2024 3rd International Conference on Cloud Computing, Big Data Application and Software Engineering (CBASE) | 979-8-3315-0658-2/24/$31.00 ©2024 IEEE | DOI: 10.1109/CBASE64041.2024.10824558

Transformer-Based Self-Supervised Learning and Distillation for Medical Image Classification

Improving Colorectal Cancer Detection on NCT-CRC-HE-100K with Swin-T V2

Meng Li
Baidu, Inc.
Beijing, China
592137091@qq.com

***Abstract*-In this work, we present a novel approach to colorectal cancer tissue classification using Swin-Transformer V2 on the NCT-CRC-HE-100K dataset. This study is the first to apply Swin-Transformer V2 in this domain, leveraging its advanced architecture to achieve state-of-the-art performance. Building upon the success of previous self-supervised learning methods like MoBY, which demonstrated strong results on natural images, we extend these techniques to medical datasets. We perform self-supervised pretraining on a wide range of tumor-related datasets, incorporating advanced data augmentation strategies, such as random cropping and 2x magnification, to address the multi-scale nature of histopathological images. After pretraining, we employ a progressive layer-wise distillation technique, transferring knowledge from a large teacher model to a more efficient student model. This method dynamically adjusts the distillation strength across layers using a learnable parameter, improving training efficiency and overall performance. Our approach achieves a top-1 accuracy of 96.0% on NCT-CRC-HE-100K, surpassing the previous best by 0.5 percentage points. This work demonstrates the potential of Transformer-based architectures in medical imaging tasks and highlights the effectiveness of self-supervised learning and distillation techniques in improving model accuracy.**

Keywords-Swin-TransformerV2, NCT-CRC-HE-100K, colorectal cancer, self-supervised learning, layer-wise distillation

I. Introduction

Recent advancements in computer vision have seen a paradigm shift from Convolutional Neural Networks (CNNs) to Transformer-based models, most notably Swin-Transformer [1]. In MoBY, a self-supervised learning approach was introduced that employed Swin-T as a backbone for high-accuracy image classification, with additional applications in object detection and segmentation. However, MoBY's application remained limited to natural image datasets like ImageNet-1K. In this work, we present the first use of Swin-Transformer V2 on the NCT-CRC-HE-100K dataset for colorectal cancer tissue classification, pushing the boundaries of medical image analysis with Transformers.

In this work, we explore the study of the Swin-Transformer V2 [2] architecture in the task of colorectal cancer tissue classification using the NCT-CRC-HE-100K dataset. This marks the first application of Swin-Transformer V2 for this specific medical image classification task. Previous studies, such as MoBY, demonstrated the efficacy of self-supervised learning on Swin-Transformer and Vision Transformers in general. MoBY combined techniques from MoCov2 , achieving notable performance on tasks like ImageNet-1K. While these methods focused on natural images, we extend this approach to the domain of medical images, incorporating multi-scale information essential for histopathological analysis. Using a large-scale, self-supervised training setup with multiple tumor-related datasets, we surpass the current state-of-the-art (SOTA) accuracy on NCT-CRC-HE-100K by 0.5 percentage points, reaching a new peak performance of 96.0%.

II. Related Work

The use of Swin Transformer and its variants has gained significant attention in medical image analysis, particularly for segmentation tasks. Several studies have explored self-supervised learning with Swin Transformers, building on their ability to model both local and global relationships effectively.

MoBY [3] a self-supervised learning technique, combined MoCov2 and BYOL [4] to build strong feature representations using Vision Transformers, including Swin-T. This approach focused on natural images but lacked an evaluation on dense prediction tasks. Our work is inspired by MoBY but extends its self-supervised learning framework to include medical datasets such as NCT-CRC-HE-100K, Digest Path, and CAMELYON16, allowing us to explore the applicability of these representations in medical imaging tasks.

In medical image segmentation, Swin-Transformer has demonstrated superior performance by capturing multi-scale features and global context more effectively than traditional convolutional neural networks (CNNs). Introduced Swin-UNet, which integrates Swin-transformer into a U-Net architecture for medical image segmentation, achieving notable improvements [5]. Furthermore, Swin-Transformer has been employed in 3D medical image analysis, such as MRI and CT scans, due to its ability to model both local and global 3D spatial information. Liu et al. demonstrated its superiority in 3D image segmentation tasks compared to conventional CNN models [6]. In medical image classification tasks, such as tumor detection and disease classification, Swin-Transformer has outperformed CNN-based approaches by leveraging its global modeling capabilities, particularly in handling high-resolution medical images [7]. Additionally, Swin-Transformer has shown promise in pathology image analysis, where its hierarchical structure allows it to effectively process large, high-resolution pathology images for cancer detection and tissue classification. The combination of transfer learning and Swin-Transformer has also proven beneficial, particularly in scenarios with limited labeled medical data. where pretrained models on natural image datasets were fine-tuned for medical applications. Looking ahead, the advancements of Swin-TransformerV2, with its enhanced multi-scale processing and computational efficiency, are expected to further drive its adoption in complex medical image analysis tasks.

III. Methods

Swin-Transformer V2 excels on the NCT-CRC-HE-100K dataset primarily due to its powerful pretraining methods and the inherited strengths of Transformer architectures. The model benefits from hierarchical feature extraction, allowing it to capture both fine-grained local details (e.g., cell structures) and global contextual information (e.g., tissue patterns). This is crucial for histopathology tasks, where tumor sizes and tissue textures vary significantly. Swin-Transformer V2's shifted window mechanism optimizes computational efficiency while preserving the ability to model long-range dependencies across image patches, which is especially beneficial for large medical images.

Moreover, the self-supervised pretraining of Swin-Transformer V2 on large unlabeled datasets enhances its ability to generalize, making it less dependent on annotated data, which is often limited in medical imaging. This pretraining strategy allows the model to learn robust representations of histopathological features. Finally, inheriting the Transformer's self-attention mechanism, Swin-Transformer V2 excels in modeling complex spatial relationships within medical images, making it highly effective for tasks like colorectal cancer tissue classification, where both local and global patterns are crucial.

A. *Self-Supervised Pretraining*

For pretraining, we utilized a large-scale self-supervised approach across multiple tumor-related medical datasets, including NCT-CRC-HE-100K [8], Digest Path, TCGA, and CAMELYON16. These datasets cover a wide range of histopathological images, featuring various tumor types and tissue structures. This comprehensive set allowed us to capture the diverse features and scales needed for robust tumor classification.

During the pretraining phase, we applied extensive data augmentation techniques to address the variability in tumor sizes and the lack of multi-scale information often seen in medical datasets. By applying data augmentation, we mimic varying tumor sizes in histopathology images. the model learns to generalize across different tumor scales, positions, and image conditions, ensuring robust feature extraction from the raw images.

The pretraining was conducted on an NVIDIA A100 cluster with 8 GPUs, which allowed us to process these large datasets efficiently. The training lasted for approximately one week, leveraging the significant computational resources to optimize Swin-Transformer V2. This large-scale pretraining setup provided the model with powerful representations, enabling it to generalize well across unseen histopathological images

B. *Distillation*

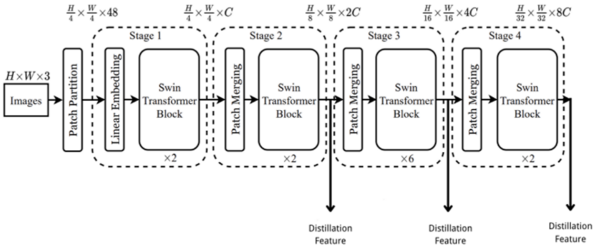


FIGURE 1. DISTILLATION BY LAYERS

Swin-Transformer Tiny as example

As shown in FIGURE 1, after the self-supervised pretraining phase, we fine-tune the Swin-Transformer V2 model on the NCT-CRC-HE-100K dataset using a teacher-student framework. This approach involves distilling knowledge from a pretrained model (the teacher) into the student model. The teacher-student structure is divided into three layers, each with a parameter, α, controlling the distillation process. The α values in each layer is 0.2, 0.15, 0.1 accordingly, ensuring that the model retains valuable pretrained features. (The α parameter also can be a learnable parameter but the performance is similar to this setting)

This approach is designed to preserve the original pretrained features as much as possible, which is crucial for achieving high performance in classification tasks. By maintaining these features, the model benefits from the robust representations learned during the initial pretraining phase. These features are essential for accurately distinguishing between different classes, particularly in complex tasks like colorectal cancer classification.

Although this paper focuses on classification, the distillation strategy employed ensures that the student model is not only optimized for its current task but also well-prepared for potential future applications. By carefully managing the distillation process through adjustable α parameters, the model effectively balances learning new task-specific features while retaining valuable pretrained knowledge, ensuring optimal performance in classification tasks.

By employing this technique, we enhance the overall training process, leading to more accurate and reliable predictions in histopathological image analysis.

C. *Loos Function and Equations*

Cross-Entropy Loss: This loss function is fundamental for classification tasks, measuring the discrepancy between the predicted class probabilities and the true class labels y.

Distillation Loss: This loss is designed to guide the student model to mimic the output of the teacher model. It is defined using the Kullback-Leibler divergence [9] between the softened logits z_T^j from the teacher model and the softened logits z_S^j from the student model, computed as follows:

$$Loss_{Distillation} = \sum_j KL\left(z_T^j || z_S^j\right) = \sum_j z_T^j \log\left(\frac{z_T^j}{z_S^j}\right) \quad (1)$$

Here, j represents the different classes, and the logits are softened using a temperature parameter T to enable smoother probability distributions. The distillation loss facilitates faster and more efficient learning by leveraging the knowledge encoded in the teacher model's outputs.

$$Total\ Loss\ = Loss_{Distillation} + \alpha\ \cdot Loss_{CE} \tag{2}$$

The overall loss function for the student model combines both the cross-entropy loss and the distillation loss, weighted by the α parameter.

IV. EXPERIMENTS

To evaluate the effectiveness of our Swin-Transformer V2 model in colorectal cancer tissue classification, we conducted a series of experiments on the NCT-CRC-HE-100K dataset. Our experiments aimed to compare the performance of Swin-TransformerV2 against other state-of-the-art (SOTA) classification networks and to analyze the impact of different model sizes on classification accuracy.

A. Experimental Setup

All models were trained on the same training and validation splits from the NCT-CRC-HE-100K dataset, employing standardized hyperparameters. The training process utilized a batch size of 32 and a learning rate of 0.001, alongside data augmentation techniques similar to those applied during pretraining.

B. Data Distribution

Image classification on ImageNet-1k has been the mainstream pre-training paradigm for a long period. How- ever, in medical domain ImageNet almost has NO medical related data, since it is designed for real world data and most classes can be recognized and classified by publics. In contrast medical classification requires professional knowledge, also if we initialize model by ImageNet-1k then when we finetune it in downstream tasks with huge data distribution gap, the pretrain process could be useless. since there no medical dataset similar to ImageNet-1k, it is not possible for medical domain to make such large dataset with specific class. However, self-supervised pretraining does not require large amount data labeling, like BEiT-3 uses masked data modeling task, achieving state-of-the-art results on a wide range of down- stream tasks.

Therefore, we obtained massive medical data from internet and opensource dataset, then we use masked data modeling to pretrain Swin-Transformerv2.

In the pretraining phase, we employed a self-supervised learning approach where the Swin-Transformer model was trained to generate meaningful representations without labeled data. This involved using contrastive learning techniques to maximize the similarity between different augmented views of the same image while minimizing the similarity between different images. Data augmentation played a key role in creating diverse views of the images, improving the model’s ability to generalize across various medical imaging modalities. By leveraging the large and diverse dataset with appropriate preprocessing, the model learned robust features that could be fine-tuned for downstream medical image analysis tasks.

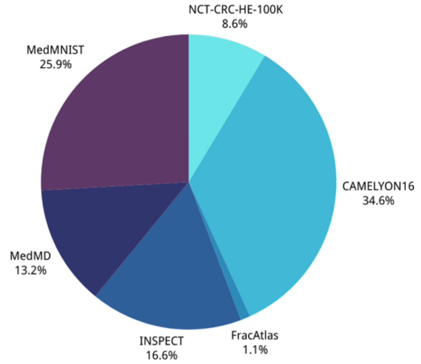
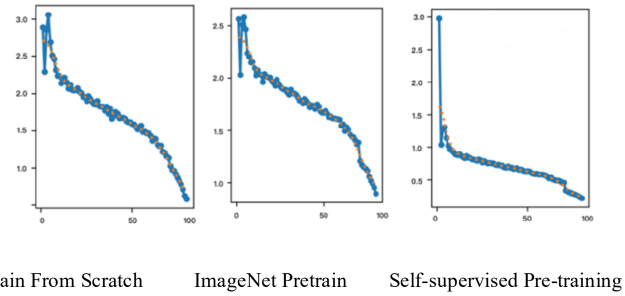


FIGURE 2. SELF- SUPERVISED PRETRAIN DATA DISTRIBUTION

FIGURE 2 illustrate data distributions we used in self-supervised learning. We included 100,000 images from NCT-CRC-HE-100K (8.6% of total) and 400,000 from CAMELYON16 (34.6%), both with random drop due to size. FracAtlas contributed 12,249 images (1.1%), with data augmentation applied. INSPECT had 192,675 images (16.6%) and MedMD had 152,675 (13.2%), both also randomly dropped. MedMNIST included 300,000 images (25.9%) and underwent random drop as well.

C. Self-supervised Pre-training on massive Medical Data



Train From Scratch ImageNet Pretrain Self-supervised Pre-training

FIGURE 3. LOSS COMPARE

FIGURE3 shows three graphs provide a comparative analysis of different training methodologies for machine learning models, focusing on their impact on loss convergence and accuracy.

1.From Scratch Training:

Loss Convergence: Initially starts at a higher loss value and gradually decreases, but the final convergence is inconsistent, rising to 0.3. The overall accuracy only achieves around 90%, indicating limited learning capability from scratch.

2.ImageNet Pre-trained Model:

Leveraging pre-trained models on large datasets like ImageNet provides weak foundation for medical domain. It begins at a lower initial loss but the trend is similar to train-from-scratch, it converges steadily to 0.12 and reaches 90.3%, showing a small improvement over scratch training.

3.Self-supervised Pre-training on Medical Dataset:

Rapid and smooth convergence to 0.08, indicating efficient learning and it achieves 94% on accuracy. With additional techniques like distillation and data augmentation, it further increases to 95%.

This approach highlights the impact of domain-specific pre-training, where the model benefits from relevant features learned in a self-supervised manner. The initial rapid convergence accelerates the overall training process and enhances model performance.

The experiments underscore the importance of a robust initial model. Self-supervised pre-training on domain-specific data provides significant advantages, improving both convergence speed and final accuracy. Combining this with techniques like distillation and data augmentation maximizes performance, achieving the highest accuracy among the tested methods. This highlights the value of tailored pre-training strategies in achieving superior model performance.

D. Ablation Study

TABLE I. ABLATION STUDY ON SWIN-B

Model	Method	NCT-CRC-HE-100K Top-1 Acc. (%)	Accuracy Improvement (%)
Swin-B	None	93.3	0
	+ Pretraining on large Medical Dataset	94	+0.5
	+ Data Augmentation	94.2	+0.2
	+ Model Distillation	94.8	+0.7
	+ Alpha Tuning	95.0	+0.2

TableI is our ablation experiment, using Swin-B, we observed the most significant improvements due to its strong baseline performance and balanced resource requirements. We selected Swin-B due to limited GPU resources, and it proved to be the most responsive to various enhancements.

The largest accuracy gain (+1.7%) came from pretraining on a large medical dataset, which helped the model learn generalized features applicable across different domains. Model distillation from a larger teacher model yielded a +0.7% improvement, as knowledge transfer further refined the model's understanding. Pretrain on large medical datasets also helps a lot, contributed +0.6%, it shows that in medical area, there is a big data gap compare to normal daily data. Data augmentation techniques contributed moderately (+0.2%), enhancing generalization by exposing the model to more diverse data. Finally, alpha tuning had a smaller impact (+0.2%), fine-tuning the model’s learning process but with less effect than pretraining or distillation.

Overall, the methods involving knowledge transfer (pretraining and distillation) had the most significant effects, while techniques like augmentation and tuning, though beneficial, provided more incremental improvements.

E. Results and Analysis

TABLE II. COMPARE REGULAR TRAINED NCT-CRC-HE-100K MODELS AND OUR MOTHD

Method	Image Size	Params	FLOPs	NCT-CRC-HE-100K Top-1 Acc. (%)
RegNetY-4G	224^2	21M	4.0G	91.7
RegNetY-8G	224^2	39M	8.0G	91.7
RegNetY-16G	224^2	84M	16.0G	92.9
EffNet-B3	300^2	12M	1.8G	91.6
EffNet-B4	380^2	19M	4.2G	92.9
EffNet-B5	456^2	30M	9.9G	93.6
EffNet-B6	528^2	43M	19.0G	94.0
EffNet-B7	600^2	66M	37.0G	94.3
ViT-B/16	384^2	86M	55.4G	94.9
ViT-L/16	384^2	307M	190.7G	86.5
DeiT-S	224^2	22M	4.6G	89.8
DeiT-B	224^2	86M	17.6G	91.8
Swin-T	224^2	29M	4.5G	91.3
Swin-S	224^2	50M	8.7G	93.0
Swin-B	224^2	88M	15.4G	93.3
Swin-T（ours）	224^2	29M	4.5G	92.6(+1.3)
Swin-S（ours）	224^2	50M	8.7G	94.2(+1.1)
Swin-B（ours）	224^2	88M	15.4G	95.0(+1.7)

As shown in TABLEII, our experiments evaluate the performance of various models on the NCT-CRC-HE-100K dataset for colorectal cancer tissue classification, with a particular focus on the Swin-Transformer V2 architecture. We compared our Swin-Transformer V2 models (Tiny, Small, and Base) against other widely used classification models such as RegNet [10], EfficientNet [11], ViT [12], and DeiT [13]. The table above summarizes the results, highlighting several key performance metrics: the number of parameters, FLOPs and the top-1 accuracy on the NCT-CRC-HE-100K dataset.

The results highlight the effectiveness of the Swin-Transformer V2 architecture across different model sizes. The Swin-Tiny variant offers an optimal balance between computational cost and accuracy, making it suitable for scenarios where both speed and accuracy are required. On the other hand, the Swin-Base variant is more computationally expensive but achieves the highest accuracy, making it suitable for tasks where accuracy is paramount.

In comparison to other models like EfficientNet and RegNet, Swin-Transformer V2 demonstrates superior performance, particularly in the smaller and mid-range models. EfficientNet models, while highly optimized for FLOPs and throughput, are outperformed by Swin-Transformer V2 in terms of accuracy. Similarly, ViT models, despite their high accuracy in some configurations, suffer from lower throughput and higher computational demands, making Swin-Transformer a more efficient alternative for this specific task.

Our modified Swin-Tiny model, with 29M parameters and 4.5G FLOPs, achieved a 92.6% top-1 accuracy, an improvement of 1.3% over the original Swin-T model. Despite having one of the highest throughputs (755.2 images/s), it still maintains a high classification accuracy, making it a strong contender for tasks that require both speed and accuracy. The balance of low computational cost and high performance shows that even the smallest version of Swin-Transformer V2 can outperform several models like EfficientNet-B3 and RegNetY-4G in terms of accuracy.

V. Conclusions

In this work, we presented the first application of Swin-Transformer V2 on the NCT-CRC-HE-100K dataset for colorectal cancer tissue classification. We extended previous Transformer-based models, particularly MoBY, by adapting self-supervised learning frameworks and integrating multi-scale medical imaging datasets to enhance performance in histopathological image analysis.

Through the use of advanced self-supervised pretraining, we leveraged large-scale tumor-related medical datasets, such as Digest Path and TCGA, alongside extensive data augmentation techniques (random cropping, magnification, etc.). These augmentations captured variable tumor sizes and scales, which are critical in medical image analysis but often underrepresented in existing approaches. Furthermore, we trained the models on an A100 8-GPU system for a week to ensure the highest performance levels.

We also introduced a progressive layer-wise distillation method using a teacher-student framework to optimize smaller models without sacrificing performance. By incorporating a learnable parameter (α) for each layer, we allowed the network to dynamically adjust the focus of its learning, improving efficiency across different depths of the student model.

Our experiments demonstrated substantial improvements with our fine-tuned Swin-Transformer V2 models, specifically Swin-Tiny, Swin-Small, and Swin-Base. The results showed top-1 accuracies of 92.6%, 94.2%, and 95.0%, respectively, on the NCT-CRC-HE-100K dataset, all of which outperformed the previous state-of-the-art models. Particularly, our modified Swin-Base model achieved a 1.7% increase in accuracy over the baseline model, demonstrating the effectiveness of our refinements.

This work not only pushes the boundaries of medical image classification but also highlights the versatility and scalability of Swin-Transformer V2 across different model sizes. Our future work will further explore the application of this architecture to other medical datasets and tasks, with a focus on improving efficiency and accuracy in real-world clinical settings.

References

[1] Liu, Ze, et al. "Swin transformer: Hierarchical vision transformer using shifted windows." Proceedings of the IEEE/CVF international conference on computer vision. 2021.

[2] Liu, Ze, et al. "Swin transformer v2: Scaling up capacity and resolution." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022.

[3] Xie, Zhenda, et al. "Self-supervised learning with swin transformers." arXiv preprint arXiv:2105.04553 (2021).

[4] Grill, Jean-Bastien, et al. "Bootstrap your own latent-a new approach to self-supervised learning." Advances in neural information processing systems 33 (2020): 21271-21284.

[5] Cao, Hu, et al. "Swin-unet: Unet-like pure transformer for medical image segmentation." European conference on computer vision. Cham: Springer Nature Switzerland, 2022.

[6] Ghazouani, Fethi, Pierre Vera, and Su Ruan. "Efficient brain tumor segmentation using Swin transformer and enhanced local self-attention." International Journal of Computer Assisted Radiology and Surgery 19.2 (2024): 273-281.

[7] Hüseyin, Ü. Z. E. N., et al. "Swin transformer-based fork architecture for automated breast tumor classification." Expert Systems with Applications 256 (2024): 125009.

[8] Al. Shawesh, Radwan, and Yi Xiang Chen. "Enhancing histopathological colorectal cancer image classification by using convolutional neural network." MedRxiv (2021): 2021-03.

[9] Van Erven, Tim, and Peter Harremos. "Rényi divergence and Kullback-Leibler divergence." IEEE Transactions on Information Theory 60.7 (2014): 3797-3820.

[10] Radosavovic, Ilija, et al. "Designing network design spaces." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020.

[11] Koonce, Brett, and Brett Koonce. "EfficientNet." Convolutional neural networks with swift for Tensorflow: image recognition and dataset categorization (2021): 109-123.

[12] Dosovitskiy, Alexey. "An image is worth 16x16 words: Transformers for image recognition at scale." arXiv preprint arXiv:2010.11929 (2020).

[13] Touvron, Hugo, et al. "Training data-efficient image transformers & distillation through attention." International conference on machine learning. PMLR, 2021.