本周工作：深度网络概览（Deep Network）

在之前学习的系数自编码器和自我学习都是浅层的神经网络（仅包含输入层、隐藏层和输出层），“浅”指特征只使用一层计算单元来得到（即隐藏层）。

但通过引入深度神经网络，即包含多个隐藏层的神经网络，那么就可以计算更为复杂的输入特征。由于每一个隐藏层都是对上一层的输出进行非线性变换，因此深度神经网络拥有比浅度神经网络更好的表达能力。注意：每一层隐藏层应该用非线性的激活函数f(x)。

深度网络的优势

主要优势是以更紧凑简洁的方式来表达比浅层网络大得多的函数集合。例如，我们可以找到一些函数，它可以用k层网络简洁地表达出来（简洁指隐藏单元的数目只需与输入单元数目呈多项式关系），但是对于一个只有k-1层的神经网络而言，除非使用输入单元数目呈指数关系的隐藏层数目来，否则不能简洁地表示这些函数。

举个例子：比如打算构建一个布尔网络来计算 个输入比特的奇偶校验码（或者进行异或运算）。假设网络中的每一个节点都可以进行逻辑“或”运算（或者“与非”运算），亦或者逻辑“与”运算。如果我们拥有一个仅仅由一个输入层、一个隐层以及一个输出层构成的网络，那么该奇偶校验函数所需要的节点数目与输入层的规模呈指数关系。但是，如果我们构建一个更深点的网络，那么这个网络的规模就可做到仅仅是的多项式函数。这也表明了深度网络在表示一些函数时具有独到的优势。

当处理的是图像时，我们就能用深度网络学习“部分——整体”的分解关系。例如，第一层可以学习如何将图像中的像素组合在一起来检测边缘（正如我们在前面的练习中做的那样）。第二层可以将边缘组合起来检测更长的轮廓或者简单的“目标的部件”。在更深的层次上，可以将这些轮廓进一步组合起来以检测更为复杂的特征。就像人的大脑啊皮层处理图像的过程一样。

训练深度网络的困难之处

训练深度网络在几十年来的进步比较少，问题原因在于研究者们主要使用的学习算法是：首先随机初始化深度网络的权重，然后使用有监督的目标函数在有标签的训练集上进行训练。例如通过使用梯度下降法来降低训练误差。

主要原因：

（1）数据获取问题

我们需要依赖于有标签的数据才能进行训练。然而有标签的数据通常是稀缺的，因此对于许多问题，我们很难获得足够多的样本来拟合一个复杂模型的参数。例如，考虑到深度网络具有强大的表达能力，在不充足的数据上进行训练将会导致过拟合。

1. 局部极值问题

利用监督学习算法对浅层网络进行训练通常能是参数收敛到合理的范围。但用来训练深层网络时，并不能取得好的效果。使用监督算法训练神经网络时，通常会涉及求解一个高度非凸的优化问题。对深度网络而言，这种非凸优化问题的搜索区域中充斥着大量“坏”的局部极值，因而使用梯度下降法（或者像共轭梯度下降法，L-BFGS等方法）效果并不好。

1. 梯度弥散问题

梯度下降算法（以及相关的L-BFGS算法）在使用随机初始化权重的深度网络上效果不好的原因是：梯度会变得非常小。在用反向传播算法计算导数时，随着网络深度的增加，反向传播的梯度的幅值会急剧减少。结果就造成了整体损失函数相对于最初几层的权重的导数非常小。这样，当使用梯度下降法的时候，最初几层的权重变化非常缓慢，以至于它们不能在样本中进行有效的学习。

与梯度弥散问题紧密相关的问题是：当神经网络中的最后几层含有足够数量神经元的时候，可能单独这几层就足以对有标签数据进行建模，而不用最初几层的帮助。因此，对所有层都使用随机初始化的方法训练得到的整个网络的性能将会与训练得到的浅层网络（仅由深度网络的最后几层组成的浅层网络）的性能相似。

逐层贪婪训练方法

简单来说，逐层贪婪算法的主要思路是每次只训练网络中的一层，即我们首先训练一个只含一个隐藏层的网络，仅当这层网络训练结束之后才开始训练一个有两个隐藏层的网络，以此类推。在每一步中，我们把已经训练好的前k-1层固定，然后增加第k层（也就是将我们已经训练好的前k-1的输出作为输入）。每一层的训练可以是有监督的（例如，将每一步的分类误差作为目标函数），但更通常使用无监督方法（例如自动编码器，我们会在后边的章节中给出细节）。这些各层单独训练所得到的权重被用来初始化最终（或者说全部）的深度网络的权重，然后对整个网络进行“微调”（即把所有层放在一起来优化有标签训练集上的训练误差）

成功的原因：

1. 数据获取

自学习方法（self-taught learning）的潜力在于它能通过使用大量的无标签数据来学习到更好的模型。具体而言，该方法使用无标签数据来学习得到所有层（不包括用于预测标签的最终分类层）Wl的最佳初始权重。相比纯监督学习方法，这种自学习方法能够利用多得多的数据，并且能够学习和发现数据中存在的模式。因此该方法通常能够提高分类器的性能。

1. 更好的局部极值

当用无标签数据训练完网络后，相比于随机初始化而言，各层初始权重会位于参数空间中较好的位置上。然后我们可以从这些位置出发进一步微调权重。从经验上来说，以这些位置为起点开始梯度下降更有可能收敛到比较好的局部极值点，这是因为无标签数据已经提供了大量输入数据中包含的模式的先验信息。

参考资料

1.UFDL深度学习教程