UNIVERSITY OF SCIENCE AND TECHNOLOGY OF HANOI

INFORMATION COMMUNICATION TECHNOLOGY DEPARTMENT

DATA SCIENCE

Introduction to Deep Learning

Midterm Report

Study the DINOv2 for medical image classification

Nguyễn Mạnh Hưng - **22BI13183**

Nguyễn Trọng Minh - **22BI13304**

Lê Thuận Ninh - **22BI13354**

Trần Lương Hoàng Anh - **22BI13039**

Nguyễn Ngọc Nhi - **22BI13351**

Vũ Hoàng Mai Nhi - **22BI13352**

Hanoi, Vietnam, 02 October 2024

TABLE OF CONTENTS

Abstract

The classification of medical images is always a critical task for providing fast information for diagnosis and treatment planning. In this report, we study DINO version 2, a Vision Transformer (ViT)-based model that leverages a robust self-supervised pre-training method, for the classification of brain tumor MRI images. We use DINOv2 to extract meaningful features from MRI scans without requiring labeled data. Preliminary results demonstrate that DinoV2 achieves competitive performance, offering a robust solution for automating the detection of brain tumors. This approach helps enhance the efficiency of quick brain tumor diagnosis.

I. Introduction

1. DINO

The DINO model is a self-supervised learning method designed for Vision Transformers and convolutional neural networks.

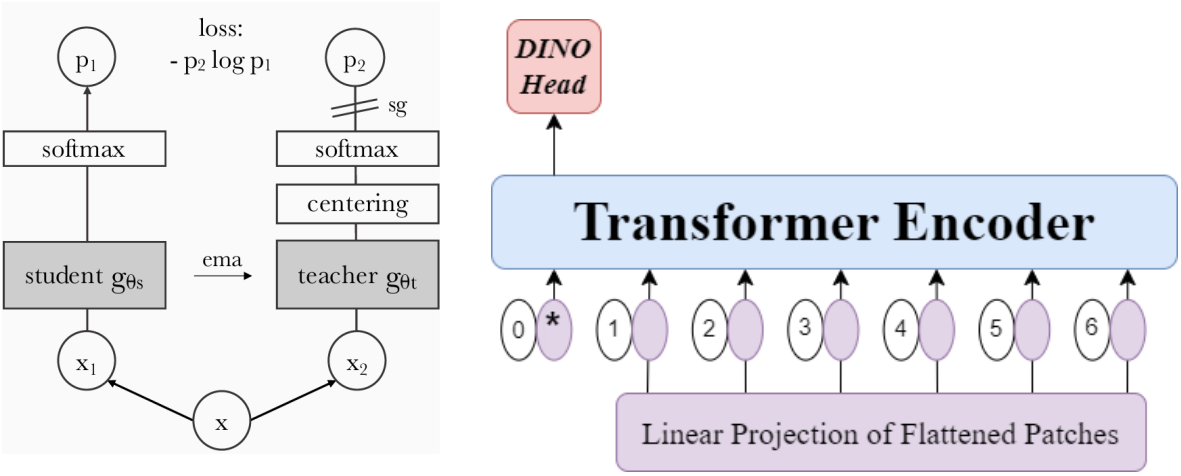


Figure 1: DINO self-supervisor and ViT wrapper

DINO shares similarities with **knowledge distillation**, training a *student* network $g_{\theta s}$ to match the output of a given *teacher* network $g_{\theta t}$. Given an input image, both networks produce a K-dimensional probability P that is normalized using a **softmax** function.

$$P_s(x)^{(i)} = \frac{exp(g_{\theta_s}(x)^{(i)}/\tau_s)}{\sum_{k=1}^{K} exp(g_{\theta_s}(x)^{(k)}/\tau_s)}$$

The objective is to **minimize** the Cross-Entropy Loss between the outputs of *student* and *teacher* so that the model can find out what's kind of the same without labels.

$$L_{DINO} = - \Sigma p_t log\, p_s$$

DINO employs a **multi-crop augmentation strategy**, where the input image is transformed into a set of different views containing *global* views at resolution 224^2 covering a large area and *local* views at resolution 96^2 covering small areas. The *teacher* network only processes *global* views, while the *student* processes both *global* views and *local* views. The goal is to encourage the *student* network to learn from both *local* and *global* information, improving feature robustness.

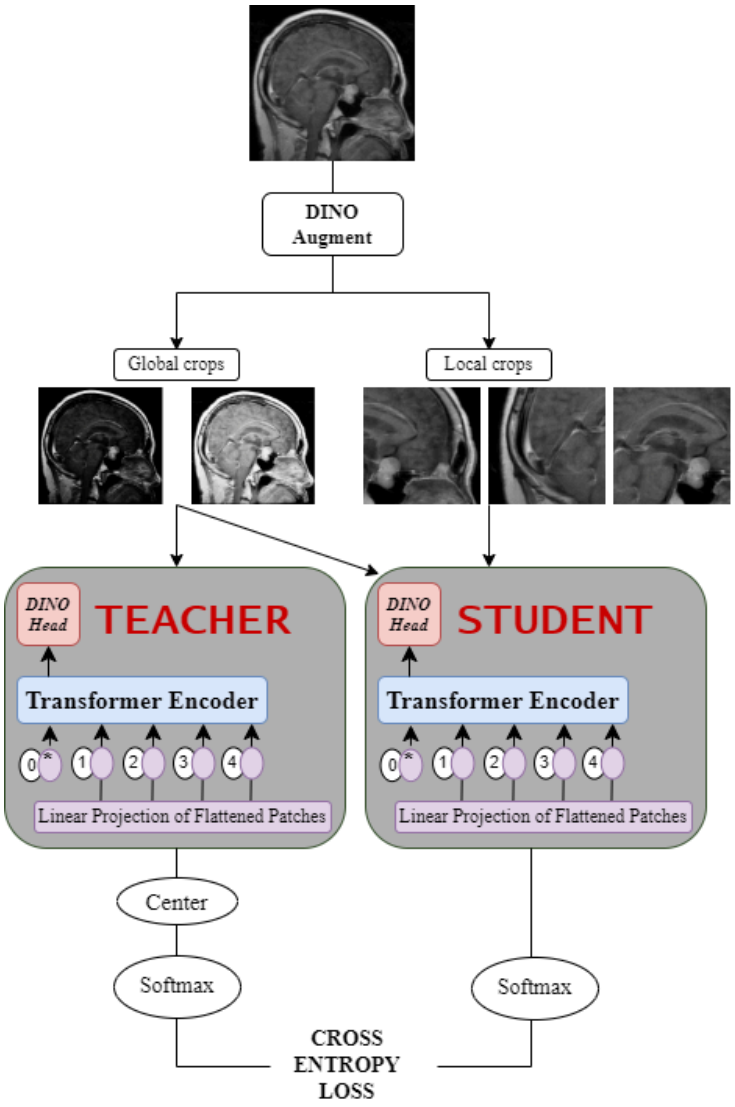


Figure 2: Multi-crop augmentation process

In the DINO model, **only** the *student* network updates itself, while the *teacher* network is built from past iterations of the *student* network using *momentum-based* update rules. The *teacher*'s parameters are an **exponential moving average** (***EMA***) of the *student*'s parameters to be dynamically updated. The ***EMA*** value (λ) follows a cosine schedule from **0.996** to 1 during training.

$$\theta_t \leftarrow \lambda\theta_t + (1 - \lambda)\theta_s$$

This update rule ensures that the *teacher* network consistently performs **better** than the *student* network, providing high-quality targets for the *student* to learn from.

To avoid collapse, DINO applies two operations: *centering* and *sharpening*. *Centering* prevents one dimension from dominating the *teacher* outputs, while *sharpening* ensures that the output distribution is not overly uniform. Together, these operations maintain balance, preventing the collapse and encouraging the model to learn diverse features.

2. DINOv2

Unlike the DINO model, DINOv2 does not require finetuning, allowing it to be suitable for use as a backbone for many different computer vision tasks. It also surpasses traditional image-text models by extracting features beyond what captions provide. DINOv2 improves significantly upon DINOv1 in **4 fields**:

- **Architecture**: It refines the ViT architecture for better feature extraction and scalability, making it highly effective for larger datasets and more complex tasks. A key improvement in DINOv2 comes from the combination of DINO and iBOT losses, which helps the model learn richer and more diverse feature representations. The iBOT loss improves localization and semantic understanding, complementing DINO's self-supervised learning method.

$$L_{iBOT} = -\ \Sigma p_{ti} log\ p_{si}$$

- **Training Efficiency**: The KoLeo regularizer is introduced to **stabilize** training, ensuring that the model doesn't overfit or collapse to trivial solutions.

$$L_{KoLeo} = -\frac{1}{n}\sum_{i=1}^{n} log(d_{n,i})$$

DINOv2 also incorporates SwAV centering, which further enhances the balance and sharpness of the learned features, promoting more diverse representations. The training process leverages advanced techniques like *mixed-precision training*, *fully sharded data parallelism*, *stochastic depth*, and *memory-efficient attention*

(*xFormers*) to enhance training **efficiency**, allowing the model to run faster (x2) and use less memory (1/3).

- **Generalization**: DINOv2 addresses several challenges from DINOv1, such as creating a large, curated training dataset and optimizing the training process. It uses the LVD-142M dataset, which consists of 142 million images, to improve robustness in self-supervised learning.

- **Performance**: DINOv2 also leverages self-distillation to compress large models (ViT-Small, ViT-Base, ViT-Large) into smaller, efficient ones, reducing inference costs and making deployment on limited hardware more feasible.

II. Methodology and Materials

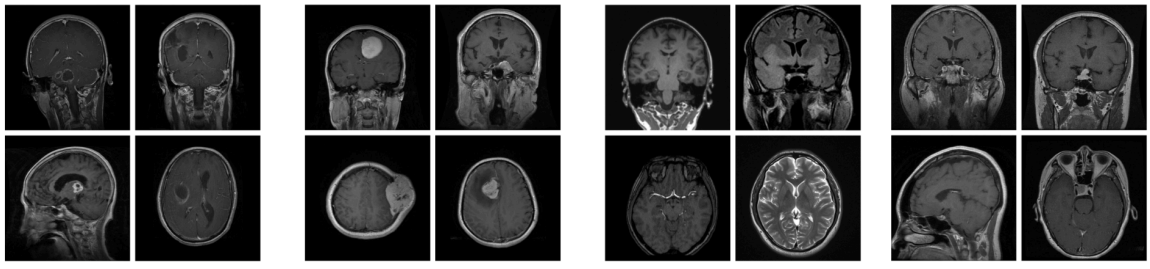
1. Overall methods

Since the DINO model is developed to demonstrate strong out-of-distribution performance and the produced features are usable without requiring any finetuning.

2. Dataset

For this project, we used the *Brain MRI Images for Brain Tumor Detection* dataset from Kaggle, which contains 7,023 Brain MRI images that detect if there's a brain tumor or not. The raw images are in different sizes, taken from different points of view, including 4 classes:

Glimora ***Meningioma***
Pituitary ***No tumor***



Class 0: ‘glioma’ Class 1: ‘meningioma’ Class 2: ‘notumor’ Class 3: ‘pituitary’

3. Model

The DINOv2 model extracts rich feature representations from input images through its self-supervised learning framework. These representations then pass into the Classification Head, which is a fully connected layer that takes the feature representations and outputs probabilities for each class. Unlike the original transfer learning approach where the DINOv2 weights are frozen, in this case, we update the DINOv2 weights during training. This allows the feature extractor to continuously adapt and improve based on the specific brain MRI classification task.

III. Implementation

1. Training settings

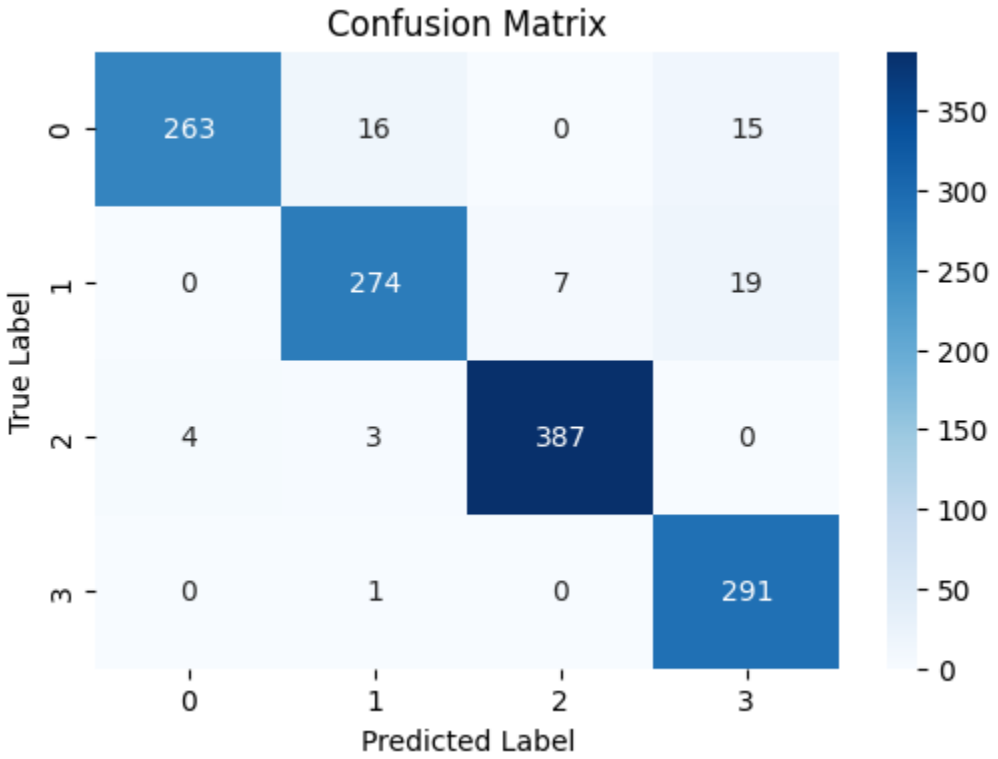
	Train set transformation	**Test set transformation**
Size	5,712 images (80%)	1,311 images (20%)
Resize	224 x 224 pixels	256 x 256 pixels
Augmentation	Randomly flip horizontally for increase the dataset's variability	Crop the center to 224 x 224 pixels
Normalize	Based on predefined mean and standard deviation → ensure the pixel values have a consistent distribution, which helps the model converge faster and more stably during training	Ensure the test data is processed in the same way as the training data, allowing for accurate evaluation
Training model setting		
Optimizer	Adam	
Learning rate	0.000001	
Loss function	CrossEntropy	
Epochs	5	

2. Evaluation & Results

Accuracy: **94.92%**

→ This accuracy is a remarkable result, indicating that our classifier model is performing well overall.

Confusion matrix



Metrics

	Precision	Recall	F1-score
Class ‘glioma’ (0)	0.99	0.89	0.94
Class ‘meningioma’ (1)	0.93	0.91	0.92
Class ‘notumor’ (2)	0.98	0.98	0.98
Class ‘pituitary’ (3)	0.90	1.00	0.94
Averaged metrics	**0.95**	**0.95**	**0.95**

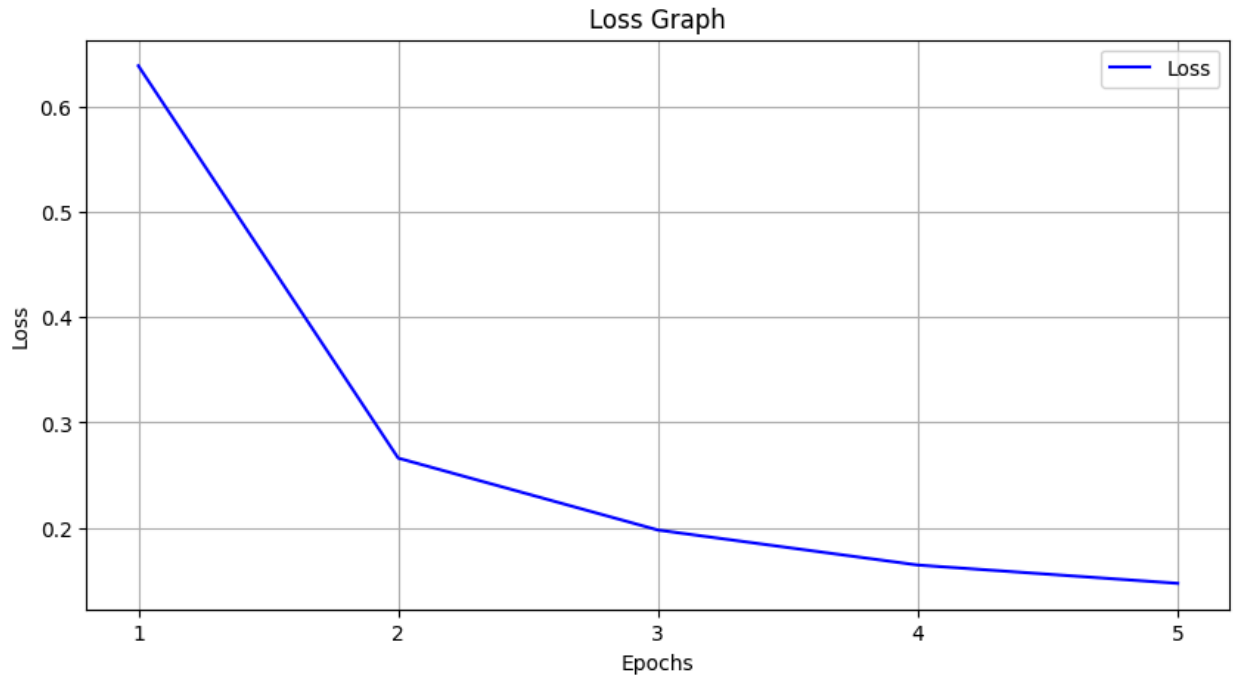
The metrics and the confusion matrix confirms that the model is performing well across all classes, with very few misclassifications.

The *averaged metrics* show that the model treats each class equally (0.95). The balanced performance across precision and recall further implies that the

model is not heavily biased toward any one class and is performing effectively in detecting and correctly classifying diverse brain MRI conditions.

The *confusion matrix* shows that most misclassifications occur between *Class 0* and *Class 1*, which may indicate some similarity or overlap in features between these classes, potentially causing confusion for the model.

Loss graph:



By the 5th epoch, the loss has a stabilized, healthy trend of convergence indicating that the model has successfully learned without overfitting. The smooth convergence without sudden spikes or drops suggests that the model is learning effectively, and the training process is well-tuned.

IV. Conclusion

The model demonstrates good performance on the Brain MRI classification test, with a strong accuracy of 94.92% and consistent precision, recall, and F1-scores across all classes. The loss graph and confusion matrix confirm that the model has mastered the art of accurately classifying MRI images.

References

Caron, M., Touvron, H., Misra, I., Jegou, H., Mairal, J., Bojanowski, P., & Joulin, A. (2021, May 24). *Emerging properties in self-supervised Vision Transformers*. arXiv.org. http://arxiv.org/pdf/2104.14294

Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P.-Y., Li, S.-W., Misra, I., Rabbat, M., Sharma, V., … Bojanowski, P. (2024, February 2). *DINOv2: Learning robust visual features without supervision*. arXiv.org. https://arxiv.org/abs/2304.07193

The DINOv2 team. (2023, April 17). *DINOv2: State-of-the-art computer vision models with self-supervised learning*. AI at Meta. https://ai.meta.com/blog/dino-v2-computer-vision-self-supervised-learning/