# 基于信息瓶颈的深度学习模型鲁棒性增强方法

董庆宽 何浚霖*

(西安电子科技大学综合业务网国家重点实验室 西安 710071)

**摘 要**：作为深度学习技术的核心算法，深度神经网络容易对添加了微小扰动的对抗样本产生错误的判断，这种情况的出现对深度学习模型的安全性带来了新的挑战。深度学习模型对对抗样本的抵抗能力被称为鲁棒性，为了进一步提升经过对抗训练算法训练的模型的鲁棒性，该文提出一种基于信息瓶颈的深度学习模型对抗训练算法。其中，信息瓶颈以信息论为基础，描述了深度学习的过程，使深度学习模型能够更快地收敛。所提算法使用信息瓶颈理论提出的优化目标推导出的结论，将模型中输入到线性分类层的张量加入损失函数，通过样本交叉训练的方式将干净样本与对抗样本输入模型时得到的高层特征对齐，使模型在训练过程中能够更好地学习输入样本与其真实标签的关系，最终对对抗样本具有良好的鲁棒性。实验结果表明，所提算法对多种对抗攻击均具有良好的鲁棒性，并且在不同的数据集与模型中具有泛化能力。

**关键词**：深度学习；对抗训练；信息瓶颈；对抗样本；鲁棒性

**中图分类号**：TN911.7; TP18 **文献标识码**：A **文章编号**：1009-5896(2023)06-2197-08

**DOI**: 10.11999/JEIT220603

# Robustness Enhancement Method of Deep Learning Model Based on Information Bottleneck

DONG Qingkuan HE Junlin

(*State Key Laboratory of Integrated Service Networks*, *Xidian University*, *Xi'an* 710071, *China*)

**Abstract**: As the core algorithm of deep learning technology, deep neural network is easy to make wrong judgment on the adversarial examples with imperceptive perturbation. This situation brings new challenges to the security of deep learning model. The resistance of deep learning model to adversarial examples is called robustness. In order to improve the robustness of the model trained by adversarial training algorithm, an adversarial training algorithm of deep learning model based on information bottleneck is proposed. Among this, information bottleneck describes the process of deep learning based on information theory, so that the deep learning model can converge faster. The proposed algorithm uses the conclusions derived from the optimization objective proposed based on the information bottleneck theory, adds the tensor input to the linear classification layer in the model to the loss function, and aligns the clean samples with the high-level features obtained when the adversarial samples are input to the model by means of sample cross-training, so that the model can better learn the relationship between the input samples and their true labels during the training process and has finally good robustness to the adversarial samples. Experimental results show that the proposed algorithm has good robustness to a variety of adversarial attacks, and has generalization ability in different data sets and models.

**Key words**: Deep learning; Adversarial training; Information bottleneck; Adversarial example; Robustness

## 1 引言

随着深度学习的发展，生活中越来越多的地方开始将各种技术与深度学习相结合，在自然语言处理、机器视觉等多个领域中取得了令人瞩目的成果。但深度学习模型实际上相当脆弱，当输入被添加了人眼无法分辨的微小的扰动时，会导致模型产生高置信度的误判。这种含有微小扰动且能干扰模型正常工作的输入被称为对抗样本。

自从Szegedy等人[1]提出对抗样本的概念后，多种对抗攻击与对抗防御算法被提出。在图像分类方面，常见的攻击算法有基于反向梯度的FGSM (Fast Gradient Sign Method)[2]、被证明是最强1阶攻击

收稿日期：2022-05-12；改回日期：2022-10-13；网络出版：2022-10-20
*通信作者： 何浚霖 425764309@qq.com
基金项目：陕西省自然科学基础研究计划(2020JM-184)
Foundation Item: The Science Basic Research Plan in Shaanxi Province of China (2020JM-184)

的PGD (Project Gradient Descent)[3]，基于超平面分类的DeepFool[4]以及基于优化的C&W (Carlini and Wagner attacks)[5]。这些算法被认为是训练鲁棒性网络的有效手段，并且被广泛用于评判深度学习模型的鲁棒性。

对抗防御算法通常通过使用更大的训练集、修改模型结构、修改损失函数等方法来提升模型的鲁棒性。这些防御算法大致可以分为对抗训练算法[3,6–12]和图像预处理算法[13–17]两大类。

图像预处理算法将对抗样本中的扰动看作噪声，通过对输入图像去噪的方式来防御对抗攻击。但是在去除输入样本中的噪声时，容易将样本中包含的信息也一同去除，会使模型对干净样本的正确率大幅下降。

对抗训练使用梯度攻击算法获得对抗样本，使用干净样本与对抗样本一同训练模型，使训练的模型拥有更好的鲁棒性，是目前最常用的方法，也是目前最优秀的对抗防御方法之一。但该类算法的训练过程耗时较长，模型收敛较慢，同时因为训练时同时使用干净样本与对抗样本，模型对干净样本的正确率也会产生一定幅度的下降。

Tishby等人[18–20]于1999年以信息论为基础首次提出信息瓶颈方法，给出了优化问题的数学定义和迭代算法，并且证明了算法的收敛性，指出深度神经网络的实质是对信息的压缩。并且尝试使用信息瓶颈理论对深度学习网络的特征拟合与特征压缩这两个阶段进行解释。Kolchinsky等人[21]与Alemi等人[22]类似，通过在网络中添加VAE编码器结构来实现信息瓶颈理论，仅使用干净样本来训练模型来增强模型的鲁棒性。他们的工作将信息瓶颈理论提出的优化目标推广到了包括离散域与连续域的更一般的领域，为本文将信息瓶颈引入对抗训练提供了理论推导基础。

为了进一步提升对抗训练的鲁棒性，本文提出一种基于信息瓶颈理论的对抗训练防御算法，主要贡献如下：

(1)将信息瓶颈理论引入了对抗训练，使深度学习模型能够更好地学习输入数据与真实标签之间的关系，将对抗样本的高层特征向干净样本的高层特征对齐，从而提升深度学习模型的鲁棒性。

(2)在多个数据集与深度学习模型中对算法进行了仿真，展示了算法在不同对抗攻击、不同深度学习模型以及不同的数据集中都具有良好的防御性能与泛化性能。

## 2 相关工作

对抗训练作为最常用的方法，无需修改模型的结构，只需要将对抗样本加入训练集，就能够使训练的模型拥有更好的鲁棒性。给定由$\theta$参数化的网络$f_\theta$，数据集$(x_i,y_i)$，损失函数$L$，扰动值$\delta$与扰动的限制范围$\varepsilon$对抗训练通常可以看作以下的优化问题。

$$\min_\theta \sum_i \max_{\delta\le\varepsilon} L\left(f_\theta\left(x_i+\delta\right),y_i\right) \tag{1}$$

对抗训练利用对抗攻击使模型获得内部最大化，同时使用梯度下降法对模型进行训练，使模型获得外部最小化，这正是Madry等人[3]提出的min-max优化框架。为了能够更好地获得内部最大化，多种对抗攻击算法被提出。

FGSM算法[2]是最早来近似内部最大化的方法，能够使用较少的时间使深度学习模型获得鲁棒性，算法的公式化呈现为式(2)。给定由$\theta$参数化的网络$f_\theta$，数据集$(x,y)$，即可算出当前输入的梯度$\nabla_x f_\theta\left(x,y\right)$。将梯度用符号函数sign()与扰动范围$\varepsilon$限制后，就得到了最终的扰动$\delta$。将扰动与输入叠加就能够获得对抗样本，使深度学习模型作出错误的判断。

$$\delta=\varepsilon\text{sign}\left(\nabla_x f_\theta\left(x,y\right)\right) \tag{2}$$

使用FGSM算法进行对抗训练所获得的深度学习模型仅对FGSM算法本身生成的样本有良好的鲁棒性，原因是FGSM算法只迭代了1次，获得的梯度并不是准确的，因而PGD算法[3]在FGSM算法的基础上对扰动进行$N$次迭代，每次迭代都将获得的扰动限制在指定的范围内，算法的公式化呈现为式(3)。这种方式能够获得更贴近于模型的梯度，因而获得的扰动也更有效，进一步提升了深度学习模型的鲁棒性。因为使用PGD进行对抗训练的每次训练都需要经过$N$次迭代，因此耗时将会是普通训练的$N$倍。

$$x_{n+1}=\text{clip}\left(x_n+\varepsilon\text{sign}\left(\nabla_x f_\theta\left(x,y\right)\right)\right) \tag{3}$$

为了避免PGD算法带来的大量资源消耗，Fast-AT算法[6]对FGSM算法进行了改进，为FGSM算法在开始前添加非零的随机初始化步骤，并在最后将扰动限制在指定范围内，算法的公式化呈现为式(4)。无论随机初始化步骤将为扰动带来什么样的值，这一操作都能够使FGSM在不牺牲速度的情况下拥有比肩使用PGD进行对抗训练的性能。

$$\delta=\text{clip}\left(\text{random}\left(-\varepsilon,\varepsilon\right)+\varepsilon\text{sign}\left(\nabla_x f_\theta\left(x,y\right)\right)\right) \tag{4}$$

在监督学习中，信息瓶颈将深度学习模型的训练过程表述为最大化地压缩输入，并且保留关于标签的信息。信息瓶颈理论使用互信息来衡量输出中含有的输入信息，将深度学习模型的训练过程看作最小化隐藏变量与输入数据之间的互信息，并最大

化隐藏变量与输出数据的互信息的过程。Alemi等人[22]将马尔可夫过程引入信息瓶颈，并使用变分推理构造优化目标的下界后，使深度学习网络在训练时能够更快地收敛，同时使输入更难通过信息瓶颈传递细小、特殊的扰动，从而使模型对对抗输入更具鲁棒性。该方法在深度学习模型中加入了新的结构，改变了模型的结构，不能简单地进行部署。

本文提出的算法将信息瓶颈引入了对抗训练中，能够在不对深度学习模型进行修改的前提下，在输入干净样本时正确率仅产生小幅下降，同时大幅提升模型对其他对抗攻击的鲁棒性。

## 3　基于信息瓶颈的模型鲁棒性增强方法

深度神经网络通常由前方的特征提取层与后方的线性分类层组成，通常能够将线性分类层的输入看作深度神经网络将输入通过特征提取层后获得的高层特征。高层特征确定了深度神经网络最终将输入样本分为哪一类，如图1。

信息瓶颈理论将深度神经网络的高层特征看作信息瓶颈，高层特征应该尽可能地保留与其对应的真实标签有关的信息，并遗忘其他无关的信息。因此，将深度神经网络中的高层特征抽象为隐藏变量$z$，记网络参数为$\theta$，输入的对抗样本为$\tilde{x}$，因而需要使隐藏变量$z$与网络输出$y$的互信息最大，使隐藏变量$z$与网络输入$\tilde{x}$的互信息最小，即最大化目标函数 $L=I(z,y;\theta)-\beta I(z,\tilde{x};\theta)$。下面将从由目标函数推导至损失函数与算法的实际应用两个方面来讲述。

### 3.1　训练优化的目标函数推导

首先通过信息瓶颈理论提出的目标函数$L=I(z,y;\theta)-\beta I(z,\tilde{x};\theta)$推导至损失函数，因为整个过程未改变网络结构且都在同一个网络中完成，下面的推导将省略网络参数$\theta$，公式变量对照如表1。

首先来推导$I(z,y)$的下界。由互信息的定义，可以得到隐藏变量$z$与网络输出$y$之间的互信息$I(z,y)$。

$$I(z,y)=H(p(y))-H(p(y|z)) \tag{5}$$

其中，$H(p(y))$是变量$y$的熵，$p(y)$为网络输出$y$的边缘概率分布。算法的目的是希望对抗样本经由目标网络后获得的高层特征与干净样本经由目标网络后获得的高层特征尽可能相似，从而使目标网络在不同的情况下都能给出正确的输出。因而在此引入与输入样本相对应真实标签的边缘分布概率$q(y)$。

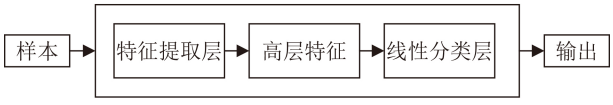


图 1 深度神经网络示意图

当$p(y|z)$与$q(y)$相同时，目标网络就能够在输入是对抗样本时获得与输入是干净样本相同的输出。对于两个概率分布，本文在这里使用KL散度来衡量二者之间的差距。

$$\mathrm{KL}\left[p(y|z)||q(y)\right]\geq 0 \tag{6}$$

且交叉熵

$$\mathrm{CE}\left[p(y)||q(y)\right]=H(p(y))+\mathrm{KL}\left[p(y)||q(y)\right] \tag{7}$$

因而有

$$\begin{aligned} I(z,y)\geq& H(p(y))-H(p(y|z))-\mathrm{KL}\left[p(y|z)||q(y)\right]\\ =&H(p(y))-\mathrm{CE}\left[p(y|z)||q(y)\right] \end{aligned} \tag{8}$$

其中，$H(p(y))$与网络参数无关，仅与真实标签相关，因此是常数，与优化过程无关。由式(4)确定了隐藏变量$z$与输出$y$的互信息$I(z,y)$的下界，下面将推导$I(z,\tilde{x})$的上界。

$p(z)$为输入对抗样本时隐藏变量的边缘分布，设$q(z)$为输入干净样本时隐藏变量的边缘分布

$$\begin{aligned} I(z,\tilde{x})=&\iint p(\tilde{x},z)\ln\frac{p(z|\tilde{x})}{p(z)}\mathrm{d}z\mathrm{d}\tilde{x}\\ =&\iint p(\tilde{x},z)\ln p(z|\tilde{x})\,\mathrm{d}z\mathrm{d}\tilde{x}\\ &-\iint p(\tilde{x},z)\ln p(z)\,\mathrm{d}z\mathrm{d}\tilde{x}\\ =&\iint p(\tilde{x},z)\ln p(z|\tilde{x})\,\mathrm{d}z\mathrm{d}\tilde{x}\\ &-\int p(z)\ln p(z)\,\mathrm{d}z \end{aligned} \tag{9}$$

由KL散度的非负性，有 $\mathrm{KL}\left[p(z)||q(z)\right]\geq 0$

$$\Rightarrow\int p(z)\ln p(z)\mathrm{d}z\geq\int p(z)\ln q(z)\mathrm{d}z$$

则

$$\begin{aligned} I(z,\tilde{x})\leq&\iint p(\tilde{x},z)\ln p(z|\tilde{x})\,\mathrm{d}z\mathrm{d}\tilde{x}\\ &-\iint p(z)\ln q(z)\,\mathrm{d}z\mathrm{d}\tilde{x}\\ =&\iint p(\tilde{x},z)\ln\frac{p(z|\tilde{x})}{q(z)}\mathrm{d}z\mathrm{d}\tilde{x}\\ =&\iint p(\tilde{x})\,p(z|\tilde{x})\ln\frac{p(z|\tilde{x})}{q(z)}\mathrm{d}z\mathrm{d}\tilde{x}\\ =&\int p(\tilde{x})\,\mathrm{KL}\left[p(z|\tilde{x})||q(z)\right]\mathrm{d}\tilde{x} \end{aligned} \tag{10}$$

**表 1　公式变量对照表**

| 公式变量 | 名称 | 公式变量 | 名称 |
|---|---|---|---|
| $L$ | 目标函数 | $p(\cdot)$ | 边缘概率分布 |
| $\tilde{x}$ | 对抗样本 | $q(\cdot)$ | 边缘概率分布 |
| $y$ | 网络输出 | $\beta$ | 信息瓶颈通过率 |
| $z$ | 隐藏变量 | $H(\cdot)$ | 熵 |
| $I(\cdot)$ | 互信息 | $\mathrm{KL}(\cdot)$ | KL散度 |
| $L_{\mathrm{IB}}$ | 损失函数 | $\mathrm{CE}(\cdot)$ | 交叉熵 |

可得 $\int p(\tilde{x})\,\mathrm{KL}\left[p(z|\tilde{x})\,||q(z)\right]\mathrm{d}\tilde{x}$ 是 $I(z,\tilde{x})$ 的上界。

将上述推导的下界和上界分别代入目标函数 $L=I(z,y;\theta)-\beta I(z,\tilde{x};\theta)$，得其下界

$$\begin{aligned} I(z,y)-\beta I(z,\tilde{x}) \geq & -\mathrm{CE}((p(y|z)||q(y)) \\ & -\beta\iint p(\tilde{x})\,p(z|\tilde{x})\ln\frac{p(z|\tilde{x})}{q(z)}\mathrm{d}z\mathrm{d}\tilde{x} \\ & +H(p(y)) \end{aligned} \tag{11}$$

可以看到式(11)确定了需要最大化的目标函数的下界，因此只需要将下界最大化就可以优化目标函数。

记$-\mathrm{CE}[(p(y|z)||q(y)]-\beta\iint p(\tilde{x})p(z|\tilde{x})\ln\frac{p(z|\tilde{x})}{q(z)}\mathrm{d}z\mathrm{d}\tilde{x}=\hat{L}$。

因为$H(p(y))$仅与真实标签相关，与网络参数无关。因此在仿真过程中，可以通过最大化$\hat{L}$来使目标函数$L$最大化。在仿真中需要对下界进行估算，使用经验概率分布

$$p(\tilde{x})=\frac{1}{N}\sum_{n=1}^{N}\delta_{\tilde{x}_n}(x) \tag{12}$$

将式(8)代入$\hat{L}$可以得到

$$\begin{aligned} \hat{L}\approx & -\mathrm{CE}\left[p(y|z)||q(y)\right] \\ & -\frac{1}{N}\sum_{n=1}^{N}\left[\beta\int p(z|\tilde{x}_n)\ln\frac{p(z|\tilde{x}_n)}{q(z)}\mathrm{d}z\right] \end{aligned} \tag{13}$$

对$\hat{L}$添加一个负号，就能够将最大化转为最小化。

$$L_{IB}=\mathrm{CE}\left[p(y|z)||q(y)\right]+\frac{1}{N}\sum_{n=1}^{N}\beta\mathrm{KL}\left[p(z|\tilde{x}_n)\,||q(z)\right] \tag{14}$$

其中，$\mathrm{CE}\left[p(y|z)||q(y)\right]$表示对抗样本进入模型后的交叉熵，这部分可以看作正常训练损失函数。

$\beta\mathrm{KL}\left[p(z|\tilde{x}_n)\,||q(z)\right]$是模型中由高层特征获得的后验分布$p(z|\tilde{x}_n)$与先验分布$q(z)$之间的KL散度，$\beta$控制了信息通过瓶颈的多少，随使用的数据集进行调整。Shamir等人[23], Still等人[24], Alemi等人[22]对$\beta$的取值进行了相关研究，通常$\beta\leq10^{-2}$。

与正常训练损失函数相比，$\beta\mathrm{KL}\left[p(z|\tilde{x}_n)\,||q(z)\right]$这一部分的增加使模型能够在训练过程中，以对齐模型的高层特征的方法使模型在面对对抗样本时能够获得与干净样本相似的高层特征，从而做出正确的判断。

### 3.2 基于信息瓶颈的深度学习防御模型训练方法

基于上述算法推导，本文提出一个基于信息瓶颈的防御模型，包含3部分：目标模型、样本交叉训练与特征对齐。目标模型可以是目前使用的任意图像分类模型。样本交叉训练可以让目标模型保持在干净样本下的正确率，同时也能更好地学习样本间的联系。特征对齐是使用基于信息瓶颈理论推导而来的结论，能够使目标模型更好地学习输入样本与真实标签之间的关系，使模型的高层特征对齐，不容易被微小的扰动干扰。

图2展示了整个算法的流程。首先使用干净样本训练一个高正确率的模型，记为DNN0，同时将目标模型记为DNN1。将干净样本输入DNN0，从模型中输入线性分类层中取得此时的高层特征$q(z)$。然后使用Fast-AT算法生成对抗样本，将对抗样本输入DNN1，从模型中输入线性分类层中取得此时的高层特征$p(z|\tilde{x}_n)$与输出$p(y|z)$。

将以上取得的数据与真实标签$q(y)$一同代入式(10)作为损失函数来更新模型的参数。同样地，将干净样本输入DNN1进行1次如上过程的模型参数更新，最后获得的模型DNN1就是所需的模型。

## 4 实验结果及分析

### 4.1 实验设置

算法使用Python3.8与Pytorch1.9.0在GeForce RTX 3070上实现与测试，优化器使用Adam[25]，对抗攻击算法使用FoolBox开源库中的实现的FGSM[2], PGD[3], DeepFool[3]与C&W[4]对模型进行测试。

实验数据集使用CIFAR10，MNIST与Fashion-MNIST数据集，Fashion-MNIST的复杂度略高于MNIST，数据集信息如表2。

实验共包含3项测试，第1项是算法与不同防御方法的比较，第2项是算法在不同深度学习模型中的泛化性，第3项是算法在不同数据集中的泛化性。在CIFAR10数据集上进行第1项与第2项的测试，在

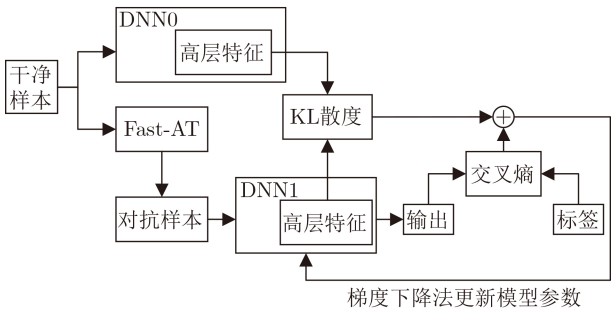


图 2 算法流程图

表 2 使用的数据集信息

| 数据集名称 | 图片大小 | 是否彩色 | 数量($10^4$张) | 类别(种) | $\beta$ |
|---|---|---|---|---|---|
| CIFAR100 | 32×32 | 是 | 6 | 20 | $10^{-5}$ |
| CIFAR10 | 32×32 | 是 | 6 | 10 | $10^{-5}$ |
| MNIST | 28×28 | 否 | 7 | 10 | $10^{-3}$ |
| Fashion-MNIST | 28×28 | 否 | 7 | 10 | $10^{-3}$ |

MNIST与Fashion-MNIST数据集上进行第2项与第3项的测试。防御算法对比实验采用了TRADES[26], ADT[8], Feature Scatter[27]与Fast_AT[5]算法进行结果对比。

### 4.2 实验结果

在CIFAR10数据集上，目标模型使用ResNet18[28]和VGG16[29]，使用FGSM, PGD, DeepFool, C&W攻击算法进行测试。超参数设置如下：ResNet18以0.1的学习率训练120代，以0.01的学习率训练5代，以0.001的学习率训练5代。VGG16以0.1的学习率训练100代，0.01的学习率训练45代，0.001的学习率训练10代。其中，对比算法中，ADT算法[8]通过对抗分布训练框架学习对抗性分布来获取对抗样本进行对抗训练。TRADES算法[26]使用分类校准损失理论提出了新的损失函数对模型进行对抗训练。Feature Scatter算法[27]使用特征散射方法，用无监督的方式生成对抗样本进行对抗训练。

表3对比了不同防御方法在CIFAR10数据集上的鲁棒性，各算法均采用ResNet18模型在CIFAR10数据集上进行训练，且扰动强度均为8，因此对比算法除Fast_AT外，均直接引用了相关论文中的数据。

实验结果表明在CIFAR10数据集上，本文提出的算法对各种攻击都具有较好的鲁棒性。本文提出的算法在干净样本上的正确率与无防御的模型相比，由93.0%下降到了85.0%，但在面对FGSM, PGD-20与PGD-100攻击时，正确率均有较大幅度的提升。

TRADES(1/λ=1)、ADT与Feature Scatter防御算法在干净样本上的正确率略微高于本文提出的算法，但在面对各种对抗攻击时，本文提出的算法的正确率均高于上述防御算法。

表4的实验结果表明在CIFAR10数据集上，C&W攻击在1 000次迭代的设置下，对二者效果均不明显，因此正确率与输入干净样本时相似，无防御时的正确率要高于本算法。对于其他攻击算法，本文提出的算法在Resnet18模型和VGG16模型上在扰动强度为0到16时，面对对抗攻击准确率的下降均小于15%，以少量降低干净样本正确率的代价使模型对多数对抗攻击算法具有了鲁棒性。证明了本文提出的算法对不同的深度学习模型具有良好的泛化性能。

图3展示了ResNet18模型中，由2幅CIFAR10图片得到的部分类激活图与特征图。其中类激活图由ResNet18.Layer4的输出获得，特征图由ResNet18.Layer1的输出获得。从类激活图可以看到，正常训练的模型关注的像素范围较小，容易受到扰动的干扰产生误判，由本文提出算法进行训练的模型将更多的像素纳入了关注的范围，能够更好地抵抗扰动，从而做出正确的判断。从特征图可以看到，正常模型输入干净样本时，特征图可以看到明显的轮廓，而输入对抗样本时，特征图的轮廓则不再清晰。由本文提出算法进行训练的模型在输入对抗样本时仍然能够得到较清晰的轮廓，使模型能够将更多信息传递下去，最终作出正确的判断。

在CIFAR100数据集20分类任务上，使用ResNet18

**表 3　不同防御方法在CIFAR10数据集上的鲁棒性(%)**

| | 干净样本 | FGSM | PGD-20 | PGD-100 | C&W | DeepFool |
|---|---|---|---|---|---|---|
| 无防御 | 93.0 | 65.9 | 54.2 | 49.7 | 92.0 | 41.9 |
| TRADES(1/$\lambda$=6) | 84.9 | 61.0 | 56.6 | 56.4 | 81.2 | 61.3 |
| TRADES(1/$\lambda$=1) | 88.6 | 56.3 | 49.1 | 48.9 | 84.0 | 59.1 |
| ADT | 86.8 | 60.4 | 52.1 | 51.6 | 52.4 | – |
| Feature Scatter | 90.0 | 78.4 | 70.5 | 68.6 | 62.6 | – |
| Fast_AT | 78.6 | 72.4 | 72.3 | 72.2 | 78.5 | 71.1 |
| 本文 | 85.0 | 79.0 | 78.8 | 78.7 | 84.9 | 73.5 |

**表 4　Resnet18与VGG16模型在CIFAR10数据集上的鲁棒性(%)**

| | 无防御 (Resnet18) | 本文 (Resnet18) | 无防御 (VGG16) | 本文 (VGG16) |
|---|---|---|---|---|
| 干净样本($\varepsilon$=0) | 93.0 | 85.0 | 92.1 | 81.4 |
| FGSM($\varepsilon$=2/8/16) | 83.1/65.9/66.4 | 84.9/79.0/78.7 | 83.6/47.8/28.3 | 81.4/79.8/75.9 |
| PGD-40($\varepsilon$=2/8/16) | 79.1/51.5/45.2 | 84.9/78.7/77.6 | 81.3/24.3/11.8 | 81.4/79.7/74.6 |
| C&W($\varepsilon$=2/8/16) | 92.7/92.0/91.0 | 85.0/84.9/84.8 | 92.0/91.5/90.7 | 81.3/81.2/81.2 |
| DeepFool($\varepsilon$=2/8/16) | 78.3/41.9/16.5 | 83.5/78.5/71.5 | 78.6/31.8/5.1 | 79.2/73.5/67.0 |

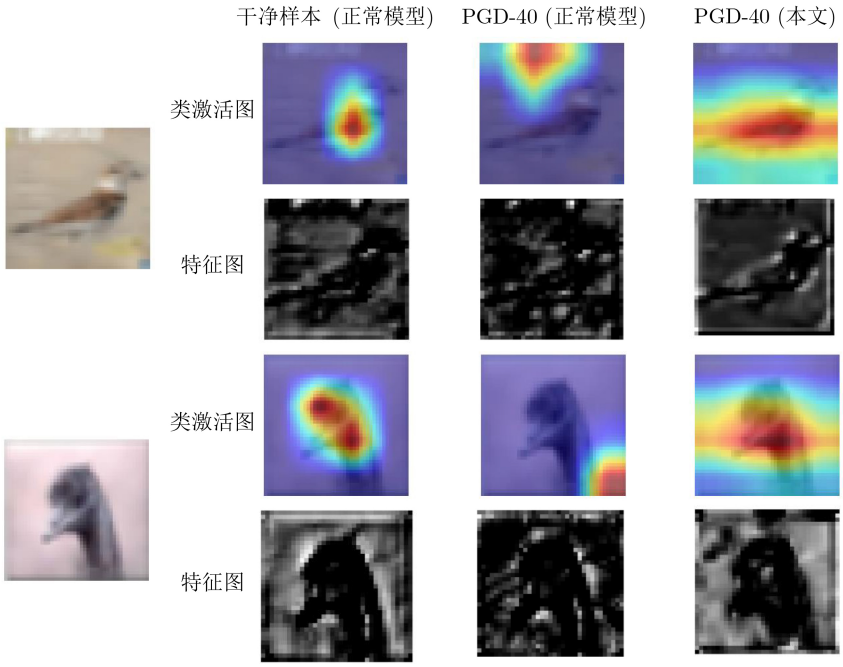


图 3 类激活图与特征图

模型进行训练。超参数设置如下：以0.1的学习率训练100代，0.01的学习率训练10代，0.001的学习率训练5代。

表5的实验结果表明，ResNet18模型使用本文提出的算法训练后，在CIFAR100数据集20分类任务上，以降低干净样本10.7%正确率的代价使模型对多数对抗攻击算法具有了鲁棒性，在扰动强度为0到16时，面对对抗攻击准确率的下降均小于16%，证明了本文提出的算法在更复杂的图像分类任务上仍能具有良好的鲁棒性。

图4展示了使用正常方法训练模型与使用本文提出算法进行训练时得到的模型在面对干净样本与FGSM(eps=8/255)对抗样本时的正确率，其中使用本文算法进行训练的正确率曲线中100代后的突变是学习率从0.1缩小至0.01导致。在图中可以看到，正常方法训练时，随着训练代数的增加，模型对干净样本的正确率有明显的上升，但是对FGSM对抗样本的正确率却上升不明显。使用本文提出算法进行训练时，随着训练代数的增加，模型对干净样本与FGSM对抗样本的正确率同步上升，证明了本文提出的算法能使模型具有良好的鲁棒性。

MNIST与Fashion-MNIST数据集由于图像构成简单，在此采用由两层CNN网络与两层线性网络，使用ReLU激活函数构建的模型进行训练与测试。超参数设置如下：均使用0.001的学习率训练60代。

表6与表7的实验结果表明，CNN网络使用本文提出的算法训练后，在MNIST数据集上，无论是干净样本还是对抗样本的正确率都高于或等于无防御的模型。在Fashion-MNIST数据集上，在干净样本的正确率下降了6%，但在对抗样本的正确率有了大幅的提升，展现出了与在CIFAR10数据集上类似的鲁棒性。证明了本文提出的算法对不同的数据集具有良好的泛化性能。

**表 5 ResNet18模型在CIFAR100数据集20分类任务上的鲁棒性(%)**

| 攻击算法 | 无防御 | 本文 |
|---|---|---|
| 干净样本($\varepsilon$=0) | 76.74 | 66.02 |
| FGSM($\varepsilon$=2/8/16) | 51.71/34.73/30.64 | 64.28/59.18/52.78 |
| PGD-20($\varepsilon$=2/8/16) | 46.10/14.34/5.25 | 64.26/58.96/51.91 |
| PGD-100($\varepsilon$=2/8/16) | 44.12/8.73/2.56 | 64.26/58.94/51.62 |
| C&W($\varepsilon$=2/8/16) | 49.64/16.55/3.66 | 64.05/58.22/50.48 |
| DeepFool($\varepsilon$=2/8/16) | 76.21/74.42/72.19 | 66.00/65.86/57.00 |

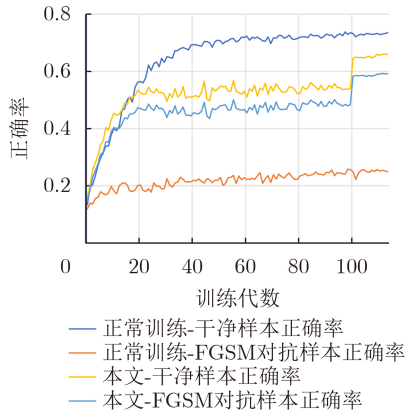


图 4 干净样本与对抗样本测试正确率对比图

**表 6　CNN网络在MNIST数据集上的鲁棒性(%)**

| 攻击算法 | 无防御 | 本文 |
|---|---|---|
| 干净样本($\varepsilon$=0) | 99.1 | 99.1 |
| FGSM($\varepsilon$=2/8/16) | 98.9/96.3/88.9 | 99.1/98.1/94.9 |
| PGD($\varepsilon$=2/8/16) | 98.8/90.8/67.0 | 99.1/97.8/91.4 |
| C&W($\varepsilon$=2/8/16) | 99.1/99.0/99.0 | 99.1/99.0/99.0 |
| DeepFool($\varepsilon$=2/8/16) | 98.4/93.4/64.2 | 98.8/97.5/93.7 |

**表 7　CNN网络在Fashion-MNIST数据集上的鲁棒性(%)**

| 攻击算法 | 无防御 | 本文 |
|---|---|---|
| 干净样本($\varepsilon$=0) | 93.47 | 87.41 |
| FGSM($\varepsilon$=2/8/16) | 80.13/48.09/35.17 | 86.18/82.74/78.40 |
| PGD-20($\varepsilon$=2/8/16) | 76.27/32.76/24.23 | 86.14/81.90/75.04 |
| PGD-100($\varepsilon$=2/8/16) | 75.41/29.11/23.88 | 86.14/81.78/74.10 |
| C&W($\varepsilon$=2/8/16) | 93.25/91.95/90.28 | 87.35/87.21/86.96 |
| DeepFool($\varepsilon$=2/8/16) | 77.67/25.64/0.36 | 86.09/82.26/76.69 |

## 5　结束语

对深度学习模型易受到对抗攻击干扰而产生高置信度误判的问题，本文将信息瓶颈引入了对抗训练中，使深度学习模型能够更好地学习输入数据与真实标签之间的关系，从而获得更好的鲁棒性。并且在多个数据集与深度学习模型中对算法进行了仿真，证明了算法的防御性能优于其他的防御算法，并且算法在不同条件下均展现了良好的防御性能，证明了算法具有良好的泛化性能。

## 参 考 文 献


[1] SZEGEDY C, ZAREMBA W, SUTSKEVER I, *et al*. Intriguing properties of neural networks[C]. The 2nd International Conference on Learning Representations (ICLR), Banff, Canada, 2014: 1–10.

[2] GOODFELLOW I J, SHLENS J, and SZEGEDY C. Explaining and harnessing adversarial examples[C]. The 3rd International Conference on Learning Representations (ICLR), San Diego, USA, 2015: 1–11.

[3] MADRY A, MAKELOV A, SCHMIDT L, *et al*. Towards deep learning models resistant to adversarial attacks[C]. 6th International Conference on Learning Representations (ICLR), Vancouver, Canada, 2018: 1–28.

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

[5] CARLINI N and WAGNER D. Towards evaluating the robustness of neural networks[C]. IEEE Symposium on Security and Privacy (SP), San Jose, USA, 2017: 39–57. doi: 10.1109/SP.2017.49.

[6] WONG E, RICE L, and KOLTER J Z. Fast is better than free: Revisiting adversarial training[C]. The 8th International Conference on Learning Representations (ICLR), Addis Ababa, Ethiopia, 2020: 1–17.

[7] ZHENG Haizhong, ZHANG Ziqi, GU Juncheng, *et al*. Efficient adversarial training with transferable adversarial examples[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, USA, 2020: 1178–1187. doi: 10.1109/CVPR42600.2020.00126.

[8] DONG Yinpeng, DENG Zhijie, PANG Tianyu, *et al*. Adversarial distributional training for robust deep learning[C]. The 34th International Conference on Neural Information Processing Systems (NeurIPS), Vancouver, Canada, 2020: 693.

[9] WANG Hongjun, LI Guanbin, LIU Xiaobai, *et al*. A Hamiltonian Monte Carlo method for probabilistic adversarial attack and learning[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022, 44(4): 1725–1737. doi: 10.1109/TPAMI.2020.3032061.

[10] CHEN Sizhe, HE Zhengbao, SUN Chengjin, *et al*. Universal adversarial attack on attention and the resulting dataset DAmageNet[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022, 44(4): 2188–2197. doi: 10.1109/TPAMI.2020.3033291.

[11] FAN Jiameng and LI Wenchao. Adversarial training and provable robustness: A tale of two objectives[C/OL]. The 35th AAAI Conference on Artificial Intelligence, 2021: 7367–7376. doi: 10.1609/aaai.v35i8.16904.

[12] GOKHALE T, ANIRUDH R, KAILKHURA B, *et al*. Attribute-guided adversarial training for robustness to natural perturbations[C/OL]. The 35th AAAI Conference on Artificial Intelligence, 2021: 7574–7582. doi: 10.1609/aaai.v35i9.16927..

[13] LI Xiaoyu, ZHU Qinsheng, HUANG Yiming, *et al*. Research on the freezing phenomenon of quantum correlation by machine learning[J]. *Computers, Materials & Continua*, 2020, 65(3): 2143–2151. doi: 10.32604/cmc.2020.010865.

[14] SALMAN H, SUN Mingjie, YANG G, *et al*. Denoised smoothing: A provable defense for pretrained classifiers[C]. The 34th International Conference on Neural Information Processing Systems (NeurIPS), Vancouver, Canada, 2020: 1841.

[15] SHAO Rui, PERERA P, YUEN P C, *et al*. Open-set adversarial defense with clean-adversarial mutual learning[J]. *International Journal of Computer Vision*, 2022, 130(4): 1070–1087. doi: 10.1007/s11263-022-01581-0.

[16] MUSTAFA A, KHAN S H, HAYAT M, *et al*. Image super-resolution as a defense against adversarial attacks[J]. *IEEE Transactions on Image Processing*, 2020, 29: 1711–1724.

doi: 10.1109/TIP.2019.2940533.

[17] GU Shuangchi, YI Ping, ZHU Ting, *et al*. Detecting adversarial examples in deep neural networks using normalizing filters[C]. The 11th International Conference on Agents and Artificial Intelligence (ICAART), Prague, Czech Republic, 2019: 164–173. doi: 10.5220/0007370301640173.

[18] TISHBY N, PEREIRA F C, and BIALEK W. The information bottleneck method[EB/OL]. https://arxiv.org/pdf/physics/0004057.pdf, 2000.

[19] TISHBY N and ZASLAVSKY N. Deep learning and the information bottleneck principle[C]. IEEE Information Theory Workshop (ITW), Jerusalem, Israel, 2015: 1–5. doi: 10.1109/ITW.2015.7133169.

[20] SHWARTZ-ZIV R and TISHBY N. Opening the black box of deep neural networks via information[EB/OL]. https://arXiv.org/abs/1703.00810, 2017.

[21] KOLCHINSKY A, TRACEY B D, and WOLPERT D H. Nonlinear information bottleneck[J]. *Entropy*, 2019, 21(12): 1181. doi: 10.3390/e21121181.

[22] ALEMI A A, FISCHER I, DILLON J V, *et al*. Deep variational information bottleneck[C]. The 5th International Conference on Learning Representations (ICLR), Toulon, France, 2017: 1–19.

[23] SHAMIR O, SABATO S, and TISHBY N. Learning and generalization with the information bottleneck[J]. *Theoretical Computer Science*, 2010, 411(29/30): 2696–2711. doi: 10.1016/j.tcs.2010.04.006.

[24] STILL S and BIALEK W. How many clusters? An information-theoretic perspective[J]. *Neural Computation*, 2004, 16(12): 2483–2506. doi: 10.1162/0899766042321751.

[25] KINGMA D P and BA J. Adam: A method for stochastic optimization[C]. 3rd International Conference on Learning Representations (ICLR), San Diego, USA, 2015: 1–15.

[26] ZHANG Hongyang, YU Yaodong, JIAO Jiantao, *et al*. Theoretically principled trade-off between robustness and accuracy[C]. The 36th International Conference on Machine Learning (ICML), Long Beach, USA, 2019: 7472–7482.

[27] ZHANG Haichao and WANG Jianyu. Defense against adversarial attacks using feature scattering-based adversarial training[C]. The 33rd International Conference on Neural Information Processing Systems (NeurIPS), Vancouver, Canada, 2019: 164.

[28] HE Kaiming, ZHANG Xiangyu, REN Shaoqing, *et al*. Deep residual learning for image recognition[C]. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, USA, 2016: 770–778. doi: 10.1109/CVPR.2016.90.

[29] LIU Shuying and DENG Weihong. Very deep convolutional neural network based image classification using small training sample size[C]. The 3rd IAPR Asian Conference on Pattern Recognition (ACPR), Kuala Lumpur, Malaysia, 2015: 730–734. doi: 10.1109/ACPR.2015.7486599.

董庆宽：男，硕士生导师，研究方向为网络与信息安全、深度学习与安全.

责任编辑：马秀强