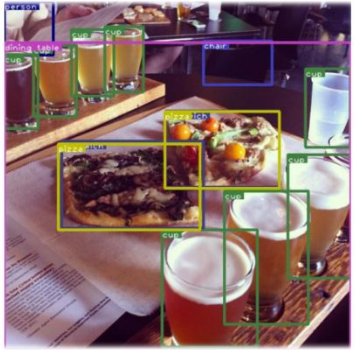
MobileViT

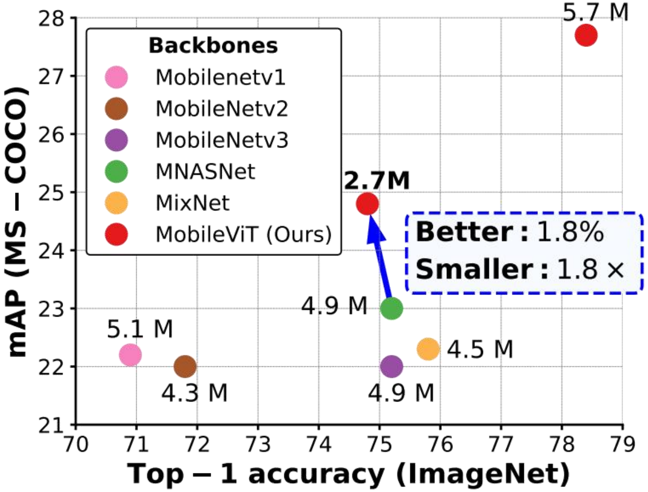
MOBILEVIT: LIGHT-WEIGHT, GENERAL-PURPOSE, AND MOBILE-FRIENDLY VISION TRANSFORMER

2021

Sachin Mehta
Apple

Mohammad Rastegari
Apple





论文下载：https://arxiv.org/abs/2110.02178

官方源码（Pytorch实现）：https://github.com/apple/ml-cvnets

自己从ml-cvnets仓库中剥离的代码：https://github.com/WZMIAOMIAO/deep-learning-for-image-processing/tree/master/pytorch_classification/MobileViT

对应博文：https://blog.csdn.net/qq_37541097/article/details/126715733

公众号：阿喆学习小记

MobileViT

目录

0. 前言

1. 模型结构解析

    1.1 Vision Transformer结构简介

    1.2 MobileViT介绍

    1.3 Patch Size对性能的影响

2. 模型详细配置

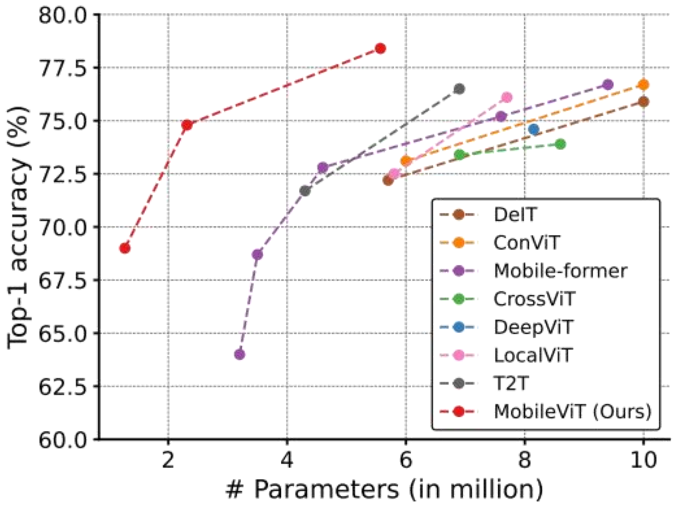
前言

当前纯Transformer模型存在的问题

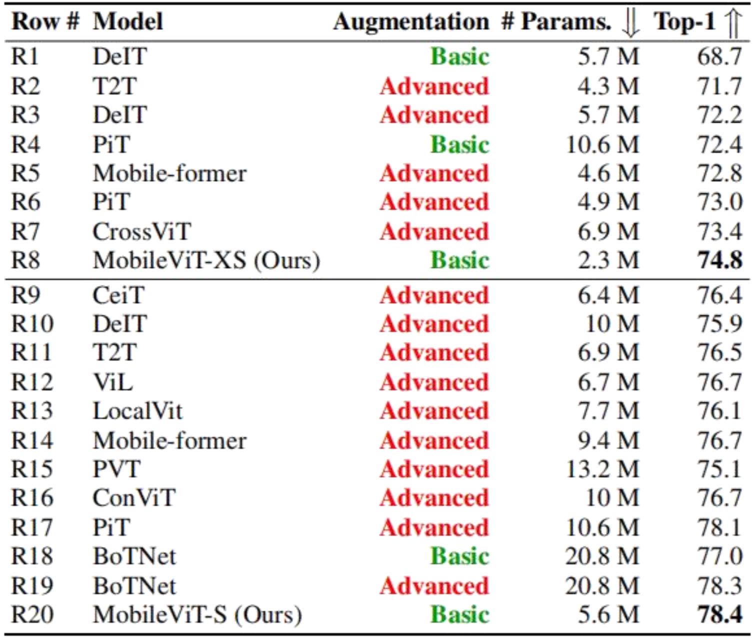
- Transformer参数多，算力要求高
- Transformer缺少空间归纳偏置
- Transformer迁移到其他任务比较繁琐
- Transformer模型训练困难

MobileViT

前言



(a)



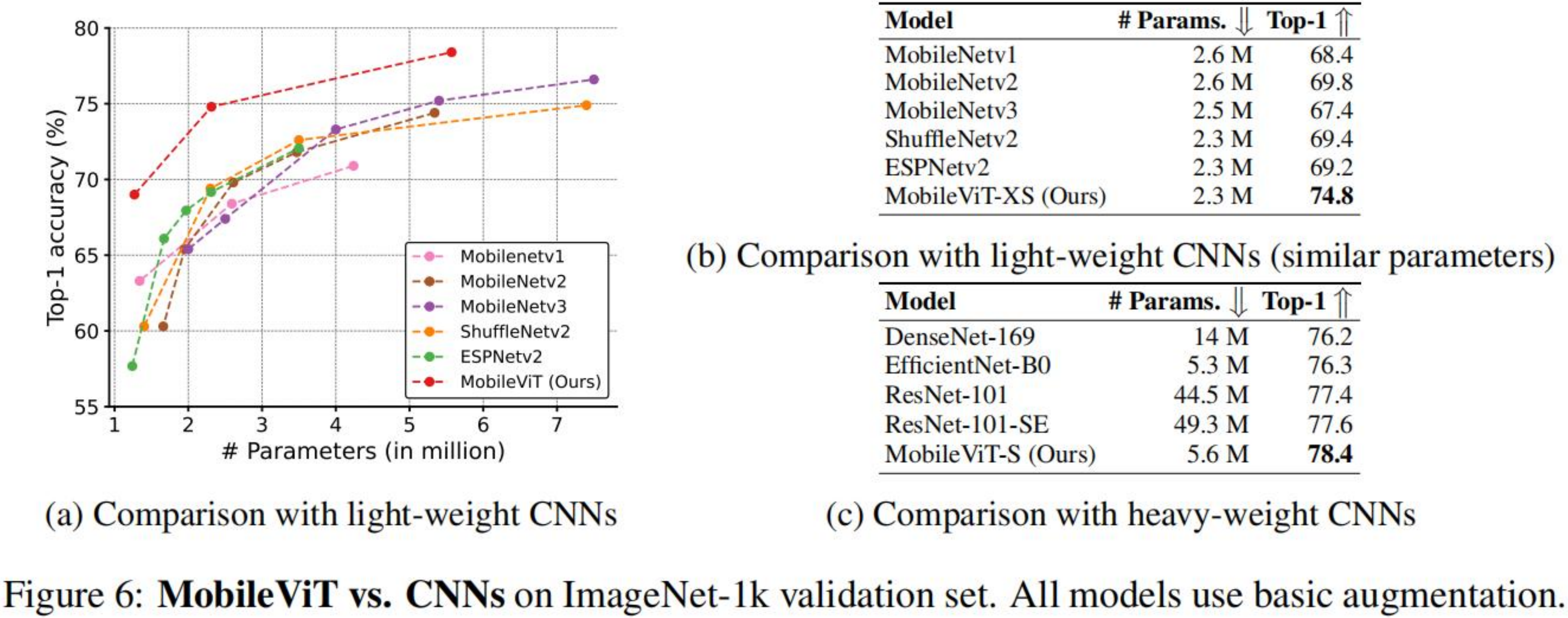
Row #	Model	Augmentation	# Params. ⇓	Top-1 ⇑
R1	DeIT	Basic	5.7 M	68.7
R2	T2T	Advanced	4.3 M	71.7
R3	DeIT	Advanced	5.7 M	72.2
R4	PiT	Basic	10.6 M	72.4
R5	Mobile-former	Advanced	4.6 M	72.8
R6	PiT	Advanced	4.9 M	73.0
R7	CrossViT	Advanced	6.9 M	73.4
R8	MobileViT-XS (Ours)	Basic	2.3 M	**74.8**
R9	CeiT	Advanced	6.4 M	76.4
R10	DeIT	Advanced	10 M	75.9
R11	T2T	Advanced	6.9 M	76.5
R12	ViL	Advanced	6.7 M	76.7
R13	LocalVit	Advanced	7.7 M	76.1
R14	Mobile-former	Advanced	9.4 M	76.7
R15	PVT	Advanced	13.2 M	75.1
R16	ConViT	Advanced	10 M	76.7
R17	PiT	Advanced	10.6 M	78.1
R18	BoTNet	Basic	20.8 M	77.0
R19	BoTNet	Advanced	20.8 M	78.3
R20	MobileViT-S (Ours)	Basic	5.6 M	**78.4**

(b)

Figure 7: **MobileViT vs. ViTs** on ImageNet-1k validation set. Here, **basic** means ResNet-style augmentation while **advanced** means a combination of augmentation methods with basic (e.g., MixUp (Zhang et al., 2018), RandAugmentation (Cubuk et al., 2019), and CutMix (Zhong et al., 2020)).

MobileViT

前言



Model	# Params. ⇓	Top-1 ⇑
MobileNetv1	2.6 M	68.4
MobileNetv2	2.6 M	69.8
MobileNetv3	2.5 M	67.4
ShuffleNetv2	2.3 M	69.4
ESPNetv2	2.3 M	69.2
MobileViT-XS (Ours)	2.3 M	**74.8**

(b) Comparison with light-weight CNNs (similar parameters)

Model	# Params. ⇓	Top-1 ⇑
DenseNet-169	14 M	76.2
EfficientNet-B0	5.3 M	76.3
ResNet-101	44.5 M	77.4
ResNet-101-SE	49.3 M	77.6
MobileViT-S (Ours)	5.6 M	**78.4**

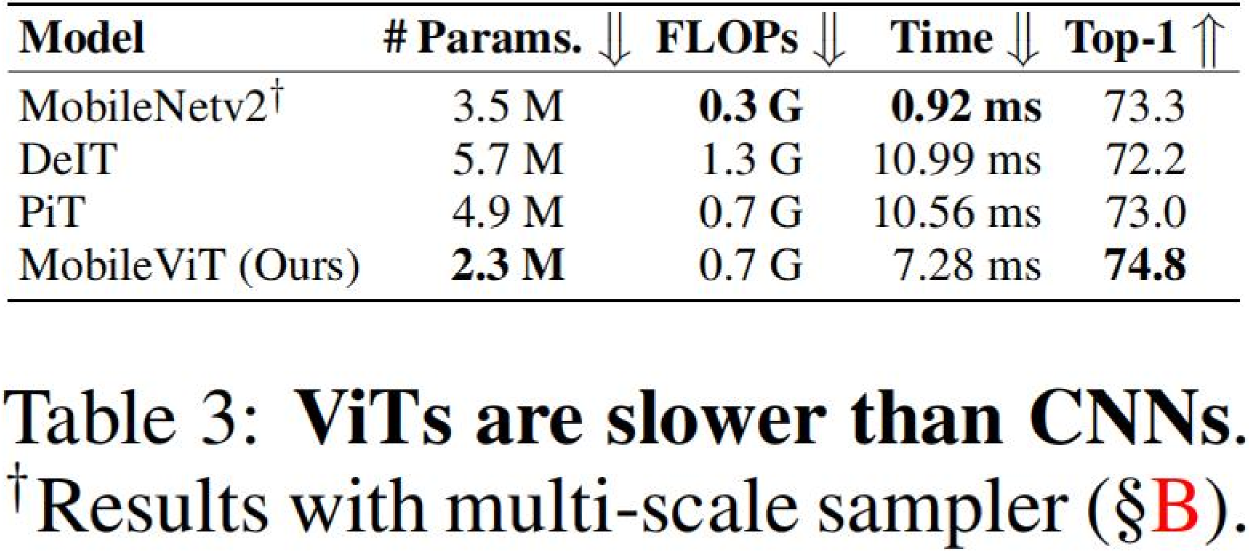
(a) Comparison with light-weight CNNs

(c) Comparison with heavy-weight CNNs

Figure 6: **MobileViT vs. CNNs** on ImageNet-1k validation set. All models use basic augmentation.

前言

移动端

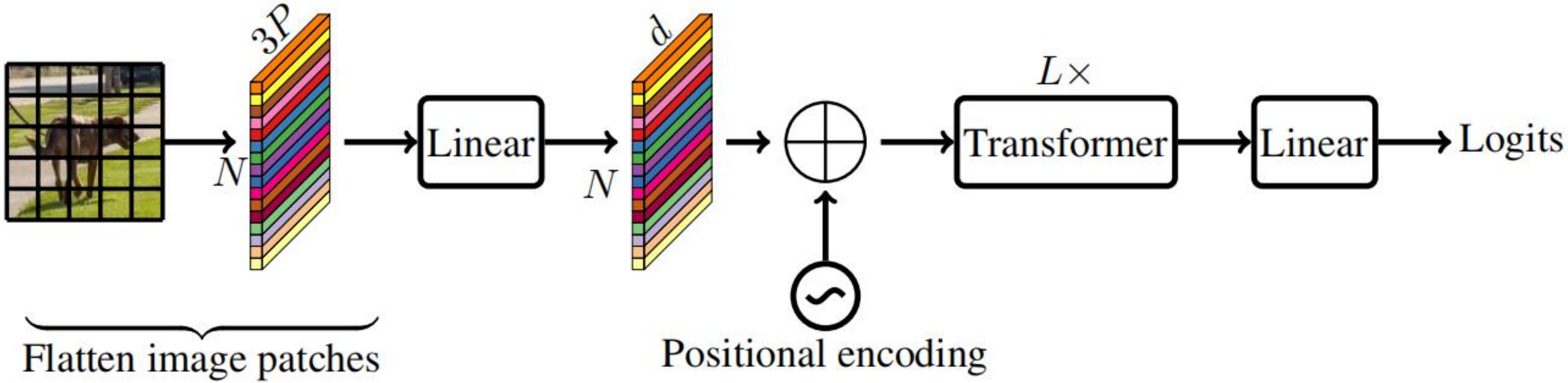


Model	**# Params. ⇓**	**FLOPs ⇓**	**Time ⇓**	**Top-1 ⇑**
MobileNetv2[†]	3.5 M	**0.3 G**	**0.92 ms**	73.3
DeIT	5.7 M	1.3 G	10.99 ms	72.2
PiT	4.9 M	0.7 G	10.56 ms	73.0
MobileViT (Ours)	**2.3 M**	0.7 G	7.28 ms	**74.8**

Table 3: **ViTs are slower than CNNs.** [†]Results with multi-scale sampler (§B).

模型结构解析

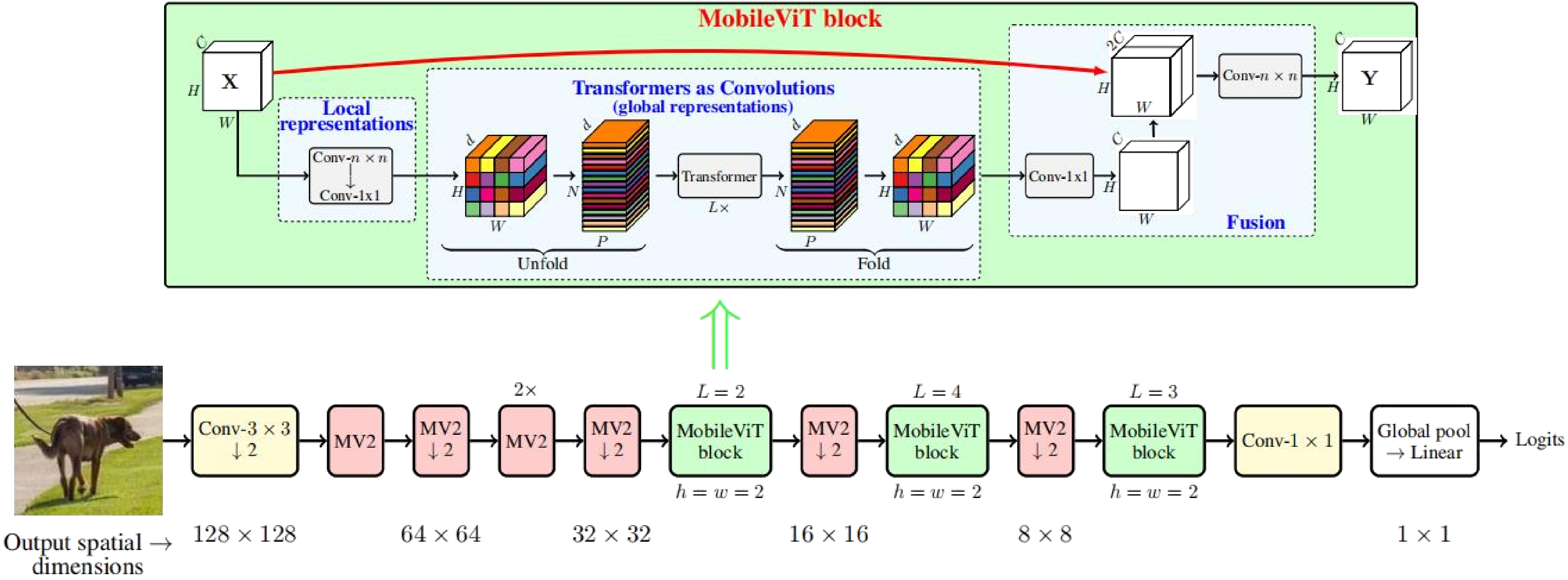
Vision Transformer



(a) **Standard visual transformer (ViT)**

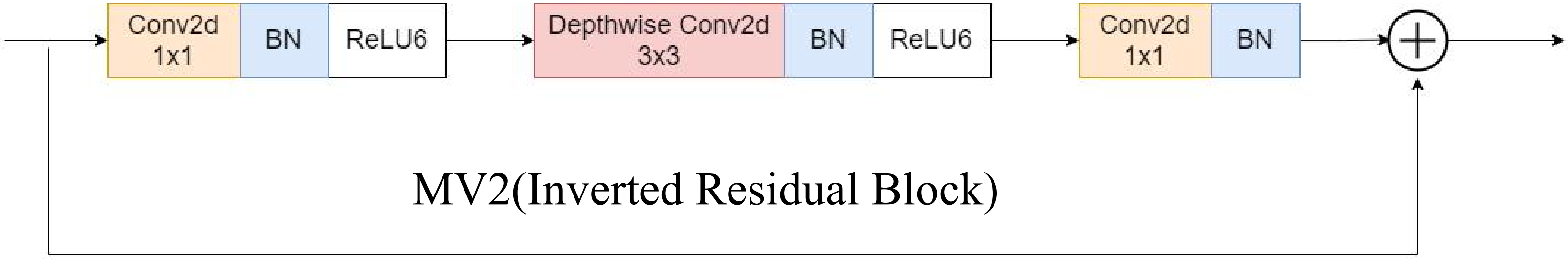
模型结构解析

MobileViT

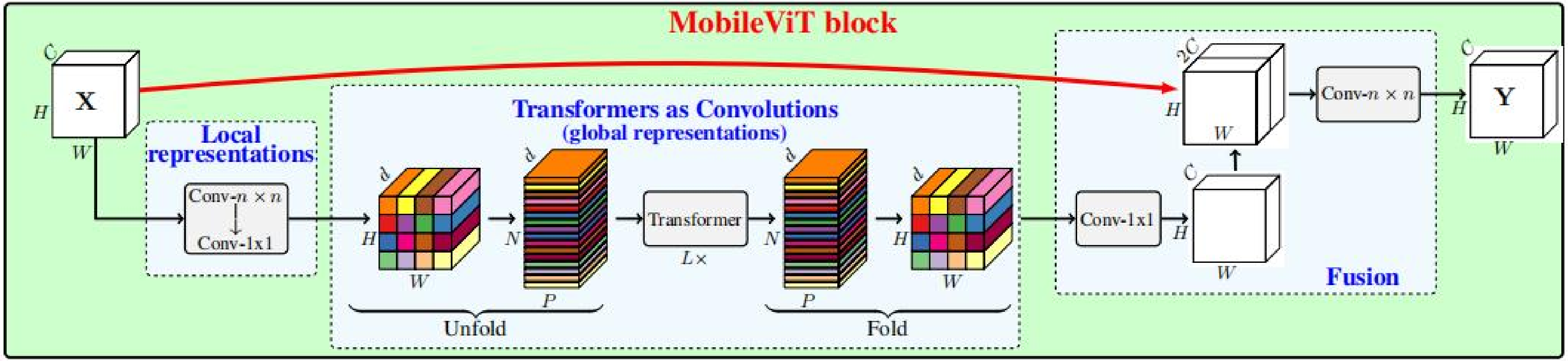


模型结构解析

MobileViT

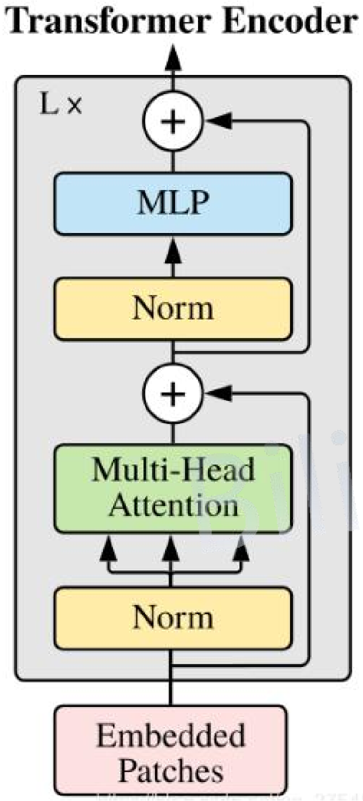


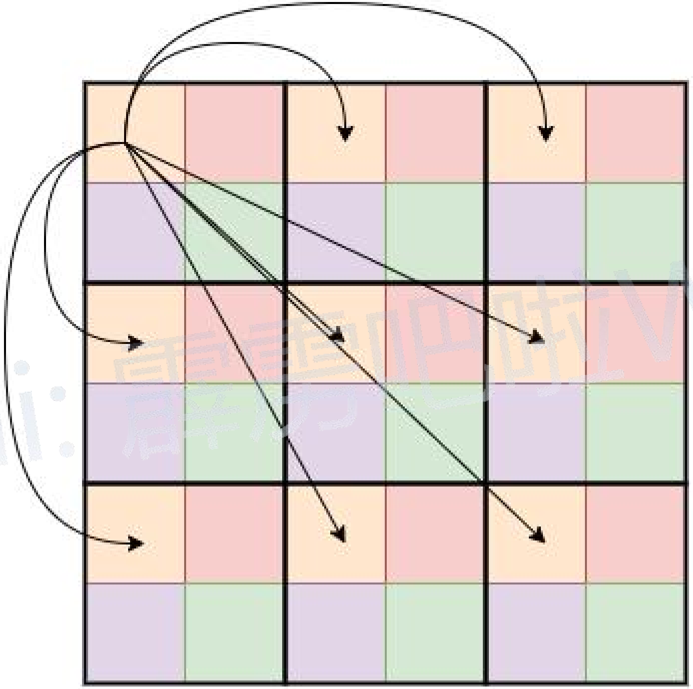
MV2(Inverted Residual Block)



模型结构解析

MobileViT





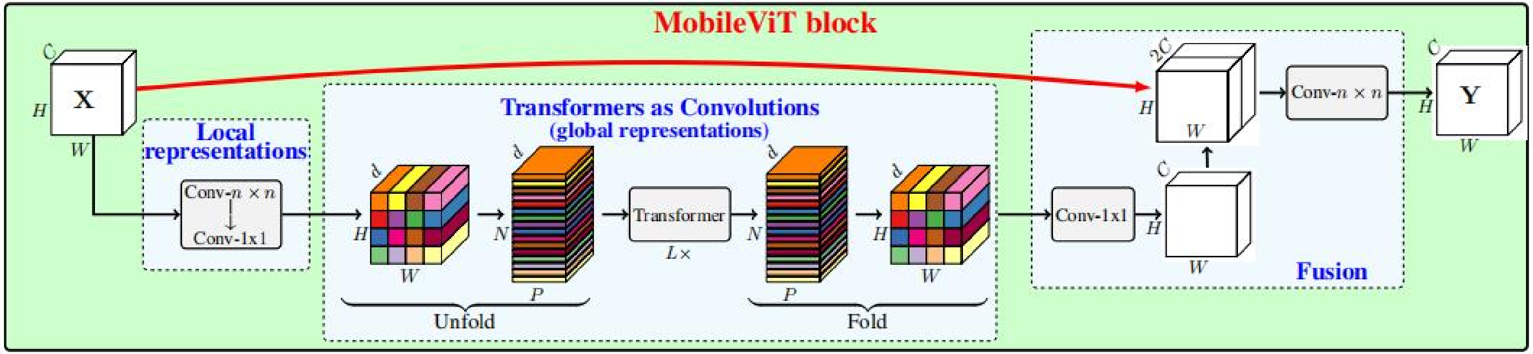
$$c1 = WHC$$

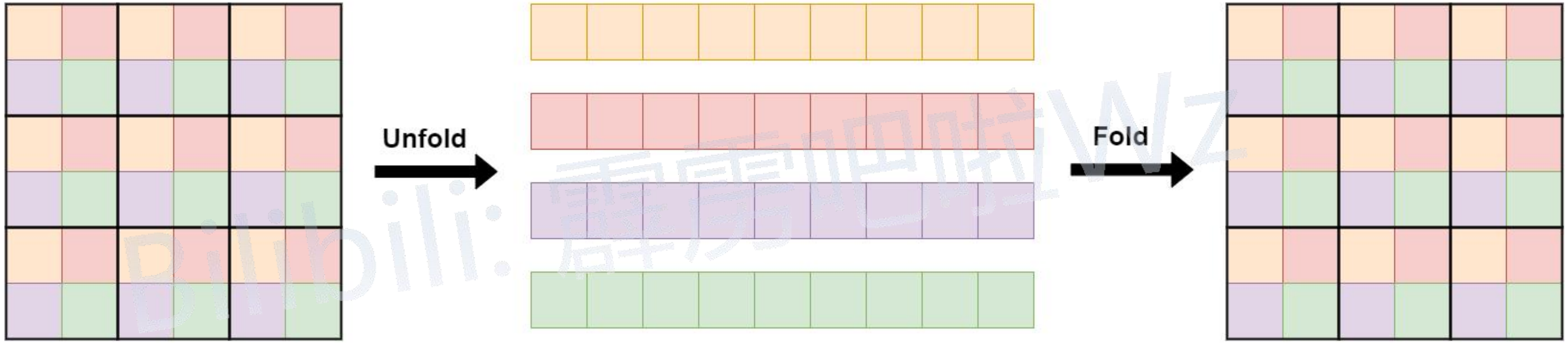
$$c2 = \frac{WHC}{4}$$

bilibili:霹雳吧啦Wz

模型结构解析

MobileViT

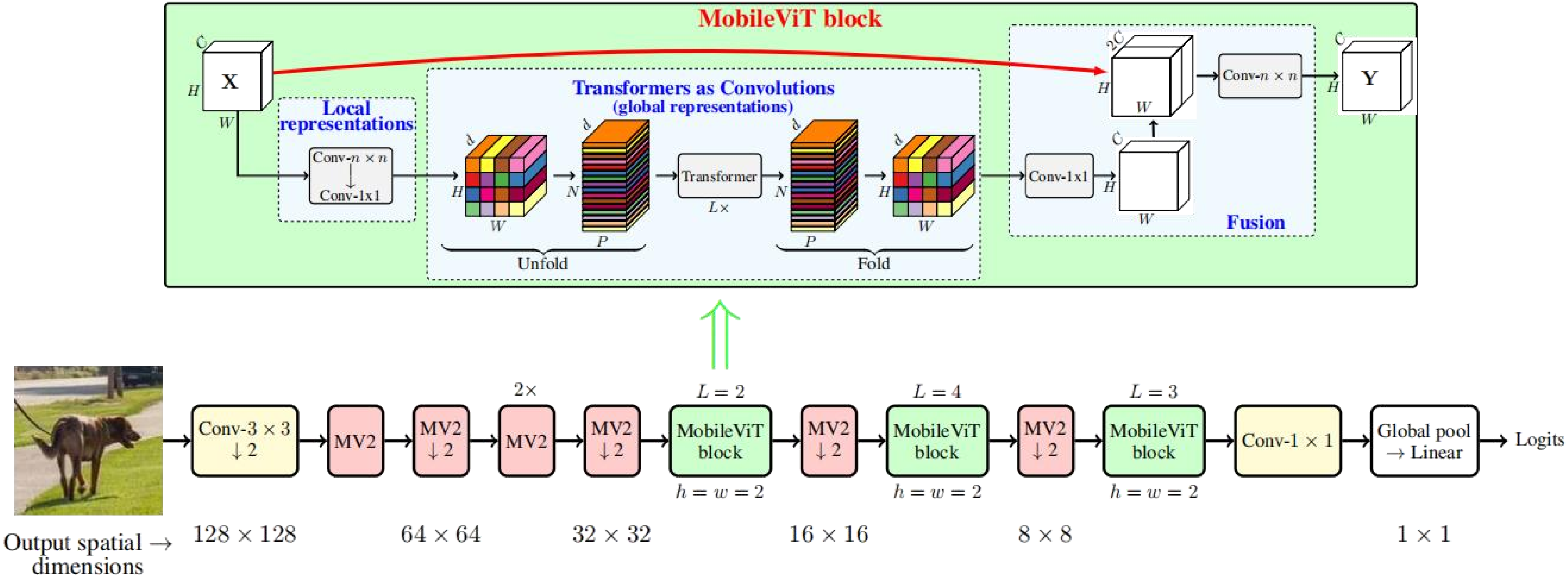




Bilibili: 霹雳吧啦Wz

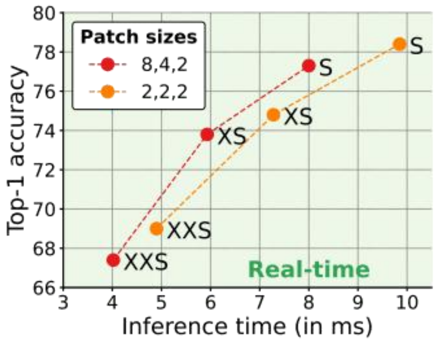
模型结构解析

MobileViT

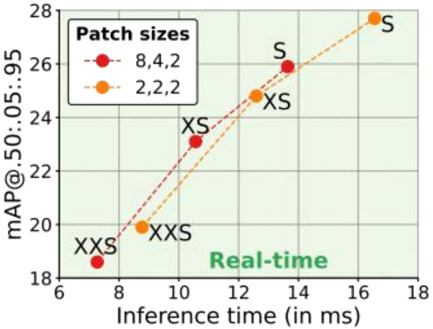


模型结构解析

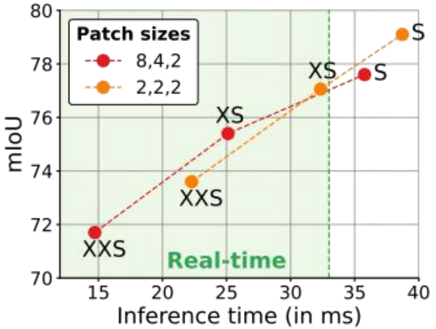
Patch Size影响



(a) Classification @ 256 × 256



(b) Detection @ 320 × 320

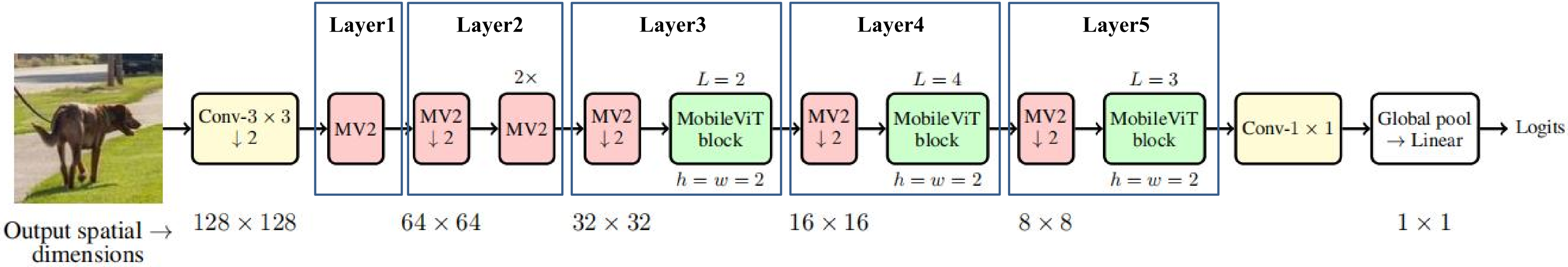


(c) Segmentation @ 512 × 512

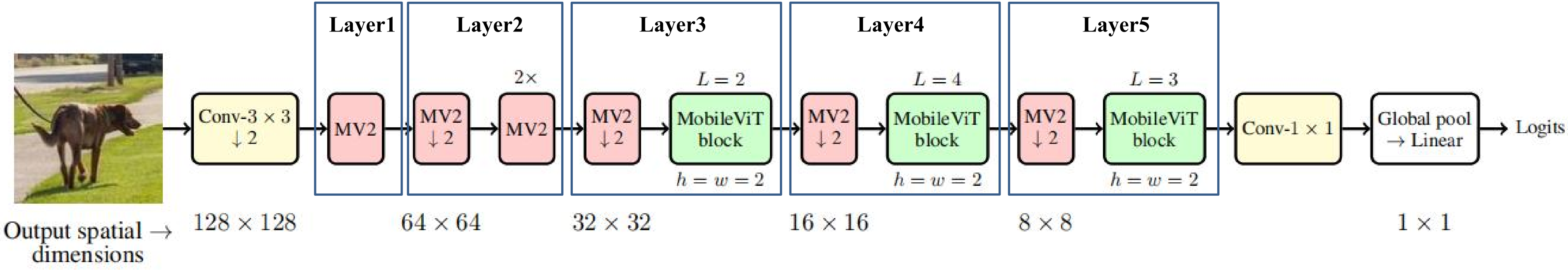
对语义细节要求逐渐提高

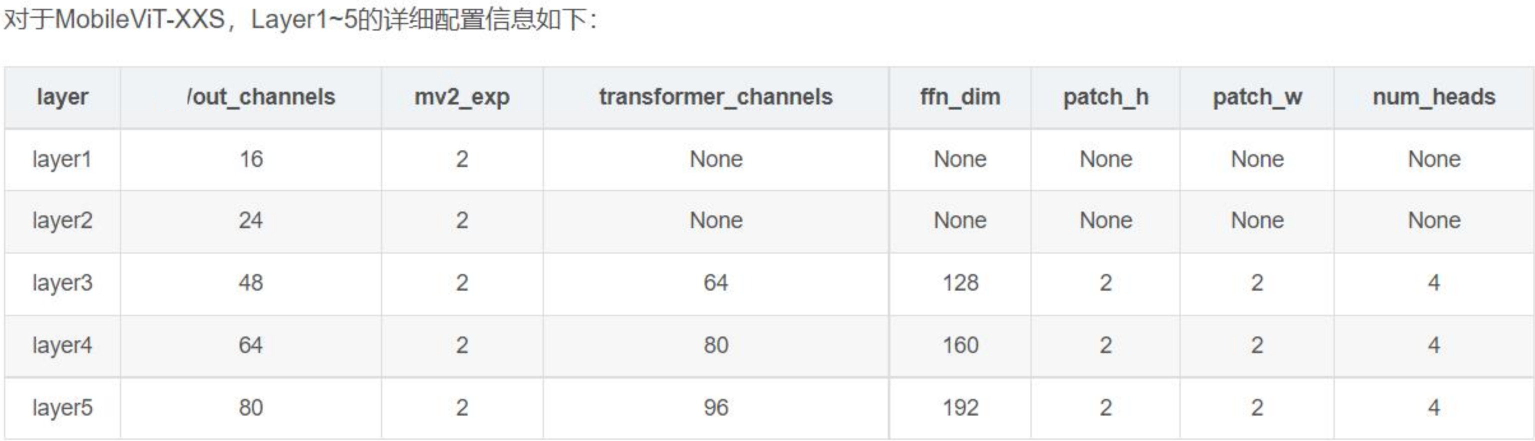
模型配置

- MobileViT-S(small)
- MobileViT-XS(extra small)
- MobileViT-XXS(extra extra small)



模型配置

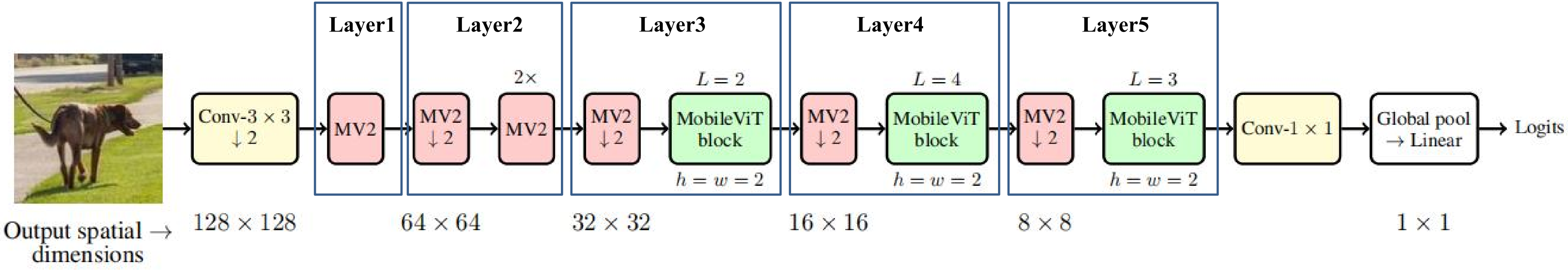




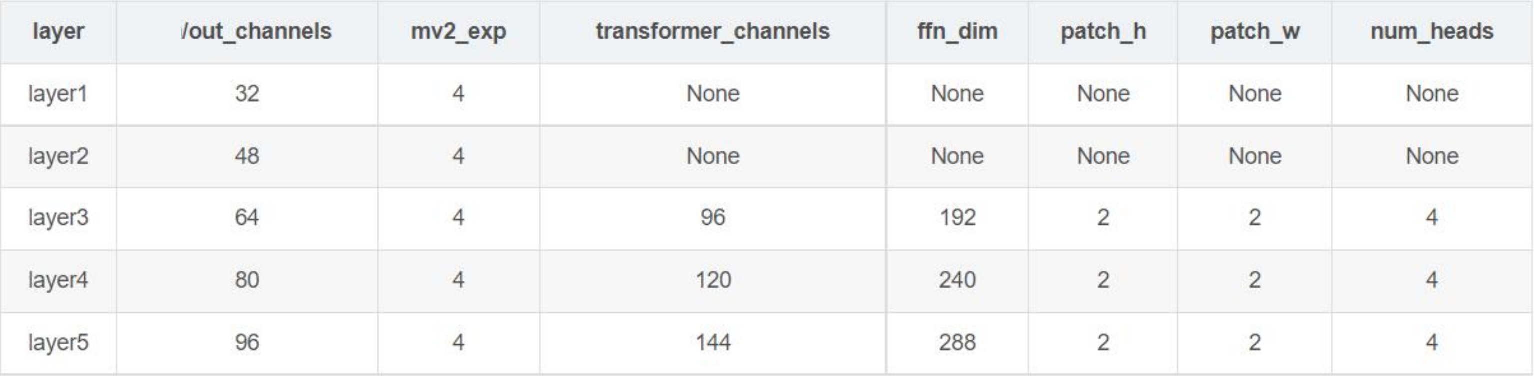
对于MobileViT-XXS，Layer1~5的详细配置信息如下：

layer	/out_channels	mv2_exp	transformer_channels	ffn_dim	patch_h	patch_w	num_heads
layer1	16	2	None	None	None	None	None
layer2	24	2	None	None	None	None	None
layer3	48	2	64	128	2	2	4
layer4	64	2	80	160	2	2	4
layer5	80	2	96	192	2	2	4

模型配置

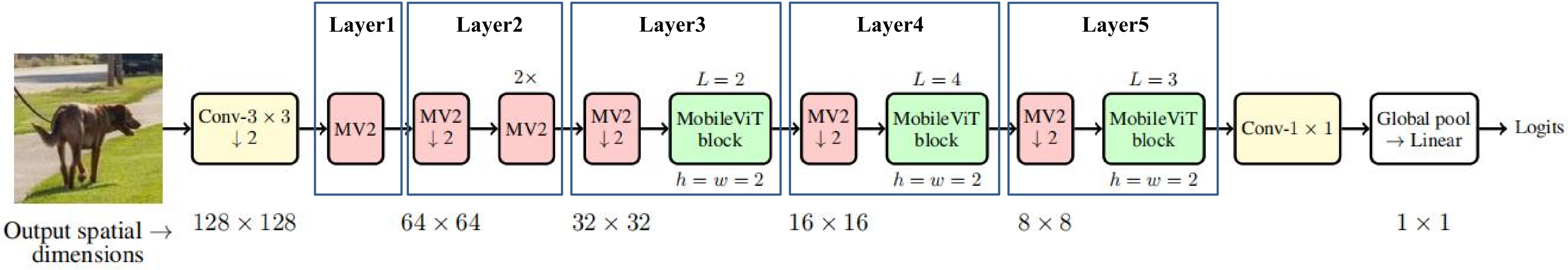


对于MobileViT-XS，Layer1~5的详细配置信息如下：

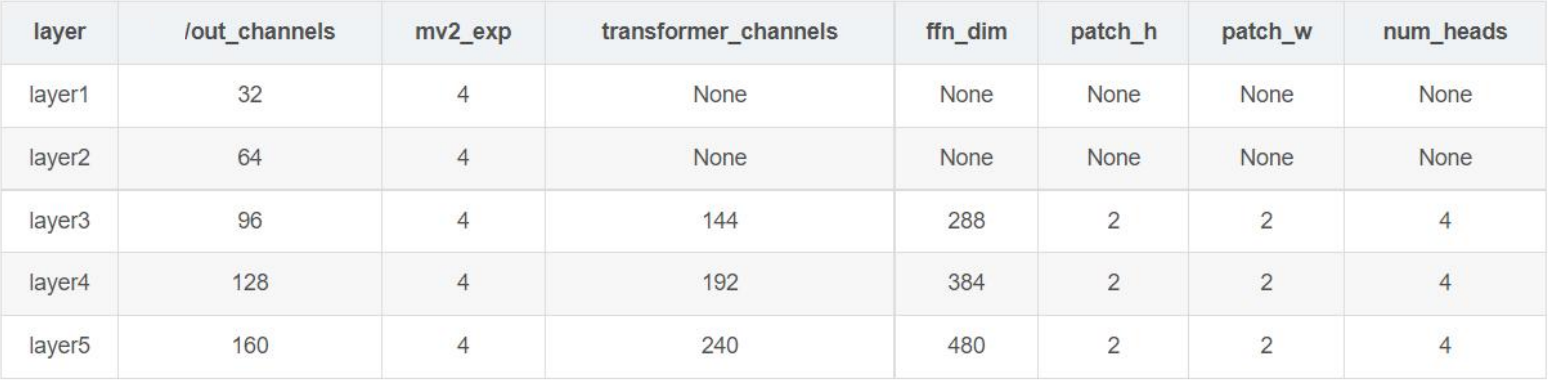


layer	/out_channels	mv2_exp	transformer_channels	ffn_dim	patch_h	patch_w	num_heads
layer1	32	4	None	None	None	None	None
layer2	48	4	None	None	None	None	None
layer3	64	4	96	192	2	2	4
layer4	80	4	120	240	2	2	4
layer5	96	4	144	288	2	2	4

模型配置



对于MobileViT-S，Layer1~5的详细配置信息如下：



layer	/out_channels	mv2_exp	transformer_channels	ffn_dim	patch_h	patch_w	num_heads
layer1	32	4	None	None	None	None	None
layer2	64	4	None	None	None	None	None
layer3	96	4	144	288	2	2	4
layer4	128	4	192	384	2	2	4
layer5	160	4	240	480	2	2	4