1. 深度学习

最近几年，深度学习(deep learning)在很多领域上都取得了很好的效果，如图像识别、语音识别、自然语言处理。深度学习从而成为近年来学术会议的焦点， 工业界的炫技利器， 甚至在创业公司中都掀起了一波深度学习热潮， 拍衣搜商品、 智能机器人、 Siri语音助手、无人车都借助这一技术提升性能。 可能你还是巴望在深度学习门槛前的一个，在读完本章后，让我们看看“深度学习”有多简单。

* 1. 横空处世？

什么是深度学习？

深度学习其实在很早的时候叫“神经网络”， 但是由于缺乏训练数据和 计算能力瓶颈， 一直很难训练好一个神经网络。 要么网络大了（网络中的参数相对数据多了），模型过拟合，要么网络小了，模型不能处理复杂问题。最近随着大数据和计算能力（GPU）的快速发展，才使神经网络改头换面为“深度学习”并变成很多任务的winner，所以说，深度学习的成功，只是由于有了更多数据和更多计算资源而已。

神经网络的雏形为perceptron,

最后，再下面介绍具体模型之前，我稍微介绍几个深度学习领域的大牛，肯定不能完全覆盖，但至少在下面内容中看到他们名字的时候能够有大概印象。

Geoffrey Hinton：深度学习教父，前多伦多大学教授，80年代后依然坚守神经网络的研究者之一，在将深度学习从边缘方向推进到各大互联网公司重要技术的路中做出很大贡献。于2013年入职google，当时64岁，一进去就以实习生的身份带Google Brain团队，目前google 安卓端语音搜索，图像搜索和Youtube推荐都是该团队技术支持下的作品。

 Michael I Jordan：UC Berkeley教授，前MIT教授。培养了一批优秀学子，如Andrew Ng, Yoshua Bengio。他在统计和机器学习领域都很有声望，推广了机器学习中的贝叶斯网络，EM算法等。

Yann Lecun： Hinton的博士，先后就职于多伦多大学（博士后），AT&T图像处理研究部门（从小弟到老大），纽约大学教授，目前到了Facebook做AI Research（人工智能研究院）的老大。他是第一个把CNN用到实际问题上并得到有效结果的人，后面在推动CNN普及应用上也有很大贡献。

Yoshua Bengio： 神经网络复兴的三大推动者之一（另外两个是Hinton和Lecun），Michael I Jordan 的学生，Lecun在AT&T 贝尔实验室时候的同事。组织编写基于符号计算的深度学习python库theano，深入研究并推动RNN的训练与应用。

Andrew Ng: 前斯坦福大学教授，和做了很久google工程师的Jeffrey Dean聊神经网络时碰撞出学术火花，在公司里建立了另一个团队，叫Google X（后来改名为X），做有长远收益（但不一定靠谱）的高端科研，如自动驾驶车（无人车），无人机，Google glass（已经差不多死了）等。估计也是Ng觉着在Google做着不爽，说能力还有Hinton跟这争；说项目收益，靠谱的都给Google Brain做了。一山不容二虎，他于2014年加盟百度，任首席科学家（T12）。

余凯：前西门子数据挖掘部门资深研究员，后跳到NEC实验室研究深度学习、图像识别、文本挖掘，多媒体检索等，做到部门主管，后加盟百度任深度学习实验室院长，目前离职创办公司“地平线”，做机器人、智能家电等设备。

颜水成：前新加坡国立大学(Dean’s Chair)副教授，现任360人工智能研究院老大。北大博士，UIUC博士后，虽然没有在神经网络空巢期坚挺这项技术，但他的团队能力扎实，跟进方向迅速，13年发表的Network in Network(NIN)在深度学习领域当时有很强影响力。

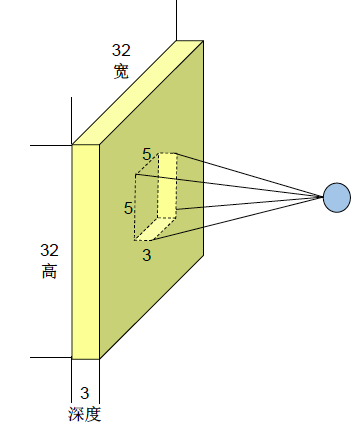
* 1. 基础CNN

在深度神经网络的众多类型中，卷积神经网络（Convolutional Neural Network， CNN）算是最广为研究的一种了，原因一是CNN结构简单，方便实现，二是CNN可以很好地处理图像相关问题。 在本节中我们首先介绍CNN的基本结构，然后深入看看近些年来都有哪些工作在CNN上进行改进。

CNN最早是Yann LeCun在98年提出用来做手写数字（0-9）识别的，在这个任务中，他们建立了一个叫LeNet-5的网络，用来从原始图像得到可以分辨图像类别的特征，此后随着CNN的发展又出现了很多CNN网络结构的变种，但其基本机构都是相似的。基本CNN网络结构包含三种类型的层：卷积层（convolutional layer）, 池化层（pooling layer）和全连接层（fully connected layer）， 以输入为图像为例。

* 卷积层

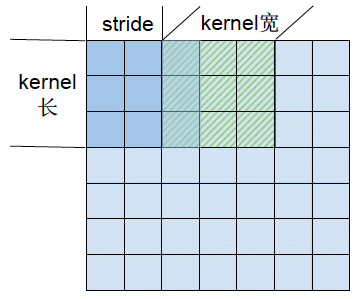
卷积神经网络是一个层级结构的前向神经网络。除了输入层和输出层，中间的隐层的目标是从原始数据学得一系列非线性组合，使得可以很好地结合当前任务表征信号。以图像为例，输入为像素级数据，每个像素本身不含有太多信息，但是组合起来所得到的更高级别的特征可以表示物体类别等信息。 以输入层和第一层隐层中的一个神经元节点连接为例，如图3.7所示。



图中黄色区域表示输入，为32\*32\*3 大小的数据，比如可表示一张深度为3（红、绿、蓝 3 通道），宽、高均为32 的图像。图中的蓝色圆点表示与输入层相连的一个隐层节点，连接项由权重参数组成。 和大脑中视觉皮层的接收域（receptive fields）一样，卷积神经网络的隐层中也设有接收域，该接收域可以视作一个在输入中进行扫描的线性滤波器，为了减少参数数量， 使得CNN网络更好训练，该滤波器中的权重参数在输入层中共享。卷积神经网络中，我们称该接收域为核（kernel）。例如图中为5\*5 的kernel，深度为3（和通道数相同）。由于CNN 的连接层中共享权值，所以该节点与输入层相连接的参数个数总共只有5\*5\*3 个。在输入层中，从原点（左上角）开始分别向右方和下方扫描，假如每隔1 个像素进行扫描，每个5\*5\*3 的区域块经5\*5\*3 的kernel 卷积得到一个标量，那么一个数据32\*32\*3 的数据经过卷积也就可以得到一个（32-5+1）\*（32-5+1）大小的特征，该特征称为feature map。

对于隐层中有多个节点的情况，我们在隐层的深度维进行扩展。假如我们希望隐层有N 个节点，就令隐层的深度为N，那么隐层即有N个feature map，每个的大小为（32-5+1）\*（32-5+1）。

刚才我们对于输入层每隔一个像素进行一次卷积，有时为了减小计算复杂度，我们还可以设置卷积的步长，也就是每隔多少个元素进行一次卷积操作，这个步长称为stride，如图3.8所示为输入大小为7\*7 的图像，stride 为2 的情况，可见此时对应的输出长和宽为。即给定输入数据大小，kernel 大小, stride 大小，要求卷积神经网络的隐层节点为，隐层大小为, 有 个参数。



在线性卷积之后，卷积层中会通过引入一个非线性激活函数（activation function）获得非线性特征（同perceptron）。

* 降采样层/池化层

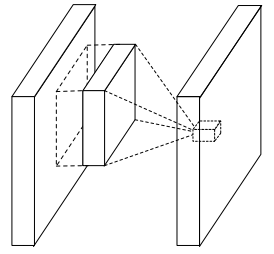
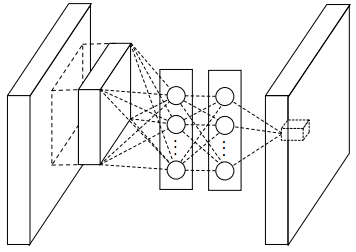
1998 年，Yan Lecun 等人提出，在网络构建时可以在CNN 网络的卷积层后面跟一个降采样层，用来对卷积结果中的每个小窗口做均值降采样（average pooling）或者最大值降采样（max pooling）。有两个原因，一来这样可以减少下一层的数据数据规模，降低计算量；二来这样在一个小区域的节点中做平均或最大化选择可以增加模型的适应性，减少网络受旋转、平移等变换的影响。将卷积层和降采样层结合实际是受Hubel 和Wiesel 在生物实验上结果的影响[84]。和卷积层类似，降采样层也有相应的窗口高、宽和stride。经过降采样层，使得一块区域的数据被压缩，分在不同的“池子”里，因此降采样层所做的事情也叫做池化。除了均值和最大值，还可以取L1, L2 范数进行降采样。

* 全连接层
  1. CNN进阶

如果你只是想粗浅的了解一下CNN的基本部件，那么上一节基础CNN就足够了。 本节中，我们将主要介绍从2012年CNN的成功范例——AlexNet （ImageNet 竞赛2012冠军）之后CNN的主要有效改进。

* 卷积层
  + Network in Network（NIN）

2013年，颜水成等人提出了Network in Network，旨在提高模型在receptive field中的判别能力。所谓NIN就是用一个微型神经网络代替一个线性卷积，也就是在卷积后跟一个mlp (multilayer perceptron)层，整个神经网络由很多这样的微型神经网络堆叠而成，如图所示。

(a).线性卷积层 (b). 微型神经网络卷积层

原本CNN卷积层中的卷积核只是广义线性模型（GLM），当特征线性可分的时候可以通过线性卷积核获得特征。可以说NIN相对线性卷积和是广义线性卷积网络结构，这样的好处是可以抽象出更强的特征。

* + Inception Layer

Inception Layer 在GoogleNet中首次提出，可以看做是NIN的一个

* 降采样层
* 损失函数
* 激活函数
* 正则化
* 优化方法
  1. Deeper than deeper?

上一节中我们已提到，从12年到15年，图像分类竞赛ImageNet冠军依次使用了越来越深的网络，那么我们要问，是不是只要不过拟合，网络越深越好使？拍拍脑袋想好像是这样，网络越深，参数拟合能力越强， 但在Highway Network一文中作者提出了一个相反的现象，［1］中称此现象为degradation, 如图所示，

* 1. Recurrent Neural Network (RNN)

参考文献：

［1］.