第51卷第5期 南京航空航天大学学报 Vol. 51 No. 5
2019年10月 Journal of Nanjing University of Aeronautics & Astronautics Oct. 2019

DOI:10.16356/j.1005-2615.2019.05.011

对抗黑盒攻击的混合对抗性训练防御策略研究

陈 慧[1] 韩科技[2] 杭 杰[3] 李 云[2,3]

(1.南京邮电大学计算机科学与技术学院，南京,210023；2.江苏省大数据安全与智能处理重点实验室，南京，210023)

摘要：随着深度学习模型在无人驾驶等安全敏感性任务中的广泛应用,围绕深度模型展开的攻防逐渐成为机器学习研究的热点。黑盒攻击是一种典型的攻击场景,在攻击者不知道模型具体使用结构和参数等情况下仍能进行有效攻击,是现实场景中最常用的攻击方法。因此,分析深度学习模型的脆弱性并设计出更加鲁棒的模型来对抗黑盒攻击成为迫切需要。而传统基于单模型的单强度和多强度对抗性训练方法,在抵御黑盒攻击时性能十分有限;基于多模型的集成对抗性训练方法在抵御高强度、多样化攻击样本效果也不理想。本文提出一种基于贪婪强度搜索的混合对抗性训练方法,实验结果表明,所提出的混合对抗性训练能够有效抵御多样化的黑盒攻击,性能优于传统的集成对抗性训练。

关键词：深度学习；黑盒攻击；贪婪搜索；对抗性训练

中图分类号：TP181 **文献标志码：**A **文章编号：**1005-2615(2019)05-0660-09

Defense Strategy Against Black-Box Attacks with Mixed Adversarial Training

CHEN Hui[1], *HAN Keji*[2], *HANG Jie*[3], *Li Yun*[2,3]

(1. School of Computer Science and Technology, Nanjing University of Posts and Telecommunications, Nanjing, 210023, China; 2. Jiangsu Key Laboratory of Big Data Security and Intelligent Processing, Nanjing, 210023, China)

Abstract: Deep learning (DL) models have been widely applied to security-sensitivity tasks, such as auto-driving, etc. Attacks and defenses concerned with the DL have gradually become hot spots in the field of machine learning. The black box attack, as a typical attack type and the most common attack method in the real context, can still perform effective attacks without knowing the specific structure of the model and parameters. Therefore, a reasonable analysis of the vulnerability of the DL model and design of a more robust model against black-box attacks has become an emergent topic. Traditional single-strength and multi-strength adversarial training methods based on single-model are infeasible to resist black-box attacks. Ensemble adversarial training based on multi-model still fails to resist attack samples that are high-intensity and diversify. In order to solve this problem, the mixed adversarial training defense strategy based on greedy search strength is proposed. Experimental results show that the proposed defensive strategy has robustness faced with the diversified black box attacks, and superior performance compared to conventional adversarial training methods.

Key words: deep learning; black-box attack; greedy search; adversarial training

深度学习技术已广泛应用于人脸识别系统、无人驾驶工程和安防监控等安全敏感性任务中,攻击者会通过一定的手段分析目标模型的脆弱性,然后设计相应的攻击算法恶意篡改输入样本,从而降低

基金项目：国家自然科学基金(61772284,61603197,41571389)资助项目。

收稿日期：2018-07-10;**修订日期：**2018-10-29

通信作者：陈慧,女,硕士研究生,E-mail:980222712@qq.com。

引用格式：陈慧,韩科技,杭杰,等.对抗黑盒攻击的混合对抗性训练防御策略研究[J].南京航空航天大学学报,2019,51(5):660-668. CHEN Hui, HAN Keji, HANG Jie. Defense Strategy Against Black-Box Attacks with Mixed Adversarial Training[J]. Journal of Nanjing University of Aeronautics & Astronautics,2019,51(5):660-668.

目标模型的性能[1-3]。目前主流的攻击算法主要以梯度计算为主，例如，快速梯度符号法（Fast gradient sign method，FGSM）[4]通过计算目标模型损失函数的梯度并沿着梯度方向寻找对抗扰动，然后添加到原始正常输入样本中来构造攻击样本。I-FGSM[5]和MI-FGSM[6]则是对FGSM进行多步迭代获得更具攻击性的攻击样本。R+FGSM[7]在计算目标模型损失函数的梯度之前，先在原始输入样本中加入高斯噪声。Carlini攻击算法[8]通过采用梯度下降优化目标函数的方法来构造攻击样本。这些攻击算法的设计都需要攻击者对目标模型的体系结构或训练数据有充分的了解（白盒）。然而在现实世界中，由于很难获取到目标模型的内部信息，这对攻击者来说，目标模型就是一个黑盒。先前的研究表明，不同学习模型之间存在着迁移性[9-13]，也就是说，采用不同攻击算法生成的攻击样本能够使多个模型同时错误分类。这一属性为攻击者在未知目标模型内部信息情景下能够实现黑盒攻击奠定了基础。

针对攻击样本的存在，为了提高深度学习模型安全性及鲁棒性，Goodfellow等首次提出了对抗性训练[14]概念来减缓深度学习模型脆弱性问题。基于单模型的对抗性训练通过在原始训练集上注入单一强度或多个强度的攻击样本，在原始训练集和对抗训练集上训练神经网络，在增强了模型抵御攻击样本的能力的同时，还能够保持模型在正常样本上的分类准确率，但在抵御黑盒攻击时性能却十分有限。Papernot等进一步提出了集成对抗性训练方法[7]，通过预先训练多个模型生成单一强度的攻击样本并注入到原始训练集中进行对抗性训练。该方法尽管能够成功抵御黑盒攻击，但疏于考虑攻击样本的强度及差异性，导致在抵御高强度、多样化的攻击样本时，防御性能仍然失效。所以，本文提出了基于贪婪强度搜索的混合对抗性训练，不仅显著降低了攻击样本的迁移性，还提高了对不同攻击算法生成的攻击样本的抵抗能力。

1　相关工作

1.1　黑盒攻击过程

假设 $\boldsymbol{X}\in\mathbf{R}^{I*D}$ 表示 I 个有 D 维特征空间的样本组成的矩阵，$\boldsymbol{Y}=\{y^1,y^2,\cdots,y^c\}$ 表示 c 个不同标记的标记集合。给定合成数据集 $\mathrm{Data}_{\mathrm{synthetic}}=\{(x_i,y_i)\,|\,1\leqslant i\leqslant I\}$，$x_i\in\boldsymbol{X}$ 表示第 i 个训练样本，$y_i\in\boldsymbol{Y}$ 表示第 i 个样本查询目标模型的反馈标记。黑盒攻击的过程是从 $\mathrm{Data}_{\mathrm{synthetic}}$ 学习一个替代模型 $F(\boldsymbol{x})$，然后选用某种攻击算法为该模型生成攻击样本 $\boldsymbol{x}^{\mathrm{adv}}$，并将该样本迁移到目标模型 $O(x)$，使目标模型也错误分类，即 $O(\boldsymbol{x})\neq O(\boldsymbol{x}^{\mathrm{adv}})$。详细攻击流程如图1所示。

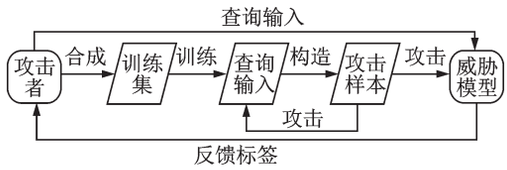

图1　黑盒攻击流程图

Fig. 1　Flow chart of black-box attack

本文着重探讨深度学习模型的脆弱性，那么替代模型也可选择深层模型，即 n 个带参函数 $f(\theta,\boldsymbol{x})$ 的分层组合来模拟高维输入 $\boldsymbol{x}$，定义为

$$F(\boldsymbol{x})=f_n(\theta_n,f_{n-1}(\theta_{n-1},\cdots,f_2(\theta_2,f_1(\theta_1,\boldsymbol{x}))))\tag{1}$$

1.2　攻击算法

当替代模型训练完成以后，攻击者会利用替代模型结构信息设计相应的攻击算法恶意篡改输入样本，从而紊乱目标模型以降低分类性能。目前主流的攻击算法包括两个方向：快速梯度符号方法和Carlini攻击算法。本文中将综合使用这两种方法生成的攻击样本来验证提出的防御策略的有效性。

快速梯度符号方法（Fast gradient sign method，FGSM）首次由Goodfellow等提出，通过计算目标模型损失函数的梯度，并沿着梯度方向最大化损失函数来寻找对抗扰动，然后添加到原始正常输入样本中，该方法定义为

$$\boldsymbol{x}^{\mathrm{adv}}=\boldsymbol{x}+\phi\cdot\mathrm{sign}(\nabla_x\mathrm{Loss}(F(\boldsymbol{x}),y_{\mathrm{true}}))\tag{2}$$

式中：ϕ 表示所添加对抗扰动的幅值，用于控制攻击样本的强度；y_{true} 表示样本 $\boldsymbol{x}$ 对应的真实标签，$\mathrm{Loss}(u,v)$ 表示交叉熵损失函数[15]，其定义为

$$\mathrm{Loss}(u,v)=-\sum u\cdot\log v\tag{3}$$

Carlini攻击算法是一种基于l-bfgs[10]攻击算法优化改进的迭代算法，从一定程度上提高了攻击样本的攻击能力。该方法通过使用辅助变量 ω 来寻找对抗扰动 r，即

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

式中 $\tanh(\cdot)$ 表示双曲正切函数，然后通过优化以下目标函数来获得辅助变量 ω，有

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

式中 $f(\cdot)$ 函数定义为

$$f(x)=\max(Z(\boldsymbol{x})_{y_{\mathrm{true}}}-\max\{Z(\boldsymbol{x})_i:i\neq y_{\mathrm{true}}\},-k)\tag{6}$$

式中：$Z(\boldsymbol{x})_i$ 表示神经网络softmax前一层的类别 i

的输出，k用于控制攻击类别标记与真实类别标记之间的置信度差值(即强度)，k值越大，攻击样本被错误分类的可能性越大。

1.3 对抗性训练防御策略

相关研究者指出对抗样本存在的主要原因是过度线性，而神经网络主要也是基于线性块构建的，整体函数也被证明了高度线性的本质[13]。当线性函数输入的维度较大时，即使每一个修改很微小，最终的线性函数值也会与之前相差甚远。如果用ε改变每个输入，那么权重为ω的线性函数可以改变$\varepsilon\|\omega\|_1$之多，若ω是高维的这会是更大的数。正因为这种线性本质的存在，才导致了攻击样本的可行性。而针对攻击样本的存在，对抗训练可以激励网络在训练数据附近的局部区域恒定，这样可以使得微小的变动对于结果的影响敏感度得到限制。

对抗性训练旨在构建攻击样本并注入到原始训练集中，在对抗扰动的训练集样本上训练网络。通过对抗训练可以减少在原有独立同分布的测试集上的错误率，同时提升了模型在抵御攻击样本时的鲁棒能力。考虑攻击样本的不同生成模式，对抗性训练又可划分为基于单模型的对抗性训练和多模型的对抗性训练。

1.3.1 单模型对抗性训练

单模型对抗性训练旨在利用模型自身去构建攻击样本，并注入到原始训练集中进行对抗性训练，其训练损失将被重新定义，除了包括对原始样本的预测值与真实值之间损失，还增添了对攻击样本的预测值与真实值之间的损失，具体定义为

$$\mathrm{Loss}=\mathrm{Loss}(F(\boldsymbol{x}),y_{\mathrm{true}})+\mathrm{Loss}(F(\boldsymbol{x}^{\mathrm{adv}}),y_{\mathrm{true}}) \tag{7}$$

其中$\mathrm{Loss}(u,v)$表示交叉熵损失，定义同式(3)。

对抗性训练很大程度上取决于攻击样本的强度选择，选择合适攻击样本强度进行对抗性训练将直接影响模型抵抗攻击样本的能力。因此，基于单模型对抗性训练又可划分为单强度和多强度对抗性训练。此时训练损失将修改为

$$\mathrm{Loss}=\mathrm{Loss}(F(\boldsymbol{x}),y_{\mathrm{true}})+\sum_{j=1}^{S}\mathrm{Loss}(F((\boldsymbol{x}^{\mathrm{adv}})_j),y_{\mathrm{true}}) \tag{8}$$

式中：S表示用于对抗性训练所选的攻击样本强度个数，当$S=1$即表示单强度对抗性训练，$S>1$即表示多强度对抗性训练。理想状态下，在给定有限强度范围内，所有强度的攻击样本都应该被用于对抗性训练，但这将带来过于耗时的问题。为了避免该问题，传统方法通过随机选择某些强度进行对抗性训练。

1.3.2 多模型集成对抗性训练

基于单模型的对抗性训练从一定程度上能够成功抵御白盒攻击，然而在抵御黑盒攻击时性能却十分有限。为了解决该问题，Papernot等在单模型对抗性训练基础上提出了多模型集成对抗性的概念[7]。该方法通过预先训练k个深度模型分离出攻击样本的生成过程，然后再注入到原始训练集中进行对抗性训练，从一定程度上有效减缓了攻击样本的迁移性，即能够成功抵御黑盒攻击。多模型集成对抗训练过程如图2所示。

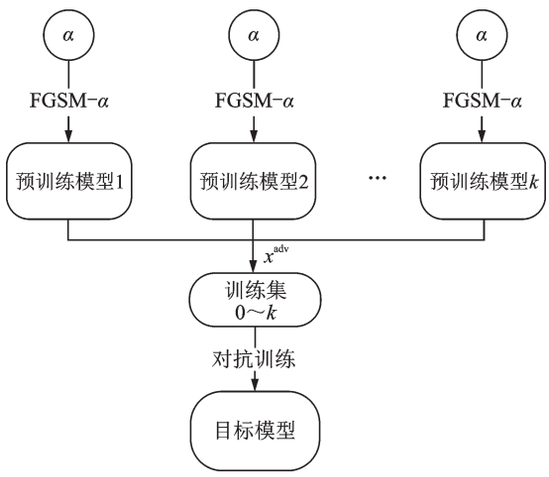

图2 多模型集成对抗训练过程

Fig. 2 Ensemble adversarial training process with multi-model

图2中参数α用于控制选用某种攻击算法生成的攻击样本的强度。假设以FGSM为例，同式(2)中的扰动幅值参数ϕ。Carlini攻击则表示置信度差值，同式(6)中的k值。本文主要以FGSM这类攻击算法为例实施对抗性训练，但对其他攻击算法仍然有效。

尽管集成对抗性训练能够有效减缓攻击样本的迁移性，但疏于考虑攻击样本的强度及差异性对对抗性训练的影响，导致在抵御高强度、多样化的攻击样本时，防御性能仍然失效。为了解决该问题，合理设计攻击样本强度搜索策略将有助于提高对抗性训练模型在抵御高强度攻击样本的鲁棒性。

2 基于贪婪搜索强度的混合对抗性训练

针对集成对抗性训练防御策略难以有效抵御高强度、多样化的攻击，为了提高模型在这种情况下的鲁棒性能，使得分类器在高强度、多样化的攻击样本前依然保证较好的抵御性能。本节将提出一种新颖的搜索强度的算法，主要使用了贪婪策略，借助这种强度搜索策略，使得在生成对抗样本

的时候，能够在不影响正常样本分类的基础上，产生尽可能攻击性强的样本。将这些样本注入到训练数据集中进行对抗性训练，从而提高模型在抵御高强度样本的鲁棒性。具体实现过程如图3所示。

该方法以多模型集成对抗性训练的思想为基础，在集成对抗性训练的前提下，需要预训练k个深度模型，利用这k个模型分离出攻击样本的生成过程，然后再注入到原始训练集中进行对抗性训练，保证这样训练得到的模型可以成功抵御黑盒攻击。

不同的是，传统集成对抗性训练方法仅生成了单一强度α的攻击样本进行对抗性训练，所以能达到的抵御性能比较有限。本文提出的贪婪搜索强度算法旨在搜索出最适合对抗性训练的攻击样本强度集合，然后从该强度集合中随机选择强度分配给每个预训练模型生成攻击样本，利用模型的迁移性，使用能够攻击预训练模型的对抗样本对目标模型进行对抗性训练。这种方法的优点是解决了传统集成对抗性训练在面对高强度攻击时的脆弱性，同时传统的集成对抗性训练仅能抵御单一的攻击方法，不能抵御多样化的攻击。虽然本文方法在训练的过程中使用的是FGSM攻击方法生成的对抗样本，但是在测试时却也能够较好抵御Carlini攻击方法生成的对抗样本。

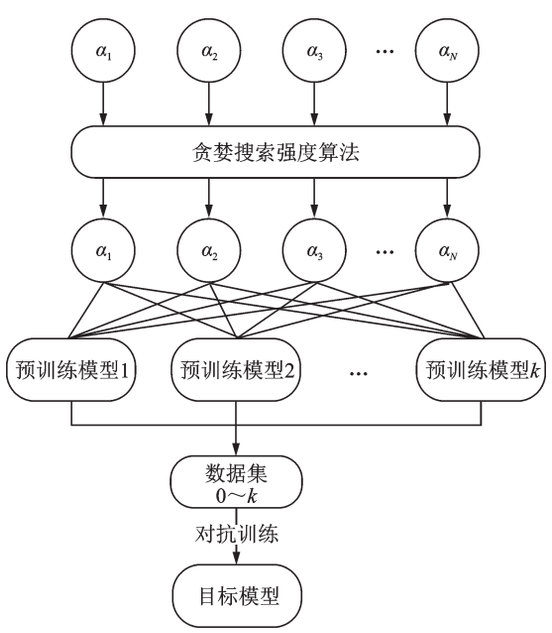

图3　基于贪婪搜索强度的混合对抗性训练过程

Fig. 3　Mixed adversarial training process based on greedy search strength

从图3可以看出，基于贪婪搜索强度的混合对抗性训练过程与多模型集成对抗训练过程的最大区别在于：在使用预训练模型生成对抗样本之前，先使用贪婪搜索强度算法从原始的N个强度中选取M个，其中$M<N$，这选取出来的M个强度能够代表很强的攻击能力，在不使用全部强度的前提下达到与使用全强度对抗性训练相差不大的抵御能力。

最重要的贪婪搜索强度策略如算法1所示。首先，需要限定用于对抗性训练的初始攻击样本强度范围，$\alpha_1,\alpha_2,\cdots,\alpha_N$，然后使用这$N$个强度中的每个强度上生成攻击样本进行基于单模型的单强度对抗性训练，经过这一步，将得到N个模型$F_1,F_2,\cdots,F_N$，这N个模型分别是使用从1到N的强度生成的对抗样本进行训练的。用验证集生成攻击样本去分别验证N个模型抵御$\alpha^1,\alpha^2,\cdots,\alpha^L$强度下的攻击样本的精度值。当$N$个模型训练完成以后，将得到一个$N\times L$的精度矩阵，然后从该精度矩阵每一列中选择最大的精度值对应的强度$\alpha_j^{\max}$并去重，添加到预先定义的强度集合中，也就是对于每一个验证攻击强度，为其选择具有最强抵御能力的模型，再将此模型对应的训练强度保留下来，供后续的多模型混合对抗性训练从该集合中选取所需的攻击样本强度。

算法1　贪婪搜索强度算法

符号说明：V 验证集

FGSM-α 将原样本变为对抗性样本

输入：$\alpha_1,\alpha_2,\cdots,\alpha_N$ 对应的单强度对抗性训练模型$F_1,F_2,\cdots,F_N$

$\alpha_1,\alpha_2,\cdots,\alpha_L$ 攻击强度

输出：选择的强度集合 set

(1) 强度集合 set=[ ]

(2) for i in 1 to L:

(3)　map V to V' with FGSM$-\alpha_i$

(4)　for j in 1 to N:

(5)　　计算F_j抵御攻击强度α_i的攻击样本的精度值

(6)　　选择最大的精度值对应的强度$\alpha_{j\max}$

(7)　add $\alpha_{j\max}$ to 强度集合，并去重

(8) end for

(9) 返回强度集合 set

算法1主要是描述了如何选取适合的强度，贪婪搜索强度算法只是整个混合对抗性训练算法中的一个步骤，也是最为重要的一步。现在将贪婪搜索强度算法与传统的集成对抗性训练结合起来，得到混合对抗性训练算法。

算法2　混合对抗性训练算法

输入：训练数据D，训练样本数目N，学习率η，攻击方法G，预训练模型集合F，目标模型O，预训练模型数目k，对抗样本强度集合set，对抗强度集

合中元素个数s。

输出:经过训练得到的模型参数θ

(1) 对目标模型参数θ进行初始化

(2) for i in 1 to N:

(3) //从训练集中每次采样一个样本

(4) $(x_i,y_i)\sim D$

(5) for j in 1 to k:

(6) for m in 1 to s:

(7) Select ε from set randomly

(8) //为每个样本构造迁移对抗样本

(9) $\boldsymbol{x}_{jm}^{*}=G(F_j(\boldsymbol{x}_i),y_i,\varepsilon)$

(10) end for

(11) //定义对抗目标函数

(12) $L=L(O(\boldsymbol{x}_i),y_i)+\sum_{j=1}^{k}\sum_{m=1}^{s}L(F_j(\boldsymbol{x}_{jm}^{*}),y_i)$

(13) //更新网络权重

(14) $\theta=\theta-\eta\nabla_{\theta}(L)$

(15) end for

(16) return θ

混合对抗性训练算法描述了使用本文算法进行训练模型的整个过程。首先,先对目标模型的参数进行初始化,然后对训练集中的每一个样本构造对抗样本,每一个对抗样本的生成都要使用k个模型,同时在每一步操作过程中都要随机从强度集合中选择一个强度,这两个步骤具体在上述算法中的内外循环中体现。

对抗样本生成结束后,要对传统的损失函数定义进行修改,在正常样本的损失函数基础上,将对抗样本的损失代价也加入进去,同时考虑到多个强度。然后,通过梯度下降法对模型参数进行更新,更新的原则是使得刚才定义的对抗目标函数的值能够越来越小,从而选择最合适的模型参数。

3 实 验

为了验证上述防御策略的有效性,在现实数据集上比较了基于贪婪搜索强度的混合对抗性训练防御策略和其他对抗性训练方法在抵御黑盒攻击的性能。所有实验在Pytorch框架实现,使用NVIDIA GTX660 GPU, 12 GB内存。

3.1 数据集

为了验证基于贪婪搜索强度的混合对抗性训练防御策略的有效性,分别在MNIST手写数字体识别数据集、Fashion MNIST商品分类数据集和GTSRB交通标志识别数据集上做了实验。其中,MNIST是对模型的防御能力进行衡量的基准数据集,之前所有的有关对抗性训练的研究都是以此数据集为基础进行的。同时还考虑了一个在现实场景中非常具有安全研究意义的数据集GTSRB,通过对交通标志识别任务的分类鲁棒性进行衡量来验证本文算法的现实使用意义。

为了更好地衡量算法的性能,还在Fashion-MNIST数据集上作了补充实验。Fashion-MNIST是一个替代MNIST手写数字集的图像数据集,包括了10种类别的共7万个不同商品的正面图片,类别不再是简单的0~9的手写数字体,而是T恤、裤子、裙子、汗衫、运动鞋以及外套等日常商品类别。Fashion-MNIST(F-MNIST)的大小格式等与原始的MNIST完全一致,在没有增加训练的难度的基础上,同时又保证了数据集的复杂度,能够代表现代机器学习任务。表1给出了实验数据集的详细信息。

表1 实验数据信息

Tab. 1 Experimental data information

数据集	样本个数	特征数	类标数	任务
MINST	60 000	784	10	手写体数字识别
F-MNIST	60 000	784	10	商品类别分类
GTSRB	51 839	3 072	43	交通标志识别

3.2 实验设置与结果分析

3.2.1 实验设置

对于MNIST,Fashion-MNIST和GTSRB数据集,统一按照50:9:1的比例将原始数据划分为训练数据集、测试数据集和验证数据集。在MNIST,Fashion-MNIST和GTSRB数据集上,分别预训练$k(k=4)$个模型用于生成攻击样本,并注入到训练集中进行混合对抗性训练得到目标模型作为黑盒模型。在对目标模型进行训练时,损失函数使用负数似然损失函数,优化方法选取的是Adam优化器,学习率设置为0.01。

通过混合对抗性训练得到的模型就是本文的防御模型,实验的主要目的就是通过对传统的模型和本文的算法训练得到的防御模型分别进行攻击,通过分析模型的准确率来验证算法的有效性。为了更加准确地衡量混合对抗性训练模型的防御能力,主要考虑在黑盒的场景中实施攻击,也就是把用于对比试验的每一个目标模型都看作黑盒模型,在目标模型训练完成后,给定输入,通过查询目标模型得到反馈,通过查询的输入和反馈的标签构造合成数据集,用合成数据集来对替代模型进行训练,通过使用替代模型实现黑盒攻击并验证提出的方法抵御黑盒攻击的性能。

3.2.2 实验过程及分析

在MNIST数据集上,首先分析了传统集成对

抗训练方法在抵御不同强度的攻击样本的防御性能，通过这样的实验分析来说明传统集成对抗性训练方法的不足，实验结果如图 4 所示。实验结果表明源模型，即不进行对抗性训练的情况，抵御攻击样本的能力最弱，说明深度学习模型的脆弱性的确存在。

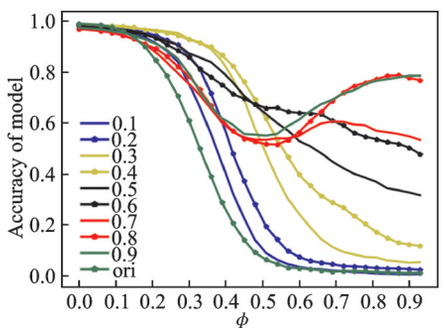

图 4　MNIST 上集成对抗性训练模型鲁棒性

Fig. 4　Robustness of a ensemble adversarial training on MNIST

然后，选用不同强度的攻击样本进行集成对抗性训练，通过实验结果可以发现，选用不同单一强度的攻击样本进行集成对抗性训练，仅在特定攻击强度区域内获得好的防御效果，却无法同时抵御多种范围的攻击样本。

使用强度为 0.1 进行单强度对抗性训练，仅在抵御 0.1～0.3 范围内的攻击时效果较好，在面对大于 0.3 的强度攻击时，模型的准确度就表现得很差，下降趋势仅优于原始模型，尤其是在面对大于 0.6 的强度攻击时，模型效果会快速下降。在使用强度 0.5 进行训练时，可以看到在面对小于 0.5 的攻击时，模型的准确度还能够保持在 70% 以上，对比使用强度为 0.1 进行训练的模型准确度，可以看到使用强度为 0.5 能够将模型的有效抵御攻击的强度范围增大。不过，对于大于 0.6 的攻击强度，模型的准确度还是会急速下降。

针对上述现象，考虑可以尽量选取较大的强度进行对抗性训练。但是，从实验结果来看，如果单纯地使用较大的强度进行对抗性训练，取得的效果也是有利有弊。在选取 0.7，0.8 和 0.9 强度的对抗样本进行对抗性训练时，模型在面对 0.6 到 0.9 的强度攻击时，准确度能够较好的保持，甚至于能够接近 80%，表现明显优于之前的使用其他强度的对抗性训练。尽管使用高强度进行对抗性训练能够抵御较强的攻击，但是，在面对较低强度的攻击时，模型效果却又表现得不如使用中等强度。

在 GTSRB 数据集上，同样可以发现不采取对抗性训练的模型鲁棒性是最差的。如图 5 所示，对于使用不同的单一强度进行训练，所得到的模型的抵御能力是有一定的作用区域的，在 GTSRB 数据集上，这种作用区域的差别看上去可能并不明显，这是由于 GTSRB 这个数据集相对而言比较复杂，所以在面对较强攻击时的模型准确度都会比较低，看上去都是一条直线。但是对于使用单一强度所产生的弊端与 MNIST 数据集上一致。

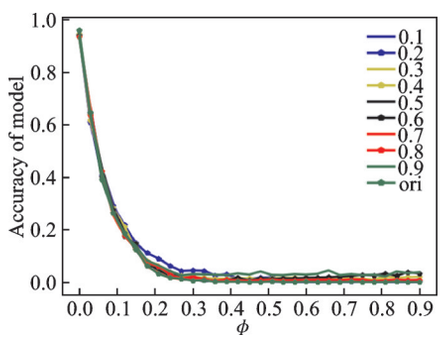

图 5　GTSRB 上集成对抗性训练模型的鲁棒性

Fig. 5　Robustness of a ensemble training model on GTSRB

在 FASHION-MNIST 数据集上，也进行了同样的分析。如图 6 所示，在面对高强度的攻击时，明显在对抗性训练时选取较大的强度能够获得更好的效果。每一个单一强度进行的防御都是有一定的作用范围的，超过这个范围之内，模型的性能就会急剧下降。

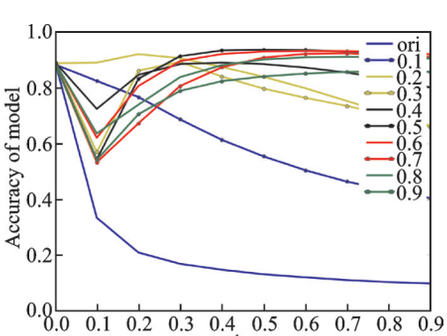

图 6　FASHION-MNIST 上集成对抗性训练模型的鲁棒性

Fig. 6　Robustness of a ensemble training model on FASHION-MNIST

为了解决这个问题，很直观的考虑是否可以在进行对抗性训练时考虑使用全部强度或随机选择强度。对于这两种方法，都在 3 个数据集上进行了对比实验，分别与传统的对抗性训练、传统的集成对抗性训练以及使用单模型的多强度方法进行对比分析。不同方法比较的具体结果如图 7—9 所

示。m_all即表示使用多模型,选择所有的强度;m_random表示使用多模型,但是随机选择强度;sm表示使用单模型多强度进行对抗性训练;ss表示使用单模型单强度进行对抗性训练,即为传统的对抗性训练;ms表示使用多模型单强度进行训练,即为传统的集成对抗性训练;ori表示不使用对抗性训练。

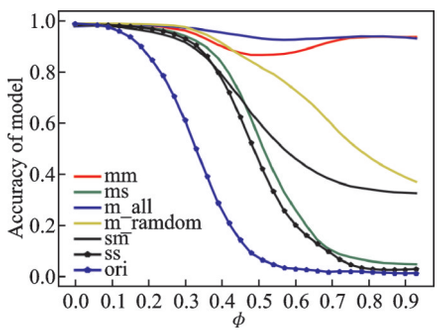

图7 MNIST上不同的对抗强度选择方法对比

Fig. 7 Comparison of different methods for selection of adversarial strength on MNIST

随机选择强度的随机性对实验效果的影响比较大,在MNIST和GTSRB上随机选择强度的方法表现得还可以,但是在FASHION-MNIST上的效果就和不使用对抗性训练的模型效果相差无几。

用全强度进行对抗性训练存在两个问题:(1)整个实验耗费时间比较长;(2)在测试集上的效果并没有很好,这可能是因为加入过多攻击强度的样本导致模型过拟合,而且考虑到的强度过多导致模型在正常样本上的分类效果变差。

在图7和图8中,ms代表的传统集成对抗性训练的效果始终不如使用多强度的m_all以及m_random,说明尽管随机选择强度和使用全部强度有上述缺点,但是使用多模型和多强度相结合的方法效果明显优于传统的多模型单强度对抗性训练。

基于此,本文提出了贪婪搜索强度的算法来进行攻击强度搜索,并将本文的算法与上述方法在3个数据集上进行了比较,mm代表的就是贪婪搜索强度算法对应的多模型多强度对抗性训练。首先分析在MNIST数据集上的效果,从图7可以看到,混合对抗性训练的效果仅次于使用全部强度,并且在面对高强度时,训练达到的效果与使用全部强度几乎持平。在图8所示的GTSRB数据集上,混合对抗性训练在面对攻击时的性能远远优于随机选择强度的多模型多强度方法以及传统的集成对抗

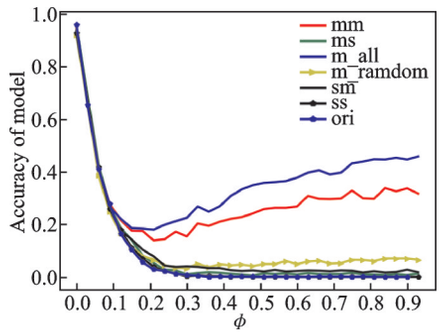

图8 GTSRB上不同的对抗强度选择方法对比

Fig. 8 Comparison of defense performance of different methods for selection of adversarial strength on GTSRB

性训练等,说明搜索出来的强度能够极大地提升模型抵御攻击的能力。在图9所示的FASHION-MNIST上有同样的实验效果,并且,在这种稍微复杂的数据集上,本文的方法的优越性能够表现得更加突出。

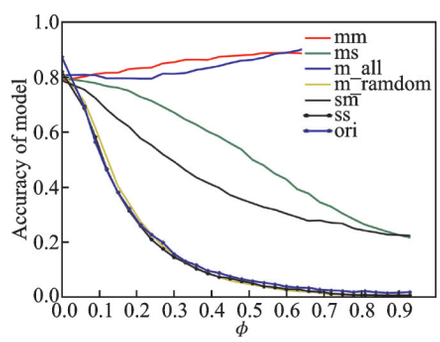

图9 FASHION-MNIST上不同的对抗强度选择方法对比

Fig. 9 Comparison of defense performance of different methods for selection of adversarial strength on FASHION-MNIST

综合而言,基于贪婪搜索强度算法既避免了使用全部强度带来的弊端,在降低模型训练时间的同时能更好地将不同的强度的工作区域结合起来,具有很好的鲁棒性。同时,在攻击强度较大时表现明显优于除了全强度以外的其余方法,鲁棒性能与使用全强度的性能最为接近。

3.2.3 多样化攻击及模型精度

本文的模型在训练过程中使用的攻击方法是FGSM算法,现在使用Carlini方法来对模型进行攻击。在MNIST数据集上结果如表2所示,表中数据为每个方法面对Carlini攻击时的分类准确性。从表2可以看出,使用贪婪搜索多强度进行对抗性训练的模型能够抵御Carlini攻击,说明本文的模型能够抵御多样化的攻击。

表 2　MNIST 上各模型抵御 Carlini 攻击的性能

Tab. 2　Accuracy of the model under Carlini attack on MNIST

mm	ms	m_all	m_random
99.2	98.8	97.0	98.6

本文还对以上几种方法训练的模型在没有攻击情况下的性能进行比较，结果如表 3 所示，性能差别不大，这证明本文的模型能够在保证分类精度的同时提高了鲁棒性。

表 3　MNIST 上强度选择方法对应模型精度

Tab. 3　Accuracy of the model corresponding to the intensity selection method on MNIST

mm	ms	m_all	m_random
98.43	98.61	97.64	98.79

3.2.4　算法性能分析

为了对混合对抗性训练的算法复杂度进行分析，首先对传统的对抗性训练的算法复杂度进行分析，相比较而言，本文的方法并没有增加算法复杂度。

首先看一下原始的集成对抗性训练，集成对抗性训练主要是为了改进传统对抗性训练不能很好的抵御黑盒攻击这一缺点，其主要做法是预训练多个模型并在此基础上生成具有迁移性的样本，然后将这些具有迁移性的对抗样本注入到训练样本中进行训练。集成对抗性训练与传统对抗性训练相比，增加了预训练替代模型的过程，并增加了由此产生的时间与空间上的花销，这个部分的时空开销是所有基于集成对抗性训练方法都会产生的。

与集成对抗性训练相比，混合对抗性训练在预训练多个替代模型的基础上，仅仅增加了通过贪婪策略选择多个强度用来生成迁移对抗样本的过程。在进行强度选择的时候，需要预先训练 N 个模型，再针对这 N 个模型分别使用 L 个强度进行攻击。选择强度算法的时间复杂度是 $O(N\times L)$，在实际实验过程中，这个 N 和 L 都是极小的整数，并且整个操作都是通过矩阵运算实现，能够并行操作，整体对于算法没有增加太多的复杂度。

4　结　　论

本文旨在探索将多模型与多强度进行结合，并运用到对抗性训练中以提高模型抵御高强度，多样化攻击样本的能力。首先，通过设计一种贪婪搜索强度算法找到最利于对抗性训练的一组攻击样本的强度集合；然后随机选出模型，在模型上运用搜索到的攻击强度来生成对抗样本。通过在 MNIST，GTSRB 和 FASHION-MNIST 数据集上进行实验，验证了所提想法的性能，并证明了所提出的方法在保持精度的同时，鲁棒性优于之前的对抗性训练模型。

参考文献：

[1] SHARIF M, BHAGAVATULA S, BAUER L, et al. Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition[C]//2016 ACM SIGSAC Conference on Computer and Communications Security (CCS). New York: ACM, 2016: 1528-1540.

[2] BARRENO M, NELSON B, JOSEPH A D, et al. The security of machine learning[J]. Machine Learning, 2010, 81(2):121-148.

[3] NICOLAS P, PATRICK M, SOMESH J, et al. The limitations of deep learning in adversarial settings [C]//2016 IEEE European Symposium on Security and Privacy (EuroS&P). [S. l.]: IEEE, 2016: 372-387.

[4] SHAHAM U, YAMADA Y, NEGAHBAN S.Understanding adversarial training: Increasing local stability of neural nets through robust optimization[J]. Neurocomputing, 2018,307: 195-204.

[5] ALEXEY KURAKIN, IAN J. Goodfellow, Samy Bengio: Adversarial examples in the physical world [C]//International Conference on Learning Representations(ICLR).[S.l.]:[s.n.],2017.

[6] DONG Y, LIAO F, PANG T, et al. Boosting adversarial attacks with momentum[C]//2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018. [S.l.]:IEEE,2018: 9185-9193.

[7] TRAMER F, KURRAKIN A, PAPERNOT N, et al.Ensemble adversarial training: Attacks and defenses [C]//The 6th International Conference on Learning Repre-sentations (ICLR). [S.l.]:[s.n.], 2018.

[8] CARLINI N, WAGNER D. Towards evaluating the robustness of neural networks [C]//Security and Privacy.[S.l.]: IEEE, 2017: 39-57.

[9] HANG J, HAN K J, LI Y.Delving into diversity in substitute ensembles and transfer-ability of adversarial examples[C]//The 25th International Conference on Neural Information Processing(ICONIP). Heidelberg: Spring, 2018: 175-187.

[10] SZEGEDY C,ZAREMBA W,SUTSKEVER I,et al. Intriguing properties of neural networks [C]//International Conference on Learning Representations (ICLR). [S.l.]:[s.n.],2014.

[11] PAPERNOT N, MCDANIEL P, GOODFELLOW I. Practical black-box attacks against machine learning [C]//Proceedings of the 2017 ACM on Asia Conference on Computer and Com-munications Security, (ASIA CCS). [S.l.]:ACM,2017: 506-519.

[12] LIU Y, CHEN X, LIU C, et al. Delving into transferable adversarial examples and black box attacks [C]// 5th International Conference on Learning Representations (ICLR). [S.l.]:[s.n.],2017.

[13] GOODFELLOW I J, SHLENS J, SZEGEDY C. Explaining and harnessing adversarial examples [C]// International Conference on Learning Representations (ICLR). [S.l.]:[s.n.],2015.

[14] KURAKIN A, GOODFELLOW I, BENGIO S. Adversarial machine learning at scale [C]// 5th International Conference on Learning Representations (ICLR). [S.l.]:[s.n.],2017.

[15] BOER P T D, KROESE D P, MANNOR S, et al. A tutorial on the cross-entropy method [J]. Annals of Operations Research, 2005, 134(1): 19-67.

(编辑:刘彦东)