实施与反抗“人脸欺诈”——计算机视觉课程报告

张效伟
21821203
DILab
zhmoll@zju.edu.cn

摘要（Abstract）

本综述从计算机视觉领域的人脸识别发展历程出发，提出了人脸识别现存的一个“人脸欺诈”问题。并从CVPR2019 中找到了两篇关于对抗与反对抗人脸识别技术的方法，即实施“人脸欺诈”和反抗“人脸欺诈”问题，就关键方法加以阐述。最后，对现有的“人脸欺诈”问题以及攻击和反抗的解决办法进行评价。

1. 引言

计算机视觉（下简称CV）是一个研究机器“看”世界的科学问题，在当下机器学习大热的情况下，CV焕发新的生机，其不少研究方法都由机器学习赋予了强大的驱动力而迅速前进，也吸引了更多人前往这个领域一探究竟。CVPR（IEEE Conference on Computer Vision and Pattern Recognition的缩写）是计算机视觉和模型识别领域的顶级会议，近年来接收的论文逐年增多，今年接收了 1300 篇论文，令人震撼。

本报告关注今年CVPR的关于人脸识别的研究成果。在人脸识别技术逐渐成熟的今天，对抗与反对抗在人脸识别之中的重要性也愈加凸显。在第二章介绍了与人脸识别与人脸欺诈相关的概念和发展历程，第三章综述人脸识别技术的反欺诈的最新成果，第四章则是评价、总结和展望。

2. 相关概念

2.1. 简介

人脸识别是一种基于人的脸部特征信息进行身份识别的生物识别技术。通常采用摄像机或摄像头采集含有人脸的图像或视频，并自动在图像中检测和跟踪人脸，进而对检测到的人脸进行脸部识别的一系列相关技术。

2.2. 发展历程

人脸识别算法大致经历了三个阶段——特征算法、人工特征+分类器算法和深度学习算法。

2.2.1 特征算法

早期的人脸识别算法主要有基于几何特征的算法、基于模板匹配的算法、子空间算法等多种类型。其代表的算法有PCA（主成分分析，降维后识别人脸）、LDA（线性判别分析，保证同一个人的不同人脸图像在投影后聚集在一起，不同人的人脸在投影后用一个大的间距分开）和HMM（隐马尔科夫模型，能够对光照变化、表情和姿态的变化更加鲁棒）等。

2.2.2 人工特征+分类器

学者们在机器学习理论的发展下相继推出了基于SVM（支持向量机）、boosting、遗传算法等分类器的人脸识别方法。

此时业界基本达成共识：基于人工构造的局部描述子进行特征提取和子空间方法进行特征选择能够取得最好的识别效果。Gabor 及 LBP 特征描述算子是人脸识别领域最为成功的两种人工设计局部描述特征。

2.2.3 深度学习

自深度学习在ILSVRC-2012 大放异彩后，很多研究者都在尝试将其应用在自己的方向，CV领域则是深度学习的最大受益者之一。卷积神经网络对于图像分类有着巨大的功效，而且其对于人脸识别技术的发展亦有决定性的作用。

CVPR2014 上Facebook提出的DeepFace是深度卷积神经网络在人脸识别问题上的奠基之作[1]。2015 年Google提出的FaceNet[2]使用三元组损失函数（Triplet Loss）代替常用的Softmax损失函数，使用了GoogLeNet的Inception模型，减少了模型参数量，提高了精度。

香港中文大学提出的DeepID系列方法[3-5]也有不错的提升效果，随后更多的研究关注如何设计损失函数来提升模型训练的速度和准确性。

在深度学习阶段的前期，学者主要在网络结构和输入数据进行研究；后期，主要的改进集中在损失函数

和训练的方法上，使得卷积网络能够更有效地学习到更显著的特征。

2.3. 人脸欺诈

人脸识别依然存在不少的问题需要解决。一个显著的问题就是“人脸欺诈”。

人脸识别通常有两个子任务:人脸验证和人脸识别。前者区分一对人脸图像是否代表相同的身份，而后者对图像进行分类。最先进的人脸识别模型利用深度神经网络来提取具有最小类内方差和最大类间方差的人脸特征，从而实现这两项任务。由于这些模型的优异性能，人脸识别在金融/支付、公共接入、刑事识别等众多应用中被广泛应用于身份认证。

尽管在各种应用中都取得了巨大的成功，但众所周知，深度神经网络很容易受到反面例子的攻击。对于人类观察者来说，这些恶意生成的对抗性例子通常通过添加小扰动与合法的例子难以区分。但是，他们可以使深层模型产生不正确的预测。基于深度神经网络的人脸识别系统也表现出了对这类对抗实例的脆弱性。例如，可以对眼镜进行对抗性的干扰，当戴上眼镜时，攻击者就可以逃避被认出或冒充另一个人。人脸识别系统在现实应用中的不安全性，尤其是那些具有敏感目的的应用，可能会导致严重的后果和安全问题。

3. 最新成果

3.1. 基于深度树学习的ZSFA[6]

尽管采用了深度学习方法的人脸识别系统识别率极高，但是基于二维图像的系统往往不能分辨真实的人脸和虚假的人脸。攻击者可以利用许多媒介来达成欺骗攻击，比如重放攻击（重放有效的视频）、打印攻击（打印有效的照片）等。目前已有不少研究专注于解决这样的“人脸欺诈”问题，如使用手工构造特征、使用基于CNN的特征，降低了这两种攻击方法的成功率。

最近出现了使用 3D蒙版的攻击手段，而上述的手工构造特征和基于CNN的特征方法对于这种攻击手段的抵抗效果较差，但是可以使用远程光体积扫描术（r-PPG）检测心率脉搏作为欺骗的提示。此外，面部化妆也可能影响识别结果。

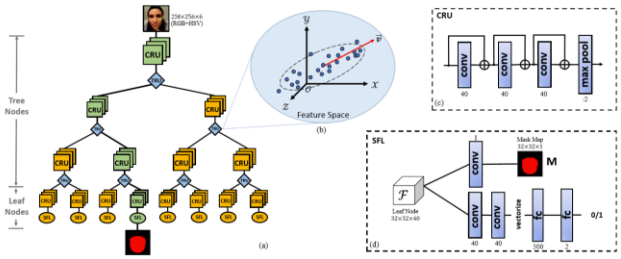
上述所有的解决方案都是针对各自的攻击方案而提出的并测试的，然而在实际引用中，攻击者往往会混合多种攻击手段，甚至采取人脸识别系统设计者所不知道的手法进行欺诈，现有的系统往往束手无策。

反“人脸欺诈”越来越受到研究者的重视，其中就有学者试图将反“人脸欺诈”的模型进行泛化，试图以“零学习”的成本解决这个问题，被称作Zero-Shot Face Anti-spoofing（ZSFA）。

有学者在CVPR2019 提出了一种基于深度树学习的方法，将ZSFA泛化到能抵御更多未知欺骗（从 2 种扩展到 13 种）攻击的程度。论文的主要贡献即将ZSFA泛化到更广泛的程度，提出了深度树网络（DTN）用于分层学习特征和检测未知的欺骗攻击，以及为ZSFA收集了新的数据库。

3.1.1 深度树网络

DTN的主要目的有两个：1、发现已知欺骗的语义子群；2、分层学习特征。



如图所示，每个树节点由卷积残差单元（CRU）和树路由单元（TRU）组成，叶子节点由CRU和监督特征学习模块（FSL）组成。CRU是一个具有卷积层和捷径连接的块。TRU定义了一个节点路由函数，用于将数据样本路由到子节点之一。路由函数沿着数据变化最大的方向对所有访问数据进行分区。SFL模块将分类监控和像素级监控相结合，学习欺骗干扰的特征。

3.1.2 欺骗树

树的目的是从所有已知的欺骗中找到语义子组，称为欺骗树。同样，也可以只训练具有真脸的活树，以及同时具有真数据和欺骗数据的通用数据树。与欺骗树相比，活树和通用数据树存在一些不足。活树并不能传达欺骗的语义，在每个节点学习到的属性不能帮助路由和更好的检测欺骗；通用数据树可能导致不平衡的子组，其中一个类的样本数量超过另一个类。这种不平衡会给下一步处理带来不少偏差。

为了使每个叶子都有一个平衡的子组，将真脸数据的响应抑制为零，以便将所有真脸数据均匀地划分到子节点。同时还抑制了不访问该节点的欺骗数据的响应，使每个节点都能模拟一个唯一的欺骗子集的分布。

3.1.3 训练方法

深度树网络(DTN)是该模型的主要框架。输入$I \in R^{256\times256\times6}$，其中 6 个通道为RGB+HSV颜色空间。将 3 个 3×3 的卷积层与 40 个信道和 1 个最大池化层连接起来，并将它们分组为一个CRU。由于网络中批量大小的动态性，每个卷积层都配备了ReLU和分组归一化层。还为每个卷积层应用了一个快捷连接。对于每个树节点，在TRU之前部署一个CRU。叶子节点，DTN产生的特征表示输入I，$F(I|\theta) \in R^{32\times32\times40}$，然后使用一层 1×1 的卷积层生成二值蒙版映射M。

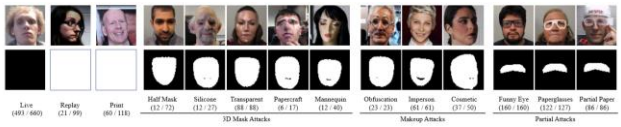
TRU模块将数据样本路由到其子CRU。首先利用1×1 卷积层对特征进行压缩，并在空间上调整响应大小。对于根节点，将CRU特征压缩到$x \in R^{32\times32\times10}$，对于以后的树节点，将CRU特征压缩到$x \in R^{16\times16\times20}$。将输入特征压缩到更小的尺寸有助于减少计算的负担，并在公式中保存协方差矩阵。之后，将输出向量化，应用路由函数。最后，将压缩后的CRU响应投影到最大的基v上，获得投影系数。然后将系数为负的样本赋值给左侧子CRU，系数为正的样本赋值给右侧子CRU。

论文提出的网络是以端到端的方式进行训练的。所有损失都是根据每个小批计算的。DTN模块与TRU模块交替优化。优化DTN时保持参数不变，反之亦然。

3.1.4 多种欺骗攻击的数据库

论文收集了具有多种欺骗攻击的野生数据库（下称SiW-M）。与已有的数据库相比，SiW-M在欺骗攻击、主体身份、环境等方面表现出很大的多样性。

对于欺骗数据的收集，考虑两种欺骗场景：模拟（被识别为其他人）和混淆（隐藏攻击者的身份）。总共收集了 968 个视频，共 13 种类型的欺骗攻击，记录了 1080P、720P的视频，按层次排列如图所示。



对于打印和重播攻击，打算从现有系统识别失败的情况中收集视频。因此，部署了一种现成的人脸反欺骗算法，并在该算法预测实时录制的恶意视频。

现场视频采集分为 3 个场景：1、房间环境，拍摄对象的姿势、灯光、表情等变化不大；2、一个不同的、大得多的房间，其他条件与 1 相同；3、手机模式，拍摄对象在移动，引入了极端位姿角和光照条件。

3.1.5 实验结果

在已有的数据集和论文所提供的数据集上都具有很好的效果。

3.1.6 结论与评价

本文研究了 13 种类型的欺骗攻击中的ZSFA。该方法利用深度树网络将未知攻击路由到最合适的叶节点进行欺骗检测。树形图以一种无监督的方式进行训练，以找到变异最大的特征基来分割假数据。论文收集的SiW-M包含的主题和欺骗类型比任何以前的数据库都多。最后，通过实验验证了该方法的优越性。

3.2. 基于决策的人脸识别黑箱对抗攻击[7]

对抗性攻击被广泛研究，因为它们可以在模型部署之前识别出模型的脆弱性。在基于决策的黑盒攻击场景中，攻击者无法访问模型参数和梯度，只能通过向目标模型发送查询来获取硬标签预测。本文对目前最先进的人脸识别模型的鲁棒性进行了评估。这种攻击设置在真实的人脸识别系统中更实用。为了提高已有方法的效率，提出了一种进化攻击算法，该算法可以对搜索方向的局部几何进行建模，减小搜索空间的维数。大量的实验证明了该方法的有效性，该方法对输入查询较少的人脸图像的扰动最小。并将该方法成功地应用于实际的人脸识别系统中。

论文的主要贡献在于：

1. 提出了一种基于决策的黑盒场景下的进化攻击方法，该方法可以对搜索方向的局部几何进行建模，同时降低了搜索空间的维数。进化攻击方法一般适用于任何图像识别任务，与现有方法相比，其效率有了显著提高。
2. 深入评估了几种最先进的人脸识别模型的鲁棒性，基于决策的黑盒攻击在不同的设置。在此背景下，展示了这些人脸模型的脆弱性。
3. 通过对实际人脸识别系统的成功攻击，验证了该方法的实用性。

3.2.1 进化攻击算法

攻击的工作原理是寻找一个识别为特定身份的对抗性图像，它可以用来逃避人脸认证系统。对于人脸验证，攻击者试图找到一个被识别为另一幅图像的相同身份的对抗性图像，而原始图像并非来自相同的身份。对于人脸识别，生成的敌对图像需要被分类为一个特定的身份。

由于无法访问模型的配置和参数，只能发送查询来探测模型，因此采用黑盒优化技术来最小化目标函数。梯度估计方法在模型给出预测概率时，利用有限差分近似目标函数的梯度，并通过梯度下降更新解，这是基于分数的黑盒攻击中常用的方法。然而，在硬标签输出的情况下，攻击目标函数是不连续的，并且输出对小的输入扰动不敏感。因此，梯度估计方法不能直接应用。一些方法成功地将不连续优化问题重新表述为一些连续优化问题，并使用梯度估计方法进行优化。但是他们需要计算一个点到决策边界的距离，或者通过硬标签输出来估计预测概率，实验证明这种方法效率较低。因此，考虑如何直接有效地优化。

本文提出了一种新的进化攻击方法来解决黑盒优化问题。该方法基于一种简单有效的协方差矩阵自适应进化策略(CMA-ES)的变体，即(1+1)-CMA-ES。在(1+1)-CMA-ES的每次更新迭代中，通过添加一个随机噪声，从其父（当前的解决方案）生成一个新的子代（候选解决方案），评估这两个解决方案的目标，并为下一个迭代选择更好的一个。该方法能够较好地解决黑盒优化问题。考虑到人脸图像基于决策的黑盒攻击的查询限制，原算法(1+1)-CMA-ES可能不可行。为了加快算法的速度，设计了一个合适的分布来采样每个迭代中的随机噪声，它可以对搜索方向的局部几何形状建模。针对该问

题的特点，提出了几种降低搜索空间维数的方法。

整个进化攻击算法如下所示。

Algorithm 1 The evolutionary attack algorithm

Input: The attack objective function $\mathcal{L}(\boldsymbol{x}^*)$; the original face image $\boldsymbol{x}$; the dimension $n \in \mathbb{N}_+$ of the input space ($\boldsymbol{x}^* \in \mathbb{R}^n$); the dimension $m \in \mathbb{N}_+$ of the search space; the number of coordinates $k \in \mathbb{N}_+$ for stochastic coordinate selection.

Input: The total number of queries T.

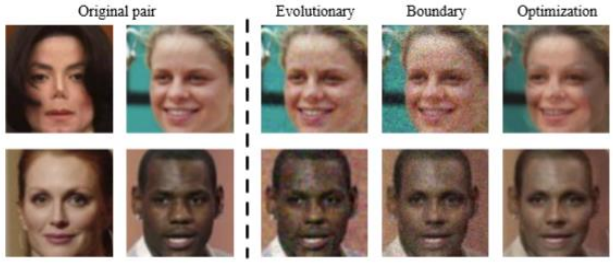
Initialize $\mathbf{C} = \mathbf{I}_m$, $\boldsymbol{p}_c = \mathbf{0}$, $\sigma, \mu, c_c, c_{cov} \in \mathbb{R}_+$, $\tilde{\boldsymbol{x}}^* \in \mathbb{R}^n$;
for $t = 1$ to T **do**
  Sample $\boldsymbol{z} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{C})$;
  Select k coordinates from m with probability proportional to each diagonal element in $\mathbf{C}$;
  Set the non-selected coordinates of $\boldsymbol{z}$ to 0;
  Upscale $\boldsymbol{z}$ to $\mathbb{R}^n$ by bilinear interpolation and obtain $\tilde{\boldsymbol{z}}$;
  $\tilde{\boldsymbol{z}} \leftarrow \tilde{\boldsymbol{z}} + \mu(\boldsymbol{x} - \tilde{\boldsymbol{x}}^*)$;
  if $\mathcal{L}(\tilde{\boldsymbol{x}}^* + \tilde{\boldsymbol{z}}) < \mathcal{L}(\tilde{\boldsymbol{x}}^*)$ **then**
    $\tilde{\boldsymbol{x}}^* \leftarrow \tilde{\boldsymbol{x}}^* + \tilde{\boldsymbol{z}}$;
    Update $\boldsymbol{p}_c$ and $\mathbf{C}$ by $\boldsymbol{z}$ according to Eq. (3) and Eq. (4);
  end if
end for
return $\tilde{\boldsymbol{x}}^*$.

总的来说，该算法由六步构成：

1. 初始化
2. 平均高斯分布
3. 协方差矩阵适应
4. 随机坐标选择
5. 降维
6. 超参数调整

3.2.2 *实验结果*

将进化攻击方法应用于腾讯AI开放平中的人脸验证API。该人脸验证API允许用户上传两张人脸图像，并输出它们的相似度评分。将阈值设置为 90，即，若相似度得分大于 90，则预测两幅图像具有相同的同一性；如果不是，他们被预测为不同的身份。从LFW数据集中选择 10 对图像来执行模拟攻击。每对人脸的原始图像来自不同的身份。为其中一个图像生成光栅，并使API识别出与另一个图像相同的对抗性图像。将查询的最大数量设置为 10,000。使用所提出的进化方法攻击人脸验证API，并将结果与边界和优化进行比较。



上图中还显示了两个例子。可以看出，对抗性图像在视觉上更接近于原始图像，而其他方法生成的对抗性图像失真较大，与原始图像有较大的区别。

3.2.3 *结论*

在基于决策的黑盒设置中，提出了一种进化攻击方法来生成反例。该方法通过对搜索方向进行局部几何建模，同时减小了搜索空间的维数，提高了搜索效率。将所提出的方法应用于几种最先进的人脸识别模型的鲁棒性综合研究，并与其他方法进行了比较。大量的实验证明了该方法的有效性。我们发现现有的人脸识别模型极易受到黑盒方式的攻击，这为开发更健壮的人脸识别模型提出了安全问题。最后利用该方法对一个实际的人脸识别系统进行了攻击，证明了该方法的实用性。

4. 总结与评价

首先，就阅读论文的过程中而言，论文均提出了现有数据集无法满足实验要求的问题。联想到最近微软删除世界上最大的人脸识别公开数据库，人脸识别的安全性和伦理问题亟待妥善解决，一个合理的方法是限制人脸识别技术的自由发展。但是，学术界对人脸识别技术依然有着更高的追求。为了解决数据集不适用问题，最近产生了不少的新数据集以供研究，其中[8]就专门提到了这个问题，并贡献了一个非常不错的数据库以继续研究人脸识别技术的相关方法，并且，论文还给出了一种基准方法作为参照。

其次，对抗与反对抗永远是一对矛盾的话题，此消而彼长。值得欣喜的是，不少“人脸欺诈”方法需要做的准备具有一定的针对性，在现实世界中无法适应复杂的生产环境，因此欺骗技术并不太容易成功。这也是目前手机、平板等用户终端设备仍然提供人脸识别作为生物身份识别功能的补充方法。而在大型场合的生物身份识别，往往有多种技术相辅相成，如虹膜识别、静脉识别、面部特征识别以及体型识别等，也同样会有人工鉴别的参与，因此实验室环境中的先进攻击方法往往难以在这种生产环境下生效。因此，我认为，下一步对抗人脸识别的技术进行“人脸欺诈”的行为，可能要着眼于能够轻易突破真实环境的人脸识别系统，才能让欺诈技术有更长足的进步。

最后，尽管人脸识别的实验室准确率已经让人叹为观止，然而生产环境的准确率却没有实验室的准确率那么乐观。可以做的是，下一步的准确性研究应该着眼更加极端光照、极端角度等的方向，甚至结合非图像感知器件的硬件策略，毕竟，不是每个人都喜欢有摄像头照在脸上。无感知的人脸识别技术，可能是下一个研究的新热点。

引用

[1] Taigman Y, Yang M, Ranzato M A, et al. Deepface: Closing the gap to human-level performance in face verification[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2014: 1701-1708.

[2] Schroff F, Kalenichenko D, Philbin J. Facenet: A unified embedding for face recognition and clustering[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2015: 815-823.
[3] Yi Sun, Xiaogang Wang, Xiaoou Tang. DeepID: Deep Learning for Face Recognition. 2014, computer vision and pattern recognition.
[4] Yi Sun, Yuheng Chen, Xiaogang Wang, Xiaoou Tang. Deep Learning Face Representation by Joint Identification-Verification. 2014, neural information processing systems.
[5] Yi Sun, Ding Liang, Xiaogang Wang, Xiaoou Tang. DeepID3: Face Recognition with Very Deep Neural Networks. 2015, Computer Vision and Pattern Recognition.
[6] Liu Y, Stehouwer J, Jourabloo A, et al. Deep Tree Learning for Zero-shot Face Anti-Spoofing[J]. arXiv preprint arXiv:1904.02860, 2019.
[7] Dong Y, Su H, Wu B, et al. Efficient Decision-based Black-box Adversarial Attacks on Face Recognition[J]. arXiv preprint arXiv:1904.04433, 2019.
[8] Zhang S, Wang X, Liu A, et al. A Dataset and Benchmark for Large-Scale Multi-Modal Face Anti-Spoofing[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019: 919-928.