KNU KYUNGPOOK NATIONAL UNIVERSITY

Fair Comparison between Efficient Attentions

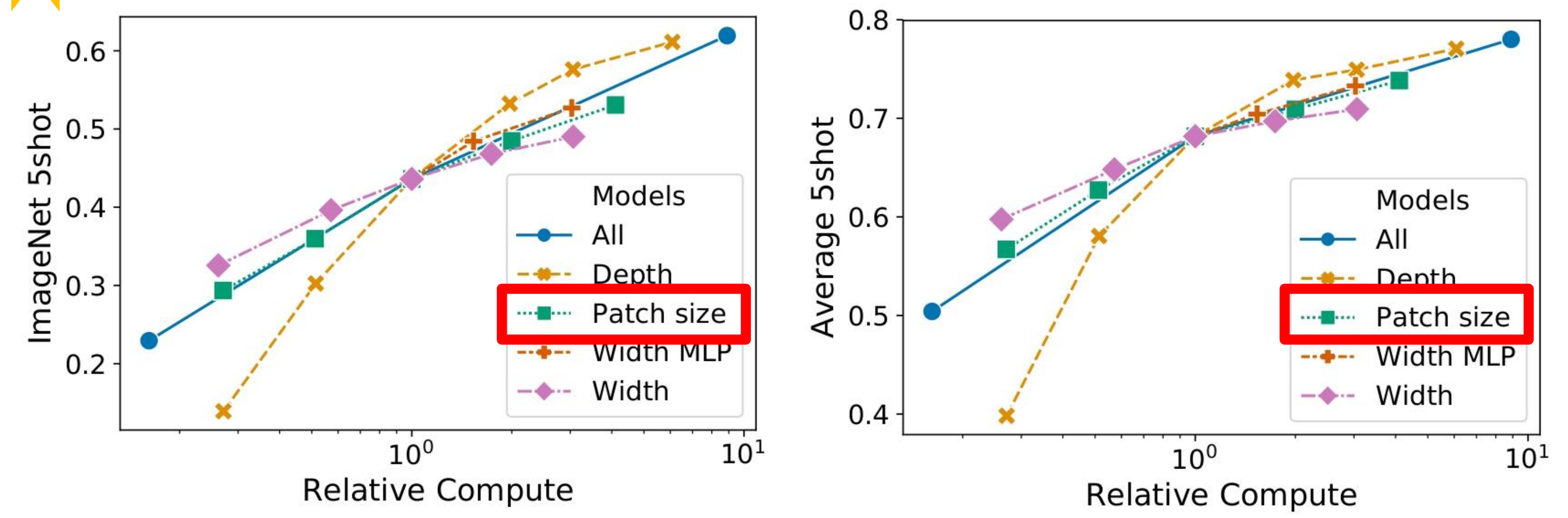
Jiuk Hong[1], Chaehyeon Lee[1], Soyoun Bang[1] and Heechul Jung[1]

[1]Department of Artificial Intelligence, Kyungpook National University

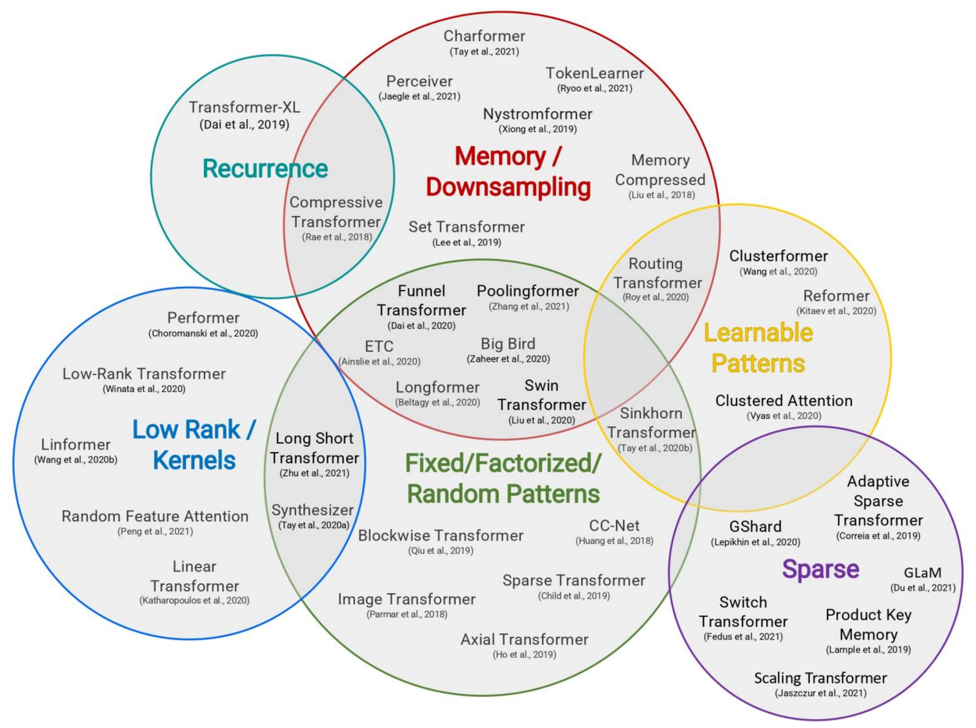
Workshop on Transformers for Vision

CVPR JUNE 19-24 2022 NEW ORLEANS • LOUISIANA

Motivation

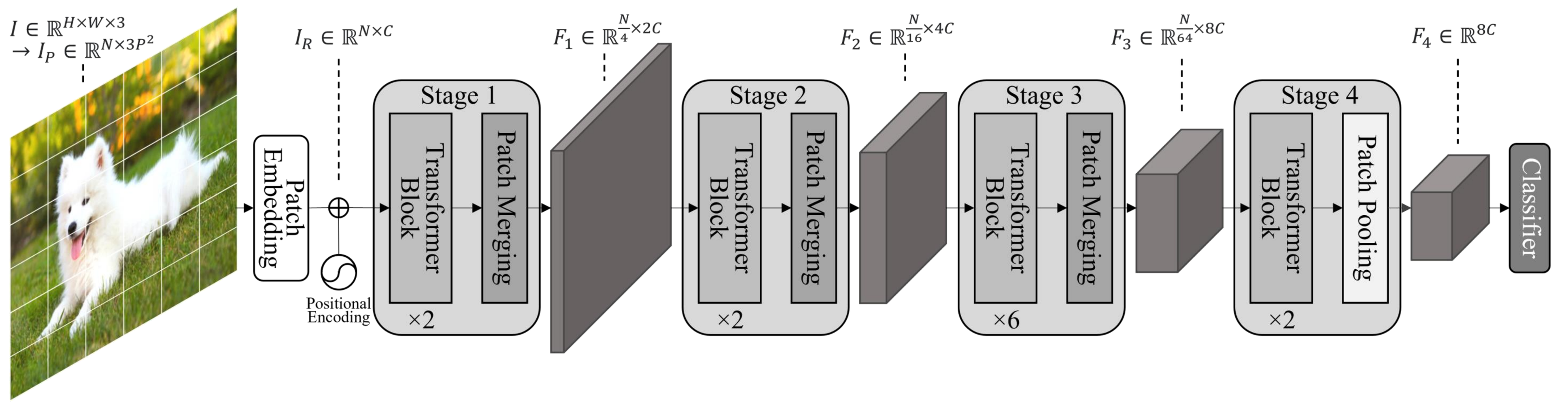


- ✓ In ViT [5], decreasing the patch size and thus **increasing the effective sequence length shows surprisingly robust improvements** without introducing parameters.
  - A well-known concern with self-attention is the quadratic time and memory complexity.
  - The quadratic complexity can occur patch size scalability issues.



- ✓ There has been many studies related to **efficient attention** [9] that addresses this problem.
  - But there are few attempts to compare the works fairly because of different model configuration and training schemes.
- ✓ Then, how about using those works to reduce patch size while decreasing computations?

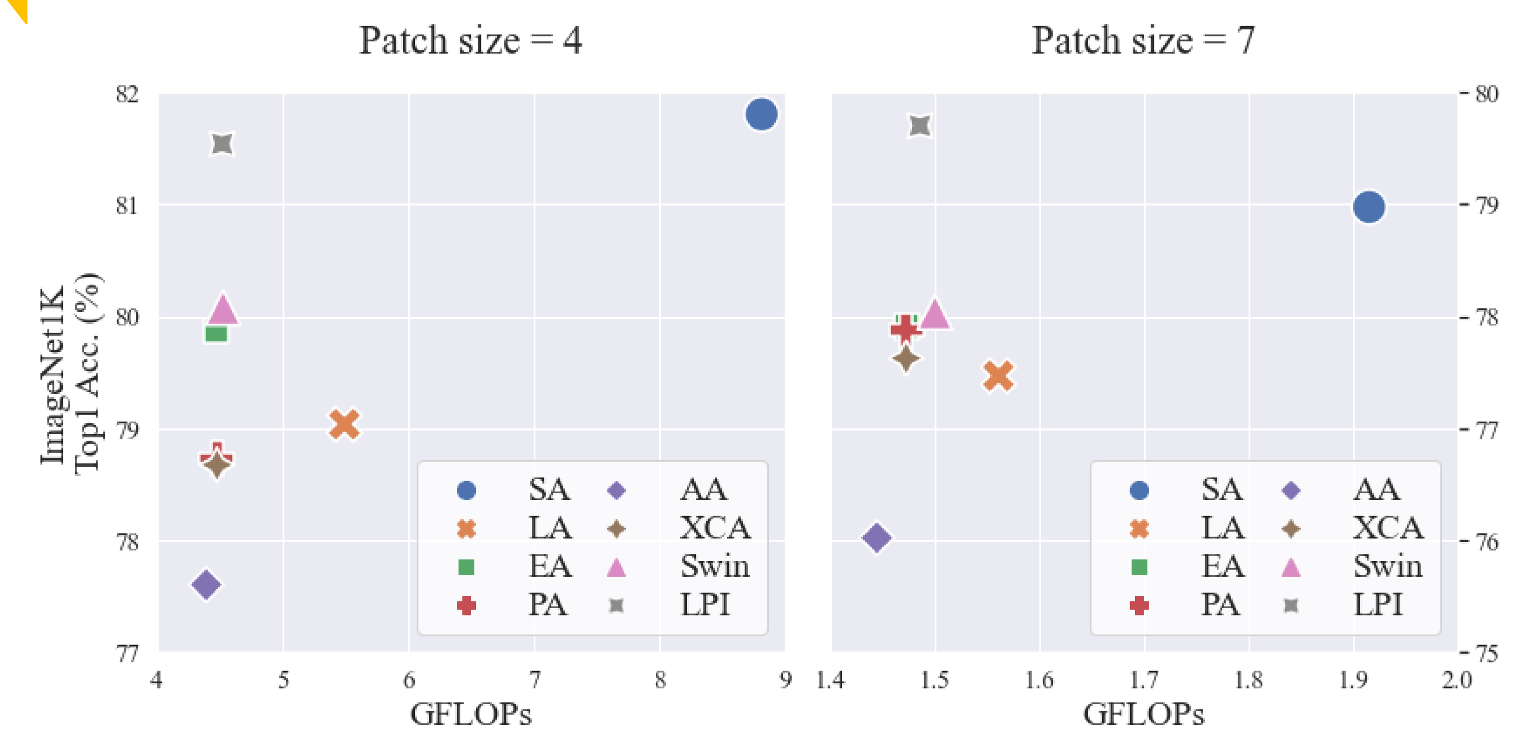
Experiment Setting



- ✓ We use the **pyramid architecture** from Pyramid Transformer [12].
  - Like CNN, it reduce spatial dimension and increase feature dimension.
  - While we conduct experiments using small size of patch, we cannot afford to handle computation with columnar architecture like ViT.
- ✓ For a fair comparison, we use same architectures and training schemes. **Only the patch size and the type of attention change**.
  - We compare several attentions, which have global token interaction and linear complexity.
  - For reference, we also evaluate Swin, which uses local token interaction, and columnar architecture with patch size of 14 and 16.
  - Below table shows efficient attention operation used in our paper and its corresponding complexity.

Model Architecture	Equation	Complexity per Self-Attention
Transformer (SA) [10]	$\mathrm{softmax}\left[\frac{\mathbf{q}\mathbf{k}^T}{\sqrt{d_k}}\right]\mathbf{v}$	$O(N^2C)$
Linformer (LA) [11]	$\mathrm{softmax}\left(\frac{\mathbf{q}[W_{proj}\mathbf{k}]^T}{\sqrt{d_k}}\right)W_{proj}\mathbf{v}.$	$O(NCm)$
Efficient Attention (EA) [8]	$\mathrm{softmax(q)}\left[\mathrm{softmax}(\mathbf{k}^T)\mathbf{v}\right]$	$O(NC^2)$
Performer (PA) [3]	$\frac{\psi(\mathbf{q})\ [\psi(\mathbf{k})^T\mathbf{v}]}{\mathrm{diag}(\psi(\mathbf{q})[\psi(\mathbf{k})^T\mathbb{I}_N])}$	$O(NCr)$
Fastformer (AA) [15]	q + [k'* v]W	$O(NC)$
XCiT (XCA) [1]	$\left[\mathrm{softmax}\left(\frac{\lVert\mathbf{q}\rVert_2{}^T\lVert\mathbf{k}\rVert_2}{\tau}\right)\mathbf{v}^T\right]^T$	$O(NC^2)$
Swin Transformer (Swin) [6]	same as SA, but using window	$O(NCw^2)$

Result on ImageNet-1K



Type of Attention	#params (M, millions)	FLOPs (G, giga)	FLOPs ratio	Top 1 Acc. (%)
Baseline				
SA-4 [10]	28.27	8.821	1	81.80
SA-7 [10]	28.28	1.915	1	78.97
Efficient Attentions				
LA-4 [11]	30.91	5.496	0.62	79.04(-2.76)
LA-7 [11]	28.56	1.561	0.81	77.47(-1.5)
EA-4 [8]	28.27	4.480	0.51	79.87**(-1.93)**
EA-7 [8]	28.28	1.473	0.77	77.91**(-1.06)**
PA-4 [3]	28.27	4.481	0.51	78.73(-3.07)
PA-7 [3]	28.28	1.473	0.77	77.87(-1.1)
AA-4 [15]	28.27	4.394	0.50	77.60(-3.93)
AA-7 [15]	28.28	1.445	0.75	76.02(-2.95)
XCA-4 [1]	28.27	4.480	0.51	78.67(-3.13)
XCA-7 [1]	28.28	1.473	0.77	77.62(-1.35)
References				
Swin-4 [6]	28.27	4.528	0.51	80.08(-1.72)
Swin-7 [6]	28.28	1.500	0.78	78.72(-0.25)
LPI-4 [1]	28.38	4.520	0.51	81.54(-0.26)
LPI-7 [1]	28.39	1.486	0.78	79.7(+0.73)
COL-14	22.00	6.117	0.69	81.30(-0.50)
COL-16	22.00	4.589	0.52	80.97(-0.83)

- ✓ Efficient attentions have lower performance, but much lower computation.
- ✓ Additional methods such as two convolution layer before attention (LPI) or shifted window attention (Swin) show comparable results with baseline.
- ✓ Pyramid architecture with small patches does not show superior performance compared to columnar architectures.

Limitation

- ✓ More experiments with columnar architecture are needed to show above results under additional model configurations.
- ✓ Attention with much smaller patches is ineffective due to its high computational cost.