面向实际场景的人工智能脆弱性分析

田 鹏[1,2],左大义[1,2],高艳春[1,2],陈海兵[1,2],丁 灏[1,2]
(1. 中国电子科技集团第三十研究所,四川 成都 610000;
2. 中电科网络空间安全研究院有限公司,北京 100191)

摘 要:人工智能技术广泛应用于自动驾驶、无人机、机器人等自主无人系统,是实现场景感知、情报获取、辅助决策等复杂功能的重要支撑。因此,研究人工智能技术的脆弱性和本身安全性问题引起了越来越多的关注。对抗机器学习(adversarial machine learning)是机器学习和计算机安全领域的交叉学科,是人工智能算法普遍面临的挑战之一。文中以实际场景下的人工智能安全性为出发点,梳理了对抗样本发展的起源,形成的机理以及发展脉络。首先从攻击、防御两个方面探究各种方法的原理和优缺点;其次,在分析研究经典算法和适用场景的基础上,研究了在实际场景下智能技术面临的脆弱性和挑战;最后,针对在图像、语音、网络和软件应用等不同领域中所面临的挑战和未来发展趋势做了进一步的分析和展望。

关键词:人工智能安全;安全威胁;深度学习;对抗样本;对抗检测
中图分类号:TP309 **文献标识码**:A **文章编号**:1673-629X(2021)11-0129-07
doi:10.3969/j.issn.1673-629X.2021.11.021

Vulnerability Analysis of Artificial Intelligence in Real World

TIAN Peng[1,2],ZUO Da-yi[1,2],GAO Yan-chun[1,2],CHEN Hai-bing[1,2],DING Hao[1,2]
(1. China Electronics Technology Group Corporation 30,Chengdu 610000,China;
2. China Electronics Technology Research Institute of Cyberspace Security Co.,Ltd.,Beijing 100191,China)

Abstract:Artificial intelligence is widely used in autonomous unmanned systems such as autonomous driving,unmanned aerial vehicles,robots,etc,which is an important support to realize complex functions such as scene perception,intelligence acquisition,assistant decision-making and so on. Therefore,more and more attention has been paid for the vulnerability and security of artificial intelligence technology. Adversarial machine learning is an interdisciplinary subject in the field of machine learning and computer security. It is one of the challenges that artificial intelligence algorithms are facing. On the basis of the security of artificial intelligence in the actual scene,we introduce the origin,principle and development of generating adversarial examples. Firstly,we explore the principles,advantages and disadvantages of attack and defense. Secondly,based on the analysis of classic algorithms and real world,the vulnerability and challenges faced in the actual scene are studied. Finally,we make a further study on the challenges and future development trend in different fields such as image,voice,network and software application.

Key words:artificial intelligence security;security thread;deep learning;adversarial example;adversarial detecting

0 引 言

人工智能技术正在迅速应用于网络空间安全、自动驾驶等关键领域,而人工智能中的一系列安全问题,并没有得到解决,不安全的人工智能技术的冒进应用,必然会带来一系列新型安全隐患。对人工智能技术中出现的安全问题的研究,形成了所谓的对抗性机器学习研究领域,其中对抗样本是一个研究的热点[1]。

对抗性机器学习技术源于 2013 年,Szegedy[2] 发现:通过对样本添加极微小、经过计算的扰动,可以使深度学习分类器得到完全不同的结果,从而提出了“对抗样本”的概念,这标志着对抗性机器学习的正式诞生。从谷歌的统计数据上[3]来看,自 2013 年提出抗性机器学习直到 2016 年底,对抗性机器学习在学术界内保持了一般的研究热度,但从 2017 年发现了用于现实世界的对抗样本,以及 OpenAI 等机构的宣传,学术界和行业界对对抗样本、对抗性机器学习的兴趣显著提升。

对抗样本是一类被恶意设计来攻击机器学习模型

收稿日期:2020-11-10 **修回日期**:2021-03-15
基金项目:国家自然科学基金青年科学基金(61803352)
作者简介:田 鹏(1981-),男,博士,高级工程师,研究方向为网络安全、人工智能。

的样本,是攻击者故意设计的,它们与真实样本的区别几乎无法用肉眼分辨,但是却会导致模型进行错误的判断。就像是让机器在视觉上产生幻觉一样。例如文献[4]中描述,在“panda”图片中,加入精心制作的微小扰动,即可使神经网络模型判断错误,以 99.3% 的高置信度识别为“gibbon”长臂猿,如图 1 所示。

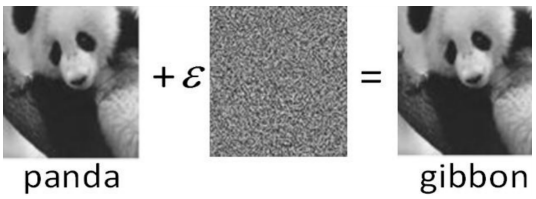

图 1 对抗样本的形成

对抗样本的形成原理,Goodfollow[4] 认为是在高维空间中,模型中存在线性化部分,而非线性化;另一个解释是认为对抗样本不是数据的主要部分,即不在数据流行内。有研究人员认为[5],内部矩阵中较大的奇异值会让分类器在面临输入中的小波动时变得更加脆弱。另外有研究认为对抗样本这种现象本质上是由数据流形的高维度造成的。

1 对抗样本攻击与防御

1.1 对抗攻击

1.1.1 白盒攻击

白盒攻击,攻击者掌握深度学习网络模型结构、模型参数等详细信息。早期研究针对特定目标的对抗样本提出了经典的基于约束的 L-BFGS 白盒算法。在随后的生成机制研究中,除了增加分类损失的方向上采取单步来干扰样本的思路外,还可以基于迭代,在每个步骤之后调整方向的同时采取多个小步骤,由此提出了一系列的经典迭代算法,四种较为主流的白盒攻击方法分别是 L-BFGS、FGSM、I-FGSM、C&W。

L-BFGS 攻击,Szegedy[6] 等人根据神经网络损失函数,构建使模型做出误分类的最小扰动模型,通过方程求解的方式得到最优攻击。但由于方程求解复杂度过高,在求解过程中通过寻找最小损失函数正则项,将原问题进行简化,利用 L-BFGS 对问题进行凸优化,具体为:

$$\min c\mid r\mid + \mathrm{loss}_f(x+r,l) \tag{1}$$

其中,$\mathrm{loss}_f(x+r,l)$ 是神经网络对应的损失函数;l 是错误分类的类别标记;c 是惩罚参数。

快速梯度符号法(FGSM)是最简单最广泛的非目标对抗攻击方法之一。基本思想是通过迭代优化的思路寻找对抗样本[4]。给定一个原始图像 x,以及一个目标分类器损失函数 $\mathrm{loss}(x,l_x)$,FGSM 的目标是在 x 的 l_x 正无穷邻域中寻找一个类似的图像 x 来欺骗分类器,将 x 分类为标签 l_x。然后将问题转化为最大化 $\mathrm{loss}(x,l_x)$,该损失是将图像 x 分类为标签 l_x 的成本,同时保持扰动较小。FGSM 通过在 ε 的图像空间中从 x 进行一步梯度更新来解决该优化问题。更新步长 ε 对于每个像素是相同的,并且更新方向由该像素处的梯度信号确定。这里模型损失函数的梯度方向为 $g=\mathrm{sign}(\nabla x' J(\theta,X,Y))$,其中设定步长 ε 实现损失函数最大化。生成对抗样本的过程为 $x_{\mathrm{adv}}=x+\varepsilon g$,其中 J 是每个样本的损失函数,$f(x,\theta)$ 是输入 x 时模型预测的输出,θ 是模型参数,Y 是正确分类。

基本迭代方法(I-FGSM)是在 FGSM 基础上进行优化的方法[7],通过扩展 FGSM 将单步的扰动变为多次迭代的扰动,迭代公式为:

$$X_{N+1}^{\mathrm{adv}}=\mathrm{ClipX},\varepsilon\{X_N^{\mathrm{adv}}+\alpha\mathrm{sign}[\nabla_X J(X_N^{\mathrm{adv}},y_{\mathrm{true}})]\} \tag{2}$$

依据 I-FGSM 迭代的思想,每次按照梯度方向移动的较小步长 $\alpha=\frac{\varepsilon}{T}$,通过 T 次累加,生成最终的扰动值。具体来说,初始 $x_0=x$,重复 FGSM 中的过程,计算损失函数的梯度方向 $g=\mathrm{sign}(\nabla x' J(\theta,X,Y))$,每次迭代生成的对抗样本为 $x_{i+1}=x_i+\alpha g$。第 T 次迭代后,x_T 就是最终生成的对抗样本。相比于 FGSM 向着梯度的方向移动一步的距离,该算法将一步切分为很多小步,逐渐优化,实现更优攻击效果。

C&W 攻击方法是目前白盒攻击中效果最好的靶向攻击方法,可以攻破防御性蒸馏等神经网络防御方法[8]。C&W 算法的损失函数中包含两部分,一是对抗样本 x 与原始输入 x 之间的范数约束;另一部分为衡量模型识别对抗样本是否是目标分类的最大差异值。C&W 攻击方法将问题转化成一个优化问题,通过最小化损失函数来寻找对抗样本。并通过样本的可转移性,实现黑盒攻击。具体方法是:

$$\begin{aligned}&\min \text{ mize } \|\delta\|_p + c\cdot f(x+\delta)\\ &\text{such that } \quad x+\delta\in[0,1]^n\end{aligned} \tag{3}$$

其中,f 是 C&W 攻击方法定义的优化的目标函数,对于给定输入图像 x,攻击方法寻求较小的扰动 δ,且能够达到欺骗分类器的目的。测试平衡二者的参数。F 和 δ 的形式为:

$$\min \text{ mize } \| \frac{1}{2}(\tanh(w)+1)-x\|_2^2 + c\cdot f(\frac{1}{2}(\tanh(w)+1)) \tag{4}$$

$$\delta_i=\frac{1}{2}(\tanh(w_i)+1)-x_i \tag{5}$$

1.1.2 黑盒攻击

在实际的攻击场景下,攻击者往往很难获得相关模型的架构、训练参数和网络超参数等关键信息,只能采取黑盒攻击方式,对模型进行有限次数的样本查询,

并基于反馈信息进行设计攻击行为。

目前常见的黑盒攻击算法主要分为两类,一类是基于一定的算法结构输入,然后根据模型的反馈不断迭代修改输入,比较典型的就是单像素攻击算法和本地搜索攻击算法;另一类是基于迁移学习的思想,使用与白盒攻击类似的开源模型,之后用生成的对抗样本进行黑盒攻击。

单像素攻击(one pixel attack)是基于改变样本中的一个像素以实现对目标模型的扰动,是一种低成本的对抗攻击策略。Su 等人[9]利用差分进化算法,通过迭代修改像素值产生变种,并将变种与母样本比较,从候选像素点中逐步筛选出稀疏像素点,最后根据选择攻击效果最好的变种作为对抗样本,有效攻击分类模型,损失函数如下:

$$\begin{aligned}&\min \text{loss}_F(x_i^{'}),x^{'} \in [0,1]^n\\&\text{使得 } \| x - x^{'} \|_0 \leqslant d\end{aligned} \tag{6}$$

基于单像素和多像素的对抗样本搜索空间如图 2 所示,三维空间中,任一坐标点即为待修改坐标。即三条平面交线组成了单像素对抗样本的搜索空间,同样,三个灰色的二维平面组成了两像素对抗样本的搜索空间,对抗样本的生成过程转化为对应空间的搜索过程。

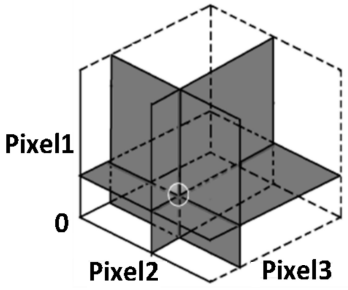

图 2　单像素攻击模型

通常的对抗样本生成方法允许扰动所有的像素,然后对所有像素的变化量之和进行整体约束来构造目标函数,而这里所考虑的攻击方法是相反的,其只关注于被修改像素的数量,但不限制单个变化量的大小。

通用对抗攻击(UAP)算法:单像素攻击、local search attack 等方法生成的对抗扰动只对某个特定的图像有效[9],而通用扰动(UAP)攻击生成的对抗扰动叠加在任何图片上均能使分类器出错,并且这些扰动对人类而言是不可见的[10]。UAP 攻击的主要思想是通过分析对抗性质将图像逐步偏离分类边界。该方法对于 ResNet 网络效果较好,并且这种扰动可以泛化到其他网络上。UAP 攻击的主要模型为:

$$P_{I_c \sim \varphi_c}(c(I_c) \neq c(I_c + \rho)) \geqslant \delta \tag{7}$$

其中,$\rho(\cdot)$ 表示概率;$\delta \in (0,1]$ 为扰动率。

基于迁移算法的黑盒攻击:可迁移能力是对抗样本的重要属性,也是对抗样本研究面对的另一个重要的理论问题。近期研究发现,对抗样本不仅在不同网络结构间存在可迁移能力,在不同算法、分类类别及数据集之间也存在迁移能力。利用可迁移能力能够在不具备对目标模型完备知识的前提下构建具有足够欺骗能力的对抗样本,从而构成黑盒攻击的基础。利用类别间的可迁移能力,能够大幅提升构建对抗样本的效率。文献[11]分别从迁移性和失真度两个标准来衡量对抗样本的迁移攻击能力。

对抗样本转移性衡量:通过计算一个模型生成的对抗样本能被另一个模型正确分类的准确率,来衡量非目标攻击的迁移性,且和非目标攻击的迁移性呈反比。反之,针对特定目标攻击,则以匹配度衡量迁移性,且呈正比表示,表 1 为不同模型之间的对抗样本的迁移性表示。

表 1　迁移准确率

	RMSD	ResNet-152	ResNet-101	ResNet-50	VGG-16	GoogLeNet
ResNet-152	22.83	0	0.13	0.18	0.19	0.11
ResNet-101	23.81	0.19	0	0.21	0.21	0.12
ResNet-50	22.86	0.23	0.2	0	0.21	0.18
VGG-16	22.51	0.22	0.17	0.17	0	0.05
GoogLeNet	22.58	0.39	0.38	0.19	0.19	0

对抗样本失真度衡量:除了可转移性之外,另一个重要因素是对抗图像与原始图像之间的扭曲度。失真度可以通过下面的公式进行计算。

$$d(x^*,x) = \sqrt{\sum_i (x_i^* - x_i)^2 / N} \tag{8}$$

其中,x^* 和 x 是对抗图像和原始图像的向量表示,N 是 x^* 和 x 的维数,x_i 是 x 在第 i 维度上的像素值(0 ~ 255)。

在进一步分析对抗图像样本的迁移能力和失真度的基础上,比较对抗样本在不同数据集、不同模型之间的迁移能力,为迁移性更强的黑盒对抗攻击方法构建及其应用奠定基础。

1.2　对抗防御

在对抗性攻击研究的同时,如何使模型更具有鲁棒性、更好地进行防护,得到了广泛关注。对抗样本的防御可分为两类:

(1)对网络本身结构进行更改,例如,针对模型中

相关函数以及网络结构本身进行变更,形成防御,如 Papernot[12] 提出的网络蒸馏法。

(2)在对抗的基础上,采用更改过的对抗样本进行再训练,在样本的输入阶段,弥补样本的多样性缺陷,使模型更加鲁棒的训练[13],如 Goodfellow 提出的对抗性训练法等。

防御蒸馏法:蒸馏是一种将复杂网络模型转化为简单网络模型的技术[14]。由于防御蒸馏技术应用了图像梯度,因此也可以看成是一种基于梯度掩模的方法。2016 年 Papernot 等人基于知识蒸馏设计了一种提高网络鲁棒性能的方法[12]。就是在进行模型训练时,使用一些平滑处理的方法,将模型梯度中陡峭的地方平滑掉,使得模型的分类输出对于输入数据的一定扰动不那么敏感。这样就可以降低模型对于对抗样本中的正常图像上增加的噪声扰动的脆弱性,从而使得训练得到的模型具备一定的对抗鲁棒能力,这种知识通过输入向量的分类概率提取,并且反馈训练原始的模型。实验表明这种方法增加了网络对于微小扰动的鲁棒性,如图 3 所示。

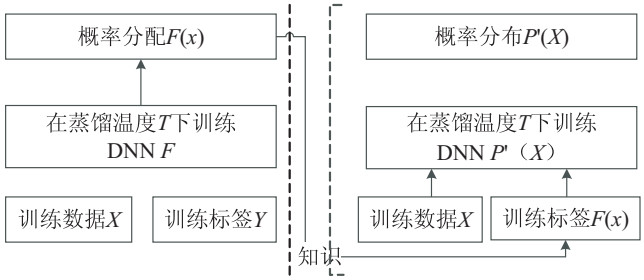

图 3 蒸馏法原理

对抗训练法:对抗训练是一种蛮力训练方法,通过对训练集添加预先构造的对抗样本,提升模型针对对抗样本的稳健性。该方法是将对抗样本与正常样本合并形成一个新的数据集[14],通过重训练,重新学习添加扰动的损失函数 $\tilde{J}(\theta,x,y)=\alpha J(\theta,x,y)+(1-\alpha)J(\theta,x+\varepsilon\text{sign}(\nabla_x J(\theta,x,y)),y)$,达到提高模型鲁棒性的目的。通过对抗训练过程的最小最大优化问题,在训练过程中加入对抗样本生成方法所生成的干扰数据,并且训练目标设定为最小化所训练出的神经网络模型的错误分类概率。这样训练得到的模型,能够在面对最强大的对抗样本生成方法的时候,仍然保持最小化错误分类概率,由此获得一定的对抗鲁棒性能力。

2 实际场景下的攻击方法分析

2.1 物理场景攻击

在实际场景中,往往攻击者无法完全控制输入模型的数据,对于模型如何预处理原始数据也一无所知。只能通过摄像头、麦克风这类物理设备,经过一系列黑盒的预处理后才能真正进入模型[15]。文中从实际场景出发,基于预处理阶段和实际应用场景,分析对抗样本攻击的方式和方法:

2.1.1 预处理阶段攻击

根据不同的应用场景,模型系统可能会采集到各种规格的输入图像,相比于实际的采集图像,经典模型在训练过程中的图像通常较小且固定,例如 Inception-v3:299×299, VGGNet:224×224, AlexNet:224×224, GoogleNet:224×224, ResNet:224×224。固定的尺寸可以确保训练和预测的效率,因此,图像采集后,往往会有缩放步骤[16],对数据进行归一化操作。它们被广泛用于深度学习框架(Tensorflow,Caffe,Pytorch)中。

在数据采集缩放过程中,实际采集和模型中图像的不匹配生成对抗样本,实现对模型的下采样攻击。如图 4 所示,即在对抗样本相似性约束的前提下,通过

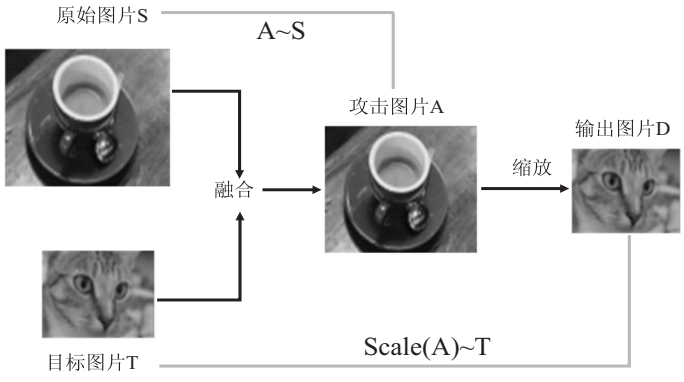

图 4 预处理攻击

将属于目标类并具有网络规定输入大小的小图像(Target Image T)嵌入到原始图像(Source Image S)中来构造缩放攻击图像(Attack Image A)[17]。

2.1.2　子块(Patch)攻击

之前的大部分工作都集中在对输入的微小或难以察觉的变化进行攻击和防御上,而在实际场景下,精确地给一个目标,定制一个对抗扰动是不切实际的。针对此缺陷,恶意攻击者会将一个与图像无关的补丁添加到输入图像中,即使能够注意此补丁,也不能理解其意图[18],或将其忽略,如图 5 中路牌的小广告、涂鸦等形式。在实际路标检测中,需要几个小小的标签,就能让 YOLOv2[15]无法检测出路标。而这些小标签能伪装成涂鸦艺术之类的东西融入到路标图像中,让人们难以察觉,即使是发现了也往往不会在意。

图 5　patch 攻击

现有的攻击策略还远远不能生成具有较强攻击能力的视觉天然斑块,为进一步提高 patch 的攻击能力,引入图像上下文的相关性和视觉注意力机制进行子块的优化处理。基于此,文献[19]提出了一种感知敏感生成对抗网络(PS-GAN),该方法提出了一种用于生成敌对 patch 的感知敏感 GAN (PSGAN)。PS-GAN 利用被攻击网络的感知敏感性,保证生成的敌对 patch 具有自然的外观,并在对抗性生成过程中耦合注意机制,保证生成的对抗性补丁具有较强的攻击能力。

2.2　对抗检测

实际场景下,数据输入的形式往往多种多样,针对此问题,单纯针对研究模型结构来克服对抗样本的干扰比较困难。基于此,基于各类样本在空间分布的差异性的检测技术引起了广泛关注[20],现有的检测手段可分为两大类,基于度量的方法与基于预测不一致的方法。

(1)基于度量的方法,对输入(和激活值)进行统计测量以检测对抗样本。这些技术的关键挑战是如何定义高质量的统计指标,使该指标可以清楚地分辨正常样本和对抗样本之间的差异。

(2)基于预测不一致的方法,许多其他工作都基于预测不一致的方法,即对抗样本具有扰动,利用其他检测手段与原输出进行比较,一致为正常样本,不一致则为对抗样本。

3　对抗样本应用领域

人工智能技术已经渗透到各个领域,以由初期的图像对抗样本领域拓展到当前阶段的针对音频、文本、生物、二进制应用和网络流量等各类数据的对抗样本:

3.1　图像领域

无人车:自动驾驶引起了业界和学术界越来越多的关注。在自动驾驶中,分类模型被广泛使用并部署在定位和感知模块中,其关键是卷积神经网络(CNN),它根据摄像机和激光雷达的传感器输出做出实时决策,为回归模型提供精细粒度的上下文信息。研究表明,CNN 易受对抗性攻击,相应地通用物体识别系统也易受对抗性攻击,但目前通用主流物体识别模型的脆弱点对于自动驾驶中物体识别系统的适用性未知。近两年优步和谷歌的自动驾驶汽车相继发生事故,主要原因可能是物体识别模块对于特定情况下(如行驶速度、天气、背景环境等复杂环境和道路条件)的识别任务不准确导致的[21]。同时,系统中回归模型强依赖于分类模型,使得分类模型结果的简单变化很容易影响回归模型,从而造成不安全危害。这些因素使得自动驾驶中物体识别系统面临的安全威胁更特殊、更复杂,同时后果更严重。

路标识别攻击:最近一项研究表明[17],只要在路上贴上几个不起眼的小贴纸,智能汽车或许就无法识别出这些路标了。研究人员对路标进行了有目标指向的全局扰动,然后将其以海报的形式全尺寸打印了出来,覆盖在原来的 STOP 路标上。在测试中,视觉感知系统从不同的距离和角度,对这个对抗样本进行识别,结果在大多数情况下,其将 STOP 路标识别为了限速标志,如图 6 所示。

图 6　路牌识别攻击

3.2　语音领域

语音领域的对抗攻击可分为两种类型:Speech-to-Label、Speech-to-Text[21]。Speech-to-Label 是指通过构造对抗样本,可以让音频识别系统将该样本分类为

任意指定的标签;这个类别的攻击和基于图像的对抗样本攻击很类似[21]。然而,由于标签的种类是有限的,因此这种类别的攻击有很大的局限性。Speech-to-Text 则是通过构造对抗样本,可以让音频识别系统将该样本转录为任意指定的字符串。

常规语音领域对抗攻击的构造如图 7 所示,其中 x 是输入的原始音频向量,δ 是在原始音频向量上添加的扰动。音频对抗样本攻击就是通过向原始音频向量 x 添加一些扰动 δ,使得语音识别系统 ASR 可以将构造的新样本 $x+\delta$ 识别为攻击者指定的文本 t,但人耳并不能分辨出新样本和原始音频的区别[21]。这个过程可以表示为 $f(x+\delta)=t$。构造音频对抗样本的过程就是通过计算损失函数 $\ell(\cdot)$ 的梯度然后不断更新 x 的过程,直到构造的对抗样本满足终止条件,其中常用的损失函数表示为:

$$l(x,\delta,t)=l\mathrm{model}(f(x+\delta),t)+c\cdot \mathrm{lmetric}(x,x+\delta) \tag{9}$$

其中,l model 是语音识别模型本身的损失函数。

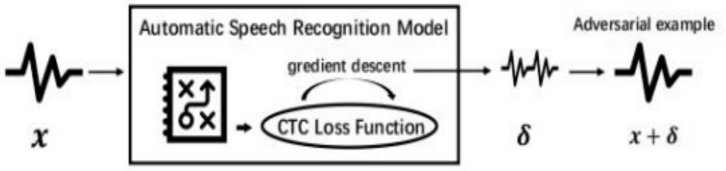

图 7 语音对抗样本生成

目前还没有方法可以在分钟级时间内构造出低噪声、高鲁棒性的音频对抗样本,生成过程要兼顾音频样本质量、鲁棒性和生成速度之间的权衡。

3.3 网络领域

人工智能技术赋能网络空间安全领域的关键因素是带来安全增益的同时,必须具有高安全性能。在网络空间领域,已发表了一系列针对现有安全相关算法造成挑战的成果:入侵检测(IDS)算法、恶意软件检测算法、动态生成域名(DGA)检测算法、恶意流量检测算法[22]等,均出现了对抗性机器学习算法。

NDSS 会议上,Kitsune 技术方案[22]作为基于深度学习的网络入侵检测的典型例子引起了广泛关注。Kitsune 异常检测功能的核心是自编码学习网络。针对 Kitsune 进行白盒攻击研究表明,基于机器学习的 Kitsune 入侵检测技术,面对经典对抗样本攻击算法 FGSM、JSMA、C&W、ENW 等表现非常脆弱,其误报率、漏报率都达到 100%,即能够产生正常网络流量使 Kitsune 识别为异常流量,以及产生异常流量使得 Kitsune 识别为正常流量。

3.4 软件应用领域

当前移动设备已广泛使用,例如手机移动端,许多都运行着 android 系统,基于 android 系统具有开放、共享等特点,快速形成了以 android 系统为基础的软件生态系统,而针对移动端软件系统的恶意攻击呈上升态势[22],而将对抗样本思想融入恶意代码检测是新型的前沿方向。Xu[23]提出利用遗传算法将干扰信息注入恶意样本,模型将其错误识别为正确样本,成功绕过检测模型。Kolosnjaji 等[24]在保持恶意样本的功能的同时,针对恶意样本数据中特定字段的修改,成功躲避恶意软件检测系统。

4 结束语

人工智能在图像识别、语音识别、自然语言处理、网络安全等领域均取得了跨越式的发展和广泛应用。而对抗性机器学习是人工智能技术实践过程中的极大威胁。如何确保人工智能技术安全、可靠、可控发展的同时,最大限度降低、规避智能应用风险是一个严峻的挑战。文中在对国内外智能安全研究调研和分析的基础上,首先从攻、防两个方面梳理了智能安全发展的脉络:攻击方面,分别从白盒攻击、黑盒攻击、全像素攻击、单像素攻击对攻击方法进行分类讨论;防御方面,从反应式和主动式分类进行了分析。同时,针对实际场景,文中分别从对抗样本产生形式和实际场景下对抗样本的检测进行了归纳总结。最后,结合基于实用场景,分别在不同领域对对抗样本的表现形式、效果影响进行了讨论。

参考文献:

[1] RIGAKI M. Adversarial deep learning against intrusion detection classifiers[D]. Luleå University of Technology, 2017.

[2] SZEGEDY C. Intriguing properties of neural networks[C]// 26th IEEE conference on computer vision and pattern recognition. Portland, Oregon, USA: IEEE, 2013.

[3] KURAKIN A, GOODFELLOW I, BENGIO S. Adversarial examples in the physical world[C]//International conference on learning representation. Palais des Congrès Neptune, Toulon: [s. n.], 2017.

[4] GOODFELLOW I, SHLENS J, SZEGEDY J. Explaining and harnessing adversarial examples[C]//International conference on learning reprsentaion. San Diego: [s. n.], 2015.

[5] MAHENDRAN A, VEDALDI A. Visualizing deep convolutional neural networks using natural pre-images[J]. International Journal of computer Vision, 2016, 120: 233-255.

[6] PIERAZZI F, PENDLEBURY F, CORTELLAZZI J. Intriguing properties of adversarial ML attacks in the problem space [C]//IEEE symposium on security & privacy. Oakland: IEEE, 2020.

[7] KURAKIN A, GOODFELLOW I, BEGIO S. Adversarial examples in the physical world [C]//IEEE conference on computer vision and pattern recognition. Honolulu, HI, USA: IEEE, 2017.

[8] CARLINI N, WAGNER D. Towards evaluating the robustness of neural networks[C]//2017 IEEE symposium on security and privacy (SP). San Jose, USA: IEEE, 2017.

[9] DEZFOOLI S M, FAWZI A, FAAWZI O, et al. Universal adversarial perturbations[C]//Proceedings of IEEE conference on computer vision and pattern recognition (CVPR). Honolulu, HI, USA: IEEE, 2017.

[10] LIU Yanpei, CHEN Xinyun. Delving into transferable adversarial examples and black-box attacks [C]//International conference on learning representations. Palais des Congres Neptune, Toulon: [s. n.], 2017.

[11] YUAN X, HE P, ZHU Q. Adversarial examples: attacks and defenses for deep learning[J]. IEEE Transactions on Neural Networks and Learning Systems, 2019, 30(9): 2805-2824.

[12] PAPERNOT N, MCDANIEL P, WU Xi, et al. Distillation as a defense to adversarial perturbations against deep neural networks[C]//2016 IEEE symposium on security and privacy (SP). San Jose, USA: IEEE, 2016: 582-597.

[13] MADRY A, MAKELOV A, SCHMIDT L, et al. Towards deep learning models resistant to adversarial attacks[C]// IEEE conference on computer vision and pattern recognition. Honolulu, HI, USA: IEEE, 2017.

[14] HE W, WEI J, CHEN X, et al. Adversarial example defenses: ensembles of weak defenses are not strong[C]//Proceedings of the 11th USENIX conference on offensive technologies. Vancouver, BC, Canada: USENIX, 2017.

[15] CARLINI N, WAGNER D. Adversarial examples are not easily detected: bypassing ten detection methods [C]//Proceedings of the 10th ACM workshop on artificial intelligence and security. Texas, USA: ACM, 2017.

[16] XIE C, WANG J, ZHANG Z. et al. Adversarial examples for semantic segmentation and object detection[C]//IEEE international conference on computer vision. Venice, Italy: IEEE, 2017.

[17] QUIRING E, KLEIN D, ARP D, et al. Adversarial preprocessing: understanding and preventing image-scaling attacks in machine learning[C]//29th USENIX security symposium. Boston, USA: USENIX, 2020.

[18] BROWN T B, MANE D, ROY A, et al. Adversarial patch [C]//31th conference on neural information processing systems. Long Beach, CA, USA: [s. n.], 2017.

[19] LIU A S, LIU X L, FAN J X, et al. Perceptual-sensitive GAN for generating adversarial patches [C]//Association for the advance of artificial intelligence. Honolulu, Hawaii, USA: [s. n.], 2019.

[20] FEINMAN R, CURTIN R R, SHINTRE S, et al. Detecting adversarial samples from artifacts [C]//34th international conference on machine learning. Sydney, Australia: [s. n.], 2017.

[21] SAHU S K, KUMAR P, SINGH A P. Dynamic routing using inter capsule routing protocol between capsules [C]//20th international conference on computer modelling and simulation. Sydney, Australia: [s. n.], 2018.

[22] MIRSKY Y, DOITSHMAN T, ELOVICI Y, et al. Kitsune: an ensemble of autoencoders for online network intrusion detection[C]//12th international conference on network and system security. Hong Kong, China: [s. n.], 2018.

[23] XU W, QI Y, EVANS D. Automatically evading classifiers [C]//Proceedings of the network and distributed security systems symposium. San Diego, USA: [s. n.], 2016.

[24] KOLOSNJAI B, DEMONTIS A, BIGGIO B, et al. Adversarial malware binaries: Evading deep learning for malware detection in executables[C]//26th European signal processing conference. Roma, Italy: [s. n.], 2018.