Generación de un set de imágenes para la detección de Deepfake en Modelos de Redes Neuronales Convolucionales

Generation of a Data Set for Convolutional Neuronal Networks in Deepfake

Jesus Antonio Cisneros Rivera[1], Misael López Ramírez[1], Luis Manuel Ledesma Carrillo[1]
[1]Departamento de Estudios Multidisciplinarios, Universidad De Guanajuato, Guanajuato, México
ja.cisnerosrivera@ugto.mx[1]

Resumen

Con los avances de la Inteligencia Artificial y el internet, es bastante común encontrarse con imágenes deepfake que tienen distintos objetivos como pueden ser: desinformación, fraudes o suplantación de identidad. Para contrarrestar estos casos, existen modelos de Redes Neuronales Convolucionales (CNN) los cuales se necesita entrenar con un conjunto de datos de rostros reales o manipulados. El objetivo del trabajo es optimizar la extracción de rostros reales y falsos para generar una base de datos. Para ello, se desarrolló un programa para la extracción de rostros tanto en videos reales como videos deepfake donde se han extraído el rostro a 2000 videos, obteniendo un total de 951, 231 rostros. Las imágenes extraídas se les aplica un acercamiento y las guarda en subcarpetas con 10 000 imágenes cada una.

Palabras clave: Deepfake, Base de datos, Detección facial, Red Neuronal Convolucional, Inteligencia Artificial, Clasificación.

Introducción

La reputación y la credibilidad son activos fundamentales en nuestra sociedad. Según Malik et al. (2022), la presencia y/o divulgación de contenido manipulado, conocido como deepfake, puede afectar negativamente la confianza de los usuarios en las distintas plataformas digitales y, por ende, es necesario tomar medidas para prevenir la difusión de información engañosa. Mukta et al. (2023) nos menciona que la palabra deepfake, proviene de Deep Learning (DL) y Fake.

Los deepfake son contenido en base a sonidos, imágenes o video, los cuales son manipulados para generar una nueva versión de los originales. Nirkin et al. (2022) investigaron que usan aplicaciones con apoyo de la inteligencia artificial, para intercambiar rostros que son demasiado realistas. Por lo tanto, estas aplicaciones permiten a cualquier usuario intercambiar los rostros, lo que viene siendo, boca, ojos, nariz, piel, cejas, basándonos en lo escrito previamente, las imágenes deepfake, en el caso de los rostros, son imágenes realistas de personas que no existen o contienen la manipulación de algún rostro que es real. Esto con el fin de que sea conveniente para la persona que este manipulando dicha imagen.

Existen modelos que actualmente ya han sido desarrollados dentro de la arquitectura de CNN. Estos modelos demuestran ser prometedores para la detección del contenido que ha sido manipulado (Tthing, 2023; Guarnera et al., 2022). Estos modelos tienen que ser entrenados con un gran volumen de datos. Sin embargo, una limitación de los conjuntos de datos existentes es la inclusión de información irrelevante, lo que provoca ruido al entrenar los modelos y la precisión disminuya. Este trabajo busca reducir el ruido y mejorar la precisión de los modelos mediante una base de datos compuesta por rostros extraídos de videos que han sido etiquetados como reales o falsos. El conjunto de datos se enfocará en las características faciales, para reducir la complejidad del análisis de la imagen completa y potenciar la efectividad de los modelos CNN para identificar manipulación de rostros.

Metodología Propuesta

Los videos provienen de una base de datos llamada FaceForensics++, contiene 1000 videos de personas reales subidos a YouTube. Además, cada uno de los videos, han sido manipulados con técnicas de manipulación de rostros, generando otros 1000 videos que son deepfake. Los videos al estar en su respectiva carpeta, real y falso, nos ayuda a facilitar la clasificación para nuestra Base de Datos.

Extracción de Rostros

La creación de la base de datos se realizó utilizando el lenguaje de programación Python y las herramientas de OpenCV, El proceso inicia cargando las carpetas donde se encuentran nuestros videos reales y deepfake. Cada uno de estos videos son procesados utilizando la librería de OpenCV. Para detectar los rostros se utilizó el algoritmo de clasificador en cascada, específicamente el clasificador “haarcascade frontal face default". Este clasificador permite detectar rostros en videos o imágenes, basándose en patrones de características faciales. Fue elegido ya que demostró ser efectivo, aunque los rostros tuvieran distintas dimensiones, iluminación, expresiones o pose.

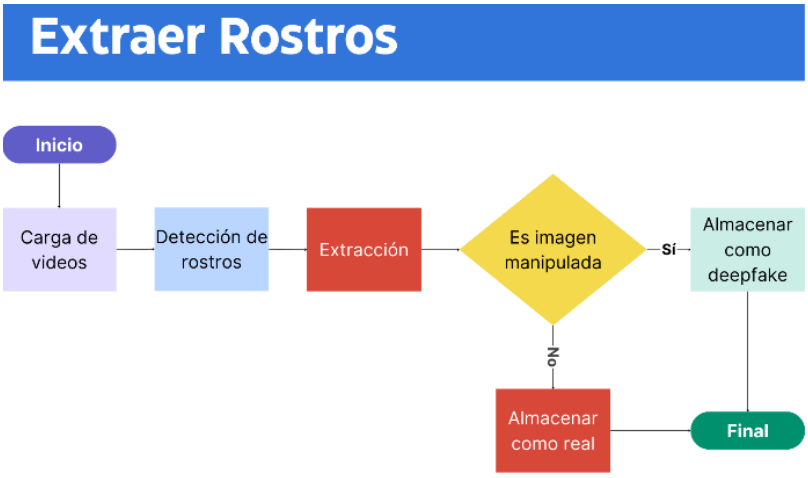

Figura 1. Pasos para la extracción de rostros
Fuente: Autoría propia

El proceso de extracción consistió en los siguientes pasos:

- **Carga de video**: Con la ayuda de la librería de OpenCV, cada video es cargado y es descompuesto en fotogramas.
- **Detección de rostros**: En cada uno de los fotogramas se ha aplicado el algoritmo de clasificador de cascada, devolviéndonos las coordenadas del rostro dentro del fotograma.
- **Extracción**: Al detectar la posición del rostro dentro del fotograma, es extraído del fotograma, se aplica un acercamiento del 40% ya que la sección extraída contiene ruido y son partes innecesarias para el modelo, el ruido viene siendo: la frente, pelo o fondo. También es redimensionado a un tamaño de 300x300 y los deja en formato de 8 bits.
- **Almacenamiento**: El almacenamiento se divide en dos etiquetas, reales y falsas. Con el fin de optimizar el aprendizaje del modelo se necesita colocar todas las imágenes extraídas en subcarpetas, por lo que las carpetas reales y falsas contendrán subcarpetas que almacenarán un máximo de 10 000 imágenes de rostros extraídos.

Resultados

Se ha obtenido una base de datos con los rostros detectados y extraídos de los videos (Figura 2). Como se puede observar, la extracción se enfoca en la parte central del rostro, disminuyendo el ruido para nuestro modelo y obtener buenos resultados al entrenar un modelo CNN.

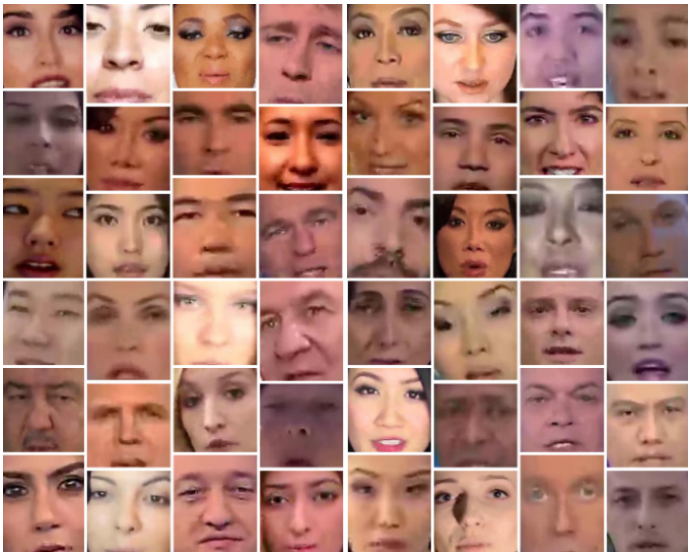

Figura 2. Imágenes de rostros extraídos
Fuente: Autoría propia

En la figura 3, se puso a prueba la eficacia de la base de datos con un modelo CNN, la base de datos se dividió en el set de entrenamiento y prueba, para comprobar la eficacia del modelo entrenado se usó un conjunto de 20 000 imágenes tanto reales como falsas. El modelo ha demostrado buenos resultados al predecir si es real o falso. Además, en la Tabla 1, se puede observar algunas métricas donde se han obtenidos resultados arriba del 0.92.

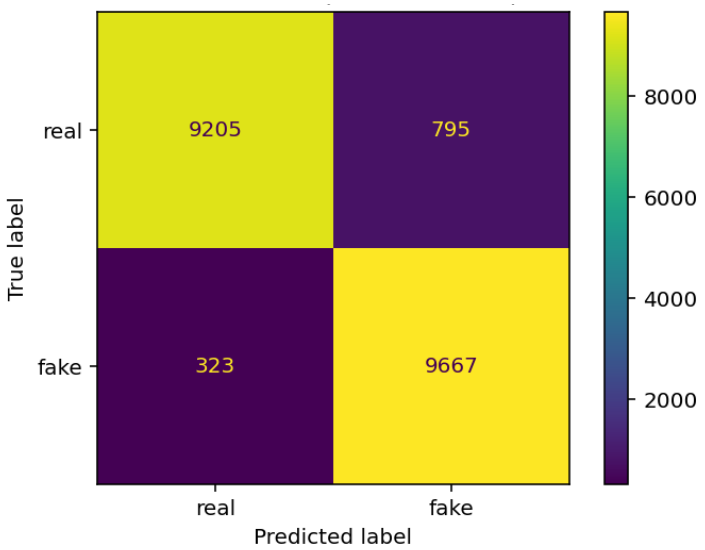

Figura 3. Matriz de confusión de un conjunto de imágenes de prueba
Fuente: Autoría propia

***Tabla 1**. Métricas del Modelo CNN al ser evaluado utilizando la Base de Datos.*

Métrica	Resultado
Recall	0.9676676676676677
F1-Score	0.9453354195188735
Accuracy	0.944072036018009
ROC AUC	0.9440838338338338
Precisión clase "real"	0.9660999160369438
Precisión clase "falso"	0.9240107054100555

Fuente: Autoría propia

Conclusión

El presente trabajo demuestra que la creación de un conjunto de datos con rostros extraídos de videos, mejora significativamente la precisión de los modelos al reducir el ruido en el entrenamiento. Al eliminar información irrelevante, se optimiza la calidad de los datos, permitiendo que el modelo se enfoque en patrones y características más representativos para la detección. Estos resultados confirman la importancia de un conjunto de datos bien estructurado y etiquetado, lo que proporciona una base sólida para futuras investigaciones en detección de deepfakes.

Referencias

1. Guarnera, L., Giudice, O., Guarnera, F., Ortis, A., Puglisi, G., Paratore, A., Bui, L. M. Q., Fontani, M., Coccomini, D. A., Caldelli, R., Falchi, F., Gennaro, C., Messina, N., Amato, G., Perelli, G., Concas, S., Cuccu, C., Orrù, G., Marcialis, G. L., & Battiato, S. (2022). The Face Deepfake Detection challenge. Journal of Imaging, 8(10), 263. https://doi.org/10.3390/jimaging8100263
2. Malik, A., Kuribayashi, M., Abdullahi, S. M., & Khan, A. N. (2022). DeepFake detection for human face images and videos: A survey. IEEE access: practical innovations, open solutions, 10, 18757–18775. https://doi.org/10.1109/access.2022.3151186
3. Mukta, M. S. H., Ahmad, J., Raiaan, M. A. K., Islam, S., Azam, S., Ali, M. E., & Jonkman, M. (2023). An investigation of the effectiveness of deepfake models and tools. Journal of Sensor and Actuator Networks, 12(4), 61. https://doi.org/10.3390/jsan12040061
4. Nirkin, Y., Wolf, L., Keller, Y., & Hassner, T. (2022). DeepFake detection based on discrepancies between faces and their context. IEEE Transactions on Pattern Analysis and Machine Intelligence, 44(10), 6111–6121. https://doi.org/10.1109/TPAMI.2021.3093446
5. Thing, V. L. L. (2023). Deepfake detection with deep learning: Convolutional neural networks versus Transformers. En arXiv [cs.CR]. https://doi.org/10.48550/ARXIV.2304.03698