Letter

Dendritic Learning-Incorporated Vision Transformer for Image Recognition

Zhiming Zhang ⃝, Zhenyu Lei ⃝, Masaaki Omura ⃝,
Hideyuki Hasegawa ⃝, *Member, IEEE*, and
Shangce Gao ⃝, *Senior Member, IEEE*

Dear Editor,

This letter proposes to integrate dendritic learnable network architecture with Vision Transformer to improve the accuracy of image recognition. In this study, based on the theory of dendritic neurons in neuroscience, we design a network that is more practical for engineering to classify visual features. Based on this, we propose a dendritic learning-incorporated vision Transformer (DVT), which outperforms other state-of-the-art methods on three image recognition benchmarks.

Introduction: Image recognition, as an upstream task of many computer vision problems, has very important research value. Many studies focus on optimizing the architecture of the feature extraction network to make it extract richer and more representative image features. In the early stages of deep learning, the convolutions are simply stacked to build feature extraction networks. While effective, this method had some limitations such as the need for large amounts of data, long training times, and limited interpretability [1]. To address these issues, researchers have introduced more effective and biologically interpretable structures. The use of residual connections [2], densely connected blocks [3], and attention mechanisms [4] have all been explored to improve the performance of image recognition models. These structures have proven to be successful in improving accuracy, reducing training time, and enhancing interpretability. More recently, the introduction of vision Transformer (ViT) has further improved the network used to extract image features [4]. ViT decomposes images into multiple patches and processes them through multiple Transformer layers, allowing the network to capture global context and long-term dependencies of images. Furthermore, a self-attention mechanism allows the model to focus on the most important regions of images, further improving its efficiency and accuracy.

However, another important aspect of the image recognition task is seldom mentioned, i.e., how to effectively classify the extracted features. Most of the aforementioned studies have focused on using multi-layer perceptron (MLP) structures for feature classification. Despite their simplicity and effectiveness, MLPs still have limitations, such as excessive parameter requirements and susceptibility to overfitting [5]. They are also less suitable for handling high-dimensional feature vectors in large-scale image recognition tasks. Inspired by the evolution of visual feature extraction networks, developing more efficient and biologically interpretable classification networks has the potential to significantly improve image recognition accuracy. Thus, opening up new possibilities for computer vision applications.

Artificial neurons, inspired by their biological counterparts, play a

crucial role in shaping neural networks. The initial McCulloch Pitt's model used a simple linear threshold function for computation [6]. MLP architectures later addressed the issue of linear inseparability [7], and spiking neural networks introduced discrete pulse signals to improve computational efficiency [8]. However, there still exists a considerable accuracy gap between current artificial neurons and biological neurons. Recently, dendritic neurons, drawing inspiration from neuroscience, have emerged as promising alternatives. With their architecture incorporating synapse, dendrite, and soma layers, dendritic networks enhance biological interpretability and exhibit superior performance in challenging tasks [5].

In this study, we propose a novel neural network architecture that combines two biologically interpretable networks for neuroscientifically aligned image recognition. To ensure practicality, we carefully design the synapse, dendrite, and soma layers of the dendritic neuron as an artificial neuron model. By integrating the Vision Transformer with our proposed dendritic network, we create DVT, a highly interpretable network. Extensive experiments on multiple benchmarks demonstrate the significant performance improvements achieved by DVT compared to state-of-the-art methods in image recognition. The accuracy results in Fig. 1 depict the performance of peer models on the CIFAR dataset, without pre-training weights. These findings indicate the potential of DVT to advance computer vision and deepen our understanding of visual perception mechanisms.
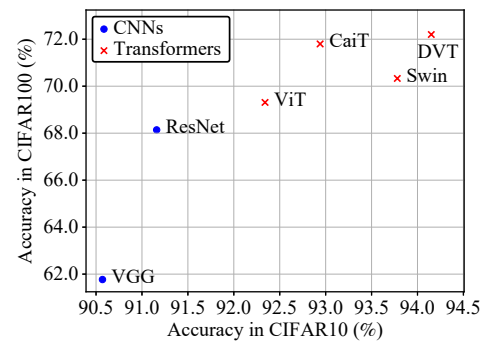

Fig. 1. The accuracy comparison in CIFAR.

Related work:

1) Vision Transformer: It, a novel class of image feature extraction networks, overcomes the limitations of traditional convolutional operators by using long-term dependency-based self-attention to extract spatial features [4]. However, despite their effectiveness in capturing global features, they still face issues such as computational complexity [9], sensitivity to hyper-parameters [10], and data dependency [11]. Besides, enhancing their expressiveness, particularly on lower-resolution datasets, remains a significant challenge.

2) Dendritic network: Inspired by the structure and function of retinal ganglion cells [12], the dendritic network has been proposed as a more biologically plausible artificial neuron [5]. It has shown remarkable results in various kinds of problems [13], [14]. However, its sophisticated architecture requires efficient learning algorithms to improve its performance [5], which presents an obstacle to its further development. In light of this, we aim to optimize the architecture of the dendritic network to enhance its practical performance. Specifically, we reinvent its synapse, dendrite, and soma layer to improve its learning stability and performance. It makes the dendritic network more practical for image recognition.

Methodology: In this study, we propose DVT, a dendritic learning-incorporated vision Transformer, aimed at enhancing the performance and interpretability of image recognition tasks. DVT combines two essential components: a vision Transformer featuring embedded attention mechanisms and a dendritic network mirroring real neuronal architecture. The vision Transformer extracts more comprehensive and representative image features, while the dendritic network ensures accurate feature classification. The overall
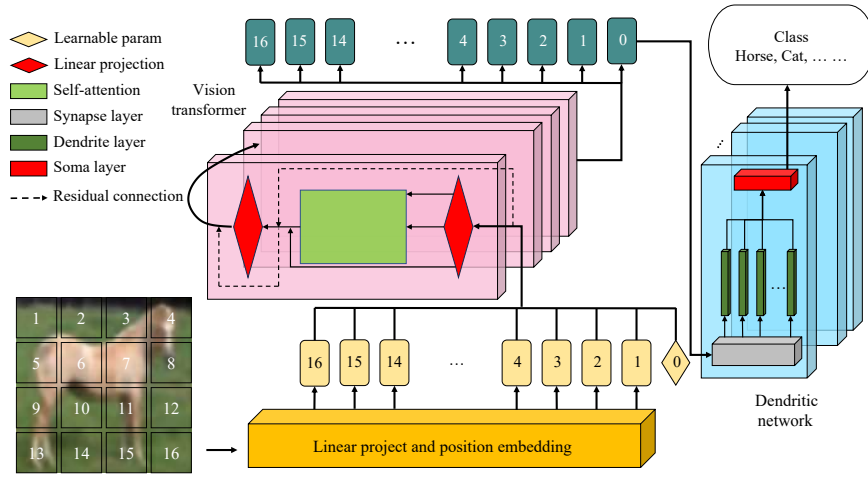
Fig. 2. The framework of DVT to recognize image.

framework of DVT is depicted in Fig. 2. Initially, the input image is sliced into smaller patches, and linear projection and position encoding are applied to each patch to preserve spatial information and optimize computational efficiency. These processed feature maps are then fed into multiple stacked Transformer blocks, wherein the self-attention mechanism enables the network to selectively focus on pertinent information and suppress irrelevant noise. Through continuous fusion and amplification of receptive fields, DVT efficiently extracts highly representative features from the entire image. The dendritic network, comprising three biomimetic layers (synapse, dendrite, and soma), takes charge of the final feature classification. To enhance its convergence·, we introduce a feature normalization operation that reduces feature discrepancies and overall improves network performance.

1) Self-attention mechanism: It plays a crucial role in visual feature extraction by incorporating local and global information to obtain more representative features [4]. By employing multi-head self-attention and stacking multiple Transformer blocks, the robustness of feature extraction is enhanced. However, the computational cost of training a vision Transformer from scratch remains a challenge, particularly when dealing with small-sized datasets. To mitigate this issue, we propose integrating locality self-attention into DVT, drawing inspiration from [15]. This modification effectively captures locally-focused attention contextual information through a self-masking matrix $m \in \mathbb{R}^{h \times n \times n}$ and a learnable parameter γ. This adaptation improves the efficiency of DVT without compromising its ability to capture relevant local information. Its formula is following:

$$z = A(q,k,v) = \delta(m \odot \frac{qk^T}{\sqrt{\gamma}})v \tag{1}$$

$$m = J_n - \infty I_n \tag{2}$$

where q, k, and $v \in \mathbb{R}^{h \times n \times d}$ are different feature vectors that obtained by linear projection of input data x. Self-attention $A(\cdot)$ integrates them via scaled dot-product. m is a all-ones matrix with negative infinity eigenvalues. It is added to A to further deepen the ability of the network to capture global features.

2) Dendritic network: Extensive neuroscience research has unequivocally demonstrated the irreplaceable nature of the theoretical model of dendritic nerves. Moreover, numerous experiments conducted in the field of information science have consistently showcased the remarkable capability of dendritic networks in effectively addressing nonlinear problems. Building upon this knowledge, in our proposed DVT, we integrate a feature normalization operation $\eta(\cdot)$ into the dendritic network, thus aligning it more closely with the practical requirements of real-world engineering applications, i.e.,

$$y^k = \sum_{i=1}^{m} \sum_{j=1}^{d} \delta(\eta(w_{i,j}^k \eta(x)^j + b_{i,j}^k)) \tag{3}$$

$$y = [y^1, y^2, \dots, y^c] \tag{4}$$

$$\eta(x) = \frac{x - \bar{x}}{\sqrt{\sigma(x) + \epsilon}} \theta + \lambda \tag{5}$$

where x is the input feature of d dimension and y^k is the predicted probability of the tth classification target by the network. c is the number of classes. First, normalized inputs are mapped to m dendritic branches through m sets of learnable parameters w and b, a process called synaptic connection. Then, feature normalization and softmax activation function $\delta(\cdot)$ are performed on each branch. Finally, soma layer conception in neuroscience is applied in the network to integrate all dendrites into the result. Notably, each y is associated with one dendritic neuron following (3), and the synapses on each branch are independent for each neuron. Such mutually exclusive connections are considered to be ubiquitous in neuroscience [16], and they have also been proved to be the basis of efficient network inference [17]. In proposed feature normalization $\eta(\cdot)$, θ and λ are learnable parameters, ϵ is constant to prevent the denominator from being 0, $\bar{x}$ and $\sigma(x)$ are mean and variance of x, respectively.

Experiment:

1) Dendrite and learning rate analysis: The number of dendrite branches directly influences the network's ability to approximate the objective function accurately. Similarly, the learning rate significantly impacts the network's adaptability and learning capacity. In this study, we comprehensively analyze these hyper-parameters to determine the optimal DVT configuration. We establish a fair baseline by comparing our findings to the original ViT. To ensure a consistent evaluation, all methods are trained for 100 epochs using the AdamW optimizer. CIFAR10 is chosen as the dataset for this experiment due to its universality in image recognition and the diversity of images it contains. Fig. 3 presents the results, where the number of dendritic branches is denoted as 0, representing the original ViT. Remarkably, incorporating the dendritic network significantly improves the performance of DVT across various learning rates. Through extensive experimentation with different numbers of dendritic branches (ranging from 2 to 64), we observe that increasing the number of branches leads to better results. For optimal prediction outcomes across different problem domains, we recommend setting the number of branches within the range of 8 to 32. Additionally, we find that a learning rate of 0.003 yields favorable outcomes for DVT.

2) Performance comparison: We evaluate the performance of DVT against state-of-the-art methods using four widely recognized and challenging datasets: SVHN, CIFAR10, CIFAR100, and Tiny-ImageNet. The compared methods include VGG19 [18], ResNet50 [2], ViT [4], Swin [9], and CaiT [11], covering the classic CNN architecture and the latest Transformer-based neural network. Notably, the biological interpretability of these methods has shown a gradual increase, transitioning from CNN to Transformer models. The results presented in Table 1 demonstrate the clear superiority of DVT over its peers. Notably, the advantages of DVT become more pronounced as the classification difficulty of the datasets increases, reinforcing our conclusion that DVT excels at approximating complex target
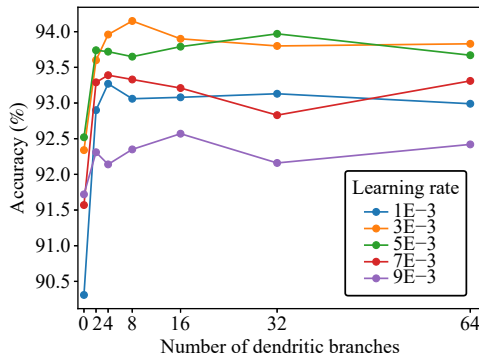
Fig. 3. The analysis of dendrite and learning rate in CIFAR10.

Table 1. Accuracy Comparison on Four Datasets

	SVHN	CIFAR10	CIFAR100	ImageNet
VGG19	96.98	90.57	61.77	41.77
ResNet50	97.30	91.16	68.14	54.38
ViT	97.31	93.01	69.31	45.14
Swin	97.42	93.78	70.33	48.13
CaiT	97.66	92.94	71.80	54.66
DVT	**97.72**	**94.15**	**72.20**	**54.78**

functions. Furthermore, the accuracy of each model increases as its interpretability improves, highlighting the advantage of employing biologically interpretable models for image recognition problems.

3) *Ablation study:* We delve deeper into the spatial and temporal complexity of DVT. It exhibits an increased number of learnable parameters and higher FLOPs compared to the original ViT. More specifically, we introduce a linear layer between the extracted features and the classification outcomes. As presented in Table 2, the sizes of the added linear layers are 160 and 576, respectively. The parameters of the ViT-160 and the FLOP of the ViT-576 are almost similar to those in the DVT. This allows us to perform comparisons to highlight the performance advantages of DVT. The backbone networks of all model architectures are the same. Therefore, only learnable parameters and FLOPs of their classification networks are counted. Notably, a simple stacking of linear layers and increasing their size not only fails to enhance accuracy but also leads to degraded network performance due to heightened learning difficulty. In contrast, DVT relies on a sophisticated architecture to perform efficient calculations with fewer parameters, thereby enhancing network performance.

Table 2. Ablation Study On IFAR10

Architecture	Params	FLOPs	Accuracy
ViT	1.92 K	1.93 K	93.01
ViT-160	32.49 K	32.32 K	92.92
ViT-576	117.32 K	117.31 K	92.68
DVT	34.19 K	108.58 K	94.15

Conclusion: In this study, we introduce DVT, a dendritic learning-incorporated vision Transformer, specifically designed for universal image recognition tasks inspired by dendritic neurons in neuroscience. The incorporation of a highly biologically interpretable dendritic architecture enables DVT to excel in handling complex nonlinear classification problems. Our experimental results highlight the substantial improvement achieved by DVT compared to the current state-of-the-art methods on four general datasets. Moreover, these findings affirm our hypothesis that networks with high biological interpretability in architecture also exhibit superior performance in image recognition tasks.

References

[1] S. Minaee, Y. Boykov, F. Porikli, A. Plaza, N. Kehtarnavaz, and D. Terzopoulos, "Image segmentation using deep learning: A survey," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 44, no. 7, pp. 3523–3542, 2021.

[2] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[3] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2017, pp. 4700–4708.

[4] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16×16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learning Representations*, 2021.

[5] S. Gao, M. Zhou, Y. Wang, J. Cheng, H. Yachi, and J. Wang, "Dendritic neuron model with effective learning algorithms for classification, approximation, and prediction," *IEEE Trans. Neural Networks and Learning Systems*, vol. 30, no. 2, pp. 601–614, 2019.

[6] W. S. McCulloch and W. Pitts, "A logical calculus of the ideas immanent in nervous activity," *The Bulletin of Mathematical Biophysics*, vol. 5, pp. 115–133, 1943.

[7] D. T. Tran, S. Kiranyaz, M. Gabbouj, and A. Iosifidis, "Heterogeneous multilayer generalized operational perceptron," *IEEE Trans. Neural Networks and Learning Systems*, vol. 31, no. 3, pp. 710–724, 2019.

[8] A. Taherkhani, A. Belatreche, Y. Li, G. Cosma, L. P. Maguire, and T. M. McGinnity, "A review of learning in biologically plausible spiking neural networks," *Neural Networks*, vol. 122, pp. 253–272, 2020.

[9] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Computer Vision*, 2021, pp. 10012–10022.

[10] T. Xiao, M. Singh, E. Mintun, T. Darrell, P. Dollár, and R. Girshick, "Early convolutions help transformers see better," *Advances in Neural Information Processing Systems*, vol. 34, pp. 30392–30400, 2021.

[11] H. Touvron, M. Cord, A. Sablayrolles, G. Synnaeve, and H. Jégou, "Going deeper with image transformers," in *Proc. IEEE/CVF Int. Conf. Computer Vision*, 2021, pp. 32–42.

[12] C. Koch, T. Poggio, and V. Torre, "Retinal ganglion cells: A functional interpretation of dendritic morphology," *Philosophical Transa. Royal Society of London*, vol. 298, no. 1090, pp. 227–263, 1982.

[13] Y. Yu, Z. Lei, Y. Wang, T. Zhang, C. Peng, and S. Gao, "Improving dendritic neuron model with dynamic scale-free network-based differential evolution," *IEEE/CAA J. Automa. Sinica*, vol. 9, no. 1, pp. 99–110, 2021.

[14] H. He, S. Gao, T. Jin, S. Sato, and X. Zhang, "A seasonal-trend decomposition-based dendritic neuron model for financial time series prediction," *Applied Soft Computing*, vol. 108, p. 107488, 2021.

[15] S. Lee, S. Lee, and B. Song, "Improving vision transformers to learn small-size dataset from scratch," *IEEE Access*, vol. 10, p. 123, 2022.

[16] R. Yuste, "Dendritic spines and distributed circuits," *Neuron*, vol. 71, no. 5, pp. 772–781, 2011.

[17] X. Wu, X. Liu, W. Li, and Q. Wu, "Improved expressivity through dendritic neural networks," *Advances in Neural Information Processing Systems*, vol. 31, pp. 8057–8068, 2018.

[18] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learning Representations*, 2015.