Artificial Intelligence
Task 2: Evaluation of Visual Encoders for Image Retrieval

Prof: José M. Saavedra & Tomás de La Sotta

April, 2025

1. Objective

To evaluate the quality of existing visual encoders on the image retrieval task.

2. Description

Image retrieval is a computer vision task that aims to retrieve the most similar images from a catalog concerning an image query. To this end, all the images are represented in a feature space (a.k.a. latent space), inferred by an encoder. Here, a similarity measure is computed between the catalog images and the query. The result is a ranking of all the images in the catalog ordered from the most to the least similar images with respect to the input query. We can use a distance function like Euclidian (ordered inverse to the similarity) or the cosine function to compute similarity.

Therefore, in this task, you should evaluate three different SOTA (state-of-the-art) encoders for retrieving images from three diverse datasets. Further details are provided below.

2.1. Visual Encoders

In this task, you should evaluate the following visual encoders :

- ResNet_18 (Imagenet): ResNet is the most representative convolutional-based architecture. ResNet proposed the use of residual connections between convolutional layers. Although modern architectures are based on attention, the residual connections are still critical components of current models. Here, ResNet18 is the smaller architecture in the ResNet family that produces vectors in $\mathbb{R}^{512}$. For more details, see the published paper `https://arxiv.org/abs/1512.03385`.
- ResNet_34 (Imagenet): this follows ResNet1s8, having now 34 layers. ResNet_34 produces vectors in $\mathbb{R}^{512}$.
- DINOv2: this encoder consists of a ViT that divides the image into a set of patches. It then computes an embedding for each patch together with a class token that absorbs global information. This encoder was trained on a dataset of 142 million images by a self-supervised strategy. In this experiment we use the smaller version which produces vectors in $\mathbb{R}^{384}$. For more details, see the published paper `https://arxiv.org/abs/2304.07193`.
- CLIP (the visual encoder): CLIP is a bimodal model aligning text and image representations. This is one of the most popular models working with images and text in a shared space. In this task, you should use the visual encoder of CLIP. In this experiment we use ViT-B/32 version that produces vectors in $\mathbb{R}^{512}$. For more details, see the published paper `https://arxiv.org/abs/2103.00020`.

2.1.1. Sample Code

In this task, we have a sample code to show how you should load the models and infer the feature vector for a given image. The code is available at `https://github.com/jmsaavedrar/encoders_pub`.

Other links of interest are:

- `https://github.com/facebookresearch/dinov2`
- `https://github.com/openai/CLIP?tab=readme-ov-file#usage`

2.2. Datasets

- Simple1K: this is a very simple dataset with 1326 images distributed among 50 classes. You can download the dataset from `https://www.dropbox.com/scl/fi/72huigwj9y6gbqci2s7j5/simple1K.zip?rlkey=tfqzribyiuktgo8zdxdqaz63t&st=3qtuv0q4&dl=0`.
- VOC-Pascal: this is a more challenging dataset, very famous around 2010. You will use the validation set containing 5823 images distributed among 20 categories. You can download the dataset from `https://www.dropbox.com/scl/fi/owu68slp2ckn56nsaxjag/VOC_val.zip?rlkey=gb8tb1pi524myeatoyj1y94te&st=w6kt7p1v&dl=0`.
- Paris: this is a traditional dataset proposed to evaluate image retrievavl. In this case we will use a reduced version of the original dataset containing 1274 images distributed among 12 classes. You can download the dataset from `https://www.dropbox.com/scl/fi/si8lp4oarcfzuitgectqa/Paris_val.zip?rlkey=7yf6hfe0u1q7yiy0avcbgyx7c&st=maofk10a&dl=0`.

For each dataset, you will find a file named *list_of_image.txt* containing the file name of each image together with the corresponding class. You should use the class to determine the relevance of an image with respect to a query. Two images from the same class are relevants each other.

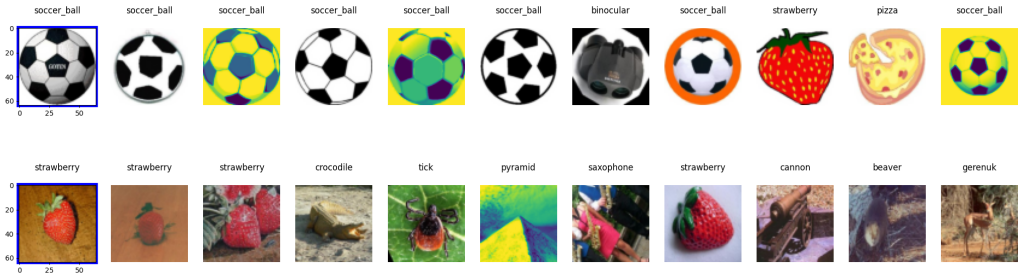

Figura 1: Sample results of image retrieval using ResNet34 as encoder. The first image or of each row is the query.

3. Evaluation Protocol

For evaluation, you should consider each image from the dataset as a query and search for the rest of the images. This strategy is called leave-one-out evaluation. Furthermore, you must use *cosine similarity* to compare vectors. Two examples of image retrieval using ResNet34 as encoder appear in Figure 1.

For each dataset you should present the following metrics:

1. Mean Average Precision (mAP) for the three evaluated encoders.
2. A Recall-Precision Graphic showing the performance of the three encoders.

In addition, for each model and dataset you should report :

- Five examples of the best retrieval results.
- Five examples of the worst retrieval results.

4. Report

In this task, you must write the report in ENGLISH containing the following sections:

1. **Abstract**: this should summarize the work, describing the goal and providing a brief description of how it was achieved. Finally, you should include the main results.

2. **Introduction**: here, you should describe the problem involved in the task, together with a general description of the tools, materials, and models used. 10 %

3. **Development/Methodology**: this is one of the two critical parts of the report. Here, you should precisely describe how you solve the task. In terms of computation, you should describe the programs you implemented to solve the task. Please avoid screenshots without a proper explanation. (40 %)

4. **Experimental Results and Discussion**: here, you should present all the requested results accompanied by the corresponding metrics and figures. Please present a discussion about the performance of the evaluated models. (40 %)

5. **Conclusions**: describe the main insights of your works. (10 %)

5. Submission

Please, submitt your solution by CANVAS until May 05th, 2025 at 23.59hrs. Be sure to send:

1. Source Code

2. Report