

Automatizando la detección de contenido (deep) fake

Dr. Alfonso Muñoz (@mindcrypt)
Jose Ignacio Escribano



\$whoami

Dr. Alfonso Muñoz – Twitter: [@mindcrypt](https://twitter.com/mindcrypt)

Global Technical Cybersecurity Lead & Head of cybersecurity lab

Telegram: t.me/criptored | alfonso@criptored.com | Github: [mindcrypt](https://github.com/mindcrypt)



CriptoCert



- **Offensive/defensive security expert (certifications, European Organisms, public bodies and companies global 500, +60 academic publications, talks, patents, IEEE, ACM, books, ...)**
- **Speaker:** STIC CCN-CERT, DeepSec, HackInTheBox, Virus Bulletin, RootedCon, 8.8, ...
- **Profesor de Máster Seguridad:** UEM, UNIR, UC3M, UPM, UJAEN, EOI
- **Cybersecurity expert - Europol European Cybercrime Centre (EC3)**
- **Co-editor CRIPTORED** – Red Temática de Criptografía y seguridad de la información – <http://www.criptored.com>
- **Co-(autor) de la certificación CriptoCert Certified Crypto Analyst** – <https://www.cryptocert.com>

\$whoami

José Ignacio Escribano

joseignacio.escribano.pablos.next@bbva.com |

Github: [@jiel](#)



- Security & Machine Learning Researcher – BBVA Next Technologies
- Vidas paralelas
 - Graduado en Matemáticas e Ingeniería del Software & Máster en Ingeniería de la Decisión.
 - Actualmente, realizando tesis sobre criptografía post-cuántica.
 - Ponente en Cybergamp y Hack In The Box
- Áreas
 - Inteligencia Artificial
 - Criptografía
 - Cutting-edge research (Offensive & defensive)

Índice

01 Introducción

02 Generación de contenido sintético

Texto, audio, imagen y video

03 Detección de contenido sintético

Texto, audio, imagen y video

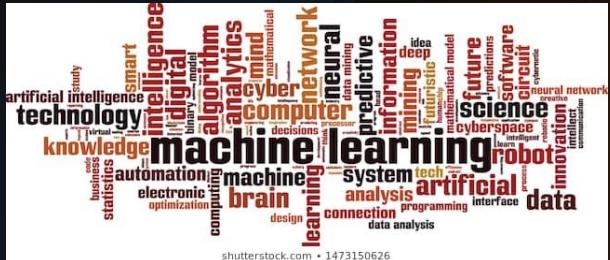
04 Futuro y conclusiones



OI Introducción

Contenido sintético y
ciberseguridad





SAY WHAT?

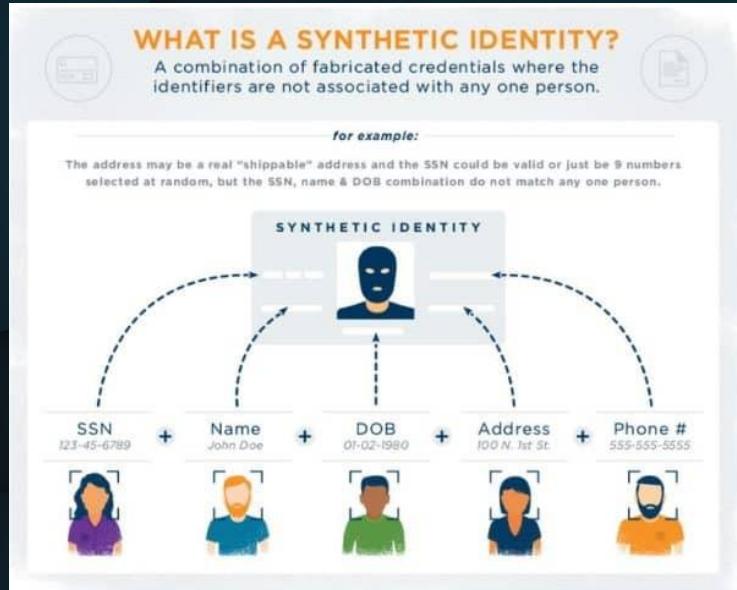


¿Qué es la identidad sintética?

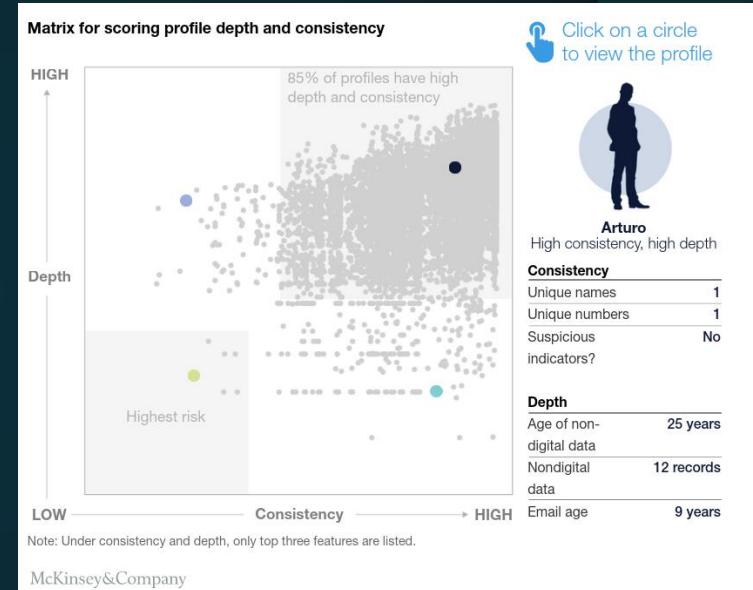
- El principal **objetivo** de esta presentación es dar una visión sobre la generación y detección de contenido sintético para engañar a una víctima.
- **No es algo novedoso** pero debido a la facilidad de acceso a **herramientas automáticas** es más sencillo.
- Diferencias dependiendo cómo se quiere generar, habría dos vertientes:
 - Graphical Computing
 - Deep Learning
- La aplicación de estas identidades para cometer **fraude** está en aumento, por lo que es necesario comprobar los métodos de autenticación.



Identidad sintética



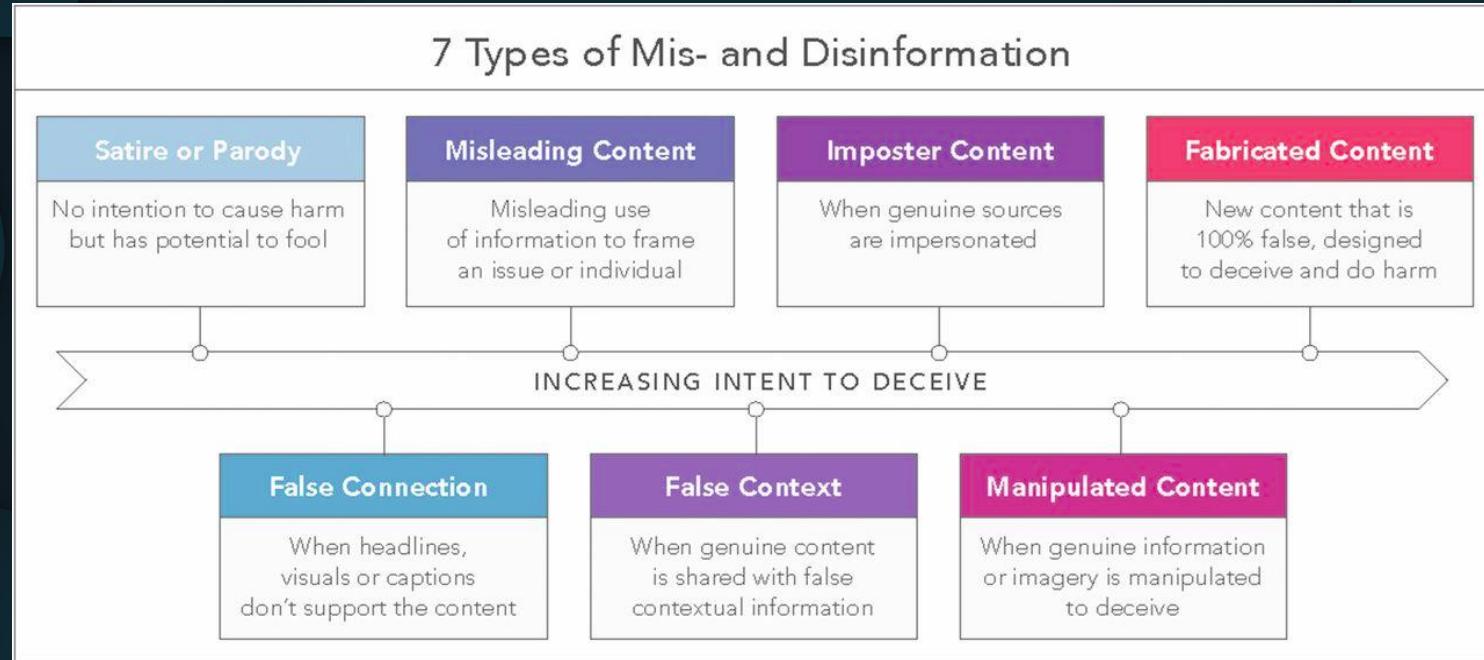
<https://www.idanalytics.com/solutions-services/fraud-risk-management/synthetic-identity-fraud/>



McKinsey&Company

<https://www.mckinsey.com/business-functions/risk/our-insights/fighting-back-against-synthetic-identity-fraud>

Threat models - ¿por qué?



<https://www.pnas.org/content/114/48/12631>

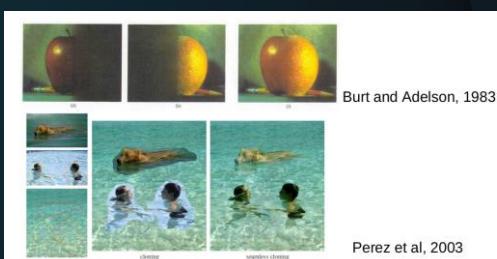
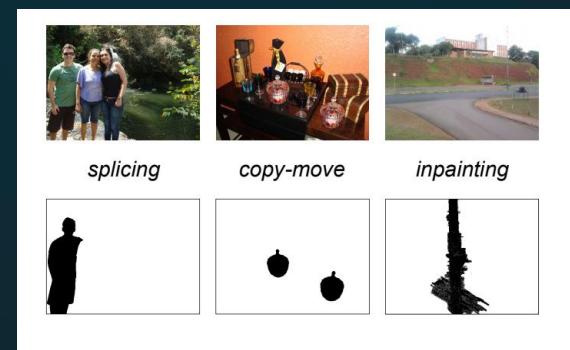


Threat models - ¿para qué?



<https://slideslive.com/38917755/applied-synthesis-and-detection>

Evolución de la manipulación



Evolución de la manipulación



Evolución de la manipulación – Fake News

Sistemas de verificación manual

- <https://www.factcheck.org>
- <https://maldita.es>
- <https://www.newtral.es>

Contenido fuera de contexto

- Otros lugares.
- Otros momentos temporales.
- Falsificaciones caseras.



02

Generación de contenido sintético

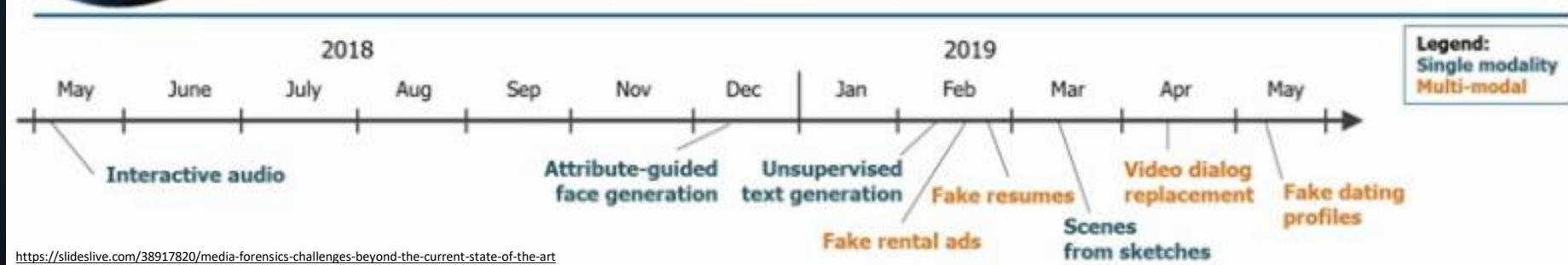
Texto, audio, imagen y vídeo



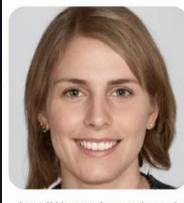
Generación de contenido



Incredible pace of synthetic media generation



<https://slideslive.com/38917820/media-forensics-challenges-beyond-the-current-state-of-the-art>



<https://thispersondoesnotexist.com/>



ENTIRE GUEST SUITE
Luxury Condo 3 Bed + 3 Bath
Port Melbourne

> 8 guests > 3 bedrooms > 4beds > 2 baths
Bathroom (with seating for 2 more people), basin and separate French garden and kitchen. 24/7 carpeted stairs. Laundry/membrainly - More balcony - Garden - Metro, Liverpool Street (15 min walk) Walking distance to Wyckofferton

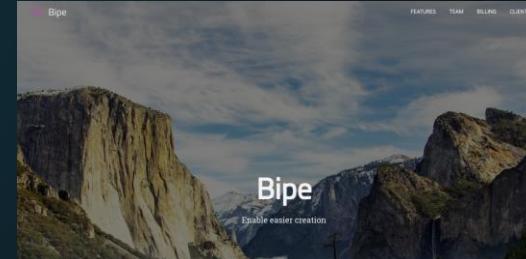
<https://thisrentaldoesnotexist.com/>



<https://thesecatsdonotexist.co.mn>



<https://thissnackdoesnotexist.com>



<https://thisstartupdoesnotexist.com>



Datos

Uno de los problemas más importantes en estos sistemas es el de tener datos de **calidad** para entrenar los modelos.



Capacidad de cómputo

Es necesario contar con **equipamiento** necesario para realizar las tareas en tiempos asumibles (**GPU**).



Conocimiento

Entender el funcionamiento de los modelos subyacentes ayuda a generar modelos personalizados.



Requisitos generales de generación

Generación de texto

- GPT-2 es el sucesor de GPT, es el modelo estado del arte de generación de texto.
- Liberado por OpenAI (<https://github.com/openai/gpt-2>).
- Predice la próxima palabra del texto.
- Consta de 1500 millones de parámetros.
- Entrenado con un conjunto de datos de 8 millones de páginas web.

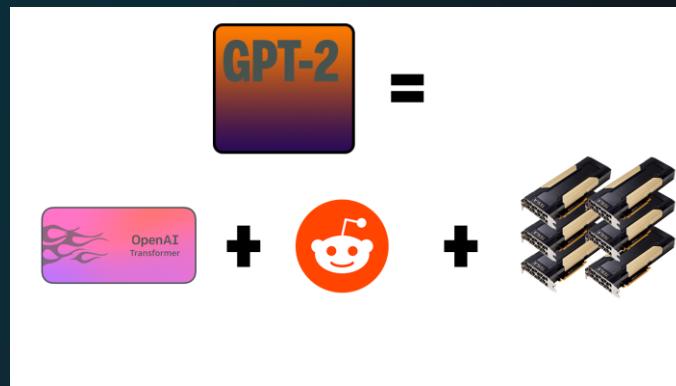


Better Language Models and Their Implications

We've trained a large-scale unsupervised language model which generates coherent paragraphs of text, achieves state-of-the-art performance on many language modeling benchmarks, and performs rudimentary reading comprehension, machine translation, question answering, and summarization—all without task-specific training.

FEBRUARY 14, 2019
24 MINUTE READ

<https://openai.com/blog/better-language-models/>



Generación de texto - Aplicaciones

Talk to Transformer

See how a modern neural network completes your text. Type a custom snippet or try one of the examples. [Learn more](#) below.

[Follow @AdamDanielKing](#)

for more neat neural networks.

Custom prompt

[COMPLETE TEXT](#)

About

Built by Adam King (@AdamDanielKing) as an easier way to play with OpenAI's new machine learning model. In February, OpenAI unveiled a language model called GPT-2 that generates coherent paragraphs of text one word at a time.

<https://talktotransformer.com/>

CTRL - A Conditional Transformer Language Model for Controllable Generation

Authors: Nitish Shirish Keskar, Bryan McCann, Lav Varshney, Caiming Xiong, and Richard Socher

This process repeats until an entire story is generated.



[Links](https://www.cnn.com/2019/08/26/tech/salesforce-new-ai) <https://www.cnn.com/2019/08/26/tech/salesforce-new-ai>

Salesforce is testing a new AI system that can help sales reps make better decisions about how to sell their products. The company has been working on the technology for more than two years, and it recently released its first public test of it.

In an interview with CNN, CEO Marc Benioff said he believes artificial intelligence will be able to do things like predict customer behavior or recommend products based on past purchases. But he's not sure if i

CTRL: A Conditional Transformer Language Model

<https://github.com/salesforce/ctrl>

AI DUNGEON

Support on Patreon

Imagine A Game With Infinite Adventures, As Unique As Your Own Life...

AI DUNGEON 2



Download on the
App Store



GET IT ON
Google Play



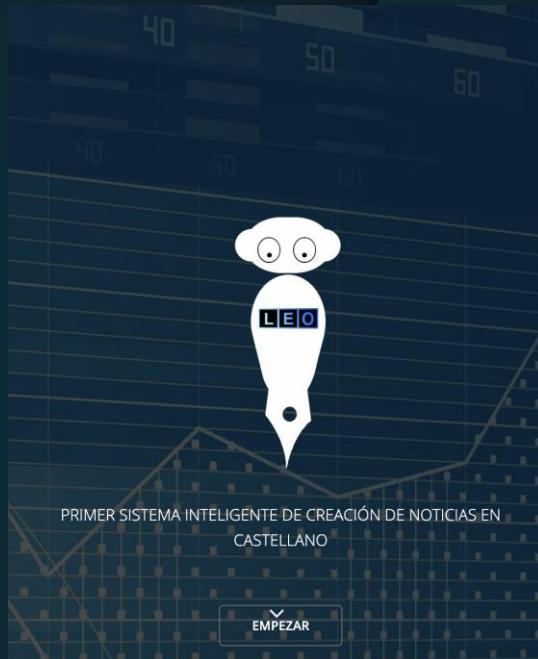
Play on your
Browser

By installing & playing our game you agree to our privacy policy and being contacted by us occasionally with the best stories, news on the development front and the odd special offer we feel you wouldn't want to miss out on. Emphasis on occasional - we hate spam probably more than you do. Opt out at your own peril at the bottom of any email.

<https://www.aidungeon.io/start>

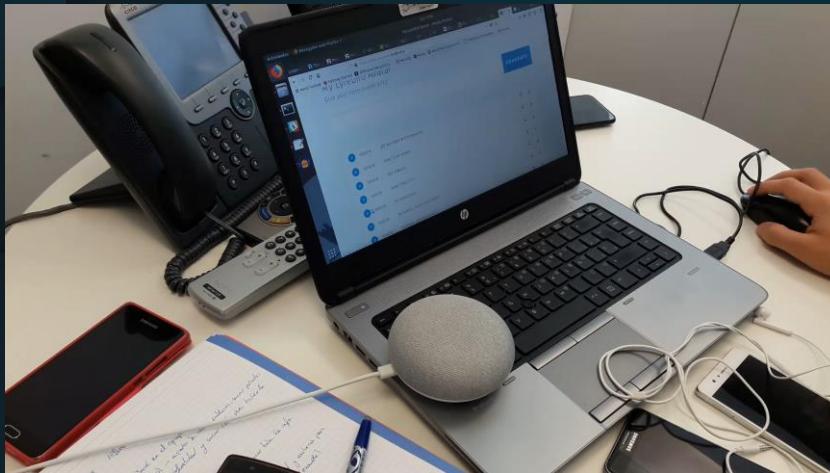
Generación de texto – Limitaciones

- Generación de textos solo en **inglés**.
 - No existen conjuntos de **datos en español** para entrenar modelos.
 - Las investigaciones se realizan en inglés y después se mejoran las traducciones, pero no tratan de generar textos nativos en **otros idiomas**.
- Textos largos con **poca coherencia**.



<http://leoia.es/>

Generación de audio



Lyrebird + Descript

You might have noticed a Descript feature in the video called "Overdub" — it deserves a few paragraphs of context.

In 2017, Alexandre de Brebisson, Kundan Kumar, Jose Sotelo, PhD students studying at MILA under [Yoshua Bengio](#), launched a startup called Lyrebird, allowing anyone to generate a realistic text-to-speech model of themself by uploading a few minutes of audio.

As soon as we saw it, we became conscious of a Lyrebird-shaped hole in Descript. In some ways, it felt like our missing half — Descript let you delete audio by deleting text, but Lyrebird would let you add audio by typing text.

A few months ago, we started talking to Lyrebird, and the more we talked, the more we realized how well we complimented each other — not only enabling personalized speech generation, but a whole new class of AI-enabled tools for media editing and synthesis.

So we merged the companies. And Overdub is our first collaboration, available today in closed beta.



Generación de audio - Datos

The screenshot shows the Mozilla Common Voice website. At the top, there's a navigation bar with links for 'COLABORAR', 'ARCHIVOS DE DATOS', 'IDIOMAS', 'SOBRE COMMON VOICE', and language selection ('ES'). Below the navigation is a sign-in button ('Iniciar sesión / Registrarse') and a dropdown for language ('ES'). The main area is divided into two sections: 'Hablar' (left) and 'Escuchar' (right). The 'Hablar' section features a red microphone icon and a large pink waveform. The 'Escuchar' section features a green play button icon and a large blue waveform. Below these sections is a URL: <https://voice.mozilla.org/>.

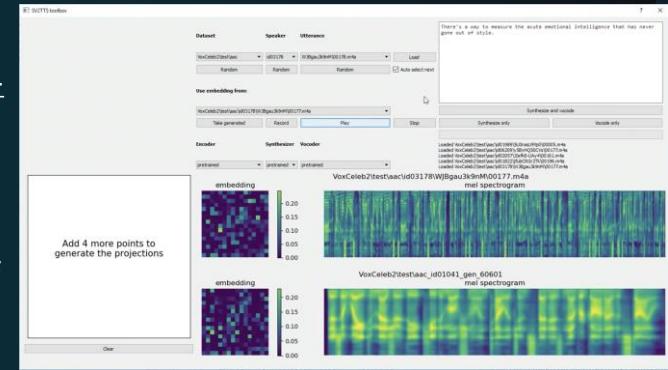
This screenshot shows the Mozilla Common Voice dataset page. It displays the following statistics:

Ítem	Español
TAMAÑO	5 Gb
VERSIÓN	es_221h_2019-12-10
TOTAL DE HORAS VALIDADAS	167
TOTAL DE HORAS	221
LICENCIA	CC-0
NÚMERO DE VOCES	8252
FORMATO DE AUDIO	MP3
PARTICIONES	Aerop. 14% España: Norte peninsular (Asturias, Castilla y León, Cantabria, País Vasco, ...) 13% España: Sur peninsular (Andalucía, Extremadura, Murcia), ... Edad 19-29, 13% 60-69, ... Género 55% Hombre, 10% Mujer



Real Time Voice Cloning

- Real-Time-Voice-Cloning es una herramienta **OpenSource** de generación de voz.
- Promete clonar la voz con sólo **5 segundos** de audio.
- Disponible en **GitHub**: <https://github.com/CorentinJ/Real-Time-Voice-Cloning>.
- Implementa **4 modelos** previos:
 - SV2TTS (TTS): <https://arxiv.org/pdf/1806.04558.pdf>
 - WaveRNN (vocoder): <https://arxiv.org/pdf/1802.08435.pdf>
 - Tacotron 2 (sintetizador):
<https://arxiv.org/pdf/1712.05884.pdf>
 - GE2E (encoder): <https://arxiv.org/pdf/1710.10467.pdf>
- Desarrollado en **Python**.
- Necesario disponer de **GPU**.



Real Time Voice Cloning (demo)

- Es necesario un **fichero de audio de la voz** que se quiere clonar sin ruido (.mp3, .wav, ...).
- Ejecutar **python demo_cli.py**.
 - Escribir ruta del fichero con el audio.
 - Escribir frase a clonar.

```
$ python demo_cli.py
```

```
Found 1 GPUs available. Using GPU 0 (GeForce GTX 1080) of compute capability 6.1 with 8.6Gb total  
memory.
```

```
Preparing the encoder, the synthesizer and the vocoder...  
Loaded encoder "pretrained.pt" trained to step 1564501  
Found synthesizer "pretrained" trained to step 278000  
Building Wave-RNN  
Trainable Parameters: 4.481M  
Loading model weights at vocoder\saved_models\pretrained\pretrained.pt  
Testing your configuration with small inputs.  
Testing the encoder...  
Testing the synthesizer... (loading the model will output a lot of text)
```

```
Reference voice: enter an audio filepath of a voice to be cloned (mp3, wav, m4a, flac, ...):  
trump.wav
```

```
Loaded file succesfully  
Created the embedding  
Write a sentence (+20 words) to be synthesized:  
That's one small step for a man, one giant leap for mankind.  
Created the mel spectrogram  
Synthesizing the waveform:  
([██████████] 66500/67200 | Batch Size: 7 | Gen Rate: 3.6kHz | )
```

```
Saved output as demo_output_00.wav
```



Shakespeare

Resemble.ai

- Aplicación similar a Real Time Voice Cloning.
- Clona la voz con al menos 50 frases prefijadas.
- Disponibles voces disponibles en varios idiomas (incluyendo español).
- Versión gratuita limitada.
- Editor avanzado de voces.

The screenshot shows a web-based application for voice cloning. At the top, there is a sidebar with a list of available voices:

- Aaron (Male - US)
- Aiden (Male - US)
- ✓ Scarlet (Female - US)
- Noah (Male - US)
- Sophia (Female US)
- Elijah (Male - US)
- William (Male - US)
- Olivia (Female Australian)
- Isaac (Male Australian)
- Lily (Female British)
- Harry (Male British)
- Elise (Female Dutch)
- Chloe (Female French)
- Alexandre (Male French)
- Hans (Male German)
- Emma (Female German)
- Maria (Female Italian)
- Akari (Female Japanese)
- Seo-yun (Female Korean)
- Zehra (Female Turkish)
- Asya (Female Turkish)
- Diego (Male European Spanish)
- Julia (Female European Spanish)** (highlighted in blue)
- Santiago (Male Latin America Spanish)
- Isabella (Female Latin America Spanish)
- Mehek (Female Indian-English)
- Diya (Female Indian-Hindi)
- Mariana (Female Portuguese - Brazilian)
- Joa (Male Portuguese - Brazilian)
- Claudia (Female Portuguese - European)
- Bruno (Male Portuguese - European)

The main interface has a blue header bar with the text: "You are creating the Voice for **Mi voz**. Speak clearly with a good microphone and no background noise. We need at least 50 samples." Below this, a progress bar indicates "You have recorded 0 samples so far (0.00%)". A large text input field contains the sentence: "The birch canoe slid on the smooth planks.". Above the input field, it says "SAY THE SENTENCE BELOW". At the bottom, there is a "TIP" button with the text "Say this like you would naturally.", a "Record" button with a microphone icon, and a red "Stop" button with a red circle icon.



Casos de uso

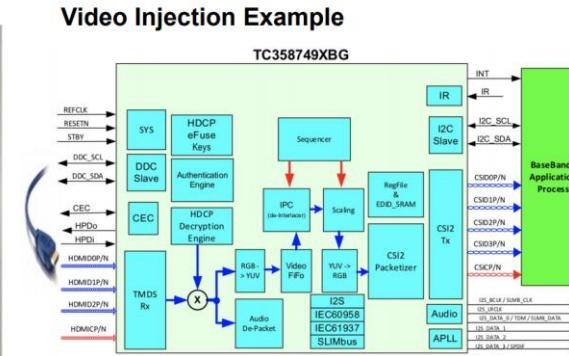
Audio Injection Example



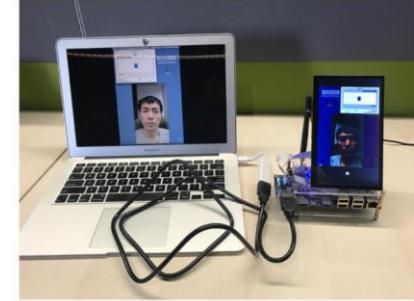
(a) For Android Devices

(b) For iOS Devices

Audio injection device based on analog circuits and sound card #BHUSA © @BLACK EAGLE



Video injection device based on TC358749XBG



#BHUSA  @BLACKHATEVENTS

<https://iblackhat.com/USA-19/Wednesday/us-19-Chen-Biometric-Authentication-Under-Threat-Liveness-Detection-Hacking.pdf>



<https://www.pandasecurity.com/mediacenter/news/deepfake-voice-fraud/>

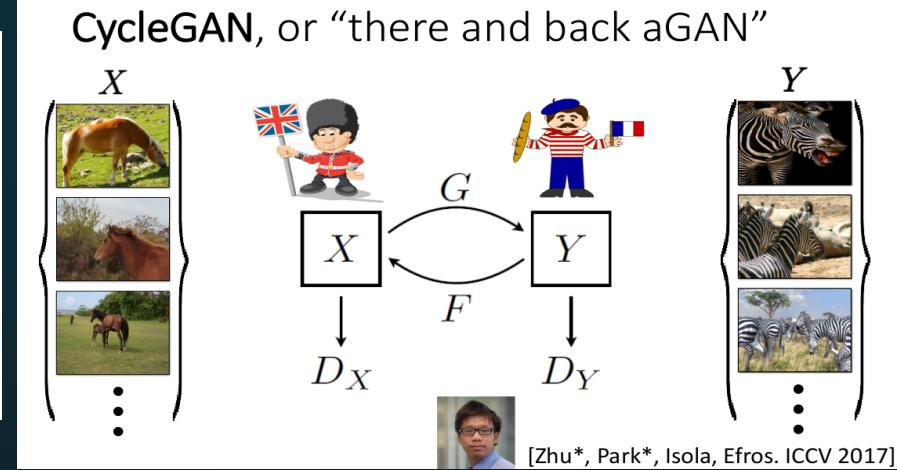
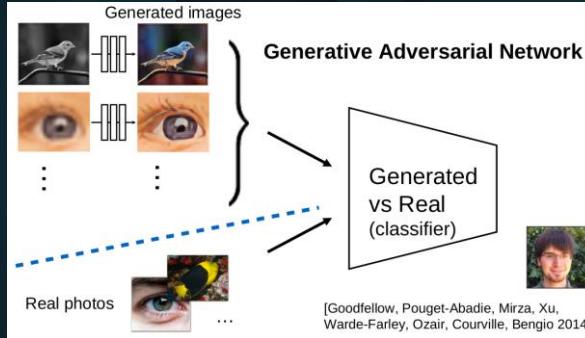
RealTalk: This Speech Synthesis Model Our Engineers Built Recreates a Human Voice Perfectly



<https://medium.com/dessa-news/real-talk-speech-synthesis-5dd0897eef7f>

Generación de imágenes

<https://tenso.rs/demos/fast-neural-style/>



Requisitos y limitaciones

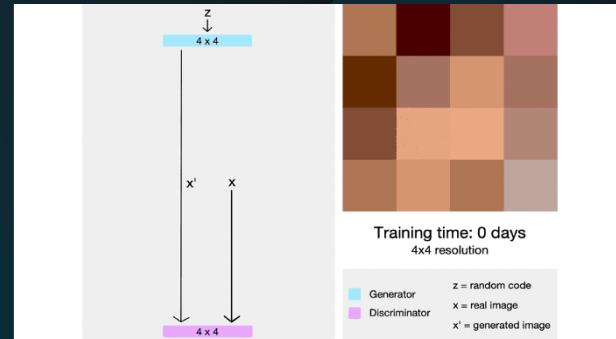
- Entrenar es **computacionalmente muy costoso**, aún con el uso de GPUs.

Configuration	Resolution	Total kimg	1 GPU	2 GPUs	4 GPUs	8 GPUs	GPU mem
config-f	1024×1024	25000	69d 23h	36d 4h	18d 14h	9d 18h	13.3 GB
config-f	1024×1024	10000	27d 23h	14d 11h	7d 10h	3d 22h	13.3 GB
config-e	1024×1024	25000	35d 11h	18d 15h	9d 15h	5d 6h	8.6 GB
config-e	1024×1024	10000	14d 4h	7d 11h	3d 20h	2d 3h	8.6 GB
config-f	256×256	25000	32d 13h	16d 23h	8d 21h	4d 18h	6.4 GB
config-f	256×256	10000	13d 0h	6d 19h	3d 13h	1d 22h	6.4 GB

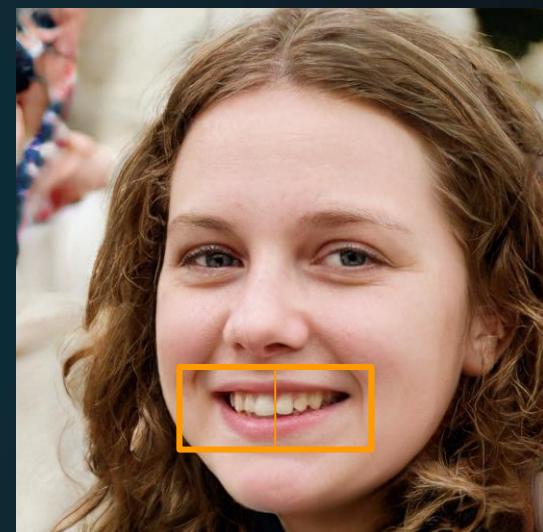
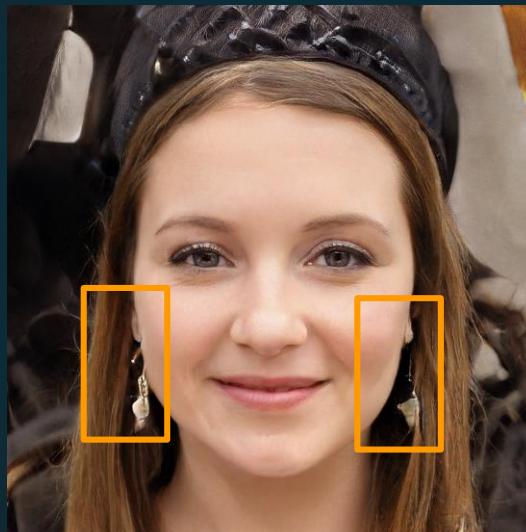
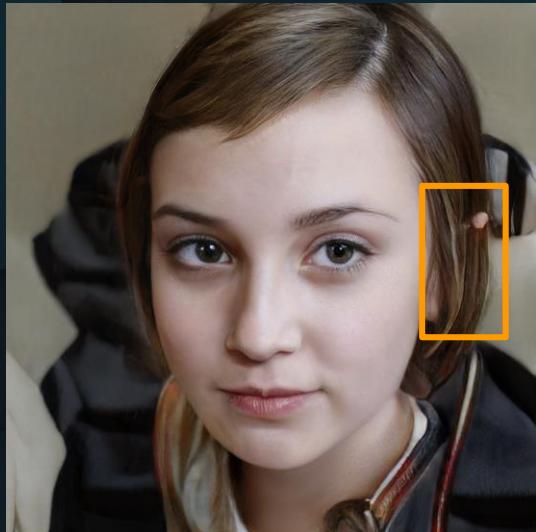
- Adaptar el modelo a nuestras necesidades, requiere disponer de **una gran cantidad de datos**.
- Imágenes **poco realistas** en ocasiones.

StyleGAN

- Generación de **caras hiperrealistas**.
- Disponible en **GitHub** (<https://github.com/NVlabs/stylegan>).
- Entrenado con **70.000 imágenes** (con licencia no permisiva) de **Flickr**.
 - Flickr-Faces-HQ (<https://github.com/NVlabs/ffhq-dataset>).
 - Tamaño de 1024x1024 píxeles.
- Desarrollado por **NVIDIA Labs**.
- <https://www.thispersondoesnotexist.com/> usa **StyleGAN** para la generación de caras.
- También aplicado a generación de **coches, habitaciones**,...

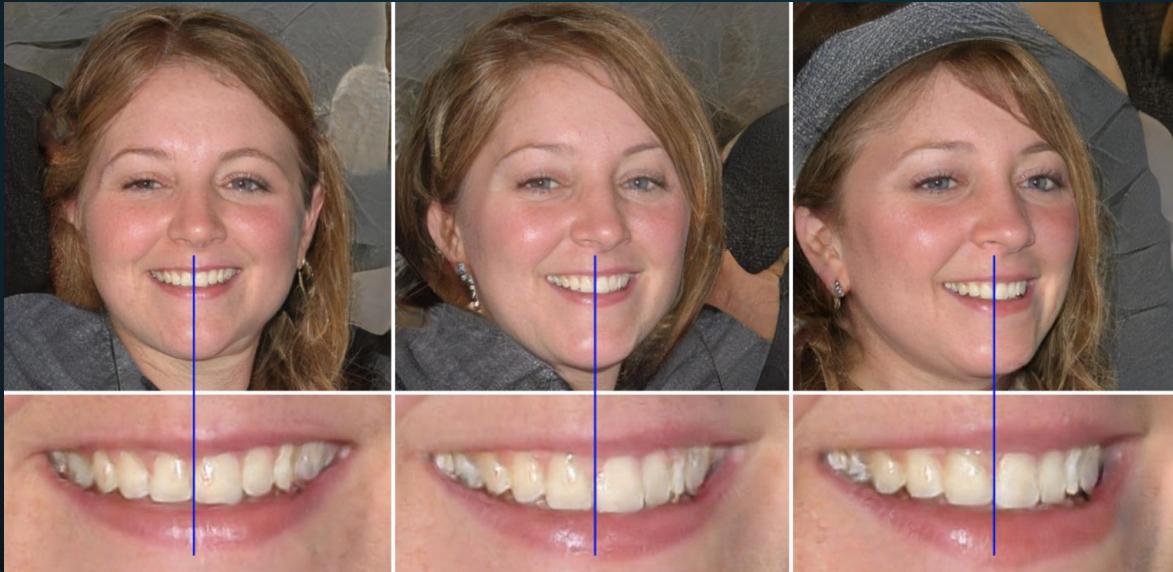


StyleGAN - limitaciones



StyleGAN2

- Evolución de StyleGAN.
- Soluciona los problemas de la versión anterior.
- Disponible en GitHub (<https://github.com/NVlabs/stylegan2>).



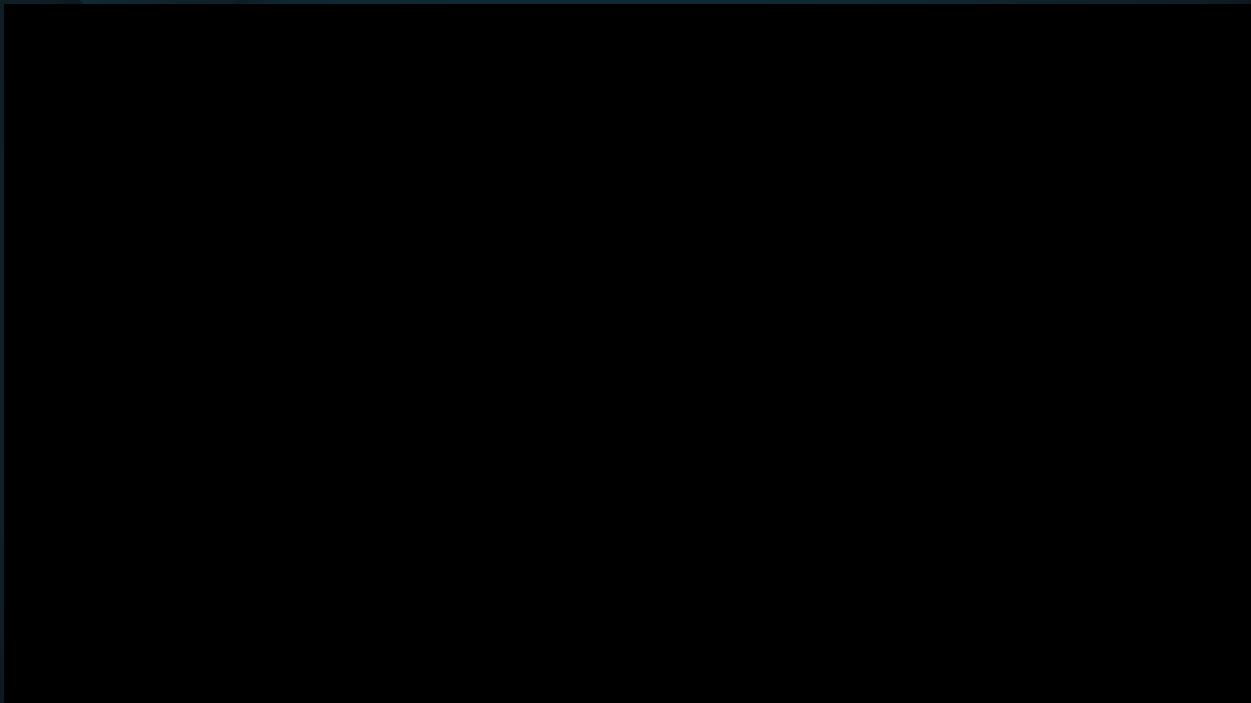
StyleGAN2

```
$ python run_generator.py generate-images --network=networks/stylegan2-ffhq-config-f.pkl \
--seeds=6600-6605 --truncation-psi=0.5
```

```
Local submit - run_dir: results\00179-generate-images
dnnlib: Running run_generator.generate_images() on
localhost...
Loading networks from "networks/stylegan2-ffhq-config-
f.pkl"...
Setting up TensorFlow plugin "fused_bias_act.cu":
Preprocessing... Loading... Done.
Setting up TensorFlow plugin "upfirdn_2d.cu":
Preprocessing... Loading... Done.
[...]
Generating image for seed 6600 (0/6) ...
Generating image for seed 6601 (1/6) ...
Generating image for seed 6602 (2/6) ...
Generating image for seed 6603 (3/6) ...
Generating image for seed 6604 (4/6) ...
Generating image for seed 6605 (5/6) ...
dnnlib: Finished run_generator.generate_images() in 23s.
```



StyleGAN2 – Proyector



Generación de imagen - casos de uso

REDES SOCIALES

Las cuentas falsas de Facebook ahora también usan rostros de personas que ni siquiera existen

Por [Donie O'Sullivan](#)

18:52 ET(23:52 GMT) 23 Diciembre, 2019



PATRÍCIA MARTINEAU

BUSINESS 12.28.2019 08:21 PM

Facebook Removes Accounts With AI-Generated Profile Photos

Researchers said it appears to be the first use of artificial intelligence to support an inauthentic social media campaign.

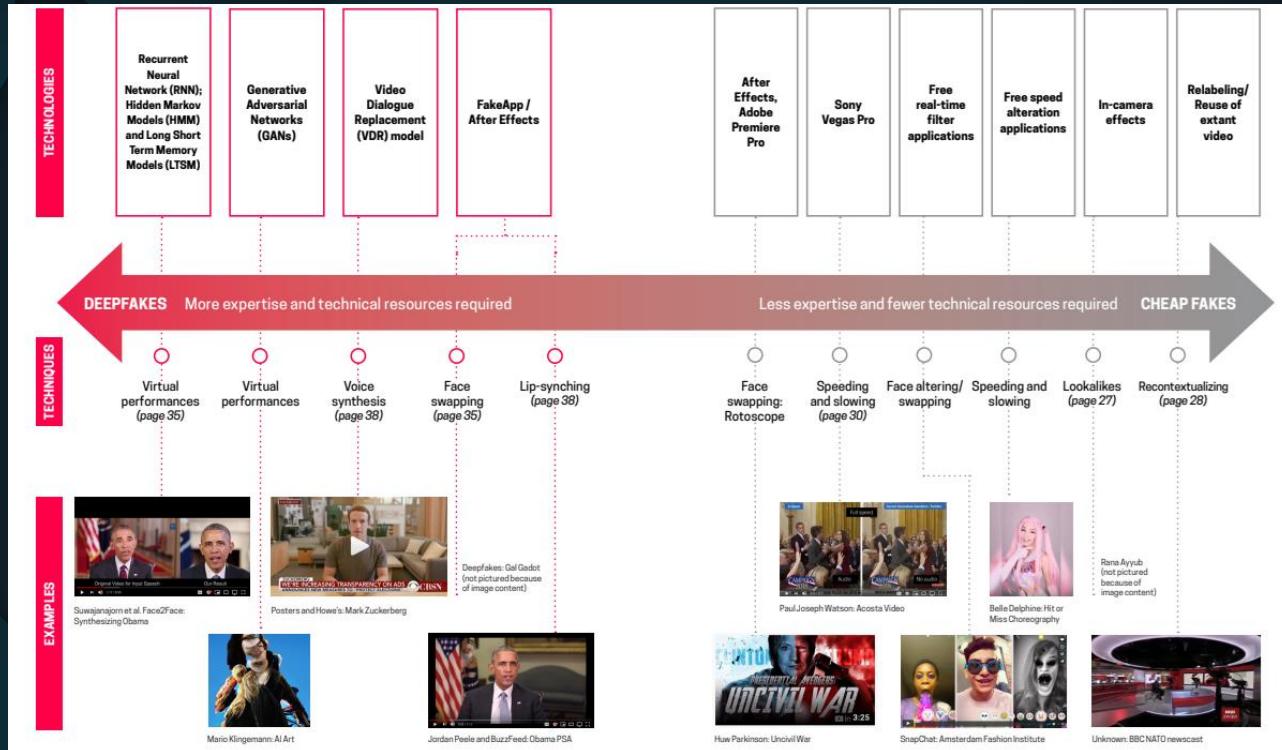


Generación de vídeo

- DeepFake Obama
- Modelo 3D de la cara
 - <https://dpo.si.edu/blog/smithsonian-creates-first-ever-3d-presidential-portrait>
- Información OSINT de Obama en discursos.
- Generación de la voz sintética.
 - <https://lyrebird.ai>
- Generación de modelo de la boca.
 - <https://nips2017creativity.github.io/docs/ObamaNet.pdf>
- Generación del vídeo completo.



Generación de vídeo



https://datasociety.net/wp-content/uploads/2019/09/DS_Deepfakes_Cheap_FakesFinal.pdf

Generación de vídeo - Limitaciones

- Es importante diferenciar si lo que se quiere es generar algo nuevo o alterar uno existente (predominante).
- Necesidad de una gran **cantidad de datos** de la persona.
- **Diversos ángulos, iluminación y expresiones** para un resultado realista.
- **Costoso en tiempo y en infraestructura** para generar la salida.
- Si no se utilizan muchas muestras en la fase de entrenamiento los resultados son **pobres y fácilmente detectables**.



Generación de vídeo

- Herramientas open source automatizadas.
 - *FaceSwap* (<https://github.com/MarekKowalski/FaceSwap>)

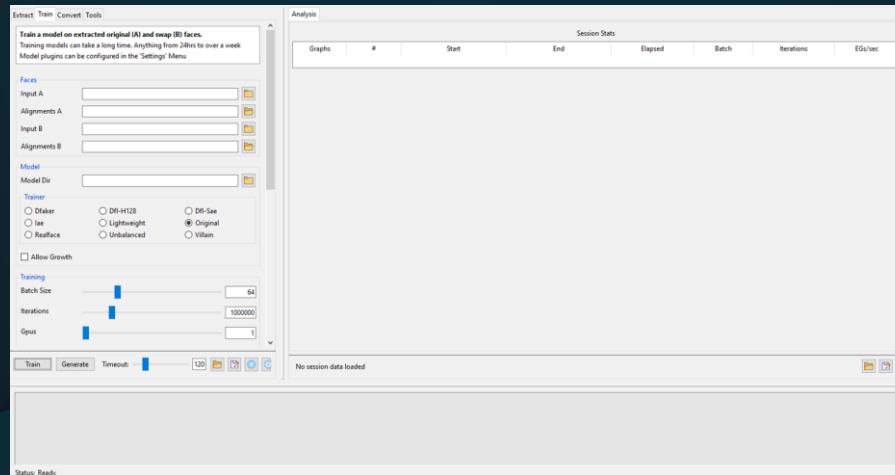


- *Face2Face* (<https://github.com/datitran/face2face-demo>)

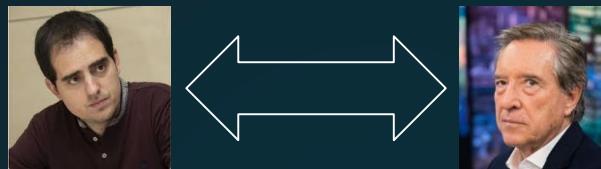


DeepFakes - FaceSwap

- Herramienta **OpenSource** para generar **Deep Fakes**.
- Disponible en **GitHub** (<https://github.com/deepfakes/faceswap>).
- Desarrollado en **Python**.
- Disponible tanto como **CLI** como **GUI**.



DeepFakes - Demo



Minutos de vídeo

40

Días de entrenamiento (GTX 1080)

1.5

Frames Iñaki Gabilondo

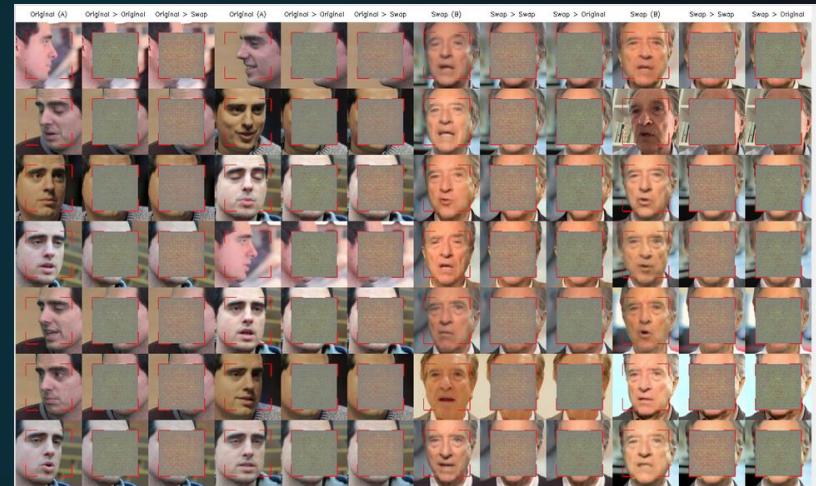
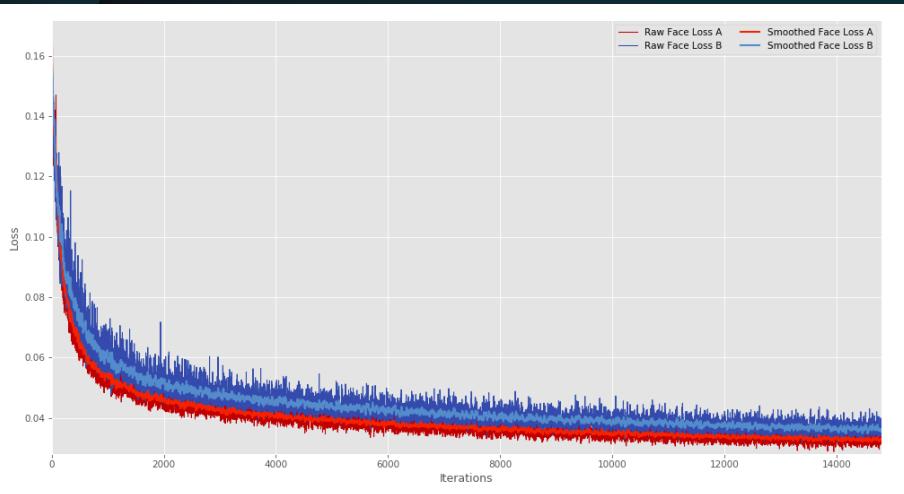
21818

Frames Alfonso

8924



DeepFakes - Demo



DeepFakes - FaceSwap

The logo for the Spanish television show "La Resistencia". It features the words "LA RESISTENCIA" in a white, sans-serif font, enclosed within a glowing yellow neon-style border that has a wavy, dynamic feel. The background is a dark, moody landscape with distant lights, suggesting a city at night.The logo for the Spanish public radio and television network SER. It consists of the letters "SER" in a large, bold, white, sans-serif font.

SER



Automatizando la detección de contenido deep fake – Dr. Alfonso Muñoz (@mindcrypt) y Jose Ignacio Escribano (Madrid, 2020)

03

Detección de contenido sintético

Texto, audio, imagen y vídeo



¿Es un problema real?

Artificial Intelligence / Machine Learning

The biggest threat of deepfakes isn't the deepfakes themselves

The mere idea of AI-synthesized media is already making people stop believing that real things are real.

by Karen Hao

Oct 10, 2019

<https://www.technologyreview.com/s/614526/the-biggest-threat-of-deepfakes-isnt-the-deepfakes-themselves/>

POLICY

DEEPFAKE PROPAGANDA IS NOT A REAL PROBLEM

We've spent the last year wringing our hands about a crisis that doesn't exist

By Russell Brandom | Mar 5, 2019, 12:25pm EST

<https://www.theverge.com/2019/3/5/18251736/deepfake-propaganda-misinformation-troll-video-hoax>

GILAD EDELMAN

BUSINESS 01.07.2020 08:58 PM

Facebook's Deepfake Ban Is a Solution to a Distant Problem

The platform has a plan to deal with tomorrow's disinformation. But what about today's?

<https://www.wired.com/story/facebook-deepfake-ban-disinformation/>

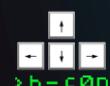


¿Es un problema real? NO

2. Detecting Manipulated Content

- Claim: GAN-based methods are not our biggest problem (yet)
 - Still easy to tell, both by (trained) humans and (trained) computers
 - Likely because the “discriminator” always wins
 - Once discriminators start to loose, there might be nothing we could do ☹
- Our strategy: address the current biggest sources of visual fakes

<https://slideslive.com/38917824/digital-image-manipulation-and-ways-to-detect-it>



¿Es un problema real? SÍ



Photo: Ryan Lash / TED

"Not only do we believe fakes, we are starting to doubt the truth," says Danielle Citron. Deepfakes pose a threat to the truth and democracy. (July 23, 2019, in Edinburgh, Scotland)

<https://www.daniellecitron.com/ted-talk>



**Sexual Privacy
Danielle Keats Citron**

University of Maryland Francis King Carey School of Law
Legal Studies Research Paper
No. 2018-25

¿Es un problema real? Depende...

The cover features a red, glowing, circular deepfake effect on a dark background. The title 'DEEPTTRACE' is at the top, followed by 'MAPPING THE DEEPFAKE LANDSCAPE'. Below the title is a large number '14,678'.

Henry Ajder, Giorgio Patrini, Francesco Cavalli, and Laurence Cullen
www.deeptracelabs.com
info@deeptracelabs.com

<https://deeptracelabs.com/mapping-the-deepfake-landscape/>

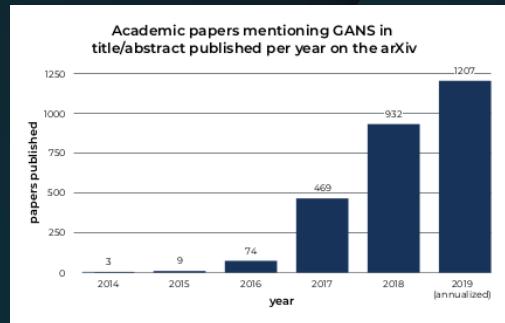
Table of Contents

The State of Deepfakes: An Overview	1
The Commodification of Deepfakes	3
Deepfake Pornography	6
Politics & Current Affairs	9
Ganware: Hacking humans and machines	12
Concluding Remarks	16
Appendix	17

Total number of deepfake videos online
14,678

percentage of deepfake videos online by pornographic and non-pornographic content
96% 4%

Total number of video views across top four dedicated deepfake pornography websites
134,364,438



Usan inteligencia artificial para cambiar la cara de actrices porno por famosas

Algoritmos de 'aprendizaje profundo' son capaces de cambiar rostros por otros, manteniendo expresiones

[Leer más](#)

Artificial Intelligence X-Ray App

Descarga Gratis Intentalo Online

Detección de contenido sintético - challenges

Join the Deepfake Detection Challenge (DFDC)

The Deepfake Detection Challenge invites people around the world to build innovative new technologies that can help detect deepfakes and manipulated media. Identifying manipulated content is a technically demanding and rapidly evolving challenge, so we're working together to build better detection tools.



facebook



ai PARTNERSHIP ON AI

Fraud Detection Contest



Ground Truth

Noiseprint

Heat map

Fraud Detection Contest: find it! (<http://find.it.univie.ac.at/>)
Artaud et al. "Find It! Fraud Detection Contest Report", IEEE ICPR 2018.

FNC

FNC-I TIMELINE FAQ FNC-I RESULTS ABOUT CONTACT



Exploring how artificial intelligence technologies could be leveraged to combat fake news.

FNC-I WINNERS AND RESULTS

OUR GITHUB REPOSITORIES

JOIN THE SLACK



ASVspoof 2019

Automatic Speaker Verification

Spoofing And Countermeasures Challenge

Future horizons in spoof/dialect audio detection



Detección de texto

AI2 ALLEN INSTITUTE for ARTIFICIAL INTELLIGENCE

GROVER - A State-of-the-Art Defense against Neural Fake News

Online disinformation, or fake news intended to deceive, has emerged as a major societal problem. Currently, fake news articles are written by humans, but recently-introduced AI technology based on Neural Networks might enable adversaries to generate fake news. Our goal is to reliably detect this "neural fake news" so that its harm can be minimized.

To study and detect neural fake news, we built a model named Grover. Our study presents a surprising result: the best way to detect neural fake news is to use a model that is also a generator. The generator is most familiar with its own habits, quirks, and traits, as well as those from similar AI models, especially those trained on similar data, i.e. publicly available news. Our model, Grover, is a generator that can easily spot its own generated fake news articles, as well as those generated by other AIs. In a challenging setting with limited access to neural fake news articles, Grover obtains over 92% accuracy at telling apart human-written from machine-written news. For more information, please [read our publication](#) as well as our [blog post with additional experiments](#). For updates, also [check out our project page](#).

Here, we demonstrate how Grover can generate a realistic-looking fake news article, and then detect that it was AI-generated.

- To generate a fake news article with Grover, use the 'Generate' tab, Fill in some article pieces, and press 'Generate' next to the piece you would like to generate. Grover will generate that piece based on the data provided. For instance, if the domain is "nytimes.com", clicking 'Generate' for the Article will generate a fake article as if it were written for the New York Times.
- To detect whether an article was written by Grover or a human, use the 'Detect' tab. Fill in the input field with article text, and click 'Detect Fake News.'

Note that, even if Grover fails to detect a given piece as fake, our findings suggest that releasing many such articles taken together would be relatively easy to spot. Thus, if a source of Neural Fake News disseminates a large number of articles, Grover will be increasingly capable of spotting these articles as malicious.

This demo is a prototype. It might take upwards of 30 seconds for Grover to finish generating or detecting, depending on how many people are using the demo right now.

Generate

Detect

Examples

FNC



Exploring how artificial intelligence technologies could be leveraged to combat fake news.

FACT CHECKERS AND RESULTS
OUR CITATION VERIFIERS
JOIN THE TEAM

Fake Article

nytimes.com

Why Bitcoin is a great investment

June 6, 2019 – Paul Krugman

As most of my readers know, I'm an optimist.

This belief applies across my life, and to various investments as well. So I am

<https://grover.allenai.org>

Giant Language model Test Room

The GLTR demo enables forensic inspection of the visual footprint of a language model on input text to detect whether a text could be real or fake. It is a collaborative effort between [Hendrik Strobelt](#), [Sebastian Gehrmann](#), and [Alexander Rush](#) from the [MIT-IBM Watson AI lab](#) and [Harvard NLP](#).

Please read the detailed [intro about GLTR](#). Source code is on [GitHub](#).

Each text is analyzed by how likely each word would be the predicted word given the context to the left. If the actual used word would be in the Top 10 predicted words the background is colored green, for Top 100 in yellow, Top 1000 red, otherwise violet. Try some sample texts from below and see for yourself if you can spot the difference between machine generated text and human generated text or try your own. (Tip: hover over the words for more detail)

The histograms show some statistic about the text: Frac(p) describes the fraction of probability for the actual word divided by the maximum probability of any word at this position. The Top 10 entropy describes the entropy along the top 10 results for each word.

Quick start - select a demo text:

machine: GPT-2 small top_k 5 temp_1 machine: GPT-2 small top_k 40 temp_.7 machine*: unicorn text (GPT2 large)

human: NYTimes article human: academic text human: woodchuck :)

or enter a text:

The cat was playing in the garden.

analyze

Tweet about GLTR
[MIT-IBM Watson AI lab](#) and [Harvard NLP](#)



<http://gltr.io/dist/index.html>

Detección de texto



Harley Barrett Smith @HarleyBarrettS1 · 26 oct. 2019
And now that? Europe and U.K. play game of Brexit Chicken. We are the affected. We see how the economy suffers from uncertainty.

<https://twitter.com/HarleyBarrettS1/status/1187975578076532739>

Fakenews & Fakeperson: Herramientas para manipular la verdad (A. Martín y Juan L. García, Sidertia)

Article

Text:

And now that? Europe and U.K. play game of Brexit Chicken. We are the affected. We see how the economy suffers from uncertainty.

Detect Fake News We are quite sure this was written by a machine.

Español

Article

Text:

¿Y ahora que? Europa y el Reino Unido juegan al Pollo Brexit. Nosotros somos los afectados. Vemos cómo la economía sufre de incertidumbre.

Detect Fake News

We think this was written by a machine (but we're not sure).



Detección de audio



DeepFake Audio Detection

With the popularity and capabilities of audio deep fakes on the rise, creating defenses against deep fakes used for malicious intent is becoming more important than ever.

We built a fake audio detection model with Foundations Atlas, for anyone to use. If you'd like to read more about why we decided to build this, [click here](#).

Here are two examples of short audio clips in `./data/example_clips/` folder. One of them is real and the other is fake.

This repository provides the code for a fake audio detection model built using Foundations Atlas. It also includes a pre-trained model and inference code, which you can test on any of your own audio files.

Setup

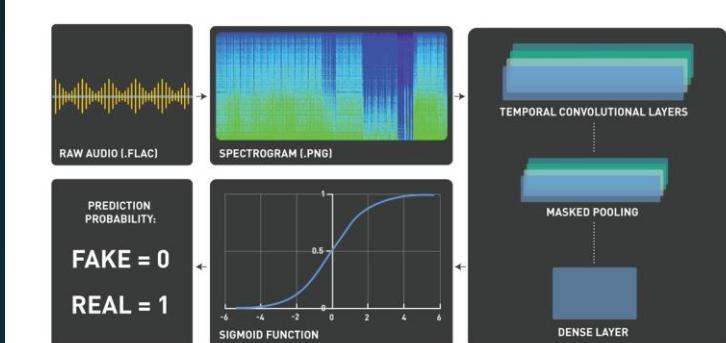
1. Git Clone this repository.
2. If you haven't already, install [Foundations Atlas Community Edition](#).
3. Once you've installed Foundations Atlas, activate your environment if you haven't already, and navigate to your project folder, and then into the `code` directory.
4. Run the following line in the terminal to install the required packages:

```
pip install -r requirements.txt
```

That's it, You're ready to go!

Note: To run the code, your system should meet the following requirements: RAM >= 32GB , GPUs >=1

<https://github.com/dessa-public/fake-voice-detection>

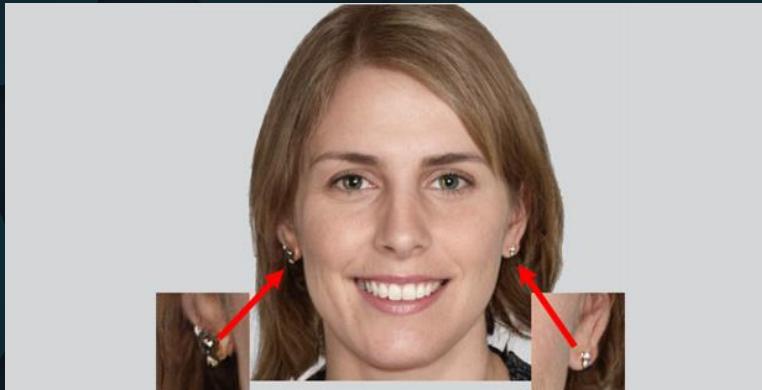


```
$ python inference.py
Using TensorFlow backend.
Loading inference data from ../data/inference_data/unlabeled
Loading pretrained model
saved_model_248_8_32_0_05_1_50_0_0.0001_100_156_2_True_
True_fitted_objects.h5
100%[=====] | 1/1
[00:00<00:00, 10.53it/s]
Using TensorFlow backend.
```

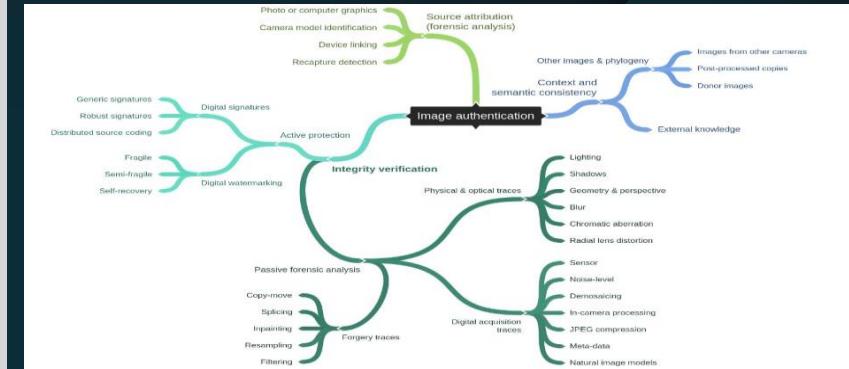
```
-----
[[[-1.        , -1.        , -0.866223   , ..., -1.        ,
   -1.        , -1.        ],
  [-0.46634382, -0.42106295, -0.37900943, ..., -1.        ,
   -1.        , -1.        ],
  [ 0.05602419,  0.18866599,  0.23696363, ..., -1.        ,
   -1.        , -1.        ],
  ...,
  [-0.82941407, -0.66641605, -0.54652816, ..., -1.        ,
   -1.        , -1.        ],
  [-0.91284347, -0.7570157 , -0.6872368 , ..., -1.        ,
   -1.        , -1.        ],
  [-1.        , -1.        , -1.        , ..., -1.        ,
   -1.        , -1.        ]], dtype=float32)]
```

```
[@ 1.17236993]
The probability of the clip being real is: 17.24%
```

Detección de imagen (manual)



<https://www.darpa.mil/news-events/2019-09-03a>



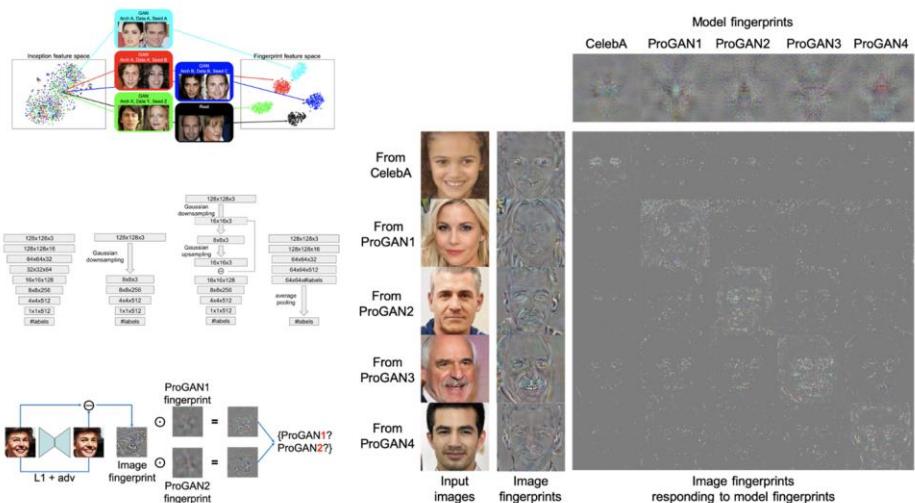
<http://kt.agh.edu.pl/~korus/files/icml19.pdf>

- A la hora de detectar este tipo de vídeos, los más **difíciles** para un humano son los que únicamente **cambian la expresión**. Las **orejas** no suelen ser iguales.
- Aparecen **pendientes diferentes** o en algunos casos desaparecen en una de ellas.
- **Las caras son asimétricas**, la mirada suele ser distinta para cada ojo.
- **Los dientes no son regulares**.
- El **fondo** de las fotos suele ser abstracto.

<https://medium.com/@kcimc/how-to-recognize-fake-ai-generated-images-4d1f6f9a2842>

Detección de imagen (automática)

GANFingerprints



<https://arxiv.org/pdf/1811.08180.pdf>

FakeSpotter: A Simple Baseline for Spotting AI-Synthesized Fake Faces

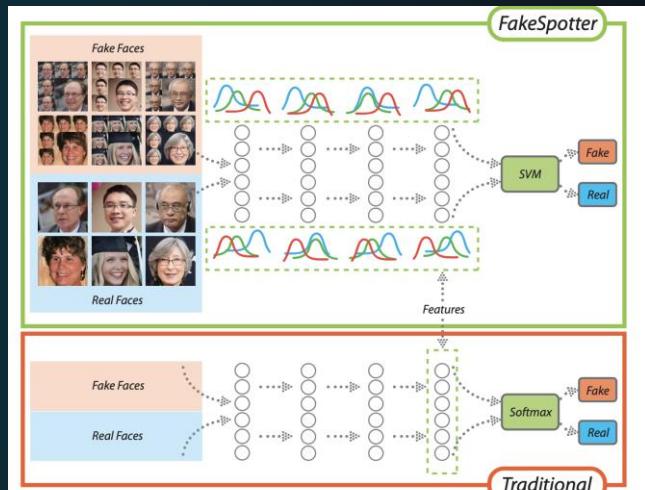


Fig. 2: An overview of the proposed FakeSpotter fake face detection method. Compared to the traditional learning based method (shown at the bottom), the FakeSpotter uses layer-wise neuron behavior as features, as opposed to final-layer neuron output.

<https://arxiv.org/pdf/1909.06122.pdf>

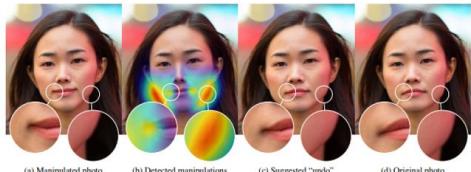
Detección de imagen (automática)

Detecting Photoshopped Faces by Scripting Photoshop

Sheng-Yu Wang¹ Oliver Wang² Andrew Owens¹ Richard Zhang² Alexei A. Efros¹

UC Berkeley¹ Adobe Research²

<https://peterwang512.github.io/FALdetector/>



<https://arxiv.org/pdf/1906.05856.pdf>

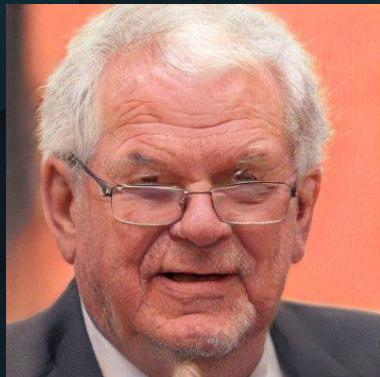


<https://github.com/peterwang512/FALdetector>

Detección de imagen (automática)



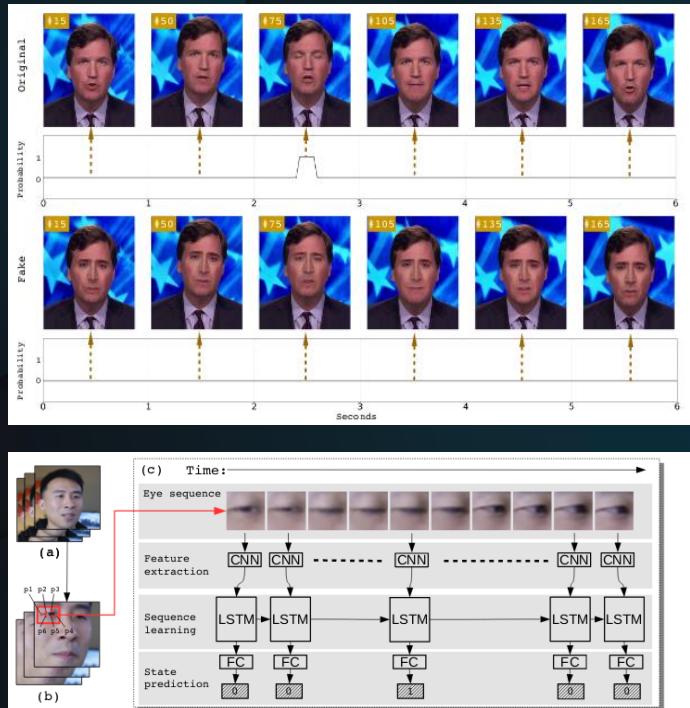
Harley Barrett Smith @HarleyBarrettS1 · 26 oct. 2019
And now that? Europe and U.K. play game of Brexit Chicken. We are the affected. We see how the economy suffers from uncertainty.



Probability being modified by Photoshop FAL:

10.70%

Detección de vídeo



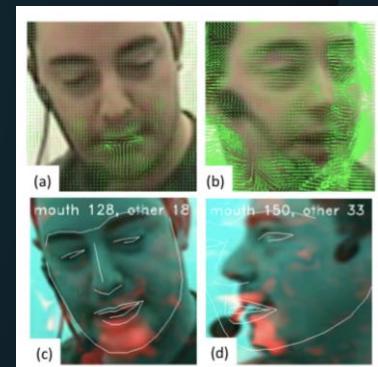
Spotting Audio-Visual Inconsistencies (SAVI) in Manipulated Video

Robert Bolles, J. Brian Burns, Martin Graciarena, Andreas Kathol, Aaron Lawson, Mitchell McLaren
SRI International

Thomas Mensink
University of Amsterdam

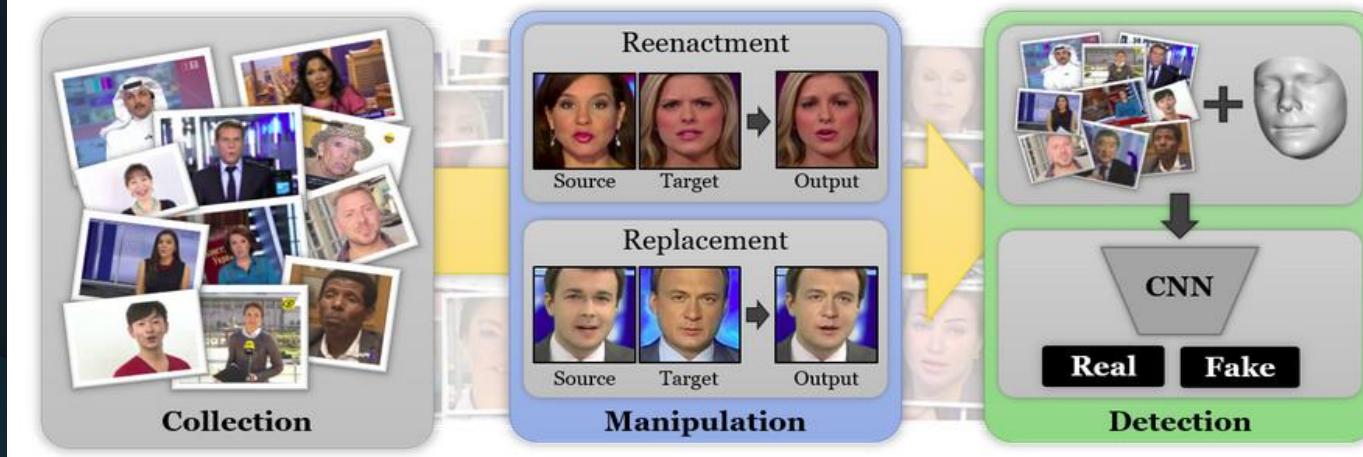
Type of inconsistency	Visual Features	Audio Features
Environmental class	Scene: indoor vs. outdoor, small room, etc.	Environmental classes, reverberation of closed vs open spaces
Speaker identity	Face recognition	Speaker ID
Lip movement	Pattern of lip movement	Speech patterns
Head movement	Change in head pose	Left-right channel balance
AV device movement	Motion relative to scene	Changes in environmental features
Missing sound	Presence of sound-producing activity	Presence of corresponding sound
Middle-level features	Visual scene signature	Audio scene signature

Table 1. Different types of audiovisual inconsistencies to be detected and characterized in SAVI. The aspects explored in this paper are shown in bold.



FaceForensics

FaceForensics++: Learning to Detect Manipulated Facial Images



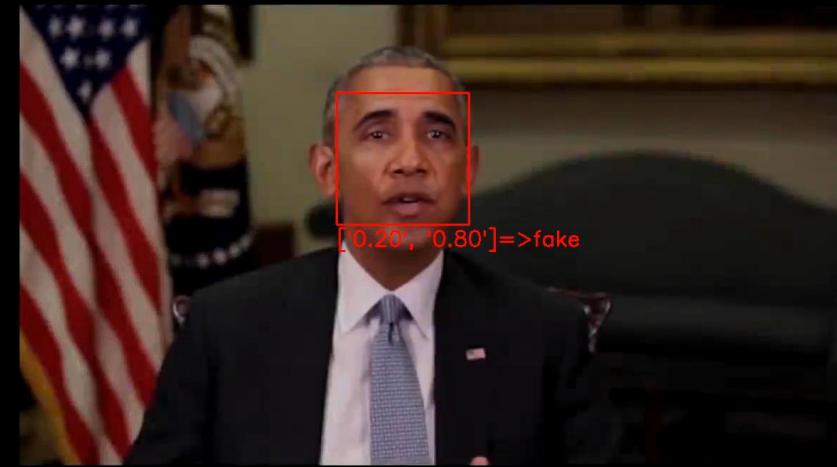
<https://github.com/ondyari/FaceForensics>

FaceForensics

- FaceForensics es una herramienta de **detección de DeepFakes**.
- Calcula la **probabilidad de que cada frame sea falso** o no.
- Disponible en **GitHub** (<https://github.com/ondyari/FaceForensics>).
- Entrenado con **vídeos de YouTube**.
- Conjunto de entrenamiento disponible **bajo solicitud**.
- Sólo detecta **3 modelos de Deep Fakes** (datos de entrenamiento).
- **Difícil de instalar**. Documentación incompleta.



FaceForensics - demo



FaceForensics - demo



Automatizando la detección de contenido deep fake – Dr. Alfonso Muñoz (@mindcrypt) y Jose Ignacio Escribano (Madrid, 2020)

FakeVideoForensics

Usage

```
usage: main.py [-h] [--model_path MODEL] [--output_path VIDEOOUT]
               [--start_frame START_FRAME] [--end_frame END_FRAME] [--cuda]
               [--fast] --video_path VIDEOIN

optional arguments:
  -h, --help            show this help message and exit
  --model_path MODEL, -mI MODEL
  --output_path VIDEOOUT, -o VIDEOOUT
  --start_frame START_FRAME
  --end_frame END_FRAME
  --cuda
  --fast

required arguments:
  --video_path VIDEOIN, -i VIDEOIN
```

Output

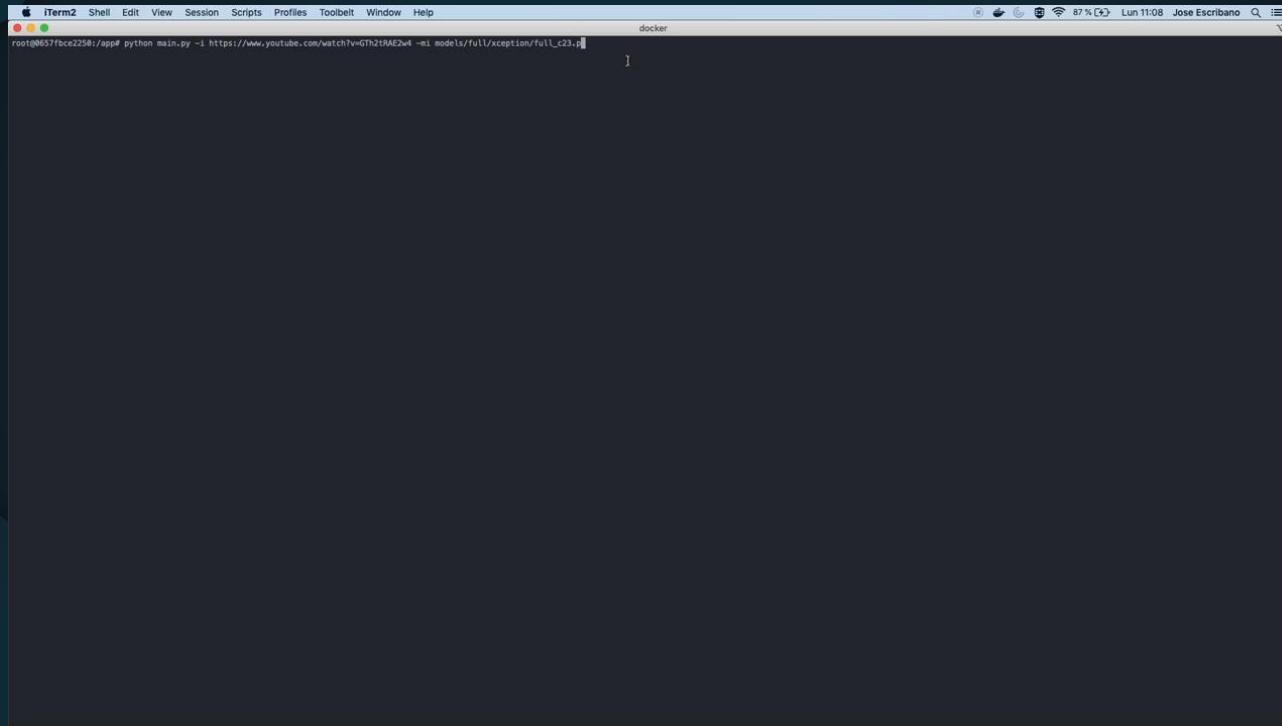
```
user@host:/app# python3 main.py -i https://www.youtube.com/watch?v=GTh2tRAE2w4 -mi models/full/xception
[youtube] GTh2tRAE2w4: Downloading webpage
[youtube] GTh2tRAE2w4: Downloading video info webpage
WARNING: Unable to extract video title
[youtube] GTh2tRAE2w4: Downloading MPD manifest
[download] video.mp4 has already been downloaded
[download] 100% of 1.75MiB
video.mp4
Starting: video.mp4
Model found in models/full/xception/full_c23.p
100%|██████████| 1.75M/1.75M [00:00<00:00, 1.75MiB/s]
The Fake Score is: 90.05763688760807%
Output video in: video.avi
```

<https://github.com/next-security-lab/fakeVideoForensics>



FakeVideoForensics - demo

<https://github.com/next-security-lab/fakeVideoForensics>



A screenshot of an iTerm2 terminal window on a Mac. The window title is "iTerm2". The menu bar includes "File", "Edit", "View", "Session", "Profiles", "Toolbelt", "Window", and "Help". The status bar at the bottom right shows battery level (87%), time (Lun 11:08), and user (Jose Escribano). The main pane displays a command-line session:

```
root@657fbce2258:/app# python main.py -i https://www.youtube.com/watch?v=6Th2tRAE2w4 -m models/full/xception/full_c23.py
```

The terminal has a dark background with light-colored text. The cursor is visible in the middle of the command line.



04

Futuro y conclusiones



Futuro

DARPA The essential elements of media integrity

Digital Integrity
Are the pixels/representations inconsistent?
Blurred edges, replicated pixels, mangled compression?

Physical Integrity
Are the laws of physics violated?
Inconsistent shadows and elongation/compression, vanishing points?

Semantic Integrity
Is a hypothesis about a claim disputable?

DARPA Media Forensics (MediFor) applied to GEOINT

Digital Integrity
Are the pixels or containers inconsistent?
• Sensor validation

Physical Integrity
Are physical laws violated?
• Inconsistent sun angles

Semantic Integrity
Is a claimed date disputable?
• Inconsistent changes
• Temporal disparities

Approved for Public Release, Distribution Unlimited

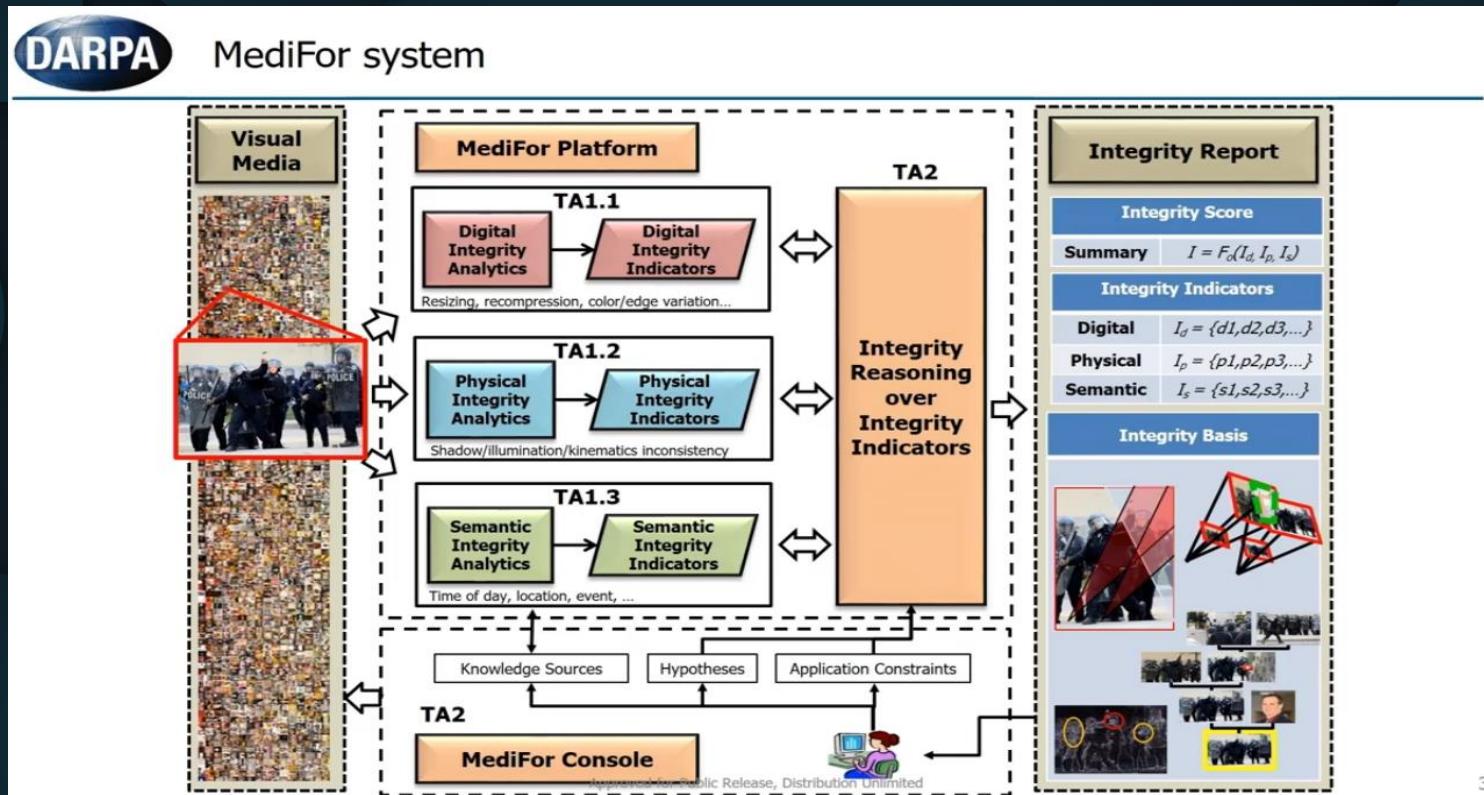
Result & examples courtesy: Kitware, Inc.

Approved for Public Release, Distribution Unlimited

Futuro



MediFor system



3

<https://slideslive.com/38917820/media-forensics-challenges-beyond-the-current-state-of-the-art>
<https://github.com/mediaforensics/medifor>

Futuro

DARPA Future threats

Targeted Personal Attacks
Peelle 2017

The diagram shows a portrait of a man in a dark shirt. Below it is a box labeled "AI Multimedia Algorithms". An arrow points down to a second portrait of the same man, labeled "Highly realistic video".

Generated Events at Scale

The diagram shows a central box labeled "AI Multimedia Algorithms" with an arrow pointing down to three overlapping boxes labeled "Text", "Video & Audio", and "Image". To the left of these boxes is the text "1000s x". Above the boxes is the text "Believable fake events".

Ransomfake concept: Identity Attacks as a service (IAaaS)
Bricman 2019

The diagram shows a central box labeled "AI Multimedia Algorithms" with an arrow pointing down to a box labeled "Forged Evidence", which in turn has an arrow pointing down to a box labeled "Identity Attacks".

Examples of possible fakes:

- Substance abuse
- Foreign contacts
- Compromising events
- Social media postings
- Financial inconsistencies
- Forging identity

Undermines key individuals and organizations

<https://slideslive.com/38917820/media-forensics-challenges-beyond-the-current-state-of-the-art>

Conclusiones

- Vivimos en una **sociedad conectada**, cada vez con más medios y más contenido y es una tendencia creciente.
- Realizar contenido sintético es cada día **más sencillo**, pero hacerlo **realmente creíble no**. Son necesarios datos, procesamiento y conocimientos técnicos.
- Es **necesario correlar** siempre la información que vemos en Internet.
- ¿Existe solución milagrosa? Nosotros creemos que no, pero siempre existirán indicadores que nos permitan saber la veracidad de la información que recibimos.



Herramientas de generación y detección



<https://github.com/next-security-lab/tools-generation-detection-synthetic-content>



@ joseignacio.escribano.pablos.next@bbva.com

@ alfonso.munoz2.next@bbva.com | alfonso@criptored.com

¡Gracias por vuestra (*fake*) atención!

