Assessing advanced computer vision and NLP for waste image classification

Juan Manuel Camara Diaz

Resumen– El reciclaje es un proceso importante para proteger el medio ambiente y preservar los recursos naturales al convertir materiales usados en nuevos productos para reducir la cantidad de residuos que se depositan en vertederos. La automatización del reciclaje, tanto a nivel doméstico como profesional, puede mejorar la eficiencia y precisión del proceso de separación de residuos. En este artículo, se exploran los nuevos avances en el campo de la visión por computación para proporcionar herramientas para la clasificación de residuos. Se discuten diversas cuestiones, como el uso de técnicas de generación de imágenes con stable diffusion, modelos de 0-shot learning para etiquetar conjuntos de datos de manera más eficiente, la posibilidad de utilizar el texto asociado a una imagen para mejorar la precisión de la clasificación.

Palabras clave– Reciclaje, automatización, visión por computación, redes neuronales, generación de imagen, lenguaje natural, 0-shot learning, clasificación, dataset, CLIP, stable diffusion.

Abstract– Recycling is an important process for protecting the environment and preserving natural resources by converting used materials into new products to reduce the amount of waste deposited in landfills. The automation of recycling, both at the household and professional level, can improve the efficiency and accuracy of the waste separation process. In this article, new advances in the field of computer vision are explored to provide tools for waste classification. Various issues are discussed, such as the use of stable diffusion image generation techniques, 0-shot learning models for more efficient data labeling, and the possibility of using text associated with an image to improve classification accuracy.

Keywords– Recycling, automation, computer vision, neural networks, image generation, natural language, 0-shot learning, classification, dataset, CLIP, stable diffusion.

1 INTRODUCCIÓN

El reciclaje es el proceso de convertir materiales usados en nuevos productos para reducir la cantidad de residuos que van a parar a los vertederos y minimizar el impacto ambiental. El reciclaje es fundamental para proteger el medio ambiente y preservar los recursos naturales.

En primer lugar, el reciclaje ayuda a reducir la cantidad de residuos que van a parar a los vertederos, lo que a su vez reduce la cantidad de espacio que se necesita para depositar-los. Los vertederos ocupan un espacio valioso y pueden ser una fuente de contaminación del aire y del agua si no se gestionan adecuadamente. Además, los residuos que se depositan en los vertederos pueden tardar décadas o incluso siglos en descomponerse, lo que aumenta aún más el problema.

En segundo lugar, el reciclaje ayuda a conservar los recursos naturales. Muchos de los productos que utilizamos diariamente, como el papel, el vidrio y el plástico, están hechos de recursos naturales limitados, como los árboles, el petróleo y el gas natural. Al reciclar estos materiales, se reduce la necesidad de extraer nuevos recursos naturales, lo que a su vez reduce la huella de carbono y ayuda a preservar los recursos para las generaciones futuras.

Por otro lado, la automatización del reciclaje, tanto a nivel doméstico como profesional, puede mejorar significativamente la eficiencia y la precisión del proceso de separación de residuos. En el caso de la automatización doméstica, existen herramientas y dispositivos que pueden ayudar a clasificar los productos y saber en qué contenedor deben ir. Por ejemplo, contenedores con sensores de luz y peso que indican al usuario si un objeto debe ir al contenedor de papel, vidrio o plástico.

En cuanto a la automatización profesional, las cintas transportadoras y otras herramientas mecánicas pueden separar los residuos en diferentes categorías de manera más rápida

• E-mail de contacto: juanma.caaz@gmail.com
• Mención realizada: Ingeniería de Computación
• Trabajo tutorizado por: Coen Antens (CVC)
• Curso 2022/2023

y precisa que si se hiciera de forma manual. Además, la automatización puede reducir la cantidad de mano de obra necesaria para separar los residuos, lo que a su vez reduce los costos y mejora la eficiencia del proceso.
Por estos motivos me he enfocado en explorar los nuevos avances que han ido saliendo en el campo de la visión por computación para aportar herramientas para la clasificación de residuos. En los últimos años, se han logrado importantes avances en el campo de la visión por computador gracias a los modelos de redes neuronales que se pueden utilizar para resolver problemas de clasificación relacionado con el reciclaje. [1]

2 Motivación

En 2020 se aprobó el Pacto Verde Europeo [2], también conocido como green deal y que tiene como objetivo lograr la neutralidad climática en Europa para el año 2050. Este plan ambicioso busca transformar la economía europea y abordar los desafíos del cambio climático y la pérdida de biodiversidad. Incluye medidas para mejorar la eficiencia energética, reducir las emisiones de gases de efecto invernadero y fomentar la transición hacia una economía circular. Además, el Pacto Verde Europeo se enfoca en la creación de empleos y en el apoyo a una transición justa para garantizar que todos los ciudadanos europeos se beneficien de una economía más sostenible. Por lo tanto creo que realizar una clasificación de estos residuos puede ayudar a cumplir con este objetivo.

La clasificación de residuos es un problema importante en la gestión de residuos y en la preservación del medio ambiente. Sin embargo, la eficiencia de los modelos de clasificación existentes todavía puede mejorarse. En este trabajo de investigación, se abordarán diversas cuestiones relacionadas con la clasificación de residuos utilizando técnicas de inteligencia artificial.

Una de las inquietudes que se quiere abordar es si se puede utilizar el uso de técnicas de generación de imagen con stable diffusion[3] para generar imágenes a partir de una descripción usando el lenguaje natural para aumentar un conjunto de datos ya existentes. Además, se pretende investigar la posibilidad de etiquetar de manera más eficiente los datasets utilizando modelos de 0-shot learning.

Por último, se reflexionará sobre si los modelos de clasificación tradicionales seguirán siendo la mejor opción en el futuro y se investigará la viabilidad de utilizar stable diffusion para generar datasets sintéticos con el conjunto de datos inicial.

En base a estas cuestiones, se plantearán tres objetivos específicos para abordar los problemas planteados:

1. Generar un gran dataset para cubrir la mayor cantidad de residuos reciclables y no reciclables utilizando modelos 0-shot learning y analizar su desempeño con modelos de clasificación clásicos.

2. Verificar si el texto en las imágenes del dataset es relevante en el desempeño de la clasificación y crear un modelo multi-modal para permitir el uso de texto e imágenes.

3. Investigar la viabilidad de utilizar stable diffusion para generar datasets sintéticos con el conjunto de datos inicial.

Con estos objetivos se busca mejorar la clasificación de residuos y contribuir a la gestión sostenible de los mismos.

3 Estado del Arte

En los últimos años, se han logrado importantes avances en el campo del Deep Learning, esto gracias a la aparición de los Transformers. Presentados en el paper de Attention is all you need [4].

Los transformers son clave en modelos de inteligencia artificial porque han demostrado ser muy efectivos en tareas de procesamiento del lenguaje natural (NLP, por sus siglas en inglés) y otras aplicaciones relacionadas con el aprendizaje automático, incluyo extrapolando el su uso a todo tipo de tareas.

3.1. Clasificación

La clasificación ha dado un cambio radical con la salida del CLIP [8]. CLIP es un modelo de aprendizaje profundo que ha cambiado el estado del arte en la clasificación de imágenes. Utiliza un modelo de lenguaje para entender tanto las imágenes como el texto relacionado y una técnica de aprendizaje por similitud para clasificar imágenes en función de su similitud con una determinada descripción textual.

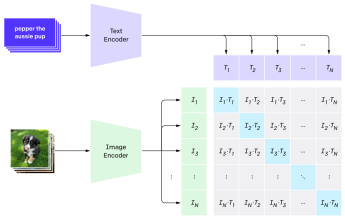


Fig. 1: Arquitetura de CLIP, por OpenAI.

En comparación con los modelos clásicos de redes neuronales convolucionales, CLIP puede comprender el contexto y el significado detrás de una imagen, lo que lo hace especialmente útil en situaciones en las que las imágenes pueden ser difíciles de clasificar para los modelos clásicos.

Model	Accuracy	Params	Tecnology
BASIC-L	91.1 %	2440M	Conv+Transf
CoCa	91.0 %	2100M	Transformer
Model soups	90.98 %	2440M	Conv+Transf
Model soups	90.94 %	1843M	Transformer
ViT-e	90.9 %	3900M	Transformer
CoAtNet-7	90.88 %	2440M	Conv+Transf
ViT-G/14	90.71 %	1843M	Transformer
CoCa	90.60 %	2100M	Transformer
CoAtNet-6	90.45 %	1470M	Conv+Transf
ViT-G/14	90.45 %	1843M	Transformer
DaViT-G	90.4 %	1437M	Transformer
Meta Pseudo Labels	90.2 %	480M	EfficientNet
DaViT-H	90.2 %	362M	Transformer
SwinV2-G	90.17 %	3000M	Transformer
Florence-CoSwin-H	90.05 %	893M	Transformer
Meta Pseudo Labels	90 %	390M	EfficientNet
RevCol-H	90.0 %	2158M	Pure CNN

Fig. 2: Comparativa de modelos de clasificación con mejor accuracy en imágenes del dataset Imagenet tabla extraída y adaptada de ”paperswithcode”. Modelos del 2020 al 2023-Q1.

En la tabla se puede ver una comparativa de los modelos de clasificación de imágenes más destacados del 2020 al 2023-Q1 y como los modelos basados en transformers (entrenados con el contexto) son superiores a los son puramente convolucionales. Esto se debe a que los modelos de redes neuronales convolucionales puros no pueden comprender el contexto y el significado detrás de una imagen, mientras que los modelos de redes neuronales basados en Transformer pueden comprender el contexto y el significado detrás de una imagen consiguiendo mejores resultados.

También es interesante ver que modelos con una cantidad de parámetros muy inferior a los modelos de redes neuronales convolucionales puros, como el modelo de ViT-G/14, pueden conseguir resultados mejores que los modelos de redes neuronales convolucionales puros.

Respecto a la clasificación del lenguaje natural, también se ha visto que los transformers son las arquitecturas que mejor funcionan dejando atrás a los modelos LSTM.

3.2. Generadores

Los generadores de texto han sido una de las aplicaciones mas destacadas de los transformers. En el paper de GPT-2 [7] se presento un modelo de lenguaje que puede generar texto de manera autoregresiva.

3.3. Modelos multimodales

También ha habido una tendencia a la inteligencia general, es decir modelos que pueden realizar múltiples tareas, se ha visto que los modelos de transformers son capaces de realizar múltiples tareas, como por ejemplo el modelo de GPT-3 [6] que puede realizar tareas de clasificación de texto, generación de texto, etc. Y como otros modelos que han sido entrenados con diferentes tipos de input pueden lograr resultados sorprendentes, como por ejemplo el modelo GATO [5] por deepmind que puede realizar tareas de clasificación de texto, clasificación de imágenes, jugar a videojuegos de Atari y contestar preguntas similar a un chatbot.

En definitiva, se esta viendo que los modelos multimodales basados en transformers están logrando resultados sorprendentes en diferentes tareas. Aun así, los modelos mas tradicionales como las redes neuronales convolucionales, recurrentes y LSTM siguen siendo muy competentes en sus respectivas tareas.

4 METODOLOGÍA

Para la realización del proyecto se han determinado unas metodologías de trabajo, desde a nivel de planificación hasta a nivel de desarrollo.

4.1. Planificación

Para la planificación he decidido realizar tareas que sean ejecutadas secuencialmente para así poder llevar un mayor control del avance del proyecto. Cada fase esta compuesta de diferentes tareas a realizar. Se puede ver en el Apèndix la planificación detallada.

La metodología a utilizar en el proyecto sera Kanban, cada tarea se representara en una tarjeta Kanban y se moverá a la columna correspondiente en función de su estado. Los estados que se han definido son:

TODO: tarea pendiente de realizar.
IN PROGRESS: tarea en proceso.
DONE: tarea realizada.

Para la implementación de la metodología Kanban se ha utilizado la herramienta Trello, que permite crear tableros y gestionar las tareas de manera sencilla. En el Apèndix se puede ver el tablero Kanban.

Para todo el código desarrollado se utilizara el sistema de control de versiones Git para llevar un control de las versiones del proyecto.

4.2. Tecnologías

Para llevar a cabo el proyecto, se han seleccionado una serie de tecnologías que permitirán realizar de manera eficiente el procesamiento de imágenes y el análisis de datos, así como el despliegue de modelos de aprendizaje automático.

- OpenCV: una biblioteca de visión por computadora de código abierto que proporciona herramientas para el procesamiento de imágenes y el análisis de vídeo en tiempo real. Esta tecnología se eligió debido a su capacidad para proporcionar una amplia gama de herramientas para el procesamiento de imágenes y su eficiente despliegue de modelos de aprendizaje profundo (DNN).
- Diversas bibliotecas de Python, incluyendo SKlearn, numpy, matplotlib y pandas, para el tratamiento de datos y visualizaciones. Estas bibliotecas ofrecen un conjunto de herramientas y funcionalidades para la manipulación de datos y su representación gráfica.
- OpenCLIP: una alternativa open-source de CLIP ofrecida por LAION, que permite la clasificación de imagenes dado una lista de las posibles clases. Esta tecnología se eligió debido a su capacidad para proporcionar una alternativa open-source de CLIP.
- PaddleOCR [9]: se utilizará para la detección y reconocimiento del texto en las imágenes. PaddleOCR es una biblioteca de aprendizaje profundo fácil de usar que se basa en la plataforma PaddlePaddle y proporciona herramientas para el procesamiento de imágenes y la extracción de texto.
- Tensorflow2 [10]: para el entrenamiento y creación de modelos multimodales. Tensorflow2 es una biblioteca de aprendizaje automático de código abierto que se utiliza ampliamente en la creación de modelos de aprendizaje profundo y su despliegue en producción.
- Stable Diffusion: para realizar la data augmentation en el entrenamiento de los modelos. Esta técnica permite la creación de nuevas imágenes a partir de las imágenes existentes para mejorar el rendimiento del modelo.
- Colab [11]: como entorno de ejecución durante la mayor parte del proyecto. La versión pro de Colab permite un mayor tiempo de ejecución y recursos computacionales,

lo que es necesario para el procesamiento de imágenes y la creación de modelos de aprendizaje profundo.

- Android Studio: para el desarrollo de la aplicación móvil. Esta plataforma de desarrollo proporciona herramientas para la creación de aplicaciones para dispositivos móviles y su despliegue en dispositivos Android.

4.3. Documentacion

La documentacion se realizara en el lenguaje de marcado LaTeX, ya que es un lenguaje de marcado que permite crear documentos de gran calidad y que es muy utilizado en la comunidad cientifica.

Parte de la documentacion se encontrara en el repositorio del proyecto en GitHub y todo el dataset generado sera publicado y de libre acceso en el repositorio.

5 Dataset

Unos de los principales problemas para la generación del dataset es que en cada región tiene un sistema de reciclaje diferente y por lo tanto las clases de residuos que se pueden reciclar también son diferentes por lo que ha sido necesario crear un dataset propio adaptado a Cataluña. El motivo de utilizar el sistema de Cataluña es porque llevan tiempo fomentando la separación de residuos y hay mucha información disponible sobre el tema.

Residuonvas.cat es una página web que tiene como objetivo informar a la población sobre la gestión de residuos en Cataluña. En su sitio web, promueven un sistema de clasificación y separación de residuos en diferentes categorías para su posterior tratamiento y reciclaje. A continuación, se presenta una lista de las diferentes clases de residuos para reciclar que se promueven en este sistema:

Envase de vidrio	Envase ligero
Medicamentos	Pilas y baterías
Aparatos eléctricos	Ropa y calzado
Punto verde	Orgánica
	Resta

Esta lista serán las clases que se utilizaran para la creación del dataset.

Referencias

[1] Por que es importante reciclar https://ecoembesdudasreciclaje.es/por-que-es-importante-reciclar/

[2] Green Deal https://commission.europa.eu/strategyandpolicy/priorities20192024/europeangreendeal_es

[3] Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2021). High-Resolution Image Synthesis with Latent Diffusion Models. doi:10.48550/ARXIV.2112.10752

[4] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., … Polosukhin, I. (2017). Attention Is All You Need. doi:10.48550/ARXIV.1706.03762

[5] GATO (A genelist Agent) https://openreview.net/pdf?id=1ikK0kHjvj

[6] Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., … Amodei, D. (2020). Language Models are Few-Shot Learners. doi:10.48550/ARXIV.2005.14165

[7] Language Models are Unsupervised Multitask Learners https://d4mucfpksywv.cloudfront.net/better-language-models/language-models.pdf

[8] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., … Sutskever, I. (2021). Learning Transferable Visual Models From Natural Language Supervision. doi:10.48550/ARXIV.2103.00020

[9] PaddleOCR https://github.com/PaddlePaddle/PaddleOCR

[10] Tensorflow https://www.tensorflow.org/

[11] Google Colab https://colab.research.google.com/

Apèndix

A.1. Planificación

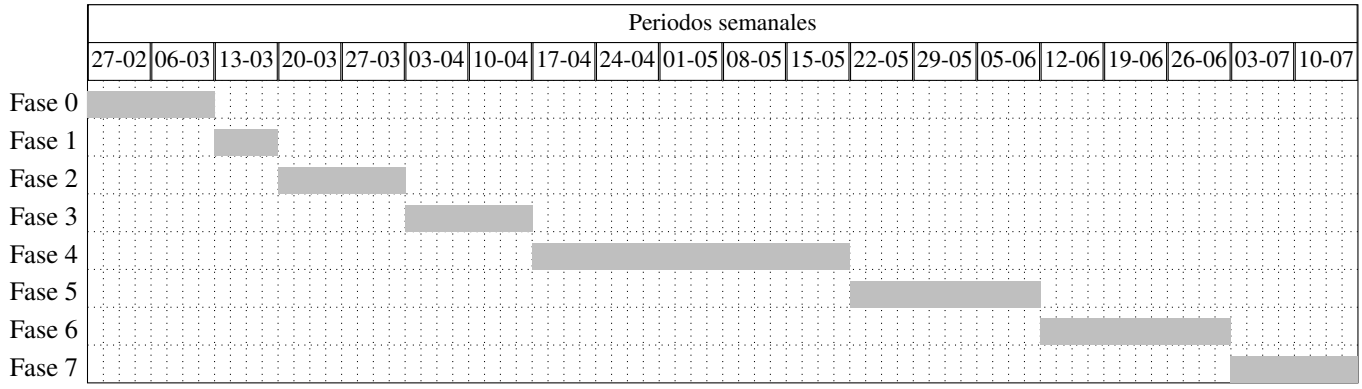
Para la planificación he decidido realizar tareas que sean ejecutadas secuencialmente para así poder llevar un mayor control del avance del proyecto. Casa fase esta compuesta de diferentes tareas a realizar.

- **Fase 0 – Planificación del proyecto (2 semanas)**
  - Pensar el enfoque del proyecto.
  - Creación de la petición de TFG.
  - Validación y adaptación del TFG.
- **Fase 1 – Creación del dataset (1 semana)**
  - Pensar qué clases se usarán para clasificar.
  - Buscar/Crear imágenes de residuos.
  - Utilizar OpenCLIP para clasificar aquellas sin etiquetas.
  - Terminar de etiquetar a mano aquellas imágenes mal etiquetadas.
  - Ver el desempeño de OpenCLIP con el dataset corregido.
- **Fase 2 – Comparativa OpenCLIP vs RNN (2 semanas)**
  - Entrenar un modelo RNN solo con información visual.
  - Realizar comparativas con OpenCLIP.
- **Fase 3 – Extracción del texto en las imágenes (2 semanas)**
  - Utilizar un OCR para extraer el posible texto de las imágenes.
  - Entrenar uno o varios modelos de clasificación de texto.
  - Evaluar resultados.

- **Fase 4 – Creación del modelo multimodal (4 semanas)**
  - Preparar los datos para el entrenamiento del modelo.
  - Crear un modelo que permita texto e imágenes como input.
  - Evaluar resultados.
- **Fase 5 – Realizar data-augmentation con Stable-diffusion (3 semanas)**
  - Instalar Stable-diffusion.
  - Probar la viabilidad para realizar el data aumentation.
  - Realizar pruebas con few show learning.
  - Evaluar resultados.
- **Fase 6 – Implementación del modelo final en una aplicación. (3 semanas)**
  - Seleccionar el mejor modelo para por ejecutarse en una arquitectura móvil.
  - Creación de una sencilla aplicación que haga uso del modelo óptimo.
- **Fase 7 – Finalización del proyecto (3 semanas)**
  - Finalizar informe.
  - Creación del póster.
  - Creación de la presentación.

A.2. Distribución temporal

Como se ha explicado comentado anteriormente, las fases se realizarán de manera secuencial. Hasta no terminar una fase no se podrá empezar la siguiente.



A.3. Trello

Para llevar un control de las tareas a realizar he decidido utilizar Trello.

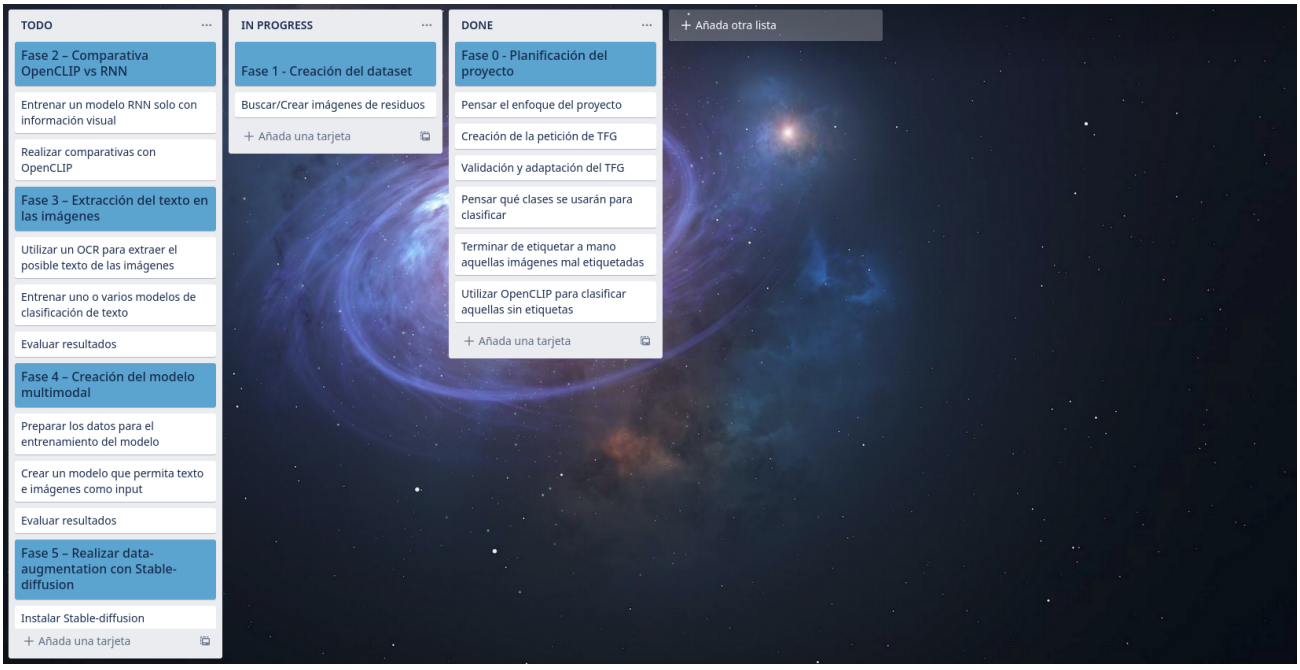


Fig. 3: Tablero de Trello