一种基于双流网络的 Deepfakes 检测技术

李旭嵘, 于 鲲

阿里巴巴 杭州 中国 310027

摘要 随着深度学习技术的飞速发展, 以 Deepfakes 为代表的深度伪造技术开始充斥在互联网上的各个角落。Deepfakes 借助于生成对抗网络和自动编码器技术,能够轻松替换人脸以及篡改人的表情信息。此类 Deepfakes 假视频可以制作虚假色情影片、谣言, 传播假新闻, 甚至影响政治选举, 带来的社会影响极其恶劣。然而, 针对此类伪造视频的检测技术还远远落后于生成技术, 已有的工作都存在一定的局限性, 并不能较好地对 Deepfakes 视频进行检测。本文首先对现有生成和检测工作进行综述, 并分析了现有工作的缺陷, 然后提出了基于 EfficientNet 的双流网络检测框架。通过在大规模开源数据集 **FaceForensics++**测试, 我们的检测技术可以在检测 Deepfakes 类假视频上平均准确率达到 99%以上, 并一定程度提高模型对抗压缩的能力。

关键词 深度学习, 深度伪造, 检测, 双流网络
中图法分类号 TP309.2 **DOI 号** 10.19363/J.cnki.cn10-1380/tn.2020.02.07

A Deepfakes detection technique based on two-stream network

LI Xurong, YU Kun

Alibaba Group, Hangzhou 310027, China

Abstract With the rapid development of deep learning technology, Deep forgery techniques, such as Deepfakes, are beginning to fill every corner of the Internet. By utilizing the generative adversarial networks and auto-encoder technology, the Deepfakes replace faces and tamper with facial expressions easily. The Deepfakes can produce fake pornography, spread rumors, spread fake news, and even influence political elections, leading to disastrous social consequences. However, the detection technology for this kind of fake videos is still far behind the generation technology, and the existing works have some limitations. This paper first summarizes the existing generation and detection works, and analyzes the defects of the existing works, then we propose the two-stream network detection framework based on the EfficientNet. By testing on a large open source dataset, FaceForensics++, our detection method was able to detect fake videos with an average accuracy of over 99%, and improve the ability of the model to resist compression to a certain extent.

Key words deep learning, deepfakes, detection, two-stream networks

1 引言

近年来, 网络上逐渐出现各类换脸视频, 此类深度伪造技术开始兴起, 一开始换脸较多的是一些影视明星, 制作虚假的色情视频。利用该技术能实现将一些公众人物的脸移转到色情明星的身体上, 伪造逼真的色情场景。这些虚假的色情视频一经传播, 受害人的名誉将严重受损, 个人隐私还随时有可能受到侵犯。逐渐的, 网络上开始有各种恶搞国家领导人的换脸视频, 不仅影响着一个国家的形象, 甚至也影响着一些国家的选举活动。伴随着深度对抗网络[1]技术的发展, 这些以 Deepfakes[2]为代表的伪造技术越发成熟, 使得假视频以假乱真, 人眼无法辨别。而这种假视频的泛滥使得谣言四起, 个人或者公司、国家利益受损。同时也使得媒体的公信力不断下降。更糟糕的是, 当国家司法机构取证时, 所用视频证据也不再可信, 对法律公正性提出了极大的挑战。CNBC 网站在其报道中称, Deepfakes 将成为“2020 年美国总统选举中的大事件”, Deepfakes 视频将在 2020 年的美国大选中, 掀起强大的血雨腥风[3]。

鉴于此, 中国互联网信息办于 2019 年 11 月 18 日印发了《网络音视频信息服务管理规定》[4], 其中

通讯作者: 于鲲, 研究生 高级算法专家, yukun.yk@alibaba-inc.com。
本课题得到阿里实人认证项目资助。
收稿日期: 2020-01-16; 修改日期: 2020-03-09; 定稿日期: 2020-03-10

明确规定网络音视频信息服务提供者和网络音视频信息服务使用者不得利用基于深度学习、虚拟现实等的新技术新应用制作、发布、传播虚假新闻信息。同时要求网络音视频信息服务提供者应当加强对网络音视频信息服务使用者发布的音视频信息的管理，部署应用违法违规音视频以及非真实音视频鉴别技术。不仅禁止个人传播假视频的要求，也对企业研发假视频检测技术提出要求，使得针对 Deepfakes 的检测尤为重要。

为了检测以 Deepfakes 为代表的假视频，研究者们开始相继提出不同的方案。然而，现有方案准确性不高，泛化能力弱，使得检测方案存在很大的局限性。因此仍然需要一些新思路解决 Deepfakes 检测问题。在本文中，我们首先综述了已有的 Deepfakes 生成和检测工作，并分析指出了现有检测方案的缺陷。接着我们提出了一种基于 EfficientNet[5]的双流检测框架，并在开源数据集 FaceForensics++[6]上进行评测，结果显示，我们模型的检测结果比现有方法更优且提升了模型一定的抗压缩能力。最后，我们讨论了 Deepfakes 检测研究面临的挑战以及未来可行的研究方向。

2 背景与相关工作

2.1 Deepfakes 生成技术总览

人脸替换技术在 3D 领域早有研究，但是复杂度高，成本大，较高的技术门槛使得基于 3D 的换脸技术很难普及。而深度学习尤其对抗生成网络技术的发展使得换脸技术更加逼真和低成本，也进一步催生了以 Deepfakes 技术为代表的一系列深度伪造技术。本节中我们将对基于图形学的生成技术进行简单的介绍，重点关注基于学习的生成算法。

2.1.1 基于图形学的 Deepfakes 生成技术

Thies 等人做了一个实时的脸部表情迁移，通过重建和追踪源和目标演员的 3D 模型[7]，将跟踪的源人脸模型应用到目标模型上，最后融合到原始的目标模型。此后，Thies 等人又提出了更高级的脸部表情替换系统，Face2Face [8]融合 3D 重建和渲染技术，能够实时改变任何来自因特网上的视频中的脸部移动。Suwajanakorn[9]等人提出用网络学习声频和嘴唇动作的映射，最后合成目标人物脸部嘴型到指定动作，但是其合成方法仍与 Face2Face 类似。基于图形学的方法已经研究了多年，但是其较高的计算复杂度使得难以全面推广。

2.1.2 基于学习的 Deepfakes 生成技术

最近两年来，越来越多的人开始采用深度学习技术来生成假脸。尤其生成对抗网络(Generative adversarial network)的发展，大力推进了 Deepfakes 技术的成熟。GAN 被用于改变人脸属性[10]，以及提升人脸图像分辨率等[11]。Deepfakes[2]是第一个融合这些技术的开源方法，基本原理如图 1 所示，核心思想是训练两个自动编码器，而两个编码器共享权重，两个解码器可以分别重建成两个人。完成训练后，人脸 A 可以被相同的编码器编码，而被编码器 B 解码，从而人脸 A 被人脸 B 替换。后续也陆续有些针对 Deepfakes 改进工作，如 FaceSwap-GAN [12]，在原有基础上增加了对抗损失和感知损失。对抗损失作为判别器使得生成的图像和被替换的图像更加接近，感知损失能够让眼球转动的方向更加真实，从而输出质量更高。

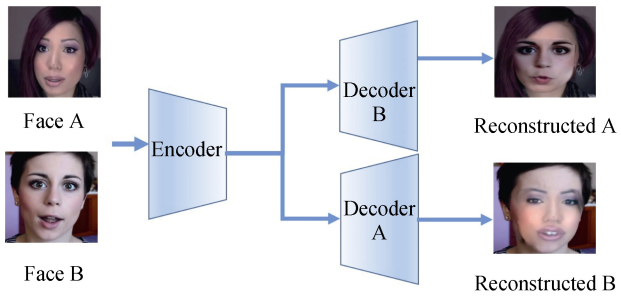

图 1 Deepfakes 基本原理

Figure 1 The principle of Deepfakes

2.1.3 现有 Deepfakes 生成工具总结

目前针对 Deepfakes 生成的工具有两大类，一类是开源的项目，不断的有社区研究爱好者不断改进该工具，这类工具需要使用者有专业的基础知识，使用门槛较高，需要 GPU 资源和大量的训练图片，训练的效果稳定性也不同。另一类是工业界开发的商业软件，不对外开源，有些不需要 GPU 资源，但是，功能单一，只支持固定的人脸替换。本节列举现有的开源工具或者应用软件，并总结他们的特点，供研究者们使用比较。

2.2 Deepfakes 检测技术总览

随着 Deepfakes 假视频的泛滥，研究者们展开了对深度伪造视频的检测工作。现有的检测工作主要分为两大类，一类是基于图片序列信息的检测，主要利用分类器学习图片序列之间的连贯性或者内生属性，如脸部自然转动的轨迹等。另一类是基于单帧图像的视频检测，判断每一帧图像的真假，最后对一个视频的所有帧或者抽取帧进行综合决策。这两类方法各有利弊，总结如表 2 所示:

表 1 Deepfakes 生成工具总览

Table 1 The overview Overview of Deepfakes generation tools

名称	特点
Deepfakes[2]	最早的 Deepfakes 开源工具, 基于自动编码器原理, 只能 1 对 1 生成
FaceSwap[13]	开源的基于 3D 模型融合方式的换脸工具
Deepface-Lab[14]	在 Deepfakes 基础上支持多个基础模型和人脸检测引擎
Dfaker[15]	基于 Keras 实现, 使用 DSSIM 损失函数重建脸。
Deepfake-tf[16]	Dfaker-tf 的 TensorFlow 版本实现
FakeApp[17]	Windows 系统安装的 APP, 只要有 GPU 即可运行两端视频的换脸。需要大量训练集和 GPU 资源。
Zao[18]	手机 APP, 提供一张图片即可替换到制定的影视小片段中。需要人脸图片少, 但是只能替换制定人物的脸。

表 2 检测方法比较

Table 2 The comparison of detection methods

方法类型	优点	缺点
基于图片序列	能够学习时序维度信息, 利用数据多	对帧长度敏感, 存在很多短视频。不能很好地关注局部特征, 无法判断单帧图像
基于单帧图像	能够捕捉图像的局部信息, 利用综合决策能够降低误判率。	不能利用时序信息, 局部特征依赖数据集。

以下将综述两类方法的代表性方法。

2.2.1 基于图片序列的 Deepfakes 视频检测

文献[19]认为个体有不一致的面部表情和移动。通过追踪面部和头部移动然后抽取特定动作集合的存在和强度。用 Openface2 抽取脸部动作单元特征进行编码, 然后训练一个 one-class SVM 分类器进行区分。

在文献[20-21]中作者们均利用一个 CNN 抽取帧内 feature, 紧跟着一个时间感知的 RNN 网络捕获由换脸带来的帧间不一致性。主要框架流程如图 2 所示, 这种利用图片序列的方法无法判断单帧的真伪, 同时对帧的长度很敏感, 但是现实世界中测试集往往长度未知。

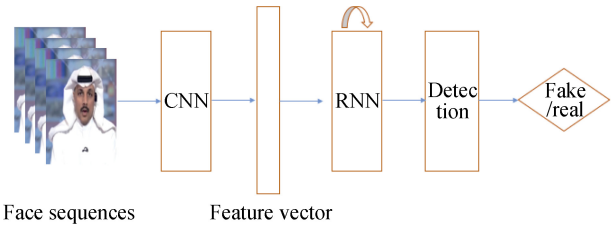

图 2 基于循环神经网络的伪造视频分类基本架构

Figure 2 Basic architecture of fake video classification based on recurrent neural network

2.2.2 基于单帧图像的 Deepfakes 视频检测

目前大多数研究者还是关注基于单帧图像的检测。基于单帧图像检测能够捕捉每一帧图像的局部特征, 既能对图像的真伪做出预测, 又能通过视频的所有帧或者部分帧对视频进行综合预测, 应用性更强。单帧图像的检测大都数是先抽取人脸然后学习真假人脸特征的分布。

Li[22]等人利用人的眨眼特征来检测假视频, Deepfakes 视频制作时缺乏眨眼素材或者生成的人眨眼频率与正常人不同。这种方法在检测初期 Deepfakes 有效, 在新版 GAN 版本 Deepfakes[23]中, 眨眼等动作生成已被改进, 眨眼检测不再有效。此外, Yang 等人[24-25]还利用人的头部姿势, 人脸关键点分布等真人的生物特征来区分真假视频的分布。Matern 等人[26]聚集生成的假脸中不对称的牙齿、眼睛等细节信息, 用神经网络学习这种局部特征。这几种方法均能检测早期比较粗糙的 Deepfakes 视频, 如存在明显篡改痕迹, 脸部歪曲等, 无法较好应对改进的新版 Deepfakes 视频。此后, 研究者开始直接采用数据驱动的方法来学习整张脸。Rössler 等人[6]用 Xcepption 网络进行预训练能在单独的数据集上达到不错的检测率, Afchar 等人[27]搭建浅层 CNN 网络学习真假人脸的微观特征, Nguyen 等人[28]搭建自己的胶囊网络架构, 来对 VGG 网络提取的特征进行分类。此类基于 CNN 的方法在单一的篡改方法上展现了较高的检测率。但同时也存在很多缺陷, 如依赖特定的篡改方法数据集, 依赖特定的压缩率, 即如果篡改方法和压缩率未知, 检测效果将大大下降, 这些仍然是 Deepfakes 检测的难点。

3 基于 EfficientNet 的双流 Deepfakes 检测方案

本文从基于单帧图像的检测入手, 致力于设计鲁棒的检测网络。Zhou 等人在[29-30]指出利用图像的噪音特征可以一定程度抵御图像压缩的影响。而 Deepfakes 视频通常作不同程度的压缩, 带来检测上的困难。因此本研究引入噪音流特征, 以此来提高图像特征对抗压缩的能力。本文实验所采用的图像噪音特征如图 3 所示。图 3 中左图为原图, 右图为噪音图。但是, 与 RGB 图不同的是, 噪音特征丧失了大部分局部信息, 不利于分类学习任务。因此融合 RGB 图的学习设计双流架构, 从而可以从不同维度学习图像的信息。此外, 考虑到 Deepfakes 视频人脸大小和分辨率的多样性, 本文采用 EfficientNet 作为双流学习任务的主干网络, 检测的主体框架如图 4 所示,

框架分为上下两条流, 输入不同的图像信息, 分别独自训练。与 Zhou 等人[29]中的双流网络不同的是, 本文所关注的提取噪声区域是脸部的改动或者替换, 一些是深度学习方法生成的无痕迹的替换, 而不是论文[29]的部分内容复制或者拼接。此外本文应对的是 Deepfakes 视频级的压缩, 整个视频的压缩与图片压缩有很大的不同。视频级的压缩会受前后帧编码的影响, 而图片的压缩是独立的。并且在文献[29]中, 作者是将两条流的中间特征层进行融合, 而本文中是在模型决策层进行融合, 两条流单独训练。因此, 本文与论文[29]从模型结构和框架到应用场景, 均不相同。本文是首次利用噪音流解决 Deepfakes 的压缩特征问题, 并同 EfficientNet 联合提升 Deepfakes 在不同复杂场景下的检测率。由于文献[22]解决的是图像拼接复制或者移除问题, 评判的是篡改区域的定位准确性, 与本文 Deepfakes 数据集相差甚大, 无法在同一维度进行对比, 因此本文后续实验中会选择 Deepfakes 领域的模型算法进行对比分析。

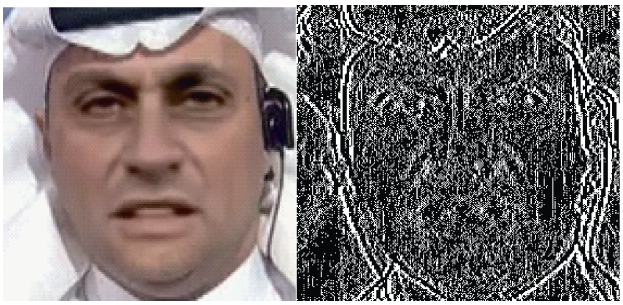

图 3　图像噪音示例

Figure 3　The sample of image noise

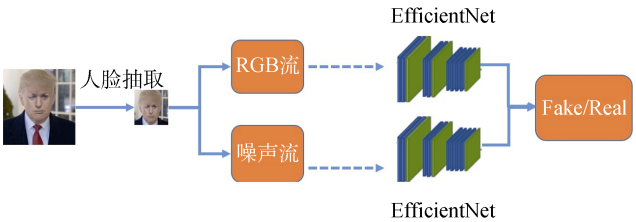

图 4　检测网络框架

Figure 4　The framework of detection networks

当输入视频时, 先将视频截取成帧序列, 然后利用人脸检测器抽取帧中人脸信息。将视频的序列人脸输入模型的两个分支, 下分支使用噪音过滤器[31]获取人脸的噪音特征, 输送到 EfficientNet 网络(E_{noise})进行单独训练, 上分支(E_{rgb})直接对整张人脸信息进行训练, 学习真假人脸特征的分布差异。最后在输出对两个分支的结果进行融合。由于 Deepfakes 的压缩特征并不是一直存在, 跟人脸特征的决策融合时需要设置权重, 融合方式如下:

$$Prob(x) = \alpha * E_{noise}(x) + (1-\alpha) * E_{rgb}(x)$$

其中, x 是帧输入, α 是可调节的平衡因子, $Prob(x)$ 代表 x 的预测概率。最后, 对视频的所有帧进行预测, 预测结果的均值作为该视频是否属于 Deepfakes 的概率。

4　实验结果与分析

4.1　评测数据集和模型

本文选取业界较全对比最多的 FaceForensics++[6]数据集对模型进行评测。FaceForensics++数据集第一个大规模深度伪造研究数据集。其素材是来自于 YouTube 上收集的 1000 个原始视频, 这些视频均有可追踪的清晰的人脸, 没有遮挡能够较好地实现人脸篡改替换等, 该数据集中对这 1000 个原始视频分别进行 Deepfakes[2], FaceSwap[13], Face2Face[8], NeuralTextures[32]四种方式的篡改, 其中, Deepfakes 和 FaceSwap 为换脸类型篡改, Face2Face 和 NeuralTextures 为表情类型篡改, 分别生成对应的 1000 个假视频, 共 4000 个。同时对真假视频均进行 H.264 编码中的 C0,C23,C40 三种参数压缩, 即整个数据集有 15000 个视频。此外, 该数据集还提供了假视频篡改区域的 mask。真假视频的视频帧示例如图 5 所示, 篡改类型为 Deepfakes 换脸。其中(a)与(c)是两个人的原图, (b)为(c)的人脸替换在(a)上的造假图, (d)为(a)的人脸替换在(c)上的假图。

在本文的实验中, 为了避免大量相似帧, 对所有的视频每秒截取 5 帧。并用人脸检测器抽取出人脸框, 截取人脸时以人脸框为基准, 向外扩展 0.3 倍大小。在模型选择上, 模型上本文选择 EfficientNet-b4 模型, 并使用预训练模型进行训练, 使用 Adam 优化器, 初始学习率为 0.001, 每 10 轮学习率下降 10 倍, 直到模型收敛。平衡因子 α 设为 0.1。在评测验证集时, 为了跟 Xception[6]中的实验结果对比, 所有指标均在视频级别评测, 即取一个视频所有抽取帧的预测值均值作为该视频的预测概率。对视频的判断阈值设置为 0.5, 大于 0.5 的认为是假视频。实验的评估指标有 *TPR* (True Positive Rate), *TNR*(True Negative Rate), *Acc*. 其中

$$Acc = \frac{TP + TN}{TP + FP + TN + FN}$$

TPR 和 *TNR* 分别代表真假视频的识别准确率。*Acc*

代表整体数据集的准确率。

(a) (b) (c) (d)

图 5 FaceForensics++数据集 Deepfakes 篡改示例

Figure 5 The examples of FaceForensics++ dataset on Deepfakes

4.2 跨压缩率实验

为了验证噪音流的有效性, 本文测试只有 RGB 单流和双流框架下在不同压缩率下的性能对比, 为了避免重复, 选取 Deepfakes 数据集作为测试集。由于 C0 参数是视频的 Raw 格式, 没经过任何压缩, 在网络上很少见, 因此采用 C0 参数训练的模型不具有泛化能力, 而 C23 参数压缩的高清视频较为常见。采用 C23 参数压缩的视频训练的模型, 如果能检测较高压缩率的假视频, 则此模型能对抗一定的压缩。因此本实验使用 Deepfakes 的 C23 压缩版本训练, 然后在 C40 版本测试。表 3 为模型跨压缩率(C40)实验结果。从表 3 中可以看出, 在同样的主干网络下, 增加噪音流能够明显提升模型在跨压缩数据集上的性能。鉴于此, 本文后续实验均在双流网络框架下进行。

表 3 Deepfake 跨压缩率实验结果

Table 3 The experimental results of cross compressions on Deepfakes

模型方法	TPR(%)	TNR(%)	Acc(%)
RGB 单流	72.7	98.5	86.5
Noise-RGB 双流	77.1	99.3	88.2

4.3 独立数据集实验

为了验证在不同数据集上的训练效果, 本文分别在 FaceForensics++数据集的四个数据集上测试双流网络的性能, 并与 Xception[6]的结果进行对比。即在四种篡改方法的 3 种压缩率版本下分别训练, 测试独立数据集上的性能。由于 NeuralTextures 是最新集成的数据集, Xception[6]中并没有提供测试结果。相关测试结果如表 4 所示, 其中"-"代表无此实验。从表 4 中可以看出, 模型在不同压缩率下的性能有差异, 压缩率高的视频丧失了很多特征使得检测难度变大, 但仍然平均达到 93%以上的准确率。在 C0 和 C23 的单独实验上, 本文提出的模型在视频级别的测试均达到了 99%以上的准确率甚至个别达到 100%, 已经可以较准确地对不同的篡改方式进行分类。再跟 Xception[6]的最好结果对比, 本文提出的双流网络在单独测试集上的性能均优于 Xception[6]。尤其在高压缩率的实验上, 双流网络展现极大的提升, 展现了良好的对抗压缩能力。

表 4 双流网络在四种独立数据集上的性能

Table 4 The performance of two-stream network on four independent datasets

数据集	Acc (%) (Two stream)	Acc(%) (Xception)
C0		
Deepfakes	99.64	99.06
FaceSwap	100	99.61
Face2Face	100	99.14
NeuralTextures	100	-
C23		
Deepfakes	99.50	98.76
FaceSwap	99.8	98.59
Face2Face	99.5	98.53
NeuralTextures	100	-
C40		
Deepfakes	93.57	93.46
FaceSwap	95.4	89.80
Face2Face	96.4	92.72
NeuralTextures	99.2	-

4.4 跨数据集实验

跨数据集检测仍然是个难点, 即在一种篡改方法上训练的模型在另一个篡改方法上测试效果会大幅度下降。现有的做法是将已有的篡改方法融合到训练集再测试。本文中将 FaceForensics++中的四种篡改方法融合到一起用双流网络训练, 不同的压缩率分别训练模型, 最后测试结果如表 5 所示。

表 5 双流网络在四种数据集上训练的性能

Table 5 The performance of two-stream network trained on four different datasets

All	Acc(%) (Two stream)	Acc(%) (Xception)
C0	100	99.41
C23	99	97.53
C40	97.57	85.49

从表 5 可以看出, 本文提出的双流网络在综合数据集上训练的性能比 Xception[6]要好, 尤其针对 C40 压缩率比较高的数据集上, 表现远远优于 Xception[6], 体现了噪音流的优越性。多种数据集的融合训练并没有降低模型在每个篡改数据集上的测试性能。另外实验结果也发现, 采取不同篡改方法的数据融合训练, 一定程度上能提高在高压缩率数据上的表现。因此针对新的篡改方法, 将新的篡改类型数据集扩充到训练集是一个有效的解决方案。

5 当前挑战和未来方向

尽管目前已有不少学者展开了针对Deepfakes的检测研究, 但是目前 Deepfakes 的检测仍然存在诸多关键问题没有解决。如视频压缩问题, 视频分辨率问题, 篡改算法问题。这些问题如今学术界仍然没有完美的解决方案, 将在本节逐一讨论。

首先是压缩问题, 互联网中的视频均会遭受不同程度的压缩, 而实验已经证明压缩会影响模型的检测效果。由第四节中的实验知道, 在低压缩率数据下训练的模型检测高压缩率数据, 检测效果会大幅度下降, 本文探索的噪音流模型能一定程度减少这个差距, 但是还不能根本上消除这个问题。甚至如果编码方式不再是H.264编码, 那么不同的压缩率又会带来训练测试很大的差距。当然, 如果视频收到高度压缩, 本身的质量会大幅下降, 影响 Deepfakes 视频的传播效果。对于压缩的解决方法不能只依赖于基于数据驱动的方法, 这样训练的模型会强烈依赖于数据集的压缩类型, 泛化能力很差。而传统取证领域使用的一些拼接检测的噪音提取或者残差[33-37]也会受压缩而导致特征提取困难和消失。本文探索的噪音流是一个研究压缩的开端, 利用传统取证领域提取的特征, 不依赖于图像的局部特征, 而是提取在不同压缩率下共存的特性, 这个是未来研究 Deepfakes 模型抗压缩的方向之一。

其次是视频分辨率问题。网络上有不同分辨率的视频, 也会导致视频中的人脸大小各异。若人脸大小跨度很大, 这样对模型测试提出很大的挑战。人脸经过统一的放缩后, 原有特征都会一定程度丢失, 使得用于检测分类的特征无分布法一致, 从而降低模型检测率。面对此类方法一种可行的方案就是搭建多尺度网络, 针对不同分辨率的数据集进行训练, 然后最后进行特征融合。

最后是篡改算法问题。未知的生成方式层出不穷, 基于数据训练的方法的缺陷之一就是对未知类型不鲁棒。这是因为不同篡改类型的篡改特征分布不一样, 使得基于学习的模型过于关注局部特征而无法适应其他篡改类型的分布。应对此类问题一种解决方案是扩充训练集, 不断融合新的篡改类型, 如本文第四节实验。但是这种做法无法从根本上解决问题, 在面临网络上未知类型篡改依然不鲁棒, 只能尽可能多地检测已有篡改类型。解决此类问题的可行方案是探索以异常检测思路来做Deepfakes视频检测。真实视频是大量存在的, 可以训练学习真实视频的分布, 然后用已有类型的假视频来确定真假视频的界限。当模型预测视频时, 只要是界限之外的均当做假视频, 此类方法需要确定精准的阈值。

最近有些研究者提出了一些新思路解决了部分问题。如论文[38-40]均是利用 mask 信息做语义分割, 并促进分类任务的提升。提升了单模型的性能, 却无法解决上述核心问题。论文[41-42]则是利用 GAN 图像的特点专注辨别 GAN 生成的图像。针对 GAN 生成的有效, 而非 GAN 形式的假视频则无法奏效。此类新思路一定程度提升了现有模型或者解决了部分问题, 但是如压缩、分辨率、跨生成算法等难点问题仍然是一个艰巨的挑战。Deepfakes 的检测是一个持久性问题, 需要相关的技术爱好者不断的投入付出, 成立相关的技术社区。当前除了开源的 FaceForensics++ 数据, 谷歌公司也开源了一批 Deepfakes 视频供大家研究[43]。此批视频由演员拍摄, 谷歌公司进行人脸修改替换等操作。Facebook 公司最近也联合学术机构展开了 Deepfakes 竞赛[44], 竞赛数据且覆盖了前述的难点问题, 如未知压缩, 多样分辨率, 未知篡改类型等。此竞赛将会有力推动检测工作的进展。想要较快较好地解决 Deepfakes 检测问

题, 仅仅靠技术当前无法完美的解决。需要的是政府、法律机制的完善。中国相关管理机构出台了《网络音视频信息服务管理规定》[4], 其中四次直接提及深度学习, 基本可以看做是针对 AI 造假视频的一次针对性管控。另外推广区块链技术和视频的融合, 如能对发布的视频进行溯源[45], 也能大幅度削减 Deepfakes 的泛滥程度。

6 结论

Deepfakes 如今变得炙手可热, 而伴随着生成技术的发展, Deepfakes 带来的负面影响越来越多, 针对 Deepfakes 类假视频的检测越发重要。本文首先总结了当前 Deepfakes 的生成工作和已有的检测工作, 并指出了检测方法的一些缺陷。此外, 本文提出了一个基于 EfficientNet 的双流检测网络, 噪音流一定程度上能够提高模型的抗压缩能力。通过在大规模数据集 FaceForensics++上的评测, 本文提出的架构无论在单数据集还是综合数据集上的测试性能均比现有方法好。最后讨论了 Deepfakes 检测领域的一些研究挑战和未来可行的研究方向。提出了一些针对 Deepfakes 领域研究的有效建议, 旨在为推动 Deepfakes 检测的研究和应用部署提供一定帮助。

参考文献

[1] Goodfellow I, Pouget-Abadie J, Mirza M, et al. Causal Categorization with Bayes Nets[M]//Advances in Neural Information Processing Systems 14. 2014: The MIT Press, 2002: 2672-2680.

[2] Deepfakes, https://github.com/deepfakes/faceswap, Sept. 2019.

[3] Deepfakes policy, https://www.csis.org/analysis/trust-your-eyes-deepfakes-policy-brief, Sept.2019.

[4] 网络音视频信息服务管理规定, http://www.cac.gov.cn/2019-11/29/c_1576561820967678.htm, November. 2019.

[5] Tan M X, Le Q. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks[EB/OL]. 2019: arXiv:1905.11946 [cs.LG]. https://arxiv.org/abs/1905.11946.

[6] Rossler A, Cozzolino D, Verdoliva L, et al. FaceForensics++: Learning to Detect Manipulated Facial Images[C]. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, October 27-November 2, 2019. Seoul, Korea (South). Piscataway, NJ: IEEE, 2019: 209-215.

[7] Thies J, Zollhöfer M, Nießner M, et al. Real-time Expression Transfer for Facial Reenactment[J]. *ACM Transactions on Graphics*, 2015, 34(6): 1-14.

[8] Thies J, Zollhofer M, Stamminger M, et al. Face2Face: Real-Time Face Capture and Reenactment of RGB Videos[C]. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 27-30, 2016. Las Vegas, NV, USA. Piscataway, NJ: IEEE, 2016: 2387-2395.

[9] Suwajanakorn S, Seitz S M, Kemelmacher-Shlizerman I. Synthesizing Obama[J]. *ACM Transactions on Graphics*, 2017, 36(4): 1-13.

[10] G. Antipov, M. Baccouche, J. Dugelay. Face aging with conditional generative adversarial networks[EB/OL]. 2017: CoRR, abs/1702.01983.

[11] Karras T, Aila, Laine S, et al. Progressive Growing of GANs for Improved Quality, Stability, and Variation[EB/OL]. 2017: arXiv:1710.10196[cs.NE].https://arxiv.org/abs/1710.10196.

[12] Faceswap-GAN, https://github.com/shaoanlu/faceswap-GAN, Sept. 2019.

[13] FaceSwap, https://github.com/MarekKowalski/FaceSwap Sept. 2019.

[14] DeepFaceLab, https://github.com/iperov/DeepFaceLab Sept. 2019.

[15] dfaker, https://github.com/dfaker/df Sept. 2019.

[16] DeepFake_tf, https://github.com/StromWine/DeepFake_tf Sept. 2019.

[17] fakeapp, https://www.malavida.com/en/soft/fakeapp/#gref Sept. 2019.

[18] Zao app, https://www.zaoapp.net/ Sept. 2019.

[19] Agarwal S, Farid H, Gu Y, et al. Protecting World Leaders Against Deep Fakes[C]. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops(CVPRW)*. 2019: 38-45.

[20] Guera D, Delp E J. Deepfake Video Detection Using Recurrent Neural Networks[C]. *2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, November 27-30, 2018. Auckland, New Zealand. Piscataway, NJ: IEEE, 2018: 1-6.

[21] Sabir E, Cheng J X, Jaiswal A, et al. Recurrent Convolutional Strategies for Face Manipulation Detection in Videos[EB/OL]. 2019: arXiv:1905.00582[cs.CV]. https://arxiv.org/abs/1905.00582.

[22] Li Y Z, Chang M C, Lyu S W. In Ictu Oculi: Exposing AI Created Fake Videos by Detecting Eye Blinking[C]. *2018 IEEE International Workshop on Information Forensics and Security (WIFS)*, December 11-13, 2018. Hong Kong, China. Piscataway, NJ: IEEE, 2018: 1-7.

[23] Korshunov P, Marcel S. Deepfakes: a new threat to face recognition? assessment and detection[EB/OL]. 2018: arXiv preprint arXiv:1812.08685.

[24] Yang X, Li Y Z, Lyu S W. Exposing Deep Fakes Using Inconsistent Head Poses[C]. *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 12-17, 2019. Brighton, United Kingdom. Piscataway, NJ: IEEE, 2019: 8261-8265.

[25] Yang X, Li Y Z, Qi H G, et al. Exposing GAN-synthesized Faces Using Landmark Locations[EB/OL]. 2019: arXiv:1904.00167 [cs.CV]. https://arxiv.org/abs/1904.00167.

[26] Matern F, Riess C, Stamminger M. Exploiting Visual Artifacts to Expose Deepfakes and Face Manipulations[C]. *2019 IEEE Winter Applications of Computer Vision Workshops (WACVW)*, January 7-11, 2019. Waikoloa Village, HI, USA. Piscataway, NJ: IEEE, 2019: 83-92.

[27] Afchar D, Nozick V, Yamagishi J, et al. MesoNet: A Compact Facial Video Forgery Detection Network[C]. *2018 IEEE International Workshop on Information Forensics and Security (WIFS)*,

December 11-13, 2018. Hong Kong, China. Piscataway, NJ: IEEE, 2018: 1-7.

[28] Nguyen H H, Yamagishi J, Echizen I. Capsule-forensics: Using Capsule Networks to Detect Forged Images and Videos[C]. *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 12-17, 2019. Brighton, United Kingdom. Piscataway, NJ: IEEE, 2019: 2307-2311.

[29] Zhou P, Han X T, Morariu V I, et al. Learning Rich Features for Image Manipulation Detection[C]. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, June 18-23, 2018. Salt Lake City, UT, USA. Piscataway, NJ: IEEE, 2018: 1053-1061.

[30] Zhou P, Han X T, Morariu V I, et al. Two-Stream Neural Networks for Tampered Face Detection[C], *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, July 21-26, 2017. Honolulu, HI, USA. Piscataway, NJ: IEEE, 2017: 1831-1839.

[31] Fridrich J, Kodovsky J. Rich Models for Steganalysis of Digital Images[J]. *IEEE Transactions on Information Forensics and Security*, 2012, 7(3): 868-882.

[32] Thies J, Zollhöfer M, Nießner M. Deferred Neural Rendering: Image Synthesis Using Neural Textures[EB/OL]. 2019: arXiv:1904.12356[cs.CV]. https://arxiv.org/abs/1904.12356.

[33] Cun X D, Pun C M. Image Splicing Localization Via Semi-global Network and Fully Connected Conditional Random Fields[M]. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2019: 252-266.

[34] Bappy J H, Roy-Chowdhury A K, Bunk J, et al. Exploiting Spatial Structure for Localizing Manipulated Image Regions[C]. *2017 IEEE International Conference on Computer Vision (ICCV)*, October 22-29, 2017. Venice. Piscataway, NJ: IEEE, 2017: 4970-4979.

[35] Cozzolino D, Gragnaniello D, Verdoliva L. Image Forgery Detection through Residual-based Local Descriptors and Block-matching[C]. *2014 IEEE International Conference on Image Processing (ICIP)*, October 27-30, 2014. Paris, France. Piscataway, NJ: IEEE, 2014: 5297-5301.

[36] Huh M, Liu A, Owens A, et al. Fighting Fake News: Image Splice Detection Via Learned Self-Consistency[M]. Computer Vision – ECCV 2018. Cham: Springer International Publishing, 2018: 106-124.

[37] Korus P, Huang J W. Multi-Scale Analysis Strategies in PRNU-Based Tampering Localization[J]. *IEEE Transactions on Information Forensics and Security*, 2017, 12(4): 809-824.

[38] Bappy J H, Simons C, Nataraj L, et al. Hybrid LSTM and Encoder–Decoder Architecture for Detection of Image Forgeries[J]. *IEEE Transactions on Image Processing*, 2019, 28(7): 3286-3300.

[39] Cozzolino D, Thies J, Rössler A, et al. ForensicTransfer: Weakly-supervised Domain Adaptation for Forgery Detection[EB/OL]. 2018: arXiv:1812.02510[cs.CV]. https://arxiv.org/abs/1812.02510.

[40] Nguyen H H, Fang F M, Yamagishi J, et al. Multi-task Learning for Detecting and Segmenting Manipulated Facial Images and Videos[EB/OL]. 2019: arXiv:1906.06876[cs.CV]. https://arxiv.org/abs/1906.06876.

[41] Xuan X S, Peng B, Wang W, et al. On the Generalization of GAN Image Forensics[M]. Biometric Recognition. Cham: Springer International Publishing, 2019: 134-141.

[42] Wang R, Juefei-Xu F, Ma L, et al. FakeSpotter: A Simple yet Robust Baseline for Spotting AI-Synthesized Fake Faces[EB/OL]. 2019: arXiv:1909.06122[cs.CR]. https://arxiv.org/abs/1909.06122.

[43] N. Dufour , A. Gully. Contributing data to Deepfake detection research, Google AI blog. Available: https://ai.googleblog.com/2019/09/contributing-data-to-deepfake-detection.html.

[44] M. Schroepfer. Creating a data set and a challenge for deep fakes, Facebook AI blog. Available: https: //ai.facebook.com/ blog/deepfake-detection-challenge

[45] Hasan H R, Salah K. Combating Deepfake Videos Using Blockchain , Smart Contracts[J]. *IEEE Access*, 2019, 7: 41596-41606.

李旭嵘 于 2015 年在南京邮电大学信息安全专业获得学士学位，现在浙江大学计算机科学与技术专业攻读博士学位，研究领域为深度学习安全，计算机视觉。研究兴趣包括对抗攻击与防御，深度伪造检测等。
Email: lixurong@zju.edu.cn

于鲲 于 2007 年毕业于华中科技大学模式识别与智能系统专业，硕士学位。现任阿里巴巴风险能力中台高级算法专家。研究领域为人脸识别、活体攻防等。研究兴趣包括：大规模人脸检索、活体检测、人脸改检测等。
Email: yukun.yk@alibaba-inc.com