Tailored ViT Slimming: Budget-Aware Multi-Dimensional Sparsity Regularization for Vision Transformers Pruning (Student Abstract)

Suwoong Lee[1,2], **Seungjae Lee**[2], **Yunho Jeon**[3*], **Junmo Kim**[1*]
[1]KAIST, 291 Daehak-ro, Yuseong-gu, Daejeon, Republic of Korea
[2]ETRI, 218 Gajeong-ro, Yuseong-gu, Daejeon, Republic of Korea
[3]Hanbat National University, 109 Jiphyeonbuk-ro, Sejong, Republic of Korea
leesw@kaist.ac.kr, seungjlee@etri.re.kr, yhjeon@hanbat.ac.kr, junmo.kim@kaist.ac.kr

Abstract

We propose Tailored ViT Slimming (TVS), a budget-aware multi-dimensional pruning framework for Vision Transformers. TVS injects learnable masks into MHSA and MLP modules and applies adaptive non-convex sparsity regularization to achieve maximal utilization of parameters under strict module-wise budgets. In addition, by retaining scaled masks after pruning, TVS avoids abrupt accuracy drops and provides stable initialization for fine-tuning. On ImageNet-1k with DeiT-S and DeiT-B, TVS consistently outperforms prior ViT compression methods. This result empirically shows that the non-convex sparsity regularizer is effective not only in CNNs but also in ViTs.

Introduction

Vision transformers have shown strong performance in various vision applications, but their large number of parameters and computational complexity make them challenging to apply in real-world applications. ViT-Slim (Chavan et al. 2022) adopts sparsity regularization to Vision transformers, enabling multidimensional pruning across various components of ViTs. However, it suffers from inefficiency due to fixed-form L_1 regularization that is independent of the budget. Recently, Tailored Channel Pruning (TCP) (Lee et al. 2025) successfully prunes convolutional neural networks using budget-aware non-convex sparsity regularization.

Inspired by TCP, we propose *Tailored ViT Slimming* (TVS), which applies budget-aware adaptive sparsity regularization to Vision Transformer pruning, aiming to achieve a better accuracy–FLOPs trade-off. TVS introduces learnable masks for heads in Multi-Head Self-Attention(MHSA) layers and MLP layers, and applies adaptive L_p regularization calculated from the target budget for maximally utilizing the parameters while keeping a strict budget constraint. Experiments on ImageNet-1k with DeiT-S and DeiT-B show that TVS outperforms prior ViT pruning methods.

Proposed Method

Masked Computation in Transformer Modules

Given a pre-trained Vision Transformer, we inject learnable sparsity masks to intermediate tensors in all major transformer modules, following the masking formulation introduced in (Chavan et al. 2022). In the MHSA module, masks $\mathbf{z}^A_{l,h}$ are applied to the query, key, and value projections, thereby pruning less important feature dimensions from attention computation. In the MLP module, a mask $\mathbf{z}^M_l$ is applied after the first linear projection to remove unimportant hidden units.

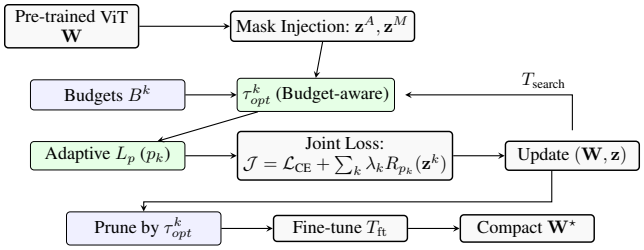

Figure 1: Overall pipeline of TVS. Masks $\mathbf{z}^A$ and $\mathbf{z}^M$ are injected, budgets B^k set module-wise optimal thresholds τ^k_{opt}, adaptive L_p adjusts gradients around τ^k_{opt}, then joint optimization is performed. After sparsity regularization, pruning and fine-tuning yield a compact ViT.

MHSA Masking: For head (l, h) with intermediate tensors $\mathbf{t}^Q_{l,h}, \mathbf{t}^K_{l,h}, \mathbf{t}^V_{l,h} \in \mathbb{R}^{N_l \times d_h}$, the masked Q, K, and V projections are:

$$\tilde{\mathbf{t}}^Q_{l,h} = \mathbf{z}^A_{l,h} \odot \mathbf{t}^Q_{l,h} \tag{1}$$

$$\tilde{\mathbf{t}}^K_{l,h} = \mathbf{z}^A_{l,h} \odot \mathbf{t}^K_{l,h} \tag{2}$$

$$\tilde{\mathbf{t}}^V_{l,h} = \mathbf{z}^A_{l,h} \odot \mathbf{t}^V_{l,h} \tag{3}$$

The masked attention output in the MHSA layer is:

$$\mathbf{t}^A_{l,h} = \text{Softmax}\left(\frac{\tilde{\mathbf{t}}^Q_{l,h}(\tilde{\mathbf{t}}^K_{l,h})^\top}{\sqrt{d_h}}\right)\tilde{\mathbf{t}}^V_{l,h} \tag{4}$$

MLP Masking: For the intermediate tensor $\mathbf{t}^M_l \in \mathbb{R}^{N_l \times M_l}$ after the first linear projection in the MLP, the masked hidden units are:

$$\tilde{\mathbf{t}}^M_l = \mathbf{z}^M_l \odot \mathbf{t}^M_l \tag{5}$$

where $\mathbf{t}$ denote an *intermediate tensor* produced within the network before masking and $\tilde{\mathbf{t}}$ denote masked result of $\mathbf{t}$ by a mask $\mathbf{z}$.

*Joint corresponding authors.

Adaptive Non-Convex Sparsity Regularization

Let z_i be the i-th scalar entry in the concatenated vector of all mask elements $\mathbf{z}$, and let $N = |\mathbf{z}|$ denote the total number of elements in $\mathbf{z}$. We define the remaining parameter ratio at threshold τ as:

$$\text{Ratio}(\tau; \mathbf{z}) = \frac{|\{z_i \in \mathbf{z} \mid |z_i| > \tau\}|}{N}. \quad (6)$$

Here, $\text{Ratio}(\tau; \mathbf{z})$ indicates the fraction of the mask whose magnitude is larger than τ. The optimal threshold τ is computed separately for each module, $\mathbf{z}^k \in \{\mathbf{z}^A, \mathbf{z}^M\}$, and the optimal threshold τ^k_{opt} is computed by:

$$\tau^k_{\text{opt}} = \underset{\tau \in \mathbf{z}^k,\ \text{Ratio}(\tau;\mathbf{z}^k) \le B^k}{\arg\max} \text{Ratio}(\tau; \mathbf{z}^k), \quad (7)$$

where B^k is the target ratio budget (e.g., proportion of remaining parameters) for module $k \in \{A, M\}$.

Instead of convex regularization (e.g., L_1), we adopt non-convex sparsity regularization, which shows sharp gradient near zero and can effectively control the sparsity across different parameter distributions. Following TCP in (Lee et al. 2025), we use the L_p regularizer with fractional p ($0 < p < 1$) on each module's mask:

$$R_p(\mathbf{z}^k) = \sum_i |z^k_i|^p, \quad 0 < p < 1, \quad k \in \{A, M\} \quad (8)$$

For a given τ^k_{opt}, p is chosen such that the gradient at $z = \tau^k_{\text{opt}}$ equals 1:

$$p^k \cdot (\tau^k_{\text{opt}})^{p^k - 1} = 1 \quad \Rightarrow \quad p^k = \frac{W(\tau^k_{\text{opt}} \cdot \log \tau^k_{\text{opt}})}{\log \tau^k_{\text{opt}}} \quad (9)$$

where $W(\cdot)$ is the Lambert-W function. The detailed derivation can be found in the Appendix.

Pruning and Fine-tuning with Scaled Mask

After the sparsity regularization, we prune the model with a *scaled mask*, which sets the mask elements above the optimal threshold τ^k_{opt} to the final value $\mathbf{z}^k$, not $\mathbf{1}$ as in (Chavan et al. 2022). The mask elements below the optimal threshold τ^k_{opt} are set to zero.

This process ensures that the pruning process only removes unimportant parameters that are near zero, without altering the important parameters. Therefore, the accuracy loss caused by pruning is minimized. We then fine-tune the pruned model for a standard number of epochs to restore the accuracy. The overall pipeline of the proposed method is illustrated in Figure 1.

Experimental Results

Table 1 summarizes the comparison with state-of-the-art ViT compression methods. For DeiT-S, our TVS method consistently outperforms WDPruning (Yu et al. 2022) and ViT-Slim under comparable computational budgets. For the larger DeiT-B model, TVS achieves comparable or superior accuracy to ViT-Slim and OFB (Ye et al. 2024) while requiring fewer parameters and FLOPs. These results highlight that TVS is effective for both small and large ViTs.

Model	#Params (M)	FLOPs (B)	Top-1	Top-5
DeiT-S baseline: 22.1M Params / 4.6B FLOPs				
DeiT-S (Touvron et al. 2021)	22.1	4.6	79.8	95.0
WDPruning (Yu et al. 2022)	-	2.6	78.4	94.1
ViT-Slim (Chavan et al. 2022)	**11.4**	**2.3**	77.9	94.1
TVS (ours)	**11.4**	**2.3**	**78.6**	**94.3**
DeiT-B baseline: 86.6M Params / 17.5B FLOPs				
DeiT-B (Touvron et al. 2021)	86.6	17.5	81.8	95.6
ViT-Slim (Chavan et al. 2022)	52.6	10.6	**82.4**	**96.1**
TVS (ours)	**51.8**	**10.5**	**82.4**	96.0
OFB (Ye et al. 2024)	43.9	8.7	81.7	**95.8**
TVS (ours)	**35.6**	**7.1**	**81.9**	**95.8**

Table 1: Comparison with State-of-the-art ViT compression methods on ImageNet-1k. The best results are highlighted in bold.

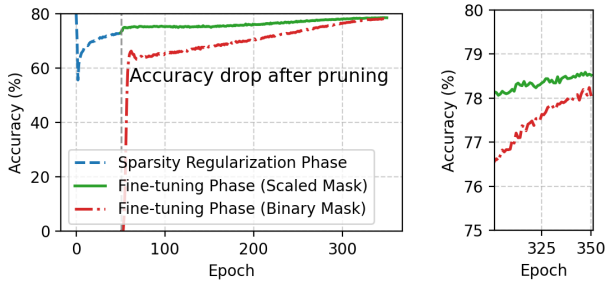

Figure 2: Top-1 Accuracy during Sparsity Regularization (0–50 epoch), Pruning (50 epoch), and Fine-tuning (50–350 epoch) of 50%/50% budget. Left: Whole View, Right: Zoom-in View (Final 50 epochs).

To investigate the effect of keeping the learned mask values during fine-tuning, we compare cases using the binary mask and the scaled mask. Figure 2 compares binary and scaled masks during sparsity regularization, pruning, and fine-tuning. Pruning with a binary mask causes an abrupt accuracy drop due to hard binarization, whereas pruning with a scaled mask preserves a well-trained scale parameter at the pruning point, providing a better initialization for fine-tuning. As training continues, the scaled mask consistently outperforms the binary mask and converges to higher final accuracy, showing its advantage of enabling smoother pruning and more effective fine-tuning.

Conclusion and Future Work

We propose Tailored ViT Slimming (TVS), a budget-aware multi-dimensional sparsity regularization method for Vision Transformers. By utilizing sophisticated budget-aware adaptive non-convex sparsity regularization and retraining with a scaled mask, TVS achieves a state-of-the-art trade-off between accuracy and FLOPs on both small and large Vision Transformers. A possible next research direction is to apply sophisticated sparsity regularization to LLM or VLM for enhancing their efficiency.

Acknowledgments

This work was supported by Culture, Sports and Tourism R&D Program through the Korea Creative Content Agency grant funded by the Ministry of Culture, Sports and Tourism in 2025. (Project Name : Development of AI Agent Technology Based on Artists Unique Characteristics for Interactive Culture Creation, Project Number : RS-2025-02312732, Contribution Rate : 50%), 2025 Cultural Heritage Smart Preservation & Utilization R&D Program of Korea Heritage Service, National Research Institute of Cultural Heritage (Project No.: RS-2024-00396158, Contribution Rate: 30%), and the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (No. RS-2023-00240379, Contribution Rate: 20%).

References

Chavan, A.; Shen, Z.; Liu, Z.; Liu, Z.; Cheng, K.-T.; and Xing, E. P. 2022. Vision transformer slimming: Multi-dimension searching in continuous optimization space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4931–4941.

Lee, S.; Jeon, Y.; Lee, S.; and Kim, J. 2025. Tailored Channel Pruning: Achieve Targeted Model Complexity Through Adaptive Sparsity Regularization. *IEEE Access*.

Touvron, H.; Cord, M.; Douze, M.; Massa, F.; Sablayrolles, A.; and Jégou, H. 2021. DeiT: Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning (ICML)*.

Ye, H.; Yu, C.; Ye, P.; Xia, R.; Tang, Y.; Lu, J.; Chen, T.; and Zhang, B. 2024. Once for both: Single stage of importance and sparsity search for vision transformer compression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5578–5588.

Yu, F.; Huang, K.; Wang, M.; Cheng, Y.; Chu, W.; and Cui, L. 2022. Width & depth pruning for vision transformers. In *Proceedings of the AAAI conference on artificial intelligence*.