Retraining-free Constraint-aware Token Pruning for Vision Transformer on Edge Devices

Yun-Chia Yu*, Mao-Chi Weng*, Ming-Guang Lin†, An-Yeu (Andy) Wu†
*Department of Electrical Engineering, National Taiwan University, Taipei, Taiwan
†Graduate Institute of Electronics Engineering, National Taiwan University, Taipei, Taiwan
chrislin@access.ee.ntu.edu.tw, andywu@ntu.edu.tw

***Abstract*—Vision transformer (ViT) and its variants have demonstrated great potential in various computer vision tasks. However, intensive computation requirements with respect to the token size hinder ViT from being deployed on edge devices with diverse computation resources. Recently, token pruning has been proven to be a promising method to exploit the redundancy of tokens. However, it often requires a laborious retraining process to meet different resource constraints. In this paper, we introduce Fisher information (FI) from tokens to evaluate token importance across different transformer blocks and propose a Retraining-free Constraint-aware Token Pruning (RCTP) framework. RCTP employs a two-step process to obtain the optimal pruning thresholds without retraining under different FLOPs constraints. In the first step, a candidate threshold table and a FLOPs-Fisher table are constructed through a three-stage pipeline to record the trade-off between FLOPs and FI loss of each candidate threshold. In the second step, a modified Viterbi algorithm determines optimal threshold sets with minimum overall FI loss under different FLOPs-constraints in one shot. Our experiment illustrates that RCTP attains better accuracy-FLOPs trade-off than prior pruning-based approaches.**

***Index Terms*—Vision Transformer, edge computing, token pruning, image classification**

I. Introduction

Vision Transformer (ViT) [1] has demonstrated remarkable performance in various computer vision tasks. However, its computational complexity has posed a significant challenge when it comes to deploying it on resource-constrained edge devices. In response to this challenge, token pruning has been a promising research area to address this problem, since ViT exhibits quadratic computation costs with the number of tokens.

Within the realm of token pruning, two main branches have emerged: pruning-based methods and merging-based methods. The former reduces tokens by discarding unimportant tokens, while the latter reduces tokens by merging similar ones. In terms of edge devices, merging-based methods like ToMe [2] need additional hardware to support operations like argsort and token similarity calculations, making them less ideal for edge devices with limited hardware resources.

Given the limitation of edge devices, we opt for pruning-based approaches in this work. Pruning-based methods can be further divided into retraining-based methods and retraining-free ones. Retraining-based methods like learnable token pruning [3] [4] use learned modules to make token pruning decisions. They can be constraint-aware but necessitate a repetitive retraining process when the constraint changes. On the other hand, ATS [5], a retraining-free method, adaptively prunes tokens by sampling over the inverse cumulative distribution function (CDF) of class attention scores. However, it requires manual adjustment of sampling time to meet varying FLOPs constraints.

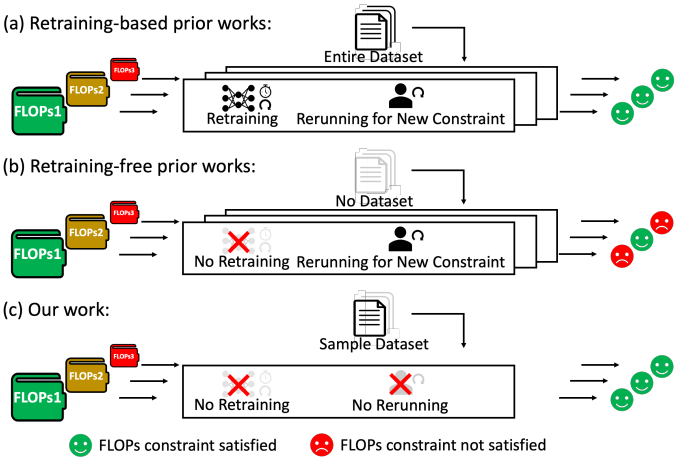


Fig. 1. Comparison between **(a)** retraining-based prior works, **(b)**retraining-free prior works, and **(c)** the proposed RCTP.

In this work, we propose Retraining-free Constraint-aware Token Pruning (RCTP) to address these issues, as shown in Fig. 1. Our approach utilizes Fisher information (FI) to estimate the information loss of candidate thresholds in each transformer block. Subsequently, our modified Viterbi algorithm [6] determines threshold sets with minimum FI loss suitable for specific FLOPs constraints.

Our contribution can be summarized as follows:

- To the best of our knowledge, we are the first to leverage Fisher information to evaluate token importance across different transformer blocks.
- We introduce a retraining-free token pruning method that achieves a competitive FLOPs-accuracy trade-off with prior pruning-based works.
- Our modified Viterbi algorithm enables one-shot optimal threshold set determinations under different FLOPs-constraints.

Yun-Chia Yu and Mao-Chi Weng contributed equally to this work.

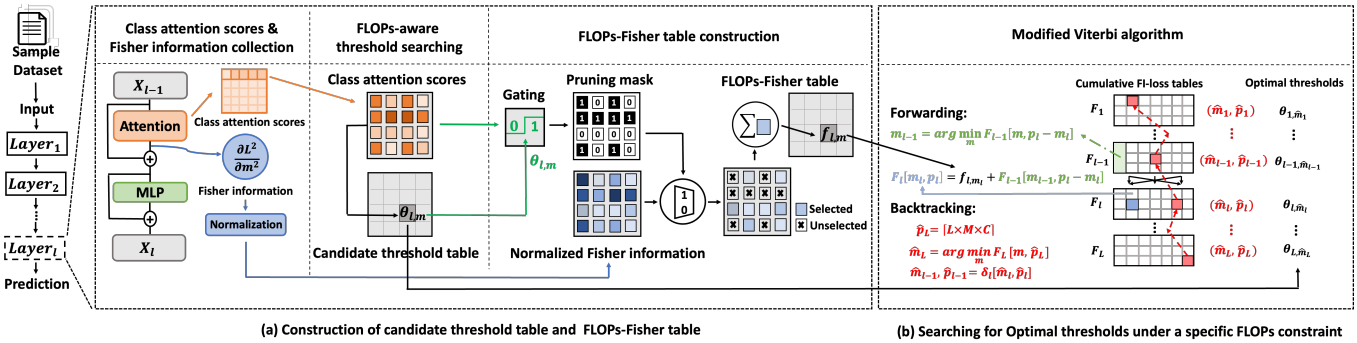


Fig. 2. Overview of the proposed RCTP framework. (a) Construction of candidate threshold table and FLOPs-Fisher table. (b) Searching for optimal threshold set under a specific FLOPs constraint.

II. Background

A. Vision Transformer

Vanilla ViT architecture mainly consists of Multi-head Self-Attention (MSA), Multilayer Perceptron (MLP), layer normalization, and residual connection. Within l-th transformer block, the input X_{l-1} undergoes layer normalization LN_1 and is then sent into MSA module. In MSA, the input X_{l-1} is first linearly transformed into query Q_l, key K_l, and value V_l. Then the output of MSA is calculated as (1):

$$Attention(Q_l, K_l, V_l) = Softmax(Q_l K_l^T / \sqrt{d}) V_l. \quad (1)$$

Lastly, the output of MSA module will conduct residual addition and then proceed through layer normalization LN_2 and MLP module sequentially. The whole process is shown in (2) and (3):

$$\tilde{X}_l = MSA(LN_1(X_{l-1})) + LN_1(X_{l-1}), \quad (2)$$

$$X_l = MLP(LN_2(\tilde{X}_l)) + LN_2(\tilde{X}_l). \quad (3)$$

B. Token pruning

Prior pruning-based works can be divided into two groups: retraining-based approaches and retraining-free ones.

While retraining-based schemes usually demonstrate minimal accuracy drop, they demand considerable time and computing resources for retraining and hyperparameter searching. The most similar retraining-based scheme to our work is learnable token pruning [3] [4] (denoted as **LTP**), which uses a set of learnable thresholds to make token pruning decisions. Given a threshold set $\Theta = \{\theta_l \mid \theta_l \in \mathbb{R}, l = 1, 2, ..., L\}$, where L is the number of transformer blocks, tokens with importance scores below θ_l in l-th transformer block are pruned. While it is possible to obtain these thresholds after a single epoch of fine-tuning, hyperparameter searching remains necessary.

For retraining-free schemes, a naive baseline is Top-K, only preserving K_l tokens with the highest class attention scores after l-th MSA module. In this paper, we use **constant Top-K**, which prunes a constant number of tokens at each block, as our baseline. Another retraining-free scheme is **ATS [5]**, which dynamically reduces tokens by inverse sampling on CDF of class attention scores. However, this inverse sampling operation will cause additional hardware overhead when being deployed on edge devices. Moreover, ATS is not constraint-aware, and hence it cannot be guaranteed to suit edge devices with specific computation resources.

C. Fisher information

The process of pruning can be perceived as applying masks to pruning targets. Liu et al. [7] utilizes Fisher information (FI), which represents the mean of square gradient of the pruning masks, to evaluate the importance of channels within a convolution neural network. Kwon et al. [8] introduces FI to guide the pruning process in transformer, where their pruning targets are heads and filters. For a sample dataset D, the important score I_i of the i-th pruning target can be calculated as (4):

$$I_i = \frac{1}{|D|} \sum_{d \in D} \left(\frac{\partial \mathcal{L}_d}{\partial m_i}\right)^2, \quad (4)$$

where m_i is the i-th mask and $\mathcal{L}_d$ is the loss associated with the sample d. In this work, we conceptualized the threshold-based token pruning process as dynamically applying pruning masks to tokens. Thus, we utilize the FI from tokens pruned by a specific threshold to estimate the FI loss with respect to that threshold.

Our work falls into the category of retraining-free token pruning. While mirroring LTP at the inference stage, we determine optimal threshold sets in a constraint-aware way without the need for retraining.

III. Proposed Retraining-free Constraint-aware Token Pruning for Vision Transformer

A. Overview

We adopt the threshold-based token pruning method, which prunes tokens with class attention scores below a specific threshold. Our goal is to determine optimal threshold sets under certain FLOPs constraints without retraining. This is accomplished through a two-step process, as illustrated in Fig 2. First, our framework takes a sample dataset as input and constructs a candidate threshold table and a FLOPs-Fisher table. These tables respectively record the candidate thresholds and the estimation of the Fisher information (FI) loss

corresponding to specific FLOPs reduction ratios in different transformer blocks. Secondly, based on these two tables, our modified Viterbi algorithm [6] outputs the optimal threshold set that minimizes the overall FI loss under a specific FLOPs constraint.

B. Construction of Candidate threshold table and FLOPs-Fisher table

As shown in Fig. 2(a), the first step consists of three stages: (1) collection of class attention scores and FI of tokens; (2) construction of threshold table; and (3) construction of FLOPs-Fisher table.

In the first stage, given a sample dataset D, we collect the class attention scores and the FI of $|D| \times (N-1)$ non-class tokens for each transformer block, where $|D|$ is the number of samples in dataset D and N is the number of tokens. Since we apply our threshold pruning modules between MSA and MLP modules, we obtain the FI from the output of MSA modules. We empirically find that the FI of tokens descends exponentially from shallow layers to deep layers, which leads to an ineffective assessment of token importance between blocks. Therefore, we normalize the FI block-wisely.

In the second stage, we select M candidate thresholds within each transformer block to construct a candidate threshold table with a size of $L \times M$. For simplicity, we expect these M candidate thresholds to be evenly distributed across the domain of FLOPs reduction ratio. Ideally, the m-th candidate threshold in l-th block, denoted as $\theta_{l,m}$, can approximately reduce FLOPs count by a fraction of m/M in the block. $\theta_{l,m}$ is obtained by a two-step process. First, based on the number of tokens n and the embedding dimension d, the FLOPs count of a transformer block ϕ_{BLK} can be calculated as (5):

$$\phi_{BLK}(n,d) = 12nd^2 + 2n^2d. \tag{5}$$

We obtain the initial threshold $\tilde{\theta}_{l,m}$ by calulating the average of k-th and (k+1)-th largest class attention scores in l-th block, which is previously collected in the first stage, where k satisfies:

$$\phi_{BLK}(\frac{k}{|D| \times N}, d) = \frac{m}{M} \times \phi_{BLK}(N,d). \tag{6}$$

Secondly, since the actual FLOPs reduction ratio of threshold $\tilde{\theta}_{l,m}$, denoted as $r(\tilde{\theta}_{l,m})$, will be slightly smaller than the desired FLOPs reduction ratio, $\frac{m}{M}$, we adjust the FLOPs reduction ratio in equation (6) from $\frac{m}{M}$ to $\frac{m}{M} + (\frac{m}{M} - r(\tilde{\theta}_{l,m}))$ and repeat the first step to obtain the candidate threshold $\theta_{l,m}$.

In the third stage, we construct a FLOPs-Fisher table f with a size of $L \times M$, where $f_{l,m}$ represents the FI loss associated with candidate threshold $\theta_{l,m}$. To obtain $f_{l,m}$, we collect tokens with class attention scores below $\theta_{l,m}$ in l-th transformer block and compute the sum of their normalized FI.

C. Searching for optimal threshold set under a specific FLOPs constraint

The optimal threshold set should minimize the overall FI loss under a specific FLOPs constraint. Since all candidate thresholds are distributed evenly across the domain of FLOPs reduction ratio, the index of a candidate threshold is proportional to its FLOPs reduction ratio. Therefore, given a FLOPs reduction ratio constraint C, the optimization problem can be formulated as finding the optimal index set, denoted as $\{\hat{m}_1, \hat{m}_2, ..., \hat{m}_L\}$, that satisfies (7):

$$\hat{m}_1, ..., \hat{m}_L = \arg\min_{m_1,...,m_L} \sum_{l=1}^{L} f_{l,m_l}, where \frac{1}{L}\sum_{l=1}^{L}\frac{m_l}{M} \geq C. \tag{7}$$

Since all indices m_l are non-negative integers and all FI loss $f_{l,m}$ are non-negative real values, the above optimization problem can be solved by Viterbi Algorithm (VA) [6].

However, the FLOPs reduction ratio in deeper layers should not be less than that in shallower ones, thus we add this constraint on the indices: $m_1 \leq m_2 \leq ... \leq m_L$. Since original VA can only obtain a suboptimal solution under this additional constraint, we modify the forwarding stage of VA, as illustrated in Fig. 2(b). With dynamic programming, we solve (7) with additional constraint by iteratively construct cumulative FI-loss table of l-th block, denoted as F_l, as shown in (8):

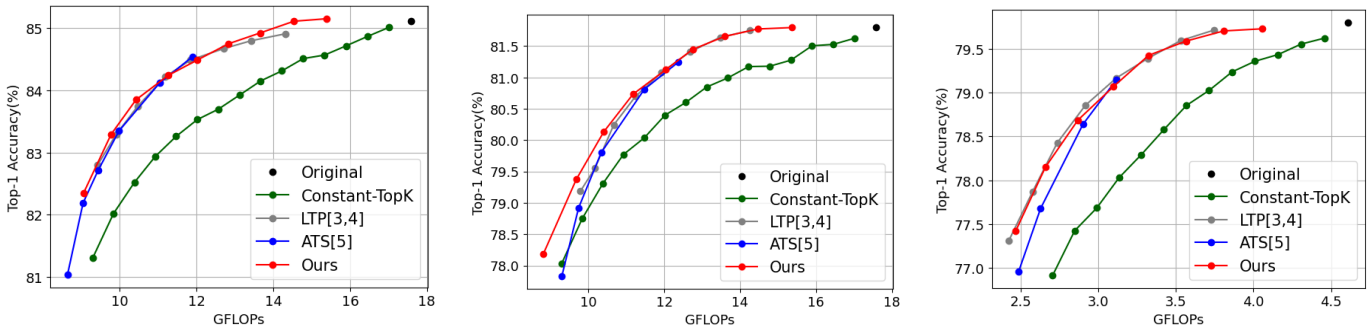
$$\begin{aligned} F_l[m_l, p_l] &:= \min_{m_1 \leq ... \leq m_l} \sum_{i=1}^{l} f_{i,m_i}, where \sum_{i=1}^{l-1} m_i = p_l - m_l \\ &= f_{l,m_l} + \min_{0 \leq m_{l-1} \leq m_l} F_{l-1}[m_{l-1}, p_l - m_l]. \end{aligned} \tag{8}$$

The modified VA would construct all cumulative FI-loss tables F_l by solving (8) from $l = 1$ to L. In the backtracking stage, for a FLOPs reduction constraint C, the minimum sum of indices, denoted as $\hat{p}_L$, should be $\lceil L \times M \times C \rceil$, and thus the optimal index in L-th block, denoted as $\hat{m}_L$, should be $\arg\min_m F_L[m, \hat{p}_L]$. Subsequently, we sequentially obtain the optimal threshold set through backtracking as the original VA process. For different FLOPs constraints, the corresponding optimal threshold sets can be obtained by backtracking on the same cumulative FI-loss table with different starting points, making our framework capable of making optimal threshold set determinations in one shot.

IV. Experiment Result

A. Experiment Setup

We evaluate our method on image classification task ImageNet-1k [9] dataset at 224×224 pixel resolution, with off-the-shelf ViT models including ViT-B, DeiT-B, and DeiT-S from Timm [10] library. We collect Fisher information (FI) from 16k training set images, and we generate the results by applying the resulting threshold set on testing sets with M = 201 and batch_size = 1.



(a) ViT-B (b) DeiT-B (c) DeiT-S

Fig. 3. Pareto fronts of accuracy-FLOPs trade-off for different token pruning methods on (a) ViT-B, (b) DeiT-B and (c) DeiT-S.

B. Experiment Result

To make a fair comparison, we reproduce **LTP [3] [4], ATS [5]**, and **constant Top-K** to be our baseline. All results except for LTP are generated without fine-tuning. The results on ViT-B, DeiT-B, and DeiT-S model are shown in Fig. 3. Across all three models, RCTP consistently outperforms constant Top-K across all FLOP constraints.

Comparing to LTP, Fig. 3 shows that RCTP achieves competitive and even better performance. On top of this, LTP needs fine-tuning from scratch to meet varying FLOPs constraints. In contrast, RCTP can generate resulting threshold sets for all FLOPs constraints with the modified Viterbi algorithm in "one-shot" by backtracking from different starting points on the same cumulative FI-loss table.

For ATS, RCTP has better performance while using less complex operations during inference. Besides, ATS lacks the capability to take FLOPs constraints as input, a feature that RCTP provides. It requires a trial-and-error process of hyperparameter tuning in order to meet a specific computation constraint.

C. Ablation study

In this section, we discuss the effectiveness of normalized FI by conducting experiments on DeiT-B model with different token importance criteria. Firstly, we compare the experiment results of FI and normalized FI to investigate the efficacy of the normalization. Secondly, we replace normalized FI with class attention scores to validate the capability of normalized FI in estimating the importance of tokens in different blocks.

a) Normalization on Fisher information: Fig. 4 demonstrates a significant accuracy drop of FI under restrictive FLOPs constraints. This decline is attributed to the substantial disparity in FI between transformer blocks, which makes our modified Viterbi algorithm over-prune tokens in deep layers.

b) Token importance assessment: In Fig. 4, it is evident that normalized FI greatly outperforms class attention scores. This observation suggests that normalized FI guides our modified Viterbi algorithm toward better threshold set selection.

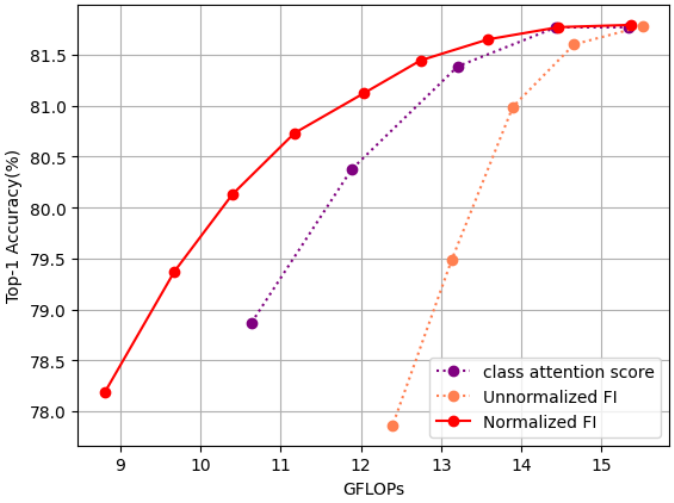


Fig. 4. Pareto fronts of accuracy-FLOPs trade-off for different token importance criteria on DeiT-B.

V. Conclusion

In this paper, we propose RCTP for edge-friendly deployment of ViT. RCTP generates an optimal threshold set to make token pruning decisions in each transformer block, which is edge-friendly due to the minimal additional hardware requirement. Meanwhile, the framework exploits Fisher information of tokens and our modified Viteri algorithm to achieve a constraint-aware trade-off between accuracy and FLOPs without retraining. Experiment results on the ImageNet benchmark show that compared with prior pruning-based works, RCTP achieves the best Pareto fronts of accuracy-FLOPs trade-off. Moreover, the ablation study proves that normalized Fisher information of tokens effectively estimates the information loss of thresholds across different blocks.

REFERENCES

[1] 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 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations*, 2021.

[2] D. Bolya, C.-Y. Fu, X. Dai, P. Zhang, C. Feichtenhofer, and J. Hoffman, "Token merging: Your ViT but faster," in *International Conference on Learning Representations*, 2023.

[3] S. Kim, S. Shen, D. Thorsley, A. Gholami, W. Kwon, J. Hassoun, and K. Keutzer, "Learned token pruning for transformers," in *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2022, pp. 784–794.

[4] M. Bonnaerens and J. Dambre, "Learned thresholds token merging and pruning for vision transformers," *Transactions on Machine Learning Research*, 2023.

[5] M. Fayyaz, S. A. Koohpayegani, F. R. Jafari, S. Sengupta, H. R. V. Joze, E. Sommerlade, H. Pirsiavash, and J. Gall, "Adaptive token sampling for efficient vision transformers," in *European Conference on Computer Vision (ECCV)*, 2022.

[6] A. Viterbi, "Error bounds for convolutional codes and an asymptotically optimum decoding algorithm," *IEEE Transactions on Information Theory*, vol. 13, no. 2, pp. 260–269, 1967.

[7] L. Liu, S. Zhang, Z. Kuang, A. Zhou, J.-H. Xue, X. Wang, Y. Chen, W. Yang, Q. Liao, and W. Zhang, "Group fisher pruning for practical network compression," in *International Conference on Machine Learning*. PMLR, 2021, pp. 7021–7032.

[8] W. Kwon, S. Kim, M. W. Mahoney, J. Hassoun, K. Keutzer, and A. Gholami, "A fast post-training pruning framework for transformers," *Advances in Neural Information Processing Systems*, vol. 35, pp. 24 101–24 116, 2022.

[9] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248–255.

[10] R. Wightman, "Pytorch image models," https://github.com/rwightman/pytorch-image-models, 2019.