原 Deep Learning（深度学习）学习笔记整理系列之（七）

分类： Deep Learning 机器学习 Linux驱动 2013-04-10 10:48 800人阅读

Deep Learning（深度学习）学习笔记整理系列

zouxy09@qq.com

http://blog.csdn.net/zouxy09

作者：**Zouxy**

version 1.0 2013-04-08

声明：

1）该Deep Learning的学习系列是整理自网上很大牛和机器学习专家所无私奉献的资料的。具体引用的资料请看参考文献。具体的版本声明也参考原文献。

2）本文仅供学术交流，非商用。所以每一部分具体的参考资料并没有详细对应。如果某部分不小心侵犯了大家的利益，还望海涵，并联系博主删除。

3）本人才疏学浅，整理总结的时候难免出错，还望各位前辈不吝指正，谢谢。

4）阅读本文需要机器学习、计算机视觉、神经网络等等基础（如果没有也没关系了，没有就看看，能不能看懂，呵呵）。

5）此属于第一版本，若有错误，还需继续修正与增删。还望大家多多指点。大家都共享一点点，一起为祖国科研的推进添砖加瓦（呵呵，好高尚的目标啊）。请联系：zouxy09@qq.com

目录：

接上

9.5、Convolutional Neural Networks卷积神经网络

卷积神经网络是人工神经网络的一种，已成为当前语音分析和图像识别领域的研究热点。它的权值共享网络结构使之更类似于生物神经网络，降低了网络模型的复杂度，减少了权值的数量。该优点在网络的输入是多维图像时表现的更为明显，使图像可以直接作为网络的输入，避免了传统识别算法中复杂的特征提取和数据重建过程。卷积网络是为识别二维形状而特殊设计的一个多层感知器，这种网络结构对平移、比例缩放、倾斜或者共他形式的变形具有高度不变性。

CNNs是受早期的延时神经网络（TDNN）的影响。延时神经网络通过在时间维

度上共享权值降低学习复杂度，适用于语音和时间序列信号的处理。

CNNs是第一个真正成功训练多层网络结构的学习算法。它利用空间关系减少需要学习的参数数目以提高一般前向BP算法的训练性能。CNNs作为一个深度学习架构提出是为了最小化数据的预处理要求。在CNN中，图像的一小部分（局部感受区域）作为层级结构的最低层的输入，信息再依次传输到不同的层，每层通过一个数字滤波器去获得观测数据的最显著的特征。这个方法能够获取对平移、缩放和旋转不变的观测数据的显著特征，因为图像的局部感受区域允许神经元或者处理单元可以访问到最基础的特征，例如定向边缘或者角点。

1）卷积神经网络的历史

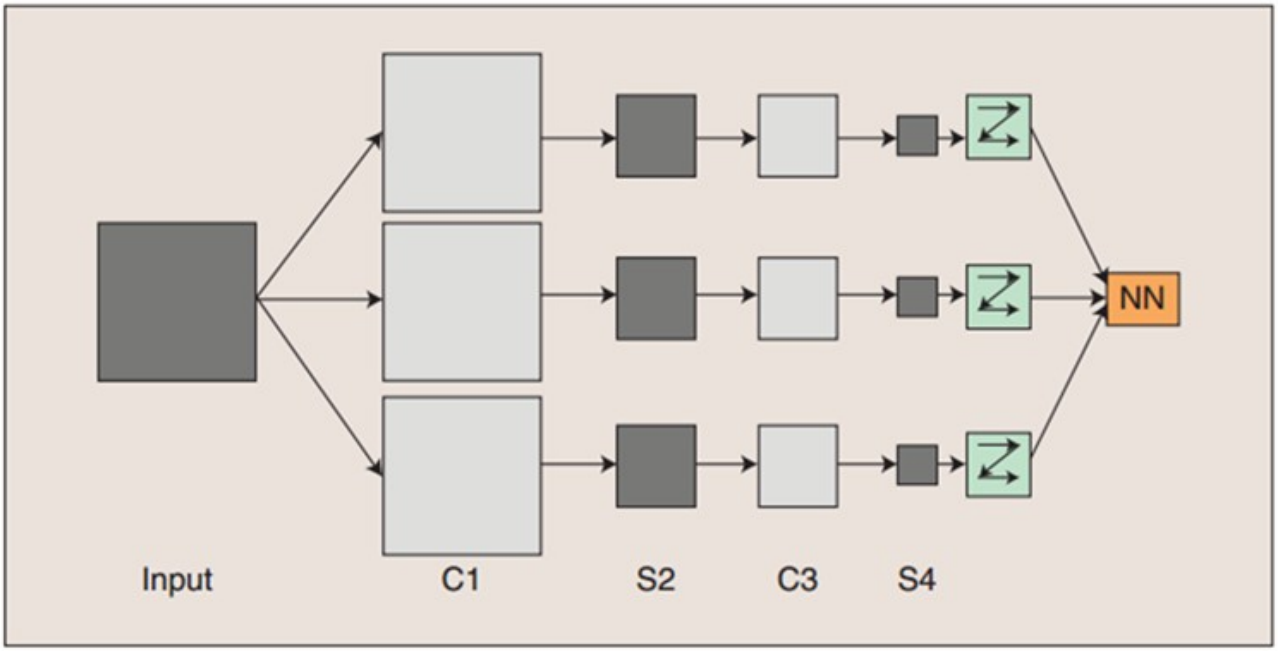
1962年Hubel和Wiesel通过对猫视觉皮层细胞的研究，提出了感受野(receptive field)的概念，1984年日本学者Fukushima基于感受野概念提出的神经认知机(neocognitron)可以看作是卷积神经网络的第一个实现网络，也是感受野概念在人工神经网络领域的首次应用。神经认知机将一个视觉模式分解成许多子模式（特征），然后进入分层递阶式相连的特征平面进行处理，它试图将视觉系统模型化，使其能够在即使物体有位移或轻微变形的时候，也能完成识别。

通常神经认知机包含两类神经元，即承担特征抽取的S-元和抗变形的C-元。S-元中涉及两个重要参数，即感受野与阈值参数，前者确定输入连接的数目，后者则控制对特征子模式的反应程度。许多学者一直致力于提高神经认知机的性能的研究：在传统的神经认知机中，每个S-元的感光区中由C-元带来的视觉模糊量呈正态分布。如果感光区的边缘所产生的模糊效果要比中央来得大，S-元将会接受这种非正态模糊所导致的更大的变形容忍性。我们希望得到的是，训练模式与变形刺激模式在感受野的边缘与其中心所产生的效果之间的差异变得越来越大。为了有效地形成这种非正态模糊，Fukushima提出了带双C-元层的改进型神经认知机。

Van Ooyen和Niehuis为提高神经认知机的区别能力引入了一个新的参数。事实上，该参数作为一种抑制信号，抑制了神经元对重复激励特征的激励。多数神经网络在权值中记忆训练信息。根据Hebb学习规则，某种特征训练的次数越多，在以后的识别过程中就越容易被检测。也有学者将进化计算理论与神经认知机结合，通过减弱对重复性激励特征的训练学习，而使得网络注意那些不同的特征以助于提高区分能力。上述都是神经认知机的发展过程，而卷积神经网络可看作是神经认知机的推广形式，神经认知机是卷积神经网络的一种特例。

2）卷积神经网络的网络结构

卷积神经网络是一个多层的神经网络，每层由多个二维平面组成，而每个平面由多个独立神经元组成。



图：卷积神经网络的概念示范：输入图像通过和三个可训练的滤波器和可加偏置进行卷积，滤波过程如图一，卷积后在C1层产生三个特征映射图，然后特征映射图中每组的四个像素再进行求和，加权值，加偏置，通过一个Sigmoid函数得到三个S2层的特征映射图。这些映射图再进过滤波得到C3层。这个层级结构再和S2一样产生S4。最终，这些像素值被光栅化，并连接成一个向量输入到传统的神经网络，得到输出。

一般地，C层为特征提取层，每个神经元的输入与前一层的局部感受野相连，并提取该局部的特征，一旦该局部特征被提取后，它与其他特征间的位置关系也随之确定下来；S层是特征映射层，网络的每个计算层由多个特征映射组成，每个特征映射为一个平面，平面上所有神经元的权值相等。特征映射结构采用影响函数核小的sigmoid函数作为卷积网络的激活函数，使得特征映射具有位移不变性。

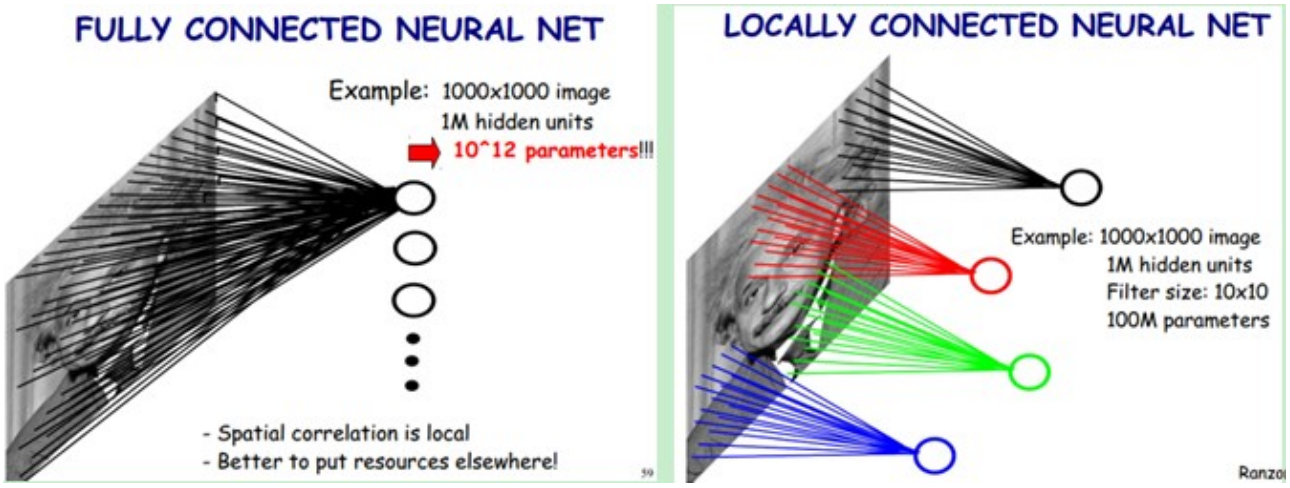
此外，由于一个映射面上的神经元共享权值，因而减少了网络自由参数的个数，降低了网络参数选择的复杂度。卷积神经网络中的每一个特征提取层（C-层）都紧跟着一个用来求局部平均与二次提取的计算层（S-层），这种特有的两次特征提取结构使网络在识别时对输入样本有较高的畸变容忍能力。

3）关于参数减少与权值共享

上面聊到，好像CNN一个牛逼的地方就在于通过感受野和权值共享减少了神经网络需要训练的参数的个数。那究竟是啥的呢？

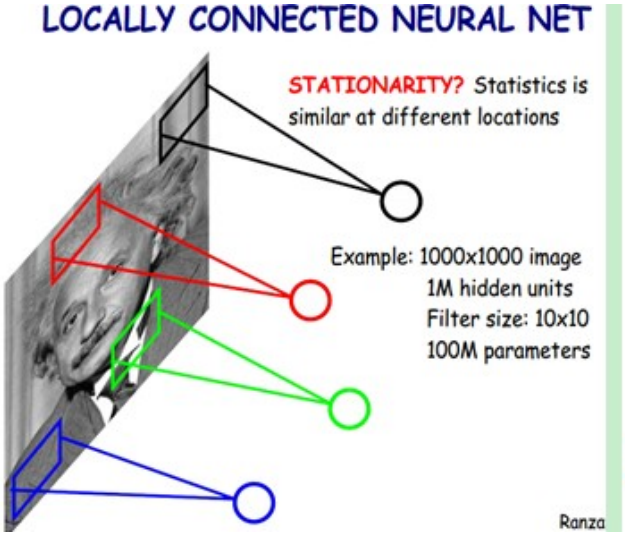
下图左：如果我们有1000x1000像素的图像，有1百万个隐层神经元，那么他们全连接的话（每个隐层神经元都连接图像的每一个像素点），就有1000x1000x1000000=10^12个连接，也就是10^12个权值参数。然而图像的空间联系是局部的，就像人是通过一个局部的感受野去感受外界图像一样，每一个神经元都不

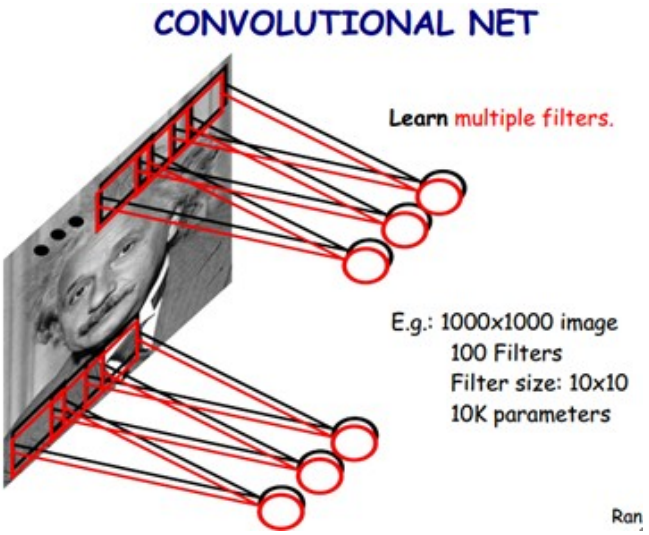
需要对全局图像做感受，每个神经元只感受局部的图像区域，然后在更高层，将这些感受不同局部的神经元综合起来就可以得到全局的信息了。这样，我们就可以减少连接的数目，也就是减少神经网络需要训练的权值参数的个数了。如下图右：假如局部感受野是10x10，隐层每个感受野只需要和这10x10的局部图像相连接，所以1百万个隐层神经元就只有一亿个连接，即10^8个参数。比原来减少了四个0（数量级），这样训练起来就没那么费力了，但还是感觉很多的啊，那还有啥办法没？



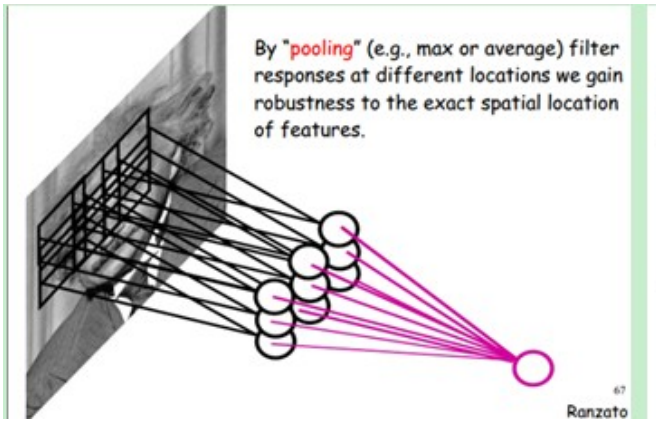
我们知道，隐含层的每一个神经元都连接10x10个图像区域，也就是说每一个神经元存在10x10=100个连接权值参数。那如果我们每个神经元这100个参数是相同的呢？也就是说每个神经元用的是同一个卷积核去卷积图像。这样我们就只有多少个参数？？只有100个参数啊！！！亲！不管你隐层的神经元个数有多少，两层间的连接我只有100个参数啊！亲！这就是权值共享啊！亲！这就是卷积神经网络的主打卖点啊！亲！（有点烦了，呵呵）也许你会问，这样做靠谱吗？为什么可行呢？这个……共同学习。

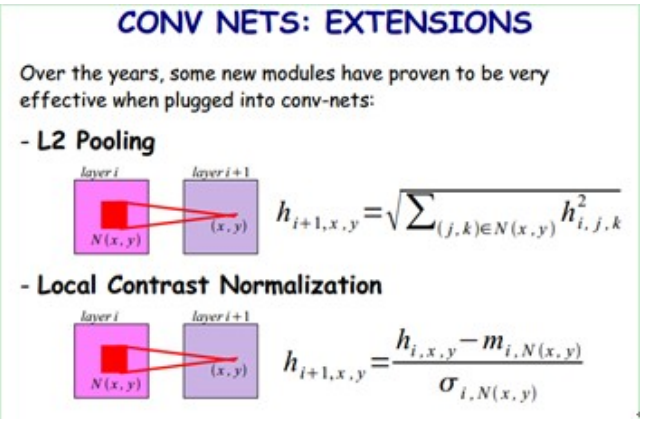
好了，你就会想，这样提取特征也忒不靠谱吧，这样你只提取了一种特征啊？对了，真聪明，我们需要提取多种特征对不？假如一种滤波器，也就是一种卷积核就是提出图像的一种特征，例如某个方向的边缘。那么我们需要提取不同的特征，怎么办，加多几种滤波器不就行了吗？对了。所以假设我们加到100种滤波器，每种滤波器的参数不一样，表示它提出输入图像的不同特征，例如不同的边缘。这样每种滤波器去卷积图像就得到对图像的不同特征的放映，我们称之为Feature Map。所以100种卷积核就有100个Feature Map。这100个Feature Map就组成了一层神经元。到这个时候明了了吧。我们这一层有多少个参数了？100种卷积核x每种卷积核共享100个参数=100x100=10K，也就是1万个参数。才1万个参数啊！亲！（又来了，受不了了！）见下图右：不同的颜色表达不同的滤波器。





嘿哟，遗漏一个问题了。刚才说隐层的参数个数和隐层的神经元个数无关，只和滤波器的大小和滤波器种类的多少有关。那么隐层的神经元个数怎么确定呢？它和原图像，也就是输入的大小（神经元个数）、滤波器的大小和滤波器在图像中的滑动步长都有关！例如，我的图像是1000x1000像素，而滤波器大小是10x10，假设滤波器没有重叠，也就是步长为10，这样隐层的神经元个数就是(1000x1000)/ (10x10)=100x100个神经元了，假设步长是8，也就是卷积核会重叠两个像素，那么……我就不算了，思想懂了就好。注意了，这只是一种滤波器，也就是一个Feature Map的神经元个数哦，如果100个Feature Map就是100倍了。由此可见，图像越大，神经元个数和需要训练的权值参数个数的贫富差距就越大。





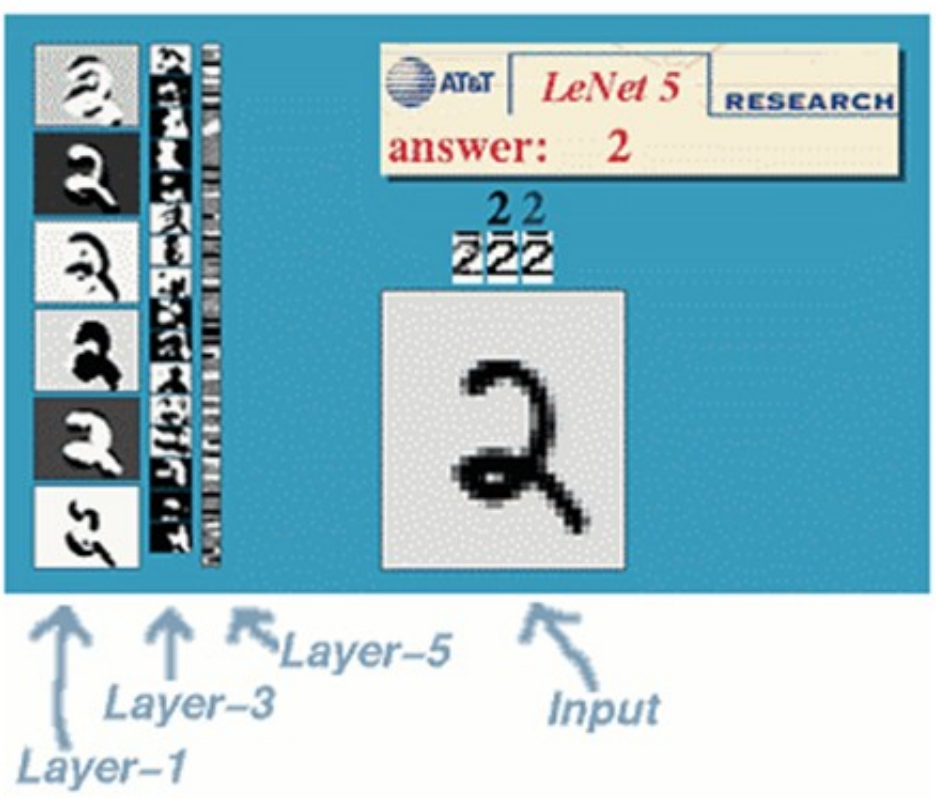
需要注意的一点是，上面的讨论都没有考虑每个神经元的偏置部分。所以权值个数需要加1 。这个也是同一种滤波器共享的。

总之，卷积网络的核心思想是将：局部感受野、权值共享（或者权值复制）以及时间或空间亚采样这三种结构思想结合起来获得了某种程度的位移、尺度、形变不变性。

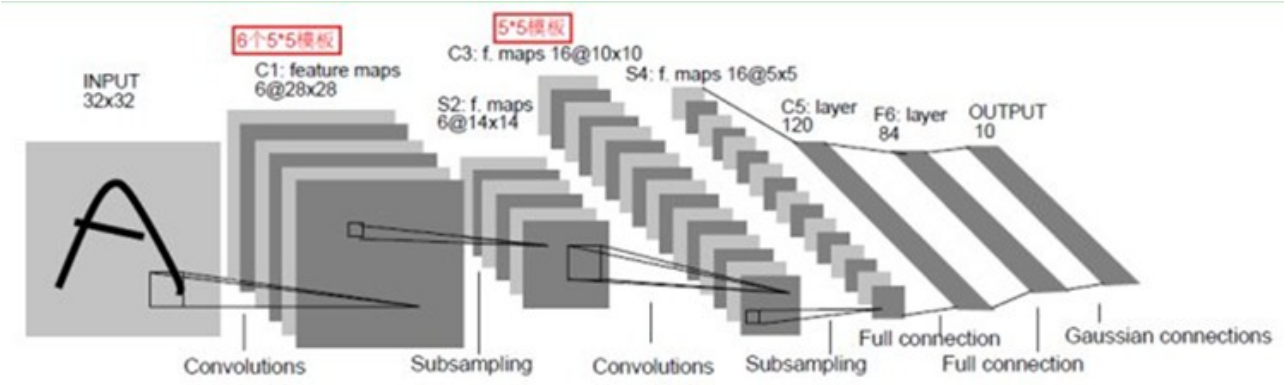
4）一个典型的例子说明

一种典型的用来识别数字的卷积网络是LeNet-5（效果和paper等见这）。当年美

国大多数银行就是用它来识别支票上面的手写数字的。能够达到这种商用的地步，它的准确性可想而知。毕竟目前学术界和工业界的结合是最受争议的。



那下面咱们也用这个例子来说明下。



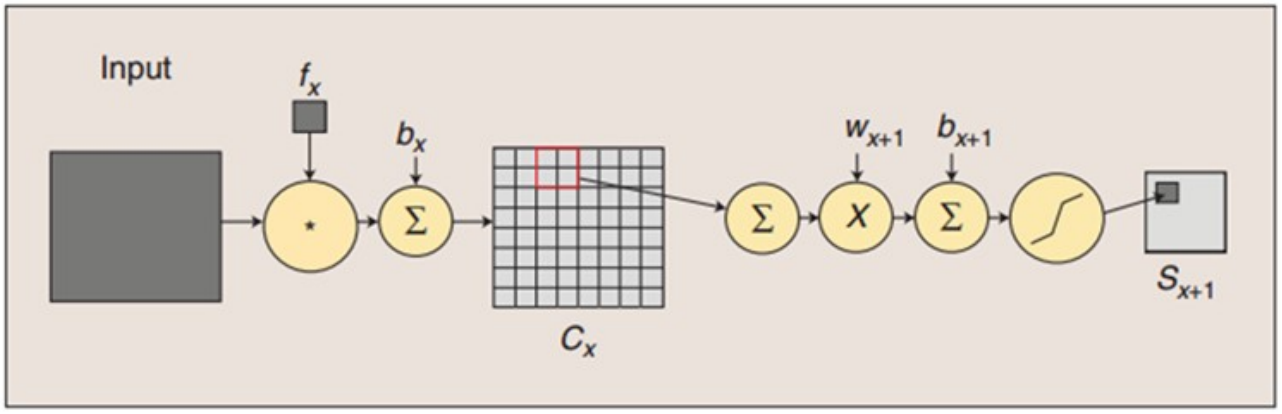
LeNet-5共有7层，不包含输入，每层都包含可训练参数（连接权重）。输入图像为32*32大小。这要比Mnist数据库（一个公认的手写数据库）中最大的字母还大。这样做的原因是希望潜在的明显特征如笔画断电或角点能够出现在最高层特征监测子感受野的中心。

我们先要明确一点：每个层有多个Feature Map，每个Feature Map通过一种卷积滤波器提取输入的一种特征，然后每个Feature Map有多个神经元。

C1层是一个卷积层（为什么是卷积？卷积运算一个重要的特点就是，通过卷积运算，可以使原信号特征增强，并且降低噪音），由6个特征图Feature Map构成。特征图中每个神经元与输入中5*5的邻域相连。特征图的大小为28*28，这样能防止输入的连接掉到边界之外（是为了BP反馈时的计算，不致梯度损失，个人见解）。C1有

156个可训练参数（每个滤波器5*5=25个unit参数和一个bias参数，一共6个滤波器，共(5*5+1)*6=156个参数），共156*(28*28)=122,304个连接。

S2层是一个下采样层（为什么是下采样？利用图像局部相关性的原理，对图像进行子抽样，可以减少数据处理量同时保留有用信息），有6个14*14的特征图。特征图中的每个单元与C1中相对应特征图的2*2邻域相连接。S2层每个单元的4个输入相加，乘以一个可训练参数，再加上一个可训练偏置。结果通过sigmoid函数计算。可训练系数和偏置控制着sigmoid函数的非线性程度。如果系数比较小，那么运算近似于线性运算，亚采样相当于模糊图像。如果系数比较大，根据偏置的大小亚采样可以被看成是有噪声的“或”运算或者有噪声的“与”运算。每个单元的2*2感受野并不重叠，因此S2中每个特征图的大小是C1中特征图大小的1/4（行和列各1/2）。S2层有12个可训练参数和5880个连接。



图：卷积和子采样过程：卷积过程包括：用一个可训练的滤波器f_x去卷积一个输入的图像（第一阶段是输入的图像，后面的阶段就是卷积特征map了），然后加一个偏置b_x，得到卷积层C_x。子采样过程包括：每邻域四个像素求和变为一个像素，然后通过标量W_{x+1}加权，再增加偏置b_{x+1}，然后通过一个sigmoid激活函数，产生一个大概缩小四倍的特征映射图S_{x+1}。

所以从一个平面到下一个平面的映射可以看作是作卷积运算，S-层可看作是模糊滤波器，起到二次特征提取的作用。隐层与隐层之间空间分辨率递减，而每层所含的平面数递增，这样可用于检测更多的特征信息。

C3层也是一个卷积层，它同样通过5x5的卷积核去卷积层S2，然后得到的特征map就只有10x10个神经元，但是它有16种不同的卷积核，所以就存在16个特征map了。这里需要注意的一点是：C3中的每个特征map是连接到S2中的所有6个或者几个特征map的，表示本层的特征map是上一层提取到的特征map的不同组合（这个做法也并不是唯一的）。（看到没有，这里是组合，就像之前聊到的人的视觉系统一样，底层的结构构成上层更抽象的结构，例如边缘构成形状或者目标的部分）。

刚才说C3中每个特征图由S2中所有6个或者几个特征map组合而成。为什么不把

S2中的每个特征图连接到每个C3的特征图呢？原因有2点。第一，不完全的连接机制将连接的数量保持在合理的范围内。第二，也是最重要的，其破坏了网络的对称性。由于不同的特征图有不同的输入，所以迫使他们抽取不同的特征（希望是互补的）。

例如，存在的一个方式是：C3的前6个特征图以S2中3个相邻的特征图子集为输入。接下来6个特征图以S2中4个相邻特征图子集为输入。然后的3个以不相邻的4个特征图子集为输入。最后一个将S2中所有特征图为输入。这样C3层有1516个可训练参数和151600个连接。

S4层是一个下采样层，由16个5*5大小的特征图构成。特征图中的每个单元与C3中相应特征图的2*2邻域相连接，跟C1和S2之间的连接一样。S4层有32个可训练参数（每个特征图1个因子和一个偏置）和2000个连接。

C5层是一个卷积层，有120个特征图。每个单元与S4层的全部16个单元的5*5邻域相连。由于S4层特征图的大小也为5*5（同滤波器一样），故C5特征图的大小为1*1：这构成了S4和C5之间的全连接。之所以仍将C5标示为卷积层而非全相联层，是因为如果LeNet-5的输入变大，而其他的保持不变，那么此时特征图的维数就会比1*1大。C5层有48120个可训练连接。

F6层有84个单元（之所以选这个数字的原因来自于输出层的设计），与C5层全相连。有10164个可训练参数。如同经典神经网络，F6层计算输入向量和权重向量之间的点积，再加上一个偏置。然后将其传递给sigmoid函数产生单元i的一个状态。

最后，输出层由欧式径向基函数（Euclidean Radial Basis Function）单元组成，每类一个单元，每个有84个输入。换句话说，每个输出RBF单元计算输入向量和参数向量之间的欧式距离。输入离参数向量越远，RBF输出的越大。一个RBF输出可以被理解为衡量输入模式和与RBF相关联类的一个模型的匹配程度的惩罚项。用概率术语来说，RBF输出可以被理解为F6层配置空间的高斯分布的负log-likelihood。给定一个输入模式，损失函数应能使得F6的配置与RBF参数向量（即模式的期望分类）足够接近。这些单元的参数是人工选取并保持固定的（至少初始时候如此）。这些参数向量的成分被设为-1或1。虽然这些参数可以以-1和1等概率的方式任选，或者构成一个纠错码，但是被设计成一个相应字符类的7*12大小（即84）的格式化图片。这种表示对识别单独的数字不是很有用，但是对识别可打印ASCII集中的字符串很有用。

使用这种分布编码而非更常用的“1 of N”编码用于产生输出的另一个原因是，当类别比较大的时候，非分布编码的效果比较差。原因是大多数时间非分布编码的输出必须为0。这使得用sigmoid单元很难实现。另一个原因是分类器不仅用于识别字母，也用于拒绝非字母。使用分布编码的RBF更适合该目标。因为与sigmoid不同，他们在输入空间的较好限制的区域内兴奋，而非典型模式更容易落到外边。

RBF参数向量起着F6层目标向量的角色。需要指出这些向量的成分是+1或-1，这正好在F6 sigmoid的范围内，因此可以防止sigmoid函数饱和。实际上，+1和-1是sigmoid函数的最大弯曲的点处。这使得F6单元运行在最大非线性范围内。必须避免sigmoid函数的饱和，因为这将会导致损失函数较慢的收敛和病态问题。

5）训练过程

神经网络用于模式识别的主流是有指导学习网络，无指导学习网络更多的是用于聚类分析。对于有指导的模式识别，由于任一样本的类别是已知的，样本在空间的分布不再是依据其自然分布倾向来划分，而是要根据同类样本在空间的分布及不同类样本之间的分离程度找一种适当的空间划分方法，或者找到一个分类边界，使得不同类样本分别位于不同的区域内。这就需要一个长时间且复杂的学习过程，不断调整用以划分样本空间的分类边界的位置，使尽可能少的样本被划分到非同类区域中。

卷积网络在本质上是一种输入到输出的映射，它能够学习大量的输入与输出之间的映射关系，而不需要任何输入和输出之间的精确的数学表达式，只要用已知的模式对卷积网络加以训练，网络就具有输入输出对之间的映射能力。卷积网络执行的是有导师训练，所以其样本集是由形如：（输入向量，理想输出向量）的向量对构成的。所有这些向量对，都应该是来源于网络即将模拟的系统的实际“运行”结果。它们可以是从实际运行系统中采集来的。在开始训练前，所有的权都应该用一些不同的小随机数进行初始化。“小随机数”用来保证网络不会因权值过大而进入饱和状态，从而导致训练失败；“不同”用来保证网络可以正常地学习。实际上，如果用相同的数去初始化权矩阵，则网络无能力学习。

训练算法与传统的BP算法差不多。主要包括4步，这4步被分为两个阶段：

第一阶段，向前传播阶段：

a）从样本集中取一个样本(X,Y_p)，将X输入网络；

b）计算相应的实际输出O_p。

在此阶段，信息从输入层经过逐级的变换，传送到输出层。这个过程也是网络在完成训练后正常运行时执行的过程。在此过程中，网络执行的是计算（实际上就是输入与每层的权值矩阵相点乘，得到最后的输出结果）：

$$O_p=F_n(\ldots(F_2(F_1(X_pW^{(1)})W^{(2)})\ldots)W^{(n)})$$

第二阶段，向后传播阶段

a）算实际输出O_p与相应的理想输出Y_p的差；

b）按极小化误差的方法反向传播调整权矩阵。

6）卷积神经网络的优点

卷积神经网络CNN主要用来识别位移、缩放及其他形式扭曲不变性的二维图形。由于CNN的特征检测层通过训练数据进行学习，所以在使用CNN时，避免了显式的特征抽取，而隐式地从训练数据中进行学习；再者由于同一特征映射面上的神经元权值相同，所以网络可以并行学习，这也是卷积网络相对于神经元彼此相连网络的一大优势。卷积神经网络以其局部权值共享的特殊结构在语音识别和图像处理方面有着独特的优越性，其布局更接近于实际的生物神经网络，权值共享降低了网络的复杂性，特别是多维输入向量的图像可以直接输入网络这一特点避免了特征提取和分类过程中数据重建的复杂度。

流的分类方式几乎都是基于统计特征的，这就意味着在进行分辨前必须提取某些特征。然而，显式的特征提取并不容易，在一些应用问题中也并非总是可靠的。卷积神经网络，它避免了显式的特征取样，隐式地从训练数据中进行学习。这使得卷积神经网络明显有别于其他基于神经网络的分类器，通过结构重组和减少权值将特征提取功能融合进多层感知器。它可以直接处理灰度图片，能够直接用于处理基于图像的分类。

卷积网络较一般神经网络在图像处理方面有如下优点： a）输入图像和网络的拓扑结构能很好的吻合；b）特征提取和模式分类同时进行，并同时在训练中产生；c）权重共享可以减少网络的训练参数，使神经网络结构变得更简单，适应性更强。

7）小结

CNNs中这种层间联系和空域信息的紧密关系，使其适于图像处理和理解。而且，其在自动提取图像的显著特征方面还表现出了比较优的性能。在一些例子当中，Gabor滤波器已经被使用在一个初始化预处理的步骤中，以达到模拟人类视觉系统对视觉刺激的响应。在目前大部分的工作中，研究者将CNNs应用到了多种机器学习问题中，包括人脸识别，文档分析和语言检测等。为了达到寻找视频中帧与帧之间的相干性的目的，目前CNNs通过一个时间相干性去训练，但这个不是CNNs特有的。

呵呵，这部分讲得太啰嗦了，又没讲到点上。没办法了，先这样的，这样这个过程我还没有走过，所以自己水平有限啊，望各位明察。需要后面再改了，呵呵。