# 分组卷积

卷积层是卷积神经网络模型的重要组成模块。现在，卷积神经网络模型的一个主要研究工作，就是在减少参数量和计算量的同时，几乎不损失模型的性能。最直观的方法，当然是减少卷积层的参数量和运算量。Aggregated Residual Transformations for Deep Neural Networks提出了分组卷积，MobileNets：Efficient Convolutional Neural Networks for Mobile Vision Applications提出了深度可分离卷积，都是这方面的工作。

## 1　卷积运算


**图 1**　卷积运算的例子：用“⊛”符号表示卷积运算

对于输入数据，卷积运算以一定间隔滑动滤波器的窗口并应用。这里所说的窗口是指图1中灰色的 3 × 3 的部分。如图1所示，将各个位置上滤波器的元素和输入的对应元素相乘，然后再求和（有时将这个计算称为乘积累加运算）。然后，将这个结果保存到输出的对应位置。将这个过程在所有位置都进行一遍，就可以得到卷积运算的输出。

## 2　3 维卷积运算

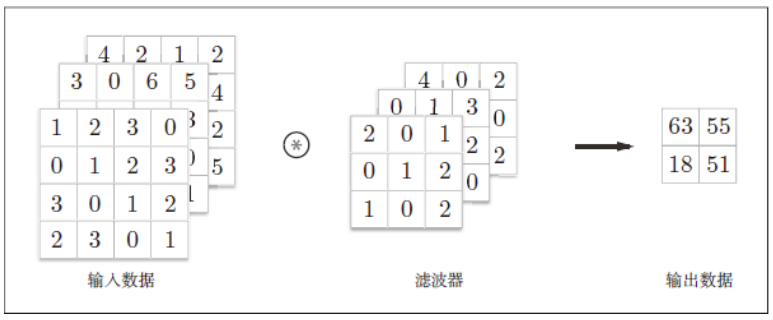



**图 2**　对 3 维数据进行卷积运算的例子

图2是卷积运算的例子，图2是计算顺序。这里以 3 通道的数据为例，展示了卷积运算的结果。和 2 维数据时（图1的例子）相比，可以发现纵深方向（通道方向）上特征图增加了。通道方向上有多个特征图时，会按通道进行输入数据和滤波器的卷积运算，并将结果相加，从而得到输出。

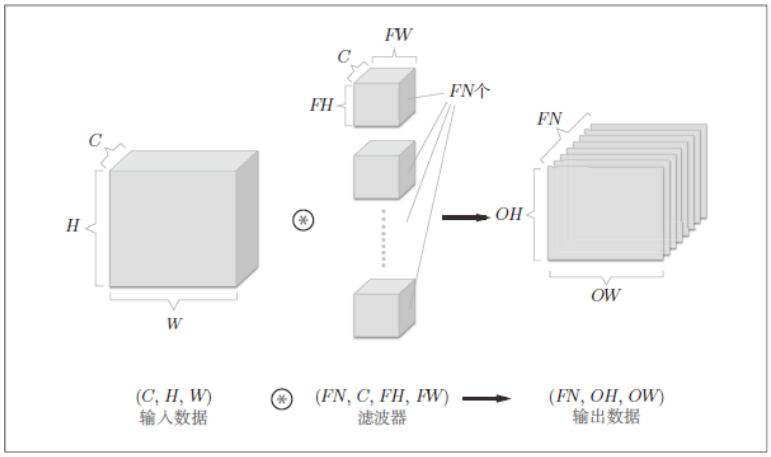



**图 3**　基于多个滤波器的卷积运算的例子

图3中，通过应用 FN 个滤波器，输出特征图也生成了 FN 个。如果将这 FN 个特征图汇集在一起，就得到了形状为 (OH, OW, FN) 的方块。

## 3　分组卷积

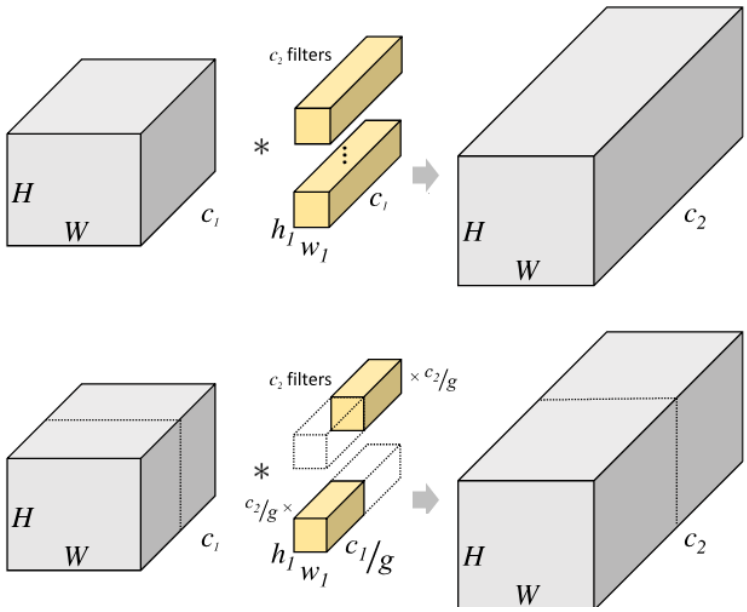



**图 4**　分组卷积

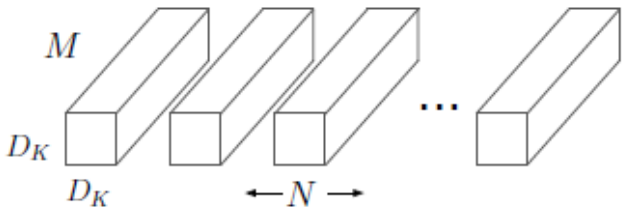
第一张图代表标准卷积。若输入特征图尺寸为 $H \times W \times c_1$，卷积核尺寸为 $h_1 \times w_1 \times c_1$，输出特征图尺寸为 $H \times W \times c_2$，标准卷积的参数量为：

$$P = (h_1 \cdot w_1 \cdot c_1) \cdot c_2 \tag{1}$$

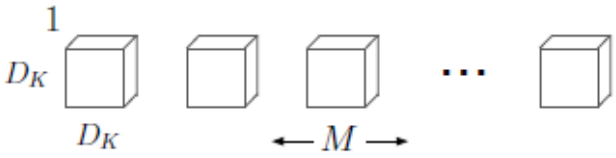
第二张图代表分组卷积。将输入特征图按照通道数分成 $g$ 组，则每组输入特征图的尺寸为 $H \times W \times \frac{c_1}{g}$，对应的卷积核尺寸为 $h_1 \times w_1 \times \frac{c_1}{g}$，每组输出特征图尺寸为 $H \times W \times \frac{c_2}{g}$，最后将 $g$ 组结果拼接起来，那么分组卷积的参数量为：

$$P = (h_1 \cdot w_1 \cdot \frac{c_1}{g}) \cdot \frac{c_2}{g} \cdot g = (h_1 \cdot w_1 \cdot c_1) \cdot c_2 \cdot \frac{1}{g} \tag{2}$$

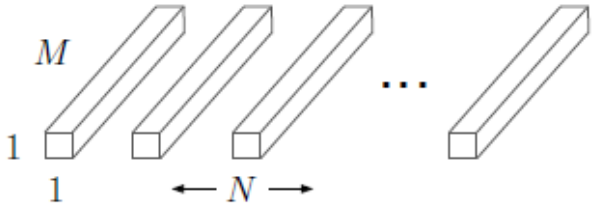
# 4 深度可分离卷积




(a) Standard Convolution Filters




(b) Depthwise Convolutional Filters




(c) $1 \times 1$ Convolutional Filters called Pointwise Convolution in the context of Depthwise Separable Convolution

Figure 2. The standard convolutional filters in (a) are replaced by two layers: depthwise convolution in (b) and pointwise convolution in (c) to build a depthwise separable filter.

**图 5** 深度可分离卷积

图 a 表示标准卷积，参数量为 $D_k \times D_k \times M \times N$。图 b 表示深度卷积，图 c 表示逐点卷积，两者结合起来就是深度可分离卷积。深度卷积负责滤波，逐点卷积负责转换通道，深度可分离卷积的参数量为：

$$P = (D_k \cdot D_k \cdot 1) \cdot M + (1 \cdot 1 \cdot N \cdot M) \tag{3}$$

# 5 附开源代码

20201225[keras] 分组卷积