

Sistemas Inteligentes para Gestión de la Empresa



# UNIVERSIDAD DE GRANADA

GENERACIÓN DE “DEEPFAKES”

**Autor**

Juan Manuel Castillo Nievas



MÁSTER PROFESIONAL EN INGENIERÍA INFORMÁTICA 2020-2021

—  
Granada, 28 de mayo de 2021

# Índice

<b>1. Introducción</b>	<b>2</b>
1.1. ¿Qué son las <i>fake news</i> ? . . . . .	2
1.2. Ejemplo de difusión . . . . .	2
1.3. Tipos de <i>fake news</i> . . . . .	4
<b>2. Deepfakes</b>	<b>7</b>
2.1. Tipos de manipulación . . . . .	8
2.1.1. Síntesis facial completa . . . . .	8
2.1.2. Intercambio de identidad . . . . .	11
2.1.3. Manipulación de atributos faciales . . . . .	13
2.1.4. Intercambio de expresión facial . . . . .	15
2.2. Casos reales de deepfakes . . . . .	15
2.2.1. Discurso de Obama . . . . .	16
2.2.2. Donald Trump se une a la serie Breaking Bad . . . . .	16
2.2.3. Remasterización de la Princesa Leia . . . . .	17
<b>3. Caso práctico</b>	<b>17</b>
<b>4. Bibliografía</b>	<b>21</b>

# 1. Introducción

En este documento se va a presentar una introducción a lo que son las conocidas como *fake news*. La finalidad de este documento es investigar sobre el término más actual de *deepfake*, muy relacionado con las *fake news*. Se van a discutir los distintos tipos de *fake news* que existen, poniendo un ejemplo de cada uno para entender todo de manera más clara, para posteriormente pasar a ver qué son las *deepfake* y cuáles son las **técnicas más utilizadas** para ello. También se verán algunos de los **casos reales** más sonados. Finalmente, se hará una pequeña **demonstración** utilizando una técnica *deepfake* para conocer su utilidad.

## 1.1. ¿Qué son las *fake news*?

Las técnicas de manipulación de imágenes permiten actualmente manipular imágenes y vídeos que **alteran o cambian la identidad de una persona**. Si bien estas técnicas y herramientas se utilizan como soporte para algunas tareas dentro de la **inteligencia artificial** que ayudan al desarrollo e investigación de ellas, cada día son más las **noticias falsas** que se distribuyen a través de Internet ayudándose de estas herramientas.

Realmente esto es solo una técnica más para la **difusión de *fake news***. El término de *noticia falsa* cada vez es más familiar en nuestra sociedad. Una noticia falsa se difunde con un contenido **supuestamente auténtico**, pero cuyo fin es **alterar y escandalizar** la opinión pública mediante la manipulación [1]. Las personas que se encargan de difundir estas noticias falsas tienen principalmente 3 objetivos: **personales, políticos y/o económicos**.

A pesar de que este término es muy reciente, el concepto que engloba es algo que lleva mucho tiempo en la sociedad. Mismamente el **periódico**, uno de los medios de comunicación impresos más antiguos, se utilizaba en muchas ocasiones para difundir mentiras a través de titulares **incompletos o fuera de contexto** y artículos con información **no contrastada o sin fuente de información**.

Hoy en día el impacto es aún mayor, pues el impacto del aumento de medios de comunicación y de las redes sociales han propiciado el **auge de la difusión de noticias falsas**. No es un secreto que los medios de comunicación televisivos difunden noticias con ciertos **prejuicios** basados en el **fin político** de cada productora.

## 1.2. Ejemplo de difusión

Casi toda la población ha oído alguna vez esa frase de “**los humanos nos comemos alrededor de 8 arañas en un año mientras estamos dormidos**”. Pues bien, este

dato es completamente un bulo. Pero esta historia es curiosa ya que hay un bulo dentro de otro bulo.

Inicialmente se desmintió este bulo apuntando a que todo fue un experimento social realizado por la periodista **Lisa Holst** en 1993 con el objetivo de **demostrar lo fácil que es difundir noticias falsas** a través de redes sociales (en aquella época la única red social que existía era el correo electrónico, y realmente no es una red social). También tenía un segundo objetivo que era el **demostrar cómo la gente se cree las noticias** que ve sin contrastar si quiera la mínima fuente de información.

Esta explicación parecía creíble y de hecho hubo otra mucha gente que asumió y se creyó este experimento social. La verdad que se esconde detrás de todo esto es que es otro bulo. A día de hoy, aún **no se han encontrado evidencias ni pruebas** de que el artículo escrito por Lisa Holst exista [2]. Tampoco existe el libro al que supuestamente se referencia en dicho artículo, escrito supuestamente en 1954. Todo esto apunta a que ha sido un **doble bulo**.

En 2018, la cuenta de Twitter llamada **@pictoline** publicó una imagen acerca de este doble bulo [3], titulada como **Fool me twice** (ver Figura 1). En dicha imagen se explica de desmienten de manera breve ambos bulos.

A pesar de no existir el supuesto experimento social, sí que se ha convertido en uno de ellos, pues se ha demostrado que este bulo se ha dado a conocer mundialmente a través de las redes sociales e Internet y que hay gente que se cree la información sin ni siquiera haber contrastado su veracidad.



Figura 1: Desmintiendo el bulo de las arañas, por @pictoline [3]

### 1.3. Tipos de *fake news*

Dentro de las *fake news*, encontramos distintas categorías de las mismas dependiendo de cómo se transmita la información y el objetivo de ellas. Estas categorías pueden variar dependiendo del punto de vista de cada uno, pero una buena clasificación es la que se encuentra en la página web de **Webwise** [4]:

- **Clickbait:** Un término que lleva relativamente poco en la sociedad, pero que cada día es más sonado. El **clickbait** tiene como objetivo **captar la atención de los visitantes** incitando a que hagan **click** en su página web porque reciben bastantes ingresos por cada click debido a toda la publicidad que anuncian.

Una de las plataformas más famosas por el clickbait es **YouTube**, ya que los usuarios reciben más dinero si reciben más visitas. Esto ha hecho que los usuarios suban sus vídeos con titulares como los que se ven en la Figura 2a, que **incitan a los consumidores a hacer click** para ver el vídeo.

- **Desinformación deliberada:** Se trata de noticias que han sido creadas para un **grupo de personas susceptible de creerse la información** sin ni siquiera contrastar su veracidad, y contribuyen a la distribución de la noticia a través de las redes sociales. Normalmente estas noticias se mueven por **intereses mayoritariamente políticos**.

Un claro ejemplo de este tipo de *fake news* es la noticia que se hizo viral en 2016 en las **elecciones de Estados Unidos** [5]. En la Figura 2b se puede ver el titular de la noticia que dice que **“Obama no investigará a los votantes ilegales y Hillary los indultará después de las elecciones”**. Esta noticia es claramente falsa pero fue muy difundida a través de Internet con fines políticos exclusivamente.

- **Parodias:** Dentro de esta categoría se encuentra aquella información falsa que se usa con el objetivo de **entretenir** al público a través de la **ironía o parodia**. Lo malo de estas noticias es que se difunden de una manera mucho más rápida, pues la gente que realmente ha captado la ironía las comparte a modo de burla y risa y la gente que se las cree las comparte a través de sus redes sociales a modo de denuncia.

Un ejemplo muy famoso es la cuenta de Twitter de **@sanchezcasrejon** [6], la cual es una **cuenta parodia** del actual Presidente del Gobierno en España. A pesar de que sus twits son absolutamente falsos y se autodenomina como una cuenta parodia, sí que es cierto que la web **Maldita**, creada para desmentir bulos y demás, ha tenido que **desmentir** varios twits de esa cuenta porque estaban empezando a llegar a gente que creía la veracidad de la noticia.

El **24 de diciembre de 2018** se difundió un twit de esta cuenta y **Maldita** tuvo que desmentirlo ya que empezó a creerse entre la gente. Dicho twit se encuentra en la Figura 2c. A pesar de que en el nombre de la cuenta aparece explícitamente la

palabra **PARODIA**, aún sigue habiendo gente que se toma en serio todo lo que se publica.

- **Noticias con información no contrastada:** Este tipo de noticias se caracterizan por contener **información no fiable o no contrastar su contenido** con fuentes de información fiables y con una información certera.

Aprovechando la **campaña electoral** que se ha hecho en Madrid debido a las **elecciones del 4-M**, ha habido un caso de este tipo provocada por el ya conocido partido polémico **Vox**, el cual utiliza las *fake news* para generar debates políticos y ganar votos. Dicho partido publicó un anuncio a través de las distintas estaciones de metro en el que se decía que “**un mena cuesta a los españoles 4.700€ al mes**” (ver Figura 2d). Por supuesto que esta información fue rápidamente contrastada con fuentes de información fiables, pero curiosamente esos carteles **no se retiraron**.

- **Titulares falsos:** Son titulares que informan de **hechos completamente falsos** cuyo objetivo final es el de **captar la atención de los lectores**. Los titulares falsos son bastante recurridos en los medios de comunicación poco serios y normalmente se descubre que el titular es falso cuando se lee el contenido de la noticia.

En la Figura 2e se muestra un titular en el que supuestamente se dice que el **Papa Francisco apoyó a Donald Trump**, ex-presidente de los Estados Unidos, en las elecciones de 2016. Estas elecciones provocaron la creación de muchas noticias falsas ya que se dieron cuenta de que era un tema muy buscado y a través del cual podían ganar muchísimo dinero.

- **Noticias sesgadas:** A través de los **algoritmos de inteligencia artificial** a través de los cuales se nos recomiendan anuncios, productos, etc. en nuestras redes sociales, también se utilizan estos algoritmos para hacernos llegar noticias basadas en nuestras **creencias o gustos** sin el filtro de saber si realmente es una noticia falsa o no. Como las personas tendemos a fiarnos que todo lo que se nos recomienda, es muy fácil tender a creer que las noticias son siempre verdaderas, ya que por algo se nos han recomendado.



(a) Clickbait

(b) Desinformación deliberada

(c) Parodia



(d) Información no contrastada

Pope Francis Shocks World, Endorses Donald Trump for President, Releases Statement

TOPICS: Pope Francis Endorses Donald Trump



(e) Titular falso

Figura 2: Casos reales de *fake news*

## 2. Deepfakes

A través del concepto de *fake news*, en los últimos años ha aparecido el término de **Deepfake**. Concretamente, el término apareció por primera vez en 2017 [7] cuando un usuario de la web **Reddit** llamado “**deepfakes**” creó un algoritmo de aprendizaje automático que le ayudó a **recrear escenas pornográficas** sustituyendo las caras de los actores por caras de *celebrities*.

Por supuesto que esto sólo fue el inicio, y desde entonces se han utilizado estas técnicas para apoyar aún más al contenido de las *fake news*, manipulando imágenes y vídeos para hacer aún más creíble una determinada noticia.

Lo bueno de la aparición de estas técnicas es que no sólo se han utilizado para crear polémica entre los humanos, si no que han servido de gran ayuda para diversas creaciones e investigaciones. Por ejemplo, a principios de este año se hizo muy famoso el anuncio de **Cruzcampo** que tenía como lema “**Con mucho acento**” en el que aparecía la gran **Lola Flores** (ver Figura 3). Mucha gente se cuestionaba cómo era posible haber hecho ese anuncio con tanta precisión, pues realmente parecía que era **Lola Flores** de verdad quién había hecho ese anuncio. La realidad es que ese anuncio fue posible gracias a las **técnicas de manipulación facial** usando herramientas de *machine learning*.



Figura 3: Anuncio de Cruzcampo 2021

Aún más se dieron a conocer las *deepfakes* cuando aparecieron distintas aplicaciones móviles que utilizaban técnicas de inteligencia artificial para **rejuvenecer o envejecer el rostro** de una persona. Quizá una de las aplicaciones más sonadas fue **FaceApp**. Estas

aplicaciones permiten a cualquier persona **crear imágenes y videos falsos** sin ni siquiera tener una mínima experiencia en el campo de *machine learning* y técnicas de manipulación facial.

## 2.1. Tipos de manipulación

De acuerdo al estudio realizado en 2020 [7], hay 4 tipos de manipulación que están muy establecidas actualmente en la comunidad:

- **Síntesis facial completa**
- **Intercambio de identidad**
- **Manipulación de atributos faciales**
- **Intercambio de expresión facial**

### 2.1.1. Síntesis facial completa

En esta manipulación se crea un rostro facial completamente nuevo usando **Generative Adversarial Networks** (redes generativas antagónicas). La técnica que se está haciendo muy popular es **StyleGAN** [8]. Básicamente en esta arquitectura se aprende a separar atributos de alto nivel, así como la pose e identidad de los rostros, sin ninguna supervisión o variación en las imágenes generadas.

Las **componentes básicas de GAN** son dos redes neuronales: una que **genera nuevas caras** desde cero, y otra que funciona como un **discriminador** que predice si una cara es real o no basándose en las imágenes de entrenamiento y las salidas del generador [9].

La **entrada** del generador es un **vector aleatorio**, es decir, es sólo ruido, por lo tanto su salida es también ruido. A medida que va recibiendo las salidas del discriminador, **aprende a crear imágenes más realistas**. El discriminador también va aprendiendo a detectar imágenes reales e imágenes generadas. En la Figura 4 se puede ver una visión general de este proceso.

El problema que generó este proceso es que era complicado generar imágenes de gran calidad, como de 1024x1024. Entonces **NVIDIA** lanzó en 2018 **ProGAN**, cuyo funcionamiento se basaba en un **entrenamiento progresivo** empezando por imágenes con resoluciones muy bajas (4x4) y aumentando cada vez dicha resolución (ver Figura 5). De esta forma, primero se crea la base de la imagen y poco a poco se van aprendiendo nuevos detalles a medida que la resolución aumenta.

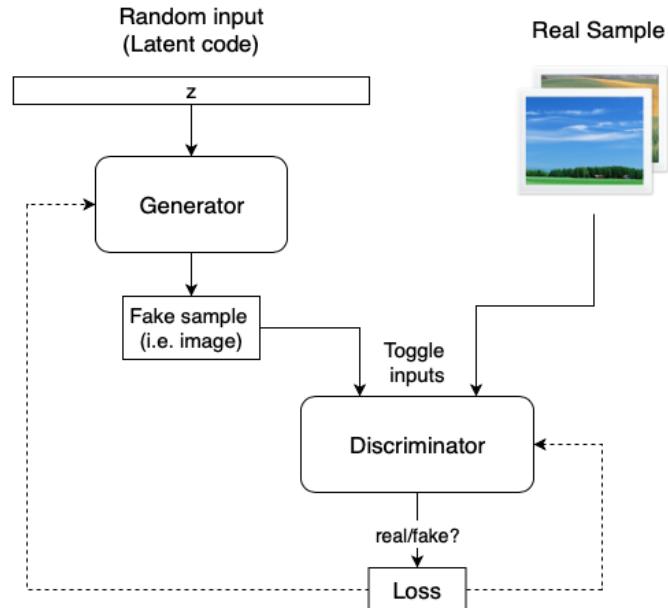


Figura 4: Visión general de las GANs

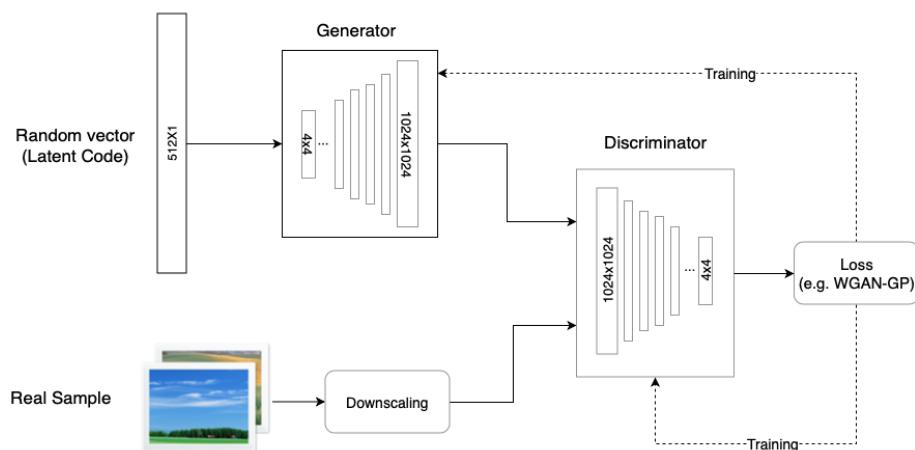


Figura 5: Visión general de las ProGANs

La aparición de **StyleGAN** viene propiciada de la limitación que tiene **ProGAN** a la hora de manipular algunas características faciales específicas, pues dichas características se confunden y a veces afectan a otras características al mismo tiempo. El principal beneficio de ProGAN es el **progreso a través de las capas**, permitiendo controlar diferentes aspectos visuales de la imagen. Algunas de las mejoras que se incorporan a **StyleGAN** son las siguientes [7]:

- **Mapping network:** el objetivo de esto es **codificar el vector inicial** en un vector intermediario en el cual los diferentes elementos controlan diferentes características. Esto se hace debido a que controlar las características visuales con únicamente el vector de entrada es un proceso muy limitado. Por ejemplo, si hay muchas imágenes de personas con pelo negro, habrá muchos elementos del mismo vector que se mapean a la misma característica, y habría una gran correlación.
- **Adaptive Instance Normalization (AdaIN):** este módulo **transfiere** la información codificada  $w$  de la *mapping network* a la imagen generada. Este módulo se utiliza en cada resolución que se genera.
- **Eliminar vector inicial:** los modelos como **ProGAN** usan el vector inicial para crear una imagen generada inicial. Con **StyleGAN** esto no es necesario ya que las características de las imágenes se controlan con el vector codificado  $w$  y AdaIN, con lo cual el **vector inicial se puede omitir**.
- **Variación estocástica:** en los generadores **GAN**, la manera de insertar características pequeñas como pecas, arrugas, etc. se hace añadiendo **ruido aleatorio** al vector inicial. En **StyleGAN** se añade un **ruido escalado** antes de pasar al módulo AdaIN y cambia un poco la expresión facial de la resolución actual.
- **Mezcla de estilos:** una forma de combinar diferentes imágenes de una forma coherente es a través de esta mejora. **StyleGAN** usa el vector  $w$  en cada nivel, y eso causa una **correlación** entre los niveles. Para reducir esta correlación, el modelo selecciona aleatoriamente dos vectores iniciales y genera un vector  $w$  para cada uno de ellos. Entonces, en algunos niveles se usa el primero y en otros se usa el segundo.

En la Figura 6 se muestra una imagen del proceso por el que pasa la generación de imágenes en **ProGAN**, que como ya se ha comentado parte de una imagen con una resolución muy baja y progresivamente va aumentando a resoluciones más altas. Como se puede apreciar en la imagen, este proceso **tarda demasiado tiempo**, tardando hasta 14 días en encontrar una imagen en alta resolución y con un rostro difícil de distinguir de uno real.

Por último, una imagen visual del modelo base de **StyleGAN** se puede ver en la Figura 7, en contraposición del modelo tradicional.

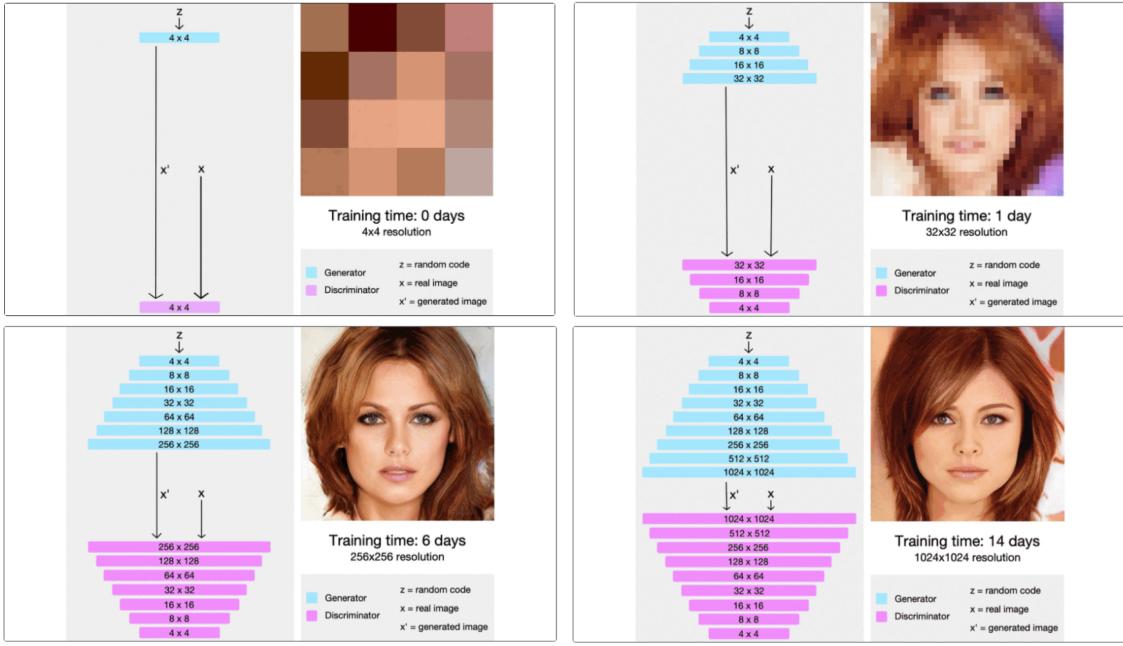


Figura 6: Ejemplo de entrenamiento con ProGAN

### 2.1.2. Intercambio de identidad

Esta manipulación consiste en **reemplazar** la cara de una persona en un vídeo con la cara de otra persona. Esta técnica es la que se utilizó en el anuncio de **Cruzcampo** que se ha mencionado anteriormente para reemplazar la cara de la actriz por la de Lola Flores. Sin embargo, esta técnica también puede ser utilizada de una mala manera, como la que se ha comentado al principio de esta sección en la que un usuario de **Reddit** sustituía la cara de actores y actrices en películas pornográficas por la cara de gente famosa.

En la **síntesis facial completa** se manipulaban las caras a **nivel de imagen**; en este tipo el objetivo final es generar imágenes o vídeos *fake* que sean **realistas**. En la Figura 8 se pueden ver varios ejemplos que mezclan a varias *celebrities* para manipular sus rostros.

Hay 3 pasos que están involucrados en el proceso de este tipo de manipulación facial [10]:

1. **Detección del rostro**
2. **Aprendizaje del rostro**
3. **Reconstrucción de la imagen**

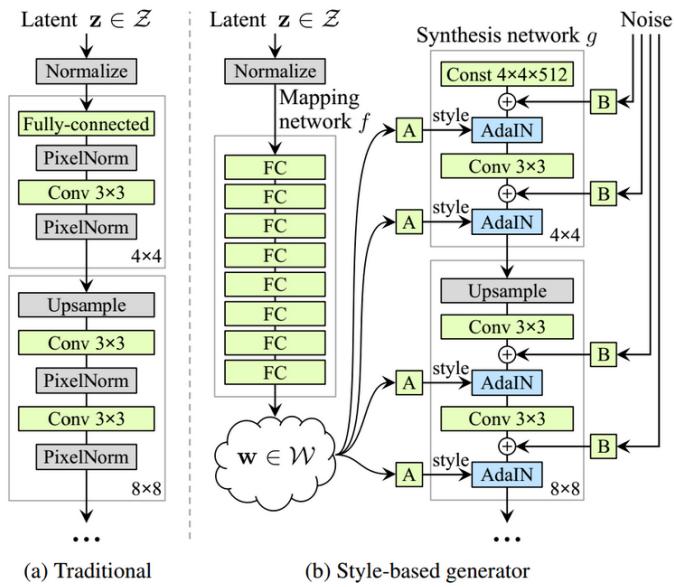


Figura 7: Modelo StyleGAN

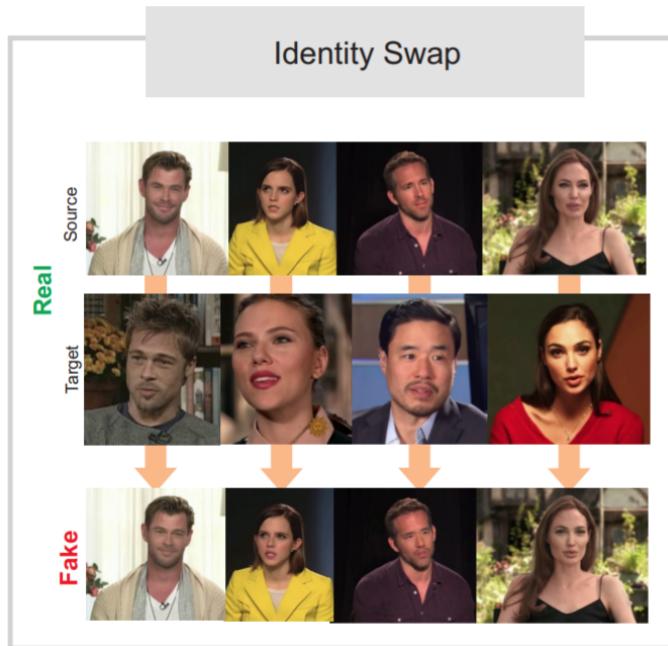


Figura 8: Ejemplos de aplicaciones usando intercambios de identidad

Para el **aprendizaje del rostro** se suelen utilizar algoritmos que usan ***autoencoders***, es decir, modelos de aprendizaje no supervisados que se usan para aprender una **representación compacta** de los datos.

En la Figura 9 se muestra el proceso por el cual se genera un *deepfake* usando esta técnica [10]. Hay 2 imágenes originales de dos personas reales, A y B. Estas imágenes se pasan a la red neuronal para aprender la representación de ambos rostros. Primero se pasan al **encoder** para **descomponer los rostros en características**, y de aquí se pueden sacar diversas características como la expresión facial, orientación, etc. Luego se pasan al **decoder** para **aprender el rostro y reconstruir la imagen original**. Intercambiando los **decoders**, este modelo puede generar el rostro de B con el rostro de A, y viceversa.

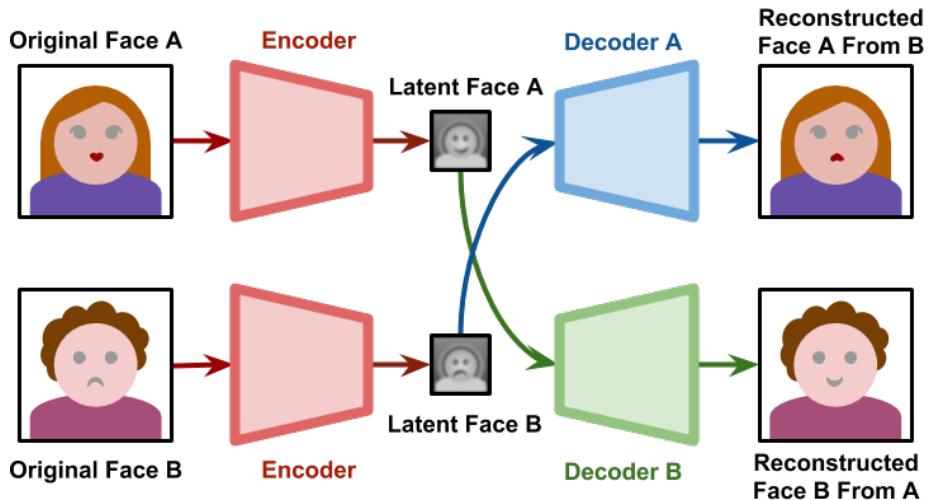


Figura 9: Proceso *autoencoder* utilizado en el intercambio de identidad

Al igual que en casi todos estos modelos, es mucho más eficiente en términos de **calidad y realismo** en comparación con otras técnicas gráficas como **FaceSwap** [11], el cual simplemente consiste en un detector de rostros y mezcla de imágenes.

### 2.1.3. Manipulación de atributos faciales

Esta técnica es un poco más simple que las demás. Consiste en modificar una imagen **cambiando atributos** tales como el color del pelo, color de la piel, género, edad, etc.

Además de ser simple, se está convirtiendo en una técnica muy popular debido a las aplicaciones comentadas anteriormente como **FaceApp** o incluso **Instagram**, cuyos filtros permiten ponerse gafas, pecas, cambiarse el color del pelo, etc.

Uno de los métodos utilizados para esta técnica se denomina **Attribute GAN** (AttGAN), que como se puede intuir es una mejora del modelo **GAN** que ya se introdujo en la primera técnica vista en esta Sección.

De forma básica, **AttGAN** se compone de dos subredes básicas: un *encoder* y un *decoder*. En la Figura 10 se muestra una vista general de cómo funciona este proceso. La idea principal es la siguiente [12]:

1. Dada una imagen de entrada que contiene  $n$  atributos, el **encoder** codifica la imagen y la representa comprimiendo dichos atributos. Por ejemplo, color de ojos, pelo, barba, etc.
2. Los atributos se editan a través de la **decodificación de los atributos** obtenidos en el paso 1).

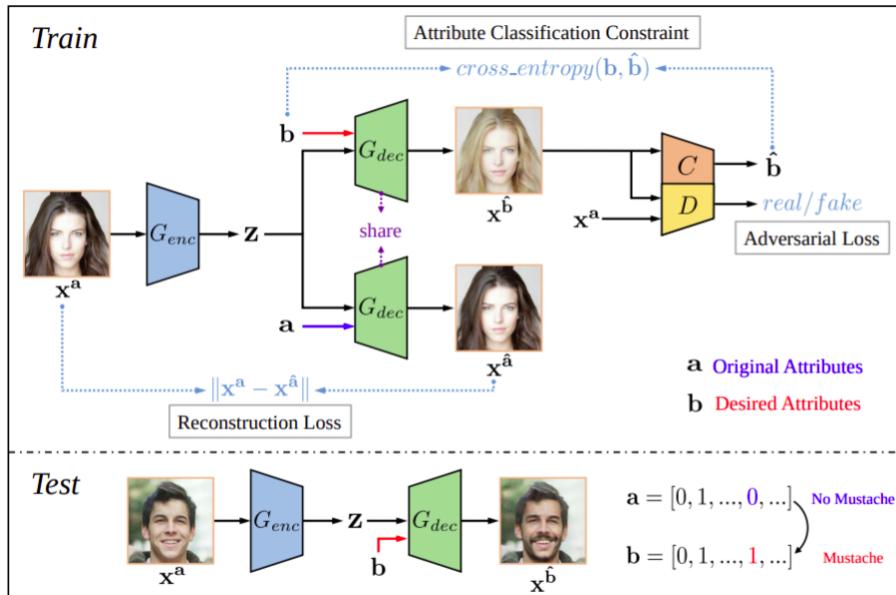


Figura 10: Vista general de **AttGAN**

Hay una extensión de este modelo que **modifica el estilo de cada atributo** [12]. En la Figura 11 se puede ver que la extensión consiste en pasar una entrada adicional al *decoder* que indica el nivel del atributo. Por ejemplo, un nivel 0 en el atributo *gafas* indicaría unas simples gafas sin cristales; si se sube un nivel, pasariamos a tener unas gafas con cristales, y así sucesivamente.

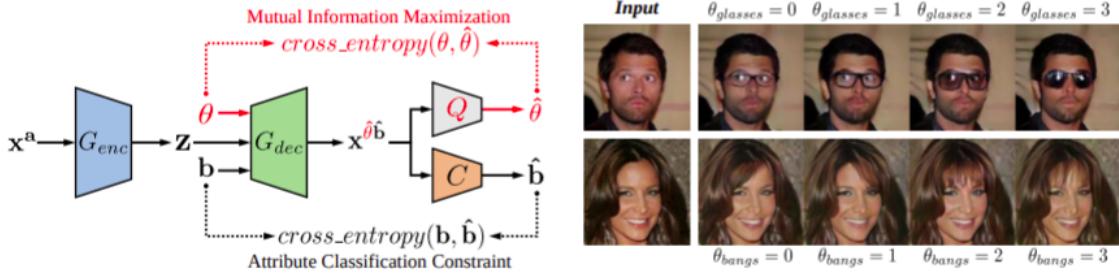


Figura 11: Visualización de la extensión de **AttGAN**

#### 2.1.4. Intercambio de expresión facial

Tal y como se puede intuir, en esta última técnica se **manipulan las expresiones faciales** de las personas a partir de otras, es decir, es muy similar al intercambio de identidad visto anteriormente pero intercambiando expresiones faciales. También es una técnica utilizada en el anuncio tan remarcado de este documento de **Campofrío**, pues en él se deben copiar las expresiones faciales de Lola Flores (que son muy características de ella).

En la Figura 12 se muestran varios ejemplos ilustrativos de cómo se manipulan las expresiones faciales a partir de otras caras [7].

El modelo **AttGAN** también es muy usado en este tipo, pero aparte de este también se pueden usar otros modelos tales como **StarGAN**, **InterFaceGAN** o **STGAN**, que al fin y al cabo son modelos **GAN** con ciertas modificaciones.

De acuerdo al estudio realizado por Shiv Kumar [13], **StarGAN** presenta algunos problemas en la manipulación de atributos sobre todo cuando se trata de cambiar el color y tono de la piel. Es por eso que a lo largo de este documento se ha decidido investigar más profundamente **AttGAN**, que obtiene aún mejor resultados.

## 2.2. Casos reales de deepfakes

Una vez que se han visto diferentes técnicas con las que es posible obtener imágenes reales a través de *machine learning*, en este apartado se van a presentar algunos de los casos más famosos recientes que han sido *deepfakes*. Esta lista se debe a **Joseph Foley** [14] y para simplificarla se van a presentar los tres que parecen más interesantes.

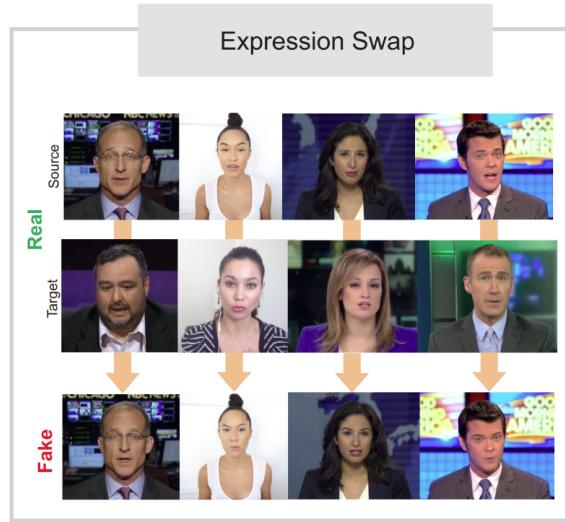


Figura 12: Ejemplos de manipulación de expresiones faciales

### 2.2.1. Discurso de Obama

En 2018 se hizo un videomontaje muy famoso publicado por **BuzzFeed** en el que se mostraba un vídeo de **Obama** afirmando que Trump “**es un completo idiota**” [15]. Este vídeo se puede encontrar en la plataforma **YouTube** y cuenta con más de 8 millones de reproducciones. El enlace del vídeo es el siguiente: <https://www.youtube.com/watch?v=cQ54GDm1eL0>.

Fue el actor **Jordan Peele**, cuñado del fundador de **BuzzFeed**, el que recreó esta escena con un motivo fundamental: advertir del peligro que supone la inteligencia artifical para la creación de *fake news*.

De acuerdo al artículo de este *deepfake*, este vídeo de tan solo 70 segundos tuvo un tiempo de montaje de más de **56 horas** de renderización usando para ello las herramientas software **Adobe After Effects** y **FakeApp**.

### 2.2.2. Donald Trump se une a la serie Breaking Bad

Este *deepfake* se hizo con una intención puramente paródica, es decir, en ningún momento se quiso engañar al público para difundir algo que es totalmente falso. Lo que se hizo fue escoger una escena de la popular serie **Breaking Bad** y reemplazó a **James McGill** con la cara de **Donald Trump**.

En esta escena, James está explicando a otro compañero cómo funciona el blanqueo de dinero.

Los creadores de esta parodia se conocen como **Ctrl Shift Face** en **YouTube** y usaron como herramienta **DeepFaceLab**. Para las voces, decidieron usar la herramienta llamada **Stable Voices**, que es un modelo cuyo entrenamiento se hace sobre discursos de voz reales.

El enlace al vídeo es el siguiente: <https://www.youtube.com/watch?v=Ho9h0ouemWQ>

### 2.2.3. Remasterización de la Princesa Leia

Esta noticia fue tan polémica que hasta gente que nunca ha visto la saga de **Star Wars** (yo incluido) sintió las críticas que le llovieron en su día.

La película de esta saga llamada **Star Wars: Rogue One** lanzada en 2016 sufrió muchísimas críticas por parte de los fans de la saga debido a que la remasterización de **Carrie Fisher** (actriz que da vida a la princesa Leia) era completamente surrealista y en ningún momento se consiguió el resultado esperado.

El creador de contenido *deepfake* llamado **Derpfakes** en **YouTube** subió un vídeo en el que él hacía su propia remasterización de la Princesa Leia y cuyo resultado era muchísimo mejor que el obtenido en la película original. Los fans de la saga comenzaron a difundir este vídeo diciendo que deseaban que la película hubiera tenido una remasterización como esta, que se puede ver en el siguiente enlace: <https://www.youtube.com/watch?v=RiBqZoVe92U>.

## 3. Caso práctico

Para poner en práctica todos los conceptos desarrollados en este documento, la intención era en un principio probar alguna implementación de los algoritmos **AttGAN** y **StyleGAN**.

En cuanto al modelo **AttGAN**, hay una implementación en **Tensorflow** hecha por el usuario de GitHub **LynnHo** [16] a través del cual se pueden editar los atributos faciales que se quieran. El problema encontrado es que debe haber un problema con la versión actual del código ya que da un error al ejecutarlo.

A continuación se ha buscado alguna implementación de **StyleGAN** y se ha encontrado un repositorio de la propia empresa **Nvidia** [17] el cual contiene la implementación oficial en **Tensorflow**. El problema es que tal y como se explicó de forma gráfica en el desarrollo

de **StyleGAN**, este modelo puede durar incluso días para mejorar la resolución de las fotos. De hecho, en el propio repositorio se muestran los tiempos de entrenamiento usando una **GPU Tesla V100** (ver Figura 13).

Expected training times for the default configuration using Tesla V100 GPUs:			
GPUs	1024×1024	512×512	256×256
1	41 days 4 hours	24 days 21 hours	14 days 22 hours
2	21 days 22 hours	13 days 7 hours	9 days 5 hours
4	11 days 8 hours	7 days 0 hours	4 days 21 hours
8	6 days 14 hours	4 days 10 hours	3 days 8 hours

Figura 13: Tiempos de entrenamiento del modelo **StyleGAN**

Finalmente, investigando un poco a través de la web, se ha encontrado una herramienta interactiva a modo de visualización para las **Generative Adversarial Networks** (GAN) que básicamente son la base de todos los modelos utilizados en los distintos tipos de *deep-fakes*. Esta herramienta se llama **GAN Lab** y se encuentra subida en un repositorio de Github a manos del usuario **poloclub** [18]. Está implementada usando **Tensorflow**, al igual que las anteriores.

**GAN Lab** no visualiza realmente la generación realística de imágenes, ya que esto es un proceso demasiado complejo. Sin embargo, se muestra el aprendizaje de una distribución de puntos en 2D. Realmente no hay ninguna aplicación que use esto tan simple, pero da una idea de cómo funciona el mecanismo de una manera muy sencilla.

Para empezar a utilizar esta herramienta, primero se debe elegir una distribución de datos. Par ello, existen diversos conjuntos de datos de ejemplo. Cuando se selecciona, el conjunto aparece en dos lugares (ver Figura 14):

- **Vista general del modelo:** se sitúa a la izquierda y muestra la arquitectura **GAN**, sus componentes principales y cómo se conectan.
- **Distribución en capas:** se visualizan las componentes del modelo.

Una vez que el conjunto de datos está seleccionado, se le da al botón *play* para empezar a entrenar el modelo. En la Figura 16 se puede ver el proceso a medida que va aumentando el número de *epochs*. En verde se muestran los **ejemplos reales** y en morado los **ejemplos fake** que genera el modelo. Se puede ver como el modelo intenta que las imágenes *fake* cada vez se vayan pareciendo más a las imágenes reales, acercándose cada vez más a ellas. También se puede visualizar de forma muy sencilla la **dirección del gradiente**.

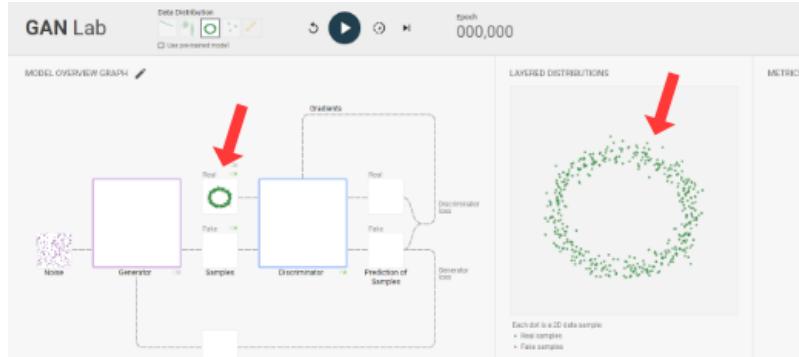
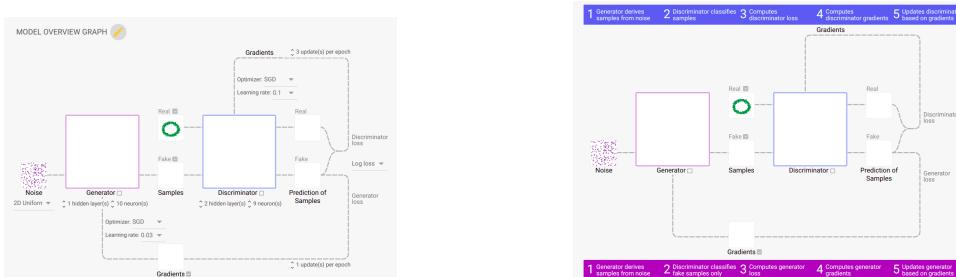


Figura 14: Selección y visualización de la distribución de datos [18]

Esto es sólo algo sencillo que se puede ver con esta herramienta, pero incluye muchas más funciones. Recordemos que en **GAN** encontramos dos subredes: el **generador** y el **discriminador**, y ambos compiten entre ellos para que las imágenes *fake* se parezcan cada vez más a las imágenes reales. Esta herramienta permite visualizar la interacción entre el generador y el discriminador.

Entre otras funciones interesantes que presenta esta herramienta, se encuentran las siguientes:

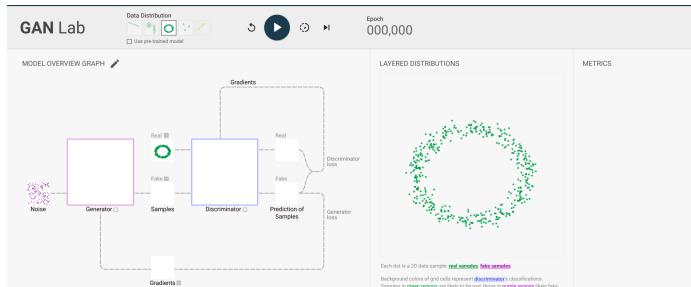
- **Ajuste de hiperparámetros:** se pueden editar los valores de los hiperparámetros (ver Figura 15a).
- **Seleccionar datos de entrada propios.**
- **Modo lento:** se puede activar el modo lento para no perder detalle de todo lo que está pasando (ver Figura 15b).
- **Ejecución paso por paso:** se puede entrenar individualmente cada iteración.



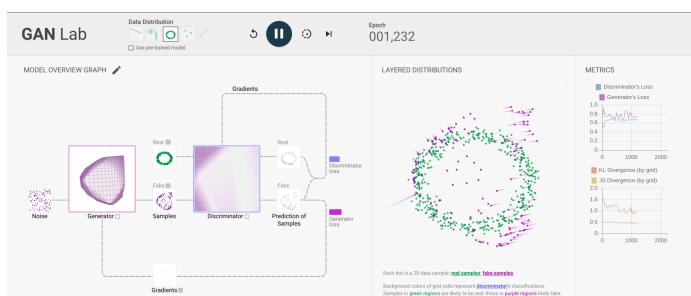
(a) Ajuste de hiperparámetros

(b) Modo lento

Figura 15: Otras funciones de **GAN Lab**



(a) Inicio (Epoch 0)



(b) Epoch 1,232



(c) Epoch 2,572

Figura 16: Proceso de entrenamiento de GAN usando GAN Lab

## 4. Bibliografía

- [1] IONOS. ¿qué son las fake news? <https://www.ionos.es/digitalguide/online-marketing/redes-sociales/que-son-las-fake-news/>, 2020.
- [2] Fernando Muñoz Gómez. Cada año te tragas ocho arañas mientras duermes: el origen del bulo. <https://abcblogs.abc.es/archivos-desclasificados/2015/08/13/comer-aranas-dormir-cama?ref=https%3A%2Fwww.google.com%2F>, 2015.
- [3] Pictoline. <https://twitter.com/pictoline/status/969419429636288512?lang=es>, 2018.
- [4] Webwise. Explained: What is false information (fake news)? <https://www.webwise.ie/teachers/what-is-fake-news/>.
- [5] Alex Cooper. Obama says illegal voters in 2016 won't be investigated, hillary announces amnesty for all illegals after election. <https://conservativedailypost.com/breaking-obama-and-hillary-now-promising-amnesty-to-any-illegal-that-votes-democrat/>, 2016.
- [6] Presidente del Gobierno \_\_\_ PARODIA Pedro Sánchez. <https://twitter.com/sanchezcasrejon>.
- [7] Ruben Tolosana, Ruben Vera-Rodriguez, Julian Fierrez, Aythami Morales, and Javier Ortega-Garcia. Deepfakes and beyond: A survey of face manipulation and fake detection. *Information Fusion*, 64:131–148, 2020.
- [8] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4401–4410, 2019.
- [9] R Horev. Explained: A style-based generator architecture for gans-generating and tuning realistic artificial faces. *towardsdatascience.com*, 30, 2018.
- [10] Nathanael Cretin. Using distributed learning for deepfake detection. <https://www.substra.ai/en/blog/deepfake1>, 2020.
- [11] Marek Kowalski. Face swap. <https://github.com/MarekKowalski/FaceSwap>, 2021.
- [12] Zhenliang He, Wangmeng Zuo, Meina Kan, Shiguang Shan, and Xilin Chen. Attgan: Facial attribute editing by only changing what you want. *IEEE Transactions on Image Processing*, 28(11):5464–5478, 2019.

- [13] Shiv Kumar Ganesh. Deepfakes — production detection using various deep learning methodologies. <https://medium.com/analytics-vidhya/deepfakes-production-detection-using-various-deep-learning-methodologies-3221e6002dd2>, 2021.
- [14] Joseph Foley. 12 deepfake examples that terrified and amused the internet. <https://www.creativebloq.com/features/deepfake-examples>, 2021.
- [15] José Ángel Plaza López. Los ‘deepfakes’ complican la lucha contra las noticias falsas. [https://elpais.com/retina/2018/09/17/innovacion/1537177382\\_367863.html](https://elpais.com/retina/2018/09/17/innovacion/1537177382_367863.html), 2018.
- [16] LynnHo. Attgan-tensorflow. <https://github.com/LynnHo/AttGAN-Tensorflow>, 2020.
- [17] NVlabs. Stylegan. <https://github.com/NVlabs/stylegan>, 2019.
- [18] poloclub. Ganlab. <https://github.com/poloclub/ganlab>, 2019.