*Texto

El contenido generado por IA puede ser incorrecto.*

**Asignatura**

Sistemas Interactivos Inteligentes

**Práctica 2. Unidad IV**

Pre-CLIP y CLIP

**Alumna:** Carlota Fernández del Riego

*Texto

El contenido generado por IA puede ser incorrecto.*Índice

1. Resumen……………………………………………………………………………………………………….3
2. Dataset propio……………….….………………………………………………………………………….3
3. Decisiones de diseño e implementación………………………………..……………………..3/4
   1. Modelo Pre-CLIP…………………………………………….…………………………………3
   2. Modelo CLIP………………………………….…………….……………………………………4
4. Resultados obtenidos……………………………………………………………….…………………..4
5. Discusión………………………………………………………………………………….…………………..5
6. Limitaciones……………….…………………………………………………………….…………………..5
7. Conclusiones…………………………………………………………………………………….…………..5
8. *Texto

   El contenido generado por IA puede ser incorrecto.***Resumen**

Esta práctica consistió en el desarrollo de un análisis comparativo entre dos enfoques de representación multimodal:

* **Pre-CLIP**, un modelo que combina redes de visión y lenguaje entrenadas de manera independiente.
* **CLIP**, un modelo multimodal entrenado conjuntamente para alinear texto e imagen en un mismo espacio vectorial.

El trabajo se basó en crear un dataset propio de 20 imágenes reales distribuidas en cuatro categorías, obtener los embeddings de texto e imagen con ambos métodos y comparar los resultados de similitud mediante métricas de coseno.

El objetivo principal fue comprender cómo los modelos multimodales logran capturar relaciones semánticas entre modalidades diferentes, mostrando la evolución desde arquitecturas separadas hasta modelos entrenados de forma conjunta.

1. **Dataset propio**

El dataset desarrollado para esta práctica, titulado **“Urban\_Scenes”**, reúne diferentes tipos de imágenes obtenidas de **Unsplash**, una plataforma en línea que ofrece fotografías reales y libres de derechos. De este modo, se garantiza que el contenido sea auténtico y no generado por inteligencia artificial.

La colección se organiza en cuatro categorías principales, cada una asociada a un tipo de entorno urbano distinto:

|  |  |
| --- | --- |
| **Categoría** | **Descripción general** |
| City Architecture | Rascacielos, fachadas y estructuras urbanas modernas. |
| Urban Mobility | Escenas con tráfico, bicicletas y transporte público. |
| Street Life | Personas en espacios urbanos, mercados o terrazas. |
| Industrial Areas | Fábricas, almacenes y zonas industriales. |

Cada imagen va acompañada de una breve descripción en inglés, redactada manualmente para asegurar la coherencia semántica entre texto e imagen. Toda esta información se recoge en el archivo **dataset\_UrbanScenes.csv**, que contiene tres columnas: *image*, *description* y *category*.  
Las imágenes se encuentran organizadas en subcarpetas por categoría, facilitando así su lectura e indexación automática por los scripts del proyecto.

1. **Decisiones de diseño e implementación**
   1. **Modelo Pre-CLIP**

Para la primera parte del trabajo se implementó un enfoque **Pre-CLIP**, combinando dos redes especializadas:

* **ResNet50**, preentrenada en *ImageNet*, con la capa de clasificación eliminada para acceder a las características visuales profundas.
* *Texto

  El contenido generado por IA puede ser incorrecto.***SentenceTransformer - distiluse-base-multilingual-cased**, un modelo ligero y multilingüe que genera embeddings semánticos de texto.

Estos modelos se eligieron por su equilibrio entre precisión y eficiencia, así como por su integración directa en librerías estándar como *torchvision* y *sentence-transformers*.  
Las representaciones vectoriales obtenidas de ambas modalidades se compararon mediante similitud coseno, generando una matriz de correlación entre imágenes y descripciones.

* 1. **Modelo CLIP**

En la segunda parte se utilizó el modelo **openai/clip-vit-base-patch32** de *Hugging Face*, que permite generar embeddings conjuntos de texto e imagen. El preprocesamiento de los datos se realizó con CLIPProcessor, mientras que CLIPModel se empleó para extraer y normalizar los vectores de características.

La principal ventaja de CLIP es que ambos tipos de datos comparten el mismo espacio semántico, lo que permite medir directamente la afinidad entre una imagen y su descripción, sin necesidad de transformaciones adicionales.

|  |  |  |
| --- | --- | --- |
| **Modelo** | **Similitud media (pares correctos)** | **Similitud media global** |
| Pre-CLIP | 0.0159 | 0.0066 |
| CLIP | 0.2813 | 0.1950 |

1. **Resultados obtenidos**

Imagen que contiene Gráfico

El contenido generado por IA puede ser incorrecto.Imagen que contiene Gráfico

El contenido generado por IA puede ser incorrecto.

El incremento en la similitud media destaca una **mejora en el rendimiento al emplear CLIP**, lo que demuestra la capacidad para establecer relaciones semánticas consistentes entre texto e imagen. Por otra parte, en el modelo **Pre-CLIP**las correlaciones son escasas y los valores de similitud se mantienen próximos al nivel de ruido, es decir, con el modelo **CLIP**se consigue una **alineación mucho más precisa**entre los pares correctos.

Esta diferencia se aprecia en la **matriz de similitud**, donde CLIP genera una **diagonal intensamente brillante**, indicando la relación directa entre cada imagen y su descripción.  
*Texto

El contenido generado por IA puede ser incorrecto.*De hecho, ese patrón muestra que el modelo ha aprendido a asociar de forma eficaz conceptos textuales con sus representaciones visuales, incluso en un conjunto de datos reducido y sin realizar un entrenamiento adicional.

En resumen, CLIP demuestra una **mayor comprensión contextual y generalización semántica**, que confirma la ventaja de los modelos multimodales entrenados de manera conjunta frente a los enfoques tradicionales basados en redes separadas.

1. **Discusión**

Los resultados confirman que los modelos multimodales preentrenados como CLIP alcanzan una **alineación semántica mucho más precisa** entre texto e imagen. Esto se debe a su **entrenamiento** sobre grandes volúmenes de pares imagen-texto, lo que le permite conseguir relaciones conceptuales de alto nivel, con palabras como “tráfico”, “edificio” o “fábrica”, incluso en diferentes contextos.

Por otra parte, el enfoque Pre-CLIP, al tratar las modalidades de manera independiente, genera vectores difíciles de comparar. Esta comparación muestra la importancia del **entrenamiento conjunto** en tareas de comprensión multimodal.

1. **Limitaciones**

* El tamaño reducido del dataset, 20 imágenes, limita la capacidad de generalización.
* No se ha realizado *fine-tuning* en ninguno de los modelos, lo que podría mejorar los resultados en dominios específicos.
* La métrica de similitud de coseno no siempre capta matices semánticos complejos.

1. **Conclusiones**

En conclusión, los resultados obtenidos muestran que **CLIP alcanza una correlación de imagen y de texto claramente superior** a la del modelo Pre-CLIP, dejando claro el potencial de los modelos entrenados de forma multimodal con el objetivo de comprender la relación entre ambas modalidades.