基于积分损失的对抗样本生成算法

章 进,李 琦
(南京邮电大学 计算机学院,江苏 南京 210023)

摘 要:随着计算机性能的飞速提升和数据量的爆炸式增长,深度学习在越来越多的领域取得了惊人的成果。然而,研究者们发现深度网络也存在对抗攻击。在图像分类领域,攻击者可以通过向原始的图片上加入人为设计的微小的扰动,来使得深度神经网络分类器给出错误的分类,而这种扰动对于人类来说是不可见的,加入了扰动之后的图片就是对抗样本。基于梯度攻击的对抗样本生成算法(projected gradient descent,PGD)是目前有效的攻击算法,但是这类算法容易产生过拟合。该文提出了积分损失快速梯度符号法,利用积分损失来衡量输入对于损失函数的重要性程度,规避梯度更新方向上可能陷入局部最优值的情况,不仅进一步提升了对抗样本的攻击成功率,而且也增加了对抗样本的迁移性。实验结果证明了所提方法的有效性,可以作为测试防御模型的一个基准。

关键词:对抗样本;白盒攻击;积分梯度;卷积神经网络;深度学习
中图分类号:TP391.41;TP183 **文献标识码**:A **文章编号**:1673-629X(2022)07-0001-07
doi:10.3969/j.issn.1673-629X.2022.07.001

Adversarial Examples Generation Algorithm Based on Integrated Loss

ZHANG Jin,LI Qi
(School of Computer Science,Nanjing University of Posts and Telecommunications,Nanjing 210023,China)

Abstract:With the rapid improvement of computer performance and the explosive growth of data,deep learning has achieved amazing results in more and more fields. However,researchers have found that deep networks are also vulnerable to adversarial attacks. In the field of image classification,the attackers can add artificially designed small perturbations to the original image to make the deep neural network classifier give the wrong classification,which is invisible to human beings. The image with perturbations is called the adversarial example. The projected gradient descent (PGD) algorithm based on gradient attack is an effective adversarial examples generation algorithm at present,but this kind of algorithm is easy to over fit. In this paper,the integrated loss fast gradient sign method is proposed,which uses the integrated loss to measure the importance of the input to the loss function,and avoids the situation that the gradient update direction may fall into the local optimal value. The proposed algorithm further improves the attack success rate of the adversarial sample. Furthermore,it also increases the transferability of the adversarial examples. The experiments results show the effectiveness of the proposed method,which can be used as a benchmark to test the defense model.

Key words:adversarial examples;white-box attack;integrated gradients;convolutional neural network;deep learning

0 引 言

目前,深度学习在许多领域取得了迅猛的发展。例如:机器视觉[1-2]、语音识别[3]、自然语言处理[4]、恶意软件检测[5]等,甚至一度超过了人类的水平。但是神经网络与其他系统一样存在安全性和鲁棒性的问题。通过添加一些精心设计的噪声到图片上,可以使得深度神经网络给出置信度非常高的错误的预测,然而这些噪声对于人类来说是不可见的。添加了这些噪声的图片就称之为对抗样本[6],对应的攻击称之为对抗攻击。在实际生活中,深度学习的部署需要较高的安全性,例如人脸识别[7]、自动驾驶[8-9]等,因此研究强有力的对抗攻击算法,对于理解深度网络内在的脆弱性,进一步提升模型的鲁棒性和安全性就变得非常有意义。

在探索深度学习可解释性的过程中,Christian Szegedy 等人[6]提出了对抗样本(adversarial examples)的概念,即在数据集中通过添加细微的扰动所形成的输入样本,将导致模型以高置信度给出一个错误的输出。他们发现许多深度学习模型,包括卷积神经网络(convolutional neural network,CNN)对于对抗样本都

收稿日期:2021-08-06 **修回日期**:2021-12-08
基金项目:科技部重点研发计划(2019YFB2101704)
作者简介:章 进(1997-),男,硕士,研究方向为深度学习对抗攻击;李 琦,博士,副教授,研究方向为信息安全。

具有极高的脆弱性。同时,对抗样本具有迁移性,很多情况下,在训练集的不同子集上训练得到的网络架构不同的模型都会对同一个对抗样本做出错误的分类。根据目标模型的架构和参数是否已知可以将对抗攻击分为:白盒攻击和黑盒攻击。最近几年,有许多的攻击算法相继提出,其中主要的是基于梯度的攻击算法。这些算法可以被进一步地划分为单步的和多步的。Goodfellow 等人[10]指出神经网络对对抗样本表现脆弱性的原因是神经网络的线性性,与早期所认为的非线性和过拟合有所不同,并提出了一种快速生成对抗样本的方法(fast gradient sign method,FGSM),这个算法是单步的并且具有较高的迁移性。Kurakin 等人[11]将 FGSM 进一步改进,提出了多步的迭代版本的 FGSM(iterative fast gradient sign method,I-FGSM),这个算法进一步提升了白盒攻击的成功率,但是对于目标模型产生了过拟合,迁移性较差。为了进一步提升攻击成功率和迁移性,Dong 等人[12]将动量引入到 I-FGSM,提出了(momentum iterative fast gradient sign method,MI-FGSM)算法。该算法可以有效避免震荡,稳定对抗样本更新方向,加速逼近最优值。之后,Jiadong Lin 等人[13]认为 MI-FGSM 算法使得梯度不断累积,无限制加速,可能会错过最优值,将 Nesterov 集成到了 MI-FGSM,提出了(Nesterov-momentum iterative fast gradient sign method,NI-FGSM)算法。该算法在每次对抗样本更新的时候粗略地估算下一次的位置,达到及时减速的目的,来避免梯度更新的太快。

由于之前的对抗样本生成算法可能会陷入局部最优值的情况,需要一种方法来有效地评估样本的梯度。Sundararajan 等人[14]认为对于非线性深度网络,输入对于输出的梯度很容易饱和,导致一个重要的输入可能会有一个很小的梯度,并提出积分梯度(integrated gradients)这一算法。该算法通过在数据点周围间隔小范围的均匀采样,并对这些采样的数据进行梯度计算,最后将这些梯度进行累加,用来表示当前样本的梯度。用这个集成的梯度更好地捕获了输入对于输出的重要性。受到这种方法的启发,该文将间隔小范围的均匀采样变为按照间隔指数增长范围进行采样,从而避免采样次数过多所带来的计算消耗和大量无用的相似样本采样。同时,将采样的样本用于损失函数的计算,这样做不仅考虑了原始样本的损失同样考虑了当前样本线性比例上的损失,可以看作是在当前样本上的损失的一个集成,将其称之为积分损失(integrated loss)。在此基础上提出了积分损失快速梯度符号法(integrated loss fast gradient sign method,IL-FGSM)。该算法利用积分损失作为优化的目标函数,一定程度上避免了梯度饱和的情况,从而更好地达到全局最优值。实验结果表明,IL-FGSM 效果较好,相比于基线方法提升了 10% ~20% 的攻击成功率。

1 相关知识

1.1 符号说明

神经网络是一个函数 $F(x)=y$,接受一个输入 $x \in R^n$,产生一个输出 $y \in R^m$。模型 F 也隐式地包含一些模型参数 θ。该文重点研究了用作 m 类分类器的神经网络。使用 softmax 函数计算网络的输出,该函数确保输出向量 y 满足 $0 \leqslant y_i \leqslant 1$, $y_1+\cdots+y_m=1$。因此,输出向量 y 被视为概率分布,即 y_i 被视为输入 x 具有类别 i 的概率。分类器将 $C(x)=\text{argmax}F(x)_i$ 作为输入 x 的标签。设 $C^*(x)$ 为 x 的正确标签。softmax 函数的输入称为 logits。定义 F 为包含 softmax 函数的全连接神经网络, $Z(x)=z$ 为除了 softmax 之外的所有层的输出,所以 z 为 softmax 的输入,即 logits,则:

$$F(x)=\text{softmax}(Z(x))=y$$

1.2 对抗样本

Szegedy 等人[6]首先指出了对抗样本的存在:给定有效的输入 x 和目标 $t \neq C^*(x)$,通常可以找到类似的输入 x',使得 $C(x')=t$,但 x, x' 根据某种距离度量是接近的。样本 x' 具有这个属性被称为有目标的对抗样本。相反是无目标的对抗样本,只寻找输入 x',满足 $C(x') \neq C(x)$,并且 x, x' 很接近,而不是将 x 分类为给定的目标类别。因此,无目标攻击比起有目标攻击实施起来更加容易。

2 白盒攻击算法

研究者们提出了许多的方法来生成对抗样本,这里进行一个简要的介绍。

2.1 I-FGSM & PGD

由于 FGSM 是在梯度的符号方向上进行一次的单个步长 ε 的扰动,更新生成的对抗样本扰动强度较大,Kurakin 等人[11]提出了基础迭代法 I-FGSM。该算法采用多个较小的步长 α 更新优化扰动强度,同时将结果裁剪到约束范围 ε,产生的对抗样本攻击能力更强。

$$X_0^{\text{adv}}=X$$

$$X_{n+1}^{\text{adv}}=\text{Clip}_X^{\varepsilon}\{X_n^{\text{adv}}+\alpha\cdot\text{sign}(\nabla_x L(X_n^{\text{adv}},y^{\text{true}}))\}$$

其中,ε 表示总的扰动大小,α 表示单步扰动的大小,通常设置为 ε/T,其中 T 表示总的迭代次数。$\text{Clip}_X^{\varepsilon}$ 表示裁剪操作,保证 X^{adv} 是一个有效的数据,例如图片的范围[0,255]。

Madry 等人[15]提出了梯度投影下降方法(projected gradient descent,PGD),一个更强的 FGSM 方法的变种,主要思想是在更新对抗样本前,使用一个

随机的起点作为对抗样本的初始值。

2.2 MI-FGSM & NI-FGSM

为了解决 I-FGSM 的迁移性较差的问题,Yong 等人[12]提出了动量迭代快速梯度符号法(MI-FGSM)。该方法将动量项加入到攻击的过程中,来稳定的更新方向,避免了迭代过程中可能出现的梯度更新震荡和落入较差的局部最优值。更新步骤类似于 I-FGSM,替换的公式如下:

$$g_{n+1} = \mu \cdot g_n + \frac{\nabla x L(X_n^{\text{adv}}, y^{\text{true}}; \theta)}{\| \nabla x L(X_n^{\text{adv}}, y^{\text{true}}; \theta) \|_1}$$

$$X_{n+1}^{\text{adv}} = \text{Clip}_X^{\varepsilon}\{X_n^{\text{adv}} + \alpha \cdot \text{sign}(g_{n+1})\}$$

其中,μ 是动量项衰减因子,通常设置为 1;$g_0 = 0$,g_n 是第 n 次的搜集的梯度。

Lin 等人[13]利用 Nesterov 来加速梯度下降并稳定梯度更新方向,在每次计算梯度前,提前使用下一次的对抗样本作为当前对抗样本,提出了 NI-FGSM 算法,公式如下:

$$X_n^{\text{nes}} = X_n^{\text{adv}} + \alpha \cdot \mu \cdot g_n$$

$$g_{n+1} = \mu \cdot g_n + \frac{\nabla x L(X_n^{\text{nes}}, y^{\text{true}}; \theta)}{\| \nabla x L(X_n^{\text{nes}}, y^{\text{true}}; \theta) \|_1}$$

$$X_{n+1}^{\text{adv}} = \text{Clip}_X^{\varepsilon}\{X_n^{\text{adv}} + \alpha \cdot \text{sign}(g_{n+1})\}$$

2.3 DIM & TIM

Xie 等人[16]将输入多样性加入对抗样本的生成过程,进一步改善了对抗样本的迁移性,提出了 DIM(diverse input method)算法。DIM 的更新步骤和 I-FGSM 相似,具有如下的替换:

$$X_{n+1}^{\text{adv}} = \text{Clip}_X^{\varepsilon}\{X_{\text{n}}^{\text{adv}} + \alpha \cdot \text{sign}(\nabla x L(T(X_n^{\text{adv}}; p), y^{\text{true}}; \theta))\}$$

$T(X_n^{\text{adv}}; p)$ 为随机变换函数,具体的公式如下:

$$T(X_n^{\text{adv}}; p) = \begin{cases} T(X_n^{\text{adv}}) & \text{with probability } p \\ X_n^{\text{adv}} & \text{with probability } 1 - p \end{cases}$$

其中,p 是概率值,表示有 p 的概率使用这个随机变换函数,p 的概率保持原始的输入。这样做的目的是为了在不减少白盒攻击成功率的情况下,进一步提升黑盒攻击的成功率。通常设置 $p = 0.5$。

他们又将 DIM 和 MI-FGSM 整合到一起提出了 M-DI^2-FGSM,直觉上,动量和多样性输入是两个完全不同的方式来缓解过拟合的现象,通过将它们自然地结合到一起形成一个更强的攻击。总体上的更新过程和 MI-FGSM 相似,其中梯度的更新替换如下:

$$g_{n+1} = \mu \cdot g_n + \frac{\nabla x L(T(X_n^{\text{adv}}; p), y^{\text{true}}; \theta)}{\| \nabla x L(T(X_n^{\text{adv}}; p), y^{\text{true}}; \theta) \|_1}$$

由于 DIM 是在单个样本上进行的优化扰动,Dong 等人[17]在计算当前对抗样本时,对该样本进行一系列的图像变换操作,形成一个表示当前对抗样本的集合,用该集合来优化对抗扰动,由于计算效率的原因,他们进一步提出了 TIM(translation-invariant method)。具体的,通过将没有变换的原始图片和一个内核矩阵(通常为高斯核)进行卷积操作,对梯度进行高斯模糊,以此来增加对抗样本的鲁棒性。DIM 和 TIM 都是增加对抗样本迁移性的方法,通过将这两个方法结合到一起是目前有效的增加对抗样本迁移性的方法。

3 IL-FGSM & ENS-IL-FGSM

这里首先介绍积分梯度的概念,然后给出所提方法 IL-FGSM 的定义。

3.1 Integrated Gradients

为了有效评估模型输入对于输出的梯度,Sundararajan 等人[14]提出了积分梯度(integrated gradients)这一算法。该算法初始时输入一个零排列的矩阵,随后让输入数据逐步向测试的目标数据转变,以此通过模型输出的变化反过来研究输入对于输出的影响,有效估计了模型输入对于输出的影响程度,一定程度上避免了输入对于输出的过饱和情况。具体的公式如下:

$$\text{IntegratedGrads}(x) = (X - X^{\text{baseline}}) \cdot \frac{1}{s} \cdot \sum_{i=1}^{s} \frac{\partial Z[X^{\text{baseline}} + \frac{i}{s} \cdot (X - X^{\text{baseline}})]}{X}$$

X^{baseline} 在他们的设置中,对于图片是纯黑的图片,对于文本数据是全为零的嵌入向量。s 是估计 X 的积分梯度需要计算的采样总数。

3.2 IL-FGSM

基于积分梯度算法,提出了积分损失快速梯度符号法 IL-FGSM。该算法将积分损失(integrated loss)作为它的损失函数,替换了原本的对抗样本生成算法中的单一的损失,一定程度上避免由单一损失计算出来的梯度出现饱和的现象,更容易地估算出当前样本对于模型输出的梯度,从而更好地达到全局最优值。

具体的,在每次的迭代过程中,IL-FGSM 依靠当前样本的积分损失来更新输入的图片:

$$\text{IL} = \frac{1}{s} \sum_{i=0}^{s-1} L[X^{\text{baseline}} + \frac{1}{2^i} \cdot (X_n^{\text{adv}} - X^{\text{baseline}}), y; \theta]$$

其中,IL 为 Integrated Loss,s 表示样本 X_n^{adv} 采样的次数,X^{baseline} 设置为 0,这里的 $L(\cdot)$ 可以是一般的交叉熵损失函数,也可是 C&W[18]攻击算法中提出的损失函数,因此可以很容易地集成到现有的攻击算法中。

3.3 ENS-IL-FGSM

同时攻击多个模型,称为集成攻击。与攻击单个模型相比,同时攻击多个模型,可以显著提高对抗样本的迁移性。集成攻击的思想十分直观,如果一个对抗

样本能同时攻击多个模型,那么它很可能对其他模型仍具有攻击性。

采用攻击多个模型的 logits 集成,由于 logits 捕捉概率预测之间的对数关系,因此由 logits 融合的模型集合汇集了所有模型的精细细节输出,这些模型的脆弱性很容易被发现。具体的,攻击 K 个模型:

$$Z(x)=\sum_{k=1}^{K}w_k z_k(x)$$

其中,$z_k(x)$ 表示第 k 个模型的 logits,w_k 表示该模型对应的权重,$w_k\geqslant 0$,$\sum_{k=1}^{K}w_k=1$。损失函数 $L(X,y;\theta)$ 定义为给定真实标签 y 和 logits 值 $Z(x)$ 的 softmax 交叉熵:

$$L(X,y;\theta)=-1_y\cdot\log(\mathrm{softmax}(Z(x)))$$

其中,-1_y 是标签 y 的 one-hot 编码的向量。将攻击多个模型的策略集成到提出的方法 IL-FGSM,并命名为 ENS-IL-FGSM(ensemble integrated loss fast gradient sign method)。相较于 IL-FGSM,ENS-IL-FGSM 具有如下的替换:

$$\mathrm{IL}=\frac{1}{s}\sum_{i=0}^{s-1}\sum_{k=1}^{K}-1_y\cdot\log(\mathrm{softmax}(w_k\cdot z_k(X^{\mathrm{baseline}}+\frac{1}{2^i}\cdot(X_n^{\mathrm{adv}}-X^{\mathrm{baseline}}))))$$

在每次的迭代中,不再采用单一模型的 logits,而是采用多个模型的 logits 的集成,来表示当前样本的 logits,然后再计算该样本的梯度,进而更新样本 X_n^{adv}。

4 实 验

通过实验来证明所提 IL-FGSM 方法和 ENS-IL-FGSM 的优势。首先,提供了实验的相关设置,然后,比较了积分损失的采样策略,接着,分析了积分损失的采样次数问题。之后,将该方法和几个基线方法在常规训练和对抗训练的模型上进行了比较。最后,将增加对抗样本迁移性的方法与该方法结合起来,与基线方法进行了进一步的比较。

4.1 设 置

数据集:攻击一个不能将原始的图片正确分类的分类器是没有意义的,所以随机选择了 ILSVRC2012 验证集上的 1 000 张属于 1 000 个类别的图片,这些图片都可以被本实验的所有分类器正确分类。

网络:考虑了 7 个模型,其中 4 个是常规训练的模型:Inception-v3(Inc-v3)[19],Inception-v4(Inc-v4),Inception-Resnet-v2(IncRes-v2)[20] 和 Resnet-v2-101(Res-101)[21],3 个是对抗训练的模型:Inc-v3-ens3,Inc-v3-ens4 和 IncRes-v3-ens[22]。

超参数:对于网络的超参数,与 Dong 等人[12]的设置相同,最大的扰动 $\varepsilon=16$,迭代次数 $T=10$,步长 $\alpha=1.6$。对于 MI-FGSM,采用默认的衰减参数 $\mu=1.0$。对于 DI-FGSM,变换概率 $p=0.5$。对于 TIM,采用高斯核,内核大小设置为 7×7。

4.2 均匀采样 OR 指数采样

在计算 Integrated Loss 的过程中,对于同一个样本计算其 IL 损失,可以分为两种方式:均匀采样和指数采样。均匀采样指的是将样本在全为零的黑色的图片到该样本空间等比例的进行样本的损失计算然后集成。指数采样指的是将样本在全为零的黑色的图片到该样本空间进行间隔指数比例的样本损失计算然后集成。

将 IL-FGSM 分别在这两种采样策略下进行了比较。具体的,使用 IL-FGSM 在这两种策略下,设置采样次数都为 5,攻击常规训练的模型,这些模型包括(Inc-v3,Inc-v4,IncRes-v2 和 Res-101)。如表 1 所示,可以看出均匀采样和指数采样几乎具有相同的攻击用时,但是指数采样的攻击成功率平均要比均匀采样高出 5% ~10%。因为均匀采样,每次采样间隔的距离较近,导致了大量相似样本的计算,所以攻击成功率偏低。因此,综合攻击用时和攻击成功率,选择指数采样作为计算 Integrated Loss 的一种采样策略。

表 1 单个模型设置下,使用均匀采样和指数采样攻击七个模型的攻击成功率(*表示白盒攻击) %

Model	Sampling strategy	*Inc-v3	*Inc-v4	*IncRes-v2	*Res-101	Inc-v3-ens3	Inc-v3-ens4	IncRes-v2-ens	Attack Time
Inc-v3	Uniform	99.3	65.9	63.0	53.9	21.4	19.0	9.4	344.8
	Exponential	100.0	75.7	72.1	67.1	30.4	31.0	15.2	344.6
Inc-v4	Uniform	75.8	99.8	64.1	56.2	25.1	22.6	11.5	625.1
	Exponential	82.7	99.8	74.1	68.9	41.0	36.4	22.0	630.9
IncRes-v2	Uniform	72.4	66.5	99.0	56.6	28.3	23.2	19.1	717.9
	Exponential	82.7	78.0	99.5	70.9	46.7	38.8	34.0	732.3
Res-101	Uniform	70.9	65.1	62.3	98.7	30.0	25.1	16.2	591.4
	Exponential	78.4	71.8	69.7	99.3	39.9	36.4	22.8	593.7

4.3　采样次数

合理的采样次数可以提供更好的梯度方向并且使计算变得更加高效。因此,研究了不同采样次数 s 的影响。使用 Inc-v3 生成对抗样本攻击 Inc-v3,Inc-v4,IncRes-v2,Res-101,使用的采样次数从 1 到 8。图 1 显示了不同采样次数的攻击成功率,可以看出攻击成功率随着采样次数的增加而不断改善,在采样次数为 5 时,所有模型都获得了较高的成功率,5 之后攻击成功率上升的相对平缓,所以综合计算消耗和攻击成功率的影响,选择采样次数为 5。

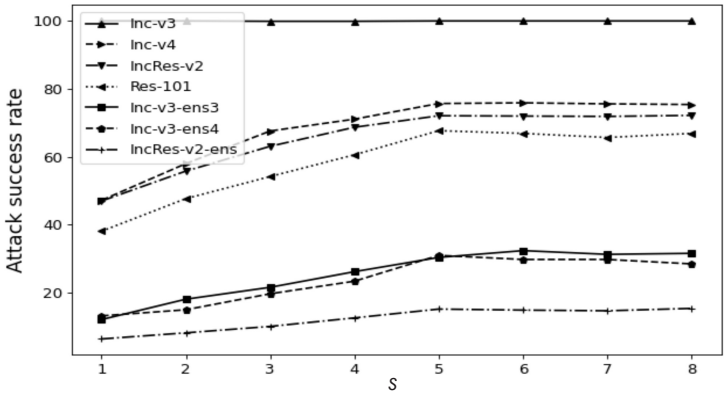

图 1　不同采样次数的攻击成功率

4.4　攻击单个模型

在这个部分,将 IL-FGSM 和其他的黑盒攻击方法(I-FGSM,MI-FGSM,NI-FGSM)进行比较,攻击单个模型。如表 2 所示,文中方法改善了所有的基线方法的攻击成功率。一般的,IL-FGSM 和其他基线方法一样都具有几乎 100% 的白盒攻击成功率,在黑盒攻击上,文中方法超过了基线攻击 10% ~20%。表明这些高级的对抗训练的模型在黑盒攻击 IL-FGSM 攻击下只是提供了微弱的防护。同样的可以观看到,使用的白盒模型的结构越复杂,生成的对抗样本的迁移性越好。

表 2　单个模型设置下,攻击七个模型的攻击成功率(*表示白盒攻击)　%

Model	Attack	*Inc-v3	*Inc-v4	*Inc Res-v2	*Res-101	Inc-v3-ens3	Inc-v3-ens4	Inc Res-v2-ens
Inc-v3	I-FGSM	100.0	20.8	16.3	14.4	7.1	7.1	3.6
	MI-FGSM	100.0	41.7	38.4	33.9	12.7	12.2	6.9
	NI-FGSM	100.0	48.2	47.4	38.3	12.9	12.2	6.2
	IL_FGSM	100.0	75.7	72.1	67.1	30.4	31.0	15.2
Inc-v4	I-FGSM	31.9	99.5	20.1	18.2	7.6	7.4	4.5
	MI-FGSM	54.6	99.6	44.1	40.6	17.4	14.9	8.1
	NI-FGSM	61.4	99.7	50.2	44.0	16.3	14.1	7.2
	IL_FGSM	82.7	99.8	74.1	68.9	41.0	36.4	22.0
IncRes-v2	I-FGSM	28.1	22.6	97.9	18.6	7.8	8.2	5.1
	MI-FGSM	56.3	48.0	97.6	40.2	19.9	17.1	12.1
	NI-FGSM	60.1	49.2	98.7	40.4	16.3	15.2	10.7
	IL_FGSM	82.7	78.0	99.5	70.9	46.7	38.8	34.0
Res-101	I-FGSM	26.6	22.5	20.2	98.1	9.4	8.6	5.6
	MI-FGSM	53.5	48.9	46.1	98.2	22.2	19.2	11.7
	NI-FGSM	61.5	56.5	54.1	98.6	25.1	19.8	10.6
	IL_FGSM	78.4	71.8	69.7	99.3	39.9	36.4	22.8

4.5　攻击集成模型

集成方法在研究中被广泛地采用来增加模型的表现和鲁棒性。集成的思想同样也可以用到对抗攻击上,因为如果一个对抗样本对于多个模型都是对抗的,那么它很有可能捕获到了内在的对抗方向,并且更容易在同一时间迁移到其他模型上,从而进一步提升黑盒的攻击成功率。目前,有三个常用的集成策略:logits 集成、预测集成、损失集成,其中 logits 集成被认

为是有效的集成策略。

考虑用 IL-FGSM 的集成模型算法 ENS-IL-FGSM 同时攻击多个模型在 logits 上的集成。具体的,使用(FGSM,MI-FGSM,NI-FGSM,ENS-IL-FGSM)攻击常规训练的模型集合,这些模型包括(Inc-v3,Inc-v4,IncRes-v2 和 Res-101),并将它们的权重设置为相等的。

如表3所示,与攻击单个模型相比,攻击集成的模型,在保持白盒攻击成功率的情况下,明显改善了黑盒攻击的成功率,并且文中方法在保持较高的白盒攻击的同时,在黑盒攻击上超过了基线攻击 10% ~20%。表明这些高级的对抗训练的模型在黑盒攻击 ENS-IL-FGSM 的攻击下只是提供了微弱的防护。

为了进一步提升 IL-FGSM 的黑盒攻击成功率,将改善样本迁移性的方法 DIM 和 TIM 集成到文中方法中。具体的,将它们与 FGSM, MI-FGSM, NI-FGSM,IL-FGSM 进行了集成,并进行了进一步的比较。如表4所示,IL-FGSM 集成了 DIM 和 TIM,在保持较高白盒攻击成功率的同时,在黑盒攻击上达到了70% ~85%的攻击成功率,这个效果堪比白盒攻击。

表3 集成模型设置下,攻击七个模型的攻击成功率(*表示白盒攻击) %

Attack	*Inc-v3	*Inc-v4	*IncRes-v2	*Res-101	Inc-v3-ens3	Inc-v3-ens4	IncRes-v2-ens
I-FGSM	98.0	97.6	96.9	97.9	17.4	14.5	9.7
MI-FGSM	98.5	97.7	97.1	97.8	35.9	31.3	20.6
NI-FGSM	98.8	98.6	98.4	98.6	36.3	31.1	19.7
ENS-IL-FGSM	99.0	98.8	98.8	98.6	43.8	39.6	26.7

表4 集成模型设置下,集成 DIM 和 TIM 方法,攻击七个模型的攻击成功率(*表示白盒攻击) %

Attack	*Inc-v3	*Inc-v4	*IncRes-v2	*Res-101	Inc-v3-ens3	Inc-v3-ens4	IncRes-v2-ens
I-FGSM(DIM&TIM)	97.1	96.7	95.1	96.3	44.2	42.5	33.5
MI-FGSM(DIM&TIM)	97.3	96.7	95.5	96.2	71.0	66.6	58.6
NI-FGSM(DIM&TIM)	98.1	97.7	97.4	97.5	65.6	61.1	50.5
IL-FGSM(DIM&TIM)	98.8	98.9	98.5	98.6	84.3	80.8	73.8

4.6 攻击用时分析

衡量一个对抗样本生成算法的好坏,不仅要考虑这个算法的攻击成功率,还要考虑这个算法的时间复杂度,也就是攻击所用时间。在这个部分,将 IL-FGSM 和其他的黑盒攻击方法(I-FGSM,MI-FGSM,NI-FGSM)在攻击用时方面进行比较。如表5所示,IL-FGSM 在攻击用时方面比其他基线方法平均多出4倍的用时。因为 IL-FGSM 使用 Integrated Loss 作为它的损失函数,相比于其他基线方法,对于每一个样本只计算了一次损失,IL-FGSM 计算了这个样本的线性比例损失的集成,所以相对的用时较多。

表5 单个模型设置下,四个方法的攻击用时 s

Attack	Inc-v3	Inc-v4	IncRes-v2	Res-101
I-FGSM	161.9	95.1	175.0	157.6
MI-FGSM	99.7	163.9	186.5	154.7
NI-FGSM	93.0	150.0	168.4	141.4
IL-FGSM	341.2	632.7	710.2	609.0

5 结束语

对抗样本的存在严重威胁到深度学习在众多安全领域的运用,因此对抗样本生成算法也成了当下研究的热点。该文提出了一种新的方法,积分损失快速梯度符号法 IL-FGSM。该方法利用了输入的间隔指数范围采样的损失函数的集成来更新对抗样本,与单个损失函数的更新相比,使用了当前输入样本的积分损失来更新当前的对抗样本,这可以更好地避免梯度出现饱和的情况,引导梯度在全局最优的方向上更新。实验结果表明,IL-FGSM 不仅提高了白盒攻击的成功率也提高了黑盒攻击的成功率。除此之外,将增加样本迁移性的方法 DIM 和 TIM 集成到了该方法,进一步提升了黑盒攻击的成功率。实验显示 IL-FGSM 提升了基线攻击 10% ~20% 的攻击成功率。与基线方法相比,该方法生成了更强的对抗样本,但是计算消耗较大。未来,将做进一步的研究来加快积分损失的计算。

参考文献:

[1] 张　顺,龚怡宏,王进军. 深度卷积神经网络的发展及其在计算机视觉领域的应用[J]. 计算机学报,2019,42(3):453-482.

[2] 高君宇,杨小汕,张天柱,等. 基于深度学习的鲁棒性视觉跟踪方法[J]. 计算机学报,2016,39(7):1419-1434.

[3] 侯一民,周慧琼,王政一. 深度学习在语音识别中的研究进展综述[J]. 计算机应用研究,2017,34(8):2241-2246.

[4] 奚雪峰,周国栋. 面向自然语言处理的深度学习研究[J]. 自动化学报,2016,42(10):1445-1465.

[5] 苏志达,祝跃飞,刘　龙. 基于深度学习的安卓恶意应用检测[J]. 计算机应用,2017,37(6):1650-1656.

[6] SZEGEDY C,ZAREMBA W,SUTSKEVER I,et al. Intriguing properties of neural networks[J]. arXiv: 13126199, 2013.

[7] 邹国锋,傅桂霞,李海涛,等. 多姿态人脸识别综述[J]. 模式识别与人工智能,2015,28(7):613-625.

[8] 张新钰,高洪波,赵建辉,等. 基于深度学习的自动驾驶技术综述[J]. 清华大学学报:自然科学版,2018,58(4):438-444.

[9] 王科俊,赵彦东,邢向磊. 深度学习在无人驾驶汽车领域应用的研究进展[J]. 智能系统学报,2018,13(1):55-69.

[10] GOODFELLOW I J, SHLENS J, SZEGEDY C. Explaining and harnessing adversarial examples[J]. arXiv: 14126572, 2014.

[11] KURAKIN A, GOODFELLOW I, BENGIO S. Adversarial examples in the physical world[J]. arXiv:160702533,2016.

[12] DONG Y, LIAO F, PANG T, et al. Boosting adversarial attacks with momentum[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. Salt Lake City:IEEE,2018:91859193.

[13] LIN J, SONG C, HE K, et al. Nesterov accelerated gradient and scale invariance for adversarial attacks[J]. arXiv: 190806281,2019.

[14] SUNDARARAJAN M, TALY A, YAN Q. Gradients of counterfactuals[J]. arXiv:161102639,2016.

[15] MADRY A, MAKELOV A, SCHMIDT L, et al. Towards deep learning models resistant to adversarial attacks[J]. arXiv:170606083,2017.

[16] XIE C, ZHANG Z, ZHOU Y, et al. Improving transferability of adversarial examples with input diversity[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. Long Beach: IEEE, 2019: 2730-2739.

[17] DONG Y, PANG T, SU H, et al. Evading defenses to transferable adversarial examples by translation-invariant attacks[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. Long Beach: IEEE, 2019: 4312-4321.

[18] CARLINI N, WAGNER D. Towards evaluating the robustness of neural networks[C]//2017 IEEE symposium on security and privacy (sp). San Jose: IEEE, 2017: 39-57.

[19] SZEGEDY C, VANHOUCKE V, IOFFE S, et al. Rethinking the inception architecture for computer vision[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. Las Vegas: IEEE, 2016: 2818-2826.

[20] SZEGEDY C, IOFFE S, VANHOUCKE V, et al. Inception-v4, inception-resnet and the impact of residual connections on learning[C]//Proceedings of the AAAI conference on artificial intelligence. San Francisco: AAAI, 2017: 4278-4284.

[21] HE K, ZHANG X, REN S, et al. Identity mappings in deep residual networks[C]//Proceedings of the European conference on computer vision. Amsterdam: Springer, 2016: 630-645.

[22] TRAMÈR F, KURAKIN A, PAPERNOT N, et al. Ensemble adversarial training: attacks and defenses[J]. arXiv: 170507204,2017.