Evo-ViT: Slow-Fast Token Evolution for Dynamic Vision Transformer

Yifan Xu*[1], Zhijie Zhang*[2], Mengdan Zhang[3], Kekai Sheng[3], Ke Li[3], Weiming Dong[1],
Liqing Zhang[2], Changsheng Xu[1], Xing Sun[3]
[1] NLPR, Institute of Automation, Chinese Academy of Sciences,
[2] Shanghai Jiao Tong University, [3] Tencent Youtu Lab

Introduction

Motivation. The computation inefficiency is still a severe issue for vision transformers (ViTs). Existing token pruning methods for redundancy reduction are restricted in two aspects: 1) inapplicability on structured compressed transformers; 2) inability to train from scratch.

Methodology. To address the limitations, we propose a slow-fast token evolution approach for dynamic vision transformers (Evo-ViT). We distinguish the informative tokens from the placeholder tokens (uninformative tokens) for each instance in an unstructured and dynamic way, and update the two types of tokens with different computation paths.

Results. Experimental results demonstrate that our approach significantly accelerates various state-of-the-art ViTs while maintaining comparable accuracy. For example, our approach accelerates DeiT-S over 60% but only sacrifices 0.4% accuracy.

Code: https://github.com/YifanXu74/Evo-ViT

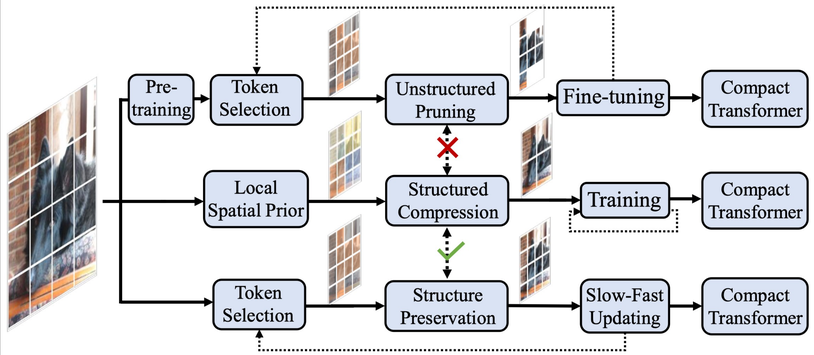
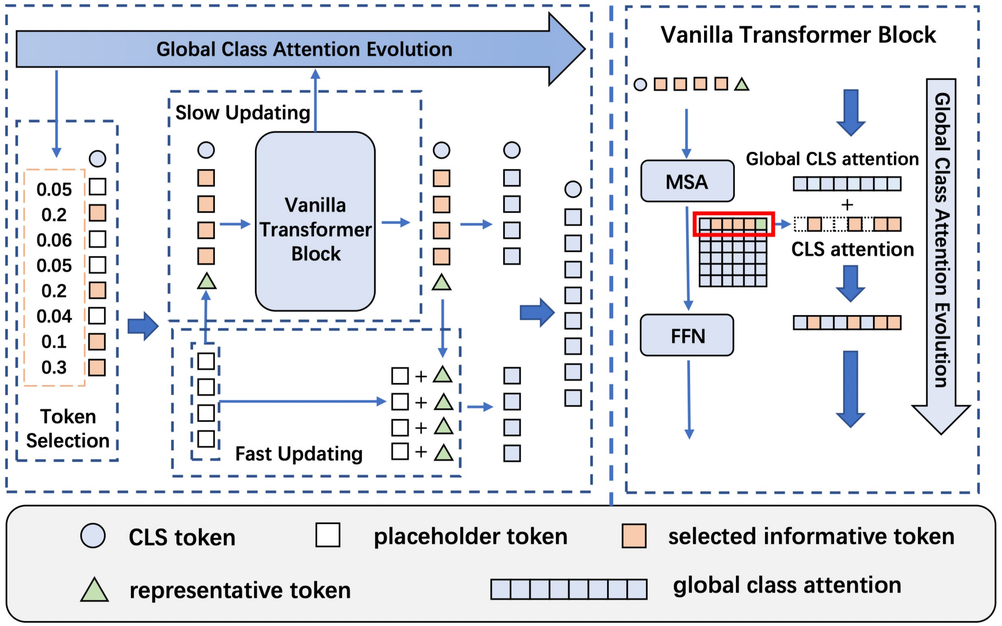


Figure 1. An illustration of technique pipelines for computation reduction via tokens. The third line is our approach.

Framework



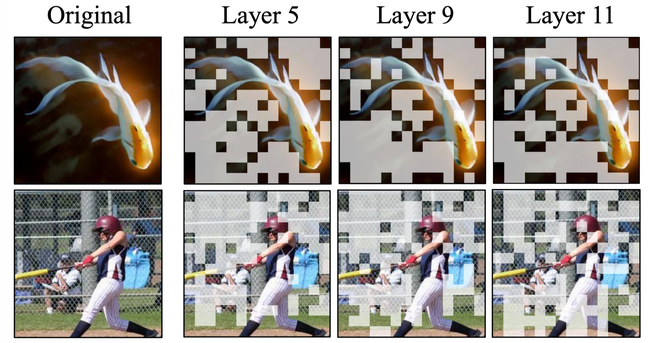
- Slow-fast token updating
- Structure preserving token selection
- Global class attention evolution

Technical Insight

Wide applicability. The Evo-ViT approach maintains the complete spatial structure. Therefore, it is able to accelerate ViTs of both flat and deep-narrow structures from the very beginning of the training process.

Potential in downstream tasks. More significant acceleration performance is found with dense input. Future work includes extending our method to dense prediction tasks, such as detection and segmentation.

Visualization



Acceleration

Method	Top-1 Acc. (%)	Throughput (img/s)	(%)
DeiT-S			
Baseline (Touvron et al. 2021)	79.8	940	-
PS-ViT (Tang et al. 2021)	79.4	1308	43.6
DynamicViT (Rao et al. 2021)	79.3	1479	57.3
SViTE (Chen et al. 2021)	79.2	1215	29.3
IA-RED^2 (Pan et al. 2021)	79.1	1360	44.7
Evo-ViT (ours)	79.4	**1510**	**60.6**

Model	Param (M)	Throughput (img/s)	Top-1 Acc. (%)
LeViT-256	18.9	3357	80.1
LeViT-256*	19.0	906	81.8
PVTv2-B2	25.4	687	82.0
PiT-S	23.5	1266	80.9
Swin-T	29.4	755	81.3
CoaT-Lite Small	20.0	550	81.9
Evo-LeViT-256	19.0	4277	78.8
Evo-LeViT-256*	19.2	1285	81.1