Report ML competion by MLOps

Camilla Bonomo
matr. 255138
camilla.bonomo@studenti.unitn.it

Paolo Fabbri
matr. xxxxxx
secondauthor@i2.org

Davide ...
matr. xxxxxx
.....@i2.org

1. Introduction

1.1. Task Description

In this project, we address the problem of image-based retrieval in a closed-set scenario, where the objective is to return the most visually and semantically similar images from a gallery set, given a query image. The metric of interest is top-k retrieval accuracy, which quantifies the proportion of times a relevant image appears among the k nearest neighbors retrieved for a given query. The dataset provided follows a structured format, separating the data into **training**, **test/query**, and **test/gallery** folders. The training data is used to fine-tune models on class-labeled images, while retrieval is performed on test queries against the gallery without using explicit class labels at inference time.

1.2. Overview of Approaches

To tackle this problem, we design a multi-stage pipeline that integrates: (i) model selection via benchmarking; (ii) supervised fine-tuning of selected architectures; (iii) embedding extraction and L2 normalization; (iv) similarity-based retrieval using cosine similarity through FAISS indexing.

We evaluate multiple backbone models pre-trained on ImageNet, including:

ResNet-50: a convolutional residual network;

EfficientNet-B0: a lightweight and efficient CNN architecture;

Vision Transformer (ViT-B/16): a transformer-based image encoder.

In addition to individual models, we implement **ensemble strategies** by averaging embeddings from multiple backbones to leverage complementary feature representations. The pipeline includes both classification-based training (to adapt the models to the dataset) and embedding-based retrieval.

1.3. Summary of Results

Our evaluation includes both single-model and ensemble methods across a consistent validation split. Using the top-10 accuracy as the key metric, we found that the X model outperformed all single models, achieving a notable improvement in retrieval performance. Fine-tuning also led to significant gains over using frozen features. Quantitative results are supported by qualitative visual inspection of retrieved samples. .... (da fare post competition)

2. Models Considered

2.1. Litterature References

ResNet-50 [He et al., 2016]: Introduces deep residual learning via identity skip connections, enabling the training of very deep convolutional networks.

EfficientNet-B0 [Tan and Le, 2019]: Scales depth, width, and resolution uniformly using a compound coefficient, providing an optimal balance between accuracy and efficiency.

Vision Transformer (ViT-B/16) [Dosovitskiy et al., 2020]: Applies transformer architectures directly to image patches, challenging CNN dominance in visual recognition tasks.

Feature Aggregation and Embedding Fusion [Zhou et al., 2021]: Justifies ensemble averaging of embeddings from heterogeneous backbones to improve generalization in retrieval.

2.2. Theoretical Model Descriptions

ResNet-50 ResNet-50 is a convolutional neural network comprising 50 layers with residual connections that mitigate vanishing gradient problems. In our pipeline, we remove the final classification layer and use the feature maps from the penultimate layer, which are then projected into a shared 512-dimensional space and L2-normalized for retrieval.

EfficientNet-B0 EfficientNet-B0 applies neural architecture search and compound scaling to achieve superior performance with fewer parameters. The extracted feature embeddings (1280-dim) from the features block are pooled and projected to 512-dim vectors using a projection head, followed by normalization.

Vision Transformer (ViT-B/16) The ViT-B/16 model splits input images into fixed-size patches and processes them via standard transformer encoders. After discarding

the classification head, we use the output of the [CLS] token for image representation, further projected and normalized. Its ability to capture global context complements the local focus of CNNs.

Each model was fine-tuned only at the classification head level to prevent overfitting and reduce training time. The training loop optimizes cross-entropy loss using Adam, with early stopping and model checkpointing based on validation loss. After fine-tuning, the backbone is frozen, and the feature extractor is reused for retrieval.

Ensemble Strategies We construct two ensembles: (i) ResNet-50 + ViT-B/16 and (ii) EfficientNet-B0 + ViT-B/16. In both cases, normalized embeddings from the respective models are averaged before similarity computation. This leverages the orthogonal strengths of each backbone, combining spatial locality from CNNs and global context from transformers.