深度学习及其在图像物体分类与检测中的应用综述

刘 栋[1,2] 李 素[1] 曹志冬[2]
(北京工商大学计算机与信息工程学院食品安全大数据技术北京市重点实验室 北京 100048)[1]
(中国科学院自动化研究所复杂系统管理与控制国家重点实验室 北京 100190)[2]

摘 要 传统的图像物体分类与检测算法及策略难以满足图像视频大数据在处理效率、性能和智能化等方面所提出的要求。深度学习通过模拟类似人脑的层次结构建立从低级信号到高层语义的映射,以实现数据的分级特征表达,具有强大的视觉信息处理能力,成为应对这一挑战的前沿技术和国内外研究热点。首先论述了深度学习的起源、发展历程及理论体系;然后分别围绕图像物体分类和检测,总结了近年来深度学习在视觉领域的发展;最后对深度学习及其在视觉领域目前存在的诸多问题以及后续的研究方向进行了分类探讨。

关键词 深度学习,特征表达,图像物体分类,图像物体检测
中图法分类号 TP181 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2016.12.003

State-of-the-art on Deep Learning and its Application in Image Object Classification and Detection

LIU Dong[1,2] LI Su[1] CAO Zhi-dong[2]
(Beijing Key Laboratory of Big Data Technology of Food Fafety, School of Computer and Information Engineering, Beijing Technology and Business University, Beijing 100048, China)[1]
(State Key Laboratory of Complex Systems Management and Control, Institute of Automation, Chinese Academy of Sciences, Bejing 100190, China)[2]

Abstract For traditional algorithms and strategies on image object classification and detection is hard to face the Challenges from efficiency, performance and intelligent of processing of image video big data. Based on the simulation of a hierarchical structure existing in human brain, deep learning can establish the mapping between the low-level signals and the high-level semantics for achieving the hierarchical expression of data characteristic. Deep learning with powerful ablility for visual information processing becomes the cutting-edge technology and research hot spot in coping with the coming challenge. At first, in this paper the basic theory of deep learning was discussed. Then, around image object classification and detection, we respectively summarized the development of deep learning in the visual field recentely. Finally, deep learning and its current problems in the visual field and the subsequent research direction were discussed in a well-informed level.

Keywords Deep learning, Feature representations, Image object classification, Image object detection

计算机视觉理论的奠基者 Marr 把视觉信息处理归结为:"认识外部世界存在什么东西及它们在什么地方",即"what is where"[1,2];亦即基于外部感知数据[3,4]构建模拟人类视觉能力的智能系统并对目标进行判断和识别[5,6]。其涉及现实复杂场景的分析与理解,首先必需判定图像中存在什么物体(分类问题)或什么位置存在什么物体(检测问题)。因此物体分类和检测是计算机视觉研究的基本问题之一,是其他高级或复杂视觉问题的基础[2]。所以,其算法和策略获得了深入研究和广泛应用,但仍存在诸多挑战。

图像视频大数据应用领域的迅速延伸对视觉信息处理技术提出了挑战。计算机与智能终端的普及和互联网的迅猛发展[7],以及随之而来的社会化媒体尤其自媒体(包括社交网站、博客、微博、论坛、内容社区、微信等)的不断创新,使网络内容主流和用户交流媒介均呈现出由文本转向图像或视频的趋势,信息传播方式发生了巨大变革,因此图像视频大数据在网络信息过滤、智能监控[8]、刑侦、遥感测绘以及远程医疗等众多领域的研究与应用得以广泛展开。图像和视频数据内部包含巨大潜在价值的同时,其产生的速度和规模也十分惊人。据不完全统计:气象卫星遥感图像每几分钟更新一次且蕴含大量时间、空间和光谱信息;2014 年 YouTube 上传的视频时长以 72h/min 的速度增加,Facebook 用户分享的信息数目(包含图像或视频)和 Google 搜索请求次数更是以每分钟百万数量级增长[9];2015－2016 年微信日均登录账号量达5.7亿,而其用户活跃时段平均每小时所分享转发的信息(包含图像或视频)数量超过 1 亿[7]。所以,针对视觉信息处理任务的技术应具有高效率、高性能甚至智能化的特点。

到稿日期:2016-05-31 返修日期:2016-09-01 本文受国家自然科学基金项目(31101088,91546112,91224008),北京市教育委员会科技计划面上项目(KM201310011010)资助。

刘 栋(1990－),男,硕士生,主要研究方向为机器学习与数据挖掘,E-mail:ton.liu@foxmail.com;李 素(1976－),女,副教授,主要研究方向为智能空间信息处理;曹志冬(1978－),男,博士,副研究员,主要研究方向为时空数据分析与预测。

特征表达是图像物体检测和分类的关键，然而传统的特征设计需要手工完成，这种方式费力、耗时并对设计者的技术要求较高；人工设计特征虽然已获得广泛应用，但不是一个可扩展的途径。

鉴于效率和性能的权衡以及智能化等需求，深度学习凭借强大的建模和数据表征能力迅速成为计算机视觉的研究热点。深度学习通过低层信号到高层特征的函数映射，来建立学习数据内部隐含关系的逻辑层次模型，以模仿人脑的视觉认知推理过程，从而使习的特征具有更强的泛化能力和表达能力[10]。为满足图像物体分类与检测问题乃至计算机视觉领域的实际应用需要，深度学习理论不断取得突破，本文结合深度学习基本原理，对其算法、模型甚至方法的演化与创新进行了重点论述，旨在引起更多研究者的关注和探讨。

1 深度学习基本框架

深度学习源于神经学启示，具有区别于浅层模型的多隐层结构，对大数据具有更好的拟合性。

1.1 深度学习的兴起

对神经网络的深入研究客观上刺激了深度学习基本思想的形成。目前，计算机视觉研究框架内均由人工设计获得底层特征，即视觉信息处理的第一步是通过手动设计特征表达算子来完成的。但人类随时面对来自现实世界的大量视觉或听觉等感知数据，却总能轻而易举地自动获取值得关注的信息。神经学研究表明[10-12]：哺乳动物大脑并未对获取到的外部感官信号直接进行特征提取处理，而是让输入信号通过一个复杂层次网络模型，并截获数据中所隐藏的模式或规则，最终使得哺乳动物感知世界，即大脑是根据经聚合和分解过程处理后的有效信息表达来识别物体。通过文献归纳研究，这一基本思想可抽象为数学命题的形式进一步论述和推导。假设一个视觉信息处理系统由 n 层($L_1,L_2,L_3,\cdots,L_{n-1},L_n$)组成(见图 1)，层际之间存在映射法则 $f:L_i\rightarrow L_{i+1}$ ($i=1,2,3,\cdots,n$)，I 和 O 代表输入和输出，此系统的信息处理过程可表示为：$I\rightarrow L_1\rightarrow L_2\rightarrow\cdots\rightarrow L_n\rightarrow O$，且 O 等价于 I，即系统的每一层处理的输入信号都不存在信息丢失。而由数据处理不等式[13,19]：

若 $X\rightarrow Y\rightarrow Z$，那么

$$I(X;Z)\leqslant I(X;Y) \tag{1}$$

图 1 深度学习基本思想

但这种系统是不存在的。因为不存在能使从数据中获得的推理得到改善的这种对数据的优良操作，即输入 I 经过层次结构处理、传播后必然存在信息丢失。但只有输出 O 与输入 I 等价或接近时，才能获得输入 I 的一系列有效的层次特征表达，即 $L_1,\cdots,L_n$。因此，如果设计一个 n 层信息处理系统(或模型)，那么必须对模型参数进行调整。涉及实际应用时，围绕调整深度模型参数进行预学习(训练)来最小化输入和实际输出的差值，从而实现输入信息的有效分级表达，使习得的特征能更深刻地表征数据内部的复杂结构，这成为深度学习理论研究的核心之一。

2006 年，Hontin 等开创性地提出了深度学习的概念[14]。研究者引入无监督学习的方法训练深度神经网络，使其相比传统方法习得的特征更有利于数据的分类、回归分析及可视化，揭示出深度学习对于大规模、高维度数据的巨大应用潜力。于是，深度学习迅速成为机器学习领域中最为活跃的研究方向，并掀起了一场新的研究热潮。

关于机器学习发展史，学术界广泛认可：对应于浅层结构和深层结构而出现的浅层学习和深度学习掀起了机器学习两次研究浪潮。随着浅层模型的不断提出并取得了理论分析和实际应用的成功，机器学习迎来了第一次浪潮的高峰[15]，产生了包括支撑向量机(SVM)、最大熵方法(比如 LR)、高斯混合模型(GMM)、条件随机场(CRF)、含单隐层的多层感知机等经典浅层结构。但相关研究表明，对于具有复杂结构的图像、视频、语音、自然声音等数据，浅层模型会出现特征表达能力不足和特征维度过多导致维数灾难等现象，并且随着大数据时代的来临，这些问题显得越发严重，浅层学习陷入了捉襟见肘的境地，而深度学习却能有效地克服这一点。

1.2 深度学习的方法与模型

随着深度学习技术的不断突破，学术界出现了一些普遍认同的观点[16,50]：深度学习可分为无监督学习方法和监督学习方法两大类(见图 2)，且两类方法间的界限划分并非严格。随着研究的不断深入，出现了结合两者技术优势的探究[17,75]，而且在具体应用时，往往会采用两者混合的策略，如深度置信度网络的训练过程[18]。

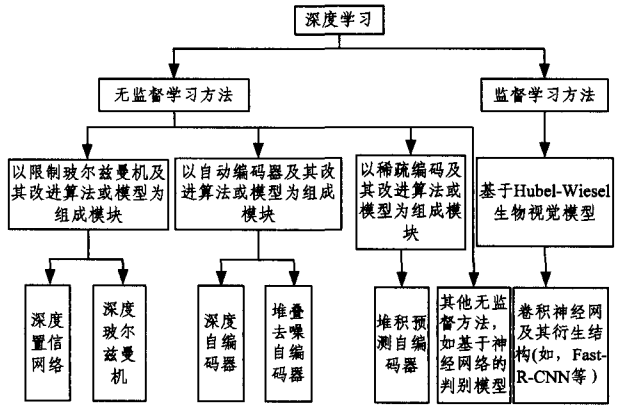

图 2 深度学习算法及模型体系结构

深度学习先驱 Hinton 和 Bengio 均指出：无监督贪婪地逐层特征学习是深度学习最初设计的核心和关键[14,18,23]。因此，根据深度模型的应用方式，可以将其分为 3 类：1)生成性结构，通常被用于表达数据的高阶相关特性或输入数据与对应类别的联合概率分布，如基于多层隐变量概率图的产生式模型、深度置信网络(DBN)等；2)区分性结构，一般用于描述数据的后验分布，以提供模式分类的区分性能，如卷积神经网络 CNN 等；3)混合型结构，其是生成性和区分性结构的融合，且输出易于优化，如基于神经网络构建的判别式模型(谷歌猫[44])等。目前，无监督学习方法训练的模型通常属于生成性结构和混合型结构，而大多监督学习方法训练的模型则属于区分性结构。

2 深度学习：无监督特征学习方法

2.1 无监督学习模型的基本架构模块

2.1.1 限制波尔兹曼机

Hinton 和 Sejnowski 根据统计力学研究提出了玻尔兹曼机(Boltzmann Machine，BM)[19]。BM 是一种能量模型(见图 3)，即参数空间的每种情况都通过能量函数这一映射法则指

向能量域中的一个能量。BM 具有优秀的无监督学习能力，但训练过程耗时且所表示的概率分布无法确切计算，甚至其概率分布的随机样本获取也很困难[20]。于是，Smolensky 提出了限制性玻尔兹曼机（Restricted Boltz Mann，machine RBM）[21]。实质上，RMB 是由 BM 简化而来（见图 3 和图 4）：任意给定可见层或隐层单元的状态时，另一层单元的激活条件独立[22]。而且 Bengio 和 Roux[23] 运用数学理论推导严格证明：只要隐层单元足够多，RBM 能拟合任意离散分布。

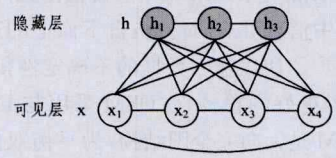

图 3　玻尔兹曼机（BM）

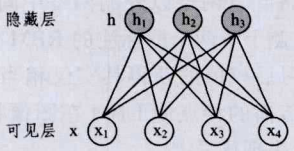

图 4　限制玻尔兹曼机（RMB）

假设一个 RBM 存在可见单元和隐单元的数量分别为 n 和 m，$\boldsymbol{x}$ 和 $\boldsymbol{h}$ 分别表示二者的状态；x_i 和 h_j 表示第 i 个可见单元和第 j 个隐单元的状态。那么，对于一组给定的状态（$\boldsymbol{x}$，$\boldsymbol{h}$），RBM 的能量函数可以定义为：

$$E(\boldsymbol{x},\boldsymbol{h}|\boldsymbol{\theta})=-\sum_{i=1}^{n}b_i x_i-\sum_{j=1}^{m}c_j h_j-\sum_{i=1}^{n}\sum_{j=1}^{m}x_i W_{ij}h_j \tag{2}$$

其中，$\boldsymbol{\theta}=\{W_{ij},b_i,c_j\}$ 且 $W_{ij},b_i,c_j\in\mathbb{R}$，$\boldsymbol{\theta}$ 是 RBM 的参数；W_{ij} 表示可见单元 i 和隐单元 j 之间的连接权重；b_i 和 c_j 分别表示可见单元 i 和隐单元 j 的偏置。当参数确定时，就可以得到（x，h）的联合分布式（3）及其边缘分布式（4）（即输入数据 x 的分布），通常式（4）也被称为似然函数。

$$P(\boldsymbol{x},\boldsymbol{h}|\boldsymbol{\theta})=\frac{e^{-E(\boldsymbol{x},\boldsymbol{h}|\boldsymbol{\theta})}}{Z(\boldsymbol{\theta})} \tag{3}$$

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

其中，$Z(\boldsymbol{\theta})=\sum_{\boldsymbol{x},\boldsymbol{h}}e^{-E(\boldsymbol{x},\boldsymbol{h}|\boldsymbol{\theta})}$ 为配分函数，需归一化处理。

为获得 $P(\boldsymbol{x}|\boldsymbol{\theta})$，需进行 2^{m+n} 次运算来确定归一化因子 $Z(\boldsymbol{\theta})$，因此通过有效计算而确定 RBM 的概率分布仍然极为困难。只能采用基于马尔可夫蒙特卡洛方法（MCMC）[24] 的吉布斯采样（Gibbs Sampling）来优化 RMB 训练（学习）过程进而获得该分布的近似值，然而 RMB 的训练效率又受采样步长所限，并且当训练数据的特征维度较高时，往往达不到预期效果。所以，2006 年 Hinton 提出对比散度算法（Contrastive Divergence，CD）[18] 来快速训练 RMB，该算法的核心是采用实验对比的方法将对数似然函数梯度的求解做近似处理；2009 年 Hinton 和 Tieleman 把 CD 算法做了进一步改进[25]。最终，基于 CD 算法的 RMB 模型获得了迅速推广和发展。

2.1.2　自动编码

自动编码器（Auto-Encoder，AE）的主要思想[26,27] 是：输入经过编码映射到特征空间（或层），特征经过解码映射回数据空间（或层），完成输入数据的重建；通过最小化重建误差的约束，学习从输入到特征空间的映射关系（见图 5）。为避免输入简单复制为重建后的输出，对 AE 增加一定的约束条件可变换为不同的形式[28]；而且，当误差函数确定时（如均方误差），如果在编码和解码过程采用相应的映射法则（如线性函数、量化编码等），那么 AE 可以等效于常见的主成分分析、K 均值聚类、稀疏编码等。图 5 中，w_y 和 w_z 表示权重参数，b_y 和 b_z 表示偏移量，$f(\cdot)$ 是激活函数，x 是输入，z 取决于参数 w、b_y、b_z，$c(x,z)$ 为误差并可以通过多种方式定义。

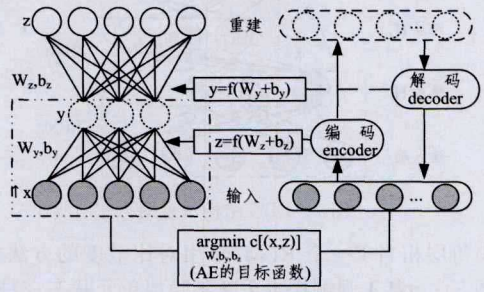

图 5　AE 核心思想示意图

2.1.3　稀疏编码

神经学研究表明，人类大脑对视觉信号的处理呈现稀疏特性。因此，稀疏编码（Sparse Coding，SC）被提出并被引入深度学习理论中。基于视觉感知稀疏特性的研究，Olshausen 和 Field 最早成形地提出稀疏编码理论[29]，并对 V1 区视觉细胞响应策略进行深入探究，进一步提出了寻找过完备基的稀疏编码算法[30]。于是，围绕寻找过完备基和系数的问题成为了稀疏编码算法的主要研究方向。为解决获取过完备基和系数过程中计算耗时等问题，Lee 和 AndrewNg 等提出一种基于递归来解决两个凸优化问题（即 L1-正则化和 L2-约束化的最小二乘问题，分别用于学习系数和基）的方法[31]。该方法在实际图像处理中学习较大规模超完备基时，其效率提升明显，且获得的稀疏编码具有端点抑制和非经典感受野抑制，给出了存在于 V1 神经元中以上两个现象的局部解释。在此基础上，Lee 和 AndrewNg 等又进一步通过增加隐层单元响应约束来缩小其范数，并对视觉皮层 V2 区响应特性进行了深入研究[32]，对其严谨详细的理论推导请查阅文献[31,32]，此处不再赘述。稀疏编码是一种能从无标签数据中发现良好过完备基向量的方法，但所获得的过完备基不唯一；如果将过完备基构成字典（dictionary），那么输入数据（如图像）可以通过字典和具有稀疏特性的系数（少数不为零）来表达[33]（见图 6）。其中，x_p 是图像块的原始像素，Dic_p 是字典，α_p 是稀疏系数，γ_p 是编码冗余，具有修正作用。

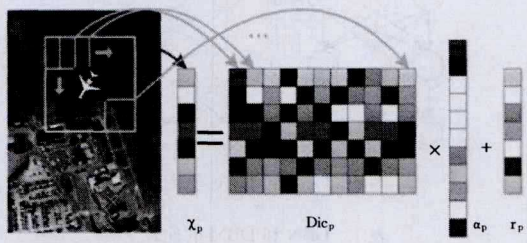

图 6　输入图像的稀疏编码表达

2.2　生成性结构

2.2.1　深信置信网络

BP 算法[34] 采用随机初始化后反向传播来调整深层网络参数，但在训练过程中很快暴露出其存在的缺陷：1）梯度扩散；2）容易收敛到局部最优，且随着模型深度的增加而越发严重；3）只能使用有标签数据作为训练集。于是，Hinton 提出了深度置信度网络（Deep Belief Nets，DBN）[18]，DBN 由一系列 RBM 逐层叠加而成（见图 7），是一种含多隐层随机变量的概率生成模型[35]。DBN 采用无监督预训练网络权值，训练

过程大致可分为 3 个阶段。

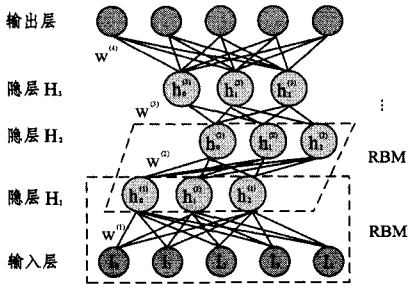

图 7　DBN 结构示意图

(1)每层相当于一个 RBM,采用对比散度的方法单独训练(见图 8):1)基于观测样本更新各隐层单元状态;2)根据 1)中已更新过的隐单元状态更新可见层单元的值;3)再根据 2)中已更新过的可见层单元的值更新隐层单元的状态;4)调整权重系数 W 及可见层和隐层的偏置 b。

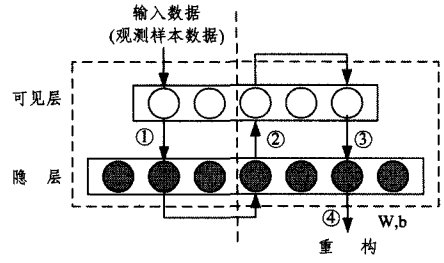

图 8　RBM 预学习(训练)示意图

(2)在相邻层之间,较低层输出作为较高层的输入(见图 7)。假设 I_i 是第一个 RBM 的输入(虚线方框所标注 RBM),其生成的隐层 $h_i^{(1)}$ ($i=0,1,2,\cdots$)被作为下一个 RBM(虚线斜框标注)的输入来训练产生 $h_i^{(2)}$;同理,可依次得到 $h_i^{(3)}$ 和 O_i (输出)以及每层的权重系数 W^j ($j=1,2,3,4,\cdots$)。DBN 只有顶层是无向图的 RBM 结构,其他各层都是自下而上的有向图结构(见图 9),这有利于对最终学习到的模式进行自上而下的概率推理(见表 1)。

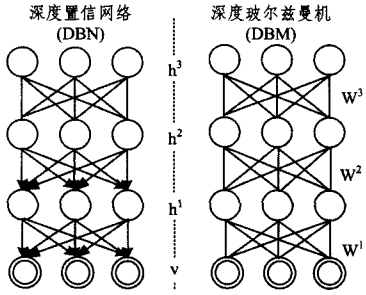

图 9　DBN 和 DBM 结构图

表 1　DBN 和 DBM 的比较

	DBN	DBM
组成单元	RBM	RBM
模型信息流传播方向	单向	双向
模型训练方法	无监督预训练	无监督预训练
概率推理方向	习得的模式可进行自上而下的概率推理传递	习得的模式可进行自下而上的近似推理和自上而下的反馈
典型应用领域	数据降维、信息检索、人体行为分析	多模数据的特征融合、图像标注和检索
模型结构	除顶层外的自上而下的有向图模型	无向完全图模型
模型分类	生成性结构(深度概率生成模型)	生成性结构(深度概率生成模型)

(3)使用 wake-sleep 算法[36]或 BP 算法结合少量有标签数据进行全局 W 优化微调。DBN 这种混合有监督和无监督方法的学习(训练)策略提高了模型的生成性能和判断力(识别能力)。

2.2.2　深度玻尔兹曼机

深度玻尔兹曼(Deep Boltzmann Machine,DBM)是另一种由 RBM 组成且结构类似于 DBN 的概率生成模型[35](见表 1)。与 DBN 不同的是,DBM 每一层边都由无向边连接(见图 9),因此 DBM 中信息可双向传播:自下而上的近似推理和自上而下的反馈[37]。所以输入数据的不确定性得以全局传播,使 DBM 获得了更好的鲁棒性;DBM 采用与 DBN 相似的训练方法,但 DBM 是无向完全图结构,为平衡双向推理所导致的变化,需将网络底部 RBM 的可视层和顶层 RBM 的隐层单元计数翻倍,并将除这两层以外的网络中间层的权值减半。如图 10 所示,左侧上下两个框标注的 RBM 代表双向推导过程的计数,预学习后权值的获得其实仅相当于一个 RBM 的情况,信息双向传播的特点使 DBM 在图像物体检索和分类以及标注问题上表现优异[38]。

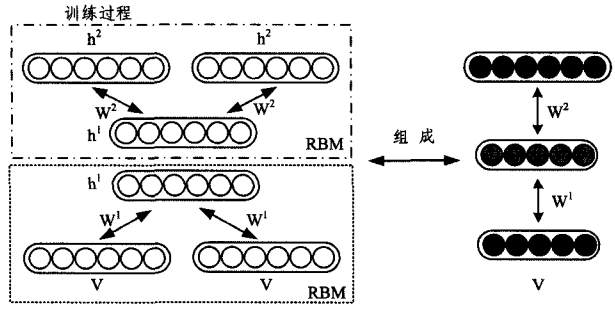

图 10　DBM 训练过程中网络权值的变化

2.2.3　深度自编码器

深度自编码器(Stacked Auto-Encoder,SAE)[27]由 AE 逐层叠加而成,是一种重要的深度模型。它通过对观测数据进行编码、解码特征表达,来获得简洁而有效的特征,并深度捕获隐藏在数据内部的规则;为充分利用数据类别、模式等隐含信息,同样要对其模型参数进行监督的微调。通常,可以通过在模型顶层加一个逻辑回归层(LR)来实现[39],另外 BP 算法[34]、牛顿插值法(Newton interpolation)[40]等均可以很好地完成参数微调。然而,现实数据普遍存在噪声,为增加 SAE 的鲁棒性,Bengio 等[41]提出去噪自编码器来替代自编码器以构建深度网络,即深度去噪自编码器。研究者们在训练数据中加入噪声,将重构信号与未加入噪声的数据进行对比,再将差异作为重构误差进行优化,使模型具备了一定的抗干扰能力。实验表明,此方法在实际应用中非常有效。

2.3　混合型结构

无监督学习还可以结合神经网络构建深度模型,属于混合型结构。作为深度学习乃至人工智能的一个重要里程碑,"谷歌猫"基于图像物体识别的研究曾轰动学术界[16]。New York Times 曾在 2012 年 6 月报道过谷歌公司的这一项目 Google Brain[42],此项目主要由著名机器学习教授 AndrewNg 负责,其技术先进性主要体现在:1)结构、模型设计。通过组合自编码和稀疏编码模块,基于 16000 个 CPU Core 并行计算节点构建机器学习阵列,该阵列是一个拥有超过 10 亿个节点的 9 层深度神经网络模型。2)数据驱动。从 YouTube 视

频中挑选 1000 万张 200×200 像素的无标记缩略图像作为训练集，训练后的该模型能从不同的目标中识别猫脸。3）基本思想。计算机科学与神经科学的结合是人工智能领域从未有过的创新。最终，这一成果[43]被发表在 2012 年国际机器学习大会上。鉴于此，AndrewNg 等继续深入研究，基于 GPU 集群设计出具有更大网络规模、更快训练速度的系统，使得再现“谷歌猫”仅需 16 台计算机，这一技术突破点燃了深度学习大规模摸拟人脑视觉功能的希望[44]。

3 深度学习：有监督特征学习方法

监督学习方法较为典型的学习模型是多层感知器（Multi-Layer Perceptrons，MLP）和卷积神经网络（Convolutional Neural Networks，ConvNets 或 CNN），且 MLP 和 CNN 属于区分性结构。

3.1 区分性结构

3.1.1 多层感知器（多层感知机）

1957 年，Rosenblatt 提出了一种用于模式分类的神经网络模型，即感知器（perceptron），而此时的感知器结构（单层感知器）仅适应线性可分数据[45]；为解决感知器非线性不可分类问题，Hinton 于 1986 年基于感知器研究引入多隐层结构来构建深层神经网络模型[46]，即多层感知机（Multi-Layer Perceptron，MLP），并与 Rumelhart 和 Williams 等完整地提出了训练深层神经网络的 BP 算法[34]。MLP 是一种前馈神经网络（Feedforward Neural Network，FNN），是目前应用最广泛的神经网络模型。实质上，深度学习源于神经网络研究，且含多隐层的多层感知机归属于深度监督学习模型。

3.1.2 卷积神经网络

ConvNets 最初的设计灵感直接源于视觉神经科学对动物大脑视觉皮层简单细胞和复杂细胞的研究[47]，即著名的 Hubel-Wiesel 生物视觉模型。实质上，ConvNets 就是通过模仿此类细胞视觉信息的处理过程而构建的多阶段 Hubel-Wiesel 结构[48]，属于典型的区分性深度模型。2015 年，为纪念人工智能的提出，深度学习领域的先驱——GeoffreyHinton，YannLeCun 和 YoshuaBengio 首次联合署名发表文献[49]。该论文对深度学习的基本原理和核心优势以及最新发展方向进行了指引性的论述，其中，监督学习方法主要围绕卷积神经网络进行了深刻阐释。ConvNets 是由多阶段组成的深度结构，根据功能的不同，可将其划分为 3 个阶段。

（1）卷积阶段（卷积层）。卷积层的作用是通过对 Hubel-Wiesel 简单细胞的模拟检测其前驱层特征的局部连接。首先，通过多卷积窗口（滤波器组）对输入图像的各种特征进行卷积操作，从而生成特征图；然后，把当前层的每个单元与其前驱层的特征图局部块通过权值（滤波器组）建立连接，并进行局部加权和非线性变换（如使用 ReLU）。其中，不同特征所使用的滤波器是不同的，因为：1）对于诸如图像等的数组数据，其值的局部组之间通常高度相关，所形成的局部主题易于区分；2）图像等的局部统计具有位置不变性，即相同主题可能出现在图像的不同部分。因此，不同位置的单元共享权值可以检测数组不同部分的相同模式。数学上，由特征映射实现的过滤操作被称为离散卷积，ConvNets 因此而得名。

（2）池化阶段（池化层）。池化层的作用是通过对 Hubel-Wiesel 复杂细胞的模拟从而进行池化操作合并具有相似语义的特征。首先，通过粗粒化各特征的位置，避免相对位置导致所形成主题的微小变化，以此实现主题的可靠检测。其次，每个典型的池化单元通常计算一个或几个特征图内局部块的最大值；而相邻单元则通过行或列平移形成的块来获取输入数据，从而降低特征表达维度，并对平移和扭曲等较小形变具有鲁棒性。

（3）全局训练阶段。根据实际需要可以串联多个卷积、非线性变换和池化阶段，通常再叠加一个全连接层（即分类器）来构建深度网络，然后通过 BP 算法等有监督地训练所有过滤器中的权值参数。

在计算机视觉领域，ConvNets 最主要的技术优势体现在：通过充分利用图像数据的层次属性，抽象或组合低层信号来构建高层特征，即局部边缘构成主题，主题聚合成部分，部分组成物体，最终使得图像中的物体易于检测或分类。因此，对于拥有复杂结构的图像、音频谱图、三维视频和体积图像等，ConvNets 具有极其优良的数据表征能力。

4 深度学习在视觉领域的研究现状

深度学习以其优异的性能引起了国内外学术界的极大关注，并迅速成为包括计算机视觉在内的众多领域的研究热点，产生了诸多研究成果。其中，学术理论以算法改进、创新及模型的优化、改良为主，而技术应用则以深度学习的硬件加速、系统研发及应用领域的拓展居多。

4.1 国外研究现状

深度学习源于神经网络研究，所以国外研究起步较早，理论发展较为成熟。因此，大致按照时间顺序，以标志性研究成果为划分对其分进行简要梳理。

随着深度学习技术不断取得突破，以优化训练深度网络为重点，出现了一些 RBM 的替代结构，如去噪自动编码器[41]等；而且，如果提供足够多的训练数据，以随机初始化权值的完全监督训练方法代替无监督预训练网络权值，在实际应用中网络模型的性能往往表现更佳[50]。同时，为克服一些图像建模、识别等特殊问题，研究者提出了一系列算法，如：三元因子玻尔兹曼机很好地解决了图像像素存在相互关系而影响识别效果的问题[51]；基于卷积限制玻尔兹曼机（Convolutional-RBM）来构建 DBN，不仅能满足计算机视觉研究对图像处理对象尺寸上的要求，而且可以自动学习目标的组成结构[17]；神经自回归分布估计器[52]的提出避开了多维数据概率分布估计需计算归一化因子所导致的计算复杂度的问题；预测稀疏（Predictive Sparse Decompo-setion，PSD）算法克服了稀疏编码推理过程复杂、耗时的问题，并消除了图像块学习造成的特征冗余，明显改善了图像分类效果[53]等。

从 2012 年 Hinton 引入 CNN 解决 ImageNet 问题并取得巨大成功以来[54]，ImageNet 成为了深度学习理论创新和技术突破的引擎（见后文的表 2、表 3），在计算机视觉领域掀起了一股研究热潮。Google，Facebook，Microsoft，IBM，yahoo，Twitter 和 Adobe 等纷纷成立研究机构将深度学习纳入实际项目进行研发，如：Microsoft 基于深度学习的视觉系统[49]；Google 在 2016 年 3 月基于深度学习技术的 AlphaGo 通过人机对弈击败人类所配置的视觉组成部分[55]等。

4.2 国内研究现状

国内深度学习研究虽然相比于国外研究起步稍晚，但近

年来发展迅速，学术界和产业界都取得了一系列阶段性成果，甚至某些技术已达国际领先水平。目前，国内深度学习研究和应用主要集中于一些科研机构和高校以及大型技术公司。

中国科学院关于深度学习的研究走在了前列。自动化研究所模式识别国家重点实验室在计算机视觉及其拓展领域取得突出成绩。其中，黄凯奇教授和谭铁牛院士等在模式识别、智能视频、机器视觉等领域基于深度学习方面的研究上取得了一系列成果，尤其在图像物体检测和分类问题上，使深度学习在视觉领域广为人知[2,8]；另外，董秋雷教授等将深度学习应用于复杂三维场景复原建模的研究也值得关注[56]。然而，在确保深度模型识别精度的基础上，如何有效提高模型训练效率及降低耗费是当前视觉领域深度学习的主要瓶颈。2014年，计算所的陈云霁教授等开创性地提出了深度学习处理器架构“寒武纪(DianNao)[57]”，DianNao是一种针对深度模型的加速器；为促进此研究成果转化为实际应用，2015年陈云霁等人又提出了指令集“电脑语(DianNaoYu)[58]”。“电脑语”是“寒武纪”的指令集，是国际首创性深度学习指令集，因此“电脑语”与“寒武纪”的结合必将大幅提高深度神经网络的训练效率，加速深度学习在视觉领域的发展。

国内高校关于视觉领域深度学习的研究也早已广泛展开。香港中文大学的汤晓鸥和王晓刚教授等注重将深度学习用于ImageNet等项目实践，进行理论和应用的相互促进式双向发展。其带领的视觉小组从2014年开始进入ImageNet竞赛，通过比赛实践有针对性地提出了一系列图像物体检测和分类算法改进及模型创新，取得了令人瞩目的成功，甚至某些研究成果已达国际领先水平(见后文表3)。清华大学学者张长水等将去噪自编码器的扩展能力和受限玻尔兹曼机的迁移能力有机结合，提出了一种混合深度神经模型，突破了单纯模型在音频谱图分类方面的局限性[59]。类似地，最近深度模型被广泛引入计算机视觉领域：文献[60]在关注语义的图像检索中构建基于卷积神经网络的多标签图像自动标注系统，其在开放数据集上与传统算法相比性能有较大提升；文献[61]则提出一种卷积神经网络模型来进行车标识别，该模型相比传统人工特征提取方法自动化程度高、识别精度高，还具有一定的鲁棒性，主要对光照变化、噪声污染甚至雾霾天气有较强抵抗能力；更值得关注的是，文献[62]借鉴稀疏编码的思想，提出了一种针对人脸识别任务的基于系数相似性的字典学习算法，在不同的人脸库数据集中取得了现有方法不可比拟的识别精度。

百度是国内最早突破深度神经网络工业化的企业，研究者们充分利用丰富的计算资源将深度神经网络模型用于声学建模、广告CTR预估、语音识别和图像识别等，获得了巨大成功[15]；另外，科大讯飞、阿里巴巴等对深度学习的研究也初见成效。

微软亚洲研究院视觉小组一直致力于机器视觉研究，其基于CNN设计的视觉系统在开放数据集上超过人眼的识别能力[63]；为解决模型深度与准确度的矛盾，该小组提出了残差学习的概念，并基于CNN构建了深达152层的模型，在ImageNet竞赛中获得巨大成功。深度残差网络通过调整深度模型内部信息流方向，有效克服了网络深度与准确度的矛盾，并具有极强的通用性(见图11)[64]。

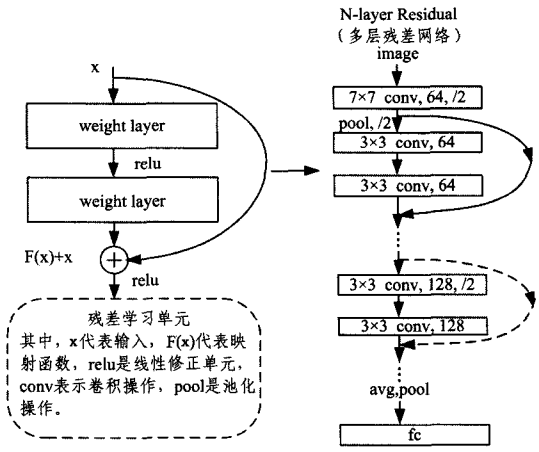

图11 残差网络示意图

5 在图像物体分类和检测中的应用

物体分类通过特征表达对图像进行全局描述，然后借助分类操作来判定图像中是否存在某类物体；而物体检测则侧重局部描述，即回答一张图像中在什么位置存在一个什么物体，所以除特征表达外，物体结构是物体检测区别于物体分类的最明显特点[2]，即特征表达是物体分类的研究主线；而结构学习则是物体检测的研究重点。

5.1 图像物体分类

物体分类关注于全局统计信息，而且对于图像数据，深度学习具有优秀的建模和特征提取能力，已被广泛应用于图像物体分类的理论分析和实际应用。

物体分类是计算机视觉的基本问题，是复杂或高级视觉问题的基础；同时，深度学习能深层表达数据内部潜藏的复杂结构和规则，因此物体分类和深度学习的契合必然被应用于复杂视觉问题的理论分析。其中，较为典型的是多维数据(3D、视频、三维手势等)分类的理论研究。针对3D物体分类，Andrew Ng等提出了一种有机组合递归和卷积神经网络(RNN和CNN)的新模型来进行特征学习和RGB-D图像分类[65]。其中，CNN层负责学习低级平移不变性特征，RNN层则结合卷积来组成高阶特征，并汇聚成特征集，与现有的架构(如双CNN)相比具有准确、快捷的特点。鉴于CNN的优良表现，Karpathy等[3]基于100万部含487个类别的YouTube视频进行大规模视频分类研究，通过多分辨率的小凹结构来加速CNN训练，使所提出的基于时空网络的模型相比传统上基于特征的模型在分类精度上具有明显提升。三维手势追踪研究的主要挑战是：1)人的手势自由度大且非常灵活；2)自然情况下，手势动作速度快，难以捕捉。为克服这些问题，Sanchez-Riera等[4]基于CNN模型训练出一系列手势作为样本来粗略预测手势的姿势和方向，并把此方法定义为一个非严格模型算法；因为，即使其违反了手势平滑动作的时间假设，此方法仍然可以获得手势的各项参数。研究表明，通过预先采集关于手势的RGB图像及相应的深度图像，基于深度学习的方法相比传统的手动特征方法取得了手势分类识别和追踪的最佳效果。

深度学习优秀的特征表达能力使得图像物体分类在实际中应用广泛：1)交通领域。文献[66]基于深度置信网络构建模型，利用方向梯度直方图(Histogram of Oriented Gradient，HOG)算子和特征很好地区分了真实交通场景图像数据中的

行人、车辆等，并且模型在光照、姿势、分辨率上都具有非常好的鲁棒性；类似地，文献[61]基于 CNN 进行车标识别，识别精度高且也具有一定的鲁棒性。2)安防领域人脸分类识别。文献[67]通过引入多任务训练机制来提升深度模型性能，明显改善了人脸分类识别效果；更值得关注的是，文献[68]基于 CNN 设计出一个 DeepID 人脸识别系统，并通过增加验证和识别信号[69]在人脸识别挑战 LFW(Labeled Faces in the Wild)数据库上取得 99.15%的识别率，首次超越同样数据集上人眼 97.52%的识别率，最终通过模型的进一步完善，使 DeepID 系统拥有非常好的遮挡鲁棒性[70]，这一成果极大地推动了具有巨大实用价值的人脸识别领域的发展。3)专业领域图像分类。基于 DBM 的医疗影像分类为克服医疗影像数据难以获取导致的训练数据不足，采用了迁移学习的方法(Transfer Learning，TL)[71]，并针对遥感图像分类构建新的特征抽取框架[33]，这些针对分类问题的构思都取得了良好的效果。

5.2 图像物体检测

有别于分类问题，检测问题更侧重于物体结构信息，物体检测的输入是包含物体的窗口，所以学术界根据窗口位置获取方法的不同，将物体检测方法划分为滑动窗口和广义霍夫投票两类。其中，滑动窗口方法已被广泛使用，并且随着以 HOG 和形变部件为代表的模型的发展，滑动窗口模型逐渐成为目前主流的物体检测方法。

深度网络训练过程需要大量的数据，只有经过充分训练的深度模型才能充分表现出优良的性能，例如，CNN 及其改进模型在一些大型开放数据集(ImageNet 数据集)上的图像物体检测任务中相比于传统方法取得了前所未有的效果[49,54]。尽管深度学习关于物体检测问题的深入研究并未广泛展开，但针对包含复杂专业信息的应用领域(如遥感、医疗等领域)图像而构建的物体检测框架却表现良好，这充分证明大数据需要深度学习。1)遥感图像物体检测研究主要围绕其所包含的丰富的地物信息展开。文献[33]提出利用高层次特征学习和弱监督学习来构建 DBM 以获得高层特征进而实现对遥感图像的物体检测，取得令人满意的效果。类似地，文献[72]提出构建深度自编码器并使用极限学习机(Extreme Learning Machine，ELM)来进行遥感图像中船舶区域快速检测的方法。在此基础上，文献[73]利用 DBN 良好的鲁棒性检测飞行器，并进一步构建并行 CNN 结构从而进行车辆识别[74]。2)医学图像理解具有信息含量大、依赖标注和特征表达、数据量不充足等特点，其物体检测研究围绕解决这些问题而展开：医学图像数据包含更加复杂的模态，所以其图像变形配准十分重要。为提高医学图像的变形配准精度，文献[75]基于据卷积深度自编码器(convolutional-SAE)构建图像变形配准框架，利用深度模型的无监督特性选择紧凑而具有良好区分性的特征，简洁、准确地描述图像模态，在变形配准精度上取得了非常好的效果；医学图像理解更需要标注和特征表达，为减少人工特征以及手动标注失误并提高效率，文献[76]提出了基于深度学习自动提取特征和多实例学习框架的方法，克服了传统方法难以解决的医疗图像精准标注和特征表达的问题；因涉及隐私，医疗图像数据获取困难从而难以满足深度模型对数据规模的需求，但可以通过基于大型公开的非医疗图像数据集训练网络模型，即通过迁移学习的方法训练深度结构来进行医学目标分类及检测[71]，这些研究为一些问题的解决提供了宝贵经验和思路借鉴。

5.3 在 ILSVRC 挑战赛中的应用

ILSVRC(ImageNet Large Scale Visual Recognition Challenge)是当前计算机视觉领域最受关注、与深度学习联系最为密切的以物体检测和分类为主的大规模任务。ILSVRC 是衡量深度学习发展的量度，且深度模型习得的特征可广泛用于计算机视觉及其他领域的研究；相反，比赛获胜的优秀深度模型将对视觉领域的发展起巨大推动作用。ILSVRC 数据库是目前计算机视觉领域内最大的数据集，规模海量，类别庞大且接近自然图像分辨率，使比赛极具挑战性。按时间顺序，深度学习成功应用于 ImageNet 比赛，可将其大致划分为以下 3 个阶段。

(1) 2012 年 ImageNet 挑战赛上深度学习被首次使用，在分类任务上取得了前所未有的成绩。Hinton 等[54]通过构建 CNN，基于 120 万张包含 1000 个类别的图像，以 84.7%的分类精度夺得 top-5 分类冠军，在分类错误率上几乎比第二名(使用 Fisher 向量编码算法)降低了一半。其模型使用了 GPU 和 ReLU(线性修正单元)，并通过 dropout(随机丢弃输入)正则化约束和大量数据生成来抑制过拟合。由于 CNN 性能卓越，2013 年几乎每个参赛小组所提出的方案都使用了 CNN，并且通过技术改进将 top-5 分类错误率降低到 11.2%，目前此技术已应用于地图标注和搜索。表 2 列出了历年 ImageNet比赛分类任务的优秀算法、模型等。

表 2 历年 ImageNet 比赛的图像分类算法

年份	所设计的模型	模型最大深度	主要分类器	优化技术	最低错误率(%)
2012	CNN(AlexNet)	8 层	SVM+多分类器	GPU、ReLU、dropout、大量数据生成	15.3
2013	改进型 CNN	大于或等于 11 层	SVM+多分类器	GPU、ReLU、dropout、正则化方法-Dropconnect、数据生成等	11.2
2014	改进型 AlexNet	20 层	SVM+多分类器	GPU(CUDA)、ReLU 等	6.66
2015	深度残差网络(DRN)	152 层	多分类器	ReLU 或 PreLU、残差学习等	3.57

(2)尽管 2014 年比赛难度增加，但关于深度学习的研究成果则进一步凸显，尤其是针对检测问题的研究成果。1)分类问题。研究者关注于 AlexNet[54]结构的改进，通过减轻过拟合导致的负面效应，并在增加网络深度的基础上，控制参数规模和计算量，最终将 top-5 分类错误率降至 6.7%(见表 2)。2)检测问题。检测任务是 ImageNet 视觉挑战中最具挑战性的项目，必须从超过 40000 张图片中检测并确定 200 类物体的具体位置，且每幅图像均包含多个不同类别的物体，表 3 列出了历年 ImageNet 比赛检测任务的优秀算法、模型等。其中，谷歌 GoogleNet 研究小组通过构建超过 20 层的深度神经网络以 43.9%的检测率夺冠；中国香港中文大学 DeepID-Net 小组首次参赛以 40.7%的检测率获第二名，经过继续研究，该小组基于 CNN 提出形变约束池化(Deformationconstrained Pooling Layer，Def-Pooling)技术，并在模型训练策略等多方面进行创新来改进其检测框架，最终以 50.3%的检测率超越谷歌取得最佳检测效果(见表 3)[77]。另外，基于区域的卷积

神经网络(Regions-CNN)[78]的方法也取得良好效果并获得广泛关注。

表3 历年ImageNet比赛的图像检测算法

年份	检测策略	模型最大深度	底层特征	分类器	检测率(%)
2012	滑动窗口	8层	多尺度HOG	SVM	33.5
2013	滑动窗口	大于或等于8层	SIFT+多特征	SVM+多分类器	22.6
2014	深度神经网络(GoogleNet)和引入形变约束池化层CNN(DeepID-Net)	20层	多特征	多分类器	43.9(GoogleNet)、50.3(DeepID-Net)
2015	深度残差网络(DRN)	152层	多特征	多分类器	62.1

(3)2015年,围绕ImageNet比赛,深度学习技术获得了更大的突破。最近,微软研究小组基于ConvNets设计的视觉系统在2012年ImageNet1000分类数据集上使分类错误率达到4.94%,首次超越相同实验中人类视觉分类识别的5.1%的错误率,研究者们将这一成果归功于两方面的技术突破[63]:1)提出并采用参数化修正线性单元(Parameter rectified Linear Unit,PreLU);2)通过对非线性特征ReLU或PreLU建模和推导,获得了新的深度模型训练策略。鉴于此,该小组进一步提出残差学习的概念并构建了深达152层的深度残差网络(Deep Residual Networks,DRN),使top-5分类错误率降至3.57%,并以绝对优势获得2015年ImageNet图像分类、定位和检测3个主要项目的冠军,因而,DRN以优良的性能表现和较强的通用性[67]迅速赢得了学术界的极大关注。残差学习通过对训练过程的重构来重定向深度结构中的信息流,有效解决了深度网络层级与准确度之间的矛盾,使模型准确度产生了质的飞跃。但值得关注的是,近期部分学者基于自相似的神经网络架构策略提出了分形网络(FractalNet),并且通过实验对比发现,FractalNet在性能上优于深度残差网络,这表明残差学习并非极深卷积神经网络取得成功的必要因素[84],再次表明深度学习在计算机视觉领域具有巨大的应用潜力。另外,中国香港中文大学欧阳万里教授等人通过集成DeepID-Net物体检测系统[77]和Fast-R-CNN[79]在2015年新增并与图像物体检测相比更为困难的视频物体检测中取得超过60%的检测率而夺冠并轰动业界,使得计算机模拟人类视觉功能向前更进一步。

ImageNet竞赛成绩的不断刷新所导致的对深度模型和算法的持续改进带来了深度学习理论和技术应用的不断突破,最终深度学习入选Science American及其中文版《环球科学》2015年度十大改变世界的创新技术之一[81]。虽然深度学习所得结果相比于人类全部的视觉能力仍有较大差距,但充分表明了其在视觉领域拥有巨大的发展潜力。

结束语 近年来深度学习理论及应用在包括计算机视觉在内的许多研究领域中已实现惊人的突破,掀起了一场革命性的研究热潮,但仍存在诸多问题和不足,大致可概括为理论问题、建模问题和工程问题。(1)理论问题。1)全局寻优问题。目前,深度学习所使用的目标函数绝大多数具有非凸性,且求解全局最优的主流方法均以梯度下降为基础展开,但非凸问题并不能通过梯度下降法彻底解决;随着深度学习理论的发展,规避局部最优而逼近全局最优的研究必将成为一个热点,因此高维参数的全局寻优是值得深入研究的问题。2)计算强度高。深层网络模型在训练阶段和测试阶段需计算确定的参数规模相当庞大。因此,探求有效降低计算复杂度的方法十分必要。如假设某一参数邻域内方差忽略不计,那么权重参数可赋值为0,这样做可明显减少计算耗费等。3)对人脑功能的模拟程度不完善。现有的深度学习技术模仿了人脑的层次结构和稀疏性,但模拟化程度不高。例如,当深度学习模型接收视觉信号刺激时,所有神经元都产生兴奋并开始计算(被激活),而人为设置的稀疏约束只能使某些神经元输出为0,但不代表该神经元处于未激活状态,因此对大脑视觉皮层稀疏响应这种非常经济的响应机制的模拟程度不够,更关键的是大脑还有复杂的视觉注意机制、多分辨率特性、联想、心理暗示等功能。4)代价函数的构建。代价函数的设计直接关系到模型最终的训练结果,是深度学习研究的核心模块。根据对已发表文献的总结,存在两种思路:误差重构和在目标函数中置入惩罚项。惩罚项的作用是防止过拟合或保证稀疏性,学术界目前还没有定论。是否能通过二者定量关系的研究统一认识是亟待解决的难题。(2)建模问题。模型结构设计是拥有巨大学术价值和应用价值的部分,“谷歌猫”项目的神经网络、DBN、DBM、CNN和SAE等深度结构各具特点,所以深度模型架构对于增强深度学习能力,促进其理论发展具有重要意义。而且,现有深度模型除底层可以通过可视化直接查看外,更高层的特征表达过程却难以知悉,使得深度结构如同一个只关注输入得到相应输出的黑盒系统,可解释性差,因此增强模型可解释性也是模型结构设计的一个重要研究方向。(3)工程问题。当前,深度学习和大数据的契合突破了训练数据量不足的瓶颈;另一方面,蕴藏于大数据内部的复杂高阶统计特性需要高容量的深度模型来深度发掘。先进的硬件设施(如高效、并行、分布式的GPU/CPU计算平台等)和新的优化技术的提出(如修正的非线性激活函数ReLU、drop-out原则等)都为海量数据的训练、处理提供了必要支撑,但如何提高海量数据的训练速度和效率以及加快发展更高速的硬件仍然是实际应用中需要不断突破的难点。

尽管图像物体检测和分类是计算机视觉的基本问题,但其仍是一个面临诸多挑战的研究领域,在实例、类别和语义层次还存在物理、语义等方面的诸多难点。在大数据驱动的深度学习背景下,在图像物体检测与分类的研究呈现出存在差异性的基础上更强调互补与统一,因此基于深度模型架构通用性和统一性的视觉识别框架是当前计算机视觉领域的发展趋势,也是目前深度学习理论在视觉领域取得更深层次技术突破的一个切入点。相信深度卷积神经网络必将对图像物体检测与分类乃至计算机视觉物体识别的深入发展做出更大的贡献。

参考文献

[1] Marr D. Vision: A Computational Investigation Into the Human-Representation and Processing of VisualInformation[M]. Cambridge: The MIT Press, 2010

[2] Huang Kai-qi, Ren Wei-qiang, Tan Tie-niu. Review on Image Object Classification and Detection[J]. Chinese Journal of Computers, 2014, 37(6): 1225-1240(in Chinese)

黄凯奇，任伟强，谭铁牛. 图像物体分类与检测算法综述[J]. 计算机学报，2014，37(6)：1225-1240

[3] Karpathy A, Toderici G, Shetty S, et al. Large-Scale Video Classification with Convolutional Neural Networks[C]// IEEE Conference on CVPR. 2014：1725-1732

[4] Sanchez-Riera J, Yuan-Sheng Hsiao, TekoingLim, et al. Arobust tracking algorithm for 3D hand gesture with rapidhandmotion through deep learning[C]// IEEE Conference on Multimedia and Expo Workshops(ICMEW). 2014：1-6

[5] Wikipedia. Computervision[EB/OL]. https://en. wikipedia. org/wiki/Computer_vision

[6] Kong B. Comparison between human vision and computer vision [J]. Nature Magazine，2002，24(1)：51-55

[7] Statistical Report on the Development of the 38th China Internet Network，CNNIC[R/OL]. [2016-08-03]. http://www. cnnic. cn/hlwfzyj/hlwxzbg/hlwtjbg/201608/t20160803_54392. htm

[8] Huang Kai-qi，Chen Xiao-tang，Tan Tie-niu，et al. Intelligent Visual Surveillance：A Review[J]. Chinese Journal of Computers，2015，38(6)：1193-1118(in Chinese)
黄凯奇，陈晓棠，谭铁牛，等. 智能视频监控技术综述[J]. 计算机学报，2015，38(6)：1193-1118

[9] James J. Data NeverSleeps 2. 0 [EB/OL]. (2014-04-23)[2014-07-20]. http://www. domo. com/blog/2014/04/data-never-sleeps-2-0

[10] Dong Yu，Hinton G，Morgan N，et al. Introduction to the Special Section on Deep Learning for Speech and Language Processing [J]. IEEE Transactions on Audio，Speech，and Language Processing，2012，20(1)：4-6

[11] Lee T，Mumford D. Hierarchical Bayesian inference in the visual cortex[J]. JOSA A，2003，20(7)：1434-1448

[12] Sun Zhi-yuan，Lu Cheng-xiang，Shi Zhong-zhi，et al. Research and Advances on Deep Learning[J]. Computer Science，2016，43(2)：1-8(in Chinese)
孙志远，鲁成祥，史忠植，等. 深度学习研究与进展[J]. 计算机科学，2016，43(2)：1-8

[13] Cover T M，Thomas J A. Elements of Information Theory(2nd Edition)[M]. New Jersey：Wiley Inter Science Publication，John Wiely & Sons，Inc. ，2006

[14] Hinton G，Salakhutdinov R. Reducing the Dimensionality of Data with Neural Networks[J]. Science，2006，313(5786)：504-507

[15] 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

[16] Guo Li-li，Ding Shi-fei. Research Progress on Deep Learning[J]. Computer Science，2015，42(5)：28-33(in Chinese)
郭丽丽，丁世飞. 深度学习研究进展[J]. 计算机科学，2015，42(5)：28-33

[17] Lee H，Grosse R，Ranganath R，et al. Convolutional deep belief-networks for scalable unsupervised learning of hierarchical representations[C]// Proceedings of the 26th Annual International-Conferenceon Machine Learning. New York，USA：ACM，2009：609-616

[18] Hinton G，Osindero S I Y. A fast learning algorithm for deepbelief nets[J]. Neural Computation，2006，18(7)：1527-1554

[19] Hinton G，Sejnowski T. Learning and relearng in Boltzmann machines[C]// Parallel Distributed Processing：Explorations in the Microstructure of Cognition. Cambridge，USA，1986：45-76

[20] Zhang Chun-xia，Ji Nan-nan，Wang Guan-wei. Restricted Boltzmann Machines[J]. Chinese Journal of Engineering Mathematics，2015，32(2)：159-173(in Chinese)
张春霞，姬楠楠，王冠伟. 受限波尔兹曼机[J]. 工程数学学报，2015，32(2)：159-173

[21] Smolensky P. Information processing in dynamical systems：foundations of harmony theory[M]// Parallel Distributed Processing：Explorations in the Microstructure of Cognition. Cambridge，USA，1986：194-281

[22] Freund Y，Haussler D. Unsupervised learning of distributions on binary vectors using two layer networks[R]. Santa Cruz：University of California，UCSC-CRL-94-25，1994

[23] Roux N，Bengio Y. Representational Power of Restricted Boltzmann Machines and Deep Belief Networks[J]. Neural Computation，2008，20(6)：1631-1649

[24] Andrieu C，de Freitas N，Doucet A，et al. An introducetion to MCMC for machine learning[J]. Machine Learning，2003，50(1/2)：5-43

[25] Tieleman T，Hinton G. Using fast weights to improve persistent contrastive divergence[C]// Proceedings of the 26th International Conference on Machine Learning. Helsinki，Finland，2008：1064-1071

[26] Hinton G，Zemel R. Autoencoders，minimum description length，and Helmholtz free energy[C]// Advances in Neural Information. Processing Systems. Burlington，USA，Morgan Kaufmann，1994：3-10

[27] Vincent P，Bengio Y，Larochelle H，et al. Stacked denoising autoencoders：learning useful representations in a deep network with a local denoising criterion[J]. Machine Learning Res，2010，11(12)：3371-3408

[28] Rifai S，Vincent P，Muller X，et al. Contractive Auto-Encoders：Explicit Invariance during Feature Extraction[C]// Proceedings of the 28th International Conference on Machine Learning. New York：ACM，2011：833-840

[29] Olshausen B，Field D. Emergence of simplecell receptive field properties by learning a sparse code for natural images[J]. Nature，1996，381(6583)：607-609

[30] Olshausen B，Field D. Sparse coding with an overcomplete basis set：a strategy employed by V1? [J]. Vision Research，1997，37(23)：3311-3326

[31] Lee H，Battle A，Ng A Y，et al. Efficient sparse coding algorithms[C]// Proceedings of the Twentieth Annual Conference on Neural Information Processing Systems，Van-couver. British Columbia，Canada，2006，19：801-808

[32] Lee H，Ekanadham C，Ng A. Sparse deep belief net model for visual area V2[C]// Proceedings of Advanvances In Neural Information Processing Systems. Cambridge，MA：MIT Press，2008：873-880

[33] Han Jun-wei，Zhang Ding-wen，Cheng Gong，et al. Object Detection in Optical Remote Sensing Images Based on Weakly Supervised Learning and High-Level Feature Learning Geoscience and Remote Sensing[J]. IEEE Journals & Magazines，2015，53(6)：3325-3337

[34] Rumelhart D, Hinton G, Williams R. Learning represention by back-propagating errors[J]. Nature, 1986, 323(6088): 533-536

[35] Deng L, Yu D. Deep Learning: Methods and Applications[R]. NOW Publishers, 2014

[36] Hinton G E, Dayan P, Frey B J, et al. The Wake-Sleep Algorithm for Self-Organizing Neural Networks[J]. Science, 1995, 268: 1158-1161

[37] Salakhutdinov R, Hinton G. Deep Boltzmann Machines [C]// Proceedings of the 12th International Conference on Artificial Intelligence and Statistics(AISTATS). 2009: 448-455

[38] Srivastava N, Salakhutdinov R. Multimodal Learning with Deep Boltzmann Machines [J]. Journal of Machine Learning Research, 2014, 15(8): 1967-2006

[39] Chen Yu-shi, Lin Zhou-han, Zhao Xing, et al. Deep Learning Based Classification of Hyperspectral Data[J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2014, 7(6): 2094-2107

[40] Zhang Yong-feng, Shang Chang-jing. Combining Newton interpolation and deep learning for image classification[J]. IET Journals & Magazines, 2015, 51(1): 40-42

[41] 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. NewYork: ACM, 2008: 1096-1103

[42] Markoff J. How many computers to identify a cat? [N]. The New York Times, 2012

[43] Le Q V, Ramzato M, Ng A, et al. Building high-level features using large scale unsupervised learning, 1112. 6209[R]. NewYork, USA: Cornell University, 2012

[44] Coates A, et al. Deep Learning with COTS HPC Systems[J]. JMLR W C P, 2013, 28(3): 1337-1345

[45] Rosenblatt F. The perceptron-a perceiving and recognizing automaton[C]// Math. Stat.. 1957

[46] Hinton G. Learning distributed representations of concept[C]// Proceedings of the Eighth Annual Conference of Thecognitive Science Society. 1986

[47] Hubel D H, Wiesel T N. Receptive Fields, Binocular Interaction. and Functional Architecture in the Cat's VisualCortex[J]. The Journal of Physiology, 1962, 160(1): 106-154

[48] LeCun Y, Kavukcuoglu K, Farabet C. Convolutional Networks and Applications in Vision[C]// Proceedings of 2010 IEEE International Symposium on Circuits and Systems(ISCAS). IEEE, 2010: 253-256

[49] LeCun Y, Bengio Y, Hinton G. Deep Learing[J]. Nature Magazines, 2015, 521(7553): 436-444

[50] Hinton G, Li Deng, Dong Yu, et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four researchgroups[J]. IEEE Signal. Process. Mag, 2012, 29(6): 82-97

[51] Ranzato M A, Hinton G E. Modeling pixel means and covariances using factorized third-order boltzmann machines [C]// IEEE Conference on Computer Vision and Pattern Recognition. SanFrancisco. CA: IEEE, 2010: 2551-2558

[52] Larochelle H, Murray I. The neural autoregressive distributeon estimator[C]// Proceedings of the 14th International Conference on Artificial Intelligence and Statistics. Fort Lauderdale, FL, United states: Microtome Publishing, 2011: 29-37

[53] Chang Hang, Zhou Yin, Borowsky A, et al. Stacked Predictive Sparse Decomposition for Classification of Histology Sections [J]. International Journal of Computer Vision (IJCV), 2015, 113(1): 3-18

[54] Krizhevsky A, Sutskever I, Hinton G E. ImageNet Classification with Deep Convolutional Neural Networks Neural Information Processing Systems 25. Neural Information [J]. Advances in Neural Information Processing Systems, 2012, 25(2): 2012

[55] Go Master Walloped by Emotionless Challenger, a Google Computer Program [N]. The New York Times. http://www.nytimes.com/2016/03/10/world/asia/google-alphago-lee-se-dol.html?partner=rss&emc=rss&_r=1

[56] Robot Vision Group. National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences [EB/OL]. http://vision.ia.ac.cn/index.html

[57] Chen T, Du Z, Sun N, et al. DianNao: a small-footprint high-throughput accelerator for ubiquitous machine-learning [J]. ACM Sigplan Notices, 2014, 49(4): 269-284

[58] Liu Shao-li, Chen Yun-ji, Chen Tian-shi, et al. DianNaoYu: An Instruction Set Architecture for Neural Networks[C]// Proceedings of the 43rd ACM/IEEE International Symposium on Computer Architecture(ISCA'16). 2016: 393-405

[59] Hu Zhen, Fu Kun, Zhang Chang-shui. Audio Classical Composer Identification by Deep Neural NetWork[J]. Journal of Computer Research and Development, 2014, 51(9): 1945-1954(in Chinese)
胡振, 傅昆, 张长水. 基于深度学习的作曲家分类问题[J]. 计算机研究与发展, 2014, 51(9): 1945-1954

[60] Li Jian-cheng, Yan Chun, Song You. Multilabel Image Annotation Based on Convolutional Neural Network [J]. Computer Science, 2016, 43(7): 41-45(in Chinese)
黎健成, 袁春, 宋友. 基于卷积神经网络的多标签图像自动标注[J]. 计算机科学, 2016, 43(7): 41-45

[61] Peng Bo, Zang Di. Vehicle Logo Recognition Based on Deep Learning[J]. Computer Science, 2015, 42(4): 268-273(in Chinese)
彭博, 臧笛. 基于深度学习的车标识别方法研究[J]. 计算机科学, 2015, 42(4): 268-273

[62] Shi Jing-lan, Chang Kan, Zhang Zhi-yong, et al. Coefficient-similarity-based Dictionary Learning Algorithm for Face Recognition [J]. Computer Science, 2016, 43(6): 298-302(in Chinese)
施静兰, 常侃, 张智勇, 等. 人脸识别中基于系数相似性的字典学习算法[J]. 计算机科学, 2016, 43(6): 298-302

[63] He K, Zhang X, Ren S, et al. Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNetClassification [J]. arXiv: 1502.01852

[64] He K, Zhang X, Ren S, et al. Deep Residual Learning for Image Recognition[J]. arXiv: 1512.03385

[65] Socher R, Huval B, Ng A Y, et al. Convolutional-Recursive Deep Learning for 3D Object Classification[C]// Neural Information Processing Systems Conference(NIPS). 2012: 665-673

[66] Sun Ning, Han Guang, Du Kun, et al. Person/vehicle classification based on deep belief networks[C]// 2014 10th International Conference on Natural Computation(ICNC). 2014: 113-117

[67] Bo Yu, Lane I. Multitask deep learning for image understanding multi-task training[C]// 2014 6th International Conference of Soft Computing and Pattern Recognition(SoCPaR). 2015: 37-42

[68] Sun Y, Wang X, Tang X. Deep learning face representation from predicting 10000 classes[C]// 2014 IEEE Conference on CVPR.

2014:1891-189

[69] Sun Y, Chen Y, Wang X, et al. Deep learning face representation by joint identification-verification[C]// Advances in N-eural Information Processing Systems. 2014:1988-1996

[70] Sun Y, Wang X, Tang X. Deeply learned face representations are sparse, selective, and robust[C]// 2015 IEEE Conference on CVPR. 2015:2892-2900

[71] Sawada Y, Kozuka K. Transfer learning method using multip-rediction deep Boltzmann machines for a small scaledataset[C]// 2015 14th IAPR International Conference MachineVision Applications (MVA). 2015:110-113

[72] Tang Jie-xiong, Deng Chen-wei, Huang Guang-bin, et al. Compressed-Domain Ship Detection on Spaceborne Optical Image Using Deep Neural Network and Extreme Learning Machine [J]. IEEE Transactions on Geoscience and Remote Sensing, 2015, 53(3):1174-1185

[73] Chen Xue-yun, Xiang Shi-ming, Liu Cheng-lin, et al. Aircraft Detection by Deep Belief Nets[C]// 2013 2nd IAPRA-sian Conference on ACPR. 2013:54-58

[74] Chen Xue-yun, Xiang Shi-ming, Liu Cheng-lin, et al. Vehicle Detection in Satellite Images by Parallel Deep Convolutional Neural Networks[C]// 2013 2nd IAPR Asian Conference on Pattern Recognition (ACPR). 2013:181-185

[75] Wu G, Kim M, Wang Q, et al. Scalable High Performance Image Registration Framework by Unsupervised Deep Feature RepresentationsLearning[J]. IEEE Transactions on Biomedical Engineering, 2016, 63(7):1

[76] Xu Yan, Mo Tao, Feng Qi-wei, et al. Deep learning of feature representation with multiple instance learning for medical image analysis[C]// IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2014:1626-1630

[77] Ouyang Wan-li, Wang Xiao-gang, Zeng Xing-yu. Deep-ID-Net: Deformable deep convolutional neural networks for object detection[C]// IEEE Conference on CVPR, 2015:2403-2412

[78] Girshick R, Donahue J, Darrell T, et al. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation[C]// IEEE Conference on CVPR, 2014:580-587

[79] Girshick R. Fast-R-CNN[J]. arXiv:1504. 08083v2

[80] Larsson G, Maire M, Shakhnarvich G. FractalNet: Ultra-Deep Neural Networks without Residuals[J]. arXiv:1605. 07648vl

[81] Gary Stix[OL]. http://www. scientificamerican. com/article/deep-learning-is-the-a-i-breakthrough-we-ve-been-waiting-for

(上接第 12 页)

[44] Cheng Kai, Zhang Hong-jun, Zhang Rui, et al. Calculation of operation task effectiveness based on extended timed influence nets [J]. Systems Engineering and Electronics, 2012, 34(12):2492-2497(in Chinese)
程恺, 张宏军, 张睿, 等. 基于扩展时间影响网络的作战任务效能计算方法[J]. 系统工程与电子技术, 2012, 34(12):2492-2497

[45] Cheng Kai, Zhang Hong-jun, Zhang Rui. A Task-resource Allocation Method Based on Effectiveness [J]. Knowledge-Based Systems, 2013, 37(1):196-202

[46] Haider S, Levis A H. Effective Course-of-Action Determination to Achieve Desired Effects[J]. IEEE Transactions On Systems Man and Cybernetics Part A: Systems and Humans, 2007, 37(6):1140-1150

[47] Levis A H, Haider S. Finding effective courses of action using particle swarm optimization[C]// IEEE World Congress on Computational Intelligence. 2008:1135-1140

[48] Helsinki A J. The systems concepts in military operations-discussion of critique[C]// IEEE International Systems Conference (SysCon). 2013

[49] Humbert M. Adopt the Effects-Based Approach to Operations? [J]. Defense Nationale et Securite Collective, 2008, 64:102-112

[50] Du Zheng-jun, Chen Chao, Jiang Xin. Modeling and solution of course of action based on influence net and sequential game[J]. Systems Engineering-Theory & Practice, 2013, 33(1):215-222 (in Chinese)
杜正军, 陈超, 姜鑫. 基于影响网络与序贯博弈的作战行动序列模型与求解[J]. 系统工程理论与实践, 2013, 33(1):215-222

[51] Du Zheng-jun, Chen Chao, Jiang Xin. Modeling and solution method of course of action based on influence net and multi-stage games with incomplete information [J]. Journal of National University of Defense and Technology, 2012, 34(3):63-67(in Chinese)
杜正军, 陈超, 姜鑫. 基于影响网络与不完全信息多阶段博弈的作战行动序列模型及求解方法[J]. 国防科技大学学报, 2012, 34(3):63-67

[52] Levchuk G M, Levchuk Y N, Pattipati K R, et al. Normative Design of Project-Based Organizations—Part III: Modeling Congruent, Robust, and Adaptive Organizations[J]. IEEE Transactions On Systems Man and Cybernetics Part A: Systems and Humans, 2004, 34(3):337-350

[53] Mou Liang. Dynamic Adaptive Optimization Methodology of C2 Organization Structure under Uncertainty Mission Environment [D]. Changsha: National University of Defense Technology, 2011(in Chinese)
牟亮. 不确定使命环境下 C2 组织结构动态适应性优化方法研究[D]. 长沙: 国防科学技术大学, 2011

[54] Corban B, Philemon S, Andrew A, et al. Robust Mission Planning[J]. Military Operations Research, 2011, 16(4):5-24

[55] Wu Yun-peng, Huang Jin-cai, Zhang Wei-ming, et al. Planning modeling and state reasoning in competitive situation[J]. Computer Engineering and Applications, 2011, 47(26):35-41(in Chinese)
武云鹏, 黄金才, 张维明, 等. 对抗条件下的计划生成过程建模及状态推理[J]. 计算机工程与应用, 2011, 47(26):35-41

[56] Li Wei-sheng. Plan recognition method based on intelligent planning[J]. Journal of Chongqing University of Posts and Telecommunications (Natural Science Edition), 2009, 21(4):538-543(in Chinese)
李伟生. 基于智能规划的计划识别方法研究[J]. 重庆邮电大学学报(自然科学版), 2009, 21(4):538-543

[57] Hong J. Plan Recognition through Goal Graph Analysis [J]. Journal of Artificial Intelligence Research, 2001, 15(1):1-30

[58] Zhou Yun, Huang Jiao-min, Huang Ke-li. A Study on Impact of "Deep Green" on Command and Control[J]. Fire Control & Command Control, 2013, 38(6):1-5(in Chinese)
周云, 黄教民, 黄柯棣. 美国"深绿"计划对指挥控制的影响[J]. 火力与指挥控制, 2013, 38(6):1-5