Beyond Attentive Tokens: Incorporating Token Importance and Diversity for Efficient Vision Transformers

Sifan Long[1,2*] Zhen Zhao[3,2*] Jimin Pi[2] Shengsheng Wang[1†] Jingdong Wang[2†]
[1]Jilin University [2]Baidu VIS [3]University of Sydney
longsf22@mails.jlu.edu.cn zhen.zhao@sydney.edu.au
wss@jlu.edu.cn {pijimin01, wangjingdong}@baidu.com

Abstract

Vision transformers have achieved significant improvements on various vision tasks but their quadratic interactions between tokens significantly reduce computational efficiency. Many pruning methods have been proposed to remove redundant tokens for efficient vision transformers recently. However, existing studies mainly focus on the token importance to preserve local attentive tokens but completely ignore the global token diversity. In this paper, we emphasize the cruciality of diverse global semantics and propose an efficient token decoupling and merging method that can jointly consider the token importance and diversity for token pruning. According to the class token attention, we decouple the attentive and inattentive tokens. In addition to preserve the most discriminative local tokens, we merge similar inattentive tokens and match homogeneous attentive tokens to maximize the token diversity. Despite its simplicity, our method obtains a promising trade-off between model complexity and classification accuracy. On DeiT-S, our method reduces the FLOPs by 35% with only a 0.2% accuracy drop. Notably, benefiting from maintaining the token diversity, our method can even improve the accuracy of DeiT-T by 0.1% after reducing its FLOPs by 40%.

1. Introduction

Transformer [29] has become the most popular architecture in both natural language processing and computer vision communities. Vision transformers (ViTs) [8] have achieved superior performance and outperformed standard CNNs in different vision tasks such as image classification [10, 28, 31, 38], semantic segmentation [17, 19, 30, 33], and object detection [1, 5]. The most remarkable advantage of transformer is its ability to effectively capture long-range dependencies between patches in the input image through the self-attention mechanism [23]. However, quadratic interactions between tokens significantly degrade the computational efficiency [36], which motivates many researches on exploring efficient transformers.

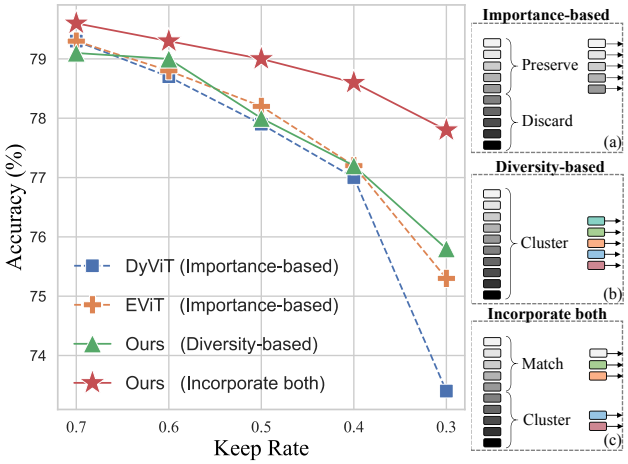


Figure 1. The ImageNet accuracy and keep rate of the pruned DeiT-S. (a) Importance-based method preserves attentive tokens based on the class token attention and masks all inattentive tokens; (b) Diversity-based method clusters similar tokens into a group and then combines tokens from the same group into a new token. (c) Incorporate method decouples and merges tokens to consider token importance and diversity simultaneously.

As one of the most direct and effective ways to reduce computational complexity, token pruning has been widely studied recently. Existing studies mainly focus on designing different importance-evaluating strategies to retain attentive tokens and prune inattentive tokens [18, 21, 23, 35, 37]. In these importance-based works, DyViT [23] introduces an extra module to estimate the importance of each token while EViT [18] reorganizes image tokens based on the class attention importance score. However, inspired by recent diversity-preserving studies in ViT variants [9, 11–13, 25, 27], we argue that promoting token diversity is also cru-

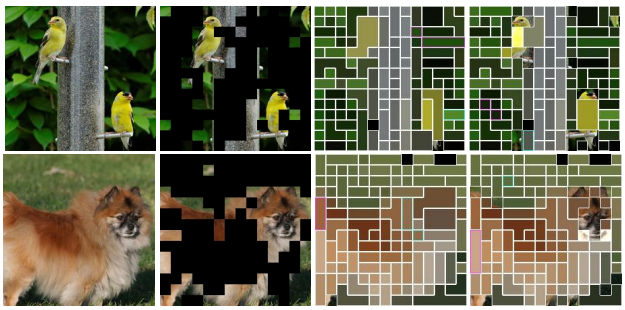
*Equal contribution.
†Corresponding authors.

cial for token pruning. Though inattentive tokens like the image background and low-level textures are not directly related to the classification objects, they can increase the token diversity and improve the expressivity of the model. As discussed in [32], image backgrounds (*e.g.*, the grass and leaves in Fig. 2) can improve the classification accuracy due to their potential relations to foreground objects. To this end, we first investigate a diversity-based pruning strategy on DeiT-S [28] with different keep rates. Specifically, instead of highlighting the token importance, it directly clusters and combines similar tokens into a single one, hereby maximizing the token diversity. Surprisingly, as shown in Fig. 1, such an intuitive strategy can achieve comparable and even better performance than SOTA importance-based pruning methods, especially at the low keep rate.

Despite its promising performance, the diversity-based strategy cannot retain original attentive tokens and may consequently weaken the discriminative ability of the model. As shown in Fig. 2 (c), the most representative tokens, *e.g.*, eyes and ears of the dog or beaks of two birds, contain critical semantic information for classification tasks but cannot be preserved by the diversity-based strategy. To address this issue, we naturally tend to keep all these dominant tokens while maintaining the token diversity, as shown in Fig. 2 (d). In short, a satisfied pruning method should jointly take the token importance and diversity into account, such that the most important local information and the diverse global information can be preserved simultaneously.

Motivated by these above observations, in this paper, we propose a novel pruning method that incorporates the token importance and diversity through efficient token decoupling and merging. As shown in Figure 1 (c), we first decouple the origin token sequence into attentive and inattentive portions based on class token attention. Instead of discarding inattentive tokens completely, we apply a simplified density peak clustering algorithm [24] to efficiently cluster similar inattentive tokens and combine these tokens from the same group into a new one. In addition, unlike existing methods that preserve all attentive tokens, we design a straightforward matching algorithm to fuse homogeneous attentive tokens and improve the calculation efficiency further. In this way, we can effectively prune tokens while maximizing the preservation of token diversity. We conduct extensive token pruning experiments to validate the effectiveness of our method. Despite its simplicity, our method achieves superior pruning performance on ImageNet [6] for two different vision transformers, DeiT [28] and LV-ViT [15]. Our main contributions are summarized as follows:

- To the best of our knowledge, we are the first to emphasize the token diversity for pruning ViT. We also demonstrate its cruciality through numerical and empirical analysis.



(a) (b) (c) (d)

Figure 2. Visualizations of pruning results of different methods on ImageNet with DeiT-S. (a) Original image. (b) Importance-based method masks inattentive tokens. (c) Diversity-based method clusters similar tokens and visualizes the same group of tokens as one colour. (d) Our method preserves the most discriminative tokens, *e.g.*, the heads of birds and dogs. In addition, we merge similar inattentive tokens and match homogeneous attentive tokens, *e.g.*, the grass and leaves.

- We propose a simple yet effective decoupling and merging method that can simultaneously preserve the most attentive local tokens and diverse global semantics without imposing extra parameters.

- Benefiting from incorporating token importance and diversity, our method achieves new SOTA performance on the trade-off between accuracy and FLOPs. It can also be deployed to other token pruning methods, achieving excellent performance improvement.

2. Related work

Vision Transformers. Different from convolution networks, the transformer has a significant ability to model long-range dependencies and minimal inductive bias [35]. Recent advances suggest that the variants of transformers could be a competitive alternative to CNNs. Visual transformer (ViT) [8] is the first work to apply transformer architecture to achieve STOA performance, but it only replaces the standard convolution in the deep neural network on large-scale image datasets. To free ViT from dependence on large datasets, DeiT [28] incorporates an additional token for knowledge distillation to improve the training efficiency of vision transformers. LV-ViT [15] further improves the performance by utilizing all the image patch tokens to calculate the training loss intensively. It is equivalent to converting the image classification problem into each token recognition problem. While ViT and its follow-ups achieve excellent performance, the complexity quadratic with the number of tokens incurs high computational costs. Token pruning aims to reduce redundant tokens and improve the inference efficiency of various ViT backbones.

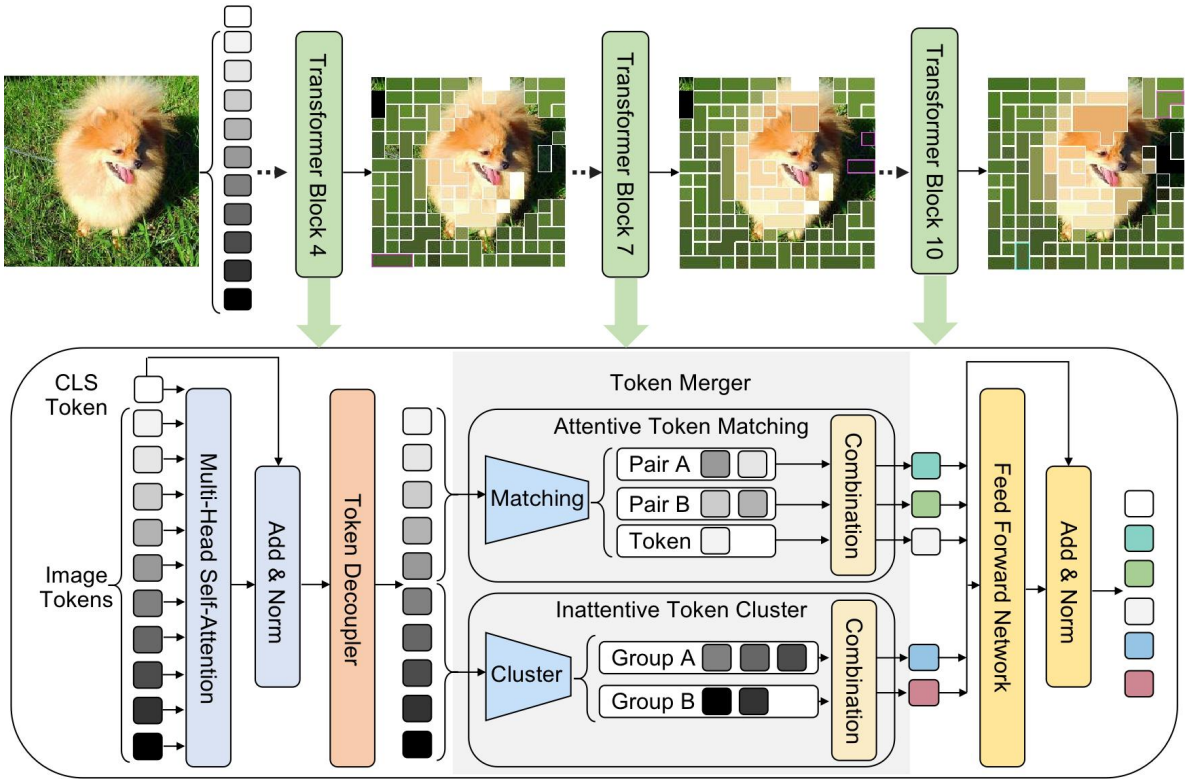


Figure 3. Illustration of our approach. (top) Employ our method at the 4th, 7th, and 10th layers of the DeiT-S model. (bottom) Model structure within a single transformer block. We decouple the attentive and inattentive tokens according to class token attention. Then, we cluster inattentive tokens and combine the tokens from the same group into a new token. Meanwhile, we match homogeneous attentive tokens and combine the same pair of tokens.

ViT Token pruning. Though ViT has achieved competitive accuracy in vision tasks [1, 5, 10, 28, 31, 38], it needs huge memory and computational resources. Therefore, how to build a more efficient transformer draws researchers' interest. Compared with CNN, the higher computing cost of the transformers is mainly due to the quadratic time complexity of multi-head self-attention (MHSA). Accordingly, some work [18, 21, 23, 35] attempts to prune tokens based on importance score in transformer. Based on whether extra parameters need to be introduced to the model, we divide the existing token pruning methods into the following two groups. One group performs token pruning by inserting prediction modules. DyViT [23] designs a lightweight prediction module inserted into different layers to estimate the importance score of each token to prune redundant tokens given the current features. IA-RED2 [21] introduces interpretable modules to dynamically delete redundant patches that are not related to the input. AdaViT [20] connects a lightweight decision network to the backbone to dynamically generate decisions. The other group leverages the class token attention to keep attentive tokens. EViT [18] divides image tokens into attentive and inattentive tokens according to class token attention, retains attentive tokens and discards inattentive image tokens to reorganize image tokens. Evo-ViT [35] distinguishes informative and uninformative tokens through global class attention for slow and fast updates, respectively. A-ViT [37] designs an adaptive token pruning mechanism based on class token attention, which dynamically adjusts the calculation cost of images with different complexity. Unlike these token pruning methods, which only focus on the importance of tokens, our method also considers the diversity of token semantic information. Therefore our method achieve incredible performance.

3. Preliminaries

In standard vision transformers [8], each input image $\mathrm{I} \in \mathbb{R}^{H\times W\times C}$ is first converted into a single-dimensional patch sequence $\mathrm{X} \in \mathbb{R}^{N\times P^2\times C}$. Then all patches are mapped into D-dimensional token embeddings via a trainable linear layer. Additionally, a learnable position embedding $\mathrm{E}_{\text{pos}} \in \mathbb{R}^{(N+1)\times D}$ is added to token embedding to retain position information. Formally, the input patch sequence can be represented as:

$$\mathrm{X} = [\mathrm{x}_{\mathrm{cls}}; \mathrm{x}_1; \ldots; \mathrm{x}_{\mathrm{N}}] + \mathrm{E}_{\mathrm{pos}}, \quad (1)$$

where $\mathrm{x}_{\mathrm{cls}}$ denotes the learnable class token that serves as the image representation, and x_i denotes the token of the i-th patch with $i \geq 0$. Afterwards, such token sequence is fed into a ViT model with L stacked transformer blocks, each of which consists of a multi-head self-attention (MHSA) module and a feed forward network (FFN).

3.1. MHSA & FFN

In MHSA, the input token sequence is linearly mapped into three different matrices of query Q, key K, and value V, respectively. MHSA can be formulated as:

$$\mathrm{MHSA}\left(\mathrm{Z}\right) = \mathrm{Concat}\left[\mathrm{softmax}\left(\frac{\mathrm{Q}^{\mathrm{h}}\left(\mathrm{K}^{\mathrm{h}}\right)^{\top}}{\sqrt{\mathrm{d}}}\right)\mathrm{V}^{\mathrm{h}}\right]_{\mathrm{h}=1}^{\mathrm{H}}, \quad (2)$$

where Z is the token sequence of $\mathrm{N}+1$ tokens. Concat$[\cdot]$ outputs the feature concatenation of H heads. Q^{h}, K^{h} and V^{h} are projection matrices of Q, K, and V in the h-th head, respectively. d is the feature dimension of the single head.

FFN typically consists of two fully-connected layers and a nonlinear mapping layer, which can expressed as:

$$\mathrm{FFN}\left(\mathrm{Z}\right) = \mathrm{Sigmoid}(\mathrm{Linear}(\mathrm{GeLU}(\mathrm{Linear}(\mathrm{Z})))), \quad (3)$$

where Linear denotes the fully-connected layers and GeLU is an non-linear activation function.

3.2. Computation Complexity

The dimension of the input token sequence is $N \times D$, where N is the number of tokens and D is the embedding dimension of each token. Thus the calculational costs of MSHA and FFN modules are $\mathcal{O}\left(4ND^2 + 2N^2D\right)$ and $\mathcal{O}\left(8ND^2\right)$, respectively. Obviously, vision transformers require very intensive computational costs, with the total computational complexity of $\mathcal{O}\left(12ND^2 + 2N^2D\right)$. Since reducing the channel dimension D only affects the calculation of current matrix multiplication, most related works tend to prune tokens, *e.g.* reducing the number of N, to reduce all operations linearly or even quadratically.

4. Methodology

4.1. Overview

Different from existing works only focus on attentive tokens, our method incorporating token importance and diversity to obtain efficient and accurate vision transformers. To this end, we propose the token decoupling and merging method, achieving promising trade-offs between the FLOPs and accuracy. As shown in Figure 3, we insert our approach at the 4th, 7th, and 10th layers of the DeiT-S model. The approach has two main components: **the token decoupler** and **the token merger**. The decoupler divides the origin token sequence into attentive and inattentive sections based on class token attention. Then the merger clusters similar inattentive tokens and matches homogeneous attentive tokens. In this section, we first demonstrate how preserving token diversity benefits token pruning and then present the two main components in detail.

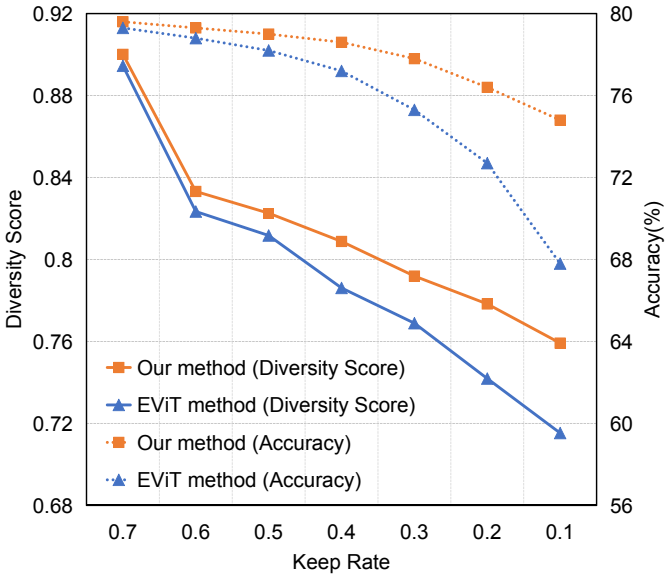


Figure 4. Comparing the pruning results of our method and the EViT method with different keep rates on DEiT-S in terms of the diversity score and classification accuracy on ImageNet. The diversity score is obtained at the final pruning layer.

4.2. Token diversity matters

In the literature, most work only emphasize retaining important tokens but directly discard all the remaining ones to achieve satisfactory token keep rates. However, inspired by the observations in [32], that even the image background can help improve foreground-instance classification, we argue that the less important tokens could also contain useful semantic information and be an effective complementary to the information diversity. Also, as discussed in [9, 11–13, 25, 27], the token diversity is very critical to optimize transformer structures. Therefore, appropriately preserving these inattentive tokens augments the diversity of semantic information, which can be beneficial to token pruning. On the contrary, blindly discarding tokens will cause irreversible loss of semantic information, especially at the low keep rates. Referring to [7, 11, 26, 27], we leverage the difference between the token and a rank-1 matrix to measure the diversity of token sequence Z. The diversity

scores $r(\mathrm{Z})$ can be calculated as:

$$r(\mathrm{Z}) = \left\|\mathrm{Z} - 1z^\top\right\|, \text{ where } \boldsymbol{z} = \operatorname{argmin}_{\boldsymbol{z}'} \left\|\mathrm{Z} - 1z'^\top\right\|, \tag{4}$$

where $\|\cdot\|$ represents l_1 norm. $\mathrm{Z} \in \mathbb{R}^{N\times C}$ is the token sequence of N tokens and $\boldsymbol{z}, \boldsymbol{z}' \in \mathbb{R}^C$ is one of the tokens. $z^\top$ is the matrix transpose of $\boldsymbol{z}$ and 1 is an all-ones vector. The rank of matrix $1z^\top$ is 1. A larger diversity score indicates a more diverse token sequence.

We investigate how the diversity scores affect the token pruning performance. In Figure 4, we examine the classification accuracy and the diversity score at different keep rates. Obviously, we can see that the token sequence diversity score is positively correlated with classification accuracy. In either the EViT method or our proposed method, higher diversity score consistently result in higher accuracy. In addition, it can also be observed that, as for the EViT method, the diversity score and classification accuracy drop rapidly as the keep rate decreases. Differently, benefiting from our diversity-preserving token merging strategy, our method can maintain relatively higher diversity scores at different keep-rates and thus consistently outperform the EViT method, especially at the low keep-rate. Therefore, maintaining higher token diversity is crucial to improve classification accuracy.

4.3. Token Decoupler

In order to fully consider token importance while maintaining diversity, we prioritize the attentive tokens to preserve the most important semantic information. Therefore, we decouple original token sequence into attentive and inattentive tokens so that we maintain token diversity and importance simultaneously. Same as [29], we split the tokens into two groups by comparing their similarities with the class tokens. Mathematically, the similarity scores $\mathrm{Attn}_{\mathrm{cls}}$ between the class token and other tokens as class token attention can be calculated by

$$\mathrm{Attn}_{\mathrm{cls}} = \mathrm{Softmax}\left(\frac{\mathrm{q}_{\mathrm{cls}} \cdot \mathrm{K}^\top}{\sqrt{\mathrm{d}}}\right), \tag{5}$$

where $\mathrm{q}_{\mathrm{cls}}$ denotes the class token of query vector. With N tokens in total and the keep rate of η, we choose the top-K tokens as attentive tokens according to attention scores. The remaining N-K tokens are identified as inattentive tokens that contain less information. Moreover, in the multi-head self-attention layer, we calculate the average of the attention scores of all heads.

4.4. Token Merger

We apply different strategies for attentive and inattentive tokens when merging tokens. For inattentive tokens, we first apply density peak clustering algorithm to cluster inattentive tokens and then combine the tokens from the same group into new token by weighted sum. In this way, we can integrate a new inattentive token sequence $\mathrm{T} = [\mathrm{t}_1; \ldots; \mathrm{t}_\mathrm{n}]$. For attentive tokens, we adapt a straightforward matching algorithm to fuse homogeneous attentive tokens. The fused token sequence is $\mathrm{P} = [\mathrm{p}_1; \ldots; \mathrm{p}_\mathrm{m}]$. We concat T and P to obtain the pruned token sequence $\mathrm{Z} = [\mathrm{z}_{\mathrm{cls}}; \mathrm{p}_1; \ldots; \mathrm{p}_\mathrm{m}; \mathrm{t}_1; \ldots; \mathrm{t}_\mathrm{n}]$.

Inattentive Token Clustering. Common K-means clustering algorithm requires multiple iterations to obtain satisfactory results, reducing throughput in practice and defeating the intent of speeding up the model. After research, we found that the density clustering algorithm can quickly find classes of arbitrary shape by exploiting the density connectivity of classes. Therefore, we simplify an efficient density peak clustering algorithm (DPC) with neither an iterative process nor more parameters. We follow the DPC algorithm proposed in [24]. It assumes that the cluster center is surrounded by low-density neighbours, and that the distance between the cluster center and any high-density points is relatively large. We calculate two variables for each token i: the density ρ and the minimum distance from the higher density token δ. Given a set of tokens x, we calculate the density of each token ρ by

$$\rho_i = \exp\left(\sum_{z_j \in \mathrm{Z}} \|z_i - z_j\|_2^2\right). \tag{6}$$

where Z denotes the set of tokens. z_i and z_j are corresponding token features.

For the token with the highest density, its minimum distance is set to the maximum distance between it and any other tokens. We define δ_i as the minimum distance between the token i and any other token with higher density. The minimum distance of each token is:

$$\delta_i = \begin{cases} \min_{j:\rho_j > \rho_i} \|z_i - z_j\|_2, \text{ if } \exists j \text{ s.t. } \rho_j > \rho_i \\ \max_j \|z_i - z_j\|_2, \text{ otherwise} \end{cases}. \tag{7}$$

We denote the clustering center score of the i-th token as $\rho_i \times \delta_i$. Higher scores mean higher potential to be cluster centers. We select top-K-scored tokens as cluster centers. The DPC algorithm assigns each remaining token to the cluster whose cluster center is closest to the token and has a higher density.

Attentive Token Matching. See example images in Figure 5. Homogeneous tokens are also present in foreground objects (attentive tokens), such as the cheeks of animals. This redundancy makes it possible to fuse homogeneous attentive tokens to reduce the number of tokens while maintaining accuracy. We could apply the same token clustering strategy as did for inattentive tokens. However, since

Model	Method	Top-1 Acc. (%)	Params (M)	FLOPs (G)	FLOPs ↓(%)	Throughput (img/s)
DeiT-T	DeiT-T [28]	72.2	5.6	1.3	0.0	2536
	DyViT [23]	71.2	5.9	0.9	30.8	3542
	PS-ViT [26]	72.0	5.6	0.9	30.8	3563
	SViTE [3]	70.1	4.2	0.9	30.8	2836
	SPViT [16]	72.2	5.6	1.0	23.1	-
	Evo-ViT [35]	72.0	5.6	0.8	38.5	3627
	Ours-DeiT-T	72.3	5.6	0.8	38.5	3641
DeiT-S	DeiT-S [28]	79.8	22.1	4.6	0.0	943
	DyViT [23]	79.3	22.8	2.9	37.0	1420
	PS-ViT [26]	79.4	22.1	2.6	43.5	1392
	IA-RED2 [21]	79.1	22.1	3.2	30.4	1362
	Evo-ViT [35]	79.4	22.1	3.0	34.8	1449
	EViT [18]	79.5	22.1	3.0	34.8	1455
	A-ViT [37]	78.6	22.1	3.6	21.7	-
	Ours-DeiT-S	79.6	22.1	3.0	34.8	1468
	EViT+Ours	79.6	22.1	3.0	34.8	1459
DeiT-B	DeiT-B [28]	81.8	86.6	17.6	0.0	302
	DyViT [23]	81.3	-	11.6	34.1	454
	PS-ViT [26]	81.5	86.6	11.6	34.1	445
	IA-RED2 [21]	80.9	86.6	11.6	34.1	453
	Evo-ViT [35]	81.3	86.6	11.6	34.1	448
	EViT [18]	81.3	86.6	11.6	34.1	450
	Ours-DeiT-B	82.0	86.6	11.6	34.1	462

Table 1. Comparisons with existing token pruning methods on DeiT. We report the top-1 classification accuracy, FLOPs, and throughput on ImageNet. 'FLOPs ↓' denotes the reduction ratio of FLOPs.

the attentive tokens contain critical semantic information for the final classification task, it would be best if we can preserve the original tokens. To address this problem, we customize a straightforward matching algorithm that keeps the most important tokens while fusing homogeneous tokens. Specifically, we define the cosine similarity metric to determine the similarity between different tokens and calculate cosine similarity scores between attentive tokens. Then we select top-K most similar token pairs as homogeneous tokens. Finally, we combine each pair of tokens into a new token and contact the remaining attentive tokens.

Although tokens in the same set have similar semantic information, the semantic importance of each token is not necessarily the same. Instead of blindly averaging the tokens in the same set, we combine these tokens by a weighted sum. By introducing a class token attention to represent the importance, we combine the same set of tokens into a new token by

$$p_i = \sum_{j \in C_i} s_j z_j, \tag{8}$$

where s_j denotes the importance score of token z_j, and C_i denotes the i-th set.

5. Experiments

5.1. Setup

Dataset and evaluation metrics. Our experiments are conducted on ImageNet-1K [6] with 1.28 million training images and 50000 validation images. We report the top-1 classification accuracy and the floating-point operations (FLOPs) to evaluate model efficiency. In addition, we measure the throughput of the models on a single NVIDIA V100 GPU with batch size fixed to 256.

Implementation details. To demonstrate the generalization of our method, we conduct token pruning on different ViT models including DeiT-T, DeiT-S, DeiT-B [28], and LV-ViT-S [15]. Following [18], we employ our method at the 4^{th}, 7^{th}, and 10^{th} layers of the DeiT-T/S/B model and at the 4^{th}, 8^{th}, and 12^{th} layers for LV-ViT-S [15]. We utilize the same training settings as the original papers of DeiT [28] and LV-ViT [15]. Following [38], we incorporate a cosine scheduler into our learning strategy where the keep-rate gradually decreases from 1 to the target value for 100 epochs. For fair comparisons, all of our models are trained from scratch for 300 epochs on 8 NVIDIA V100.

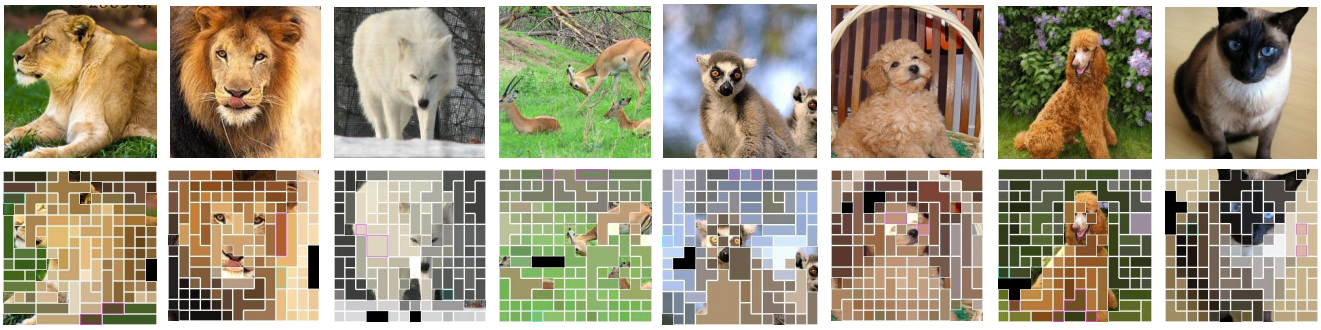


Figure 5. Visualization of token merger results on DeiT-S. The masked areas of different colours represent the inattentive tokens are divided into dissimilar token groups. Our method clusters similar inattentive tokens into a group and matches homogeneous attentive tokens. We visualize the same groups/pairs of tokens as the same colour.

Method	Top-1 Acc (%)	FLOPs (G)	Params (M)
ViT-Base/16 [8]	77.9	17.6	86.6
DeiT-S [28]	79.8	4.6	22.1
DeiT-Base/16 [28]	81.8	17.6	86.6
PVT-Small [30]	79.8	3.8	24.5
PVT-Medium [30]	81.2	6.7	44.2
CPVT-Small-GAP [4]	81.5	4.6	23.0
CoaT Mini [34]	80.8	6.8	10.0
CoaT-Lite Small [34]	81.9	4.0	20.0
CrossViT-S [2]	81.0	5.6	26.7
CrossViT-B [2]	82.3	14.1	64.1
Swin-T [19]	81.3	4.5	29.0
Swin-S [19]	83.0	8.7	50.0
Swin-B [19]	83.3	15.4	88.0
T2T-ViT-14 [38]	81.5	5.2	22.0
T2T-ViT-19 [38]	81.9	8.9	39.2
T2T-ViT-24 [4]	82.2	21.2	104.7
RegNetY-8G [22]	81.7	8.0	39.0
RegNetY-16G [22]	82.9	16.0	84.0
LV-ViT-S [14]	83.3	6.6	26.2
DyViT-LV-S	83.0	4.6	26.2
EViT-LV-S	83.0	4.7	26.2
Ours-LV-S	**83.1**	**4.7**	26.2

Table 2. Comparisons with state-of-the-art vision transformers on ImageNet. We prune the LV-ViT-S as the base model.

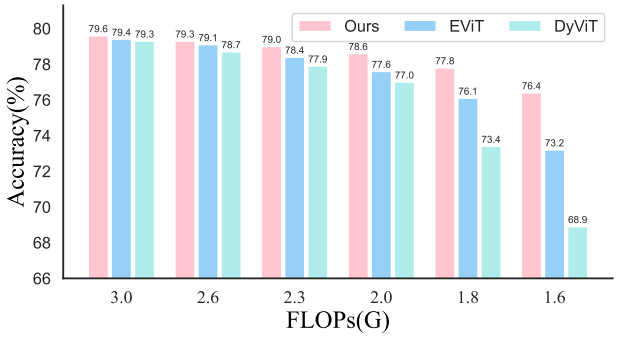


Figure 6. Performance comparisons of DyViT, EViT, and our method with different FLOPs.

5.2. Main Results

Comparisons with the-state-of-the-arts. We compare our method with SOTA token pruning methods, results are shown in Table 1. We leveraged the η indicates the keep rate. We report the top-1 accuracy, FLOPs, and throughput for each model. Compared to previous work, our method achieves new SOTA performance with similar computation costs. Specifically, on the classic model DeiT-S [28], the top-1 accuracy degradation of our pruned models is controlled within 0.2% when the computation costs decreases by 35%. In addition, the superiority of our method is more obvious at lower keep-rates. When the compression ratio of DeiT-S is increased to 50%, we improve 0.5% compared to the best counterpart. In particular, the compression ratio of our method is close to 40% on the DeiT-T [28] model, and the accuracy is even better than the original model. Owing to the token diversity and importance are orthogonal for token pruning, our method can be plugged into EViT and achieve a performance improvement of 0.1%. As shown in Table 2, we further conduct experiments on the deep-narrow transformer LV-ViT [15]. We observe that our method achieves better accuracy-complexity trade-offs on different keep rates compared to the current foremost CNN and ViT architectures.

Performance of existing methods on each keep rate. As shown in Figure 6, our method consistently achieves the best performance while the other two methods obtain close performance. In addition, with the decrease of keep rate,

Strategy	Acc (%)	FLOPs (G)
Deit-S	79.8	4.6
DeiT-S/η=0.7		
+ attentive token preservation	79.3	3.0
+ inattentive token pack	79.3	3.0
+ inattentive token clustering	79.5	3.0
+ attentive token matching	79.6	3.0
DeiT-S/η=0.5		
+ attentive token preservation	78.2	2.3
+ inattentive token pack	78.4	2.3
+ inattentive token clustering	78.8	2.3
+ attentive token matching	79.0	2.3

Table 3. Ablation study on our method with different keep-rate η.

Method	Acc (%)	FLOPs (G)	Throughput (img/s)
Pooling strategy			
Average pooling	78.1	2.3	1630
Max pooling	78.1	2.3	1623
Spatial pooling	78.2	2.3	1605
Sub-sampling strategy			
Convolution	78.2	2.3	1454
MLP	78.3	2.3	1447
Cluster strategy			
K-means(1 iter)	78.6	2.3	1386
K-means(3 iter)	78.8	2.3	1231
Ours	79.0	2.3	1670

Table 4. Different token merger strategies on DeiT-S.

the classification accuracy of existing token pruning methods drops sharply. Fortunately, our method alleviates the phenomenon by preserving the diversity of token semantic information. Especially when the FLOPs of DyViT is reduced to 1.6G, the classification accuracy drops by more than 10%. This is because completely discarding inattentive tokens greatly decreases token diversity, resulting in the reduction of the semantic information of the original token sequence. We apply the token decoupling and merging to simultaneously consider the token importance and diversity, achieving incredible accuracy at low keep rates. Intuitively, when we only keep a few tokens, merging tokens obviously makes more sense than keeping only top-K attentive tokens.

Visualization of the token merger results. To further show the interpretability of our method, we visualize the final token merger results back to its original input patches in Figure 5. We notice that our method pays attention to the regions that contribute more to image prediction instead of uninformative backgrounds. *e.g.*, the animal's five sense organs are preserved. It demonstrates that our method effectively decouple the attentive and inattentive tokens. Unlike other methods that mask all inattentive tokens, our method combines background patches with similar semantics. *e.g.*, the animal's fur is merged into a single token. It implies that our method not only focuses on attentive tokens but also maintains the diversity of token semantics. It is also worth pointing out that the paired eyes are matched and combined into a token in the fifth and sixth image. It indicates that our method effectively fuse homogeneous attentive tokens and reduces the potential redundancy.

5.3. Ablation Analysis

Effectiveness of each module. As shown in Table 3, we add the sub-modules one by one to evaluate the effectiveness of each module. i) Attentive token preservation. Discard inattentive tokens based on the class token attention in pruning layers; ii) Inattentive token pack. Pack all inattentive tokens into one token; iii) Inattentive token clustering. Cluster inattentive tokens and combine the tokens of the same group into a new token; iv) Attentive token matching. Match attentive tokens and combine the tokens of the same pair into a new token; It is evident that since the clustering module preserves token diversity, the accuracy is improved by 0.2% and 0.8% at keep rates of 0.7 and 0.5, respectively. Noteworthy, the lower keep rates, the better our method works. In addition, the attentive token matching module further reduces the FLOPs of the model while maintaining accuracy by fusing homogeneous attentive tokens.

Different Token Merger Strategy. As presented in Table 4, we compare several common token merging strategies to assess the effectiveness of our approach. i) Pooling strategy. Utilize the pooling operation to reduce the number of tokens. ii) Sub-sampling strategy. Adding a series of convolution layers between MHSA and FFN to decrease the token dimension. iii) Cluster strategy. Cluster the tokens and combine the tokens of the same group into a new token. We observe that the clustering strategy generally improves the accuracy by 0.4% compared to other token merging strategies. A possible reason is that the clustering strategy obtains more inductive bias at the same computational cost. However, we find that the throughput of K-means algorithm is lower, indicating that it does not perform well in terms of model acceleration in practice. Furthermore, we discover that the throughput of the K-means algorithm decreases with the number of iterations. Therefore we simplify an efficient DPC algorithm with neither an iterative process nor more parameters, which outperforms other strategies on both accuracy and efficiency.

6. Conclusion

In this paper, we propose a token decoupling and merging method to simultaneously consider the token importance and diversity. Since token importance and diversity are orthogonal for token pruning, our method can be employed into exisiting token pruning methods to further improve the performance. We demonstrate that our method achieved the SOTA performance trade-off between accuracy and FLOPs without imposing extra parameters. We hope that this paper, which incorporates token importance and diversity, will provide insights for the future work of pruning visual transformers.

References

[1] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020. 1, 3

[2] Chun-Fu Richard Chen, Quanfu Fan, and Rameswar Panda. Crossvit: Cross-attention multi-scale vision transformer for image classification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 357–366, 2021. 7

[3] Tianlong Chen, Yu Cheng, Zhe Gan, Lu Yuan, Lei Zhang, and Zhangyang Wang. Chasing sparsity in vision transformers: An end-to-end exploration. *Advances in Neural Information Processing Systems*, 34:19974–19988, 2021. 6

[4] Xiangxiang Chu, Zhi Tian, Bo Zhang, Xinlong Wang, Xiaolin Wei, Huaxia Xia, and Chunhua Shen. Conditional positional encodings for vision transformers. *arXiv preprint arXiv:2102.10882*, 2021. 7

[5] Zhigang Dai, Bolun Cai, Yugeng Lin, and Junying Chen. Up-detr: Unsupervised pre-training for object detection with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1601–1610, 2021. 1, 3

[6] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 2, 6

[7] Yihe Dong, Jean-Baptiste Cordonnier, and Andreas Loukas. Attention is not all you need: Pure attention loses rank doubly exponentially with depth. In *International Conference on Machine Learning*, pages 2793–2803. PMLR, 2021. 4

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 1, 2, 3, 7

[9] Chengyue Gong, Dilin Wang, Meng Li, Vikas Chandra, and Qiang Liu. Vision transformers with patch diversification. *arXiv preprint arXiv:2104.12753*, 2021. 1, 4

[10] Benjamin Graham, Alaaeldin El-Nouby, Hugo Touvron, Pierre Stock, Armand Joulin, Hervé Jégou, and Matthijs Douze. Levit: a vision transformer in convnet's clothing for faster inference. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 12259–12269, 2021. 1, 3

[11] Kai Han, Yunhe Wang, Jianyuan Guo, Yehui Tang, and Enhua Wu. Vision gnn: An image is worth graph of nodes. *arXiv preprint arXiv:2206.00272*, 2022. 1, 4

[12] Kai Han, Yunhe Wang, Qi Tian, Jianyuan Guo, Chunjing Xu, and Chang Xu. Ghostnet: More features from cheap operations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1580–1589, 2020. 1, 4

[13] Kai Han, An Xiao, Enhua Wu, Jianyuan Guo, Chunjing Xu, and Yunhe Wang. Transformer in transformer. *Advances in Neural Information Processing Systems*, 34:15908–15919, 2021. 1, 4

[14] Zihang Jiang, Qibin Hou, Li Yuan, Daquan Zhou, Xiaojie Jin, Anran Wang, and Jiashi Feng. Token labeling: Training a 85.5% top-1 accuracy vision transformer with 56m parameters on imagenet. *arXiv preprint arXiv:2104.10858*, 3(6):7, 2021. 7

[15] Zi-Hang Jiang, Qibin Hou, Li Yuan, Daquan Zhou, Yujun Shi, Xiaojie Jin, Anran Wang, and Jiashi Feng. All tokens matter: Token labeling for training better vision transformers. *Advances in Neural Information Processing Systems*, 34:18590–18602, 2021. 2, 6, 7

[16] Zhenglun Kong, Peiyan Dong, Xiaolong Ma, Xin Meng, Wei Niu, Mengshu Sun, Bin Ren, Minghai Qin, Hao Tang, and Yanzhi Wang. Spvit: Enabling faster vision transformers via soft token pruning. *arXiv preprint arXiv:2112.13890*, 2021. 6

[17] Zhiqi Li, Wenhai Wang, Enze Xie, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, Ping Luo, and Tong Lu. Panoptic segformer: Delving deeper into panoptic segmentation with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1280–1289, 2022. 1

[18] Youwei Liang, Chongjian Ge, Zhan Tong, Yibing Song, Jue Wang, and Pengtao Xie. Not all patches are what you need: Expediting vision transformers via token reorganizations. *arXiv preprint arXiv:2202.07800*, 2022. 1, 3, 6

[19] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021. 1, 7

[20] Lingchen Meng, Hengduo Li, Bor-Chun Chen, Shiyi Lan, Zuxuan Wu, Yu-Gang Jiang, and Ser-Nam Lim. Adavit: Adaptive vision transformers for efficient image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12309–12318, 2022. 3

[21] Bowen Pan, Rameswar Panda, Yifan Jiang, Zhangyang Wang, Rogerio Feris, and Aude Oliva. Ia-red2: Interpretability-aware redundancy reduction for vision trans-

formers. *Advances in Neural Information Processing Systems*, 34:24898–24911, 2021. 1, 3, 6

[22] Ilija Radosavovic, Raj Prateek Kosaraju, Ross Girshick, Kaiming He, and Piotr Dollár. Designing network design spaces. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10428–10436, 2020. 7

[23] Yongming Rao, Wenliang Zhao, Benlin Liu, Jiwen Lu, Jie Zhou, and Cho-Jui Hsieh. Dynamicvit: Efficient vision transformers with dynamic token sparsification. *Advances in neural information processing systems*, 34:13937–13949, 2021. 1, 3, 6

[24] Alex Rodriguez and Alessandro Laio. Clustering by fast search and find of density peaks. *science*, 344(6191):1492–1496, 2014. 2, 5

[25] Michael S Ryoo, AJ Piergiovanni, Anurag Arnab, Mostafa Dehghani, and Anelia Angelova. Tokenlearner: What can 8 learned tokens do for images and videos? *arXiv preprint arXiv:2106.11297*, 2021. 1, 4

[26] Yehui Tang, Kai Han, Yunhe Wang, Chang Xu, Jianyuan Guo, Chao Xu, and Dacheng Tao. Patch slimming for efficient vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12165–12174, 2022. 4, 6

[27] Yehui Tang, Kai Han, Chang Xu, An Xiao, Yiping Deng, Chao Xu, and Yunhe Wang. Augmented shortcuts for vision transformers. *Advances in Neural Information Processing Systems*, 34:15316–15327, 2021. 1, 4

[28] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*, pages 10347–10357. PMLR, 2021. 1, 2, 3, 6, 7

[29] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 1, 5

[30] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 568–578, 2021. 1, 7

[31] Haiping Wu, Bin Xiao, Noel Codella, Mengchen Liu, Xiyang Dai, Lu Yuan, and Lei Zhang. Cvt: Introducing convolutions to vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22–31, 2021. 1, 3

[32] Kai Xiao, Logan Engstrom, Andrew Ilyas, and Aleksander Madry. Noise or signal: The role of image backgrounds in object recognition. *arXiv preprint arXiv:2006.09994*, 2020. 2, 4

[33] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems*, 34:12077–12090, 2021. 1

[34] Weijian Xu, Yifan Xu, Tyler Chang, and Zhuowen Tu. Co-scale conv-attentional image transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9981–9990, 2021. 7

[35] Yifan Xu, Zhijie Zhang, Mengdan Zhang, Kekai Sheng, Ke Li, Weiming Dong, Liqing Zhang, Changsheng Xu, and Xing Sun. Evo-vit: Slow-fast token evolution for dynamic vision transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022. 1, 2, 3, 6

[36] Huanrui Yang, Hongxu Yin, Pavlo Molchanov, Hai Li, and Jan Kautz. Nvit: Vision transformer compression and parameter redistribution. *arXiv preprint arXiv:2110.04869*, 2021. 1

[37] Hongxu Yin, Arash Vahdat, Jose M Alvarez, Arun Mallya, Jan Kautz, and Pavlo Molchanov. A-vit: Adaptive tokens for efficient vision transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10809–10818, 2022. 1, 3, 6

[38] Li Yuan, Yunpeng Chen, Tao Wang, Weihao Yu, Yujun Shi, Zi-Hang Jiang, Francis EH Tay, Jiashi Feng, and Shuicheng Yan. Tokens-to-token vit: Training vision transformers from scratch on imagenet. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 558–567, 2021. 1, 3, 6, 7