Pruning One More Token is Enough

Nick John Eliopoulos[1], Purvish Jajal[1], James C. Davis[1], Gaowen Liu[2], George K. Thiravathukal[3], Yung-Hsiang Lu[1]

PURDUE UNIVERSITY 1 CISCO 2 LOYOLA UNIVERSITY CHICAGO 3

WACV 2025 TUCSON, ARIZONA • FEB 28 - MAR 4

Background

- Vision Transformers (ViTs) have quadratic complexity with respect to input size.
- ***Token Pruning / Merging:*** Identify and remove / merge unnecessary tokens

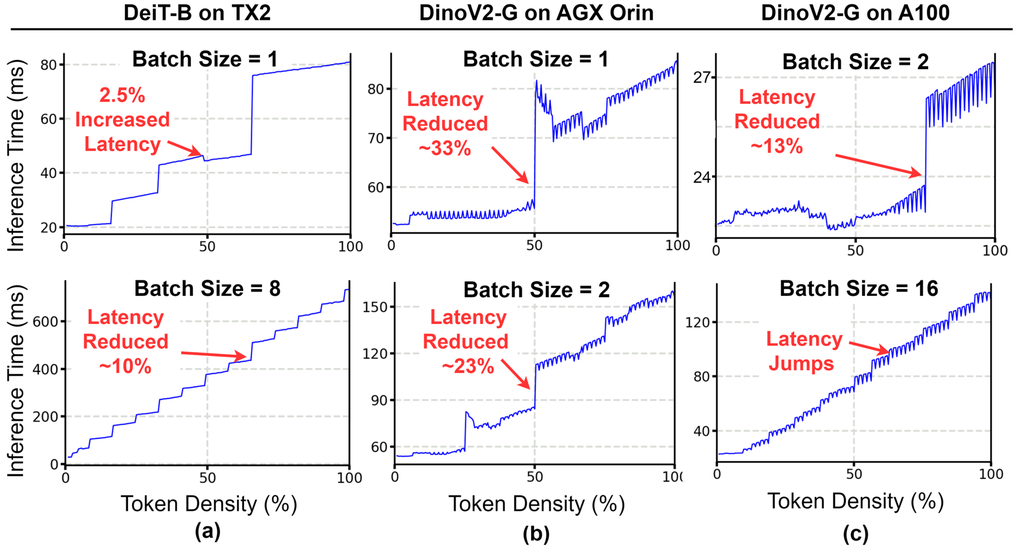
Contributions

- We consider hardware and workload size to determine how many tokens to prune
- Optimize token pruning for small workloads, where low-latency is more important than high-throughput

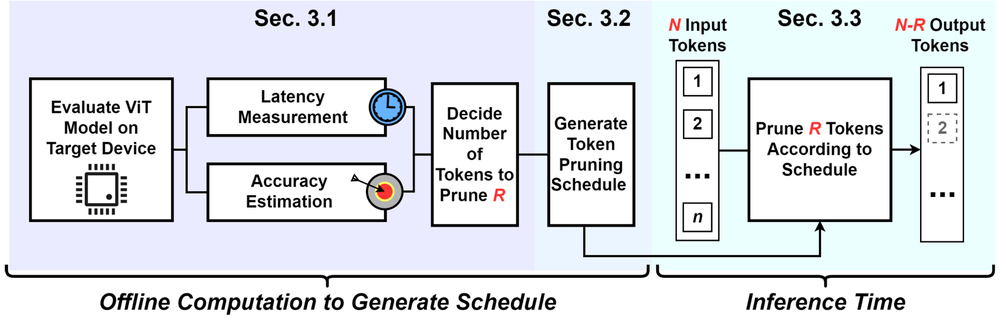
Paper

Code

Why Care About Device & Workload Size?



Our Approach



ImageNet1K - Equal Tokens Pruned

Device	Batch Size	Model	↓Top-1 Loss	↓Median Latency (ms)
TX2	1	DeiT-S		35.32
		Top-K	0.73	(+18.6%) 43.38
		ToMe	0.27	(+30.3%) 50.70
		Ours	0.99	(-9.06%) 32.12
A100	4	ViT-L		9.49
		Top-K	0.47	(+51.4%) 14.36
		ToMe	0.29	(+134.4%) 22.24
		Ours	0.85	(+40.0%) 13.29
A100	16	ViT-L		29.41
		w/ Top-K	0.77	(+2.48%) 30.14
		w/ ToMe	0.51	(+7.65%) 31.66
		w/ Ours	2.26	(-26.3%) 21.68

Table 8. Experiment that illustrate pruning overhead for certain workload sizes. Pruning parameters were chosen such that a similar number of tokens were pruned as our method. Token pruning methods may *increase* latency due to the overhead of pruning itself. We reduce overhead through single-layer pruning.

Other methods may *INCREASE* latency with token pruning!