Evaluation of Different Computer Vision Classification Techniques on Medical Datasets

Tsung-Hsiang Ma
t8ma@ucsd.edu

Sharvari Deshmukh
shdeshmukh@ucsd.edu

Andy Nguyen
ahn063@ucsd.edu

Abstract

In the evolving field of medical imaging analysis, Convolutional Neural Networks (CNNs) and Transformer models have played an essential role in image classification, offering promising avenues for enhancing diagnostic accuracy and patient care. This paper presents a comprehensive evaluation of CNNs and Transformer based-based models applied to medical datasets, aiming to highlight their respective strengths, limitations, and applicability. Through a series of experiments conducted on two datasets, MHIST and LC25000, we assess these models based on accuracy, interpretability and robustness to variations in medical datasets. We explore different techniques to enhance their ability to learn through transfer learning, augmentation techniques, and architectural changes. Our findings reveal that while CNNs exhibit exceptional performance in capture spatial hierarchies and local patterns within medical images, Transformer models leverage their self-attention mechanism, demonstrate a great capability in handling long-range dependencies and complex pattern. These two differing perspectives, capturing local and global contexts of a medical image have proved to be useful in medical image analysis. The paper concludes with insights into the future trajectory of model developmenet in medical imaging, emphasizing the need for models that not only excel in performance metrics but also align with clinical requirements.

1. Introduction

Among the various architectures made available, we decided to baseline our results with two prominent CNN architectures, VGG-16 and ResNet, alongside the Vision Transformer (ViT). These models were selected for their diverse architectural principals and have proven effectiveness in image classification. We explore and implemented different fine-grained models to enhance the ability for our model to discern and classify medical images by using API-net, and SWIN-T across these datasets. These fine-grained models are specifically designed to pay attention to subtle details within the data. By fine-tuning VGG-16, ResNet, ViT, API-net, and SWIN-T on various medical datasets, we aim to exploit the detailed feature extraction capability of both CNNs and Transformers. This fine-grained analysis allows for nuanced evaluation of each model's performance.

1.1. Motivation

The challenge in medical image classification lies in accurately identifying various pathological and anatomical structures within complex medical images. This task is pivotal for disease detection. However, due to the inherent variability in medical images, coupled with the need for high precision, presents a unique set of challenges that demand robust and adaptable solutions. This is why we felt CNNs and Transformers were the architectures to focus on in these series of experiments.

2. Materials and Methods

2.1. Datasets

Our project involves utilizing multiple histopathology image datasets for the development and evaluation of machine learning models for classification tasks. The first dataset, MHIST[7], consists of 3,152 fixed-size images of colorectal polyps with gold-standard labels determined by seven board-certified gastrointestinal pathologists, along with annotator agreement levels.The LC25000[1] dataset comprises 25,000 color images across five classes, including colon adenocarcinoma, lung adenocarcinoma, and benign tissue, each containing 5,000 images. Leveraging both of these datasets, we are able to evaluate the effectiveness of each model and their ability to learn based on these differing dataset sizes.

2.2. Model Architecture

2.2.1 CNN

Convolutional Neural Network (CNN), which is a specialized type of artificial neural network commonly used in tasks related to image recognition and classification. CNNs are inspired by the organization of the animal visual cortex,

particularly the receptive fields of neurons.
In this paper, VGG16 was applied as the baseline model, and ResNet was introduced as advanced CNN model.

VGG[4]: The VGG architecture was proposed by the Visual Geometry Group at the University of Oxford. It is characterized by its simplicity and uniformity. VGG typically consists of a series of convolutional layers with small 3x3 filters, followed by max-pooling layers to reduce spatial dimensions. The key idea behind VGG is to stack more layers with smaller filters to learn complex patterns and features. Despite its simplicity, VGG achieved remarkable performance in image classification tasks and served as a baseline for subsequent architectures.

ResNet[3]: ResNet is a revolutionary architecture introduced by Microsoft Research. It addresses the problem of vanishing gradients in deep neural networks by introducing skip connections or shortcuts, allowing the network to learn residual functions instead of directly learning desired mappings. This architecture enables training of very deep networks (up to hundreds of layers) without suffering from degradation in performance. ResNet has significantly pushed the boundaries of deep learning performance and has been widely adopted in various computer vision tasks, winning numerous competitions and benchmarks.

Both VGG and ResNet have made significant contributions to the field of deep learning and are widely used as building blocks in many state-of-the-art neural network architectures.

2.2.2 Transformer

Using the Transformer architecture, we are hoping to add a new perspective to medical image analysis that CNNs cannot provide. Seeing that CNNs are "short-sighted" they only capture local spatial hierarchies between images. This oversight is where the Transformer architecture, with its roots deeply embedded in natural language processing, shines. Originally designed to discern the nuanced relationships between words in a sentence, where the significance of a word can pivot dramatically based on its context, Transformers apply a similar principle to image analysis. By treating image pixels or regions as analogous to words in a text, Transformers are adept at identifying relationships and dependencies across distant parts of an image. This capability is especially crucial in medical imaging, where the presence or absence of disease can hinge on subtle, distributed features across the image.

ViT[2] The Vision Transformer adapts the Transformer architecture for image classification by treating images as a sequence of patches. This approach allows ViT to capture complex patterns and relationships within the image, irrespective of the spatial distance between features. Its global receptive field contrasts with the local views of CNNs. ViT's self-attention mechanism allows it to flexibly adjust to various input dimensions without a redesign of the network architecture. This flexibility allows ViT to handle the diversity and variability within the image dataset.

SWIN-T[5] SWIN Transformer also adapts the Transformer model and introduces a hierarchical structure that mirrors the biases found in CNNs. This hybrid approach allows SWIN-T to process images by dividing them into smaller, non-overlapping patches, similar to the ViT. The SWIN-T introduces shifted windows which brings greater efficiency by limiting self-attention computation to non-overlapping local windows while also allowing for cross-window connection. SWIN-T combines the depth and complexity of global context modeling inherent in Transformers with the efficient, structured approach of CNNs, offering a new architecture for medical image analysis.

2.3. Improvement Technique

2.3.1 Transfer Learning

Transfer learning is a technique in machine learning where knowledge gained from solving one problem is applied to a different but related problem. In the context of neural networks, transfer learning involves using a pre-trained model on one task and fine-tuning it on a new task, rather than training a new model from scratch.

The main idea behind transfer learning is that models trained on large datasets for tasks such as image classification or natural language processing have learned useful features and representations that can be generalized to other tasks. By leveraging pre-trained models, transfer learning can significantly reduce the amount of labeled data and computational resources required to train a new model, while often achieving better performance than training from scratch, especially when the new task has a limited amount of data available.

2.3.2 Data Augmentation

Image augmentation is a technique used to artificially increase the size of a dataset for training machine learning models, particularly in computer vision tasks such as image classification, object detection, and segmentation. It involves applying a variety of transformations to existing images in the dataset, creating new variations while preserving their semantic meaning.

Image augmentation techniques applied in this paper include:

1. Random Rotation: Rotating the image by a certain angle (-45° to $+45^\circ$).

2. Vertical Flipping: Mirroring the image vertically with probability of 0.5.

3. Horizontal Flipping: Mirroring the image horizontally with probability of 0.5.

By applying these transformations to the training images, image augmentation helps improve the generalization capability of machine learning models by exposing them to a wider variety of data variations. This can help prevent overfitting and improve the model's robustness to variations in real-world data.

Image augmentation is a crucial preprocessing step in training deep learning models, especially when the dataset is limited in size or diversity. It allows the model to learn more robust and invariant representations of the underlying data distribution, ultimately leading to better performance on unseen data.

2.3.3 Fine-Grained Model

A fine-grained model refers to a type of machine learning model, typically used in computer vision tasks, that is designed to distinguish between categories that are closely related or similar in appearance. Unlike traditional image classification tasks where the goal is to differentiate between broad categories (e.g., different types of animals), fine-grained models focus on distinguishing between subcategories within a particular class (e.g., different species of birds or breeds of dogs).

Fine-grained models require a high level of detail and precision in their predictions, often needing to recognize subtle differences in features such as shape, texture, or color. These models typically require large amounts of labeled data for training, as well as sophisticated architectures capable of capturing fine-grained details.

In this paper, API Net applied as fine-grained model.

API Net: Drawing inspiration from this human capability of discerning contrasting cues through the comparison of pairs of images, Attentive Pairwise Interaction Network (API-Net) [9] was created. The model architecture is shown in figure 1. This model progressively distinguishes between a pair of fine-grained images by fostering interaction. Initially, API-Net learns a shared feature vector to capture semantic distinctions within the input pair. Subsequently, it juxtaposes this shared vector with individual vectors to generate gating mechanisms for each image input. These distinct gating vectors encapsulate mutual context regarding semantic disparities, enabling API-Net to adeptly discern contrasting cues through the pairwise interaction between the two images. Moreover, we train API-Net in an end-to-end fashion, incorporating a score ranking regularization method that enhances generalization by considering feature priorities.

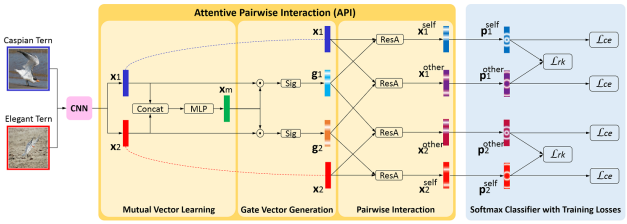

Figure 1. API Net Architecture[9]

2.4. Hybrid Approach

Swin-Transformer The main design element of SWIN-T is its shifted windows technique. This is where the window partition between consecutive self-attention layers form a connection of the windows of the preceding layer, providing connections between them that enhance modeling power.

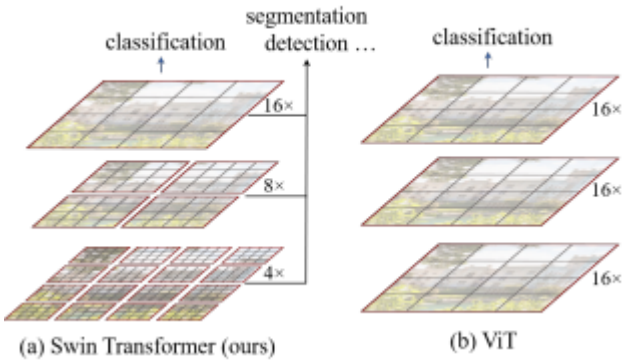

Figure 2. The Swin Transformer, merges image patches which is shown in gray in deeper layer and has linear computational complexity relative to input image size due to computation of self-attention within each local window (red). This is different from ViT which produces feature maps of a single low resolution and has quadratic computation complexity relative to input size.[5]

2.5. Experiment

Our study evaluates the performance of the various architectures on medical imaging datasets. We divide our exploration into two primary categories: CNNs and Transformer models. For each of the architectures explained from before (VGG-16, ResNet, API-Net, ViT, and SWIN-T), we used two optimizers, Adam and stochastic gradient descent. Within these optimizers we changed the parameters of learning rate, betas, weight decay, and momentum. The learning rate influences the model's ability to learn, betas are specific to Adam and control the running averages of gradient and its square, weight decay is used as a regularization technique to penalize larger weights, and momentum is used with SGD to accelerate gradients in the right direction. Then we used transfer learning[8] and augmen-

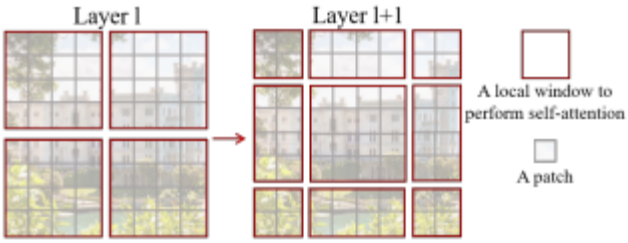

Figure 3. This figure shows the shifted window approach for computing self-attention. On the left layer (l), a regular window window partitioning scheme is shown and self attention is computed within each window. On the right layer (l+1), the window partitioning is shifted which results in new windows. The self-attention computation in the new windows provides connections of the left layer.[5]

tation techniques on all variation to enhance robustness and generalization of the models.

2.6. Evaluation Metrics

To assess our networks, we used these key evaluation metrics: Precision, Recall, Test Accuracy, and F1 scores. Precision is important in minimizing false positives, recall is critical for ensuring no condition is overlooked, as missing a true positive can mean a missed diagnosis with potentially life-threatening implications, test accuracy gives us a snapshot of the model's overall performance, and F1 score ensures our model is both precise and sensitive.

3. Results

3.1. CNN

The performance of the models on the MHIST and LC25000 datasets is compared in Figure 4 and Figure 5. As evident from the results, transfer learning leads to significant improvements in CNN performance across all architectures. The models exhibit an approximately 10% enhancement following the application of transfer learning. Additionally, Image Augmentation enhances model performance by approximately 3%. Notably, the performance of the ResNet18 model improves by 10% post-image augmentation, demonstrating that this technique increases data variance and enhances model robustness. Moreover, the fine-grained model, when combined with transfer learning and image augmentation, exhibits the most superior performance across all evaluation metrics. This suggests that the fine-grained architecture, along with transfer learning and image augmentation, significantly enhances CNN model performance on small datasets. From Figure 5, we observe a similar trend of improvements in model performance on the large dataset. Transfer learning boosts the model by approximately 2%, while image augmentation enhances it by about 0.5%. Furthermore, the combination of the fine-grained model with transfer learning and image augmentation demonstrates superior performance across all evaluation metrics. This underscores the significant enhancement of CNN model performance on large datasets through the integration of fine-grained architecture, transfer learning, and image augmentation.

Model	VGG16		ResNet18		VGG16		ResNet18		API-Net	
Pretrained	X		X		O		O		O	
Augmentation	X	O	X	O	X	O	X	O	X	O
Accuracy	71.14	73.39	67.35	77.07	84.44	85.36	83.52	86.18	82.19	**86.90**
Precision	56.75	62.43	47.96	60.00	78.47	78.94	75.27	79.58	76.49	**81.87**
F1-score	51.39	60.00	40.83	46.67	76.94	74.44	68.06	73.06	78.61	**80.28**
Recall	63.36	65.06	58.10	84.00	80.06	84.01	84.19	87.38	74.47	**83.53**

Figure 4. CNN Architecture Evaluations on MHIST

Model	VGG16		ResNet18		VGG16		ResNet18		API-Net	
Pretrained	X		X		O		O		O	
Augmentation	X	O	X	O	X	O	X	O	X	O
Accuracy	96.94	97.02	96.92	96.98	97.90	99.42	99.60	99.76	99.71	**99.80**
Precision	96.94	97.02	96.92	96.98	97.90	99.42	99.60	99.76	99.71	**99.80**
F1-score	96.94	97.02	96.92	96.98	97.90	99.42	99.60	99.76	99.71	**99.80**
Recall	96.94	97.02	96.92	96.98	97.90	99.42	99.60	99.76	99.71	**99.80**

Figure 5. CNN Architecture Evaluations on LC25000

Model	ViT		ViT		Swin-T		Swin-T	
Pretrained	X		O		X		O	
Augmentation	X	O	X	O	X	O	X	O
Accuracy	62.2	63.13	81.88	79.30	62.20	62.20	**83.08**	80.08
Precision	48.08	49.89	73.89	74.07	76.73	76.73	**82.98**	80.01
F1-score	64.94	66.32	76.18	70.61	48.89	48.89	**82.94**	79.37
Recall	100.0	62.41	78.61	67.46	63.15	63.15	**82.91**	80.04

Figure 6. Transformer Architecture Evaluations on MHIST

Model	ViT		ViT		Swin-T		Swin-T	
Pretrained	X		O		X		O	
Augmentation	X	O	X	O	X	O	X	O
Accuracy	65.48	70.36	99.00	97.34	92.75	92.22	**99.94**	99.28
Precision	66.50	68.54	99.80	99.90	90.83	92.48	**99.94**	99.28
F1-score	68.12	73.62	99.9	99.60	90.46	92.34	**99.94**	99.28
Recall	69.82	79.52	1.00	99.30	90.10	92.36	**99.94**	99.28

Figure 7. Transformer Architecture Evaluations on LC25000

3.2. Transformers

As shown in Figure 6 and 7, we can see that there is significant performance improvements for both architectures when data augmentation and pretraining strategies were employed. These enhancements were important in adapting models, originally designed for general image analysis, to the domain of medical imaging. We can see that training on the LC25000 dataset had better metrics overall because of the larger dataset size, as was the case for the CNN architectures. When using data augmentation, ViT had about a 1-5% boost in performance for MHIST and pretraining gave it about 20%. SWIN-T had about a 2% boost for data augmentation and pretraining gave it about a 8% increase. This shows that our transformer architectures require much more data in order for it to be effective compared to the CNNs.

4. Conclusion

In our evalutaion of various neural network architectures applied to medical imaging datasets, we observed distinct performance patterns that highlight the strengths and limitations of CNNs and transformer models.

The performance of CNN models on MHIST and LC25000 datasets is analyzed in Figure 4 and Figure 5. Transfer learning notably improves CNN performance across various architectures, with models seeing around a 10% enhancement. Additionally, Image Augmentation boosts performance by approximately 3%, particularly evident in ResNet18 with a 10% improvement, indicating increased data variance and model robustness. Combining the fine-grained model with transfer learning and image augmentation yields superior performance across all metrics, highlighting the significant enhancement of CNN model performance on both small and large datasets through these integrated techniques.

ViT's reliance on large amounts of data to effectively learn and generalize became apparent, as it struggled to reach the performance benchmarks set by its CNN counterparts without extensive pretraining or data augmentation. This difficulty underscores the inherent trade-off with Transformer models: while they possess the capability to model complex, long-range dependencies within data, their performance is heavily dependent on the volume and diversity of the training data available.

Conversely, CNN architectures, with their ability to capture local spatial hierarchies and patterns, exhibited better performance on the MHIST dataset, which is comparatively smaller than the LC2500 dataset. This advantage can be attributed to the CNNs' efficiency in learning from limited data, a result of their architectural biases towards spatial locality, which seems particularly suited to the fine-grained analysis required by such datasets.

However, when evaluating models on the LC25000 dataset, SWIN-T was the best performer out of all of the CNN networks and ViT. The success can be attributed to its hybrid design which combines the best aspects of CNNs and Transformers. Our findings suggest that there is an importance of dataset characteristics in determining the most effective neural network architecture for medical image analysis. While CNNs continue to excel in scenarios with limited data, SWIN-T showed us that with a large enough dataset, it can capture intricate patterns of medical images and surpass CNN limitations.

5. Future work

A critical next step involves scaling our experiments to encompass significantly larger medical datasets, potentially exceeding 100,000 samples. A larger dataset would provide a more robust evaluation for our models. We predict that a larger dataset will benefit transformer-based models like ViT and SWIN-T which have shown the capacity for improved performanced with increased data availability. We also would like to explore meta-learning which enhances a model's ability to generalize from a limited number of examples. This technique would be really beneficial in medical imaging, where acquiring large annotated datasets can be challenging. In our experiments, we did not leverage domain-specific pretrained models which can significantly boost performance. Since pretraining on non-related images significantly boosted our model's performance metrics, we predict that pretraining the models on past medical images would boost the model's performance even more. We would like to explore more hybrid architectures that leverage the strengths of both CNNs and transformers, specifically the U-Net[6] structure. This hybrid model would combine CNN's proficiency in capturing spatial hierarchies with the transformer's ability to model long-range dependencies. We see that SWIN-T have leveraged this idea in the past, and we predict with more hybrid architectures they will reach similar performance metrics.

6. Contribution

- Tsung-Hsiang Ma: CNN experiments, paper
- Sharvari Deshmukh: Dataset and Swin-T experiments
- Andy Nguyen: Evaluation metrics, paper, ViT experiments

References

[1] Andrew A Borkowski, Marilyn M Bui, L Brannon Thomas, Catherine P Wilson, Lauren A DeLand, and Stephen M Mastorides. Lung and colon cancer histopathological image dataset (lc25000). *arXiv preprint arXiv:1912.12142*, 2019.

[2] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner,

Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2021.

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[4] Andrew Zisserman Karen Simonyan. Very deep convolutional networks for large-scale image recognition. *International Conference on Learning Representations*, 2015.

[5] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

[6] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015.

[7] Jerry Wei, Arief Suriawinata, Bing Ren, Xiaoying Liu, Mikhail Lisovsky, Louis Vaickus, Charles Brown, Michael Baker, Naofumi Tomita, Lorenzo Torresani, et al. A petri dish for histopathology image analysis. In *International Conference on Artificial Intelligence in Medicine*, pages 11–24. Springer, 2021.

[8] Ross Wightman. Pytorch image models. `https://github.com/rwightman/pytorch-image-models`, 2019.

[9] Peiqin Zhuang, Yali Wang, and Yu Qiao. Learning attentive pairwise interaction for fine-grained classification. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(07):13130–13137, Apr 2020.