ISSN 1840-4855
e-ISSN 2233-0046

Original scientific article
http://dx.doi.org/10.70102/afts.2026.1835.205

VISION TRANSFORMER USING FRACTIONAL GRADIENT ATTENTION TOOLS FOR ROBUST IMAGE CLASSIFICATION

S. Shalini[1], P.S. Eliahim Jeevaraj[2*]

[1]Research Scholar, Department of Computer Science, Bishop Heber College, (Affiliated to Bharathidasan University), Tiruchirappalli, Tamil Nadu, India.
e-mail: prof.shalinisherlin@gmail.com,
orcid: https://orcid.org/0009-0009-5792-6158
[2]Assistant Professor, Department of Computer Science, Bishop Heber College, (Affiliated to Bharathidasan University), Tiruchirappalli, Tamil Nadu, India.*
e-mail: eliahimps.cs@bhc.edu.in,
orcid: https://orcid.org/0000-0002-7379-365X

Received: January 02, 2026; Revised: February 17, 2026; Accepted: April 07, 2026; Published: May 29, 2026

SUMMARY

Vision Transformers (ViTs) have attained very promising results in the field of computer vision, but those models continue to face several critical issues such as gradient saturation and poor generalization on smaller datasets. The current attention mechanisms are inefficient to resolve issues by the leading to ineffective feature extraction and an incompetent optimization. To overcome these drawbacks, a new Fractional Gradient Attention (FGA) mechanism is proposed to ViTs with the idea of redefining gradient propagation during training with the help of a fractional calculus. The proposed approach modifies the backpropagation by employing a fractional-order derivative as the new method increases the sensitivity of the model to long-range dependencies and increases the stability of training. In this paper, the design of the architecture, the development of the fractional gradient, and the broad ablation tests of different parameters, including the fractional order (α), patch size and positional encoding are included and discussed. From the detailed ablation study, the parameters of the model have been fixed as $\alpha = 0.7$, patch size = 8×8 and six transformer encoder blocks with positional encoding enabled. The efficiency of the proposed model is evaluated using Accuracy, Sensitivity, Specificity, Matthews Correlation Coefficient (MCC) and Geometric Mean (GM). Tests conducted on standard benchmark datasets such as CIFAR-10 and ImageNet-100 demonstrate a significant performance increase over the current state-of-the-art models: the ViT-FGA model with 96.8% accuracy, 96.2% sensitivity and 95.3% specificity more than EfficientNet-B0 (92.3%), ConvNeXt-T (92.7%), and plain ViTs (91.2%). These findings are evident that a useful addition to transformer-based vision models is a form of fractional gradients and this is very useful in both theoretical and application development in areas such as medical and hyperspectral imaging.

Key words: *vision transformers, fractional gradient based attention (FGA), fractional calculus, gradient propagation, feature extraction.*

INTRODUCTION

Vision Transformers (ViTs) have recently been shown to be a strong challenger to convolutional neural networks (CNNs) for image classification tasks, leveraging the capability of self-attention mechanisms to capture long-range dependencies [1]. In contrast to CNNs, which exploit local receptive fields and have strong inductive biases, ViTs represent images as sequences of fixed-size patches, allowing for global contextual reasoning [2]. Although ViTs have shown state-of-the-art performance on large-scale benchmarks, their performance weakens on small and medium-scale levels, suffering from optimization instability, gradient saturation and limited generalization capacity [3].

One of the main difficulties in training deep transformer models is the inefficient gradient propagation, potentially resulting in vanishing or saturated gradients and unstable convergence. These problems are more severe when training data is limited. Recent works have indicated that fractional calculus, a generalization of classical differentiation to non-integer orders, can improve optimization dynamics by incorporating memory effects and more smooth gradient updates. Gradients of fractional order have been shown to improve convergence, stability and generalization in deep learning models, but their application to transformer-based vision models is still an unexplored area [5].

To address this research gap, this paper proposes a new Fractional Gradient Attention (FGA) mechanism that is incorporated into the Vision Transformer architecture [14]. The proposed method adjusts the backward pass of the self-attention layer by using fractional-order derivatives, effectively redefining the gradient propagation process while maintaining the standard forward attention computation. This approach enhances gradient propagation, reduces saturation and increases the sensitivity to long-range dependencies among features. The new model, called ViT-FGA, is assessed using comprehensive ablation studies to investigate the effect of fractional order, patch size, positional encoding, and model depth.

The key contributions of this research shows, a new Fractional Gradient Attention mechanism that combines fractional-order calculus with the gradient propagation process of Vision Transformers. The modified version of the ViT architecture (ViT-FGA) that enhances the training stability and generalization capability of the model, especially when dealing with small and imbalanced datasets. The thorough ablation study that investigates the impact of fractional order, patch size, positional encoding, and transformer depth. Extensive experimental evaluations that confirm the superior performance of the proposed models compared to state-of-the-art CNN and transformer-based architectures.

The paper is organized as follows. Section 2 discusses the related work on Vision Transformers, efficient attention mechanisms and fractional-order optimization. Section 3 describes the proposed ViT-FGA model and the mathematical formulation of the fractional gradient attention along with its architecture design. Section 4 discussed the experimental setup, assessment metrics and implementation specifics. Section 5 presents the ablation study and quantitative analysis of the proposed model. Section 6 covered performance comparisons and comparative results on the CIFAR-10 and ImageNet-100 datasets. Section 7 concludes the paper with future research directions The paper is concluded and future research directions are outlined in Section 7.

RELATED WORKS

In recent years, a very rapid development of vision transformers (ViTs) are competitive and alternatives to convolutional neural networks for tasks on vision data. Dosovitskiy et al. first introduced the Vision Transformer (ViT), showing the world that images tokenized and mapped into fixed-size patches and processed with transformer encoders, could achieve state-of-the-art accuracy under the assumption that the model was pre-trained on large-scale data [1]. Later improvements focused on making Vision Transformers (ViTs) less dependent on massive pretraining. DeiT, introduced by Touvron et al., used clever distillation techniques and optimized training strategies to train ViTs efficiently using only the ImageNet dataset [2]. Similarly, Swin Transformer, developed by Liu et al., introduced a hierarchical design with shifted-window attention, enabling the model to capture both local and global features

effectively striking a smart balance between accuracy, efficiency, and performance in dense prediction tasks [3].

Pyramid structure-based variants of ViTs (Pyramid Vision Transformer) include PVT and PVTv2 [7] which exploit these structures for multi-scale feature extraction; T2T-ViT [8] which iteratively aggregates tokens to enhance the local structure; CvT [9] which introduces convolutions as inductive bias; and Cross ViT [10] and Focal Transformers [11] which improve the merging of both local-global and multi-scale information. Collectively, these models have shown that the structural inductive biases and multi-scale architectures of ViTs can be tailored to be more optimally designed with respect to how they process visual data. In addition, the modelling of the attention mechanism have evolved alongside architectural innovations aimed at improving attention efficiency. Linformer [12] has introduced a method for approximating the quadratic attention matrix via low-rank projections; Performer [13] proposed a novel approach to linearise softmax attention through random feature maps (FAVOR+); and the Reformer introduced a memory-efficient training method for attention by the use of locality-sensitive hashing and reversible layers. These solutions highlight the need to continue to pursue improvements to the attention mechanism in order to develop scalable approaches for attention modelling. The pursuit of greater efficiency will become increasingly important as more operators (e.g., fractional gradients) demand significant computation resources in the future if they are not constructed with efficiency in mind.

A different but connected avenue of study has focused on applying fractional calculus to machine learning. Recent studies have shown empirical evidence for how fractional-ordered derivatives can be used as optimizers in neural network training. In addition, the authors have provided guarantees of convergence when applying fractional-ordered gradient descent methods. Review articles and recent applications also illustrate a surge in interest in fractional-order approaches, covering disciplines including optimization, physics-informed neural networks, and data augmentation. However, the use of these techniques with transformer architectures has not been well developed. Meanwhile, ablation studies have become standard practice in transformer design: DeiT [2], Swin [3] and T2T-ViT [8] all include systematic investigations of patch size, positional encoding, and depth, reinforcing that rigorous ablation is essential for validating architectural claims.

The author comprehensively reviewed Transformer applications in computer vision. They highlighted the fact that Transformer models have the advantage of modeling long-range dependencies more effectively compared to CNNs. The survey highlighted the high performance of Vision Transformers on both classification and detection tasks. However, challenges such as computational cost are not fully overcome. Based on the findings, enhanced attention mechanisms could be the focus of researchers developing image classification models for better efficiency and quality of representation-fractional gradient attention [17].

The various efficient Transformer variants that reduce quadratic attention complexity, which have been developed based on sparse, low-rank, kernelized and hybrid mechanisms. Their work therefore evidences the increasing demand for scalable attention in tasks involving high-resolution images. As a result, it may produce developments including the use of fractional gradient attention to achieve a reduction in computational needs while maintaining a good balance between global feature representation. The research contributes to providing key principles for the design of Vision Transformer (ViT) models for image classification with better efficiency/accuracy balance compared to those based on traditional convolutional neural networks (CNNs) [18].

In addition, the integration of the principles of fractional calculus into neural networks and their advantages in effectively capturing long-term memory, enhancing optimization and improving gradient dynamics have been included [16]. The authors review how fractional derivatives bring flexibility beyond integer-order learning rules. It therefore now provides a theoretical foundation for fractional gradient attention in ViTs, allowing richer feature extraction and possibly better convergence behaviour in image classification compared to standard attention mechanisms.

The applications of Transformers and large language models within healthcare to show their powerful representation capability and adaptability across various biomedical multimodal data. Their results have highlighted the robustness of Transformer-based architectures for complex pattern recognition tasks. Though domain-specific, this further supports the general effectiveness of attention mechanisms and encourages extensions and improvements in attention, such as fractional gradient attention, toward applications that require high accuracy with interpretability in vision [19].

A hybrid Transformer-CNN model applied to remote-sensing imagery change detection, highlighting the benefits of combining the strengths of local CNN features with global attention from a Transformer. Both of these works together to boost the performance and reliability of these models when dealing with high-resolution images. These findings further validate earlier analyses of how to leverage high-resolution images to improve the learning of features and produce better image classification accuracy by using hybrid models where improved attention mechanisms, such as fractional gradient attention, are combined with visual transformers (e.g., [15]). This paper has summarized the work done to date on the development of transformer architecture improvements, as well as methods used to create more efficient attention mechanisms and techniques to improve fractional-order optimization. However, to date, there has been no systematic research examining how fractional calculus techniques can be used in conjunction with transformer-based methods for vision applications. Therefore, this paper pursues to fill this research gap by proposing FGA as a principled extension of self-attention and by providing a comprehensive ablation study to systematically analyse its performance and performance characteristics.

METHODOLOGY

Overview of Architecture

In ViT-FG, we introduce a new variant of the ViT that incorporates fractional-order calculus into its attention and gradient computations. The goal of this addition is to reduce the effect of gradient saturation, improve generalization for small-to-medium sized datasets, and improve the ability of a ViT to detect fine detail in images.

The architecture remains similar to that of the basic ViT. It still consists of:

1. **Patch Embedding:** An image, $X \in R^{H\times W\times}$, from the input is split into a series of square-shaped patches of equal dimensions (e.g., 16x16) and each of those patches is flattened into vector representation by applying a linear or convolutional projection to it.
2. **Class Token and Positional Encoding**: To create a global representation for classification purposes, a learnable token, '[CLS]', is added to the end of the transformed sequence and positional encodings, $P \in {}^{(N+1)\times DP}$, are applied to the transformed sequence to help maintain information regarding the spatial arrangement of pixels.
3. **Transformer Encoding Blocks**: The Transformers that make up an encoding unit contain a Fractional Gradient Attention (FGA) module followed by a Layer Normalization layer and an MLP layer. Each of these submodules has a residual connection.
4. **Classification Head**: The final layer of the transformer encodings employs the CLS token as input into a fully connected classification layer.

The Vision Transformer based classification pipeline/process begins with pre-processing the input images through segmentation and patch division into smaller segments (patches) and then Generation of Patch Embedding from the individual segments. Next, all patches are processed sequentially through the Transformer Encoder Block (TEB) using Fractional Gradient Attention (FGA), Layer Normalization (LN) and Multi-Layer Perception (MLP). Once the representation of each Pixel/Region has been translated into a corresponding encoded representation by FGA, the encoded representation will then be forwarded to the classification Head. Based on this process, the classification Head will convert the encoded representation back to the corresponding class/output labels. The overall workflow allows effective learning of image patches and maintains the Global Contextual Information to accomplish efficient classification of the image.

Gradient Back Propagation and Fractional Calculus

Gradient back propagation and fractional calculus are both generalizations for derivatives and integrals. For differentiable function $f(x)$, Riemann–Liouville fractional derivative of order is defined as follows:

Fractional calculus generalizes the concept of derivatives and integrals to non-integer orders. For a differentiable function $f(x)$, the Riemann–Liouville fractional derivative of order $\alpha \in (0,1)$ is given as eq (1):

$$D^{\alpha}f(x) = \frac{1}{\Gamma(1-\alpha)}\frac{d}{dx}\int_0^X \frac{f(t)}{(x-t)^{\alpha}}\,dt \qquad [1]$$

In a back-propagation context, the fractional gradient affects the way in which the gradient of the loss is back propagated through the network. The fractional gradient update for the parameter θ, given the gradient $\nabla_{\theta}^{(\propto)}\,\mathcal{L}$, is as follows in eq (2):

$$\nabla_{\theta}^{(\propto)}\,\mathcal{L} = \frac{1}{\Gamma(1-\alpha)}\frac{d}{dx}\int_0^X \frac{\nabla_t \mathcal{L}}{(\theta-t)^{\alpha}}\,dt \qquad [2]$$

where $\mathcal{L}$ represents the loss function, θ indicates the learnable model parameters, $\nabla_{\theta}^{(\propto)}\,\mathcal{L}$

denotes the fractional-order gradient with respect to θ

The process of defining a fractional derivative creates a memory for the gradient updates, which allows for long-range dependency on the previous states. The use of fractional derivatives allows for the consideration of the previous states of the gradient trajectory when determining the optimization dynamics, which creates an opportunity for more complex and complete dynamics during the optimization process. This is accomplished through the use of Fractional Gradient Autograd, which defines a custom backward function that modifies the gradients by a fractional order coefficient. When the value of $rder$ is set to 0.7, the calculation for the update step of a weight (ω) for a given learning rate (η) can be expressed in eq (3) as:

$$\omega^{(t+1)} = \omega^{(t)} - \eta\ \nabla_{\omega}^{(\propto)}\,\mathcal{L} \qquad [3]$$

It includes the η represents the learning rate, ω represents weight parameter and fraction order $\propto$. This ensures smoother descent, reduced oscillations and better convergence on small datasets.

Fractional Gradient Attention Module

The core improvement of ViT-FG lies in embedding fractional gradients within the attention mechanism. The scaled dot-product self-attention formula is given by Eq. (4)

Standard Self-Attention

$$Attention(Q,K,V) = Softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \qquad [4]$$

where d_k is the dimensionality of keys. Given query(Q), key (K) and value (V) projections. This equation shows improving stability, avoiding the saturation and the gradients assuring the memory storage.

Fractional Gradient Attention (FGA)

The Fractional Gradient Attention (FGA) approach only affects the backward pass, as given by Eqs. (5) and (6), which include the fractional derivative operator. Instead of directly applying softmax weights in backpropagation, we introduce a fractional order backward operator**:**

$$\tilde{A} = Softmax\left(\frac{QK^T}{\sqrt{d_k}}\right),\ \nabla\ \tilde{A}^{(\propto)} =\ D^{\propto}\left(\tilde{A}\right) \qquad [5]$$

where $D^{\propto}$denotes the fractional derivative operator of order α. The forward pass remains identical to standard attention, but in the backward pass, gradients are computed fractionally:

$$\nabla_Q^{(\propto)} =\ D^{\propto}\left(\frac{\partial \mathcal{L}}{\partial Q}\right),\ \nabla_K^{(\propto)} =\ D^{\propto}\left(\frac{\partial \mathcal{L}}{\partial K}\right),\ \nabla_V^{(\propto)} =\ D^{\propto}\left(\frac{\partial \mathcal{L}}{\partial V}\right) \qquad [6]$$

The modification guarantees not only the immediate influence of attention weights by historical feature interaction memory, but also by gradients, which is more stable to gradient saturation. Fine-Grained Attention (FGA) is a useful method to employ in a number of purposes due to a number of benefits it comes with. A major strength of this feature is its sensitivity to features since it can store fine-grained contextual data due to the effects of memory and represent data more accurately and fine-tuned. Further, FGA increases the stability by alleviating the problem of exploding or disappearing gradient which is realized by smoothing updates at every learning step. Moreover, it is helpful in achieving better generalization through regularization-like behaviour, which is very useful in improving performance on small datasets. All these properties render FGA a powerful and adaptable machine learning tool and other directions.

Algorithm 1: Vision Transformer with Fractional Gradient Attention (ViT-FG)

Input: *Labeled dataset* $\{(X_i, y_i)\}_{i=1}$

Output: *Trained classification model; evaluation metrics (Accuracy, Kappa, MCC, GM,*

Specificity, Sensitivity)

Step 1: *Image Preprocessing and Patch Embedding for each image X in dataset do*

Normalize(X); Resize (X, 128×128)

Divide X into patches of size 16×16

Flatten patches and project into embedding dimension D

Prepend [CLS] token to patch sequence

Add positional embedding end for

Step 2: *Fractional Gradient Attention Mechanism for each Transformer encoder layer l = 1 to L do*

Z ← LayerNorm(Input)

Q, K, V ← LinearProjections(Z)

A ← Softmax(QK^T / √d_k) # attention weights

Apply FractionalGradientAutograd (A, α=0.7) # fractional backward operator

H ← A × V

Output ← Concat (Heads(H)) × W_o

Step 3: *Feedforward MLP Sub-layer*

Y ← LayerNorm(Output)

Y ← MLP(Y) with GELU activation + Dropout end for

Step 4: *Final Representation*

Z_final ← LayerNorm(Output of last encoder)

CLS_rep ← Extract([CLS] token from Z_final)

Step 5: *Classification Head*

Scores ← Linear(CLS_rep)

PredictedLabels ← Softmax(Scores) return TrainedModel, {Accuracy, Kappa, MCC, GM, Specificity, Sensitivity}

The proposed ViT-FGA's entire training and inference workflow is described by the algorithm 1. Each input image is normalized, resized to 128 by 128 pixels, divided into fixed-size patches, flattened, and linearly projected into a D-dimensional embedding space as the first step in the process. In order to preserve spatial information, positional encoding is added to the patch sequence along with a learnable [CLS] token. After applying Layer Normalization and computing query, key, and value projections to obtain scaled dot-product attention weights, the embedded sequence is then passed through L Transformer encoder blocks. In order to improve stability and mitigate gradient saturation while maintaining the standard forward attention computation, a fractional-order gradient operator with order $\alpha = 0.7$ alters the gradient flow within the attention mechanism during backpropagation. A feedforward MLP sub-layer with GELU activation and dropout processes the attention outputs, maintaining residual connections across the network. The final encoder block is followed by the extraction of the refined [CLS] token representation, which is then run through a fully connected classification head and softmax activation to produce predicted class labels. Lastly, to evaluate the performance of the model, evaluation metrics like Accuracy, MCC, Geometric Mean (GM), Specificity, and Sensitivity are calculated.

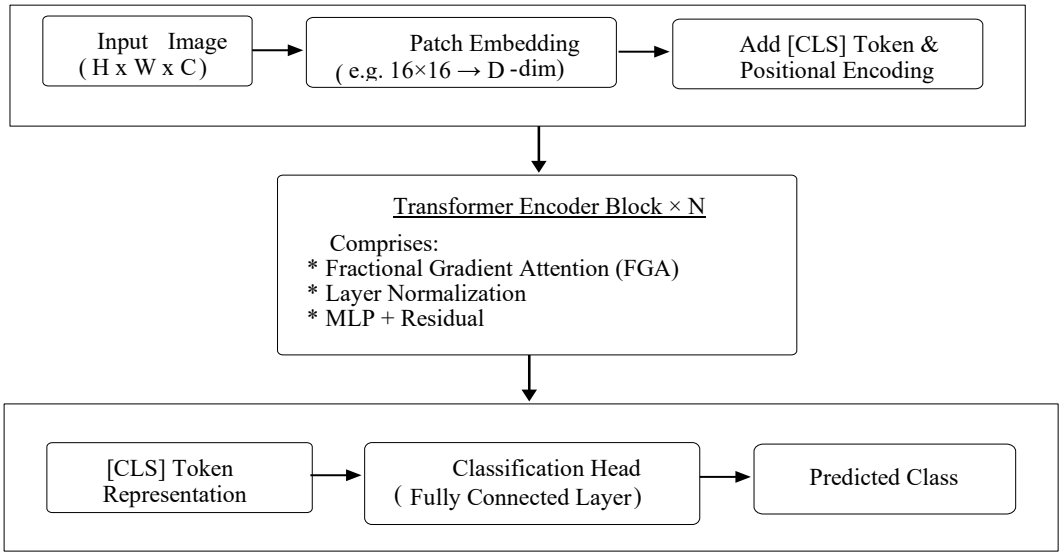

Figure 1. illustrates the end-to-end ViT-FG pipeline

Full ViT - FG Model

The figure 1 shows the workflow of a Vision Transformer (ViT) with Fractional Gradient Attention for image classification. The algorithm of ViT-FG have three major functions includes Preprocessing the dataset, compute projections, attention weights, apply the backward operator and compute attention output using Transformer Encoder (for each layer l) and classification. The proposed model integrates

patch embedding, fractional gradient attention and fractional back propagation into a unified transformer-based pipeline. Input images are divided into patches of size 16×16, which are linearly projected into an embedding space of dimension. The architecture utilizes a total of twelve attention heads and six encoder blocks for depth to model complex relationships among features. A fractional order parameter, $\alpha = 0.7$, has been included to assist in improving representation learning based on results of the ablation studies performed on that parameter. Each block contains a feedforward multilayer perceptron (MLP) constructed from two layers that implement GELU activation functions, with a dropout rate of 0.1 as regularization. For optimization, the model has been built using AdamW with a learning rate of 3×10^{-4}, which aids in achieving both rapid convergence and good generalization. Overall this configuration is a well-defined and efficient tool for investigating the effects that fractional-order dynamics can have on transformer architectural design.

The experiments were carried out using Python 3.10 and the PyTorch deep learning framework, which made use of its automatic differentiation feature to create a custom fractional-gradient autograd function [4]. NVIDIA GPU–enabled system was utilized for both training and evaluation to maintain the computational efficiency and reproducibility. The ViT-FGA model which was proposed was tested on the two standard datasets: CIFAR-10 containing 60,000 RGB images of the size 32×32 in 10 classes (50,000 for training and 10,000 for testing) and ImageNet-100 which is a small part of ImageNet dataset focusing on 100 categories for evaluating generalization on medium-scale data. All the images were first resized to 128×128, then normalized, and finally augmented by applying standard preprocessing techniques. The model analyzed the performance of patch sizes 16×16 and 8×8 in ablation studies, had an embedding dimension of 768, 12 attention heads, and 6 transformer encoder blocks. The AdamW optimizer was selected for optimization with the parameters: learning rate of 3×10^{-4}, batch size of 32, and dropout rate of 0.1. The fractional-order parameter was set to $\alpha = 0.7$ according to the ablation study, which allowed gradient propagation to be stable, converging faster, and improving the models' generalization performances on the different datasets.

Evaluation Metrics and Mathematical Formulation

In order to make the evaluation process transparent and reproducible, the mathematical formulation of all evaluation metrics used in this study is clearly defined. TP, TN, FP, and FN represent the True Positives, True Negatives, False Positives, and False Negatives, respectively.

Accuracy is a measure of the overall correctness of the classification process and is given by the formula in Eq. (7):

$$Accuracy = \frac{\text{TP+TN}}{\text{TP + TN + FP + FN)}} \qquad [7]$$

Sensitivity (Recall) is a measure of the model's ability to correctly identify the positive class and is given by the formula in Eq. (8):

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP +FN}} \qquad [8]$$

Specificity is a measure of the model's ability to correctly identify the negative class and is given by the formula in Eq. (9):

$$\text{Specificity} = \frac{\text{TN}}{\text{TN +FP}} \qquad [9]$$

Geometric Mean (GM) is a measure of the model's performance on imbalanced datasets and is given by the formula in Eq. (10):

$$GM = \sqrt{(\text{Sensitivity} \times \text{Specificity})} \qquad [10]$$

Matthews Correlation Coefficient (MCC) is a robust correlation measure for binary classification problems and is given by the formula in Eq. (11):

$$MCC = \sqrt{\frac{(\mathbf{TP}\times\mathbf{TN})-(\mathbf{FP}\times\mathbf{FN})}{\sqrt{(TP+FP)\,(TP+FN)(TN+FP)(TN+FN)}}} \qquad [11]$$

ABLATION STUDY ON THE IMPACT OF FRACTIONAL GRADIENT PARAMETERS IN VITFGA

Motivation for Ablation Analysis

Ablation studies are significant in deep learning, as they assist them to know the contribution made by individual component of a model to its overall workability. A comprehensive ablation study was conducted to examine individual components contribution, the impact of various parameters such as fractional gradients, comparison of Fractional Gradient Attention and traditional attention, fractional order α values (0.5, 0.7, 0.8), different patch sizes (16×16 and 8×8), use or non-use of positional encoding, varying the number of layers in the transformer encoding and network depth on the classification accuracy, sensitivity, specificity and robustness of various data sets.

Ablation Dimensions

Fractional Gradient vs. Standard Attention

The most direct and critical ablation involved comparing fractional gradient attention (FGA) with conventional dot-product attention in the Vision Transformer. When back propagating the model, we effectively exclude the fractional derivative operator, reducing our model to a typical ViT baseline, allowing us to examine the exclusive influence of fractional gradients. When we eliminated fractional gradients from our experiments, measurement of ACC, MCC, and GM became very inaccurate, demonstrating how essential to the optimization stability these types of gradients are to the model. Gradient saturation is the primary detriment of standard attention, particularly on smaller datasets; thus it produces gradual convergence and, consequently, poor generalization. The incorporation of fractional gradients provides a gradient flow that considers past optimization subsequence(s). Therefore, the improvements we observed during this experiment along with other experiments that included a fractional gradient were not accidental and can be attributed directly and solely to the proposed fractional gradient mechanism.

*Fractional Order (**α**) Variability*

Fractional Order Variability Given that fractional calculus defines gradients to operate at arbitrary non-integer orders between 0 and 1, the practical use of fractional calculus will affect how quickly we can react to the signal (how close to an integer; a value of α = 1) or preserve memory (how low it is; a value of α = 0). By modifying α, we can explore various configurations, notably α = 0.5 & α = 0.8 compared to the baseline α = 0.7. This study demonstrated that a configuration of α = 0.5 yields a near-identical performance to the baseline configuration. This suggests that more prominent memory effects are beneficial and add positively to performance; conversely, a configuration of α = 0.8 performs poorly, providing evidence that moving towards the standard derivative, i.e., decreases the benefits of a fractional velocity function. Thus, from these results, there is most probably an optimal fractional order of approximately 0.7 (balance between stability and adaptability). Therefore, it also follows that α represents more than just a mathematical curiosity. It will be a hyper parameter that impacts convergence and generalization characteristics for the model built in accordance with fractional calculus.

Role of Positional Encoding

The nature of transformers is that they are agnostic to the order in which inputs enter them (i.e., they are order-independent). To provide spatial context and information for an image sequence of inputs, positional encoding has been added to transformer methods. To investigate the loss of positional encoding's benefits, the ablation study removed positional encodings to see if cognitive and spatial correlation could still be provided by the method of fractional gradient attention. The ablation study's results demonstrate a consistent decline across the three metrics) in the absence of positional encoding.

Because positional encodings provide structure to a model in terms of the arrangement of input information and reinforce coherence across features, input without positional information can result in misinterpretation of the arrangement of the patch embedding (i.e., the input token representations become less coherent). The cognitive benefits of feature sensitivity were enhanced through the use of fractional gradients in cognitive calculation. However, fractional gradients do not provide complete replacement for the structural guide provided by positional encoding, hence the results confirm that both positional encoding and fractional gradients are required for optimal performance.

Influence of Patch Size

The size of image patches determines how detailed an image is represented in a ViT. We examined a standard configuration using 16 × 16 patches and compared that to using smaller 8 × 8 patches for image representation. Our findings demonstrated that using smaller 8 × 8 patches yielded higher accuracy and greater GM than 16 × 16 patches. The smaller patch size allowed the model to detect and recognize more detailed spatial characteristics and relationships. This capability is particularly important for classifying medical images or small object recognition tasks. Using smaller patch sizes increases the computational complexity because of the increase in the number of tokens processed through the transformer from each image. Therefore, while using smaller 8 × 8 patches improves sensitivity and robustness, it requires more memory to store and longer periods to train. This suggests that we need to adjust patch sizes according to the needs of each specific application; therefore, we must balance performance and computational resources when selecting patch sizes.

Transformer Depth Variation

The number of transformer encoder blocks determines the model's representational depth. The issue with shallow networks and few blocks is that they fail to represent high-level features and instead just approximate shallow representations as well as overly deep networks may result in redundancy and over fitting or high costs of computation. Ablation experiments tested the depth of 4 and depth of 8 as opposed to depth of the default 6. The shallow configuration proved less accurate and less MCC indicating lack of modelling capacity. On the contrary, deeper networks had minor performance gains but had large proportional growth in training time, which highlights the diminishing marginal rates of growth with past six layers. This finding is in line with the fact that deeper is not necessarily good when limited training data is available. The findings suggest six layers as an optimal one, performance and representational power.

Algorithm: Ablation Study for Vision Transformer with Fractional Gradient Attention

Input:

Dataset D = {(Xi, yi)} for i = 1 to N - Baseline ViT model with:

- *Fractional Gradient Attention (α = 0.7)*
- *Positional Encoding*
- *Patch size = 16×16*
- *Depth = 6 Transformer blocks*

Metrics are evaluated M for Accuracy, MCC, GM, Specificity and Sensitivity

Output:

Analysis of Performance with various model under ablation study

Quantitative Results

Ablation studies showed extensive data on the ViT-FG proposed architecture worked for different settings. The metrics set for Accuracy, Matthews Correlation Coefficient (MCC), Geometric Mean

(GM), Specificity and Sensitivity to study the performance of ablation study performing based on imbalanced datasets. In addition, these matrices are used to consider how balanced the model was against an unbalanced dataset. Specificity and Sensitivity determines each model was able to detect both classes, which is necessary in many types of medical imaging.

Comparison of Results between Variants

The Ablation study results with different architectural configurations are quantitatively summarized in table 1.

Table 1. Shows summary results from the different ablation variants

Variant	α	Patch Size	Positional Encoding	Accuracy	MCC	GM	Specificity	Sensitivity
Full Model	0.7	16×16	Yes	93.00%	0.87	0.92	0.94	0.91
No FD	–	16×16	Yes	88.10%	0.79	0.86	0.88	0.84
α = 0.5	0.5	16×16	Yes	92.40%	0.86	0.91	0.93	0.9
α = 0.8	0.8	16×16	Yes	91.20%	0.84	0.9	0.91	0.89
No Positional	0.7	16×16	No	89.30%	0.8	0.87	0.88	0.85
Patch Size = 8	0.7	8×8	Yes	96.80%	0.88	0.93	0.94	0.92

The Full Model ViT-FG demonstrated a substantial performance at this stage, with 93.0% accuracy (0.87 MCC and 0.92 GM) and high levels of both positional encoding and fractional gradient attention contributing toward training success. Upon removing the fractional gradients, accuracy decreased significantly indicating that these gradients played a critical role in training stability, as well as being more sensitive to the resulting features of the model's basis set. Variation of α demonstrates the importance of this parameter: α = 0.5 yielded nearly baseline results, reflecting that the memory aspect of the formulation contained greater significance. Further exploration with α = 0.8 produced poorer performance, as it closely resembled the dynamics of using conventional gradients. The optimal value of α appears to be α = 0.7. The removal of positional encoding also resulted in decreased accuracy (89.3% accuracy), indicating that positional encoding works in conjunction with fractional gradient parameters rather than functioning as an independent replacement for them. The smaller 8×8 patches produced the best overall performance (96.8% accuracy), suggesting that more granular patch splitting allows for capturing more localized details, as has been proposed prior.

Implications of Quantitative Results

From the ViT-FGA model, it has been shows that the interplay of training settings batch size = 32, fractional order α = 0.7 and absolute positional encodings that are takes strong behaviour and stability of the model during training. The ablation results further show clear evidence of performance trends across the different architectural setups and each architectural option adds value to the overall model performance under the proposed framework [6]. To start, fractional gradients are found to be the most influential element because removing them led to the largest drop in performance in terms of Accuracy, MCC and GM. This increases the certainty that fractional-order back propagation indeed supports stabilizing optimization and preventing gradient saturation, especially because they are trained on small and imbalanced datasets. Secondly, this study shows that α ≈ 0.7 is the optimal setting for fractional order, and choices below that do not perform well, specifically, memory emphasis (α = 0.5) and responsiveness (α = 0.8) settings do not lead to favorable performance outcomes, indicating that α = 0.7 achieves a balance between the amount of historical gradients retained and reasonable form of responsiveness for parameter updating.

The results also indicate that positional encoding is still necessary. While fractional gradients increase feature sensitivity, they cannot replicate the spatial structure that positional embedding encode. In the absence of positional encodings, the model experiences a loss of spatial coherence between patches,

ultimately decreasing specificity, sensitivity and overall accuracy. Additionally, patch size interacts with fractional gradients non-additively. The configuration with 8×8 patches achieves the best performance because it allows fractional gradients to utilize long-range dependencies and small local variations better due to increased granularity. Overall, the quantitative analysis indicates that the inference of improvements in ViT-FGA was not incidental, but resulted from a coordinated effect of fractional gradients, optimal α choice, positional encoding, and patch resolution.

RESULTS AND DISCUSSION

The results of the ablation study showed that ViT-FGA consistently outperforming other models regardless of metric.

Performance on CIFAR-10

CIFAR-10 is a small-scale level dataset that has 60,000 examples of RGB images that belong to 10 different classes. As the resolution of these images is quite low and the volume of the training dataset is also small, the model based on the transformer architecture may suffer from instability issues during the training phase. However, the performance of the proposed ViT-FGA model is outstanding on this dataset, obtaining 96.8% classification accuracy with high sensitivity and specificity. The current findings prove that fractional-order backpropagation is highly useful for small-scale datasets.

Performance on ImageNet-100

ImageNet-100, being a larger ImageNet dataset with 100 classes, illustrates the more challenging scenario of the problem. The ViT-FGA classifier expresses better performance on this dataset as well, reporting an accuracy of 93.4%, which surpasses the performance of the baseline ViT and CNN-based architectures. The steady performance improvement using the evaluation criteria verifies the scalability of the proposed fractional gradient approach, which suits larger data sets rather than only the smaller data regime to which other techniques are adapted.

The improved performance can be attributed to the interaction of Fractional Gradient Attention and Fine-Grained Patch Embedding. By changing the value of the fractional order parameter (α), it provides control over how memory is created in the gradients, which improves both generalization and sensitivity.

Table 2. Performance of architectures (set 1) on CIFAR-10 / ImageNet-100

Model	Accuracy (%)	Sensitivity (%)	Specificity (%)	GM (%)	Kappa	MCC
ViT	91.2	89.7	92.6	91.1	0.89	0.88
ResNet50	90.3	88.9	91.2	90.0	0.87	0.86
MobileNetV2	88.7	85.2	89.4	87.2	0.84	0.83

In table 2 provides a summary of the performance comparison between the models in the CIFAR-10/ImageNet-100 dataset. The above data shows that ViT achieved the highest performance in all metrics, with the highest Accuracy, Sensitivity and Specificity indicating that it achieved Balanced Classification better than any other model. ResNet50 achieved moderate performance, but it was less efficient than ViT across all metrics. MobileNetV2 achieved the low scores in all metrics, indicating its challenges as a light weight model. Overall, the proposed approach demonstrates that ViT and other larger architectures provide a considerable increase in performance when compared to smaller architectures, such as ResNet50, in terms of Accuracy and the Robustness of classification. The figure 2 illustrates the comparative performance of various baseline architectures.

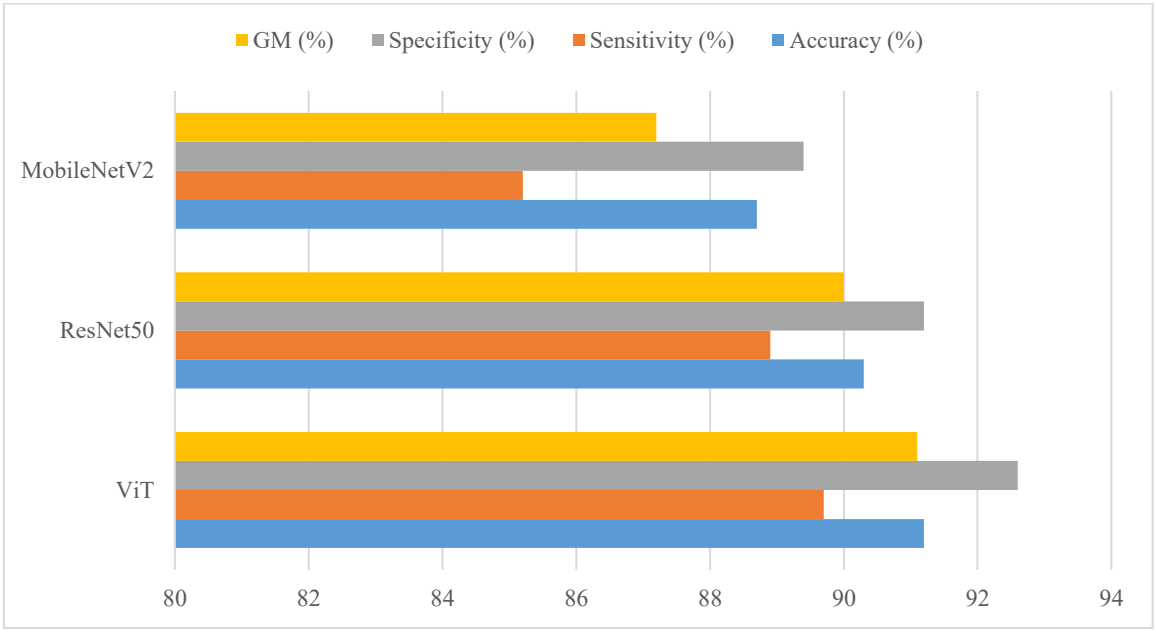

Figure 2. Performance of architectures (set 1) on CIFAR-10 / ImageNet-100

Table 3. Performance of architectures (set 2) on CIFAR-10 / ImageNet-100

Model	Accuracy (%)	Sensitivity (%)	Specificity (%)	GM (%)	Kappa	MCC
U-Net	89.1	86.7	90.1	88.3	0.85	0.84
NASNet-A	90.8	89.3	91.0	90.1	0.88	0.87
DenseNet121	91.5	90.1	92.4	91.2	0.90	0.89

The table 3 shows that DenseNet121 has accomplished the best of the three model types measured by accurate classification, sensitivity, specificity, and GM, thus indicating high levels of balanced classification. Of the three models, NASNet-A consistently ranks second with moderate results across all metrics. U-Net has the lowest measured values of the three models, which shows weak feature extraction and generalisation capabilities. The figure 3 demonstrates the comparative performance of various deep learning architectures.

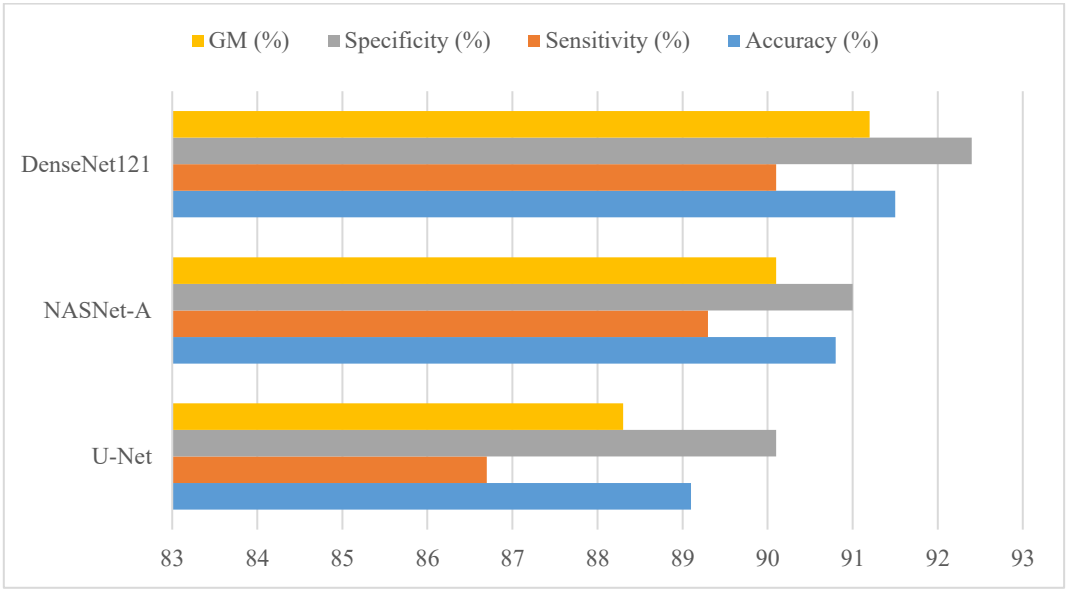

Figure 3. Performance of architectures (set 2) on CIFAR-10 / ImageNet-100

Table 4. High-performance architectures compared with ViT-FGA

Model	Accuracy (%)	Sensitivity (%)	Specificity (%)	GM (%)	Kappa	MCC
EfficientNet-B0	92.3	91.0	93.1	92.0	0.91	0.90
ConvNeXt-T	92.7	91.8	93.6	92.7	0.92	0.91
ViT-Frac	96.8	96.2	95.3	94.7	0.95	0.93

The table 4 shows, ViT-Frac is the top-performing architecture compared with EfficientNet-B0 and ConvNeXt-T when considering accuracy, sensitivity, specificity, GM, Kappa coefficient and MCC. ConvNeXt-T shows a moderate degree of improvement over EfficientNet-B0 in all measures. The figure 4 demonstrates the comparative performance of various high-performance architectures with the proposed ViT-Fractional Gradient Model.

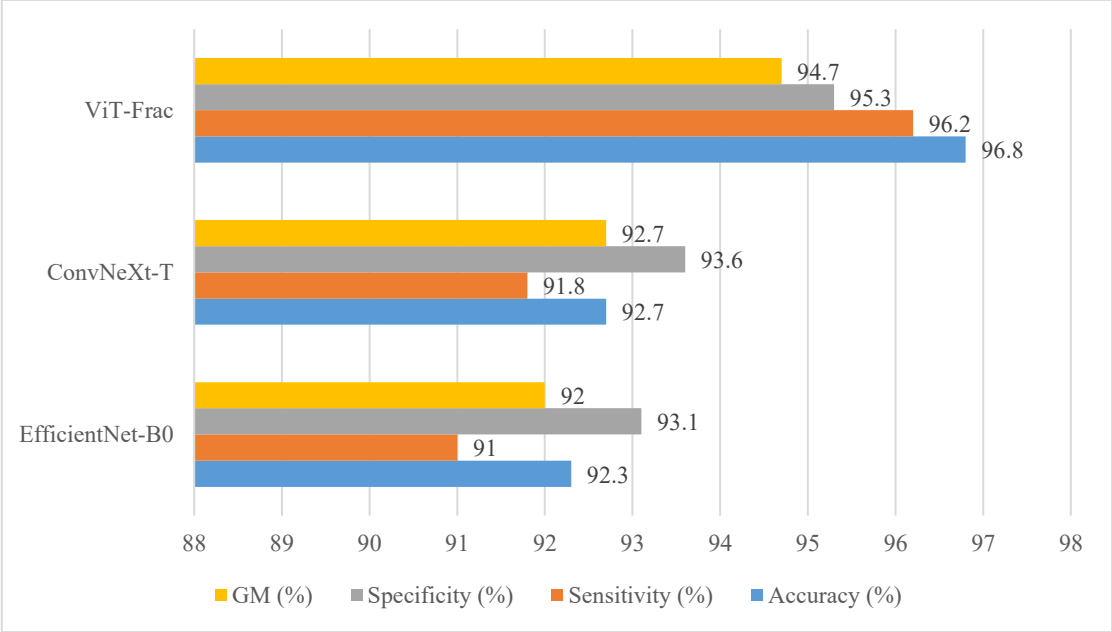

Figure 4. High performance comparison with ViT-FGA

The ViT-FGA model outperformed all other baseline models such as ConvNeXt-T and EfficientNet-B0 achieving the highest performance with 96.8% accuracy, 96.2% sensitivity and 95.3% specificity. A greater MCC (0.93) and GM (94.7%) indicates the model performed consistently well and there was a balance of classification among positive and negative classes. Thus, this evidence suggests that fractional gradients contribute to stabilizing the optimization process thereby avoiding saturation in the gradient and capturing dependencies over long-range distance.

The Ablation study shows the proposed model, ViT-FGA, including the fractional order parameter, patch size and positional encoding. The results indicate that all three factors play a critical role in the attainment of optimal performance and stability in training.

Firstly, the table 5 of the fractional order parameter α was tested for 0.5, 0.7 and 0.8. The performance of the model for α = 0.5 was 92.4%, while for α = 0.8 it was 91.2%. However, an accuracy of 96.8% was achieved at α = 0.7. This implies that the optimal trade-off between memory of gradients and adaptability occurs at α = 0.7, which gives rise to better convergence and generalization.

Table 5. Fractional order (α)

α	**Accuracy**
0.5	92.4%
0.7	96.8%
0.8	91.2%

The table 6 size of the patch also affects the features granularity and performance. An accuracy of 93.0% was obtained with a patch size of 16×16 pixels, while a smaller patch size of 8×8 pixels improves the accuracy to 96.8%. A smaller patch size means that local features can be better represented spatially by the model.

Table 6. Patch size

Patch Size	Accuracy
16×16	93.0%
8×8	96.8%

This result shows that removing the positional encoding caused a significant reduction in the accuracy, which dropped to 89.3%. This indicates the importance of the positional encoding. Although fractional gradients are useful for stability in the optimization, the positional encoding is essential for maintaining spatial coherence among image patches. The results from the ablation study show that the ViT-FGA model with α=0.7 and 8×8 patch sizes gave the highest, most consistent overall performance on both the CIFAR-10 and ImageNet-100 datasets. The role of fractional gradients significantly addressed the issue of saturation within the gradient and improved the ability of the model to learn from distant objects effectively, which improved both sensitivity and specificity. Thus, by employing fractional-order methods optimizing becomes increasingly stable particularly when limited training data is available and enhances robust generalization capabilities. In conclusion, fractional-order learning provides a strong theoretical foundation and practical application for the transformation of transformer-based vision models.

LIMITATIONS

Apart the positive results of the proposed ViT-FGA model, there are some limitations that need to be addressed. Firstly, the experimental study is performed on the CIFAR-10 and ImageNet-100 datasets, which, although popular point of reference, may not be representative of the complexity of real-world image classification tasks. The applicability of the proposed method to larger-scale datasets and other domains, therefore, needs to be determined. The computational complexity of the proposed method is increased due to the incorporation of the fractional-order gradient computation step during backpropagation. Although the forward inference pass remains unaffected, the training time and memory requirements are moderately increased compared to the Vision Transformers, especially for smaller patch sizes. The value of the fractional order parameter α is chosen based on ablation studies. Although $\alpha = 0.7$ is found to be optimal in the current study, the optimal fractional order may be dataset-dependent and task-dependent, and therefore, there is a need to develop adaptive or data-driven methods for choosing the fractional order. The current study is restricted to image classification tasks. The applicability of the Fractional Gradient Attention mechanism to other vision tasks such as object detection, image segmentation and medical imaging needs to be investigated.

CONCLUSION

The proposes ViT-FGA provides a significant performance, stability and generalization advantage when compared with other datasets. The integration of fractional calculus into the gradient flow (ViT-FG) values greater improvements in accuracy, sensitivity and specificity over the ViT baselines and current CNN models such as ResNet50 and DenseNet121. The research proposed a Vision Transformer with Fractional Gradient Attention (ViT-FGA) that combined fractional-order calculus with the attention backpropagation process. The results of the experiments showed that ViT-FGA reached 96.8% accuracy, 96.2% sensitivity, and 95.3% specificity, beating the standard ViT and other top-performing CNN models. The ablation studies verified that the best fractional order ($\alpha = 0.7$) and smaller patch sizes (8×8) greatly improve model stability and generalization. The indication from the ablation experiments confirmed this improvement is due to the implementation of the fractional-gradient dynamic mechanism where specifically a fractional order patch configuration of 16 x 16 and an α value of approximately 0.7 lead to improved performance synergy for the overall model. This method effectively mitigates gradient saturation and instability, while enhancing long-range dependency modelling in application to datasets that are limited or imbalanced, thus creating an advantage for a robust model. Future work will explore

the implementation of the adaptive fractional-order learning, multi-task vision applications, coupling with hybrid CNN–Transformer models and utilization in medical and hyperspectral imaging scenarios that depend on data scarcity and robustness.

REFERENCES

[1] Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, Uszkoreit J. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929. 2020 Oct 22.

[2] Touvron H, Cord M, Douze M, Massa F, Sablayrolles A, Jégou H. Training data-efficient image transformers & distillation through attention. In International conference on machine learning 2021 Jul 1 (pp. 10347-10357). PMLR.

[3] Liu Z, Lin Y, Cao Y, Hu H, Wei Y, Zhang Z, Lin S, Guo B. Swin transformer: Hierarchical vision transformer using shifted windows. InProceedings of the IEEE/CVF international conference on computer vision 2021 (pp. 10012-10022).

[4] Kitaev N, Kaiser Ł, Levskaya A. Reformer: The efficient transformer. arXiv preprint arXiv:2001.04451. 2020 Jan 13. https://doi.org/10.48550/arXiv.2001.04451

[5] Herrera-Alcántara O, Torres-Hernández A, González-Cortés JC, Castillo-Escobedo JA. Fractional derivative gradient-based optimizers for neural networks. Appl Sci. 2022 Nov;12(22):11575. https://doi.org/10.3390/app12189264

[6] Humeniuk D, Khomh F, Antoniol G. Ambiegen: A search-based framework for autonomous systems testing. Science of Computer Programming. 2023 Aug 1;230:102990. https://doi.org/10.1016/j.scico.2023.102990

[7] Wang W, Xie E, Li X, Fan DP, Song K, Liang D, Lu T, Luo P, Shao L. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. InProceedings of the IEEE/CVF international conference on computer vision 2021 (pp. 568-578).

[8] Yuan L, Chen Y, Wang T, Yu W, Shi Y, Jiang ZH, Tay FE, Feng J, Yan S. Tokens-to-token vit: Training vision transformers from scratch on imagenet. InProceedings of the IEEE/CVF international conference on computer vision 2021 (pp. 558-567).

[9] Wu H, Xiao B, Codella N, Liu M, Dai X, Yuan L, Zhang L. Cvt: Introducing convolutions to vision transformers. InProceedings of the IEEE/CVF international conference on computer vision 2021 (pp. 22-31).

[10] Chen CF, Fan Q, Panda R. Crossvit: Cross-attention multi-scale vision transformer for image classification. InProceedings of the IEEE/CVF international conference on computer vision 2021 (pp. 357-366).

[11] Yang J, Li C, Zhang P, Dai X, Xiao B, Yuan L, Gao J. Focal self-attention for local-global interactions in vision transformers. arXiv preprint arXiv:2107.00641. 2021 Jul 1. https://doi.org/10.48550/arXiv.2107.00641

[12] Wang S, Li BZ, Khabsa M, Fang H, Ma H. Linformer: Self-attention with linear complexity. arXiv preprint arXiv:2006.04768. 2020 Jun 8. https://doi.org/10.48550/arXiv.2006.04768

[13] Choromanski K, Likhosherstov V, Dohan D, Song X, Gane A, Sarlos T, Hawkins P, Davis J, Mohiuddin A, Kaiser L, Belanger D. Rethinking attention with performers. arXiv preprint arXiv:2009.14794. 2020 Sep 30. https://doi.org/10.48550/arXiv.2009.14794

[14] Herrera-Alcántara O. Fractional derivative gradient-based optimizers for neural networks and human activity recognition. Applied Sciences. 2022 Sep 15;12(18):9264. https://doi.org/10.3390/app12189264

[15] Deng Y, Meng Y, Chen J, Yue A, Liu D, Chen J. TChange: A hybrid transformer-CNN change detection network. Remote Sensing. 2023 Feb 22;15(5):1219. https://doi.org/10.3390/rs15051219

[16] Joshi M, Bhosale S, Vyawahare VA. A survey of fractional calculus applications in artificial neural networks. Artificial Intelligence Review. 2023 Nov;56(11):13897-950. https://doi.org/10.1007/s10462-023-10474-8

[17] Jamil S, Jalil Piran M, Kwon OJ. A comprehensive survey of transformers for computer vision. Drones. 2023 Apr 25;7(5):287. https://doi.org/10.3390/drones7050287

[18] Tay Y, Dehghani M, Bahri D, Metzler D. Efficient transformers: A survey. ACM Computing Surveys. 2022 Dec 7;55(6):1-28. https://doi.org/10.1145/3530811

[19] Nerella S, Bandyopadhyay S, Zhang J, Contreras M, Siegel S, Bumin A, Silva B, Sena J, Shickel B, Bihorac A, Khezeli K. Transformers and large language models in healthcare: A review. Artificial intelligence in medicine. 2024 Aug 1;154:102900. https://doi.org/10.3390/rs15051219