Token Turing Machines are Efficient Vision Models

PURDUE UNIVERSITY [1] LOYOLA UNIVERSITY CHICAGO [2]

Purvish Jajal[1], Nick John Eliopoulos[1], Benjamin Shiue-Hal Chou[1], George K. Thiruvathukal[2], James C. Davis[1], Yung-Hsiang Lu[1]

WACV 2025 TUCSON, ARIZONA • FEB 28 - MAR 4

Background

- Vision Transformers (ViTs) have quadratic complexity with respect to input size.
- Computational costs can be reduced by processing fewer tokens *but* this compromises accuracy.

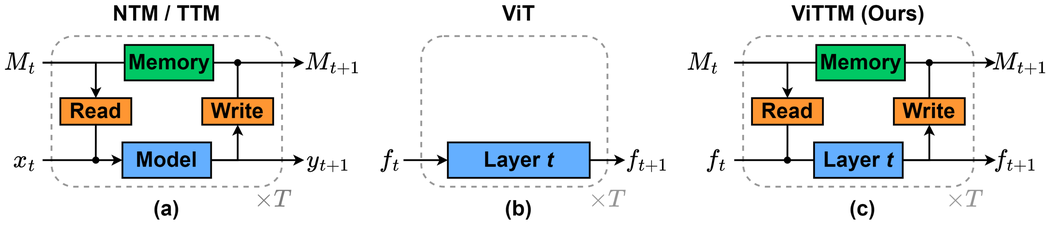
Contributions

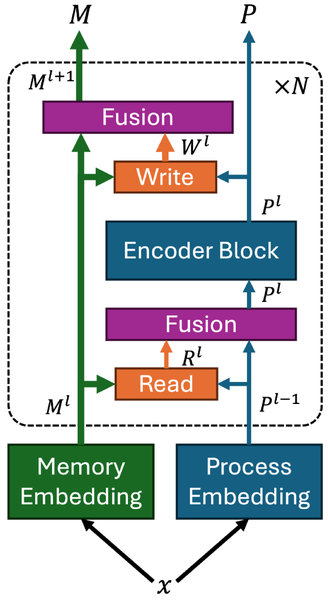
- A novel efficient memory-augmented ViT (ViTTM) with competitive accuracy-latency trade-offs for classification and segmentation.
- A comprehensive ablation exploring various designs and identify key decisions that impact the quality of the architecture.

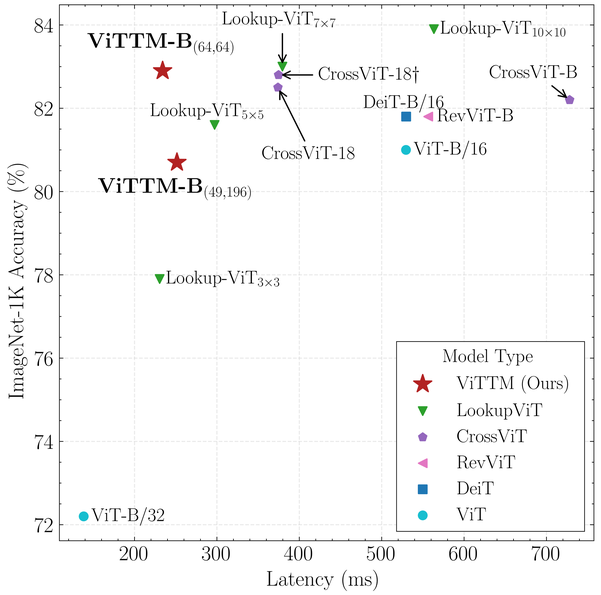
Paper

Code

Token Turing Machines + ViTs = *Efficient* ViTs







ImageNet 1K, *56% faster*

Model Class	Model	Params (M)	GFLOPs ↓	Latency (ms)↓	Top-1(%)↑
ViT/DeiT	ViT-S/16	22	4.25	149.5	74.2
	DeiT-S/16	22	4.25	152.0	79.8
	ViT-B/32	88	4.37	138.3	72.2
	ViT-B/16	87	16.87	529.5	81.0
	DeiT-B/16	87	16.87	529.7	81.8
Two-Stream	CrossViT-S	27	5.63	235.7	81.0
	CrossViT-15	28	5.81	249.1	82.3
	CrossViT-15†	28	6.13	252.3	81.5
	CrossViT-B	105	21.22	728.1	82.2
	CrossViT-18	43	9.05	374.1	82.5
	CrossViT-18†	44	9.50	378.2	82.8
	Rev-ViT-B	86	17.49	556.5	81.8
	LookupViT$_{3\times3}$	90	5.26	230.5	77.9
	LookupViT$_{5\times5}$	90	6.94	297.2	81.6
	LookupViT$_{7\times7}$	90	9.45	379.5	83.0
	LookupViT$_{10\times10}$	90	14.80	563.4	83.9
Token Merging [1]	ViT-S/16$_{(r=2)}$	22	4.31	172.7	74.0
	ViT-S/16$_{(r=4)}$	22	4.02	161.6	73.8
	ViT-S/16$_{(r=8)}$	22	3.41	138.2	73.1
	ViT-S/16$_{(r=10)}$	22	3.14	125.9	72.5
	ViT-S/16$_{(r=12)}$	22	2.85	115.2	71.6
	ViT-S/16$_{(r=14)}$	22	2.57	103.3	70.4
	ViT-S/16$_{(r=16)}$	22	2.30	94.0	68.1
	ViT-B/16$_{(r=2)}$	86	16.46	551.1	81.0
	ViT-B/16$_{(r=4)}$	86	15.34	515.7	80.9
	ViT-B/16$_{(r=8)}$	86	13.12	440.6	80.4
	ViT-B/16$_{(r=10)}$	86	12.02	402.0	80.1
	ViT-B/16$_{(r=12)}$	86	10.93	367.0	79.6
	ViT-B/16$_{(r=14)}$	86	9.84	330.2	78.9
	ViT-B/16$_{(r=16)}$	86	8.78	296.4	77.6
Ours	ViTTM-S$_{(64,64)}$	33	1.84	77.7	79.2
	ViTTM-B$_{(64,64)}$	127	7.08	234.1	82.9
	ViTTM-B$_{(49,196)}$	125	7.10	251.5	80.9

Semantic Segmentation, *+94% FPS*

Model (M)	FPS (A100) ↑	FPS (A30) ↑	mIoU ↑
ViT-B/16$_{384}$ [35]	23.8	13.8	48.06
ViT-B/16$_{224}$	23.8	13.8	45.65
ViTTM-B$_{(64,64)}$	*(+37%)* 32.5	*(+94%)* 26.8	45.17
ViTTM-B$_{(49,196)}$	*(+38%)* 32.8	*(+94%)* 26.7	43.60

Measured on the ADE20K dataset using Segmenter.