深度学习研究综述

尹宝才，王文通，王立春
（北京工业大学 城市交通学院 多媒体与智能软件技术北京市重点实验室，北京 100124）

摘 要： 鉴于深度学习在学术界和工业界的重要性，依据数据流向对目前有代表性的深度学习算法进行归纳和总结，综述了不同类型深度网络的结构及特点. 首先介绍了深度学习的概念；然后根据深度学习算法的结构特征，概述了前馈深度网络、反馈深度网络和双向深度网络3类主流深度学习算法的网络结构和训练方法；最后介绍了深度学习算法在不同数据处理中的最新应用现状及其发展趋势. 可以看到：深度学习在不同应用领域都取得了明显的优势，但仍存在需要进一步探索的问题，如无标记数据的特征学习、网络模型规模与训练速度精度之间的权衡、与其他方法的融合等.

关键词： 深度学习；深度神经网络；卷积神经网络；反卷积网络；深度玻尔兹曼机
中图分类号： TP 391.41 **文献标志码：** A **文章编号：** 0254－0037(2015)01－0048－12
doi： 10.11936/bjutxb2014100026

Review of Deep Learning

YIN Bao-cai, WANG Wen-tong, WANG Li-chun
(Beijing Key Laboratory of Multimedia and Intelligent Software Technology, College of Metropolitan Transportation, Beijing University of Technology, Beijing 100124, China)

Abstract: Considering deep learning's importance in academic research and industry application, this paper reviews methods and applications of deep learning. First, the concept of deep learning is introduced, and the main stream deep learning algorithms are classified into three classes: feed-forward deep networks, feed-back deep networks and bi-directional deep networks according to the architectural characteristics. Second, network architectures and training methods of the three types of deep networks are reviewed. Finally, state-of-the-art applications of mainstream deep learning algorithms is illustrated and trends of deep learning is concluded. Although deep learning algorithms outperform traditional methods in many fields, there are still many issues, such as feature learning on unlabeled data; the balance among network scale, training speed and accuracy; and model fusion.

Key words: deep learning; deep neural networks; convolutional neural network; deconvolutional network; deep Boltzmann machines

1 深度学习

深度学习是机器学习领域一个新的研究方向，近年来在语音识别、计算机视觉等多类应用中取得突破性的进展[1-20]. 其动机在于建立模型模拟人类大脑的神经连接结构，在处理图像、声音和文本这些信号时，通过多个变换阶段分层对数据特征进行描述[21-22]，进而给出数据的解释. 以图像数据为例，灵

收稿日期： 2014-09-05
基金项目： 国家自然科学基金资助项目(61390512)
作者简介： 尹宝才(1963—)，男，教授，主要从事数字多媒体技术、多功能感知技术、虚拟现实与图形学方面的研究，E-mail: ybc@bjut.edu.cn

长类的视觉系统中对这类信号的处理依次为：首先检测边缘、初始形状，然后再逐步形成更复杂的视觉形状[22]，同样地，深度学习通过组合低层特征形成更加抽象的高层表示、属性类别或特征，给出数据的分层特征表示.

深度学习之所以被称为“深度”，是相对支撑向量机(support vector machine，SVM)、提升方法(boosting)、最大熵方法等“浅层学习”方法而言的，深度学习所学得的模型中，非线性操作的层级数[21]更多. 浅层学习依靠人工经验抽取样本特征，网络模型学习后获得的是没有层次结构的单层特征[23-25]；而深度学习通过对原始信号进行逐层特征变换，将样本在原空间的特征表示变换到新的特征空间，自动地学习得到层次化的特征表示，从而更有利于分类或特征的可视化[26]. 深度学习理论的另外一个理论动机是：如果一个函数可用 k 层结构以简洁的形式表达，那么用 $k-1$ 层的结构表达则可能需要指数级数量的参数(相对于输入信号)，且泛化能力不足[21,27].

深度学习的概念最早由多伦多大学的 G. E. Hinton 等[26]于 2006 年提出，指基于样本数据通过一定的训练方法得到包含多个层级的深度网络结构的机器学习过程[21]. 传统的神经网络随机初始化网络中的权值，导致网络很容易收敛到局部最小值，为解决这一问题，Hinton 提出使用无监督预训练方法优化网络权值的初值，再进行权值微调的方法，拉开了深度学习的序幕.

深度学习所得到的深度网络结构包含大量的单一元素(神经元)，每个神经元与大量其他神经元相连接，神经元间的连接强度(权值)在学习过程中修改并决定网络的功能. 通过深度学习得到的深度网络结构符合神经网络的特征[28]，因此深度网络就是深层次的神经网络，即深度神经网络(deep neural networks，DNN).

深度神经网络是由多个单层非线性网络叠加而成的[21,29]，常见的单层网络按照编码解码情况分为 3 类：只包含编码器部分、只包含解码器部分、既有编码器部分也有解码器部分. 编码器提供从输入到隐含特征空间的自底向上的映射，解码器以重建结果尽可能接近原始输入为目标将隐含特征映射到输入空间[30]. 深度神经网络分为以下 3 类(如图 1 所示).

1) 前馈深度网络(feed-forward deep networks，FFDN)，由多个编码器层叠加而成，如多层感知机(multi-layer perceptrons，MLP)[31-32]、卷积神经网络(convolutional neural networks，CNN)[33-34]等.

2) 反馈深度网络(feed-back deep networks，FBDN)，由多个解码器层叠加而成，如反卷积网络(deconvolutional networks，DN)[30]、层次稀疏编码网络(hierarchical sparse coding，HSC)[35]等.

3) 双向深度网络(bi-directional deep networks，BDDN)，通过叠加多个编码器层和解码器层构成(每层可能是单独的编码过程或解码过程，也可能既包含编码过程也包含解码过程)，如深度玻尔兹曼机(deep Boltzmann machines，DBM)[36-37]、深度信念网络(deep belief networks，DBN)[26]、栈式自编码器(stacked auto-encoders，SAE)[38]等.

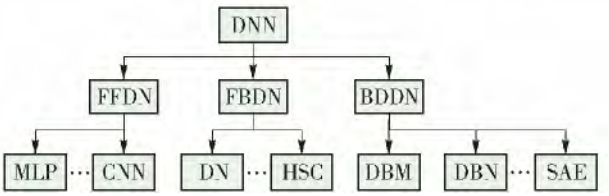


图 1　深度神经网络分类结构

Fig. 1　Classification of deep neural networks

2　前馈深度网络

前馈神经网络是最初的人工神经网络模型之一. 在这种网络中，信息只沿一个方向流动，从输入单元通过一个或多个隐层到达输出单元，在网络中没有封闭环路. 典型的前馈神经网络有多层感知机[29-30]和卷积神经网络[32-33]等.

F. Rosenblatt[39]提出的感知机是最简单的单层前向人工神经网络，但随后 M. Minsky 等[40]证明单层感知机无法解决线性不可分问题(如异或操作)，这一结论将人工神经网络研究领域引入到一个低潮期，直到研究人员认识到多层感知机可解决线性不可分问题[31-32]，以及反向传播算法与神经网络结合的研究[41-43]使得神经网络的研究重新开始成为热点. 但是由于传统的反向传播算法[41-43]具有收敛速度慢、需要大量带标签的训练数据、容易陷入局部最优等缺点，多层感知机的效果并不是十分理想.

1984 年日本学者 K. Fukushima 等基于感受野概念[45]提出的神经认知机可看作卷积神经网络的一种特例[45]，Y. Lecun 等[33-34]提出的卷积神经网络是神经认知机的推广形式. 卷积神经网络是由多个单层卷积神经网络组成的可训练的多层网络结构. 每个单层卷积神经网络包括卷积、非线性变换

和下采样 3 个阶段[46]，其中下采样阶段不是每层都必需的. 每层的输入和输出为一组向量构成的特征图(feature map)(第一层的原始输入信号可以看作一个具有高稀疏度的高维特征图). 例如，输入部分是一张彩色图像，每个特征图对应的则是一个包含输入图像彩色通道的二维数组(对于音频输入，特征图对应的是一维向量；对于视频或立体影像，对应的是三维数组)；对应的输出部分，每个特征图对应的是表示从输入图片所有位置上提取的特定特征.

2.1 单层卷积神经网络

卷积、非线性变换和下采样 3 个阶段构成的单层卷积神经网络如图 2 所示.

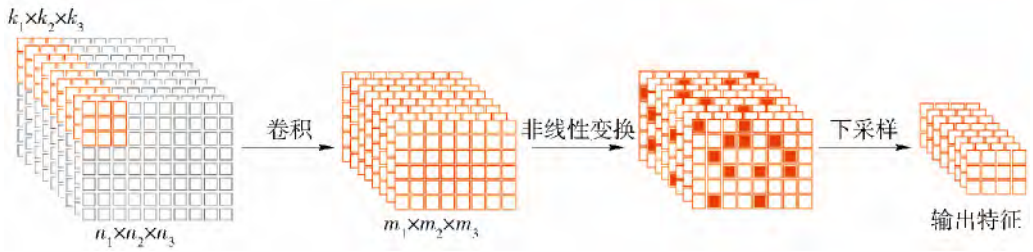


图 2 单层卷积神经网络的 3 个阶段

Fig. 2 Three phases of a single layer convolutional neural network

卷积阶段，通过提取信号的不同特征实现输入信号进行特定模式的观测. 其观测模式也称为卷积核，其定义源于由 D. H. Hubel 等[44]基于对猫视觉皮层细胞研究提出的局部感受野概念. 每个卷积核检测输入特征图上所有位置上的特定特征，实现同一个输入特征图上的权值共享[34]. 为了提取输入特征图上不同的特征，使用不同的卷积核进行卷积操作.

卷积阶段的输入是由 n_1 个 $n_2 \times n_3$ 大小的二维特征图构成的三维数组. 每个特征图记为 x_i. 该阶段的输出 y 也是个三维数组，由 m_1 个 $m_2 \times m_3$ 大小的特征图构成. 在卷积阶段，连接输入特征图 x_i 和输出特征图 y_j 的权值记为 w_{ij}，即可训练的卷积核(局部感受野[44,46])，卷积核的大小为 $k_2 \times k_3$. 输出特征图为

$$y_j = b_j + \sum_i w_{ij} * x_i \tag{1}$$

式中：* 为二维离散卷积运算符；b_j 是可训练的偏置参数.

非线性阶段，对卷积阶段得到的特征按照一定的原则进行筛选，筛选原则通常采用非线性变换的方式，以避免线性模型表达能力不够的问题.

非线性阶段将卷积阶段提取的特征作为输入，进行非线性映射 $R = h(y)$. 传统卷积神经网络中非线性操作采用 sigmoid、tanh 或 softsign 等饱和非线性(saturating nonlinearities)函数[47]，近几年的卷积神经网络中多采用不饱和非线性(non-saturating nonlinearity)函数 ReLU(rectified linear units)[1,48-50]. 在训练梯度下降时，ReLU 比传统的饱和非线性函数有更快的收敛速度，因此在训练整个网络时，训练速度也比传统的方法快很多[1]. 4 种非线性操作函数的公式为

sigmoid:

$$R = \frac{1}{1 + e^{-y}} \tag{2}$$

tanh:

$$R = \frac{e^{y} - e^{-y}}{e^{y} + e^{-y}} \tag{3}$$

softsign:

$$R = \frac{y}{1 + |y|} \tag{4}$$

ReLU:

$$R = \max(0, y) \tag{5}$$

其函数形态如图 3 所示.

下采样阶段，对每个特征图进行独立操作，通常采用平均池化(average pooling)或者最大池化(max pooling)的操作. 平均池化依据定义的邻域窗口计算特定范围内像素的均值 P_A，邻域窗口平移步长大于 1(小于等于池化窗口的大小)；最大池化则将均值 P_A 替换为最值 P_M 输出到下个阶段. 池化操作后，输出特征图的分辨率降低，但能较好地保持高分辨率特征图描述的特征. 一些卷积神经网络完全去掉下采样阶段，通过在卷积阶段设置卷积核窗口滑动步长大于 1 达到降低分辨率的目的[33,51].

2.2 卷积神经网络

如图 4 所示，将单层的卷积神经网络进行多次堆叠，前一层的输出作为后一层的输入，便构成卷积神经网络. 其中每 2 个节点间的连线，代表输入节

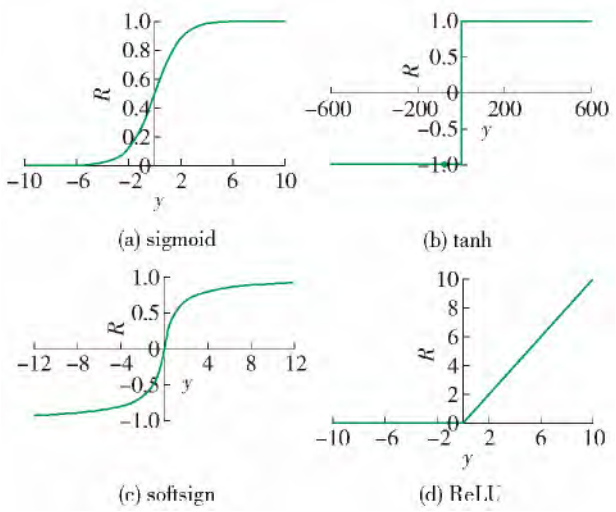


图 3　4 种非线性操作函数

Fig. 3　Four nonlinear operation functions

点经过卷积、非线性变换、下采样 3 个阶段变为输出节点，一般最后一层的输出特征图后接一个全连接层和分类器. 为了减少数据的过拟合，最近的一些卷积神经网络，在全连接层引入“Dropout”[1,52] 或“DropConnect”[53] 的方法，即在训练过程中以一定概率 P 将隐含层节点的输出值(对于“DropConnect”为输入权值)清 0，而用反向传播算法更新权值时，不再更新与该节点相连的权值. 但是这 2 种方法都会降低训练速度[1,48,53].

在训练卷积神经网络时，最常用的方法是采用反向传播法则[42-43,54] 以及有监督的训练方式，算法流程如图 5 所示. 网络中信号是前向传播的，即从输入特征向输出特征的方向传播，第 1 层的输入 X，经过多个卷积神经网络层，变成最后一层输出的特征图 O. 将输出特征图 O 与期望的标签 T 进行比较，生成误差项 E. 通过遍历网络的反向路径，将误差逐层传递到每个节点，根据权值更新公式(式(6))，更新相应的卷积核权值 w_{ij}. 在训练过程中，网络中权值的初值通常随机初始化(也可通过无监督的方式进行预训练[55])，网络误差随迭代次数的增加而减少，并且这一过程收敛于一个稳定的权值集合，额外的训练次数呈现出较小的影响.

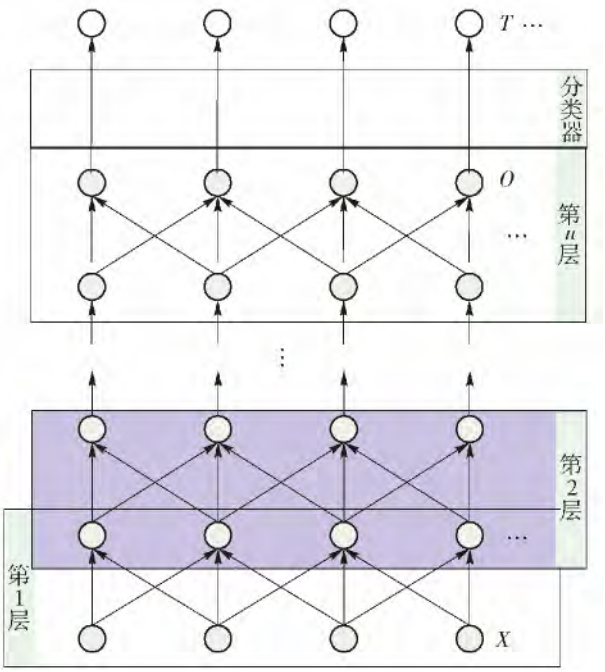


图 4　卷积神经网络模型

Fig. 4　Convolutional neural network model

对于卷积网络的任意一层 L，其第 i 个输入特征 X_i 和第 j 个输出特征 Y_j 之间的权值 w_{ij} 的更新公式[43] 为

$$\Delta w_{ij} = \alpha \delta_j X_i \tag{6}$$

当 L 层是卷积网络的最后一层时，如图 6(a) 所示，δ_j 为

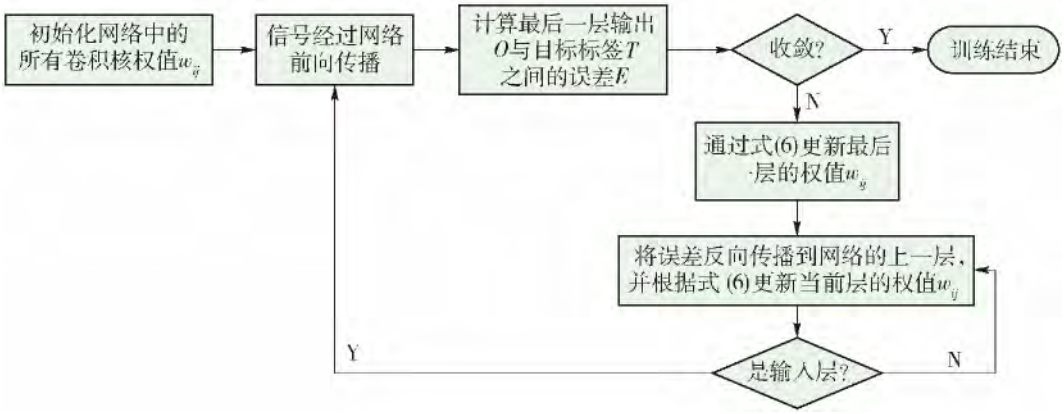


图 5　卷积神经网络训练过程

Fig. 5　Training convolutional neural network

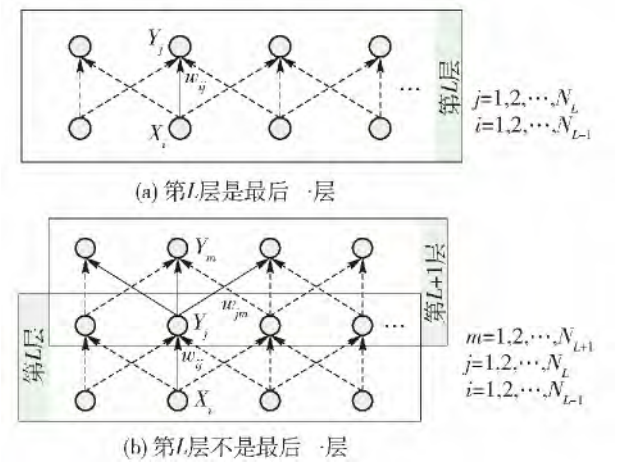
$$\delta_j = (T_j - Y_j)\, h_L'(X_i) \tag{7}$$

式中: T_j 为第 j 个预期标签; $h'(x)$ 为非线性映射函数的导数; $j = 1, 2, \cdots, N_L$.

式(6)中,当 L 层不是最后一层时,如图 6(b)所示, $L+1$ 层是其下一层,则 δ_j 为

$$\delta_j = h_L'(X_i) \sum_{m=1}^{N_{L+1}} \delta_m w_{jm} \tag{8}$$

式中: N_{L+1} 为第 $L+1$ 层输出特征的数目; $m = 1, 2, \cdots, N_{L+1}$; w_{jm} 为 L 层的第 j 个输出(作为 $L+1$ 层的第 j 个输入)与 $L+1$ 层第 m 个输出之间的权值.



(a) 第 L 层是最后一层

(b) 第 L 层不是最后一层

图 6 卷积神经网络第 L 层权值 w_{ij} 更新
(实线为与计算相关的连接关系)

Fig. 6 Update w_{ij}, weight of layer L in CNN

2.3 卷积神经网络的特点

卷积神经网络的特点在于,采用原始信号(一般为图像)直接作为网络的输入,避免了传统识别算法中复杂的特征提取和图像重建过程; 局部感受野方法获取的观测特征与平移、缩放和旋转无关. 卷积阶段利用权值共享结构减少了权值的数量进而降低了网络模型的复杂度,这一点在输入特征图是高分辨率图像时表现得更为明显. 同时,下采样阶段利用图像局部相关性的原理对特征图进行子抽样,在保留有用结构信息的同时有效地减少数据处理量.

3 反馈深度网络

与前馈网络不同,反馈网络并不是对输入信号进行编码,而是通过解反卷积[30]或学习数据集的基[35,56],对输入信号进行反解. 前馈网络是对输入信号进行编码的过程,而反馈网络则是对输入信号解码的过程.

典型的反馈深度网络有反卷积网络[30]、层次稀疏编码网络[35]等.

以反卷积网络为例,M. D. Zeiler 等[30]提出的反卷积网络模型和 Y. LeCun 等[33-34]提出的卷积神经网络思想类似,但在实际的结构构件和实现方法上有所不同. 卷积神经网络是一种自底向上的方法,该方法的每层输入信号经过卷积、非线性变换和下采样 3 个阶段处理,进而得到多层信息. 相比之下,反卷积网络模型的每层信息是自顶向下的,组合通过滤波器组学习得到的卷积特征来重构输入信号. 层次稀疏编码网络和反卷积网络非常相似,只是在反卷积网络中对图像的分解采用矩阵卷积的形式,而在稀疏编码中采用矩阵乘积的方式[35].

3.1 单层反卷积网络

反卷积网络是通过先验学习,对信号进行稀疏分解和重构的正则化方法. 图 7 所示是一个单层反卷积网络模型,输入信号 y 由 K_0 个特征通道 y_1, $y_2, \cdots, y_{K_0}$ 组成,其中任意一个通道 y_c 可看作 K_1 个隐层特征图 z_k 与滤波器组 $f_{k,c}$(个数为 $K_0 \times K_1$)的卷积.

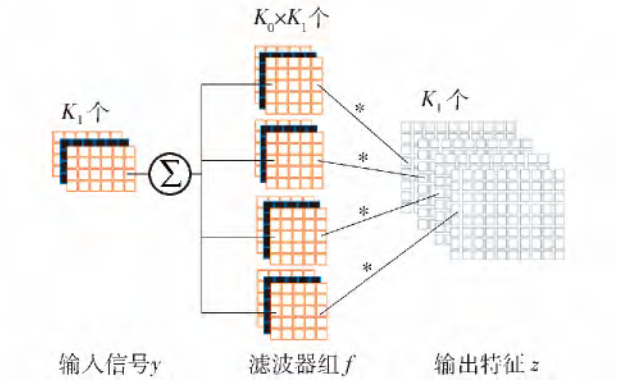


图 7 单层反卷积模型

Fig. 7 Single layer of deconvolutional net

$$\sum_{k=1}^{K_1} z_k * f_{k,c} = y_c \tag{9}$$

由于式(9)是一个欠定(未知数的个数多于方程个数)的函数,为了求得其唯一解,需要引入一个关于特征图 z_k 的正则项,且该正则项使得特征图 z_k 趋于稀疏. 于是代价函数为

$$C_1(y) = \frac{\lambda}{2} \sum_{c=1}^{K_2} \left\| \sum_{k=1}^{K_1} z_k * f_{k,c} - y_c \right\|_2^2 + \sum_{k=1}^{K_1} |z_k|^p \tag{10}$$

式中: 第 1 项为输入图像与重建结果的误差; 第 2 项为特征图的稀疏程度,为 p 范数,一般取 $p = 1$; λ 为平衡重建误差和特征图稀疏度的权重系数.

3.2　反卷积网络

通过将 3.1 节所述单层反卷积网络进行多层叠加，可得到反卷积网络，如图 8 所示. 多层模型中，在学习滤波器组的同时进行特征图的推导，第 L 层的特征图和滤波器是由第 $L-1$ 层的特征图通过反卷积计算分解获得.

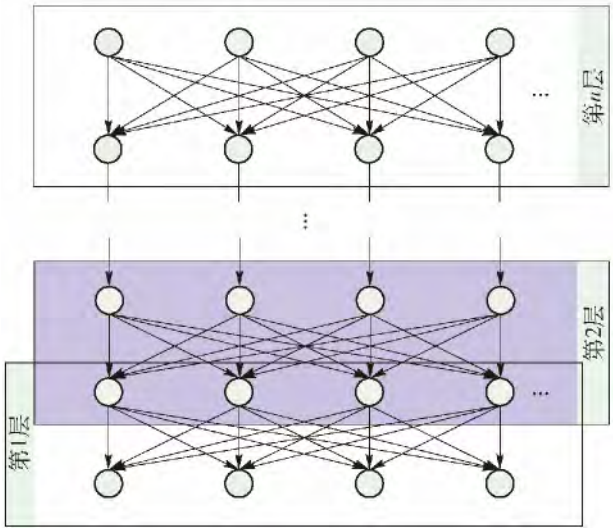


图 8　反卷积网络模型

Fig. 8　Decovolutional network model

反卷积网络训练时，使用一组不同的信号 $y=\{y^1,y^2,\cdots,y^l\}$，求解 $\mathrm{argmin}_{f,z}C_l(y)$，利用式(11)，进行滤波器组 f 和特征图 z 的迭代交替优化[30]. 训练从第 1 层开始，采用贪心算法，逐层向上进行优化，各层间的优化是独立的.

在反卷积网络中，单层网络的代价函数(为当前层所有输入信号的代价函数之和)为

$$C_l(y)=\frac{\lambda}{2}\sum_{i=1}^{l}\sum_{c=1}^{K_{l-1}}\left\|\sum_{k=1}^{K_l}g_{k,c}^{l}(z_{k,i}^{i}*f_{k,c}^{l})-z_{c,l-1}^{i}\right\|_2^2+\sum_{i=1}^{l}\sum_{k=1}^{K_l}|z_{k,l}^{i}|^p \tag{11}$$

式中第 1 项为前一层与当前层重建目标的误差. 其中：$z_{k,l}^{i}$ 是当前层的特征图；$f_{k,c}^{l}$ 是当前层的滤波器组；$z_{c,l-1}^{i}$ 是前一层的特征图；$g_{k,c}^{l}$ 表示同一层中输入特征图与输出特征图之间的连通情况，是一个固定的二值矩阵. 通常假定第 1 层是全连接的，后边的层为稀疏连接. 第 2 项为特征图的稀疏程度；λ 是平衡重建误差和特征图稀疏度的权重系数.

3.3　反卷积网络的特点

反卷积网络的特点在于，通过求解最优化输入信号分解问题计算特征，而不是利用编码器进行近似，这样能使隐层的特征更加精准，更有利于信号的分类或重建.

4　双向深度网络

双向网络由多个编码器层和解码器层叠加形成，每层可能是单独的编码过程或解码过程，也可能同时包含编码过程和解码过程. 双向网络的结构结合了编码器和解码器 2 类单层网络结构，双向网络的学习则结合了前馈网络和反馈网络的训练方法，通常包括单层网络的预训练和逐层反向迭代误差 2 个部分，单层网络的预训练多采用贪心算法：每层使用输入信号 I_L 与权值 w 计算生成信号 I_{L+1} 传递到下一层，信号 I_{L+1} 再与相同的权值 w 计算生成重构信号 I_L' 映射回输入层，通过不断缩小 I_L 与 I_L' 间的误差，训练每层网络；网络结构中各层网络结构都经过预训练之后，再通过反向迭代误差对整个网络结构进行权值微调. 其中单层网络的预训练是对输入信号编码和解码的重建过程，这与反馈网络训练方法类似；而基于反向迭代误差的权值微调与前馈网络训练方法类似.

典型的双向深度网络有深度玻尔兹曼机[36-37]、深度信念网络[26]、栈式自编码器[38]等.

以深度玻尔兹曼机为例，深度玻尔兹曼机由 R. Salakhutdinov 等[36]提出，它由多层受限玻尔兹曼机(restricted Boltzmann machine，RBM)[57-59]叠加构成.

4.1　受限玻尔兹曼机

玻尔兹曼机(Boltzmann machine，BM)是一种随机的递归神经网络，由 G. E. Hinton 等[60-62]提出，是能通过学习数据固有内在表示、解决复杂学习问题的最早的人工神经网络之一. 玻尔兹曼机由二值神经元构成，每个神经元只取 0 或 1 两种状态，状态 1 代表该神经元处于激活状态，0 表示该神经元处于抑制状态. 然而，即使使用模拟退火算法，这个网络的学习过程也十分慢.

Hinton 等提出的受限玻尔兹曼机[57-59]去掉了玻尔兹曼机同层之间的连接，从而大大提高了学习效率. 受限玻尔兹曼机分为可见层 v 以及隐层 h，可见层和隐层的节点通过权值 w 相连接，2 层节点之间是全连接，同层节点间互不相连，如图 9 所示.

受限玻尔兹曼机一种典型的训练方法如图 10 所示，首先随机初始化可见层，然后在可见层与隐层之间交替进行吉布斯采样：用条件分布概率 $P(h|v)$ 计算隐层；再根据隐层节点，同样用条件分布概率 $P(v|h)$ 来计算可见层；重复这一采样过程直到可见

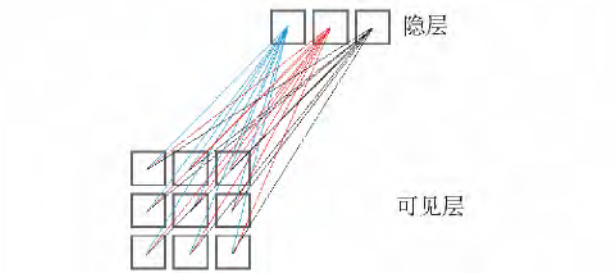


图 9 受限玻尔兹曼机(单层深度玻尔兹曼机)

Fig. 9 Restricted Boltzmann machine (single layer of deep Boltzmann machines)

层和隐层达到平稳分布. 而 Hinton 提出了一种快速算法,称作对比离差(contrastive divergence, CD)学习算法[37,59,63]. 这种算法使用训练数据初始化可见层,只需迭代 k 次上述采样过程(即每次迭代包括从可见层更新隐层,以及从隐层更新可见层),就可获得对模型的估计(通常 $k=1$).

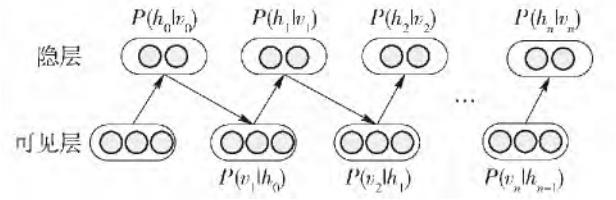


图 10 受限玻尔兹曼机的训练过程

Fig. 10 Training procedure of restricted Boltzmann machine

4.2 深度玻尔兹曼机

将多个受限玻尔兹曼机堆叠,前一层的输出作为后一层的输入,便构成了深度玻尔兹曼机,如图 11 所示. 网络中所有节点间的连线都是双向的.

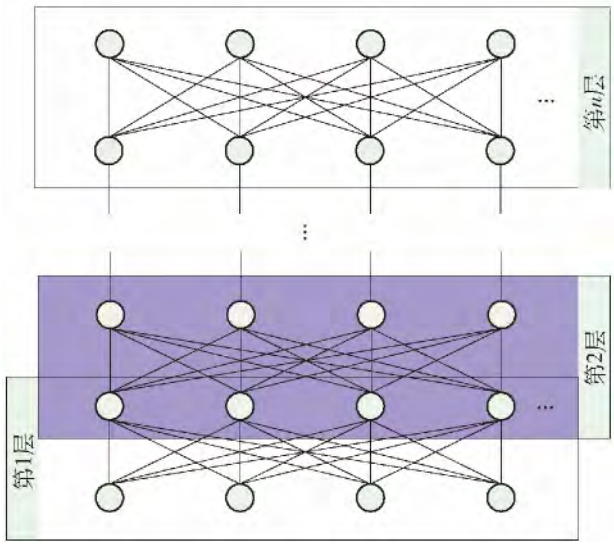


图 11 深度玻尔兹曼机

Fig. 11 Deep Boltzmann machines

深度玻尔兹曼机训练分为 2 个阶段: 预训练阶段和微调阶段,如图 12 所示.

在预训练阶段,采用无监督的逐层贪心训练方法来训练网络每层的参数,即先训练网络的第 1 个隐含层,然后接着训练第 2,3,…个隐含层,最后用这些训练好的网络参数值作为整体网络参数的初始值. 预训练之后,将训练好的每层受限玻尔兹曼机叠加形成深度玻尔兹曼机,利用有监督的学习对网络进行训练(一般采用反向传播算法).

由于深度玻尔兹曼机随机初始化权值以及微调阶段采用有监督的学习方法,这些都容易使网络陷入局部最小值. 而采用无监督预训练的方法,有利

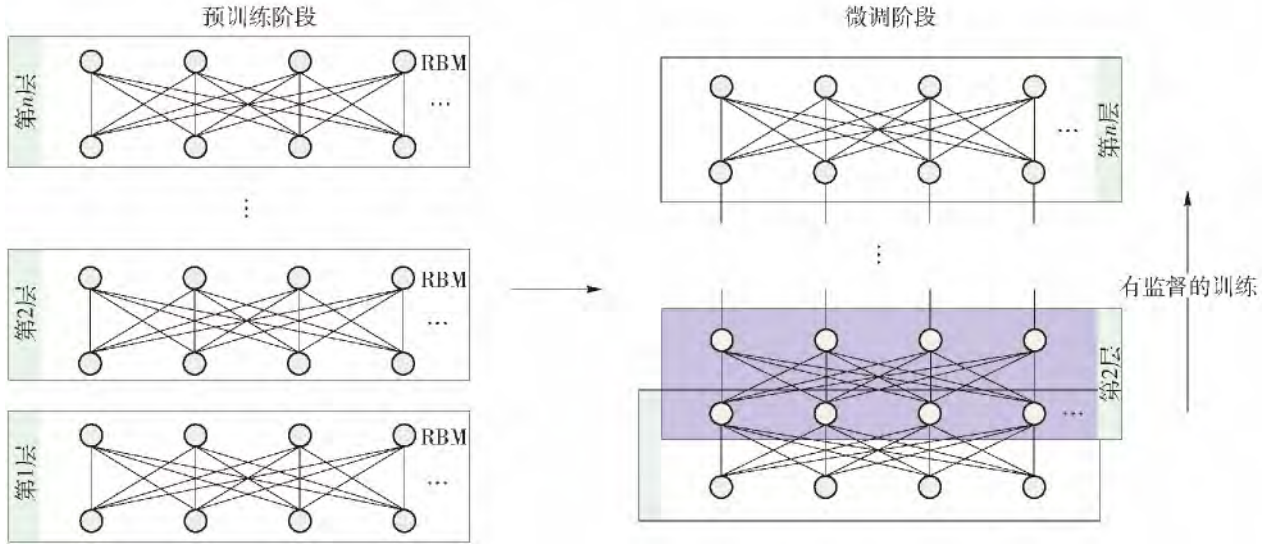


图 12 深度玻尔兹曼机逐层贪心训练方法

Fig. 12 Greedy layer-wise pre-training of DBM

于避免陷入局部最小值问题[64].

5　深度学习应用

深度学习目前在很多领域都优于过去的方法，下面根据所处理数据类型的不同，对深度学习的应用进行介绍.

5.1　深度学习在语音识别、合成及机器翻译中的应用

微软研究人员使用深度信念网络对数以千计的 senones(一种比音素小很多的建模单元)直接建模，提出了第 1 个成功应用于大词汇量语音识别系统的上下文相关的深层神经网络-隐马尔可夫混合模型(CD-DNN-HMM)[2]，比之前最领先的基于常规 CD-GMM-HMM 的大词汇量语音识别系统相对误差率减少 16% 以上.

随后又在含有 300 h 语音训练数据的 Switchboard 标准数据集上对 CD-DNN-HMM 模型进行评测[65]. 基准测试字词错误率为 18.5%，与之前最领先的常规系统相比，相对错误率减少了 33%.

H. Zen 等[3]提出一种基于多层感知机的语音合成模型. 该模型先将输入文本转换为一个输入特征序列，输入特征序列的每帧分别经过多层感知机映射到各自的输出特征，然后采用文献[66]中的算法，生成语音参数，最后经过声纹合成生成语音. 训练数据包含由一名女性专业演讲者以美国英语录制的 3.3 万段语音素材，其合成结果的主观评价和客观评价均优于基于 HMM 方法的模型.

K. Cho 等[67]提出一种基于循环神经网络(recurrent neural network，RNN)的向量化定长表示模型(RNNenc 模型)，应用于机器翻译. 该模型包含 2 个 RNN，一个 RNN 用于将一组源语言符号序列编码为一组固定长度的向量，另一个 RNN 将该向量解码为一组目标语言的符号序列.

在该模型的基础上，D. Bahdanau 等[4]克服了文献[67]中固定长度的缺点(固定长度是其效果提升的瓶颈)，提出了 RNNsearch 的模型. 该模型在翻译每个单词时，根据该单词在源文本中最相关信息的位置以及已翻译出的其他单词，预测对应于该单词的目标单词. 该模型包含一个双向 RNN 作为编码器，以及一个用于单词翻译的解码器. 在进行目标单词位置预测时，使用一个多层感知机模型进行位置对齐. 采用 BLEU 评价指标，RNNsearch 模型在 ACL2014 机器翻译研讨会(ACL WMT 2014)提供的英/法双语并行语料库[68]上的翻译结果评分均高于 RNNenc 模型的评分，略低于传统的基于短语的翻译系统 Moses[69](本身包含具有 4.18 亿个单词的多语言语料库). 另外，在剔除包含未知词汇语句的测试预料库上，RNNsearch 的评分甚至超过了 Moses.

5.2　深度学习在图像分类及识别中的应用

5.2.1　深度学习在大规模图像数据集中的应用

A. Krizhevsky 等[1]首次将卷积神经网络应用于 ImageNet 大规模视觉识别挑战赛(ImageNet large scale visual recognition challenge，ILSVRC)[70]中，所训练的深度卷积神经网络[1]在 ILSVRC—2012 挑战赛中，取得了图像分类和目标定位任务的第一. 其中，图像分类任务中，前 5 选项错误率为 15.3%，远低于第 2 名的 26.2% 的错误率；在目标定位任务中，前 5 选项错误率 34%，也远低于第 2 名的 50%.

在 ILSVRC—2013 比赛中，M. D. Zeiler 等[5]采用卷积神经网络的方法，对文献[1]的方法进行了改进，并在每个卷积层上附加一个反卷积层用于中间层特征的可视化[5,30]，取得了图像分类任务的第一名. 其前 5 选项错误率为 11.7%，如果采用 ILSVRC—2011 数据进行预训练，错误率则降低到 11.2%. 在目标定位任务中，P. Sermanet 等[6]采用卷积神经网络结合多尺度滑动窗口的方法，可同时进行图像分类、定位和检测，是比赛中唯一一个同时参加所有任务的队伍. 多目标检测任务中，获胜队伍的方法在特征提取阶段没有使用深度学习模型，只在分类时采用卷积网络分类器进行重打分[7].

在 ILSVRC—2014 比赛中，几乎所有的参赛队伍都采用了卷积神经网络及其变形方法[7]. 其中 GoogLeNet 小组采用卷积神经网络结合 Hebbian 理论提出的多尺度的模型，以 6.7% 的分类错误，取得图形分类“指定数据”组的第一名；CASIAWS 小组采用弱监督定位和卷积神经网络结合的方法，取得图形分类“额外数据”组的第一名，其分类错误率为 11%.

在目标定位任务中，VGG 小组在深度学习框架 Caffe 的基础上，采用 3 个结构不同的卷积神经网络进行平均评估，以 26% 的定位错误率取得“指定数据”组的第一名；Adobe 组选用额外的 2 000 类 ImageNet 数据训练分类器，采用卷积神经网络架构进行分类和定位，以 30% 的错误率，取得了“额外数据”组的第一名.

在多目标检测任务中，NUS 小组采用改进的卷积神经网络——网中网(network in network，NIN)[8]与多种其他方法融合的模型，以 37% 的平均准确率(mean average precision，mAP)取得“提供数据”组的第一名；GoogLeNet 以 44% 的平均准确率取

得“额外数据”组的第一名.

从深度学习首次应用于 ILSVRC 挑战赛并取得突出的成绩,到 2014 年挑战赛中几乎所有参赛队伍都采用深度学习方法,并将分类识错率降低到 6.7%,可看出深度学习方法相比于传统的手工提取特征的方法在图像识别领域具有巨大优势.

5.2.2 深度学习在人脸识别中的应用

基于卷积神经网络的学习方法,香港中文大学的 DeepID 项目[9] 以及 Facebook 的 DeepFace 项目[10] 在户外人脸识别(labeled faces in the wild, LFW)数据库[71] 上的人脸识别正确率分别达 97.45% 和 97.35%,只比人类识别 97.5%[72] 的正确率略低一点点. DeepID 项目采用 4 层卷积神经网络(不含输入层和输出层)结构,DeepFace 采用 5 层卷积神经网络(不含输入层和输出层,其中后 3 层没有采用权值共享以获得不同的局部统计特征)结构.

之后,采用基于卷积神经网络的学习方法,香港中文大学的 DeepID2 项目[11] 将识别率提高到了 99.15%,超过目前所有领先的深度学习[9-10] 和非深度学习算法[73] 在 LFW 数据库上的识别率以及人类在该数据库的识别率[72]. DeepID2 项目采用和 DeepID 项目类似的深度结构,包含 4 个卷积层,其中第 3 层采用 2×2 邻域的局部权值共享,第 4 层没有采用权值共享,且输出层与第 3、4 层都全连接.

5.3 深度学习在视频分类及行为识别中的应用

A. Karpathy 等[12] 基于卷积神经网络提供了一种应用于大规模视频分类上的经验评估模型,将 Sports-1M 数据集[12] 的 100 万段 YouTube 视频数据分为 487 类. 该模型使用 4 种时空信息融合方法用于卷积神经网络的训练,融合方法包括单帧(single frame)、不相邻两帧(late fusion)、相邻多帧(early fusion)以及多阶段相邻多帧(slow fusion);此外提出了一种多分辨率的网络结构,大大提升了神经网络应用于大规模数据时的训练速度. 该模型在 Sports-1M 上的分类准确率达 63.9%,相比于基于人工特征的方法(55.3%),有很大提升. 此外,该模型表现出较好的泛化能力,单独使用 slow fusion 融合方法所得模型在 UCF-101 动作识别数据集[74] 上的识别率为 65.4%,而该数据集的基准识别率为 43.9%.

S. Ji 等[13] 提出一个三维卷积神经网络模型用于行为识别. 该模型通过在空间和时序上运用三维卷积提取特征,从而获得多个相邻帧间的运动信息. 该模型基于输入帧生成多个特征图通道,将所有通道的信息结合获得最后的特征表示. 该三维卷积神经网络模型在 TRECVID 数据上优于其他方法,表明该方法对于真实环境数据有较好的效果;该模型在 KTH 数据上的表现,逊于其他方法,原因是为了简化计算而缩小了输入数据的分辨率.

M. Baccouche 等[14] 提出一种时序的深度学习模型,可在没有任何先验知识的前提下,学习分类人体行为. 模型的第一步,是将卷积神经网络拓展到三维,自动学习时空特征. 接下来使用 RNN 方法训练分类每个序列. 该模型在 KTH 上的测试结果优于其他已知深度模型,KTH1 和 KTH2 上的精度分别为 94.39% 和 92.17%.

事实上,深度学习的应用远不止这些,但是本文只是分别从数据的维度上(音频文本,一维;图像,二维;视频,三维)对深度学习的典型应用进行详细介绍,目的在于突出深度学习带来的优越性能以及其对不同数据的应用能力. 其他应用还包括图像超分辨率重建[15-16]、纹理识别[17]、行人检测[18]、场景标记[19]、门牌识别[20] 等.

6 深度学习的问题及趋势

深度学习算法在计算机视觉(图像识别、视频识别等)和语音识别中的应用,尤其是大规模数据集下的应用取得突破性的进展,但仍有以下问题值得进一步研究:

1) 无标记数据的特征学习. 目前,标记数据的特征学习仍然占据主导地位[1,7],而真实世界存在着海量的无标记数据,将这些无标记数据逐一添加人工标签,显然是不现实的. 所以,随着数据集和存储技术的发展,必将越来越重视对无标记数据的特征学习,以及将无标记数据进行自动添加标签技术的研究.

2) 模型规模与训练速度、训练精度之间的权衡. 一般地,相同数据集下,模型规模越大,训练精度越高,训练速度会越慢. 例如一些模型方法采用 ReLU 非线性变换、GPU 运算,在保证精度的前提下,往往需要训练 5~7 d[1,4]. 虽然离线训练并不影响训练之后模型的应用,但是对于模型优化,诸如模型规模调整、超参数设置、训练时调试等问题,训练时间会严重影响其效率. 故而,如何在保证一定的训练精度的前提下,提高训练速度,依然是深度学习方向研究的课题之一.

3) 与其他方法的融合. 从上述应用实例中可发现,单一的深度学习方法,往往并不能带来最好的

效果，通常融合其他方法或多种方法进行平均打分，会带来更高的精确率. 因此，深度学习方法与其他方法的融合，具有一定的研究意义.

参考文献：

[1] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. Imagenet classification with deep convolutional neural networks [C] // Advances in Neural Information Processing Systems. Red Hook, NY: Curran Associates, 2012: 1097-1105.

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

[3] ZEN H, SENIOR A, SCHUSTER M. Statistical parametric speech synthesis using deep neural networks [C] // Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on. Piscataway, NJ: IEEE, 2013: 7962-7966.

[4] BAHDANAU D, CHO K, BENGIO Y. Neural machine translation by jointly learning to align and translate [J]. CoRR, 2014: abs/1409. 0473.

[5] ZEILER M D, FERGUS R. Visualizing and understanding convolutional neural networks [J]. CoRR, 2013: abs/1311. 2901.

[6] SERMANET P, EIGEN D, ZHANG X, et al. Overfeat: integrated recognition, localization and detection using convolutional networks [J]. CoRR, 2013: abs/1312. 6229.

[7] RUSSAKOVSKY O, DENG J, SU H, et al. ImageNet large scale visual recognition challenge [J]. CoRR, 2014: abs/1409. 0575.

[8] LIN M, CHEN Q, YAN S. Network in network [J]. CoRR, 2013: abs/1312. 4400.

[9] SUN Y, WANG X, TANG X. Deep learning face representation from predicting 10, 000 classes [C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2014: 1891-1898.

[10] TAIGMAN Y, YANG M, RANZATO M A, et al. Deepface: closing the gap to human-level performance in face verification [C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2014: 1701-1708.

[11] SUN Y, WANG X, TANG X. Deep learning face representation by joint identification-verification [J]. CoRR, 2014: abs/1406. 4773.

[12] KARPATHY A, TODERICI G, SHETTY S, et al. Large-scale video classification with convolutional neural networks [C] // IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ: IEEE, 2014: 1725-1732.

[13] JI S, XU W, YANG M, et al. 3D convolutional neural networks for human action recognition [J]. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 2013, 35(1): 221-231.

[14] BACCOUCHE M, MAMALET F, WOLF C, et al. Sequential deep learning for human action recognition [C] // Human Behavior Understanding. Berlin: Springer, 2011: 29-39.

[15] DONG C, LOY C C, HE K, et al. Learning a deep convolutional network for image super-resolution [C] // Computer Vision-ECCV 2014. Cham: Springer International Publishing, 2014: 184-199.

[16] CUI Z, CHANG H, SHAN S, et al. Deep network cascade for image super-resolution [C] // Computer Vision-ECCV 2014. Cham: Springer International Publishing, 2014: 49-64.

[17] BADRI H, YAHIA H, DAOUDI K. Fast and accurate texture recognition with multilayer convolution and multifractal analysis [C] // Computer Vision-ECCV 2014. Cham: Springer International Publishing, 2014: 505-519.

[18] ZENG X, OUYANG W, WANG M, et al. Deep learning of scene-specific classifier for pedestrian detection [C] // Computer Vision-ECCV 2014. Cham: Springer International Publishing, 2014: 472-487.

[19] FARABET C, COUPRIE C, NAJMAN L, et al. Learning hierarchical features for scene labeling [J]. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 2013, 35(8): 1915-1929.

[20] GOODFELLOW I J, BULATOV Y, IBARZ J, et al. Multi-digit number recognition from street view imagery using deep convolutional neural networks [J]. CoRR, 2013: abs/1312. 6082.

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

[22] SERRE T, KREIMAN G, KOUH M, et al. A quantitative theory of immediate visual recognition [J]. Progress in Brain Research, 2007, 165: 33-56.

[23] LOWE D G. Distinctive image features from scale-invariant keypoints [J]. International Journal of Computer Vision, 2004, 60(2): 91-110.

[24] DALAL N, TRIGGS B. Histograms of oriented gradients for human detection [C] // Computer Vision and Pattern Recognition, 2005. IEEE Computer Society Conference on. Piscataway, NJ: IEEE, 2005: 886-893.

[25] OJALA T, PIETIKAINEN M, MAENPAA T. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns [J]. Pattern

Analysis and Machine Intelligence, IEEE Transactions on, 2002, 24(7): 971-987.

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

[27] HÅSTAD J, GOLDMANN M. On the power of small-depth threshold circuits [J]. Computational Complexity, 1991, 1(2): 113-129.

[28] PSALTIS D, SIDERIS A, YAMAMURA A. A multilayered neural network controller [J]. IEEE Control Systems Magazine, 1988, 8(2): 17-21.

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

[30] ZEILER M D, KRISHNAN D, TAYLOR G W, et al. Deconvolutional networks [C] // Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on. Piscataway, NJ: IEEE, 2010: 2528-2535.

[31] HORNIK K, STINCHCOMBE M, WHITE H. Multilayer feedforward networks are universal approximators [J]. Neural Networks, 1989, 2(5): 359-366.

[32] GARDNER M W, DORLING S R. Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences [J]. Atmospheric Environment, 1998, 32(14/15): 2627-2636.

[33] LeCun Y, BOSER B, DENKER J S, et al. Handwritten digit recognition with a back-propagation network [C] // Advances in Neural Information Processing Systems. San Francisco, CA: Morgan Kaufmann Publishers, 1990: 396-404.

[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] YU K, LIN Y, LAFFERTY J. Learning image representations from the pixel level via hierarchical sparse coding [C] // Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on. Piscataway, NJ: IEEE, 2011: 1713-1720.

[36] SALAKHUTDINOV R, HINTON G E. Deep Boltzmann machines [C] // JMLR Workshop and Conference Proceedings Volume 5: AISTATS 2009. Brookline, MA: Microtome Publishing, 2009: 448-455.

[37] **刘建伟，刘媛，罗雄麟. 玻尔兹曼机研究进展** [J]. **计算机研究与发展**, 2014, 51(1): 1-16.
LIU Jian-wei, LIU Yuan, LUO Xiong-lin. Research and development on Boltzmann machine [J]. Journal of Computer Research and Development, 2014, 51(1): 1-16. (in Chinese)

[38] 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. New York, NY: ACM, 2008: 1096-1103.

[39] ROSENBLATT F. The perceptron: a probabilistic model for information storage and organization in the brain [J]. Psychological Review, 1958, 65(6): 386.

[40] MINSKY M, PAPERT S. Perceptrons [M]. Cambridge, MA: MIT Press, 1969: 105-110.

[41] RUMELHART D E, HINTON G E, WILLIAMS R J. Learning representations by back-propagating errors [J]. Nature, 1986, 323: 533-536.

[42] LeCun Y. Une procedure d'apprentissage pour reseau a seuil assymetrique [J]. Proceedings of Cognitiva, 1985, 85: 599-604.

[43] HINTON G E. How neural networks learn from experience [J]. Scientific American, 1992, 267(3): 145-151.

[44] 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(1): 106.

[45] FUKUSHIMA K, MIYAKE S. Neocognitron: a new algorithm for pattern recognition tolerant of deformations and shifts in position [J]. Pattern Recognition, 1982, 15(6): 455-469.

[46] LeCun Y, KAVUKCUOGLU K, FARABET C. Convolutional networks and applications in vision [C] // Circuits and Systems (ISCAS), Proceedings of 2010 IEEE International Symposium on. Piscataway, NJ: IEEE, 2010: 253-256.

[47] GLOROT X, BENGIO Y. Understanding the difficulty of training deep feedforward neural networks [C] // International Conference on Artificial Intelligence and Statistics. Brookline, MA: 2010: 249-256.

[48] DAHL G E, SAINATH T N, HINTON G E. Improving deep neural networks for LVCSR using rectified linear units and dropout [C] // Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on. Piscataway, NJ: IEEE, 2013: 8609-8613.

[49] NAIR V, HINTON G E. Rectified linear units improve restricted Boltzmann machines [C] // Proceedings of the 27th International Conference on Machine Learning (ICML-10). Madison, WI: Omnipress, 2010: 807-814.

[50] GLOROT X, BORDES A, BENGIO Y. Deep sparse rectifier networks [C] // JMLR Workshop and Conference Proceedings Volume 15: AISTATS 2011. Brookline, MA: Microtome Publishing, 2011: 315-323.

[51] SIMARD P Y, STEINKRAUS D, PLATT J C. Best practices for convolutional neural networks applied to visual document analysis [C] // Document Analysis and

Recognition，2003. Proceedings Seventh International Conference on. Washington，DC：IEEE Computer Society，2003，2：958-963.

[52] HINTON G E，SRIVASTAVA N，KRIZHEVSKY A，et al. Improving neural networks by preventing co-adaptation of feature detectors [J]. CoRR，2012：abs/1207.0580.

[53] WAN L，ZEILER M，ZHANG S，et al. Regularization of neural networks using dropconnect [C] // Proceedings of the 30th International Conference on Machine Learning. Brookline，MA：Microtome Publishing，2013，28(3)：1058-1066.

[54] BOUVRIE J. Notes on convolutional neural networks [R]. Massachusetts：Center for Biological and Computational Learning，2006：38-44

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

[56] OLSHAUSEN B A，FIELD D J. Sparse coding with an overcomplete basis set：a strategy employed by V1? [J]. Vision Research，1997，37(23)：3311-3325.

[57] SMOLENSKY P. Information processing in dynamical systems：foundations of harmony theory [M] // Rumelhart D E，McClelland J L. Parallel Distributed Processing，Cambridge，MA：MIT Press，1986：194-281.

[58] FREUND Y，HAUSSLER D. Unsupervised learning of distributions of binary vectors using two layer networks [C] // Advances in Neural Information Processing Systems. San Francisco，CA：Morgan Kaufmann Publishers，1994：912-919.

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

[60] HINTON G E，SEJNOWSKI T J. Optimal perceptual inference [C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway，NJ：IEEE，1983：448-453.

[61] HINTON G E，SEJNOWSKI T J. Analysing cooperative computation [C] // Proceedings of the Fifth Annual Conference of the Cognitive Science Society. Rochester，NY：Lawrence Erlbaum Associates，1983.

[62] HINTON G E，SEJNOWSKI T J，ACKLEY D H. Boltzmann machines：constraint satisfaction networks that learn [M]. Pennsylvania：Department of Computer Science，1984.

[63] HINTON G. A practical guide to training restricted Boltzmann machines [R]. Toronto：University of Toronto，2010.

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

[65] SEIDE F，LI G，YU D. Conversational speech transcription using context-dependent deep neural networks [C] // International Speech Communication Association. Annual Conference. 12th 2011.(Interspeech 2011). Red Hook，NY：Curran Associates，2011：437-440.

[66] TOKUDA K，YOSHIMURA T，MASUKO T，et al. Speech parameter generation algorithms for HMM-based speech synthesis [C] // Acoustics，Speech，and Signal Processing，2000. Proceedings 2000 IEEE International Conference on. Piscataway，NJ：IEEE，2000：1315-1318.

[67] CHO K，van MERRIENBOER B，GULCEHRE C，et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation [J]. CoRR，2014：abs/1406.1078.

[68] ACL 2014 Ninth Workshop on Statistical Machine Translation [DB/OL]. [2014-9-23]. http：// www.statmt.org/wmt14/translation-task.html.

[69] KOEHN P，HOANG H，BIRCH A，et al. Moses：open source toolkit for statistical machine translation [C] // Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions. Stroudsburg，PA：Association for Computational Linguistics，2007：177-180.

[70] DENG J，DONG W，SOCHER R，et al. Imagenet：a large-scale hierarchical image database [C] // Computer Vision and Pattern Recognition，2009. IEEE Conference on. Piscataway，NJ：IEEE，2009：248-255.

[71] HUANG G B，MATTAR M，BERG T，et al. Labeled faces in the wild：a database forstudying face recognition in unconstrained environments [C] // Workshop on Faces in 'Real-Life' Images：Detection，Alignment，and Recognition. Marseille：Erik Learned-Miller and Andras Ferencz and Frédéric Jurie，2008.

[72] KUMAR N，BERG A C，BELHUMEUR P N，et al. Attribute and simile classifiers for face verification [C] // Computer Vision，2009 IEEE 12th International Conference on. Piscataway，NJ：IEEE，2009：365-372.

[73] LU C，TANG X. Surpassing human-level face verification performance on LFW with GaussianFace [J]. CoRR，2014：abs/1404.3840.

[74] SOOMRO K，ZAMIR A R，SHAH M. Ucf101：a dataset of 101 human actions classes from videos in the wild [J]. CoRR，2012：abs/1212.0402.

（责任编辑　吕小红）