**Mobilenet** 顧名思義為應用於手機，嵌入式設備而設計的模型，主要克服應用場景為low latency, speed fast目標。在傳統的CNN中，根據模型深度越多，複雜度也越高，參數也隨之增加，而面臨內存不足問題，mobilenet就是為了解決此問題以減少參數而不影響accuracy之下，提出 **depthwise separable convolution** 深度可分離卷積，包含depthwise convolution以及pointwise convolution兩個計算理念，通常大部分模型架構由convolution 和fully connect組成，而Global average pooling(對每個feature map取全局average，輸出node與class類別)取代FC，目的為減少參數、避免overfitting(沒有params)

首先depthwise convolution: 將卷積核拆分為單channel形式，對每channel進行convolution，因此輸入和輸出channel不變

結合pointwise convolution，將input channel融合，對feature map進行升維或是降維，

標準convolution計算量為filter\_wight\*filter\_height\*channel\*filter\_count\*feature\_map\_width\*feature\_map\_height，經過depthwise convolution與pointwise convolution後計算量為filter\_width\*filter\_height\*channel\*feature\_map\_width\*feature\_map\_height + 1\*1\*channel\*filter\_count\*feature\_map\_width\*feature\_map\_height = 相比於mobilenet只有1/N(filter\_count) + 1/(filter\_width\*filter\_height) = 通常使用3x3 filter 約計算量下降到 1/9- 1/8

模型架構部分是3x3 depthwise conv. -> BN -> ReLU6 -> 1x1 Conv. -> BN -> ReLU6，

ReLU6在移動端 float16低精度有很好的數值分辨率，

而有些在特定應用需要更小又快的模型，因此提出兩個alpha寬度因子控制模型大小的分辨因子rho，alpha(width\_Multiplier) 對每層網路input output channel 進行縮減，使用alpha提高1/alpha^2，rho(resolution multiplier)控制feature map分辨率，

-------------------------------------------------------------------------------------------------------

**MobilenetV2** --- MobilenetV2提出，Linear Bottleneck解決depthwise conv. 計算過程input channel和output channel不變，因此若input channel 很少，Depthwise只能在低維度計算，在MobilenetV1的Depthwise前面增加一個pointwise conv. 升維，因此Depthwise獲得了高維feature map，進行特徵獲取，並且將Depthwise conv.後接著的pointwise conv.的ReLU6改成Linear，此時的pointwise作為降維，而ReLU對於channel較低造成大量信息損耗，ReLU在input負值時，output=0，高維後ReLU信息丟失少

MobilenetV2提出了 Inverted Residuals 倒殘差，

殘差最經典就是ResNet，架構以壓縮 pointwise --- 傳統CNN ----擴張 pointwise 架構，目的減少3x3計算量

倒殘差 架構以擴張 pointwise --- depthwise ---- 壓縮pointwise ，增加channel使depthwise能在高維提取特徵

使用short cut 將進行兩者feature map相加，解決傳統上CNN layers 越多 Degradation退化問題，使用Identity Mapping將前一層輸出到後一層想法，避免信息丟失或導致gradient vanishing or explore ，訓練目的為殘差結果趨近於0

**MobileNetV3 -----**

1. MobileNetV1 引入 Depthwise separable convolution
2. MobileNetV2 引入具有Linear Bottleneck and Inverted Residual
3. MnasNet 引入基於squeeze and excitation結構的輕量級注意力模型

**Network Architecture Search(NAS)**

經典使用RNN作為controller產生child network，再對child network進行training and evaluate，得到網路性能(accuracy)，最後更新controller params. 但無法直接對controller進行優化(child network不可導)，透過RL，採用policy gradient方法直接更新controller params.

**Activation function = Swish**



Swish在深層模型上效果優於ReLU，替換後 mobile NASNETA在ImageNet分類準確提高0.9%，Inception-ResNet-v 分類準確提高0.6%

但計算量太大，對swish進行數值近，提出h-swish(hard version of swish)

\*\*\*\*feature map = (in\_w - kernel\_w + 2 \* padding) / stride + 1