**Brain Tumor Classification and Segmentation with DINOv2: Fine-Tuning Unsupervised Feature Extractor on MRI Scans**

Hod Fluger

307868596

[hodfluger@mail.tau.ac.il](mailto:hodfluger@mail.tau.ac.il)

**Abstract**:

Recent advancements in fast, automated medical image analysis have made it increasingly important in both medical research and clinical practice. This study focuses on applying transfer learning techniques for the classification and segmentation of brain tumors from MRI images. The pretrained DINOv2 model, a state-of-the-art self-supervised learning framework, is used as the backbone for both tasks. As a self-supervised model, DINOv2 is particularly valuable in medical imaging, where annotated datasets are often scarce and expensive to obtain. By leveraging this method, I achieved classification results that are comparable to those of weakly-supervised and some fully-supervised models (accuracy of 95.4%). Additionally, the segmentation results demonstrated promising performance (mIoU of 54.4%), showcasing the potential of self-supervised learning in this domain.

**1. Introduction**:

Medical image analysis, particularly in the field of brain tumor detection, is crucial for early diagnosis and treatment planning. MRI scans are commonly used to visualize brain tumors, but manually annotating these images for classification and segmentation is time-consuming, expensive, and often impractical. The challenge lies in developing efficient, accurate methods that can perform these tasks automatically while minimizing the reliance on large annotated datasets.

In this study, I address this issue by utilizing transfer learning techniques, specifically the DINOv2 model, a self-supervised learning framework that has shown great promise in various computer vision tasks. DINOv2 is particularly well-suited for medical imaging because it can learn meaningful representations from unannotated data, significantly reducing the need for costly labeled datasets. By applying this model, I aim to improve both the classification and segmentation of brain tumors from MRI scans.

This study results show that the transfer learning approach with DINOv2 achieves classification performance comparable to weakly-supervised and some fully-supervised models. Additionally, the segmentation results indicate significant promise, demonstrating that self-supervised learning can be a powerful tool in medical image analysis. These findings suggest that self-supervised techniques can pave the way for more scalable and efficient methods in medical imaging, ultimately supporting faster and more accurate clinical decision-making.

**2. Related Work**:

Classifying and segmenting brain tumor images from MRI scans have become key challenges in the computer vision field in recent years. Most studies have focused on convolutional neural networks (CNNs) for classification and Encoder-Decoder architectures, such as UNet, for semantic segmentation. For instance, Srinivasan et al. (2024)1 introduced a deep CNN for multi-class brain tumor classification, achieving superior results compared to popular architectures like ResNet-101, VGG-16, and AlexNet. Similarly, Buda et al. (2019)2 and Al Nasim et al. (2022)3 proposed enhanced UNet architectures for brain tumor segmentation, yielding impressive results.

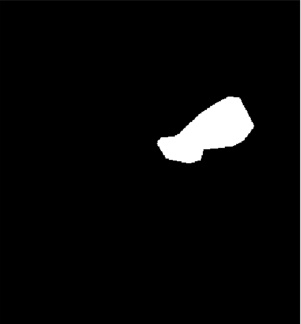
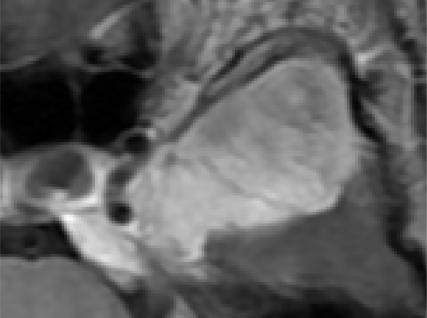
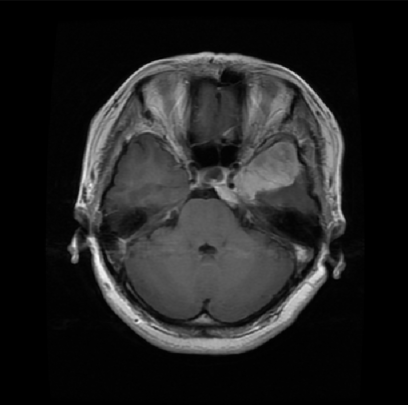
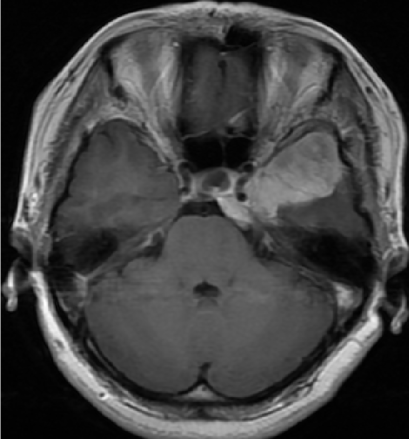
Vimala et al. (2023)4 addressed the brain tumor classification problem by applying transfer learning techniques, fine-tuning the model with MRI data, and achieving excellent accuracy with the same dataset used in this study. They utilized the supervised EfficientNet model as the backbone. Their work inspired this study, though I approached the problem with a different backbone. In this research, I aim to replace the supervised EfficientNet backbone with an unsupervised, pre-trained DINOv2 model. While it is challenging to replicate the results of a supervised model using an unsupervised one, achieving similar performance would highlight the advantages of the unsupervised approach for this task. Additionally, I extended the study by training a semantic segmentation model with the same DINOv2 backbone to detect tumor pixels, a feature not explored in the original work (As will be explained later, having the segmented tumor can improve classification results significantly).

In the original DINOv2 paper, Oquab et al. (2024)4, Meta AI proposed a self-supervised model as a backbone for various computer vision tasks, including both classification and semantic segmentation. The method was evaluated on 12 image and video classification benchmarks, achieving an average accuracy of 90.1% with the “ViT Base” model I also used, and 92.1% with the “ViT Giant” model. Moreover, it demonstrated significant improvements in semantic segmentation, outperforming other self-supervised and weakly supervised models, with a mean Intersection over Union (mIoU) greater than 64%.

**3. Data**:

I used a publicly available dataset consisting of 3,064 brain MRI T1-weighted contrast-enhanced images from 233 patients, featuring three types of brain tumors: meningioma (708 slices), glioma (1,426 slices), and pituitary tumor (930 slices)5. Those 3 types differ in their origin, characteristics, and the parts of the brain they affect. The dataset also includes binary masks, with 1s indicating the tumor regions in the entire brain scan.

The following three pre-processing steps were performed:

1. Brain Cropping: Using Otsu’s global thresholding, I extracted the brain's bounding box and cropped the images to minimize the number of background pixels. The same cropping operation was applied to the corresponding binary masks. In most brain tumor classification studies, skull filtering is a common preprocessing step done as well, where the brain scan is masked to remove pixels corresponding to the skull surrounding the brain. However, I chose to skip this stage.
2. Tumor Cropping: Based on the binary masks, I created a secondary dataset containing cropped images around the tumor region, with a 20-pixel margin in every dimension from the tumor's bounding box. This dataset was used in addition to the "Brain Cropped" dataset, not as a replacement. It essentially resulted in each data sample having a unique spatial ratio..
3. Dataset Preparation: Both the brain-cropped and tumor-cropped datasets were randomly split into training (70%), validation (15%), and test (15%) sets, with the same split applied to both datasets. 

(a) (b) ( c) ( d)

Figure 1. Example image during the pre-processing various steps: a- Original Image, b-”Brain Cropped” , c- Tumor Mask and d- “Tumor Cropped”.

**4. Pre-trained unsupervised DinoV2 Model as a Feature Extractor**:

DINOv2 (Self-Distilled Vision Transformer) is a state-of-the-art, foundational self-supervised learning model that enhances its predecessor, DINO ("Distillation with NO labels"). Designed to learn high-quality visual representations from unlabeled data, DINOv2 is especially valuable for tasks where annotated data is scarce or unavailable. The model utilizes a self-distillation technique and Vision Transformers (ViTs) to leverage large volumes of unlabeled data for effective feature learning.

Like other Vision Transformers, the model divides an image into fixed-size patches, typically 16x16 or 32x32 pixels. These patches are treated as individual tokens, which are then embedded into a high-dimensional space. The transformer architecture processes these patches by learning contextual relationships between them, helping the model capture both local and global information. The model employs a self-distillation approach, where a “student” model is trained to match the output of a “teacher” model (which is updated more slowly) using contrastive learning. This process does not require labeled data, as the model learns to differentiate between different views of the same image. During training, a subset of the patches from an image is randomly masked, meaning they are hidden from the model. The model is tasked with learning to predict or infer the missing information based on the context provided by the unmasked patches. This forced reconstruction of the masked patches helps the model learn more robust and generalized features from the available data.

The resulting visual features are rich in discriminative power and can be directly applied to various computer vision tasks, often with minimal additional training (e.g., using simple linear classifiers).

DINOv2 is available in four model variants, each with a different number of parameters: "small" (21M), "base" (86M), "large" (300M), and "giant" (1100M). These models were pretrained on a dataset of 142 million images, sourced from both curated and uncurated datasets, all without labels or annotations. During the dataset creation, deduplication and clustering operations (using techniques like KNN, cosine similarity, etc) were applied to the embedded vectors of the raw images, enabling smart image selection for better generalization.

Using DINOv2 as a back-bone offers several advantages:

* No fine-tuning is needed - only the "head" of the model (concatenated over the backbone), tailored for a specific task, requires training, while the backbone weights can remain frozen.
* Training smaller "head" models significantly reduces hardware requirements.
* Self-supervised learning improves pixel-level understanding.
* Transfer learning has shown to deliver superior results compared to training smaller models from scratch.

In this study, I used the pretrained "Base" model as the backbone and incorporated transfer learning techniques by adding simple linear and/or convolutional layers to fine-tune the model for brain MRI classification and segmentation tasks.

Leveraging a self-supervised model like DINOv2 for MRI tasks offers several distinct advantages. This approach allows high performance despite the limited availability of labeled data, which is particularly valuable in medical imaging, where annotating datasets is costly and time-consuming. Additionally, as mentioned, self-supervised learning encourages the model to learn generalizable features that are not task- or dataset-specific. This is crucial in MRI tasks, where variations in anatomy and imaging conditions can differ across datasets.

**5. Tumor Classification**:

Inspired by the transfer learning approach using EfficientNet, as described at Babu et al., 2023, I incorporated a Global Average Pooling (GAP) layer applied to the DINOv2 feature map of size 768×256. This operation reduces the feature map to a 768×1 vector for each image, effectively compressing the spatial dimensions while retaining the most informative features. The GAP layer reduces the computational load for subsequent layers, introduces translation invariance, and enhances the model's ability to generalize to unseen data. To further mitigate overfitting during training, a Dropout layer with a drop probability of 0.2 is applied. Finally, a fully connected layer is added to map the 768-dimensional feature vector to three output channels, corresponding to the three classes: meningioma, glioma, and pituitary tumor.

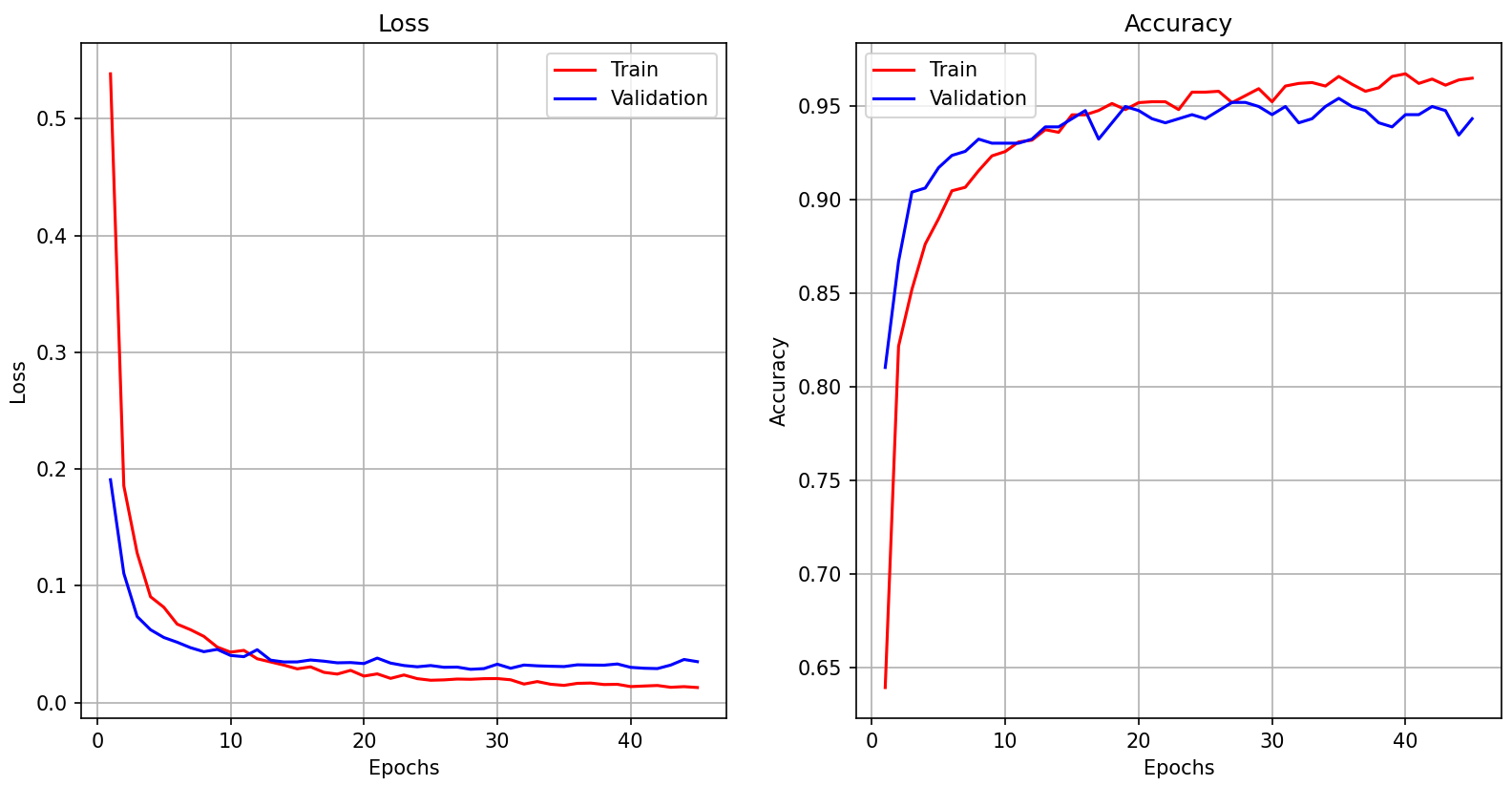


Figure 2. Architecture of the classification model based on DINOv2 model as a backbone

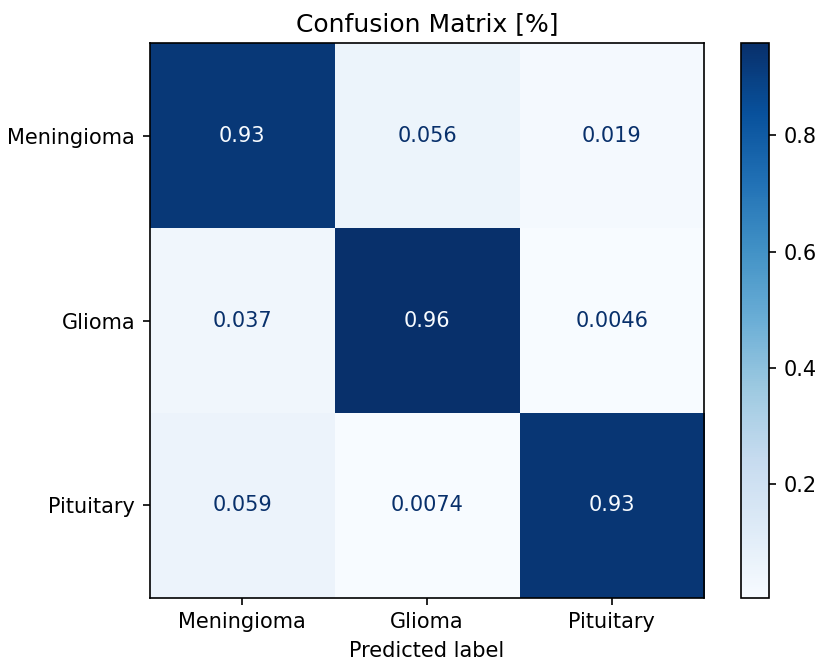
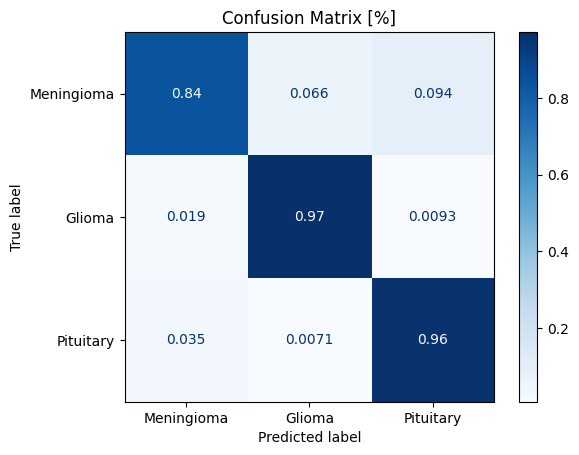
Training: During the training process, I used the Adam optimizer with an initial learning rate of 0.0005. The weights of the backbone were frozen, and only the classifier weights were updated. I experimented with three variants of the cross-entropy loss function, which is commonly used for classification tasks. These variants include:

* The standard cross entropy: where represents the score output by the model for class i.
* A weighted version of cross-entropy: Here, the weights are determined based on the class distribution in the training set, with less frequent classes receiving larger weights. Specifically, the weight for each class is given by . For example, we found that the “Meningioma” class had fewer labeled samples and required balancing.
* Focal loss function: , introduced by Lin et al, (2017)6,is designed to give more focus to harder-to-classify examples, making it especially useful for imbalanced class distributions.

I fine-tuned the hyperparameters by experimenting with various combinations of batch sizes, learning rates, and loss functions. Several image augmentations were tested, including random horizontal flips, random scaling, and 'tumor cropping,' as outlined in the data section. Each image was resized to 224x224 pixels, expanded into three channels to match the input size of the DINOv2 model, and normalized to the ImageNet statistics (mean and standard deviation) since the backbone was trained on ImageNet and other datasets.

The best model achieved validation accuracy of  **95.4%** and test accuracy of **93.7%**. The model was trained with an initial learning rate of 0.0005, a batch size of 16, Focal loss with a of 1, and the 'tumor cropped' dataset. Tumor cropping resulted in an accuracy improvement of approximately 4%. As expected, smaller batch sizes (16, 28, 32) yielded better results than larger batch sizes, likely due to the relatively small size of our dataset. These classification results outperform most of the 12 fine-grained classification benchmarks described by Oquab et al. (2024), all of which employed the DINOv2 base model as a backbone. Moreover, this model outperformed all the self-supervised methods and weakly-supervised methods mentioned in the DINOv2 paper, Oquab et al. (2024)

(a) (b)

Figure 3. a. Loss per epochs, b.Classification accuracy per epochs.

(a) (b)

Figure 4 . Confusion matrix of the **test** set, b. Confusion matrix of the **validation** set

We can see in the confusion matrix that in spite of the usage of Focal loss, the model has some difficulties in differentiating Meningioma tumor from the two other types.

**6. Tumor Segmentation**:

For the segmentation task, I used the same DINOv2 ViT Base as the backbone, but with a modified head. I added a classifier on top of the backbone, where the feature map (768x256) was resized. The 768 channels were reduced to 2 channels using a 1x1 convolutional layer, defining two classes for each pixel: "Tumor" and "Background." Every 256x1 feature was resized into 16x16 token, used as probability maps for the classifier. Next, the 16x16 probability map for each class was bi-linearly interpolated back to the original DINOv2 input size (224x224), and the class with the higher probability was selected for each pixel. This step is essentially equivalent to the decoder in a UNet architecture. Choosing this architecture was inspired by the [Transformers-Tutorials](https://github.com/NielsRogge/Transformers-Tutorials/blob/master/DINOv2/Train_a_linear_classifier_on_top_of_DINOv2_for_semantic_segmentation.ipynb) by Neils Rodge (on git-hub).7

The resulting output mask (contains '1' for tumor and '0' for background) was then refined using morphological operations to filter out small, irrelevant false positives, followed by a Conditional Random Fields (CRF) algorithm for fine-tuning8 (I used the CRF implementation from [Mr-TalhaIlyas](https://github.com/Mr-TalhaIlyas) git-hub9).

Figure 5. Architecture of the semantic segmentation model based on DINOv2 model as a backbone

Training: During the training process, I used the AdamW optimizer with an initial learning rate of \_\_. The weights of the backbone were frozen, and only the classifier weights were updated. I used a combination of two loss functions:

* Binary cross entropy (BCE), refers to the segmentation task as 224x224 classification tasks (for each pixel), with two labels - “Tumor” and “Background”.
* Dice Loss - measures the similarity between two sets of data. It is commonly used in image segmentation to evaluate the overlap between a predicted segmentation mask and the target segmentation mask. The most popular segmentation evaluation metric is the Mean Intersection Over Union (mIoU) and it needs to be maximized during training, thus the Dice loss function that needs to be minimized is defined as: . where are the probabilities of every class per every pixel, calculated by applying SoftMax on the logits.

The combined loss function is defined as follows: where is a hyper parameter.

I fine-tuned the hyperparameters by experimenting with various combinations of batch sizes, learning rates, and loss function weights. Several image augmentations were tested, including random horizontal flips, “brain cropping” (as detailed in the data section), and additional random cropping to focus on different regions of the frame. Each image was resized to 224x224 pixels and expanded into three channels to match the input size of the DINOv2 model. The images were then normalized in one of two ways: using the mean and standard deviation of the ImageNet dataset (as in the classification task) or normalizing to a mean of 0.5 and a standard deviation of 0.5, which is also a common practice.

The best performance was achieved after training with a learning rate of 0.0005, a batch size of 16, normalization using ImageNet statistics, and a combined loss function with a weight of =0.5.

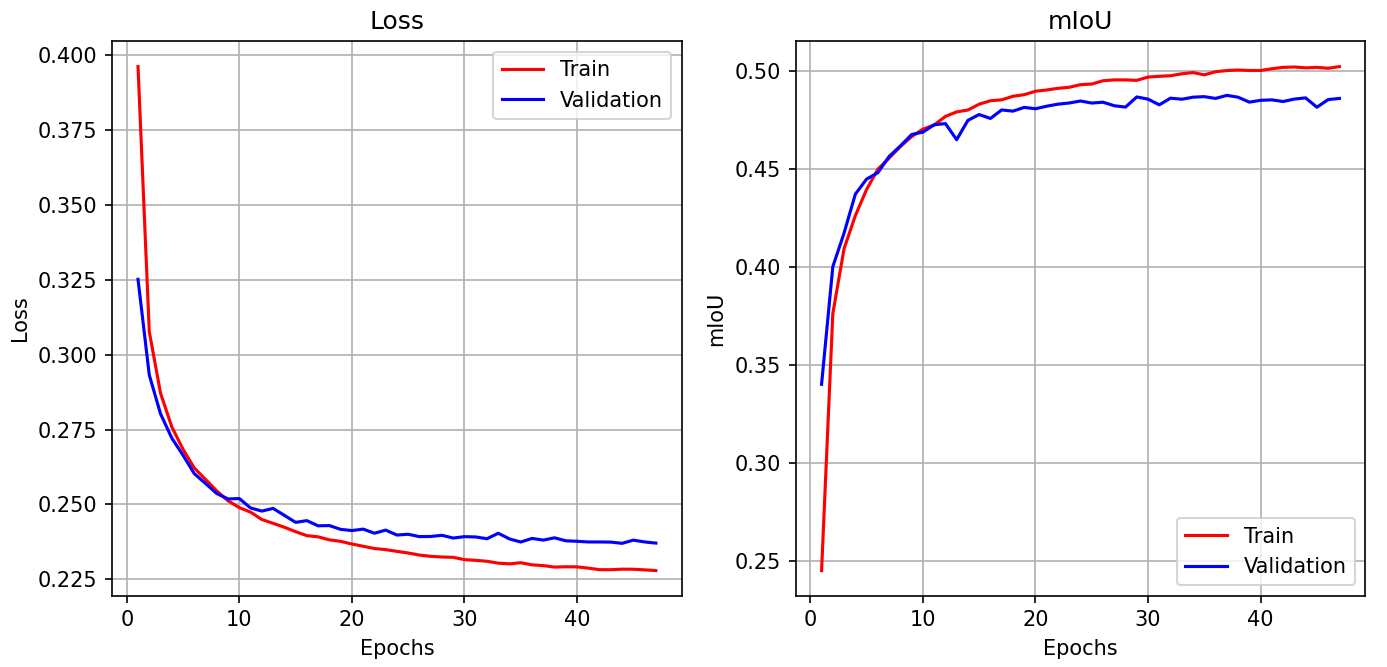


Figure 6. Loss and nIoU per epochs.

I found that, rather than predicting based on the max score, lowering the threshold for the “Tumor” class to 0.17 (instead of the default 0.5) slightly degraded the model’s direct performance but improved post-processing results. This suggests that the data suffers from class imbalance (as expected, since the tumor occupies far fewer pixels than the background), and the original model, with its conservative threshold of 0.5, was not sensitive enough to detect the tumor class effectively. It seems the model struggled to capture the subtle or smaller features of the tumor, leading to lower classification confidence.

| **Metric /**  **Segmentation Predictions** | **Validation**  **mIoU [%]** | **Test**  **mIoU [%]** | **Validation**  **F1 Score [%]** | **Test**  **F1 Score [%]** |
| --- | --- | --- | --- | --- |
| Original Model Tumor Predictions (Th=0.5) | 48.6 | 47.6 | 61.7 | 59.9 |
| Original Model + Post Processing  (Morphological Filtering + CRF) | 52.4 | 51.4 | 63.3 | 61.9 |
| Lowering Tumor Prediction Threshold (to 0.17)  +Post Processing  (Morphological Filtering + CRF) | 54.4 | 52.8 | 65.9 | 63.9 |

Table 1 - Results of the best segmentation model, before and after post-processing.

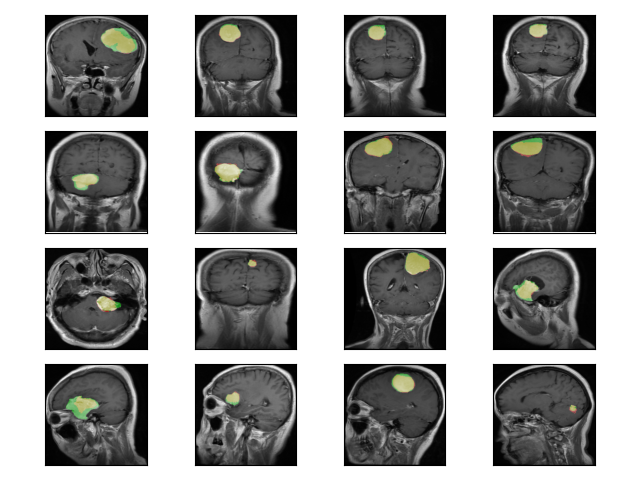
****

Figure 7. Samples of tumor segmentations - green: the ground truth tumor, red: the tumor final prediction, yellow: true positive pixels (the overlapping between prediction and target)

**7. Conclusions**:

In this study, I experimented with DINOv2 as a feature extractor backbone for both the segmentation and classification of brain tumors in MRI scans. I achieved a classification accuracy of 95.5%, which outperformed the average classification accuracy across 12 benchmarks from the original DINOv2 paper. Additionally, I achieved a segmentation mIoU of 54%, which is a reasonable result but can be further improved.

Brain tumor classification from MRI scans is a complex task that typically requires large datasets. However, by using a self-supervised model like DINOv2 as the backbone, I was able to achieve high accuracy with fast training of a simple linear classifier "head." The accuracy can be further improved with a more extensive and balanced dataset, as the dataset I used had a limited number of meningioma tumor images.

For the semantic segmentation task, I found that self-supervised models are generally not as finely tuned for specific tasks like segmentation compared to supervised models. Nonetheless, they can still achieve an mIoU in the range of 50%-60% for certain tasks. My segmentation performance (54%) using DINOv2 could be improved by adding more layers to the decoder "head" or by expanding the dataset.

Since tumor segmentation improves classification accuracy, training a combined model for both segmentation and classification using the same backbone is a promising future direction.

**8. References:**

1. Srinivasan, S., Francis, D., Mathivanan, S.K. et al. (2024) “A hybrid deep CNN model for brain tumor image multi-classification.” *BMC Med Imaging* 24, 21 .<https://doi.org/10.1186/s12880-024-01195-7>
2. Buda, Mateusz, Ashirbani Saha, and Maciej A. Mazurowski. "Association of Genomic Subtypes of Lower-Grade Gliomas with Shape Features Automatically Extracted by a Deep Learning Algorithm." <https://doi.org/10.48550/arXiv.1906.03720>
3. Al Nasim, M. A., Al Munem, A., Islam, M., Haque Palash, M. A., Haque, M. A., & Shah, F. M. (2022). “Brain tumor segmentation using enhanced U-Net model with empirical analysis”. <https://arxiv.org/abs/2210.13336>
4. Vimala, B., Srinivasan, S., Mathivanan, S.K. et al. (2023) “Detection and classification of brain tumor using hybrid deep learning models”. <https://doi.org/10.1038/s41598-023-50505-6>
5. Cheng, Jun (2017). brain tumor dataset. figshare. Dataset. <https://doi.org/10.6084/m9.figshare.1512427.v5>
6. T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, (2017) "Focal loss for dense object detection" [10.1109/ICCV.2017.324](https://doi.org/10.1109/ICCV.2017.324)
7. DINOv2/Train\_a\_linear\_classifier\_on\_top\_of\_DINOv2\_for\_semantic\_segmentation.ipynb

<https://github.com/NielsRogge/Transformers-Tutorials/blob/master/DINOv2/Train_a_linear_classifier_on_top_of_DINOv2_for_semantic_segmentation.ipynb>

1. Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., & Yuille, A. L. (2017). DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs”, [arXiv:1412.7062](https://arxiv.org/abs/1412.7062)
2. [Mr-TalhaIlyas](https://github.com/Mr-TalhaIlyas)/[Conditional-Random-Fields-CRF](https://github.com/Mr-TalhaIlyas/Conditional-Random-Fields-CRF) <https://github.com/Mr-TalhaIlyas/Conditional-Random-Fields-CRF?tab=readme-ov-file>

**9. Appendix:**

All my code is available on:

<https://github.com/hodfluger/Brain-Tumor-Classification-and-Segmentation-with-DINOv2>

|  |
| --- |

.