Theory and Practice of Science and Technology
2026, VOL. 7, NO. 3, 100-103
DOI: 10.47297/taposatWSP2633-456920.20260703

Research on Static Structured Pruning of Vision Transformers Based on L1 Norm and Uniform Layer-wise Constraint: A Case Study of DeiT

Siyu Chen

Sun Yat-sen University, School of Science, Department of Physics, Shenzhen, Guangdong, 518106, China

ABSTRACT

Vision Transformers (e. g., DeiT) have demonstrated exceptional performance in image classification tasks, yet their massive parameter counts heavily limit their deployment on resource-constrained edge devices. Focusing on the fine-grained image classification task (CIFAR-100), this paper investigates the structural redundancy within the multi-head attention mechanism of the DeiT-small model. We propose a static structured pruning method based on the L_1 norm combined with a uniform layer-wise constraint. This approach evaluates the importance of the output projection weights of attention heads statically and removes redundant heads uniformly across each Transformer layer, effectively preventing the tensor dimension mismatch that occurs when all heads in a single layer are pruned. Experimental results indicate that removing 1 attention head per layer (16.7% globally) reduces the parameter count by 5.45% (down to 20.52M), while the post-finetuning accuracy reaches 86.12%. When the pruning ratio is scaled to 3 heads per layer (50.0% globally), the parameters are reduced by 16.34% (down to 18.16M), and the accuracy is maintained at 82.08%. This study successfully quantifies the redundancy boundary of attention heads in DeiT for fine-grained tasks, providing an empirical reference for model lightweighting.

KEYWORDS

Vision transformer; Structured pruning; Multi-head attention

1 Introduction

The Vision Transformer (ViT) and its derivative architectures, such as Data-efficient Image Transformers (DeiT), have emerged as baseline models in computer vision [1-2]. These models rely on the Multi-Head Self-Attention (MSA) mechanism to establish global feature representations, which intrinsically require massive parameters (e. g., DeiT-small contains approximately 21.7M parameters). This substantial computational and storage overhead poses a significant barrier to deployment on edge devices.

Structured pruning reduces parameters and FLOPs by directly eliminating specific network components, such as attention heads. However, when processing fine-grained image classification tasks (e.g., CIFAR-100), the model often relies on specific attention heads to capture localized detailed features. Furthermore, global magnitude-based pruning strategies often encounter physical risks at high pruning rates: an entire layer of attention heads might be removed, leading to dimension mismatch in residual connections and subsequent structural breakdown.

To address these architectural conflicts, this paper explores the redundancy of DeiT in fine-grained tasks and implements a static structured pruning strategy based on the L_1 norm and uniform layer-wise constraints. The static L_1 norm is employed to evaluate head importance without extra computational overhead [4-5]. Simultaneously, by restricting the maximum number of pruned heads per layer, the structural integrity of the network's information flow is strictly maintained.

2 Related Work

2.1 Vision Transformers and DeiT

ViT divides images into fixed-size patches and feeds them into a standard Transformer encoder, proving the feasibility of pure self-attention architectures in vision tasks [1,6]. Building upon this, DeiT introduces a knowledge distillation mechanism and more efficient training strategies, improving convergence performance without relying on ultra-large-scale datasets *e.g.*,$JFT-300M$ [2]. Despite optimizing data efficiency, the parameter scale of its fundamental components (MSA and MLP) remains unchanged.

2.2 Attention Head Redundancy and Structured Pruning

Extensive redundancy exists within the multi-head attention mechanism. Michel et al. demonstrated that removing a large number of attention heads during inference does not significantly degrade the final performance of Transformers [3]. Existing pruning evaluation metrics include first-order gradient estimation, dynamic entropy evaluation, and static magnitude evaluation [4]. Considering computational overhead constraints, the static evaluation method based on the L_1 norm weight magnitude demonstrates high execution efficiency, as it bypasses the need for forward and backward propagation [5].

2.3 Parallel Paradigms in Model Compression and Edge Deployment Constraints

To comprehensively contextualize the necessity of static structured pruning, it is crucial to analyze the physical constraints and algorithmic limitations of alternative model compression paradigms, namely unstructured pruning, network quantization, and knowledge distillation.

Unstructured Pruning and Hardware Misalignment: Unstructured pruning aims to induce sparsity by zeroing out individual weight parameters based on specific criteria, such as absolute magnitude thresholds or Taylor expansion approximations. While this paradigm can theoretically achieve extremely high pruning ratios (e.g., removing over 90% of parameters) with minimal accuracy degradation, it fundamentally generates unstructured sparse matrices. In physical edge deployment scenarios, standard hardware accelerators, such as Graphical Processing Units (GPUs) or Neural Processing Units (NPUs), rely on highly optimized dense matrix multiplication routines (e. g., cuBLAS). Consequently, without specialized sparse computing architectures, the theoretical reduction in floating-point operations (FLOPs) does not translate into a proportional reduction in inference latency. In many practical cases, the additional memory indexing overhead required to store and process sparse formats (such as Compressed Sparse Row, CSR) can exacerbate memory bandwidth bottlenecks, rendering unstructured pruning highly inefficient for off-the-shelf edge devices.

Network Quantization and Activation Sensitivity: Quantization reduces the physical bit-width of model weights and activations, typically compressing 32-bit floating-point (FP32) representations to 16-bit (FP16) or 8-bit integers (INT8). While highly effective for traditional Convolutional Neural Networks (CNNs), the direct application of Post-Training Quantization (PTQ) to Vision Transformers frequently triggers severe performance collapse. This vulnerability stems from the extreme variations and outliers in the activation ranges of the Multi-Head Self-Attention mechanism. Specifically, the attention probability maps computed via $softmax(QK^T/\sqrt{d_k})$ often exhibit long-tailed distributions, making linear mapping to low-bit integers highly lossy. To recover baseline accuracy, complex Quantization-Aware Training (QAT) pipelines must be introduced, which significantly increases the engineering complexity and computational training overhead without altering the topological macro-structure of the model.

Knowledge Distillation and Architectural Stagnation: Knowledge Distillation (KD) involves transferring the generalized representational capacity from a cumbersome, high-capacity "Teacher" model to a compact "Student" model by minimizing the Kullback-Leibler (KL) divergence between their output logits or intermediate feature maps. Although DeiT itself heavily utilizes token-based distillation during its pre-training phase, KD is fundamentally a training strategy rather than a structural optimization technique. It enhances the parameter efficiency of a given architecture but cannot physically alter the tensor dimensions or the computational graph of the Student model. If the baseline Student architecture (e.g., DeiT-small) already exceeds the SRAM capacity or latency threshold of a specific edge device, KD alone is physically incapable of breaching that deployment barrier.

The Imperative for Static Structured Pruning: Synthesizing the aforementioned hardware and algorithmic constraints, static structured pruning emerges as the most deterministic pathway for deploying Vision Transformers on edge terminals. By physically stripping entire functional blocks—specifically, the redundant Query, Key, Value, and Output projection matrices associated with specific attention heads—this approach strictly maintains the dense matrix format. This physical structural reduction ensures seamless alignment with standard hardware acceleration libraries, delivering a proportional and immediate reduction in both memory footprint and physical inference latency, thereby justifying its selection as the core methodology of this study.

3 Methodology

3.1 Underlying Mathematical Mapping of Vision Transformer Architecture and Multi-Head Attention Mechanism

Before delving into the evaluation metrics and physical stripping mechanisms of structured pruning, it is imperative to clarify the information flow essence of the DeiT architecture from the underlying logic of tensor dimensions. DeiT inherits the topological structure of the standard Vision Transformer (ViT). Its core concept is to abandon localized receptive fields and construct feature representations through global attention. The computational pipeline of the entire architecture can be decomposed into three sequential physical stages: Patch Embedding, Multi-Head Self-Attention (MSA) encoding, and Multi-Layer Perceptron (MLP) feature mapping.

3.1.1 Patch Embedding and Positional Encoding Tensors

For any given input image tensor $x \in \mathrm{R}^{H\times W\times C}$ (where H and W are the spatial resolution, and C is the number of channels, typically $C = 3$), the model first rigidly divides it spatially into a sequence of non-overlapping 2D patches. Let the resolution of each patch be $P \times P$; the total number of generated patches is $N = \frac{HW}{P^2}$. These 2D patches are flattened into a 1D sequence $x_p \in \mathrm{R}^{N\times(P^2\cdot C)}$.

Subsequently, a learnable linear projection matrix $E \in \mathrm{R}^{(P^2\cdot C)\times D}$ maps it to a unified hidden dimension D (in DeiT-small,

$D = 384$). To introduce the classification objective and retain the global topological position information of the image, an additional Class Token $x_{class} \in \mathrm{R}^{1 \times D}$ is appended to the head of the sequence, and a 1D learnable positional encoding matrix $E_{pos} \in \mathrm{R}^{(N+1) \times D}$ is added element-wise. The construction formula for the initial input feature sequence z_0 is as follows:

$$z_0 = [x_{class}; x_p^1 E; x_p^2 E; \cdots; x_p^N E] + E_{pos}$$

3.1.2 Computational Graph of the Single-Head Self-Attention Mechanism

The sequence z_0 sequentially passes through L deeply stacked Transformer encoder layers. In the l-th layer, the core feature interaction is driven by the self-attention mechanism. For a single attention head, the input tensor X (i.e., the output z_{l-1} from the previous layer after LayerNorm) is projected into three independent orthogonal subspaces to generate the Query, Key, and Value matrices:

$$Q = XW^Q, K = XW^K, V = XW^V$$

where the projection weight matrices are $W^Q, W^K, W^V \in \mathrm{R}^{D \times d_k}$, and d_k is the channel dimension of a single attention head.

The physical significance of self-attention lies in calculating the global correlation of all element features in the sequence. Its measurement adopts the Scaled Dot-Product; that is, by calculating the inner product of Q and K, dividing by the scaling factor $\sqrt{d_k}$ to prevent vanishing or exploding gradients, and finally applying the Softmax function to obtain the normalized attention probability distribution matrix. This distribution is then used to perform a weighted recombination of the feature values V:

$$Attention(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

3.1.3 Multi-Head Mechanism (MSA) and Physical Constraints of the Output Projection Matrix

To enable the model to jointly attend to information from different representation subspaces at different positions, DeiT employs a multi-head attention mechanism. Assuming a single layer contains H independent heads (in DeiT-small, $H = 6, d_k = D/H = 64$), the independent computational path for the h-th head is:

$$\mathrm{h}ead_{\mathrm{h}} = Attention(XW_{\mathrm{h}}^Q, XW_{\mathrm{h}}^K, XW_{\mathrm{h}}^V)$$

After the output tensors of each independent head are concatenated along the feature dimension, they must undergo a linear transformation through a global Output Projection Matrix $W^O \in \mathrm{R}^{D \times D}$ to map them back to the original dimensional space, thereby participating in the subsequent Residual Connection:

$$MSA(X) = Concat(\mathrm{h}ead_1, \mathrm{h}ead_2, \cdots, \mathrm{h}ead_H)W^O$$

3.1.4 Multi-Layer Perceptron (MLP) Feature Dimensionality Expansion

After the attention module establishes the global correlation of features, each token must undergo non-linear feature mapping through a Multi-Layer Perceptron (MLP). The MLP module of DeiT consists of two fully connected layers with an intermediate dimension expansion (typically by a factor of 4), and its non-linear activation function employs the GELU mechanism:

$$MLP(x) = GELU(xW_1 + b_1)W_2 + b_2$$

where $W_1 \in \mathrm{R}^{D \times 4D}$ and $W_2 \in \mathrm{R}^{4D \times D}$. Through the dual processing of MSA and MLP, coupled with residual superposition, the forward propagation of features in a single Transformer layer is completed.

The aforementioned concatenation and projection formula of $\text{MSA}(X)$ not only establishes the foundation for information transfer within the network but also constitutes the critical physical prerequisite for executing static pruning in this paper: regardless of how the attention weight matrices are distributed within a single head_h, the absolute intensity of its extracted features before entering the main feature stream is directly and physically constrained by the magnitude of the parameters in the corresponding slice of the output matrix W^O. This provides a rigorous mathematical basis for evaluating head importance based on the L_1 norm.

3.2 Uniform Layer-wise Pruning Strategy

Global sorting pruning at a high pruning rate can easily eliminate all attention heads in a specific layer (i.e., the number of surviving heads is 0). This causes the tensor feature dimension input to the Multi-Layer Perceptron (MLP) to collapse to zero, triggering matrix multiplication errors at the residual connections.

To circumvent this architectural calculation error, a uniform layer-wise constraint strategy is designed. The pruning action is strictly limited within a single Transformer layer:

Intra-layer Sorting: Calculate $S(l,h)$ for all 6 heads in the l-th layer and sort them in ascending order.

Uniform Removal: Set an absolute number of heads to be pruned per layer as N_{prune} $N_{prune} \in 1,2,3$. In each layer, the Query, Key, Value, and Output weight slices corresponding to the N_{prune} heads with the lowest scores are physically removed. This mechanism ensures that at least $(6 - N_{prune})$ active heads are retained in each layer, thereby maintaining the consistency of the underlying dimensions of DeiT.

4 Experiments and Results

4.1 Experimental Setup

The evaluation is conducted on the CIFAR-100 fine-grained image classification dataset. The hardware platform utilizes a single RTX 4070 Ti Super (16GB) GPU. The experimental pipeline consists of three stages:

4.1.1 Baseline Construction

The pre-trained DeiT-small model is loaded and fine-tuned on CIFAR-100 for 20 epochs to obtain the baseline model.

4.1.2 Static Pruning

Using the baseline model as input, uniform layer-wise pruning with $N_{prune} = 1,2,3$ is executed respectively to generate three sets of model architectures with reduced volumes.

4.1.3 Finetune-to-Recover

The pruned models are re-trained on CIFAR-100 for 20 epochs with a constant learning rate of 3×10^{-5} to rebuild feature mapping channels and record the validation accuracy.

4.2 Redundancy Boundary Analysis

The experimental data, as shown in Table 1, reveals the parameter reduction and accuracy relationship of DeiT-small in fine-grained visual tasks.

Table 1 Structure and Accuracy Relationship under Different Pruning Ratios in DeiT-small

Model State	Pruned Heads per Layer (N_{prune})	Total Parameters (M)	Parameter Drop Ratio	Validation Accuracy
Baseline	0	21.70	0.00%	85.31%
Group 1	1	20.52	5.45%	86.12%
Group 2	2	19.34	10.89%	84.44%
Group 3	3	18.16	16.34%	82.08%

The total parameter count of the unpruned baseline model is 21.70M.

When $N_{prune} = 1$ (removing 16.7% of attention heads globally), the total parameter count drops to 20.52M (a decrease of 5.45%). After recovery finetuning, the validation accuracy reaches 86.12%. Notably, the accuracy of the pruned and finetuned model (86.12%) surpasses the unpruned baseline (85.31%). This phenomenon can be attributed to the regularization effect of moderate pruning, which mitigates overfitting on the smaller CIFAR-100 dataset.

When $N_{prune} = 2$ (removing 33.3% of attention heads), the parameter count drops to 19.34M (a decrease of 10.89%), and the finetuned accuracy is 84.44%.

When $N_{prune} = 3$ (removing 50.0% of attention heads), the parameters are reduced to 18.16M (a decrease of 16.34%). At this point, the feature extraction channels are significantly restricted. After recovery finetuning, the accuracy drops to 82.08%. This data establishes the limit of parameter reduction achieved solely by pruning attention heads under the current mechanism.

5 Conclusion

This experiment tested the mechanism of the pruning strategy based on the L_1 norm and uniform layer-wise constraint on the DeiT architecture. The empirical data verifies the existence of separable structural redundancy within the multi-head attention mechanism. Under the setting of removing 50% of the attention heads in a single layer, the parameter count decreased by 16.34%, while the recovery finetuning accuracy was maintained at 82.08%. This variable relationship clarifies the parameter reduction boundary achievable by solely removing attention heads. To further decrease the overall volume of the model, a prerequisite is the introduction of dimensionality reduction mechanisms targeting the Multi-Layer Perceptron (MLP) layers.

References

[1] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... & Houlsby, N. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. ICLR.

[2] Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., & Jégou, H. (2021). Training data-efficient image transformers & distillation through attention. International Conference on Machine Learning.

[3] Michel, P., Levy, O., & Neubig, G. (2019). Are sixteen heads really better than one?. Advances in Neural Information Processing Systems.

[4] Han, S., Pool, J., Tran, J., & Dally, W. (2015). Learning both weights and connections for efficient neural network. Advances in Neural Information Processing Systems.

[5] Li, H., Kadav, A., Durdanovic, I., Samet, H., & Graf, H. P. (2016). Pruning filters for efficient convnets. ICLR.

[6] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems.