Desarrollo de una herramienta digital para la detección de videos generados por ‘DeepFakes’

Bojalil Abiti, David

2024

Desarrollo de una herramienta digital para la detección de videos generados por 'DeepFakes'

Bojalil Abiti David (séptimo semestre en Ingeniería en Sistemas Computacionales)[1,*], Quintana Delgadillo Iñigo (séptimo semestre en Ingeniería en Sistemas Computacionales)[1], Díaz Hurtado Naomi (séptimo semestre en Ingeniería en Sistemas Computacionales)[1], López Cruz Lesther Emmanuel (profesor responsable)[1], Acevedo Escalante Manuel F. (profesor asesor)[1], Pérez Aguirre Rafael (profesor asesor)[1].

[1] *Universidad Iberoamericana Puebla, San Andrés Cholula, Puebla, México*

Palabras clave: CNN, DeepFake, Detección, IA, Tensorflow.
[*]**Autor Corresponsal:** david.bojalil @iberopuebla.mx

Introducción

El avance en el desarrollo de sistemas que emulan las funciones cognitivas humanas ha sido fundamental en la evolución de la inteligencia artificial (IA). Durante el último siglo, el interés por construir sistemas inteligentes ha dado lugar a la creación de modelos inspirados en el cerebro humano [1]. Estos modelos, destacan por su capacidad para procesar grandes volúmenes de datos y tomar decisiones informadas en tiempo real, lo que ha favorecido su creciente implementación en múltiples campos.

A pesar de estos progresos, el uso de inteligencia artificial plantea desafíos éticos significativos, particularmente en lo referente a las redes generativas adversarias (GAN). Estas redes, que pueden generar datos realistas a partir de datos previamente entrenados, se encuentran en una controversia creciente debido a su mal uso [2]. Un caso particularmente preocupante es el de los DeepFakes, videos hiperrealistas manipulados digitalmente para hacer que personas aparezcan diciendo o haciendo cosas que en realidad no sucedieron [2]. Más allá de la simple distorsión de la realidad, los DeepFakes se han convertido en herramientas para la desinformación, la manipulación de la opinión pública, la difamación y otros actos fraudulentos.

Dada a la falta de marcos éticos rigurosos en estas innovaciones tecnológicas, se vuelve imperativo desarrollar soluciones efectivas que permitan detectar los casos de DeepFakes. Este proyecto se enfoca en desarrollar una herramienta digital basada en técnicas de aprendizaje automático, diseñada para la detección de patrones en videos generados por DeepFakes. Esta iniciativa pretende contribuir a la seguridad y la ética en el ámbito digital, proponiendo una respuesta concreta a los riesgos de manipulación audiovisual que afectan tanto a individuos como a la sociedad en su conjunto.

Metodología

El desarrollo del proyecto, ~~se~~ implementó una metodología estructurada que comenzó con la adquisición y análisis del conjunto de datos, clasificando los videos en dos carpetas denominadas "original" y "deepfake". Se garantizó un balance entre ambas categorías, lo cual fue esencial para minimizar sesgos y obtener un modelo confiable.

En la fase de diseño del modelo de inteligencia artificial, se investigaron diversas arquitecturas de redes neuronales convolucionales (CNN). La selección de la arquitectura se basó en cuatro criterios esenciales: capacidad de detección de características finas, rendimiento óptimo, flexibilidad adaptativa y eficiencia computacional.

Para el entrenamiento del modelo, los datos fueron preprocesados y segmentados en conjuntos de entrenamiento y prueba. Se seleccionaron y ajustaron cuidadosamente los hiperparámetros clave, como el número de épocas, la tasa de aprendizaje y el tamaño de lote, para garantizar un aprendizaje eficiente del modelo.

La evaluación y validación del sistema se realizó mediante diversas métricas complementarias, como la matriz de confusión, exactitud, precisión, sensibilidad, especificidad y puntuación F1. Estas métricas permitieron evaluar su desempeño desde diferentes perspectivas, asegurando un modelo robusto y efectivo para la detección de DeepFakes.

Resultados

El modelo de detección de DeepFakes fue evaluado a través de varias métricas clave. La precisión del modelo en el conjunto de prueba fue del 63.86%, lo que indica que aproximadamente dos de cada tres predicciones fueron correctas. El valor de la pérdida en el conjunto de prueba fue de 0.5945, mientras que la media del error cuadrático (MSE) fue de 0.2095, lo que sugiere que el modelo tiene margen para mejorar su rendimiento. En cuanto a las métricas de clasificación, la precisión fue del 61.92%, lo que significa que el modelo clasificó correctamente poco más del 60% de los casos positivos, pero aún cometió algunos falsos positivos. El recall alcanzó un 72.03%, indicando que el modelo fue

capaz de identificar más del 70% de los DeepFakes presentes en el conjunto de prueba. La puntuación F1 fue de 66.59%, lo que refleja un equilibrio razonable entre precisión y recall.

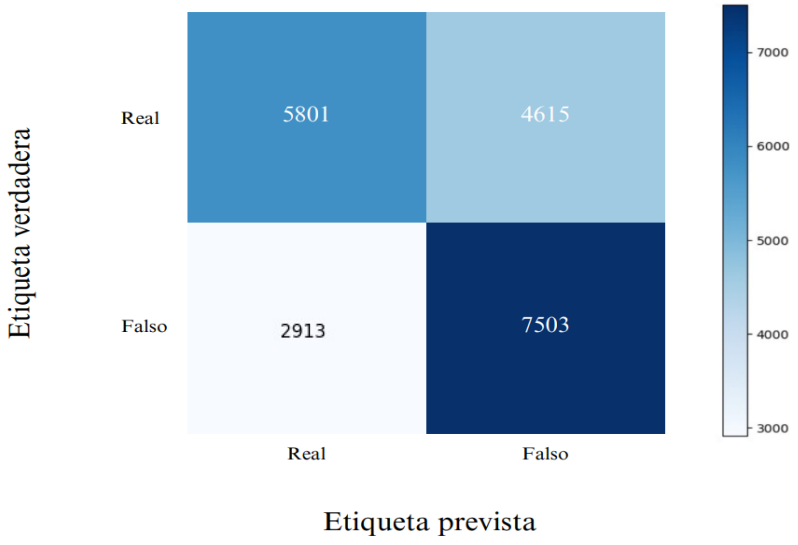

figura 1. Matriz de confusión del modelo sobre el conjunto de datos balanceado. Los valores indican clasificaciones correctas e incorrectas para las etiquetas "Real" y "Falso". La escala de color indica la densidad de los valores en cada celda, destacando las áreas con mayor frecuencia.

La matriz de confusión muestra la distribución del modelo para la predicción de real y falso. En ella, los valores en la diagonal principal (5801 y 7503) representan las predicciones correctas, mientras que los valores fuera de la diagonal principal (4615 y 2913) corresponden a los errores del modelo. El número 5801 corresponde a las predicciones correctas de videos originales clasificados como originales, y el 7503 corresponde a los DeepFakes correctamente clasificados como tal. Sin embargo, el modelo también cometió errores, clasificando 4615 videos originales como deepfakes y 2913 deepfakes como videos originales.

Análisis de resultados

Los resultados obtenidos indican que el modelo tiene un rendimiento moderado en la detección de DeepFakes. Con una precisión del 61.92% y un recall del 72.03%, el modelo es eficaz para identificar la mayoría de los DeepFakes, pero todavía presenta una tasa considerable de falsos positivos. La precisión moderada y el recall más alto sugieren que el modelo prioriza la detección sobre evitar falsos positivos. La exactitud global del 63.86% y la pérdida de 0.5945 indican que el modelo puede mejorarse para reducir las equivocaciones de sus predicciones entre real y falso. La puntuación F1 de 66.59% demuestra que hay un equilibrio aceptable entre la capacidad de detectar DeepFakes y minimizar los falsos positivos, aunque optimizar el modelo para mejorar estas métricas sigue siendo crucial, especialmente en DeepFakes más complejos o en condiciones difíciles como baja iluminación.

Conclusiones

El prototipo desarrollado demuestra capacidad funcional para la detección de DeepFakes, aunque existen oportunidades para su optimización. Las limitaciones principales incluyen la pérdida de precisión en el seguimiento durante movimientos rápidos, una mayor susceptibilidad a falsos negativos en condiciones de baja iluminación y ciertas dificultades para identificar DeepFakes de alta complejidad tecnológica. No obstante, el sistema actual exhibe una efectividad notable en la identificación de DeepFakes comunes en redes sociales, especialmente aquellos que comparten similitudes con el conjunto de datos utilizado en el entrenamiento.

La escalabilidad del proyecto presenta un alto potencial, ya que la implementación de una plataforma que permita la recopilación continua de nuevos datos contribuiría a enriquecería el conjunto de entrenamiento, aumentando así la precisión y robustez del modelo de detección. Con futuras mejoras, el sistema podría llegar a convertirse en una herramienta de detección más completa y adaptable a los avances tecnológicos en la generación de DeepFakes.

Referencias

1. R. López, **Algunas reflexiones sobre el presente y futuro de la Inteligencia Artificial**. Csic.es, Sep. 2016, doi: issn: 0211-2124.
2. F. J. García-Ull. (2021). ***Deepfakes*: el próximo reto en la detección de noticias falsas**. Anàlisi: Quaderns de Comunicació i Cultura, 64, 103-120. DOI: https://doi.org/10.5565/rev/analisi.3378
3. **Introducción**. *Google for Developers*, 2022. https://developers.google.com/machine-learning/gan?hl=es-419
4. J. J. Bird, **DEEP-VOICE: DeepFake Voice Recognition**. *Kaggle.com*, 2023. https://www.kaggle.com/datasets/birdy654/deep-voice-deepfake-voice-recognition
5. L. Gergely Ferenc; G. Gergely. **Deepfake y desinformación –¿Qué puede hacer el derecho frente a las noticias falsas creadas por deepfake?** IDP. Revista de Internet, Derecho y Política, 2024, n.º 41, pp. 1-13, https://www.raco.cat/index.php/IDP/article/view/n41-lendvai
6. L. Rouhiainen, **INTELIGENCIA ARTIFICIAL 101 COSAS QUE DEBES SABER HOY SOBRE NUESTRO FUTURO.** 2018. Available: https://planetadelibrosec0.cdnstatics.com/libros_contenido_extra/40/39308_Inteligencia_artificial.pdf
7. R. González Duque. **Python para todos**. Madrid, https://repositorio.uci.cu/bitstream/123456789/10206/1/Python_para_todos.pdf
8. S. Tiwarekar, **Deep Fake Detection (DFD) Entire Original Dataset**. *Kaggle.com*, Jun. 12, 2024, https://www.kaggle.com/datasets/sanikatiwarekar/deep-fake-detection-dfd-entire-original-dataset
9. L. Gergely Ferenc; G. Gergely. **Deepfake y desinformación –¿Qué puede hacer el derecho frente a las noticias falsas creadas por deepfake?**. IDP. Revista de Internet, Derecho y Política, 2024, n.º 41, pp. 1-13, https://www.raco.cat/index.php/IDP/article/view/n41-lendvai
10. A. Valle Barrio, **Aplicación de Tensorflow en deep learning**. Tesis, Universidad Politécnica de Madrid, Madrid, 2018. https://oa.upm.es/53815/
11. T. O. Ayodele, **Machine learning overview**. New Advances in Machine Learning, vol. 2, no. 9-18, p. 16, 2010.
12. L. Rouhiainen, **Inteligencia artificial.** in Inteligencia artificial. Madrid: Alienta Editorial, 2018, pp. 20-21.
13. A. Maisueche. **UTILIZACIÓN DEL MACHINE LEARNING EN LA INDUSTRIA 4.0**. Sept. 2019. Valladolid. Available: https://uvadoc.uva.es/bitstream/handle/10324/37908/TFM-I-1372.pdf?sequence=1
14. **Diccionario sobre inteligencia artificial: 100 conceptos claves sobre sistemas inteligentes**, 1ª ed. TN Editorial, TN University. Marzo 2024. [Online]. Available: https://www.tnuniversity.edu.mx/docs/newsletter/Portadas/Diccionario%20sobre%20Inteligencia%20Artificial.pdf
15. ***Deepfake, N.*** Oxford English Dictionary, 2023. https://doi.org/10.1093/OED/9547101155
16. M. R. Costa-Jussá and J. A. Fonollosa. **DeepVoice: tecnologías de aprendizaje profundo aplicadas al procesado de voz y audio**. Procesamiento del Lenguaje Natural, vol. 59, pp. 117-120, 2017. http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/viewFile/5500/3259.