Entregable 1.
Métodos generadores y detectores de Deepfake

Estudiantes:

Bayron Daymiro Campaz Hurtado

Juan David Diaz Monsalve

Santiago Gutierrez Bolaños

Tutor:

Christian Camilo Urcuquí Lopez, Msc

Universidad ICESI
Facultad de Ingeniería
Ingeniería de Sistemas
Cali

2020

INTRODUCCIÓN

Actualmente, existen métodos y herramientas para la generación de imágenes y videos no reales basados en inteligencia artificial, entre estos podemos encontrar desde herramientas open source tales como Face2Face que permiten hacer deepfake en vivo hasta aplicaciones de escritorios como Faceswap que permiten hacer un intercambio de caras entre dos videos. A continuación, una descripción de cada uno de los métodos y/o herramientas más conocidas y un cuadro comparativo entre ellas.

Métodos generadores de Deepfake

Actualmente, existen métodos y herramientas para la generación de imágenes y vídeos no reales basados en inteligencia artificial, entre estos podemos encontrar desde herramientas de código abierto tales como Face2Face que permiten hacer Deepfake en vivo hasta aplicaciones de escritorios como Faceswap que permiten hacer un intercambio de caras entre dos vídeos. A continuación, una descripción de cada uno de los métodos y/o herramientas más conocidas y un cuadro comparativo entre ellas.

Faceswap

Es un programa para escritorio que usa una red neuronal (NN) para intercambiar un rostro original por un rostro de intercambio. Para hacer esto cuenta con:

- **Codificador compartido:** Que se entrena con dos conjuntos de rostros; conjunto A y B. Conjunto A (las caras originales que queremos reemplazar) y el conjunto B (las caras de intercambio que deseamos colocar en el original). Compartiendo el codificador se logra que este genere un algoritmo único para ambos conjuntos de rostros.

- **Decodificadores diferentes:** El decodificador A se encarga de reconstruir los rostros de conjunto A y de igual forma trabaja el decodificador B. Cuando se quiere hacer el intercambio de rostro (luego de entrenar al modelo con suficientes imágenes de cada conjunto), el decodificador A reconstruirá los rostros del conjunto B y el decodificador B reconstruirá los rostros del conjunto A. Dando como resultado una cara intercambiada. Lo anterior está esquematizado en la Figura 5.

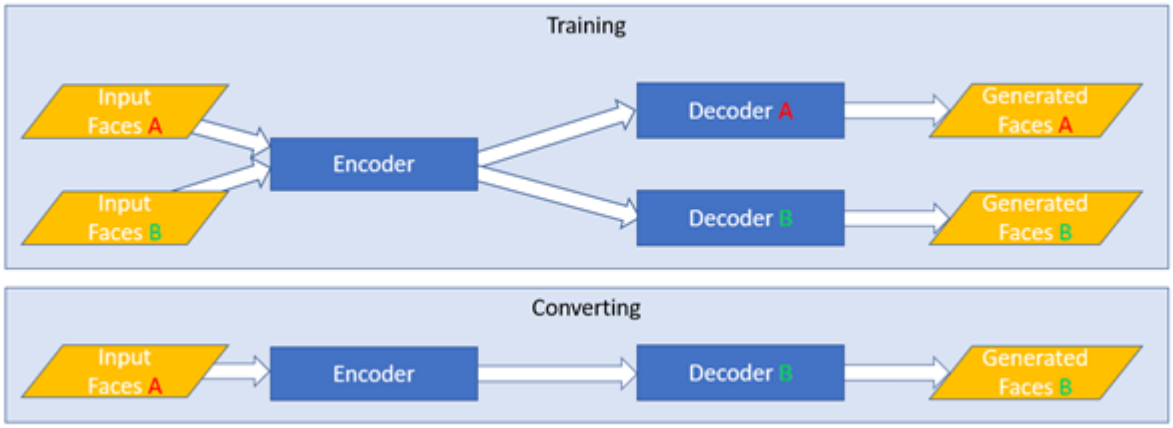

Figura 1. Esquema de funcionamiento de Faceswap.

Resulta importante mencionar que permite el uso de diferentes configuraciones de parámetros y modelo. El anteriormente descrito es el modelo original cuya entrada y salida son imágenes de rostros de 64 x 64 píxeles, sin embargo, existen modelos más sofisticados, que permiten entradas y producen salidas en diversas resoluciones.

Face2Face

Face2Face es una herramienta de código abierto que opera a través de dos entradas, en primer lugar, el sistema procesa una transmisión de vídeo en vivo a través de una cámara web del sujeto a reemplazar, en segundo lugar, un vídeo en el que este presente el rostro del objetivo (vídeo del sujeto que se quiere suplantar) utilizado. Las imágenes de los rostros se sintetizan usando un modelo multi-linear de rostros y un modelo de transformación rígida.

El anterior método permite encontrar parámetros de los rostros, que posteriormente se someterán a la aplicación de un método de minimización de energía variacional que permitirá optimizar los parámetros. Para minimizar esta energía se usa un solucionador de mínimos cuadrados iterativos reponderados iterativamente (Iteratively Reweighted Least Squares IRLS).

Lo anterior permite obtener la identidad facial entre otros datos relevantes de los actores fuente y objetivo. En ejecución las animaciones se reconstruyen usando un seguimiento a cuadro con una formulación energética similar a la anterior. Para la recreación se usa una deformación rápida que opera en el estadístico usado anteriormente.

FakeAPP

Es una aplicación de escritorio que permite crear vídeos Deepfake. Se caracteriza por usar un modelo de red neuronal profunda (DNN). Las representaciones comprimidas o vectores latentes que los autoencoders convolucionales actuales aprenden, son la

piedra angular detrás de las capacidades de intercambio de caras de esta aplicación (Yuezun Li, 2019).

FaceApp

Es una aplicación para dispositivos móviles. Usa una foto del rostro de una persona y posteriormente se le realiza la edición que se seleccione. La versión gratuita ofrece diferentes ediciones como, aplicar una sonrisa básica, añadir una barba básica, realizar una mezcla de fotos de rostro, aplicar un aumento de edad, cambiar el color del cabello, agregar accesorios, entre otras. La versión Pro ofrece herramientas de edición más potentes y al ser una aplicación comercial no es de código abierto.

Faceswap-GAN

Este método está basado en el uso de redes generativas antagónicas (GAN), específicamente en una variación denominada Red Adversaria Generativa de Auto-Atención SAGAN). Este tipo de red introduce un mecanismo de auto atención en las GAN convolucionales. El módulo de auto-atención es complementario a las convoluciones y ayuda a modelar dependencias de múltiples niveles a largo plazo en las regiones de la imagen. Armado con atención propia, el generador puede dibujar imágenes en las que los detalles finos en cada ubicación se coordinan cuidadosamente con los detalles finos en partes distantes de la imagen. (Han Zhang, 2019)

Visto de manera general la arquitectura del modelo de Faceswap-GAN se compone principalmente de 4 partes: un codificador, un decodificador, una red generativa y una red discriminadora (Shaoanlu, 2019).

Por otro lado, dentro de las características importantes de este método tenemos:

- La pérdida perceptual de cara VGG que mejora la dirección de los globos oculares permitiendo al Deepfake ser más realista y consistente con la cara de entrada.
- Una resolución de salida configurable. El modelo admite resoluciones de salida de 64x64, 128x128 y 256x256.
- Alineación de rostros usando MTCNN y filtro de Kalman en conversión de vídeo. Se introduce MTCNN para detecciones más estables y una alineación facial confiable.

Style-Based Generator Architecture GAN.

Este método está basado en una red generativa antagónica con una arquitectura basada en transferencia de estilos. La transferencia de estilos se basa en representar el contenido de una imagen en el estilo de otra. Esta arquitectura conduce a una

separación automática y sin supervisión de los atributos de alto nivel (por ejemplo, pose e identidad cuando se entrena en rostros humanos) y a la variación estocástica en las imágenes generadas (por ejemplo pecas, cabello) (Tero Karras, 2019).

El generador parte de una entrada constante aprendida y ajusta el "estilo" de la imagen en cada capa de convolución en función del código latente, por lo tanto, controla directamente la fuerza de las características de la imagen a diferentes escalas. Combinado con el ruido inyectado directamente en la red, este cambio arquitectónico es el que conduce a la separación automática y sin supervisión de los atributos de alto nivel de la variación estocástica en las imágenes generadas, permitiendo una mezcla específica a escala intuitiva, así como operaciones de interpolación.

Cabe resaltar que esta herramienta no solo sirve para generar Deepfakes sino que también permite “transferencia de estilo” entre diversos tipos de imágenes ya sea que estas contengan objetos, animales o cualquier tipo de representación de forma visual. El modelo bajo el cual se basa este método permite que haya un entrenamiento con cualquier tipo de resolución. Sin embargo, este método solo usa como ejemplo un modelo pre entrenado para imágenes de 1024 x 1024 píxeles.

Tabla 1. Comparación entre métodos generadores.

Método/ Característica	**Faceswap**	**Face2Face**	**FakeAPP**	**Faceswap-GAN**	**Style-Based Generator Architecture GAN.**	**FaceAPP**
Tipo de dato de entrada	Conjunto de imágenes en formato PNG y/o JPG	Transmisión de vídeo en vivo a través de una cámara web y vídeo monocular	Vídeo fuente y objetivo en cualquier formato de vídeo	Vídeo fuente y vídeo objetivo en formato MP4	Conjunto de imágenes en formato PNG	Imagen en formato JPG o PNG
Resolución de dato de entrada	Cualquier resolución	1280 x 720 (Vídeo monocular) 640 x 480 (Vídeo en vivo)	Cualquier resolución	Cualquier resolución	1024 x 1024 píxeles	Cualquier resolución
Tipo de dato de salida	Imagen en formato PNG o vídeo en MP4 que refleja el intercambio de un rostro	Vídeo en tiempo real del objetivo con los gestos de la fuente	Vídeo y conjunto de imágenes con el rostro del vídeo objetivo en el vídeo fuente	Vídeo en formato MP4 que refleja el intercambio de rostro	Imagen en formato PNG que refleja la transferencia de estilo	Imagen en formato PNG o JPG con una modificación facial escogida
Resolución de dato salida	La misma que la resolución del vídeo o imagen ingresada	1280 x 720 píxeles	La misma que el vídeo objetivo	64 x 64 128 x 128 256 x 256 pixeles	1024 x 1024 píxeles	La misma que la resolución de la imagen ingresada

Tipo de modelo	Red Neuronal	Consta de dos modelos: un modelo multi-linear de rostros y un modelo de transformación rígida	No disponible	Red Generativa Antagónica (GAN) + Auto encoder	Red Generativa Antagónica (GAN) + normalización de instancia adaptativa (AdaIN)	No disponible
Uso	Se encuentra integrado en una aplicación de escritorio	No aplica	Es en si una aplicación de escritorio	Se encuentra el código fuente organizado en Notebooks que permite realizar Deepfakes a través del navegador con el uso de la plataforma Colab de Google	Se encuentra el código fuente del modelo con instancias pre-entrenadas del modelo. Se puede usar para transferencia de estilos entre 2 imágenes	No aplica
Acceso a código fuente (Open Source)	Si	No	No	Si	Si	No

MÉTODOS DETECTORES DE DEEPFAKE

Two-Stream Neural Networks for Tampered Face Detection

En este proyecto se propone una red de dos flujos para la detección de alteraciones faciales. Entrenan una red mediante el servicio de Google GoogLeNet para detectar artefactos de manipulación en una secuencia de clasificación de caras, se entrena esta red con el fin de aprovechar las características que capturan los residuos de ruido local y las características de la cámara como una segunda secuencia. Además, utilizan dos aplicaciones diferentes de intercambio de caras para crear un nuevo conjunto de datos que consta de imágenes manipuladas. Este método no puede detectar caras manipuladas muy pequeñas, esto se debe a que la secuencia de clasificación de caras necesita cambiar el tamaño de la cara de entrada a 299 × 299, y en el muestreo de caras pequeñas se pierde información visual crucial para la detección de alteraciones.

FWA: Exposing Deepfake Vídeos By Detecting Face Warping Artifacts.

En este trabajo, se describe un método basado en Deep Learning que puede distinguir los vídeos falsos generados por IA de los vídeos reales. Este método se basa en que los algoritmos de Deepfake deben realizar unas transformaciones que dejan elementos distintivos en los vídeos resultantes, y muestran que estos pueden ser capturados por redes neuronales convolucionales (CNN). Tiene la ventaja que en comparación con otros métodos no necesita imágenes generadas por Deepfake para el entrenamiento.

Mesonet: a compact facial vídeo forgery detection network.

Este método busca detectar de manera automática y eficiente la manipulación de la cara en vídeos. Sigue un enfoque de aprendizaje profundo y presenta dos redes, ambas con un bajo número de capas, para enfocarse en las propiedades mesoscópicas de las imágenes. Se evalúan esas redes tanto en un conjunto de datos existente como en un conjunto de datos que ellos han constituido a partir de vídeos en línea. Las pruebas demuestran una tasa de detección muy exitosa con más del 98% para Deepfake y 95% para Face2Face para tamaños de imágenes 256 x 256 píxeles.

HeadPose: Exposing deep fakes using inconsistent head poses.

Este método busca detectar imágenes o vídeos de caras falsas generadas por Deepfake. Se basa en las observaciones de que el contenido alterado se genera al empalmar la región de la cara sintetizada en la imagen original y, al hacerlo, se introducen errores que pueden revelarse cuando se estiman las posturas de la cabeza en 3D a partir de las imágenes de la cara. Se realizan experimentos para demostrar este fenómeno y usando características basadas en este indicio, se evalúa un clasificador SVM (Máquinas de Vectores de Soporte) usando un conjunto de imágenes de caras reales y falsificadas.

Multi-task Learning For Detecting and Segmenting Manipulated Facial Images and Vídeos

En este proyecto diseñan una red neuronal convolucional que utiliza el enfoque de aprendizaje de tareas múltiples para detectar simultáneamente imágenes y vídeos manipulados y ubicar las regiones manipuladas. La información obtenida al realizar una tarea se comparte con la otra tarea y, por lo tanto, mejora el rendimiento de ambas tareas. Se utiliza un enfoque de aprendizaje semi-supervisado. La red incluye un codificador y un decodificador en forma de Y. La activación de las características codificadas se utiliza para la clasificación binaria. La salida de una rama del decodificador se usa para segmentar las regiones manipuladas, mientras que la de la otra rama se usa para reconstruir la entrada, lo que ayuda a mejorar el rendimiento general.

En la Tabla 2, se realiza un resumen de lo que se encuentra en el estado del arte. Las columnas representan los trabajos discutidos y en la última (izquierda a derecha) se incluye el trabajo realizado en este proyecto, por último, las filas representan los criterios de comparación entre cada uno de los trabajos.

Tabla 2. Tabla comparativa de métodos de detección.

Método/ características	**Two-Stream**	**FWA**	**Mesonet-4**	**HeadPose**	**Multitask**	**Proyecto de Grado**
Tecnología usada	Red neuronal profunda GoogLeNet	Red neuronal convolucional (ResNet50)	Red neuronal convolucional	Modelos matemáticos para estimar la postura de la cabeza en 3D a partir de imágenes de la cara	Red neuronal convolucional	Red Neuronal Convolucional Xception + Imagenet

Número de imágenes usadas en el entrenamiento	705 alterada + 1.400 reales	15.185 alteradas + 15.185 reales	5.111 alteradas + 7.250 reales	10.847 alteradas + 10.847 reales	218.179 alteradas + 218.179 reales	4.073 alteradas + 2.597 reales
Conjunto de datos de entrenamiento	Datos generados con SwapMe	Datos recopilados por los autores en distintas plataformas de vídeo	Datos recopilados por los autores en distintas plataformas de vídeo	Conjunto de datos UADFV	Conjunto de datos FF-DF	Conjunto de datos FF-DF, DFD, DFDC, Celeb-DF
Conjunto de datos público	No	Si	No	Si	Si	Si
AUC mayor a 80% sobre el conjunto de datos de PRIMERA generación.	No	Si	No	No	No	Si
AUC mayor a 80% sobre el conjunto de datos de SEGUNDA generación.	No	No	No	No	No	Si
Acceso a código fuente (Open Source)	No	Si	Si	Si	Si	Si

CONCLUSION

Este documento permite vislumbrar que existen diferentes métodos de generación y de detección de Deepfake que emplean diferentes técnicas y tecnologías, reciben y producen diferentes tamaños y tipos de contenido (algunos videos otras imágenes), algunos son open source. Hablando específicamente de los métodos detectores, son entrenados con distintas cantidades de datos.

Resulta relevante destacar que las redes neuronales son ampliamente usadas tanto en generación como en detección; ya sean redes neuronales clásicas, profundas o GAN. Pudimos evidenciar que los métodos detectores no funcionan de manera adecuada cuando son puestos a prueba con Deepfake de alta resolución, por lo que es necesario trabajar en métodos que mejoren la detección para este tipo de resolución.

REFERENCIAS

Github. (3 de octubre de 2019). Obtenido de Github: https://github.com/shaoanlu/faceswap-GAN

Han Zhang, I. G. (14 de junio de 2019). *Self-Attention Generative Adversarial Networks.* Obtenido de Cornell University: https://arxiv.org/abs/1805.08318

Tero Karras, S. L. (29 de marzo de 2019). *A Style-Based Generator Architecture for Generative Adversarial Networks.* Obtenido de Cornell University: https://arxiv.org/abs/1812.04948

Li, Y., Yang, X., Sun, P., Qi, H., & Lyu, S. (2019). *Celeb-DF: A New Dataset for DeepFake Forensics*. *1*, 1–6. Obtenido de http://arxiv.org/abs/1909.12962

Thies, J., Zollhöfer, M., Stamminger, M., Theobalt, C., & Nießner, M. (2018). Face2Face: real-time face capture and reenactment of RGB videos. Obtenido de https://niessnerlab.org/papers/2019/8facetoface/thies2018face.pdf

torzdf. (19 de septiembre de 2019). *Training in Faceswap.* Obtenido de forum.faceswap.dev: https://forum.faceswap.dev/viewtopic.php?t=146

torzdf. (2019) NN de Faceswap [Figura 1]. Recuperado de https://forum.Faceswap.dev/viewtopic.php?t=146