Optimizaing DinoV2 for Medical Imaging

Matan Lange matanlange@mail.tau.ac.il

Daniel Shekel danielshekel@mail.tau.ac.il

Deep Learning in Medical Imaging - 0553554201
September, 2024
Tel-Aviv University

Abstract

In this study, we explored the potential of adapting DinoV2, a self-supervised vision transformer, to medical imaging tasks. By pre-training DinoV2 on domain-specific medical datasets and comparing its performance to RadImageNet, a CNN-based medical imaging model, we demonstrated that self-supervised pre-trained transformers models have the potential to serve as competitive foundation models in the healthcare domain. While DinoV2 initially lagged behind RadImageNet in some tasks, particularly in Thyroid imaging, the model showed stronger performance after domain-specific pre-training, narrowing the performance gap in datasets such as ACL and Meniscus MRI. Code and evaluation data can be found at: https://github.com/Matan-Lange/MediDino

1 Introduction

Medical imaging plays a critical role in modern healthcare, providing essential tools for diagnosing and treating various diseases. With the advent of deep learning, automated medical image analysis has shown great potential for improving diagnostic accuracy and reducing the workload of healthcare professionals. However, the complex nature of medical images, including variations in anatomy and pathology, presents unique challenges for model selection and optimization.

In recent years, self-supervised learning approaches have gained traction for their ability to produce generalizable models without the need for extensive labeled data. DinoV2 is one of the most powerful self-supervised Vision Transformer (ViT) foundation models, capable of producing universal features that are suitable for various downstream tasks. These tasks range from image-level applications like classification and retrieval to pixel-level applications such as depth estimation and semantic segmentation. The strength of DinoV2 lies in its unique teacher-student training method, where the teacher model processes the entire image while the student receives only a random crop. In this project, Inspired by recent advancements in domain-specific pre-training, such as those seen in models tailored for medical applications[3], our objective is to assess how pre-training DINOv2 on medical imaging data enhances its performance in downstream medical tasks.

We evaluated the effectiveness of DinoV2 in medical imaging tasks by comparing its performance against RadImageNet, a medical image foundation model based on ResNet architecture. RadImageNet has been trained specifically for medical image analysis, making it an ideal benchmark for assessing DinoV2's potential. We utilized four of the eight datasets provided by RadImageNet: Thyroid, Breast, ACL, and Meniscus, excluding the larger datasets due to resource limitations. For each dataset, we pre-trained DinoV2 on similar data from RadImageNet database and performed fine-tuning on the specific task at hand.

Our results suggest that DinoV2, when adapted for the medical domain, can serve as a strong foundation model for medical image analysis, demonstrating competitive performance across the selected datasets. This research offers new insights into leveraging self-supervised vision transformers for medical imaging tasks, highlighting the potential of tailoring such models to meet domain-specific challenges.

2 Related Work

The advancement of self-supervised learning has significantly impacted the field of computer vision, enabling models to learn robust representations from unlabeled data. Vision Transformers (ViTs), introduced by Dosovitskiy et al., have been pivotal in this progress due to their ability to cap-

ture global relationships within images. Building on this foundation, Oquab et al. [1] proposed DINOV2, a self-supervised framework that enhances the learning of visual features without supervision. DINOV2 introduces improvements over its predecessor DINO by incorporating multi-scale feature processing, momentum teacher updates, and a more effective objective function. The authors demonstrated that DINOV2 achieves state-of-the-art performance on various benchmarks, outperforming both supervised and unsupervised counterparts in tasks such as image classification and object detection.

In the medical imaging domain, the scarcity of large-scale labeled datasets poses a significant challenge for training deep learning models. Transfer learning from models pre-trained on natural images, like those from ImageNet, often leads to sub-optimal performance due to domain discrepancies. To address this issue, Mei et al. [2] developed RadImageNet, a large-scale radiology image dataset comprising over 1.35 million images across 35 categories and five imaging modalities, including X-ray, CT, MRI, ultrasound, and PET. By pre-training convolutional neural networks (CNNs) on RadImageNet, the authors facilitated more effective transfer learning for medical imaging tasks. Their experiments showed that models pre-trained on RadImageNet significantly outperformed those pre-trained on ImageNet in various radiologic classification and segmentation tasks, emphasizing the importance of domain-specific pre-training.

The potential of self-supervised ViTs in medical imaging has been explored in recent studies. Koch et al. [3] introduced DinoBloom, a foundation model designed to generate generalizable cell embeddings in hematology. By leveraging the DINOV2 framework, they trained ViTs on a large corpus of unlabeled blood smear images. The resulting embedding captured intricate morphological features of blood cells, enabling improved performance in downstream tasks such as cell classification and anomaly detection. Their work demonstrated that self-supervised ViTs could effectively learn representations in specialized medical domains, even surpassing models trained with supervised learning.

Similarly, Roth et al.[4] investigated the application of foundation models in histopathology. They fine-tuned self-supervised ViTs on histopathological datasets with limited annotations and compared the performance against state-of-the-art CNNs. Their findings revealed that the fine-tuned ViTs not only achieved superior accuracy but also required fewer labeled samples, highlighting the efficiency of self-supervised transformers in low-resource settings. This study underscored the viability of adapting foundation models to specialized medical imaging tasks with minimal annotation effort.

Our work extends these advancements by exploring the efficacy of DINOV2 in a broader range of medical imaging tasks and directly comparing its performance with models based on RadImageNet. Unlike Koch et al. [3], who focused on hematology, we pre-train DINOV2 on medical imaging data similar to that used in RadImageNet, encompassing multiple modalities and anatomical regions. Specifically, we utilize four datasets—Thyroid, Breast, ACL, and Meniscus — which are part of the RadImageNet collection. By doing so, we aim to assess the generalizability and robustness of self-supervised ViTs across diverse medical imaging applications.

Our approach differs from previous studies in several key aspects:

Model Architecture: While RadImageNet leverages CNN architectures like ResNet, our work employs the transformer-based DINOV2 model. The self-attention mechanisms in ViTs allow for capturing long-range dependencies and complex patterns in medical images, which may be advantageous over the locality-focused convolutions in CNNs.

Pre-training Strategy: We adopt a self-supervised pre-training approach on unlabeled medical images, in contrast to the supervised pre-training utilized in RadImageNet. This strategy enables the model to learn inherent data structures without relying on extensive annotations, which are costly and time-consuming to obtain in the medical field.

Dataset Utilization: By selecting four representative datasets from RadImageNet, we balance the need for diverse medical imaging tasks with the practical constraints of computational resources. This selection allows for a direct comparison between DINOV2 and RadImageNet-based models under similar conditions.

In summary, our work contributes to the ongoing exploration of transformer models in medical imaging by providing empirical evidence on the effectiveness of self-supervised ViTs pre-trained on domain-specific data. By comparing our results with those obtained from RadImageNet-based models, we aim to highlight the potential advantages of transformer architectures and self-supervised learning in advancing medical image analysis.

3 Data

In this study, we focused on evaluating the performance of the DINOV2 model in medical imaging classification tasks by utilizing four specific

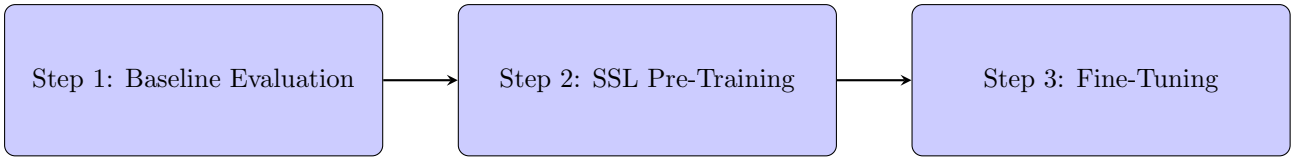

Figure 1: Overview of the Methodology Process

datasets: thyroid ultrasound images, breast ultrasound images, and knee magnetic resonance imaging (MRI) scans focusing on anterior cruciate ligament (ACL) and meniscal tears. These datasets are not part of the RadImageNet database; however, they were also used by the RadImageNet team in their evaluation experiments. Additionally, we leveraged the extensive RadImageNet database, comprising over 1.3 million medical images, for self-supervised pre-training of the DINOV2 model.

3.1 Evaluation Datasets

The four datasets used for reporting classification results are as follows: Thyroid - 349 ultrasound (US) images of the thyroid gland, collected from an open-access thyroid image repository. Breast - 780 breast ultrasound images, this dataset includes images acquired for the detection and diagnosis of breast cancer. It encompasses both normal breast tissue images and images exhibiting pathological findings indicative of malignancies. ACL - includes 1,021 MRI images of the knee, specifically focusing on ACL tears. Meniscal - Containing 4,201 MRI images, this dataset focuses on meniscal tears in the knee joint.

3.2 Pre-training Data

For the self-supervised pre-training of the DINOV2 model, we utilized a targeted subset of the RadImageNet database, specifically selecting images related to thyroid, breast, ACL, and meniscus. Although RadImageNet consists of over 1.3 million medical images across various imaging modalities, anatomical regions, and pathologies, we focused on these particular categories to align the pre-training data closely with our evaluation datasets and to manage computational resource constraints. This subset provided a substantial amount of unlabeled medical imaging data suitable for self-supervised learning, encompassing modalities such as ultrasound (for thyroid and breast) and MRI (for ACL and meniscus).

4 Method

In this study, we aimed to evaluate the effectiveness of the DINOV2 model, specifically the ViT-S/14 architecture, in medical imaging classification tasks. Our approach involved a systematic evaluation of the model's performance under different training regimes, leveraging both general and domain-specific data. We employed a combination of self-supervised learning (SSL) and fine-tuning techniques, along with rigorous cross-validation, to assess the model's capability in comparison to the existing benchmark RadImageNet.

Our methodology consisted of three primary phases: 1.Baseline Evaluation: Assessing the performance of the pre-trained DINOV2 ViT-S/14 model without any additional self-supervised training on medical images. 2.Domain-Specific Self-Supervised Pre-training: Continuing the self-supervised pre-training of DINOV2 on a custom data comprising medical images relevant to our target tasks. 3.Fine-Tuning and Evaluation: Fine-tuning the model on specific classification tasks using the prepared datasets and evaluating its performance using 5-fold cross-validation.

4.1 Model Architecture

We utilized the Vision Transformer Small with 14x14 patch size (ViT-S/14) architecture from the DINOV2 family of models. The ViT-S/14 model consists of a transformer encoder with 12 layers, an embedding dimension of 384, and a total of approximately 22 million parameters. The choice of ViT-S/14 was guided by its balance between performance and computational efficiency, making it suitable for training with limited resources.

4.2 Baseline Evaluation without Pretraining

Initially, we evaluated the DINOV2 ViT-S/14 model pre-trained on ImageNet without any additional self-supervised learning on medical images. This step aimed to establish a baseline performance metric to compare against subsequent experiments. The model was fine-tuned on each of the four datasets—thyroid ultrasound, breast ultrasound, ACL MRI, and meniscal tear MRI—using the same data augmentation pipeline and training settings as in RadImageNet's experiments.

4.3 Domain-Specific SSL Pretraining

Recognizing that medical images possess characteristics distinct from natural images, we hypothesized that further pre-training the model on domain-specific data would enhance its performance. Therefore, we conducted self-supervised

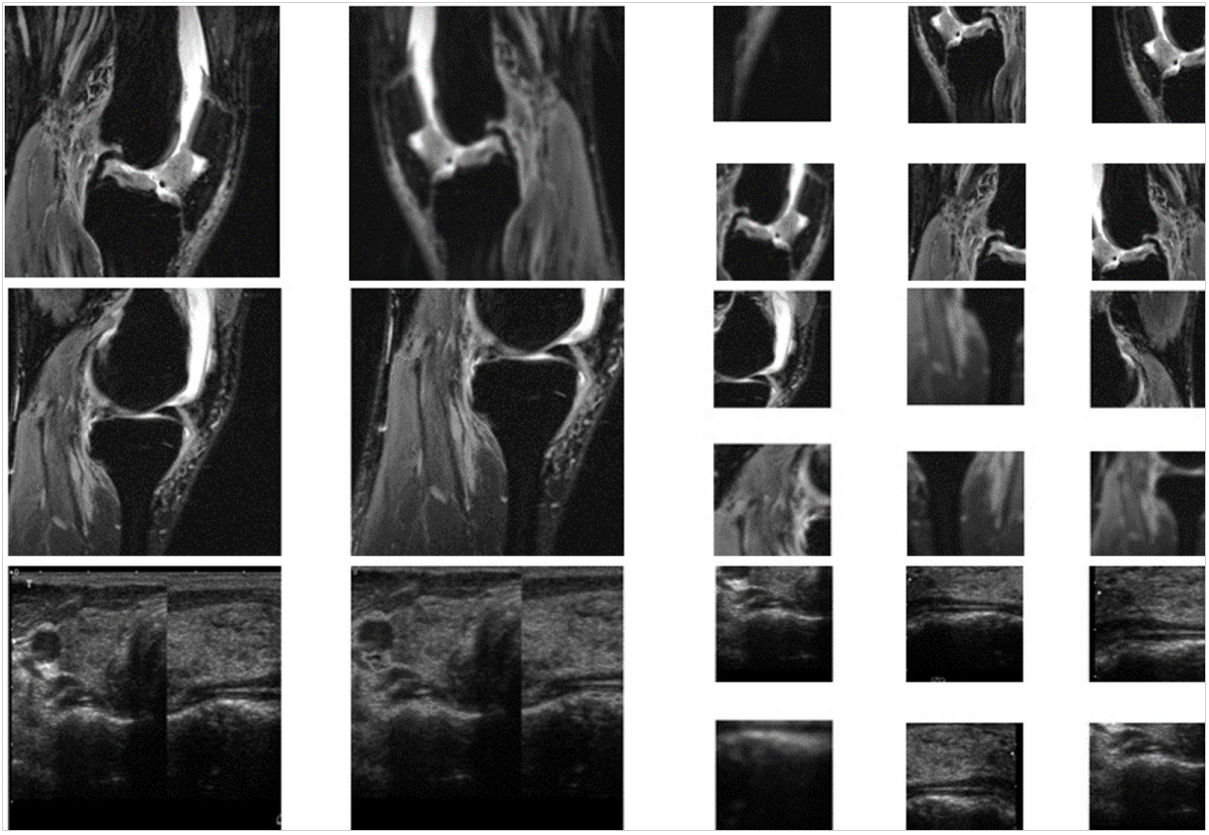

Figure 2: Local and global crop. upper image: ACL, middle image: Meniscus, lower image: Breast

pre-training of the DINOV2 ViT-S/14 model on a custom dataset extracted from RadImageNet, consisting of images related to thyroid, breast, ACL, and meniscus. This dataset included a substantial number of unlabeled images across relevant imaging modalities (ultrasound and MRI), providing the model with exposure to domain-specific features.

The self-supervised pre-training followed the DINOV2 framework, utilizing the multi-crop strategy and teacher-student setup:

Multi-Crop Strategy: Each image was augmented to produce multiple views at different scales, including several global crops (covering most of the image) and local crops (smaller patches). This encouraged the model to learn scale-invariant and context-rich representations. see example in Figure 2.

Teacher-Student Setup: A momentum encoder updated via an exponential moving average of the student encoder's weights acted as the teacher. The student network learned to match the teacher's output representations for the same images, promoting consistency across different views.

4.4 Fine-Tuning SSL custom model

After the domain-specific self-supervised pre-training, we fine-tuned the model on each of the four classification tasks

5 Experiments

A summary of our result you can find in the next table.

Dataset	IN	RIN	DinoV2	SSL
Thyroid	0.76±0.14	**0.85±0.09**	0.71±0.01	0.74±0.03
Breast	0.9±0.1	0.94±0.05	**0.954±0.005**	NA
ACL	0.91±0.08	**0.97±0.03**	0.957±0.005	0.966±0.002
Meniscal	0.92±0.06	**0.96±0.02**	0.91±0.01	0.927±0.007

Table 1: AUC Best Results for Different Model Starting Points in Fine-tuning for Specific Dataset Tasks: Dinov2 (Pretrained on ImageNet) and Custom SSL (Trained on Dinov2 with Similar Data to the Specific Task) IN (ImageNet) RIN (RadImageNet), SSL (Self-Supervised Learning)

5.1 SSL Custom Model Training

The custom SSL model training was conducted on a single A100 GPU with 80 GB VRAM. Each model was trained for 200 epochs, with training times of at least 4 hours per model. The batch size used during training was 256. Most of the other parameters were retained from the original Dinov2 training configuration. Training augmentations were similar to those used in the work by *Low-resource fine-tuning of foundation models beats state-of-the-art in histopathology.*[4]

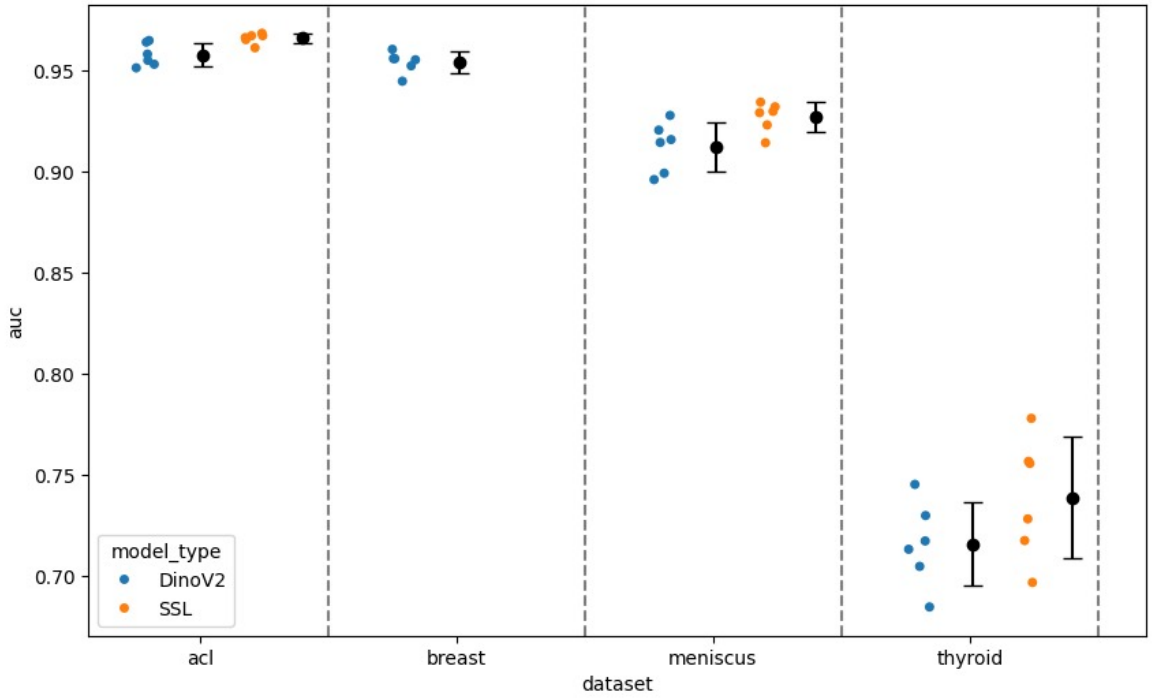

Figure 3: AUC Detailed Results

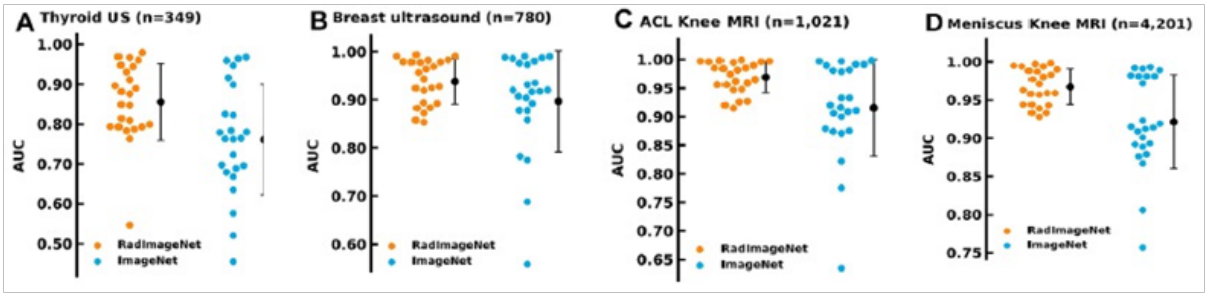

Figure 4: RadImageNet and ImageNet Detailed AUC Results

5.2 Experimental Setup for Fine-tuning

The fine-tuning experiments were conducted similarly to the RadImageNet methodology. Models were trained for 30 epochs using fivefold cross-validation, ensuring the use of the same folds as in RadImageNet. During training, the models with the lowest validation loss were saved for further evaluation and comparison. Full fine-tuning was performed without freezing layers, in contrast to the approach taken in RadImageNet.

Learning rates of 1×10^{-5} and 1×10^{-6}, along with different batch sizes of 8, 16, and 32, were tested across six settings. The average AUC was compared across these configurations. Binary cross-entropy was employed as the loss function, and the input images were resized to 224X224 pixels.

For the vision transformer architecture, a linear layer with ReLU activation was added, followed by an output layer activated by the sigmoid function to accommodate binary classification.

6 Conclusion

The detailed results presented in Figures 3 and 4 highlight that the self-supervised learning (SSL) approach showed an overall improvement compared to DinoV2. Both methods, DinoV2 and SSL, exhibited smaller standard deviations in their performance compared to RadImageNet, indicating a higher level of robustness. However, when examining the performance across the different modalities, DinoV2 pre-trained on ImageNet underperformed RadImageNet in 3 out of the 4 modalities tested. The exception was the breast dataset, where DinoV2 outperformed RadImageNet, suggesting that specific modalities may benefit more from DinoV2's transformer-based architecture. Notably, the customized SSL approach improved the area under the curve (AUC) in both the Thyroid and ACL datasets, but it still fell short of matching RadImageNet's performance. It is important to emphasize that the self-supervised pre-training on the medical datasets was focused, requiring approximately 4 hours of computation. Despite the relatively short training duration, the SSL model was able to exceed the performance of the original ImageNet pre-trained DinoV2 and approach the performance of RadImageNet. These findings underscore the potential of domain-specific self-supervised pre-training for enhancing model performance in medical imaging, though further refinements are necessary to consistently surpass established models like RadImageNet across all tasks.

Self-supervised learning (SSL) in medical imaging warrants further investigation, particularly with the integration of augmentation techniques tailored to the unique characteristics of medical data. Specifically, methods such as fan-shaped augmentations and artifact-based techniques, as suggested by Ramarkers et al. [5] and Goceri et al. [6], offer promising avenues for enhancing model performance. We hypothesize that incorporating these domain-specific augmentations may lead to performance improvements comparable to, or exceeding, current results in medical image analysis. Further exploration of these strategies could provide valuable insights into optimizing SSL for medical applications.

Our experiments underline the importance of customizing pre-training strategies for medical data. The inherent differences between natural and medical images, such as lower spatial variability and the critical nature of small anatomical details, make domain-specific pre-training essential for improving model performance. This approach holds particular promise for medical imaging applications where labeled data is scarce, as self-supervised learning can leverage vast amounts of unlabeled medical data to enhance model generalization.

Overall, this research contributes to the growing body of evidence supporting the use of transformer architectures in medical image analysis. It also paves the way for future studies exploring deeper optimizations of self-supervised vision transformers in specialized medical fields. Further exploration of model architecture, training strategies, like unique augmentation for medical imaging will be crucial in advancing this work toward clinical applicability.

References

[1] Oquab, M.et al (2023). DINOV2: Learning robust visual features without supervision. *arXiv.org*. https://arxiv.org/abs/2304.07193

[2] Mei, X., et al (2022). RaDImageNet: An open radiologic deep learning research dataset for effective transfer learning. Radiology Artificial Intelligence, 4(5). *RSNA.org* https://doi.org/10.1148/ryai.210315

[3] Koch, V., et al (2024, April 7). DinoBloom: a foundation model for generalizable cell embeddings in hematology. *arXiv.org. https://arxiv.org/abs/2404.05022*

[4] Roth, B., et al (2024, January 9). Low-resource finetuning of foundation models beats state-of-the-art in histopathology. *arXiv.org. https://arxiv.org/abs/2401.04720*

[5] Ramakers, Florian., et al (2024, June). UltraAugment: fan-shape and artifact-based data augmentation for 2D ultrasound images *Ramakers2024CVPR.* https://openaccess.thecvf.com

[6] Goceri, E. (2023). Medical image data augmentation: techniques, comparisons and interpretations. Artificial Intelligence Review, 56(11), 12561–12605. *https://doi.org/10.1007/s10462-023-10453-z*