

DeepFakes

Creation, Detection, and Impact

EDITED BY

Loveleen Gaur



CRC Press
Taylor & Francis Group

DeepFake

DeepFakes là một phương tiện tổng hợp tận dụng các kỹ thuật Trí tuệ nhân tạo (AI) và máy học (ML) mạnh mẽ để tạo ra nội dung âm thanh và hình ảnh giả cực kỳ chân thực, do đó khiến con người rất khó phân biệt nội dung đó với nội dung gốc. Ngoài phần giới thiệu công nghệ về khái niệm DeepFakes, cuốn sách còn trình bày chi tiết các thuật toán để phát hiện DeepFakes, kỹ thuật xác định nội dung bị thao túng và xác định hoán đổi khuôn mặt, mạng lưới thần kinh đối nghịch tổng quát, kỹ thuật điều tra phương tiện, kiến trúc học sâu, phân tích điều tra DeepFakes, v.v.

- Cung cấp phần giới thiệu kỹ thuật về DeepFakes, lợi ích của nó và tiềm năng tác hại nhỏ
- Trình bày các phương pháp tạo và phát hiện DeepFake thực tế bằng cách sử dụng Kỹ thuật học sâu (DL)
- Thu hút sự chú ý đến các vấn đề thách thức khác nhau và tác động xã hội của DeepFakes với các giải pháp hiện có của họ
- Bao gồm phân tích nghiên cứu trong lĩnh vực giả mạo DL để hỗ trợ tạo và phát hiện các ứng dụng DeepFakes
- Thảo luận về các hướng nghiên cứu trong tương lai với sự xuất hiện của DeepFakes công nghệ

Cuốn sách này dành cho sinh viên mới tốt nghiệp, nhà nghiên cứu và chuyên gia về khoa học dữ liệu, trí tuệ nhân tạo, thị giác máy tính và học máy.



Taylor & Francis
Taylor & Francis Group
<http://taylorandfrancis.com>.

DeepFake

Sáng tạo, phát hiện và tác động

Sửa bởi
Bò tót Loveleen



CRC Press
Taylor & Francis Group
Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group, an **informa** business

Ấn bản đầu tiên xuất bản năm 2023

bởi CRC Press

6000 Broken Sound Parkway NW, Suite 300, Boca Raton, FL 33487-2742

và bởi CRC Press

4 Quảng trường Công viên, Công viên Milton, Abingdon, Oxon, OX14 4RN

CRC Press là chi nhánh của Taylor & Francis Group, LLC

© 2023 tuyển chọn và biên tập, Loveleen Gaur; các chương riêng lẻ, những người đóng góp

Những nỗ lực hợp lý đã được thực hiện để xuất bản dữ liệu và thông tin đáng tin cậy, nhưng tác giả và nhà xuất bản không thể chịu trách nhiệm về tính hợp lệ của tất cả các tài liệu hoặc hậu quả của việc sử dụng chúng. Các tác giả và nhà xuất bản đã cố gắng truy tìm chủ sở hữu bản quyền của tất cả các tài liệu được sao chép trong ấn phẩm này và xin lỗi chủ sở hữu bản quyền nếu không xin phép xuất bản dưới hình thức này. Nếu bắt kỳ tài liệu bản quyền nào chưa được thừa nhận, vui lòng viết thư và cho chúng tôi biết để chúng tôi có thể khắc phục trong bắt kỳ lần tái bản nào trong tương lai.

Trừ khi được cho phép theo Luật Bản quyền Hoa Kỳ, không phần nào của cuốn sách này có thể được in lại, sao chép, truyền tải hoặc sử dụng dưới bất kỳ hình thức nào bằng bất kỳ phương tiện điện tử, cơ học hoặc phương tiện nào khác, hiện được biết đến hoặc sau này được phát minh, bao gồm sao chụp, quay vi phim và ghi âm, hoặc trong bất kỳ hệ thống lưu trữ hoặc truy xuất thông tin nào mà không có sự cho phép bằng văn bản của nhà xuất bản.

Để được phép sao chụp hoặc sử dụng tài liệu điện tử từ tác phẩm này, hãy truy cập www.copyright.com hoặc liên hệ với Copyright Clearance Center, Inc. (CCC), 222 Rosewood Drive, Danvers, MA 01923, 978-750-8400. Đối với các tác phẩm không có trên CCC, vui lòng liên hệ mpkbookspermissions@tandf.co.uk

Thông báo nhãn hiệu: Tên sản phẩm hoặc tên công ty có thể là nhãn hiệu hoặc nhãn hiệu đã đăng ký và chỉ được sử dụng để nhận dạng và giải thích mà không có ý định vi phạm.

Dữ liệu Biên mục của Thư viện Quốc hội

Một số danh mục đã được yêu cầu cho tiêu đề này.

ISBN: 978-1-032-13920-3 (hbk)

ISBN: 978-1-032-13923-4 (pbk)

ISBN: 978-1-003-23149-3 (ebk)

DOI: 10.1201/9781003231493

Sắp chữ trong Times

bởi Nhà xuất bản Newgen Vương quốc Anh

Cuốn sách này dành tặng cho gia đình, bạn bè, học sinh của tôi và tất cả những
người đã truyền cảm hứng trực tiếp và gián tiếp cho tôi.



Taylor & Francis
Taylor & Francis Group
<http://taylorandfrancis.com>

nội dung

Lời nói đầu	ix
Lời cảm ơn.....	xiii Tiêu sữ
bìa tập viên.....	xv Người
đóng góp	xvii
Chương 1 Giới thiệu về Công nghệ DeepFake	1
Bò tót Loveleen, Saurav Mallik và Noor Zaman Jhanjhi	
Chương 2 DeepFakes: Đánh giá có hệ thống và phân tích thư mục	9
Bò tót Loveleen, Jyoti Rana và Amlan Chakrabarti	
Chương 3 Kỹ thuật Deep Learning để tạo DeepFake.....	23
Bò tót Loveleen, Gursimar Kaur Arora và Noor Zaman Jhanjhi	
Chương 4 Phân tích các video DeepFake bằng Face Warping Artifacts.....	35
Ajantha Devi Vairamani	
Chương 5 Phát triển mô hình dịch ảnh để chống lại các cuộc tấn công của đối thủ	57
Bò tót Loveleen, Mohan Bhandari và Tanvi Razdan	
Chương 6 Phát hiện DeepFake bằng cách sử dụng các tính năng cục bộ và mạng thần kinh tích chập	73
Shreya Rastogi, Amit Kumar Mishra và Loveleen Gaur	
Chương 7 DeepFakes: Các trường hợp tích cực	91
Bò tót Loveleen và Gursimar Kaur Arora	
Chương 8 Các mối đe dọa và thách thức của công nghệ DeepFake.....	99
Mamta Sareen	

Chương 9 DeepFakes, Truyền thông và Tác động Xã hội	115
Shubha Mishra, Piyush Kumar Shukla và Ratish Agrawal	
Chương 10 Phát hiện tin giả bằng máy học.....	121
Sonali Raturi, Amit Kumar Mishra và Srabanti Maji	
Chương 11 Tương lai của DeepFakes và Ectypes.....	135
Bò tót Loveleen, Mansi Ratta, và Bò tót Adesh	
Mục lục.....	147

lời nói đầu

Tôi rất vui khi đặt tựa đề cuốn sách “DeepFake: Sáng tạo, Phát hiện và Tác động”. Một trong những ứng dụng đáng sợ nhất của thuật toán học sâu tạo ra tiếng vang ngày nay là “DeepFakes”. Nó là một ứng dụng hỗ trợ học tập sâu được phát triển gần đây. DeepFakes có thể là nội dung hình ảnh, âm thanh hoặc video trông cực kỳ chân thực đối với con người, đặc biệt khi được sử dụng để tạo và thay đổi/hoán đổi hình ảnh khuôn mặt. Các thuật toán có thể tạo ra ảnh và video giả mà con người không phân biệt được với ảnh gốc. Nó là sự kết hợp giữa “deep learning và fake”. Các video giả mạo được phát triển với sự trợ giúp của Trí tuệ nhân tạo (AI) và phần mềm Mạng đối thủ sáng tạo (GAN) cực kỳ ấn tượng, để mô tả những người làm hoặc nói những điều mà họ chưa bao giờ làm. Những video này có vẻ khá thực tế.

Việc sử dụng rộng rãi các nền tảng truyền thông xã hội chia sẻ hình ảnh cung cấp một lượng lớn dữ liệu kết hợp với các kỹ thuật học sâu, đặc biệt là GAN, để tạo ra các DeepFakes có vẻ chân thực. Nó có khả năng tạo ra nội dung gây hiểu lầm và rõ ràng, với những nhân vật có ảnh hưởng chủ yếu như người nổi tiếng, chính trị gia và thậm chí cả các nhà lãnh đạo tôn giáo là mục tiêu. Khái niệm này đang phát triển nhanh chóng và trở nên nguy hiểm, không chỉ đối với danh tiếng của nạn nhân mà còn đối với sự an toàn của họ.

Trong kích bản đầy thách thức này, các thuật toán là cần thiết để vạch mặt DeepFakes, phát hiện chúng hoặc ít nhất là giảm thiểu tác hại và lạm dụng tiềm ẩn có thể xảy ra bằng cách sử dụng các nội dung đa phương tiện này.

Cuốn sách này rất thích hợp trong những thời điểm đặc biệt này. Cuốn sách tóm tắt các xu hướng và tác động của việc nhận các phiên bản đã thay đổi của chính bạn và của những người khác bằng cách sử dụng công nghệ DeepFakes. Cuốn sách cũng tóm tắt việc tiếp xúc với DeepFakes làm suy yếu niềm tin vào phương tiện truyền thông như thế nào; cách công nghệ tương tác với DeepFakes; độ bền và khả năng phục hồi của DeepFakes; các chiến lược xung quanh việc gỡ lỗi hoặc chống lại DeepFakes; và hiểu việc sử dụng DeepFakes để tự trình bày trong quá trình tương tác xã hội. Tôi tin chắc rằng nguồn tài liệu tham khảo quan trọng này là lý tưởng cho các nhà nghiên cứu, học giả, học viên, công ty và sinh viên.

CHƯƠNG 1: GIỚI THIỆU CÔNG NGHỆ DEEPFAKE

Chương này giới thiệu khái niệm về các công cụ/công nghệ DeepFakes để tạo và phát hiện DeepFakes. Nó xây dựng nền tảng để người đọc hiểu cách sử dụng DeepFakes hiện tại trong kinh doanh.

CHƯƠNG 2: DEEPFAKES: ĐÁNH GIÁ HỆ THỐNG VÀ PHÂN TÍCH BIBLIOGRAPHIC

Chương này tập trung vào việc xem xét tài liệu một cách có hệ thống, xác định các phương pháp chính hiện có để xác định phương tiện giả mạo và áp dụng chúng trong các tình huống khác nhau. Chương này tập trung vào các tạp chí hàng đầu và các quốc gia hàng đầu có ảnh hưởng khoa học về DeepFake.

CHƯƠNG 3: CÁC KỸ THUẬT HỌC SÂU ĐỂ SÁNG TẠO CỦA DEEPFAKES

Chương này tập trung vào các kỹ thuật học sâu nâng cao để tạo DeepFakes. Chương này nhằm mục đích giúp người đọc hiểu rõ hơn về việc triển khai học sâu trong DeepFakes.

CHƯƠNG 4: PHÂN TÍCH VIDEO DEEPFAKES BẰNG KHÔN MẶT HIỆN TƯỢNG VỐNG

Mục tiêu của chương này là sử dụng các tạo tác cong vênh khuôn mặt để phân biệt video DeepFake với video thật một cách hiệu quả và để hiểu sự khác biệt giữa hai loại này.

CHƯƠNG 5: PHÁT TRIỂN DỊCH ẢNH MÔ HÌNH CHỐNG LẠI CÁC TẤN CÔNG ĐỐI THỦ

Mục tiêu của chương này là phát triển một mô hình để đối phó với việc lạm dụng mô hình tạo sâu bằng cách sử dụng các cuộc tấn công đối nghịch để tạo ra các cảnh báo tinh vi có thể khiến các thuật toán tạo sâu không thể tạo ra hình ảnh giả ngay từ đầu.

CHƯƠNG 6: PHÁT HIỆN DEEPFAKES BẰNG CÁC TÍ NH NĂNG ĐỊA PHƯƠNG VÀ MẠNG NEURAL XOAY CHIỀU

Chương này phân tích DeepFakes của khuôn mặt người để phát hiện dấu vết pháp y ẩn trong hình ảnh bằng cách sử dụng các tính năng cục bộ và quy trình tạo tích chập.

CHƯƠNG 7: DEEPFAKES: CÁC TRƯỜNG HỢP TÍ CH CỰC

Chương này tập trung vào điểm tiềm quan trọng của DeepFakes có hai hướng; có nhiều trường hợp sử dụng tích cực khác nhau của DeepFakes. Chương này nhằm mục đích xác định và phân tích mặt tích cực của DeepFakes, tức là giáo dục, trở ngại về lời nói, v.v.

CHƯƠNG 8: CÁC MỐI ĐE DỌA VÀ THÁCH THỨC CỦA DEEPFAKE CÔNG NGHỆ

Chương này nhằm mục đích phân tích tác động của các mối đe dọa và thách thức khác nhau do DeepFakes gây ra cho xã hội.

CHƯƠNG 9: DEEPFAKES, TRUYỀN THÔNG VÀ TÁC ĐỘNG XÃ HỘI

Chương này mang đến một nghiên cứu toàn diện về DeepFakes, phương tiện truyền thông và tác động của chúng đối với địa chính trị.

CHƯƠNG 10: DÙNG MÁY PHÁT HIỆN TIN GIẢ HỌC HỎI

Chương này tập trung vào các kỹ thuật DeepFakes nâng cao để tạo tin tức giả mạo.

CHƯƠNG 11: TƯƠNG LAI CỦA DEEPFAKES VÀ ECTYPES

Chương này tập trung vào các khía cạnh trong tương lai nơi các kỹ thuật DeepFakes có thể được triển khai và cung cấp các ectype nơi DeepFakes đã được triển khai.

Các công nghệ kỹ thuật số như trí tuệ nhân tạo, học máy và học sâu sẽ rất quan trọng trong cách DeepFakes mang lại lợi thế kinh doanh. Tôi cảm ơn các tác giả đáng kính của chúng tôi đã thể hiện sự tin tưởng vào cuốn sách và coi nó như một nền tảng để giới thiệu và chia sẻ tác phẩm gốc của họ.

biên tập viên:

Loveleen Gaur, Trường Kinh doanh Quốc tế Amity (AIBS)

Đại học Amity, Noida, Ấn Độ



Taylor & Francis
Taylor & Francis Group
<http://taylorandfrancis.com>

Sự nhìn nhận

Không có gì quan trọng và khẩn cấp hơn là ơn. Nhiều cá nhân đã đưa ra những gợi ý và phê bình có giá trị của họ, giúp xuất bản lần xuất bản đầu tiên. Hàng chục sinh viên đã tham gia vào các cuộc thảo luận về các chương khác nhau, ứng dụng phần mềm, phân tích vấn đề và thu thập tài liệu. Không thể nêu tên tất cả những người đã tham gia vào dự án này, nhưng tôi xin gửi lời cảm ơn tới tất cả họ.

Các cá nhân cụ thể đã có những đóng góp đáng kể và do đó họ xứng đáng được ghi nhận đặc biệt.

Đầu tiên, tôi đánh giá cao nỗ lực của những cá nhân đã cung cấp đánh giá chính thức về cuốn sách xuất bản đầu tiên.

Adesh Gaur, Có vẫn cấp cao về phát triển phần mềm, NTT Data

Gurinder Singh, Phó hiệu trưởng nhóm, Đại học Amity

Gurmeet Singh, Đại học Nam Thái Bình Dương, Fiji

Noor Zaman Jhanjhi, Đại học Taylors, Subang Jaya, Selangor, Malaysia

Tarun Kumar Singhal, Christ (Được coi là Đại học), Cơ sở Delhi NCR, Ấn Độ

Saurav Mallik, Đại học Harvard, MA, Hoa Kỳ

Tôi cũng biết ơn Gagandeep Singh và Aditi Mittal tại Taylor và Francis vì đã tin tưởng vào công việc của tôi. Nếu không có sự giúp đỡ của họ, việc tạo ra cuốn sách này sẽ không thể thực hiện được.

Ngoài ra, tôi muốn bày tỏ lòng biết ơn trước sự hỗ trợ và tình yêu thương của gia đình tôi-mẹ tôi (Amarjeet Kaur), chồng tôi (Adesh Gaur), con gái tôi (Devanshi Gaur) và con trai tôi (Raghav Gaur). Chúng là trụ cột và sức mạnh giúp tôi tiếp tục-cuối cùng, lòng biết ơn của tôi đối với quyền năng thiêng liêng và các phước lành của NGÀI.

Bò tót Loveleen



Taylor & Francis
Taylor & Francis Group
<http://taylorandfrancis.com>

tiểu sử biên tập viên

Loveleen Gaur hiện đang làm Giáo sư kiêm Giám đốc Chương trình (Trí tuệ nhân tạo và Phân tích dữ liệu) tại Trường Kinh doanh Quốc tế Amity, Đại học Amity, Ấn Độ. Bà có hơn 20 năm kinh nghiệm giảng dạy, nghiên cứu và quản lý quốc tế. Cô là giám đốc sáng lập chương trình MBA về Trí tuệ nhân tạo và Phân tích dữ liệu tại Trường Kinh doanh Quốc tế Amity. Cô đang giám sát một số học giả tiến sĩ và sinh viên sau đại học, chủ yếu về Trí tuệ nhân tạo và Phân tích dữ liệu cho kinh doanh và quản lý. Dưới sự hướng dẫn của cô ấy, cụm nghiên cứu AI/Phân tích dữ liệu đã xuất bản rộng rãi trên các tạp chí có yếu tố tác động cao và đã thiết lập sự hợp tác nghiên cứu sâu rộng trên toàn cầu với một số chuyên gia nổi tiếng.

Cô ấy là thành viên cấp cao của IEEE và là Biên tập viên sê-ri với CRC và Wiley. Cô ấy có các án phẩm được lập chỉ mục cao trong SCI/ABDC/WoS/Scopus và có một số Bằng sáng chế/bản quyền trên tài khoản của cô ấy và đã biên tập/tác giả hơn 20 cuốn sách nghiên cứu được xuất bản bởi các nhà xuất bản tầm cỡ thế giới. Cô có kinh nghiệm tuyệt vời trong việc giám sát và đồng giám sát các sinh viên sau đại học quốc tế. Rất nhiều nghiên cứu sinh tiến sĩ và thạc sĩ đã tốt nghiệp dưới sự hướng dẫn của cô. Cô là người kiểm tra/đánh giá luận án tiến sĩ/thạc sĩ bên ngoài cho một số trường đại học trên toàn cầu. Cô đã hoàn thành thành công các khoản tài trợ nghiên cứu được tài trợ liên quốc gia. Cô cũng đã từng là diễn giả chính cho một số hội nghị quốc tế, trình bày một số hội thảo trên web trên toàn thế giới và chủ trì các phiên hội nghị quốc tế. Bò tốt đã góp phần đáng kể vào việc nâng cao hiểu biết khoa học bằng cách tham gia hơn 300 hội nghị khoa học, hội nghị chuyên đề và hội thảo chuyên đề, bằng cách chủ trì các phiên họp kỹ thuật và phát biểu toàn thể và các cuộc nói chuyện được mời.



Taylor & Francis
Taylor & Francis Group
<http://taylorandfrancis.com>

người đóng góp

Ratish Agrawal Phó giáo sư, Khoa CNTT, UIT RGPV Bhopal, Ấn Độ	Srabanti Maji trợ lý giáo sư, Đại học DIT, Dehradun, Ấn Độ
Mohan Bhandari Giảng viên NCIT, Népal	Saurav Mallik nghiên cứu sinh sau tiến sĩ Đại học Harvard, MA, Mỹ
Amlan Chakrabarti Giáo sư kiêm Giám đốc Trường Công nghệ Thông tin AKChoudhury, Đại học Calcutta, Calcutta, Ấn Độ	Amit Kumar Mishra Phó Giáo sư kiêm Trưởng phòng trường máy tính, Đại học DIT, Dehradun, Ấn Độ
Ajantha Devi Vairamani Trưởng phòng Nghiên cứu, Giải pháp AP3 Chennai, Tamil Nadu, Ấn Độ	Shubha Mishra trợ lý giáo sư Đại học Công nghệ Lakshmi Narain, Bhopal, Ấn Độ
Bò tót Adesh Kiến trúc sư phần mềm cao cấp, Dữ liệu NTT Khu vực 125, Noida, Ấn Độ	Jyoti Rana Học giả nghiên cứu, Amity College of Thương mại và Tài chính, Khu vực 125, Đại học Amity, Noida, Ấn Độ
Bò tót Loveleen Giáo sư và Giám đốc Chương trình Kinh doanh Quốc tế Amity Trường học, Khu vực 125, Đại học Amity, Noida, Ấn Độ	Shreya Rastogi Kỹ sư phần mềm, ngân xếp xenon Mohali, Chandigarh, Ấn Độ
Noor Zaman Jhanjhi Phó Giáo sư, Giám đốc Trung tâm Xã hội Thông tỉnh 5.0 [CSS5], Đại học Taylor's, Malaysia	Mansi Ratta Phân tích kinh doanh, Phần mềm Prospecta, Noida, Ấn Độ
Gursimar Kaur Arora Tư vấn, Deloitte Mỹ Delhi, Ấn Độ	Sonali Raturi Học giả, Đại học DIT, Uttarakhand, Ấn Độ
tanvi razdan Chuyên viên phân tích kinh doanh, Hyundai Motor Ấn Độ Giới hạn (HMIL) Noida, Ấn Độ	

xviii

người đóng góp

Mamta Sareen
Trưởng Bộ phận
Đại học Delhi, Delhi, Ấn Độ

Piyush Kumar Shukla
Phó Giáo sư, Khoa học Máy tính
& Phòng Kỹ thuật, Ấn Độ
Viện Đại học Công nghệ,
Rajiv Gandhi Proudyogiki
Vishwavidyalaya, Ấn Độ

1 Giới thiệu về Công nghệ DeepFake

Bò tót Loveleen, Saurav Mallik, và
Noor Zaman Jhanjhi

NỘI DUNG

1.1 Giới thiệu	1
1.2 Làm sáng tỏ DeepFake.....	2
1.3 Nguồn gốc và lịch sử	2
1.4 Xu hướng ngày càng tăng của DeepFake	3
1.5 Tại sao nó lại là một vấn đề cần quan tâm?.....	4
1.6 DeepFakes hoạt động như thế nào?.....	4
1.6.1 Phân tích công nghệ	5
1.7 Ảnh hưởng của DeepFake.....	5
1.8 Tóm tắt	6
Người giới thiệu.....	6

1.1 GIỚI THIỆU

Ký nguyên hiện nay có thể được đặc trưng rõ ràng bởi sự thống trị của kỹ thuật số, trong đó việc tạo ra, truyền thông và phổ biến thông tin được điều khiển bằng kỹ thuật số. Nó đã đặt ra một tình trạng đáng báo động và đầy thách thức về sự tin cậy và xác minh nội dung kỹ thuật số có sẵn cho công dân. AI là một công nghệ thay đổi mô hình do các chức năng thực dụng đa dạng của nó được thể hiện trong những năm qua.

Lĩnh vực con của AI là Deep Learning (DL), đóng vai trò to lớn trong việc phát triển nhiều ứng dụng. Câu khẩu hiệu DF bắt nguồn từ sự đổi mới được che giấu, tức là nhận thức sâu sắc-một loại AI [1]. Công nghệ DF đang tiết lộ một thời đại sản xuất phương tiện khác. Giống như tất cả các công nghệ khác, cá trường hợp sử dụng mang tính xây dựng và nguy hiểm đều xảy ra.

Như tên gợi ý, DL được dạy từ một lượng lớn dữ liệu và có nhiều cấp độ (sâu) tạo điều kiện thuận lợi cho việc học. Tương tự như cách con người học hỏi từ kinh nghiệm, thuật toán DL sẽ thực hiện một nhiệm vụ lặp đi lặp lại, mỗi lần điều chỉnh nó để nâng cao kết quả. DL cho phép các máy giải quyết các vấn đề phức tạp đối với các bộ dữ liệu khác nhau, không có cấu trúc và được kết nối với nhau. Hiệu suất của các thuật toán phụ thuộc vào việc học chuyên sâu. DL có những khả năng tinh vi để tạo hoặc thay đổi hình ảnh, chữ viết và biểu thức một cách cực kỳ thiết thực.

Về bản chất, DL đã thúc đẩy việc tạo ra các văn bản giả mạo, âm thanh nhân tạo, video giả mạo và ảnh giả mạo, tất cả đều có vẻ hợp pháp và chính xác một cách đáng kinh ngạc. Tuy nhiên, chúng không phải [2].

Những tiến bộ hiện đại trong AI và DL đã thúc đẩy xu hướng DF (chính thức là hình ảnh hoặc video được chỉnh sửa tài liệu). Xu hướng hình thành các tác phẩm trực tuyến được kiểm chứng đáng tin cậy đang tăng cao trên các nền tảng truyền thông xã hội với rất nhiều hình ảnh giả mạo và vô số video về những người nổi tiếng, chính trị gia và nhân vật nổi tiếng. Rủi ro và các tác động xã hội là đáng kể và có sức tàn phá lớn, đặc biệt là với trình độ kỹ thuật tối thiểu và các thiết bị cần thiết để sản xuất DF.

Nội dung như vậy có thể được tạo ra dễ dàng bởi bất kỳ ai và đồng minh điện tử phân tán. Do đó, điều này bắt buộc phải điều tra kỹ lưỡng về DF thông qua nhiều lăng kính khác nhau, bao gồm phương tiện truyền thông, xã hội, nền tảng kỹ thuật số, người xem, đặc điểm tình dục, luật pháp, quy định và niềm tin chính trị. Để hiểu tầm quan trọng của DF, người ta phải hiểu khái niệm cơ bản. Đầu tiên, chương này sẽ thảo luận về ý tưởng, nguồn gốc, lịch sử, xu hướng và tác động đối với xã hội.

1.2 GIẢI QUYẾT DEEPFAKES

DeepFake là một tập hợp "học sâu" và "giả mạo", sử dụng thuật toán DL để sửa đổi hình ảnh, âm thanh và video nhằm tạo phương tiện tổng hợp/giả mạo. Đây là một quy trình không độc lập áp dụng các thuật toán AI cho chủ đề, tạo ra hình ảnh, video và âm thanh được chỉnh sửa.

Công nghệ cơ bản có thể phủ hình ảnh khuôn mặt, tạo chuyển động khuôn mặt, chuyển đổi khuôn mặt, điều khiển nét mặt, tạo khuôn mặt và tổng hợp lời nói của một cá nhân mục tiêu vào video của người phát ngôn để tạo video về cá nhân mục tiêu hành động tương tự như người nguồn. Việc mạo danh tiếp theo thường thực tế không thể phân biệt được với bản gốc. Những công cụ này thường được áp dụng để mô tả các cá nhân tuyên bố hoặc thực hiện điều gì đó mà họ không bao giờ làm trong các tình huống điển hình. DL có các ứng dụng quan trọng trong nhiều loại khó khăn phức tạp trong thế giới thực, thay đổi từ độ nhạy thị giác máy tính phân tích dữ liệu lớn đến các hệ thống điều khiển tự động. Thật không may, với sự phát triển của các kỹ thuật DL, các rủi ro đối với quyền riêng tư, sức mạnh và sự an toàn của các hệ thống dựa trên Học máy (ML) cũng đã phát triển.

1.3 NGUỒN GỐC VÀ LỊCH SỬ

Thị giác máy tính là một lĩnh vực phức tạp chủ yếu liên quan đến xử lý hình ảnh, cung cấp cho PC kỹ năng nhận biết kiến thức từ hình ảnh. Các ứng dụng phổ biến bao gồm xe tự lái, chẩn đoán bệnh trong chăm sóc sức khỏe và nhận diện khuôn mặt của Facebook để gợi ý gắn thẻ ảnh. Công nghệ DFs nằm dưới sự bảo trợ của thị giác máy tính.

Nền tảng của DF được đặt ra vào năm 1997 khi Bregler et al. đã thiết lập "Chương trình ghi lại video" [3] mang tính đột phá có thể tạo ra các mô phỏng khuôn mặt mới từ đầu ra âm thanh. Tuy nhiên, bài báo này vẫn là bản gốc để đặt ba khái niệm này và làm sinh động chúng một cách thực tế. Nó được coi là một trong những công việc thiết yếu trong việc tạo ra nền tảng công nghệ DFs [4].

Giới thiệu về công nghệ DeepFake

Năm 2001, một bài báo nổi tiếng khác của Cootes et al. trên thuật toán mô hình xuất hiện hoạt động (AAM) đã được xuất bản, sử dụng nguyên mẫu thống kê toàn diện để khớp hình dạng với hình ảnh; một đóng góp đáng chú ý trong lĩnh vực theo dõi và đổi khuôn mặt [5].

Theis et al. (2016), trong Face2Face của họ, cố gắng tạo hoạt ảnh tức thời, hoán đổi vùng miêng của video mục tiêu với một diễn viên; video này không có giọng nói. Tương tự, trong bài báo tổng hợp về Obama, Suwajanakorn et al. (2017) đã nâng cao các cải tiến về đồ họa với nhiều hoạt ảnh, kết cấu và biểu thức hơn. Mặc dù mục tiêu của cả hai bài báo là khác nhau, nhưng những bài báo này đã cải thiện thời gian xử lý và dịch thuật trong khi những bài báo khác đang đổi mới sự phù hợp về đồ họa để trông giống như ảnh thực. Những bài báo này là những cột mốc quan trọng trong việc phát triển DF [6,7]. Thuật ngữ DF được phát minh vào năm 2017 bởi một khách hàng Reddit có tên giống hệt. Anh ta bắt đầu đăng những hình ảnh và video khiêu dâm của những người nổi tiếng bằng công nghệ hoán đổi khuôn mặt mã nguồn mở. Sau đó, Reddit đã cấm người dùng và cập nhật chính sách nội dung của họ. Kể từ đó, thuật ngữ này đã được mở rộng để kết hợp “các ứng dụng truyền thông tổng hợp” và các sáng tạo đổi mới như StyleGAN (hình ảnh của những người trông giống thật nhưng không tồn tại). Xu hướng gần đây đang hướng tới thao túng lời nói được ghi âm [8,9].

1.4 XU HƯỚNG PHÁT TRIỂN CỦA DEEPFAKES

Với sự tiến bộ của AI và thị giác máy tính, xu hướng đang chuyển từ mối quan hệ của người nổi tiếng với công chúng sang chòng các hình ảnh và video phổ biến lên các hình ảnh hoặc video nguồn bằng kỹ thuật mạng đối nghịch chung (GAN) [10].

Xu hướng ngày càng tăng đối với các video giả mạo vì mục đích chính trị hoặc khiêu dâm. Theo báo cáo do Sensity công bố vào năm 2019, có tổng cộng 14.678 video DF được tìm thấy trực tuyến. Người ta cũng quan sát thấy rằng 96% đã được sử dụng trong các vấn đề khiêu dâm. Xu hướng ngày càng tăng và số lượng DF đang tăng lên. Ngoài những người nổi tiếng, những người có ảnh hưởng trên mạng xã hội nổi tiếng và những nhân vật nổi tiếng trên mạng cũng được biết đến rộng rãi. Ngoài ra, những người khởi tạo cũng nhắm đến mọi người, thường là phụ nữ, những người dùng tích cực. Mỗi đe dọa rõ ràng nhất được đặt ra cho phụ nữ ngay bây giờ với sự raphy phim khiêu dâm không có sự đồng thuận. Ngoài ra, xu hướng khiêu dâm báo thù giả đang phát triển. Mối nguy hiểm không chỉ giới hạn ở phụ nữ; phong trào có thể lan rộng trong trường học hoặc nơi làm việc, vì bất kỳ ai cũng có thể đặt mọi người vào những tình huống vô lý, nguy hiểm hoặc thỏa hiệp. Những lo lắng khác liên quan đến DF là tổng tiền, gian lận danh tính, lừa đảo của các tập đoàn lớn và nguy cơ đổi với nền dân chủ [11].

Ngoài việc sử dụng rộng rãi nội dung khiêu dâm được thiết kế, không có sự đồng thuận, ngày càng có nhiều trường hợp DF được sử dụng để bắt chước ai đó đang cố gắng mở tài khoản ngân hàng. Mọi người có thể giả mạo ID và sự xuất hiện của họ trong video bằng cách thuật toán rất tinh vi. Mặc dù việc áp dụng cách lừa dối như vậy không phổ biến, nhưng nó biểu thị một ứng dụng xấu cho DF. Xu hướng và tác động lan rộng của nó được cho là sẽ tiếp tục trong thời gian tới với khả năng tiếp cận công nghệ dễ dàng; các quy định và chính sách rõ ràng về việc sử dụng AI là rất cần thiết để kiểm soát mặt tiêu cực [12].

1.5 TẠI SAO NÓ LÀ VẤN ĐỀ CẦN QUAN TÂM?

Những tiếng kêu gào xung quanh DF không phải là không có lý. Nó quan trọng do những điều sau đây mối quan tâm:

- **Thấy là tin:** Việc chúng ta tin vào những gì mắt thấy, tai nghe là điều khá thuyết phục đối với chúng ta. Không có khả năng không tin vào những điều bạn đã quan sát chính mình. Giờ đây, việc đánh lừa hệ thống thị giác của não bộ bằng nhận thức sai lầm bằng cách sử dụng những công nghệ mới nhất đang bùng nổ này trở nên dễ dàng hơn nhiều.
- **Tính khả dụng:** Với các ứng dụng mới dễ sử dụng, việc tạo nội dung lừa đảo như vậy dưới dạng hình ảnh, video hoặc bất kỳ hình thức phương tiện nào khác sẽ dễ dàng hơn nhiều. Khả năng truy cập ngày càng tăng cùng với sự gia tăng của các công cụ và công nghệ. Ví dụ: Ứng dụng Zao cho phép người dùng đặt khuôn mặt của họ vào các đoạn phim/TV.

1.6 DEEPFAKES HOẠT ĐỘNG NHƯ THẾ NÀO?

Quy trình ba bước của bất kỳ DF nào bao gồm:

1. Trích xuất ảnh gốc từ khung gốc.
2. Sử dụng hình ảnh được trích xuất này làm đầu vào cho các thuật toán DL sẽ tự động tạo ra phù hợp chính xác cho hình ảnh ban đầu.
3. Ảnh kết xuất sau đó được chèn vào ảnh tham chiếu ban đầu để tạo ra một bức ảnh giả.

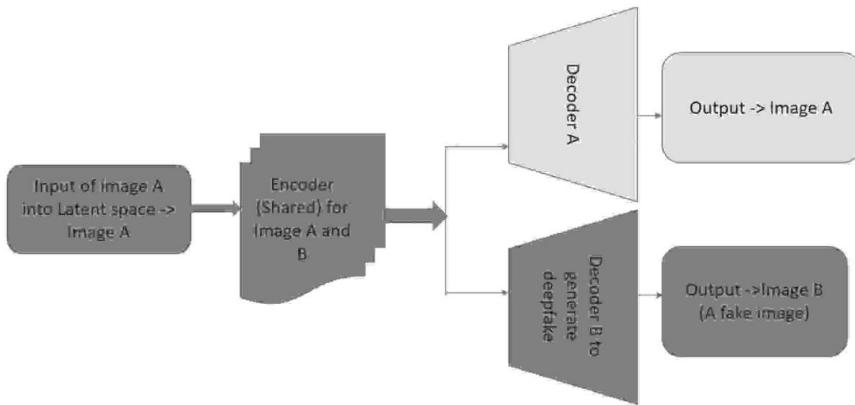
DF chủ yếu được xây dựng bằng cách sử dụng “bộ mã hóa tự động”, một kiến trúc mạng sâu [13,14].

Bộ mã hóa tự động được huấn luyện để nhận biết các đặc điểm chính của hình ảnh đầu vào để sau đó tái tạo nó thành đầu ra của chúng. Trong quá trình này, mạng thực hiện nén dữ liệu nặng. Sau đây là ba phần con của bộ mã hóa tự động:

- **Bộ mã hóa:** Có nhiệm vụ trích xuất các đặc điểm quan trọng từ hình ảnh đầu vào. Bộ mã hóa nén ảnh gốc từ hàng nghìn pixel thành hàng trăm pixel. Các phép đo này liên quan đến các đặc điểm trên khuôn mặt như chuyển động của mắt, tư thế đầu, màu da, biểu cảm, v.v.
- **Không gian tiềm ẩn đại diện** cho các đặc điểm khuôn mặt duy nhất mà hình ảnh được đào tạo. Nó tập trung hơn vào các đặc điểm quan trọng trên khuôn mặt. Nó loại bỏ phần nhiễu/không quan trọng của hình ảnh, cho biết hình ảnh là một phiên bản nén, giúp ích tối đa cho việc ghi nhớ các đặc điểm cơ bản.
- **Bộ giải mã** giải nén thông tin trong không gian tiềm ẩn để tái tạo lại hình ảnh giống với hình ảnh gốc. Việc so sánh các hình ảnh đầu vào và đầu ra cung cấp hiệu suất của bộ mã hóa tự động. Hình ảnh đầu vào và đầu ra càng giống nhau thì hiệu suất của bộ mã hóa càng cao [13,14].

Nếu hai bộ mã hóa tự động riêng biệt được đào tạo trên những người khác nhau, thì việc tích hợp sẽ khó đạt được. Bí quyết để tạo DF là giao tiếp bộ mã hóa qua hai mạng để duy trì tính nhất quán. Nó có nghĩa là hình ảnh của một người có thể được sử dụng để

Giới thiệu về công nghệ DeepFake



HÌNH 1.1 Hình minh họa “Cách hoạt động của DF?”

tính toán một biểu diễn không gian tiềm ẩn được nén, từ đó bộ giải mã của người khác được sử dụng để tạo ra hình ảnh được chỉnh sửa/giả mạo.

1.6.1 Phân tích Công nghệ

Các nhà nghiên cứu đã chứng minh rằng có thể dễ dàng xác định các tiêu chí để xác định DF dựa trên số lượng tham số. Các thông số cụ thể như sau:

- Tổng số hình ảnh
- Tinh huống chiểu sáng/chiểu sáng
- Quy mô và chất lượng đầu vào
- Vị trí của hình ảnh đầu vào
- Thay đổi hình dạng khuôn mặt
- Các đối tượng giao nhau

1.7 ẢNH HƯỞNG CỦA DEEPFAKES

DF là hoạt động khai thác công nghệ AI và ML nghiêm trọng có thể ảnh hưởng và đe dọa mọi người cũng như tổ chức. Họ sẽ gây xáo trộn cho xã hội bằng cách thao túng cảm xúc và quan điểm của mọi người. Việc sử dụng công nghệ một cách phi đạo đức có những hậu quả lâu dài trong tương lai đối với xã hội và người dân nói chung [15].

DFs chính trị là đối tượng nổi bật của thông tin sai lệch dựa trên video có sẵn trực tuyến. Ví dụ: nếu không bị thách thức, các DF phổ biến của Barack Obama và Donald Trump có tác động sâu sắc đến báo chí, năng lực công dân và chất lượng của nền dân chủ. Bằng chứng gián tiếp chỉ ra rằng khả năng sản xuất hàng loạt và phổ biến DF bởi những kẻ tinh quái có thể là một thách thức nghiêm trọng đối với tính hợp pháp của các cuộc đối thoại chính trị trực tuyến. Nó có thể đóng một vai trò quan trọng trong việc tác động đến niềm tin của công chúng vào quá trình chuẩn bị cho các cuộc bầu cử. Một hệ quả nghiêm trọng khác của DF là việc nó được sử dụng rộng rãi trong việc tạo ra các hình ảnh/video khiêu dâm nhằm làm sai lệch

người nổi tiếng trong giả mạo. Chẳng hạn, Emma Watson, Natalie Portman và Gal Gadot là những người nổi tiếng phổ biến bị ảnh hưởng bởi DF [16,17].

Các nền tảng truyền thông xã hội là nền tảng được ưa thích nhất để đăng nội dung độc hại như vậy. Với việc số hóa mọi lĩnh vực, các cá nhân sử dụng phương tiện truyền thông xã hội và nền tảng kỹ thuật số để lấy thông tin. Với sự gia tăng của DFs, lòng tin của mọi người đã được khơi dậy và cuối cùng nó sẽ đặt ra những thách thức thực sự cho xã hội [18,19,20,21,22,23,24].

Các hệ quả tiềm tàng từ công nghệ DF có thể tàn khốc hơn.

Các công nghệ dựa trên AI sẽ mang tiếng xấu và có thể cản trở sự phát triển và tăng trưởng xung quanh công nghệ tiềm năng này [25,26,27]. Tuy nhiên, các quan chức có thể bảo vệ an toàn cho bản thân và doanh nghiệp của họ với cái nhìn sâu sắc về công nghệ.

Do đó, các doanh nghiệp đã bắt đầu cung cấp dịch vụ phát hiện DF, giúp mọi người phân biệt hình ảnh, âm thanh hoặc video giả với hình gốc với độ chính xác chấp nhận được cao nhất [28,29,30].

1.8 TÓM TẮT

Thao tác dữ liệu không có gì mới; đó là xu hướng lâu đời khi Joseph Stalin sử dụng đòn áp và chỉnh sửa hình ảnh để gây ảnh hưởng đến tính cách và chính phủ của ông ta vào đầu thế kỷ 20. Sự phát triển vượt bậc của máy tính đã tiếp thêm nhiên liệu, và giờ đây việc thao tác chỉ là vấn đề của một vài cú nhấp chuột. Ngoài ra, các nhà nghiên cứu đã chứng minh rằng các cá nhân có khuynh hướng tin tưởng vào đôi mắt và đôi tai của chính họ.

Do đó, khi các phương tiện truyền thông đại chúng mà họ sử dụng có vẻ quá tuyệt vời để có thể là giả mạo, thì thật dễ dàng để trở thành nạn nhân của trò bịa bợm. Trong khi các bức ảnh tĩnh "chỉnh sửa ảnh" đã trở thành một thành trì của văn hóa kỹ thuật số, các bức ảnh và video đã được chỉnh sửa của các cá nhân hiện đang diễn khám phá ra cách của chúng trực tuyến dưới dạng DF.

Chương tiếp theo sẽ tập trung vào công việc liên quan trong DF. Nó cũng sẽ tập trung vào các kỹ thuật AI được các nhà nghiên cứu sử dụng để tạo và phát hiện DF.

NGƯỜI GIỚI THIỆU

- [1] Meredith, S. (2019). DF, Giải thích, <https://mitsloan.mit.edu/ideas-made-to-matter/DFs-explained> (Truy cập vào ngày 17 tháng 8 năm 2021).
- [2] Cole, S. (24 tháng 1 năm 2018). "Chúng tôi thực sự chết tiệt: Bây giờ mọi người đang làm phim khiêu dâm giả do AI tạo ra." Hành vi xấu xa. Bản gốc lưu trữ ngày 7 tháng 9 năm 2019. Truy cập ngày 4 tháng 5 năm 2019.
- [3] Bregler, C., Covell, M., và Slaney, M. (1997). "Viết lại video: Thúc đẩy bài phát biểu trực quan bằng âm thanh," Kỷ yếu của Hội nghị thường niên lần thứ 24 về Đò họa máy tính và Kỹ thuật tương tác, 24: 353-360. doi:10.1145/258734.258880
- [4] Karnouskos, S. (tháng 9 năm 2020). "AIin Digital Media: Kỹ nguyên của DF," IEEE Transactions on Technology and Society, 1(3): 138-147. doi:10.1109/TTS.2020.3001312
- [5] Cootes, TF, Edwards, GJ, và Taylor, CJ (tháng 6 năm 2001). "Người mẫu ngoại hình năng động," Giao dịch của IEEE về Phân tích mẫu và trí thông minh của máy, 23(6): 681-682.
- [6] Westerlund, M. (2019). "Sự xuất hiện của Công nghệ DF: Đánh giá," Đánh giá Quản lý Đổi mới Công nghệ, 9(11): 39-54. <http://doi.org/10.22215/timreview/1282>; <https://timreview.ca/article/1282>

Giới thiệu về công nghệ DeepFake

- [7] Thies, J., Zollhöfer, M., Stamminger, M., Theobalt, C., và Nießner, M. (2016). "Face2Face: Chụp và tái hiện khuôn mặt trong thời gian thực của video RGB," Proc. Thị giác Máy tính và Nhận dạng Mẫu (CVPR), IEEE, 62(1): 96-104. doi:10.1145/3292039
- [8] Thies, J., Zollhöfer, M., Stamminger, M., Theobalt, C., và Nießner, M. (tháng 1 năm 2019). "Face2Face: Chụp và tái hiện khuôn mặt trong thời gian thực của video RGB," Truyền thông của ACM, 62(1): 96-104. doi:10.1145/3292039
- [9] Suwananakorn, S., Seitz, SM, và Kemelmacher-Shlizerman, I., (2017). "Tổng hợp về Obama: Học hát nhèo từ âm thanh," Đại học Washington, https://grail.cs.washington.edu/projects/AudioToObama/siggraph17_obama.pdf
- [10] Vaccari, C. và Chadwick, A. (2020). "DFs và thông tin sai lệch: Khám phá tác động của video chính trị tổng hợp đối với sự lừa dối, sự không chắc chắn và niềm tin vào tin tức," Social Media and Society 6(1): 1-13. <https://doi.org/10.1177/2056305120093408>
- [11] <https://blog.gao.gov/2020/10/20/deconstructing-DFs-how-do-they-work-and-what-are-the-risk/> (Truy cập ngày 17 tháng 8 năm 2021).
- [12] www.nortonlifelock.com/blogs/norton-labs/DFs-terror-era-ai (Truy cập vào tháng 8 17, 2021).
- [13] Sharma, DK, Gaur, L., và Okunbor, D. (2007). "Nén hình ảnh và trích xuất tính năng với mạng thần kinh," Kỷ yếu của Viện Khoa học Quản lý và Thông tin, 11(1): 33-38.
- [14] Singh, G., Kumar, B., Gaur, L. và Tyagi, A. (2019). "So sánh giữa Multinomial và Bernoulli Naïve Bayes để phân loại văn bản," Hội nghị quốc tế về tự động hóa, tính toán và quản lý công nghệ (ICACTM), trang 593-596, doi:10.1109/ICACTM.2019.8776800
- [15] <https://interculturaltalk.com/2019/11/05/cool-but-scaryDFs-are-here/> (Truy cập ngày 17 tháng 8 năm 2021).
- [16] <https://medium.com/@songda/a-short-history-of-DFs-604ac7be6016> (Truy cập ngày 17 tháng 8 năm 2021).
- [17] www.Discovermagazine.com/technology/DFs-the-dark-origins-of-fake-videos-and-their-potential-to-wreak-havoc (Truy cập vào ngày 17 tháng 8 năm 2021).
- [18] Anshu, K., Gaur, L., và Khazanchi, D. (2017). "Đánh giá mức độ hài lòng của các nhà bán lẻ tạp hóa điện tử bằng mô hình TOPSIS và ECCSI mở trực quan," Hội nghị quốc tế về Công nghệ Infocom và Hệ thống không người lái (Xu hướng và Định hướng Tương lai) (ICTUS), trang 276-284, doi:10.1109/ICTUS.2017. 8286019
- [19] Gaur, L., và Anshu, K. (2018). "Phân tích Sở thích của Người tiêu dùng đối với Trang web Sử dụng e TailQ và AHP," Tạp chí Kỹ thuật & Công nghệ Quốc tế, 7(2.11): 14-20.
- [20] Gaur, L., Singh, G., Solanki, A., Jhanjhi, NZ, Bhatia, U., Sharma, S., . và Kim, W. (2021). "Sứ mệnh của thanh niên trong việc dự đoán sự phát triển bền vững Các mục tiêu sử dụng thuật toán xứng ngẫu nhiên và thần kinh mở," Khoa học thông tin và điện toán lấy con người làm trung tâm, 11(NA): 1-19.
- [21] Gaur L., Agarwal V., và Anshu K. (2020). "Phương pháp tiếp cận DEMATEL mở để xác định các yếu tố ảnh hưởng đến hiệu quả của ngành bán lẻ Ấn Độ," Đảm bảo hệ thống chiến lược và Phân tích kinh doanh. Phân tích nội dung (Quản lý hiệu suất và an toàn). Springer, Singapore.

- [23] Ramakrishnan, R., Gaur, L., và Singh, G. (2016). "Tính khả thi và hiệu quả của các thiết bị BLE Beacon IoT trong quản lý hàng tồn kho tại phân xưởng," Tạp chí quốc tế về kỹ thuật điện và máy tính, 6(5): 2362-2368. doi:10.11591/ijece.v6i5.10807
- [24] Afaq, A., Gaur, L., Singh, G., và Dhir, A. (2021). "COVID-19: Thay đổi hành vi của hành khách đi máy bay và định hình lại kỳ vọng của họ đối với ngành hàng không," Nghiên cứu giải trí du lịch. doi:10.1080/02508281.2021.2008211
- [25] Rana, J., Gaur, L., Singh, G., Awan, U., và Rasheed, MI (2021). "Cung cấp hành trình của khách hàng thông qua trí tuệ nhân tạo: Chương trình nghiên cứu và đánh giá," Tạp chí quốc tế về các thị trường mới nổi, Tập. trước khi in (No. trước khi in). <https://doi.org/10.1108/IJOM-08-2021-1214>
- [26] Gaur, L., Afaq, A., Singh, G., và Dwivedi, YK (2021). "Vai trò của trí tuệ nhân tạo và người máy trong việc thúc đẩy du lịch không chạm trong thời kỳ đại dịch: Chương trình đánh giá và nghiên cứu," Tạp chí quốc tế về quản lý khách sạn đương đại, 33(11): 4079-4098. <https://doi.org/10.1108/IJCHM-11-2020-1246>
- [27] Sharma, S., Singh, G., Gaur, L., và Sharma, R. (2022). "Khoảng cách tâm lý và tôn giáo có ảnh hưởng đến hành vi lừa đảo của khách hàng không?" Tạp chí Nghiên cứu Người tiêu dùng Quốc tế. doi:10.1111/ijcs.12773
- [28] <https://spectrum.ieee.org/what-is-DF> (Truy cập ngày 17 tháng 8 năm 2021).
- [29] www.technologyreview.com/2020/12/24/1015380/best-ai-DFs-of-2020/ (Truy cập ngày 17 tháng 8 năm 2021).
- [30] <https://latinamericanpost.com/34481-what-are-DFs-and-why-are-they-such-a-big-problem> (Truy cập ngày 20 tháng 8 năm 2021).

2 DeepFake

Đánh giá có hệ thống và Phân tích thư mục

Bò tót Loveleen, Jyoti Rana, và
Amlan Chakrabarti

NỘI DUNG

2.1 Giới thiệu	9
2.2 Thu thập dữ liệu	10
2.2.1 Mục tiêu nghiên cứu	10
2.3 Kết quả và thảo luận	11
2.3.1 Xu hướng xuất bản.....	11
2.3.2 Đồng tác giả với tác giả	12
2.3.3 Đồng tác giả của Tổ chức	14
2.3.4 Đồng tác giả của các quốc gia	14
2.3.5 Sự xuất hiện đồng thời của các từ khóa	15
2.3.6 Liên kết thư mục của các tác giả.....	15
2.3.7 Danh sách các tạp chí tiêu biểu	15
2.3.8 Danh sách các Hội nghị Tiêu biểu	18
2.4 Thảo luận	18
2.5 Tóm tắt	19
Người giới thiệu.....	20

2.1 GIỚI THIỆU

DeepFakes (DF) đã xuất hiện từ lâu nhưng gần đây mới thu hút được nhiều sự quan tâm vì những lý do tốt và xấu. Các án phẩm sau đây đã cung cấp một quan điểm tốt về DF chính xác là gì và một bức tranh chi tiết hơn. Kiëtzmann và cộng sự. [1] đã nói về DF là gì, các loại DF hiện có khác nhau, công nghệ đằng sau chúng cũng như những cơ hội và vấn đề mà nó đặt ra và có thể tạo ra trong tương lai. Để giải quyết vấn đề về DF, họ cũng đã đề xuất “khuôn khổ THỰC SỰ”, một bộ nguyên tắc được thiết kế để quản lý các rủi ro liên quan đến DF. Khuôn khổ này nhằm mục đích sớm vạch trần DF, bảo vệ các cá nhân và tận dụng niềm tin để chống lại sự đáng tin cậy. Các tác giả [2] đã giải thích cách DF đã góp phần làm gia tăng tin giả và thông tin sai lệch trên mạng. Sau khi phân tích một nhóm dân số mẫu có trụ sở tại Vương quốc Anh, họ phát hiện ra rằng mọi người có nhiều khả năng cảm thấy không chắc chắn hơn là bị DF lừa dối, nhưng sự không chắc chắn này làm giảm niềm tin của họ vào các nguồn tin tức.

Chuyển sang phần kỹ thuật của DF, tức là cách chúng được tạo và có thể được phát hiện. Korshunov và Marcel [3] đã giới thiệu ngắn gọn về DF; họ

sau đó đã trình bày một bộ dữ liệu có sẵn công khai cho DF và tạo chúng thông qua mạng đối thủ tạo ra (GAN). Người ta thấy rằng các hệ thống nhận dạng khuôn mặt FaceNet và VGG dễ bị ảnh hưởng bởi DF và phương pháp nhất quán mô dựa trên âm thanh-hình ảnh không hiệu quả trong việc phát hiện DF. Sau khi tiến hành và phân tích nhiều quy trình hơn, người ta kết luận rằng phương pháp tốt nhất là, dựa trên các chỉ số chất lượng hình ảnh, tạo ra tỷ lệ lỗi là 8,97% cho DF chất lượng cao. Nghiên cứu của Kumar và cộng sự [4] đã cung cấp một cái nhìn tổng quan ngắn gọn về các nghiên cứu gần đây và các bộ dữ liệu được sử dụng để hỗ trợ nghiên cứu. Họ đã xem xét sự xuất hiện của DF và các cách chống lại chúng để tạo điều kiện phát triển một công nghệ và phương pháp tốt hơn và hiệu quả hơn để chống lại các vấn đề về DF. Xu et al. [5] đã cung cấp một cách duy nhất bằng cách xử lý DF như một vấn đề phân loại chi tiết và đề xuất một mạng phát hiện DF đa chú ý. Mô hình của họ bao gồm ba khái niệm quan trọng làm tăng hiệu quả tổng thể của mô hình. Với các thí nghiệm cường độ cao và dữ liệu đào tạo, họ đã cung cấp một số kết quả đầy hứa hẹn. Tương tự, Huang et al. (2020) [6] đã giới thiệu một khung đơn giản nhưng mạnh mẽ giúp giảm các mẫu hình ảnh giả mà không ảnh hưởng đến chất lượng hình ảnh. Quan sát chính của họ là "việc thêm nhiều vào hình ảnh giả có thể giảm thành công các mẫu tạo tác trong cả miền không gian và tần số". Họ đã sử dụng dữ liệu có sẵn về DF và tạo thêm dữ liệu từ nhiều GAN khác nhau để đào tạo và thử nghiệm với khuôn khổ của họ. Phương pháp của họ nhằm mục đích cải thiện độ trung thực của DF và phơi bày các vấn đề với các phương pháp phát hiện DF hiện có. Cuối cùng, Hernandez-Ortega et al. [7] đã đề xuất một phương pháp phát hiện DF dựa trên phát hiện nhịp tim. Họ đã sử dụng "chụp ảnh thể tích từ xa" (rPPG) để xem nhịp tim trong video.

Mạng chú ý tích chập (CAN), được sử dụng trong mô hình của họ, thu thập thông tin không gian và thời gian từ các khung hình video, phân tích và kết hợp cả hai nguồn để phát hiện phim giả hiệu quả hơn. Phương pháp nhận dạng này được thử nghiệm bằng cách sử dụng các cơ sở dữ liệu công cộng mới nhất hiện có trong lĩnh vực này: Celeb-DF và DFDC.

Kết quả rất khả quan, chứng tỏ sự thành công của máy dò giả dựa trên sinh lý học trong việc phát hiện các DF gần đây nhất.

Án phầm [8] của Ahmed đã giới thiệu ngắn gọn về DF, công nghệ cơ bản. Anh ấy đã tập trung làm nổi bật những lợi ích và mối đe dọa mà DF gây ra cho doanh nghiệp, xã hội và thế giới. Bài báo kết luận với các định hướng trong tương lai cho DF.

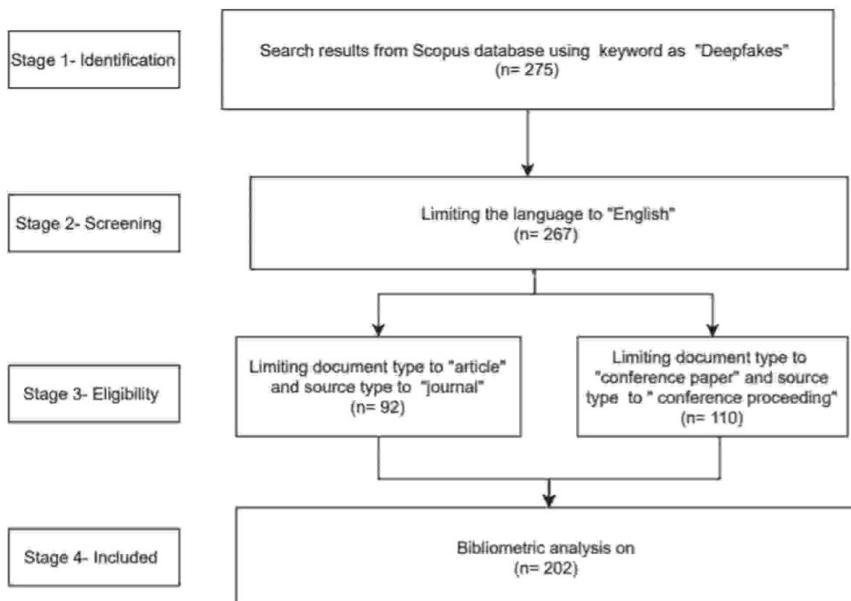
2.2 THU THẬP DỮ LIỆU

Các tác giả đã sử dụng một từ khóa duy nhất, "DeepFakes," để thu thập dữ liệu từ cơ sở dữ liệu Scopus. Kết quả tìm kiếm ban đầu là 254 đối với các loại tài liệu như bài báo và tài liệu hội nghị, các loại nguồn như tạp chí và kỹ yếu hội nghị, và Ngôn ngữ giải hạn đối với tiếng Anh. Dữ liệu được truy xuất vào ngày 26 tháng 10 năm 2021 lúc 2:42 chiều theo Giờ chuẩn Ấn Độ. Hình 2.1 mô tả quy trình tìm kiếm Scopus bằng cách sử dụng sơ đồ Mục báo cáo ưu tiên cho Đánh giá hệ thống và phân tích tổng hợp (PRISMA) [9].

2.2.1 Mục tiêu nghiên cứu

Trọng tâm của nghiên cứu là xác định các lĩnh vực chính và vai trò của DF trong kinh doanh bằng cách trả lời các câu hỏi nghiên cứu (RQ) sau đây.

Đánh giá có hệ thống và phân tích thư mục của DeepFakes



HÌNH 2.1 Sơ đồ quy trình tìm kiếm PRISMA Scopus.

RQ1: Tìm xu hướng xuất bản và các tạp chí quan trọng, xuất bản các bài báo nghiên cứu về ứng dụng của DF trong kinh doanh và lĩnh vực.

RQ2: Các nghiên cứu của DF đã tăng cường dòng nghiên cứu của họ như thế nào, chỉ ra các tác giả và quốc gia có ảnh hưởng nhất?

RQ3: Tiến hành phân tích từ khóa và phân tích đồng đồng trích dẫn cho các bài báo được xuất bản để tích hợp DF trong kinh doanh.

Một nghiên cứu thư mục (như trong Hình 2.2) cho phép nhà nghiên cứu tự do lập kế hoạch, tổ chức và điều tra tài liệu một cách có hệ thống và đạt được kiến thức toàn diện về lĩnh vực này.

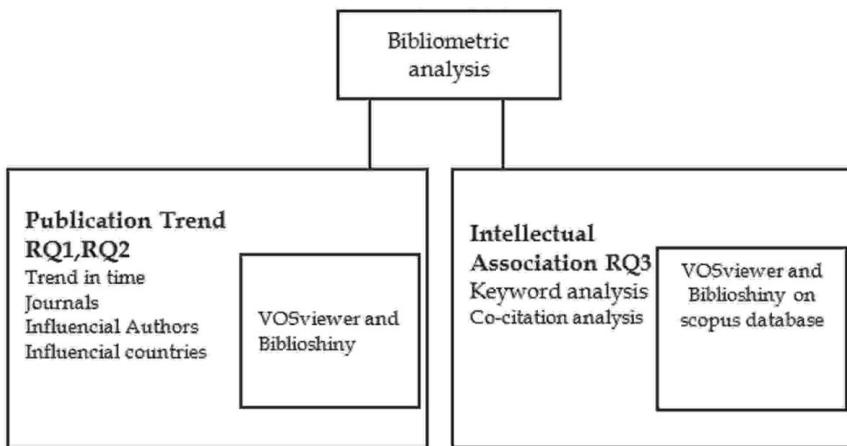
Bibliometrics là một cách tiếp cận đầy đủ và chi tiết để theo dõi giải phẫu tri thức trong bất kỳ lĩnh vực nghiên cứu nào [10].

2.3 KẾT QUẢ VÀ THẢO LUẬN

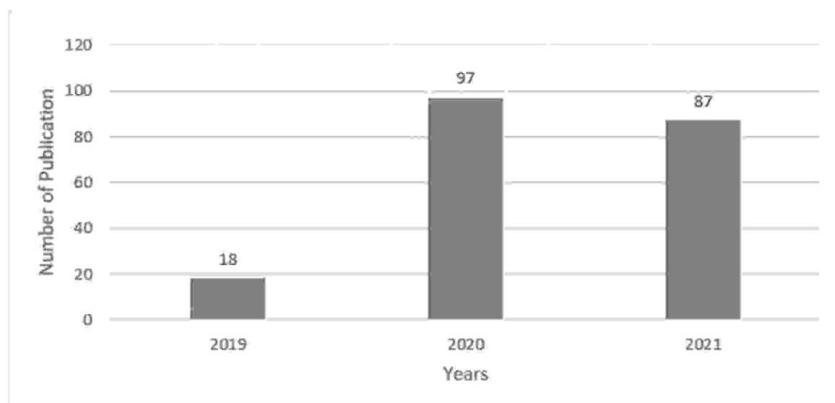
2.3.1 Xu hướng xuất bản

Xu hướng xuất bản có gắng tìm ra xuất bản diễn ra trong lĩnh vực DF trong những năm qua. Sự xuất hiện của DF là một khái niệm mới lạ; do đó nó có các nghiên cứu hạn chế. Xu hướng công bố khoa học về DF trong ba năm qua được thể hiện trong Hình 2.3 [11,12].

Năm 2019 đã mua án phẩm DF trên các tạp chí và hội nghị, với 18 án phẩm, tiếp theo là 97 án phẩm vào năm 2020. Sự gia tăng số lượng xuất bản là do đại dịch COVID-19, và trong năm 2021, cho đến giữa tháng 10 năm 2021, đã có 87 án phẩm [13].



HÌNH 2.2 Phương pháp và công cụ phân tích.



HÌNH 2.3 Xu hướng xuất bản.

2.3.2 Đồng tác giả với Tác giả

Đồng tác giả là khi hai hoặc nhiều người có những đóng góp đáng kể cho một bài báo. Đồng tác giả chia sẻ trách nhiệm và giải trình về kết quả.

Ở đây, phần mềm VOSviewer được sử dụng để trực quan hóa cơ sở dữ liệu Scopus. Giới hạn số tài liệu tối thiểu của một tác giả được đặt là 3 và số trích dẫn tối thiểu cũng được đặt là 3; trong số 582 tác giả chỉ có 23 tác giả đạt ngưỡng.

Bảng 2.1 cho thấy danh sách các tác giả đầu tiên làm việc trong lĩnh vực DF [14].

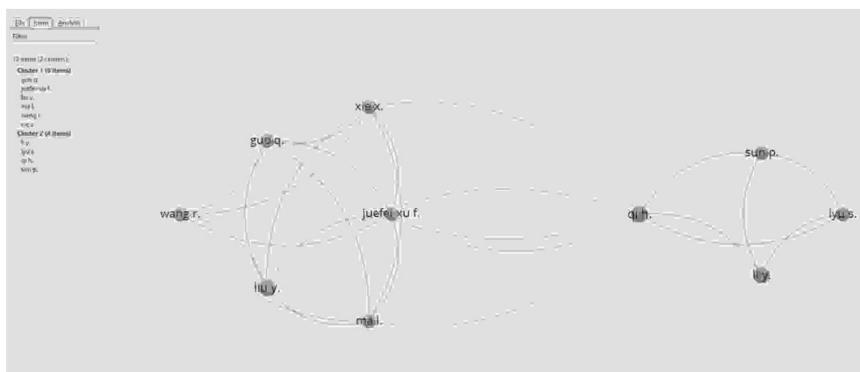
Điều tương tự cũng được thể hiện trong Hình 2.4, cung cấp mười tác giả được kết nối. Hình này được tạo bằng cách sử dụng phần mềm VOSviewer [15].

Chúng tạo thành hai cụm quan trọng. Cụm 1, màu đỏ, có sáu tác giả: Rui Wang, Yang Liu, Lei Ma, Felix Juefei-Xu, Xiaofei Xie, và Xi Gou. Cụm 2, màu xanh lục, có bốn tác giả: Pu Sun, Hua Qi, Yuezun Li và Siwei Lyu.

BẢNG 2.1

Đồng tác giả cho các tác giả

tác giả	Các tài liệu	trích dẫn	Sức mạnh liên kết
Honggang Qi	4	68	14
Yuezun Li	4	67	9
Jiarui	4		0
Liu Yan	4	3	15
Liu Siwei	3	11	9
Lyu Pu	3	65	9
Sun Ruben	3	65	3
Tolosana	3	40	3
Julian Fierrez	3	40	6
Oliver Giudice	3	23	6
Luca Guarnera	3	23	6
Sebastiano	3	23	0
Battiatto	3	14	15
Simson S. Woo	3	11	15
Qing	3	11	15
Guo Felix	3	11	15
Juefei-Xu Lei	3	11	6
Ma Xiaofei Xie	3	11	6
Akash	3		10
Chintha Matthew	3	11	6
Wright Rui	3		0
Wang Raymond	3	9	0
Ptucha Zahid Akhtar Saifuddin Ahmed Jan Kietmann	11 8 9 6		0



HÌNH 2.4 Đồng tác giả phân tích tác giả của mười tác giả được kết nối.

2.3.3 Đồng tác giả của Tổ chức

Đồng tác giả của phân tích tổ chức là đồng tác giả của các tác giả thuộc tổ chức nào làm việc song song với các tác giả của tổ chức kia. Giới hạn số lượng tài liệu tối thiểu của một tổ chức được quy định là 3 và số lượng trích dẫn tối thiểu của một tổ chức cũng được quy định là 3; trong số 326, chỉ có 5 người có thể đáp ứng ngưỡng. Bảng 2.2 cho thấy danh sách các tổ chức.

2.3.4 Đồng tác giả của các quốc gia

Bảng 2.3 mô tả danh sách các quốc gia. Giới hạn số tài liệu tối thiểu mà một quốc gia có được đặt là 3, và số lượng trích dẫn tối thiểu cũng được giới hạn là 3; trong số 51 quốc gia, chỉ có 16 quốc gia có thể đáp ứng ngưỡng.

BẢNG 2.2

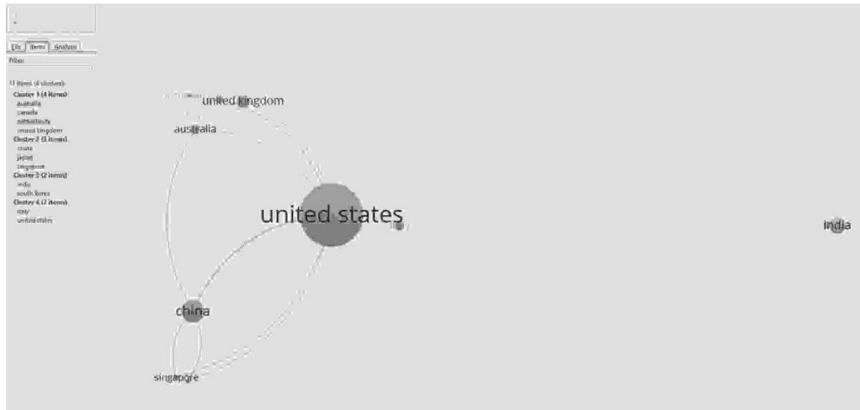
Đồng tác giả cho các tổ chức

tổ chức	Tài liệu	Trích dẫn	Độ mạnh liên kết
Đại học Học viện Khoa học Trung Quốc, Trung Quốc	4	65	0
Đại học Michigan, Hoa Kỳ	3	49	0
Tập đoàn Alibaba, San Mateo, Hoa Kỳ	3	11	3
Đại học Kyushu, Fukuoka, Nhật Bản	3	11	3
Đại học Công nghệ Nanyang, Singapore	3	9	0

BẢNG 2.3

Đồng tác giả cho các quốc gia

Quốc gia	Các tài liệu	trích dẫn	Sức mạnh liên kết
Hoa Kỳ	70	269	26
Trung Quốc	25	91	15
Ấn Độ	18	27	1
Vương quốc Anh	14	78	6
Châu Úc	11	21	4
Canada	9	36	10
Nước Ý	10	88	1
Hàn Quốc		35	2
Tây Ban Nha	8 7	47	0
Liên Bang Nga	6		0
Singapore	6	9	9
nước Đức	5	20	0
nước Hà Lan	4	155 11	3
Nhật Bản	3	11	9
Ireland	3		0
Na Uy	3	3 5	0



HÌNH 2.5 Phân tích đồng tác giả đối với các quốc gia của 11 quốc gia được kết nối.

Trong số 16 quốc gia, chỉ có 11 quốc gia được kết nối, như trong Hình 2.5.

2.3.5 Sự xuất hiện đồng thời của các từ khóa

Từ khóa có nghĩa là một từ hoặc khái niệm có ý nghĩa; ở đây, các tác giả cung cấp phân tích từ khóa cùng xuất hiện dựa trên số lần tối thiểu một từ khóa xuất hiện, được đặt là 10. Trong số 1263 từ khóa, chỉ có 19 từ khóa có thể đáp ứng ngưỡng; tương tự được thể hiện trong Bảng 2.4. Theo VOSviewer, cường độ liên kết là một giá trị số dương của mỗi liên kết. Một kết nối mạnh mẽ có nghĩa là một giá trị cao hơn.

Các từ khóa xuất hiện ít nhất mười lần trong tập dữ liệu được hiển thị trong Bảng 2.4; tương tự được mô tả trong Hình 2.6, tạo thành ba cụm từ khóa quan trọng.

2.3.6 Liên kết thư mục của các tác giả

Khớp nối thư mục được sử dụng trong tất cả các lĩnh vực. Nó giúp các nhà nghiên cứu tìm thấy các công việc liên quan của các nghiên cứu trước đây. Khi hai tài liệu đề cập đến một tác phẩm tiêu biểu thứ ba trong các tài liệu tham khảo của chúng, thì có khả năng tồn tại hai phần chia sẻ cùng một miền. Ở đây, Bảng 2.5 có một danh sách các tác giả được tìm thấy trong danh sách tài liệu tham khảo. Số lượng tài liệu tối thiểu mà một tác giả có được giới hạn ở 3 và số lượng trích dẫn tối thiểu mà một tác giả có cũng được giới hạn ở 3.

Trong số 582 tác giả, chỉ có 23 tác giả đạt ngưỡng. Hình 2.7 cũng vậy; 23 tác giả có ba cụm quan trọng.

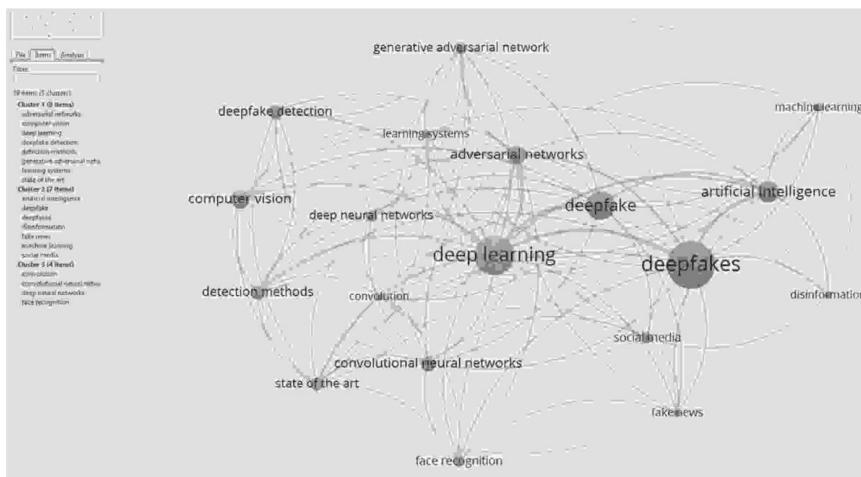
2.3.7 Danh sách các tạp chí tiêu biểu

Danh sách tạp chí nổi bật nhất dựa trên số lượng ấn phẩm. Danh sách này được lập bằng phần mềm Biblioshiny. Ở đây, các tác giả có tình chỉ đề cập đến tên của các tạp chí có nhiều hơn một bài báo được xuất bản trong lĩnh vực DF. Bảng 2.6 cho thấy danh sách 16 tạp chí xuất bản nhiều hơn một bài báo nghiên cứu trong lĩnh vực DF.

BẢNG 2.4

Sự xuất hiện đồng thời của từ khóa

từ khóa	lần xuất hiện	Tổng sức mạnh liên kết
DeepFake	68	82
Học Kì càng	57	144
giả sâu	41	80
Trí tuệ nhân tạo	31	59
Mạng đối thủ	26	79
Mạng thần kinh tích chập	22	47
Tầm nhìn máy tính	27	48
Phát hiện DF	21	41
Phương pháp phát hiện	21	52
Nhà nước của nghệ thuật	18	49
Nhận dạng khuôn mặt	15	32
Truyền thông xã hội	15	26
Mạng thần kinh sâu	16	33
GAN	13	39
máy học	12	23
Tin giả	11	29
Hệ thống học tập	12	36
tích chập	10	31
thông tin sai lệch	10	19

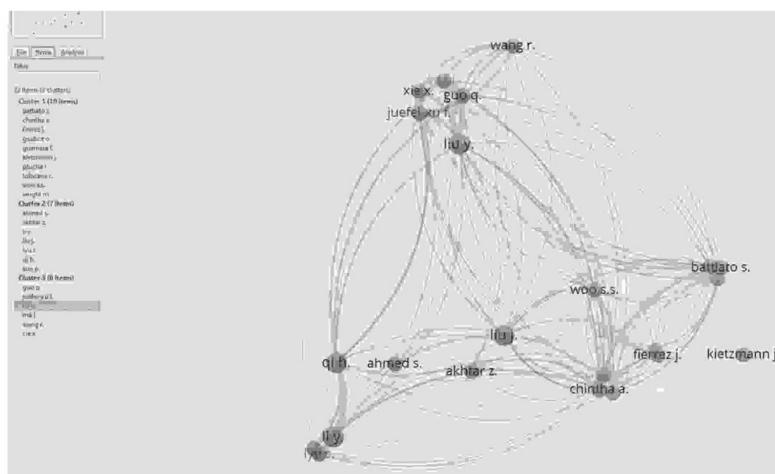


HÌNH 2.6 Sự xuất hiện đồng thời của các từ khóa.

BẢNG 2.5

Khớp nối thư mục của các tác giả

tác giả	Các tài liệu	trích dẫn	Tổng sức mạnh liên kết
Cụm 1 (10)			
Ruben Tolosana	3	40	331
Julian Fierrez	3	40	331
Sebastiano Battiato	3	23	710
Oliver Giudice	3	23	710
Luca Guarnera	3	23	710
Simson S. Woo	3	14	451
Matthew Wright	3	11	1104
Raymond Ptucha	3	11	1104
Akash Chinthia	3	11	1104
Jan Kiëtmann	3	6	22
Cụm 2 (7)			
Hồng Cương Tè	4	68	1023
Nhạc Tôn Lý	4	67	551
Gia Thụy Lưu	4	3	436
Tư Vệ Lyu	3	65	482
Pu Sun	3	65	482
Saifuddin Ahmed	3	9	14
Zahid Akhtar	3	11	196
Cụm 3 (6)			
Dương Liễu	4	11	1620
Felix Juefei-Xu	3	11	1471
Thanh Quách	3	11	1471
Lôi Mã	3	11	1471
Tiêu Phi Tạ	3	11	1471
Thụy Vương	3	9	1011



HÌNH 2.7 Khớp nối thư mục của các tác giả.

BÀNG 2.6

Danh sách các tạp chí xuất bản nhiều hơn một bài báo

Tên tạp chí	Bài báo đã xuất bản
hội tụ	9
Nghiên cứu Khoa học	7
Truy cập IEEE	4
Hành vi tâm lý học mạng và mạng xã hội	3
Tạp chí IEEE về các chủ đề được chọn trong xử lý tín hiệu	3
Tạp chí quốc tế về các xu hướng tiên bộ trong khoa học máy tính và Kỹ thuật	2
Tạp chí Báo chí/Chính trị Quốc tế	2
Dân IT chuyên nghiệp	2
Tạp chí Truyền thông Trực quan và Biểu diễn Hình ảnh	2
Tạp chí hình ảnh	2
Tạp chí Truyền thông Trực quan và Biểu diễn Hình ảnh	2
Truyền thông và Truyền thông	2
Triết học đạo đức và chính trị	2
Truyền thông và xã hội mới	2
Triết học và Công nghệ	2
Nga trong các vấn đề toàn cầu	2

Tạp chí hội tụ được ghi nhận là nổi bật nhất, đã xuất bản chính bài báo về lĩnh vực DF.

2.3.8 Danh sách Hội nghị Tiêu biểu

Danh sách các hội nghị lớn được lập giống như danh sách các tạp chí ở đây. Ngoài ra, trong Bảng 2.7, chỉ những hội nghị đó đã xuất bản nhiều hơn một bài báo hội nghị và chỉ có 15 bài báo có thể tạo nên danh sách.

2.4 THẢO LUẬN

Phần này của chương cung cấp một cuộc thảo luận ngắn gọn về việc phân tích các lĩnh vực nói trên. Các xu hướng xuất bản cung cấp cho độc giả một cái nhìn tổng quan về tài liệu được thực hiện trong DF. Năm 2018 đã đánh dấu sự khởi đầu của DF chỉ với hai án phẩm, tiếp theo là 18 án phẩm vào năm 2019. Năm 2020 chứng kiến sự bùng nổ với 98 án phẩm do đại dịch COVID-19 [16,17,18,19,20]. Đồng tác giả để phân tích tác giả cung cấp cho các tác giả đã đóng góp đáng kể cho một bài báo. Điều tương tự cũng xảy ra với đồng tác giả của các quốc gia và tổ chức. Sự xuất hiện đồng thời của các từ khóa cho biết số lần một từ khóa xuất hiện trong tập dữ liệu. Tại đây, từ khóa "DF" xuất hiện nhiều nhất 65 lần. Khớp nối thư mục của các tác giả để cập đến sự chồng chéo trong danh sách tài liệu tham khảo cho án phẩm. Hai tài liệu được kết hợp theo thư mục khi một án phẩm thứ ba đã trích dẫn cả hai. Tạp chí hội tụ là tạp chí nổi bật nhất, đã xuất bản chính bài báo về

BẢNG 2.7

Danh sách hội nghị xuất bản nhiều hơn một tài liệu

tên hội nghị	Các tài liệu Được phát hành
Hội nghị Hiệp hội Máy tính IEEE về Hình ảnh và Hình mẫu Máy tính	..
Hội thảo công nhận	
MM 2020 - Ký yếu của Hội nghị Quốc tế ACM lần thứ 28 về đa phương tiện	5
Ký yếu - Nhận dạng mẫu hình ảnh ứng dụng	5
Ký yếu hội thảo CEUR	5
IS và T Hội nghị chuyên đề quốc tế về khoa học hình ảnh điện tử và Công nghệ	4
Ký yếu của Hội nghị Hiệp hội Máy tính IEEE về Tầm nhìn Máy tính và Nhận dạng mẫu	3
Chuỗi Ký yếu Hội nghị Quốc tế ACM	2
Hội nghị Yêu tố con người trong Hệ thống máy tính-Ký yếu	2
ICASSP IEEE Hội nghị quốc tế về âm thanh và tín hiệu xử lý-ký yếu	2
Ký yếu - Hội nghị IEEE mùa đông 2021 về ứng dụng máy tính Tầm nhìn WAVV 2021	2
Ký yếu của Hội nghị chung quốc tế về mạng lưới thần kinh	2
Ký yếu - Hội nghị quốc tế lần thứ 5 về phương pháp tính toán và Truyền thông ICCMC 2021	2
Ký yếu Hội nghị Quốc tế lần thứ 4 về IoT trong Di động Xã hội Phân tích và đám mây ISMAC 2020	2
Ký yếu của Hội nghị Châu Âu về Tác động của AI và Người máy Éclair 2020	2
Web Conference 2021 - Ký yếu của World Wide Web Conference Www 2021	2

miền của DF, tiếp theo là Nghiên cứu Khiêu dâm và Truy cập IEEE [21,22,23]. Các hội nghị có ảnh hưởng là Hội nghị IEEE Computer Society về Thị giác Máy tính và Hội thảo Nhận dạng Mẫu, MM 2020 - Ký yếu của Hội nghị Quốc tế ACM lần thứ 28 về Đa phương tiện, Ký yếu - Nhận dạng Mẫu Hình ảnh Ứng dụng và Ký yếu Hội thảo CEUR, xuất bản năm tài liệu.

2.5 TÓM TẮT

Trong kết luận của bài phân tích trước đó, có thể nói rằng mặc dù DFs mới bắt đầu ra đời từ năm 2019, nhưng nó đã được ứng dụng trong nhiều lĩnh vực như Khoa học Máy tính, Kỹ thuật, Khoa học Xã hội, Nghệ thuật, Nhân văn, Khoa học Quyết định, Tâm lý học, Vật lý và Thiên văn học [24-30]. DF vẫn là một khái niệm đang phát triển và các quốc gia trên toàn thế giới, dù đã phát triển hay đang phát triển, đều đang sử dụng nó; do đó, những người tìm kiếm lại có thể thu được kiến thức về ứng dụng của DF, lưu ý đến các vấn đề về quyền riêng tư. Cũng giống như hai mặt của một đồng xu, sử dụng DF trong kinh doanh sẽ có những mặt lợi.

và khuyết điểm. Vì vậy, các nhà kinh doanh nên có một phân tích chi tiết và tiến hành một nghiên cứu thử nghiệm để hiểu được cảm nhận và sự tin tưởng của khách hàng đối với DF.

Chương tiếp theo sẽ thảo luận về các thuật toán được sử dụng để tạo DF.

NGƯỜI GIỚI THIỆU

- [1] Kietzmann, J., Mills, AJ và Plangger, K. (2021). DFs: Quan điểm về "Thực tế" trong tương lai của Quảng cáo và Thương hiệu. Tạp chí Quảng cáo Quốc tế, 40(3), 473–485. doi:10.1080/02650487.2020.1834211
- [2] Vaccari, C. và Chadwick, A. (2020). DF và thông tin sai lệch: Khám phá tác động của video chính trị tổng hợp đối với sự lừa dối, sự không chắc chắn và niềm tin vào tin tức. Truyền thông Xã hội và Xã hội, 6 (1). doi:10.1177/2056305120903408
- [3] Korshunov, P. và Marcel, S. (2019). Đánh giá lỗi hỏng và phát hiện các video DF. Hội nghị quốc tế về sinh trắc học năm 2019. doi:10.1109/ICB45273.2019.8987375
- [4] Kumar, P., Vatsa, M. và Singh, R. (2020). Phát hiện tái hiện khuôn mặt Face2Face trong video. Kỷ yếu - Hội nghị IEEE mùa đông 2020 về ứng dụng thị giác máy tính, WACV 2020. doi:10.1109/WACV45572.2020.9093628
- [5] Xu, Z., Liu, J., Lu, W., Xu, B., Zhao, X., Li, B. và Huang, J. (2021). Phát hiện các video được thao tác trên khuôn mặt dựa trên các mạng thần kinh chuyển đổi tập hợp. Tạp chí Biểu diễn Hình ảnh Giao tiếp Trực quan, 77 doi:10.1016/j.jvcir.2021.103119
- [6] Huang, Y., Juefei-Xu, F., Guo, Q., Xie, X. & Ma, L., Miao, W., Liu, Y., và Pu, G. (2020). FakeRetouch: Tránh phát hiện DeepFakes thông qua Hướng dẫn về tiếng ồn có chủ ý, 70.
- [7] Hernandez-Ortega, J., Tolosana, R., Fierrez, J., và Morales, A. (2021). DFSON Phys: Phát hiện DF dựa trên ước tính nhịp tim. Kỷ yếu Hội thảo CEUR.
- [8] Ahmed, S. (2021). Bị lừa bởi hàng giả: Sự khác biệt về nhận thức trong nhận thức về độ chính xác của yêu cầu bồi thường và chia sẻ ý định của các DF phi chính trị. Sự khác biệt về Tính cách và Cá nhân, 182(111074). doi:10.1016/j.pay.2021.111074
- [9] Liberati, A., Altman, DG, Tetzlaff, J., Mulrow, C., Gotzsche, PC, Ioannidis, JPA, Clarke, M., Devereaux, PJ, Kleijnen, J., và Moher, D. (2009). Tuyên bố Prisma để báo cáo Đánh giá hệ thống và phân tích tổng hợp các nghiên cứu đánh giá các can thiệp chăm sóc sức khỏe: Giải thích và soạn thảo. BMJ, 339.
- [10] Rana, J., Gaur, L., Singh, G., Awan, U., và Rasheed, MI (2021). Củng cố Hành trình của Khách hàng Thông qua Trí tuệ Nhân tạo: Chương trình Nghiên cứu và Đánh giá. Tạp chí quốc tế về các thị trường mới nổi, Tập. trước khi in (No. trước khi in). <https://doi.org/10.1108/IJOEM-08-2021-1214>
- [11] Pavis, M. (2021). Tái cân bằng Phản hồi theo quy định của chúng tôi đối với DF với Quyền của Người biểu diễn. Hội tụ, 27(4), 974–998. doi:10.1177/13548565211033418
- [12] Lees, D., Bashford-Rogers, T., và Keppel-Palmer, M. (2021). Sự hồi sinh kỹ thuật số của Margaret Thatcher: Những khó khăn về sáng tạo, công nghệ và pháp lý trong việc sử dụng DF trong phim truyền hình. Hội tụ, 27(4), 954–973. doi:10.1177/13548565211030452
- [13] Bonomi, M., Pasquini, C., và Boato, G. (2021). Phân tích kết cấu động để phát hiện khuôn mặt giả trong chuỗi video. Tạp chí Truyền thông Trực quan và Biểu diễn Hình ảnh, 79 . doi:10.1016/j.jvcir.2021.103239
- [14] Jung, H., Green, A., Morales, J., Silva, M., Martinez, B., Cattaneo, A., Yang, Y., Park, G., McClean, J., và Mascareñas , D. (2021). Giao thức bảo mật vật lý-mạng toàn diện để xác thực nguồn gốc và tính toàn vẹn của cấu trúc

Đánh giá có hệ thống và phân tích thư mục của DeepFakes

Dữ liệu hình ảnh theo dõi sức khỏe. Giám sát Sức khỏe Kết cấu, 20(4), 1657-1674.
doi:10.1177/1475921720927323

- [15] Xu, Z., Liu, J., Lu, W., Xu, B., Zhao, X., Li, B. và Huang, J. (2021). Phát hiện các video được thao tác trên khuôn mặt dựa trên các mạng thần kinh chuyển đổi tập hợp. Tạp chí Biểu diễn Hình ảnh Truyền thông Trực quan, 77. doi:10.1016/j.jvcir.2021.103119
- [16] Singh, G., Kumar, B., Gaur, L. và Tyagi, A. (2019). "So sánh giữa Multinomial và Bernoulli Naïve Bayes để phân loại văn bản," Hội nghị quốc tế về tự động hóa, tính toán và quản lý công nghệ (ICACTM), trang 593-596. doi:10.1109/ICACTM.2019.8776800
- [17] Gaur, L., Singh, G., Solanki, A., Jhanjhi, NZ, Bhatia, U., Sharma, S., . và Kim, W. (2021). Bố trí của thanh niên trong việc dự đoán các mục tiêu phát triển bền vững bằng cách sử dụng các thuật toán rừng ngẫu nhiên và thần kinh mở. Khoa học thông tin và máy tính lấy con người làm trung tâm, 11, NA.
- [18] Sharma, S., Singh, G., Gaur, L. và Sharma, R. (2022). Khoảng cách tâm lý và tôn giáo có ảnh hưởng đến hành vi lừa đảo của khách hàng không? Tạp chí Nghiên cứu Người tiêu dùng Quốc tế. doi:10.1111/ijcs.12773
- [19] Sahu, G., Gaur, L., và Singh, G. (2021). Áp dụng phương pháp tiếp cận lý thuyết thích hợp và hàm lồng để kiểm tra niềm đam mê của người dùng đối với các nền tảng vượt trội và truyền hình thông thường. Viễn thông và Tin học, 65. doi:10.1016/j.tele.2021.101713
- [20] Gaur, L., Ujjjan, RMA và Hussain, M., 2022. Ảnh hưởng của học sâu trong việc phát hiện các cuộc tấn công mạng vào các ứng dụng chính phủ điện tử. Trong Các biện pháp an ninh mạng cho khung chính phủ điện tử (trang 107-122). IGI toàn cầu.
- [21] Ramakrishnan, R., Gaur, L., và Singh, G. (2016). Tính khả thi và hiệu quả của các thiết bị BLE Beacon IoT trong quản lý hàng tồn kho tại Cửa hàng. Tạp chí Quốc tế về Kỹ thuật Điện và Máy tính, 6(5), 2362-2368. doi:10.11591/ijece.v6i5.10807
- [22] Afaq, A., Gaur, L., Singh, G., và Dhir, A. (2021). COVID-19: Thay đổi hành vi của hành khách đi máy bay và định hình lại kỳ vọng của họ đối với ngành hàng không. Nghiên cứu giải trí du lịch. doi:10.1080/02508281.2021.2008211
- [23] Gaur, L., Bhatia, U., Jhanjhi, NZ, Muhammad, G. và Masud, M., (2021). Phát hiện COVID-19 dựa trên hình ảnh y tế bằng cách sử dụng mạng nơ-ron tích chập sâu. Hệ thống đa phương tiện, trang 1-10.
- [24] Anshu, K., Gaur, L. và Singh, G., (2022). Tác động của trải nghiệm khách hàng đối với thái độ và ý định mua lại trong bán lẻ hàng tạp hóa trực tuyến: Cơ chế kiểm duyệt đồng sáng tạo giá trị. Tạp chí Bán lẻ và Dịch vụ Tiêu dùng, 64, tr. 102798.
- [25] Mahbub, MK, Biswas, M., Gaur, L., Alenezi, F. và Santosh, KC, (2022). Các tính năng chuyên sâu để phát hiện các bất thường về phổi khi chụp X-quang ngực do bệnh truyền nhiễmX: Covid-19, Viêm phổi và Lao. Khoa học Thông tin, 592, trang 389-401.
- [26] Gaur, L., Afaq, A., Solanki, A., Singh, G., Sharma, S., Jhanjhi, NZ, My, HT and Le, DN, (2021). Tận dụng Dữ liệu lớn và Công nghệ 5G mang tính cách mạng: Trích xuất và trực quan hóa Xếp hạng và Đánh giá về Chuỗi Khách sạn Toàn cầu. Máy tính & Kỹ thuật điện, 95, tr. 107374.
- [27] Anshu, K., Gaur, L. và Solanki, A., (2021). Tác động của Chatbot trong việc thay đổi bộ mặt bán lẻ - Một mô hình thực nghiệm về tiền đề và kết quả. Những tiền đề gần đây trong khoa học máy tính và truyền thông (Trước đây: Bằng sáng chế gần đây về khoa học máy tính), 14(3), trang 774-787.
- [28] Kaswan, KS, Gaur, L., Dhamterwal, JS và Kumar, R., (2021). Xử lý ngôn ngữ tự nhiên dựa trên AI để tạo ra thông tin có ý nghĩa Sức khỏe điện tử

- Ghi lại dữ liệu (của cô ấy). Trong Kỹ thuật AI nâng cao và ứng dụng trong tin sinh học (trang 41-86). Nhà xuất bản CRC.
- [29] Gaur, L., Jhanjhi, NZ, Bakshi, S. và Gupta, P., (2022), Tháng Hai. Phân tích Hậu quả của Tri tuệ Nhân tạo đối với Công việc bằng cách sử dụng Lập mô hình Chủ đề và Trích xuất Từ khóa. Vào năm 2022, Hội nghị Quốc tế lần thứ 2 về Thực tiễn Đổi mới trong Công nghệ và Quản lý (ICIPTM) (Tập 2, trang 435-440). IEEE.
- [30] Bùi Tốt, L., (2022). Internet vạn vật trong chăm sóc sức khỏe. Trong Khoa học dữ liệu không gian địa lý ở Chăm sóc sức khỏe cho xã hội 5.0 trang 131-140. Springer, Singapore.

3 Học kĩ càng Kỹ thuật tạo DeepFakes

Bò tót Loveleen, Gursimar Kaur Arora, và
Noor Zaman Jhanjhi

NỘI DUNG

3.1 Giới thiệu	23
3.2 Hàng giả giá rẻ so với DeepFakes	24
3.2.1 Mô hình hóa sâu	24
3.2.2 Bộ mã hóa tự động	24
3.2.3 Mạng đổi thủ chung	26
3.3 Ứng dụng/Phần mềm/Chương trình tạo DeepFake	26
3.4 Đi sâu vào các bài báo liên quan để tạo phương tiện tổng hợp	26
3.4.1 GAN	26
3.4.2 Hoán đổi khuôn mặt	28
3.4.3 Âm thanh	28
3.4.4 Ánh động	29
3.5 Tóm tắt	31
Người giới thiệu	31

3.1 GIỚI THIỆU

Sự tiến bộ và cải tiến mới trong Trí tuệ nhân tạo (AI) đã khai sinh ra DF. Sự tăng trưởng theo cấp số nhân này đã chứng kiến nhiều chức năng phức tạp được thực hiện bằng một kỹ thuật duy nhất, đặc biệt là trong Machine Learning (ML). kỹ thuật ML dài trong rặng; có thể tạo nội dung hiện đại, ngoài các chức năng chung như dự đoán. Các thuật toán để tạo phương tiện tổng hợp sử dụng các thuật toán của DL. DL, một tập hợp con của ML, hoạt động dựa trên khái niệm mạng thần kinh thuật toán học tập không giám sát, còn được gọi là mạng thần kinh nhân tạo (ANN).

Một mạng thần kinh hoạt động giống như các tế bào thần kinh của bộ não của chúng ta. Chúng nổi như cồn vào những năm 1980, nhưng do thiếu dữ liệu và sức mạnh xử lý, chúng không thể được thực hiện cho đến những phát triển gần đây. Tương tự như cách một sợi trực chuyền thông điệp đến các nơron khác trong khi vai trò của dây đeo gai là thu thập thông tin đầu vào từ các nơron, mạng nơron có một quy trình phức tạp với nhiều lớp đơn vị được kết nối với nhau. Các lớp được kết nối thông qua các khớp thần kinh. Mỗi đơn vị có một trọng số nhất định.

Perceptron chia sẻ tín hiệu cho chức năng kích hoạt. Chức năng kích hoạt giúp xác định các mẫu trong dữ liệu ghép kenne để đưa ra đầu ra chính xác. Nó quyết định

yếu tố về thông tin nào sẽ được chuyển đến tế bào thần kinh tiếp theo. Loại mạng thần kinh khác là mạng thần kinh tích chập (CNN), có ứng dụng rộng rãi trong thị giác máy tính để xử lý và phát hiện hình ảnh và có vai trò đáng kể trong việc tạo DF, sẽ được thảo luận trong phần tiếp theo.

Hai chức năng chính được coi là quan trọng nhất trong mạng lưới thần kinh là mắng mỉa và trọng lượng. Trong DF, mạng thần kinh tự chấm điểm bằng cách so sánh đầu ra được tạo với đầu vào và cập nhật trọng số theo điểm số. Các mạng thần kinh kiểm tra xem các trọng số có được thêm vào đúng giám đốc hay không. Họ sẽ kiểm tra điều này thông qua đầu ra được tạo sau khi điều chỉnh trọng số. Quá trình này tiếp tục cho đến khi giá trị tổng thắt giảm đáng kể. Nó không ngừng cải thiện bản thân, một kỹ thuật siêu học. Kỹ thuật tự học này sẽ giúp tạo và tự đánh giá đầu ra của hàng giả.

Phong trào tiền bộ trong công nghệ đã qua lại. Tiền bộ trong mô hình tổng quát để phát hiện DF cũng sẽ mở đường cho các phương pháp tạo tốt hơn. Các đánh giá và nghiên cứu gần đây cho thấy rằng DF là một từ khác để chỉ vấn đề nghiêm trọng, một triệu chứng bắt buộc của các bệnh hiện có [1]. Tuy nhiên, điểm khác biệt duy nhất ở đây là triệu chứng nặng hơn bệnh. Nhưng điều gì sẽ xảy ra nếu điều này thúc đẩy những vết nứt đó được lắp đầy và trở thành lực lượng thay đổi [2] . Do đó, chương này sẽ khám phá nghiên cứu đã thực hiện và các kỹ thuật được sử dụng để tạo DF như Mạng đối thủ chung (GAN), Bộ mã hóa tự động, Hoán đổi khuôn mặt, Hoạt ảnh hình ảnh, v.v..

3.2 GIÁ RẺ VS. DEEPFAKES

Mặc dù cả hai đều có thuật ngữ chung là kỹ thuật thao tác nghe nhìn, nhưng Hàng nhái thường bị lẫn lộn với DF. Hàng giả giá rẻ được sản xuất "rẻ tiền", trong đó giá rẻ có nghĩa là không có công nghệ tiên tiến nào được sử dụng cho thê hệ của nó.

Cheapfakes đang đe dọa nhiều hơn. Nó có thể được tạo ra một cách nhanh chóng, không cần card đồ họa cao cấp bộ dữ liệu lớn, không tốn chi phí [3].

Cheapfakes có thể được sản xuất bằng Photoshop, bất kỳ phần mềm, ứng dụng nào (ví dụ: Snapchat) hoặc đơn giản như giảm hoặc tăng tốc độ thông qua các tính năng chỉnh sửa có sẵn cho máy ảnh và video. DF được tạo thông qua công nghệ AI và ML. Một kỹ thuật thao tác khác là rotoscoping kỹ thuật số, giống như DF, nhưng yêu cầu phác thảo thủ công các tính năng [4].

3.2.1 Mô hình sâu

Bộ xương để tạo DF bắt đầu với Autoencoder (bộ mã hóa-giải mã), một loại CNN và Mạng đối thủ chung (GAN). Phần này sẽ giới thiệu về các kỹ thuật này và giải thích công việc của họ.

3.2.2 Bộ mã hóa tự động

Autoencoder được đặt tên như vậy, vì nó lấy đầu vào, nén nó xuống bộ mã hóa và kết quả đầu ra là đầu vào được tạo lại. Để hiểu rõ hơn về bộ mã hóa tự động, hãy tưởng tượng một kệ dành cho luồng biến thể và dữ liệu hép ở giữa và rộng ở hai đầu. Hình ảnh, giả sử Hình ảnh A ở đầu bên trái của bộ mã hóa, là đầu vào.

Các thuật toán khác nhau có các lớp ẩn khác nhau đang được xử lý. Các

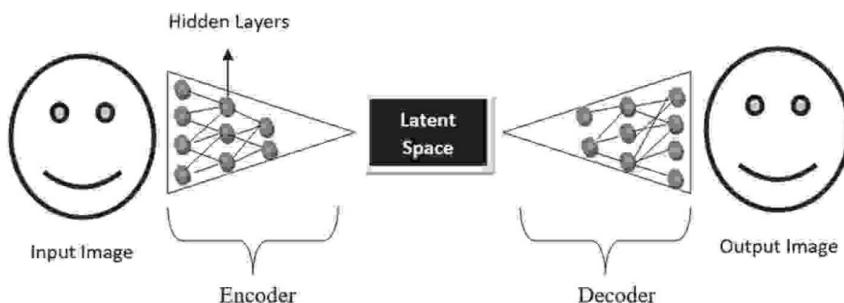
lớp ẩn đầu tiên sẽ xem xét các mảng pixel, thay vì từng pixel một, ví dụ: một mảng 3×3 , để xác định các tính năng cần thiết. Lớp ẩn tiếp theo sẽ tuân theo quy trình tương tự, dựa trên ma trận pixel, được gọi là cửa sổ, được tạo bởi lớp trước đó [5,6,7,8,9].

Tuy nhiên, mạng này trở nên sâu sắc hơn, vì có nhiều tờ biến thể tùy thuộc vào các tính năng và mỗi tính năng tạo thành một lớp. Nhiều tính năng dẫn đến nhiều lớp, được gọi là lớp tích chập. Sau đó, điều này được đơn giản hóa hơn nữa thông qua phương pháp tổng hợp. Trong phương pháp này, người ta đã đạt đến lớp sâu nhất, nơi các tính năng đã được trích xuất thành công. Nó cung cấp cho CNN thuộc tính bất biến về mặt không gian, vì nó không tiếp nhận vị trí của đối tượng. Khi hình ảnh đến lớp tổng hợp, không có thông tin về các tính năng liên quan đến hình ảnh đầu vào và kích thước pixel ban đầu bị mất.

Xem xét ví dụ trước đó về bộ mã hóa tự động như một kênh, từ không gian rộng hơn thông qua các lớp ẩn khác nhau, nén và trích xuất tính năng đến không gian hẹp ở giữa, tức là không gian tiềm ẩn (Hình 3.1). Nó chứa tất cả các thông tin cần thiết để mô hình tái tạo lại đầu vào thông qua bộ giải mã. Chẳng hạn, nếu hình ảnh đầu vào là một khuôn mặt, thì không gian tiềm ẩn sẽ có thông tin liên quan đến mắt, nháy hay mờ, biểu cảm của khuôn mặt, v.v. Bộ giải mã xây dựng hình ảnh đầu ra với các đặc điểm được xác định từ không gian tiềm ẩn. Bộ mã hóa tự động có phương pháp không giám sát, vì nó so sánh đầu ra với đầu vào và thực hiện các sửa đổi giữa các lớp cho đến khi hình ảnh đầu vào tương tự hoặc gần giống như gương được hình thành.

Trong DF, hoạt động của bộ mã hóa tự động thêm một bước vì có một bộ mã hóa và hai bộ giải mã. Trong trường hợp này, đầu vào sẽ là hai hình ảnh, chẳng hạn như Hình ảnh 1 và Hình ảnh 2. Mục đích của việc có một bộ mã hóa cho hai hình ảnh là để giữ lại các tính năng thiết yếu của cả hai hình ảnh trong không gian tiềm ẩn, nút cỗ chai.

Mục đích của nút cỗ chai là cung cấp cho mô hình khả năng tạo lại thay vì chỉ trả về cùng một giá trị như đầu vào. Hai bộ giải mã được đào tạo riêng biệt để xác định các tính năng tiêu chuẩn trong cả hai khuôn mặt. Bộ giải mã 1 sẽ học cách tái tạo lại khuôn mặt của Hình ảnh 1 từ đầu vào nhiều. Sau khi hoàn thành giai đoạn huấn luyện, giai đoạn thử nghiệm sẽ cho kết quả là Ánh 1 có đặc điểm của Ánh 2 và Ánh 2 có đặc điểm của Ánh 1, một kỹ thuật tráo đổi khuôn mặt. Kỹ thuật này không giới hạn chỉ tổng hợp hình ảnh. Với việc sử dụng thuật toán thích hợp, có thể tổng hợp hình ảnh thành video, trong đó đầu ra sẽ là hình ảnh được chuyển thành video, với



HÌNH 3.1 Hoạt động của bộ mã hóa tự động.

các tính năng và chuyển động của người trong video cũng được kết hợp trong người trong hình ảnh và tổng hợp từ video sang video, với những hạn chế nhất định. Để đạt được điều ra thực tế và chính xác, mô hình cần được đào tạo trên một tập dữ liệu lớn và mô hình càng phức tạp thì kết quả càng tốt.

3.2.3 Mạng đối thủ chung

Mặc dù bộ mã hóa tự động dễ truy cập hơn và dễ tính toán hơn, nhưng GAN đã tạo ra kết quả thực tế hơn. Trong GAN, có hai mạng nơ-ron, một mạng là “Đối thủ”, mặt khác, để tạo ra đầu ra tổng hợp. Nó được giới thiệu trong bài báo “General Adversarial Network” của Ian J. Goodfellow et al. [10], một ý tưởng năng động được đề xuất trong lịch sử của ML. Cái tên GAN bắt nguồn từ hai kỹ thuật cơ bản, phương pháp học sâu “Tạo ra”, để tự tạo ra một thứ gì đó hoàn toàn mới, đây sẽ là kết quả tham chiếu cho kỹ thuật “Đối thủ” kiểm tra và so sánh đầu ra được tạo với đầu vào. Hai mạng thần kinh được gọi là bộ tạo và bộ phân biệt. Hệ thống phân biệt đối xử hoạt động trên nguyên tắc xác suất. Hoạt động tương tự như hoạt động của con lắc.

Xét máy phát và máy phân biệt ở hai đầu đối diện của con lắc.

Trình tạo tạo ra một hình ảnh mới cố gắng lừa người phân biệt đối xử chuyển phương tiện tổng hợp thành phương tiện thật. Tuy nhiên, mạng phân biệt nhằm mục đích ước tính xem hình ảnh là do máy tính tạo ra hay là thật. Do đó, theo cách này, quá trình di chuyển qua lại, với mục đích của trình tạo là để cho bộ phân biệt mắc lỗi trong việc thiết lập xác suất. Một bộ phân biệt lý tưởng sẽ khiến trình tạo cực kỳ khó tạo ra hình ảnh chân thực, giúp tạo ra đầu ra tốt hơn.

3.3 ỨNG DỤNG/PHẦN MỀM/CHƯƠNG TRÌNH TẠO DEEPFAKE

Bảng 3.1 trình bày tổng quan về ứng dụng, phần mềm và chương trình hiện tại có sẵn để sử dụng (miễn phí và trả phí) để tạo các loại DF khác nhau. Nó cung cấp một loạt các mô hình tùy thuộc vào tập dữ liệu, nhóm đào tạo và sức mạnh của card đồ họa.

3.4 ĐI SÀU VÀO CÁC GIẤY TỜ LIÊN QUAN ĐẾN TẠO

PHƯƠNG TIỆN TỔNG HỢP

3.4.1 GAN

GAN là phương pháp phổ biến nhất để tạo phương tiện tổng hợp vì nó mở đường cho nhiều kỹ thuật phức tạp khác. Ruthotto và Haber [11] cung cấp cái nhìn sâu sắc về các nguyên tắc toán học liên quan đến Mô hình tạo sâu (DGM), mô hình này giám sát kiến trúc cơ bản của GAN và bộ mã hóa tự động. Để xác định các khả năng cho DGM, ba kỹ thuật chính được sử dụng là GAN, Bộ mã hóa tự động biến đổi [12] và Chuẩn hóa luồng. Natsume và cộng sự. [13] đã sử dụng RSGAN, sử dụng hai bộ mã hóa tự động đa dạng để tổng hợp các đặc điểm khuôn mặt và tóc ở không gian tiềm ẩn. GAN là cách tiếp cận phổ biến nhất để tạo DF [14], yêu cầu tập dữ liệu lớn và card đồ họa có sức mạnh xử lý cao.

BẢNG 3.1**Các loại mô hình Deep Fakes khác nhau**

Tên	người mẫu	Trả phí/Miễn phí	Loại DF
web DF	Mô hình dựa trên hoán đổi khuôn mặt và giá trị tần số	Tiền	Hoán đổi khuôn mặt
DeepFaceLab	H64, Hình đại diện, SAE	Miễn phí	Hoán đổi khuôn mặt
mô tả	mô hình dựa trên ML	Đóng thử miễn phí/trả phí	Lời nói thành văn bản/thính lực đồ
trùng lặp	GAN/Tái tạo	Miễn phí	Hoán đổi khuôn mặt, Giới tính tráo đổi
Hiệu ứng nghệ thuật sâu GAN		Miễn phí và trả tiền phiên bản	Truyền thông vào nghệ thuật
Nỗi Nhớ Sâu Thầm by đi săn của tôi	Video lái xe được cài đặt sẵn cho hình ảnh động	Đóng thử miễn phí/trả phí Phiên bản	Hình ảnh động
Đổi mặt với nó	GAN	Miễn phí	Video hoán đổi khuôn mặt
Hoán đổi khuôn mặt	Nhẹ, Dfaker, Không cân bằng, DFL-H128, DFL-SAE, Dlight, Realface	Miễn phí	Hoán đổi khuôn mặt
Face2Face	mô hình dựa trên ML	Miễn phí	khuôn mặt thời gian thực tái hiện
FaceApp	GAN	Miễn phí và trả tiền phiên bản	Thao tác hình ảnh
FSGAN (Khuôn mặt Tráo đổi GAN)	Mô hình dựa trên RNN và Mắt pha trộn Poisson	Miễn phí	Hoán đổi khuôn mặt và tái hiện
Phong cách ảnh đã tạo GAN		Miễn phí	tạo ảnh
Morphin	ánh xạ hình ảnh	Miễn phí	Lập bản đồ khuôn mặt trong phương tiện gif
quá tài	Mô tả/mô hình dựa trên AI	Đóng thử miễn phí/trả phí Phiên bản	Giọng nói chuyển văn bản thành giọng nói nhân bản / Thính lực đồ
đổi mặt	Nhúng khuôn mặt dựa trên mô hình Miễn phí và trả phí phiên bản		Hoán đổi khuôn mặt người nổi tiếng
phản ánh bởi NeoContext	mô hình dựa trên ML	Miễn phí	Hoán đổi khuôn mặt
tacotron 2	Trình tự lập lại để dự đoán tính năng trình tự, SóngNet	Miễn phí	Tổng hợp giọng nói từ chữ
điệu nhảy	Video lái xe được cài đặt sẵn và ghi lại chuyển động, lập bản đồ hình ảnh	Cả hai đều miễn phí & Tiền phiên bản	Hát nhép
SóngNet	Mô hình dựa trên CNN	Miễn phí	Lời nói từ văn bản
Tảo	Mô hình dựa trên AI	Miễn phí	Video hoán đổi khuôn mặt

Một kỹ thuật khác để hạn chế trở ngại của bộ dữ liệu không lồ đã được nghiên cứu bởi Zakharov et al. [15] bằng cách đào tạo thuật toán đầu tiên (đào tạo meta) với một tập dữ liệu lớn.

Sau đó, thuật toán có thể tạo DF đầu tiên chỉ với một vài đầu vào. Lưu và cộng sự. [16] cũng khám phá các kỹ thuật khác nhau để ổn định GAN, chẳng hạn như độ dốc ngẫu nhiên giảm dần hoặc tăng dần, v.v. Một hạn chế khác đối với việc tạo hình ảnh thông qua GAN là độ dốc biến mất. Để khắc phục điều này, Mao et al. [17] đã đề xuất một mô hình GAN bình phương nhỏ nhất, trong đó bộ phân biệt chiếm hòn đảo bình phương nhỏ. Để tạo phương tiện tổng hợp, Cycle GAN sử dụng hai mạng GAN. Một khung duy nhất, được gọi là MaskGAN, được giới thiệu bởi Lee et al. [18], mang lại cho các lập trình viên sự tự do và thuận lợi hơn trong việc tạo DF để tạo kiểu sao chép và chuyển giao thuộc tính.

3.4.2 Hoán đổi khuôn mặt

Kỹ thuật hoán đổi khuôn mặt [19], mặc dù hiệu quả, nhưng cũng có những hạn chế. Một trong những cách để giải mã điều này là thông qua chuyển kiều thần kinh bằng cách sử dụng CNN đã được đào tạo. Thông qua GAN dựa trên kiều dáng và chuyển giao cấu trúc [20], các tác giả từ mô hình 3D tạo ra hình ảnh 2D bằng cách ánh xạ trên một bề mặt ổn định. Để nhận được đầu ra thực tế hơn cho phương pháp truyền kiếu, Korshunova et al. [21] đã pha chế một chức năng mắt tiên tiến bằng cách giải quyết những hạn chế trong công việc nghiên cứu trước đây bằng cách cải thiện kỹ thuật hoán đổi khuôn mặt. Một trở ngại khác đối với các thuật toán DL mới này, chẳng hạn như GAN và bộ mã hóa tự động biến thể, là tạo ra một DF trông chân thực với quần áo và nền phù hợp. Một giải pháp khả thi có thể là đào tạo GAN về kỹ thuật chuẩn hóa quang phổ [22], tạo ra một hình ảnh mới ở tư thế và nền mong muốn. Và để khắc phục những vấn đề gặp phải trong quá trình chuyển kiều cho kỹ thuật hoán đổi khuôn mặt, Guo et al. [23] đã sử dụng kỹ thuật mã hóa tự động được cập nhật. Natsume và cộng sự. [24] đã sử dụng kiến trúc FSNet để tạo ra kết quả hoán đổi khuôn mặt.

Bản dịch từ hình ảnh sang hình ảnh [25] cũ hơn DF nhưng đã tạo thành công phương tiện tổng hợp. Phương pháp tạo phương tiện này dựa trên các mạng đối nghịch có điều kiện [26], nghĩa là các điều kiện của hình ảnh đầu ra dựa trên hình ảnh đầu vào. Kỹ thuật này hơi khác so với hoạt động thông thường của GAN.

Đầu vào cho bộ phân biệt là ảnh nguồn và bộ phân biệt phải xác định xem ảnh đích có được tạo từ ảnh nguồn hay không. Một phương pháp cập nhật đã được giới thiệu bởi Lombardi et al. [27] để tiếp nhận và mã hóa các đặc điểm khuôn mặt và thể hiện chúng trong thời gian thực, trong khi Nirkin et al. [28] mạng tích chập để thay thế và phân đoạn khuôn mặt, bằng cách mã hóa riêng các đặc điểm khuôn mặt của ảnh nguồn và ảnh đích, trong không gian tiềm ẩn.

3.4.3 Âm thanh

Shimba và cộng sự. [29] đã tạo ra những cái đầu biết nói, tức là tạo video từ âm thanh tổng hợp bằng cách sử dụng mô hình hồi quy và đào tạo nó bằng Bộ nhớ ngắn hạn dài (LSTM), trong khi Santha [30] đã sử dụng GAN dựa trên LSTM để tạo DF. Trong Suwajanakorn và cộng sự. [31], các tác giả đã sử dụng âm thanh của Obama để tạo video có chất lượng tối ưu với tính năng hát nhép hoàn hảo để thêm vào video mục tiêu, sử dụng mạng thần kinh thuỷ lập lại (RNN). Mô hình đã học các ghi chú âm thanh và chuyển động của miệng với mọi khung hình đi qua bằng cách tổng hợp phần dưới của khuôn mặt. Phương pháp này đã được cập nhật và âm thanh thành video được tạo trong thời gian thực bằng Autoencoder và

Kỹ thuật học sâu để tạo DeepFakes

kỹ thuật CNN. Một CNN khác [6,7,8,9] được sử dụng để giữ chất lượng của video bằng cách duy trì các tính năng sắc nét [32].

Tương tự, Prajwal et al. [33] đã tạo ra một mô hình với một người phân biệt được đào tạo tập trung vào hát nhép; đầu ra, sử dụng đoạn âm thanh, là một video hát nhép hoàn hảo. Bát nhã et al. [34] đã tạo một video khuôn mặt bằng cách lấy đầu vào là hình ảnh và âm thanh và tổng hợp LipGAN. Trong khuôn khổ này, vai trò của bộ phân biệt đổi xử lý là xem liệu khuôn mặt và âm thanh có đồng bộ hóa hay không. Một hệ thống khác được giới thiệu bởi Arik et al. [35] sử dụng hai phương pháp để tổng hợp và học hỏi từ một giọng nói nhất định, chỉ yêu cầu một vài mẫu âm thanh. Hai cách tiếp cận được sử dụng là mã hóa người nói và sử dụng người nói. Đó là một phương pháp nhân bản giọng nói thần kinh. Một phô khán khác để khám phá trong tổng hợp âm thanh là tổng hợp giọng nói thành văn bản, thông qua Tacotron 2 [36], có kiến trúc dựa trên mô hình cập nhật của WaveNet [37] và kỹ thuật dựa trên dự đoán tính năng lặp lại. Đầu ra là một bài phát biểu giống như con người, trong đó mô hình có thể được đào tạo trực tiếp từ dữ liệu. Sau phương pháp nhân bản giọng nói [38] là tổng hợp giọng hát [39], bằng cách thu thập chất liệu để tạo ra một nguyên mẫu đa loa. Các tác giả nhằm mục đích tạo ra một mô hình có thể đào tạo trên tập dữ liệu nhỏ hơn và điều chỉnh theo giọng nói mới. Giọng nói thu được có chất lượng âm thanh tối ưu so với giọng thật.

3.4.4 Ảnh động

Tái hiện khuôn mặt là một loại DF khuôn mặt khác, ngoài hoán đổi khuôn mặt, trong đó có ảnh nguồn và ảnh đích, mục đích là chuyển biến cảm của khuôn mặt trong ảnh nguồn sang ảnh đích. Trong phương pháp tái hiện khuôn mặt [40,41] này, nhóm tác giả đã lấy một video mục tiêu và một video nguồn ghi trực tiếp qua webcam.

Hình thành cơ sở tạo DF, Suwajanakorn et al. [42] đã tạo phương tiện tổng hợp cho nghệ sĩ múa rối, trong đó video của Người B có thể điều khiển hình ảnh của Người A. Một tiến bộ năng động khác trong việc tạo DF của Kim et al. [43] đã sử dụng video đầu vào để tạo hiệu ứng lại cho video dọc. Hạn chế của phương pháp này là thao tác chỉ giới hạn ở nét mặt. Tuy nhiên, các tác giả đã tạo DF với chuyển động toàn bộ đầu mà không có bất kỳ thay đổi nào đối với nền hoặc đặc điểm của một người, chẳng hạn như tóc/cơ thể.

Với sự tiến bộ trong nghiên cứu và công nghệ [44,45,5], các tác giả của Siarohin et al. [46] đã tạo DF thông qua hình ảnh động. Họ đã sử dụng hình ảnh nguồn và video lái xe. Do đó, hai loại kết quả đầu ra được phát triển. Đầu tiên, khuôn mặt của hình ảnh nguồn sẽ ổn định dựa trên vị trí điểm chính tương đối, nhưng chuyển động của mắt và miệng sẽ giống như video lái xe. Một hình khác có khuôn mặt được hợp nhất của hình ảnh đầu vào và video nguồn và khuôn mặt đó đang chuyển động giống như người trong video. Đó là một vị trí quan trọng, nhưng đầu ra, trong trường hợp này, hơi bị biến dạng. Balakrishnan et al. [47] đã sử dụng mạng lưới thần kinh tổng quát để điều khiển hình ảnh ở những tư thế không nhìn thấy được. Kết quả có cùng một nền nhưng chỉ là một tư thế khác. Một công cụ phân biệt đổi nghịch đã được sử dụng để thấy trước rằng một hình ảnh thực tế sẽ được tái tạo. Caporaso [48] đã đề xuất một khung cho ứng dụng cho quy trình ba chiều (chẳng hạn như thu thập dữ liệu, giai đoạn xử lý và cuối cùng là tạo nội dung tổng hợp) để giúp người dùng dễ tiếp cận hơn. Nó được sử dụng để thu thập dữ liệu từ một người ở các giai đoạn khác nhau trong cuộc đời của họ và thêm dữ liệu đó vào mô hình ML để đạt được một mô hình hiệu quả nhằm tạo ra Digital Twins.

Để tóm tắt nghiên cứu được thực hiện bởi cộng đồng AI, Bảng 3.2 cung cấp tổng quan về quan điểm của nghiên cứu đã xuất bản, các kỹ thuật được sử dụng và loại DF được tạo ra.

BẢNG 3.2

Việc áp dụng mô hình DeepFakes bởi các tác giả khác nhau (So sánh theo năm)

Tác giả	Kỹ thuật/Mô hình	năm DF
Metri và Mamatha [25]	GAN Lanham [19]	2021 Dịch từ hình ảnh sang hình ảnh
	GAN	2021 Hoán đổi khuôn mặt
Liu và cộng sự. [16]	GAN	2021 Tổng hợp hình ảnh và video
Kurupathi và cộng sự. [22]	tình trạng GAN	2020 Tạo ảnh người ở các tư thế khác nhau
Caporusso [48]	Ứng dụng quy trình ba chiều	Bản sao kỹ thuật số 2020
Siarohin et al. [46]	Mô hình chuyển động bậc nhất	2020 Hình ảnh động
Zhao và Chen [38]	nhân bản giọng nói	Hát tổng hợp 2020
Bát nhã et al. [33]	GAN	Hát nhép 2020
Santa [30]	GAN dựa trên LSTM	VDQG 2020
Lee và cộng sự. [15]	mặt nạ GAN	2020 Thao tác hình ảnh khuôn mặt
Bát nhã et al. [34]	Môi GAN	Video khuôn mặt 2019 từ hình ảnh và âm thanh
Blauw và cộng sự. [39]	Mô hình nhiều loa	Hát tổng hợp 2019
Nirkin et al. [14]	FSGAN	Hoán đổi và tái hiện khuôn mặt 2019
Zakharov và cộng sự. [15]	GAN	Mô hình đầu nói thần kinh 2019
Jamaludin và cộng sự. [32]	Bộ mã hóa tự động, CNN	Tổng hợp âm thanh thành video 2019
Balakrishnan et al. [47]	Thần kinh thế hệ mạng	2018 Tái hiện hình ảnh con người trong tư thế không nhìn thấy
Kim et al. [43]	Video chân dung hồi sinh 2018
Shen và cộng sự. [36]	Sóng mạng, Tacotron 2	Tổng hợp chuyển văn bản thành giọng nói 2018
Arik và cộng sự. [35]	Bộ mã hóa loa và sử dụng loa	Nhân bản giọng nói thần kinh 2018
Nirkin et al. [28]	GAN	2018 Phân đoạn khuôn mặt và hoán đổi khuôn mặt
Natsume và cộng sự. [24]	Fsnet	2018 Hoán đổi khuôn mặt dựa trên hình ảnh
Quách và cộng sự. [23]	Chuyển giao phong cách, bộ mã hóa tự động	Hoán đổi khuôn mặt 2018
Lombardi et al. [27]	CNN, Có điều kiện Mạng đối thủ	2018 Thao tác hình ảnh khuôn mặt
Natsume và cộng sự. [13]	RSGAN	Hoán đổi khuôn mặt 2018
Suwajanakorn và cộng sự. [31]	RNN	Âm thanh 2017 để hát nhép
Isola et al. [26]	Đổi thứ có điều kiện Mạng	Dịch từ hình ảnh sang hình ảnh năm 2017
Korshunova và cộng sự. [21]	chuyển kiều	Hoán đổi khuôn mặt 2017
Thies et al. [40]	GAN	Tái hiện gương mặt 2016
Vuong và cộng sự. [20]	Phong cách và cấu trúc GAN dựa trên chuyển giao	Hoán đổi khuôn mặt 2016
Suwajanakorn [42]	bộ mã hóa tự động	Múa rối truyền thông tổng hợp 2015
Thies et al. [41]	GAN	Tái hiện khuôn mặt 2015
Shimba và cộng sự. [29]	LSTM, Mô hình hồi quy	2015 Tạo video từ tổng hợp âm thanh

3.5 TÓM TẮT

DF là phương tiện tổng hợp được tạo thông qua các thuật toán siêu học của ML và DL. Chương này đọc về hai thuật toán tạo thành khung để tạo thành các DF: Autoencoder và GAN. Rất ít ứng dụng, phần mềm và chương trình được sử dụng với các thuật toán được đào tạo dành cho người dùng mới hoặc những người không có nền tảng kỹ thuật.

Nghiên cứu kỹ lưỡng đã được thực hiện trên các bài báo liên quan và các kỹ thuật được sử dụng bởi các nhà nghiên cứu. Các bài viết liên quan được phân chia rộng rãi dựa trên GAN, Hoán đổi khuôn mặt, tổng hợp âm thanh và hoạt ảnh. Nghiên cứu của chúng tôi kết luận rằng vẫn tồn tại hạn chế về việc đào tạo chất lượng mô hình của tập dữ liệu hoặc thời gian cần thiết để đào tạo mô hình. Do đó, nghiên cứu trong tương lai cần tập trung vào các khía cạnh này để tạo ra một thế hệ DF dễ dàng hơn, nhanh hơn và thực tế hơn.

Chương tiếp theo sẽ tập trung vào việc sử dụng các tạo tác công vênh mặt để phân biệt DF video từ người thật một cách hiệu quả.

NGƯỜI GIỚI THIỆU

- [1] Brownlee, J. (2019). Cách phát triển GAN Pix2Pix để dịch từ hình ảnh sang hình ảnh. <https://machinelearningmastery.com/how-to-develop-a-pix2pix-gan-for-image-to-image-translation/>, Sử dụng tích cực DF (Bài viết).
- [2] Kietzmann, J., Lee, LW, McCarthy, IP, & Kietzmann, TC (2020). DFs: Lừa hay đái? Chân trời kinh doanh, 63(2), 135-146. <https://doi.org/10.1016/j.bushor.2019.11.006>
- [3] Loukides, M. (2021, ngày 11 tháng 5). DeepCheapFake. Truyền thông O'Reilly. www.oreilly.com/radar/deepcheapfakes/ (Bài báo).
- [4] Paris, B., & Donovan, J. (2019, ngày 18 tháng 9). DF và hàng giả giá rẻ. Dữ liệu & Xã hội. <https://datasociety.net/library/DFs-and-cheap-fakes/>. Báo cáo của Trường Harvard Kennedy.
- [5] Gaur, L., Afqaq, A., Singh, G. & Dwivedi, YK (2021). Vai trò của trí tuệ nhân tạo và người máy trong việc thúc đẩy du lịch không chạm trong thời kỳ đại dịch: Chương trình nghiên cứu và đánh giá. Tạp chí Quốc tế về Quản lý Khách sạn Đương đại, 33(11), 4079-4098. <https://doi.org/10.1108/IJCHM-11-2020-1246>
- [6] Singh, G., Kumar, B., Gaur, L., & Tyagi, A. (2019). "So sánh giữa Multinomial và Bernoulli Naïve Bayes để phân loại văn bản," Hội nghị quốc tế về tự động hóa, tính toán và quản lý công nghệ (ICACTM), trang 593-596. doi:10.1109/ICACTM.2019.8776800.
- [7] Gaur, L., Singh, G., Solanki, A., Jhanjhi, NZ, Bhatia, U., Sharma, S., . & Kim, W. (2021), "Sứ mệnh của thanh niên trong việc dự đoán sự phát triển bền vững các mục tiêu bằng cách sử dụng thuật toán rừng ngẫu nhiên và thần kinh mờ" Khoa học thông tin và điện toán lấy con người làm trung tâm, 11, NA.
- [8] Rana, J., Gaur, L., Singh, G., Awan, U. & Rasheed, MI (2021). Củng cố hành trình của khách hàng thông qua trí tuệ nhân tạo: Chương trình nghiên cứu và đánh giá. Tạp chí quốc tế về các thị trường mới nổi, Tập. trước khi in (No. trước khi in). <https://doi.org/10.1108/IJOEM-08-2021-1214>
- [9] Sharma, DK, Gaur, L., & Okunbor, D. (2007). Nén ảnh và trích xuất đặc trưng với mạng nơ-ron. Ký yếu của Viện Khoa học Thông tin và Quản lý, 11(1), 33-38.
- [10] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014) . GAN. Truyền thông của ACM, 63(11), 139-144. <https://doi.org/10.1145/3422622>

- [11] Ruthotto, L., & Haber, E. (2021). Giới thiệu về mô hình thế hệ sâu. GAMM-Mitteilungen, 44(2). <https://doi.org/10.1002/gamm.202100008>
- [12] Pinheiro Cinelli, L., Araújo Marins, M., Barros da Silva, EA, & Lima Netto, S. (2021). Bộ mã hóa tự động đa dạng. Các phương pháp đa dạng cho ML với các ứng dụng cho mạng sâu, 111-149. lò xo. https://doi.org/10.1007/978-3-030-70679-1_5
- [13] Natsume, R., Yatagawa, T., & Morishima, S. (2018). "Rsgan: Hoán đổi khuôn mặt và chỉnh sửa bằng cách sử dụng biểu diễn khuôn mặt và tóc trong không gian tiềm ẩn," bản in trước của arXiv arXiv:1804.03447. <https://doi.org/10.1145/3230744.3230818>
- [14] Nirkin, Y., Keller, Y., & Hassner, T. (2019). "FSGAN: Hoán đổi khuôn mặt bắt khả tri và tái hiện chủ đề," trong Kỷ yếu của Hội nghị Quốc tế IEEE về Tầm nhìn Máy tính, trang 7184-7193 <https://doi.org/10.1109/ICCV.2019.000728>
- [15] Zakharov, E., Shysheya, A., Burkov, E., & Lempitsky, V. (2019). Học hỏi về các mô hình cái đầu biết nói thần kinh thực tế. Hội nghị quốc tế IEEE/CVF 2019 về thị giác máy tính (ICCV). <https://doi.org/10.1109/iccv.2019.00955>
- [16] Liu, M.-Y., Huang, X., Yu, J., Wang, T.-C., & Mallya, A. (2021). GAN để tổng hợp hình ảnh và video: Thuật toán và ứng dụng. Ký yếu của IEEE, 1-24. <https://doi.org/10.1109/jproc.2021.3049196>
- [17] Mao, X., Li, Q., Xie H., Lau R., Zhen W., & Smolley, S. (2017). GAN bình phương nhỏ nhất. <https://doi.org/10.1109/ICCV.2017.304>
- [18] Lee, C.-H., Liu, Z., Wu, L., & Luo, P. (2020). Maskgan: Hướng tới thao tác hình ảnh khuôn mặt đa dạng và tích cực. Hội nghị IEEE/CVF 2020 về Thị giác máy tính và Nhận dạng mẫu (CVPR). <https://doi.org/10.1109/cvpr42600.2020.00559>
- [19] Lanham, M. (2021). DF và hoán đổi khuôn mặt. Tạo ra một thực tế mới, 255-285. lò xo. https://doi.org/10.1007/978-1-4842-7092-9_9
- [20] Wang, X., & Gupta, A. (2016). Mô hình hóa hình ảnh tổng quát sử dụng phong cách và cấu trúc mạng lưới đối nghịch. Trong Hội nghị Châu Âu về Tầm nhìn Máy tính, trang 318-335. lò xo. https://doi.org/10.1007/978-3-319-46493-0_20
- [21] Korshunova, I., Shi, W., Dambre, J., & Theis, L. (2017). Hoán đổi khuôn mặt nhanh bằng cách sử dụng mạng thần kinh tích chập. Hội nghị quốc tế IEEE 2017 về thị giác máy tính (ICCV). <https://doi.org/10.1109/iccv.2017.397>
- [22] Kurupathi, S., Murthy, P., & Stricker, D. (2020). Tạo hình ảnh con người với quần áo bằng cách sử dụng GAN có điều kiện nâng cao. Ký yếu Hội nghị Quốc tế lần thứ nhất về Lý thuyết và Ứng dụng DL. <https://doi.org/10.5220/0009832200300041>
- [23] Guo, Y., He, W., Zhu, J., & Li, C. (2018). Mạng tự động mã hóa nhẹ cho ping hoán đổi khuôn mặt. Ký yếu Hội nghị Quốc tế lần thứ 2 về Khoa học Máy tính và AI- CSAI '18. <https://doi.org/10.1145/3297156.3297210>
- [24] Natsume, R., Yatagawa, T., & Morishima, S. (2018). "Fsnet: Một mô hình tổng quát nhận biết danh tính cho hoán đổi khuôn mặt dựa trên hình ảnh," trong Hội nghị Châu Á về Tầm nhìn Máy tính, trang 117-132. lò xo. https://doi.org/10.1007/978-3-030-20876-9_8
- [25] Metri, O., & Mamatha, HR (2021). Tạo hình ảnh bằng GAN. GAN cho dịch từ hình ảnh sang hình ảnh, 235-262. Elsevier. <https://doi.org/10.1016/b978-0-12-823519-5.00007-5>
- [26] Isola, P., Zhu, J.-Y., Zhou, T., & Efros, AA (2017). Dịch từ hình ảnh sang hình ảnh với các mạng đối thủ có điều kiện. Hội nghị IEEE 2017 về Thị giác máy tính và Nhận dạng mẫu (CVPR). <https://doi.org/10.1109/cvpr.2017.632>
- [27] Lombardi, S., Saragih, J., Simon, T., & Sheikh, Y. (2018). Các mô hình xuất hiện sâu để vẽ khuôn mặt. Giao dịch ACM trên Đồ họa, 37(4), 1-13. <https://doi.org/10.1145/3197517.3201401>

- [28] Nirkin, Y., Masi, I., Tuan, AT, Hassner, T., & Medioni, G. (2018). "Về phân đoạn khuôn mặt, hoán đổi khuôn mặt và nhận thức khuôn mặt," trong Hội nghị quốc tế lần thứ 13 của IEEE về Nhận dạng khuôn mặt và cử chỉ tự động (FG 2018), trang 98-105: IEEE. <https://doi.org/10.1109/FG.2018.00024>
- [29] Shimba, T., Sakurai, R., Yamazoe, H., & Lee, J.-H. (2015). Tổng hợp đầu nói từ âm thanh với mạng lưới thần kinh sâu. Hội nghị chuyên đề quốc tế IEEE/SICE 2015 về tích hợp hệ thống (SII). <https://doi.org/10.1109/sii.2015.7404961> FaceSwap GAN: <https://github.com/shaoanlu/faceswap-GAN>
- [30] Santha, A. (2020). Tạo DF bằng cách sử dụng GAN dựa trên LSTM (luận văn). Học viện Công nghệ Rochester. Truy cập từ <https://scholarworks.rit.edu/theses/10447>
- [31] Suwajanakorn, S., Seitz, SM, & Kemelmacher-Shlizerman, I. (2017). Tổng hợp Obama: Học hát nhép từ audio. Giao dịch ACM trên Đồ họa, 36(4), Điều 95. <https://doi.org/10.1145/3072959.3073640>
- [32] Jamaludin, A., Chung, JS, & Zisserman, A. (2019). Bạn đã nói rằng? Tổng hợp khuôn mặt biết nói từ âm thanh. Tạp chí Quốc tế về Tầm nhìn Máy tính, 127(11-12), 1767-1779. <https://doi.org/10.1007/s11263-019-01150-y>
- [33] Prajwal, KR, Mukhopadhyay, R., Namboodiri, VP, & Jawahar, CV (2020). Một chuyên gia hát nhép là tất cả những gì bạn cần để tạo lời nói thành môi trong tự nhiên. Kỷ yếu Hội nghị Quốc tế ACM lần thứ 28 về Đa phương tiện. <https://doi.org/10.1145/3394171.3413532>
- [34] Prajwal, KR, Mukhopadhyay, R., Philip, J., Jha, A., Namboodiri, V., & Jawahar, C. V. (2019). Tiến tới dịch tự động trực diện. Kỷ yếu Hội nghị Quốc tế ACM lần thứ 27 về Đa phương tiện. <https://doi.org/10.1145/3343031.3351066>
- [35] Arik, S., Chen, J., Peng, K., Ping, W., & Zhou, Y. (2018). "Nhân bản giọng nói thần kinh với một vài mẫu," trong Những tiến bộ trong Hệ thống xử lý thông tin thần kinh, trang 10019-10029.
- [36] Shen, J., Pang, R., Weiss, RJ, Schuster, M., Jaitly, N., Yang, Z., . Wu, Y. (2018). Tổng hợp TTS tự nhiên bằng cách điều chỉnh Wavenet trên Dự đoán quang phổ MEL. Hội nghị quốc tế IEEE 2018 về xử lý âm thanh, giọng nói và tín hiệu (ICASSP). doi:10.1109/icassp.2018.8461368
- [37] WaveNet: Một mô hình chung cho âm thanh khô. Tâm trí sâu sắc. (nd). <https://deepmind.com/blog/bài viết/wavenet-generative-model-raw-audio>.
- [38] Zhao, L., & Chen, F. "Nghiên cứu về nhân bản giọng nói với một vài mẫu," 2020 Hội nghị quốc tế về Mạng máy tính, Điện tử và Tự động hóa (ICCNEA), 2020, trang 323-328. <https://doi.org/10.1109/ICCNEA50255.2020.00073>.
- [39] Blaauw, M., Bonada, J., & Daido, R. (2019). Dữ liệu nhân bản giọng nói hiệu quả để tổng hợp giọng hát thần kinh. ICASSP 2019-2019 Hội nghị quốc tế IEEE về xử lý âm thanh, giọng nói và tín hiệu (ICASSP). <https://doi.org/10.1109/icassp.2019.8682656>
- [40] Thies, J., Zollhofer, M., Stamminger, M., Theobalt, C., & Niessner, M. (2016). Face2Face: Chụp và tái hiện khuôn mặt trong thời gian thực của video RGB. Hội nghị IEEE 2016 về Tầm nhìn Máy tính và Nhận dạng Mẫu (CVPR). <https://doi.org/10.1109/cvpr.2016.262>
- [41] Thies, J., Zollhöfer, M., Nießner, M., Valgaerts, L., Stamminger, M., & Theobalt, C. (2015). Chuyển biểu cảm theo thời gian thực để tái hiện khuôn mặt. Giao dịch ACM trên Đồ họa, 34(6), 1-14. <https://doi.org/10.1145/2816795.2818056>
- [42] Suwajanakorn, S., Seitz, SM, & Kemelmacher-Shlizerman, I. (2015). Điều gì khiến Tom Hanks trông giống Tom Hanks. Hội nghị quốc tế về thị giác máy tính (ICCV) năm 2015 của IEEE. <https://doi.org/10.1109/iccv.2015.450>

- [43] Kim, H., Garrido, P., Tewari, A., Xu, W., Thies, J., Nießner, M., Pérez, P., Richardt, C., Zollhöfer, M., & Theobalt , C. (2018). Chân dung video sâu. <https://doi.org/10.1145/3197517.3201283>
- [44] Gaur, L., Afaq, A., Solanki, A., Singh, G., Sharma, S., Jhanjhi, NZ, . Le, D. (2021). Tận dụng dữ liệu lớn và công nghệ 5G mang tính cách mạng: Trích xuất và trực quan hóa xếp hạng cũng như đánh giá về chuỗi khách sạn toàn cầu. Máy tính và Kỹ thuật điện, 95. doi:10.1016/j.compeleceng.2021.107374
- [45] Gaur, L., Bhatia, U., Jhanjhi, NZ, Muhammad, G., & Masud, M. (2021). Phát hiện COVID-19 dựa trên hình ảnh y tế bằng cách sử dụng mạng thần kinh tích chập sâu. Hệ thống đa phương tiện. doi:10.1007/s00530-021-00794-6
- [46] Siarohin, A., Lathuilière, S., Tulyakov, S., Ricci, E., & Sebe, N. (2020). Mô hình chuyển động đặt hàng đầu tiên cho hoạt ảnh hình ảnh. Một phần của Những tiến bộ trong Hệ thống xử lý thông tin thần kinh 32 (NeurIPS 2019).
- [47] Balakrishnan, G., Zhao, A., Dalca, AV, Durand, F., & Guttag, J. (2018). Tổng hợp những hình ảnh người trong tư thế không che. Hội nghị IEEE/CVF 2018 về Thị giác máy tính và Nhận dạng mẫu. <https://doi.org/10.1109/cvpr.2018.00870>
- [48] Caporaso, N. (2020). DFs for The good: Một ứng dụng hữu ích của công nghệ AI đang gây tranh cãi. Những tiến bộ trong Hệ thống Thông minh và Điện toán, 235-241. Springer Hoa Kỳ. https://doi.org/10.1007/978-3-030-51328-3_33

4 Phân tích DeepFakes Video của Face Warping Hiện vật

Ajantha Devi Vairamani

NỘI DUNG

4.1 Giới thiệu	36
4.2 Ảnh hưởng của DFs.....	
4.3 Bộ dữ liệu DeepFakes.....	37
4.3.1	
UADFV	
4.3.2 DeepFakes-TIMIT (DF-TIMIT)	38
4.3.3 FaceForensics ++.....	38
4.3.4 Phát hiện Google DeepFakes-DFD.....	38
Thử thách phát hiện DeepFakes trên Facebook-DFDC	38
4.3.5 Celeb-DF	38
4.4 DF	38
Phát hiện DF.....	39
4.5 Các phương pháp xử lý video dựa trên khuôn mặt.....	41
4.5.1 Các tính năng tạm thời trên các khung	42
4.5.1.1 Sử dụng mạng nơ-ron hồi quy.....	42
mặt	42
quan bên trong khung	43
4.5.2.1 Phân loại sâu	43
4.5.2.1.1 Face Warping Artifacts.....	44
4.5.2.2 Shallow Classifiers.....	45
4.5.2.2.1 Sử dụng các tư thế đầu không nhất quán	45
4.5.2.3 Khó phát hiện	45
DeepFake.....	47
liệu DeepFake.....	48
suất.....	48
dạng thiếu tính hợp lý.....	49
thời	49
mạng xã hội	49
tắt	50
4.6 Khó phát hiện	
4.6.1 Chất lượng của bộ dữ liệu DeepFake.....	
4.6.2 Đánh giá hiệu suất.....	
4.6.3 Các chiến lược nhận dạng	
4.6.4 Tập hợp tạm thời	
4.6.5 Rửa tiền trên mạng xã hội	
4.7 Tóm tắt	



HÌNH 4.1 Ảnh do DF tạo ra (phải) dựa trên một cá nhân (trái).

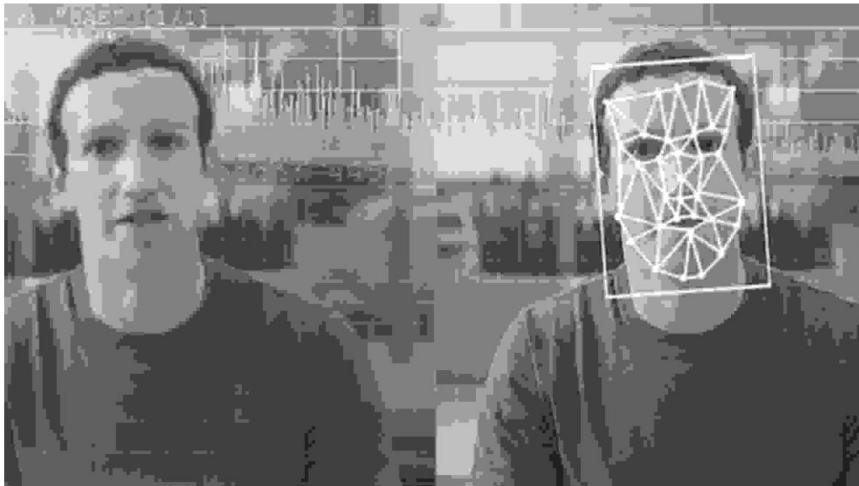
4.1 GIỚI THIỆU

DeepFakes (DF) trộn các từ “Deep learning” và “fake”. DFs là một sự đổi mới áp đặt hình ảnh khuôn mặt của cá nhân khách quan lên video của cá nhân nguồn, như được minh họa trong Hình 4.1, khiến cho cá nhân khách quan có tất cả các đặc điểm nhận biết khi thực hiện các bài tập tương tự như cá nhân nguồn trong video. Các mô hình như Bộ mã hóa tự động và Mạng đối thủ chung (GAN) thường được sử dụng trong thị giác PC để giải quyết các vấn đề về phân chia hình ảnh, nhận dạng khuôn mặt và hợp nhất hình ảnh khuôn mặt nhiều chế độ xem. Tính toán DFs cũng được sử dụng để xem xét cảm xúc và sự phát triển trên khuôn mặt của một cá nhân, giống như sự kết hợp hình ảnh khuôn mặt của người khác, tạo ra các khớp nối và sự phát triển tương đương [1].

Sự đổi mới của DF đã được đề xuất một cách thú vị vào cuối năm 2017. Do hệ thống Bộ mã hóa-Giải mã, sự đổi mới của GAN đã được triển khai để trình bày sự đổi mới của DF. Bằng cách sử dụng nguyên tắc nâng cao của giả thuyết trò chơi, phép tính GAN không chỉ giảm số lượng ranh giới mô hình và độ phức tạp của mô hình trong các điều kiện tương tự mà còn làm cho khuôn mặt được tạo ra trở nên hợp lý một cách đáng kinh ngạc, do đó làm giảm sự phụ thuộc vào bức ảnh đầu tiên và làm việc dựa trên tác động của khuôn mặt đang phát triển, trong bất kỳ trường hợp nào, gây nhầm lẫn giữa khuôn mặt lừa đảo với khuôn mặt xác thực [2].

4.2 TÁC DỤNG CỦA DFS

DFs là cải tiến Trí tuệ nhân tạo (AI) phát triển nhanh nhất [3] đã đạt được danh tiếng về việc sử dụng nó trong việc sản xuất các bản ghi sâu. Ví dụ: một số cá nhân sử dụng sự đổi mới này để tạo nội dung tục tĩu hoặc giả mạo các bài diễn văn chính thức của chính phủ và những thứ khác. Thành thật mà nói, ngoài việc ảnh hưởng đến tính hợp lệ của video, sự đổi mới của DF có thể được sử dụng để tạo ra bằng chứng. Chẳng hạn, những kẻ gian lận có thể tạo ra những bộ phim chuyển động giả mạo liên quan đến trò nghịch ngợm của những người đứng đầu nó để đe dọa và



HÌNH 4.2 Tác động tàn phá của DF đã làm dây lên nhiều lo ngại và đã tấn công bờ biển Ánh Độ (Ảnh: CNBC).

ép buộc hợp tác. Đáng tiếc hơn, sự phô biến của các bản ghi được tạo ra bởi sự đổi mới của DF đã dẫn đến hậu quả to lớn là không ai chấp nhận các bản ghi chính hãng [4,5].

Mặt khác, DF cung cấp một số lợi ích; chẳng hạn như giúp đỡ những người bị mất tiếng khi gây ồn ào hoặc làm mới các đoạn phim mà không cần ghi âm lại [1].

Các cá nhân đang lái xe, cũng như các khung máy tính tinh chỉnh, đang gặp phải vấn đề phát hiện ra các bức ảnh và bản ghi giả như trong Hình 4.2 khi các mạng thần kinh sâu (DNN) tiên tiến được cải thiện và mở ra một lượng thông tin khổng lồ [6]. Ngày nay, việc cung cấp ảnh và phim giả ngày càng trở nên đơn giản hơn và mọi thứ cần thiết là một bức ảnh nhân vật hoặc một đoạn video ngắn từ cá nhân khách quan. Do đó, DF ảnh hưởng đến cả những người lưu ý và những người theo phong tục [7]. Tiếng nói của Giám đốc điều hành công ty mẹ của một công ty năng lượng Anh ở Đức đã lừa đảo nhiều cộng sự và đồng phạm với số tiền 220.000 Euro chỉ trong một ngày [8]. Một tình tiết nữa đã xảy ra khi sử dụng AI để tạo hình ảnh DF và hồ sơ trên LinkedIn, đánh lừa nhiều cá nhân, bao gồm cả các cơ quan chính phủ [9].

4.3 BỘ DỮ LIỆU DEEPFAKES

Do sự lạm dụng của DF và sự phát triển nhanh chóng, việc nghiên cứu các phương pháp phát hiện DF ngày càng trở nên phức tạp. Tính khả dụng của bộ dữ liệu quy mô lớn DF là một thành phần quan trọng trong sự tiến bộ của các thuật toán phát hiện DF.

4.3.1 UDFV

Bộ dữ liệu UADFV [10] là một trong những cơ sở dữ liệu công khai đầu tiên phát hiện DF. Có 49 video YouTube thực trong cơ sở dữ liệu. Những phim này được sử dụng để tạo 49 video gerie cho khuôn mặt mục tiêu bằng thiết bị di động FakeApp.

4.3.2 DeepFakes-TIMIT (DF-TIMIT)

Đại học Queensland (UQ) ở Úc đã tạo bộ dữ liệu video âm thanh Vid-TIMIT [11]. Bộ dữ liệu khác là bộ dữ liệu DFs-TIMIT từ Viện Idiap Thụy Sĩ, được tổ chức bằng bộ dữ liệu Vid-TIMIT [12]. Mỗi đối tượng trong số 43 đối tượng trong cơ sở dữ liệu Vid-TIMIT đã sàng lọc 13 video thực tế. Bộ dữ liệu DF-TIMIT chứa 32 chủ đề và 620 video DF từ bộ dữ liệu Vid-TIMIT. Faceswap-GAN đã được sử dụng để thực hiện các video tổng hợp này. Có 10.537 ảnh thực tế trong tập dữ liệu và 34.023 ảnh tổng hợp được tạo từ 320 video.

4.3.3 Pháp y khuôn mặt ++

FaceForensics ++ [13] là bộ dữ liệu giả mạo khuôn mặt mà các nhà điều tra có thể sử dụng để đào tạo các phương pháp dựa trên học sâu có giám sát. DF, Face2Face, NeuralTextures và FaceSwap là bốn kỹ thuật thao tác khuôn mặt tự động được sử dụng để thay đổi 1.000 chuỗi video bắt đầu. Những số liệu này dựa trên 977 video trên YouTube, tất cả đều có thể theo dõi và chủ yếu là mặt trước không có vỏ bọc, cho phép tạo ra các giả mạo thực tế bằng các phương pháp giả mạo tự động.

4.3.4 Phát hiện Google DeepFake- DFD

Nghiên cứu của Google đã tạo tập dữ liệu Phát hiện DF (DFD) [14] bằng cách quay 100 video trong 28 tình huống khác nhau với các diễn viên tình nguyện và trả phí. Sau đó, bằng cách sử dụng phương pháp tạo DF có thể truy cập miễn phí, hơn 3.000 DF đã được tạo từ những video này. Bộ dữ liệu DFD, bao gồm video tự nhiên và video giả, phát hiện DF.

4.3.5 Thử thách phát hiện DeepFake trên Facebook- DFDC

Facebook đã thu thập và sản xuất bộ dữ liệu DF Detection Challenge (DFDC) [15], bao gồm các video 5K với hai thuật toán sửa đổi khuôn mặt. Facebook thu hút một nhóm các diễn viên chuyên nghiệp để nắm bắt các công nghệ sinh học sâu khác nhau. Mỗi người tham gia phải gửi một bộ phim để thực hiện một bộ nhiệm vụ được xác định trước. Những bộ phim này bao gồm nhiều tình huống ánh sáng, tư thế đứng và sự đa dạng của các trực (giới tính, màu da và tuổi tác).

4.3.6 Người nói tiếng-DF

Bộ dữ liệu Celeb-DF-v1 [16] chứa các video tổng hợp thực và DF có chất lượng tương đương với các video được tìm thấy trên internet. Bộ dữ liệu Celeb-DF bao gồm 408 video thực tế trên YouTube thuộc nhiều giới tính, độ tuổi và chủng tộc khác nhau cũng như 795 DF được tạo từ các bản ghi này.

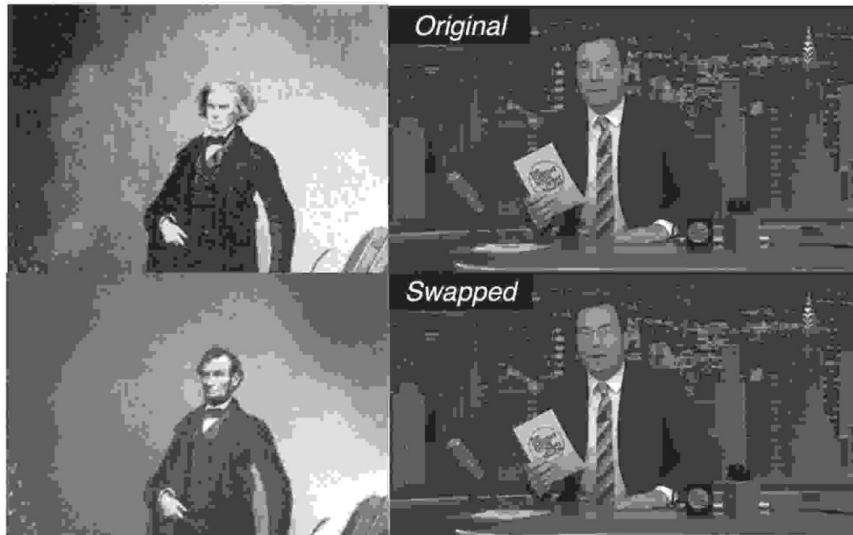
Các video tổng hợp thực tế và DF với chất lượng video có thể so sánh với các video khuếch tán internet được bao gồm trong bộ dữ liệu Celeb-DF-v2 [17]. Bộ dữ liệu Celeb-DF-v2 quan trọng hơn đáng kể so với bộ dữ liệu Celeb-DF-v1 trước đó, chỉ chứa 795 DF. Celeb-DF hiện có 590 video trên YouTube với nhiều sắc tộc, giới tính, độ tuổi và người nói tiếng khác nhau.

4.4 PHÁT HIỆN DFS

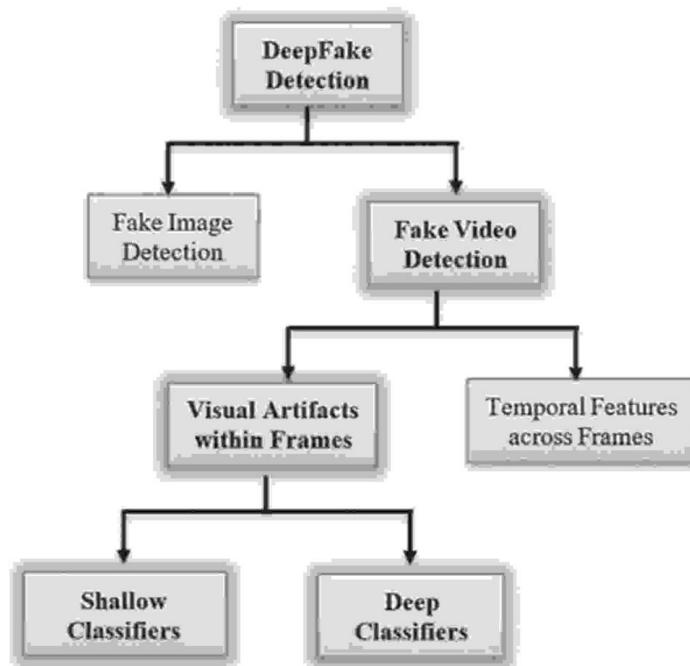
Một trong những tác phẩm nghệ thuật nổi tiếng nhất của Tổng thống Hoa Kỳ Abraham Lincoln, có từ khoảng năm 1865, chứa đựng nỗi lực sớm nhất được ghi nhận trong giao dịch trực tiếp. Như được hiển thị trong Hình 4.3, bản in thạch bản nổi đầu của Abraham Lincoln và cơ thể của nhà tiên phong miền Nam John Calhoun. Sau cái chết của Lincoln, các bản in thạch bản của ông đã trở nên phổ biến, và các bản khắc đầu của ông trên nhiều thi thể khác nhau hầu như chỉ xuất hiện trong thời gian ngắn [18].

Như minh họa trong Hình 4.4, các kỹ thuật phát hiện DF được chia thành hai loại: phát hiện ảnh giả và phát hiện video giả. Lý do dằng sau sự phân nhóm này là do sự suy yếu nghiêm trọng do áp lực video tạo ra, hầu hết các quy trình nhận dạng hình ảnh không thể được sử dụng trực tiếp cho vị trí video. Bên cạnh đó, các bản ghi kết hợp chất lượng lập kế hoạch thay đổi theo các vỏ bọc khác nhau, làm cho các phương pháp nhận dạng hình ảnh tĩnh được thử nghiệm để nhận ra [19].

Các nâng cấp đang diễn ra trong việc thay đổi hình ảnh và video [20,21] đã thay đổi hoàn toàn chiến trường. Việc dân chủ hóa các tiền bối ngày nay đã thúc đẩy sự thay đổi thế giới quan này, chẳng hạn như TensorFlow [22] và Keras [23], cũng như sự thừa nhận cởi mở đối với văn bản chuyên ngành muộn và thiết bị đăng ký chi phí tối thiểu. Họ đang chỉnh sửa ảnh và bản ghi, vốn đã dễ dàng truy cập được đối với các chuyên gia về chủ đề được chuẩn bị đặc biệt như trong Hình 4.5, sử dụng các mô hình mã hóa tự động tích chập [19,24] và GAN [25,26]. Các ứng dụng điện thoại di động và PC phụ thuộc vào phương pháp này bao gồm FaceApp [27] và FakeApp [28].



HÌNH 4.3 Di chuyển khuôn mặt: Quan chức chính phủ John Calhoun được giao dịch với Hoa Kỳ Tổng thống Abraham Lincoln (trái). Đầu của Jimmy Fallon và John Oliver được trao đổi trong FakeApp (phải) [28].



HÌNH 4.4 Phân loại các phương pháp phát hiện DF.



HÌNH 4.5 Hình ảnh được đánh dấu bằng ô chữ nhật màu đỏ là hình ảnh bị mạo.

Phân tích video DeepFakes bằng Face Warping Artifacts

FaceApp là một chương trình cơ giới hóa thời đại chụp ảnh các điều chỉnh quảng cáo trên khuôn mặt thực tế. Bạn có thể thay đổi khuôn mặt, kiểu tóc, tuổi tác, khuynh hướng tình dục và các điều khoản khác bằng điện thoại di động của mình. FaceApp là một ứng dụng cho phép khách hàng tạo các bản ghi "DFs".

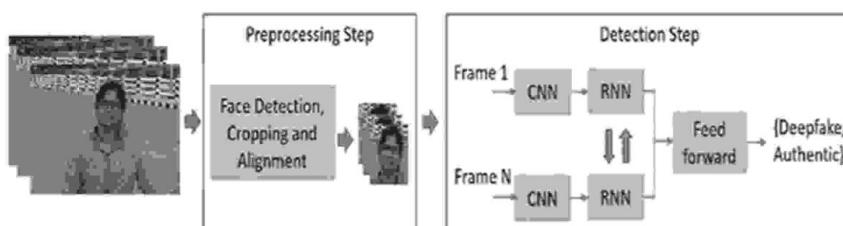
Mặc dù một số video của DF là vô thường vô phạt, nhưng chúng chỉ là thiểu số. Cho đến thời điểm này, các công nghệ video DF [28] đã được sử dụng để làm giả các bản ghi khêu dâm có tên tuổi lớn và nội dung khêu dâm trả thù [29]. Trên các nền tảng như Reddit, Twitter và Pornhub, loại nội dung khêu dâm này hiện đã bị loại trừ. Các bản ghi DF là mục tiêu cho tài liệu tình dục Ped, phim trinh sát giả mạo, tin tức giả mạo và các nội dung khủng khiếp khác do bản chất chính xác của chúng. Các tổ chức chính phủ đang xem xét nghiêm túc những bộ phim giả mạo này, những bộ phim gần đây đã được sử dụng để xoa dịu căng thẳng chính trị [30].

4.5 CÁC PHƯƠNG PHÁP XỬ LÝ VIDEO DỰA TRÊN MẶT

Một vài phương pháp để nhận biết các thay đổi trên khuôn mặt trong các nhóm video đã được phát triển trong những năm 1990 [31,32]. Thies et al. đã nhanh chóng thực hiện các bước di chuyển liên tục cho khuôn mặt. Sau đó, họ đã giới thiệu Face2Face [19], một khung tái hiện khuôn mặt nhất quán có thể thay đổi diễn biến khuôn mặt trên các video khác nhau. Các tùy chọn trái ngược với Face2Face cũng đã được giới thiệu [33]. Các nhà khoa học [34] chỉ ra rằng một số phép tính tuổi hình ảnh khuôn mặt dựa trên học sâu cũng đã được phân tích. GAN đã được sử dụng để làm giả khuôn mặt [26] và sửa đổi các thuộc tính trên khuôn mặt như màu da [35].

Phép nội suy đặc điểm sâu [33] tạo ra kết quả tuyệt vời liên quan đến việc thay đổi các đặc điểm trên khuôn mặt như tuổi tác, sự phát triển của râu và cảm giác miệng. Lample et al. [34] có được những khám phá có thể so sánh được với các bổ sung thuộc tính. Mục tiêu hình ảnh của hầu hết các tính toán kết hợp hình ảnh dựa trên DL là không phô trương. Karras và cộng sự. [35] hiển thị pha trộn khuôn mặt tuyệt vời bằng cách sử dụng GAN vừa phải để nâng cao chất lượng hình ảnh hơn nữa.

Recurrent Neural Networks (RNNs), như trong Hình 4.6, là một loại công việc của mạng nơ-ron tiếp tục lặp lại cùng một thứ. Các mạng Bộ nhớ ngắn hạn dài (LSTM) đã được Hochreiter và Schmidhuber [36] khám phá như một loại RNN [37] để học các điều kiện đường dài trong sắp xếp đầu vào. Các mô hình DL sử dụng cả LSTM và mạng thần kinh tích chập (CNN) được ám chỉ là “ở đâu đó trong không gian” và “ở đâu đó đúng lịch trình”, là hai phương thức làm việc khung khác nhau.



HÌNH 4.6 Phát hiện thao tác khuôn mặt bằng RNN.

4.5.1 Các tính năng tạm thời trên các khung hình

Các tính năng tạm thời trên các đường viền đưa ra các lựa chọn phụ thuộc vào các thuộc tính liên quan đến thời gian trong video, chẳng hạn như tần suất nhấp nháy của con người và hình dạng miệng, thường sử dụng các chiến lược mô tả đặc điểm lặp lại.

4.5.1.1 Sử dụng mạng thần kinh hồi quy

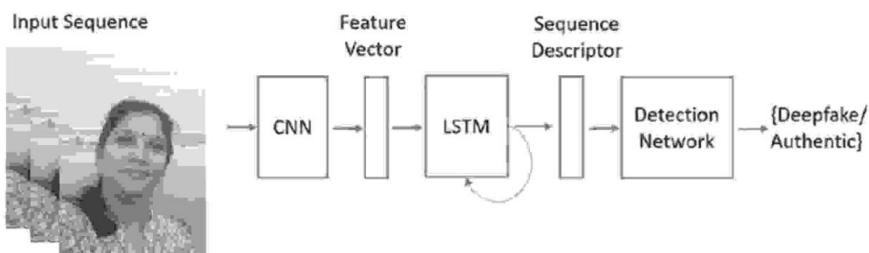
Khung bắt đầu đến kết thúc được khám phá bởi các chuyên gia [38]. Mạng CNN trước đó sử dụng ImageNet để chuẩn bị trước cho mô hình InceptionV3 [39,40] trong chuỗi video nhất định. Tuy nhiên, nó sẽ xóa lớp kết nối cuối cùng để xây dựng 2048-chiều nhãn hiệu vector cho mỗi vỏ bọc. Mạng LSTM lấy vectơ thương hiệu làm thông tin. Xác suất hợp lệ và giả mạo cuối cùng đã được xử lý bằng cách sử dụng softmax, như trong Hình 4.7, sau lớp liên kết 512 chiều.

Các chuyên gia đã thử những cách khác nhau với các đoạn phim có khung cảnh khác nhau sau khi thu thập 300 bản ghi DF từ trang web. Từ kết quả, chúng ta có thể thấy rằng phép tính này có thể dự đoán chính xác liệu một phần bị hỏng có phải từ một video được tạo sâu sắc trong chưa đầy hai giây của video hay không (40 cạnh của video được kiểm tra ở 24 trường hợp mỗi giây) với tốc độ chính xác lên đến 97%. Tuy nhiên, nó có nhược điểm là yêu cầu cả hình ảnh thật và giả để chuẩn bị thông tin, gây lãng phí.

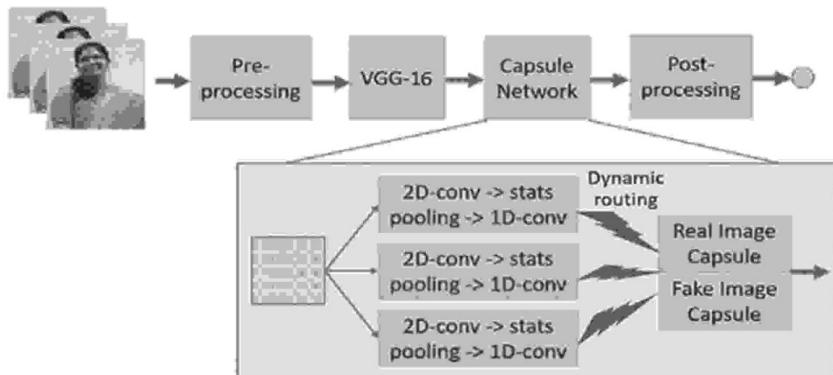
4.5.1.2 Nháy mắt

Sự lặp lại nhấp nháy [41,42] của vị trí khuôn mặt video xác định mí mắt nháy trong các thử nghiệm chuẩn bị phụ thuộc vào video được thiết kế của DF. Để bắt đầu, hãy phân biệt từng cạnh của khuôn mặt, tìm vùng trên khuôn mặt và tập trung vào các chi tiết cơ bản của khuôn mặt, chẳng hạn như đầu mắt, môi, mũi và hình dạng của má. Sau đó, sử dụng chiến lược sắp xếp khuôn mặt dựa trên các móc để điều chỉnh vùng khuôn mặt thành không gian được tổ chức thống nhất, tránh trở ngại do những thay đổi về hướng phát triển của đầu và hướng khuôn mặt trong đường viền video gây ra. Sau đó, tại thời điểm đó, để thiết lập một sự sắp xếp nhất quán, hãy tìm và loại bỏ mắt tự nhiên và gửi nó đến mạng Mạng kết hợp lặp lại dài hạn (LRCN).

Mạng LRCN: để quyết định tình trạng của mắt, trước tiên tập trung các điểm sáng mắt tự nhiên bằng cách sử dụng VGG16 [43], như trong Hình 4.8, sau đó, tại điểm đó, nhập các đơn vị RNN và LSTM, sau đó, tại điểm đó, gửi kết quả tới lớp liên kết hoàn toàn,



HÌNH 4.7 Một phương pháp phát hiện DF sử dụng mạng thần kinh tích chập (CNN).



HÌNH 4.8 Các tính năng sâu thu được từ mạng VGG-16.

trong đó tính ra khả năng mở hoặc nhắm mắt. Cuối cùng, đào tạo mạng CNN, LSTM-RNN và các lớp được liên kết hoàn toàn bằng cách sử dụng riêng công việc bắt hạnh chéo entropy mô tả đặc tính của hai lớp.

Trong một thử nghiệm, các nhà khoa học thay thế kỹ thuật LRCN [44] bằng mạng nhóm hai lớp VGG16 và hệ thống dựa trên tỷ lệ góc mắt (EAR). Sau đó, khi so sánh với CNN 0,98 và EAR 0,79, LRCN có triển lãm tốt nhất (0,99) cho đến quận thuộc ROC (AUC). Cuối cùng, video thực tế có tần số nhấp nháy lặp lại là 34:1/phút. Ngược lại, video sâu có 3,4 lần nheo mắt/phút, vì vậy tôi đặt cạnh lặp lại nhấp nháy của một người bình thường là 10 lần/phút. Chúng tôi có thể xác định xem video này có phải là giả mạo hay không.

4.5.2 Tạo tác trực quan bên trong khung hình

Việc điều tra các đồ cỗ trực quan bên trong vỏ có xu hướng trong phân khúc này, sử dụng các điểm không hoàn hảo trên cạnh của bức tranh cũng như các chi tiết không tự nhiên như các yếu tố trên khuôn mặt và bóng trên khuôn mặt để đánh giá, loại bỏ các phẩm chất rõ ràng và hoàn thành việc khám phá bằng các phân loại sâu hoặc nông. Sau đó, các thành phần này được phân lập thành các bộ phân loại bè ngoài sâu, tách biệt các bản ghi thật và giả.

4.5.2.1 Phân loại sâu

Do mục tiêu của các bản ghi DF liên tục bị hạn chế, nên chúng ta nên sử dụng chiến lược uốn bẻ mặt tương đối (ví dụ: chia tỷ lệ, xoay và cắt). Vì các phần mặt xoắn không phù hợp với khía hậu chung, nên một bóng tối sâu được tạo ra mà mô hình CNN có thể nhận ra. Tiếp theo, chúng ta nên xem các kỹ thuật nhận dạng được kết nối phụ thuộc vào bộ phân loại sâu.

Chẳng hạn, kiểm tra thu nhỏ phụ thuộc vào xung động hình ảnh sẽ không hoạt động trong các trường hợp video được nén trong đó tiếng ồn hình ảnh phải được giảm bớt. Ngoài ra, các bức ảnh mô phỏng [45] khó có thể xác định bằng mắt thường ở mức độ ngữ nghĩa cao hơn, đặc biệt khi một hình ảnh mô tả khuôn mặt người. Do đó, các nhà khoa học đã đề xuất một kỹ thuật chuyển tiếp cho DNN với số lượng lớp được xác định trước.

Mạng Meso-4 có bốn mạng thần kinh rối rắm đang phát triển, mỗi mạng đều có Chuẩn hóa hàng loạt [46] và Tổng hợp tối đa [47]. Cuối cùng, chúng được sắp xếp theo thứ tự bằng cách sử dụng hai lớp liên kết hoàn chỉnh và sigmoid.

MesoInception-4: Với mô-đun nguồn gốc biến thể v1, hai mạng lưới thần kinh rối rắm hoạt động trước khi Meso-4 được đặt lại. Chuyên gia đề cập rằng việc sử dụng mô-đun Inception để thay thế nhiều hơn một vài lớp sẽ không tạo ra kết quả sắp xếp lý tưởng.

Kế hoạch của mô-đun hiện tại này nhằm mục đích chia sẻ chất sản lượng của hai lớp rối với các hình dạng mảnh khác nhau, mở rộng không gian dung lượng có thể truy cập để nâng cao mô hình. Các chuyên gia đã thử khám phá của họ bằng cách sử dụng bộ dữ liệu mà họ đã nghiên cứu về DF và Face2Face. Để tạo ra sự đơn giản hóa quá mức và sự chân thành, các bản vá đầu vào được thực hiện với những thay đổi nhỏ bất thường, bao gồm chia tỷ lệ, cách mạng, lật mức cũng như thay đổi độ chói và bóng. Trong các điều kiện phân tán mạng tự nhiên, tốc độ khám phá bình thường của phương pháp này đối với video DF là 90% và tốc độ nhận dạng bình thường của video Face2Face là 95%.

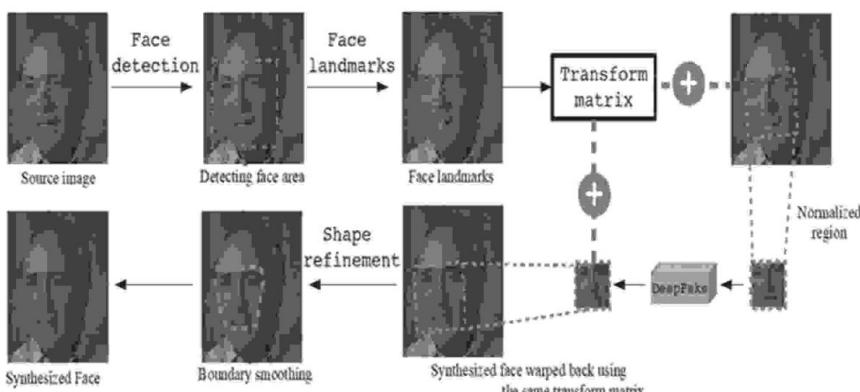
4.5.2.1.1 Tạo tác cong vênh mặt

Phương pháp này phụ thuộc vào các đặc điểm trong Hình 4.9 của video DF [48].

Do không có thời gian tạo và nội dung tính toán, kỹ thuật DF có thể hợp nhất các hình ảnh khuôn mặt có mục tiêu thấp, những hình ảnh này sẽ trải qua một sự thay đổi tương đối để phối hợp với sự sắp xếp khuôn mặt nguồn. Do các mục tiêu của diện tích bề mặt xoắn và các vùng bao quanh nhanh chóng của nó xung đột với nhau, nên sự uốn cong này mang lại các bóng sâu cù thể trong video DF, mà các mô hình DNN thông thường có thể bắt được một cách hiệu quả (chẳng hạn như VGG, ResNet [49], v.v.).

Ưu điểm: Sử dụng các mẫu âm làm dữ liệu huấn luyện là một thao tác xử lý ảnh đơn giản, giúp tiết kiệm thời gian và tài nguyên tính toán.

Nhược điểm: Có thể trang bị quá mức video DF với một bản phân phối cụ thể.



HÌNH 4.9 Tổng quan về quy trình sản xuất DF.

Phân tích video DeepFakes bằng Face Warping Artifacts

4.5.2.2 Bộ phân loại nồng

Các bộ phân loại nồng đòi hỏi một trình trích xuất đặc trưng có khả năng giải quyết vấn đề về tính chọn lọc-bất biến của sự hoàn hảo. Một trình trích xuất tính năng hữu ích có thể tạo ra một tính năng duy nhất; nghĩa là, nó có thể trích xuất thông tin có giá trị để nhận ra nội dung của bức ảnh trong khi loại bỏ dữ liệu không cần thiết chẳng hạn như vị trí của con vật. Phần trước đã đề cập đến các bộ phân loại sâu và phần này sẽ đề cập đến các phương pháp liên quan dựa trên bộ phân loại nồng.

4.5.2.2.1 Sử dụng các tư thế đầu không nhất quán

DF được tạo bằng cách ghép khu vực khuôn mặt đã tạo vào ảnh gốc. Nó sẽ đưa ra các lỗi khi tính toán tư thế đầu ba chiều (chẳng hạn như hướng và vị trí của đầu) từ hình ảnh khuôn mặt hai chiều. Các nhà nghiên cứu sử dụng bộ phân loại Máy vectơ hỗ trợ (SVM) [50] để phân loại tính năng này và thực hiện các thí nghiệm để chứng minh hiện tượng. Các nhà nghiên cứu đã so sánh các tư thế đầu ước tính dựa trên các điểm tọa độ trên toàn bộ khuôn mặt, với các tư thế đầu chỉ đơn giản là ở khu vực trung tâm của khuôn mặt, nhận thấy rằng chúng rất giống nhau trên khuôn mặt thực tế.

Các chuyên gia coi vectơ hướng đầu để cải thiện vấn đề, thu được vectơ đơn vị ba chiều hai đầu được xác định từ toàn bộ mặt và mặt giữa và khoảng cách cosin tương phản.

Khoảng cách cosin của một vài vectơ vị trí đầu được xác định từ các lần xuất hiện tự nhiên được tập trung xung quanh phạm vi tiếp cận hạn chế (tối đa 0,02). Mặt khác, hai giá trị khoảng cách cosine vectơ của DF nằm rải rác trong phạm vi 0,02-0,08, chứng tỏ rằng chúng có thể được phân tách khỏi nhau dọc theo các đường này. Bộ phân loại SVM đã được chuẩn bị bổ sung trong bộ dữ liệu UADDV [51] và bộ dữ liệu DARPA GAN Challenge [52].

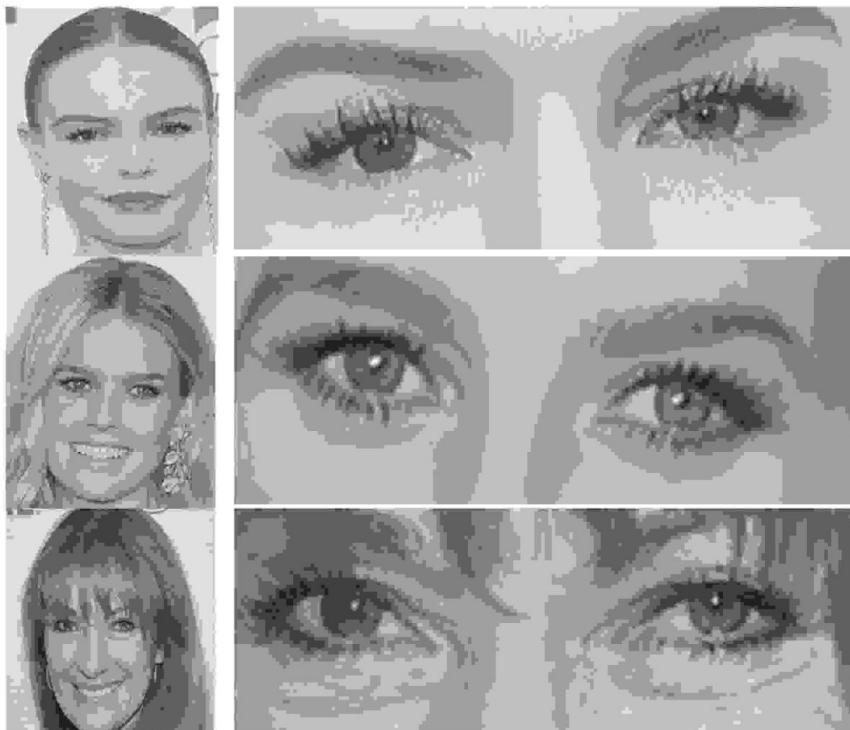
4.5.2.3 DeepFakes và thao tác trên khuôn mặt

Có thể tạo các tạo tác trực quan giả đéo [53,54] bằng cách sử dụng các thuật toán sản xuất video hiện có. Những đặc điểm này có thể nhận thấy bằng cách kiểm tra mắt, rằng và các đường nét trên khuôn mặt. Các nhà nghiên cứu phân loại các hiện vật trực quan được sản xuất này thành các loại sau:

Tính nhất quán toàn cầu: Thuật ngữ "tính nhất quán toàn cầu" đề cập đến sự không nhất quán về màu sắc của móng mắt ở mắt trái và mắt phải. Dị sắc tố khá hiếm trong thực tế, nhưng mức độ xuất hiện này ở khuôn mặt được tạo ra rất khác nhau, như được minh họa trong Hình 4.10.

Ước tính độ sáng: Mặc dù quy trình Face2face [55] sử dụng rõ ràng đánh giá độ sáng, đánh giá toán học và đưa ra các ví dụ để chứng minh, lỗi hoặc đánh giá sai ánh sáng tập có thể sẽ tạo ra các vết tích liên quan trên khuôn mặt do học sâu tạo ra. Ví dụ, đồ cổ này thường tạo ra không gian xung quanh mũi, khiến một bên trở nên buồn tẻ một cách đáng ngạc nhiên. Ngoài ra, hình ảnh phản chiếu trong mắt hoặc bị mất hoặc được sắp xếp lại bằng cách học sâu, như thể hiện trong Hình 4.10 và 4.11.

Ước tính hình học: Đánh giá sai về mặt toán học đối với khuôn mặt người gây ra các giới hạn rõ ràng, giống như đồ cổ giả có độ tương phản cao, xuất hiện trên các giới hạn của khuôn mặt và trang bìa của con người. Bên cạnh đó, ở một mức độ nào đó, phần khuôn mặt bị cản trở, chẳng hạn



HÌNH 4.10 Sai số ước tính độ chiếu sáng và nhát quán toàn cục.



HÌNH 4.11 Lỗi ước lượng độ sáng và ước lượng hình học.



HÌNH 4.12 Sai số ước lượng hình học.

núi một kiểu tóc không phù hợp, sẽ mang lại sự “trống rỗng”. Hơn nữa, răng đôi khi không được khắc họa theo bất kỳ cách nào, như được tìm thấy trong một số bộ phim. Răng xuất hiện trong các phim này dưới dạng các đốm trắng nhỏ thay vì từng răng riêng lẻ, như minh họa trong Hình 4.11 và 4.12.

Bằng cách tách các thành phần này để xây dựng các tập hợp vectơ nổi bật, các chuyên gia tìm cách chuẩn bị KNN [56], MLP [57], mô hình lặp lại chiến lược [58] và các bộ phân loại khác trên toàn bộ khuôn mặt do thông tin GAN, DF và Face2face tạo ra. Vì các điểm nổi bật đặc trưng cho các mặt hàng giả cợ thể, những bộ phân loại nhỏ này cũng có thể thực hiện các nhiệm vụ đặt hàng tùy thuộc vào kết quả. Đây cũng là một lợi ích quan trọng của quy trình này so với các quy trình khác sử dụng bộ phân loại sâu để chuẩn bị thông tin và thời gian, bắt kể kết quả có phải là bất thường hay không.

4.6 KHÓ KHĂN TRONG PHÁT HIỆN DEEPFAKES

Mặc dù tiềm bộ quan trọng đã được thực hiện trong việc trưng bày các thiết bị định vị DF, nhưng có nhiều vấn đề khác nhau với các tính toán nhận dạng hiện tại. Phân đoạn này xem xét một phần các thách thức mà thủ tục xác định vị trí DF gặp phải, như thể hiện trong Hình 4.13.



HÌNH 4.13 So sánh tập dữ liệu.

4.6.1 Chất lượng của bộ dữ liệu DeepFake

Khả năng tiếp cận các cơ sở thông tin quan trọng của DF là một quan điểm quan trọng trong việc cải thiện tính toán vị trí của DF. Tuy nhiên, việc so sánh bản chất của các bản ghi từ các bộ dữ liệu này với nội dung thực sự đã bị thay đổi chiếm đoạt trên web đã phát hiện ra những mâu thuẫn lớn. Bạn có thể tìm thấy những điều gì mà về hình ảnh kèm theo trong các cơ sở thông tin sau:

- Tôi. thế gian lắp lánh trong khi nói chuyện,
- thứ hai. mơ hồ xung quanh các khu vực trên khuôn mặt,
- iii. sự hoàn hảo trên bề mặt/không có sự tinh tế trên bề mặt,
- v.v. sự vắng mặt của người đứng đầu thể hiện sự phát triển hoặc cuộc cách mạng,
- v. không có các mặt hàng sắp xảy ra như kính, sét đánh, và vân vân
- vi. nhạy cảm với những thay đổi về tư thế hoặc dáng vẻ khi nhập liệu, nước da không đều và rò rỉ ký tự
- vii. khả năng truy cập hạn chế của bộ dữ liệu DF đa phương tiện hàng đầu đã tham gia. Sự mơ hồ trong tập dữ liệu được trình bày trước đó là do sai lầm trong phương pháp kiểm soát.

Bên cạnh đó, các chất kém chất lượng đã điều chỉnh có thể khó thuyết phục hoặc thiết lập kết nối được chứng nhận. Do đó, bất kể khung khám phá có đánh bại các bản ghi như vậy hay không, không có gì đảm bảo rằng các kỹ thuật này sẽ hoạt động tốt trong thực tế.

Hơn nữa, nội dung chất lượng thấp đã được sửa đổi có thể khó thuyết phục hoặc tạo ảo tưởng chân thực. Do đó, ngay cả khi các hệ thống phát hiện vượt trội hơn các video như vậy, không có gì đảm bảo rằng các phương pháp này sẽ hoạt động tốt trong thế giới thực.

4.6.2 Đánh giá hiệu suất

Tính toán vị trí DF hiện được thể hiện dưới dạng vấn đề mô tả đặc điểm tương tự, trong đó mỗi ví dụ có thể là thật hoặc giả. Thứ tự như vậy đơn giản hơn để tạo trong

một cài đặt được kiểm soát, nơi chúng tôi phát triển và thử nghiệm các khung nhận dạng DF sử dụng nội dung phương tiện có sẵn thực hoặc được sản xuất. Tuy nhiên, trong điều kiện thực tế, phim có thể được sửa đổi theo cách khác với DF; do đó, bộ phim không được kiểm soát nói chung là không trung thực.

Bên cạnh đó, vì nội dung DF có thể được thay đổi theo nhiều cách khác nhau, chẳng hạn như phương tiện truyền thông chung, nên một cái tên đơn độc có thể không hoàn toàn đúng. Hơn nữa, diện mạo của ít nhất một cá nhân thường bị thay đổi với DF trên một số cạnh trong phim trực quan, bao gồm cả diện mạo của nhiều cá nhân. Chiến lược đặt hàng song song nên được chuyển sang mô tả/khám phá đặc tính vùng lân cận ở cấp vỏ bọc để quản lý những khó khăn trong việc cài đặt chính xác.

4.6.3 Các chiến lược nhận dạng thiếu tính hợp lý

Các chiến lược nhận dạng DF hiện có thường hoạt động để phân tích một tập dữ liệu quan trọng theo nhóm. Tuy nhiên, khi các nhà văn hoặc yêu cầu luật sử dụng các công cụ này trong lĩnh vực này, có thể chỉ có một số tài khoản có sẵn để kiểm tra. Giả sử không thể kiểm tra điểm số toán học liên quan đến xác suất âm thanh hoặc video chính hãng hoặc giả mạo bằng chứng hợp pháp về điểm số. Trong trường hợp đó, nó không quan trọng đối với các chuyên gia. Trong các trường hợp cụ thể, thông thường người ta tìm cách làm rõ toàn bộ điểm toán để bài kiểm tra được chấp nhận trước khi nó được phân phối hoặc sử dụng tại tòa án. Tuy nhiên, hầu hết các tính toán vị trí của DF, chủ yếu phụ thuộc vào DL đang đến gần, cần được làm rõ như vậy vì bản chất khám phá của chúng.

4.6.4 Tập hợp tạm thời

Vị trí DF di chuyển về phía hiện đang được sử dụng tùy thuộc vào nhóm được ghép nối ở cấp độ vỏ; ví dụ: quyết định xác suất của mọi phác thảo video là xác thực hoặc được kiểm soát. Mặt khác, các phương pháp này không xem xét tính nhất quán thoáng qua giữa các phương pháp, điều này có thể dẫn đến hai vấn đề:

- i) Nội dung DF có thể hiển thị các cổ vật thoáng qua và
- ii) các cạnh chính hãng hoặc giả mạo có thể xảy ra trong các nhịp liên tiếp.

Bên cạnh đó, các chiến lược này yêu cầu một giai đoạn khác để xử lý điểm số độ tin cậy của video, vì chúng phải phối hợp điểm số từ mỗi trường hợp để hiển thị ở sản phẩm cuối cùng.

4.6.5 Rửa tiền trên mạng xã hội

Các mạng web chính được sử dụng để truyền bá dữ liệu truyền thông chung giữa xã hội nói chung là các nền tảng xã hội như Twitter, Facebook và Instagram. Trước khi chuyển giao, nội dung đó không chứa thông tin siêu dữ liệu, được kiểm tra kỹ lưỡng và nén nghiêm ngặt để giảm dung lượng truyền mạng hoặc đảm bảo khả năng bảo vệ của khách hàng. Những sửa đổi này, còn được gọi là rửa phương tiện trực tuyến, loại bỏ các dấu hiệu của giả mạo cơ bản và do đó, làm tăng số lượng mặt trái không có thật được phân biệt. Rửa phương tiện trực tuyến ảnh hưởng đáng kể đến hầu hết các chiến lược nhận dạng DF sử dụng

vẫn đề trung tâm mức tín hiệu. Việc xem xét tái tạo các tác động này trong quá trình chuẩn bị thông tin, cũng như mở rộng bộ dữ liệu đánh giá để ghi nhớ thông tin cho tài liệu trực quan được rửa trên phương tiện truyền thông dựa trên web, là một cách để cải thiện độ chính xác của các tính toán DF ID đối với quá trình rửa phương tiện trực tuyến.

4.7 TÓM TẮT

Các bản ghi giao dịch khuôn mặt là mục tiêu được công nhận rộng rãi nhất của các phương pháp nhận dạng DF muộn và hầu hết các bản ghi giả mạo được chuyển giao đều thuộc loại này.

Tôi. xác định cổ vật còn sót lại từ sự tương tác thời đại; ví dụ, có sự bất thường ở đầu [59], không néo mắt [60], các dạng bóng mờ trên mặt [61] và cách sắp xếp răng,

thứ hai. phát hiện các xét nghiệm do GAN tạo ra không rõ ràng,

iii. cung cấp tạm thời không gian, và

v.v. các tín hiệu tâm lý như xung [62] và giá trị hành vi của một số ít ards [63].

Mặc dù có bao nhiêu công việc đã được thực hiện trong nhận dạng bằng rô-bốt, nhưng vẫn còn chỗ để cải thiện.

- Các phương pháp tiếp cận hiện tại không thể chống lại các chu kỳ xử lý sau như áp lực, tác động chấn động và thay đổi ánh sáng, trong số những thứ khác. Hơn nữa, chỉ một số lượng khám phá hạn chế đã được đưa ra để xác định DF âm thanh và hình ảnh.
- Hầu hết các hệ thống muộn đã tập trung vào nhận dạng giao dịch khuôn mặt bằng cách sử dụng các vết bẩn như cổ vật trực quan. Tuy nhiên, khi sự đổi mới thúc đẩy, các giao dịch khuôn mặt hiện đại hơn, chẳng hạn như đóng giả ai đó có hình dạng khuôn mặt, tính cách và kiểu tóc tương tự, sớm muộn gì cũng có thể xảy ra. Các loại DF khác nhau, chẳng hạn như tái hiện khuôn mặt và hát nhép, đang trở nên nổi tiếng hơn.
- Bằng cách mở rộng các vấn đề trung tâm ở mức tín hiệu được sao chép được sử dụng bởi các chiến lược bằng chứng có thể nhận dạng được hiện có, các phương pháp luận trái pháp luật có thể đặt tên một video duy nhất là DF, một trạng thái, chúng tôi gọi là DF giả mạo.
- Hơn nữa, để chống lại DF, một số nhà tiêu luận đã đề xuất sử dụng các ý tưởng về chuỗi khôi và hợp đồng thông minh để nhận biết các thay đổi pháp lý được thực hiện bên trong nội dung trực quan [64,65]. Thực vậy, ngay cả trong tầm nhìn của các cuộc tấn công kiểm soát khác, Sutton [65] đã sử dụng các thỏa thuận quan trọng của Ethereum để tìm và theo dõi nguồn gốc cũng như lịch sử của dữ liệu đã điều chỉnh và nguồn của nó. Hợp đồng thông minh này đã sử dụng hàm băm của khung bản ghi liên hành tinh để lưu các bản ghi cùng với siêu dữ liệu của chúng. Kỹ thuật này có thể khá thi đẻ nhận dạng DF, nhưng sẽ hữu ích nếu siêu dữ liệu video có thể truy cập được. Sử dụng các bộ dữ liệu không đáng kể để xác định DF bằng khoa học vật liệu AI: Công cụ tìm DF gấp phải những xác rối do thông tin rời rạc, ít ỏi và ôn ào trong chu trình chuẩn bị. Các cấu trúc, tính toán và phương pháp tiếp cận AI giàu trí tưởng tượng “chuẩn bị cho” khoa học vật liệu,

số học và báo cáo trước đó phù hợp với DF nên được điều tra.

Việc sử dụng thông tin được đưa vào để tìm ra cách cài đặt khoa học vật liệu và phát biểu trước đó vào AI sẽ giúp giải quyết những khó khăn do thông tin không đầy đủ và hoạt động với cấu trúc của các mô hình tạo ra logic và nhân quả.

- Các số nhận dạng DF hiện có về cơ bản phụ thuộc vào các phần cố định của các cuộc tấn công kỹ thuật số hiện có thông qua các thủ tục AI; ví dụ: nhóm không được hỗ trợ và nhóm được quản lý tiến gần hơn, khiến chúng ít có xu hướng nhận ra các DF tối nghĩa. Sau đó, các chiến lược học tăng cường (RL) [66] có thể đảm nhận một phần quan trọng trong bằng chứng phân biệt về DF sau này.
- Bởi vì nhiều DF lõi xôn bao gồm các chuỗi thực hành động nhất thời, các chiến lược như [67] có thể được sử dụng để lập kế hoạch thử thách khám phá đối với nhiệm vụ dự báo giá trị trạng thái chuỗi Markov. Mô hình kỳ vọng về giá trị trạng thái có thể là phép tính RL tương phản thoáng qua (TD) trực tiếp [67], với đầu ra tương phản với một cạnh định trước để xác định hàng thật và hàng lừa DF. Sau đó, một lần nữa, bạn có thể sử dụng thủ tục RL dựa trên bit với TD bình phương nhỏ nhất. Khả năng suy đoán của TD RL được cải thiện bằng cách áp dụng các mảnh kéo gần, đặc biệt là trong các trường phán tử chiều cao và phi tuyến tính. Sau đó, thủ tục TD bình phương nhỏ nhất có thể được sử dụng để đánh giá một cách đáng tin cậy các xác suất đặc thù, do đó mở rộng khả năng tồn tại của mã định danh DF.
- Các điều khoản dự kiến (ví dụ: các âm thanh hình ảnh và âm thanh khác nhau, v.v.) cần thiết để đánh giá khả năng tồn tại của các phép tính phát hiện DF mạnh mẽ hơn bị thiếu trong bộ dữ liệu DF hiện có. Các nhà nghiên cứu có uy tín đã bỏ qua cách phim DF chứa các hình ảnh bị đặt và điều chỉnh âm thanh. Bộ dữ liệu DF là các giả mạo hình ảnh khảo sát có thể truy cập được và bỏ qua các giả mạo âm thanh. Việc nhân bản và phát lại giọng nói chế nhạo có thể sớm có tác động đáng kể hơn trong thời đại video DF. Trong bản ghi âm DF, giả âm thanh nồng có thể được kết hợp với giả âm thanh sâu. Chúng tôi đã thực hiện một ngữ liệu chế giễu công nhận diễn ngôn [67]. Hiện tại, chúng tôi đang xử lý việc tạo một bộ dữ liệu DF đa phương tiện và sao chép giọng nói vững chắc có thể được sử dụng để kiểm tra tính phù hợp của các khung nhận dạng DF phương tiện tiên tiến có sẵn.

Chương tiếp theo sẽ tập trung vào việc phát triển mô hình dịch ảnh để chống lại các cuộc tấn công của đối thủ.

NGƯỜI GIỚI THIỆU

- [1] M. Westerlund, "Sự xuất hiện của công nghệ DF: Đánh giá," Đánh giá quản lý đổi mới công nghệ, tập. 9, không. 11, trang 39-52, 2019.
- [2] D. Güera và EJ Delp, "Phát hiện video DF bằng cách sử dụng mạng thần kinh tái phát," trong Hội nghị quốc tế lần thứ 15 của IEEE về giám sát dựa trên tín hiệu và video nâng cao (AVSS), trang 1-6, 2018.
- [3] K. Saxena và N. John, "DF là một con quái vật phát triển nhanh. Đây là lý do tại sao con người và máy móc phải chung tay để chế ngự nó," The Economic Times, 2019 [Trực tuyến]. Có sẵn tại: <https://prime.economictimes.indiatimes.com/news/69050623/technology-and-startups/DF-is-a-fast-Growing-monster-heres-why-man-and-machine-must-join-hands-to-meme-it->

- [4] D. Harris, "DFs: Nội dung khiêu dâm sai trái ở đây và luật pháp không thể bảo vệ bạn," Duke L. Tech. Rev., tập. 17, trang 99-127, 2018.
- [5] J. Fletcher, "AI, and some kind of dystopia: The new face of online post-fact performance anxiety," Tạp chí Sân khấu, tập. 70, không. 4, trang 455-471, 2018.
- [6] C. Vaccari và A. Chadwick, "DF và thông tin sai lệch: Khám phá tác động của video chính trị tổng hợp đối với sự lừa dối, sự không chắc chắn và sự tin tưởng vào tin tức," Social Media+ Xã hội, tập. 6, không. 1, trang 1-13, 2020.
- [7] M. Albahee và J. Almaliki, "DFs: Đánh giá có hệ thống về các mối đe dọa và biện pháp đối phó," Tạp chí Công nghệ thông tin lý thuyết và ứng dụng, tập. 97, không. 22, trang 3242-3250, 2019.
- [8] S. Catherine, "Những kẻ lừa đảo đã sử dụng AI để bắt chước giọng nói của Giám đốc điều hành trong trường hợp tội phạm mạng bất thường," Tạp chí Phố Wall, 2019 [Trực tuyến]. Có tại: www.wsj.com/articles/fraudsters-use-ai-to-mimic-ceos-voice-in-unusual-cybercrime-case-11567157402?mod=searchresults&page=1&pos=1
- [9] S. Raphael, "Các chuyên gia: Gián điệp đã sử dụng khuôn mặt do AI tạo ra để kết nối với các mục tiêu," TIN TỨC AP, 2019 [Trực tuyến]. Có tại: <https://apnews.com/bc2f19097a4c4fffaa00de6770b8a60d>
- [10] X. Yang, Y. Li, và S. Lyu, "Phơi bày những trò giả tạo sâu sắc bằng cách sử dụng những tư thế đầu không phù hợp," ICASSP 2019-2019 Hội nghị quốc tế IEEE về Âm học, Lời nói và Xử lý Tin hiệu (ICASSP), trang 8261-8265, 2019.
- [11] P. Korshunov và S. Marcel, "DFs: Một mối đe dọa mới đối với nhận dạng khuôn mặt? Đánh giá và phát hiện," bản in trước của arXiv arXiv:1812.08685, 2018.
- [12] "Vidtimit," [Trực tuyến]. Có tại: <http://conradsanderson.id.au/vidtimit/>
- [13] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies và M. Nießner, "Faceforensics++: Học cách phát hiện hình ảnh khuôn mặt bị thao túng," trong Kỷ yếu của Hội nghị Quốc tế IEEE về Tầm nhìn Máy tính, trang 1-11, 2019.
- [14] "Đò tìm DF," [Trực tuyến]. Có tại: <https://ai.googleblog.com/2019/09/contributing-data-to-DFs-detection.html>
- [15] "Thử thách phát hiện DF," [Trực tuyến]. Có tại: <https://DFdetectionchallenge.ai>
- [16] YZ Li, X. Yang, P. Sun, HG Qi và SW Lyu, "Celeb-df: Một bộ dữ liệu đầy thách thức quy mô lớn dành cho pháp y DF," bản in lại arXiv arXiv:1909.12962, 2019.
- [17] "Celeb-df(v2)," [Trực tuyến]. Có tại: www.cs.albany.edu/~lsw/celebDFforensics.html
- [18] S. Lorant, Lincoln; một câu chuyện hình ảnh của cuộc sống của mình. Norton, 1969 [Trực tuyến]. Có sẵn tại: www.amazon.com/dp/0393074463
- [19] J. Thies, M. Zollhofer, M. Stamminger, C. Theobalt và M. Nießner, "Face2Face: Chụp và tái hiện khuôn mặt trong thời gian thực của video RGB," Kỷ yếu của Hội nghị IEEE về Hình ảnh và Thị giác Máy tính Sự công nhận, trang 2387-2395, tháng 6 năm 2016, Las Vegas, NV [Trực tuyến]. Có tại: <https://doi.org/10.1109/CVPR.2016.262>
- [20] DK Sharma, L. Gaur và D. Okunbor, "Nén hình ảnh và trích xuất đặc trưng bằng mạng nơ-ron," Kỷ yếu của Viện Khoa học Thông tin và Quản lý, 11(1), trang 33-38, 2007.
- [21] JY Zhu, T. Park, P. Isola, và AA Efros, "Bản dịch từ hình ảnh sang hình ảnh không ghép nối bằng cách sử dụng các mạng đối nghịch nhất quán theo chu kỳ," Kỷ yếu của Hội nghị Quốc tế IEEE về Thị giác Máy tính, trang 2242-2251, Tháng 10 năm 2017, Venice, Ý [Trực tuyến].
Có tại: <https://doi.org/10.1109/ICCV.2017.244>
- [22] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard và cộng sự, "Tensorflow: Một hệ thống dành cho ML quy mô lớn." Kỷ yếu của Hội nghị USENIX về Thiết kế và Triển khai Hệ điều hành, tập. 16, tr. 265-283, tháng 11 năm 2016, Savannah, GA [Trực tuyến]. Có tại: www.usenix.org/conference/osdi16/kỹ thuật phiên/bản trình bày/abadi

- [23] F. Chollet, và cộng sự, "Keras," <https://keras.io>, 2015.
- [24] A. Tewari, M. Zollhofer, H. Kim, P. Garrido, F. Bernard, P. Perez, và C. Theobalt, "MoFA: Bộ mã hóa khuôn mặt xoán sâu dựa trên mô hình cho một mắt không giám sát tái thiết," Kỷ yếu của Hội nghị Quốc tế IEEE về Thị giác Máy tính, trang 3735-3744, tháng 10 năm 2017, Venice, Ý [Trực tuyến]. Có tại: <https://doi.org/10.1109/ICCV.2017.401>
- [25] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, và Y. Bengio, "Mạng đối thủ sáng tạo," Những tiến bộ trong thông tin thần kinh Hệ thống xử lý, trang 2672-2680, tháng 12 năm 2014, Montréal, Canada [Trực tuyến]. Có tại: <http://papers.nips.cc/paper/5423-generative-adversarial-nets>
- [26] G. Antipov, M. Baccouche, và J.-L. Dugelay, "Lão hóa khuôn mặt với GAN có điều kiện," [arXiv:1702.01983v2](https://arxiv.org/abs/1702.01983v2), tháng 2 năm 2017 [Trực tuyến]. Có tại: <https://arxiv.org/abs/1702.01983v2>
- [27] K. Dale, K. Sunkavalli, MK Johnson, D. Vlasic, W. Matusik và H. Pfister, "Thay thế khuôn mặt video," Giao dịch ACM trên Đồ họa, tập. 30, không. 6, trang 1-130, tháng 12 2011 [Trực tuyến]. Có tại: <https://doi.org/10.1145/2070781.2024164>
- [28] "Ứng dụng giả," www.fakeapp.org/
- [29] C. Bregler, M. Covell và M. Slatney, "Viết lại video: Thúc đẩy lời nói bằng hình ảnh bằng âm thanh," Kỷ yếu Hội nghị Thường niên ACM về Đồ họa Máy tính và Kỹ thuật Tương tác, trang 353-360, tháng 8 năm 1997, Los Angeles, CA [Trực tuyến]. Có tại: <https://doi.org/10.1145/258734.258880>
- [30] H. Averbuch-Elor, D. Cohen-Or, J. Kopf, và MF Cohen, "Đưa chân dung vào cuộc sống," Giao dịch ACM trên Đồ họa, tập. 36, không. 6, trang 196:1-196:13, tháng 11 năm 2017 [Trực tuyến]. Có tại: <https://doi.org/10.1145/3130800.3130818>
- [31] Z. Lu, Z. Li, J. Cao, R. He và Z. Sun, "Tiến bộ gần đây của luận điểm tổng hợp hình ảnh khuôn mặt," [arXiv:1706.04717v1](https://arxiv.org/abs/1706.04717v1), tháng 6 năm 2017 [Trực tuyến]. Có tại: <https://arxiv.org/abs/1706.04717v1>
- [32] Y. Lu, Y.-W. Tai, và C.-K. Tang, "Tạo khuôn mặt theo hướng dẫn thuộc tính bằng cách sử dụng CycleGAN có điều kiện," Kỷ yếu của Hội nghị Châu Âu về Tầm nhìn Máy tính, trang 293-308, tháng 10 năm 2018, Munich, Đức [Trực tuyến]. Có tại: <https://doi.org/10.1007/978-3-030-01258-8n18>
- [33] P. Upchurch, J. Gardner, G. Pleiss, R. Pless, N. Snavely, K. Bala và K. Weinberger, "Nội suy đặc trưng sâu cho các thay đổi nội dung hình ảnh," Kỷ yếu của Hội nghị IEEE về Tầm nhìn Máy tính và Nhận dạng mẫu, trang 6090-6099, tháng 7 năm 2017, Honolulu, HI [Trực tuyến]. Có tại: <https://doi.org/10.1109/CVPR.2017.645>
- [34] G. Lample, G. Lample, N. Zeghidour, N. Usunier, A. Bordes, L. Denoyer, và M. Ranzato, "Mạng Fader: Thảo tác hình ảnh bằng thuộc tính trượt," Những tiến bộ trong Hệ thống xử lý thông tin thần kinh, trang 5967-5976, tháng 12 năm 2017, Long Beach, CA [Trực tuyến]. Có tại: <https://papers.nips.cc/paper/7178-fader-networks-manipulating-images-by-sliding-attributes>
- [35] T. Karras, T. Aila, S. Laine và J. Lehtinen, "Sự phát triển dần dần của GAN để có chất lượng, sự ổn định và biến thể đã được chứng minh," Kỷ yếu của Hội nghị Quốc tế về Biểu diễn Học tập, tháng 4 năm 2018 [Trực tuyến]. Có sẵn tại: <https://openreview.net/forum?id=Hk99zCeAb>
- [36] S. Hochreiter và J. Schmidhuber, "Trí nhớ ngắn hạn dài," Tính toán thần kinh, tập. 9, không. 8, trang 1735-1780, tháng 11 năm 1997 [Trực tuyến]. Có tại: <https://doi.org/10.1162/neco.1997.9.8.1735>
- [37] D. Güera và EJ Delp, "Phát hiện video DF bằng cách sử dụng mạng thần kinh tái phát," Hội nghị quốc tế lần thứ 15 của IEEE về giám sát dựa trên tín hiệu và video nâng cao (AVSS), trang 1-6, 2018.

- [38] K. Simonyan và A. Zisserman, "Mạng tích chập rất sâu cho quy mô lớn nhận dạng hình ảnh," bản in trước của arXiv arXiv:1409.1556, 2014.
- [39] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens và Z. Wojna, "Suy nghĩ lại về kiến trúc ban đầu cho thị giác máy tính," Kỷ yếu của Hội nghị IEEE về Thị giác máy tính và Nhận dạng mẫu, trang 2818- 2826, 2016.
- [40] J. Donahue, L.A Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko và T. Darrell, "Mạng tích chập định kỳ dài hạn để mô tả và nhận dạng trực quan," trong CVPR, trang 2625-2634, 2015.
- [41] Y. Li, MC Chang và S. Lyu, "In ictu oculi: Phơi bày AI đã tạo video giả bằng cách phát hiện chớp mắt," Hội thảo quốc tế IEEE 2018 về pháp y và bảo mật thông tin (WIFS), trang 1-7, 2018.
- [42] D. Afchar, V. Nozick, J. Yamagishi và I. Echizen, "Mesonet: Mạng phát hiện giả mạo video khuôn mặt nhỏ gọn," Hội thảo quốc tế IEEE 2018 về pháp y và bảo mật thông tin (WIFS), trang 1-7 , 2018.
- [43] S. Ioffe và C. Szegedy, "Chuẩn hóa hàng loạt: Tăng tốc đào tạo mạng sâu bằng cách giảm sự dịch chuyển đồng biến bên trong," bản in trước của arXiv arXiv:1502.03167, trang 1-7, 2015.
- [44] N. Murray và F. Perronnin, "Tổng hợp tối đa được tổng quát hóa," trong Kỷ yếu của Hội nghị IEEE về Tầm nhìn Máy tính và Nhận dạng Mẫu, trang 2473-2480, 2014.
- [45] Y. Li và S. Lyu, "Hiển thị các video DF bằng cách phát hiện các tạo tác cong vênh trên khuôn mặt," arXiv trước in arXiv:1811.00656, 2018.
- [46] K. He, XY Zhang, SQ Ren và J. Sun, "Học tập dư sâu để nhận dạng hình ảnh," trong Kỷ yếu hội nghị IEEE về thị giác máy tính và nhận dạng mẫu, trang 770-778, 2016.
- [47] CC Chang và CJ Lin, "Libsvm: Một thư viện cho các máy vectơ hỗ trợ," ACM giao dịch trên các hệ thống và công nghệ thông minh (TIST), tập. 2, không. 3, trang 1-27, 2011.
- [48] J. Wu, K. Feng, X. Chang, X và T. Yang, "Phương pháp pháp y cho hình ảnh DF dựa trên nhận dạng khuôn mặt," trong Kỷ yếu của Hội nghị công nghệ cụm và máy tính hiệu suất cao lần thứ 4 năm 2020 lần thứ 3 năm 2020 Hội nghị quốc tế về Dữ liệu lớn và AI, trang 104-108, 2020.
- [49] X. Yang, Y. Li, và S. Lyu, "Phơi bày những trò giả tạo sâu sắc bằng cách sử dụng các tư thế đầu không nhất quán," ICASSP 2019-2019 Hội nghị quốc tế IEEE về Âm học, Lời nói và Xử lý Tin hiệu (ICASSP), trang 8261-8265, 2019.
- [50] F. Matern, C. Riess và M. Stamminger, "Khai thác các đồ tạo tác trực quan để phơi bày DF và các thao tác trên khuôn mặt," Hội thảo Ứng dụng thị giác máy tính mùa đông năm 2019 của IEEE (WACVW), trang 83-92, 2019.
- [51] C. Chen, MN Do và J. Wang, "Làm mờ hình ảnh và video mạnh mẽ với khả năng triệt tiêu tạo tác thị giác thông qua giảm thiểu dư lượng gradient," trong Hội nghị Châu Âu về Thị giác Máy tính, trang 576-591, 2016.
- [52] J. Thies, M. Zollhöfe, M. Stamminger, C. Theobalt và M. Nießner, "Face2face: Chụp và tái hiện khuôn mặt trong thời gian thực của video rgb," Kỷ yếu của Hội nghị IEEE về Thị giác máy tính và Nhận dạng mẫu, trang 2387-2395, 2016.
- [53] G. Singh, B. Kumar, L. Gaur và A. Tyagi (2019), "So sánh giữa Đa thức và Bernoulli Naïve Bayes để phân loại văn bản," Hội nghị quốc tế 2019 về Quản lý tự động hóa, tính toán và công nghệ (ICACTM), trang .593-596. doi:10.1109/ICACTM.2019.8776800.
- [54] T. Windeatt, "Thiết kế bộ phân loại mlp tập hợp," trong Computational Intelligence Các khung mẫu, trang 133-147, 2008.
- [55] MA Mansournia, A. Geroldinger, S. Greenland và G. Heinze, "Sự tách biệt trong hồi quy logistic: Nguyên nhân, hậu quả và kiểm soát," Tập chí Dịch tễ học Hoa Kỳ, tập. 187, không. 4, trang 864- 870, 2018.

- [56] L. Gaur, U. Bhatia, NZ Jhanjhi, G. Muhammad và M. Masud (2021). Phát hiện COVID-19 dựa trên hình ảnh y tế bằng cách sử dụng mạng thần kinh tích chập sâu.
Hệ thống đa phương tiện. doi:10.1007/s00530-021-00794-6
- [57] T. Soukupova và J. Cech, "Phát hiện chớp mắt bằng cách sử dụng các mốc trên khuôn mặt," trong Hội thảo mùa đông về thị giác máy tính lần thứ 21, Rimske Toplice, Slovenia, 2016.
- [58] S. McCloskey và M. Albright, "Phát hiện hình ảnh do GAN tạo bằng cách sử dụng tín hiệu màu," bản in trước arXiv arXiv: .08247, 2018.
- [59] S. Fernandes, và cộng sự, "Dự đoán các biến thể nhịp tim của video DF bằng cách sử dụng ODE thần kinh," trong Kỷ yếu của Hội thảo Quốc tế IEEE về Hội thảo Thị giác Máy tính, 2019.
- [60] S. Agarwal, H. Farid, Y. Gu, M. He, K. Nagano và H. Li, "Bảo vệ các nhà lãnh đạo thế giới chống lại Deep Fakes," trong Kỷ yếu Hội nghị IEEE về Thị giác Máy tính và Hội thảo Nhận dạng Mẫu , trang 38-45, 2019.
- [61] P. Fraga-Lamas và TM Fernández-Caramés, "Tin giả, thông tin sai lệch và DF: Tận dụng công nghệ số cái phân tán và chuỗi khôi để chống lại sự lừa dối kỹ thuật số và thực tế giả," IT Professional, tập. 22, không. 2, trang 53-59, 2020.
- [62] HR Hasan và K. Salah, "Chống lại các video DF bằng cách sử dụng chuỗi khôi và hợp đồng thông minh," Truy cập IEEE, tập. 7, trang 41596-41606, 2019.
- [63] M. Feng và H. Xu, "Bảo vệ tối ưu dựa trên học tăng cường sâu cho hệ thống vật lý không gian mạng trước cuộc tấn công mạng không xác định," trong Chuỗi hội nghị chuyên đề của IEEE về Trí tuệ tính toán (SSCI) năm 2017, trang 1-8, 2017 : IEEE.
- [64] X. Xu và T. Xie, "Một cách tiếp cận học tăng cường để phát hiện xâm nhập dựa trên máy chủ sử dụng các chuỗi lệnh gọi hệ thống," trong Hội nghị Quốc tế về Máy tính Thông minh, trang 995-1003, 2005: Springer.
- [65] RS Sutton, "Học cách dự đoán bằng phương pháp sai biệt thời gian," ML, tập. 3, không. 1, trang 9-44, 1988.
- [66] Gaur, L., Singh, G., Solanki, A., Jhanjhi, NZ, Bhatia, U., Sharma, S., . and Kim, W. (2021), "Sử mệnh của thanh niên trong việc dự đoán sự phát triển bền vững các mục tiêu bằng cách sử dụng thuật toán rừng ngẫu nhiên và thần kinh mờ" Khoa học thông tin và điện toán lấy con người làm trung tâm, 11, NA.
- [67] R. Baumann, KM Malik, A. Javed, A. Ball, B. Kujawa, và H. Malik, "Văn bản phát hiện giả mạo giọng nói để phát lại âm thanh đơn và đa thứ tự," Ngôn ngữ giọng nói máy tính, tập. 65, tr. 101132, 2021.



Taylor & Francis
Taylor & Francis Group
<http://taylorandfrancis.com>

5 Phát triển hình ảnh

Dịch mô hình sang đối thủ truy cập tấn công

Bò tót Loveleen, Mohan Bhandari, và
tanvi razdan

NỘI DUNG

5.1 Giới thiệu57
5.2 Công việc liên quan58
5.3 Bộ dữ liệu61
5.4 Tiền xử lý dữ liệu61
5.5 Phương pháp61
5.6 Nhiều64
5.7 Độ dốc65
5.8 Kết quả và thảo luận66
5.8.1 Cài đặt tấn công66
5.9 Phân tích kết quả66
5.10 Tóm tắt69
Người giới thiệu69

5.1 GIỚI THIỆU

DeepFakes (DF) hoặc ảnh và video được sửa đổi khuôn mặt có thể được sử dụng một cách xúc phạm để truyền bá thông tin sai lệch hoặc làm mất uy tín của ai đó. DF sử dụng AI để chồng lên giọng nói và chân dung, cho phép họ đưa lời nói của người khác vào miệng họ một cách cơ bản. DF sinh sôi nảy nở trên khắp các phương tiện truyền thông xã hội và chính thống, và những nguồn này đang tranh giành để kiểm soát việc lan truyền thông tin có khả năng gây hiểu lầm trên nền tảng của họ [1]. Do đó, việc nhận dạng DF là rất quan trọng để cải thiện độ tin cậy của các nền tảng truyền thông xã hội và các trang web chia sẻ phương tiện truyền thông khác. Các phương pháp phát hiện DF hiện đang được sử dụng dựa trên các mô hình phân loại dựa trên mạng thần kinh đã được tiết lộ là yếu tố với các ví dụ đối nghịch [2].

DL là cốt lõi của sự phát triển hiện tại của AI. Nó đã trở thành xương sống của lĩnh vực thị giác máy tính [3]. Với những tiến bộ gần đây trong lĩnh vực thị giác máy tính [4,5] và xử lý ngôn ngữ tự nhiên [6], họ cùng nhau đưa các bộ phân loại được đào tạo vào trung tâm của các hệ thống bảo mật quan trọng. Các ứng dụng từ

ô tô tự lái để giám sát và an ninh đều là những ví dụ điển hình. Do những phát triển này, bảo mật ML ngày càng trở nên quan trọng. Đặc biệt, khả năng chống lại các đầu vào được lựa chọn một cách bất lợi đang trở thành một mục tiêu thiết kế quan trọng. Mặc dù các mô hình được đào tạo thường khá giỏi trong việc phân loại các đầu vào vô thường vô phạt, nhưng nghiên cứu mới cho thấy chúng không phải lúc nào cũng giống nhau. Một số công trình [7,8,9] chứng minh cách một đối thủ có thể thường xuyên thay đổi đầu vào để khiến mô hình cung cấp đầu ra không chính xác. Những thay đổi nhỏ đối với hình ảnh đầu vào có thể đánh lừa mạng thần kinh hiện đại hoạt động với độ tin cậy cao, khiến thị giác máy tính trở thành một thử thách đặc biệt hấp dẫn [10].

Các mạng nơ-ron sâu (DNN) cực kỳ hiện đại và tiên tiến có hiệu quả cao trong việc giải quyết nhiều vấn đề phức tạp trong thế giới thực nhưng lại là đối tượng của các ví dụ bất lợi, gây rủi ro bảo mật cho các thuật toán này do các kết quả có thể nghiêm trọng.

Trước khi các mô hình DL được triển khai, các cuộc tấn công đối nghịch được sử dụng làm proxy để kiểm tra khả năng phục hồi của chúng. Mặt khác, hầu hết các cuộc tấn công đối nghịch chỉ có thể đánh lừa một mô hình hộp đen với tỷ lệ thành công thấp [11]. Nó vẫn đúng mặc dù trường hợp lành tính đã được phân loại chính xác và sự thay đổi không thể phát hiện được đối với con người. Bên cạnh những cân nhắc về bảo mật, bằng chứng này giải thích rằng các mô hình hiện tại của chúng tôi không liên tục học các ý tưởng cơ bản. Các cuộc tấn công đối nghịch là các thuật toán tìm các hình ảnh có độ tương đồng cao để đánh lừa bộ phân loại. Huấn luyện các bộ phân loại dưới sự tấn công của kẻ thù đã trở thành một trong những cách có lợi nhất để tăng cường độ bền của các bộ phân loại [12]. Mạng đối thủ chung (GAN) là một mô hình tổng quát trong đó trình tạo học cách chuyển đổi nhiễu tráng thành hình ảnh trông xác thực đối với bộ phân biệt đối xử [13,14].

Hiện tượng các ví dụ đối nghịch-đầu vào được tạo có chủ ý nhằm đánh lừa các mô hình ML được đào tạo đã thu hút sự quan tâm của cộng đồng học thuật trong những năm gần đây, chủ yếu khi chỉ giới hạn ở những thay đổi nhỏ đối với đầu vào được diễn giải chính xác.

Người ta cũng thấy rằng các bộ phân loại hình ảnh cũng hoạt động kém hiệu quả trên các hình ảnh bị biến dạng ngẫu nhiên, chẳng hạn như hình ảnh có nhiễu Gaussian phụ gia [15]. Nhiều kỹ thuật phức tạp dựa trên khoảng cách Lp để xử lý các nhiễu loạn đã được phát triển để tạo ra các ví dụ đối nghịch. Để bảo vệ chống lại các cuộc tấn công đối nghịch như vậy, các nhà nghiên cứu đã xem xét các cơ chế phòng thủ khác nhau. Việc sử dụng khoảng cách Lp làm thước đo chất lượng cảm nhận vẫn đang được nghiên cứu [16].

Do tính phức tạp của chúng, rất khó để xác định cách các mô hình ML có thể hoạt động sai hoặc bị lạm dụng khi triển khai. Nghiên cứu gần đây về các trường hợp nghịch cảnh, hoặc trong trường hợp có những thay đổi khiêm tốn dẫn đến các dự đoán mô hình khác biệt đáng kể, đã giúp đánh giá mức độ vững chắc của các mô hình này bằng cách làm nổi bật các tình huống thù địch khi chúng thất bại. Mặt khác, những lỗi loạn có ý này thường là giả tạo, thiếu ý nghĩa ngữ nghĩa và không thể áp dụng cho các lĩnh vực phức tạp như ngôn ngữ [17].

5.2 CÁC CÔNG VIỆC LIÊN QUAN

Ruiz et al. [18] đã đề xuất và áp dụng thành công các cuộc tấn công đối nghịch có thể chuyển đổi lớp, khái quát hóa cho các lớp khác nhau và đào tạo đối thủ cho GAN như là bước đầu tiên hướng tới các mạng dịch hình ảnh mạnh mẽ. Họ đã nghiên cứu một cuộc tấn công đối thủ trên phỏ rỗng có khả năng vượt qua sự phòng thủ mờ nhạt. phương pháp của họ đạt được tốt hơn

hiệu suất trên các kịch bản mờ Gaussian với cường độ mờ cao. Giải pháp của họ vượt trội so với đối thủ trong các trường hợp mờ Gaussian với cường độ mờ lớn.

Kỹ thuật phô trá rỗng lặp đi lặp lại của họ nhanh hơn khoảng K lần so với kỹ vọng so với chuyển đổi (EoT) vì mỗi lần lặp lại của Phương pháp kỹ hiệu độ dốc nhanh lặp lại (I-FGSM) chỉ yêu cầu một lần chuyển tiếp-lùi thay vì K để tính toán tồn thắt.

Qiu et al. [19] tóm tắt toàn diện tiến trình nghiên cứu mới nhất về các công nghệ phòng thủ và tấn công đối thủ trong học sâu. Các phát hiện cho thấy rằng khi so sánh với các cuộc tấn công hộp đen, các cuộc tấn công hộp trắng có tỷ lệ thành công cao hơn, cho phép mô hình mục tiêu đạt được tỷ lệ lỗi khoảng 89%-99%.

Mặc dù tỷ lệ thành công của các cuộc tấn công hộp đen, dẫn đến tỷ lệ lỗi khoảng 84%-96%, không hợp lý bằng các cuộc tấn công hộp trắng, vì các cuộc tấn công hộp đen không cần biết bất kỳ thông tin nào của hộp đen. mô hình mục tiêu, các cuộc tấn công sarial của đối thủ có thể được thực hiện bằng cách sử dụng khả năng chuyển đổi của các mẫu đối thủ, đảo ngược mô hình và trích xuất mô hình.

Finlayson và cộng sự. [20] đề xuất và triển khai các mô hình đại diện cho tình trạng hiện tại của nghệ thuật trong thị giác máy tính y tế. Tất cả các mô hình cơ sở đều đạt được hiệu suất tương đối phù hợp với kết quả được báo cáo trong các bản thảo gốc về hình ảnh tự nhiên: AUROC là 0,910 đối với bệnh võng mạc do tiểu đường, AUROC là 0,936 đối với tràn khí màng phổi và AUROC là 0,86 đối với khối u ác tính. Dự kiến các cuộc tấn công giảm dần độ dốc, nhắm mục tiêu vào câu trả lời không chính xác trong mọi trường hợp, tạo ra AUROC hiệu quả là 0,000 và độ chính xác là 0% cho tất cả các cuộc tấn công hộp trắng. Các cuộc tấn công hộp đen tạo ra AUROC dưới 0,10 cho tất cả các tác vụ và độ chính xác nằm trong khoảng từ 0,01% khi soi đáy mắt đến 37,9% khi soi da. Về chất lượng, tất cả các cuộc tấn công đều không thể nhận thấy được bằng con người. Các cuộc tấn công và lỗi đối thủ cũng đạt được AUROC hiệu quả là 0,000 và độ chính xác <1% đối với các cuộc tấn công hộp trắng trên tất cả các tác vụ. Các cuộc tấn công bẩn và đối thủ hộp đen đạt AUROC dưới 0,005 cho tất cả các tác vụ và độ chính xác dưới 10%. Các điều khiển "bẩn và tự nhiên" được tạo bằng cách thêm các bẩn và được tạo từ hình ảnh được phân loại mạnh nhất của loại mong muốn dẫn đến AUROC nằm trong khoảng từ 0,48 đến 0,83 với độ chính xác nằm trong khoảng từ 67,5% đến 92,1%.

Samangouei et al. [21] đã đề xuất Defense-GAN, một khung mới tận dụng khả năng biểu cảm của các mô hình tổng quát để bảo vệ DNN trước các cuộc tấn công như vậy. Nó được đào tạo để mô hình hóa việc phân phối các hình ảnh không bị xáo trộn. Nó tìm thấy một dấu ra gần giống với một hình ảnh nhất định không chứa các thay đổi bất lợi. Đầu ra này sau đó được đưa đến bộ phân loại. Phương pháp được đề xuất có thể được sử dụng với bất kỳ mô hình phân loại nào và không sửa đổi cấu trúc phân loại hoặc quy trình đào tạo. Hiệu suất của Defense-GAN-Rec và của Defense-GAN-Orig rất gần nhau và MagNet đạt được độ chính xác thấp hơn Defense-GAN.

Santhanam và Grnarova [22] đã đề xuất cao bồi, một cách tiếp cận để phát hiện và bảo vệ chống lại các cuộc tấn công của kẻ thù bằng cách sử dụng bộ phân biệt đối xử và trình tạo GAN được đào tạo trên cùng một bộ dữ liệu. Người phân biệt đối xử luôn cho điểm các mẫu sarial của đối thủ thấp hơn các mẫu thực qua nhiều cuộc tấn công và bộ dữ liệu theo cách tiếp cận này. Họ cũng đưa ra một phương pháp làm sạch sử dụng cả bộ phân biệt và bộ tạo của GAN để chiếu các mẫu trở lại đa tạp dữ liệu. Quy trình làm sạch này độc lập với bộ phân loại và kiểu tấn công và do đó có thể được triển khai trong các hệ thống hiện có.

Hu và Tan [23] cung cấp MalGAN, một cách tiếp cận dựa trên GAN để tạo các phiên bản phần mềm độc hại đối nghịch có thể phá vỡ các mô hình phát hiện dựa trên ML hộp đen. Để phù hợp với hệ thống phát hiện phần mềm độc hại hộp đen, MalGAN sử dụng một trình phát hiện thay thế. Một mạng chung được đào tạo để giảm xác suất độc hại được dự đoán bởi bộ phát hiện thay thế trong các mẫu đối nghịch được tạo. MalGAN vượt trội so với các kỹ thuật tạo mẫu đối thủ dựa trên độ dốc điển hình bằng cách hạ thấp tỷ lệ phát hiện xuống thực tế bằng 0 và tạo ra các phương pháp phòng thủ dựa trên đào tạo lại chống lại các mẫu đối thủ khó thực hiện. Theo dữ liệu thực nghiệm, các ví dụ đối nghịch được tạo ra có thể tránh máy dò hộp đen một cách hiệu quả. Các nhà sản xuất phần mềm độc hại có thể bẻ khóa ngay bộ phát hiện hộp đen sau khi nó được cập nhật.

Gandhi và Jain [24] đã tạo ra các nhiễu loạn đối nghịch bằng cách sử dụng Phương pháp dấu hiệu chuyển màu nhanh (FGSM) và chuẩn L2 của Carlini và Wagner để tạo ra các nhiễu loạn đối nghịch. Trên các DF không bị xáo trộn, máy dò thu được độ chính xác hơn 95% nhưng độ chính xác thấp hơn 27% trên các DF bị nhiễu. Họ đã phát hiện ra rằng hệ thống bảo vệ Xử lý hình ảnh kỹ thuật số (DIP) loại bỏ các nhiễu không được giám sát bằng cách sử dụng các mạng thần kinh tích chập tổng quát. Trung bình, việc chính quy hóa đã tăng khả năng phát hiện trên mỗi DF bị xáo trộn, với mức cải thiện 10% trong kịch bản hộp đen. Trên một mẫu con gồm 100 hình ảnh, hệ thống phòng thủ DIP đạt được độ chính xác 95% đối với các DF bị nhiễu đánh lừa máy dò ban đầu trong khi vẫn duy trì độ chính xác 98% trong các trường hợp khác.

Liu và Hsieh [25] tạo khung Rob-GAN để cùng nhau tối ưu hóa trình tạo và trình phân biệt đối xử khi đối mặt với các cuộc tấn công đối nghịch-trình tạo tạo ra các hình ảnh giả để đánh lừa trình phân biệt. Ngược lại, kẻ tấn công đối thủ làm nhiễu các bức ảnh xác thực để đánh lừa người phân biệt đối xử và người phân biệt đối xử muốn giảm thiểu tổn thất dưới cả hình ảnh giả mạo và đối thủ. Theo các phát hiện, trình phân loại được tạo ra có khả năng phục hồi tốt hơn so với chiến lược đào tạo đối thủ hiện đại và trình tạo đánh bại SN-GAN trên ImageNet-143.

Croce và Hein [26] đề xuất một phương pháp hộp đen để tạo các mẫu nghịch cảnh nhằm giảm khoảng cách 10 so với ảnh nguồn. Thủ nghiệm rộng rãi đã tiết lộ rằng cuộc tấn công vượt trội hoặc cạnh tranh với tình trạng hiện tại của nghệ thuật. Hơn nữa, nó có thể kết hợp các ràng buộc nhiễu loạn thành phần bổ sung. Các ví dụ về quảng cáo trên thực tế không thể phát hiện được vì các tác giả chỉ cho phép các pixel thay đổi trong các vùng có độ biến thiên cao và tránh các thay đổi dọc theo các cạnh được căn chỉnh theo trực. Cuộc tấn công Projected Gradient Descent cũng đã được sửa đổi để giải thích cho giới hạn khôn ngoan của thành phần cách tử tích hợp 10 chuẩn, cho phép mô hình thực hiện đào tạo đối thủ để cải thiện khả năng phục hồi của bộ phân loại trước những thay đổi đối nghịch thưa thớt và không thể nhận thấy.

Zhang và Wang [27] trình bày một phương pháp đào tạo đối thủ dựa trên phân tán tính năng để tăng khả năng phục hồi của mô hình trước các cuộc tấn công của đối thủ. Phương pháp được đề xuất sử dụng phân tán tính năng trọng không gian tiềm ẩn để tạo ra các hình ảnh đối nghịch để đào tạo, không được giám sát và loại bỏ rò rỉ nhãn. Quan trọng hơn, phương pháp mới lạ này tạo ra các hình ảnh được thay đổi một cách cộng tác, xem xét các tương tác giữa các mẫu.

Liao et al. [28] cung cấp bộ khử nhiễu hướng dẫn biểu diễn mức cao (HGD) để bảo vệ việc phân loại hình ảnh. Sử dụng hàm măt măt được định nghĩa là sự khác biệt giữa đầu ra của mô hình đích được kích hoạt bởi hình ảnh sạch và hình ảnh khử nhiễu, HGD tránh được

hiện tượng khuếch đại lỗi, trong đó tiếng ồn đối nghịch còn sót lại khiêm tốn được khuếch đại dần dần và dẫn đến phân loại không chính xác. HGD cung cấp một số lợi ích so với đào tạo đồng bộ đối thủ, cách tiếp cận áp phòng thủ tiên tiến nhất hiện nay trên các bức tranh lớn.

Mustafa và cộng sự. [29] giới thiệu một phương pháp nâng cao hình ảnh hiệu quả về mặt tính toán với cơ chế bảo vệ mạnh mẽ để hạn chế thành công ảnh hưởng của nhiễu loạn đối nghịch. Họ chỉ ra rằng các mạng khôi phục hình ảnh sâu có thể học các chức năng ánh xạ di chuyển các mẫu đối nghịch từ bên ngoài đa tạp sang đa tạp hình ảnh tự nhiên, khôi phục phân loại về các lớp thích hợp. Kỹ thuật này độc đáo ở chỗ nó cải thiện chất lượng hình ảnh trong khi vẫn duy trì hiệu suất mô hình trên hình ảnh sạch và mang lại khả năng phục hồi trước các cuộc tấn công. Hơn nữa, giải pháp được đề xuất không yêu cầu một hệ thống riêng biệt để phát hiện các hình ảnh bất lợi và không yêu cầu bất kỳ sửa đổi nào đối với bộ phân loại. Các thử nghiệm mở rộng đã cho thấy hiệu quả của chương trình trong các tình huống hộp xám.

Bài hát và cộng sự. [30] đã đề xuất sử dụng biến đổi Saak bị mất dữ liệu như một công cụ tiền xử lý để bảo vệ chống lại các cuộc tấn công đối nghịch vào các hình ảnh bị nhiễu đối nghịch. Họ phát hiện ra rằng kết quả đầu ra của biến đổi Saak rất tốt trong việc phân biệt giữa các mẫu đối nghịch và mẫu sạch [31,32,33]. Hình ảnh sau khi xử lý được chứng minh là có khả năng chống nhiễu đối nghịch. Trên cả bộ dữ liệu CIFAR10 và ImageNet, mô hình của họ đánh bại các chiến lược phòng thủ tiên tiến nhất của đối thủ với một biên độ đáng kể mà không ảnh hưởng đến hiệu suất phán đoán trên hình ảnh rõ ràng. Điều quan trọng là, những phát hiện của chúng tôi ngụ ý rằng những nhiễu loạn đối nghịch [34,35,36] có thể được chống lại một cách hiệu quả và hiệu quả bằng cách sử dụng phân tích tần số tiên tiến nhất.

5.3 BỘ DỮ LIỆU

Trong nghiên cứu này, chúng tôi đã sử dụng bộ dữ liệu MNIST có sẵn công khai. Cơ sở dữ liệu MNIST chứa 60.000 hình ảnh huấn luyện và 10.000 hình ảnh kiểm tra. Một nửa tập huấn luyện và một nửa tập kiểm tra được lấy từ tập dữ liệu huấn luyện của NIST, trong khi nửa còn lại của tập huấn luyện và nửa còn lại của tập kiểm tra được lấy từ tập dữ liệu thử nghiệm của NIST.

5.4 XỬ LÝ SẼ DỮ LIỆU

Hình ảnh MNIST có kích thước 28×28 pixel, nhưng chúng không được đệm thành 32×32 pixel và được chuẩn hóa trước khi đưa vào mạng. Phần còn lại của mạng không sử dụng bất kỳ phần đệm nào, vì vậy kích thước tiếp tục giảm khi hình ảnh tiến triển qua web.

5.5 PHƯƠNG PHÁP

FGSM liên quan đến việc thêm nhiễu (không phải nhiễu ngẫu nhiên) có cùng hướng với độ dốc của hàm chi phí liên quan đến dữ liệu. Tiếng ồn được điều chỉnh bởi epsilon, thường bị ràng buộc bởi định mức tối đa là một số nguyên nhỏ. Một ví dụ đối nghịch là một phần của dữ liệu đầu vào đã bị thay đổi một chút để khiến thuật toán phân loại sai nó.

Trong thử nghiệm của chúng tôi, chúng tôi có đầu vào là hình ảnh gốc $F(X)$ và đầu ra là Hình ảnh đối nghịch $F(X')$ ở giữa là quá trình mạng thần kinh trong đó thuật toán, lúc đầu, thu thập dấu hiệu phần tử khôn ngoan của độ dốc dữ liệu từ $F(X)$ và sau đó tạo ảnh nhiễu $F(X')$ bằng cách điều chỉnh từng pixel của ảnh đầu vào, sau đó nó thêm clipping để duy trì phạm vi $[0,1]$; điều này tiếp tục cho đến khi thuật toán phân loại sai hình ảnh "chính xác" và chỉ sau đó nó mới quay lại hình ảnh gấp sự cố.

Nếu nó không thể phân loại sai hình ảnh, thuật toán thuật toán sẽ lặp lại các bước tương tự và mạng thần kinh sẽ điều chỉnh các trọng số [37,38,39] mỗi lần cho đến khi thuật toán thuật toán cuối cùng có thể tạo ra hình ảnh đối nghịch.

Giả sử chúng ta nói về vai trò của mạng thần kinh, một lớp ẩn giữa đầu vào và đầu ra của thuật toán, áp dụng các trọng số cho đầu vào và hướng dẫn chúng thông qua chức năng kích hoạt làm đầu ra. Các lớp ẩn thay đổi đầu vào mạng theo kiểu phi tuyến tính. Mục đích của mạng thần kinh xác định các lớp ẩn và bản thân các lớp có thể thay đổi dựa trên các trọng số liên quan của chúng. Các lớp ẩn là một loạt các hàm toán học, mỗi hàm cho một đầu ra cụ thể cho kết quả mong muốn. Ví dụ, các chức năng nén là một số loại lớp ẩn. Bởi vì chúng chấp nhận đầu vào và tạo giá trị đầu ra trong khoảng từ 0 đến 1, phạm vi xác định xác suất, các hàm này có lợi khi kết quả mong muốn của thuật toán là xác suất. Các lớp ẩn cho phép chức năng của mạng thần kinh được chia thành các sửa đổi dữ liệu cụ thể. Mỗi chức năng trong lớp ẩn được điều chỉnh để tạo ra một kết quả tóm tắt (xem Hình 5.1).

Sự có tổng thể của quá trình bao gồm những điều sau đây:

1. Tính toán tốn thắt sau khi lan truyền chuyển tiếp,
2. Tính độ dốc liên quan đến các pixel của hình ảnh,
3. Di chuyển các điểm ảnh của hình ảnh thật nhẹ theo hướng của các gradien được tính toán để tối đa hóa tốn thắt được tính toán ở trên.

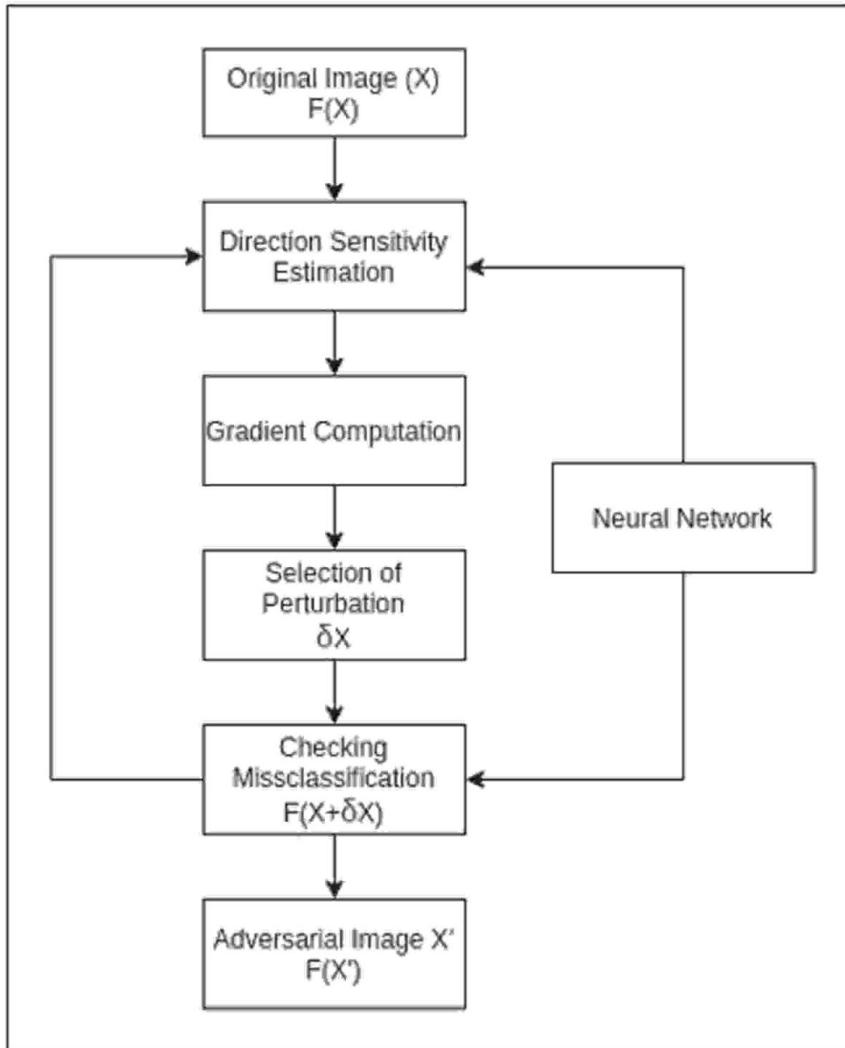
Trong khi tính toán tốn thắt sau khi truyền chuyển tiếp, chúng tôi sử dụng hàm tốn thắt khả năng âm để ước tính dự đoán của mô hình của chúng tôi gần với lớp thực tế như thế nào.

Tính toán độ dốc, chúng tôi xác định [40,41,42] hướng di chuyển các trọng số của bạn để giảm giá trị tốn thắt. Chúng tôi điều chỉnh các pixel hình ảnh đầu vào theo hướng của độ dốc để tối đa hóa giá trị mất mát.

Khi huấn luyện mạng thần kinh, cách phổ biến nhất để xác định hướng điều chỉnh một trọng số cụ thể nằm sâu trong mạng (nghĩa là độ dốc của hàm mất mát liên quan đến trọng số cụ thể đó) là lan truyền ngược các độ dốc ngay từ đầu (đầu ra một phần) đến trọng lượng. Chúng tôi truyền ngược độ dốc từ lớp đầu ra sang hình ảnh đầu vào [43,44,45]. Trong ML, để điều chỉnh các trọng số nhằm giảm giá trị tốn thắt, chúng tôi sử dụng phương trình đơn giản sau:

```
new_weights = old_weights - learning_rate * độ dốc
```

Chúng tôi áp dụng khái niệm tương tự cho FGSM, nhưng chúng tôi muốn tối đa hóa tốn thắt, vì vậy chúng tôi tăng giá trị pixel của hình ảnh theo phương trình sau:



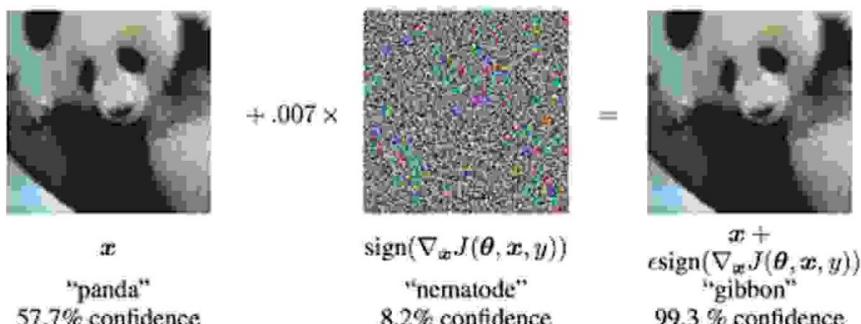
HÌNH 5.1 Lưu đồ.

```
new_pixels = old_pixels + epsilon * độ dốc
```

X đại diện cho hình ảnh đầu vào mà chúng ta muốn mô hình dự đoán sai trong hình trên (xem Hình 5.2).

Phần thứ hai của hình ảnh đại diện cho độ dốc của hàm mất liên quan đến hình ảnh đầu vào.

Hãy nhớ rằng gradient chỉ là một tensor định hướng (nó cung cấp thông tin về hướng di chuyển). Chúng tôi nhân các gradient với một giá trị rất nhỏ, epsilon



HÌNH 5.2 Kết quả của mô hình.

(0,007 trong hình ảnh), để thực thi hiệu ứng thúc đẩy. Sau đó, chúng tôi thêm kết quả vào hình ảnh đầu vào.

Biểu thức đáng báo động dưới hình ảnh kết quả có thể được diễn đạt đơn giản theo cách này:

$$\text{input_image_pixels} + \text{epsilon} * \text{độ dốc của hàm mất đối với} \\ \text{input_image_pixels}$$

5.6 NHIỄM KHUẨN

Đưa ra một phân loại tuyển tính đơn giản

$$w \cdot x$$

trong đó w là ma trận trọng số, chúng ta có thể nghĩ về một ví dụ nghịch cảnh có chứa một nhiễu loạn nhỏ, không thể nhận thấy đối với đầu vào. Hãy để chúng tôi biểu thị sự e ngại là η .

$$x' = x + \eta$$

Sau đó, nhật ký của classifier sẽ là

$$w^T x' = w^T(x + \eta)$$

$$= w^T x + w^T \eta$$

Điều đó có nghĩa là, với một nhiễu loạn nhỏ η , tác động thực tế của nhiễu loạn đối với nhật ký của bộ phân loại được cho bởi $w^T \eta$. Ý tưởng cơ bản đằng sau FGSM là chúng ta có thể tìm thấy một số η gây ra sự thay đổi không thể nhận biết được và bề ngoài có vẻ vô hại đối với mắt người, nhưng lại đủ phá hoại và bất lợi cho bộ phân loại đến mức dự đoán của nó không còn chính xác nữa.

Hãy để chúng tôi đặt điều này vào bối cảnh bằng cách xem xét một ví dụ. Giả sử chúng ta có một số trình phân loại hình ảnh nhận hình ảnh RGB làm đầu vào. Các hình ảnh RGB diễn hình có các giá trị pixel nguyên nằm trong khoảng từ 0 đến 255. Các giá trị này thường được xử lý trước thông qua

Mô hình dịch hình ảnh để chống lại các cuộc tấn công của đối thủ

chia cho 255. Do đó, độ chính xác của dữ liệu bị giới hạn bởi nút cỗ chai tám bit này.

Điều đó có nghĩa là, đối với các nhiễu loạn dưới 1/255, chúng ta không nên mong đợi bộ phân loại đưa ra một dự đoán khác. Nói cách khác, việc bỏ sung $w\eta$ sẽ không làm cho mô hình hoạt động bỏ sung nếu không có bất kỳ nhiễu loạn nào.

Một ví dụ nghịch cảnh tối đa hóa giá trị của $w\eta$ để khiến mô hình đưa ra dự đoán sai. Tất nhiên, có một ràng buộc đối với η ; nếu không, chúng ta chỉ có thể áp dụng một nhiễu loạn lớn cho hình ảnh đầu vào, nhưng sau đó nhiễu loạn có thể hiển thị đủ để thay đổi nhãn sự thật cơ bản. Do đó, chúng tôi áp dụng một ràng buộc sao cho

$$\eta \leq \epsilon \|\eta\| \leq \epsilon$$

Định mức vô cực được định nghĩa là

$$A = \max_{1 \leq i \leq m} |a_{ij}|$$

nói một cách đơn giản hơn, có nghĩa là giá trị tuyệt đối lớn nhất của phần tử trong ma trận hoặc vectơ. Trong bối cảnh này, điều đó có nghĩa là độ lớn lớn nhất của phần tử trong η không vượt quá giới hạn độ chính xác ϵ .

Sau đó, Goodfellow tiến hành giải thích giới hạn tối đa của nhiễu loạn này.

Cụ thể, cho rằng

$$\eta = \epsilon \text{ sign}(w)$$

giới hạn tối đa của việc hủy kích hoạt thay đổi có thể được viết là

$$w - \eta = \epsilon w T \text{ dấu}(w)$$

$$= \epsilon w - 1$$

$$= \epsilon mn$$

trong đó độ lớn trung bình của một phần tử của w được cho bởi m , và $w \in \mathbb{R}^n$.

Nó cho chúng ta biết rằng sự thay đổi trong kích hoạt được cung cấp bởi nhiễu loạn tăng lên sớm liên quan đến n hoặc kích thước. Nói cách khác, trong các ngữ cảnh có đủ chiều cao, chúng ta có thể mong đợi ngay cả một nhiễu loạn nhỏ giới hạn ở ϵ cũng tạo ra một nhiễu loạn đủ lớn để khiến mô hình dễ bị tấn công bởi đối thủ. Các ví dụ nhiễu loạn như vậy được gọi là các ví dụ nghịch cảnh.

5.7 TUYẾT VỜI

Các phương trình trên đã chứng minh rằng mức độ nhiễu loạn tăng lên khi tính đa chiều tăng lên. Nói cách khác, chúng tôi đã xác định rằng việc tạo ra các ví dụ đối lập là có thể thông qua các nhiễu loạn vô cùng nhỏ. Trong phần này, chúng ta hãy đi sâu vào các chi tiết cụ thể của FGSM.

Ý tưởng đằng sau FGSM đơn giản một cách đáng ngạc nhiên: chúng tôi làm ngược lại với giảm dần độ dốc điển hình để tối đa hóa tổn thất do nhầm lẫn mô hình là mục tiêu cuối cùng của một cuộc tấn công đối nghịch. Do đó, chúng tôi coi \mathbf{x} , đầu vào của mô hình, là một tham số có thể huấn luyện được. Sau đó, chúng tôi thêm gradient vào biến đầu vào ban đầu của nó để tạo nhiễu loạn. Về mặt toán học, điều này có thể được thể hiện như sau:

$$\eta = \epsilon \operatorname{sign}(\mathbf{x} \mathbf{J}(\mathbf{w}, \mathbf{x}, \mathbf{y}))$$

trong đó \mathbf{J} đại diện cho hàm chi phí. Sau đó, chúng ta có thể tạo một ví dụ về đối thủ thông qua

$$\mathbf{x}' = \mathbf{x} + \epsilon \operatorname{sign}(\mathbf{x} \mathbf{J}(\mathbf{w}, \mathbf{x}, \mathbf{y}))$$

Đó là mâu chốt của FGSM: chúng tôi sử dụng ký hiệu độ dốc, nhận nó với một giá trị nhỏ nào đó và thêm nhiễu loạn đó vào đầu vào ban đầu để tạo ra một ví dụ nghịch cảnh.

Một cách để xem xét điều này là theo xấp xỉ bậc nhất. nhớ lại rằng

$$\mathbf{f}(\mathbf{x}') = \mathbf{f}(\mathbf{x}) + (\mathbf{x}' - \mathbf{x})^T \mathbf{f}'(\mathbf{x})$$

Trong bối cảnh này, chúng ta có thể coi \mathbf{f}' là hàm chi phí J , sau đó biến thành

$$\mathbf{J}(\mathbf{w}, \mathbf{x}') = \mathbf{J}(\mathbf{x}, \mathbf{w}) + (\mathbf{x}' - \mathbf{x})^T \mathbf{x} \mathbf{J}(\mathbf{w}, \mathbf{x})$$

Sau đó, mục tiêu của một cuộc tấn công đối nghịch là tối đa hóa thuật ngữ thứ hai bổ sung.

Vì có một ràng buộc chuẩn vô hạn đối với nhiễu loạn, cụ thể là

$$\|\mathbf{x}' - \mathbf{x}\| \leq \epsilon$$

với một số suy nghĩ, chúng ta có thể thuyết phục bản thân rằng ví dụ nhiễu loạn tối đa hóa hàm mất mát được đưa ra bởi

$$\mathbf{x}' = \mathbf{x} + \epsilon \operatorname{sign}(\mathbf{x} \mathbf{J}(\mathbf{w}, \mathbf{x}, \mathbf{y}))$$

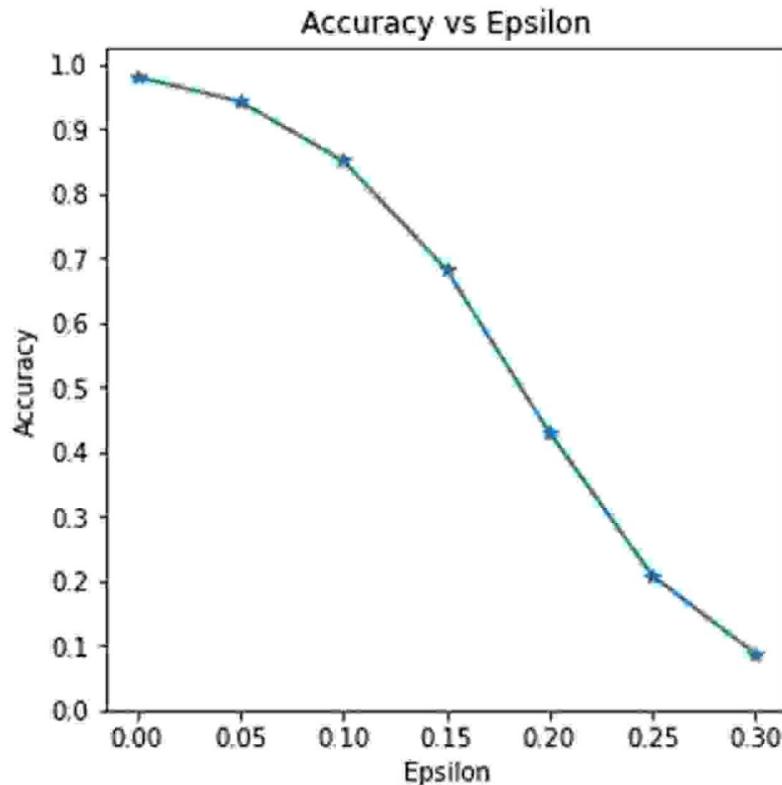
5.8 KẾT QUẢ VÀ THẢO LUẬN

5.8.1 Cài đặt tấn công

Chúng tôi tập trung vào độ khó của một cuộc tấn công đối nghịch vào hình ảnh MNIST. Độ khó của cuộc tấn công được đo bằng độ nhiễu tối đa ít nhất cần thiết để hầu hết (ví dụ: >99%) cuộc tấn công thành công. Cụ thể, chúng tôi thay đổi kích thước nhiễu tối đa của phạm vi $[0, .05, .1, .15, .2, .25, .3]$ và trực quan hóa sự giảm độ chính xác của mô hình trên các ví dụ đối nghịch trong Hình 5.3.

5.9 PHÂN TÍCH KẾT QUẢ

FGSM hoạt động bằng cách sử dụng độ dốc của mạng thần kinh để tạo ra một ví dụ đối nghịch. Đối với một hình ảnh đầu vào, phương pháp này sử dụng độ dốc của sự mất mát liên quan đến



HÌNH 5.3 Biểu đồ đường.

hình ảnh đầu vào để tạo ra một hình ảnh mới tối đa hóa sự mất mát. Hình ảnh mới này được gọi là hình ảnh nghịch cảnh. Nó có thể được tóm tắt bằng biểu thức sau:

$$\text{adv_x} = \text{x} + \epsilon \cdot \text{sign}(\text{xJ}(\theta, \text{x}, \text{y}))$$

Ở đây

- adv_x : Hình ảnh nghịch cảnh.
- x : Ảnh đầu vào
- ϵ : Hỗn số nhân để đảm bảo nhiễu loạn nhỏ.
- $\text{sign}(\text{xJ}(\theta, \text{x}, \text{y}))$: Tồn thắt.

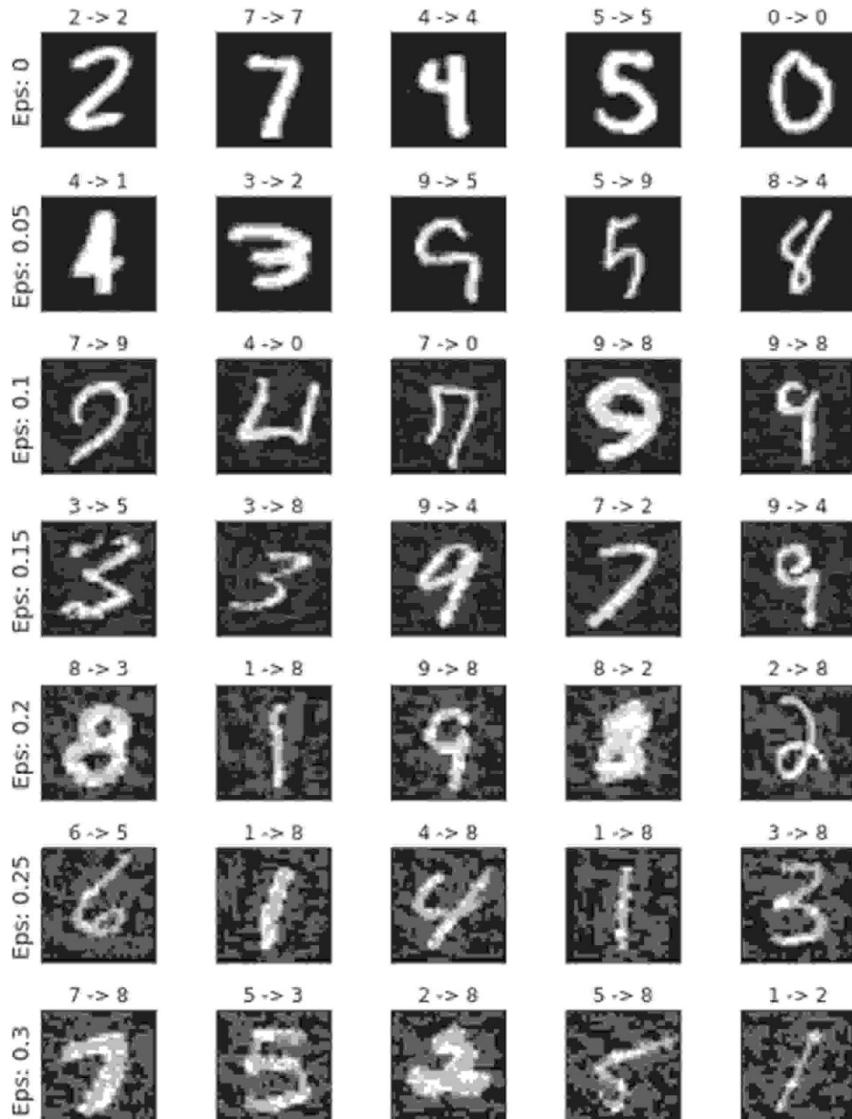
Thông số mô hình. • J

• :

Tồn thắt.

Một thuộc tính hấp dẫn ở đây là độ dốc được thực hiện liên quan đến hình ảnh đầu vào. Nó được thực hiện vì mục tiêu là tạo ra một hình ảnh tối đa hóa tổn thất.

Một phương pháp để thực hiện điều này là tìm xem mỗi pixel trong ảnh đóng góp bao nhiêu vào giá trị mờ mịt và thêm nhiễu tương ứng. Nó hoạt động khá nhanh vì có thể dễ dàng tìm ra cách mỗi pixel đầu vào góp phần vào sự mờ mịt bằng cách sử dụng quy tắc chuỗi và tìm độ dốc cần thiết. Do đó, độ dốc được thực hiện liên quan đến hình ảnh. Ngoài ra, do mô hình không còn được đào tạo (do đó, độ dốc không được tính liên quan đến các biến có thể đào tạo, tức là các tham số mô hình), nên các tham số mô hình không đổi. Mục tiêu duy nhất là đánh lừa một người mẫu đã được đào tạo.



HÌNH 5.4 Ví dụ về các mẫu đối thủ.

Chúng tôi đã thử các giá trị epsilon khác nhau, quan sát hình ảnh thu được và nhận thấy rằng khi giá trị của epsilon tăng lên, việc đánh lừa mạng trở nên dễ dàng hơn. Tuy nhiên, đây là một sự đánh đổi, dẫn đến nhiều loạn trở nên dễ nhận biết hơn.

Chúng ta có thể thấy rằng giá trị của epsilon càng lớn, việc học mô hình bị suy giảm và bắt đầu phân loại sai các chữ số.

Cuối cùng, chúng tôi đã vẽ một số ví dụ về các mâu nghịch cảnh ở mỗi epsilon (như thể hiện trong Hình 5.4).

5.10 TÓM TẮT

Chương này nhằm mục đích phát triển một mô hình để chống lại việc sử dụng sai mô hình thế hệ sâu bằng cách sử dụng các cuộc tấn công đối nghịch để tạo ra các cảnh báo tinh vi có thể khiến các thuật toán thế hệ sâu không thể tạo ra hình ảnh giả ngay từ đầu.

Một trong những mục tiêu chính của các cuộc tấn công đối nghịch vào mạng thần kinh là phân loại sai, trong đó chúng ta chỉ muốn thúc đẩy mô hình đưa ra dự đoán sai mà không cần lo lắng về loại dự đoán thực tế; trong thử nghiệm của chúng tôi, chúng tôi có thể thấy rằng khi giá trị epsilon bằng 0 ($\epsilon=0$), quá trình đào tạo diễn ra bình thường và với mỗi lần tăng giá trị của epsilon (ϵ), quá trình học mô hình bị suy giảm và cuối cùng bắt đầu phân loại sai các chữ số.

NGƯỜI GIỚI THIỆU

- [1] Plangger, K. (2020). DFs: Quan điểm về "thực tế" trong tương lai của quảng cáo và xây dựng thương hiệu. Tạp chí Quảng cáo Quốc tế. <https://doi.org/10.1080/02650487.2020.1834211>
- [2] Neekhara, P., Dolhansky, B., Bitton, J., & Ferrer, CC (1970, ngày 1 tháng 1). Các mối đe dọa đối nghịch đối với việc phát hiện DF: Một góc nhìn thực tế. Truy cập mở CVF. Truy cập ngày 11 tháng 1 năm 2022, từ https://openaccess.thecvf.com/content/CVPR2021W/WMF/html/Neekhara_Adversarial_Threats_to_DF_Detection_A_Practical_Perspective_CVPRW_2021_paper.html
- [3] Akhtar, N., & Mian, A. (2018). "Mối đe dọa của các cuộc tấn công đối nghịch vào DL trong thị giác máy tính: Một cuộc khảo sát," trong IEEE Access, tập. 6, trang 14410-14430. doi:10.1109/TRUY CẬP.2018.2807385
- [4] Krizhevsky, A., Sutskever, I., & Hinton, GE (2012). Phân loại ImageNet với Mạng nơ-ron tích chập sâu. Trong F. Pereira, CJC Burges, L. Bottou, & K. Q. Weinberger (Quý Đô), Những tiến bộ trong Hệ thống Xử lý Thông tin Thần kinh (Tập 25). Văn phòng điều hành, <https://proceedings.neurips.cc/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf>
- [5] He, K., Zhang, X., Ren, S., & Sun, J. (2015, ngày 6 tháng 2). Đè sâu vào các bộ chỉnh lưu: Vượt qua hiệu suất ở cấp độ con người trong phân loại ImageNet. arXiv.org. Truy cập ngày 11 tháng 1 năm 2022, từ <https://arxiv.org/abs/1502.01852>
- [6] Mý, RCNECL, Collobert, R., Mý, NECL, Mý, JWN ECL, Weston, J., University, CM, Amherst, U. of M., Google, U. of T. and, & Metrics, OMVA (2008, ngày 1 tháng 7). Kiến trúc hợp nhất để xử lý ngôn ngữ tự nhiên: Mạng lưới thần kinh sâu với khả năng học tập đa nhiệm. Kiến trúc thống nhất cho xử lý ngôn ngữ tự nhiên | Kỷ yếu của Hội nghị quốc tế lần thứ 25 về ML. Truy cập ngày 11 tháng 1 năm 2022, từ <https://dl.acm.org/doi/10.1145/1390156.1390177>

- [7] Biggio, B., Corona, I., Maiorca, D., Nelson, B., Srndic, N., Laskov, P., Giacinto, G., & Roli, F. (2017, ngày 21 tháng 8). Các cuộc tấn công trốn tránh chống lại ML tại thời điểm thử nghiệm. arXiv.org. Truy cập ngày 11 tháng 1 năm 2022, từ <https://arxiv.org/abs/1708.06131>
- [8] Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., & Fergus, R. (2014, ngày 19 tháng 2). Các thuộc tính hấp dẫn của mạng lưới thần kinh. arXiv.org. Truy cập ngày 11 tháng 1 năm 2022, từ <https://arxiv.org/abs/1312.6199>
- [9] Nguyen, A., Yosinski, J., & Clune, J. (2015, 2 tháng 4). Mạng lưới thần kinh sâu dễ dàng bị đánh lừa: Dự đoán có độ tin cậy cao đối với hình ảnh không thể nhận dạng được. arXiv.org. Truy cập ngày 11 tháng 1 năm 2022, từ <https://arxiv.org/abs/1412.1897>
- [10] Moosavi-Dezfooli, S.-M., Fawzi, A., & Frossard, P. (2016, ngày 4 tháng 7). DeepFool: Một phương pháp đơn giản và chính xác để đánh lừa Deep Neural Networks. arXiv.org. Truy cập ngày 11 tháng 1 năm 2022, từ <https://arxiv.org/abs/1511.04599>
- [11] Dong, Y., Liao, F., Pang, T., Su, H., Zhu, J., Hu, X., & Li, J. (1970, 1 tháng 1). Tăng cường các cuộc tấn công đổi thủ với đà. Truy cập mở CVF. Truy cập ngày 11 tháng 1 năm 2022, từ https://openaccess.thecvf.com/content_cvpr_2018/html/Dong_Boosting_Adversarial_Attacks_CVPR_2018_paper.html
- [12] Madry, A., Makelov, A., Schmidt, L., Tsipras, D., & Vladu, A. (2019, ngày 4 tháng 9). Hướng tới các mô hình DL chống lại các cuộc tấn công của đối thủ. arXiv.org. Truy cập ngày 10 tháng 1 năm 2022, từ <https://arxiv.org/abs/1706.06083>
- [13] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., WardeFarley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Lưới đổi thủ sáng tạo. Trong Những tiến bộ trong Hệ thống Xử lý Thông tin Thần kinh, trang 2672-2680.
- [14] Odena, A., Olah, C., & Shlens, J. (2017, ngày 20 tháng 7). Tổng hợp hình ảnh có điều kiện với bộ phân loại phụ trợ Gans. arXiv.org. Truy cập ngày 11 tháng 1 năm 2022, từ <https://arxiv.org/abs/1610.09585>
- [15] Gilmer, J., Ford, N., Carlini, N., & Cubuk, E. (2019, ngày 24 tháng 5). Các ví dụ về đổi thủ là hệ quả tự nhiên của lỗi thử nghiệm trong tiếng ồn. PMLR. Truy cập ngày 11 tháng 1 năm 2022, từ <http://proceedings.mlr.press/v97/gilmer19a.html>
- [16] Xiao, C., Zhu, J.-Y., Li, B., He, W., Liu, M., & Song, D. (2018, ngày 9 tháng 1). Các ví dụ về đổi thủ được biến đổi không gian. arXiv.org. Truy cập ngày 11 tháng 1 năm 2022, từ <https://arxiv.org/abs/1801.02612>
- [17] Zhao, Z., Dua, D., & Singh, S. (2018, ngày 23 tháng 2). Tạo các ví dụ quảng cáo tự nhiên. arXiv.org. Truy cập ngày 11 tháng 1 năm 2022, từ <https://arxiv.org/abs/1710.11342>
- [18] Ruiz, N., Bargal, SA, & Sclaroff, S. (2020, ngày 27 tháng 4). Phá vỡ các DF: Các cuộc tấn công đổi nghịch chống lại các mạng dịch hình ảnh có điều kiện và các hệ thống thao tác trên khuôn mặt. arXiv.org. Truy cập ngày 5 tháng 1 năm 2022, từ <https://arxiv.org/abs/2003.01279>
- [19] Qiu, S., Liu, Q., Zhou, S., & Wu, C. (2019). Đánh giá các công nghệ phòng thủ và tấn công đổi thủ AI. Khoa học Ứng dụng, Vol. 9, số 5, tr. 909. MDPI AG. Lấy từ <http://dx.doi.org/10.3390/app9050909>
- [20] Finlayson, SG, Chung, HW, Kohane, IS, & Beam, AL (2019, ngày 4 tháng 2). Các cuộc tấn công đổi nghịch chống lại các hệ thống DL y tế. arXiv.org. Truy cập ngày 6 tháng 1 năm 2022, từ <https://arxiv.org/abs/1804.05296>
- [21] Samangouei, P., Kabkab, M., & Chellappa, R. (2018, ngày 18 tháng 5). Defense-GAN: Bảo vệ các bộ phân loại chống lại các cuộc tấn công của đối thủ bằng cách sử dụng các mô hình tổng quát. arXiv.org. Truy cập ngày 6 tháng 1 năm 2022, từ <https://arxiv.org/abs/1805.06605>
- [22] Santhanam, GK, & Grnarova, P. (27/5/2018). Bảo vệ chống lại các cuộc tấn công của đối thủ bằng cách tận dụng toàn bộ GAN. arXiv.org. Truy cập ngày 6 tháng 1 năm 2022, từ <https://arxiv.org/abs/1805.10652>

Mô hình dịch hình ảnh để chống lại các cuộc tấn công của đối thủ

- [23] Hu, W., & Tan, Y. (2017, ngày 20 tháng 2). Tạo các ví dụ về phần mềm độc hại đối nghịch cho các cuộc tấn công hợp đồng dựa trên GAN. arXiv.org. Truy cập ngày 6 tháng 1 năm 2022, từ <https://arxiv.org/abs/1702.05983>
- [24] Gandhi, A. & S. Jain, "Adversarial Perturbations Đánh lừa Máy dò DF," Hội nghị Chung Quốc tế về Mạng nơ-ron năm 2020 (IJCNN), 2020, trang 1-8. doi:10.1109/IJCNN48605.2020.9207034.
- [25] Liu, X., & Hsieh, C.-J. (2019, ngày 15 tháng 4). Rob-Gan: Kẻ tạo ra, kẻ phân biệt đối xử và kẻ tấn công đối thủ. arXiv.org. Truy cập ngày 11 tháng 1 năm 2022, từ <https://arxiv.org/abs/1807.10454>
- [26] Croce, F., & Hein, M. (2019, tháng 10). Các cuộc tấn công đối thủ thua thót và không thể nhận ra. Ký yếu của Hội nghị Quốc tế IEEE/CVF về Tâm nhìn Máy tính (ICCV).
- [27] Zhang, H., & Wang, J. (2019). Phòng thủ chống lại các cuộc tấn công của đối thủ bằng cách sử dụng Huấn luyện đối thủ dựa trên phân tán tính năng. Trong H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, & R. Garnett (eds.), Nhũng tiến bộ trong Hệ thống xử lý thông tin thần kinh (Tập 32). Văn phòng điều hành, <https://proceedings.neurips.cc/paper/2019/file/d8700cbd38cc9f30cecb34f0c195b137-Paper.pdf>
- [28] Liao, F., Liang, M., Dong, Y., Pang, T., Hu, X., & Zhu, J. (2018). Bảo vệ chống lại các cuộc tấn công của đối thủ bằng cách sử dụng bộ khử nhiễu có hướng dẫn đại diện cấp cao. arXiv [cs.CV]. Opgehaal van, <http://arxiv.org/abs/1712.02976>
- [29] Mustafa, A., Khan, SH, Hayat, M., Shen, J., & Shao, L. (2020). Hình ảnh siêu phân giải như một biện pháp bảo vệ chống lại các cuộc tấn công của đối thủ. Giáo dịch của IEEE về Xử lý hình ảnh, Tập. 29, trang 1711-1724. doi:10.1109/TIP.2019.2940533
- [30] Song, S., Chen, Y., Cheung, N.-M., & Kuo, C.-CJ (2018). Phòng thủ chống lại các cuộc tấn công của đối thủ với Saak Transform. arXiv [cs.CV]. Opgehaal van, <http://arxiv.org/abs/1808.01785>
- [31] Sharma, DK, Gaur, L., & Okunbor, D. (2007). "Nén hình ảnh và trích xuất đặc trưng bằng mạng thần kinh," Ký yếu của Viện Khoa học Thông tin và Quản lý, 11(1), 33-38.
- [32] Anshu, K., Gaur, L., & Khazanchi, D. (2017) "Đánh giá mức độ hài lòng của các nhà bán lẻ điện tử tạp hóa bằng cách sử dụng mô hình TOPSIS và ECCSI mờ trực giác," Hội nghị quốc tế về Công nghệ Infocom và Hệ thống không người lái (Xu hướng và Định hướng Tương lai) (ICTUS), trang 276-284. doi:10.1109/ICTUS.2017.8286019
- [33] Gaur, L., & Anshu, K. (2018). Phân tích sở thích của người tiêu dùng cho các trang web sử dụng e-TailQ và AHP. Tạp chí Kỹ thuật & Công nghệ Quốc tế, Vol. 7, số 2.11, trang 14-20.
- [34] Rana, J., Gaur, L., Singh, G., Awan, U. & Rasheed, MI (2021). Củng cố hành trình của khách hàng thông qua trí tuệ nhân tạo: Chương trình nghiên cứu và đánh giá. Tạp chí quốc tế về các thị trường mới nổi, Tập. trước khi in (No. trước khi in). <https://doi.org/10.1108/IJOEM-08-2021-1214>
- [35] Gaur, L., Singh, G., Solanki, A., Jhanjhi, NZ, Bhatia, U., Sharma, S., . & Kim, W. (2021), "Sứ mệnh của thanh niên trong việc dự đoán sự phát triển bền vững các mục tiêu bằng cách sử dụng thuật toán rừng ngẫu nhiên và thần kinh mờ" Khoa học thông tin và điện toán lấy con người làm trung tâm, 11, NA.
- [36] Singh, G., Kumar, B., Gaur, L., & Tyagi, A. (2019). "So sánh giữa Multinomial và Bernoulli Naïve Bayes để phân loại văn bản," Hội nghị quốc tế về tự động hóa, tính toán và quản lý công nghệ (ICACTM), trang 593-596. doi:10.1109/ICACTM.2019.8776800
- [37] Gaur, L., Agarwal, V., & Anshu, K. (2020). Phương pháp tiếp cận DEMATEL mờ để xác định các yếu tố ảnh hưởng đến hiệu quả của ngành bán lẻ Ấn Độ. Trong PK Kapur, O. Singh, S. Khatri, & A. Verma (eds.) Đảm bảo Hệ thống Chiến lược và Phân tích Kinh doanh. Phân tích nội dung

- (Quản lý Hiệu suất và An toàn). Springer, Singapore. 69-83. <https://doi.org/10.1007/978-981-15-3647-2>
- [38] Gaur, L., Afaq, A., Singh, G. & Dwivedi, YK (2021). Vai trò của trí tuệ nhân tạo và người máy trong việc thúc đẩy du lịch không chạm trong thời kỳ đại dịch: Chương trình nghiên cứu và đánh giá. Tạp chí Quốc tế về Quản lý Khách sạn Đương đại, Vol. 33, Số 11, trang 4079-4098. <https://doi.org/10.1108/IJCHM-11-2020-1246>
- [39] Sharma, S., Singh, G., Gaur, L., & Sharma, R. (2022). Khoảng cách tâm lý và tôn giáo có ảnh hưởng đến hành vi lừa đảo của khách hàng không? Tạp chí Nghiên cứu Người tiêu dùng Quốc tế. doi:10.1111/ijcs.12773
- [40] Sahu, G., Gaur, L., & Singh, G. (2021). Áp dụng phương pháp tiếp cận lý thuyết thích hợp và hài lòng để kiểm tra niềm đam mê của người dùng đối với các nền tảng vượt trội và truyền hình thông thường. Viễn thông và Tin học, 65. doi:10.1016/j.tele.2021.101713
- [41] Ramakrishnan, R., Gaur, L., & Singh, G. (2016). Tính khả thi và hiệu quả của các thiết bị IoT đèn hiệu BLE trong quản lý hàng tồn kho tại khu vực cửa hàng. Tạp chí Quốc tế về Kỹ thuật Điện và Máy tính, Vol. 6, Số 5, trang 2362-2368. doi:10.11591/ijece.v6i5.10807
- [42] Afaq, A., Gaur, L., Singh, G., & Dhir, A. (2021). COVID-19: Thay đổi hành vi của hành khách đi máy bay và định hình lại kỳ vọng của họ đối với ngành hàng không. Nghiên cứu giải trí du lịch. doi:10.1080/02508281.2021.2008211
- [43] Mahbub, Md. K., Biswas, M., Gaur, L., Alenezi, F., & Santosh, K. (2022). Các tính năng sâu để phát hiện các bất thường về phổi trên phim X-quang phổi do bệnh truyền nhiễm X: Covid-19, viêm phổi và bệnh lao. Khoa học thông tin, Vol. 592, trang 389-401. <https://doi.org/10.1016/J.INS.2022.01.062>
- [44] Sharma, S., Singh, G., Gaur, L., & Sharma, R. (2022). Khoảng cách tâm lý và tôn giáo có ảnh hưởng đến hành vi gian lận của khách hàng không? Tạp chí Nghiên cứu Người tiêu dùng Quốc tế. 10.1111/ijcs.12773.
- [45] Zaman, N., & Gaur, L. (2022). Phương pháp tiếp cận và ứng dụng của Deep Learning trong chăm sóc y tế áo. IGI. doi:10.4018/978-1-7998-8929-8.ch002.

6 Phát hiện DeepFake

Sử dụng các tính năng cục bộ và thần kinh tích chập

Mạng

Shreya Rastogi, Amit Kumar Mishra, và
Bò tót Loveleen

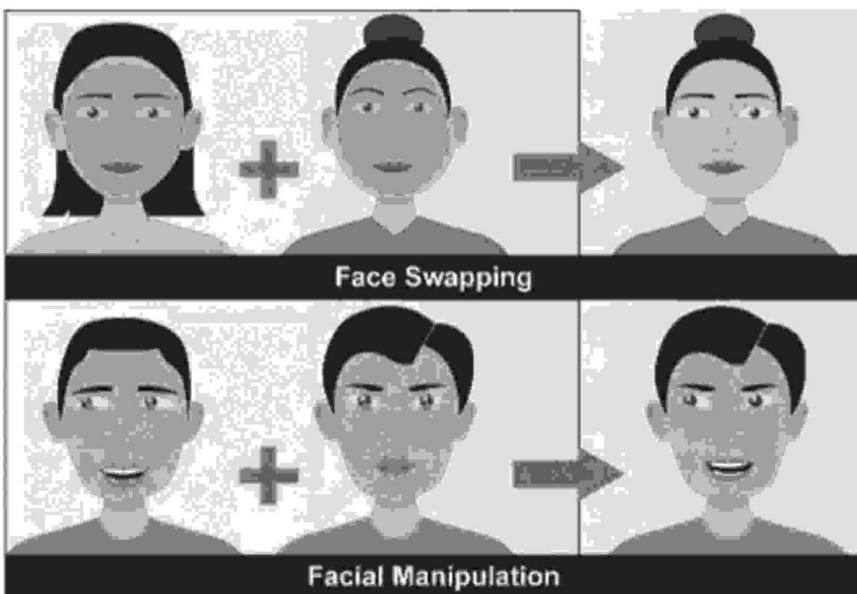
NỘI DUNG

6.1 Giới thiệu	73
6.2 Lịch sử của DF	74
6.3 Các ứng dụng của DeepFake	75
6.4 Ưu điểm của DeepFake	77
6.5 Nhược điểm của DeepFake	77
6.6 Tạo DeepFakes	77
6.7 Các phương pháp phát hiện DeepFake	80
6.8 Các tính năng cục bộ và toàn cục của hình ảnh	82
6.8.1 Phát hiện DeepFakes bằng các tính năng cục bộ	86
6.9 Tóm tắt	86
Người giới thiệu	86

6.1 GIỚI THIỆU

DeepFake (DF) nổi tiếng có thể nhanh chóng tạo hoặc thay đổi khuôn mặt của người trong ảnh và bản ghi bằng các kỹ thuật dựa trên công nghệ học sâu (DL).

Đó là một trong những hiện tượng phát triển nhanh chóng [1]. Có thể đạt được kết quả nổi bật bằng cách sản xuất các mặt hàng da phương tiện mới khó phân biệt là thật hay giả bằng mắt thường. Tuy nhiên, thuật ngữ DF đề cập đến tất cả các tài liệu nghe nhìn đã được xử lý tổng hợp hoặc tạo ra bằng cách sử dụng các mô hình tổng quát ML [2]. Nhờ công nghệ mới, bất kỳ ai cũng có thể tạo DF từ một vài bức ảnh, vì vậy phim giả sẽ vượt ra ngoài phạm vi nổi tiếng và thúc đẩy khiêu dâm trả thù. Danielle Citron, giáo sư luật tại Đại học Boston [3] cho biết: "Nology tech DF đang được vũ khí hóa để chống lại phụ nữ. Những lo ngại về hậu quả có hại của DF đã làm dậy lên sự tò mò về việc nhận biết DF [4]. DF đòi hỏi phải biến khuôn mặt của một người thành khuôn mặt của người khác theo những cách mà một người biên tập viên sẽ không xem xét hoặc không thể nhận thấy [5]. Các khuôn mặt thường xuyên bị tráo đổi hoặc các nét mặt được thao tác trong phim DF, như trong Hình 6.1. Mặt ở đây bên trái được trao đổi với



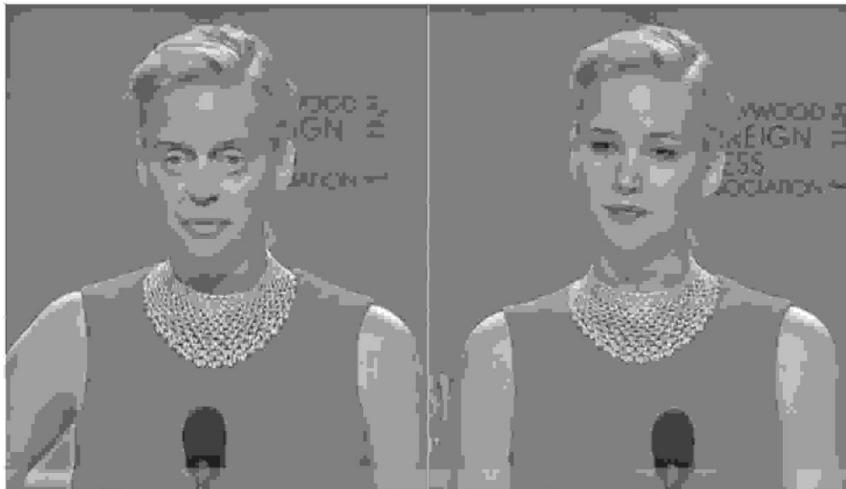
HÌNH 6.1 Hoán đổi và thao tác khuôn mặt trong video DF [6].

cơ thể của một số cá nhân khác. Các đặc điểm của khuôn mặt bên trái được sao chép bằng khuôn mặt bên phải trong thao tác trên khuôn mặt [6]. Trong Hình 6.2, thao tác trên khuôn mặt của người nổi tiếng là một ví dụ về cách hoạt động của DF trong đời thực.

ML là yếu tố quan trọng của DF và nó đã cho phép chúng được sản xuất nhanh hơn đáng kể và với giá rẻ hơn [7]. Để xây dựng một bộ phim DF về bất kỳ ai, trước hết, nhà phát triển sẽ chuẩn bị một mạng nơ-ron trên video giám sát thực tế kéo dài vài phút về người cung cấp cho nó "sự hiểu biết" thực tế về cách họ hoặc xuất hiện từ nhiều góc độ và điều kiện ánh sáng khác nhau. Mô hình thần kinh sau đó sẽ được kết hợp với các hiệu ứng hình ảnh để phủ một bão sao của con người lên trên một người khác [8]. Mặc dù việc kết hợp AI tăng tốc quá trình, nhưng vẫn cần thời gian để tạo ra một hỗn hợp hợp lý đặt con người vào một tình huống tưởng tượng. Để tránh trực trặc và biến dạng trong hình ảnh, nhà thiết kế phải thay đổi một số cài đặt của hệ thống được đào tạo phù hợp riêng lẻ [9]. Sau phần giới thiệu, lịch sử và ứng dụng của DF sẽ được thảo luận. Tiếp theo là những ưu điểm và nhược điểm của DF. Sau đó, các kỹ thuật tạo và phát hiện DF được giải thích.

6.2 LỊCH SỬ CỦA DF

Christoph Bregler, Michele Covell và Malcolm Slaney đã tạo ra phần mềm Visual Rewrite Offsite Link vào năm 1997, phần mềm này đã chuyển đổi cảnh quay video hiện có của một diễn giả để phản ánh các cụm từ có trong một tín hiệu âm nhạc riêng biệt. Lần đầu tiên, đổi mặt



HÌNH 6.2 Hình bên trái là hình gốc và hình bên phải được sản xuất bằng công nghệ DF [10].

reanimation hoàn toàn tự động [11]. Để đạt được điều này, nó đã sử dụng các kỹ thuật ML để tương quan tiếng ôn do nhân vật chính trong phim tạo ra với hình dạng đầu của họ. Phần mềm này được tạo ra để sử dụng trong chuyển mã điện ảnh, cho phép các cử chỉ trên khuôn mặt của người biểu diễn khớp với một chủ đề động. Quá trình tổng hợp thị giác của con người được cải thiện khi thị giác máy tính và AI trưởng thành. Bằng cách sử dụng phương pháp ML được gọi là Mạng đối thủ sáng tạo (GAN), Liên kết ngoại vi cho phép nhiều người hơn áp đặt ánh và phim có sẵn lên ảnh hoặc video thô. DF Offsite Link, sự kết hợp giữa "học sâu" và "giả mạo", được phát minh vào năm 2017 [12].

6.3 ỨNG DỤNG CỦA DEEPFAKES

Có nhiều ứng dụng khác nhau của DF như sau:

1. Giáo dục

DF có thể giúp giáo viên đưa ra những bài giảng hấp dẫn. Hơn nữa, những bài giảng này sẽ vượt qua các phương tiện đồ họa và đa phương tiện tiêu chuẩn. Trong nghiên cứu, phương tiện tổng hợp do AI phát triển thực sự có thể đưa người cổ đại trở thành hiện thực. Các lớp học trở nên thú vị hơn và kết quả là có sự tham gia. Một bộ phim tái hiện hoặc âm thanh và video về một nhân vật thần thoại cũng sẽ có tác động đáng kể hơn. Nó có thể thúc đẩy sự nhiệt tình và làm cho việc giảng dạy trở nên vô cùng hữu ích. Việc sử dụng lời nói và video tổng hợp ở cấp độ toàn cầu và hợp lý có thể làm tăng hiệu quả và đạt được mục tiêu [13].

2. Nghệ thuật

DF có thể làm cho các công nghệ "VFX" đắt tiền trở nên dễ tiếp cận hơn. Nó cũng có thể là một công cụ tuyệt vời cho những người sáng tạo tự do với một phần chi phí đáng kể. DFs có thể là một kỹ thuật hiệu quả để đưa sự hài hước hoặc châm biếm vào cuộc sống một cách thực tế. Nó có thể là sự tái hiện, mở rộng, biến dạng hoặc khai thác các sự kiện trong đời thực.

Phương tiện tổng hợp do AI tạo ra cũng có nhiều hứa hẹn. Nó có thể mở ra những cánh cửa trong thế giới giải trí. Một số nhà sản xuất cá nhân và người dùng YouTube cũng đang sử dụng khả năng này [14].

3. Giải phóng và tự chủ

Những người ủng hộ nhân quyền và phóng viên có thể sử dụng các phương tiện truyền thông tổng hợp ở các quốc gia độc đoán và khắc nghiệt để ẩn danh. Bằng cách sử dụng công nghệ để tố giác tội phạm trên các trang truyền thống hoặc mạng xã hội, các nhà báo và nhà hoạt động công dân có thể có nhiều ảnh hưởng. DF đôi khi được sử dụng để bảo vệ quyền của mọi người bằng cách che giấu âm thanh và tính năng của họ [13].

4. Không có rào cản ngôn ngữ.

DF là một cách tiếp cận tuyệt vời để sử dụng công nghệ AI nhằm vượt qua rào cản ngôn ngữ. Quảng cáo mang tính biểu tượng trong đó David Beckham đọc thuộc lòng chín ngôn ngữ khác nhau để truyền đạt một tuyên bố cho Sáng kiến Phải chết vì Sốt rét là một minh họa tuyệt vời về những gì DF có thể đạt được [14].

5. Nghiên cứu

Công nghệ DF cũng có thể tạo ra các bản sao được cá nhân hóa ngoài thông tin giải trí và học tập. Sau đó, chúng có thể được triển khai trong các ứng dụng cho phép mọi người mặc nhiều quần áo hoặc cắt tóc một cách thoải mái và các ứng dụng học tập chuyên biệt. Một vài trong số các lĩnh vực này là y học, nơi đổi mới sáng tạo đã được sử dụng để tạo ra hình ảnh thần kinh "giả" dựa trên dữ liệu bệnh nhân thực tế. Những hình ảnh hư cấu này đang được sử dụng để xây dựng các chương trình phát hiện khối u trong hình ảnh thực tế [15].

6. Tiết kiệm chi phí

Một số ý kiến cho rằng đổi mới sáng tạo mang lại khả năng vượt trội để cách mạng hóa các lĩnh vực khác nhau. Công nghệ chung có thể giúp mọi người và doanh nghiệp tham gia vào các ngành này với chi phí ít hơn bằng cách cho phép phát triển mọi thứ từ phim ảnh đến quảng cáo và trò chơi với chi phí thấp.

Sự đổi mới không gây hại như các ứng dụng trong tương lai. Nếu các tổ chức sử dụng công nghệ DF tuân thủ các tiêu chuẩn đạo đức cao và tránh được các ứng dụng nguy hiểm một cách hiệu quả, thì công nghệ này có rất nhiều tiềm năng.

6.4 ƯU ĐIỂM CỦA DEEPFAKE

Có nhiều ưu điểm khác nhau của DF như:

1. Cho phép các quảng cáo trên toàn thế giới được điều chỉnh và “dịch chính xác” cho các tên thương hiệu và phương ngữ bản địa.
2. Tăng cường các quy trình riêng tư bằng cách cho phép các giám đốc giao tiếp với nhân viên của họ trên toàn thế giới bằng phương ngữ bản địa lý tưởng.
3. Người ta có thể đóng phim ngay cả khi họ chết ngoài đời thực.
4. Làm một bộ phim sử dụng phiên bản trẻ hơn của chính bạn hoặc một diễn viên khác vào vai chính [16,17].

6.5 NHƯỢC ĐIỂM CỦA DEEPFAKE

Có nhiều nhược điểm khác nhau của DF như:

1. Lừa đảo ở cấp quản lý

Các cuộc tấn công DF là loại tấn công phổ biến nhất. Những kẻ giả mạo không còn có gắng thuyết phục một thành viên của công ty chuyển tiền qua email giả mạo. Họ thuyết phục họ bằng một cuộc trò chuyện qua điện thoại từ một người có vẻ giống như một nhân viên tài chính hoặc giám đốc.

2. Ép buộc tài trợ từ các công ty hoặc người tiêu dùng

Khuôn mặt và âm thanh được sao chép vào tệp phương tiện bằng DF tiết lộ những cá nhân tạo điểm đánh giá giả. Thật dễ dàng để ghi lại một CEO đưa ra những tuyên bố hư cấu. Việc cố tình kích động họ xò rỉ phim cho các cơ quan báo chí hoặc phát trên mạng xã hội có thể được sử dụng để ép buộc một công ty.

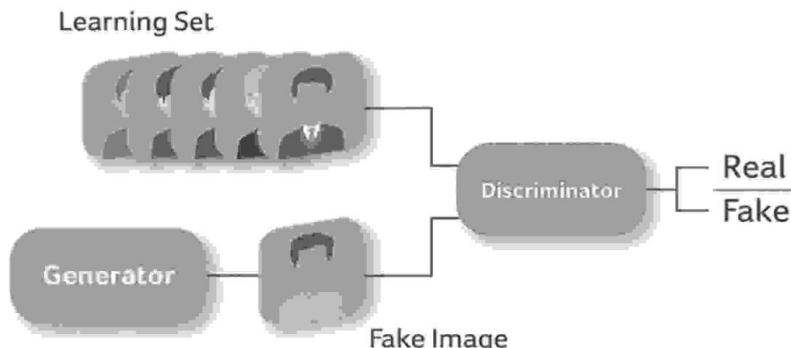
3. Người xem bị quan sát không chấp nhận hoặc đánh giá cao bất kỳ thông tin hoặc bằng chứng nào bắt gặp trên phim hoặc truyền hình.
4. Kiểm soát một sinh trắc học

6.6 THẾ HỆ DEEPFAKES

Việc tạo DF có thể được thực hiện bằng cách sử dụng các kỹ thuật sau:

1. Sử dụng GAN

GAN được sử dụng để xây dựng DF. Chúng có vẻ rất thuyết phục vì mục đích duy nhất của khung GAN là xây dựng một khung có thể đánh lừa nó tin tưởng vào tính hợp pháp của nó. Như thể hiện trong Hình 6.3, một thế hệ trong kiến trúc tạo ra các bức tranh và sau đó, bộ phân biệt sẽ kiểm tra sự khác biệt. Trình tạo hiểu biết bắt cứ điều gì mà bộ phân biệt đối xử có vẻ không thích và thay đổi để tạo ra bản giả tốt hơn đáng kể bằng cách sử dụng ML [18].



HÌNH 6.3 Hoạt động của GAN [18].

2. Sử dụng Autoencoder

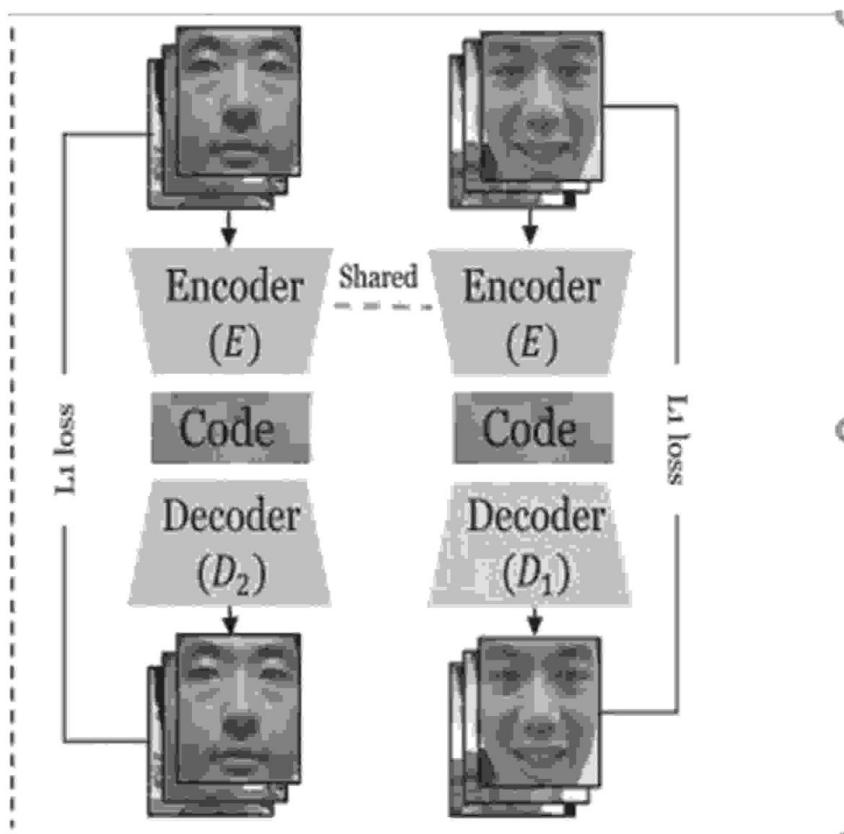
Khả năng của DL để biểu diễn thông tin phức tạp và nhiều chiều đã được biết rõ [19]. Bộ mã hóa tự động sâu, một loại mạng sâu có khả năng như vậy, đã được sử dụng rộng rãi để giảm kích thước và nén hình ảnh [20].

Yuezun Li et al. [21], trong nghiên cứu của họ, đã giải thích một kỹ thuật tạo ra DF. Hình 6.4 mô tả quy trình làm việc hoàn chỉnh của quá trình tạo DF đơn giản. Khuôn mặt của đối tượng được nhận dạng trong một clip sắp tới và các điểm đặc trưng được thu thập sau đó một cách nhanh chóng. Các mốc được sử dụng để phối hợp các tính năng theo một mẫu nhất quán. Sau đó, các tính năng được liên kết được thay đổi kích thước và gửi qua bộ mã hóa tự động để tạo ra các bức ảnh của người đóng góp với các biểu hiện cảm xúc gần giống như các tính năng của người dùng ban đầu. Bộ mã hóa tự động thường được tạo thành từ hai Mạng thần kinh chuyển đổi (CNN): một "bộ mã hóa" và một "bộ giải mã". Bộ mã hóa E biến khuôn mặt của mục tiêu nguồn thành một vectơ mã. Chỉ có một "bộ mã hóa" để đảm bảo rằng các tính năng độc lập với danh tính, bao gồm các cử chỉ trực quan, được ghi lại bất kể bộ nhận dạng của các cá nhân. Mỗi danh tính đều có bộ giải mã, "Di", sử dụng mã để tạo khuôn mặt của người đó. Theo cách tiếp cận không giám sát, bộ mã hóa và bộ giải mã được đào tạo phù hợp cùng nhau, sử dụng các bộ sưu tập ảnh không liên quan của nhiều cá nhân.

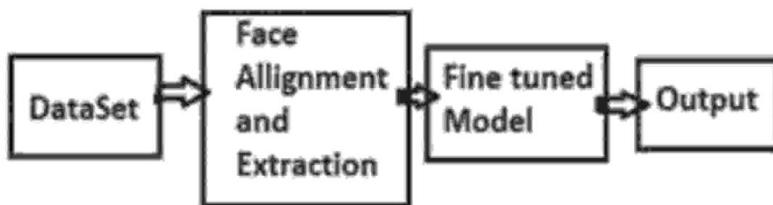
Cụ thể, như minh họa trong Hình 6.5, một cặp bộ giải mã-mã hóa được tạo bằng cách sử dụng "E" và "Di" cho hình ảnh đầu vào của từng đối tượng và các tính năng của chúng được tối ưu hóa để giảm thiểu lỗi xây dựng lại (sự khác biệt L1 giữa hình ảnh nguồn và hình ảnh được xây dựng lại). Truyền ngược được sử dụng để điều chỉnh các tham số cho đến khi đạt được độ bão hòa. Do đó, các đặc điểm tổng hợp được quay trở lại bối cảnh khuôn mặt của người dùng ban đầu và được cắt bớt một lần nữa từ điểm đặc điểm bằng bộ lọc. Giai đoạn cuối cùng giảm thiểu sự chuyển đổi giữa các khung hình tổng hợp và thực tế. Toàn bộ quá trình có thể được tự động hóa và cần rất ít sự tham gia của con người.



HÌNH 6.4 Thuật toán tạo DF cơ bản đã được tổng hợp [21].



HÌNH 6.5 Thuật toán tạo DF ban đầu đang được đào tạo [21].



HÌNH 6.6 Quy trình cơ bản để phát hiện DF [22].

6.7 CÁC PHƯƠNG PHÁP PHÁT HIỆN DEEPFAKES

Quy trình cơ bản để phát hiện DF được mô tả trong Hình 6.6. Có bốn bước liên quan đến toàn bộ quá trình. Bước đầu tiên là chọn tập dữ liệu. Trong bước thứ hai, căn chỉnh và trích xuất khuôn mặt được thực hiện. Bước thứ ba liên quan đến việc áp dụng một mô hình để phát hiện DF. Hơn nữa, trong bước cuối cùng, đầu ra được xác định.

Phát hiện DeepFakes bằng các tính năng địa phương và CNN

Các phương pháp khác nhau để phát hiện DF được mô tả như sau:

1. Phát hiện DF bằng Tín hiệu Sinh học

Một trong những phương pháp phát hiện DF là tín hiệu sinh học. Trong phương pháp này, các tín hiệu sinh học được trích xuất từ cả dữ liệu đầu vào hoặc dữ liệu thực và dữ liệu giả. Sau đó, chuyển đổi tín hiệu được áp dụng để tính toán tính nhất quán không gian và tính nhất quán thời gian. Hơn nữa, các vectơ đặc trưng được tạo và các bộ phân loại được sử dụng để tính toán tính xác thực [23].

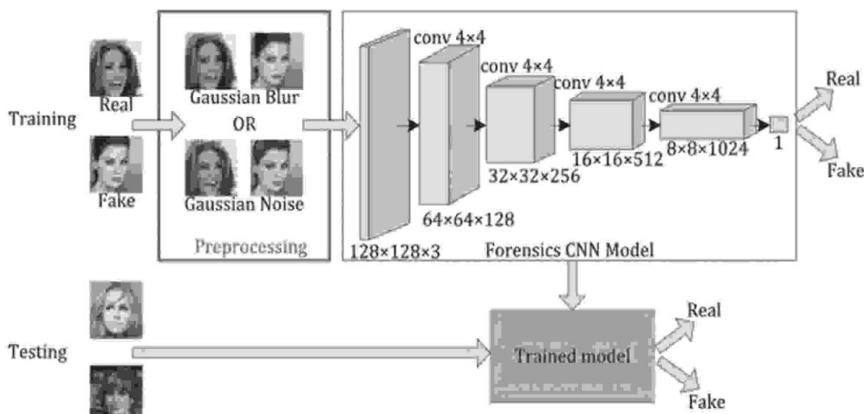
2. Phát hiện DF bằng cách sử dụng âm vị-Viseme không phù hợp

Phương pháp này sử dụng khái niệm "visemes", mô tả các chuyển động của cấu trúc hàm, có thể khác biệt hoặc mâu thuẫn với hình thái nói. Ví dụ, có thể có sự khác biệt về hình vị khi phát âm các cụm từ như mama, baba và papa, những từ này có thể được sử dụng để xác định những thay đổi thậm chí còn hạn chế về mặt địa lý và thống kê trong phim DF [24].

3. Phát hiện DF bằng Mạng thần kinh chuyển đổi

Một trong những phương pháp để phát hiện DF là sử dụng CNN. Karandikar [25], như trong Hình 6.7, đã mô tả một mô hình để phát hiện DF. Trong mô hình này, CNN là một bộ phân loại nhị phân với bốn lớp tích chập. Mô hình này loại bỏ các tín hiệu nhiễu được giảm đều đặn bằng cách áp dụng quá trình tiền xử lý ở mức hình ảnh giống nhau cho cả hình ảnh thực và hình ảnh giả. Kích hoạt các mô hình pháp y để hiểu các tính năng vốn có bổ sung để xác định các ảnh nổi bật được sản xuất và thực tế.

Ngoài mô hình này, có nhiều mô hình khác nhau trong đó CNN được sử dụng để phát hiện DF; ví dụ, Montserrat et al. [26] đã phát triển mô hình để phát hiện DF và trong mô hình này, CNN được sử dụng để trích xuất các đặc điểm trên khuôn mặt. Rana



HÌNH 6.7 Phát hiện DF bằng CNN [25].

và Sung [27] đã đề xuất một mô hình gọi là “DFStack” sử dụng CNN như một công cụ phân loại để phát hiện DF. Shad et al. [28] đã trình bày một nghiên cứu so sánh về các mô hình phát hiện DF bằng CNN. Trong nghiên cứu của ông, các bộ dữ liệu khác nhau được sử dụng để đào tạo tám CNN khác nhau. Ahmed và Sonuç [29] đã trình bày một mô hình phát hiện DF sử dụng CNN trên nền tảng MATLAB. Tóm tắt các mô hình phát hiện DF khác được giải thích trong Bảng 6.1.

6.8 CÁC TÍNH NĂNG ĐỊA PHƯƠNG VÀ TOÀN CẦU CỦA HÌNH ẢNH

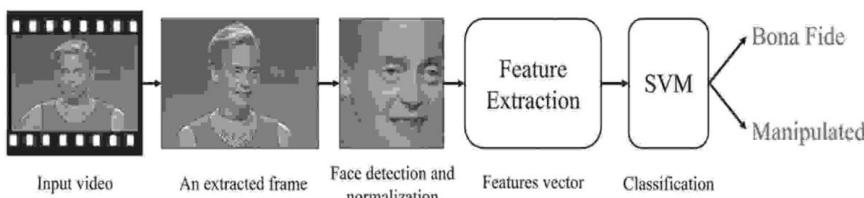
Thông tin được thu thập từ các hình ảnh dưới dạng số thường như là thách thức đối với con người để nắm bắt và giải thích được gọi là các đặc điểm. Giả sử rằng hình ảnh đại diện cho thông tin, thông tin được thu thập từ số liệu thống kê được gọi là tính năng. Kích thước của các tính năng lấy từ ảnh thường nhỏ hơn đáng kể so với ảnh thực tế. Chi phí phân tích nhóm ảnh giảm do giảm kích thước.

Nói chung, hai loại tính năng được lấy từ ảnh tùy thuộc vào cách sử dụng.

Đây là những tính năng địa phương và toàn cầu. Mô tả là một thuật ngữ được sử dụng để mô tả các tính năng. Bộ mô tả cục bộ được sử dụng để xác nhận mục, trong khi bộ mô tả toàn cục được sử dụng để trích xuất ảnh, nhận dạng đối tượng và phân loại. Sự khác biệt giữa phát hiện và mô tả đặc tính là rất quan trọng. Phát hiện là quá trình xác định sự hiện diện của một vật (ví dụ: xác định xem một vật có tồn tại trong ảnh hoặc phim hay không). Đồng thời, công nhận là quá trình lựa chọn danh tính của một thực thể (ví dụ: công nhận một cá nhân) [39].

Các đặc điểm cục bộ biểu thị các phần hình ảnh (các vị trí quan trọng trong ảnh) của một mục, trong khi các đặc điểm toàn cục thể hiện bức tranh hoàn chỉnh để mô tả toàn bộ sự vật. Biểu diễn đường viền, định danh biểu mẫu và tính năng kết cấu là các tính năng toàn cục, trong khi các tính năng cục bộ biểu thị cấu trúc trong một phần hình ảnh. Các bộ mô tả toàn cầu bao gồm “Ma trận hình dạng”, “Gradient định hướng theo biểu đồ (HOG)” và “Co-HOG”. Các bộ mô tả cục bộ bao gồm “SIFT,” “SURF,” “LBP,” “BRISK,” và “FREAK” [27].

Các tính năng toàn cầu thường được sử dụng cho các ứng dụng cấp thấp như nhận dạng và phân loại đối tượng, trong khi các tính năng cục bộ được sử dụng cho các ứng dụng cấp cao hơn như nhận dạng đối tượng. Việc sử dụng kết hợp các tính năng toàn cầu và cục bộ giúp cải thiện độ chính xác của nhận dạng đồng thời tăng chi phí xử lý.



HÌNH 6.8 Luồng phát hiện khuôn mặt DF sử dụng các tính năng cục bộ [41].

BẢNG
Tóm tắt mô hình Phân tích video ngắn “Chia sẻ”
Tác giả: Tùng Dương
Liên kết: <https://www.semanticscale.com/>

Guera và

phải được thay đổi để đảm bảo rằng các tệp trong bộ
tín dụng dữ liệu.
đưa ra quyết định “CNN” để phân loại video có
biết theo cách “DF” để phân loại video có
phim một cách tự động.

Đã sử dụng “CNN” để trích xuất các tính năng (mức khung hình).
Đã sử dụng RNN để đào tạo mô hình để tìm xem có video hay không
đã bị thao túng.

Sohrawardi
et al. [32]

2019 Bài báo nhằm mục đích giải quyết các vấn đề từ quan điểm của một phóng viên và hoạt động hiệu quả trong việc thiết kế một thiết bị có thể kết hợp dễ dàng.
Đã đề xuất một phương pháp cho phép người xem xác định xem video được chia sẻ trên internet có phải là DF một cách an toàn và hiệu quả hay không.

Irene Mý
et al. [33]

2019 -> Cung cấp một chiến lược pháp y mới để phân biệt
giữa các video giám sát giả và xác thực, trái ngược với các kỹ thuật trước đây dựa trên các khung hình đơn lẻ.
-> Họ đã sử dụng các trường dòng quang học để điều tra khả năng xảy ra sự không giống nhau giữa các khung.

Hứa Kỳ và công
sự. [34]

2020 -> Đề xuất mô hình có tên “DeepRhythm” để phát hiện
của DF bằng cách theo dõi nhịp tim.
-> Hệ thống thích ứng với khuôn mặt không ngừng phát triển và các loại sai
bằng cách sử dụng sự chú ý kép không gian-thời gian.

Độ chính xác
tác giả: 99,5%
Độ chính xác xác
thực: 96,9
Độ chính xác kiểm
trú: 96,7

Tỷ lệ lỗi tương đương
cho “ResNeXTspoof”: 5,4%
Tỷ lệ lỗi bằng nhau cho
“tích chập
LSTM”: 6,4%

Độ chính xác
của “VGG16”: 81,61%

Độ chính xác: 0,98

Để đánh giá các
tính năng của thuật toán
với phim đã chỉnh sửa bằng
cách sử dụng các chiến lược chưa được
khám phá trước đây trong giai đoạn đào
tạo.

Để tiến hành thử nghiệm beta trực
tiếp, cho phép nhiều người tham
gia hơn.

Để đánh giá thuật toán
độ chính xác bằng cách sử dụng các bộ
dữ liệu và mạng lưới thần kinh khác nhau.

Để sử dụng mô hình để xuất trong
các lĩnh vực khác như chống lại
các đối thủ phi truyền thống
các cuộc tấn công.

(còn tiếp)

BẢNG 6.1 (Tiếp theo)

Tóm tắt Mô hình Phát hiện DF Sử dụng "CNN"

Tác giả	Tóm chia khóa năm	Sự chính xác	Hạn chế/Công việc trong tương lai
Gandhi và đạo Jain [35]	2020 -> Trong cả hai tình huống hộp đen và hộp trắng, nhiễu loạn đổi phương đã được thực hiện bằng cách sử dụng "Phương pháp ký hiệu chuyển màu nhanh" và tấn công chuẩn "Caillin" và "Wagner L2". -> CNN tổng quát không giám sát được sử dụng trong DIP.	Độ chính xác: 98%	Không đề cập đến
Vương và cộng sự. [36]	2020 -> Nghiên cứu cho thấy rằng với dữ liệu và quy trình xử lý phù hợp tăng cường, một trình phát hiện hình ảnh điển hình được đào tạo chỉ trên một trình tạo CNN có thể khai thác tốt một cách bất ngờ đối với các cấu trúc, bộ dữ liệu và phương pháp đào tạo chưa được khám phá trước đây. -> Đề xuất một ý tưởng hấp dẫn rằng các bức ảnh do CNN tạo ra có một số khiếm khuyết hệ thống phổ biến khiến chúng không đạt được chất lượng cao.	Độ chính xác: 98,2	Không đề cập đến
Tarasiou và cộng sự. [43]	2020 -> Trình bày "Celeb-DF", một DF rộng lớn đầy sáng tạo có vấn đề bộ dữ liệu video với 5.639 phim DF chất lượng cao của các siêu sao được phát triển bằng cách sử dụng phương pháp tổng hợp nâng cao. -> Kiểm tra kỹ lưỡng các phương pháp và dữ liệu phát hiện DF để làm nổi bật mức độ khó tăng lên do Celeb-DF cung cấp.	"Xception-c40" cho điểm AUC là 65,5 trên bộ dữ liệu "Celeb-DF".	Celeb-DF nên bao gồm anti phương pháp pháp y.
Zi et al. [37]	Năm 2020 -> "WildDF" được đề xuất, một bộ dữ liệu mới chứa 7.314 mẫu khuôn mặt bắt nguồn từ 707 phim DF được lấy hoàn toàn từ web. -> Thực hiện phân tích toàn diện về một tập hợp các hệ thống phát hiện cơ bản.	Độ chính xác phát hiện mức trình tự: 65,50%	Không đề cập đến
Lý và cộng sự. [38]	2020 -> Trình làng một Nơ-ron chuyển đổi cập và độc đáo Phương pháp mạng (PPCNN) để phân biệt phim hoặc ảnh DF với nội dung thực.	điểm AUC trên "Faceforensics++": 99,4	Đề xuất ra một hình dạng hơn để nhận dạng các DF khác nhau với mật độ pixel và nền khác nhau.

Deepfake

BẢNG 6.2

Tóm tắt về phát hiện DF bằng cách sử dụng

Tác giả / Nguồn

Akhter et al.

Dutta et al.

Agarwal v

công sự. [44]

Mô hình / Công nghệ

Người dùng / Xuất bản

Đến từ / Mô tả

Nhóm / Phân loại

Phát triển / Phân tích

Thử nghiệm / Kết quả

Thứ tự / Tác giả

Thời gian / Năm

Mô hình / Công nghệ

Phương pháp / Kỹ thuật

Kết quả / Kết luận

DF / FID

Kawa et al.

2020

-> Đã phát triển và thử nghiệm chức năng kích hoạt mới "Pish" cho phép độ tin cậy cao hơn nữa với mức giá của các nguồn thời gian nhò.

Tỷ lệ lỗi: 0,28

Để có được EER giảm, mục tiêu là mở rộng hệ thống dựa trên LFD cách tiếp cận.

[45]

-> Hiển thị những phát hiện ban đầu về phương pháp phát hiện DeepFake dựa trên "Bộ mô tả tính năng cục bộ", cho phép thiết lập nền tảng nhanh hơn đáng kể cũng như thay vì sử dụng GPU.

Abdullah et al.

2020

-> Đã xuất một phương pháp mới để xác định phân loại nhịp ECG.

Độ chính xác: 99,9% -> Phương pháp lựa chọn tính năng và kết hợp cục bộ

[46]

Phương pháp được đề xuất dựa trên phân tích hình ảnh.

mô tả sẽ được kiểm tra.

-> Các ảnh chụp nhanh xung ECG trước tiên được lưu trữ dưới dạng ảnh xung ECG, sau đó việc thu thập đặc điểm từ các ảnh xung ECG được thực hiện bằng cách sử dụng bộ mô tả đặc điểm cục bộ.

-> Đề sử dụng các kỹ thuật ML khác trong kiến trúc được đề xuất.

Vương và cộng

2021

-> "FFR_FD" được đề xuất như một vectơ chuyên sâu để thể hiện định nghĩa hình ảnh của khuôn mặt, được xây dựng bằng cách sử dụng các bộ mô tả tính năng cục bộ thông qua các vùng khuôn mặt được phân đoạn.

điểm AUC trên

-> Đề tạo một khuôn mặt pháp y định danh đặc trưng loại trừ.

sự. [47]

"DFTIMIT"

LQ- 99.9

6.8.1 Phát hiện DeepFakes bằng các tính năng cục bộ

Phương pháp trích xuất các tính năng dựa trên pixel được sử dụng trong phát hiện DF dựa trên tính năng cục bộ. Nó trích xuất các đặc điểm từ mọi pixel. Phát hiện dựa trên tính năng cục bộ có hiệu quả tốt hơn so với phát hiện dựa trên tính năng trực quan vì nó nắm bắt được tất cả các phần của hình ảnh. Các tính năng kết quả là có hứa và thường không thể nhìn thấy bằng mắt thường [40].

Việc phát hiện các DF “khuôn mặt” thường được coi là một vấn đề phân loại nhị phân, trong đó dữ liệu nguồn phải được phân loại là đã sửa đổi hoặc vô hại. Mục tiêu chính của quy trình là thu được một tập hợp các đặc điểm duy nhất, khi được kết hợp với phương pháp phân loại, sẽ làm tăng khả năng dữ liệu nguồn là thực tế. KN

Ramadhan và R. Munir [41] đã đề xuất một luồng hệ thống được minh họa trong Hình 6.8 để phát hiện DF bằng cách sử dụng các bộ mô tả cục bộ. Kiến trúc bắt đầu bằng cách trích xuất một hình ảnh duy nhất từ clip hình ảnh được phát hiện. Hình ảnh sau đó được xác định bằng cách sử dụng phương pháp “Viola Jones”. Hình ảnh được điều chỉnh và các đặc điểm được truy xuất bằng cách sử dụng bộ mô tả hình ảnh cục bộ. Các tính năng đã thu thập được đặt vào bộ phân loại SVM, ngăn chặn việc khai thác xem dữ liệu có bị thay đổi (DF) hay chính hãng hay không. Tóm tắt các mô hình phát hiện DF đã được sử dụng để phát hiện DF bằng cách sử dụng các đặc điểm cục bộ của ảnh được trình bày trong Bảng 6.2.

6.9 TÓM TẮT

Mục đích của chương này là cung cấp cho các nhà nghiên cứu mới một cái nhìn đơn giản và ngắn gọn về việc tạo và trích xuất DF bằng cách sử dụng “các tính năng cục bộ” và “CNN”. Nó giới thiệu các DF, các ứng dụng của chúng, cùng với những ưu điểm và nhược điểm. Các kỹ thuật được sử dụng thường xuyên nhất để tạo DF cũng được thảo luận. Hơn nữa, tổng quan về các kỹ thuật phát hiện DF được giải thích, cùng với bản tóm tắt công việc được thực hiện trong lĩnh vực đó (các phương pháp). Các kỹ thuật chính được mô tả ở đây là phát hiện DF bằng cách sử dụng “các tính năng cục bộ” và “CNN”.

NGƯỜI GIỚI THIỆU

- [1] TT Nguyen, QVH Nguyen, CM Nguyen, D. Nguyen, DT Nguyen, and S. Nahavandi, “DL for DF Creation and Detection: A Survey,” trang 1-16, 2019. [Trực tuyến].
Có sẵn: <http://arxiv.org/abs/1909.11573>.
- [2] J. Kietzmann, LW Lee, IP McCarthy, và TC Kietzmann, “DFs: Trick or Treat?,” Xe buýt. Chân trời., tập. 63, không. 2, trang 135-146, 2020, doi:<https://doi.org/10.1016/j.bushor.2019.11.006>
- [3] “DF là gì và nó đánh lừa mọi người như thế nào? - Skeptikai,” Skeptikai - Bao gồm Địa chính trị, Tâm lý học, Ngôn ngữ và hơn thế nữa, 2022. [Trực tuyến]. Có sẵn: <https://skeptikai.com/what-is-DF-and-how-is-it-fooling-people/>. [Truy cập: ngày 02 tháng 1 năm 2022].
- [4] G. Shao, 2022. [Trực tuyến]. Có sẵn: www.cnbc.com/2019/10/14/what-is-DF-and-how-nó-có-thể-nguy-hiểm.html. [Truy cập: ngày 02 tháng 1 năm 2022].
- [5] T. Biggs và R. Moran, “Deep Fake là gì và chúng được tạo ra như thế nào?” smh.com.au, 2022. [Trực tuyến]. Có sẵn: www.smh.com.au/technology/what-is-the-difference-between-a-fake-and-a-DF-20200729-p55ghi.html. [Truy cập: ngày 02 tháng 1 năm 2022].
- [6] “Giải cấu trúc DF—Chúng hoạt động như thế nào và rủi ro là gì?” XemBlog, 2022.

- [7] HS Shad và cộng sự, "Phân tích so sánh phương pháp phát hiện hình ảnh DF bằng mạng thần kinh chuyển đổi.", Máy tính. thông minh. Khoa học thần kinh., tập. 2021, tr. 3111676, 2021, doi:10.1155/2021/3111676
- [8] M. Westerlund, "Sự xuất hiện của công nghệ DF: Đánh giá," Technol. đổi mới. quản lý. Rev., tập. 9, trang 40-53, 2019, doi:<http://doi.org/10.22215/timreview/1282>
- [9] S. Adee, "DF là gì và chúng được tạo ra như thế nào?" IEEE Spectrum, 2022. [Trực tuyến]. Có sẵn: <https://spectrum.ieee.org/what-is-DF>. [Truy cập: ngày 14 tháng 1 năm 2022].
- [10] R. Gonzales, "DFs: How to Tell What's Real When nothing Is Is," GIANT FREAKIN ROBOT, 2022. [Trực tuyến]. Có sẵn: www.giantfreakinrobot.com/tech/DFs.html. [Truy cập: ngày 2 tháng 1 năm 2022].
- [11] D. Song, "A Short History of DFs," Medium, 2022. [Trực tuyến]. Có sẵn: <https://medium.com/@songda/a-short-history-of-DFs-604ac7be6016>. [Truy cập: ngày 4 tháng 1 năm 2022].
- [12] R. Toews, "DFs sẽ tàn phá xã hội. Chúng Tôi Chưa Chuẩn Bị," Forbes, 2022. [Trực tuyến]. Có sẵn: www.forbes.com/sites/robtoews/2020/05/25/DFs-dang-sap-tan-phá-tan-hai-xa-hoi-chung-toi-khong-chuan-bi/?sh=60ac43bb7494. [Truy cập: ngày 4 tháng 1 năm 2022].
- [13] A. Jaiman, "Các trường hợp sử dụng tích cực của DF," Trung bình, 2022. [Trực tuyến]. Có sẵn: <https://directiondatascience.com/positive-use-cases-of-DFs-49f510056387>. [Truy cập: ngày 4 tháng 1 năm 2022].
- [14] Technology.org, 2022. [Trực tuyến]. Có sẵn: www.technology.org/2021/11/26/DF-in-dien-anh/. [Truy cập: ngày 4 tháng 1 năm 2022].
- [15] "Những ứng dụng này của công nghệ DF sẽ làm bạn ngạc nhiên! - Trường AI của Ấn Độ," Trường AI của Ấn Độ, 2022. [Trực tuyến]. Có sẵn: <https://aischoolofindia.com/parents/these-application-of-DF-technology-will-amaze-you/>. [Truy cập: ngày 4 tháng 1 năm 2022].
- [16] "Ưu và nhược điểm: Công nghệ DF và Hình đại diện AI - Springwise," Springwise, 2022. [Trực tuyến]. Có sẵn: www.springwise.com/pros-cons/DF-technology-ai-avatars. [Truy cập: ngày 4 tháng 1 năm 2022].
- [17] "Phương tiện tổng hợp / giả mạo sâu sắc trong tiếp thị - Ưu điểm và nhược điểm - PR Smith," PR Smith, 2022. [Trực tuyến]. Có sẵn: <https://prsmith.org/2021/10/14/deep-fake-synthetic-media-in-marketing-advans-disadvans/>. [Truy cập: ngày 6 tháng 1 năm 2022].
- [18] "DFs là tin giả mới | Công nghệ kinh doanh hàng đầu," Công nghệ kinh doanh hàng đầu, 2022. [Trực tuyến]. Có sẵn: <https://bttech.co/news/DFs-are-the-new-fake-news/>. [Truy cập: ngày 14 tháng 1 năm 2022].
- [19] A. Voulodimos, N. Doulamis, A. Doulamis và E. Protopapadakis, "DL for Computer Vision: A Brief Review," Comput. thông minh. Khoa học thần kinh., tập. 2018, tr. 1-14, 2018, doi:10.1155/2018/7068349
- [20] BU Mahmud và A. Sharmin, "Những hiểu biết sâu sắc về công nghệ DF: Đánh giá," tập. 5, trang 13-23, 2021. [Trực tuyến]. Có sẵn: <http://arxiv.org/abs/2105.00192>.
- [21] Y. Li, X. Yang, P. Sun, H. Qi và S. Lyu, "Celeb-DF: Bộ dữ liệu đầy thách thức quy mô lớn dành cho pháp y DF," Proc. Điện toán IEEE. Sóc. Conf. Điện toán. xem Nhận dạng mẫu, trang 3204-3213, 2020, doi:10.1109/CVPR42600.2020.00327
- [22] A. Karandikar, "Phát hiện video DF bằng mạng thần kinh tích chập," Int. J. Adv. Máy tính xu hướng. Khoa học. Anh, tập. 9, không. 2, trang 1311-1315, 2020, doi:10.30534/ijatcse/2020/62922020
- [23] UA Ciftci, I. Demir và L. Yin, "FakeCatcher: Phát hiện các video chân dung tổng hợp bằng tín hiệu sinh học," IEEE Trans. Mẫu hậu môn. máy móc. Trí tuệ, tập. X, không. X, tr. 1-1, 2020, doi:10.1109/tpami.2020.3009287
- [24] S. Agarwal, H. Farid, O. Fried và M. Agrawala, "Phát hiện video giả mạo sâu từ sự không khớp giữa âm vị-Visme," Hội thảo IEEE/CVF 2020 về Thị giác máy tính và Hội thảo nhận dạng mẫu (CVRW), 2020 , trang 2814-2822, doi:10.1109/CVPRW50498.2020.00338

- [25] A. Karandikar, "Phát hiện video DF bằng mạng thần kinh tích chập," Int. J. Adv. Máy tính xu hướng. Khoa học. Anh, tập. 9, không. 2, trang 1311-1315, 2020, doi:10.30534/ijatcse/2020/62922020
- [26] DM Montserrat và cộng sự, "Phát hiện DF với trọng số khuôn mặt tự động," IEEE Comput. Sóc. Conf. Điện toán. xem Nhận dạng mẫu. Công việc., tập. Tháng 6 năm 2020, trang 2851-2859, 2020, doi:10.1109/CVPRW50498.2020.00342
- [27] ThS Rana và AH Sung, "DFStack: Kỹ thuật học tập dựa trên tập hợp sâu để phát hiện DF," Proc. - 2020 lần thứ 7 IEEE Int. Conf. An ninh mạng. Điện toán đám mây. 2020 lần thứ 6 IEEE Int. Conf. Điện toán cạnh. Đám mây có thể mở rộng, CSCloud-EdgeCom 2020, số. Tháng 8, tr. 70-75, 2020, doi:10.1109/CSCloud-EdgeCom49738.2020.00021
- [28] HS Shad và cộng sự, "Phân tích so sánh phương pháp phát hiện hình ảnh DF bằng cách sử dụng Mạng nơ-ron tích chập," tập. 2021, 2021.
- [29] SRA Ahmed và E. Sonuç, "Phát hiện DF bằng cách sử dụng Mạng nơ-ron tích chập tăng cường cơ sở lý luận," Appl. Nanosci., không. Tháng 9 năm 2021, doi:10.1007/s13204-021-02072-3
- [30] Y. Li và S. Lyu, "Hiển thị video DF bằng cách phát hiện tạo tác công vênh khuôn mặt," 2018. [Trực tuyến]. Có sẵn: <http://arxiv.org/abs/1811.00656>
- [31] D. Guera và EJ Delp, "Phát hiện video DF bằng mạng thần kinh tái phát," Proc. AVSS 2018-2018 Lần thứ 15 IEEE Int. Conf. quảng cáo Giám sát dựa trên tín hiệu video., 2019, doi:10.1109/AVSS.2018.8639163
- [32] SJ Sohrabwardi và cộng sự, "Poster: Hướng tới khả năng phát hiện DF mạnh mẽ trong thế giới mở," Proc. Hội nghị ACM Điện toán. công đồng. An toàn., không. Tháng 2 năm 2020, trang 2613-2615, 2019, doi:10.1145/3319535.3363269
- [33] Irene Amerini và cộng sự, "Phát hiện video DF thông qua đệ trình ICCV ẩn danh trên CNN dựa trên dòng quang," ICCV Work., số. Micc, trang 1-3, 2019. [Trực tuyến]. Có sẵn: http://openaccess.thecvf.com/content_ICCVW_2019/html/HBU/Amerini_DF_Video_Detection_through_Optical_Flow_Based_CNN_ICCVW_2019_paper.html.
- [34] H. Qi và cộng sự, "DeepRhythm: Phơi bày DF với Nhịp điệu nhịp tim có chú ý trực quan," MM 2020 - Proc. ACM lần thứ 28 Conf. Multimed., trang 1318-1327, 2020, doi:10.1145/3394171.3413707
- [35] A. Gandhi và S. Jain, "Adversarial Perturbations Đánh lừa Máy dò DF," Proc. quốc tế Jt. Conf. Mạng lưới thần kinh, không. Ijcn, 2020, doi:10.1109/IJCNN48605.2020.9207034
- [36] SY Wang, O. Wang, R. Zhang, A. Owens, và AA Efros, "Hình ảnh do CNN tạo ra dễ phát hiện một cách đáng ngạc nhiên. Hiện tại," Proc. Điện toán IEEE. Sóc. Conf. Điện toán. xem Nhận dạng mẫu, trang 8692-8701, 2020, doi:10.1109/CVPR42600.2020.00872
- [37] B. Zi, M. Chang, J. Chen, X. Ma và YG Jiang, "WildDF: Bộ dữ liệu thế giới thực đầy thách thức để phát hiện DF," MM 2020 - Proc. ACM lần thứ 28 Conf. Multimed., trang 2382-2390, 2020, doi:10.1145/3394171.3413769
- [38] X. Li, K. Yu, S. Ji, Y. Wang, C. Wu và H. Xue, "Chiến đấu chống lại DF: Mạng thần kinh chuyển đổi và & ghép nối (PPCNN)," Web Conf . 2020 - Đồng hành World Wide Web Conf. WWW2020, không. Tháng 10, tr. 88-89, 2020, doi:10.1145/3366424.3382711
- [39] D. Tyagi, "Giới thiệu về phát hiện và đổi sảnh đặc điểm," Trung bình, 2022. [Trực tuyến]. Có sẵn: <https://medium.com/data-breach/introduction-to-feature-detection-and-Matching-65e27179885d>. [Truy cập: ngày 14 tháng 1 năm 2022].
- [40] AI Awad và M. Hassaballah, Trình phát hiện và mô tả tính năng hình ảnh, tập. 630, không. Tháng 10 năm 2017, Springer, 2016.
- [41] KN Ramadhani và R. Munir, "Một nghiên cứu so sánh về phương pháp phát hiện video DF," 2020 3rd Int. Conf. thông tin liên lạc công đồng. công nghệ. ICOIACT 2020, trang 394-399, 2020, doi:10.1109/ICOIACT50329.2020.9331963

- [42] Z. Akhtar và D. Dasgupta, "Đánh giá so sánh các bộ mô tả tính năng cục bộ để phát hiện DF," 2019 IEEE Int. Triệu chứng công nghệ. quê hương. bao mật. HST 2019, tr. 0-4, 2019, doi:10.1109/HST47167.2019.9033005
- [43] M. Tarasiou và S. Zafeiriou, "Trích xuất các đặc điểm cục bộ sâu để phát hiện các hình ảnh khuôn mặt người bị thao túng," Proc. - Quốc tế Conf. Quá trình hình ảnh. ICIP, tập. Tháng 10 năm 2020, tr. 1821-1825, 2020, doi:10.1109/ICIP40778.2020.9190714
- [44] A. Agarwal, R. Singh, M. Vatsa, và A. Noore, "MagNet: Detecting Digital Presentation Attacks on Face Recognition," Front. nghệ thuật. Trí tuệ, tập. 4, không. Tháng 12, trang 1-19, 2021, doi:10.3389/frai.2021.643424
- [45] P. Kawa và P. Syga, "Ghi chú về phát hiện DF với nguồn lực thấp," 2020. [Trực tuyến]. Có sẵn: <http://arxiv.org/abs/2006.05183>.
- [46] DA Abdullah, MH Akpinar, và A. Şengür, "Phân loại nhịp ECG dựa trên các bộ mô tả tính năng cục bộ," Heal. thông tin liên lạc Khoa học. Hệ thống, tập. 8, không. 1, 2020, doi:10.1007/s13755-020-00110-y
- [47] G. Wang, Q. Jiang, X. Jin và X. Cui, "FFR_FD: Phát hiện nhanh và hiệu quả các DF dựa trên lỗi đặc điểm đặc trưng," 2021. [Trực tuyến]. Có sẵn: <http://arxiv.org/abs/2107.02016>.



Taylor & Francis
Taylor & Francis Group
<http://taylorandfrancis.com>

7 DeepFake

Trường hợp tích cực

Bò tót Loveleen và Gursimar Kaur Arora

NỘI DUNG

7.1 Giới thiệu	91
7.2 Các ứng dụng của DeepFakes	92
7.2.1 Ngành: Y tế và Dược phẩm	92
7.2.1.1 Trường hợp 1	92
7.2.1.2 Trường hợp 2	92
7.2.1.3 Trường hợp 3	92
7.2.2 Ngành: Bán lẻ, Thương mại điện tử, Tư vấn	93
7.2.2.1 Trường hợp 4	93
7.2.3 Ngành: Thời trang	94
7.2.3.1 Trường hợp 5	94
7.2.3.2 Trường hợp 6	94
7.2.4 Ngành: Truyền thông và Giải trí	94
7.2.4.1 Trường hợp 7	94
7.2.4.2 Trường hợp 8	94
7.2.4.3 Trường hợp 9	95
7.2.5 Ngành: Giáo dục	95
7.2.5.1 Trường hợp 10	95
7.3 Tóm tắt	96
Người giới thiệu	96

7.1 GIỚI THIỆU

DeepFakes (DF) là phương tiện tổng hợp được tạo bằng thuật toán DL, như đã thảo luận trong các chương trước. Điều cần thiết là nghiên cứu sâu hơn được tiến hành để giới thiệu các ứng dụng tích cực của nó, ngay cả khi câu hỏi về sự tồn tại của nó là một nguyên nhân vẫn đang được tiến hành [1]. Godull và cộng sự [2] giới thiệu nghiên cứu có liên quan đứng ở vị trí 37 trong số có khoảng 6 nghiên cứu tập trung vào các cơ hội liên quan đến DF và 13 nghiên cứu về rủi ro của nó. Do đó, trong các phần tiếp theo, chúng tôi thảo luận về cách DF có thể giúp xã hội sử dụng các ngành công nghiệp khác nhau như các trường hợp khác nhau trong các hạng mục nghiên cứu, sáng tạo nội dung, sáng tạo và chiến lược rộng hơn trong các ngành như chăm sóc sức khỏe và dược phẩm, thời trang, bán lẻ, Thương mại điện tử, phương tiện truyền thông, giải trí, v.v. Sau mỗi ngành, một bảng sẽ tóm tắt các trường hợp được bao gồm cùng với kỹ thuật và cách sử dụng nó.

7.2 ỨNG DỤNG CỦA DEEPFAKES

7.2.1 Ngành: Y tế và Dược phẩm

7.2.1.1 Trường hợp 1

Nghiên cứu là xương sống của ngành chăm sóc sức khỏe và dược phẩm. Bất kỳ nghiên cứu nào trong ngành này đều có khả năng điều trị các bệnh đe dọa đến tính mạng, khám phá thuốc hoặc dịch tễ học. Thu thập dữ liệu và tính sẵn có của dữ liệu là một điểm gây trở ngại cho các nhà nghiên cứu y học, vì nó tốn kém và mất thời gian. Zhu et al. đề xuất nghiên cứu sử dụng DF để bảo vệ danh tính bệnh nhân, sử dụng kỹ thuật hoán đổi khuôn mặt, giữ lại điểm trọng yếu là khuôn mặt và cơ thể. Phương pháp này có thể là một bước đột phá trong nghiên cứu y học, vì dữ liệu của bệnh nhân có thể được chia sẻ mà không có vấn đề riêng tư nào [3]. Dữ liệu tổng hợp về chăm sóc sức khỏe là dữ liệu được tạo (medGAN) [4] từ việc đào tạo thuật toán trên một nguồn tự nhiên. Dữ liệu này có nhiều ứng dụng khác nhau, chẳng hạn như phát hiện khối u, bằng cách cung cấp tia X của những bệnh nhân có khối u trước đó. Việc sử dụng khác là nghiên cứu sâu rộng bằng cách tạo ra tia X tổng hợp. Ví dụ, Zhao et al. [5] đã tạo ra các hình ảnh vòm mạc giống như thật, sử dụng GAN để khuyến khích nghiên cứu về các ứng dụng của hình ảnh tế bào thần kinh. Dữ liệu tổng hợp có thể được sử dụng để tạo bộ dữ liệu hợp lý và tính linh hoạt cho các mô hình AI [6]. Khả năng tương thích với mô hình là một khía cạnh thiết yếu trong việc phê duyệt dữ liệu theo quy định lâm sàng. Dữ liệu này có thể hữu ích trong các chuyên đổi kỹ thuật số hiện tại trong chăm sóc sức khỏe [7].

7.2.1.2 Trường hợp 2

Để tạo thuận lợi cho nghiên cứu, chuột và động vật gặm nhấm thường được sử dụng để thử nghiệm các loại thuốc mới hoặc tìm kiếm thông tin mở rộng về một số vi rút hoặc tế bào vi khuẩn. Tận dụng công nghệ DL mô hình, Coffey et al. [8] đã trình bày phân tích nghiên cứu của họ về Phát âm siêu âm (USV) để hiểu rõ hơn về trạng thái tâm lý và chức năng thần kinh của động vật được sử dụng trong giai đoạn thử nghiệm trong phòng thí nghiệm. Do đó, mô hình được tạo bằng thuật toán CNN khu vực có thể chuyển đổi tín hiệu âm thanh sang định dạng trực quan như hình ảnh. Nghiên cứu này sẽ có tác động đáng kể bằng cách cung cấp các phân tích chuyên sâu [9,10].

7.2.1.3 Trường hợp 3

Ngoài nghiên cứu, DF giờ đây có thể lên tiếng cho những người mắc bệnh thần kinh vận động (MND). Thật không may, bệnh nhân mắc chứng rối loạn này sẽ mất khả năng nói, vì nó ảnh hưởng đến dây thần kinh não, nơi chỉ đạo bộ não của chúng ta giao tiếp. Trong Phòng thí nghiệm dữ liệu R2 của mình, Rolls Royce đã hợp tác với Hiệp hội MND, Microsoft, Accenture, Computacenter, Intel và Dell Technologies [11] để tạo ra Quips. Đó là một công nghệ ngân hàng giọng nói trong đó giọng nói được lưu trữ trong ngân hàng giọng nói cho đến khi bệnh nhân MND không thể nói được. Thông qua các giọng nói thu thập được, giọng nói tổng hợp được tạo ra với cùng một giọng nói của bệnh nhân. Công việc của Quip là nghe cuộc trò chuyện và gợi ý câu trả lời của bệnh nhân để lựa chọn mà không cần người gõ.

Một phần mềm khác cung cấp ngân hàng giọng nói sử dụng công nghệ tiên tiến này là VocaliD, phần mềm này cũng cung cấp gói giọng nói và di sản giọng hát. Một viên nang giọng nói hoặc tiêu sử âm thanh có thể được định nghĩa là một trang web nơi một người có thể ghi lại giọng nói của họ bằng cách tường thuật

những câu chuyện và trải nghiệm của họ và lưu lại cho thế hệ tương lai sau khi mua một hộp thoại. Di sản giọng hát là một phương pháp gián tiếp đảm bảo giọng nói của một người bằng cách tạo ra giọng nói kỹ thuật số riêng cho từng người, đặc biệt đối với những người mắc bệnh MND hoặc bệnh xơ cứng teo cơ một bên (AML).

Phần mềm ứng dụng/

Chương trình học	Trường hợp sử dụng	Kỹ thuật	Trang web, nếu có
Zhu et al. [3]	Bảo vệ sự riêng tư của bệnh nhân	Hoán đổi khuôn mặt	
Cà phê và công sự. [số 8]	Nghiên cứu và phân tích chuyên sâu	Phân tích siêu âm phát âm bằng cách sử dụng CNN bằng cách chuyển đổi tín hiệu âm thanh sang định dạng hình ảnh	
Lời châm biếm [11]	Ngôn hàng thoại	Tổng hợp âm thanh	
VocaliD	Ngôn hàng thoại, giọng nói Viên nang, Thành nhạc Di sản	Tổng hợp âm thanh	https://vocalid.ai/
Triệu và cộng sự. [5]	Hình ảnh vòng mạc tổng hợp để khuyến khích nghiên cứu về tế bào thần kinh	GAN	
	hình ảnh		

7.2.2 Ngành: Bán lẻ, Thương mại điện tử, Tư vấn

7.2.2.1 Trường hợp 4

Bán lẻ và Thương mại điện tử luôn tìm kiếm các phương pháp tốt hơn để thu hút người tiêu dùng. Phương tiện truyền thông tổng hợp cung cấp một chiến lược đơn giản hơn, rẻ hơn, nhanh hơn và hiệu quả hơn [12]. Như đã thấy trong các chương trước, DF không chỉ là thao tác hình ảnh hoặc video mà còn là thao tác âm thanh [13,14,15].

Synthesia, một công ty AI đến từ London, đã tạo ra phần mềm để chọn Người thuyết trình AI, nhập văn bản mà hình đại diện đã chọn sẽ nói và thế là xong! Các công ty doanh nghiệp đã sử dụng phần mềm này để tạo các video đào tạo, có thể thay đổi, cập nhật và thậm chí chuyển tải bằng các ngôn ngữ khác nhau-có thể dịch đóng một lợi thế lớn cho các công ty đa quốc gia trải rộng trên toàn cầu nhưng phải đối mặt với rào cản ngôn ngữ, đặc biệt là khi thuê công dân địa phương. Nó cũng tạo ra một video với các ngôn ngữ khác nhau để nâng cao nhận thức về bệnh sốt rét có sự góp mặt của David Beckham; do đó, DF là một công cụ tuyệt vời để nâng cao nhận thức của công chúng.

Phần mềm ứng dụng/

Chương trình học	Trường hợp sử dụng	Kỹ thuật	Trang web, nếu có
tổng hợp	Video đào tạo doanh nghiệp	Tạo hình ảnh và tổng hợp âm thanh	www.synthesia.io/

7.2.3 Ngành: Thời trang

7.2.3.1 Trường hợp 5

Tính độc đáo và tầm nhìn thúc đẩy ngành công nghiệp thời trang. DF có thể tạo các bản in và thiết kế mới thông qua GAN bằng cách huấn luyện thuật toán theo các phương pháp hiện tại. DF cũng thêm yếu tố cá nhân hóa. Vì có một sự thay đổi đáng kể khi mua hàng trên các trang Thương mại điện tử nên một bất lợi lớn là tưởng tượng mình đang mặc trang phục đó.

Do đó, thông qua DF, người ta có thể cá nhân hóa người mẫu trên màn hình theo kích thước cơ thể, màu sắc và kiểu tóc trước khi mua. Ứng dụng này đã được triển khai trong ứng dụng Superpersonal, một công ty của Anh [9,10,16,17,18,19].

7.2.3.2 Trường hợp 6

Một công ty Nhật Bản, DataGrid, đã tạo ra các DF giống như thật của các người mẫu thời trang bằng cách sử dụng kỹ thuật GAN và nghiên cứu sâu hơn về việc tạo chuyển động thông qua AI [15,20,21,22,23]. Nếu các mô hình kỹ thuật số được triển khai, nó có thể là một giải pháp thay thế kinh tế và hiệu quả cho ngành này. Theo nghiên cứu của Forbes về thị trường và khách hàng, người ta suy luận rằng người mẫu kỹ thuật số có thể giảm chi phí tới 75% so với chi phí chụp sản phẩm [24].

Phần mềm ứng dụng/

Chương trình học	trường hợp sử dụng	Kỹ thuật	Trang web, nếu có
Siêu cá nhân [16]	Cá nhân hóa cho người mua	Tạo hình ảnh www.superpersonal.com/ sắm trực tuyến	
Lưới dữ liệu [24]	Mô hình kỹ thuật số GAN		https://datagrid.co.jp/en/

7.2.4 Ngành: Truyền thông và Giải trí

7.2.4.1 Trường hợp 7

Ngành công nghiệp này dự kiến sẽ thu được nhiều lợi nhuận nhất với việc tạo ra DF. Có thể giải thích các ứng dụng DF cho ngành truyền thông và giải trí thông qua ví dụ về việc tạo và phát sóng một bộ phim hoặc video. Với sự ra đời của DF, việc thực hiện lại sẽ không còn khó chịu và chủ yếu được giảm về số lượng. Các cảnh có thể được thao tác với DF của các nghệ sĩ cảnh. Ngoài ra, hình ảnh do máy tính tạo ra (CGI) khá tốn kém, dẫn đến một số bộ phim quảng cáo lớn có các hiệu ứng tiêu chuẩn phụ do máy tính tạo ra. Như vậy, DF bổ sung khía cạnh tiết kiệm tài chính cho ngành này. Có thể thấy điều này khi một DF gồm các nhà báo Hàn Quốc được tạo ra để trình bày tin tức ngày hôm đó; công chúng đã được thông báo trước về nó [13,14,25,26,27].

7.2.4.2 Trường hợp 8

Khi các bộ phim được phát hành trên toàn thế giới, việc lồng tiếng cần được thực hiện. Nhưng âm thanh và chuyển động của miệng không đồng bộ và giọng nói của các diễn viên không đủ tốt. Nó làm giảm toàn bộ trải nghiệm xem phim. Tuy nhiên, với DF, âm thanh và video có thể được tổng hợp và giọng nói sẽ vẫn còn

giống với diễn viên gốc và hát nhép. TrueSync, một sản phẩm của Flawless, là công cụ thay đổi cuộc chơi trong nghệ thuật lồng tiếng bằng cách mang lại kết quả chất lượng cho các video hát nhép.

Một nhà đổi mới khác sử dụng công nghệ tổng hợp âm thanh là Marvel.ai của Veritone, cung cấp dịch vụ Voice as a Service (VaaS) để tạo và định hướng giọng nói tổng hợp cũng như dễ dàng cá nhân hóa giọng nói đó liên quan đến giới tính, giọng nói hoặc ngôn ngữ. Ứng dụng này có nhiều trường hợp sử dụng, chẳng hạn như kỹ thuật tiếp thị để tạo quảng cáo sáng tạo và nhắm mục tiêu đối tượng lớn hơn. Một công ty truyền thông có thể sử dụng điều này để chuyển đổi văn bản thành âm thanh, và bởi chính phủ và các cơ quan công cộng để tăng cường giao tiếp bằng các ngôn ngữ hoặc phương ngữ khác nhau và để tạo ra nội dung học tập điện tử thú vị.

7.2.4.3 Trường hợp 9

Deep Nostalgia, một dịch vụ AI của MyHeritage do Gilad Japhet thành lập, chuyển đổi hình ảnh (khuôn mặt) thành các video ngắn thực tế. Chỉ có chuyển động cơ bản của khuôn mặt, mắt và miệng. Ứng dụng này đã cho kết quả sống động đến mức khó có thể tin rằng đó là một video chưa từng tồn tại được tạo từ một hình ảnh. Phần mềm này đã được các nhà khoa học sử dụng trên các đạo luật để xem nó hoạt động. Xu hướng nhất ở Ấn Độ là DF của chiến binh tự do Bhagat Singh. Hình ảnh động này dựa trên bộ dữ liệu video trình điều khiển được thiết lập sẵn, từ đó phần mềm sẽ chọn hình ảnh động phù hợp nhất có thể. Do đó, hoạt ảnh hơi khác nhau đối với mọi hình ảnh được chuyển đổi thành video.

Phần mềm ứng dụng/

Chương trình theo Công ty/

Học	trường hợp sử dụng	Kỹ thuật	Trang web, nếu có
Nỗi Nhớ Sâu Thảm by di sản của tôi	Làm sống lại ký ức và khoảnh khắc	Hoạt ảnh hình ảnh sử dụng video lái xe được cài đặt sẵn	www.myheritage.com/ nỗi nhớ sâu
Marvel.ai của Veritone	Giọng nói như một dịch vụ (VaaS)	Tạo giọng nói tổng hợp	www.veritone.com/ ứng dụng/marvelai
TrueSync của Flawless	Lồng tiếng, hát nhép video	Tổng hợp âm thanh	www.flawlessai.com/ sản phẩm

7.2.5 Ngành: Giáo dục

Nếu một bức ảnh đáng giá cả ngàn lời nói, vậy còn một video thì sao? Điều gì sẽ xảy ra nếu chúng ta có thể nhìn và nghe trực tiếp các nhân vật/nhân vật nổi tiếng mà chúng ta đọc được trong sách của mình!

7.2.5.1 Trường hợp 10

Một video DF do nghệ sĩ nổi tiếng thế kỷ 20 Salvador Dali tạo ra trong một cuộc triển lãm ở Hoa Kỳ có tên Dali Lives trong một liên minh với công ty quảng cáo có tên Goodby, Silverstein & Partners (GS&P). DF nói bằng giọng nói và trọng âm của anh ấy, được tuyển chọn bằng hàng giờ nghiên cứu, phát hiện và làm việc chăm chỉ [28,29]. Cuối cùng, Salvador quay lại để chụp ảnh tự sướng với người xem. Theo mô tả của người xem, đó là một trải nghiệm siêu thực. Giọng nói trong sách nói có thể được điều chỉnh theo giới tính, độ tuổi, ngôn ngữ và giọng điệu, đặc biệt đối với những người hiểu ngôn ngữ mẹ đẻ của họ [30]. Do đó, ngôn ngữ sẽ không phải là rào cản đối với việc học.

Phần mềm ứng dụng/

Chương trình học

trường hợp sử dụng

Trang web kỹ thuật, nếu có

Triển lãm Cuộc sống Đại Lý bởi The Dali và GS&P [14,15]	DF của Salvador đại lý	Hình ảnh https://thedali.org/exhibit/dali-lives/ hoạt hình
---	---------------------------	---

7.3 TÓM TẮT

Chúng tôi đã thảo luận rất nhiều triển vọng của DF trong các ngành công nghiệp khác nhau. Ngoài ra, không nhất thiết phải có một kỹ sư phần mềm có kinh nghiệm để tạo DF. Nó được thực hiện theo sự thay đổi mô hình của mã thấp/không mã. Tuy nhiên, việc thực hiện nó nằm trong việc khắc phục những hạn chế hiện có. Những trở ngại chính đòi hỏi một bộ dữ liệu lớn với hình ảnh độ phân giải cao và card đồ họa chắc chắn. Mặc dù đã có nghiên cứu về việc tạo DF bằng một vài bức ảnh, nhưng mô hình này trước tiên được đào tạo và sau đó được sử dụng trên một tập dữ liệu nhỏ. Nếu các ngành có thể tận dụng kỹ thuật này, nó sẽ sớm mang lại một làn sóng thay đổi về cách thức hoạt động của từng ngành.

NGƯỜI GIỚI THIỆU

- [1] DFs for the Good: Một ứng dụng hữu ích của công nghệ AI gây tranh cãi, tháng 1 2021. doi:10.1007/978-3-030-51328-3_33
- [2] Godulla I., Hoffmann C. và Seibert D. (2021) Đối phó với DF – Một cuộc kiểm tra tuyển tính liên ngành về tình trạng nghiên cứu và ý nghĩa đối với nghiên cứu giao tiếp trong nghiên cứu SCM. Truyền thông và Phương tiện, 10(1), 72-96. doi:10.5771/2192-4007-2021-1-72, ISSN trực tuyến: 2192-4007.
- [3] Zhu, B., Fang, H., Sui, Y., và Li, L. (2020). DF cho nhận dạng video y tế: Bảo vệ quyền riêng tư và lưu giữ thông tin chẩn đoán. Kỷ yếu của Hội nghị AAAI/ACM về AI, Đạo đức và Xã hội, 414-420. <https://doi.org/10.1145/3375627.3375849>
- [4] Choi, E., Biswal, S. , Malin, B. , Công tước, J. , Stewart, W. và Sun, J. (2017). Tạo hồ sơ bệnh nhân rời rạc nhiều nhãn bằng cách sử dụng GAN. Trong Hội nghị Học máy cho Chăm sóc sức khỏe, 286-305. PMLR.
- [5] Zhao, H., Li, H., Maurer-Stroh, S. và Cheng, L. (2018). Tổng hợp hình ảnh vòng mạc và tế bào thần kinh với các lối đối nghịch chung. Phân tích hình ảnh y tế, 49, 14-26. doi:10.1016/j.media.2018.07.001
- [6] Chen, RJ, Lu, MY, Chen, TY, Williamson, DF và Mahmood, F. (2021). Dữ liệu tổng hợp trong ML cho y học và chăm sóc sức khỏe. Kỹ thuật y sinh tự nhiên, 5(6), 493-497. <https://doi.org/10.1038/s41551-021-00751-8>
- [7] Joyce, KE (2020, ngày 6 tháng 10). Dữ liệu tổng hợp trong chăm sóc sức khỏe thúc đẩy phân tích bệnh nhân. Quản lý dữ liệu tìm kiếm. <https://searchdataManagement.techtarget.com/feature/Dữ liệu tổng hợp trong chăm sóc sức khỏe-nâng cao-phân tích bệnh nhân>
- [8] Coffey, KR, Marx, RG và Neumaier, JF (2019). DeepSqueak: Một hệ thống dựa trên học tập sâu để phát hiện và phân tích cách phát âm siêu âm. Khoa tâm thần kinh, 44, 859-868. <https://doi.org/10.1038/s41386-018-0303-6>
- [9] Gaur, L., Afaq, A., Solanki, A., Singh, G., Sharma, S., Jhanjhi, NZ, . Le, D. (2021). Tận dụng dữ liệu lớn và công nghệ 5G mang tính cách mạng: Trích xuất và trực quan hóa xếp hạng cũng như đánh giá về chuỗi khách sạn toàn cầu. Máy tính và Kỹ thuật điện, 95, 107374, ISSN 0045-7906. doi:10.1016/j.compeleceng.2021.107374

- [10] Gaur, L., Bhatia, U., Jhanjhi, NZ, Muhammad, G., và Masud, M. (2021). Phát hiện COVID-19 dựa trên hình ảnh y tế bằng cách sử dụng mạng thần kinh tích chập sâu. Hệ thống đa phương tiện, 1-10. doi:10.1007/s00530-021-00794-6
- [11] Thông cáo báo chí Bước đột phá về công nghệ của Rolls Royce mang lại hy vọng cho những người bị câm lặng vì khuyết tật. Bánh cuốn. (nd). www.rolls-royce.com/media/press-releases/2019/18-12-2019-cong-nghiep-dot-bac-cung-cap-hy-vong-cho-nhung-nguoi-bi-khuynh-tam-im-lang.aspx
- [12] Thông tin chi tiết về CB. (2021, ngày 29 tháng 6). Các thương hiệu và nhà bán lẻ có nên áp dụng phương tiện tổng hợp - AI cho nội dung số không? Nghiên cứu chuyên sâu về CB. www.cbinsights.com/research/what-is-synthetic-media/
- [13] Anshu, K., Gaur, L., và Khazanchi, D. (2017). "Đánh giá mức độ hài lòng của các nhà bán lẻ điện tử tạp hóa bằng cách sử dụng mô hình TOPSIS và ECCSI mờ trực quan," Hội nghị quốc tế về Công nghệ Infocom và Hệ thống không người lái (Xu hướng và Định hướng Tương lai) (ICTUS), trang 276-284. doi:10.1109/ICTUS.2017.8286019
- [14] Gaur, L., và Anshu, K. (2018). Phân tích sở thích của người tiêu dùng cho các trang web sử dụng e TailQ và AHP. Tạp chí Kỹ thuật & Công nghệ Quốc tế, 7(2.11), 14-20.
- [15] Gaur L., Agarwal V., và Anshu, K. (2020). "Phương pháp tiếp cận DEMATEL mờ để xác định các yếu tố ảnh hưởng đến hiệu quả của ngành bán lẻ Ấn Độ," Đàm bảo hệ thống chiến lược và Phân tích kinh doanh. Phân tích nội dung (Quản lý hiệu suất và an toàn). Springer, Singapore. https://doi.org/10.1007/978-981-15-3647-2_
- [16] Grey, C. (2020, ngày 25 tháng 8). Thêm vào giỏ hàng: Tại sao DF lại tốt cho bán lẻ. Tin tức quảng cáo. www.adnews.com.au/news/add-to-cart-why-DFs-are-good-for-retail
- [17] Sahu, G., Gaur, L., và Singh, G. (2021). Áp dụng phương pháp tiếp cận lý thuyết thích hợp và hài lòng để kiểm tra niềm đam mê của người dùng đối với các nền tảng vượt trội và truyền hình thông thường. Viễn thông và Tin học, 65, 101713. doi:10.1016/j.tele.2021.101713
- [18] Ramakrishnan, R., Gaur, L., và Singh, G. (2016). Tính khả thi và hiệu quả của các thiết bị IoT đèn hiệu BLE trong quản lý hàng tồn kho tại khu vực cửa hàng. Tạp chí Quốc tế về Kỹ thuật Điện và Máy tính, 6(5), 2362-2368. doi:10.11591/ijece.v6i5.10807
- [19] Afaq, A., Gaur, L., Singh, G., và Dhir, A. (2021). COVID-19: Thay đổi hành vi của hành khách đi máy bay và định hình lại kỳ vọng của họ đối với ngành hàng không. Nghiên cứu Giải trí Du lịch, 1-9. doi:10.1080/02508281.2021.2008211
- [20] Gaur, L., Singh, G., Solanki, A., Jhanjhi, NZ, Bhatia, U., Sharma, S., . và Kim, W. (2021). Xu hướng của giới trẻ trong việc dự đoán các mục tiêu phát triển bền vững bằng thuật toán rứng ngẫu nhiên và thần kinh mờ. Khoa học thông tin và máy tính lấy con người làm trung tâm, 11, NA.
- [21] Singh, G., Kumar, B., Gaur, L. và Tyagi, A. (2019). "So sánh giữa Multinomial và Bernoulli Naïve Bayes để phân loại văn bản," Hội nghị quốc tế về tự động hóa, tính toán và quản lý công nghệ (ICACTM), trang 593-596. doi:10.1109/ICACTM.2019.8776800
- [22] Gaur, L., Afaq, A., Singh, G., và Dwivedi, YK (2021). Vai trò của trí tuệ nhân tạo và người máy trong việc thúc đẩy du lịch không chặng trong thời kỳ đại dịch: Chương trình nghiên cứu và đánh giá. Tạp chí Quốc tế về Quản lý Khách sạn Đương đại, 33(11), 4079-4098. <https://doi.org/10.1108/IJCHM-11-2020-1246>
- [23] Sharma, S., Singh, G., Gaur, L., và Sharma, R. (2022). Khoảng cách tâm lý và tôn giáo có ảnh hưởng đến hành vi lừa đảo của khách hàng không? Tạp chí Nghiên cứu Người tiêu dùng Quốc tế. doi:10.1111/ijcs.12773
- [24] Dietmar, J. (2019, ngày 21 tháng 5). Bài đăng của hội đồng: GAN và DF có thể cách mạng hóa ngành thời trang. Forbes. www.forbes.com/sites/forbestechcouncil/2019/05/21/gans-and-DFs-could-revolutionize-the-fashion-industry/?sh=4f638513d17f
- [25] Debusmann Jr, B. (2021, ngày 8 tháng 3). "DF là tương lai của sáng tạo nội dung." BBC Tin tức. www.bbc.com/news/business-56278411

- [26] Sharma, DK, Gaur, L., và Okunbor, D. (2007). Nén ảnh và trích xuất đặc trưng với mạng nơ-ron. Kỷ yếu của Viện Khoa học Thông tin và Quản lý, 11(1), 33-38.
- [27] Rana, J., Gaur, L., Singh, G., Awan, U., và Rasheed, MI (2021). Củng cố hành trình của khách hàng thông qua trí tuệ nhân tạo: Chương trình nghiên cứu và đánh giá. Tạp chí quốc tế về các thị trường mới nổi, Tập. trước khi in (No. trước khi in). <https://doi.org/10.1108/IJOEM-08-2021-1214>
- [28] Lee, D. (2019, ngày 10 tháng 5). DF Salvador Dalí chụp ảnh tự sướng với khách tham quan bảo tàng. The Verge. www.theverge.com/2019/5/10/18540953/salvador-dali-lives-DF-museum
- [29] Westerlund, M. (2019). Sự xuất hiện của công nghệ DF: Đánh giá. Đánh giá quản lý đổi mới công nghệ, 9(11), 39-52. <https://doi.org/10.22215/timreview/1282>
- [30] Silbey, J. và Hartzog, W. (2018). Mặt trái của Deep Fakes. 78 Md. L. Rev. 960 (2019).

Các mối đe dọa và thách thức 8 bởi Công nghệ DeepFake

Mamta Sareen

NỘI DUNG

8.1 Giới thiệu	99
8.2 DF và sự phát triển của chúng	100
8.3 Mối đe dọa của DF	102
8.4 Các mối đe dọa đối với An ninh Quốc gia.....	102
8.5 Mối đe dọa đối với các cá nhân	103
8.6 Mối đe dọa đối với xã hội	103
8.7 Các mối đe dọa đối với hệ thống tư pháp.....	104
8.8 Các mối đe dọa đối với chính trị hoặc diễn ngôn dân chủ.....	105
8.9 Đe dọa Bầu cử.....	105
8.10 Các mối đe dọa đối với doanh nghiệp.....	106
8.11 Mối đe dọa làm xói mòn niềm tin vào các tổ chức	107
8.12 Các biện pháp đối phó	108
8.13 Các biện pháp pháp lý.....	108
8.14 Các biện pháp đối phó về công nghệ.....	108
8.15 Tóm tắt	111
Người giới thiệu.....	111

8.1 GIỚI THIỆU

Những tiến bộ gần đây trong công nghệ AI và DL đã tạo ra những bước tiến quan trọng trong nhận dạng hình ảnh. Những công nghệ này, với khả năng tạo bộ dữ liệu hình ảnh, nâng cao độ phân giải hình ảnh, có nhiều dự đoán video đặc biệt hơn và biểu đồ ảnh chân thực, đồng thời đã mang lại sự thay đổi đáng kể trong lĩnh vực xử lý hình ảnh.

Một trong những công nghệ như vậy là công nghệ DeepFake (DF) sử dụng Mạng đối thủ chung (GAN), nơi hai mạng thần kinh sâu được đào tạo để hoạt động song song.

Các mạng này lặp đi lặp lại giữa hình ảnh chính hàng/mẫu và hình ảnh được tạo theo thông kê cho đến khi sự khác biệt của chúng là tối thiểu. Phương pháp này cho thấy khả năng phát triển ảnh giả hoặc tạo ảnh giả bằng cách thay thế khuôn mặt của mọi người bằng khuôn mặt khác. Công nghệ này đã mang đến một cách tiếp cận mang tính cách mạng để tạo video biến hình nhanh hơn nhiều. Một ví dụ điển hình về công nghệ DF là của một cầu thủ bóng đá nổi tiếng, David Beckham, người nói thông thạo 9 thứ tiếng [1]. Mục đích ban đầu của việc tạo những video như vậy là vô hại và nhằm mục đích cung cấp một công cụ nghiên cứu tạo video nhân tạo có thể hữu ích cho phim ảnh, kể chuyện và các dịch vụ đa phương tiện hiện đại khác.

Tuy nhiên, công nghệ này vẫn phổ biến cho đến giữa những năm 2010.

Tuy nhiên, một video tổng hợp của cựu tổng thống Hoa Kỳ Barack Obama được phát hành vào năm 2017, cho thấy ông ấy nói những lời từ một bản nhạc thay thế, đã trở nên lan truyền và mở đường cho sự phát triển bùng nổ của DF trong thời điểm hiện tại [2]. Công nghệ này đã nở rộ như một bông hoa đột và có mặt ở khắp mọi nơi, ngay cả đối với những người nghiệp dư. Nó cho họ cơ hội để nghịch ngợm. Một ví dụ điển hình là một video giả mạo cho thấy cựu Tổng thống Hoa Kỳ Barack Obama đưa ra một số tuyên bố khác thường đã nhận được 5 triệu lượt xem và hơn 83.000 lượt chia sẻ trên Facebook và các nền tảng truyền thông xã hội khác [3,4,5]. Tuy nhiên, video này là một video giả mạo sử dụng trí tuệ nhân tạo AI tổng hợp khuôn mặt diễn viên hài người Mỹ Jordan Peele đóng giả Obama bằng giọng nói của ông.

Một ví dụ khác là một video liên quan đến Tổng thống Mỹ Donald Trump, trong đó ông đưa ra tuyên bố đầy thách thức để Bỉ bị khiêu khích rút khỏi hiệp định Paris. Một đảng chính trị của Bỉ đã công bố đoạn video trên Twitter và Facebook nhưng cuối cùng đã bị Lead Stories [6] vạch trần. Người tạo có ý định thu hút sự chú ý của mọi người và hướng họ đến một bản kiến nghị trực tuyến. Cuộc biểu tình quy mô nhỏ này đủ để gây lo ngại về cách DF có thể đe dọa các hệ thống toàn cầu vốn đã dễ bị tổn thương. DF, cùng với sự lan truyền rộng rãi của chúng thông qua các nền tảng truyền thông xã hội, đã gây ra những lo ngại đáng kể trên toàn thế giới về độ tin cậy của phương tiện kỹ thuật số. Tin tức liên tục về tác động của các video DF này đối với cá nhân và tổ chức đã đưa ra nhiều thách thức bảo mật khác nhau. Bài báo này là một nỗ lực khiêm tốn để chỉ ra nhiều mối đe dọa và thách thức do công nghệ DF đặt ra.

8.2 DFS VÀ SỰ TĂNG TRƯỞNG CỦA CHÚNG

Năm 1990, Gruber, thông qua thí nghiệm mang tính bước ngoặt của mình, đã chứng minh tác động của giao tiếp bằng hình ảnh là rất lớn và có tầm ảnh hưởng [7]. Grabe và Bucy vào năm 2009 [8] và Trước đó vào năm 2013 [9] đã chứng minh thêm ý tưởng của họ về giao tiếp bằng hình ảnh, điều này thể hiện mức độ nhở lại và tác động cao hơn của những người dùng được cung cấp hỗ trợ trực quan. Những nghiên cứu này đã chứng minh rằng hình ảnh gây ra phản ứng mạnh hơn lời nói và giúp người dùng tương tác với nội dung, từ đó ảnh hưởng đến việc lưu giữ thông tin. Mọi người thấy hình ảnh và nội dung nghe nhìn dễ hiểu hơn văn bản viết vì điều này mang lại cho họ "trải nghiệm siêu nhận thức" [10]. Trong thế giới ngày nay, thông tin đóng một vai trò quan trọng và mọi người thường không có thời gian và năng lượng để kiểm tra tính xác thực của thông tin họ có được. Vì vậy, nhận thức thị giác trực tiếp thường được coi là đáng tin cậy. Thông qua những hình ảnh này, công chúng hoặc người tiêu dùng có được kiến thức cụ thể giúp họ đưa ra quyết định quan trọng về con người, thể chế hoặc tổ chức và thậm chí cả các nhà lãnh đạo. Người tiêu dùng của những video này được cho là chấp nhận bằng chứng video sâu sắc hơn các nguồn thông tin khác. Do đó, các video tỏ ra có lợi khi cần có thỏa thuận tập thể về một chủ đề [11].

Công nghệ DF sử dụng thẩm quyền giao tiếp bằng hình ảnh này để tạo ra một dạng thông tin sai lệch về hình ảnh bằng cách tác động đến cảm xúc và nhận thức của mọi người.

Một video giả mạo đúng thời điểm về một nhà lãnh đạo có ảnh hưởng nói về phân biệt chủng tộc hoặc khủng bố có khả năng gây chấn động khắp các phương tiện truyền thông thế giới. Một số yếu tố có xu hướng thúc đẩy ngọn lửa DF. Một số trong số họ có thể được nêu như sau:

Dòng thông tin: David et al. tảng thông tin được mô tả là cơ sở của nội dung lan truyền. Nó xảy ra khi mọi người không tin vào thông tin của họ mà dựa nhiều hơn vào việc xác thực thông tin của người khác và sau đó chuyển thông tin đó đi xa hơn [12].

Các nền tảng truyền thông xã hội cung cấp môi trường phù hợp cho các tảng thông tin truyền bá nội dung của nó và đưa nội dung đó đến với đông đảo khán giả. Các nhà hoạt động Black Lives Matter hay phong trào Never Again của học sinh trường trung học Parkland chỉ là một vài ví dụ về dòng thông tin [13,14]. Theo nghiên cứu, những trò lừa bịp và tin đồn giả có thể đến với mọi người nhanh gấp 10 lần so với những câu chuyện có thật [15]. Người ta đã chứng minh rằng khả năng ghi nhớ của thông tin tiêu cực cao hơn thông tin tích cực [15]. Được hỗ trợ bởi xu hướng tự nhiên của con người đối với một số kích thích nhất định như khiêu dâm, buôn chuyện và bạo lực, DF chỉ cung cấp "môi trường phù hợp" để thúc đẩy xu hướng này. Những cá nhân chơi ác ý với DF có thể nhanh chóng tiếp cận một lượng lớn khán giả, thậm chí trên toàn cầu.

Thời gian lưu thông: Thời gian nguy hiểm là ý tưởng cơ bản đằng sau việc tạo ra DF.

Trong mọi trường hợp, một số thời điểm nhất định là quan trọng hoặc nhạy cảm và có thể nâng cao kết quả của sự kiện theo bất kỳ hướng nào. Giả sử bất kỳ video giả mạo nào hiển thị một số sự thật không thể kiểm chứng ngay lập tức được lưu hành giữa những người dùng vào thời điểm nhạy cảm. Trong trường hợp đó, rất có thể nó có thể để lại tác động tức thời và to lớn lên tâm trí của người nhận và có thể là công cụ thay đổi quan điểm của họ về sự việc được hiển thị. Việc phân phối các video giả mạo được thiết lập để hạn chế đến mức vào thời điểm video bị lật tẩy, tác động không thể thay đổi đã được tạo ra. Những video bị bóp méo có chủ ý này được lưu hành trên mạng xã hội có thể che mờ thực tế tại một thời điểm. Ví dụ, thời gian ngay trước ngày bỏ phiếu trong bất kỳ cuộc bầu cử nào là rất quan trọng. Giả sử bất kỳ video giả mạo nào cho thấy hành vi tham nhũng của một ứng cử viên (mà anh ta không làm) được lan truyền giữa các cử tri trước khi được xác minh công khai. Trong trường hợp đó, nó có thể ảnh hưởng đến suy nghĩ của cử tri và có thể có tác động hệ quả đến kết quả bầu cử.

Tốc độ lan truyền: Đã qua rồi cái thời mà khả năng phân phối thông tin hình ảnh dưới dạng âm thanh hoặc video của một cá nhân hoặc tổ chức bị hạn chế. Cuộc cách mạng thông tin đã thay đổi mạnh mẽ mô hình phân phối nội dung. Ngày nay, nhiều nền tảng trực tuyến tạo điều kiện thuận lợi cho kết nối toàn cầu, từ đó dân chủ hóa quyền truyền thông tin ở một mức độ chưa từng có. Nội dung được phân phối có thể đến với nhiều khán giả quốc tế. Tốc độ lưu hành video giả mạo trực tuyến đặt nền tảng cho hiệu quả của nó.

Kiến thức kỹ thuật số: Một số nhà nghiên cứu tin rằng những quốc gia đang phát triển nơi tỷ lệ hiểu biết kỹ thuật số tương đối thấp sẽ dễ trở thành nạn nhân của thông tin sai lệch do DF gây ra. Người tiêu dùng phương tiện truyền thông ở những quốc gia này không thông thạo các kỹ thuật phân biệt thật giả và khả năng họ chấp nhận giả tạo là sự thật là rất cao. Trong một vụ việc như vậy, bạo lực gây ra cho cộng đồng người Rohingya ở Myanmar là do các bài viết giả mạo lan truyền trên Facebook [16]. Tương tự như vậy, DF có thể được sử dụng như một công cụ khuất phục bởi các chế độ độc tài hoặc các nhóm cực đoan nhằm kích động sự chia rẽ xã hội.

Vào năm 2021, Cục Điều tra Liên bang (FBI) đã đưa ra cảnh báo về nội dung truyền thông giả mạo như một cuộc tấn công mạng mới được xác định gây thiệt hại đáng kể về tài chính.

và ảnh hưởng uy tín đến các tổ chức và xã hội nói chung [17]. Hiểu các mối đe dọa cơ bản do DF gây ra là điều cần thiết để xây dựng bất kỳ biện pháp đối phó nào để chống lại chúng.

8.3 NGUY HIỂM CỦA DFS

Động cơ chính của người tạo ra DF là truyền bá thông tin sai lệch và khiêu người tiêu dùng tin vào những gì đã được hiển thị cho họ. Ngành công nghiệp điện ảnh sử dụng công nghệ này cho các hiệu ứng đặc biệt và hình ảnh động dường như vô hại. Tuy nhiên, công nghệ này hiện đang được sử dụng cho mục đích bất chính bởi những tên tội phạm am hiểu công nghệ. Công nghệ DF dường như đã giới thiệu một loại phương tiện mới mà những người dùng độc hại đang sử dụng để làm lợi cho họ. DF có thể đe dọa các cuộc bầu cử chính trị, an ninh mạng, tài chính cá nhân và doanh nghiệp, danh tiếng, v.v., trong số những rủi ro tiềm ẩn.

Dưới đây là một số mối đe dọa có thể xảy ra do công nghệ DF.

8.4 Đe dọa đến AN NINH QUỐC GIA

DF có thể gây ra mối đe dọa đáng kể đối với bất kỳ an ninh quốc gia nào nếu được triển khai một cách ác ý bởi bàn tay của những kẻ xấu. Nhiều nhà nghiên cứu, học viên và đại diện chính phủ của các quốc gia khác nhau đang lo lắng về những tác động liên quan đến an ninh quốc gia của DF. Các nhà lập pháp ở Hoa Kỳ đang bày tỏ quan ngại về các chiến dịch tung tin thất thiệt trong các cuộc bầu cử ở Hoa Kỳ có khả năng làm trầm trọng thêm sự chia rẽ chính trị, xã hội trong xã hội và đe dọa đến an ninh quốc gia [18]. Theo Giám đốc Trung tâm AIC của Lầu Năm Góc, "DFs là một vấn đề an ninh quốc gia" [19].

Ngoài ra, Giám đốc Tình báo Quốc gia, Daniel R. Coats, đã trích dẫn rằng "Các đối thủ và đồng minh cạnh tranh chiến lược có thể sẽ cố gắng sử dụng các công nghệ giả sâu hoặc công nghệ máy học tương tự để tạo ra các tệp hình ảnh, âm thanh và video thuyết phục - nhưng sai sự thật nhằm tăng cường các chiến dịch gây ảnh hưởng. chống lại Hoa Kỳ và các đồng minh và đối tác của chúng ta" [20]. Nhiều nhà bình luận, chuyên gia và nhà phân tích cũng tán thành những kết luận này.

Thông tin sai lan truyền qua DF có thể gây nguy hiểm cho an ninh quốc gia theo nhiều cách. Ví dụ: thông tin sai lệch có thể gây nguy hiểm cho sự an toàn của lực lượng quân sự của bất kỳ quốc gia nào làm việc với bất kỳ dân thường nước ngoài nào nếu bất kỳ DF nào cho thấy các thành viên quân đội tấn công hoặc giết thường dân được lưu hành. Bất kỳ kẻ ác ý nào cũng có thể lợi dụng sự bất ổn của một khu vực (ngày nay khá phổ biến) bằng cách phát tán nội dung giả mạo thông qua DF để làm trầm trọng thêm người dân địa phương, dẫn đến thương vong cho dân thường. Các chế độ nước ngoài thù địch cũng có thể sử dụng DF để tuyên truyền, cho thấy các nhà lãnh đạo thế giới đang hành xử công kích hoặc thù địch. Ví dụ, Chesney và Citron, trong bài báo của họ, đã đặt ra một kịch bản về hậu quả của một video giả mạo về một vị tướng Mỹ ở Afghanistan đốt một cuốn kinh Koran. Hiện tại, chúng ta đang sống trong một thế giới sẵn sàng cho bạo lực; thông tin sai lệch như vậy có thể trở thành một công cụ mạnh mẽ để kích động tiềm năng [21]. Đây chỉ là một số trường hợp được nêu trong đó một DF thực tế cao có thể gây ra mối đe dọa duy nhất đối với an toàn công cộng và an ninh quốc gia.

8.5 Đe dọa đối với các cá nhân

Lịch sử đã chứng kiến hậu quả của những lời dối trá về những gì con người đã làm hoặc nói.

Với độ tin cậy vốn có và khả năng che giấu vai trò sáng tạo của kẻ nói dối, công nghệ DF trở thành một công cụ mạnh mẽ để khai thác hoặc phá hoại người khác. Vì công nghệ này cho phép điều khiển giọng nói, khuôn mặt và cơ thể của một cá nhân trong video, nên nó sẽ hạn chế ít cơ hội hơn cho những người dùng ác ý khai thác danh tính của họ để thỏa mãn ác ý.

Nó có thể giáng một đòn mạnh vào bất kỳ cá nhân nào trong bất kỳ lĩnh vực cạnh tranh nào, cho dù đó là nơi làm việc, thể thao, chính trị, chuyện tình cảm hay cuộc sống cá nhân. Sự kết hợp đúng đắn giữa tăng cường thuật toán, xu hướng nhận thức và các công cụ tìm kiếm ngày càng cải tiến làm tăng lưu lượng cho những tin giả tai tiếng như vậy.

Thế giới ảo thường xuyên phải đối mặt với các video sex DF được thiết kế chủ yếu nhằm mục đích thỏa mãn tình dục hoặc tài chính của người sáng tạo. Một ví dụ điển hình là nỗ lực bôi nhọ nhân phẩm của nhà báo Rana Ayyub bằng cách biến khuôn mặt của cô ấy thành cơ thể của một phụ nữ khác và tạo ra một video khiêu dâm giả dài 2 phút [22]. Điều này là để trả đũa bài báo của cô ấy chống tham nhũng trong nền chính trị theo chủ nghĩa dân tộc của người theo đạo Hindu. Một hành động như vậy gây tổn hại tâm lý trực tiếp cho bất kỳ ai và phá hoại danh tiếng của họ, do đó gây ra thiệt hại ở các không gian khác. Các video DF có thể là một công cụ hữu ích để những kẻ tống tiền khai thác và trích xuất thứ gì đó có giá trị từ chúng. Ngay cả khi ai đó không còn gì để mất, việc hoàn tác thiệt hại ban đầu do DF gây ra có thể buộc mọi người phải khuất phục trước mối đe dọa và cung cấp tiền, thông tin, v.v. Như đã thấy trong ví dụ trước, DF có thể khai thác bản dạng giới tính của một cá nhân. Người ta sợ rằng các video sex DF có thể buộc các cá nhân phải quan hệ tình dục ảo và biến các mối đe dọa hiếp dâm thành một thực tế áo đáng sợ. Ngoài ra, DF mô tả hành vi bạo lực hoặc lạm dụng của một cá nhân có thể được sử dụng để đe dọa, bắt nạt hoặc gây tổn hại tâm lý cho cá nhân được nhắm mục tiêu.

Với sự hỗ trợ của công nghệ giúp dễ dàng sao chép và lưu trữ trong các phiên tòa từ xa, việc loại bỏ những tin giả mạo này trở nên khó khăn hơn một khi chúng được đăng và chia sẻ. Tùy thuộc vào thời gian, tình huống và lưu hành các video của DF, các tác động có thể rất nghiêm trọng. Nó có thể dẫn đến việc đánh mất cơ hội duy nhất hoặc mất đi sự hỗ trợ của bạn bè hoặc bị từ chối thăng chức hoặc bị từ chối trong chuyện tình cảm hoặc hủy bỏ cơ hội kinh doanh, v.v. Đôi khi, việc vạch trần tính xác thực của hàng giả có thể đến quá muộn để khắc phục tác hại ban đầu.

8.6 Đe dọa đối với xã hội

Công nghệ DF tác động đến các cá nhân và có thể gây ra những hậu quả tàn khốc đối với xã hội nếu không được kiểm soát kịp thời. DF có thể ảnh hưởng đến tình cảm và nhận thức của mọi người và nếu được thực hiện một cách ác ý, có thể tác động tiêu cực đến cộng đồng.

Vào năm 2018, một video đã lan truyền trên mạng xã hội Ấn Độ quay cảnh một đứa trẻ đang chơi ở Bangalore thì bị hai người đàn ông bắt cóc [23]. Mặc dù cái gọi là vụ bắt cóc được thể hiện trong video này là không có thật, nhưng nó đã tạo ra sự hoang mang và hoảng loạn lan rộng, dẫn đến bạo lực đám đông kéo dài 8 tuần cướp đi sinh mạng của ít nhất 9 người vô tội. Hãy xem xét các tình huống sau đây và tác động đáng kinh ngạc của chúng đối với xã hội:

- Video giả mạo có cảnh các quan chức nhà nước tham gia vào các hoạt động phi pháp như nhận hối lộ, ngoại tình, phát biểu thù địch, v.v.
- Các video giả mạo nhắm vào những nhân vật nổi tiếng hoặc có uy tín trong xã hội bằng cách tạo ra các clip khiêu dâm đã được chỉnh sửa.
- Phát tán các video giả mạo liên quan đến các chính trị gia và các quan chức chính phủ khác trong các tình huống hoặc địa điểm mà họ không có mặt và nói hoặc làm những điều xấu xa mà họ không làm. Ví dụ: nó có thể tạo ra các video họ hợp tác với gián điệp hoặc tội phạm.
- Các video giả mạo về những người lính sát hại những người vô tội có thể gây ra làn sóng bạo lực trong nhân dân và trong trường hợp xấu nhất có thể dẫn đến bất tuân dân sự.
- Các video giả mạo thể hiện sự tàn bạo đối với một đẳng cấp hoặc chủng tộc cụ thể có thể làm tăng thêm sự chia rẽ xã hội hiện có và kích hoạt các hành động hoặc thậm chí là bạo lực.
- Âm thanh hoặc video giả mạo có thể kích động giới trẻ nhận ra mức độ huy động hành động mà chỉ bằng văn bản không thể đạt được.

DF có thể khiến mọi người có niềm tin sai lầm, tức là họ có thể tin rằng những thứ giả mạo này là thật. Điều này có thể gây ra những hậu quả thực tế nghiêm trọng vì DF không chỉ gây ra những tiền đề sai lầm mà còn khắc sâu thói quen phá hoại sự biện minh cho niềm tin chân chính.

Do đó, có thể có mức độ tin cậy thấp đối với các video thực từ các phương tiện truyền thông tin tức hợp pháp [21,24]. Sự hiện diện của những video giả mạo như vậy trên mạng xã hội và các phương tiện điện tử khác có thể làm mất đi nền tảng cơ bản của bất kỳ xã hội nào và gây ra những cái giá phải trả. Những chi phí này có thể là vô hình trong việc duy trì/cải thiện sự phân chia xã hội hoặc chi phí hữu hình cho những người đã bị lừa vào một số hành động nhất định và những người phải chịu đựng những hành động đó.

8.7 Đe dọa đối với HỆ THỐNG TƯ PHÁP

DF vẫn là một chủ đề mới nổi trong luật. Mặc dù nghiên cứu của DF bắt đầu từ đầu năm 2014, nhưng tác động thực sự đối với hệ thống tư pháp hiện đang trở thành vấn đề được các học giả pháp lý và các nhà hoạch định chính sách quan tâm. Bất kỳ hệ thống tư pháp nào cũng phụ thuộc rất nhiều vào các mẫu bằng chứng. Một trong những mối quan tâm chính là mối đe dọa do DF đặt ra liên quan đến việc giả mạo bằng chứng [25]. Điều này vô tình có thể đặt ra vấn đề cho tòa án trong việc chấp nhận tính hợp pháp của các đương sự hoặc nhân chứng. Làn thé nào DF giả mạo, bằng chứng thậm chí có thể kết án người vô tội hoặc tha thứ cho kẻ có tội. Có cảm giác rằng các vấn đề không lường trước có thể phát sinh trong quá trình kiểm tra chéo khi một bên liên quan làm chứng có lợi cho các chi tiết của video DF và bên kia (đối lập) phủ nhận tính xác thực của video đó. Điều này sẽ tác động tiêu cực đến các vụ kiện của tòa án và áp đặt thêm chi phí cả về thời gian và tiền bạc trong việc xác thực bằng chứng được đưa ra. Một ví dụ như vậy là một trường hợp về quyền nuôi con ở Vương quốc Anh, trong đó người mẹ đã tạo một tệp âm thanh DF của người cha để hiện hành vi bạo lực để làm bằng chứng [26]. Mặc dù bằng chứng này đã bị loại bỏ sau khi kiểm tra pháp y, nhưng nó đặt ra nhiều câu hỏi về tính hợp lý của một phán quyết sai lầm.

DF cũng có thể gây ra khối lượng vụ việc bổ sung tại các tòa án với tổng chi phí của các trường hợp trong đó video DF làm phát sinh khiếu nại về tội nhẹ. Họ cũng có thể ảnh hưởng đến các tòa án bằng cách đóng vai trò phụ trong các tranh chấp mà họ không gây ra bằng cách chỉ thêm một bằng chứng khác vào vụ kiện tụng và làm sai lệch quá trình phán quyết. Các video giả mạo có thể vô tình hoặc cố ý kết thúc quá trình lưu trữ được coi là đáng tin cậy trong lịch sử, chẳng hạn

như những tin tức truyền thông. Giả sử người quản lý các bản ghi phương tiện đó không phát hiện kịp thời DF. Trong trường hợp đó, có rủi ro là người giám sát có thể vô tình chứng minh cho DF khi được yêu cầu xác thực bằng chứng trong một thủ tục tố tụng tại tòa án. Hơn nữa, vì DF rất phổ biến nên quy trình tổng thể để xác minh tính xác thực của bằng chứng thực tế có thể khiến bồi thẩm đoàn/thẩm phán nghi ngờ. Do đó, nếu bồi thẩm đoàn/thẩm phán được giao phó vai trò quan trọng của sự thật bắt đầu nghi ngờ và hoài nghi, thì nền tảng của hệ thống tư pháp có thể bị lung lay.

8.8 CÁC ĐE DỌA ĐỐI VỚI CHÍNH TRỊ HOẶC DIỄN GIẢI DÂN CHỦ

Bất kỳ nền dân chủ nào cũng chủ yếu dựa vào tuyên truyền trực tuyến vì đây là công cụ rẻ nhất để tạo ra mức độ ảnh hưởng cao. DF được tạo nhanh chóng và dễ dàng lưu hành tới nhiều đối tượng và do đó có thể được sử dụng một cách có ý hoặc vô tình để cung cấp thông tin sai lệch cho công chúng vì lợi ích chính trị. Bởi vì một người bình thường khó phân biệt giữa video thật và video DF, những trò giả mạo này có thể làm thay đổi cảm giác thực tế của video. Một đoạn video đã được chỉnh sửa của một chính trị gia người Mỹ, Nancy Pelosi, đã trở nên lan truyền trên mạng xã hội, cho thấy cô ấy đã phát âm sai lời nói của mình trong tình trạng say xỉn. Khi được chia sẻ bởi Tổng thống Hoa Kỳ lúc bấy giờ là Donald J. Trump, tin giả này đã nhận được hơn 2,5 triệu lượt chia sẻ chỉ trên Facebook [27]. Những hàng giả như vậy có thể làm thay đổi nhận thức của công chúng và vào thời điểm chúng được chứng minh là hàng giả, thiệt hại dường như đã được thực hiện. Một màn bắt chước Joe Biden gần đây cho thấy anh ta không biết mình đang ở bang nào đã nhận được 1 triệu lượt xem trên Twitter [28].

Công nghệ của DF có thể gieo rắc lượng thông tin sai lệch chưa từng có và hạ thấp mức độ tin tưởng của cử tri vào nền dân chủ [29]. DF tấn công vào chính nền tảng của nền dân chủ và có thể được gọi là phản dân chủ vốn có. Mặc dù những hình ảnh huyễn hoặc thường được sử dụng để hạ bệ các đối thủ chính trị, nhưng DF mang đến một ảo giác thuyết phục hơn những hình ảnh đơn lẻ. Không khó để một cơ quan truyền thông của chính phủ được tài trợ tốt để tạo ra các DF nhục nhã của những cá nhân đặt ra các rào cản chính trị. Những giả mạo như vậy có thể gây khó khăn cho các phong trào đối lập.

Hơn nữa, diễn ngôn công khai thường được sử dụng để hiểu rõ hơn về mọi người về các vấn đề chính sách khác nhau. Đôi khi những lời đối trả do DF lan truyền có ý định thách thức tính chính trực và độ tin cậy của những người tham gia trong các cuộc tranh luận như vậy. Vào những thời điểm khác, những lời nói đối này mạnh đến mức có thể làm xói mòn nền tảng thực tế của diễn ngôn chính sách. DF có thể cung cấp “bằng chứng” cho những người đang tìm cách khắc phục thêm sự bất đồng về nhận thức của họ xung quanh thông tin sai lệch. Diễn ngôn dân chủ hoạt động tốt nhất khi các cuộc thảo luận/tranh luận dựa trên sự thật được chia sẻ và sự thật được hỗ trợ bởi bằng chứng thực nghiệm. Trong trường hợp không có những sự thật xác minh như vậy, những nỗ lực giải quyết các vấn đề lớn của quốc gia vẫn bị vướng vào những câu hỏi cấp một không cần thiết. DF đã gây ra sự xói mòn trên quy mô lớn niềm tin của công chúng vào các sự kiện và số liệu thống kê được trình bày và có thể gây khó khăn trong các thủ tục tranh luận dân chủ. Thật khó để các sự thật trung thực xuất hiện từ mớ hỗn độn của các DF tràn ngập.

8.9 NGUY CƠ BẦU CỬ

Các cuộc bầu cử là thời điểm có rủi ro cao và một DF thực tế có khả năng tác động trực tiếp đến cử tri. Một DF đúng lúc có thể can thiệp vào các cuộc bầu cử tiểu bang hoặc trung tâm bằng cách lan truyền những điều hư cấu và mơ hồ đáng kể về cuộc sống cá nhân và chính sách của các ứng cử viên

vị trí trong quá trình bầu cử. Sự không chắc chắn như vậy có khả năng làm mất ổn định hệ quả của bất kỳ cuộc bầu cử nào. Các DF này được tính thời gian trong một khoảng thời gian cụ thể, cho phép đủ thời gian để nó lưu hành nhưng không đủ khung thời gian để gỡ lỗi nó một cách hiệu quả.

Khoảng thời gian giới hạn này có khả năng che mờ thực tế và nghiêng kết quả của bất kỳ cuộc bầu cử nào theo hướng có lợi cho bất kỳ ai. Một nghiên cứu tại Vương quốc Anh cho thấy DF có xu hướng khiến mọi người nhầm lẫn về thông tin thật hay giả lưu hành trên internet và do đó, họ có xu hướng ít tin tưởng hơn vào tin tức trên mạng xã hội [30].

Hơn nữa, càng ít người tin tưởng vào các phương tiện truyền thông tin tức thì họ càng có nhiều khả năng rơi vào thông tin sai lệch trôi nổi trên mạng xã hội hoặc phương tiện điện tử, điều này có thể ảnh hưởng đến hành vi bỏ phiếu của họ [31]. Nó có thể có tác động khá lớn nếu DF được kết hợp với các kỹ thuật nhầm mục tiêu vi mô chính trị. Một số tình huống mà DF có thể ảnh hưởng đến các cuộc bầu cử có thể là:

- Một DF mạo danh người đưa tin để cung cấp thông tin bỏ phiếu sai, từ đó tạo ra sự lộn xộn trong ngày bầu cử.
- Các tác nhân chính trị ác ý có thể sử dụng DF để giả mạo bằng chứng nhằm thúc đẩy phản đối của họ các bộ phận' cáo buộc sai và tưởng thuật giả.
- Việc DF bắt chước một ứng viên và cho họ thấy những từ nhất định có thể ảnh hưởng nghiêm trọng đến danh tiếng của ứng viên.
- DF có thể hữu ích trong việc tạo nội dung hư cấu mới gây tranh cãi hoặc tuyên bố thù hận để kích động chia rẽ chính trị hoặc thậm chí bạo lực.
- Một video DF chiếu các ứng viên trong các tình huống thử thách hoặc với mọi người, điều này có thể tạo ra sự ngỡ vực.
- Ngược lại, các ứng cử viên hoặc các chủ thể chính trị/các bên liên quan khác nhau có thể nghi ngờ sự thật thông tin cuối cùng có hại cho danh tiếng của họ bằng cách gọi đó là DF.
- Các tác nhân chính trị có thể sử dụng DF như một mồi đe dọa giả định để đưa ra những tuyên bố không có cơ sở và gây nhầm lẫn cho cử tri-ví dụ: sự cố bầu cử quốc hội Gruzia năm 2020.

Trong một nghiên cứu ở Singapore, mặc dù 54% số người được hỏi ý thức được nhận thức về DF, nhưng một phần ba số người được hỏi vẫn chia sẻ nội dung của những DF đó trên mạng xã hội [32]. Nếu có một chương trình khuyến mãi quy mô lớn cho các DF như vậy, nó có thể chứng minh là thành công trong việc gây ảnh hưởng đến mọi người. Một cử tri bị cung cấp thông tin sai/không được cung cấp thông tin không thể bỏ phiếu một cách sáng suốt và do đó gián tiếp tước bỏ quyền bầu cử của họ, theo nghĩa thực tế, dựa trên sự hiểu biết chính xác về quan điểm của các chính trị gia [29]. Ngoài ra, trong các tình huống phân cực chính trị, mọi người có xu hướng tin tưởng nhiều hơn vào những "thông tin" tán thành quan điểm của họ. Do đó, cơ hội trờ nên dễ bị ảnh hưởng bởi thông tin sai lệch tăng lên đáng kể. DF có thể ảnh hưởng đến kết quả của bất kỳ cuộc bầu cử nào, đặc biệt nếu thời điểm được chọn để phân phối các phương tiện bị thao túng đó đủ để lưu hành nhưng còn ít thời gian hơn để nạn nhân vạch trần nó một cách hiệu quả.

8.10 Đe dọa đối với doanh nghiệp

DF gây ra mối đe dọa đáng kể cho các doanh nghiệp trong thời đại ngày nay khi các tương tác ảo và phương tiện kỹ thuật số là hình thức giao tiếp tiêu chuẩn được hầu hết các tổ chức áp dụng.

Công nghệ DF không còn được sử dụng cho nội dung liên quan đến khiêu dâm nữa, nhưng giờ đây nó đã

đi chuyền một cách ác ý đến các tổ chức mục tiêu bằng cách tiết lộ thông tin sai lệch và thông tin sai lệch. Nhiều ví dụ có sẵn công khai mô tả tội phạm sử dụng DF hình ảnh và âm thanh để thực hiện hành vi lừa đảo hoặc phạm tội, bao gồm tống tiền, đánh cắp danh tính và kỹ thuật xã hội. Người ta có thể tưởng tượng tác động đối với danh tiếng của tổ chức hoặc thiệt hại không thể khắc phục đối với lòng tin của các cổ đông khi người ta thấy một DF cho thấy Giám đốc điều hành của một công ty thú nhận gian lận tài chính hoặc nhận hối lộ. Một số tình huống có khả năng gây hại cho doanh nghiệp có thể là:

- Lừa đảo lừa đảo: DF có thể được sử dụng để mạo danh đồng nghiệp hoặc khách hàng một cách tổng hợp để tiết lộ thông tin hoặc chi tiết nhạy cảm về bất kỳ dự án nào hoặc cấp quyền truy cập vào cơ sở dữ liệu của công ty.
- Lừa đảo giao dịch: Có nhiều trường hợp những kẻ tạo lừa đảo đã dụ được nhân viên của công ty và thuyết phục họ thông qua DF âm thanh hoặc video để thực hiện một số khoản thanh toán nhất định. Trong số các ví dụ khác nhau, một ví dụ như vậy là về một công ty năng lượng có trụ sở tại Vương quốc Anh thực hiện chuyển khoản 243.000 đô la sau khi bị DF lừa.
- Lừa đảo tống tiền: DF như một công cụ tống tiền để tống tiền một công ty bằng cách đe dọa phát hành video DF có nội dung vi phạm.

Trên đây chỉ là một số trường hợp được nêu có thể đe dọa nghiêm trọng đến uy tín của bất kỳ doanh nghiệp nào. Các tổ chức hiện đang phải đối phó với các cuộc tấn công kỹ thuật xã hội khác nhau và gánh nặng bổ sung của việc chống lại các DF có thể có tác động bất lợi đến ngân sách đối với các doanh nghiệp. Theo kết quả nghiên cứu của Forrester Research, khoản lỗ trị giá 250 triệu USD được dự đoán sẽ xảy ra vào cuối năm 2020 do công nghệ DF.

8.11 MỐI ĐE DỌA LÀM XÓI MỎN NIỀM TIN VÀO CÁC TỔ CHỨC

Các thế chế công cộng, chẳng hạn như thẩm phán, bồi thẩm đoàn, các quan chức được bầu, các nhà lập pháp, các quan chức được bổ nhiệm, v.v., tạo thành nền tảng của xã hội. Công chúng tin tưởng các tổ chức này và những người liên quan đến họ. Nếu bất kỳ tổ chức nào đóng một vai trò quan trọng trong bất kỳ tổ chức nào, thì đó là mục tiêu tiềm năng cho công nghệ DF. Một video giả mạo và lan truyền cho thấy một thẩm phán lạm dụng quyền hạn của mình hoặc một quan chức cảnh sát cấp cao liên quan đến tội phạm hoặc một thành viên quốc hội nói ngôn ngữ phân biệt chủng tộc là nguyên nhân đủ để làm xói mòn lòng tin của công chúng vào các tổ chức đó và đôi khi là nền tảng của bất kỳ xã hội.

Các tổ chức như vậy hoặc những người có liên quan đến họ đã phải chịu các cuộc tấn công uy tín và công nghệ DF càng khiến việc giảm bớt/xóa bỏ những tuyên bố như vậy trở nên khó khăn hơn.

Vào năm 2020, một nhà hoạt động chính trị bảo thủ Theodore Dickinson đã chia sẻ một video với dòng tweet của cô ấy yêu cầu: "Để đáp trả các cuộc tấn công vào Nhà thờ Hồi giáo ở New Zealand, những người Hồi giáo đã đốt phá một nhà thờ Thiên chúa giáo ở Pakistan. Tại sao điều này không được hiển thị trên @BBCNews?!" [33]. Tuy nhiên, video này là giả mạo và thuộc về một cuộc tấn công vào nhà thờ ở Ai Cập vào năm 2013. Thông tin sai lệch/thông tin sai lệch như vậy có thể kích động sự phân biệt chủng tộc hoặc sự bất an trong cộng đồng nói chung, đặc biệt nếu những người đáng tin cậy trong xã hội chia sẻ chúng. Các video giả mạo và khiêu khích dễ dàng tìm thấy khán giả ưu tiên, đặc biệt là ở những nơi đã có sẵn những câu chuyện rõ ràng về sự ngõ趣 vực đối với các chính sách hoặc các nhà lãnh đạo chính trị hoặc các quy định xã hội hoặc đức tin tôn giáo hoặc bất kỳ luật nào sắp ban hành.

Chúng ta đang sống trong một xã hội mà nền tảng vốn đã rất mong manh, và bất kỳ sự cố mất lòng tin nào như vậy đều có thể gây hậu quả nghiêm trọng.

8.12 BIỆN PHÁP ĐỐI PHÓ

Khả năng sử dụng công nghệ DF một cách bất chính để gây thiệt hại và gây hại cho xã hội, hệ thống bầu cử, cá nhân và tổ chức đã gióng lên hồi chuông cảnh báo để tìm kiếm các biện pháp đối phó hiệu quả để chống lại chúng. Các nghiên cứu về tác động của DF và các cách giảm thiểu rủi ro mà chúng gây ra đã xác định các lĩnh vực biện pháp đối phó quan trọng sau đây.

8.13 CÁC BIỆN PHÁP PHÁP LUẬT

Hiện tại, hầu hết các quốc gia chưa sẵn sàng về mặt pháp lý để đối phó với DF. Không có luật hoặc chế độ trách nhiệm dân sự ở hầu hết các quốc gia chống lại việc tạo ra hoặc phân phối DF. Các quốc gia cần chấp nhận DF là một mối đe dọa tiềm tàng và đưa ra luật truy tố hình sự những người vi phạm các phiên bản sự thật đã được đặt ra. Việc cấm DF đơn thuần là điều không mong muốn vì việc thao túng kỹ thuật số vốn dĩ cũng đã được hưởng lợi. Điều cần thiết là xây dựng một luật cấm các ứng dụng phá hoại của công nghệ DF và đồng thời loại trừ những ứng dụng có lợi. Chỉ sự tồn tại của luật cấm DF cũng đủ để tạo ra một cái bóng đáng kể và đóng vai trò ngăn cản việc sử dụng nó.

Nhận thức theo hướng này đã bắt đầu và các nhà lập pháp ở nhiều quốc gia đang nỗ lực xây dựng luật mới cho thông tin sai lệch. Các nhà lập pháp của New York đang nghiên cứu dự luật cấm sử dụng cụ thể "bản sao kỹ thuật số" của một người [34]. Luật pháp Hoa Kỳ được cho là xây dựng một dự luật thiết lập các hình phạt hình sự đối với bất kỳ ai bị kết tội sử dụng/sản xuất DF và yêu cầu các nhà sản xuất DF tuân thủ các kỹ thuật xác minh cụ thể như hình mờ kỹ thuật số để đảm bảo tính xác thực của phương tiện [35]. Singapore gần đây đã thông qua luật trao quyền cho chính phủ ra lệnh cho các nền tảng truyền thông xã hội xóa bất kỳ nội dung nào mà chính phủ cho là sai [36].

Tuy nhiên, các luật như vậy có hậu quả của chúng. Chúng cần nghiên cứu kỹ lưỡng về khả năng lạm dụng có thể xảy ra vì rất có thể, nếu bị lạm dụng, chúng cũng có thể được sử dụng để ngăn chặn tự do ngôn luận báo chí dưới vỏ bọc giải quyết thông tin sai lệch. Chính phủ phải khuyến khích các nền tảng truyền thông xã hội cảnh giác hơn, chú ý hơn đến nội dung được đăng trên nền tảng của họ và thông qua luật/quy định. Hơn nữa, cần phải xem xét việc chia sẻ hình ảnh tình dục không có sự đồng thuận để hạn chế nội dung khiêu dâm DF.

Luật nên được thực hiện để hình sự hóa việc tạo hoặc chia sẻ DF với mục đích xấu. Tuy nhiên, các biện pháp pháp lý thường được áp dụng sau thực tế và do đó sẽ có tác động hạn chế trong việc giải quyết quy mô thiệt hại có thể xảy ra.

8.14 BIỆN PHÁP CÔNG NGHỆ

Có các giải pháp kỹ thuật cụ thể như các biện pháp đối phó khả thi đối với DF. DF là sản phẩm của AI và chỉ có công nghệ mới có thể chống lại các mối đe dọa do chúng gây ra. Trong khi công nghệ là tác nhân khởi xướng vấn đề, thì công nghệ cũng đưa ra một giải pháp tiềm năng để chống lại mối đe dọa ngày càng tăng của chúng. Các biện pháp đối phó kỹ thuật có thể được phân loại thành:

Phát hiện: Một công cụ ngăn chặn như vậy sử dụng các kỹ thuật phát hiện đa phương thức để tìm bất kỳ hành vi giả mạo nào trong phương tiện đích. Sau đây là một số giải pháp kỹ thuật có sẵn trên thị trường:

- Công cụ xác thực video: Được Microsoft ra mắt vào năm 2020, công cụ này phát hiện ranh giới gam màu hỗn hợp của DF và các phần tử thang độ xám được đánh giá thấp, đồng thời cung cấp điểm tin cậy của thao tác.
- Tín hiệu sinh học: Một công cụ do các nhà nghiên cứu từ Đại học Binghamton và Intel phát triển nhằm tìm kiếm các tín hiệu nhiễu tự nhiên và tổng quát duy nhất do các video mô hình DF để lại và có thể đạt được độ chính xác khoảng 97,29% để phát hiện video giả mạo.
- Không khớp âm vị-vị: Các nhà nghiên cứu từ Đại học Stanford và Đại học California đã phát triển một công cụ khác khớp các âm vị, động lực học của hình dạng miệng với âm vị, lời nói để phát hiện các thao tác thậm chí nhỏ về mặt không gian và cục bộ theo thời gian trong các video DF.
- Mô hình tích chập lặp lại (RCM): Kỹ thuật này sử dụng thông tin tạm thời từ các luồng hình ảnh trên các miền và phát hiện thao tác khuôn mặt trong video. Trong các luồng video, nó có thể phát hiện các khuôn mặt được giả mạo DF, Face2Face và FaceSwap.

Các kỹ thuật đã nêu ở trên có gắng phát hiện những điểm không nhất quán nhất định do DF tạo ra mà con người khó nhận thấy khi sử dụng ML và AI. Do đó, nếu các công cụ dễ tiếp cận tràn ngập thị trường để tạo video giả, thì công nghệ cũng cung cấp các công cụ dễ tiếp cận tương tự để phát hiện chúng. Nó luôn luôn chỉ đơn thuần là một cuộc giằng co giữa cả những “diễn viên” công nghệ xấu và tốt.

Xác thực phương tiện: Đây là một trong những công cụ có thể xác thực nguồn gốc và nội dung của người tạo phương tiện. Việc xác thực này có thể đạt được bằng cách sử dụng nhật ký chuỗi lưu ký thủy văn hoặc các phương tiện sẵn có khác. Điều này sẽ cho phép người dùng xem liệu phương tiện có bị giả mạo hoặc thao túng hay không và do đó làm chậm quá trình tạo video giả. Nhiều công cụ khác nhau như FotoForensics, Jeffrey's Exif Viewer, TinEye, v.v. có sẵn trên thị trường để hỗ trợ pháp y cho bất kỳ phương tiện nào.

Xuất xứ phương tiện: Việc thêm thông tin xuất xứ và đính kèm thông tin đó vào phương tiện giúp nội dung đáng tin cậy dễ xác định hơn. Thông tin xuất xứ bao gồm thông tin cơ bản trên phương tiện truyền thông bắt đầu từ nguồn gốc của phương tiện truyền thông đến các trang web xuất bản khác. Bất kỳ hình ảnh giả nào cũng có thể dễ dàng bị phát hiện bằng cách sử dụng tìm kiếm hình ảnh ngược trên internet. Nếu DF sử dụng một phương án khác ở đâu đó trên internet để tạo ra nó, thì hình ảnh gốc sẽ xuất hiện trong tìm kiếm. Có thể liệt kê một số giải pháp truyền thông xuất xứ trên internet hiện nay như sau:

- ID Nội dung YouTube: YouTube cung cấp ID Nội dung cho chủ sở hữu bản quyền để xác định và quản lý nội dung của họ. ID này được YouTube theo dõi để tìm bắt ký kết quả trùng khớp nào với cơ sở dữ liệu đã tồn tại để kiểm tra mọi hành vi vi phạm bản quyền.
- Adobe Content Authenticity Initiative: Adobe cung cấp nguồn gốc bằng cách cung cấp cho người tạo tùy chọn xác nhận quyền tác giả và ủy quyền cho người tiêu dùng đánh giá độ tin cậy của nội dung.
- Microsoft Aether Media Provenance (AMP): Microsoft AMP cho phép người dùng tạo các bảng kê khai đã ký cho phương tiện đã tạo. Các tệp kê khai này cũng có thể được đăng ký và ký bởi số cái chuỗi hành trình sản phẩm như chuỗi khối. Nếu AMP được yêu cầu

manifest được định vị thành công, nó được thông báo tới người tiêu dùng thông qua các phần tử trực quan trong trình duyệt.

- FuJo Provenance: Một dự án của EU có tên PROVENANCE đã được khởi xướng vào năm 2018 để phát triển một giải pháp xác minh nội dung kỹ thuật số không cần trung gian. Lớp xác minh PROVENANCE nhằm mục đích sử dụng các công cụ nâng cao như nâng cao ngữ nghĩa, pháp y hình ảnh và phân tích theo tầng để phát hiện bất kỳ thay đổi nào đối với nội dung phương tiện.

Giám sát: Hiệu quả của tất cả các biện pháp đối phó nêu trên sẽ được nâng cao nếu nó được kết hợp với nhận thức của xã hội. DF là một vấn đề xã hội hơn là vấn đề công nghệ vì chúng phổ biến hơn ở những nơi phân cực. Các tổ chức cần điều chỉnh lại các chính sách bảo mật của họ và thêm nhiều điểm kiểm tra vào các tình huống có thể xảy ra khi DF có thể gây ra sự cố. Tiến hành các bài tập quản lý khủng hoảng có thể hạn chế các nỗ lực DF thành công. Những cá nhân có nguy cơ cao nhất, chẳng hạn như người nổi tiếng hoặc giám đốc điều hành cấp cao của các công ty lớn, cần tuân theo các biện pháp cụ thể để đối phó với những nguy cơ do cuộc tấn công DF gây ra. Giám sát web là biện pháp tốt nhất để xác định và chống lại sự lây lan của DF ở giai đoạn sớm nhất.

Nâng cao kiến thức truyền thông: Kiến thức truyền thông là một công cụ hiệu quả khác để chống lại thông tin sai lệch do DF gây ra. Cải thiện kiến thức truyền thông để thu hút công chúng sáng suốt là một trong những tiền đề để chống lại những thách thức do DF đặt ra. Là người tiêu dùng phương tiện truyền thông, mọi người cần phát triển khả năng giải mã, hiểu và phân biệt giữa giả và thực. Mọi người cần phải tạm dừng trước khi mù quáng tin vào những gì họ đang nhìn thấy. Và điều này có thể đạt được nhờ nhận thức về phương tiện truyền thông hiệu quả của người tiêu dùng. Các DF được sản xuất với giá rẻ để lại một số đồ tạo tác nhất định có thể giúp phát hiện ra sự thao túng khi quan sát kỹ lưỡng. Có những đặc điểm nhận biết cụ thể của một video DF rẻ tiền mà nếu được quan sát kỹ lưỡng, có thể giúp xác định chúng. Ví dụ: một video bị chỉnh sửa sẽ có làn da trông quá mịn hoặc quá nhăn, bóng gần mắt và lông mày thường không đồng nhất, độ tuổi của da, tóc và mắt không khớp nhau, cử động mắt không tự nhiên, nét mặt, cơ thể khác lạ hình dạng hoặc tóc, màu da bất thường, vị trí đầu không phù hợp, v.v.; đây là một số gợi ý, nếu được công khai, có thể hạn chế việc phân phối lan truyền các video giả mạo.

Các biện pháp đã nêu ở trên là một số cách mà người ta có thể ngăn chặn nếu không chống lại hoàn toàn tác động do DF gây ra; tuy nhiên, những ngờ vực vẫn tồn tại về hiệu quả của các can thiệp khác nhau. Vì DF được sinh ra thông qua đào tạo đối thủ, nên có khả năng cao là khả năng tránh các phương pháp phát hiện dựa trên AI của những DF này có thể cải thiện khi chúng quen thuộc hơn với các kỹ thuật phát hiện như vậy. Dường như có một quy trình “mèo vờn chuột” có lợi cho chuột, đặc biệt là về lâu dài. Hầu hết các biện pháp đối phó phát hiện DF đều tập trung vào việc đạt được các giải pháp ngắn hạn, với kỳ vọng rằng các kỹ thuật giám sát hoặc lập pháp có thể tỏ ra có lợi hơn như một giải pháp lâu dài đối với mối đe dọa do DF gây ra. Chương này nêu bật một số mối nguy hiểm và hậu quả có thể xảy ra của công nghệ DF. Tuy nhiên, các nhà hoạch định chính sách, nhà nghiên cứu và người tiêu dùng rất cần phải tìm kiếm những cải tiến công nghệ và phi kỹ thuật mới có thể đóng vai trò ngăn chặn mức độ đe dọa ngày càng tăng của thông tin sai lệch hình ảnh do DF tạo ra. khám phá liên tục

việc thử nghiệm các biện pháp ngăn chặn mới là cần thiết với tốc độ phù hợp với tốc độ mà công nghệ DF đang phát triển. Hơn nữa, đã đến lúc người tiêu dùng không nên tin hoàn toàn vào những gì họ nhìn thấy hoặc nghe thấy mà thay vào đó hãy vận dụng trí tuệ của mình để xác nhận độ tin cậy của nó.

8.15 TÓM TẮT

Giả sử mọi video đều có khả năng là giả mạo. Trong trường hợp đó, nó mang đến cho mọi người cơ hội để thách thức tính toàn vẹn của cảnh quay chân thực, dẫn đến sự sụp đổ niềm tin của mọi người. Người ta nói rằng công nghệ DF đang dần đưa mọi người đến một "ngày tận thế thông tin", nơi rất khó để phân biệt giữa thật và giả [37]. Nó có thể dẫn đến một tình huống mà thực tế và niềm tin phổ biến sụp đổ và sự hỗn loạn nguy trị. Con người nói chung [38,39,40] bị thu hút nhiều hơn bởi những thông tin tiêu cực và mới lạ. Rất cần có các hành động hợp tác trong các quy định pháp lý, chính sách nền tảng, can thiệp công nghệ và hiểu biết về phương tiện truyền thông để có thể đưa ra các biện pháp đối phó nhằm giảm thiểu mối đe dọa của các DF độc hại.

NGƯỜI GIỚI THIỆU

- [1] Sốt Rét Phái Chết. (2019, ngày 9 tháng 4). David Beckham nói được 9 thứ tiếng để phát động kiến nghị Malaria Must Die Voice. YouTube. Truy cập ngày 9 tháng 6 năm 2021, từ www.youtube.com/watch?v=QiiSAvKJlHo
- [2] BuzzFeed/YouTube. (2018, ngày 17 tháng 4). Bạn sẽ không tin những gì Obama nói trong Video này!
- [3] YouTube. (2018, ngày 17 tháng 4). Bạn sẽ không tin những gì Obama nói trong video này đâu! www.youtube.com/watch?v=cQ54GDm1eL0 và www.youtube.com/watch?v=cQ54GD (video không có sẵn hiện nay).
- [4] Twitter. (2018, ngày 17 tháng 4). Bạn sẽ không tin những gì Obama nói trong video này đâu! https://twitter.com/BuzzFeed/status/9862579917_99222272
- [5] Facebook. (2018, ngày 17 tháng 4). Bạn sẽ không tin những gì Obama nói trong video này đâu! www.facebook.com/watch/?v=10157675_129905329
- [6] Lytvynenko, J. (2018). "Một đảng chính trị của Bi đang lưu hành một video Trump DF," Tin tức BuzzFeed, ngày 20 tháng 5 năm 2018. [Trực tuyến]. Có sẵn: www.buzzfeednews.com/article/janelytvynenko/a-bi-chinh-dang-vua-xuat-ban-a-df-video
- [7] Graber, DA (1990). Nhìn là ghi nhớ: Hình ảnh đóng góp như thế nào vào việc học từ tin tức truyền hình. Tạp chí Truyền thông, 40(3), 134-156.
- [8] Grabe, TÔI, & Bucy, EP (2009). Hình ảnh Bite Chính trị: Tin tức và Khung trực quan của Bầu cử. Nhà xuất bản Đại học Oxford.
- [9] Trước đó, M. (2013). Kiến thức chính trị trực quan: Con đường khác dẫn đến năng lực? Tạp chí Chính trị, 76(1), 41-57.
- [10] Schwarz, N., Sanna, LJ, Skurnik, I., & Yoon, C. (2007). Kinh nghiệm siêu nhận thức và sự phức tạp của việc thiết lập mọi người ngay thẳng: Hệ quả đối với các chiến dịch thông tin đại chúng và gõ lỗi. Những tiên bộ trong Tâm lý xã hội thực nghiệm, 39, 127-161.
- [11] Rini, R. (2019). DF và backstop epistemia. <https://philpapers.org/rec/RINDAT>
- [12] David Easley & Jon Kleinberg, Mạng lưới, đám đông và thị trường: Lý luận về một thế giới kết nối cao. (2010). (Khám phá các xu hướng nhận thức trong thị trường thông tin); Cass Sunstein, Republic.Com 2.0 (2007).

- [13] Anderson, M., & Hitlin, P., Hashtag #BlackLivesMatter nỗi lên: Hoạt động xã hội trên Twitter, PEW RES. CTR. (ngày 15 tháng 8 năm 2016), www.pewinternet.org/2016/08/15/the-hashtag-blacklivesmatter-emerges-xã-hội-hoạt-dong-trên-twitter
- [14] Bromwich, JE (ngày 7 tháng 3 năm 2018). Làm thế nào mà sinh viên Parkland trở nên giỏi về mạng xã hội, The New York Times, www.nytimes.com/2018/03/07/us/parkland-students-social-media.html [<https://perma.cc/7AW9-4HR2>
- [15] Vosoughi, S. và cộng sự, Sự lan truyền tin tức đúng và sai trực tuyến, 359 KHOA HỌC 1146, (2018), <http://science.sciencemag.org/content/359/6380/1146/tab-pdf> [<https://perma.cc/5U5DUHPZ>
- [16] www.nytimes.com/2018/11/06/technology/myanmar-facebook.html
- [17] www.wilmerhale.com/en/insights/client-alerts/20200316-fbi-warns-companies-of-môis-de-doga-gan-nhu-chắc-chắn-tu-DF
- [18] <https://schiff.house.gov/imo/media/doc/2018-09%20ODNI%20Deep%20Fakes%20thu.pdf>
- [19] Hạ viện Hoa Kỳ, Phiên điều trần về những thách thức an ninh quốc gia của AI, Phương tiện bị thao túng và DF, 13/VI/2019.
- [20] Đánh giá Mối đe dọa Toàn cầu của Cộng đồng Tình báo Hoa Kỳ, Statement for the Record, 29/I/2019, tr. 7.
- [21] Chesney, R., & Citron, DK (2019). DF và cuộc chiến thông tin mới. Foreign Affairs, Tháng Giêng/ Tháng Hai, 147-155.
- [22] Rana Ayyub, Ở Ấn Độ, các nhà báo phải đối mặt với những lời đe dọa làm dấy và hiếp dâm, The New York Times (22 tháng 5 năm 2018), www.nytimes.com/2018/05/22/opinion/india-journalists-slut-shamingrape.html; [<https://perma.cc/A7WR-PF6L>]; 'Tôi không thể nói chuyện hoặc ngủ trong ba ngày': Thủ thách khủng khiếp trên mạng xã hội của nhà báo Rana Ayyub đối với Tweet giả, Daily O (26 tháng 4 năm 2018), www.dailyo.in/variety/rana-ayyub-trolling-fake-tweet-social-media-harassmentinindia/story/1/23733.html
- [23] www.bbc.com/news/world-asia-india-44435127
- [24] Toews, R. (2020). DFs sẽ tàn phá xã hội. Chúng tôi không chuẩn bị. Forbes. www.forbes.com/sites/robtoews/2020/05/25/DFs-are-going-to-wreak-havoc-on-society-weare-not-prepared/
- [25] Pfefferkorn, R. (2020). " 'DFs' trong phòng xử án," Tạp chí Luật Đại học Boston, 29(2), 245-276.
- [26] Ryan, P. 'DF' Bằng chứng âm thanh được sử dụng tại Tòa án Vương quốc Anh để làm mất uy tín của Dubai Dad, THE NATIONAL (ngày 8 tháng 2 năm 2020), www.thenational.ae/uae/courts/DF-audio-evidence-used-in-uk-court-against-dad-1.975764
- [27] Owen, LH Chúng tôi làm gì với video "Shallowfake" của Nancy Pelosi và những người khác thích nó?, NIEMAN LAB (ngày 31 tháng 5 năm 2019), www.niemanlab.org/2019/05/what-do-we-do-about-the-shallowfake-nancy-pelosi-video-and-others-like-it/
- [28] Kelly, M. (2019). "Các video Nancy Pelosi bị bóp méo cho thấy các nền tảng chưa sẵn sàng chống lại các mánh khóc chiến dịch bẩn thỉu." [Trực tuyến]. Có sẵn: www.theverge.com/2019/5/24/18637771/nancy-pelosi-congress-DF-video-facebook-twitter-youtube
- [29] Màu xanh lá cây, R. (2019). Bài phát biểu chiến dịch giả mạo. Tạp chí Luật Hastings, 70(6), 1445-1490.
- [30] Vaccari, C., & Chadwick, A. (2020). DF và thông tin sai lệch: Khám phá tác động của video chính trị tổng hợp đối với sự lừa dối, sự không chắc chắn và sự tin tưởng vào tin tức. Mạng xã hội + Xã hội, 1-13. <https://doi.org/10.1177/2056305120903408>
- [31] Zimmermann, F., & Kohring, M. (2020). Sự nghi ngờ, tin xuyên tạc và lựa chọn bỏ phiếu: Một cuộc khảo sát của ban hội thẩm về nguồn gốc và hậu quả của việc tin vào thông tin sai lệch trong Cuộc bầu cử Nghị viện Đức năm 2017. Truyền thông Chính trị, 37, 215-237. doi:10.1080/10584609.2019.1686095

- [32] <https://sea.mashable.com/tech/13337/think-you-can-spot-a-DF-survey-proves-that-even-the-best-get-fooled>
- [33] [www.theguardian.com/technology/ng-interactive/2019/jun/22/the-rise-of-the-DF_and-the-threat-to-democracy](http://www.theguardian.com/technology/ng-interactive/2019/jun/22/the-rise-of-the-DF-and-the-threat-to-democracy); www.nysenate.gov/legislation/bills/2017/a8155
- [34] www.bu.edu/bulawreview/files/2021/04/LANGA.pdf
- [35] www.congress.gov/bill/116th-congress/house-bill/3230
- [36] www.bbc.com/news/world-asia-48196985
- [37] Rothman, J. (2018). Thời đại AI, thấy còn tin? Người New York. www.nguoiNewYork.com/magazine/2018/11/12/in-the-age-of-ai-is-seeing-van-tin
- [38] Mahbub, Md. K., Biswas, M., Gaur, L., Alenezi, F., & Santosh, K. (2022). Các tính năng sâu để phát hiện các bất thường về phổi trên phim X-quang phổi do bệnh truyền nhiễm X: Covid-19, viêm phổi và bệnh lao. Khoa học Thông tin, 592, 389-401. <https://doi.org/10.1016/J.IJCS.2022.01.062>
- [39] Sharma, S., Singh, G., Gaur, L., & Sharma, R. (2022). Khoảng cách tâm lý và tôn giáo có ảnh hưởng đến hành vi gian lận của khách hàng không? Tạp chí Nghiên cứu Người tiêu dùng Quốc tế. 10.1111/ijcs.12773
- [40] Zaman, N., & Gaur, L. (2022). Phương pháp tiếp cận và ứng dụng của Deep Learning trong Chăm sóc y tế áo. IGI. doi:10.4018/978-1-7998-8929-8.ch002



Taylor & Francis
Taylor & Francis Group
<http://taylorandfrancis.com>

9 DeepFakes, Phương tiện và Tác động xã hội

Shubha Mishra, Piyush Kumar Shukla, và
Ratish Agrawal

NỘI DUNG

9.1 Giới thiệu115
9.2 Thiếu thông tin về thử nghiệm	115
9.3 Một ít kiến thức từ điều tra giao dịch kép	116
9.4 Tác động xã hội	117
9.5 Tóm tắt	118
Người giới thiệu.....	119

9.1 GIỚI THIỆU

Các tính toán xây dựng DeepFakes (DF) đơn giản hơn để xây dựng so với nhận dạng. Dựa trên bản chất đặc biệt của Hệ thống đổi khảng sáng tạo được sử dụng đồng ý với Goodfellow, các mô hình này được xây dựng bằng cách thiết lập “những kẻ giả mạo” chống lại “cảnh sát” và các mô hình hiệu quả theo định nghĩa cho đến nay, có vẻ như hàng giả có thể đánh bại các chiến lược vị trí. Không còn nghi ngờ gì nữa, vì các DF đã chuyển từ các cơ sở nghiên cứu khoa học máy tính thông thường sang các giai đoạn chương trình giá rẻ trên toàn thế giới, nên các nhà phân tích cũng đang tập trung vào các tính toán cần thận có thể xác định hướng đi sai (xem Tolosana và cộng sự [1] để khảo sát sau). Nhưng Seitz không chắc chắn về kỹ thuật này và so sánh việc đi sai hướng và vị trí quanh co với một cuộc chạy đua vũ trang [2,3], với những tính toán mà kẻ lừa đảo có lợi thế sớm so với những tính toán xác định được.

Điều mở rộng tầm mắt của thời điểm này là vô số câu hỏi về mặt xã hội và tinh thần mà những người DF này nêu ra: việc trình bày trước những người DF có làm suy yếu niềm tin trong giới truyền thông không? Làm thế nào DF có thể được sử dụng trong bối cảnh trực giác xã hội? Có phương pháp nào để lật tẩy hoặc chống lại DF không? Tuy nhiên, cho đến nay, đã có một nhóm khiêm tốn các nhà nghiên cứu xã hội đã kiểm tra tác động xã hội của sự đổi mới. Đã đến lúc hiểu những tác động tiềm ẩn của DF đối với các cá nhân và cách áp dụng những suy đoán về tinh thần và phương tiện truyền thông.

9.2 NGẮN HỎI THỦ NGHIỆM VỀ

Có vẻ như tại thời điểm viết bài này, một số điều đã kiểm tra tác động xã hội của DF [4], bắt chấp sự nổi tiếng của việc đổi đầu với các giai đoạn hoán đổi (ví dụ: ứng dụng Zao [5]). Vẫn đề đặc biệt này hướng đến kỷ nguyên đầu tiên của cuộc điều tra DF phân tích các gợi ý về tinh thần, xã hội và cách tiếp cận của một thế giới trong đó các cá nhân

có thể dễ dàng cung cấp và phổ biến các bản ghi âm của các sự kiện chưa bao giờ thực sự xảy ra, nhưng điều này không rõ ràng so với các bản ghi âm chính hãng.

Mặc dù đã có một số cân nhắc về việc bảo vệ bộ nhớ sai và tác động xã hội từ các ảnh tĩnh bị thay đổi (ví dụ: Garry và Swim [3]), các hình thái tinh thần và kết quả của việc xem video được sửa đổi thông tin chi tiết giả (AI) nói chung vẫn tồn tại. Không học. Thật đáng ngạc nhiên, điểm khởi đầu hàng đầu để hiểu tác động của DF là thực tế ảo (VR) nhập vai. Trong VR, người ta có thể xây dựng các mô hình "doppelgangers", ba chiều (3D) của một cá nhân nhất định, dựa trên phép chụp ảnh và các chiến lược khác để tạo cấu trúc 3D từ sự sắp xếp các hình ảnh hai chiều (2D). Sau khi doppelganger được xây dựng, điều cần thiết là áp dụng các hình ảnh động có sẵn lên các mô hình 3D. Sau đó, các cá nhân xuất hiện trong cảnh VR DF, trong màn hình gần đầu hoặc được hiển thị dưới dạng hoạt động video 2D diễn hình. VR DF có tác động mạnh mẽ.

So với việc quan sát cảnh của một cá nhân khác, việc quan sát dop pelganger yêu cầu của bạn gây ra mã hóa những ký ức không có thật trong đó các thành viên chấp nhận rằng họ đã thực hiện hành động DF, hành vi tập thể dục nhiều hơn sau khi quan sát thấy kết quả sức khỏe tích cực và xu hướng thương hiệu đối với các vật phẩm được sử dụng bởi bản thân ảo trong DF [6]. Rất có thể các thành phần và kết quả tinh thần không được sử dụng đang diễn ra khi một video DF về mặt nhận thức là mơ hồ so với một video thực tế.

9.3 MỘT SỐ BIẾT KIẾM THỨC TỪ ĐIỀU TRA GIAO DỊCH NHÂN ĐÔI

Hành động hai mặt tại trung tâm của DF bao gồm cô ý, cố ý và cố ý lừa dối một cá nhân khác [7]. Việc viết sai vị trí cho thấy rằng các cá nhân không đặc biệt giỏi trong việc xác định sự trùng lặp khi đánh giá các thông điệp và có thể nhanh chóng nhận được kết luận sai lầm. Các phân tích tổng hợp về địa điểm giao dịch kép xem xét khuyến nghị rằng các cá nhân nên hành động như thế nào hời có cơ hội khi đánh giá một tuyên bố là xác thực hoặc gây hiểu lầm. Một cách bắt buộc, mức độ chính xác này không bị ảnh hưởng bởi phương tiện mà thông điệp được truyền đi [8]. Có nhiều ý kiến cho rằng việc phát hiện đánh lạc hướng gần như giống nhau cho dù tin nhắn được truyền qua nội dung (ví dụ: bản ghi của tòa án, nhật ký trò chuyện trên Web), bản ghi âm (ví dụ: thư thoại, chương trình radio) hoặc video (ví dụ: video kiểm tra chéo).

Mặc dù điều này có vẻ gây sốc khi xem xét chi tiết phong phú hơn có sẵn trong video, nhưng độ chính xác có xu hướng khó xảy ra bất kể phương tiện vì không có tín hiệu chắc chắn nào cho việc con người xử lý hai mặt (tức là không có mũi của Pinocchio). Chúng ta có xu hướng tin những gì người khác nói. Nhưng phần lớn, quan trọng hơn của cuộc điều tra về giao dịch kép dựa trên video đã kiểm tra nội dung lời nói của một diễn ngôn, chẳng hạn, một cá nhân nói dối, trái ngược với sự phát triển và hình dạng cơ thể của một người. Một trong những khía cạnh thú vị nhất của vấn đề không phổ biến này là điều tra sự trùng lặp không chỉ dựa trên những lời nói dối được nói ra bằng lời nói mà còn cần trả toàn bộ quá trình sản xuất các hành vi bằng lời nói và phi ngôn ngữ.

Mặc dù tỷ lệ khám phá có khả năng cao hơn so với các phương tiện khác, nhưng tác động của việc xử lý kép của DF có thể nổi bật hơn so với cảm giác trùng lặp bằng lời nói về ưu thế của giao tiếp bằng hình ảnh đối với nhận thức của con người. DFs không phải là nó thay đổi miệng

chất, nhưng chúng cũng thay đổi các thuộc tính quang học của cách thông điệp được truyền đi, cho dù điều này bao gồm việc phát triển miệng của một người nói điều gì đó mà họ không nói, hay hành vi của một cá nhân làm điều gì đó mà họ không làm. Sự thống trị của các tín hiệu hình ảnh trong nhận dạng của con người được xây dựng tốt. Ví dụ, trong nhiều trường hợp, mọi người phụ thuộc nhiều vào dữ liệu trực quan hơn là các dạng dữ liệu xúc giác khác, một điều kỳ diệu được gọi là tác động thống trị thi giác của Colavita.

Trong thế giới quan cơ bản của Colavita, các thành viên phải đưa ra phản ứng nhanh chóng đối với sự sắp xếp bất thường của các phương tiện liên quan đến âm thanh, hình ảnh hoặc các phương tiện khác nhau. Các thành viên được dạy tạo một phản ứng đối với mục tiêu liên quan đến âm thanh, một phản ứng khác đối với mục tiêu hình ảnh và tạo cả hai phản ứng tại bất kỳ thời điểm nào mà tar liên quan đến âm thanh và hình ảnh được hiển thị đồng thời. Các thành viên không gặp vấn đề gì khi phản ứng độc lập với các mục tiêu âm thanh và video, nhưng khi chúng được hiển thị cùng nhau, họ thường thất bại trong việc phản hồi các mục tiêu liên quan đến âm thanh. Nó giống như trường hợp tăng cường hình ảnh dập tắt tăng cường âm thanh. Người ta quan sát thấy rằng các cá nhân có nhiều khả năng xem xét các thông điệp bằng hình ảnh hơn là thông điệp bằng lời nói và việc đánh lừa dữ liệu hình ảnh có nhiều khả năng tạo ra sự nhận dạng sai hơn là đánh lừa nội dung bằng lời nói do "heuristic chủ nghĩa hiện thực", trong đó các cá nhân có nhiều khả năng tin vào các phương thức truyền thông khác nhau hơn lời nói vì bản chất bao gồm một sự tương đồng cao hơn với thế giới thực.

Video là ranh giới cuối cùng-là ranh giới mà khách hàng có thể quan sát và do đó không mong đợi bị làm giả [9]. Nhưng điều gì sẽ xảy ra khi chúng ta biết rằng một video có thể được "chỉnh sửa bằng photoshop" một cách dễ dàng như những bức ảnh? Chúng tôi có thể chấp nhận bất kỳ phương tiện truyền thông nào mà chúng tôi xem không? Người theo chủ nghĩa duy lý Wear Fallis ám chỉ điều này là rõ ràng nhận thức của DFs [5]. Sự tranh chấp của anh ấy truyền từ việc kiểm soát phương tiện trực quan sang mang dữ liệu, dữ liệu này cho biết lượng cờ được chuyển qua một tin nhắn. Do sự thống trị của khung hình ảnh, các bản ghi có tiềm năng mang dữ liệu cao-nghĩa là chúng ta có xu hướng chấp nhận những gì chúng ta thấy trong video và kết quả là các bản ghi đã trở thành "tiêu chuẩn vàng" của sự thật. Nhưng khi DF nhận lên và lưu ý rằng các bản ghi có thể bị làm giả lan rộng trong cộng đồng, thì tổng dữ liệu mà các bản ghi mang đến cho người xem sẽ giảm đi.

Thật vậy, nếu một video là xác thực và người xem sẽ có niềm tin thực sự, thì sự nghi ngờ do DF gây ra sẽ ngăn cản một cá nhân thực sự chấp nhận những gì họ đã xem [1,7]. Rủi ro về mặt nhận thức đối với Fallis là DF sẽ can thiệp vào khả năng thu thập thông tin về thế giới của chúng ta bằng cách quan sát các phương tiện truyền thông. Những gợi ý về sự hiểu biết chung của chúng ta về thế giới và vai trò mà tin tức cũng như các phương tiện truyền thông khác đóng vai trò trong việc xây dựng thế giới đó có thể thực sự bị hủy hoại.

9.4 TÁC ĐỘNG XÃ HỘI

Thật đáng ngạc nhiên, một trong số ít nghiên cứu quan sát về DF đưa ra một vài bằng chứng ban đầu chứng minh điều đáng lo ngại về giải thích triết học này. Khi cân nhắc xem xét tác động của DF đối với niềm tin vào tin tức [10], Vaccari và Chadwick nhận thấy rằng mặc dù mọi người không thể bị DF đánh lừa (ít nhất là với công nghệ mà họ đang sử dụng), phần giới thiệu về DF được mở rộng sự bất ổn của họ xung quanh các phương tiện truyền thông chung. Khẳng định những mong muốn khủng khiếp nhất, cảm giác bất ổn đó đã khiến các thành viên giảm bớt niềm tin vào tin tức, giống như lời kể của Fallis về mối nguy hiểm xuất phát từ tiền sử đã dự đoán.

DF, hơn nữa, có kết quả giữa các cá nhân. Khi VR cảm nhận để xuất hiện được mô tả, DF video có khả năng điều chỉnh ký ức của chúng ta và thực sự nhúng những ký ức tồi tệ. Chúng có thể thay đổi thái độ của một người đối với mục tiêu của DF. Một nghiên cứu sau đó đã phát hiện ra rằng bài thuyết trình đó cho một DF mô tả một nhân vật chính trị về cơ bản đã từ chối tâm trạng của những người tham gia đối với nhà lập pháp đó. Thực sự đáng lo ngại hơn, với khả năng nhúng mục tiêu nội dung vào các nhóm chính trị hoặc thông kê cụ thể của phương tiện truyền thông xã hội, nghiên cứu đã phát hiện ra rằng việc nhúng mục tiêu vi mô DF vào các nhóm có nhiều khả năng bị xúc phạm nhất (ví dụ: Cơ đốc nhân) đã làm tăng tác động này so với việc chia sẻ DF với công chúng bình thường.

Mặc dù những đe dọa này vạch ra một đại diện đáng thất vọng về một tương lai với sự đổi mới của DF, nhưng điều này đòi hỏi một người mua phương tiện không hoạt động ở mức độ vừa phải [11, 12]. Điều quan trọng là phải xem xét rằng mọi người đã thích nghi với các hình thức sai hướng mới trong nhiều thiên niên kỷ. Các cá nhân có xu hướng tin tưởng lẫn nhau cho đến khi họ có một vài lý do để trở nên nghi ngờ hoặc cảnh giác hơn, một trạng thái mà Levine ám chỉ là một sự mặc định về niềm tin. Chúng ta thoát khỏi niềm tin mặc định của mình khi chúng ta biết về những dữ liệu xung đột, một bên thứ ba cảnh báo chúng ta, hoặc chúng ta được dạy những thủ tục phức tạp gần như mới lạ. Ví dụ, thử xác ít thành công hơn nhiều so với khi nó bắt đầu được phát triển vì mọi người đều quan tâm đến nó.

Theo cách tương tự, các cá nhân có thể tạo ra sức mạnh cho các hình thức sai lệch mới như DF. Để minh họa, việc quảng cáo theo thói quen phụ thuộc vào dữ liệu thị giác đánh lừa (ví dụ: uống loại bia này, có những người bạn đồng hành tuyệt vời; hút điếu thuốc này, gặp gỡ bên ngoài tuyệt vời). Theo thời gian, người mua chú ý theo dõi và không bị lừa bằng cách công khai, một phần vì họ tạo ra một mô hình mong muốn quảng cáo. Không còn nghi ngờ nữa, chúng tôi đưa ra những mong muốn như thế này đối với hầu hết các phương tiện truyền thông mà chúng tôi ngẫu nhiên. Ví dụ, công nghệ DF hiện được sử dụng trong các bộ phim Hollywood. Ví dụ, mô tả của Công chúa Leia trong Chiến tranh giữa các vì sao VIII sau khi nghệ sĩ biểu diễn Carrie Fisher đã đái cái xô. Hầu hết mọi người đánh dấu DF là hư cấu vì họ xem một bộ phim chuyển động mang tính giải trí. Tuy nhiên, một vấn đề quan trọng là liệu bằng chứng trực quan có bắt đầu ghi nhớ của người xem rằng nghệ sĩ biểu diễn đã qua đời hay không, trong bất kỳ trường hợp nào họ biết rằng đó có thể là một bộ phim chuyển động?

Tuy nhiên, một tồn thương bắt buộc mà chúng tôi chưa xem xét là tai nạn vô cớ được mô tả trong DF là làm hoặc nói điều gì đó mà họ không làm. Một trong những hình thức ban đầu phổ biến nhất của DF là sự sửa đổi của sự tục tĩu, mô tả những người không có sự đồng thuận khóa chặt trong một hành vi tình dục không bao giờ xảy ra thường xuyên bằng cách đặt một người đối đầu với cơ thể của người khác. Với khả năng kiểm soát khung hình trực quan có tác dụng sửa đổi niềm tin của chúng ta như được miêu tả hiện nay và tác động mà những DF như vậy có thể gây ra đối với bản thân nhân vật, tác động đối với cuộc sống của nạn nhân có thể bị phá hủy. Mặc dù điều tra quan sát về ngày bị hạn chế, nhưng không khó để giả định cách DF có thể được sử dụng để tổng tiền, hành xác hoặc gây khó chịu cho thương vong.

9.5 TÓM TẮT

Trong vấn đề hiềm gãy này, chúng tôi khuyến khích các nhà phân tích suy ngẫm về các vấn đề xã hội liên quan đến đổi mới DF. Những suy nghĩ trong tập này đã thực hiện một công việc đáng kinh ngạc trong việc vạch ra câu hỏi về các câu hỏi, áp dụng giả thuyết cho điều kỳ diệu và tạo ra các công cụ không sử dụng để áp dụng cho các câu hỏi trong tương lai. Nhưng suy nghĩ này là chuẩn bị, và

chúng tôi khuyến khích các nhà nghiên cứu xây dựng dựa trên cân nhắc này vì DF sử dụng tiền thu được để phát triển.

Tuy nhiên, có một vùng hoang dã liên quan khác cần được xem xét. Hiện tại, khi chúng tôi kiểm tra DF, chúng tôi đang đề cập đến video đã quay. Nhưng ML đã phát triển đầy đủ để trao quyền cho các DF thời gian thực: các kênh do AI cung cấp. Các kênh này cho phép thay đổi hoặc tối ưu hóa nội dung video của hội nghị truyền hình trong thời gian thực, chẳng hạn như làm cho mắt của một người hiển thị mặc dù nó thực sự hướng vào máy ảnh, mặc dù nó được gởi ý ở một nơi khác trên màn hình. Để mở rộng sang việc giám sát [13] sự cân nhắc chung trong cài đặt video vô duyên, các kênh DF khác đang được tạo để tối ưu hóa cho các yếu tố giữa các cá nhân khác, chẳng hạn như sự nồng nhiệt hoặc sự mê hoặc giữa các cá nhân. Đối với trường hợp, Goodness et al. đã sử dụng một phương tiện thời gian thực để cải thiện tổng số nụ cười toe toét [14,15] trong các cặp đôi, chứng minh rằng những người đồng phạm trong điều kiện cười toe toét được nâng cấp cảm thấy rõ ràng hơn sau cuộc thảo luận và sử dụng nhiều từ tích cực hơn trong cuộc thảo luận của họ dựa trên nghiên cứu từ nguyên.

Điều cơ bản cần lưu ý là những tác động xuôi dòng này đã thực sự xảy ra, mặc dù các thành viên không chú ý đến kênh cười toe toét và gần như không bao giờ được nhận ra.

Các nhà phân tích tại cơ sở nghiên cứu MIT Media đang tạo ra “các mô hình phần được cá nhân hóa” bằng cách sử dụng công nghệ DF để sửa đổi luồng video thời gian thực nhằm cho phép người nói nhìn thấy sự thích ứng của chính họ vượt quá mong đợi khi thực hiện các nhiệm vụ nói chuyện theo một cách nhất định. Họ đang minh họa những tác động không phải đối với tính khí mà là trí tưởng tượng về nhiệm vụ. Việc sử dụng AI này để điều chỉnh phần tư giới thiệu của một người trong hội nghị truyền hình có thể là một khung giao tiếp do AI làm trung gian, ám chỉ đến “giao tiếp giữa các cá nhân trong đó một chuyên gia xuất sắc làm việc vì lợi ích của người giao tiếp bằng cách điều chỉnh, mở rộng hoặc tạo thông điệp cho đạt được mục tiêu giao tiếp.”

Mặc dù công nghệ DF có thể làm suy yếu niềm tin của chúng ta vào phương tiện truyền thông hoặc tác động sai lầm đến niềm tin của chúng ta về thế giới, nhưng nó có thể trở nên phổ biến và bình thường hơn khi các cá nhân sử dụng công nghệ DF để thúc đẩy hoạt động giao tiếp hàng ngày của họ. Khi bài diễn văn cụ thể làm rõ và các bài báo trong vấn đề hiếm gặp này nổi bật, có rất nhiều vấn đề cần thiết về tinh thần, xã hội và đạo đức đòi hỏi phải có những kiểm tra quan sát thận trọng và giàu trí tưởng tượng về tác động xã hội của những tiến bộ của DF.

NGƯỜI GIỚI THIỆU

- [1] Tolosana R, Vera-Rodriguez R, Fierrez J, et al. DFs và hơn thế nữa: Một cuộc khảo sát về thao túng khuôn mặt và phát hiện giả mạo. Hợp nhất thông tin 2020; 64:131-148.
- [2] Cáo J, & Bailenson JN. Tự lập mô hình ảo: Tác động của việc cung cấp và xác định gian tiếp đối với các hành vi tập thể dục. Tâm lý truyền thông 2009; 12:1-25.
- [3] Garry M, & Wade KA. Trên thực tế, một bức ảnh có giá trị dưới 45 từ: Những câu chuyện kể tạo ra nhiều ký ức sai lầm hơn những bức ảnh. Bản tin Tâm lý & Đánh giá 2005; 12:359-366.
- [4] Ahmed S. Ai vô tình chia sẻ DF? Phân tích vai trò của lợi ích chính trị, khả năng nhận thức và quy mô mạng lưới xã hội. Viễn thông và Tin học 2021; 57:101508.
- [5] Doffman Z. Ứng dụng DF ZAO của Trung Quốc lan truyền chóng mặt, quyền riêng tư của hàng triệu người 'có nguy cơ'. Tạp chí Forbes. 2019. www.forbes.com/sites/zakdoffman/2019/09/02/chinese-best-ever-DF-app-zao-sparks-huge-faceapp-like-privacy-storm/?sh=2951ebb88470 (truy cập ngày 27 tháng 1, 2021).

- [6] Ahn SJ, & Bailenson J. Quảng cáo tự xác nhận: Khi bản thân thuyết phục bản thân. *Tạp chí Lý thuyết và Thực hành Marketing* 2014; 22:135-136.
- [7] Levine TR. (2019) Bị lừa: Lý thuyết mặc định về sự thật và Khoa học xã hội về dối trá và lừa dối. Tuscaloosa, AL: Nhà xuất bản Đại học Alabama.
- [8] Bond Jr CF, & DePaulo BM. Độ chính xác của các phán đoán lừa dối. *Đánh giá Nhân cách và Tâm lý Xã hội* 2006; 10:214-234.
- [9] Segovia KY, & Bailenson JN. Hầu như đúng: Trẻ em có được những ký ức sai trong thực tế ảo. *Tâm lý truyền thông* 2009; 12:371-393.
- [10] Hướng dẫn Goodfellow I. Nips 2016: GANs 2016; bản in trước arXiv arXiv:1701.00160. Tác động xã hội của DFS.
- [11] Hancock JT, Woodworth MT, & Goorha S. Không thấy điều ác: Ánh hưởng của phương tiện giao tiếp và động lực đối với việc phát hiện lừa dối. *Đàm phán quyết định nhóm* 2010; 19:327-343.
- [12] Suwajanakorn S, Seitz SM, & Kemelmacher-Shlizerman I. Tổng hợp Obama: Học hát nhép từ âm thanh. *Giao dịch ACM trên Đồ họa (ToG)* 2017; 36:1-13.
- [13] Mahbub Md. K, Biswas M, Gaur L, Alenezi F, & Santosh K. Các tính năng sâu giúp phát hiện các bất thường ở phổi trên phim X-quang ngực do bệnh truyền nhiễm X: Covid-19, viêm phổi và lao. *Khoa học Thông tin* 2022; 592: 389-401. <https://doi.org/10.1016/J.INS.2022.01.062>.
- [14] Sharma S, Singh G, Gaur L, & Sharma R. Khoảng cách tâm lý và tín ngưỡng có ảnh hưởng đến hành vi gian lận của khách hàng không? *Tạp chí Nghiên cứu Người tiêu dùng Quốc tế* 2022. [10.1111/ijcs.12773](https://doi.org/10.1111/ijcs.12773).
- [15] Zaman N & Gaur L. Phương pháp tiếp cận và ứng dụng của học sâu trong chăm sóc y tế ảo. *IGI toàn cầu*. 2022. doi:[10.4018/978-1-7998-8929-8.ch002](https://doi.org/10.4018/978-1-7998-8929-8.ch002)

10 Phát hiện tin giả

Sử dụng máy học

Sonali Raturi, Amit Kumar Mishra, và
Srabanti Maji

NỘI DUNG

10.1 Giới thiệu	121
10.2 Lý do sử dụng mạng xã hội cho tin giả.....	122
10.3 Lý do lan truyền tin giả.....	122
10.4 Tài khoản độc ác trên mạng xã hội để vận động chính sách.....	122
10.5 Các phương pháp phát hiện tin tức giả mạo.....	123
10.6 Hồi quy tuyến tính	123
10.7 Rừng ngẫu nhiên	123
10.8 Cây quyết định	123
10.9 Bộ phân loại tăng cường độ dốc.....	124
10.10 Trình phân loại thư động-tích cực	124
10.11 Công việc liên quan.....	124
10.12 Mục tiêu	125
10.13 Phương pháp	127
10.14 Bộ dữ liệu	127
10.15 Tiền xử lý dữ liệu	127
10.16 Đánh giá mô hình.....	128
10.17 Kết quả và thảo luận	128
10.18 Tóm tắt	131
Người giới thiệu.....	131

10.1 GIỚI THIỆU

Chúng ta đang sống trong một xã hội nơi mọi người thường dựa vào các nguyên tắc của mạng xã hội, nơi nhiều người có thể tìm kiếm và nhận tin tức hoặc bài đăng từ mạng xã hội chứ không phải từ các tin tức thông thường như báo chí. Tin giả là tin kém chất lượng, chứa tin sai sự thật được tạo ra một cách có chủ ý. Sự lan truyền rộng rãi của tin tức giả ngày nay qua ngày khác có khả năng gây ra những tác động xấu to lớn đối với xã hội hoặc bất kỳ cá nhân nào [1]. Tin giả được viết ra nhằm đánh lừa người đọc để họ tin vào những thông tin sai sự thật được tạo ra một cách có chủ ý, dẫn đến việc chỉ dựa vào nội dung đưa tin sẽ khó phát hiện tin giả; do đó, chúng tôi cần liên quan đến thông tin dành riêng [2] có thể là sự tham gia xã hội của người dùng trên phương tiện truyền thông xã hội giúp hình thành kết luận.

10.2 LÝ DO SỬ DỤNG MẠNG XÃ HỘI CHO TIN GIẢ

Trong trường hợp phương tiện truyền thông xã hội, nó phải được cung cấp kịp thời và không quá tốn kém để người tiêu dùng tiếp nhận tin tức hơn là các phương tiện truyền thông tin tức truyền thống khác như báo chí. Phương tiện truyền thông xã hội giúp dễ dàng chia sẻ thêm tin tức hoặc bình luận và chúng tôi có thể xem xét cập nhật với sự trợ giúp của những độc giả khác một cách dễ dàng hơn.

Tuy nhiên, các bài báo được [3] sản xuất trực tuyến vì nó ít tốn kém hơn và nhanh hơn để phát hành tin tức thông qua phương tiện truyền thông xã hội. Chúng được lấy trực tuyến cho các mục đích khác nhau như lợi ích chính trị và tài chính. Trong tình hình đại dịch này, tin đồn lan truyền với tốc độ nhanh hơn. Dữ liệu giả đang lan truyền trên mạng xã hội cùng với các biện pháp khắc phục. Cách để phân biệt tin tức thực với thông tin sai lệch là liên kết các thuộc tính và lý thuyết đa dạng trên các phương tiện truyền thông, tức là phương tiện thông thường cũng như phương tiện truyền thông xã hội.

Bây giờ, những hạn chế trong dự đoán tin giả sẽ được xác định và các phương pháp sẽ được xem xét. Tiếp theo, các bộ dữ liệu sẽ được sử dụng trong phương pháp này và việc đánh giá một mô hình mới được sử dụng bởi các phương pháp hiện có sẽ được xác định. Chủ yếu có hai tính năng cơ bản: tính xác thực và mục đích. Đầu tiên, bằng chứng sai có thể được xác minh. Thứ hai, nó được tạo ra để đánh lừa người tiêu dùng với ý định không trung thực.

10.3 LÝ DO LUYỆN TIN GIẢ

Tin giả có thể là tin đồn thường không được tạo ra từ bất kỳ sự kiện tin tức nào, chỉ vì lợi ích chính trị hoặc bất kỳ lợi ích tài chính nào. Tin tức giả mạo có thể là thông tin sai lệch được tạo ra mà không được tính toán trước. Tin giả có thể được tạo ra để mua vui hoặc để mua chuộc một người cụ thể. Gần đây, tin tức giả mạo rất năng động khi thay đổi giai đoạn từ phương tiện truyền thông truyền thống sang phương tiện truyền thông xã hội hoặc tin tức trực tuyến. Đây là hai thành phần khiến người dùng gặp nguy hiểm với những tin tức hoặc bài đăng sai sự thật:

Chủ nghĩa hiện thực ngày thơ: Trong trường hợp này, người dùng bắt đầu tin rằng quan điểm của họ về thực tế là quan điểm duy nhất chính xác [4] và những người có quan điểm khác nhau được coi là định kiến.

Xu hướng xác nhận: Trong trường hợp này, người dùng tin rằng chỉ nhận được thông tin rằng các chế độ xem hiện tại của họ được xác nhận.

10.4 TÀI KHOẢN NỘI TRÊN MẠNG XÃ HỘI CHO ÚNG HỘ

Lý do chính cho các tài khoản độc hại có thể là hiệu quả chi phí của việc tạo tài khoản trên mạng xã hội. Việc tạo bot trực tuyến cho phương tiện truyền thông xã hội sẽ ít tốn kém hơn.

Một bot có thể là một tài khoản trên phương tiện truyền thông xã hội và được quản lý bởi các thuật toán máy tính khác nhau để nó có thể tự động tạo nội dung và liên kết với bot hoặc mọi người trên phương tiện truyền thông xã hội [5]. Các bot xã hội được cho là thực thể độc hại khi nó được thiết kế với mục đích cụ thể, về cơ bản là để gây hại, chẳng hạn như phát tán hoặc thao túng tin tức lừa bịp trên mạng xã hội. Mọi người bắt đầu tin vào những tin tức lừa bịp vì những yếu tố sau:

Do độ tin cậy trên mạng xã hội, điều đó có nghĩa là người dùng có thể coi một nguồn tin giả là đáng tin cậy nếu những người khác coi cùng một nguồn là đáng tin cậy. Và

họ làm như vậy khi không có đủ thông tin để quyết định xem nguồn đó là giả hay thật, hoặc tính trung thực của bất kỳ nguồn nào.

Do kinh nghiệm tàn sói, có nghĩa là người dùng [6] bắt đầu hỗ trợ một cách tự nhiên thông tin mà họ nghe đài nghe lại thậm chí nó có thể là tin giả.

10.5 CÁC PHƯƠNG PHÁP PHÁT HIỆN TIN GIẢ

Đó là một cách để xác định tin giả. Và công việc này dựa trên việc phát hiện tin tức giả mạo trên mạng xã hội bằng cách sử dụng Machine Learning (ML). Có một số thuật toán ML khác nhau, bao gồm Naïve Bayes và RNN [7]. Sử dụng các thuật toán được đề xuất này, chúng tôi có thể tạo ra một công cụ để xác định tin giả này trên mạng xã hội.

10.6 HỒI QUY TUYẾN TÍ NH

Đó là một phương pháp mô hình hóa một giá trị cuối cùng dựa trên các yếu tố dự đoán độc lập. Kỹ thuật độc đáo này là một thuật toán ML được giám sát. Nhiệm vụ hồi quy được thực hiện trong thuật toán này. Ứng dụng của kỹ thuật này là dự báo. Phương pháp này khác biệt dựa trên số lượng giá trị độc lập và mối liên kết giữa giá trị phụ thuộc và giá trị độc lập [8].

10.7 RỪNG NGẦU NHIÊN

Rừng ngẫu nhiên được tạo thành từ một số lượng lớn các cây quyết định. Một cây riêng biệt là một thuật toán loại bỏ dự đoán lớp [9]. Giống như cách kết hợp cổ phiếu và trái phiếu để xây dựng một danh mục đầu tư lớn hơn tổng số. Một số cây có thể đúng, hoặc những cây khác có thể sai, vì vậy cây có thể đi đúng đường. Vì vậy, các yếu cầu để hoạt động đúng của rừng ngẫu nhiên như sau:

Yêu cầu một số tín hiệu xác thực trong các tính năng để các mô hình được xây dựng thông qua các tính năng đó có thể hoạt động tốt hơn ước tính ngẫu nhiên [10]. Các dự đoán được thực hiện bởi các cây đơn lẻ phải có mối tương quan thấp với cây khác.

10.8 CÂY QUYẾT ĐỊNH

Nó là khái cơ bản của thiết kế rừng ngẫu nhiên. Ngoài ra, họ cực kỳ bẩn nết.

Chẳng hạn, việc hiểu chính xác cây quyết định hoạt động như thế nào trong trường hợp này sẽ dễ dàng hơn rất nhiều. Giả sử rằng bộ dữ liệu của chúng tôi có hai số 1 và năm số 0 và chúng tôi muốn phân loại. Điều này có thể được thực hiện với các tính năng như màu đỏ so với màu xanh lam cũng như liệu nhận xét có được nhấn mạnh hay không. Vì vậy, là nó có thể? Màu sắc dương như là một đặc điểm để phân chia tất cả các phương tiện, nhưng một số số 0 được biểu thị bằng màu xanh lam. Do đó, chúng ta có thể nói rằng 'Màu này có hiển thị là màu đỏ không?' Giả sử một nút trong cây giống như điểm mà đường mòn được chia thành hai, tức là Có nhánh và Không có nhánh. Nhánh Không có màu xanh lam là tất cả 0 và một nhánh khác

dấu vết sẽ có thể phân chia ngoài [11,12]. Những cái đại diện cho có, và số 0 đại diện cho dấu vết phụ chính xác.

10.9 PHÂN LOẠI TĂNG TỐC TĂNG TỐC

Để giải quyết các vấn đề về phân loại và hồi quy, kỹ thuật ML này được sử dụng.

Điều này có liên quan đến thuật toán cây quyết định; bắt cứ khi nào cây quyết định tạo ra hiệu suất như một người học yếu, thì thuật toán được gọi là cây tăng cường độ dốc. Nó tạo ra mô hình theo mô hình giai đoạn [13].

10.10 PHÂN LOẠI TÍ CH CỰC THỤ ĐỘNG

Thuật toán ML này có thể hữu ích cho một số ứng dụng nhất định. Nó được sử dụng cho việc học quy mô lớn. Trong trường hợp này, dữ liệu là nối tiếp và ML được cập nhật từng cái một trong đó toàn bộ dữ liệu đào tạo được sử dụng cùng một lúc.

Cách thức hoạt động của bộ phân loại Thụ động-Tích cực:

Bị động: Nếu dự đoán chính xác thì không thay đổi dữ liệu.

Tích cực: Nếu dự đoán không chính xác, thì có thể thực hiện các thay đổi đối với mô hình.

10.11 CÔNG VIỆC LIÊN QUAN

Sau đây là tổng quan tài liệu:

Mykhailio et al. (2017) đã nâng cao một phương pháp đơn giản để phát hiện tin tức giả cùng với việc sử dụng bộ phân loại Naïve Bayes. Đối với điều này, anh ấy đã sử dụng Tin tức BuzzFeed; anh ấy đã sử dụng nó để biết và kiểm tra bộ phân loại Naïve Bayes. Cody và cộng sự. (2017) đã đề xuất một kỹ thuật tự động nhận dạng tin giả; nó hoạt động trên Twitter. Họ đã áp dụng kỹ thuật này cho nguồn Twitter và truy cập dữ liệu để tự động phát hiện tin giả từ bộ dữ liệu tin giả của BuzzFeed trên Twitter. Marco và cộng sự.

(2017) nâng cao một bài báo. Trong bài báo này, họ thuật lại cách các mạng xã hội và tiện ích sử dụng và nghiên cứu các chiến lược ML khác nhau có thể được sử dụng để phát hiện tin tức giả.

Họ đã sử dụng và hoàn thành phương pháp phát hiện tin giả này bên trong bot Facebook Messenger và bắt đầu nó với các ứng dụng khác. Rishabh và cộng sự. (2015) đã hoàn thành ba thuật toán ML bao gồm Naïve Bayes và các thuật toán khác dưới dạng phần cụm trên một số tính năng. Các tính năng này có thể là mức độ của một tweet hoặc lượt thích người theo dõi, bất kỳ từ Spam nào được tạo có chủ ý, nhận xét của người dùng và thẻ bắt đầu bằng # được sử dụng trên mạng xã hội. Saranya và cộng sự. (2018) đã trình bày một khái niệm trong đó họ sử dụng một khung nâng cao hơn có thể phát hiện nội dung thông tin sai lệch. Lúc đầu, họ đã rút ra nội dung có chức năng khả năng sử dụng Twitter API. Và sau đó, tất cả các chức năng này được vận hành cùng một lúc với một số phân tích của Twitter; sau đó, nó đảo ngược việc tìm kiếm hình ảnh và xác thực tin giả được sử dụng bởi các thuật toán khác nhau; xác thực này được thực hiện cho lớp và phân tích. Shloka (2017) đã thực hiện một khái niệm trong đó NLP được áp dụng để ngăn chặn thông tin sai lệch. Trong đó, anh ấy đã áp dụng thời gian, một số khái niệm về bigram và cũng sử dụng khả năng phát hiện ngữ pháp phi ngữ cảnh trong khái niệm của mình. Marco và cộng sự.

(2018) đã trình bày một phương pháp phát hiện tin giả ML

Phát hiện tin tức giả bằng máy học

trong một chatbot Facebook Messenger. Ngoài ra, họ đã áp dụng phương pháp này với một ứng dụng trong thế giới vật chất. Họ đã kết hợp các phương pháp dựa trên xã hội và dựa trên nội dung phụ thuộc vào quy tắc ngưỡng. Stefan và cộng sự. (2018) đã trình bày một cách tiếp cận được giám sát yếu. Cách tiếp cận bóc đồng này thu thập một lượng lớn nhưng chứa một tập dữ liệu đào tạo ở ào ào nặng nề. Họ đã sử dụng Giám sát yếu, Học máy và Phân loại. Tháng và cộng sự. (2018) đã đề xuất rằng clickbait can thiệp vào tin tức giả mạo, có khả năng người dùng sẽ phát hiện ra thông tin hữu ích. Họ đã sử dụng phân loại trong cách tiếp cận của họ. Sagar et al. (2017) đã trình bày một mô hình phụ thuộc vào hai tính năng mới: kiểm tra ngôn ngữ và nhận dạng các tweet spam. Họ đã sử dụng NLP thống kê trong mô hình của họ và cả ML. Lourdes và cộng sự. (2010) đã đề xuất hệ thống phát hiện giả được tổ chức tốt, phụ thuộc vào bộ phân loại hợp nhất các tính năng dựa trên liên kết mới nhất và sử dụng các tính năng này với các tính năng dựa trên mô hình ngôn ngữ (LM) bằng cách sử dụng phân tích nội dung. Ye et al. (2019) đã trình bày phần trợ giúp Học sâu về Hệ thống Ước tính Nội dung Ngôn ngữ Tự nhiên. Hệ thống này phát hiện tin giả xác định thông tin sai sử dụng NLP.

Vanya et al. (2020) đã đề xuất một hệ thống phát hiện tin tức giả mạo để phân loại các tiêu đề hoặc văn bản tin tức là sai hay không sai bằng cách đánh giá các tiêu đề tin tức theo nhãn sử dụng ML, NLP và k-láng giêng gần nhất. Hoa nhài và cộng sự. (2020) đã đề xuất một kỹ thuật phát hiện tin giả sử dụng nhiều phương pháp phân loại khác nhau. Họ đã áp dụng các phương pháp phân loại như Passive-Aggressive classifier, SVM, và Naïve Bayes.

Kết quả của trình phân loại SVM có độ chính xác 95% và phương pháp được sử dụng để trích xuất tính năng là TF-IDF. Mykhailo và cộng sự. (2020) đã trình bày một mô hình mà họ áp dụng bộ phân loại Naive Bayes. Ở đây, mô hình này được thực thi dưới dạng một hệ thống phần mềm và bộ dữ liệu về các bài đăng tin tức trên Facebook được sử dụng để thử nghiệm. Độ chính xác phân loại là 74% đã đạt được trên dữ liệu thử nghiệm. Alia et al. (2018) đã đề xuất Hệ thống theo dõi tin đồn sử dụng phương pháp quét web. Mô hình này khám phá các nguồn trên các trang web và xác nhận nội dung bằng cách sử dụng Thuật toán khớp thông tin. Youngkyung Seo và cộng sự. (2015) đã trình bày một hệ thống phát hiện tin tức sai phụ thuộc vào học sâu và dự đoán liệu bài đăng là giả hay thật bằng cách sử dụng phép biến đổi ngữ pháp. Hệ thống này bao gồm một số lớp: lớp đối sánh, lớp suy luận, lớp nhúng và lớp tạo ngữ cảnh. Chúng tôi cũng tóm tắt tổng quan tài liệu ở định dạng bảng (Bảng 10.1).

10.12 MỤC TIÊU

Mục tiêu chính của mô hình này là nhận ra các vấn đề về tin giả trên mạng xã hội, để mọi người có thể dễ dàng phân biệt giữa tin giả và tin thật. Mô hình này đã sử dụng quá trình xử lý Ngôn ngữ tự nhiên và bộ phân loại Thủ động-Tích cực và thu được độ chính xác. Việc lan truyền tin giả gây tác động tiêu cực đến mọi người [14,15,16].

Do đó, việc phát hiện tin tức giả mạo là cần thiết để giảm sự lan truyền của những tin tức không liên quan như vậy. Việc sử dụng TF-IDF để chuyển đổi văn bản thành một mô tả số có ý nghĩa được sử dụng để dự đoán bằng các thuật toán ML.

Chúng tôi tóm tắt mục tiêu của chúng tôi ở đây:

Mục tiêu chính của mô hình này là nhận ra các vấn đề về tin giả trên mạng xã hội để mọi người có thể dễ dàng phân biệt giữa tin giả và tin thật. Cái này

BẢNG 10.1**Tóm tắt khảo sát văn học**

S. số	Tác giả	Tiếp cận	Kết quả
1	Kyeong-Hwan Kim et al.	Ghép câu, Học sâu, NLP (2017)	Máy dò tin giả của Hàn Quốc được tạo ra và có thể được cập nhật bởi sự phán xét của con người
2	A Di Đà Dey et al.	Phân cực, Ngôn ngữ học (2017)	Một khuôn khổ có thể được sử dụng để đưa ra quyết định tốt hơn
3	Marco L. Della Vedova et al.	Kết hợp dựa trên xã hội và kỹ thuật dựa trên nội dung (2017)	Phát hiện tin tức giả mạo Machine Learning được sử dụng trong một Facebook Messenger
4	Stefan Helmstetter et al. (2015)	Giám Sát Yếu, Máy Học hỏi	Một cách tiếp cận được giám sát yếu. Ở đây, nó thu thập dữ liệu lớn nhưng tập dữ liệu huấn luyện ồn ào.
5	Tháng Aldwairi et al. (2018)	phân loại	Một mô hình với khả năng của người dùng để phân biệt hữu ích thông tin
6	Palagati Bhanu, Prakash Reddy và cộng sự. (2017)	NLP, rừng ngẫu nhiên, KNN, SVM, Cây quyết định, Naïve Bayes	Trình phát hiện tin giả bằng thuật toán bồi phiêu đã thức
7	Sagar Ghouse và cộng sự. (2018)	Thống kê NLP, Máy móc Học hỏi	Một kỹ thuật phụ thuộc vào hai nguyên tắc: phát hiện các tweet giả mạo và một nguyên tắc khác phụ thuộc vào việc phân tích ngôn ngữ
8*	Burak et al. (2018)	Xử lý ngôn ngữ tự nhiên	Một mô hình cho thuật toán NER
9	Alina Campan và cộng sự. (2018)	Thông tin khuếch tán, tối đa hóa ảnh hưởng	Trình bày tin tức giả như thế nào tràn lan trên các mạng xã hội online hiện nay
10	Bhavika Bhutan et al. (2017)	Naive Bayes, Random Forest	Hệ thống phát hiện tin giả bao gồm tính cảm
			phân tích để cải thiện sự chính xác
11	Lourdes Araujo và cộng sự. (2010)	phân tích nội dung	Mô hình phát hiện tin giả đã sử dụng các tính năng dựa trên liên kết dựa trên mô hình ngôn ngữ
12	Ye-Chan Ahn và cộng sự. (2019)	Xử lý ngôn ngữ tự nhiên	Hệ thống phát hiện tin giả để đánh giá thông tin không chính xác đó bằng cách sử dụng Natural Nội dung ngôn ngữ Sự đánh giá
13	Va Vanya Tiwari và cộng sự. (2019)	NLP, KNN, học máy	Trình phát hiện tin tức giả mạo phân loại tiêu đề tin tức hoặc nhãn tin là giả hay không giả bằng cách phân tích các tiêu đề tin tức có nhãn

mô hình đã sử dụng xử lý Ngôn ngữ tự nhiên và bộ phân loại Thu động-Tích cực và tìm độ chính xác.

Việc lan truyền tin giả gây ảnh hưởng tiêu cực đến mọi người. Do đó, việc phát hiện tin tức giả mạo là cần thiết để giảm [17,18,19] sự lan truyền của những tin tức không phù hợp.

Sử dụng TF-IDF để chuyển đổi văn bản thành một mô tả có ý nghĩa về số bers được sử dụng để dự đoán với các thuật toán học máy.

10.13 PHƯƠNG PHÁP

Hệ thống được đề xuất của chúng tôi dự định đánh giá phân loại kiện trình phân loại Thu động-Tích cực bằng cách sử dụng bộ dữ liệu liên quan đến tin tức. Tập dữ liệu kết quả được chia thành hai loại phụ. Đầu tiên, tập Huấn luyện sử dụng 80% tập dữ liệu và thứ hai, tập Kiểm tra sử dụng 20% tập dữ liệu. Việc thu thập tập dữ liệu, tiền xử lý dữ liệu và phân loại được áp dụng cho mô hình.

10.14 BỘ DỮ LIỆU

Mô hình này sử dụng tập dữ liệu được thu thập từ Kaggle. Bộ dữ liệu này liên quan đến tin tức bao gồm ID, tiêu đề, tác giả và văn bản, tức là chủ đề của tin tức cùng với nhãn [20,21,22] của các giá trị đúng và sai, trong đó các giá trị bị thiếu được bỏ qua.

Tập dữ liệu cuối cùng không thể bị xáo trộn gồm 18.285 bài đăng có nhãn, một số ít được dán nhãn là giả và một số được dán nhãn là thật. Thông tin bổ sung liên quan đến tập dữ liệu được cung cấp trong Bảng 10.2. Các văn bản tin tức đã được đánh giá cho các tác vụ tiền xử lý được bao gồm trong phần tiền xử lý dữ liệu. Ngoài ra, kiến thức từ dữ liệu đã được thu thập để hiểu rõ hơn, sử dụng kỹ thuật ML.

10.15 XỬ LÝ SẼ DỮ LIỆU

Phần này bao gồm mã thông báo, từ vựng, loại bỏ các từ dừng, loại bỏ dấu chấm câu, bộ kiểm tra null và chuẩn bị dữ liệu được gửi đến mô hình để thực hiện các bước tiếp theo. Làm sạch dữ liệu là quá trình chuẩn bị dữ liệu để đánh giá, được thực hiện bằng cách loại bỏ dữ liệu không liên quan hoặc không phù hợp hoặc thay đổi dữ liệu. Dữ liệu đó có thể không phù hợp, không đầy đủ hoặc không chính xác

BẢNG 10.2

tập dữ liệu

	Tiêu đề	Tác giả	Chữ	Nhãn
0	House Dem Aide:	Darrell Lucas	House Dem Aide:	1
1	FLYNN: Hillary	Daniel J. Flynn	Bao giờ có được	0
2	Tại Sao Là Sự Thật	Consortiumnews.com	Elton in ra Jessica	1
3	15 thường dân thiệt mạng	Purkiss	Video 15 thường dân	1
4	phụ nữ Iran	Howard Portnoy	In Một phụ nữ Iran	1

được định cấu hình [23,24,25]. Vì nó cung cấp kết quả không chính xác nên dữ liệu được sử dụng ở đây không chính xác để kiểm tra dữ liệu. Có nhiều kỹ thuật khác nhau để làm sạch dữ liệu sẽ khác nhau tùy thuộc vào cách dữ liệu được lưu trữ cùng với kết quả. Tại đây, một mô hình sẽ được tạo ra có thể báo trước tính trung thực của tin tức theo thời gian thực. Từ vựng và mã thông báo được áp dụng bằng cách sử dụng Bộ công cụ ngôn ngữ tự nhiên (NLTK).

Lemmatization là quá trình thu thập các hình thức sửa đổi khác nhau của một từ để chúng có thể được kiểm tra như một mục riêng lẻ. Do đó, nó liên kết các từ có nghĩa chung với một từ duy nhất. Tokenization là một bước quan trọng trong các phương pháp Xử lý ngôn ngữ tự nhiên như Count Vectorizer cũng như Transformers. Đó là một cách để chia một đoạn văn bản thành các đơn vị nhỏ hơn được gọi là mã thông báo.

10.16 ĐÁNH GIÁ MÔ HÌNH

Sau khi xử lý trước dữ liệu, tập dữ liệu được chia thành hai, tức là dữ liệu thử nghiệm và dữ liệu huấn luyện, và thu được độ chính xác bằng cách sử dụng trình phân loại Thu động-Tích cực. Xử lý ngôn ngữ tự nhiên sử dụng thống kê, ML, học sâu và ngôn ngữ điện toán. Cùng với nhau, nó cho phép xử lý ngôn ngữ của con người dưới dạng văn bản hoặc giọng nói [26,27,28]. NLP thực thi các chương trình chuyển đổi văn bản từ ngôn ngữ này sang ngôn ngữ khác. Trình phân loại thu động-tích cực còn được gọi là thuật toán trực tuyến. Chúng tôi đã sử dụng thuật toán này khi chúng tôi có một luồng dữ liệu lớn.

Ở đây, nếu dự đoán thì áp dụng dự đoán nhân với vectơ trọng số là:

$$\text{dwT} > 0 \quad (10.1)$$

Nếu dương thì $y = +1$

Nếu âm thì $y = -1$

trong đó y = quan sát các lớp thực, khi nó nhỏ hơn 1, thì chúng tôi biểu diễn nó bằng cách thêm một phần nhỏ của D, đó là vectơ chúng tôi nhận được và vectơ trọng số. Hung hăng cho mắt mà chúng tôi đã nhận được.

10.17 KẾT QUẢ VÀ THẢO LUẬN

Chúng tôi đã sử dụng công cụ vector hóa TF-IDF để lấy các đặc điểm và phê duyệt chúng cho bộ phân loại. Như trong mô hình này, trình phân loại thu động-tích cực được áp dụng bằng cách sử dụng mã thông báo và

BẢNG 10.3

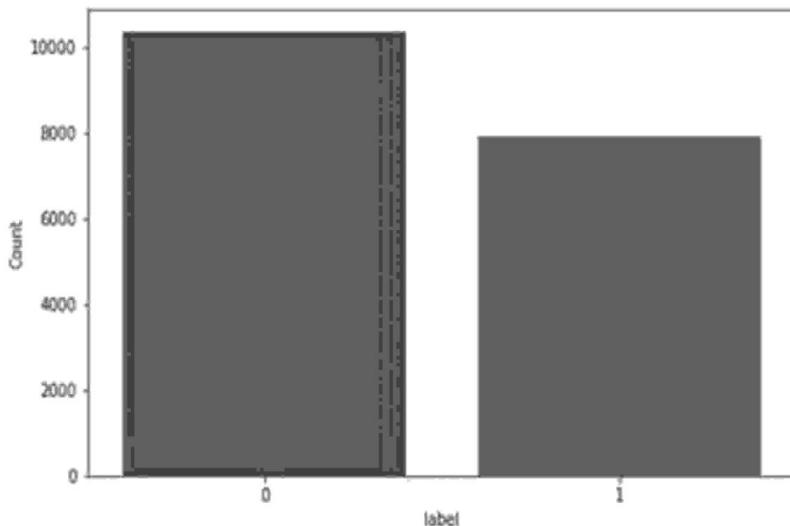
Dọn dẹp dữ liệu

	Nhãn
0 1	1
2	0
	1
3 4	1

BẢNG 10.4

Sự chính xác

Sự chính xác	Giá trị đích thực	Giá trị sai
95%	43,3%	56,7%



HÌNH 10.1 Trực quan hóa tập dữ liệu giả và thật.

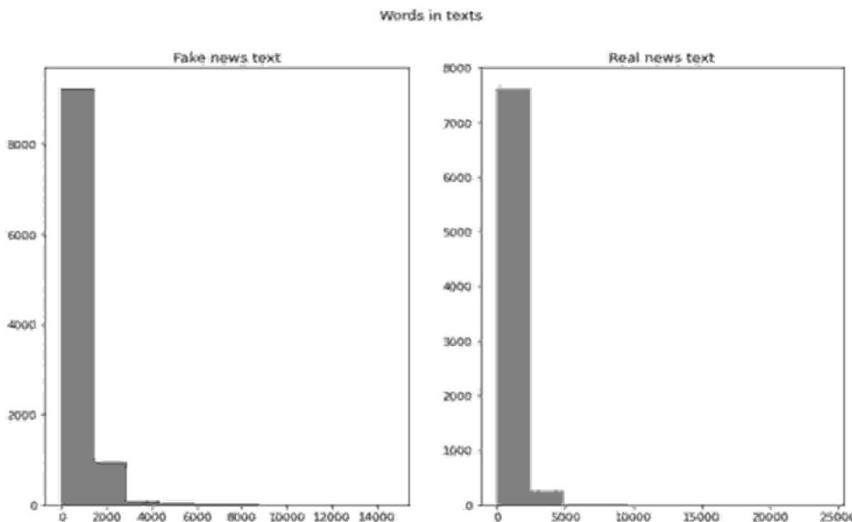
từ vựng với Xử lý ngôn ngữ tự nhiên và thu được độ chính xác. Mô hình này kết hợp với độ chính xác cung cấp một xấp xỉ của kỹ thuật định dạng tốt nhất. Xử lý ngôn ngữ tự nhiên (NLP) và nhiều công cụ của nó được sử dụng để làm sạch dữ liệu (Bảng 10.3). Nó là một phần quan trọng của phân loại văn bản.

Độ chính xác mà chúng tôi nhận được, như trong Bảng 10.4, rõ ràng đã phát minh ra rằng thuật toán phân loại Tích cực-Thụ động đang hoạt động tốt hơn với độ chính xác là 95%.

Bộ dữ liệu, đại diện cho tin tức, được thu thập từ Kaggle. Tập dữ liệu chứa tiêu đề, tác giả và văn bản cùng với nhãn, tức là '0' và '1', 'false' và 'true' tương ứng.

Nó chứa dữ liệu liên quan đến tin tức. Các giá trị còn thiếu được bỏ qua khỏi tập dữ liệu. Mô tả thêm về bộ dữ liệu được mô tả trong Bảng 10.2. Bộ dữ liệu chúng tôi đã sử dụng chứa 56,7% dữ liệu giả và 43,3% dữ liệu thật, như thể hiện trong Bảng 10.4. Thử nghiệm được thực hiện bằng nền tảng Python.

Trong quá trình tiền xử lý, chúng tôi đã xóa tất cả HTML, dấu chấm câu và từ dừng tiếng Anh. Thuật toán ML mà chúng tôi đã sử dụng trong mô hình của mình là thuật toán Tích cực thụ động để phân loại [29,30,31]. Thuật toán này được triển khai bằng cách sử dụng gói python scikit-learning. Một thử nghiệm đã được thực hiện trên hệ thống xử lý 64-bit. Các giá trị thiếu được loại bỏ bằng cách xóa các cột chứa giá trị null như trong Bảng 10.2. Trực quan hóa dữ liệu được thực hiện trong Hình 10.1. Nó cho thấy



HÌNH 10.2 Biểu đồ về số từ trong mỗi văn bản.

số lượng dữ liệu giả và thật trong các bộ dữ liệu đơn lẻ. Vì vậy, dữ liệu giả chúng tôi nhận được ở đây là 56,7% và dữ liệu thật là 43,3%.

Sau khi các quy trình làm sạch dữ liệu và loại bỏ dấu câu, HTML và từ dừng được thực hiện, chúng tôi nhận được dữ liệu hoàn toàn không có giá trị null và chúng tôi đã đo số lượng từ trong văn bản như trong Hình 10.2. Sau đó, chúng tôi chia tập dữ liệu thành dữ liệu thử nghiệm và dữ liệu huấn luyện. Trình phân loại Thụ động-Tích cực được áp dụng cho tập dữ liệu kết quả [32,33,34].

Như thể hiện trong Bảng 10.4, chúng tôi đã xử lý trước dữ liệu bằng Ngữ tự nhiên Xử lý với [35,36,37] Passive-Aggressive classifier với độ chính xác 95%.

Độ chính xác được thực hiện để đánh giá các mô hình phân loại. Một cách tự tin, độ chính xác là tỷ lệ dự báo mà mô hình của chúng tôi thu được chính xác. Chính thức, một công thức chính xác như được đề cập trong phần sau:

$$\text{độ chính xác} = \frac{\text{Số dự đoán đúng}}{\text{Tổng số dự đoán}} \quad (10.2)$$

Phương trình để đo lường độ chính xác về mặt tích cực và tiêu cực được mô tả như sau:

$$\text{độ chính xác} = \frac{\text{TP TN}}{\text{TP TN FP FN} + + +} \quad (10.3)$$

trong đó TP là Kết quả xác thực đúng, FP là Kết quả xác thực sai, FN là Kết quả phủ định sai và TN là Kết quả phủ định thực.

Chúng tôi đã thu được độ chính xác của các kỹ thuật và liệu mức trung bình của chúng có gần với giá trị thực đang được nêu hay không.

10.18 TÓM TẮT

Công việc phân loại tin giả một cách thủ công cần có kiến thức sâu về chuyên môn, lĩnh vực để nhận ra những điểm bất thường trong văn bản. Trong nghiên cứu này, chúng tôi tạo ra một mô hình phát hiện tin tức giả mạo bằng cách sử dụng NLP và bộ phân loại Thu động-Tích cực. Dữ liệu chúng tôi sử dụng trong công việc của mình được thu thập thông qua Kaggle. Nghiên cứu nhằm xác định các mẫu trong văn bản giúp phân biệt tin giả với tin thật. Chúng tôi đã rút ra các cách làm sạch dữ liệu khác nhau bằng cách sử dụng Trình phân loại xử lý ngôn ngữ tự nhiên và thu động tích cực để phân loại với độ chính xác 95%. Việc phát hiện tin giả còn nhiều vấn đề cần nhận thức của các nhà nghiên cứu. Ví dụ, chúng ta cần nhận ra các yếu tố chính để giảm sự lan truyền của tin giả. Các phương pháp ML có thể được sử dụng để xác định các yếu tố liên quan đến sự phát triển của tin giả. Ngoài ra, phát hiện tin tức sai bằng học sâu có thể là một hướng đi khác trong tương lai bằng cách sử dụng nhiều tính năng khai thác khác nhau.

NGƯỜI GIỚI THIỆU

- [1] Araujo, L. và Martinez, RJ (2010). Phát hiện thư rác trên web: Các tính năng phân loại mới dựa trên các mô hình ngôn ngữ và phân tích liên kết đủ điều kiện. Giao dịch của IEEE về Điều tra và Bảo mật Thông tin, 581-590.
- [2] Ahn, Y. và Jeong, C. (2019). Hệ thống đánh giá nội dung ngôn ngữ tự nhiên để phát hiện tin tức giả bằng Deep Learning. Hội nghị chung quốc tế lần thứ 16 về Khoa học máy tính và Kỹ thuật phần mềm (JCSSE).
- [3] Buntain, C. và Golbeck, J. (2017). Tự động xác định tin tức giả mạo trong các chủ đề phổ biến trên Twitter. Hội nghị Quốc tế IEEE về Đám mây Thông minh.
- [4] Chin, P.-Y., Choo, KKR và Evans, N. (2015). Khám phá các yếu tố ảnh hưởng đến việc sử dụng mạng xã hội doanh nghiệp trong các công ty dịch vụ chuyên nghiệp đa quốc gia. Tạp chí Máy tính Tổ chức và Thương mại Điện tử, 25, 289-315.
- [5] Davis, C., Ferrara, E., Flammini, A. và Varol, O. (2016). Sự gia tăng của các bot xã hội. Truyền thông của ACM, 96-104.
- [6] Granik, M. và Mesyura, V. (2017). Phát hiện tin tức giả sử dụng Bayes Classifier, IEEE First Ukraine Conference on Electrical and Computer Engineering (UKRCON).
- [7] Gupta, A. và Kaushal, R. (2015). Cải thiện khả năng phát hiện thư rác trong mạng xã hội trực tuyến Mạng, 978-1-4799-7171.
- [8] Gilda, S. (2017). Đánh giá các thuật toán học máy để phát hiện tin tức giả mạo. Hội nghị Sinh viên lần thứ 15 về Phát triển Nghiên cứu (SCoReD) của IEEE.
- [9] Granik, M. và Mesyura, V. (2017). Phát hiện tin tức giả bằng Trình phân loại Naïve Bayes. IEEE Hội nghị Ukraine đầu tiên về Kỹ thuật Điện và Máy tính (UKRCON), 900-903.
- [10] Chavan, M. và Gharge, S. (2017). Một cách tiếp cận tích hợp để phát hiện các tweet độc hại bằng cách sử dụng NLP. Hội nghị quốc tế về truyền thông sáng tạo và công nghệ tinh toán (ICICCT) năm 2017.
- [11] Helmstetter, S. và Paulheim, H. (2018). Học tập được giám sát yếu để phát hiện tin tức giả mạo trên Twitter. Hội nghị quốc tế IEEE/ACM về những tiến bộ trong xã hội Phân tích và khai thác mạng (ASONAM), 274-277.
- [12] Chen, M. và Krishnan, S. (2018). Xác định Tweets với Fake News. Tích hợp và tái sử dụng thông tin IEEE cho khoa học dữ liệu.
- [13] Ali, A. và Monther, A. (2018). Phát hiện tin giả trên mạng truyền thông xã hội, mạng quốc tế lần thứ 9. Hội nghị quốc tế lần thứ 9 về Hệ thống phổ biến mới nổi và Mạng phổ biến (EUSPN).

- [14] Ghani, V., Mochamad, A., và Setijadi, P., A. (2015). Thiết kế và Triển khai Xử lý Ngôn ngữ Tự nhiên với Phân tích Ngữ nghĩa và Cú pháp để Trích xuất Điều kiện Giao thông từ Dữ liệu Mạng Xã hội. Hội nghị quốc tế lần thứ 5 về Kỹ thuật và Công nghệ hệ thống của IEEE.
- [15] Nickerson, RS (1998). Xu hướng xác nhận: Một hiện tượng phổ biến dưới nhiều chiêu bài. Đánh giá về Tâm lý học đại cương, 2(2), 175.
- [16] Patil, R. và Shaikh (2020). Phát hiện tin tức giả bằng Machine Learning. Hội nghị chuyên đề quốc tế của IEEE về Năng lượng bền vững, Xử lý tín hiệu và An ninh mạng (ISSSC), 1-5.
- [17] Samreen, A., Ahmad, A., và Zeshan, F. (2020). Tìm kiếm Sự thật trong Thời đại Hậu Sự thật. Hội nghị quốc tế lần thứ 3 về những tiến bộ trong khoa học tính toán (ICACS), 1-5.
- [18] Seo, Y. và Jeong, C. (2018). FaGoN: Mô hình phát hiện tin tức giả sử dụng phép biến đổi ngữ pháp trên mạng nơ-ron. Hội nghị quốc tế lần thứ mười ba về Hệ thống Hồ trợ Tri thức, Thông tin và Sáng tạo (KICSS), 1-5.
- [19] Tiwari, V., Lennon, RG và Dowling T. (2020). Không phải mọi bạn đọc đều đúng! Phát hiện tin tức giả bằng Thuật toán máy học. Hội nghị hệ thống và tín hiệu Ireland lần thứ 31 (ISSC), 1-4.
- [20] Ballarin, G., de Massimo, AL, Moret, S., Vendova, MLD, và Tacchini, E. (2017). Tự động phát hiện tin giả trực tuyến kết hợp nội dung và tín hiệu xã hội, 2305-7254.
- [21] Ballarin, G., de Alfaro, L., DiPierro, M., Moret, S., Vedova, DML, và Tacchini, E. (2018). Tự động phát hiện tin tức giả trực tuyến kết hợp nội dung và tín hiệu xã hội. Hội nghị Hiệp hội Sáng tạo Mở (FRUCT), 272-279.
- [22] Brown, T., Ward, A., Ross, L., Reed, E., và Turiel, E. (1997). Chủ nghĩa hiện thực ngây thơ trong cuộc sống hàng ngày: Hệ lụy đối với xung đột và hiểu lầm xã hội. Giá trị và Kiến thức, 103-135. https://web.mit.edu/curhan/www/docs/Articles/15341_Readings/Negotiation_and_Conflict_Management/Ross_Ward_Naive_Realism.pdf
- [23] Juan, C., Jiebo, L., Yongdong, Z., và Zhiwei, J. (2016). Xác minh tin tức bằng cách khai thác các quan điểm xã hội mâu thuẫn trong các tiêu blog. Trong AAAI'16.
- [24] Anshu, K., Gaur, L., và Khazanchi, D. (2017). "Đánh giá mức độ hài lòng của các nhà bán lẻ điện tử tạp hóa bằng cách sử dụng mô hình TOPSIS và ECCSI mờ trực quan," Hội nghị quốc tế về Công nghệ Infocom và Hệ thống không người lái (Xu hướng và Định hướng Tương lai) (ICTUS), trang 276-284. doi:10.1109/ICTUS.2017.8286019
- [25] Gaur, L. và Anshu, K. (2018). Phân tích sở thích của người tiêu dùng cho các trang web sử dụng e TailQ và AHP. Tạp chí Kỹ thuật & Công nghệ Quốc tế, 7(2.11), 14-20.
- [26] Rana, J., Gaur, L., Singh, G., Awan, U., và Rasheed, MI (2021). Cung cấp hành trình của khách hàng thông qua trí tuệ nhân tạo: Chương trình nghiên cứu và đánh giá. Tạp chí quốc tế về các thị trường mới nổi, Tập. trước khi in (No. trước khi in). <https://doi.org/10.1108/IJOM-08-2021-1214>
- [27] Gaur, L., Singh, G., Solanki, A., Jhanjhi, NZ, Bhatia, U., Sharma, S., . và Kim, W. (2021). Xu hướng của giới trẻ trong việc dự đoán các mục tiêu phát triển bền vững bằng thuật toán rừng ngẫu nhiên và thần kinh mờ. Khoa học thông tin và máy tính lấy con người làm trung tâm, 11, NA.
- [28] Singh, G., Kumar, B., Gaur, L. và Tyagi, A. (2019). "So sánh giữa Multinomial và Bernoulli Naïve Bayes để phân loại văn bản," Hội nghị quốc tế về tự động hóa, tính toán và quản lý công nghệ (ICACTM), trang 593-596. doi:10.1109/ICACTM.2019.8776800
- [29] Gaur, L., Agarwal, V., và Anshu, K. (2020), "Phương pháp tiếp cận DEMATEL mờ để xác định các yếu tố ảnh hưởng đến hiệu quả của ngành bán lẻ Ấn Độ," Đàm bảo hệ thống chiến lược và

- Phân tích kinh doanh. Phân tích nội dung (Quản lý hiệu suất và an toàn). Springer, Singapore. https://doi.org/10.1007/978-981-15-3647-2_
- [30] Gaur, L., Afaq, A., Singh, G., và Dwivedi, YK (2021). Vai trò của trí tuệ nhân tạo và người máy trong việc thúc đẩy du lịch không chạm trong thời kỳ đại dịch: Chương trình nghiên cứu và đánh giá. Tạp chí Quốc tế về Quản lý Khách sạn đương đại, 33(11), 4079-4098. <https://doi.org/10.1108/IJCHM-11-2020-1246>
- [31] Sharma, S., Singh, G., Gaur, L., và Sharma, R. (2022). Khoảng cách tâm lý và tôn giáo có ảnh hưởng đến hành vi lừa đảo của khách hàng không? Tạp chí Nghiên cứu Người tiêu dùng Quốc tế. doi:10.1111/ijcs.12773
- [32] Sahu, G., Gaur, L., và Singh, G. (2021). Áp dụng phương pháp tiếp cận lý thuyết thích hợp và hài lòng để kiểm tra niềm đam mê của người dùng đối với các nền tảng vượt trội và truyền hình thông thường. Viễn thông và Tin học, 65. doi:10.1016/j.tele.2021.101713
- [33] Ramakrishnan, R., Gaur, L., và Singh, G. (2016). Tính khả thi và hiệu quả của các thiết bị IoT đèn hiệu BLE trong quản lý hàng tồn kho tại khu vực cửa hàng. Tạp chí Quốc tế về Kỹ thuật Điện và Máy tính, 6(5), 2362-2368. doi:10.11591/ijece.v6i5.10807
- [34] Afaq, A., Gaur, L., Singh, G., và Dhir, A. (2021). COVID-19: Thay đổi hành vi của hành khách đi máy bay và định hình lại kỳ vọng của họ đối với ngành hàng không. Nghiên cứu giải trí du lịch. doi:10.1080/02508281.2021.2008211
- [35] Mahbub, Md. K., Biswas, M., Gaur, L., Alenezi, F., và Santosh, K. (2022). Các tính năng sâu để phát hiện các bất thường về phổi trên phim X-quang phổi do bệnh truyền nhiễm X: Covid-19, viêm phổi và bệnh lao. Khoa học Thông tin, 592, 389-401. <https://doi.org/10.1016/J.INS.2022.01.062>
- [36] Sharma, S., Singh, G., Gaur, L., & Sharma, R. (2022). Khoảng cách tâm lý và tôn giáo có ảnh hưởng đến hành vi gian lận của khách hàng không? Tạp chí Nghiên cứu Người tiêu dùng Quốc tế. 10.1111/ijcs.12773.
- [37] Zaman, N., & Gaur, L. (2022). Phương pháp tiếp cận và ứng dụng của Deep Learning trong Chăm sóc y tế áo. IGI. doi:10.4018/978-1-7998-8929-8.ch002



Taylor & Francis
Taylor & Francis Group
<http://taylorandfrancis.com>

11 Tương lai của DeepFakes và Ectypes

Bò tót Loveleen, Mansi Ratta, và Bò tót Adesh

NỘI DUNG

11.1 Giới thiệu	135
11.2 DeepFake và thực tế.....	136
11.3 Giả mạo và Ectypes.....	137
11.4 Trình phát hiện giả mạo hình ảnh với DeepFakes	137
11.5 Các vấn đề pháp lý với DeepFakes và các chiến lược trong tương lai (Quyền sở hữu).....	138
11.6 Chiến lược trong tương lai: Thông qua các khía cạnh tổ chức	138
11.7 Các chiến lược trong tương lai: Thông qua các khía cạnh xã hội.....	139
11.8 Các chiến lược trong tương lai: Thông qua các khía cạnh của chính phủ	139
11.9 Truyền thông xã hội: Nhận liệu cho nội dung giả mạo?.....	140
11.10 Đấu tranh với “Phản hồi sẵn sàng”	140
11.11 Hệ thống pháp luật và chính sách (giữa các quốc gia khác nhau).....	141
11.12 DeepFakes có tồn tại hay không?/Dự báo tương lai.....	142
11.13 Kỳ vọng trong tương lai từ DeepFake.....	142
Người giới thiệu.....	143

11.1 GIỚI THIỆU

Một giáo sư từ Đại học Đông Bắc đã thành lập công ty tái tạo giọng nói.

Nó nhằm mục đích cung cấp giọng nói tùy chỉnh cho những người không thể nói. Ngay cả một người bình thường cũng có thể tùy chỉnh giọng nói của họ nếu họ có nguy cơ mất nó trong tương lai. Dự án tiên phong này có thể trả lại cho mọi người một phần thiết yếu trong danh tính của họ [1]. Trong một trường hợp, giám đốc điều hành của một công ty năng lượng ở Vương quốc Anh đã lừa chuyển 200.000 euro cho một nhà cung cấp Hungary vì anh ta tin rằng ông chủ của mình đang chỉ đạo anh ta làm như vậy. Những trường hợp như vậy làm cho DeepFakes (DF) trở nên không an toàn cho xã hội và người dân của nó. Mỗi đe dọa thực sự là việc sử dụng công nghệ này để truyền bá thông tin sai lệch. Ý nghĩa đạo đức và cảm nhận đối với công nghệ như DF là rất lớn. Hãy tưởng tượng điều này: một người bị buộc tội và tìm thấy một đoạn video từ cảnh quay giám sát ở một thành phố hoàn toàn khác. Đoạn phim sau đó được đưa vào quy trình DF để đưa khuôn mặt và cơ thể của người đó vào đoạn phim. Người vô tội sẽ bị buộc tội ngay lập tức.

Hoặc điều gì sẽ xảy ra nếu các cuộc gọi điện thoại có thể được tạo bằng giọng nói chính xác của một người khác [2]. Điều đó có thể dẫn đến hành vi trộm cắp danh tính, lừa đảo, v.v. Chúng tôi đã thấy DF có khả năng gây nghi ngờ về tính hợp pháp của bất kỳ hình thức truyền thông kỹ thuật số nào. Khó phân biệt được đâu là thật, đâu là giả.

Hai năm trước, một tập đoàn gồm hơn 100 công ty công nghệ, bao gồm AWS, Facebook và Microsoft, đã tạo ra một thử thách phát hiện DF với giải thưởng 1 triệu đô la cho bất kỳ ai tìm ra cách phát hiện DF hiệu quả. Một ý tưởng khác đang được đưa ra xung quanh là một số dạng hình mờ kỹ thuật số được phát hành và lưu trữ bằng cách sử dụng chuỗi khối.

Khi tài sản kỹ thuật số được tạo, nó sẽ được gán một ID duy nhất [3]. ID này giống như dấu vân tay được tạo từ bản gốc. Điều này sau đó có thể được sử dụng để xác minh tính xác thực của nội dung đó sau này hoặc phát hiện ra hàng giả. Dự kiến sẽ có hai dòng chính của công nghệ DF trong tương lai. Một bên sẽ tăng cường việc sử dụng và độ tin cậy của nó [4]. Người kia sẽ phát hiện đâu là DF và đâu là không. Cả hai sẽ quan trọng như nhau.

11.2 DEEPFAKES VÀ SỰ THẬT

Ranh giới giữa thực và ảo kỹ thuật số ngày càng mờ nhạt. Đồng thời, nó mờ ra các hình thức sáng tạo và giải trí mới cũng như lừa đảo và tuyên truyền. Một số nhà lập pháp lo ngại rằng DF có thể được sử dụng để tấn công nền dân chủ. Hãy tưởng tượng một video xuất hiện vào ngày bầu cử cho thấy một ứng cử viên đang ở vị trí thỏa hiệp. Có thể không có thời gian để cảnh sát hoặc báo chí điều tra nghiêm ngặt và công nghệ phát hiện DF vẫn đang trong quá trình hoàn thiện. Vào thời điểm sự thật được đưa ra ánh sáng, có thể đã quá muộn. Về mặt sáng sủa hơn, nó cho phép các hình thức sáng tạo mới. Phương tiện truyền thông xã hội đang tràn ngập các ví dụ về DF được tạo ra để giải trí. Giống như một loạt phim trong đó Tom Cruise xuất hiện để tập đánh gôn và chơi ghi-ta [5]. Nó có một tương lai thường được sử dụng trong các video quảng cáo và đào tạo của công ty. Công nghệ này đang tiến đến điện thoại và mạng xã hội của chúng ta.

Một công ty khởi nghiệp tên là Rosebud đã tạo ra các ứng dụng có thể thay đổi tuổi của bạn trong một bức ảnh selfie hoặc chuyển nét mặt của bạn sang khuôn mặt của Leonardo DiCaprio. Tất cả những thứ có vẻ tự nhiên này có thể khiến một người cảm thấy hơi khó chịu. Mọi người lo lắng rằng trong thời đại của DF, sự thật sẽ bị mất mãi mãi. Tuy nhiên, chúng ta cũng có thể lập luận rằng chúng ta đã sống trong một thế giới kỹ thuật số hậu hiện thực được một thời gian rồi. Kể từ khi Photoshop ra đời vào những năm 1990, mọi người đã quen với ý tưởng rằng hình ảnh có thể bị làm giả một cách thuyết phục. Thế giới chưa chính xác kết thúc. Chúng tôi chỉ biết rằng không coi bức ảnh chụp người ngoài hành tinh hay Quái vật hồ Loch Ness là bằng chứng xác thực [6]. Các luật mới và công nghệ phát hiện cuối cùng có thể hạn chế một số mặt tối của DF. Nhưng hầu hết sự thích nghi sẽ đến từ chúng ta, con người. Chúng tôi đã quen với ý tưởng rằng ảnh có thể bị thao túng và chúng tôi tìm kiếm các dấu hiệu đáng tin cậy khác, chẳng hạn như nguồn hoặc các ngữ cảnh khác nhau. Nhưng chúng ta sẽ phải phát triển các chuẩn mực mới cho một thế giới nơi phương tiện truyền thông tổng hợp ở khắp mọi nơi và học cách chung sống với DF.

DF có thể gây ra sự thao túng toàn bộ quan điểm của khán giả. Những quan điểm này có thể là về lựa chọn mua của một người, lựa chọn bỏ phiếu, những gì họ muốn và những gì họ không muốn. Và nó tinh tế đến mức khán giả thậm chí sẽ không biết rằng họ đã được hướng dẫn mua một thứ hoặc bỏ phiếu cho ai đó [7]. Trong các tình huống chính trị, DF có khả năng thao túng cử tri do dự bỏ phiếu cho một vấn đề cụ thể một cách dễ dàng.

11.3 GIẢ MẠO VÀ LOẠI NGOẠI HÌNH

Kể từ vài thập kỷ trước, chúng tôi đã thấy mọi người bắt chước các tác phẩm gốc bằng cách tạo ra các bản sao hoặc bản sao. Những tác phẩm này có thể bao gồm việc bắt chước một bức ảnh nổi tiếng, tạo ra một bản sao giả mạo của một bức tranh nổi tiếng, hoặc thậm chí là sao chép một thiết kế đồ họa mà không đề cập đến tình trạng không có nguồn gốc và tính xác thực của nó. Nếu tôi vẽ lại bức tranh sơn dầu "Những đêm đầy sao" của Vincent Van Gogh và ký tên nó giống như một tác phẩm gốc nói rằng họa sĩ cũ đã vẽ nó, thì đó sẽ là một cách bán bức tranh rõ ràng, không trung thực. Gọi một bản sao giả là bản gốc là phi đạo đức [8]. Lấy cảm hứng từ công việc của người khác và tạo lại nó với các khoản tín dụng cần thiết hoặc tinh minh bạch là cách đúng đắn để thực hiện mọi việc. Đó là nơi mà từ ectypes xuất hiện. Nó chỉ đơn giản đề cập đến công việc sao chép nhưng phần nào được phân biệt với công việc thực tế.

Một ví dụ có thể giải thích rõ hơn về ectypes. Vài năm trước, Microsoft, liên kết với Bảo tàng Nhà Rembrandt, đã sản xuất tác phẩm nghệ thuật làm mờ xanh giới giữa tác phẩm gốc và tác phẩm không nguyên bản. Một hệ thống Trí tuệ nhân tạo (AI) đã được tạo ra để có một bức tranh lấy cảm hứng từ các tác phẩm trước đây của Rembrandt. Nó đã xác định những đặc điểm chung nhất trong các bức tranh gốc của ông và tạo ra một tác phẩm lưu giữ những chi tiết nhô về các bức tranh của Rembrandt. Tác phẩm này không phải là bản gốc của Rembrandt cũng không phải là bản sao giả [9].

Đây là nơi mà từ ectypes có thể được sử dụng. Là một ectype làm cho nó trở nên hợp đạo đức và làm rõ tình trạng xác thực của một tác phẩm. Nó có thể là âm thanh, hình ảnh hoặc thậm chí là video.

Trong khi đó, việc sử dụng thuật ngữ ectypes làm danh tính của tác phẩm là một giải pháp cho vấn đề về hành vi không trung thực do DF gây ra. Các cách khác cũng nên có sẵn, chẳng hạn như để phát hiện các video giả mạo về các chính trị gia được lưu hành trước khi nó tạo ra bầu không khí lộn xộn giữa khán giả. Khi thời gian trôi qua, tương lai của chúng ta dường như có nhiều ứng dụng của công nghệ DF [10]. Với việc công nghệ này trở nên phổ biến trong những năm tới, nhu cầu về một giải pháp hiệu quả cũng tăng lên. Chúng ta, với tư cách là một xã hội, đã phải vật lộn với những hình ảnh giả mạo và bị thao túng trong một thời gian dài. Và ngày nay, có nhiều giải pháp thiết thực cho vấn đề này.

11.4 MÁY PHÁT HIỆN HÌNH ẢNH GIẢ MẠO BẰNG DEEPFAKES

Ví dụ: "Công cụ phát hiện giả mạo hình ảnh" của Scorto Corporation là một doanh nghiệp hoàn toàn tập trung vào việc xác định hình ảnh giả mạo. Họ phân loại ảnh do khách hàng cung cấp là đã phát triển, đáng ngờ và xác thực. Tương tự như vậy, nhiều giải pháp khác tồn tại trong thị trường vũ hội. Vì vậy, trong những năm tới, với những kỹ thuật và thuật toán phù hợp, có thể sẽ có một giải pháp hiệu quả cho mặt tối của DF. Trong những năm gần đây, rất nhiều nghiên cứu đã được thực hiện trong lĩnh vực này. Một trong số đó là một công cụ do Microsoft tạo ra [11]. Công cụ này lấy ảnh hoặc video làm thông tin đầu vào và cung cấp điểm tin cậy cho biết ảnh/video đó có phải là giả mạo hay không. Công cụ này xác định các khu vực được pha trộn trong ảnh/video mà một người nhanh chóng không chú ý đến.

Một ví dụ khác là nghiên cứu của Đại học Binghamton và Intel, nơi họ đã thành lập một hệ thống phát hiện giúp xác định mô hình được sử dụng cùng với việc phát hiện ảnh/video bị thao túng. Những công cụ này chỉ có thể cung cấp trợ giúp ở một mức độ nào đó vì chúng không phải lúc nào cũng chính xác. Ngoài ra, theo thời gian, có sự cập nhật liên tục trong việc hình thành DF. Những tiến bộ này cũng đòi hỏi các công cụ và giải pháp mạnh mẽ.

Do đó, khi các ứng dụng và phương pháp của công nghệ này tăng lên, các giải pháp rắn cũng được cho là sẽ loại bỏ mặt tiêu cực của DF.

11.5 CÁC VẤN ĐỀ PHÁP LÝ VỚI DEEPFAKES VÀ CHIẾN LƯỢC TƯƠNG LAI (QUYỀN SỞ HỮU)

(Quyền sở hữu có đủ để chống lại DF trong tương lai không?)

Đến thời điểm này, chúng tôi đã xác định rằng có một nhược điểm đi kèm với một số ứng dụng DF. Nó có khả năng vi phạm quyền bảo vệ của chúng tôi và sử dụng sai dữ liệu của bất kỳ cá nhân nào bằng cách thao túng các tính năng của nó. Một ví dụ nổi tiếng là việc tạo ra các video khiêu dâm sử dụng hình ảnh khuôn mặt của phụ nữ. Mặc dù có khá nhiều sai sót trong việc sử dụng DF trong xã hội của chúng ta, mối quan tâm đầu tiên này sinh liên quan đến quyền dữ liệu của một người. Người tạo một DF nhất định có được phép sử dụng dữ liệu ban đầu thuộc về người khác không? Anh ta có được quyền thao túng nó không? Đây là những câu hỏi mà người sáng tạo cần ghi nhớ trước khi tạo DF. Vào năm 2019, Tổ chức Sở hữu Trí tuệ Thế giới (WIPO) đã đưa ra một dự thảo liên quan đến quyền sở hữu đối với AI, trong đó họ đã đưa vào một tiêu mục dành cho DF. Nó đã giải quyết các vấn đề liên quan đến quyền sở hữu trí tuệ do công nghệ này gây ra [12]. Nó đã đề cập rằng bản quyền của một DF nhất định thuộc về người tạo ra nó sau khi họ đã nhận được sự cho phép các quyền từ người có dữ liệu đang được sử dụng (âm thanh/

ảnh/video). Đó là bởi vì ý tưởng và sự đổi mới thuộc về người sáng tạo, và anh ta nên được ghi công cho nó. Ngoài ra, nếu nội dung được tạo nhằm mục đích hiển thị người (có dữ liệu đang được sử dụng) dưới ánh sáng xáu, thì người tạo không nên được chỉ định bất kỳ loại quyền nào đối với tài sản. Trong khi nó cũng nói rằng các vấn đề với công nghệ này không thể được giải quyết chỉ bằng cách tạo ra một khuôn khổ bản quyền. Các vấn đề về quyền riêng tư và quyền quản lý hình ảnh hoặc trạng thái của một số người khác ở nơi công cộng không thể chỉ giới hạn trong các chính sách bản quyền.

Theo EU, Quy định bảo vệ dữ liệu chung cho Liên minh châu Âu đảm bảo rằng dữ liệu cá nhân được sử dụng để tạo DF phải xác thực và chính xác [13]. Trong trường hợp nó được phát hiện là cũ hoặc không chính xác, thì người tạo có nghĩa vụ xóa hoặc sửa nó ngay lập tức. Nó cũng đảm bảo xóa hoàn toàn nội dung DF nếu nội dung đó không liên quan hoặc gây hiểu lầm.

Bên cạnh việc đảm bảo tính xác thực của dữ liệu, luật này trao quyền cho nạn nhân của DF thực hành quyền xóa nội dung của họ mà không có bất kỳ sự chậm trễ nào. Những quy định này làm cho DF không quá tệ đối với cư dân châu Âu. Những quy tắc và quy định như vậy là cần thiết nếu chúng ta muốn duy trì mối quan hệ tốt hơn với DF trong tương lai.

11.6 CÁC CHIẾN LƯỢC TƯƠNG LAI: QUA CÁC KHÓA TỔ CHỨC

Những người trong công ty có hiểu những khả năng AI này mà tôi phạm mạng có thể sử dụng không? Họ có sẵn sàng đầu tư vào nghiên cứu để tạo ra một hệ thống mạnh mẽ có thể giải quyết những vấn đề này không? Mặc dù nhu cầu về các giải pháp được cá nhân hóa ngày càng tăng, nhưng mối lo ngại về các vấn đề riêng tư cũng ngày càng tăng. Làm thế nào các tổ chức có thể gặp nhau ở giữa?

Trong khi các công ty công nghệ lớn cần nghiên cứu các giải pháp kỹ thuật để giải quyết vấn đề này đồng thời, họ cũng cần giải quyết vấn đề này từ các góc độ khác. Một trong những bước cần thiết là

bỗ nhiệm đúng người có thể tạo ra nhận thức sâu sắc trong nhân viên. Các nhân viên nên biết việc rơi vào bẫy của một DF bị lỗi có thể ảnh hưởng đến tổ chức của họ như thế nào. Nếu một tổ chức muốn duy trì trong tương lai khi các cuộc tấn công DF gia tăng, họ cần hướng dẫn các nhóm của mình cách phát hiện nội dung giả mạo như vậy. Các tổ chức cũng có thể kết hợp các bài học kỹ thuật xã hội cho nhân viên của họ. Nên có sự chuẩn bị trước. Các khu vực dễ bị tổn thương có thể bị ảnh hưởng bởi DF nên sử dụng các thuật toán phát hiện. Xử lý rủi ro là một khái niệm khác cần được chú ý trong khi tạo nhận thức. Cho dù thuật toán của họ có mạnh đến đâu hay mức độ nhận thức của mọi người, thì vẫn có khả năng chúng ta mắc lỗi [14]. Do đó, các nhóm của tổ chức nên biết các bước sau để khắc phục thiệt hại đã tạo ra. Ngoài ra, các nhóm nên biết cách điều tra các động sau sự kiện của rủi ro. Cần có một nhóm được hướng dẫn tốt có thể phân tích tổn thất một cách chính xác và xem mức độ ảnh hưởng của nó đối với khách hàng. Vì hoạt động bị lỗi này không phải là vấn đề đau đầu của riêng một công ty, tất cả các tổ chức nên chung tay và tạo ra một hệ thống để cung cấp các rào cản ngăn chặn mọi nội dung tổng hợp. Hệ thống này có thể là một tiêu chuẩn và sau đó được các công ty khác làm theo [15]. Một hành vi có trách nhiệm và có trách nhiệm từ phía tổ chức có thể là một DF-mỗi trường an toàn.

11.7 CHIẾN LƯỢC TƯƠNG LAI: THÔNG QUA CÁC KHÍ A SẮC XÃ HỘI

Chắc chắn rồi, DF có thể được tạo ra để chống lại ai đó-tổ chức, cá nhân, cộng đồng, v.v. Trách nhiệm đầu tiên của họ là duy trì một hệ thống an toàn không thể bị tổn hại, nhưng điều đó không có nghĩa là người dùng/người tiêu dùng/xã hội không cần chịu trách nhiệm. Nó có thể không gây ra bất kỳ tác hại trực tiếp nào cho cá nhân xem nó nhưng ý kiến hoặc quan điểm của anh ta đã thay đổi dựa trên một mẫu tin giả. Ví dụ: một video giả mạo về một chính trị gia sẽ không gây hại cho cử tri đã xem video đó, nhưng ý kiến của cử tri này bị ảnh hưởng bởi video đó có thể khiến anh ta bỏ phiếu cho nhầm người, người cuối cùng có hại cho xã hội của chúng ta. Trong tương lai, những sự cố như thế này sẽ không thành vấn đề nếu cộng đồng của chúng ta được hướng dẫn và giáo dục tốt về những vấn đề như vậy [16]. Hiểu biết về phương tiện truyền thông là một yếu tố thiết yếu trong sự lan truyền và tác động của DF. Tạo ra một xã hội hiểu biết về truyền thông sẽ giúp mọi người nhận ra sự khác biệt giữa thông tin chính xác và nội dung tổng hợp. Mô hình này giúp chống lại DF và những bất ngờ kỳ lạ mà các công nghệ tương lai mang lại cho chúng ta.

11.8 CÁC CHIẾN LƯỢC TƯƠNG LAI: THÔNG QUA CÁC PHƯƠNG TIỆN CHÍ NH PHỦ

Các chính sách và sự tham gia của chính phủ cũng quan trọng không kém. Chính phủ nên hợp tác với các tổ chức để phát minh ra các giải pháp kỹ thuật hiệu quả để chống lại vấn đề này.

Một khía cạnh nữa là truyền bá kiến thức. Các chính phủ nên rõ ràng và minh bạch hơn với công chúng về bất cứ khi nào họ sử dụng DF trong bất kỳ chiến dịch hoặc quảng cáo nào của họ. Nó sẽ hữu ích theo nhiều cách. Đầu tiên, điều này sẽ lan truyền thông tin về DF tới khán giả thuộc mọi lứa tuổi (thậm chí cả công dân) bởi vì hầu hết mọi nhóm tuổi đều xem tin tức (ít nhất là tin tức liên quan đến chính phủ). Ngay cả khi nhiều người không biết ý nghĩa của nó, thì ít nhất điều này sẽ khiến họ tò mò muốn biết DF là gì. Thứ hai, điều này sẽ giới thiệu cho họ các ứng dụng của DF và làm quen với cách nó đang được phổ biến. Thứ ba, điều này sẽ khiến họ đặt câu hỏi về bất kỳ loại

truyền thông, lưu ý, chịu trách nhiệm và điều tra trước khi chia sẻ thêm bất kỳ nội dung nào [17]. Hơn nữa, đây sẽ là một bước tiến trong kiến thức truyền thông của xã hội chúng ta. Nó cũng có thể giúp khuyến khích cộng đồng của chúng tôi đi sâu vào nghiên cứu sâu và tìm ra giải pháp vì theo thời gian, chất lượng của DF sẽ ngày càng tốt hơn.

11.9 TRUYỀN THÔNG XÃ HỘI: NHIÊN LIỆU CHO NỘI DUNG GIẢ MẠO?

Phương tiện truyền thông xã hội là nguồn chính của tất cả nội dung ngày nay. Ngày nay, nội dung chúng ta tiêu thụ chủ yếu đến từ các nền tảng truyền thông xã hội. Ở Ấn Độ, mỗi khi bạo loạn xảy ra, rất nhiều bạo lực được thúc đẩy bởi các tin nhắn WhatsApp được chuyển tiếp và các video phóng đại được gửi với số lượng quá nhiều. Khi công chúng tìm thấy điều gì đó kỳ lạ hoặc gây tranh cãi, họ sẽ chia sẻ điều đó với những người quen biết của họ và chu kỳ chuyển tiếp nội dung này tiếp tục diễn ra. Hầu hết thời gian, mọi người thậm chí không xác minh tính xác thực của nội dung. Đôi khi nó quá thật để nghi ngờ tính xác thực của nó. Nếu nguyên nhân gốc rễ được xử lý, nó có thể ngăn chặn quá nhiều thiệt hại xảy ra. Đây là nơi các nền tảng truyền thông xã hội và hành động của nó phát huy tác dụng. Tạo một hệ thống phát hiện trên các nền tảng truyền thông xã hội không dễ dàng như vẻ ngoài của nó. Nó đòi hỏi các thuật toán mạnh mẽ hơn được thiết kế đặc biệt. Tại sao nó không dễ dàng? Bởi vì việc xác định khi nào nên loại bỏ DF và khi nào nên để nó ở lại là một công việc tẻ nhạt. Vấn đề là các nền tảng truyền thông xã hội đối xử với tất cả các loại nội dung có vấn đề như nhau. Chúng tôi biết rằng không phải tất cả các DF đều bị lỗi và gây tranh cãi. Do đó, chúng không cần phải được gỡ xuống. Một vấn đề khác là không phải tất cả các DF sai sự thật đều bị lỗi.

Lấy ví dụ thế này: Một người hát không hay nhưng lại sử dụng DF; anh ấy đã làm cho mình có âm thanh tốt trong một video. Điều này không có vấn đề gì và không cần phải gỡ xuống. Những tình huống như thế này đòi hỏi hệ thống phát hiện phải chính xác. Ngoài ra, sẽ rất lạ nếu nhầm mục tiêu DF trong khi có sẵn nội dung giả mạo khác. Những trường hợp như thế này gợi ý rằng cần có một hệ thống phát hiện được xác định chính xác và có hướng dẫn rõ ràng. Chỉ xác định các DF “bị lỗi”, “có vấn đề” hoặc “độc hại” là một nhiệm vụ và thách thức cần được các nền tảng truyền thông xã hội quan tâm trong tương lai.

Nếu các chính sách quan trọng không được điều chỉnh kịp thời, điều đó có thể dẫn đến việc giảm niềm tin vào tin tức được truyền qua Twitter, Facebook hoặc bất kỳ nền tảng truyền thông xã hội nào khác [19].

11.10 CHIẾN ĐẤU VỚI “PHẢN ỨNG SẴN SÀNG”

Chúng tôi không thể chờ đợi một phép màu bất ngờ xảy ra, tạo ra một hệ thống phát hiện chính xác 100%. Đây là một chủ đề nóng và rất nhiều nghiên cứu đang diễn ra để tìm ra cách giải quyết vấn đề này. Khi chúng tôi không có hệ thống phát hiện để dựa vào, có thể khó xác minh hoặc chứng minh điều gì đó sai. Đây là lý do tại sao để chuẩn bị sẵn sàng cho các cuộc tấn công trong tương lai, một tổ chức nên tập trung vào việc tạo ra phản hồi trước thời hạn. Ví dụ: nếu nội dung tổng hợp được lưu hành trong trường hợp khẩn cấp, sẽ không có thời gian để ngồi chờ nó biến mất vì không thể làm gì bây giờ. Hiểu nó với ví dụ này: Trong các cuộc bạo loạn, lũ lụt hoặc thảm chí là lốc xoáy, một video được lưu hành trong đó một số nguồn đáng tin cậy được hiển thị nói rằng mọi thứ bên ngoài đều ổn và không có gì khủng khiếp đang xảy ra bên ngoài. Nghe vậy, những người ở trong nhà có thể ra ngoài và gặp rắc rối [20]. Bây giờ, không có thời gian để lập một kế hoạch để hoàn tất mọi thứ

trong một tình huống như thế này. Đó là lý do tại sao các tổ chức hoặc bất kỳ ai khác nên có sẵn “kế hoạch ứng phó”. Trong phản hồi sẵn sàng này, họ có thể đã thiết lập sẵn các tập lệnh video có thể được lưu hành để giải thích cho mọi người biết DF là gì, nó diễn ra như thế nào, đề xuất một số nguồn xác thực và tuyên bố rằng họ không tạo ra nó. Nó sẽ giúp vạch trần cuộc tranh cãi và ngăn ngừa thiệt hại thêm. Điều này là cần thiết bởi vì tin tức càng cảng tồn tại lâu mà không có phản hồi thì nó sẽ càng ở trong vòng lưu thông lâu hơn.

Phản hồi rõ ràng được đưa ra càng sớm thì càng tốt. Việc duy trì dữ liệu của các bản ghi sự kiện liên quan đến việc có mặt trên máy ảnh hoặc phát biểu trước đám đông có thể được sử dụng làm bằng chứng khi tình huống như vậy phát sinh. Nó có thể xác minh tuyên bố phản hồi của bạn và hoàn tác thiệt hại ở mức độ lớn.

Trong khi điều cần thiết là phải phản hồi càng sớm càng tốt, thì cũng cần phải xác định kịp thời hành vi nguy hiểm này [21]. Chúng ta càng sớm biết rằng có DF thì càng tốt cho chúng ta. Thật khó để phát hiện ra nó bởi vì những người tạo ra nó có gắng để nó tránh xa tầm mắt của những người có liên quan. Trong một tương lai đầy DF, các tổ chức có thể sử dụng bot khẩn cấp hoặc cảnh báo. Nó sẽ làm việc như thế nào? Nó sẽ tập trung vào những người trong tổ chức luôn theo dõi camera và gửi cảnh báo mỗi khi họ tìm thấy điều gì đó liên quan đến người này. Nó có thể là các bài báo, blog ngắn nhiên hoặc các hình thức nền tảng truyền thông xã hội. Nếu bạn xác định rằng mọi người đang nói về người này thông qua các thẻ bắt đầu bằng # hoặc trong các bình luận, điều đó cho thấy có điều gì đó không ổn.

11.11 HỆ THỐNG PHÁP LUẬT VÀ CHÍNH SÁCH (TRÊN CÁC QUỐC GIA KHÁC NHAU)

Trước đó, chúng tôi đã đọc về các quy định được Liên minh Châu Âu thông qua về quyền dữ liệu. Chính phủ của tất cả các quốc gia phải bắt đầu can thiệp bằng cách tạo ra các quy tắc và chính sách mới. Nó sẽ xây dựng một khuôn khổ pháp lý sẽ cung cấp cho xã hội một số định hướng, bao gồm cả những người phát minh ra nội dung đó. Hãy tìm hiểu xem các chính phủ khác đang làm gì để chống lại điều này.

Rất ít quốc gia đã tạo ra các quy định chống lại nội dung lan truyền thông tin sai lệch. Mỹ, Anh, Úc và Trung Quốc là một vài trong số đó. Nhưng liệu những luật này có khả năng xử lý khả năng giả mạo cấp cao của nội dung DF không? Vào năm 2019, Hoa Kỳ đã đưa ra luật chống lại DF. Theo luật này, bắt buộc phải thêm dấu nước vào nội dung tổng hợp để dễ nhận biết. Ngoài ra, Virginia đã thông qua luật thực hiện các cáo buộc hình sự đối với người phát tán video DF khiêu dâm mà không có sự đồng ý của người trong đó. Texas là một ví dụ khác về luật hạn chế/cấm tạo bất kỳ nội dung DF nào bị lỗi và có vấn đề liên quan đến các vấn đề chính trị.

Vào năm 2019, California đã cấm sử dụng nội dung DF liên quan đến các chính trị gia trước 60 ngày diễn ra cuộc bầu cử để ngăn chặn mọi nội dung gây hiểu lầm. Mặc dù không có luật tự nhiên nào chống lại DF ở Ấn Độ, nhưng một luật (được thông qua vào năm 2019) đảm bảo quyền bảo vệ dữ liệu của một người và hạn chế việc sử dụng dữ liệu của một cá nhân. Nó bảo vệ ở một mức độ nào đó, nhưng không có luật nào bảo vệ dữ liệu của một người không còn sống nữa.

Một thực tế rất nổi tiếng là Ấn Độ có một hệ thống chính trị rất rộng lớn-rất nhiều đảng phái chống lại nhau, những cuộc chiến hàng ngày, chỉ trích những người khác dưới ánh sáng xấu chỉ là một vài khía cạnh. Trong một môi trường như vậy, DF có thể phạm sai lầm. Ví dụ,

một video bị chỉnh sửa về người chết của bên A có thể gây ra tranh cãi lớn và làm lung lay niềm tin của công chúng [22]. Nó sẽ làm cho bên A trở nên tồi tệ trong mắt công chúng. Đó là lý do tại sao các luật cụ thể để bảo vệ dữ liệu của người chết cũng nên tồn tại. Đôi với các kịch bản tương tự, một luật của Tây Ban Nha cũng đã được thông qua vào năm 2018, trao “quyền xóa nội dung” cho người thừa kế của người đã chết. Có, những luật này không trực tiếp tập trung vào nội dung DF, nhưng nó mang lại một số an toàn. Tương lai sẽ cho biết mức độ hiệu quả của những luật này bởi vì internet là một không gian rộng lớn khiến cho việc dự đoán bất cứ điều gì trờ nên khó khăn. Với tất cả các luật này của các quốc gia khác nhau, Trung Quốc đã đi trước một bước. Trong các chính sách trước đó vào năm 2019, Trung Quốc đã cấm phân phối thông tin sai lệch và nội dung sai lệch được tạo ra bằng kỹ thuật AI, bao gồm cả VR. Các quy tắc mới đi trước một bước và áp đặt lệnh cấm đối với bất kỳ ứng dụng nào đẩy nội dung giả mạo hoặc bị xâm phạm về phía người dùng bằng thuật toán chất lượng cao. Nó đảm bảo rằng các nền tảng truyền thông xã hội sử dụng các thuật toán mạnh mẽ để cung cấp nội dung tùy chỉnh không lan truyền nội dung giả mạo tới người dùng. Chúng mở rộng chủ yếu đối với các nền tảng tin tức. Bất kỳ ứng dụng truyền thông xã hội nào cũng không được phép có nội dung bị thao túng theo luật này ở Trung Quốc. Luật này cũng gây áp lực lên những người tạo ra các nền tảng truyền thông xã hội khác. Các bước hợp tác để hình thành các quy định cũng có thể hữu ích [23]. Sẽ có ý nghĩa trong một số trường hợp khi thông tin được truyền qua nhiều biên giới (vì một số nền tảng tồn tại trên toàn cầu). Nó có thể trở nên phức tạp hơn để phát hiện và hiểu trong tương lai. Việc thông qua luật là cần thiết và có thể chứng minh tính hiệu quả vì công nghệ này vẫn đang trong giai đoạn đầu.

11.12 DEEPFAKES CÓ TỒN TẠI HAY KHÔNG?/DỰ BÁO TƯƠNG LAI

Vâng, không có câu trả lời trực tiếp và chắc chắn cho câu hỏi này. Bằng cách xem xét tất cả các vấn đề và các ứng dụng tích cực liên quan đến nó, chúng ta có thể hiểu rằng công nghệ này có thể đi theo ba con đường-tích cực, tiêu cực hoặc thuận lợi. Chỉ có tương lai mới làm sáng tỏ liệu có thể chỉ có những tác động tích cực hay không. Có vẻ như khá khó khăn (không phải là hoàn toàn không thể) vào ngày hôm nay. Kịch bản này trong tương lai có thể xảy ra với các khuôn khổ và quy định nghiêm ngặt. Nếu DF được sử dụng một cách tích cực, chúng có thể tạo ra vô số cơ hội và ứng dụng. Một số lĩnh vực là ngành công nghiệp giải trí (bao gồm cả phim), cá nhân hóa cấp độ tiếp theo (nâng cao mối quan hệ với khách hàng), quảng cáo cho người nổi tiếng, ngành thời trang, mạng xã hội, lồng tiếng (cho phim hoặc quảng cáo), ngành trò chơi, mục đích giáo dục, hoặc thậm chí ngành y tế.

Đó sẽ là một tình huống tốt nhất [24]. Trường hợp xấu nhất có thể xảy ra nếu tất cả các biện pháp không được thực hiện đúng cách. Nó có thể xảy ra khi an ninh không chặt chẽ. Nó có thể dẫn đến một tương lai đầy lửa đao, mạo danh nhân vật của công chúng (người nổi tiếng và chính trị gia), mất tiền, tạo ra các vấn đề về lòng tin, vi phạm dữ liệu cá nhân, ID giả, v.v. Nó sẽ dẫn đến việc mất lòng tin vào AI nói chung [25]. Trường hợp thứ ba rất dễ xảy ra phần nào nằm ở cả tiêu cực và tích cực. Đó là trạng thái hiện tại của nó - tiêu cực và tích cực đồng thời [26].

11.13 KÝ VỌNG TƯƠNG LAI TỪ DEEPFAKE

Giống như các kỹ thuật AI khác đã xuất hiện trong cuộc sống hàng ngày của chúng ta, chúng ta có thể mong đợi điều tương tự với DF. Có thể hiểu được nếu ai đó nói rằng AI sẽ đóng một vai trò rất lớn trong

ngành công nghiệp giải trí và truyền thông trong tương lai. Mặc dù nó đã bắt đầu mở đường vào lĩnh vực tiếp thị thông qua quảng cáo và các kỹ thuật tiếp thị khác, nhưng chúng ta có thể mong đợi nhiều hơn nữa trong những năm tới. Bỏ mặt tối của nó sang một bên, chúng tôi biết rằng nó là một công cụ thay đổi cuộc chơi để đưa người tiêu dùng gần hơn với người bán và thương hiệu của họ vì nó tạo ra tình huống đôi bên cùng có lợi. AI được xã hội chấp nhận vì nó dễ dàng thực hiện trong các công việc hàng ngày.

Tương tự như vậy, DF sẽ giúp người bán bằng cách giảm bớt nhiệm vụ tạo hoạt động tiếp thị được cá nhân hóa và tùy chỉnh cho nhiều đối tượng khách hàng của mình. Nhiều nhà nghiên cứu tham gia vào việc phát minh ra các lĩnh vực mới để sử dụng công nghệ này. Một trong những lĩnh vực là ngành công nghiệp game. Nó được kỳ vọng sẽ tận dụng tối đa công nghệ này. Những người tạo ra trò chơi đã phải vật lộn để làm cho chúng gần với thực tế trong một thời gian dài. Trí tưởng tượng của mọi game thủ đều có hình ảnh siêu thực trong một trò chơi.

DF làm cho nó có vẻ như có thẻ. Có khuôn mặt của một người nổi tiếng trên trình phát của họ, chuyển động cơ thể thực tế, biểu cảm thời gian thực hoặc thậm chí là các vật thể thực tế trong nền sẽ mang lại trải nghiệm tuyệt vời cho bất kỳ người chơi nào. Nó sẽ giúp giảm chi phí cho sản xuất. Không chỉ cộng đồng game thủ hiện có mà còn thu hút người dùng mới. Một lĩnh vực đáng ngạc nhiên khác đang được nghiên cứu là chăm sóc sức khỏe. Ý tưởng là tạo một bệnh nhân và dữ liệu sức khỏe của anh ta bằng DF. Nó có thể được sử dụng để thử nghiệm cho các mục đích điều trị khác nhau. Sử dụng nó có thể tốt vì bệnh nhân thực tế sẽ không phải trải qua bất kỳ thử nghiệm nào và tác dụng phụ của nó. Ngoài ra, sẽ không sử dụng được dữ liệu chính xác, đây là mối quan tâm chính của DF.

Một khía cạnh trái của công nghệ này bắt đầu bị phủ nhận, mọi người sẽ bắt đầu chấp nhận nó. Nó có tiềm năng to lớn được sử dụng trong các ứng dụng sáng tạo và độc đáo.

Hãy cùng chờ xem mọi thứ sẽ diễn ra như thế nào trong những năm tới!

NGƯỜI GIỚI THIỆU

- [1] Somers, M., DFs, giải thích <https://mitsloan.mit.edu/ideas-made-to-matter/DFs-explained> (Truy cập ngày 17 tháng 8 năm 2021).
- [2] Cole, S. (24 tháng 1 năm 2018). "Chúng tôi thực sự chết tiệt: Bây giờ mọi người đang làm phim khiêu dâm giả do AI tạo ra." Hành vi xấu xa. Bản gốc lưu trữ ngày 7 tháng 9 năm 2019. Truy cập ngày 4 tháng 5 năm 2019.
- [3] Karousos, S. (tháng 9 năm 2020). "AI in Digital Media: Ký nguyên của DF," trong IEEE Transactions on Technology and Society, Vol. 1, Số 3, trang 138-147. doi:10.1109/TTS.2020.3001312
- [4] Thies, J., Zollhöfer, M., Stamminger, M., Theobalt, C., và Nießner, M. Face2Face: Chụp và tái hiện khuôn mặt trong thời gian thực của video RGB" và đã được xuất bản trong Proc. Thị giác máy tính và Nhận dạng mẫu (CVPR), 2016, IEEE. doi:10.1145/3292039
- [5] Thies, J., Zollhöfer, M., Stamminger, M., Theobalt, C., và Nießner, M. (tháng 1 năm 2019). Hợp tác Mở trong Thời đại Không tin tưởng, Truyền thông của ACM, Tập. 62, số 1, trang 96-104. doi:10.1145/3292039
- [6] Vaccari, C., và Chadwick, A. DF và thông tin sai lệch: Khám phá tác động của video chính trị tổng hợp đối với sự lừa dối, sự không chắc chắn và niềm tin vào tin tức, Tập: 6, Số phát hành: 1, <https://doi.org/10.1177/2056305120903408>
- [7] www.tcs.com/content/dam/tcs/pdf/discover-tcs/Research-and-Innovation/Deepfakes_Envisioning-Prospects-and-Perils.pdf (Truy cập vào ngày 17 tháng 8 năm 2021).
- [8] <https://restofworld.org/2022/china-steps-up-efforts-to-ban-deepfakes/> (Truy cập ngày 17 tháng 8 năm 2021).

- [9] <https://interculturaltalk.com/2019/11/05/cool-but-scaryDFs-are-here/> (Truy cập ngày 17 tháng 8 năm 2021).
- [10] <https://medium.com/@songda/a-short-history-of-DFs-604ac7be6016> (Truy cập ngày 17 tháng 8 năm 2021).
- [11] www.Discovermagazine.com/technology/DFs-the-dark-origins-of-fake-videos-and-their-potential-to-wreak-havoc (Truy cập vào ngày 17 tháng 8 năm 2021).
- [12] <https://medium.com/predict/why-deepfakes-will-make-you-play-video-games-instead-of-films-99ee5c2d7c9e> (Truy cập ngày 17 tháng 8 năm 2021).
- [13] www.technologyreview.com/2020/12/24/1015380/best-ai-DFs-of-2020/ (Truy cập ngày 17 tháng 8 năm 2021).
- [14] www.thinkautomation.com/bots-and-ai/ai-in-healthcare-how-artificial-intelligence-can-help-us-fight- (Truy cập ngày 20 tháng 8 năm 2021).
- [15] Sharma, DK, Gaur, L., và Okunbor, D. (2007). "Nén hình ảnh và trích xuất đặc trưng với mạng thần kinh," Kỷ yếu của Học viện Khoa học Quản lý và Thông tin, Tập. 11, Số 1, trang 33-38.
- [16] Anshu, K., Gaur, L., và Khazanchi, D. (2017). "Đánh giá mức độ hài lòng của các nhà bán lẻ điện tử tạp hóa bằng cách sử dụng mô hình TOPSIS và ECCSI mờ trực quan," Hội nghị quốc tế về Công nghệ Infocom và Hệ thống không người lái (Xu hướng và định hướng tương lai) (ICTUS), trang 276-284, doi:10.1109/ICTUS.2017.8286019
- [17] Gaur, L., và Anshu, K. (2018). Phân tích sở thích của người tiêu dùng cho các trang web sử dụng e TailQ và AHP. Tạp chí Kỹ thuật & Công nghệ Quốc tế, Vol. 7, số 2.11, trang 14-20.
- [18] Rana, J., Gaur, L., Singh, G., Awan, U., và Rasheed, MI (2021). Củng cố hành trình của khách hàng thông qua trí tuệ nhân tạo: Chương trình nghiên cứu và đánh giá. Tạp chí quốc tế về các thị trường mới nổi, Tập. trước khi in (No. trước khi in). <https://doi.org/10.1108/IJOEM-08-2021-1214>
- [19] Gaur, L., Singh, G., Solanki, A., Jhanjhi, NZ, Bhatia, U., Sharma, S., and Kim, W. (2021), Xu hướng của giới trẻ trong việc dự đoán các mục tiêu phát triển bền vững sử dụng các thuật toán rừng ngẫu nhiên và mờ thần kinh. Khoa học thông tin và máy tính lấy con người làm trung tâm, 11, NA.
- [20] Singh, G., Kumar, B., Gaur, L. và Tyagi, A. (2019). "So sánh giữa Multinomial và Bernoulli Naïve Bayes để phân loại văn bản," Hội nghị quốc tế về tự động hóa, tính toán và quản lý công nghệ (ICACTM), trang 593-596. doi:10.1109/ICACTM.2019.8776800
- [21] Gaur, L., Agarwal, V., và Anshu, K. (2020). "Phương pháp tiếp cận DEMATEL mờ để xác định các yếu tố ảnh hưởng đến hiệu quả của ngành bán lẻ Ấn Độ," Đảm bảo hệ thống chiến lược và Phân tích kinh doanh. Phân tích nội dung (Quản lý hiệu suất và an toàn). Springer, Singapore. https://doi.org/10.1007/978-981-15-3647-2_
- [22] Gaur, L., Afaq, A., Singh, G., và Dwivedi, YK (2021). Vai trò của trí tuệ nhân tạo và người máy trong việc thúc đẩy du lịch không chạm trong thời kỳ đại dịch: Chương trình nghiên cứu và đánh giá. Tạp chí Quốc tế về Quản lý Khách sạn Đương đại, Vol. 33, Số 11, trang 4079-4098. <https://doi.org/10.1108/IJCHM-11-2020-1246>
- [23] Sharma, S., Singh, G., Gaur, L., và Sharma, R. (2022). Khoảng cách tâm lý và tôn giáo có ảnh hưởng đến hành vi lừa đảo của khách hàng không? Tạp chí Nghiên cứu Người tiêu dùng Quốc tế. doi:10.1111/ijcs.12773
- [24] Sahu, G., Gaur, L., và Singh, G. (2021). Áp dụng phương pháp tiếp cận lý thuyết thích hợp và hài lòng để kiểm tra niềm đam mê của người dùng đối với các nền tảng vượt trội và truyền hình thông thường. Viễn thông và Tin học, 65. doi:10.1016/j.tele.2021.101713

- [25] Ramakrishnan, R., Gaur, L., và Singh, G. (2016). Tính khả thi và hiệu quả của các thiết bị IoT đèn hiệu BLE trong quản lý hàng tồn kho tại khu vực cửa hàng. *Tạp chí Quốc tế về Kỹ thuật Điện và Máy tính*, 6(5), 2362-2368. doi:10.11591/ijcece.v6i5.10807
- [26] Afaq, A., Gaur, L., Singh, G., và Dhir, A. (2021). COVID-19: Thay đổi hành vi của hành khách đi máy bay và định hình lại kỳ vọng của họ đối với ngành hàng không. *Nghiên cứu giải trí du lịch*. doi:10.1080/02508281.2021.2008211



Taylor & Francis
Taylor & Francis Group
<http://taylorandfrancis.com>

Mục lục

A

Tấn công đối thủ, 58, 65, 67, 69
 Sự nhiễu loạn của đối thủ, 60-61
 Bot cảnh báo, 141
 Bộ mã hóa tự động, 4, 24-31
 Tự động nhận dạng tin giả, 74-75

B

Khớp nối thư mục, 9, 15, 17
 Nghiên cứu thư mục, 11
 Tín hiệu sinh học, 81

C

Trích dẫn, 11, 13-15
 Ngôn ngữ học tính toán, 128
 Thị giác máy tính, 2, 3, 16, 19, 24, 57-59, 75
 Hàng Giả, 77, 115
 Biện pháp đổi phó, 99, 102, 108, 110, 111
 Uy tín, 100, 103, 122
 Mạng thần kinh tích chập, 16, 24, 41, 42, 60, 81

D.

Làm sạch dữ liệu, 127
 Quyền bảo vệ dữ liệu, 141
 Cây quyết định, 121, 123, 124, 126
 Bộ giải mã, 4, 5, 23, 25, 36, 78
 DeepFakes, 4, 5, 9-142
 Học sâu, 1-26, 36, 38, 41, 59, 73, 125, 128
 Mạng lưới thần kinh sâu, 16, 58, 99
 Kiến thức kỹ thuật số, 101
 Phương tiện kỹ thuật số, 100, 106, 135
 Ý định bắt lương, 122
 Thông tin sai lệch, 15, 16, 100-110, 135

E

Ectypes, 136, 137
 Bầu cử, 5, 99, 101, 102, 105

F

Hoán đổi khuôn mặt, 23, 24, 27-31, 93
 Dữ liệu thật và giả, 130
 Tin giả, 16, 41, 121-132
 Thông tin sai sự thật, 124, 125
 Phương pháp dấu gradient nhanh, 59

G

Mạng đối thủ chung, 36
 Hệ thống đối kháng sáng tạo, 115
 Phân loại tăng cường độ dốc, 124
 Độ dốc của hàm mất mát, 64

-

Dịch ảnh, 28, 30, 58

M

Học máy, 22, 123, 126, 127
 Xác thực phương tiện, 109
 Kiến thức truyền thông, 110, 111, 139
 Truyền thông xuất xứ, 109

N

Xử lý ngôn ngữ tự nhiên, 125-131
 Mạng lưới thần kinh, 16, 19, 23, 26, 37, 41-69

P

Nhiều loạn, 57-61, 64-68
 Hình ảnh nhiễu loạn, 59, 61, 62
 Các vấn đề về quyền riêng tư, 138
 Các cuộc tấn công giảm độ dốc dự kiến, 59

I

Các mô hình tích chập hồi quy, 109

S

Mạng xã hội, 1, 2, 118, 121-142
 Hội, 2, 5, 10, 18, 49, 91, 99, 103, 107
 Phương tiện tổng hợp, 31, 76, 91, 93, 136

T

Các mối đe dọa, 99, 100, 102-105
 Token hóa, 127, 128

V

Công cụ xác thực video, 109
 Hội nghị truyền hình, 119
 Thực tế ảo, 116
 Thế giới ảo, 103
 Dập tắt tăng cường thị giác, 117
 Tự quan hóa dữ liệu, 129
 Trình xem VOS, 12



Taylor & Francis
Taylor & Francis Group
<http://taylorandfrancis.com>