全连接网络到卷积神经网络逐步推导

云栖君导读：在图像分析中，卷积神经网络（Convolutional Neural Networks, CNN）在时间和内存方面优于全连接网络（Full Connected, FC）。这是为什么呢?

卷积神经网络优于全连接网络的优势是什么呢?

卷积神经网络是如何从全连接网络中派生出来的呢?

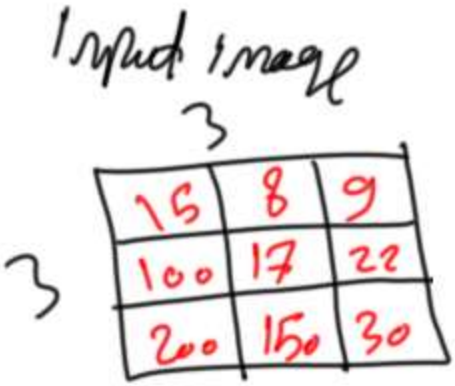
卷积神经网络这个术语又是从哪里而来?

一、介绍

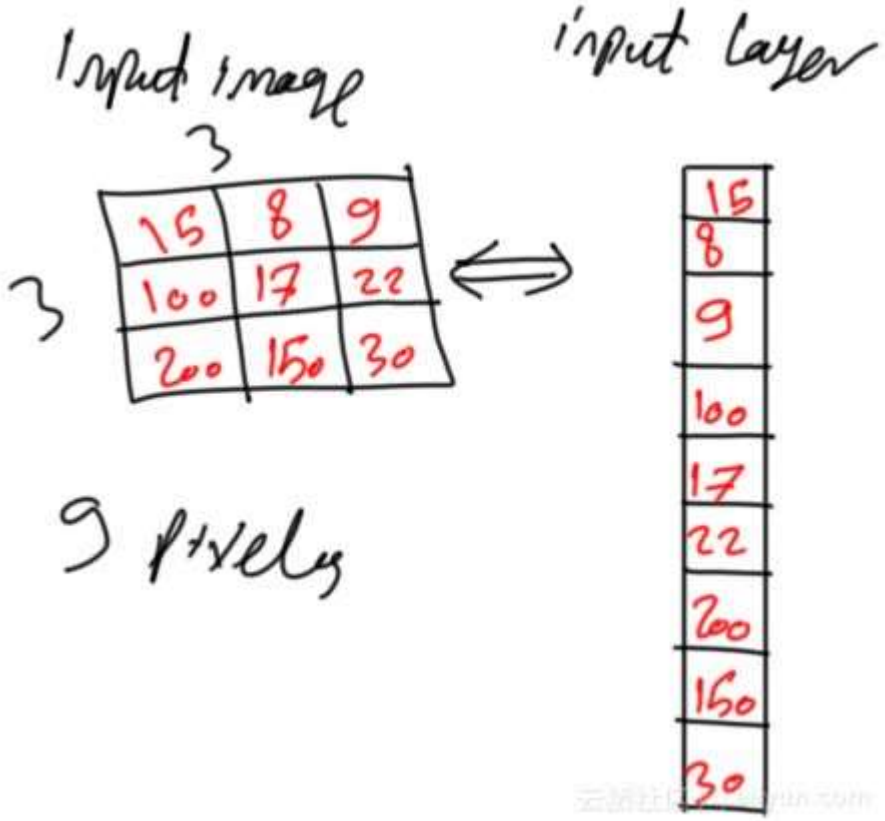
对于图像分析而言，具体可以将其划分为很多类型的任务，比如分类、对象检测、识别、描述等。对于图像分类器而言，即使在诸如遮挡、照明变化、视觉等变化的情况下，也应该能够以高精度的性能工作。以特征工程为主要步骤的传统图像分类方法不适合在丰富环境中工作，即使是该领域的专家也不能给出一组能够在不同变化下达到高精度的特征，无法保证手工所选的特征是否合适。在这个问题的启发下，特征学习的思想应运而生，通过自主学习来获得合适的图像特征，这也是人工神经网络（ANN）对于图像分析任务鲁棒性的原因之一。基于梯度下降算法（GD）等学习算法，ANN可以自动学习到图像特征，将原始图像输入人工神经网络后，ANN能够自动地生成描述它的特征。

二、基于全连接网络的图像分析

现在一起看看人工神经网络是如何对进行处理的，以及CNN为什么在时间和内存上相较于全连接网络更高效。如图1所示，输入的是一个3x3大小的灰度图。例子中使用小尺寸的图像是为了方便讲解，而不是表明ANN只能处理小尺寸的图像。

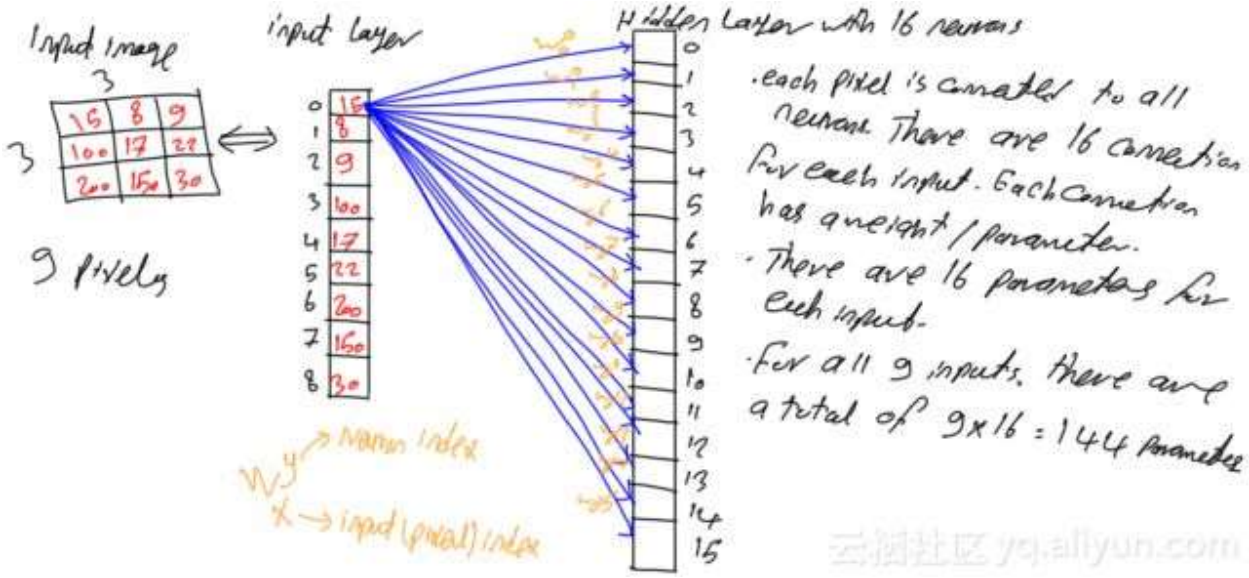


在输入ANN时，图像会转变为像素矩阵。由于ANN使用的是一维向量，而不是二维矩阵，所以将输入的二维灰度图转换成一维向量，其中每个像素点代表一个输入神经元节点。

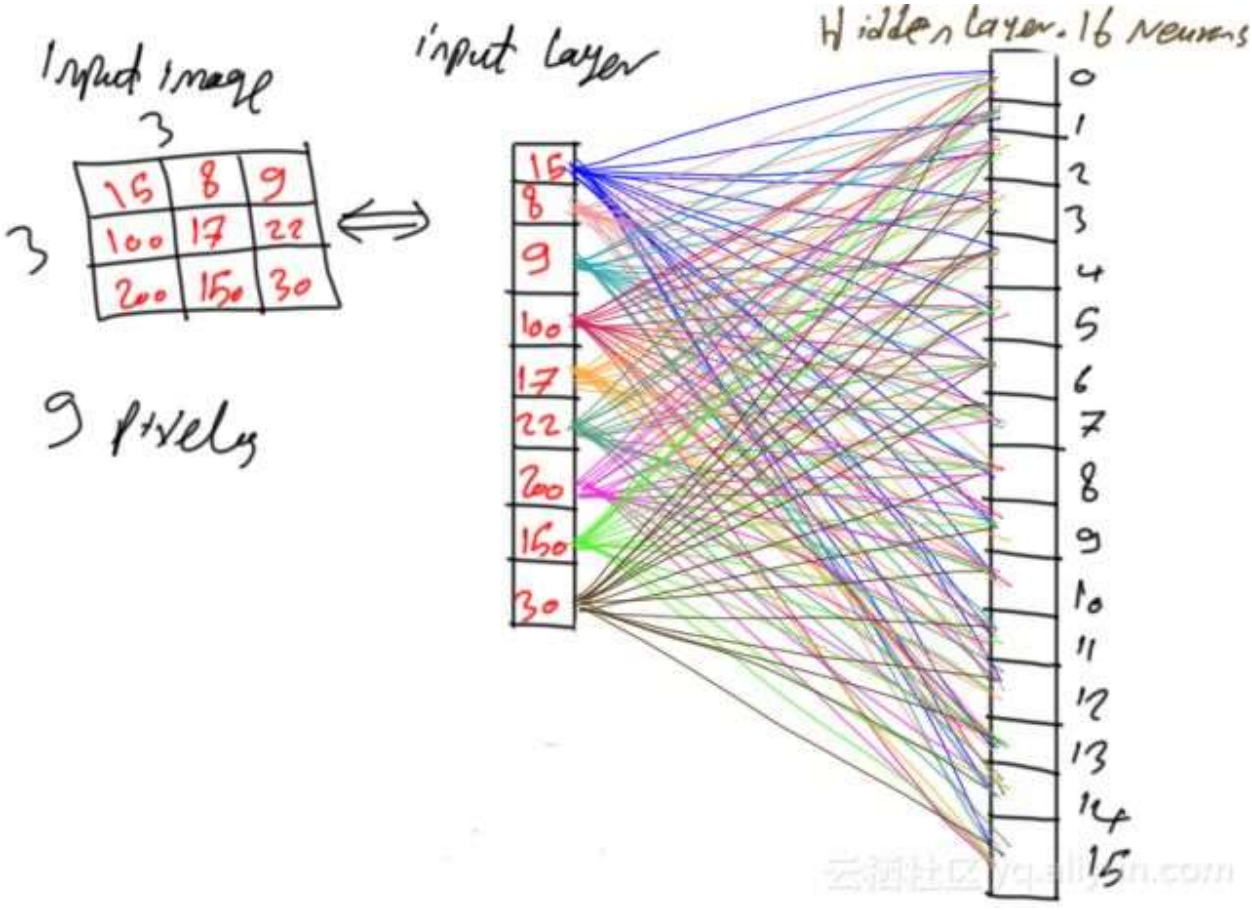


每个像素被映射为向量元素，向量中的每个元素又代表ANN中的神经元。由于图像有3x3=9个像素点，那么输入层（Input Layer）将有9个神经元。由于ANN结构通常水平延伸，因此每层被表示为列向量。

输入层与隐藏层（Hidden Layer）相连，输入层的输出又输入给隐藏层，隐藏层学习如何将图像像素转换为代表性特征。假设在图3中有一个具由16个神经元的单个隐藏层。



由于网络是全连接网络，这意味着第i层的每个神经元与第i-1层中的所有神经元相连。即隐藏层中的每个神经元都与输入层中9个神经元相连。换句话说，每个输入像素与隐藏层中的16个神经元相连，其中每条连接都具有相应的参数（权重）。通过将每个像素与隐藏层中的所有神经元相连，如图4所示，该网络具有9x16=144个参数（权重）。



图像4

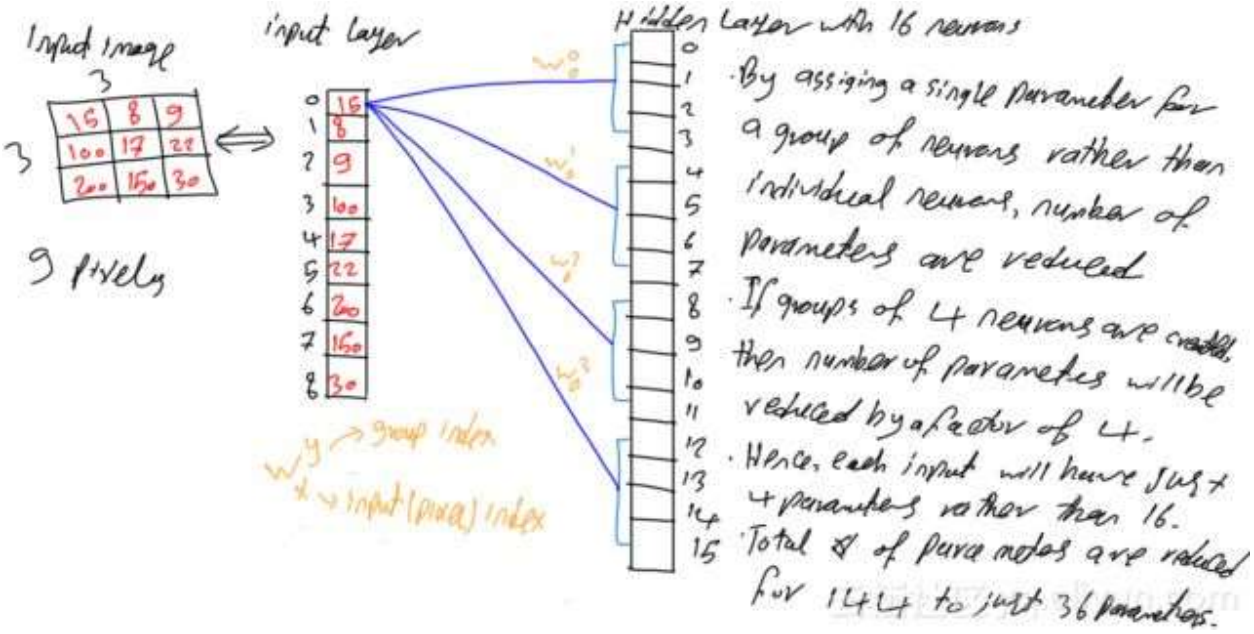
三、大量参数

上面例子中的参数数目似乎还可以接受，但是随着输入图像尺寸变大以及隐藏层数量增加，网络参数将大大增加。

例如，若网络具有两个隐层，分别有90和50个神经元，那么输入层和第一隐藏层之间的参数数目是9x90=810，两个隐藏层之间的参数数目为90x50=4500，该网络的参数总数为810+4500=5310。对于这样简单的网络结构就有这么多的参数数量，显然是不合适的；另外一种情况是输入图像尺寸较大，比如32x32大小的图像（1024个像素），如果网络使用单个隐藏层（含有500个神经元），则总共有1024x500=512000个参数（权重），这对于只含单个隐藏层的网络而言是一个巨大的数字。因此，必须有一个解决方案来减少网络参数，那么针对于此，卷积神经网络（CNN）应运而生，虽然它网络模型通常比较大，但大大降低了参数数量。

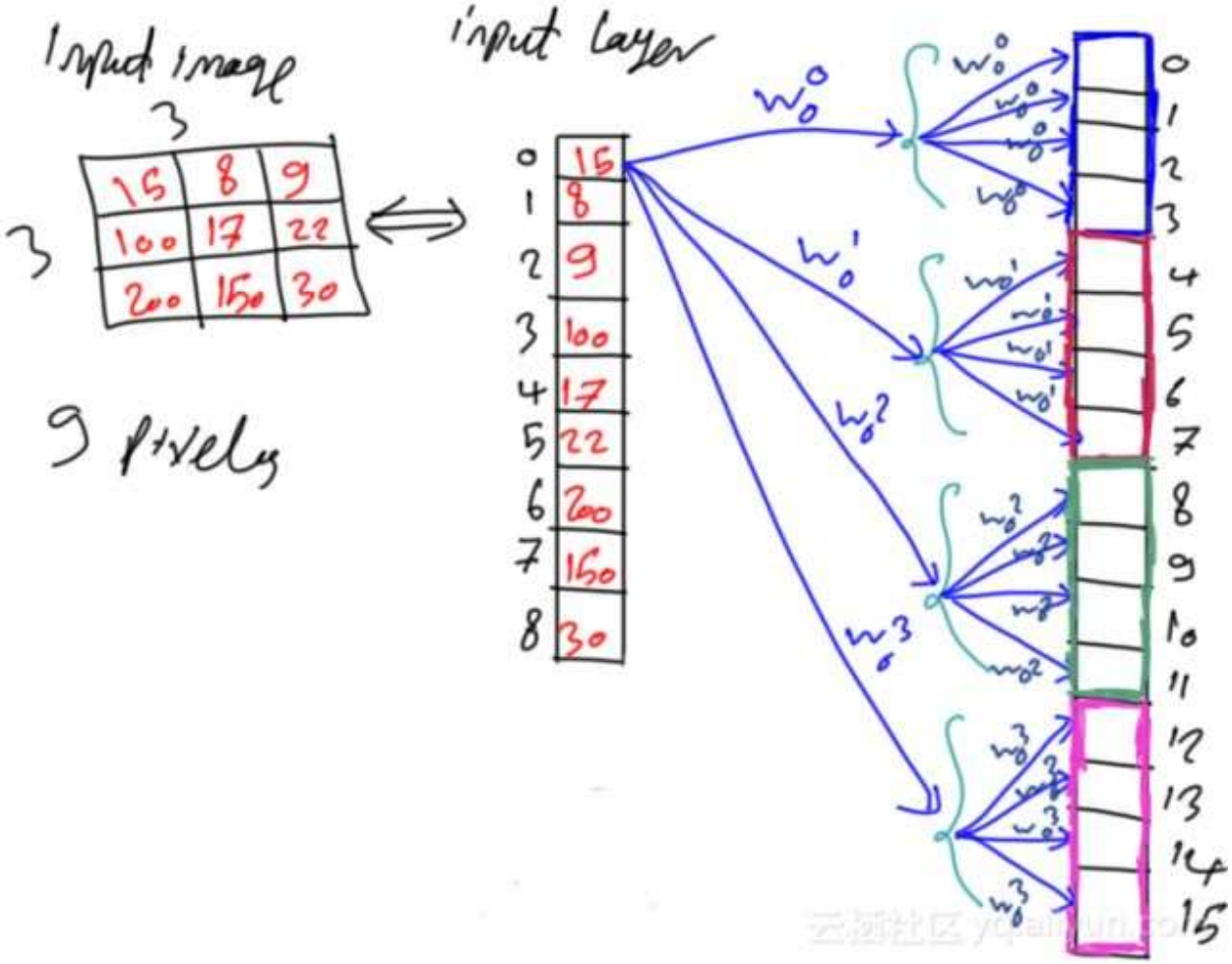
四、神经元组群

即使是很小的全连接网络，网络参数数目变得非常大的原因在于其层与层之间神经元每条连接上都是不同的参数。因此，可以考虑给一组神经元提供相同的参数，如图5所示，一组神经元内的神经元都将分配同一个参数。

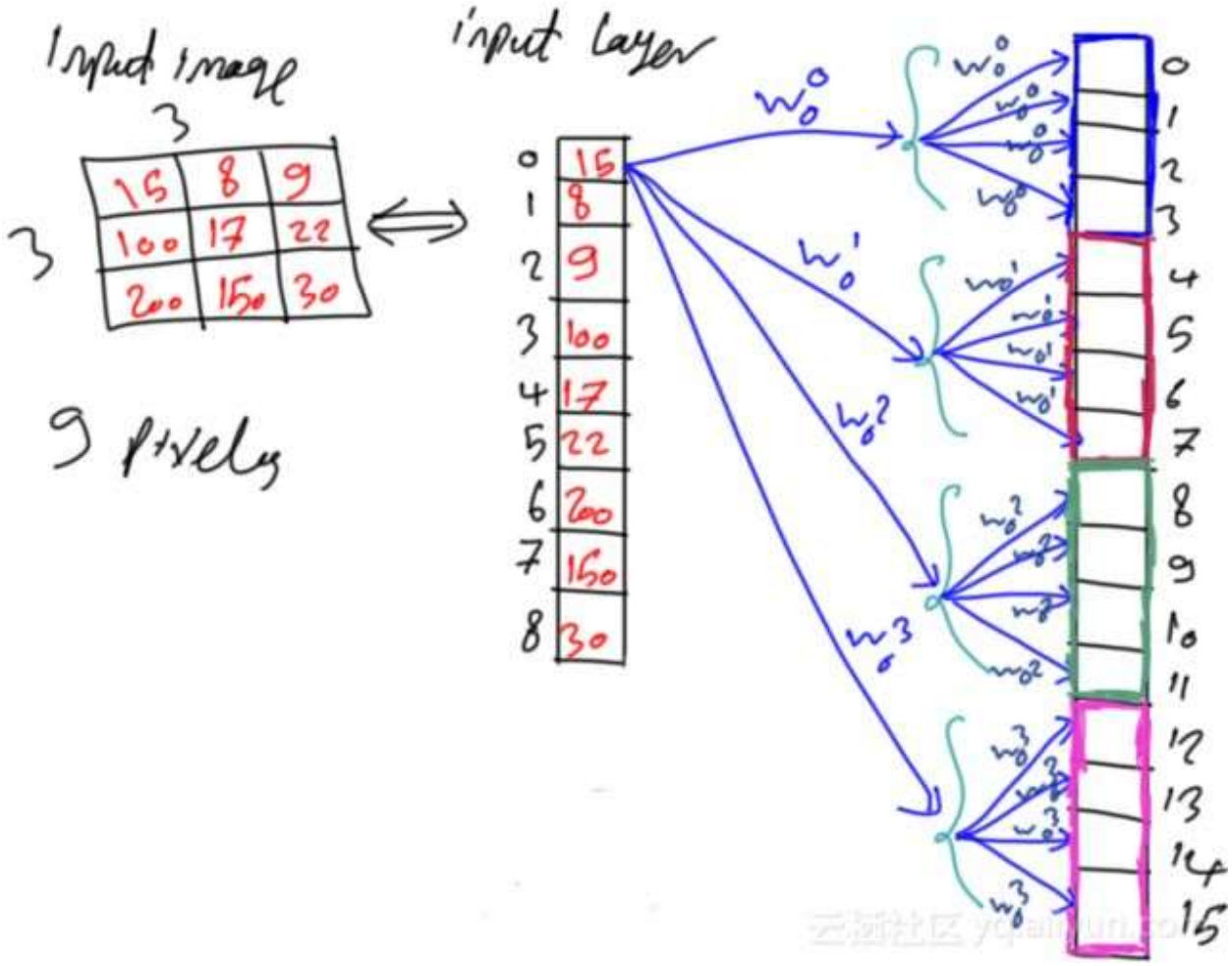


这样处理以后，网络参数数量大大降低。以图4为例，比如每4个连续神经元作为一组，其结果是参数

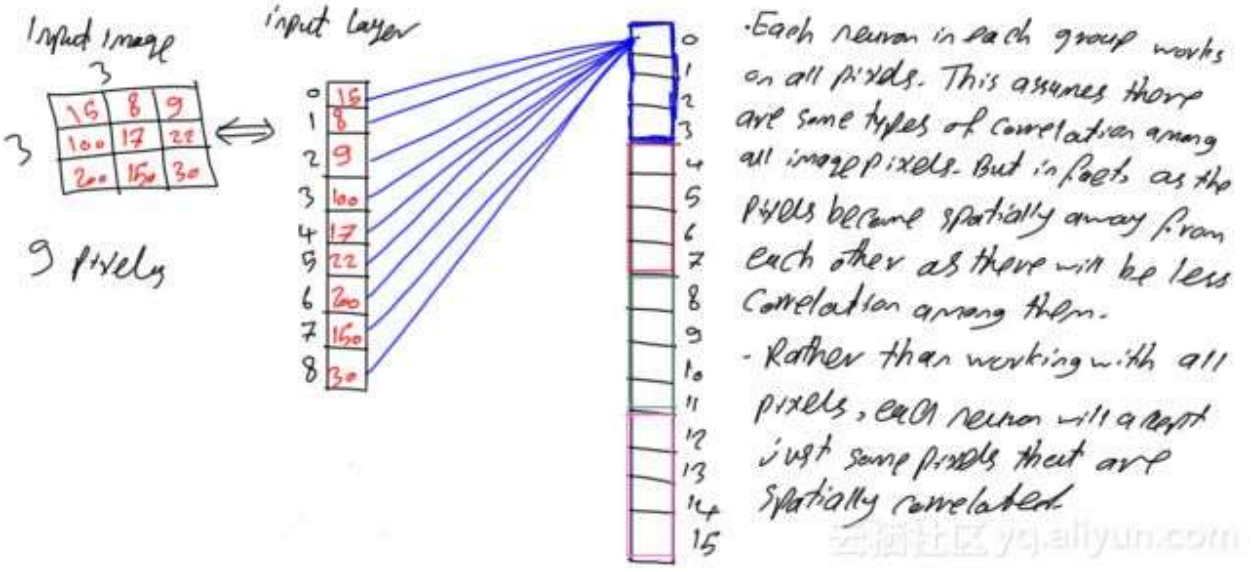
数量减少了4倍。每个输入神经元将具有16/4=4个参数。整个网络将具有144/4=36个参数，参数数量减少了75%。可以看到，其效果很好，但仍然有可优化的地方。



下图显示了每个像素到每个组中第一个神经元的连接，但每组中的每个像素与每个神经元还是相互连接，该网络仍然是全连接网络。



为了简单起见，只挑选出一组并忽略其它组，如下图所示。从图中可以看到，每个组仍然与输入层所有的9个神经元有所连接，因此具有9个参数。

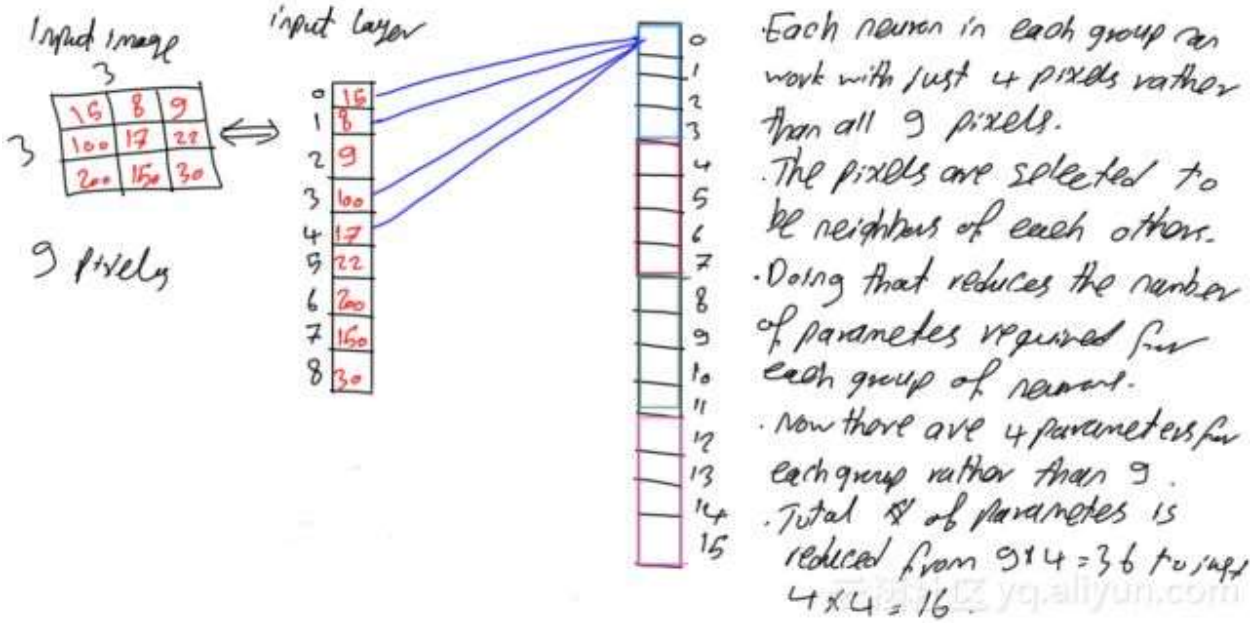


五、像素空间相关性

之前所述内容使得每个神经元接受所有像素，若存在接受4个输入的函数f(x1,x2,x3,x4)，则这意味着要基于所有这4个输入来进行决定。如果只有2个输入，但其输出结果与使用4个输入的结果相同，那么将不必使用所有的这4个输入，只需给出对结果有影响的2个输入即可。借鉴该思想，每个神经元接受输入的9个像素，**若能使用更少的像素获得相同或更好的结果就大大降低了参数数量，因此可以朝着这个方向优化网络参数。**

通常，在图像分析中，输入图像被转换为像素矩阵，像素矩阵中的每个像素与它周围的像素高度相关，两个像素之间的距离越远，二者越不相关。例如，如图9所示，面部的像素与面部周围的像素相关，但它与天空、地面等像素的相关性较低。

基于这样的假设，上述示例中的每个神经元只接受彼此空间相关的像素，而不是将所有9个像素点都应用到每个输入神经元中，因此可以选择4个空间相关像素，如图10所示。对于像素矩阵位置（0,0），那么空间上最相关的像素是坐标点（0,1）、（1,0）以及（1,1）。同一组中的所有神经元共享相同的权重，那么每组中的4个神经元将只有4个参数而不是9个。总的参数变为4x4=16。与图4中的全连接网络相比，减少了128个参数（减少了88.89%）。

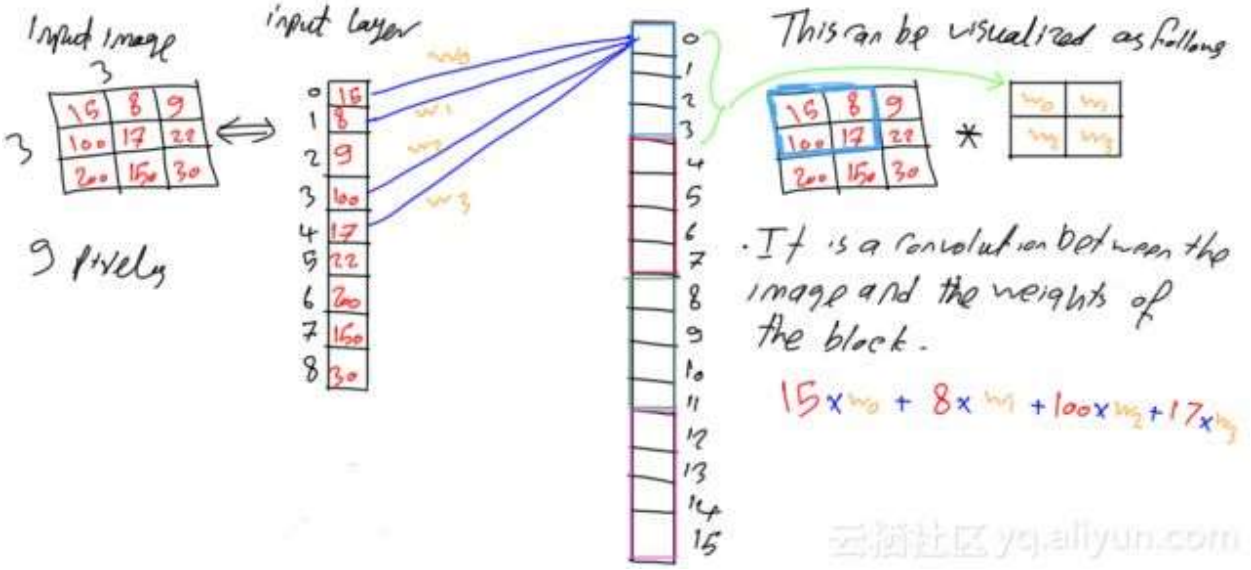


六、卷积神经网络（CNN）

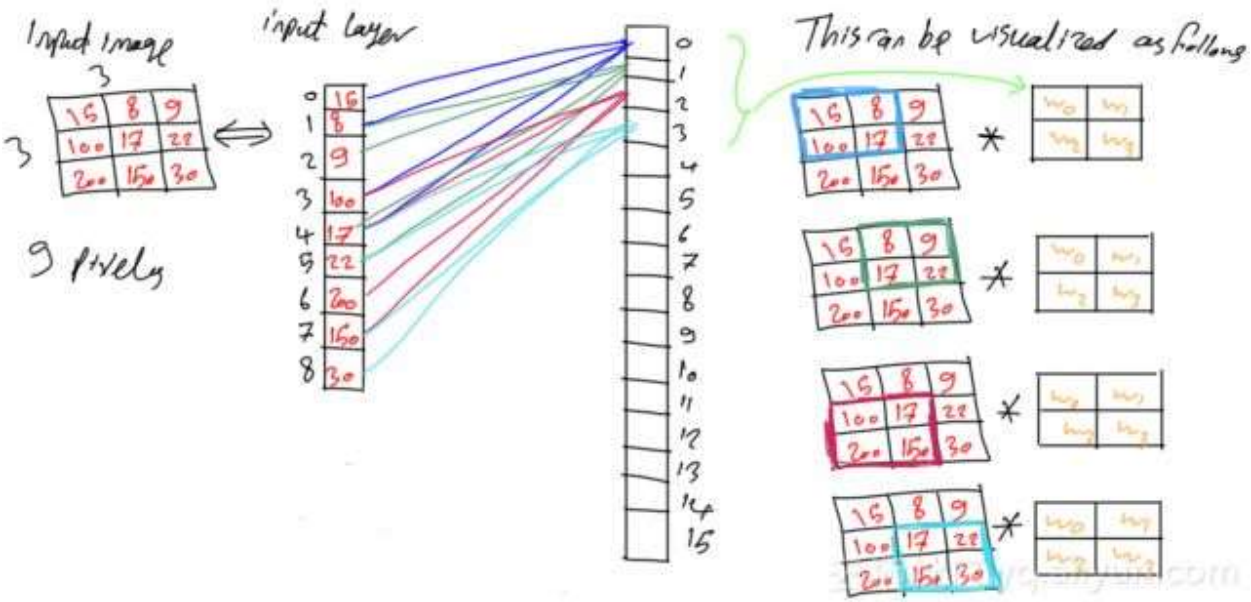
由于CNN使用权重共享，使用较少的参数，这使得CNN网络结构一般层数比较多，这是全连接网络无法具有的特性。

现在只有4个权重分配给同一组中的所有神经元，那么这4个权重如何涵盖9个像素点呢？让我们看看这是如何处理的吧！

下图展示了上图中的一个网络，并为每条连接添加了权重标记。在神经元内部，4个输入像素中的每一个都与其相应的权重相乘，如下图中公式所示。



假设这里每次移动的步长设置为1（步长可以自己设置），每次相乘后将像素点索引移动一位，权重矩阵与另外一组像素相乘。以此类推，直到整个像素矩阵都与权重矩阵进行了相乘运算。整个过程与卷积运算相同，组的权重与图像矩阵之间进行卷积运算，这也是CNN有“卷积”一词的原因。



剩余的神经元组也会进行同样的操作，从像素矩阵的左上角开始，直到像素矩阵的右下角都与权重矩阵相乘。

Ahmed Gad，教师、专注于深度学习、计算机视觉

本文由阿里云云栖社区组织翻译。

文章原标题《Derivation of Convolutional Neural Network from Fully Connected Network Step-By-Step》

译者：海棠

审校：Uncle_LLD