# Fine-Tuning Pretrained Models on Non-ImageNet Datasets for Visual Recognition Tasks

Darius Dabert
MVA Master
darius.dabert@polytechnique.edu

## Abstract

*This assignment involves participating in a Kaggle competition alongside classmates, using the ImageNet-Sketch dataset. The goal is to develop a model that achieves the highest accuracy on a test dataset containing the same categories as the training set.*

## 1. Introduction and Objective

In this work, we investigate the fine-tuning of four state-of-the-art models—Clip, Dinov2, Beit, and Deit—on a custom visual recognition task. The objective of this study is to determine the best-performing model and provide insights into the fine-tuning process on ImageNet data.

## 2. Methodology

We fine-tuned four pretrained models, originally trained on large datasets other than ImageNet, for visual recognition tasks on a custom dataset with diverse categories. The dataset was preprocessed by reshuffling the training and validation sets to create complementary model configurations, ensuring better generalization.

Initially, we trained a classifier with a frozen backbone, allowing the new layers to adapt to the task. We then fine-tuned the entire model, unfreezing the network to further refine the pretrained weights on our dataset. Hyperparameters such as learning rates, batch sizes, and regularization techniques were tuned to optimize performance.

To enhance accuracy, we employed majority voting as an ensemble technique, combining predictions from multiple models. Finally, the models were evaluated on the Kaggle Public dataset to assess their generalization to real-world scenarios.

## 3. Results and Discussion

The following table summarizes the performance of each model on the Kaggle Public dataset. The results are presented in terms of accuracy and loss for both the training and validation sets. Additionally, we provide an analysis of the models' strengths and weaknesses based on the observed trends in the learning curves and final performance.

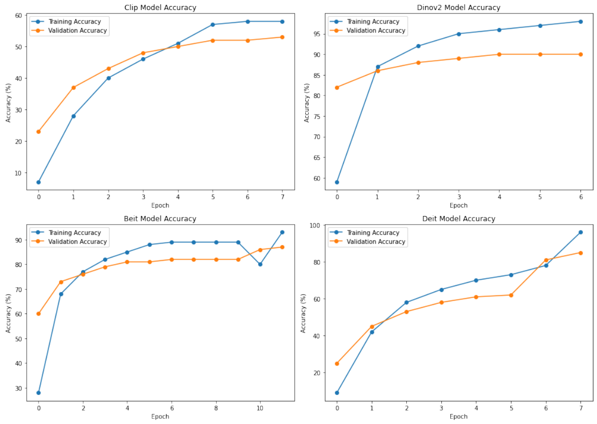

Figure 1. Learning Curves for Clip, Dinov2, Beit, and Deit Models

| **Model** | **Training (%)** | **Validation (%)** | **Test (%)** |
|---|---|---|---|
| Dinov2 | 98% | 90% | 91.11% |
| Deit | 96% | 85% | 85 % |
| Mixture | - | - | 91.15 % |

Table 1. Performance of models on the Kaggle Public dataset

The Dinov2 model outperformed the others, achieving high validation accuracy and low loss on both the custom dataset and the Kaggle Public dataset. Clip, however, struggled with subpar performance, particularly in terms of generalization to the Kaggle dataset.

Training the models in two stages—first with a frozen backbone and then fine-tuning the entire model—helped improve performance. Initially freezing the backbone allows the newly added layers to adapt to the task, which stabilizes learning early on. When the entire model was fine-tuned, it led to a slight improvement in accuracy for models like Beit and Deit, as the pretrained weights were further refined for the specific dataset. This two-step process helps balance the benefits of pretrained features while fine-tuning the model for optimal performance on the task at hand.

While fine-tuning was beneficial, model architecture significantly influenced accuracy and generalization. The use of majority voting with these models helped improve results, but further training and more models would have been needed to fully leverage this approach, requiring additional computational resources.