使用 InceptionResNetV2 和 LSTM 进行深度假货检测

Priti Yadav[1]、Ishani Jaswal[2]、Jaiprakash Maravi[3]、Vibhash Choudhary[4] 和 Gargi Khanna[5]

[1,2,3,4] 印度哈米尔普尔国立技术学院学生

[5] 印度哈米尔普尔国立技术学院副教授

摘要

"眼见为实 "已不再是真理，它对我们生活的许多方面都产生了巨大影响。随着技术的不断进步，深度伪造变得越来越容易。事实上，有些甚至可以通过应用程序信手拈来。要识破深度伪造并不容易。人眼已很难识别深度伪造。但与此同时，一些研究人员正在努力寻找识别深度伪造的方法。深度伪造是利用人工智能算法合成的媒体。人工智能算法可以学习目标图像和源图像的属性。然后将目标图像叠加到源图像上。我们的目标是使用 LSTM 和 InceptionResNetV2 等深度学习神经网络检测视频深度伪造。我们成功地利用迁移学习建立了深度伪造检测模型，其中预训练的 InceptionResNetV2 CNN 用于提取特征和形成向量。LSTM 层使用特征进行训练，由此产生的混淆矩阵为我们提供了验证和测试精度。在 20 次和 40 次历时中，相应模型的准确率分别达到了 84.75% 和 91.48%。

关键词

Deepfake、InceptionResNetV2、深度学习、神经网络、LSTM、生成式对抗网络。

1. 导言

在过去的几年里，数字技术已经发展到我们可以在网上大幅改变任何事物的外观的地步。人们可以把缺陷想象成 Photoshop 视频，但其实还有很多其他的东西。这些缺陷实际上是指利用人工智能教会机器如何反应、读取和模仿人的面部表情和声音。这是通过向机器提供人物的真实照片、视频和声音样本来实现的。系统从提供的数据中学习，然后就能生成一个完全虚构的人物视频。这有两种方法：第一种方法是使用另一个演员代替第一人称创建深度伪造视频，机器学习如何编码或处理两个视频中的数据，找到两者之间的相似之处，然后压缩这些数据，并通过解码过程交换两个视频中的信息。因此，当你看到和听到一个人的声音时，你接收到的实际上是演员的信息。另一种方法是使用生成式对抗网络（GAN）[1]。

新兴技术国际会议：面向科学技术应用的人工智能、物联网和 CPS 国际会议，2021 年 9 月 6-7 日，印度昌

迪加尔 NITTTR

pritinith@gmail.com (P. Yadav); ishanijaswal8@gmail.com (I. Jaswal); Jai516843@gmail.com (J. Maravi); vibhash18.vc@gmail.com (V. Choudhary); gargi@nith.ac.in (G. Khanna)

在此过程中，使用深度学习算法从噪声中创建合成图像，然后将其添加到真实图像流中。反复多次处理后，就能得到不存在的人的逼真面孔。基本上，通过深度学习，我们现在有能力制作出令人信服的视频，让人们看到和做我们想要的任何事情。深度伪造分为三种类型：人脸合成、人脸互换和人脸表情处理。在人脸合成中，最流行的人脸合成方法是 StyleGAN。生成器模型经过训练，可以将高级特征与其他特征分离开来。检测这些人脸合成的一种方法是提取人脸的操纵区域。该系统会给出图像真假的二进制输出。人脸互换是另一种深度伪造，它是将目标人物的脸与真人的脸互换后得到的。它利用图像混合、人脸对齐、裁剪等技术来交换人脸，生成换脸深度伪造图像。为了检测人脸互换，主要对 CNN 和 RNN 进行训练，以识别生成过程中留下的痕迹。人脸属性和表情处理意味着对性别、脸部颜色、皮肤或头发颜色、人的年龄或使人悲伤或快乐等属性进行一些修改[2]。

2. 文献调查

近来，深度伪造的数量呈爆炸式增长。目前，有许多软件都为这些深度伪造提供了便利。它们正在成为对隐私、民主和信任的威胁。因此，对深度伪造分析的需求也在增加。我们将列出一些深度假冒检测方法。

A.Jadhav 等人，[3] 开发了一个基于网络的平台，用户通过上传视频就能轻松地对其进行真假分类。该模型使用了 ResNeXt 和 LSTM。该方法对用户友好且可靠。该方法的基础是使用 ResNeXt 从帧级特征中提取特征，并使用 LSTM 进行序列处理。该方法包括将视频划分为帧，然后对帧进行跨脸裁剪。其中一些选定的帧被组合在一起，形成视频中的人脸，从而创建了一个包含所有视频人脸裁剪的新数据集。然后，他们利用用于模型评估的混淆矩阵计算出了相当高的准确率。

Y.Li, S. Lyu, [4] 提出了一种使用传统神经网络比较生成的人脸区域及其周围区域的新方法。该方法基于观察有限资源的图像是否能通过 DF 算法生成。

U.A. Ciftci、I. Demir 和 L. Yin，[5] 以特征提取为目标，然后计算相干性和时间一致性。该方法从假视频和真视频中的面部区域提取生物信号。对 SVN 和 CNN 进行了训练，以找出真伪概率。

D.Guera 和 J. Delp，[6] 使用识别管道自动检测深度伪造。他们提出了两步分析法。在第一阶段，使用 CNN 提取帧级特征。第二阶段由 RNN 组成，它将捕捉由于人脸交换过程而引入的不规则帧。分析的数据集包含从各种在线来源收集的 600 个视频。其模型的准确率达到 94%。

Y.Li, MC.Chang 和 S. Lyu，[7] 基于使用神经网络生成的眉毛连动，提出了一种揭露深度

伪造的新系统。论文重点分析了视频中的眼动，因为眼动是一种自然信号，在合成媒体中无法很好地呈现。在该方法中，首先对视频进行预处理，定位每帧中的人脸区域，然后使用长期循环卷积网络（LRCN）找出时间上的不协调。

3. 建议的工作

生成 deepfake 的基本架构是编码器-解码器架构，其中编码器获取目标和源人脸的特征，解码器的任务是获取目标人脸的编码特征，然后生成假视频 [8]。通过高级处理，视频的质量得到了提高，残留物也被去除，但仍会留下一些肉眼无法看到的痕迹。这些残留痕迹是我们检测模型的关键特征 [9]。所提议的模型包括用于特征提取的 InceptionResnetV2。这些提取的特征用于训练一个递归神经网络，以分析视频是否经过处理。只有一小部分视频被篡改，这意味着深度伪造的时间较短，因此，视频被分割成小帧，这些帧被作为检测模型的输入[10]。

3.1. 数据集和预处理

数据集收集自 Kaggle、FaceForenscis 和 Celeb-deepfakeforensics [11]上的深度伪造检测挑战数据集。数据集包含约 6458 个视频。这些视频还包括真实视频，这些视频被有偿演员进一步篡改，然后通过不同的 deepfake 生成器方法制作成 deepfake 视频。数据集的 70% 用于训练，30% 用于测试系统。在训练期间，我们还向机器提供了系统所收到的视频文件的标签。原始视频转换为深度伪造视频的点被捕获为一帧，然后在预处理过程中进行分析。在视频预处理过程中，平均要提取 147 个帧。由于计算能力较低，我们使用了有限的帧数来训练模型。预处理后的帧将进一步分批发送，用于训练和测试。

3.2. 建模

该系统的模型对从视频中提取的每个帧进行图像分类分析。我们使用了一个名为 InceptinResNetV2 [12] 和 RNN 以及 LSTM 的预训练 CNN 模型。我们还需要定义损失函数、优化器和训练过程所需的其他超参数。根据训练模型的状态，应调整学习率以最小化损失值。

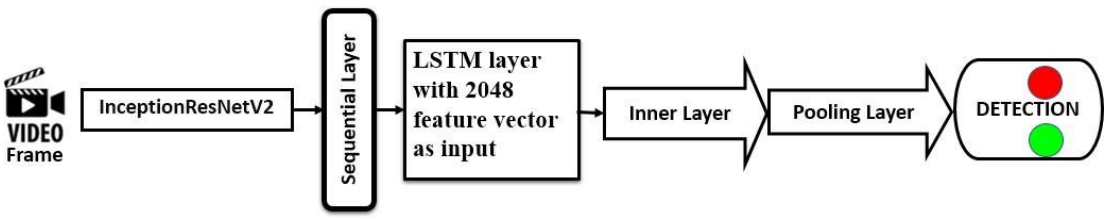

图 1：模型结构。

3.3. 用于特征可视化和分类的 InceptionResnetV2

InceptionResNetV2 是 Inception 和 ResNet 系列的结合体，共有 164 层，用于从图像中检测物体和提取特征。只添加了最后一层来分析结果。为了训练 InceptionResNetV2 CNN 模型，我们在处理视频时直接模拟了仿射人脸封装的分辨率不一致性。使用训练好的模型有助于减小体积和降低训练难度[13]。InceptionResNetV2 在预处理过程中提取每帧的特征，经过最后一次池化后考虑 2048 时维的特征向量，然后将 LSTM 作为下一个序列层。由于 CNN 不考虑暂时的不连续性，它只关注面部提取和检测，因此我们考虑用 LSTM 进行序列处理。

3.4. 用于顺序处理的 LSTM

长短期记忆（LSTM）是递归神经网络（RNN）的一种，具有前馈连接。它们是 RNN 的一个特殊版本，可以解决记忆较短的问题。LSTM 消除了 RNN 中的梯度消失问题，其设计方式是学习数据的长期依赖关系并按顺序处理数据。CNN 网络的输出是 2048 LSTM 层的输入[6]。LSTM 按顺序处理帧，然后比较不同时间帧的特征 [3]。通过比较帧的特征，就能判定视频是否深度伪造。训练完成后，任何视频都可以传给模型进行预测。

4. 实施情况

Python 是机器学习应用中最著名的语言，因此我们使用 Python 来加载数据集和进行人脸提取。数据集包含视频文件，这些文件在不同的 Cvv 文件中被标记为假视频或真视频。之后，代码会将数据集与标签文件进行匹配，并找出是否有遗漏的文件。确认唯一视频的确切数量后。从视频中提取图像并以帧的形式存储。此时，OpenCV 被用于图像识别和解释。捕获的帧被发送到模型中进行预处理。预处理后，Inception-ResNetV2 将作为迁移学习模块发挥作用。Inception-ResNetV2 会移除损失层，取而代之的是检测深度伪造损失的输出层，称为深度伪造检测损失输出层，该层已在预处理过程中定义好[12]。
网络的微调限制了数据集中已识别数据或预测性能的变体。为掌握训练数据集，该模型已编译了 20 个历元和 40 个历元。模型中使用了对神经网络非常有用的 Sigmoid 激活函数。该函数将图形中的所需数据映射为介于 0 和 1 之间的值。

5. 探测模型和结果

由于运行时间限制，对所设计的模型进行了 20 epoch 和 40 epoch 测试，结果表明准确率分别为 84.75% 和 91.48%。实施后得到的结果图表明，随着历时次数的增加，验证和测试准确率也在提高。混淆矩阵结果有助于评估系统的测试准确性。

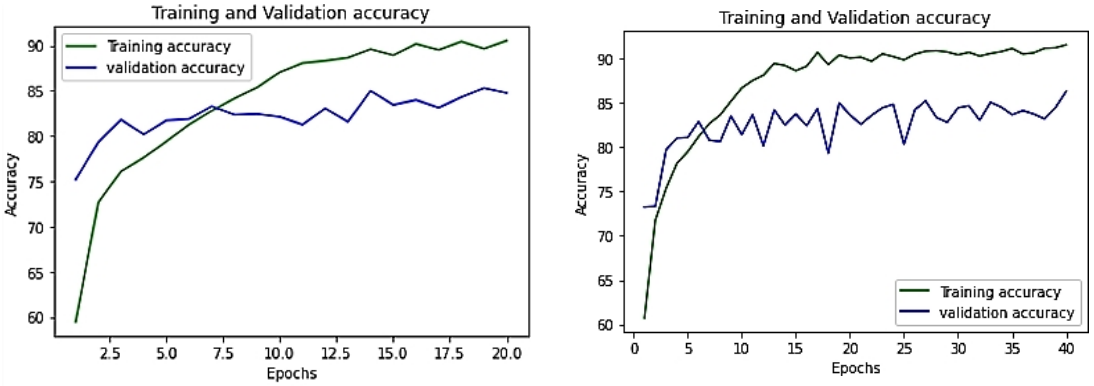

图 2：20 和 40 个纪元的训练和验证精度图

6. 结论和未来展望

由于流媒体内容似乎不再真实可信，大众的信仰开始因深度伪造而瓦解。在我们的论文中，我们提出了一种基于深度学习概念、能自动检测深度伪造的方法。在深度伪造中，目标人脸会在视频中短暂出现，因此模型会将用户视频分成若干帧，然后使用 InceptionResNetV2 和 LSTM 对这些帧进行进一步预处理。该方法具有良好的准确性和可靠性。所提出的方法能够使用卷积 LSTM 系统分析任何视频，还有助于检测被操纵的深度假脸，从而防止个人诽谤。我们还可以使用更多的历时和学习率进行实验，以获得更高的准确率。未来，我们可以通过探索更多的架构来扩展这项工作，这将有助于实施新的检测技术来检测深度假脸。

致谢

我们借此机会向我们的导师、哈米尔布尔国家理工学院欧洲经委会系副教授 Gargi Khanna 博士（夫人）表示感谢，她在整个项目过程中为我们提供了指导。这是一项团队工作，因此要特别感谢所有团队成员的合作、辛勤工作和对项目的奉献。

参考资料

[1] M.-Y. Liu, X. Huang, J. Yu, T.-C.Liu, X. Huang, J. Yu, T.-C. Wang, A. Mallya, Generative adversarial networks for image and video synthesis: A. M. -Y.Wang, A. Mallya, Generative adversarial networks for image and video synthesis：算法与应用》，《电气和电子工程师学会论文集》 109 (2021) 839-862 。doi:10.1109/JPROC.2021.3049196。

[2] Bouarara, H. Ahmed, Recurrent neural network (rnn) to analyse mental behaviour in social media, 2021. doi:10.4018/IJSSCI.2021070101.

[3] A.Jadhav, A. Patange, H. Patil, J. Patel, M. Mahajan, Deep residual learning for image recognition, International Journal for Scientific Research and Development 8 (2020) 1016- 1019.

[4] Y.Li, S. Lyu, Exposing deepfake videos by detecting face warping artifacts, 2019. arXiv:1811.00656.

[5] U.A. Ciftci, I. Demir, L. Yin, Fakecatcher：Detection of synthetic portrait videos using biological signals, IEEE Transactions on Pattern Analysis and Machine Intelligence (2020) 1-1. doi:10.1109/TPAMI.2020.3009287.

[6] D.Güera, E. J. Delp, Deepfake video detection using recurrent neural networks, in：2018 第 15 届 IEEE 基于先进视频和信号的监控国际会议 （AVSS），2018 年，第 1-6 页。doi:10.1109/AVSS.2018.8639163。

[7] Y.Li, M. -C.Chang, S. Lyu, In ictu oculi: Exposing ai created fake videos by detecting eye blinking, in： 2018 IEEE International Workshop on Information Forensics and Securit y (WIFS)，2018，pp. 1-7. doi:10.1109/WIFS.2018.8630787.

[8] I.J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville、Y.Bengio, Generative adversarial networks, 2014. arXiv:1406.2661.

[9] K.He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in：doi:10.1109/CVPR.2016.90。

[10] R.Raghavendra, K. B. Raja, S. Venkatesh, C. Busch, Transferable deep-cnn features for detecting digital and print-scanned morphed face images, in ： doi:10.1109/CVPRW.2017.228.

[11] B.Dolhansky, J. Bitton, B. Pflaum, J. Lu, R. Howes, M. Wang, C. C. Ferrer, The deepfake detection challenge (dfdc) dataset, 2020. ArXiv:2006.07397.

[12] R.K. Singh, P. V. Sarda, S. Aggarwal, D. K. Vishwakarma, Demystifying deepfakes using deep learning, in：2021 第五届计算方法与 通信国际会议（ICCMC (ICCMC)、 2021, pp. 1290–1298. doi:10.1109/ICCMC51019.2021.9418477.

[13] C.Szegedy, S. Ioffe, V. Vanhoucke, A. Alemi, Inception-V4, inception-resnet and the impactof residual connections on learning, 2016. ArXiv:1602.07261.