IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

收到日期:2021 年 5 月 25 日,接受日期:2021 年 6 月 20 日,出版日期:2021 年 6 月 25 日,当前版本日期:2021 年 7 月 5 日。

数字对象标识符 10.1109/ACCESS.2021.3092646

针对人脸识别的对抗性攻击：综合研究

法蒂玛·瓦赫什特[1]，艾哈迈德·尼卡巴迪[2]，和拉加文德拉 拉玛钱德拉3，（IEEE 高级会员）
1阿米尔卡比尔理工大学生物医学工程系,德黑兰 1591634311,伊朗
2阿米尔卡比尔理工大学计算机工程与信息技术系,德黑兰 1591634311,伊朗
3挪威生物识别实验室（NBL），挪威科技大学信息安全与通信技术系
技术（NTNU i Gjovik），2815 Gjøvik,挪威

通讯作者:Raghavendra Ramachandra (raghavendra.ramachandra@ntnu.no)

此项研究并未涉及人类受试者或动物。

摘要人脸识别 (FR) 系统已展示出可靠的验证性能,表明其适用于从社交媒体中的照片标记到自动边境管制 (ABC) 等各种实际应用。然而,在具有基于深度学习架构的高级 FR 系统中,仅提高识别效率是不够的,系统还应抵御潜在类型的攻击。最近的研究表明,(深度) FR 系统对难以察觉或可察觉但看起来自然的对抗性输入图像表现出有趣的脆弱性,这会导致模型输出预测错误。在本文中,我们全面调查了针对 FR 系统的对抗性攻击,并详细阐述了针对它们的新对策的能力。此外,我们根据不同的标准提出了现有攻击和防御方法的分类。我们比较了攻击方法的方向、评估过程和属性,以及防御方法的类别。最后,我们讨论了挑战和潜在的研究方向。

索引术语生物识别、面部识别、对抗性攻击、对抗性干扰、深度学习。

1.引言

人脸识别（FR)是一种流行的身份验证生物识别技术,广泛应用于金融、军事、公共安全和日常生活等多个领域。

典型的 FR 系统的最终目标是从数字图像或视频帧中识别或验证一个人。研究人员将 FR 描述为一种基于生物识别人工智能的应用程序,它可以通过分析人的面部特征模式来专门识别一个人。

使用面部作为生物特征的想法始于 20 世纪 60 年代,第一个成功的面部识别系统的设计可以追溯到 20 世纪 60 年代初 [1]。近年来,深度学习的最新进展以及大量硬件和大量数据的使用,使得面部识别算法得到了长足的发展,并且性能准确 [2]–[4]。这种性能使得面部识别技术可以广泛应用于更多样化的应用中,从社交媒体中的照片标记到自动边境控制 (ABC) 系统中的可疑身份识别。

然而,在先进的 FR 模型中,仅仅提高识别效率是不够的,系统还应该能够抵御各种潜在的攻击。最近,研究人员发现 (深度) FR 系统容易受到不同类型的攻击,这些攻击会创建数据变化来欺骗分类器。这些攻击可以通过 (a) 物理攻击（在拍摄图像之前修改面部外观)或 (b) 数字攻击（在拍摄的面部图像中实施修改)发起 [5]。

演示攻击也称为欺骗攻击[6],是物理攻击的主要技术之一。

演示攻击旨在通过展示面部生物特征来颠覆人脸识别系统,包括打印的照片、面部照片的电子显示、使用电子显示器重放视频和 3D 面罩 [7]。

最近有研究表明,化妆也可能被滥用来发起演示攻击[8]。

相比之下,对抗性攻击 [9] 和变形攻击 [10]、[116] 所导致的变化是数字入侵的关键技术。典型的对抗性攻击可以通过精心设计的扰动（称为对抗性示例 [11])来欺骗 FR 系统。应该注意的是

负责协调本稿件审阅的副主编和批准出版的是林毓达

对抗攻击主要属于数字攻击类,例如,对抗性示例生成方法大多以数字方式在人脸图像上实现,但也有一些方法旨在通过对人脸外观进行物理更改,然后捕获修改后的图像来物理实现 [12]。已经提出了多种方法来克服此类攻击的破坏性后果,包括针对 FR 系统的 [13]–[16] 和针对该区域之外的攻击的 [17]–[19]。另一方面,变形攻击的目标是通过变形和混合两个或更多不同主体(例如,罪犯和同伙)来生成假脸,以将罪犯登记为 FR 系统的合法身份模板 [20],[21]。同样,在这方面已经做出了很多努力来应对破坏性后果,从人脸变形检测方法 [22]–[25] 到同伙面部恢复方法 [26]–[28]。

在不同的攻击中,对抗性攻击最引人注目,因为它们通常以深度神经网络 (DNN) 为目标,并且可能特别针对卷积神经网络 (CNN),而最先进的 FR 模型正是基于该网络建立的。对抗性示例生成领域每年发表的论文数量大幅增长证明了这种类型的攻击(见图 1)。

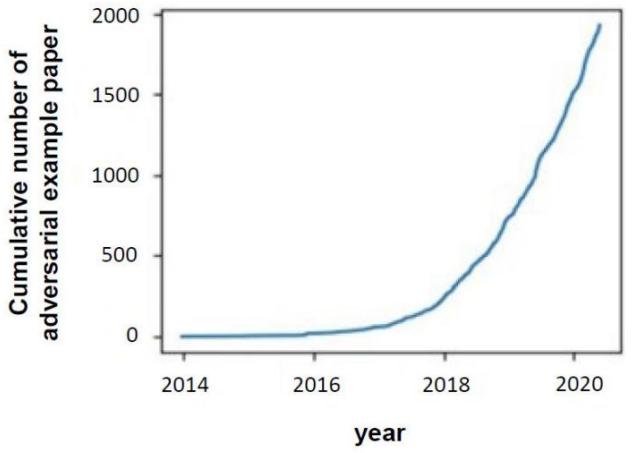

图 1.近年来发表的对抗性示例论文累计数量[29]。

本研究全面调查了旨在欺骗 FR 系统的不同对抗性攻击生成技术,以及针对这些技术制定的潜在对策。这是我们首次尝试回顾 FR 系统的对抗性攻击和防御策略。由于 FR 可能指代人脸识别或人脸验证这两种应用,因此我们在本研究中对这两种应用都进行了回顾。

本文的主要贡献是: ·我们回顾了最近关于 FR 系统上的对抗性示例生成方法的研究,并根据这些方法给出了相应方法的说明性分类法对其方向进行分析,并在方向、评估过程和属性上比较这些方法。 ·我们回顾了针对 FR 系统的新对抗性检测方法,对所提出的算法进行分类,并展示了描述性分类法。 ·我们根据四个主要问题概述了针对 FR 模型的对抗性示例的主要挑战和潜在解决方案:对抗性示例的特殊化/规范、FR 模型的不稳定性、与人类视觉系统的偏差以及与图像无关的扰动生成。

本文的其余部分安排如下:第二部分介绍了 FR 技术、架构和数据集的背景。在第三部分中,我们将在 FR 课程中描述与对抗性攻击和防御相关的标准术语,表示攻击的属性,解释实验标准,并讨论生成攻击的先驱方法。我们将在第四部分回顾旨在欺骗 FR 任务的对抗性示例生成方法。我们讨论了这些方法,并根据方向、评估过程和属性比较了这些方法。在第五部分中,我们研究了相应的对策。我们在第六部分讨论了当前的挑战和未来的潜在研究方向。第七部分总结了本文的工作。

II. 背景

本节简要介绍基本的 FR 系统,并阐述深度学习时代的融合模型。接下来,我们介绍这方面广泛使用的架构和标准数据集。

A. 人脸识别简介

人脸识别是计算机视觉界一个古老的研究课题,首次成功可追溯到 20 世纪 60 年代。从那时起,这条研究道路经历了四个决定性的科学飞跃。用于识别的人脸表征经历了整体学习、局部特征学习、浅层学习和深度学习的顺序形式 [30]。

20 世纪 90 年代初,历史性的特征脸方法 [1] 被引入,此后不久,FR 研究开始流行起来。从那时起到 21 世纪,基于某些分布假设从人脸图像中提取低维表示的整体方法 [31]–[34] 主导了 FR 社区。然而,这些方法在解决偏离先前考虑的假设的不受控制的面部修改方面表现不佳。21 世纪初,基于局部特征的 FR 技术被引入,手工制作的描述符(如 Gabor [35] 和 LBP [36])变得流行起来。然而,独特性和紧凑性是这些局部特征所缺乏的两个属性。21 世纪 10 年代初,基于局部学习的特征被引入 [37]–[39],以学习局部滤波器和编码码本,以获得更好的独特性和紧凑性。虽然解决了缺乏必要属性的问题,但这些

浅层表示表现出对复杂非线性面部外观变化的鲁棒性的丧失。

这些传统方法试图通过一层或两层表示来识别人脸,并提高 FR 准确率,目标是分别探索不受约束的面部变化的各个方面,包括照明、姿势、表情或遮挡。深度学习方法的出现解决了传统方法的局限性。

在基于深度学习的面部识别方法中,多层处理单元学习与不同抽象级别相对应的多种表示。有趣的是,更高级别的抽象表示已表现出对面部照明、姿势、表情和遮挡变化的强大不变性,并以非凡的稳定性表示面部身份。2014 年,DeepFace [3] 在 Labeled Faces in the Wild (LFW) 数据集 [40] 上获得了最先进的准确率。在不受约束的条件下,它首次成功与人类表现相媲美,并通过在 400 万张面部图像上训练 9 层网络接近所需的准确率。深度学习技术几乎在各个方面都彻底改变了面部识别的研究视野,从算法设计和训练/测试数据集到应用程序设置和评估协议。

B. 人脸识别器的典型架构

DeepFace [3] 是第一个引入 FR 社区的杰出深度架构。它具有深度 CNN 架构和多个局部连接层。随后,基于深度学习的 FaceNet [41] 和 VGG-Face [2] 模型相继问世,它们分别用于在大规模人脸数据集上训练流行的 GoogleNet [42] 和 VGGNet [43]。这些模型通过三重损失函数对网络进行微调,并将其应用于通过在线三重挖掘方法创建的脸部补丁。

后来,基于 ResNet 架构 [45] 提出了 SphereFace [44],并提出了一种新颖的角度 softmax 损失,通过角度边缘来学习判别性特征。与此网络类似,分别基于余弦和角度边缘损失引入了 CosFace [46] 和 ArcFace [47]。

这些模型的设计方式是将学习到的特征与较大的余弦和角度距离分开。然后提出了轻量级网络来克服 GPU 功率和内存大小不足的问题,并使其适用于许多移动设备和嵌入式设备。LightCNN [48] 具有新颖的最大特征图 (MFM) 激活函数,是此类的一个著名示例,它可产生紧凑的表示并降低计算成本。

C.标准人脸识别数据集

2007 年,LFW 数据集由网络上 3K 张无约束条件下的人脸图像组成,

为其他测试数据库在不同任务中的使用开辟了一条新道路。拥有足够大的训练数据集来评估深度FR模型的有效性,导致不断开发更复杂的数据集以促进

FR 研究。早期的深度 FR 模型(例如 DeepFace、FaceNet 和 DeepID [49])都是在私有、受控或小规模的训练数据集上进行训练的,因此无法与新模型进行比较。为了解决这个问题,CASIA-Webface [50](包含 10,000 名名人的 0.5M 图像集合)被引入作为第一个广泛使用的公共训练数据集。后来,MS-Celeb-1M [51]、VGGface2 [52] 和 Megaface [53](包含超过 1M 图像的集合)被引入作为公共大规模训练数据集,供许多先进的深度学习方法使用。

III.对抗性攻击生成对抗性攻击包括对原始图像进行精细修改,目的是使改变几乎无法被人眼察觉,以欺骗特定的分类器。

在数字攻击领域,这可以实现为在输入图像x上添加一个最小向量n ,即(x + n),这样深度学习模型F就会对改变后的输入x + n 预测出一个错误的输出,这被称为对抗性示例。这样,生成对抗性示例x的盒约束优化问题通常可以描述为 [9]:

$$\begin{aligned} &\min_{x} \quad \|x - x\|_2 \\ &F(x) = l \; F(x) = l \\ &\qquad l = l \\ &x \in [0, 1] \end{aligned} \tag{1}$$

其中l和l表示x的输出标签, x表示根据L2 范数计算的两个图像样本之间的距 ,和 .2 离。

如图 2 所示,为了欺骗 FR 模型(本例中为 VGG16) ,输入图像被改变,以便人类仍然可以预测正确的类别。然而,深度学习网络会感到困惑并被误导到错误的类别。

Szegedy等人[9] 首次证明了 CNN 模型容易受到通过在输入图像中引入微小噪声而产生的对抗性攻击。GoogleNet 和 VGG-Face 模型的准确度也因色彩平衡操纵而降低。请注意,对抗性攻击的隐蔽性和深度学习算法的广泛应用可能会在现实场景中造成严重损害 [54]。例如,如果在自动驾驶中更改了路牌,对抗性示例可能会对汽车、行人和其他车辆造成过度威胁。

类似地,在 FR 应用中,无法验证改变的输入可能会导致性能下降,而这在闭集验证/识别场景中是可以受益的。

A. 术语和定义

本节简要介绍与(深度)FR 模型的对抗性攻击相关的标准术语。我们对词语的定义对于理解技术至关重要

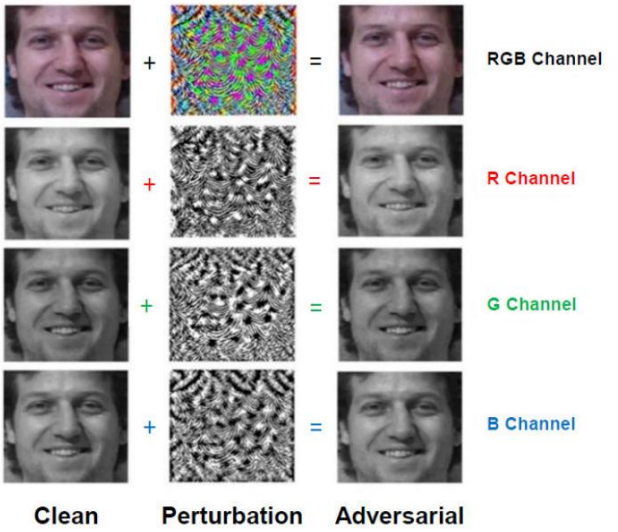

图 2.原始人脸图像（第一列）、VGG-16 的对抗噪声向量（第二列)和改变后的图像（最后一列)的可视化。
从上到下,四行表示在原始 RGB 图像和对应的 R、G、B 颜色通道灰度表示中添加对抗性噪声。对抗性噪声被放大 4 倍以增强可见性 [13]。

所审查研究的组成部分。本文的其余部分遵循相同的术语定义。

·对抗性示例/图像是故意改变（例如,通过添加噪声)的干净图像版本,以欺骗机器学习 (ML) 模型,例如 FR 模型。 ·对抗性训练是使用对抗性图像和干净图像的训练过程。 ·对手是一个代理,他根据案例研究创建对抗性示例或示例本身。 ·躲避攻击发生在攻击者试图将一张脸误认为任何其他任意脸部时。它在文献中也称为混淆攻击[55],[56]。 ·逃避攻击试图通过在测试阶段改变样本但不影响训练数据来逃避系统。 ·冒充攻击试图将一张脸伪装成一张特定的（授权的)脸部。 ·投毒攻击发生在训练期间,以污染训练数据。在这种攻击中,攻击者试图通过插入精心设计的样本来毒害数据,最终破坏整个学习过程。 ·威胁模型是一种将有关攻击者的目标、攻击策略、被攻击系统的知识等假设形式化的模型。

B. 对抗性攻击的属性

在本节中,我们讨论对抗性示例生成方法的主要属性。

1) 对抗能力对抗能力由攻击者可以
获得的有关模型的知识量决定。威胁模型
深度FR系统中,根据攻击能力不同,分为以下几种类型。

白盒攻击假设目标模型具有完整的知识,即其参数、架构、训练方法,甚至在某些情况下,还有其训练数据。

黑盒攻击将对抗性示例（在测试期间)输入目标模型,而这些示例是在不了解该模型的情况下创建的（例如,其训练过程、其架构或参数）。尽管攻击者无法获得该模型的知识,但攻击者可以利用对抗性示例的可迁移性与该模型进行交互（第 III-B.3 节）。

2) 对抗性特异性对抗性特异性被定义
为攻击允许特定入侵/破坏或造成一般混乱的能力。深度 FR 系统中的威胁模型可以根据攻击的特异性分为以下类型。

定向攻击会欺骗模型,使其错误地预测对抗样本的特定标签。在 FR 或生物识别系统中,这是通过冒充知名人士来实现的。

非针对性攻击会不相关地预测对抗性示例的标签,只要结果不是正确的标签。在 FR/生物识别系统中,这是通过面部躲避来实现的。非针对性攻击比针对性攻击更容易实施,因为它有更多的选择和空间来改变输出。

3) 对抗性可迁移性对抗性可迁移性是指对抗性
示例能够继续影响除创建它所使用的模型之外的其他模型的能力。它对于黑盒攻击至关重要,因为黑盒攻击时可能无法访问目标模型、训练数据集和其他学习参数。在这种情况下,可以训练一个替代神经网络模型,然后针对替代模型生成对抗性示例。由于可迁移性,目标模型将容易受到这些对抗性示例的攻击。对抗性示例的可迁移性可以从易到难定义,根据具有相同的神经网络架构但不同的数据集或从一开始就具有不同的神经网络架构 [11]。

4) 对抗性扰动对抗性扰动是一种破坏,可以以
高概率欺骗特定图像上的给定模型。小扰动是对抗性示例的核心前提。

在对抗性机器学习领域,目标是最小化最小对抗性扰动的范数,从而使目标模型分类错误。明确地说,给定一个输入图像x,扰动向量n旨在改变 x 的标签,对应于从x到分类器决策边界的最小距离[9]:

$$\begin{aligned} &\min_{n \in Rd} \quad n2 \\ &\text{st } F(x)F(x+n) \leqslant 0 \end{aligned} \tag{2}$$

其中d是输入图像和扰动向量的维数。扰动可根据其实现范围分为以下类型。

可以根据给定的输入图像明确生成特定于图像的扰动。

无需了解给定图像的底层细节即可生成通用扰动。请注意,通用性是指扰动具有良好的可迁移性以及能够统一应用于所有输入数据的特性。尽管通用扰动使得在实际应用中更容易制造对手,但大多数现有攻击都会生成特定于图像的扰动。

其目的是朝着这个方向发展,并创建在输入样本改变时不需要重新形成的通用扰动(第六节)。

C. 实验标准

对抗性攻击对 FR 系统的性能评估基于不同的数据集和目标模型。这种评估范围使得评估对抗性攻击和量化 FR 模型的鲁棒性变得复杂。大型数据集和复杂模型通常会使攻击和防御变得更加困难。

1) 数据集LFW、

CASIA-WebFace、MegaFace、VGGFace2 和 CelebA [57] 是用于评估针对 FR 系统的对抗性攻击的最广泛使用的图像分类数据集。

2)目标模型。对手广泛攻击几个

著名的深度 FR 模型,例如 DeepFace、FaceNet、VGG-Face、DeepID、SphereFace、CosFace ArcFace、OpenFace [58]、dlib1和 LResNet100E-IR Face ID 模型。2根据这些数据集和目标模型,在以下章节中,我们将根据这些数据集和目标模型检查针对对抗性示例的 FR 模型的最新研究。

D. 先驱者

在本节中,我们回顾了几种生成对抗样本的先锋方法,包括L-BFGS [9]、快速梯度符号法(FGSM) [59]、基本和最小似然迭代类方法[54]、[60]、基于雅可比矩阵的显著性图攻击(JSMA) [61]、单像素攻击[62]、 Deep-Fool [63]、通用对抗扰动[64] 和Carlini & Wagner 攻击(C&W) [65]。这些方法几乎每一种都构成了现实世界攻击的基础

并且在实践中具有显著影响机器学习目标模型的能力。这里提供的描述将显示对抗攻击的逐步改进以及最先进的对抗攻击能够达到的程度

1http://dlib.net

2https://github.com/deepinsight/insightface/wiki/Model-Zoo

实现。我们将重点介绍一般攻击 DNN 的主要方法,并按时间顺序进行回顾,以保持讨论的流畅性。

1)L-BFGS

Szegedy等人[9] 首先使用L-BFGS方法生成对抗样本。盒约束的L-BFGS用于近似解决以下问题:

$$\begin{aligned} &\text{最小}_{x} \; c\,n2 + L(x, l) \\ &\text{stx} \in [0, 1] \end{aligned} \tag{3}$$

其中L(x , l) 计算分类器的损失,通过线搜索近似计算出最小值c > 0 以满足上述条件。作者表明,上述方法可以计算出在添加到干净图像中时会欺骗神经网络的扰动,而人眼却无法察觉。

2)快速梯度符号法(FGSM)

Goodfellow等人[59] 提出了一种快速而直接的方法,称为快速梯度符号法(FGSM),通过有效地解决以下问题来计算对抗性扰动:

$$n = \text{符号}(\nabla x J(\theta, x, l)) \tag{4}$$

其中,表示扰动幅度,sign (.)表示符号函数, ∇xJ (., ., .)表示成本函数围绕与x有关的模型参数当前值求得的梯度。生成的对抗样本x的计算公式为x = x+n。应用FGSM方法后,对抗样本不再以迭代方式计算,而是一步一步地沿每个像素的梯度符号方向进行梯度更新。Miyato等人[66]提出了一种密切相关的方法,将其命名为快速梯度L2。使用此方法,扰动计算如下:

$$n = \frac{\nabla x J(\theta, x,}{l) \; \nabla x J(\theta, x, l)2} \tag{5}$$

如图所示,计算出的梯度用其L2 范数进行归一化。Kurakin等人[67]提出了一种使用 L∞ 范数进行归一化的替代方法,称为快速梯度L∞方法。在文献中,所有这些方法都归类为一步法。

3)基本和最小似然迭代类方法Kurakin等人[54] 扩展了一步梯度上升的思想,提出

了基本迭代方法 (BIM)。BIM通过运行多个小步骤来迭代调整增加分类器损失的方向。在每次迭代中,图像像素的值被裁剪如下:

$$x^{(i+1)} = \text{剪辑}\{x^{(i)} + \alpha \cdot \text{符号}\nabla x\,(i)J\,\theta, x^{(i)}\}, \tag{6}$$

其中x (i)表示迭代中生成的对抗性示例, Clip {.}限制其在每次迭代中的变化,并且

α 是步长。BIM算法的初始化通过设置x = x来完成(0) ,其终止由 min (+ 4, 1.25) 确定的迭代次数控制。

该方法在文献中也称为迭代快速梯度符号法 (I-FGSM) 。根据此方法论,提出了迭代快速梯度值法 (I-FGVM) ,其不同之处在于它使用∇x (i)J梯度的方式 [54], [60]。具体而言， I-FGVM沿梯度方向改变输入x ,而I-FGSM仅使用符号梯度。在I-FGSM 的每次迭代中,图像像素的值按如下方式裁剪:

$$x^{(i+1)} = \text{剪辑}_{x}\left\{x^{(i)} + \alpha \cdot \nabla x\,(i) J\,\theta, x^{(i)}, l\right\} \tag{7}$$

在另一次尝试中,Kurakin等人[54] 将BIM扩展为迭代最小似然类方法 (ILCM),类似于他们将FGSM扩展为“一步目标类”。他们用分类器预测的最小似然类（比如l2)替换(6) 中图像的标签l ,并尝试最大化交叉熵损失。

4)基于雅可比显著性图攻击（JSMA）

Papernot等人[61] 通过限制扰动的 L0 范数设计了一种对抗性攻击。与扰动整个图像不同,他们计划扰动图像中的几个像素,这可能会导致输出发生重大变化。

因此,他们定义了一个显著性对抗图,称为基于雅可比矩阵的显著性图攻击 (JSMA),通过它可以监控改变干净图像的每个像素对最终分类的影响。重复提出的算法,直到对抗图像中允许的最大像素数被改变,从而成功欺骗神经网络。

5)单像素攻击Su等人[62] 提出

了一种仅通过改变每幅图像的一个像素就能欺骗不同神经网络的成功方法。

优化问题变为:

$$\begin{aligned} &\text{分钟}_{x} \quad J\theta, F(x), l \\ &\text{st } n0 \leqslant 0 \end{aligned} \tag{8}$$

为了仅修改一个像素,将0设置为 1,因此使优化问题变得困难。因此,作者应用了差分进化 [68] 的概念来寻找最优解。该技术需要目标模型预测的概率标签,而不需要有关网络参数值或梯度的任何信息。

它通过简单的进化策略实现,但却成功欺骗了网络。

6)DEEPFOOL

Moosavi-Dezfooli等人[63] 提出了一种称为DeepFool的迭代方法,用于为干净的输入图像找到最小范数对抗扰动。所提出的算法初始化时假设输入图像位于仿射决策边界所限制的区域中分类器,初步决定输入的类标签。

每次迭代时,图像都会受到一个小矢量的扰动。

它试图将生成的扰动图像引向通过线性近似图像所在区域边界而获得的边界。在每次迭代中,扰动都会添加到图像中并累积以计算最终扰动,这会根据图像区域的原始决策边界改变输入图像标签。与FGSM和JSMA相比， DeepFool已被证明可以提供较小的扰动,同时具有相似的欺骗率。

7) 通用对抗扰动与计算特定于图像的扰动的DeepFool方法不同,

Moosavi-Dezfooli等人[64] 提出了一种较新的算法来生成与图像无关的通用对抗扰动,以成功欺骗任何图像上的网络。他们试图找到满足以下约束的通用扰动:

$$P\,(F(x) = F(x + n)) \geqslant \delta \text{ st } np \leqslant \xi \tag{9}$$

其中P (.) 表示概率,δ 控制欺骗,指的是Lp 范数,ξ 限制单速率、通用扰动的大小。因此,ξ 的值的深度 .p 越小,对抗性示例对人眼越难以察觉。结果表明,通用对抗性扰动可以在流行学习架构（例如 VGG、CaffeNet、GoogLeNet、ResNet)中很好地推广。

8)卡里尼和瓦格纳的袭击（C&W）

Carlini 和 Wagner [65] 引入了一组对抗性攻击来击败防御性蒸馏。根据他们的研究,准不可察觉扰动的L0、 L1和L2 范数受到限制,导致目标网络的防御性蒸馏失败。还证明了使用未蒸馏网络生成的对抗性示例可以很好地转移到蒸馏网络,使得生成的扰动适合黑盒攻击。关于定义,蒸馏是指将更复杂网络的知识转移到较小网络的训练过程。

这一概念最初由 Hinton等人提出[69]。

后来,Papernot等人[70] 引入了该程序的变体,利用网络知识来提高其鲁棒性。

IV. 针对人脸识别的对抗性示例生成在本节中,我们回顾针对 FR 系统生成的对抗性示例。我们首先解释文献

中介绍的主要攻击生成方法。接下来,我们根据不同的攻击方向进行比较。最后,我们根据对抗能力、特异性、可转移性和扰动类型的属性再次进行比较。

A. 方法

在本节中,我们回顾了针对 FR 模型的主要对抗性示例生成方法。我们回顾了不同的研究,并在后续章节中对它们进行比较,以保持讨论的连贯性。

1)基于图像级网格遮挡的失真不特定于面部并且可以应用于任何物体图像的失真被归类为图像级失真。

Goswami等人[71] 提出了一种称为基于网格的遮挡的图像级失真。在这种方法中,根据参数 ρgrids沿图像上边界(y = 0)和左边界(x = 0)选择点P = {p1, p2, ..., pn},其中网格指的是基于网格的遮挡。ρgrids参数确定用于以更高的值改变给定图像以产生更密集的网格(即更多网格线)的网格数量。对于每个点pi = (xi, yi),选择图像对面边界上的一个点p i ,条件是i , y i ,如果yi = 0 则y i = H,如果xi = 0 则x = W,其中W ×H是输入图像大小。一旦选择了一组对点P和P ,就会创建一个像素宽的线来连接每一对。最后,将这些线上的像素设置为 0 灰度值。

= x

2) 图像级最显著位噪声 (XMSB) 失真图像级最显著位噪声是 Goswami等人引入的另一种图像级失真[71]。

在这种方法中,从图像中随机选择三组像素X1、 X2、 X3 ,使得 |Xi | = ∅i × W × H。

这里W × H是输入图像的大小,参数∅i th表示第i位被翻转的像素的比例。因此, ∅i的值越高,第 i 个最高有效位的像素失程度就越大。对于每个Pj ∈ Xi, ∀i ∈ [1, 3],执行以下操作: Pkj = Pkj ⊕ 1 (10) 真 最重要的

日

其中Pkj表示集合中的k ,⊕ 表示按位异 日 j的最高有效位 日 像素或运算。另外,需要注意的是,集合Xi可能会重叠;因此,受噪声影响的像素总数小于或等于|X1| + | X2| + |X3|,具体取决于随机选择。

3) 人脸级失真除了图像级失真之外,Goswami等人[71] 还引入了人脸级失真。这种失真显然需要特定于人脸的信息,例如面部标志的位置。因此,这种方法通常在执行自动人脸和面部标志检测后应用。一旦检测到面部标志,就会利用它们及其边界来执行遮罩步骤。

为了遮挡眼睛区域,在脸部图像上绘制一个奇异的阻挡带,如下所示:

$$我\{x, y\} = 0, \forall x \in [0, W], \quad y \in ye - deye\,\psi, ye + deye\,\psi \quad (11)$$

其中ye = (yle + yre) 2, (xle, yle)和(xre, yre)分别是左眼中心和右眼中心的位置。

deye是两眼之间的距离,计算公式为xre − xle, ψ 是决定遮挡带宽度的参数。可以使用额头和眉毛区域的面部特征点作为遮罩,以类似的趋势遮挡额头和眉毛,从而实现眼部区域遮挡(ERO)过程。还可以使用外部面部特征点和鼻子和嘴巴坐标来遮挡胡须区域,从而将遮罩创建为单独遮挡区域的组合。

4) 进化攻击Dong等[72] 提出了基于 (1+1)-CMA-ES [73] 的进化攻击方法,该方法是协方差矩阵自适应进化策略 (CMA-ES) [74] 的一个有用且直接的变体。在 (1+1)-CMA-ES 的每次更新迭代中,通过添加随机噪声从其父级 (当前解) 生成一个新的后代 (候选解),评估这两个解的目标,并选择更优的解进行下一次迭代。该方法可以解决黑箱优化问题:

$$\min_{x} L(x) = x - x \quad {}_{2} + \delta CF\ (x) = 1 \quad (12)$$

其中 C(.) 是对抗标准,如果满足攻击要求则取 1,否则取 0;如果a为真则 δ (a)取 0,否则取 +∞。然而,由于x的维度较高,作者没有应用 (1+1)-CMA-ES 来优化 (12) 。

为了加速该算法,他们提出了一种适当的分布来在每次迭代中对随机噪声进行采样,这可以对搜索方向的局部几何形状进行建模。他们从有偏高斯分布中采样随机噪声,以最小化采样的对抗图像与原始图像的距离。这个添加的偏差项是一个关键的超参数,它控制着靠近原始图像的强度。作者还针对这个问题的特点提出了一些降低搜索空间维数的技术。他们在低维空间 R 中采样随机噪声,其中m < d,其中d是输入空间的维数。然后他们采用了一个上采样算子,确切地说是双线性插值法,将噪声向量投影到原始空间。因此,输入图像的维度得以保留,而搜索空间的维数则减少了。

5)特征快速和迭代攻击方法给定一对人脸和一个深度人脸模型,[75] 提出了特征级攻击,通过计算它们的归一化深度表示之间的距离来比较这对人脸。

这些表示类似于嵌入特征,只是它们是从深度人脸模型中归一化和提取的。为了发现深度人脸模型的脆弱性,作者提出在其中一张人脸图像上添加扰动来生成对抗性示例并欺骗人脸模型。根据他们的概念,正负

定义了人脸对,其对应的输出标签分别相同和不同。用x 1, x表示人脸对,用x = x + n, 2表示对抗样本,对于正人脸对， l = l ,优化的目标函数和损失函数公式如下：

$$\begin{aligned} & n=\max_{n} F(x^{1}+n)-F(x^{2})_{2}, n_{\infty}<\varepsilon \\ & J(x^{1}+n, x^{2})=F(x^{1}+n)-F(x^{2})_{2} \end{aligned} \tag{13}$$

而对于负样本对{x }， l优化目标和损失函数公式如下：

$$\begin{aligned} & n=\max_{n} F(x^{1}+n)-F(x^{2})_{2}, n_{\infty}<\varepsilon \\ & J(x^{1}+n, x^{2})=-F(x^{1}+n)-F(x^{2})_{2} \end{aligned} \tag{14}$$

其中F(x和 ε 限)表示归一化后的深度表示
制扰动的最大偏差。
基于（13)和（14)的损失函数形成对抗性扰动称为特征快速攻击方法（FFM） ,定义为：

$$x^{1}+n=Gx^{1},\varepsilon x^{1}+符号\nabla x^{1}J(x^{1},x^{2}) \tag{15}$$

考虑到迭代方式,作者提出了
迭代攻击方法（FIM)的特点是：

$$\begin{aligned} & n0=0 \\ & gN+1=\nabla x^{1}+nN\,J(x^{1+nN},x^{2}) \\ & nN+1=(x^{1}+nN,符号（gN+1）) \end{aligned} \tag{16}$$

其中Gx,ε(x) = min(255, x + ε, max(0, x − ε, x));可以启发式地选择迭代min(ε + 4, 1.25ε)。

6) 眼镜配饰打印Sharif等人[76] 提出了一种在数字环境
中可物理实现的冒充或躲避攻击。为了实现物理可实现性,第一步是通过 3D 甚至 2D 打印技术,仅使用面部配饰（具体来说,眼镜架)实施攻击。具体来说,他们使用了一种特定的现成的眼镜架数字模型,并利用商用喷墨打印机（Epson XP-830)将眼镜架的前平面打印在光面纸上,然后将光面纸贴在真正的眼镜架上。对齐后,眼镜架占据了 224 × 224 人脸图像像素的约 6.5%,这意味着攻击最多会扰乱图像中 6.5% 的像素。为了找到实现冒充或躲避所需眼镜架的颜色,将它们的颜色初始化为纯色（例如黄色） ,然后将眼镜架渲染到主体图像上。它们的颜色通过梯度下降过程迭代更新,以产生能够容忍佩戴镜架时轻微自然运动的对抗性扰动。

第二步涉及调整攻击者目标的数学公式,以关注对抗性扰动,这些扰动既能对观看条件的微小变化具有鲁棒性,又能像自然图像一样平滑。

为了找到独立于精确成像条件的扰动,旨在增强扰动的普遍性,作者寻找可能导致一组输入中的任何图像被错误分类的扰动。为此,攻击者收集一组图像X,并找到一个可以优化其目标的扰动,用于每个图像x ∈ X。对于模仿,这被形式化为以下优化问题（躲避是类似的）：

$$\max_{n} \text{精氨酸} \sum_{x\in X} \text{softmaxloss}(F(x+n), l) \tag{17}$$

其中n表示扰动。为了保持扰动的平滑度,优化过程进行了更新,以最小化总变差 (TV) [77],其定义为：

$$电视（n）=\sum_{我,} (n_{i,j}-n_{i+1,j})^{2}+(n_{i,j}-n_{i,j+1})^{2})^{1/2} \tag{18}$$

其中ni,j表示n中坐标(i, j)处的像素。当相邻像素的值彼此接近（即扰动平滑)时， TV (n)较低,否则较高。因此,通过最小化TV (n),扰动图像的平滑度以及物理可实现性得到改善。

7)基于可见光的攻击（VLA）

Shen等人[78] 提出了一种针对 FR 系统的基于可见光的攻击(VLA) ,其中制作基于可见光的对抗性扰动并投射到人脸上。

对于每个对抗性示例,作者建议生成一个扰动框和一个隐藏框,它们被投射到用户的脸上。扰动框包含有关如何将输入用户的面部特征更改为目标用户或非目标用户的特征的信息,而隐藏框旨在隐藏扰动框中的扰动,不让人类的眼睛观察到。

对于扰动帧的生成,该方法将像素级的图像修改扩大到区域级,以避免物理场景中可能出现的扰动损失。

相应地,根据包含颜色值的相似性将扰动帧划分为排他范围。

Manshift 聚类将所有颜色进行划分,将邻近的相似颜色划分为相同的区域,将图像中每组具有相同颜色的邻近像素视为一个扰动区域。然后,在第二步中,采用区域过滤策略,确保相机能够成功捕捉到扰动帧内的所有投影细节,并且不会在物理场景中捕获的图像中丢失小的颜色区域。设n = x − x为扰动帧， n的聚类和过滤结果表示为Cx,x ，定义如下：

$$Cx,x=\{Gi\,(p)\,,\,Ri\mid 0\leqslant i\leqslant m\} \tag{19}$$

其中， Gi (p)表示像素p的颜色是否应该设置为Ri ， m为颜色区域总数。对于

图像Cx,x中的每个像素p ,如果p位于Ri内,则Gi (p)为 1,否则为 0。接下来定义生成函数H (·),将聚类结果Cx,x转换为扰动框架n,如 (20) 所示:

$$n = H\,C_{x,x} = [R_i \text{如果} G_i\,(p) = 1] \tag{20}$$

为了将扰动帧隐藏在人眼之外,根据视觉暂留 (POV) 效应生成隐藏帧 [79]。根据视觉暂留效应,两种不同的颜色频繁交换会导致人脑无法在变化发生的那一刻直接处理这些变化,从而使人眼将新颜色视为这些颜色的融合。基于这一知识,通过交替投影扰动帧和隐藏帧,即交替显示生成的图像的相应两种颜色,人眼可能难以感受到扰动帧,并且这些颜色的融合将被感知为图像的基色/背景色。

8) ADVHAT 攻击Komkov 和

Petiushko [80] 提出了一种可重现的对抗攻击生成方法,称为AdvHat。他们在标准彩色打印机上打印了一张矩形纸贴纸,并使用离平面变换算法将其贴在帽子上。所提出的算法分为两个步骤:(1)贴纸的离平面弯曲,模拟为 3D 空间中的抛物线变换,将贴纸的每个点映射到抛物线圆柱上的新点;(2)贴纸的俯仰旋转,通过对获得的新点应用 3D 仿射变换来模拟。作者将生成的贴纸投影到高质量的人脸图像上,投影参数的扰动很小。他们将新的人脸图像转换为 ArcFace 输入的标准模板,并将其传递给优化步骤。关于优化步骤,两个参数(TV 损失和两个嵌入之间的余弦相似度)的总和按如下方式最小化,以实现用于修改贴纸图像的梯度符号:

$$\text{低温} \quad , a = L_{\text{sim}}\,x \qquad , a + \lambda \cdot TV \ (\text{补丁}) \tag{21}$$

其中LT是总损失， patch表示贴纸， x是贴有补丁的照片,λ 是 TV 损失的权重,在本文中假设为 1e−4。这里， Lsim是两个嵌入之间的余弦相似度,定义如下:

$$\text{模拟} \quad , a = \cos(e_x, e_a) \tag{22}$$

其中ex是获得的攻击者面部图像的嵌入， ea是指通过 ArcFace 计算的所需人的面部图像的嵌入。

9)惩罚快速梯度值方法（P-FGVM）

Chatzikyriakidis等人[81] 提出了一种惩罚快速梯度值方法(P-FGVM)对抗攻击技术,该技术在图像空间域上运行并生成与原始图像类似的对抗性去识别面部图像。

该技术受到I-FGVM 的启发,但在其梯度下降更新方程中结合了对抗性损失和“真实主义”损失项。

在该方法中,通过以下梯度下降更新方程生成有针对性的对抗示例x：

$$x^{(i+1)} = x^{(i)} - \alpha \cdot \nabla x\,(i) J\,\theta, x_x^{(i)} \quad , \text{丹}$$

$$\text{世）} + \lambda\,x - x^{(-} \tag{23}$$

其中 λ 是权重系数， (x (i) − x) 是真实性
损失金额。

10) 人脸安全攻击Kwon等人[82] 提出了人脸安

全对抗样本生成方法,该方法生成的对抗样本会被敌人 FR 系统误认,但会被朋友 FR 系统正确识别,且失真程度最小。该方法由一个转换器、一个朋友分类器Mfriend和一个敌人分类器 Menemy 组成,用于生成对抗人脸图像。给定预先训练的Mfriend和Menemy以及原始输入x ∈ X,生成对抗人脸样本x的优化问题如下:

$$\text{酸}_{x} \quad L(x, x)\text{精氨}$$

$$\text{st } g_{\text{friend}}(x) = l \text{ 和} g_{\text{enemy}}(x) = l \text{其中} g_{\text{friend}}(x) \text{和} \tag{24}$$

g enemy (x)分别表示朋友分类器Mfriend和敌人分类器Menemy的操作函数。L (.) 是人脸原始样本x和人脸变换示例x之间测量的距离。

Transformer 生成对抗性面部示例x ,采用原始样本x及其对应的输出标签。

Mfriend和Menemy对x的分类损失返回到 Transformer,然后 Transformer 计算总损失 LT ,并重复上述过程以生成对抗性面部示例x ,同时最小化LT 。

总损失定义如
下:

$$LT = L_{\text{friend}} + L_{\text{enemy}} + L_{\text{distortion}} \tag{25}$$

其中Lfriend是Mfriend的分类损失函数， Lenemy是Menemy的分类损失函数， Ldistortion是变换后示例的扭曲,定义为x和x之间的距离。

11) 快速地标操作 (FLM) 方法Dabouei等人[83] 提出了一种快速地标操作方法来

制作对抗性面孔。他们提出通过空间变换原始图像来生成对抗性示例。使用地标检测函数将人脸图像x映射到一组k 个2D 地标位置P = {p1, . . . , pk }, pi = (ui, vi),假设p i是pi的变换版本,位于相应对抗性图像x中。

,

$$\text{义i} \qquad = (u\,\text{th}\text{并}_i, \text{我}) \qquad \text{定} \ \text{地标}$$

为了根据 P 来
处理人脸图像,每个地标流

（位移） f被定义为产生相应对抗性地标的位置。因此,可以从原始地标pi和优化的特定位移向量fi = (ui, vi)获得对抗性地标p i ,如下所示：

$$= \quad = pi + fi\ p\ i\ )$$

$$(你 \quad 维\ (ui + ui, vi + vi) \qquad (26)$$

与参考文献 [84] 相比,该文献通过为输入图像中的所有像素位置定义场f来实现此目的,而Dabouei等人[83] 仅为k 个特征点定义了场 f,与输入图像中的像素数量相比,这个数字非常小,尤其是在实际应用中,如 FR 问题。这种有限的控制点数量还减少了空间变换引入的失真。使用变换T将良性人脸图像空间变换为对抗性人脸图像,如下所示：

,

$$x = T\ (P, P\ , x) \qquad (27)$$

其中P表示目标控制点。结合 softmax 成本作为正确分类的衡量标准,作者将生成对抗性面孔的总损失定义为：

$$L\ P, P\ , x, l = \text{softmaxloss}\ F\ T\ P, P \qquad , X \qquad , \text{卉}$$

$$— \lambda\ 流量\ L\ 流量 P — P \quad (28)$$

其中λflow是用于控制位移幅度的正系数， Lflow是用于限制位移场的项。这样,地标位移场f可以使用预测的梯度方向迭代找到,称为FLM方法。作者还扩展了这种方法,提出了分组快速地标操作(GFLM)方法,该方法对地标进行语义分组并操作组属性,而不是扰动每个地标。这个想法是为了解决FLM生成的对抗性面孔的严重扭曲并保留所创建图像的整个结构而形成的。

B. 不同对手的取向比较

图 3 给出了考虑对手方向的针对 FR 系统的现有对抗性示例生成技术的一般分类。根据不同研究中所采用的策略或用于发起对抗性攻击的工具,不同的技术主要可分为四类,即 (1) 面向 CNN 模型;(2) 面向物理攻击;(3) 面向几何。本节的其余部分按照此分类进行构建。

1)以 CNN 模型为导向

如前所述,深度学习范式在 FR 任务中得到了显著传播。一些模型是基于深度 CNN 的架构,具有许多隐藏层和

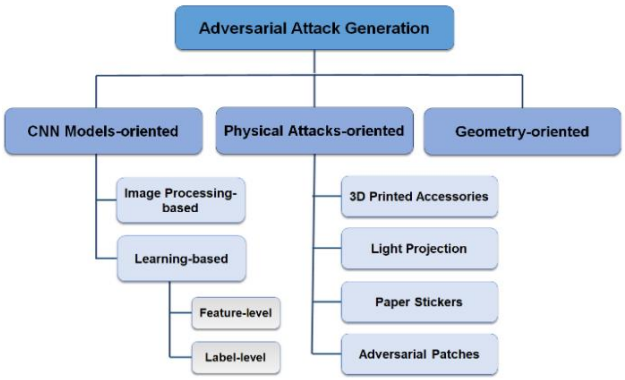

图 3.旨在欺骗 FR 系统的对抗性攻击生成方法的广泛分类。

数百万个参数,这些模型旨在在不同数据库上测试时实现非常高的准确率。尽管据报道此类模型的效率在逐步提高,但它们被证明容易受到对抗性攻击。意识到这一点,许多研究人员已经开始设计方法来利用此类算法的弱点。

Goswami等人[71] 考虑了几种基于深度 CNN 的 FR 算法在图像处理失真（1)图像级别和（2)人脸级别)的情况下的脆弱性。他们证实,对系统的攻击不需要基于复杂的学习。相反,随机噪声,甚至是人脸图像中绘制的水平和垂直黑色网格线,都会严重降低人脸验证的准确性。

图 4 描述了此项努力的例子。

图 4.干净的输入图像（a)通过基于图像处理的 xMSB 扭曲（b）、基于网格的遮挡（c）、前额和眉毛遮挡（FHBO）(d）、眼部区域遮挡（ERO）(e)和面包状遮挡(f)修改而来[71]。

Dong等人[72] 提出了进化攻击算法,以评估多个高级 FR 模型在基于决策的攻击环境中针对标签级对抗性示例的鲁棒性。

Zhong 和 Deng [85] 定义了Dropout Face Attacking Networks (DFANet)技术来探索深度 CNN 针对特征级对抗性示例的脆弱性。

他们在卷积层中加入了 dropout

对抗生成过程的迭代步骤,以提高对抗示例的可迁移性。具体来说,对于由卷积层组成的人脸模型,给定第i 个卷积层的输出,他们提出生成一个掩码,其元素独立地从伯努利分布中采样。然后利用该掩码来

日

修改其中第i 个乘积的输出。作者 日 通过 Hadamard 实现的卷积层提出将此方法应用于FIM的生成,并将其与可迁移性增强方法相结合 [86]–[88]。他们在 LFW 数据集上进行了实践,生成了一组新的对抗性人脸对,以攻击亚马逊、微软、百度和 Face++ 的商业 API。

6 它提供高度准确的面部分析和面部搜索功能,可检测、分析和比较各种应用的面部。他们向公众开放了这个 TALFW 数据库,以供将来研究。

Garofalo等人[89] 专注于人脸认证系统的安全性,旨在让冒名顶替者逃避 FR 模型。作者对基于 OpenFace FR 框架的认证器部署了投毒攻击,该框架通过支持向量机 (SVM) 分类器进行了扩展。他们针对底层 SVM 模型实施了攻击,以对 FaceNet 模型提取的人脸模板进行分类。在另一项具有类似目的的研究中,Chatzikyriakidis等人[81] 提出在人脸去识别的情况下使用对抗性示例。他们针对基于 CNN 的人脸分类器引入了P-FGVM对抗性攻击技术。图 5 显示了实施此方法生成对抗性图像的示例。

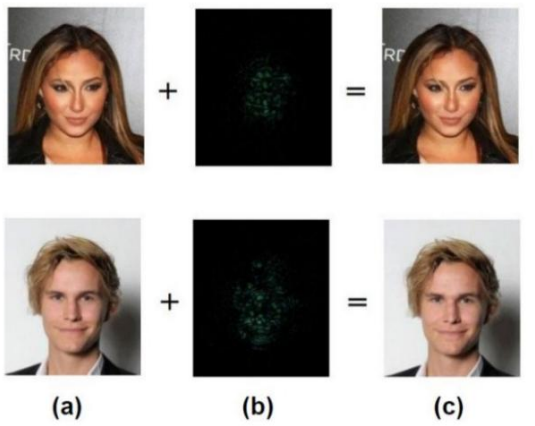

图 5.通过对抗攻击方法 P-FGVM [81] 将干净的面部图像 (a) 通过对抗扰动 (b) 修改,以生成去识别的面部图像 (c)。扰动的绝对值被放大了 10 倍。

最近,Kwon等人[82] 提出了Face Friend-safe对抗样本生成方法,成功地

3https://aws.amazon.com/rekognition
4https://azure.microsoft.com
5 https://ai.baidu.com https://
6 www.faceplusplus.com.cn

误导敌方 FR 系统,但能够被友方 FR 系统正确识别。

最近,提出了一种基于 Python 的新工具箱 Advbox,用于生成对抗性示例 [90]。使用 Advbox,可以欺骗 PaddlePad-dle、PyTorch、Caffe2、MxNet、Keras 和 TensorFlow 中的神经网络,并具有对 ML 模型的稳健性进行基准测试的附加功能。与以前的研究相比,该平台支持实际的攻击场景,例如 FR 攻击。

2) 面向物理攻击面部生物识别系统的入侵者

经常会遇到两种挑战:(1) 他们无法精确控制面部识别系统的（数字)输入;相反,他们可能能够控制自己的外表;(2) 当他们通过操纵外表来逃避识别时,例如化浓妆,可能会被警察等传统手段轻易发现。面对这些挑战,出现了一种基于攻击者身体状况的新型对抗性攻击。

Sharif等人[76] 开发了眼镜配件打印方法,以生成一类物理上可实现但不显眼的攻击。在 [91] 中,作者提出了对抗生成网络(AGN)来生成会导致错误分类的制品（例如眼镜)图像。此类神经网络生成的制品类似于一组参考制品（例如真实的眼镜设计）,并满足不显眼的目标。与 GAN 类似， AGN经过对抗性训练,以学习如何生成逼真的图像。与 GAN 不同的是， AGN还经过训练以生成对抗性输出,这些输出可以在数字和物理层面误导给定的 FR 模型以达到逃避目的。在本研究中,FR 算法在数字层面上受到传统攻击,例如 Szegedy 的L-BFGS方法 [9],并在物理层面上受到欺骗,要求个人佩戴 3D 打印的太阳镜框架。图 6 说明了通过佩戴此类配件来产生模仿攻击。

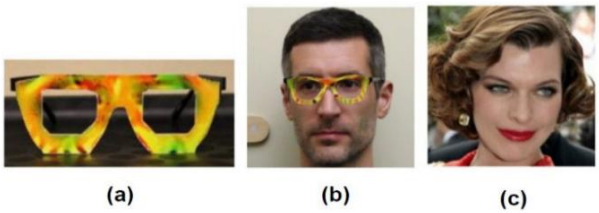

图 6.卢乔·鲍尔 (b) 使用该眼镜框 (a) 模仿米拉·乔沃维奇 (c) [76]。

Zhou等人[92] 设计了一顶帽子,帽子顶部装有几枚一分钱大小的红外 LED,可通过红外点照射佩戴者的脸部来产生不显眼的物理对抗攻击。本研究中的损失通过根据攻击者照片上的模型调整光点来优化。然后,攻击者可以通过调整点的位置、大小和强度来逃避检测。

受相机和人眼成像原理差异的启发,Shen等人[78] 提出了针对 FR 模型的VLA攻击。在一项类似的研究中,Nguyen等人[55] 研究了使用网络摄像头和投影仪通过对抗性光投影对 FR 系统进行实时物理攻击的可行性。在这种方法中,作者使用摄像头捕捉对手的面部图像,并使用一个或多个目标图像 (1) 根据攻击环境调整摄像头-投影仪设置和 (2) 创建数字对抗模式。然后使用投影仪将数字模式投影到物理域中的对手脸上以逃避识别。虽然这项工作的目标与基于红外的对抗性攻击 [92] 相同,但它不需要创建可穿戴设备,因此提供了一种更舒适的替代设置来直接对 FR 模型进行物理攻击。

另一项研究 [80] 提议通过AdvHat攻击生成方法在固定（均匀光线下的全脸照片)和可变（不同角度的脸部旋转和光线条件)设置下针对公共 Face ID 模型 LResNet100E-IR、ArcFace@ms1m-refine-v2 进行攻击。同样,Pautov等人[93] 研究了相同识别系统的安全性,并提议打印、添加（作为脸部属性)和拍摄对抗性补丁;然后将具有此类属性的个人快照传送给分类器,以将正确识别的类别更改为所需的类别。在这项工作中,补丁要么是攻击者脸部的各个部分,如鼻子或前额,要么是一些可穿戴配件,如眼镜。

3)面向几何的流行的基于强度的对抗

性攻击方法,直接操纵输入图像的强度,计算成本低,但对空间变换很敏感。

输入图像的微小旋转、平移或缩放变化都可能导致这些方法的相似性发生剧烈变化。由于这一限制,人们发起了一类新的攻击来生成基于几何的对抗性示例。

Daboueі等人[83] 提出了FLM方法,该方法可以比传统的几何攻击快近 200 倍地制作对抗性面孔。他们进一步引入了GFLM作为快速几何扰动生成算法的扩展版本。图 7 展示了所提出的基于几何的快速对抗性攻击的概览 [83]。

Song等人[94] 专注于误导 FR 网络将某人检测为目标人物的攻击,而不是不引人注意地进行错误分类。他们引入了注意力对抗攻击生成网络(A 3GN) ,以生成与原始图像相似但具有与目标人脸相同的特征表示的对抗性示例。为了捕获目标人的语义信息,他们附加了条件变分自动编码器和注意模块来学习人脸之间的实例级对应关系。

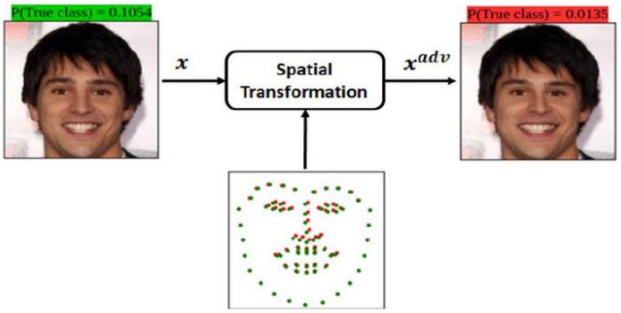

图 7.快速地标操作方法应用,生成对抗性地标位置,将地面真实图像空间转换为自然对抗性图像。如绿色和红色所示,地面真实图像被正确分类,而对抗性图像被错误分类为错误类别 [83]。

Deb等人[56]利用 GAN制作了自然人脸图像,与目标人脸图像几乎没有区别。他们提出了AdvFaces对抗人脸合成方法,以在突出的面部区域制作最小的扰动。该方法包括生成器、鉴别器和人脸匹配器,用于自动生成对抗性蒙版并将其添加到图像中以获得对抗性人脸图像。表 1 概述了不同对抗性示例生成方法的方向。

3.3 不同对手的比较

评估过程

本节从评估过程和相应利用指标的角度比较了不同的对抗性示例生成技术。

Goswami等人[71] 在 PaSC [96] 和 MEDS [97] 数据库上评估了基于 CNN 的 FR 算法（包括 OpenFace、VGG-Face、LightCNN 和 L-CSSE [95])以及一种商用现货识别器 (COTS) 在存在基于图像处理的对抗性扭曲的情况下的验证性能。他们报告了基于 1% 错误接受率 (FAR) 的攻击的真实接受率 (GAR) (%) 的实验结果。总体而言,他们证明,当数据中引入任何扭曲时,基于深度学习的算法可能会比非基于深度学习的 COTS 经历更高的性能下降。

Dong等人[72] 将进化攻击方法与所有现有的基于决策的黑盒攻击生成方法的性能进行了比较,包括边界攻击方法 [98]、基于优化的方法 [99] 和仅标签设置中的 NES 扩展（NES-LO)[100]。

在 LFW 和 MegaFace 数据集上,作者与 SphereFace、CosFace 和 ArcFace FR 模型进行了比较。对于所有方法,他们通过均方误差 (MSE) 测量对抗图像和原始图像之间的失真,以评估不同方法的性能。实验结果表明,所提出的

表 1.不同对抗攻击生成算法在方向和评估过程上的比较。

Representative study	Attack orientation	Method/Description	Evaluation metrics
[71]	CNN models	*(1) Image-level Grid-based Occlusion, (2) Image-level Most Significant Bit-based Noise (xMSB) Distortion, (3) Face-level Distortion*	GAR (%) @ 1% FAR
[72]	CNN models	*Evolutionary Attack*	MSE
[85]	CNN models	*DFANet*	Hit rate
[89]	CNN models	Poisoning attack on an authenticator based on OpenFace extended with an SVM classifier	FNR, FPR, CE
[81]	CNN models	*P-FGVM*	MSSIM
[82]	CNN models	*Face Friend-safe Attack*	SR of the enemy classifier, the accuracy of the friend classifier, and average distortion
[90]	CNN models	Advbox toolbox	
[76], [91]	Physical	*Eyeglass Accessory Printing*	SR
[92]	Physical	Physical adversarial example generation via an infrared LEDs-equipped cap	L_2 distance between feature vectors of given pair of faces
[78]	Physical	*VLA*	SR
[55]	Physical	Physical adversarial example generation via real-time light projection	Similarity score threshold @ 0.01% FAR
[80]	Physical	*AdvHat attack*	Baseline similarity and final similarity
[93]	Physical	Adversarial example generation by printing, adding, and photographing adversarial patches of nose, forehead, and eyeglasses of the attacker	Cosine similarity between embeddings of the given pair of faces
[83]	Geometric	*FLM*	SR, computation time
[94]	Geometric	Adversarial example generation via A^3GN	Physical likeness, similarity score, recognition accuracy
[56]	Geometric	Adversarial face generation via *AdvFaces* method	SR, SSIM

与其他方法相比,该方法可以在两个任务（即人脸验证和识别）、两种攻击设置（即躲避和模仿)和所有面部模型中始终如一地实现更快的收敛和更小的扭曲。

钟和邓 [85] 评估了在 CASIA-WebFace、MS-Celeb-1M、VGGFace2 和 IMDb-Face [101] 四个数据集上训练的 ResNet-50 模型之间的针对性攻击的可迁移性。他们将攻击的目标定义为从源图像生成对抗性示例,并计划获得比 FR 系统的距离阈值更接近目标图像的人脸嵌入表示的源图像。因此,他们计算了归一化深度特征的欧几里得距离以获得 ROC 曲线,并确定了判断一对源/目标图像是正面还是负面的距离阈值。

在本研究中,当源图像和目标之间的嵌入距离小于阈值时,攻击被定义为成功（命中）。作者使用快速目标梯度符号法（FTGSM） [41]和迭代目标梯度符号法（ITGSM） [41]生成标签级对抗样本,使用FFM和FIM生成特征级对抗样本。由于在可迁移性方面更有效,作者选择FIM作为基线方法,并通过结合可迁移性增强方法进一步改进它[86]-[88]。然后将创建的强基线方法与提出的DFANet方法进行比较。基于比较,作者验证了DFANet方法的优越性,并且该方法生成的对抗样本的大多数成功命中率在四种深度 FR 模型之间,该方法的准确性可以提高到约 90%。

Garofalo等人[89] 利用 Facescrub 数据集 [102] 进行了深入评估,因为该数据集提供了大量身份和每个身份的样本。他们通过假阴性率 (FNR)、假阳性率 (FPR) 和分类错误率 (CE) 描述了身份验证器的强度。实验结果表明,在最成功的攻击中,平均 CE 可以达到令人印象深刻的 40.11%,与非目标系统相比,平均身份验证错误率增加了近 37%,而平均 FPR 增加了 40% 以上。此外,最成功的攻击部署导致测试集上的 CE 达到 51.23%,使人脸身份验证系统完全失效。

Chatzikyriakidis等人[81] 在两个基于 CNN 的人脸分类器上评估了所提出的P-FGVM方法：(1)一个简单的架构模型；(2)基于预训练的 VGG-Face CNN 描述符的迁移学习微调模型,使用 VGG-16 架构 [43]。他们计算了去识别化图像和原始人脸图像之间的平均结构相似性指数 (MSSIM) 以及对抗性扰动的L2范数作为衡量结果视觉质量的指标。与基线I-FGVM和I-FGSM方法相比,针对上述人脸分类器和 CelebA 数据集的子集,作者证明所提出的方法可以生成更接近原始图像的去识别化图像,同时具有比竞争方法更好的误分类误差（误分类误差分别增加了 3% 和 1.7%）。

与I-FGVM和I-FGSM方法相比,错误分类率更高)。

Kwon等人[82] 将 FaceNet 识别系统视为目标模型,在 VGGFace2 上训练方法,并在 LFW 数据集上进行测试。作者通过测量敌人分类器的攻击成功率 (SR)、朋友分类器的准确率和平均失真度来评估所提方法的效率,结果分别为 92.2%、91.4% 和 64.22。

通过报告这些数值,他们声称他们的工作目标已经成功实现。

Sharif等人[76] 在数字环境和物理可实现性实验中评估了他们的对抗性示例生成方法。他们将攻击的 SR 测量为实现目标的尝试比例。为了计算超出单个图像的统计数据,他们对每个受试者的三张图像进行了每次攻击,并报告了这些图像的平均 SR。在数字环境实验中,在白盒场景下攻击不同的 DNN,攻击者几乎在所有尝试中都能够躲避识别或冒充目标,平均 SR 为 100%。在物理可实现性实验中,要求受试者戴上眼镜框,然后拍摄他们的图像,前三位作者参与其中,每人考虑了五个环节。在第一个环节中,受试者没有戴眼镜框,非对抗性图像被正确分类,分类尝试中正确类别的平均概率高于 0.85。在第二和第三个环节中,他们戴上眼镜框试图躲避 DNN。考虑到不同情况,分配给受试者类别的平均概率从 0.85 以上显著下降到 0.03 以下。这相当于实现 100% 的 SR(除了一个实验,其 SR 为 97.22%)。在第四和第五个环节中,受试者戴上镜框,试图模仿 DNN。

考虑到不同情况,这些会话中收集的图像中有超过 87.87% 被 DNN 错误分类(目标的平均概率大于 0.75)。

在 [91] 中,Sharif等人评估了针对 VGG-Face 和 OpenFace 模型的躲避和冒充攻击。在评估阶段,他们报告了 DNN 和 SR 的攻击准确率。使用AGN,在数字领域,所有尝试均成功,所有躲避情况下的平均 SR 为 100%,所有冒充攻击的平均 SR 大于 88%。

在物理可实现性实验中,对于躲避攻击,作者报告称AGN在最坏和最好情况下的 SR 分别为 81% 和 100%,分配给正确类别的平均概率分别为 0.40 和 0.01。对于冒充攻击,他们报告称AGN 的SR 为 53%,分配给目标的平均概率为 0.22。

Zhou等人[92] 在 LFW 数据集上检验了他们提出的技术对 FaceNet 模型的有效性。他们使用L2距离来加权模型生成的两个特征向量之间的距离,并采用在 LFW 数据集上,阈值为 1.242。这样,距离低于阈值的一对人脸被识别为来自同一个人,否则就是两个不同的人。

作者观察到原始距离,即发起攻击之前攻击者和受害者之间的嵌入距离都高于阈值。

因此,身份验证系统可以识别出相应照片中没有受害者。另一方面,该算法可能导致对抗性示例在理论上使距离低于阈值。在这项工作中,理论距离是指计算出的对抗性示例与受害者之间的距离。更重要的是,作者证明了攻击者确实可以使用所提出的设备来实现这些对抗性示例,从而欺骗身份验证系统。他们通过测量攻击后低于阈值的距离来验证这一点。

Shen等人[78] 在 CusFace [78] 和 LFW 数据集上以及 FaceNet、SphereFace 和 dlib 模型上进行了广泛的实验。作者分别使用FGSM和VLA方法生成对抗样本。

在 FaceNet 模型上,他们证明了对于物理场景中的非针对性攻击, VLA可以显著提高 SR 优于FGSM。然而,对于针对性攻击,所提出的方法可以实现合理的 SR。实验结果表明, VLA生成的扰动帧中的区域级颜色区域更稳健,有助于获得更有效的对抗性示例。

生成的对抗样本也用于评估 SphereFace 和 dlib 等其他人脸识别器的性能。FGSM 的结果表明,针对 dlib 和 SphereFace 的攻击SR 低于针对 FaceNet 的攻击 SR,因为FGSM是一种白盒方法,针对 FaceNet 的对抗样本可能不适合其他识别器。然而,由于 VLA对人脸识别器不敏感,因此它可以对这三个识别器表现出类似的性能。

Nguyen等人[55] 针对 FaceNet、SphereFace 和一款商用现成 FR 系统评估了他们的方法,并确认了这些模型容易受到光投影攻击。他们使用与 FAR 0.01% 对应的相似度得分阈值来确定攻击是否成功。在对真人受试者和 FaceNet 系统进行模仿和混淆实验后,作者报告的最高 SR 分别为 93.3% 和 100%。而针对商用现成 FR 系统的 SR 值最低,这表明深度 FR 系统更容易受到生成的攻击。

Komkov 和 Petiushko [80] 评估了成功与否
在固定和可变条件下攻击的特征。在 CASIA-WebFace 数据集上,他们验证了他们的方法很容易混淆 LResNet100E-IR Face ID 模型。作为评估指标,他们探索了基线相似度和最终相似度,他们将其定义为带帽子照片的真实嵌入和嵌入之间的余弦相似度,以及真实嵌入和真实嵌入之间的余弦相似度

分别对带有对抗性贴纸的照片进行嵌入和嵌入。在固定条件的实验中,他们观察到对抗性贴纸可以显著降低与地面真实类的相似度。在各种条件下的实验中,旨在检查所提出的方法对不同拍摄条件的稳健性,尽管最终相似度在每种情况/条件下都有所增加,但观察到的攻击有效,并且几乎所有最终相似度都低于基线相似度。

Pautov等人[93] 在 CASIA-WebFace 数据集和本文第一作者和第二作者的照片上针对 Arc-Face 评估了他们的方法。他们表明,使用简单的攻击技术,他们可以在数字和物理世界中欺骗 FR 系统。实验结果表明,尽管攻击者照片对应的嵌入与应用的具有地面真实类别的补丁的相似度可以略低于该嵌入与所需类别的相似度,但 FR 模型无法将攻击者识别为地面真实类别。

作者还发现,补丁的位置及其大小极大地影响了物理领域攻击的成功。

Dabouei等人[83] 评估了所提出的FLM和GFLM方法在白盒攻击场景中的性能。他们在 VGGFace2 和 CASIA-WebFace 数据集上训练了 FaceNet 模型,并在 CASIA-WebFace 数据集上评估了其性能。作者设计了几个实验来研究面部不同区域的重要性。从结果中,他们观察到,通过这些方法引导攻击,可以实现 99.86% 以上的 SR。这些算法的计算时间也很明显。观察到 FLM和GFLM生成对抗面孔的平均时间分别为 125 和 254 毫秒,这比 stAdv [84] 的计算时间(平均 27.177 秒)短得多。

Song等人[94] 通过在 CASIA-WebFace 上训练模型并在 LFW 数据集上对其进行评估来检验所提出的方法。他们将自己的方法与 stAdv 和GFLM方法进行了比较,并观察到通过他们提出的方法可以存档令人满意的攻击 SR。总体而言,作者通过一组评估标准证明了A 3GN在身体相似性、相似度得分和不同目标面部识别准确度方面的出色表现。

Deb等人[56] 通过攻击 SR 和结构相似性指数 (SSIM) 量化了他们提出的对抗性示例生成方法的有效性。作者在 CASIA-WebFace 上训练了AdvFaces ,并在 LFW 上进行了测试。

他们发现,与FGSM、 PGD、 A 3GN和GFLM等最先进的对抗性示例生成方法相比,AdvFaces可以生成与测试图像相似的对抗性面孔,并与图库图像进行匹配。在避开最先进的 FR 模型(FaceNet、SphereFace、ArcFace)和两个商用现成机器(COTS-A 和 COTS-B)的同时,生成的图像的攻击 SR 被证明分别高达 97.22% 和 24.30%(用于混淆和冒充攻击)。他们报告了对抗图像和测试图像之间的结构相似性以及生成单个对抗图像所需的时间,并证明使用他们提出的AdvFaces方法,可以分别实现 0.01 秒的计算时间和 0.95 和 0.92 的 MSSIM(用于混淆和冒充攻击)。报告的 SSIM 值和计算时间分别高于和低于其他方法,表明 AdvFaces 方法优于它们。表 1 的最后一列列出了所审查研究中使用的不同评估指标。

D. 不同对手的属性比较

本节从容量、特异性、可转移性和所采用的扰动类型等攻击属性方面比较了不同的对抗性示例生成技术。

1)容量表2总结了两个主

要属性信息,即容量和攻击方法的特殊性。关于容量属性,我们发现大多数攻击生成技术都是白盒攻击。在黑盒攻击场景中,Dong等人[72] 专注于 CNN 模型导向,考虑了一种基于决策的黑盒攻击设置,并证明他们的方法可以快速收敛并通过精细的扭曲欺骗目标模型。在 [85] 中,针对商业 API 生成了有效的黑盒对抗攻击,并进一步探索了特征级对抗示例对基于深度 CNN 的 FR 模型的可迁移性(第 IV-B.1 节)。

Goodman等人[90] 提出了 Advbox 工具箱,展示了其支持针对 FR 系统的黑盒攻击的能力。在物理攻击方面,[78] 中的作者提出了针对黑盒 FR 系统的VLA 。

Nguyen等人[55] 专注于实时光投影攻击,同时考虑了白盒和黑盒攻击设置。在面向几何的攻击中,Deb等人[56] 证明,AdvFaces对抗人脸合成方法生成的人脸可以逃避多种黑盒当代人脸匹配技术,同时实现前所未有的攻击 SR。

2) 特殊性考虑到对抗性示例

生成技术的特殊性,表 2 表示大多数攻击方法都是有针对性的,也是无针对性的。因此,实际上考虑了关于此属性的泛化。

在更容易实施的非针对性攻击场景中,Garofalo等人[89] 专注于投毒攻击设计,Komkov 和 Petiushko [80] 则专注于在屏幕上投影纸质贴纸以达到逃避攻击的目的。

表 2.不同对抗性攻击对容量和特异性属性的比较。

Representative study	Adversarial capacity	Adversarial Specificity
[71]	None	None
[72]	Black-box	Both
[85]	Black-box	Targeted
[89]	White-box	Non-targeted
[81]	White-box	Targeted
[82]	White-box	Targeted
[90]	Both	Both
[76], [91]	White-box	Both
[92]	White-box	Both
[78]	Black-box	Both
[55]	Both	Both
[80]	White-box	Non-targeted
[93]	White-box	Both
[83]	White-box	Non-targeted
[94]	White-box	Targeted
[56]	Black-box	Both

hats 和 Dabouei等人[83] 优先考虑基于地标的对抗性示例生成算法的速度。

3)可迁移性一些研究探索了攻击方法

的可迁移性[56]、[80]、[85]、[91]。钟和邓[85]探讨了基于CNN的FR模型对可迁移攻击的脆弱性。他们观察到,他们提出的DFANet技术可以增强现有攻击方法的可迁移性。

Sharif等人[91] 发现,针对 OpenFace 架构的攻击只能在有限次数（10% 到 12%)的尝试中成功欺骗 VGG 架构,而躲避 VGG 的攻击至少有 63% 的尝试可以成功躲避 OpenFace。他们还认为,生成的通用攻击可以在架构之间迁移,并且成功率相似。Komkov 和 Petiushko [80] 使用他们提出的可复现的AdvHat方法证明,纸质贴纸在帽子上的投影很容易混淆 Face ID 模型 LResNet100E-IR。他们表示,提出的方法可以迁移到取自 InsightFace Model Zoo[7] 的其他 Face ID 模型,与 LResNet100E-IR 相比,这些模型具有不同的架构、损失函数和训练数据集。Deb等人[56] 验证了使用他们的 AdvFaces对抗性人脸合成方法生成的人脸与模型无关且可转移,并且可以逃避几种黑盒新人脸匹配技术。

4) 扰动虽然通用扰动使得在实际

应用中更容易制造对手,但本文中除一种攻击方法外,所有方法都已证明可以生成特定于图像的扰动。在 [89] 中,作者使用少量眼镜生成通用躲避,许多受试者可以使用它来逃避识别。尽管针对 FR 模型的通用扰动生成似乎是一条潜在的研究路径,值得投资以避免在输入样本发生改变时重新产生噪声（第 VI-D 节）。

V. 对抗性示例的防御

随着提出制作对抗性示例的新方法,研究也针对对抗对手,旨在缓和他们对目标深度网络性能的影响。因此,已经定义了几种防御策略来提高处于危险中的 FR 模型的安全性。

A. 防御目标

防御战略目标一般可分为以下几类:

在构建任何针对对抗样本的防御技术时,模型架构保存是首要考虑因素。为此,应尽量减少对模型架构的改动。

精度维护是考虑的主要因素
保持分类输出几乎不受影响。

模型速度保持是另一个在大型数据集上部署防御技术进行测试时不应受到影响的因素。

B. 防御战略

一般来说,对抗攻击的防御策略可以分为三类:(1)在学习过程中改变训练,例如,通过将对抗性示例注入训练数据或在整个测试过程中合并改变的输入,(2)改变网络,例如,通过改变层数、子网络、损失和激活函数,以及 (3)通过外部网络补充主模型,以关联对看不见的样本进行分类。第一类方法与学习模型无关。然而,其他两类直接处理 NN 本身。“改变”网络和通过外部网络“补充”网络之间的区别在于,前者在训练过程中改变原始深度网络架构/参数。同时,后者保持原始模型完整并在测试中将外部模型附加到它。所述类别的分类也显示在图 8 中。本节的其余部分与此分类一致。

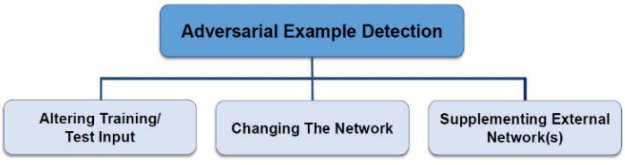

图 8.旨在保护 FR 系统免受对抗性攻击的对抗性检测方法的一般分类。

1)改变训练/测试输入Agarwal等人[13] 提出了一种

有效的对抗性检测方法来识别与图像无关的通用

[7] https://github.com/deepinsight/insightface/wiki/Model-Zoo

扰动。该方法对 (1) 像素值和 (2) 从主成分分析 (PCA) 特征获得的投影进行操作,作为测试输入,并与 SVM 分类器结合以检测扰动。由于平坦化,所提出的解决方案被视为第一类,因此改变了训练数据库的图像以形成行向量,该行向量可用作像素值或降维向量。

作者通过两种扰动算法(通用扰动算法和它的一个变体,称为快速特征欺骗 [103])评估了这种方法的有效性。他们对三个不同的数据库(MEDS、PaSC 和 Multi-PIE [104])和四种不同的 DNN 架构(VGG-16、GoogLeNet、ResNet-152 [45] 和 CaffeNet [105])进行了实验,结果表明,更直接的方法(例如本文提出的方法)可以对与图像无关的对抗性扰动产生更高的检测率。另一项研究 [106] 提出了一种基于域转换输入数据的分类集成的防御策略。根据这种方法,输入图像被转换为灰度格式,裁剪并旋转以通过分类器,分类器的预测组合起来以创建集成决策。这项研究的目的是发现一种不需要任何再训练的方法。在VGGface2数据集上,实验表明,领域变换有助于抑制对抗性攻击对人脸验证任务的影响。

2) 改变网络Goswami等人[14] 提出了两种
防御算法:(1)一种对抗性扰动检测算法,该算法利用 CNN 中间滤波器响应;(2)一种缓解算法,该算法结合了特定的 dropout 技术。在前一种算法中,作者比较了原始图像与每一层对应的失真图像的中间表示模式。他们应用这两种模式的差异来训练一个分类器,该分类器可以将看不见的输入归类为原始/失真图像。

在后者中,他们有选择地删除了 CNN 模型中受影响最严重的滤波器响应,即对噪声数据最敏感的中间层的滤波器响应,以减轻对抗性噪声的影响。随后,他们与未受影响的滤波器图进行了比较。使用 VGG-Face 和 LightCNN 网络,作者根据跨数据库协议评估了检测和缓解算法;他们仅使用 Multi-PIE 数据库进行训练,并完成了 MEDS、PaSC 和 MBGC [107] 数据库的测试。在三个数据库的所有扭曲中,结果表明,即使在低假阳性率下,所提出的检测算法也能保持较高的真阳性率,这对系统来说是理想的。此外,还观察到,通过使用所提出的缓解算法丢弃一定比例受影响最严重的中间表示,可以获得更好的识别输出。

在另一项研究中,提出了一种区块链安全机制来防御 FR 模型的攻击 [108]。任何深度学习模型(例如 CNN)的传统块都会转换为类似于区块链块的块,以在分布式环境中提供容错访问。

通过这种方式,对一个特定组件的篡改就会向整个系统发出警报,并轻松检测到“任何”可能的更改。

实验使用 Multi-PIE 和 MEDS 数据库证明了所提出的网络对深度学习模型和生物识别模板的弹性。

Su等人[109] 提出了一种深度残差生成网络(ResGN)来清除人脸验证中的对抗性扰动。他们提出了一种由ResGN、 VGG-Face 和 FaceNet组成的创新训练框架;他们提出了三种损失的联合:像素损失、纹理损失和验证损失,以优化ResGN参数。VGG-Face 和 FaceNet 网络分别通过提供纹理和验证损失来促进学习过程,从而从根本上提高清理图像的验证性能。实证结果验证了所提方法在 LFW 基准数据集上的有效性。Zhong 和 Deng [75] 提出通过将基于边缘的三重嵌入正则化 (MTER) 项集成到分类目标中来恢复表示空间的局部平滑度,从而使获得的模型学会抵抗对抗性示例。正则化项由两阶段优化组成,该优化检测可能的扰动并以迭代方式惩罚使用较大边距的扰动。在 CASIA-WebFace、VGGFace2 和 MS-Celeb-1M 上的实验结果表明,所提出的方法提高了网络对深度 FR 模型中特征级和标签级对抗攻击的鲁棒性。

根据 [110] 中探索的特征距离空间概念,Massoli等人[111] 提出了一种基于内部表征轨迹(即隐藏层的神经元激活,也称为深度特征)的检测方法。他们认为对抗性输入的表征遵循与真实输入不同的演变。

具体来说,他们在目标模型的前向传播过程中收集深度特征,对深度特征应用平均池化以在每个选定层获得单个特征向量,并计算每个向量与每层每个类的类质心之间的距离,以获得表示输入图像在特征空间中轨迹的嵌入。这样的轨迹最终被输入到二元分类器或对抗检测器。作为对抗检测器,本研究考虑了多层感知器 (MLP) 和长短期记忆 (LSTM) 网络两种不同的架构。作者在 VGGFace2 数据集和最先进的 Se-ResNet-50 [52] 上进行了实验。为了评估所提出方法的效率,他们展示了考虑每种架构的针对性和非针对性攻击的对抗检测的接收操作特性 (ROC) 曲线。他们报告了相对于每次攻击的曲线下面积 (AUC) 值。

因此,针对性攻击的 AUC 值非常接近,而对于非针对性攻击,

事实证明,LSTM 的性能明显优于 MLP。

最近,Kim等人[112] 提出了一种低功耗、高度安全的始终在线 FR 处理器,用于移动设备上的验证应用。该处理器基于三个关键特性运行:(1) 基于分支网络的早期停止 FR (BESF) 方法,可防止对抗性攻击并消耗低功耗;(2) 统一处理单元 (PE),用于点和深度卷积,并具有层融合以减少外部存储器访问;(3) 瓶颈层之间合并的噪声注入层 (NIL),使网络在外部存储器访问较少的情况下更能抵御对抗性攻击。他们证明,在 FGSM 和 PGD 下,BESF 可以实现高识别准确率,同时显著降低平均功耗。他们还表明,PE 减少了外部存储器访问,而 NIL 可以进一步减少 FGSM 和 PGD 攻击的 SR。总体而言,该处理器在 LFW 数据集的带标签人脸中实现了 95.5% 的 FR 准确率。

3) 补充外部网络Xu等人[113] 提出了一种特征压缩策略,通过将

原始空间中对应于不同特征向量的样本合并为一个样本,来调整攻击者可用的搜索空间。通过在分类器网络中添加两个外部模型,他们探索了两种特征压缩方法:(1) 降低每个像素的颜色位深度和 (2) 空间平滑。Goswami等人[14] 表示,这种方法对于具有详细数据的高分辨率图像简单有效,但对于经常在 FR 设置中使用的低分辨率裁剪人脸,它可能不起作用。在 [114] 中,提出了一个基于 Python 的开源工具箱 SmartBox,用于根据 FR 模型对对抗攻击检测和缓解算法的功能进行基准测试。此工具箱中包含的检测方法包括:“通过卷积滤波器统计进行检测”、“基于 PCA 的检测”、“伪像学习”和“自适应降噪”,它们分别属于“更改网络”、“更改训练/测试输入”和“补充外部网络”防御类别。我们将这项研究归入“补充外部网络”类别,因为它涵盖了后两种方法,因此涵盖了大多数 SmartBox 检测方法。

尽管大多数当前的防御方法要么假设预先了解特定的攻击,要么由于其潜在的假设而可能无法在复杂模型上运行良好,但通过利用 DNN 的可解释性,为对抗性检测技术打开了一扇新的窗户 [15]。

Tao等人[15]在 FR 实践中提出了一种称为攻击与可解释性(AmI)的检测技术。

该技术采用创新的双向对应推理,在面部属性和内部神经元之间进行推理,使用属性级突变和神经元强化/弱化。更准确地说,识别单个属性的关键神经元,并增强激活值以放大计算的推理部分。

相反,其他神经元的激活值被削弱,以抑制不可解释的部分。采用三个不同的数据集,VGG-Face、LFW 和 CelebA,将AmI应用于 VGG-Face,并采用七种不同的攻击。大量实验表明,所提出的技术可以成功检测对抗样本,平均真阳性率为 94%,明显高于使用称为特征压缩的最先进的参考技术所取得的成果 [113]。同样, AmI技术的 FPR 低于参考工作,证明了其在这方面的高效性。表 3 概述了不同的对抗性示例检测方法及其类别。

VI. 挑战与讨论尽管在 FR 领域已经提出和开发了几种对

抗性示例生成方法和防御策略,但仍需要解决各种问题和挑战。本节总结了威胁该领域的潜在挑战。根据上述文献,我们将挑战分为四类。

A.对抗样本的具体化/规范

如本研究所述,已经提出了多种图像、人脸和特征级对抗性示例生成方法来欺骗 FR 系统;然而,这些方法很难构建广义的对抗性示例,并且只能在某些评估指标中取得良好的性能。这些评估指标主要分为三类:生成对抗性示例的 SR、FR 模型的鲁棒性以及攻击的特定属性,例如扰动量和可迁移程度。简而言之,攻击的 SR 与扰动量成反比,被称为最直接和最有效的评估标准。

FR 模型的鲁棒性与分类准确率有关。FR 模型设计得越好,越不容易受到对抗样本的攻击。关于攻击的属性,对原始样本的扰动过小则难以构造对抗样本,而扰动过大则人眼容易分辨。

因此,对抗样本的构建与人类视觉系统之间需要取得平衡。另一方面,在一定的扰动范围内,对抗样本的迁移率与对抗样本的扰动大小成正比,即对原始样本的扰动越大,构建的对抗样本的迁移率越高。考虑到这些事实,在原始图像上需要考虑的扰动量以及模型架构的设计变得至关重要。

类似地,不同研究中所研究的成像条件的变化比实际中遇到的要小。也就是说,它们恰好是在受控的照明、距离等条件下。这些条件可以应用于一些

表 3.对抗性示例检测方法。

Representative study	Defense strategies	Description
[13]	Altering training/test input	Image pixels + PCA + SVM
[106]	Altering training/test input	An ensemble of classification results from domain transformed (grayscale, cropped and rotated) input data
[14]	Changing the network	Filter responses of CNN; dropout of filter responses
[108]	Changing the network	Conversion of traditional blocks of deep learning models into blocks similar to the blocks in the blockchain
[109]	Changing the network	Design of *ResGN* model + employment of a pixel loss, a texture loss, and a verification loss for parameter optimization
[75]	Changing the network	Integration of MTER term into the classification objective for detection and punishment of perturbations
[111]	Changing the network	Exploration of the adversary's evolution by tracking the trajectory of deep features representations
[112]	Changing the network	Design of a low-power and highly secure always-on FR processor
[113]	Supplementing external network(s)	Feature squeezing strategies of (1) pixel's color bit depth decreasing and (2) spatial smoothing via the addition of two external models to the classifier
[114]	Supplementing external network(s)	SmartBox toolbox
[15]	Supplementing external network(s)	Bi-directional correspondence inference amongst face attributes and internal neurons via *AmI* technique
[115]	Supplementing external network(s)	Defending black-box FR classifiers via iterative adversarial image purifiers

实际情况（例如,建筑物内部署的 FR 系统）。然而,其他实际情况更具挑战性,需要攻击能够容忍更广泛的成像条件。

这些问题阻碍了防御者设计通用的检测技术,并鼓励他们提出针对有限攻击的有效防御措施。为了克服这些挑战,应该考虑一个全面的实验设置,可能可以通过设计一个标准平台作为基准设置,以便同时测量所有评估指标来报告生成的对抗性示例的效率。此外,研究空间应该更多地集中在 (1) 原始图像上要考虑的扰动量、(2) 要针对的 FR 模型架构的设计,以及 (3) 生成的对抗性示例的可迁移性水平。如表 2 所示,现有 FR 模型对黑盒方式的对抗性攻击的脆弱性研究较少,表明缺乏可迁移性探索。

B. FR 模型的不稳定性

虽然深度 FR 系统的引入带来了好处,但也增加了此类系统的攻击面。

例如,实施基于图像失真的对抗性攻击,与针对相同评估数据应用基于浅层学习的商用现成匹配器相比,基于深度学习的系统的性能会大幅下降。因此,强烈建议只集成那些对逃避攻击具有鲁棒性的架构。上一段已经表达了开发鲁棒模型以提高对抗性示例的普遍性的必要性,以及其他影响因素。然而,这项义务被单独重申,以强调其在采取措施生成更多黑盒攻击时的重要性。在这种情况下,将提出开发更强大的 FR 模型的安全问题。

C. 与人类视觉系统的偏差

对视觉系统的对抗性攻击利用了这样一个事实:系统对图像中的细微变化很敏感,而人类却不敏感。开发能够推理出更类似于人类的图像的算法将是一个好主意。特别是,那些根据图像属性而不是像素强度对图像进行分类的方法可能会变得更加实用。这种方法可以训练分类器识别视觉外观中可描述方面（如性别、种族、年龄和头发颜色)的存在与否,并提取和比较面部图像中对姿势、照明、表情和其他成像条件不敏感的高级视觉特征或特性。

对人类视觉生理学的深入关注也可能为研究空间打开另一扇窗户。例如,VLA展示了物理对抗攻击的成功实施,其设计试图模拟人类的视觉系统。

D. 图像不可知扰动生成

现有的对抗性示例生成方法明显与图像无关,并且强烈地注意到缺乏针对 FR 模型的通用扰动生成。FR 模型同时攻击不同目标面部的能力将是生成通用扰动的副产品,这是在这方面进行的众多研究中的一个重要关注点。

VII. 结论本文对针对智能

深度 FR 系统的对抗性攻击过程进行了全面的概述。

尽管先进的 FR 模型性能出色,但它们很容易受到难以察觉的对抗性输入图像的影响,导致它们完全修改输出。这一事实为设计 FR 系统中的对抗性攻击和对策的众多最新贡献打开了一扇新的窗户。本文回顾了这些贡献,主要

专注于文献中最有效和最鼓舞人心的作品。根据不同的标准提出了现有攻击和防御方法的分类。我们还讨论了针对 FR 模型的对抗性示例中的当前挑战和潜在解决方案。希望这项工作能够阐明一些关键概念,以促进未来的进步。

参考

[1] MA Turk 和 AP Pentland,《使用特征脸进行人脸识别》,《 Int. J. Comput. Appl.》,第 118 卷,第 5 期,第 586-591 页,1991 年,doi: 10.5120/ 20740-3119。

[2] OM Parkhi、A. Vedaldi 和 A. Zisserman,《深度人脸识别》,载于Proc. Brit. Mach. Vis. Conf.,第 1 卷,2015 年,第 6 页。

[3] Y. Taigman、M. Yang、M. Ranzato 和 L. Wolf,《DeepFace:缩小与人类水平的人脸验证差距》, IEEE Conf. 会议论文集。
Comput. Vis. Pattern Recognit., 2014 年 6 月,第 1701-1708 页。

[4] Y. Wen, K. Zhang, Z. Li 和 Y. Qiao,“深度人脸识别的判别特征学习方法”,载于Proc. Eur. Conf. Comput. Vis., 2016 年,第 499-515 页。

[5] R. Singh、A. Agarwal、M. Singh、S. Nagpal 和 M. Vatsa,《论人脸识别算法抵御攻击和偏见的稳健性》, 2020 年, arXiv:2002.02942。 [在线]。可访问:http://arxiv.org/abs/2002.02942 [6] S. Marcel、MS Nixon 和 SZ Li,《生物特征反欺骗手册》。美国纽约州纽约:Springer-Verlag,2014 年。

[7] R. Raghavendra 和 C. Busch,“人脸识别系统的演示攻击检测方法:全面调查”, ACM Comput。
Surv.,第 50 卷,第 1 期,第 1-37 页,2017 年 3 月。

[8] C. Rathgeb、P. Drozdowski 和 C. Busch,“化妆演示攻击:审查和检测性能基准”, IEEE Access,第 8 卷,第 224958-224973 页,2020 年。

[9] C. Szegedy、W. Zaremba、I. Sutskever、J. Bruna、D. Erhan、I. Goodfellow 和 R. Fergus,《神经网络的迷人特性》,2013 年, arXiv:1312.6199。 [在线]。可访问网址:http://arxiv.org/abs/1312.6199 [10] M. Ferrara、A. Franco 和 D. Maltoni,《神奇护照》, Proc.
IEEE 国际联合会议生物识别技术, 2014 年 9 月,第 1-7 页。

[11] X. Yuan, P. He, Q. Zhu 和 X. Li,《对抗性示例:深度学习的攻击与防御》, IEEE 神经网络学习系统汇刊,第 30 卷,第 9 期,第 2805-2824 页,2019 年 9 月。

[12] H. Xu、Y. Ma、H.-C. Liu、D. Deb、H. Liu、J.-L. Tang 和 AK Jain,《图像、图表和文本中的对抗性攻击与防御:回顾》,《 Int. J. Autom. Comput.》,第 17 卷,第 2 期,第 151-178 页,2020 年 4 月。

[13] A. Agarwal、R. Singh、M. Vatsa 和 N. Ratha,“用于人脸识别的图像不可知通用对抗性扰动是否难以检测?”,载于IEEE 第 9 届国际会议生物识别理论与应用系统 (BTAS) 会议论文集, 2018 年 10 月,第 1-7 页。

[14] G. Goswami、A. Agarwal、N. Ratha、R. Singh 和 M. Vatsa,《检测和减轻对抗性扰动以实现稳健的人脸识别》,《 Int.
J. Comput. Vis.,第 127 卷,第 6-7 期,第 719-742 页,2019 年 6 月。

[15] G. Tao、S. Ma、Y. Liu 和 X. Zhang,“攻击与可解释性:对抗性样本的属性引导检测”,载于Proc. Adv. Neural Inf. Process. Syst., 2018 年,第 7717-7728 页。

[16] L. Guarnera、O. Giudice 和 S. Battiato,《通过在图像上暴露卷积痕迹来对抗深度伪造》, 《IEEE Access》,第 8 卷,第 165085-165098 页,2020 年。

[17] Y. Zhou、X. Hu、L. Wang、S. Duan 和 Y. Chen,《基于马尔可夫链的计算机视觉对抗示例有效防御》,《 IEEE Access》,第 7 卷,第 5695-5706 页,2019 年。

[18] Y. Bakhti、SA Fezza、W. Hamidouche 和 O. Deforges,《DDSA:使用深度去噪稀疏自动编码器防御对抗攻击》,《 IEEE Access》,第 7 卷,第 160397-160407 页,2019 年。

[19] X. Liu、L. Xie、Y. Wang、J. Zou、J. Xiong、Z. Ying 和 AV Vasilakos,《深度学习中的隐私和安全问题:一项调查》,《 IEEE Access》,第 9 卷,第 4566-4593 页,2021 年。

[20] A. Makrushin、T. Neubert 和 J. Dittmann,《自动生成和检测视觉上完美的面部形态》,第 12 届国际会议论文集。
联合会议计算机视觉、图像计算机图形理论应用, 2017 年,第 39-50 页。

[21] DJ Robertson、RSS Kramer 和 AM Burton,《利用脸部变形进行欺诈性身份识别:对人类和自动识别的实验》, PLoS ONE,第 12 卷,第 3 期,2017 年 3 月,文章编号 e0173319。

[22] R. Raghavendra、KB Raja、S. Venkatesh 和 C. Busch,《用于检测数字和印刷扫描变形人脸图像的可迁移深度 CNN 特征》,载于IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW), 2017 年 7 月,第 1822-1830 页。

[23] C. Seibold、W. Samek、A. Hilsmann 和 P. Eisert,《通过深度学习检测人脸变形攻击》,载于Proc. Int. Workshop Digit. Water-marking, 2017 年,第 107-120 页。

[24] L. Debiasi、U. Scherhag、C. Rathgeb、A. Uhl 和 C. Busch,《基于 PRNU 的变形人脸图像检测》,载于国际生物识别取证研讨会 (IWBF) 文集, 2018 年 6 月,第 1-7 页。

[25] L.-B. Zhang、F. Peng 和 M. Long,《利用传感器模式噪声的傅里叶谱进行人脸变形检测》, IEEE 国际会议论文集。
多媒体博览会(ICME), 2018 年 7 月,第 1-6 页。

[26] M. Ferrara、A. Franco 和 D. Maltoni,《人脸变形》, IEEE Trans.
Inf. Forensics Security,第 13 卷,第 4 期,第 1008-1017 页,2018 年 4 月。

[27] F. Peng、L.-B. Zhang 和 M. Long,《FD-GAN:用于恢复同谋面部图像的面部去变形生成对抗网络》,《 IEEE Access》,第 7 卷,第 75122-75131 页,2019 年。

[28] D. Ortega-Delcampo、C. Conde、D. Palacios-Alonso 和 E. Cabello,《使用卷积神经网络去变形方法检测边境控制变形攻击》,《 IEEE Access》,第 8 卷,第 92301-92313 页,2020 年。

[29] N. Carlini。(2019 年)。所有 (arXiv) 对抗性示例论文的完整列表。 [在线]。可用:https://nicholas.carlini.com/writing/2019/all-adversarial-example-papers.html

[30] M. Wang 和 W. Deng,“深度人脸识别:一项调查”,2018 年, arXiv:1804.06655。 [在线]。可访问:http://arxiv.org/abs/1804.06655 [31] W. Deng、J. Hu 和 J. Guo,“扩展的 SRC:通过类内变体词典进行欠采样人脸识别”,IEEE模式分析与机器翻译汇刊。
智力,卷34,没有。 9,第 1864-1870 页,2012 年 9 月。

[32] X. He、S. Yan、Y. Hu、P. Niyogi 和 H.-J. Zhang,《使用拉普拉斯人脸进行人脸识别》, IEEE 模式分析机器智能汇刊,第 27 卷,第 3 期,第 328-340 页,2005 年 3 月。

[33] B. Moghaddam、W. Wahid 和 A. Pentland,《超越特征脸:人脸识别的概率匹配》,载于第三届 IEEE 国际会议论文集。
自动面部手势识别, 1998 年 4 月,第 30-35 页。

[34] L. Zhang, M. Yang 和 X. Feng,“稀疏表示还是协作表示:哪个有助于人脸识别?”,载于Proc. Int. Conf. Comput. Vis., 2011 年 11 月,第 471-478 页。

[35] C. Liu 和 H. Wechsler,《基于 Gabor 特征的分类使用增强型 Fisher 线性判别模型进行人脸识别》, IEEE 图像处理汇刊,第 11 卷,第 4 期,第 467-476 页,2002 年 4 月。

[36] T. Ahonen、A. Hadid 和 M. Pietikainen,《利用局部二元模式进行人脸描述:应用于人脸识别》, IEEE 模式分析机器智能汇刊,第 28 卷,第 12 期,第 2037-2041 页,2006 年 12 月。

[37] Z. Cao、Q. Yin、X. Tang 和 J. Sun,《基于学习的描述符的人脸识别》,载于Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., 2010 年 6 月,第 2707-2714 页。

[38] T.-H. Chan、K. Jia、S. Gao、J. Lu、Z. Zeng 和 Y. Ma,《PCANet:用于图像分类的简单深度学习基线?》IEEE 图像处理汇刊,第 24 卷,第 12 期,第 5017-5032 页,2015 年 12 月。

[39] Z. Lei、M. Pietikäinen 和 SZ Li,《学习判别人脸描述符》, IEEE 模式分析机器智能汇刊,第 36 卷,第 2 期,第 289-302 页,2014 年 2 月。

[40] GB Huang, M. Mattar, T. Berg 和 E. Learned-Miller,《自然界中的标记人脸:用于研究无约束环境中人脸识别的数据库》,载于《 Faces 真实生活图像、检测、对齐和识别研讨会论文集》, 2008 年,第 1-15 页。

[41] F. Schroff、D. Kalenichenko 和 J. Philbin,《FaceNet:用于人脸识别和聚类的统一嵌入》,载于Proc. IEEE Conf. Comput.
Vis. Pattern Recognit. (CVPR), 2015 年 6 月,第 815-823 页。

[42] C. Szegedy、W. Liu、Y. Jia、P. Sermanet、S. Reed、D. Anguelov、D. Erhan、V. Vanhoucke 和 A. Rabinovich,《深入研究卷积》,载于IEEE 计算机视觉模式识别会议论文集 (CVPR), 2015 年 6 月,第 1-9 页。

[43] K. Simonyan 和 A. Zisserman,《用于大规模图像识别的超深卷积网络》,2014 年, arXiv:1409.1556。 [在线]。
来源:http://arxiv.org/abs/1409.1556

[44] W. Liu、Y. Wen、Z. Yu、M. Li、B. Raj 和 L. Song,《SphereFace:用于人脸识别的深度超球面嵌入》, IEEE Conf. 会议论文集。
计算机视觉模式识别(CVPR), 2017 年 7 月,第 212-220 页。

[45] K. He、X. Zhang、S. Ren 和 J. Sun,《深度残差学习用于图像识别》,载于IEEE Conf. Comput. Vis. Pattern Recognit.
(CVPR), 2016 年 6 月,第 770-778 页。

[46] H. Wang, Y. Wang, Z. Zhou, X. Ji, D. Gong, J. Zhou, Z. Li 和 W. Liu,《CosFace:深度人脸识别的大边距余弦损失》,载于Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.,2018 年 6 月,第 5265-5274 页。

[47] J. Deng、J. Guo、N. Xue 和 S. Zafeiriou,《ArcFace:深度人脸识别的附加角度边缘损失》,载于Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2019 年 6 月,第 4690-4699 页。

[48] X. Wu、R. He、Z. Sun 和 T. Tan,《一种用于带噪声标签的深度人脸表示的轻量级 CNN》,IEEE 信息取证安全汇刊,第 13 卷,第 11 期,第 2884-2896 页,2018 年 11 月。

[49] Y. Sun, Y. Chen, X. Wang 和 X. Tang,《通过联合识别-验证进行深度学习人脸表示》,《Proc. Adv. Neural Inf. Process. Syst.》, 2014 年,第 1988-1996 页。

[50] D. Yi、Z. Lei、S. Liao 和 SZ Li,《从头学习人脸表征》,2014 年, arXiv:1411.7923。 [在线]。可访问网址:http://arxiv.org/abs/1411.7923 [51] Y. Guo、L. Zhang、Y. Hu、X. He 和 J. Gao,《MS-celeb-1M:大规模人脸识别的数据集和基准》,载于Proc. Eur. Conf. Comput. Vis., 2016 年,第 87-102 页。

[52] Q. Cao、L. Shen、W. Xie、OM Parkhi 和 A. Zisserman,《VGGFace2:用于识别跨姿势和年龄人脸的数据集》,载于第 13 届 IEEE 国际会议自动人脸手势识别 (FG) 会议论文集, 2018 年 5 月,第 67-74 页。

[53] I. Kemelmacher-Shlizerman、SMSeitz、D. Miller 和 E. Brossard,《MegaFace 基准:用于大规模识别的 100 万张人脸》,载于IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2016 年 6 月,第 4873-4882 页。

[54] A. Kurakin、I. Goodfellow 和 S. Bengio,《物理世界中的对抗性示例》,2016 年, arXiv:1607.02533。 [在线]。网址:http://arxiv.org/abs/1607.02533

[55] D.-L. Nguyen、SS Arora、Y. Wu 和 H. Yang,“对抗性光投射攻击人脸识别系统:可行性研究”, Proc. IEEE/CVF Conf. 计算机视觉模式识别研讨会 (CVPRW), 2020 年 6 月,第 814-815 页。

[56] D. Deb、J. Zhang 和 AK Jain,《AdvFaces:对抗性人脸合成》,2019 年, arXiv:1908.05008。[在线]。可访问网址:http://arxiv.org/abs/1908.05008

[57] Z. Liu, P. Luo, X. Wang 和 X. Tang,《深度学习自然人脸属性》, IEEE 国际计算机视觉大会 (ICCV) 论文集, 2015 年 12 月,第 3730-3738 页。

[58] B. Amos、B. Ludwiczuk 和 M. Satyanarayanan,《Openface:具有移动应用程序的通用人脸识别库》,卡内基梅隆大学计算机科学学院,第 6 卷,第 2 页,2016 年 6 月。

[59] IJ Goodfellow、J. Shlens 和 C. Szegedy,《解释和利用对抗性样本》,2014 年, arXiv:1412.6572。 [在线]。可访问网址:http://arxiv.org/abs/1412.6572 [60] A. Rozsa、EM Rudd 和 TE Boult,《对抗性多样性和硬正例生成》, IEEE Conf. Comput. Vis. Pattern Recognit. 会议记录。研讨会(CVPRW), 2016 年 6 月,第 25-32 页。

[61] N. Papernot、P. McDaniel、S. Jha、M. Fredrikson、ZB Celik 和 A. Swami,《对抗环境中深度学习的局限性》,载于《IEEE 欧洲安全隐私研讨会文集》(EuroS&P), 2016 年 3 月,第 372-387 页。

[62] J. Su、DV Vargas 和 K. Sakurai,《欺骗深度神经网络的单像素攻击》, IEEE 计算演化学报,第 23 卷,第 5 期,第 828-841 页,2019 年 10 月。

[63] S.-M. Moosavi-Dezfooli、A. Fawzi 和 P. Frossard,《DeepFool:一种简单而准确的欺骗深度神经网络的方法》,载于IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2016 年 6 月,第 2574-2582 页。

[64] S.-M. Moosavi-Dezfooli、A. Fawzi、O. Fawzi 和 P. Frossard,《通用对抗性扰动》,载于IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2017 年 7 月,第 1765-1773 页。

[65] N. Carlini 和 D. Wagner,《评估神经网络的稳健性》,载于《IEEE 安全隐私 (SP) 研讨会文集》, 2017 年 5 月,第 39-57 页。

[66] T. Miyato、S.-I. Maeda、M. Koyama 和 S. Ishii,《虚拟对抗训练:监督和半监督学习的正则化方法》,《IEEE 模式分析机器智能汇刊》,第 41 卷,第 8 期,第 1979-1993 页,2019 年 8 月。

[67] A. Kurakin、I. Goodfellow 和 S. Bengio,《大规模对抗机器学习》,2016 年, arXiv:1611.01236。 [在线]。可访问:http://arxiv.org/abs/1611.01236 [68] S. Das 和 P. Nagaratnam Suganthan,《差异进化:最新进展调查》, IEEE 进化计算学报,第 15 卷,第 1 期,第 4-31 页,2011 年 2 月。

[69] G. Hinton、O. Vinyals 和 J. Dean,“提炼神经网络中的知识”,2015 年, arXiv:1503.02531。[在线]。可访问网址:http://arxiv.org/abs/1503.02531 [70] N. Papernot、P. McDaniel、X. Wu、S. Jha 和 A. Swami,“提炼作为对深度神经网络对抗扰动的防御”, IEEE 安全隐私研讨会论文集 (SP), 2016 年 5 月,第 582-597 页。

[71] G. Goswami、N. Ratha、A. Agarwal、R. Singh 和 M. Vatsa,“揭示基于深度学习的人脸识别对抗攻击的稳健性”,载于Proc. AAAI Conf. Artif. Intell., 2018 年,第 1-8 页。

[72] Y. Dong、H. Su、B. Wu、Z. Li、W. Liu、T. Zhang 和 J. Zhu,“基于决策的有效人脸识别黑盒对抗攻击”,载于IEEE Conf. Comput. Vis. Pattern Recognit. 会议论文集, 2019 年 6 月,第 7714-7722 页。

[73] C. Igel、T. Suttorp 和 N. Hansen,《用于进化策略的计算有效协方差矩阵更新和 (1+1)-CMA》,载于《第 8 届遗传进化计算会议论文集》, 2006 年,第 453-460 页。

[74] N. Hansen 和 A. Ostermeier,《进化策略中的完全去随机自适应》,《Evol. Comput.》,第 9 卷,第 2 期,第 159-195 页,2001 年 6 月。

[75] Y. Zhong 和 W. Deng,《基于边缘的三重嵌入正则化的对抗性学习》, IEEE 国际计算机视觉会议论文集, 2019 年 10 月,第 6549-6558 页。

[76] M. Sharif、S. Bhagavatula、L. Bauer 和 MK Reiter,《犯罪附属品:对最先进人脸识别技术的真实而隐秘的攻击》,载于Proc. ACM SIGSAC Conf. Comput. Commun. Secur., 2016 年,第 1528-1540 页。

[77] A. Mahendran 和 A. Vedaldi,《通过反转理解深度图像表示》,载于IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2015 年 6 月,第 5188-5196 页。

[78] M. Shen、Z. Liao、L. Zhu、K. Xu 和 X. Du,“VLA:一种针对物理世界中人脸识别系统的实用可见光攻击”, Proc. ACM Interact.,移动、可穿戴无处不在的技术,第 3 卷,第 3 期,第 1-19 页,2019 年。

[79] L. Zhang, C. Bo, J. Hou, X.-Y. Li, Y. Wang, K. Liu 和 Y. Liu,《Kaleido:你可以观看但不能录制》,载于《第 21 届国际会议论文集》。移动计算网络, 2015 年,第 372-385 页。

[80] S. Komkov 和 A. Petiushko,《AdvHat:对 ArcFace 人脸识别系统的真实对抗性攻击》,2019 年, arXiv:1908.08705。 [在线]。可用:http://arxiv.org/abs/1908.08705

[81] E. Chatzikyriakidis、C. Papaioannidis 和 I. Pitas,《对抗性面部去识别》,载于IEEE 国际图像处理会议论文集 (ICIP), 2019 年 9 月,第 684-688 页。

[82] H. Kwon, O. Kwon, H. Yoon 和 K.-W. Park,《人脸识别系统中的好友安全对抗示例》,第 11 届国际会议论文集。无处不在的未来网络 (ICUFN), 2019 年 7 月,第 547-551 页。

[83] A. Dabouei、S. Soleymani、J. Dawson 和 N. Nasrabadi,《快速几何扰动对抗面孔》, IEEE Winter Conf. 会议纪要。Appl. Comput. Vis. (WACV), 2019 年 1 月,第 1979-1988 页。

[84] C. Xiao、J.-Y. Zhu、B. Li、W. He、M. Liu 和 D. Song,《空间转换的对抗性示例》,2018 年, arXiv:1801.02612。 [在线]。网址:http://arxiv.org/abs/1801.02612 [85] Y. Zhong 和 W. Deng,“面向可转移对抗性深度人脸识别攻击”,2020 年, arXiv:2004.05790。 [在线]。网址:http://arxiv.org/abs/2004.05790 [86] Y. Dong、F. Liao、T. Pang、H. Su、J. Zhu、X. Hu 和 J. Li,“通过动量推动对抗性攻击”,载于Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2018 年 6 月,第 9185-9193 页。

[87] Y. Liu, X. Chen, C. Liu 和 D. Song,《深入研究可转移对抗样本和黑盒攻击》,2016 年, arXiv:1611.02770。 [在线]。可访问网址:http://arxiv.org/abs/1611.02770 [88] C. Xie、Z. Zhang、Y. Zhou、S. Bai、J. Wang、Z. Ren 和 AL Yuille,“通过输入多样性提高对抗性示例的可转移性”,载于Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2019 年 6 月,第 2730-2739 页。

[89] G. Garofalo、V. Rimmer、D. Preuveneers 和 W. Joosen,《鱼尾纹:制作对抗性图像以毒害人脸身份验证》,载于第 12 届 USENIX 进攻性技术研讨会文集, 2018 年,第 1-12 页。

[90] D. Goodman、H. Xin、W. Yang、W. Yuesheng、X. Junfeng 和 Z. Huan,“Advbox:一种用于生成欺骗神经网络的对抗性示例的工具箱”,2020 年, arXiv:2001.05574。 [在线]。可访问网址:http://arxiv.org/abs/2001.05574 [91] M. Sharif、S. Bhagavatula、L. Bauer 和 MK Reiter,“具有目标的对抗性示例的通用框架”, ACM 隐私安全汇刊,第 22 卷,第 3 期,第 1-30 页,2019 年 7 月。

[92] Z. Zhou、D. Tang、X. Wang、W. Han、X. Liu 和 K. Zhang，“隐形面具:对红外人脸识别的实际攻击”，2018 年， arXiv:1803.04683。 [在线]。可访问网址:http://arxiv.org/abs/1803.04683 [93] M. Pautov、G. Melnikov、E. Kaziakhmedov、K. Kireev 和 A. Petiushko，“关于对抗性补丁:对 ArcFace-100 人脸识别系统的真实攻击”，载于Proc. Int. Multi-Conf. Eng., Comput. Inf. Sci.

（SIBIRCON）， 2019 年 10 月,第 391-396 页。

[94] Q. Song、Y. Wu 和 L. Yang，“使用注意力对抗攻击生成网络对最先进人脸识别的攻击”，2018 年， arXiv:1811.12026。 [在线]。可访问网址:http://arxiv.org/abs/1811.12026 [95] A. Majumdar、R. Singh 和 M. Vatsa，“通过基于类稀疏性的监督编码进行人脸验证”， IEEE 模式分析与机器翻译汇刊。

智力,卷39,没有。 6,第 1273–1280 页,2017 年 6 月。

[96] JR Beveridge、KW Bowyer、PJ Flynn、S. Cheng、PJ Phillips、DS Bolme、BA Draper、GH Givens、YM Lui、MN Teli、H. Zhang 和 WT Scruggs，《数码傻瓜相机带来的脸部识别挑战》，载于《 IEEE 第 6 届生物识别国际会议:理论与应用系统 (BTAS)》会议论文集， 2013 年 9 月，第 1-8 页。

[97] AP Founds、N. Orlans、W. Genevieve 和 CI Watson，《NIST 特殊数据库 32-多次遭遇数据集 II (MEDS-II)》，美国马里兰州盖瑟斯堡国家标准技术研究所,技术报告,2011 年。

[98] W. Brendel、J. Rauber 和 M. Bethge，《基于决策的对抗攻击:针对黑盒机器学习模型的可靠攻击》，2017 年，arXiv:1712.04248。 [在线]。可访问网址:http://arxiv.org/abs/1712.04248

[99] M. Cheng、T. Le、P.-Y. Chen、J. Yi、H. Zhang 和 C.-J. Hsieh，“查询高效硬标签黑盒攻击:一种基于优化的方法”，2018 年， arXiv:1807.04457。 [在线]。可访问网址:http://arxiv.org/abs/1807.04457 [100] A. Ilyas、L. Engstrom、A. Athalye 和 J. Lin，“使用有限查询和信息的黑盒对抗性攻击”，2018 年，
arXiv:1804.08598。

[在线]。可访问网址:http://arxiv.org/abs/1804.08598 [101] F. Wang, L. Chen, C. Li, S. Huang, Y. Chen, C. Qian 和 CC Loy，“人脸识别的魔鬼就在噪音中”，载于 Proc. Eur. Conf.
计算。可见。 （ECCV）， 2018,第 765-780 页。

[102] H.-W. Ng 和 S. Winkler，《一种数据驱动的清理大型人脸数据集的方法》， IEEE 国际图像处理会议论文集（ICIP）， 2014 年 10 月,第 343-347 页。

[103] KR Mopuri、U. Garg 和 RV Babu，《快速特征欺骗:一种独立于数据的通用对抗性扰动方法》，2017 年， arXiv:1707.05572。 [在线]。可访问网址:http://arxiv.org/abs/1707.05572 [104] R. Gross、I. Matthews、J. Cohn、T. Kanade 和 S. Baker，《Multi-PIE》，《图像可视化计算》,第 28 卷,第 5 期,第 807-813 页,2010 年。

[105] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama 和 T. Darrell,《Caffe:用于快速特征嵌入的卷积架构》,第 22 届 ACM 国际多媒体会议论文集, 2014 年 11 月, 第 675-678 页。

[106] L. Kurnianggoro 和 K.-H. Jo，“增强输入的预测集合作为人脸验证系统的对抗性防御”，载于《亚洲会议情报数据库系统文集》， 2019 年,第 658-669 页。

[107] PJ Phillips、PJ Flynn、JR Beveridge、WT Scruggs、AJ O toole、D. Bolme、KW Bowyer、BA Draper、GH Givens、YM Lui、H. Sahibzada、JA Scallan III 和 S. Weimer，《多重生物识别技术大挑战概述》，载于Proc. Int. Conf. Biometrics， 2009 年,第 705-714 页。

[108] A. Goel、A. Agarwal、M. Vatsa、R. Singh 和 N. Ratha，《使用区块链保护 CNN 模型和生物识别模板》，载于《IEEE BTAS 论文集》， 2019 年 9 月,第 1-7 页。

[109] Y. Su, G. Sun, W. Fan, X. Lu 和 Z. Liu，“通过残差生成网络清除人脸验证中的对抗性扰动”，

IEEE 国际声学会议,语音信号处理。（ICASSP）， 2019 年 5 月,第 2597-2601 页。

[110] F. Carrara、R. Becarelli、R. Caldelli、F. Falchi 和 G. Amato，《特征距离空间中的对抗性示例检测》，载于Proc. Eur. Conf.
计算。可见。 （ECCV）， 2018,第 1-15 页。

[111] FV Massoli、F. Carrara、G. Amato 和 F. Falchi，《人脸识别对抗攻击检测》，2019 年，arXiv:1912.02918。 [在线]。
可访问网址:http://arxiv.org/abs/1912.02918 [112]
Y. Kim、D. Han、C. Kim 和 H.-J. Yoo，“具有对抗攻击预防功能的 0.22–0.89 mW 低功耗、高安全性始终开启人脸识别处理器”， IEEE 电路系统汇刊 II,实验简报,第 67 卷,第 5 期,第 846–850 页，2020 年 5 月。

[113] W. Xu、D. Evans 和 Y. Qi，《特征压缩:在深度神经网络中检测对抗性示例》，2017 年，arXiv:1704.01155。 [在线]。
可访问网址:http://arxiv.org/abs/1704.01155 [114]
A. Goel、A. Singh、A. Agarwal、M. Vatsa 和 R. Singh，“SmartBox:人脸识别对抗检测和缓解算法的基准测试”，载于IEEE 第 9 届国际会议生物识别理论、应用系统会议论文集。

（BTAS）， 2018 年 10 月,第 1-7 页。

[115] R. Theagarajan 和 B. Bhanu，“保护黑盒面部识别分类器免受对抗性攻击”， IEEE/CVF Conf. Comput. 会议纪要。
Vis. 模式识别研讨会 (CVPRW)， 2020 年 6 月,第 812-813 页。

FATEMEH VAKHSHITEH获得了伊朗德黑兰 Sharif 理工大学 (SUT) 的应用数学学士学位、加拿大安大略省渥太华卡尔顿大学 (Carleton University) 的生物医学工程硕士学位以及德黑兰 Amirkabir 理工大学 (AUT) 的生物医学工程博士学位。她的主要研究兴趣包括模式识别、机器学习、图像处理和深度学习,以及应用人脸识别和对抗性攻击。

AHMAD NICKABADI分别于 2004 年、2006 年和 2011 年获得伊朗德黑兰阿米尔卡比尔理工大学 (AUT) 计算机工程学士学位以及人工智能硕士和博士学位。自 2012 年以来,他一直担任 AUT 计算机工程系的助理教授。他的研究兴趣包括使用深度学习和概率图模型分析图像和视频内容,特别关注关于活动识别、面部识别和面部合成。

RAGHAVENDRA RAMACHANDRA （IEEE 高级会员)获得印度迈索尔大学（UOM)学士学位、印度贝尔高姆 Visvesvaraya 科技大学电子与通信专业硕士学位以及 UOM、电信学院和法国埃夫里 Telecom Sudparis 计算机科学与技术专业博士学位（合作完成）。他目前被任命为挪威约维克挪威科技大学 (NTNU) 挪威生物特征识别实验室的教授。他曾是意大利热那亚意大利理工学院 (IIT) 的研究员。他撰写过多篇文章并拥有专利。他成功领导了由欧盟、美国 IARPA、印度 MHRD 和挪威挪威研究委员会资助的多个项目。他的主要研究兴趣包括统计模式识别、应用机器学习、深度学习、数据融合方案和随机优化,应用于生物识别、多模态生物特征融合、人类行为分析和群体行为分析。他是多个国际会议和期刊的审稿人/组织者/编辑委员会。他是 ISO/IEC SC37 的编辑,并为多模态生物特征识别标准 TR 24722 做出了贡献。