Samsung AI Center -Cambridge

香港中文大學 The Chinese University of Hong Kong

EdgeViTs: Competing Light-weight CNNs on Mobile Devices with Vision Transformers

ECCV TEL AVIV 2022

Junting Pan, Adrian Bulat, Fuwen Tan, Xiatian Zhu, Lukasz Dudziak, Hongsheng Li, Georgios Tzimiropoulos and Brais Martinez

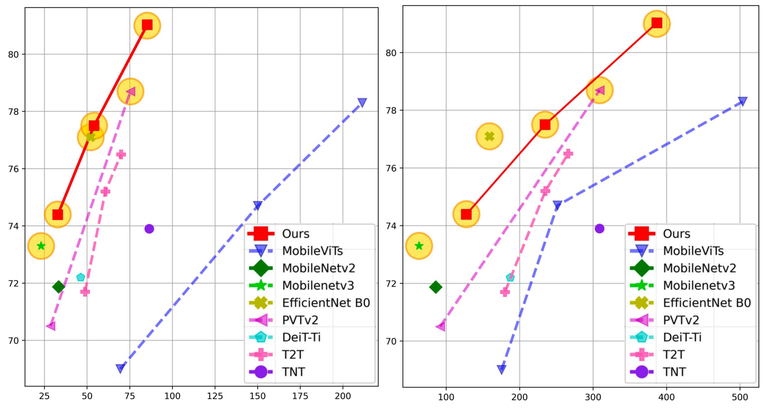
Source code: https://github.com/saic-fi/edgevit

1. Contributions

(1) We investigate the **design of light-weight ViTs** from the **practical on-device deployment** and execution perspective.

(2) We present **EdgeViTs**, based on an optimal decomposition of self-attention using standard primitive operations.

(3) We consider **latency** and **energy consumption** of different models rather than the number of FLOPs or parameters.



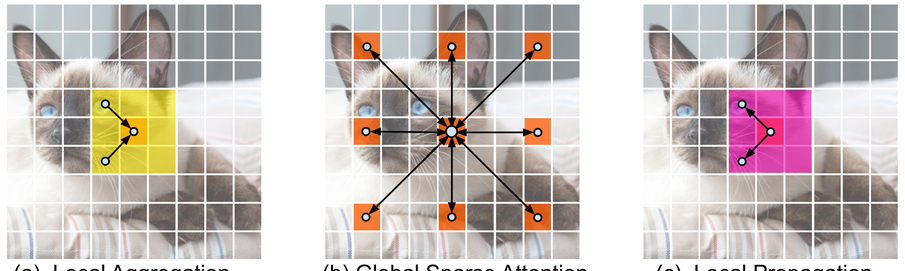
Accuracy/Latency and **Accuracy/Energy** trade off on ImageNet-1K. Note that, **all three variants of our EdgeViTs are pareto-optimal**, are highlighted with amber cercle.

Testing device: Samsung Galaxy S21 (latency), Snapdragon 888 Hardware Development Kit (energy).

2. EdgeViTs

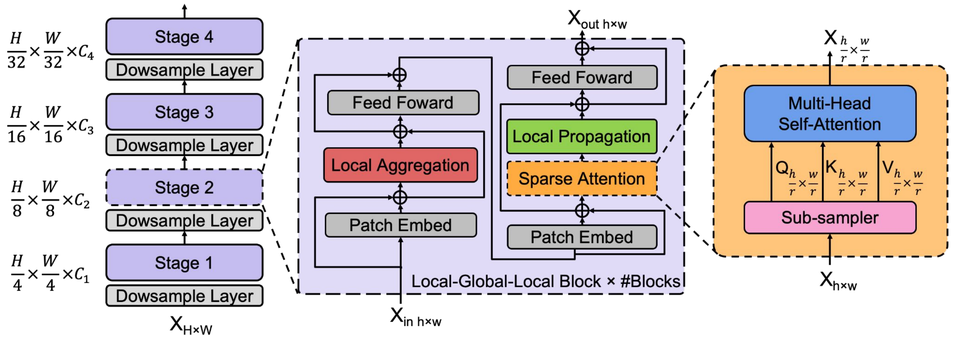
EdgeViTs are based on a factorization of the standard self-attention by introducing a **light-weight and easy-to-implement** local-global-local (LGL) bottleneck with three operations:

(1) Local information aggregation from neighbor tokens with depth-wise convolutions;
(2) Forming a sparse set of evenly distributed delegate tokens for long-range information exchange by self-attention;
(3) Diffusing updated information from delegates to the non-delegate tokens in local neighborhoods via transposed convolutions.



(a) Local Aggregation (b) Global Sparse Attention (c) Local Propagation

$$
\begin{aligned}
X &= \texttt{LocalAgg}(\texttt{Norm}(X_{in})) + X_{in}, \\
Y &= \texttt{FFN}(\texttt{Norm}(X)) + X, \\
Z &= \texttt{LocalProp}(\texttt{GlobalSparseAttn}(\texttt{Norm}(Y))) + Y, \\
X_{out} &= \texttt{FFN}(\texttt{Norm}(Z)) + Z.
\end{aligned}
$$



3. Results

On-device evaluation on ImageNet-1K

Model	Top-1 (%)	CPU (ms)	Energy (mJ)	Power(W)	Efficiency (%/msW)
MobileNet-v2	72.0	33.3	85.7±7.4	3.31±0.26	0.841
MobileNet-v3 0.75	**73.3**	**23.0**	**63.0**±9.6	**3.46**±0.4	**1.164**
EfficientNet-B0	**77.1**	**52.1**	**159.0**±26.2	**3.62**±0.45	**0.485**
PVT-v2-B0	70.5	26.0	91.7±19.7	3.94±0.68	0.769
PVT-v2-B1	**78.7**	75.4	309.0±65.8	4.63±0.71	0.255
Twins-SVT-Tiny*	71.2	36.9	114.5±17.3	3.71±0.24	0.622
DeiT-Tiny	72.2	46.2	187.2±7.6	4.77±0.21	0.386
Uniformer-Tiny*	74.1	40.5	134.7±27.3	4.1±0.71	0.55
T2T-ViT-12	76.5	69.9	266.2±42.6	4.37±0.36	0.287
TNT-Tiny	73.9	86.4	308.7±70.5	3.94±0.63	0.239
LeViT-384†	**79.5**	**71.3**	**455.2**±125.8	**6.18**±0.74	**0.173**
MobileViT-XXS	69.0	69.5	175.3±28.7	2.77±0.24	0.394
MobileViT-XS	74.7	150.1	251.5±81.1	2.63±0.61	0.297
MobileViT-S	78.3	221.3	503.6±117.0	2.76±0.21	0.155
EdgeViT-XXS	**74.4**	**32.8**	**127.4**±27.3	**4.27**±0.67	**0.584**
EdgeViT-XS	**77.5**	**54.1**	**234.6**±44.0	**4.77**±0.84	**0.33**
EdgeViT-S	**81.0**	**85.3**	**386.7**±43.5	**4.8**±0.26	**0.209**

We define an energy-aware efficiency metric as the average gain in top-1 accuracy from each 1W run for 1ms (equivalent to consuming 1mJ of energy)

COCO Object Detection

	RetinaNet 1×							Mask R-CNN 1×						
Backbone	#Par.	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L	#Par.	AP^b	AP^b_{50}	AP^b_{75}	AP^m	AP^m_{50}	AP^m_{75}
PVTv2-B0	**13.0**	37.2	57.2	39.5	**23.1**	40.4	49.7	**23.5**	38.2	60.5	40.7	36.2	57.8	38.6
EdgeViT-XXS	13.1	**38.7**	**59.0**	**41.0**	22.4	**42.0**	**51.6**	23.8	**39.9**	**62.0**	**43.1**	**36.9**	**59.0**	**39.4**
EdgeViT-XS	16.3	40.6	61.3	43.3	25.2	43.9	54.6	26.5	41.4	63.7	45.0	38.3	60.9	41.3
ResNet18	**21.3**	31.8	49.6	33.6	16.3	34.3	43.2	**31.2**	34.0	54.0	36.7	31.2	51.0	32.7
PVTv1-Tiny	23.0	36.7	56.9	38.9	22.6	38.8	50.0	32.9	36.7	59.2	39.3	35.1	56.7	37.3
PVTv2-B1	23.8	41.2	61.9	43.9	25.4	44.5	54.3	33.7	41.8	64.3	45.9	38.8	61.2	41.6
EdgeViT-S	22.6	**43.4**	**64.9**	**46.5**	**26.9**	**47.5**	**58.1**	32.8	**44.8**	**67.4**	**48.9**	**41.0**	**64.2**	**43.8**

COCO Instance Segmentation

Backbone	Semantic FPN		
	#Param (M)	GFLOPs	mIoU (%)
PVTv2-B0	**7.6**	**25.0**	37.2
EdgeViT-XXS	7.9	24.4	**39.7**
EdgeViT-XS	10.6	27.7	41.4
ResNet18	**15.5**	32.2	32.9
PVTv1-Tiny	17.0	33.2	35.7
PVTv2-B1	17.8	34.2	42.5
EdgeViT-S	16.9	**32.1**	**45.9**