深度学习研究进展*

刘建伟，刘 媛，罗雄麟
（中国石油大学 自动化研究所，北京 102249）

摘 要：鉴于深度学习的重要性，综述了深度学习的研究进展。首先概述了深度学习具有的优点，由此说明了引入深度学习的必要性；然后描述了三种典型的深度学习模型，包括卷积神经网络模型、深度信任网络模型和堆栈自编码网络模型，并对近几年深度学习在初始化方法、网络层数和激活函数的选择、模型结构、学习算法和实际应用这四个方面的研究新进展进行了综述；最后探讨了深度学习在理论分析、数据表示与模型、特征提取、训练与优化求解和研究拓展这五个方面中有待进一步研究解决的问题。

关键词：深度学习；神经网络；模型；表示；堆栈；预训练
中图分类号：TP181 **文献标志码**：A **文章编号**：1001-3695(2014)07-1921-10
doi:10.3969/j.issn.1001-3695.2014.07.001

Research and development on deep learning

LIU Jian-wei, LIU Yuan, LUO Xiong-lin
(*Research Institute of Automation, China University of Petroleum, Beijing* 102249, *China*)

Abstract: In view of the significance of deep learning, this paper reviewed the research and development on deep learning. Firstly, this paper summarized the advantage of deep learning, and illustrated the necessity of introducing deep learning. Secondly, it described three kinds of typical deep learning models, included convolutional neural network model, deep belief network model, and stacked auto-encoder network model. Thirdly, it reviewed new research and development on deep learning in recent years, included the choice of initialization methods, the number of network layers, and activation function, model structure, learning algorithms, and practical application. Finally, it presented the problems to be solved in aspects of theoretical analysis, representation and model of data, feature extraction, training and optimization, and research extension.

Key words: deep learning; neural network; model; representation; stacking; pre-training

0 引言

许多研究表明，为了能够学习表示高阶抽象概念的复杂函数，解决目标识别、语音感知和语言理解等人工智能相关的任务，需要引入深度学习(deep learning)。深度学习架构由多层非线性运算单元组成，每个较低层的输出作为更高层的输入，可以从大量输入数据中学习有效的特征表示，学习到的高阶表示中包含输入数据的许多结构信息，是一种从数据中提取表示的好方法，能够用于分类、回归和信息检索等特定问题中。

深度学习的概念起源于人工神经网络的研究，有多个隐层的多层感知器是深度学习模型的一个很好的范例。对神经网络而言，深度指的是网络学习得到的函数中非线性运算组合水平的数量。当前神经网络的学习算法多是针对较低水平的网络结构，将这种网络称为浅结构神经网络，如一个输入层、一个隐层和一个输出层的神经网络；与此相反，将非线性运算组合水平较高的网络称为深度结构神经网络，如一个输入层、三个隐层和一个输出层的神经网络。深度学习与浅学习相比具有许多优点，说明了引入深度学习的必要性：

a)在网络表达复杂目标函数的能力方面，浅结构神经网络有时无法很好地实现高变函数等复杂高维函数的表示，而用深度结构神经网络能够较好地表征。

b)在网络结构的计算复杂度方面，当用深度为 k 的网络结构能够紧凑地表达某一函数时，在采用深度小于 k 的网络结构表达该函数时，可能需要增加指数级规模数量的计算因子，大大增加了计算的复杂度。另外，需要利用训练样本对计算因子中的参数值进行调整，当一个网络结构的训练样本数量有限而计算因子数量增加时，其泛化能力会变得很差。

c)在仿生学角度方面，深度学习网络结构是对人类大脑皮层的最好模拟。与大脑皮层一样，深度学习对输入数据的处理是分层进行的，用每一层神经网络提取原始数据不同水平的特征。

d)在信息共享方面，深度学习获得的多重水平的提取特征可以在类似的不同任务中重复使用，相当于对任务求解提供了一些无监督的数据，可以获得更多的有用信息。

深度学习比浅学习具有更强的表示能力，而由于深度的增加使得非凸目标函数产生的局部最优解是造成学习困难的主要因素。反向传播基于局部梯度下降，从一些随机初始点开始运行，通常陷入局部极值，并随着网络深度的增加而恶化，不能很好地求解深度结构神经网络问题。2006年，Hinton等人[1]提出的用于深度信任网络(deep belief network，DBN)的无监督

收稿日期：2013-09-17；修回日期：2013-11-05 基金项目：国家“973”计划资助项目(2012CB720500)；国家自然科学基金资助项目(21006127)；中国石油大学(北京)基础学科研究基金资助项目(JCXK-2011-07)

作者简介：刘建伟(1966-)，男，副研究员，博士，主要研究方向为智能信息处理、复杂系统分析、预测与控制、算法分析与设计(liujw@cup.edu.cn)；刘媛(1989-)，女，硕士研究生，主要研究方向为机器学习；罗雄麟(1963-)，男，教授，博士，主要研究方向为智能控制.

学习算法，解决了深度学习模型优化困难的问题。求解 DBN 方法的核心是贪婪逐层预训练算法，在与网络大小和深度呈线性的时间复杂度上优化 DBN 的权值，将求解的问题分解成为若干更简单的子问题进行求解。从具有开创性的文献[1]发表之后，Bengio、Hinton、Jarrett、Larochelle、Lee、Ranzato、Salakhutdinov、Taylor 和 Vincent 等大量研究人员[2~13]对深度学习进行了广泛的研究以提高和应用深度学习技术。Bengio 和 Ranzato 等人[2,12]提出用无监督学习初始化每一层神经网络的想法；Erhan 等人[14]尝试理解无监督学习对深度学习过程起帮助作用的原因；Glorot 等人[15]研究深度结构神经网络的原始训练过程失败的原因。许多研讨会，如 the 2009 ICML Workshop on Learning Feature Hierarchies，the 2008 NIPS Deep Learning Workshop: Foundations and Future Directions，the 2009 NIPS Workshop on Deep Learning for Speech Recognition and Related Applications 以及 the 2010 IEEE Trans on Audio，Speech，and Language Processing 等，都致力于深度学习及其在信号处理领域的应用。文献[16]对深度学习进行了较为全面的综述，基于无监督学习技术提出贪婪逐层预训练学习过程用于初始化深度学习模型的参数，从底层开始训练每层神经网络形成输入的表示，在无监督初始化之后，堆栈各层神经网络转换为深度监督前馈神经网络，用梯度下降进行微调。用于深度学习的学习方法主要集中在学习数据的有用表示，在神经网络较高层中使学习到的特征不随变化的因素而变化，对实际数据中的突发变化具有更强的鲁棒性[17]。文献[18]给出了训练深度学习模型的相关技巧，尤其是受限玻尔兹曼机(restricted Boltzmann machine，RBM)，许多来自神经网络训练的想法也可以用于深度结构神经网络学习[19,20]。Bengio 在文献[21]中给出了用于不同种类深度结构神经网络的训练方法的指导意见。深度学习方法已经被成功用于文本数据学习任务和视觉识别任务上[8,9,22~28]。

鉴于深度学习的理论意义和实际应用价值，国内对深度结构的研究尚处于起步阶段，这方面已经发表的文献相对较少而且多是侧重于应用领域，与国外已有综述文献[16,21]相比，本文系统综述了深度学习的最新研究进展，为进一步深入研究深度学习理论和拓展其应用领域奠定了一定的基础。

1 深度学习概述

1.1 深度学习表示模型和网络结构

深度学习方法试图找到数据的内部结构，发现变量之间的真正关系形式。大量研究表明，数据表示的方式对训练学习的成功产生很大的影响，好的表示能够消除输入数据中与学习任务无关因素的改变对学习性能的影响，同时保留对学习任务有用的信息[29,30]。

深度学习中数据的表示有局部表示(local representation)、分布表示(distributed representation)[31,32]和稀疏分布表示(sparse distributed representation)[33~35]三种表示形式。学习输入层、隐层和输出层的单元均取值 0 或 1。举个简单的例子，整数 $i \in \{1,2,\cdots,N\}$ 的局部表示为向量 $r(i)$，该向量有 N 位，由 1 个 1 和 $N-1$ 个 0 组成，即 $r_j(i) = 1_{i=j}$。分布表示中的输入模式由一组特征表示，这些特征可能存在相互包含关系，并且在统计意义上相互独立。对于例子中相同整数的分布表示有 $\log_2 N$ 位的向量，这种表示更为紧凑，在解决降维和局部泛化限制方面起到帮助作用。稀疏分布表示介于完全局部表示和非稀疏分布表示之间，稀疏性的意思为表示向量中的许多单元取值为 0。对于特定的任务需要选择合适的表示形式才能对学习性能起到改进的作用。当表示一个特定的输入分布时，一些结构是不可能的，因为它们不相容。例如在语言建模中，运用局部表示可以直接用词汇表中的索引编码词的特性，而在句法特征、形态学特征和语义特征提取中，运用分布表示可以通过连接一个向量指示器来表示一个词。分布表示由于其具有的优点，常常用于深度学习中表示数据的结构。由于聚类簇之间在本质上互相不存在包含关系，因此聚类算法不专门建立分布表示，而独立成分分析(independent component analysis，ICA)[36]和主成分分析(principal component analysis，PCA)[37]通常用来构造数据的分布表示。

典型的深度学习模型有卷积神经网络(convolutional neural network)、DBN 和堆栈自编码网络(stacked auto-encoder network)模型等，下面对这些模型进行描述。

1.1.1 卷积神经网络模型

在无监督预训练出现之前，训练深度神经网络通常非常困难，而其中一个特例是卷积神经网络。卷积神经网络受视觉系统的结构启发而产生。第一个卷积神经网络计算模型是在 Fukushima[38]的神经认知机中提出的，基于神经元之间的局部连接和分层组织图像转换，将有相同参数的神经元应用于前一层神经网络的不同位置，得到一种平移不变神经网络结构形式。后来，LeCun 等人[39,40]在该思想的基础上，用误差梯度设计并训练卷积神经网络，在一些模式识别任务上得到优越的性能。至今，基于卷积神经网络的模式识别系统是最好的实现系统之一，尤其在手写体字符识别任务上表现出非凡的性能。

LeCun 的卷积神经网络由卷积层和子抽样层两种类型的神经网络层组成。每一层有一个拓扑图结构，即在接收域内，每个神经元与输入图像中某个位置对应的固定二维位置编码信息关联。在每层的各个位置分布着许多不同的神经元，每个神经元有一组输入权值，这些权值与前一层神经网络矩形块中的神经元关联；同一组权值和不同输入矩形块与不同位置的神经元关联。卷积神经网络是多层的感知器神经网络，每层由多个二维平面块组成，每个平面块由多个独立神经元组成[41]。为了使网络对平移、旋转、比例缩放以及其他形式的变换具有不变性，对网络的结构进行一些约束限制：

a)特征提取。每一个神经元从上一层的局部接收域得到输入，迫使其提取局部特征。

b)特征映射。网络的每一个计算层由多个特征映射组成，每个特征映射都以二维平面的形式存在，平面中的神经元在约束下共享相同的权值集。

c)子抽样。该计算层跟随在卷积层后，实现局部平均和子抽样，使特征映射的输出对平移等变换的敏感度下降。

图 1 是一个用于手写体字符识别的卷积神经网络，由一个输入层、四个隐层和一个输出层组成。由图 1 可以看出，与完全连接的多层前馈感知器网络相比，卷积神经网络通过使用接收域的局部连接，限制了网络结构。卷积神经网络的另一个特点是权值共享，图中包含大量连接权值，但是由于同一隐层的神经元共享同一权值集，大大减少了自由参数的数量。

卷积神经网络本质上实现一种输入到输出的映射关系，能够学习大量输入与输出之间的映射关系，不需要任何输入和输

出之间的精确数学表达式,只要用已知的模式对卷积神经网络加以训练,就可以使网络具有输入输出之间的映射能力。卷积神经网络执行的是有监督训练,在开始训练前,用一些不同的小随机数对网络的所有权值进行初始化。

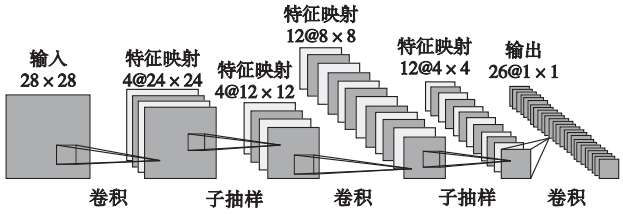

图1　用于手写体字符识别的卷积神经网络

卷积神经网络的训练分为两个阶段:

a)向前传播阶段。从样本集中抽取一个样本(X,Y_p),将X输入给网络,信息从输入层经过逐级变换传送到输出层,计算相应的实际输出:

$$O_P = F_n(\cdots(F_2(F_1(XW_1)W_2)\cdots)W_n) \tag{1}$$

b)向后传播阶段,也称为误差传播阶段。计算实际输出O_P与理想输出Y_P的差异:

$$E_p = \frac{1}{2}\sum_j (y_{pj} - o_{pj})^2 \tag{2}$$

并按最小化误差的方法调整权值矩阵。

卷积神经网络的特征检测层通过训练数据来进行学习,避免了显式的特征提取,而是隐式地从训练数据中学习特征,而且同一特征映射面上的神经元权值相同,网络可以并行学习,这也是卷积神经网络相对于其他神经网络的一个优势。权值共享降低了网络的复杂性,特别是多维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。

卷积神经网络的成功依赖于两个假设:a)每个神经元有非常少的输入,这有助于将梯度在尽可能多的层中进行传播;b)分层局部连接结构是非常强的先验结构,特别适合计算机视觉任务,如果整个网络的参数处于合适的区域,基于梯度的优化算法能得到很好的学习效果。卷积神经网络的网络结构更接近实际的生物神经网络,在语音识别和图像处理方面具有独特的优越性,尤其是在视觉图像处理领域进行的实验,得到了很好的结果。

1.1.2　深度信任网络模型

DBN 可以解释为贝叶斯概率生成模型,由多层随机隐变量组成,上面的两层具有无向对称连接,下面的层得到来自上一层的自顶向下的有向连接,最底层单元的状态为可见输入数据向量。DBN 由若干结构单元堆栈组成,如图 2 所示,结构单元通常为 RBM。堆栈中每个 RBM 单元的可视层神经元数量等于前一 RBM 单元的隐层神经元数量。根据深度学习机制,采用输入样例训练第一层 RBM 单元,并利用其输出训练第二层 RBM 模型,将 RBM 模型进行堆栈通过增加层来改善模型性能。在无监督预训练过程中,DBN 编码输入到顶层 RBM 后解码顶层的状态到最底层的单元实现输入的重构。

RBM 的无向图模型如图 3 所示,作为 DBN 的结构单元,RBM 与每一层 DBN 共享参数。

RBM 是一种特殊形式的玻尔兹曼机(Boltzmann machine,BM),变量之间的图模型连接形式有限制,只有可见层节点与隐层节点之间有连接权值,而可见层节点与可见层节点及隐层节点与隐层节点之间无连接。BM 是基于能量的无向图概率模型,用输入 x 和隐变量 h 的能量函数定义联合概率分布为

$$p(x,h) = \frac{\mathrm{e}^{-\mathrm{energy}(x,h)}}{Z} \tag{3}$$

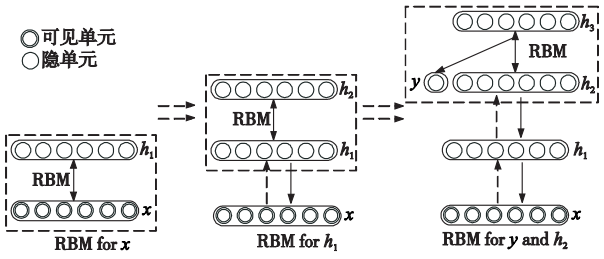

图2　DBN的生成过程

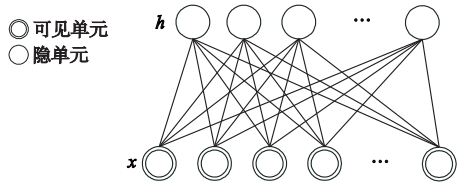

图3　RBM的无向图模型

将式(3)中的归一化常量 $Z = \sum_{x,h} \mathrm{e}^{-\mathrm{energy}(x,h)}$ 称为配分函数。可观察的输入 x 的边际概率分布为

$$p(x) = \sum_h p(x,h) = \sum_h \frac{\mathrm{e}^{-\mathrm{energy}(x,h)}}{Z} \tag{4}$$

引入自由能将式(4)变为

$$p(x) = \frac{\mathrm{e}^{-\mathrm{freeEnergy}(x)}}{Z} \tag{5}$$

式(5)中 $Z = \sum_x \mathrm{e}^{-\mathrm{freeEnergy}(x)}$,即

$$\mathrm{freeEnergy}(x) = -\log \sum_h \mathrm{e}^{-\mathrm{energy}(x,h)} \tag{6}$$

引入 θ 表示模型的参数,对式(5)取对数并求导可得

$$\frac{\partial \log p(x)}{\partial \theta} = -\frac{\partial \mathrm{freeEnergy}(x)}{\partial \theta} + \frac{1}{Z}\sum_{\tilde{x}} \mathrm{e}^{-\mathrm{freeEnergy}(\tilde{x})} \frac{\partial \mathrm{freeEnergy}(\tilde{x})}{\partial \theta} = -\frac{\partial \mathrm{freeEnergy}(x)}{\partial \theta} + \sum_{\tilde{x}} p(\tilde{x}) \frac{\partial \mathrm{freeEnergy}(\tilde{x})}{\partial \theta} \tag{7}$$

BM 的配分函数求解非常困难,因此用对数似然梯度 $\frac{\partial \log p(x)}{\partial \theta}$ 的近似值来训练[42],用服从数据分布的样例 $x \sim p(x)$ 和服从模型分布的样例 $\tilde{x} \sim p(\tilde{x})$ 上的自由能梯度定义模型参数更新规则

$$E_{\hat{p}}\left[\frac{\partial \log p(x)}{\partial \theta}\right] = -E_{\hat{p}}\left[\frac{\partial \mathrm{freeEnergy}(x)}{\partial \theta}\right] + E_p\left[\frac{\partial \mathrm{freeEnergy}(\tilde{x})}{\partial \theta}\right] \tag{8}$$

其中:$\hat{p}$ 是训练数据集的经验概率分布,p 是模型概率分布,$E_{\hat{p}}$和 E_p 是在相应概率分布下的期望值。式(8)中的第一项很容易计算得到,通常用子训练样本的均值近似代替;第二项包含从模型 p 中采样得到的样本,通常采用一些近似采样出的样本代替算法。可以用近似极大似然随机梯度下降算法训练 BM,通常用蒙特卡罗马尔可夫链(Monte-Carlo Markov chain,MCMC)方法来得到模型样例,更详细的描述内容参见文献[16,18]。

BM 的典型训练算法有变分近似法、随机近似法(stochastic approximation procedure,SAP)[43,44]、对比散度算法(contrastive divergence,CD)[45]、持续对比散度算法(persistent contrastive divergence,PCD)、快速持续对比散度算法(fast persistent contrastive divergence,FPCD)[46] 和回火 MCMC 算法[47] 等。

1.1.3 堆栈自编码网络模型

堆栈自编码网络的结构与 DBN 类似，由若干结构单元堆栈组成，不同之处在于其结构单元为自编码模型（auto-encoder）而不是 RBM。

自编码模型是一个两层的神经网络，第一层称为编码层，第二层称为解码层。如图 4 所示，训练该模型的目的是用编码器 $c(\cdot)$ 将输入 x 编码成表示 $c(x)$，再用解码器 $g(\cdot)$ 从 $c(x)$ 表示中解码重构输入 $r(x)=g(c(x))$。因此，自编码模型的输出是其输入本身，通过最小化重构误差 $L(r(x),x)$ 来执行训练。当隐层是线性的，并且 $L(r(x),x)=\|r(x)-x\|^2$ 是平方误差时，$c(x)$ 训练网络将输入投影到数据的主分量空间中，此时自编码模型的作用等效于 PCA；当隐层非线性时，与 PCA 不同，得到的表示可以堆栈成多层，自编码模型能够得到多模态输入分布[2,17,48]。重构误差的概率分布可以解释为非归一化对数概率密度函数这种特殊形式的能量函数[13]，意味着有低重构误差的样例对应的模型具有更高的概率。给定 $c(x)$，将均方差准则推广到最小化重构负对数似然函数的情况：

$$\mathrm{RE}=-\log p(x\mid c(x)) \tag{9}$$

能量函数中的稀疏项可用于有固定表示的情形[3,12]，并用于产生更强的保持几何变换不变性的特征。当输入 x_i 是二值或者二项概率时，损失函数为

$$-\log p(x\mid c(x))=-\sum_i x_i\log g_i(c(x))+(1-x_i)\log(1-g_i(c(x))) \tag{10}$$

$c(x)$ 并不是对所有 x 都具有最小损失的压缩表示，而是 x 的失真压缩表示，因此学习的目的是使编码 $c(x)$ 为输入的分布表示，可学习到数据中的主要因素，使其输出成为所有样例的有损压缩表示。

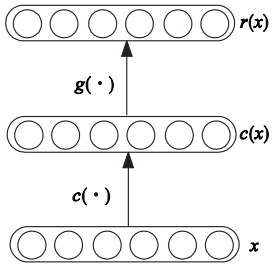

图4 自编码模型结构

在无约束的情况下，有 n 维输入并且编码维数至少为 n 的自编码模型只能学习恒等函数，没有编码效果。文献[2]中进行的实验表明，在实际情况下，用随机梯度下降方法训练目标函数，当隐单元个数多于输入数据个数时，非线性自编码模型能够产生有用的表示。自编码模型的重构误差的梯度与 RBM 的 CD 更新规则表达式存在对应关系。

堆栈自编码网络的结构单元除了上述的自编码模型之外，还可以使用自编码模型的一些变形，如降噪自编码模型和收缩自编码模型等。

降噪自编码模型避免了一般的自编码模型可能会学习得到无编码功能的恒等函数和需要样本的个数大于样本的维数的限制，尝试通过最小化降噪重构误差，从含随机噪声的数据中重构真实的原始输入[49]。降噪自编码模型使用由少量样本组成的微批次样本执行随机梯度下降算法，这样可以充分利用图处理单元（graphical processing unit，GPU）的矩阵到矩阵快速运算使得算法能够更快地收敛。文献[50]中说明降噪自编码模型与得分匹配方法直接相关。得分匹配是一种归纳原理，当所求解的问题易于处理时，可以用来代替极大似然求解过程。得分匹配的目标函数是模型得分和产生数据的真实概率密度的得分之间的平方差，其中模型得分是似然函数对输入的梯度 $\frac{\partial\log p(x)}{\partial x}$，而生成数据的概率密度是未知的。当把自编码模型的重构误差作为能量函数时，来自自编码模型的样例重构表示 $r(\tilde{x})$ 通常从低概率（高能量）结构输入到附近的高概率（低能量）结构，因此重构和输入之间的差值 $r(\tilde{x})-\tilde{x}$ 是概率即模型得分（model's score）的一个最大增长方向；另一方面，取一个训练样本 x 并将其随机加入噪声后记为 $\tilde{x}$，通常得到一个较低概率的近邻，即向量 $x-\tilde{x}$ 表示概率快速增长的方向。在重构损失函数选为平方误差的情况下，这两个差值 $r(\tilde{x})-\tilde{x}$ 和 $x-\tilde{x}$ 的平方差就是降噪重构误差 $(r(\tilde{x})-x)^2$。

收缩自编码模型的训练目标函数是重构误差和收缩罚项（contraction penalty）的总和，通过最小化该目标函数使已学习到的表示 $c(x)$ 尽量对输入 x 保持不变[51]。为了避免出现平凡解，编码器权值趋于零而解码器权值趋于无穷，并且收缩自编码模型采用固定的权值，令解码器权值为编码器权值的置换阵。与其他自编码模型相比，收缩自编码模型趋于找到尽量少的几个特征值，特征值的数量对应局部秩和局部维数。收缩自编码模型可以利用隐单元建立复杂非线性流形模型。

1.2 深度学习训练算法

实验结果表明，对深度结构神经网络采用随机初始化的方法，基于梯度的优化使训练结果陷入局部极值，而找不到全局最优值，并且随着网络结构层次的加深，更难以得到好的泛化性能，使得深度结构神经网络在随机初始化后得到的学习结果甚至不如只有一个或两个隐层的浅结构神经网络得到的学习结果好[52]。由于随机初始化深度结构神经网络的参数得到的训练结果和泛化性能都很不理想，在 2006 年以前，深度结构神经网络在机器学习领域文献中并没有进行过多讨论[1,2]。

通过实验研究发现，用无监督学习算法对深度结构神经网络进行逐层预训练，能够得到较好的学习结果。最初的实验对每层采用 RBM 生成模型，后来的实验采用自编码模型来训练每一层，两种模型得到相似的实验结果。一些实验和研究结果证明了无监督预训练相比随机初始化具有很大的优势，无监督预训练不仅初始化网络得到好的初始参数值，而且可以提取关于输入分布的有用信息，有助于网络找到更好的全局最优解[53,54]。对深度学习来说，无监督学习和半监督学习是成功的学习算法的关键组成部分，主要原因包括以下几个方面：

a）与半监督学习类似，深度学习中缺少有类标签的样本，并且样例大多无类标签。

b）逐层的无监督学习利用结构层上的可用信息进行学习，避免了监督学习梯度传播的问题，可减少对监督准则函数梯度给出的不可靠更新方向的依赖。

c）无监督学习使得监督学习的参数进入一个合适的预置区域内，在此区域内进行梯度下降能够得到很好的解。

d）在利用深度结构神经网络构造一个监督分类器时，无监督学习可看做学习先验信息，使得深度结构神经网络训练结果的参数在大多情况下都具有意义。

e）在深度结构神经网络的每一层采用无监督学习将一个问题分解成若干与多重表示水平提取有关的子问题，是一种常用

的可行方法,可提取输入分布较高水平表示的重要特征信息。

基于上述思想,Hinton 等人在 2006 年引入了 DBN 并给出了一种训练该网络的贪婪逐层预训练算法[1]。贪婪逐层无监督预训练学习的基本思想为:首先采用无监督学习算法对深度结构神经网络的较低层进行训练,生成第一层深度结构神经网络的初始参数值;然后将第一层的输出作为另外一层的输入,同样采用无监督学习算法对该层参数进行初始化。在对多层进行初始化后,用监督学习算法对整个深度结构神经网络进行微调,得到的学习性能具有很大程度的提高。

以堆栈自编码网络为例,深度结构神经网络的训练过程如下所示:

a)将第一层作为一个自编码模型,采用无监督训练,使原始输入的重建误差最小。

b)将自编码模型的隐单元输出作为另一层的输入。

c)按步骤 b)迭代初始化每一层的参数。

d)采用最后一个隐层的输出作为输入施加于一个有监督的层(通常为输出层),并初始化该层的参数。

e)根据监督准则调整深度结构神经网络的所有参数,堆栈所有自编码模型组成堆栈自编码网络。

基本的无监督学习方法在 2006 年被 Hinton 等人提出用于训练深度结构神经网络[1,2,12,55],该方法的学习步骤如下:

a)令 $h_0(x)=x$ 为可观察的原始输入 x 的最低阶表示。

b)对 $l=1,\cdots,L$,训练无监督学习模型,将可观察数据看做 $l-1$ 阶上表示的训练样例 $h_{l-1}(x)$,训练后产生下一阶的表示 $h_l(x)=R_l(h_{l-1}(x))$。

随后出现了一些该算法的变形拓展,最常见的是有监督的微调方法[1,2,12,56],该方法的学习步骤如下所示:

a)初始化监督预测器。

(a)用参数表示函数 $h_L(x)$。

(b)将 $h_L(x)$ 作为输入得到线性或非线性预测器。

b)基于已标记训练样本对 (x,y) 采用监督训练准则微调监督预测器,在表示阶段和预测器阶段优化参数。

2　深度学习研究的新进展

由于深度学习能够很好地解决一些复杂问题,近年来许多研究人员对其进行了深入研究,出现了许多有关深度学习研究的新进展。下面分别从初始化方法、网络层数和激活函数的选择、模型结构、学习算法和实际应用这四个方面对近几年深度学习研究的新进展进行介绍。

2.1　初始化方法、网络层数和激活函数的选择

研究人员试图搞清网络初始值的设定与学习结果之间的关系。Erhan 等人[14]在轨迹可视化研究中指出即使从相近的值开始训练深度结构神经网络,不同的初始值也会学习到不同的局部极值,同时发现用无监督预训练初始化模型的参数学习得到的极值与随机初始化学习得到的极值差异比较大,用无监督预训练初始化模型的参数学习得到的模型具有更好的泛化误差。Bengio 与 Krueger 等人[57,58]指出用特定的方法设定训练样例的初始分布和排列顺序可以产生更好的训练结果,用特定的方法初始化参数,使其与均匀采样得到的参数不同,会对梯度下降算法训练的结果产生很大的影响。Glorot 等人[15]指出通过设定一组初始权值使得每一层深度结构神经网络的 Jacobian 矩阵的奇异值接近 1,在很大程度上减小了监督深度结构神经网络和有预训练过程设定初值的深度结构神经网络之间的学习结果差异。另外,用于深度学习的学习算法通常包含许多超参数,文献[21]给出了这些超参数的选择指导性意见,推荐一些常用的超参数,尤其适用于基于反向传播的学习算法和基于梯度的优化算法中;并讨论了如何解决有许多可调超参数的问题,描述了实际用于有效训练常用的大型深度结构神经网络的超参数的影响因素,指出深度学习训练中存在的困难。

选择不同的网络隐层数和不同的非线性激活函数会对学习结果产生不同的影响。Glorot 等人[15,59]研究了隐层非线性映射关系的选择和网络的深度相互影响的问题,讨论了随机初始化的标准梯度下降算法用于深度结构神经网络学习得到不好的学习性能的原因。Glorot 等人观察不同非线性激活函数对学习结果的影响,得到逻辑斯蒂 S 型激活单元的均值会驱使顶层和隐层进入饱和,因而逻辑斯蒂 S 型激活单元不适合用随机初始化梯度算法学习深度结构神经网络;并据此提出了标准梯度下降算法的一种新的初始化方案来得到更快的收敛速度,为理解深度结构神经网络使用和不使用无监督预训练的性能差异作出了新的贡献。Bengio 等人[60,61]从理论上说明深度学习结构的表示能力随着神经网络深度的增加以指数的形式增加,但是这种增加的额外表示能力会引起相应局部极值数量的增加,使得在其中寻找最优值变得困难。

2.2　模型结构

1)DBN 的结构及其变种　采用二值可见单元和隐单元 RBM 作为结构单元的 DBN,在 MNIST 等数据集上表现出很好的性能。近几年,具有连续值单元的 RBM,如 mcRBM[28]、mPoT 模型[62]和 spike-and-slab RBM[63]等已经成功应用。Spike-and-slab RBM 中 spike 表示以 0 为中心的离散概率分布,slab 表示在连续域上的稠密均匀分布,可以用吉布斯采样对 spike-and-slab RBM 进行有效推断,得到优越的学习性能。

2)和—积网络　深度学习最主要的困难是配分函数的学习,如何选择深度结构神经网络的结构使得配分函数更容易计算?Poon 等人[64]提出一种新的深度模型结构——和—积网络(sum-product network,SPN),引入多层隐单元表示配分函数,使得配分函数更容易计算。SPN 是有根节点的有向无环图,图中的叶节点为变量,中间节点执行和运算与积运算,连接节点的边带有权值,它们在 Caltech-101 和 Olivetti 两个数据集上进行实验证明了 SPN 的性能优于 DBN 和最近邻方法。

3)基于 rectified 单元的学习　Glorot 与 Mesnil 等人[59,65,66]用降噪自编码模型来处理高维输入数据。与通常的 S 型和正切非线性隐单元相比,该自编码模型使用 rectified 单元,使隐单元产生更加稀疏的表示。在此之前,文献[67]已经对随机 rectified 单元进行了介绍;对于高维稀疏数据,Dauphin 等人[68]采用抽样重构算法,训练过程只需要计算随机选择的很小的样本子集的重构和重构误差,在很大程度上提高了学习速度,实验结果显示提速了 20 倍。Glorot 等人[59]提出在深度结构神经网络中,在图像分类和情感分类问题中用 rectified 非线性神经元代替双曲正切或 S 型神经元,指出 rectified 神经元网络在零点产生与双曲正切神经元网络相当或者有更好的性能,能够产生有真正零点的稀疏表示,非常适合本质稀疏数据的建模,在理解训练纯粹深度监督神经网络的困难,搞清使用或不使用无

监督预训练学习的神经网络造成的性能差异方面，可以看做新的里程碑；Glorot 等人还提出用增加 L1 正则化项来促进模型稀疏性，使用无穷大的激活函数防止算法运行过程中可能引起的数值问题。在此之前，Nair 等人[67] 提出在 RBM 环境中 rectified 神经元产生的效果比逻辑斯蒂 S 型激活单元好，他们用无限数量的权值相同但是负偏差变大的一组单元替换二值单元，生成用于 RBM 的更好的一类隐单元，将 RBM 泛化，可以用噪声 rectified 线性单元（rectified linear units）有效近似这些 S 型单元。用这些单元组成的 RBM 在 NORB 数据集上进行目标识别以及在数据集上进行已标记人脸实际验证，得到比二值单元更好的性能，并且可以更好地解决大规模像素强度值变化很大的问题。

4）卷积神经网络　文献 [9] 研究了用生成式子抽样单元组成的卷积神经网络，在 MNIST 数字识别任务和 Caltech-101 目标分类基准任务上进行实验，显示出非常好的学习性能。Huang 等人[68] 提出一种新的卷积学习模型——局部卷积 RBM，利用对象类中的总体结构学习特征，不假定图像具有平稳特征，在实际人脸数据集上进行实验，得到性能很好的实验结果。

2.3　学习算法

1）深度费希尔映射方法　Wong 等人[70] 提出一种新的特征提取方法——正则化深度费希尔映射（regularized deep Fisher mapping，RDFM）方法，学习从样本空间到特征空间的显式映射，根据 Fisher 准则用深度结构神经网络提高特征的区分度。深度结构神经网络具有深度非局部学习结构，从更少的样本中学习变化很大的数据集中的特征，显示出比核方法更强的特征识别能力，同时 RDFM 方法的学习过程由于引入正则化因子，解决了学习能力过强带来的过拟合问题。在各种类型的数据集上进行实验，得到的结果说明了在深度学习微调阶段运用无监督正则化的必要性。

2）非线性变换方法　Raiko 等人[71] 提出了一种非线性变换方法，该变换方法使得多层感知器（multi-layer perceptron，MLP）网络的每个隐神经元的输出具有零输出和平均值上的零斜率，使学习 MLP 变得更容易。将学习整个输入输出映射函数的线性部分和非线性部分尽可能分开，用 shortcut 权值（shortcut weight）建立线性映射模型，令 Fisher 信息阵接近对角阵，使得标准梯度接近自然梯度。通过实验证明非线性变换方法的有效性，该变换使得基本随机梯度学习与当前的学习算法在速度上不相上下，并有助于找到泛化性能更好的分类器。用这种非线性变换方法实现的深度无监督自编码模型进行图像分类和学习图像的低维表示的实验，说明这些变换有助于学习深度至少达到五个隐层的深度结构神经网络，证明了变换的有效性，提高了基本随机梯度学习算法的速度，有助于找到泛化性更好的分类器。

3）稀疏编码对称机算法　Ranzato 等人[13] 提出一种新的有效的无监督学习算法——稀疏编码对称机（sparse encoding symmetric machine，SESM），能够在无须归一化的情况下有效产生稀疏表示。SESM 的损失函数是重构误差和稀疏罚函数的加权总和，基于该损失函数比较和选择不同的无监督学习机，提出一种与文献 [12] 算法相关的迭代在线学习算法，并在理论和实验上将 SESM 与 RBM 和 PCA 进行比较，在手写体数字识别 MNIST 数据集和实际图像数据集上进行实验，表明该方法的优越性。

4）迁移学习算法　在许多常见学习场景中训练和测试数据集中的类标签不同，必须保证训练和测试数据集中的相似性进行迁移学习。Mesnil 等人[65] 研究了用于无监督迁移学习场景中学习表示的不同种类模型结构，将多个不同结构的层堆栈使用无监督学习算法用于五个学习任务，并研究了用于少量已标记训练样本的简单线性分类器堆栈深度结构学习算法。Bengio[60] 研究了无监督迁移学习问题，讨论了无监督预训练有用的原因，如何在迁移学习场景中利用无监督预训练，以及在什么情况下需要注意从不同数据分布得到的样例上的预测问题。

5）自然语言解析算法　Collobert[72] 基于深度递归卷积图变换网络（graph transformer network，GTN）提出一种快速可扩展的判别算法用于自然语言解析，将文法解析树分解到堆栈层中，只用极少的基本文本特征，得到的性能与现有的判别解析器和标准解析器的性能相似，而在速度上有了很大提升。

6）学习率自适应方法　学习率自适应方法可用于提高深度结构神经网络训练的收敛性并且去除超参数中的学习率参数，其中包括全局学习率[73]、层次学习率、神经元学习率和参数学习率[74] 等。最近研究人员提出了一些新的学习率自适应方法，如 Duchi 等人[75] 提出的自适应梯度方法和 Schaul 等人[76] 提出的学习率自适应方法；Hinton[18] 提出了收缩学习率方法使得平均权值更新在权值大小的 1/1000 数量级上；Le Roux 等人[77,78] 提出自然梯度的对角低秩在线近似方法，并说明该算法在一些学习场景中能加速训练过程。

2.4　实际应用

2.4.1　语音和音频

Yu 等人在文献 [79] 中介绍了深度学习的基本概念、DBN 等常用的深度学习模型以及流行且有效的深度学习算法，包括 RBM 和基于降噪自编码模型的预训练方法，并指出在许多信号处理应用中，特别是对语音和音频信号处理，深度学习技术有好的学习结果。通过综合深度学习模型强大的判别训练和连续建模能力，深度学习已成功应用于大规模词汇连续语音识别任务。卷积 DBN 和堆栈自编码网络等深度结构神经网络已经被用于语音和音频数据处理中，如音乐艺术家流派分类、说话者识别、说话者性别分类和语音分类等，得到非常好的学习结果。堆栈多层条件随机场（conditional random field，CRF）等其他深度结构神经网络结构模型也成功用于语言识别、语音识别、序列标记[80] 和置信度校准等语音相关任务。Lee 等人[10] 首次用无监督卷积神经网络方法将 DBN 用于声学信号处理，说明该方法在讲话者、性格和音素检测上表现出比梅尔倒谱系数（Mel frequency cepstrum coefficient，MFCC）更优越的性能。Hamel 等人[81] 将 DBN 用于音乐类型识别和自动标记问题，将原始级光谱作为 DBN 的输入，用贪婪预训练和监督微调方法进行训练，得到的分类精度比 MFCC 有很大改进。Schmidt 等人[82] 用基于回归的 DBN 直接从光谱中学习特征，将系统应用于特定的音乐情感识别问题，并且该系统也可以应用于任何基于回归的音频特征学习问题。Deng 等人[83] 将堆栈自编码网络用于语音特征编码问题，以最小的重构误差将数据压缩到预先设定长度的表示。

2.4.2　图像和视频

1）手写体字符识别　Bengio 等人[84] 运用统计学习理论和

大量的实验工作证明了深度学习算法非常具有潜力,说明数据中间层表示可以被来自不同分布而相关的任务和样例共享,产生更好的学习效果,并且在有 62 个类别的大规模手写体字符识别场景上进行实验,用多任务场景和扰动样例来得到分布外样例,并得到非常好的实验结果。Lee 等人[11] 对 RBM 进行拓展,学习到的模型使其具有稀疏性,可用于有效地学习数字字符和自然图像特征。Hinton 等人关于深度学习的研究说明了如何训练深度 S 型神经网络来产生对手写体数字文本有用的表示,用到的主要思想是贪婪逐层预训练 RBM 之后再进行微调[22,55]。

2)人脸识别　Nair 等人[67] 用噪声 rectified 线性单元组成的深度结构神经网络将深度学习应用于目标识别和人脸验证;Ranzato 等人[83] 提出深度产生式模型用于人脸识别;Susskind 等人[84] 将因式分解的三路 RBM 用于建立成对人脸图像的模型。Luo 等人[87] 研究如何从局部遮挡的人脸图像解析面部成分,提出一种新的人脸解析器,将人脸成分分割重构为重叠的形态数据过程,首先在块等级和组等级上检测人脸,在 DBN 上执行产生式训练过程,再用逻辑斯蒂回归进行判别式调整,然后计算对像素敏感的标记映射。从 LFW、BioID 和 CUFSF 三个数据集中挑选 2 239 个图像进行实验,说明了该方法的有效性,该方法不仅对局部遮挡的人脸图像具有鲁棒性,而且也为人脸分析和人脸合成提供了更丰富的信息。

3)图像识别和检索　DBN 和堆栈自编码网络在单个图像识别任务中表现出很好的性能,成功用于生成紧凑而有意义的图像检索表示形式,并且已用于大型图像检索任务中,得到非常好的结果[11]。图像识别方面比 DBN 更一般的方法在文献[88] 中有所描述。Taylor 等人[89] 将条件 DBN 用于视频排序和人类动作合成,条件 DBN 使得 DBN 的权值与之前的数据相关联,可以提高训练的有效性。Lee 和 Raina 等人[9,90] 用稀疏编码和 DBN 从自然图像中学习有效特征表示。Nair 等人[91] 提出改进的 DBN,该模型的顶层模型用三阶 BM,他们将这种模型用于三维目标识别任务 NORB 数据集上,实验结果显示出训练得到了很低的预测误差率。Tang 等人[92] 提出两种策略来提高 DBN 的鲁棒性,首先将 DBN 的第一层具有稀疏连接结构引入正则化方法,接着提出一种概率降噪算法,这些技术在高噪声图像识别任务和随机噪声的鲁棒性方面显示出其有效性。Lee 等人[93] 提出一种深度学习方法使脑图像分割自动化,用卷积神经网络建立用于脑图像分割的判别特征,能自动从人类专家提供的类标签中进行学习,通过实验验证该方法在自动多类脑图像分割方面显示出优越的性能,表明该方法可以替代已有的模板图像分割方法,减少了图像分割过程对人类专家的干预和对先验信息的需求。

2.4.3　语言处理和信息检索

文献[94] 阐述了深度学习用于自然语言处理的基本动机、思路、模型和学习算法,提出基本的神经网络模型和训练算法,并指出这些方法在语言模型、POS 标记、命名实体识别和情感分析等任务中表现出很好的性能。深度学习已经被成功应用于文本、图像和音频等单模态无监督特征学习。Salakhutdinov 与 Lecun 等人[22,88] 将 DBN 和堆栈自编码网络用于对文档建立索引以便检索;Deng 等人[83] 也将该想法应用于音频文档检索问题中;Collobert 等人[23] 提出用卷积 DBN 作为模型同时解决许多经典问题,如词性标记、名实体标记、语义角色识别和相似词识别等;Deselaers 等人[95] 将 DBN 用于解决机器音译问题。Ngiam 等人[96] 提出深度学习的一种新应用——在多模态上学习特征,提出一系列学习任务,说明如何训练深度结构神经网络,证明在特征学习时使用多个模态有助于学习得到更好的特征,并说明如何学习模态之间的共享表示,在视听语音分类 CUAVE 和 AVLetters 数据集上对提出的方法进行了验证。在线评价和舆情分析的需求使得情感分类问题成为热点研究问题,Glorot 等人[66] 将深度学习方法用于域自适应情感分类器设计问题,用基于有稀疏 rectified 单元的堆栈降噪自编码网络从无监督的在线评价和建议中提取有意义的特征表示,在四种亚马逊产品的评价数据上进行实验,结果说明用高阶特征表示训练的情感分类器的学习性能明显优于当前的其他方法;另外该方法允许在 22 个领域的更大的工业级数据集上成功执行域自适应学习方法,在很大程度上提高了分类器的泛化性能。

3　未来研究方向

经过近几十年来大量研究人员对人工神经网络的理论和实验研究,深度学习领域的研究取得了一定进展,实验结果表明了其良好的学习性能。但是目前深度学习领域的研究仍然存在许多有待进一步解决的问题,未来深度学习的研究在理论分析、数据表示与模型、特征提取、训练与优化求解以及研究拓展这五个方面需要进一步研究。

3.1　理论分析

需要更好地理解深度学习及其模型,进行更加深入的理论研究。深度学习模型的训练为什么那么困难?这仍然是一个开放性问题。一个可能的答案是深度结构神经网络有许多层,每一层由多个非线性神经元组成,使得整个深度结构神经网络的非线性程度更强,减弱了基于梯度的寻优方法的有效性;另一个可能的答案是局部极值的数量和结构随着深度结构神经网络深度的增加而发生定性改变,使得训练模型变得更加困难。造成深度学习训练困难的原因究竟是由于用于深度学习模型的监督训练准则大量存在不好的局部极值,还是因为训练准则对优化算法来说过于复杂,这是值得探讨的问题。此外,对堆栈自编码网络学习中的模型是否有合适的概率解释,能否得到深度学习模型中似然函数梯度的小方差和低偏差估计,能否同时训练所有的深度结构神经网络层,除了重构误差外,是否还存在其他更合适的可供选择的误差指标来控制深度结构神经网络的训练过程,是否存在容易求解的 RBM 配分函数的近似函数,这些问题还有待未来研究。可以参考文献[97,98],考虑引入退火重要性抽样来解决局部极值问题,不依赖于配分函数的学习算法也值得尝试。

3.2　数据表示与模型

数据的表示方式对学习性能具有很大的影响,除了局部表示、分布表示和稀疏分布表示外,可以充分利用表示理论研究成果。是否还存在其他形式的数据表示方式,是否可以通过在学习的表示上施加一些形式的稀疏罚从而对 RBM 和自编码模型的训练性能起到改进作用,以及如何改进,这方面可以参考文献[13,99~101] 中的内容。是否可以用便于提取好的表示并且包含更简单优化问题的凸模型代替 RBM 和自编码模型;不增加隐单元的数量,用非参数形式的能量函数能否提高

RBM 的容量等，未来还需要进一步探讨这些问题。此外，除了卷积神经网络、DBN 和堆栈自编码网络之外，是否还存在其他可以用于有效训练的深度学习模型，有没有可能改变所用的概率模型使训练变得更容易，是否存在其他有效的或者理论上有效的方法学习深度学习模型，这也是未来需要进一步研究的问题。现有的方法，如 DBN-HMM 和 DBN-CRF，在利用 DBN 的能力方面只是简单的堆栈叠加基本模型，还没有充分发掘出 DBN 的优势，需要研究 DBN 的结构特点，充分利用 DBN 的潜在优势，找到更好的方法建立数据的深度学习模型，可以考虑将现有的社会网络、基因调控网络、结构化建模理论以及稀疏化建模等理论运用其中。

3.3 特征提取

除了高斯—伯努利模型之外，还有哪些模型能用来从特征中提取重要的判别信息，未来需要提出有效的理论指导在每层搜索更加合适的特征提取模型。自编码模型保持了输入的信息，这些信息在后续的训练过程中可能会起到重要作用，未来需要研究用 CD 训练的 RBM 是否保持了输入的信息，在没有保持输入信息的情况下如何进行修正。树和图等结构的数据由于大小和结构可变而不容易用向量表示其中包含的信息，如何泛化深度学习模型来表示这些信息，也是未来需要研究的问题。尽管当前的产生式预训练加判别式微调学习策略看起来对许多任务都运行良好，但是在某些语言识别等其他任务中却失败了，对这些任务，产生式预训练阶段的特征提取似乎能很好地描述语音变化，但是包含的信息不足以区分不同的语言，未来需要提出新的学习策略，对这些学习任务提取合适的特征，这可以在很大程度上减小当前深度学习系统所需模型的大小。

3.4 训练与优化求解

为什么随机初始化的深度结构神经网络采用基于梯度的算法训练总是不能成功，产生式预训练方法为什么有效？未来需要研究训练深度结构神经网络的贪婪逐层预训练算法到底在最小化训练数据的似然函数方面结果如何，是否过于贪婪，以及除了贪婪逐层预训练的许多变形和半监督嵌入算法[102]之外，还有什么其他形式的算法能得到深度结构神经网络的局部训练信息。此外，无监督逐层训练过程对训练深度学习模型起到帮助作用，但有实验表明训练仍会陷入局部极值并且无法有效利用数据集中的所有信息，能否提出用于深度学习的更有效的优化策略来突破这种限制，基于连续优化的策略能否用于有效改进深度学习的训练过程，这些问题还需要继续研究。二阶梯度方法和自然梯度方法在理论研究中可证明对训练求解深度学习模型有效，但是这些算法还不是深度结构神经网络优化的标准算法，未来还需要进一步验证和改进这些算法，研究其能否代替微批次随机梯度下降类算法。当前的基于微批次随机梯度优化算法难以在计算机上并行处理，目前最好的解决方法是用 GPU 来加速学习过程，但是单个机器的 GPU 无法用于处理大规模语音识别和类似的大型数据集的学习，因此未来需要提出理论上可行的并行学习算法来训练深度学习模型。

3.5 研究拓展

当深度模型没有有效的自适应技术，在测试数据集分布不同于训练集分布时，它们很难得到比常用模型更好的性能，因此未来有必要提出用于深度学习模型的自适应技术以及对高维数据具有更强鲁棒性的更先进的算法。文献 [102] 中指出，目前的深度学习模型训练算法包含许多阶段，而在在线学习场景中一旦进入微调阶段就有可能陷入局部极值，因此目前的算法对于在线学习环境是不可行的。未来需要研究是否存在训练深度学习的完全在线学习过程能够一直具有无监督学习成分。DBN 模型很适合半监督学习场景和自教学习场景，当前的深度学习算法如何应用于这些场景并且在性能上优于现有的半监督学习算法，如何结合监督和无监督准则来学习输入的模型表示，是否存在一个深度使得深度学习模型的计算足够接近人类在人工智能任务中表现出的水平，这也是未来需要进一步研究的问题。

4 结束语

深度学习作为机器学习的一个研究领域在近几年受到了越来越多的关注，许多学者对深度学习进行了广泛的研究。本文详细概述了深度学习相对于浅学习的优点，说明了引入深度学习的必要性，描述了深度学习的数据表示形式以及卷积神经网络、DBN 和堆栈自编码网络这几种典型的深度学习模型，对可能造成深度学习训练困难的原因进行了说明，并介绍了有效的训练方法，从初始化方法、网络层数和激活函数的选择、模型结构、学习算法和实际应用这四个方面对近几年深度学习研究的新进展进行了综述，并从理论分析、数据表示与模型、特征提取、训练与优化求解和研究拓展这五个方面指出了深度学习中有待进一步解决的问题。可以预见，随着深度学习理论和方法研究的深入，深度学习将被更加广泛地应用在各个领域。

参考文献：

[1] HINTON G E, OSINDERO S, TEH Y W. A fast learning algorithm for deep belief nets [J]. Neural Computation, 2006, 18(7): 1527-1554.

[2] BENGIO Y, LAMBLIN P, POPOVICI D, *et al*. Greedy layer-wise training of deep networks [C] //Advances in Neural Information Processing Systems. Cambridge: MIT Press, 2007: 153-160.

[3] VINCENT P, LAROCHELLE H, BENBIO Y, *et al*. Extracting and composing robust features with denoising autoencoders [C] //Proc of the 25th International Conference on Machine Learning. New York: ACM Press, 2008: 1096-1103.

[4] LAROCHELLE H, BENGIO Y, LOURADOUR J, *et al*. Exploring strategies for training deep neural networks [J]. Journal of Machine Learning Research, 2009, 10(12): 1-40.

[5] TAYLOR G, HINTON G E. Factored conditional restricted Boltzmann machines for modeling motion style [C] //Proc of the 26th Annual International Conference on Machine Learning. New York: ACM Press, 2009: 1025-1032.

[6] SALAKHUTDINOV R, HINTON G E. Deep Boltzmann machines [C] //Proc of the 12th International Conference on Artificial Intelligence and Statistics. 2009: 448-455.

[7] TAYLOR G, SIGAL L, FLEET D J, *et al*. Dynamical binary latent variable models for 3D human pose tracking [C] //Proc of IEEE Conference on Computer Vision and Pattern Recognition. 2010: 631-638.

[8] JARRETT K, KAVUKCUOGLU K, RANZATO M, *et al*. What is the best multi-stage architecture for object recognition? [C] //Proc of the 12th International Conference on Computer Vision. 2009: 2146-2153.

[9] LEE H, GROSSE R, RANGANATH R, *et al*. Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations [C] //Proc of the 26th International Conference on Machine Learning. New York: ACM Press, 2009: 609-616.

[10] LEE H, PHAM P, LARGMAN Y, *et al*. Unsupervised feature learning for audio classification using convolutional deep belief networks [C] //Advances in Neural Information Processing Systems. Cambridge: MIT Press, 2009: 1096–1104.

[11] LEE H, EKANADHAM C, NG A Y. Sparse deep belief net model for visual area V2 [C] //Advances in Neural Information Processing Systems. Cambridge: MIT Press, 2008: 873–880.

[12] RANZATO M, POULTNEY C, CHOPRA S, *et al*. Efficient learning of sparse representations with an energy-based model [C] //Advances in Neural Information Processing Systems. Cambridge: MIT Press, 2007: 1137–1144.

[13] RANZATO M, BOUREAU Y L, LeCUN Y. Sparse feature learning for deep belief networks [C] //Advances in Neural Information Processing Systems. Cambridge: MIT Press, 2008: 1185–1192.

[14] ERHAN D, BENGIO Y, COURVILLE A, *et al*. Why does unsupervised pre-training help deep learning? [J]. Journal of Machine Learning Research, 2010, 11(2): 625–660.

[15] GLOROT X, BENGIO Y. Understanding the difficulty of training deep feedforward neural networks [C] //Proc of the 13th International Conference on Artifical Intelligence and Statistics. 2010: 249–256.

[16] BENGIO Y. Lèarning deep architectures for AI [J]. Foundations and Trends in Machine Learning, 2009, 2(1): 1–127.

[17] GOODFELLOW I, LE Q, SAXE A, *et al*. Measuring invariances in deep networks [C] //Advances in Neural Information Processing Systems. Cambridge: MIT Press, 2009: 646–654.

[18] HINTON G E. A practical guide to training restricted Boltzmann machines [M] //Neural Networks: Tricks of the Trade. Berlin: Springer-Verlag, 2013: 599–619.

[19] MONTAVON G, ORR G B, MULLER K R. Neural networks: tricks of the trade [M]. Berlin: Springer-Verlag, 1998.

[20] LeCUN Y, BOTTOU L, ORR G B, *et al*. Efficient backProp [M] //Neural Networks: Tricks of the Trade. Berlin: Springer-Verlag, 1998: 9–50.

[21] BENGIO Y. Practical recommendations for gradient-based training of deep architectures [M] //Neural Networks: Tricks of the Trade. Berlin: Springer-Verlag, 2012: 437–478.

[22] SALAKHUTDINOV R, HINTON G E. Semantic hashing [J]. International Journal of Approximate Reasoning, 2009, 50(7): 969–978.

[23] COLLOBERT R, WESTON J. A unified architecture for natural language processing: deep neural networks with multitask learning [C] //Proc of the 25th Internationnal Conference on Machine Learning. New York: ACM Press, 2008: 160–167.

[24] RANZATO M, SZUMMER M. Semi-supervised learning of compact document representations with deep networks [C] //Proc of the 25th International Conference on Machine Learning. New York: ACM Press, 2008: 792–799.

[25] RANZATO M, HUANG Fu-jie, BOUREAU Y L, *et al*. Unsupervised learning of invariant feature hierarchies with applications to object recognition [C] //Proc of International Conference on Computer Vision and Pattern Recognition. 2007: 1–8.

[26] ZEILER M, KRISHNAN D, TAYLOR G, *et al*. Deconvolutional networks [C] //Proc of International Conference on Computer Vision and Pattern Recognition. 2010.

[27] YANG Jian-chao, YU Kai, GONG Yi-hong, *et al*. Linear spatial pyramid matching using sparse coding for image classification [C] //Proc of International Conference on Computer Vision and Pattern Recognition. 2009.

[28] RANZATO M, HINTON G E. Modeling pixel means and covariances using factorized third-order Boltzmann machines [C] //Proc of IEEE Conference on Computer Vision and Pattern Recognition. 2010: 2551–2558.

[29] ERHAN D, COURVILLE A, BENGIO Y. Understanding representations learned in deep architectures, TR 1355 [R/OL]. (2010-10-19). http://www.dumitru.ca/files/publications/invarianes_techreport.pdf.

[30] PINTO N, DOUKHAN D, DICARLO J J, *et al*. A high-throughput screening approach to discovering good forms of biologically inspired visual representation [J]. PLoS Computational Biology, 2009, 5(11): e1000579.

[31] BENGIO Y, DUCHARME R, VINCENT P. A neural probabilistic language model [C] //Advances in Neural Information Processing Systems. Cambridge: MIT Press, 2001: 933–938.

[32] HINTON G E. Learning distributed representations of concepts [C] //Proc of the 8th Annual Conference of Cognitive Science Society. 1986: 1–12.

[33] BAGNELL J A, BRADLEY D M. Differentiable sparse coding [C] //Advances in Neural Information Processing Systems. Cambridge: MIT Press, 2009: 113–120.

[34] COATES A, NG A Y. The importance of encoding versus training with sparse coding and vector quantization [C] //Proc of the 28th International Conference on Machine Learning. New York: ACM Press, 2011: 921–928.

[35] GOODFELLOW I, COURVILLE A, BENGIO Y. Spike-and-slab sparse coding for unsupervised feature discovery [C] //Proc of NIPS Workshop on Challenges in Learning Hierarchical Models. 2011.

[36] BELL A J, SEJNOWSKI T J. An information maximisation approach to blind separation and blind deconvolution [J]. Neural Computation, 1995, 7(6): 1129–1159.

[37] HOTELLING H. Analysis of a complex of statistical variables into principal components [J]. Journal of Educational Psychology, 1933, 24(6): 417–441, 498–520.

[38] FUKUSHIMA K. Neocognitron: a self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position [J]. Biological Cybernetics, 1980, 36(4): 193–202.

[39] LeCUN Y, BOTTOU L, BENGIO Y, *et al*. Gradient-based learning applied to document recognition [J]. Proceedings of the IEEE, 1998, 86(11): 2278–2324.

[40] LeCUN Y, BOSER B, DENKER J S, *et al*. Backpropagation applied to handwritten zip code recognition [J]. Neural Computation, 1989, 1(4): 541–551.

[41] LE Q, NGIAM J, CHEN Zheng-hao, *et al*. Tiled convolutional neural networks [C] //Advances in Neural Information Processing Systems. Cambridge: MIT Press, 2010.

[42] BENGIO Y, DELALLEAU O. Justifying and generalizing contrastive divergence [J]. Neural Computation, 2009, 21(6): 1601–1621.

[43] NEAL R M. Connectionist learning of belief networks [J]. Artificial Intelligence, 1992, 56(1): 71–113.

[44] TIELEMAN T. Training restricted Boltzmann machines using approximations to the likelihood gradien [C] //Proc of the 25th International Conference on Machine Learning. New York: ACM Press, 2008: 1064–1071.

[45] HINTON G E. Training products of experts by minimizing contrastive divergence [J]. Neural Computation, 2002, 14(8): 1771–1800.

[46] TIELEMAN T, HINTON G E. Using fast weights to improve persistent contrastive divergence [C] //Proc of the 26th Annual International Conference on Machine Learning. New York: ACM Press, 2009: 1033–1040.

[47] DESJARDINS G, COURVILLE A, BENGIO Y, *et al*. Parallel tempering for training of restricted Boltzmann machines [C] //Proc of the 13th International Conference on Artificial Intelligence and Statistics. 2010:145-152.

[48] BENGIO Y, ALAIN G, RIFAI S. Implicit density estimation by local moment matching to sample from auto-encoders [R/OL]. (2012-06-30). http://arxiv.org/abs/1207.0057.

[49] VINCENT P, LAROCHELLE H, LAJOIE I, *et al*. Stacked denoising autoencoders: learning useful representations in a deep network with a local denoising criterion [J]. Journal of Machine Learning Research, 2010, 11(3): 3371-3408.

[50] VINCENT P. A connection between score matching and denoising autoencoders [J]. Neural Computation, 2011, 23(7): 1661-1674.

[51] RIFAI S, VINCENT P, MULLER X, *et al*. Contracting auto-encoders: explicit invariance during feature extraction [C] //Proc of the 28th International Conference on Machine Learning. 2011.

[52] BOTTOU L. Large-scale machine learning with stochastic gradient descent [C] //Proc of the 19th International Conferece on Computational Statistics. Berlin: Springer-Verlag, 2010: 177-186.

[53] SAXE A M, KOH P W, CHEN Z, *et al*. On random weights and unsupervised feature learning [C] //Proc of the 28th International Conference on Machine Learning. New York: ACM Press, 2011: 1089-1096.

[54] SWERSKY K, CHEN Bo, MARLIN B, *et al*. A tutorial on stochastic approximation algorithms for training restricted Boltzmann machines and deep belief nets [C] //Proc of Information Theory and Applications Workshop. 2010: 1-10.

[55] HINTON G E, SALAKHUTDINOV R. Reducing the dimensionality of data with neural networks [J]. Science, 2006, 313(5786): 504-507.

[56] LAMBLIN P, BENGIO Y. Important gains from supervised fine-tuning of deep architectures on large labeled sets [C] //Proc of NIPS Deep Learning and Unsupervised Feature Learning Workshop. 2010.

[57] BENGIO Y, LOURADOUR J, COLLOBERT R, *et al*. Curriculum learning [C] //Proc of the 26th International Conference on Machine Learning. New York: ACM Press, 2009: 41-48.

[58] KRUEGER K A, DAYAN P. Flexible shaping: how learning in small steps helps [J]. Cognition, 2009, 110(3): 380-394.

[59] GLOROT X, BORDES A, BENGIO Y. Deep sparse rectifier neural networks [C] //Proc of the 14th International Conference on Artificial Intelligence and Statistics. 2011: 315-323.

[60] BENGIO Y. Deep learning of representations for unsupervised and transfer learning [C] //Proc of Workshop on Unsupervised and Transfer Learning. 2011: 17-37.

[61] BENGIO Y, DELALLEAU O. On the expressive power of deep architectures [C] //Proc of the 14th International Conference on Discovery Science. Berlin: Springer-Verlag, 2011: 1.

[62] RANZATO M, MNIH V, HINTON G E. Generating more realistic images using gated MRF's [C] //Advances in Neural Information Processing Systems. Cambridge: MIT Press, 2010: 2002-2010.

[63] COURVILLE A, BERGSTRA J, BENGIO Y. Unsupervised models of images by spike-and-slab RBMs [C] //Proc of the 28th International Conference on Machine Learning. 2011.

[64] POON H, DOMINGOS P. Sum-product networks: a new deep architecture [C] //Proc of IEEE International Conference on Computer Vision Workshops. 2011: 689-690.

[65] MESNIL G, DAUPHIN Y, GLOROT X, *et al*. Unsupervised and transfer learning challenge: a deep learning approach [C] //Proc of Workshop on Unsupervised and Transfer Learning. 2011: 1-15.

[66] GLOROT X, BORDES A, BENGIO Y. Domain adaptation for large-scale sentiment classification: a deep learning approach [C] //Proc of the 28th International Conference on Machine Learning. 2011: 513-520.

[67] NAIR V, HINTON G E. Rectified linear units improve restricted Boltzmann machines [C] //Proc of the 27 th International Conference on Machine Learning. 2010: 807-814.

[68] DAUPHIN Y, GLOROT X, BENGIO Y. Sampled reconstruction for large-scale learning of embeddings [C] //Proc of the 28th International Conference on Machine Learning. 2011: 945-952.

[69] HUANG G B, LEE H, LEARNED-MILLER E. Learning hierarchical representations for face verification with convolutional deep belief networks [C] //Proc of IEEE Conference on Computer Vision and Pattern Recognition. 2012: 2518-2525.

[70] WONG W K, SUN M. Deep learning regularized Fisher mappings [J]. IEEE Trans on Neural Networks, 2011, 22(10): 1668-1675.

[71] RAIKO T, VALPOLA H, LeCUN Y. Deep learning made easier by linear transformations in perceptrons [C] //Proc of the 15th International Conference on Artificial Intelligence and Statistics. 2012: 924-932.

[72] COLLOBERT R. Deep learning for efficient discriminative parsing [C] //Proc of the 14th International Conference on Artificial Intelligence and Statistics. 2011: 224-232.

[73] CHO K, RAIKO T, ILIN A. Enhanced gradient and adaptive learning rate for training restricted Boltzmann machines [C] //Proc of the 28th International Conference on Machine Learning. 2011: 105-112.

[74] BORDES A, BOTTOU L, GALLINARI P. SGD-QN: careful quasi-Newton stochastic gradient descent [J]. Journal of Machine Learning Research, 2009, 10(12): 1737-1754.

[75] DUCHI J, HAZAN E, SINGER Y. Adaptive subgradient methods for online learning and stochastic optimization [J]. Journal of Machine Learning Research, 2011, 12(2): 2121-2159.

[76] SCHAUL T, ZHANG S, LeCUN Y. No more pesky learning rates [C] //Proc of International Conference on Machine Learning. 2013: 343-351.

[77] Le ROUX N, MANZAGOL P A, BENGIO Y. Topmoumoute online natural gradient algorithm [C] //Proc of the 32nd Conference on Neural Information Processing Systems. 2008: 849-856.

[78] Le ROUX N, BENGIO Y, FITZGIBBON A. Improving first and second-order methods by modeling uncertainty [M] //Optimization for Machine Learning. Cambridge: MIT Press, 2011.

[79] YU Dong, DENG Li. Deep learning and its applications to signal and information processing [J]. IEEE Signal Processing Magazine, 2011, 28(1): 145-154.

[80] YU Dong, WANG Shi-zhen, DENG Li. Sequential labeling using deep-structured conditional random fields [J]. IEEE Journal of Selected Topics in Signal Processing, 2010, 4(6): 965-973.

[81] HAMEL P, ECK D. Learning features from music audio with deep belief networks [C] //Proc of the 11th International Society for Music Information Retrieval Conference. 2010: 339-344.

[82] SCHMIDT E M, KIM Y E. Learning emotion-based acoustic features with deep belief networks [C] //Proc of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics. 2011: 65-68.

[83] DENG Li, SELTZER M, YU Dong, *et al*. Binary coding of speech spectrograms using a deep auto-encoder [C] //Proc of the 11th Annual Conference on International Speech Communication Association. 2010: 1692-1695.

[84] BENGIO Y, BASTIEN F, BERGERON A, *et al*. Deep learners benefit more from out-of-distribution examples [C] //Proc of the 14th International Conference on Artificial Intelligence and Statistics. 2011.

(下转第 1942 页)

puter Applications,2009,29(5):1233-1240.
[15] 冯琳函,钱志鸿,金冬成. 增强型的无线 Mesh 网络信道分配方法[J]. 通信学报,2012,33(10):44-50.
[16] 姬文江,马建峰,田有亮,等. 无线 Mesh 网络中一种基于博弈论的公平性路由协议[J]. 通信学报,2012,33(11):17-23.
[17] KIM K, PARK J. Performance analysis of the multi-channel wireless Mesh networks [C]//Proc of International Conference on Computer Applications for Communication, Networking, and Digital Contents. Berlin: Springer-Verlag, 2012: 161-166.
[18] NAVEED A, KANHEE S S. NIS07-7: security vulnerabilities in channel assignment of multi-radio multi-channel wireless Mesh networks [C]//Proc of Global Telecommunications Conference. 2006: 1-5.
[19] MOHSENIAN-RAD A H, WONG V W S. Joint logical topology design, interface assignment, channel allocation, and routing for multi-channel wireless Mesh networks [J]. IEEE Trans on Wireless Communications, 2007, 6(12): 4432-4440.
[20] CONTI M, DAS S K, LENZINI L, *et al*. Channel assignment strategies for wireless Mesh networks [M]//Wireless Mesh Networks. Berlin: Springer-Verlag, 2008: 161-166.
[21] 王嶔琦. 无线 Mesh 网络路由协议关键技术的研究[D]. 长沙:国防科学技术大学,2009.
[22] JING Tao, SHI Hong-bin, HUO Yan, *et al*. A novel channel assignment scheme for multi-radio multi-channel wireless Mesh networks [C]//Proc of the 6th International Conference on Wireless Algorithms, Systems, and Applications. Berlin: Springer-Verlag, 2011: 261-270.
[23] BIANCHI G. Performance analysis of the IEEE 802.11 distributed coordination function [J]. IEEE Journal on Selected Areas in Communications, 2000, 18(3): 535-547.
[24] SUBRAMANIAN A P, GUPTA H, DAS S R. Minimum interference channel assignment in multi-radio wireless Mesh networks [C]//Proc of the 4th Annual IEEE Conference on Sensor Mesh and Ad hoc Communications and Networks. 2007: 1459-1473.
[25] LIU Jun, XIE Xiu-feng. Cognitive network channel allocation method based on the queuing delay and game analysis [J]. Journal on Communications, 2012, 33(6): 73-81.
[26] 刘玉涛. 认知无线电中基于博弈相关理论的频谱分配算法研究[D]. 哈尔滨:哈尔滨工业大学,2010.
[27] GAO Lin, WANG Xin-bing, XU You-yun. Multi-radio channel allocation in multi-hop wireless networks [J]. IEEE Trans on Mobile Computing, 2009, 8(11): 1454-1468.
[28] FISCHER S, VOCKING B. Evolutionary game theory with applications to adaptive routing [C]//Proc of European Conference on Complex Systems. 2005: 1-6.
[29] LUONG T T, LEE B S, YEO C K. Channel allocation for multiple channels multiple interfaces communication in wireless Ad hoc networks [C]//Proc of the 7th International Conference on Ad hoc and Sensor Networks, Wireless Networks, Next Generation Internet. Berlin: Springer-Verlag, 2008: 87-98.
[30] DAS A K, ALAZEMI H M K, VJAYKUMAR R, *et al*. Optimization models for fixed channel assignment in wireless Mesh networks with multiple radios [C]//Proc of the 2nd Annual IEEE Conference on Sensor and Ad hoc Communications and Networks. 2005: 463-474.
[31] 沈士根,马绚,蒋华,等. 基于演化博弈论的 WSNs 信任决策模型与动力学分析[J]. 控制与决策,2012,27(8):1133-1138.
[32] 周静波. 演化博弈论的基本方法及应用[J]. 中国城市经济,2011,2(3):234-236.
[33] 孙庆文,陆柳,严广乐,等. 不完全信息条件下演化博弈均衡的稳定性分析[J]. 系统工程理论与实践,2003,7(6):11-16.

(上接第 1930 页)

[85] RANZATO M, SUSSKIND J, MNIH V, *et al*. On deep generative models with applications to recognition [C]//Proc of IEEE Conference on Computer Vision and Pattern Recognition. 2011: 2857-2864.
[86] SUSSKIND J, HINTON G E, MEMISEVIC R, *et al*. Modeling the joint density of two images under a variety of transformations [C]//Proc of IEEE Conference on Computer Vision and Pattern Recognition. 2011: 2793-2800.
[87] LUO Ping, WANG Xiao-gang, TANG Xiao-ou. Hierarchical face parsing via deep learning [C]//Proc of IEEE Conference on Computer Vision and Pattern Recognition. Washington DC: IEEE Computer Society, 2012: 2480-2487.
[88] LeCUN Y, CHOPRA S, RANZATO M, *et al*. Energy-based models in document recognition and computer vision [C]//Proc of the 9th International Conference on Document Analysis and Recognition. 2007: 337-341.
[89] TAYLOR G, HINTON G E, ROWEIS S. Modeling human motion using binary latent variables [C]//Advances in Neural Information Processing Systems. Cambridge: MIT Press, 2007: 2007-2014.
[90] RAINA R, BATTLE A, LEE H, *et al*. Self-taught learning: transfer learning from unlabeled data [C]//Proc of the 24th International Conference on Machine Learning. New York: ACM Press, 2007: 759-766.
[91] NAIR V, HINTON G E. 3D object recognition with deep belief nets [C]//Proc of the 24th Annual Conference on Neural Information Processing Systems. 2009: 1339-1347.
[92] TANG Yi-chuan, ELIASMITH C. Deep networks for robust visual recognition [C]//Proc of the 27th International Conference on Machine Learning. 2010: 1055-1062.
[93] LEE N, LAINE A F, KLEIN A. Towards a deep learning approach to brain parcellation [C]//Proc of IEEE International Symposium on Biomedical Imaging: from Nano to Macro. 2011: 321-324.
[94] SOCHER R, BENGIO Y, MANNING C D. Deep learning for NLP (without magic) [EB/OL]. (2012-07-07). http://nlp.stanford.edu/courses/NAACL2013/NAACL2013-Socher-Manning-DeepLearning.pdf.
[95] DESELAERS T, HASAN S, BENDER O, *et al*. A deep learning approach to machine transliteration [C]//Proc of the 4th Workshop on Statistical Machine Translation. 2009: 233-241.
[96] NGIAM J, KHOSLA A, KIM M, *et al*. Multimodal deep learning [C]//Proc of the 28th International Conference on Machine Learning. 2011.
[97] SALAKHUTDINOV R, MURRAY I. On the quantitative analysis of deep belief networks [C]//Proc of the 25th International Conference on Machine Learning. New York: ACM Press, 2008: 872-879.
[98] MURRAY I, SALAKHUTDINOV R. Evaluating probabilities under high-dimensional latent variable models [C]//Advances in Neural Information Processing Systems. Cambridge: MIT Press, 2009: 1137-1144.
[99] LEE H, BATTLE A, RAINA R, *et al*. Efficient sparse coding algorithms [C]//Advances in Neural Information Processing Systems. 2006: 801-808.
[100] RAO N S, NOWAK R D, WRIGHT S J, *et al*. Convex approaches to model wavelet sparsity patterns [C]//Proc of the 18th IEEE International Conference on Image Processing. 2011: 1917-1920.
[101] LI Ji-ming, QIAN Yun-tao. Dimension reduction of hyperspectral images with sparse linear discriminant analysis [C]//Proc of IEEE International Geoscience and Remote Sensing Symposium. 2011: 2927-2930.
[102] WESTON J, RATLE F, COLLOBERT R. Deep learning via semi-supervised embedding [C]//Proc of the 25th International Conference on Machine Learning. New York: ACM Press, 2008: 1168-1175.