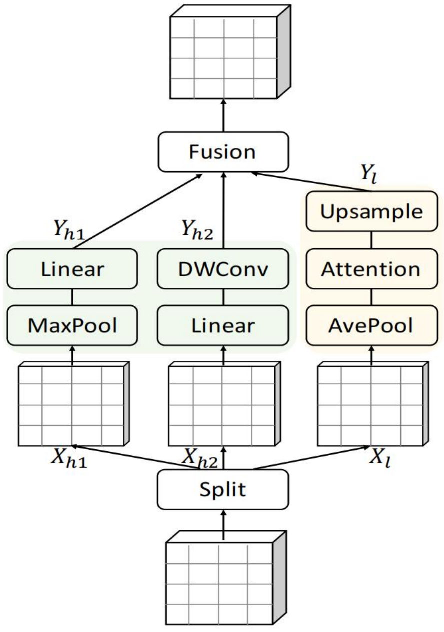
Lightweight Vision Transformer with Bidirectional Interaction

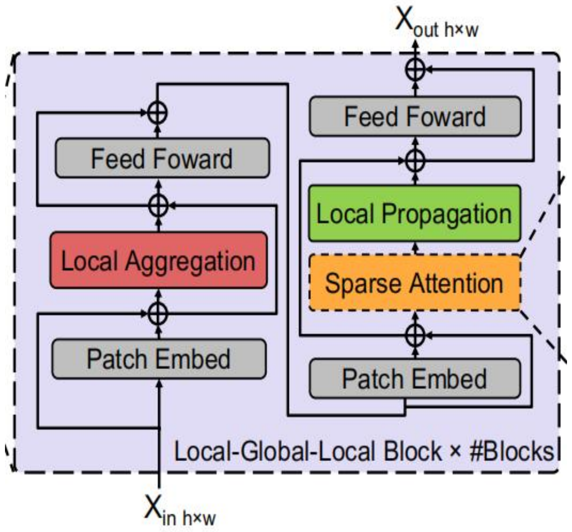
Qihang Fan[1,2], Huaibo Huang[1], Xiaoqiang Zhou[1,3], Ran He[1,2]

[1]MAIS & CRIPAC, Institute of Automation, Chinese Academy of Sciences, Beijing, China
[2]School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China
[3]University of Science and Technology of China, Hefei, China

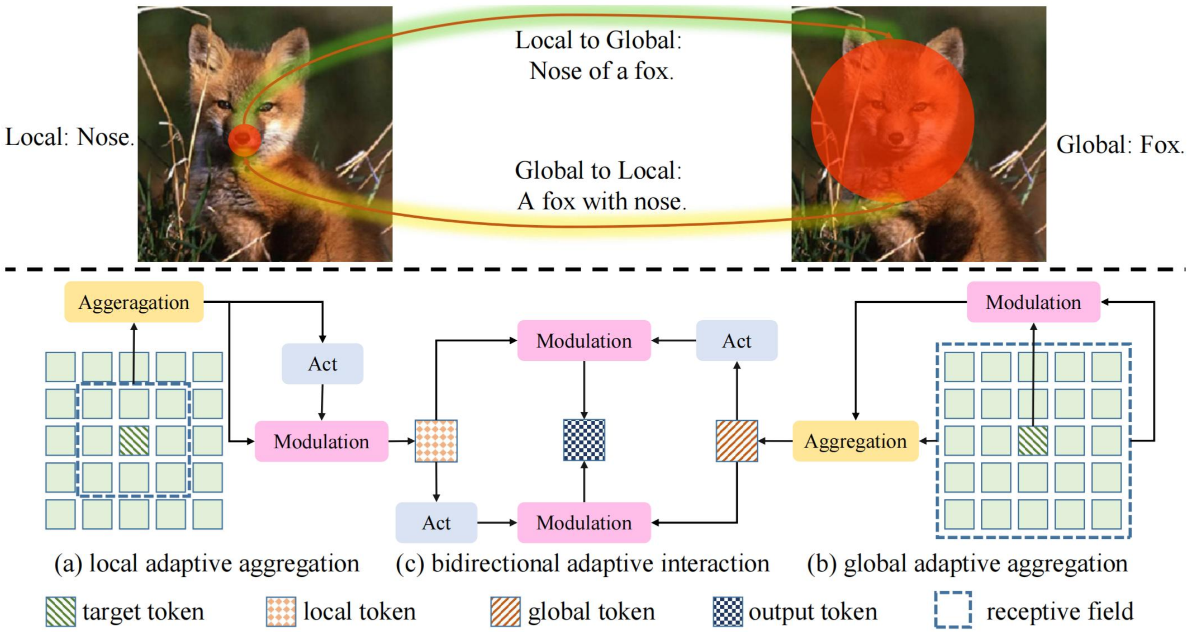
Introduction



Inception Transformer



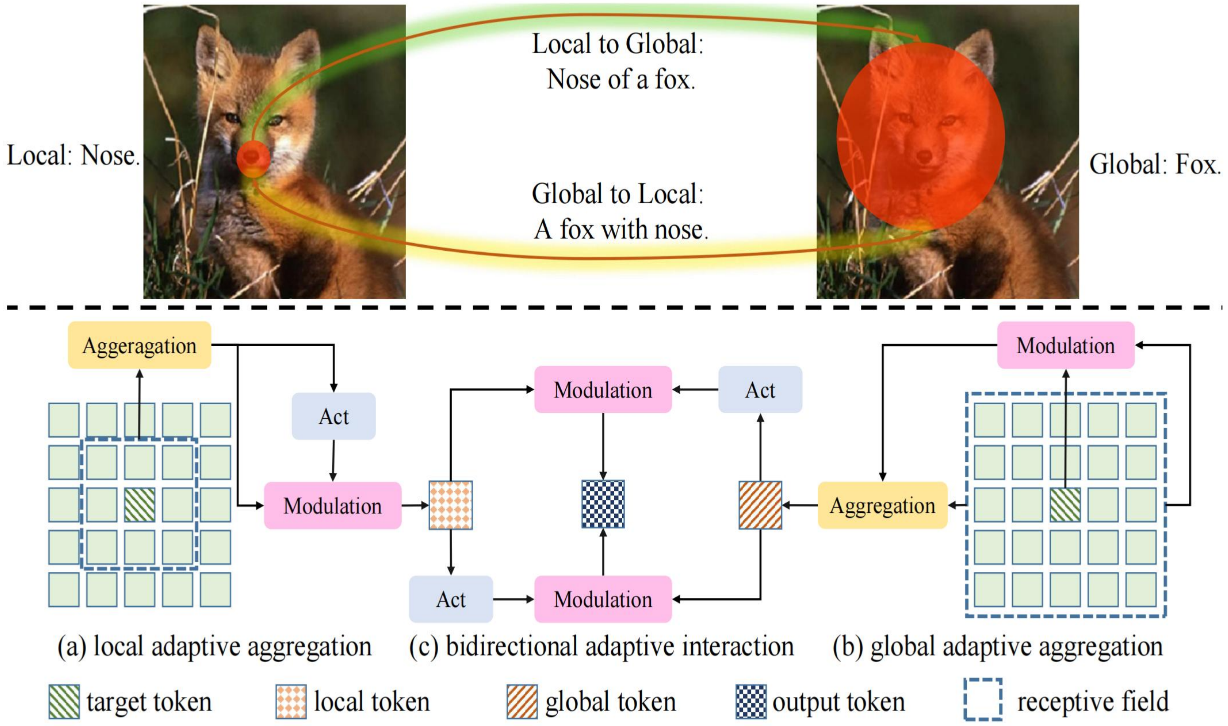
EdgeViT



Parallel/Sequential Local-Global

Local-Global Bidirectional Interaction

Fully Adaptive Self–Attention(FASA)



Three operators in FASA

1. Global Adaptive Aggregation
2. Local Adaptive Aggregation
3. Bidirectional Adaptive Interaction

All based on Context–Aware Feature Aggregation

Global Adaptive Aggregation：

$$y_i = \mathcal{A}(\mathcal{F}(\mathcal{M}(x_i, X)), X),$$

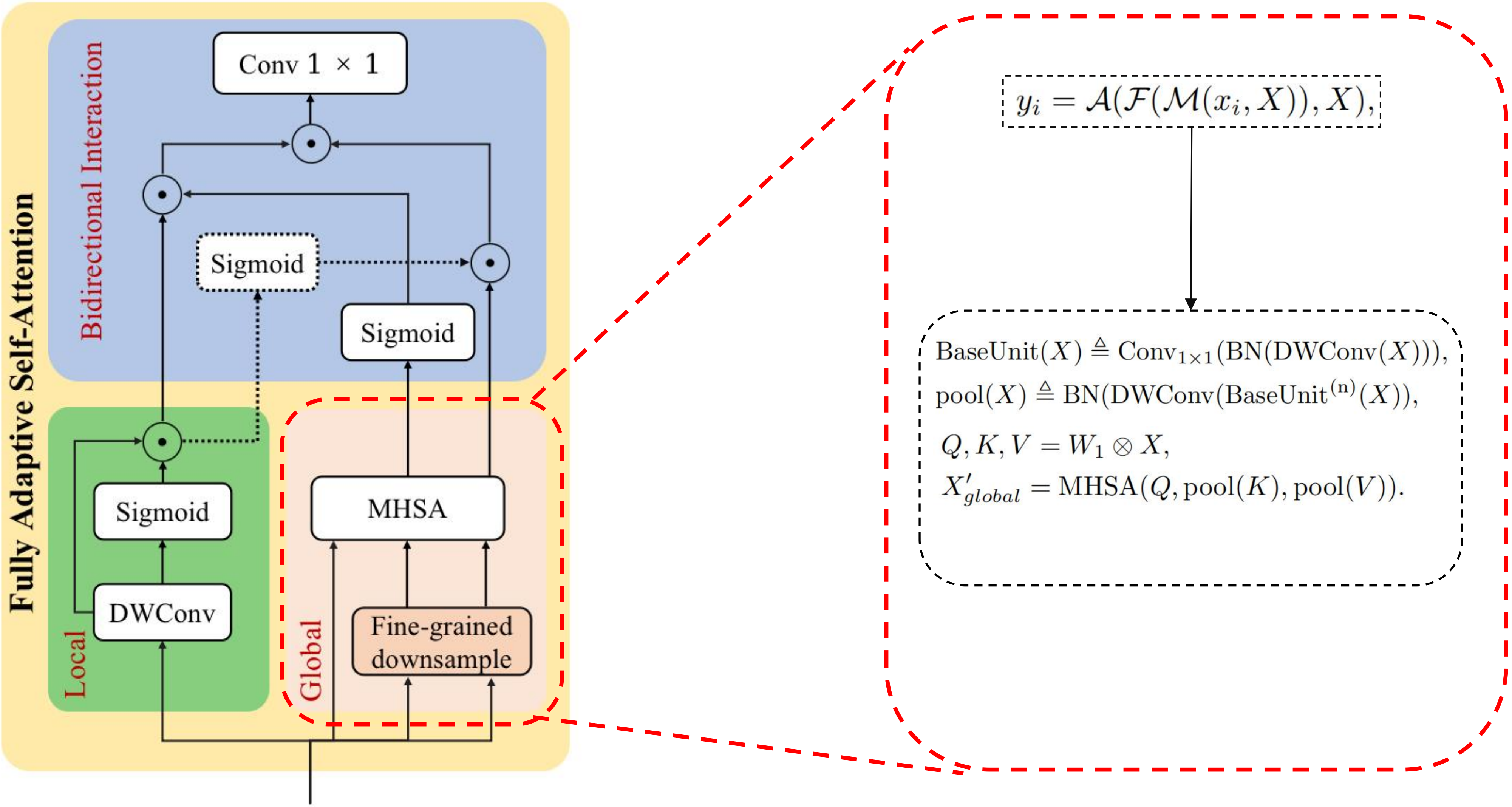
Local Adaptive Aggregation：

$$y_i = \mathcal{M}(\mathcal{F}(\mathcal{A}(x_i, X)), \mathcal{A}(x_i, X)),$$

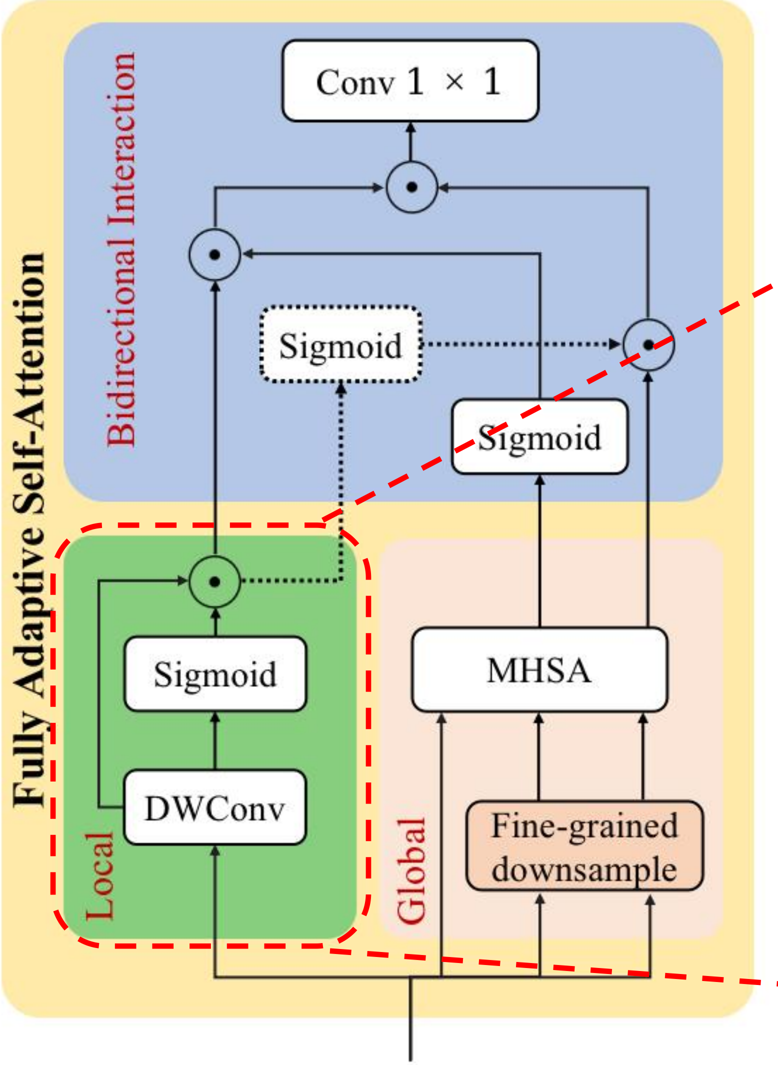
Bidirectional Adaptive Interaction：

$$y_{i2} = \mathcal{M}(\mathcal{F}(\mathcal{A}_1(x_i, X)), \mathcal{A}_2(x_i, X)),$$
$$y_{i1} = \mathcal{M}(\mathcal{F}(\mathcal{A}_2(x_i, X)), \mathcal{A}_1(x_i, X)).$$

Global Adaptive Aggregation



Local Adaptive Aggregation

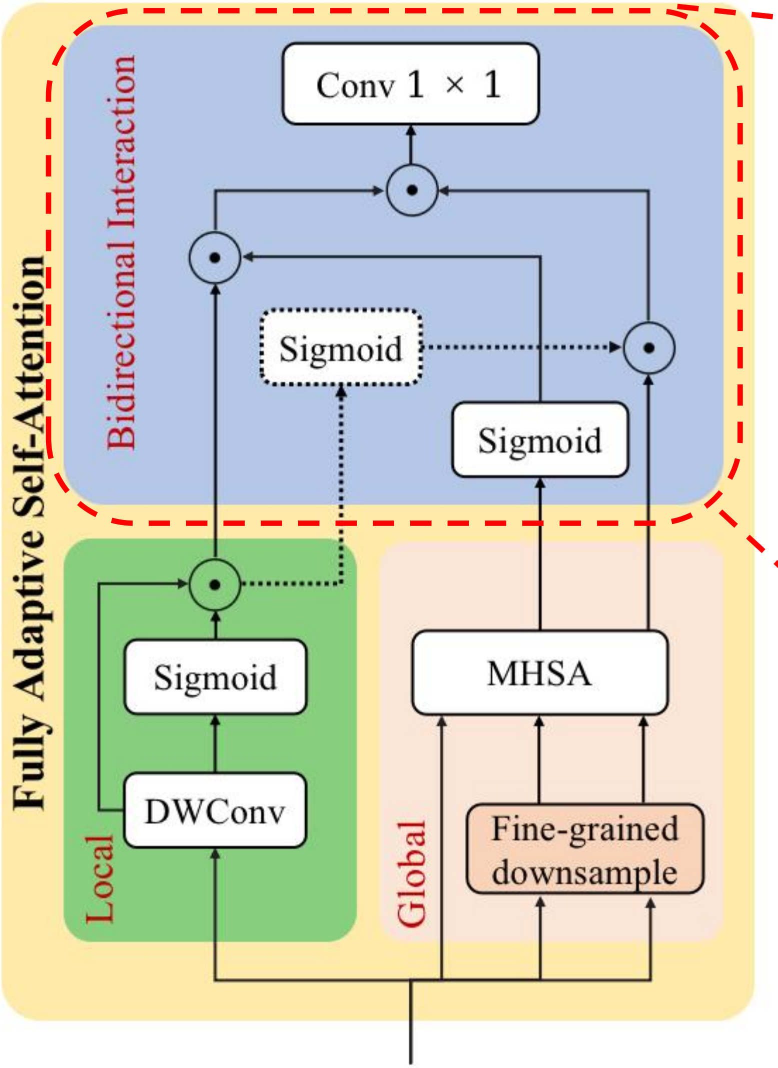


$$y_i = \mathcal{M}(\mathcal{F}(\mathcal{A}(x_i, X)), \mathcal{A}(x_i, X)),$$

$$Q' = \mathrm{DWConv}(Q),$$

$$X'_{local} = Q' \odot \mathrm{Sigmoid}(Q') = Q' \odot \frac{1}{1 + e^{-Q'}}.$$

Bidirectional Adaptive Interaction



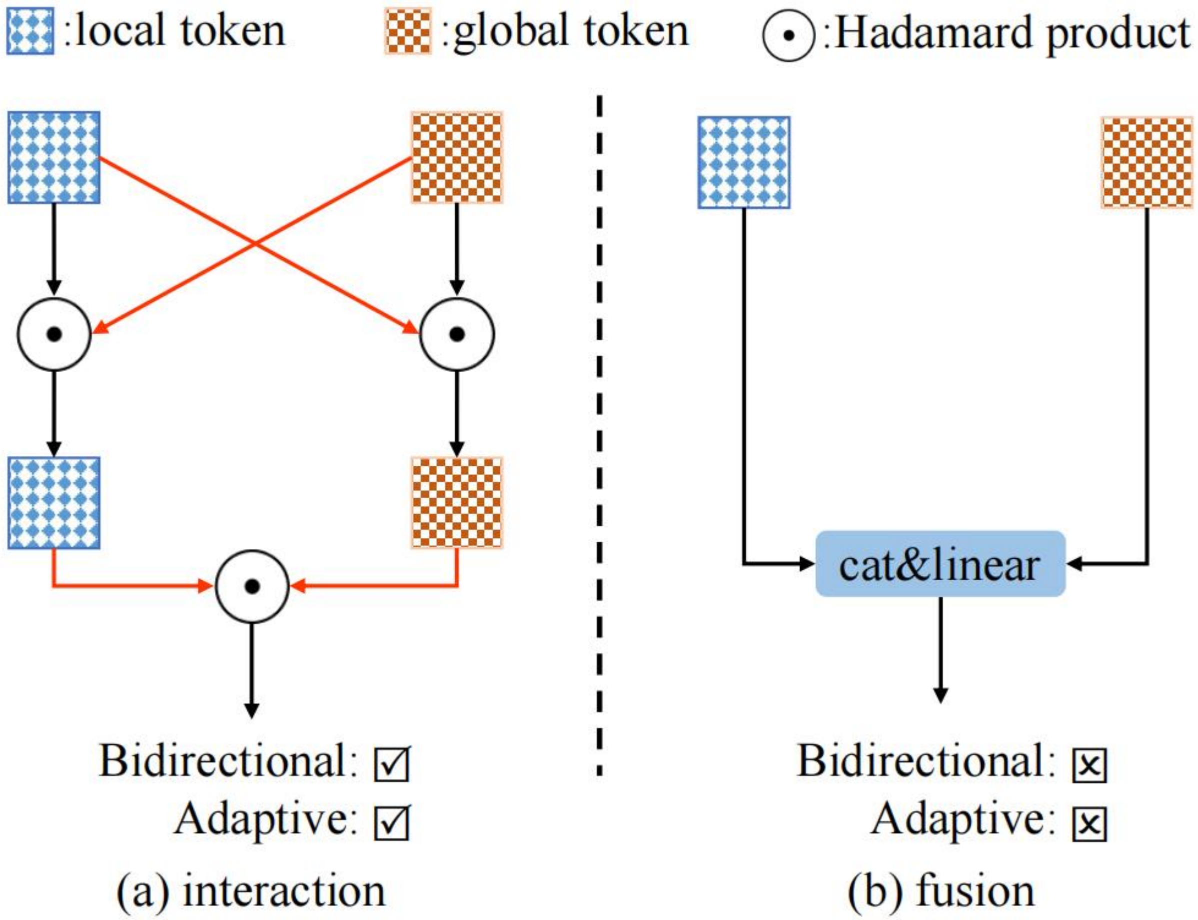
$$y_{i2} = \mathcal{M}(\mathcal{F}(\mathcal{A}_1(x_i, X)), \mathcal{A}_2(x_i, X)),$$
$$y_{i1} = \mathcal{M}(\mathcal{F}(\mathcal{A}_2(x_i, X)), \mathcal{A}_1(x_i, X)).$$

$$X_{local} = X'_{local} \odot \text{Sigmoid}(X'_{global}) = X'_{local} \odot \frac{1}{1+e^{-X'_{global}}}$$
$$= Q' \odot \frac{1}{1+e^{-Q'}} \odot \frac{1}{1+e^{-X'_{global}}},$$

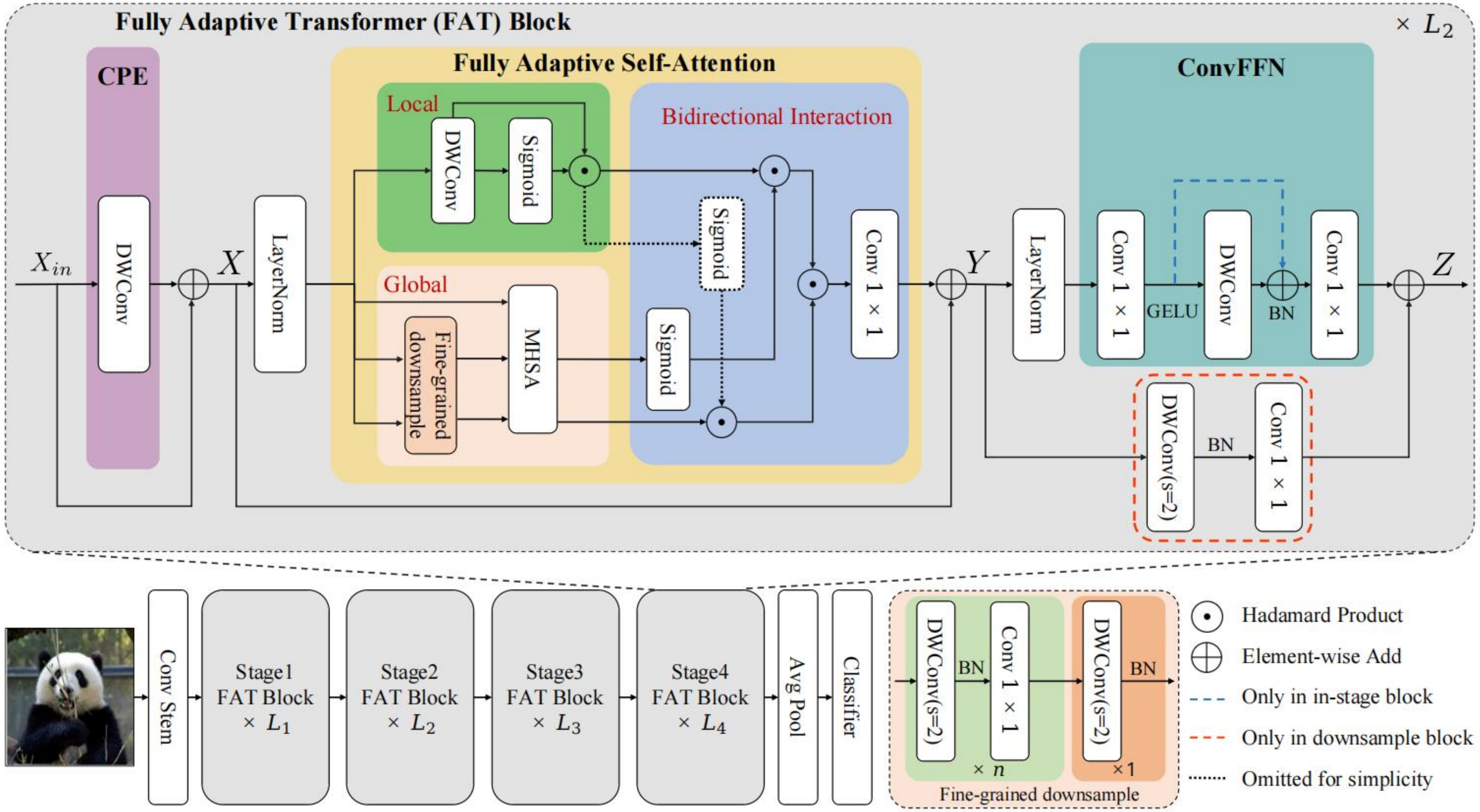
$$X_{global} = X'_{global} \odot \text{Sigmoid}(X'_{local}) = X'_{global} \odot \frac{1}{1+e^{-X'_{local}}}$$
$$= X'_{global} \odot \frac{1}{1+e^{-Q' \odot \frac{1}{1+e^{-Q'}}}},$$

$$Y = W_2 \otimes (X_{global} \odot X_{local})$$

“Interaction” v.s. “Fusion”



Overall Architecture



1. Hierarchical Stages
2. Conditional Positional Encoding (CPE)
3. Fully Adaptive Self-Attention (FASA)
4. Convolutional Feed-Forward Network (ConvFFN)

For one block：

$$X = \mathrm{CPE}(X_{in}) + X_{in},$$
$$Y = \mathrm{FASA}(\mathrm{LN}(X)) + X,$$
$$Z = \mathrm{ConvFFN}(\mathrm{LN}(Y)) + \mathrm{ShortCut}(Y).$$

Experiments

Size (M)	Model	Input	Params (M)	FLOPs (G)	Throughput (img/s)	Top-1 (%)
0 ~ 5	T2T-ViT-7 [73]	224^2	4.3	1.1	1762	71.7
	QuadTree-B-b0 [52]	224^2	3.5	0.7	885	72.0
	TNT-Tiny [18]	224^2	6.1	1.4	545	73.9
	Ortho-T [25]	224^2	3.9	0.7	—	74.0
	EdgeViT-XXS [40]	224^2	4.1	0.6	1926	74.4
	MobileViT-XS [38]	**256^2**	2.3	1.1	1367	74.8
	LVT [66]	224^2	5.5	0.9	1265	74.8
	CPVT-Ti-GAP [6]	224^2	5.8	1.3	—	74.9
	EdgeNeXt-XS [37]	**256^2**	2.3	0.5	1417	75.0
	PVT-T [57]	224^2	13.2	1.6	1233	75.1
	ViT-C [63]	224^2	4.6	1.1	—	75.3
	VAN-B0 [17]	224^2	4.1	0.9	1662	75.4
	ViL-Tiny [77]	224^2	6.7	1.4	857	76.3
	CeiT-T [72]	224^2	6.4	1.2	—	76.4
	FAT-B0	224^2	4.5	0.7	1932	77.6
5 ~ 10	T2T-ViT-12 [73]	224^2	6.9	1.9	1307	76.5
	Rest-lite [78]	224^2	10.5	1.4	1123	77.2
	XCiT-T12 [14]	224^2	6.7	1.2	1676	77.1
	EdgeViT-XS [40]	224^2	6.7	1.1	1528	77.5
	CoaT-Lite-T [64]	224^2	5.7	1.6	1045	77.5
	SiT-Ti w/o FRD [83]	224^2	15.9	1.0	1057	77.7
	RegNetY-1.6GF [44]	224^2	11.2	1.6	1241	78.0
	MPViT-T [28]	224^2	5.8	1.6	737	78.2
	MobileViT-S [38]	**256^2**	5.6	2.0	898	78.4
	ParC-Net-S [76]	**256^2**	5.0	1.7	1321	78.6
	PVTv2-B1 [58]	224^2	13.1	2.1	1007	78.7
	PerViT-T [39]	224^2	7.6	1.6	1402	78.8
	CoaT-Lite-Mi [64]	224^2	11.0	2.0	963	79.1
	EfficientNet-B1 [51]	**240^2**	7.8	0.7	1395	79.2
	FAN-T-ViT [81]	224^2	7.3	1.3	1181	79.2
	EdgeNext-S [37]	224^2	5.6	1.3	1243	79.4
	FAT-B1	224^2	7.8	1.2	1452	80.1

Size (M)	Model	Input	Params (M)	FLOPs (G)	Throughput (img/s)	Top-1 (%)
10 ~ 15	XCiT-T24 [14]	224^2	12.1	2.3	1194	79.4
	ResT-S [78]	224^2	13.7	1.9	918	79.6
	Shunted-T [47]	224^2	11.5	2.1	957	79.8
	DeiT-S [54]	224^2	22.1	4.6	899	79.9
	QuadTree-B-b1 [52]	224^2	13.6	2.3	543	80.0
	RegionViT-Ti [2]	224^2	13.8	2.4	710	80.4
	Wave-MLP-T [53]	224^2	17.0	2.4	1052	80.6
	MPViT-XS [28]	224^2	10.5	2.9	597	80.9
	EdgeViT-S [40]	224^2	11.1	1.9	1049	81.0
	VAN-B1 [17]	224^2	13.9	2.5	995	81.1
	Swin-T [34]	224^2	29.0	4.5	664	81.3
	CrossFormer-T [59]	224^2	27.8	2.9	929	81.5
	ResT-B [78]	224^2	30.3	4.3	588	81.6
	EfficientNet-B3 [51]	**300^2**	12.0	1.8	634	81.6
	FAT-B2	224^2	13.5	2.0	1064	81.9
25 ~ 30	DAT-T [61]	224^2	29	4.6	577	82.0
	FocalNet-T [67]	224^2	29	4.4	610	82.1
	Focal-T [68]	224^2	29	4.9	301	82.2
	CrossFormer-S [59]	224^2	31	4.9	601	82.5
	RegionViT-S [2]	224^2	31	5.3	460	82.6
	DaViT-T [10]	224^2	28	4.5	616	82.8
	WaveViT-S [70]	224^2	20	4.3	482	82.7
	QuadTree-B-b2 [52]	224^2	24	4.5	299	82.7
	CSWin-T [12]	224^2	23	4.3	591	82.7
	MPViT-S [28]	224^2	23	4.7	410	83.0
	HorNet-T [46]	224^2	23	4.0	586	83.0
	FAT-B3-ST	224^2	29	4.7	641	83.0
	ConvNeXt-S [35]	224^2	50	8.7	405	83.1
	LITv2-M [41]	224^2	49	7.5	436	83.3
	EfficientFormer-L7 [30]	224^2	82	10.2	368	83.3
	iFormer-S [48]	224^2	20	4.8	471	83.4
	FAT-B3	224^2	29	4.4	474	83.6

Comparison with the state-of-the-art on ImageNet-1K classification

Experiments

Model	Params(M)	FLOPs(G) ↓	CPU(ms) ↓	GPU(ms) ↓	Trp(imgs/s) ↑	Top1-acc(%)
EdgeViT-XXS [40]	4.1	0.6	43.0	14.2	1926	74.4
MobileViT-XS [38]	2.3	1.1	100.2	15.6	1367	74.8
tiny-MOAT-0 [65]	3.4	0.8	61.1	14.7	1908	75.5
FAT-B0	4.5	0.7	44.3	14.4	1932	77.6
EdgeViT-XS [40]	6.7	1.1	62.7	15.7	1528	77.5
ParC-Net-S [76]	5.0	1.7	112.1	15.8	1321	78.6
EdgeNext-S [37]	5.6	1.3	86.4	14.2	1243	79.4
FAT-B1	7.8	1.2	62.6	14.5	1452	80.1
ParC-ResNet50 [76]	23.7	4.0	160.0	16.6	1039	79.6
tiny-MOAT-2 [65]	9.8	2.3	122.1	15.4	1047	81.0
EfficientNet-B3 [51]	12.0	1.8	124.2	25.4	624	81.6
FAT-B2	13.5	2.0	93.4	14.6	1064	81.9

Comparison with the state-of-the-art lighweight model on efficiency and performance

Experiments

Model	Params(M)	FLOPs(G)	Top1-acc(%)	AP^b	AP^m
Swin-S [34]	50	8.7	83.0	44.8	40.9
Focal-S [68]	51	9.1	83.5	47.4	42.8
CMT-B [16]	46	9.3	84.5	–	–
FAT-B4	52	9.3	84.8	49.7	44.8
CSwin-B [12]	78	15.0	84.2	–	–
MOAT-2 [65]	73	17.2	84.7	–	–
CMT-L [16]	75	19.5	84.8	–	–
FAT-B5	88	15.1	85.2	–	–

Comparison with general backbones.

Experiments

Backbone	RetinaNet 1×							Mask R-CNN 1×						
	Params(M)	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L	Params(M)	AP^b	AP^b_{50}	AP^b_{75}	AP^m	AP^m_{50}	AP^m_{75}
DFvT-T [15]	-	-	-	-	-	-	-	25	34.8	56.9	37.0	32.6	53.7	34.5
PVTv2-B0 [58]	13	37.2	57.2	39.5	23.1	40.4	49.7	24	38.2	60.5	40.7	36.2	57.8	38.6
QuadTree-B-b0 [52]	13	38.4	58.7	41.1	22.5	41.7	51.6	24	38.8	60.7	42.1	36.5	58.0	39.1
EdgeViT-XXS [40]	13	38.7	59.0	41.0	22.4	42.0	51.6	24	39.9	62.0	43.1	36.9	59.0	39.4
FAT-B0	14	40.4	61.6	42.7	24.0	44.3	53.1	24	40.8	63.3	44.2	37.7	60.2	40.0
DFvT-S [15]	-	-	-	-	-	-	-	32	39.2	62.2	42.4	36.3	58.9	38.6
EdgeViT-XS [40]	16	40.6	61.3	43.3	25.2	43.9	54.6	27	41.4	63.7	45.0	38.3	60.9	41.3
ViL-Tiny [77]	17	40.8	61.3	43.6	26.7	44.9	53.6	27	41.4	63.5	45.0	38.1	60.3	40.8
MPViT-T [28]	17	41.8	62.7	44.6	27.2	45.1	54.2	28	42.2	64.2	45.8	39.0	61.4	41.8
FAT-B1	17	42.5	64.0	45.1	26.9	46.0	56.7	28	43.3	65.6	47.4	39.6	61.9	42.8
PVTv1-Tiny [57]	23	36.7	56.9	38.9	22.6	38.8	50.0	33	36.7	59.2	39.3	35.1	56.7	37.3
ResT-Small [78]	23	40.3	61.3	42.7	25.7	43.7	51.2	33	39.6	62.9	42.3	37.2	59.8	39.7
PVTv2-B1 [58]	24	41.2	61.9	43.9	25.4	44.5	54.3	34	41.8	64.3	45.9	38.8	61.2	41.6
DFvT-B [15]	-	-	-	-	-	-	-	58	43.4	65.2	48.2	39.0	61.8	42.0
QuadTree-B-b1 [52]	24	42.6	63.6	45.3	26.8	46.1	57.2	34	43.5	65.6	47.6	40.1	62.6	43.3
EdgeViT-S[40]	23	43.4	64.9	46.5	26.9	47.5	58.1	33	44.8	67.4	48.9	41.0	64.2	43.8
MPViT-XS [28]	20	43.8	65.0	47.1	28.1	47.6	56.5	30	44.2	66.7	48.4	40.4	63.4	43.4
FAT-B2	23	44.0	65.2	47.2	27.5	47.7	58.8	33	45.2	67.9	49.0	41.3	64.6	44.0
Swin-T [34]	38	41.5	62.1	44.2	25.1	44.9	55.5	48	42.2	64.6	46.2	39.1	61.6	42.0
DAT-T [61]	38	42.8	64.4	45.2	28.0	45.8	57.8	48	44.4	67.6	48.5	40.4	64.2	43.1
DaViT-Tiny [10]	39	44.0	-	-	-	-	-	48	45.0	-	-	41.1	-	-
CMT-S [16]	44	44.3	65.5	47.5	27.1	48.3	59.1	45	44.6	66.8	48.9	40.7	63.9	43.4
MPViT-S [28]	32	45.7	57.3	48.8	28.7	49.7	59.2	43	46.4	68.6	51.2	42.4	65.6	45.7
QuadTree-B-b2 [52]	35	46.2	67.2	49.5	29.0	50.1	61.8	45	46.7	68.5	51.2	42.4	65.7	45.7
CSWin-T [12]	-	-	-	-	-	-	-	42	46.7	68.6	51.3	42.2	65.6	45.4
Shunted-S [47]	32	45.4	65.9	49.2	28.7	49.3	60.0	42	47.1	68.8	52.1	42.5	65.8	45.7
FAT-B3	39	45.9	66.9	49.5	29.3	50.1	60.9	49	47.6	69.7	52.3	43.1	66.4	46.2

Comparison to other backbones using RetinaNet and Mask-RCNN on COCO val2017 object detection and instance segmentation.

Ablation Study

Model	ImageNet-1K Params(M)	ImageNet-1K FLOPs(G)	ImageNet-1K Top-1(%)	COCO AP^b	COCO AP^m	ADE20K mIoU
add+linear	4.5	0.72	76.2	39.0	35.8	39.6
cat+linear	4.8	0.77	76.6	39.6	36.3	40.2
mul+linear	4.5	0.72	77.1	40.3	37.1	40.9
interaction	4.5	0.72	77.6	40.8	37.7	41.5
pool down	4.4	0.71	77.2	40.2	36.9	40.6
conv w/o overlap	4.4	0.71	77.2	40.3	36.9	40.8
conv w/ overlap	4.5	0.71	77.3	40.5	37.3	40.9
refined down	4.5	0.72	77.6	40.8	37.7	41.5
w/o conv. pos	4.4	0.70	77.4	40.5	37.3	41.2
conv. pos	4.5	0.72	77.6	40.8	37.7	41.5

Main components analysis

Method	Params (M)	FLOPs (G)	Top1-acc (%)
Max-SA [55]	4.3	0.8	75.7
WSA/S-WSA [34]	4.3	0.8	76.1
SRA [57]	4.4	0.7	76.2
LSA/GSA [5]	4.3	0.7	76.6
CSWSA [12]	4.3	0.8	76.8
FASA	4.5	0.7	**77.6**

Different self-attention mechanisms comparison

Model	Blocks	Channels	Params(M)	FLOPs(G)	Top-1 acc(%)
Swin-T [34]	[2, 2, 6, 2]	[96, 192, 384, 768]	29	4.5	81.3
DAT-T [61]	[2, 2, 6, 2]	[96, 192, 384, 768]	29	4.6	82.0
FocalNet-T [67]	[2, 2, 6, 2]	[96, 192, 384, 768]	28	4.4	82.1
Focal-T [68]	[2, 2, 6, 2]	[96, 192, 384, 768]	29	4.9	82.2
FAT-B3-ST	[2, 2, 6, 2]	[96, 192, 384, 768]	29	4.7	**83.0**

Comparison with four baseline models when use the same layout with Swin-T.

Summary

- an efficient vision Transformer backbone
- Fully Adaptive Self-Attention
- superior performance on many downstream vision tasks

References

- Inception Transformer. In NeurIPS, 2022.
- EdgeViTs：Competing Light—weight CNNs on Mobile Devices with Vision Transformers. In ECCV, 2022.
- Lightweight Vision Transformer with Bidirectional Interaction. In *NeurIPS*, 2023.

Lightweight Vision Transformer with Bidirectional Interaction

Thanks

Code Link

Group's Homepage