ECCV 2024

Embedding-Free Transformer with Inference Spatial Reduction for Efficient Semantic Segmentation

Sogang University

Vision & Display Systems Lab, Dept. of Electronic Engineering

Presented by

Hyunwoo Yu, Yubin Cho, Beoungwoo Kang, Seunghun Moon, Kyeongbo Kong, Suk-Ju Kang

Outline

- Background

- Method
    - Embedding-Free Attention (EFA) structure
    - Inference Spatial Reduction (ISR) method

- Experiment

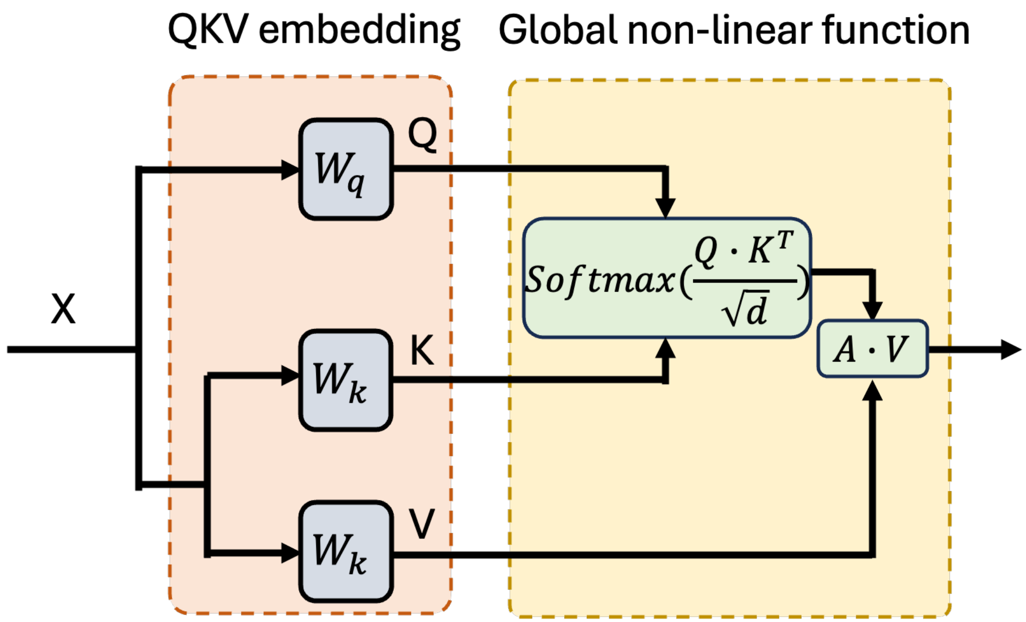
Background

- Transformer-based architecture shows great success in computer vision
- Large amount of computation and parameter in transformer structure

  Especially, the computational cost of transformer is crucial in high resolution task such as semantic segmentation

- In this paper, we analyze the general self-attention mechanism as two parts.

  The first is QKV embedding phase and the second is global non-linear functioning

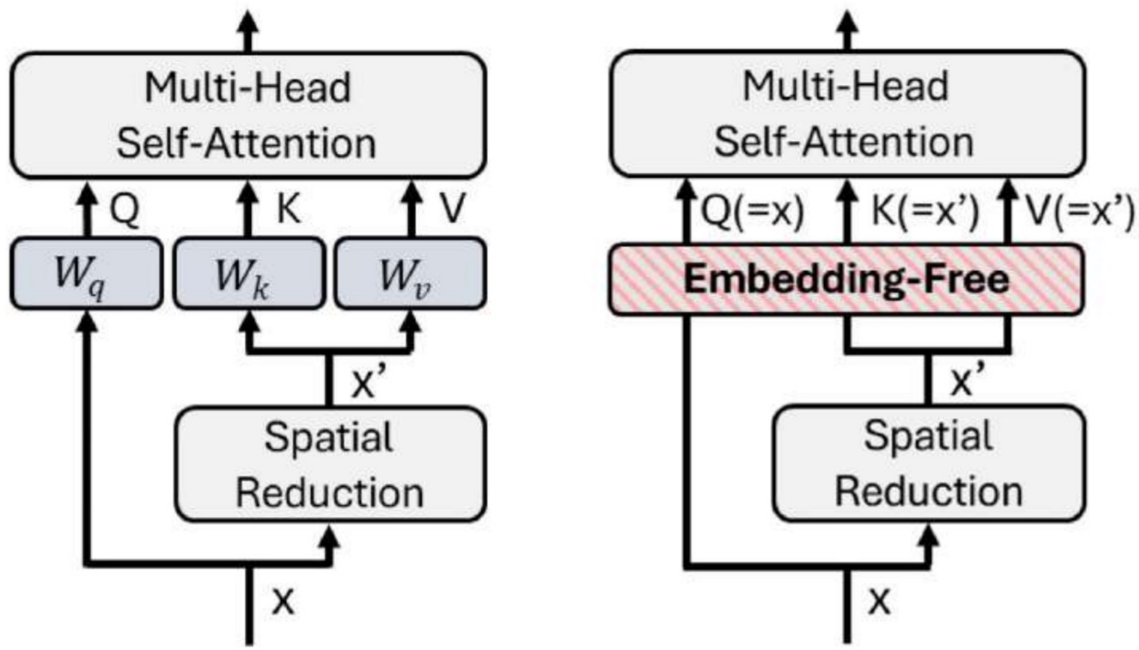


Self-attention structure

Method

- Embedding-Free Attention (EFA) structure

  Remove the query, key, value embedding phase and focus on the non-linear global functioning



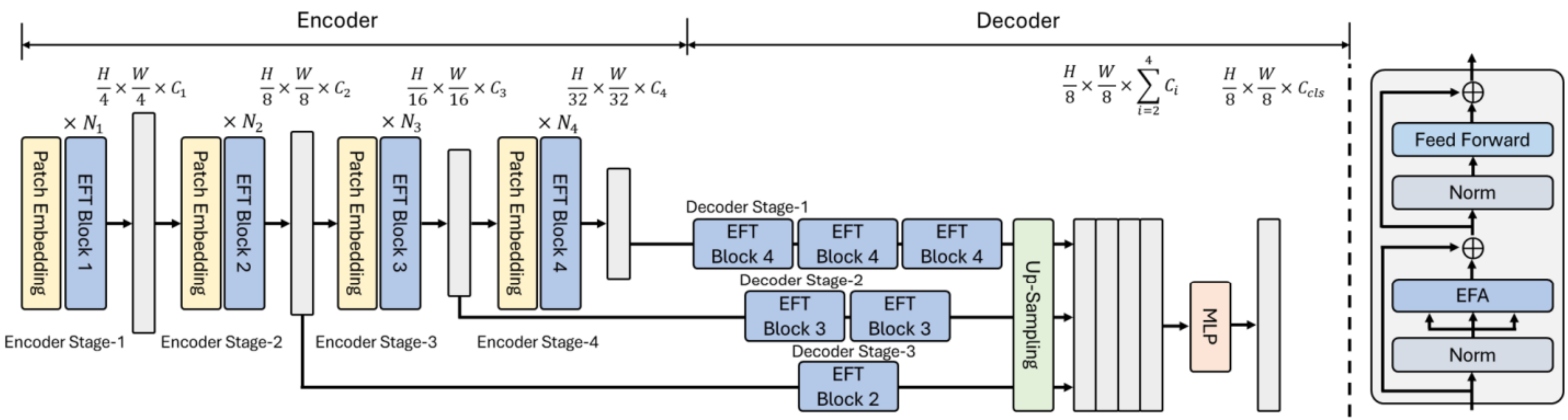
(a) Embedding-based SRA (b) Embedding-Free Attention (Our EFA)

Method

- Encoder-Decoder Attention Transformer (EDAFormer) architecture

  Based on our powerful EFA module, we design the semantic segmentation model

  - EDAFormer composed with EFA transfomrer block (EFT) in encoder-decoder.
  - The decoder leverage the more number of EFA module to the high-level features



(a) Overall architecture of EDAFormer

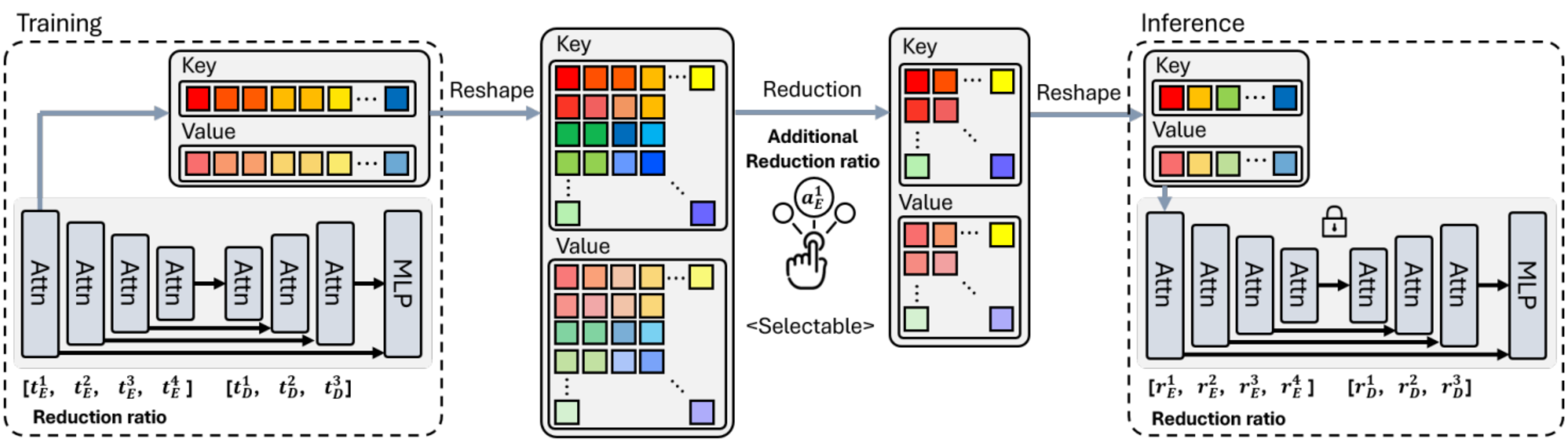
(b) EFT block

Method

- Inference Spatial Reduction(ISR) method

  Reduce the key-value resolution in inference phase.

  - Reduce the computation with little performance degradation
  - Segmentation specific method by maintained the input-output resolution
    - In self-attention mechanism, the reduction of key-value resolution does not affect to the output resolution



Overview of Inference Spatial Reduction(ISR) method

Experiment

Method	Params (M)	ADE20K GFLOPs ↓	ADE20K mIoU (%) ↑	Cityscapes GFLOPs ↓	Cityscapes mIoU (%) ↑	COCO-Stuff GFLOPs ↓	COCO-Stuff mIoU (%) ↑
Segformer-B0 [65]	3.8	8.4	37.4	125.5	76.2	8.4	35.6
FeedFormer [50]	4.5	7.8	39.2	107.4	77.9	-	-
VWFormer-B0 [66]	3.7	5.1	38.9	-	77.2	5.1	36.2
EDAFormer-T (w/o ISR)	4.9	5.6	42.3	151.7	78.7	5.6	40.3
EDAFormer-T (w/ ISR)	4.9	**4.7**	**42.1**	**94.9**	**78.7**	**4.7**	**40.3**
OCRNet [17]	70.5	164.8	45.6	1296.8	81.1	-	-
Swin UperNet-T [40]	60.0	236.0	44.4	-	-	-	-
ContrastiveSeg [57]	58.0	-	-	-	79.2	-	-
SenFormer [2]	144.0	179.0	46.0	-	-	-	-
Segformer-B2 [65]	27.5	62.4	46.5	717.1	81.0	62.4	44.6
ProtoSeg [80]	90.5	-	48.6	-	80.6	-	42.4
MaskFormer [10]	42.0	55.0	46.7	-	-	-	-
Mask2Former [9]	47.0	74.0	47.7	-	-	-	-
FeedFormer-B2 [50]	29.1	42.7	48.0	522.7	81.5	-	-
VWFormer-B2 [66]	27.4	38.5	48.1	-	81.7	38.5	45.2
EDAFormer-B (w/o ISR)	29.4	32.0	49.0	605.9	81.6	32.0	45.9
EDAFormer-B (w/ ISR)	29.4	**29.4**	**48.9**	**452.9**	**81.6**	**29.4**	**45.8**

Table 1. Comparison with semantic segmentation model

Models	Params (M)	GFLOPs	Top-1 Acc. (%)
RSB-ResNet-18 [29, 61]	12	1.8	70.6
PVTv2-B0 [59]	3.4	0.6	70.5
MiT-B0 [65]	3.7	0.6	70.5
EFT-T (Ours)	3.7	0.6	**72.3**
ResNet50 [29]	25.5	4.1	78.5
RSB-ResNet-152 [29, 61]	60.0	11.6	81.8
DeiT-S [54]	22.0	4.6	79.8
PVT-Small [58]	25.0	3.8	79.8
PVTv2-B2 [59]	25.4	4.0	82.0
MiT-B2 [65]	25.4	4.0	81.6
T2T-ViT-14 [74]	21.5	4.8	81.5
TNT-S [26]	23.8	4.8	81.5
ResMLP-S24 [53]	30.0	6.0	79.4
Swin-Mixer-T/D6 [40]	23.0	4.0	79.7
Visformer-S [8]	40.2	4.8	82.1
gMLP-S [37]	20.0	4.5	79.6
PoolFormer-S36 [71]	31.0	5.0	81.4
EfficientFormer-L3 [35]	31.3	3.9	82.4
FasterViT-0 [27]	31.4	3.3	82.1
EFT-B (Ours)	25.4	4.2	**82.4**

Table 2. Comparison with classification model

Mechanism	QKV Embedding MFLOPs ↓	QKV Embedding Params (K)	Global Functioning MFLOPs ↓	Global Functioning Params (K)	Output Projection MFLOPs ↓	Output Projection Params (K)	Others MFLOPs ↓	Others Params (K)	Total MFLOPs ↓	Total Params (K)
Ω (SRA)	4.82	49.6	2.46	0.0	3.21	16.5	0.83	16.5	11.32	82.6
Ω (EFA w/o ISR)	0.00	0.0	2.46	0.0	3.21	16.5	0.83	16.5	**6.50** (-42.6%)	**33.0** (-60.0%)
Ω (EFA w/ ISR)	0.00	0.0	0.61	0.0	3.21	16.5	0.18	16.5	**4.00** (-64.7%)	**33.0** (-60.0%)

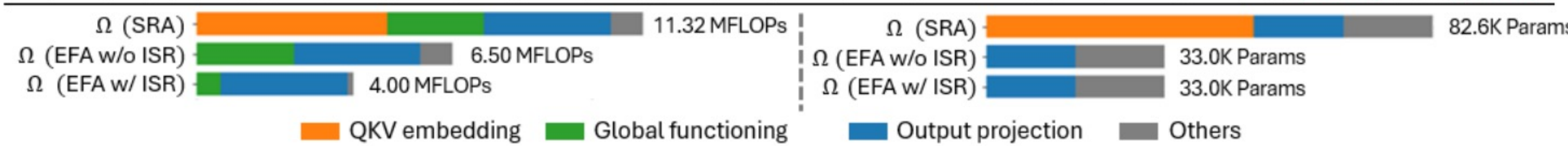


Table 3. Computation analysis of attention block. The FLOPs and parameters were computed on stage 3 features of 224 × 224 size

Experiment

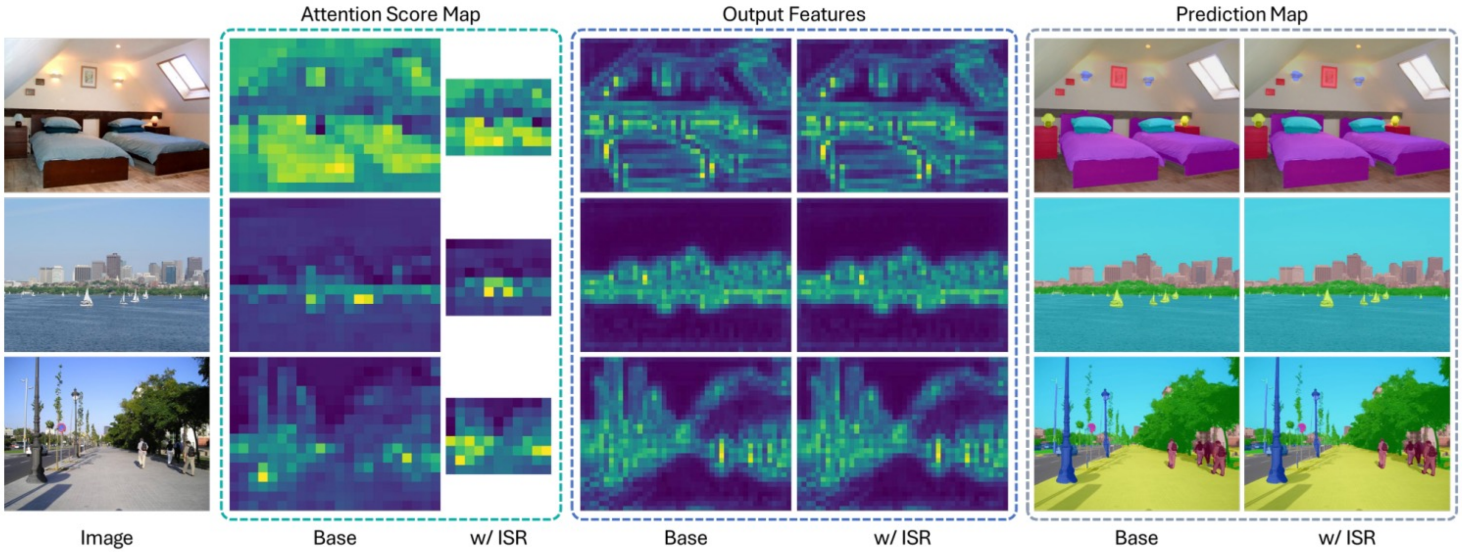


Figure 1. Visualization of the attention map, output features and prediction map on ADE20K

$[r_E^1, r_E^2, r_E^3, r_E^4]$-$[r_D^1, r_D^2, r_D^3]$ Reduction ratio Train	Inference	Params (M)	ADE20K GFLOPs ↓	ADE20K mIoU (%) ↑	Cityscapes GFLOPs ↓	Cityscapes mIoU (%) ↑	COCO-Stuff GFLOPs ↓	COCO-Stuff mIoU (%) ↑
(a) EDAFormer-T with the different reduction ratio at inference.								
[8, 4, 2, 1]-[1, 2, 4]	[8, 4, 2, 1]-[1, 2, 4]†	4.9	5.6	42.3	151.7	78.7	5.6	40.3
	[8, 4, 2, 1]-[2, 4, 8]	4.9	5.3 (-5.4%)	42.2 (-0.1)	133.6 (-11.9%)	78.7 (-0.0)	5.3 (-5.4%)	40.3 (-0.0)
	[16, 8, 2, 1]-[2, 4, 8]	4.9	4.7 (-16.1%)	42.1 (-0.2)	94.9 (-37.4%)	78.7 (-0.0)	4.7 (-16.1%)	40.3 (-0.0)
	[16, 8, 4, 2]-[2, 4, 8]	4.9	4.1 (-26.8%)	41.3 (-1.0)	59.1 (-61.0%)	78.1 (-0.6)	4.1 (-26.8%)	39.1 (-1.2)
	[16, 8, 4, 2]-[2, 4, 8]*	4.9	4.1 (-26.8%)	42.1 (-0.2)	59.1 (-61.0%)	78.6 (-0.1)	4.1 (-26.8%)	40.2 (-0.1)
(b) EDAFormer-B with the different reduction ratio at inference.								
[8, 4, 2, 1]-[1, 2, 4]	[8, 4, 2, 1]-[1, 2, 4]†	29.4	32.0	49.0	605.9	81.6	32.0	45.9
	[8, 4, 2, 1]-[2, 4, 8]	29.4	31.3 (-2.2%)	48.9 (-0.1)	569.0 (-6.1%)	81.6 (-0.0)	31.3 (-2.2%)	45.8 (-0.1)
	[16, 8, 2, 1]-[2, 4, 8]	29.4	29.4 (-8.1%)	48.9 (-0.1)	452.9 (-25.3%)	81.6 (-0.0)	29.4 (-8.1%)	45.8 (-0.1)
	[16, 8, 4, 2]-[2, 4, 8]	29.4	26.6 (-16.9%)	48.3 (-0.7)	298.1 (-50.8%)	81.4 (-0.2)	26.6 (-16.9%)	45.0 (-0.9)
	[16, 8, 4, 2]-[2, 4, 8]*	29.4	26.6 (-16.9%)	48.7 (-0.3)	298.1 (-50.8%)	81.6 (-0.0)	26.6 (-16.9%)	45.7 (-0.2)

Table 4. Comparison with different reduction ration condition of our ISR method