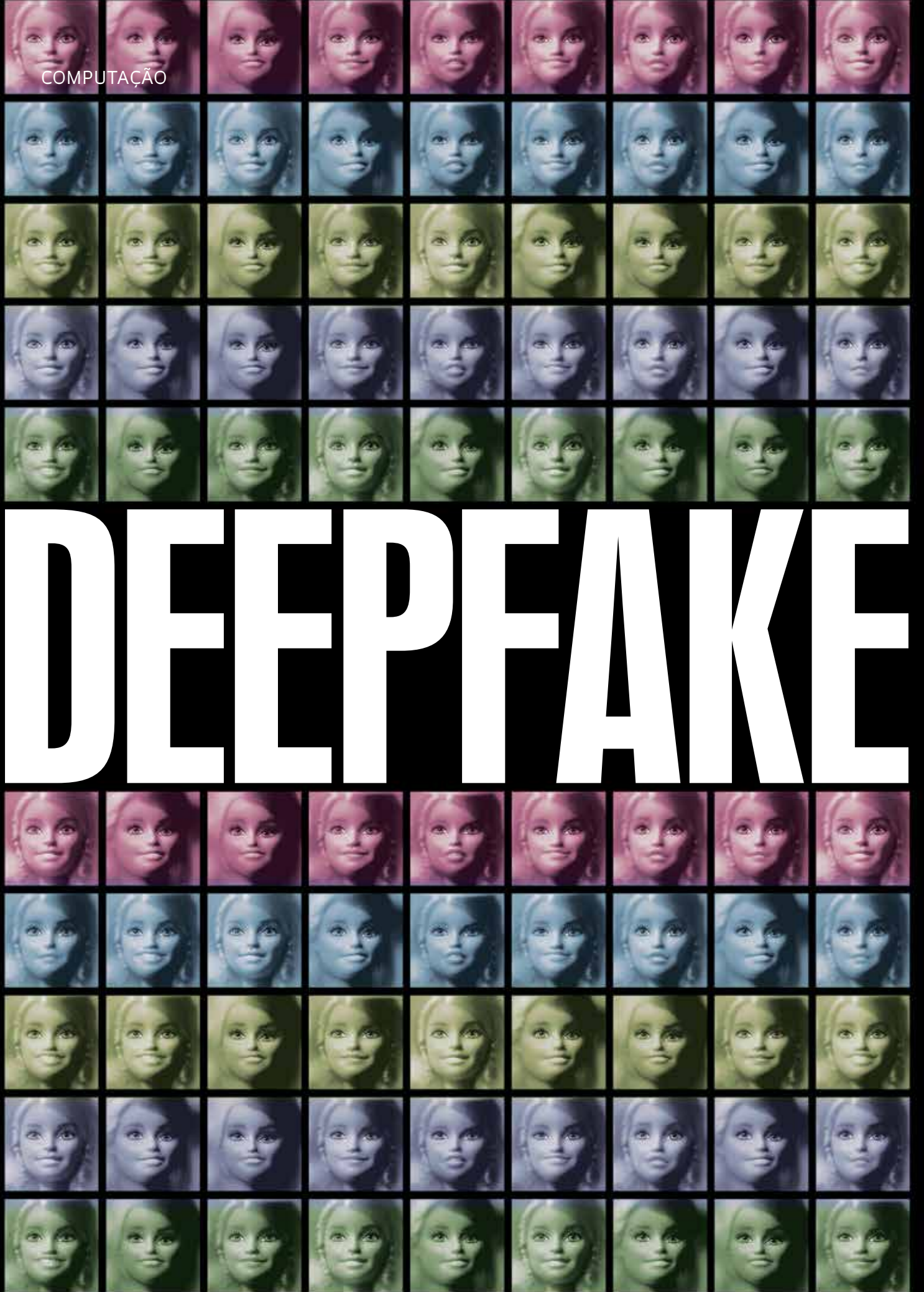
COMPUTAÇÃO



DEEPFAKE

Algoritmo detecta imagens e vídeos alterados com inteligência artificial, o novo estágio tecnológico para disseminação de desinformação

Sarah Schmidt

Em setembro, um vídeo adulterado do *Jornal Nacional*, principal noticiário da rede Globo de televisão, ganhou as redes sociais. Nele, os apresentadores William Bonner e Renata Vasconcellos mostravam os resultados de uma pesquisa de intenção de votos para a Presidência, mas os dados estavam invertidos sobre quem era o candidato favorito, tanto nos gráficos quanto nas falas dos apresentadores. No dia seguinte, o telejornal fez um esclarecimento alertando que o vídeo estava sendo usado para desinformar a população e afirmando que se tratava de deepfake, técnica que usa inteligência artificial para fazer edições profundas no conteúdo. Com ela, é possível, por exemplo, trocar digitalmente o rosto de uma pessoa ou simular sua voz, fazendo com que ela faça o que não fez ou diga o que não disse.

Em agosto, outro vídeo do telejornal com edição semelhante, que também invertia os resultados de uma pesquisa para a Presidência, foi postado na rede social de vídeos TikTok, onde alcançou 2,5 milhões de visualizações, segundo o Projeto Comprova, iniciativa que reúne jornalistas de 43 veículos de comunicação do país para checar desinformação.

"Pode ser que tenha sido usada alguma técnica de deepfake nesses vídeos, mas é preciso uma análise mais detalhada. Para nós, o importante é saber que são falsos", observa o cientista da computação Anderson Rocha, diretor do Instituto de Computação da Universidade Estadual de Campinas (Unicamp), onde coordena o Laboratório de Inteligência Artificial (Recod.ai). O pesquisador tem estudado maneiras de detectar adulterações maliciosas em fotos e vídeos, inclusive em deepfakes, também chamadas de mídia sintética.

Ainda em março deste ano, logo após o início da guerra entre a Rússia e a Ucrânia, o presidente ucraniano, Volodymyr Zelensky, foi vítima de deepfake. Um vídeo em que ele parecia pedir aos ucranianos que largassem as armas e voltassem para suas casas, como se o país estivesse se rendendo, circulou nas redes sociais, obrigando o Facebook e o YouTube a removê-lo assim que se constatou que era falso. Nas imagens, o rosto do presidente aparecia em um corpo que quase não se mexia, vestido com uma camiseta verde.

Em alguns casos, como nos vídeos do *Jornal Nacional*, não é tão difícil perceber que foram alterados de alguma forma, porque as notícias originais estão facilmente disponíveis para verificação. Mas nem sempre é o caso. Diante das mídias sintéticas, o ditado "ver para crer" vai perdendo sentido, e a própria inteligência artificial pode ser uma aliada.

"Geralmente os vídeos sintéticos são feitos em duas etapas: primeiro, com o uso de uma plataforma de deepfake, para trocar os rostos ou fazer a sincronia da boca, e depois é feita uma edição em programas editores de vídeo", explica Rocha. Quem sabe o que procurar costuma detectar alguma falha do programa usado para produzir a farsa, como um jogo de luzes diferentes, um contraste entre o vídeo original e a nova face que foi inserida.

É como recortar um rosto de uma foto e colocar sobre outra: a incidência de luz e a forma como

ILUSTRAÇÃO FELIPE BRAZ COM FOTO DE ALEXA/ PIXABAY

a câmera captou as duas imagens são diferentes. Esses vestígios são pistas que ficam pelo caminho, identificadas pelas técnicas de computação forense, área de pesquisa que tem crescido nos últimos anos e da qual Rocha faz parte.

Com colegas da Universidade de Hong Kong, o pesquisador desenvolveu um algoritmo que ajuda a detectar, de forma simultânea nos vídeos, se houve manipulação de rostos e, em caso positivo, a localizar quais regiões foram mudadas. Pode, por exemplo, ter sido a face inteira ou apenas a boca, a região dos olhos ou o cabelo. "A média de acertos foi de 88% para vídeos de baixa resolução e de 95% para vídeos com resolução maior", explica Rocha, sobre um universo de 112 mil faces testadas: metade verdadeira, metade manipulada e gerada por quatro programas de deepfake. O método também indica se a imagem foi criada do zero, ou seja, sem ter como base uma fotografia existente. Os resultados foram publicados em abril de 2022 na revista *Transactions on Information Forensics and Security*.

Segundo o cientista da computação, outros algoritmos desenvolvidos conseguem detectar traços de alteração nos vídeos de deepfake, mas, em sua maioria, trabalham com base nas pistas deixadas por programas de manipulação mais conhecidos, basicamente divididos em duas categorias: os que permitem a troca dos rostos e aqueles que possibilitam a edição das expressões faciais. Um desses softwares é conhecido por deixar sempre alguma imperfeição na sincronização da boca – o algoritmo detector, então, é programado para buscar esse erro específico. "Há um problema nisso: se não conhecermos o software de deepfake, fica mais difícil identificar esses traços. E sempre surgem aplicativos novos", observa Rocha.

Por isso, ele e seus colegas treinaram o algoritmo que desenvolveram para que detectasse pistas sem pressupor conhecimento do aplicativo gerador de deepfake. "Trabalhamos com a ideia de que, independentemente do programa, ele vai deixar um ruído, algo que não é coerente com o resto da imagem." O método atua em duas frentes: procura por assinaturas de ruído, ou seja, mudanças sutis na borda do rosto, por exemplo, e mapeia a chamada assinatura semântica, que pode ser uma falha de cor, de textura ou de forma.

"O algoritmo automatiza o processo que um especialista humano faz, que é procurar incoerências, como os contrastes de luz", diz ele. "O próximo passo é testá-lo com vídeos falsos gerados por um número maior de programas, para confirmar esse potencial."

Esse tipo de algoritmo detector pode ser usado para diversos fins que envolvam o combate ao uso mal-intencionado de deepfakes. Rocha integra um projeto internacional, chamado Semantic Forensics, ao lado de pesquisadores das universidades de Siena e Politécnica de Milão, na Itália, e de Notre Dame, nos Estados Unidos, que conta com o apoio do Departamento de Defesa dos Estados Unidos. O objetivo é desenvolver ferramentas automatizadas que detectem essas manipulações. "Já vimos casos de vídeos alterados de exercícios militares de outros países, que multiplicam o número de mísseis para mostrar um poder bélico maior", conta ele.

Esses algoritmos também podem ajudar nos casos de deepfakes políticas, como no episódio do presidente ucraniano, ou mesmo os pornográficos. Foi usando os filmes de sexo que a técnica ganhou fama, no final de 2017. Na época, alguns usuários da internet passaram a inserir o rosto de celebridades do cinema em cenas de filmes com conteúdo sexual. Em setembro de 2019, segundo um levantamento do DeepTrace Labs, uma companhia holandesa de cibersegurança, 96% dos vídeos de deepfake mapeadas na rede eram de pornografia não consensual. As principais vítimas eram mulheres, sobretudo atrizes, mas havia

Em vídeo falsificado, o presidente ucraniano, Volodymyr Zelensky, pedia que seus compatriotas largassem as armas

É POSSÍVEL TROCAR O ROSTO DE UMA PESSOA OU SIMULAR SUA VOZ, FAZENDO COM QUE ELA DIGA O QUE NÃO DISSE

também registros de casos com pessoas que não eram famosas. Em julho deste ano, a cantora Anitta também foi vítima de uma deepfake pornô. O vídeo original já havia sido usado para produzir mídias falsas com o rosto da atriz Angelina Jolie.

Segundo a jornalista Cristina Tardáguila, diretora de programas do Centro Internacional para Jornalistas (ICFJ) e fundadora da Agência Lupa, especializada em checagem de fatos, o Brasil já tem lidado com deepfakes que precisam ser desmentidas. Por isso, programas que ajudem a detectar mídia sintética podem ser aliados dos jornalistas e checadores de fatos, que trabalham contra o tempo. "Ao lidar com desinformação, é preciso ser rápido. Por isso, é importante investir em inteligência artificial, em ferramentas que possam ajudar a detectar e mapear esse tipo de conteúdo falso de forma mais veloz. Assim, conseguimos encurtar o tempo entre a propagação do conteúdo falso e a entrega da checagem", avalia.

"As deepfakes são o auge das fake news. Elas têm o potencial de enganar mais facilmente, porque, se é um vídeo, a pessoa está vendo aquela cena", observa a jornalista Magaly Prado, que faz estágio de pós-doutorado no Instituto de Estudos Avançados da Universidade de São Paulo (IEA-USP). "O áudio, inclusive, também pode ser gerado de forma sintética", diz ela, autora do livro *Fake news e inteligência artificial: O poder dos algoritmos na guerra da desinformação*, lançado em julho pela Edições 70.

Ela avalia que, apesar de serem menos lembradas e menos comuns, as deepfakes exclusivamente em formato de áudio têm potencial para se espalhar por plataformas como o WhatsApp, mídia muito usada pelos brasileiros. Eles seguem lógica parecida com a dos vídeos: com aplicativos acessíveis que ficam cada vez melhores, é possível simular a voz de alguém. As vítimas mais fáceis são figuras públicas, cuja voz está amplamente disponível na rede. A técnica pode ser usada também para golpes financeiros. "Já houve casos como o de um funcionário de empresa de tecnologia que recebeu mensagem de voz de um alto executivo pedindo uma transferência em dinheiro. Ele desconfiou, o caso foi analisado por uma empresa de segurança e verificou-se que era uma mensagem construída com inteligência artificial", conta.

O jornalista Bruno Sartori, diretor da empresa FaceFactory, explica que produzir deepfakes bem-feitas, tanto de áudio quanto de vídeo, não é tão simples – ainda. É o que ele faz, profissionalmente: sua empresa cria mídia sintética para uso comercial e fornece conteúdo para programas de humor dos canais de televisão Globo e SBT.

Em 2021, ele trabalhou em um comercial para a Samsung em que a apresentadora Maísa, já adulta, interagia com sua versão criança. Esta última, criada com técnicas de deepfake. A menininha virtual dança, brinca e joga um notebook para cima. Em outra ocasião, ele precisou inserir o rosto de um ator em um dublê. "Para treinar bem a inteligência artificial, é importante ter um bom banco de imagens e de áudio da pessoa que se quer imitar. Os bons programas que fazem processamento de alta qualidade também precisam ter configurações avançadas. Caso contrário, pode haver falhas visíveis no rosto ou, no caso do áudio, uma voz robotizada", explica.

Em sua avaliação, os vídeos manipulados do *Jornal Nacional* que trocam os dados da pesquisa não chegaram a ser alterados com uso de inteligência artificial. "Na minha análise, eles passaram por uma edição comum, com o corte e a inversão da ordem dos áudios. São o que chamamos de *shallowfake*. Mas, como está bem-feita, o potencial de enganar as pessoas é o mesmo", avalia Sartori. Para ele, daqui a poucos anos esses programas estarão mais leves, mais inteligentes e mais acessíveis.

Há alguns caminhos para se proteger da desinformação criada com o auxílio da tecnologia. Um deles é ficar atento às licenças de uso e privacidade dos mais diversos aplicativos gratuitos usados no dia a dia – desde aqueles que pedem acesso às fotos do usuário para gerarem efeitos divertidos, passando pelos que podem armazenar a voz. Segundo Rocha, da Unicamp, muitos deles guardam uma quantidade grande de dados que podem ser compartilhados para outros fins, como treinar programas de deepfake.

Outro ponto importante é a educação midiática. "Por mais que existam softwares que nos ajudem a apontar o que é falso, o primeiro passo é desconfiar daquilo que se recebe nas redes sociais. E conferir as fontes de informação, pesquisar sobre elas", conclui. ■

Projeto

Déjà vu: Coerência temporal, espacial e de caracterização de dados heterogêneos para análise e interpretação de integridade (nº 17/12646-3); **Modalidade** Projeto Temático; **Pesquisador responsável** Anderson Rocha; **Investimento** R$ 1.912.168,25.

Artigo científico

KONG, C. *et al.* Detect and locate: Exposing face manipulation by semantic- and noise-level telltales. **Transactions on Information Forensics and Security**. v. 17. abr. 2022.

Livro

PRADO, M. **Fake news e inteligência artificial: O poder dos algoritmos na guerra da desinformação**. São Paulo: Edições 70, 2022.

REPRODUÇÃO