基于卷积 LSTM 的视频中 Deepfake 检测方法

李永强，白　天

（中国科学技术大学 软件学院，安徽 合肥 230026）

摘　要： 以 Deepfake 为代表的伪造人脸技术，使用少量的人脸数据就能将视频中的人脸替换成为目标人脸，从而达到伪造视频的目的。此类技术的滥用将带来恶劣的社会影响，需要使用检测技术加以制裁。针对这一问题，已有若干检测算法被提出。现有方法具有一定局限性，单帧检测算法忽略了 Deepfake 动态缺陷；当数据存在缺陷时，模型可能会陷入“学会特定脸”的陷阱中。提出了一种对视频数据中的 Deepfake 检测方法，使用结合 CNN 和 LSTM 的卷积 LSTM，判断视频真伪。提出了一种基于人脸特征点的 cutout 方法，能抑制网络学会特定脸。实验表明，在不同场景下，准确度对比基准算法均有提升。

关键词： Deepfake 检测；计算机视觉；深度学习

中图分类号： TP18　　**文献标识码：** A　　**DOI：** 10.19358/j.issn.2096-5133.2021.04.005

引用格式： 李永强，白天. 基于卷积 LSTM 的视频中 Deepfake 检测方法[J].信息技术与网络安全，2021，40(4)：28-32.

Deepfake detection method in videos based on convolutional LSTM

Li Yongqiang，Bai Tian

（School of Software Engineering，University of Science and Technology of China，Hefei 230026，China）

Abstract： The face forgery technology represented by deepfake can replace the face in video with the target face by using a small amount of face data，so as to achieve the purpose of forgery video. The abuse of this kind of technology will bring adverse social effects，which need to be punished by using detection technology. The existing methods have some limitations，single frame detection algorithm ignores the dynamic defect of deepfake；when the data has defects，the model may fall into the trap of "learning specific face". In this paper，we propose a forgery face detection method in video，which uses the convolutional LSTM combined with CNN and LSTM to judge if a video is original or manipulated by deepfake. In addition，we propose a cutout method based on landmarks，which can inhibit the network from learning specific face. Experiments show that the accuracy of the baseline algorithm is improved in different scenes.

Key words： Deepfake detection；computer vision；deep learning

0 引言

近年来，基于深度学习技术的图像生成技术迅速发展，视频人脸伪造技术也随之日趋成熟。利用此类技术的人脸伪造技术已经可以欺骗普通人类[1]。但这些技术的滥用也引发了一些社会问题，因为这些技术可以利用公众人物公开的视频、图像素材，伪造公众人物出场的虚假视频，发布虚假的言论，或伪造色情影片，破坏名誉。由于 Deepfakes 项目[2]的广泛流传，这一类技术常被通称为 Deepfake。为了避免 Deepfake 技术的滥用，许多研究团体做出了卓越的贡献。ROSSLER A 等人发布了包含大量 Deepfake 数据的公开数据集 FaceForensics++[1]，以帮助研究人员研究检测算法。Facebook 开展了 DFDC（Deepfake Detection Challenge）比赛并公布了训练数据集[3]。

早期的研究主要是从视频中随机提取帧，使用基于卷积神经网络(Convolutional Neural Networks，CNN)的二分类器进行检测[1]。这样的方法存在两个问题。一是只使用了单帧信息，忽略了 Deepfake 技术的动态缺陷，在低质量场景下容易出错。二是分类器与训练数据高度相关，不具备通用性，在对数据的生成模型未知的情况下，效果将会大打折扣。

另外有研究从频域角度出发试图解决问题。KORSHUNOV P[4]等人通过构造图片的频率特征或统计特征等方法，构造图像质量指标(Image Quality Measures，IQM)，作为特征供支持向量机(Support Vector Machine，SVM)学习，但是通过构造特征的方式需要大量专业知识，不能很好地泛化问题。QIAN Y[5]等人从频域提取到 Deepfake 模型留下的特定频率特征，在特定数据集上获得了较好的效果。然而无法保证不同的模型能产生类似的频率特征，并且不同的有损压缩方式也会带来频率噪声，对频域特征存在干扰，缺乏鲁棒性。

针对以上问题，本文提出了一种基于深度学习的视频中 Deepfake 检测方法。

本文的主要工作如下：

(1)提出卷积 LSTM 的模型架构，结合 CNN 和长短期记忆网络(Long Short-Term Memory，LSTM)的模型架构，融合了存在于帧间的时间信息，用于视频中 Deepfake 检测。

(2)提出一种帧抽取方法，提高了 Deepfake 动态缺陷的显著性。

(3)提出一种基于人脸特征点进行 cutout 的数据增强方法，抑制了模型学会特定脸的现象。

(4)在公开数据集上进行测试，并与文献中其他算法进行对比。

1 本文方法

通过对相关数据的分析可以发现，使用 Deepfake 伪造人脸的视频在动态过程中会出现异常抖动。对于帧之间独立分析的方法无法发现这种抖动，仅局限于发现单帧画面中的瑕疵。而视频质量较差或使用有损视频压缩算法也会带来许多瑕疵，当模型无法区分这两类瑕疵时，模型的性能将大大降低。本文提出卷积 LSTM 架构，将 CNN 与 LSTM 进行融合，用于解决传统模型忽略时序特征的问题，并提出一种基于人脸标记点(landmarks)的 cutout 方法，以抑制模型学会特定脸的现象。

1.1 模型架构

卷积 LSTM 分为 CNN block 和 LSTM block，CNN block 负责获取空间信息，LSTM block 则从特征图序列中获取时间信息。如图 1(a)所示，从视频中提取 N 帧后，使用现有的人脸提取器对这些帧进行人脸提取，得到人脸图片序列，调整大小到 CNN block 对应的大小，在训练时，还需进行动态数据增强。假设

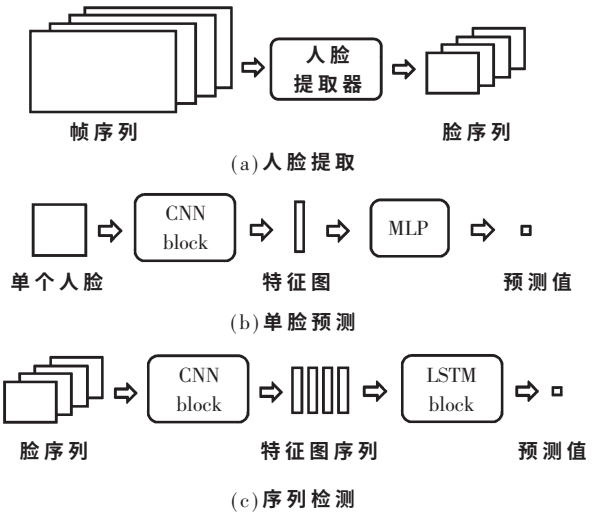


图 1 模型架构图

使用的 CNN block 输出 512 维的特征图，那么将得到 N 个 512 维向量，在训练阶段，每个向量还将经过多层感知机得到标签的独热编码，用于约束特征图。将 N 个 512 维的向量输入 LSTM block 中，模型最终按独热编码的形式输出预测标签。

1.1.1 帧抽取

由于生成模型的训练过程中没有唯一标准的答案，因此生成结果具有一定的不确定性。即对于相似的输入，模型可能生成不相似的结果，尤其是处理毛发、斑点等特征的情况下，无法保证每次随机生成的结果相同，从而导致了视频中的抖动现象。记 R_i 表示原视频的第 i 帧，F_i 表示 R_i 经过 Deepfake 处理过的图像，$d_r(i,j)$表示原始视频中第 i 和 j 帧的差异，即 R_i 和 R_j 的差异，$d_f(i,j)$表示 F_i 和 F_j 的差异。通常，$d_r(i,j)$主要受视频中人脸的姿态、光照等条件影响，而 $d_f(i,j)$还额外受到模型不确定性 $\mathrm{error}(i,j)$ 影响，如式(1)所示。若$|i-j|$过于小，d_r 和 d_f 均很小，造成信息冗余。若$|i-j|$过于大，则 d_f 主要取决于 d_r，模型的抖动将难以捕捉。因此，应当在保证一定最小间隔的前提下，选取相对紧凑的选取帧，本文实验中，采用在视频中随机选取时间点，以 0.2 s 为间隔采样，采样总长度不超过 32 帧。

$$d_f(i,j)=d_r(i,j)+\mathrm{error}(i,j)>d_r(i,j) \tag{1}$$

1.1.2 人脸提取

现有的人脸提取算法已经能满足实际需要。常见的人脸提取器有 MTCNN[6]、dlib 等。本文后续实验中，将使用 MTCNN 提取人脸框及人脸特征点。

1.1.3 CNN-block

ImageNet 竞赛极大推动了深度卷积网络的发展，即使 ImageNet 早已结束，图像领域的新模型都会在 ImageNet 上进行基准测试，并发布预训练模型。基于这样的预训练模型在其他任务上训练，可以加快训练收敛的速度，并且一般会使得模型最终效果更好、更稳定。针对不同的应用场景，有许多开箱即用的模型可以使用。Resnet[7]系列因为其普遍性，具有良好的可移植性，几乎所有平台都能使用。Mobilenet[8]系列针对边缘节点算力较弱的场景，在可接受的准确率损失的前提下，极大地减少了计算力。Efficientnet[9]则相反，使用更大的模型，更大的输入尺寸，获得更好的拟合效果。本文提出的架构设计中，可以根据实际场景轻松地切换 CNN-block。为了后续模块中能保留充足的信息，CNN-block 将会向后输出一个较大维度的向量，如 512 维向量，同时为了保证这个向量中包含了有关于训练目标的信息，在训练过程中，对于每个向量，都会经过一个浅层神经网络，输出单帧的标签预测值，从对特征向量本身进行约束，以提高特征本身对于标签的相关性。

1.1.4 LSTM-block

单独使用 CNN-block 无法处理变长数据和有序数据，因此需要结合使用循环神经网络(Recurrent Neural Network，RNN)来处理时间信息。LSTM 是一种特殊的 RNN，如图 2 所示，本文所用的结构在原始 LSTM 的基础上加入了实例正则化(Instance Normalization，IN)，这是由于不同的数据可能使用的替换人脸不同，每个图像实例之间独立地进行正则化，可以加快模型收敛。结合 CNN 与 LSTM 后，模型拥有融合时空信息的能力，能同时挖掘数据中 Deepfake 的动态缺陷和静态缺陷，提升了数据信息利用率。

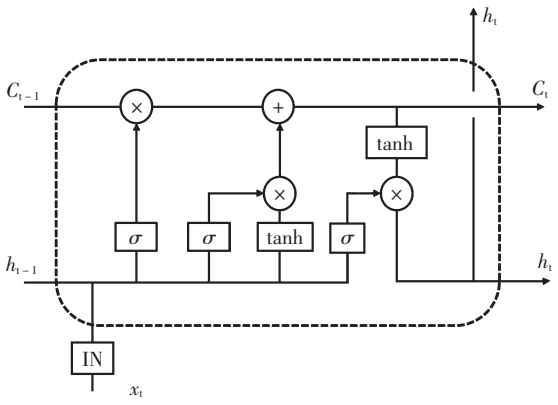


图 2 LSTM 单元结构图

1.2 数据处理

1.2.1 数据增强

常规的数据增强方法依然适用，但是需要注意一点的是，部分增强方法在同一组数据中需要保持一致。除了 1.2.3 节中将要介绍的 cutout 方法，本文实验中使用到的数据增强方法如表 1 所示。

表 1 数据增强方法及参数说明

方法名称	方法说明	参数范围
水平翻转	以概率 p 随机进行左右翻转	$p=0.5$
旋转	在角度范围内进行旋转	-30°~30°
缩放	对图像进行比例缩放	100%~120%
JPEG 压缩攻击	使用不同质量因子进行 JPEG 压缩	0.5~1.0

1.2.2 基于人脸特征点的 cutout

cutout[10]是一种数据增强技术，在图像中随机选择一个正方形区域，进行全 0 填充。Deepfake 伪造的痕迹主要存在于面部及交界处，直接使用 cutout 技术，有可能将面部覆盖，从而引入有害噪声，影响模型的学习。人脸标记点用于定位人脸不同的区域，dlib 能检测 68 个特征点，如图 3(a)所示。本文实验中，利用这些点划分了 6 个区域，如图 3(b)所示，区域所用的特征点序号如表 2 所示。每个区域按均等概率进行 cutout。

表 2 不同区域所用特征点序号

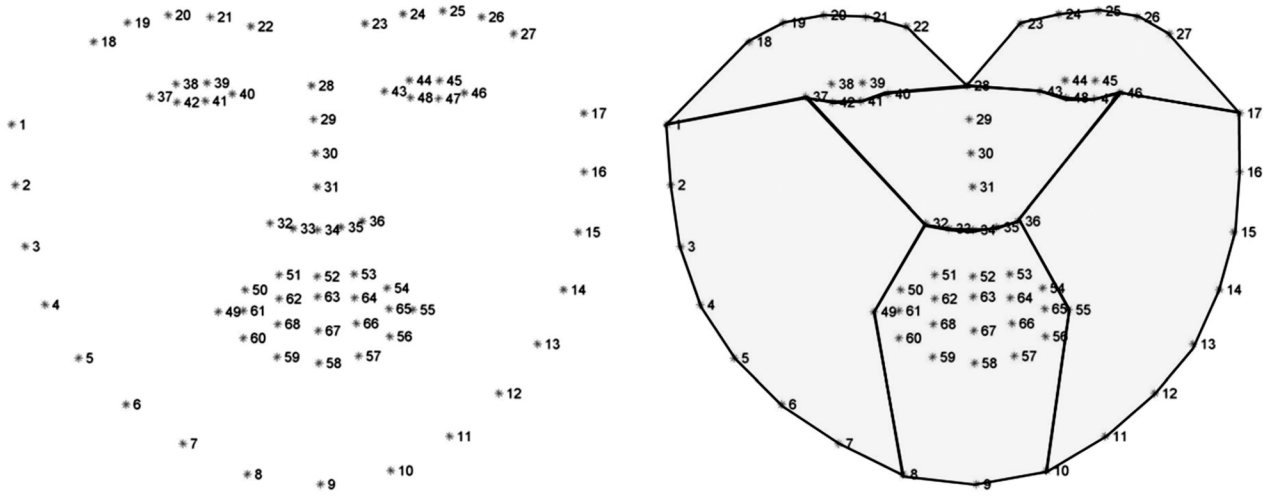
区域名称	特征点序号
左脸颊	1,2,3,4,5,6,7,8,49,32,37
右脸颊	10,11,12,13,14,15,16,17,46,36,55
嘴及下巴	8,9,10,55,36,35,34,33,32,49
左眼	1,18,19,20,21,22,28,40,41,42,37
右眼	28,23,24,25,26,27,17,46,47,48,43
鼻	32,33,34,35,36,46,47,48,43,28,40,41,42,37

2 实验结果及分析

本文实验使用了 2 个数据集，FaceForensics++和 DFDC。实验中，将利用 DFDC 数据训练 CNN 模型，使用迁移学习技术，在完整模型架构中，使用 FaceForensics++对应数据子集进行模型微调，最后在 FaceForensics++对应的子集上进行效果验证。

2.1 数据集介绍

FaceForensics++数据集包含了1 000 个从 YouTube 上筛选的视频片段，片段以单人视频为主，视频物



(a)68 个人脸特征点　　(b)6 个 cutout 区域

图 3　人脸特征点及 cutout 区域

理分辨率从 480p 到 1 080p 不等。通过 Deepfakes[2]、Face2Face[11]、FaceSwap[12]以及 NeuralTextures[13]四种算法生成伪造视频以及对应的模型。对于每个视频，根据 H264 编码时使用的参数分为无损 (RAW)、低压缩(C23)和高压缩(C40)三种版本，对于伪造视频，还提供了伪造区域的 mask 信息。

DFDC(Deepfake－Detection－Challenge)数据集来源于 Kaggle 上的算法竞赛，由 AWS、Facebook 等共同创建，其中的伪造视频由视频、音频以及音视频同时伪造。数据集共 471.84 GB，分为 50 个相互独立的分卷，每个分卷中有若干视频和一个标签文件，每个文件对应的标签中标明了数据是否是造假视频，对于伪造视频，还提供了其原视频的标签，相比 FaceForensics++，没有公布进行伪造的算法和模型，也没有提供伪造区域的 mask 信息。

2.2 训练细节

为了方便对比基线，实验的 CNN 模型选用了 Xception[14]，单帧输入大为 224×224，对应输出 feature map 的大小为 512。使用带有梯度裁切的 adam 优化器进行优化，学习率设置为 0.000 1，损失函数使用 focal loss[15]，其表达式为式(2)，α_t 设置为 0.25，γ 设置为 2，focal loss 可以缓解数据中的不平衡，使模型更专注于难样本。训练时存在两种约束，需要交替训练。训练 CNN 的阶段，设置batch size 为 128，迭代次数为 10 个 epoch，使用 DFDC 数据作为预训练。综合训练阶段，使用FaceForensics++的某个子集进行训练，设置 batch size 为 32，迭代次数为 20 个 epoch。

$$\text{focal\_loss}(p_t)=-\alpha_t\times(1-p_t)^{\gamma}\times\log(p_t) \tag{2}$$

2.3 实验结果及对比

本文方法使用 DFDC 数据预训练 CNN 模块，然后迁移到 FaceForensics++对应的数据子集上进行后续的完整训练。数据子集包括无损、低压缩和高压缩三种质量下的四种算法生成的伪造视频和对应的真实视频进行混合的 12 种集合。结果如表 3 所示。相比于文献[1]中给出的基于 Xception 的基线算法，在 12 种集合上的效果均有提升，尤其是在低视频质量的情况下，提升较为明显。

表 3　在 FaceForensics++数据集上准确率对比(%)

数据子集	本文方法	Xception[1]
Raw		
Deepfakes	99.67	99.59
FaceSwap	99.71	99.61
Face2Face	99.32	99.14
NeuralTextures	99.41	99.36
Compressed 23		
Deepfakes	98.96	98.85
FaceSwap	98.55	98.36
Face2Face	98.46	98.23
NeuralTextures	95.17	94.5
Compressed 40		
Deepfakes	95.53	94.28
FaceSwap	93.05	91.56
Face2Face	95.16	93.7
NeuralTextures	92.56	82.11

3 结论

本文提出了一种用于检测视频的Deepfake检测方法。提出将CNN和LSTM结合的卷积LSTM，充分利用了视频中帧的空间信息和Deepfake的动态缺陷这一时间信息。针对任务目标，提出了一种帧提取方法，提高了Deepfake动态缺陷的显著性。提出一种基于人脸特征点的cutout方法用于数据增强，同时抑制模型学会特定脸的现象。在FaceForensics++数据集上的实验表明，算法在各种压缩质量和换脸算法下，对比基线算法均有提升。

参考文献

[1] ROSSLER A, COZZOLINO D, VERDOLIVA L, et al. Faceforensics++: learning to detect manipulated facial images[C]. Proceedings of the IEEE International Conference on Computer Vision, 2019.

[2] Deepfakes. Deepfakes github[EB/OL]. [2021-01-12]. https://github.com/deepfakes/faceswap.

[3] Facebook. Deepfake-detection-challenge[EB/OL]. [2021-01-12]. https://www.kaggle.com/c/deepfake-detection-challenge.

[4] KORSHUNOV P, MARCEL S. Deepfakes: a new threat to face recognition? assessment and detection[J]. arXiv preprint arXiv: 1812.08685, 2018.

[5] QIAN Y, YIN G, SHENG L, et al. Thinking in frequency: face forgery detection by mining frequency-aware clues[C]. European Conference on Computer Vision. Springer, Cham, 2020.

[6] ZHANG K, ZHANG Z, LI Z, et al. Joint face detection and alignment using multitask cascaded convolutional networks[J]. IEEE Signal Processing Letters, 2016, 23(10): 1499-1503.

[7] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[C]. Proceedings of the IEEE conference on computer vision and pattern recognition, 2016: 770-778.

[8] HOWARD A G, ZHU M, CHEN B, et al. Mobilenets: efficient convolutional neural networks for mobile vision applications[J]. arXiv preprint arXiv: 1704.04861, 2017.

[9] TAN M, LE Q V. Efficientnet: rethinking model scaling for convolutional neural networks[J]. arXiv preprint arXiv: 1905.11946, 2019.

[10] DEVRIES T, TAYLOR G W. Improved regularization of convolutional neural networks with cutout[J]. arXiv preprint arXiv: 1708.04552, 2017.

[11] THIES J, ZOLLHOFER M, STAMMINGER M, et al. Face2face: realtime face capture and reenactment of rgb videos[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016.

[12] MAREKKOWALSKI. FaceSwap github[EB/OL]. [2021-01-12]. https://github.com/MarekKowalski/FaceSwap.

[13] THIES J, ZOLLHÖFER M, NIE?NER M. Deferred neural rendering: image synthesis using neural textures[J]. ACM Transactions on Graphics (TOG), 2019, 38(4): 1-12.

[14] CHOLLET F. Xception: deep learning with depthwise separable convolutions[C]. Proceedings of the IEEE conference on computer vision and pattern recognition, 2017.

[15] LIN T Y, GOYAL P, GIRSHICK R, et al. Focal loss for dense object detection[C]. Proceedings of the IEEE International Conference on Computer Vision, 2017.

(收稿日期：2021-01-20)

作者简介：

李永强(1995-)，男，硕士研究生，主要研究方向：计算机视觉。

白天(1975-)，男，博士，讲师，主要研究方向：图像处理与分析、计算机视觉。