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

Sifan Long[1] Zhen Zhao[2] Jimin Pi[2] Shengsheng Wang[1] Jingdong Wang[2]

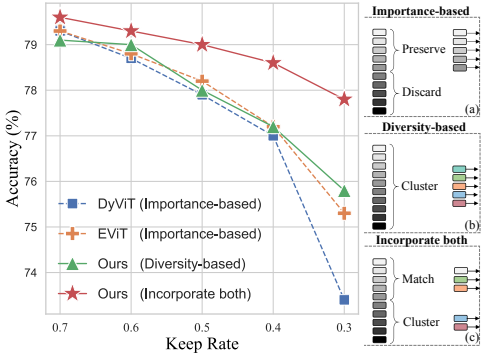
[1]CCST, Jilin University, Jilin, China

[2]Baidu VIS

CVPR 2023

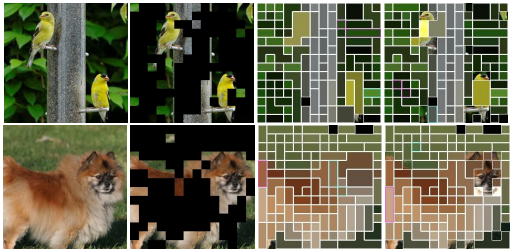
Outline

Importance-based *vs.* Diversity-based *vs.* Incorporate both



- Importance-based method preserves attentive tokens based on the class token attention and masks all inattentive tokens.
- Diversity-based method clusters similar tokens into a group and then combines tokens from the same group into a new token.
- ncorporate method decouples and merges tokens to consider token importance and diversity simultaneously.

Visualizations of pruning results 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.

The diversity score and classification accuracy on ImageNet

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

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

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.

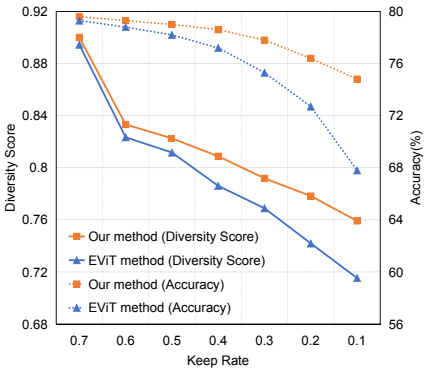
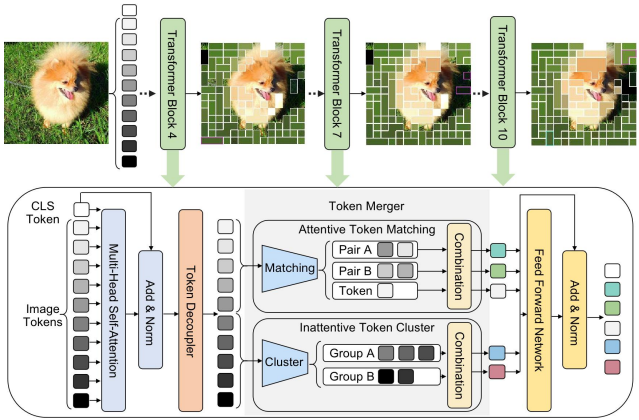


Illustration of our approach

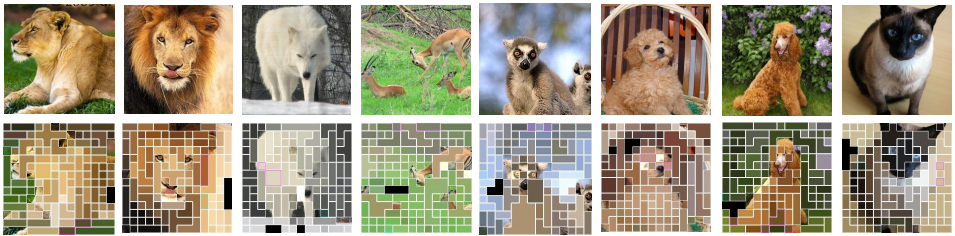


- (top) Employ our method at the 4th, 7th, and 10th layers of the DeiT-S model.
- (bottom) Model structure within a single transformer block.

Comparisons with existing token pruning methods on DeiT.

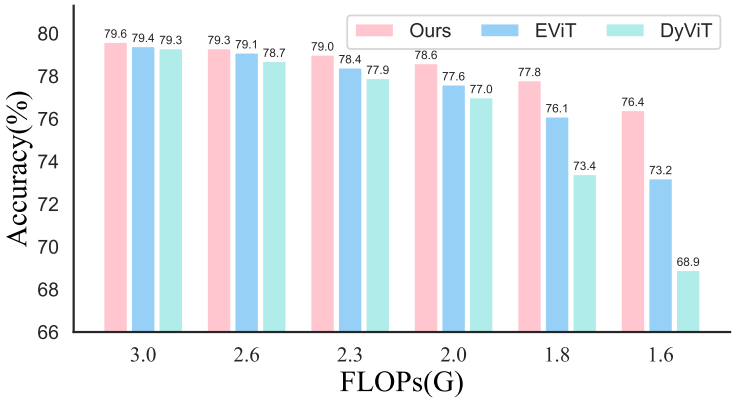
Model	Method	Top-1 Acc. (%)	Params (M)	FLOPs (G)	FLOPs ↓(%)	Throughput (img/s)
DeiT-T	DeiT-T	72.2	5.6	1.3	0.0	2536
	DyViT	71.2	5.9	0.9	30.8	3542
	PS-ViT	72.0	5.6	0.9	30.8	3563
	SViTE	70.1	4.2	0.9	30.8	2836
	SPViT	72.2	5.6	1.0	23.1	-
	Evo-ViT	72.0	5.6	0.8	38.5	3627
	Ours-DeiT-T	72.3	5.6	0.8	38.5	3641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

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.

Performance of existing methods on each keep rate



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

Conclusion

- 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.