STViT: Vision Transformer with Super Token Sampling

Huaibo Huang[1,2], Xiaoqiang Zhou[1,4], Jie Cao[1,2], Ran He[1,2,3*], Tieniu Tan[1,2,4,5]

[1]MAIS&CRIPAC, Institute of Automation, Chinese Academy of Sciences, China
[2]2School of Artificial Intelligence, University of Chinese Academy of Sciences, China
[3]School of Information Science and Technology, ShanghaiTech University, China
[4]University of Science and Technology of China, [5]Nanjing University

Key Milestones of Transformer

Foundation models in NLP and CV:

Before 2018

local feature: HOG, SIFT
sparse rep.: half-quadratic (our group)
CNN: AlexNet, ResNet, **LightCNN (our group)**

2017.6 | Transformer

Solely based on attention mechanism, the Transformer is proposed and shows great performance on NLP tasks.

2018.10 | BERT

Pre-training transformer models begin to be dominated in the field of NLP.

2020.5 | GPT-3

A huge transformer with 170B parameters, takes a big step towards general NLP model.

2020.05 | DETR

A simple yet effective framework for high-level vision by viewing object detection as a direct set prediction problem

2020.10 | ViT

Pure transformer architectures work well for visual recognition.

End of 2020 | IPT/CLIP

Applications of transformer model on low-level vision, segmentation and multimodality tasks, respectively.

2021- Now | ViT Variants

Variants of ViT models, e.g., DeiT, PVT, Swin, Uniformer, **OrthoT (our group), STViT (our group)**

2022-Now | Stable Diffusion, ChatGPT

Some breakthrough transformer models for practical applications, such as AIGC, Conversational AI.

The vision Transformer models are marked in red.

Introduction

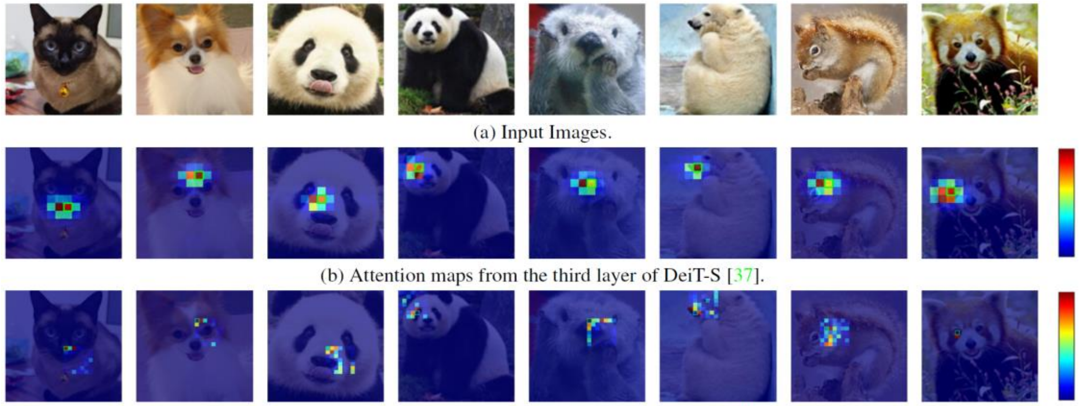
- **Task:** general vision Transformer backbone design.

- **Applications**: image classification, object detection, instance segmentation and semantic segmentation, etc.

- **Research Highlight**:
  - a general vision transformer backbone, STViT,
  - access efficient and effective global context modeling at the early stages of a neural network

Motivation

- **Existing methods:** suffer from high redundancy in capturing local features for shallow layers. Local self-attention or early-stage convolutions sacrifice the capacity to capture long-range dependency



(a) Input Images.

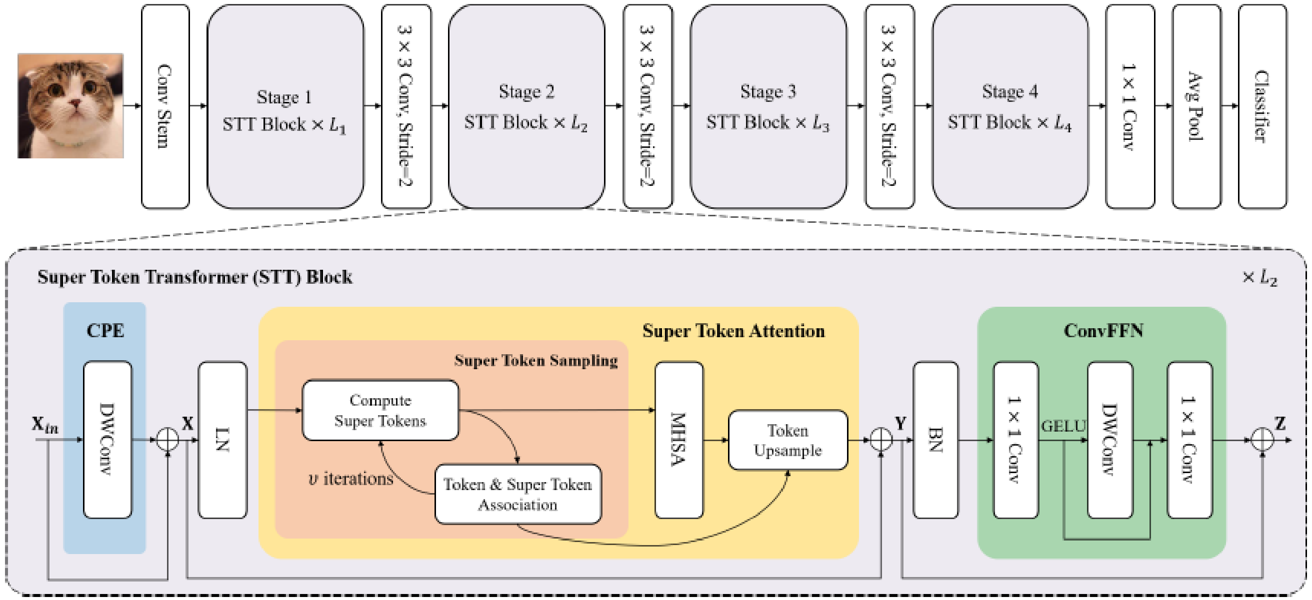
(b) Attention maps from the third layer of DeiT-S [37].

(c) Attention maps from the third layer of Swin-T [30].

- **Challenge:** Can we access efficient and effective global context modeling at the early stages of a neural network?

- **We draw inspiration from the design of superpixels,** and propose a simple yet strong super token attention (STA) mechanism with three steps: **1) Super Token Sampling, 2) Self-Attention for Super Tokens, and 3) Token Upsampling**

Method



The architecture of Super Token Vision Transformer (STViT).

Model	Blocks	Channels	Heads	Params	FLOPs
STViT-S	[3,5,9,3]	[64,128,320,512]	[1,2,5,8]	25M	4.4G
STViT-B	[4,6,14,6]	[96,192,384,512]	[2,3,6,8]	52M	9.9G
STViT-L	[4,7,19,8]	[96,192,448,640]	[2,3,7,10]	95M	15.6G

Architecture variants of STViT

Token & Super Token Association

$$Q^t = \mathrm{Softmax}\left(\frac{X {S^{t-1}}^{\mathrm{T}}}{\sqrt{d}}\right)$$

Super Token Update

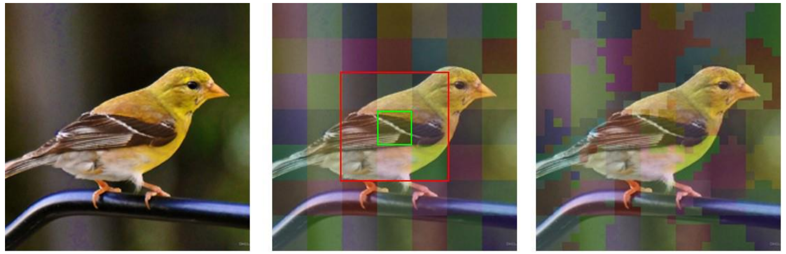
$$S = (\hat{Q}^t)^{\mathrm{T}} X$$

Super Token Attention

$$\mathrm{Attn}(S) = \mathrm{Softmax}\left(\frac{\mathbf{q}(S)\mathbf{k}^{\mathrm{T}}(S)}{\sqrt{d}}\right)\mathbf{v}(S)$$

Token Upsampling

$$\mathrm{TU}(\mathrm{Attn}(S)) = Q\mathrm{Attn}(S)$$

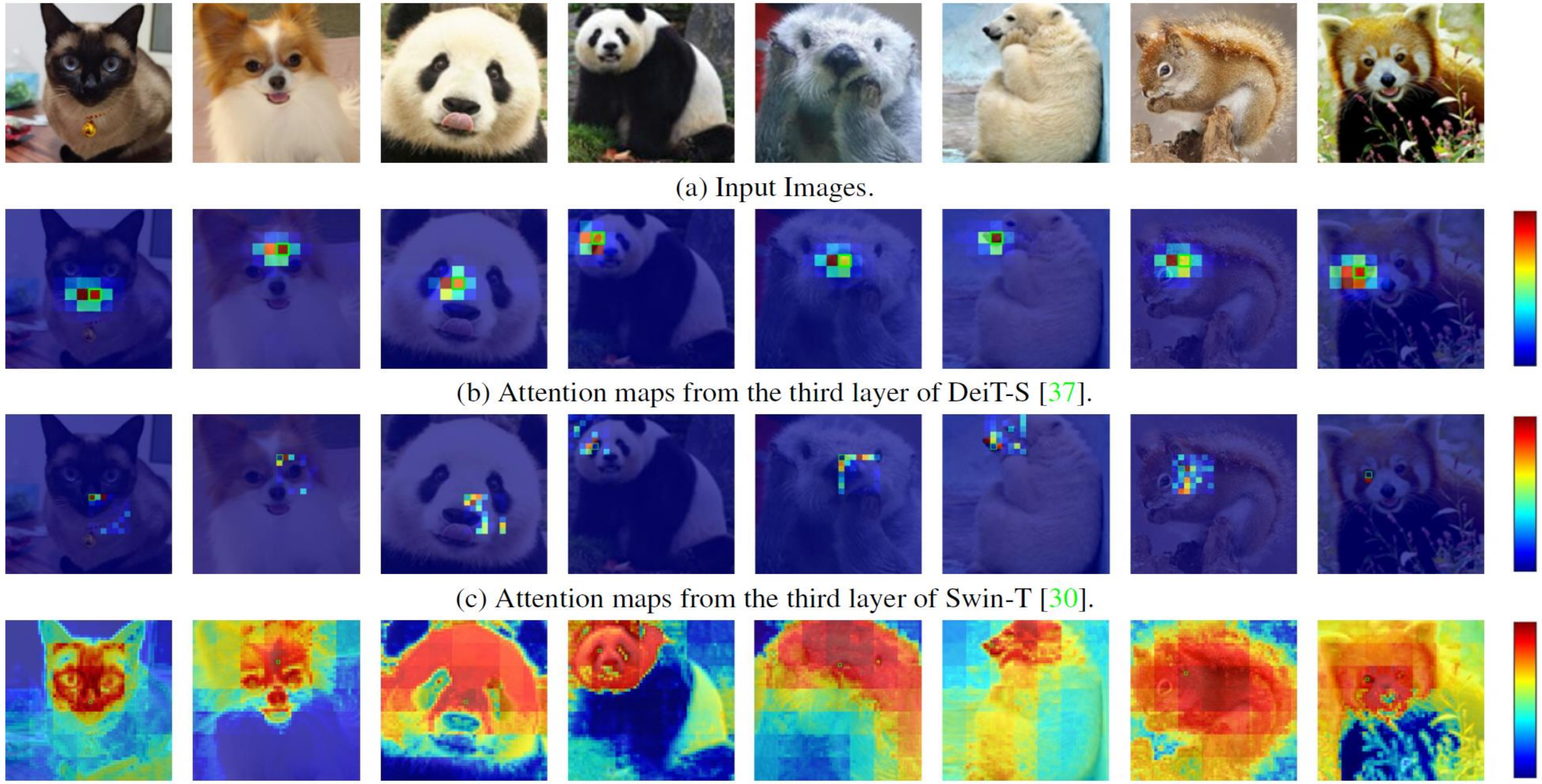


(a) Input Image (b) Initial Super Tokens (c) Learned Super Tokens

Visualization of super tokens from initial grid to learned ones.

Method

access efficient & effective global context modeling at early stages



(a) Input Images.

(b) Attention maps from the third layer of DeiT-S [37].

(c) Attention maps from the third layer of Swin-T [30].

(d) Attention maps from the third layer of STViT-S (Ours).

Visualization of early-stage attention maps for different vision transformers

Experiments

Model Size	Model	#Param	Flops	Throughput	Test Size	Top-1
small model size (∼25M)	ConvNeXt-T [31]	28M	4.5G	720	224	82.1
	DeiT-S [37]	22M	4.6G	922	224	79.9
	PVT-S [41]	25M	3.8G	722	224	79.8
	Swin-T [30]	29M	4.5G	712	224	81.3
	CoAtNet-0 [9]	25M	4.2G	943	224	81.6
	Focal-T [49]	29M	4.9G	323	224	82.2
	DAT-T [45]	29M	4.6G	575	224	82.0
	CSwin-T [12]	23M	4.3G	515	224	82.7
	UniFormer-S [25]	24M	4.2G	824	224	82.9
	MPViT-S [24]	23M	4.7G	352	224	83.0
	Ortho-S [19]	24M	4.5G	435	224	83.4
	CMT-S [15]	25M	4.0G	481	224	83.5
	STViT-S	25M	4.4G	564	224	**83.6**
	CvT-13 [44]	20M	16.3G	-	384	83.0
	CaiT-xs24 [38]	27M	19.3G	86	384	83.8
	CoAtNet-0 [9]	25M	13.4G	262	384	83.9
	CSwin-T [12]	23M	14.0G	204	384	84.3
	STViT-S	25M	14.1G	188	384	**85.0**
medium model size (∼50M)	ConvNeXt-S [31]	50M	8.7G	404	224	83.1
	PVT-L [41]	61M	9.8G	321	224	81.7
	Swin-S [30]	50M	8.7G	406	224	83.0
	CaiT-s24 [38]	47M	9.4G	326	224	82.7
	CoAtNet-1 [9]	42M	8.4G	471	224	83.3
	Focal-S [49]	51M	9.1G	191	224	83.5
	CrossFormer-B [42]	52M	9.2G	377	224	83.4
	CSwin-S [12]	35M	6.9G	315	224	83.6
	DAT-S [45]	50M	9.0G	312	224	83.7
	UniFormer-B [25]	50M	8.3G	375	224	83.9
	Ortho-B [19]	50M	8.6G	286	224	84.0
	MViTv2-B [26]	52M	10.2G	250	224	84.4
	CMT-B [15]	46M	9.3G	259	256	84.5
	STViT-B	52M	9.9G	286	224	**84.8**
	CaiT-s24 [38]	47M	32.2G	60	384	84.3
	CSwin-S [12]	35M	22.0G	128	384	85.0
	CoAtNet-1 [9]	42M	27.4G	124	384	85.1
	MViTv2-B [26]	52M	36.7G	-	384	85.6
	STViT-B	52M	31.5G	95	384	**86.0**
large model size (∼100M)	ConvNeXt-B [31]	89M	15.4G	252	224	83.8
	DeiT-B [37]	86M	17.5G	298	224	81.8
	Swin-B [30]	88M	15.4G	258	224	83.3
	CaiT-s48 [38]	90M	18.6G	162	224	83.5
	Focal-B [49]	90M	16.0G	136	224	83.8
	CrossFormer-L [42]	92M	16.1G	246	224	84.0
	DAT-B [45]	88M	15.8G	211	224	84.0
	CoAtNet-2 [9]	75M	15.7G	298	224	84.1
	CSwin-B [12]	78M	15.0G	216	224	84.2
	Ortho-L [19]	88M	15.4G	180	224	84.2
	MPViT-B [24]	75M	16.4G	185	224	84.3
	CMT-L [15]	75M	19.5G	-	288	84.8
	STViT-L	95M	15.6G	192	224	**85.3**
	Swin-B [30]	88M	47.0G	85	384	84.2
	CaiT-s48 [38]	90M	63.8G	30	384	85.1
	CSwin-B [12]	78M	47.0G	69	384	85.5
	CoAtNet-2 [9]	75M	49.8G	88	384	85.7
	STViT-L	95M	49.7G	65	384	**86.4**

Performance comparison on ImageNet-1K classification

Model	ImageNet-1K			COCO		ADE20K
	#Param	FLOPs	Acc.	AP^b	AP^m	mIoU
DeiT-S	22M	4.6G	79.9%	-	-	-
Swin-T	28M	4.5G	81.3%	42.2	39.1	44.5
STA-4stage	24.1M	3.97G	82.3%	43.6	40.1	46.0
+ Conv Stem	24.2M	4.26G	82.6%	44.6	41.1	46.9
+ Projection	25.2M	4.29G	82.9%	44.6	41.2	47.0
+ CPE	25.3M	4.30G	83.3%	46.8	42.5	47.7
w/o CPE, + APE	24.2M	4.26G	83.1%	45.2	41.5	47.3
w/o CPE, + RPE	25.3M	4.29G	83.2%	45.6	41.8	47.5
+ ConvFFN	25.4M	4.37G	83.6%	47.6	43.1	48.6
w/o shortcut	25.4M	4.37G	83.4%	47.5	43.0	48.4

Ablations of STViT

Backbone	#Param (M)	FLOPs (G)	mIoU (%)	MS mIoU (%)
Swin-T [30]	60	945	44.5	45.8
CSWin-T [12]	60	959	**49.3**	**50.7**
UniFormer-S [25]	52	1008	47.6	48.5
STViT-S	54	926	48.6	49.0
Res101 [17]	86	1029	-	44.9
Twins-B [7]	89	1020	47.7	48.9
Swin-S [30]	81	1038	47.6	49.5
Focal-T [49]	85	1130	48.0	50.0
CrossFormer-B [42]	84	1079	49.2	50.1
UniFormer-B [25]	80	1106	49.5	50.7
CSWin-S [12]	65	1027	50.4	51.5
STViT-B	80	1036	**50.7**	**51.9**
Swin-B [30] [30]	121	1188	48.1	49.7
Focal-B [49] [49]	126	1354	49.0	50.5
CSWin-B [12]	109	1222	51.1	52.2
STViT-L	125	1151	**52.4**	**53.2**

Semantic segmentation with Upernet on ADE20K

Experiments

Backbone	#Param (M)	FLOPs (G)	Mask R-CNN 1× schedule						Mask R-CNN 3× + MS schedule					
			AP^b	AP^b_{50}	AP^b_{75}	AP^m	AP^m_{50}	AP^m_{75}	AP^b	AP^b_{50}	AP^b_{75}	AP^m	AP^m_{50}	AP^m_{75}
Res50 [17]	44	260	38.0	58.6	41.4	34.4	55.1	36.7	41.0	61.7	44.9	37.1	58.4	40.1
PVT-S [41]	44	245	40.4	62.9	43.8	37.8	60.1	40.3	43.0	65.3	46.9	39.9	62.5	42.8
Swin-T [30]	48	264	42.2	64.6	46.2	39.1	61.6	42.0	46.0	68.2	50.2	41.6	65.1	44.8
Focal-T [49]	49	291	44.8	67.7	49.2	41.0	64.7	44.2	47.2	69.4	51.9	42.7	66.5	45.9
CMT-S [15]	45	249	44.6	66.8	48.9	40.7	63.9	43.4	48.3	70.4	52.3	43.7	67.7	47.1
UniFormer-S [25]	41	269	45.6	68.1	49.7	41.6	64.8	45.0	48.2	70.4	52.5	43.4	67.1	47.0
STViT-S	44	252	**47.6**	**70.0**	**52.3**	**43.1**	**66.8**	**46.5**	**49.2**	**70.8**	**54.4**	**44.2**	**68.0**	**47.7**
Res101 [17]	63	336	40.4	61.1	44.2	36.4	57.7	38.8	42.8	63.2	47.1	38.5	60.1	41.3
PVT-M [41]	64	302	42.0	64.4	45.6	39.0	61.6	42.1	44.2	66.0	48.2	40.5	63.1	43.5
Swin-S [30]	69	354	44.8	66.6	48.9	40.9	63.4	44.2	48.5	70.2	53.5	43.3	67.3	46.6
Focal-S [49]	71	401	47.4	69.8	51.9	42.8	66.6	46.1	48.8	70.5	53.6	43.8	67.7	47.2
DAT-S [45]	69	378	47.1	69.9	51.5	42.5	66.7	45.4	49.0	70.9	53.8	44.0	68.0	47.5
UniFormer-B [25]	69	399	47.4	69.7	52.1	43.1	66.0	46.5	50.3	**72.7**	55.3	44.8	69.0	48.3
STViT-B	70	359	**49.7**	**71.7**	**54.7**	**44.8**	**68.9**	**48.7**	**51.0**	72.3	**56.0**	**45.4**	**69.5**	**49.3**
Swin-B	107	496	46.9	-	-	42.3	-	-	48.5	69.8	53.2	43.4	66.8	46.9
CSWin-B	97	526	48.7	70.4	53.9	43.9	67.8	47.3	50.8	72.1	55.8	44.9	69.1	48.3
STViT-L	114	470	**50.8**	**72.5**	**56.3**	**45.5**	**69.7**	**49.1**	**51.7**	**73.0**	**56.9**	**45.9**	**70.4**	**49.9**

Object detection and instance segmentation with Mask R-CNN on COCO val2017

STViT: Vision Transformer with Super Token Sampling

Huaibo Huang, Xiaoqiang Zhou, Jie Cao, Ran He, Tieniu Tan

Thanks

Code Link

Group's Homepage