**模型的存储问题和模型进行预测的速度问题**

卷积神经网络（简称 CNN）在图像分类、图像分割、目标检测等领域获得广泛应用。随着性能要求越来越高，AlexNet 已经无法满足大家的需求，于是乎各路大牛纷纷提出性能更优越的 CNN 网络，如 VGG、GoogLeNet、ResNet、DenseNet 等。由于神经网络的性质，为了获得更好的性能，网络层数不断增加，从 7 层 AlexNet 到 16 层 VGG，再从 16 层 VGG 到 GoogLeNet 的 22 层，再到 152 层 ResNet，更有上千层的 ResNet 和 DenseNet。虽然网络性能得到了提高，但随之而来的就是效率问题。

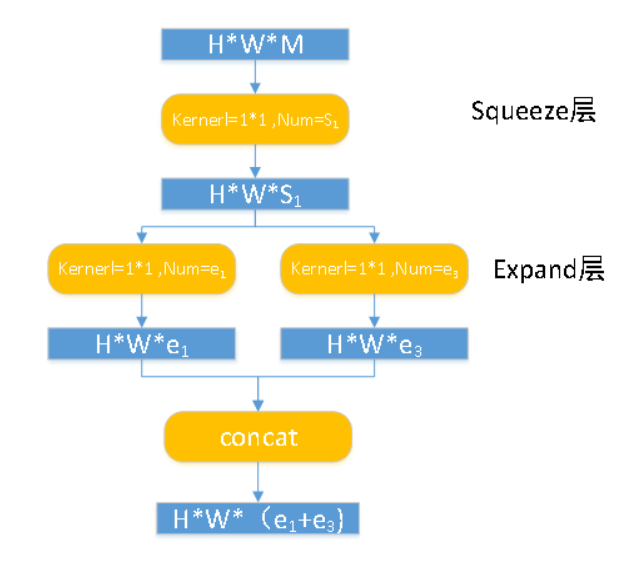
**四个轻量化模型**

SqueezeNet、MobileNet、ShuffleNet、Xception

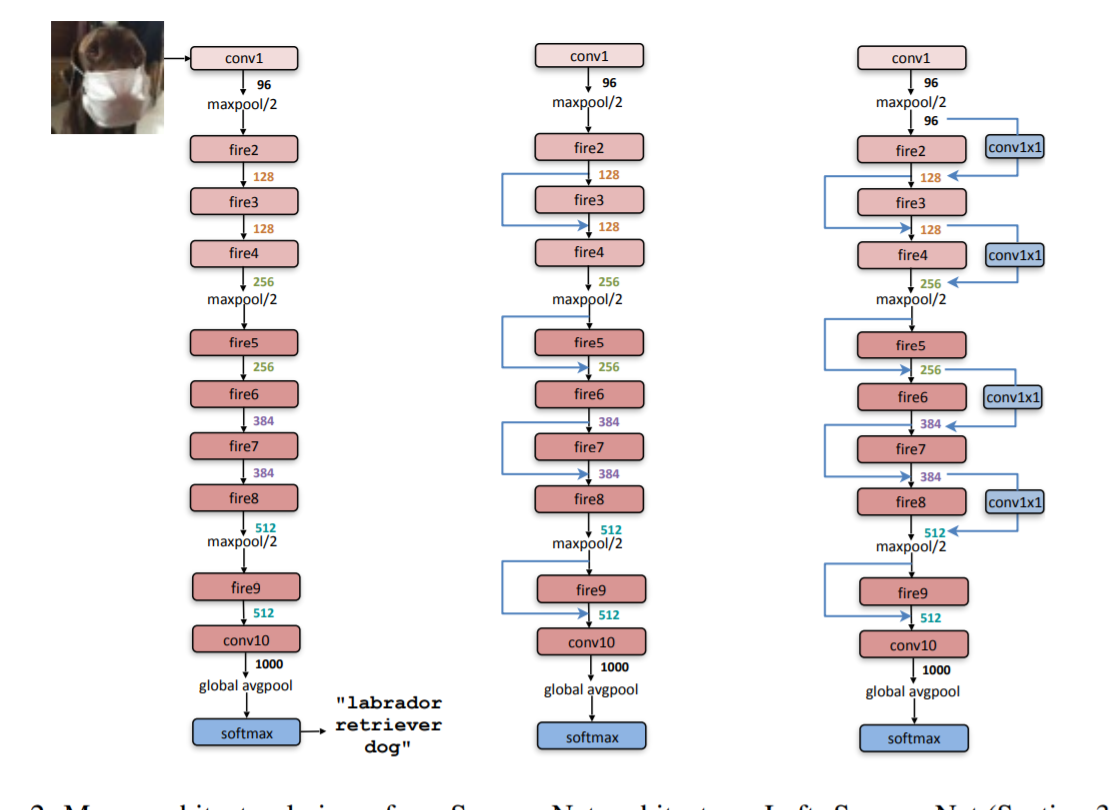
**这四种轻量化模型仅是在卷积方式上做了改变**

**SqueezeNet（SQUEEZENET: ALEXNET-LEVEL ACCURACY WITH 50X FEWER PARAMETERS AND <0.5MB MODEL SIZE）（1DeepScale∗ & UC Berkeley 2Stanford University）**

采用不同于传统的卷积方式，提出 fire module；fire module 包含两部分：squeeze 层+expand 层



**在原文中使用，e1=e3=4s1**

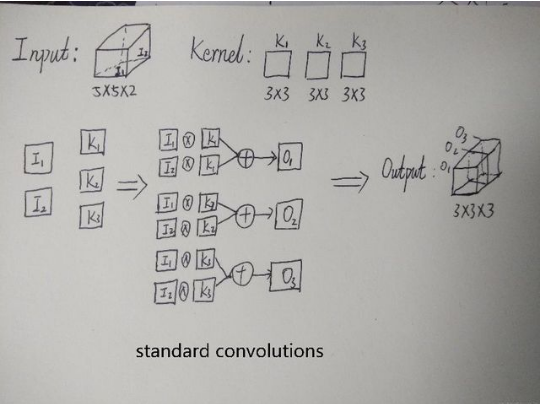
Fire module 与 GoogLeNet 思想类似，采用 1\*1 卷积对 feature map 的维数进行「压缩」，从而达到减少权值参数的目的；

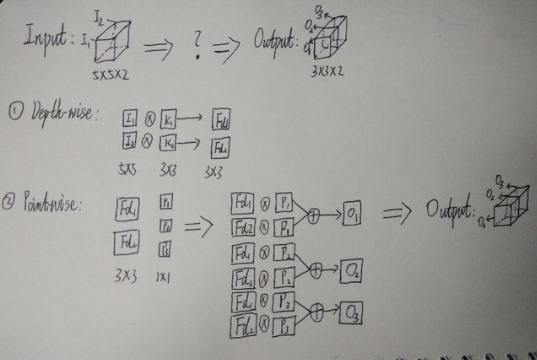
**MobileNet（MobileNets: Efficient Convolutional Neural Networks for Mobile Vision**

**Applications）**

采用名为 **depth-wise separable convolution** 的卷积方式代替传统卷积方式，以达到减少网络权值参数的目的。通过采用 depth-wise convolution 的卷积方式，达到：1. 减少参数数量 2. 提升运算速度。（这两点是要区别开的，参数少的不一定运算速度快！还要看计算方式！）。

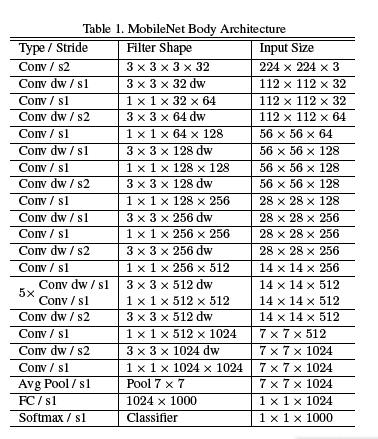
Width Multiplier 和 Resolution Multiplier 这两个**超参**只是方便于设置要网络要设计为多小，方便于量化模型大小。



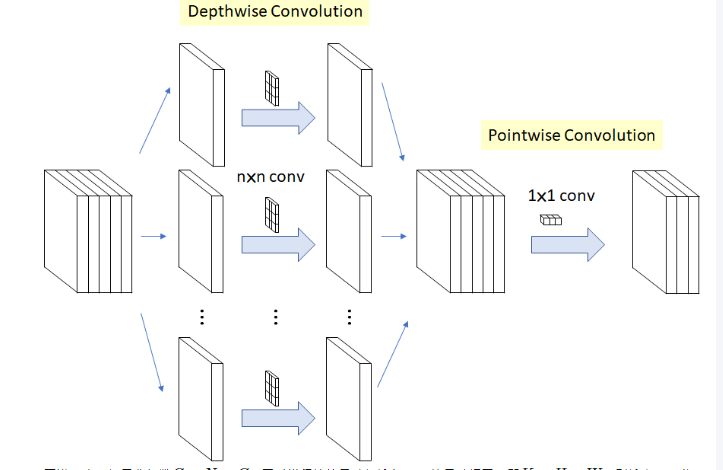


All layers are followed by a batchnorm [13] and ReLU nonlinearity with the exception

of the final fully connected layer which has no nonlinearity and feeds into a softmax layer for classification.



核心思想是采用 depth-wise convolution 操作，在相同的权值参数数量的情况下，相较于 standard convolution 操作，可以减少数倍的计算量，从而达到提升网络运算速度的目的。输出的 feature map 仅包含输入的 feature map 的一部分，在这里，MobileNet 采用了 point-wise convolution 解决这个问题。



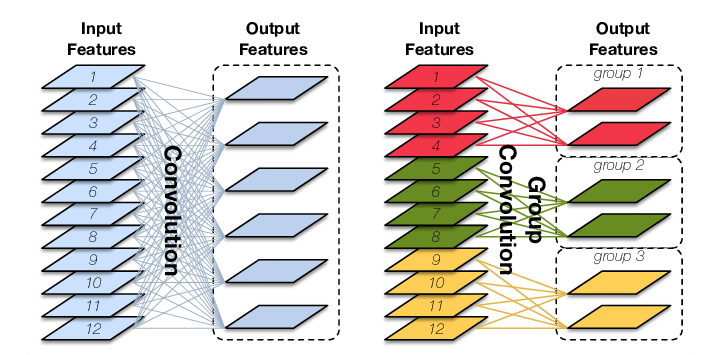
**ShuffleNet（ShuffleNet: An Extremely Efficient Convolutional Neural Network for Mobile**

**Devices Face++）**

shuffle 具体来说是 channel shuffle，是将各部分的 feature map 的 channel 进行有序的打乱，构成新的 feature map，以解决 group convolution 带来的「信息流通不畅」问题。

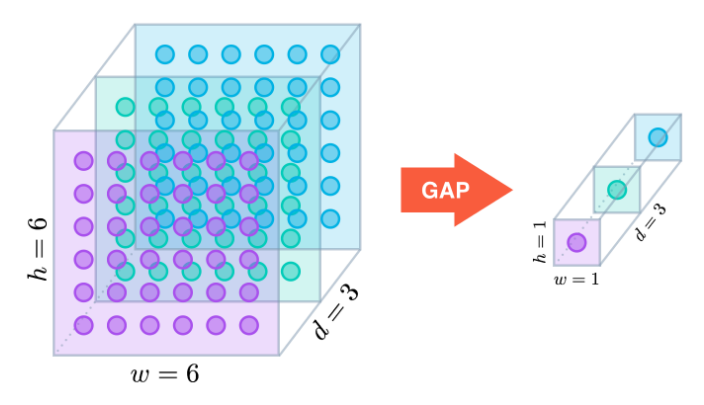
利用 group convolution 和 channel shuffle 这两个操作来设计卷积神经网络模型, 以减少模型使用的参数数量。采用 group convolution 会导致信息流通不当，因此提出 channel shuffle

**group convolution**

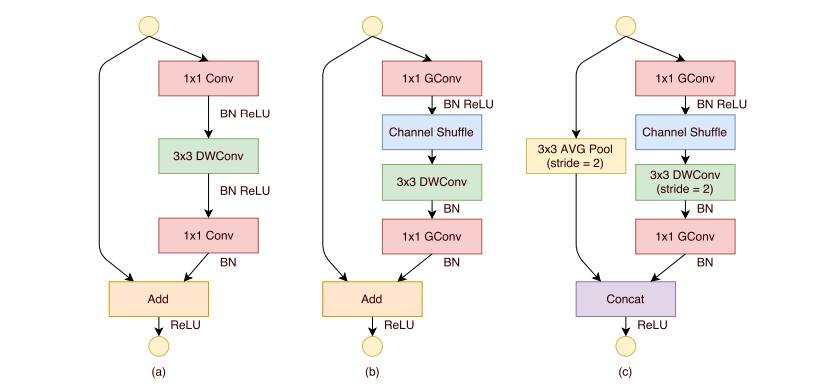


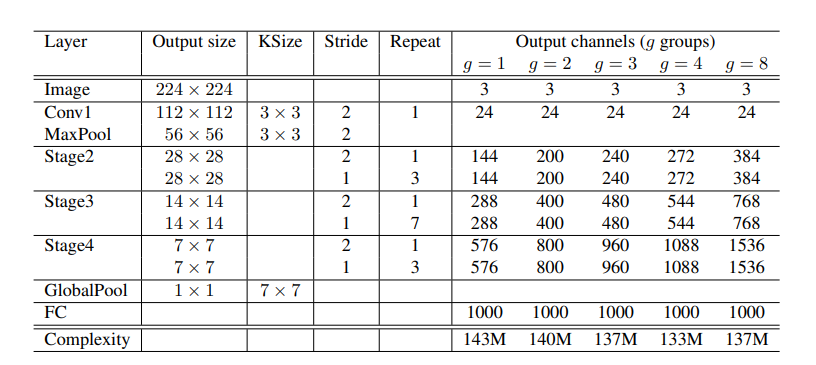
参数相比为传统卷积的1/G（G为分组数）

更进一步，如果分组数G=N=C，同时卷积核的尺寸与输入map的尺寸相同，即K=H=W，则输出map为C∗1∗1即长度为C的向量，此时称之为Global Depthwise Convolution（GDC），见MobileFaceNet，可以看成是全局加权池化，与 Global Average Pooling（GAP） 的不同之处在于，GDC 给每个位置赋予了可学习的权重（对于已对齐的图像这很有效，比如人脸，中心位置和边界位置的权重自然应该不同），而GAP每个位置的权重相同，全局取个平均，如下图所示：



ShuffleNets Units

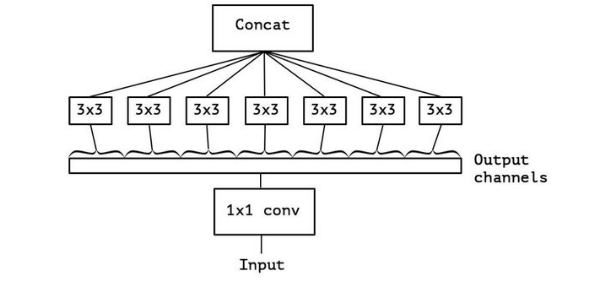




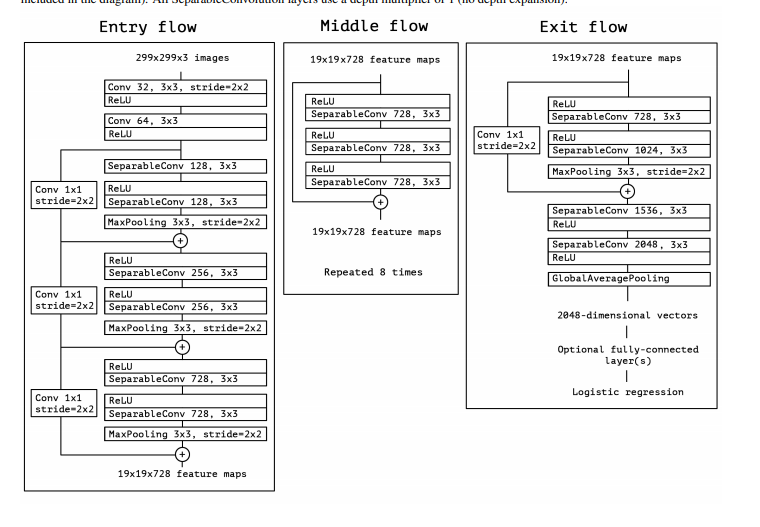
**Xception(Xception: Deep Learning with Depthwise Separable Convolutions,** **Google)**

A depthwise separable convolution can be understood as an Inception module with a maximally large number of towers. This observation leads us to propose a novel deep convolutional neural network architecture inspired by Inception

we make the following hypothesis: that the mapping of cross-channels correlations and spatial correlations in the feature maps of convolutional neural networks can be entirely decoupled

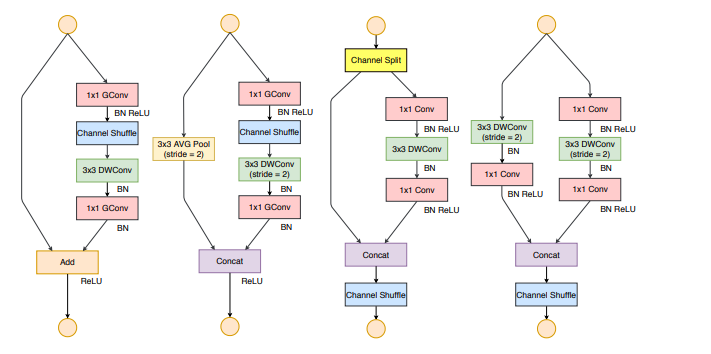


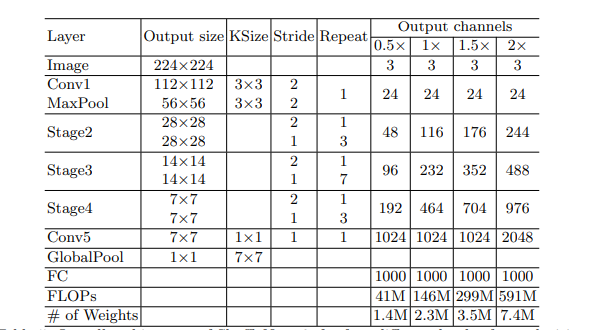
Xception 中Depth-wise与MobileNets中DW的区别在于前者是先通过1x1的卷积进行通道融合，再逐通道卷积



**ShuffleNetV2（Practical Guidelines for Efficient CNN Architecture Design** (Face++**）**

**https://zhuanlan.zhihu.com/p/48261931**





为了改善v1的缺陷，v2版本引入了一种新的运算：channel split。具体来说，在开始时先将输入特征图在通道维度分成两个分支：通道数分别为c1 和 c2 ，实际实现时 c1=c/2 。