卷积神经网络知识总结

学号:20192131089　姓名: 吴宇涛

2021 年 12 月 3 日

目录

摘要

卷积神经网络（Convolutional Neural Networks, 简称 CNN）是一类特殊的人工神经网络, 区别于神经网络其他模型（如, 递归神经网络、Boltzmann 机等）, 其最主要的特点是卷积运算操作（convolution operators）. 因此,CNN 在诸多领域应用特别是图像相关任务上表现优异, 诸如, 图像分类（image classification）、图像语义分割（image semantic segmentation）、图像检索（image retrieval）、物体检测（object detection）等计算机视觉问题. 此外, 随着 CNN 研究的深入, 如自然语言处理（natural language processing）中的文本分类, 软件工程数据挖掘（software mining）中的软件缺陷预测等问题都在尝试利用

卷积神经网络解决, 并取得了相比传统方法甚至其他深度网络模型更优的预测效果. 本文将总结卷积神经网络的发展, 思想与技术的演化,

关键词: 卷积神经网络; CNN; GPU; pytorch

1 卷积神经网络

1.1 卷积神经网络的概念

上世纪 60 年代,Hubel 等人通过对猫视觉皮层细胞的研究, 提出了感受野这个概念, 到 80 年代,Fukushima 在感受野概念的基础之上提出了神经认知机的概念, 可以看作是卷积神经网络的第一个实现网络, 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.

卷积神经网络是多层感知机（MLP）的变种, 由生物学家休博尔和维瑟尔在早期关于猫视觉皮层的研究发展而来, 视觉皮层的细胞存在一个复杂的构造, 这些细胞对视觉输入空间的子区域非常敏感, 称之为感受野.

CNN 由纽约大学的 Yann Lecun 于 1998 年提出, 其本质是一个多层感知机, 成功的原因在于其所采用的局部连接和权值共享的方式:

- 一方面减少了权值的数量使得网络易于优化
- 另一方面降低了模型的复杂度, 也就是减小了过拟合的风险

该优点在网络的输入是图像时表现的更为明显, 使得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的过程, 在二维图像的处理过程中有很大的优势, 如网络能够自行抽取图像的特征包括颜色、纹理、形状及图像的拓扑结构, 在处理二维图像的问题上, 特别是识别位移、缩放及其他形式扭曲不变性的应用上具有良好的鲁棒性和运算效率等.

1.2 发展过程

1986 年 Rumelhart 等人提出了人工神经网络的反向传播算法, 掀起了神经网络在机器学习中的热潮, 神经网络中存在大量的参数, 存在容易发生过拟合、训练时间长的缺点, 但是对比 Boosting、Logistic 回归、SVM 等基于统计学习理论的方法（也可以看做具有一层隐层节点或不含隐层节点的学习模型, 被称为浅层模型）来说, 具有较大的优越性.

1.2.1 浅层模型的缺点

浅层学习模型通常要由人工的方法来获得好的样本特性, 在此基础上进行识别和预测, 因此方法的有效性在很大程度上受到特征提取的制约.

1.2.2 深度学习的提出

2006 年,Hinton 提出了深度学习, 两个主要的观点是:

- 多隐层的人工神经网络具有优异的特征学习能力, 学习到的数据更能反映数据的本质特征有利于可视化或分类
- 深度神经网络在训练上的难度, 可以通过逐层无监督训练有效克服.

1.2.3 深度学习取得成功的原因

- 大规模数据（例如 ImageNet）: 为深度学习提供了好的训练资源
- 计算机硬件的飞速发展: 特别是 GPU 的出现, 使得训练大规模上网络成为可能

1.3 什么是卷积神经网络

卷积神经网络是一种带有卷积结构的深度神经网络, 卷积结构可以减少深层网络占用的内存量, 其三个关键的操作, 其一是局部感受野, 其二是权值共享, 其三是 pooling 层, 有效的减少了网络的参数个数, 缓解了模型的过拟合问题.

1.3.1 网络结构

卷积神经网络整体架构：卷积神经网络是一种多层的监督学习神经网络, 隐含层的卷积层和池采样层是实现卷积神经网络特征提取功能的核心模块. 该网络模型通过采用梯度下降法最小化损失函数对网络中的权重参数逐层反向调节, 通过频繁的迭代训练提高网络的精度. 卷积神经网络的低隐层是由卷积层和最大池采样层交替组成, 高层是全连接层对应传统多层感知器的隐含层和逻辑回归分类器. 第一个全连接层的输入是由卷积层和子采样层进行特征提取得到的特征图像. 最后一层输出层是一个分类器, 可以采用逻辑回归,Softmax 回归甚至是支持向量机对输入图像进行分类.

卷积神经网络结构包括：卷积层, 降采样层, 全链接层. 每一层有多个特征图, 每个特征图通过一种卷积滤波器提取输入的一种特征, 每个特征图有多个神经元.

输入图像统计和滤波器进行卷积之后, 提取该局部特征, 一旦该局部特征被提取出来之后, 它与其他特征的位置关系也随之确定下来了, 每个神经元的输入和前一层的局部感受野相连, 每个特征提取层都紧跟一个用来求局部平均与二次提取的计算层, 也叫特征映射层, 网络的每个计算层由多个特征映射平面组成, 平面上所有的神经元的权重相等.

通常将输入层到隐藏层的映射称为一个特征映射, 也就是通过卷积层得到特征提取层, 经过 pooling 之后得到特征映射层.

1.3.2 局部感受野与权值共享

卷积神经网络的核心思想就是局部感受野、是权值共享和 pooling 层, 以此来达到简化网络参数并使得网络具有一定程度的位移、尺度、缩放、非线性形变稳定性.

- 局部感受野：由于图像的空间联系是局部的, 每个神经元不需要对全部的图像做感受, 只需要感受局部特征即可, 然后在更高层将这些感受得到的不同的局部神经元综合起来就可以得到全局的信息了, 这样可以减少连接的数目.
- 权值共享：不同神经元之间的参数共享可以减少需要求解的参数, 使用多种滤波器去卷积图像就会得到多种特征映射. 权值共享其实就是对图像用同样的卷积核进行卷积操作, 也就意味着第一个隐藏层的所有神经元所能检测到处于图像不同位置的完全相同的特征. 其主要的能力就能检测到不同位置的同一类型特征, 也就是卷积网络能很好的适应图像的小范围的平移性, 即有较好的平移不变性（比如将输入图像的猫的位置移动之后, 同样能够检测到猫的图像）

1.4 卷积层、下采样层、全连接层

1.4.1 卷积层

因为通过卷积运算我们可以提取出图像的特征, 通过卷积运算可以使得原始信号的某些特征增强, 并且降低噪声.

用一个可训练的滤波器 fx 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 bx, 得到卷积层 Cx.

1.4.2 下采样层

因为对图像进行下采样, 可以减少数据处理量同时保留有用信息, 采样可以混淆特征的具体位置, 因为某个特征找出来之后, 它的位置已经不重要了, 我们只需要这个特征和其他特征的相对位置, 可以应对形变和扭曲带来的同类物体的变化.

每邻域四个像素求和变为一个像素, 然后通过标量 Wx+1 加权, 再增加偏置 bx+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 Sx+1.

1.4.3 全连接层

采用 softmax 全连接, 得到的激活值即卷积神经网络提取到的图片特征.

1.5 卷积神经网络在图像处理中的优点

- 网络结构能够较好的适应图像的结构
- 同时进行特征提取和分类, 使得特征提取有助于特征分类
- 权值共享可以减少网络的训练参数, 使得神经网络结构变得简单, 适应性更强

2 卷积和池化的意义

2.1 CNN 中卷积的意义

假如有一幅 1000*1000 的图像, 如果把整幅图像作为向量, 则向量的长度为 1000000（10^6）. 在假如隐含层神经元的个数和输入一样, 也是 1000000; 那么, 输入层到隐含层的参数数据量有 10^{12}, 所以, 我们还得降低维数, 同时得以整幅图像为输入.

2.1.1 局部感知

卷积神经网络有两种神器可以降低参数数目, 第一种神器叫做局部感知野. 一般认为人对外界的认知是从局部到全局的, 而图像的空间联系也是局部的像素联系较为紧密, 而距离较远的像素相关性则较弱. 因而, 每个神经元其实没有必要对全局图像进行感知, 只需要对局部进行感知, 然后在更高层将局部的信息综合起来就得到了全局的信息. 网络部分连通的思想, 也是受启发于生物学里面的视觉系统结构. 视觉皮层的神经元就是局部接受信息的（即这些神经元只响应某些特定区域的刺激）. 如下图所示: 左图为全连接, 右图为局部连接.

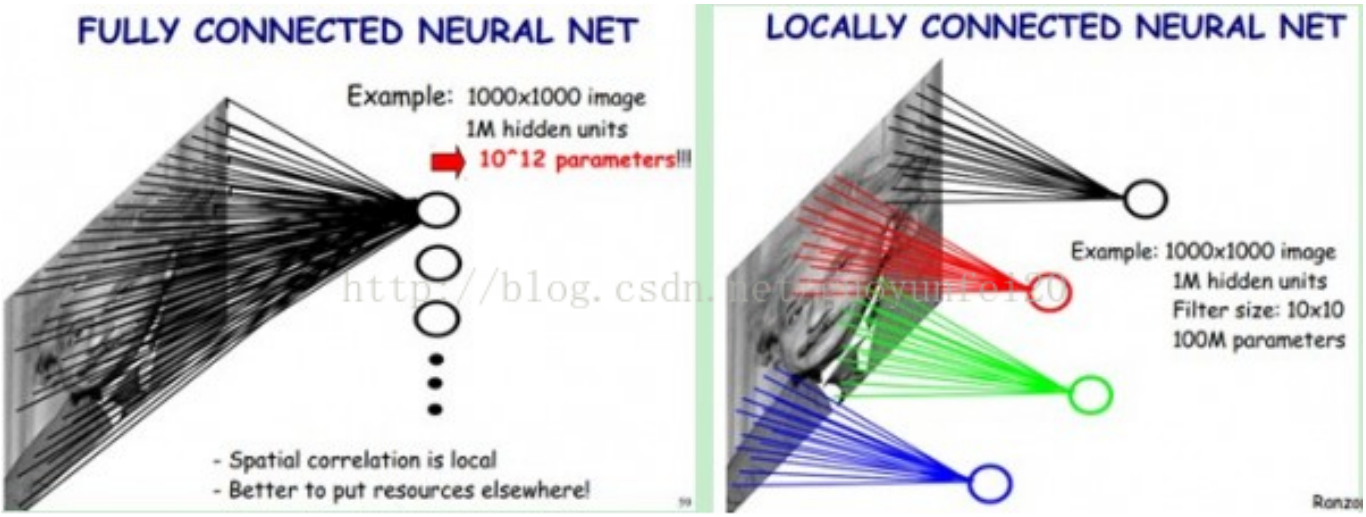

图 1: 全连接和局部连接

在上右图中, 假如每个神经元只和 10×10 个像素值相连, 那么权值数据为 1000000×100 个参数, 减少为原来的千分之一. 而那 10×10 个像素值对应的 10×10 个参数, 其实就相当于卷积操作.

2.1.2 权值共享

在上面的局部连接中, 每个神经元都对应 100 个参数, 一共 1000000 个神经元, 如果这 1000000 个神经元的 100 个参数都是相等的, 那么参数数目就变为 100 了.

我们可以这 100 个参数（也就是卷积操作）看成是提取特征的方式, 该方式与位置无关. 这其中隐含的原理则是: 图像的一部分的统计特性与其他部分是一样的. 这也意味着我们在这一部分学习的特征也能用在另一部分上, 所以对于这个图像上的所有位置, 我们都能使用同样的学习特征. 更直观一些, 当从一个大尺寸图像中随机选取一小块, 比如说 8×8 作为样本, 并且从这个小块样本中学习到了一些特征, 这时我们可以把从这个 8×8 样本中学习到的特征作为探测器, 应用到这个图像的任意地方中去. 特别是, 我们可以用从 8×8 样本中所学习到的特征跟原本的大尺寸图像作卷积, 从而对这个大尺寸图像上的任一位置获得一个不同特征的激活值.

如下图所示, 展示了一个 3*3 的卷积核在 5*5 的图像上做卷积的过程. 每个卷积都是一种特征提取方式, 就像一个筛子, 将图像中符合条件（激活值越大越符合条件）的部分筛选出来.

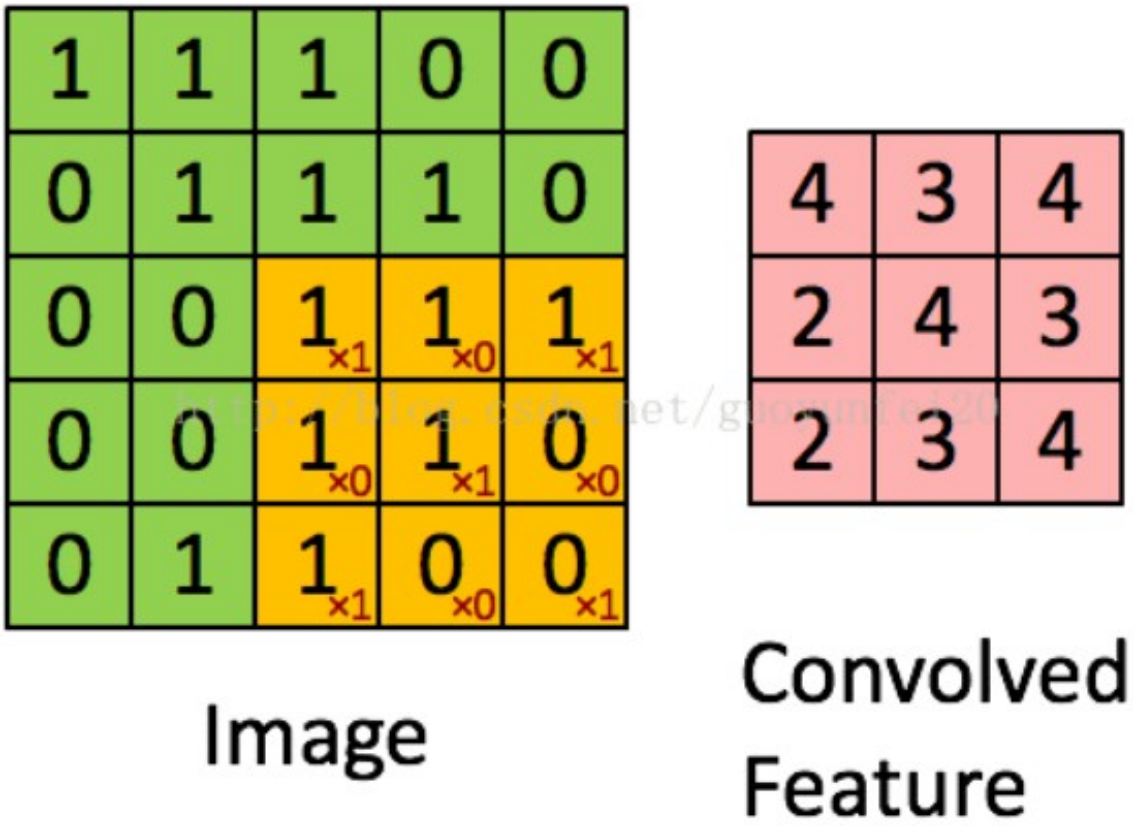

图 2: 卷积操作

2.1.3 多核卷积

上面所述只有 100 个参数时, 表明只有 1 个 100*100 的卷积核, 显然, 特征提取是不充分的, 我们可以添加多个卷积核, 比如 32 个卷积核, 可以学习 32 种特征. 在有多个卷积核时, 如下图所示

上图右, 不同颜色表明不同的卷积核. 每个卷积核都会将图像生成为另一幅图像. 比如两个卷积核就可以将生成两幅图像, 这两幅图像可以看做是一张图像的不同的通道. 如下图所示:

2.2 池化的意义

池化的定义比较简单, 最直观的作用便是降维, 常见的池化有最大池化、平均池化和随机池化. 池化层不需要训练参数.

2.2.1 三种池化示意图

最大池化是对局部的值取最大; 平均池化是对局部的值取平均; 随机池化是根据概率对局部的值进行采样, 采样结果便是池化结果. 概念非常容易理解, 其示意图如下所示:

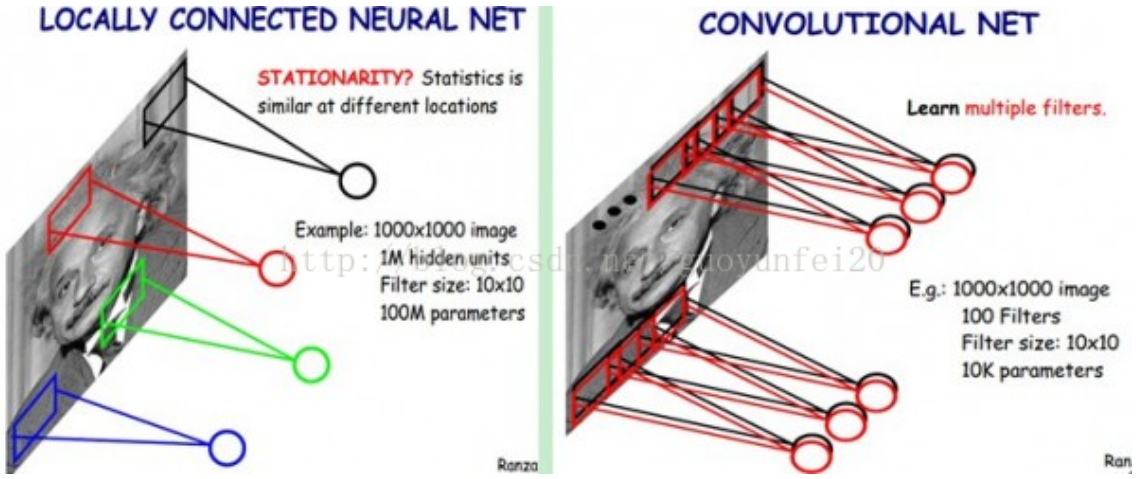

图 3: 多核卷积

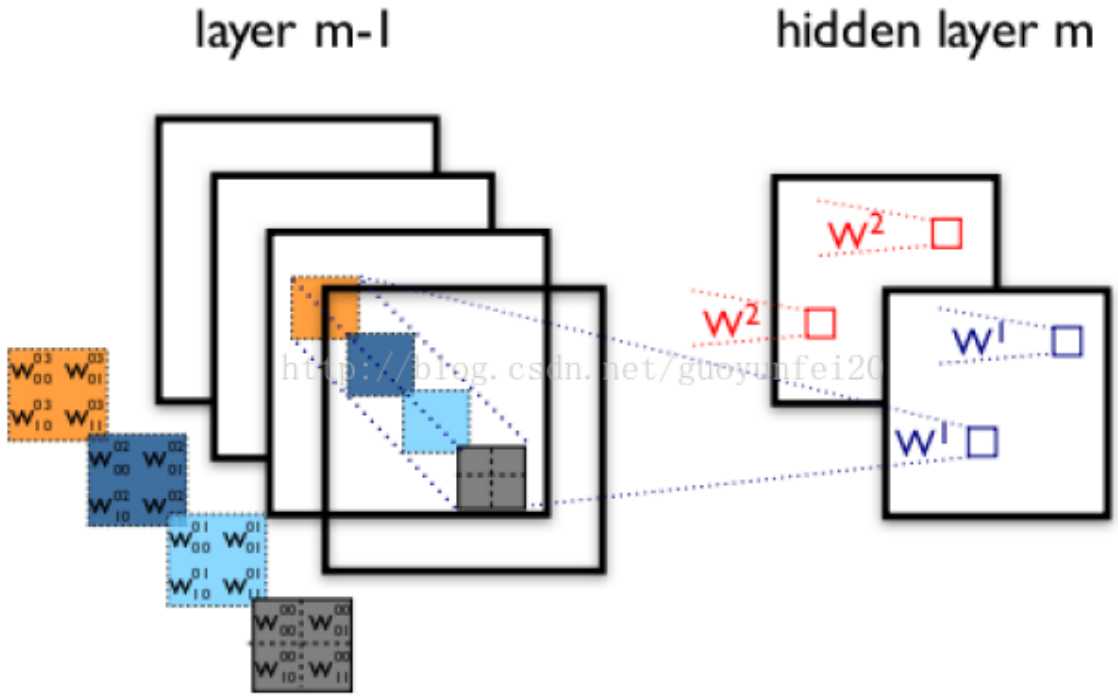

图 4: 图像的不同通道

2.2.2 三种池化的意义

- 最大池化可以获取局部信息, 可以更好保留纹理上的特征. 如果不用观察物体在图片中的具体位置, 只关心其是否出现, 则使用最大池化效果比较好.
- 平均池化往往能保留整体数据的特征, 能凸出背景的信息.
- 随机池化中元素值大的被选中的概率也大, 但不是像最大池化总是取最大值. 随机池化一方面最大化地保证了 Max 值的取值, 一方面又确保了不会完全是 max 值起作用, 造成过度失真. 除此之外, 其可以在一定程度上避免过拟合.

2.2.3 池化在 CNN 中的作用

- 不变性 (invariance). 包括平移、旋转、尺度、(translation、rotation、scale)
- 减少维度 (少参数就是少计算量) 并可以保留主要特征 (降维, 效果类似 PCA).
- 最常用的目的是–防止过拟合. 提高模型的泛化能力.

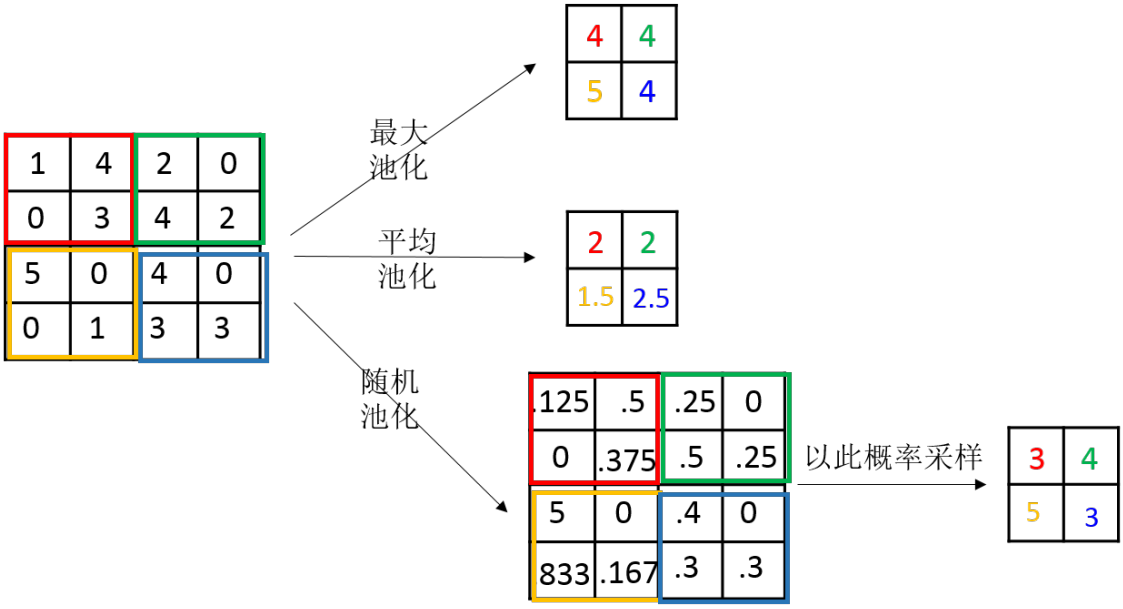

图 5: 示意图

3 动量优化

3.1 动量法

随机梯度下降的方法很难通过峡谷区域（也就是在一个维度梯度变化很大, 另一个维度变化较小）, 这个很好理解, 因为梯度下降是梯度更新最大的反方向, 如果这个时候一个维度梯度变化很大, 那么就很容易在这个方向上振荡, 另一个方向就更新很慢, 如下图:

图 6:

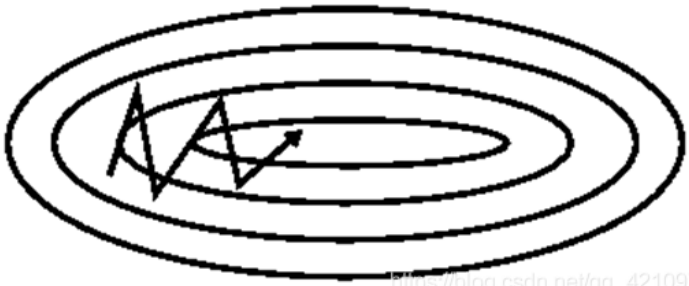

图 7:

图 6 没有加动量, 图 7 加了动量的方法, 可以看到有动量可以在变化小的维度上加快更新, 使得加快收敛. 该方法是通过添加一个参数 β 构建一个一阶动量 m, 其中 m 有下列表达式:

$$m_t = \beta m_{t-1} + (1-\beta)\frac{\partial Loss}{\partial \theta_t}$$

而对于其二阶动量 V=1, 所以其参数更新公式为:

$$\theta_{t+1} = \theta_t - \eta m_t$$

其中 β 一般取 0.9, 我们来看看其它表达式, 相信在搜索动量梯度下降时, 有时候在其它地方也会看到下面这种表达式:

$$v_t = \gamma v_{t-1} + \eta \nabla \theta_t J(\theta_t)$$

$$\theta_{t+1} = \theta_t - v_t$$

这里的 γ 一般也是等于 0.9, 看起来这两种表达式有很大不一样, 其实是差不多的, 只不过第一种我觉得看起来更容易理解, 第二种我觉得就不是那么明显的去理解, 下面我将根据这两种表达式对比并分析动量梯度下降原理, 这样更容易理解, 将表达式继续拆开可以得到:

$$\eta m_t = \eta[\beta^3 m_{t-3} + \beta^2(1-\beta)\frac{\partial Loss}{\partial \theta_{t-2}} + \beta(1-\beta)\frac{\partial Loss}{\partial \theta_{t-1}} + (1-\beta)\frac{\partial Loss}{\partial_t}]$$

$$v_t = \gamma^3 v_{t-3} + \gamma^2 \eta \nabla \theta_{t-2} J(\theta_{t-2}) + \gamma\eta\nabla\theta_{t-1}J(\theta_{t-1}) + \eta\nabla\theta_t J(\theta_t)$$

解释下动量梯度更新的现实意义理解, 首先来看看 “An overview of gradient descent optimization algorithms” 这篇论文中的比喻: “从本质上说, 动量法, 就像我们从山上推下一个球, 球在滚下来的过程中累积动量, 变得越来越快 (直到达到终极速度, 如果有空气阻力的存在, 则 γ<1). 同样的事情也发生在参数的更新过程中: 对于在梯度点处具有相同的方向的维度, 其动量项增大, 对于在梯度点处改变方向的维度, 其动量项减小. 因此, 我们可以得到更快的收敛速度, 同时可以减少摇摆.” , 这样解释视乎就能够解释之前表达式 2 的含义了, 把其当做动量的累加, 比如小球在下坡山坡上, 那么根据表达式 2, 梯度方向是一直向下的, 自然参数更新幅度也就是一直累加的, 也就变得越来越大; 而当遇到山沟, 越过山沟此时就在另一边山坡, 这个时候梯度方向是跟之前相反的, 此时由于之前梯度大小的累加, 在两个山坡间的变化就会被互相抵消掉, 也就不会一直在两个山坡振荡, 容易朝山沟向下走, 也就是减少摇摆了.

3.2 程序运行结果对比

在本次选做的作业中, 我对 CIFAR10data 这个数据集使用了三个不同的动量参数, 分别使用了 0.5, 0.85, 0.9, 对比发现 0.9 的效果最好, 其次是 0.85,0.5 的效果比较差.

同理在手写数字识别的部分, 测试结果也和上面所述一致, 动量参数作为超参数, 也是一个值得研究的参数.

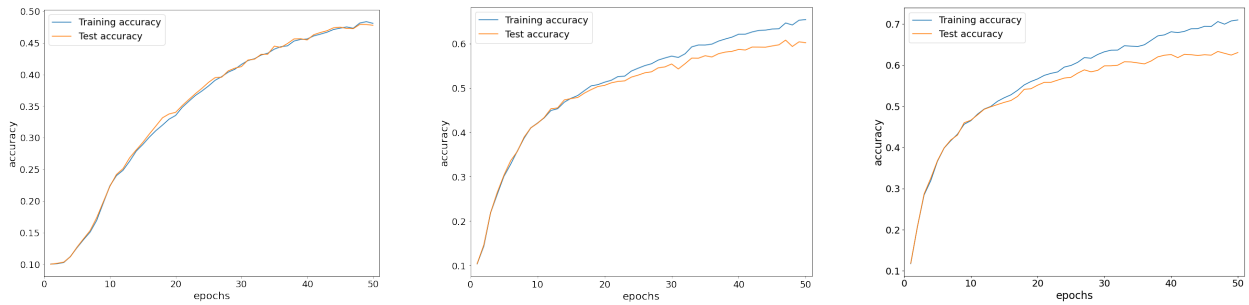

图 8: 动量参数分别为 0.5,0.85.0.9

4 总结

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.

CNN 一个非常重要的特点就是头重脚轻（越往输入权值越小, 越往输出权值越多）, 呈现出一个倒三角的形态, 这就很好地避免了 BP 神经网络中反向传播的时候梯度损失得太快.

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.

参考文献

[1] Dive Into Deep LearningV1.1

[2]《从机器学习到深度学习基于 scikit-learn 与 TensorFlow 的高效开发实战》