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 recent years, top- image retrieval has emerged as a crucial task in computer vision, with applications in visual search engines, recommendation systems, and digital archiving. The goal of this project is to develop and evaluate an efficient pipeline that retrieves the most visually similar images from a gallery, given a query image. The evaluation metric of interest is top- accuracy, which reflects the proportion of correct class matches among the k nearest retrieved images. 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

The pipeline follows a modular architecture combining classification-based fine-tuning and embedding-based retrieval. We evaluate several deep neural networks—ResNet50, EfficientNet-B0, and Vision Transformer (ViT-B/16)—as backbone architectures for feature extraction. These models are initially pre-trained on ImageNet and subsequently fine-tuned on a classification task using Cross-Entropy Loss, allowing for the adaptation of the feature space to the domain-specific dataset. Following fine-tuning, we strip the classification heads and extract L2-normalized feature embeddings for both query and gallery images. These embeddings are indexed using FAISS, a similarity search library optimized for large-scale datasets. The final retrieval is performed using cosine similarity, and both single-model and ensemble-based strategies are considered. The best model configuration is automatically selected through a benchmarking routine based on validation accuracy and inference efficiency.

1.3. Summary of Results

Our experimental evaluation, summarized in Table **??**, reveals that ensemble models achieve the highest top-k accuracy, significantly outperforming single backbone baselines. The performance trade-offs between accuracy and inference time are further discussed in Section 3.

Table 1. Top-3 accuracy and inference time for different models.

Model	**Top-3 Accuracy**	**Avg. Time/Image (s)**	**Typ**
ResNet50	0.811	0.0064	Sing
EfficientNet-B0	0.794	0.0052	Sing
ViT-B/16	0.807	0.0098	Sing
ResNet50 + ViT	0.835	N/A	Ensem
EfficientNet + ViT	0.828	N/A	Ensem

2. Models Considered

2.1. Litterature References

The models evaluated in this project represent three state-of-the-art families of image encoders:

- **Convolutional Neural Networks (CNNs)**: These models utilize convolutional layers to capture local patterns and hierarchical features, making them effective for image classification and retrieval tasks.
- **Vision Transformers (ViTs)**: These models leverage self-attention mechanisms to process images as sequences of patches, enabling them to capture long-range dependencies and global context.
- **Hybrid Models**: These architectures combine the strengths of CNNs and transformers, allowing for both local feature extraction and global context modeling.

The following specific models were selected for evaluation based on their performance in image classification and retrieval tasks, as well as their architectural diversity:

- **ResNet50** (He et al., 2016): A deep residual network featuring skip connections that mitigate the vanishing gradient problem, achieving robust performance on classification and feature embedding tasks.
- **EfficientNet-B0** (Tan and Le, 2019): A compound-scaled architecture designed to balance depth, width, and resolution while maintaining computational efficiency.

- **ViT-B/16** (Dosovitskiy et al., 2020): A transformer-based architecture that applies self-attention mechanisms to image patches, enabling global feature extraction with minimal inductive bias.

2.2. Theoretical Model Descriptions

Each model architecture was leveraged as follows:

ResNet50 We removed the final fully-connected (fc) layer and applied a projection head composed of a linear layer followed by L2 normalization. This design facilitates cosine similarity-based retrieval in the projected feature space.

EfficientNet-B0 Similar to ResNet, the classifier head is discarded, and a custom projection head maps the 1280-dimensional features to a 512-dimensional space.

ViT-B/16 The classification head is replaced with an identity mapping, and the output embeddings are normalized to ensure uniformity in cosine similarity.

All models are wrapped with a custom *ProjectionHead* module to unify the output embedding dimensionality to 512, enabling consistent comparisons and ensemble averaging. Ensemble models average the normalized embeddings from two distinct backbones before similarity computation, leveraging the complementary nature of CNN-based and transformer-based features.