文章编号：2095-2163(2019)06-0111-08 中图分类号：TP391 文献标志码：A

信息安全领域中鲁棒的深度学习及其应用研究

王赛男

(江苏联合职业技术学院 南京工程分院，南京 211135)

摘 要：本文初步探索了深度学习模型脆弱性，存在的潜在原因之一归结于其网络结构中高度敏感的局部线性行为。而对抗性训练的提出，旨在对抗扰动的训练集上训练深度学习模型，是一种有效的正则化方法，可缓解其脆弱性问题。由于传统对抗性训练算法依赖于已知攻击算法，在抵御其攻击时性能十分有限，而基于特征掩膜(Feature Mask)和特征填补(Feature Padding)的对抗性训练防御策略的提出，不仅不依赖于对抗样本，还能提高深度学习模型的鲁棒性及安全性，并在公开交通标识识别和人脸识别数据集上，验证了所提对抗性训练防御策略在对抗环境下较优的防御性能。

关键词：深度学习；脆弱性；局部线性；对抗性训练；特征掩膜；特征填补

Robust Deep learning and its application in information security

WANG Sainan

(Nanjing Engineering Vocational College, Jiangsu Union Technical Institute, Nanjing 211135, China)

【Abstract】 This article has initially explored the vulnerability of deep learning models. One of the potential reasons for this is due to its highly sensitive local linear behavior in the network structure. The adversarial training proposed to train deep learning models against the perturbed training set is an effective regularization method that can alleviate its vulnerability. While traditional adversarial training algorithms rely on known attack algorithms, their performance is very limited when resisting them. Advance adversarial training defense strategies based on feature mask and feature padding without relying on adversarial examples are proposed in this paper to improve the robustness and security of the DL models, and verifies the better defense performance of our proposed adversarial training defense mechanisms on public traffic sign and face recognition datasets in the adversarial scenario.

【Key words】 deep learning; vulnerability; local linearity; adversarial training; feature mask; feature padding

0 引 言

目前人工智能和深度学习技术在信息安全领域得到了广泛应用，具体场景包括人脸识别、无人驾驶、垃圾邮件过滤、抵抗恶意代码攻击、网络攻击等[1-4]。然而在对抗环境中，可能潜在一些恶意的攻击者试图挖掘深度学习模型的脆弱性，并设计出相应的攻击算法来恶意篡改输入样本，从而降低深度学习模型的性能[5-12]。随着语音、图像作为新兴的人机输入手段，其便捷和实用性被大众所欢迎。同时随着移动设备的普及，以及移动设备对这些新兴的输入手段的集成，使得这项技术被大多数人所亲身体验。然而，语音、图像识别的准确性对机器理解并执行用户指令的有效性至关重要，往往这一过程也是最容易被攻击者利用。攻击者通过对原始输入样本进行细微地修改，达到用户感知不到，而机器接受了该类样本之后却会以较大概率做出错误的判断，从而导致计算设备被入侵，错误命令被执行以及执行后的连锁反应造成的严重后果。

图 1 展示了攻击者通过在正常输入样本(第一行图像)中注入细微的、精心设计的对抗扰动，生成不易察觉的对抗样本(第二行图像)，来紊乱 Resnet50[13] 和 FaceNet[1] 等经典深度学习模型，使其产生错误的输出结果(识别结果在每行图像下方)。目前主流的对抗扰动生成算法主要以梯度计算为主，例如，快速梯度符号方法(Fast Gradient Sign Method, FGSM)[5] 以及基于 FGSM 改进的 I-FGSM[5]。R+FGSM[14] 通过计算目标模型损失函数的梯度，并沿着梯度方向寻找对抗扰动来构造对抗样本。C&W 攻击算法[13] 通过采用梯度下降优化对抗目标函数来构造对抗样本。

针对对抗样本的存在，为了提高深度学习模型在安全敏感性相关任务中的鲁棒性，Goodfellow 等人首次发现深度学习模型的网络结构中存在着局部线性的现象，极易被攻击者利用，从而获取隐私信息。幸运的是，深度学习模型却不同于线性模型(如机器学习中的简单线性回归、逻辑回归等)纯粹的线性，可以通过在模型训练损失函数中添加正则

作者简介：王赛男(1980-)，女，硕士，讲师，主要研究方向：机器学习与模式识别。

收稿日期：2019-09-30

化项(即惩罚函数),来消除这种局部线性本质,从而达到局部区域恒定。基于该思想,对抗性训练(Adversarial Training, AT)[5]概念应运而生,旨在对抗扰动的训练集上训练深度模型,在不损失原有独立同分布测试集上准确率的同时,能够减缓深度学习模型脆弱性问题。然而,传统的对抗性训练思想过于依赖已知攻击算法来构造对抗样本,并注入到训练集中进行对抗性训练,从而导致防御不同类型的攻击时具有一定的局限性,即泛化能力较弱。例如采用由 FGSM 攻击算法构造的对抗样本,注入到正常样本中进行对抗性训练,得到的模型却无法防御由 C&W 构造的对抗样本。

图 1 正常样本(上)与对抗样本图像(下)

Fig. 1 Legitimate samples (top) and adversarial samples (bottom)

因此,本文提出了两类更为有效的基于特征掩膜(Feature Mask, FM)和特征填补(Feature Padding, FP)的对抗性训练防御策略。不仅不依赖于对抗样本,同时能够防御多样化的对抗样本,具有较好的泛化能力。由于防御者无法预知所有潜在的攻击样本,而 FM 和 FP 对抗性训练策略颠覆了传统对抗性训练思想,并非一定要构建对抗样本进行训练才可以抵御攻击。本文通过构建特征变换操作之后的样本进行训练,即通过模糊化输入样本来增加攻击者构建对抗样本的难度,从而实现抵御多样化攻击样本的能力。

1 相关工作

1.1 深度学习模型

深度神经网络(Deep Neural Network, DNN)作为最为常见的深度学习模型,其最基本的体系结构如图 2 所示。DNN 由输入层、隐藏层(包括卷积层、池化层、全连接层等)、Softmax 层及输出层等部分组成[16]。其中每一层通过使用 n 个带参函数的分层组合来拟合高维的输入 x,其建模函数可以形式化为:

$$F(w;x) = f_n(w_n, f_{n-1}(w_{n-1},\dots, f_2(w_2, f_1(w_1,x)))). \quad (1)$$

其中,每个函数 $\{f_i \mid f_i = \sigma(w_i, f_{i-1}), i \in [1,n]\}$ 表示每层神经元。这些神经元是由激活函数 σ 应用于前一层输入的加权表示,以生成新表示的基本计算单元。每一层由权重向量 w_i 参数化,从而影响每个神经元的激活。

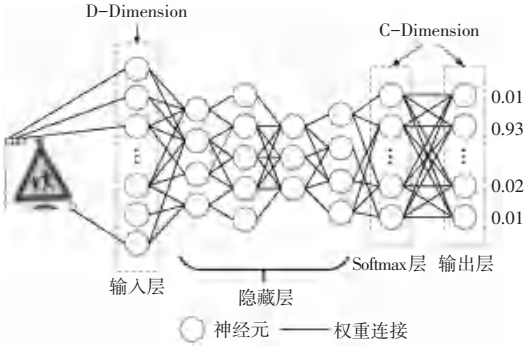

图 2 深度神经网络结构图

Fig. 2 Architecture of deep neural network

1.2 脆弱性分析

目前研究表明,对抗样本存在的潜在原因之一是由于深度学习模型欠拟合,由深度学习模型的局部线性性质导致[17]。从公式(1)中可以明显看出,深度神经网络主要是基于线性块构建,设计的目的主要便于优化模型训练所定义的损失函数。但是,当一个线性函数面临高维输入,那么其权重向量可能会面临严重倾斜的风险。

假设,攻击者精心设计了一个微小的扰动向量 r 来改变原始输入 x,那么权重向量为 w 的线性函数会产生 $r \parallel w \parallel_1$ 之多,如果 x 是高维的,那么该值将会是一个非常大的数。这也就意味着细微地修改却能从很大程度上影响深度神经网络的分类。

图 3 将有助于进一步理解线性本质带来的弊端。假设存在一个二分线性分类器,能够很好地拟合训练集。但这个超平面没有掌握训练集真正的结果,正类正常样本的分布明显是一个弧形,沿着弧线继续采样,却越过了超平面被误分;负类正常样本的分布也容易越过超平面被误分。线性模型在没有训练集出现的地方,做出的预测通常是有问题的,这一点是由线性模型的特点导致的。当一个数据点沿着一个固定方向,在训练集中移动,当移出到训练集分布之外的区域时,模型输出的变化方向也是不变的。进一步来说,在高维空间,每个像素值只需要非常小的改变,这些改变会通过和线性模型的参数进行点

乘累计造成很明显的变化。也就是说，只要方向正确，图像只要迈一小步，而在特征空间上就是一大步，就能很大程度地跨越决策层，从而迷惑模型的识别。

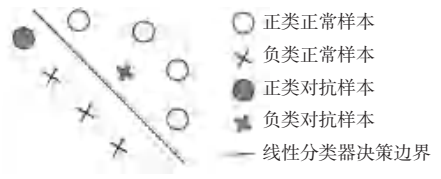

图 3　局部线性下的对抗样本实例

Fig. 3　Adversarial example under local linearity

因此，攻击者精心设计的细微且不易察觉的扰动向量即可影响整个深度神经网络判别的原因是欠拟合。由于输入空间维度过高，模型过于线性的结果，即深度神经网络高度敏感的局部线性行为。

1.3　对抗样本生成算法

由上一节所知，深度神经网络极大可能存在局部线性的特性，因其高度敏感的局部线性行为带来的潜在安全威胁，则一系列对抗样本生成算法应运而生。目前主流的对抗样本生成算法主要包括：基于梯度和基于优化的两大派别。其中基于梯度的以 FGSM[5] 及其改进的 I-FGSM[5]、R+FGSM[14] 最为典型，基于优化的以 C&W 算法[15] 性能最优。

1.3.1　快速梯度符号方法

快速梯度符号方法（Fast Gradient Sign Method，FGSM）[5] 首次由 Goodfellow 等人提出，因其可以快速生成对抗样本而著称。FGSM 通过计算目标模型损失函数的梯度，并沿着梯度方向寻找对抗扰动，然后添加到原始正常输入样本中来构造对抗样本。假设给定一个深度神经网络 $F(w;x)$，输入样本 x 及其对应的真实标签 y，FGSM 构建对抗样本过程如下：

$$x^* = x + \varepsilon \cdot \mathrm{sgn}(\tilde{\nabla}_x loss(y, F(w;x))), \quad (2)$$

其中，ε 用于控制生成的对抗样本攻击能力；sgn 表示符号函数，用于将向量中的每个维度值转到 $\{-1,0,1\}$ 范围；*loss* 函数表示深度神经网络预测值与真实值 y 之间的损失函数，一般深度神经网络主要以交叉熵损失函数[18] 应用居多，其定义如下：

$$loss(y, F(w;x)) = -\sum y \cdot \log(F(w;x)). \quad (3)$$

FGSM 是基于梯度攻击算法的典型，该算法尽管能够快速生成大量对抗样本，却无法保证所有对抗样本都行之有效。其根本原因在于，沿着梯度方向寻找扰动不一定会跨越模型决策边界或者过于跨过边界导致样本失真，极易被检测为对抗样本。

因此，为了提高对抗样本的攻击成功率，基于 FGSM 改进的迭代式算法 I-FGSM，以及引入随机噪声的 R+FGSM 等优化攻击算法被相继提出。该类算法虽提高了攻击成功率，又增加了样本失真的概率，即轻易被检测器检测为对抗样本的风险。

1.3.2　C&W 攻击方法

针对 FGSM 算法性能上的不足，Carlini 等人提出了一种基于优化的迭代算法。C&W 攻击算法（Carlini & Wagner Attack，C&W Attack）[13] 不仅提高了生成对抗样本的攻击成功率，同时避免了样本失真被检测器检测出来的风险。从模型决策边界角度分析，C&W 攻击算法生成的所有对抗样本相比于 FGSM 而言均分布在模型决策边界附近，即对正类样本构造的对抗样本均分布在负类样本一侧，负类样本反之。故而具有较高的攻击成功率，这是优化算法特有的能力。

C&W 攻击算法是一种基于 L-BFGS[9] 攻击算法优化改进的迭代算法，该算法通过辅助变量 ω 来寻找对抗扰动向量 r。

$$r = \frac{1}{2}(\tanh(\omega) + 1) - x, \quad (4)$$

其中，tanh 函数表示双曲正切函数，通过优化以下目标函数来获取 ω 值：

$$\min_{\omega} \left\| \frac{1}{2}(\tanh(\omega) + 1) - x \right\|_2^2 + c \cdot f\left(\frac{1}{2}(\tanh(\omega) + 1)\right), \quad (5)$$

f 函数定义如下：

$$f(x) = \max(Z(x)_y - \max\{Z(x)_i : i \neq y\}, -k). \quad (6)$$

其中，$Z(x)_i$ 表示神经网络 softmax 前一层类别 i 对应的输出，k 用于控制攻击类别标记与真实类别标记之间的置信度差值（即强度），等效于公式（2）中的 ε 值。k 值越大，攻击样本被错误分类的可能性越大。

2　对抗性训练防御策略

2.1　局部区域恒定

对抗性训练（Adversarial Training，AT）[5] 是防御对抗样本攻击的一种有效正则化方法，通过将对抗样本和正常样本一起训练，不仅可以提高模型的准确度，同时也能有效降低对抗样本的攻击成功率。

对抗性训练通过激励神经网络在训练数据附近的局部区域,恒定来限制深度神经网络高度敏感的局部线性行为,如图 4 所示。由于其被限制为线性而无法抵抗对抗样本,而神经网络能够将函数从接近线性转化为局部区域恒定,从而可以灵活地捕获到训练数据中的线性趋势,同时学习抵抗局部扰动。

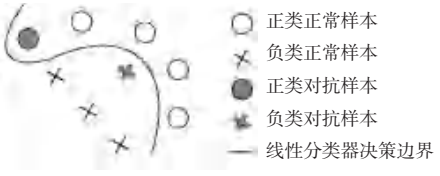

图 4 对抗性训练下的局部区域恒定

Fig. 4 Nearly locally constant under adversarial training

然而,传统对抗性训练得到的模型泛化性能较弱,即依赖于已知的攻击算法,在抵御其它对抗样本时性能不佳。例如使用 FGSM 攻击算法构建对抗样本进行对抗性训练,得到的模型却无法抵御 C&W 算法构造的对抗样本。由于防御者知识是有限的,无法预知所有潜在的攻击样本,而 FM 和 FP 对抗性训练防御策略的提出,颠覆了传统对抗性训练思想。并非一定要构建对抗样本进行训练才可以抵御攻击,也可以对原始输入样本进行特征变换来迷惑攻击者,使之无法针对性地设计对抗样本,从而增加构建对抗样本的难度。基于该思想,本文对原始输入样本采取了特征掩膜与特征填补两种特征变换操作。通过模糊化输入样本后,再同正常样本一起做对抗性训练,在保证不损失模型精度的情况下,增强抵御多样化攻击样本的能力。

2.2 基于特征掩膜的对抗性训练

特征掩膜(Feature Mask, FM)是一种常用的特征变换操作,且在语义分割、目标检测等领域拥有着广泛应用,例如经典的目标检测模型 Mask R-CNN[18]。Mask 实现机制如图 5 所示。假设存在一张 3×3 维的原始图像,通过与自定义的 Mask 矩阵进行点乘,即对原图中的每个像素和 Mask 矩阵中的每个对应元素做哈达马内积(Hadamard product)[19],从而得到用户所需的 Mask 变换图。该 Mask 矩阵的设计决定了最终变换图的效果,同时也决定了对抗性训练 DNN 的性能。

基于 FM 的对抗性训练策略,通过以固定比例(即图 4 中 Mask 矩阵中 0 的个数)随机废除部分特征来得到新的样本进行对抗性训练。图 6 展示了以 30%的随机废除率进行特征掩膜。其中为了方便理解并未体现随机性,实际实验中是对原始 $W \times H$ 维特征进行随机废除。为了不影响最终 DNN 模型在测试集上的识别精度,一般随机废除率不宜超过 50%,否则会大大降低模型准确率。

由于随机性的引入,对于攻击者而言,很难猜测到原始输入图像是如何变换的,从一定程度上增加了探索性攻击(Exploratory Attack)[20]的难度,即直接修改原始输入样本进行攻击。

图 5 特征掩膜机制

Fig .5 Feature mask mechanism

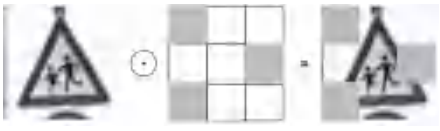

图 6 特征掩膜变换实例

Fig. 6 Feature mask transformation example

2.3 基于特征填补的对抗性训练

特征填补(Feature Padding, FP)是另一种普遍使用的特征变换操作,其初始被熟知是卷积神经网络(Convolution Neural Network, CNN)的问世,目前主流的图像识别模型(例如 VGG, ResNet, GoogleNet 等)都来源于 CNN。对原始图像进行卷积,操作前后会发生维度的缩减,若用户并不需要每一次都进行降维操作,就需要采用 FP 操作进行补 0、边界复制填补、镜像填补、块填补等诸多方式扩充到原始维度,其实现机制如图 7 所示。

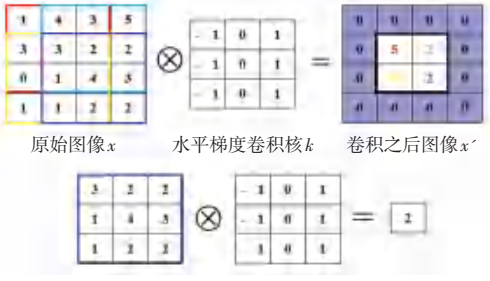

图 7 特征填补机制

Fig. 7 Feature padding mechanism

假设存在一张 4×4 维的原始图像 x, 通过与自定义的 3×3 水平梯度卷积核 k 进行卷积⊗操作,并以步长为 1 向右滑动窗口,最终将得到一张压缩为 2×2 的

图像，为了获得与初始维度一致的图像 x'，需要对其进行 FP 操作，如图 7 中所示。本文采用补 0 的方式进行特征维度扩充。其中⊗操作定义如下：

$$\otimes = \sum x_{ij} \times k_{ij}. \tag{7}$$

基于 FP 的对抗性训练策略设计理念来源于卷积运算中的填补机制，与 FM 随机废除特征的不同在于该方法以固定比例随机位置扩充特征维度 $W \times H$ 到同一维度 $W' \times H'$，从而得到多个新样本，并从中随机选择一张作为变换结果进行对抗性训练，其实现过程如图 8 所示。

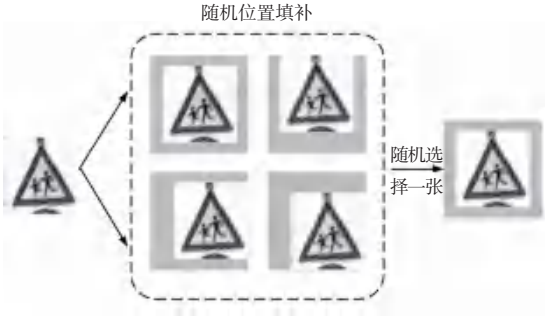

图 8　特征填补变换实例

Fig .8　Feature padding transformation example

该策略设计初衷同 FM 一致，通过引入随机性，在不影响模型识别精度的情况下，增加攻击者直接对输入样本修改的难度，从而达到更好地防御效果。同样，为了避免模型精度过多的损失，建议填补比例控制在 $[W, W+2]$，$[H, H+2]$ 范围内。

3　实　验

3.1　实验设置

无人驾驶工程和人脸识别是信息安全领域重要的两个应用场景。本文在公开的交通标识识别数据集(GTSRB, Belgium)和人脸识别(ORL)数据集上进行了数据扩充与预处理操作，特征维度压缩到 32×32，然后对提出的特征掩膜 FM 和特征填补 FP 对抗性训练算法验证其在对抗环境下的防御性能，其详细信息见表 1。

表 1　实验数据集总揽表

Tab. 1　The datasets used in experiment

名称	特征维度	标签数	任务
GTSRB	32×32	43	交通标识识别
Belgium	32×32	62	交通标识识别
ORL	32×32	40	人脸识别
YALE	32×32	10	人脸识别

针对不同任务，将选用不同的深度学习模型。在交通标识识别数据集中，本文自定义了 4 个 DNN 模型用于对抗性训练，其体系结构见表 2。而在人脸识别应用中，选用了 2 个比较简单的人脸识别模型(MTCNN[21] 和 FaceNet[1])进行对抗性训练测试。

表 2　不同深度神经网络体系结构

Tab. 2　Different deep neural network architectures

DNN_1	DNN_2
Dropout(0.2)	Conv(128,3,3)+Relu
Conv(64,8,8)+Relu	Conv(64,3,3)+Relu
Conv(128,6,6)+Relu	Dropout(0.25)
Conv(128,5,5)+Relu	FC(128)+Relu
Dropout(0.5)	Dropout(0.5)
FC(43)+Softmax	FC(43)+Softmax
DNN_3	DNN_4
S.Dropout(0.2)	Conv(128,3,3)+Relu
Conv(64,8,8)+Relu	Conv(64,3,3)+Relu
Conv(128,6,6)+Relu	Dropout(0.25)
Conv(128,5,5)+Relu	FC(128)+Relu
Dropout(0.5)	Dropout(0.5)
FC(43)+Softmax	FC(43)+Softmax

3.2　实验结果分析

首先，使用表 1 定义的不同深度神经网络结构及人脸识别模型 MTCNN 和 FaceNet，分别在对应数据集上进行模型训练，然后使用 FGSM 和 C&W 等攻击算法来构造对抗样本，去验证深度学习模型的脆弱性确实存在。针对两个任务训练得到的模型测试准确率及其在对抗样本下的测试准确率见表 3。其中 No-attack 表示在纯净测试样本上的测试准确率。FGSM 和 C&W Attack 两行表示由该两种攻击算法对纯净测试样本进行对抗样本构造。

实验结果显示，拥有高测试准确率的深度神经网络，在对抗环境中是非常脆弱的，极易受到对抗样本的影响。在由 FGSM 和 C&W 攻击算法生成的对抗样本下，各模型的识别准确率发生了大幅度的降低，尤其是基于优化的 C&W 攻击算法，获得了较高的攻击成功率。

为了验证传统对抗性训练算法在抵御多样性攻击样本性能中的不足，本文采用了 C&W 攻击算法

来构造对抗样本进行对抗性训练,并对得到的模型在不同的攻击样本上进行鲁棒性测试。实验结果见表4,其中每一列 DNN+C&W 表示由 C&W 攻击算法构造对抗样本并注入到原始训练样本中进行 DNN 训练。实验结果表明,由 C&W 构造的对抗样本进行对抗性训练,得到的模型在抵御 C&W 攻击时较为鲁棒,而在抵御由 FGSM 及其改进算法生成的对抗样本却显得力不从心。这也充分证明,传统依赖已知攻击算法构造的对抗样本进行对抗性训练,在抵御其它攻击算法时性能十分有限,通用性较差。

表3 深度学习模型脆弱性分析

Tab. 3 The analysis of the vulnerability of deep learning models %

	GTSRB				Belgium				ORL		YALE	
	DNN_1	DNN_2	DNN_3	DNN_4	DNN_1	DNN_2	DNN_3	DNN_4	MTCNN	FaceNet	MTCNN	FaceNet
No-attack	98.30	97.59	99.10	89.00	95.00	93.10	95.50	87.90	95.5	93.8	98.5	98.8
FGSM	44.80	48.45	50.25	33.10	45.10	50.25	52.00	30.00	30.2	45.5	45.0	50.5
C&W Attack	2.00	0.88	3.70	0	1.00	0.38	0.50	0.50	0	0.5	0	0

表4 基于 C&W 对抗样本的对抗性训练实验结果

Tab. 4 Results of adversarial training based on C&W adversarial examples %

	GTSRB				Belgium				ORL		YALE	
	DNN_1+ C&W	DNN_2+ C&W	DNN_3+ C&W	DNN_4+ C&W	DNN_1+ C&W	DNN_2+ C&W	DNN_3+ C&W	DNN_4+ C&W	MTCNN+ C&W	FaceNet+ C&W	MTCNN+ C&W	FaceNet+ C&W
No-attack	98.50	98.00	99.15	90.20	97.50	99.00	98.50	91.00	96.00	97.50	90.50	91.02
FGSM	84.00	88.50	86.55	73.50	80.05	81.00	83.00	71.00	64.00	68.55	70.50	67.00
I-FGSM	70.50	65.00	70.15	66.80	66.20	62.50	56.15	50.80	45.20	55.20	58.95	55.10
R-FGSM	73.28	68.50	72.50	68.00	63.8	60.50	62.50	58.00	53.21	58.05	63.25	60.50
C&W Attack	92.00	90.58	93.45	90.00	89.50	87.58	90.05	87.50	90.50	89.00	90.50	91.08

为了验证本文提出的两类基于 FM 和 FP 对抗性防御策略在抵御多样化攻击算法的有效性,本文进一步对比了在两种对抗性训练防御策略下,训练得到的目标模型在抵御其它攻击样本时的性能。

基于特征掩膜对抗性训练所得模型,在交通标识识别和人脸识别任务中的测试准确率见表5。首先,从测试准确率可以看出,随着废除率的增加,对抗性训练所得模型的测试准确率没有明显的下降;其次测试了 DNN_1 模型在不同攻击策略下抵御对抗样本的能力。实验结果表明基于特征掩膜的对抗性训练可以抵御多样化的对抗样本,相比于传统对抗性训练思想,通用性得到了一定程度的提高。

表5 基于特征掩膜的对抗性训练实验结果

Tab. 5 The results of adversarial training based on feature mask %

废除率	DNN_1+FM(GTSRB)					MTCNN+FM(ORZ)				
	No-attack	FGSM	I-FGSM	R-FGSM	C&W Attack	No-attack	FGSM	I-FGSM	R-FGSM	C&W Attack
10	98.17	93.00	89.50	91.00	90.50	98.78	92.89	90.00	90.10	89.90
20	98.09	91.20	89.88	90.25	90.30	97.50	90.50	88.00	89.55	91.00
30	97.89	90.50	90.58	91.28	92.10	98.09	91.00	90.01	90.05	90.50
40	97.53	89.90	90.00	89.50	89.50	97.88	89.00	89.98	88.59	90.25

基于特征填补的对抗性训练实验结果见表 6。首先，从测试准确率可以看出随机位置变化控制在一定范围内，不会大幅度降低模型的精度；其次在抵御多样化攻击时，模型仍然具有较好的鲁棒性。

表 6　基于特征填补的对抗性训练实验结果

Tab. 6　The results of adversarial training based on feacture padding

	GTSRB				MTCNN+FP	FaceNet+FP
	DNN_1+FP	DNN_2+FP	DNN_3+FP	DNN_4+FP		
No-attack	97.80	98.50	98.05	89.25	98.12	98.55
FGSM	90.50	89.20	88.28	85.89	90.10	92.35
C&W Attack	90.00	91.52	91.55	89.50	89.50	89.00

总体来看，基于特征掩膜与特征填补的对抗性训练所得模型可以抵御多样化的攻击，通用性更佳。

4　结束语

本文揭露了深度学习模型的脆弱性，即极易受到对抗样本的影响，尤其是 C&W 攻击算法，并合理分析了深度学习模型脆弱性存在的潜在原因是 DNN 高度敏感的局部线性行为。为了解决深度学习模型脆弱性问题，即提高深度学习模型在对抗环境中的鲁棒性及安全性。本文借鉴了 Mask 机制和卷积特征填补思想，颠覆了传统对抗性训练算法过于依赖已知攻击样本，而导致在抵御多样化攻击样本时的性能不佳，提出了基于 FM 和 FP 两种不依赖于对抗样本的对抗性训练防御策略，并在公开交通标识识别数据集 GTSRB、Belgium 和人脸识别数据集 ORL、YALE 上，验证了所提出的对抗性训练防御策略的优越性和较好的抵御多样化攻击的性能。

参考文献

[1] SCHROFF F, KALENICHENKO D, PHILBIN J. FaceNet: A Unified Embedding for Face Recognition and Clustering [C]. Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition, Boston, June 8-10, 2015. New York: IEEE Press, 2015.

[2] VISALIN S. Traffic Sign Recognition Using Convolution Neural Networks [J]. International Journal of Innovative Research in Computer and Communication Engineering, 2017, 5(6):11259-11266.

[3] SAAD O, DARWISH A, FARAI A. A survey of machine learning techniques for Spam filtering [J]. International Journal of Computer Science and Network Security, 2012, 12(2):66-73.

[4] TSAIA C F, HSU Y F, LIN C Y, LIN W Y. Intrusion detection by machine learning: A review [J]. Expert System with Applications, 2009, 36(10):11994-12000.

[5] GOODFELLOW I, SHLENS J, SZEGEDY C. Explaining and harnessing adversarial examples [C]. Proceedings of the 3rd International Conference on Learning Representations, San Diego, May 7-9, 2015. OpenReview.net, 2015.

[6] KURAKIN K, GOODFELLO I, BENGIO S. Adversarial examples in the physical world [C]. Proceedings of the 5th International Conference on Learning Representations, Toulon, April 24-26, 2017. OpenReview.net, 2017.

[7] BIGGIO B. Evasion attacks against machine learning at test time [C]. Proceedings of the 24th Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Prague, September 23-27, 2013. Berlin:Spring, 2013.

[8] PAPERNOT N, MCDANIEL P, JHA S, FREDRIKSON M, BERKAY CELIK Z, SWAMI A. The limitations of deep learning in adversarial settings [C]. Proceedings of the 1st European Symposium on Security and Privacy, Saarbrücken, March 21-24, 2016. New York: IEEE Press, 2016.

[9] SZEGEDY C, ZAREMBA W. Intriguing properties of neural networks[C]. Proceedings of the 2nd International Conference on Learning Representations, Banff, April 14-16, 2014. OpenReview.net, 2014.

[10] XU W L, QI Y J, DAVID E. Automatically Evading Classifiers A Case Study on PDF Malware Classifiers [C]. Proceedings of the 23rd Network and Distributed System Security Symposium, San Diego, February 21-24,2016. New York: IEEE Press, 2016.

[11] NEDIM S, PAVEL L. Practical Evasion of a Learning-Based Classifier: A Case Study [C]. Proceedings of the 35th IEEE Symposium on Security and Privacy, San Jose, May 18-21,2014. New York: IEEE Press, 2014.

[12] CORONA I, GIACINTO G, ROLI F. Adversarial attacks against intrusion detection systems: Taxonomy, solutions and open issues [J]. Information Sciences, 2013, 239: 201-225.

[13] He K M, ZHANG X Y, REN S Q, SUN J. Deep Residual Learning for image recognition [C]. Proceedings of the 2016 International Conference on Machine Learning, New York, June 20-22, 2016.

[14] Florian T, Alexey K, Papernot N, Goodfellow I, Boneh D, McDaniel P. Ensemble Adversarial Training: Attacks and Defenses [C]. Proceedings of the 6th International Conference on Learning Representations, Vancouver, Apr 30- May 3rd, 2018.

[15] CARLINI N, WAGNER D. Towards Evaluating the Robustness of Neural Networks [C]. Proceedings of the 38th IEEE Symposium on Security and Privacy, San Jose, May 22-24,2017. New York: IEEE Press, 2017.

（下转第 120 页）