Detección de Videos Multi-Rostro Manipulados por Deepfake a Partir de Aprendizaje Profundo

Santiago Ballesteros Pastrán

Catalina Jiménez Varón

Asesora: Haydemar María Núñez Castro

Departamento de Ingeniería de Sistemas y Computación

Facultad de Ingeniería

Universidad de los Andes

Abstract

Resumen

Contenido

[1 Introducción 2](#_Toc103077913)

[2 Descripción General 2](#_Toc103077914)

[2.1 Objetivos 2](#_Toc103077915)

[2.2 Antecedentes 3](#_Toc103077916)

[3 Diseño y Especificaciones 3](#_Toc103077917)

[3.1 Definición del Problema 3](#_Toc103077918)

[3.2 Especificaciones 3](#_Toc103077919)

[3.3 Restricciones 3](#_Toc103077920)

[4 Desarrollo del Diseño 3](#_Toc103077921)

[4.1 Alternativas de Diseño 3](#_Toc103077922)

[5 Implementación 3](#_Toc103077923)

[5.1 Entendimiento de Datos 3](#_Toc103077924)

[5.2 Preparación de Datos 3](#_Toc103077925)

[5.3 Modelaje y Evaluación 3](#_Toc103077926)

[5.4 Despliegue 3](#_Toc103077927)

[6 Conclusiones 3](#_Toc103077928)

[7 Trabajo Futuro 3](#_Toc103077929)

[8 Referencias 3](#_Toc103077930)

# Introducción

La cantidad de datos creados en todo el mundo en 2018 alcanzó los 33 ZB (un ZetaByte (ZB) equivale a 1.000 millones de TeraBytes (TB)), 16,5 veces más que solo hace nueve años. No obstante, gracias a los nuevos avances tecnológicos como el internet de las cosas y las redes sociales se estima que la cantidad de información digital generada en 2035 ascenderá a los 2.142 ZB [1]. Por otro lado, gracias a la facilidad con la cual la información es compartida por medio del internet y las redes sociales, ha surgido un gran inconveniente relacionado con la veracidad de dicha información. Una encuesta de alrededor de 25.000 participantes realizada por Ipsos en nombre del Centre *for International Governance Innovation* (CIGI) revelan que alrededor del 86% de los encuestados creen que han estado expuestos a noticias falsas. Las principales fuentes de información falsa son redes sociales como Facebook y YouTube e inclusive la televisión y los sitios web [2].

Del mismo modo, gran parte del contenido disponible se encuentra en formato de imágenes o vídeos, y gracias a la aparición de tecnologías como la Inteligencia Artificial, se han desarrollado nuevas técnicas para crear contenido multimedia falso. Una de las técnicas más recientes desarrolladas es DeepFake, la cual consiste en la creación de videos o imágenes hiperrealistas utilizando inteligencia artificial [3]. El contenido multimedia generado por medio de DeepFake consiste en videos de personas haciendo y diciendo cosas irreales, a tal punto que es difícil identificar si el contenido es real o generado por DeepFake. Un ejemplo del resultado de la alteración de imágenes por medio de DeepFake se puede observar en la figura 1.

A picture containing person, posing

Description automatically generated

Figura 1. (I) Imagen original de Allison Brie. (D) Imagen modificada usando DeepFake para poner la cara de Jim Carrie en lugar de la cara de Allison Brie [24].

Debido a esta problemática se han desarrollado varios modelos de DeepLearning para combatir la desinformación, esto a través de la clasificación de los videos como reales o alterados por DeepFake. Una de las implementaciones ya realizadas fue la del proyecto “Desarrollo de un algoritmo de Machine Learning para la detección de contenido multimedia generado por medio de DeepFake” [4] realizado por Santiago Ballesteros, Daniel Barón, Magda Mejia y Larry Portocarrero en la cual se obtuvieron las siguientes métricas de desempeño para los distintos modelos construidos:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Modelo** | **LOSS** | **ACCURACY** | **PRECISION** | **RECALL** |
| Red Preentrenada VGG16 + MLP | 0,524 | 0,833 | 0,84 | 0,833 |
| Red Preentrenada InceptionV3 + MLP | 0,576 | 0,826 | 0,83 | 0,83 |
| Modelo Baseline | 0,585 | 0,67 | 0,67 | 0,66 |

Frente a este proyecto se ven posibles mejoras tales como el reconocimiento de múltiples rostros en un mismo video para la clasificación de estos, el procesamiento y análisis temporal a través de RNN y la validación étnica. Debido a esto, el proyecto descrito a continuación busca implementar estas mejoras.

# Descripción General

## Objetivos

General

El objetivo de este proyecto es identificar si un video ha sido manipulado a través de técnicas de DeepFake. Esto a través del uso de redes neuronales convolucionales, las cuales permiten la extracción de características o “features”, desde las más sencillas como formas de las caras, gestos, iluminación hasta algunas más definidas como vectores de incorporación de modalidad y emoción para la cara y el habla.

Adicionalmente en caso de que el video contenga varios rostros, se podrá caracterizar cada uno de estos como manipulado o no.

Específicos

* Separar los rostros de los videos usados para entrenamiento y validación.
* Entrenar y usar diferentes redes convolucionales.
* Comparar el rendimiento de diferentes modelos de aprendizaje profundo.
* Ofrecer el servicio de clasificación de un video a través de una página web.

## Antecedentes

El contenido generado a partir de los DeepFakes es cada vez más perjudicial para la privacidad, la seguridad de la sociedad e incluso la democracia [8]. Se han propuesto métodos para detectar DeepFakes desde el inicio de esta amenaza. Los primeros métodos se basaron en características evidentes obtenidas de inconsistencias del proceso de síntesis del video falso. Los métodos más recientes aplican el aprendizaje profundo para extraer características que revelan la presencia de alteraciones en un vídeo [9].

La detección de contenido generado a partir de DeepFake se ha considerado como un problema de clasificación binaria, estableciendo las clases de vídeos auténticos y manipulados. Los modelos basados en una clasificación binaria requieren una gran cantidad de datos en su fase de entrenamiento, estos datos están representados en datasets de videos reales y manipulados. El inicio de estos dataset fue dado por Korshunov y Marcel [10], quienes crearon un conjunto de datos de DeepFake que incluía 620 videos basados en el modelo GAN utilizando la herramienta de faceswap-GAN de código abierto [11]. Este conjunto de datos se alimentó de videos de la base de datos de VidTIMIT disponible de manera pública [12] para generar videos alterados de tipo DeepFake de baja y alta calidad, imitando expresiones faciales, movimientos de la boca y parpadeo de ojos.

Con estos videos disponibles se ponen a prueba varios métodos de detección de DeepFake. Los resultados de estas pruebas muestran que populares sistemas de reconocimiento facial basados en VGG [13] y Facenet [14][15] son incapaces de diferenciar rostros generados por DeepFake de un rostro real.

Así mismo otros métodos, como los de enfoques de sincronización de labios [16][17] y las métricas de calidad de imagen con máquina de vectores de soporte (SVM) [18], producen tasas de error muy altas en la detección de videos DeepFake del conjunto de datos de Korshunov y Marcel.

En intentos más recientes, se encuentra el método basado en la observación de la coherencia temporal de un video de Sabir et al [19]. Este método realiza la comparación de características de las secuencias cuadro por cuadro de un video, identificando manipulaciones y así detectar DeepFakes. Se propuso un modelo convolucional recurrente (RCN) basado en la integración de la red convolucional DenseNet [20] y celdas unitarias recurrentes [21] y así identificar las diferencias temporales entre cuadros. (Figura 2). Este modelo fue probado con un nuevo dataset, FaceForensics++ [22], el cual incluye 1000 videos y se obtiene una mejora comparada con los anteriores modelos.

Graphical user interface, application

Description automatically generated

Figura 2. Método basado en la detección de modificaciones en una secuencia de fotogramas.[22]

# Diseño y Especificaciones

## Definición del Problema

La generación de contenido multimedia falso por medio de DeepFake tiene un alto potencial dañino en el uso y manipulación de la imagen pública de la persona afectada debido a la calidad del contenido creado, que llega al punto de que es casi imposible para un ser humano diferenciar si un video es real o fue creado por medio de DeepFake. Por ejemplo, en 2017 una celebridad se enfrentó a una situación controversial debido a la circulación de un video pornográfico falso en el ciberespacio [4], y hoy en día el uso de esta tecnología está en manos del público por lo cual cualquier persona que sepa usar herramientas como DeepFaceLab [5], está en la capacidad de generar DeepFakes sobre actores, personas famosas, y en general cualquier persona vulnerando sus derechos personales y derechos de propiedad intelectual [6]. Debido al potencial dañino que tiene DeepFake es importante tener una herramienta para detectar contenido multimedia generado por medio de DeepFake.

## Especificaciones

Requerimientos Funcionales

* Carga de un video a una plataforma web.
* Identificación de la cantidad de caras en el video.
* Partición del video en marcos y detección de las caras en los marcos seleccionados.
* Categorización de las caras de acuerdo con la similitud entre los diferentes marcos.
* Clasificación de las caras individualmente a través de su probabilidad en cada marco de ser alterada por técnicas de Deep Fake.

Requerimientos No Funcionales

## Restricciones

# Desarrollo del Diseño

## Fuentes de Información

## Alternativas de Diseño

# Implementación

## Entendimiento de Datos

Descripción y estadísticas de los datos.

## Preparación de Datos

Procesamiento de los videos en marcos y luego caras a través de modelo de similitud

## Modelaje y Evaluación

Redes Convolucionales Pre-Entrenadas

Uso de las redes pre-entrenadas para categorizar.

Evaluación de Modelos

Selección de Modelos

## Despliegue

Diagrama de despliegue del modelo. (Vertex AI, página web con resultados)

# Conclusiones

# Trabajo Futuro

# Referencias

1. Statista. (2019, April 17). Infografía: A la espera de un big bang de datos. Statista Infografías. Retrieved October 7, 2021, from <https://es.statista.com/grafico/17734/cantidad-real-y-prevista-de-datos-generados-en-todo-el-mundo/>.
2. Ipsos. (2019, June 11). Fake News: A Global Epidemic Vast Majority (86%) of Online Global Citizens Have Been Exposed to it. Ipsos. Retrieved October 7, 2021, from <https://www.ipsos.com/en-us/news-polls/cigi-fake-news-global-epidemic>.
3. Westerlund, M. (2019). The emergence of deepfake technology: A review. Technology Innovation Management Review.
4. Ballesteros, S., Baron, D., Mejia, M., Portocarrero, L. (Dic, 2021). Desarrollo de un algoritmo de Machine Learning para la detección de contenido multimedia generado por medio de DeepFake.
5. Masood, M., Nawaz, M., Malik, K. M., Javed, A., & Irtaza, A. (2021). Deepfakes Generation and Detection: State-of-the-art, open challenges, countermeasures, and way forward. arXiv preprint arXiv:2103.00484.
6. Colaner, N. & Quinn, .M.J (2020). Deepfakes and the Value-Neutrality Thesis Retrieved October 7, 2021 from <https://www.seattleu.edu/ethics-and-technology/viewpoints/deepfakes-and-the-value-neutrality-thesis.html>
7. Colak, B. (2021). Legal Issues of DeepFakes. Retrieved October 7, 2021 from <https://www.internetjustsociety.org/legal-issues-of-deepfakes>
8. Chesney, R., and Citron, D. K. (2018). Deep fakes: a looming challenge for privacy, democracy, and national security. <https://dx.doi.org/10.2139/ssrn.3213954>.
9. De Lima, O., Franklin, S., Basu, S., Karwoski, B., and George, A. (2020). Deepfake detection using spatiotemporal convolutional networks. arXiv preprint arXiv:2006.14749.
10. Korshunov, P., and Marcel, S. (2019). Vulnerability assessment and detection of deepfake videos. In The 12th IAPR International Conference on Biometrics (ICB), pp. 1-6.
11. Faceswap-GAN. Available at <https://github.com/shaoanlu/faceswap-GAN>.
12. VidTIMIT database. Available at <http://conradsanderson.id.au/vidtimit/>
13. Parkhi, O. M., Vedaldi, A., and Zisserman, A. (2015, September). Deep face recognition. In Proceedings of the British Machine Vision Conference (BMVC) (pp. 41.1-41.12).
14. FaceNet. Available at <https://github.com/davidsandberg/facenet>.
15. Schroff, F., Kalenichenko, D., and Philbin, J. (2015). Facenet: A unified embedding for face recognition and clustering. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 815-823).
16. Chung, J. S., Senior, A., Vinyals, O., and Zisserman, A. (2017, July). Lip reading sentences in the wild. In 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 3444-3453).
17. Korshunov, P., and Marcel, S. (2018, September). Speaker inconsistency detection in tampered video. In 2018 26th European Signal Processing Conference (EUSIPCO) (pp. 2375- 2379). IEEE.
18. Galbally, J., and Marcel, S. (2014, August). Face anti-spoofing based on general image quality assessment. In 2014 22nd International Conference on Pattern Recognition (pp. 1173- 1178). IEEE.
19. Sabir, E., Cheng, J., Jaiswal, A., AbdAlmageed, W., Masi, I., and Natarajan, P. (2019). Recurrent convolutional strategies for face manipulation detection in videos. In Proceedings of the IEEE Conference on Comp
20. Huang, G., Liu, Z., Van Der Maaten, L., and Weinberger, K. Q. (2017). Densely connected convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 4700-4708).
21. Cho, K., van Merrienboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. (2014, October). Learning phrase representations using RNN encoder–decoder for statistical machine translation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP) (pp. 1724-1734)
22. Hsu, C. C., Lee, C. Y., and Zhuang, Y. X. (2018, December). Learning to detect fake face images in the wild. In 2018 International Symposium on Computer, Consumer and Control (IS3C) (pp. 388-391). IEEE.