[RE] A Comprehensive Analysis of ViT Fine-tuning: Exploring resolution, dataset size, domain shift, and ablation study

Seonjin, Hwang
passiona2z@seoultech.ac.kr

Jewoo, Kwak
kwak2963@gmail.com

Jiwon, Park
oolong0205@gmail.com

Abstract

While the Transformer architecture has become the de-facto standard for natural language processing tasks, its successful application to computer vision is exemplified by Vision Transformer (ViT). ViT, when pretrained on large-scale datasets and applied to medium- or small-sized recognition benchmarks, has surpassed the performance of CNN-based image classification tasks. In our work, we aim to reproduce the performance of ViT pretrained on ImageNet and transferred to other datasets (e.g. CIFAR-10, CIFAR-100), as demonstrated in ViT paper. Furthermore, we conducted a comprehensive analysis of ViT models by exploring experiments not explicitly addressed in the original ViT paper, such as adjusting resolution and dataset size, and applying the model to different domains. Additionally, we performed an ablation study of ViT to gain a deeper understanding of the model.[1]

1 Introduction

1.1 Transformer

The Transformer architecture, introduced by [2], has revolutionized the field of natural language processing (NLP). Unlike recurrent neural networks (RNNs) that process data sequentially, Transformers leverage a mechanism called self-attention to process data in parallel, allowing for more efficient training and superior performance on a variety of tasks. The core component of the Transformer is the multi-headed self-attention mechanism, which enables the model to focus on different parts of the input sequence simultaneously and capture long-range dependencies more effectively.

While Transformers have become the de-facto standard in NLP, their application to computer vision tasks has been less straightforward due to the inherent differences between text and images. However, the Vision Transformer (ViT) has successfully adapted the Transformer architecture to the vision domain, achieving state-of-the-art results on image classification tasks. The key innovation in ViT is the way it processes images. Unlike convolutional neural networks (CNNs) that operate directly on the raw pixel values of the image using convolutional filters, ViT divides the input image into a sequence of smaller, non-overlapping patches. Each patch is treated as a token, similar to a word in NLP.

1.2 Inductive Bias

Inductive bias refers to the ability to predict outputs for unseen inputs. In other words, it can be seen as additional assumptions made to improve generalization performance. Despite ViT's innovative approach, initial experiments showed that ViT trained on mid-sized datasets like ImageNet exhibited slightly lower accuracy compared to ResNet models. This was attributed to the lack of inductive biases inherent in CNNs, such as translation equivariance and locality. Inductive biases in CNNs, such as the assumption that nearby pixels are more related (exploited through small kernels like 3x3), are not naturally present in the Transformer architecture. Transformers, by design, do not have any inherent understanding of the order or position of tokens in a sequence. This is because the self-attention mechanism treats input tokens independently, considering their positions only through the learned attention weights. In NLP, this lack of positional information is mitigated by adding positional embeddings to the input tokens, enabling the model to understand the sequence order.

[1] Our code are available at https://github.com/passiona2z/RE_ViT

In the context of ViT, where the input is an image divided into patches, positional embeddings play a similar role. It encodes spatial Information. When an image is divided into patches, the sequence of patches loses the explicit spatial arrangement of the image. Positional embeddings are added to each patch embedding to encode its position in the original image grid. This helps the model retain information about the relative positions of patches, which is crucial for understanding the spatial structure of the image. Most of all, positional embeddings introduce a form of inductive bias by explicitly encoding the positional information. This bias helps the model understand the spatial relationships between patches, somewhat analogous to how convolutional kernels in CNNs capture local spatial patterns through their structure.

1.3 Proposed Method

The ViT paper argued that large-scale training can compensate for the lack of inductive biases inherent in the transformer architecture. When pre-trained on large-scale datasets and fine-tuned on medium- or small-sized image recognition benchmarks, ViT models have demonstrated superior performance compared to traditional CNN-based approaches.

As part of the replication task, we tested the ViT model pre-trained on ImageNet-21K and compared its performance with the results reported in the original ViT paper. In addition to reproducing the original results, we explored new avenues of ViT research by conducting experiments not previously addressed. For instance, the ViT paper suggested that it is often beneficial to fine-tune at higher resolution than pre-training, but did not provide any supporting experiments. We investigated this claim by examining whether fine-tuning at a higher resolution leads to improved performance. Additionally, the paper stated that ViT attains excellent results when pre-trained at sufficient scale and transferred to tasks with fewer datapoints. We tested this assertion by further reducing the number of data points during fine-tuning.

Furthermore, we explored whether the ViT model remains competitive when fine-tuned on domain images not included in the pre-training dataset. Also, we conducted a Ablation study using grid search on learning rate and optimizer to gain a deeper understanding of the ViT model.

2 ViT Architecture

The ViT architecture deviates from the traditional CNN approach of processing images. Instead, it treats an image as a sequence of patches, analogous to tokens in NLP. Here's a step-by-step breakdown of the ViT structure:

1. Patch Embedding:
    - Image to Patches: The input image is divided into fixed-size patches, for instance, a 224x224 image might be divided into 16x16 patches, resulting in a sequence of 196 patches (14x14 grid).
    - Linear Projection: Each patch is flattened into a 1D vector and passed through a linear projection layer, converting it into a latent vector of a specified dimension.
2. Positional Embeddings:
    - Spatial Encoding: Since Transformers lack an inherent understanding of the spatial relationships between patches, positional embeddings are added to each patch embedding. These embeddings encode the position of each patch in the original image, allowing the model to maintain spatial context.
3. Transformer Encoder:
    - Multi-Headed Self-Attention: The sequence of patch embeddings, augmented with positional embeddings, is fed into a Transformer encoder. The encoder consists of multiple layers of multi-headed self-attention and feed-forward neural networks. Self-attention allows the model to learn relationships between different patches, capturing both local and global context.
    - Layer Normalization and MLP: Each layer in the encoder includes normalization and multi-layer perceptron (MLP) blocks to enhance learning dynamics.

3 Implementation Detail

We implemented our experiments using PyTorch and Hugging Face frameworks and closely followed the implementation details outlined in the ViT paper to the best of our ability. However, due to GPU resource constraints in our experimental environment, we had to reduce the batch size (from 512 to 128). Additionally, for practical reasons, we adjusted the number of steps (CIFAR dataset: from 10,000 to 2,500) and used a fixed learning rate (0.01) without tuning. The parameter study and ablation study were conducted in individual experimental environments. The specific details of each environment will be provided in the respective sections.

4 Replication Experiments

We conduct experiments related to reproduction and further proceed with additional experiments that involve adjusting the resolution and the amount of data, applying the model to different domains, parameter study and ablation study.

4.1 Replication experiments

Due to limited computational resources, we focused on reproducing the supplementary experiments presented in the ViT paper, specifically fine-tuning the ViT-B/16 model (pre-trained on ImageNet-21k) on CIFAR10, CIFAR100, and Flower102 datasets. Our results aligned with the performance reported in the original paper.

	CIFAR-10	CIFAR-100	Oxford Flowers-102
ViT paper	98.95	91.67	99.38
Our experiments	98.48	90.06	98.53

Table 1: Replication experiments (ViT-B/16 pretrained with ImageNet-21k)

4.2 Resolution adjustment experiments

The original ViT paper suggests that fine-tuning at a higher resolution than pre-training may be beneficial, but provides no experimental evidence. We investigated this claim by fine-tuning ViT-B/16 (pretrained with ImageNet-21k) on CIFAR100 and Flower102 datasets, varying the resolution during fine-tuning. Our results indicate that increasing resolution does not lead to a linear improvement in performance. For CIFAR100, the best performance was achieved when fine-tuning at the same resolution as pre-training. For Flower102, a resolution of 384 outperformed 512. This suggests that while fine-tuning at an equal or higher resolution than pre-training may be necessary for optimal performance, higher resolution does not guarantee improved results.

Resolution	512	384	224	128	64
CIFAR-100	90.17	90.06	**91.19**	89.02	47.96
Oxford Flowers-102	98.43	**98.53**	98.04	92.94	20.98

Table 2: Resolution adjustment experiments (ViT-B/16 pretrained with ImageNet-21k)

4.3 Data sampling experiments

While the original ViT paper highlights the benefits of larger pre-training datasets, it does not explore the impact of fine-tuning dataset size on model performance. We investigated this aspect by examining how varying the amount of fine-tuning data affects ViT performance. Our experiments on two datasets revealed a notable sensitivity of ViT models to data scarcity during fine-tuning. Specifically, using only 10% or 25% of the full dataset resulted in a substantial decrease in performance compared to utilizing the complete dataset. This performance degradation was particularly pronounced when the fine-tuning data was extremely limited, such as with 10% of the Flower dataset.

Data Sampling	100%	50%	25%	10%	**diff.**
CIFAR-100	(50000)	(25000)	(12500)	(5000)	
ViT-B/16	90.81	88.11	82.95	64.24	**+26.57**
BiT-50x1	84.85	82.51	77.70	69.04	**+15.81**
Oxford Flowers-102	(7160)	(3584)	(1792)	(716)	
ViT-B/16	98.63	85.39	55.20	27.55	**+71.08**
BiT-50x1	98.62	97.94	97.45	85.29	**+13.33**

Table 3: Fine-Tuning Data Sampling Experiments (ViT-B/16 vs BiT-50x1) : (value) indicates the number of images used for fine-tuning, and diff. represents the performance difference between using 100% and 10% of the dataset.

4.4 Domain shifting experiment

We conducted a fine-tuning experiment using a domain-specific dataset with unique characteristics that distinguish it from general images. Specifically, we aimed to verify whether the Vision Transformer (ViT) outperforms the ResNet model, such as Big Transfer (BiT), when pre-trained on a large dataset like ImageNet-21K, as mentioned in the paper. Following the paper's recommendations, we used a resolution of 384 for fine-tuning. Due to GPU limitations, we set the batch size to 128 and the maximum number of epochs to 7.

Domain	Medical	Horticulture	Military
ViT-B/16	**0.9639**	0.9192	**0.5907**
BiT-50x1	0.9611	**0.9697**	0.4852

Table 4: Domain shifting experiment (acc)

4.4.1 Medical Domain : Brain Tumor MRI Classification

We experimented to evaluate the performance of ViT and BiT models in the medical domain, specifically focusing on the classification of brain tumor MRI images. Our dataset comprised 7,023 MRI images categorized into four distinct classes based on the tumor's location: 'glioma', 'meningioma', 'pituitary', and 'no tumor'.

The results of the experiment revealed that the ViT model outperformed the BiT model by a margin of approximately 0.27%. This finding aligns with the claims presented in the original ViT paper, which asserted that ViT models outperform ResNet models when pre-trained on large-scale datasets such as ImageNet-21K and subsequently fine-tuned on specific downstream tasks.

Our results demonstrate that the advantages of the ViT model are not limited to general image classification tasks but also extend to the specialized domain of medical imaging. This is significant because it suggests that the architectural strengths and pre-training benefits of the ViT model are applicable across diverse domains, including medical diagnostics, which often require precise and accurate image classification capabilities.

4.4.2 Horticulture Domain : Plant Health Status Classification

In the field of horticulture, we carried out experiments using a dataset designed to measure plant health. This dataset comprises 877 images, each classified into one of three categories based on the health status of the plant: 'healthy,' 'mildew,' and 'spots.'

Contrary to our findings in the medical domain experiment, where ViT outperformed BiT, the results in the horticulture experiment showed a different trend. Here, BiT outperformed ViT by a significant margin of approximately 5.05%. This notable difference can be attributed to the impact of data sparsity on the performance of ViT

In the course of our data sampling reproduction experiments, we observed that ViT's performance was highly susceptible to the effects of limited data availability. This susceptibility was particularly evident in domain-specific experiments, such as those in horticulture, where the number of dataset was relatively small. These findings highlight a crucial consideration for the application of ViT: while they excel with large and diverse datasets, their performance may decline in scenarios where data is sparse.

4.4.3 Military Domain : Artificial Weapon Classification

Artificial weapons classification experiments were conducted in the military domain using a dataset consisting of 2,502 images. This task involved binary classification to determine whether an image depicted a weapon or not. The dataset predominantly featured images of pistols for the weapon category and images of scrap metal resembling pistols for the non-weapon category.

In our study, ViT outperformed BiT by a substantial margin of approximately 10.55%, marking the largest gap observed in our domain-specific experiments. However, both models exhibited significantly lower accuracy compared to previous experiments, which can be attributed to the challenging nature of the dataset. The images of non-weapons were often very similar in appearance to the weapons, making it difficult even for the general public to distinguish between them with the naked eye.

These findings suggest that while ViT demonstrated superior performance, the inherent difficulty of distinguishing between very similar images poses a significant challenge for classification tasks. This underscores the need for further

advancements in model accuracy and robustness, particularly in scenarios where visual similarities between classes are subtle and deceptive.

4.5 Parameter Study

We conducted a parameter study on the optimizer and learning rate scheduling using the CIFAR-10 dataset. This experiment was performed on a ViT-B/16 model pretrained on ImageNet. The base parameters for our study included: batch size of 32, weight decay of 0, gradient clipping at a global norm of 1, dropout rate of 0.1, resolution of 224, cosine learning rate scheduling, and the SGD optimizer with a momentum of 0.9. The differences between this parameter study and the fine-tuning described in paper [1] are that paper [1] used a batch size of 512 and a resolution of 384. All other parameters are identical to those in the paper [1].

optimizer	SGD	Adam	Learning rate	Cosine	Linear	Exponential	ExponentialReduceLROnPlateau
	98.72	71.80	scheduling	98.72	98.67	98.72	**98.75**

Table 5: parameter study on the optimizer and learning rate scheduling

4.6 Ablation Study

We ablated positional embedding in the ViT model. After declaring the position embedding, we omitted the step of adding sequential information and created a model that directly passes the image patches through a linear projection to the Transformer encoder. We compared the results between ViT model with adding positional embedding and the Vit model without adding positional embedding.

The ablation study was conducted with the following setup: image size of 32x32, patch size of 4x4, batch size of 128, dropout rate of 0.1, 8 heads, 12 layers, and CIFAR-10 as the training data.

	acc
ViT with adding positional embedding	**84.64**
ViT without adding positional embedding	69.86

Table 6: ablation study on positional embedding

4.6.1 Linear Projection

According to [1] The first layer of the Vision Transformer linearly projects the flattened patches into a lower-dimensional space. Figure 1 shows RGB embedding filters which is lower dimension mapping of linear projection. Through the RGB embedding filter, it indicates that the ViT also learns in a similar manner to CNNs.

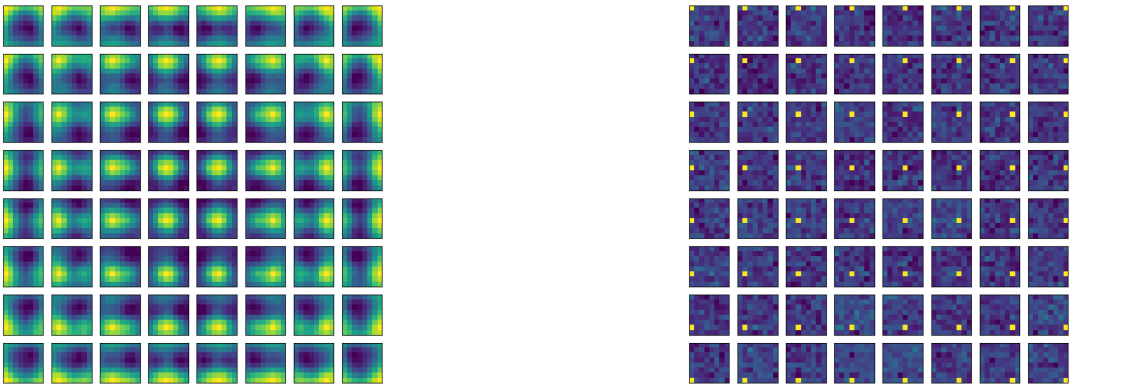
(a) RGB embedding filter with adding positional embedding

(b) RGB embedding filter w/o adding positional embedding

Figure 1: Visualization of RGB embedding filter

4.6.2 Positional embedding

After the linear projection, a learned position embedding is added to the patch representation. Figure 2 shows that the model learns to encode distance within the image through the similarity of position embeddings. Figure 2 (a) shows that closer patches tend to have more similar position embeddings. A row-column structure appears where patches in the same row or column have similar embeddings. As shown in Figure 2(b), the training results indicated that when visualizing the cosine similarity of the positional embedding, the similarity between a given patch and its surrounding patches was learned to be low.



(a) cosine similarity with positional embedding

(b) cosine similarity w/o adding positional embedding

Figure 2: Visualization of cosine similarity

4.6.3 Attention Matrix

Attention matrices in the Vision Transformer (ViT) represent how the model distributes its attention across different parts of the input image when processing it. Specifically, these matrices indicate which regions of the image the model considers most relevant when making decisions. Moreover, in a multi-head self-attention mechanism, different heads might focus on different parts of the image. So visualizing attention matrices for each head and each layer can reveal how different parts of the image are processed at various stages. Figure 3 shows an image in CIFAR-10 and visualizes the 4th rows of attention matrices in the 0-7th heads. Each head exhibits a distinct focus on disparate image regions, indicative of the multifaceted analysis performed by the model.

(a) ViT model with adding positional embedding

(b) ViT model w/o adding positional embedding

Figure 3: Visualization of attention matrices

5 Conclusion

In conclusion, our experiments confirm that while the Vision Transformer model excels in various tasks and domains, its performance is significantly influenced by factors such as data size, resolution, and positional embeddings. These findings underscore the importance of carefully considering these factors when applying ViT models to different tasks, particularly in domain-specific scenarios with unique challenges. Further research and advancements are necessary to address the model's limitations and enhance its robustness and accuracy across diverse applications.

References

[1] Dosovitskiy, Alexey, et al. "An image is worth 16x16 words: Transformers for image recognition at scale." International Conference on Learning Representations (2021).

[2] Vaswani, Ashish, et al. "Attention is all you need." Advances in neural information processing systems 30 (2017).

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

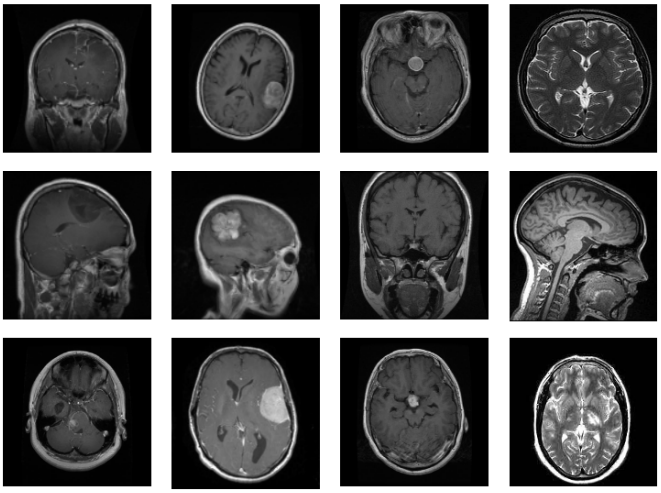
[4] Paul, Sayak, and Pin-Yu Chen. "Vision transformers are robust learners." Proceedings of the AAAI conference on Artificial Intelligence. Vol. 36. No. 2. 2022.

[5] Mao, Xiaofeng, et al. "Towards robust vision transformer." Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition. 2022.

[6] Han, Kai, et al. "A survey on vision transformer." IEEE transactions on pattern analysis and machine intelligence 45.1 (2022): 87-110.

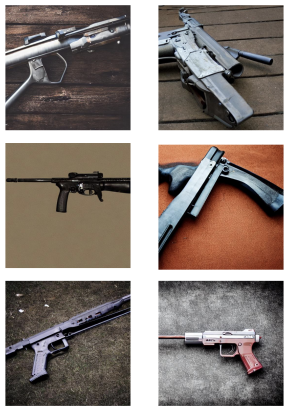
[7] Manzari, Omid Nejati, et al. "MedViT: a robust vision transformer for generalized medical image classification." Computers in Biology and Medicine 157 (2023): 106791.

Appendix



(a) Medical : 'glioma', 'meningioma', 'pituitary', and 'no tumor'

(b) Horticulture : 'healthy', 'mildew', and 'spots'



(c) Military : 'weapon' and 'non-weapon'

Figure 4: Domain shifting dataset samples (labels from Left to Right)