adversarial Attacks Using High-Level Representation Guided Denoiser[C]. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018: 1778-1787.

[9] Gong Z T, Wang W L, Ku W S. Adversarial and Clean Data are not Twins[EB/OL]. 2017: arXiv: 1704.04960. https://arxiv.org/abs/1704.04960

[10] Papernot N, McDaniel P, Wu X, et al. Distillation as a Defense to Adversarial Perturbations Against Deep Neural Networks[EB/OL]. 2015: arXiv: 1511.04508. https://arxiv.org/abs/1511.04508

[11] Cihang Xie, Yuxin Wu, Laurens van der Maaten, et al. Feature denoising for improving adversarial robustness[C]. *The IEEE Conference on Computer Vision and Pattern Recognition*, 2019: 501-509.

[12] Metzen J H, Genewein T, Fischer V, et al. On Detecting Adversarial Perturbations[EB/OL]. 2017: arXiv: 1702.04267. https://arxiv.org/abs/1702.04267

[13] Li X R, Ji S L, Han M, et al. Adversarial Examples Versus Cloud-Based Detectors: A Black-Box Empirical Study[J]. *IEEE Transactions on Dependable and Secure Computing*, 2021, 18(4): 1933-1949.

[14] Carlini N, Wagner D. Adversarial Examples are not Easily Detected: Bypassing Ten Detection Methods[C]. *The 10th ACM Workshop on Artificial Intelligence and Security*, 2017: 3-14.

[15] Grosse K, Manoharan P, Papernot N, et al. On the (Statistical) Detection of Adversarial Examples[EB/OL]. 2017: arXiv: 1702.06280. https://arxiv.org/abs/1702.06280

[16] Carlini N, Wagner D. Towards Evaluating the Robustness of Neural Networks[C]. *2017 IEEE Symposium on Security and Privacy*, 2017: 39-57.

[17] Lu J J, Issaranon T, Forsyth D. SafetyNet: Detecting and Rejecting Adversarial Examples Robustly[C]. *2017 IEEE International Conference on Computer Vision*, 2017: 446-454.

[18] Feinman R, Curtin R R, Shintre S, et al. Detecting Adversarial Samples from Artifacts[EB/OL]. 2017: arXiv: 1703.00410. https://arxiv.org/abs/1703.00410

[19] Dongyu Meng and Hao Chen. Magnet: a two-pronged defense against adversarial examples[C]. *The 2017 ACM SIGSAC Conference on Computer and Communications Security*, 2017: 135-147.

[20] Tianyu Pang, Chao Du, Yinpeng Dong, et al. Towards robust detection of adversarial examples[C]. *In Advances in Neural Information Processing Systems*, 2018: 4579-4589.

[21] Gene H Golub and F Charles. van loan[J]. *Matrix computations*, 2, 1996.

[22] Brandon Tran, Jerry Li, and Aleksander Madry. Spectral signatures in backdoor attacks[C]. *In Advances in Neural Information Proc-*

essing Systems, 2018: 8000-8010.

[23] Hitchcock F L. The Distribution of a Product from Several Sources to Numerous Localities[J]. *Journal of Mathematics and Physics*, 1941, 20(1/2/3/4): 224-230.

[24] Rizzo M L, Székely G J. Energy Distance[J]. *Wiley Interdisciplinary Reviews*: *Computational Statistics*, 2016, 8(1): 27-38.

[25] Krizhevsky A, Hinton G. Learning multiple layers of features from tiny images[J]. *Handbook of Systemic Autoimmune Diseases*, 2009, 1(4).

[26] Nicolae M I, Sinn M, Tran M N, et al. Adversarial Robustness Toolbox V1.0.0[EB/OL]. 2018: arXiv: 1807.01069. https://arxiv.org/abs/1807.01069

[27] Kurakin A, Goodfellow I, Bengio S. Adversarial Machine Learning at Scale[EB/OL]. 2016: arXiv: 1611.01236. https://arxiv.org/abs/1611.01236

[28] Moosavi-Dezfooli S M, Fawzi A, Frossard P. DeepFool: A Simple and Accurate Method to Fool Deep Neural Networks[C]. *2016 IEEE Conference on Computer Vision and Pattern Recognition*, 2016: 2574-2582.

[29] Kotyan S, Vargas D V. Evolving Robust Neural Architectures to Defend from Adversarial Attacks[EB/OL]. 2019: arXiv: 1906.11667. https://arxiv.org/abs/1906.11667

[30] Zantedeschi V, Nicolae M I, Rawat A. Efficient Defenses Against Adversarial Attacks[C]. *The 10th ACM Workshop on Artificial Intelligence and Security*, 2017: 39-49.

[31] Xu W L, Evans D, Qi Y J. Feature Squeezing: Detecting Adversarial Examples in Deep Neural Networks[EB/OL]. 2017: arXiv: 1704.01155. https://arxiv.org/abs/1704.01155

韩蒙 于 2017&2021 年在佐治亚州立大学&佐治亚理工大学计算机科学&工商管理专业获得博士&MBA 学位。现任浙江大学滨江研究院数智融合研究中心主任。研究领域为人工智能安全、生成内容安全。研究兴趣包括: 可信智能、生成智能。Email: mhan@zju.edu.cn

俞伟平 于 2011 年在浙江工商大学通信工程专业获得硕士学位。现任杭州涂鸦科技有限公司技术副总裁。研究领域为人工智能、物联网、计算机网络。研究兴趣包括: 智能物联网、计算机视觉、对抗样本。Email: ywp123@gmail.com

周依云 于 2020 年在 Kennesaw State University 数据科学专业获得博士学位。现任 CareerBuilder 科研团队的数据科学家。研究领域为推荐系统, 图论, 自然语言处理, 深度学习。研究兴趣包括: 对抗学习, 安全系统。Email: yiyunzhouds@icloud.com

杜文涛 于 2022 年在中国科学技术大学网络空间安全学院获得硕士学位。现任浙江大学滨江研究院助理研究员, 研究领域为隐私保护, 研究兴趣包括: 图神经网络和差分隐私。Email: duwentao@mail.ustc.edu.cn

孙彦斌 于 2016 年在哈尔滨工业大学计算机科学与技术专业获得博士学位。现任广州大学计算机科学与网络工程学院副教授, 研究领域为网络安全、工业控制系统安全、未来网络, 研究兴趣包括: 工业互联网安全。Email: sunyanbin@gzhu.edu.cn

林昶廷 于 2018 年在浙江大学计算机科学与技术专业获得博士学位。现任浙江大学滨江研究院副研究员。研究领域为网络空间安全、人工智能、物联网。研究兴趣包括: 人工智能安全和深度学习。Email: linchangting@zju.edu.cn