GLOBAL CONTEXT VISION TRANSFORMERS

ALI HATAMIZADEH, HONGXU (DANNY) YIN, GREG HEINRICH, JAN KAUTZ, PAVLO MOLCHANOV

MOTIVATION

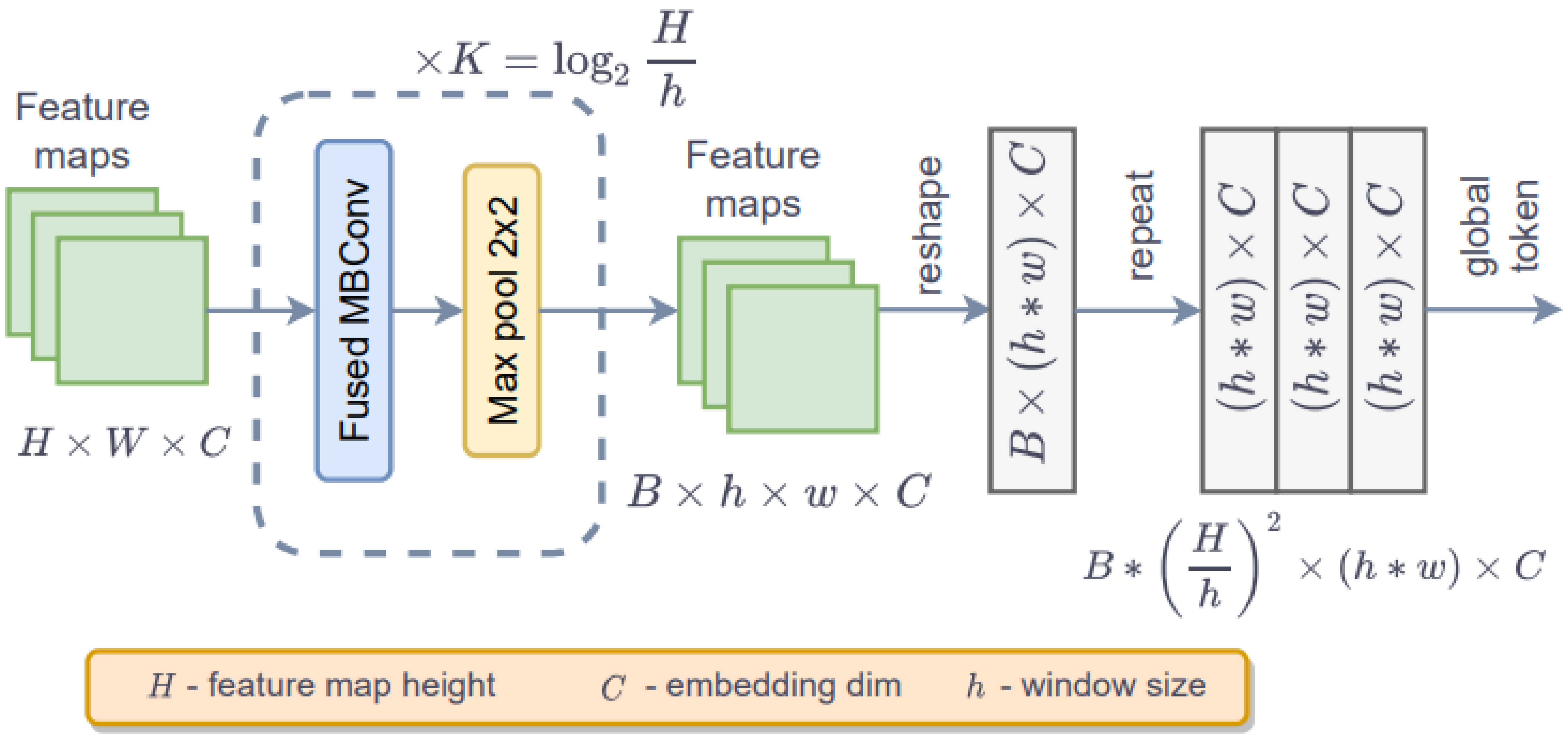
Can we model global context more efficient ?

- Our goal is to create a new transformer model which can accurately model both local and global information without imposing computational constrains.
- We propose Global Context (GC) Vision Transformers which models both long and short-range spatial interactions, without the need for expensive operations such as computing attention masks or shifting local windows.
- Every GC ViT stage is composed of alternating local and global self-attention modules to extract spatial features. Both operate in local windows like Swin Transformer.

METHODOLOGY

Global Token Generator

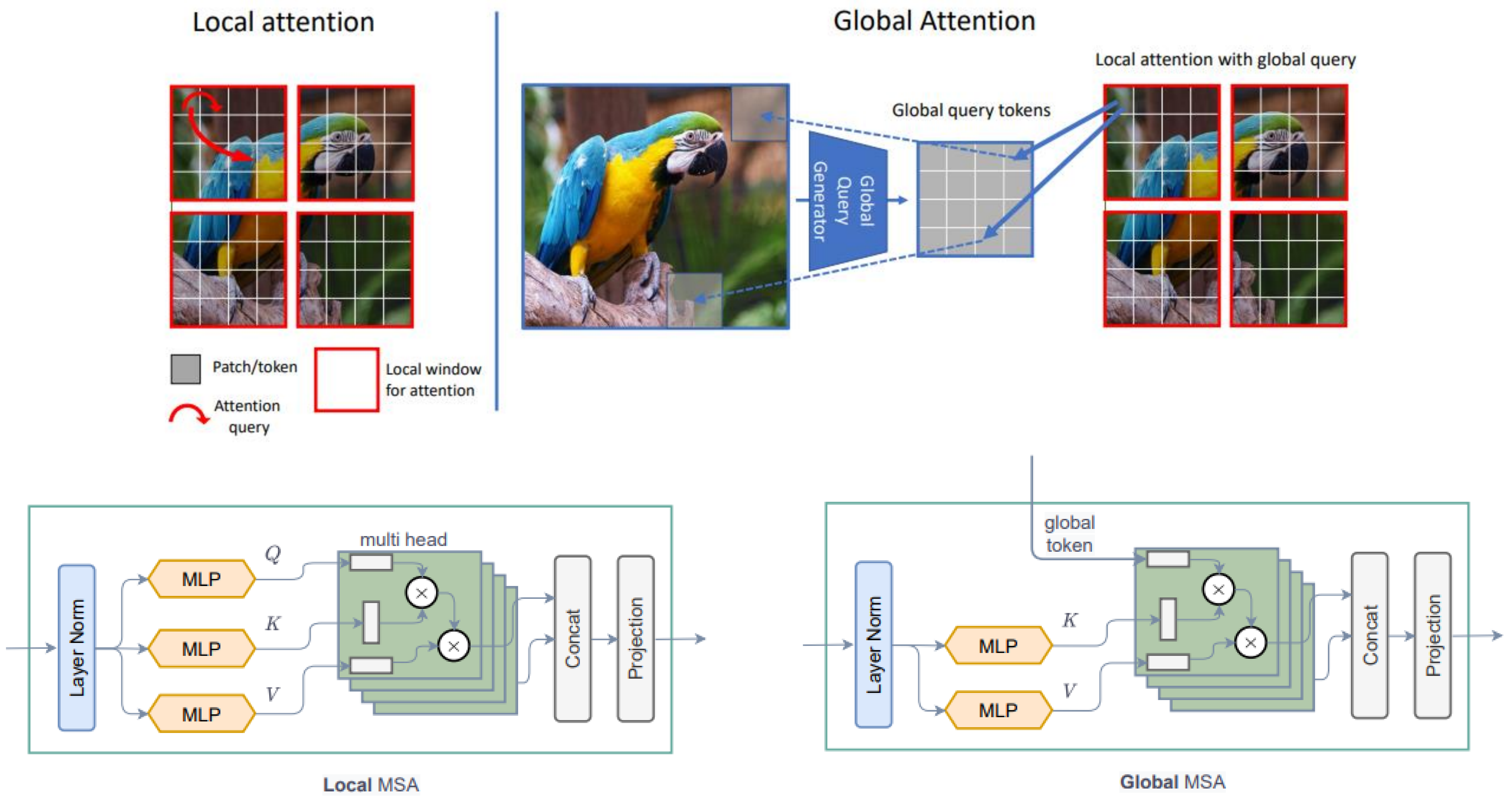
- We propose to generate global query tokens that encompass information across the entire input feature maps for interaction with local key and value features.



METHODOLOGY

Global Attention

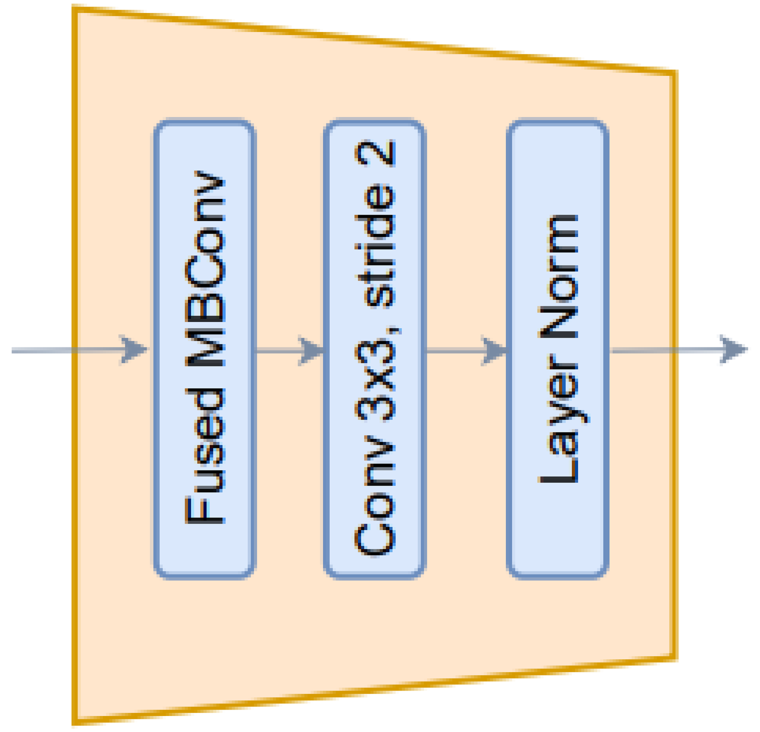
- Local self-attention can only query patches within a local window, whereas with global attention can query image globally while still operating in the window



METHODOLOGY

Downsampling

- We borrow an idea of spatial feature contraction from CNN models that imposes locality bias and cross channel communication while reducing dimensions.



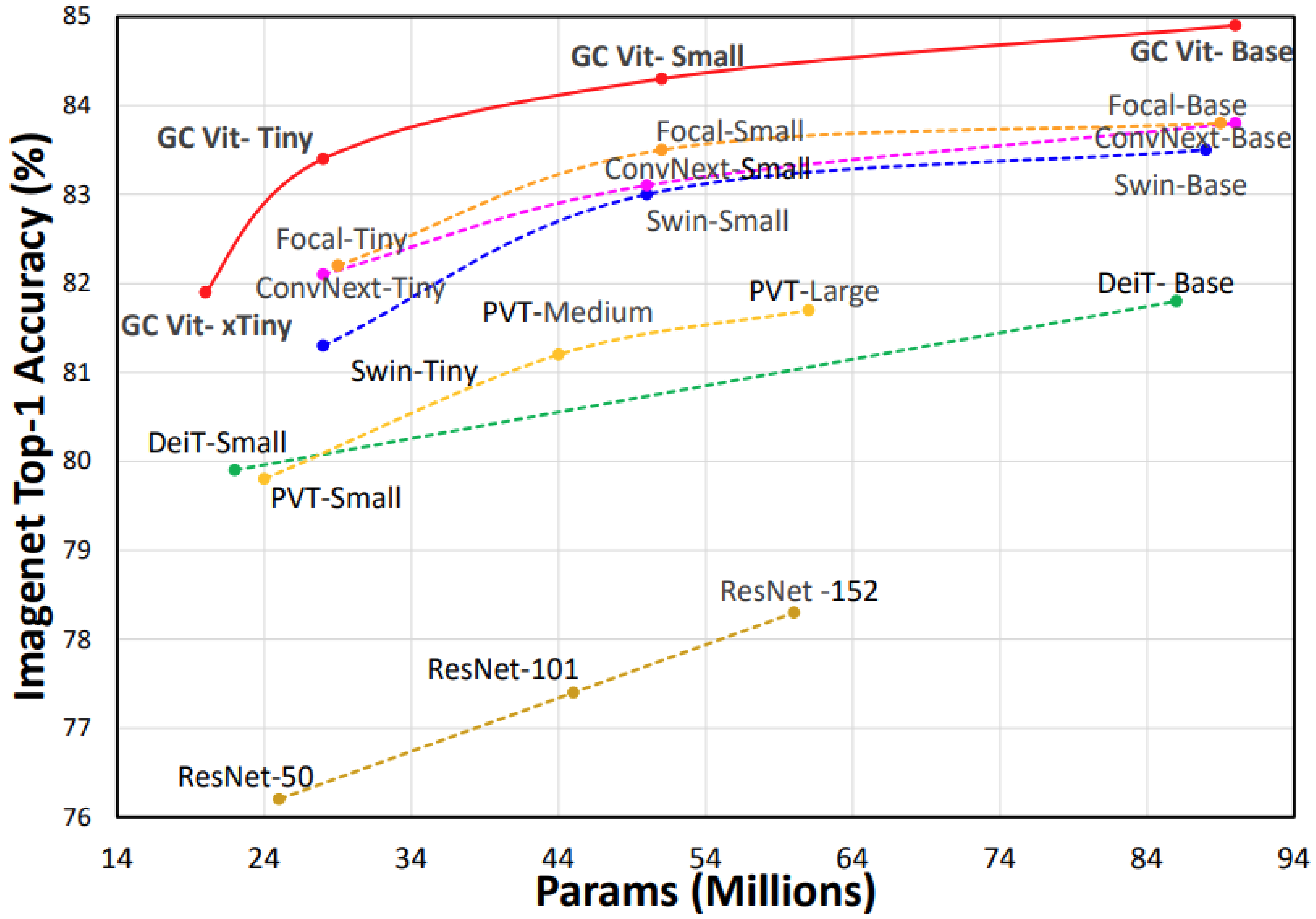
- We use a modified Fused-MBConv block, followed by a max pooling layer with a kernel size of 3 and stride of 2 as a downsampling operator according to:

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

EXPERIMENTS

ImageNet-1K Classifcation

- Our model achieves new SOTA benchmarks for accuracy vs number of parameters/FLOPs tradeoff.



EXPERIMENTS

MS COCO Detection/Instance Segmentation

- Models with GC ViT backbones archive strong performance for object detection and instance segmentation on MS COCO dataset.

Backbone	Param (M)	FLOPs (G)	AP^{box}	AP^{box}_{50}	AP^{box}_{75}	AP^{mask}	AP^{mask}_{50}	AP^{mask}_{75}
Mask-RCNN 3× schedule								
Swin-T (Liu et al., 2021)	48	267	46.0	68.1	50.3	41.6	65.1	44.9
ConvNeXt-T (Liu et al., 2022b)	48	262	46.2	67.9	50.8	41.7	65.0	44.9
GC ViT-T	48	291	**47.9**	**70.1**	**52.8**	**43.2**	**67.0**	**46.7**
Cascade Mask-RCNN 3× schedule								
DeiT-Small/16 (Touvron et al., 2021)	80	889	48.0	67.2	51.7	41.4	64.2	44.3
ResNet-50 (He et al., 2016)	82	739	46.3	64.3	50.5	40.1	61.7	43.4
Swin-T (Liu et al., 2021)	86	745	50.4	69.2	54.7	43.7	66.6	47.3
ConvNeXt-T (Liu et al., 2022b)	86	741	50.4	69.1	54.8	43.7	66.5	47.3
GC ViT-T	85	770	**51.6**	**70.4**	**56.1**	**44.6**	**67.8**	**48.3**
X101-32 (Xie et al., 2017)	101	819	48.1	66.5	52.4	41.6	63.9	45.2
Swin-S (Liu et al., 2021)	107	838	51.9	70.7	56.3	45.0	68.2	48.8
ConvNeXt-S (Liu et al., 2022b)	108	827	51.9	70.8	56.5	45.0	68.4	49.1
GC ViT-S	108	866	**52.4**	**71.0**	**57.1**	**45.4**	**68.5**	**49.3**
X101-64 (Xie et al., 2017)	140	972	48.3	66.4	52.3	41.7	64.0	45.1
Swin-B (Liu et al., 2021)	145	982	51.9	70.5	56.4	45.0	68.1	48.9
ConvNeXt-B (Liu et al., 2022b)	146	964	52.7	71.3	57.2	45.6	68.9	49.5
GC ViT-B	146	1018	**52.9**	**71.7**	**57.8**	**45.8**	**69.2**	**49.8**

EXPERIMENTS

Interpretability

- The associated feature maps uncovered by the global self-attention modules align with image semantics.

(a) Original images from ImageNet-1K validation set.

(b) **Global attention** maps from GC ViT model (ours).

(c) Corresponding **Grad-CAM** maps.

CONCLUSION

- In this work, we introduced a novel hierarchical ViT, referred to as GC ViT, which can efficiently capture global context by utilizing global query tokens and interact with local regions.
- We achieve new SOTA benchmarks for image classification across various model sizes on ImageNet-1K dataset,and surpasses both CNN and ViT-based counterparts by a significant margin.
- We have also achieved SOTA or competitive performance for downstream tasks of obect detection and instance segmentation on high-resolution images using MS COCO datasets.
- Code and pre-trained models are available:

https://github.com/NVlabs/GCVit

NVIDIA.


NVIDIA