Evolution of Vision Transformers. From ***ViT*** to ***GC ViT*** and ***Next-ViT***

Dmytro Kuzmenko

& Data Science UA

About me

- ML Researcher @UofT, ML Engineer @Infopulse. Mentor @Projector Institute and @WWCode.
- Education: BSc Software Engineering, MSc Applied Maths @NaUKMA.
- More than 4 years in AI/ML. Devised multiple elaborate pipelines with tabular, time series, textual and spatio-temporal data.
- Current research focus lies in Adversarial Attacks and Defenses. Finalizing the first scientific work about Semi-Supervised Learning in Vision-Based Automated Stair Recognition.
- Playing shogi, looking to make an RL-based engine for the game, drinking tea, riding scooters.

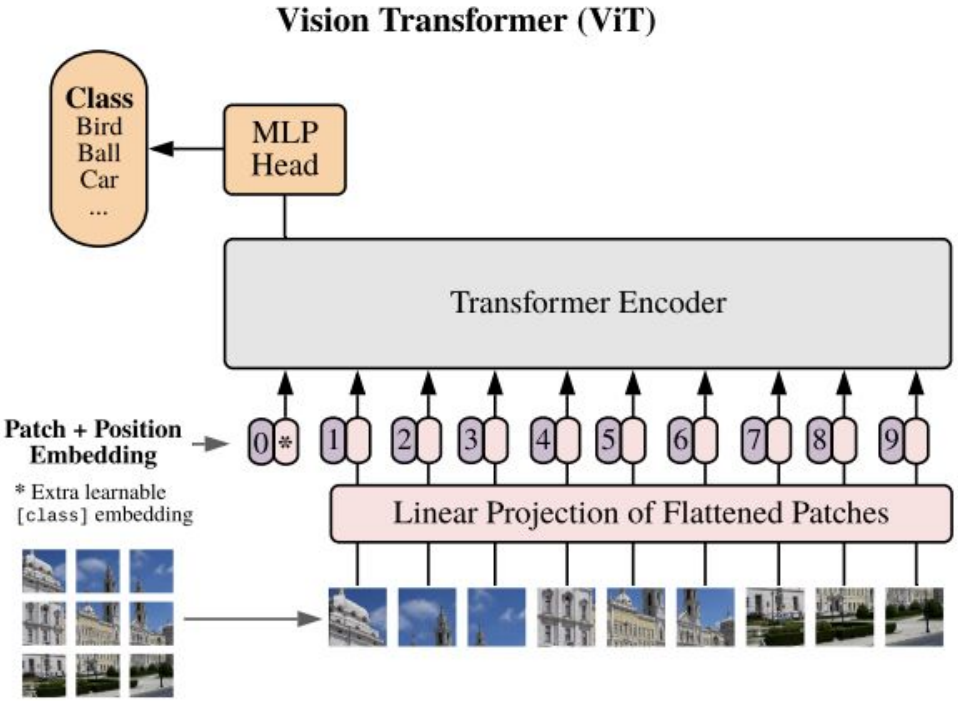
LinkedIn, Web-page

Meetup Outline

- About me
- Introduction to ViT
- GC ViT
- Next-ViT
- Summary and future provisions
- Reference literature
- Q/A

Intro to ViTs

The **Vision Transformer** (ViT) was first proposed as an alternative to CNNs with the advantage of an **enlarged receptive field**, due to its self-attention layers. However, it **lacked** desirable properties of CNNs such as **inductive biases** and **translation invariance** and required large-scale training datasets to achieve competitive performance.



ViT Bottlenecks

Transformers have achieved SOTA performance in NLP benchmarks and become the de facto model for various tasks. A key element in the success of Transformers is the **self-attention mechanism** which allows for capturing contextual representations via attending to both distant and nearby tokens. Making use of this idea, **Vision Transformers** (ViTs) proposed to **utilize image patches as tokens** in a monolithic architecture very similar to the encoder part of the original Transformers.

Even though CNNs were historically dominant in computer vision, ViT-based models have achieved SOTA or competitive performance in various vision tasks. In essence, the **self-attention mechanism** in ViT allows learning more uniform **short and long-range information unlike in CNN**.

However, the **monolithic architecture** of ViT and **quadratic computational complexity of self-attention** hindered their successful application to high-resolution images in which capturing multi-scale long-range information is crucial for accurate representation modeling.

Intro to ViTs - Successors

- ➢ **Data-efficient Image Transformers** (DeiT) introduced a *distillation-based training strategy* with significantly improved performance.
- ➢ **LeViT** proposed a *hybrid model* with a re-designed multi-layer perceptron (MLP) and self-attention modules that are highly optimized for fast inference.
- ➢ **Cross-covariance Image Transformer** (XCiT) introduced a *transposed self-attention module* for modeling the interactions of feature channels.
- ➢ **Convolutional vision Transformer** (CvT) introduced *convolutional token embedding layer and transformer block in a hierarchical architecture* to improve the efficiency and accuracy of ViTs
- ➢ **Conditional Position encoding Vision Transformer** (CPVT) showed great performance by *conditioning the position encoding on localized patch tokens*.

Intro to ViTs - Successors

- **Tokens-To-Token Vision Transformer** (T2T-ViT) proposed a t*ransformation layer for aggregating adjacent tokens* and establishing images prior by *exploiting spatial correlations*.
- **Focal Transformer** introduced the *Focal self-attention* to capture long-range spatial interactions.
- **CoAtNet** proposed an *optimized hybrid model comprising convolutional and transformer layers* in the earlier and later stages, respectively.
- **PiT** incorporates a *pooling layer* into ViT, and demonstrates advantageous outcomes of such approach.

Today, researchers pay more attention to **efficiency**, including efficient **self-attention**, **training strategy**, **pyramid design**, etc.

Global Context Vision Transformer – GC ViT [2022-06-20 NVIDIA]

The authors introduced a hierarchical ViT architecture that combines local and **global self-attention** modules, and a **Fused-MBConv** blocks. Such an approach eliminates sophisticated and computationally expensive operations and ensures the effectiveness of self-attention when applied to high-resolution images.

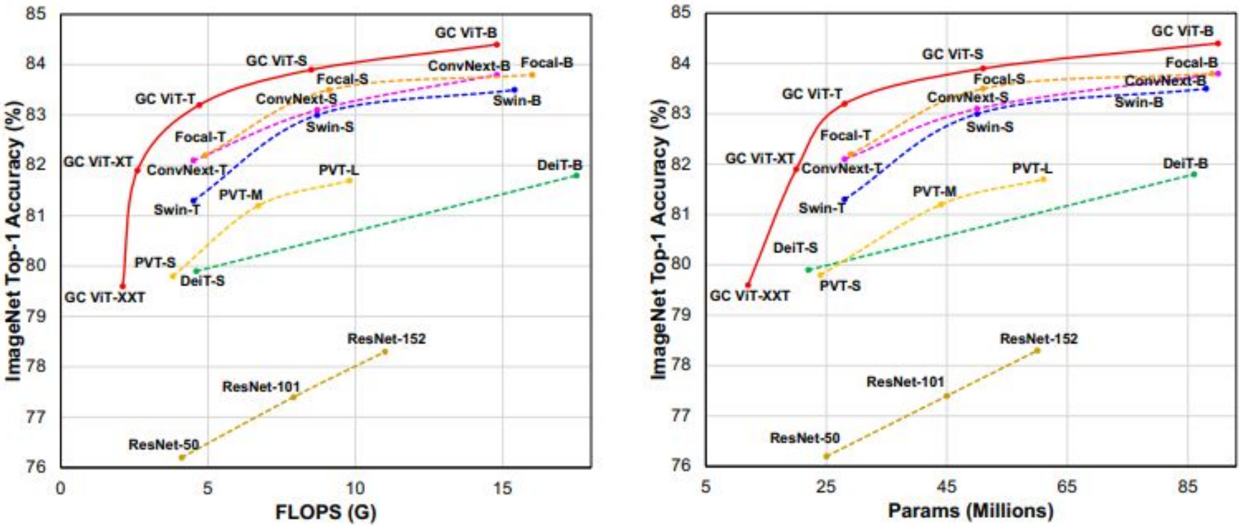


Figure 1 – Top-1 accuracy *vs.* model FLOPs/parameter size on ImageNet-1K dataset. GC ViT achieves new SOTA benchmarks for different model sizes as well as FLOPs, outperforming competing approaches by a significant margin.

GC ViT – Architecture

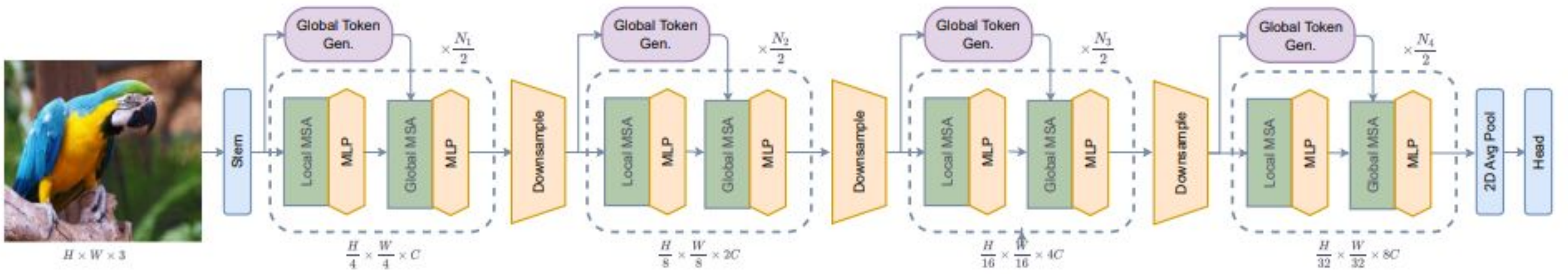


Figure 2 – Architecture of the proposed Global Context ViT. We use alternating blocks of local and global context self attention layers in each stage of the architecture.

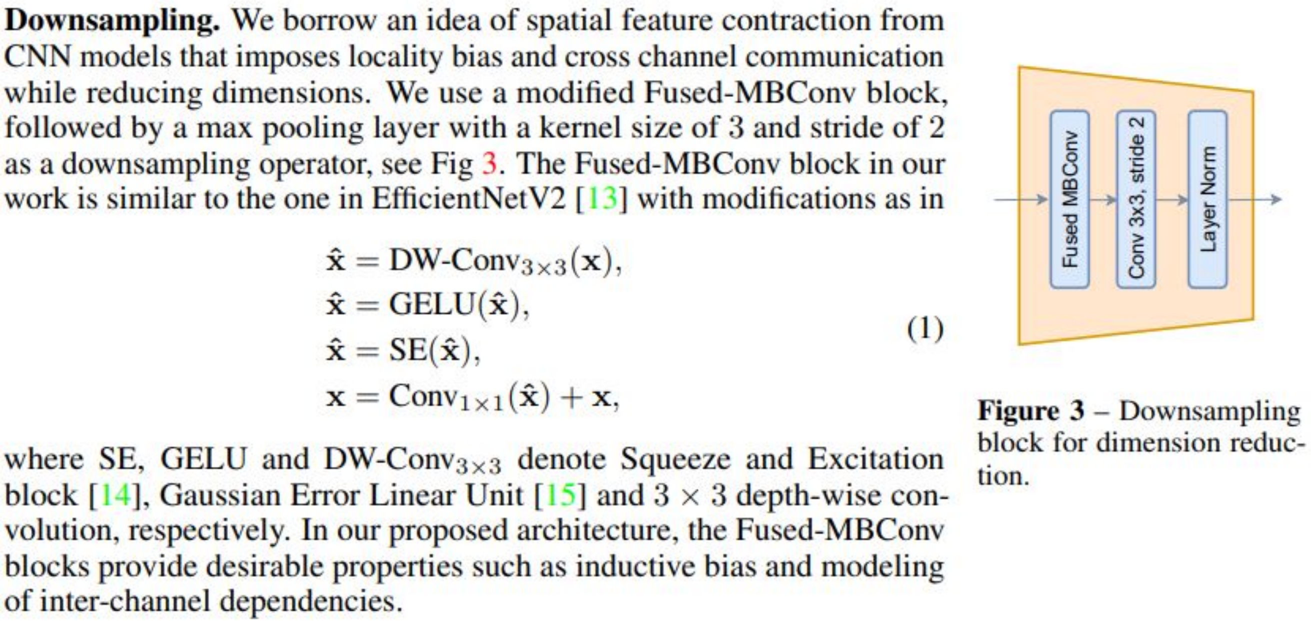
Authors use a hierarchical framework to obtain feature representations at **several resolutions** (aka “**stages**”) by decreasing the spatial dimensions while expanding the embedding dimension by factors of 2 and 2, respectively.

Each such stage is composed of alternating LSA and GSA modules to extract spatial features. Both operate in local windows like Swin Transformer, however, the GSA accesses global features extracted by **Global Token Generator** (GTG).

The GTG is a CNN-like module that extracts features from the entire image only once at every stage. The spatial resolution is decreased by 2 while increasing the number of channels by a downsampling block after every stage. The resulting features are passed through average pooling and linear layers to create an embedding for a downstream task.

GC ViT – Downsampling with Fused-MBConv

The GC ViT architecture relies on novel blocks such as a downsampling operator, a global query generator, and a global self-attention module described in the next sections.



Downsampling. We borrow an idea of spatial feature contraction from CNN models that imposes locality bias and cross channel communication while reducing dimensions. We use a modified Fused-MBConv block, followed by a max pooling layer with a kernel size of 3 and stride of 2 as a downsampling operator, see Fig 3. The Fused-MBConv block in our work is similar to the one in EfficientNetV2 [13] with modifications as in

$$\begin{aligned}\hat{\mathbf{x}} &= \text{DW-Conv}_{3\times3}(\mathbf{x}),\\ \hat{\mathbf{x}} &= \text{GELU}(\hat{\mathbf{x}}),\\ \hat{\mathbf{x}} &= \text{SE}(\hat{\mathbf{x}}),\\ \mathbf{x} &= \text{Conv}_{1\times1}(\hat{\mathbf{x}}) + \mathbf{x},\end{aligned} \tag{1}$$

where SE, GELU and DW-Conv$_{3\times3}$ denote Squeeze and Excitation block [14], Gaussian Error Linear Unit [15] and 3 × 3 depth-wise convolution, respectively. In our proposed architecture, the Fused-MBConv blocks provide desirable properties such as inductive bias and modeling of inter-channel dependencies.

Figure 3 – Downsampling block for dimension reduction.

GC ViT – “Attention is all you need”

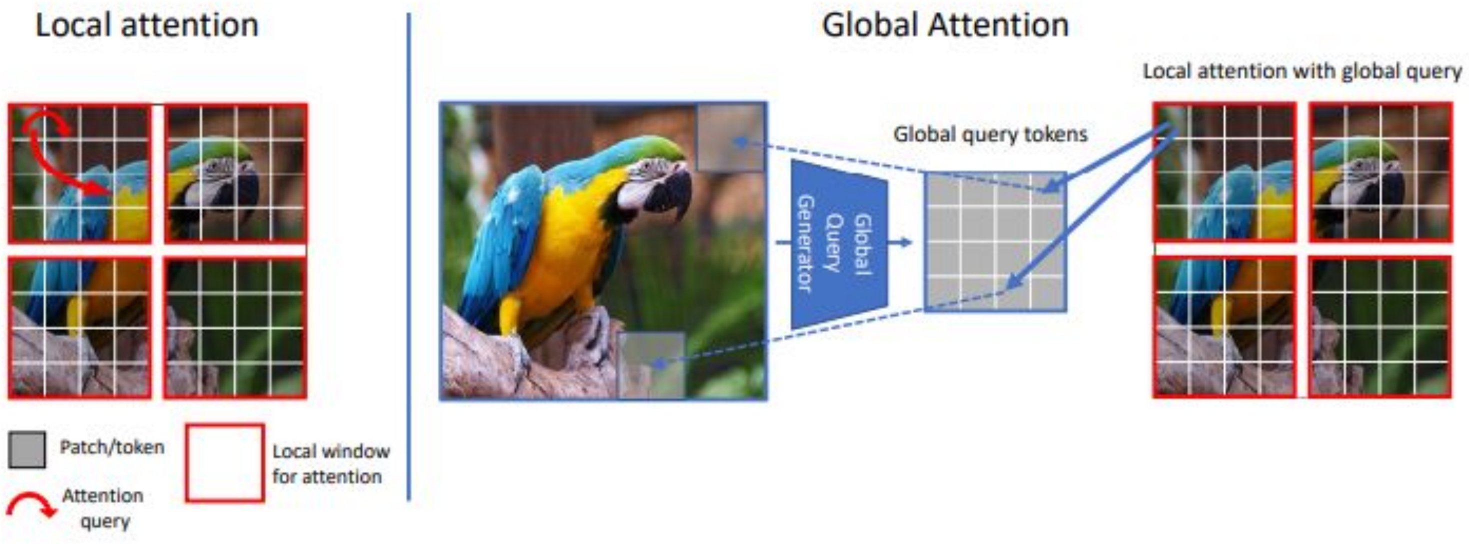


Figure 4 – Attention formulation. Local attention is computed on feature patches within local window only (left). On the other hand, the global features are extracted from the entire input features and then repeated to form global query tokens. The global query is interacted with local key and value tokens, hence allowing to capture long-range information via cross-region interaction.

GC ViT – “Attention is all you need”

- Multi-head self-attention (MSA) is the core computational operator in GC ViT to extract semantic information from the image. There are local and global self-attention modules in GC ViT.
- Similar to Swin Transformer, images are split into windows and LSA is performed within them. This leads to linear complexity scaling with image size. The LSA extracts local, short-range, information.
- In order to facilitate long-range dependencies, the authors proposed to use novel GSA to allow cross-patch communication with the ones situated outside the local window. GSA attends to other regions in the image via a global query token that represents image embedding extracted with GTG.

GC ViT – Global Self-Attention

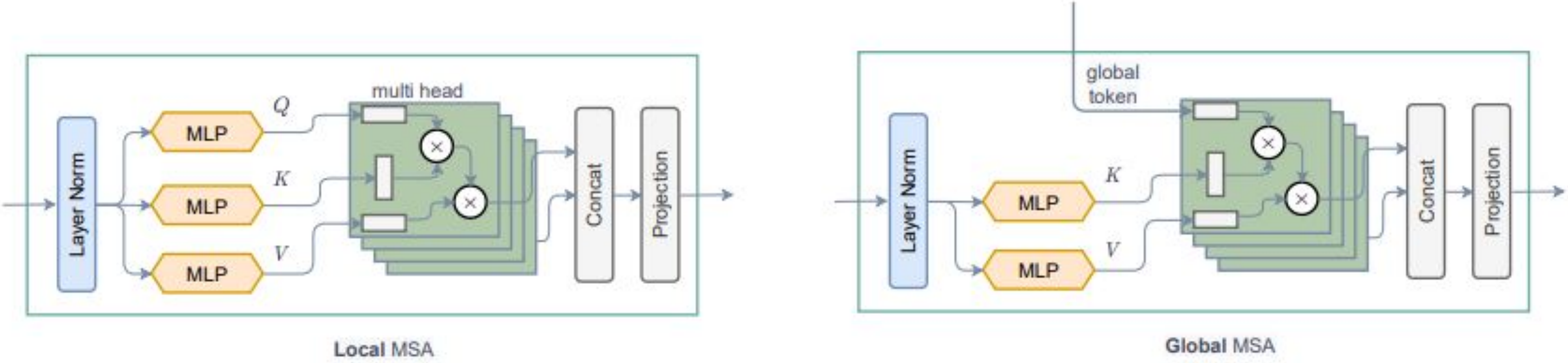


Figure 6 – Local and global attention blocks. Global attention block does not compute query vector and and reuses global query computed via Global Token Generation.

Incorporating local ***and*** global self-attention types is the main idea of our contribution. LSA can only query patches within a local window, whereas GSA can query image globally while still operating in the window. The only difference in implementation is that the query component is pre-computed in the case of global attention and the rest is the same as in local attention. In each stage, GC ViT employs alternating local and global self-attention blocks to effectively capture both local and global spatial information.

GC ViT – Global Query Generator

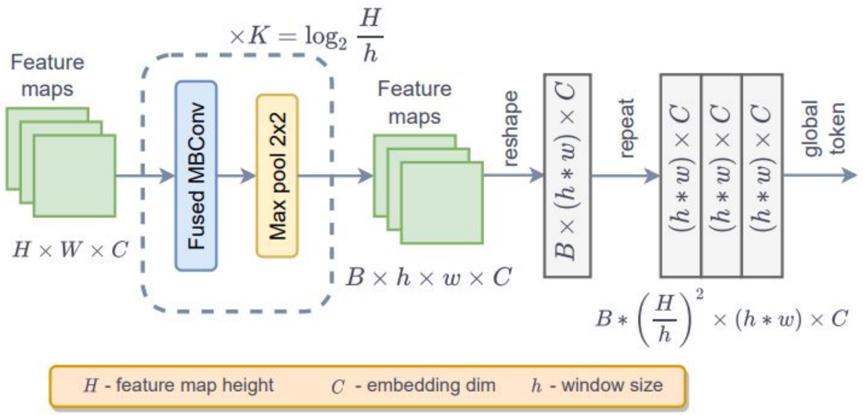


Figure 5 – Global Query Generator. Given feature map it computes a patch summarization obtained with imposing inductive bias via convolutional layers.

Such query tokens are computed once at every stage of the pipeline and shared across all global attention blocks, thus decreasing the number of parameters and FLOPs and improving the generalizability.

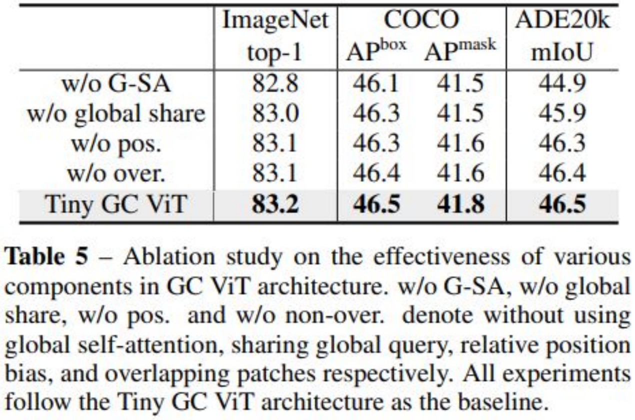
The authors propose to generate global query tokens that encompass information across the entire input feature maps for interaction with local key and value features. Specifically, as shown in figure, a layer in the generator consists of a Fused-MBConv block followed by a max pooling layer.

GC ViT – Performance, Results, and Ablations

GC ViT consistently outperforms both ConvNeXt and Swin Transformer models by a significant margin. GC ViT **base** (90M parameters), **small** (51M parameters), and **tiny** (28M parameters) achieve new SOTA results of **84.4%, 83.9%, and 83.2%** Top-1 accuracy on ImageNet-1k classification.

GC ViT demonstrates great scalability for high-resolution images in various downstream tasks, validating the effectiveness of the proposed framework in capturing both short and long-range information.

- ➢ **GSA** proved to be **vital** in the performance of classification, detection, and segmentation tasks.
- ➢ **Sharing global context query features** consistently **improves** the performance for all tasks.
- ➢ **Relative position bias** is also shown to **benefit** the pipeline.



	ImageNet	COCO		ADE20k
	top-1	AP^{box}	AP^{mask}	mIoU
w/o G-SA	82.8	46.1	41.5	44.9
w/o global share	83.0	46.3	41.5	45.9
w/o pos.	83.1	46.3	41.6	46.3
w/o over.	83.1	46.4	41.6	46.4
Tiny GC ViT	**83.2**	**46.5**	**41.8**	**46.5**

Table 5 – Ablation study on the effectiveness of various components in GC ViT architecture. w/o G-SA, w/o global share, w/o pos. and w/o non-over. denote without using global self-attention, sharing global query, relative position bias, and overlapping patches respectively. All experiments follow the Tiny GC ViT architecture as the baseline.

Next-ViT: Next Generation Vision Transformer for Efficient Deployment in Realistic Industrial Scenarios – [ByteDance 2022-07-13]

TL;DR

- ➢ Powerful **convolution block** and **transformer block** are introduced, i.e. NCB and NTB, with deployment-friendly mechanisms. Next-ViT stacks NCB and NTB to build advanced **CNN-Transformer hybrid architecture**.
- ➢ An innovative CNN-Transformer **hybrid strategy** is designed from a new insight that boosts **performance** with high **efficiency**.
- ➢ Next-ViT is presented with a multitude of variations. It achieves **SOTA latency/accuracy trade-off** on image classification, object detection, and semantic segmentation on TensorRT and CoreML.

Next-ViT – Architecture and Novelties

1) The authors introduce the **Next Convolution Block** (NCB), which is skilled at *capturing short-term dependency information* in visual data with a novel deployment-friendly **Multi-Head Convolutional Attention** (MHCA).

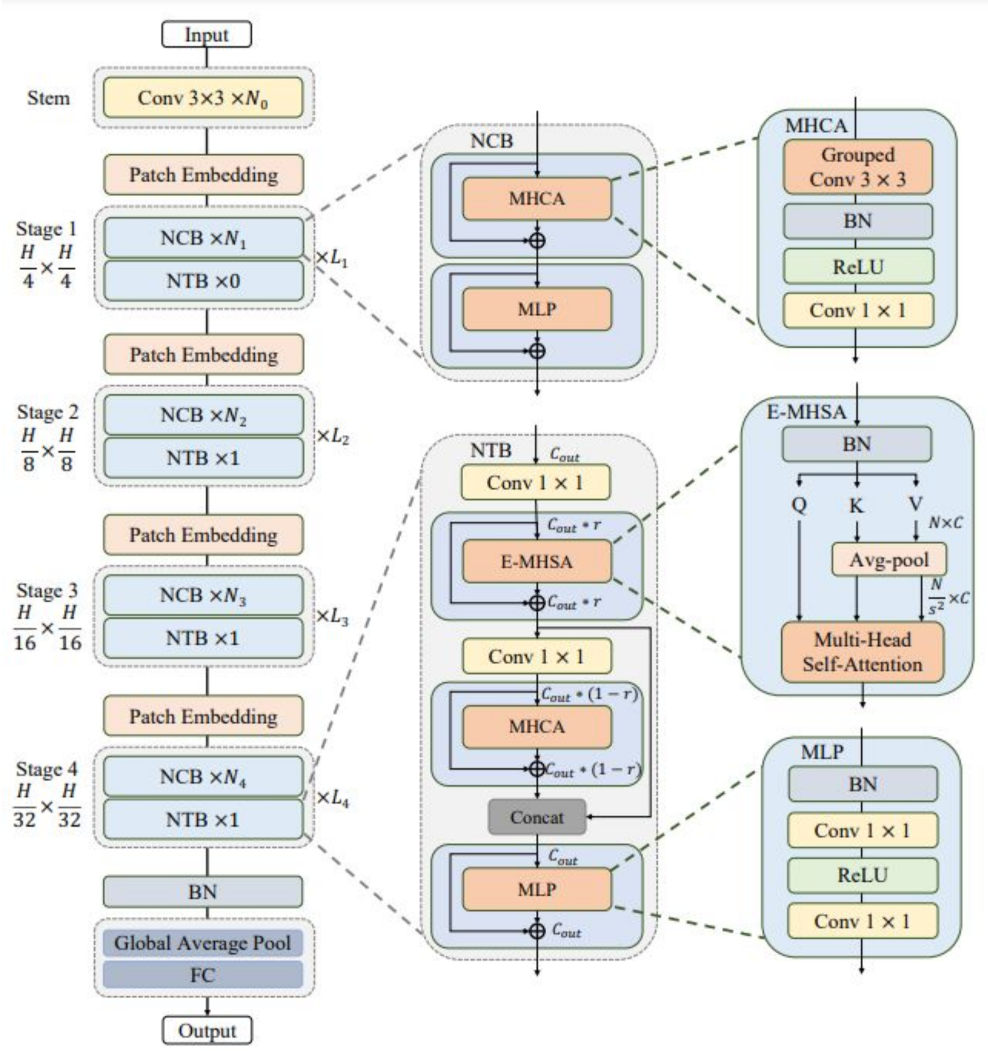
2) They build the **Next Transformer Block** (NTB), NTB is not only an expert in *capturing long-term dependency information* but also works as a *lightweight and high-and-low frequency signal mixer* to enhance modeling capability.

3) They also introduced the **Next Hybrid Strategy** (NHS) to stack NCB and NTB in a novel hybrid paradigm in each stage, which *greatly reduce the proportion of the transformer block* and *retaining the high precision* of the vision transformer network in various downstream tasks.

Next-ViT – Architecture and Novelties

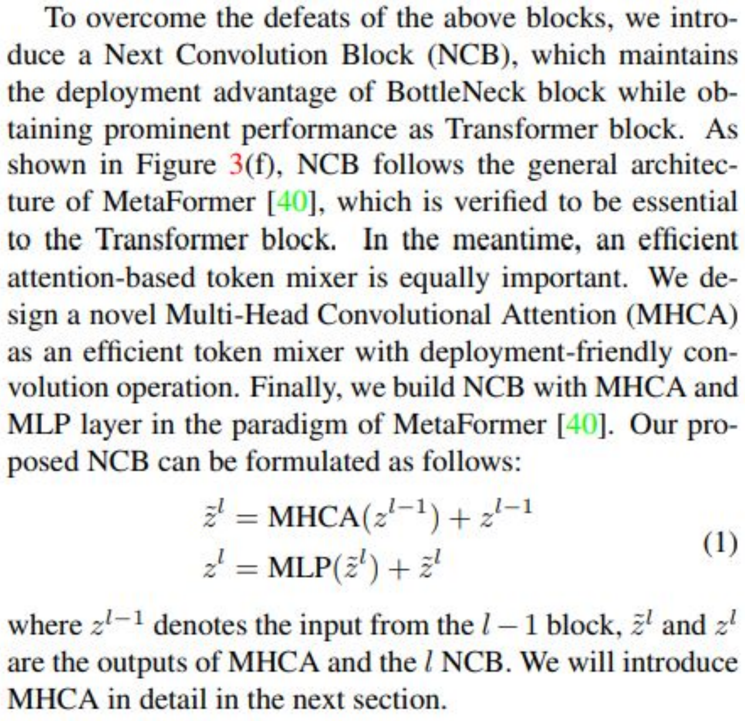
4-stage pipeline with details on ***Efficient Multi-Head Self-Attention*** (E-MHSA) in Transformer Block, ***Multi-Head Convolutional Attention*** (MHCA) in Convolutional Block.

Re-introduction of *Batch Norm* (BN) instead of traditional *Layer Norm* (LN), and *ReLU* instead of *GELU* in MLP.



Next-ViT – Next Convolution Block (NCB)

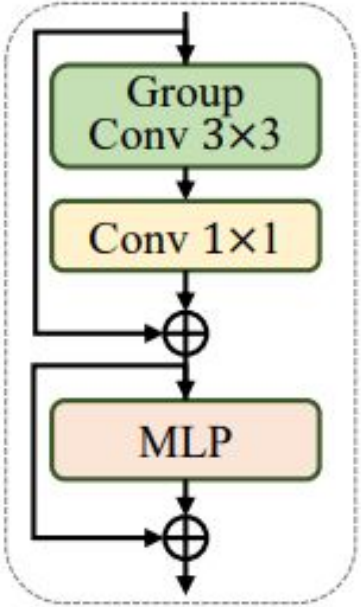
Blocks introduced in previous works, e.g. *Bottle-Neck block*, *ConvNeXt block*, *Transformer block*, all have their downsides that have to do with either attention computation complexity or too large depth-wise convolutions and thus severely limited inference speed.



To overcome the defeats of the above blocks, we introduce a Next Convolution Block (NCB), which maintains the deployment advantage of BottleNeck block while obtaining prominent performance as Transformer block. As shown in Figure 3(f), NCB follows the general architecture of MetaFormer [40], which is verified to be essential to the Transformer block. In the meantime, an efficient attention-based token mixer is equally important. We design a novel Multi-Head Convolutional Attention (MHCA) as an efficient token mixer with deployment-friendly convolution operation. Finally, we build NCB with MHCA and MLP layer in the paradigm of MetaFormer [40]. Our proposed NCB can be formulated as follows:

$$\begin{aligned} \tilde{z}^l &= \mathrm{MHCA}(z^{l-1}) + z^{l-1} \\ z^l &= \mathrm{MLP}(\tilde{z}^l) + \tilde{z}^l \end{aligned} \tag{1}$$

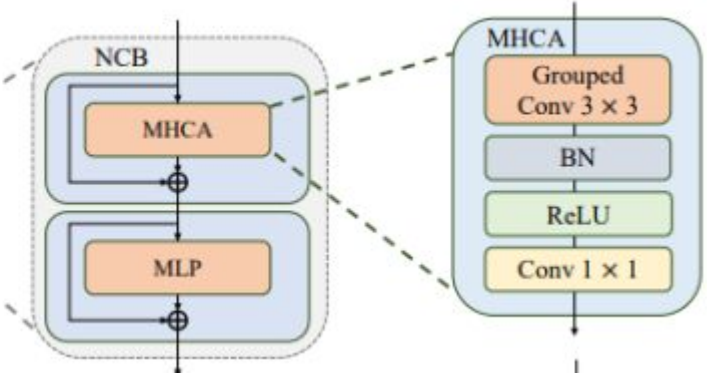
where z^{l-1} denotes the input from the $l-1$ block, $\tilde{z}^l$ and z^l are the outputs of MHCA and the l NCB. We will introduce MHCA in detail in the next section.



(f) NCB (ours)

Next-ViT – Multi-Head Convolutional Attention (MHCA)

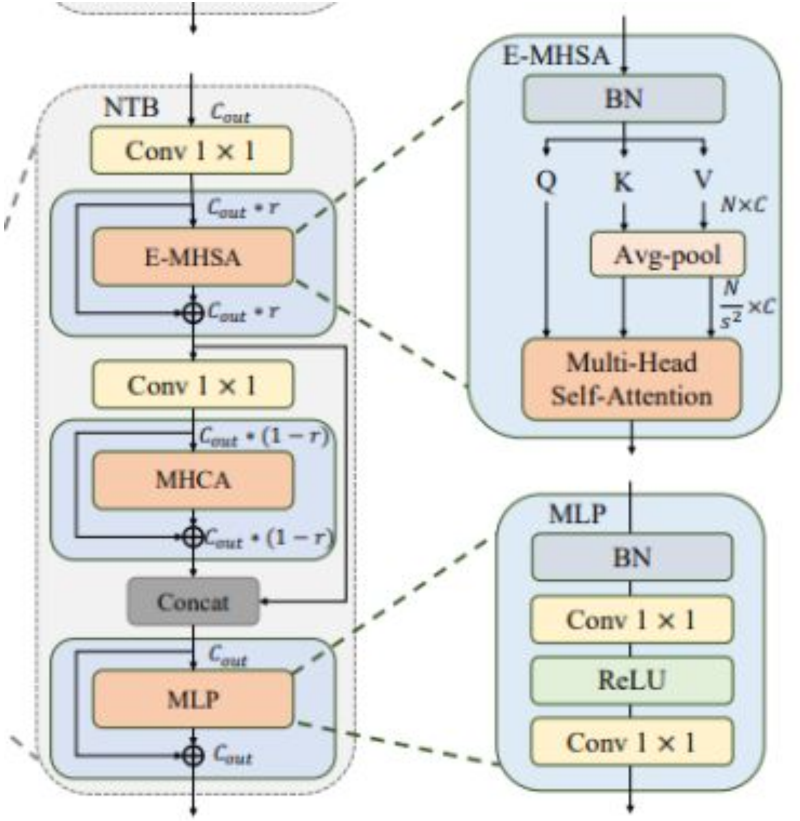
To free the existing attention-based token mixer from the high latency dilemma, the authors design a **novel attention mechanism with efficient convolution operation**, i.e. **Convolutional Attention** (CA), for fast inference speed. In the meantime, inspired by the effective multi-head design in MHSA, they build their **convolutional attention with a multi-head paradigm** that jointly attends to information from different representation subspaces at a different positions for effective local representation learning.



The implementation of MHCA is carried out with a **group convolution** (multi-head convolution) and a **point-wise convolution**. Additionally, they adopt an efficient **BatchNorm** (BN) and **ReLU** activation function in NCB rather than **LayerNorm** (LN) and **GELU** in traditional Transformer blocks, which further accelerates inference speed. Experimental results in the ablation study show the superiority of NCB compared with existing blocks.

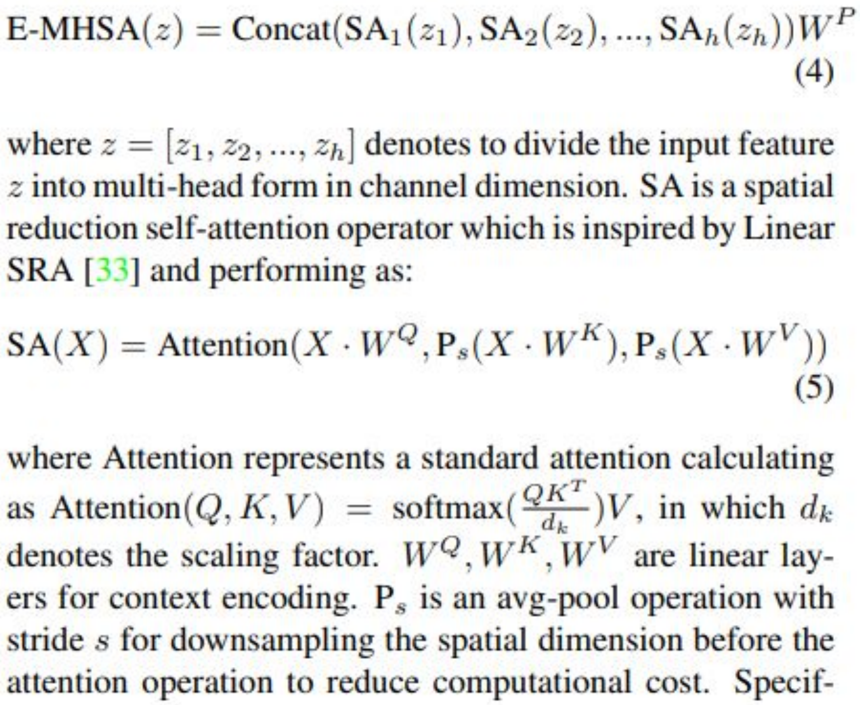
Next-ViT – Next Transformer Block (NTB)

NCB effectively captures the local representations, but the global information is yet to be addressed. The authors develop a **Next Transformer Block** (NTB) to capture multi-frequency signals in the lightweight mechanism. Furthermore, NTB works as an effective multi-frequency signals mixer to further enhance the overall modeling capability.



Next-ViT – Efficient Multi-Head Self Attention (E-MHSA)

NTB captures low-frequency signals with an E-MHSA. Batch Normalization is also utilized in the E-MHSA module for extremely efficient deployment.

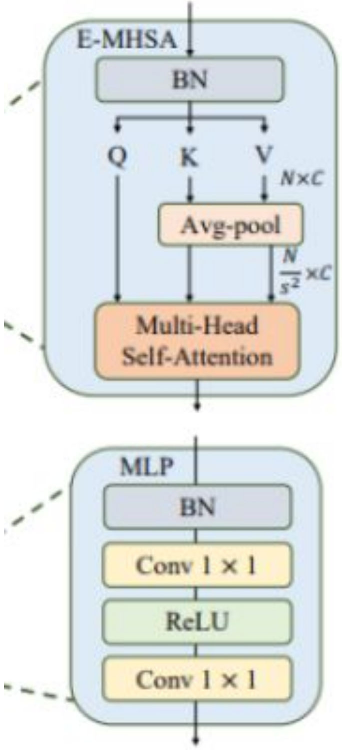


$$\text{E-MHSA}(z) = \text{Concat}(\text{SA}_1(z_1), \text{SA}_2(z_2), ..., \text{SA}_h(z_h))W^P \tag{4}$$

where $z = [z_1, z_2, ..., z_h]$ denotes to divide the input feature z into multi-head form in channel dimension. SA is a spatial reduction self-attention operator which is inspired by Linear SRA [33] and performing as:

$$\text{SA}(X) = \text{Attention}(X \cdot W^Q, \text{P}_s(X \cdot W^K), \text{P}_s(X \cdot W^V)) \tag{5}$$

where Attention represents a standard attention calculating as $\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{d_k})V$, in which d_k denotes the scaling factor. W^Q, W^K, W^V are linear layers for context encoding. P_s is an avg-pool operation with stride s for downsampling the spatial dimension before the attention operation to reduce computational cost. Specif-



Next-ViT – Next Hybrid Strategy (NHS)

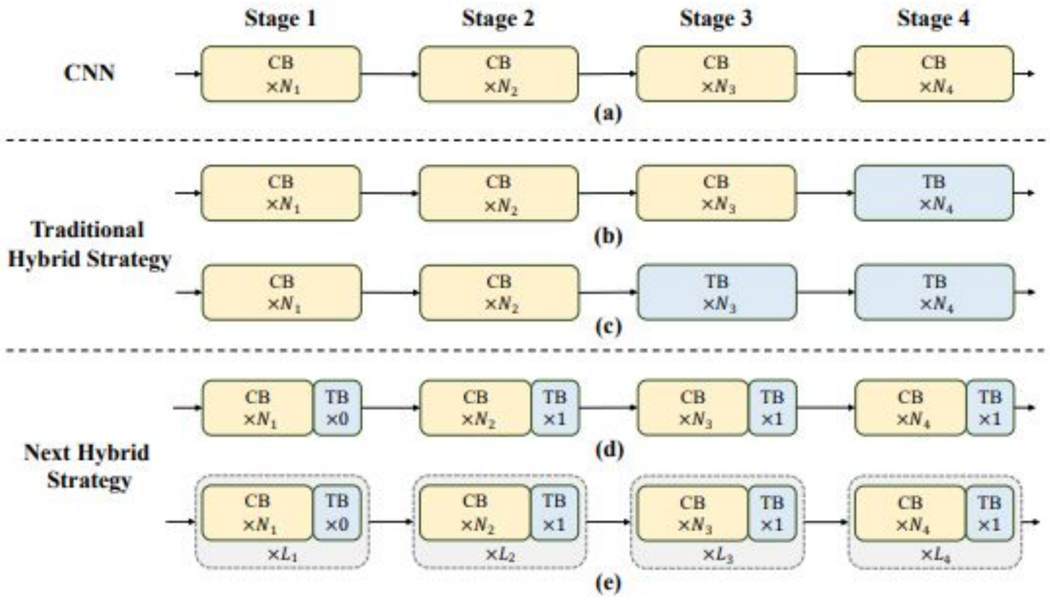


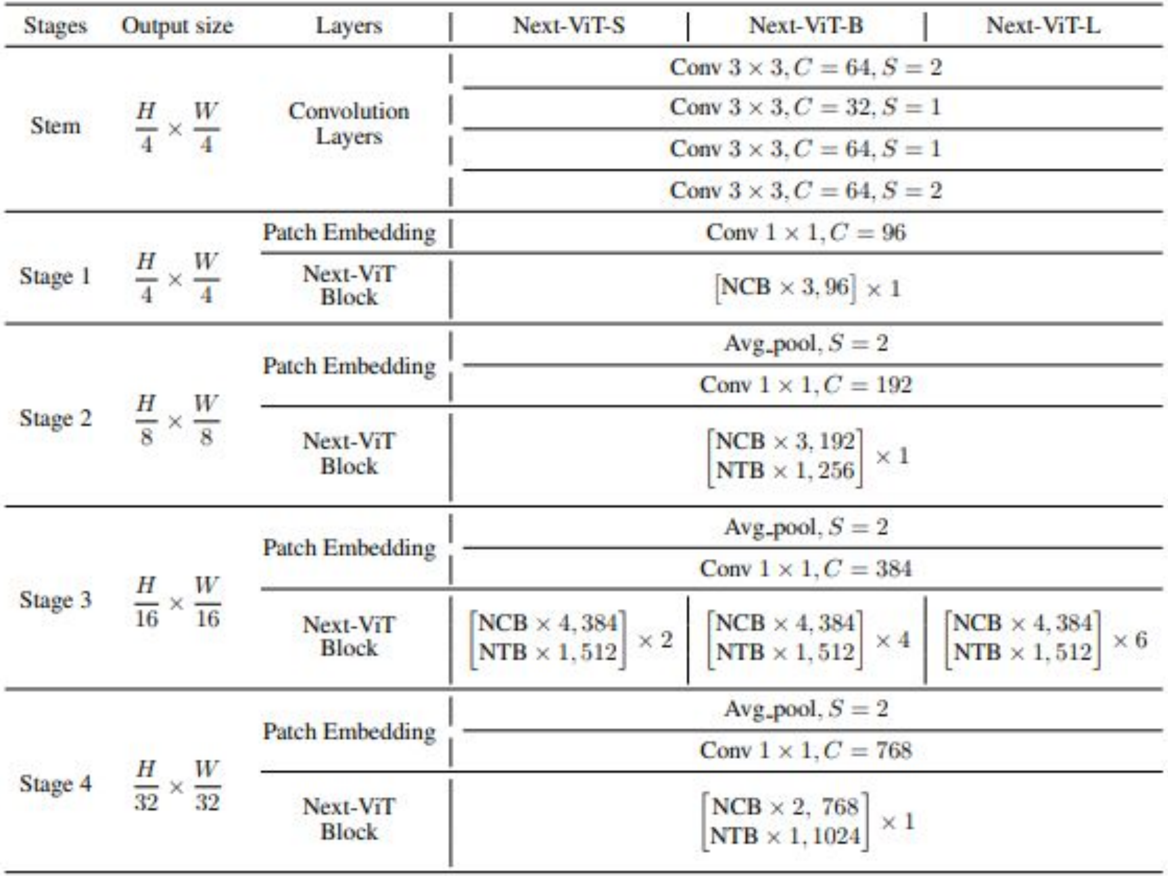
Figure 4. Comparison of traditional hybrid strategies and NHS.

To address the performance of existing hybrid methods on downstream tasks, the authors propose a **Next Hybrid Strategy** (NHS) from a new insight, which creatively stacks NCB and NTB with $(N + 1) * L$ hybrid paradigm. NHS significantly promotes model performance in downstream tasks by controlling the proportion of Transformer block for efficient deployment

The authors present a **novel hybrid strategy** in (NCB × N + NTB × 1) pattern, which sequentially stacks N NCB and one NTB in each stage.

Next-ViT – Configuration variants

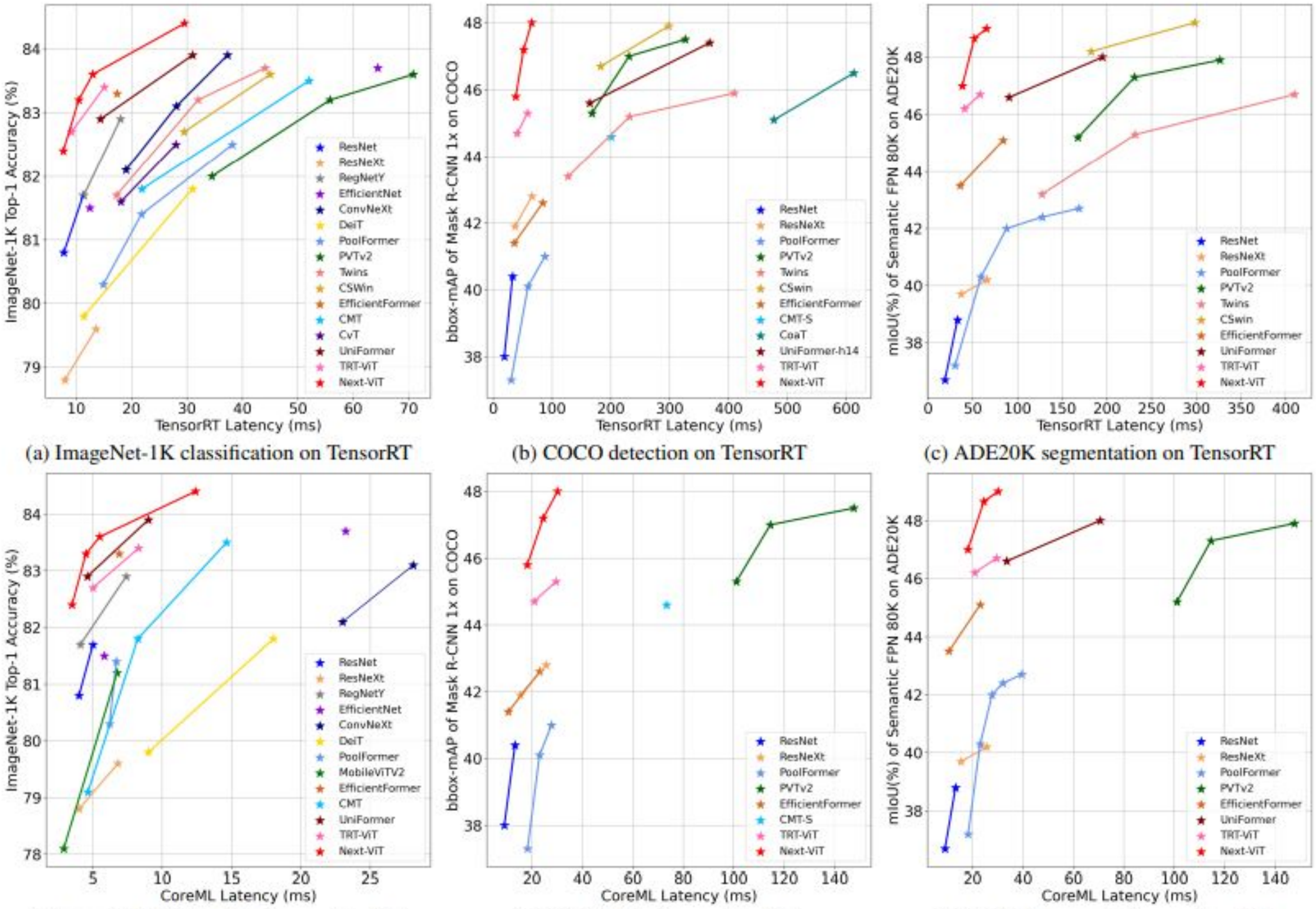
Table 3. Detailed configurations of Next-ViT variants.



Stages	Output size	Layers	Next-ViT-S	Next-ViT-B	Next-ViT-L
Stem	$\frac{H}{4} \times \frac{W}{4}$	Convolution Layers	Conv $3 \times 3, C = 64, S = 2$		
			Conv $3 \times 3, C = 32, S = 1$		
			Conv $3 \times 3, C = 64, S = 1$		
			Conv $3 \times 3, C = 64, S = 2$		
Stage 1	$\frac{H}{4} \times \frac{W}{4}$	Patch Embedding	Conv $1 \times 1, C = 96$		
		Next-ViT Block	$\begin{bmatrix} \text{NCB} \times 3, 96 \end{bmatrix} \times 1$		
Stage 2	$\frac{H}{8} \times \frac{W}{8}$	Patch Embedding	Avg_pool, $S = 2$		
			Conv $1 \times 1, C = 192$		
		Next-ViT Block	$\begin{bmatrix} \text{NCB} \times 3, 192 \\ \text{NTB} \times 1, 256 \end{bmatrix} \times 1$		
Stage 3	$\frac{H}{16} \times \frac{W}{16}$	Patch Embedding	Avg_pool, $S = 2$		
			Conv $1 \times 1, C = 384$		
		Next-ViT Block	$\begin{bmatrix} \text{NCB} \times 4, 384 \\ \text{NTB} \times 1, 512 \end{bmatrix} \times 2$	$\begin{bmatrix} \text{NCB} \times 4, 384 \\ \text{NTB} \times 1, 512 \end{bmatrix} \times 4$	$\begin{bmatrix} \text{NCB} \times 4, 384 \\ \text{NTB} \times 1, 512 \end{bmatrix} \times 6$
Stage 4	$\frac{H}{32} \times \frac{W}{32}$	Patch Embedding	Avg_pool, $S = 2$		
			Conv $1 \times 1, C = 768$		
		Next-ViT Block	$\begin{bmatrix} \text{NCB} \times 2,\ 768 \\ \text{NTB} \times 1, 1024 \end{bmatrix} \times 1$		

Next-ViT – Performance and Impact

Next-ViT shows a more **significant latency/accuracy trade-off superiority** on downstream tasks. This work builds a stable bridge between *academic research and industrial deployment* in terms of visual neural network design.



(a) ImageNet-1K classification on TensorRT

(b) COCO detection on TensorRT

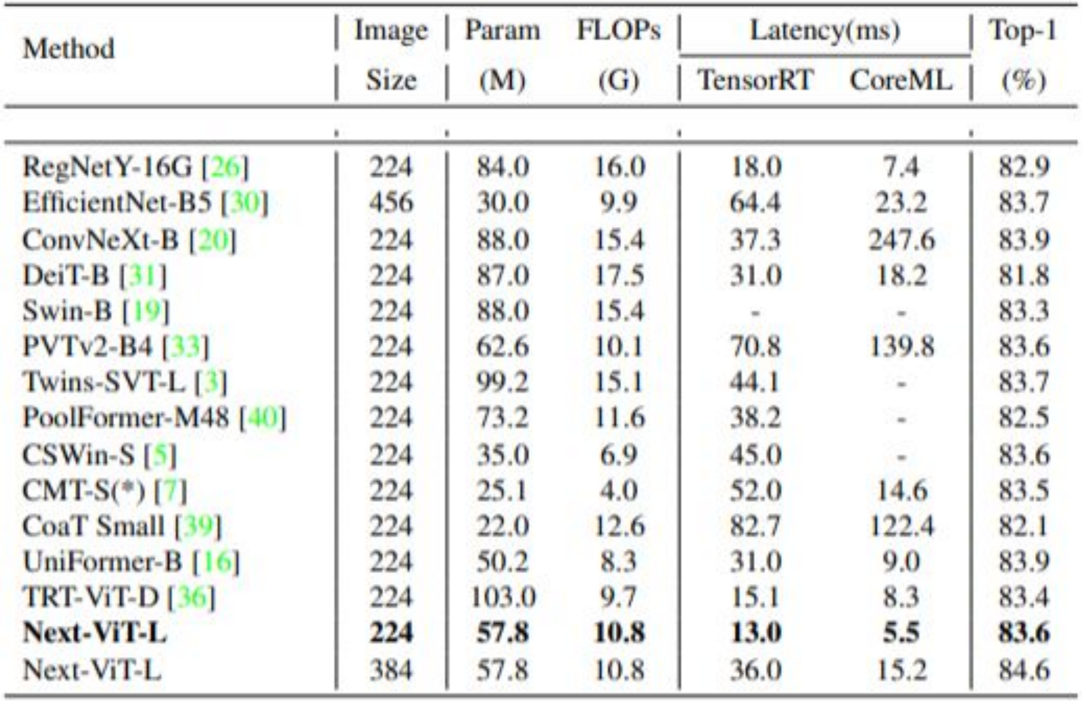
(c) ADE20K segmentation on TensorRT

(d) ImageNet-1K classification on CoreML

(e) COCO detection on CoreML

(f) ADE20K segmentation on CoreML

Next-ViT – Benchmarking



Method	Image	Param	FLOPs	Latency(ms)		Top-1
	Size	(M)	(G)	TensorRT	CoreML	(%)
RegNetY-16G [26]	224	84.0	16.0	18.0	7.4	82.9
EfficientNet-B5 [30]	456	30.0	9.9	64.4	23.2	83.7
ConvNeXt-B [20]	224	88.0	15.4	37.3	247.6	83.9
DeiT-B [31]	224	87.0	17.5	31.0	18.2	81.8
Swin-B [19]	224	88.0	15.4	-	-	83.3
PVTv2-B4 [33]	224	62.6	10.1	70.8	139.8	83.6
Twins-SVT-L [3]	224	99.2	15.1	44.1	-	83.7
PoolFormer-M48 [40]	224	73.2	11.6	38.2	-	82.5
CSWin-S [5]	224	35.0	6.9	45.0	-	83.6
CMT-S(*) [7]	224	25.1	4.0	52.0	14.6	83.5
CoaT Small [39]	224	22.0	12.6	82.7	122.4	82.1
UniFormer-B [16]	224	50.2	8.3	31.0	9.0	83.9
TRT-ViT-D [36]	224	103.0	9.7	15.1	8.3	83.4
Next-ViT-L	**224**	**57.8**	**10.8**	**13.0**	**5.5**	**83.6**
Next-ViT-L	384	57.8	10.8	36.0	15.2	84.6

TensorRT and CoreML benchmark comparison on ImageNet-1k.

Summary and future provisions

- The field of Vision Transformers is growing rapidly as many researchers dedicate time to hone, fine-tune, and even novelize these recently emerged yet very powerful architectures.
- Introduction of optimized hybrid models that balance the usage of both Convolutional and Transformer blocks gives chances to finally productionalize ViTs in industry.
- More lower-level details are getting addressed, namely the variants of self-attention, norm layers and activation functions are rethought, completely new modules get devised.
- Hybrid ViTs are without a doubt what State-of-the-Art situation will look like in the nearest future. Therefore, we can expect to get even more novelties shortly, while this rather fascinating area gets explored.

Reference literature

1. Vision Transformer paper – https://arxiv.org/abs/2010.11929v2.
2. GC ViT paper – https://arxiv.org/pdf/2206.09959.pdf.
3. GC ViT official implementation – https://github.com/NVlabs/GCViT.
4. Next-ViT paper – https://arxiv.org/pdf/2207.05501.pdf.
5. Video Swin – https://arxiv.org/abs/2106.13230.

Q/A Section

Thank you for your time!