高技术通讯 2024 年 第34卷 第9期:935-944

doi:10.3772/j.issn.1002-0470.2024.09.003

# 基于沙漏状数据处理单元和分组 RBF 单元的对抗性免疫防御方法①

丁伟杰②* ** 郑文浩* *** 方 怡* *** 王琦晖**** 李小薪③* ***

(* 浙江工业大学计算机科学与技术学院 杭州 310023)

(** 浙江警察学院计算机与信息安全系 杭州 310053)

(*** 浙江省可视媒体智能处理技术研究重点实验室 杭州 310023)

(**** 杭州师范大学钱江学院 杭州 311121)

**摘 要** 针对深度神经网络(DNN)容易受到对抗样本攻击的问题,研究人员提出了许多防御方法,可分为外部防御方法(EDM)和免疫防御方法(IDM)。外部防御方法试图在将对抗性样本输入 DNN 之前去除其中存在的对抗干扰,而免疫防御方法则致力于提升 DNN 本身的鲁棒性,本文重点研究免疫防御方法。现有的免疫防御方法主要基于鲁棒优化策略来提升 DNN 的鲁棒性,为 DNN 构建鲁棒模块的工作较少。本文在 DNN 中引入了 2 个新的鲁棒单元:基于特征压缩和精度注入的沙漏状数据处理单元,用以减小对抗性扰动的干扰;分组径向基函数单元,用于增强 DNN 的非线性和适应类内变化的能力。在优化过程中使用标签平滑、退火策略和权值衰减来进一步提高鲁棒性。在 2 个数据集(MNIST 和 CIFAR-10)以及 2 个流行的 DNN 模型(LeNet5 和 VGG16)上的实验表明,将所提出的鲁棒单元集成到 DNN 中可以大幅提高其对对抗性攻击的免疫能力,同时保持其在干净样本上的识别性能。

**关键词** 免疫防御;精度注入;分组径向基函数(RBF);权重衰减

近年来,深度神经网络(deep neural network,DNN)在图像识别领域取得了极大的进展。然而,随着对抗攻击方法的出现[1-8],DNN 的安全性受到了极大的挑战,由对抗攻击方法所产生的对抗样本容易导致 DNN 识别性能的崩溃。对抗攻击方法通过在输入图像中添加微小的、人眼难以察觉的扰动来生成对抗样本。对抗攻击方法的存在使得 DNN 模型容易出现安全漏洞,严重限制了其在有较高安全性要求领域的应用,如身份认证和自动驾驶。

为了增强 DNN 的鲁棒性,研究者们提出了许多防御方法。这些方法可分为外部防御方法(external defense methods,EDM)和免疫防御方法(immune defense methods,IDM)。EDM[9-19]通过对输入图像进行预处理来消除对抗干扰所造成的影响,从而达到防御攻击方法的目的。EDM 独立于目标 DNN 模型,只须训练防御模型本身,因此,可以很容易地应用于任意现有 DNN 模型的防御。然而,EDM 本质上是一种面向黑盒攻击的防御方法,即 EDM 假设防御模型本身的结构和权重对攻击者而言是未知的,其自身的安全性没有得到攻击方法的检验。

IDM 试图通过重新训练 DNN 模型或局部修改 DNN 模型的结构来增强 DNN 本身的鲁棒性。IDM 通常采用先进的优化策略,梯度掩蔽[9,20-22]和标签平滑[22-24]是 2 种广泛使用的优化策略。正如 Liao 等人[14]所指出的那样,这些基于优化的防御方法尽管可以使白盒攻击变得更加困难,但对黑盒攻击方

① 国家自然科学基金(62271448),浙江省自然科学基金(LGF22F020027),公安部理论及软科学计划(2022LL18),公安部科技强警基础工作专项(2020GABJC35)和教育部人文社会科学研究项目(22YJA840004)资助。

② 男,1981 年生,博士,副教授;研究方向:图像处理与模式识别;E-mail:dingweijie@zjjcxy.cn。

③ 通信作者,E-mail:mordecai@163.com。

(收稿日期:2023-05-09)

法[20-21]的防御能力仍然是较弱的。这里,白盒攻击是指攻击者能够获取目标模型的所有信息,包括网络模型的结构、参数以及训练数据集等;黑盒攻击是指攻击者无法获取目标网络模型的相关信息。黑盒攻击成功的难度远大于白盒攻击,但也更符合实际攻击情况。与鲁棒优化方法不同,对抗训练[1,20,25-28]通过将对抗性样本纳入训练数据来增强 DNN 模型的鲁棒性。严格来说,将对抗样本纳入训练的外部防御方法,如高阶表征指导的去噪器(high-level representation guided denoiser,HGD)[14]和 ComDefend[15]也属于对抗训练。对抗训练是基于大数据的训练方法,对于从未见过的攻击方法所产生的对抗样本的防御仍然较差,并且训练过程更为耗时[18,20]。

IDM 还致力于将新的鲁棒单元应用于 DNN 模型。Goodfellow 等人[1]指出 DNN 对对抗性扰动的脆弱性主要归因于其在高维空间中的线性行为。尽管 DNN 使用了许多非线性单元,如 ReLUs、Sigmoid 等激活函数,但为了易于优化,它们通常都被设计成以接近于线性模型的方式运行。在高维线性空间中,众多非常微小的扰动会被累加起来,从而引发“蝴蝶效应”,对网络的输出产生巨大影响。为了消除这种影响,可以在 DNN 模型中纳入非线性单元。最为广泛使用的非线性单元是径向基函数(radial basis function,RBF)[1,29-31]。尽管 RBF 单元增强了模型的鲁棒性,但存在 2 个问题。首先,与线性模型相比,RBF 单元对较为显著的类内变化的适应性较差,因此会弱化 DNN 模型的泛化能力[1];其次,使用 RBF 单元容易产生梯度消失问题。除了非线性单元外,Yu 等人[32]提出在 DNN 中添加多个辅助模块(AuxBlock)。然而,Yu 等人[32]假定辅助模块对攻击者是未知的。这使得基于 AuxBlock 的 DNN 模型更适合黑盒或灰盒攻击。Li 等人[29]建议将 RBF 单元和数据变换技术(如 EDM 中使用的特征压缩[12]和高斯噪声注入[9])结合起来增强 DNN 模型的鲁棒性。他们的模型旨在减少对抗性扰动,同时提高深度神经网络模型的非线性分类能力。然而,由于只使用了非常有限的数据变换以及基于受限的正则化策略优化基于 RBF 单元的 DNN 模型,他们的模型对对抗性扰动的鲁棒性仍然非常有限。

本文重点研究免疫防御模型的构建。鲁棒的 DNN 模型应该满足以下 3 个条件。首先,它应该是攻击不可知论[13];其次,DNN 模型的网络结构和参数可以被攻击者随意访问;最后,保持原始 DNN 在干净图像上的预测精度。

为了满足这些条件,本文建议在考虑 IDM 中现有的鲁棒模块的基础上,结合 EDM 和 IDM 中鲁棒优化策略的思想来构建新的鲁棒单元。受先进的 EDM 中广泛采用的自编码器结构[9,14-17,19]的启发,本文首先提出了沙漏状数据处理单元(沙漏单元)。沙漏单元由一个特征压缩模块[10,12]和一个精度注入模块组成,分别用于缩小和扩大给定数据的动态范围。减小取值范围可以减少对抗性扰动,而扩大取值范围可以破坏对抗性扰动。本文在输入层和所有隐藏层之后添加沙漏单元,以进一步缩小对抗性扰动对所有网络层的影响。

进一步增加非线性模块,即 RBF 单元,以提高 DNN 模型的鲁棒性。由于 RBF 单元处理类内变化的能力有限,本文引入分组 RBF 单元和标签平滑来解决这一问题。通过分组 RBF 单元,为每一类建立了多个类心,从而提高模型的泛化性能。本文还将标签平滑的优化策略[22-24]转化为防御网络的鲁棒元素。具体来说,本文将由 RBF 单元组成的 RBF 层附加在原始网络的 Softmax 层之后,而不是直接替换 Softmax 层。也就是说,本文让 RBF 层通过标签平滑来学习 Softmax 层的输出。本文还注意到,Softmax 网络与 RBF 网络有截然相反的特性:前者更容易优化,并且对类内变化具有鲁棒性。因此,通过使用标签平滑,本文可以将 Softmax 网络和 RBF 网络的优势结合起来。

在 2 个基线网络、3 种攻击方法和 2 个公共数据集上进行的大量实验表明,本文提出的模型可以有效地对抗白盒攻击和黑盒攻击中的对抗样本。

## 1 Softmax 网络和 RBF 网络

本节讨论 Softmax 网络和 RBF 网络。为简单起见,假设这 2 种网络没有隐藏层。令 $\boldsymbol{x} \in R^n$ 表示待识别图像,且其类别是 $C$ 个类别中的某一个。Soft-

max 网络按如下方式预测 $x$ 的类别:

$$\hat{y}_{\text{Softmax}}(\boldsymbol{x}) = \text{argmax}_c \boldsymbol{w}_c^{\text{T}} \boldsymbol{x} \tag{1}$$

其中,$\boldsymbol{w}_c \in R^n$ 是类 $c$ 的法向量。RBF 网络由 $C$ 个 RBF 单元组成,每个 RBF 单元定义如下:

$$\phi(\boldsymbol{x};\boldsymbol{\mu}_c,\sigma) \triangleq \exp\left(-\frac{\|\boldsymbol{x}-\boldsymbol{\mu}_c\|_2^2}{2\sigma^2}\right) \tag{2}$$

其中,$\boldsymbol{\mu}_c \in R^n$ 表示类 $C$ 的质心,$\sigma$ 表示标准差。RBF 网络的决策基于最近质心准则(nearest centroid criterion,NCC)

$$\hat{y}_{\text{RBF}}(\boldsymbol{x}) = \text{argmin}_c \frac{1}{2\sigma^2}\|x-\boldsymbol{\mu}_c\|_2^2 \tag{3}$$

Softmax 网络和 RBF 网络对加入到图像 $x$ 中的较小的扰动 $\eta \in R^n$ 的响应是不同的。由于高斯基函数主要关注类质心所在的局部区域,因此,$\eta$ 对 RBF 网络的最终输出仅有非常微弱的影响。然而,由于 Softmax 网络的线性特性,较小的扰动累加起来可能会对最终决策产生巨大的影响。

## 2 本文方法

本文提出的 DNN 结构如图 1 所示。与常见的 DNN 模型相比,增加了 2 个新单元:由特征压缩和精度注入组成的沙漏状数据处理单元以及分组 RBF 单元。接下来将深入讨论这 2 个单元。

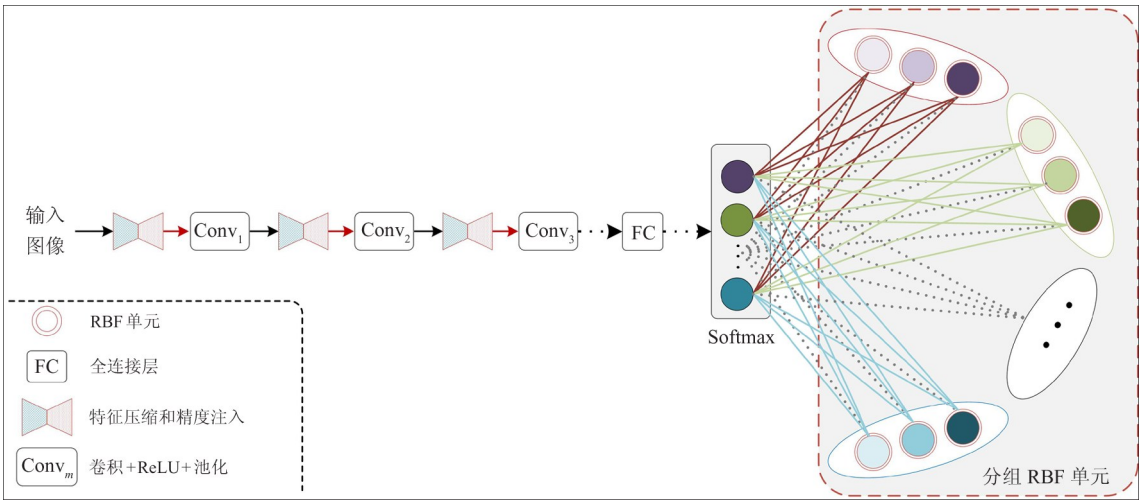


**图 1 本文所提出的 DNN 结构**

### 2.1 沙漏状数据处理单元

正如 Goodfellow 等人[1]所指出的,对抗性扰动主要位于输入数据的精度边界上。因此,消除对抗性扰动的自然方法是降低输入精度。然而,精度收缩也牺牲了输入图像中的一些高频信息,例如纹理和边缘。为了克服这个问题,许多现有的预处理(即外部防御)方法[9,14-17,19]采用沙漏结构来净化对抗样本,例如自编码器[33]和 U-Net[34]。沙漏结构由 1 个编码器和 1 个解码器组成。编码器将输入图像压缩到低维空间,而解码器则逐步重建输入图像。尽管这些沙漏状结构网络在净化对抗性扰动方面表现良好,但它们对攻击者是不可见的,这违反了白盒攻击中攻击者可以自由获取网络结构和网络参数的前提条件。因此,本文将沙漏状数据处理单元放入 DNN 模型内部,以便它们可以被攻击者自由访问,以充分验证其鲁棒性。然而,直接使用现有的沙漏结构网络会导致过高的计算代价,本文提出了一种轻量化的沙漏结构。具体来说,首先使用特征压缩[10,12,35]来降低输入精度,然后使用精度注入来抵消对抗干扰所造成的影响[1,13]。

(1)特征压缩。Xu 等人[12]提出了 3 种典型的特征压缩方法:位深度压缩、局部中值滤波和全局中值滤波,并验证了它们在不同数据集上对于不同攻击方法的性能。为了动态选择最优的攻击方法,Li 等人[29]建议通过度量输入图像/特征图的信息熵[36]来自动选择特征压缩方法。具体来说,用 $Q$ 表示输入数据的信息熵,若 $Q$ 小于某个阈值 $\tau$ 则意味着输入数据的信息主要位于低频区域,可以用较低

的颜色深度来表示;反之,若 $Q > \tau$ 则图像意味着输入数据的信息可能同时覆盖了高频区域和低频区域。在此情况下,为了保持信息的多样性,可以采用中值滤波。本文只使用局部中值滤波,因为它的平均性能优于全局中值滤波[12]。值得注意的是,本文选择的特征压缩方案与 Li 等人[29]的不同:当 $Q > \tau$ 时,本文采用中值滤波,而 Li 等人[29]是直接使用输入数据。由于中值滤波是一种非线性变换,在更强的攻击下,它比位深度压缩方法的防御效果更好。

(2)精度注入。Gu 和 Rigazio[9]使用高斯噪声注入(Gaussian noise injection,GNI)将对抗性输入移动到分类网络的"盲点"之外。然而,高斯噪声可能不是最佳的注入方式。由于对抗性扰动主要位于精度边界 $\pm\epsilon$ 上[1],因此本文使用来自于精度邻域的噪声 $n$ 注入输入数据。这里,精度邻域定义为 $[-\epsilon, -\epsilon+\iota]$ 和 $[\epsilon-\iota, \epsilon]$,$\iota$ 是一个非常接近 $\epsilon$ 的常数。将这一噪声注入方法称为"精度注入"。注意,这里的精度 $\pm\epsilon$ 表示的是数据在经过压缩处理后的精度。此外,本文的精度注入与 Goodfellow 等人[1]所采用的噪声注入策略不同,Goodfellow 等人[1]从集合 $\{-\epsilon, +\epsilon\}$ 中均匀地提取噪声,并且仅在模型训练期间使用这一策略用于数据增强,而本文是将噪声注入策略嵌入网络模型内部,可以认为是一种在线的噪声注入策略。在线噪声注入有如下 2 个优点:(1)它模拟对抗样本的扰动行为,以帮助在训练过程中构建鲁棒模型;(2)增强了网络在预测阶段的不确定性,从而提高了攻击的难度。

如图 1 所示,特征压缩和精度注入的沙漏结构附加在输入层和所有卷积层之后。正如 Gu 和 Rigazio[9]所指出的那样,与仅在输入层进行数据转换相比,该策略能能使网络具有更好的防御性能。

### 2.2　分组 RBF 单元

RBF 网络的非线性性质使其对对抗性攻击具有天然的鲁棒性[1]。然而,RBF 网络难以优化,并且对类内变化的适应能力较差。与 RBF 网络相比,Softmax 网络的性能正好相反,易于优化且对类内变化是鲁棒的,但易受对抗扰动的影响。因此,本文将 Softmax 网络和 RBF 网络结合起来,即将 RBF 网络附加在网络的 Softmax 层之后,如图 1 所示。将 RBF 附加在 Softmax 之后,除了利用 2 种网络的互补特性外,在本质上也是一种标签平滑技术。此外,为了进一步提升 RBF 网络的鲁棒性,本文进一步提出了分组 RBF 单元。

(1)标签平滑。标签平滑[22-24]对于构建免疫防御系统非常重要。数据的类别标签一般是离散的,然而,离散的标签不能表示类之间的关联性。因此,Papernot 等人[22]提出知识蒸馏法:首先,构建教师网络学习数据各类别之间的关系;然后,构建学生网络,由教师网络所给出的平滑标签进行网络权重学习。然而,现有文献中所使用的教师网络是独立于学生网络的,本文的学生网络(即 RBF 网络)是直接追加在教师网络之后的,它们一起工作。

(2)分组 RBF 单元。如式(3)所示,RBF 网络的决策主要依赖于 RBF 单元所提供的类的质心。在现实中,来自同一类的图像通常具有一定的平移、旋转和光照变化等。因此,如果 RBF 分类器的每个类别只有一个质心,则很难对类内变化具有鲁棒性。因此,本文为每个类使用一组 RBF 单元。假设每个类别都具有 $\alpha(\alpha \geqslant 1)$ 个质心。如图 1 所示,可以为整个 RBF 层构建 $\alpha C$ 个 RBF 单元,每 $\alpha$ 个质心可以用来构建其所代表的类的流形子空间。对于基于分组 RBF 单元结构的 RBF 网络来说,其决策规则可以基于 K 最近质心准则(K-nearest centroid criterion,K-NCC),即首先搜索目标图像的前 $K$ 个最近质心,然后通过投票做出最终决策。

标签平滑和分组 RBF 单元共同协作,不仅可以抵抗扰动,还可以保证网络对干净数据的预测精度与原始网络的预测精度一致。Vidnerova 和 Neruda[30]也研究了 RBF 和 Softmax 的组合,所提出的网络称为 DNRBF,在 RBF 层之后附加了另一个 Softmax 层,但该 Softmax 层的存在仅仅是为了便于优化,却削弱了 RBF 层所创建的非线性鲁棒性。

### 2.3　损失函数与优化

本文将 DNN 模型的整个参数集 $\boldsymbol{\Theta}$ 分成 $\boldsymbol{\Theta}_{\mathrm{DNN}}$ 和 $\boldsymbol{\Theta}_{\mathrm{RBF}}$ 这 2 个子集。具体来说,$\boldsymbol{\Theta}_{\mathrm{DNN}}$ 表示 RBF 层之前的参数集合,$\boldsymbol{\Theta}_{\mathrm{RBF}} \triangleq \{\boldsymbol{\mu}_c^{(k)}\}_{\substack{k=1,\cdots,\alpha \\ c=1,\cdots,C}}$ 表示 RBF 层的参数,这里 $\boldsymbol{\mu}_c^{(k)}$ 表示类 $c$ 的第 $k$ 个质心。令 $d_c^{(k)}(v) \triangleq 1/(2\sigma^2) \cdot \| v - \boldsymbol{\mu}_c^{(k)} \|_2^2$。对于标签为 $y$ 的输入图

像 $x$, 用 $\hat{y}$ 表示其在 Softmax 层的输出: $\hat{y} = f_{\mathrm{DNN}}(x; \boldsymbol{\Theta}_{\mathrm{DNN}})$。

网络的总体损失是 RBF 损失和 Softmax 损失之间的平衡,定义如下:

$$J(\hat{y}, y, \boldsymbol{\Theta}) = J_{\mathrm{RBF}}(\hat{y}, y, \boldsymbol{\Theta}_{\mathrm{RBF}}) + \omega \cdot J_{\mathrm{Softmax}}(\hat{y}, y, \boldsymbol{\Theta}_{\mathrm{DNN}}) + \lambda \cdot \| \boldsymbol{\Theta} \|_2 \tag{4}$$

其中, $\omega$ 和 $\lambda$ 是 2 个正则化参数, $J_{\mathrm{Softmax}}$ 表示交叉熵损失, $J_{\mathrm{RBF}}$ 具体定义如下[31]:

$$J_{\mathrm{RBF}}(\hat{y}, y, \boldsymbol{\Theta}_{\mathrm{RBF}}) = \sum_{k=1}^{\alpha} \left( d_{y_x}^{(k)}(\hat{y}) + \sum_{c \neq y_x} \log(1 + \exp(\lambda - \hat{d}_c^{(k)}(\hat{y}))) \right) \tag{5}$$

其中 $\lambda > 0$, 并且 $y_x$ 是输入图像 $x$ 的正确类。注意,为了避免梯度消失,RBF 损失式(5)基于度量学习定义[37]:其中的第 1 项旨在减小与正确类的距离,第 2 项基于铰链损失(hinge loss),是为了增大与错误类别之间的距离。

为了实现良好的性能,优化策略非常重要[38]。本文主要考虑了以下 2 个因素:正则化参数 $\omega$ 和权重衰减系数 $\lambda$。受 Liu 等人[39]工作的启发,本文在优化过程中使用余弦退火策略: $\omega$ 首先被设置为较大的值,如 0.6,以加快初始阶段的训练速度,然后使用余弦退火[40]来收缩 $\omega$ 的值以增强 RBF 损失的作用。参数的权重衰减以及 L2 正则化[38,41]对于提高鲁棒性也非常重要。Goodfellow 等人[1]指出了对抗训练和权重衰减之间的联系,并建议使用较小的权重衰减系数,如.002 5。然而,他们发现较小的权重衰减系数并不会增强模型的鲁棒性。然而,当权重衰减策略与分组 RBF 网络相结合时,情况发生了变化,较小的 $\lambda$ 对于提升鲁棒性具有较为关键的作用。根据 Pang 等人[38]的建议,本文将权重衰减系数 $\lambda$ 设置为 $5 \times 10^{-4}$。

## 3 实 验

### 3.1 实验设置

为了验证本文方法的有效性,选取 MNIST[42]和 CIFAR-10[43]2 个公共数据集进行实验。对于 MNIST,以 LeNet 5[42]为目标网络,并选取 60 000 个样本作为训练集和 10 000 个样本作为测试集。对于 CIFAR-10,以 VGG 16[44]为目标网络,分别选取 50 000 张图像和 10 000 张图像作为训练集和测试集。

使用即快速梯度符号法(fast gradient sign method,FGSM)[1]、投影梯度下降方法(project gradient descent,PGD)[27]和$\mathrm{CW}_2$[3,45]等 3 种攻击方法生成对抗样本。FGSM[1]是基于梯度的攻击方法,通过最小化施加在输入图像上的加性扰动而得到对抗样本,由于这一扰动可以通过计算损失函数关于输入图像的梯度的正负符号而得到,因此称为梯度符号方法。FGSM 及其变体是基于梯度攻击的代表性方法。2015 年,Goodfellow 等人[1]提出了 FGSM。2016 年,Kurakin 等人[2]对 FGSM 方法进行了改进,并提出基础迭代方法(basic iterative method,BIM),该方法又被称为 I-FGSM(iterative FGSM)方法。I-FGSM 方法生成的对抗样本容易过拟合到局部极值点,从而影响对抗样本的迁移性。Madry 等人[27]提出的 PGD 是一种效果较好的基于梯度攻击的方法。PGD 也是 I-FGSM 方法的一种改进,通过加入一层随机化处理,增加了迭代次数,极大改善了攻击效果。

2017 年,Carlini 和 Wagner 等人[3,45]提出了一种新的对抗攻击方法,该方法被称为 CW 方法,是最有代表性的基于优化的攻击方法。CW 方法自定义了不同的目标函数,并通过实验数据选择最优的目标函数来实现对抗攻击;该方法可通过改变目标函数中的变量来增加最优解的空间大小,从而显著提高攻击成功率。

根据数据精度,将 FGSM 的扰动大小 $\epsilon$ 在 MNIST 和 CIFAR-10 上分别设置为 0.30 和 0.05;PGD 中的 2 个超参数 $(\beta, \epsilon)$ 在 MNIST 上设置为(0.1,0.3),在 CIFAR-10 上设置为(0.01,0.03)。$\mathrm{CW}_2$ 是指 CW 方法的 L2 攻击版本,其搜索步长和最大迭代次数分别设置为 5 和 1 000。

为清晰起见,相关方法所使用的主要超参数在攻击方法后面的括号中进行了标注,例如,FGSM(0.25)表示 FGSM 的扰动大小 $\epsilon$ 设置为 0.25。使用随机梯度下降(stochastic gradient descent,SGD)作为优化器。批量大小、初始学习率、初始 $\omega$、动量系数和权重衰减系数分别设置为 128、0.1、0.6、0.9 和 $5 \times 10^{-4}$。MNIST 和 CIFAR-10 的训练周期分别设

置为 20 和 200。用于估计输入图像的熵水平的阈值 $\tau$ 设置为 3。在每个训练周期后,使用余弦退火来降低 $\omega$ 和学习率。

### 3.2　消融实验

本节对本文所提出的各鲁棒单元的性能进行消融实验。除非特别说明,本节所报告的实验结果都是针对 MNIST 上的 LeNet5 的。

(1)不同策略对 $\omega$ 的影响。在网络优化过程中,可以采用 2 种策略来设置 $\omega$(见式(4))的值:固定 $\omega$ 的值和退火策略。图 2 比较了 $\omega = 0.2$ 和余弦退火 $\omega$(初始值设置为 $\omega_0 = 0.6$)2 种策略。显然,对于对抗样本的识别而言,退火策略相对于固定值策略的优越性是显著的。

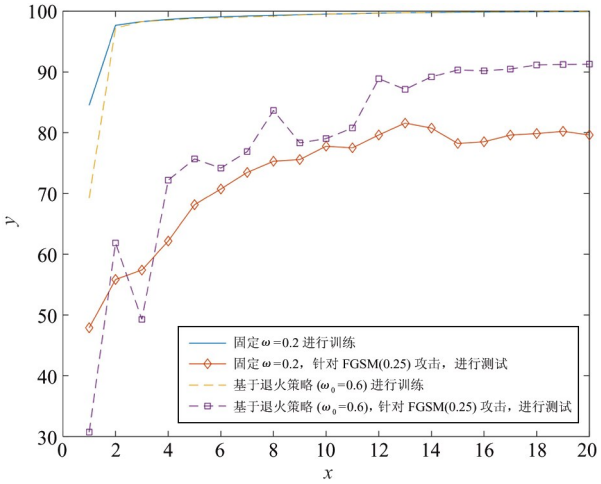


**图 2　定值策略和退火策略对 $\omega$ 的影响**

(2) $\alpha$ 的影响。用于定义 RBF 层的每个类的质心的数量 $\alpha$ 是导致非线性的重要因素。从理论上讲,更大的 $\alpha$ 意味着对类内变化的更大容忍度,也更有助于提升鲁棒性。如表 1 的第 1 行所示,在 MNIST 上使用 LeNet 5 进行性能测试,可以验证这一观点。然而,对于使用 CIFAR-10 数据集的 VGG 16 而言,情况就不同了,如表 1 的第 2 行所示,$\alpha = 3$ 时已导致网络的性能达到最优,较大的 $\alpha(\alpha > 3)$ 反而会使网络的性能有轻微的降低。在不同环境下 $\alpha$ 对网络性能的影响将留待未来进一步研究。

(3)特征压缩(FS)、精确注入(PI)、RBF 及其组合的效果。为简单起见,将 $\omega$ 固定为 0.2,使用 L1 表示仅在输入层使用 FS/PI,$L^*$ 表示在输入层以及所有隐藏层使用 FS/PI。比较高斯噪声注入(GNI)[9]和本文提出的 PI 方法的有效性,设置 GNI 的噪声水平为 0.3。表 2 报告了相关结果。可以看出,对对抗样本的识别性能随着所集成的鲁棒单元的增多而逐渐提升。其中,RBF 和 FS 是最重要的 2 个鲁棒单元。RBF 产生的结果显著优于现有文献[1,29]中的结果。这主要归因于 $\alpha$ 和 $\lambda$ 的设置。本文使用较大的 $\alpha$ 值($\alpha = 30$)和较小的权重衰减值 $\lambda$。较大的 $\alpha$ 增强了对类内样本变化的容忍度,而较小的 $\lambda$ 导致小的权重,从而抑制了对抗性扰动。

**表 1　在 Clean/FGSM 样本上的准确率/%**

| 网络模型 | 数据集 | $\alpha = 1$ | $\alpha = 3$ | $\alpha = 10$ | $\alpha = 30$ | $\alpha = 40$ |
|---|---|---|---|---|---|---|
| LeNet5 | MNIST | 81.18/17.64 | 97.22/61.51 | 99.04/63.45 | 99.08/69.51 | 99.10/70.32 |
| VGG16 | CIFAR10 | 46.19/ 9.56 | 93.50/46.72 | 93.54/45.53 | 93.50/45.75 | 93.50/46.57 |

**表 2　LeNet5 + (FS/PI/RBF) 对于干净和对抗样本的准确率/%**

| FS | GNI | PI | RBF | Clean | | FGSM(.25) | | PGD(.1,.3) | | CW2 | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | L1 | $L^*$ | L1 | $L^*$ | L1 | $L^*$ | L1 | $L^*$ |
| | | | | 99.04 | | 8.12 | | 0.74 | | 5.99 | |
| | | | √ | 99.08 | | 81.58 | | 75.53 | | 98.80 | |
| √ | | | | 98.89 | 99.01 | 89.43 | 90.21 | 85.59 | 86.72 | 84.92 | 84.95 |
| √ | | | √ | 99.05 | 99.07 | 92.61 | 92.75 | 89.74 | 90.03 | 98.85 | 98.97 |
| | √ | | √ | 99.02 | 99.03 | 65.81 | 67.72 | 59.14 | 62.35 | 63.24 | 64.53 |
| | | √ | √ | 99.04 | 99.05 | 66.08 | 68.15 | 60.03 | 62.67 | 63.35 | 64.66 |
| √ | √ | | √ | 99.05 | 99.07 | 93.01 | 93.53 | 90.23 | 90.35 | 98.70 | 98.75 |
| √ | | √ | √ | 99.05 | 99.10 | 93.51 | 93.75 | 90.45 | 90.52 | 98.91 | 98.93 |

与 RBF 和 FS 的显著效果相反,噪声注入的效果,即 GNI 和 PI 对对抗样本的影响是有限的,尽管 PI 优于 GNI。这主要是因为噪声注入不仅破坏了对抗性扰动所产生的不良结构,而且破坏了输入图像的原始结构。在表 2 的最后 2 行中看到,这种破坏可以在特征压缩空间中被大幅减少。

### 3.3 防御结果

通过与 5 个前沿方法的对比,进一步验证本文方法的有效性。具体来说,选择 ComDefend[15] 和 DIPDefend[19] 这 2 种外部防御方法以及 3 种免疫防御方法进行比较。3 种免疫防御方法主要包括:基于 PGD 的对抗训练(PGD-AT)[27]、AuxBlocks[32] 和 DT-RBF(数据转换加 RBF)[29]。在白盒攻击和黑盒攻击 2 种情形下验证各对比方法的防御能力。

(1)白盒攻击的结果。表 3 比较了所提出的方法和其他 5 种防御方法对 3 种白盒攻击的防御能力。可以看出,本文方法对 FGSM 具有最佳的防御能力,同时仍然保持了对干净样本的识别性能。这主要是因为本文方法所使用的鲁棒单元及鲁棒优化策略,如精度注入和优化过程中所使用的权值衰减策略,主要是针对 FGSM 的攻击特性的。此外,对于 PGD 和 $CW_2$ 这 2 种攻击方法,本文方法与前沿的外部防御方法 DIPDefend 几乎保持相同的精度。唯一的例外是,当在 CIFAR-10 上防御 $CW_2$ 时,本文方法显著优于大多数对比方法。其主要原因可能是在 CIFAR-10 上分组 RBF 单元只提供了有限的鲁棒性。从表 1 中可以看出,较大的 $\alpha$(主要是 RBF 单元组的鲁棒性)并不能提高网络在 CIFAR-10 上的性能。值得注意的是,大多数前沿方法,例如,ComDefend、DIPDefend 和 AuxBlocks 实际上对攻击者隐藏了部分防御模块,而 PGD-AT 需要建立在对攻击者所产生的对抗样本已知的基础上。这意味着,如果所比较的各方法的识别性能处于同一水平,则本文方法具有更强的防御能力。

**表 3 不同防御方法在白盒攻击下的准确率/%**

| | MNIST | | | | CIFAR-10 | | | |
|---|---|---|---|---|---|---|---|---|
| | Clean | FSGM | PGD | CW2 | Clean | FSGM | PGD | CW2 |
| DNN | 99.04 | 8.12 | 0.74 | 5.99 | 93.52 | 21.58 | 14.21 | 0.38 |
| ComDefend | 99.04 | 91.36 | 90.67 | 98.92 | 89.00 | 36.72 | 25.97 | 68.42 |
| DIPDefend | 99.04 | 92.75 | **90.84** | **98.95** | 90.00 | 49.45 | **55.32** | **72.53** |
| PGD-AT | 99.06 | 72.54 | 71.43 | 70.67 | 76.32 | 48.78 | 45.54 | 33.68 |
| AuxBlocks | 98.40 | 68.65 | 23.35 | 62.90 | 92.34 | 27.00 | 14.34 | 66.80 |
| DT-RBF | 98.72 | 91.00 | 89.86 | 98.70 | 92.75 | 64.19 | 52.27 | 41.44 |
| Proposed | **99.08** | **93.75** | 90.52 | 98.93 | **93.54** | **67.52** | 55.27 | 47.63 |

(2)黑盒攻击的结果。进一步比较本文方法与其他 5 个前沿方法对黑盒攻击的防御能力。黑盒攻击意味着用来制作对抗样本的源模型与目标模型不同。对于 CIFAR-10 和 MNIST,本文分别使用 ResNet 32/VGG 16 和 VGG 16/LeNet 5 作为源/目标模型。防御结果报告于表 4 中。显然,黑盒攻击的设置要公平得多:所有方法都对攻击者隐藏了它们的权重和结构。首先,由于攻击成本增加了,几乎所有的对比方法都比它们在白盒攻击下的识别性能更好。其次,由于基于更公平的攻击设置,本文方法优于几乎所有的对比方法,除了对 PGD(.01,.03)的防御效果弱于 DIPDefend。值得注意的是,当从白盒攻击转移到黑盒攻击时,DIPDefend 对 $CW_2$ 的防御性能下降了。正如 Dai 等人[19]所指出的那样,JSMA 方法也出现了类似的性能下降趋势,但并未给出深入的分析。

(3)沙漏单元的效果。沙漏单元是本文提出的防御对抗样本攻击的最重要的数据处理单元,通过对比使用沙漏单元前后特征图的变化可以观察沙漏单元的作用。具体地,可视化 LeNet5 在未使用/使用沙漏单元的情况下生成的特征图来评估沙漏单元的效果。图 3 绘制了白盒 FGSM 攻击下 MNIST 中 3 个数字 5、0、4 的特征图(第 2/第 3 列显示了 LeNet5 的 Conv2 在未使用/使用了沙漏单元的情况下生成

的特征图,矩形框表示相应特征图之间的主要差异)。不使用沙漏单元,LeNet5 分别错误地将对抗样本 5、0、4 识别为 3、2、9,而有沙漏单元的 LeNet5 可以成功识别这 3 个对抗数字。通过观察图 3 的第 2 列中原始 LeNet5 的特征映射,可以看到,在语义丰富的区域,对抗性扰动可以引起非常小的变化:一些重要的特征被削弱,而一些对抗性特征被增加。通过使用沙漏单元,如图 3 第 3 列所示,输入图像/特征图中的重要特征被残差连接保留,而对抗性特征被沙漏数据处理单元平滑或削弱。由于沙漏单元的作用,鲁棒化的 LeNet5 可以很好地识别这些对抗数字。

**表 4　不同防御方法在黑盒攻击下的准确率/%**

| | MNIST | | | | CIFAR-10 | | | |
|---|---|---|---|---|---|---|---|---|
| | Clean | FSGM | PGD | CW2 | Clean | FSGM | PGD | CW2 |
| DNN | 99.04 | 21.52 | 19.43 | 15.89 | 93.52 | 25.78 | 24.82 | 15.28 |
| ComDefend | 99.04 | 93.56 | 93.46 | 99.01 | 89.00 | 59.65 | 67.73 | 58.39 |
| DIPDefend | 99.04 | 95.84 | 92.35 | 99.03 | 90.00 | 67.57 | **70.55** | 57.88 |
| PGD-AT | 99.06 | 80.93 | 76.55 | 77.49 | 76.32 | 66.53 | 66.36 | 22.74 |
| AuxBlocks | 98.40 | 70.87 | 59.59 | 76.85 | 92.34 | 46.12 | 29.57 | 69.35 |
| DT-RBF | 98.72 | 92.31 | 90.78 | 99.02 | 92.75 | 78.32 | 52.72 | 55.89 |
| Proposed | **99.08** | **97.25** | **94.52** | **99.43** | **93.54** | **86.64** | 69.50 | **59.68** |

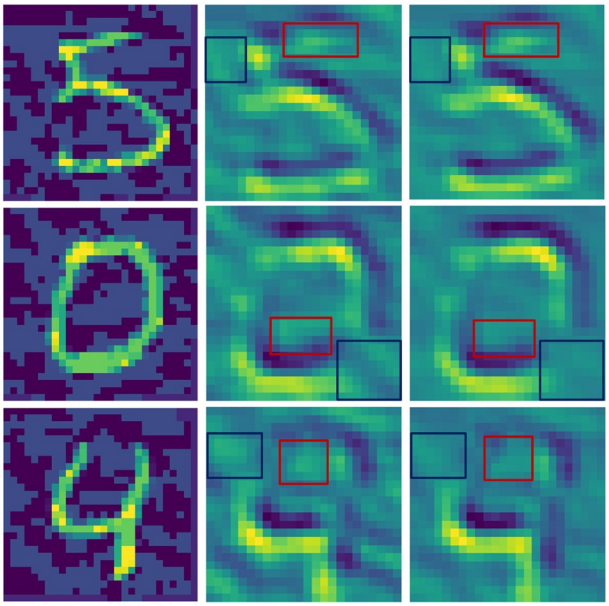

**图 3　白盒 FGSM 攻击下 MNIST 三位数 5、0、4 的特征映射对比**

## 4　结　论

本文专注于免疫防御方法(IDM)。通过添加 2 个新的鲁棒单元来增强 DNN:沙漏状数据处理单元和分组 RBF 单元。由特征压缩和精度注入组成的沙漏状单元用于减少对抗性扰动,而分组 RBF 单元用于增强非线性和处理类内变化。鲁棒单元和鲁棒优化策略之间存在一定的联系:权值衰减策略可以很好地与分组 RBF 单元协作。本文还成功地在训练过程中使用了退火策略来增强鲁棒性。这些鲁棒结构和优化策略未来将在更多的 DNN 模型和更多的数据集上进一步研究。此外,在分组 RBF 单元中使用更多类质心并不总是提高性能的原因也有待于进一步研究。

**参考文献**


[1] GOODFELLOW I J, SHLENS J, SZEGEDY C. Explaining and harnessing adversarial examples[C]//Proceedings of the International Conference on Learning Representations. San Diego, USA: ICLR, 2015:1-11.

[2] KURAKIN A, GOODFELLOW I, BENGIO S. Adversarial examples in the physical world[C]//Proceedings of the International Conference on Learning Representations. Toulon, France: ICLR, 2016:1-14.

[3] ATHALYE A, CARLINI N, WAGNER D. Obfuscated gradients give a false sense of security: circumventing defenses to adversarial examples[C]//Proceedings of the International Conference on Machine Learning. Stockholm, Sweden: PMLR, 2018:274-283.

[4] TOLOSANA R, VERA-RODRIGUEZ R, FIERREZ J, et al. Deepfakes and beyond: a survey of face manipulation and fake detection[J]. Information Fusion, 2020, 64:131-148.

[5] 潘文雯, 王新宇, 宋明黎, 等. 对抗样本生成技术综述[J]. 软件学报, 2020,31(1):67-81.
[6] 胡欣毅. 深度学习中的对抗样本机理分析及神经网络鲁棒性[D]. 南京: 东南大学, 2021.
[7] 张思思, 左信, 刘建伟. 深度学习中的对抗样本问题[J]. 计算机学报, 2019,42(8):1886-1904.
[8] 陈梦轩, 张振永, 纪守领, 等. 图像对抗样本研究综述[J]. 计算机科学, 2022,49(2):92-106.
[9] GU S, RIGAZIO L. Towards deep neural network architectures robust to adversarial examples[C]//Proceedings of the International Conference on Learning Representations. San Diego, USA: ICLR, 2015:1-9.
[10] OSADCHY M, HERNANDEZ-CASTRO J, GIBSON S, et al. No bot expects the deepcaptcha! introducing immutable adversarial examples, with applications to captcha generation[J]. IEEE Transactions on Information Forensics and Security, 2017,12(11):2640-2653.
[11] GUO C, RANA M, CISSE M, et al. Countering adversarial images using input transformations[C]//Proceedings of the International Conference on Learning Representations. Vancouver, Canada: ICLR, 2018:1-12.
[12] XU W, EVANS D, QI Y. Feature squeezing: detecting adversarial examples in deep neural networks[C]//Proceedings of the Annual Network and Distributed System Security Symposium. San Diego, USA: NDSS, 2018:1-15.
[13] SONG Y, KIM T, NOWOZIN S, et al. PixelDefend: leveraging generative models to understand and defend against adversarial examples[C]//Proceedings of the International Conference on Learning Representations. Vancouver, Canada: ICLR, 2018:1-20.
[14] LIAO F, LIANG M, DONG Y, et al. Defense against adversarial attacks using high-level representation guided denoiser[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE, 2018:1778-1787.
[15] JIA X, WEI X, CAO X, et al. Comdefend: an efficient image compression model to defend adversarial examples[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE, 2019:6084-6092.
[16] MENG D, CHEN H. MagNet: a two-pronged defense against adversarial examples[C]//Proceedings of the ACM SIGSAC Conference on Computer and Communications Security. Dallas, USA: ACM, 2017:135-147.
[17] SRINIVASAN V, ROHRER C, MARBAN A, et al. Robustifying models against adversarial attacks by langevin dynamics[J]. Neural Networks, 2021,137:1-17.
[18] QIU H, ZENG Y, ZHENG Q, et al. An efficient preprocessing-based approach to mitigate advanced adversarial attacks[J]. IEEE Transactions on Computers, 2024,73(3):645-655.
[19] DAI T, FENG Y, CHEN B, et al. Deep image prior based defense against adversarial examples[J]. Pattern Recognition, 2022,122:108-249.
[20] TRAMÈR F, KURAKIN A, PAPERNOT N, et al. Ensemble adversarial training: attacks and defenses[C]//Proceedings of the International Conference on Learning Representations. Vancouver, Canada: ICLR, 2018:1-20.
[21] PAPERNOT N, MCDANIEL P, GOODFELLOW I, et al. Practical black-box attacks against machine learning[C]//Proceedings of the ACM on Asia Conference on Computer and Communications Security. Abu Dhabi, Emirate: ACM, 2017:506-519.
[22] PAPERNOT N, MCDANIEL P, WU X, et al. Distillation as a defense to adversarial perturbations against deep neural networks[C]//Proceedings of the IEEE European Symposium on Security and Privacy. San Jose, USA: IEEE, 2016:582-597.
[23] WARDE-FARLEY D, GOODFELLOW I. 11 Adversarial perturbations of deep neural networks[J]. Perturbations, Optimization, and Statistics, 2016,311:5-12.
[24] HAZAN T, PAPANDREOU G, TARLOW D. Adversarial perturbations of deep neural networks[J]. Perturbations, Optimization, and Statistics, 2017:311-342.
[25] KANNAN H, KURAKIN A, GOODFELLOW I. Adversarial logit pairing[EB/OL]. (2018-03-16) [2023-05-08]. https://browse.arxiv.org/pdf/1803.06373.
[26] XIE C, WU Y, MAATEN L V D, et al. Feature denoising for improving adversarial robustness[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE, 2019:501-509.
[27] MADRY A, MAKELOV A, SCHMIDT L, et al. Towards deep learning models resistant to adversarial attacks[C]//Proceedings of the International Conference on Learning Representations. Vancouver, Canada: ICLR, 2018:1-28.
[28] 钱亚冠, 张锡敏, 王滨, 等. 基于二阶对抗样本的对抗训练防御[J]. 电子与信息学报, 2021,43(11):3367-3373.
[29] LI J, GAO J, LI X X. Adversarial examples defense via combining data transformations and RBF layers[C]//Proceedings of the Pacific Rim International Conference on Artificial Intelligence. Hanoi, Vietnam: Springer, 2021:349-361.
[30] VIDNEROVÁ P, NERUDA R. Deep networks with RBF layers to prevent adversarial examples[C]//Proceedings of the International Conference on Artificial Intelligence and Soft Computing. Zakopane, Poland: Springer, 2018:257-266.
[31] ZADEH P H, HOSSEINI R, SRA S. Deep-RBF networks revisited: robust classification with rejection[EB/OL]//(2018-12-07)[2023-05-08]. https://browse.arxiv.org/pdf/1812.03190.
[32] YU Y, YU P, LI W. AuxBlocks: defense adversarial examples via auxiliary blocks[C]//Proceedings of the In-

ternational Joint Conference on Neural Networks. Budapest, Hungary: IEEE, 2019:1-8.
[33] VINCENT P, LAROCHELLE H, BENGIO Y, et al. Extracting and composing robust features with denoising autoencoders[C]//Proceedings of the International Conference on Machine learning. Helsinki, Finland: ACM, 2008:1096-1103.
[34] RONNEBERGER O, FISCHER P, BROX T. U-Net: convolutional networks for biomedical image segmentation [C]//Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention. Munich, Germany: Springer, 2015:234-241.
[35] 韩蒙, 俞伟平, 周依云, 等. 基于特征分布差异的对抗样本检测 [J]. 信息安全学报, 2023,8(3):1-11.
[36] LI Q, WANG H, LI B Y, et al. IIE-segNet: deep semantic segmentation network with enhanced boundary based on image information entropy[J]. IEEE Access, 2021, 9:40612-40622.
[37] BELLET A, HABRARD A, SEBBAN M. A survey on metric learning for feature vectors and structured data [EB/OL]//(2014-02-12)[2023-05-08]. https://arxiv.org/pdf/1306.6709.
[38] PANG T, YANG X, DONG Y, et al. Bag of tricks for adversarial training[EB/OL]//(2014-02-12)[2023-05-08]. https://browse.arxiv.org/pdf/1306.6709.
[39] LIU W, WEN Y, YU Z, et al. SphereFace: deep hypersphere embedding for face recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE, 2017:212-220.
[40] LOSHCHILOV I, HUTTER F. SGDR: stochastic gradient descent with warm restarts [EB/OL] // (2017-05-03) [2023-05-08]. https://browse.arxiv.org/pdf/1608.03983.
[41] TANAY T, GRIFFIN L D. A new angle on L2 regularization[EB/OL]//(2018-06-28)[2023-05-08]. https://browse.arxiv.org/pdf/1806.11186.
[42] LECUN Y, BOTTOU L, BENGIO Y, et al. Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE, 1998,86(11):2278-2324.
[43] KRIZHEVSKY A. Learning multiple layers of features from tiny images[D]. Toronto: University of Toronto, 2009.
[44] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[C]//Proceedings of the International Conference on Learning Representations. San Diego, USA: ICLR, 2015:1-14.
[45] CARLINI N, WAGNER D. Towards evaluating the robustness of neural networks [C] // Proceedings of the IEEE Symposium on Security and Privacy. San Jose, USA: IEEE, 2017:39-57.

# Immune defense against adversarial attacks via hourglass data-processing units and group RBF units

DING Weijie[*,**], ZHENG Wenhao[*,***], FANG Yi[*,***], WANG Qihui[****], LI Xiaoxin[*,***]
([*] College of Computer Science and Technology, Zhejiang University of Technology, Hangzhou 310023)
([**] Department of Computer and Information Security, Zhejiang Police College, Hangzhou 310053)
([***] Key Laboratory of Visual Media Intelligent Processing Technology of Zhejiang Province, Hangzhou 310023)
([****] Qianjiang College, Hangzhou Normal University, Hangzhou 311121)

**Abstract**

Deep neural network (DNN) is vulnerable to adversarial examples with imperceptible perturbation to clean images. To counter this issue, researchers proposed many powerful defensive methods, which can be categorized into external defense methods (EDMs) and immune defense methods (IDMs). EDMs try to purify the adversarial examples before they are fed into DNNs, while IDMs try to robustify the DNNs per se. This work focuses on IDMs. Most of the existing IDMs boost robustness mainly via using robust optimization strategies rather than building robust modules for DNNs. This work introduces two new robust units into DNNs: the hourglass data-processing units, based on feature squeezing and precision injection, for reducing adversarial perturbations, and the group RBF units for enhancing nonlinearity and handling intra-class variations. This work also uses label smoothing, annealing strategy and weight decay during optimization to further boost robustness. Extensive experiments on two public datasets, MNIST and CIFAR-10, and two popular DNNs, LeNet5 and VGG16, demonstrate that integrating the proposed robust units into DNNs could greatly improve their immune abilities against adversarial attacks while keeping their original recognition performance on clean samples.

**Key words**: immune defense, precision injection, group radial basis function (RBF), weight decay