Contents lists available at ScienceDirect

Engineering

journal homepage: www.elsevier.com/locate/eng

News & Highlights

人工智能增强型媒体——我们还能相信新闻吗？

Ramin Skibba

Senior Technology Writer

虽然消息、信件甚至照片的内容和意图都有可能被篡改，但是人们通常不会认为它们被篡改了，因为这并不是件容易的事情。但在当今的数字世界中，情况已不再如此。随着计算机、互联网以及近年来智能手机和社交媒体的出现，用来处理照片和其他媒体的工具也迅速出现。现在，人工智能（AI）正在用更复杂的程序进一步改变数字媒体，出于各种目的，人们可以使用这些程序近乎完美地处理各种视频、照片、音频和文本。

纽约州立大学奥尔巴尼分校计算机科学教授兼计算机视觉和机器学习实验室主任Siwei Lyu说："处理照片的历史与摄影本身一样古老。最近的变化则是其与AI结合，从而扩大照片处理的操作范围。过去，处理照片需要大量的时间、精力以及特殊的培训和设备。"Lyu说，有了功能强大的计算机和足够的知识来运行算法，现在就可以在更大的范围内处理视频。

计算机工程师也在努力完善用于"自然语言处理"的AI系统，该系统可以生成与人类语言非常接近的文本和语音。例如，在2019年年初，位于旧金山的研究实验室OpenAI宣布他们已经开发出一种最先进的文本生成器，叫做GPT-2，该生成器可以根据少量提示，用英语写出连贯的句子，甚至写出短篇小说和诗歌。研究人员最初不愿发布该软件的完整模型，因为他们担心该软件因效果太好，而被恶意利用，如被用于生成"假新闻"[1]。但是在看到"没有强有力的滥用证据"[2]之后，他们在2019年11月放松了限制。但是，在这种媒体和其他媒体中，老话"眼见为实"似乎已成为了假新闻。

Photoshop等用来修改照片的软件已经存在了一段时间（图1），现在，人们也可以轻松地操纵视频了。最常见的处理方法是深度造假（deepfake），通常指的是将一个人（目标）的脸与另一个人（供体）的脸交换。深度造假的另一种类型是"口型同步"，指的通过是修改源视频，使得讲话者嘴部的动作与另一个音频保持一致。如果处理得好，输出的视频将会非常逼真，看起来讲话者说出了一些实际上他们从未说过的话。此类欺骗性视频可以并且曾经被用来操纵公众舆论，实施欺诈以及抹黑他人[3]。

在实践中，要生成深度造假视频，需要将数据（大量图片或者文本）输入到一种叫做生成对抗网络（GAN）的机器学习工具中。最简单的生成对抗网络包含两个神经网络，用来开发和改进模型将输入数据转化成新图片和新视频的能力。早期算法使用海量数据集进行训练，这些数据来自政客和名人等容易获得的图像。虽然这一过程曾经需要程序员进行某种程度上的监督，但最新的程序几乎是完全自动化的。

"不需要大量的训练数据，短短10 s的视频就足够了。"亚利桑那州立大学（位于坦佩市）计算机科学与工程学教授Subbarao Kambhampati说，他也是人类感知AI的专家。但是使用较长的视频训练模型并使用具有至少1000个高质量帧的源视频，将得到质量更好的输出视频。对于视频中的每一帧，算法都能够绘制出人头上的"标记"，以及人的头部姿势、视线，以及更

英文原文: Engineering 2020, 6(7): 723–724
引用本文: Ramin Skibba. Media Enhanced by Artificial Intelligence: Can We Believe Anything Anymore? *Engineering*, https://doi.org/10.1016/j.eng.2020.05.011

图1. 使用Adobe Photoshop软件，用16张不同的照片创建了这个奇特但逼真的风景。由AI算法驱动的软件现在提供了工具，让人们可以更轻松地创建逼真但被处理过和（或）模拟生成的视频、文本和语音。图片来源：Wikimedia Commons (CC BY-SA 3.0)。

详细的特征，包括眉毛、眨眼、眼睑、上下嘴唇、脸颊、下巴和酒窝[4]。

输出视频中，人的运动看起来像人类视觉所期望的那样流畅。但是，如果处理不当，输出视频可能会有破绽，这些内容可能会使敏锐的观看者怀疑视频被修改过。“有时候会出现奇怪的现象，例如，面部特征的拉伸或扭曲与正常面部特征不完全匹配。”弗莱彻·琼斯计算学者、美国加利福尼亚州克莱蒙特市斯克里普斯学院媒体研究专业的访问教授Doug Goodwin说。例如，如果训练数据的分辨率不足，则输出视频可能具有模糊的区域，在嘴中出现白色条纹，而不是单个的牙齿，或者面部毛发没有按照应有的方式运动。Goodwin说，使用包含各种面部表情和吐字的数据训练后，算法的效果会更好。

处理技术的进步促使了计算机科学家和工程师开发AI算法（取证软件）来检测视频和音频是否被修改[5]。“取证工具可以检测合成的媒体，并判断它是由机器还是由人生成的。但是，如果不对这些工具保密，那么总是可以制作出绕过工具的媒体。”加利福尼亚大学圣地亚哥分校计算机科学博士Paarth Neekhara说，他的研究方向包括音频和视频的深度造假。

处理和检测之间的拉锯战类似于病毒和防病毒软件的计算机安全军备竞赛，其中，补丁程序阻止了黑客，而黑客又找到了绕过补丁程序的方法[6]。专家发现了一个缺陷，使他们能够检测出被修改过的媒体，随后媒体的生成者调整算法，生成更逼真的假媒体。例如，第一代的深度造假软件会生成不定期眨眼的脸，导致造假很容易被检测出来，而下一代深度造假软件便修复了这一问题。Kambhampati说，另一个例子是，一个包含时任美国总统巴拉克·奥巴马的视频被人为修改，使其看起来像是他说了一些实际上没说过的话，但视频中他的眉毛运动与嘴唇运动不符。但在后来的深度造假视频中，奥巴马的眉毛如预期般正常地动了起来。由于可以训练AI来检测和修复此类差异，因此最新一代的深度造假软件几乎没有破绽。

出现了许多AI的负面应用[3,7]，但是也有许多正面的应用，它们推动了技术的进步。例如，改善有言语障碍的人的视频或音频记录，为电影添加更逼真的外语配音，甚至在电影中重现已故演员饰演的角色。例如，在《星球大战外传：侠盗一号》中，重现了已故演员卡丽·费雪饰演的莱娅公主[8]。结合了该项技术的虚拟现实游戏或其他娱乐活动看起来很有发展前景[9]。

正如上面提到的OpenAI，计算机科学家也在使用AI来生成可靠的文本和语音[1]。像修改视频一样，这种技术也使用了GAN来生成逼真的句子[10]。例如，谷歌翻译现在就使用了这种AI算法[11]。这些算法足够复杂精妙，可以以特定人物的风格生成文本，如生成看似出自已故作家简·奥斯汀之手的新故事[12]。程序员也在社交媒体等平台上创造了聊天机器人，该聊天机器人具有足够的阅读和真实听觉，可以像真人一样与潜在客户互动。亚马逊的Alexa和苹果的Siri可能是使用最广泛的AI通信的商业应用，它们基于云的语音服务被设定为模仿与客户的真实对话。虽然Alexa和Siri不是真人，但它们的确能够给出问题的真实答案。

Goodwin说，迄今为止，程序员在生成逼真的视频和图像方面取得了更大的进步。他说，如果当前的趋势继续发展下去，可能很快就可以构建AI算法，来创造全新且可信的语音，并自动将其与模拟音频和视频融合。这种前景及其在诈骗中的潜在用途，促使研究人员开发自动检测深度造假视频的代码，并呼吁社交媒体网站将此类媒体标识为被篡改过的媒体[13]。2020年12月，Facebook与Microsoft、亚马逊和包括Lyu在内的学术界

计算机科学家合作发起了深度造假检测挑战赛，号召研究人员提交自己的自动检测工具，并有机会赢得100万美元的奖金[14]。美国国防高级研究计划局的工程师也在研究自动检测视频或照片是否被篡改的工具[15]。

References

[1] Schwartz O. For centuries, people dreamed of a machine that could produce language. Then OpenAI made one [Internet]. New York: IEEE Spectrum; 2019 Dec 2 [cited 2020 Apr 18]. Available from: https://spectrum.ieee.org/tech-talk/ artificial-intelligence/machine-learning/for-centuries-people-dreamed-ofa- machine-that-can-produce-language-then-openai-made-one.

[2] OpenAI. GPT-2: 1.5B release [Internet]. OpenAI; 2019 Nov 5 [cited 2020 Apr 18]. Available from: https://openai.com/blog/gpt-2-1-5b-release/.

[3] Verdoliva L. Media forensics and Deepfakes: an overview. 2020. arXiv:2001.06564.

[4] Agarwal S, Farid H, Gu Y, He M, Nagano K, Li H. Protecting world leaders against deep fakes. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops; 2019 Jun 16–20; Long Beach, CA, USA; 2019. p. 38–45.

[5] Li Y, Yang X, Sun P, Qi H, Lyu S. Celeb-DF: a large-scale challenging dataset for Deepfake forensics. 2020. arXiv:1909.12962.

[6] Chesney R, Citron DK. Deep fakes: a looming challenge for privacy, democracy, and national security [Internet]. Berkeley: California Law Review; 2019 Dec 17 [cited 2020 Apr 18]. Available from: https://doi.org/10.2139/ssrn.3213954.

[7] Graham J. It's not just phishing emails, now we have to worry about fake calls, too [Internet]. Tysons Corner: USA Today; 2020 Feb 27 [cited 2020 Apr 18]. Available from: https://www.usatoday.com/story/tech/2020/02/27/phishingdeepfake- audio-scams-increasing-fake-calls/4876171002/.

[8] Winick E. How acting as Carrie Fisher's puppet made a career for Rogue One's Princess Leia [Internet]. Cambridge: MIT Technology Review; 2018 Oct 16 [cited 2020 Apr 18]. Available from: https://www.technologyreview.com/2018/10/16/139739/how-acting-as-carrie-fishers-puppet-made-a-careerfor-rogue-ones-princess-leia/.

[9] National Academies of Sciences, Engineering, and Medicine. Implications of artificial intelligence for cybersecurity: proceedings of a workshop. Washington, DC: The National Academies Press; 2019.

[10] Wang K, Wan X. Automatic generation of sentimental texts via mixture adversarial networks. Artif Intell 2019;275:540–58.

[11] Wu Y, Schuster M, Chen Z, Le QV, Norouzi M, Macherey W, et al. Google's neural machine translation system: bridging the gap between human and machine translation. 2016. arXiv:1609.08144.

[12] Poole S. The rise of robot authors: is the writing on the wall for human novelists? [Internet]. London: The Guardian; 2019 Mar 25 [cited 2020 Apr 18]. Available from: https://www.theguardian.com/books/2019/mar/25/the-riseof-robot-authors-is-the-writing-on-the-wall-for-human-novelists.

[13] Eggerton J. Hill calls for social media standards from Facebook, Reddit, others on combatting deepfakes [Internet]. Bath: Multichannel News; 2019 Oct 2 [cited 2020 Apr 24]. Available from: https://www.multichannel.com/news/ hill-calls-social-media-standards-facebook-reddit-others-combating-deepfakes.

[14] Deepfake detection challenge [Internet]. Menlo Park: Facebook; c2019 [cited 2020 Apr 24]. Available from: https://deepfakedetectionchallenge.ai/.

[15] Turek M. Media Forensics (MediFor) [Internet]. Arlington: DARPA; c2016 [cited 2020 Apr 24]. Available from: https://www.darpa.mil/program/mediaforensics.