Contents lists available at ScienceDirect

Engineering

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

News & Highlights

Facebook Deepfake 检测挑战赛

Ramin Skibba

Senior Technology Writer

人工智能（AI）的力量越来越强大，尤其在越来越逼真的视频和数字媒体上表现得最明显[1]，如新兴的AI换脸视频技术——深度伪造（Deepfake）[2]。在这一背景下，脸书（Facebook）公司赞助并发起了一项人脸视频深度伪造检测挑战赛。Deepfake检测挑战赛（DFDC）于2019年12月启动，于2020年3月报名截止[3]。比赛结果排名见参考文献[3–5]。虽然比赛结果并不惊人，但其强调了识别深度伪造视频日益增长的挑战性，为自动检测策略提供了一个基准，并为进一步研究提供了有益的指导。

在近乎没有人指导的情况下，深度伪造技术的先进计算机算法可以自动处理视频和文本，达到以假乱真的效果[1,6,7]。这种技术在创造许多积极应用的同时也带来了极大的潜在风险，如无意或恶意误导受众，传播虚假新闻和虚假信息[8]。这引发了计算机科学家和数字公民自由主义倡导者的广泛关注。

美国纽约州布法罗的纽约州立大学计算机科学教授兼媒体取证实验室主任Siwei Lyu表示："在我看来，这些工具的发展极其迅速，并正向着更高质量、更真实、更快速的趋势发展，只需要一些人脸数据，算法就能实时制作出换脸视频。"

Facebook公司与AI合作组织（Partnership on AI，位于美国加利福尼亚州旧金山，致力于AI研究和倡导的组织，成员包括谷歌公司和亚马逊公司）、微软公司以及来自美国、英国、德国和意大利的大学的科学家[3]联合组织了DFDC。大赛学术顾问Lyu说："这项挑战赛引起了研究界的广泛关注。"

大赛提供了10万多个新创建的10 s换脸视频短片（DFDC伪造人脸视频数据集）。来自学术界和行业的2114名参赛者用此数据集来测试各自的检测模型[4,9]。参赛者的任务是识别数据集中的深度伪造视频，其中包括用不同技术合成的虚假视频，以及一些不易被现有的检测模型识别的虚假视频[3,4]。随后，用包含4000多个视频短片的黑盒数据集测试算法，这些数据集包括训练数据集中未出现的一些高级加强版伪造视频短片。比赛结果排行榜和100万美元奖金获得者名单已于2020年6月公布。

最优的算法模型在已知的训练数据集中识别伪造视频的准确率达80%以上，但是，在更真实的、不可预见的黑盒测试中，最优算法模型的识别准确率只有65%，远低于在已知训练数据集中的识别准确性[4]。其他4支获胜队伍的检测模型的识别准确率与第一名差距不大。Facebook公司发言人Kristina Milian表示："比赛结果显示的较低准确率强调了一点，即建立一套能够概括归纳深度伪造技术未知领域的系统，仍然是一个亟待解决的难题。"

美国印第安纳州西拉法叶普渡大学计算机工程学教授Edward Delp说，"低端伪造"（cheapfakes）几乎用任何一种机器都可以实现，并且易于发现，而目前最高端的深度伪造需要复杂的计算机硬件（包括图形处理单元）创建。现实中，每当我们开口说话时，我们的头或嘴唇会产生微妙但独特的运动，而在伪造的视频里这些

英文原文: Engineering 2020, 6(12): 1339–1340
引用本文: Ramin Skibba. Accuracy Eludes Competitors in Facebook Deepfake Detection Challenge. *Engineering*, https://doi.org/10.1016/j.eng.2020.10.008

动作表现得并不明显。位于白俄罗斯明斯克的地图公司Mapbox的机器学习工程师Selim Seferbekov提交的代码获得了DFDC的第一名。该算法借助机器学习工具，识别出人的头部在移动时与背景不一致的像素，从而识别虚假视频。Delp认为这是一个相当复杂的方法。

Delp表示，现在的深度伪造代码会产生干扰因素，例如，调整或裁剪视频帧、略微模糊或重新压缩视频帧都可能会引入一些伪影，使检测复杂化。因此，如DFDC结果所示，检测算法的准确性取决于样本的多样性和样本的质量。

美国弗吉尼亚州阿灵顿市美国国防部高级研究计划局（DARPA）信息创新办公室项目经理Matt Turek指出，准确检测的关键在于正确识别不一致性。除了检测数码赝像，检测算法还可以检查视频的物理完整性，如光线和阴影是否正确匹配，还可以查找语义上的不一致，如视频中的天气是否与独立已知的天气匹配，还可以分析深度伪造创作和发现的社会背景，推断发布人的意图[10]。DARPA已经在其新的语义取证项目中开始了这一领域的专门研究[11]。

对所有的检测工作来说，最大的问题可能不是漏识几个伪造视频样本，而是错误地标记了原始的真实视频样本。美国纽约州纽约大学的计算机科学教授Nasir Memon说："误报是导致失败的原因。"他表示，假如大多数事件都是良性的，那么基率谬误总会让检测出现问题。例如，在人们每天上传到YouTube的数百万个视频中，可能只有少部分是篡改伪造的。在这样的前提下，即使是精确度达到99%的检测算法也会错误地标记成千上万的良性视频，从而很难快速捕捉到真正恶意篡改的视频。Memon说："对所有问题都做出回应是不可能的。"

一些数字取证专家，为了减少误报的影响，将重点放在了问题的另一面，这是DFDC竞赛未涉及的。美国纽约州约翰·杰伊刑事司法学院的计算机科学教授Shweta Jain表示："我一直在努力确定非伪造视频的来源，而不是寻找伪造视频。"

Jain利用区块链技术开发了E-Witness方法，该方法可为图像或视频文件注册唯一的"hash"或指纹，"hash"或指纹可以重新计算文件并验证其完整性[12]。Jain解释，这个过程类似在照片上使用水印，但不同的是原始"hash"标志始终存在于区块链中，不容易被篡改。"hash"可以包括相关文件的"元数据"，如制作图像或视频的设备的信息、位置数据以及数据压缩算法。Turek透露，DARPA研究人员也在研究媒体特定来源的安全方法，但仍处于早期开发阶段。

同时，创建形成虚假视频的算法的能力也越来越强大。这种算法不仅能创建越来越逼真的视频而且还能逃避检测[9]。Memon说："人们总是假想自己的技术会被对手知晓，这就类似于猫和老鼠的游戏。"在这个游戏角逐中，微软公司已经开发出深度伪造检测工具[13]，TikTok公司紧随Facebook、Twitter等其他社交媒体公司[14,15]，开始采取措施禁止深度伪造技术在其平台上使用[16]。

References

[1] Skibba R. Media enhanced by artificial intelligence: can we believe anything anymore? Engineering 2020;6(7):723–4.

[2] Adee S. What are deepfakes and how are they created? [Internet]. New York: IEEE Spectrum; 2020 Apr 29 [cited 2020 Aug 30]. Available from: https://spectrum.ieee.org/tech-talk/computing/software/what-are-deepfakes-how-arethey- created.

[3] Ferrer CC, Dolhansky B, Pflaum B, Bitton J, Pan J, Lu J. Deepfake detection challenge results: an open initiative to advance AI [Internet]. Menlo Park: Facebook AI Blog; 2020 Jun 12 [cited 2020 Sep 15]. Available from: https://ai.facebook.com/blog/deepfake-detection-challenge-results-an-open-initiativeto- advance-ai/.

[4] Dolhansky B, Bitton J, Pflaum B, Lu J, Howes R, Wang M, et al. The deepfake detection challenge dataset. 2020. arXiv:2006.07397.

[5] KnightW.Deepfakes aren't very good. Nor are the tools to detectthem [Internet]. San Francisco: Wired; 2020 Jun 12 [cited 2020 Sep 20]. Available from: https:// www.wired.com/story/deepfakes-not-very-good-nor-tools-detect/.

[6] Manjoo F. How do you know a human wrote this? [Internet]. New York: New York Times; 2020 Jul 29 [cited 2020 Sep 20]. Available from: https://www.nytimes.com/2020/07/29/opinion/gpt-3-ai-automation.html?smid=em-share.

[7] Brockman G, Murati M, Welinder P; OpenAI. OpenAI API [Internet]. San Francisco: OpenAI; 2020 Jun 11 [cited 2020 Sep 15]. Available from: https://openai.com/blog/openai-api/.

[8] Lyu S. Deepfakes and the new AI-generated fake media creation-detection arms race [Internet]. New York: Scientific American; 2020 Jul 20 [cited 2020 Sep 10]. Available from: https://www.scientificamerican.com/article/detecting-deepfakes1/.

[9] Dolhansky B, Howes R, Pflaum B, Baram N, Ferrer CC. The deepfake detection challenge (DFDC) preview dataset. 2019. arXiv:1910.08854v2.

[10] Nguyen TT, Nguyen CM, Nguyen DT, Nguyen DT, Nahavandi S. Deep learning for deepfakes creation and detection: a survey. 2019. arXiv:1909.11573.

[11] Turek M. Semantic forensics (SemaFor) [Internet]. Arlington: DARPA; c2020 [cited 2020 Sep 10]. Available from: https://www.darpa.mil/program/semantic-forensics.

[12] Samanta P, Jain S. E-Witness: preserve and prove forensic soundness of digital evidence. In: Proceedings of the 24th Annual International Conference on Mobile Computing and Networking; 2018 Oct 29–Nov 2; New Delhi, India; 2018. p. 832–4.

[13] Kelion L. Deepfake detection tool unveiled by Microsoft [Internet]. London: BBC News; 2020 Sep 1 [cited 2020 Sep 25]. Available from: https://www. bbc.com/news/technology-53984114.

[14] Kelly M. Facebook bans deepfake videos ahead of the 2020 election [Internet]. New York: The Verge; 2020 Jan 7 [cited 2020 Sep 15]. Available from: https://www.theverge.com/2020/1/7/21054504/facebook-instagram-deepfake-ban-videos-nancy-pelosi-congress.

[15] Robertson A. Twitter will ban 'deceptive' faked media that could cause 'serious harm' [Internet]. New York: The Verge; 2020 Feb 4 [cited 2020 Sep 15]. Available from: https://www.theverge.com/2020/2/4/21122661/twitter-deepfake- manipulated-media-policy-rollout-date.

[16] Statt N. TikTok is banning deepfakes to better protect against misinformation [Internet]. New York: The Verge; 2020 Aug 5 [cited 2020 Sep]. Available from: https://www.theverge.com/2020/8/5/21354829/tiktok-deepfakes-ban-misinformation- us-2020-election-interference.