SparseFormer: Sparse Visual Recognition via Limited Latent

Ziteng Gao[1] Zhan Tong[2] Limin Wang[3] Mike Zheng Shou[1]

[1] Show Lab, National University of Singapore

[2] Ant Group

[3] Nanjing University

Motivation

Common-used vision networks involve dense units,
pixels in ConvNets or patches in vision transformers,

But here comes some issues, including

1) redundant compute for uninformative backgrounds;
2) soaring compute and memory footprint w.r.t. scaling resolutions, especially in vision transformers;

Can we avoid dense units in vision modeling?

Sparse Latent Tokens

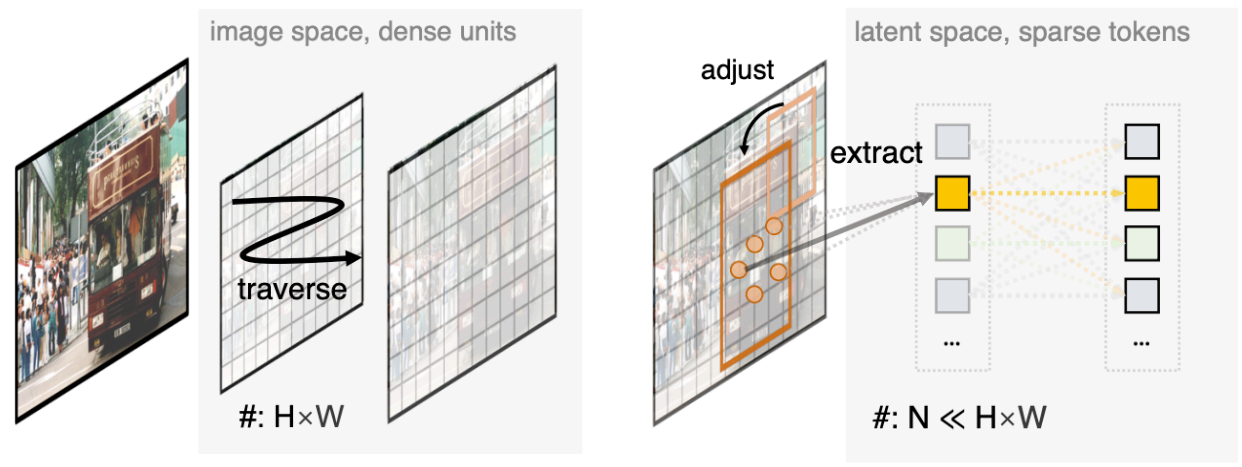
Inspired by Perceivers and detection transformers,

1) We move basic units in vision transformers into the latent space
2) We exploit limited latent tokens to perform vision transformers.

The number of latent tokens is highly limited

We can even use 9 latent tokens to recognize an img!

i.e., 74.5 top-1 acc on ImageNet-1K

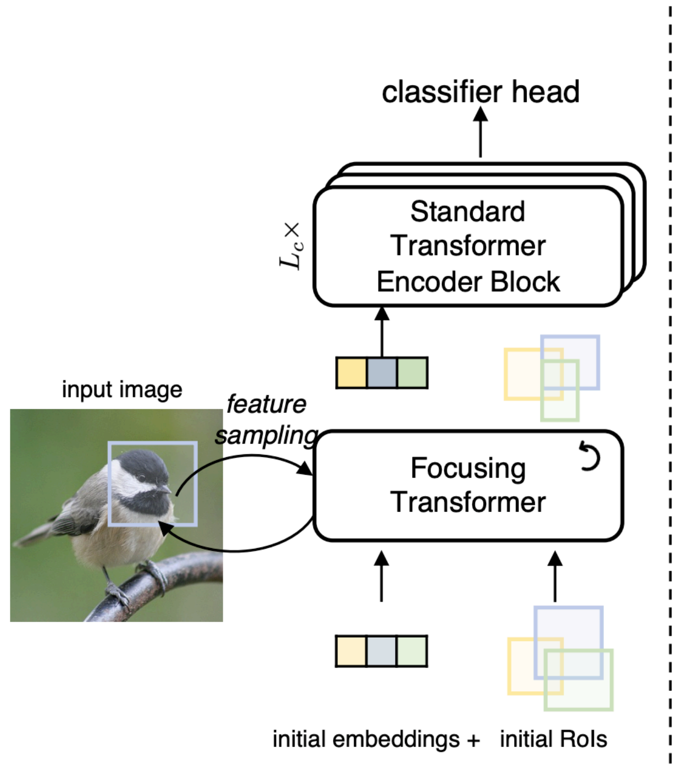


dense paradigm our **sparse** paradigm

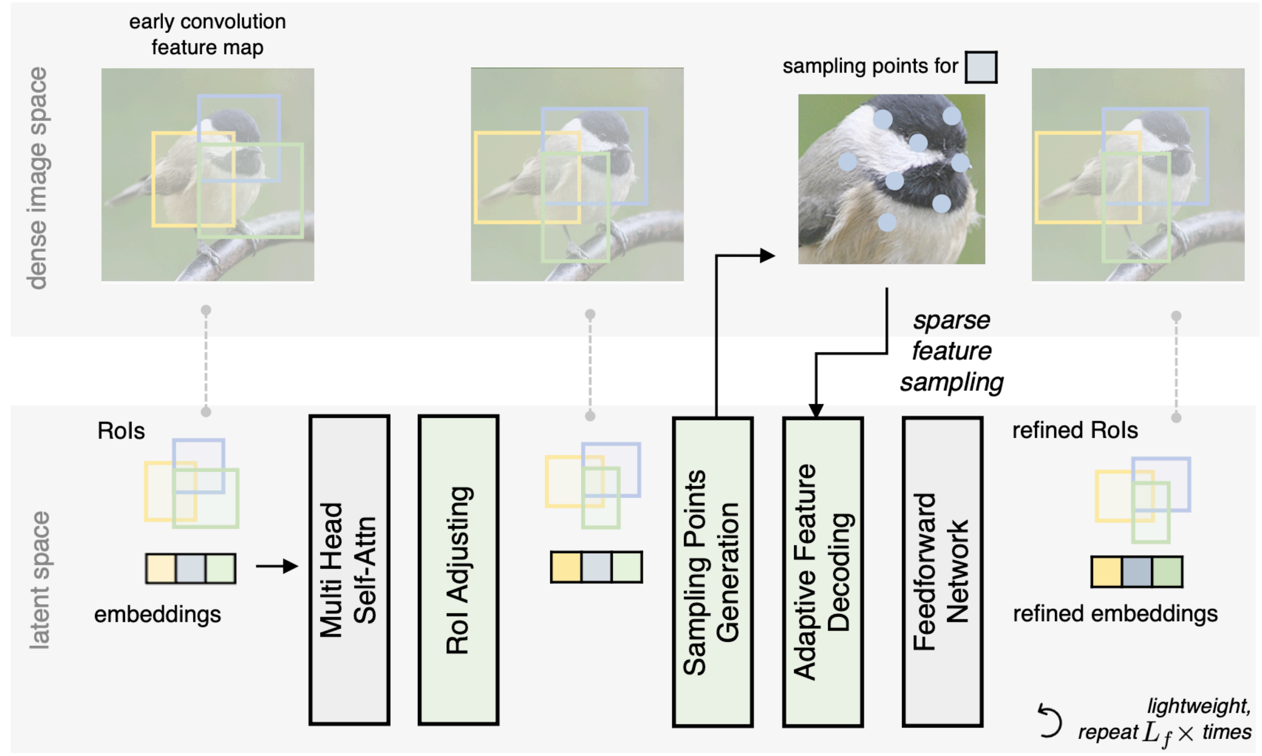
SparseFormer

SparseFormer = latent tokens + focusing transformer + cortex transformer

1) Latent tokens = Latent embeddings + RoIs $\mathbf{T} = \{(\mathbf{t}_1, \mathbf{b}_1), (\mathbf{t}_2, \mathbf{b}_2), \cdots, (\mathbf{t}_N, \mathbf{b}_N)\},$
2) Focusing transformer = MHSA + RoI Adjusting + Feature Sampling + FFN
3) Cortex transformer = standard transformer = MHSA + FFN



(a) overall architecture



(b) details of a focusing transformer block

method	top-1	FLOPs	#params	throughput (img/s)
ResNet-50 (Wightman et al., 2021)	80.4	4.1G	26M	1179
ResNet-101 (Wightman et al., 2021)	81.5	7.9G	45M	691
DeiT-S (Touvron et al., 2021)	79.8	4.6G	22M	983
DeiT-B (Touvron et al., 2021)	81.8	17.5G	86M	306
Swin-T (Liu et al., 2021a)	81.3	4.5G	29M	726
Swin-S (Liu et al., 2021a)	83.0	8.7G	50M	437
Perceiver (Jaegle et al., 2021)	78.0	707G	45M	17
Perceiver IO (Jaegle et al., 2022)	82.1	369G	49M	30
SparseFormer-T	81.0	2.0G	32M	1270
SparseFormer-S	82.0	3.8G	48M	898
SparseFormer-B	82.6	7.8G	81M	520

(I) SparseFormer on ImageNet-1K

method	top-1	pre-train	#frames	GFLOPs	#params
NL I3D (Wang et al., 2018)	77.3	ImageNet-1K	128	359×10×3	62M
SlowFast (Feichtenhofer et al., 2019)	77.9	-	8+32	106×10×3	54M
TimeSFormer (Bertasius et al., 2021)	75.8	ImageNet-1K	8	196×1×3	121M
Video Swin-T (Liu et al., 2021b)	78.8	ImageNet-1K	32	88×4×3	28M
ViViT-B FE (Arnab et al., 2021)	78.8	ImageNet-21K	32	284×4×3	115M
MViT-B (Fan et al., 2021)	78.4	-	16	71×5×1	37M
VideoSparseFormer-T	77.9	ImageNet-1K	32	22×4×3	31M
VideoSparseFormer-S	79.1	ImageNet-1K	32	38×4×3	48M
VideoSparseFormer-B	79.8	ImageNet-21K	32	74×4×3	81M

(II) VideoSparseFormer on Kinetics-400

Experiments

variant	pre-training data	resolution	top-1	FLOPs	throughput (img/s)
B	IN-1K	224^2	82.6	7.8G	520
B	IN-21K	224^2	83.6	7.8G	520
B	IN-21K	384^2	84.1	8.2G	444
B	IN-21K	512^2	84.0	8.6G	419
B, $N = 144 \uparrow$	IN-21K	384^2	84.6	14.2G	292
B, $N = 196 \uparrow$	IN-21K	384^2	84.8	19.4G	221

(III) Scaling Up SparseFormers

N	9	16	25	36	49	64	81
top-1	74.5	77.4	79.3	80.1	81.0	81.4	81.9
GFLOPs	0.5	0.8	1.1	1.6	2.0	2.7	3.3

(a)

method	SF	ViT/32	ViT/32*	conv×4	swin
top-1	81.0	72.8	74.3	79.4	79.7
GFLOPs	2.0	1.4	1.7	2.2	2.0

(b)

L_f	top-1	GFLOPs
nil	77.8	1.6
1	79.7	1.7
4	81.0	2.0
8	81.0	2.5

(c)

P	top-1	GFLOPs
16	80.3	1.9
36	81.0	2.0
64	81.3	2.3

(d)

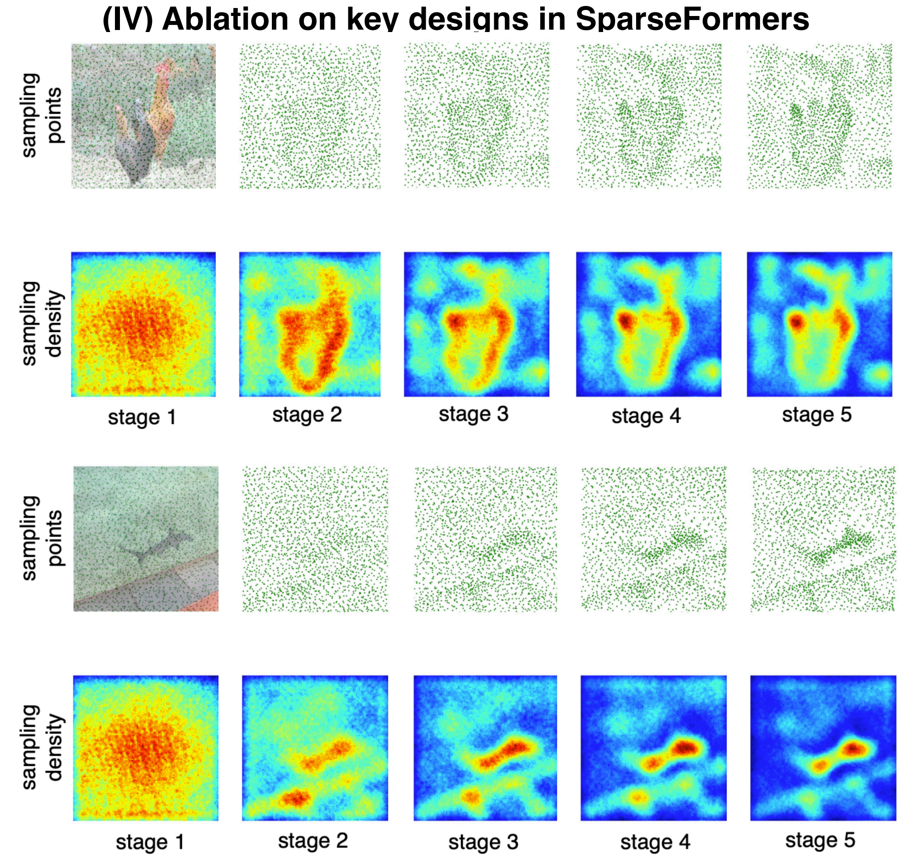
img feat.	top-1	GFLOPs
RGB	fail	1.5
ViT/8-embed	78.4	1.9
early conv	81.0	2.0
ResNet C1+C2	82.2	3.1

(e)

decode	top-1	GFLOPs
linear	78.5	1.9
static, mix	80.1	1.9
adaptive, mix	81.0	2.0

(f)

(IV) Ablation on key designs in SparseFormers



Experiments

<- SparseFormer OpenReview link

<- Code and checkpoints are available at https://github.com/showlab/sparseformer

Ziteng's twitter: @ziteng_v

Kindly also check our follow-up work!!

"Bootstrapping SparseFormers from Vision Foundation Models" in CVPR 2024