◎热点与综述◎

深度学习的研究进展与发展

史加荣[1,2,3],马媛媛[3]
SHI Jiarong[1,2,3], MA Yuanyuan[3]

1.西安建筑科技大学 建筑学院,西安 710055
2.省部共建西部绿色建筑国家重点实验室,西安 710055
3.西安建筑科技大学 理学院,西安 710055
1.School of Architecture, Xi'an University of Architecture and Technology, Xi'an 710055, China
2.State Key Laboratory of Green Building in Western China, Xi'an 710055, China
3.School of Science, Xi'an University of Architecture and Technology, Xi'an 710055, China

SHI Jiarong, MA Yuanyuan. Research progress and development of deep learning. Computer Engineering and Applications, 2018, 54(10):1-10.

Abstract: Deep learning is a broader class of machine learning method based on data representation. Its emergence has not only promoted the development of machine learning, but also accelerated the innovation of artificial intelligence. This paper discusses and compares several typical models of deep learning. It first investigates restricted Boltzmann machine, deep belief network and auto-encoder, and explores their structure, principle, advantages and disadvantages of these unsupervised learning models in detail. Secondly, it discusses several supervised learning models including convolutional neural network, recurrent neural network and deep stacked network. And it also evaluates and analyzes the model structure and working principle. Then it makes a contrastive analysis of typical deep learning models and performs comparative experiments. Deep belief network and convolutional neural network are applied to the handwriting digits recognition task and experimental results show that deep learning models have better recognition performance than traditional neural network. Finally, it discusses the developments and challenges of deep learning in the future.

Key words: deep learning; convolutional neural network; deep belief network; auto-encoder; recurrent neural network; deep stacked network

摘 要:深度学习是基于数据表示的一类更广的机器学习方法,它的出现不仅推动了机器学习的发展,而且促进了人工智能的革新。对深度学习的几种典型模型进行研究与对比。首先介绍受限玻尔兹曼机、深度置信网络、自编码器等无监督学习模型,对其结构、原理和优缺点进行了详细探讨。讨论卷积神经网络、循环神经网络和深度堆叠网络等监督学习模型,分别从模型架构和工作原理来评价与分析。对深度学习的典型模型进行对比分析,将深度置信网络和卷积神经网络应用在手写体数字识别任务中,结果证实深度学习比传统的神经网络具有更好的识别性能。最后探讨深度学习未来的发展与挑战。

关键词:深度学习;卷积神经网络;深度置信网络;自编码器;循环神经网络;深度堆叠网络
文献标志码:A **中图分类号:**TP18 doi:10.3778/j.issn.1002-8331.1712-0418

1 引言

机器学习是人工智能的核心研究领域之一,其最初的研究动机是为了让计算机系统具有人的学习能力以实现人工智能[1]。深度学习(深度结构学习或分层学习)

基金项目:中国博士后科学基金(No.2017M613087);国家自然科学基金青年科学基金(No.61403298)。
作者简介:史加荣(1979—),男,博士后,教授,主要研究方向为机器学习,E-mail:shijiarong@xauat.edu.cn;马媛媛(1991—),女,硕士研究生,主要研究方向为机器学习。
收稿日期:2017-12-28 **修回日期:**2018-03-21 **文章编号:**1002-8331(2018)10-0001-10

是基于数据表示的一类更广的机器学习方法,它通过组合低级特征形成更加抽象的高级表示特征,以发现数据的分布式特征[2]。深度学习使机器学习能够实现更多的应用,并拓展了人工智能的服务范围,已成为诸多领域新的研究热点,如:语音识别[3]、视频识别[4]、图像识别[5]、自然语言处理[6]和信息检索[7]等。

Hinton等人于2006年提出了一种无监督学习模型:深度置信网络,该模型解决了深度神经网络训练的难题,掀起了深度学习的浪潮[8]。此后,深度学习发展非常迅速,涌现出诸多模型。深度置信网络、自编码器[9]、卷积神经网络[10]和循环神经网络[11]构成了早期的深度学习模型,随后由这些模型演变出许多其他模型,主要包括稀疏自编码器[12]、降噪自编码器[13]、堆叠降噪自编码器[14]、深度玻尔兹曼机[15]、深度堆叠网络[16]、深度对抗网络[17]和卷积深度置信网络[18]等。本文主要探讨了深度学习的几种典型模型以及研究与发展。

2　深度学习简介

为简化表示,下面给出深度学习几种典型模型的名称表,如表1所示。

表1　深度学习典型模型名称表

模型	英文名	英文缩写
受限玻尔兹曼机	Restricted Boltzmann Machine	RBM
深度置信网络	Deep Belief Network	DBN
深度玻尔兹曼机	Deep Boltzmann Machine	DBM
自编码器	Auto-Encoder	AE
稀疏自编码器	Sparse Auto-Encoder	SAE
降噪自编码器	Denoising Auto-Encoder	DAE
卷积神经网络	Convolutional Neural Network	CNN
循环神经网络	Recurrent Neural Network	RNN
深度堆叠网络	Deep Stacked Network	DSN

深度学习的概念不仅起源于对人工神经网络的研究[19],而且受到统计力学的启发[20]。1986年,Smolensky提出了一种以能量为基础的模型:RBM,该模型由BM发展而来[21],主要用于语音识别[22]和图像分类[23]。2006年,Hinton和Salakhutdinov提出了一种贪婪的逐层学习网络:DBN,它由多个RBM堆叠而成[24],避免了梯度消失[2,8],主要用于图像识别和信号处理[25];2009年,他们又提出了另一种贪婪的逐层学习模型:DBM[15],该模型也是由多个RBM堆叠而成,主要应用于目标识别和信号处理[26]。

与RBM的发展相独立,Rumelhart于1986年提出了一种无监督学习算法:AE,该算法通过编码器和解码器工作完成训练[12],主要用于语音识别和特征提取[27]。随着AE的发展,它的衍生版本不断出现,如:SAE和DAE。SAE是另一种无监督学习算法,它在AE的编码层上加入了稀疏性限制,主要用于图像处理和语音信号处理[28]。DAE在AE的输入上加入了随机噪声,用来预测缺失值[13]。

与前述模型不同,CNN是一种较流行的监督学习模型,它受猫的视觉皮层研究的启发[10],已成为图像识别[29]和语音识别[30]领域的研究热点。RNN是另一种重要的监督学习模型,专门用来处理序列数据[11],通常用于语音识别、文本生成和图像生成[31]。DSN是一种深度堆叠神经网络,是为研究伸缩性问题而设计的[16]。

机器学习有无监督学习与监督学习之分,不同学习框架下的模型有很大的差异。根据结构和技术应用领域的不同,可以将深度学习分为无监督(生成式)、监督(判别式)和混合深度学习网络[32],而无监督学习可为监督学习提供预训练[2]。最常见的无监督学习模型有RBM,DBN,DBM,AE,SAE,DAE,其中前3个模型以能量为基础,后两个模型以AE为基础。典型的监督学习模型有CNN、RNN和DSN等。混合深度学习通常以生成式或者判别式深度学习网络的结果作为重要辅助,克服了生成式网络模型的不足[33],其代表模型有混合深度神经网络[34](如:DNN-HMM和DNN-CRF)和混合深度置信网络[35](DBN-HMM)。

3　无监督学习模型

先引入以能量为基础的无监督学习模型:RBM、DBN和DBM,再介绍以AE为基础的模型:SAE和DAE。

3.1　RBM

作为一种特殊类型的马尔可夫随机场,RBM由一个可视层和一个隐层组成[2],如图1所示,其中 v 和 h 分别表示可视层和隐层,可视单元和隐单元间均存在连接,而同层单元间无连接。记可视层和隐层的神经元个数分别为 I 和 J,可视单元 $v_i \in \{0,1\}$ 和隐单元 $h_j \in \{0,1\}$ 之间的连接权值为 w_{ij},a_i 和 b_j 分别为可视层和隐层的偏置,$\theta = \{w_{ij}, a_i, b_j\}$。

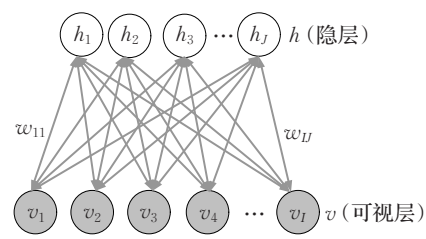

图1　RBM的网络结构

通常假设RBM的隐单元服从伯努利分布,可视单元服从伯努利分布或高斯分布。为了学习模型参数 θ,先定义可视单元不同分布下的两种能量函数[2]:

$$E_1(v,h;\theta) = -\sum_{i=1}^{I}\sum_{j=1}^{J} w_{ij}v_i h_j - \sum_{i=1}^{I} a_i v_i - \sum_{j=1}^{J} b_j h_j \quad (1)$$

$$E_2(v,h;\theta) = -\sum_{i=1}^{I}\sum_{j=1}^{J} w_{ij}v_i h_j - \frac{1}{2}\sum_{i=1}^{I}(v_i - a_i)^2 - \sum_{j=1}^{J} b_j h_j \quad (2)$$

其中 E_1 关于 v、h 是双线性的,E_2 是 h 的线性函数、v 的二次函数。对于一般形式的能量函数 $E(v,h;\theta)$,可视单元和隐单元的联合概率分布为[21]:

$$p(v,h;\theta)=\exp\left(-E(v,h;\theta)\right)/Z(\theta) \tag{3}$$

其中 $Z(\theta)$ 是归一化因子。

RBM模型关于可视单元的边缘分布为[2]:

$$p(v;\theta)=\sum_h p(v,h;\theta) \tag{4}$$

当可视层 v 给定时,第 j 个隐层节点被激活的条件概率为[2]:

$$p(h_j=1|v;\theta)=\mathrm{sigm}(\sum_{i=1}^{I}w_{ij}v_i+b_j) \tag{5}$$

式中,$\mathrm{sigm}(x)=1/\left(1+\exp(-x)\right)$。当隐层 h 给定时,在伯努利分布和高斯分布假设下第 i 个可视层节点被激活的条件概率分别为[2]:

$$p(v_i=1|h;\theta)=\mathrm{sigm}(\sum_{j=1}^{J}w_{ij}h_j+a_i) \tag{6}$$

$$p(v_i|h;\theta)=N(\sum_{j=1}^{J}w_{ij}h_j+a_i,1) \tag{7}$$

其中式(7)右边表示高斯分布。

对式(4)取负对数并对 θ 求偏导有[21]:

$$-\frac{\partial\ln p(v;\theta)}{\partial\theta}=\sum_h p(h|v)\frac{\partial E(v,h;\theta)}{\partial\theta}-\sum_{v,h}p(v,h)\frac{\partial E(v,h;\theta)}{\partial\theta}\triangleq E_{\hat{p}}\left(\frac{\partial E(v,h;\theta)}{\partial\theta}\right)-E_p\left(\frac{\partial E(v,h;\theta)}{\partial\theta}\right) \tag{8}$$

在上式中,$E_{\hat{p}}$ 是在 $p(h|v)$ 下的期望,被称为正向位的期望,它降低了训练数据的能量;E_p 是在 $p(v,h)$ 下的期望,被称为负向位的期望,它提高了模型所有可视单元的能量。

正向位易于计算,而负相位计算相对复杂。可根据采样近似计算负相位,即给定可视层状态,更新隐层状态;给定隐层状态,更新可视层状态[2,21]。为了更好地计算负相位,先根据 k 步吉布斯采样得到 $v^{(k)}$,再利用式(8)对权值 w_{ij} 求偏导:

$$\frac{\partial\ln p(v;\theta)}{\partial w_{ij}}\approx P(h_j=1|v^{(0)})v_i^{(0)}-p(h_j=1|v^{(k)})v_i^{(k)} \tag{9}$$

最后采用对比散度对权值进行更新。类似可计算 a_i 和 b_j。

RBM使用隐变量来描述输入数据的分布,而未涉及数据的标签信息。当有可利用的标签数据时,可将标签信息与数据一起使用,并计算与数据相关的近似目标函数[23]。一般而言,RBM主要用来对神经网络进行预训练,其目的是初始化权值,从而使网络尽可能拟合输入数据。

3.2 DBN

DBN是由多个RBM堆叠而成的神经网络,通常由一个可视层和多个隐层组成,最高的两个隐层存在无向对称边连接,其余隐层形成一个有向的无环图[2,36],如图2所示。该图由一个可视层 v 和三个隐层 h^1、h^2、h^3 组成,连接方式是自上向下,可以看出:DBN的每一层有两个作用,即前一层的隐层和后一层的输入层。

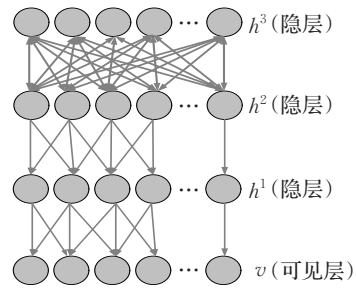

图2 DBN示意图

考虑有 l 个隐层的DBN,令 $h^0=v$,$p(h^k|h^{k+1})$ 是与第 $k+1$ 层相关联的RBM的条件分布,$k=0,1,\cdots,l-1$。DBN最高两个隐层间的连接相当于一个RBM,满足如下公式[20]:

$$P(h^{l-1},h^l)\propto\exp(-E(h^{l-1},h^l;\theta))/Z(\theta) \tag{10}$$

于是DBN关于可视层与隐层的联合概率分布为[20]:

$$P(h^0,h^1,\cdots,h^l)=P(h^{l-1},h^l)\prod_{k=0}^{l-2}P(h^k|h^{k+1}) \tag{11}$$

DBN可以通过无监督预训练(自上向下)和有监督反向微调(自下而上)来训练整个网络[7,8,29],其训练过程如下。先使用无标签数据训练第一层,学习该层参数。再分层训练各层参数,此无监督学习的训练过程相当于网络参数的初始化。最后利用有标签数据进行训练,并使用BP算法将实际输出与预计输出的误差逐层向后传播,此监督学习的训练过程相当于网络参数的微调。作为一种快速贪婪的逐层学习算法,DBN结合了有监督学习与无监督学习各自的优点,能更好地挖掘出有价值的特征[8-9,36]。在预训练过程中,DBN能高效地计算出最深的隐层变量,且能有效地克服过拟合、欠拟合问题。

3.3 DBM

DBM由多个RBM堆叠而成,是一个完整的无向图模型。与RBM相比,DBM可有多层隐变量[2,37-38],且每一层中不同节点都是相互独立的。图3给出了由一个

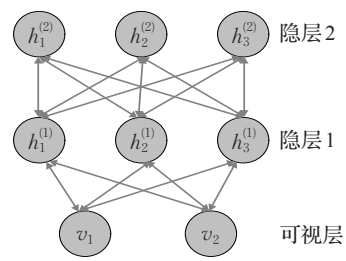

图3 DBM示意图

可视层和两个隐层组成的DBM。为简化表示,此处省略偏置。

对于图3所示的模型,定义能量函数[15]:

$$E(\boldsymbol{v},\boldsymbol{h}^{(1)},\boldsymbol{h}^{(2)};\theta))=-\boldsymbol{v}\boldsymbol{W}^{(1)}\boldsymbol{h}^{(1)}-\boldsymbol{h}^{(1)}\boldsymbol{W}^{(2)}\boldsymbol{h}^{(2)} \tag{12}$$

式中 $\boldsymbol{W}^{(1)}$ 和 $\boldsymbol{W}^{(2)}$ 分别表示可视层到隐层和隐层到隐层的对称连接权值矩阵,$\theta=\{\boldsymbol{W}^{(1)},\boldsymbol{W}^{(2)}\}$。因此,关于可视单元和隐单元的联合概率分布为[15]:

$$p(\boldsymbol{v},\boldsymbol{h}^{(1)},\boldsymbol{h}^{(2)};\theta)=\exp\left(-E\left(\boldsymbol{v},\boldsymbol{h}^{(1)},\boldsymbol{h}^{(2)};\theta\right)\right)/Z(\theta) \tag{13}$$

于是有DBM关于可视单元的边缘分布:

$$p(\boldsymbol{v};\theta)=\sum_{\boldsymbol{h}^{(1)},\boldsymbol{h}^{(2)}}p(v,\boldsymbol{h}^{(1)},\boldsymbol{h}^{(2)};\theta) \tag{14}$$

下面给出可视层和隐层的条件分布[15]:

$$p(v_i=1|\boldsymbol{h}^{(1)})=\mathrm{sigm}(\sum_j w_{ij}^{(1)}h_j^{(1)}) \tag{15}$$

$$p(h_j^{(1)}=1|\boldsymbol{v},\boldsymbol{h}^{(2)})=\mathrm{sigm}(\sum_i w_{ij}^{(1)}v_i+\sum_m w_{jm}^{(2)}h_m^{(2)}),$$

$$p(h_j^{(2)}=1|\boldsymbol{h}^{(1)})=\mathrm{sigm}(\sum_m w_{jm}^{(2)}h_m^{(1)}) \tag{16}$$

作为一种贪婪的逐层学习算法,DBM的训练过程与DBN相似,其学习算法对复杂的输入结构有一个很好的表示[2,37]。但由于直接计算DBM的后验分布较复杂,故采用KL散度和EM算法来计算后验分布,具体计算过程可参考文献[39]。在训练时,以RBM的后验分布对样例进行建模。

3.4　AE

AE通常由三层构成:数据(特征向量)的输入层,特征转换的隐层,用于重构信息的输出层[12]。AE由编码器(encoder)和解码器(decoder)来完成训练[2],其原理如图4所示。将输入向量 $\boldsymbol{x}$ 映射到隐层向量 $\boldsymbol{h}$ 的过程叫做编码,将隐层向量 $\boldsymbol{h}$ 映射到输出向量 $\boldsymbol{r}$ 的过程叫做解码,分别定义如下形式的编码函数和解码函数[61]:

$$\boldsymbol{h}=\mathrm{sigm}(\boldsymbol{W}_1\boldsymbol{x}+\boldsymbol{b}_1),\boldsymbol{r}=\mathrm{sigm}(\boldsymbol{W}_2\boldsymbol{h}+\boldsymbol{b}_2) \tag{17}$$

其中 $\boldsymbol{W}_1$ 和 $\boldsymbol{b}_1$ 分别表示编码器的权值矩阵和偏置向量,$\boldsymbol{W}_2$ 和 $\boldsymbol{b}_2$ 分别表示解码器的权值矩阵和偏置向量。

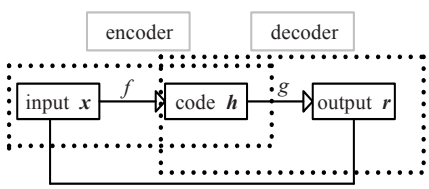

图4　AE编码与解码原理图

AE一般不能复制输入本身,只能让输出尽可能地逼近输入,可通过最小化损失函数求出网络参数[61]:

$$\min_{\theta}\sum_{i=1}^{N}L(\boldsymbol{x}^{(i)},\boldsymbol{r}^{(i)})/N \tag{18}$$

其中,N 为训练样例个数,L 为损失函数。通常要求AE的输入维度与输出维度相等,隐层的维度小于输入维度[16-17]。此时,AE对应的变换就是降维。如果隐层的维度大于输入维度,则很难学习数据中的特征,这时可以给AE加入稀疏性[27]等限制性条件来发现数据中的结构。

AE模型结构简单,训练过程与RBM类似,可以充分利用无标签数据得到网络的初始化权值,从而有效地提取特征[2,40]。训练AE的目的是让输出尽可能逼近输入,但当训练样本与预测样本不符合相同分布时,所提取到的特征往往较差。

3.5　SAE

SAE是在AE的编码层上加入稀疏项[12,41]。当隐层节点被激活的节点数远远小于被抑制的节点数目时,隐层才具有稀疏响应特征[41-42]。SAE正则化的重构误差为[40]:

$$L(\boldsymbol{x},g(\boldsymbol{h}))+\lambda(\boldsymbol{h}) \tag{19}$$

其中 $g(\boldsymbol{h})$ 为输出向量,$\lambda(\boldsymbol{h})$ 为稀疏项。可将KL散度作为稀疏性约束[42],即:

$$\lambda(\boldsymbol{h})=\lambda\sum_{i=1}^{m}KL(p||p_i) \tag{20}$$

式中 λ 是惩罚因子,m 是隐层神经元的个数,p 是隐层神经元激活程度的一个稀疏性参数,p_i 是第 i 个隐层神经元的平均活跃度。p_i 的计算公式如下[42]:

$$p_i=\sum_{j=1}^{m_j}f_i(\boldsymbol{x}^{(j)})/m_j \tag{21}$$

其中,$f_i(\cdot)$ 表示第 i 个隐层神经元的激活函数,m_j 为与此神经元连接的数目。

SAE实现了降维的目的[41],可以为监督学习提供预训练。与多层BP神经网络相比,SAE只是在反向传播时添加了一个稀疏项,从而抑制了大多数神经元的输出。

3.6　DAE

DAE是在AE的输入中加入了随机噪声,将含噪数据经过一个编码器使其形成输入信号的压缩表示,再经过一个解码器得到不含噪声的输出数据,然后计算期望输出与原始输入的误差,最后采用随机梯度下降法来更新网络权值[13]。图5绘出了DAE的原理图。在该图中,$\tilde{x}$ 表示加入噪声后的输入,f 和 y 分别为编码函数和解码函数,z 表示解码层的输出,$L\left(\boldsymbol{x},y(f(\tilde{x}))\right)$ 为损失函数。DAE与AE的编码函数和解码函数相同,只是输入了含有噪声的数据。

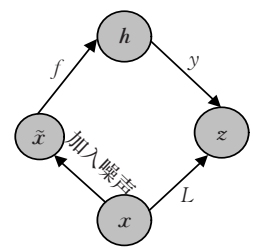

图5　DAE的原理图

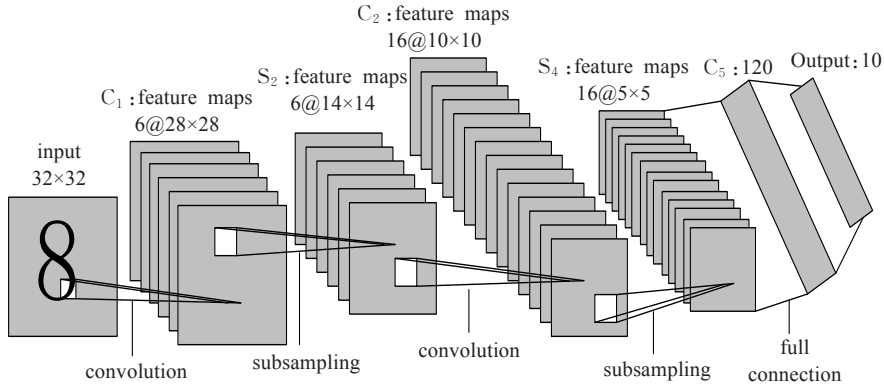

图6　CNN架构图

训练DAE是为了去除随机噪声以获得没有被噪声污染的输入,这就迫使DAE学习比输入信号更加鲁棒的表示,从而更好地预测夹杂在数据中的噪声。因此,DAE也被用来预测缺失值[13,42]。

4　监督学习模型

本章将研究三种典型的监督学习模型:CNN、RNN和DSN。

4.1　CNN

CNN是一种特殊类型的深度前馈神经网络,由输入层、隐层、全连接层和输出层组成。隐层由卷积层和下采样层交替连接组成,即通过卷积操作提取特征,再通过下采样操作得到更加抽象的特征,并将其输入到一个或多个全连接层。最后一个全连接层连接到输出层[43-44],典型的CNN架构如图6所示。卷积层和下采样层构成了CNN的主要模块,下面对它们进行研究。

4.1.1　卷积层

在卷积层中,先将输入图像与卷积核进行卷积,再传递给非线性函数 f,从而得到输出特征图[43]。假设第 $l-1$ 层为下采样层,第 l 层为卷积层,则第 l 层的第 j 个特征图的激活值为[43]:

$$x_j^l=f\left(\sum_{i\in M_j}x_i^{l-1}*k_{ij}^l+b_j^l\right) \tag{22}$$

其中 M_j 是某个特征图像的子集,x_i^{l-1} 是第 $l-1$ 层的第 i 个特征映射所对应的像素值,$k_{ij}{}^l$ 是卷积核,b_j^l 是第 j 个单元所对应的偏置,"*"代表卷积运算。当卷积层提取的特征维数过高时,很容易出现过拟合现象,而下采样层的加入可以在一定程度上减少该现象的发生。

4.1.2　下采样层

下采样层可以减少像素信息,实现图像压缩[45-46]。该层一般采用最大池化或平均池化方法。假设第 $l-1$ 层为卷积层,第 l 层为下采样层。下采样层的输入特征图与输出特征图数目相同,只是特征图变小了。下采样层的计算公式如下[43]:

$$x_j^l=f(\beta_j^l\mathrm{down}(x_j^{l-1},N^l)+b_j^l) \tag{23}$$

其中 N^l 表示第 l 层输入特征图的大小,β_j^l 和 b_j^l 分别为乘性偏置和加性偏置,down(·)表示下采样函数。

CNN有三个重要的特性:稀疏连接、权值共享和池采样[43-47],这些特性可以帮助改善机器学习系统,并使得CNN在一定程度上具有平移、缩放和扭转不变性。

(1)稀疏连接

CNN采用了前向传播计算输出值,反向传播调整权值和偏置。CNN的相邻层之间的(去掉)是稀疏连接,这既减少了模型的内存需求,又提高了计算效率。假设CNN模型有 m 个输入节点和 n 个输出节点,全连接共有 $m\times n$ 个参数;在稀疏连接中,限制每个输出可能具有的连接数为 $k(k\ll m)$,则有 $k\times n$ 个参数[46]。

(2)权值共享

当计算某层的输出时,传统的神经网络仅使用一次权值矩阵。但在CNN中,卷积核共享相同的权值矩阵和偏置向量。图7给出了一个二维卷积操作的例子,其中:左上角为输入数据(4×4矩阵),右上角为卷积核(2×2滤波器),下方为卷积操作结果。由此可以看出:卷积核被重复应用于整个输入数据中。这种权值共享降低了网络复杂度[44]。

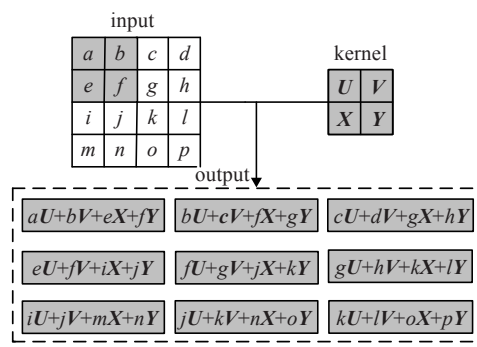

图7　卷积运算示意图

(3)池化

在卷积层获得图像特征后,再对特征进行分类,这通常会产生极大的计算量。采用池化(或下采样)方法对卷积特征进行降维,可在一定程度上保留一些重要或者有用的信息[43-44]。

与传统的图像处理方法相比,CNN避免了前期对图像的预处理。但CNN的特征受到特定的网络结构、学习算法及训练集等诸多因素影响,对其原理的分析与解释更加抽象和困难[2,47]。卷积层的权值共享和下采样层的池化策略降低了网络模型的复杂度,但在训练过程中耗费大量的时间和计算资源,也会出现过拟合现象[45]。模型结构的合理设置及训练速度的提升是CNN亟待解决的问题。

4.2 RNN

RNN是指一个随着时间推移而重复发生的结构,即为时间轴上的循环神经网络[2,48]。它是由输入层、隐层和输出层组成的有向无环结构。隐层是循环实现的基础,其取值不仅取决于本次的输入,还取决于上次隐层的输出,且层级较高的隐层不会向较低的隐层传播。RNN中的“循环”会把系统隐层的输出保留在网络中,再与下一时刻的输入共同决定输出[49]。

给定输入序列 $\{\boldsymbol{x}^{(i)}\}_{i=1}^{t}$ 和预测序列 $\{\boldsymbol{y}^{(i)}\}_{i=1}^{t}$。记 $\boldsymbol{h}_{t-1}$ 和 $\boldsymbol{h}_t$ 分别为 $t-1$ 时刻和 t 时刻所对应的隐变量的状态,$\boldsymbol{O}_t$ 表示 t 时刻所对应的输出,建立如下模型[49]:

$$\boldsymbol{h}_t=f(\boldsymbol{U}\boldsymbol{x}^{(t)}+\boldsymbol{W}\boldsymbol{h}_{t-1}+\boldsymbol{b}),\boldsymbol{O}_t=g(\boldsymbol{V}\boldsymbol{h}_t+\boldsymbol{c}) \tag{24}$$

其中 $\boldsymbol{U}$ 和 $\boldsymbol{V}$ 分别表示从输入层到隐层和隐层到输出层的连接权值,$\boldsymbol{W}$ 表示从隐层到隐层的循环连接权值,$\boldsymbol{b}$ 和 $\boldsymbol{c}$ 分别表示输入层和隐层的偏置,f 和 g 是预先定义的激活函数。一般取 f 为tanh或ReLU函数,g 为softmax函数。将 $\boldsymbol{h}_t$ 和 $\boldsymbol{h}_{t-1}$ 带入 $\boldsymbol{O}_t$ 得[50]:

$$\boldsymbol{O}_t=g\left(\boldsymbol{V}f(\boldsymbol{U}\boldsymbol{x}^{(t)}+\boldsymbol{W}f(\boldsymbol{U}\boldsymbol{x}^{(t-1)}+\boldsymbol{W}\boldsymbol{h}_{t-2}+\boldsymbol{b})+\boldsymbol{b})+\boldsymbol{c}\right) \tag{25}$$

由上式可以看出:输出值 $\boldsymbol{O}_t$ 依赖于 $\boldsymbol{x}^{(t)}$,$\boldsymbol{x}^{(t-1)}$,$\boldsymbol{x}^{(t-2)}$,…,即存在长期依赖问题。

在训练RNN时,仍使用反向传播算法,且在每一个时刻均共享参数。每次的梯度不仅依赖于当前时刻的值,也依赖于之前所有时刻的结果,称此为时间的反向传播(BPTT)[48-49]。BPTT导致参数与隐层状态之间的高度不稳定,从而对梯度下降产生直接影响,即出现“梯度消失问题”。长短时记忆网络(LSTM)是RNN的一种修改结构[50],在学习时仍具有长期依赖性。LSTM通过门的开关来实现时间上的记忆功能,并防止了梯度消失问题。对于多任务学习,LSTM优于RNN。目前,LSTM已被成功应用于语音和手写体识别中。

图8是RNN在时间轴的展开示意图,其中 L_t 表示 t 时刻所对应的损失函数。在每一时步,RNN先接受一个输入向量,再通过非线性函数来更新隐层状态,最后对输出进行预测。RNN常用的损失函数有均方误差函数和交叉熵函数。

由于RNN在所有时刻都共享参数 $\boldsymbol{U}$、$\boldsymbol{V}$ 和 $\boldsymbol{W}$,这极大地减少了需要学习的参数[2,51]。在应用RNN时,往往只需回顾之前的几步,不需要每一刻的输出。虽然RNN在理论上可以建立长时间的间隔状态之间的依赖关系,但由于梯度消失问题,只能学习到短期的依赖关系。

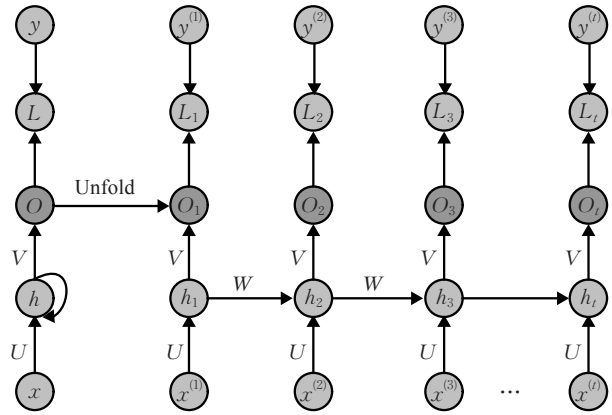

图8　RNN在时间轴的展开图

4.3 DSN

DSN(或深度凸网络)强调学习网络的凸性质。它由多个模块堆叠而成,每一个模块都是一种特殊类型的神经网络且具有相同的结构,即线性输入层、非线性隐层和线性输出层。但每一个模块的输入有所不同,它们将原始输入单元与低层模块中的输出单元连接起来[52-53]。

DSN的最底层模块是构建模型的基础,也由输入单元的线性层、隐单元的非线性层和输出单元的线性层组成[16,52]。记训练样例 $\boldsymbol{x}^{(i)}$ 为 B 维列向量,对应的输出标签 $\boldsymbol{t}^{(i)}$ 为 C 维列向量。最底层模块输出的计算公式为[2]:

$$\boldsymbol{h}_i=\mathrm{sigm}(\boldsymbol{W}_1^{\mathrm{T}}\boldsymbol{x}^{(i)}),\boldsymbol{y}_i=\boldsymbol{U}_1^{\mathrm{T}}\boldsymbol{h}_i\triangleq G_i(\boldsymbol{U}_1,\boldsymbol{W}_1) \tag{26}$$

其中下层权值矩阵 $\boldsymbol{W}_1$ 为 $B\times A$ 维,上层权值矩阵 $\boldsymbol{U}_1$ 为 $A\times C$ 维,$\boldsymbol{h}_i$ 表示隐层的输出单元,$\boldsymbol{y}_i$ 表示底部模块的输出,A 为隐单元的数量。采用均方误差来学习模型参数 $\boldsymbol{U}_1$ 和 $\boldsymbol{W}_1$,其公式如下[2]:

$$E(\boldsymbol{U}_1,\boldsymbol{W}_1)=\sum_i\left\|\boldsymbol{y}_i-\boldsymbol{t}^{(i)}\right\|^2/N \tag{27}$$

其中 N 表示训练样例的总数目。在计算 E 之前,需要先对 $\boldsymbol{W}_1$ 进行经验性设置,下面给出两种方法:随机生成各种分布,将结果用于设置 $\boldsymbol{W}_1$;使用对比散度算法训练RBM,将权值用于设置 $\boldsymbol{W}_1$。

令 E 关于 $\boldsymbol{U}_1$ 的偏导数为0,得 $\boldsymbol{U}_1=F(\boldsymbol{W}_1)$。而在传统的反向传播中,$\boldsymbol{U}_1$ 和 $\boldsymbol{W}_1$ 是相互独立的。构造拉格朗日函数[2]:

$$\sum_i\left\|G_i(\boldsymbol{U}_1,\boldsymbol{W}_1)-\boldsymbol{t}_i\right\|^2+\lambda\left\|\boldsymbol{U}_1-F(\boldsymbol{W}_1)\right\| \tag{28}$$

通过最小化上述函数,得到最优化的参数 $\boldsymbol{W}_1$。

图9绘出了DSN示意图,它由3个模块相互堆叠而成,且构造非常相似,仅在输入层有一个扩展。以块堆叠的目的是从大数据中学习复杂的函数,而学习复杂函数的方法是把简单函数组合在一起形成一个链[52-53]。

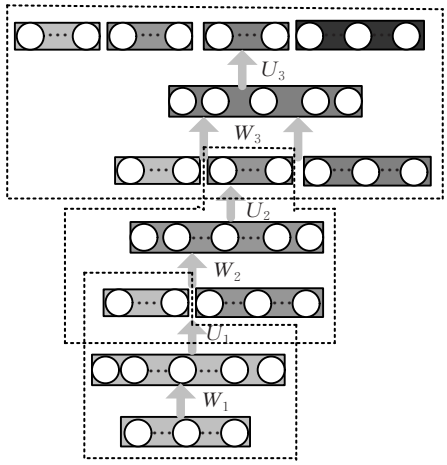

图9 DSN示意图

5 深度学习典型模型对比及在MNIST数据集上的实验

5.1 深度学习典型模型对比

随着深度学习的发展,不断涌现出各种衍生模型。它们都基于深度学习的几种典型模型,因此快速地理解深度学习典型模型及它们之间的关系是至关重要的。表2汇总了深度学习的几种典型模型,该表包括模型、模型结构、训练方式和相关算法等[54-59]。

神经网络(NN)是深度学习的基础;DBN的出现不仅掀起了深度学习的浪潮,而且加快了深度学习的发展;CNN是深度学习最具有代表性的模型。下面在MNIST数据集上对上述三种模型进行评价和对比。

5.2 MNIST数据集与实验参数设计

本文实验使用MNIST手写体数字数据集(http://yann.lecun.com/exdb/mnist/)。该数据集由Google实验室的Corinna和Facebook人工智能负责人Yann LeCun建立,其训练集和测试集分别由60 000和10 000个样例组成[60-61]。每个样本是一幅0~9的手写体数字图片,分辨率为28×28。本文主要使用DeepLearn Toolbox程序,其下载网址如下:https://github.com/rasmusbergpalm/DeepLearnToolbox。此程序使用MATLAB语言编写,在2.9 GHz CPU的个人电脑上运行。

NN由输入层、隐层和输出层组成,每层节点个数分别设置为784、100和10,其中"784"为输入样本的维数(28×28),"10"为类别数目。DBN由输入层、第一隐层、第二隐层和输出层等四层组成,每层节点个数分别设置为784、100、100和10。将CNN设置为一个含输入层在内的五层网络,包含两个卷积层和两个下采样层。CNN的卷积层C_1和C_3分别包含6个和12个大小均为5×5的卷积核,下采样层S_2和S_4对应的采样核大小均为2×2。

5.3 实验结果分析

5.3.1 不同策略下的NN

为了更好地验证NN的有效性,对NN采用了dropout技术[62]和权值衰减策略[61]。Dropout技术是指在模型训练时随机让网络某些隐层节点的权值不工作,此处将dropout的概率设置为0.5。权值衰减是为了避免由于权值越来越大而出现的过拟合现象,设置惩罚因子为10^{-4}。此外,令迭代次数$epoch=1$,批大小$minibatch=100$。

NN、NN+dropout技术、NN+权值衰减策略对应的误分率分别为7.41%、8.65%、1.86%。可以看出:采用权值衰减策略,误分率降低了5.55%;而采用dropout技术,误分率反而增加了1.24%。因此,权值衰减策略可明显提升神经网络的性能。

5.3.2 学习率和*epoch*对DBN的影响

学习率(LearnRate)是深度学习技术的重要参数[59],它决定了每次循环训练过程中所产生的权值变化量。学习率过大或过小都会对实验结果造成影响。通常需要多次调节学习率,或者基于先验知识对其进行设置。

表2 深度学习的典型模型汇总

模型	模型结构	训练方式及算法	模型评价	适用领域
RBM	层内无连接,层间全连接	无监督训练;吉布斯采样,对比散度算法	为监督学习模型提供预训练;所表示的分布无法有效计算	语音识别[22],图像分类[23]
DBN	最高两层全连接,其他层有向连接	无监督预训练和有监督微调;BP算法	防止过拟合;随机梯度下降法,不能有效地进行训练	图像识别[25],信号处理[42]
DBM	层内为全连接(多个隐层)	无监督预训练和有监督微调;EM算法及BP算法	为监督学习模型提供预训练,解决贝叶斯变分推断问题;效率低	目标识别和信号处理[26]
AE	由编码器和解码器组成	无监督训练;BP算法和梯度下降法	用于特征提取与降维;学习速度慢	语音识别[27]
SAE	AE+稀疏性限制	无监督训练;BP算法和梯度下降法	用于降维与编码;不及监督学习的性能	图像处理[13],语音处理[28]
DAE	在AE的输入中加入随机噪声	无监督训练;BP算法和梯度下降法	用来预测缺失值;具有鲁棒性	语音处理[13],信号处理[14]
CNN	卷积层与下采样层交替连接	有监督训练;BP算法和梯度下降法	用于处理图像数据;解决过拟合问题	图像识别[29],语音识别[30]
RNN	有向无环结构,隐层引入定向循环	有监督训练;BPTT和梯度下降法	专门用于处理序列数据,具有较强的计算和建模能力	语音分类、识别[11],文本、图像生成[31]
DSN	以块堆叠	有监督训练;BP算法和梯度下降法	凸优化问题;解决过拟合问题	信息检索[6]

表3 不同学习率和*epoch*下DBN的实验结果

epoch	1		10		50	
	误分率/%	时间/s	误分率/%	时间/s	误分率/%	时间/s
LearnRate=0.1	15.74	5.44	3.33	7.58	2.46	7.96
LearnRate=0.5	15.36	5.40	3.23	10.10	2.47	7.88
LearnRate=1	7.56	7.42	3.35	7.89	2.56	8.04

表4 不同学习率和*epoch*下CNN的实验结果

epoch	1		10		50	
	误分率/%	时间/s	误分率/%	时间/s	误分率/%	时间/s
LearnRate=0.1	80.96	137.05	11.01	148.20	4.47	148.53
LearnRate=0.5	16.48	135.57	4.36	148.54	2.05	189.60
LearnRate=1	11.11	188.08	2.57	138.61	1.39	168.09

一次迭代(epoch)就是将训练集中的全部样例训练一次。分别考虑三种不同的学习率和epoch,DBN的识别率和运行时间如表3所示。

从表3可以看出:当 $epoch=1$ 时,网络的误分率随学习率的增加而降低;当学习率固定时,网络的识别能力随 *epoch* 的增加而增强;随 *epoch* 或学习率的增加,实验运行时间往往也变长。

5.3.3 学习率和 *epoch* 对CNN的影响

对于CNN模型,同样考虑不同学习率和 *epoch* 组合下的识别结果,如表4所示。从表4可以看出,当学习率一定时,网络的误分率随着 *epoch* 的增加而降低;当 *epoch* 固定时,网络的误分率随着学习率的增加而降低。当 $LearnRate=1$、$epoch=50$ 时,网络的识别效果最佳。

6 发展趋势

本文主要探讨了深度学习的几种典型模型,阐述了它们的模型结构、建立、求解和评价,并对这些典型模型进行了总结和对比。DBN等无监督学习模型通常用来协助随后的监督学习,并为其提供预训练;预训练结束后,再使用监督学习进行反向微调。虽然深度学习已被成功应用于语音、视频、图像、自然语言处理和信息检索等诸多科学领域,但仍面临一些挑战[2,33,40,42,55,63-64]:

(1)数学理论的缺乏。对于深度学习框架,业界普遍存在一系列疑问,例如:算法的收敛性与稳定性;深度学习需要多少隐层;在大规模网络中,需要多少有效参数。不管是构建更好的深度学习系统,还是提供更好的解释,深度学习都需要完善的理论支持。

(2)深度学习的应用推广。在应用经典的深度学习模型时,实验结果可能不理想,这就要求根据特定的问题与数据来制定和优化深度学习的网络结构。

(3)深度网络训练的求解问题。这些问题主要包括:随网络层数增加而带来的梯度消失问题;如何有效地设置深度学习的模型参数和进行大规模并行训练。

(4)新模型对人工智能发展的影响。深度学习不断涌现出新的模型,如:生成对抗网络和胶囊网络等。这些模型可能会从观念上挑战传统的深度学习,也可能会改变计算机视觉传输的方式,重塑人工智能。

随着人工智能的蓬勃发展,我国越来越多的学者开始关注深度学习。深度学习将智能技术从实验室带到了产业及应用层面,但许多学者仍将深度学习当做一种工具来使用,忽略了它的分类及基础概念、技术的历史进程和发展方向,从而导致人们对此人工智能技术的整体发展趋势及可用性缺乏宏观认识。因此,为了加深对深度学习的理解,需要完善深度学习的数学理论,并将深度学习技术应用于大数据相关问题的求解上,尤其是数据的高维度、学习算法的可扩展性及分布式计算等。

参考文献:

[1] Arel I, Rose D C, Karnowski T P.Deep machine learning-a new frontier in artificial intelligence research[J].IEEE Computational Intelligence Magazine, 2010, 5(4): 13-18.

[2] Deng L, Yu D.Deep learning: methods and applications[J]. Foundations and Trends in Signal Processing, 2014, 7(3/4): 197-387.

[3] 王山海,景新幸,杨海燕.基于深度学习神经网络的孤立词语音识别的研究[J].计算机应用研究,2015,32(8):2289-2291.

[4] 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(NIPS), 2009: 1096-1104.

[5] 许可.卷积神经网络在图像识别上的应用的研究[D].杭州:浙江大学,2012.

[6] 林奕鸥,雷航,李晓瑜,等.自然语言处理中的深度学习:方法及应用[J].电子科技大学学报,2017,46(6):913-919.

[7] Deng L, He X, Gao J.Deep stacking networks for information retrieval[C]//IEEE International Conference on Acoustics, Speech and Signal Processing(ICASSP), 2013: 3153-3157.

[8] Hinton G E,Osindero S,Teh Y W.A fast learning algorithm for deep belief nets[J].Neural Computation,2006,18(7):1527-1554.

[9] Bengio Y,Lamblin P,Popovici D,et al.Greedy layer-wise training of deep networks[C]//Advances in Neural Information Processing Systems,2007:153-160.

[10] Abdel-Hamid O,Deng L,Yu D.Exploring convolutional neural network structures and optimization techniques for speech recognition[C]//Interspeech,2013:3366-3370.

[11] Martens J,Sutskever I.Learning recurrent neural networks with hessian-free optimization[C]//Proceedings of the 28th International Conference on Machine Learning (ICML),2011:1033-1040.

[12] Sainath T N,Kingsbury B,Ramabhadran B.Auto-encoder bottleneck features using deep belief networks[C]//IEEE International Conference on Acoustics,Speech and Signal Processing(ICASSP),2012:4153-4156.

[13] Vincent P,Larochelle H,Bengio Y,et al.Extracting and composing robust features with denoising autoencoder[C]//Proceedings of the 25th International Conference on Machine Learning(ICML),2008.

[14] 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,2010:3371-3408.

[15] Salakhutdinov R,Hinton G.Deep Boltzmann machines[C]//Artificial Intelligence and Statistics,2009:448-455.

[16] Deng L,Yu D,Platt J.Scalable stacking and learning for building deep architectures[C]//IEEE International Conference on Acoustics,Speech and Signal Processing (ICASSP),2012:2133-2136.

[17] Goodfellow L,Pouget-Abadie J,Mirza M,et al.Generative adversarial networks[C]//Advances in Neural Information Processing Systems(NIPS),2014.

[18] Lee H,Grosse R,Ranganath R,et al.Unsupervised learning of hierarchical representations with convolutional deep belief networks[J].Communications of the ACM,2011,54(10):95-103.

[19] Ajith A.Artifical neural networks[M].Sydenham P H,Thorn R.Handbook of measuring system design.New York:John Wiley & Sons,2005.

[20] Bengio Y.Learning deep architectures for AI[J].Foundations and trends in Machine Learning,2009,2(1):1-127.

[21] Hinton G.A practical guide to training restricted Boltzmann machines[J].Momentum,2012,9(1):926.

[22] Mohamed A R,Hinton G.Phone recognition using restricted Boltzmann machines[C]//IEEE International Conference on Acoustics,Speech and Signal Processing(ICASSP),2010:4354-4357.

[23] Larochelle H,Bengio Y.Classification using discriminative restricted Boltzmann machines[C]//Proceedings of the 25th International Conference on Machine Learning (ICML),2008:536-543.

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

[25] Mohamed A R,Yu D,Deng L.Investigation of full-sequence training of deep belief networks for speech recognition[C]//Eleventh Annual Conference of the International Speech Communication Association,2010.

[26] Ngiam J,Chen Z.Learning deep energy models[C]//Proceedings of the 28th International Conference on Machine Learning(ICML),2011:1105-1112.

[27] Deng L,Seltzer M L,Yu D,et al.Binary coding of speech spectrograms using a deep auto-encoder[C]//Eleventh Annual Conference of the International Speech Communication Association,2010.

[28] Bengio Y,Courville A,Vincent P.Representation learning:A review and new perspectives[J].IEEE Transactions on Pattern Analysis and Machine Intelligence,2013,35(8):1798-1828.

[29] Lawrence S,Giles C L,Tsoi A C,et al.Face recognition:A convolutional neural-network approach[J].IEEE Transactions on Neural Networks,1997,8(1):98-113.

[30] 张晴晴,刘勇,王智超,等.卷积神经网络在语音识别中的应用[J].网络新媒体技术,2014(6):39-42.

[31] Graves A.Sequence transduction with recurrent neural networks[J].arXiv:1211.3711,2012.

[32] Deng L.An overview of deep-structured learning for information processing[C]//Proceedings of Asian-Pacific Signal & Information Processing Annual Summit and Conference(APSIPA-ASC),2011.

[33] Bengio Y.Deep learning of representations for unsupervised and transfer learning[C]//Proceedings of ICML Workshop on Unsupervised and Transfer Learning,2012:17-36.

[34] Dahl G E,Yu D,Deng L,et al.Context-dependent pre-trained deep neural networks for large vocabulary speech recognition[J].IEEE Transactions on Audio,Speech,and Language Processing,2012,20(1):30-42.

[35] Dahl G E,Yu D,Deng L,et al.Context-dependent DBN-HMMs in large vocabulary continuous speech recognition[C]//Proceedings of International Conference on Acoustics,Speech and Signal Processing(ICASSP),2011.

[36] Mohamed A R,Dahl G E,Hinton G E.Acoustic modeling using deep belief networks[J].IEEE Transactions on Audio,Speech,and Language Processing,2012,20(1):14-22.

[37] Goodfellow L,Mirza M,Courville A,et al.Multi-prediction deep Boltzmann machines[C]//Advances in Neural Information Processing Systems(NIPS),2013:548-556.

[38] Salakhutdinov R R,Hinton G E.A better way to pretrain deep boltzmann machines[C]//Advances in Neural Information Processing Systems(NIPS),2012:2447-2455.

[39] Tzikas D G,Likas A C,Galatsanos N P.The variational approximation for Bayesian inference[J].IEEE Signal Processing Magazine,2008,25(6):131-146.

[40] 焦李成,赵进,杨淑媛,等.稀疏认知学习,计算与识别的研究进展[J].计算机学报,2016,39(4):835-851.

[41] Coates A,Ng A Y.The importance of encoding versus training with sparse coding and vector quantization[C]//Proceedings of the 28th International Conference on Machine Learning(ICML),2011:921-928.

[42] 焦李成,赵进,杨淑媛,等.深度学习、优化与识别[M].北京:清华大学出版社,2017:100-120.

[43] Bouvrie J.Notes on convolutional neural networks[J/OL].(2006).http://cogprints.org/5869/1/cnn_tutorial.pdf.

[44] Deng L,Abdel-Hamid O,Yu D.A deep convolutional neural network using heterogeneous pooling for trading acoustic invariance with phonetic confusion[C]//IEEE International Conference on Acoustics,Speech and Signal Processing(ICASSP),2013:6669-6673.

[45] Zeiler M D,Fergus R.Visualizing and understanding convolutional networks[C]//European Conference on Computer Vision(ECCV).Cham:Springer,2014:818-833.

[46] Goodfellow L,Bengio Y,Courvile A.Deep learning[M].[S.l.]:MIT Press,2016.

[47] 李彦冬,郝宗波,雷航.卷积神经网络研究综述[J].计算机应用,2016,36(9):2508-2515.

[48] 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.

[49] Gulcehre C,Cho K,Pascanu R,et al.Learned-norm pooling for deep feedforward and recurrent neural networks[C]//Joint European Conference on Machine Learning and Knowledge Discovery in Databases.Berlin,Heidelberg:Springer,2014:530-546.

[50] Hochreiter S,Schmidhuber J.Long short-term memory[J].Neural Computation,1997,9(8):1735-1780.

[51] 邓力,俞栋.深度学习方法及应用[M].谢磊,译.北京:机械工业出版社,2015:48-57.

[52] Huang P S,Deng L,Hasegawa-Johnson M,et al.Random features for kernel deep convex network[C]//IEEE International Conference on Acoustics,Speech and Signal Processing(ICASSP),2013:3143-3147.

[53] Hutchinson B,Deng L,Yu D.A deep architecture with bilinear modeling of hidden representations:Applications to phonetic recognition[C]//IEEE International Conference on Acoustics,Speech and Signal Processing(ICASSP),2012:4805-4808.

[54] 马世龙,乌尼日其其格,李小平.大数据与深度学习综述[J].智能系统学报,2016,11(6):728-742.

[55] 刘帅师,程曦,郭文燕,等.深度学习方法研究新进展[J].智能系统学报,2016,11(5):567-577.

[56] 孙志军,薛磊,许阳明,等.深度学习研究综述[J].计算机应用研究,2012,29(8):2806-2810.

[57] Yu D,Deng L.Deep learning and its applications to signal and information processing[J].IEEE Signal Processing Magazine,2011,28(1):145-154.

[58] Schmidhuber J.Deep learning in neural networks:An overview[J].Neural Networks,2015,61:85-117.

[59] Huang F J,Boureau Y L,LeCun Y.Unsupervised learning of invariant feature hierarchies with applications to object recognition[C]//IEEE Conference on Computer Vision and Pattern Recognition(CVPR),2007:1-8.

[60] Deng L.The MNIST database of handwritten digit images for machine learning research[J].IEEE Signal Processing Magazine,2012,29(6):141-142.

[61] Palm R B.Prediction as a candidate for learning deep hierarchical models of data[J].Technical University of Denmark,2012,5.

[62] Ba J,Frey B.Adaptive dropout for training deep neural networks[C]//Advances in Neural Information Processing Systems(NIPS),2013:3084-3092.

[63] 范竣翔,李琦,朱亚杰,等.基于RNN 的空气污染时空预报模型研究[J].测绘科学,2017,42(7):76-83.

[64] 尹宝才,王文通,王立春.深度学习研究综述[J].北京大学学报,2015,41(1):49-58.