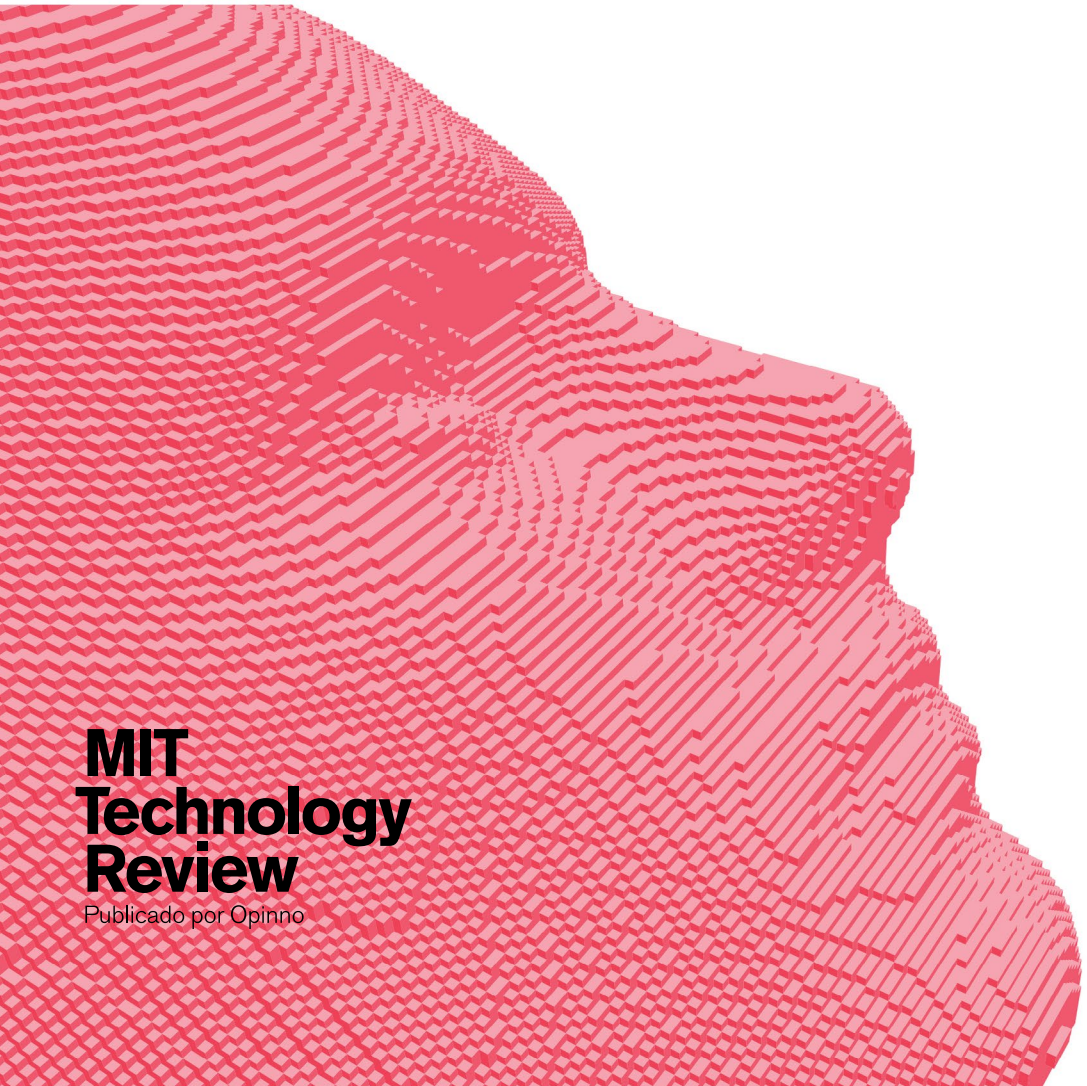
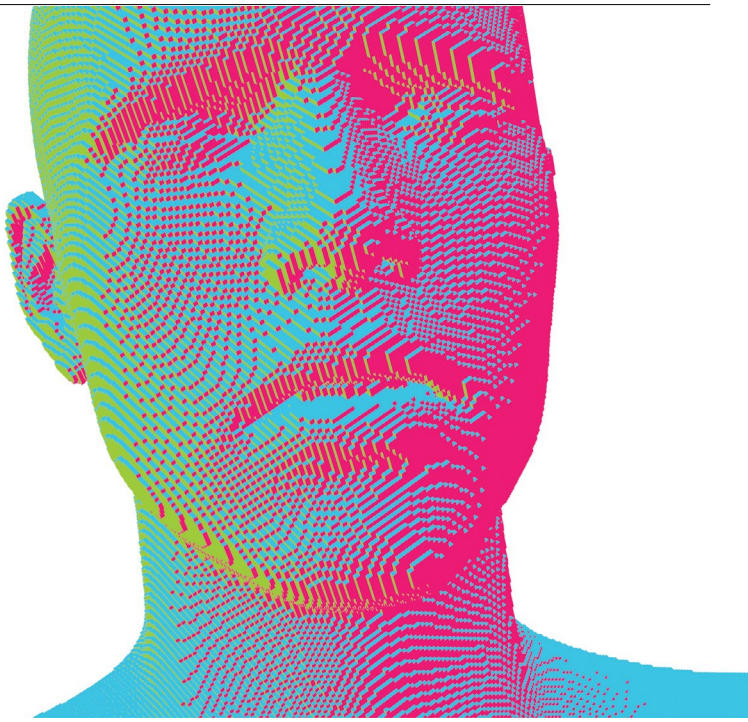
Cómo **hackear una GAN** para **destapar** las identidades de su **entrenamiento**

Dos técnicas diferentes permiten acceder a los datos originales, como las caras, que se usan para entrenar redes generativas antagónicas capaces de crear imágenes ultrarrealistas, pero falsas, conocidas como *deepfakes*. Ambas investigaciones son una alerta más de los riesgos de privacidad de este tipo de inteligencia artificial.



MIT Technology Review
Publicado por Opinno



WILL DOUGLAS HEAVEN
TRADUCIDO POR **ANA MILUTINOVIC**
15 OCTUBRE, 2021

Si entramos en la página web This Person Does Not Exist veremos un rostro humano, casi perfecto en su realismo, pero totalmente falso. Si volvemos a cargar la página, la red neuronal subyacente generará otro rostro similar, y otro, y otro más. La secuencia interminable de las caras creadas por inteligencia artificial (IA) se produce mediante la red generativa antagónica (GAN, sus siglas en inglés de *generative adversarial network),* un tipo de IA que aprende a generar ejemplos realistas, pero falsos, de los datos con los que se entrena.

Pero esos rostros creados, que han empezado a usarse en películas y anuncios generados por ordenador o CGI (del inglés *Computer Generated Imagery),* podrían no ser tan únicos como parecen. En un artículo titulado *Esta persona (probablemente) existe,* los autores muestran que muchas caras producidas por GAN tienen un parecido sorprendente con personas reales que aparecen en los datos de entrenamiento.

Los rostros falsos pueden desenmascarar los individuos reales en los que se entrenó la GAN, lo que descubre la identidad de esas personas. Este trabajo es el último de una serie de estudios que ponen en duda la popular idea de que las redes neuronales son «cajas negras» que no revelan nada sobre lo que ocurre en su interior.

Para destapar los datos de entrenamiento ocultos, el investigador de la Universidad de Caen (Francia) Ryan Webster y sus colegas utilizaron una técnica conocida como inferencia de membresía, que puede averiguar si ciertos datos se usaron para entrenar un modelo de red neuronal. En general, estos ataques aprovechan las sutiles diferencias entre la forma en la que un modelo trata los datos en los que se entrenó (y, por eso, los ha visto miles de veces antes) y los datos nunca vistos.

Por ejemplo, un modelo puede identificar con precisión una imagen que no había visto antes, pero la confianza es algo menor que la de una con la que se entrenó. Un segundo modelo de ataque puede aprender a detectar tales señales en el comportamiento del primer modelo y usarlos para predecir cuándo ciertos datos, como una foto, están en el conjunto de entrenamiento o no.

Estos ataques pueden provocar graves problemas de seguridad. Por ejemplo, descubrir que los datos médicos de alguien se utilizaron para entrenar un modelo relacionado con una enfermedad podría revelar que esta persona tiene esa enfermedad.

El equipo de Webster extendió esta idea de tal modo que, en vez de identificar las fotos exactas utilizadas para entrenar una GAN, detectaron las fotos del conjunto de entrenamiento de las GAN que no eran idénticas, pero parecían retratar a la misma persona. En otras palabras, rostros creados con la misma identidad. Para llevarlo a cabo, los investigadores primero generaron caras con la GAN y luego utilizaron una IA de reconocimiento facial aparte para averiguar si la identidad coincidía con la identidad de alguno de los rostros de los datos de entrenamiento.

Los resultados son sorprendentes. En muchos casos, el equipo encontró varias fotos de personas reales en los datos de entrenamiento que parecían coincidir con las caras falsas generadas por la GAN, revelando la identidad de las personas en las que se había entrenado la IA.

la secuencia interminable de las caras creadas por inteligencia artificial (IA) se produce mediante la red generativa antagónica (GAN, sus siglas en inglés de *generative adversarial network).*

los rostros falsos pueden desenmascarar los individuos reales en los que se entrenó la GAN, lo que descubre la identidad de esas personas.

La columna de la izquierda en cada bloque muestra caras generadas por la GAN. Estas caras falsas van seguidas de tres fotos de personas reales identificadas en los datos de entrenamiento.
Créditos: Universidad de Caen Normandy

El trabajo plantea serias preocupaciones sobre privacidad. «La comunidad de IA tiene una falsa sensación de seguridad cuando comparte modelos de redes neuronales profundas entrenadas», opina el vicepresidente de investigación de aprendizaje y percepción de Nvidia, Jan Kautz.

En teoría, este tipo de ataque se podría aplicar a otros datos vinculados a una persona, como biométricos y médicos. Por otro lado, Webster señala que las personas también podrían usar esta técnica para verificar si sus datos se han utilizado para entrenar a una IA sin su consentimiento.

Los artistas pueden averiguar si su trabajo se ha utilizado para entrenar a una GAN en una herramienta comercial. Webster detalla: «Un método como el nuestro se podría usar para buscar evidencias de infracción de derechos de autor».

El proceso también se podría utilizar para asegurarse de que las GAN no expongan los datos privados en primer lugar. Una GAN puede comprobar si sus creaciones se parecen a los ejemplos reales en sus datos de entrenamiento, antes de publicarlas, con la misma técnica desarrollada por los investigadores.

Pero, Kautz explica que esto supone que esos datos de entrenamiento se pueden conseguir. Él y sus colegas de Nvidia han ideado una forma diferente de revelar los datos privados, incluidas las imágenes de rostros y otros objetos, datos médicos, etcétera, que no requiere en absoluto el acceso a los datos de entrenamiento.

Desarrollaron un algoritmo que puede recrear los datos a los que ha estado expuesto un modelo entrenado al invertir los pasos por los que pasa el modelo al procesar dichos datos. Por ejemplo, una red de reconocimiento de imágenes: para identificar qué hay en una imagen, la red la pasa por una serie de capas de neuronas artificiales. Cada capa extrae diferentes niveles de información, desde los bordes hasta las formas y otras características más reconocibles.

El equipo de Kautz descubrió que podían interrumpir un modelo en medio de estos pasos e invertir su dirección, recreando la imagen de entrada o *input* a partir de los datos internos del modelo. Probaron su técnica en una variedad de modelos comunes de reconocimiento de imágenes y en otras GAN. En una prueba,

Imágenes de ImageNet (arriba) junto con las recreaciones de esas imágenes realizadas al retroceder un modelo entrenado en ImageNet (abajo). **Créditos:** Nvidia

demostraron que podían recrear con precisión las imágenes de ImageNet, uno de los más conocidos conjuntos de datos de reconocimiento de imágenes.

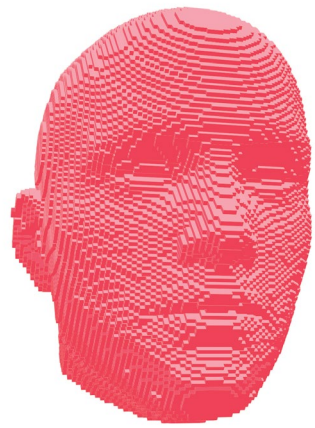
Al igual que en el trabajo de Webster, las imágenes recreadas se parecen mucho a las reales. Kautz admite: «Nos sorprendió la calidad final».

Los investigadores argumentan que este tipo de ataque no es solo hipotético. Los teléfonos inteligentes y otros dispositivos pequeños empiezan a usar más IA. Debido a las limitaciones de la batería y la memoria, a veces los modelos solo se procesan a medias en el dispositivo y se envían a la nube para el procesamiento final, un enfoque conocido como computación distribuida. La mayoría de los investigadores asumen que la computación distribuida no revelará ningún dato privado del teléfono de una persona porque solo se comparte el modelo, dice Kautz. Pero su ataque demuestra que no es así.

Kautz y sus colegas están trabajando para evitar que los modelos filtren datos privados, y explica que deben comprender los riesgos para minimizar las vulnerabilidades.

Aunque utilizan técnicas muy diferentes, Kautz cree que su trabajo y el de Webster se complementan bien. El equipo de Webster mostró que se pueden encontrar datos privados en el *output* de un modelo. El grupo de Kautz descubrió que los datos privados se pueden revelar dando marcha atrás, recreando así el *input*. «Explorar ambas direcciones es importante para comprender mejor cómo prevenir los ataques», resalta Kautz. **</>**

una GAN puede comprobar si sus creaciones se parecen a los ejemplos reales en sus datos de entrenamiento, antes de publicarlas, con la misma técnica desarrollada por los investigadores.



El autor es editor senior de Inteligencia Artificial en *MIT Technology Review*.

El artículo original «Cómo hackear una GAN para destapar las identidades de su entrenamiento» pertenece a la edición digital de *MIT Technology Review*.