

深度学习

前言

这部分依旧是论属性东西，依旧是深度学习范畴，依旧是看两眼有个印象，尽可能的看懂就好（简述的内容很浅，并且随便单拎出来一个点都是一篇文章）。

来自西瓜书：

深度学习

理论上来说，参数越多的模型复杂度越高、“容量”(capacity)越大，这意味着它能完成更复杂的学习任务。但一般情形下，复杂模型的训练效率低，易陷入过拟合，因此难以受到人们青睐(大型深度学习模型中甚至有上百亿个参数)。而随着云计算、大数据时代的到来，计算能力的大幅提高可缓解训练低效性，训练数据的大幅增加则可降低过拟合风险，因此，以“深度学习”(deep learning)为代表的复杂模型开始受到人们的关注。

典型的深度学习模型就是很深层的神经网络。显然，对神经网络模型，提高容量的一个简单办法是增加隐层的数目。隐层多了，相应的神经元连接权、阈值等参数就会更多。模型复杂度也可通过单纯增加隐层神经元的数目来实现，前面我们谈到过，单隐层的多层前馈网络已具有很强大的学习能力；但从增加模型复杂度的角度来看，增加隐层的数目显然比增加隐层神经元的数目更有效，因为增加隐层数不仅增加了拥有激活函数的神经元数目 θ ，还增加了激活函数嵌套的层数边。然而，多隐层神经网络(这里所说的“多隐层”是指三个以上隐层；深度学习模型通常有八九层甚至更多隐层)难以直接用经典算法(例如标准BP算法)进行训练，因为误差在多隐层内逆传播时，往往“发散”(diverge)而不能收敛到稳定状态。

无监督逐层训练(unsupervised layer-wise training)是多隐层网络训练的有效手段，其基本思想是每次训练一层隐结点，训练时将上一层隐结点的输出作为输入，而本层隐结点的输出作为下一层隐结点的输入，这称为“预训练”(pre-training)；在预训练全部完成后，再对整个网络进行“微调”(finetuning)训练。例如，在深度信念网络(deep belief network, 简称 DBN) [Hinton et al., 2006] 中，每层都是一个受限 Boltzmann 机，即整个网络可视为若干个 RBM 堆叠而得。在使用无监督逐层训练时，首先训练第一层，这是关于训练样本的 RBM 模型，可按标准的 RBM 训练；然后，将第一层预训练好的隐结点视为第二层的输入结点，对第二层进行预训练；……各层预训练完成后，再利用 BP 算法等对整个网络进行训练。

事实上，“预训练+微调”的做法可视为将大量参数分组，对每组先找到局部看来比较好的设置，然后再基于这些局部较优的结果联合起来进行全局寻优。这样就在利用了模型大量参数所提供的自由度的同时，有效地节省了训练开销。

另一种节省训练开销的策略是“权共享”(weight sharing)，即让一组神经元使用相同的连接权。这个策略在卷积神经网络(Convolutional Neural Network, 简称 CNN) [LeCun and Bengio, 1995; LeCun et al., 1998] 中发挥了重要作用。以 CNN 进行手写数字识别任务为例 [LeCun et al., 1998]，如

图所示，网络输入是一个 32×32 的手写数字图像，输出是其识别结果，CNN 复合多个“卷积层”和“采样层”对输入信号进行加工，然后在连接层实现与输出目标之间的映射(近来人们在使用 CNN 时常将 Sigmoid 激活函数替换为修正线性函数。这样的神经元称为ReLU(Rectified Linear Unit)：此外，汇合层的操作常采用“最大”或“平均”这更接近于集成学习中的一些操作)。每个卷积层都包含多个特征映射(feature map)，每个特征映射是一个由多个神经元构成的“平面”，通过一种卷积滤波器提取输入的一种特征。例如，图中第一个卷积层由 6 个特征映射构成，每个特征映射是一个 28×28 的神经元阵列，其中每个神经元负责从 5×5 的区域通过卷积滤波器提取局部特征。采样层亦称为“汇合”(pooling)层，其作用是基于局部相关性原理进行亚采样，从而在减少数据量的同时保留有用信息。例如图中第一个采样层有 6 个 14×14 的特征映射，其中每个神经元与上一层中对应特征映射的 2×2 邻域相连，并据此计算输出。通过复合卷积层和采样层，图中的 CNN 将原始图像映射成 120 维特征向量，最后通过一个由 84 个神经元构成的连接层和输出层连接完成识别任务。CNN 可用 BP 算法进行训练，但在训练中无论是卷积层还是采样层，其每一组神经元(即图中的每个“平面”)都是用相同的连接权，从而大幅减少了需要训练的参数数目。

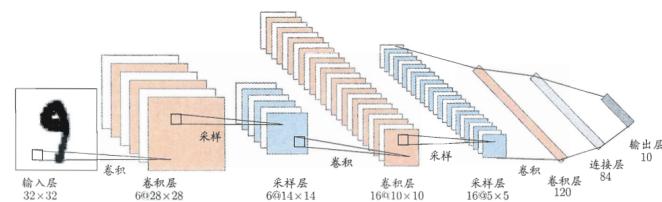


图 5.15 卷积神经网络用于手写数字识别 [LeCun et al., 1998]

我们可以从另一个角度来理解深度学习。无论是 DBN 还是 CNN，其多隐层堆叠、每层对上一层的输出进行处理的机制，可看作是在对输入信号进行逐层加工，从而把初始的、与输出目标之间联系不太密切的输入表示，转化成与输出目标联系更密切的表示，使得原来仅基于最后一层输出映射难以完成的任务成为可能。换言之，通过多层处理，逐渐将初始的“低层”特征表示转化为“高层”特征表示后，用“简单模型”即可完成复杂的分类等学习任务。由此可将深度学习理解为进行“特征学习”(feature learning)或“表示学习”(representation learning)。

注：若将网络中前若干层处理都看作是在进行特征表示，只把最后一层处理看作是在进行“分类”，则分类使用的就是一个简单模型。

以往在机器学习用于现实任务时，描述样本的特征通常需由人类专家来设计，这称为“特征工程”(feature engineering)。众所周知，特征的好坏对泛化性能有至关重要的影响，人类专家设计出好特征也并非易事；特征学习则通过机器学习技术自身来产生好特征，这使机器学习向“全自动数据分析”又前进了一步。