

# 深度学习结构和算法比较分析

李海峰<sup>1</sup>, 李纯果<sup>2</sup>

(1. 河北大学 教务处, 河北 保定 071002; 2. 河北大学 数学与计算机学院, 河北 保定 071002)

**摘要:** Hinton 等人提出的深度机器学习, 掀起了神经网络研究的又一个浪潮. 介绍了深度机器学习的基本概念和基本思想. 对于目前比较成熟的深度机器学习结构深度置信网 DBNs 和约束 Boltzmann 机 (RBM) 的结构和无监督贪婪学习算法作了比较详细的介绍和比较, 并对算法的改进方向提出了有建设性的意见, 对深度机器学习的未来发展方向和目前存在的问题进行了深刻的分析.

**关键词:** 深度机器学习; 无监督贪婪学习算法; DBNs; RBMs

**中图分类号:** TP391      **文献标志码:** A      **文章编号:** 1000-1565(2012)05-0538-07

## Note on deep architecture and deep learning algorithms

LI Haifeng<sup>1</sup>, LI Chunguo<sup>2</sup>

(1. Department of Academic Affairs, Hebei University, Baoding 071002, China;

2. College of Mathematics and Computer Science, Hebei University, Baoding 071002, China)

**Abstract:** Deep architectures proposed by Hinton et al stir up another study wave in neural networks. This paper introduced the idea and basic concepts in deep learning. DBNs and RBMs are the advanced structures of deep learning, whose structures and effective learning algorithm are also introduced in detail in this paper. In addition, open questions in deep learning are also briefly displayed so that researchers who are interested in can devote themselves into those questions and solve them.

**Key words:** deep learning; greedy learning algorithm; DBNs; RBMs

随着电脑的普及与发展, 智能化、机械化成为人们关注的热点. 机器学习是仿照人类大脑工作的方式, 让电脑进行计算, 学习到类似于大脑的工作方式. 为此, 研究学者需要构建计算机能够运作的模型, 例如, 神经网络就是根据人类的大脑神经的激活或抑制的信号传输构建的模型<sup>[1]</sup>. 神经网络的基本组成单位就是神经元, 神经元的构造方式完全模拟了人类大脑细胞的结构, 如图 1. 但是, 显而易见, 人工神经元只是简单的结构的模拟, 要想达到与生物神经元有相同的功能, 还远远不够. 科研工作者就其训练的方式对其进行训练, 试图让人工神经网络的运算功能尽可能的与人类接近. 简单的网络已经可以进行基本的运算, 甚至有 2 个隐含层的非线性神经网络已经能够对任意的函数进行平滑的逼近. 从 1943 年 McCulloch 和 Pitts 提出的简单神经元开始, 神经网络经历了几度兴衰. 神经网络已经深入到各个领域, 技术相对比较成熟, 然而也很难再有

收稿日期: 2012-04-05

基金项目: 保定市科学技术研究与发展指导计划项目(12ZG005); 河北省高等学校科学研究计划项目(JYGH2011011)

第一作者: 李海峰(1980-), 男, 河北唐县人, 河北大学讲师, 主要从事机器学习、教学信息化等研究.

E-mail: lihf@hbu.edu.cn

新突破. 人类完成的日常生活中的各种简单的动作, 如果让计算机来完成, 就需要高度复杂的神经网络来完成. 因此, Hinton 等人提出了深度学习, 掀起了神经网络研究的又一次浪潮.

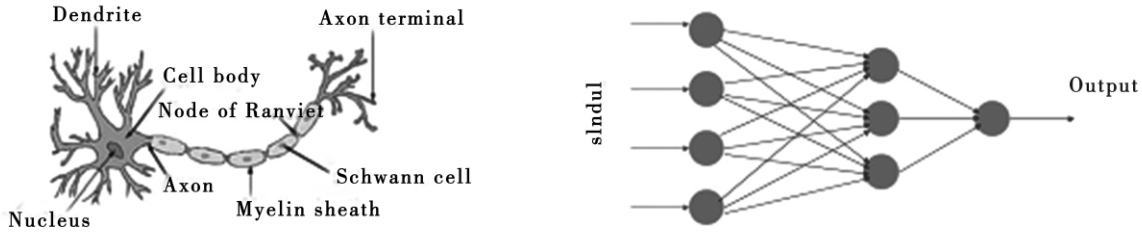


图 1 生物神经元与人工神经元  
Fig. 1 Structure of biological and artificial neurons

### 1 深度学习

深度学习是为了能够得到有助于理解图片、声音、文本等的数据所表述的意义而进行的多层次的表示和抽取的学习<sup>[2]</sup>. 例如, 给定图 2 中的图片, 大脑做出的反应是: “许多黄色的郁金香.” 同样的图片, 输入到计算机中, 是描述图片的最原始数据, 那就是用向量表示的像素. 用简单的机器学习, 例如用含 2 个或 3 个隐含层的神经网络, 是不可能达到与人类类似的判别决策的. 这就需要多层的学习器, 逐层学习并把学习到的知识传递给下一层, 以便下层能够得到更高级别的表述形式, 期望可以得到与人类类似的结论<sup>[2]</sup>.

#### 1.1 学习的深度

学习器的深度, 决定于学习器的构造. 假设学习器为一个有向流通图, 那么深度就是从开始结点到结束结点(或从输入结点到输出结点)的最长路径. 例如, 一个支撑向量机的深度是 2, 是输入经过一个核变换到核空间, 再加上一个线性组合. 再如多层前传神经网络的深度是隐含层层数加 1(输出层). 如果说学习到一次知识, 就是一个深度的话, 那么, 学习的深度是原始数据被逐层学习的次数.

根据学习的深度, 机器学习可以分为浅度学习和深度学习. 对于简单的计算, 浅度学习可以有效地进行计算, 例如二进制数据的逻辑运算. 显然, 如果想让机器达到人脑的反应效果, 浅度学习是远远不够的, 必须进行深度的机器学习, 才有可能得到与人脑反应近似的结果. 实际上, 深度的机器学习正是模拟了人脑的工作方式. 对于图 2 中的图片, 先由视网膜接受数据信号, 视网膜通过神经链接, 把看到的图片转化成脑波信号传输到大脑中, 由于大脑的不同部位处理不同的问题, 信号不可能一下子就传到相应位置, 需要层层传输. 同时, 在信号传输过程中, 大脑会提取不同的信息, 例如, 花的颜色、形状、个数、位置、个体差异等等. 因此, 深度的机器学习模型需要具备类似的特征, 也即, 深度的机器学习模型可以提取观察对象的不同方面的特征. 为此, 深度的机器学习模型通常为分层结构, 每一层提取数据的 1 个或多个不同方面的特征, 并把提取出的特征作为下一层的输入. 图 3 是一个典型的深度学习模型.



图 2 待识别的图片  
Fig. 2 Picture for recognition

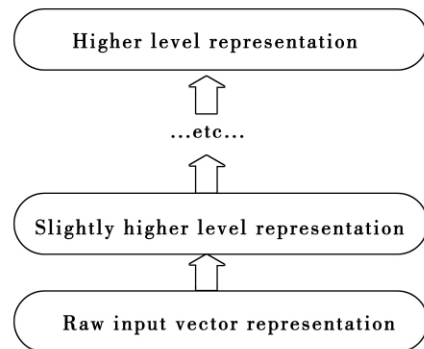


图 3 深度机器学习模型  
Fig. 3 Deep learning model

## 1.2 深度学习的动机

从早期的神经网络的学习,到现在的深度学习,究其机制,都是从模拟大脑的构架并辅以一定的学习算法,从而使计算机的工作方式尽可能地接近人类的工作方式.机器学习从仅有 2 层左右的学习构架,要向有多层的结构发展,不仅有生物神经元的启示,也是对现有的机器学习结构的弊端的改进.

首先,人类大脑的神经元系统是一个庞大的结构,由无数个神经元共同组成,完成一定的生理功能.例如,从视网膜到处理视网膜的大脑区域,需要经过无数层的神经元层层传递视觉信息,最终到达大脑的视觉处理区域,然后再经过信息处理,把信息反馈到肌肉神经,或语言区域.这个过程在生物神经元系统只不过是瞬间的事情,但是,完成这个过程,是由已经训练好的神经系统完成的,神经系统对整个过程的处理,与从出生到成人的认知过程是分不开的.而这一切,要用电脑来完成,不是构造简单的人工神经元就能够完成的,需要大规模的神经元组织和链接,并经过来自于外界信息的不断强化和训练.故从结构上,神经网络结构要加深.

其次,如果神经网络的层次不够深,那么,多项式级的参数个数能解决的问题就有可能需要指数级的参数个数.例如,要计算  $\prod_{i=1}^n \sum_{j=1}^m a_{ij}$ ,计算方式不同,计算复杂度会有很大的不同.如果计算和的积,计算复杂度为  $O(mn)$ ;如果计算积的和,计算复杂度为  $O(n^m)$ .参数多的结构不仅训练复杂,训练时间长,而且泛化性能也很差,也容易产生过拟合问题.

再次,有很多学习结构的学习算法使得到的学习器是局部估计算子.例如,由核方法构造的学习器,  $f(x) = b + \sum_{i=1}^n \alpha_i K(x, x_i)$ ,是由对模板的匹配度加权构成.对于这样的问题,往往要求被拟合函数都是平滑的,如果不平滑,该学习器就不能很好地拟合数据.而在实际应用中的很多例子,往往不满足平滑这个条件,或者对于被拟合函数一无所知.此时,模板匹配方法就不再可行.

最后,深度机器学习是数据分布式表示的必然结果<sup>[2]</sup>.分布式表示系统是由一组信息计算单元共同完成的,每一个概念都由不止 1 个信息计算单元表示,而每一个信息计算单元也参与计算不止 1 个概念<sup>[3]</sup>.例如,表示整数  $i(i=1, 2, \dots, N)$ ,有 2 种表示方式,一种用 1 个  $N$  维向量  $r(i)$  表示,  $r_j(i) = \begin{cases} 1, & j=i \\ 0, & j \neq i \end{cases}$ ,另一种表示方式是用  $\log_2 N$  维的向量表示,表示方式类似于二进制.显然,后一种表示方式更紧凑,这种表示方式就是分布式表示.分布式表示不仅可以很好地描述概念间的相似性,而且合适的分布式表示在有限的数据下能体现出更好的泛化性能.人类的认知活动包括理解和处理接受到的信息,这些信息的结构通常很复杂,因此,有必要构造深度结构的学习机器去实现一些人类的认知活动.

由于其自身的复杂性,深度学习算法很多年都没有新的进展.就监督的多层神经网络来说,无论是测试精度还是训练精度,深度学习的结果远远不如有 1 个或 2 个隐含层的神经网络的结果.直到 2006 年, Hinton 等人提出了贪婪无监督逐层学习算法<sup>[4]</sup>,深度学习的问题才有所突破.

## 2 深度学习的方法

同机器学习方法一样,深度机器学习方法也有监督学习与无监督学习之分.不同的学习框架下建立的学习模型很是不同.例如,卷积神经网络(Convolutional neural networks, 简称 CNNs)就是一种深度的监督学习下的机器学习模型,而深度置信网(Deep Belief Nets, 简称 DBNs)就是一种无监督学习下的机器学习模型.

### 2.1 卷积神经网络

20 世纪 60 年代, Hubel 和 Wiesel 在研究猫脑皮层时,发现了一种独特的神经网络结构,可以有效地降低反馈神经网络的复杂性,进而提出了卷积神经网络<sup>[5]</sup>.现在,卷积神经网络已经发展成一种高效的图像识别方法<sup>[6]</sup>.

卷积神经网络的基本结构如图 4 所示.卷积神经网络是一种非全连接的神经网络结构,包含 2 种特殊的结构层:卷积层和次抽样层.卷积层由多个特征平面构成,完成抽取特征的任务.每个特征平面由神经元构成,在同一个特征平面上的所有神经元具有相同的连接权重.对于每个神经元,定义了相应的接受域<sup>[7-8]</sup>,只

接受从其接受域传输的信号,如图 5. 同一个特征平面的神经元的接受域具有相同的大小,故而每个神经元的连接矩阵相同. 根据卷积的数学定义,  $(f * g)(x) = \int f(u)g(x-u) du$ , 或  $(f * g)(n) = \sum_m f(m)g(n-m)$ , 网络的卷积层应用了卷积的离散表示. 设所考查层的第  $k$  个特征平面记作  $h^k$ , 与前一层的连接权重矩阵为  $W^k$ , 偏差为  $b^k$ , 对于非线性的正切函数, 可以得到特征映射如下:

$$h_{ij}^k = \tanh((W^j * x)_{ij} + b^k), \tag{1}$$

其中,  $i$  和  $j$  标注了该神经元在特征平面上的位置.

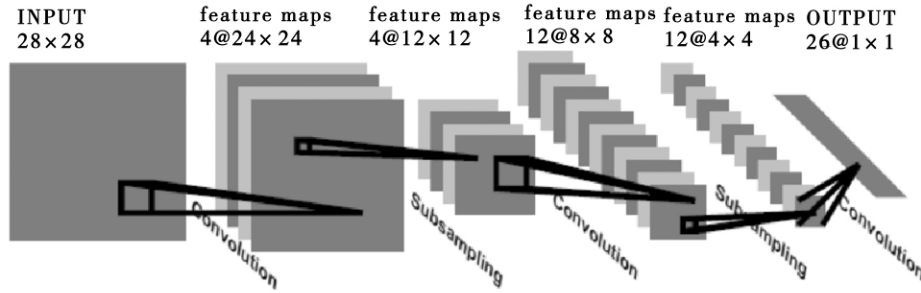


图 4 用于图像识别的卷积神经网络结构

Fig. 4 Convolution network for image recognition

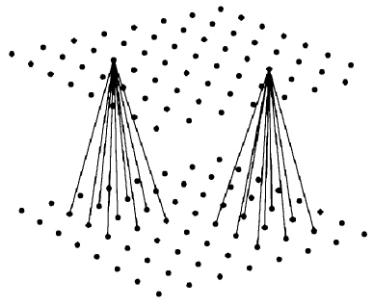


图 5 卷积层的接受域

Fig. 5 Receptive field of one neuron in a convolution layer

每个卷积层都会紧跟 1 个次抽样层. 输入数据经过卷积后, 进入高维空间, 换句话说, 卷积层进行了升维映射. 如果不断地进行升维, 则不可避免地陷入维数灾难. 同卷积层类似, 次抽样层的每个特征平面上的神经元也共享连接权重, 且每个神经元都从其接受域中接受数据. 卷积层的每个特征平面都对应了次抽样层的 1 个特征平面, 次抽样层中的神经元对其接受域中的数据进行了抽样(例如, 取大, 取小, 取平均值, 等等), 因此次抽样层的特征平面上的神经元的个数往往会减半.

卷积层的每一个平面都抽取了前一层某一个方面的特征. 每个卷积层上的每个结点, 作为特征探测器, 共同抽取输入图像的某个特征, 例如 45°角、反色、拉伸、翻转、平移等. 图像经过一层卷积, 就由原始空间被影射到特征空间, 在特征空间中进行图像的重构. 卷积层的输出, 为图像在特征空间中重构的坐标, 作为下一层也就是次抽样层的输入.

LeCun 从 1998 年开始, 专注于卷积神经网络的研究, 提出了 LeNet 模型<sup>[8]</sup>(如图 6), 用于识别手写和机打字体, 逐渐已经适用识别很多类图形问题. 由图 6 所示, 输入层后有 2 组隐含层抽取输入图像的特征, 最后有一个全连接的隐含层完成对输入图像的认识. LeNet 模型在识别手写数字上达到很高的识别率, 而且具有拉伸、挤压、反转的不变性, 而且抗噪能力很强. 模型用传统的 BP 进行训练.

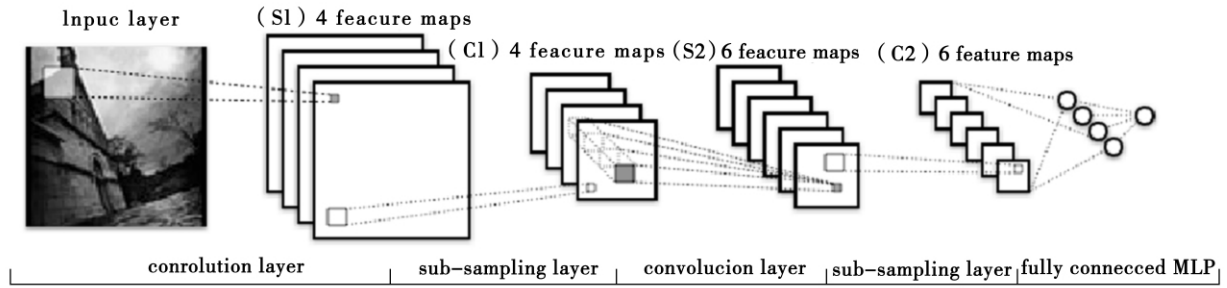


图 6 LeNet 模型

Fig. 6 LeNet Model

### 2.2 深度置信网

深度置信网(Deep Belief Networks, 简称 DBNs)是另一种深度学习结构<sup>[4]</sup>, 是多个 RBM 的累加. 一个 Boltzmann 机是基于能量理论的概率模型, 通过热力学的能量函数定义了一个概率分布: Boltzmann 分布. 设状态随机变量  $x$ , 能量函数为  $E(x)$ , 则该状态出现的概率分布密度函数为  $p(x) = e^{-E(x)} / Z$ , 其中  $Z = \sum_x e^{-E(x)}$ . 一个典型的 Boltzmann 机是一个无向循环图, 其能量函数定义为

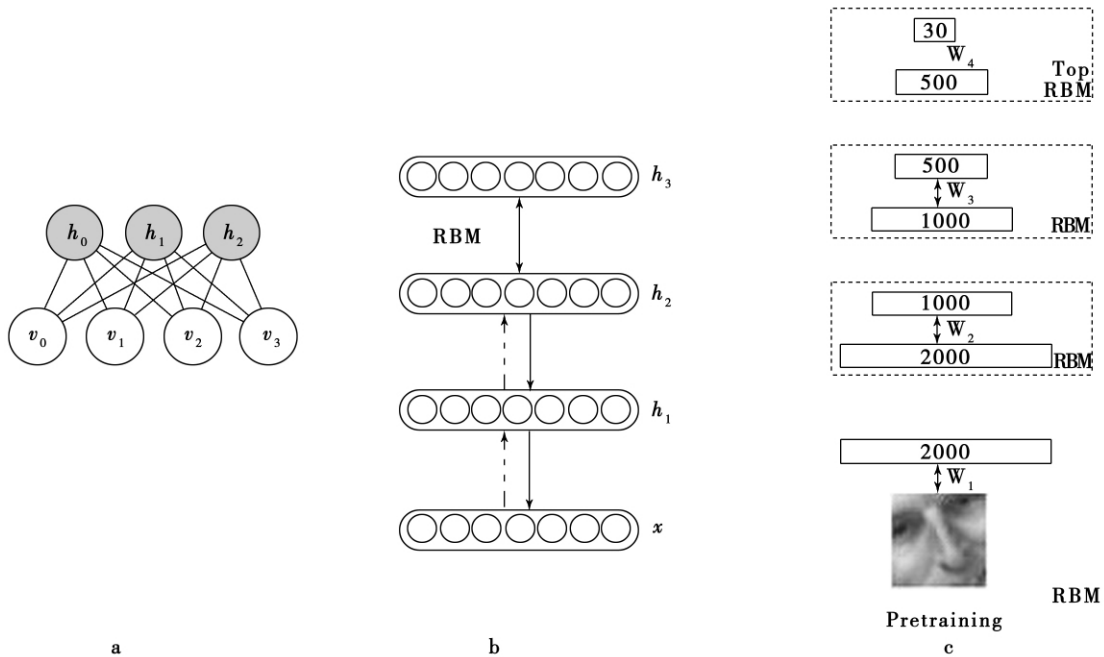
$$E(x, h) = -b'x - c'h - x'Wh - x'Ux - h'Vh, \tag{2}$$

其中  $x$  是可见变量, 描述可以观察的数据;  $h$  是隐含变量, 无法观察到其实际取值;  $b$  和  $h$  分别是可见变量和隐含变量的阈值;  $W, U, V$  是结点之间的连接权重. 如果对 Boltzmann 机加以约束条件, 令其自身不与自身连接, 则得到一个有向无环图 RBM(如图 7a), 其能量函数定义为  $E(x, h) = -b'x - c'h - x'Wh$ .

一个典型的置信网可以看成是由多个随机变量组成的有向无环图, 也可以看成是多个 RBM 的累加, 而深层置信网就是一个复杂度很高的有向无环图. Hinton 等人<sup>[3]</sup>认为, 一个有  $l$  个隐含层的典型的 DBN, 可以用联合概率分布刻画输入向量  $x$  和隐含向量  $h$  的关系

$$P(x, h^2, \dots, h^l) = (\prod_{k=1}^{l-2} P(h^k | h^{k+2})) P(h^{l-2}, h^l), \tag{3}$$

其中  $x = h^0$ ,  $P(h^k | h^{k+2})$  是条件概率分布. DBN 学习的过程, 就是学习联合概率分布的过程. 而联合概率分布的学习是机器学习中的产生式学习方式.



a. RBM 模型; b. DBN 典型结构; c. 用 DBN 进行图像识别示例.

图 7 深度学习模型

Fig. 7 Deep Learning Model

对于深度的机器学习,由于参数变量很多,所以合适的训练算法直接决定了学习器的性能.以往的基于最速梯度下降的 BP 算法,在经典的神经网络中被广泛应用,可以得到泛化性能很好的网络结构,但是 BP 算法对于深度学习器的训练却存在一定的困难.这主要是 BP 算法本身的约束所在.首先,BP 算法是监督学习,训练数据必须是有类标数据.但是,实际能得到的数据大都是无类标数据.其次,BP 算法不适合有很多隐含层的学习结构,一是计算偏导数很困难,二是误差需要层层逆传,收敛速度很慢.最后,BP 算法经常会陷入到局部最优解,不能到达全局最优解.因此,Hinton 等人提出了贪婪的逐层无监督训练算法<sup>[4]</sup>.

贪婪无监督学习算法的基本思想是,把一个 DBN 网络分层,对每一层进行无监督学习,最后对整个网络用监督学习进行微调.把一个 DBN 网络分层,每层都由若干计算单元(常常是几百个或几千个)组成(如图 7b),各自独立计算该层接受到的数据,每个层的节点之间没有连接.与外界环境连接的节点层为输入层,输入层接受来自于外界输入,例如图像数据.第 1 层(即输入层)与第 2 层构成一个典型的 RBM,根据无监督学习调节网络参数,使得 RBM 达到能量平衡.然后,第 1 层的输出作为第 2 层与第 3 层构成一个新的 RBM,第 1 层的输出作为外界输入,继续调节参数,使当前 RBM 结构达到能量平衡.如此进行下去,直到最后一层(如图 7c).当完成无监督逐层训练学习后,再以原始外界输入和目标输出对整个网络进行有监督学习,以最大似然函数为目标,精调网络各层的参数.

Gibbs 抽样技术是在训练每个 RBM 时采用的有效随机抽样技术<sup>[11]</sup>.设需要从未知的联合概率分布  $f(x_1, \dots, x_k)$  中抽取  $n$  个样本  $X^{(1)}, X^{(2)}, \dots, X^{(n)}$ . 由于

$$p(x_j | x_1, \dots, x_{j-1}, x_{j+1}, \dots, x_k) = \frac{p(x_1, \dots, x_k)}{p(x_1, \dots, x_{j-1}, x_{j+1}, \dots, x_k)} \propto p(x_1, \dots, x_k),$$

从而对某一个变量抽取样本时,假设其他变量已知,在以其他变量为条件的概率分布下进行抽取,直到抽取所有样本.换句话说,对于样本  $X^{(i)}$  的第  $j$  个变量,是从分布  $p(x_j^{(i)} | x_1^{(i)}, \dots, x_{j-1}^{(i)}, x_{j+1}^{(i)}, \dots, x_k^{(i)})$  中抽取,直到抽取  $n$  个样本.

在贪婪学习算法中,也采用了 Wake-Sleep 算法的基本思想<sup>[12]</sup>.算法在觉醒阶段,采用学习到的权重,按照自底朝上的顺序,为下一层产生训练需要用的数据,而在睡眠阶段,按照自顶朝下,用权重对数据进行重建,如表 1.

表 1 贪婪学习算法实现步骤  
Tab. 1 Implementing procedure of greedy layer-wise learning

步骤	内容
第 1 步	以 $x=h^0$ 为外界输入,训练第 1 个 RBM,达到能量平衡;
第 2 步	用第 1 层学习到的输入的联合分布,作为第 2 层的 RBM 的输入,继续训练;
第 3 步	重复第 2 步,直到最后一层;
第 4 步	以最大似然函数为目标函数,精调网络参数,使网络达到最优.

### 3 总结

神经网络是人工智能领域的一个重要分支,利用神经网络可以任意精度逼近任意光滑的曲线,这使得神经网络成为人工智能、数据挖掘等领域的一个重要工具.本文主要是简要介绍了深度机器学习的主要思想,以及有效的学习算法.深度机器学习是神经网络又一次兴起的标志.但是,深度机器学习的训练时间过长,常常需要几个星期的训练时间,如果能合并训练,提高训练速度,则会大大提高深度机器学习的实用性.另外,深度机器学习学习到的知识表示的物理意义很不明确,如果能把各层学习到的知识表示成有物理意义的知识,则会增加学习到知识的可理解性.这些问题都有待解决.

## 参 考 文 献:

- [1] HAYKIN S. Neural Networks: A comprehensive foundation[M]. 2nd ed. New York: Prentice-Hall, 1999.
- [2] BENGIO Y. Learning deep architectures for AI[J]. Foundations and Trends in Machine Learning, 2009, 2(1):1-127.
- [3] HINTON G E, MCCLELLAND J L, RUMELHART D E. Distributed Representations[M]. Cambridge: MIT Press, 1986.
- [4] HINTON G E, OSINDERO S. A fast learning algorithm for deep belief nets[J]. Neural Computation, 2006, 18:1527-1554.
- [5] HUBEL D, WIESEL T. Receptive fields, binocular interaction, and functional architecture in the cat's visual cortex[J]. Journal of Physiology, 1962, 160:106-154.
- [6] LECUNY, KAVUKCUOGLUK, FARABET C. Convolutional networks and applications in vision[Z]. International Symposium on Circuits and Systems, Paris, 2010.
- [7] LASERSON J. From neural networks to deep learning: zeroing in on the human brain [J]. XRDS, 2011, 18(1):29-34.
- [8] LECUNY, BOTTOU L, BENGIO Y, et al. Gradient-based learning applied to document recognition[J]. Proceedings of IEEE, 1998, 86(11): 2278-2324.
- [9] ERHAND, BENGIO Y, COURVILE A, et al. Why does unsupervised pre-training help deep learning[J]. Journal of Machine Learning Research, 2010, 11:625-660.
- [10] BENGIO Y, LAMBLIN P, POPOVICI D, et al. Greedy layer-wise training of deep networks[J]. Advances in Neural Information Processing Systems, 2007, 19:153-160.
- [11] BISHOP C M. Pattern recognition and machine learning[M]. New York: Springer, 2006.
- [12] HINTON G E, DAYAN P, FREY B, et al. The wake-sleep algorithm for unsupervised neural network[J]. Science, 1995, 268:1158-1161.

(责任编辑:孟素兰)