UNIVERSIDAD SAN FRANCISCO DE QUITO USFQ

Colegio de Ciencias e Ingenierías

Aplicación de redes generativas antagónicas GAN (Generative Adversarial Networks) para la comprensión de Deepfakes y sus repercusiones sociales.

Juan Daniel Salazar Díaz

Ingeniería en Ciencias de la Computación

Trabajo de fin de carrera presentado como requisito
para la obtención del título de
Ingeniero en Ciencias de la Computación

Quito, 20 de Diciembre de 2022

UNIVERSIDAD SAN FRANCISCO DE QUITO USFQ

Colegio de Ciencias e Ingenierías

HOJA DE CALIFICACIÓN
DE TRABAJO DE FIN DE CARRERA

Aplicación de redes generativas antagónicas GAN (Generative Adversarial Networks) para la comprensión de Deepfakes y sus repercusiones sociales.

Juan Daniel Salazar Díaz

Nombre del profesor, Título académico **Felipe Grijalva, Ph.D.**

Quito, 20 de Diciembre de 2022

© DERECHOS DE AUTOR

Nombres y apellidos: Juan Daniel Salazar Díaz

Código: 00130124

Cédula de identidad: 1717463747

Lugar y fecha: Quito, 20 de Diciembre de 2022

ACLARACIÓN PARA PUBLICACIÓN

Nota: El presente trabajo, en su totalidad o cualquiera de sus partes, no debe ser considerado como una publicación, incluso a pesar de estar disponible sin restricciones a través de un repositorio institucional. Esta declaración se alinea con las prácticas y recomendaciones presentadas por el Committee on Publication Ethics COPE descritas por Barbour et al. (2017) Discussion document on best practice for issues around theses publishing, disponible en http://bit.ly/COPETheses.

UNPUBLISHED DOCUMENT

Note: The following capstone project is available through Universidad San Francisco de Quito USFQ institutional repository. Nonetheless, this project – in whole or in part – should not be considered a publication. This statement follows the recommendations presented by the Committee on Publication Ethics COPE described by Barbour et al. (2017) Discussion document on best practice for issues around theses publishing available on http://bit.ly/COPETheses.

RESUMEN

Deepfakes son videos híper realistas que permiten reemplazar el rostro de una persona por el rostros de otra. Esta tecnología es una implementación de redes generativas antagónicas o también conocidas como GANs (por sus siglas en inglés), la cual es una arquitectura de aprendizaje profundo introducida por Ian Goodfellow en 2014. Desde sus primeras implementaciones en 2017, los videos híper realistas deepfakes han causado gran revuelo en la sociedad debido a sus impactantes resultados a la hora de realizar un reemplazo de rostros. A causa de a esta razón, el estado del arte sobre este tema ha tenido un gran enfoque en métodos de detección y contramedidas que puedan asistir en mitigar potenciales amenazas producidas por la mala práctica de esta herramienta. En el presente trabajo se desarrolla una aplicación de redes GAN a través de la implementación de un deepfake de tipo de intercambio de rostro, con el fin de poner en contexto las distintas problemáticas sociales asociadas a esta tecnología. Para el desarrollo del deepfake, se realizó una adaptación del proyecto DeepFaceLab. Para analizar el progreso del modelo se desarrollaron 4 videos deepfakes durante distintas épocas de entrenamiento del modelo. Los resultados obtenidos, tras entrenar el modelo durante 1 millón de épocas y reducir la función de pérdida aproximadamente a 0.1, producen un video deepfake híper realista que pone en contexto las distintas inquietudes relacionadas a esta tecnología.

Palabras clave: Deepfake, redes generativas antagónicas, GANs, inteligencia artificial, aprendizaje automático, aprendizaje, profundo.

ABSTRACT

Deepfakes are hyper-realistic videos that allow a person's face to be replaced by the face of another person. This technology is an implementation of generative adversarial networks, also known as GANs, which is a deep learning architecture introduced by Ian Goodfellow in 2014. Since its first implementation in 2017, hyper-realistic deepfake videos have caused a stir in society due to its impressive results when performing face replacement. Due to this reason, the state of the art on this subject has had a great focus on detection methods and countermeasures that can assist in mitigating potential threats caused by the malpractice of this tool. In the present work, a GAN network application is developed through the implementation of a deepfake of the face swap type in order to put into context the different social problems associated with this technology. For the development of the deepfake, an adaptation of the DeepFaceLab project was put together. To analyze the progress of the model, 4 deepfake videos were developed during different epochs of the training of the model. The results obtained after training the model for 1 million epochs and reducing the loss function values to approximately 0.1, produce a hyper-realistic deepfake video that puts the different concerns related to this technology into context.

Key words: Deepfake, generative adversarial networks, GANs, artificial intelligence, machine learning, deep learning.

TABLA DE CONTENIDO

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

INTRODUCCIÓN

Desde su introducción, uno de los campos de la inteligencia artificial mayormente explorados es el aprendizaje automático. Debido a que esta área de la inteligencia artificial busca desarrollar algoritmos computacionales, se han desarrollado una gran cantidad de modelos y metodologías de aprendizaje automático. Uno de estos modelos es de las redes neuronales artificiales, las cuales modelan la relación entre un conjunto de señales de entrada y señales de salida, utilizando un modelo derivado de nuestro entendimiento de cómo un cerebro biológico responde a estímulos de entradas sensoriales.

El modelo de redes neuronales artificiales fue muy exitoso, ya que intenta simular el cerebro humano, por lo tanto el modelo resultó novedoso debido a que computadoras podrían llegar a simular comportamientos similares a los del ser humano. De esta forma, el modelo de redes neuronales artificiales fue adoptado y diversas iteraciones del modelo surgieron. Tal es el caso del modelo de aprendizaje profundo. Este modelo permite a las computadoras aprender de experiencias y entender el mundo en términos de una jerarquía de conceptos, con cada concepto definido en términos de su relación a un concepto más simple. Al conseguir conocimiento a partir de la experiencia, este enfoque evita la necesidad de operadores humanos para especificar todo el conocimiento que el computador necesita. La jerarquía de conceptos permite al computador aprender conceptos complicados al construirlos a partir de conceptos más simples. Tal y como menciona Goodfellow en su libro de Deep Learning, si se construyera un gráfico de la jerarquía de conceptos, este sería “profundo” y con muchas capas. Por esta razón, denominaron a este modelo, inspirado en redes neuronales artificiales, aprendizaje profundo.

Pero ¿ qué tienen que ver el aprendizaje automático, redes neuronales artificiales, y el aprendizaje profundo? Tal y como se puede apreciar, el campo de la inteligencia artificial ha llevado una trayectoria jerárquica, en la cual distintos modelos se derivan a partir de otros, y de estos surgen nuevas aplicaciones. Entre los modelos de aprendizaje profundo más conocidos, se encuentran: redes neuronales convolucionales, redes de memoria a corto y largo plazo, redes neuronales recurrentes, redes de creencias profundas, y redes neuronales antagónicas.

Los algoritmos de aprendizaje profundo se pueden clasificar en: supervisados, semi supervisados, o no supervisados. El aprendizaje supervisado necesita supervisión humana para el desarrollo de su proceso, por lo tanto no es un proceso completamente autónomo. Es a partir de esta cuestión, que muchos investigadores desempeñaron su trabajo dentro del aprendizaje no supervisado, en donde se originan los modelos 'generativos'. "El objetivo del algoritmo de un modelo generativo es el aprender un modelo que se aproxime a una serie de datos lo más cercano posible [9]." Es a partir de estos modelos generativos, que se han desarrollado las redes generativas antagónicas, o mejor conocidas como GANs (por sus siglas en inglés). Las GANs se originan a partir de la teoría de juegos, en donde una red neuronal conocida como 'generadora' se encarga de crear muestras que serán evaluadas por otra red neuronal denominada 'discriminadora'. La evaluación consiste en determinar si la muestra creada por la generadora es real o falsa. El proceso sigue un ciclo iterativo que tiene como finalidad engañar a la red discriminadora a partir de las muestras creadas por la red generadora. Y, una vez, que el proceso finaliza, las redes GAN puede producir muestras como, por ejemplo, imágenes híper realistas.

A medida que las aplicaciones de redes generativas antagónicas se fueron refinando y diversificando, surge una implementación, la cual ha tenido un gran impacto desde su introducción. Esta tecnología es conocida como Deepfake. “Deefakes son el producto de aplicaciones de inteligencia artificial que mezclan, combinan, reemplazan, y sobreponen imágenes y video para crear videos falsos que parecen auténticos [5].” Los deepfakes fueron introducidos por primera vez en 2017 y poco a poco han comenzado a propagarse en el mundo digital. En un inicio, las primeras implementaciones se encargaban de generar videos falsos de celebridades y políticos. Hoy en día, los deefakes se han se han utilizado en varias campos como son: el cine, los videojuegos, medios educativos, redes sociales, cuidados de la salud, ciencia de materiales y campos financieros.

A pesar de la contribución a varias industrias por parte de esta herramienta, existe cierta preocupación del mal uso que se le pueda dar a los Deepfakes. “Los beneficios de los Deepfakes a la humanidad pueden incrementar, pero no tanto como sus pérdidas [16].” Por ejemplo, con el desarrollo de esta herramienta, robar identidades de personas se facilitará, se podrán sabotear elecciones políticas y en campos como el deporte en donde existan competencias. Contenido con el propósito de humillar, sobornar, o adulterar a otra persona podrá ser creado. Figuras públicas podrán ser vistas en lugares a los que nunca han ido o decir cosas que nunca han dicho. Organizaciones o individuos podrán crear contenido con finalidad terrorista y las noticias falsas tienen el potencial de difundir desinformación que puede llegar a tener serias repercusiones. A causa de esto, existe una gran preocupación dentro del campo del periodismo, ya que la credibilidad de estos medios se ve comprometida: “se afirma que esta nueva etapa en la era de la desinformación puede causar el fin del periodismo fiable [16].” En conclusión, el campo de la

inteligencia artificial ha desarrollado una herramienta sofisticada con aplicaciones positivas para varios sectores de la sociedad. Pero, antes se debe entender qué significa deepfake. A pesar de la asociación que comúnmente se hace con los videos de reemplazo de rostro, existen otro tipo de deepfakes, por lo cual se debe definir adecuadamente este concepto.

DESARROLLO DEL TEMA

Para el desarrollo del proyecto, se ha considerado los fundamentos del diseño de ingeniería, basándose en el libro "Engineering Design Process" de los autores Yousef Haik y Tamer M. Shahin [22].

1. Deepfakes

1.1. ¿Qué es un deepfake?

Para poder definir qué es un deepfake, hay que tomar en cuenta las distintas definiciones que se le otorgan al término. Por lo general, se asume que deepfake se origina de las palabras 'deep' por aprendizaje profundo o 'deep learning' y 'fake' por su capacidad para producir contenido falso. "La frase deepfake hace referencia a una técnica basada en aprendizaje profundo que genera falsos filmes en donde el rostro de una persona es intercambiado por el rostro de otra persona [1]."

Por lo general, en la literatura, lo que comúnmente se conoce como deepfake, es la creación de un contenido visual en el que se intercambia el rostro de una persona por el rostro de otra persona. Pero, existen otras aplicaciones de este concepto. Por ejemplo, ciertos investigadores del tema consideran que los deepfakes no solo abarcan el intercambio de rostros, también pueden utilizar técnicas de aprendizaje profundo para generar discursos auditivos falsos o imágenes falsas, a las cuales también se refieren como deepfakes. Evidentemente, el contenido no se limita a videos, sino que puede expandirse a otras áreas. Es por esta razón que se ha propuesto un término que pueda abarcar todos estos casos. El término propuesto es conocido como 'síntesis falsa'. Este concepto tiene como finalidad llegar a un punto de neutralidad para poder hacer referencia a las

distintas implementaciones de la definición. Pero, esta alternativa no ha sido adoptada por la comunidad en general. También se intentaron introducir dos términos nuevos como son el "shallowfake y cheapfake [4]" para hacer referencia a la manipulación de bajo nivel de contenido audiovisual a partir de software de fácil acceso.

También cabe mencionar que legisladores han intentado proponer una definición en caso de requerir combatir a deepfakes maliciosos mediante la legislatura. Por ejemplo, la siguiente acta de los Estados Unidos define a un deepfake como: "Un registro audiovisual creado o alterado de tal manera que el registro parecería falsamente, a un observador razonable, como un registro auténtico del discurso o la conducta de un individuo[4]."

Como se puede evidenciar, existen diferentes conceptos y definiciones de lo que pude llegar a ser un deepfake. Pero, en el siguiente documento, se utilizará la definición comúnmente asociada a este concepto, la cual hace referencia a las técnicas de aprendizaje profundo y a la creación de videos híper realistas falsos, para ser más específico, se referirá a deepfake como: técnica de aprendizaje profundo derivada de redes generativas antagónicas (GANs, por sus siglas en inglés). "Para generar tales videos falsos, se necesitan dos redes neuronales: (i) una red generativa y (ii) una red discriminativa. La red generativa crea imágenes falsas usando un codificador y decodificador. La red discriminativa define la autenticidad de las imágenes generadas. La combinación de estas redes es llamada Redes Generativas Antagónicas (GANs) [3]."

Figura 1: Ejemplos de deepfakes
Fuente: [16]

1.2. Tipos de deepfakes.

Como se mencionó anteriormente, el término deepfake se utiliza para distintos tipos de aplicaciones. A continuación se presentan los distintos tipos de implementaciones referidos como deepfakes:

- 1) Reemplazo de rostro: Comúnmente conocido como intercambio de rostro. Este tipo de deepfake incluye un rostro fuente que se fija en un rostro de destino. La identidad y atención están enfocadas en el rostro fuente. Es el tipo de deepfake utilizado para generar videos híper realistas falsos.
- 2) Reconstrucción facial: También conocida como marioneta. Incluye el manejo de características de un rostro objetivo, incluyendo el movimiento de labios, cejas, ojos y el movimiento de la cabeza. La reconstrucción facial no está destinada a reemplazar personalidades como otros tipos de deepfakes, sino busca distorsionar las expresiones de un individuo.
- 3) Generación facial: El objetivo de la generación facial es producir rostros completamente nuevos a partir de un rostro fuente.
- 4) Síntesis de voz: incluye la creación de un modelo de voz de alguien que puede leer el texto como la entidad objetivo de la misma manera, entonación y cadencia.

En el caso de este proyecto se hace referencia al reemplazo de rostro cuando se utiliza el término deepfake.

2. Redes generativas antagónicas (GANs)

2.1. Modelado generativo.

El modelado generativo es un enfoque al entrenamiento no supervisado, en el cual muestras de entrenamiento son extraídas de una distribución de datos, y a partir de un algoritmo de un modelo generativo, se busca aprender un modelo que se aproxime a la distribución de datos lo más cerca

posible. Es decir, el modelado generativo implica descubrir y aprender automáticamente las regularidades o patrones en los datos de entrada, para que el modelo pueda generar muestras que podrían haberse extraído de los datos originales. “Una forma sencilla de aprender una aproximación de p_{datos} es escribir explícitamente una función $p_{modelo}(x; \theta)$, controlada por el parámetro θ y buscar un valor para el parámetro, el cual haga que p_{datos} y p_{modelo} sean lo más similares posibles [9].” En resumen, el modelado generativo tiene como fin el tratar de encontrar un modelo que permita emular a una distribución de datos lo más aproximado posible.

Aparte de tomar un punto x como entrada y retornar un estimado de la probabilidad de generar aquel punto, un modelo generativo puede ser útil para poder crear una muestra a partir de la distribución del modelo p_{modelo}. En algunos casos, generar muestras puede llegar a tener un costo computacionalmente alto, por lo tanto algunos modelos generativos evitan este problema y aprenden solo un proceso de generación de muestras tratable. Estos son llamados los modelos generativos implícitos, los cuales abarcan a las GANs. “Hoy en día los modelos generativos se pueden dividir en tres categorías: Redes Generativas Antagónicas (GANs), Auto Codificadores Variacionales (VAE por sus siglas en inglés), y Redes Auto Regresivas [12].”

2.2. ¿Qué es una GAN?

Comúnmente conocidas como GANs (por sus siglas en inglés), las redes generativas antagónicas son un modelo de inteligencia artificial propuesto en 2014 por Ian Goodfellow. Este modelo fue creado a partir de la teoría de juegos entre dos modelos de aprendizaje automático, por lo general implementados usando redes neuronales. En este juego, se encuentra una red denominada generadora, la cual es capaz de crear datos a partir de un conjunto de datos para el entrenamiento

de la red. A partir de la muestra creada por la red generadora, aparece el siguiente jugador involucrado en el juego: la red discriminadora. Esta red se encarga de realizar una evaluación binaria (verdadero o falso) de la muestra creada por la generadora. Para realizar esta evaluación, utiliza los datos de entrenamiento utilizados por la red generadora para crear la muestra. Tras cada evaluación, cada jugador genera una peso. El objetivo de los jugadores es reducir este peso al máximo.

"Se puede pensar en GANs como policías y falsificadores: los falsificadores tratan de generar dinero falso, mientras los policías tratan de arrestar a los falsificadores. La competencia entre falsificadores y policía conduce a dinero falso cada vez más realista, hasta que eventualmente la policía no puede diferenciar entre el dinero falso y el verdadero [9]." Teniendo en cuenta la estructura básica del modelo, Ian Goodfellow creó las redes generativas antagónicas como un algoritmo diseñado para resolver el problema del modelado generativo.

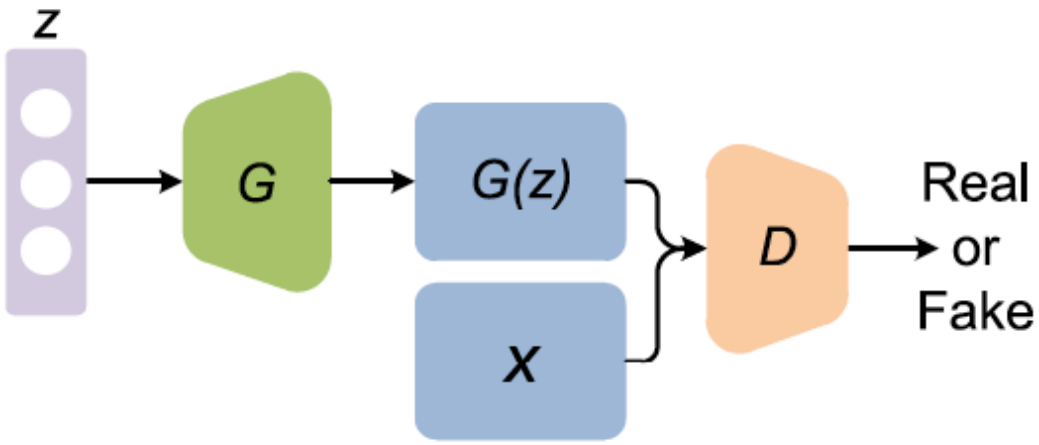

Figura 2: Estructura de una GAN

Fuente: [12]

2.2.1. Red Generadora.

Tal y como se puede apreciar en la figura 2, la red generadora *G* define el modelo p_{modelo} y tiene como entrada un vector aleatorio de ruido *z*. "El vector de entrada *z* puede ser pensado como una fuente de aleatoriedad en un sistema determinístico, análoga a una semilla pseudoaleatoria de un generador de números aleatorios [9]." Las muestras del vector z, son por lo tanto 'ruido'. Este ruido es mapeado a un nuevo espacio de datos a través de la red generadora para obtener una muestra falsa *G(z)*. Por lo tanto, la función principal del generador es aprender la función que transforma las muestras de ruido *z* en muestra realistas.

2.2.2. Red Discriminadora.

Es una clasificador binario que toma como entradas: la muestra *x* del conjunto de datos y la muestra falsa *G(z)* creada por el generador. La red discriminadora examina la muestra generada *G(z)* y retorna una probabilidad de que la muestra es real o falsa. Cuando el discriminador no puede distinguir si la muestra proviene del conjunto de datos o del generador, se puede decir que se ha llegado al estado óptimo. Es en este punto, se obtienen un modelo generador que ha aprendido la distribución de los datos reales.

2.2.3. Entrenamiento antagónico.

Anteriormente se mencionó que el modelaje generativo de las GANs era un problema de entrenamiento no-supervisado. Por el contrario, el entrenamiento de esta arquitectura está planteado como un tipo de entrenamiento supervisado. Las dos redes se entrenan juntas, y tras la evaluación de la red discriminadora, tanto la red generadora como la discriminadora ajustan sus parámetros respectivos en un proceso de 'backpropagation'. La discriminadora es actualizada para

diferenciar mejor entre muestras reales y falsas en la siguiente ronda. Mientras que la red generadora se actualiza basándose en qué tan bien sus muestras pudieron engañar a la red discriminadora.

De esta forma, los dos modelos están compitiendo el uno contra el otro, y son antagónicos desde el punto de vista de la teoría de juegos, ya que se encuentran jugando un juego de suma cero. Esto significa que cuando el discriminador logra identificar correctamente las muestras, sus parámetros no son ajustados, mientras que el generador es penalizado y actualizará los parámetros del modelo para generar muestras más cercanas a la de los datos de entrenamiento. Al contrario, cuando el generador logra engañar al discriminador, los parámetros del generador se mantienen y el discriminador actualiza sus parámetros.

Tras seguir este proceso iterativo del entrenamiento, el generador va a crear réplicas idénticas a la del conjunto de datos originales. “Cuando el discriminador no pueda determinar las diferencias entre ambas distribuciones, en este estado, el modelo alcanzará la solución óptima global [12].” En otras palabras, se habrá llegado al objetivo del algoritmo, el cual es encontrar el equilibrio de Nash local, el cual se define como: “Un punto que es el mínimo local del costo de cada jugador con respecto a los parámetros de ese jugador [9].”

3. Repercusiones sociales derivadas de deepfakes

A pesar de haber sido introducida en 2017, los deepfakes han tenido un impacto social considerable. Al navegar la web, se pueden encontrar una serie de implementaciones de gran calidad que han sido desarrolladas por varios usuarios en la web. Pero, el hecho de poder crear

contenido híper realista, con el cual se pueda manipular la imagen de una persona ajena con gran control, lleva a plantear una serie de cuestiones: ¿Qué amenazas puede conllevar? ¿Qué beneficios se pueden extraer de estas tecnologías? ¿Podrán desarrollarse métodos capaces de detectar esta clase de tecnologías? A medida que estas herramientas se siguen refinando y optimizando, esta serie de preguntas es importante planteárselas, ya que los deepfakes pueden llegar a generar grandes beneficios o consecuencias negativas sino se preparan los mecanismos adecuados para afrontar estas situaciones emergentes.

3.1. Beneficios.

Uno de los sectores de la sociedad de mayor importancia que puede llegar a beneficiarse de la tecnología deepfake es el sector de la educación. Los deepfakes pueden llegar crear un escenario educativo mucho más didáctico y entretenido para los estudiantes. Tal y como menciona Buo: "La investigación en curso está explorando formas de desarrollar un sistema de inteligencia artificial que automatice el proceso de producción de contenido utilizando tecnología deepfake. Un sistema en particular, conocido como 'LumièreNet', agilizará el proceso de creación de videos educativos [19]." Es decir, que con este sistema en desarrollo, se podrá realizar videos didácticos, por ejemplo, de figuras históricas para presentar clases de historia.

En otras industrias, como la del entretenimiento, los deepfakes llevan siendo utilizados durante varios años. Por ejemplo, en la película 'Furious 7' del año 2017, el rostro del actor fallecido, Paul Walker, fue recreado en la última escena de la película mediante deepfakes. Otro caso es el de la industria de los videojuegos, la cual puede utilizar estas herramientas para mejorar la experiencia

del jugador. Un deepfake del tipo síntesis de voz, puede ser utilizado para replicar voces o hasta desarrollar entornos virtuales con sonoridades realistas y naturales.

Otra de las áreas en las que se encuentran incorporando aplicaciones para esta tecnología es en el de la salud, para ser más específico, la salud mental. En concreto, se pueden ayudar a individuos a lidiar con la muerte de un ser querido al desarrollar una versión digital de dicho ser querido. Otro instancia interesante es la de la asistencia en procesos de rehabilitación de individuos que sufren de adicciones. “La organización mundial de la salud ha desarrollado una solución basada en inteligencia artificial llamada ‘Florence’, la cual permite a individuos a sobrellevar la adicción al tabaco. Los usuarios pueden tener una conversación con ‘Florence’ para construir confianza en su proceso de rehabilitación”[19].

En fin, los posibles beneficios de esta aplicación de redes GAN no se limita a la industria del entretenimiento o la investigación. De momento, los servicios que se encuentran en desarrollo están en sus etapas iniciales, pero son esperanzadores de cara a futuro.

3.2. Amenazas emergentes.

Puede deberse al éxito con el que consigue replicar el rostro de una persona, pero las consecuencias que puede traer esta herramientas podrían ser graves y de mucho cuidado. Dentro de la esfera de los sistemas judiciales, estas aplicaciones pueden llegar a ser utilizadas como instrumentos para la manipulación de evidencia. “Durante interrogatorios, cuando una parte testifica afirmativamente sobre los detalles de un video falso, la parte contraria pueda negar el contenido del video [19].” La irrupción de este elemento dentro de un caso judicial puede llegar a complicar el mismo, llegando

a costar más recursos y tiempo a las partes involucradas. Por ejemplo, en el Reino Unido, en un caso de retención de custodia, una madre presentó un deepfake de tipo síntesis de voz como evidencia en el caso. Siguiendo instrucciones en línea, la madre logró crear un audio en el que trataba de incriminar al padre de conducta violenta [19]. Eventualmente, se logró identificar la manipulación, pero en algunos casos las medidas de detección han fallado. Por lo tanto, es de vital importancia desarrollar contramedidas para prevenir que el sistema judicial se puede ver comprometido.

Uno de los mayores afectados puede llegar a ser el área política. Es una de las que más susceptible se encuentra a las consecuencias negativas de la mala práctica de esta herramienta, y debido a esto es uno de los principales incentivos detrás de las investigaciones para contramedidas y detección. Gracias a la era de las redes sociales, los videos deepfakes pueden ser difundidos fácilmente a una gran audiencia y ser usados para crear desinformación con el fin de crear ventajas políticas para una parte o perjudicar a otra. Uno de los ejemplos más notables fue la circulación, en redes sociales, de un video manipulado de la política estadounidense, Nancy Pelosi. En el video aparentaba estar intoxicada con alcohol y mal pronunciando sus palabras. El presidente estadounidense de aquel momento, Donal Trump, compartió el video en su cuenta de Twitter para cambiar la percepción pública de su oponente, y el video rápidamente consiguió 2.5 millones de visitas en Facebook. “A pesar de llamados bipartidistas para que se elimine el video, un portavoz de Facebook confirmó que no se eliminarían porque la plataforma no poseía políticas que dicten la eliminación de información falsa [19].” Este caso, fue uno de los puntos de inflexión para que políticas de legislación en contra de este tipo de videos sea desarrollada en distintas plataformas. En las etapas iniciales del desarrollo del video deepfake de este proyecto, se buscó utilizar la plataforma Google

Colab para el entrenamiento del modelo, pero la plataforma tenía prohibido la utilización de su plataforma para la creación de videos deepfakes.

En adición a los sistemas judiciales y políticos, los sistemas económicas también pueden verse afectados. Estafadores podrían intentar personificar a individuos de alta importancia dentro de una compañía para obtener acceso a información sensible o solicitar transferencias de dinero sin llegar a ser detectados. En un caso dentro del Reino Unido, estafadores defraudaron una firma al suplantar la identidad del CEO mediante un video falso para llegar a convencer a empleados del departamento de finanzas de realizar una transferencia de dinero. “Forrester Research predijo una pérdida monetaria de $250 millones para fines de 2020 debido a fraudes causados por deepfakes [19].”

Es evidente que las consecuencias negativas de la mala práctica de estas herramientas puede conllevar a graves consecuencias sociales, políticas, económicas y de otras cualidades aún por verse. Para prevenir tales escenarios, es necesario el desarrollo e investigación de contramedidas.

3.3. Métodos de detección.

Las investigaciones enfocadas en métodos de detección de tecnologías deepfakes ha sido desarrollada utilizando redes neuronales artificiales. Para el caso de videos manipulados, un estudio reciente desveló que la mayoría de este contenido es de baja resolución debido al alto costo de recursos que conlleva desarrollar un deepfake de alta resolución. Los videos deepfake de baja resolución son fáciles de detectar utilizando redes neuronales convolucionales (CNN por su siglas en inglés). “Los investigadores pudieron detectar e identificar con éxito el 99.1% de los deepfakes

[19].” A pesar del éxito de estas investigaciones, también se recalca que a medida videos con alta resolución aparezcan, las CNN se volverán inefectivas.

Otra de las soluciones que se evalúan es la de equipos de forenses digitales que realicen evaluaciones del contenido. Por ejemplo, expertos forenses pueden utilizar técnicas computacionales para aislar elementos de una imagen como sombras o reflexiones, y así poder detectar manipulación de pixeles. También pueden investigar los metadatos para verificar si el archivo ha sido alterado o revisar el historial de edición y verificar la cantidad de veces que el archivo ha sido comprimido. Un obstáculo para adoptar esta solución es que llegar a tener un equipo dedicado de forenses digitales puede llegar a ser costoso. Además, a medida que se siga produciendo y difundiendo deepfakes, el utilizar un equipo de forenses puede llegar a ser ineficiente. Por lo tanto, esta solución puede ser de gran utilidad para casos puntuales.

Una solución que puede ser adoptada a gran escala es el uso de blockchain. Esta tecnología puede actuar como una firma transparente digital en contenido digital, incluyendo deepfakes. “Esta solución se basa en registros de secuencia de tiempo para rastrear el historial de contenidos multimedia, monitoreando dónde se utilizó para posteriormente determinar sus orígenes [19].” Esta propuesta puede ser fácilmente integrada en navegadores web para verificar la autenticidad del contenido. Pero, presenta una serie inconvenientes. Esta firma digital es propensa a errores como la falsa identificación del contenido como falso. Además, blockchain es una tecnología relativamente nueva que puede resultar costosa y difícil de implementar.

Las regulaciones gubernamentales y sanciones a la creación de deepfakes socialmente nocivos es otra de las contramedidas que se exploran. Pero, uno de los argumentos en contra de esta medida es que la libertad de expresión se vería comprometida y los límites de la censura judicial se verían afectados. Por lo tanto, es crucial evitar implementaciones generalizadas que puedan infringir la libertad de expresión [19].

A pesar de disponer de una amplia gama de propuestas, las contramedidas en contra del uso malicioso de deepfakes aún se encuentra en sus etapas iniciales. Actualmente, investigaciones y medidas siguen en desarrollo y progresan a la vez que el contenido de deepfakes sigue produciéndose.

4. Implementación de un deepfake

4.1. Estándar ISO/IEC TR 24372:2021

La implementación fue desarrollada tomando en cuenta el estándar para la inteligencia artificial: ISO/IEC TR 24372:2021. Este estándar fue escogido ya que menciona que tiene como intención mirar distintos métodos computacionales dentro de los sistemas de la inteligencia artificial. Además, el estándar proporciona especificaciones que incluyem términos y definiciones que se encuentren directamente relacionados con el alcance de este proyecto: GAN, generador, discriminador, entre otros.

4.2. Materiales

4.2.1. Datos de origen

Para poder iniciar el proceso de creación de un deepfake, es necesario disponer de datos de origen, en este caso un video, a partir del cual se podrá crear una máscara del rostro que será superpuesta sobre los datos de destino. A continuación se presentan, las características del video:

Dimensiones	1920 x 1080
Marcos por Segundo	30
Duración	494 [segundos]
Imágenes Extraídas	14820

Tabla 1: Características datos de origen

Cabe recalcar que la calidad de los datos de origen es de vital importancia para poder conseguir los resultados deseados. Para este fin, se tiene que conseguir datos que sean de la mayor resolución posible, para así poder captar todos los detalles del rostro. Además es muy importante que la numerosidad sea extensa. Es por esto que se produjo un video de una duración de casi 500 segundos, del cual se pudieron extraer 14820 imágenes del rostro. También, se tiene que buscar una amplia variabilidad de los datos, lo cual significa que se debe captar el rostro en diversos ángulos e iluminaciones. Para esta implementación, se replicaron las mismas condiciones de iluminación y tomas tanto en los datos de origen como en los datos de destino, para así poder obtener un mejor resultado.

Figura 3: Datos de origen

Figura 4: Datos de origen – ángulo izquierdo

Figura 5: Datos de origen – ángulo derecho

4.2.2. Datos de destino

En el caso de los datos de destino se tienen que tomar en cuenta menos variables, ya que el deepfake final dependerá en gran parte de la calidad de los datos de origen. Al igual que en los datos de origen, los datos de destino son un video en formato mp4 que servirá para superponer la máscara creada a partir de los datos de origen. A continuación se presentan las características de los datos de destino:

Dimensiones	1920 x 1080
Marcos por Segundo	30
Duración	88 [segundos]
Imágenes Extraídas	2640

Tabla 2: Características de datos de destino

A diferencia de los datos de origen, no es necesario una amplia variabilidad o numerosidad de los datos de destino. Queda a disposición del usuario. En este caso se produjo un video de 88 segundos sobre el cual se aplicará la máscara creada a partir de los datos de origen.

4.3. Métodos

A continuación, el método utilizado para la producción del video deepfake es el marco de trabajo open source DeepFaceLab. Esta implementación nace en el 2018 con el objetivo de "despertar la conciencia de la gente sobre videos de manipulación facial y brindar comodidad a los investigadores de detección de falsificaciones [20]." DeepFaceLab es una implementación desarrollada en el lenguaje de programación Python que utiliza GANs como base de su desarrollo.

Además, el proyecto produce deepfakes del tipo ‘reemplazo de rostro’. La finalidad del proyecto es la del entretenimiento, pero principalmente la de colaborar al desarrollo de la detección de falsificaciones al proporcionar datos de falsificación (deepfakes) de alta calidad.

Al ser un proyecto open source, el código puede ser adaptado a las conveniencias del usuario. Esto permite generar una alta personalización del producto final. Para el desarrollo de este proyecto, se añadió un módulo para el registro de los valores de las funciones de pérdida de las redes generadora y discriminadora. El código de DeepFaceLab modificado se lo puede encontrar en el ANEXO D.

A continuación, se procede con el pipeline de trabajo del proyecto que consiste en tres etapas de producción: extracción, entrenamiento y conversión.

4.3.1. Etapa de extracción

La primera fase del proceso busca extraer los rostros de los datos de origen y destino. Esta etapa consiste de tres pasos para obtener el resultado deseado: detección del rostro, alineamiento del rostro y la segmentación del rostro.

Para iniciar la detección, DeepFaceLab utiliza el algoritmo de extracción S3FD (Single Shot Scale-invariant Face Detector). El algoritmo producirá una imagen de formato png, la cual tendrá la sección del rostro detectado por el algoritmo. Además, se especificó que la resolución de los rostros extraídos sea de 512x512 pixeles. Este proceso se lo debe realizar tanto para los rostros de los datos de origen, como para los rostros de los datos de destino. Se obtuvieron 14820 imágenes del rostro de origen y 2640 del rostro de destino.

Figura 6: Rostro extraído de datos de origen mediante S3FD

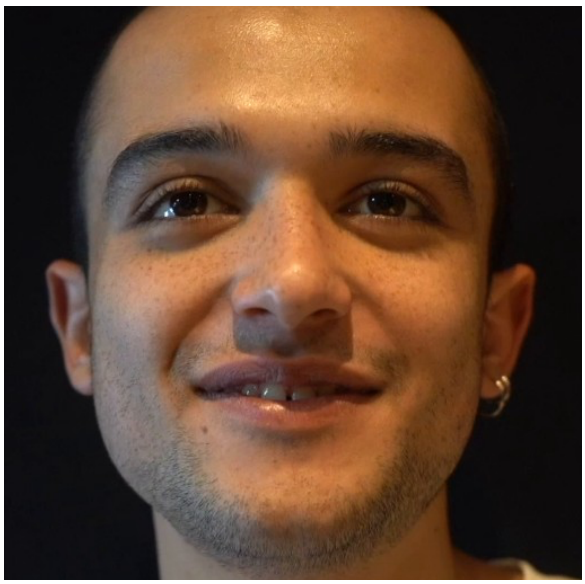

Figura 7: Rostro extraído de datos de destino mediante S3FD

Posteriormente, sigue el paso de alineamiento del rostro. Este paso consiste en realizar un mapeo de los rostros para poder identificar los puntos de referencia de los mismos. “Los puntos de referencia son clave para mantener la estabilidad en el tiempo [20].” Es decir, estos puntos ayudan con contenido que tiene un metraje sucesivo y contiene movimiento del rostro, como son los

videos. Para este proceso, se utiliza el algoritmo 2DFAN (2D Face Alignment). Finalmente, utilizando un método propuesto por Umeyama [20], se calcula la matriz de transformación de similitud para finalizar con el mapeo del rostro.

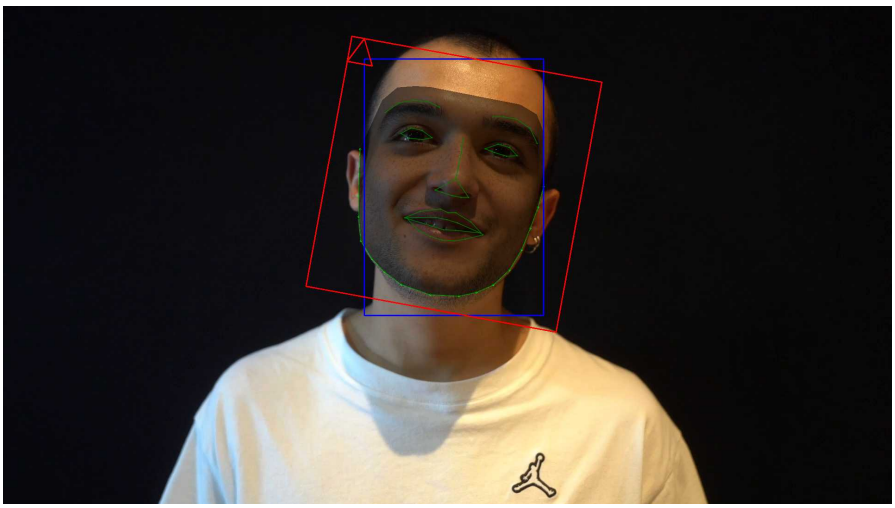

Figura 8 : Mapeo del rostro de destino usando 2DFAN

Como último paso de la etapa de extracción, se encuentra la segmentación del rostro. Este paso tiene como objetivo trazar máscaras de segmentación de los rostros. Se utilizan las imágenes generadas en el proceso de alineamiento del rostro con una red de segmentación de rostros denominada TernausNet, mediante la cual "un rostro con cabello, dedos, o lentes puede ser segmentado exactamente [20]." El producto final serán las máscaras de segmentación de los rostros:

Figura 9: Máscara de segmentación de rostro de destino usando TernausNet

Con la obtención de las máscaras, se habrá finalizado la etapa de extracción y se puede proceder al entrenamiento del modelo GAN.

4.3.2. Etapa de entrenamiento

Para la implementación de la GAN, se utiliza una estructura denominada DF. Esta estructura consiste de un codificador y un modelo intermedio con pesos compartidos entre los datos de origen y destino. Le siguen dos decodificadores que pertenecen a los datos de origen y destino separadamente.

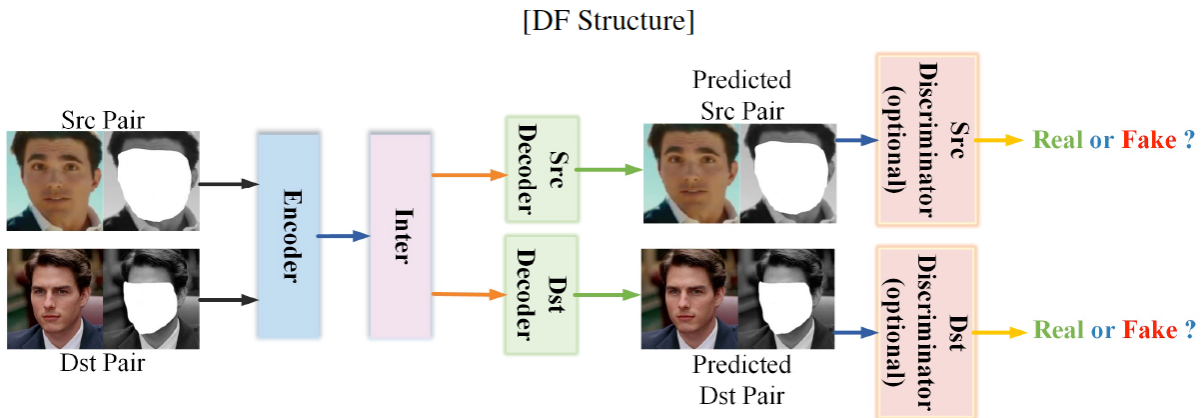

Figura 10: Estructura GAN DF

Fuente: [20]

Como funciones de pérdida, se utiliza un híbrido entre la función SSIM (Structural Similarity Index) y MSE (Mean Squared Error). El MSE corresponde al error cuadrático medio de los valores de los pixeles entre dos imágenes, mientras que el SSIM calcula un peso derivado de tres componentes de cada imagen: contraste, similaridad estructural y luminancia. A continuación, se presenta la fórmula de la misma:

$$S(\mathbf{a},\mathbf{b}) = \left(\frac{2\mu_a\mu_b}{\mu_a^2+\mu_b^2}\right)^{\beta}\left(\frac{2\sigma_a\sigma_b}{\sigma_a^2+\sigma_b^2}\right)^{\alpha}\left(\frac{\sigma_{ab}}{\sigma_a\sigma_b}\right)^{\gamma}$$

Figura 11 : Función de error SSIM
Fuente: [21]

Donde, los tres componente de *S,* corresponden a luminancia, contraste y similaridad estructural entre dos imágenes. Además, μ, σ y σ_{ab} corresponden a la media, desviación estándar y covariancia respectivamente. “La motivación para esta combinación es el obtener beneficios de ambas funciones: SSIM generaliza los rostros humanos mejor, mientras MSE provee de mejor claridad. Esta combinación de funciones de pérdida sirve para encontrar un compromiso entre generalización y claridad [20].”

4.3.3. Etapa de conversión

Finalmente, la etapa de conversión es en donde se utiliza el modelo para intercambiar el rostro desde el origen al destino o viceversa. Para iniciar, se transforman el rostro creado por el generador junto a la máscara de segmentación desde el decodificador de los datos de destino a la posición original de la imagen deseada en los datos de origen.

A continuación, sigue la mezcla entre el rostro recreado para que encaje completamente con el rostro de destino. “Para mantener la consistencia de la complexión, DeepFaceLab provee de cinco algoritmos de transferencia de color (Reinhard color Transfer. RCT, iterative distribution transfer, IDT, etc) [20].” Para este caso se utilizó el algoritmo de transferencia de color RCT. Este algoritmo permite aproximar el color del rostro recreado al rostro destinado.

Por último, se realiza un proceso de agudizamiento de la imagen utilizando una red neuronal pre entrenada de súper resolución, ya que los rostros recreados mediante los métodos de intercambio de rostros se suavizan y se pierden detalles del mismo.

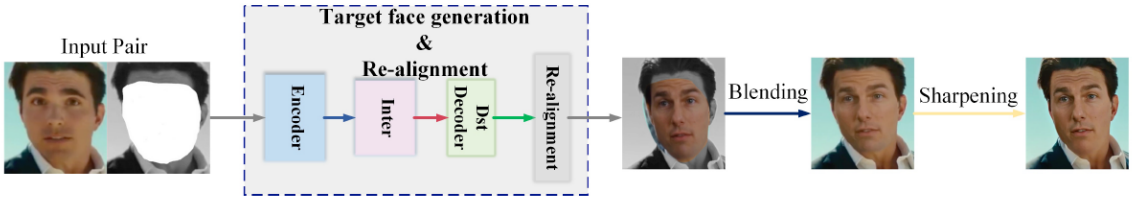

Figure 4. Overview of conversion phase in DeepFaceLab(DFL).

Figura 12: Etapa de conversión
Fuente: [20]

5. Resultados y discusión

Para poder realizar una evaluación apropiada del modelo, anteriormente se mencionó que se introdujo un módulo en la adaptación del proyecto DeepFaceLab, el cual consistía en registrar los valores de las funciones de pérdida tanto del generador, como del discriminador. A continuación se presentan los resultados de los errores en un gráfico loss vs epochs:

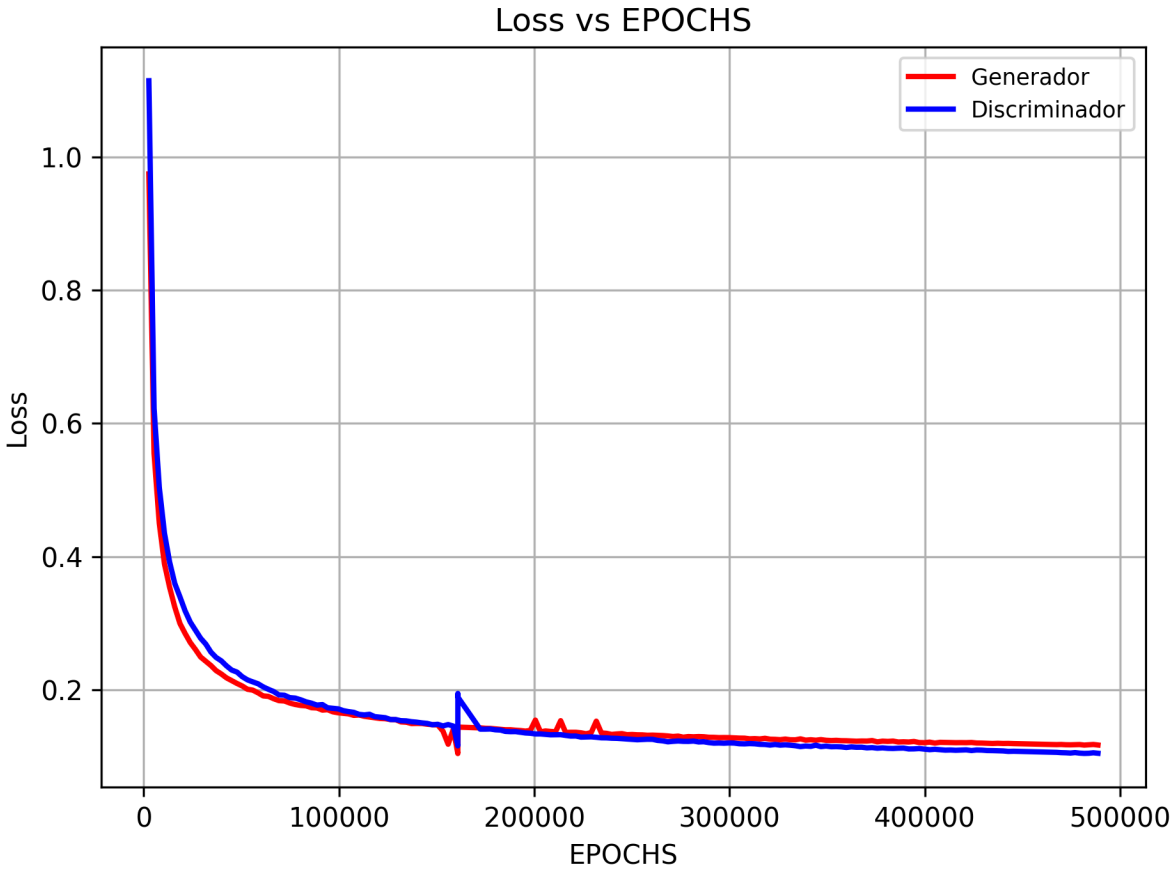

Figura 13: Loss vs Epochs modelo DeepFaceLab

Tal y como se puede apreciar, durante los primeras épocas de entrenamiento del modelo, se obtiene un error que se aproxima o hasta excede el valor de 1.0. Esto se debe gracias a que el modelo de redes se inicializa con ruido en sus primeras iteraciones, hasta que el modelo generativo vaya aproximándose a las imágenes suplementadas en la etapa de extracción. Alrededor de las diez mil épocas, el modelo va reduciendo su error hasta aproximadamente un valor de 0.4. Es en este punto en donde se detuvo el entrenamiento, y se procedió a realizar el proceso de conversión para verificar visualmente el rendimiento del modelo. A continuación, se puede apreciar una imagen del rostro de destino en contraste con el rostro tras el proceso de conversión:

Figura 14: Datos de destino vs 10K epochs

Las imágenes son tomadas del mismo segundo del video de destino, y se puede apreciar claramente que el rostro generado por la red generadora omite muchos detalles que hacen que el rostro tenga una apariencia de suavizado. Hasta podría decirse que el rostro podría aparentar una menor resolución que la resolución de 512x512 que se utilizó en la etapa de extracción de los rostros. Tras continuar el proceso de entrenamiento hasta las cien mil épocas, se puede apreciar en la figura 13, que los valores de las funciones de error van por debajo del valor de 0.2, y el resultado se puede visualizar en la siguiente figura:

Figura 15: Datos de destino vs 100K epochs

Al igual que en la figura 14, las capturas fueron tomadas del mismo segundo en el video. Y se puede apreciar que la red generativa va captando más detalles y creando rostros cada vez más realistas. Además, existe una correlación entre la reducción del error y la mejora del rostro generado. A continuación, se presentan en la figura 16 los resultados tras haber detenido el modelo tras quinientas mil épocas:

Figura 16: Datos de destino vs 500k epochs

En la figura 13 se puede apreciar que tras quinientas mil épocas de entrenamiento, el modelo tiene un error aproximado de 0.1. En este punto, el rostro creado por la red no tiene una mejora sustancial como fue el cambio de diez mil épocas a cien mil épocas. Y es evidente, ya que la reducción del error pasó de aproximadamente 0.2 a aproximadamente 0.1. La correlación es notable, y las mejores a partir de este punto no van a ser sustanciales. Y, cabe mencionar, que el error del modelo tiene reducciones mínimas, y el comportamiento de la curva se mantiene igual hasta llegar al millón de épocas. Finalmente, se muestran los resultados tras terminar el entrenamiento:

Figura 17: Datos de destino vs 1M epochs

Tras entrenar por un millón de épocas, los resultados visuales son evidentes. Se logró con éxito implementar un deepfake de tipo reemplazo de rostro. La cantidad de épocas de entrenamiento tuvo una influencia importante para poder llegar a este resultado, ya que a diferencia de la figura 16 , las épocas se duplicaron y el modelo pudo captar los detalles del rostro del video de origen. Igualmente, vale destacar que el rostro superpuesto es creado a partir del modelo generativo de las GANs. Conjuntamente, la etapa de entrenamiento es la de mayor duración de todo el proceso, ya que se debe entrenar el modelo durante varias épocas para obtener resultados híper realistas de

calidad. Para este caso en específico, cada época tarda en entrenar en un rango de [450 – 580] mili segundos. Por lo tanto para llegar a entrenar el modelo durante un millón de épocas, se necesitó de aproximadamente de 6.3 días de entrenamiento. Además, los desarrolladores del proyecto DeepFaceLab recomiendan entrenar el modelo en el rango de quinientas mil épocas a un millón de épocas. Como se puede verificar con la figura 17, en este caso, fue necesario entrenar el modelo durante el millón de épocas para obtener los resultados deseados.

Para revisar los resultados en su totalidad, verificar el ANEXO A.

CONCLUSIONES Y RECOMENDACIONES

El objetivo de este proyecto tiene como enfoque la implementación de un deepfake del tipo reemplazo de rostro mediante redes generativas antagónicas para poner en contexto distintas implicaciones sociales. Se puede concluir que se pudo implementar el deepfake con éxito, tal y como se puede verificar en el ANEXO A.

Se pudieron conseguir resultados híper realistas al realizar el reemplazo de un rostro desde un video de origen hacia un rostro en un video de destino. Pero, cabe recalcar que los resultados no fueron impecables a pesar de haber entrenado el modelo GAN durante 1 millón de épocas tal y como recomiendan los desarrolladores del proyecto DeepFaceLab. Las imperfecciones se pueden apreciar en el archivo *result_1M.mp4* en el ANEXO A. Ciertas secciones de sombras y cabello de la parte superior del rostro de origen también fueron creadas en los rostros generados por las GANs. Estos detalles pueden ser apreciados en el deepfake resultante, ya que la red generativa incluyó estos detalles externos al rostro en la superposición sobre el rostro de destino. Por lo tanto, se recomienda que los materiales visuales, tanto de origen como de destino, tengan condiciones de iluminación lo más cercanas posibles, ya que las redes utilizarán todos los detalles que se encuentren en la superficie del rostro, esto incluye: cabello, luz, vello facial, sombras, entre otros. Por lo tanto, se debe utilizar material que contenga la zona del rostro lo más despejada posible.

Cabe recalcar que el tiempo de entrenamiento del modelo total fue de 6.3 días, con un tiempo por época en un rango de [450 – 580] mili segundos. El tiempo para producir un deepfake híper realista es extenso, tomando en cuenta que se utilizó una unidad de procesamiento gráfico de última

generación y alta gama. Esto implica, que usuarios que buscan resultados similares van a tener que emplear una suma considerable de recursos y tiempo para obtener resultados similares a los del experimento. Por lo tanto, la producción a gran escala de deepfakes aún no es posible por las limitaciones a la accesibilidad de hardware de alta gama para el público en general. Esto implica que los investigadores y desarrolladores que buscan métodos de prevención se encuentran a tiempo de seguir implementando soluciones que puedan mejorar la detección de deepfakes. Y proyectos como DeepFaceLab cumplen su objetivo en proporcionar datos falsos de calidad para el desarrollo de estas investigaciones

También es importante mencionar que, a pesar de todas las limitaciones para la producción de un deepfake, los resultados son a tomar en cuenta. Lo que hace aún más relevante las cuestiones planteadas en la sección 3, ya que las amenazas son latentes. Todos las consecuencias, anteriormente plateadas, a nivel político, económico y social pueden seguir expandiéndose a medida que las arquitecturas GANs sigan refinándose y la Ley de Moore siga con su tendencia, como la ha hecho durante décadas. Esto hace suponer, que eventualmente, el continuo desarrollo del hardware y el refinamiento de estas herramientas, harán plausible que los deepfakes se vuelvan de fácil acceso, y usuarios puedan desarrollar un deepfake en un tiempo significativamente menor al que tomó este proyecto. Por lo tanto, es imprescindible que las investigaciones y métodos de detección sigan progresando a la par para prevenir escenarios perjudiciales para la sociedad.

Desde sus primeras implementaciones en 2017, los deepfakes han ido incorporando métodos cada vez más sofisticados que permiten resultados más impresionantes. Tal y como menciona Mika Westerlund: “Esto se está desarrollando más rápidamente de lo que pensaba. Pronto, va a llegar al

punto en el que no habrá forma de que podamos detectar (deepfakes), así que vamos a tener que buscar otro tipo de soluciones [5].” En conclusión, el desarrollo de estas herramientas produce resultados híper realistas, tal y como se pudo implementar en este proyecto, por lo tanto, se deberá tomar una serie de medidas en los campos de la investigación para prevenir amenazas que pueden generar graves consecuencias para la sociedad.

REFERENCIAS BIBLIOGRÁFICAS

[1] A. Abdulreda y A. Obai. 2022. "A landscape view of deepfake techniques and detection methods". *International Journal of Nonlinear Analysis and Applications* 13(1): 745-755.

[2] G. Bansal y M. Joshi. 2022. "Deepfake: A Systematic Review". *Kalyan Bharati* 36(8): 71-79.

[3] S. Rana, M. Nobi, B. Murali y A. Sung. 2022. "Deepfake Detection: A Systematic Literature Review". *IEEE Access* 10: 25494-25513.

[4] E. Altuncu, V. Franqueira y S. Li. 2022. "Deepfake: Definitions, Performance Metrics and Standards, Datasets and Benchmarks, and a Meta-Review". En *arxiv.* Cornell University, [documento digital]. Disponible en: https://arxiv.org [Accedido el 4 de Diciembre del 2022].

[5] M. Westerlund. 2019. "The Emergence of Deepfake Technology: A Review". Technology Innovation Management Review 9(11): 39-52.

[6] R. Astor, K. Costello, J. DeOliveira y A. Lewis. 2022. "Building a Customizable GAN Package Tool in Python. Helping Users Make and Test GANs in an Easy to Use Format". Tesis de pregrado. Worcester Polytechnic Institute. Worcester, Massachusetts.

[7] D. Bau, J. Zhu, H. Strobelt, B. Zhou, J. Tenenbaum, W. Freeman y A. Torralba. 2018. "GAN Dissection: Visualizing and Understanding Generative Adversarial Networks". En *arxiv.* Cornell University, [documento digital]. Disponible en: https://arxiv.org [Accedido el 4 de Diciembre del 2022].

[8] A. Aggarwal, M. Mitta y G. Battineni. 2021. "Generative adversarial network: An overview of theory and applications". International Journal of Information Management Data Insight: 1-9.

[9] I. Goodfellow, J. Pouget-Abadie, M. Mirzra, B. Xu, D. Warde-Farley, S. Ozair, A. Courville y Y. Bengio. 2020. “Generative Adversarial Networks”. Communications of the ACM 63(11): 139-144.

[10] B. Ghojogh, A. Ghodsi, F. Karray y M. Crowley. 2021. “Generative Adversarial Networks and Adversarial Autoencoders: Tutorial and Survey”. En *arxiv.* Cornell University, [documento digital]. Disponible en: https://arxiv.org [Accedido el 4 de Diciembre del 2022].

[11] Z. Cai, Z. Xiong, H. Xu, P. Wang y W. Li. 2021. “Generative Adversarial Networks: A Survey Toward Private and Secure Applications”. *ACM Computing Surveys* 54(6): 1-31.

[12] Z. Pan, W. Yu, X. Yi, A. Khan, F. Yuan y Y. Zheng. 2019. “Recent Progress on Generative Adversarial Networks (GANs): A Survey”. IEEE Access 7: 36332-36333.

[13] A. Gupta, J. Johnson, L. Fei-Fei, S. Sacarese y A. Alahi. 2018. “Social GAN: Socially Acceptable Trajectories with Generative Adversarial Networks”. En *arxiv.* Cornell University, [documento digital]. Disponible en: https://arxiv.org [Accedido el 4 de Diciembre del 2022].

[14] T. Karras, M. Aittala, J. Hellsten, S. Laine, J. Kehtinen y T. Aila. 2020. “Training Generative Adversarial Networks with Limited Data”. Presentado en la 34ª Conferencia de Sistemas de Procesamiento de Información Neuronal (NeurIPS 2020), Vancouver, Canadá.

[15] V. Jones. 2020. “Artificial Intelligence Enabled – Deepfake Technology. The Emergence of a New Threat”. Tesis de M.Sc. Faculty of Utica College. Utica, New York.

[16] E. Temir. 2020. “Deepfake: New Era in The Age of Disinformation & End of Reliable Journalism”. Journal of Selcuk Communication 13(2): 1009-1024.

[17] E. Hine y L. Floridi. 2022. “New deepfake regulations in China are a tool for social stability, but at what cost?”. *Nature Machine Intelligence* 4: 608-610.

[18] F. Repez y M. Popescu. 2021. “Social Media and the Threats Against Human Security Deepfake and Fake News”. *Romanian Military Thinking International Scientific Conference:* 44-55.

[19] S. Buo. 2020. “The Emerging Threats of Deepfake Attacks and Countermeasures”. En *arxiv.* Cornell University, [documento digital]. Disponible en: https://arxiv.org [Accedido el 4 de Diciembre del 2022].

[20] I. Petrov, D. Gao, N. Chervoniy, K. Liu, S. Marangonda, C. Umé, J. Jiang y L. RP. 2021. “DeepFaceLab: Integrated, flexible and extensible face-swapping framework”. En *arxiv*, Cornell University, [document digital]. Disponible en: https://arxiv.org [Accedido el 4 de Diciembre del 2022].

[21] A. Loza, L Mihaylova, N. Canagarajah y D. Bull. 2006. “Structural Similarity-Based Object Tracking in Video Sequences”. En *ResearchGate*, [documento digital]. Disponible en: https://www.researchgate.net [Accedido el 4 de Diciembre del 2022].

[22] Y. Haik y T. M. Shahin, *Engineering Design Process.* Stamford, CT: Global Engineering, 2011.

ANEXO A: RESULTADOS DEEPFAKES

En el siguiente enlace se encuentra los archivos mp4 con los deepfakes resultantes.

ANEXO B: DATOS DE ORIGEN

En el siguiente enlace se encuentra el archivo mp4 que contiene los datos de origen utilizados para la implementación del deepfake.

ANEXO C: DATOS DE DESTINO

En el siguiente enlace se encuentra el archivo mp4 que contiene los datos de destino utilizados para la implementación del deepfake.

ANEXO D: CÓDIGO

En el siguiente enlace se encuentra el repositorio en GitHub con la implementación del código utilizado para la elaboración del proyecto.