计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.200800081

基于高斯增强和迭代攻击的对抗训练防御方法

王丹妮　陈　伟　羊　洋　宋　爽
电子科技大学信息与软件工程学院(软件工程)　成都 610054
(1248404073@qq.com)

摘　要　近年来,现有的深度学习网络模型已经能在各种分类任务中达到很高的准确率,但它们仍然极易受到对抗样本的攻击。目前,对抗训练是防御对抗样本攻击的最好方法之一。但已知的单步攻击对抗训练方法仅对单步攻击有着良好的防御效果,对迭代攻击的防御性能却很差,而迭代攻击对抗训练方法只提升了对迭代攻击的防御性能,对单步攻击的防御效果却不够理想。为了同时提高深度学习网络模型对单步攻击与迭代攻击的鲁棒性,文中提出了一种综合高斯增强和迭代攻击 ILLC(Iteration Least-Likely Class)的对抗训练防御方法 GILLC(Gaussian Iteration Least-Likely Class)。首先,在干净样本中添加了一个高斯扰动,用于提高深度学习网络模型的泛化能力;然后,使用 ILLC 产生的对抗样本进行对抗训练,近似解决对抗训练的内部最大化问题。文中以 CIFAR10 为数据集进行了白盒攻击实验,结果表明,通过与基线、单步攻击对抗训练和迭代攻击对抗训练的方法相比,GILLC 方法有效提高了深度学习网络模型对单步攻击和迭代攻击的鲁棒性,同时不会显著降低对干净样本的分类性能。

关键词:迭代攻击;高斯增强;单步攻击;对抗训练;对抗样本;深度学习
中图法分类号　TP391

Defense Method of Adversarial Training Based on Gaussian Enhancement and Iterative Attack

WANG Dan-ni,CHEN Wei,YANG Yang and SONG Shuang
School of Information and Software Engineering(Software Engineering),University of Electronic Science and Technology of China,Chengdu 610054,China

Abstract　In recent years,the existing deep learning network models have been able to achieve high accuracy in various classification tasks,but they are still extremely vulnerable to be attacked by adversarial samples. At present,adversarial training is one of the best methods to defend against adversarial sample attacks. However,the known single-step attack adversarial training methods only have a good defensive effect against single-step attacks,but have poor defense performance against iterative attacks. The iterative attack adversarial training methods only improve the defense performance against iterative attacks,but the defense effect of single-step attacks is not ideal. In order to improve the robustness of the deep learning network model against single-step attacks and iterative attacks at the same time,this paper proposes GILLC,an adversarial training defense method that combines Gaussian enhancement and ILLC iterative attacks. First,a Gaussian perturbation is added to the clean samples to improve the generalization ability of the deep learning network model. Then,the adversarial samples generated by ILLC are used for adversarial training,which approximately solves the internal maximization problem of adversarial training. In this paper,a white box attack experiment is conducted with CIFAR10 as the data set. The results show that the GILLC method effectively improves the robustness of the deep learning network model against single-step attacks and iterative attacks by comparing with the baseline,single-step attack adversarial training and iterative attack adversarial training methods,without significantly reducing the classification performance of the clean samples.

Keywords　Iterative attacks,Gaussian enhancement,Single-step attacks,Adversarial training,Adversarial samples,Deep learning

1　引言

深度学习网络在解决许多具有挑战性的人工智能问题方面取得了进展,如图像分类[1]、目标检测[2]和语义分割[3]等。然而,最近的研究[4]表明,深度学习网络模型容易受到对抗样本的影响,例如在图像识别中,攻击者可以向测试样本添加一个人类不可察觉的小扰动,从而使该样本被深度学习网络模型错误分类。其中,带有扰动的测试样本称为对抗样本[4],相反地,原始样本称为干净样本。产生对抗样本的攻击方法主要分成两类:单步攻击(只执行一步梯度计算)和迭代攻击(执行多步梯度计算)。在实际场景中,对抗样本给深度学习系统带来了很大的安全隐患,例如自动驾驶汽车的一个基本功能是自动识别停车标志和交通灯,假设一个攻击者制作了一个停车标志的对抗样本,如果自动驾驶汽车不会将其识别为停

基金项目:国家自然科学基金国际(地区)合作与交流项目(61520106007)
This work was supported by the Funds for International Cooperation and Exchange of the National Natural Science Foundation of China (61520106007).
通信作者:陈伟(chenwei@uestc.edu.cn)

车标志，就不会在停车标志处停车，且可能会与其他车辆相撞，造成严重的交通事故。因此，研究对抗样本的防御显得尤其重要。

目前，对抗样本攻击的防御技术可分成完全防御和仅检测[5]两类，完全防御方法旨在使对抗样本在目标网络上实现正确的分类，要求分类器以可接受的精度预测对抗样本的标签，而仅检测方法只能筛选出干净样本。但是，分类器的最终目标在于无论输入是否为对抗样本，都要获取该输入样本的真实分类标签。由此可见，仅检测防御方法在实际运用中具有一定的局限性。在完全防御方法中，对抗训练[6]是防御对抗样本的最佳方案之一[7]，它通过将对抗样本注入训练数据中来提高深度学习模型的鲁棒性。针对产生对抗样本的攻击方法的类型，本文将对抗训练分为单步攻击对抗训练与迭代攻击对抗训练。

Kurakin 等[8]提出了一种大规模机器学习对抗训练的方法，他们将单步攻击对抗样本注入训练集中，并由此进行训练。该方法将对抗训练扩展到大型模型和数据集的应用中，并对单步攻击生成的对抗样本具有很高的分类准确度，而对迭代攻击生成的对抗样本的分类准确率仅略高于正常训练模型。Madry 等[6]提出 PGD 对抗训练，将 PGD 迭代攻击对抗样本作为训练集进行训练，提高了模型对迭代攻击对抗样本的分类精度，但是对单步攻击对抗样本的分类精度明显不如 Kurakin 等的对抗训练方法。后来提出的对抗训练方法大多都是以 PGD 对抗训练为基础进行改进的，如 Song 等[9]进一步提高了模型对迭代攻击的鲁棒性，Shafahi 等[10]提升了模型对抗训练的速度，但是这些方法都忽略了模型对单步攻击的防御性能。

因此，本文提出了一种基于高斯增强和迭代攻击 ILLC 的对抗训练方法 GILLC，该方法改进了 Kurakin 等[8]的单步攻击 Step-LL 对抗训练方法，使用添加了高斯扰动的 ILLC，接着将产生的对抗样本作为训练集进行对抗训练，从而同时提升了模型对单步攻击和迭代攻击的鲁棒性。实验结果证明，在白盒攻击的条件下，与正常训练模型相比，GILLC 方法不会明显降低对干净样本的分类性能；与 Kurakin 等[8]的方法相比，GILLC 显著提升了对迭代攻击的鲁棒性；与 Madry 等[6]的方法相比，GILLC 有效提升了对单步攻击的鲁棒性。

2 相关工作

2.1 生成对抗样本的攻击方法

Szegedy 等[4]首先发现了对抗样本的存在，相关研究者由此提出了基于深度学习分类网络模型产生对抗样本的攻击方法，本文选取了其中的 7 种攻击算法进行介绍，并在后续的实验中用于对抗训练或评估。

(1)FGSM

快速符号梯度法(Fast Gradient Sign Method，FGSM)是 Goodfellow 等[11]提出的第一种单步攻击。其基本原理是最大化损失函数，使模型产生误分类，如式(1)所示：

$$x^{adv}=x+\varepsilon\cdot sign(\nabla_x L(h_\theta(x),y_{\text{true}})) \tag{1}$$

其中，x^{adv} 表示对抗样本；x 表示正常样本；ε 表示添加的扰动大小；$sign(\cdot)$表示符号函数；$L(\cdot)$表示损失函数；$h_\theta(\cdot)$表示参数为 θ 的训练模型；y_{true}表示干净样本对应的真实标签。

(2)Step-LL

Kurakin 等[12]提出了单步最不可能分类法(Single-Step Least-Likely Class Method，Step-LL)。Step-LL 以最不可能的分类作为目标标签，用于代替真实标签来最小化损失函数，使模型朝着最不可能的目标类进行分类，如式(2)所示：

$$x^{adv}=x-\varepsilon\cdot sign(\nabla_x L(h_\theta(x),y_{\text{target}})) \tag{2}$$

其中，y_{target}表示模型对干净样本的最不可能分类标签，即 $\arg\min\{h_\theta(x)\}$。

(3)ILLC

同时，Kurakin 等[12]提出了一种 Step-LL 的迭代增强版本，即迭代最不可能分类法(Iteration Least-Likely Class Method，ILLC)，如式(3)所示：

$$x_0^{adv}=x \tag{3}$$

$$x_{N+1}^{adv}=Clip_{x,\varepsilon}\{x_N^{adv}-\alpha\cdot sign(\nabla_{x_N^{adv}}L(h_\theta(x_N^{adv}),y_{\text{target}}))\}$$

其中，α 表示每一次迭代添加的扰动步长；$Clip_{x,\varepsilon}$表示将对抗样本裁剪到$[x-\varepsilon,x+\varepsilon]$的范围内。

(4)BIM

基本迭代法[12](Basic Iterativa Method，BIM)是 FGSM 的迭代增强版本，如式(4)所示：

$$x_0^{adv}=x \tag{4}$$

$$x_{N+1}^{adv}=Clip_{x,\varepsilon}\{x_N^{adv}+\alpha\cdot sign(\nabla_{x_N^{adv}}L(h_\theta(x_N^{adv}),y_{\text{true}}))\}$$

(5)PGD

Madry 等[6]在迭代扰动之前添加了一个随机扰动，并据此提出了 BIM 的增强版本，即投影梯度下降法(Projected Gradient Decent，PGD)，具体计算式如式(5)所示：

$$x_0^{adv}=x+U(-\varepsilon,\varepsilon) \tag{5}$$

$$x_{N+1}^{adv}=Clip_{x,\varepsilon}\{x_N^{adv}+\alpha\cdot sign(\nabla_{x_N^{adv}}L(h_\theta(x_N^{adv}),y_{true}))\}$$

其中，$U(-\varepsilon,\varepsilon)$表示$[-\varepsilon,\varepsilon]$的均匀分布。

(6)DeepFool

Moosavi-Dezfooli 等[13]提出了一种迭代攻击 DeepFool。DeepFool 的关键思想是迭代地向干净样本中添加扰动，直到使分类器产生误分类或达到最大迭代次数，并找到将对抗样本移动到线性分类边界所需的最小扰动。

(7)CW

Carlini 等[14]提出的 CW L_2 攻击是一种强大而复杂的攻击方法，该攻击方法通过使用辅助变量 w 直接产生对抗样本 x^{adv}，如式(6)所示：

$$x^{adv}=0.5\cdot(\tanh(w)+1) \tag{6}$$

优化辅助变量的目标函数如式(7)、式(8)所示：

$$F(x^{adv})=\max(\max\{Z_i(x^{adv}):i\neq y_{\text{target}}\}-Z_{y_{\text{target}}}(x^{adv}),-k) \tag{7}$$

$$\min_w \|x^{adv}-x\|_2+c\cdot F(x^{adv}) \tag{8}$$

其中，$Z(x)$是 softmax 函数前一层的输出；常数 k 用于控制对抗样本类和干净样本类之间的置信区间；超参数 c 用于平衡两个目标函数。

2.2 对抗样本的防御方法

与生成对抗样本的攻击相反，相关学者在研究对抗样本的防御方面也取得了一定进展。现有的对抗样本防御方法可分为以下两类。

(1)仅检测

仅检测是一种能检测出对抗样本并拒绝对该样本分类的防御方法。MagNet[15]是 2017 年提出的一种仅检测防御方法。根据深度学习的流行假设(Manifold Hypothesis)，干净样本位于分类任务的 Manifold 空间内，而对抗样本则远离或接近分类任务的 Manifold 边界，Meng 和 Chen 由此设计了 MagNet 防御系统。该系统包括一个或多个检测器(Detec-

tor)和一个恢复器(Reformer)，Detector 和 Reformer 都通过自编码器[16](Autoencoder)实现。Detector 判断输入样本是否远离 Manifold 边界，若远离，则直接拒绝对该样本分类；若接近 Manifold 边界，则利用 Reformer 对该样本进行恢复，再输入到分类器进行分类。这有利于提高神经网络的分类准确度，但与所有的仅检测方法类似，MagNet 依赖于手动标记检测到的对抗样本，失去了自动决策的优势。

(2)完全防御

完全防御是一种不仅能将干净样本正确分类，还能使一部分对抗样本也被正确分类的防御方法。2016 年，Papernot 等[17]提出了一种基于防御蒸馏的深度学习网络模型训练方法。他们发现，使用防御蒸馏训练的网络模型可以有效防御对抗样本。然而，Carlini 等[14]的实验证明，CW 方法仍然可以对防御蒸馏训练的网络模型达到 100% 的攻击成功率。2017 年，Xu 等[18-19]提出的特征压缩是目前对抗攻击的最有效防御手段，但会大大降低模型对干净样本的分类准确率。在每个批次的训练集中，Kurakin 等[8]采用干净样本和相应对抗样本的混合来替换原始干净样本进行训练，以提高模型的鲁棒性，但该方法对迭代攻击对抗样本的分类准确率很低。Madry 等[6]于 2018 年提出的 PGD 对抗训练，仅利用 PGD 迭代攻击产生的对抗样本进行训练，该方法大大提升了模型对迭代攻击的防御效果，对单步攻击的防御效果却不如 Kurakin 等的方法。2019 年，Shafahi 等[10]采用对训练批次进行回放的方式减少训练次数，加快了 PGD 对抗训练的速度。2020 年，Wong 等[20]将 PGD 对抗训练中的迭代攻击用单步攻击代替，进一步加速了对抗训练的过程。Song 等[9]在 PGD 对抗训练之前随机打乱和重组 PGD 对抗样本的局部特征，提高了模型对迭代攻击的鲁棒性。Xiao 等[21]将激活函数改为非连续性函数 k-WTA，k-WTA 保留 k 个最大值或比例为 k 的最大值，有效地防御了基于梯度的攻击。

尽管现有的防御方法只能在有限的环境和一定的程度上提高模型的鲁棒性，但是 Li 等[7]的研究结果表明，对抗训练依然是最好的防御方法之一。

3 基于高斯增强和迭代攻击的对抗训练防御方法

3.1 添加高斯扰动的对抗训练方法

高斯增强是一种数据增强方法，通过向训练集中添加高斯扰动来实现。高斯增强通常用于减轻深度学习网络中高频特征(经常出现且没有意义的特征)对模型训练的影响，从而避免发生过拟合(Overfitting)现象，虽然这种方式也会对低频特征(重点关注的特征)产生影响，但是深度学习网络能够通过学习来忽略这些影响。高斯扰动能在所有的频率上都产生数据点，通过添加适量的高斯扰动能够有效提高深度学习网络的学习能力。

数据增强防御(如对抗训练)的依据是，限制模型对干净样本做出相同的预测，并对其添加扰动的版本进行预测，从而提高模型的泛化能力。一般的对抗训练方法只能让模型在几个特定的方向上得到加强(通常每个输入样本让模型在一个方向上得到加强)，但它在其他方向上则容易遭受攻击。利用添加高斯扰动的样本来增强训练集，一方面，可以加强模型的多个方向，另一方面，可以平滑模型的置信度[22]。虽然前者的特性可以通过任何类型的扰动(如均匀扰动)来实现，但后者的独特之处在于采用高斯分布的扰动，能够鼓励模型逐渐降低对输入样本分类的置信度。

由于高斯增强具有提高模型学习与泛化的能力，本文预设在对抗训练产生对抗样本之前先对干净样本添加一个高斯扰动会从整体上提高模型的鲁棒性，故在 Kurakin 等[8]的对抗训练中加入了一个高斯扰动进行改进，并将其称为 GLLC 对抗训练，如算法 1 所示。算法 1 中，$N(\mu,\sigma^2)$ 表示均值为 μ、方差为 σ^2 的高斯分布。

算法 1 GLLC 对抗训练

输入：对抗训练次数 T；训练集大小 M；训练集 (x_i, y_i^{true})；目标标签 y_i^{target}，$i\in\{1,2,\cdots,M\}$；扰动 ε；迭代扰动步长 α；均值 μ；方差 σ^2

输出：模型参数 θ

1. for t←1 to T do
2. for i←1 to M do
3. / * 为干净样本添加高斯扰动并将其标准化 * /
4. $x_i^{adv}=x_i+\max(\min(N(\mu,\sigma^2)/255,\varepsilon),-\varepsilon)$
5. / * 产生 LLC 对抗样本 * /
6. $x_i^{adv}=x_i^{adv}-\alpha\cdot\mathrm{sign}(\nabla_{x_i^{adv}}L(h_\theta(x_i^{adv}),y_i^{target}))$
7. $x_i^{adv}=\max(\min(x_i^{adv},x_i+\varepsilon),x_i-\varepsilon)$
8. / * 梯度下降更新模型参数 θ * /
9. $\theta=\theta-\nabla_\theta L(h_\theta(x_i^{adv}),y_i^{true})$
10. end for
11. end for

3.2 基于迭代攻击的对抗训练方法

对抗训练是一种网络学习防御对抗样本攻击的方法，使网络具有较高的鲁棒性。对抗训练的基本思想是解决最小最大优化问题，如式(9)所示：

$$\min_{\theta}\max_{x^{adv},\|x^{adv}-x\|_\infty\leqslant\varepsilon}L(h_\theta(x^{adv}),y_{\text{true}}) \tag{9}$$

其中，x^{adv} 表示对抗样本；x 表示正常样本；ε 表示添加的扰动大小；$L(\cdot)$ 表示损失函数；$h_\theta(\cdot)$ 表示参数为 θ 的训练模型；y_{true} 表示干净样本对应的真实标签。对抗训练的过程是利用在干净样本中添加一定的对抗扰动 ε 来逼近损失函数的内部最大值，然后对模型参数 θ 进行梯度下降。

对抗训练的最早版本之一是使用单步攻击来近似式(9)中的内部最大化，如 Kurakin 等[8]的对抗训练，利用 Step-LL 单步攻击产生的对抗样本来近似内部最大化。随着迭代攻击的产生，可以推测内部最大化的一个更好的近似是采取多个更小的 Step-LL 步骤来代替单步攻击。因此，本文提出了一种通过 Step-LL 攻击的迭代版本来近似内部最大化的方法，即 ILLC 对抗训练。算法 2 描述了 ILLC 对抗训练。

算法 2 ILLC 对抗训练

输入：对抗训练次数 T；迭代次数 I；训练集大小 M；训练集 (x_i, y_i^{true})；目标标签 y_i^{target}，$i\in\{1,2,\cdots,M\}$；扰动 ε；迭代扰动步长 α

输出：模型参数 θ

1. for t←1 to T do
2. for i←1 to M do
3. $x_i^{adv}=x_i$
4. for j←1 to I do
5. / * 产生 ILLC 对抗样本 * /
6. $x_i^{adv}=x_i^{adv}-\alpha\cdot\mathrm{sign}(\nabla_{x_i^{adv}}L(h_\theta(x_i^{adv}),y_i^{target}))$
7. $x_i^{adv}=\max(\min(x_i^{adv},x_i+\varepsilon),x_i-\varepsilon)$
8. end for
9. / * 梯度下降更新模型参数 θ * /
10. $\theta=\theta-\nabla_\theta L(h_\theta(x_i^{adv}),y_i^{true})$
11. end for
12. end for

3.3 GILLC 对抗训练方法

为了兼备对单步攻击的良好防御效果[8]与迭代攻击的防御性能[6]，本文提出了 GILLC 对抗训练方法，采用 ILLC 迭代攻击产生的对抗样本用于对抗训练，并在 ILLC 迭代攻击之前添加高斯扰动来增强训练模型的泛化能力，以进一步提高深度学习网络模型对单步攻击和迭代攻击的鲁棒性，如算法 3 所示。其中，$N(\mu,\sigma^2)$表示均值为 μ、方差为 σ^2 的高斯分布。GILLC 是一种全新的对抗训练方法，首次使用高斯扰动与 ILLC 迭代攻击结合产生对抗样本进行训练，为对抗训练方法提供了一种新的思路，本文将在后面的实验中证明 GILLC 方法的有效性。

算法 3 GILLC 对抗训练

输入：对抗训练次数 T；迭代次数 I；训练集大小 M；训练集(x_i, y_i^{true})，目标标签 y_i^{target}，$i\in\{1,2,\cdots,M\}$；扰动 ε；迭代扰动步长 α；均值 μ；方差 σ^2

输出：模型参数 θ

1. for t←1 to T do
2. for i←1 to M do
3. /* 为干净样本添加高斯扰动并将其标准化 */
4. $x_i^{adv}=x_i+\max(\min(N(\mu,\sigma^2)/255,\varepsilon),-\varepsilon)$
5. for j←1 to I do
6. /* 产生 ILLC 对抗样本 */
7. $x_i^{adv}=x_i^{adv}-\alpha\cdot\mathrm{sign}(\nabla_{x_i^{adv}}L(h_\theta(x_i^{adv}),y_i^{target}))$
8. $x_i^{adv}=\max(\min(x_i^{adv},x_i+\varepsilon),x_i-\varepsilon)$
9. end for
10. /* 梯度下降更新模型参数 θ */
11. $\theta=\theta-\nabla_\theta L(h_\theta(x_i^{adv}),y_i^{true})$
12. end for
13. end for

4 实验结果与分析

4.1 实验设置

实验使用的数据集为 CIFAR10 基准数据集，有 5 万张训练图像和 1 万张测试图像以及 10 个类别。基于 CIFAR10 数据集，将 GILLC 对抗训练模型与 Kurakin 等[8]的对抗训练模型、Madry 等[6]的对抗训练模型和正常训练模型（仅使用干净样本进行训练）进行对比实验，分别将这 4 种训练模型记作 GILLC，NAT，PAT 和 Natural。

攻击设置如下：本文只讨论白盒攻击（攻击者知道模型的所有参数，直接攻击模型产生对抗样本），使用 1 种经典的单步攻击和 4 种先进的迭代攻击来生成对抗样本，单步攻击为 FGSM 攻击，迭代攻击包括 BIM 攻击、PGD 攻击、DeepFool 攻击和 CW 攻击。首先，从 CIFAR10 测试集中随机选取了 1000 个样本，再根据以上攻击生成不同的对抗样本。对于 FGSM 攻击，扰动大小 ε 为 2/255，4/255 和 6/255。对于 BIM 和 PGD 攻击，迭代次数为 7，扰动大小 ε 为 1/255，2/255，…，8/255，对应的迭代扰动步长为 $\alpha=\varepsilon/4$。对于 DeepFool 攻击，最大迭代次数为 20。对于 CW 攻击，二分搜索步骤为 5，最大迭代次数为 20，学习率为 0.01，初始常数为 0.01，置信度为 10。

训练设置如下：所有模型均在 ResNet18 网络上采用 Adam 优化器进行训练，训练集分组大小为 32，训练次数 epochs 为 120，第 1epoch 至 80epoch 的学习率为 0.001，第 80epoch 至 120epoch 的学习率为 0.0001，最后选择对干净测试集分类准确率最高的训练模型进行实验。对于 GILLC 对抗训练方法，迭代次数为 7，扰动大小 ε 为 8/255，扰动步长 α 为 2/255，均值 μ 为 0，方差 σ^2 为 1。对于 NAT 和 PAT 对抗训练方法，遵循文献[6]和文献[8]中的相同参数设置。

评估标准如下：所有的实验均采用模型对干净样本或对抗样本的分类准确率进行评估，分类准确率越高，防御效果就越好，模型的鲁棒性也就越强。

4.2 干净样本的分类结果

表 1 列出了正常训练模型 Natural、对抗训练模型 NAT、PAT 和 GILLC 的分类准确率。与 Natural 相比，NAT 和 GILLC 对干净样本的分类准确率都略有下降，但其幅度在可接受的范围内，说明了这两者都不会显著降低模型的性能。PAT 对干净样本的分类准确率最低，说明了 GILLC 优于 PAT。此外，由于 NAT 在对抗样本中混合加入了干净样本进行训练，GILLC 对干净样本的分类准确率稍低于 NAT。

表 1 干净样本的分类准确率

Table 1 Classification accuracy of clean samples

Training	Accuracy/%
Natural	93.38
NAT	91.43
PAT	84.98
GILLC	87.56

4.3 白盒攻击下的实验结果

表 2 列出了 3 种不同训练模型在 FGSM 白盒攻击下的准确率。扰动 ε 的值越大，攻击强度就越大。与 Natural 相比，NAT，PAT 和 GILLC 大幅度提升了对 FGSM 单步攻击的防御效果。当 ε 较小时，GILLC 的防御效果优于 NAT 和 PAT。随着 ε 的增加，GILLC 对 FGSM 对抗样本的分类准确率略低于 NAT，但依然保持着较高的分类精度，如攻击强度 ε 高达 6 时，GILLC 下降的准确率仅为 3.7%，说明了 GILLC 基本保持了与 NAT 相同的单步攻击防御效果；同时，PAT 对 FGSM 对抗样本的分类准确率急剧下降。从总体上来说，这反映了 GILLC 有效提高了模型对单步攻击的防御效果。

表 2 FGSM 白盒攻击下的准确率

Table 2 Classification accuracy under FGSM white box attack

（单位：%）

Training	Accuracy (ε=2/255)	Accuracy (ε=4/255)	Accuracy (ε=6/255)
Natural	34.9	20.8	17.6
NAT	76.4	73.5	71.2
PAT	77.0	68.1	57.6
GILLC	77.9	72.9	67.5

不同训练模型在 BIM 白盒攻击下的对抗样本分类准确率如图 1 所示。图 1 中的横轴为不同扰动大小 ε 的 BIM 迭代攻击，将扰动大小 1/255，2/255，…，8/255 用 1，2，…，8 表示。图 1 中的纵轴为模型对 BIM 对抗样本的分类准确率。其中，Natural 未采取任何防御措施，因此将其作为对照基线。从图中可以看出，NAT，PAT 和 GILLC 都提高了模型对 BIM 对抗样本的分类准确率，但是 GILLC 和 PAT 的防御效果明显优于 NAT，且 GILLC 可以达到与 PAT 几乎同样好的防御效果，说明与 NAT 相比，GILLC 有效提升了模型对迭代攻击的防御性能。另外，随着攻击强度 ε 的增加，NAT 对 BIM 对抗样本的分类准确率急剧下降，而 GILLC 曲线则比较平缓，如 ε 为 4 时，NAT 对 BIM 对抗样本的分类准确率已经降低到 17%，而 GILLC 的分类准确率却仍然高达 64.6%。同时，本文还发现攻击强度 ε 越大，越能凸显 GILLC 的防御效果优于

NAT，即使是在 ε 为 8 的攻击强度下，相比 NAT 对 BIM 对抗样本 1.8%的分类准确率，GILLC 也能达到 40.1%。

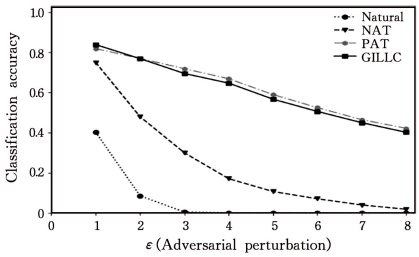


图 1 BIM 白盒攻击下的分类准确率

Fig. 1 Classification accuracy under BIM white box attack

图 2 给出了不同训练模型在 PGD 白盒攻击下的分类准确率。图 2 中的横轴表示从 1/255 到 8/255 的 8 种不同 ε 强度的 PGD 迭代攻击，纵轴表示模型对 PGD 对抗样本的分类准确率。由图 2 可知，该实验结果与 BIM 白盒攻击的实验结果相似，NAT，PAT 和 GILLC 都使模型对 PGD 对抗样本的分类准确率得到了提高，且 GILLC 与 PAT 的防御效果相当。虽然 NAT 和 GILLC 都增强了模型的鲁棒性，但是 GILLC 对不同 ε 强度的 PGD 对抗样本的分类精度都远高于 NAT，尤其是当 ε 为 4 时，GILLC 的分类准确率为 65.9%，NAT 的分类准确率为 17.6%，GILLC 使分类准确率得到了 48.3%的提升，同样证明了 GILLC 对迭代攻击的防御效果优于 NAT。

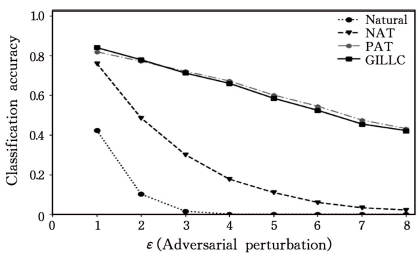


图 2 PGD 白盒攻击下的分类准确率

Fig. 2 Classification accuracy under PGD white box attack

表 3 列出了不同训练模型在 DeepFool 白盒攻击下的准确率。从表中可以看出，GILLC 对 DeepFool 对抗样本的分类精度最高，比 PAT 高了约 5%，而 NAT 基本没有起到防御效果，对 DeepFool 对抗样本的分类准确率在 3 种防御方法中最低，说明了 GILLC 的防御效果最好。

不同训练模型在 CW 白盒攻击下的分类准确率如表 4 所列。由于 CW 是一种比较复杂而强大的迭代攻击方法，产生 CW 对抗样本需要耗费较长的时间，因此本文仅产生置信度为 10 的 CW 对抗样本用于实验。从表中可以看出，NAT 并没有显著提升对 CW 对抗样本的分类准确率，而 PAT 的防御效果比 GILLC 稍差。相反地，GILLC 使分类准确率从正常训练模型 Natural 的 0.6%提升到了 52.2%，证明了即使在强大的迭代攻击下，GILLC 依然具有较高的分类准确率和较好的防御性能。

表 3 DeepFool 白盒攻击下的分类准确率

Table 3 Classification accuracy under DeepFool white box attack

Training	Accuracy/%
Natural	1.2
NAT	7.9
PAT	42.6
GILLC	46.5

表 4 CW 白盒攻击下的分类准确率

Table 4 Classification accuracy under CW white box attack

Training	Accuracy/%
Natural	0.6
NAT	5.9
PAT	51.9
GILLC	52.2

结束语 本文提出了一种防御对抗样本的对抗训练方法 GILLC，它在添加高斯扰动的基础上，利用迭代攻击产生对抗样本来替换干净训练样本，并进行训练。本文采用了 FGSM，BIM，PGD，DeepFool 和 CW 5 类白盒攻击方法，基于 CIFAR10 数据集，与基线 Natural、对抗训练模型 NAT 和 PAT 进行了对比实验，证实了 GILLC 方法在保持对干净样本良好分类效果的同时，有效提高了模型对单步攻击和迭代攻击的防御性能，从整体上提高了深度学习网络模型的鲁棒性。本文方法仍存在一定的局限性，仅从高斯增强与迭代攻击方面去逼近最大化损失函数的对抗样本，并不是最小最大优化问题的最优解，该方法对迭代攻击的防御效果不如对单步攻击的防御效果，对迭代攻击的防御性能还有待提高，未来的研究将致力于从更全面的角度去寻找对抗训练中内部最大化问题的最优解决方案，并且使用更多的攻击方法去衡量其防御性能。

参 考 文 献

[1] HE K，ZHANG X，REN S，et al. Deep Residual Learning for Image Recognition[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016：770-778.

[2] ZHANG Z，QIAO S，XIE C，et al. Single-shot Object Detection with Enriched Semantics[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018：5813-5821.

[3] CHEN L，PAPANDREOU G，KOKKINOS I，et al. DeepLab：Semantic Image Segmentation with Deep Convolutional Nets，Atrous Convolution，and Fully Connected CRFs[J]. IEEE Annals of the History of Computing，2018(4)：834-848.

[4] SZEGEDY C，ZAREMBA W，SUTSKEVER I，et al. Intriguing Properties of Neural Networks[C]// International Conference on Learning Representations. 2014.

[5] AKHTAR N，MIAN A. Threat of Adversarial Attacks on Deep Learning in Computer Vision：A Survey[J]. IEEE Access，2018：14410-14430.

[6] MADRY A，MAKELOV A，SCHMIDT L，et al. Towards Deep Learning Models Resistant to Adversarial Attacks[C]// International Conference on Learning Representations. 2018.

[7] LI Y，LI L，WANG L，et al. NATTACK：Learning the Distributions of Adversarial Examples for an Improved Black-Box Attack on Deep Neural Networks[C]// International Conference on Machine Learning. 2019：3866-3876.

[8] KURAKIN A，GOODFELLOW I，BENGIO S，et al. Adversarial Machine Learning at Scale[C]// International Conference on Learning Representations. 2017.

[9] SONG C，HE K，LIN J，et al. Robust Local Features for Improving the Generalization of Adversarial Training[C]// International Conference on Learning Representations. 2020.

[10] SHAFAHI A，NAJIBI M，GHIASI M A，et al. Adversarial training for free[C]// Neural Information Processing Systems. 2019：3358-3369.

（下转第 537 页）

参 考 文 献

[1] GUO B,SUN Z T,WANG Y,et al. Resisting power analysis attacks algorithm of scalar multiplication based on factorial expansions form [J]. Bulletin of Science and Technology, 2016, 32(6):149-153.

[2] WU K K,LI H Y,YAN L J. Homogeneous mapping model of ECC for preventing differential power analysis[J]. Computer Engineering,2017,43(10):115-119.

[3] LIANG F,SHEN J N. Resisting power analysis attacks scheme for ellipse curve cryptography based on odd-only Comb method [J]. Computer Applications and Software, 2016, 33(3): 288-290.

[4] MA B,BAO S G,DAI X Y. Efficiency improvement of ECC resisting power attack scheme in smart card[J]. Computer Engineering,2010,36(16):113-115.

[5] WANG Z Y,ZHAO J G. Resisting power analysis attack scheme based on signed double-based number system[J]. Journal of Computer Applications,2011,31(11):2973-2974.

[6] YANG B. Secure and efficient scalar multiplication algorithm with power analysis attack resistance[J]. Control Engineering of China,2017,24(12):2462-2465.

[7] LI Y,WANG J L,ZENG X W,et al. A segmented Montgomery scalar multiplication algorithm with resistance to simple power analysis SPA attacks[J]. Computer Engineering and Science, 2017,30(1):92-101.

[8] SHI L,XU M. DWNAF:a dynamic window NAF scalar multiplication with threshold[J]. Computer Science, 2017, 44(10): 159-164.

[9] PHILLIPS B J,BURGESS N. Implementing 1024-bits RSA exponentiation on a 32-bits processor core[C]// Proceeding of the Application Specific-Systems,Architecture and Processor. 2000: 127-137.

[10] WEI G H,WANG Y,ZHANG H G. ECC point multiplication lightweight improvement for RFID applications over GF(2^m) [J]. Computer Engineering and Science,2017,39(1):81-85.

[11] WANG Y X,ZHANG C R,ZHANG B H,et al. Efficient scalar multiplication of ECC based on composite operations over prime fields[J]. Application Research of Computers, 2013, 30(11): 3365-3387.

[12] LIU G Z,QI H X. Efficient NAF scalar multiplication algorithm with low storage [J]. Science Technology and Engineering, 2013,13(19):5683-5686.

[13] BARUA R,PANDEY S K,PANKAJ R. Efficient window-based scalar multiplication on elliptic curves using double-base number system [J]. Lecture Notes in Computer Science,2007,4859(12): 351-360.

GONG Jian-feng, postgraduate, lecturer. His main research interests include computer network technology and information security.

(上接第 513 页)

[11] GOODFELLOW I,SHLENS J,SZEGEDY C,et al. Explaining and Harnessing Adversarial Examples[C]// International Conference on Learning Representations. 2015.

[12] KURAKIN A,GOODFELLOW I,BENGIO S,et al. Adversarial examples in the physical world[C]// International Conference on Learning Representations. 2017.

[13] MOOSAVI-DEZFOOLI S M,FAWZI A,FROSSARD P. Deepfool:A Simple and Accurate Method to Fool Deep Neural Networks[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016:2574-2582.

[14] CARLINI N,WAGNER D. Towards Evaluating the Robustness of Neural Networks[C]// IEEE Symposium on Security and Privacy. 2017:39-57.

[15] MENG D,CHEN H. MagNet:A Two-Pronged Defense against Adversarial Examples[C]// Computer and Communications Security. 2017:135-147.

[16] GU S,RIGAZIO L. Towards Deep Neural Network Architectures Robust to Adversarial Examples[J]. arXiv: Learning, 2014.

[17] PAPERNOT N,MCDANIEL P,WU X,et al. Distillation as a Defense to Adversarial Perturbations Against Deep Neural Networks[C]// IEEE Symposium on Security and Privacy. 2016: 582-597.

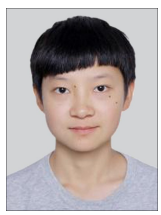
[18] XU W,EVANS D,QI Y,et al. Feature Squeezing Mitigates and Detects Carlini/Wagner Adversarial Examples[J]. arXiv:Cryptography and Security,2017.

[19] XU W,EVANS D,QI Y,et al. Feature Squeezing:Detecting Adversarial Examples in Deep Neural Networks[C]// Network and Distributed System Security Symposium. 2018.

[20] WONG E,RICE L,KOLTER J Z,et al. Fast is Better than Free:Revisiting Adversarial Training[J]. arXiv preprint arXiv: 2001.03994,2020.

[21] XIAO C,ZHONG P,ZHENG C,et al. Enhancing Adversarial Defense by k-Winners-Take-All[J]. arXiv preprint arXiv:1905.10510,2019.

[22] ZANTEDESCHI V,NICOLAE M,RAWAT A,et al. Efficient Defenses Against Adversarial Attacks[J]. arXiv: Learning, 2017.



WANG Dan-ni, born in 1995, postgraduate. Her main research interest includes information security of artificial intelligence.

CHEN Wei, born in 1978, Ph.D, associate professor. His main research interest includes network security.