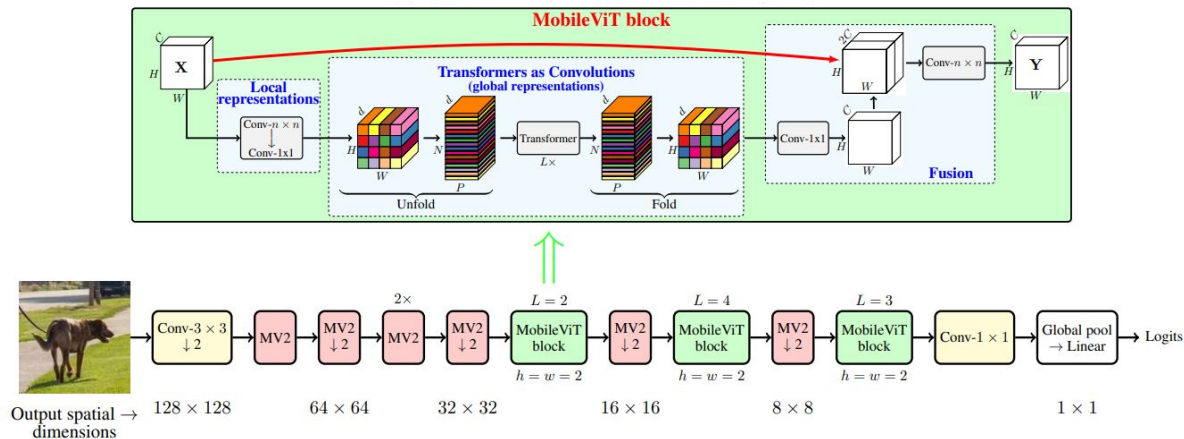# MobileViT
Mobile Vision Transformer

What is MobileViT?

MobileViT is a light-weight and general-purpose vision transformer for mobile vision tasks. It combines the strength of the standard CNN and the ViT. It has outperforms several CNNs and ViT-based network across different tasks and datasets.

What is the need for MobileViT

1. ViT has huge parameters.
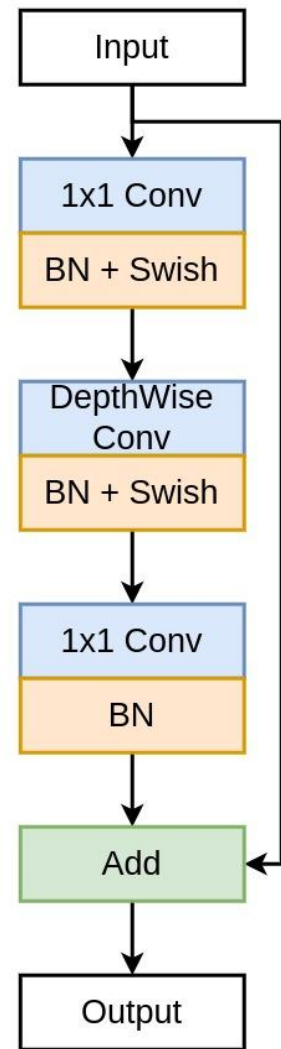
2. Cannot work with real-world applications

Architecture

1. Inverted Residual block (MV2 block)

2. MobileViT block

Inverted Residual Block

1. These blocks are mainly used for used for downsampling the feature maps by applying strided convolution.

MobileViT Block

1. MobileViT block aims to model the local and global information in an input tensor with fewer parameters.

2. P = wh and N = HW / P

Input	
3x3 Conv	H x W x C
BN + Swish	H x W x C
1x1 Conv	H x W x d
BN + Swish	H x W x d
Reshape	P x N x d
Transformer	P x N x d
Reshape	H x W x d
1x1 Conv	H x W x C
BN + Swish	H x W x C
Concatenate	H x W x C*2
3x3 Conv	H x W x D
BN + Swish	H x W x D
Output	