深度学习研究与进展

孙志远[1,2] 鲁成祥[1,3] 史忠植[1] 马 刚[1,2]
(中国科学院计算技术研究所智能信息处理重点实验室 北京 100190)[1]
(中国科学院大学 北京 100049)[2] (曲阜师范大学信息科学与工程学院 日照 276826)[3]

摘 要 深度学习是机器学习领域一个新兴的研究方向,它通过模仿人脑结构,实现对复杂输入数据的高效处理,智能地学习不同的知识,而且能够有效地解决多类复杂的智能问题。近年来,随着深度学习高效学习算法的出现,机器学习界掀起了研究深度学习理论及应用的热潮。实践表明,深度学习是一种高效的特征提取方法,它能够提取数据中更加抽象的特征,实现对数据更本质的刻画,同时深层模型具有更强的建模和推广能力。鉴于深度学习的优点及其广泛应用,对深度学习进行了较为系统的介绍,详细阐述了其产生背景、理论依据、典型的深度学习模型、具有代表性的快速学习算法、最新进展及实践应用,最后探讨了深度学习未来值得研究的方向。

关键词 深度学习,机器学习,深层神经网络,图像识别,语音识别,自然语言处理
中图法分类号 TP391 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2016.2.001

Research and Advances on Deep Learning

SUN Zhi-yuan[1,2] LU Cheng-xiang[1,3] SHI Zhong-zhi[1] MA Gang[1,2]
(Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China)[1]
(University of Chinese Academy of Sciences, Beijing 100049, China)[2]
(School of Information Science and Engineering, Qufu Normal University, Rizhao 276826, China)[3]

Abstract Deep learning (DL) is a recently-developed field belonging to machine learning. It tries to mimic the human brain, which is capable of processing the complex input data fast, learning different knowledge intellectually, and solving different kinds of complicated human intelligence tasks well. Recently, with the advent of a fast learning algorithm for DL, the machine learning community set off a surge to study the theory and applications of DL since it has many advantages. Practice shows that deep learning is a kind of high efficient feature extraction method, which can detect more abstract characteristics and realize the essence of the data, and the model constructed by DL tends to have stronger generalization ability. Due to the advantages and wide applications of deep learning, this paper attempted to provide a started guide for novice. It presented a detailed instruction of the background and the theoretical principle of deep learning, its emblematic models, its representative learning algorithm, the latest progress and applications. Finally, some research directions of deep learning that are deserved to be further studied were discussed.

Keywords Deep learning, Machine learning, Deep neural network, Image recognition, Speech recognition, Natural language processing

机器学习研究的主要任务是设计和开发可以智能地根据实际数据进行"学习"的算法,这些算法可以自动地挖掘隐藏在数据中的模式和规律。目前,各种机器学习算法在科研、工业、金融、医药等诸多领域都扮演着非常重要的角色。人工神经网络(ANN)[1,2]作为一种通过模仿生物神经网络建立起来的计算模型,是很具有代表性的一类机器学习方法。ANN因其较好的自学习、建模能力和较强的鲁棒性能等,受到学界的广泛关注。

在ANN中,感知机[3]是较早被提出的一种ANN模型,它的学习能力非常有限。随后,多隐层神经网络(MNN)[4]被提出,其具有较强的无监督学习能力,能够挖掘数据中隐藏的复杂模式和规则。但是,MNN训练时间较长,而且极易陷入局部最优解[5];使用传统学习算法训练MNN时,存在误差信号逐层衰减等问题。

2006年,Hinton等[6]提出了深度置信网络(DBN)和相应的高效学习算法。这个算法成为了其后至今深度学习算法的主要框架。该算法中,一个DBN是由多个受限波尔兹曼机(RBM)[7,8]以串联的方式堆叠而形成的一种深层网络,训练

到稿日期:2015-01-15 返修日期:2015-04-25 本文受国家"九七三"重点基础研究计划(2013CB329502),国家自然科学基金(61035003)资助。

孙志远(1991-),男,硕士生,主要研究方向为人工智能、机器学习、数据挖掘、图像处理等,E-mail:sunzy@ics.ict.ac.cn;鲁成祥(1988-),男,硕士生,主要研究方向为多智能体、强化学习;史忠植(1941-),男,博士生导师,主要研究方向为智能科学、人工智能、机器学习、多智能体,E-mail:shizz@ics.ict.ac.cn;马 刚(1986-),男,博士生,主要研究方向为人工智能、机器学习、神经计算、复杂网络。

时通过自低到高逐层训练 RBM 将模型参数初始化为较优值，再使用少量传统学习算法对网络微调，使得模型收敛到接近最优值的局部最优点。由于 RBM 可以通过对比散度(CD)[9]等算法快速训练，这一框架避开了直接训练 DBN 的高计算量，将模型化简为对多个 RBM 的训练问题。这个学习算法解决了模型训练速度慢的问题，能够产生较优的初始参数，有效地提升了模型的建模、推广能力。自此，深层神经网络难以有效训练的僵局被成功打破[6,10,11]，机器学习界掀起了深度学习[12-14]的研究热潮。自 2006 年至今，深度学习研究对机器学习领域产生了非常大的影响[14-18]，很多顶级会议和专题报告如 NIPS、ICML 等[19]都对深度学习及其在不同领域的应用给予了很大关注。

当前，深度学习已经成为机器学习领域富有生命力的研究方向。鉴于深度学习占据的核心地位以及其本身优良的性质，为了给深度学习初学者提供入门指导，便于其对深度学习展开进一步的研究与实践，本文将对深度学习进行较为系统的介绍，阐述其产生背景、理论依据、典型的深层网络模型、主流的学习算法、最新进展及实践应用，最后对深度学习未来值得研究的方向进行探讨。

本文第 1 节讲述深度学习的研究背景和理论依据；第 2 节阐述深度学习的 3 类模型；第 3 节详细介绍当前深度学习训练方法的研究进展；第 4 节讨论深度学习在不同领域的主要应用以及面对的挑战；最后总结与展望，主要探讨深度学习在未来值得研究的方向。

1 深度学习的研究背景与理论依据

1.1 研究背景

深度学习是机器学习领域一个比较年轻的研究方向。机器学习的目标是让机器像人一样去感知环境中的声音、图像等信息。1958 年，Rosenblatt[14]提出感知机模型，它能够对一些简单形状分类，这掀起了神经网络研究的第一个高潮。但是它的特征提取层是人工构造的，单隐层结构限制了它的学习能力，使得它与智能感知相悖。1986 年，Hinton[20]基于感知机提出用多隐层构造深层神经网络。相对于感知机，深层神经网络能够学习更加复杂的功能。深层神经网络使用反向传播算法(BP)[21]训练模型，但是 BP 训练效果较差[13,22]。MNN 存在如下问题：(1)不能训练未标注数据；(2)误差信号存在梯度扩散；(3)学习速率缓慢；(4)易于过拟合。这些问题使得 MNN 研究陷入低潮。随后，Vapnik 等[23]提出支持向量机(SVM)。SVM 使用核方法实现输入数据向高维空间的映射，使得模型的训练变得高效。但是 SVM 要求对数据有一定的先验知识，虽然可以人工向 SVM 加入先验[24]，但是先验不是机器主动学习得出，与机器学习的目标相悖。尽管 SVM 使用核方法取代人工挑选特征，使用序列最小优化(SMO)算法[25]等训练模型，但是 SVM 本质上是一种局部估计算子，要求数据具有平滑性和足够的先验知识。因此，当面向大数据或数据高度复杂时，SVM 不再适用。

1.2 理论依据

在神经学方面，1981 年，Hubel[26]等发现了视觉系统的信息处理在可视皮层是分级的，激发了对神经系统的信息处理机制的探索：神经-中枢-大脑的工作过程，即通过从原始信号进行低级抽象，并逐渐向高级抽象迭代。特征提取从低到高逐级抽象，从而更有利于特征表示。深度学习通过模仿人脑，建立了一个深层神经网络，通过输入层输入数据，由低到高逐层提取特征，建立起低级特征到高级语义之间复杂的映射关系。

Bengio 等[27]发现，很多使用浅层结构需要指数级的参数才能有效表示的问题，而在深层结构中只需要多项式级的参数。这从统计学的角度验证了深度学习的高效性。Hinton[6]提出一种贪心无监督逐层学习算法，对深层结构逐层初始化参数，再使用有监督算法对模型参数微调。这种学习方法克服了传统训练遇到的问题，使得深层神经网络的学习变得高效。Bengio[10]通过实验验证了深层神经网络和贪心学习算法的高效性：(1)在用自动编码机取代 RBM 构建的深层神经网络中，使用贪心训练算法能取得良好的效果，这表明这种学习算法是一种具有较强推广能力的学习算法；(2)在深层结构中，使用贪心无监督学习算法相对于浅层结构以及参数随机初始化的深层结构能够获得更高的分类准确率；(3)这种贪心学习算法能够得到更好的推广的原因是它对深层结构的每一层进行单独训练后，使得每一层的参数已经比较接近最优化参数；(4)使用贪心算法训练模型时，采用无监督方式比有监督方式得到的模型参数推广性能更好，因为无监督策略相对于有监督策略能够学习更多体现数据本身特征的知识。

2 深度学习的 3 类模型

深度学习是机器学习领域技术和模型较为丰富的一个研究方向，代表了以使用深层神经网络实现数据拟合的一类机器学习方法[28]。根据深层神经网络的构造方式、训练方法等因素，深度学习可分为 3 大类别：生成深层结构、判别深层结构以及混合深层结构。

生成深层结构是通过学习观测数据高阶相关性，或观测数据和关联类别之间的统计特征分布来实现模式分类的一类深层结构。判别深层结构是通过直接学习不同类别之间的区分表达能力来实现模式分类的一类深层结构。混合深层结构是将生成模块和判别模块相结合而成的一类深层结构。

2.1 生成深层结构

深层生成模型根据网络结构中是否存在方向性可以分为深层有向网络、深层无向网络以及深层混合网络 3 种[13]。本文以 DBN[6]为代表详细介绍生成深层结构。DBN 模型是一种深层混合网络，如图 1 所示，它以 RBM[7]为基本单元串联堆叠构成。DBN 的训练是通过先逐层训练 RBM，再使用传统学习算法进行微调。因此，介绍 DBN 之前需要重点介绍 RBM 的结构、原理和训练方法。

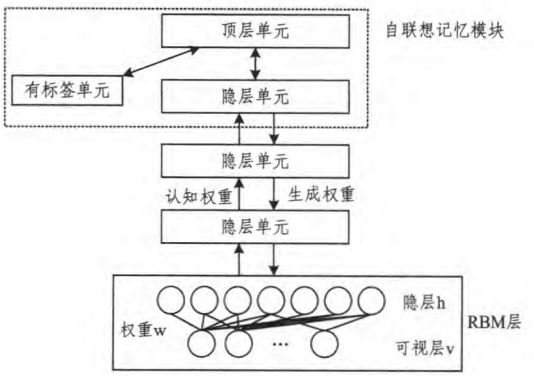

图 1 DBN 模型

RBM是一类具有对称连接、无自反馈的双层、无向随机神经网络模型，层间全连接，层内无连接，如图2所示。其中，v为可见层，表示观测数据；h为隐层，表示特征提取器；w为两层之间的连接权重。Welling[29]指出，RBM中的隐单元和可见单元可以为任意的指数族单元。

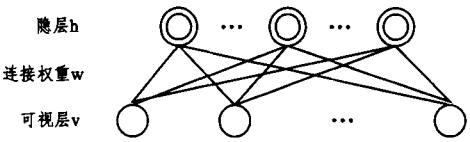

图2 RBM模型

设所有可见单元和隐单元均为二值变量，即$\forall i,j,U_{ij}^{j}\in\{0,1\},h_j\in\{0,1\}$。

这里假定RBM有n个可见单元和m个隐单元，用向量v和h分别表示可见单元和隐单元的状态向量。其中，v_i表示第i可见单元的状态，h_j表示第j隐单元的状态。对于一组给定状态(v,h)，RBM作为一个系统所具备的能量[30]，定义为：

$$E(v,h)=-\sum_{i=1}^{n}a_iv_i-\sum_{j=1}^{m}b_jh_j-\sum_{i=1}^{n}\sum_{j=1}^{m}v_iw_{ij}h_j \tag{1}$$

其中，w_{ij}表示可见单元i与隐单元j之间的连接权重，a_i表示可见单元i的偏置，b_j表示隐单元j的偏置。假定RBM参数为θ，$\theta=\{w_{ij},a_i,b_j\}$。当参数$\theta$确定时，基于该能量函数，可以得到$(v,h)$的联合概率分布：

$$P(v,h)=\frac{e^{-E(v,h)}}{Z(\theta)} \tag{2}$$

$$Z(\theta)=\sum_{v}\sum_{h}e^{-E(v,h)} \tag{3}$$

其中，$Z(\theta)$为配分函数，在这里实现归一化。对于一个实际问题，需要求解由RBM关于观测数据v的分布$P(v)$，即联合概率分布$P(v,h)$的边缘分布：

$$P(v)=\frac{1}{Z(\theta)}\sum_{h}e^{-E(v,h)} \tag{4}$$

为了确定该分布，需要计算配分函数$Z(\theta)$。因此，即使通过训练可以得到模型的参数θ，但仍然无法有效地计算由这些参数确定的分布。

由RBM层内无连接、层间全连接可知，当给定v时，各隐单元之间的状态条件独立，即h_j的激活概率为：

$$P(h_j=1)=\sigma(b_j+\sum_i v_iw_{ij}) \tag{5}$$

其中，$\sigma(z)=1/(1+\exp(-z))$为sigmoid激活函数。

由于RBM的结构具有对称性，当给定隐单元状态时，各可见单元的激活状态也条件独立，即v_i的激活概率为：

$$P(v_i=1)=\sigma(a_i+\sum_i v_iw_{ij}) \tag{6}$$

学习RBM的任务是求出参数θ的值，以拟合给定观测数据。参数θ可以通过最大化RBM在训练集上（假设包含K个样本）的对数似然函数学习得到，即

$$\theta^*=\arg\max_{\theta}L(\theta)=\arg\max_{\theta}\sum_{k=1}^{K}\log P(v^{(k)}) \tag{7}$$

为了获得最优参数θ^*，可以使用随机梯度上升(SGA)算法求$L(\theta)=\sum_{k=1}^{K}\log(v^{(k)})$的最大值。其中，关键步骤是计算$\log(v^{(k)})$关于各个模型参数的偏导数。

由于

$$L(\theta)=\sum_{k=1}^{K}\log P(v^{(k)})=\sum_{k=1}^{K}\log P(v^{(k)},h)$$

$$=\sum_{k=1}^{K}\log\frac{\sum_h\exp[-E(v^{(k)},h)]}{\sum_v\sum_h\exp[-E(v,h)]}$$

$$=\sum_{k=1}^{K}(\langle\log\sum_h\exp[-E(v^{(k)},h)]\rangle-\langle\log\sum_v\sum_h\exp[-E(v,h)]\rangle) \tag{8}$$

令θ_1表示θ中的某一个参数，则对数似然函数关于θ_1的梯度为：

$$\frac{\partial L}{\partial\theta_1}=\sum_{k=1}^{K}\frac{\partial}{\partial\theta_1}(\langle\log\sum_h\exp[-E(v^{(k)},h)]\rangle-\langle\log\sum_v\sum_h\exp[-E(v,h)]\rangle)$$

$$=\sum_{k=1}^{K}(\langle\sum_h\frac{\exp[-E(v^{(k)},h)]}{\sum_h\exp[-E(v^{(k)},h)]}\times\frac{\partial(-E(v^{(k)},h))}{\partial\theta_1}\rangle-\langle\sum_v\sum_h\frac{\exp[-E(v,h)]}{\sum_v\sum_h\exp[-E(v,h)]}\times\frac{\partial(-E(v,h))}{\partial\theta_1}\rangle)$$

$$=\sum_{k=1}^{K}(\langle\frac{\partial(-E(v^{(k)},h))}{\partial\theta_1}\rangle_{P(h)}-\langle\frac{\partial(-E(v,h))}{\partial\theta_1}\rangle_{P(v,h)}) \tag{9}$$

其中，$\langle\cdot\rangle_P$表示求关于分布P的数学期望；$P(h)$表示在可见单元限定为训练样本$v^{(k)}$时隐层的概率分布；$P(v,h)$表示可见单元与隐单元的联合分布，由于配分函数$Z(\theta)$，该分布很难获取，只能通过采样方法获取其近似值。

下面假设只有一个训练样本，分别用“data”和“model”来简记$P(h)$和$P(v,h)$这两个概率分布，则对数似然函数关于w_{ij}、a_i和b_j的偏导数分别为：

$$\frac{\partial\log P(v)}{\partial w_{ij}}=\langle v_ih_j\rangle_{data}-\langle v_ih_j\rangle_{model} \tag{10}$$

$$\frac{\partial\log P(v)}{\partial a_i}=\langle v_i\rangle_{data}-\langle v_i\rangle_{model} \tag{11}$$

$$\frac{\partial\log P(v)}{\partial b_j}=\langle h_j\rangle_{data}-\langle h_j\rangle_{model} \tag{12}$$

这里，通过Gibbs采样[31]得到服从RBM分布的样本。在RBM中，进行k步Gibbs采样的算法为：用一个观测样本初始化可见层状态v_0，交替进行如下操作：$h_0\sim P(h|v_0)$，$v_1\sim P(v|h_0)$，$h_1\sim P(h|v_1)$，$v_2\sim P(v|h_1)$，…，$v_{k+1}\sim P(v|h_k)$。当采样步数k足够大时，得到服从RBM分布的样本。

由于DBN由多个RBM串联堆叠组成，在训练时，从低到高逐层训练RBM[10]：

(1)低层RBM使用观测样本进行训练；

(2)低层RBM的输出作为高层RBM的输入进行训练；

(3)迭代(1)和(2)，重复训练所有RBM，实现模型参数的初始化。

使用上述算法对模型参数进行初始化，再使用传统学习算法进行网络微调，这就是DBN模型的整个学习过程。

使用Gibbs采样可以得到对数似然关于参数梯度的近似估计，但一般要求较大的采样步数，导致RBM学习时间过长。有关RBM学习算法的改进详见3.2节。

以上模型训练方法适用于大多数以RBM为基本单元构造的深层神经网络。

2.2 判别深层结构

判别深层结构包括深层堆叠网络、卷积神经网络等。本文以卷积神经网络(CNN)[32]为代表详细介绍判别深层结构。

1962年，Hubel等[26]通过研究猫视觉机理，提出感受野的概念。1984年，Fukushima[33]基于感受野提出神经感知机，这是第一个成功实现的CNN模型，也是感受野在ANN

领域的首次成功应用。LeCun 等[34]基于 CNN，将 BP 应用于 CNN 模型的训练，并成功应用于图像识别等问题。CNN 因具有位移、畸变鲁棒性和并行性等而受到广泛关注。

CNN 是一种局部连接、权值共享的前馈式多层神经网络，每层神经网络由一对二维平面组成，每个二维平面由多个独立神经元组成。其中，每对二维平面包括对位移鲁棒的卷积层和提取特征的子采样层，二维平面上的每个神经元只与前一层的局部感受野相连接，并提取局部特征。卷积层和子采样层都由多个特征平面构成，同一特征平面共享特征参数，不同的特征平面使用不同特征参数提取不同特征，每一个特征映射为一个特征平面，因而减少了网络中自由参数的个数，降低了网络中参数选择的复杂度[35]，提高了网络的泛化能力。特征映射结构采用影响函数较小的 sigmoid 函数作为卷积网络的激活函数，使得特征映射具有位移鲁棒性。卷积网络中每一个卷积层都连接着一个二次特征提取的子采样层，并有多个这样的二维结构串联在卷积神经网络中，这样的二次特征提取结构使得网络在识别时对输入样本具有较强的畸变鲁棒性，同时多层串联结构满足层与层之间空间分辨率逐层递减，每层提取出来的特征平面数量递增，使得 CNN 能够有效地检测更多特征信息。

卷积层中，前一层的特征图与一个可学习卷积核进行卷积运算，卷积结果经过激活函数后的输出形成下一层特征图的神经元，从而构成下一层对应某一种特征的特征图。使用不同的卷积核进行卷积可提取前一层特征图的不同特征，这些代表不同特征的特征图共同作为下一层子采样层的输入数据。

如图 3 所示，卷积层为 C2 和 C4，卷积层与子采样层间隔出现，卷积层每一个输出特征图与前一层的特征图的卷积结果建立关系。一般地，卷积层的计算方式为：

$$X_j^l=\sigma\left(\sum_{m_j} X^{l-1} * Kernel_j^l+b_l\right) \tag{13}$$

其中，l 表示网络的层数，$Kernel$ 表示卷积核，每个特征图对应不同的卷积核，m_j 为输入特征图的一个选择，每一层有共享的偏移 b_l，$\sigma(z)=1/(1+\exp(-z))$。

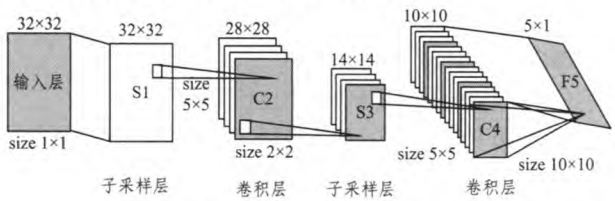

图 3　CNN 模型

子采样层的功能比较简单，通过提取局域野的特征来减小数据规模。子采样层的主要作用是降低网络的空间分辨率，从而实现畸变、位移鲁棒性。抽样层上的神经元的计算公式为：

$$X_j^l=\sigma\left(\underset{m_j}{g}(X^{l-1})+b_l\right) \tag{14}$$

其中，$g(z)$表示定义在数据 z 上的一种操作，可以定义为取子区域的最大值、平均值等。

CNN 本质上是一种输入到输出的非线性映射，它不需要输入和输出之间精确的数学表达关系，却能够有效地学习输入与输出之间的非线性映射关系，这正是判别模型的重要依据。CNN 通过用已知模式训练卷积网络，使得卷积网络获得输入与输出之间的映射关系。传统的卷积神经网络采用有监督学习方法，训练模型的样本集形如(X,Y_i)，其中 X 为输入数据，Y_i 为理想输出数据。开始训练前，所有数据都使用不同的小随机数进行初始化。“小随机数”用来约束网络不会因权值过大而陷入饱和状态；使用不同的随机数是为了约束网络，使其具备正常的学习能力[36]。

训练算法[35]主要包括 4 个步骤，分为两个阶段。

第一阶段，前向传播阶段：

(1)从样本集取出一个样本(X,Y_i)，将 X 输入网络；

(2)计算网络的实际输出 O_i。

在此阶段，信息从输入层开始前向逐层传播，直至输出层。在此过程中，网络执行如下计算：

$$O_i=f_n(\cdots(f_2(f_1(Xw_1)w_2))w_n) \tag{15}$$

其中，$f_i(\cdot)$，$i=\{1,2,\cdots,n\}$，表示 CNN 第 i 层激活函数；$w_i(\cdot)$，$i=\{1,2,\cdots,n\}$，表示 CNN 第 i 层转换矩阵。

第二阶段，反向传播阶段：

(3)计算网络实际输出 O_i 与理想输出 Y_i 的差；

(4)按照极小化误差方法调整网络权值。

假设有 K 个训练样本，使用误差的二范数作为误差的测度。

$$Error_k=\frac{1}{2}\|Y_k-O_k\|^2 \tag{16}$$

$$\theta^*=\arg\min_{\theta}\sum_{k=1}^{K}Error_k \tag{17}$$

其中，$\theta=\{w_i,b_i\}$，$i=\{1,2,\cdots,n\}$，w_i 表示第 i 层权重，b_i 表示第 i 层偏置，n 表示 CNN 层数。

这里，主要使用 BP 训练 CNN。由于 BP 训练时间过长、易于过拟合等，不少学者对其进行了改进，详见 3.3 节。

由于传统 CNN 对训练集没有合理的使用方法，模型结构难以确定，同时传统 CNN 结构固定，导致模型推广性能较差等，一些研究人员对其进行了改进，请参见 3.1 节。

2.3　混合深层结构

混合深层结构是一类由生成单元和判别单元组合而成的混合深层网络。混合深层结构将生成单元对模型强大的表达能力和判别单元高效的分类能力结合起来，从而有效地提高混合模型的判别能力。生成单元在面向高度非线性参数估计问题时，能够将模型参数初始化为近似最优解，有效地控制模型复杂度。在诸多混合深层模型训练中，生成单元首先将模型参数初始化为近似最优解，再使用判别单元全局微调，有效地解决了高度复杂问题的建模与推广问题。

3　深度学习研究的新进展

3.1　模型结构

传统 RBM 主要面向离散数据进行建模，当面向连续数据时，效果不理想。有学者[37,38]提出连续受限波尔兹曼机(CRBM)对连续数据建模，并取得了良好效果。CRBM 采用最小化对比散度(MCD)训练准则取代仅依靠 Gibbs 采样的 RBM 松弛搜索，大幅度减少了计算量。传统 RBM 学习到的特征表示是分布、非稀疏的，由于稀疏表示符合生物视觉系统特性，且能够更加有效地提取图像的高级特征，Lee[39]提出稀疏受限波尔兹曼机(SRBM)，通过把稀疏惩罚加入对数似然，来惩罚隐单元的平均激活概率偏离给定水平 p 所引起的损失。

给定训练数据 $v^{(1)},v^{(2)},\cdots,v^{(K)}$，SRBM 的目标函数为：

$$\underset{\theta}{\text{minimize}}\{\langle -\sum_{k=1}^{K}\log\sum_{h}P(v^{(k)},h^{(k)})\rangle+\langle\lambda\sum_{j=1}^{m}|p-\frac{1}{K}\sum_{k=1}^{K}E[h_j^{(k)}|v^{(k)}]|^2\rangle\},\theta=\{w_{ij},a_i,b_j\} \quad (18)$$

这里，$E[\cdot]$表示数据已知时的条件期望，λ是正则系数，p是隐单元稀疏度约束项。在学习过程中，先使用CD给出对数似然的梯度近似，再使用正则进行梯度下降，直至算法收敛。文中实验表明，稀疏RBM提取的特征与人脑V_1区简单细胞感受野很相似，堆叠稀疏RBM可以提取高级的抽象特征。实验结果表明，对于自然图像，堆叠稀疏RBM可以提取轮廓、拐角等特征，与人脑V_2区细胞的感受野特征很相似。相比之前的稀疏表示方法，堆叠稀疏RBM不但可以提取类似V_1区简单细胞的感受野特征，而且能够提取类似V_2区细胞的感受野特征。

直接学习所有隐单元的统计关系十分困难，特别是面向高维观测数据建模的场景。为简化该问题，Luo[40]将组稀疏与RBM结合，提出了稀疏组受限波尔兹曼机(SGRBM)。与SRBM相比，SGRBM可以学习到更局部化的特征。实验表明，SGRBM用于训练深层神经网络可以取得更高的识别率。

由于RBM采用无监督学习方法，学习到的特征并不完全适合分类任务，Larochelle[41,42]提出ClassRBM，使其可直接用于解决有监督学习问题，其主要思想是利用包含二值随机变量的隐单元来拟合输入特征与类标签的联合概率分布。ClassRBM使得分类过程无需额外训练分类器，保证了学习特征的判别能力，并且可以使用在线学习方法实时监测学习特征的判别性能。

针对传统CNN学习效率低、推广性能差等问题，Mrazova[43]提出了一种可增长方法来构造和训练CNN模型。与传统CNN[32]相比，它有如下改进：(1)根据输入数据的不同维数和内部结构自动调整网络拓扑结构；(2)有效处理海量高维数据；(3)分层特征提取，逐层迭代，提取高级抽象特征。实验表明，相对于传统CNN，可增长CNN速度更快，分类效果更佳。

针对传统CNN依赖有标签数据，Ranzato[44]提出一种无监督学习方法，将传统CNN和贪心逐层无监督学习算法结合起来。这种算法采用编码器-解码器结构，将CNN的特征图作为编码器和解码器，使用一种变形的期望最大(EM)算法[45]，在无标签数据集上训练CNN。实验表明，这种无监督学习方法在有标签数据稀少时能够训练出性能良好的特征提取器。

Kaiming He等[46]提出的一种空间金字塔池化方法突破了传统CNN应用于图像识别时要求固定图像尺寸的约束，改善了几乎所有基于CNN的图像分类方法。

3.2 预训练算法

2.1节指出了RBM训练效率较低。2002年，Hinton[9]提出CD算法，成功解决了RBM训练效率低的问题。区别于Gibbs采样，当使用训练数据初始化v_0时，CD仅需使用k步Gibbs采样便可有效估计最大似然。CD首先根据观测数据初始化可视单元，再根据可视层状态更新隐层状态，然后基于更新的隐层状态反向更新可视层状态，因而产生可见层的一个重构。这样，在使用SGA最大化对数似然时，各参数的更新规则为：

$$\Delta w_{ij}=\varepsilon(\langle v_i h_j\rangle_{data}-\langle v_i h_j\rangle_{recon}) \quad (19)$$

$$\Delta a_i=\varepsilon(\langle v_i\rangle_{data}-\langle v_i\rangle_{recon}) \quad (20)$$

$$\Delta b_j=\varepsilon(\langle h_j\rangle_{data}-\langle h_j\rangle_{recon}) \quad (21)$$

这里，ε是学习率，$\langle h_j\rangle_{data}$表示给定观测数据后模型的分布，$\langle h_j\rangle_{recon}$表示重构后模型的分布。目前，训练RBM主要是使用CD，如何更加有效地设置RBM模型参数，可参考文献[36]，这里不再详述。上述针对RBM单元均为二值变量提出的CD算法，可以推广到可见层和隐层单元为高斯变量等情形。此外，还有一些学者基于CD做了进一步的改进。例如Tieleman[47]提出持续对比散度(PCD)算法，该算法和CD的区别在于：(1)PCD不再使用观测数据初始化CD算法Gibbs采样后的马氏链；(2)PCD算法中学习率较小且不断衰减。

Tieleman和Hinton[48]进一步改进了PCD，通过引入辅助参数加快PCD中马氏链的混合率，提出快速持续对比散度(FPCD)算法。关于RBM的其它学习算法，有兴趣的读者可以参考文献[49]。另外，Bengio[50]提出在使用无监督分层预训练深层模型时，通过在可视层引入随机噪声，使得模型能够学习到更加鲁棒的特征。Hinton[51]提出通过阻止特征检测器共同作用来提高神经网络的性能。

3.3 全局优化算法

针对2.2节指出的BP收敛速度慢、易于过拟合等问题，一些学者进行了改进。Jacob[52]针对BP使用固定学习率不利于收敛等问题提出了Delta-Bar-Delta(DBD)算法，其中学习率根据权值的变化而改变，通过增加动量项等方法实现。它对网络性能有较大改善，是一种有效的方法。这里，增加了动量项的连接权值调整公式为：

$$\Delta w_{ij}(k+1)=\eta\frac{\partial E}{\partial w_{ij}(k)}+\alpha\Delta w_{ij}(k) \quad (22)$$

其中，$\alpha(0<\alpha<1)$为动量常数，η为学习步速，$w_{ij}(k+1)$为本次校正信号量，$\Delta w_{ij}(k)$为上次校正信号量。

有学者[53]提出用模拟退火(SA)算法结合BP，SA用于调整学习步速η。这种方法大大减少了计算量，加快了收敛速度，进一步改善了局部极小问题，应用于权值参数较多的深层神经网络时效果更加明显。也有学者[54]提出将遗传(GA)算法与BP结合，GA通过随机互换学习率的修正量，使得深层神经网络快速收敛，成功克服了陷入局部最小等问题，同时不增加额外的计算量和存储空间，克服了对参数梯度过于敏感等问题。

此外，还有一些学者对BP算法的目标函数进行了改进。例如Abid[55]提出修正反向传播(MBP)算法，该算法的误差函数是由输出单元的线性误差和非线性误差的平方和得到。Yamamoto和Nikiforuk[56]提出使用代数方法取代传统的梯度法来修改网络权值的算法。Yu[57]进一步提出一种基于前馈神经网络的广义BP算法，利用时变非梯度代数方法取代一般的代数方法。Yam和Chow[58]提出一种基于增广最小二乘算法的前馈神经网络训练方法，其中，连接最后隐层和输出层的权值由最小二乘法构造，其他网络层之间的权重通过使用一种改进的梯度下降法来计算，最小二乘形式的逐层优化方法有效改善了网络学习的延迟问题。

Andrew[59]概述了深度学习的主流优化算法。LeCun[60]使用随机梯度下降(SGD)结合BP对深层神经网络全局微调。相对于梯度下降(GD)，SGD易于实现，面向海量数据时

更加高效，但存在人工干预参数和难以并行化等问题。为克服 SGD 存在的问题，提出了具有线性搜索过程的批处理方法，比如约束存储 BFGS(L-BFGS)算法以及共轭梯度(CG)算法。L-BFGS 和 CG 相对于 SGD 更易于模型训练以及收敛性检测，同时 L-BFGS 和 CG 可以利用 GPUs 或者分布式计算等实现并行化，使得模型训练速度大幅度提高。由于L-BFGS 和 CG 需要计算全部数据的梯度实现数据更新，当面向海量数据时，小批量 L-BFGS 与 CG 学习速率快于大批量模式。文中还提到 L-BFGS 相对于 SGD 更适用于低维观测数据和 CNN。由于篇幅限制，对优化算法细节感兴趣的读者请参见文献[59]。

4 深度学习的实际应用以及面对的挑战

4.1 深度学习的实际应用

4.1.1 语音识别

Mohamed[61]在 2009 年用一个 5 层 DBN 替换高斯混合-隐马尔科夫(GMM-HMM)模型中的 GMM，并将单音素状态作为基本状态构建了一个 DBN-HMM 模型，DBN-HMM 模型的语音识别准确度超过了传统 GMM-HMM 模型的最佳状态。DBN-HMM 模型应用于 Google 语音检索、YouTube 等语音识别任务时，取得了显著的识别效果。2010 年，Mohamed[62]将最大互信息(MMI)用于训练 DBN 模型，并成功地应用于 TIMIT 语音识别任务。MMI-DBN 应用于语音识别任务时能够比 DBN-CRF 高出 5%的准确率。2011 年，Yu[17]将深层堆叠网(DSN)应用于语音识别任务，得到了比 DBN 更高的识别准确率。2012 年，Hutchinson[63]将 tensorized-DSN 应用于语音识别，取得了非常理想的识别效果。

4.1.2 图像识别

2006 年，Hinton[12]使用 DBN 和深层自动编码机在 MNIST 数据集上进行简单的图像识别和降维任务，取得了良好的实验效果，证明了深层神经网络应用于图像识别的可行性。2008 年，Taralba 等[64]将 DBNs 成功应用于产生较为完备同时又有意义的图像表达，表明 DBN 可以应用于图像检索。在应用于大规模图像检索任务中，深度学习方法取得了非常好的实验效果。2012 年，在 ImageNet LSVRC 国际测评大会上，Krizhevsky 等[65]提出的深层 CNN 模型在提供的标准数据集上取得了非常高的测评准确率，打破了当时的最高纪录。

4.1.3 其他应用

2003 年 Bengio 等[66]使用 embedding 方法将词映射到一个矢量表示空间，然后使用非线性神经网络来表示 N-Gram 模型。2008 年，Collobert 等[67]将 embedding 和多层一维卷积结构应用于 POS tagging、Chunking、Named Entity Recognition、Semantic Role Labeling 4 个典型的自然语言处理(NLP)问题。2012 年，Mikolov[68]将 RNN 模型应用到词 embedding，取得了良好的性能。2009 年，Deselaers[69]将 DBN 模型应用于多任务学习解决机器学习翻译遇到的多模态感知问题。2011 年，Sarikaya[70]将 DBNs 应用于自然语言 call-routing 任务。2011 年，Socher[71]将递归神经网络用于构造深层结构并应用于 NLP，取得了良好的应用效果。2013 年，百度尝试将 DNN 应用于广告搜索，并取得了一定效果[72]。2014 年，Hasan 等[73]将稀疏自动编码机与主动学习相结合，实现了对连续无标签视频流的在线学习，并取得了良好效果。

4.2 深度学习面对的挑战

4.2.1 理论上的挑战

目前，深度神经网络通过仿照人类大脑皮层的网状神经结构进行建模，实际构造的模型都是简化的 MNN，主要通过邻接层之间的连接来表达非线性映射关系。如果非邻接层或同层神经元之间也建立连接，能否提高深层网络的学习和表达能力？能否从神经学找到依据？能否构造一个深层神经网络，有效处理和人类智力水平相当的机器学习问题？如何构造深层神经网络，使得每一层提取特征的物理意义比较明确？相对于主流的两段式训练算法，能否找到一种完全无监督的在线训练算法？

4.2.2 建模上的挑战

如果允许非邻接层或同层神经元存在连接，深层神经网络模型应该如何构造？如何对深层模型进行改进，使输入数据只需简单预处理即可输入模型，同时能够直接处理多模态数据？如何构造深层模型，使其减轻对有标签数据的依赖？如何改造深层模型使其实现并行加速？

4.2.3 工程实现上的挑战

深层神经网络训练时间过长，易于过拟合，使得模型建模及推广能力较差，如何改造深层神经网络的训练算法，使其能够快速收敛到最优解，从而大幅度减少训练时间，而且模型推广性能良好，是一个需要解决的重要问题。如何改造深层模型，使其适用于多种类型的输入数据甚至多模态混合数据？如何改造深层模型，使其能够有效地结合 GPUs 以及分布式计算等并行加速技术？

结束语 贪心无监督逐层学习算法较好地解决了深度学习的效率问题，让深度学习更好地实现了对多类复杂智能问题的有效建模，近些年在许多领域得到了广泛的研究与应用。本文对深度学习产生的背景、理论依据、典型的深度学习模型、快速学习算法、深度学习的最新进展以及实践应用等内容作了较为详细的介绍。深度学习为人类通过模仿大脑结构解决智能问题提供了一种强有力的工具，并为其他相关领域的研究与实践提供了新技术和新思路，研究前景广阔。尤其是随着对深度学习研究的深入，各种更加高效的深度学习模型的提出以及学习算法的改进，借助深度学习来解决多种智能问题逐渐成为机器学习研究的主流，也使得深度学习在机器学习领域逐渐占据核心地位。然而，在深度学习相关理论和学习算法的研究中，仍然有许多问题值得我们进一步探讨。例如，如何提高深度学习在无监督学习场景下所提取的各层特征的辨别能力和解释能力？如何平衡深度学习网络层数与单层神经单元个数的关系，使得深度学习面向不同应用场景时提高其建模和推广性能？深度学习能否用于图像分割、多模态感知、缺失数据恢复等更加广泛的实际应用？这些问题的研究和探讨都将具有十分重要的理论和实际意义。

参考文献

[1] Haykin S. Neural Networks: A Comprehensive Foundation (second edition) [M]. N. J.: Prentice Hall, 1999

[2] Haykin S. Neural Networks & Learning Machines [M]. Upper Saddle River: Pearson Education, 2009

[3] Rosenblatt F. The perceptron: a probabilistic model for information storage and organization in the brain [J]. Psychological Re-

view,1958,65(6):386
[4] Mo D. A survey on deep learning: one small step toward AI [R]. 2012
[5] Gori M, Tesi A. On the problem of local minima in backpropagation [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1992, 14(1): 76-86
[6] Hinton G E, Osindero S, Teh Y W. A fast learning algorithm for deep belief nets [J]. Neural Computation, 2006, 18(7): 1527-1554
[7] Smolensky P. Information processing in dynamical systems: Foundations of harmony theory [M]// Parallel Distributed Processing: Explorations in the Microstructure of Cognition. 1986: 194-281
[8] Freund Y, Haussler D. Unsupervised learning of distributions of binary vectors using two layer networks [R]. Santa Cruz: Computer Research Laboratory, University of California, 1994
[9] Hinton G E. Training products of experts by minimizing contrastive divergence [J]. Neural Computation, 2002, 14(8): 1771-1800
[10] Bengio Y, Lamblin P, Popovici D, et al. Greedy layer-wise training of deep networks [M]// Advances in Neural Information Processing Systems. 2007: 153-160
[11] Poultney C, Chopra S, Cun Y L. Efficient learning of sparse representations with an energy-based model [C]// Proceedings of the 2006 Conference on Advances in Neural Information Processing Systems. 2007: 1137-1144
[12] Hinton G E, Salakhutdinov R R. Reducing the dimensionality of data with neural networks [J]. Science, 2006, 313(5786): 504-507
[13] Bengio Y. Learning deep architectures for AI(Foundations and Trends in Machine Learning)[M]. 2009
[14] Arel I, Rose D C, Karnowski T P. Deep machine learning—a new frontier in artificial intelligence research [J]. Computational Intelligence Magazine, IEEE, 2010, 5(4): 13-18
[15] Hinton G, Deng L, Yu D, et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups [J]. Signal Processing Magazine, IEEE, 2012, 29(6): 82-97
[16] 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
[17] Yu D, Deng L. Deep learning and its applications to signal and information processing [J]. Signal Processing Magazine, IEEE, 2011, 28(1): 145-154
[18] Bengio Y, Boulanger-Lewandowski N, Pascanu R. Advances in optimizing recurrent networks [C]// 2013 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP). Vancouver, Canada, 2013
[19] 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
[20] Hinton G E. Learning distributed representations of concepts [C]// Proceedings of the eighth annual conference of the cognitive science society. 1986
[21] Rumelhart D E, Hinton G E, Williams R J. Learning Representations by Back-Propagating Errors [J]. Nature, 1986, 323(6088): 533-536
[22] Glorot X, Bengio Y. Understanding the difficulty of training deep feedforward neural networks [C]// Proceedings of the International Conference on Artificial Intelligence and Statistics. 2010
[23] Cortes C, Vapnik V. Support-vector networks [J]. Machine Learning, 1995, 20(3): 273-297
[24] Lauer F, Bloch G. Incorporating prior knowledge in support vector machines for classification: A review [J]. Neurocomputing, 2008, 71(7-9): 1578-1594
[25] Barbero A, Dorronsoro J R. Momentum sequential minimal optimization: an accelerated method for support vector machine training [C]// The 2011 International Joint Conference on Neural Networks (IJCNN). San Jose, 2011: 370-377
[26] Hubel D H, Wiesel T N. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex [J]. The Journal of Physiology, 1962, 160: 106-154
[27] Bengio Y, LeCun Y. Scaling learning algorithms towards AI [J]. Large-Scale Kernel Machines, 2007, 34: 1-41
[28] Schmidhuber J. Deep Learning in Neural Networks: An Overview [J]. Neural Networks, 2014, 61: 85-117
[29] Welling M, Rosen-Zvi M, Hinton G E. Exponential family harmoniums with an application to information retrieval [M]// Advances in Neural Information Processing Systems. 2004: 1481-1488
[30] Hopfield J J. Neurons with graded response have collective computational properties like those of two-state neurons [J]. Proceedings of the National Academy of Sciences, 1984, 81(10): 3088-3092
[31] Liu J S. Monte Carlo strategies in scientific computing [M]. Springer, 2008
[32] Bouvrie J. Notes on Convolutional Neural Networks[D]. Cambridge: MIT, 2006
[33] 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
[34] 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
[35] LeCun Y, Boser B, Denker J S, et al. Backpropagation applied to handwritten zip code recognition [J]. Neural Computation, 1989, 1(4): 541-551
[36] Hinton G. A practical guide to training restricted Boltzmann machines [J]. Momentum, 2010, 9(1): 926
[37] Chen H, Murray A. A continuous restricted Boltzmann machine with a hardware-amenable learning algorithm [C]// Artificial Neural Networks—ICANN 2002. Springer, 2002: 358-363
[38] Chen H, Murray A F. Continuous restricted Boltzmann machine with an implementable training algorithm [J]. Iee Proceedings-Vision Image And Signal Processing, 2003, 150(3): 153-158
[39] Lee H, Ekanadham C, Ng A Y. Sparse deep belief net model for visual area V2 [C]// Proceedings of the Advances in neural information processing systems. 2008
[40] Luo H, Shen R, Niu C, et al. Sparse Group Restricted Boltzmann Machines [C]// Proceedings of the AAAI. 2011
[41] Larochelle H, Bengio Y. Classification using discriminative restricted Boltzmann machines [C]// Proceedings of the 25th in-

ternational conference on Machine learning. 2008:536-543

[42] Larochelle H, Mandel M, Pascanu R, et al. Learning Algorithms for the Classification Restricted Boltzmann Machine [J]. Journal of Machine Learning Research, 2012, 13:643-669

[43] Mrazova I, Kukacka M. Image Classification with Growing Neural Networks [J]. International Journal of Computer Theory & Engineering, 2013, 5(3):422-427

[44] Ranzato M, Huang F J, Boureau Y L, et al. Unsupervised learning of invariant feature hierarchies with applications to object recognition[C]// IEEE Conference on Computer Vision and Pattern Recognition. 2007:1429-1436

[45] Dempster A P, Laird N M, Rubin D B. Maximum Likelihood From Incomplete Data Via Em Algorithm [J]. Journal of the Royal Statistical Society Series B-Methodological, 1977, 39(1): 1-38

[46] He K, Zhang X, Ren S, et al. Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition[M]// Computer Vision—ECCV 2014: 13th European Conference, Zurich, Switzerland, Sep. 6-12, 2014, Proceedings, Part III. 2014:346-361

[47] Tieleman T. Training restricted Boltzmann machines using approximations to the likelihood gradient [C]// Proceedings of the 25th international conference on Machine learning. ACM, 2008: 1064-1071

[48] Tieleman T, Hinton G. Using fast weights to improve persistent contrastive divergence [C]// Proceedings of the 26th Annual International Conference on Machine Learning. ACM, 2009: 1033-1040

[49] Bengio Y, Courville A C, Vincent P. Unsupervised feature learning and deep learning: A review and new perspectives [Z]. CoRR abs/12065538, 2012

[50] Vincent P, Larochelle H, Bengio Y, et al. Extracting and composing robust features with denoising autoencoders [C]// Proceedings of the 25th international conference on Machine learning. ACM, 2008: 1096-1103

[51] Hinton G E, Srivastava N, Krizhevsky A, et al. Improving neural networks by preventing co-adaptation of feature detectors [J]. arXiv preprint arXiv: 12070580, 2012

[52] Jacobs R A. Increased Rates Of Convergence Through Learning Rate Adaptation [J]. Neural Networks, 1988, 1(4):295-307

[53] Sexton R S, Dorsey R E, Johnson J D. Optimization of neural networks: A comparative analysis of the genetic algorithm and simulated annealing [J]. European Journal of Operational Research, 1999, 114(3):589-601

[54] Montana D J, Davis L. Training Feedforward Neural Networks Using Genetic Algorithms [C]// IJCAI. 1989:762-767

[55] Abid S, Fnaiech F, Najim M. A fast feedforward training algorithm using a modified form of the standard backpropagation algorithm [J]. IEEE Transactions on Neural Networks, 2001, 12 (2):424-430

[56] Yamamoto Y, Nikiforuk P N. A new supervised learning algorithm for multilayered and interconnected neural networks [J]. IEEE Transactions on Neural Networks, 2000, 11(1):36-46

[57] Yu X H, Efe M O, Kaynak O. A general backpropagation algorithm for feedforward neural networks learning [J]. IEEE Transactions on Neural Networks, 2002, 13(1):251-254

[58] Yam J Y F, Chow T W S. Extended least squares based algorithm for training feedforward networks [J]. IEEE Transactions on Neural Networks, 1997, 8(3):806-810

[59] Ngiam J, Coates A, Lahiri A, et al. On optimization methods for deep learning [C]// Proceedings of the 28th International Conference on Machine Learning. 2011

[60] LeCun Y, Bottou L, Orr G B, et al. Efficient backprop [M]// Neural Networks: Tricks Of the Trade. 1998:9-50

[61] Mohamed A R, Sainath T N, Dahl G, et al. Deep Belief Networks Using Discriminative Features for Phone Recognition [C]// 2011 IEEE International Conference on Acoustics, Speech, And Signal Processing (ICASSP). 2011:5060-5063

[62] Mohamed A R, Yu D, Deng L. Investigation of Full-Sequence Training of Deep Belief Networks for Speech Recognition [C]// The 11th Annual Conference of the International Speech Communication Association 2010 (Interspeech 2010). 2010: 2850-2853

[63] Hutchinson B, Deng L, Yu D. A Deep Architecture with Bilinear Modeling of Hidden Representations: Applications To Phonetic Recognition [C]// 2012 IEEE International Conference on Acoustics, Speech And Signal Processing (ICASSP). 2012: 4805-4808

[64] Torralba A, Fergus R, Weiss Y. Small codes and large image databases for recognition [C]// 2008 IEEE Conference on Computer Vision And Pattern Recognition. 2008:2269-2276

[65] Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks [C]// Proceedings of the Advances in neural information processing systems. 2012

[66] Bengio Y, Ducharme R, Vincent P, et al. A neural probabilistic language model [J]. Journal of Machine Learning Research, 2003, 3(6):1137-1155

[67] Collobert R, Weston J. A unified architecture for natural language processing: Deep neural networks with multitask learning [C]// Proceedings of the 25th international conference on Machine learning. ACM, 2008:160-167

[68] Mikolov T. Statistical language models based on neural networks [D]. Brno University of Technology, 2012

[69] Deselaers T, Hasan S, Bender O, et al. A deep learning approach to machine transliteration [C]// Proceedings of the Proceedings of the Fourth Workshop on Statistical Machine Translation. Association for Computational Linguistics, 2009:233-241

[70] Sarikaya R, Hinton G E, Ramabhadran B. Deep Belief Nets for Natural Language Call-Routing [C]// 2011 IEEE International Conference on Acoustics, Speech, And Signal Processing. 2011: 5680-5683

[71] Socher R, Manning C D, Ng A Y. Learning continuous phrase representations and syntactic parsing with recursive neural networks [C]// Proceedings of the NIPS-2010 Deep Learning and Unsupervised Feature Learning Workshop. 2010:1-9

[72] Yu Kai, Jia Lei, Chen Yu-qiang, et al. Deep Learning: Yesterday, Today, and Tomorrow [J]. Journal of Computer Research and Development, 2013, 50(9):1799-1804 (in Chinese)
余凯, 贾磊, 陈雨强, 等. 深度学习的昨天、今天和明天[J]. 计算机研究与发展, 2013, 50(9):1799-1804

[73] Hasan M, Roy-Chowdhury A K. Continuous Learning of Human Activity Models using Deep Nets[C]// European Conference on Computer Vision. Springer International Publishing, 2014: 705-720