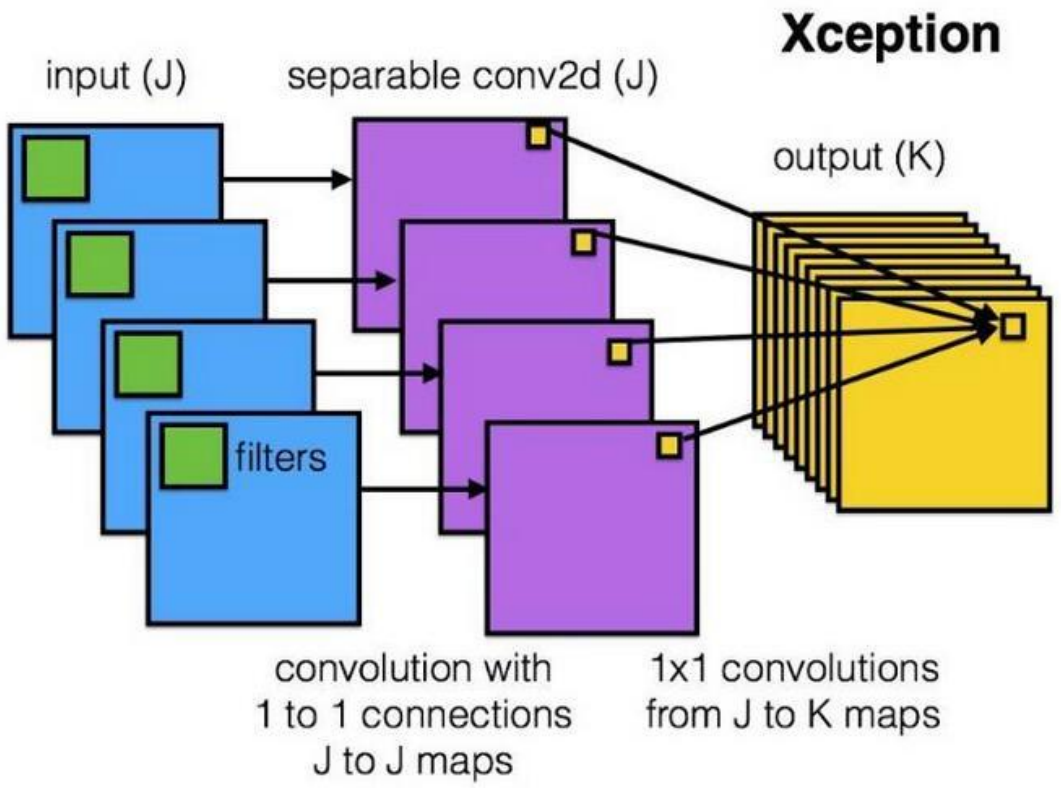


变形卷积核、可分离卷积？卷积神经网络中十大拍案叫绝的操作。

Professor ho · 4 个月前

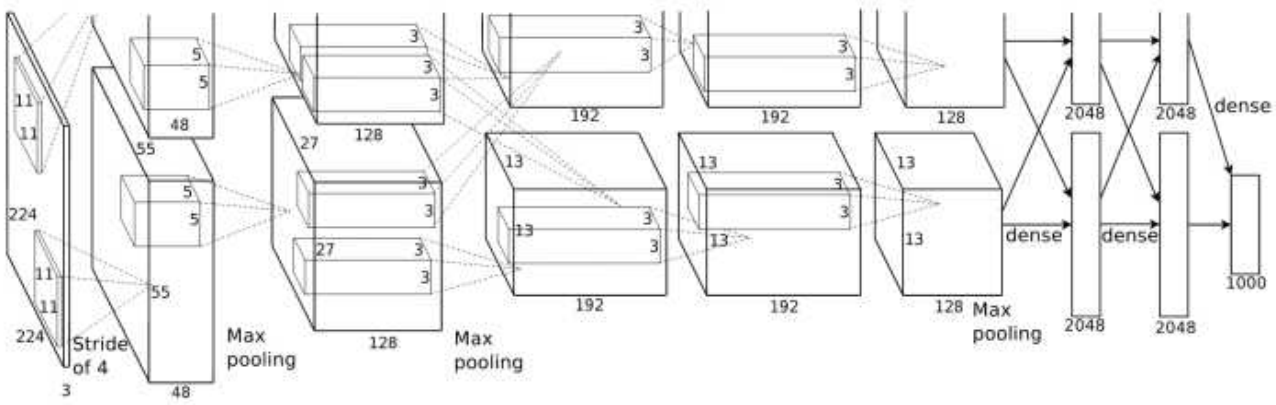
CNN从2012年的AlexNet发展至今，科学家们发明出各种各样的CNN模型，一个比一个深，一个比一个准确，一个比一个轻量。我下面会对近几年一些具有变革性的工作进行简单盘点，从

这些充满革新性的工作中探讨日后的CNN变革方向。

注：水平所限，下面的见解或许有偏差，望大牛指正。另外只介绍其中具有代表性的模型，一些著名的模型由于原理相同将不作介绍，若有遗漏也欢迎指出。

一、卷积只能在同一组进行吗？-- Group convolution

Group convolution 分组卷积，最早在AlexNet中出现，由于当时的硬件资源有限，训练AlexNet时卷积操作不能全部放在同一个GPU处理，因此作者把feature maps分给多个GPU分别进行处理，最后把多个GPU的结果进行融合。



alexnet

分组卷积的思想影响比较深远，当前一些轻量级的SOTA（State Of The Art）网络，都用到了分组卷积的操作，以节省计算量。但题主有个疑问是，如果分组卷积是分在不同GPU上的话，

每个GPU的计算量就降低到 1/groups，但如果依然在同一个GPU上计算，最终整体的计算量是否不变？找了pytorch上有关组卷积操作的介绍，望读者解答我的疑问。

```
:attr:`groups` controls the connections between inputs and outputs.
`in_channels` and `out_channels` must both be divisible by `groups`.
        At groups=1, all inputs are convolved to all outputs.
        At groups=2, the operation becomes equivalent to having two conv
             layers side by side, each seeing half the input channels,
             and producing half the output channels, and both subsequently
             concatenated.
        At groups=`in_channels`, each input channel is convolved with its
             own set of filters (of size `out_channels // in_channels`).
```

pytroch github

EDIT：

关于这个问题，知乎用户朋友 @蔡冠羽 提出了他的见解：

> 我感觉group conv本身应该就大大减少了参数，比如当input channel为256，output channel也为256，kernel size为3*3，不做group conv参数为256*3*3*256，若group为8，每个group的input channel和output channel均为32，参数为8*32*3*3*32，是原来的八分之一。这是我的理解。

我的理解是分组卷积最后每一组输出的feature maps应该是以concatenate的方式组合，而不是element-wise add，所以每组输出的channel是 input channels / #groups，这样参数量就大大减少了。

二、卷积核一定越大越好？-- 3×3卷积核

AlexNet中用到了一些非常大的卷积核，比如11×11、5×5卷积核，之前人们的观念是，卷积核越大，receptive field（感受野）越大，看到的图片信息越多，因此获得的特征越好。虽说如此，但是大的卷积核会导致计算量的暴增，不利于模型深度的增加，计算性能也会降低。于是在VGG（最早使用）、Inception网络中，利用2个3×3卷积核的组合比1个5×5卷积核的效果更佳，同时参数量（3×3×2+1 VS 5×5×1+1）被降低，因此后来3×3卷积核被广泛应用在各种模型中。

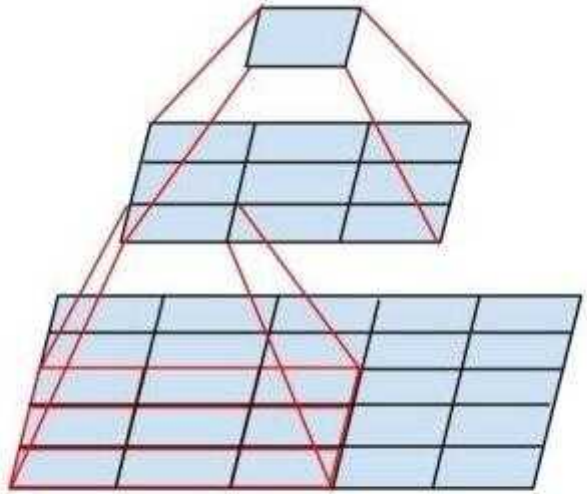
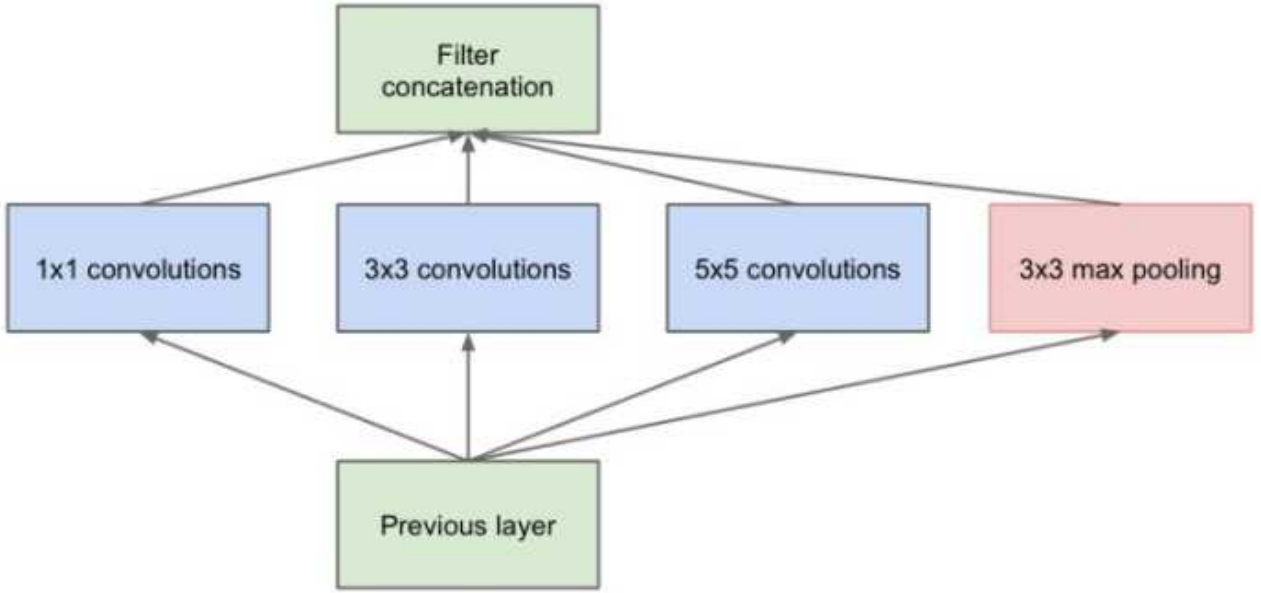


Figure 1. Mini-network replacing the 5×5 convolutions.

三、每层卷积只能用一种尺寸的卷积核？-- Inception结构

传统的层叠式网络，基本上都是一个个卷积层的堆叠，每层只用一个尺寸的卷积核，例如VGG结构中使用了大量的3×3卷积层。事实上，同一层feature map可以分别使用多个不同尺寸的卷积核，以获得不同尺度的特征，再把这些特征结合起来，得到的特征往往比使用单一卷积核的要好，谷歌的GoogleNet，或者说Inception系列的网络，就使用了多个卷积核的结构：

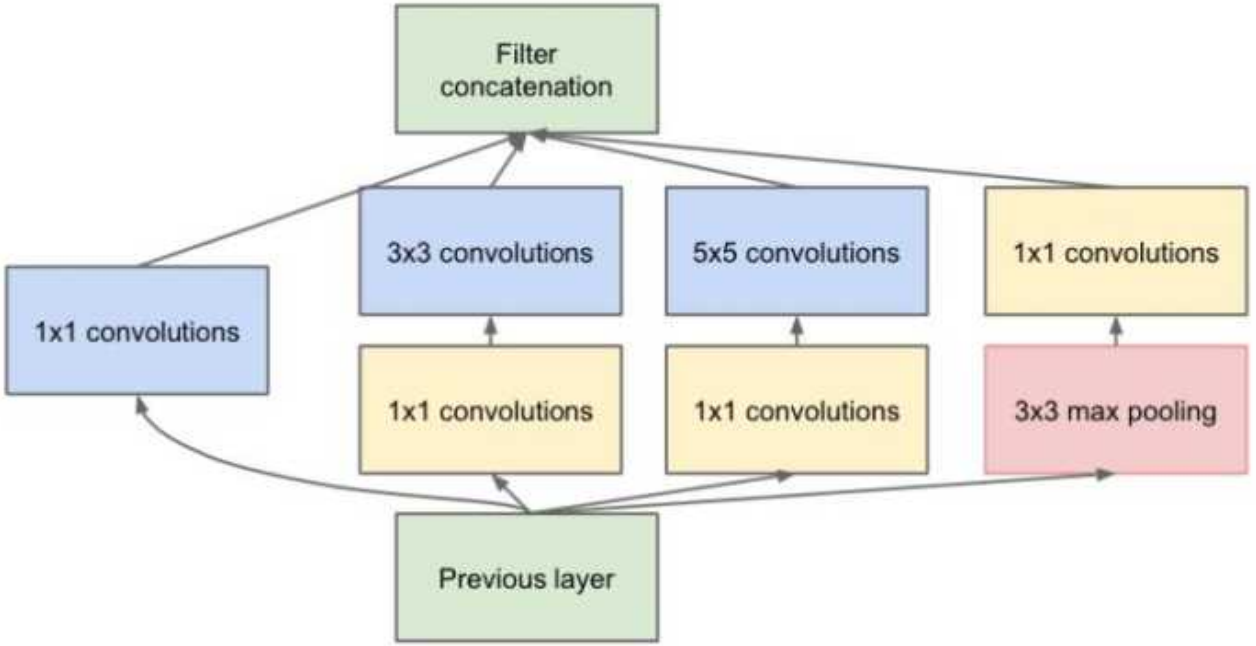


最初版本的Inception结构

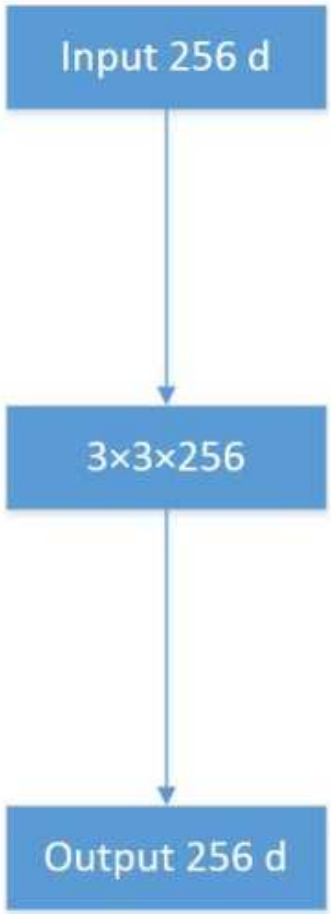
如上图所示，一个输入的feature map分别同时经过1×1、3×3、5×5的卷积核的处理，得出的特征再组合起来，获得更佳的特征。但这个结构会存在一个严重的问题：参数量比单个卷积核要多很多，如此庞大的计算量会使得模型效率低下。这就引出了一个新的结构：

四、怎样才能减少卷积层参数量？-- Bottleneck

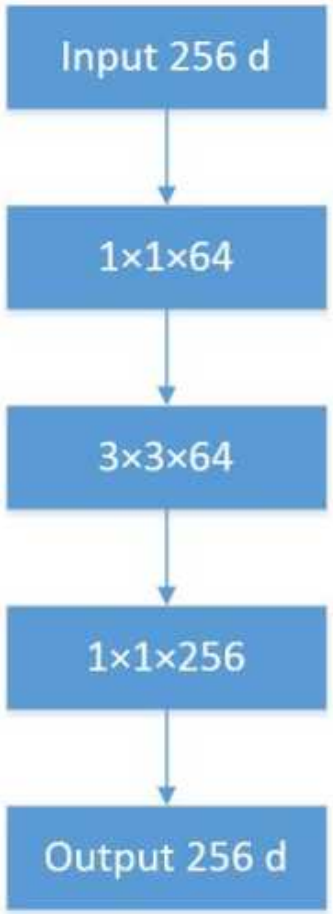
发明GoogleNet的团队发现，如果仅仅引入多个尺寸的卷积核，会带来大量的额外的参数，受到Network In Network中1×1卷积核的启发，为了解决这个问题，他们往Inception结构中加入了一些1×1的卷积核，如图所示：



加入1×1卷积核的Inception结构



第一种



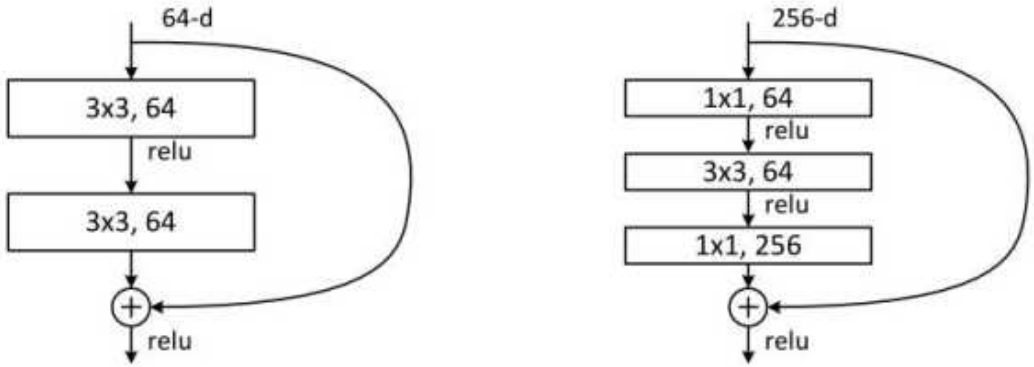
第二种

根据上图，我们来做个对比计算，假设输入feature map的维度为256维，要求输出维度也是256维。有以下两种操作：

1. 256维的输入直接经过一个3×3×256的卷积层，输出一个256维的feature map，那么参数量为：256×3×3×256 = 589,824
2. 256维的输入先经过一个1×1×64的卷积层，再经过一个3×3×64的卷积层，最后经过一个1×1×256的卷积层，输出256维，参数量为：256×1×1×64 + 64×3×3×64 + 64×1×1×256 = 69,632。足足把第一种操作的参数量降低到九分之一！

1×1卷积核也被认为是影响深远的操作，往后大型的网络为了降低参数量都会应用上1×1卷积核。

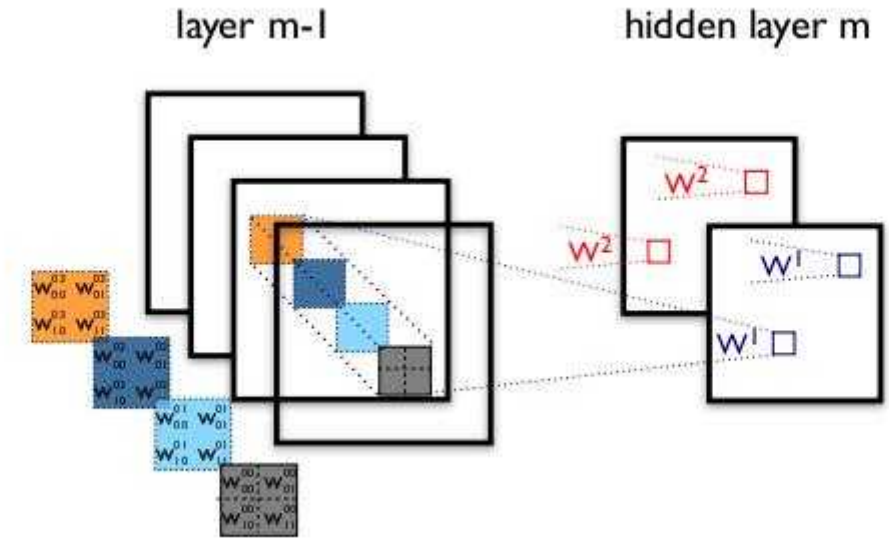
五、越深的网络就越难训练吗？-- Resnet残差网络



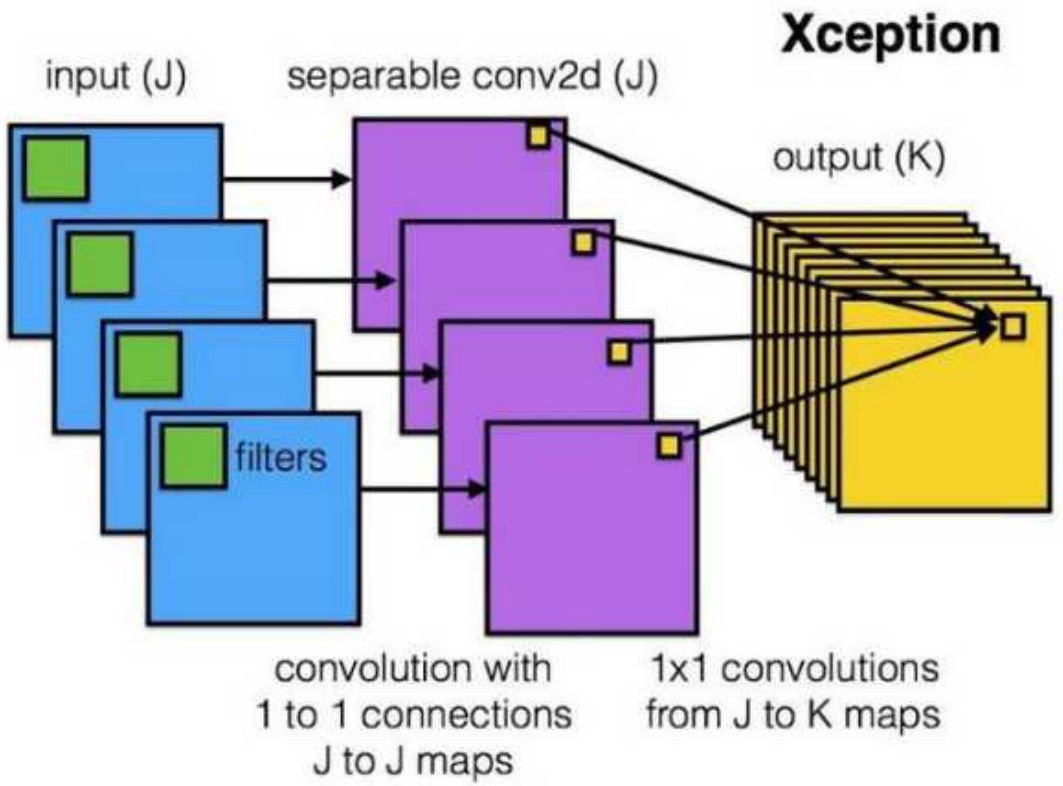
ResNet skip connection

传统的卷积层层叠网络会遇到一个问题，当层数加深时，网络的表现越来越差，很大程度上的原因是因为当层数加深时，梯度消散得越来越严重，以至于反向传播很难训练到浅层的网络。为了解决这个问题，何凯明大神想出了一个“残差网络”，使得梯度更容易地流动到浅层的网络当中去，而且这种“skip connection”能带来更多的好处，这里可以参考一个PPT：极深网络（ResNet/DenseNet）: Skip Connection为何有效及其它 ，以及我的一篇文章：为什么ResNet和DenseNet可以这么深？一文详解残差块为何能解决梯度弥散问题。 ，大家可以结合下面的评论进行思考。

六、卷积操作时必须同时考虑通道和区域吗？-- DepthWise操作



标准的卷积过程可以看上图，一个2×2的卷积核在卷积时，对应图像区域中的所有通道均被同时考虑，问题在于，为什么一定要同时考虑图像区域和通道？我们为什么不能把通道和空间区域分开考虑？



Xception网络就是基于以上的问题发明而来。我们首先对每一个通道进行各自的卷积操作，有多少个通道就有多少个过滤器。得到新的通道feature maps之后，这时再对这批新的通道feature maps进行标准的1×1跨通道卷积操作。这种操作被称为 **“DepthWise convolution”** ，缩写“DW”。

这种操作是相当有效的，在imagenet 1000类分类任务中已经超过了InceptionV3的表现，而且也同时减少了大量的参数，我们来算一算，假设输入通道数为3，要求输出通道数为256，两种做法：

1.直接接一个3×3×256的卷积核，参数量为：3×3×3×256 = 6,912

2.DW操作，分两步完成，参数量为：3×3×3 + 3×1×1×256 = 795，又把参数量降低到九分之一！

因此，一个depthwise操作比标准的卷积操作降低不少的参数量，同时论文中指出这个模型得到了更好的分类效果。

EDIT：2017.08.25

本文在发出12小时后，一位知乎用户私信了我，向我介绍了Depthwise和Pointwise的历史工作，而Xception和Mobilenet也引用了他们16年的工作，就是Min Wang et al 的Factorized Convolutional Neural Networks，这篇论文的Depthwise中，每一通道输出的feature map（称为“基层”）可以不止一个，而Xception中的Depthwise separable Convolution， 正是这篇工作中“单一基层”的情况。推荐有兴趣的读者关注下他们的工作，这里有篇介绍博文：【深度学习】卷积层提速Factorized Convolutional Neural Networks。而最早关于separable convolution的介绍，Xception作者提到，应该追溯到Lau- rent Sifre 2014年的工作 Rigid-Motion Scattering For Image Classification 6.2章节。

七、分组卷积能否对通道进行随机分组？-- ShuffleNet

在AlexNet的Group Convolution当中，特征的通道被平均分到不同组里面，最后再通过两个全连接层来融合特征，这样一来，就只能在最后时刻才融合不同组之间的特征，对模型的泛化性是相当不利的。为了解决这个问题，ShuffleNet在每一次层叠这种Group conv层前，都进行一次channel shuffle，shuffle过的通道被分配到不同组当中。进行完一次group conv之后，再一次channel shuffle，然后分到下一层组卷积当中，以此循环。

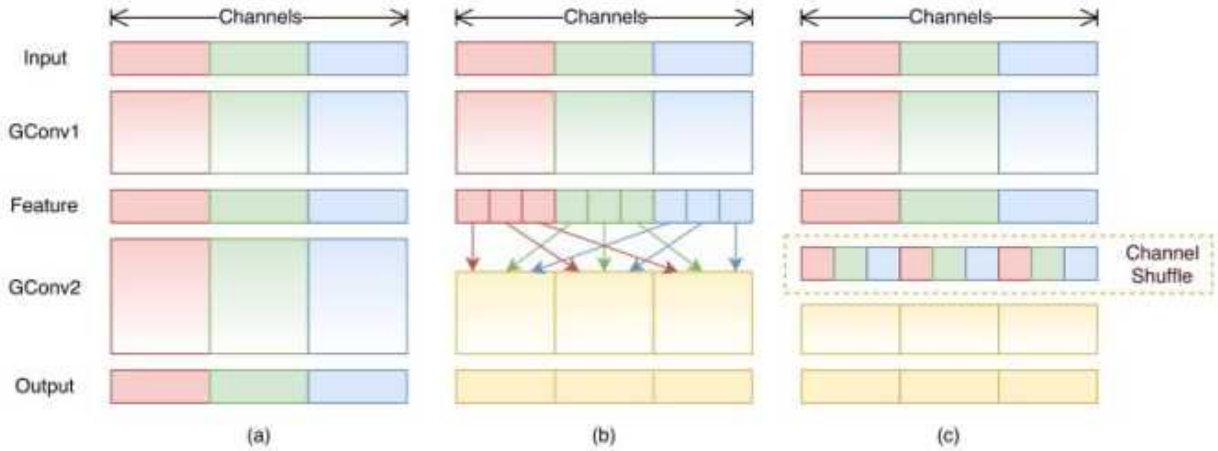


Figure 1: Channel shuffle with two stacked group convolutions. GConv stands for group convolution. a) two stacked convolution layers with the same number of groups. Each output channel only relates to the input channels within the group. No cross talk; b) input and output channels are fully related when GConv2 takes data from different groups after GConv1; c) an equivalent implementation to b) using channel shuffle.

来自ShuffleNet论文

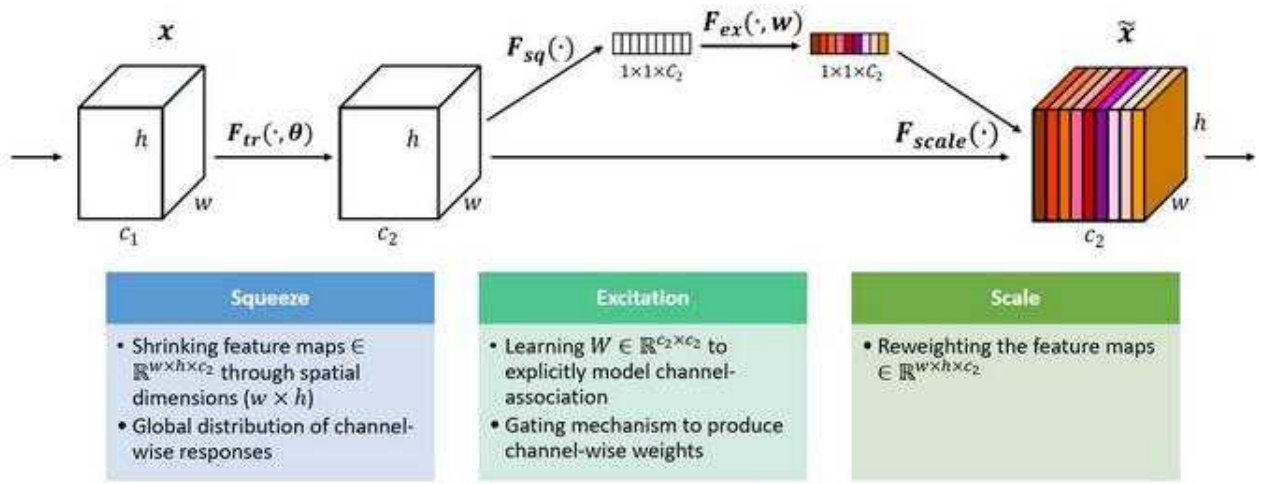
经过channel shuffle之后，Group conv输出的特征能考虑到更多通道，输出的特征自然代表性就更高。另外，AlexNet的分组卷积，实际上是标准卷积操作，而在ShuffleNet里面的分组卷积操作是depthwise卷积，因此结合了通道洗牌和分组depthwise卷积的ShuffleNet，能得到超少量的参数以及超越mobilenet、媲美AlexNet的准确率！

另外值得一提的是，微软亚洲研究院MSRA最近也有类似的工作，他们提出了一个IGC单元（Interleaved Group Convolution），即通用卷积神经网络交错组卷积，形式上类似进行了两次组卷积，Xception 模块可以看作交错组卷积的一个特例，特别推荐看看这篇文章：王井东详解ICCV 2017入选论文：通用卷积神经网络交错组卷积

要注意的是，**Group conv**是一种**channel**分组的方式，**Depthwise +Pointwise**是卷积的方式，只是**ShuffleNet**里面把两者应用起来了。因此**Group conv**和**Depthwise +Pointwise**并不能划等号。

八、通道间的特征都是平等的吗？ -- SEnet

无论是在Inception、DenseNet或者ShuffleNet里面，我们对所有通道产生的特征都是不分权重直接结合的，那为什么要认为所有通道的特征对模型的作用就是相等的呢？ 这是一个好问题，于是，ImageNet2017 冠军SEnet就出来了。

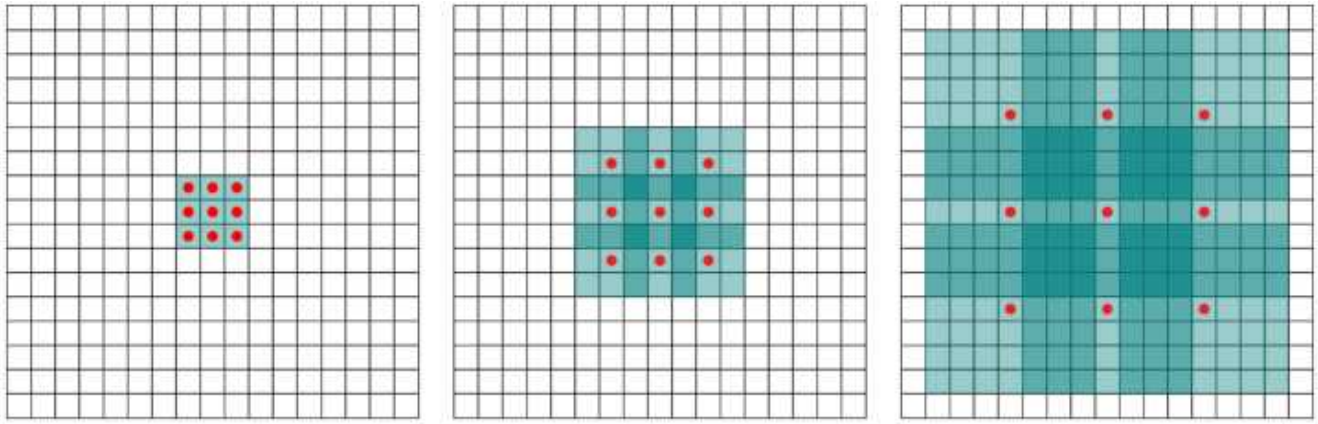


SEnet 结构

一组特征在上一层被输出，这时候分两条路线，第一条直接通过，第二条首先进行Squeeze操作（Global Average Pooling），把每个通道2维的特征压缩成一个1维，从而得到一个特征通道向量（每个数字代表对应通道的特征）。然后进行Excitation操作，把这一列特征通道向量输入两个全连接层和sigmoid，建模出特征通道间的相关性，得到的输出其实就是每个通道对应的权重，把这些权重通过Scale乘法通道加权到原来的特征上（第一条路），这样就完成了特征通道的权重分配。作者详细解释可以看这篇文章：专栏 | Momenta详解ImageNet 2017夺冠架构SENet

九、能否让固定大小的卷积核看到更大范围的区域？-- Dilated convolution

标准的3×3卷积核只能看到对应区域3×3的大小，但是为了能让卷积核看到更大的范围，dilated conv使其成为了可能。dilated conv原论文中的结构如图所示：



上图b可以理解为卷积核大小依然是3×3，但是每个卷积点之间有1个空洞，也就是在绿色7×7区域里面，只有9个红色点位置作了卷积处理，其余点权重为0。这样即使卷积核大小不变，但它看到的区域变得更大了。详细解释可以看这个回答：如何理解空洞卷积（dilated convolution）？

十、卷积核形状一定是矩形吗？-- Deformable convolution 可变形卷积核

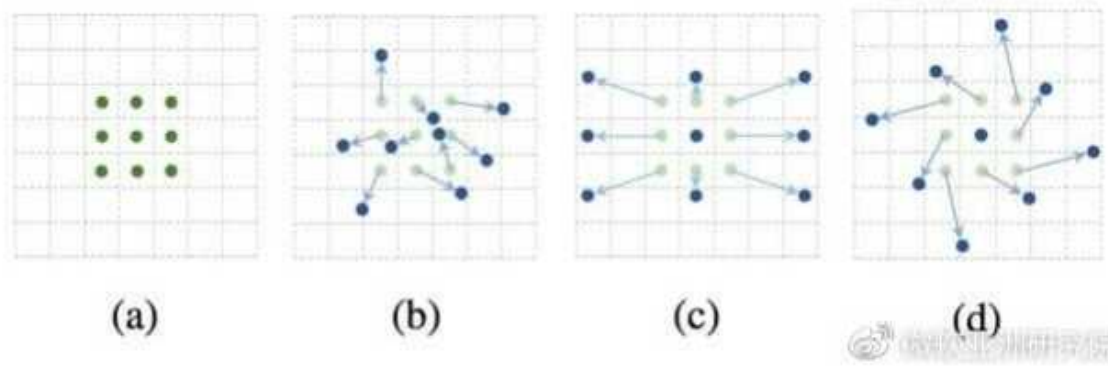


图1：展示了卷积核大小为 3x3 的正常卷积和可变形卷积的采样方式，(a) 所示的正常卷积规律的采样 9 个点（绿点），(b)(c)(d) 为可变形卷积，在正常的采样坐标上加上一个位移量（蓝色箭头），其中(c)(d) 作为 (b) 的特殊情况，展示了可变形卷积可以作为尺度变换，比例变换和旋转变换的特殊情况

图来自微软亚洲研究院公众号

传统的卷积核一般都是长方形或正方形，但MSRA提出了一个相当反直觉的见解，认为卷积核的形状可以是变化的，变形的卷积核能让它只看感兴趣的图像区域 ，这样识别出来的特征更佳。

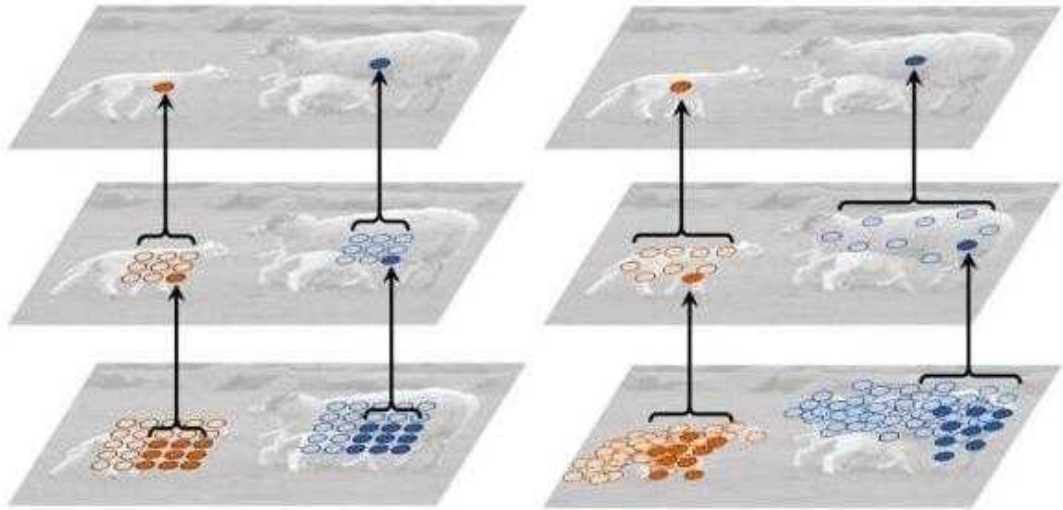


图2：两层3*3的标准卷积和可变形卷积的区别。(a) 标准卷积中固定的感受野和卷积核采样点。(b) 可变性卷积中自适应的感受野和卷积核采样点。

图来自微软亚洲研究院公众号要做到这个操作，可以直接在原来的过滤器前面再加一层过滤器，这层过滤器学习的是下一层卷积核的位置偏移量（offset），这样只是增加了一层过滤器，或者直接把原网络中的某一层过滤器当成学习offset的过滤器，这样实际增加的计算量是相当少的，但能实现可变形卷积核，识别特征的效果更好。详细MSRA的解读可以看这个链接：可变形卷积网络：计算机新“视”界。

启发与思考

现在越来越多的CNN模型从巨型网络到轻量化网络一步步演变，模型准确率也越来越高。现在工业界追求的重点已经不是准确率的提升（因为都已经很高了），都聚焦于速度与准确率的trade off，都希望模型又快又准。因此从原来AlexNet、VGGnet，到体积小一点的Inception、Resnet系列，到目前能移植到移动端的mobilenet、ShuffleNet（体积能降低到0.5mb！），我们可以看到这样一些趋势：

卷积核方面：

1. 大卷积核用多个小卷积核代替；
2. 单一尺寸卷积核用多尺寸卷积核代替；
3. 固定形状卷积核趋于使用可变形卷积核；
4. 使用1×1卷积核（bottleneck结构）。

卷积层通道方面：

1. 标准卷积用depthwise卷积代替；
2. 使用分组卷积；
3. 分组卷积前使用channel shuffle；
4. 通道加权计算。

卷积层连接方面：

1. 使用skip connection，让模型更深；
2. densely connection，使每一层都融合上其它层的特征输出（DenseNet）

启发

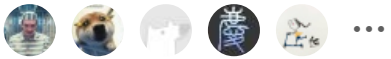
类比到通道加权操作，卷积层跨层连接能否也进行加权处理？bottleneck + Group conv + channel shuffle + depthwise的结合会不会成为以后降低参数量的标准配置？

如果你有更多的想法或意见，欢迎评论留言，好的idea值得交流传播。另外本人的简书号是：人工智豪 - 简书 ，简书上会发一些比较技术性的文章，如GPU降温等，知乎上会发比较理论性的见解文章，欢迎关注。

卷积神经网络（CNN）

深度学习（Deep Learning）

计算机视觉



猫的薛定谔

请问

256维的输入先经过一个1×1×64的卷积层，再经过一个3×3×64的卷积层，最后经过一个3×3×256的卷积层，输出256维，参数量为：256×1×1×64 + 64×3×3×63 + 64×1×1×256 = 69,632

这块是不是写错了？为什么不是256×1×1×64 + 64×3×3×64 + 64×3×3×256 呢

4 个月前

猫的薛定谔

两种操作：

256维的输入直接经过一个3×3×256的卷积层，输出一个256维的feature map，那么参数量为：256×3×3×256 = 589,824256维的输入先经过一个1×1×64的卷积层，再经过一个3×3×64的卷积层，最后经过一个3×3×256的卷积层，输出256维，参数量为：256×1×1×64 + 64×3×3×63 + 64×1×1×256 = 69,632。足足把第一种操作的参数量降低到九分之一！

请问如何保证方法二在减少计算量的同时不会影响卷积效果

4 个月前

Professor ho（作者） 回复 猫的薛定谔 查看对话

抱歉，编辑的时候漏了一张图，有些数字写错了，现已更正，谢谢你的指出。

4 个月前

Professor ho（作者） 回复 猫的薛定谔 查看对话

卷积效果换个说法是特征提取效果如何，这个可以用channel数体现，第一种方法只有256个channel，第二种方法经过了64+64+256=384个channel，通道数多了，representation更

佳，也就是卷积效果更好了。

4 个月前

pby5 回复 **Professor ho**（作者） 查看对话

我觉得group conv好的效果在于后面还跟了一个1x1 conv，1x1 conv 弥补了之前3x3 group conv 中所没有的不同channel之间的信息融合

4 个月前 1 赞

Professor ho（作者） 回复 **pby5** 查看对话

嗯，depthwise然后pointwise，信息融合了

4 个月前

taigw

列一下参考文献就更好了

4 个月前 2 赞

奥修特白

Mark

4 个月前

James Liu

现已加入拍案叫绝系列套餐

4 个月前 2 赞

NAUYL

貌似两个3x3的比5x5的卷积和好，是VGG提出来的，印象中是这样，，，

4 个月前

1 2 3 4 5 下一页