

Tarea 3 - Recuperación de Imágenes Usando Codificadores Visuales

Diego Alejandro Valdés Fernández

Departamento de Ingeniería Informática

Universidad de Santiago de Chile, Santiago, Chile

diego.valdes.f@usach.cl

I. RESUMEN

Se evalúa el desempeño de tres codificadores visuales del estado del arte (DINOv2, CLIP y ResNet18) en la tarea de recuperación de imágenes sobre tres conjuntos de datos: SimpleIK, Paris y VOC-Pascal. El proceso consiste en representar las imágenes en un espacio latente mediante cada codificador, calcular la similitud coseno entre imágenes y recuperar las más similares a una consulta dada. Se utiliza evaluación leave-one-out y se reportan métricas como precisión promedio media (mAP), curvas precisión-recall (RP) y ejemplos visuales de las cinco mejores y peores recuperaciones. El objetivo es comparar el rendimiento y comportamiento de los codificadores en distintos dominios visuales.

II. INTRODUCCIÓN

La recuperación de imágenes es una tarea fundamental en visión por computador que busca identificar, dentro de un conjunto de datos, las imágenes más similares a una imagen dada como consulta. Este proceso se basa en representar cada imagen como un vector en un espacio de características donde se puede medir su similitud mediante una métrica, como la similitud coseno.

En esta tarea se implementaron y compararon tres codificadores visuales del estado del arte: DINOv2, CLIP y ResNet18. Estos modelos permiten extraer representaciones vectoriales de imágenes que capturan información semántica relevante, sin necesidad de entrenar un modelo clasificador sobre el conjunto de datos objetivo.

Los codificadores se aplicaron a tres conjuntos de datos visuales de distinta naturaleza: *SimpleIK* (categorías visuales simples), *VOC-Pascal* (objetos naturales y artificiales) y *Paris* (escenas urbanas). Para cada uno, se siguió un protocolo de evaluación leave-one-out, calculando la similitud coseno entre imágenes, la métrica de *mean Average Precision* (mAP), curvas de precisión-recall (RP), y ejemplos visuales de las mejores y peores recuperaciones por codificador.

El propósito principal de este trabajo es evaluar la calidad de las representaciones generadas por distintos codificadores visuales y su impacto en tareas de recuperación de imágenes, destacando su desempeño comparativo en distintos dominios visuales.

III. DESARROLLO

El desarrollo de esta tarea se estructuró en una serie de etapas claramente definidas, desde la preparación de los datos hasta la evaluación de los resultados. El proceso completo se ilustra en la Figura 1, que representa un flujo general de codificación, recuperación y análisis.



Figura 1: Diagrama de bloques del proceso de codificación, recuperación y evaluación.

Se trabajó con tres conjuntos de datos visuales: *SimpleIK*, *VOC-Pascal* y *Paris*. Cada conjunto incluye imágenes etiquetadas por clase, lo cual permitió definir la relevancia entre imágenes para el proceso de recuperación.

Primero se preprocesaron las imágenes para que fueran compatibles con cada codificador. En el caso de ResNet18 y DINOv2, se utilizaron transformaciones estándar basadas en recorte central, redimensionamiento y normalización. Para CLIP, se aplicó la transformación provista por la propia librería oficial.

Luego, se extrajeron los vectores de características (representaciones latentes) utilizando tres codificadores visuales:

- **ResNet18:** Red convolucional clásica entrenada en ImageNet, con la capa final de clasificación removida para obtener solo la codificación visual.

- **CLIP (ViT-B/32):** Modelo que aprende alineaciones imagen-texto. Se usó únicamente el codificador visual basado en Vision Transformer.
- **DINOv2 (ViT-S/14):** Modelo auto-supervisado que produce representaciones robustas sin necesidad de anotaciones manuales.

Cada imagen se utilizó como consulta y se calcularon las similitudes coseno con todas las demás, implementando así la evaluación *leave-one-out*. A partir de esto se generaron rankings de recuperación para cada consulta.

Posteriormente se calcularon las métricas principales:

- **mAP (mean Average Precision):** Promedio del promedio de precisión por imagen, útil para evaluar la calidad general de recuperación.
- **Curvas Precision-Recall (RP):** Permiten visualizar la relación entre cobertura y precisión en la recuperación.
- **Ejemplos visuales:** Se mostraron los 5 mejores y 5 peores resultados de recuperación para cada codificador, facilitando una evaluación cualitativa.

Todo el proceso fue automatizado mediante una función `run_pipeline(...)` que ejecuta la codificación, calcula similitudes, evalúa mAP y RP, y muestra visualmente los resultados destacados para cada combinación de dataset y codificador.

IV. EXPERIMENTOS REALIZADOS

Se utilizaron tres codificadores visuales (DINOv2, CLIP y ResNet18) sobre tres conjuntos de datos: Simple1K, Paris y VOC_Pascal. Cada imagen fue usada como consulta, y se calcularon las similitudes coseno con el resto del conjunto. A partir de esto se generaron rankings de recuperación y se evaluó el desempeño usando métricas cuantitativas y cualitativas.

Resultados para Simple1K

IV-0a. DINOv2

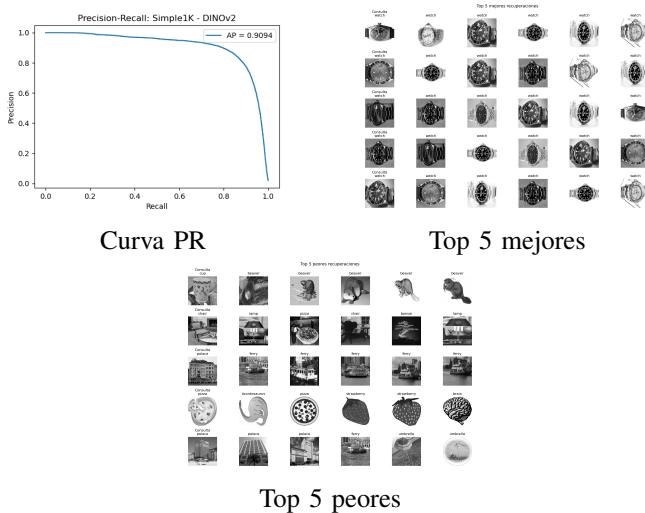


Figura 2: Resultados con DINOv2 en Simple1K.

IV-0b. CLIP

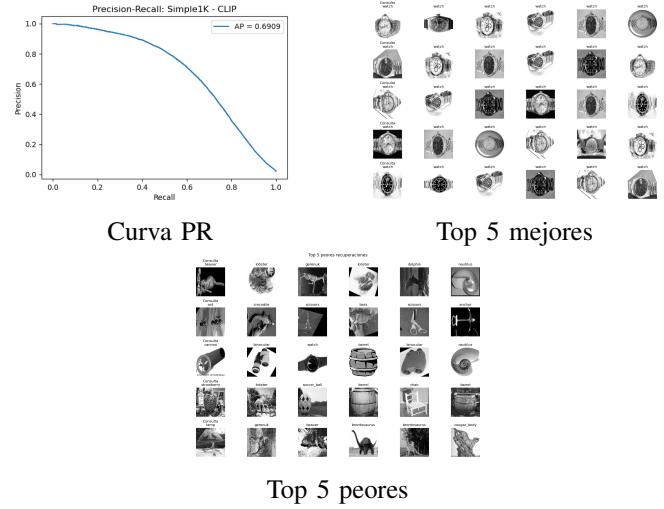


Figura 3: Resultados con CLIP en Simple1K.

IV-0c. ResNet18

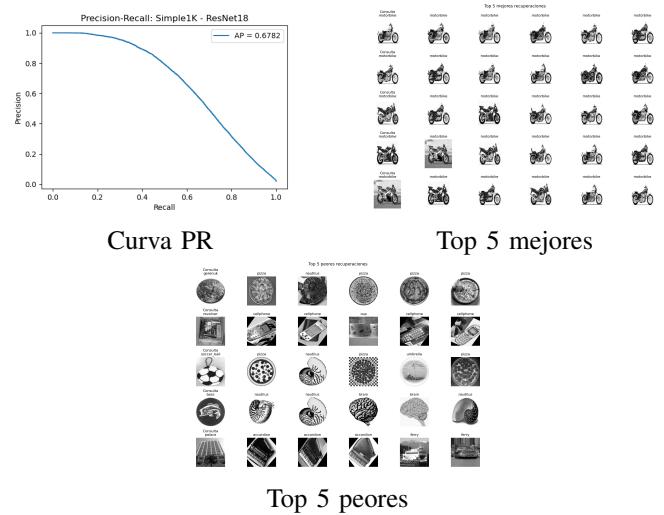


Figura 4: Resultados con ResNet18 en Simple1K.

Resultados para Paris

IV-0d. DINOv2

Resultados para Paris

IV-0d. DINOv2

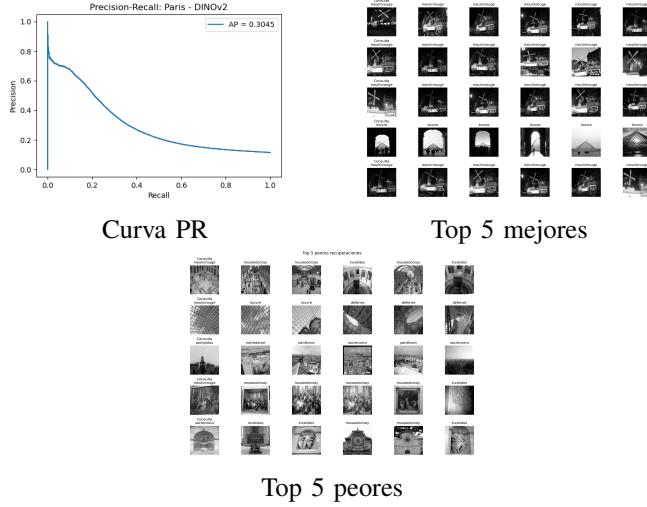


Figura 5: Resultados con DINOv2 en Paris.

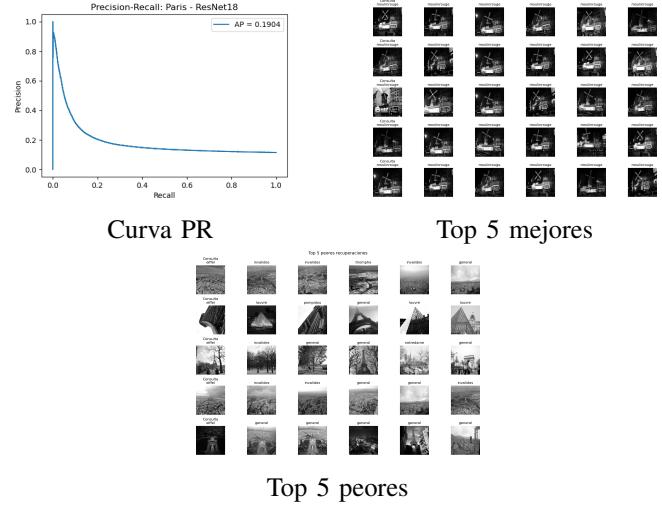


Figura 7: Resultados con ResNet18 en Paris.

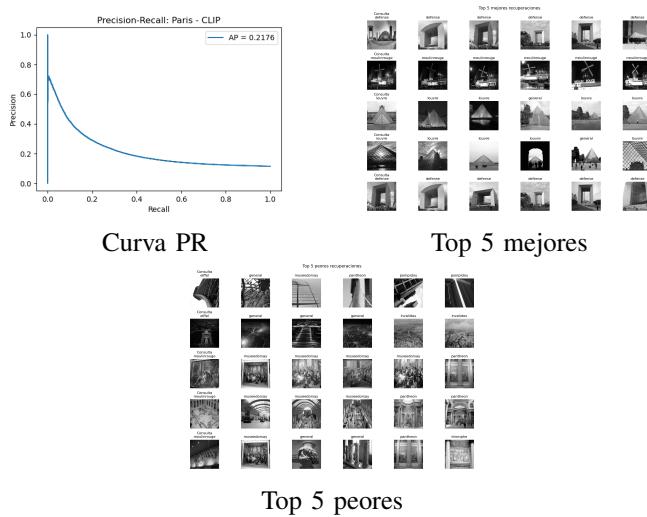
IV-0e. CLIP

Figura 6: Resultados con CLIP en Paris.

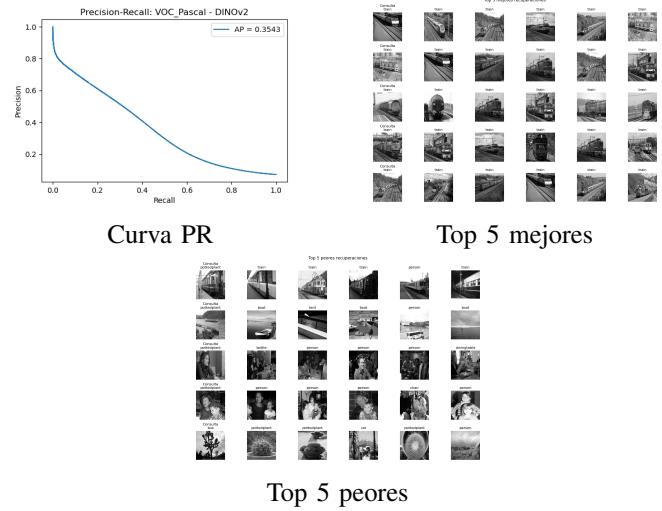
*IV-0f. ResNet18**Resultados para VOC_Pascal**IV-0g. DINOv2*

Figura 8: Resultados con DINOv2 en VOC_Pascal.

IV-0h. CLIP

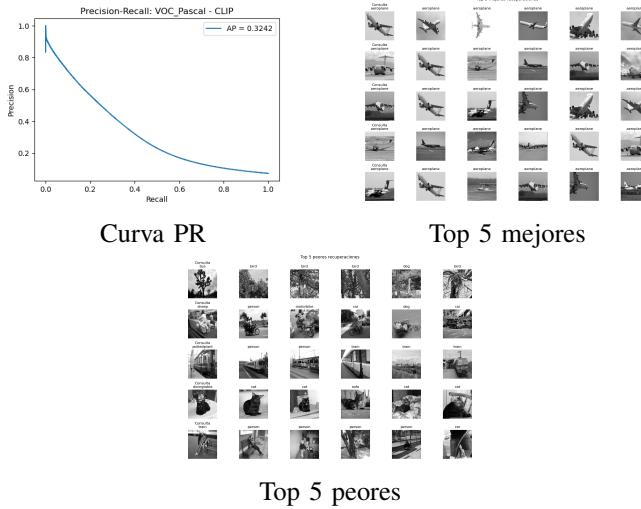


Figura 9: Resultados con CLIP en VOC_Pascal.

IV-Oi. ResNet18

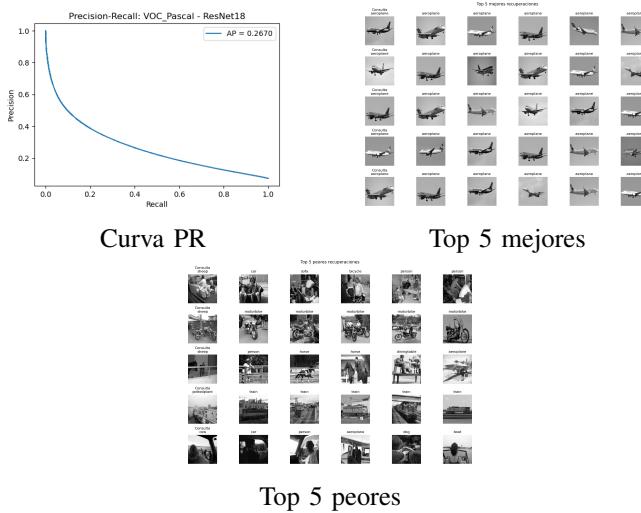


Figura 10: Resultados con ResNet18 en VOC_Pascal.

Tabla I: mAP obtenido por cada codificador en los distintos datasets

Codificador	Simple1K	Paris	VOC_Pascal
DINOv2	0.9254	0.3934	0.4644
CLIP	0.7692	0.3068	0.4346
ResNet18	0.6977	0.2489	0.3522

Resumen de las curvas Precision-Recall (PR)

Las curvas PR permiten visualizar el compromiso entre precisión y recall para distintas configuraciones de umbral. A partir de estas curvas se extrajo el valor de *mAP* (mean Average Precision), mostrado previamente en la Tabla I.

Los resultados muestran que:

- **DINOv2** presenta la mejor separación en todos los conjuntos de datos, alcanzando un **mAP de 0.9254** en

Simple1K, lo que indica un excelente desempeño. Aunque su rendimiento disminuye en conjuntos más complejos como VOC_Pascal, sigue manteniendo el mayor *mAP* entre los codificadores.

- **CLIP** obtiene resultados intermedios, con *mAP* más bajos que DINOv2 pero relativamente consistentes. Su comportamiento sugiere buena generalización, especialmente en VOC_Pascal (0.4346).
- **ResNet18** es el codificador con menor rendimiento en los tres datasets, lo que concuerda con su menor capacidad comparado con modelos preentrenados más recientes como CLIP y DINOv2.

En general, las curvas PR muestran una caída más pronunciada para los conjuntos *Paris* y *VOC_Pascal*, lo que indica una mayor dificultad del sistema para mantener alta precisión a medida que aumenta el recall en estos casos.

V. CONCLUSIONES

En esta tarea se evaluó la capacidad de tres codificadores visuales —**DINOv2**, **CLIP** y **ResNet18**— para la recuperación de imágenes en tres conjuntos de datos: Simple1K, Paris y VOC_Pascal. Se utilizó la similitud coseno como métrica de comparación y se evaluó el desempeño a través de *mAP* y curvas Precision-Recall.

- **DINOv2** demostró ser el codificador más efectivo, obteniendo los mayores valores de *mAP* en todos los datasets. Esto confirma su superioridad como extractor de características visuales autorregresivo.
- **CLIP** mostró un rendimiento intermedio, siendo más robusto que ResNet18 y manteniendo buena generalización en distintos tipos de imágenes, aunque por debajo de DINOv2.
- **ResNet18**, aunque más eficiente computacionalmente, quedó rezagado en términos de *mAP* y calidad de recuperación, mostrando limitaciones frente a codificadores más modernos.
- Los resultados confirmaron que modelos con entrenamiento contrastivo (como CLIP y DINOv2) superan ampliamente a arquitecturas tradicionales como ResNet18 para tareas de recuperación visual.

Como trabajo futuro, sería interesante evaluar nuevos codificadores más recientes (como SigLIP o EVA), incorporar texto en la consulta, o aplicar técnicas de indexación para acelerar la recuperación en catálogos más grandes.