深度机器学习——一个新的人工智能研究的前沿

1. 介绍

模仿出人脑表征信息的高效和稳健一直是近几十年来人工智能研究中的一个核心的挑战。而人类不仅每时每刻都暴露在由感官接收的无数的数据中，而且还能够通过某些方法捕捉到这些数据关键的部分来以简单的方式在未来使用。而早在五十年前，提出了动态规划理论以及开创了最优化控制领域的 Richard Bellman 就曾断言，数据的高维度是在人工智能科学和工程应用中的根本障碍。而其中主要的困难，尤其是在模式分类的应用中，随着数据维度的线性增长，学习的难度会发生指数级的增长[1][[1]](#endnote-0)。而克服这个“诅咒”的主流方法就是以一定的方式（比如分类引擎）对数据进行预处理来减少数据的维度，这样数据就可以被有效地处理。这种减少维度的做法一般被称为特征提取。因此，可以认为，在很多模式识别系统的背后的智能其实被转移到了人工的特征提取处理上去了，而这种人工的做法有时会很有挑战性而且会高度依赖于具体的应用。此外，如果不完整或者错误的特征被提取出来了，那么分类处理的表现就会从根本上受到限制。

最近神经科学在哺乳类动物大脑上的新发现给信息表征提供了指导原则，给信息表征的系统的设计提供了新的思路。其中一项关键的发现是，和多种认知能力有关的大脑皮层其实没有直接对感官的信号进行预处理，而是允许它们通过一些模块的复杂层次结构进行传播，这些模块随着时间的推移会根据自己表现出来的规律来学会表征这些观察得到的信息。这项发现促使了深度机器学习的一个分支的出现，而这个分支关注的是表现得具有类似大脑皮层特性的信息表征的计算模型。

在现实世界中的数据里，除了其中关于空间的维度，其中的关于时间的部分也扮演着一个重要的角色。一个可观测到的模式的先后顺序往往也在给观测者传达着一些意义，而这些意义是很难孤立地通过这个序列里互相独立的片段来解读的。而意义往往是通过时间上相近的收到的观察信息或者事件来推理得到的。因此，给观察到的信息的时序部分建模在信息表征里起到了关键的作用。因此，根据观察中的规律捕捉到其中的时空相关性，被视为深度学习系统的一个基本目标。

假设一个鲁棒的深度学习被实现了，那么将有可能在一个巨大的数据集上训练出来一个分层的网络，然后从这个网络中提取出信号来交给一个相对简单的分类引擎，从而实现具有鲁棒性的模式分类的目的。而这里的鲁棒性是指能够表现出来对各种各样的变换和失真的分类不变性，其中包括噪声、放缩、旋转、多变的光照条件、位移等。

这篇文章将提供一个最近十年来深度学习的主流方法和研究方向的概论。在此有必要强调一下，每一种方法都有它的长处和短处，这取决于它的应用领域和上下文环境。因此，这篇文章提出了一个关于深度机器学习领域的概况和一些关于它将如何发展的观点。卷积神经网络（CNNs）和深度信念网络（DBNs）（还有它们各自的变形）是最备受关注的，因为他们在深度学习的领域中都是被良好地构造出的，而且在未来的工作中表现出了极大的潜力。第二节介绍了 CNNs ，而随后的是第三节里的 DBNs 的细节。如果想要对这些技术有更深入的了解，读者可以去看。第四节包括其他目前提出的深度架构。第五节包括一个简短的关于这些研究成果是如何影响政府或者工业创新的记录。结论部分提供了一个关于深度分层架构的潜在影响的视角和一些有待回答的关键问题。

1. 卷积神经网络

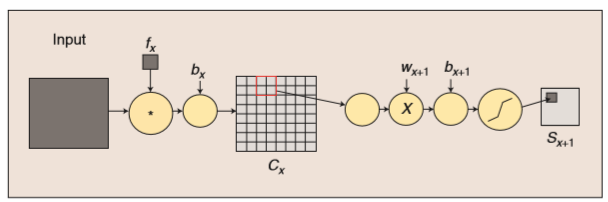


Figure 1 卷积和二次采样处理：卷积处理就是，对一个输入（第一阶段的输入是图像，之后的阶段里输入是特征图）使用一个可学习的滤波器 fx 进行卷积运算，然后加上一个可学习的偏置 bx 来产生一个卷积层 Cx 。二次采样处理就是对一个邻近元素求和（图中是四个像素），通过标量 Wx+1 进行加权，然后加上可学习的偏置 bx+1 最后通过一个 sigmoid 函数产生一个大约小两倍的特征图 Sx+1。

CNNs 是一个专门为二维数据（例如图片和视频）设计的多层神经网络的系列。 CNNs 受到了时间-延时神经网络（TDNN）的影响，而 TDNN 是一种通过时间维度的共享权值减少了学习时计算量的需求的方法，它被用于语音、时间序列数据的处理。CNNs 是第一种真正意义上成功的深度学习方法，它能够成功地用一种鲁棒的方式训练出层次结构中多层。CNN 是一种利用空间关系于减少必须训练的参数量的拓扑结构或者架构的选择，因此它能够提高前向传播和后向传播的效果。CNNs 这个深度学习框架的提出是为了尽可能地让数据的预处理的需求尽量的少。在 CNNs 中，图片中的各个小的局部（称之为局部感受野）都会输入到分层结构中的最底层。信息一般会通过应用每一层的数字滤波器获得所观察到的数据的显著特征来在网络的不同层次间传播。这种方法提供了具有一定程度上的位移、放缩和旋转的不变性的局部感受野，这允许神经元或者处理单元访问得到一些基本的特征，例如边缘和边角。

关于这个专题的其中一篇开创性的论文阐述了一个 CNNs 的对 MNIST 数据库里的手写数字进行分类的应用。本质上，图片其实是被N个小的，参数可以学习或者通过一些标准预先确定的滤波器进行过卷积了。因此，网络的第一层（或者说最底层）就是带有额外的偏置的卷积处理的结果——“特征图”，这些特征还有可能带有额外的压缩或者归一化。紧跟在初始阶段后面的，是一个进一步减少维度和提供一些空间位移不变性的二次采样（一般是一个 2x2 的平均运算），具体可以参见 Figure 1。经过二次采样的特征图将会乘以一个权值，再加上一个可学习的偏置，最后通过一个激活函数进行传播。这里面的一些变量应尽可能的少，少到一层只有一个图或者只有几个图的求和。

当权值很小的时候，激活函数将会表现得近乎于线性，结果将会是对原图像的一个模糊；其他的权值会造成激活输出类似于一个 AND 或者 OR 函数。这些输出构成了一个新的特征图，而这个图将会通过另一个卷积、二次采样和激活函数的序列的处理，就像 Figure 2 所说的一样。这个处理可以被重复任意次。值得说明的是，子序列的层可以由之前说明的层里的一个或者多个组成；例如，在 【8】 中，初始的6个特征图组合构成子序列的层里的 16 个特征图。正如 【33】 中所阐述的，CNNs 通过一种叫“特征池化”（ Figure 2 中的 S 层）的方法来创造他们的对象变换的不变性。但是，特征池化是由网络的构造者手动完成的，而不是通过系统自己去学习和训练出来的；在 CNNs 中，池化是由参数在学习的过程中“协调”出来的，但是它主要的机制（例如S层的输入的组合方式）是由网络的设计者设置好的。最后，在整个处理的最后阶段，激活的输出将转发到一个常规的前馈神经网络去产生一个系统最终的输出。

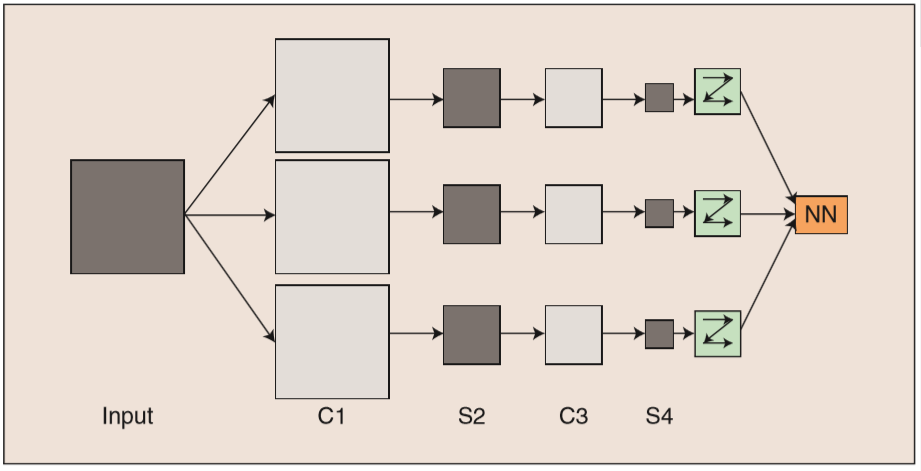


Figure 2 卷积神经网络的概念上的例子。输入图像通过三个可训练的滤波器和偏置（见 Figure 1）在 C1 的位置产生了三个特征图。特征图中的每四个像素（2x2）为一组进行求和、加权、偏置然后通过一个 sigmoid 函数来产生 S2 中的三个特征图。这些将再一次滤波得到 C3，然后再通过 S2 的方法产生出S4 。最终这些像素的值进行光栅化得到一个向量输入到一个常规的神经网络，最终由这个神经网络进行结果输出。

CNNs 中各个层和空间信息的紧密关系让它非常适合进行图片的处理和理解，而且他们在图片里自主提取显著特征的表现也十分良好。有时候，为了模拟人类的视觉对视觉激励的响应， Gabor 滤波器会被用作一个初始的预处理的步骤。在最近的一些更多的工作里，研究人员把 CNNs 应用在了各种各样的机器学习问题中，其中包括人脸识别，文档分析和语音识别。最近，CNNs 为了时间相关的目的被训练出来了，用于发现在视频中帧与帧之间的连贯性，虽然这个目的不需要一定要用CNNs 。

1. 深度信念神经网络（DBNs）

DBNs 最初在 [14] 里被提出，是一个概率生成模型，相对于传统神经网络的判断模型。生成模型在数据和标签之上提供了一个联合概率分布，在 P(观察值|标签) 和 P(标签|观察值) 上计算出一个更好的估计，而判断模型只局限于P(标签|观察值)。DBNs 是用来解决深层的神经网络在传统地使用反向传播的时候遇到的问题，也就是：（1）需要用于训练的大量标签数据；（2）缓慢的学习（即收敛）时间；（3）不适当的参数选择方法会导致学习收敛于局部最优解。

DBNs 由几层的限制玻耳兹曼机（一种神经网络，具体见Figure 3）组成。这些网络被“限制”为单个隐含层和单个可视层，而这里面的层与层之间形成有连接（同一个层里的单元之间没有连接）。这些隐含单元被训练成了能够捕捉得到来自可视单元观察到的高阶数据的相关性。开始的时候，不考虑上面的形成了联想记忆的两个层，一个 DBN 的连接只由自顶向下的生成权值所指导确定。 RBMs 由于它比起传统的和深层的 sigmoid 信念网络更易于连接权值的学习，所以它更加适合成为构建学习框架的模块。一开始通过一个非监督贪婪逐层方法去预学习来获得生成权值的做法，被Hinton证明是有效的，这种方法叫做对比分歧（contrastive divergence）。在这个训练的阶段，可视单元会产生一个向量 v ，通过它把值 传递到后面的隐含层里。反过来，可视单元会随机地输入，尝试去重构出原始的输入。最后，这些新的可视神经的激活会传递到这一步重构出来的隐含单元的激活，最后得到了h。重复地轮流执行这两步就是被称为 Gibbs 采样的过程，而隐含层激活单元和可视层的输入的相关性差别会作为权值更新的依据。训练时间显著地减少意味着，其实只需要一个步骤就可以近似地得到一个近似极大似然学习。网络里每增加一层都能提高训练数据的对数概率，我们可以理解为模型越来越接近真实的表达。这个有意义的与无标签的数据相结合的拓展，是所有深度学习应用中的关键的一部分。

在最高的两层里，权值是链接在一起的，因此，这样子低层的输出会给高层提供一个参考的线索或者链接去与它的内存里的值联系起来。我们经常遇到的问题里最关心的就是判别的性能，比如在分类的任务里。一个 DBN 在预训练之后可以通过 BP 算法利用有标签数据来进行微调以提高判别的性能。在这里上，一组标签会被附加到最高层（推广联想记忆），通过由底向上的识别的已经被学习好的权值来在网络中获得一个能够分清楚类别的分类面。这个的性能会比传统的完全只用 BP 算法来训练的网络要好。这个在直觉上解释为，因为在DBNs 中 BP 算法只需要对权值（参数）空间进行一个局部的搜索，因此它的训练速度和收敛时间都比传统的前向神经网络要优秀。

通过把 DBNs 应用在 MNIST 手写字符识别任务上，已经证明显著提高了前向网络得到的性能结果。在 DBNs 被提出来之后不久，一个更彻底的分析在 【17】 中被提出，证实了他们在无监督学习中的使用，并且将一直会重视其投入使用。【18】【19】中的进一步测试说明了 DBNs 的（以及其他深度学习的架构）在变量增长的问题中的弹性。

随着最近一个叫卷积深度信念网络（CDBNs）的概念的提出， DBNs 的灵活性被进一步地扩大了。 DBNs 本质上没有包含一副输入图像的二维结构的信息，也就是说输入是一个转换为简单的向量格式的图像矩阵。与此相反，CDBNs 采用所谓的卷积 RBMs 把相邻像素的空间关系利用起来了，提供了一个具有变换不变性的生成模型，这个模型能够很好地推广到高维图像。 DBNs 没有明确地处理观察量之间的时间关系，虽然已经有这方面的研究，例如堆叠时间 RBMs 或者它的推广，针对序列学习的被称为时间卷积机（Temporal Convolution Machines）的方法。这些序列学习的应用给语音信号处理问题，一个 DBNs 取得新进展地领域，带来了一个让人激动的未来研究的方向。

DBNs 和 CNNs 的静态图像的测试最通常使用在 MNIST ——手写数字数据集和 Caltech-101 ——有着101种类别的各种对象的数据集上。每种网络的分类错误率可以在【19】[20】【21】里找到。一个综合的，最新的关于各种机器学习技术应用在 MNIST 的比较在【27】中可以找到。

最近与DBNs相关的研究包括了，堆叠自动编码器，它是通过堆叠自动编码器来替换掉 DBNs 里的 RBMs。这就使得可以通过同样的规则来训练产生深度多层神经网络架构，但它缺少层的参数化的严格要求。不像 DBNs ，自动编码器使用判别模型，这个模型导致输入采样空间无法被采样，这就使得网络更难捕捉得到输入采样空间的内部表达。但是降噪自动编码器却能很好地避免这个问题，它通过在训练过程中加入随机的噪声，并进行堆叠以产生泛化性能。这种方法不仅能跟 DBNs 媲美，而且在一些情况下甚至比 DBNs 更好。而训练一个单一地降噪自动编码器的过程跟 RBMs 训练生成模型的过程一样。

1. 最近提出的深度学习的框架

有几个尝试着对大脑皮层进行建模的计算框架。这些模型收到一些来自类似【42】的启发，它们都试图把各种计算的阶段都映射到大脑皮层的各个区域。随着时间的推移，这些模型已经被细化了；但是视觉处理需要建立在层次结构上的这个中心概念还一直保持着。这些模型参考的由Hubel and Weisel【44】提出来的从简单到复杂的细胞组织结构，是建立在对猫的视觉皮层细胞地研究上的。

类似的组织结构也被 CNNs 和其他的深层模型所利用（例如 Neocognitron 【40】【41】【43】和 HMAX【32】【45】），然而更多的“显式”的皮层模型还在它们的架构到由生物启发的模型之间寻找一个更强的映射关系。尤其是当他们试图通过不同的机制去解决学习和不变形的问题的时候，例如时间被认为是学习过程中不可分割的要素的时序分析。

一个突出的例子是在 Numenta公司开发的层次结构的时序记忆（HTM）【30】【33】。HTMs 有一个基于在【39】里面描述的概念的层次结构，而且和其他属于皮层神经元环路建模的工作很类似。HTM 的重点是在视觉信息的表达，所以它的层次结构的最底层是负责接收来自一个输入图像的一小部分区域的输入的。层次结构中的更高层次对应的是更大的区域（或者感受野），它们是通过加入多个更低层次的感受野来构造出来的表达。除了层次结构中层与层之间比例地变化，每一层都有一个重要的基于时间的由输入图像本身变换或者扫描出来的形态。

在学习的阶段，第一层编译了绝大部分的输入的模式，并且给它们都分配好了索引。时序关系被建模成了从一个输入的序列到另一个的概率转移的形式，并通过图划分的技术聚集到了一块儿。当本阶段的学习进行总结时，随后的（第二）层从它的子模块连接当前输入的观察量的引用，并且学习到作为字母表（另一组最常见的但是在更高层次）的最一般的串联方式（类似马尔科夫链的形式）。高层的表征可以作为反馈传递到低层的模块里。反过来的，较低的层次会将这个更广泛的代表性信息纳入自己的推理的依据中。这个过程会在整个架构里的每一层里重复。当网络进行训练以后，图像识别会通过贝叶斯置信传播算法【46】去识别出最有可能的输入的模式，这个模式会在结构中的最高层给出结果（对应于其范围最广泛的图像）。其他在文献中提出的类似于 HTMs 的架构的，包括有Miller 和 Lommel 的分层绗缝【47】，这个采用了使用自组织图的两级的空间聚类和时间聚类，另外还有 Behnke 的神经抽象金字塔【48】。

深度时空推理网络模型（ DeSTIN ）是被提出来的，用来实现信息表达的鲁棒性的框架。在这个框架中，常见的神经元环路（或者节点）填充了整个架构，而且这里面的每一个节点的操作都是独立的，而且并行与其他的所有节点。

1. [↑](#endnote-ref-0)