TECNOLÓGICO DE MONTERREY®

ITESM JOURNAL OF COMPUTER SCIENCE

Submitted to *ITESM Journal of Computer Science*

Vol. 00, No. 1, May 2025, pp. 000–000

ISSN 2694-4022, EISSN 2694-4030

Detección de Deep Fakes

Sebastián Marines
Department of Computer Science, Tecnológico de Monterrey, a01383056@tec.mx

Fernando Martínez
Department of Computer Science, Tecnológico de Monterrey, a01568818@tec.mx

Santiago Posada
Department of Computer Science, Tecnológico de Monterrey, a01383419@tec.mx

Abstract. Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum."

1. Introducción

Dentro de los múltiples retos en la era digital, la verificación de identidad se ha convertido en una de las principales preocupaciones y con ello, un desafío importante. Los *deepfakes* (vídeos manipulados mediante aprendizaje automático) representan una amenaza importante para la autenticidad de identificación remota, lo cual abre la puerta a la suplantación de identidad mediante la mencionada técnica, pues se producen resultados altamente realistas.

Un estudio titulado "Deepfake detection by human crowds, machines, and machine-informed crowds"(11) concluye que las personas por sí solas son menos eficientes que los algoritmos de detección. Por otro lado, en el artículo "Fooled twice: People cannot detect deepfakes but think they can", descubrió que las personas son menos capaces de detectar *deepfakes* de lo que creen.

Los sistemas de verificación de identidad han implementado la prueba de vida (*liveness detection*), que consiste en determinar si el sujeto frente a la cámara está físicamente presente o si se trata de una suplantación. Incluso con estos medios de verificación, los *deepfakes* han sido capaces de engañar a los sistemas de identificación.

Con lo anterior en consideración, se observa que la amenaza es no sólo real, sino que es de alta urgencia, pues los sistemas de IA solamente se harán mejores, lo cuál puede complicar aún más la problemática en el futuro cercano.

Para nuestra investigación planteamos las siguientes hipótesis:

- Los videos e imágenes alterados mediante técnicas de deep fakes contienen inconsistencias sutiles como la desalinización del movimiento de los labios con el audio, artefactos en las texturas faciales, etc. Si consideramos imágenes en las cuales existan estas desalineaciones entonces es posible detectar *deepfakes* mediante una red neuronal con alta precisión.
- Si un algoritmo de reconocimiento facial es lo suficientemente preciso, entonces sí se puede comparar una selfie tomada como prueba de vida con la imagen de un documento oficial como una INE.

2. Marco Teórico

El problema de la investigación se centra en desarrollar un sistema para la detección de intentos de fraude de identidad basados en video, abordando tres puntos fundamentales que se encuentran interrelacionados: (1) detección de *deepfakes*, (2), verificación de liveness y (3) comparación facial entre la persona y su documento de identidad.

Esta investigación resulta altamente relevante debido a la incidencia en crecimiento de intentos de fraude mediante herramientas de inteligencia artificial (incremento de 10x entre 2022 y 2023)(13), que únicamente se tornará más fina conforme avance la tecnología(3). Este proyecto busca diseñar e implementar un sistema basado en aprendizaje automático que integre estas las mencionadas capas de protección, ofreciendo una solución más robusta que los enfoques actualmente disponibles.

Dependiendo del sistema de detección de deepfakes, se emplean diferentes metodologías para la detección de los mismos. Esfuerzos tempranos se enfocaron en distintas pistas como que un ojo no parpadee, o poses de cara inconsistentes. Otros modelos más sofisticados funcionan con CNNs, que detectan cuestiones más particulares causadas por el proceso de falsificación como *face-blending*. Es prometedora la investigación que busca detectar los *deepfakes* mediante inconsistencia en el movimiento, particularmente en movimientos causados por el habla(1).

Un problema particular que encuentran los modelos de detección de deepfakes, es que, a pesar de su alta efectividad cuando tienen un entrenamiento con una base de datos con fotos y videos falsos, estas pueden ser propensas a un ataque de adversario (*adversary attacks*), que consiste en manipular la base de datos y por tanto atrofiar el modelo final de detección(9).

Una cuestión que también resulta problemática actualmente con los modelos de detección, es que tienen dificultades para generalizar (funcionar ante datos nuevos) ante las diversas técnicas utilizadas en la manipulación de contenido. Este problema surge principalmente de la compleja interacción entre texturas (patrones finos) y artefactos (errores o distorsiones) en los datos de deepfake, aspectos que los métodos de detección tradicionales frecuentemente pasan por alto. (2)

Se ha buscado a su vez encontrar maneras que los sistemas de detección de deepfakes puedan ser funcionales cuando las imágenes sean de baja calidad. Esto es una situación que ha sido tema de estudio por el tema de que en ciertos países los teléfonos celulares que predominan son de gama baja, que provienen de China, con cuyas fotografías no fueron entrenados los modelos de detección. En su mayoría, los modelos de detección fueron entrenados con imágenes de alta calidad. (14)

En cuanto la detección de voces falsas, los modelos del estado del arte utilizan modelos de aprendizaje profundo como el aprendizaje autosupervisado, donde modelos como Wav2vec extraen características del audio y un clasificador define si es falso o no. El problema es que, como otros modelos de detección, se encuentran en problemas cuando se topan con deepfakes generados con nuevas tecnologías. Se propone un nuevo modelo llamado SLIM (Style-Linguistics Mismatch), el cual detecta una desalineación artificial entre el contenido lingüístico y el estilo de la voz. Este modelo novedoso ofrece mayor generalización, no necesita de más datos y resulta más explicable que otros modelos, lo cuál eleva el nivel de confianza (10).

3. Metodología

Variable dependiente: Se clasificará entre 1 o 0, para detectar entre video real, o *deepfake*, respectivamente.

Variables independientes:

- La detección de artefactos visuales (errores visuales o anomalías, como deformidades, texturas extrañas, etc.) en los videos.
- Distancia vectorial entre rasgos faciales en la selfie y el documento de identidad que se tomará en cuenta.

Covariables:

- Calidad de imagen/video, en donde influye también la iluminación y los ángulos.
- El generador de video/imagen con la cual se hicieron los *deepfakes*.

3.1. Recolección de datos

- Se emplean los siguientes *datasets*:
  - FaceForensics++;
  - DFDC;
  - Deepfake Detection Challenge 2020;
  - Celeb-DF;
  - Dataset propio: Se recolectaron selfies y documentos de identificación. Por cuestiones prácticas, se tomaron 50 muestras, correspondientes a 25 mujeres y 25 hombres.
- En cuanto a la limpieza de datos:
  - Funciones automatizadas para organizar y dividir datasets en conjuntos de entrenamiento y evaluación
  - Se hará un proceso de filtración para archivos corruptos, incompletos o duplicados;
  - Habrá un proceso de equidad dentro del conjunto de datos, en cuanto a equidad entre deepfakes/no deepfakes, para evitar sesgos en el entrenamiento;
  - Normalización de resoluciones.

3.2. Análisis exploratorio

- Se compararán similitudes faciales entre imágenes/videos reales y *deepfakes*;
- Se evaluarán características similares entre *deepfakes* e imágenes/videos reales.

3.3. Modelos e inferencia

- Implementación de una red neuronal convolucional basada en transfer learning con la arquitectura EfficientNet-B0 pre-entrenada.
- Optimización mediante Adam con tasa de aprendizaje adaptiva
- Se entrenará una red neuronal convolucional para detectar *deepfakes*;
- Para reducir el tiempo de entrenamiento del modelo y aumentar la precisión del mismo, se usa transfer learning en modelos como ImageNet con arquitectura Xception, lo cual permitió utilizar las capacidades de ese modelo para entrenar un modelo a la medida;
- En cuanto la comparación facial, se usa un modelo pre-entrenado, ArcFace, para extraer puntos de referencia del rostro en el documento proporcionado (INE, pasaporte, etc.) y el rostro extraído del video. Después se calcula la distancia vectorial entre ambos resultados para determinar una coincidencia.

- Se tienen las siguientes métricas de resultados:
  - Accuracy (exactitud)
  - Precision (precisión)
  - Recall (sensibilidad)
  - F1 Score
- Implementación de un sistema de seguimiento del proceso de entrenamiento
- Monitorización de métricas como pérdida y precisión durante el entrenamiento
- Detección temprana de problemas como sobreajuste mediante validación continua
- Almacenamiento del modelo entrenado (evitar reentrenamiento)

4. Resultados y discusión

Código

Se implementó InsightFace que es la implementación oficial, -o por lo menos, más mantenida y robusta- de ArcFace, respaldada por la comunidad de deep learning.

Se realizó la prueba en donde se pueden subir imágenes, con sus respectivas identificaciones, para ver si se trata de la misma persona o no. Los resultados fueron muy positivos incluso tomando en cuenta la mala calidad de los documentos de identidad y el hecho de que no no necesariamente son fotos recientes.

Los resultados fueron especialmente exitosos en diferenciar cuando no se trata de la misma persona.

Los resultados son representados en matrices de similitud, que usan similitud coseno que da un valor entre 0 y 1, este último tratándose de rostros idénticos. Estas se obtienen vectorizando (vectores de características) las imágenes, y entonces se comparan las similitudes.

Comparando una selfie con una identificación de la misma persona, bajo iluminaciones muy diferentes y diferencias físicas encontramos 61 % de similitud, lo cual dada las circunstancias mencionadas, resulta excelente.

```
[[1.         0.61308205]
 [0.61308205 1.0000002 ]]
```

Figure 1 Matriz de similitud para la misma persona

Comparando una selfie con una INE de diferente persona, encontramos aún mejores resultados, puesto que con gran seguridad afirma que no es la persona que "pretende" ser (.03 % de que sean la misma persona).

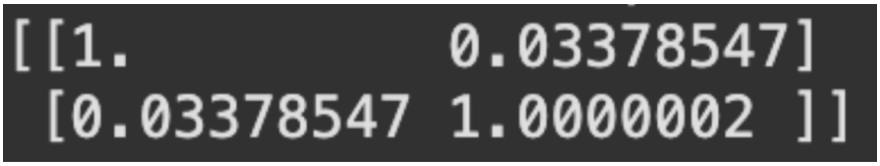

Figure 2 Matriz de similitud para diferentes personas

Ahora mismo, nos encontramos obteniendo los datos necesarios para crear un dataset, el cual contendrá fotos de los participantes, junto con su respectiva identificación. Por el momento tenemos contemplados cincuenta participantes por motivos de agenda. Con este conjunto de datos, realizaremos pruebas para ver los resultados (en gráficos y demás), cuando se trata de lidiar con un conjunto de datos mayor.

Código

Desarrollamos un sistema de detección basado en características visuales utilizando el dataset Celeb-DF. Implementamos una función para organizar y dividir automáticamente los datasets en conjuntos de entrenamiento y evaluación, con una distribución del 80 % y 20 % respectivamente. Este desequilibrio será un punto a tener en cuenta en el dataset utilizado para las siguientes entregas.

Utilizamos transfer learning con la arquitectura EfficientNet-B0 pre-entrenada, adaptando la capa de clasificación para la detección binaria de deepfakes. Esta aproximación nos permitió aprovechar las capacidades de extracción de características de una red convolucional ya entrenada en millones de imágenes, acelerando el proceso de entrenamiento y mejorando la capacidad de generalización del modelo.

Desarrollamos un sistema de entrenamiento con optimización Adam, scheduler de tasa de aprendizaje adaptativo, y monitoreo de métricas de entrenamiento. Este pipeline completo permite el seguimiento detallado del proceso de entrenamiento y facilita la detección temprana de problemas como el sobreajuste.

Implementamos un módulo de evaluación que genera informes detallados incluyendo matriz de confusión, precisión, recall y F1-score para la detección de videos reales versus manipulados. También desarrollamos herramientas de visualización para analizar la evolución del entrenamiento y las métricas de rendimiento a lo largo de las épocas, permitiendo una interpretación intuitiva de los resultados.

También implementamos la funcionalidad para guardar el modelo entrenado, permitiendo su uso posterior en sistemas de verificación sin necesidad de reentrenamiento. Este modelo persistente puede ser integrado fácilmente con otros componentes del sistema de verificación de identidad.

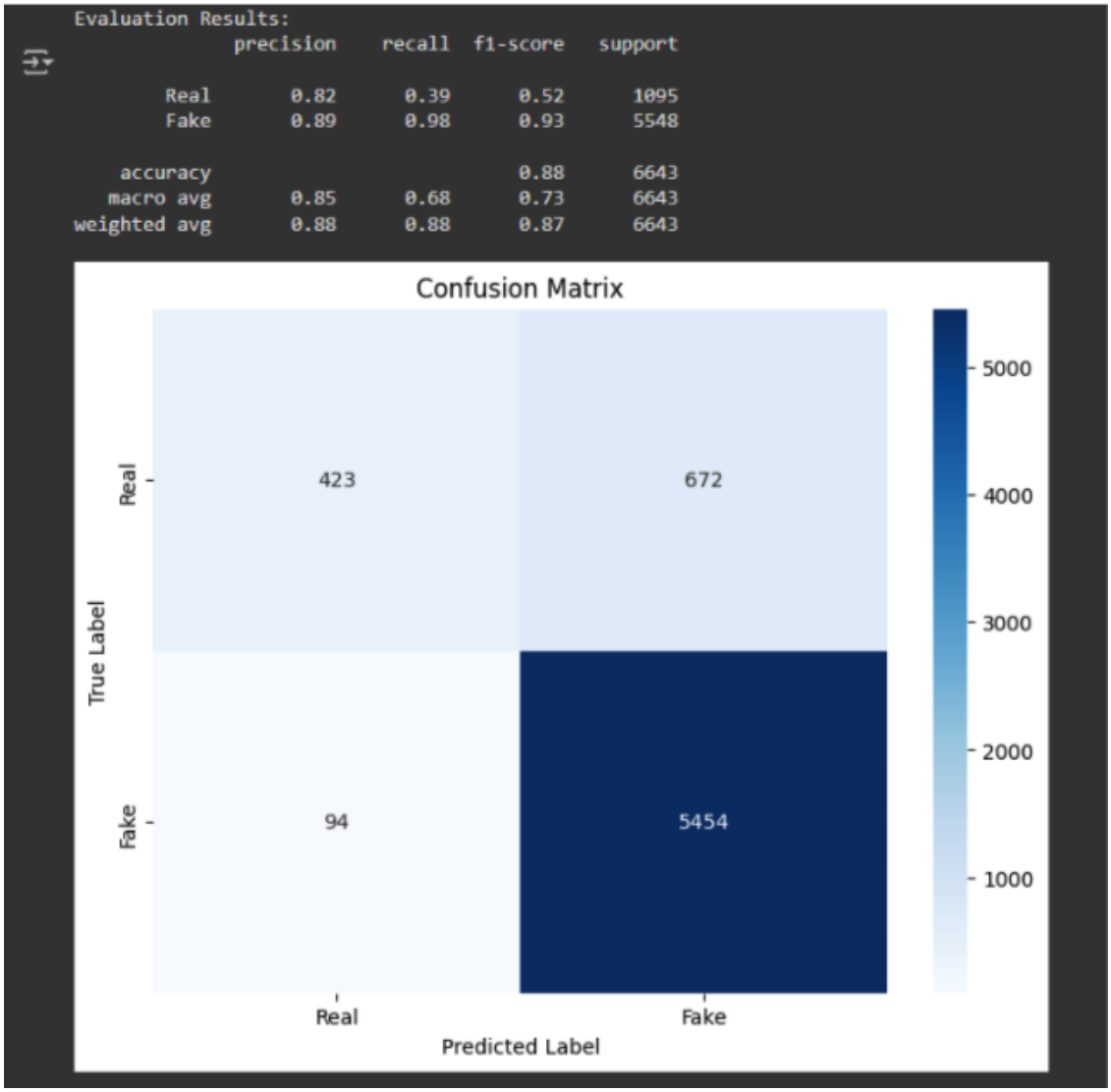

Figure 3 Resultados de la evaluación y matriz de confusión

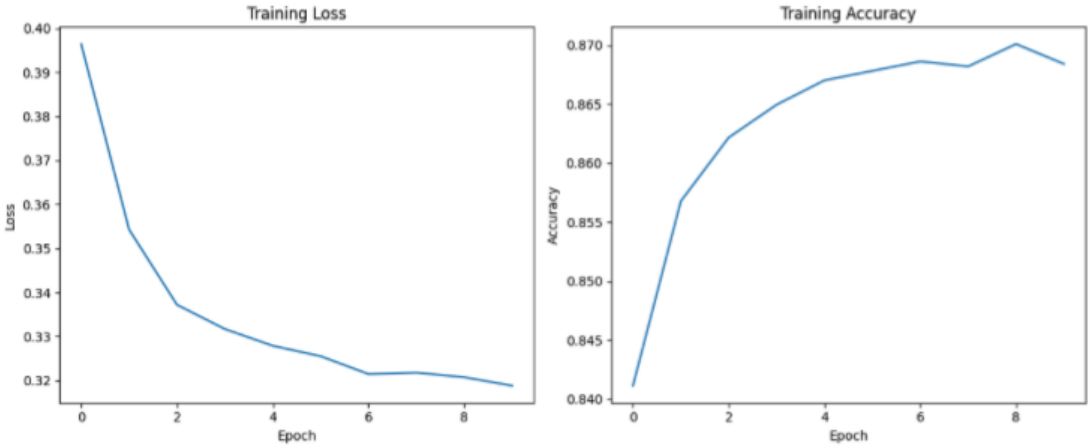

Figure 4 Gráficos de pérdida y precisión durante el entrenamiento

5. Discusión sobre hallazgos

Los resultados de ambos enfoques demuestran la viabilidad de nuestras técnicas para la detección de deepfakes, complementando el sistema de verificación facial desarrollado previamente con InsightFace. La integración de estos componentes proporciona una solución más robusta para la verificación de identidad.

Referencias

[1] https://arxiv.org/pdf/2305.05282

[2] https://arxiv.org/pdf/2408.00388v1

[3] https://www.theguardian.com/technology/2024/apr/08/time-is-running-out-can-a-future-of

[4] https://github.com/deepinsight/insightface

[5] https://github.com/deepinsight/insightface/blob/master/examples/demo_analysis.py

[6] https://www.iproov.com/es/press/study-reveals-deepfake-blindspot-detect-ai-generated-c

[7] https://keras.io/guides/transfer_learning/

[8] https://keras.io/api/applications/xception/

[9] https://kth.diva-portal.org/smash/get/diva2:1795907/FULLTEXT01.pdf

[10] https://proceedings.neurips.cc/paper_files/paper/2024/file/7d6930fd71740eae21224a5ffb70cb8c-Paper-Conference.pdf

[11] https://www.pnas.org/doi/epdf/10.1073/pnas.2110013119

[12] https://www.sciencedirect.com/science/article/pii/S2589004221013353

[13] https://sumsub.com/newsroom/sumsub-research-global-deepfake-incidents-surge-tenfold-fr

[14] https://www.wired.com/story/generative-ai-detection-gap/?utm_source=chatgpt.com