Detección de imágenes o vídeos modificados mediante Redes Neuronales

Miguel del Arco Marquez

Resum– Resum del projecte, màxim 10 línies. ....
.. ..
... .
....
....
....
....
....
.. ..
... .

Palabras claves– Aprendizaje profundo, Detección de imágenes modificadas, Conjuntos de datos, Preprocesamiento de imágenes, Detección de manipulación específica.

Abstract– Versió en anglès del resum . ...
....
....
....
....
....
.. ..
....
.. ..
... .

Keywords– Deep learning, Detection of modified images, Data sets, Image preprocessing, Detection of specific manipulation.

1 Introducción - Contexto del trabajo

En la actualidad, el uso de imágenes en la comunicación y en la toma de decisiones es cada vez más común. Sin embargo, existe la posibilidad de que algunas de estas imágenes sean manipuladas o modificadas para engañar o tergiversar la información que se presenta. Por esta razón, la detección de imágenes modificadas se ha convertido en una tarea importante en el campo de la inteligencia artificial.

- E-mail de contacto: migueldemollet10@gmail.com
- Mención realizada: Computación.
- Trabajo tutorizado por: Jordi Serra Ruiz (departamento)
- Curs 2022/23

Para abordar esta problemática, se han desarrollado diversas técnicas basadas en el aprendizaje profundo que permiten detectar si una imagen ha sido modificada o no. El aprendizaje profundo, también conocido como deep learning, es una rama de la inteligencia artificial que utiliza redes neuronales artificiales para aprender y realizar tareas complejas, como el reconocimiento de objetos y el procesamiento de lenguaje natural.

En este trabajo, se abordará la detección de imágenes modificadas mediante el uso de técnicas de aprendizaje profundo. Se discutirán los diferentes modelos de deep learning que se pueden utilizar para este fin, así como los pasos necesarios para llevar a cabo el entrenamiento y la evaluación de los modelos. Asimismo, se destacará la importancia de contar con un conjunto de datos etiquetado y representativo para el entrenamiento de los modelos, así como la necesidad de ajustar y actualizar periódicamente el modelo para

mejorar su rendimiento.

En definitiva, este trabajo tiene como objetivo brindar una visión general de cómo se puede utilizar el aprendizaje profundo para detectar imágenes modificadas y destacar la relevancia de esta tarea en la actualidad.

2 Objetivos

Los objetivos que supone un trabajo de esta magnitud son los siguientes.

1. Identificar y evaluar los modelos de aprendizaje profundo más efectivos para detectar imágenes modificadas en un conjunto de datos específico.
2. Implementar y entrenar un modelo de aprendizaje profundo para la detección de imágenes modificadas, utilizando un conjunto de datos representativo.
3. Analizar la efectividad del modelo en la detección de diferentes tipos de modificaciones, como la manipulación de la información visual, la eliminación de objetos o la inserción de objetos.
4. Comparar el rendimiento del modelo de aprendizaje profundo con otras técnicas de detección de imágenes modificadas, como el análisis forense de imágenes o la detección de patrones.
5. Proporcionar recomendaciones para mejorar la precisión y la eficacia del modelo de aprendizaje profundo en la detección de imágenes modificadas, como la incorporación de nuevos datos de entrenamiento o la adaptación del modelo a nuevos tipos de modificaciones.

3 Metodología

Durante el desarrollo del proyecto, se utilizará GitHub como herramienta de gestión de versiones y almacenamiento del código. Esto permitirá tener un seguimiento detallado del progreso del trabajo y una trazabilidad adecuada de los cambios realizados. Además, se ha establecido un proceso de revisión regular con una frecuencia de 15 días, en el cual se compartirá el avance del trabajo con el tutor designado, Jordi Serra Ruiz, del departamento correspondiente, y se recibirá feedback para asegurar que se está avanzando adecuadamente hacia los objetivos del TFG. Este enfoque garantizará una comunicación fluida entre el estudiante y el tutor, lo que resultará en un trabajo de alta calidad.

4 Estado del arte

El campo de la detección de imágenes modificadas mediante el uso de técnicas de aprendizaje profundo se encuentra en constante evolución y desarrollo. A continuación, se presentan algunos de los avances más relevantes y actuales en esta área:

- **Herramientas de pago**: Existe una herramienta que podemos encontrar en la suit de *Adobe Analytics*[1][2], esta pretende detectar si una imagen ha sido modificado mediante algoritmos de Deep Learning.

  Esta tecnologia permite a los creadores de contenido proteger sus derechos de autor y verifiifcar la autenticidad de la imagen. Por otro lado permite detectar si dicha imagen es original o ha sido modificada.

  Para tener una mejor idea de como esta funciona, podemos observar como en la figura 1 esta es capaz de detectar objetos añadidos a imagenes.

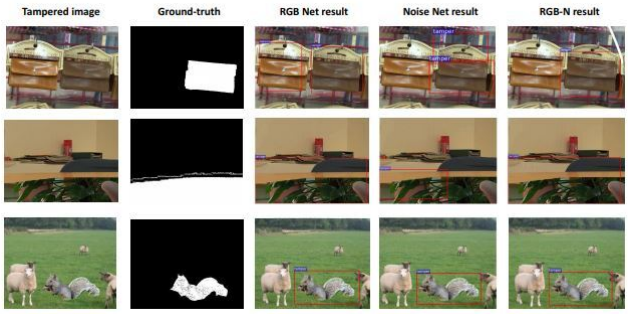

Fig. 1: Ejemplo de funcionamiento de la IA de Adobe

- **Redes Neuronales Convolucionales (CNN)**: Las redes neuronales convolucionales son el modelo de deep learning más utilizado en la detección de imágenes modificadas, y han demostrado tener un alto nivel de precisión en la detección de imagenes modificadas por scripts de *Adobe Photoshop*®.

  Podemos ver trabajos ya realizados en el cual usan este tipo de red neuronal, un caso seria el trabajo de **Detecting Photoshopped Faces by Scripting Photoshop**[3]

Fig. 2: Resultados Detecting Photoshopped Faces by Scripting Photoshop

  Para entender mejor la figura 2 debemos de ver que esta imagen esta compuesta compuesta por 3 subimagenes de izquierda a derecha seria lo siguiente.

  1. Imagen original
  2. Imagen modificada
  3. Output del modelo

- **Uso de tecnologias existentes como Deep Fake**: Es una técnica de inteligencia artificial que se utiliza para crear videos o imágenes manipuladas que parecen ser auténticas, pero en realidad son falsas. Esta técnica utiliza algoritmos de aprendizaje profundo para entrenar modelos de redes neuronales que pueden analizar, sintetizar y manipular el contenido de una imagen o un video. Con esta técnica, es posible crear videos y fotos que parecen ser reales, pero que en realidad son

el resultado de la manipulación de contenido existente, como el rostro de una persona. Los deepfakes se han utilizado en algunos casos para crear noticias falsas, difamar a personas o para fines de entretenimiento. Es importante destacar que los deepfakes pueden ser utilizados de manera engañosa, por lo que es importante ser cautelosos al consumir contenido generado por esta técnica [4].

Para entender mejor esta tecnica podemos ver en la figura 3 el resultado de aplicar esta tecnica.

Fig. 3: Resultado del Deep Fake (Output—Input)

Existen papers como **FaceForensics++ - Learning to Detect Manipulated Facial Images**[5] en el cual trata sobre la preocupación creciente acerca de la capacidad de generar y manipular imágenes sintéticas con un alto grado de realismo, lo que puede tener implicaciones graves en la sociedad, ya que puede conducir a una pérdida de confianza en el contenido digital y difundir información falsa o noticias falsas.

Este propone la detección de manipulaciones faciales y da un benchmark automatizado para la evaluación de la eficacia de los métodos de detección.

El benchmark se basa en representantes destacados de manipulaciones faciales y contiene una base de datos de más de 1.8 millones de imágenes manipuladas. Los autores muestran que el uso de conocimiento específico del dominio mejora significativamente la detección de falsificaciones y supera claramente a los observadores humanos. En resumen, el trabajo aborda un problema importante en el ámbito de la generación y manipulación de imágenes sintéticas y propone una solución útil para la detección de manipulaciones faciales.

En la figura 4 podemos observar el Pipeline del proyecto.

Fig. 4: Pipeline FaceForensics++

El estado del arte en la detección de imágenes modificadas mediante el uso de técnicas de aprendizaje profundo se centra en el desarrollo y mejora de los modelos existentes, la exploración de nuevas técnicas de aprendizaje profundo, el uso de grandes conjuntos de datos de entrenamiento y la investigación en detección de manipulación específica.

5 DATASET

La elección de un conjunto de datos adecuado es una parte crítica en la creación de un modelo. La calidad y la representatividad de los datos de entrenamiento son factores clave en la precisión y la eficacia del modelo.

Este punto del proyecto es bastante delicado, ya que este influira de una forma considerable a nuestro futuro modelo, por lo tanto debemos de tener en cuenta lo siguiente.

1. **El conjunto de datos determina la capacidad de generalización del modelo**: Si el conjunto de datos de entrenamiento es demasiado pequeño o no es representativo de las diferentes posibilidades de modificación de imágenes, el modelo puede tener dificultades para detectar imágenes modificadas en el mundo real. Un conjunto de datos amplio y variado puede mejorar la capacidad de generalización del modelo y permitir que se ajuste mejor a diferentes situaciones.

2. **El conjunto de datos influye en la precisión**: Si el conjunto de datos de entrenamiento contiene imágenes mal etiquetadas o ruidosas, el modelo puede ser menos preciso en la detección de imágenes modificadas. Por lo tanto, es importante elegir un conjunto de datos de alta calidad y con etiquetas precisas.

3. **El conjunto de datos influye en el rendimiento**: Un conjunto de datos grande y variado puede mejorar el rendimiento del modelo, permitiendo una mejor detección de imágenes modificadas y reduciendo el riesgo de sobreajuste.

4. **El conjunto de datos puede influir en el tipo de modelo que se utiliza**: El conjunto de datos puede influir en la elección del tipo de modelo que se utiliza para la detección de imágenes modificadas. Por ejemplo, si el conjunto de datos contiene una gran cantidad de imágenes con modificaciones sutiles, puede ser más apropiado utilizar un modelo basado en GAN o en aprendizaje por transferencia.

5.1. Dataset elegido

Después de analizar varios datasets, he decidido elegir **Casia-2** [6]. Este conjunto de imágenes es idóneo para crear un modelo robusto, ya que contiene distintas categorías tanto originales como modificadas, y en cada una de ellas se aplican diferentes tipos de modificaciones, como recortes y pegados de la misma imagen (**Same**) o de otra imagen (**Different**).

Las categorías incluidas en el dataset son las siguientes:

- Animales.
- Arquitectura.
- Arte.
- Personas.
- Interiores.
- Plantas.
- Textil.

En la figura 5, se muestra un ejemplo de cada una de estas categorías en el conjunto de originales.

Fig. 5: Ejemplos de cada categoría en el conjunto original.

Para tener una idea del conjunto modificado, se puede observar la figura 6.

Fig. 6: Ejemplos de cada categoría en el conjunto modificado.

Sin embargo, este dataset presenta un problema, ya que está algo desequilibrado en cuanto a las categorías, no tenemos la misma cantidad de imágenes para cada categoría. Este problema se puede apreciar en la figura 7.

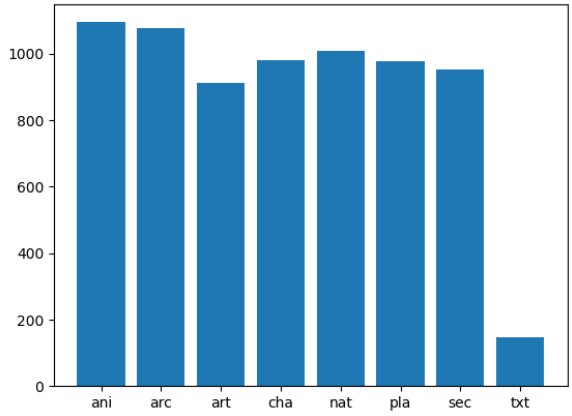

Fig. 7: Cantidad de imágenes por categoría en el conjunto original.

Afortunadamente, esto no representa una gran complicación, ya que se ha decidido recortar el dataset limitando a 600 imágenes por cada categoría. De esta forma, se equilibra el conjunto. Sin embargo, la categoría de **Textil** cuenta con muy pocas imágenes, por lo que se ha optado por eliminarla para simplificar el problema. El resultado se muestra en la figura 8.

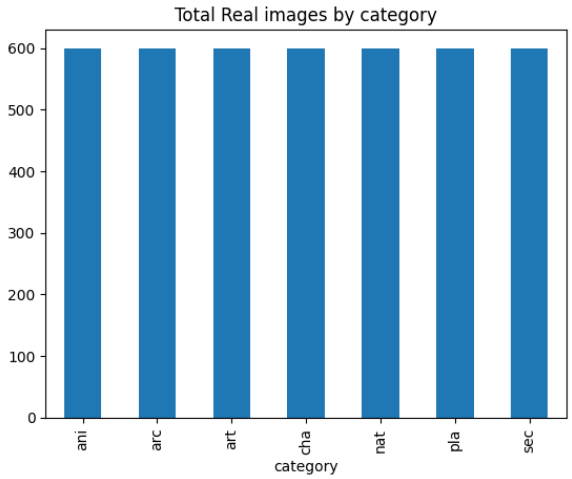

Fig. 8: Cantidad de imágenes por categoría en el conjunto original después de corregirlo.

El mismo problema ocurre en las imágenes de cada categoría en el conjunto modificado, pero en este caso, hay una variable adicional: no todas las categorías tienen la misma cantidad de imágenes con distintas modificaciones. Este desequilibrio se puede apreciar mejor en la figura 9.

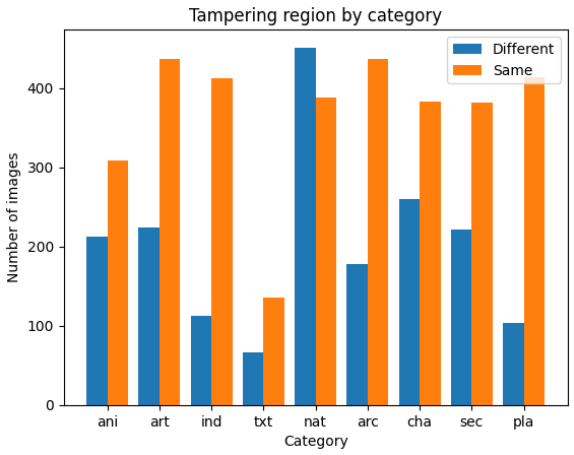

Fig. 9: Cantidad de imagenes por categoria y metodo del conjunto modificado.

Aplicando la misma logica que en el set de las originales, recortamos a 600 imagens por categoria pero ademas de estas 600 la mitad seran de una modificacion u otra.

En el cual tenemos el resultado que muestra la figura 10

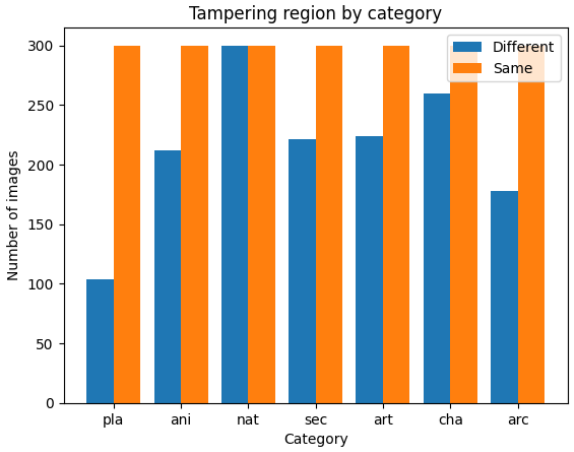

Fig. 10: Cantidad de imagenes por categoria y metodo del conjunto modificado tras corregirlo.

El resultado no es el ideal, pero es el mejor que se puede obtener con estos datos.

5.2. Alternativas

Lamentablemente, encontrar conjuntos de datos que cumplan con los requisitos necesarios para desarrollar un modelo robusto ha sido un desafío. La mayoría de los conjuntos de datos disponibles eran de pago o demasiado específicos para nuestra investigación. Por lo tanto, se optó por utilizar el conjunto de datos **Casia-2** [6], que cumplió con nuestras necesidades.

Aunque existe una alternativa al conjunto de datos seleccionado, que es la versión anterior, **Casia-1** [7], se decidió no utilizarla. Esto se debe a que los propios creadores del conjunto de datos indican que la calidad del contenido de esta versión es inferior a la de la versión 2.

REFERENCIAS

[1] Página web de Adoble Analytics. https://business.adobe.com/uk/products/analytics/adobe-analytics.html

[2] Noticia relevante del medio de Xataka sobre el software que usa adobe para la detección de imágenes modificadas mediante Photoshop. https://www.xataka.com/robotica-e-ia/adobe-creador-photoshop-esta-desarrollando-software-para-detectar-imagenes-manipuladas-photoshop

[3] Papper de detección de imágenes modificadas mediante Photoshop. https://paperswithcode.com/paper/detecting-photoshopped-faces-by-scripting

[4] Papper explicativo de como funciona la técnica de Deep Fake. https://arxiv.org/pdf/1909.11573.pdf

[5] Papper donde proponen una resolución a las imágenes modificadas apoyándose en DeepFakes, Face2Face, FaceSwap y NeuralTextures. https://arxiv.org/abs/1901.08971

[6] URL del repositorio de Github donde se ha obtenido el Dataset. https://github.com/namtpham/casia2groundtruth

[7] URL del repositorio de una alternativa al Dataset escogido. https://github.com/namtpham/casia1groundtruth

APÈNDIX

A.1. Secció d'Apèndix

... .

A.2. Secció d'Apèndix

... .