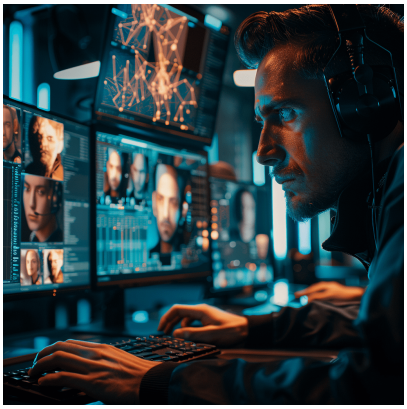

Deepfake Hunters. Cazadores de mentiras

Descripción

Introducción

En un mundo cada vez más digitalizado, la distinción entre realidad y ficción se vuelve borrosa.

Los avances recientes en inteligencia artificial han llevado al surgimiento de la tecnología deepfake, capaz de generar imágenes, videos y audios realistas que son difíciles de distinguir de los auténticos.

La tecnología de deepfake ha emergido como una herramienta de doble filo, capaz de crear desde divertidas imitaciones hasta peligrosas falsificaciones.

El poder de los deepfakes plantea serios desafíos éticos y prácticos. Frente a esta amenaza, la inteligencia artificial (IA) se perfila no solo como parte del problema, sino también como la solución principal. Sin embargo, ¿estamos realmente preparados para enfrentar los desafíos que los deepfakes representan con las herramientas actuales de IA?

Qué es deepfake

Un deepfake es un tipo de contenido multimedia generado artificialmente que utiliza técnicas avanzadas de inteligencia artificial y aprendizaje automático para crear imágenes, audios o videos que parecen auténticos.

La palabra «deepfake» combina «deep learning» (aprendizaje profundo) con «fake» (falso), y refleja el uso de redes neuronales profundas para manipular o generar contenido digital realista.

Los deepfakes se crean principalmente mediante dos técnicas de aprendizaje profundo: las Redes Generativas Antagónicas (GANs) y los Autoencoders.

Las GAns consisten en dos redes neuronales, un generador y un discriminador, que se entrenan en conjunto. El generador aprende a crear imágenes o videos falsos que imitan el estilo y las características de un conjunto de datos de entrenamiento, mientras que el discriminador aprende a diferenciar entre las creaciones del generador y los ejemplos reales. A través de este proceso competitivo, el generador mejora progresivamente su capacidad para producir contenidos cada vez más convincentes.

Los autoencoders son redes neuronales que aprenden a codificar y decodificar un conjunto de datos. En el contexto de los deepfakes, se utilizan dos autoencoders: uno para aprender a reconstruir el rostro de la persona objetivo y otro para la persona que se quiere simular. Al combinar la codificación de un rostro con la decodificación del otro, se puede transponer la expresión facial de una persona a otra en un video, creando así una ilusión convincente de que la persona objetivo está diciendo o haciendo algo que nunca ocurrió.

El proceso de creación de un deepfake de alta calidad se lleva a cabo en varias etapas.

El primer paso es recopilar una cantidad suficiente de datos de entrenamiento. Esto generalmente significa obtener muchas imágenes o videos del sujeto que se quiere falsificar y del sujeto que se quiere imitar. Por ejemplo, si se desea crear un video falso de una celebridad, se necesitarán horas de material de video de esa persona hablando y expresándose para capturar una amplia gama de expresiones faciales, gestos y matices vocales.

Los datos recopilados deben ser procesados para garantizar que sean utilizables en el entrenamiento de la red. Esto incluye tareas como localizar y extraer rostros de videos, ajustar los rostros para que tengan una orientación y tamaño consistentes y ajustar el color y el contraste de las imágenes para que las condiciones de iluminación no afecten el aprendizaje.

Mediante el uso de una GAN compuesta por dos redes neuronales que compiten entre sí, un generador y un discriminador, se lleva a cabo el entrenamiento.

El generador Intenta crear imágenes o videos falsos que parezcan reales mientras que el discriminado evalúa si las salidas del generador son reales o falsas y proporciona retroalimentación al generador.

En el contexto de los deepfakes, el generador aprenderá gradualmente a crear imágenes o secuencias de video que imiten a la persona objetivo, mientras que el discriminador se vuelve mejor para detectar diferencias entre los datos reales y los generados.

El entrenamiento complejo y refinamiento de la GAN a través de un proceso iterativo adelante (forward pass), en el que el generador produce una salida, evaluación, en el que el discriminador evalúa esta salida y retroalimentación (backpropagation), basado en la evaluación del discriminador, el generador ajusta sus parámetros para mejorar su capacidad de imitar los datos reales.

Este proceso se repite miles o incluso millones de veces, con el generador mejorando continuamente su capacidad para crear falsificaciones convincentes a medida que el discriminador se vuelve más experto en detectar las sutilezas que diferencian lo real de lo falso.

Una vez que el modelo está suficientemente entrenado, se puede usar para generar deepfakes. Esto implica alimentar al generador con nuevos datos de entrada (por ejemplo, un video de una persona diferente) y utilizar el modelo para transferir las expresiones y características del sujeto objetivo a esta nueva entrada.

El último paso es refinar los videos o imágenes generados para mejorar su realismo. Esto puede incluir ajustar la sincronización del labio si el video incluye habla, mejorar la resolución de las imágenes mediante técnicas de super-resolución, o ajustar la iluminación y el color para que coincidan con los entornos naturales.

El contenido deepfake generado por IA es indudablemente impresionante. Pero también plantea riesgos significativos.

El realismo de los deepfakes puede ser utilizado para fines de entretenimiento, formativos, informativos o divulgativos, pero también para crear desinformación a una escala masiva, afectando desde la política hasta la vida personal de individuos. El potencial de los deepfakes para influir en elecciones, desestabilizar economías y difundir noticias falsas es alarmante.

A la caza de deepfakes

Los primeros deepfakes eran relativamente fáciles de detectar con una simple exploración visual o con el uso de herramientas forenses muy simples. Cualquiera que prestara un poco de atención o que tuviera unas nociones muy básicas en el uso de herramientas de análisis digital podría detectar estas producciones artificiales.

Sin embargo, a medida que la tecnología de deepfakes evoluciona, también lo hace la necesidad de estrategias avanzadas y multidisciplinarias para combatirlas eficazmente.

Las direcciones futuras en este campo no solo abarcan el desarrollo tecnológico sino también aspectos regulatorios, educativos y de cooperación internacional.

En los últimos años han surgido herramientas basadas en IA para detectar, precisamente, creaciones artificiales desarrolladas por IA. Entre ellas desatacan las Redes Neuronales Convolucionales (CNNs), las Redes Neuronales Recurrentes (RNNs) y las GANs

Las CNNs han demostrado ser especialmente eficaces en identificar irregularidades sutiles que no corresponden a patrones humanos normales, como errores en la textura de la piel o en la sincronización de los labios. Por otro lado, las RNNS analizan secuencias de

cuadros para identificar inconsistencias típicamente dejadas por los algoritmos deepfake. Finamente las GANs, aunque usadas inicialmente para crear deepfakes, se han reconfigurado para detectar anomalías mediante el entrenamiento de redes competitivas.

Pero las redes neuronales son sólo la punta del iceberg en la lucha contra los deepfakes. La adaptabilidad de los deepfakes requiere que estas tecnologías estén en constante evolución. Los desarrolladores deben anticipar y responder rápidamente a las nuevas técnicas de falsificación.

La detección de la «huella digital» de un deepfake se refiere al proceso de identificar características únicas en los archivos de video o imagen que han sido generados o alterados por algoritmos específicos de inteligencia artificial. Estas características pueden ser sutiles artefactos visuales o patrones consistentes que son introducidos inadvertidamente por las herramientas de creación de deepfakes.

Las técnicas forenses digitales para la detección de deepfakes se centran en el análisis de a huela digital mediante técnicas de análisis del activo digital. Estas técnicas aprovechan tanto las limitaciones de los algoritmos de generación de deepfakes como los patrones sutiles que dejan atrás, que no son evidentes para el ojo humano.

Veamos algunos de los métodos forenses más efectivos utilizados en la detección de deepfakes:

1. Análisis de consistencia en imágenes o videos

Los algoritmos de deepfake pueden tener dificultades para mantener la consistencia a través de diferentes cuadros de un video o partes de una imagen.

Las técnicas forenses pueden analizar aspectos como:

- Textura y patrones de piel: los deepfakes a menudo fallan al tratar de replicar con precisión la textura de la piel humana a lo largo de varios cuadros de un video.
- Consistencia de iluminación: analizar si la iluminación y las sombras en diferentes partes de la imagen o entre cuadros consecutivos son físicamente plausibles.
- Geometría facial: verificar la coherencia de las proporciones y características geométricas del rostro, que pueden deformarse o variar anormalmente en los deepfakes.

2. Análisis de frecuencia

El análisis de frecuencia implica descomponer una imagen o video en sus componentes de frecuencia para identificar irregularidades.

- Transformada de Fourier: se utiliza para identificar patrones inusuales en el dominio de la frecuencia que pueden sugerir manipulación.
- Espectros de frecuencia: los deepfakes pueden dejar artefactos específicos en ciertas bandas de frecuencia debido a cómo se procesan los datos durante la generación.

3. Detección de artefactos y anomalías

Los deepfakes suelen crear imperfecciones o artefactos que son invisibles a simple vista pero detectables mediante análisis detallados.

- Compresión de video e imagen: los algoritmos de compresión pueden comportarse de manera diferente en imágenes reales versus manipuladas, dejando patrones distintivos que pueden ser analizados.
- Errores de pixel y color: la inconsistencia en la representación de colores o en la estructura de los píxeles puede indicar una manipulación.
- Esteganografía y análisis de ruido: los patrones de ruido inherentes a las imágenes y videos originales se alteran cuando se crean deepfakes. Detectar estos cambios mediante técnicas esteganográficas puede ayudar a identificar manipulaciones.

4. Biometría, microexpresiones y lenguaje no verbal

Los detalles finos de cómo se mueve un rostro humano real, incluyendo las microexpresiones y la sincronización de movimientos labiales, pueden ser difíciles de replicar perfectamente en los deepfakes.

- Análisis de movimientos faciales: los deepfakes suelen tener dificultades para replicar las microexpresiones humanas, que son breves, involuntarias y revelan emociones verdaderas. Los sistemas forenses pueden comparar los movimientos faciales con bases de datos de movimientos humanos normales para detectar anomalías.
- Sincronización labial: verificar si los movimientos de los labios en los videos coinciden adecuadamente con el audio.
- Análisis de movimiento: la incongruencia en el movimiento natural del cuerpo y las expresiones gestuales puede ser un indicador de manipulación. Los sistemas de IA están siendo entrenados para identificar patrones de movimiento que parezcan antinaturales o que no coincidan con las emociones o sonidos expresados.

- Análisis de la voz y lingüística: tan importante es lo que se dice como el cómo se dice. Los analistas también analizan la frecuencia de la voz en rastro de indicadores de congruencia, no tanto en la veracidad que se pretende dar al mensaje sino en la composición tonal. Así mismo, se analiza la estructura lingüística del mensaje y se compara con piezas reales contrastables. Cabe destacar que el surgimiento de sistemas de clonación de la voz es un riesgo adicional en la generación de contenidos artificiales.

5. Machine learning y redes neuronales

Además de las técnicas tradicionales, como ya hemos visto, se utilizan métodos de aprendizaje automático para entrenar modelos capaces de distinguir entre imágenes genuinas y manipuladas.

- Modelos de clasificación: redes neuronales convolucionales y otros modelos de aprendizaje profundo se entrenan con conjuntos de datos de imágenes reales y falsificadas para aprender a identificar las diferencias.
- Redes siamesas: usadas para comparar pares de imágenes o videos para determinar si uno de ellos ha sido manipulado.
- Autoencoders anómalos: los autoencoders se pueden entrenar para reconstruir solo imágenes auténticas, de modo que cuando se les presenta un deepfake, la reconstrucción presenta errores significativos, lo que indica una posible falsificación.

6. Análisis de consistencia de meta-información

Los deepfakes no solo alteran visualmente el contenido de los medios, sino que también pueden introducir inconsistencias en los metadatos o en la estructura del archivo. Analizar la consistencia de los metadatos y la estructura del archivo puede revelar signos de manipulación.

Los metadatos de los archivos pueden ser inspeccionados en busca de anomalías como fechas de creación incoherentes, formatos de archivo inconsistentes o firmas digitales que no coinciden.

La validez de la prueba digital en tiempos modernos

En un juzgado, la validez de un deepfake, o de cualquier otro tipo de evidencia digital manipulada, es sumamente cuestionable y, en la mayoría de los casos, no sería aceptada como evidencia legítima debido a varias consideraciones legales y éticas.

El problema es que la proliferación de deepfakes pone en jaque el uso de cualquier tipo de prueba digital susceptible de poder haber sido manipulada.

Para que cualquier tipo de evidencia sea admitida en un tribunal, primero debe ser autenticada. Esto significa que la parte que presenta la evidencia debe ser capaz de demostrar que es lo que afirma ser. En el caso de los deepfakes, que son inherentemente falsificaciones, probar la autenticidad es intrínsecamente problemático.

Los deepfakes son creados específicamente para alterar la realidad, lo cual es un factor disuasorio significativo contra la admisibilidad en un tribunal de cualquier prueba digital que pueda generar duda y no pueda asegurar su autenticidad. Las leyes y reglas sobre evidencia digital exigen que los materiales no estén alterados y que mantengan su integridad original.

Cualquier evidencia debe tener una cadena de custodia clara y no comprometida. Si la cadena de custodia de una evidencia digital no puede garantizar su integridad desde su

creación hasta su presentación en el tribunal podrían fácilmente ser descalificada.

Presentar un deepfake como evidencia real podría considerarse un acto de fraude o engaño, lo cual tiene implicaciones legales graves, incluyendo posibles cargos por perjurio o intento de manipular los resultados judiciales.

Utilizar evidencia falsificada compromete la integridad del proceso judicial y podría llevar a un fallo erróneo, lo que afecta la justicia y la equidad del sistema legal.

Es por ello tan importante tener en cuenta la tecnología de forense digital para examinar y verificar la integridad de la evidencia digital.

Muchas jurisdicciones ya tienen leyes que regulan la admisibilidad de evidencias electrónicas y digitales, las cuales incluyen requisitos estrictos.

Sin embargo, ante el surgimiento de tecnologías como los deepfakes, es necesario considerar e implementar legislaciones específicas para abordar los desafíos que estas tecnologías presentan.

Autentificación del activo digital

Autentificar un video o una imagen, especialmente ante un juzgado, es crucial.

La autenticación de evidencia digital implica demostrar que el contenido es original y no ha sido alterado de ninguna manera.

Existen diversos métodos y técnicas que se pueden utilizar para autentificar videos e imágenes en un contexto legal.

Una opción es la cadena de custodia. Para ello debe mantenerse un registro detallado de dónde, cuándo y cómo se obtuvo el video o la imagen, y quién ha tenido acceso a él desde su creación. Esto incluye documentar cada transferencia, copia o análisis que se haya realizado.Es muy importante asegurar que el acceso a la evidencia digital se limite a personas autorizadas, y que este acceso esté debidamente registrado y verificado.

También es posible analizar la autenticidad de un archivo digital gracias a los metadatos asociados con el archivo digital, como fechas de creación y modificación, información del dispositivo de captura, ubicación (GPS), y otros datos técnicos que pueden probar la autenticidad. Además, se puede verificar que los metadatos no presenten anomalías o inconsistencias que podrían indicar manipulación.

En lo que respecta a las técnicas forenses digitales, se utilizan herramientas forenses para detectar cualquier signo de manipulación en los archivos. Esto puede incluir la detección de artefactos de edición, inconsistencias en la compresión de archivos, y análisis de hash para confirmar que el archivo no ha cambiado desde un punto de verificación anterior. Si es posible, se compara la evidencia digital con originales o con otras copias verificadas para identificar discrepancias.

Utilizar tecnologías de watermarking (marcas de agua) permite incrustar información oculta o visible en el archivo, lo cual puede servir como prueba de autenticidad. De manera similar, pueden aplicarse firmas digitales que usan criptografía para verificar el creador del archivo y confirmar que no ha sido alterado desde que se firmó.

Menos tecnológico, pero igualmente válido puede resultar el hecho de obtener declaraciones de testigos que puedan confirmar la creación y la procedencia del video o la imagen, aunque esto es más controvertido y de difícil conclusión.

Los expertos en análisis de multimedia, o peritos multimedia, pueden proporcionar testimonio sobre la autenticidad de la evidencia y sobre cualquier signo de manipulación detectado.

Con la llegada de la tecnología blockchaine posible crear un «pasaporte digital» para cada pieza de contenido multimedia que verifica su origen y autenticidad. Cada cambio o edición en el archivo se registraría en un libro de contabilidad distribuido, proporcionando un historial transparente e inmutable.

El uso de la tecnología blockchain para autenticar imágenes o videos representa una aplicación innovadora que ayuda a garantizar la integridad y la procedencia de los contenidos digitales.

Blockchain también puede usarse para gestionar los derechos de autor y las licencias de imágenes y videos.

Conclusión

Es indudable el enorme potencial que presenta la tecnología deepfake pero igual de enorme son los importantes desafíos éticos y legales que plantea, especialmente en lo que respecta a la privacidad, el consentimiento y la veracidad de la información.

Existe una creciente necesidad de marcos regulatorios y soluciones tecnológicas para detectar y mitigar los efectos de los deepfakes, asegurando que esta poderosa tecnología se use de manera responsable y ética.

El rápido avance de la tecnología deepfake exige métodos de detección igualmente sofisticados. A medida que estas tecnologías continúan evolucionando, la investigación y el desarrollo continuos son cruciales para asegurar la autenticidad de los medios digitales y prevenir el mal uso.

Las técnicas forenses digitales para la detección de deepfakes son un campo en rápido desarrollo que combina metodologías tradicionales de análisis de imágenes con tecnologías de vanguardia basadas en IA.

La eficacia de estos métodos depende de mantenerse al día con las técnicas de generación de deepfakes, que están en constante evolución.

La colaboración entre investigadores, desarrolladores de tecnología y agencias legales es fundamental para mejorar continuamente las herramientas forenses y contrarrestar las amenazas que presentan los deepfakes.

Referencias

Westerlund, M. (2019). El surgimiento de la tecnología deepfake: una revisión. Revista de Gestión de Innovación Tecnológica.

https://timreview.ca/sites/default/files/article_PDF/TIMReview_November2019 – D – Final.pdf

Gong, D., et al. (2020). Forense de deepfakes, una detección sintetizada por IA con redes generativas adversariales convolucionales profundas. Revista Internacional.

Haz clic para acceder a ijatcse58932020202007 12-81442-1h6s4fi.pdf

Shad, H.S., et al. (2021). Análisis comparativo del método de detección de imágenes deepfake utilizando redes neuronales convolucionales. Inteligencia Computacional y Neurociencias. https://www.hindawi.com/journals/cin/2021/3111676/

Paris, B., & Donovan, J. (2019). Deepfakes and Cheap Fakes: The Manipulation of Audio and Visual Evidence. En Proceedings of the 2019 Data & Society Report. Data & Society

Research Institute.