使用卷积视觉转换器的 Deepfake 视频检测

德雷萨·沃达霍
吉马大学
deressa.wodajo@ju.edu.et

所罗门·阿特纳夫
亚的斯亚贝巴大学
solomon.atnafu@aau.edu.et

抽象的

深度学习模型的快速发展
可以生成和合成超逼真的视频,称为
Deepfake 及其易于访问的问题引起了人们的关注
可能的恶意使用。深度学习技术
现在可以生成面孔,在两个主体之间交换面孔
在视频中,改变面部表情,改变性别,以及
改变面部特征,仅举几例。这些强大的视频
操纵方法在许多领域都有潜在的用途。
然而,如果它们对所有人构成迫在眉睫的威胁,如果
用于身份盗窃、网络钓鱼等有害目的,
和骗局。在这项工作中,我们提出了一种用于检测 Deepfakes 的卷积视觉转换
器。卷积视觉转换器有两个组件:卷积神经网络 (CNN) 和视觉转换器

（维生素）。 CNN 提取可学习的特征,而 ViT
将学习到的特征作为输入并对其进行分类
使用注意力机制。我们在
DeepFake 检测挑战数据集 (DFDC) 并拥有
达到 91.5% 的准确率,AUC 值为 0.91,并且
损失值为 0.32。我们的贡献是我们添加了
一个 CNN 模块到 ViT 架构,并实现了
DFDC 数据集上的竞争结果。

一、简介

改变图像、视频和音频的技术是
发展迅速 [12, 62]。创建和操作数字内容的技术和技术专长也很容易获得。目前,
可以用很少的资源无缝生成超逼真的数字图像[28]

以及在线提供的简单操作说明 [30, 9]。
Deepfake 是一种旨在替换人脸的技术
通过视频中其他人的脸来定位目标 [1]。它
通过将合成的面部区域拼接到原始图像中来创建[62]。该术语还可以表示代表

创建的炒作逼真视频的最终输出。 Deepfakes 可以
用于创建超逼真的计算机生成
图像 (CGI)、虚拟现实 (VR) [7]、增强现实 (AR)、教育、动画、艺术和电影 [13]。

然而,由于 Deepfakes 具有欺骗性,它们可以
也可用于恶意目的。

自 Deepfake 现象以来,各种作者
提出了不同的机制来区分真实视频
从假的。正如 [10] 所指出的,尽管每种提出的机制都有其优势,但当前的检测方
法缺乏普遍性。作者指出,当前的现有模型专注于 Deepfake 创建工具来解决

通过研究他们假定的行为。比如,悦尊
等。 [33] 和 TackHyun 等人。 [25] 使用的不一致
眨眼以检测 Deepfakes。然而,使用工作
康斯坦丁诺斯等人的。 [58]和海等人。 [46] 现在是
可以模仿眨眼。 [58] 中的作者提出了一个系统,该系统可以生成说话头像的视
频
自然的面部表情,例如眨眼。作者
在[46]中提出了一个可以生成面部表情的模型
从肖像。他们的系统可以合成一张静止图片
表达情绪,包括眨眼的幻觉
议案。

我们的工作基于 [10, 11] 指出的 Deepfake 检测方法的两个弱点:数据预
处理和通用性。 Polychronis 等人。 [11] 指出,当前的 Deepfake 检测系统
主要集中在呈现

他们提出的架构,并不太重视数据
预处理及其对最终检测模型的影响。
作者强调了数据预处理对 Deepfake 检测的重要性。约书亚等人。 [10] 专注

关于面部伪造检测的一般性,发现
大多数提议的系统缺乏通用性。作者将普遍性定义为可靠地检测多种欺骗技术
和可靠地欺骗看不见的检测技术。

乌穆尔等人。 [13] 使用生物信号（内部

图像生成器和合成器的表示）。他们
使用了一个简单的卷积神经网络（CNN)分类器,只有三层。作者使用了 3000
个视频
用于训练和测试。但是,他们没有详细说明他们如何预处理数据。从 [31, 52, 21],
它
很明显,非常深的 CNN 具有卓越的性能
比图像分类任务中的浅层 CNN。因此,
仍有空间进行另一种通用的 Deepfake 检测

Tor 具有广泛的数据预处理管道,并且
在一个非常深的神经网络模型上进行训练,以捕捉到
尽可能多的 Deepfake 工件。
因此,我们提出了一种通用的卷积视觉转换器（CViT)架构来检测 Deepfake
使用卷积神经网络和变压器架构的视频。我们称我们的方法为广义的

三个主要原因。 1)我们提出的模型可以学习局部
使用CNN和Transformer架构的全局图像特征,使用注意力机制

变压器[6]。 2)我们在训练和分类过程中对我们的数据预处理给予同等重视。 3) 我们建议
在一组不同的人脸图像上训练我们的模型
目前可用于检测在不同设置、环境和方向中创建的 Deepfakes 的最大数据集。

二、相关工作

随着 CNN [4、20]、生成对抗网络 (GAN) [18] 及其变体 [22] 的快速发展,现在可以创建超逼真的图像 [32]、视频 [61] 和音频信号 [53, 15] 是

更难检测和区分真正的未篡改
视听。创造看似真实的声音的能力,
图像和视频已引起各种有关利益相关者的引导,以阻止此类事态发展

被对手用于恶意目的 [12]。为此,研究界目前迫切需要

带有 Deepfake 检测机制。

2.1。 Deepfake 视频的深度学习技术一代

Deepfake 由深度生成模型（例如 GAN 和自动编码器 (AE) [18, 37])生成和合成。
Deepfake 是通过在两个身份之间交换来创建的
图像或视频中的主题 [56]。 Deepfake 可以
也可以通过使用不同的技术来创建,例如面部
swap [43]、puppet-master [53]、lip-sync [49、47]、面部重演 [14]、合成图像或视频生成,以及
语音合成[48]。监督 [45, 24, 51] 和无监督的图像到图像翻译 [19] 和视频到视频翻译 [59, 35] 可用于创建高度逼真的 Deepfake。

第一个 Deepfake 技术是 FakeAPP [42],它
使用了两个AE网络。 AE 是一种前馈神经网络 (FFNN),其编码器-解码器架构是
训练以重建其输入数据[60]。 FakeApp 的编码器
提取潜在的人脸特征,其解码器重建
人脸图像。两个 AE 网络共享相同的编码器以在源和目标人脸之间交换,并使用不同的解码器进行训练。

大多数 Deepfake 的创建机制都集中在
面部交换和逐像素编辑的面部区域
常用[28]。在换脸时,一张脸
源图像在目标图像的面上交换。在
puppet-master,创建视频的人控制
视频中的人。在口型同步中,源人控制
目标视频中的鼠标移动,以及在面部重演中,面部特征被操纵[56]。
Deep fake 创建机制通常使用源图像和目标图像的特征图表示。一些

特征图表示是面部动作编码系统（FACS）、图像分割、面部标志、

和面部边界[37]。 FACS是人类的分类
面部表情,定义了 32 个名为动作单位 (AU) 的原子面部肌肉动作和 14 个动作描述符
(AD) 用于杂项操作。面部地标是
面部的一组定义位置,例如眼睛、鼻子和
嘴的位置 [36]。

2.1.1 人脸合成

图像合成处理从
样本训练示例[23]。人脸图像合成技术用于人脸老化、人脸正面化和姿势

引导一代。 GAN 主要用于人脸合成。 GAN 是一种生成模型,旨在创建

来自样本的数据生成模型 [3, 18]。 GAN
包含两个对抗网络,生成模型G和判别模型D ,
.

产生类似真实的样本[22]。生成器的目标是
捕获数据分布。鉴别器的目标
是确定一个样本是来自模型分布还是数据分布[18]。人脸正面化 GAN

更改图像中的面部方向。姿势引导面
图像生成将输入图像的姿态映射到其他图像。 GAN 架构,例如
StyleGAN [26] 和
FSGAN [43],合成高度逼真的图像。

2.1.2 换脸

人脸交换或身份交换是一种基于 GAN 的方法,
创建逼真的 Deepfake 视频。换脸过程
将源图像的人脸插入到目标图像中
该主题从未出现过[56]。它是最受欢迎的
用于在各种电影剪辑中插入著名演员 [2]。
换脸可以使用 GAN 和传统的
CV 技术,例如 FaceSwap（用于交换 ping 人脸的应用程序)和 ZAO（中国移动应用程序）
将任何人的脸交换到任何视频剪辑上)[56]。人脸交换 GAN (FSGAN) [43] 和区域分离 GAN
(RSGAN) [39] 用于人脸交换、人脸重演、属性编辑和人脸部分合成。深度伪造
FaceSwap 使用两个带有共享编码器的 AE,重建源和目标人脸的训练图像 [56]。
这些过程涉及一个可以裁剪和对齐的面部检测器
使用面部地标信息的面部[38]。一个训练有素的

源人脸的编码器和解码器交换特征
源图像到目标人脸。然后使用泊松将自动编码器输出与图像的其余部分混合

编辑[38]。

面部表情（面部重演)交换改变一个人
面部表情或转换面部表情
人。表达重演将身份转变为
傀儡 [37]。使用表情交换,一个人可以转移
一个人对另一个人的表达[27]。多年来,人们提出了各种面部重演。 Cycle
GAN 由 Jun-Yan 等人提出。 [63] 用于两个视频源之间的面部重演,无需任
何一对训练示例。 Face2Face 操纵面部表情

源图像并投影到另一个目标面上
实时[54]。 Face2Face 创建密集重建
在源图像和使用的目标图像之间
用于在不同光照条件下合成人脸图像[38]。

2.2. Deepfake 视频的深度学习技术检测

Deepfake检测方法分为三类
[34, 37]。第一类方法侧重于物理
或视频的心理行为,例如跟踪
眨眼或头部姿势运动。第二类
专注于 GAN 指纹和生物信号
图像,例如可以在图像中检测到的血流。第三类侧重于视觉伪影。方法

专注于视觉工件是数据驱动的,并且需要
大量数据用于训练。我们提出的模型落在
属于第三类。在本节中,我们将讨论为检测视觉而设计和开发的各种架构

Deepfakes 的人工制品。

大流士等人。 [1] 提出了一种称为 MesoNet 的 CNN 模型
网络自动检测超现实伪造
使用 Deepfake [40] 和 Face2Face [54] 创建的视频。
作者使用了两种网络架构（Meso-4 和
MesoInception-4) 专注于介观性质
的图像。 Yuezun 和 Siwei [34] 提出了一种利用图像变换的 CNN 架构（即,

在创建 Deepfakes 期间产生的缩放、旋转和剪切)不一致。他们的方法针对

仿射面部翘曲中的伪影作为显着特征
区分真假图像。他们的方法比较
Deepfake 人脸区域与相邻像素的区域,以发现在人脸过程中出现的分辨率不一致
翘曲。

休伊等人。 [41] 提出了一种新颖的深度学习方法
检测伪造的图像和视频。作者专注于
重播攻击,面部交换,面部重演和完全
计算机生成的图像欺骗。 Daniel Mas Montser 大鼠等。 [38]
提出了一种从视频中存在的面部提取视觉和时间特征的系统。他们的方法
结合 CNN 和 RNN 架构来检测 Deepfake
视频。

Md. Sohel Rana 和 Andrew H. Sung [50] 提出了一个
DeepfakeStack,一种集成方法（不同的堆栈
DL 模型)用于 Deepfake 检测。集成由 XceptionNet、InceptionV3、
InceptionResNetV2、
MobileNet、ResNet101、DenseNet121 和 DenseNet169
开源深度学习模型。 Junyaup Kim 等人。 [29] 提出
一个分类器,将目标个体与一组
类似的人使用 ShallowNet、VGG-16 和 Xception
预训练的 DL 模型。他们系统的主要目标
是评估三个DL的分类性能
楷模。

3. 卷积视觉转换器

在本节中,我们将介绍我们检测 Deep fake 视频的方法。 Deepfake 视
频检测模型包括
由两个组件组成:预处理组件和
检测组件。预处理组件包括人脸提取和数据增强。这

检测组件由训练组件组成,
验证组件和测试组件。这
训练和验证组件包含一个卷积
视觉转换器 (CViT)。 CViT 有一个特征学习组件,可以学习输入图像的特征
和
一种 ViT 架构,用于确定特定视频是否
是假的还是真的。测试组件应用 CViT
在输入图像上学习模型以检测 Deepfakes。我们的
提出的模型如图 1 所示。

3.1。预处理

预处理组件的功能是准备
用于训练、验证和测试我们的原始数据集
CViT 模型。预处理组件有两个子组件:人脸提取和数据增强

零件。人脸提取组件负责
用于从224 x 224 RGB视频中提取人脸图像
格式。图 2 和图 3 显示了提取面的示例。

3.2.检测

Deepfake 检测过程由三个子组件组成:训练、验证和测试

成分。训练部分是主要部分
的建议模型。这是学习发生的地方。深度学习
模型需要大量时间来设计和微调
将特定的问题域拟合到其模型中。在我们的案例中,
最重要的考虑是寻找一个最佳的 CViT
学习 Deepfake 视频特征的模型。为了这,
我们需要寻找合适的参数
训练我们的数据集。验证组件类似
到培训部分。验证组件是一个微调我们模型的过程。它用于

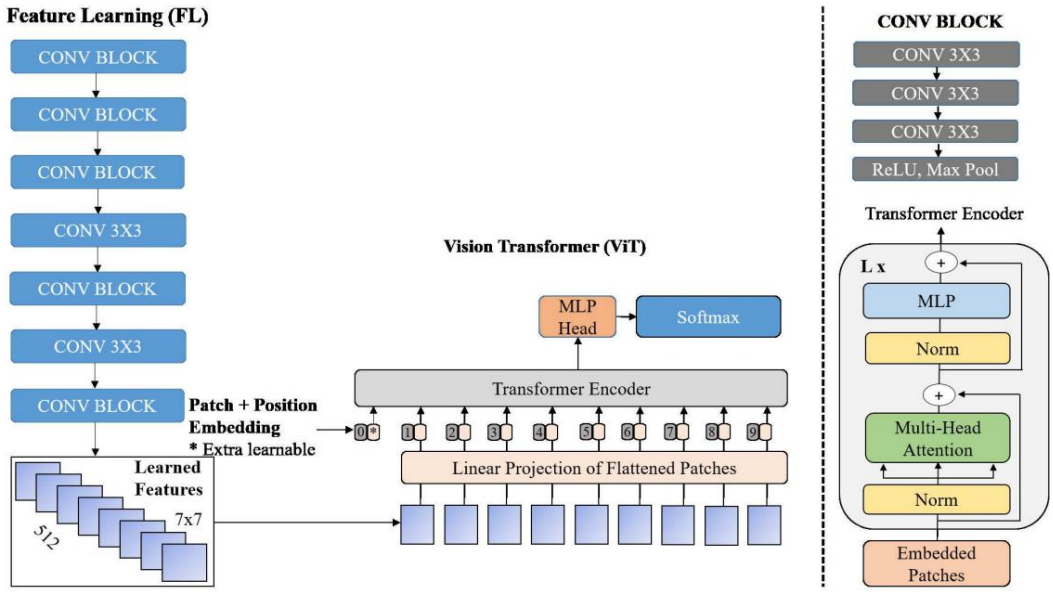

图 1. 卷积视觉转换器。

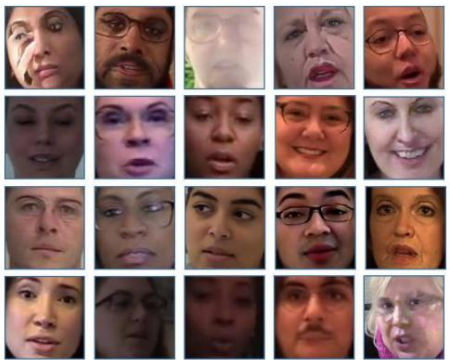

图 2. 提取的假人脸图像样本。

图 3. 提取的真实人脸图像样本。

评估我们的 CViT 模型并帮助 CViT 模型更新其内部状态。它可以帮助我们跟踪 CViT 模型的训练进度及其 Deepfake 检测精度。这测试组件是我们分类和确定在特定视频中提取的人脸类别。因此,这子组件解决了我们的研究目标。

提出的 CViT 模型由两个部分组成:特征学习 (FL) 和 ViT。 FL 从人脸图像中提取可学习的特征。 ViT 接受 FL作为输入并将它们转换为图像像素序列最后的检测过程。

特征学习 (FL) 组件是一堆卷积操作。 FL 组件遵循 VGG 架构 [52] 的结构。 FL 分量不同与 VGG 模型相比,它没有 VGG 架构中的全连接层,其目的是不是为了分类而是为了提取人脸图像特征ViT 组件。因此,FL 组件是一个 CNN没有全连接层。

FL 组件有 17 个卷积层,其中一个3 x 3 的内核。卷积层提取低人脸图像的水平特征。所有卷积层步幅和填充为 1。批量归一化以标准化输出特征和 ReLU 激活函数非线性适用于所有层。批次归一化函数归一化前层分布的变化[41],作为之间的变化这些层将影响 CNN 架构的学习过程。 2 x 2 像素窗口的五个最大池化也使用步幅等于 2。最大池化操作将图像尺寸减少了一半。每次之后max-pooling 操作,卷积层的宽度(channel) 增加了 2 倍,第一层

有 32 个通道和最后一层 512。
FL分量有三个连续的卷积
每一层的操作,除了最后两层,
它有四个卷积操作。我们称那些
为简单起见,三个卷积层作为 CONV Block。
每个卷积计算之后都是批量归一化和 ReLU 非线性。 FL 组件

有 1080 万个可学习参数。 FL 接受了一个
大小为224 x 224 x 3 的图像,然后在
每个卷积操作。 FL内部状态可以是
表示为(C, HW) 张量,其中C是通道,
H是高度, W是宽度。的最终输出
FL 是512 x 7 x 7空间相关的低级特征
输入图像,然后将其馈送到 ViT 架构。
我们的 Vision Transformer (ViT) 组件与
[16] 中描述的 ViT 架构。视觉转换器
(ViT) 是基于 [57] 工作的变压器模型。
转换器及其变体（例如,GPT-3 [44])主要用于 NLP 任务。 ViT 将
Transformer 的应用从 NLP 问题领域扩展到

CV 问题域。 ViT 使用与原始变压器模型相同的组件,但对输入信号稍作修改。 FL 组件和 ViT

组件构成了我们的卷积视觉转换器
(CViT) 模型。我们将模型命名为 CViT,因为模型
是基于堆栈的卷积运算和
ViT 架构。
ViT 组件的输入是
人脸图像。特征图被分成七个补丁
然后嵌入到1 x 1024线性序列中。
然后将嵌入的补丁添加到位置嵌入中以保留图像的位置信息

特征图。位置嵌入的尺寸为2 x 1024 。

ViT 组件接受位置嵌入和
补丁嵌入并将它们传递给 Transformer。
ViT Transformer 只使用一个编码器,与原始的 Transformer 不同。 ViT 编码器由 MSA 和
MLP 块。 MLP 块是一个 FFN。 Norm 规范化了变压器的内部层。变压器

有 8 个注意力头。 MLP 头部有两个线性层
和 ReLU 非线性。 MLP 头任务相当于典型 CNN 架构的全连接层。第一层有
2048个通道,最后一层

有两个通道代表 Fake 或 Real 类
人脸图像。 CViT 模型共有 20 个加权
层和 3860 万个可学习参数。 Softmax 是
应用于 MLP 头部输出以将权重值压缩在 0 和 1 之间,以用于最终检测目的。

4. 实验

在本节中,我们介绍了工具和实验
我们用来设计和开发原型的设置
修改模型。我们将展示从
模型的实现并给出解释
的实验结果。

4.1。数据集

DL 模型从数据中学习。因此,仔细的数据集
准备对于他们的学习质量和预测准确性至关重要。 BlazeFace
神经人脸检测器 [5],MTCNN
[55] 和人脸识别 [17] DL 库用于提取人脸。 BlazeFace 和人脸识别都是
–
快速处理大量图像。三个DL
库一起使用以增加人脸检测的准确性。人脸图像以JPEG文件格式存储

224 x 224图像分辨率。 90% 压缩
比例也适用。我们在火车上准备了我们的数据集,
验证集和测试集。我们使用了 162,174 张分类图像
112,378 用于训练,24,898 用于验证和 24,898
分别以70:15:15的比例进行测试。每一个真实的和
假类在所有集合中具有相同数量的图像。
我们使用 Albumentations 进行数据增强。 Albu mentations
是一个 Python 数据增强库,具有
一大类图像变换。百分之九十
人脸图像被增强,使我们的总数据集
为 308,130 张面部图像。

4.2.评估

CViT 模型使用二元交叉熵损失函数进行训练。使用[0.485 , 0.456 , 0.406] 的平均值和[0.229 , 0.224 , 0.225]的标准偏差对 32 个图像的小批量进行归一化。标准化的

然后在将人脸图像输入到
每次训练迭代的 CViT 模型。亚当优化器
学习率为0.1e-3 ,权重衰减为0.1e-6
用于优化。该模型总共训练了
50 个时代。学习率降低了 0.1 倍
每个步长为 15。
分类过程需要 30 张面部图像和
将其传递给我们训练有素的模型。为了确定我们模型的分类精度,我们使用了对数损失函数。一个
等式 1 中描述的 log loss 将网络分类为
从 0 到 1 的概率分布,其中0 > y < 0.5
表示真实类, 0.5 ⩾ y < 1表示
假课。我们选择了对数损失分类指标,因为
它高度惩罚随机猜测和自信的错误预测。

$$\text{对数损失} = -\frac{1}{n}\sum_{\text{我}=1}^{n}[yi\log(\hat{y}i) + \log(1 - yi)\log(1 - \hat{y}i)] \quad (1)$$

我们用来衡量模型容量的另一个指标
是 ROC 和 AUC 指标 [8]。 ROC用于可视化分类器以选择分类阈值。曲线下面积

是 ROC 曲线所覆盖的区域。 AUC 测量
分类器的准确性。

我们使用准确性、AUC 分数和
损失值。我们在 400 个看不见的 DFDC 上测试了模型
视频并达到 91.5% 的准确率,AUC 值为
0.91,损失值为 0.32。损失值表示如何
我们模型的预测与实际目标值相差甚远。
对于 Deepfake 检测,我们使用了来自每个人脸的 30 张人脸图像
视频。我们使用的帧数会影响
Deepfake 检测的机会。但是,准确性可能
并不总是检测 Deepfakes 的正确措施,因为我们
可能会遇到来自假视频的所有真实面部图像
（假视频可能包含真实帧）。

我们将我们的结果与其他 Deepfake 检测进行了比较
模型,如表 1、2 和 3 所示。从表 1 中,
2和3,我们可以看到我们的模型在
DFDC、UADFV 和 FaceForensics++ 数据集。然而,我们的模型在
FaceForensics++ 上表现不佳
FaceShifter 数据集。这样做的原因是因为视觉
工件很难学习,我们提出的模型可能
没学好那些神器。

数据集	准确性
FaceForensics++ 换脸 69% FaceForensics++ DeepFakeDetection 91% FaceForensics++ Deepfake 93% FaceForensics++ FaceShifter 46% FaceForensics++ 神经纹理 60%	

表 1. FaceForensics++ 上的 CViT 模型预测准确率
数据集

方法	验证测试	
CNN 和 RNN-GRU [38] [47] 92.61% CViT		91.88%
	87.25	91.5

表 2. 我们的模型和其他 Deepfake 检测的准确性
DFDC 数据集上的模型

方法	验证 FaceSwap	Face2Face	
MesoNet 84.3%		96%	92%
MesoInception 82.4%		98%	93.33%
CViT 93.75		69%	69.39%

表 3. 我们的模型和其他 Deepfake 检测模型在 UADFV 数据集上的 AUC 性能。 *
面部取证++

4.3.分类过程中数据处理的影响

影响我们模型准确性的一个主要潜在问题是人脸检测中的固有问题

DL 库（MTCNN、BlazeFace 和人脸识别）。 –

图 4、图 5 和图 6 显示了被 DL 库错误分类的图像。这些数字总结了我们的

对选定的 200 个视频进行初步数据预处理测试
随机从 10 个文件夹中。我们选择了我们的测试集视频
我们可以在 DFDC 数据集中找到的所有设置:室内、室外、暗室、明亮的房间、主体
定位、主体站立、对侧说话、在前面说话、主体移动

说话时,性别,肤色,一人视频,二人
人物视频、靠近相机的被摄体和被摄体
远离相机。对于初步测试,我们提取了视频的每一帧,发现了 637 非人脸

地区。

图 4. 人脸识别非人脸区域检测。

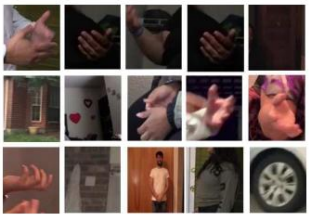

图 5. BlazeFace 非人脸区域检测。

图 6. MTCNN 非人脸区域检测。

我们测试了我们的模型以检查其准确性如何受到影响,而没有任何尝试删除
这些图像,并且
我们的模型准确率下降到 69.5%,损失
值增加到 0.4。

为了最小化非人脸区域并防止错误预测,我们使用了三个 DL 库并选择了最
好的
为我们的模型执行库,如表 4 所示。作为
解决方案,我们使用人脸识别作为人脸的“过滤器”
BlazeFace 检测到的图像。我们选择了人脸识别 –
因为,在我们的调查中,它拒绝了更多的假阳性

比其他两个模型。我们使用人脸识别 –
最终的 Deepfake 检测。

数据集	Blazeface f rec	– ** MTCNN
DFDC 83.40% FaceSwap 56%		91.50% 90.25%
FaceShifter 40% Neural Textures	57%	69% 63%
DeepFakeDetection 82%		46% 44%
Deepfake 87% Face2Face 54%	LeNet	60% 60%
74.50%		91% 79.59
		93% 81.63%
		61% 69.39%
		93.75% 88.16%

表 4. 深度学习库对 Deepfake 检测精度的比较。
** 人脸识别

5. 结论

Deepfakes 为数字媒体、VR、
机器人、教育等诸多领域。在另一个范围内,它们是可能造成破坏和不信任的技术

对公众。有鉴于此,我们设计并
使用 CNN 和 Transformer 开发了一个用于 Deepfake 视频检测的通用模型,我们将其命名为 Convolutional Vison Transformer。出于三个原因,我们将我们的模型称为广义模型。 1)我们的第一个原因出现

来自CNNs和Transformer的组合学习能力。 CNN 擅长学习局部特征,而

Transformers 可以从局部和全局特征映射中学习。
这种组合能力使我们的模型能够关联图像的每个像素并理解非局部特征之间的关系。 2)我们同样重视我们的

训练和分类期间的数据预处理。 3) 我们
使用最大和最多样化的数据集进行 Deepfake 检测。

CViT 模型在不同的集合上进行了训练
从 DFDC 数据集中提取的面部图像。
该模型在 400 个 DFDC 视频上进行了测试,并具有
达到91.5%的准确率。尽管如此,我们的模型有一个
有很大的改进空间。将来,我们打算通过添加其他已发布的数据集来扩展我们当前的工作
让 Deepfake 研究更多样化、更准确,
和健壮。

参考

[1] Darius Afchar、文森特·诺齐克、山岸纯一和 Isao
越前。 MesoNet:一个紧凑的面部视频伪造检测网络。第 1-7 页,2018 年。

[2] Shruti Agarwal, Hany Farid, Yuming Gu, Mingming He,
Koki Nagano 和 Hao Li。保护世界领导人免受
深假货。在 2019 年 CVPR 研讨会上。

[3] 查鲁 C. 阿加瓦尔。神经网络和深度学习:
一本教科书。施普林格国际出版社,瑞士,
2020 年。

[4] Md Zahangir Alom、Tarek M. Taha、Chris Yakopcic、Ste fan
Westberg、Paheding Sidike、Mst Shamima Nasrin、Mah mudul
Hasan、Brian C. Van Essen、Abdul AS Awwal 和
Vijayan K. Asari。深度学习理论和架构的最新调查。电子,8（3）:292,2019。

[5] 瓦伦丁·巴扎列夫斯基、尤里·卡廷尼克、安德烈·瓦库诺夫、
Karthik Raveendran 和 Matthias Grundmann。火焰脸:
移动 GPU 上的亚毫秒级神经人脸检测。
arXiv 预印本 arXiv:1907.05047v2, 2019。

[6] 欧文·贝洛、巴雷特·佐夫、阿什什·瓦斯瓦尼、乔纳森·施伦斯、
和 Quoc V. Le。注意力增强卷积网络。在 2019 年 IEEE/CVF 国际会议上

计算机视觉 (ICCV),第 3285-3294 页,2019 年。

[7] Avishek Joey Bose 和 Parham Aarabi。虚拟赝品:虚拟现实的深度赝品。
2019 年 IEEE 第 21 届国际会议
多媒体信号处理 (MMSP) 研讨会,页数
1-1。 IEEE,2019。

[8] 安德鲁·P·布拉德利。 ROC曲线下面积的使用
在机器学习算法的评估中。图案
承认,30（7）:1145-1159,1997。

[9] 约翰·布兰登。 2018 可怕的高科技色情片:令人毛骨悚然
年，“deepfake”视频呈上升趋势。在 https:// 可用的
www.foxnews.com/tech/terrifying-high-tech-pornreepy-
deepfake-videos-are-on-the-rise。

[10] Joshua Brockschmidt, Jiacheng Shang, and Jie Wu. On the
面部伪造检测的一般性。 2019年IEEE第16届
移动特设和传感器系统研讨会国际会议 (MASSW),第 43-47 页。 IEEE,2019。

[11] Polychronis Charitidis, Giorgos Kordopatis-Zilos, Symeon
帕帕多普洛斯和伊奥尼斯·康帕夏里斯。调查
预处理和预测聚合对 DeepFake 检测任务的影响。 arXiv 预印本

arXiv:2006.07084v1,2020。

[12] 鲍比切斯尼和丹妮尔香橼。 Deep Fakes 对隐私、民主和国家安全的迫在眉睫的
挑战,2019 年。可在 https://ssrn.com/abstract=3213954 获取。

[13] Umur Aybars Ciftci、Ilke Demir 和 Lijun Yin。 Fake Catcher:使用生物信号
检测合成肖像视频。 arXiv 预印本 arXiv:1901.02212v2, 2019。

[14] Sourabh Dhere、Suresh B. Rathod、Sanket Aarankalle、Yash
小伙子和梅格甘地。人脸再现述评
技巧。 2020年国际工业大会
4.0 技术 (I4Tech),第 191-194 页,印度浦那,2020 年。
IEEE。

[15] 克里斯·多纳休、朱利安·J·麦考利和米勒·S·帕克特。
对抗性音频合成。在第七届国际学习代表大会上,ICLR 2019,新奥尔良,洛杉矶,
美国,2019 年 5 月 6-9 日,纽约,纽约,美国,2019。

OpenReview.net。

[16] 阿列克谢·多索维茨基、卢卡斯·拜尔、亚历山大·科列斯尼科夫、
Dirk Weissenborn、翟晓华、Thomas Unterthiner
Mostafa Dehghani、Matthias Minderer、Georg Heigold、Sylvain Gelly、
Jakob Uszkoreit 和 Neil Houlsby。图像是
值得 16x16 字:用于图像识别的变形金刚
规模。 arXiv 预印本 arXiv:2010.11929v1, 2020。

[17] 亚当·盖吉。世界上最简单的 Python 和命令行面部识别 api。可在

https://github.com/ageitgey/face_recognition。

[18] Ian J. Goodfellow、Jean Pouget-Abadie、Mehdi Mirza、Bing 徐大卫·沃德-法利、舍吉尔·奥扎里、亚伦·库维尔 和约书亚·本吉奥。生成对抗网络。在第 27 届国际神经信息处理系统会议论文集 - 第 2 卷,第 2672-2680 页,

美国马萨诸塞州剑桥市,2014 年。麻省理工学院出版社。

[19] Arushi Handa、Prerna Garg 和 Vijay Khare。蒙面 使用卷积神经网络的神经风格迁移。 2018年电气、电子通信工程最新创新国际会议

(ICRIEECE),第 2099-2104 页,2018 年。

[20] Rahul Haridas 和 Jyothi R L. 卷积神经网络:综合调查。国际期刊

应用工程研究 (IJAER), 14(03):780–789, 2019 年。

[21] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 用于图像识别的深度残差学习。 2016年 IEEE 计算机视觉和模式识别会议 (CVPR),第 770-778 页。 IEEE,2016 年。

[22] Yongjun Hong、Uiwon Hwang、Jaeyoon Yoo 和 Sungroh 尹。如何生成对抗网络及其 变体工作:概述。第 52 卷,纽约,纽约, 美国,2019 年。计算机协会。

[23] 何煌、Phillip S. Yu、王长虎。使用生成对抗网络进行图像合成的介绍。

arXiv 预印本 arXiv:1803.04469v1, 2018。

[24] Xun Huang, Ming-Yu Liu, Serge Belongie, and Ming-Yu 刘。多模态无监督图像到图像转换。 在计算机视觉 – ECCV 2018,第 179–196 页,Cham, 2018. 施普林格国际出版社。

[25] TackHyun Jung、SangWon Kim 和 KeeCheon Kim。深度视觉:使用人眼眨眼检测 Deepfake 图案。 IEEE 访问,8:83144–83154,2020。

[26] 特罗·卡拉斯、萨穆利·莱恩和蒂莫·艾拉。一种基于样式的生成对抗生成器架构

网络。 arXiv 预印本 arXiv:1812.04948, 2018。

[27] 哈萨姆·哈立德和西蒙·吴。 OC-FakeDect:使用一类变分自动编码器对 Deepfake 进行分类。在 2020 年 IEEE/CVF 计算机视觉和模式识别研讨会 (CVPRW) 会议,第 2794–2803 页, 2020 年。

[28] Ali Khodabakhsh、Raghavendra Ramachandra、Kiran Raja、和Pankaj Wasnik。假人脸检测方法:他们能 泛化?在 2018 年生物指标特别兴趣小组 (BIOSIG) 国际会议上,第 1-6 页。 IEEE, 2018 年。

[29] Junyaup Kim、Siho Han 和 Simon S. Woo。分类 伪装人脸图像中的真实人脸图像。 2019年 IEEE国际大数据会议（大数据）, 第 6248-6250 页,2019 年。

[30] 帕维尔·科尔舒诺夫和塞巴斯蒂安·马塞尔。 DeepFakes:一个新的 威胁到人脸识别?评估和检测。 arXiv 预印本 arXiv:1812.08685, 2018。

[31] Alex Krizhevsky、Ilya Sutskever 和 Geoffrey E. Hinton。 使用深度卷积神经网络进行 ImageNet 分类 网络。交流。 ACM, 60(6):84–90, 2017。

[32] 克里斯蒂安·莱迪格、卢卡斯·泰斯、费伦茨·胡萨尔、何塞·卡巴列罗、 安德鲁·坎宁安、亚历杭德罗·阿科斯塔、安德鲁·艾特肯、 Alykhan Tejani、Johannes Totz、Zehan Wang 和文哲史。逼真的单图像超分辨率 使用生成对抗网络。 arXiv 预印本 arXiv:1609.04802v5,2017。

[33] Yuezun Li, Ming-Ching Chang, and Siwei Lyu. In Ictu Oculi:通过检测眨眼来暴露 AI 生成的假人脸视频。 arXiv 预印本 arXiv:1806.02877v2, 2018。

[34] Yuezun Li and Siwei Lyu. Exposing DeepFake Videos 通过检测面部扭曲伪影。 arXiv 预印本 arXiv:1811.00656v3,2019。

[35] Arun Mallya、Ting-Chun Wang、Karan Sapra 和 Ming-Yu 刘。世界一致的视频到视频合成。在计算机视觉 – ECCV 2020,第 359-378 页, Cham,2020 年。 施普林格国际出版社。

[36] Brais Martinez、Michel F. Valstar、Bihan Jiang 和 Maja 惊慌失措。面部动作的自动分析:一项调查。 IEEE 情感计算汇刊,10（3）:325–347, 2019 年。

[37] Yisroel Mirsky 和 Wenke Lee。创建与检测 Deepfakes:一项调查。 ACM 计算。生存,54（1）,2021。

[38] Daniel Mas Montserrat, Hanxiang Hao, SK Yarlagadda, Sriram Baireddy、Ruiting Shao、Janos Horvath、Emily Bar tusiak、Justin Yang、David Guera、Fengqing Zhu 和 Ed ward J. Delp。 Deepfake 自动人脸检测 加权。在 2020 年 IEEE/CVF 计算机视觉和模式识别研讨会 (CVPRW) 会议上, 页数 2851–2859,2020 年。

[39] Ryota Natsume、Tatsuya Yatagawa 和 Shigeo Morishima。 RSGAN:使用面部和头发进行面部交换和编辑 潜在空间中的表示。在 ACM SIGGRAPH 2018 海报,SIGGRAPH 18,纽约,纽约,美国,2018 年。计算机协会。

[40] Huy H. Nguyen、Ngoc-Dung T. Tieu、Hoang-Quoc Nguyen Son、Vincent Nozick、Junichi Yamagishi 和 Isao Echizen。 用于判别的模块化卷积神经网络 在计算机生成的图像和摄影图像之间。在第十三届国际会议论文集上

可用性、可靠性和安全性,纽约,纽约,美国, 2018. 计算机协会。

[41] Huy H. Nguyen、Junichi Yamagishi 和 Isao Echizen。 胶囊取证:使用胶囊网络进行检测 伪造的图像和视频。在 ICASSP 2019 - 2019 IEEE 声学、语音和信号国际会议 处理 (ICASSP),第 2307-2311 页,2019 年。

[42] Thanh Thi Nguyen, Cuong M. Nguyen, Dung Tien Nguyen, Duc Thanh Nguyen 和 Saeid Nahavandi。用于 Deepfake 创建和检测的深度学习。 arXiv 预印本 arXiv:1909.11573v1,2019。

[43] 尤瓦尔·尼尔金、约西·凯勒和塔尔·哈斯纳。 FSGAN:主题不可知的人脸交换和重演。 2019年 IEEE/CVF 计算机视觉国际会议, ICCV 2019,韩国首尔（南）,10 月 27 日 - 11 月 2,2019 年,第 7183-7192 页。 IEEE,2019。

[44] 开放人工智能。 OpenAI API,2020 年。https:// 可在 openai.com/blog/openai-api。

[45] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan 朱。具有空间自适应的语义图像合成

正常化。 2019年IEEE/CVF计算机会议
视觉和模式识别 (CVPR),第 2332-2341 页。
IEEE,2019。

[46] Hai X. Pham、Yuting Wang 和 Vladimir Pavlovic。生成对抗性谈话头:将肖像带入生活
具有弱监督神经网络。 arXiv 预印本
arXiv:1803.07716,2018。

[47] KR Prajwal、Rudrabha Mukhopadhyay、Vinay P. Nambood iri 和 CV Jawahar。唇形同步专家就是您所需要的
在野外对 Lip Generation 的演讲,第 484-492 页。计算机协会,纽约,纽约,美国,
2020 年。

[48] 迈克·普莱斯和马特·普莱斯。 2019 年使用 Deepfakes 玩防御和防御。
可用 lhttps://www.blackhat.com/us
19/简报/时间表/使用 deepfakes-14661 玩进攻和防守。

[49] Prajwal KR、Rudrabha Mukhopadhyay、Jerin Philip、Ab hishek Jha、Vinay Namboodiri 和 CV Jawahar。迈向自动面对面翻译。在第 27 届 ACM
国际多媒体会议（MM 19），页
1428–1436,纽约,纽约,美国,2019 年。
计算机。

[50] Md. Shohel Rana 和 Andrew H. Sung。 DeepfakeStack:一个
基于深度集成的 Deepfake 检测学习技术。 2020年第七届IEEE国际网络会议
安全与云计算 (CSCloud)/2020 第六届 IEEE
边缘计算和可扩展国际会议
云 (EdgeCom),第 70-75 页,2020 年。

[51] 齐藤国明、凯特·萨恩科和刘明宇。 COCO FUNIT:Few-Shot 无监督图像翻译
内容条件样式编码器。在计算机视觉中
ECCV 2020,第 382–398 页,Cham,2020 年。施普林格国际出版社。

[52] 凯伦西蒙扬和安德鲁齐瑟曼。用于大规模图像识别的非常深的卷积网络。在

Yoshua Bengio 和 Yann LeCun,编辑,3rd International
学习表征会议,ICLR 2015,圣
美国加利福尼亚州迭戈,2015 年 5 月 7 日至 9 日,会议跟踪论文集,
2015 年。

[53] Supasorn Suwajanakorn、Steven M. Seitz 和 Ira
凯梅尔马赫-施利泽曼。综合奥巴马:学习
来自音频的唇形同步。 ACM 翻译。图,36（4）:780–789,
2017 年。

[54] Justus Thies、Michael Zollhofer、Marc Stamminger、Chris tian Theobalt 和 Matthias Nießner。 Face2Face:实时
RGB 视频的面部捕捉和重演。交流。
ACM, 62(1):96–104, 2018。

[55] 时代勒。预训练的 Pytorch 人脸检测 (MTCNN)
和识别（InceptionResnet)模型。可在
https://github.com/timesler/facenet-pytorch。

[56] 鲁本·托洛萨纳、鲁本·维拉-罗德里格斯、朱利安·费雷斯、
艾萨米·莫拉莱斯和哈维尔·奥尔特加-加西亚。 DeepFakes 和
超越:人脸处理和虚假检测的调查。
信息。融合,64:131-148,2020。

[57] Ashish Vaswani、Noam Shazeer、Nick Parmar、Jacob Uszko Reit、Llion Jones、Aidan N. Gomez、Lukasz Kaiser 和 Illia Polosukhin。注意力就是你所需要的。在诉讼中
第31届神经信息国际会议
处理系统,NIPS 17,第 6000-6010 页。 Curran Associates Inc.,2017 年。

[58] Konstantinos Vougioukas、Stavros Petridis 和 Maja Pantic。具有
GAN 的逼真的语音驱动面部动画。
国际计算机视觉杂志,128:1398–1413,
2020 年。

[59] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Guilin Liu,
Andrew Tao、Jan Kautz 和 Bryan Catanzaro。视频到视频合成。在第
32 届国际神经信息处理系统会议论文集上,NIPS 18,第 1152-1164 页,
美国纽约州雷德胡克,

2018. Curran Associates Inc.

[60] M. Arif Wani、Farooq Ahmad Bhat、Saduf Afzal 和
阿西夫·伊克巴尔·汗。深度学习进展,第 57 卷
大数据研究。施普林格自然,新加坡,2020。

[61] 叶戈尔·扎哈罗夫、阿利亚克桑德拉·希舍亚、叶戈尔·布尔科夫和
维克多·伦皮茨基。 Few-Shot 对抗性学习
逼真的神经说话头模型。 arXiv 预印本
arXiv:1905.08233v2,2019。

[62] 郑丽蕾、张颖和 Vrizlynn LL Thing。一项调查
关于图像篡改及其在真实照片中的检测。
爱思唯尔,58:380–399,2018。

[63] Jun-Yan Zhu、Taesung Park、Phillip Isola 和 Alexei A.
埃弗罗斯。使用循环一致对抗网络的未配对图像到图像转换。在 2017 年
IEEE 计算机视觉国际会议 (ICCV),第 2242 页–

2251 年,2017 年。