深度学习:多层神经网络的复兴与变革

山世光,阚美娜,刘昕,刘梦怡,邬书哲

中国科学院计算技术研究所,北京 100190

摘要 人工智能(AI)已经进入一个新的蓬勃发展期。推动这一轮AI狂澜的是三大引擎,即深度学习(DL)、大数据和大规模并行计算,其中又以DL为核心。本文回顾本轮"深度神经网络复兴"的基本情况,概要介绍常用的4种深度模型,即:深度信念网络(DBN)、深度自编码网络(DAN)、深度卷积神经网络(DCNN)及长短期记忆递归神经网络(LSTM-RNN)。简要介绍深度学习在语音识别和计算机视觉领域几个重要任务上的应用效果情况。为便于应用DL,介绍了几种常用的深度学习开源平台。对深度学习带来的启示和变革做了一些开放式的评述,讨论了该领域的开放问题和发展趋势。

关键词 深度神经网络:深度信念网络;深度自编码网络;深度卷积神经网络;长短期记忆递归神经网络;语音识别;计算机视觉

1 深度学习:神经网络的复兴

深度学习是以不少于2个隐含层的神经网络对输入进行非线性变换或表示学习的技术。其以层级连接的方式,实现渐进抽象的非线性信息处理,尤其擅长于求解从原始输入信号到期望输出的复杂非线性变换,并以此实现对原始数据的表示学习或非线性建模。深度学习尤其强调直接从原始数据开始进行"端到端(end-to-end)"的学习,而不像过去一样要从人工设计的特征开始进行学习,因此深度学习在很多场合下也被称为表示学习。

深度学习本质上是包含多个隐含层的人工神经网络,而人工神经网络的研究可以追溯到20世纪40年代。1943年,McCulloch和Pitts提出了第一个神经元的数学模型[1]。1949年,Hebb提出了神经元的学习准则[2]。1957年,Rosenblatt提出了感知机(perceptron)模型,开启了人工神经网络研究的第一次热潮[3]。但随后人工智能领域的知名学者Minsky和Papert等指出感知机模型是线性模型,无法解决线性不可分问题(如异或问题),导致人工神经网络的研究进入了第一次低潮期。此后,1986年Rumelhart、Hinton和Williams在《Nature》发表了著名的误差反向传播(back propagation,BP)算法,用于训练多隐含层神经网络[4],从而使得求解具有非线性学习能力的多层感知机(multi-layer perceptron,MLP)成为可能,带动了人工神经网络的第二次研究热潮。事实上,BP算法[4]作为训练多层神经网络的标准算法,直到今天仍被广泛应用。1989年Hornik等从理论上证明了多层感知机可以逼近任意复杂的连续函数[5],进一步激励了非线性感知机的发展。

虽然BP算法貌似"解决"了多层神经网络的训练问题,但对于训练较深的网络依然存在优化难题,加之当时数据规模都较小,不足以支撑多层神经网络大量参数(权重)的学习,因此多层神经网络并未在许多领域取得突破性结果,研究热潮仅持续了很短的几年,便再次进入了研究的低谷期。此后至2006年的10多年间,多数研究者转向了采用SVM、AdaBoost等在当时看来更加简单高效的分类方法,处理非线性的手段大多依赖于经验驱动的分段、局部线性逼近(如流形学习),或者采用Kernel技巧进行隐式的非线性映射。同时,研究者还基于经验或专家知识设计了众多"人造特征"。例如,在语音识别等语音信号处理领域,LPCC、MFCC等倒谱系数特征被大量采用,而在计算机视觉领域,描述图像局部梯度或纹理性质的SIFT、HOG、Gabor、LBP等也被广泛应用,并在很多问题上取得了还不错的性能。但人工设计特征通用性差,设计新的特征需要很强的先验知识。最重要的是,即使拥有大量数据,这些以经验或专家知识驱动的模型和方法也无法借力大数据,不能从大数据中学习出其中蕴含的丰富知识和规律。

今天所谓的深度学习作为多层神经网络的复兴始于2006年。这一年,多伦多大学的Hinton等在《Science》及《Neural Computation》上发表文章[6,7],强调多隐层深度神经网络相比浅层网络具有更优异的特征学习能力,并可以通过分层、无监督的预训练有效解决深度神经网络训练困难的问题。与此同时,蒙特利尔大学的Bengio等在国际会议NIPS2006上发表论文[8],也强调了分层(layer-wise)训练深度网络的做法。这些工作重新开启了多层神经网络研究

收稿日期:2016-05-30;修回日期:2016-06-30

作者简介:山世光,研究员,研究方向为图像处理、计算机视觉、模式识别、人机交互,电子信箱:sgshan@ict.ac.cn

引用格式:山世光, 阚美娜, 刘昕, 等. 深度学习:多层神经网络的复兴与变革[J]. 科技导报, 2016, 34(14): 60-70; doi: 10.3981/j.issn.1000-7857.2016.14.007

热潮。从2009年到2011年,谷歌和微软研究院均采用深度网络,配合大规模的训练数据,将语音识别系统的错误率降低了20%以上。此后,深度神经网络在诸多领域尤其是语音处理和计算机视觉领域取得了巨大的成功,显著提升了语音识别、搜索、推荐、图像分类、物体检测、视频分析、人脸识别、行人检测、语义分割、机器翻译等众多智能处理任务的性能。

以计算机视觉为例,点燃深度学习在该领域应用热潮的爆点发生在2012年。这一年,Hinton研究组设计了深度卷积神经网络(DCNN)模型AlexNet,利用ImageNet提供的大规模训练数据,并采用两块GPU卡进行训练,将ImageNet大规模视觉识别竞赛(ILSVRC)之"图像分类"任务的Top5错误率降低到15.3%,而传统方法的错误率高达26.2%(且仅比2011年降低了2个百分点)[9]。这一结果让研究者看到了深度学习的强大威力,以致2013年这个竞赛再次举行时,成绩靠前的队伍几乎全部采用了深度学习方法,其中图像分类任务的冠军来自纽约大学Fergus研究组,他们将Top 5错误率降到了11.7%,所采用的模型亦是进一步优化的深度CNN。2014年,在同一竞赛中,Google则依靠一个加深为22层的深度卷积网络GoogLeNet将Top 5错误率降低到了6.6%[10]。到2015年,微软亚洲研究院的何凯明等则设计了一个深达152层的ResNet模型将这一错误率刷新到了3.6%[11]。4年内,ImageNet图像分类任务的Top5错误率从26.2%到11.7%再到6.6%最后到3.6%,几乎每年错误率都下降50%,这显然是一次跨越式(如果不是革命性)的进步。

然而,回溯历史不难发现深度学习是复兴而非革命。前面提到,20世纪90年代之后神经网络研究陷入了低潮期,但实际上神经网络的研究并未完全中断。特别的,LeCun等在1989年就提出了卷积神经网络(CNN)[12],并在此基础上于1998年设计了LeNet-5卷积神经网络[13],通过大量数据的训练,该模型成功应用于美国邮政手写数字识别系统中。前文所述引爆深度学习在计算机视觉领域应用热潮的AlexNet即是LeNet-5网络的扩展和改进。还需要说明的是,LeNet-5等CNN结构设计在一定程度上受到了Fukushima于1975年提出的Cognitron模型[14]和1980年提出的Neocognitron模型[15]的启发。它们与CNN在网络学习方法上有较大差异:Neocognitron采用的是无导师、自组织的学习,而CNN则依赖于有导师信号的大量数据进行参数学习。但二者都试图模拟诺贝尔奖获得者Hubel和Wiesel于1962年提出的视觉神经系统的层级感受野模型[16],即从简单特征提取神经元(简单细胞)到渐进复杂的特征提取神经元(复杂细胞,超复杂细胞等)的层级连接结构。前述的AlexNet、GoogleNet和ResNet等深度模型在基本结构上都是CNN,只是在网络层数、连接方式、非线性激活函数、优化方法等方面有了新的发展。从这个意义上讲,深度学习的种子在20世纪80年代已经生根发芽,此次爆发可以认为是丰沛的雨水(大数据)和日日肥沃的土壤(并行计算设备如GPU等)共同作用的结果。

2 常见深度模型

过去10年,深度学习在不同维度快速发展,形成原理、结构和适用范围具有较大差异的多种深度模型,其中既包括2006年Hinton、Bengio等关注较多的深度信念网络(DBN)和栈式自编码器网络(deep auto-encoder network,DAN),也包括目前在语音和视觉信息处理领域炙手可热的深度卷积神经网络(DCNN),还包括在时序信号处理上具有更大优势的长短时记忆递归神经网络(LSTM-RNN)。以下介绍DBN、DAN、DCNN、LSTM-RNN等常见深度模型及其发展。

2.1 深度信念网络(deep belief network)

简单地说,深度信念网络(deep belief network,DBN)就是层叠多个受限波尔兹曼机(restricted boltzmann machine,RBM)组成的深度神经网络。作为一种产生式模型,它可以学习到训练数据的概率分布,使得网络可以最大概率产生训练数据。DBN的基本组件是RBM,一种可通过输入数据集学习概率分布的产生式随机神经网络。受限玻尔兹曼机通过堆叠(stack)多个RBM得到深度玻尔兹曼机(DBM)。RBM和DBM网络中节点之间为无向连接,如果靠近数据层的部分层之间的连接为有向连接,即为深度信念网络[17]。

RBM最初由Paul Smolensky于1986年提出并命名为Harmonium。顾名思义,受限玻尔兹曼机是玻尔兹曼机的变体,与玻尔兹曼机中通常的全连接图相比,RBM限定模型必须为二分图,即仅在可见单元和隐单元之间有无向连接,而可见单元之间以及隐单元之间都没有连接,如图1所示。该"限制"使得RBM相比一般BM有更高效的训练算法成为可能,2006年Hinton及其合作者发明了快速学习算法后[7],才使受限玻尔兹曼机逐渐广为人知。

DBN的训练过程通常是贪婪式的逐层训练,即首先无监督地训练第一个RBM,然后将其隐单元层作为后一个RBM的可见单元层,训练下一个RBM,以此类推层叠多个RBM。这种逐层的无监督训练可以解释为预训练或权重初始化过程,使得DBN的高效学习成为可能,而且可以避免网络收敛到不良的局部极值。逐层训练结束后,还可以对整个深层网络作进一步的微调(fine-tune)。微调的目标函数可以是无监督的,如图2(a)所示,也可以是有

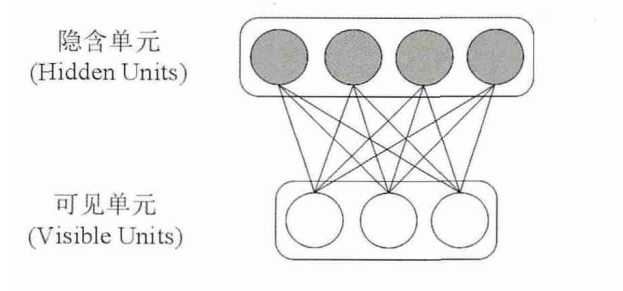
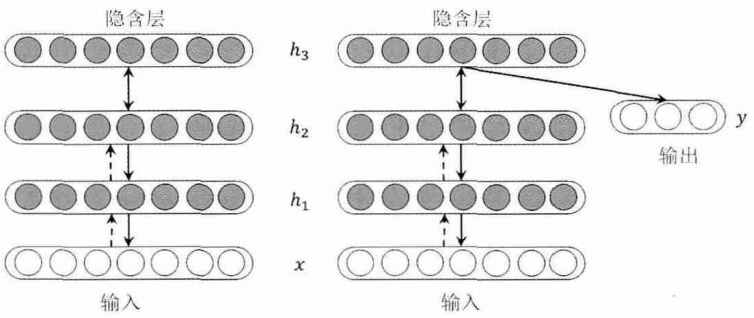


图1 受限玻尔兹曼机网络结构



(a) 无监督深度信念网络　　(b) 有监督深度信念网络

图2　深度信念网络结构

监督的,如图2(b)所示。微调时,最上层的梯度不仅影响其下一层,还将继续传播到下面所有层,从而对网络进行整体优化。无监督的深度信念网络可以在不同粒度上对输入数据进行抽象,因而一个自然的应用即是对数据进行压缩或者降维。另外一个常见的应用是用作特征提取,可以对数据进行有效的刻画。

2.2　深度自编码网络(deep auto-encoder network)

自编码器(auto-encoder,AE)是一种无监督的、以重构输入为目标的人工神经网络,主要用于学习压缩的或过完备的特征表示。结构上,自编码器是一个前向的无环网络,包含一个输入层、一个隐层、一个输出层。当自编码器包含多个隐层时即形成了深度自编码网络。深度自编码器的隐含层结点数通常明显少于输入节点数,形成一个压缩式网络结构,因此最后一个隐层的激活响应可以被看作是对输入样本的压缩表示。如果隐层节点数比输入层节点数多,则深度自编码器学习到的可能是恒等函数(identity function),不具有任何意义,因而通常可对隐层加入额外的约束如稀疏性,从而使得深度自编码网络学习具有特定属性或者过完备的特征表示。

深度自动编码器的训练常采用反向传播的方法,例如共轭梯度下降、最速下降等。但这些方法对于层数较多的深度自编码器网络存在优化难题,例如,梯度在反向传播时会逐渐变小,导致整个网络无法继续优化。Hinton等提出的逐层预训练方法可以有效地解决这个问题[7],即将相邻的两层当作单隐层的自动编码器进行优化,所有层的权重经过逐层预训练之后可得到较为合理的初始值,然后再对整个网络进行微调即可得到一个有效的深度自编码器网络。与卷积神经网络类似,深度自编码器网络学习过程中也容易遇到过拟合问题,一种有效的解决方式是Vincent等提出的去噪自动编码器(denoising autoencoder)[18],即在训练样本上人为施加噪声作为网络输入,但输出依然为无噪声的样本,由此学到的网络会对噪声具有很好的鲁棒性,提升推广能力。

除了用于无监督的特征学习,自编码特性使得深度自编码网络亦可用于动态纹理预测、信号恢复、去噪等目的。与DBN类似,为使深度自编码网络具备分类判别能力,可以将其输出层替换为类别,形成一个Softmax层,并通过类似误差后向传播(BP)的方法对网络权值进行精细调整,从而得到一个分类或回归网络。

2.3　深度卷积神经网络(deep convolutional neural network)

尽管前述DBN和DAN及其相应的逐层、无监督预训练策略敲响了深度学习时代的大钟,并在语音识别等问题上取得重要进步,但深度学习真正杀手级的成功应用更多来自于深度卷积网络(DCNN),DCNN在计算机视觉领域的巨大成功全面引爆了深度学习的狂潮。自2012年,基于LeCun等在1998年设计的深度卷积神经网络LeNet-5[13],研究者设计了许多新的深度学习模型。以下首先简单介绍LeNet-5,并介绍几个重要的DCNN模型,即:AlexNet[9],GoogLeNet[10],VGGNet[19]以及最近的ResNet[20]等。

2.3.1　CNN及其LeNet-5

与全连接网络相比,LeCun于1989年提出的卷积神经网络最重要的特征就是引入了卷积层,它将非0连接权重限制在局部时空范围内,也可以认为全连接中的非局部连接权重被强制为0了,以此提取局部特征并有效保留空间位置信息,因此卷积本质上是局部滤波操作。相比全连接,卷积显然大大降低了需要学习的参数量。而更为重要的是,CNN还引入了权值共享的策略,即一个响应图内的所有神经元共享相同的卷积核,从而进一步大大降低了需要学习的参数量。为提取更为丰富的局部特征,CNN通常在不同层上设置不同大小和数量的卷积核。

在LeCun于1989年首次提出CNN模型的9年后,1998年LeCun在贝尔实验室设计了LeNet-5网络,并成功应用于美国邮政手写数字识别系统[13]。如图3所示,LeNet-5网络有2个卷积层和2个全连接层,而每个卷积层实际包括卷积、非线性激活函数映射和下采样(pooling)3个步骤,其中非线性激活函数(通常采用Sigmoid,tanh等)则赋予了网络非线性映射能力,而下采样则兼具特征降维和获得对平移等局部变化不变性的双重作用。从图3[13]可以看出,LeNet-5在2个卷积层上使用了不同数量的卷积核:第一层是6个,第二层则是16个。

2.3.2　AlexNet

AlexNet是由Hinton研究组于2012年设计的深度卷积神经网络[9],其参加当年的ImageNet竞赛获得图像分类任务最好成绩,大幅降低了图像分类的错误率,从而引爆了深度学习在计算机视觉中的应用热潮。如图4所示,AlexNet包含5个卷积层和3个全连接层(最后一层是输出层)。与LeNet-5相比,AlexNet在设计上有几个重要特征:1)隐含层数更多;2)引入了新的非线性激活函数ReLU替代了之前普遍采用的饱和型非线性激活函数(如Sigmoid,tanh等),实践表明ReLU这样的非饱和型激活函数有利于更快速的收敛,大大减少训练时间;3)采用了Dropout防止

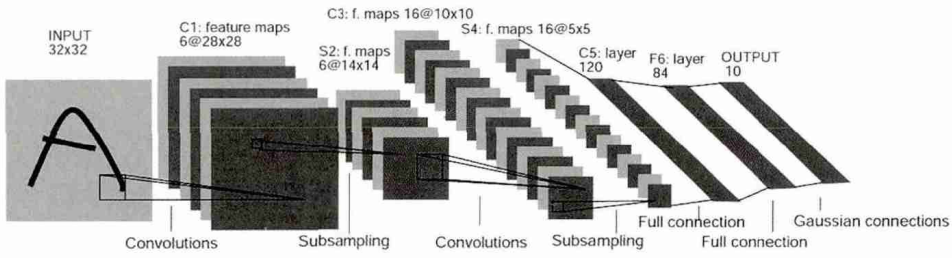


图3　LeNet-5卷积网络结构

过学习的训练策略;4）如图4所示,为解决存储和训练速度问题,AlexNet采用了双GPU卡两路训练结构,分别处理输入图像的上下部分。

2.3.3　GoogLeNet

GoogLeNet是2014年Google研发团队设计的一个22层的深度卷积神经网络[10],取得了当年ImageNet竞赛物体分类任务的最好成绩,其将Top-5的错误率从2013年的11.7%降低到了6.6%。相比AlexNet,GoogLeNet的核心不同体现在:1）更深的网络结构;2）一个被称之为Inception的多尺度卷积结构。Inception这个名字取自电影《盗梦空间》,意指嵌套式的卷积结构,如图5所示,其作用是通过多尺度的卷积操作提取响应图中的多尺度信息,增强卷积模块的功能,提高网络的非线性建模能力。

如图6所示,整个GoogLeNet叠加了多个上述Inception结构,以致总的网络深度达到22层。为了缓解深层网络训练过程中的梯度消失问题,GoogLeNet在中间层引入了2个辅助损失层,即图5中间的2个Softmax损失层引导的子结构。

2.3.4　VGG

VGG是牛津大学的Simonyan和Zisserman等于2014年设计的[19],取得了2014年ImageNet ILSVRC竞赛物体检测任务的第一名,物体分类任务的第二名。如图7所示,VGG的网络深度为16~19层,其卷积核大小为3×3,卷积采样间隔1×1,Max Pooling间隔为2×2,随着层数的增高,卷积核的数目则从64个逐渐增加到了512个。需要注意的是,VGG和GoogLeNet中都采用了卷积层紧随卷积层的做法,而不是一定要卷积和pooling交替出现。VGG与GoogLeNet均表明更深的网络对分类、识别等视觉任务更有利。

2.3.5　ResNet

ResNet的全称是Deep Residual Network(深度残差网络),是微软亚洲研究院的何凯明等于2015年底的ImageNet竞赛中提出的一种DCNN网络结构[11]。虽然加大网路的深度被证明是提高实际问题性能的一种有效手段,但直接增加层数会出现所谓的"退化"现象:以CIFAR-10数据集为例,网络深度从20层增加到56层,性能反而明显下降。为此,ResNet中引入了一种shortcut结构,如图8所示。每个building block的输出是输入信号x和卷积层输出$F(x)$的叠加。ResNet消除了超深网络训练过程中的"退化"现象,可以训练深达152层的网络,在ImageNet 2015年的竞赛中取得了不依赖外部数据条件下的物体识别与定位,以及物体检测2个子竞赛的冠军。最新的ResNet改进工作,即恒等映射深度残差网络,去掉了building block中ReLU操作,使得不同层building block

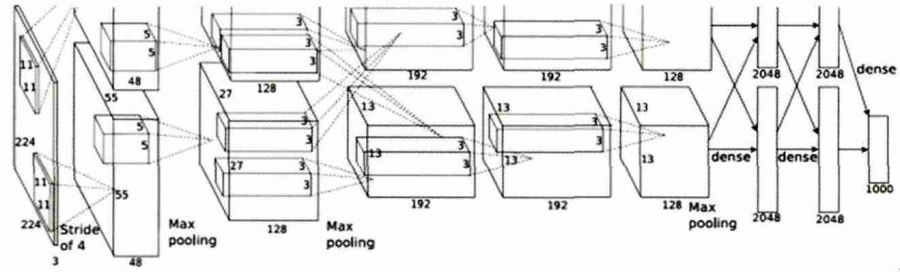


图4　AlexNet网络结构示意

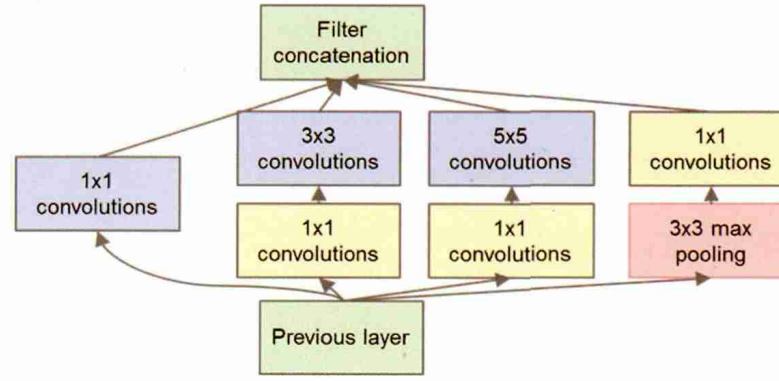
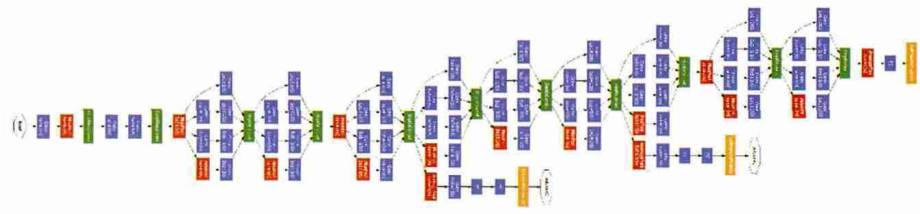


图5　GoogLeNet中的Inception结构示意



蓝色模块表示卷积层;红色模块表示MaxPooling层;绿色模块为局部响应值归一化或DepthConcat(响应图串联);橙色模块为Softmax层

图6　GoogLeNet网络结构示意

图7 VGG网络结构

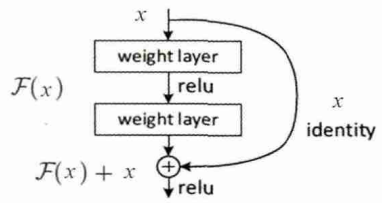


图8 ResNet中具有shortcut结构的building block

中的残差项可加,网络的深度可以达到1001层,在CIFAR-10数据集上取得了当前最好的4.62%的分类错误率。

2.3.6 全卷积网络(FCN)

对于像素级的分类和回归任务,例如图像分割或边缘检测,目前代表性的深度网络模型是全卷积网络(fully convolutional network,FCN)[20]。经典的DCNN在卷积层之后使用了全连接层,而全连接层中单个神经元的感受野是整张输入图像,破坏了神经元之间的空间关系,因此不适用于像素级的视觉处理任务。为此,如图9所示,FCN去掉了全连接层,代之以1×1的卷积核和反卷积层,从而能够在保持神经元空间关系的前提下,通过反卷积操作获得与输入图像大小相同的输出。进一步,FCN通过不同层、多尺度卷积特征图的融合为像素级的分类和回归任务提供了一个高效的框架。

2.4 递归神经网络RNN与长短期记忆网络LSTM

人类对世界的理解很大程度上基于脑海中已有的信息与认知,以阅读文章或观看电影为例,上下文对内容的理解非常关键。受此启发,研究者们对传统神经网络进行了结构改进,添加了循环递归模块用于信息的保持与传递,这就是递归神经网络(recurrent neural network, RNN)。如图10左半部分所示,神经网络单元A接收输入x_t、输出h_t、下标t指示时间。这里,模块A的循环结构实现了网络中信息从t到$(t+1)$时刻的保持与传递。图10右半部分为左图循环结构按照时间顺序的展开形式,这种链式属性体现了RNN与(时间)序列有密切关联,是分析处理此类数据最自然的网络结构。

基于RNN的展开式网络结构,可以利用前向传播算法(forward propagation)依次按照时间顺序计算输入输出,然后利用反向传播算法(BP)将累积残差从最后一个时间步骤传递回前面的步骤。这样的方法在处理"长期依赖关系"时,后面的时间节点对前面时间节点的感知能力会下降,出现"梯度消失(vanishing gradient)"的问题。

长短期记忆网络(LSTM)是一种特殊的RNN,由Hochreiter和Schmidhuber(1997)提出[21],之后很多研究者进行了诸多改进。LSTM引入了新机制对"记忆模块"进行改进,使其能够有效学习长期依赖关系。在传统的RNN中,循环模块A一般由一个简单的激活层构成,这里以tanh()函数为例;LSTM同样具有RNN的链式结构,区别在于循环模块的构成,它包含4个神经网络层以特定的方式相互作用,如图11所示。

LSTM的核心思想在于单元模块的状态,即图中的C_t。从C_{t-1}到C_t的水平线表示每个时间步骤下的单元状态沿整个链式结构传递下去,中间过程只包含部分线性操作,因此容易以较小的损失保持信息。此外,LSTM可以通过"门限(gates)"结构向单元状态中添加或从中移除信息。该操作可由一个sigmoid神经网络层和点乘运算实现,输出0~1之间的数值,用于描述信息通过的程度,0表示全部拒绝,1表示全部接受。具体地,LSTM利用3种门限:遗忘门限(forget gate)、输入门限(input gate)和输出门限(output gate)来实现保护和控制单元状态。对于某时刻下的单元模块,第一步需要确定哪些上文信息将从单元状态中移除,该决策由"遗忘门限"的sigmoid层决定(对应图11中最左侧的蓝色方框);第二步需要确定在单元状态中存储哪些新信息:首先,由"输入门限"的sigmoid层(对应图11中左二的蓝色方框)对上文信息进行更新,之后由tanh()激活层(对应图11中左三的蓝色方框)创建新的候选状态:这需要结合上文信息与新信息共同确定新的单元状态,其中"遗忘门限"和"输入门限"分别决定了旧信息和新信息的权重(即接受程度);第三步需要确定单元的输出,该输出取决于更新后的单元状态。这需要由"输出门限"(对应图11中左四的蓝色方框)结合输入信息确定一个接受权重。在此基础上,更新后的单元状态经过激活层(对应上图中最右侧的蓝色方框),乘以该权重

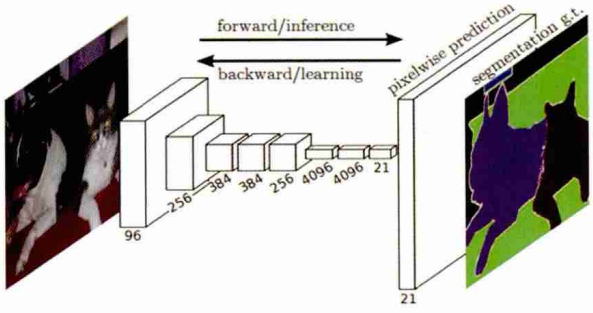


图9 FCN结构示意

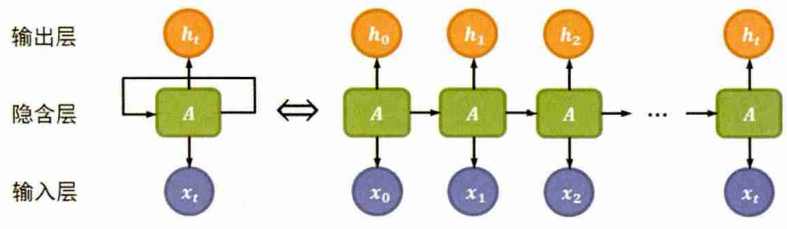


图10 递归神经网络RNN及其展开式

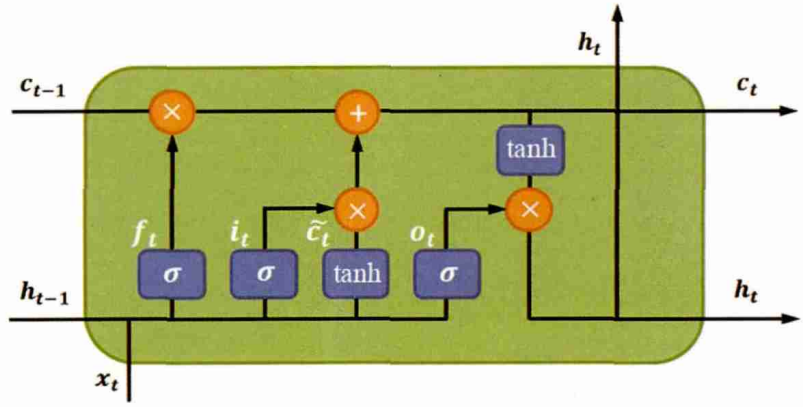


图11 LSTM的循环模块构成

即可得到输出。经此3个步骤,这利用3种门限有效地控制了增加和移除信息的频率与程度,从而实现长时或短时的信息记忆。

在过去几年中,许多序列问题,如语音识别、语言建模、机器翻译、图像分割、图像描述生成等,使用RNN/LSTM取得了显著成效。在此基础上,也有一些研究者根据任务的实际特性设计了LSTM的变种:Gers和Schmidhuber提出了"窥视孔连接"[22],将单元状态也同时作为门限的输入,并引入"耦合遗忘"机制,避免独立的对信息移除进行判断。最近较为流行的被称为"门限递归单元(gated recurrent unit,GRU)"的变种由Cho等提出[23],它将遗忘和输入门限联合输入到单个"更新门限"中,同样还将单元状态和隐藏状态合并,得到更为简化高效的LSTM模型。

3 深度学习典型任务的成功应用

近年来,深度学习在大量智能计算任务中取得了广泛而深刻的应用,在很多问题上带来了超出预期的效果,这不仅包括语音处理和视觉计算,还包括自然语言理解、搜索、推荐,乃至围棋这样极其复杂的任务。特别是计算机视觉方面,深度学习在图像分类、图像分割、目标检测、事件检测、动作识别、场景识别、图像标注、图像检索、图像超分辨率、图像去模糊、显著性检测等视觉处理任务上均取得了显著的进步。下面介绍语音识别和计算机视觉中的几个典型任务的深度学习应用情况。

3.1 语音识别

在深度学习流行之前,语音识别系统的精度已经长期没有本质突破。2009年Lee等首次用无监督卷积神经网络方法将DBN用于声学信号处理,在讲话者、性格和音素检测上表现出比梅尔倒谱系数(MFCC)更优越的性能[24]。2009年开始,微软与多伦多大学的Hinton合作,研究基于深度学习的新一代语音识别技术。2011年,微软的邓力和俞栋等将深度神经网络成功应用于语音识别并降低了20%~30%的错误率,彻底改变了语音识别原有的技术框架[25]。百度和Google也采用了深度学习进行语音识别,均上线了基于深度学习技术的商用语音识别与搜索系统。

3.2 图像分类与物体识别

图像分类与物体识别是一类经典计算机视觉问题,也是深度学习战果最为辉煌的战场。1989年卷积神经网络首次提出时,就被用于美国邮政手写数字识别。2010年,斯坦福大学李飞飞等牵头组织了ILSVRC(ImageNet large scale vision recognition)大规模视觉识别竞赛,提供了大量标注丰富的图像数据,并已成为事实上的图像分类与物体识别的测试基准。2012年,深度学习技术在当年的ILSVRC2012竞赛之图像分类任务中取得了冠军,并随后成为图像分类与物体识别领域的代表性方法。图12显示了2010—2015年6年间ILSVRC竞赛图像分类任务的冠军方法及Top 5的错误率,深度学习方法已经将2011年冠军(基于传统浅层模型方法)25.8%的Top 5错误率下降到2015年3.57%(基于ResNet)。

3.3 目标检测

深度学习给目标检测方法带来了深刻的变化,从最初基于刚性模板的级联分类器和基于可变形模板的部件模型,到现在基于卷积神经网络的端到端的检测器,滑动窗口被候选区域生成方法所取代,手工设计的特征让位给了自动学习的特征,浅层的模型被替换为深层的模型。方法上的变革带来了检测精度的跨越式提升,目前最具代表性的基于深度学习的目标检测器Faster R-CNN[26],在各大竞赛上均拔得头筹:1)在Pascal VOC 2012的检测任务comp4上排名第一,在全部20个类别获得最高的mAP;2)在ILSVRC 2015的检测任务上获得冠军,在200个类别中的194个类别上获得最高的mAP;3)在MS COCO 2015的检测任务上高居榜首。可以说,以Faster R-CNN为代表的基于深度学习的方法在目标检测任务上取得了空前的成功。

Faster R-CNN是R-CNN[27]系列方法中最先进的一个。R-CNN将新的检测框架展现在人们面前,通过Selective Search生成候选窗口,用深层卷积神经网络提取特征,用SVM进行分类,并引入边框回归,在检测精度上大幅超过旧有方法;Fast R-CNN[28]将分类和边框回归以多任务学习的方式集成进统一的网络,并借鉴SPP的做法设计了RoIPooling层,在提升精度的同时显著加快了检测速度;Faster R-CNN进一步抛弃了单独的候选窗口生成模块,将其作为网络的一个分支,构建起了端到端的检测器,进一步提升了精度和速度。Faster R-CNN现在仍在继续发展,通过设计在线的难例挖掘策略,其性能能够得到进一步的提升,相信后续在网络结构、优化方法等方面的进步会带来这一框架的进一步改进。

在深度学习的影响下,目标检测器的设计正朝着化零为整的方向迈进:从不同窗口不同过程,转变为以全图作为输入端到端的检测方式;从不同样例不同模型,转变为用所有样例训练单个模

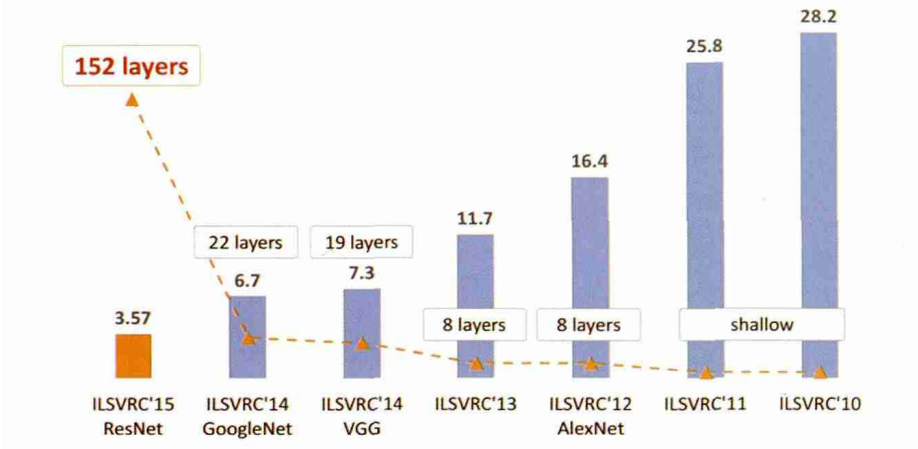


图12　2010—2015年ILSVRC竞赛图像分类任务冠军方法及Top 5错误率

型处理所有的情况;从不同功能不同模块,转变为将所有功能集成到同一个模型;从不同任务不同系统,转变为多任务同时执行相辅相成。

3.4　语义分割

基于深度学习的语义分割方法通过端到端学习,将图像的像素表示和像素的分类统一到同一个框架中,因而取得了更好的分割效果。其中代表性的方法有前面介绍过的全卷积网络FCN[20]。DeepLab则进一步将CRF与FCN结合,利用双线性差值上采样得到粗的得分图,再基于全连接CRF精化分割结果[29]。Zheng[30]提出的CRF-RNN则将CRF中的mean-field迭代实现为栈式CNN结构,从而可以等价于RNN网络[31]。该方法在PASCAL VOC图像分割测试集上,基于深度学习的方法取得了显著优越的性能,进一步将分割精度(平均交并比)提高到了75%左右(FCN方法仅62%左右,而传统方法低于50%)。

3.5　人脸识别

人脸识别是一种细粒度的物体识别。由于受到光线、姿态、表情等因素的影响从而使得人脸的变化分布是非线性的而且极为复杂,是极具挑战的一个任务。Labeled faces in the wild(LFW)是目前最具挑战的非可控人脸识别数据集[31]。在深度学习浪潮"爆发"之前,2013年传统的基于手工设计特征的识别方法在LFW上取得的最好性能为95.17%(平均分类精度)[32]。之后,深度学习技术迅速提升了非可控条件下人脸识别的性能。2014年,2个采用深度学习的团队即来自Facebook的团队[33]和香港中文大学的团队[34]分别报告了97.35%和97.45%的平均分类精度。这些结果比前述高维LBP特征方法的分类错误率降低了50%。上述2个团队均采用了卷积神经网络(CNN)的变种架构,其中Facebook的DeepFace方法强调了前端的人脸3D对齐和虚拟正面化预处理以削弱姿态变化的影响。而香港中文大学的DeepID方法则强调采用多个人脸区块分别训练卷积神经网络,并最终融合形成人脸特征表示。需要特别注意的是:这2个系统能够取得优异性能的另一个重要原因是均采用了大规模的人脸标注数据进行训练。例如,DeepFace采用了来自4030人的440万幅人脸图像(均来自社交网络);而DeepID则使用了来自10177人的共约20万人脸图像(均为网络名人图像)。也就是说,这2个系统均采用了大规模的人脸标注数据进行学习,而且其训练图像的分布与LFW测试图像(名人图像)有一定的相似性。最近香港中文大学团队进一步改进开发了DeepID2+系统,在上述测试环境下单个深度模型取得了98.7%的正确分类精度[35]。进而,谷歌公司通过挖掘大规模图像(800万人的2亿幅照片)并结合Triplet损失函数单模型取得了99.63%的正确分类精度[36]。

需要注意的是,在LFW上取得了99.63%的正确分类精度并不意味着人脸识别问题的解决。人脸识别有10余种不同的应用场景,其成熟度迥异。成熟度较高的应用包括超大规模证件照的查重比对、基于主动近红外的人脸考勤和门禁、名人人脸识别与检索、基于身份证原图的身份核验等。成熟度尚与应用预期有较大差距的场景主要是基于监控摄像头的黑名单目标人监控等。

3.6　图像自动标题

图像自动标题(image captioning)的目标是生成输入图像的文字描述[37,38],即人们常说的"看图说话",是深度学习最近才取得了重要进展的一个研究方向。深度学习方法应用于该问题的代表性思路是采用CNN学习图像表示,采用RNN/LSTM学习语言模型,并采用CNN的特征输入初始化RNN/LSRM的隐层节点,组成混合网络进行端到端的训练。代表性的工作有Google的Vinyals提出的Show and Tell方法[39],斯坦福大学Karpathy等提出的Neural Talk方法[40]和德克萨斯州大学奥斯汀分校的Donahue提出的long-term recurrent convolutional networks(LRCN)方法[41]等。这些方法均取得了令人印象深刻的结果,例如Google的系统在MSCOCO数据集上的很多结果甚至已经优于人给出的描述。

4　深度学习开源平台

深度学习得以快速发展还受益于学术界和产业界对于相关研究的开放和开源。特别是若干开源平台的出现,为深度学习的快速普及提供了良好的基础。其中,代表性的开源平台有:伯克利大学发起的DCNN开源项目Caffe,Facebook发起的Torch,DMLC发起的MXNet,Bengio等发起的包含CNN、DBN等深度学习算法的python深度学习项目Theano,以及Google发起的深度学习开源项目Tensorflow等。

4.1　Caffe

Caffe最早是由伯克利大学的贾扬清发起的DCNN开源平台,基于C++开发,提供Matlab和Python的调用。后续采用了Github全球协作的方式进行开发。Caffe是目前使用最为广泛的DL

开源平台。Caffe平台的优点包括:支持灵活的硬件架构,只需要简单改变GPU或CPU标志位即可支持不同平台;可扩展性好,其底层采用C++编写,架构灵活,可以实现间接的第三方模块嵌入;速度快等。强大的开发者社区支持,Caffe为学术研究、初创公司甚至大规模的计算机视觉、多媒体、语音识别等应用提供了强大支持,支持GitHub全球协作开发。

4.2 Torch

Torch是由Facebook发起的深度学习平台,底层基于C++编写,上层提供Lua代码调用。同时Torch包含了一系列的开发插件,包括iTorch、fbcunn、fbnn、fbcuda和fblualib等。这些插件能够在很大程度上提升神经网络的性能,可用于计算机视觉和自然语言处理(NLP)等场景。其中的fbcunn包含了Facebook用于GPU的高度工程化深度学习模块,该模块可用来加快深度学习的速度,fbcunn对物体识别、自然语言的处理以及其他大规模的深度学习系统(如卷积神经网络)有很大的帮助,目前Torch已被Google、Twitter、Intel、AMD、NVIDIA等公司采用,同时也作为一个易用的深度学习原型化平台被学术界采用。

4.3 MXNet

MXNet是百度牵头发起的深盟项目的一部分,强调提高内存使用的效率,甚至能在智能手机上运行诸如图像识别等任务。MXNet的优点包括:1)采用symbolic接口,使得用户可以快速构建一个神经网络;2)支持更多binding,目前支持比较好的是python,也支持Julia和R;3)支持多卡和多机运行;4)性能上更优。

4.4 Theano

Theano开源项目于2007年诞生于加拿大蒙特利尔大学,由Bengio领衔发起。Theano是一个强大的Python类库,支持用户基于多维矩阵高效地定义、优化和求解数学表达式。Theano的特性包括:1)与Python数值计算库Numpy高度整合,用户只需要一行代码即可调用Numpy全部功能;2)对GPU的透明使用,在float32精度下,相比CPU可以取得140倍的加速比;3)高效的符号积分,以及高效的一元或多元函数求导;4)速度和稳定性优化;5)自动生成C语言函数,可以更快地求解表达式;6)扩展的单元测试与自动验证功能:自动检测和定位多种类型的错误。

4.5 Tensorflow

TensorFlow是Google基于DistBelief研发的第二代人工智能学习系统,其命名来源于本身的运行原理:Tensor(张量)意味着N维数组,Flow(流)意味着基于数据流图的计算。TensorFlow可被用于语音识别或照片识别等领域,它可在小到一部智能手机、大到数千台数据中心服务器的各种设备上运行。TensorFlow完全开源,任何人都可以用,而且具备更好的灵活性和可延展性。TensorFlow的另一大亮点是支持异构设备分布式计算,从单个CPU/GPU到成百上千GPU卡组成的分布式系统,均可方便地运行。

5 深度学习成功的启示

深度学习的成功不仅仅带来了人工智能相关技术的快速进步,解决了许多过去被认为难以解决的难题,更重要的是它为人们带来了思想观念的变革。

5.1 优化方法的变革是开启深度学习复兴之门的钥匙

回顾自2006年(所谓深度学习元年)以来深度学习的10年大发展,必须首先注意到优化方法不断进步的重要作用。前面多次提到,业界现在热捧的深度模型(如CNN)在20世纪80年代就已经基本成型,当时未能普及的原因很多,其中之一是长期缺少有效的优化多层网络的高效方法,特别是对多层神经网络进行初始化的有效方法。从这个意义上讲,Hinton等2006年的主要贡献是开创了无监督的、分层预训练多层神经网络的先河,从而使众多研究者重拾了对多层神经网络的信心。但实际上最近3年来DCNN的繁荣与无监督、分层预训练并无多大关系,而更多的与优化方法或者有利于优化的模块有关,如Mini-Batch SGD、ReLU激活函数、Batch Normalization等,特别是其中处理梯度消失问题的手段,对DCNN网络不断加深、性能不断提升功不可没。

5.2 从经验驱动的人造特征范式到数据驱动的表示学习范式

在深度学习兴起之前,专家知识和经验驱动的AI范式主宰了语音处理、计算机视觉和模式识别等众多领域很多年,特别是在信息表示和特征方面,过去大量依赖人工的设计,严重影响了智能处理技术的有效性和通用性。深度学习彻底颠覆了这种"人造特征"的范式,开启了数据驱动的表示学习范式。具体体现在:1)所谓的经验和知识也在数据中,在数据量足够大时无需显式的经验或知识的嵌入,直接从数据中可以学到;2)可以直接从原始信号开始学习表示,而无需人为转换到别的空间再进行学习。数据驱动的表示学习范式使得研发人员无需根据经验和知识针对不同问题设计不同的处理流程,从而大大提高了AI算法的通用性,也有效降低了解决新问题的难度。

5.3 从"分步分治"到"端到端的学习"

分治或分步法,即将复杂问题分解为若干简单子问题或子步骤,曾经是解决复杂问题的常用思路,在AI领域,也是经常被采用的方法论。比如,为了解决图像模式识别问题,过去经常将其分解为预处理、特征提取与选择、分类器设计等若干步骤。再如,为了解决非线性问题,可以采用分段线性方式来逼近全局的非线性。这样做的动机是很清晰的,即:子问题或子步骤变得简单、可控,更易解决。但从深度学习的视角来看,其劣势也同样明显:子问题最优未必意味着全局的最优,每个子步骤是最优的也不意味着全过程是最优的。相反,深度学习更强调端到端的学习(end-to-end learning),即:不去人为的分步骤或者划分子问题,而是完全交给神经网络直接学习从原始输入到期望输出的映射。相比分治策略,端到端的学习具有协同增效(synergy)的优势,有更大的可能获得全局上更优的解。当然,如果一定要把分层看成是"子步骤或子问题"也是可以的,但它们各自完

成什么功能并不是预先设定好的,而是通过基于数据的全局优化来自动学习的。

5.4 深度学习具备超强的非线性建模能力

众多复杂问题本质上是高度非线性的,而深度学习实现了从输入到输出的非线性变换,这是深度学习在众多复杂问题上取得突破的重要原因之一。在深度学习之前,众多线性模型或近似线性模型曾多年大行其道。特别是从20世纪90年代开始,以判别式降维为目的的线性子空间方法得到大家的重视,如主成分分析、Fisher线性判别分析、独立成分分析等。之后,为了处理非线性问题,Kernel技巧、流形学习等非线性处理方法相继得到重视。其中Kernel方法试图实现对原始输入的非线性变换,但却无法定义显式的非线性变换,只能借助有限种类的kernel函数,定义变换空间中的点积,间接实现非线性。2000年之后曾一度广受重视的流形学习方法则试图通过对样本点之间测地距离或局部邻域关系的保持来学习非线性映射,遗憾的是这类方法难以真正实现对非训练样本的显式非线性变换。深度学习则通过作用于大量神经元的Sigmoid或ReLU等非线性激活函数,获得了可以适配足够复杂的非线性的能力。

5.5 大模型未必总是不好的

奥卡姆剃刀原理在诸多领域特别是机器学习领域广为人知,它告诫人们:"如无必要,勿增实体",换句话说,求解问题的模型能简单最好不要复杂。这一原理在机器学习领域是提高模型推广能力的重要原则,也使得复杂的大模型往往不被看好。深度学习恰恰在这一点上是令人费解的,以AlexNet为例,其需要学习的参数(权重)多达6000万个,如此之巨的参数量表明这是一个非常复杂(如果不是过分复杂的话)的模型。当然,模型中需要学习的参数的多少并不直接反应模型的复杂度,但毋庸置疑的是,深度学习乍看起来是"复杂度"非常高的。那么,奥卡姆剃刀原理失效了吗?抑或看似复杂的深度学习模型的复杂度并不高?目前似乎尚无明确的理论支撑。最近的一些工作表明,很多已经训练好的、复杂的深度学习模型可以通过剪枝等手段进行约简,其性能并不降低甚至可以提高。

这里的关键也许在于"大数据"带来的"红利"。一种可能是:科研人员过去长期面对着"小数据"问题,因而过于偏爱简单模型。而在数据量陡增的今天,适度复杂的模型变得更加适应科研人员面对的复杂问题,当训练数据量大到与测试数据同分布,甚至测试数据基本"跑不出"训练数据时,在训练数据上的"过拟合"就变得不那么可怕了。

5.6 脑神经科学启发的思路值得更多的重视

深度学习作为多层神经网络是受脑神经科学的启发而发展起来的。特别是卷积神经网络,其根源于Fukushima在1980年底提出的认知机模型,而该模型的提出动机就是模拟感受野逐渐变大、逐层提取由简及繁的特征、语义逐级抽象的视觉神经通路。在Hubel和Wiesel的共同努力下,该通路从20世纪60年代开始逐渐清晰,为CNN的诞生提供了良好的参照。但值得注意的是,生物视觉神经通路极其复杂,神经科学家对初级视觉皮层区中简单神经细胞的边缘提取功能是清晰的,对此后的部分复杂神经细胞的功能也有一些探索,但对更高层级上的超复杂细胞的功能及其作用机制尚不清晰。这意味着CNN等深度模型是否真的能够模拟生物视觉通路还不得而知。但可以确定的是,生物神经系统的连接极为复杂,不仅仅有自下而上的前馈和同层递归,更有大量的自上而下的反馈以及来自其他神经子系统的外部连接,这些都是目前的深度模型尚未建模的。

但无论如何,脑神经科学的进步可以为深度模型发展提供更多可能性,是非常值得关注的。例如,最近越来越多的神经科学研究表明,曾一度被认为功能极为特异化的神经细胞其实具有良好的可塑性,例如,视觉皮层的大量神经细胞在失去视觉处理需求后不久,即被"重塑"转而处理触觉或其他模态的数据。神经系统这种可塑性意味着不同智能处理任务具有良好通用性,为通用人工智能的发展提供了参照。

6 开放性问题及发展趋势

大数据支撑的深度学习固然是AI领域的里程碑式进步,但并不意味着深度学习具有解决全部AI问题的潜力。未来,在深度学习领域还会遇到开放性问题,有必要对其可能的未来发展思路和趋势进行讨论。

6.1 大数据是否学习之必需

大数据是深度学习成功的重要基础。越来越多的应用领域正持续积累着日趋丰富的应用数据,这对深度学习的进一步发展和应用至关重要。然而,过分倚重有标注的大数据也恰恰是深度学习的局限性之一。数据收集是有成本的,而且标注成本已经开始水涨船高,而且还有一些领域存在着难以收集数据的问题,例如在医疗诊断领域,一些较为罕见疾病的相关数据收集是困难的。

更重要的,当将人的智能作为参照系时,自然就会问:人的智能是否是大数据学习的结果呢?其答案并不显然。从人类个体的角度来说,答案很可能是否定的:我们甚至可以只见过一个苹果(甚至只是一张苹果图片)就学会了识别苹果,而无需观察成百上千个不同的苹果。但是,这样批判深度学习看似有理有据,却未必是公平的:人类作为一个种群,进化过程中已经见过了何止成百上千个苹果?

无论如何,"小数据"如何驱动深度学习是一个值得探索的新方向。在此意义上,基于无监督数据的学习、相似领域的迁移学习、通用模型的领域适应、知识与经验的嵌入等方法值得关注。

6.2 无师自通:如何获取无监督学习能力

获取有标注数据的时间和金钱成本很高,但大量无监督数据的获取成本却是微乎其微的。目前深度学习对无监督数据的学习能力严重不足,以致大

量无监督数据就像富含黄金的沙海,人们却没有高效淘金的利器。有趣的是,回顾深度学习的历史,2006年Hinton等倡导的却恰恰是利用无监督学习来对深层神经网络进行预训练。但此后,特别是DCNN兴起后,无监督的预训练也已经被很多研究者所抛弃。

直接从大量无监督数据中学习模型确实非常困难,即便是人这部"机器",也有"狼孩"的例子警告我们"无师自通"似乎是不现实的。但"少量有导师数据+大量无导师数据"的模式也许是更值得大力研究的。

6.3 从参数学习到结构学习

深度学习以数据驱动范式颠覆了"人造特征"范式,这是一个重大的进步。但与此同时,它自己又陷入了一个"人造结构"窠臼中。无论Hinton教授组最初设计的AlexNet,还是后来的VGG、GoogLeNet、ResNet等,都是富有经验的专家人工设计出来的。给定一个新问题,到底什么样的网络结构是最佳的(如多少卷积层)却不得而知,这严重阻碍着深度学习在更多视觉任务上的普及和应用。因此,同时学习网络结构和网络参数是一个值得大力关注的研究方向。从计算的角度看,全面的学习网络结构是极其复杂的。尽管近期已经有一些这方面的尝试,如剪枝算法、网络约简等,可以在一定程度上调整网络结构,但尚处于起步阶段。

6.4 如何赋予机器演绎推理能力

基于大数据的深度学习可以认为是归纳法,而从一般原理出发进行演绎是人类的另一重要能力,特别是在认知和决策过程中,我们大量依赖演绎推理。演绎推理在很多时候似乎与数据无关。例如,即使不给任何样例,我们也可以依赖符号(语言)描述,学会识别之前从未见过的某种物体。这样的学习问题看似超出了深度学习的触角范畴,但也许未必不可企及。例如,近年来越来越多的基于深度学习的产生式模型正在努力实现从符号(概念)到图像的生成。

参考文献(References)

[1] McCulloch W S, Pitts W. A logical calculus of the ideas immanent in nervous activity[J]. Bulletin of Mathematical Biophysics, 1943, 5(4): 115-133,

[2] Hebb D O. The organization of behavior[M]. New York: Wiley, 1949.

[3] Rosenblatt F. The perceptron: A probabilistic model for information storage and organization in the brain[J]. Psychological Review, 1958, 65(6): 386-408,.

[4] Rumelhart D E, Hinton G E, Williams R J. Learning internal representations by error propagation[J]. Nature, 323, 1986, doi: 10.1016/B978-1-4832-1446-7.50035-2.

[5] Hornik K, Stinchcombe M, White H. Multilayer feedforward networks are universal approximators[J]. Neural Networks, 1989, 2(2): 359-366,.

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

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

[8] Bengio Y, Lamblin P, Popovici D, et al. Greedy Layer-Wise training of deep networks[M]//Advances in Neural Information Processing Systems 19 (NIPS' 06), Cambridge MA: MIT Press, 2007: 153-160.

[9] Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks[M]//Advances in Neural Information Processing Systems (NIPS). Cambridge MA: MIT Press, 2012: 1097-1105.

[10] Szegedy C, Liu W, Jia Y Q. et al. Vincent vanhoucke and andrew rabinovich. Going deeper with convolutions[C]. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, June 7-12, 2015: 1-9.

[11] He K M, Zhang X Y, Ren S Q, et al. Deep residual learning for image recognition[C]. IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, June 26-July 1, 2016.

[12] LeCun Y, Boser B, Denker J S, et al. Backpropagation applied to handwritten zip code recognition[J]. Neural Computation, 1989, 1(4): 541-551.

[13] LeCun Y, Bottou L, Bengio Y, et al. Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE, 1998, 86: 2278-2324,.

[14] Cognitron F K. A self-organizing multilayered neural network[J]. Biological Cybernetics 1975, 20: 121-136.

[15] Neocognitron F K. A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position[J]. Biological Cybernetics 1980, 36: 193-202.

[16] Hubel D H, Wiesel T N. Receptive fields, binocular interaction and functional architecture in cat's visual cortex[J]. Journal of Physiology(london), 1962, 160: 106-154.

[17] Salakhutdinov R, Hinton G E. Deep boltzmann machines[C]. International Conference on Artificial Intelligence and Statistics (AISTATS), Clearwater, April 16-18, 2009: 448-455.

[18] Vincent P, Larochelle H, Lajoie I, et al. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion[J]. Journal of Machine Learning Research (JMLR), 2010, 11: 3371-3408,.

[19] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[C]. International Conference on Learning Representations (ICLR), San Diego, CA, May 7-9, 2015.

[20] Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation[C]. The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, June 7-12, 2015: 3431-3440.

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

[22] Gers F A. Schmidhuber J. Recurrent nets that time and count[C]. Proceedings of the IEEE-INNS-ENNS International Joint Conference on Neural Networks. Italy, July 24-27, 2000.
[23] Cho K, van Merrienboer B, Bahdanau D, et al. On the properties of neural machine translation: Encoder-decoder approaches[R]. Semantics and Structure in Statistical Translation, Doha, Qatar, October 25, 2014.
[24] Lee H, Pham P T, Largman Yan, et al. Unsupervised feature learning for audio classification using convolutional deep belief networks[C]. Advances in Neural Information Processing Systems (NIPS), Vancouver, British Columbia, Canada. December 7-10, 2009.
[25] Dahl G E, Yu D, Deng L, et al. Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition[J]. IEEE Transactions on Audio, Speech, and Language Processing, 2012, 20(1): 30-42.
[26] Ren S Q, He K M, Girshick R, et al. Faster R-CNN: Towards real-time object detection with region proposal networks[C]. IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, June 26-July 1, 2016.
[27] Girshick R, Donahue J, Darrell T, et al. Region-based convolutional networks for accurate object detection and semantic segmentation[M]. Redmond: IEEE Computer Society, 2015: 142-158.
[28] Girshick R. Fast R-CNN[C]. IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, December 13-16, 2015.
[29] Chen L C. Semantic image segmentation with deep convolutional nets and fully connected CRFs[C]. ICLR 2015, San Diego, May 7-9, 2015.
[30] Zheng S. Conditional random fields as recurrent neural networks[C]. IEEE International Conference on Computer Vision (ICCV), , Santiago, Chile, December 13-16, 2015.
[31] Huang G B. Ramesh M, Berg T, et al. Labeled faces in the wild: A database for studying face recognition in unconstrained environments[R]. Technical Report 07-49, Amherst: University of Massachusetts , 2007.
[32] Chen D, Cao X, Wen F, et al. Blessing of dimensionality: High-dimensional feature and its efficient compression for face verification[C]. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Portland, June 23-28, 2013.
[33] Taigman Y, Yang M. Marc´aurelio ranzato and lior wolf. DeepFace: closing the gap to human-level performance in face verification[C]. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Columbus, June 24-29, 2014.
[34] Sun Y, Chen Y H, Wang X G, et al. Deep learning face representation by joint identification-verification[C]//Advances in Neural Information Processing Systems (NIPS), 2014: 1988-1996,
[35] Sun Y, Wang X G, Tang X O. Deeply learned face representations are sparse, selective, and robust[C]. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, June 7-12, 2015.
[36] Schroff F, Kalenichenko D, Philbin J. FaceNet: A unified embedding for face recognition and clustering[C]. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, June 7-12, 2015.
[37] Ali F. Every picture tells a story: Generating sentences from images[C]//Proceedings of the 11th European conference on Computer vision: Part IV. Berlin: Springer-Verlag, 2015: 15-29.
[38] Gaurav K. Babytalk: Understanding and generating simple image descriptions[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013, 35(12): 2891-2903.
[39] Vinyals O. Show and tell: A neural image caption generator[C]. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, June 7-12, 2015.
[40] Karpathy A. Deep visual-semantic alignments for generating image descriptions[C]. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, June 7-12, 2015.
[41] Donahue J. Long-term recurrent convolutional networks for visual recognition and description[C]. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, June 7-12, 2015.

Deep learning: The revival and transformation of multi layer neural networks

SHAN Shiguang, KAN Meina, LIU Xin, LIU Mengyi, WU Shuzhe

Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China

Abstract Artificial intelligence (AI) has entered a new period of vigorous development. This round of AI topsy is driven by three engines, namely the depth of learning (DL), big data and massively parallel computing, with DL as the core. This article reviews from a historical perspective the basic situation of the round "deep neural networks renaissance", then summarizes the four common depth models: deep belief network (DBN), depth from network coding (DAN), deep convolutional neural networks (DCNN) and long short term memory recurrent neural network LSTM-RNN. After that, this paper briefly introduces the application effects of deep learning in speech recognition and computer vision. In order to facilitate the application of DL, it also introduces several commonly used deep learning platforms. Finally, the enlightenment and reform of deep learning are commented, and the open problems and development trend in this field are discussed.

Keywords multilayer neural networks; DBN; DAN; DCNN; LSTM-RNN; speech recognition; computer vision

(责任编辑 王媛媛)