MnasNet: Platform-Aware Neural Architecture Search for Mobile

From EfficientNet

Introduction

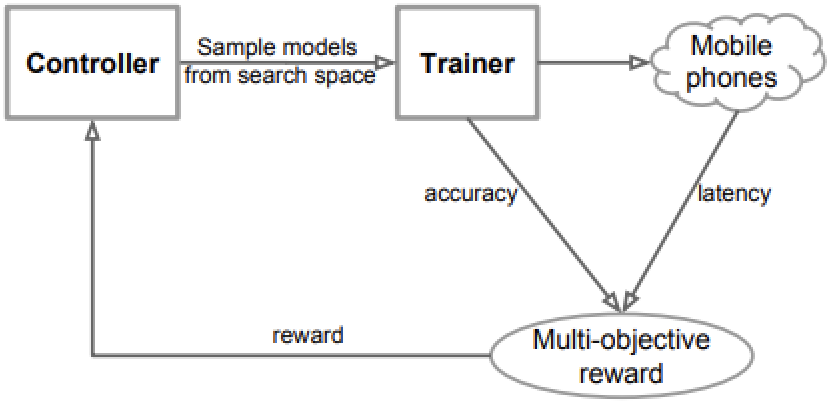


Figure 1: **An Overview of Platform-Aware Neural Architecture Search for Mobile.**

Automated neural architecture search approach

Multi-objective optimization

Directly measure real-world latency

Factorized hierarchical search space

Aim : find CNN model with both high accuracy and low inference latency

Objective Function

$$\begin{aligned} \underset{m}{\text{maximize}} \quad & ACC(m) \\ \text{subject to} \quad & LAT(m) \leq T \end{aligned} \qquad (1)$$

Pareto Optimal

$$\underset{m}{\text{maximize}} \quad ACC(m) \times \left[\frac{LAT(m)}{T}\right]^{w} \qquad (2)$$

$$w = \begin{cases} \alpha, & \text{if } LAT(m) \leq T \\ \beta, & \text{otherwise} \end{cases} \qquad (3)$$

Pareto Optimal:

Highest accuracy without increasing latency or lowest latency without decreasing accuracy

= simultaneously considers both accuracy and latency

Aim : find CNN model with both high accuracy and low inference latency

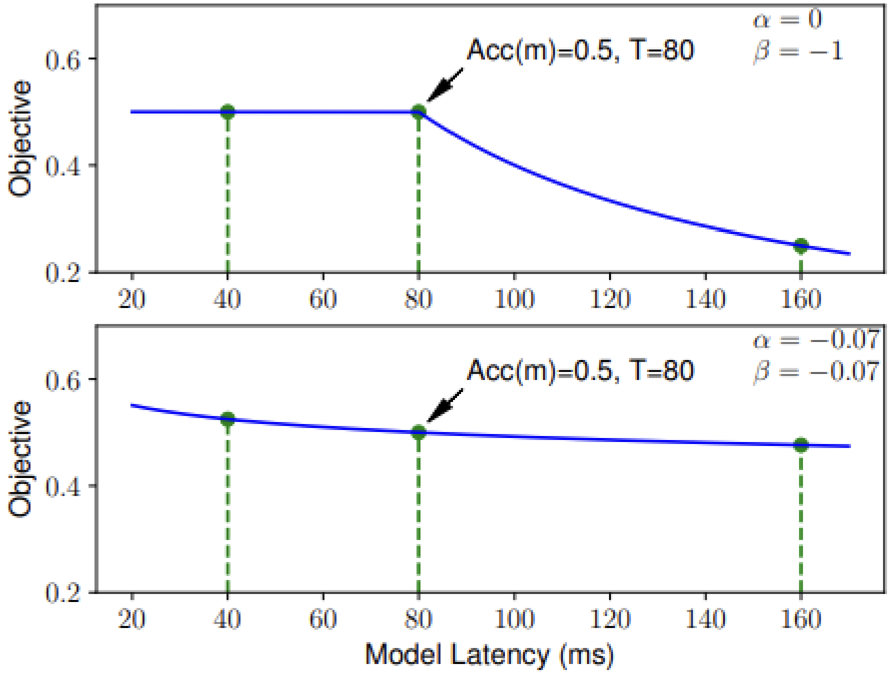


Figure 3: **Objective Function Defined by Equation 2**, assuming accuracy $ACC(m)$=0.5 and target latency T=80ms: (top) show the object values with latency as a hard constraint; (bottom) shows the objective values with latency as a soft constraint.

$$\underset{m}{\text{maximize}} \quad ACC(m) \times \left[\frac{LAT(m)}{T}\right]^{w} \tag{2}$$

$$w = \begin{cases} \alpha, & \text{if } LAT(m) \leq T \\ \beta, & \text{otherwise} \end{cases} \tag{3}$$

MnasNet in detail...

1. Factorized Hierarchical Search Space

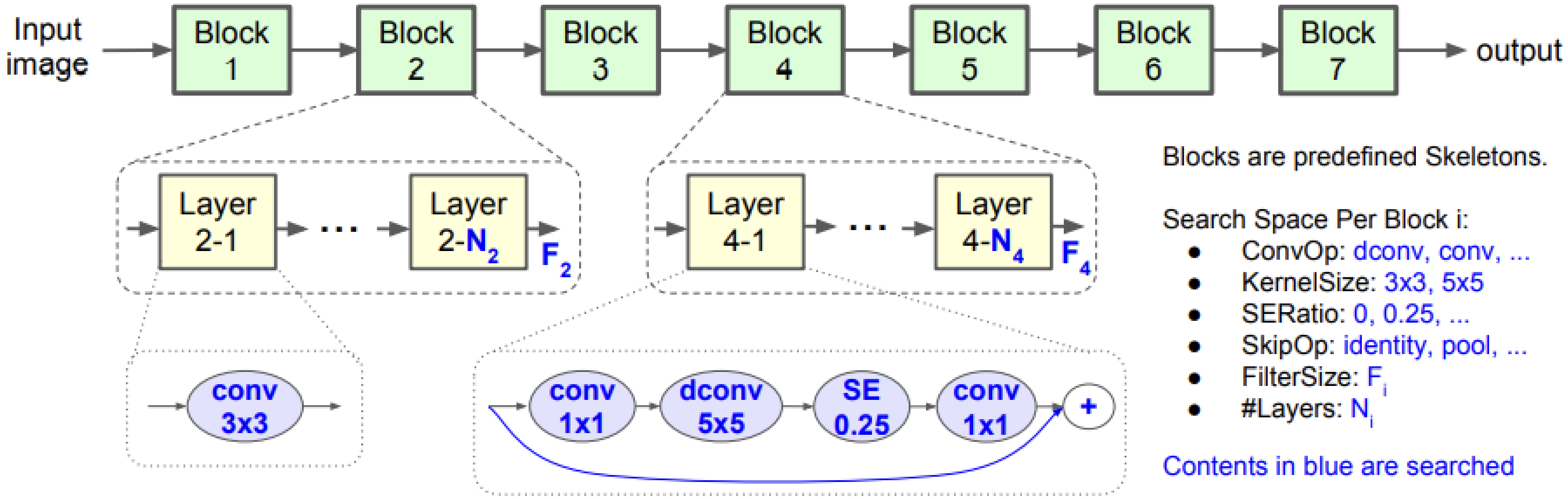
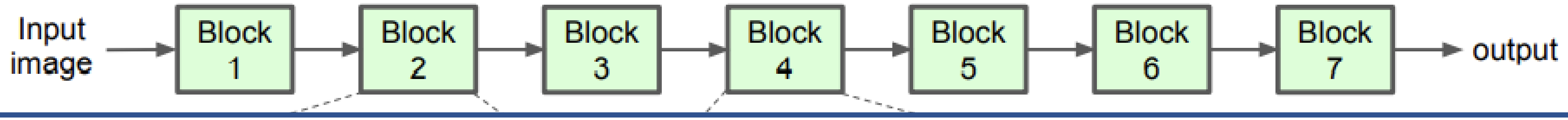


Figure 4: **Factorized Hierarchical Search Space.** Network layers are grouped into a number of predefined skeletons, called blocks, based on their input resolutions and filter sizes. Each block contains a variable number of repeated identical layers where only the first layer has stride 2 if input/output resolutions are different but all other layers have stride 1. For each block, we search for the operations and connections for a single layer and the number of layers N, then the same layer is repeated N times (e.g., Layer 4-1 to 4-N_4 are the same). Layers from different blocks (e.g., Layer 2-1 and 4-1) can be different.

MnasNet in detail…

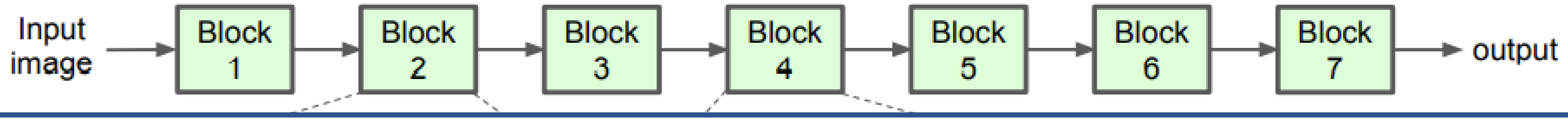
1. Factorized Hierarchical Search Space



- Convolutional ops $ConvOp$: regular conv (conv), depthwise conv (dconv), and mobile inverted bottleneck conv [29].
- Convolutional kernel size $KernelSize$: 3x3, 5x5.
- Squeeze-and-excitation [13] ratio $SERatio$: 0, 0.25.
- Skip ops $SkipOp$: pooling, identity residual, or no skip.
- Output filter size F_i.
- Number of layers per block N_i.

MnasNet in detail...

1. Factorized Hierarchical Search Space



Intuition: Layer diversity is critical for achieving both high accuracy and low latency.

Ex) based on input shapes and output shapes, specific sequence of operations return better results than fixed operations

MnasNet in detail…

2. Search Algorithm

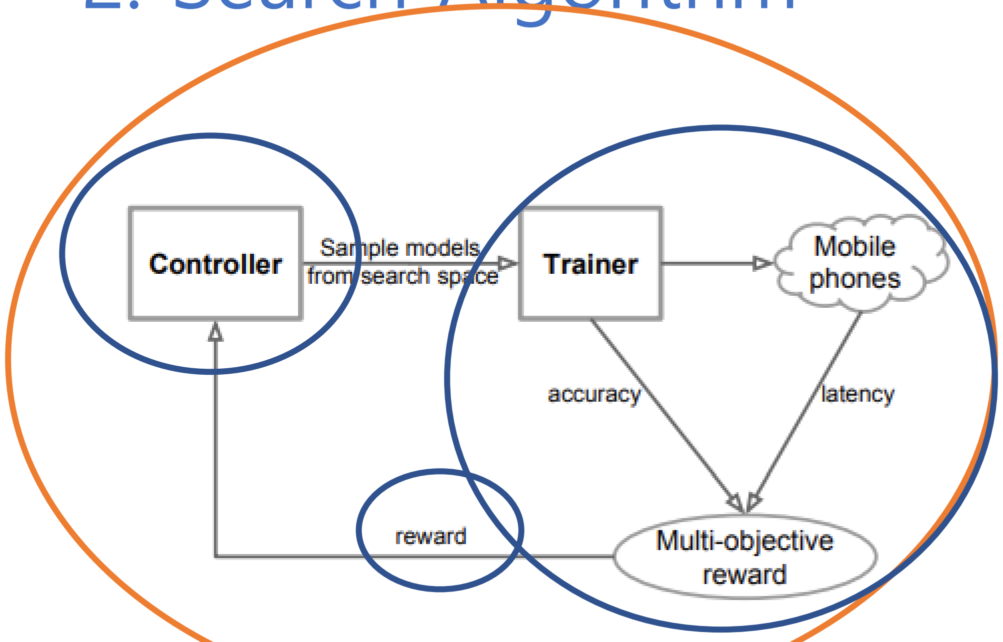
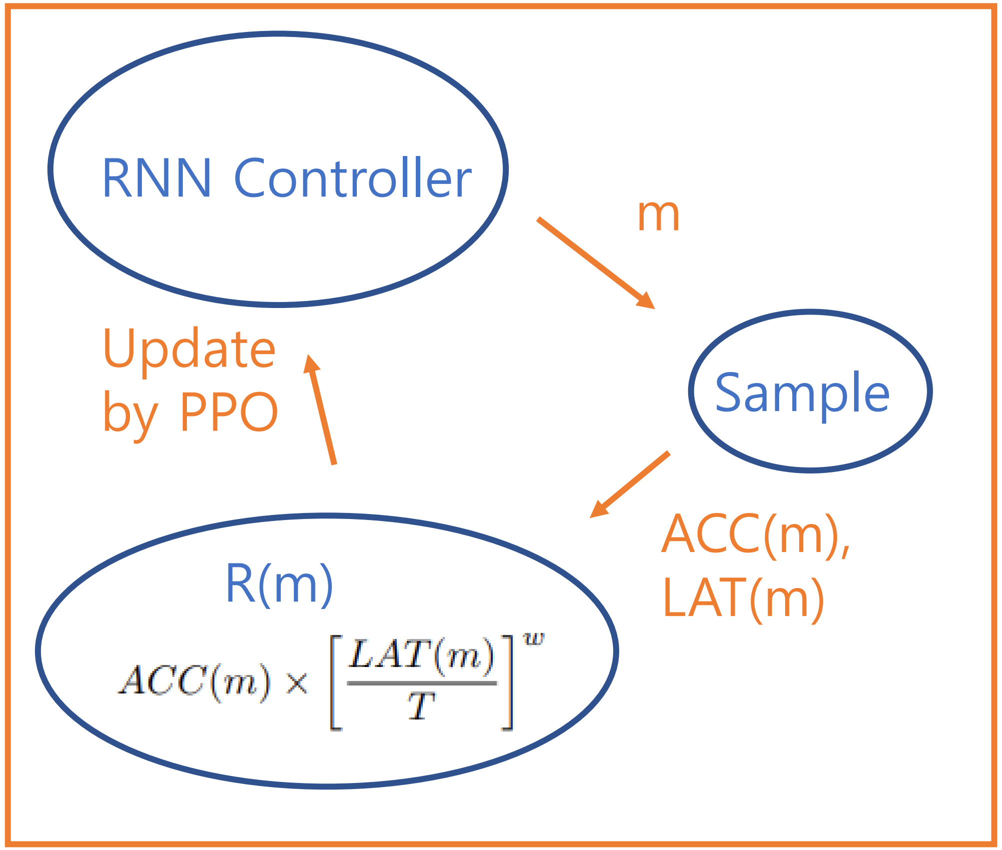
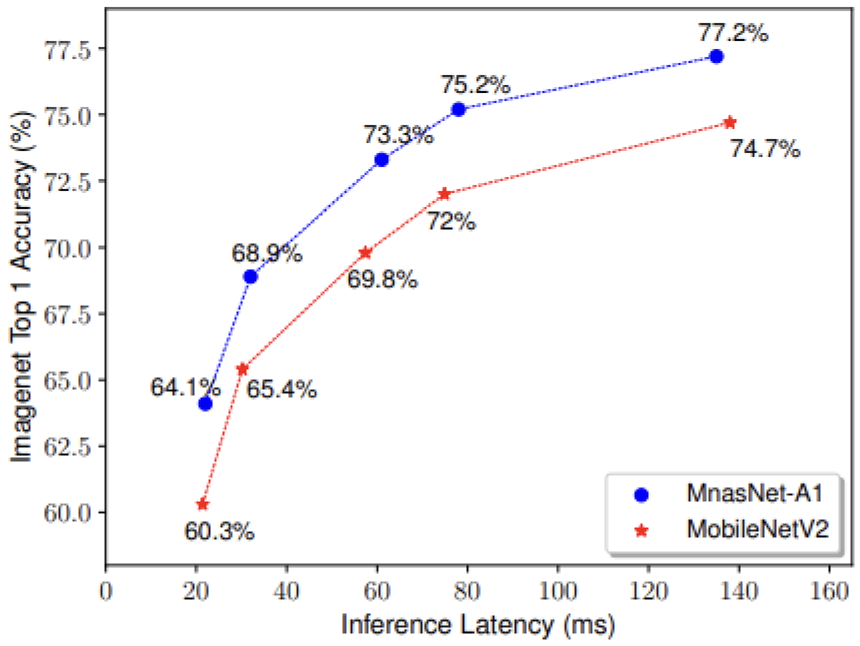


Figure 1: **An Overview of Platform-Aware Neural Architecture Search for Mobile.**

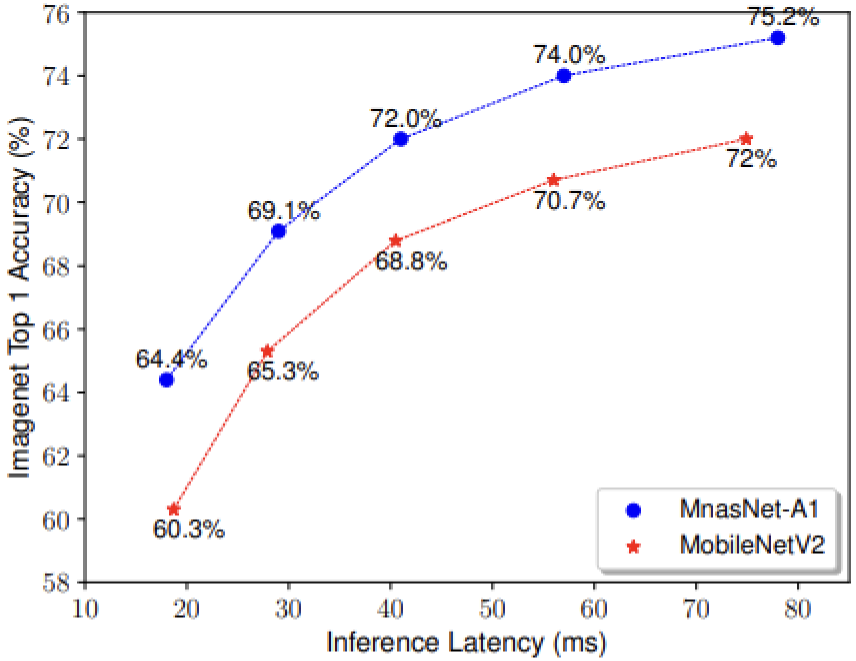
Sample-eval-update



Scaling Performance



(a) Depth multiplier = 0.35, 0.5, 0.75, 1.0, 1.4, corresponding to points from left to right.



(b) Input size = 96, 128, 160, 192, 224, corresponding to points from left to right.

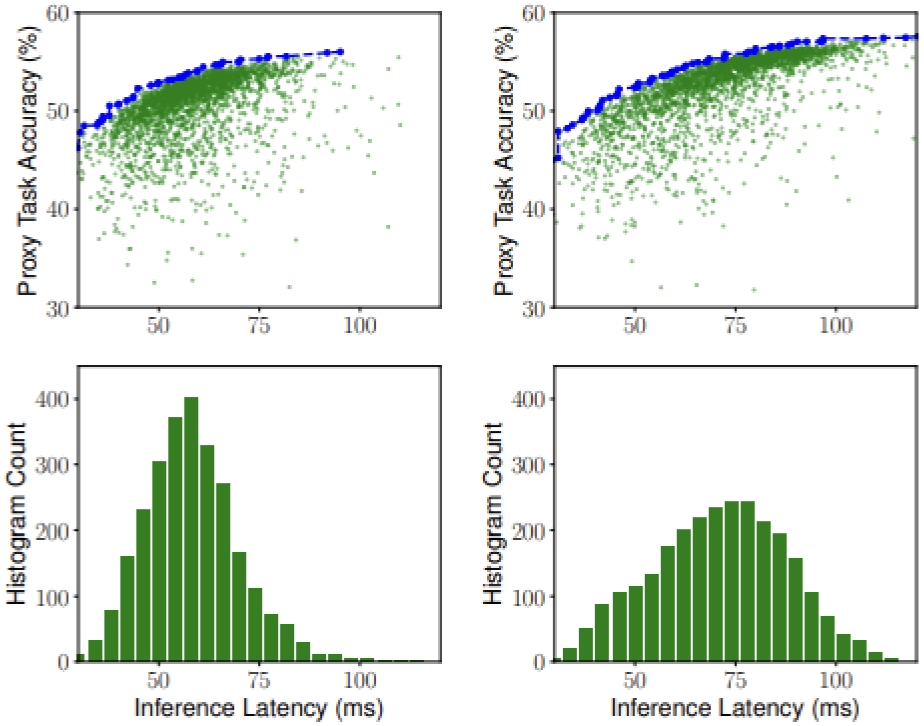
Figure 5: **Performance Comparison with Different Model Scaling Techniques**. MnasNet is our baseline model shown in Table 1. We scale it with the same depth multipliers and input sizes as MobileNetV2.

Ablation Study

Soft vs. Hard Latency Constraint

$$\underset{m}{\text{maximize}} \quad ACC(m) \times \left[\frac{LAT(m)}{T}\right]^w \tag{2}$$

$$w = \begin{cases} \alpha, & \text{if } LAT(m) \leq T \\ \beta, & \text{otherwise} \end{cases} \tag{3}$$



(a) $\alpha = 0, \beta = -1$ (b) $\alpha = \beta = -0.07$

Figure 6: **Multi-Objective Search Results** based on equation 2 with (a) α=0, β=-1; and (b) α=β=-0.07. Target latency is T=75ms. Top figure shows the Pareto curve (blue line) for the 3000 sampled models (green dots); bottom figure shows the histogram of model latency.

Ablation Study

Reward and Search Space

Starting from NASNet

Reward	Search Space	Latency	Top-1 Acc.
Single-obj [36]	Cell-based [36]	183ms	74.0%
Multi-obj	Cell-based [36]	100ms	72.0%
Multi-obj	**MnasNet**	**78ms**	**75.2%**

Ablation Study

Layer Diversity

	Top-1 Acc.	Inference Latency
MnasNet-A1	**75.2%**	**78ms**
MBConv3 (k3x3) only	71.8%	63ms
MBConv3 (k5x5) only	72.5%	79ms
MBConv6 (k3x3) only	74.9%	116ms
MBConv6 (k5x5) only	75.6%	146ms

Table 6: **Performance Comparison of MnasNet and Its Variants** – *MnasNet-A1* denotes the model shown in Figure 7(a); others are variants that repeat a single type of layer throughout the network. All models have the same number of layers and same filter size at each layer.

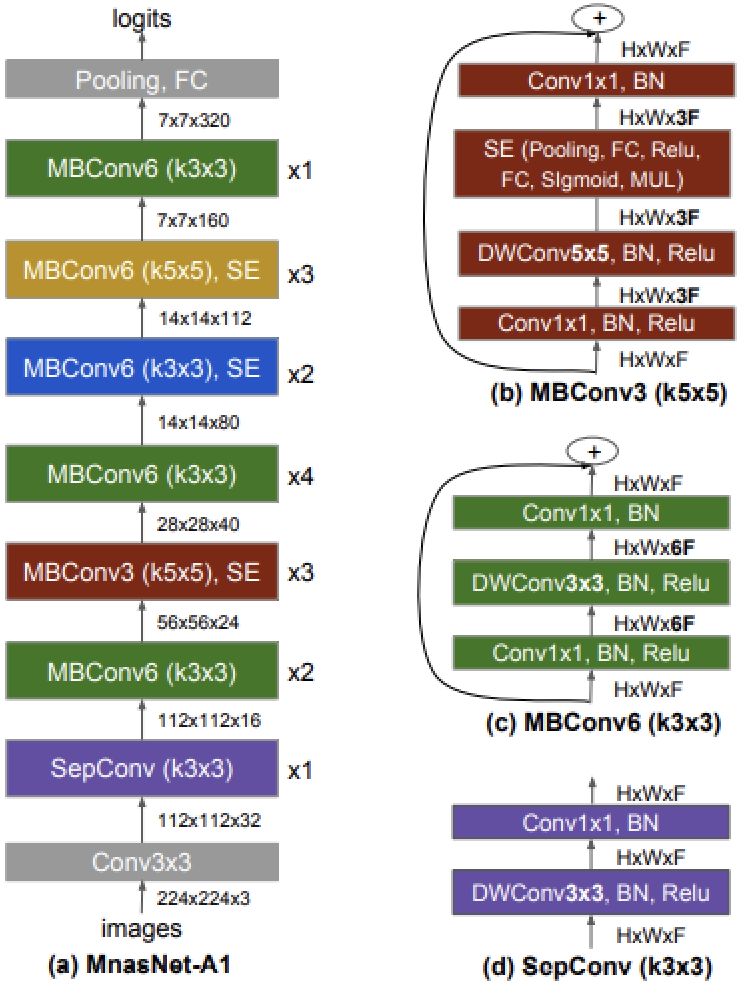


Figure 7: **MnasNet-A1 Architecture** – (a) is a representative model selected from Table 1; (b) - (d) are a few corresponding layer structures. *MBConv* denotes mobile inverted bottleneck conv, *DWConv* denotes depthwise conv, k3x3/k5x5 denotes kernel size, *BN* is batch norm, HxWxF denotes tensor shape (height, width, depth), and $\times 1/2/3/4$ denotes the number of repeated layers within the block.