基于二阶对抗样本的对抗训练防御

钱亚冠[①] 张锡敏[①] 王 滨*[②] 顾钊铨[③] 李 蔚[①] 云本胜[①]

[①](浙江科技学院理学院/大数据学院 杭州 310023)

[②](杭州海康威视网络与信息安全实验室 杭州 310052)

[③](广州大学网络空间先进技术研究院 广州 510006)

摘 要：深度神经网络(DNN)应用于图像识别具有很高的准确率，但容易遭到对抗样本的攻击。对抗训练是目前抵御对抗样本攻击的有效方法之一。生成更强大的对抗样本可以更好地解决对抗训练的内部最大化问题，是提高对抗训练有效性的关键。该文针对内部最大化问题，提出一种基于2阶对抗样本的对抗训练，在输入邻域内进行2次多项式逼近，生成更强的对抗样本，从理论上分析了2阶对抗样本的强度优于1阶对抗样本。在MNIST和CIFAR10数据集上的实验表明，2阶对抗样本具有更高的攻击成功率和隐蔽性。与PGD对抗训练相比，2阶对抗训练防御对当前典型的对抗样本均具有鲁棒性。

关键词：对抗样本；对抗训练；2阶泰勒展开

中图分类号：TN915.08; TP309.2 **文献标识码**：A **文章编号**：1009-5896(2021)11-3367-07

DOI: 10.11999/JEIT200723

Adversarial Training Defense Based on Second-order Adversarial Examples

QIAN Yaguan[①] ZHANG Ximin[①] WANG Bin[②] GU Zhaoquan[③]
LI Wei[①] YUN Bensheng[①]

[①](*School of Science/School of Big-data Science, Zhejiang University of Science and Technology, Hangzhou* 310023, *China*)

[②](*Network and Information Security Laboratory of Hangzhou Hikvision Digital Technology Co., Ltd. Hangzhou* 310052, *China*)

[③](*Cyberspace Institute of Advanced Technology* (*CIAT*), *Guangzhou University, Guangzhou* 510006, *China*)

Abstract: Although Deep Neural Networks (DNN) achieves high accuracy in image recognition, it is significantly vulnerable to adversarial examples. Adversarial training is one of the effective methods to resist adversarial examples empirically. Generating more powerful adversarial examples can solve the inner maximization problem of adversarial training better, which is the key to improve the effectiveness of adversarial training. In this paper, to solve the inner maximization problem, an adversarial training based on second-order adversarial examples is proposed to generate more powerful adversarial examples through quadratic polynomial approximation in a tiny input neighborhood. Through theoretical analysis, second-order adversarial examples are shown to outperform first-order adversarial examples. Experiments on MNIST and CIFAR10 data sets show that second-order adversarial examples have high attack success rate and high concealment. Compared with PGD adversarial training, adversarial training based on second-order adversarial examples is robust to all the existing typical attacks.

Key words: Adversarial examples; Adversarial training; The second-order Taylor expansion

收稿日期：2020-08-06；改回日期：2021-08-20；网络出版：2021-09-16

*通信作者： 王滨 32874546@qq.com

基金项目：国家重点研发计划项目(2018YFB2100400)，国家自然科学基金(61902082)

Foundation Items: The National Research and Development Program of China (2018YFB2100400), The National Natural Science Foundation of China (61902082)

1 引言

深度神经网络(DNN)在生物信息学[1,2]、语音识别[3,4]和计算机视觉[5,6]等领域获得成功应用的同时，研究者们发现DNN容易受到对抗样本的攻击[7]，即在自然图像中添加微小的扰动，可以欺骗DNN做出错误预测。由于对抗样本具有较好的隐蔽性，不易被人眼发现，给安全敏感的应用带来很大的破坏性。例如，在自动驾驶领域，研究者们通过在道路交通标志图片上添加微小扰动得到对抗样本，导致采用DNN进行道路交通标志识别的自动驾驶汽车做出错误判断，引起交通事故的发生[8]。自动驾驶系统可能会遇到的道路交通标志图片及其对应的对抗样本，对于人眼来说，两张图片是相同的，同为注意危险标志。而自动驾驶系统中的DNN则把对抗样本判断为让行标志。这意味着难以察觉的扰动有可能使一辆毫无故障的自动驾驶汽车做出危险的行为。因此，对于对抗样本的防御研究具有现实意义。

自Szegedy等人[7]发现DNN中存在对抗样本以来，研究者们提出了一系列对抗样本的生成与防御方法。生成对抗样本的过程通常被建模为一个有约束优化的问题，其目标是在约束条件下最大化损失函数。现有的典型对抗样本包括C&W[9], Deepfool[10], FGSM[11], PGD[12], M-DI^2-FGSM[13]等。同时，研究者们提出了多种防御对抗样本的方法，如防御蒸馏[14]、对抗训练[15]、强化网络[16]及对抗样本检测[17]等。

在大部分防御方法被文献[18]证实防御效果有限的情况下，对抗训练是少数被经验证明为目前最为有效的防御方法。对抗训练最早由Szegedy等人[7]提出，通过将对抗样本注入训练过程，以增强DNN的鲁棒性。随着研究的深入，Madry等人[13]将对抗训练形式化为由内部最大化问题和外部最小化问题组成的鞍点问题，即存在对抗样本最大化损失函数的情况下，优化模型参数实现损失函数最小化。按照Madry等人的鞍点理论，解决内部最大化问题需要更强的对抗样本，他们提出了基于PGD(1阶梯度投影)的对抗训练方法，实验证明能够防御大部分1阶梯度攻击。但是1阶梯度对于DNN的逼近能力有限，无法进一步找到更强大的对抗样本，因而也无法训练出更鲁棒的DNN。基于这个思路，本文提出于基于2阶梯度的对抗样本生成方法。与以往线性逼近方法不同，在输入样本的微小邻域内，对DNN损失函数进行2阶多项式逼近。本文提出的方法优点是，利用Hesse矩阵可提取到损失函数在输入邻域内的更多信息，从而更好地解决内部最大化问题。

本文分别从理论和实验角度证明了2阶对抗样本强于PGD对抗样本。本文提出将对抗样本的扰动下界，即攻击成功所需的最少扰动，用于衡量不同对抗样本的强度。计算结果显示，2阶对抗样本的扰动下界低于PGD，即2阶对抗样本攻击成功所需的最少扰动少于PGD，这意味着2阶对抗样本强于PGD。在MNIST和CIFAR10上的实验结果验证了本文的理论分析：(1) 相较于包括PGD在内的现有典型对抗样本，2阶对抗样本能够在添加更少扰动同时，达到更高的攻击成功率；(2) 基于2阶对抗样本的对抗训练能够防御现有的典型1阶对抗攻击。

2 预备知识

2.1 深度神经网络

DNN一般可以表示为映射函数$F:\mathbb{X}\mapsto\mathbb{Y}$，$\boldsymbol{X}\in\mathbb{X}$是$d$维输入变量，$\boldsymbol{Y}\in\mathbb{Y}$是一个$m$维概率向量，分别表示$m$个类的置信度。一个$N$层DNN接收一个输入$\boldsymbol{X}$后产生相应的输出，即

$$F(\boldsymbol{X})=F^{(N)}(F^{(N-1)}\cdots(F^{(1)}(\boldsymbol{X}))) \tag{1}$$

$F^{(i)}$代表DNN第i层的计算输出。这些层可以是卷积、池化或者其他形式的神经网络层。DNN的最后一层一般采用Softmax层，定义为$F^{(N)}(\boldsymbol{Z})_i=\mathrm{Softmax}(\boldsymbol{Z})_i=\exp(z_i)/\sum_{i=1}^{m}\exp(z_i)$，$\boldsymbol{Z}=F^{(N-1)}(\cdot)$则是前一层(又称最后一个隐藏层)的输出向量。最后的预测标签则由$y=\mathrm{argmax}_{i=1,\cdots,m}F(\boldsymbol{X})_i$得到，其中$F(\boldsymbol{X})=\mathrm{Softmax}(\boldsymbol{Z})$。DNN的损失函数定义为$L(\boldsymbol{X})=-\sum_{i=1}^{m}y\cdot\lg(F^{(N)}(\boldsymbol{Z})_i)$。

2.2 对抗样本

Szegedy等人[7]等人首先发现图像中存在对抗样本的现象。即在$\mathbb{R}^{m\times n}$空间中，对于图像$\boldsymbol{X}$，在ℓ_p范数内存在$\boldsymbol{X}'=\boldsymbol{X}+\boldsymbol{\delta}$, $\|\boldsymbol{\delta}_p\|<C$, $\boldsymbol{X}'$与$\boldsymbol{X}$在人类视觉上几乎没有区别，但DNN会将其判断为与$\boldsymbol{X}$不同的类，即

$$F(\boldsymbol{X}')=F(\boldsymbol{X}+\boldsymbol{\delta})=\boldsymbol{y}',\ \text{s.t.}\ \boldsymbol{y}'\neq\boldsymbol{y} \tag{2}$$

本文将这样的$\boldsymbol{X}'$称为对抗样本。通常情况下，扰动$\boldsymbol{\delta}$通过ℓ_p范数来约束$(p\in\{2,\infty\})$，即$\|\boldsymbol{X}'-\boldsymbol{X}\|_p\leq C$。

2.3 威胁模型

目前有很多对抗样本的生成方法，但这些方法都是在一定的假设限制下进行的[9]。由于对手的攻击行为很大程度上决定了对抗样本的强度。如果攻击行为不被限制，对手甚至可以使用任意图像替换给定的图像，这就违背了对抗样本的定义。为此，我们把这些攻击行为定义为威胁模型，通常包含攻击目标和攻击能力。

(1) 攻击目标

威胁模型中的攻击目标可以被定义为一个需要被检测和防御的具体式子。在DNN中，对于攻击目标的划分有利于我们明确这个具体式子。因此，威胁模型中，对于攻击目标的划分至关重要。可以将攻击目标具体划分为2类，包括无目标攻击和有目标攻击。无目标攻击是指改变对抗样本的类别至任意一个非正确类。有目标攻击是指改变对抗样本的类别至指定的一个非正确类。正式地说，有目标攻击是无目标攻击的一个子集，而对于对抗样本的防御方法来说，防御两者的难易程度并不会有所区别。因此，本文提出的2阶对抗样本属于目前更为主流的无目标攻击。

(2) 攻击能力

对抗样本还可以根据对手掌握目标分类器信息的多少来定义攻击能力，分为白箱攻击和黑箱攻击。白箱攻击是指攻击者几乎知道关于DNN的所有信息，包括训练数据、激活函数、拓扑结构、权重系数等。黑箱攻击则假设攻击者无法获得已训练的DNN内部信息，仅能获得模型的输出，包含标签和置信度。因为需要掌握目标DNN的梯度信息，2阶对抗样本属于白箱攻击。

3　对抗训练防御方法

3.1　问题的提出

目前最有效的对抗训练方法是由Madry等人[13]提出的PGD对抗训练。从优化的观点出发，对抗训练被定义为关于鞍点的优化问题：

$$\boldsymbol{\theta}^{*}=\underset{\boldsymbol{\theta}}{\operatorname{argmin}}\,\mathbb{E}_{(\boldsymbol{X},y)\in D}\left[\max_{||\boldsymbol{X}'-\boldsymbol{X}||_{\infty}\leq\varepsilon}L(f(\boldsymbol{X}';\boldsymbol{\theta}),y)\right] \tag{3}$$

其中，$(\boldsymbol{X},y)\in D$作为原始训练数据，在约束ε下可获得对抗样本$\boldsymbol{X}'=\boldsymbol{X}+\boldsymbol{\delta}$, $L(\cdot)$为损失函数。

可以发现式(3)是一个内部最大化问题和一个外部最小化问题的组合。内部最大化问题是找到令DNN产生最大损失的对抗样本。外部最小化问题是在某种对抗攻击下，寻找使对抗损失最小的模型参数。由此可见，对抗训练是模型精度和鲁棒性之间的一种最佳平衡。Madry等人[13]认为存在更强大的对抗样本可以更好地解决内部最大化问题，从而训练出更加鲁棒的DNN，但基于一阶对抗样本不能很好地解决这个问题。为此，本文提出2阶对抗样本解决式(3)中的内部最大化问题。

3.2　2阶对抗样本

本节给出了一种对抗样本的2阶生成方法。我们将生成对抗样本的过程定义为一个箱约束的优化问题：

$$\left.\begin{aligned}&\max\ L(\boldsymbol{X}+\boldsymbol{\delta})\\&\text{s.t. }F(\boldsymbol{X}+\boldsymbol{\delta})\neq y,\boldsymbol{X}+\boldsymbol{\delta}\in[0,1]^{n},\|\boldsymbol{\delta}\|_{p}\leq C\end{aligned}\right\} \tag{4}$$

其中，$L(\cdot)$是DNN损失函数的输出。现有的方法一般考虑$\|\boldsymbol{\delta}\|_{p}$中$p\in\{2,\infty\}$。注意到$\ell_{\infty}$范数由于只关注$\boldsymbol{\delta}$中最大的值，导致在梯度下降的过程中极易在两个次优解之间振荡。因此，我们使用ℓ_{2}范数作为距离度量。为了便于计算，将约束条件修改为$\|\boldsymbol{\delta}\|_{2}^{2}\leq C$，得到

$$\left.\begin{aligned}&\max\ L(\boldsymbol{X}+\boldsymbol{\delta})\\&\text{s.t. }F(\boldsymbol{X}+\boldsymbol{\delta})\neq y,\boldsymbol{X}+\boldsymbol{\delta}\in[0,1]^{n},\|\boldsymbol{\delta}\|_{2}^{2}\leq C\end{aligned}\right\} \tag{5}$$

式(5)的目的是在模型错误分类的条件下，得到约束$\|\boldsymbol{\delta}\|_{2}^{2}\leq C$内能够最大化$L(\boldsymbol{X}+\boldsymbol{\delta})$的扰动$\boldsymbol{\delta}$。然而，考虑到DNN的高度非线性，式(5)难以直接进行优化。为此，我们在以X为中心的微小邻域内利用2阶泰勒展开近似$L(\boldsymbol{X}+\boldsymbol{\delta})$的输出，得到局部邻域内满足凸性的目标函数。首先，计算损失函数关于$\boldsymbol{X}$的梯度矩阵$\nabla L(\boldsymbol{X})$和Hesse矩阵$\nabla^{2}L(\boldsymbol{X})$：

$$\nabla L(\boldsymbol{X})=\left[\frac{\partial L(\boldsymbol{X})}{\partial \boldsymbol{X}_{i}}\right]_{n\times 1} \tag{6}$$

$$\nabla^{2}L(\boldsymbol{X})=\left[\frac{\partial^{2}L(\boldsymbol{X})}{\partial \boldsymbol{X}_{i}\partial \boldsymbol{X}_{j}}\right]_{n\times n} \tag{7}$$

对损失函数$L(\boldsymbol{X})$进行求导的过程中，与训练过程中的反向传播不同，式(6)与式(7)是对$\boldsymbol{X}$而非DNN的参数进行求导。假设DNN为K层的全连接神经网络，根据链式求导法则，可得

$$\frac{\partial L(\boldsymbol{X})}{\partial \boldsymbol{X}_{i}}=\frac{\partial L(\boldsymbol{X})}{\partial F^{K}(\boldsymbol{X})}\cdot\frac{\partial F^{K}(\boldsymbol{X})}{\partial \boldsymbol{X}_{i}} \tag{8}$$

在此之后，递归地对每一个网络层进行微分：

$$\begin{aligned}\frac{\partial L(\boldsymbol{X})}{\partial \boldsymbol{X}_{i}}=&\frac{\partial L(\boldsymbol{X})}{\partial F^{K}(\boldsymbol{X})}\cdot\frac{\partial R_{k,y}(\boldsymbol{w}_{k,y}F^{K-1}(\boldsymbol{X})+\boldsymbol{b}_{k,y})}{\partial X_{i}}\\=&\boldsymbol{w}_{k,y}\frac{\partial L(\boldsymbol{X})}{\partial F^{K}(\boldsymbol{X})}\frac{\partial F^{K-1}(\boldsymbol{X})}{\partial \boldsymbol{X}_{i}}\\&\cdot\frac{\partial R_{k,y}(\boldsymbol{w}_{k,y}F^{K-1}(\boldsymbol{X})+\boldsymbol{b}_{k,y})}{\partial \boldsymbol{X}_{i}}\end{aligned} \tag{9}$$

其中，$F^{K-1}(\boldsymbol{X})$代表第$k-1$层的输出，$R_{k,y}$代表第k层中第y个神经元的激活函数，$\boldsymbol{w}_{k,y}$以及$\boldsymbol{b}_{k,y}$分别代表第k层中第y个神经元的权重与偏置。在式(9)中，$\dfrac{\partial F^{K-1}(\boldsymbol{X})}{\partial \boldsymbol{X}_{i}}$作为唯一的未知量可以用链式法则递归得到。将所求出的结果代回式(9)，就可以得到$\nabla L(\boldsymbol{X})$的值。类似地，$\nabla^{2}L(\boldsymbol{X})$可在$\nabla L(\boldsymbol{X})$的基础上利用链式法则计算得到

$$\begin{aligned}\nabla^2 L(\boldsymbol{X}) =& \frac{\partial^2 L(\boldsymbol{X})}{\partial \boldsymbol{X}_i \partial \boldsymbol{X}_j} = \frac{\partial}{\partial \boldsymbol{X}_j}\left[\frac{\partial L(\boldsymbol{X})}{\partial \boldsymbol{X}_i}\right] \\ =& \frac{\partial}{\partial \boldsymbol{X}_j}\left[\boldsymbol{w}_{k,y}\frac{\partial L(\boldsymbol{X})}{\partial F^K(\boldsymbol{X})}\frac{\partial F^{K-1}(\boldsymbol{X})}{\partial \boldsymbol{X}_i}\right. \\ &\left.\times \frac{\partial R_{k,y}(\boldsymbol{w}_{k,y}F^{K-1}(\boldsymbol{X})+\boldsymbol{b}_{k,y})}{\partial \boldsymbol{X}_i}\right] \end{aligned} \tag{10}$$

在计算得到$\boldsymbol{X}$的梯度矩阵$\nabla L(\boldsymbol{X})$和Hesse矩阵$\nabla^2 L(\boldsymbol{X})$后，可得到$\boldsymbol{X}$为中心的微小邻域内$L(\boldsymbol{X}+\boldsymbol{\delta})$的近似输出

$$L(\boldsymbol{X}+\boldsymbol{\delta}) \approx T(\boldsymbol{\delta}) = L(\boldsymbol{X}) + \nabla L(\boldsymbol{X})^T\boldsymbol{\delta} + \frac{1}{2}\boldsymbol{\delta}^T\nabla^2 L(\boldsymbol{X})\boldsymbol{\delta} \tag{11}$$

$T(\boldsymbol{\delta})$在$\boldsymbol{X}$邻域内近似$L(\boldsymbol{X}+\boldsymbol{\delta})$的输出，在式(5)的基础上，得到新的目标函数

$$\left.\begin{aligned}&\max\ T(\boldsymbol{\delta}) \\ &\text{s.t. } F(\boldsymbol{X}+\boldsymbol{\delta}) \neq y, \boldsymbol{X}+\boldsymbol{\delta} \in [0,1]^n, \|\boldsymbol{\delta}\|_2^2 \le C\end{aligned}\right\} \tag{12}$$

其中，C是对$\boldsymbol{\delta}$的约束，C的值越大意味着对抗样本越强，但对抗样本的隐蔽性会变弱，反之亦然。从经验上来说，最为合适的C是满足式(12)的最小C。图1展示了MNIST和CIFAR10数据集中C与优化过程中损失函数的关系。获得2阶对抗样本后，我们将它们用于对抗训练，如表1所示。

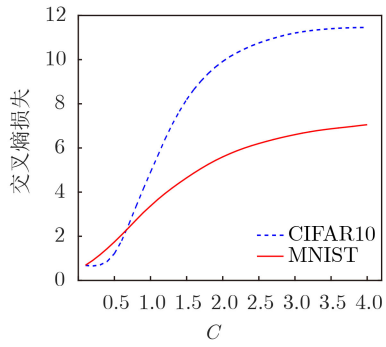

图 1 C与优化过程中交叉熵损失函数的关系

4 理论分析

本文将攻击成功需要的最少扰动定义为扰动下界，衡量2阶对抗样本与最强1阶对抗样本PGD的强度。假设DNN是一个简单的0-1二分类器，

表 1 基于2阶对抗样本的对抗训练算法

输入 $\boldsymbol{X}$为数据集；T为训练批次；M为训练集大小；n为梯度下降迭代次数；τ为学习率

输出 q为模型参数

初始化模型参数 q
for epoch $= 1, 2, \cdots, T$ **do**
  for $m = 1, 2, \cdots, M$ **do**
    $\nabla L(X) \leftarrow \left[\frac{\partial L(\boldsymbol{X})}{\partial \boldsymbol{X}_i}\right]_{n\times 1}$
    $\nabla^2 L(\boldsymbol{X}) \leftarrow \left[\frac{\partial^2 L(\boldsymbol{X})}{\partial \boldsymbol{X}_i \partial \boldsymbol{X}_j}\right]_{n\times n}$
    $T(\boldsymbol{\delta}) = L(\boldsymbol{X}) + \nabla L(\boldsymbol{X})^T\delta + \frac{1}{2}\boldsymbol{\delta}^T\nabla^2 L(\boldsymbol{X})\boldsymbol{\delta}$
    for $k = 1, 2, \cdots, n$ **do**
      $\boldsymbol{\delta} \leftarrow \boldsymbol{\delta} + \alpha \cdot \text{sign}(\nabla_\delta T(\boldsymbol{\delta}))$
    end for
    $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \tau \cdot \nabla_\theta L(\boldsymbol{X}+\boldsymbol{\delta})$
  end for
end for

$L(\boldsymbol{X}) = -\lg(F(\boldsymbol{X}))$代表DNN对于类1输入的损失函数。其中，$F(\boldsymbol{X})$是Softmax层关于类1的输出。当且仅当$L(\boldsymbol{X}) < \lg(2)$时，$\boldsymbol{X}$会被正确分类为1。为了克服DNN高度非线性导致的局部最优值问题，我们在输入邻域内用2阶泰勒展开式对$L(\boldsymbol{X})$进行近似，进一步计算得到$T(\boldsymbol{\delta}) = L(\boldsymbol{X}) + \nabla L(\boldsymbol{X})^T\boldsymbol{\delta} + \boldsymbol{\delta}^T\nabla^2 L(\boldsymbol{X})\boldsymbol{\delta}/2$，其中，$\nabla L(\boldsymbol{X})$和$\nabla^2 L(\boldsymbol{X})$分别是$L(\boldsymbol{X})$关于$\boldsymbol{X}$的梯度矩阵和Hesse矩阵。

对于2阶对抗样本，定义目标函数为

$$\left.\begin{aligned}&\min \|\boldsymbol{\delta}\|_2 \\ &\text{s.t. } L(\boldsymbol{X}) + \nabla L(\boldsymbol{X})^{\mathrm{T}}\delta + \frac{1}{2}\boldsymbol{\delta}^{\mathrm{T}}\nabla^2 L(\boldsymbol{X})\boldsymbol{\delta} \ge \lg(2)\end{aligned}\right\} \tag{13}$$

根据式(13)计算2阶对抗样本扰动$\|\delta\|_2$的下界，可以得到

$$\begin{aligned}&L(\boldsymbol{X}) - \lg(2) + \|\nabla L(\boldsymbol{X})\|_2 \cdot \|\boldsymbol{\delta}\|_2 \\ &+ \frac{1}{2}\lambda_{\max}\left(\nabla^2 L(\boldsymbol{X})\right)\|\boldsymbol{\delta}\|_2^2 \ge L(\boldsymbol{X}) \\ &- \lg(2) + \nabla L(\boldsymbol{X})^{\mathrm{T}}\boldsymbol{\delta} + \frac{1}{2}\boldsymbol{\delta}^{\mathrm{T}}\nabla^2 L(\boldsymbol{X})\boldsymbol{\delta} \ge 0\end{aligned} \tag{14}$$

式(14)求解得到$\|\boldsymbol{\delta}\|_2$的下界，即$\|\boldsymbol{\delta}\|_2 \ge \dfrac{\|\nabla L(\boldsymbol{X})\|_2}{\lambda_{\max}(\nabla^2 L(\boldsymbol{X}))}\left(\sqrt{1+\dfrac{2\lambda_{\max}\left(\nabla^2 L(\boldsymbol{X})\right)(L(\boldsymbol{X})-\lg(2))}{\|\nabla L(\boldsymbol{X})\|_2^2}}-1\right)$。

对于步长约束为$\varepsilon = 0.3$的多步迭代PGD，为了便于计算，设置其随机初始化扰动为$\boldsymbol{0}$，通过计算得到$\|\boldsymbol{\delta}\|_2$的下界，即$\|\boldsymbol{\delta}\|_2 \ge \|0.3\text{sign}(\nabla L(\boldsymbol{X}))\|_2$。我们将2阶对抗样本的下界与PGD的下界进行比较：

$$P=\frac{\|\nabla L(\boldsymbol{X})\|_2}{\lambda_{\max}(\nabla^2 L(\boldsymbol{X}))}\left(\sqrt{1+\frac{2\lambda_{\max}(\nabla^2 L(\boldsymbol{X}))\,(L(\boldsymbol{X})-\lg(2))}{\|\nabla L(\boldsymbol{X})\|_2^2}}-1\right)-\|0.3\mathrm{sign}(\nabla L(\boldsymbol{X}))\|_2$$
$$=\frac{\sqrt{\|\nabla L(\boldsymbol{X})\|_2^2+2\lambda_{\max}(\nabla^2 L(\boldsymbol{X})\,(L(\boldsymbol{X})-\lg(2))}-\|\nabla L(\boldsymbol{X})\|_2}{\lambda_{\max}(\nabla^2 L(\boldsymbol{X}))}-\|0.3\mathrm{sign}(\nabla L(\boldsymbol{X}))\|_2 \tag{15}$$

由于DNN的高度非线性，$\lambda_{\max}\left(\nabla^2 L(\boldsymbol{X})\right)\geq 0$。又由于DNN能正确分类原始样本$\boldsymbol{X}$, $L(\boldsymbol{X})-\lg(2)<0$。由此我们得出结论，$P\leq 0$恒成立，即2阶对抗样本的下界始终低于PGD的下界，可证明2阶对抗样本强于PGD。

5　实验

通过实验进行验证：(1)相比于以C&W, Deepfool, FGSM, PGD以及M-DI^2-FGSM为例的典型对抗样本，2阶对抗样本在具有更高隐蔽性的同时具有更高攻击成功率；(2)相比于PGD对抗训练，基于2阶对抗训练抗御对当前典型对抗样本都具有鲁棒性且具有更高的分类准确率。

5.1　实验设置

本文实验的数据集为MNIST和CIFAR10。MNIST是一个包含从数字0到9的10个类的手写体数据集，共包含70000张手写体数字图像，每个图像的大小为28×28像素。实验选取60000张图像作为训练数据，10000张图像作为测试数据。CIFAR-10数据集由60000个32×32彩色图像组成，包含10个类。实验选取50000个图像作为训练数据和10000图像作为测试数据。训练集分为5个训练批次，每个批次有10000个图像。对于MNIST我们使用精度为98.79%的标准LeNet网络。对于CIFAR10我们使用精度为76.97%的标准AlexNet网络。

5.2　评估指标

实验使用4个评估指标，包括ℓ_2，ℓ_∞，PSNR以及ASR。现有的研究普遍用ℓ_2的值来衡量全局添加的扰动量，ℓ_∞来衡量局部(单个像素)添加的扰动量。峰值信噪比(PSNR)作为最广泛使用的评价图片质量的客观度量，可以对对抗样本的隐蔽性进行有效评估。ASR称为对抗样本的攻击成功率，目前被大多数文献用于衡量攻击能力。若生成对抗样本的成功率不是100%，那么这些数据仅取了成功的那部分作为基数。

5.3　评估2阶对抗样本

本文采用机器学习模型攻防库Cleverhans[19]中的C&W，Deepfool，FGSM，以及由原作者给出代码的PGD和M-DI^2-FGSM作比较实验。为保证评估的严谨性，实验采用相同模型架构和测试数据集。其中，C&W的扰动上限$\delta=0.3$，$\varepsilon=0.3$，学习率为0.1；Deepfool的参数设置与C&W相同；而FGSM作为单步迭代法，$\varepsilon=0.3$；PGD作为FGSM的衍生方法，迭代扰动固定为$\varepsilon=0.3$；M-DI^2-FGSM中，$\varepsilon=0.3$。

实验中，从MNIST与CIFAR中随机取出500张可以被DNN正确判断的图片进行测试，实验结果如表2所示。实验证明，在不同数据集中，相比于现有的典型攻击方法，2阶对抗样本不但攻击成功率更高，而且添加的扰动更少。

5.4　对抗训练2阶对抗样本

本文分别采用自然样例、2阶对抗样本，C&W, Deepfool, M-DI^2-FGSM, FGSM以及PGD进行攻击和对抗训练，用于对比攻击效果与防御效果。从MNIST与CIFAR10中随机取出200张可以被初始模型正确判断的图片进行测试。图2是实验结果的热力图表示，横轴表示各种方法产生的对抗样本，纵轴表示用不同对抗样本进行对抗训练得到的对抗训练模型，图中每一个数字代表某一个对抗训练模型对于某一类对抗样本的分类准确率。图2的结果表明：(1)对抗训练产生的对抗训练模型对于特定攻击具有鲁棒性；(2)相比于PGD对抗训练，基于2阶

表 2　不同的对抗样本在MNIST和CIFAR10的对比

	MNIST				CIFAR10			
	ℓ_2	ℓ_∞	PSNR	ASR(%)	ℓ_2	ℓ_∞	PSNR	ASR(%)
本文	1.97	0.24	76.0	100	1.84	0.24	81.4	100
C&W	2.56	0.27	71.4	100	2.42	0.30	79.3	100
Deepfool	3.25	0.30	73.8	88.1	2.92	0.30	74.2	81.4
M-DI^2-FGSM	2.75	0.29	75.6	95.1	3.12	0.25	77.1	91.7
FGSM	3.26	0.30	74.2	54.1	2.34	0.30	75.0	51.3
PGD	2.25	0.26	72.3	100	2.18	0.58	78.7	100

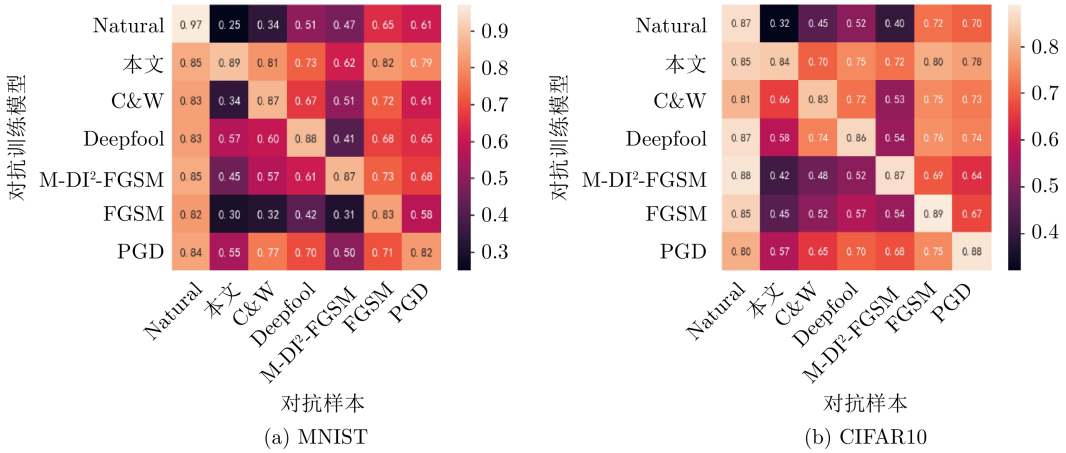

图 2 对抗训练DNN对于对抗样本的分类准确率

对抗样本的对抗训练防御能够防御现存典型1阶对抗样本，且具有更高的分类准确率。

6 结束语

通过理论分析可知，2阶对抗样本的扰动下界低于1阶最强对抗样本PGD，表明2阶对抗样本强于PGD，能够更好地解决对抗训练的内部最大化问题。在MNIST和CIFAR10数据集上的实验表明，相较于现有的典型1阶对抗样本，2阶对抗样本拥有更高的攻击成功率和更高的隐蔽性。相比于PGD对抗训练，基于2阶对抗样本的对抗训练防御能够防御现存典型1阶对抗样本，且具有更高的分类准确率。2阶对抗样本中参数经验值的选取是通过大量实验得到的，将来对参数的选取机制有待进一步研究。目前还未有研究者对对抗样本的在线攻击与线下攻击进行分析，在未来的工作中，我们将进一步研究2阶对抗样本与其他对抗样本在线攻击与线下攻击的不同特征。

参 考 文 献

[1] CHICCO D, SADOWSKI P, and BALDI P. Deep autoencoder neural networks for gene ontology annotation predictions[C]. Proceedings of the 5th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics, Newport Beach, America, 2014: 533–554.

[2] SPENCER M, EICKHOLT J, and CHENG Jianlin. A deep learning network approach to *ab initio* protein secondary structure prediction[J]. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2015, 12(1): 103–112. doi: 10.1109/TCBB.2014.2343960.

[3] MIKOLOV T, DEORAS A, POVEY D, *et al*. Strategies for training large scale neural network language models[C]. 2011 IEEE Workshop on Automatic Speech Recognition & Understanding, Waikoloa, America, 2011: 196–201.

[4] HINTON G, DENG Li, YU Dong, *et al*. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups[J]. *IEEE Signal Processing Magazine*, 2012, 29(6): 82–97. doi: 10.1109/MSP.2012.2205597.

[5] LECUN Y, KAVUKCUOGLU K, FARABET C, *et al*. Convolutional networks and applications in vision[C]. Proceedings of 2010 IEEE International Symposium on Circuits and Systems, Paris, France, 2010: 253–256.

[6] KRIZHEVSKY A, SUTSKEVER I, and HINTON G E. ImageNet classification with deep convolutional neural networks[C]. Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1, Lake Tahoe Nevada, America, 2012: 1097–1105.

[7] SZEGEDY C, ZAREMBA W, SUTSKEVER I, *et al*. Intriguing properties of neural networks[C]. 2nd International Conference on Learning Representations, ICLR 2014, Banff, Canada, 2014.

[8] STALLKAMP J, SCHLIPSING M, SALMEN J, *et al*. Man vs. computer: Benchmarking machine learning algorithms for traffic sign recognition[J]. *Neural Networks*, 2012, 32: 323–332. doi: 10.1016/j.neunet.2012.02.016.

[9] CARLINI N and WAGNER D. Towards evaluating the robustness of neural networks[C]. 2017 IEEE Symposium on Security and Privacy (SP), San Jose, America, 2017: 39–57.

[10] MOOSAVI-DEZFOOLI S M, FAWZI A, and FROSSARD P. DeepFool: A simple and accurate method to fool deep neural networks[C]. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, America, 2016: 2574–2582. doi: 10.1109/CVPR.2016.282.

[11] XIE Cihang, ZHANG Zhishuai, ZHOU Yuyin, *et al*. Improving transferability of adversarial examples with input diversity[C]. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, America, 2019: 2725–2734.

[12] LEE J G, JUN S, CHO Y W, *et al*. Deep learning in

medical imaging: General overview[J]. *Korean Journal of Radiology*, 2017, 18(4): 570–584. doi: 10.3348/kjr.2017.18.4.570.

[13] MADRY A, MAKELOV A, SCHMIDT L, *et al*. Towards deep learning models resistant to adversarial attacks[C]. ICLR 2018 Conference Blind Submission, Vancouver, Canada, 2018.

[14] GOODFELLOW I J, SHLENS J, and SZEGEDY C. Explaining and harnessing adversarial examples[C]. 3rd International Conference on Learning Representations, San Diego, America, 2015.

[15] ARAUJO A, MEUNIER L, PINOT R, *et al*. Robust neural networks using randomized adversarial training[EB/OL]. https://arxiv.org/pdf/1903.10219.pdf, 2020.

[16] LAMB A, BINAS J, GOYAL A, *et al*. Fortified networks: Improving the robustness of deep networks by modeling the manifold of hidden representations[C]. ICLR 2018 Conference Blind Submission, Vancouver, Canada, 2018.

[17] XU Weilin, EVANS D, and QI Yanjun. Feature squeezing: Detecting adversarial examples in deep neural networks[C]. Network and Distributed Systems Security Symposium (NDSS), San Diego, America, 2018. doi: 10.14722/ndss.2018.23198.

[18] BELINKOV Y and BISK Y. Synthetic and natural noise both break neural machine translation[C]. ICLR 2018 Conference Blind Submission, Vancouver, Canada, 2018.

[19] YANG Yuzhe, ZHANG Guo, KATABI D, *et al*. ME-net: Towards effective adversarial robustness with matrix estimation[C]. Proceedings of the 36th International Conference on Machine Learning, Long Beach, America, 2019.

钱亚冠：男，1976年生，副教授，研究方向为人工智能安全.
张锡敏：女，1996年生，硕士生，研究方向为对抗机器学习.
王　滨：男，1978年生，研究员，研究方向为网络与信息安全.
顾钊铨：男，1989年生，教授，研究方向为人工智能安全.
李　蔚：女，1978年生，副教授，研究方向为计算机视觉.
云本胜：男，1980年生，副教授，研究方向为数据挖掘.

责任编辑：陈　倩