DOI:10.13878/j.cnki.jnuist.2019.06.003

杨弋鋆[1]　邵文泽[1]　王力谦[1]　葛琦[1]　鲍秉坤[1]　邓海松[2]　李海波[1,3]

面向智能驾驶视觉感知的对抗样本攻击与防御方法综述

摘要

现如今,深度学习已然成为机器学习领域最热门的研究方向之一,其在图像识别、目标检测、语音处理、问答系统等诸多领域都取得了巨大成功.然而通过附加经过特殊设计的细微扰动而构造出的对抗样本,能够破坏深度模型的原有性能,其存在使许多对安全性能指标具有极高要求的技术领域,特别是以视觉感知为主要技术优先的智能驾驶系统,面临新的威胁和挑战.因此,对对抗样本的生成攻击和主动防御研究,成为深度学习和计算机视觉领域极为重要的交叉性研究课题.本文首先简述了对抗样本的相关概念,在此基础上详细介绍了一系列典型的对抗样本攻击和防御算法.随后,列举了针对视觉感知系统的多个物理世界攻击实例,探讨了其对智能驾驶领域的潜在影响.最后,对对抗样本的攻击与防御研究进行了技术展望.

关键词

对抗样本;目标检测;语义分割;智能驾驶

中图分类号 TP391

文献标志码 A

收稿日期 2019-10-10

资助项目 国家自然科学基金(61771250,61602257,61972213,11901299,61872424,61930003088)

作者简介

杨弋鋆,男,硕士生,研究方向为计算机视觉对抗样本攻击与防御.1018010626@njupt.edu.cn

1　南京邮电大学　通信与信息工程学院,南京,210003

2　南京审计大学　统计与数学学院,南京,211815

3　瑞典皇家理工学院　计算机科学与通信学院,斯德哥尔摩,10044

0　引言

得益于深度学习[1]技术的巨大突破以及计算机性能的快速提高,人工智能相关研究被提到了一个前所未有的高度.大量的新技术如语言翻译、人脸识别、图像生成、场景检测等迅速出现并被广泛应用.

在深度学习领域,研究者们不断追求着更快的速度、更高的精度以及更广的应用范围,然而在这一片欣欣向荣之景的角落,却隐藏着一个“幽灵”,它难以被人发现却能轻松“破坏”研究者们引以为傲的智能机器.尽管这些智能机器的精确度已在诸多应用中远超人类,但在这个“幽灵”面前,很可能立刻变成低能儿.而这个“幽灵”就叫做对抗样本(Adversarial Examples)[2].

2013 年,Szegedy 等[3]在研究图像分类问题时首次发现了这个奇怪的现象:在测试图片上附加一些经过特殊设计且人眼难以察觉的轻微扰动,并将其输入基于深度神经网络(Deep Neural Network,DNN)的图像分类系统后,会得到错误的输出结果,而这个错误的输出甚至可以被他们任意指定.简单来说,对于这个分类系统,干净样本(未附加扰动的原始图像)与附加了扰动的样本有着巨大的差异,但在人类观察者眼中,两者几乎毫无差别.如图 1 所示,深度模型将加入了细微扰动的“熊猫”错误地识别为了“长臂猿”.图 1a 是干净样本,可以看到图中是一只熊猫.图 1b 就是经过特殊设计生成的对抗扰动,它好似一团毫无意义的噪声.而图 1c 就是干净样本附加扰动之后生成的对抗样本了,我们会认为它和图 1a 完全一样,但分类器却将它识别为了长臂猿.

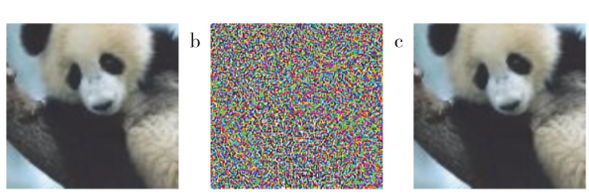


图 1　对抗样本示意图[3]

Fig. 1　The adversarial example[3]

这个发现很快引起了各方研究者的注意.他们将这个被附加轻微扰动从而具有“攻击性”的输入样本定义为对抗样本,而生成这些扰动的算法就叫做对抗样本生成算法,也叫攻击算法(Adversarial Attack).

随着对对抗样本的深入研究,针对其他模型或任务的攻击算法也随后出现.各种深度学习模型如DNN模型、强化学习模型、循环神经网络模型等,以及各类任务包括图像分类、场景检测、语义分割等,无一例外都被“量身定制”的对抗样本成功攻击.

既然存在着对抗样本这么一支锋利的矛,那么就需要一个坚固的盾来抵御它.事实上,针对防御算法(Adversarial Defense)[4]的研究早在对抗样本发现初期就开始了.尽管目前在防御方面确实取得了不少的成果,提出了许多切实可行的防御思路,但始终存在着难以突破的局限与挑战,很多时候这些防御方法无法得到令人满意的结果.就目前来讲,现有防御算法仍无法有效抵御大部分攻击算法.

1 对抗样本攻击算法

对于对抗样本的研究也不过6年时间,针对不同模型或不同任务的攻击算法却有不少.简单来说,大致可以将这些攻击算法分为两类,分别是有目标指向的攻击和无目标指向的攻击.前者是指在对抗样本输入模型之后,会获得攻击者指定好的错误结果,比如让受到扰动的汽车图像统一错分类为风筝.而后者表示获得的结果只要是错误的就行,具体内容无所谓.另外,如果进一步细分,还可以分为单步攻击和迭代攻击两种.表1给出了部分比较典型的攻击算法.

表1 代表性对抗样本攻击算法

Table 1 Representative adversarial attack algorithms

	单步攻击	迭代攻击
无目标指向	FGSM[5],R+FGSM[6]	BIM[7],DeepFool[8],UAP[9],PGD[10]
有目标指向	LLC[7],R+LLC[6]	ILLC[7],JSMA[11],C&W[12],EAD[13]

1.1 无目标指向的攻击

首先介绍最为经典的对抗样本生成方法,快速梯度符号算法(Fast Gradient Sign Method,FGSM),该方法在2014年由Goodfellow等[5]提出,也是最早的攻击算法之一.该算法利用分类器输出结果与真实标签间的损失构建对抗扰动生成模型的目标函数,通过对干净样本附加上扰动 $\boldsymbol{\sigma}$,让损失函数的值尽可能大,从而使分类器的预测结果发生改变,即:

$$f(\boldsymbol{x}+\boldsymbol{\sigma})\neq f(\boldsymbol{x}), \tag{1}$$

$$\max \mathrm{Loss}(\boldsymbol{x}+\boldsymbol{\sigma},f(\boldsymbol{x})), \tag{2}$$

其中,$\boldsymbol{x}$ 表示干净样本,$\boldsymbol{x}+\boldsymbol{\sigma}$ 即为对抗样本.式(1)表示分类器对两个样本的分类结果不一致,式(2)则展示了对抗扰动 $\boldsymbol{\sigma}$ 的生成思路,即最大化对抗样本的分类结果与真实标签之间的损失.

与此同时,扰动本身要限制在一个人眼无法察觉或对干净样本无法产生实质性破坏的范围内.FGSM采用最大化损失函数的方式产生扰动.显然,扰动发生在梯度方向是最有效的,因此FGSM通过在干净样本的梯度方向平移较小量级的步长来得到扰动 $\boldsymbol{\sigma}$:

$$\boldsymbol{\sigma}=\epsilon\mathrm{sign}(\nabla_x J(\boldsymbol{\theta},\boldsymbol{x},y)), \tag{3}$$

式中的 y 表示 $\boldsymbol{x}$ 对应的正确标签,$\boldsymbol{\theta}$ 是网络的权重参数,$\mathrm{sign}(\nabla_x J(\cdot))$ 则描述了 $\boldsymbol{x}$ 处损失函数的梯度方向,ϵ 是在该方向上的偏移量级,通过向分类器损失的梯度方向迈出一大步来生成对抗扰动.

FGSM是无目标指向的单步攻击算法,因此其训练速度很快,但由于其攻击思路相对简单,导致其攻击效果不是很理想,而且目前许多防御算法都能高效抵御FGSM的攻击.

对FGSM的直接改进便是放弃向梯度方向跨固定步长的做法,转而迭代地进行许多次的小步幅扰动,在每次扰动后调整扰动方向以达到攻击的目的,这就是基本迭代算法(Basic Iterative Method,BIM)[7].

此外,无目标指向的迭代攻击还有DeepFool[8]、UAP[9]、PGD[10]等.此处再介绍一下DeepFool.DeepFool由Moosavi-Dezfooli等[8]提出,该算法主要根据网络的原始决策边界来迭代生成对抗扰动,将位于分类边界内的图像逐步推到边界外,直到出现错分类.形象来说,它和BIM的差异在于:BIM通过多次小步幅的调整,将受扰动的图像沿着任意路径远离正确的类直至出错,而DeepFool则是让其从正确的类指向类决策边界进行移动,以最短路径进入其他的类别区域.这种做法使产生的扰动更加细微,进一步增强了其不可见性.

1.2 有目标指向的攻击

对于有目标指向的单步攻击,比较典型的有LLC算法[7].事实上,这种算法是FGSM的一个扩

展,它用 DNN 分类器预测的最低概率的类别标签来替代 FGSM 中使用的真实标签,并最小化损失函数,然后从原始图像中减去计算出来的扰动从而得到对抗样本.

类似于 BIM,LLC 自然也存在其迭代版本 ILLC[7].其他有目标指向的迭代攻击还有 JSMA[11]、C&W 攻击[12]、EAD[13]等.其中 C&W 攻击算法由 Carlini 和 Wagner 提出,他们针对同时期提出的用于抵抗对抗样本的防御蒸馏法[14],引入了 3 种攻击算法.C&W 算法的优势在于其可以根据自身需求调节置信度,且生成的扰动更小,同时它可以破解包括防御蒸馏法在内的多种防御算法,使其适用于黑盒攻击.当然,这个攻击算法的缺点是计算量太大.

除此以外,已有的攻击算法还可以从其他角度进行分类,比如数字化攻击(Digital Attack)和真实世界攻击(Real-world Attack).从字面上就能理解,前者只是在同一计算机内以纯粹数字的形式进行计算和攻击,而后者则是能在真实世界将对抗样本打印出来,并对第三方识别系统如手机、广场监控等进行攻击.另外还可以分类为白盒攻击(White-box Attack)和黑盒攻击(Black-box Attack).前者表示在对模型和训练集完全了解的前提下进行扰动的生成,而后者则是在对模型和训练集知之甚少的情况下进行对抗攻击.

2 对抗样本防御算法

对抗样本的存在使一些极具安全敏感性的技术领域受到严重威胁,研究能有效抵御对抗样本的防御机制成为当前深度学习安全领域的重要课题.

从防御方思路上看,目前大致可以将防御算法分为三类(表 2):

1) 对数据集进行修改或预处理;

2) 对原模型进行修改;

3) 添加外部模型而不改动原模型.

第一种方式不直接涉及模型本身而是把关注点放在数据集上,第二和第三种方式则更关注模型本身的优化.

另外,从防御效果上看,可以把防御算法分成两类:完全防御和检测防御.

前者意在使模型完全抵御对抗样本的攻击从而恢复原有性能,后者的目的则在于让模型“意识”到自己受到了攻击,从而发出警报,但并不能实现真正意义的防御.

表 2 对抗样本防御算法部分总结

Table 2 Representative adversarial defense algorithms

	完全防御	检测防御
从数据集入手	对抗训练[15]、数据压缩[16]、数据增强[17]、随机化[18]	
修改模型自身	梯度正则化[19]、梯度掩蔽[19]、防御蒸馏[14]、DeepCloak[20]、Parseval 网络[4]	SafetyNet[21]、探测子网[22]、Aca 扰动检测[23]
添加外部模型	扰动校正网络[24]、GAN-based 防御[25]	特征挤压[26]、MagNet[27]、自适应降噪[28]

2.1 对数据集进行修改

首先介绍对抗训练(Adversarial Training)[15].在原有的训练集上加入对应的对抗样本数据集,即让原图与对抗样本一起作为训练集输入来训练模型.实验证明,这种对抗训练能有效提升模型的鲁棒性.同时,对抗训练能使网络进一步规范化,减少过度拟合.当然,这种方法的局限性非常大,已有相关实验证明,对于已经接受对抗训练的网络,仍然可以构造出其他有效的扰动,从而获得新的对抗样本进行攻击.

Dziugaite 等[16]受到图片 JPG 压缩的启发,将这种压缩方式运用在了对抗样本上.实验证明,JPG 压缩可以在很大程度上扭转 FGSM 扰动下分类精度下降的趋势.然而,这种压缩方式,包括后来尝试的 DCT、JPEG 及 PCA 等方式,远远不能产生高效的防御,同时压缩也导致了图片本身精度的下降,有些得不偿失.

其他还有如在训练过程中做高斯数据增强[17],在测试时对图像作随机填充[18]等,也能在一定程度上增强鲁棒性.

2.2 修改网络模型本身

Ross 和 Doshi-Velez[19]将输入梯度正则化作为一种防御思路.他们的方法是,在训练可微模型的同时,惩罚导致输出相对于输入变化的变化程度.这意味着,一个小的对抗性扰动不太可能大幅改变训练模型的输出.实验也表明,该方法和对抗训练相结合,对 FGSM 和 JSMA 等攻击具有很好的鲁棒性,但这种方法成倍增加了网络的训练复杂度.

Papernot 等[14]利用蒸馏的概念使深层神经网络对对抗样本攻击具有鲁棒性.蒸馏是 Hinton 等[29]引入的一种训练过程,用于将更复杂的网络知识转移

到更小的网络之中.Papernot 等[14]引入的蒸馏过程的变体,实质上就是利用网络的知识来提高自身的鲁棒性,这种算法会以概率向量的形式从训练数据中提取知识,并反馈给原模型进行训练.实验表明,这种做法可以提高网络对图像中微小扰动的恢复能力.

Lu 等[21]曾假设,在网络的后期阶段,对抗样本相比于干净样本,会产生不同的 ReLU 激活模式.因此,他们提出了 SafetyNet,其将 SVM 分类器添加到目标模型中,让 SVM 使用网络后期 ReLU 计算的离散码来检测图像中是否存在扰动.

其他还有 Gao 等[20]的 DeepCloak,Metzen 等[22]的探测子网(Detector Subnetwork)以及 Grosse 等[23]的 Aca 扰动检测等防御算法.

2.3 使用附加模块

Akhtar 等[24]提出了一种针对全局扰动的防御框架.该框架向目标网络添加额外的扰动校正网络(Perturbance Revise Network,PRN),训练它们对扰动后的图像进行校正,使分类器的预测指向正确结果.而训练 PRN 网络的过程不影响原有网络的内部参数.

此外,Lee 等[25]使用近年流行的 GAN 框架[30]来训练一个对 FGSM 之类的攻击具有鲁棒性的防御网络.他们利用试图产生对抗扰动的生成器来训练分类器.在训练过程中,生成器不断尝试生成具有更强攻击能力的对抗扰动,而分类器则不断尝试正确地对干净样本和对抗样本进行分类.这样,经过多次迭代训练,其训练出的分类器将会具有更强的鲁棒性.

Xu 等[26]提出使用特征压缩来检测图像存在的扰动.他们在分类器网络中增加了两个外部模型,这些模型降低了图像中每个像素的颜色位深度,并对图像进行空间平滑.在此过程中,附加模型对原始图像和压缩图像进行预测比较,如果差异值超过特定阈值,则代表该图像受到了攻击.

其他还有 Meng 等[27]提出的 MagNet 以及 Liang 等[28]提出的自适应降噪算法等.

3 智能驾驶与对抗样本

前文介绍了一些典型的对抗样本攻击与防御算法,在此基础上进一步介绍一下对抗样本对智能驾驶视觉感知的影响.

在此之前,关于对抗样本是否在物理世界真实存在的问题,学术界曾展开过激烈的讨论.而 OpenAI 经过深入研究,在其博客(https://openai.com)上发表了他们的研究成果,并给出了明确的结论:物理世界存在稳定的对抗样本.此后,关于对抗样本的研究范围进一步扩大,针对真实世界的攻击算法接连出现,当下热门的智能驾驶(Automatic Driving),因其超高的安全性要求,更是成为研究者们的重要关注对象.

智能驾驶汽车利用雷达装置、视觉装置、定位系统等部件协同合作,让计算机可以在没有人类介入的情况下,自动安全地操作机动车辆.尽管自动驾驶技术已经研究了多年,许多公司也致力于汽车智能化发展方向,但多次死亡事故[31]依然时刻警示着研究人员.其中最近的一起特斯拉的死亡事故引发了业界对于智能驾驶视觉感知系统安全性能的深思.这次事故中车辆上的视觉系统错误地将前方的大型货车车厢识别为了天空,使车辆高速地撞上了货车.很多人会认为这无非就是视觉模型还存在漏洞,说明模型还没达到绝对的精准性,只要做进一步训练就能避免.

然而真的这么简单就能解决问题吗?当然,就这起事故而言,确实是由模型本身存在的缺陷导致.但是否有可能在视觉系统确实足够精准的情况下,通过某种手段让其失效呢?对抗样本给出了明确的答案:完全可以.事实上,已经有研究人员针对如路标识别[32]、行人检测[33]等智能驾驶常用的视觉感知技术进行了攻击实验,并得到了明显的成效.

3.1 路标识别攻击

路标识别对智能驾驶系统来说,是一个必需且基本的技术,智能汽车应能够准确识别出路标所表示的内容,从而采取相应的机动措施.例如,当识别到 STOP 路标时应进行停车动作,再如当识别到限速 40 的路标时就要进行相应的限速措施.

但最近一项研究表明[32],只要在路标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.这项研究由华盛顿大学、密歇根大学、斯托尼布鲁克大学、加州大学伯克利分校、斯坦福大学和三星集团美国研究所的研究人员共同合作完成.论文展示了两种不同的攻击方式,而在第二种攻击中,他们只需要几个小小的标签,就能让 YOLO v2[34]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.

首先介绍是第一种攻击,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.

前文在介绍对抗样本时提到过,只要在原图上附加非常轻微的扰动,就能产生攻击效果.事实上,这种做法目前只局限于数字化攻击中,如果要将对抗样本带入真实世界,那么这种扰动必须是较为显眼的.因为打印过程、再摄像过程、再保存过程都会产生信息丢失,原本细微的扰动极容易在这些过程中被"消去".因此不难发现,图 2a 中假路标上的扰动确实过于明显,真的拿它去欺骗视觉感知系统的话,虽然效果很好,但很容易就被人怀疑.

a. 全图扰动　　b. 对抗补丁

图 2　针对 STOP 路标的对抗样本展示图[32]

Fig. 2　The adversarial examples on STOP signs[32]

因此,第二种攻击所考虑的问题就是如何让扰动尽可能不被人警觉.在第二种攻击中,研究人员使用了一种新的对抗样本形式,叫做对抗补丁(Adversarial Patch),也就是上文所说的小标签.他们对 RP2 算法[35]进行了改进,加入自己设计的"Disappearance Attack Loss",成功制作出了可以将 STOP 路标"隐藏起来"的对抗补丁,如图 2b 所示.尽管这些小标签本身还是比较显眼,图案内容也和路标有些格格不入,但相比于第一种的全局扰动,其覆盖面积大大减少,且在人类眼中其对路标内容的影响基本上可以无视,因此可以认为在很大程度上达到了不被人警觉的要求.

下面重点介绍他们的第二种攻击算法.该攻击算法针对的是基于 YOLO v2 的物体检测模型.实验使用了专门设计的损失函数来对对抗补丁进行训练,通过迭代训练使该损失函数的值最小化,最终得到相应的补丁图案,使附加上该补丁的 STOP 路标无法被识别系统检测出来.

训练对抗补丁所使用的损失函数由三部分构成:

1) Disappearance Attack Loss

$$J_d(\boldsymbol{x},y)=\max_{s\in S^2,b\in B}P(s,b,y,f_{\boldsymbol{\theta}}(\boldsymbol{x})),\tag{4}$$

该损失函数是最核心的部分,其中 $\boldsymbol{x}$ 代表附加了对抗补丁的输入图像,$f_{\boldsymbol{\theta}}(\boldsymbol{x})$ 表示深度模型的输出,y 表示指定的类别,该实验自然指向的是 STOP 路标,s 表示网格单元,b 表示锚点.这里解释一下:YOLO v2 网络的输出会以 19×19(5(4+1+n))的三维张量表示,意思是 19×19 的网格单元,每个单元有 5 个锚点,每个锚点包含 4 个边界框信息、1 个物体存在置信度和 n 个类别各自的概率,$P(\cdot)$ 则表示给定网格单元和锚点的物体存在置信度.

该损失函数的目的,是尽可能降低指向 STOP 路标的最大存在置信度,通过多次迭代训练,使检测器最终无法检测出 STOP 路标.

2) Tatal Variation Loss

$$\mathrm{TV}(M_x\cdot\boldsymbol{\delta})=\sum_{i,j}|(M_x\cdot\boldsymbol{\delta})_{i+1,j}-(M_x\cdot\boldsymbol{\delta})_{i,j}|+|(M_x\cdot\boldsymbol{\delta})_{i,j+1}-(M_x\cdot\boldsymbol{\delta})_{i,j}|,\tag{5}$$

该损失函数计算的是对抗补丁相邻像素间的色值差,其目的在于让扰动更加平滑,让对抗信息得以区块化.若不使用该损失,补丁信息将显得噪声化,从而降低真实世界攻击的成功率.函数中 $\boldsymbol{\delta}$ 表示对抗补丁,M_x 表示对补丁做一定的随机变换如亮度调节、旋转、加噪等.

3) Non-printability Loss

$$\mathrm{NPS}(M_x\cdot\boldsymbol{\delta}),\tag{6}$$

该函数针对的是对抗补丁的可打印性.由于打印机的色域有限,某些计算机中存在的颜色打印机中却没有,从而无法被打印出来.该函数的目的在于让对抗补丁每个像素的色值尽量接近打印机所拥有的色值.

将三部分损失函数相加,进行迭代训练,最终可以得到能高效攻击 STOP 路标的对抗补丁.除此以外,研究人员还设计了名为"Creation Attack Loss"的第四个损失函数.与之前的工作不同,该函数的目的是让检测器在无法检测出 STOP 路标的前提下,在相同的位置能够检测出其他原本不存在的物体,相应的实验也显示出了较好的结果.图 3 展示了该实验的部分结果.

另外,尽管这项实验针对的是 YOLO v2 模型,但研究人员通过进一步设计,成功将攻击迁移至了 Faster-RCNN.

图 3　STOP 路标攻击实验展示图[32]

Fig. 3　Attack experiments on the STOP signs[32]

这项研究充分表明了智能驾驶的视觉感知系统仍有很长的路要走,如果不能有效抵御对抗样本的侵扰,不排除会有不法分子对路边的路标进行恶意攻击,从而使智能汽车无法正确识别,引发交通事故.

3.2　行人检测攻击

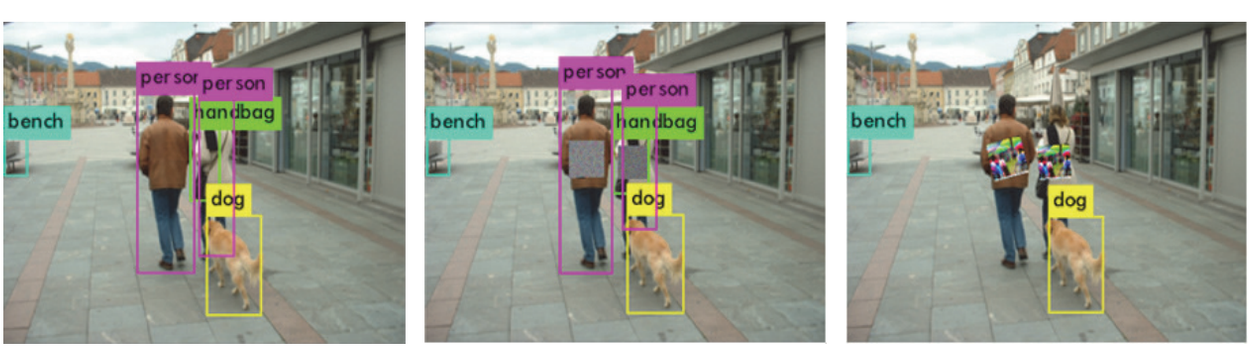
来自比利时鲁汶大学的几位研究人员研究发现[33],借助一张简单打印出来的图案,就可以完美避开行人检测系统.如图 4 所示,视觉系统成功检测到了左边的人,但却没能发现右边的人.可以看到,右边的人身上挂了一块彩色的纸板,也就是前文所说的对抗补丁,正是这块补丁欺骗了视觉感知系统,让系统无法发现这块补丁所在之处还存在着一个人.

该对抗补丁的生成算法和上文介绍的路标识别攻击算法有着相同的思路,都是通过设计好的损失函数,去训练一个固定形状的对抗补丁,使附加了补丁的对象在输入检测系统后得到尽可能低的物体存在置信度.

图 4　对抗补丁把人“隐藏”了起来[33]

Fig. 4　The adversarial patch "hides" the person[33]

行人检测攻击的负面影响是非常大的.比如,可以通过这种补丁恶意地绕过监控系统,或者更糟糕地,让别人穿上带有对抗补丁的衣物,使其在路上行走时无法被智能车辆检测出来,从而大大增加事故概率.图 5 展示了该项实验针对室外场景的一部分数字化攻击实验.



a. 干净样本　　b. 附加噪声补丁　　c. 附加对抗补丁

图 5　对抗补丁行人检测攻击实验[33]

Fig. 5　Attack experiments on pedestrians[33]

因此,对于智能汽车将要普及的未来时代,务必需要找到能高效抵御这类对抗样本的方法,这样才能进一步保障智能汽车的安全性.

3.3 道路场景的其他恶意攻击

上文介绍的两项研究成果表明,现有的攻击算法已经可以通过补丁的形式,使智能驾驶视觉感知系统针对路标和行人的检测失效.这些诡异的对抗补丁,时刻威胁着视觉系统的正常工作.

可以做一个设想,回想一下之前提到的特斯拉追尾事故,是否有可能存在一些不法分子,恶意地在大大小小的货车车厢后面贴上一些不起眼的对抗补丁,从而让这些货车"消失"在路上呢? 又或者在一些弯道很多的山路上,在转弯处的护栏上贴上这些诡异的补丁,让智能汽车误以为那里是一条直路而直冲而去呢? 虽然相信这些反人类的行为不太可能出现,但只要还没研究出高效的防御机制,这种隐患就绝对不能视而不见.

除了针对物体检测和识别的攻击,目前也有文献报道了针对语义分割、实例分割等视觉任务的攻击方法.Xie 等[36]成功使用自己设计的密集对抗生成算法(Dense Adversary Generation,DAG),对图像生成全局扰动,使视觉系统对其做出错误的语义分割,如图 6 所示.

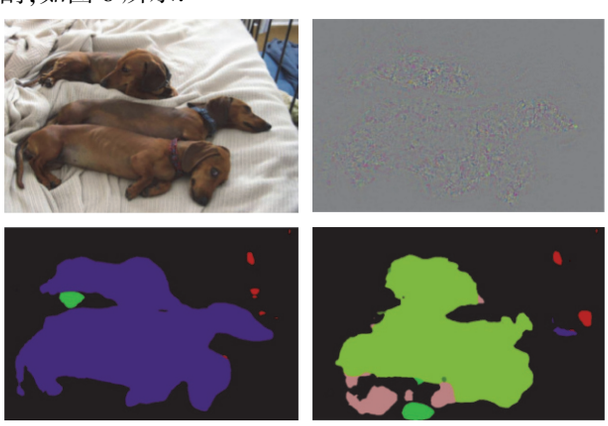


图 6 语义分割模型错误地分割了附有扰动的图像[36]

Fig. 6 The semantic segmentation model incorrectly segmented the image with disturbance[36]

图 6 中的左上图是干净样本,左下图是干净样本的语义分割结果,右上图展示的是 DAG 算法生成的对抗扰动,右下图则展示了干净样本受到攻击后所得到的语义分割结果.

语义分割技术可以让智能汽车在检测出前方物体的同时,能进一步勾画出物体的轮廓,从而获得更加精细的路况信息,更顺利地进行超车、绕行、规避等动作.

若上述针对语义分割任务的攻击算法移植到智能驾驶系统,其影响可想而知.事实上,研究人员完全可以设计出一系列面向道路场景的对抗补丁,将行人、树木等物体与背景"融为一体",使视觉系统无法将其正确分割出来.

4 研究展望

到目前为止,面对各种各样的对抗样本攻击,仍没有足够有效的防御机制来抵抗它们.更何况这些对抗样本或补丁在进一步的研究之中,将会变得更小、更隐蔽、更具破坏力与迁移力,并对所能想到的经常出现于道路场景的任何物体进行攻击,使智能汽车的视觉感知系统无法检测出它们的存在.

可以说,对抗样本的存在,对智能驾驶的安全性提出了巨大的挑战.尽管目前大部分的攻击属于白盒攻击,市面上许多智能驾驶企业也对其使用的视觉感知模型处于保密状态,但对抗样本的多任务化、强迁移化、高隐蔽化、强攻击化正是目前攻击算法的必然趋势,如果不能研究出可以高效抵御这些对抗样本的防御方法,智能汽车将始终藏着一枚定时炸弹,时刻威胁着乘客和行人的生命安全.

针对对抗样本攻击与防御方法的研究将是一个长期的任务,它不仅有趣而且至关重要.一方面,研究对抗样本可以让人们从一个新的角度去剖析深度神经网络的运行机制;另一方面,高效抵御对抗样本也必然是未来达成人工智能终极目标——通用人工智能(AGI)所必须跨越的鸿沟.当然,就目前来讲,其最重要的影响就是能进一步保障深度学习模型的安全性,尤其是针对智能驾驶这类有着极高安全性要求的技术领域,毕竟对于一切的人类活动,安全问题乃重中之重.

参考文献

References

[1] Lecun Y,Bengio Y,Hinton G.Deep learning[J].Nature,2015,521(7553):436-444

[2] Akhtar N,Mian A.Threat of adversarial attacks on deep learning in computer vision:a survey[J].IEEE Access,2018,6:14410-14430

[3] Szegedy C,Zaremba W,Sutskever I,et al.Intriguing properties of neural networks[J].arXiv Preprint,2013,arXiv:1312.6199

[4] Yuan X,He P,Zhu Q,et al.Adversarial examples:attacks and defenses for deep learning[J].IEEE Transactions on

Neural Networks and Learning Systems, 2019, 30(9): 2805-2824
[5] Goodfellow I J, Shlens J, Szegedy C. Explaining and harnessing adversarial examples[J]. arXiv Preprint, 2015, arXiv: 1412.6572
[6] Tramèr F, Kurakin A, Papernot N, et al. Ensemble adversarial training: attacks and defenses[J]. arXiv Preprint, 2018, arXiv: 1705.07204
[7] Kurakin A, Goodfellow I J, Bengio S. Adversarial examples in the physical world[J]. arXiv Preprint, 2017, arXiv: 1607.02533
[8] Moosavi-Dezfooli S M, Fawzi A, Frossard P. DeepFool: a simple and accurate method to fool deep neural networks[C]// 29th IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016: 2574-2582
[9] Moosavi-Dezfooli S M, Fawzi A, Fawzi O, et al. Universal adversarial perturbations[C]// 30th IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017: 1765-1773
[10] Madry A, Makelov A, Schmidt L, et al. Towards deep learning models resistant to adversarial attacks[J]. arXiv Preprint, 2018, arXiv: 1706.06083
[11] Papernot N, McDaniel P, Jha S, et al. The limitations of deep learning in adversarial settings[C]// 2016 IEEE European Symposium on Security and Privacy (Euro S & P), 2016: 372-387
[12] Carlini N, Wagner D. Towards evaluating the robustness of neural networks[C]// 2017 IEEE Symposium on Security and Privacy (S&P), 2017: 39-57
[13] Chen P Y, Sharma Y, Zhang H, et al. EAD: elastic-net attacks to deep neural networks via adversarial examples[C]// 32th AAAI Conference on Artificial Intelligence, 2018: 118-129
[14] Papernot N, McDaniel P, Wu X, et al. Distillation as a defense to adversarial perturbations against deep neural networks[C]// 2016 IEEE Symposium on Security and Privacy (S&P), 2016: 582-597
[15] Kurakin A, Goodfellow I, Bengio S. Adversarial machine learning at scale[J]. arXiv Preprint, 2017, arXiv: 1611.01236
[16] Dziugaite G K, Ghahramani Z, Roy D M. A study of the effect of JPG compression on adversarial images[J]. arXiv Preprint, 2016, arXiv: 1608.00853
[17] Zantedeschi V, Nicolae M I, Rawat A. Efficient defenses against adversarial attacks[C]// 10th ACM Workshop on Artificial Intelligence and Security, 2017: 39-49
[18] Buckman J, Roy A, et al. Thermometer encoding: one hot way to resist adversarial examples[C]// Proceedings of the 5th International Conference on Learning Representations (ICLR), 2018
[19] Ross A S, Doshi-Velez F. Improving the adversarial robustness and interpretability of deep neural networks by regularizing their input gradients[C]// 32th AAAI Conference on Artificial Intelligence, 2018: 1660-1669
[20] Gao J, Wang B L, Lin Z M, et al. DeepCloak: masking deep neural network models for robustness against adversarial samples[J]. arXiv Preprint, 2017, arXiv: 1702.06763
[21] Lu J J, Issaranon T, Forsyth D. SafetyNet: detecting and rejecting adversarial examples robustly[C]// IEEE International Conference on Computer Vision (ICCV), 2017: 446-454
[22] Metzen J H, Genewein T, Fischer V, et al. On detecting adversarial perturbations[J]. arXiv Preprint, 2017, arXiv: 1702.04267
[23] Grosse K, Manoharan P, Papernot N, et al. On the (statistical) detection of adversarial examples[J]. Algorithms, 2017, 11(26): 27-38
[24] Akhtar N, Liu J, Mian A. Defense against universal adversarial perturbations[C]// 31th IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018: 3389-3398
[25] Lee H, Han S, Lee J. Generative adversarial trainer: defense to adversarial perturbations with GAN[J]. arXiv Preprint, 2017, arXiv: 1705.03387
[26] Xu W L, Evans D, Qi Y J. Feature squeezing: detecting adversarial examples in deep neural networks[J]. arXiv Preprint, 2017, arXiv: 1704.01155
[27] Meng D Y, Chen H. MagNet: a two-pronged defense against adversarial examples[C]// 2017 ACM Conference on Computer and Communications Security (CCS), 2017: 135-147
[28] Liang B, Li H C, Su M Q, et al. Detecting adversarial image examples in deep neural networks with adaptive noise reduction[J]. IEEE Transactions on Dependable and Secure Computing, 2018, 25(12): 111-127
[29] Hinton G, Vinyals O, Dean J. Distilling the knowledge in a neural network[J]. Computer Science, 2015, 14(7): 38-39
[30] Goodfellow I J, Pouget-Abadie J, Mirza M, et al. Generative adversarial networks[J]. Advances in Neural Information Processing Systems, 2014, 3: 2672-2680
[31] Mcallister R, Gal Y, Kendall A, et al. Concrete problems for autonomous vehicle safety: advantages of Bayesian deep learning[C]// 26th International Joint Conference on Artificial Intelligence (IJCAI), 2017: 4745-4753
[32] Eykholt K, Evtimov I, Fernandes E, et al. Physical adversarial examples for object detectors[J]. arXiv Preprint, 2018, arXiv: 1807.07769
[33] Thys S, van Ranst W, Goedemé T. Fooling automated surveillance cameras: adversarial patches to attack person detection[C]// 32th IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019: 821-828
[34] Redmon J, Farhadi A. YOLO9000: better, faster, stronger[C]// 30th IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017: 7263-7271
[35] Evtimov I, Eykholt K, Fernandes E, et al. Robust physical-world attacks on machine learning models[J]. arXiv Preprint, 2017, arXiv: 1707.08945
[36] Xie C H, Wang J, Zhang Z, et al. Adversarial examples for semantic segmentation and object detection[C]// IEEE International Conference on Computer Vision (ICCV), 2017: 1369-1378

南京信息工程大学学报(自然科学版),2019,11(6):651-659
Journal of Nanjing University of Information Science and Technology(Natural Science Edition),2019,11(6):651-659

A survey of adversarial attacks and defenses on visual perception in automatic driving

YANG Yijun[1] SHAO Wenze[1] WANG Liqian[1] GE Qi[1] BAO Bingkun[1] DENG Haisong[2] LI Haibo[1,3]

1 College of Telecommunications and Information Engineering, Nanjing University of Posts and Telecommunications, Nanjing 210003

2 School of Statistics and Mathematics, Nanjing Audit University, Nanjing 211815

3 School of Computer Science and Communication, KTH Royal Institute of Technology, Stockholm Sweden 10044

Abstract Nowadays, deep learning has become one of the hottest research directions in the field of machine learning. It has achieved great success in a wide range of fields such as image recognition, target detection, voice processing, and question answering system. However, the emergence of adversarial examples has triggered new thinking on deep learning. The performance of deep learning models can be destroyed by adversarial examples constructed by adding specially designed subtle disturbance. The existence of adversarial examples makes many technical fields with high requirements on safety performance face new threats and challenges, especially the automatic driving system which uses visual perception as the main technology priority. Therefore, the research on adversarial attack and active defense has become an extremely important cross-cutting research topic in the field of deep learning and computer vision. In this paper, relevant concepts on adversarial examples are summarized firstly, and then a series of typical adversarial attack methods and defense algorithms are introduced in detail. Subsequently, a number of physical world attacks against visual perception are introduced along with discussions on their potential impact on the field of automatic driving. Finally, we give a technical outlook on the future study of adversarial attacks and defenses.

Key words adversarial examples; object detection; semantic segmentation; automatic driving