Adaptive Token Sampling For Efficient Vision Transformers

ECCV 2022

Paper Reading by Yixuan Zhang

智能多媒体内容计算实验室
Intelligent Multimedia Content Computing Lab

- 作者介绍
- 研究背景
- 本文方法
- 实验效果
- 总结反思

作者介绍

Mohsen Fayyaz
University of Bonn
PhD Candidate
Computer Vision
Machine Learning

	All	Since 2018
Citations	1833	1787
h-index	17	16
i10-index	19	19

Soroush Abbasi
Koohpayegani
University of California
PhD Student
Backdoor Attacks
Self-supervised learning

	总计	2018 年至今
引用	177	177
h 指数	6	6
i10 指数	5	5

作者介绍

Mohsen Fayyaz
University of Bonn
Professor
Action recognition and video understanding
Anticipation and forecasting
Human pose estimation

	总计	2018 年至今
引用	17703	10594
h 指数	67	55
i10 指数	132	108

Hamed Pirsiavash
University of California, Davis
Associate Professor
Backdoor Attacks
Self-supervised learning

	总计	2018 年至今
引用	8452	5955
h 指数	32	27
i10 指数	46	38

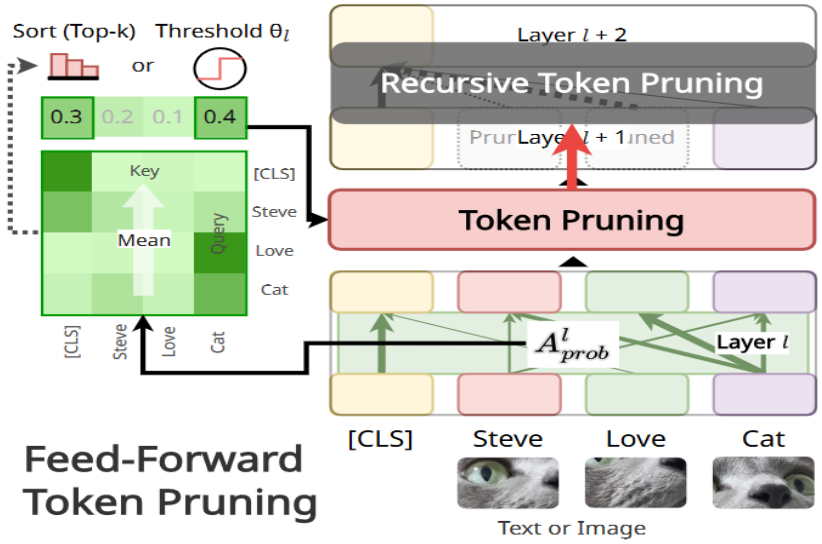
- 作者介绍
- 研究背景
- 本文方法
- 实验效果
- 总结反思

研究背景

- Transformer模型的缺点
  - 计算量/参数量大，对显存要求高，推断耗时
- Efficient Transformer模型
  - 剪枝
    - 注意力/权重：复杂度输入序列长度线性变化 $O(N^2)$ --- $O(N)$
    - Token：同时减少attenion和FFN的复杂度
  - 量化
    - K-Means的量化，binary的量化
  - 蒸馏
    - TinyBERT

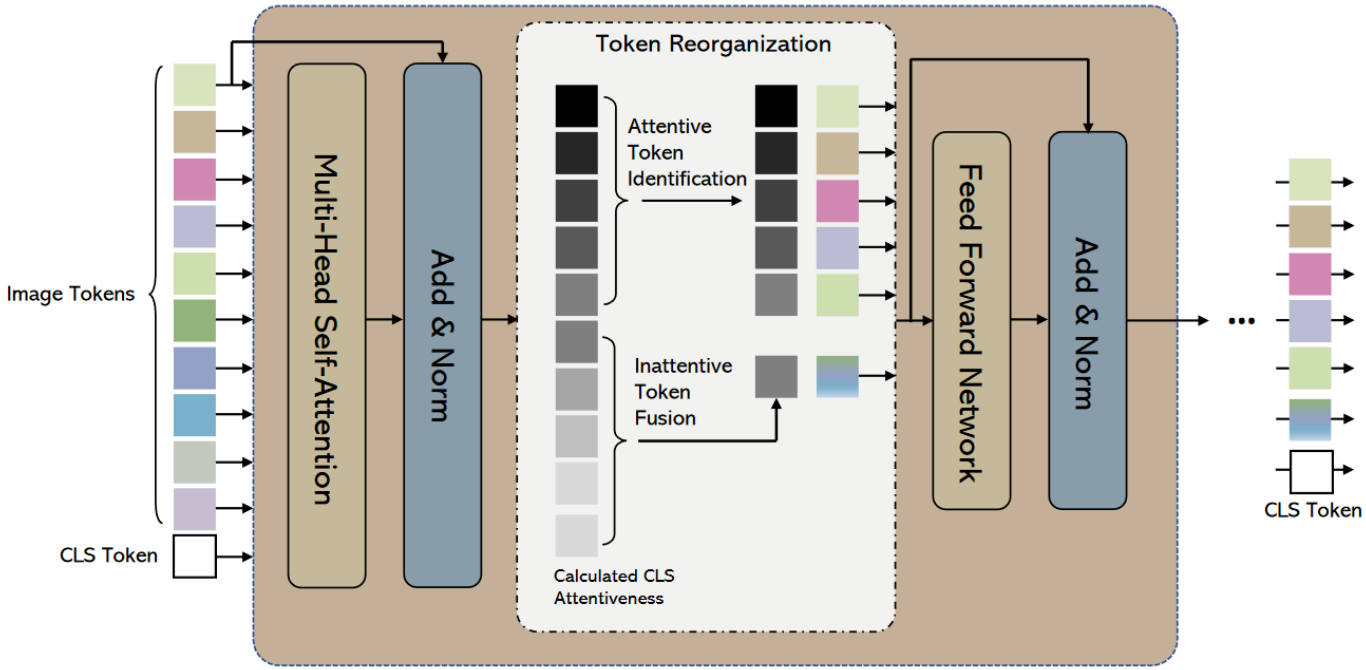
研究背景

- Token剪枝
  - 如何选取：单独模块计算重要性/attention+ topk/阈值
  - 如何减少：丢弃/合并
- 存在问题1：可能会在较早的层中删除对最终表示和任务损失很重要的token。



研究背景

- 存在问题1：可能会在较早的层中删除对最终表示和任务损失很重要的token。
  - 解决方法：融合剪掉的特征



Liang Y, Ge C, Tong Z, et al. Not all patches are what you need: Expediting vision transformers via token reorganizations[J]. arXiv preprint arXiv:2202.07800, 2022.

研究背景

- 存在问题1：可能会在较早的层中删除对最终表示和任务损失很重要的token。
  - 解决方法：融合相似的特征

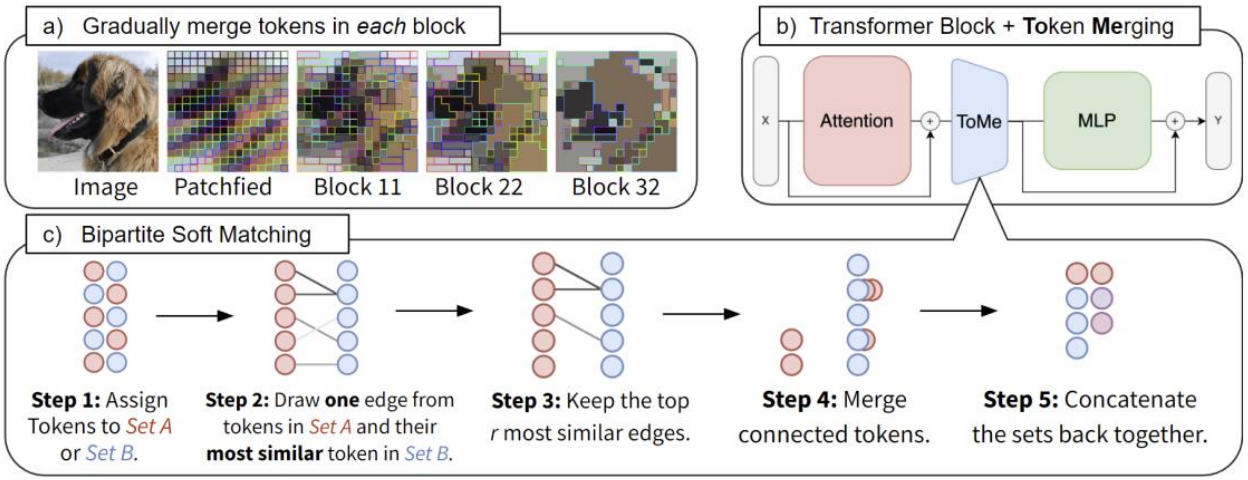
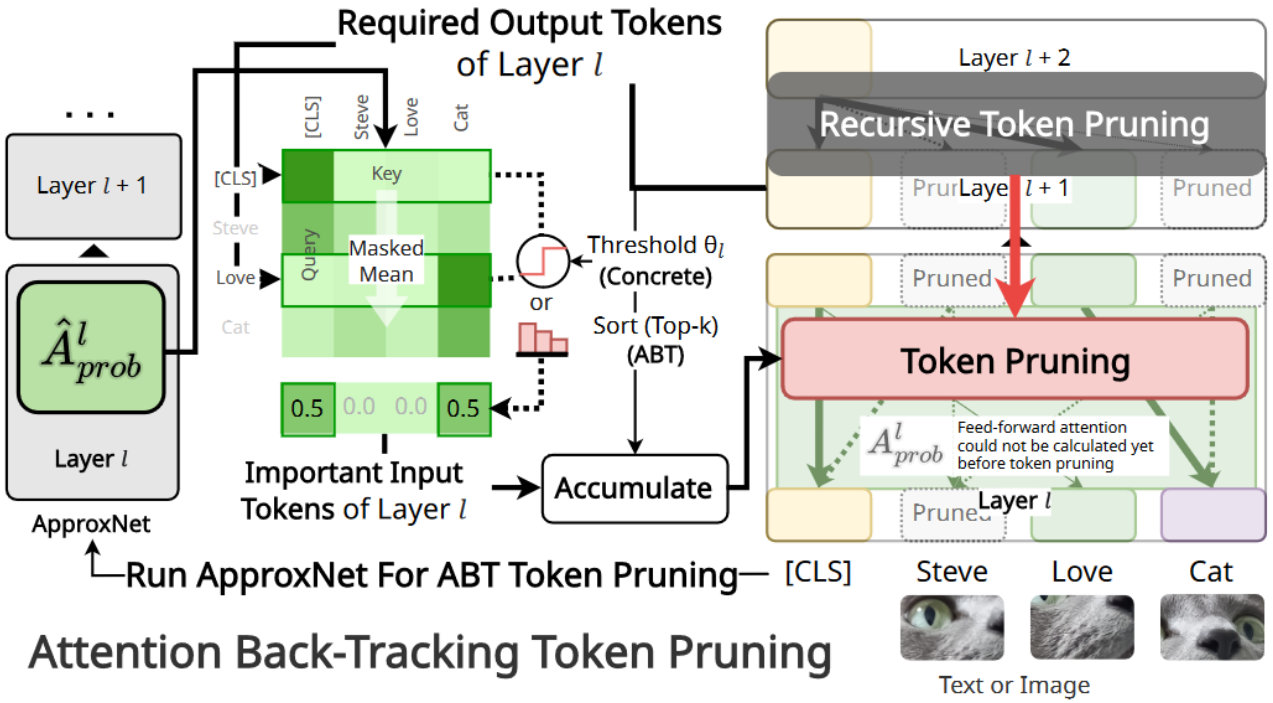


Figure 1: **Token Merging.** (a) With ToMe, similar patches are merged in each transformer block: for example, the dog's fur is merged into a single token. (b) ToMe is simple and can be inserted inside the standard transformer block. (c) Our fast merging algorithm, see Appendix D for implementation.

Bolya D, Fu C Y, Dai X, et al. Token Merging: Your ViT But Faster[J]. arXiv preprint arXiv:2210.09461, 2022.

研究背景

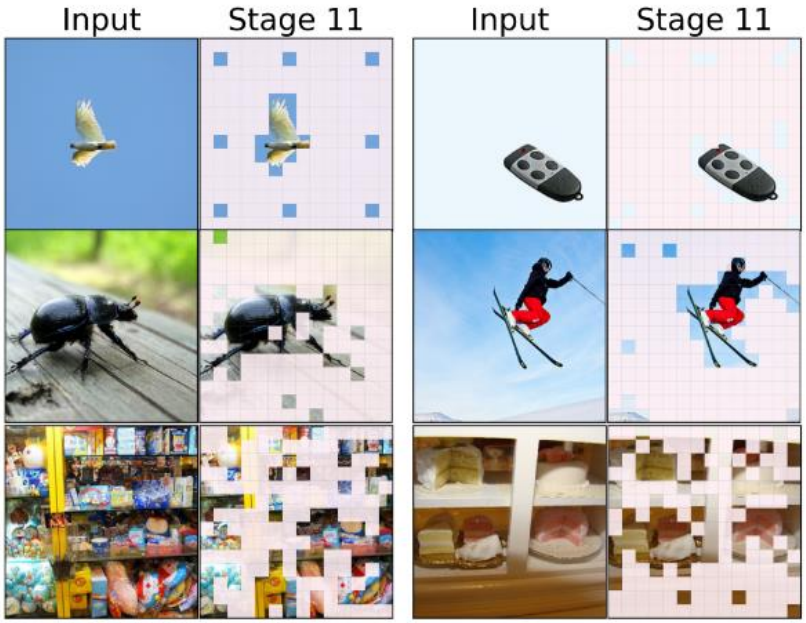
- 存在问题1：可能会在较早的层中删除对最终表示和任务损失很重要的token。
  - 解决方法：注意力回溯



Lee H, Kang M, Lee Y, et al. Sparse Token Transformer with Attention Back Tracking[C]//The Eleventh International Conference on Learning Representations.

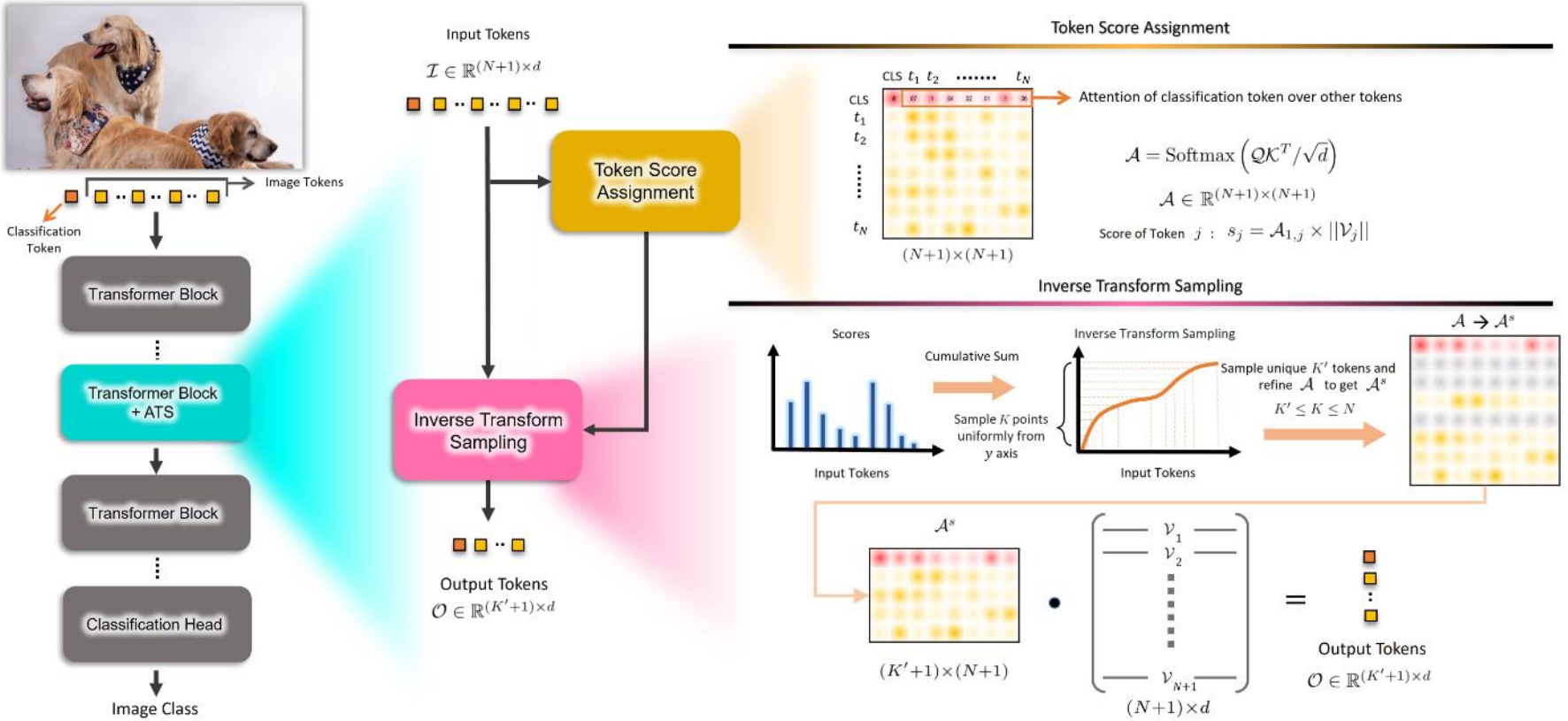
研究背景

- 存在问题2：token保留比率设置问题
  - 现有方法往往固定token保留率为定值，然而视觉图片/视频所有部分对最终分类分数的贡献并不相同，有些部分包含不相关或冗余的信息。相关信息的数量因图像或视频的内容而异。
  - 如果现有方法token保留率需要更改（例如，由于部署在不同的设备上），则可能需要重新训练。



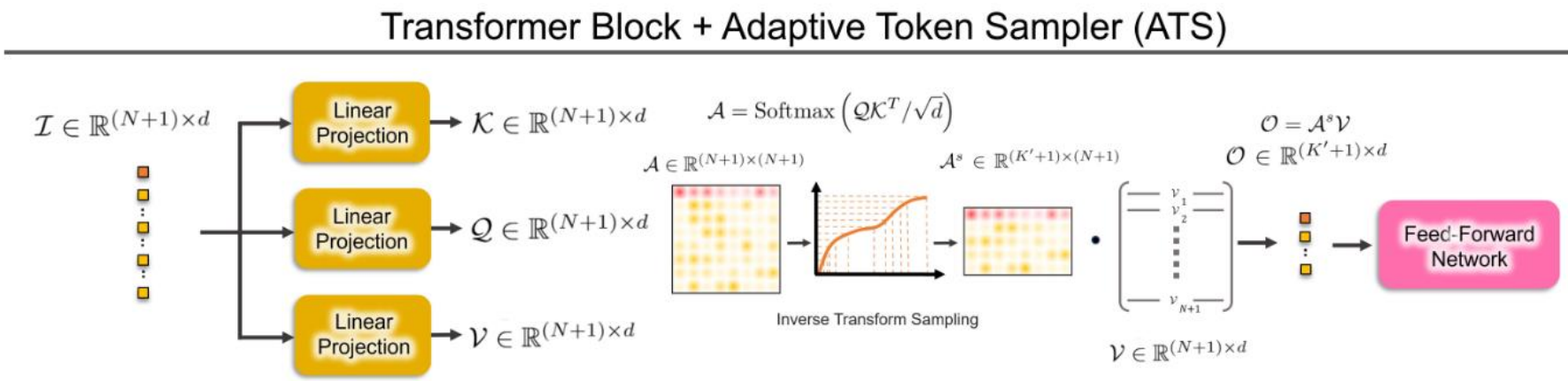
- 作者介绍
- 研究背景
- 本文方法
- 实验效果
- 总结反思

Adaptive Token Sampler



- K：定义采样的token数量的最大值，进而实现对GFLOPs 的上限的控制。
- 实际输出token数目：$K' + 1$ ($K' \leq K \leq N$)

Token Score Assignment



$$\mathcal{A} = \text{Softmax}\left(\mathcal{Q}\mathcal{K}^T/\sqrt{d}\right). \qquad \mathcal{O} = \mathcal{A}\mathcal{V}.$$

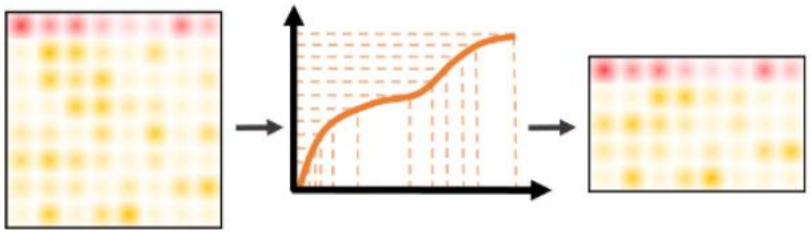
- Significant score: $\mathcal{A}_{1,2}, \ldots, \mathcal{A}_{1,N+1}$ $\qquad \mathcal{O} = \mathcal{A}^s\mathcal{V}.$

$$\mathcal{S}_j = \frac{\mathcal{A}_{1,j} \times ||\mathcal{V}_j||}{\sum_{i=2} \mathcal{A}_{1,i} \times ||\mathcal{V}_i||}$$

Inverse Transform Sampling

$$\mathcal{A} = \text{Softmax}\left(\mathcal{Q}\mathcal{K}^T/\sqrt{d}\right)$$

$\mathcal{A} \in \mathbb{R}^{(N+1)\times(N+1)}$ $\mathcal{A}^s \in \mathbb{R}^{(K'+1)\times(N+1)}$



Inverse Transform Sampling

$$\text{CDF}_i = \sum_{j=2} \mathcal{S}_j.$$

$$\Psi(k) = \text{CDF}^{-1}(k) \quad k \in [0,1]$$

- 早期特征的辨别力较低，可能有多个具有相似key的多个token，Softmax 函数将降低它们相应的注意力权重。--- 存在问题1
- Topk方法：不能自适应地选择 K‘ ≤ K 个标记。--- 存在问题2
- 采样结果：
  - K′ < K $S_j \geq 2/K$
  - K′ = K token之间非常相似
  - K′ = 1 token之间非常独立

  --- 早期阶段比后期阶段选择更多token

  --- 判别性高的图片选择token少

  不同阶段、不同图片的自适应采样

- U [0, 1]均匀采样K次 $k = \{\frac{1}{2K}, \frac{3}{2K} \cdots, \frac{2K-1}{2K}\}$
- 最接近score分数的token即为采样token，多次采到相同token时只保留一次。

$\mathcal{O} = \mathcal{A}^s\mathcal{V}. \quad \mathcal{A}^s \in \mathbb{R}^{(K'+1)\times(N+1)}$

- 作者介绍
- 研究背景
- 本文方法
- 实验效果
- 总结反思

SOTA

- ImageNet validation set

Model	Params (M)	GFLOPs	Resolution	Top-1
ViT-Base/16 [13]	86.6	17.6	224	77.9
HVT-S-1 [42]	22.09	2.4	224	78.0
IA-RED^2 [41]	-	2.9	224	78.6
DynamicViT-DeiT-S (30 Epochs) [46]	22.77	2.9	224	79.3
EViT-DeiT-S (30 epochs) [36]	22.1	3.0	224	79.5
DeiT-S+ATS (Ours)	**22.05**	**2.9**	224	79.7
DeiT-S [53]	22.05	4.6	224	79.8
PVT-Small [60]	24.5	3.8	224	79.8
CoaT Mini [64]	10.0	6.8	224	80.8
CrossViT-S [5]	26.7	5.6	224	81.0
PVT-Medium [60]	44.2	6.7	224	81.2
Swin-T [39]	29.0	4.5	766	81.3
T2T-ViT-14 [67]	22.0	5.2	224	81.5
CPVT-Small-GAP [8]	23.0	4.6	817	81.5
CvT-13 [63]	20.0	4.5	224	81.6
CvT-13+ATS (Ours)	20.0	**3.2**	224	81.4
PS-ViT-B/14 [68]	21.3	5.4	224	81.7
PS-ViT-B/14+ATS (Ours)	21.3	**3.7**	224	81.5
RegNetY-8G [44]	39.0	8.0	224	81.7
DeiT-Base/16 [53]	86.6	17.6	224	81.8
CoaT-Lite Small [64]	20.0	4.0	224	81.9
T2T-ViT-19 [67]	39.2	8.9	224	81.9
CrossViT-B [5]	104.7	21.2	224	82.2
T2T-ViT-24 [67]	64.1	14.1	224	82.3
PS-ViT-B/18 [68]	21.3	8.8	224	82.3
PS-ViT-B/18+ATS (Ours)	21.3	**5.6**	224	82.2
CvT-21 [63]	32.0	7.1	224	82.5
CvT-21+ATS (Ours)	32.0	**5.1**	224	82.3
TNT-B [22]	66.0	14.1	224	82.8
RegNetY-16G [44]	84.0	16.0	224	82.9
Swin-S [39]	50.0	8.7	224	83.0
CvT-13$_{384}$ [63]	20.0	16.3	384	83.0
CvT-13$_{384}$+ATS (Ours)	20.0	**11.7**	384	82.9
Swin-B [39]	88.0	15.4	224	83.3
LV-ViT-S [30]	26.2	6.6	224	83.3
CvT-21$_{384}$ [63]	32.0	24.9	384	83.3
CvT-21$_{384}$+ATS (Ours)	32.0	**17.4**	384	83.1

SOTA

Table 2. Comparison with state-of-the-art on Kinetics-400.

Model	Top-1	Top-5	Views	GFLOPs
STC [10]	68.7	88.5	112	-
bLVNet [15]	73.5	91.2	3×3	840
STM [37]	73.7	91.6	-	-
TEA [35]	76.1	92.5	10×3	2,100
TSM R50 [29]	74.7	-	10×3	650
I3D NL [62]	77.7	93.3	10×3	10,800
CorrNet-101 [58]	79.2	-	10×3	6,700
ip-CSN-152 [55]	79.2	93.8	10×3	3,270
HATNet [11]	79.3	-	-	-
SlowFast 16×8 R101+NL [18]	79.8	93.9	10×3	7,020
X3D-XXL [17]	80.4	94.6	10×3	5,823
TimeSformer-L [1]	80.7	94.7	1×3	7,140
TimeSformer-L+ATS (Ours)	80.5	94.6	1×3	**3,510**
ViViT-L/16x2 [1]	80.6	94.7	4×3	17,352
MViT-B, 64×3 [14]	81.2	95.1	3×3	4,095
X-ViT (16×) [2]	80.2	94.7	1×3	425
X-ViT+ATS (16×) (Ours)	80.0	94.6	1×3	**259**
TokenLearner 16at12 (L/16) [49]	82.1	-	4×3	4,596

Table 3. Comparison with state-of-the-art on Kinetics-600.

Model	Top-1	Top-5	Views	GFLOPs
AttentionNAS [61]	79.8	94.4	-	1,034
LGD-3D R101 [43]	81.5	95.6	10×3	-
HATNET [11]	81.6	-	-	-
SlowFast R101+NL [18]	81.8	95.1	10×3	3,480
X3D-XL [17]	81.9	95.5	10×3	1,452
X3D-XL+ATFR [16]	82.1	95.6	10×3	768
TimeSformer-HR [1]	82.4	96	1×3	5,110
TimeSformer-HR+ATS (Ours)	82.2	96	1×3	**3,103**
ViViT-L/16x2 [1]	82.5	95.6	4×3	17,352
Swin-B [39]	84.0	96.5	4×3	3,384
MViT-B-24, 32×3 [14]	84.1	96.5	1×5	7,080
TokenLearner 16at12(L/16) [49]	84.4	96.0	4×3	9,192
X-ViT (16×) [2]	84.5	96.3	1×3	850
X-ViT+ATS (16×) (Ours)	84.4	96.2	1×3	**521**

Ablation Experiments

- Significant scores

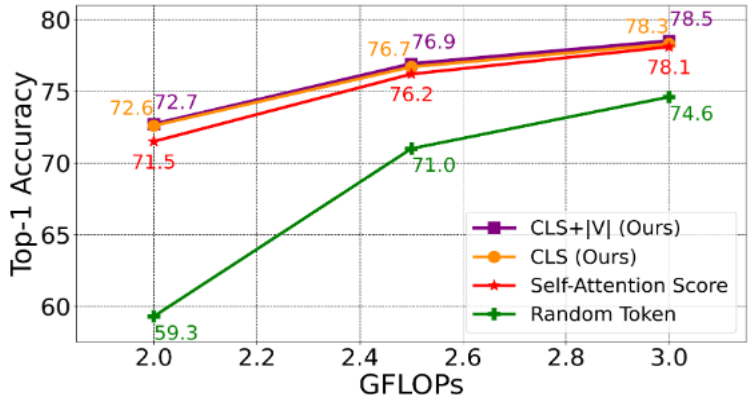


Fig. 3. Impact of different score assignment methods. To achieve different GFLOPs levels, we bound the value of K from above such that the average GFLOPs of our adaptive models over the ImageNet validation set reaches the desired level. For more details, please refer to the supplementary material.

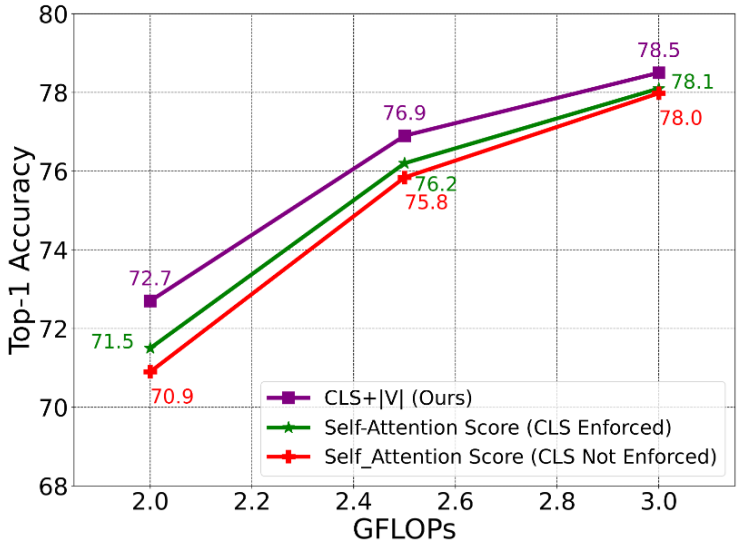


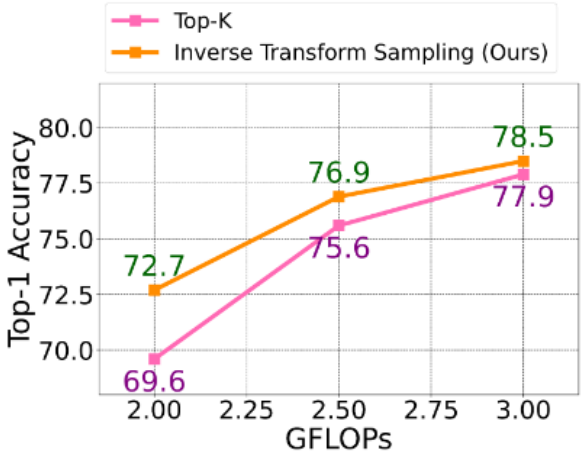
Fig. A.4. Impact of allowing ATS to discard the classification token on the network's accuracy. The model is a single stage DeiT-S+ATS without finetuning.

Ablation Experiments

- Candidate Tokens Selection

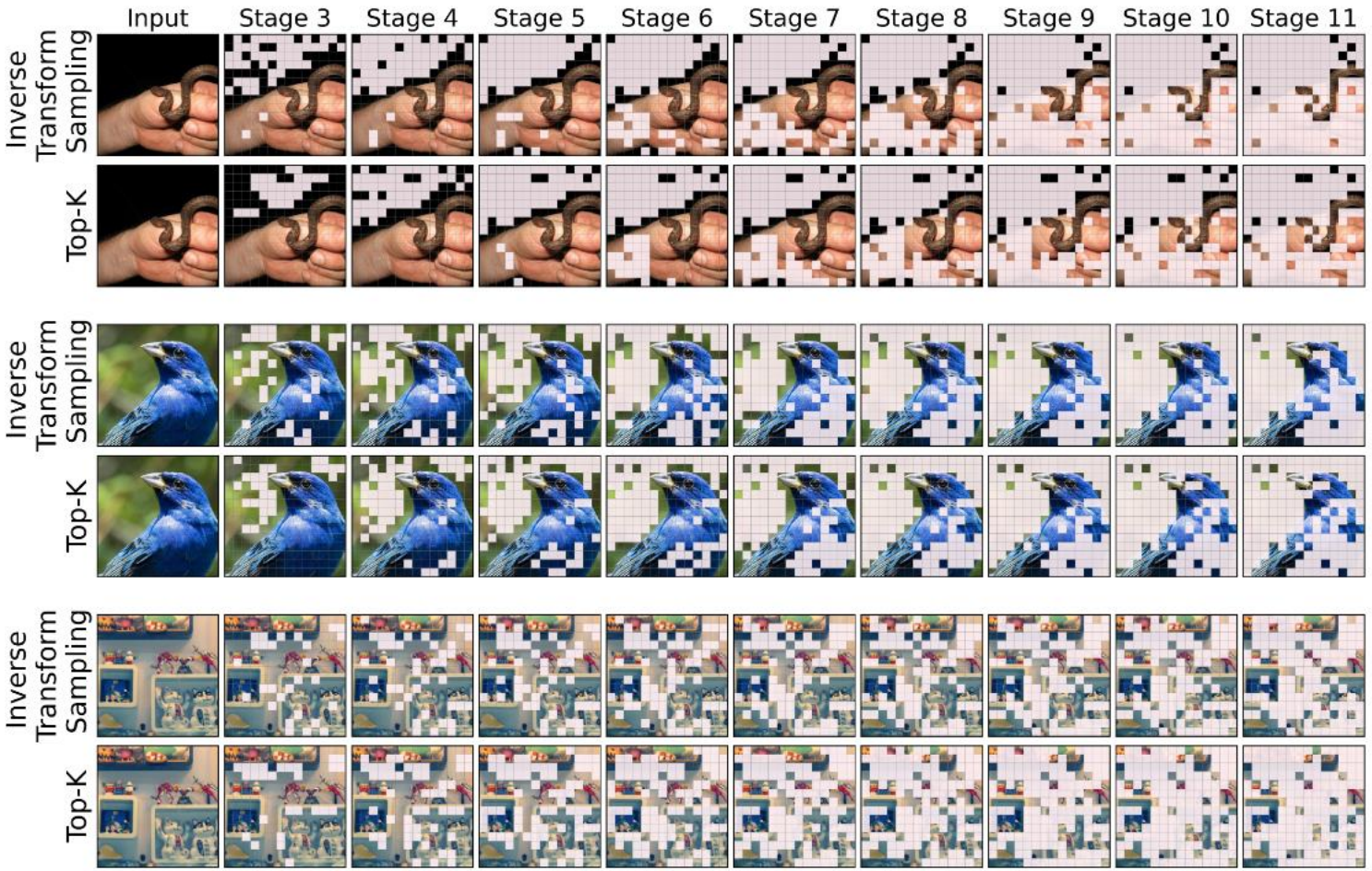
Table A.3. Comparison of the inverse transform sampling approach with the top-K selection. We finetune and test two different versions of the multi-stage DeiT-S+ATS model: with (1) top-K token selection and (2) inverse transform token sampling. We report the top-1 accuracy of both networks on the ImageNet validation set. For the model with the top-K selection approach, we set $K_n = \lfloor 0.865 \times \#InputTokens_n \rceil$ where n is the stage index. For example, $K_3 = 171$ in stage 3.

Method	Top-1 acc	GFLOPs
Top-K	78.9	2.9
Inverse Transform Sampling	**79.7**	2.9

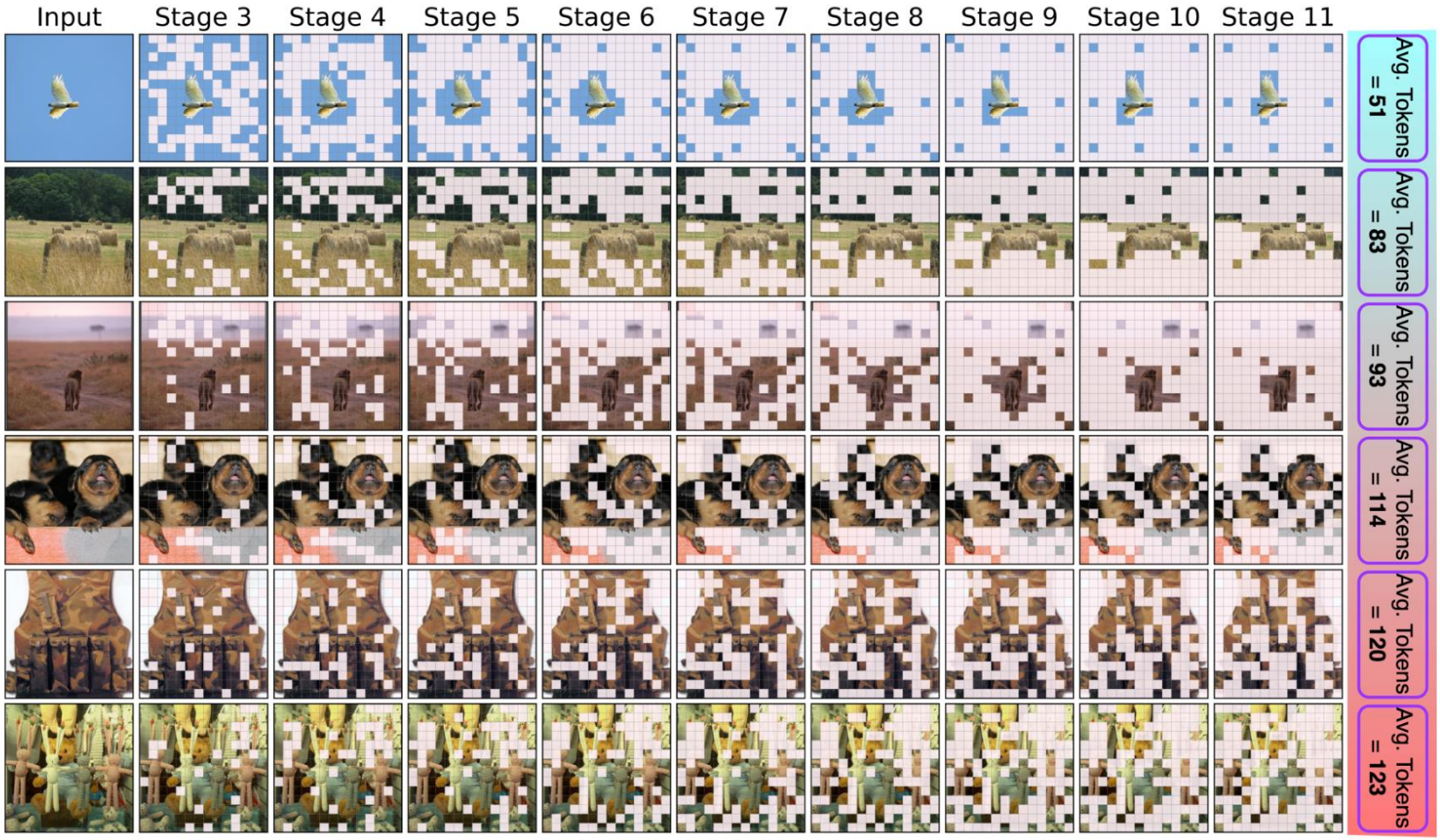


(a) Sampling Methods

Visualization



Visualization



Visualization

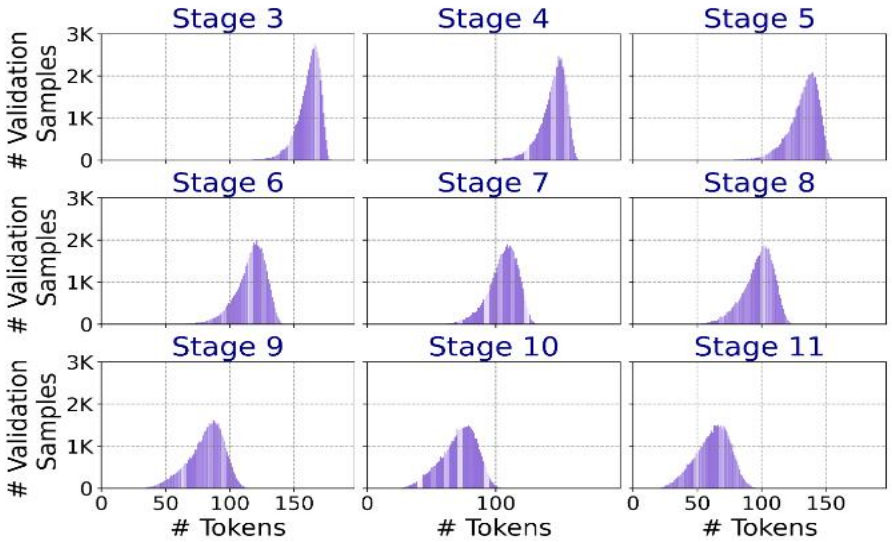


Fig. 6. Histogram of the number of sampled tokens at each ATS stage of our multi-stage DeiT-S+ATS model on the ImageNet validation set. The y-axis corresponds to the number of images and the x-axis to the number of sampled tokens.

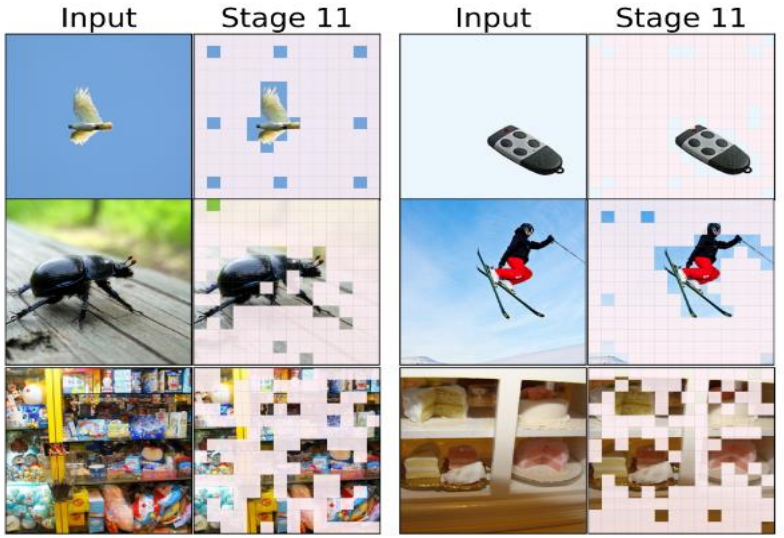


Fig. 7. ATS samples less tokens for images with fewer details (top), and a higher number of tokens for more detailed images (bottom). We show the token downsampling results after all ATS stages. For this experiment, we use a multi-stage Deit-S+ATS model.

Ablation Experiments

- Model scaling & Fine-tuning

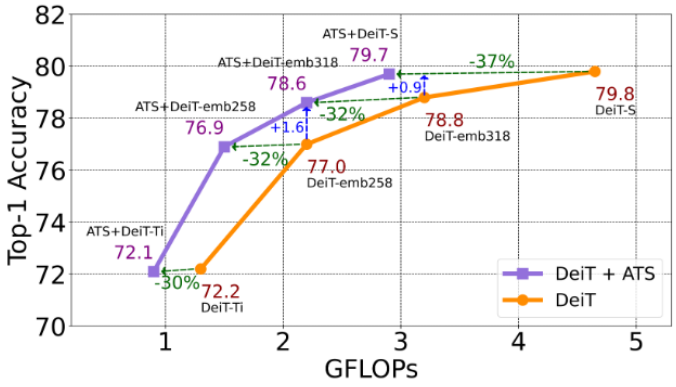
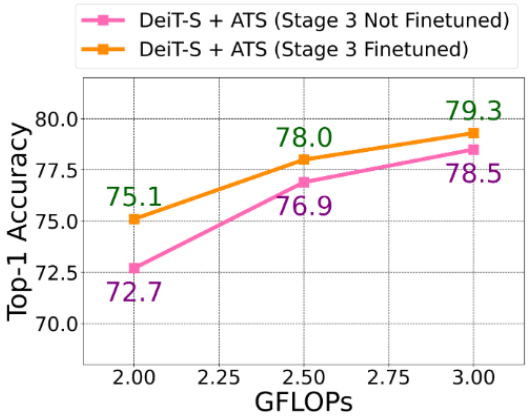
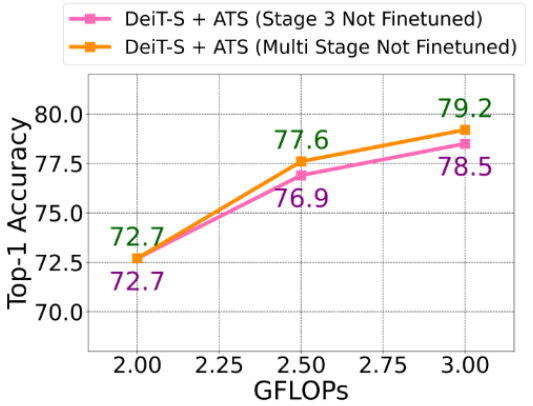


Fig. 4. Performance comparison on the ImageNet validation set. Our proposed adaptive token sampling method achieves a state-of-the-art trade-off between accuracy and GFLOPs. We can reduce the GFLOPs of DeiT-S by 37% while almost maintaining the accuracy.



(b) Fine-tuning



(c) Multi vs. Single Stage

Ablation Experiments

Table A.1. Runtime comparison: We run the models on a single RTX6000 GPU (CUDA 11.0, PyTorch 1.8, image size: 224×224). We average the value of throughput over 20 runs. We add ATS to multiple stages of the DeiT-S model and fine-tune the network on the ImageNet dataset.

Model	Params (M)	GFLOPs	Throughput	Top-1
Deit-S [53]	22.05	4.6	1010	79.8
Deit-S+ATS	22.05	**2.9**	**1403**	79.7

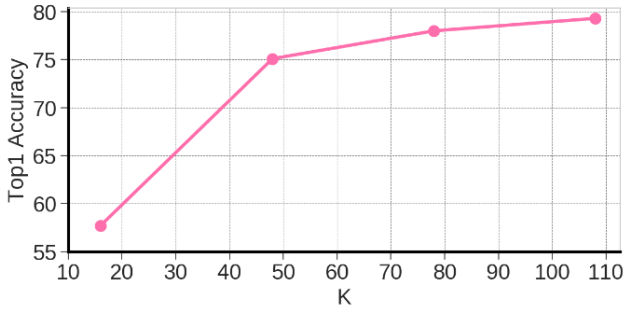


Fig. A.1. Effect of K: We varied the value of K in the ATS module to study the effect of K on the top-1 accuracy. K=48 corresponds to 2 GFLOPs. The backbone model is DeiT-S pre-trained on ImageNet-1K.

Ablation Experiments

Stage(s)	0	3	6	3-11
Top-1 Accuracy	73.1	78.5	77.4	**79.2**

Table A.4. Evaluating the integration of the ATS module into different stages of DeiT-S [53].

Model	Top-1 acc	GFLOPs
EViT-DeiT-S (30 Epochs) [36]	79.5	3.0
EViT-DeiT-S (30 Epochs)+ATS	79.5	**2.5**
EViT-DeiT-S (100 Epochs) [36]	79.8	3.0
EViT-DeiT-S (100 Epochs)+ATS	79.8	**2.5**

Table A.5. Evaluating the EViT-DeiT-S [36] model's performance when integrating the ATS module into it with $K_n = \lfloor 0.7 \times \#InputTokens_n \rceil$ where n is the stage index.

- 作者介绍
- 研究背景
- 本文方法
- 实验效果
- 总结反思

总结反思

- 总结
  - 设计思路简单，无需额外参数，扩展性好
  - 能够实现一定的自适应，K的取值仍然需要一定的限制。

Thanks for Attention!