对抗样本生成在人脸识别中的研究与应用

张加胜 刘建明 韩 磊 纪 飞 刘 煌
(桂林电子科技大学计算机与信息安全学院 广西 桂林 541000)

摘 要 随着深度学习模型在人脸识别、无人驾驶等安全敏感性任务中的广泛应用,围绕深度学习模型展开的攻防逐渐成为机器学习和安全领域研究的热点。黑盒攻击作为典型的攻击类型,在不知模型具体结构、参数、使用的数据集等情况下仍能进行有效攻击,是真实背景下最常用的攻击方法。随着社会对人脸识别技术的依赖越来越强,在安全性高的场合里部署神经网络,往往容易忽略其脆弱性带来的安全威胁。充分分析深度学习模型存在的脆弱性并运用生成对抗网络,设计一种新颖的光亮眼镜贴片样本,能够成功欺骗基于卷积神经网络的人脸识别系统。实验结果表明,基于生成对抗网络生成的对抗眼镜贴片样本能够成功攻击人脸识别系统,性能优于传统的优化方法。

关键词 深度学习 黑盒攻击 脆弱性 生成对抗网络 眼镜贴片

中图分类号 TP181 **文献标识码** A DOI: 10.3969/j.issn.1000-386x.2019.05.027

RESEARCH AND APPLICATION OF ADVERSARIAL SAMPLE GENERATION IN FACIAL RECOGNITION

Zhang Jiasheng Liu Jianming Han Lei Ji Fei Liu Huang
(*School of Computer Science and Information Security, Guilin University of Electronic Technology, Guilin 541000, Guangxi, China*)

Abstract Deep learning (DL) models have been widely applied into security-sensitivity tasks, such as facial recognition, automated driving, etc. Attacks and defenses associated with the DL have gradually become hot spots in the field of machine learning and security. 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 and parameters of the model, including data sets. With the increasing dependence on facial recognition technology, it is easy to ignore the security threats caused by its vulnerability when deploying neural networks in high security situations. This paper fully analyzed the vulnerability of the deep learning model and used the generated adversarial network (GAN) to design a novel bright glasses patch sample, which could successfully deceive the facial recognition system based on convolutional neural network. The experimental results show that the adversarial eyeglass patches generated by GAN can successfully attack the face recognition system, and the performance is better than the traditional optimization methods.

Keywords Deep learning Black-Box attack Vulnerability Generative adversarial network (GAN) Eyeglass patches

0 引 言

随着深度学习技术的发展,深度神经网络 DNN (Deep Neural Network)在人脸识别,交通标识识别,安防监控等安全敏感性任务中得到了广泛的应用[1-3]。然而 Szegedy 等首次揭示了深度神经网络脆弱性的存在,极易受到对抗性样本的影响,即通过对原始输入样

收稿日期: 2018-11-21。张加胜,硕士生,主研领域: 机器学习,模式识别。刘建明,教授。韩磊,硕士生。纪飞,硕士生。刘煌,硕士生。

本进行不可察觉的微小的扰动,可以使深度神经网络以较高的置信度错误分类[10]。进而,Goodfellow 解释了对抗样本存在的根本原因,深度模型高度线性的本质[14],并提出了相关对抗样本生成策略,如 FGSM[13]、I – FGSM[5]、RAND + FGSM[7] 等,通过在熊猫的图片中加入微小的扰动向量,在人眼不易察觉的情况下成功使神经网络以高置信度分类为长臂猿。然而这些攻击算法的设计都需要攻击者对目标系统的体系结构或训练参数有充分的了解(白盒)。由于在现实世界中很难获取到目标系统的内部信息,对于攻击者来说,目标系统完全就是一个黑盒。先前的研究表明,不同学习模型之间存在着迁移性[8-9,11],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系统内部信息情景下能够实现黑盒攻击奠定了基础。那么,现实世界中基于深度学习的黑盒系统的安全性受到了一定威胁[6],例如,在人脸识别系统中,攻击者通过对人脸图像做精心设计的改动,使系统误认为是攻击者想要的身份,从而侵入系统导致安全威胁; 对于无人驾驶系统而言,稍微修改“STOP”标识图像使得深度神经网络错误识别为其他标识而不及时停车,将造成安全事故; 再比如在安防监控系统中,攻击者往往通过化妆、装饰物(包括眼镜、假发、首饰等)进行伪装迷惑监控系统,从而混入非法分子。所以,研究对抗样本的生成过程,分析基于深度学习的系统存在的安全漏洞,这将有助于设计更加安全有效的防御机制。

对于人脸识别系统,攻击者往往利用合法用户的照片,通过添加微小的、不可察觉的扰动试图入侵系统,如图 1 所示。然而目前大部分人脸识别系统都针对此类攻击设计了相应的防御机制,来抵御微小扰动的干扰,例如通过对抗性训练新的网络,MagNet 等检测器的设计[4,16]。对于攻击者而言,唯有增加扰动量却极易被系统检测出对抗样本的可能性。进而,相关研究者提出了高亮贴片的概念[12,15],这类贴片样本摆脱了以往的束缚,即不要求生成的样本在人眼看来仍然是原来的图片,贴片样本示例如图 2 所示。攻击者通过打印生成的对抗贴片样本贴在交通标识牌、眼镜框等显眼位置,不仅不会引起警觉,同样可以欺骗深度学习系统。基于该思想,本文通过生成对抗网络原理设计出一种新颖的对抗眼镜贴片样本,并且可以应用于任何真实场景中,从而揭示了现实世界中深度学习模型的脆弱性依然存在。

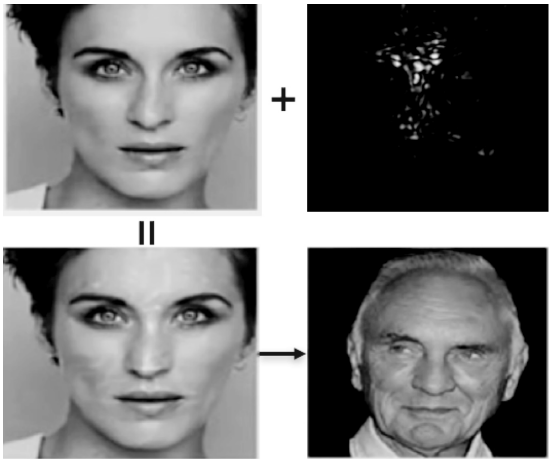

图 1　对抗扰动

图 2　对抗贴片样本

1　相关工作

1.1　黑盒攻击

在现实世界中,由于攻击者很难获取到目标模型(即攻击对象)的内部结构信息,包括训练数据、模型体系结构以及训练参数等。那么,对于攻击者而言,目标系统完全就是一个黑盒,攻击者只能通过原始的输入获得对应的反馈结果。这种情况下,攻击者就很难设计出具有很强攻击性的对抗样本。然而,先前的研究表明,不同的深度学习模型之间存在着迁移性,这为攻击者间接地去实现同样的攻击效果提供了基础。比如,攻击者可以根据已知任务去训练一个替代模型,此时的替代模型对于攻击者而言即为白盒,进而一系列的攻击策略可以设计去生成对抗样本,然后利用模型迁移性的性质,将生成的对抗样本应用于未知模型的黑盒系统。

假设 $X \in R^D$ 表示 D 维的特征空间,$Y = \{y^1, y^2, \cdots, y^c\}$ 表示 c 个不同标签的标签集合。给定合成数据集 $data = \{(x_i, y_i) \mid 1 \leqslant i \leqslant N\}$,$x_i \in X$ 表示第 i 个训练样本,$y_i \in Y$ 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 $data$ 学习一个替代模型 $F(x)$，然后选用某种攻击算法为该模型生成攻击样本 x^*，即通过最优化以下目标损失函数使得神经网络模型以最大概率错误分类为目标标签 y_t:

$$Loss = \max\{F(x^* \mid x+r)_y : y \neq y_t\} - F(x^* \mid x+r)_{y_t} \quad (1)$$

式中: r 表示添加到原始输入样本的微小的扰动向量。

然后将该样本迁移到目标模型 $O(x)$，使目标模型也错误分类，即 $O(x) \neq O(x^*)$。详细攻击流程如图 3 所示。

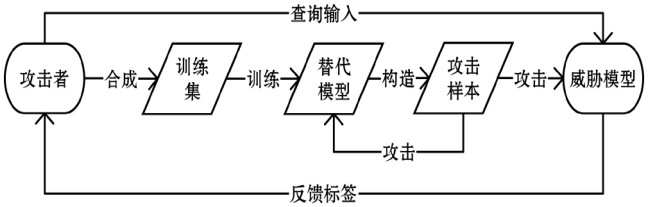

图 3 黑盒攻击流程图

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 $f(\theta, x)$ 的分层组合，来模拟高维输入 x，定义如下:

$$F(x) = f_n(\theta_\theta, f_{n-1}(\theta_{n-1}, \cdots, f_2(\theta_2, f_1(\theta_1, x)))) \quad (2)$$

1.2 生成对抗网络

生成对抗网络 GAN 是一种深度学习模型，主要由生成器(Generator)和判别器(Discriminator)两个模块构成。其中生成器的作用是尽可能地学习真实的数据分布，输入变量 z，则 G 尽可能地生成服从真实数据分布的样本 $G(z)$。判别器的作用是判别其输入数据是来自生成器 G，还是来自真实的数据 x，如果输入来自 $G(z)$，则标注为 0，并判别为伪，否则标注为 1，并判别为真。这里生成器 G 的目标是使其生成的伪数据 $G(z)$ 在判别器 D 上的表现和真实数据 x 在 D 上的表现一致。G 和 D 互相博弈学习并迭代优化的过程使得它们的性能不断提升，随着 D 的判别能力提升，并且无法判别其数据来源时，就认为 G 已学到真实的数据分布，如图 4 所示。在实际应用中一般均使用深度神经网络作为 G 和 D，一个较好的生成式对抗网络(GAN)应用需要有良好的训练方法，否则可能由于神经网络模型的自由行而导致输出不理想。

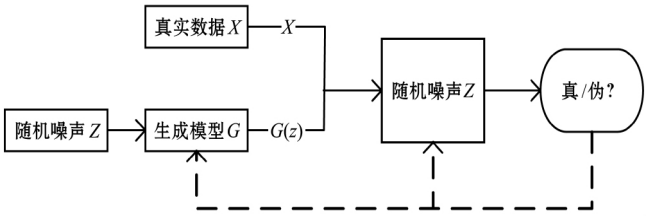

图 4 生成对抗网络框架图

GAN 的训练需迭代进行，通过反向传播更新 G 和 D 的参数。G 的训练通过最小化以下目标函数:

$$Loss_G(Z, D) = \sum_{z \in Z} \lg(1 - D(G(z))) \quad (3)$$

当 G 误导 D(即当 $D(G(z))$ 取值为 1)，该损失达到最小值。D 的训练通过最大化以下目标函数:

$$Gain_D(G, Z, data) = \sum_{x \in data} \lg(D(x)) + \sum_{z \in Z} \lg(1 - D(G(z))) \quad (4)$$

当 D 对真实样本发出 1 时，$Gain_D$ 最大化，否则为 0。目前已经提出了几种 GAN 体系结构和训练过程来训练 GAN，本文主要以 Wasserstein GAN 为主。

2 基于 GAN 对抗样本生成策略

本节主要描述提出的基于 GAN 的对抗样本生成策略以迷惑深度神经网络。第一部分明确了我们的威胁模型，第二部分详细描述攻击框架的设计。

2.1 威胁模型

假设一个攻击者能够成功侵入一个已经训练好的人脸识别系统来发动攻击。由于攻击者不能通过注入错误标记的数据或修改训练数据来危害人脸识别系统的参数。因此，只能通过改变输入以紊乱深度神经网络的分类。

攻击者的目标是通过不易察觉地伪装自己并呈现在人脸识别系统前，然后被错误地归类为除了自己以外的人。考虑这种攻击的两种变体，在定向攻击中，攻击者试图使人脸识别系统错误分类为特定的其他人；在非定向攻击时，攻击者试图使人脸识别系统错误分类为任意其他人，不用特定到某个人。通常假设攻击者在白盒场景下运行。也就是说，攻击者知道特征空间(RGB 图像表示，这在 DNN 图像分类中是常用的)以及被攻击系统的内部(体系结构和参数)。在白盒假设条件下研究 DNN 的脆弱性是本文的重点。此外，黑盒攻击可以使用本地替代模型进行白盒攻击，然后将其转移到黑盒中。

2.2 攻击框架

除了少数几种攻击方法以外，在传统深度神经网络逃避攻击中，攻击者直接改变正常输入样本来最大化或最小化一个预先定义的函数与预期的错误分类相关的函数。与以往的攻击不同，本文提出了通过训练神经网络来产生可用于达到预期目的的输出。也就是说，与其反复调整正常样本输入，使之成为对抗性样本，不如尝试迭代更新深度神经网络的权值来调整将产生导致错误分类的输出。

更具体地说，本文通过训练生成对抗神经网络来

生成眼镜的贴片并可以打印贴在眼镜镜框上,当攻击者佩戴时,不易被察觉,却能使人脸识别系统紊乱,产生定向或非定向的攻击效果。为了达到不显眼的攻击效果,我们要求这些神经网络产生的眼镜图片与真实的眼镜设计相似。与传统的 GAN 训练类似,不同于 GAN 的是,还需要产生对抗性的输出(即包含眼镜的人脸图像),能够误导用于人脸识别的神经网络模型。

所以,本文提出的方法需要训练三个深度神经网络: 一个生成器 G、一个判别器 D 和一个预训练的分类函数为 $F(\cdot)$。当输入 x 到 DNN 时,通过最小化 G 来生成不引人注目的对抗输出,能够迷惑 $F(\cdot)$,优化目标如下:

$$Loss_G(Z,D) - \kappa \cdot \sum_{z \in Z} Loss_F(x + G(z)) \tag{5}$$

其中损失函数 $Loss_G$ 同式(1)定义,通过最小化该损失的目的是希望生成器生成真实的不显眼的眼镜图像输出,能够误导判别器 D,使判别器认为生成的眼镜图像即为真实图像。$Loss_F$ 是在 DNN 的分类中定义的损失函数,在训练 G 过程中,通过最大化该损失,从而使得生成的眼镜图像付在人脸图像之上能够成功欺骗深度神经网络 F。$Loss_F$ 的定义又分为两类,分别针对定向和非定向而言。对于非定向攻击而言,$Loss_F$ 损失定义如下:

$$Loss_F(x + G(z)) = \sum_{i \neq x} F_{y_i}(x + G(z)) - F_{y_x}(x + G(z)) \tag{6}$$

通过最大化 $Loss_F$,样本 x 被 DNN 识别成真实标签 y_x 的概率大大降低,使识别成其他标签的概率增加,从而实现非定向攻击的效果。对于定向攻击而言,$Loss_F$ 损失定义如下:

$$Loss_F(x + G(z)) = F_{y_t}(x + G(z)) - \sum_{i \neq t} F_{y_i}(x + G(z)) \tag{7}$$

通过最大化 $Loss_F$,样本 x 被 DNN 识别成攻击者指定的目标标签 y_t 的概率会增加。

判别器 D 的训练作为训练过程的一部分,目标使式(4)中定义的收益 $Gain_D$ 最大化。通过调整判别器 D 的权重进而激励生成器 G 生成更真实的实例,两者相互博弈,实现互利共赢。与 D 和 G 相反,$F(\cdot)$ 的权重在训练过程中是不需要改变(因为希望在测试阶段,生成的对抗输出能够成功欺骗相同的 DNN)。整个训练过程如算法 1 所示。

算法 1　基于 GAN 对抗眼镜贴片样本生成过程

输入: $X,G,D,F(\cdot),data,Z,N_e,s_b,\kappa \in \{0,1\}$

输出: 对抗输出 G

1: $e \leftarrow 0$;
2: **for** e **in** range(1,Ne)
3: 按照大小 s_b 分块 $data$ 获得 $mini\text{-}batches$
4: **for** $batch$ **in** $mini\text{-}batches$:
5: $z \leftarrow$从 Z 中抽样 s_b 个样本;
6: $gen \leftarrow G(z)$;
7: $batch \leftarrow$合并(gen,$batch$)
8: **if** $even\ iteration$: //更新生成器 D
9: 通过反向传播求 $\partial Gain_D/\partial batch$ 更新 D;
10: **elif** $F(\cdot)$ 错误输出:
11: **return**;
12: $d_1 \leftarrow -\partial Gain_D/\partial gen$;
13: $x \leftarrow$从 X 中抽样 s_b 个样本;
14: $x \leftarrow x + gen$;
15: $d_2 \leftarrow -\partial Loss_F/\partial gen$;
16: $d_1,d_2 \leftarrow$归一化(d_1,d_2)
17: $d \leftarrow \kappa \cdot d_1 + (1-\kappa) \cdot d_2$;
18: 通过反向传播求 d 更新 G
19: **end for**

该算法以一组正常样本 X 作为输入,一个预初始化的生成器和鉴别器,一个用于人脸识别的神经网络,实际示例的数据集(生成器的输出应该类似于这些数据集; 在我们的例子中这是一个眼镜的数据集),一个可以从 G 的潜在空间(Z)中采样的函数,最大训练周期(N_e),mini-batches 的大小 s_b,和 κ(0 和 1 之间的值)。这个训练过程的结果是一个对抗的生成器,它创建输出(即眼镜图像)以迷惑用于人脸识别的深度神经网络 $F(\cdot)$。在每次训练迭代中,D 或 G 都使用随机选择的数据子集进行更新。D 的权值通过梯度上升来更新,以增加增益。相反,G 的权值通过梯度下降来更新,使式(3)定义的损失值最小化。为了平衡生成器的两个目标,分别将 $Gain_D$ 和 $Loss_F$ 的导数进行结合,通过对这两个导数进行归一化得到这两个值的欧几里德范数(算法中第 16 行)。

然后通过设置 κ 来控制这两个目标中的哪一个该获得更多的权重(算法中的第 17 行)。当 κ 接近零的时候,会有更多的权重用以迷惑 $F(\cdot)$,分配更少的权重使 G 的输出更真实。相反,当 κ 接近一个 1 时,会分配更大的权重使 G 的输出类似于真实的例子。当达到最大训练周期时,训练结束,或者当 $F(\cdot)$ 被迷惑时,即定向或非定向攻击已完成。

3　实　验

为了验证上述对抗样本生成策略的有效性,在收集的眼镜数据集上进行了对抗眼镜样本的合成,并在经典的人脸识别模型 OpenFace,VGG 上验证了提出的方法的攻击性能。需要如下准备: (1) 收集训练中使

用的真实眼镜数据集;(2) 选择生成器和鉴别器的架构,实例化它们的权值;(3) 训练可用于评估攻击的 DNNs;(4) 设定攻击的参数。

3.1 数据集采集

一个真实的眼镜框架设计数据集是必要的,以训练生成器创建真实的攻击。我们使用谷歌搜索的搜索"眼镜"及其同义词(如"眼镜"、"护目镜"),有时使用形容词修饰,我们使用的形容词主要包括颜色(如"棕色"、"蓝色")、趋势(如"极客"、"龟甲")和品牌(如"拉夫·劳伦"、"普拉达")。总共做了 430 个独特的 API 查询,收集了 26 520 张图片。

收集的图像不仅仅是眼镜;我们还发现了杯子、花瓶和眼镜品牌的标识阻碍了训练过程;此外,这些图像还包括模特佩戴的眼镜和深色背景下的眼镜图像。我们发现这些图像很难用生成网络进行建模。因此,我们训练了一个分类器来帮助我们检测和保存在白色背景下的眼镜图像,从而不包括模特们佩戴时的图像。使用 250 张手工标记的图片,训练了一个分类器,并将其调整为 100% 的精确度和 65% 的召回率。在对数据集中的所有图像进行应用后,仍有 8 340 幅眼镜图像,手动检查这些图片并没有发现假阳性的一个子集。

使用这个数据集的原始图像,可以训练一个能发出不同图案、形状和方向的眼镜的生成器。不幸的是,形状和方向的变化使得这种眼镜在运行算法 1 时难以有效和合理地对准人脸图像。因此对数据集中的图像进行预处理,并将模式从它们的帧转移到一个固定的形状,可以很容易地对齐到人脸图像。我们使用的形状的剪影如图 5 所示,然后我们训练生成器生成这种形状的眼镜的图像,但是它们有不同的颜色和质地。将眼镜的颜色和纹理转换成固定的形状,对图像进行了识别,以检测帧的区域。然后使用纹理合成技术将框架的纹理合成到固定形状上,图 6 显示了纹理合成结果的示例。

图 5 眼镜生成轮廓

图 6 原始眼镜图像(左)与合成眼镜图像(右)

3.2 实验设置与结果分析

3.2.1 预训练生成器与判别器

当训练 GANs 时,我们希望生成器构造出清晰的、真实的、多样的图像。只发射一小组图像表明生成器的功能不能很好地接近底层数据分布。为了实现这些目标,也为了能够进行有效的训练,我们选择了深度卷积生成对抗网络 DCGAN(Deep Convolutional GAN),一个具有少量参数的极简主义架构。然后探索了发生器潜在空间的各种可能性,输出维度,以及 G 和 D 中的权重数(通过调整过滤器的深度)。最终发现一个潜在的空间 $[-1;\ 1]^{25}$,(即二十五维 25-dimensional 向量之间的实数 -1 和 1),和输出的包含 64×176 像素的图像可产生最好看、最多样化的效果。

为了确保攻击能够迅速融合,将 G 和 D 初始化到一个状态,在这个状态中,生成器才能够有效生成真实的眼镜图像。为此选择了 200 次迭代进行预训练,并存储它们以初始化以后的运行。此外,本文还借鉴了 Salimans 提出的基于软标签来训练生成器。图 7 和图 8 展示了在训练结束时生成器器生成的眼镜图像。

图 7 对抗眼镜图像

图 8 原始人脸图像(左)与对抗人脸图像(右)

3.2.2 人脸识别模型

本文评估了对两个体系结构中的 DNNs 的攻击。一个神经网络是建立在超分辨率测试序列(VGG)神经网络上[17]。最初的 VGG DNN 在人面数据库(LFW)基准测试中,在被标记的面孔上显示了最先进的结果,以 98.95% 的准确率进行面部验证[18]。另外一个 DNN 是在 OpenFace 神经网络上构建的,它使用了谷歌 FaceNet 体系结构[19]。OpenFace 的主要设计考虑是提供高精度 DNN,低训练和预测时间,使 DNN 可以部署在移动设备和物联网设备上。

VGG 网络训练:原始的 VGG 网络需要一个 $224 \times$

224 像素对齐的脸部图像用以输入,并产生了一个具有高度鉴别性的 4 096 维的面部描述符(即用矢量表示法展示脸部图像)。在欧几里德空间中,两个描述同一人的图像的描述符比两个描述不同人的图像的描述符更接近。使用描述符来训练两个简单的神经网络,将图幅面积描述符映射到身份集合上的概率提高。如此,原来的 VGG 网络就有效地充当了特征提取器的角色。

OpenFace 网络训练: 原始的 OpenFace 网络需要一个 96 ×96 像素对齐的脸部图像作为输入和输出 128 维的描述符。与 VGG 网络相似,同一个人的图像描述符在欧几里得空间中接近,而不同人物形象的描述符却相差甚远。与 VGG 相反,OpenFace 的描述符位于一个单位球面上。首先尝试训练神经网络,使用与 VGG DNNs 相似的架构将 OpenFace 描述符映射到身份数据集,找到了这些神经网络来得到有竞争力的精确度。然而,与 VGG DNNs 类似,它们也很容易受到闪避的攻击。与 VGG DNNs 不同,简单的数据增加并不能提高 DNNs 的稳健性。我们认为这可能是由于使用线性分隔符对球面数据进行分类的局限性。

3.2.3　结果分析

为了评估提出的攻击策略的攻击性能,随机为 VGG 和 OpenFace 选取 10 个攻击者。对于每个攻击者和 DNN 的组合,使用攻击者的单个人脸图像来创建定向或非定向攻击。在非定向攻击中,目标是随机选择的。为了避免评估成功率的不确定性,本文使用攻击者的三个不同的图像重复每一次攻击。表 1 描述了模型的测试准确率以及被攻击时的平均成功率和标准误差。实验结果表示基于 GAN 生成的眼镜贴片成功攻击了用于人脸识别的 VGG 和 OpenFace 模型。对于非定向而言更具挑战性,成功率有所降低。

表 1　人脸识别模型性能

DNN 模型	受试者数	测试准确率	定向攻击成功率	非定向攻击成功率
VGG	10	98.9%	97.5% ±0.2%	100% ±0.0%
OpenFace	10	99.2%	96.3% ±0.3%	100% ±0.1%

然后进一步探讨了生成的攻击样本的迁移性,即在黑盒攻击场景中,提出的基于 GAN 的对抗眼镜贴片的有效性,实验结果如表 2 所示。实验结果表明提出的饿攻击策略在黑盒攻击领域仍获得了不错的攻击效果。另外,还发现 OpenFace 体系结构的攻击仅在有限次的尝试中就成功地愚弄了 VGG 体系结构(10% ~12%)。相比之下,在 63.33% 的尝试中,成功攻击 VGG 的同时也成功攻击了 OpenFace。

表 2　黑盒攻击性能

DNN 模型	VGG	OpenFace
VGG	–	63.33%
OpenFace	10.00%	–

4　结　语

本文旨在探索深度学习模型存在的脆弱性并运用生成对抗网络,设计出一种新颖的光亮眼镜贴片样本,能够成功欺骗基于卷积神经网络的人脸识别系统。通过在收集的眼镜数据集上进行合成实验,并在 OpenFace、VGG 等常用的人脸识别模型上验证了所提想法的性能,并证明了现实世界中深度学习模型的脆弱性依然不容忽视,设计有效的防御机制成为未来研究的重点。

参　考　文　献

[1] Papernot N, Mcdaniel P, Sinha A, et al. Towards the Science of Security and Privacy in Machine Learning [EB]. arXiv: 1611.03814, 2016.

[2] Barreno M, Nelson B, Joseph A D, et al. The security of machine learning [J]. Machine Learning, 2010, 81(2): 121 –148.

[3] 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 (EuroS&P). IEEE, 2016: 372 –387.

[4] Feinman R, Curtin R R, Shintre S, et al. Detecting Adversarial Samples from Artifacts [EB]. arXiv: 1703.00410, 2017.

[5] Kurakin A, Goodfellow I, Bengio S. Adversarial examples in the physical world [EB]. arXiv: 1607.02533, 2016.

[6] Dong Y, Liao F, Pang T, et al. Boosting Adversarial Attacks with Momentum [EB]. arXiv: 1710.06081, 2017.

[7] Tramèr F, Kurakin A, Papernot N, et al. Ensemble Adversarial Training: Attacks and Defenses [EB]. arXiv: 1705.07204, 2017.

[8] Carlini N, Wagner D. Towards Evaluating the Robustness of Neural Networks [C]//2017 IEEE Symposium on Security and Privacy(SP). IEEE, 2017: 39 –57.

[9] Papernot N, Mcdaniel P, Goodfellow I. Transferability in Machine Learning: from Phenomena to Black-Box Attacks using Adversarial Samples [EB]. arXiv: 1605.07277, 2016.

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

[11] Tramèr F, Papernot N, Goodfellow I, et al. The Space of Transferable Adversarial Examples [EB]. arXiv: 1704.03453, 2017.

[12] Liu Y, Chen X, Liu C, et al. Delving into Transferable Adversarial Examples and Black-box Attacks [EB]. arXiv: 1611.02770, 2016.
[13] Goodfellow I J, Shlens J, Szegedy C. Explaining and Harnessing Adversarial Examples [EB]. arXiv: 1412.6572, 2014.
[14] Kurakin A, Goodfellow I, Bengio S. Adversarial Machine Learning at Scale [EB]. arXiv: 1611.01236, 2016.
[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.
[16] Meng D, Chen H. Magnet: a two-pronged defense against adversarial examples [C]//Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security. ACM, 2017: 135 - 147.
[17] Parkhi O M, Vedaldi A, Zisserman A. Deep Face Recognition [C]//British Machine Vision Conference 2015. 2015.
[18] Huang G B, Ramesh M, Berg T, et al. Labeled faces in the wild: A database for studying face recognition in unconstrained environments [R]. Technical Report 07 - 49, University of Massachusetts, Amherst, October 2007.
[19] Amos B, Ludwiczuk B, Satyanarayanan M. OpenFace: A general-purpose face recognition library with mobile applications [R]. Technical report, CMU-CS-16-118, CMU School of Computer Science, 2016.

(上接第 79 页)

IP 地址、目的 IP 地址、源端口、目的端口)、平均包大小、包大小的标准差、数据流传输速率、包到达平均时间间隔、包到达时间间隔的标准差、PSH 标志数量、包大小变化次数、下上行数据流总字节数之比。通过以上信息可以清楚分辨数据流是在线视频或下载,并能掌握数据流分类的有关特征信息。

3.3.2 分类性能测试分析与比较

本部分主要分析本文测试时在线视频流量和下载流量的分类性能并与文献[13]进行对比。

通过图 9 可见,在测试的 10 个网站中,视频与下载流量综合分类准确率约为 91.5%。有 8 个网站分类准确率等于或高于 85%,这 8 个网站中有 5 个网站的分类准确率达到 100%。所有测试网站中仅有 2 个网站分类准确率偏低,一个为 75%,另一个为 80%。经数据分析这 2 个网站的不同种类数据流特征信息非常相似,导致分类应用产生了一定程度的误判。

同时与文献[13]中对于在线视频流量和下载流量的分类性能进行对比分析。本文对于在线视频流量和下载流量的分类相对于文献[13]有以下改进:

文献[13]中当数据包阈值为 1 000 时,全局正确率在 85% 左右且波动较大;阈值到达 4 000 时,分类准确率才能稳定在 92% 左右。但本文通过优化特征集并结合 SDN 的架构优势,实现了在数据包数量阈值仅为 1 500 时,也可以将分类准确率稳定保持在 91.5% 左右的效果。这样既能优化处理效率,又能保持较高的分类准确率。

4 结 语

SDN 控制层由于需要处理大量数据流,如何在处理前对流量进行分类,以便在后续操作中减轻 SDN 控制层处理压力,从而实现细粒度管控,已经成为 SDN 研究的热点之一。本文设计的基于机器学习的 SDN 实时流量分类应用可以实时、有效地区分在线视频流量和下载流量。今后的研究重点在于如何将分类结果应用到 SDN 中实现细粒度的数据流量管控,进而优化 QoS。

参 考 文 献

[1] 许晨辉. 面向 QoS 保证的软件定义网络资源管控技术研究[D]. 南京: 南京航空航天大学, 2016.
[2] 蔡远俊. 基于 SDN 和 OpenFlow 的流量分析系统的研究与设计[D]. 北京: 北京邮电大学, 2015.
[3] 许廷伟. 一种基于 SDN 的流量管理系统设计与实现[J]. 电脑知识与技术. 2015, 11(33): 33 - 36.
[4] 严骏驰. 基于 SDN 的数据中心流量管理研究[D]. 北京: 北京邮电大学, 2016.
[5] 房亚明. 基于 SDN 的数据中心网络流量调度技术研究[D]. 合肥: 安徽大学, 2017.
[6] 任燕凯. 基于 SDN 的物联网智能流量管理机制的设计与实现[D]. 北京: 北京邮电大学, 2016.
[7] 程光, 陈玉祥. 基于支持向量机的加密流量识别方法[J]. 东南大学学报(自然科学版), 2017, 47(4): 655 - 659.
[8] 吴辉. 基于模糊 K-Means 的网络流分类系统研究与实现[D]. 广州: 广东工业大学, 2016.
[9] Shafiq M, Yu X, Wang D. Network Traffic Classification Using Machine Learning Algorithms [C]//International Conference on Intelligent and Interactive Systems and Applications, 2017: 621 - 627.
[10] Tapaswi S, Gupta A S. Flow-Based P2P Network Traffic Classification Using Machine Learning [C]//International Conference on Cyber-enabled Distributed Computing & Knowledge Discovery. IEEE, 2013: 402 - 406.
[11] Bujlow T, Riaz T, Pedersen J M. Classification of HTTP traffic based on C5.0 Machine Learning Algorithm [C]//Proceedings of the 2012 IEEE Symposium on Computers and Communications(ISCC). IEEE, 2012: 882 - 887.
[12] 李贺. 网络视频业务流的特征选择与识别研究[D]. 南京: 南京邮电大学, 2016.
[13] 赵小祥. 基于特征选取的网络游戏与视频业务分类研究[D]. 南京: 南京邮电大学, 2016.