MPViT: Multi-Path Vision Transformer for Dense Prediction

Youngwan Lee[1,2] Jonghee Kim[1] Jeff Willette[2] Sung Ju Hwang[2,3]

[1]Electronics and Telecommunications Research Institute (ETRI), South Korea
[2]Korea Advanced Institute of Science and Technology (KAIST), South Korea
[3]AITRICS, South Korea

CVPR 2022

2022.04.11
Presenter: Sohee Jeong

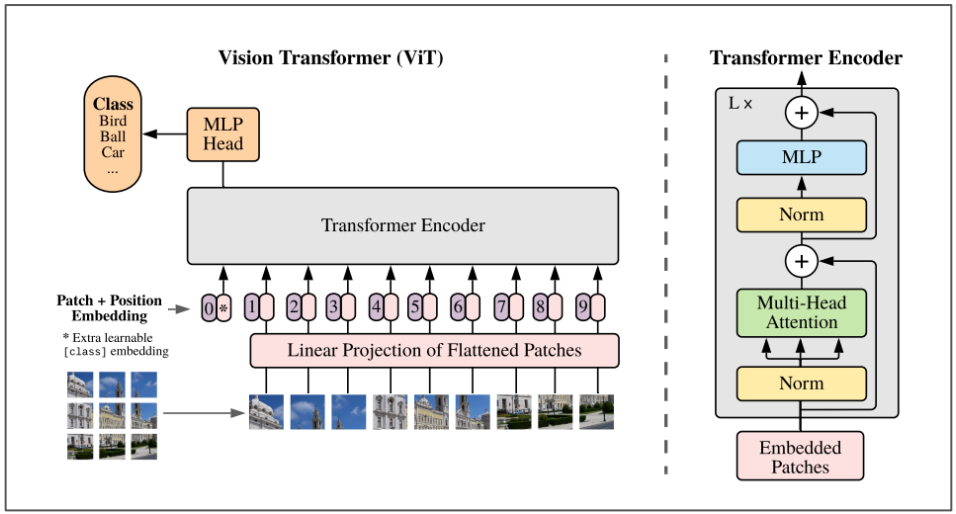
Contents

1. Vision Transformer
2. MPViT (main topic)
    - Motivation
    - Architecture
    - Methods
    - Experiments
    - Discussion

Vision Transformer

An Image is worth 16x16 words: Transformers for Image Recognition at scale

- Try to adapt pure transformer into vision task
- Patch embedding: divide image by 16x16 patch



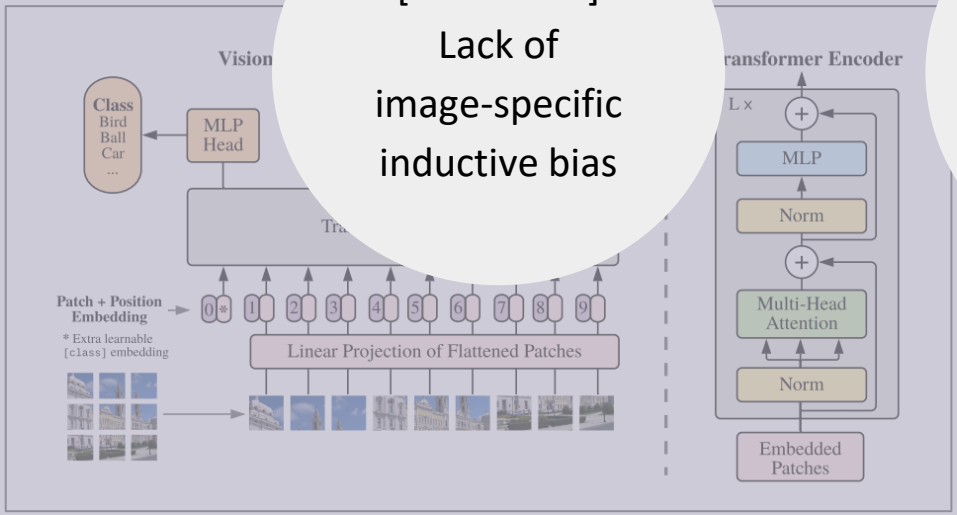
Vision Transformer

An Image is worth 16x16 words: Transformers for Image Recognition at scale

- Try to adapt pure transformer into vision task
- Patch embedding: divid… by 16x16 patch
- Have the advanta… mputational efficie…

[Problem 1] Lack of image-specific inductive bias

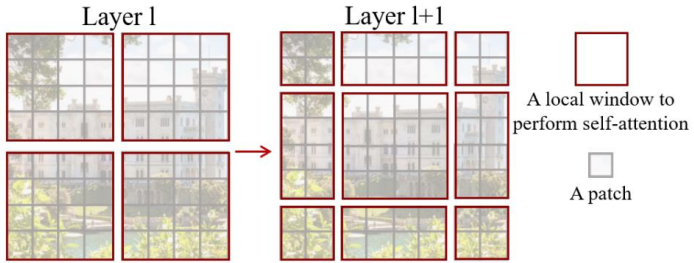
[Problem 2] High computational & memory cost



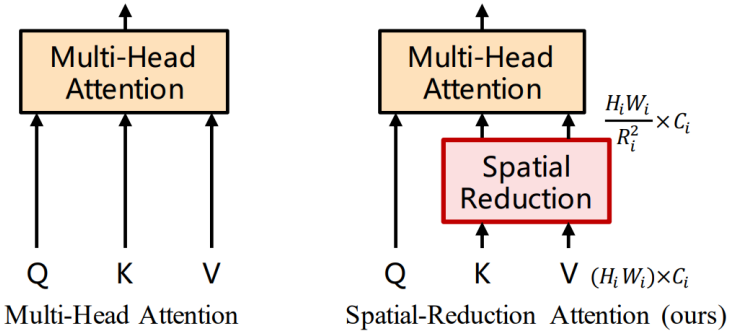
Vision Transformer

Existing Vision Transformer models focus on efficiency

- By constraining the attention range with fine-grained patches (Swin Transformer)
- Or reducing sequence length with spatial reduction (Pyramid Vision Transformer)



Swin Transformer –
Window self-attention



Pyramid Vision Transformer –
Spatial reduction attention

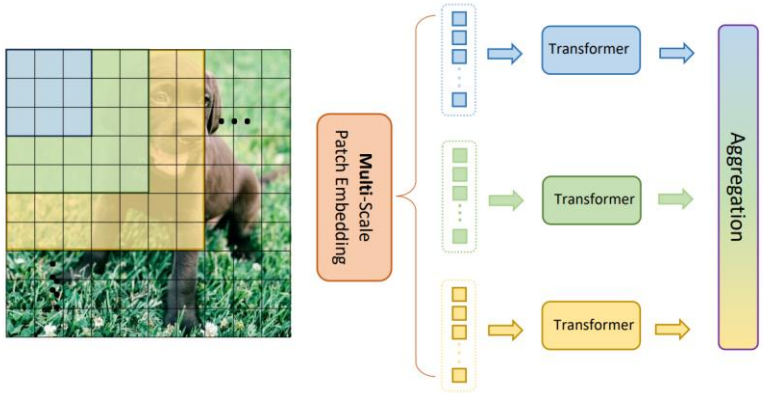
MPViT

Motivation

Multi-Path Vision Transformer for Dense Prediction (MPViT)

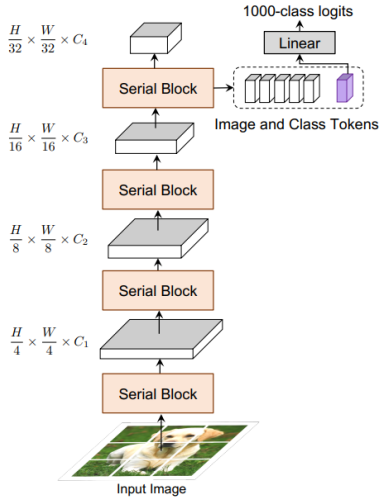
- Dense prediction task: crucial to represent features at multiple scales
- [CoaT] Consider Cross-scale interaction but high computational and memory overhead
- Improve in multi-scale feature representation for ViT architectures

MPViT focus on how to effectively represent multi-scale features with Vision Transformers for dense prediction tasks

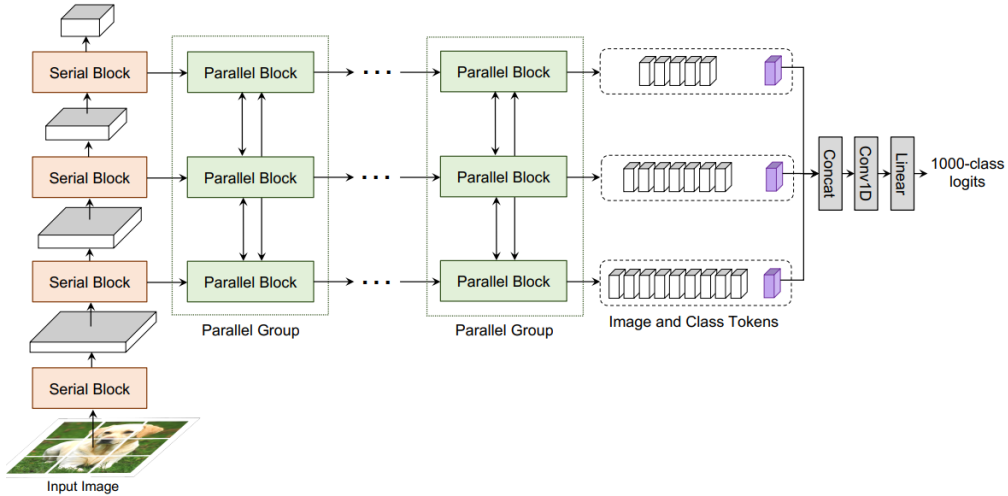


CoaT (*Co-scale Conv-Attentional Image Transformers)*

- [serial block] Enhanced multi-scale image representation: reduce resolution
- [parallel block] Cross-scale interaction: solve single-scale coarse patch embedding problem



CoaT-Lite

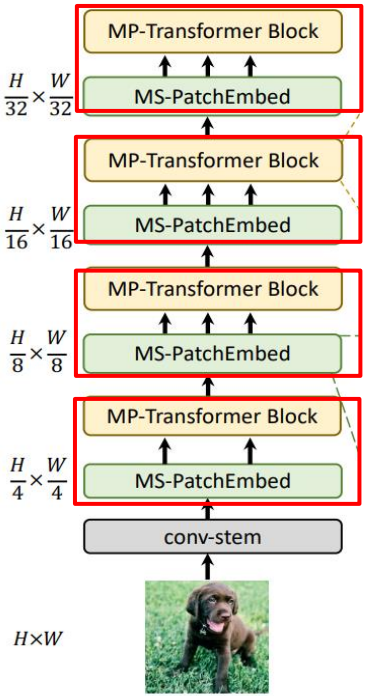


CoaT

MPViT

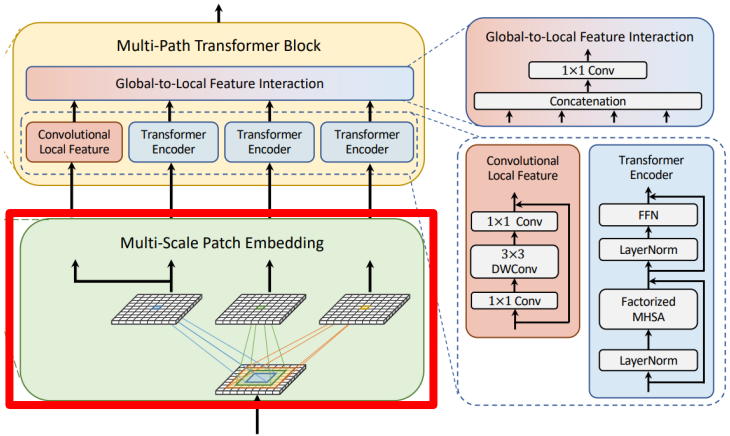
Architecture

MPViT Architecture



- Multi-Stage structure : generates feature map of different scales
- Each stage consists of MS-PatchEmbed + MP-Transformer Block
- Convolutional stem block : for lower-level representation
- Consists of two 3x3 convolution layers with stride 2

Multi-scale patch embedding (MS-PatchEmbed)



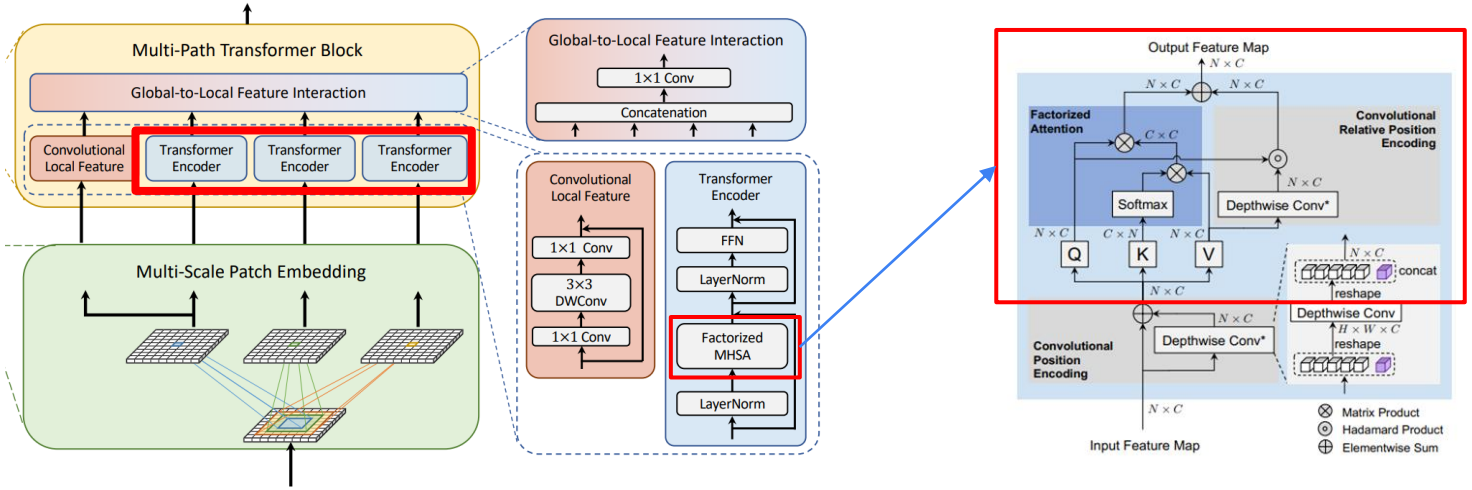

```python
# dw
self.dwconv = nn.Conv2d(
    in_ch,
    out_ch,
    kernel_size,
    stride,
    (kernel_size - 1) // 2,
    groups=out_ch,
    bias=False,
)
# pw-linear
self.pwconv = nn.Conv2d(out_ch, out_ch, 1, 1, 0, bias=False)
```

- 3x3 Depth-wise convolutions: less param #, less cost
- All convolution layers are followed by BatchNorm and Hardswish activation

- 2D convolution operation $\mathrm{F}_{\mathrm{k}\times\mathrm{k}}$ has kernel size $\mathrm{k} * \mathrm{k}$, stride s, padding p
- **Input** feature map from previous stage $\boldsymbol{X_i} \in \mathbb{R}^{\boldsymbol{H_{i-1}}\times\boldsymbol{W_{i-1}}\times\boldsymbol{C_{i-1}}}$
- **Output** feature map of MS-PatchEmbed is $\boldsymbol{F_{k\times k}(X_i)} \in \mathbb{R}^{\boldsymbol{H_i}\times\boldsymbol{W_i}\times\boldsymbol{C_i}}$,
- where $\mathrm{H_i} = \left\lfloor \frac{\mathrm{H_{i-1}}-\mathrm{k}+2\mathrm{p}}{\mathrm{s}} + 1 \right\rfloor, \mathrm{W_i} = \left\lfloor \frac{\mathrm{W_{i-1}}-\mathrm{k}+2\mathrm{p}}{\mathrm{s}} + 1 \right\rfloor$

Multi-Path Transformer (MP-Transformer)

- Each transformer block can be substituted to other transformer methods
- Use Factorized attention method from CoaT (Co-scale Conv-Attentional Image Transformers)



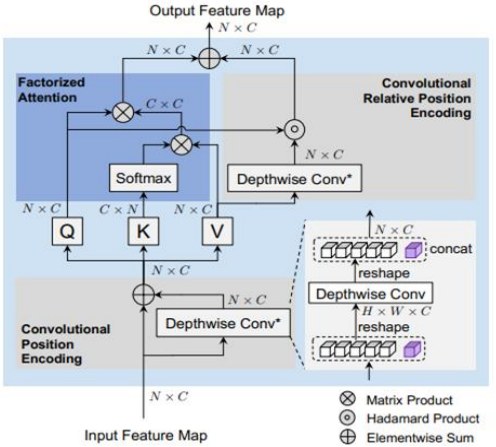
MPViT

Methods

Factorized Attention

$$FactorAtt(Q,K,V) = \frac{Q}{\sqrt{C}}(softmax(K)^T V), Q,K,V \in \mathbb{R}^{N\times C}$$

- token # N, channel # C, $N \gg C$ in general
- transformer encoder # h, layer # L
- Time complexity: $O(LhNC^2)$
- Space complexity: $O(LhC^2 + LhNC)$
- Reduce memory & computational cost

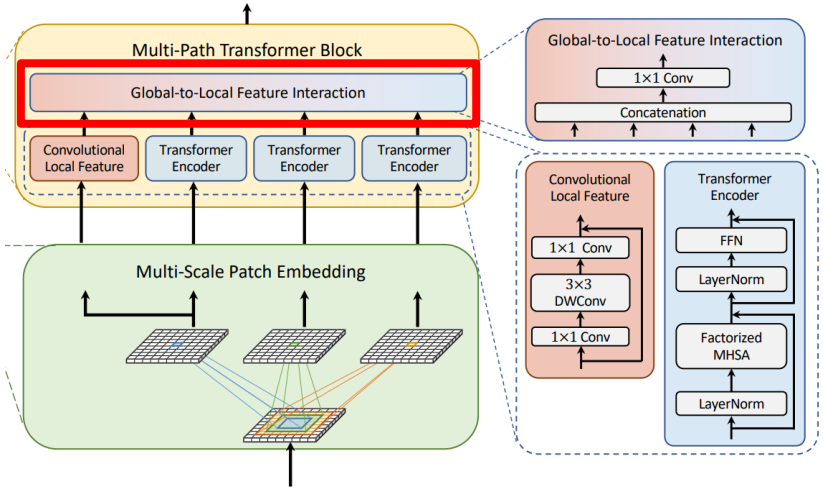



```
B, N, C = x.shape

# Generate Q, K, V.
qkv = (self.qkv(x).reshape(B, N, 3, self.num_heads,
                           C // self.num_heads).permute(2, 0, 3, 1, 4))
q, k, v = qkv[0], qkv[1], qkv[2]

# Factorized attention.
k_softmax = k.softmax(dim=2)
k_softmax_T_dot_v = einsum("b h n k, b h n v -> b h k v", k_softmax, v)
factor_att = einsum("b h n k, b h k v -> b h n v", q,
                    k_softmax_T_dot_v)

# Convolutional relative position encoding.
crpe = self.crpe(q, v, size=size)
```

2. MPViT

Global-to-Local Feature Interaction (GLI)

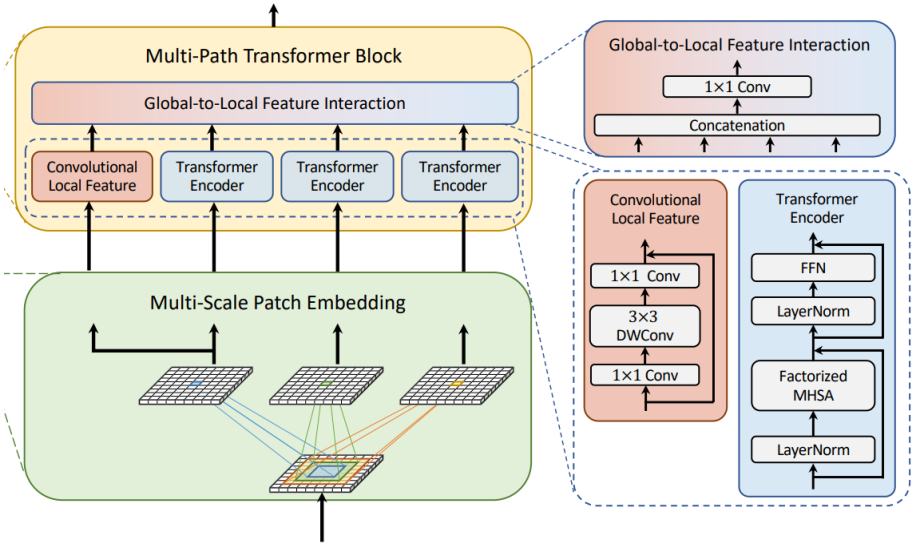
- CNN : local connectivity – each patch is processed by the same weights
- Transformer : capture long-range information, ignore local relationship within each patch
- Learns to interact between local & global features



2. MPViT

Methods

- At stage i | Convolutional Local Feature $L_i \in \mathbb{R}^{H_i x W_i x C_i}$, transformer output of path j $G_{i,j} \in \mathbb{R}^{H_i x W_i x C_i}$
- Aggregation of local and global features | $A_i = Concat([L_i, G_{i,0}, G_{i,1}, \dots, G_{i,j}])$, $A_i \in \mathbb{R}^{H_i x W_i x (j+1) C_i}$
- Final feature after 1x1 Convolution | $X_{i+1} = H(A_i)$, $X_{i+1} \in \mathbb{R}^{H_i x W_i x C_{i+1}}$



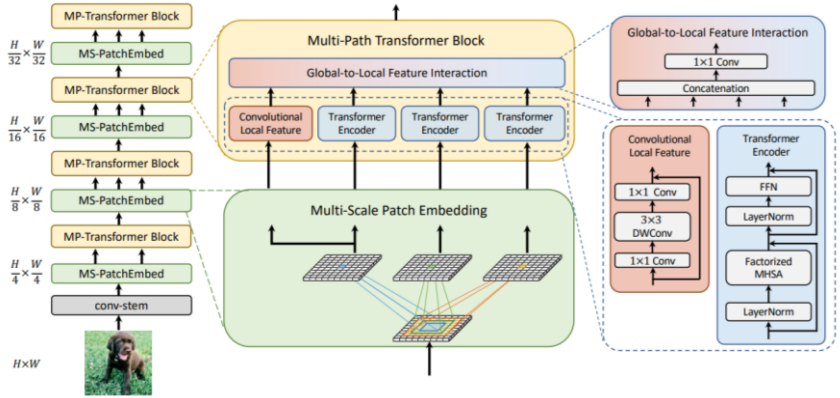
$$A_i = \mathrm{Concat}([L_i, G_{i,0}, G_{i,1}, ..., G_{i,j}])$$

$$X_{i+1} = H(A_i),$$

Model Configuration

Time complexity: $O(LhNC^2)$

Space complexity: $O(LhC^2 + LhNC)$



- Multi-path requires a reduction of C or L
- Reducing C is better than reducing L
- Triple-path structure: MPViT models use paths [2,3,3,3] in each of the 4 stages.

MPViT	#Layers	Channels	Param.	GFLOPs
Tiny (T)	[1, 2, 4, 1]	[64, 96, 176, 216]	5.7M	1.5
XSmall (XS)	[1, 2, 4, 1]	[64, 128, 192, 256]	10.5M	2.9
Small (S)	[1, 3, 6, 3]	[64, 128, 216, 288]	22.8M	4.7
Base (B)	[1, 3, 8, 3]	[128, 224, 368, 480]	74.8M	16.4

2. MPViT

Experiments

ImageNet-1K classification

Model	Param.(M)	GFLOPs	Top-1	Reference
DeiT-T [50]	5.7	1.3	72.2	ICML21
XCiT-T12/16 [17]	7.0	1.2	77.1	NeurIPS21
CoaT-Lite T [65]	5.7	1.6	76.6	ICCV21
MPViT-T	5.8	1.6	**78.2 (+1.6)**	
ResNet-18 [23]	11.7	1.8	69.8	CVPR16
PVT-T [58]	13.2	1.9	75.1	ICCV21
XCiT-T24/16 [17]	12.0	2.3	79.4	NeurIPS21
CoaT Mi [65]	10.0	6.8	80.8	ICCV21
CoaT-Lite Mi [65]	11.0	2.0	78.9	ICCV21
MPViT-XS	10.5	2.9	**80.9 (+2.0)**	
ResNet-50 [23]	25.6	4.1	76.1	CVPR16
PVT-S [58]	24.5	3.8	79.8	ICCV21
DeiT-S/16 [50]	22.1	4.6	79.9	ICML21
Swin-T [37]	29.0	4.5	81.3	ICCV21
CvT-13 [60]	20.0	4.5	81.6	ICCV21
XCiT-S12/16 [17]	26.0	4.8	82.0	NeurIPS21
Focal-T [67]	29.1	4.9	82.2	NeurIPS21
CoaT S [65]	22.0	12.6	82.1	ICCV21
CrossViT-15 [6]	28.2	6.1	82.3	ICCV21
CvT-21 [60]	32.0	7.1	82.5	ICCV21
CrossViT-18 [6]	43.3	9.5	82.8	ICCV21
CoaT-Lite S [65]	20.0	4.0	81.9	ICCV21
MPViT-S	22.8	4.7	**83.0 (+1.1)**	
ResNeXt-101 [64]	83.5	15.6	79.6	CVPR17
PVT-L [58]	61.4	9.8	81.7	ICCV21
DeiT-B/16 [50]	86.6	17.6	81.8	ICML21
XCiT-M24/16 [17]	84.0	16.2	82.7	NeurIPS21
Swin-B [37]	88.0	15.4	83.3	ICCV21
XCiT-S12/8 [17]	26.0	18.9	83.4	NeurIPS21
Focal-B [67]	89.8	16.0	83.8	NeurIPS21
MPViT-B	74.8	16.4	**84.3**	

2. MPViT

Experiments

Object Detection and Instance Segmentation

- IN1K pretrain, COCO dataset

Backbone	Params. (M)	GFLOPs	Mask R-CNN 3× schedule + MS						RetinaNet 3× schedule + MS					
			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^b_S	AP^b_M	AP^b_L
XCiT-T12/16 [17]	26	200	42.7	64.3	46.4	38.5	61.2	41.1	-	-	-	-	-	-
XCiT-T12/8 [17]	26	266	44.5	66.4	48.8	40.4	63.5	43.3	-	-	-	-	-	-
MPViT-T	28 (17)	216 (196)	**44.8**	66.9	49.2	**41.0**	64.2	44.1	**44.4**	65.5	47.4	29.9	48.3	56.1
PVT-T [58]	33 (23)	240 (221)	39.8	62.2	43.0	37.4	59.3	39.9	39.4	59.8	42.0	25.5	42.0	52.1
CoaT Mini [65]	30	307	46.5	67.9	50.7	41.8	65.3	44.8	-	-	-	-	-	-
CoaT-Lite Mini [65]	31	210	42.9	64.7	46.7	38.9	61.6	41.7	-	-	-	-	-	-
MPViT-XS	30 (20)	231 (211)	**46.6**	68.5	51.1	**42.3**	65.8	45.8	**46.1**	67.4	49.3	31.4	50.2	58.4
PVT-S [58]	44 (34)	305 (226)	43.0	65.3	46.9	39.9	62.5	42.8	42.2	62.7	45.0	26.2	45.2	57.2
XCiT-S12/16 [17]	44	285	45.3	67.1	49.5	40.8	64.0	43.8	-	-	-	-	-	-
Swin-T [37]	48 (39)	267 (245)	46.0	68.1	50.3	41.6	65.1	44.9	45.0	65.9	48.4	29.7	48.9	58.1
XCiT-S12/8 [17]	43	550	47.0	68.9	51.7	42.3	66.0	45.4	-	-	-	-	-	-
Focal-T [67]	49 (39)	291 (265)	47.2	69.4	51.9	42.7	66.5	45.9	45.5	66.3	48.8	31.2	49.2	58.7
CoaT S [65]	42	423	**49.0**	70.2	53.8	43.7	67.5	47.1	-	-	-	-	-	-
CoaT-Lite S [65]	40	249	45.7	67.1	49.8	41.1	64.1	44.0	-	-	-	-	-	-
MPViT-S	43 (32)	268 (248)	48.4	70.5	52.6	**43.9**	67.6	47.5	**47.6**	68.7	51.3	32.1	51.9	61.2
PVT-M [58]	64 (54)	392 (283)	44.2	66.0	48.2	40.5	63.1	43.5	43.2	63.8	46.1	27.3	46.3	59.9
PVT-L [58]	81 (71)	494 (345)	44.5	66.0	48.3	40.7	63.4	43.7	43.4	63.6	46.1	26.1	46.0	59.5
XCiT-M24/16 [17]	101	523	46.7	68.2	51.1	42.0	65.5	44.9	-	-	-	-	-	-
XCiT-S24/8 [17]	65	892	48.1	69.5	53.0	43.0	66.5	46.1	-	-	-	-	-	-
XCiT-M24/8 [17]	99	1448	48.5	70.3	53.4	43.7	67.5	46.9	-	-	-	-	-	-
Swin-S [37]	69 (60)	359 (335)	48.5	70.2	53.5	43.3	67.3	46.6	46.4	67.0	50.1	31.0	50.1	60.3
Swin-B [37]	107 (98)	496 (477)	48.5	69.8	53.2	43.4	66.8	49.6	45.8	66.4	49.1	29.9	49.4	60.3
Focal-S [67]	71 (62)	401 (367)	48.8	70.5	53.6	43.8	67.7	47.2	47.3	67.8	51.0	31.6	50.9	61.1
Focal-B [67]	110 (101)	533 (514)	49.0	70.1	53.6	43.7	67.6	47.0	46.9	67.8	50.3	31.9	50.3	61.5
MPViT-B	95 (85)	503 (482)	**49.5**	70.9	54.0	**44.5**	68.3	48.3	**48.3**	69.5	51.9	32.3	52.2	62.3

Object Detection and Instance Segmentation

- IN1K pretrain, COCO dataset

Backbone	Params. (M)	GFLOPs	Mask R-CNN 3× schedule + MS						RetinaNet 3× schedule + MS					
			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^b_S	AP^b_M	AP^b_L
XCiT-T12/16 [17]	26	200	42.7	64.3	46.4	38.5	61.2	41.1	-	-	-	-	-	-
XCiT-T12/8 [17]	26	266	44.5	66.4	48.8	40.4	63.5	43.3	-	-	-	-	-	-
MPViT-T	28 (17)	216 (196)	**44.8**	66.9	49.2	**41.0**	64.2	44.1	**44.4**	65.5	47.4	29.9	48.3	56.1
PVT-T [58]	33 (23)	240 (221)	39.8	62.2	43.0	37.4	59.3	39.9	39.4	59.8	42.0	25.5	42.0	52.1
CoaT Mini [65]	30	307	46.5	67.9	50.7	41.8	65.3	44.8	-	-	-	-	-	-
CoaT-Lite Mini [65]	31	210	42.9	64.7	46.7	38.9	61.6	41.7	-	-	-	-	-	-
MPViT-XS	30 (20)	231 (211)	**46.6**	68.5	51.1	**42.3**	65.8	45.8	**46.1**	67.4	49.3	31.4	50.2	58.4
PVT-S [58]	44 (34)	305 (226)	43.0	65.3	46.9	39.9	62.5	42.8	42.2	62.7	45.0	26.2	45.2	57.2
XCiT-S12/16 [17]	44	285	45.3	67.1	49.5	40.8	64.0	43.8	-	-	-	-	-	-
Swin-T [37]	48 (39)	267 (245)	46.0	68.1	50.3	41.6	65.1	44.9	45.0	65.9	48.4	29.7	48.9	58.1
XCiT-S12/8 [17]	43	550	47.0	68.9	51.7	42.3	66.0	45.4	-	-	-	-	-	-
Focal-T [67]	49 (39)	291 (265)	47.2	69.4	51.9	42.7	66.5	45.9	45.5	66.3	48.8	31.2	49.2	58.7
CoaT S [65]	42	423	**49.0**	70.2	53.8	43.7	67.5	47.1	-	-	-	-	-	-
CoaT-Lite S [65]	40	249	45.7	67.1	49.8	41.1	64.1	44.0	-	-	-	-	-	-
MPViT-S	43 (32)	268 (248)	48.4	70.5	52.6	**43.9**	67.6	47.5	**47.6**	68.7	51.3	32.1	51.9	61.2
PVT-M [58]	64 (54)	392 (283)	44.2	66.0	48.2	40.5	63.1	43.5	43.2	63.8	46.1	27.3	46.3	59.9
PVT-L [58]	81 (71)	494 (345)	44.5	66.0	48.3	40.7	63.4	43.7	43.4	63.6	46.1	26.1	46.0	59.5
XCiT-M24/16 [17]	101	523	46.7	68.2	51.1	42.0	65.5	44.9	-	-	-	-	-	-
XCiT-S24/8 [17]	65	892	48.1	69.5	53.0	43.0	66.5	46.1	-	-	-	-	-	-
XCiT-M24/8 [17]	99	1448	48.5	70.3	53.4	43.7	67.5	46.9	-	-	-	-	-	-
Swin-S [37]	69 (60)	359 (335)	48.5	70.2	53.5	43.3	67.3	46.6	46.4	67.0	50.1	31.0	50.1	60.3
Swin-B [37]	107 (98)	496 (477)	48.5	69.8	53.2	43.4	66.8	49.6	45.8	66.4	49.1	29.9	49.4	60.3
Focal-S [67]	71 (62)	401 (367)	48.8	70.5	53.6	43.8	67.7	47.2	47.3	67.8	51.0	31.6	50.9	61.1
Focal-B [67]	110 (101)	533 (514)	49.0	70.1	53.6	43.7	67.6	47.0	46.9	67.8	50.3	31.9	50.3	61.5
MPViT-B	95 (85)	503 (482)	**49.5**	70.9	54.0	**44.5**	68.3	48.3	**48.3**	69.5	51.9	32.3	52.2	62.3

2. MPViT

Experiments

Semantic Segmentation

- IN1K pretrain MPViT + UperNet
- ADE20k dataset

Backbone	Params.	GFLOPs	mIoU
Swin-T [37]	59M	945	44.5
Focal-T [67]	62M	998	45.8
XCiT-S12/16 [17]	54M	966	45.9
XCiT-S12/8 [17]	53M	1237	46.6
MPViT-S	52M	943	**48.3**
XCiT-S24/16 [17]	76M	1053	46.9
Swin-S [37]	81M	1038	47.6
XCiT-M24/16 [17]	112M	1213	47.6
Focal-S [67]	85M	1130	48.0
Swin-B [37]	121M	1841	48.1
XCiT-S24/8 [17]	74M	1587	48.1
XCiT-M24/8 [17]	110M	2161	48.4
Focal-B [67]	126M	1354	49.0
MPViT-B	105M	1186	**50.3**

Ablation study

Exploring path dimension

Path	Spec	Param.	GFLOPs	Memory	img/sec	Top-1	AP^{box}	AP^{mask}
Single	[1,1,1,1]P_[2,2,2,2]L_[64, 128, 320, 512]C	11.0M	1.9	9216	1195	78.9	40.2	37.3
(a) Dual	[2,2,2,2]P_[1,2,4,1]L_[64, 128, 256, 320]C	10.9M	2.6	6054	945	80.7+1.8	42.6+2.4	39.1+1.8
(b) Triple	[2,3,3,3]P_[1,1,2,1]L_[64, 128, 256, 320]C	10.8M	2.3	6000	1080	79.8+0.9	41.4+1.2	38.0+0.7
(c) Triple	[2,3,3,3]P_[1,2,4,1]L_[64, 128, 192, 256]C	10.1M	2.7	5954	803	80.5+1.6	**43.0**+2.8	**39.4**+2.1
(d) Quad	[2,4,4,4]P_[1,2,4,1]L_[64, 96, 176, 224]C	10.5M	2.6	5990	709	80.5+1.6	42.4+2.2	38.8+1.5

Table 5. **Exploring the path dimension.** Spec means [#path_per_stage]P, [#layer_per_stage]L and [dimension_per_stage]C. We measure inference throughput and peak GPU memory usage on V100 GPU with batch size of 256. Note that the single-path is CoaT-Lite Mini [65].

- Reducing C is better than reducing L – for both performance and memory
- Triple-path structure

2. MPViT

Experiments

Ablation study

Path	Param.	GFLOPs	Top-1	AP^b/AP^m
Single (CoaT-Lite Mini)	11.01M	1.99	78.9	40.2 / 37.3
+ Triple (p=[3,5,7], parallel)	10.18M	2.78	80.3	41.7 / 38.4
+ Triple (p=[3,3,3], series)	10.15M	2.67	**80.5**	**43.0 / 39.4**
+ GLI (Sum)	10.13M	2.82	80.3	43.0 / 39.5
+ GLI (Concat.)	10.57M	2.97	**80.8**	**43.3 / 39.7**

➢ Multi-Scale Embedding

➢ GLI aggregation method

Multi-Scale Embedding

- better to use three convolution layers in series - for performance, model size, and FLOPs

GLI aggregation method

- Concatenation shows improvement on both classification and detection tasks
- Summing naively mixes the features, but concatenation preserves them

2. MPViT

Discussion

Qualitative Analysis – Attention maps

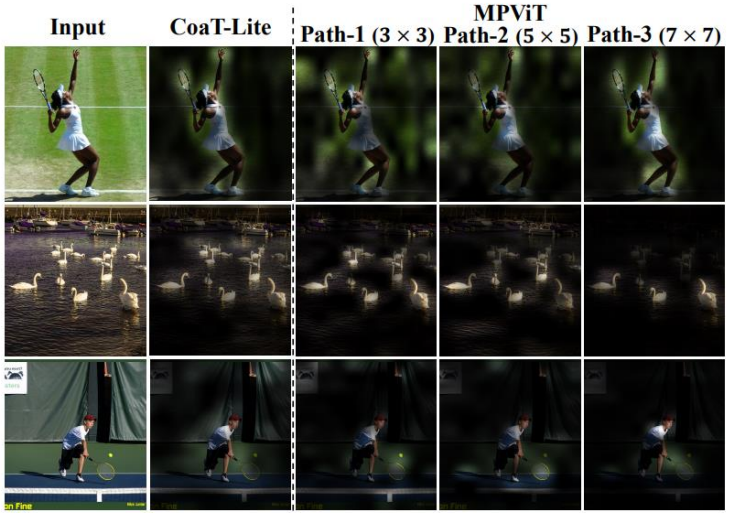


Figure 4. **Attention maps** generated by CoaT-Lite and MPViT at stage4. MPViT has a triple-path structure with patches of various sizes (*e.g.*, $3 \times 3, 5 \times 5, 7 \times 7$), leading to fine and coarse features.

Path	Spec	Param.	GFLOPs	Memory	img/sec	Top-1	AP^{box}	AP^{mask}
Single	[1,1,1,1]P_[2,2,2,2]L_[64, 128, 320, 512]C	11.0M	1.9	9216	1195	78.9	40.2	37.3
(a) Dual	[2,2,2,2]P_[1,2,4,1]L_[64, 128, 256, 320]C	10.9M	2.6	6054	945	80.7+1.8	42.6+2.4	39.1+1.8
(b) Triple	[2,3,3,3]P_[1,1,2,1]L_[64, 128, 256, 320]C	10.8M	2.3	6000	1080	79.8+0.9	41.4+1.2	38.0+0.7
(c) Triple	[2,3,3,3]P_[1,2,4,1]L_[64, 128, 192, 256]C	10.1M	2.7	5954	803	80.5+1.6	**43.0**+2.8	**39.4**+2.1
(d) Quad	[2,4,4,4]P_[1,2,4,1]L_[64, 96, 176, 224]C	10.5M	2.6	5990	709	80.5+1.6	42.4+2.2	38.8+1.5

- Visualize attention maps for each path
- Attention maps from CoaT-Lite and path-1 have similar patch sizes and show similar attention maps

- Path-1 : fine patches, small objects, low-level representation
- Path-3 : larger patches, object centric, high-level representation

2. MPViT

Discussion

Qualitative Analysis – Attention maps

- Three paths act in a complementary manner, which is beneficial for dense prediction tasks
- Difficult to capture all objects with a single path

Qualitative Analysis - Failure cases

Figure 7. **Attention Maps of failure cases on ImageNet *validation* images.** The input image and corresponding attention maps from each path are illustrated. In the rightmost column, we show the ground truth labels and predicted labels colored with red and blue, respectively.

References

- MPViT: https://arxiv.org/pdf/2112.11010.pdf
- MPViT code: https://github.com/youngwanLEE/MPViT/blob/e2d86e6c465abc4847e601f8a401bdd1692836f7/mpvit.py#L128
- CoaT: https://arxiv.org/pdf/2104.06399.pdf
- SwinT: https://arxiv.org/abs/2103.14030
- Pyramid Vision Transformer: https://arxiv.org/abs/2102.12122
- XCiT: https://arxiv.org/abs/2106.09681

Thank you