

基于卷积神经网络隐空间的虚拟对抗学习

邵琦琦,钱亚冠,王佳敏,李思敏,梁小玉

(浙江科技学院 理学院,杭州 310023)

摘要: 对抗训练存在计算效率低的缺点,对此提出一种虚拟对抗学习的方法。在 CIFAR-10 和 ImageNet(30) 数据集上验证本方法,首先,建立阈值机制来挑选对抗源样本;然后,在对抗源样本的 logits 上添加扰动生成虚拟对抗样本,而非对抗源样本保持不变;最后,计算虚拟对抗样本和非对抗源样本的损失,通过反向传播更新网络权重。试验结果表明,与传统的对抗训练相比,本文方法在干净样本的测试精度上提升了大约 7~14 百分点,在扰动样本的测试精度上不亚于投影梯度下降(projected gradient descent, PGD)对抗训练的效果,尤其是在 ImageNet(30) 数据集上提升了 4.62 百分点。在训练效率上,与最慢的 PGD 对抗训练相比,本文方法的训练时间缩短了 2/3 左右。这些结果均证明了虚拟对抗学习既能提升对干净样本的预测精度,又能提高模型的鲁棒性;同时加快对抗训练过程,为对抗训练在工业环境的运用提供有效方法。

关键词: 训练效率;对抗训练;虚拟对抗学习;虚拟对抗样本

中图分类号: TP389.1

文献标志码: A

文章编号: 1671-8798(2022)05-0426-09

Virtual adversarial learning based on latent space of convolutional neural network

SHAO Qiqi, QIAN Yaguan, WANG Jiamin, LI Simin, LIANG Xiaoyu

(School of Sciences, Zhejiang University of Science and Technology, Hangzhou 310023, Zhejiang, China)

Abstract: In response to the disadvantage of low computational efficiency of adversarial training, a virtual adversarial learning method was proposed to perform and verify on the CIFAR-10 and ImageNet(30) data sets. Firstly, a threshold mechanism was established to select adversarial source examples. Secondly, perturbations were added to the logits of adversarial source examples to generate virtual adversarial examples, while non-adversarial source examples remained unchanged. Finally, the losses of virtual adversarial examples and non-adversarial source examples were calculated, and the network weights were updated by

收稿日期: 2021-04-22

基金项目: 浙江省自然科学基金项目(LY17F020011)

通信作者: 钱亚冠(1976—),男,浙江省嵊州人,教授,博士,主要从事机器学习与大数据处理、对抗机器学习研究。

E-mail: qianyaguan@zust.edu.cn.

back propagation. Experimental results show that compared with traditional adversarial training, this method improves the test accuracy of clean examples by about 7% to 14%, and is not inferior to PGD (projected gradient descent) adversarial training in the test accuracy of perturbed examples, especially on ImageNet(30) data set, which is increased by 4.62%. In terms of training efficiency, the training time of this method is shortened by about 2/3 compared with the slowest PGD adversarial training. These results demonstrate that virtual adversarial learning can not only improve the prediction accuracy of clean examples, but also enhance robustness of the model, and at the same time speed up the adversarial training process, which provides an effective reference for the landing of adversarial training in industrial environment.

Keywords: training efficiency; adversarial training; virtual adversarial learning; virtual adversarial example

深度学习在图像识别^[1]、语音识别^[2]、自然语言处理^[3]等领域取得了成功的应用,但是近几年的研究发现,深度神经网络(deep neural network,DNN)很容易受到对抗样本的攻击。对抗样本就是在干净样本上添加微小的扰动,从而导致DNN分类错误,例如在“停止”的交通标志上添加微小的扰动,就能使自动驾驶汽车辨别为“加速”等其他交通标志,从而造成重大交通事故。自Szegedy等^[4]发现对抗样本的存在后,后续研究者提出了很多对抗样本的生成方法,如快速梯度符号法(fast gradient sign method, FGSM)^[5]、聚焦图像的无目标攻击^[6]、基于生成对抗网络合成对抗样本^[7]、补丁攻击^[8]等,对对抗样本产生的原因也进行了探索^[4-5],但由于DNN难以解释的特性,目前仍然没有完全了解对抗样本的产生机理。与此同时,相应的防御方法也被广泛研究。许笑等^[9]提出了冗余信息压缩方案以消除附加扰动,有效地防御对抗攻击。范宇豪等^[10]根据插值算法能够在图像缩小和放大过程中较好地保留图片信息这一特性,提出基于插值法的对抗防御算法。目前对抗训练被认为是极有效的防御方法,它利用对抗样本来训练模型,不同的对抗样本生成方法构成不同的对抗训练方法。

FGSM对抗训练是对抗训练的早期形式,随后更为有效的对抗训练一直在探究中。Madry等^[11]提出投影梯度下降(projected gradient descent,PGD)对抗训练并将其表示为最小最大的鞍点优化问题,其内部最大问题是寻找攻击最强的对抗样本,外部最小问题是在最强对抗样本的干扰下,获得损失最小的DNN。该方法至今仍然保持着较好鲁棒性。现有研究表明,对抗训练的计算复杂度非常高,其中生成对抗样本占了总耗时的主要部分^[12]。Zhang等^[13]认为,当执行多步PGD攻击时,在反向传播计算对抗样本期间可以减少冗余的计算来获得额外的加速。Shafahi等^[12]提出了自由对抗训练(简称Free),通过使用单次反向传播同时更新模型权重和输入扰动,以提高对抗训练的效率。在此基础上,Zhu等^[14]提出大批次自由对抗训练(简称FreeLB),与Free不同的是,FreeLB在内部损失最大化的 k 步过程中没有更新模型参数,而是利用 k 步之后积累的梯度求平均,再对模型参数进行更新,该方法进一步增强了模型的鲁棒性。Wong等^[15]提出了Fast(快速)对抗训练,利用先前的小批量扰动或从均匀随机扰动中初始化一个扰动添加到干净样本上,并使用比扰动约束更大的步长,加上循环学习率和混合精度计算等技术,使得Fast对抗训练在增强模型鲁棒性的同时实现加速。

上述对抗训练方法需要在输入空间生成真实的对抗样本,再用对抗样本进行训练。然而,生成对抗样本的过程需要更新对抗扰动,这需要通过反向传播计算损失对样本的梯度来实现,因此占用了大量训练时间。此外,对抗样本训练会影响模型对干净样本的预测精度。针对这些问题,我们提出一种虚拟对抗学习方法,也称之为虚拟对抗训练,将寻找输入空间中的扰动来生成对抗样本的问题,转化为寻找隐空间中的扰动来生成虚拟对抗样本的问题,从而有效地避免计算损失对样本的梯度,大幅提高对抗训练的速度,同时引入阈值机制来保证训练后的模型对干净样本的预测精度影响较小。

1 基本符号和定义

在本研究中,定义训练数据 $D = \{(\mathbf{x}_i, y_i)\}_{i=1}^N, \mathbf{x}_i \in \mathbb{R}^d, y_i \in \{1, \dots, C\}, y_i$ 为 \mathbf{x}_i 的类标签。 $F_\theta(\mathbf{x})$ 表示一个参数为 θ 的预训练神经网络模型, $F_\theta(\mathbf{x}) = F^{(N)}(\dots(F^{(2)}(F^{(1)}(\mathbf{x}))))$ 。分类器 $F_\theta(\mathbf{x})$ 的最后一层为 softmax 层,输出概率向量 $\mathbf{P} = (p_1, p_2, \dots, p_C)$,样本被判别为第 j 类的置信度 p_j 可表示为

$$p_j = \text{softmax}(\mathbf{z})_j = \frac{e^{z_j}}{\sum_{k=1}^C e^{z_k}}, j = 1, 2, \dots, C. \quad (1)$$

式(1)中:向量 $\mathbf{z} = (z_1, z_2, \dots, z_C)$ 为分类器 $F_\theta(\mathbf{x})$ 的倒数第二层的输出 logits。记 $\hat{y} = \arg \max_j \{p_j\}$ 是样本 \mathbf{x} 的预测分类标签。

定义 1 对抗样本:对于干净样本 \mathbf{x} ,它的正确类标签为 y 。如果存在扰动 $\delta, \|\delta\|_2 < \epsilon, \epsilon$ 为扰动大小约束,使得 $\mathbf{x}' = \mathbf{x} + \delta$ 满足 $\arg \max_j F_\theta(\mathbf{x}') \neq y$,那么称 \mathbf{x}' 为对抗样本。

由于对抗扰动 δ 受到约束,并不是任何一个干净样本都能获得对应的对抗样本。为了更好地阐述相关方法,定义了对抗源样本的概念。

定义 2 对抗源样本:对于干净样本 \mathbf{x} ,它的正确类标签为 y 。假设 G 为某个对抗样本生成算法,满足 $F_\theta(G(\mathbf{x})) \neq y$,那么该样本称为当前模型 F_θ 下 G 算法的对抗源样本。

定义 3 虚拟对抗样本:干净样本 \mathbf{x} 通过网络倒数第二层的输出 logits 为 \mathbf{z} ,在 logits 上添加扰动 Δ ,即 $\mathbf{z}' = \mathbf{z} + \Delta$,如果在输入空间存在一个对应的对抗样本 \mathbf{x}' ,满足 $F_\theta(\mathbf{x}') = \text{softmax}(\mathbf{z}')$,那么把这样的 \mathbf{x}' 称为虚拟对抗样本。

2 虚拟对抗学习的实现

2.1 问题的描述

传统的对抗训练需要在输入空间寻找一个最小扰动 δ 生成对抗样本 $\mathbf{x}' = \mathbf{x} + \delta$,再用对抗样本来训练分类器,而本研究在特征空间寻找一个最小的扰动 Δ 生成虚拟对抗样本进行训练。假定干净样本 \mathbf{x} 的真实标签 y 与预测标签 \hat{y} 一致, $y = \hat{y}$,虚拟对抗样本 \mathbf{x}' 的 softmax 输出为 $\mathbf{P}' = (p'_1, p'_2, \dots, p'_C)$,样本被判别为第 j 类的置信度 p'_j 可表示为

$$p'_j = \text{softmax}(\mathbf{z} + \Delta)_j = \frac{e^{z_j + \Delta_j}}{\sum_{k=1}^C e^{z_k + \Delta_k}}, j = 1, 2, \dots, C. \quad (2)$$

式(2)中:扰动 $\Delta = (\Delta_1, \Delta_2, \dots, \Delta_C)$,为与 \mathbf{z} 同维度的向量。记 $y' = \arg \max_j \{p'_j\}$ 为虚拟对抗样本的输出标签。由对抗样本的定义可知: $y' \neq y \cap y' \neq \hat{y}$,即有 $y \neq \arg \max_j \{p'_j\}$ 。因此, $\arg \max_j F_\theta(\mathbf{x}') \neq y$ 等价于 $\text{softmax}(\mathbf{z} + \Delta)_y < \max\{\text{softmax}(\mathbf{z} + \Delta)_j \mid 1 \leq j \leq C \cap j \neq y\}$ 。由于 softmax 函数是一个单调递增函数,上述不等式又等价于 $z_y + \Delta_y < \max\{z_j + \Delta_j \mid 1 \leq j \leq C \cap j \neq y\}$ 。因此,可以把生成虚拟对抗样本的过程表示为如下的优化问题:

$$\begin{aligned} & \min_{\Delta} \|\Delta\|_2 \\ \text{s. t. } & z_y + \Delta_y < \max\{z_j + \Delta_j \mid 1 \leq j \leq C \cap j \neq y\}. \end{aligned} \quad (3)$$

把上述虚拟对抗样本的生成和对抗训练结合在一起,就是本研究提出的虚拟对抗学习。

2.2 阈值机制的建立

传统的对抗训练只用对抗样本训练模型,虽然能提高模型的鲁棒性,但也会严重降低对干净样本的预测精度。同样地,为了更好地保证虚拟对抗训练的分类效果,需要给定一些约束条件来选择,让一些干净样本生成虚拟对抗样本来参加训练。根据 1.1 节中对抗源样本的定义,可以认为对抗源样本是容易生成对抗样本的干净样本,因此,本研究选择在对抗源样本上添加扰动 Δ 使其生成虚拟对抗样本。

记样本的 logits 中的第二大值为 z_s ,最大值与第二大值之差(logits distance, LD)为 $z_y - z_s$ 。通过

FGSM 攻击获得了 CIFAR-10 和 ImageNet(30)数据集上对抗源样本的 LD 和非对抗源样本的 LD,观察到二者 LD 分布近似服从正态分布,如图 1 所示。显然,对抗源样本的 LD 均值(mean of logits distance, MLD)小于非对抗源样本的 MLD,利用这种差异性建立阈值机制,设定阈值 γ ,把 $LD < \gamma$ 作为约束条件来挑选对抗源样本,就能获得大量对抗源样本。

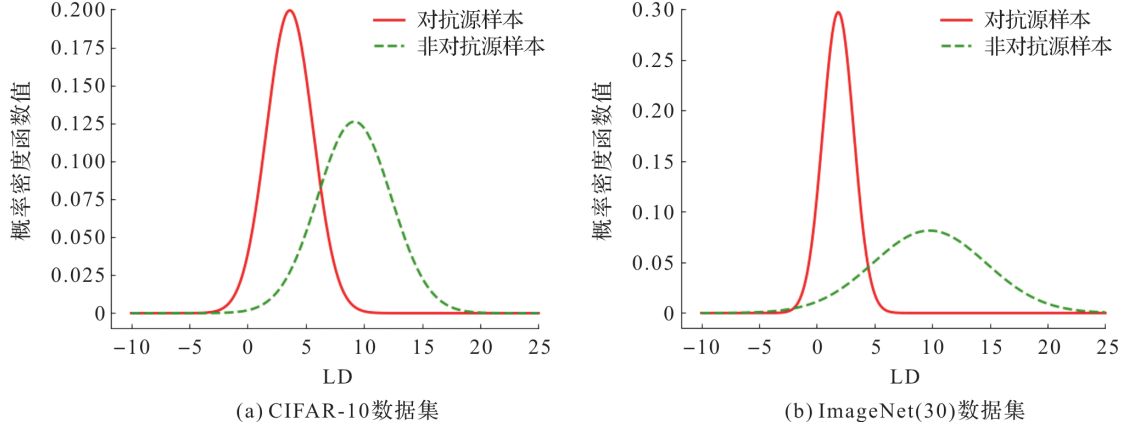


图 1 对抗源样本与非对抗源样本的 LD 分布

Fig. 1 LD distribution of adversarial source examples and non-adversarial source examples

2.3 寻找最小扰动

为了实现式(3)中的目标函数最小化,找到最小扰动,同时满足其约束条件,只需在 z_j ($1 \leq j \leq C$) 中找到第二大值,即 $z_s = \max\{z_j \mid 1 \leq j \leq C \cap j \neq y\}$,使得 $z_y + \Delta_y < z_s + \Delta_s$ 。显然,当 $\{\Delta_j = 0 \mid 1 \leq j \leq C \cap (j \neq y \cap j \neq s)\}$ 时, $\|\Delta\|_2$ 最小。此时, $\|\Delta\|_2 = \sqrt{\Delta_y^2 + \Delta_s^2}$,由于 $\min_{\Delta_y, \Delta_s} f = \Delta_y^2 + \Delta_s^2$ 的解也是 $\min_{\Delta_y, \Delta_s} g = \sqrt{\Delta_y^2 + \Delta_s^2}$ 的解,因此,式(3)又可以转换为

$$\begin{aligned} & \min_{\Delta_y, \Delta_s} \Delta_y^2 + \Delta_s^2 \\ & \text{s. t. } z_y + \Delta_y < z_s + \Delta_s. \end{aligned} \quad (4)$$

由于在约束条件 $z_y + \Delta_y < z_s + \Delta_s$ 下不能得到可行域的边界,故进一步放松约束条件,要求 $z_y + \Delta_y \leq z_s + \Delta_s$,此时,所求的 Δ_y 和 Δ_s 为式(4)的近似最优解。当 $z_y + \Delta_y = z_s + \Delta_s$ 时, $z_y + \Delta_y$ 与 $z_s + \Delta_s$ 通过 softmax 层后输出的置信值相等,则该样本数据位于第一大类和第二大类的决策边界上,即各有 50% 的概率归于两类之一。虚拟对抗训练并不严格要求每个样本都生成虚拟对抗样本,若该样本数据添加扰动后被判别为第一大类,则它作为正确样本参与训练;若被判别为第二大类,则它作为虚拟对抗样本参与训练。根据不等式的求解方法,可以得到 $\Delta_y = -(z_y - z_s)/2$, $\Delta_s = (z_y - z_s)/2$, $\lambda = z_y - z_s$ 。因此,式(3)的近似最优解为 $\Delta^* = (\Delta_1, \dots, \Delta_j, \dots, \Delta_C)$,即

$$\Delta^* = \begin{cases} -\frac{(z_y - z_s)}{2}, & j = y; \\ \frac{(z_y - z_s)}{2}, & j = s; \\ 0, & j \neq y \cap j \neq s. \end{cases} \quad (5)$$

2.4 虚拟对抗训练

利用虚拟对抗样本和干净样本训练模型的方法称为虚拟对抗训练。图 2 为虚拟对抗训练方法示意图,其具体做法是:1) 设定 T 个训练周期,把训练集分为 M 个批次,将第一个批次的训练数据输入参数为 θ_0 的模型,通过神经网络倒数第二层得到 logits,即 z ;选出满足条件 $LD < \gamma$ 的 z ,即对抗源样本的 z ,为其添加相应的扰动 Δ ,非对抗源样本保持不变;通过 softmax 层计算损失,进行反向传播更新参数 θ_0 为 θ_1 。2) 对剩余的 $M-1$ 个批次重复上述过程,直至参数由 θ_{M-1} 更新为 θ_M ,即完成一个周期的训练。3) 重复上述流程,直至完成 T 个周期的训练。

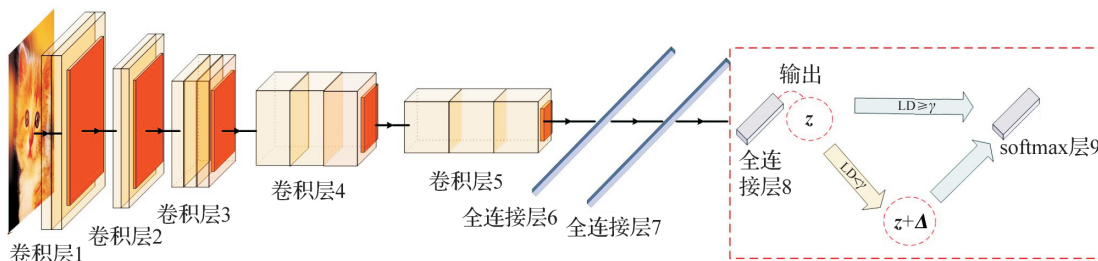


图 2 虚拟对抗训练方法示意图

Fig. 2 Schematic diagram of virtual adversarial training method

3 试验评估

在深度神经网络架构上测试虚拟对抗训练算法,应用于 CIFAR-10 和 ImageNet(30) 图片分类数据集。

3.1 试验设置

本研究的试验采用 2 个基准数据集: CIFAR-10 和 ImageNet(30)。CIFAR-10 包含 60 000 张彩色图片,其中 50 000 张作为训练集,10 000 张作为测试集,图片大小为 32×32 像素,总共 10 类。ImageNet(30) 是在 ImageNet 上随机挑选的 30 类彩色图片,每类 1 350 张,其中 39 000 张作为训练集,1 500 张作为测试集,图片大小为 224×224 像素。由于现实生活中很难知道对抗样本的来源,为了更切合真实情况,试验采用黑盒攻击,即对手不了解目标模型的内部知识,具体而言,用于对抗训练的网络架构与用于生成对抗样本的网络架构不同。

在 CIFAR-10 上几种对抗训练的网络架构为 ResNet-18,测试集为攻击 VGG-16 网络架构训练的模型生成的扰动样本。在 ImageNet(30) 上使用 Pytorch 提供的预训练模型进行再训练,几种对抗训练的网络架构为 AlexNet,测试集为攻击 ResNet-18 网络架构训练的模型生成的扰动样本。

本试验硬件如下:显卡型号为 Titan XP,有 4 个 GPU,内存为 12 GB;服务器操作系统为 Ubuntu16.04.6 LTS。模型的训练使用了循环学习率^[16],可以大幅减少深度神经网络训练所需要的训练周期。

3.2 阈值对虚拟对抗训练的影响

阈值 γ 作为一个超参数,会影响经过虚拟对抗训练后的模型的效用。以下从 3 个方面验证阈值对虚拟对抗训练的影响:训练所需的时长;训练后的模型对干净样本的测试精度;训练后的模型对扰动样本的测试精度。

在 CIFAR-10 上设定虚拟对抗训练的训练周期为 4,网络架构为 ResNet-18,循环学习率最小值为 0,最大值为 0.2,干净样本集为测试集上被 F_1 分类正确的样本,其中 F_1 是使用干净样本对 ResNet-18 进行训练得到的模型,扰动样本集为对 F_1 进行 FGSM 攻击生成的样本,扰动大小为 0.05。在 ImageNet(30) 上设定虚拟对抗训练的训练周期为 4,网络架构为 AlexNet,循环学习率最小值为 0,最大值为 0.02,干净样本集为测试集上被 F_3 分类正确的样本,其中 F_3 是使用干净样本对 AlexNet 进行训练得到的模型,扰动样本集为对 F_3 进行 FGSM 攻击生成的样本,扰动大小为 0.007。图 3 为阈值对虚拟对抗训练时长、干净样本预测精度和扰动样本预测精度的影响。

由图 3(a)和(d)可知,随着阈值的增加,虚拟对抗训练所需的时间呈现增加的趋势。在 CIFAR-10 上,当 $\gamma < 6$ 时,随着阈值的增加,虚拟对抗训练的训练时间增加得比较快;而 $\gamma > 6$ 时,训练时间增加得较慢,基本上稳定在 6.25 min。类似地,在 ImageNet(30) 上,当 $\gamma > 15$ 时,训练时间趋于稳定。阈值较大时,训练时间趋于稳定是因为当阈值大于某个值后,大多数样本都改变其 logits,此时再增加阈值,也不会改变更多的样本,因而也不会再延长训练时间。

由图 3(b)和(e)可知,无论在 CIFAR-10 上还是在 ImageNet(30) 上,随着阈值增大,虚拟对抗训练后

的模型会影响对干净样本的分类效果。尤其是在 CIFAR-10 上,预测精度的最大值与最小值之差约为 5.9%。

由图 3(c)和(f)可以看出,随着阈值的增大,虚拟对抗训练后的模型对扰动样本的预测精度提高。但从预测精度的最大值和最小值来看,在 CIFAR-10 上,二者之差不超过 1.6%,测试精度相差不大;而在 ImageNet(30)上,二者之差大约为 3.5%,这也说明阈值增大能够增强模型的鲁棒性。

综上所述,阈值越小,训练时间越短,对干净样本的预测效果越好,对扰动样本的防御效果影响不大。因此,在应用虚拟对抗训练时可以选取一个较小的阈值。

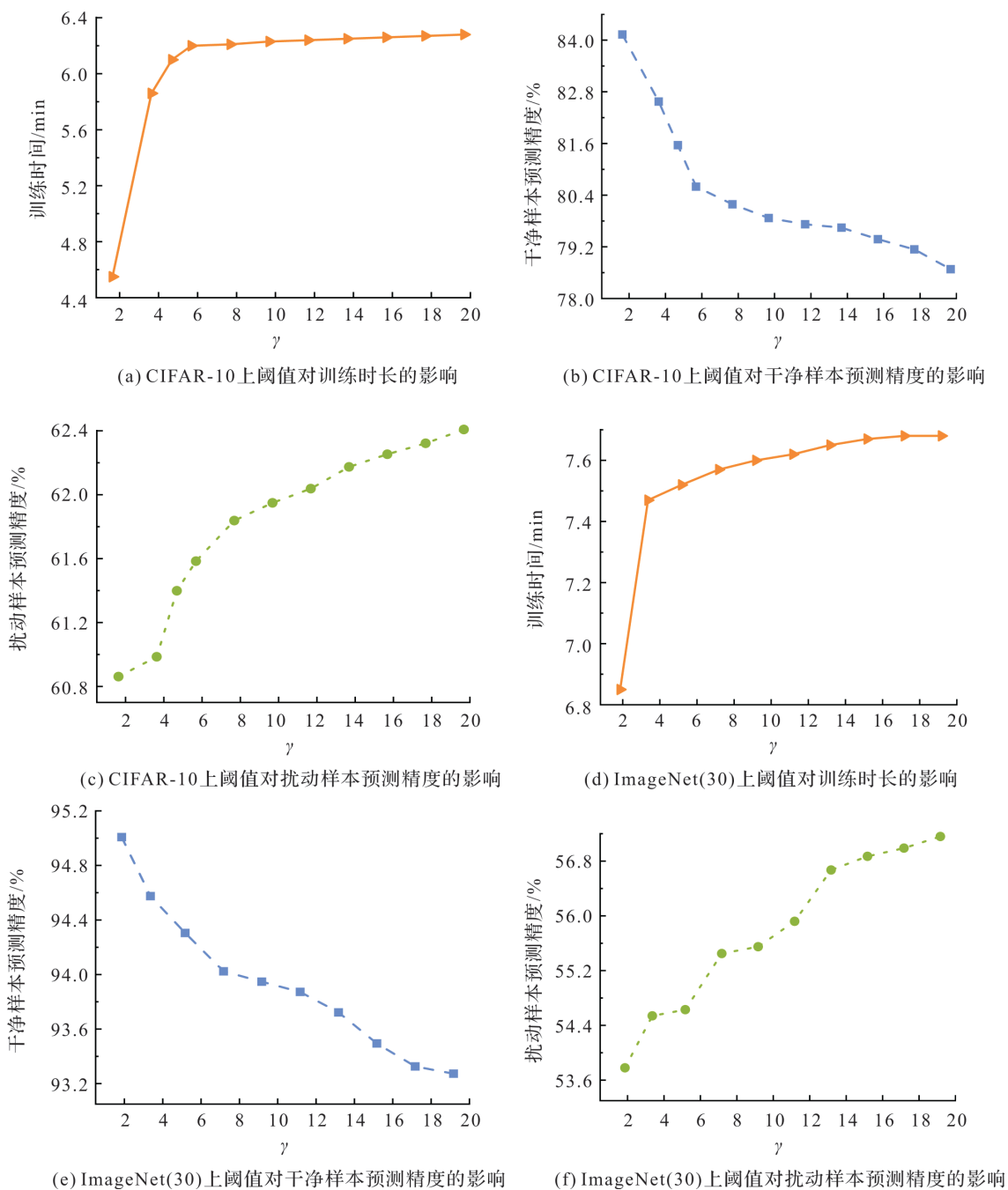


图3 阈值对虚拟对抗训练时长、干净样本预测精度和扰动样本预测精度的影响

Fig. 3 Influence of threshold on training time, prediction accuracy of clean examples and perturbed examples for virtual adversarial training

3.3 虚拟对抗训练与其他对抗训练的比较

把虚拟对抗训练应用于 CIFAR-10 和 ImageNet(30)数据集上,以对比几种对抗训练后的模型对于干净样本的预测精度、对扰动样本的预测精度及模型训练时间。在 CIFAR-10 和 ImageNet(30)上,分别使

用虚拟对抗训练、FGSM 对抗训练、Fast 对抗训练及 PGD 对抗训练进行试验。

在 CIFAR-10 上,4 种对抗训练的网络架构为 ResNet-18,训练周期为 4,循环学习率最小值为 0,最大值为 0.2。对于虚拟对抗训练,阈值 $\gamma=3.63$;对于 FGSM 对抗训练,扰动大小为 $16/255$;对于 Fast 对抗训练,扰动大小为 $16/255$,步长为 $20/255$;对于 PGD 对抗训练,扰动大小和步长分别为 $16/255$ 和 $8/255$,迭代次数为 4。为了说明对抗训练能提升对扰动样本的防御效果,只用干净样本训练的 ResNet-18 模型(即 ResNet-18 标准训练)作为对照,训练周期也为 4。此外,使用干净样本训练 VGG-16 模型(即 F_2),用于得到干净测试集和扰动测试集。干净测试集为 CIFAR-10 测试集上被 F_2 预测正确的样本,扰动测试集为攻击 F_2 生成的扰动样本,攻击方法分别为 FGSM、基本迭代法(basic iterative methon, BIM)^[17]、PGD、RFGSM^[18]、Fast、C&W(Carlini&Wagner)攻击^[19]和动量差分输入迭代快速梯度符号法(momentum diverse input iterative fast gradient sign method, M-DI²-FGSM)^[20]。

在 ImageNet(30)数据集上,4 种对抗训练的网络架构为 AlexNet,训练周期为 4,循环学习率最小值为 0,最大值为 0.02。对于虚拟对抗训练,阈值 $\gamma=2.0$;对于 FGSM 对抗训练,扰动大小为 $2/255$;对于 Fast 对抗训练,扰动大小为 $2/255$,步长为 $3/255$;对于 PGD 对抗训练,扰动大小和步长分别为 $2/255$ 和 $1/255$,迭代次数为 3。使用干净样本训练的 AlexNet 模型(即 AlexNet 标准训练)作为对照,训练周期为 4。使用干净样本训练 ResNet-18 模型(即 F_4),用于得到干净测试集和扰动测试集,干净测试集为 ImageNet(30)测试集上被 F_4 预测正确的样本,扰动测试集为攻击 F_4 生成的扰动样本,攻击方法分别为 FGSM、BIM、PGD、RFGSM、Fast、C&W 和 M-DI²-FGSM。

3.3.1 对抗训练后的模型对干净样本分类效果的比较

图 4 为标准训练和对抗训练后的模型对干净样本的预测精度。从图 4 中可以看出,与标准训练相比,对抗训练后的模型会降低对干净样本的分类精度,而本研究提出的虚拟对抗训练优于其他对抗训练,在 CIFAR-10 上,预测精度比其他对抗训练的模型预测精度高出 7.21~14.68 个百分点,在 ImageNet(30)上,预测精度高出 8.31~11.47 百分点,甚至高于标准训练模型的预测精度。

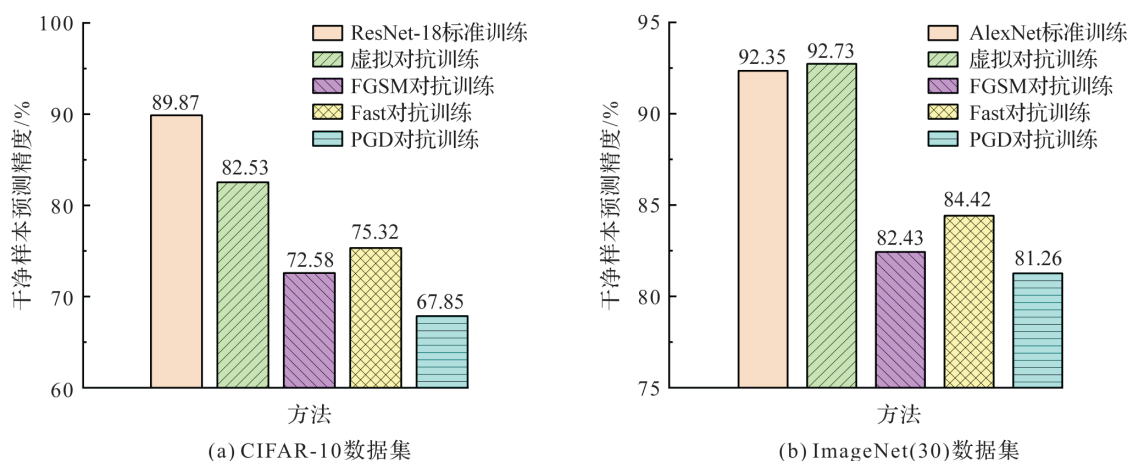


图 4 标准训练和对抗训练后的模型对干净样本的预测精度

Fig. 4 Prediction accuracy of clean examples tested by models after standard training and adversarial training

在 CIFAR-10 数据集上标准训练和对抗训练后的模型对扰动样本的预测精度见表 1。在表 1 中,FGSM 中的扰动大小为 $12/255$;BIM 和 PGD 中的扰动大小、步长和迭代次数分别为 $12/255$ 、 $4/255$ 、4;RFGSM 中的扰动大小为 $12/255$,步长为 $8/255$,迭代次数为 2;Fast 的参数为扰动大小 $12/255$,步长为 $15/255$;C&W 为无目标攻击,其中箱型约束为 1,置信度为 1,迭代次数为 30,Adam 学习率为 0.01;M-DI²-FGSM 中扰动大小为 $12/255$,迭代次数为 2,衰减率和变换可能性均为 1。最后一列为 F_2 在各种攻击下的预测精度,是为了判别扰动测试集中对抗样本所占的比例,以更直观地反映其他模型对 F_2 生成的扰动样本的防御效果。

表1 在 CIFAR-10 数据集上标准训练和对抗训练后的模型对扰动样本的预测精度

Table 1 Prediction accuracy of perturbed examples tested by models after standard training and adversarial training on CIFAR-10 data set

攻击方法	虚拟对抗训练	FGSM 对抗训练	Fast 对抗训练	PGD 对抗训练	ResNet-18 标准训练	VGG-16(F_2)
FGSM	72.10	62.15	68.18	72.45	55.31	37.14
BIM	71.47	62.85	67.32	73.22	61.37	40.30
PGD	71.40	60.57	67.19	72.46	46.54	14.68
RFGSM	72.23	61.63	68.17	72.09	55.46	30.83
Fast	71.49	61.23	68.20	72.51	54.68	30.49
C&W	46.01	27.64	32.58	37.26	31.26	31.28
M-DI ² -FGSM	68.07	58.12	65.17	70.45	48.80	22.42

在 ImageNet(30)上标准训练和对抗训练后的模型对扰动样本的预测精度见表2。在表2中,FGSM 中的扰动大小为 $2/255$;BIM 和 PGD 中的扰动大小为 $2/255$,步长为 $1/255$,迭代次数为 3;RFGSM 中的扰动大小为 $2/255$,步长为 $1/255$,迭代次数为 2;Fast 中的扰动大小为 $2/255$,步长为 $3/255$;C&W 为无目标攻击,其中箱型约束为 5,置信度为 1,迭代次数为 20,Adam 学习率为 0.01;M-DI²-FGSM 中的扰动大小为 $2/255$,迭代次数为 2,衰减率和变换可能性均为 1。

表2 在 ImageNet(30)数据集上标准训练和对抗训练后的模型对扰动样本的预测精度

Table 2 Prediction accuracy of perturbed examples tested by models after standard training and adversarial training on ImageNet(30) data set

攻击方法	虚拟对抗训练	FGSM 对抗训练	Fast 对抗训练	PGD 对抗训练	AlexNet 标准训练	ResNet-18(F_4)
FGSM	90.61	85.66	86.19	86.60	80.23	27.85
BIM	90.93	85.22	86.11	86.30	79.64	11.82
PGD	91.37	84.88	85.89	86.75	79.26	12.78
RFGSM	90.84	85.81	85.81	86.67	80.30	17.49
Fast	91.25	85.73	85.91	86.53	80.23	29.24
C&W	77.59	41.51	44.32	47.38	40.50	16.16
M-DI ² -FGSM	86.30	77.81	80.04	83.83	65.30	17.56

3.3.2 对抗训练后的模型对扰动样本分类效果的比较

通过表1和表2可知,与标准训练后的模型相比,对抗训练极大地增强了模型的鲁棒性。通过比较4种对抗训练模型的防御效果可以看出,在 CIFAR-10 上,虚拟对抗训练的防御精度仅低于 PGD 对抗训练 1.06 百分点,优于 FGSM 对抗训练和 Fast 对抗训练 4.21~10.83 百分点,但在 C&W 攻击下,虚拟对抗训练的防御效果最好。在 ImageNet(30)上,虚拟对抗训练的防御效果优于其他3种对抗训练。

3.3.3 对抗训练时间的比较

表3为4种对抗训练所需的时间。显然,虚拟对抗训练由于不用生成物理意义上的对抗样本,无须计算损失对样本的梯度,从而提高了训练速度。在 CIFAR-10 和 ImageNet(30)数据集上,与目前较先进的 Fast 对抗训练相比,本研究提出的方法分别缩减了 24.19%、26.84% 的训练时间,与最慢的 PGD 对抗训练方法相比,虚拟对抗训练缩减了将近 $2/3$ 的时间。

表3 4种对抗训练所需的时间

Table 3 Time required for four kinds of adversarial training

数据集	虚拟对抗训练	FGSM 对抗训练	Fast 对抗训练	PGD 对抗训练
CIFAR-10	5.86	7.63	7.73	17.37
ImageNet(30)	7.25	9.64	9.91	21.37

3.3.4 虚拟对抗训练有效的原因

通过试验证明,与传统的对抗训练方法相比,虚拟对抗训练使得满足 $LD < \gamma$ 的样本生成虚拟对抗样本,而这些样本位于第一大类和第二大类的决策边界上,每当模型更新参数时,相当于对决策边界进行微

调,因此对于干净精度的影响较小。同时,由于训练过程不断更新模型参数,这种微调的效果放大,就能有效地防御对抗样本。此外,本研究方法无须考虑计算损失对样本的梯度以更新扰动来生成真实的对抗样本,从而缩短了训练时间。

4 结 语

针对传统对抗训练耗时且严重降低对于干净样本预测精度的缺点,本研究提出了一种不需要生成对抗样本的虚拟对抗学习方法,通过试验研究了超参数阈值对虚拟对抗训练的影响,并比较了虚拟对抗训练与其他对抗训练的优劣。试验结果证明,本文方法不仅能加速对抗训练过程,在速度上优于传统的对抗训练方法,而且能够增强模型的鲁棒性,同时对于干净样本的分类效果影响不大。这是一种具有启发意义的对抗训练方法,未来,我们将借鉴虚拟对抗样本的概念,结合真正的对抗样本生成方法来探索新的攻击与防御方法。

参考文献:

- [1] 郑远攀,李广阳,李晔. 深度学习在图像识别中的应用研究综述[J]. 计算机工程与应用,2019,55(12):20.
- [2] 鱼昆,张绍阳,侯佳正,等. 语音识别及端到端技术现状及展望[J]. 计算机系统应用,2021,30(3):14.
- [3] 王睿怡,罗森林,吴舟婷,等. 深度学习在汉语语义分析的应用与发展趋势[J]. 计算机技术与发展,2019,29(9):110.
- [4] SZEGEDY C, ZAREMBA W, SUTSKEVER I, et al. Intriguing properties of neural networks[C]//International Conference on Learning Representations (ICLR). Banff: OpenReview.net,2014:1.
- [5] GOODFELLOW I J, SHLENS J, SZEGEDY C. Explaining and harnessing adversarial examples[C]//International Conference on Learning Representations (ICLR). San Diego: OpenReview.net,2015:1.
- [6] 吴立人,刘政浩,张浩,等. 聚焦图像对抗攻击算法 PS-MIFGSM[J]. 计算机应用,2020,357(5):112.
- [7] 张高志,刘新平,邵明文. 用于白盒目标攻击的 GAN 对抗样本生成[J]. 模式识别与人工智能,2020,33(9):830.
- [8] 钱亚冠,刘新伟,顾钊铨,等. 一种基于二维码对抗样本的物理补丁攻击[J]. 信息安全学报,2020,5(6):79.
- [9] 许笑,陈奕君,冯诗羽,等. 基于冗余信息压缩的深度学习对抗样本防御方案[J]. 网络空间安全,2020,11(8):11.
- [10] 范宇豪,张铭凯,夏仕冰. 基于插值法的对抗攻击防御算法[J]. 网络空间安全,2020,11(4):74.
- [11] MADRY A, MAKELOV A, SCHMIDT L, et al. Towards deep learning models resistant to adversarial attacks [C]//International Conference on Learning Representations (ICLR). Vancouver: OpenReview.net,2018:1.
- [12] SHAFABI A, NAJIBI M, GHIASI A, et al. Adversarial training for free! [C]//Advances in Neural Information Processing Systems (NeurIPS). Vancouver: MIT Press,2019:3358.
- [13] ZHANG D H, ZHANG T Y, LU Y P, et al. You only propagate once: painless adversarial training using maximal principle[C]//Conference on Neural Information Processing Systems (NeurIPS). Vancouver: MIT Press,2019:1.
- [14] ZHU C, CHENGY, GAN Z, et al. FreeLB: Enhanced adversarial training for natural language understanding[C]//International Conference on Learning Representations (ICLR). Addis Ababa: OpenReview.net,2020:1535.
- [15] WONG E, RICE L, KOLTER J Z. Fast is better than free: revisiting adversarial training [C]//International Conference on Learning Representations (ICLR). Addis Ababa: OpenReview.net,2020:1095.
- [16] SMITH L N. Cyclical learning rates for training neural networks[C]//IEEE Winter Conference on Applications of Computer Vision (WACV). Santa Rosa: IEEE,2017:464.
- [17] KURAKIN A, GOODFELLOW I, BENGIO S. Adversarial examples in the physical world [C]//International Conference on Learning Representations (ICLR). Toulon: OpenReview.net,2017:1.
- [18] TRAMER F, KURAKIN A, PAPERNOT N, et al. Ensemble adversarial training: attacks and defenses [C]//International Conference on Learning Representations (ICLR). Vancouver: OpenReview.net,2018:1.
- [19] CARLINI N and WAGNER D. Towards evaluating the robustness of neural networks [C]//IEEE Symposium on Security and Privacy (SP). San Jose: IEEE,2017:39.
- [20] XIE C H, ZHANG Z S, ZHOU Y Y, et al. Improving transferability of adversarial examples with input diversity [C]//IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach: IEEE,2019:2725.