Appendix – Efficient Transformer Adaptation with Soft Token Merging

Xin Yuan[1], Hongliang Fei[2], Jinoo Baek[2]
[1]University of Chicago [2]Google
yuanx@uchicago.edu {hongliangfei,jinoo}@google.com

1. More results in DeiT-S

We further compare ours and recently proposed progressive token pruning approaches on Deit-S by showing additional Top-1 accuracy on ImageNet-1K, FLOPs, and inference throughput. Table 1, 2, 3 and 4 demonstrate that our approach outperforms all the competitors consistently.

Table 1. Comparisons on ImageNet for fine-tuning DeiT-S. For competing methods, we set the token kept ratio as 0.4 while for our approach the merging position l are set as 3.

Method	Top-1 Acc(%)	FLOPs(G)	Infer Tput.(imgs/s)
IdleViT	78.4	2.1	4363
DyViT	76.0	1.9	**5741**
EViT	77.6	2.0	3717
Evo-ViT	77.5	2.1	3548
ATS	76.4	2.0	2580
Ours	**78.7**	2.0	4843

Table 2. Comparisons on ImageNet for fine-tuning DeiT-S. For competing methods, we set the token kept ratio as 0.6 while for our approach the merging position l are set as 5.

Method	Top-1 Acc(%)	FLOPs(G)	Infer Tput.(imgs/s)
IdleViT	79.3	2.7	3693
DyViT	78.5	2.5	**4474**
EViT	78.9	2.6	3045
Evo-ViT	78.0	2.6	2998
ATS	78.9	2.7	2229
Ours	**79.6**	2.7	4002

Table 3. Comparisons on ImageNet for fine-tuning DeiT-S. For competing methods, we set the token kept ratio as 0.7 while for our approach the merging position l are set as 6.

Method	Top-1 Acc(%)	FLOPs(G)	Infer Tput.(imgs/s)
IdleViT	79.6	3.1	3361
DyViT	79.3	3.0	3390
EViT	79.5	3.0	2621
Evo-ViT	78.2	3.0	2606
ATS	79.2	3.1	2161
Ours	**79.7**	3.1	**3408**

2. More results in LV-ViT-S

We detail more results in terms of Top-1 accuracy and FLOPs, as shown in Table 5, 6, and 7. We additionally provide inference throughput to demonstrate the wall-clock acceleration.

Table 4. Comparisons on ImageNet for fine-tuning DeiT-S. For competing methods, we set the token kept ratio as 0.8 while for our approach the merging position l are set as 7.

Method	Top-1 Acc(%)	FLOPs(G)	Infer Tput.(imgs/s)
IdleViT	**79.9**	3.5	3031
DyViT	79.6	3.4	**3405**
EViT	79.8	3.5	2286
Evo-ViT	78.4	3.5	2293
ATS	79.6	3.4	2036
Ours	**79.9**	3.5	3321

Table 5. Comparisons on ImageNet for fine-tuning LV-ViT-S. For competing methods, we set the token kept ratio as 0.8 while for our approach the merging position l are set as 7.

Method	Top-1 Acc(%)	FLOPs(G)	Infer Tput.(imgs/s)
IdleViT	83.2	5.1	855
DyViT	83.2	5.1	958
Ours	**83.3**	5.0	**970**

Table 6. Comparisons on ImageNet for fine-tuning LV-ViT-S. For competing methods, we set the token kept ratio as 0.7 while for our approach the merging position l are set as 6.

Method	Top-1 Acc(%)	FLOPs(G)	Infer Tput.(imgs/s)
IdleViT	83.1	4.5	938
DyViT	83.0	4.6	**1077**
Ours	**83.2**	4.5	1002

Table 7. Comparisons on ImageNet for fine-tuning LV-ViT-S. For competing methods, we set the token kept ratio as 0.6 while for our approach the merging position l are set as 5.

Method	Top-1 Acc(%)	FLOPs(G)	Infer Tput.(imgs/s)
IdleViT	82.9	4.0	1040
DyViT	82.6	4.2	**1206**
Ours	**83.0**	4.0	1188