LcT: Locally-Enhanced Cross-Window Vision Transformer

Canhui Wei, Huiwei Wang

Southwest University, China

Introduction

Previous works

- Vision Transformer's (ViT)[1] computational complexity is quadratic to feature size, which makes it unsuitable for processing high-resolution images or hierarchical representations. Here, the feature map contains $h \times w$ patches, and C is the attention dimension.

$$\Omega(\mathrm{MSA}) = 4hwC^2 + 2(hw)^2C$$

- MobileViT[2] applies cross-window multi-headed self-attention (CW-MSA) to constrain the attention region. However, convolution with large kernels for aggregating local information also leads to too much computation cost. Here, M is the window size.

$$\Omega(\mathrm{CW\_MSA}) = 4hwC^2 + 2N(hw)C$$
$$N = hw/M^2$$

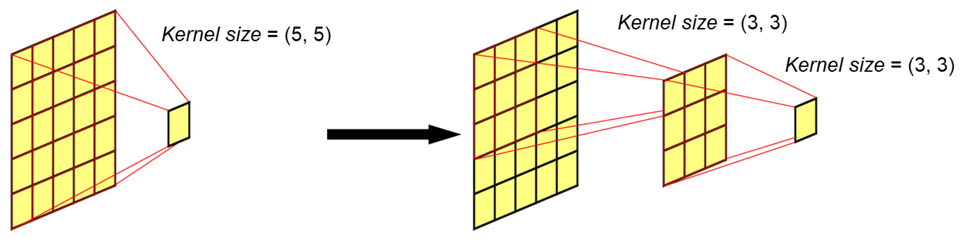
Existing problem

How aggregate information within a window in a time-efficient way?

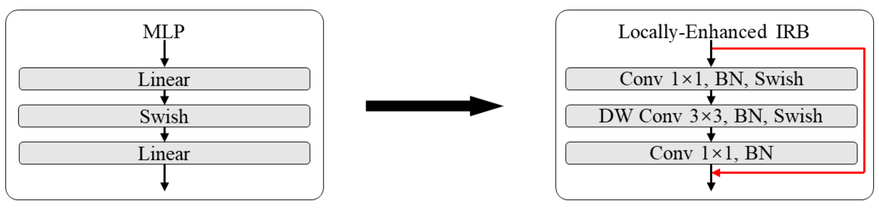
Methods

Proposed solutions

- Replacing large convolutional kernels with multiple small ones to obtain the equivalent receptive field without increasing the computational complexity.

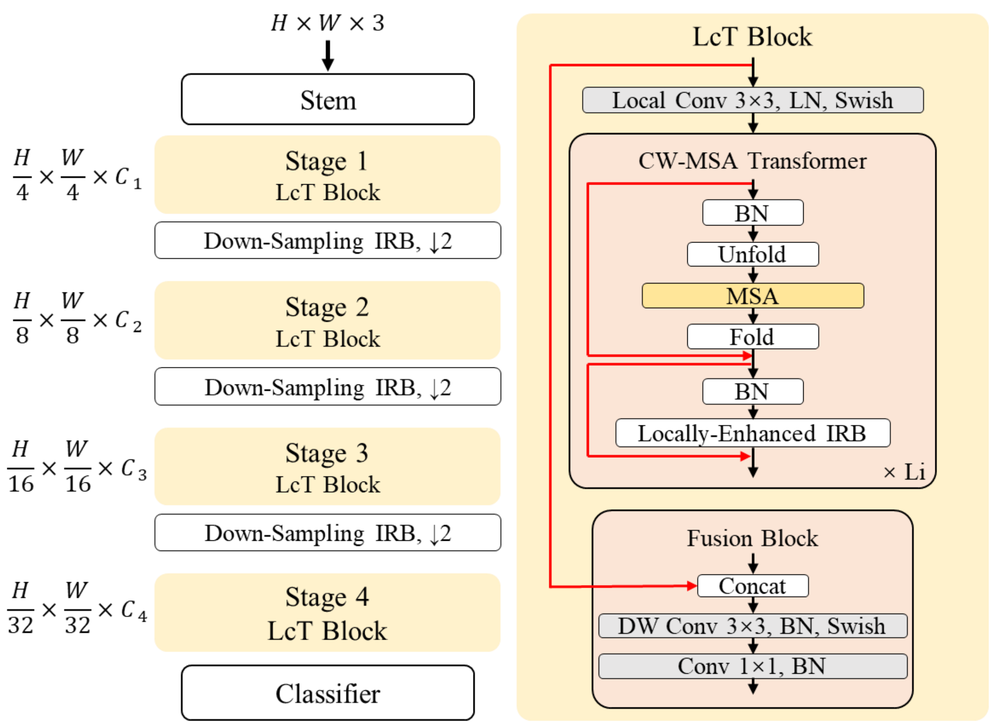


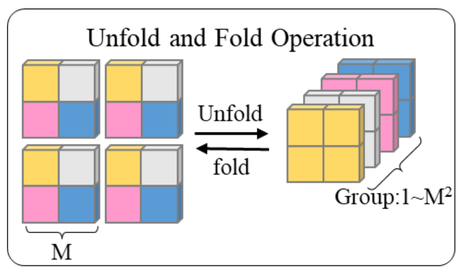
- Utilizing locally-enhanced inverted residual blocks (Locally-Enhanced IRB) instead of multi-layer perceptron (MLP) to learn local representations.



Model

- Replacing layer normalization with batch normalization to improve the inference speed without compromising performance.
- Stacking locally-enhanced inverted residual blocks to make each patch fully encode information with other patches.





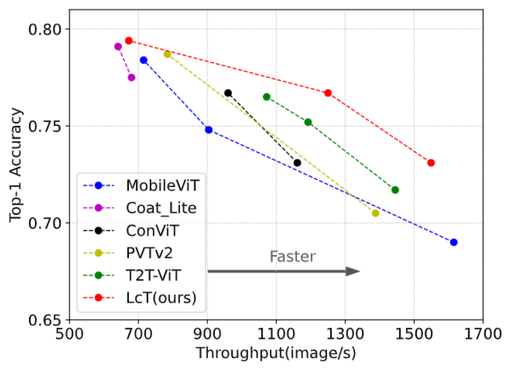
The unfold operation extracts patches from different windows at the corresponding locations. The fold operation restores all patches to the complete feature map according to their original positions.

Results

Classification

Model	Input	#param.	FLOPs	Throughput (images/s)	Top-1
MobileViT-XXS	256^2	1.3M	0.4G	1615	69.0
PVTv2-B0	224^2	3.4M	0.6G	1388	70.5
T2T-ViT-7	224^2	4.3M	1.1G	1445	71.7
LcT-Tiny(ours)	224^2	1.9M	0.6G	1549	**73.1**
MobileViT-XS	256^2	2.3M	1.1G	904	74.8
ConViT-Tiny	224^2	6.0M	1.0G	1161	73.1
ConViT-Tiny+	224^2	10.0M	2.0G	960	**76.7**
T2T-ViT-10	224^2	5.9M	1.5G	1192	75.2
T2T-ViT-12	224^2	6.9M	1.8G	1072	76.5
LcT-Small(ours)	224^2	3.2M	1.1G	1250	**76.7**
MobileViT-S	256^2	5.6M	2.0G	715	78.4
PVTv2-B1	224^2	13.1M	2.1G	784	78.7
CoaT-Lite Tiny	224^2	5.7M	1.6G	680	77.5
CoaT-Lite Mini	224^2	11.0M	2.0G	641	79.1
LcT-Base(ours)	224^2	6.5M	2.7G	672	**79.4**

(a) Comparison with ViTs on ImageNet-1K validation set. Throughput is measured on an RTX8000 GPU.



(b) LcT consistently outperforms the state-of-the-art models.

Object detection and semantic segmentation

Method	Backbone	Input	Param.	FLOPs	test-dev	mini-val
SSDLite	MobileNetV1	320^2	5.1M	1.3G	22.2	-
	MobileNetV2	320^2	4.3M	0.8G	22.1	†21.3
	MobileNetV3	320^2	4.9M	0.6G	22.0	-
	MnasNet-A1	320^2	4.9M	0.8G	23.0	-
	MobileViT-XS	320^2	2.7M	-	-	24.8
	LcT-Small(ours)	320^2	3.5M	2.7G	-	**26.9**
	MobileViT-S	320^2	5.7M	-	-	27.7
	LcT-Base(ours)	320^2	6.7M	6.1G	-	**31.4**
SSD	ResNet-50	300^2	22.9M	-	-	‡25.2
	VGG	300^2	34.3M	34.4G	-	†25.5
	VGG	512^2	36.0M	98.8G	-	†29.5

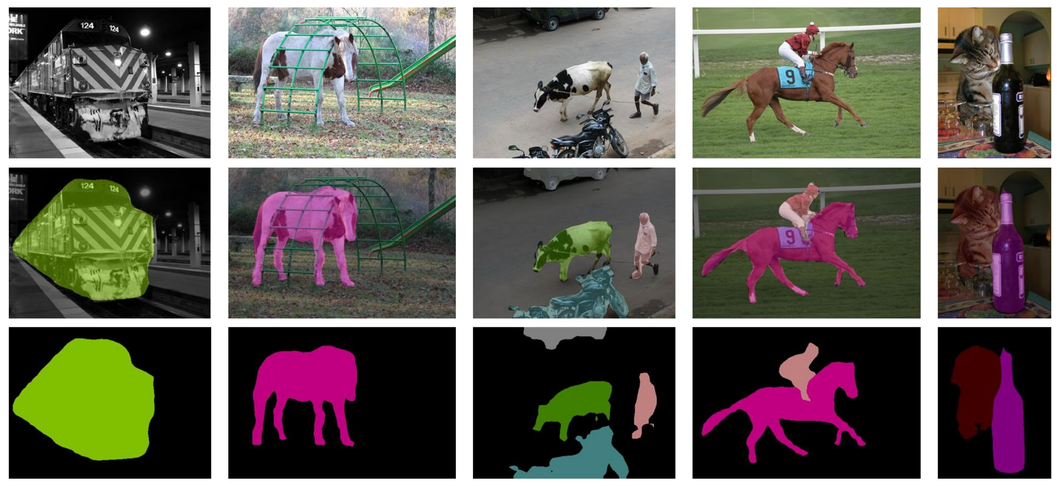
(c) Object detection results w/ SSDlite on MS COCO 2017 dataset.

Backbone	Input	Param.	FLOPs	mIOU
MobileViT-XS	512^2	2.9M	-	77.1
MobileNetV1	512^2	11.2M	14.2G	75.3
MobileNetV2	512^2	4.5M	5.8G	75.7
MobileViT-S	512^2	6.4M	-	79.1
LcT-Small(ours)	512^2	3.8M	9.1G	**79.4**
ResNet101	512^2	58.2M	81.0G	80.5
LcT-Base(ours)	512^2	7.4M	19.5G	**81.0**

(d) Semantic segmentation results w/ DeepLabV3 on the VOC 2012 validation set.

Qualitative results in downstream tasks

(e) Qualitative results on the MS COCO 2017 validation set.



(f) Qualitative results on the VOC 2012 validation set.

References

[1] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In International Conference on Learning Representations, 2021.

[2] Sachin Mehta and Mohammad Rastegari. Mobilevit: light-weight, general-purpose, and mobile-friendly vision transformer. arXiv preprint arXiv:2110.02178, 2021.