深伪视频检测综述

清华大学软件学院2020010108徐浩博

摘要:深伪(deepfake)一般指基于深度学习(deep learning)的模型修改原图片或视频的人脸,获得伪造(fake)的图片或视频的一种技术。伴随深度学习的方法技术不断发展,伪造出的图片和视频越来越逼真;虽然在一些电影特效或者社交媒体上娱乐等方面,深伪起到一定积极的作用;然而由于伪造照片和视频极易与金融诈骗、名誉侵权、新闻伪造相关,一些照片和视频甚至会伪造国家领导人的相关肖像,从而对国家安全等产生威胁,因此深伪检测技术的发展就有很强的现实意义。本文将从深伪视频数据集出发,较为系统地介绍并比较深伪检测常用的数据集;然后我们对深伪检测的算法进行分类并逐一介绍。最后,文章将会对深伪检测领域的未来发展作简单的展望。

关键词:深伪检测;数据集;深度学习

Deepfake Video Detection: A Survey

XU Haobo 2020010108

(School of Software, Tsinghua University,

Beijing 100084, China)

Abstract: Deepfake generally refers to the technology of modifying the face of the original picture or video based on the deep learning model to obtain fake pictures or videos. With the continuous development of deep learning methods and technologies, forged pictures and videos are becoming more and more realistic; although in some movie special effects or social media entertainment, deep forgery plays an active role; however, it is very easy to forge photos and videos. Involving financial fraud, reputation damage, news fraud, etc., some photos and videos may even forge relevant portraits of national leaders, thereby posing a threat to national security. Therefore, the development of deep forgery detection technology has strong practical significance. This article will start with the deep pseudo-video data set, systematically introduce and compare the commonly used data sets for deep pseudo-detection; next we will classify and introduce the deep pseudo-detection algorithms one by one. Finally, the article will give a brief outlook on the future development of the deep fraud detection field.

Keyword: Deepfake Detection; Dataset; Deep Learning  
1引言

深伪算法通常采用自动编码器(Autoencoding)或生成对抗网络(GAN)来达到伪造的目的。具体来说,深伪可以分为以下几种模式:1)面孔互换;2)整脸伪造;3)脸部属性修饰[1]。因此无论是深伪数据集的生成还是深伪检测算法,都围绕这几种伪造方法来展开。

深伪检测技术的发展离不开两大基石:一是数据集的构建,二是检测算法的更新。

首先是数据集,数据集无论是作为检测效果的测试,还是作为基于一些深度学习的模型的训练集,都有着关键的意义。第一代数据集包括UADFV、FaceForensics++等;然而第一代数据集存在的问题是数据规模小、可变性不强(如光照条件、拍摄场景单一等)、图片/视频质量过低等。基于此,如DeepFakeDetection、DFDC Preview等第二代数据集应运而生;它们一定程度上弥补了原有数据集的漏洞,并且还大量扩充了深伪视频的内容[2]。我们将在第2节详细讲述这些内容。

其次是检测算法,大致可归结为基于深度学习模型的检测、基于时序特征的检测、基于视觉细节的检测、基于生物特征的检测等四类。这些内容我们将在第3节展开。

2深伪视频数据集

2.1 UADFV

UADFV是2018年为了检测人像眨眼率提出的视频数据集[3]。作为较早的视频数据集,它在Youtube上采集了49个视频,并利用dlib库提取人脸地标,再用相似变换矩阵将假人像仿射回图像,最后进行细节处理。但由于分辨率过低,数据量小,主要采用面孔互换方法伪造,丰富性不足等,仍然有诸多缺陷。

2.2 FF/FF++/DFD

FaceForensics是第一代数据集,于2018年发布,它同时也是最早大规模运用深伪生成算法构建出的数据集之一[4]。它主要从Youtube等视频网站上采集千余个主持人新闻播报视频,并采用Face2Face的面孔互换方法。视频质量有了进一步提升,但伪造方法单一且落后,判断真伪性难度较低。大多数深伪检测算法AUC都可以接近100%。

FaceForensics++是对于FaceFornsics数据集的延拓[5]。FF++于2019年发布,它充分利用了常见的Deepfakes、FaceSwap、NeuralTextures 和 Face2Face四种视频伪造方法,对约一千段原始视频进行了伪造,生成约四千段伪造视频;这几种方法包含了自动编码器、图形学算法和神经纹理渲染的深度学习模型,方法较为丰富。此外,数据集还提供了Raw,Crf 23和Crf40三种分辨率,是目前应用较广的数据集之一。

在Google公司的支持下,DeepFakeDetection (DFD)数据集在2019年也被纳入FF++数据集范畴[6]。该数据集生成算法更优具有更高的真实感,且充分模拟了社交网络上视频处理技术,提供了若干常见视频格式和视频质量。

2.3 DFFD

DFFD为美国密歇根州立大学于2019年发布的数据集[7];该数据集的优势在于深伪生成方法广泛。他们充分利用了面孔互换、整脸伪造、脸部属性修饰、脸部表情动作修饰这四种生成方法,每种方法均运用当时最前沿的开源技术生成。除此之外,他们采样数据的男女比例均衡、年龄分布较为广泛。他们还充分考虑到了图片质量大小、图片旋转角度的多样性,因此DFFD是较为丰富且全面的数据库。

2.4 Deepfake-TIMIT

Deepfake-TIMIT数据集由Korshunov和Marcel于2018年发布[8]。该数据集包括来自VidTIMIT数据库的六百余个假视频,通过基于GAN的换脸算法生成。数据集还充分考虑了不同场景和不同视频质量,针对不同视频质量,他们还特别修改了混合算法,使得视频伪造性更强。

2.5 DFDC Preview/DFDC

The Deepfake Detection Challenge(DFDC)是Facebook公司面向全球举行的深伪检测竞赛。2019年,比赛前夕,Facebook公布了预演数据集DFDC Preview,由一千余真实视频和四千余伪造视频组成[9]。2020年,DFDC竞赛上,Facebook公布了DFDC数据集[10].它基于多种深度伪造算法,如基于GAN的star-GAN算法等和非深度学习伪造算法生成,还通过多重图像处理技术增强伪造效果。总数据数量超过十万。

2.6其他常见数据集

除以上数据集外,常见数据集还有纽约州立大学奥法罗分校发布的Celeb-DF[11],商汤公司制作的ForgeryNet[12]等。其中ForgeryNet是目前公开的最大深伪数据集,采用的伪造方法、扰动方法数目也是较为丰富的。

表1:常见深伪数据集对比表

数据集发布时间图片伪造视频真实视频

UADFV REF \_Ref90069806\r \h [3]2018√4949

Deepfake-TIMIT [8]2018×620-

Celeb-DF [11]2018×5639890

FaceForensics [4]2019×10001000

FaceForensics++[5]2019×40001000

DeepFakeDetection [6]2019×3608363

DFFD REF \_Ref90070261\r \h [7]2019√30001000

DFDC Preview [9]2019×41191131

DFDC [10]2020×10450023564

ForgeryNet [12]2021√12161799630

3深伪视频检测方法

3.1基于深度学习模型的检测

此种检测方法实际上是深度学习模型中分类(classification)问题的延拓。对于图像,深度学习方法已经较为成熟;对于视频,可以将其视为一个图像集合。取视频中的若干帧序列,对于每帧图像均做处理,最终对于整个视频,也即图像集合,可以采用平均或投票的方法,判断最终的真伪性[13]。

Chollet et al.基于Inception V3卷积神经网络进行进一步优化改进,提出XceptionNet[14]。Rossler et al.利用Xception对FF++数据集进行了评估,结果优于其他大多数网络[15]。Tan et al.则在卷积时通过设定CNN深度、宽度和图像分辨率三个维度的系数,从而提高了CNN的精度与速度,这种网络也被称为EfficientNet[16]。以上两种方法则是深伪检测中识别帧序列最为常用的两种backbone,除此之外还有若干种先进的卷积神经网络,如ResnNet等,它们共同为深伪检测奠定了基石。Dang et al.在CNN基础上引入了注意力机制,取得了较好效果[17]。 Wang et al.建立的是浅层神经网络[28],但他们监视了神经元的活动,最终也取得了不错的效果。

McCloskey和Albright 发现基于GAN生成的伪造视频和图像在颜色细节上会存在瑕疵,他们基于此利用SVM作为分类器进行深伪检测[18]。然而这种方法的局限性在于可以检测的图片生成种类单一,且伴随GAN伪造图片技术的改进,这一检测手段显然会面临淘汰的窘境。

Zhou et al.利用一种双流(Two-stream)方式来进行深伪检测[19],如图1。其中一个流是人脸分类流;它基于CNN的GoogLeNet,通过真实人脸和被标记的伪造人脸的学习被训练成为一个二分类器。另一个流是基于SVM对于图片隐写分析特征(steganalysis features)的训练,具体来说,首先选择一个三元组(锚定区域、真图片中的区域和假图片的区域),通过对三元组的学习使得SVM成为可以发掘图片特定区域背景噪声等细节的分类器,之后通过分类器对于人脸区域的判断,作出图片真伪的判别。两个流的分数加权作为最终的检测分数,并给出判断。

图1:Zhou et al.提出的双流深伪检测方式:一个流进行人脸分类,一个流进行隐写分析特征[19]

3.2基于时序特征的检测

该种检测方法首先由Güera et al.提出[20],文章首先提出由于视频伪造时基本会对每帧进行伪造,因此帧与帧之间就会存在诸如抖动等差异,检测时序上帧的差异,也就可以分辨出视频真伪。Güera et al.首先利用Inception V3的CNN对于视频的各个帧进行分析,然后利用LSTM的RNN对于序列进行整合分析。最后他们用自动编码器生成自己的数据集并进行了测试,具有不错的效果。但如果伪造的视频针对检测的特征进行优化改进,则该检测方法准确率就会降低。

Masi et al.对上述方法进行了改进,提出双分支方法[21]。其中一个分支对颜色域中的数据进行处理,另一个分支学习频域信息,利用高斯拉普拉斯算子(LoG)抑制面孔的信息,从而放大伪影。两个分支汇总在DenseNet的CNN中,再通过双向LSTM的RNN。最后,他们还对损失函数(Loss Function)进行了改进。他们的方法架构如图2所示。

图2:Masi et al.提出的双分支方法:一个分支处理颜色域,一个分支处理空间频域[21]

3.3基于视觉细节的检测

由于人脸合成必须要将新面孔的图像混合在原图像上,所以混合后的边界可能出现边界模糊、亮度不一致等等问题,通过这些问题的检测可以达到深伪检测的目的。除此之外,由于目标图像和原图像可能采用不同的拍摄相机,一些相机指纹(camera fingerprint)也可能让深伪检测成为可能。

Li和Lyu[3]认为人像是将目标面孔通过仿射变换映射到原图片上,会引起图片扭曲,使得图片的分辨率和颜色出现差异。然后他们通过CNN捕捉这种伪影差异,即可分辨图片真伪。最终通过随机选取帧进行检测来投票判断视频真伪。这样做的好处是不需要生成负样本测试,因为图片经过简单的变换即可作为负样本进行训练,所以可以避免大规模生成深伪视频的时间和硬件资源浪费。最终对比多个种类的CNN,他们选取ResNet50作为最终的选择,并取得了不错的效果。但该方法未在大量压缩视频上进行测试,因此处理多样性强的样本效果可能欠佳。

Li et al.定义脸部X光(face X-ray)为伪造人脸和背景图片的边界[22]。他们也观察到大多数图片或视频伪造时均会将目标人像混合进背景图像中,并由此产生图像差异;通过检测这种差异并试图通过发现face X-ray分离混合的人像和背景图片,从而达到鉴别真伪的目的,如图3所示。但此种方法只关注边界,对于全伪造或是特别处理的图片就束手无策了,因此泛化能力有待提高。

图3:Li et al.提出的脸部X光检测,预测的混合图片边界与实际情况较为接近[23]

另一方面,Cozzolino et al.从视频的拍摄工具相机入手[23],他们提出在拍摄时,所有的相机处理输入的光场时都会进行压缩、插值、伽马校正等一系列操作,从而留下伪影噪音。而每个相机的处理都存在差异,他们称之为相机指纹;因此可以利用图片上的噪音差异分析,也即噪声打印(noiseprint)。分析噪声时,他们采用来自相同相机或不同相机的成对图片训练一个孪生CNN(Siamese-CNN)网络,并取得了良好效果。然而最新研究显示,模拟相机指纹可以欺骗此种检测方法,因此方法的鲁棒性还需进一步完善。

3.4基于生物特征的检测

Ciftci et al.通过检测心率差异来判别视频真伪[24]。他们结合已有的生物医学文献[25],提出用一种rPPG的技术,从视频中动作和RGB颜色的细微变化检测出心率。他们在实验中发现,伪造视频中的心率并不能在空间和时间上很好地保持一致。于是他们利用简单的分类器就就完成了视频检测,并且取得了出色效果。

Mittal et al.同时利用一段视频的音频和视频两个方面特征,对于音频中的语音和视频中的面部动作表露出的感情同时进行识别[26]。他们采用了一个孪生CNN网络来完成架构。训练时,他们将真实视频和原视频同时通过网络,并且获得两个向量,计算出损失函数值;我们通过最小化真实视频的损失函数值(最大化真实视频的音频与视频表露感情的相似性)和最大化伪造视频的损失函数值(最小化相似性)来达到训练目的。

Korshunov et al.利用说话时人脸嘴部的特征变化与视频中语音之间的差异检测伪造说话的视频[27]。他们首先标定嘴部的若干地标,地标之间的距离变化可以认为是嘴部的运动;他们使用。然后他们将语音通过DNN识别特征(也可以直接用数学方法提取MFCC声学特征),最后利用不同帧之间的这些特征训练LSTM,从而实现时序分析。最后通过判断嘴部动作和语音是否匹配,鉴定视频真伪。然而这种方法存在的最大局限性在于任何嘴部的活动都会被识别为讲话的动作,而且也诚如作者自己所说,标定嘴部地标时,他们只考虑了正脸的情况,因此对于侧脸,准确率会大大下降。同时,该种方法也只能识别语音内容与嘴型相差较大的视频,若视频制作针对此方面作出更新,则鲁棒性也会较差。

表2:常见深伪检测方法

作者发表状况方法数据集效果\*

Zhou et al. REF \_Ref90071501\r \h [19]2018, CVPRW 面部分类CNN+隐写分析特征SVM SwapMe and FaceSwap 0.927AUC

UADFV 0.851AUC

Celeb-DF 0.557AUC

McCloskey et al.[18]2018, arXiv 基于颜色细节的SVM GAN Full 0.61AUC

Li & Lyu REF \_Ref90069806\r \h [3]2018, arXiv 面孔扭曲检测 CNN Faceforensics++(c40)0.911AUC

UADFV 0.974AUC

Deepfake‐TIMIT(LQ)0.999AUC

Deepfake‐TIMIT(HQ)0.932AUC

Güera et al.[20]2018, AVSS CNN+LSTM 自建0.971ACC

Cozzolino et al. REF \_Ref90074103\r \h [23]2019, TIFS 相机指纹Siamese-CNN FaceForensics++0.9214ACC

Rossler et al.[15]2019, ICCV Xception FaceForensics++(raw)0.9926ACC

FaceForensics++(c23)0.9573ACC

FaceForensics++(c40)0.8100ACC

Wang et al. REF \_Ref90076756\r \h [28]2019, IJCAI 监测神经元活动的浅层神经网络 Celeb-DF 0.668AUC

Masi et al.[21]2020, ECCV 双分支DenseNet+LSTM Faceforensics++(c23)0.991AUC

Li et al.[22]2020, CVPR 混合图像边界检测 HRNet FaceForensics++(DeepFake)0.9912AUC

FaceForensics++(FaceSwap)0.9909AUC

FaceForensics++(Face2Face)0.9931AUC

FaceForensics++(NeuralTexture)0.9927AUC

Ciftci et al.[24]2020, TPAMI 心率特征 CNN FaceForensics++(DeepFake)0.9375ACC

FaceForensics++(FaceSwap)0.9625ACC

FaceForensics++(Face2Face)0.9525ACC

FaceForensics++(NeuralTexture)0.8125ACC

Dang et al. REF \_Ref90072904\r \h [17]2020, CVPR 注意力机制+CNN DFFD 0.712AUC

Mittal et al. REF \_Ref90074123\r \h [26]2020, ACM MM 音/视频感情相似性Siamese-CNN DFDC 0.844AUC

DF-TIMIT 0.966AUC

Korshunov et al.[27]2019, ICML 嘴部特征与语音DNN+LSTM 自建0.167EER

\*注:AUC= Area Under Curve,ROC曲线下面积;ACC=Accuracy,准确率;EER=Equal Error Rate,等错误率.

4总结与展望

近年来,随着深度学习方法和技术的完善,越来越多的深伪视频生成,对个人和国家的隐私和安全都造成了一定程度的威胁。然而我们也看到,越来越丰富的深伪检测技术正在极力逆转这一局面。然而深伪检测仍面临诸多问题,除了泛化能力、鲁棒性和正确率有待提高外,不可解释的机器学习方法更类似于黑盒模型,在法庭作证等方面效力欠缺,不能令人信服[29]。但我们也看到,伴随着技术手段的不断发展,深伪技术也将越来越趋近于成熟,更好地服务于我们的生活。

参考文献 （Reference）

[1]Du M, Pentyala S, Li Y, et al. Towards generalizable deepfake detection with locality-aware autoencoder[C]//Proceedings of the 29th ACM International Conference on Information & Knowledge Management(CIKM). 2020: 325-334.

[2]Tolosana R, Vera-Rodriguez R, Fierrez J, et al. Deepfakes and beyond: A survey of face manipulation and fake detection[J]. Information Fusion, 2020, 64: 131-148.

[3]Li Y, Lyu S. Exposing deepfake videos by detecting face warping artifacts[J]. arXiv preprint arXiv:1811.00656, 2018.

[4]Rössler A, Cozzolino D, Verdoliva L, et al. Faceforensics: A large-scale video dataset for forgery detection in human faces[J]. arXiv preprint arXiv:1803.09179, 2018.

[5]Rossler A, Cozzolino D, Verdoliva L, et al. Faceforensics++: Learning to detect manipulated facial images[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision(ICCV). 2019: 1-11.

[6]Dufour N, Gully A, Karlsson P, et al. Deepfakes Detection Dataset[J]. 2019.

[7]Dang H, Liu F, Stehouwer J, et al. On the detection of digital face manipulation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern recognition(CVPR). 2020: 5781-5790.

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

[9]Dolhansky B, Howes R, Pflaum B, et al. The deepfake detection challenge (dfdc) preview dataset[J]. arXiv preprint arXiv:1910.08854, 2019.

[10]Dolhansky B, Bitton J, Pflaum B, et al. The deepfake detection challenge (dfdc) dataset[J]. arXiv preprint arXiv:2006.07397, 2020.

[11]Li Y, Yang X, Sun P, et al. Celeb-df: A large-scale challenging dataset for deepfake forensics[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR). 2020: 3207-3216.

[12]He Y, Gan B, Chen S, et al. ForgeryNet: A Versatile Benchmark for Comprehensive Forgery Analysis[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR). 2021: 4360-4369.

[13]Yu P, Xia Z, Fei J, et al. A Survey on Deepfake Video Detection[J]. IET Biometrics, 2021.

[14]Chollet F. Xception: Deep learning with depthwise separable convolutions[C]//Proceedings of the IEEE conference on computer vision and pattern recognition(CVPR). 2017: 1251-1258.

[15]Rossler A, Cozzolino D, Verdoliva L, et al. Faceforensics++: Learning to detect manipulated facial images[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision(ICCV). 2019: 1-11.

[16]Tan M, Le Q. Efficientnet: Rethinking model scaling for convolutional neural networks[C]//International Conference on Machine Learning(ICML). PMLR, 2019: 6105-6114.

[17]Dang H, Liu F, Stehouwer J, et al. On the detection of digital face manipulation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern recognition(CVPR). 2020: 5781-5790.

[18]McCloskey S, Albright M. Detecting gan-generated imagery using color cues[J]. arXiv preprint arXiv:1812.08247, 2018.

[19]Zhou P, Han X, 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). IEEE, 2017: 1831-1839.

[20]Güera 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). IEEE, 2018: 1-6.

[21]Masi I, Killekar A, Mascarenhas R M, et al. Two-branch recurrent network for isolating deepfakes in videos[C]//European Conference on Computer Vision(ECCV). Springer, Cham, 2020: 667-684.

[22]Li L, Bao J, Zhang T, et al. Face x-ray for more general face forgery detection[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR). 2020: 5001-5010.

[23]Cozzolino D, Verdoliva L. Noiseprint: A CNN-based camera model fingerprint[J]. IEEE Transactions on Information Forensics and Security(TIFS), 2019, 15: 144-159.

[24]Ciftci U A, Demir I, Yin L. Fakecatcher: Detection of synthetic portrait videos using biological signals[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence(TPAMI), 2020.

[25]De Haan G, Jeanne V. Robust pulse rate from chrominance-based rPPG[J]. IEEE Transactions on Biomedical Engineering, 2013, 60(10): 2878-2886.

[26]Mittal T, Bhattacharya U, Chandra R, et al. Emotions Don't Lie: An Audio-Visual Deepfake Detection Method using Affective Cues[C]//Proceedings of the 28th ACM international conference on multimedia(ACM MM). 2020: 2823-2832.

[27]Korshunov P, Halstead M, Castan D, et al. Tampered speaker inconsistency detection with phonetically aware audio-visual features[C]//International Conference on Machine Learning(ICML). 2019 (CONF).

[28]Wang R, Juefei-Xu F, Ma L, et al. FakeSpotter: A simple yet robust baseline for spotting AI-synthesized fake faces[C]//29th International Joint Conference on Artificial Intelligence, IJCAI 2020. International Joint Conferences on Artificial Intelligence, 2020: 3444-3451.

[29]Nguyen T T, Nguyen C M, Nguyen D T, et al. Deep learning for deepfakes creation and detection: A survey[J]. arXiv preprint arXiv:1909.11573, 2019.