

深度学习- 卷积神经网络

Guangrui Qian

图像感知



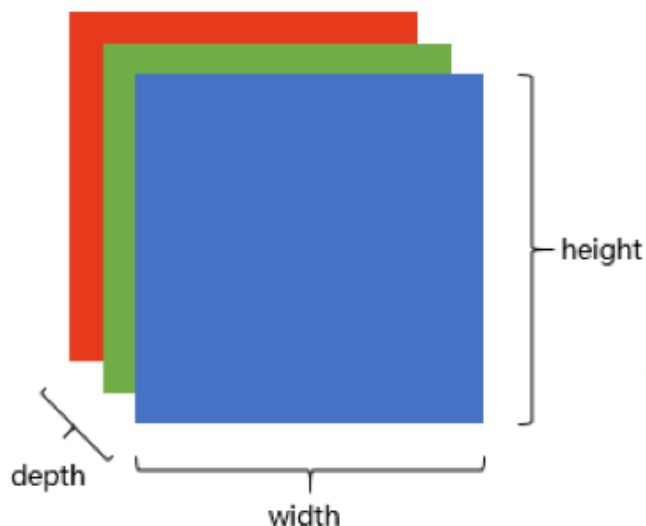
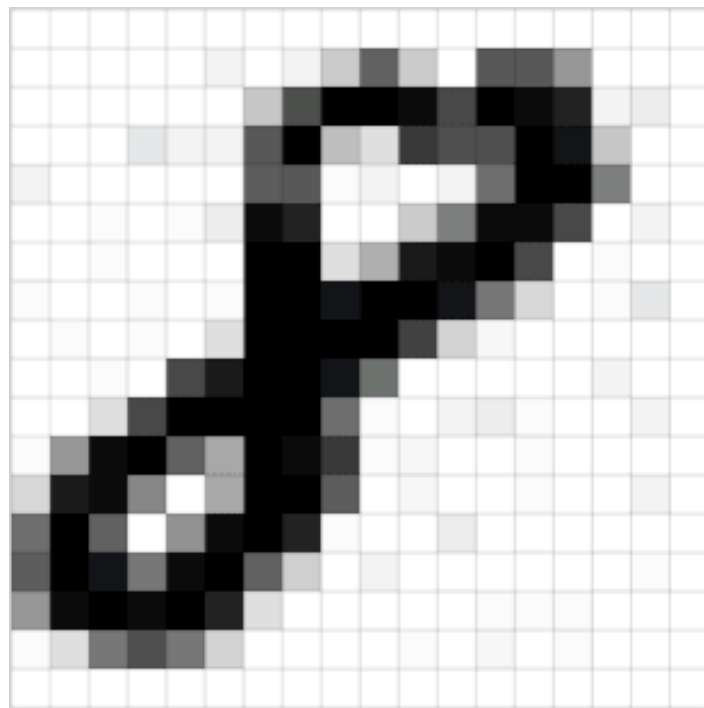
相同的图片经过不同的视觉系统，也会得到不同的感知



卷积神经网络最初是服务于画面识别

图像表达

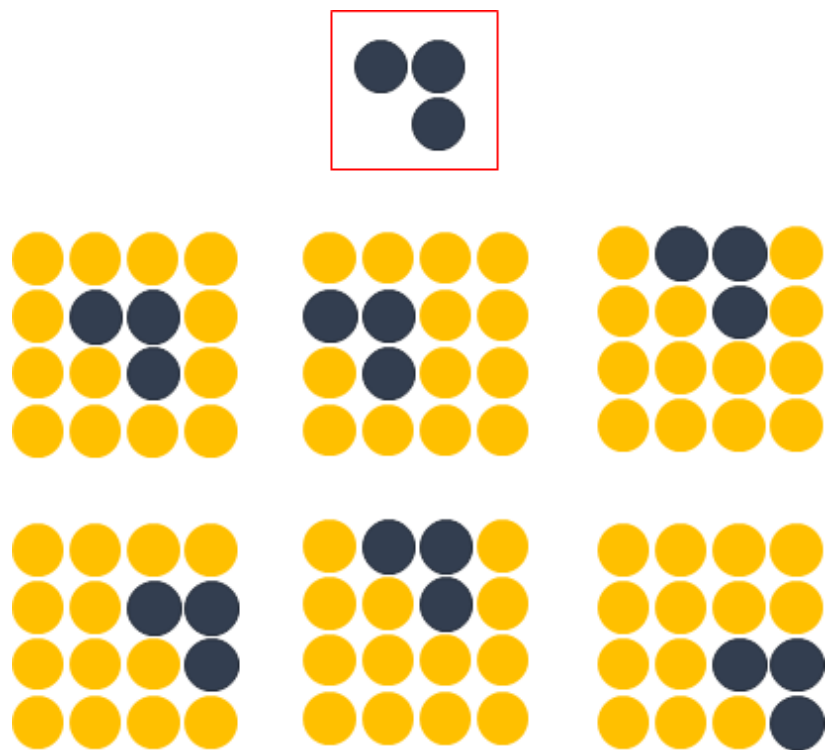
- 图像在计算机中是一堆按顺序排列的数字，数值为0到255。0表示最暗，255表示最亮。你可以把这堆数字用一个长长的向量来表示，
- 灰度图只有黑白颜色，而更普遍的图片表达方式是RGB颜色模型即红（Red）、绿（Green）、蓝（Blue）三原色的色光以不同的比例相加，以产生多种多样的色光。
- RGB颜色模型中，单个矩阵就扩展成了有序排列的三个矩阵，也可以用三维张量去理解，其中的每一个矩阵又叫这个图片的一个channel。
- 在电脑中，一张图片是数字构成的“长方体”。可用宽width, 高height, 深depth 来描述。



画面识别的输入x是shape为
(width, height, depth)的
三维张量

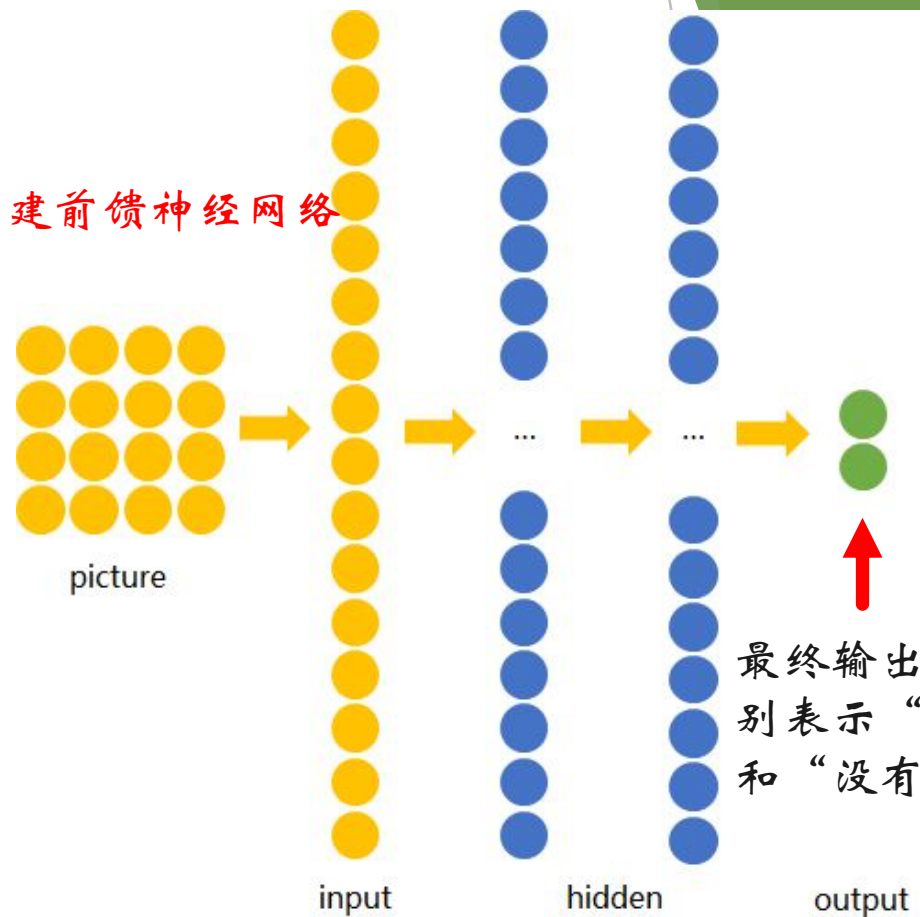
[illegible]

图片识别--前馈神经网络 - 1



想要完成的任务是：在宽长为4x4的图片中识别是否有下图所示的“横折”。

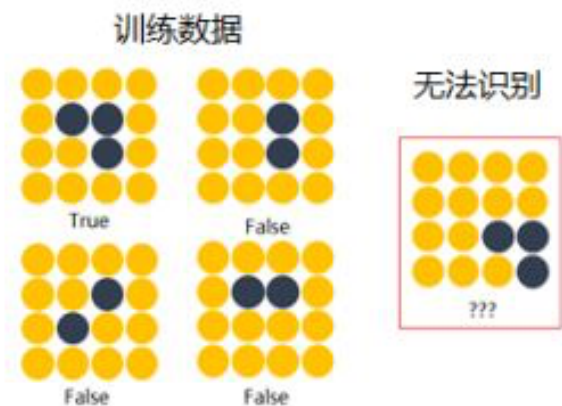
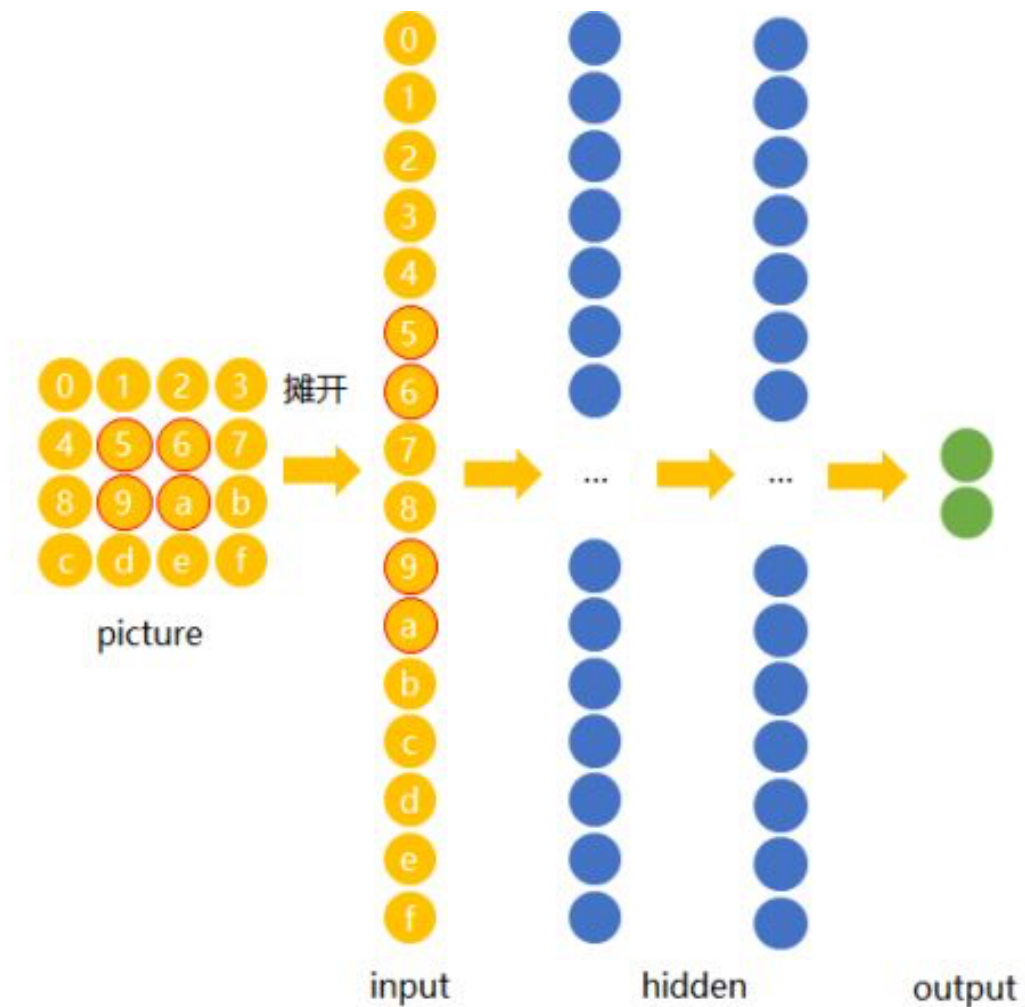
构建前馈神经网络



最终输出两个节点，分别表示“有横折的概率”和“没有横折的概率”。

表达图像的三维张量将会被摊平成一个向量，作为网络的输入，即(width, height, depth)为(4, 4, 1)的图片会被展成维度为16的向量作为网络的输入层。

图片识别--前馈神经网络 - 2



下面我们用数字（16进制）对图片中的每一个像素点（pixel）进行编号。当使用右侧那种物体位于中间的训练数据来训练网络时，**网络就只会对编号为5,6,9,a的节点的权重进行调节**。若让该网络识别位于右下角的“横折”时，则无法识别。

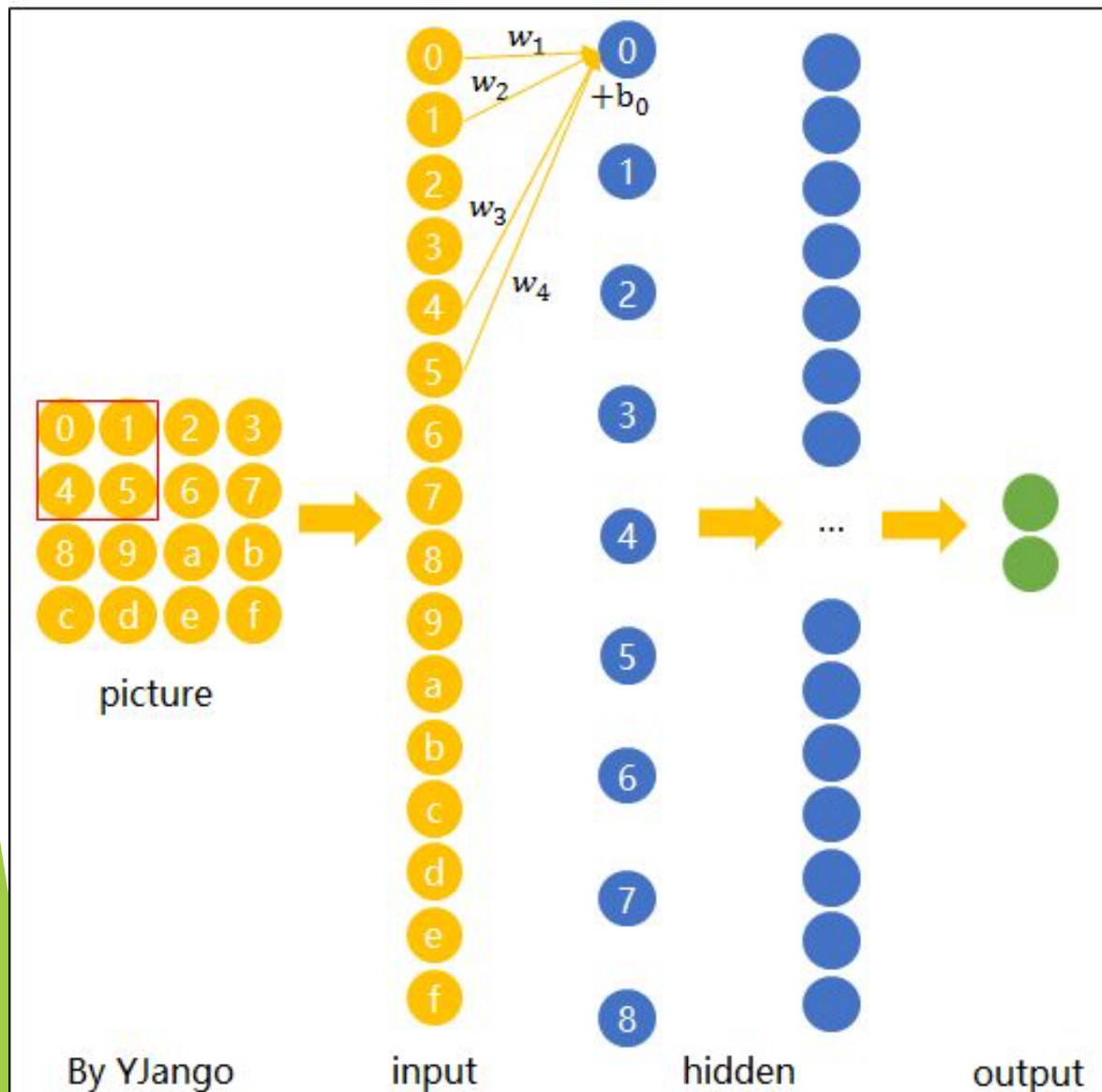
权重共享图片识别解决办法

- 用大量物体位于不同位置的数据训练，同时增加网络的隐藏层个数从而扩大网络学习这些变体的能力。
- 构建神经网络，让相同的東西在位置变了之后不需要重新学习，将中间所学到的规律也运用在其他的位置。换句话说，也就是**让不同位置用相同的权重**。

卷积神经网络

卷积神经网络就是让权重在不同位置共享的神经网络。

图片识别--卷积神经网络 - 局部连接



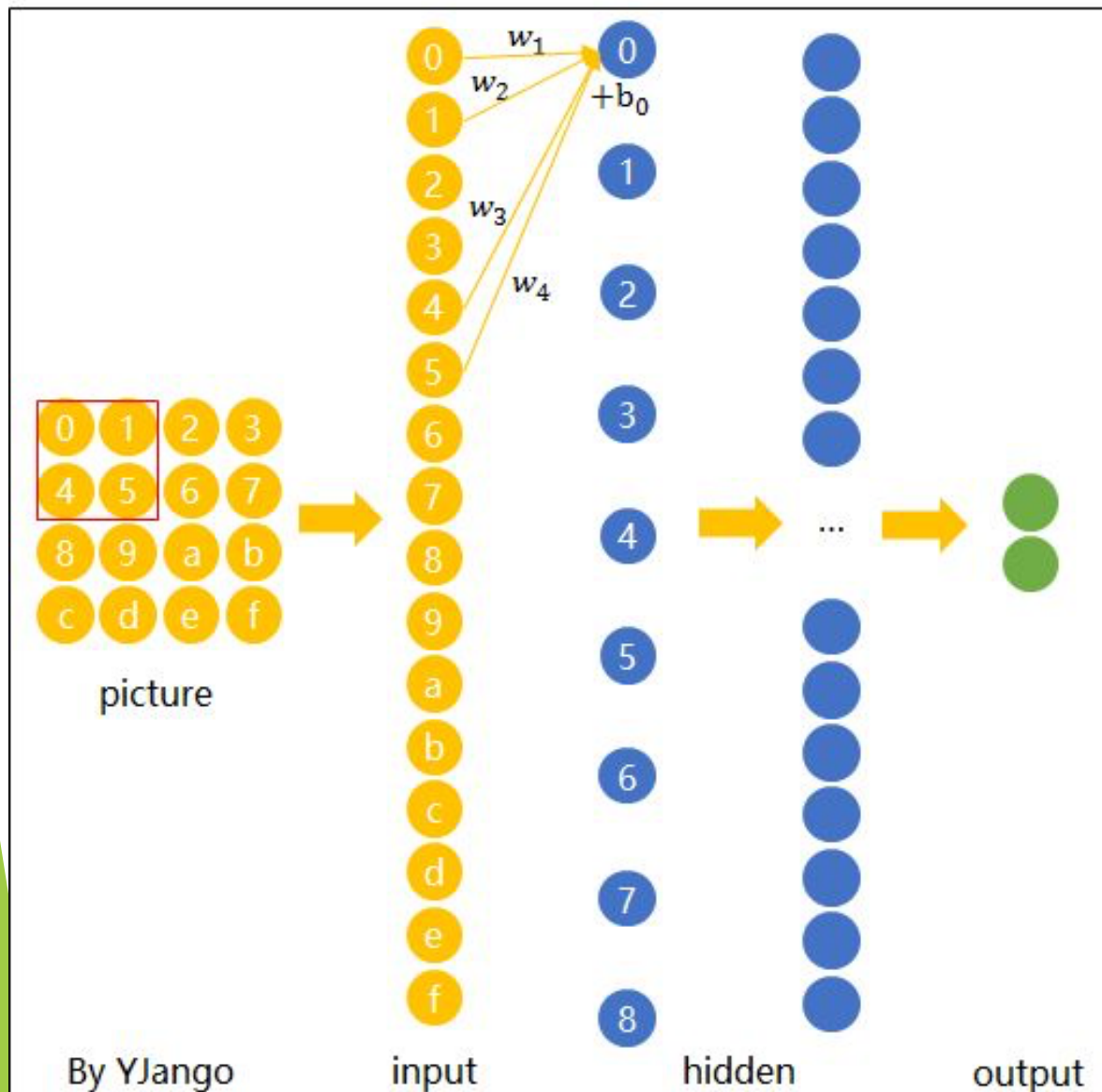
在卷积神经网络中，先选择一个局部区域，用这个局部区域去扫描整张图片。局部区域所圈起来的所有节点会被连接到下一层的一个节点上。

为了更好的和前馈神经网络做比较，将这些以矩阵排列的节点展成了向量。左图展示了被红色方框所圈中编号为0,1,4,5的节点是如何通过 连接到下一层的节点0上的。

$$\begin{bmatrix} w_1 & w_2 \\ w_3 & w_4 \end{bmatrix}$$

这个带有连接强弱的红色方框就叫做 **filter** 或 **kernel** 或 **feature detector**。而filter的范围叫做**filter size**，这里所展示的是2x2的filter size。

图片识别--卷积神经网络 - 局部连接

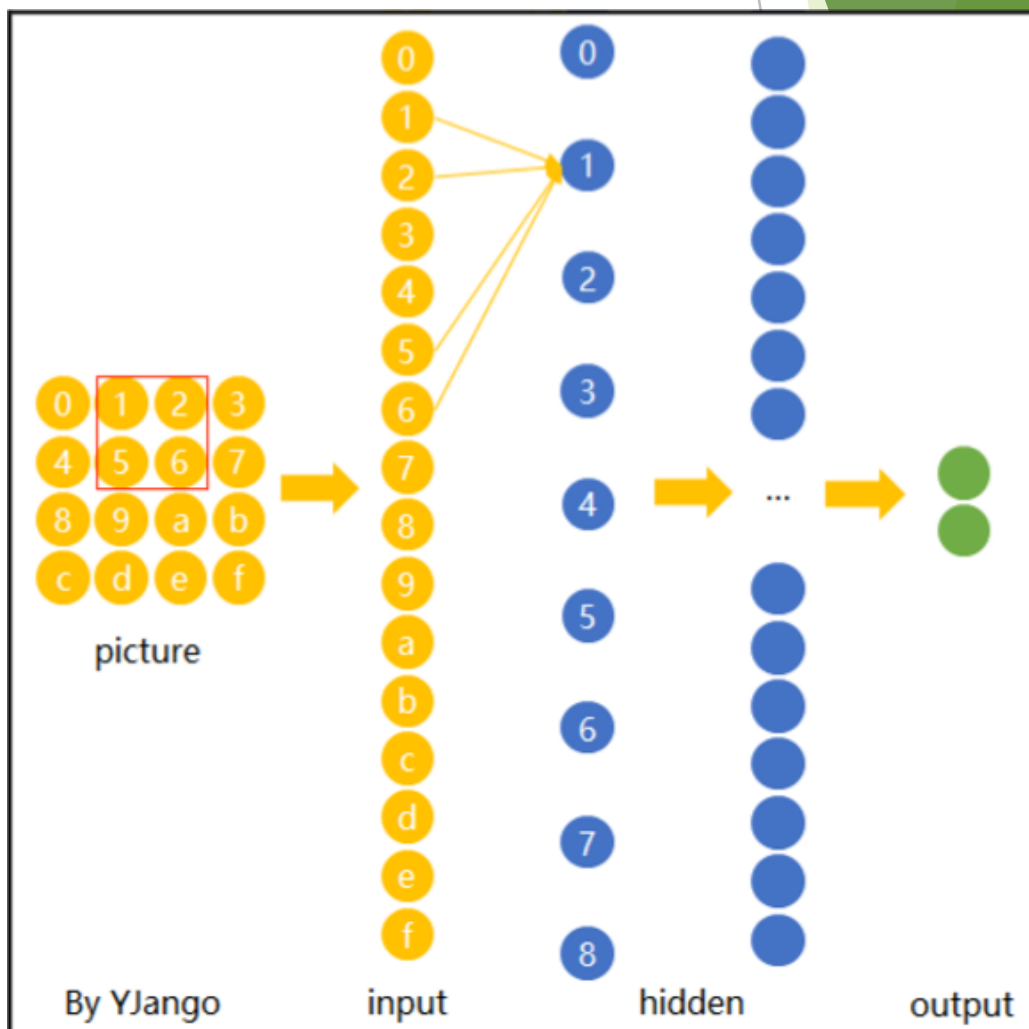
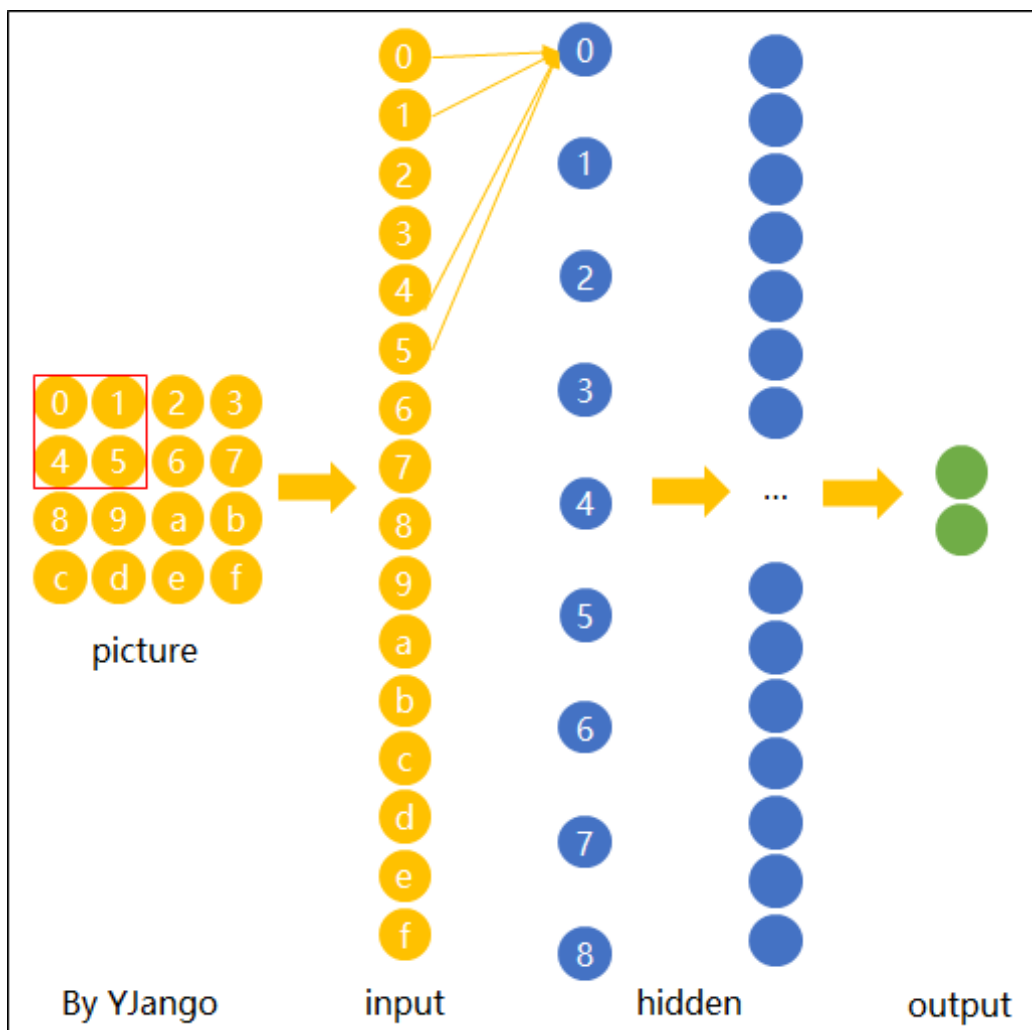


第二层的节点0的数值就是局部区域的线性组合，即被圈中节点的数值乘以对应的权重后相加。用 x 表示输入值， y 表示输出值，用图中标注数字表示角标，则下面列出了两种计算编号为0的输出值 y_0 的表达式。

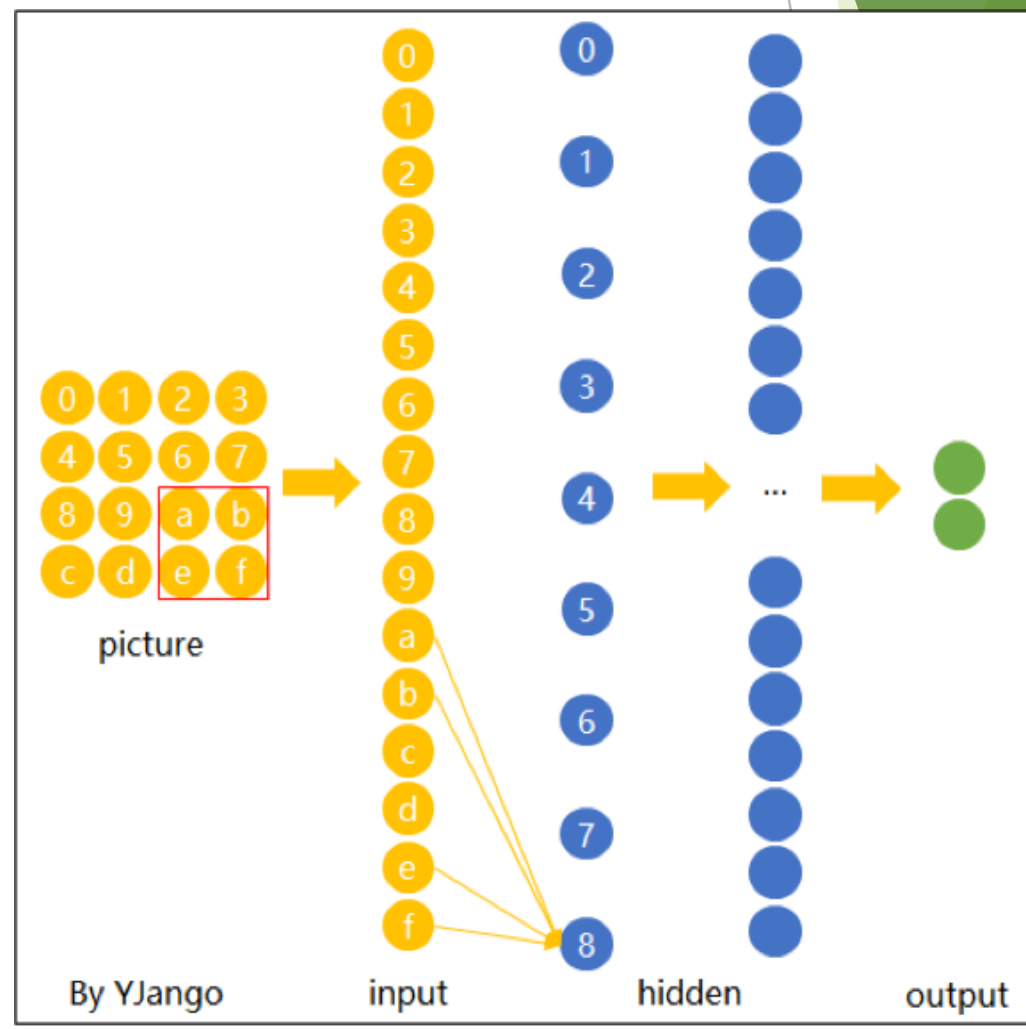
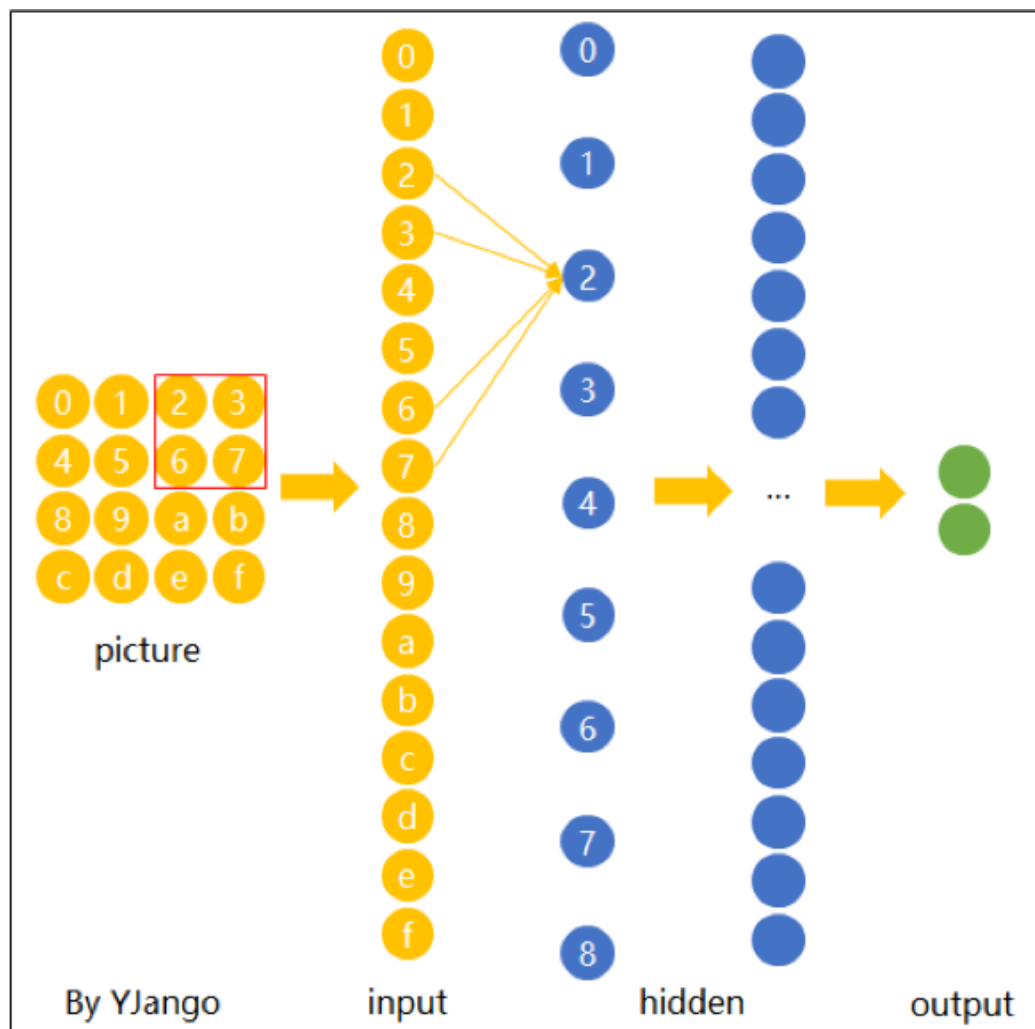
注：在局部区域的线性组合后，也会和前馈神经网络一样，加上一个偏移量 b_0 。

$$y_0 = x_0 * w_1 + x_1 * w_2 + x_4 * w_3 + x_5 * w_4 + b_0$$
$$y_0 = [w_1 \quad w_2 \quad w_3 \quad w_4] \cdot \begin{bmatrix} x_0 \\ x_1 \\ x_4 \\ x_5 \end{bmatrix} + b_0 \quad (2)$$

图片识别--卷积神经网络 - 空间共享1



图片识别--卷积神经网络 - 空间共享2



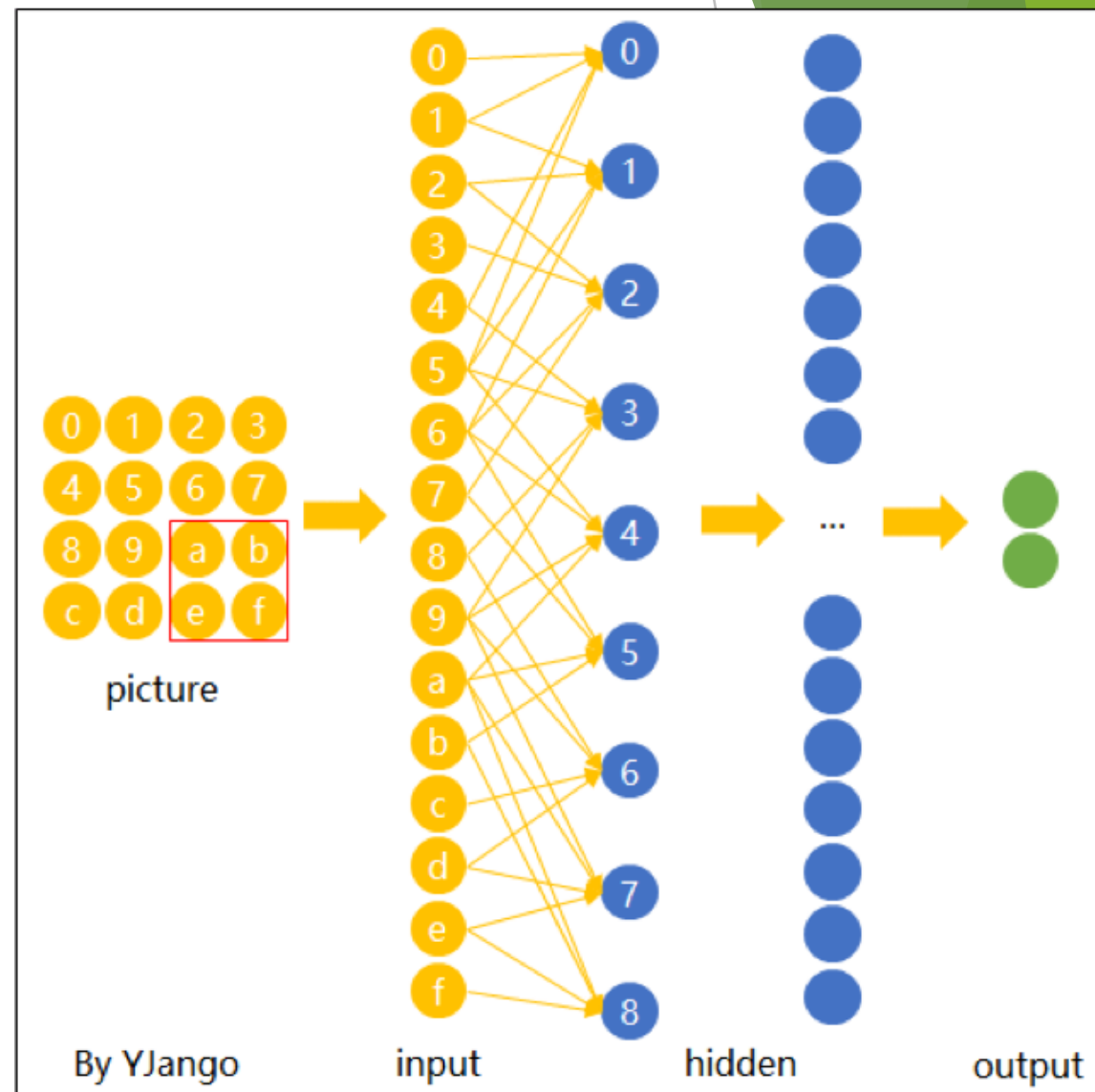
图片识别--卷积神经网络 - 空间共享3

当filter扫到其他位置计算输出节点 y_i 时，

w_1, w_2, w_3, w_4 ，包括 b_0 是共用的。

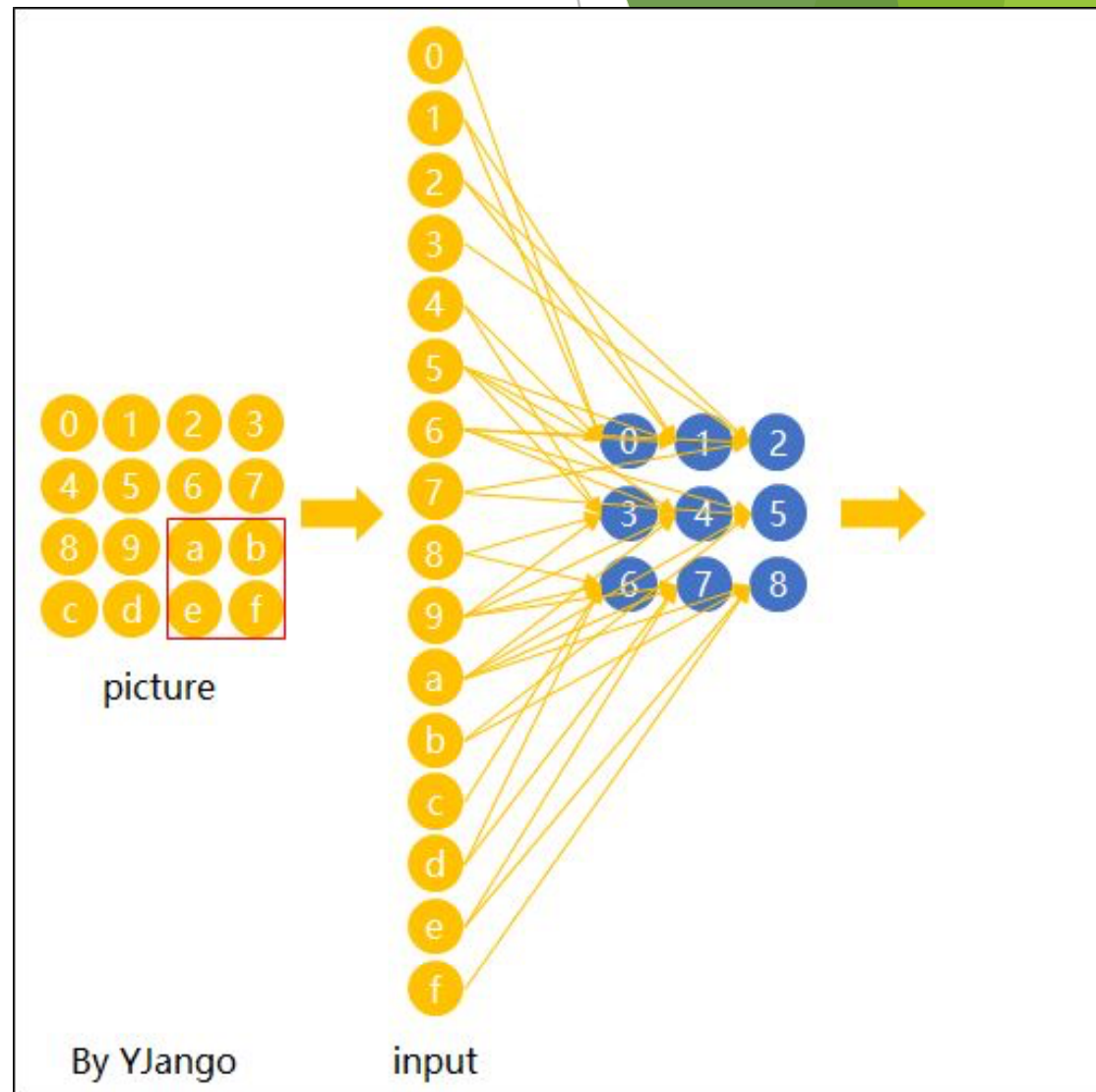
最后一帧则显示了所有连接。可以注意到，每个输出节点并非像前馈神经网络中那样与全部的输入节点连接，而是部分连接。这也就是为什么大家也叫前馈神经网络（feedforward neural network）为fully-connected neural network。图中显示的是一步步的移动filter来扫描全图，一次移动多少叫做stride。

空间共享也就是卷积神经网络所引入的先验知识



图片识别--卷积神经网络 - 输出表达

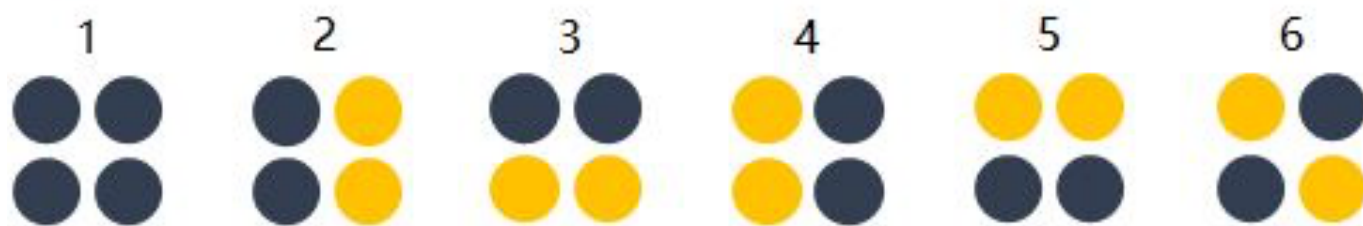
如先前在图像表达中提到的，图片不用向量去表示是为了保留图片平面结构的信息。同样的，卷积后的输出若用上图的排列方式则丢失了平面结构信息。所以我们依然用矩阵的方式排列它们，就得到了右图所展示的连接。



卷积神经元的形状、概念抓取

不同的形状都可由细小的“零件”组合而成的。比如右图中，用 2×2 的范围所形成的16种形状可以组合成格式各样的“更大”形状。

卷积的每个filter可以探测特定的形状。又由于Feature Map保持了抓取后的空间结构。若将探测到细小图形的Feature Map作为新的输入再次卷积后，则可以由此探测到“更大”的形状概念。比如右图的第一个“大”形状可由2,3,4,5基础形状拼成。第二个可由2,4,5,6组成。第三个可由6,1组成。



2x2的几个形状的例子










由上面6个形状为“零件”形成的“大”形状

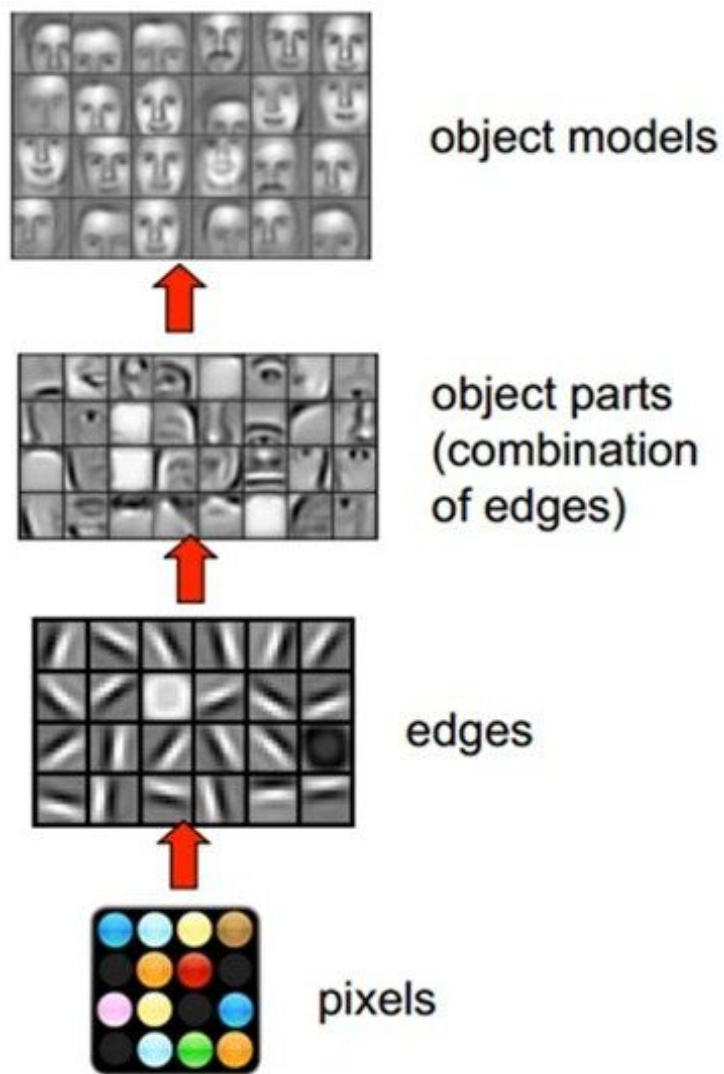
卷积神经元和图像处理关系

除了基础形状之外，颜色、对比度等概念对画面的识别结果也有影响。卷积层也会根据需要去探测特定的概念。

可以从右面这张图中感受到不同数值的filters所卷积过后的Feature Map可以探测边缘，棱角，模糊，突出等概念。

Operation	Filter	Convolved Image
Identity	$\begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}$	
Edge detection	$\begin{bmatrix} 1 & 0 & -1 \\ 0 & 0 & 0 \\ -1 & 0 & 1 \end{bmatrix}$	
	$\begin{bmatrix} 0 & 1 & 0 \\ 1 & -4 & 1 \\ 0 & 1 & 0 \end{bmatrix}$	
	$\begin{bmatrix} -1 & -1 & -1 \\ -1 & 8 & -1 \\ -1 & -1 & -1 \end{bmatrix}$	
Sharpen	$\begin{bmatrix} 0 & -1 & 0 \\ -1 & 5 & -1 \\ 0 & -1 & 0 \end{bmatrix}$	
Box blur (normalized)	$\frac{1}{9} \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}$	
Gaussian blur (approximation)	$\frac{1}{16} \begin{bmatrix} 1 & 2 & 1 \\ 2 & 4 & 2 \\ 1 & 2 & 1 \end{bmatrix}$	

卷积神经网络工作原理



图片被识别成什么不仅仅取决于图片本身，还取决于图片是如何被观察的。

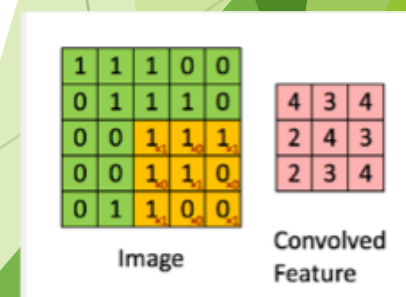
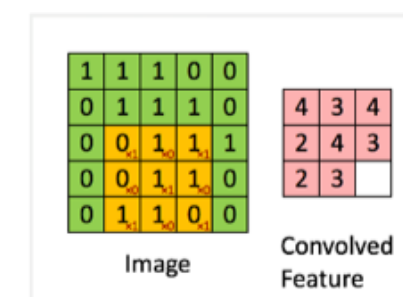
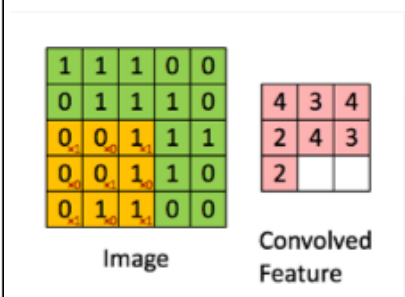
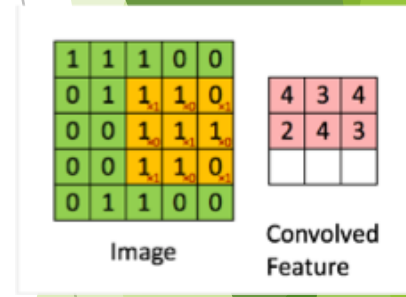
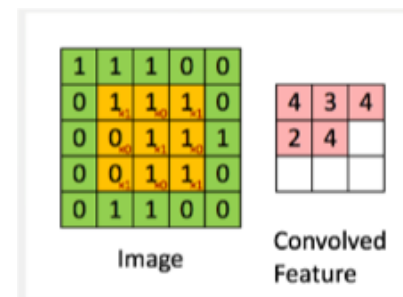
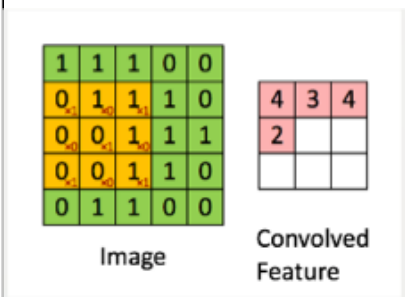
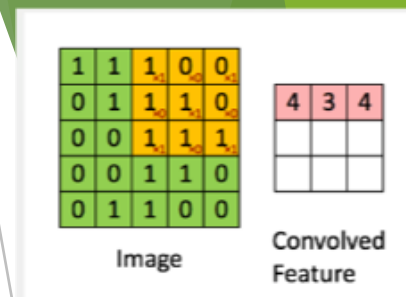
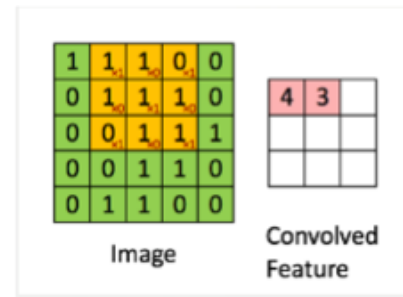
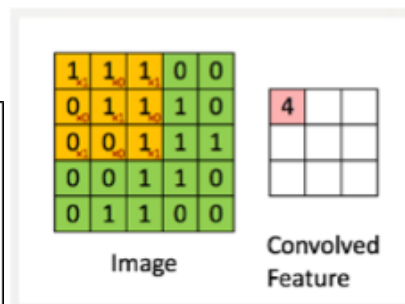
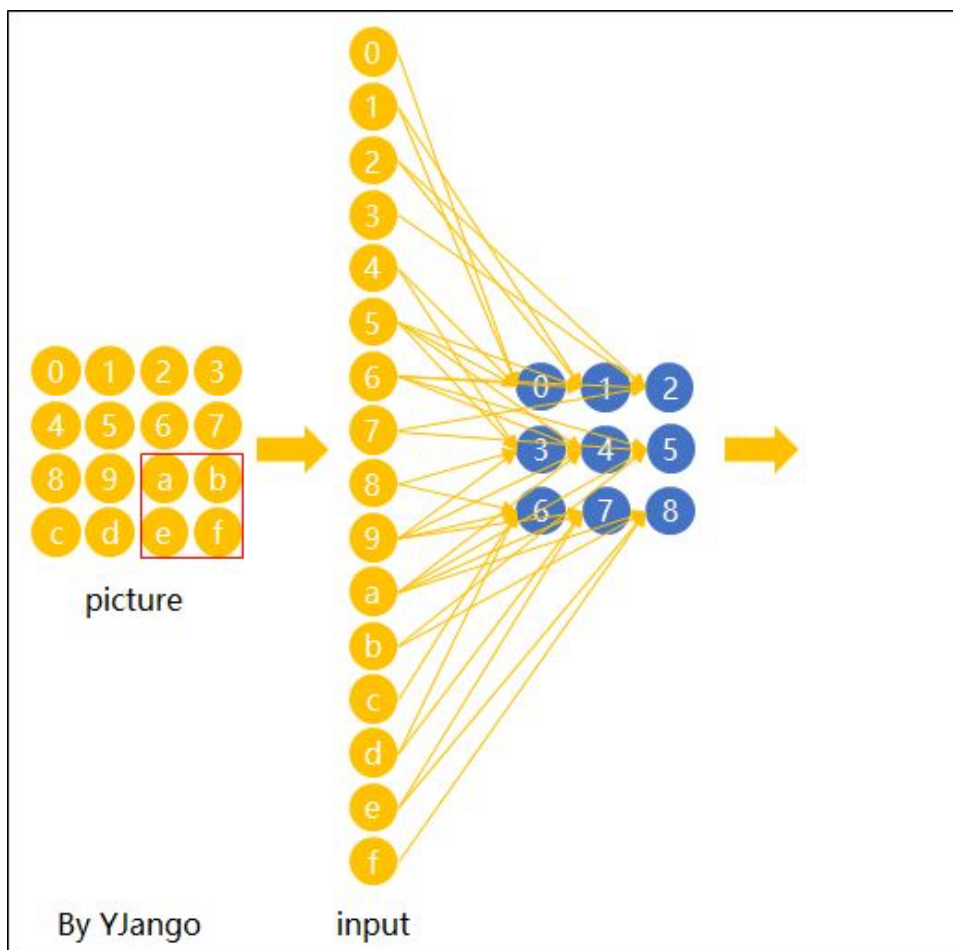
而**filter**内的权重矩阵 W 是网络根据数据学习得到的，也就是说，我们让神经网络自己学习以什么样的方式去观察图片。

拿老妇与少女的那幅图片举例，当标签是少女时，卷积网络就会学习抓取可以成少女的形状、概念。当标签是老妇时，卷积网络就会学习抓取可以成老妇的形状、概念。

左图展现了在人脸识别中经过层层卷积后，所能够探测的形状、概念也变得越来越抽象和复杂。

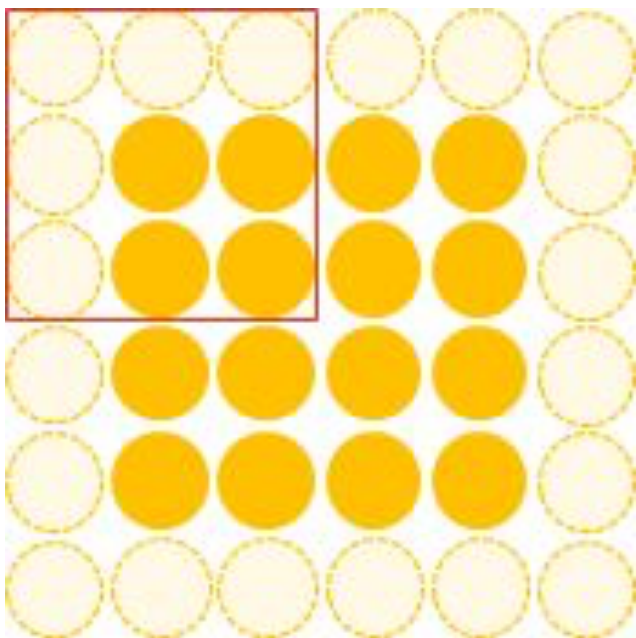
卷积神经网络会尽可能寻找最能解释训练数据的抓取方式。

卷积操作



经过一个feature detector计算后得到的粉红色区域也叫做一个“Convolved Feature”或“Activation Map”或“Feature Map”

ZERO Padding



通常大家都想要在卷积时保持图片的原始尺寸。选择3x3的filter和1的zero padding，或5x5的filter和2的zero padding可以保持图片的原始尺寸。这也是为什么大家多选择3x3和5x5的filter的原因。另一个原因是3x3的filter考虑到了像素与其距离为1以内的所有其他像素的关系，而5x5则是考虑像素与其距离为2以内的所有其他像素的关系。

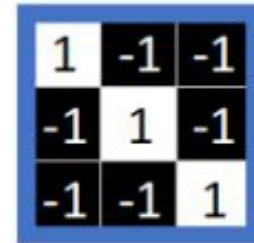
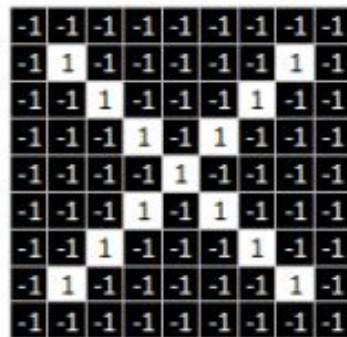
Feature Map的尺寸等于：

$$(\text{input_size} + 2 * \text{padding_size} - \text{filter_size}) / \text{stride} + 1$$

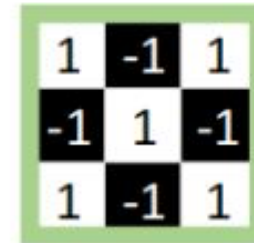
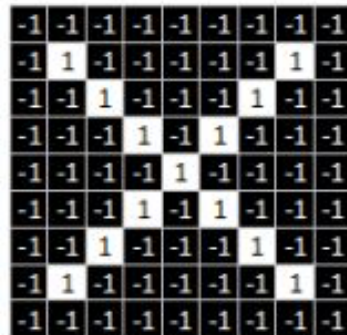
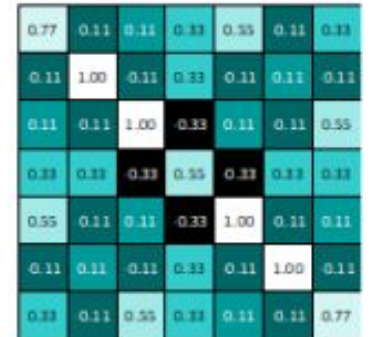
4x4的图片在边缘Zero padding一圈后，再用3x3的filter卷积后，得到的Feature Map尺寸依然是4x4不变。

卷积效果示意图

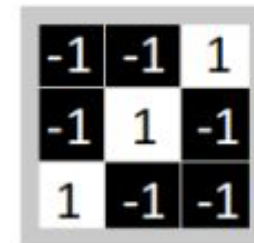
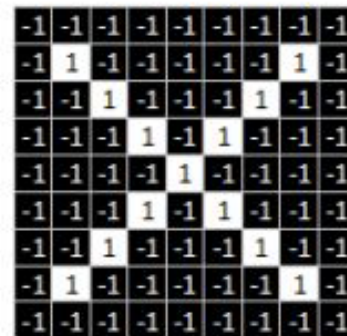
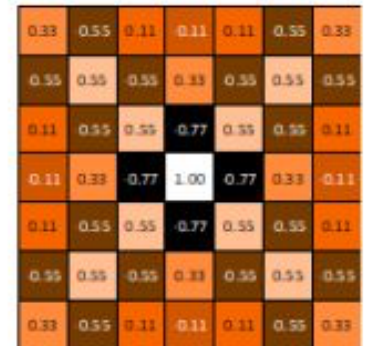
feature map是每一个feature从原始图像中提取出来的“特征”。其中的值，越接近为1表示对应位置和feature的匹配越完整，越是接近-1，表示对应位置和feature的反面匹配越完整，而值接近0的表示对应位置没有任何匹配或者说没有什么关联。



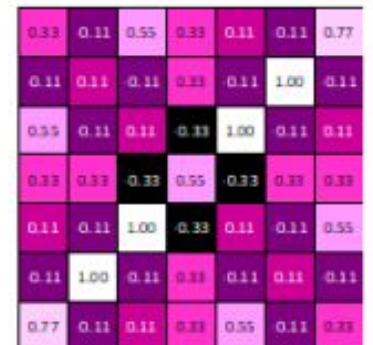
=



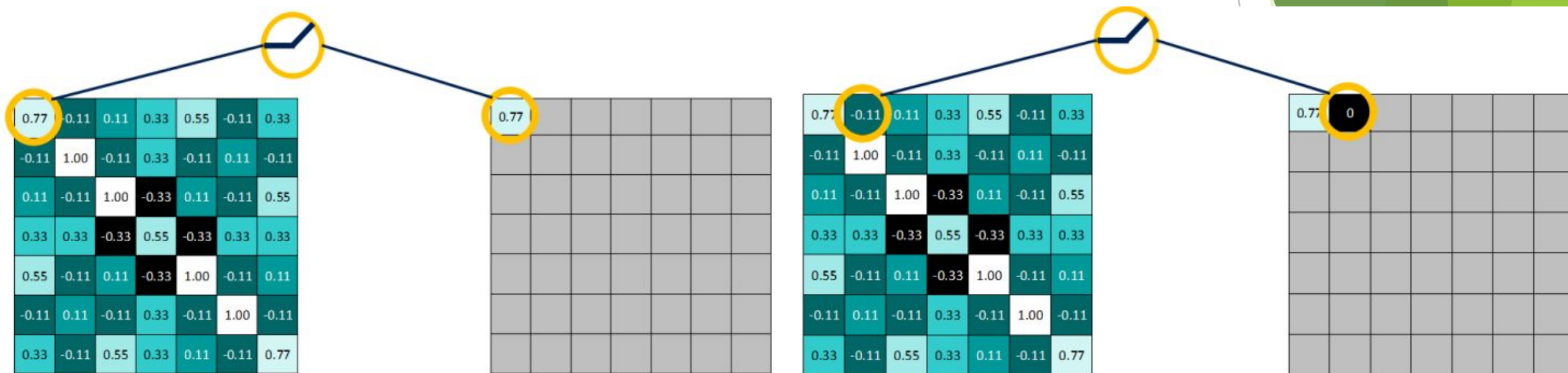
=



=



激活函数非线性 - 1



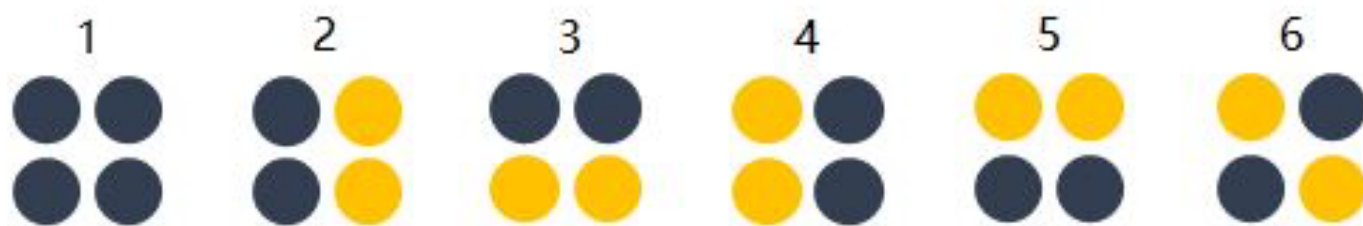
卷积后产生的特征图中的值，越靠近1表示与该特征越关联，越靠近-1表示越不关联，而我们进行特征提取时，为了使得数据更少，操作更方便，就直接舍弃掉那些不相关联的数据。



卷积神经元的形状、概念抓取

不同的形状都可由细小的“零件”组合而成的。比如右图中，用 2×2 的范围所形成的16种形状可以组合成格式各样的“更大”形状。

卷积的每个filter可以探测特定的形状。又由于Feature Map保持了抓取后的空间结构。若将探测到细小图形的Feature Map作为新的输入再次卷积后，则可以由此探测到“更大”的形状概念。比如右图的第一个“大”形状可由2,3,4,5基础形状拼成。第二个可由2,4,5,6组成。第三个可由6,1组成。

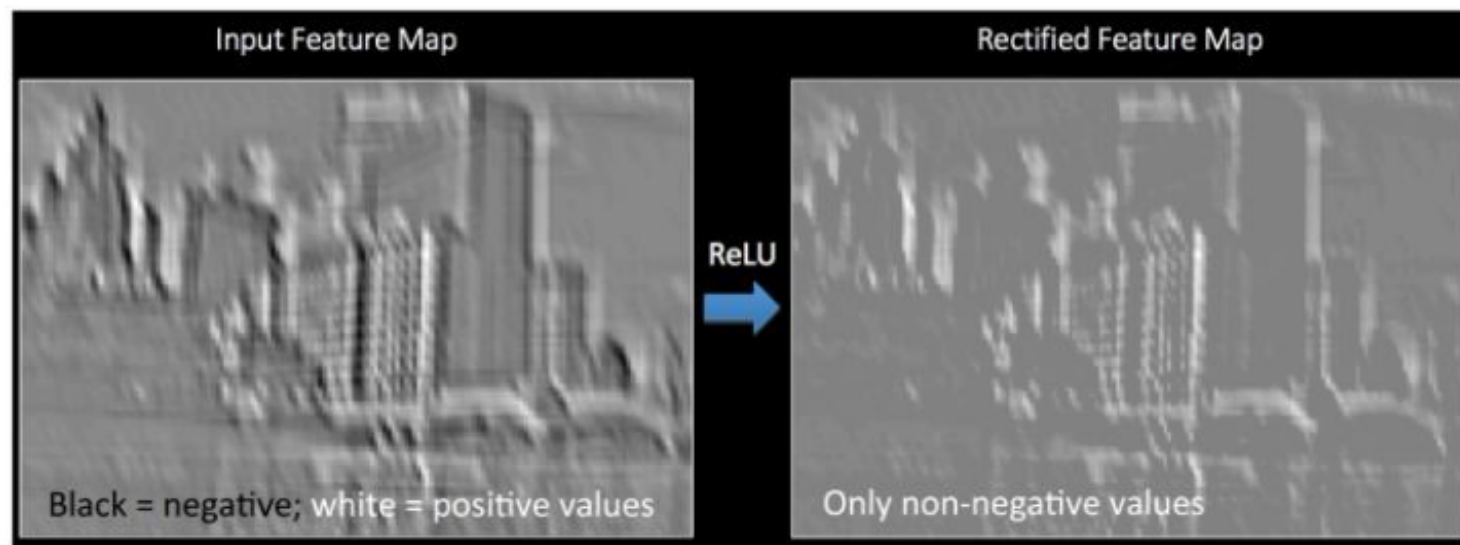
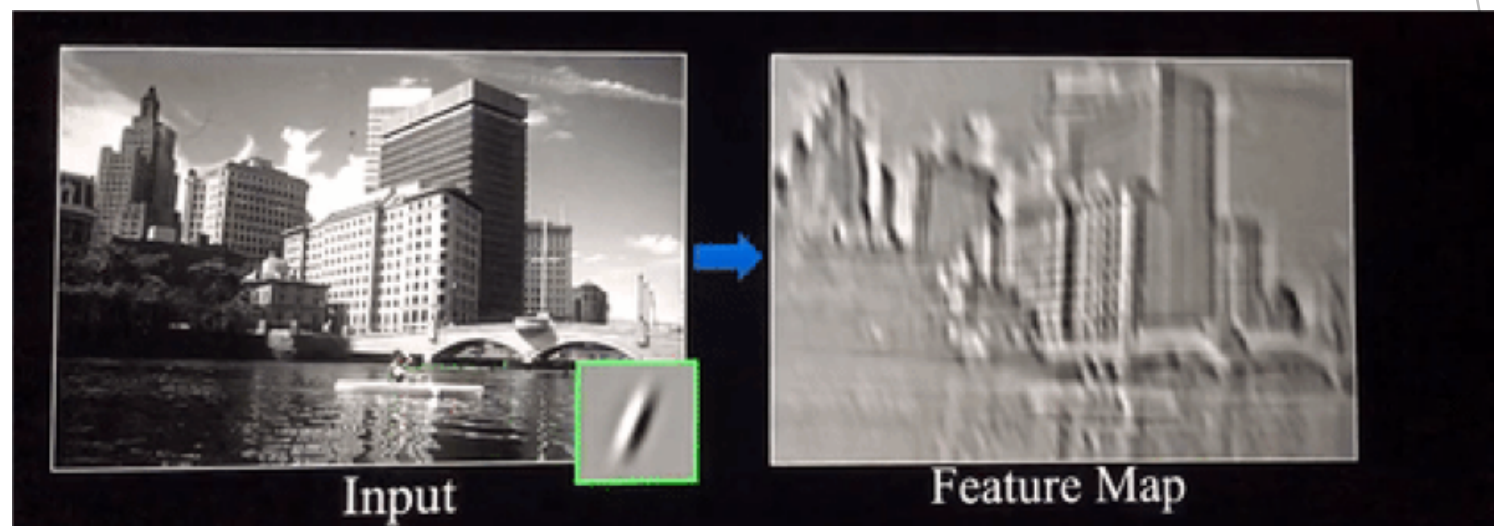


2x2的几个形状的例子

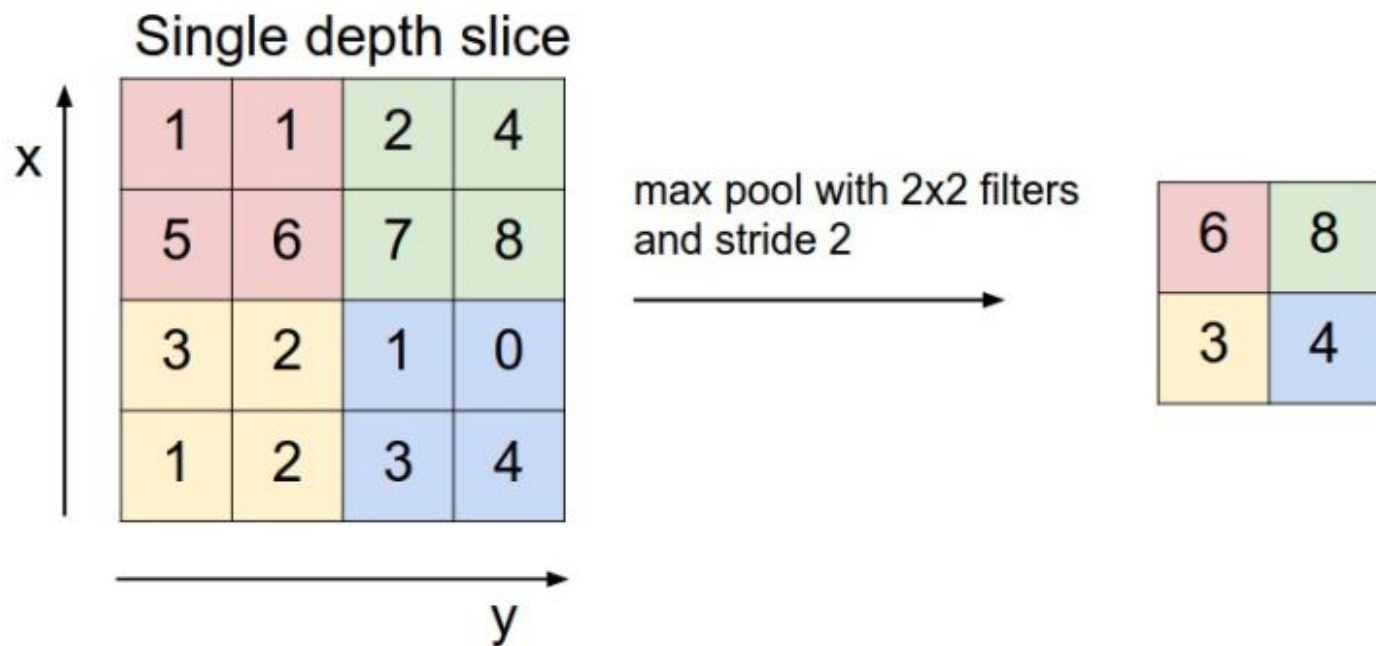


由上面6个形状为“零件”形成的“大”形状

激活函数非线性 - 2



MaxPooling(最大池化)

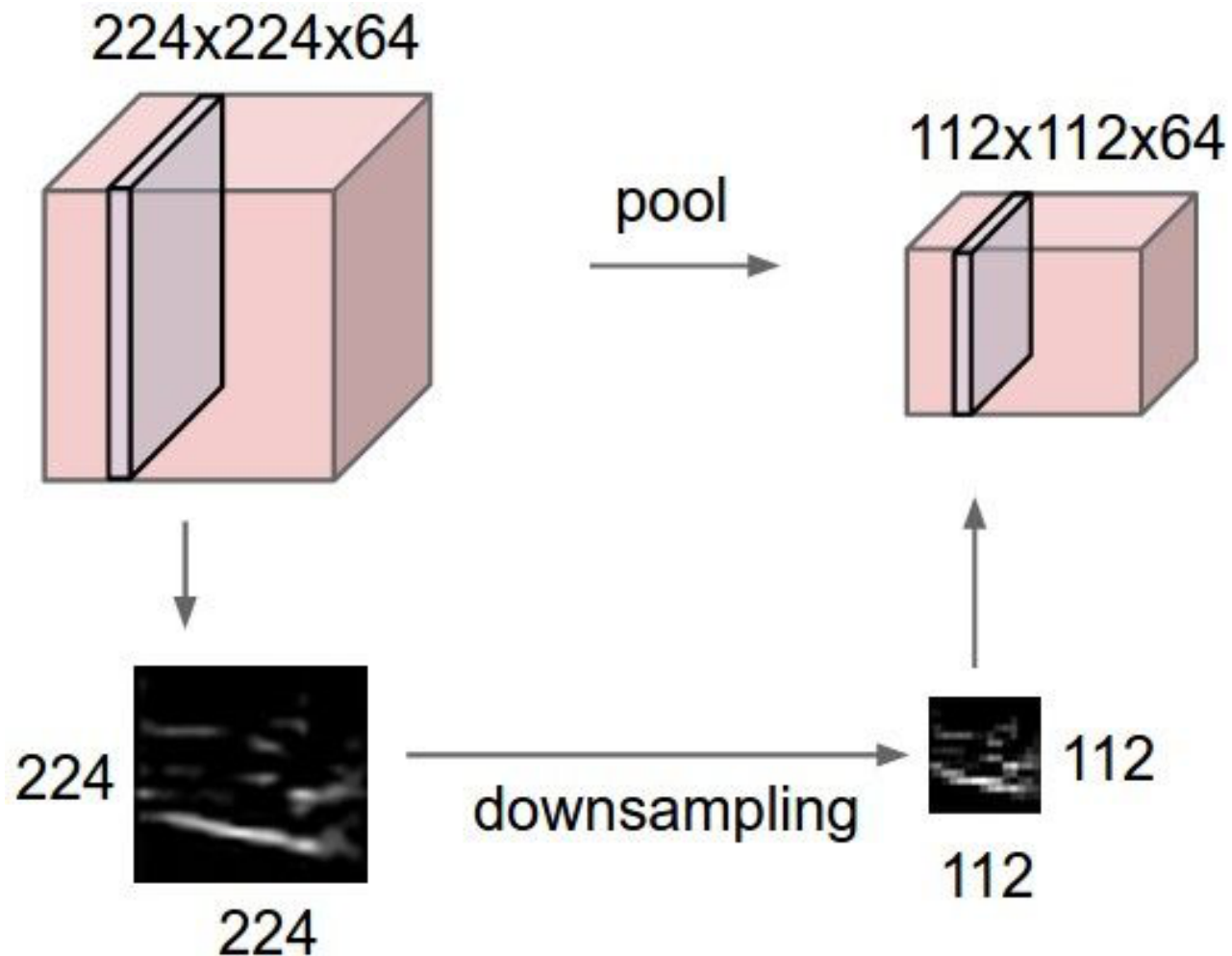


max pooling的操作如图所示：整个图片被不重叠的分割成若干个同样大小的小块（pooling size）。每个小块内只取最大的数字，再舍弃其他节点后，保持原有的平面结构得出output。

其他Pooling操作：average pooling

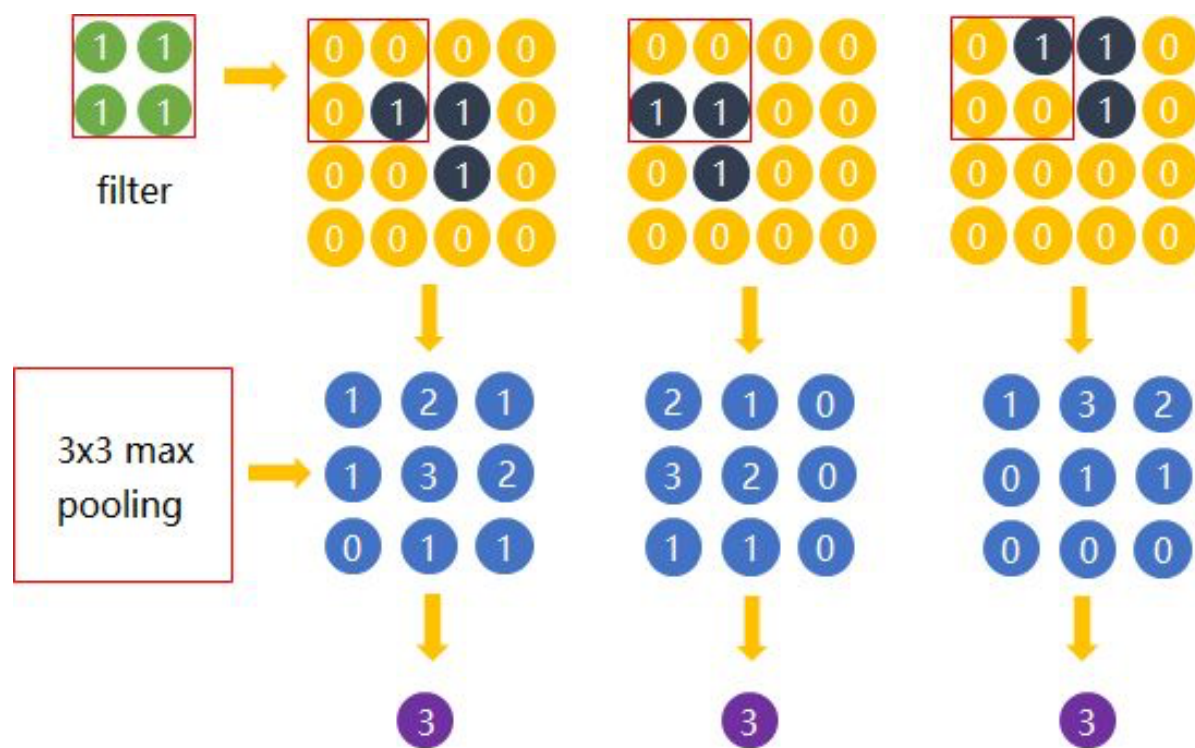
不需要参数控制

Max Pooling特性 - 1



Max pooling 的主要功能是 downsampling，却不会损坏识别结果。这意味着卷积后的 Feature Map 中有对于识别物体不必要的冗余信息。

Max Pooling特性 - 2

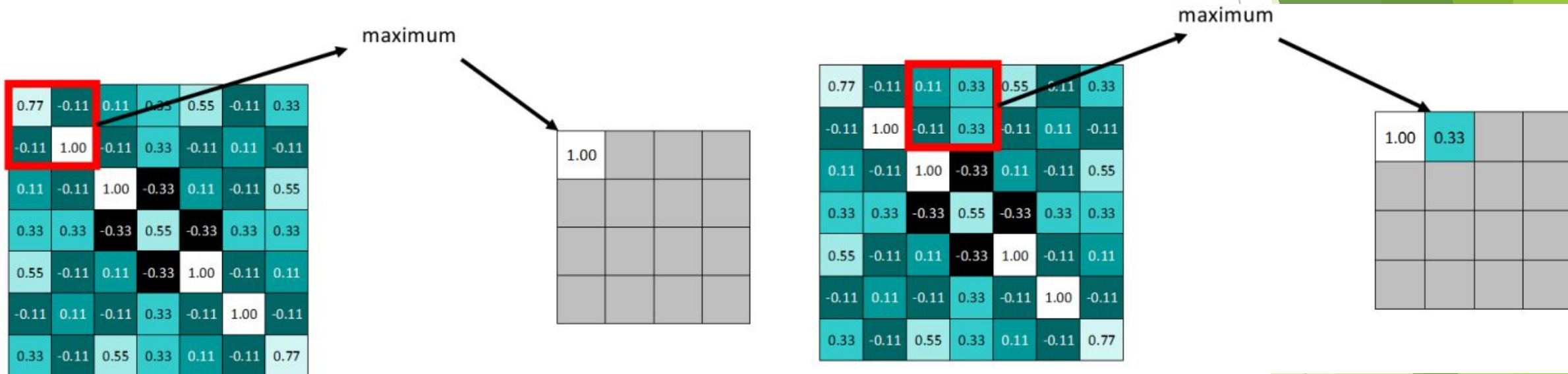


不需要参数控制

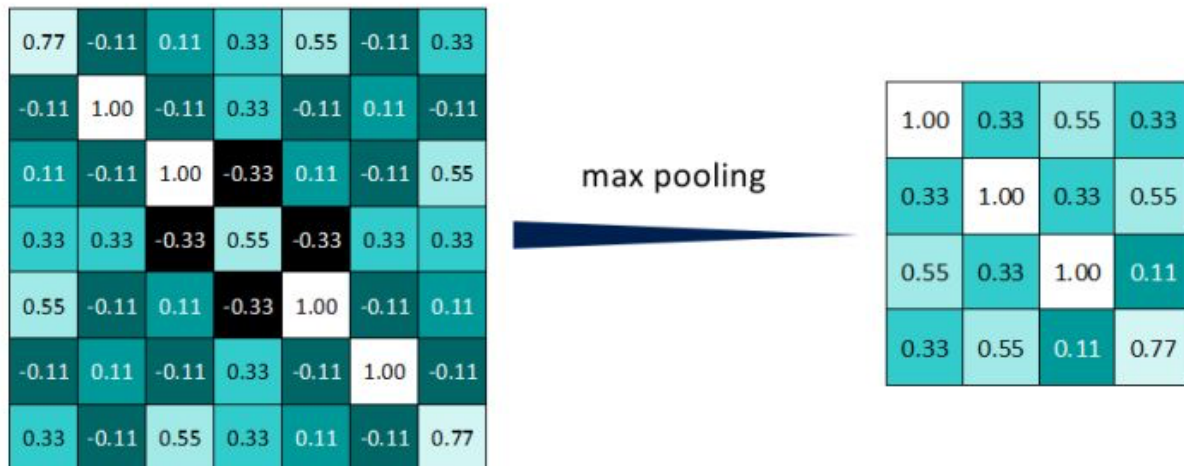
考虑探测“横折”这个形状。卷积后得到3x3的Feature Map中，真正有用的就是数字为3的那个节点，其余数值对于这个任务而言都是无关的。所以用3x3的Max pooling后，并没有对“横折”的探测产生影响。

如果不使用Max pooling，而让网络自己去学习。网络也会去学习与Max pooling近似效果的权重。因为是近似效果，增加了更多的parameters的代价，却还不如直接进行Max pooling。

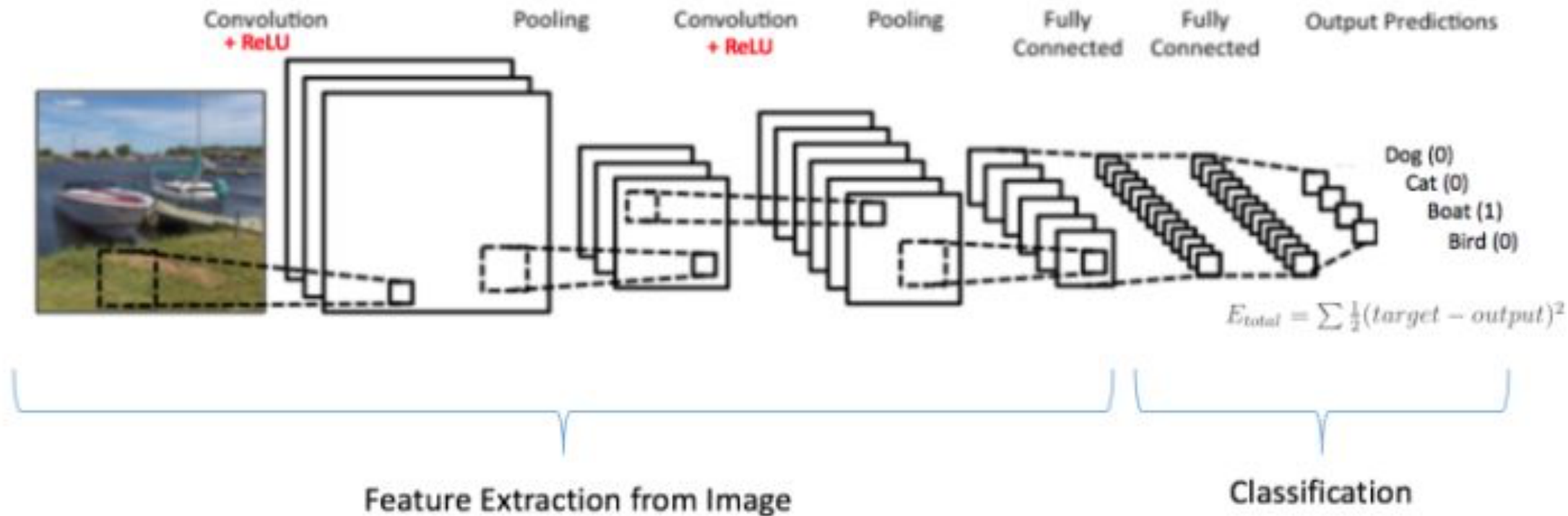
最大池化示意图



最终得到池化后的feature map。可明显发现数据量减少了很多。



全链接层

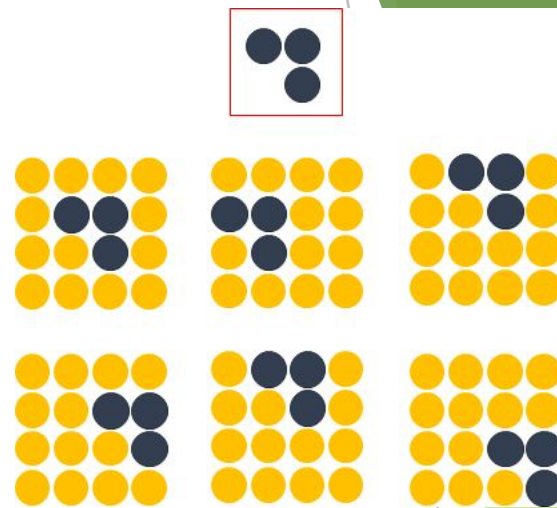


全连接层（也叫前馈层）就可以用来将最后的输出映射到线性可分的空间。通常卷积网络的最后会将末端得到的长方体平摊(**flatten**)成一个长长的向量，并送入全连接层配合输出层进行分类。

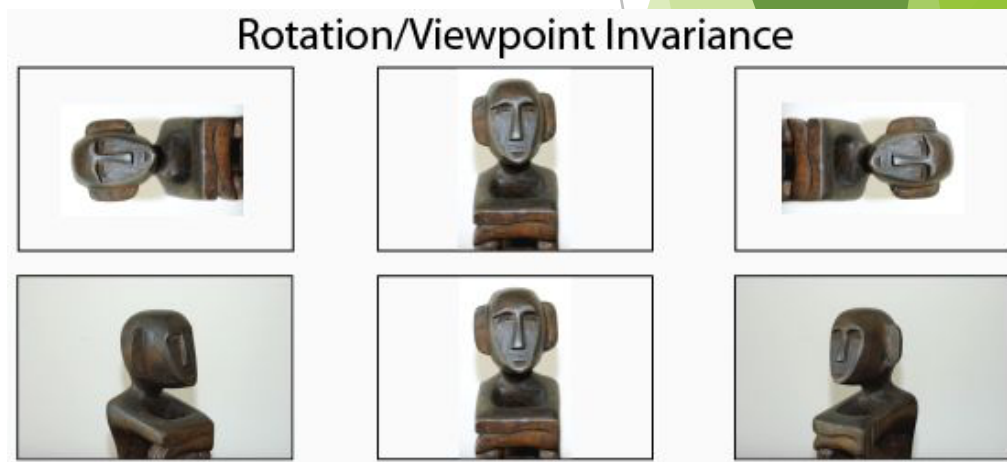
不变性的满足

- 平移不变性：局部连接和空间共享

- 因为空间共享，在不同位置的同一形状就可以被等价识别，所以不需要对每个位置都进行学习



- 旋转和视角不变性：靠大量的数据

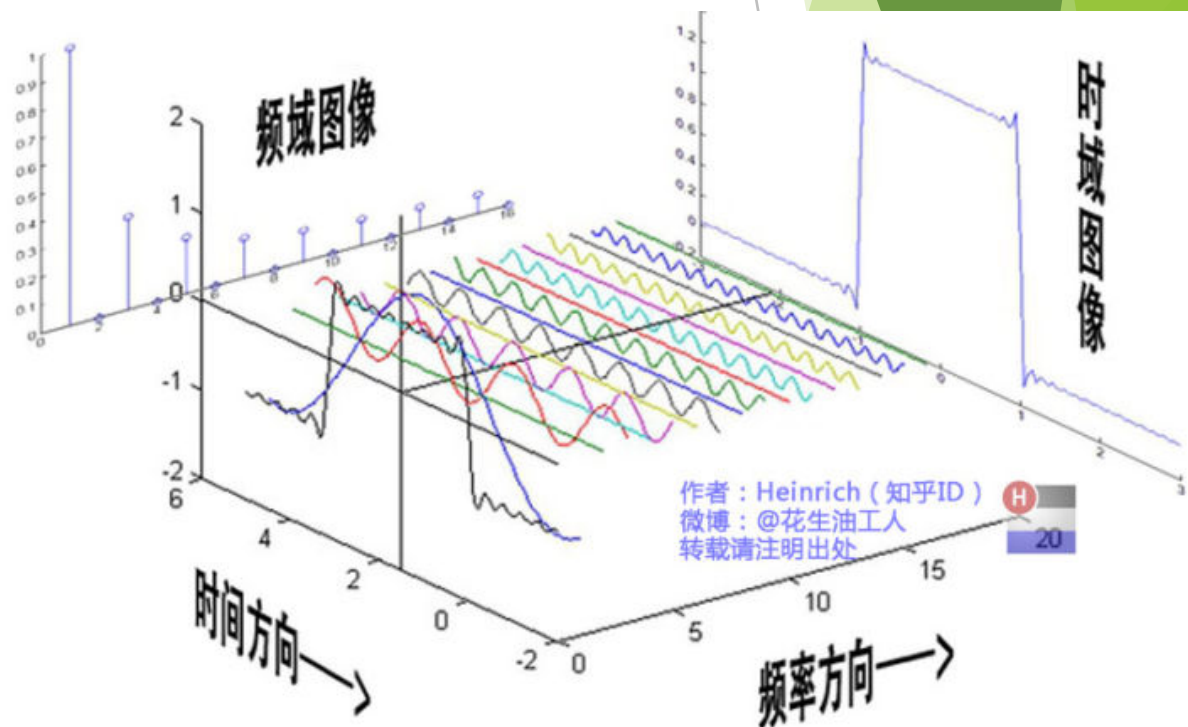
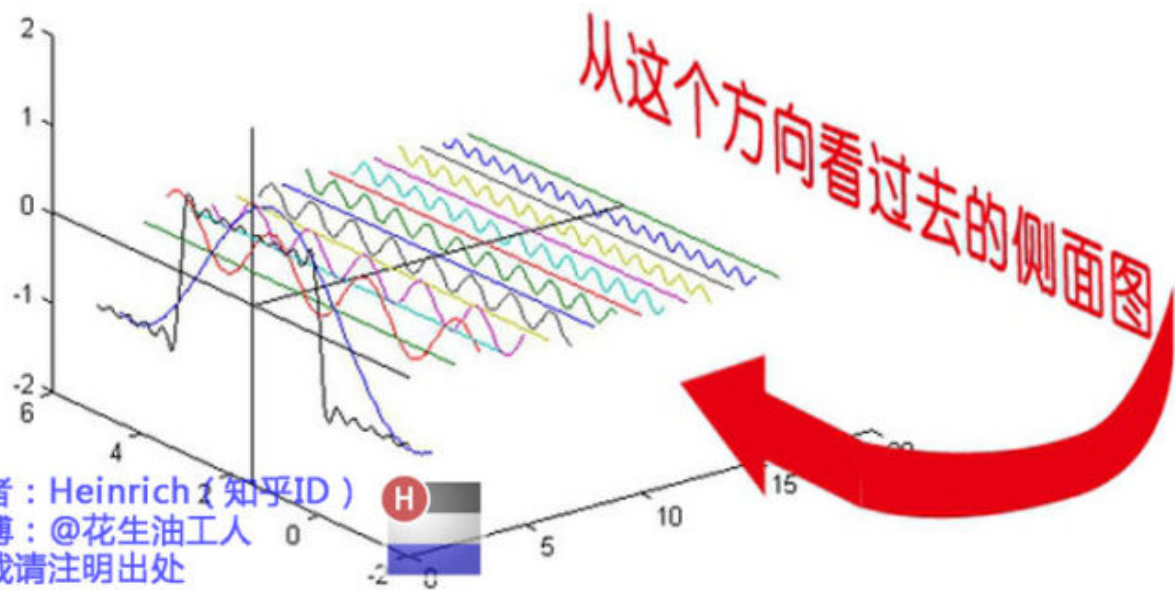


- 尺寸不变性（构造网络：多层卷积，网络结构）

- 与平移不变性不同，最初的卷积网络并没有明确照顾尺寸不变性这一特点

卷积的物理含义 - 1

- 图像『卷积』可以和『信号处理』联系起来
- 图像卷积的本质，是提取图像不同『频段』的特征



卷积的物理含义 - 2

$$(f * g)(n) = \int_{-\infty}^{\infty} f(\tau)g(n - \tau)d\tau$$

Diagram illustrating the integral form of convolution. Blue wavy lines underline $f(\tau)$ and $g(n - \tau)$. Blue arrows point from these underlined terms to the equation $n = \tau + (n - \tau)$ below.

$$n = \tau + (n - \tau)$$

$$(f * g)(n) = \sum_{\tau=-\infty}^{\infty} f(\tau)g(n - \tau)$$

Diagram illustrating the summation form of convolution. Blue wavy lines underline $f(\tau)$ and $g(n - \tau)$. Blue arrows point from these underlined terms to the equation $n = \tau + (n - \tau)$ above.



卷积的物理含义 - 3



f

1	2	3	4	5	6
---	---	---	---	---	---

f 表示第一枚骰子
 $f(1)$ 表示投出1的概率
 $f(2)$ 、 $f(3)$ 、 \dots 以此类推

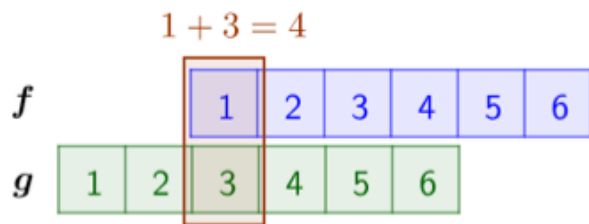
g

1	2	3	4	5	6
---	---	---	---	---	---

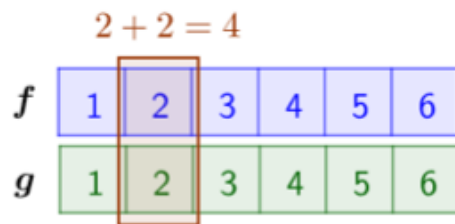
g 表示第二枚骰子

两枚骰子点数加起来为4的概率是多少？

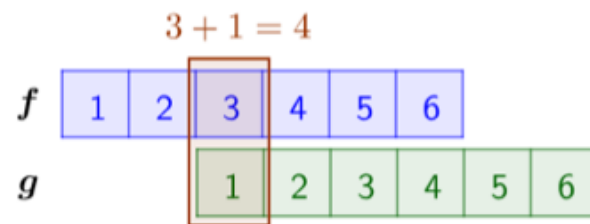
卷积的物理含义 - 4



出现概率为: $f(1)g(3)$



出现概率为: $f(2)g(2)$



出现概率为: $f(3)g(1)$

因此, 两枚骰子点数加起来为4的概率为:

$$f(1)g(3) + f(2)g(2) + f(3)g(1)$$

符合卷积的定义, 把它写成标准的形式就是:

$$(f * g)(4) = \sum_{m=1}^3 f(m)g(4 - m)$$

卷积的物理含义 - 5



这些噪点，属于高频信号

卷积平滑 →



$$g = \begin{bmatrix} \frac{1}{9} & \frac{1}{9} & \frac{1}{9} \\ \frac{1}{9} & \frac{1}{9} & \frac{1}{9} \\ \frac{1}{9} & \frac{1}{9} & \frac{1}{9} \end{bmatrix}$$



$$\Rightarrow \begin{bmatrix} a_{0,0} & a_{0,1} & a_{0,2} & \cdots & a_{0,n} \\ a_{1,0} & a_{1,1} & a_{1,2} & \cdots & a_{1,n} \\ a_{2,0} & a_{2,1} & a_{2,2} & \cdots & a_{2,n} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ a_{m,0} & a_{m,1} & a_{m,2} & \cdots & a_{m,n} \end{bmatrix}$$

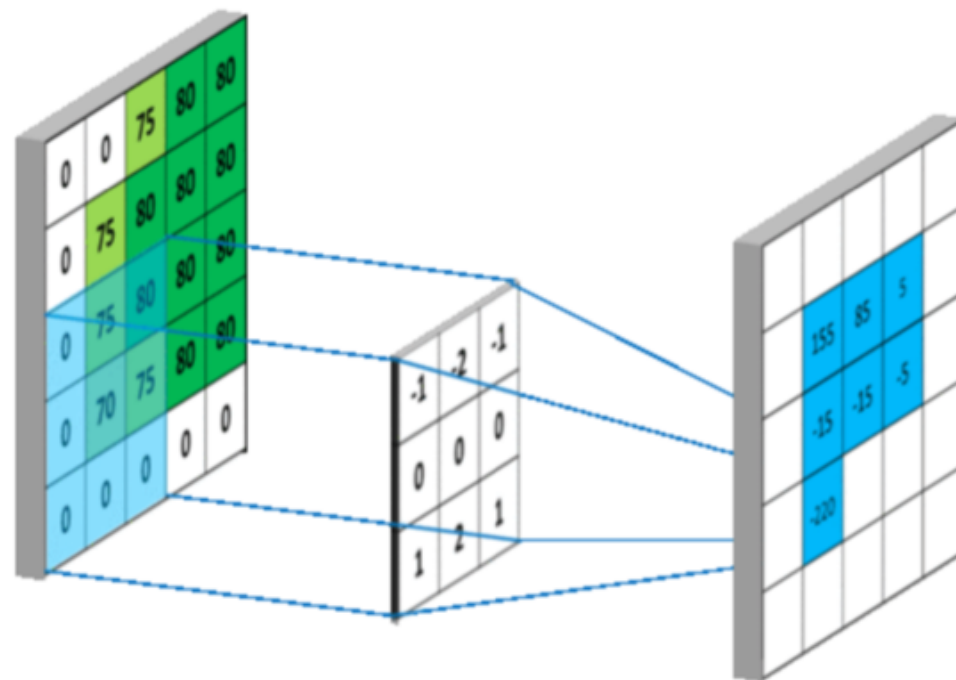
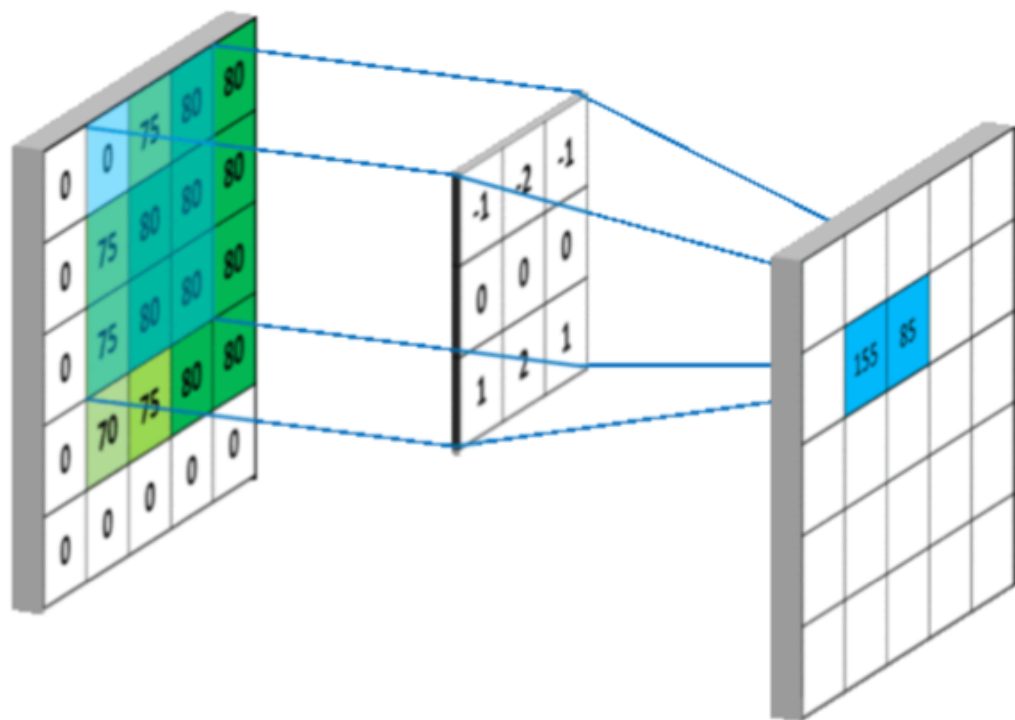
$$f = \begin{bmatrix} a_{0,0} & a_{0,1} & a_{0,2} \\ a_{1,0} & a_{1,1} & a_{1,2} \\ a_{2,0} & a_{2,1} & a_{2,2} \end{bmatrix}$$

$$\begin{bmatrix} a_{0,0} & a_{0,1} & a_{0,2} & \cdots & a_{0,n} \\ a_{1,0} & a_{1,1} & a_{1,2} & \cdots & a_{1,n} \\ a_{2,0} & a_{2,1} & a_{2,2} & \cdots & a_{2,n} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ a_{m,0} & a_{m,1} & a_{m,2} & \cdots & a_{m,n} \end{bmatrix}$$

$$\begin{bmatrix} c_{0,0} & c_{0,1} & c_{0,2} & \cdots & c_{1,n} \\ c_{1,0} & c_{1,1} & c_{1,2} & \cdots & c_{1,n} \\ c_{2,0} & c_{2,1} & c_{2,2} & \cdots & c_{2,n} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ c_{m,0} & c_{m,1} & c_{m,2} & \cdots & c_{m,n} \end{bmatrix}$$

$$\Uparrow \\ c_{1,1} = f * g$$

卷积的物理含义 - 6

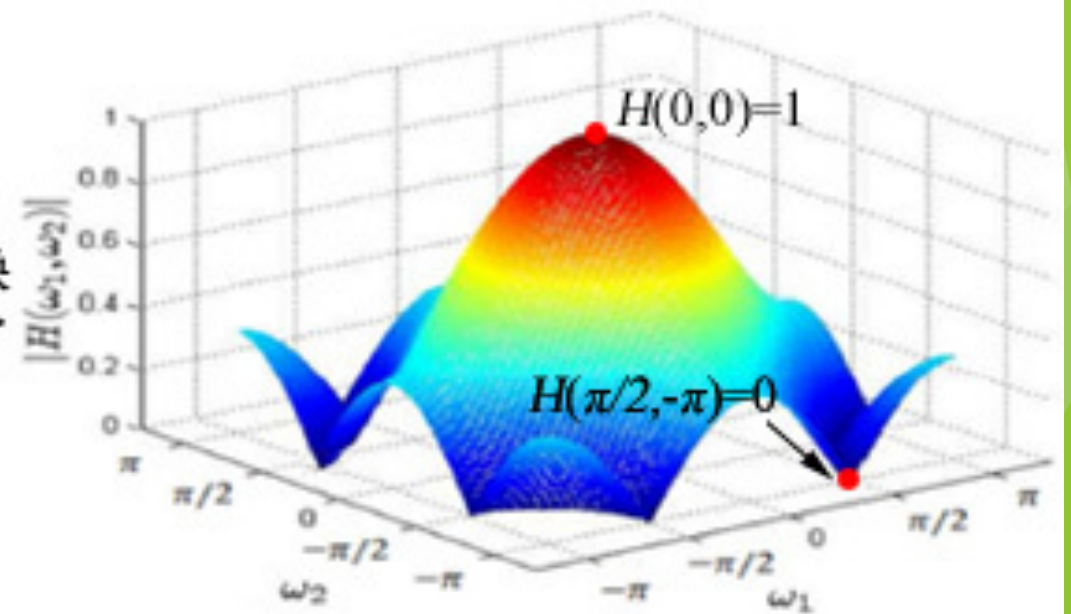


卷积的物理含义 - 7

卷积核本质上是一个二维函数，有对应的频谱函数，因而可以看成某种『滤波器』

$$h(n_1, n_2) = \begin{bmatrix} 0 & 1/6 & 0 \\ 1/6 & 1/3 & 1/6 \\ 0 & 1/6 & 0 \end{bmatrix}$$

傅里叶变换



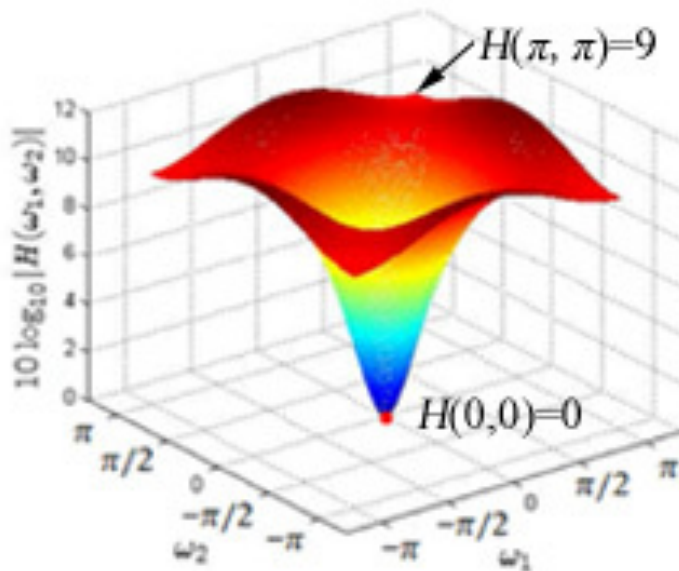
这是一个低通滤波器，频率接近原点附近的幅值很大(频率低的通过)，越往两边越小(频率高的过滤)。

卷积的物理含义 - 8

卷积核本质上是一个二维函数，有对应的频谱函数，因而可以看成某种『滤波器』

$$h(n_1, n_2) = \begin{bmatrix} -1 & -1 & -1 \\ -1 & 9 & -1 \\ -1 & -1 & -1 \end{bmatrix}$$

傅里叶变换
→



这是一个高通滤波器，频率接近原点附近的幅值很小(频率低的过滤)，越往两边越高(频率高的通过)。

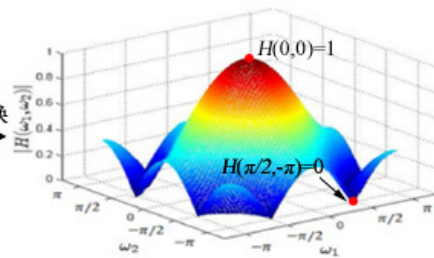
卷积的物理含义 - 9

波形里的「棱角」其实是一种突变信号，它里面包含了很多高频分量。我们沿用上面『水池』的类比，图像像素值变化陡峭的地方，反映在图像上，就是那块区域明暗变化明显，而类比到『水池』里，就是水波在该区域快速振动，『棱角』分明。

当我们将图像跟『高通滤波器』做卷积时，明暗变化会被保留，而缓和的变化会被过滤

$$h(n_1, n_2) = \begin{bmatrix} 0 & 1/6 & 0 \\ 1/6 & 1/3 & 1/6 \\ 0 & 1/6 & 0 \end{bmatrix}$$

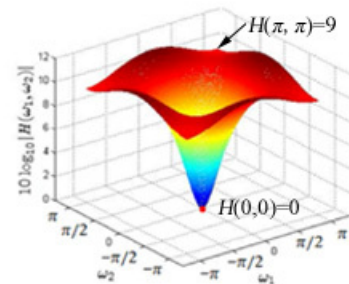
傅里叶变换



低通滤波器

$$h(n_1, n_2) = \begin{bmatrix} -1 & -1 & -1 \\ -1 & 9 & -1 \\ -1 & -1 & -1 \end{bmatrix}$$

傅里叶变换



高通滤波器

卷积的物理含义 - 10

- 当我们将图像跟『高通滤波器』做卷积时，明暗变化会被保留，而缓和的变化会被过滤。反映到图像上，就是『锐化』效果，即图像的边缘被加强，大色块的背景被过滤
- 低通滤波器做卷积，效果相反。

$f(x,y)$

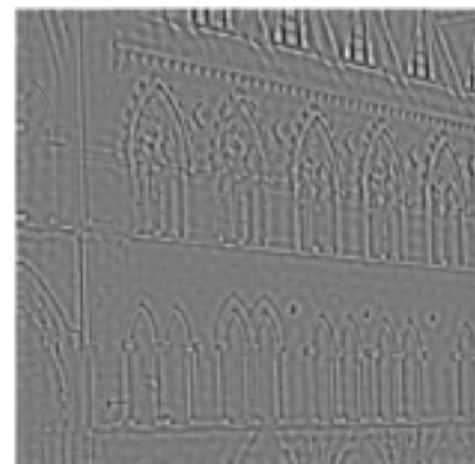
original







low pass






high pass



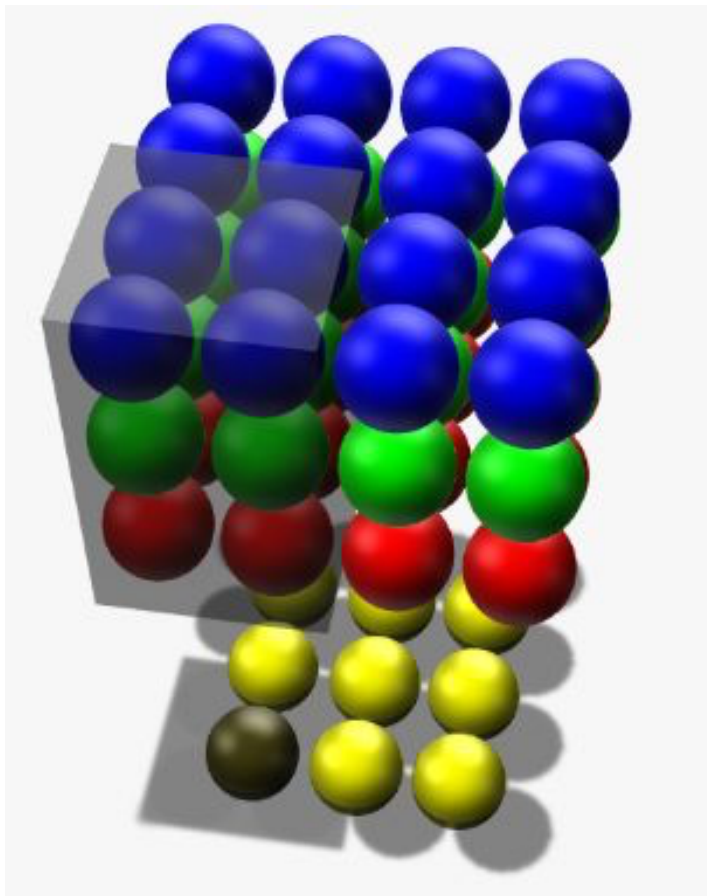
卷积神经元和图像处理关系

Operation	Filter	Convolved Image
Identity	$\begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}$	
Edge detection	$\begin{bmatrix} 1 & 0 & -1 \\ 0 & 0 & 0 \\ -1 & 0 & 1 \end{bmatrix}$	
	$\begin{bmatrix} 0 & 1 & 0 \\ 1 & -4 & 1 \\ 0 & 1 & 0 \end{bmatrix}$	
	$\begin{bmatrix} -1 & -1 & -1 \\ -1 & 8 & -1 \\ -1 & -1 & -1 \end{bmatrix}$	

Operation	Filter	Convolved Image
Sharpen	$\begin{bmatrix} 0 & -1 & 0 \\ -1 & 5 & -1 \\ 0 & -1 & 0 \end{bmatrix}$	
Box blur (normalized)	$\frac{1}{9} \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}$	
Gaussian blur (approximation)	$\frac{1}{16} \begin{bmatrix} 1 & 2 & 1 \\ 2 & 4 & 2 \\ 1 & 2 & 1 \end{bmatrix}$	

图像卷积的本质，是提取
图像不同『频段』的特征

三维卷积 - 1

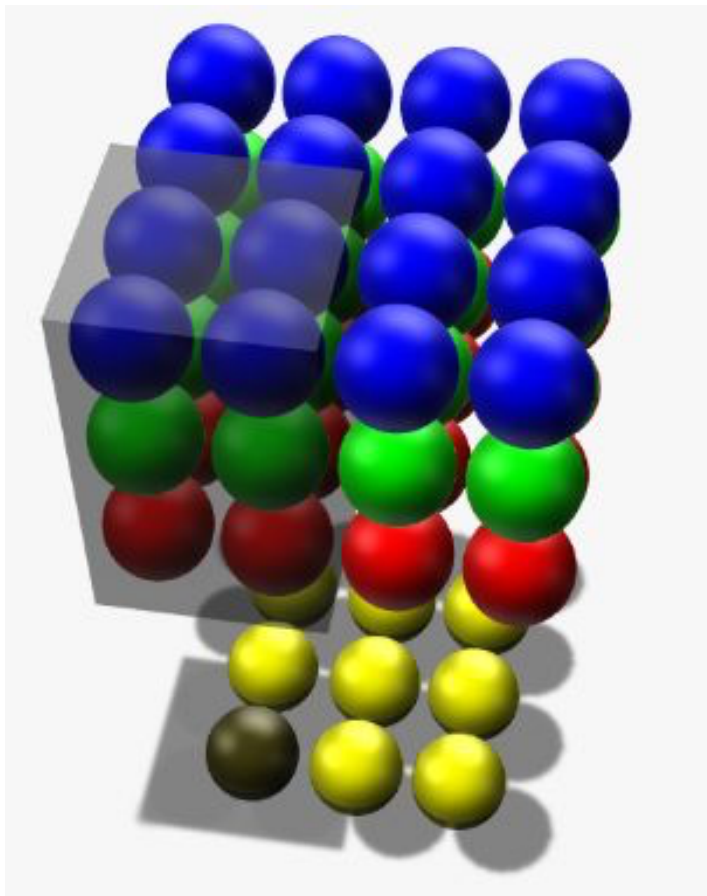


图中红、绿、蓝颜色的节点表示3个channels。黄色节点表示一个feature detector卷积后得到的Feature Map。其中被透明黑框圈中的12个节点会被连接到黄黑色的节点上。

- 在输入depth为1时：被filter size为 2×2 所圈中的4个输入节点连接到1个输出节点上。
- 在输入depth为3时：被filter size为 2×2 ，但是贯串3个channels后，所圈中的12个输入节点连接到1个输出节点上。
- 在输入depth为n时： $2 \times 2 \times n$ 个输入节点连接到1个输出节点上。

3个channels的RGB颜色模型

三维卷积 - 2



三个channels的权重并不共享。即当深度变为3后，权重也跟着扩增到了三组，不同channels用的是自己的权重。式子中增加的角标r,g,b分别表示red channel, green channel, blue channel的权重。

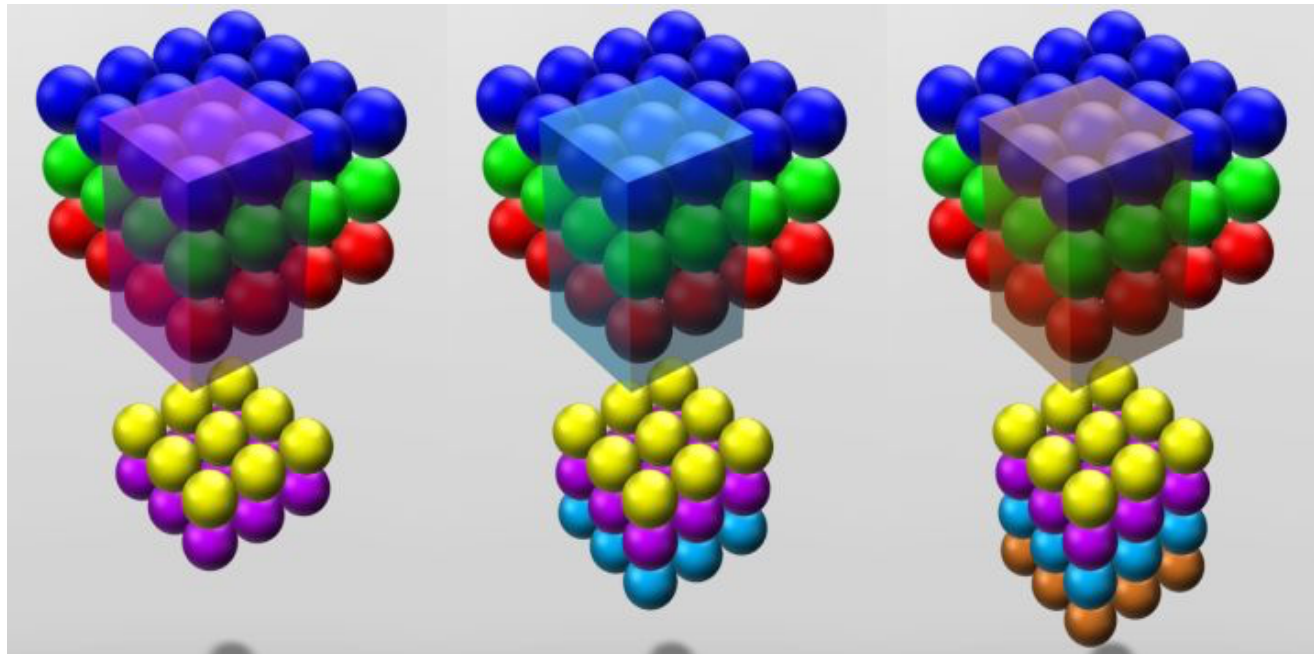
$$\begin{bmatrix} w_{r1} & w_{r2} \\ w_{r3} & w_{r4} \end{bmatrix}, \begin{bmatrix} w_{g1} & w_{g2} \\ w_{g3} & w_{g4} \end{bmatrix}, \begin{bmatrix} w_{b1} & w_{b2} \\ w_{b3} & w_{b4} \end{bmatrix}$$

$$y_0 = [w_{r1} \quad w_{r2} \quad w_{r3} \quad w_{r4}] \cdot \begin{bmatrix} x_{r0} \\ x_{r1} \\ x_{r4} \\ x_{r5} \end{bmatrix} + [w_{g1} \quad w_{g2} \quad w_{g3} \quad w_{g4}] \cdot \begin{bmatrix} x_{g0} \\ x_{g1} \\ x_{g4} \\ x_{g5} \end{bmatrix} + [w_{b1} \quad w_{b2} \quad w_{b3} \quad w_{b4}] \cdot \begin{bmatrix} x_{b0} \\ x_{b1} \\ x_{b4} \\ x_{b5} \end{bmatrix} + b_0$$

3个channels的RGB颜色模型

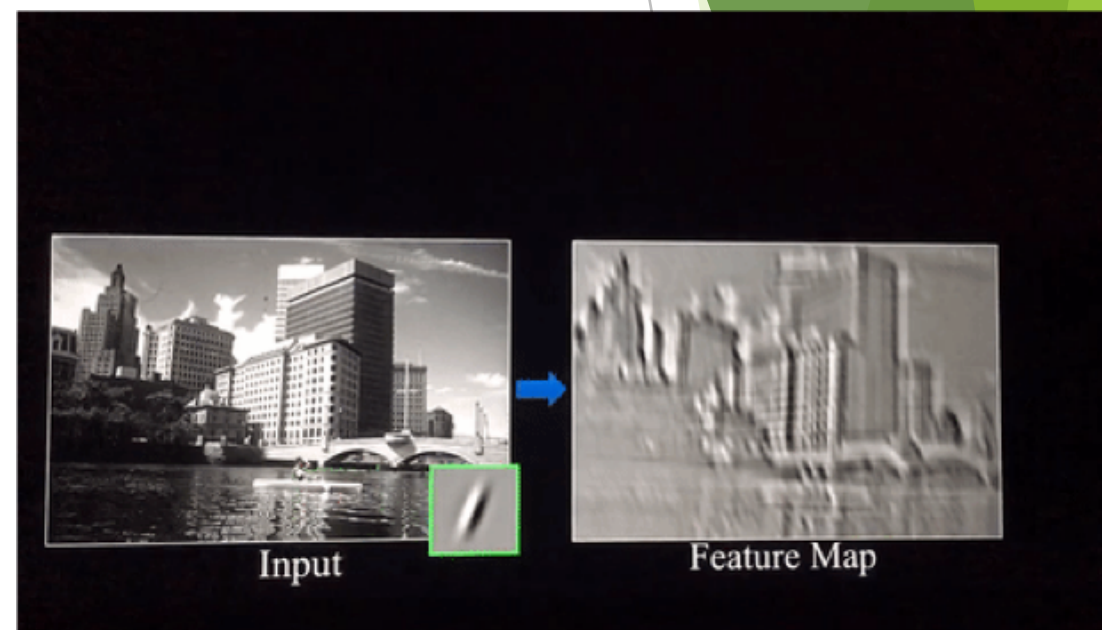
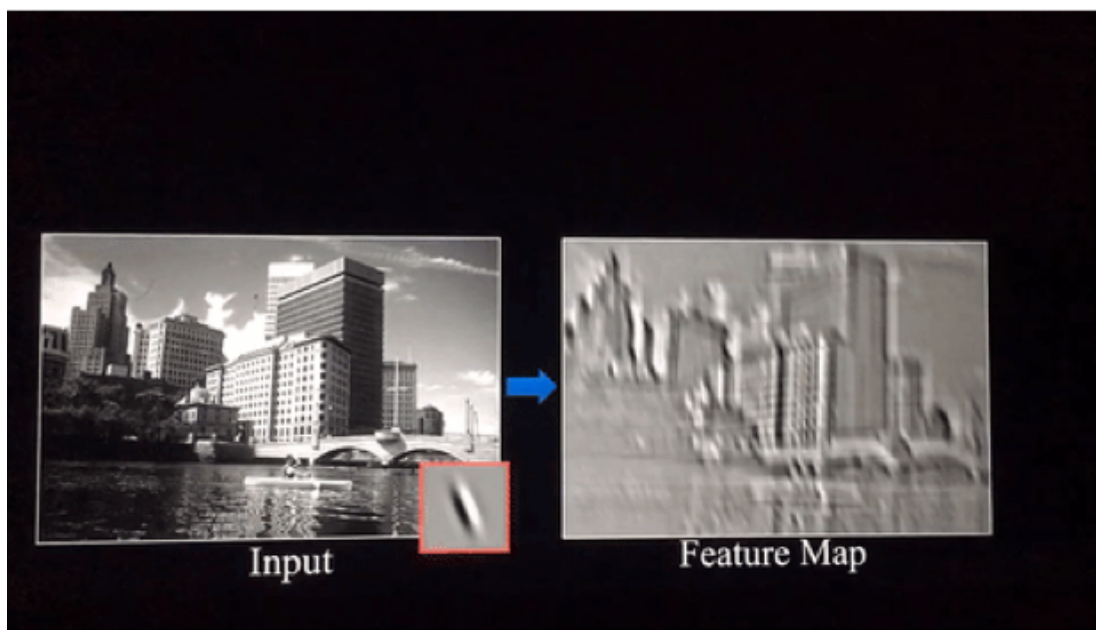
三维卷积 - 3

用多个不同的filters对同一个图片进行多次抓取



同一个图片，经过两个（红色、绿色）不同的filters扫描过后可得到不同特点的Feature Maps。每增加一个filter，就意味着你想让网络多抓取一个特征。

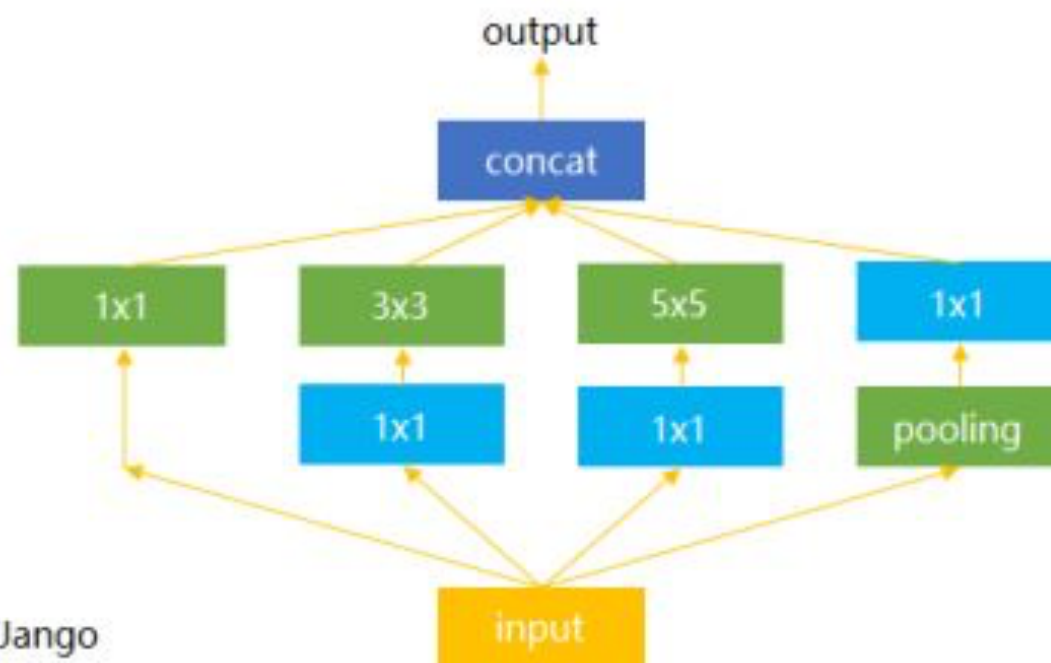
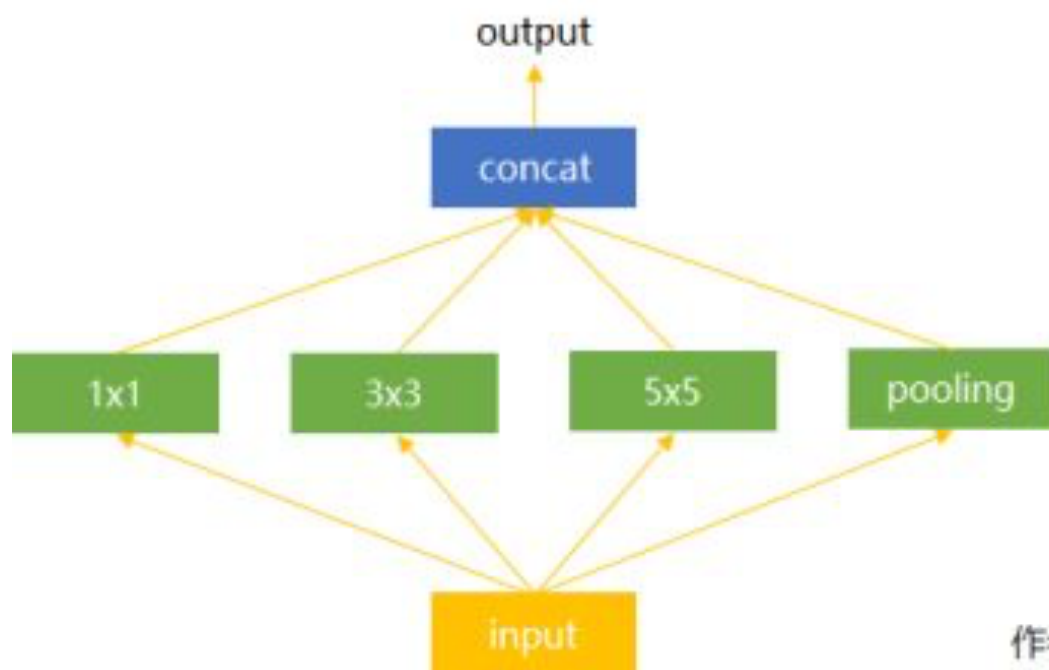
多特征抓取



实际应用中会选择用多个不同的filters对同一个图片进行多次抓取。同一个图片，经过两个（红色、绿色）不同的filters扫描过后可得到不同特点的Feature Maps。每增加一个filter，就意味着你想让网络多抓取一个特征。

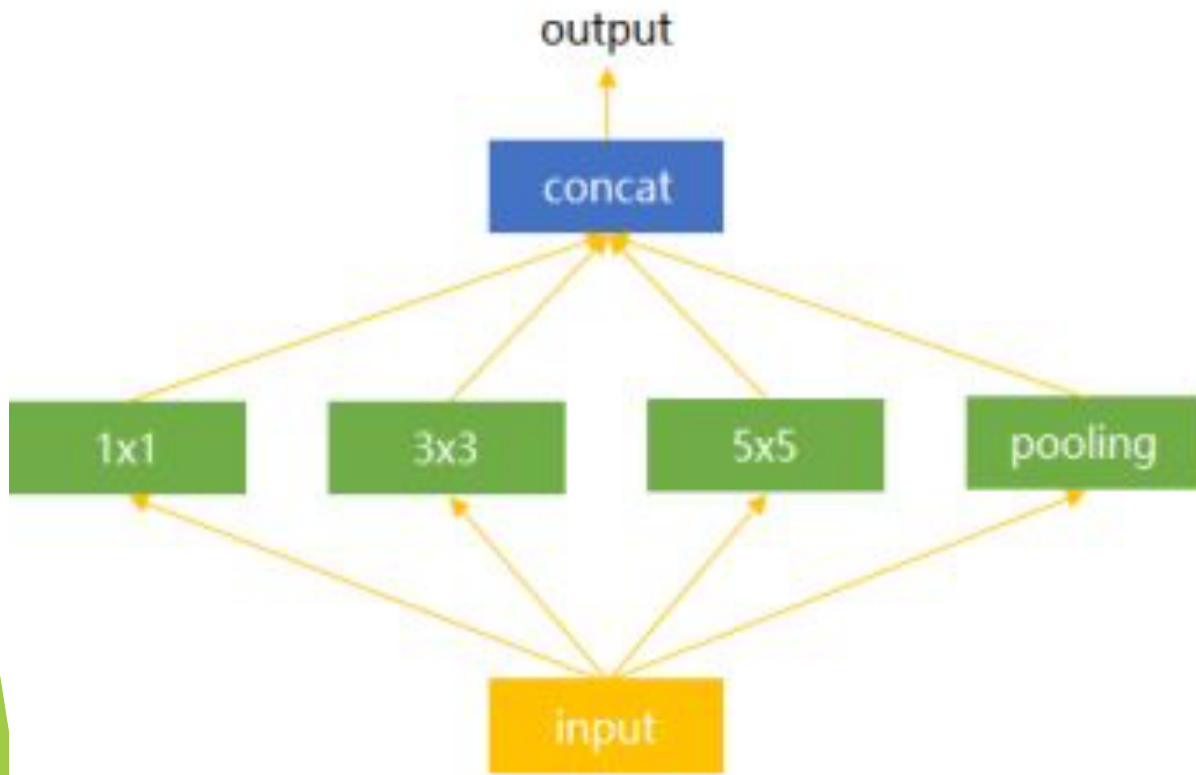
尺寸不变形 1 - Inception网络

消除尺寸对于识别结果的影响，一次性使用多个不同filter size来抓取多个范围不同的概念，并让网络自己选择需要的特征。



作者：YJango

尺寸不变形 2 - Inception网络

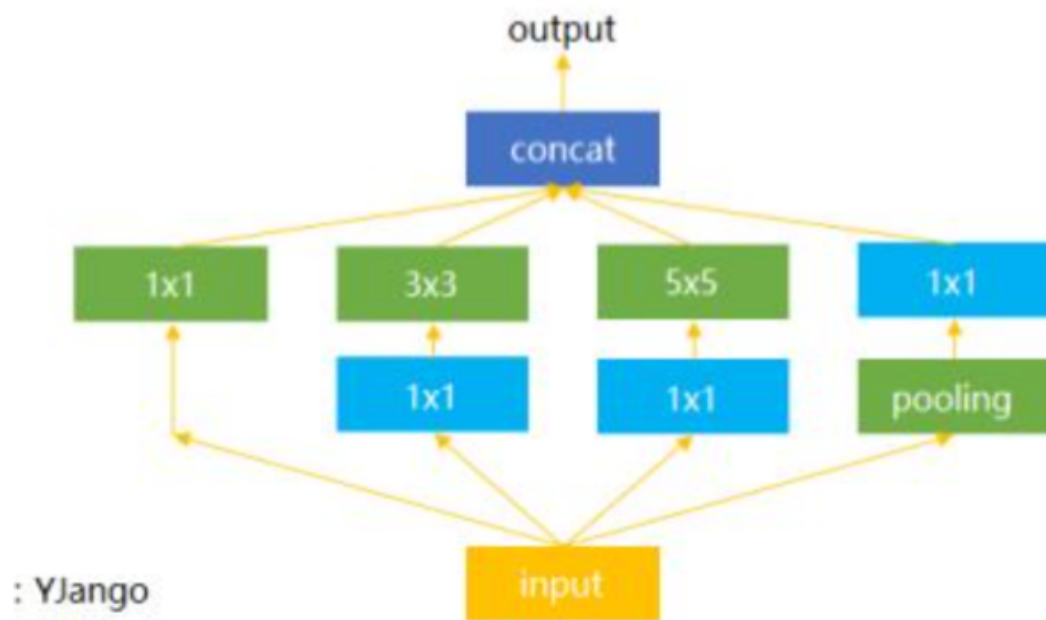


输入（可以是被卷积完的长方体输出作为该层的输入）进来后，通常我们可以选择直接使用像素信息(1x1卷积)传递到下一层，可以选择3x3卷积，可以选择5x5卷积，还可以选择max pooling的方式downsample刚被卷积后的feature maps。但在实际的网络设计中，究竟该如何选择需要大量的实验和经验的。

Inception就不用我们来选择，而是将4个选项给神经网络，让网络自己去选择最合适的解决方案。

尺寸不变形 2 - Inception网络

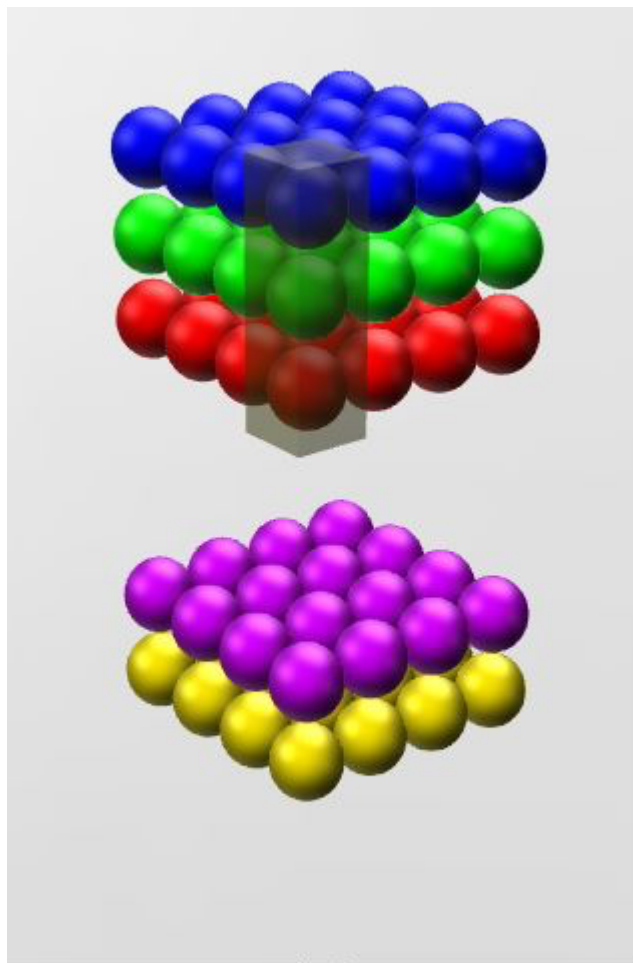
1x1卷积核



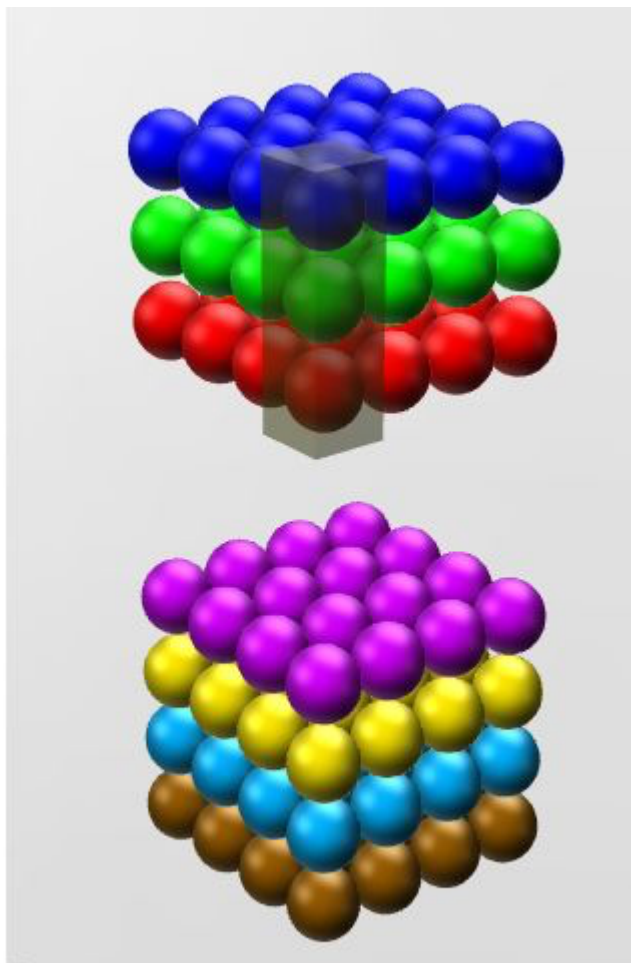
多了很多蓝色的1x1卷积。这些1x1卷积的作用是为了让网络根据需要能够更灵活的控制数据的depth的。

如果卷积的输出输入都只是一个平面，那么1x1卷积核并没有什么意义，它是完全不考虑像素与周边其他像素关系。但卷积的输出输入是长方体，所以1x1卷积实际上是对每个像素点，在不同的channels上进行线性组合（信息整合），且保留了图片的原有平面结构，调控depth，从而完成升维或降维的功能。

多filter的升维和降维



降维



升维

如果选择2个filters的1x1卷积层，那么数据就从原本的depth 3 降到了2。若用4个filters，则起到了升维的作用。