ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 31,No. 1,Jan. 2016,pp. 1－17
DOI:10.16337/j.1004-9037.2016.01.001

http://sjcj.nuaa.edu.cn
E-mail:sjcj@nuaa.edu.cn
Tel/Fax:+86-025-84892742

深度卷积神经网络在计算机视觉中的应用研究综述

卢宏涛　张秦川

(上海交通大学计算机科学与工程系,上海,200240)

摘　要:随着大数据时代的到来,含更多隐含层的深度卷积神经网络(Convolutional neural networks,CNNs)具有更复杂的网络结构,与传统机器学习方法相比具有更强大的特征学习和特征表达能力。使用深度学习算法训练的卷积神经网络模型自提出以来在计算机视觉领域的多个大规模识别任务上取得了令人瞩目的成绩。本文首先简要介绍深度学习和卷积神经网络的兴起与发展,概述卷积神经网络的基本模型结构、卷积特征提取和池化操作。然后综述了基于深度学习的卷积神经网络模型在图像分类、物体检测、姿态估计、图像分割和人脸识别等多个计算机视觉应用领域中的研究现状和发展趋势,主要从典型的网络结构的构建、训练方法和性能表现3个方面进行介绍。最后对目前研究中存在的一些问题进行简要的总结和讨论,并展望未来发展的新方向。

关键词:深度学习;卷积神经网络;图像识别;目标检测;计算机视觉

中图分类号:TP391　　**文献标志码**:A

Applications of Deep Convolutional Neural Network in Computer Vision

Lu Hongtao, Zhang Qinchuan

(Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai, 200240, China)

Abstract: Deep learning has recently achieved breakthrough progress in speech recognition and image recognition. With the advent of big data era, deep convolutional neural networks with more hidden layers and more complex architectures have more powerful ability of feature learning and feature representation. Convolutional neural network models trained by deep learning algorithm have attained remarkable performance in many large scale recognition tasks of computer vision since they are presented. In this paper, the arising and development of deep learning and convolutional neural network are briefly introduced, with emphasis on the basic structure of convolutional neural network as well as feature extraction using convolution and pooling operations. The current research status and trend of convolutional neural networks based on deep learning and their applications in computer vision are reviewed, such as image classification, object detection, pose estimation, image segmentation and face detection etc. Some related works are introduced from the following three aspects, i.e., construction of typical network structures, training methods and performance. Finally, some existing problems in the present research are briefly summarized and discussed and some possible new directions for future development are prospected.

Key words: deep learning; convolutional neural network; image recognition; object detection; computer vision

---

基金项目:国家自然科学基金(61272247)资助项目。
收稿日期:2015-12-10;**修订日期**:2016-01-10

引 言

图像识别是一种利用计算机对图像进行处理、分析和理解，以识别各种不同模式的目标和对象的技术，是计算机视觉领域的一个主要研究方向，在以图像为主体的智能化数据采集与处理中具有十分重要的作用和影响。使用图像识别技术能够有效地处理特定目标物体的检测和识别(如人脸、手写字符或是商品)、图像的分类标注以及主观图像质量评估等问题。目前图像识别技术在图像搜索、商品推荐、用户行为分析以及人脸识别等互联网应用产品中具有巨大的商业市场和良好的应用前景，同时在智能机器人、无人自动驾驶和无人机等高新科技产业以及生物学、医学和地质学等众多学科领域具有广阔的应用前景。早期的图像识别系统主要采用尺度不变特征变换(Scale-invariant feature transform, SIFT[1])和方向梯度直方图(Histogram of oriented gradients, HOG[2])等特征提取方法，然后将提取到的特征输入至分类器中进行分类识别。这些特征本质上是一种手工设计的特征，针对不同的识别问题，提取到的特征好坏对系统性能有着直接的影响，因此需要研究人员对所要解决的问题领域进行深入的研究，以设计出适应性更好的特征，从而提高系统的性能。这个时期的图像识别系统一般都是针对某个特定的识别任务，且数据的规模不大，泛化能力较差，难以在实际应用问题当中实现精准的识别效果。

深度学习是机器学习的一个分支，是近些年来机器学习领域取得的重大突破和研究热点之一。2006 年，加拿大多伦多大学教授、机器学习领域的泰斗 Geoffery Hinton 和他的学生 Ruslan Salakhutdinov 在国际顶尖学术刊物《Science》上发表了一篇文章[3]，第一次提出了深度学习的思想。这篇文章主要提出了两个观点：(1) 含多个隐层的人工神经网络具有十分强大的特征学习能力，通过训练模型所提取的特征对原始输入数据具有更抽象和更本质的表述，从而有利于解决特征可视化或分类问题；(2) 通过使用无监督学习算法实现一种称作“逐层初始化”的方法，实现对输入数据信息进行分级表达，从而可以有效地降低深度神经网络的训练难度。随后，深度学习在学术界和工业界持续升温，在语音识别、图像识别和自然语言处理等领域获得了突破性的进展。2011 年以来，研究人员首先在语音识别问题上应用深度学习技术，将准确率提高了 20% ～ 30%，取得了十多年来最大的突破性进展。仅仅一年之后，基于卷积神经网络的深度学习模型就在大规模图像分类任务上取得了非常大的性能提高，掀起了深度学习研究的热潮。文献[4]提出了两种基于深度神经网络的声学建模方法，相比于传统建模方法提取到了更有效的声学特征，并在维吾尔语的大词汇量连续语音识别应用上取得了较大的性能提升。目前，谷歌、微软和 Facebook 等众多国际互联网科技企业争相投入大量的资源，研发布局大规模的深度学习系统。

1 卷积神经网络

20 世纪 60 年代初期，Hubel 和 Wiesel 等通过对猫的大脑视觉皮层系统的研究，提出了感受野[5]的概念，并进一步发现了视觉皮层通路中对于信息的分层处理机制，由此获得了诺贝尔生理学或医学奖。到了 80 年代中期，Fukushima 等基于感受野概念提出的神经认知机[6]，可以看作是卷积神经网络(Convolution neural networks, CNNs)的第一次实现，也是第一个基于神经元之间的局部连接性和层次结构组织的人工神经网络。神经认知机是将一个视觉模式分解成许多子模式，通过逐层阶梯式相连的特征平面对这些子模式特征进行处理，使得即使在目标对象产生微小畸变的情况下，模型也具有很好的识别能力。在此之后，研究人员开始尝试使用一种被称作多层感知器[7]的人工神经网络(实际上是只含一层隐含层节点的浅层模型)来代替手工提取特征，并使用简单的随机梯度下降方法来训练该模型，于是进一步提出了用于计算误差梯度的反向传播算法，这一算法随后被证明十分有效[8]。1990 年，LeCun 等[9]在研究手写数字识别问题时，首先提出了使用梯度反向传播算法训练的卷积神经网络模型，并在 MNIST[10]手写数字数据集上表现出了相对于当时其他方法更好的性能。梯度反向传播算法和卷积神经网络的成功给机器学习领域带来了新的希望，开启了基于统计学习模型的机器学习浪潮，同时也带动

了人工神经网络进入到蓬勃发展的新阶段。目前,卷积神经网络已成为当前语音分析和图像识别领域的研究热点,它是第一个真正意义上的成功训练多层神经网络的学习算法模型,对于网络的输入是多维信号时具有更明显的优势。随着深度学习掀起的新的机器学习热潮,卷积神经网络已经应用于语音识别、图像识别和自然语音处理等不同的大规模机器学习问题中。

1.1 概念

卷积神经网络是一种为了处理二维输入数据而特殊设计的多层人工神经网络,网络中的每层都由多个二维平面组成,而每个平面由多个独立的神经元组成,相邻两层的神经元之间互相连接,而处于同一层的神经元之间没有连接。CNNs 受到早期的时延神经网络(Time-delay neural networks[11], TDNNs)的启发,TDNN 通过在时间维度上共享权值来降低网络训练过程中的计算复杂度,适用于处理语音信号和时间序列信号。CNNs 采用了权值共享网络结构使之更类似于生物神经网络,同时模型的容量可以通过改变网络的深度和广度来调整,对自然图像也具有很强的假设(统计的平稳性和像素的局部相关性)。因此,与每层具有相当大小的全连接网络相比,CNNs 能够有效降低网络模型的学习复杂度,具有更少的网络连接数和权值参数,从而更容易训练。

1.2 网络结构

一个简单的卷积神经网络模型的结构示意图如图 1 所示,该网络模型由两个卷积层(C_1, C_2)和两个子采样层(S_1, S_2)交替组成。首先,原始输入图像通过与 3 个可训练的滤波器(或称作卷积核)和可加偏置向量进行卷积运算,在 C_1 层产生 3 个特征映射图,然后对每个特征映射图的局部区域进行加权平均求和,增加偏置后通过一个非线性激活函数在 S_1 层得到 3 个新的特征映射图。随后这些特征映射图与 C_2 层的 3 个可训练的滤波器进行卷积,并进一步通过 S_2 层后输出 3 个特征映射图。最终 S_2 层的 3 个输出分别被向量化,然后输入到传统的神经网络中进行训练。

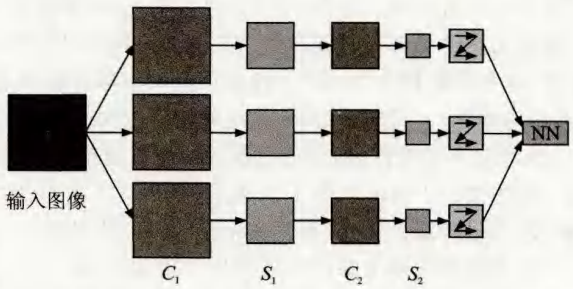

图 1 简化的卷积神经网络结构

Fig. 1 Simplified structure of convolutional neural network

1.3 卷积特征提取

自然图像有其固有特性,即对于图像的某一部分,其统计特性与其他部分相同。这意味着在这一部分学习到的特征也能用在另一部分上,因此对于图像上的所有位置,可以使用同样的学习特征。换句话说,对于大尺寸的图像识别问题,首先从图像中随机选取一小块局域作为训练样本,从该小块样本中学习到一些特征,然后将这些特征作为滤波器,与原始整个图像作卷积运算,从而得到原始图像中任一位置上的不同特征的激活值。给定分辨率为 $r\times c$ 的大尺寸图像,将其定义为 x_{large},首先从 x_{large} 中抽取 $a\times b$ 的小尺寸图像样本 x_{small},通过训练稀疏自编码器得到 k 个特征和激活值 $f(W^{(1)}x_{\text{small}}+b^{(1)})$,其中 $W^{(1)}$ 和 $b^{(1)}$ 是训练得到的参数。然后对于 x_{large} 中每个 $a\times b$ 大小的 x_s,计算对应的激活值 $f_s(W^{(1)}x_{\text{small}}+b^{(1)})$,进一步使用 x_{small} 的激活值与这些激活值 f_s 作卷积运算,就可以得到 $k\times(r-a+1)\times(c-b+1)$ 个卷积后的特征映射图。二维卷积计算的示意图如图 2 所示。

例如,对于分辨率为 128×128 的原始输入图像,假设经过预训练已经得到了该图像的 200 个 8×8 大小的特征碎片。那么,通过使用这 200 个特征碎片对原始图像中每个 8×8 的小块区域进行卷积运算,每个特征碎片均可以得到 121×121 的卷积特征映射图,最终整幅图像可以得到 200×121×121 的卷积特征映射图。

1.4 池化操作

通过将卷积层提取到的特征输入至分类器中进行训练，可以实现输出最终的分类结果。理论上可以直接将卷积层提取到的所有特征输入至分类器中，然而这将需要非常大的计算开销，特别是对于大尺寸高分辨率图像。例如：对于一个输入为 96×96 大小的图像样本，假设在卷积层使用 200 个 8×8 大小的卷积核对该输入图像进行卷积运算操作，每个卷积核都输出一个 $(96-8+1)\times(96-8+1)=7\,921$ 维的特征向量，最终卷积层将输出一个 $7\,921\times200=1\,584\,200$ 维的特征向量。将如此高维度的特征输入至分类器中进行训练需要耗费非常庞大的计算资源，同时也会产生严重的过拟合问题。然而，由于图像具有一种“静态性”的属性，在图像的一个局部区域得到的特征极有可能在另一个局部区域同样适用。因此，可以对图像的一个局部区域中不同位置的特征进行聚合统计操作，这种操作称为“池化”。比如计算该局部区域中某个卷积特征的最大值(或平均值)，称作最大池化(或平均池化)。具体来说，假设池化的区域大小为 $m\times n$，在获得卷积特征后，将卷积特征划分为多个 $m\times n$ 大小的不相交区域，然后在这些区域上进行池化操作，从而得到池化后的特征映射图。如图 3 所示，在一幅图像的 4 块不重合子区域上使用 3×3 大小的窗口对其进行最大池化，得到池化后的特征映射图。

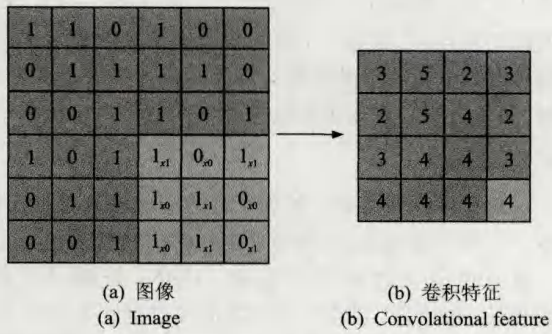

图 2 二维卷积运算操作示意图

Fig. 2 Illustration of two-dimensional convolution operation

如果选择图像中的连续范围作为池化区域，同时只对相同的隐含神经元产生的卷积特征使用池化，则这些池化后的特征单元具有平移不变性。也就是说，即使原始图像中的物体产生了一个较小的平移，依然可以得到相同的池化特征，分类器也依然能够输出相同的分类结果。与直接使用卷积后的特征相比，这些概要统计特征不仅能够极大地降低特征向量的维度，进一步降低训练分类器所需的计算量，而且能够有效地扩充训练数据，有利于防止过拟合。

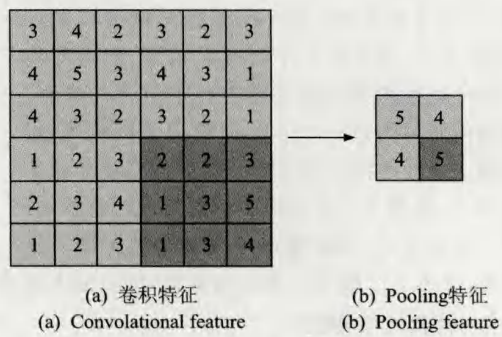

图 3 最大池化运算操作示意图

Fig. 3 Illustration of max pooling operation

2 图像分类

图像分类问题是通过对图像的分析，将图像划归为若干个类别中的某一种，主要强调对图像整体的语义进行判定。当下有很多用于评判图像分类算法的带标签的数据集，比如 CIFAR-10/100[12]，Caltech-101/256[13-14]和 ImageNet[15]，其中 ImageNet 包含超过 15 000 000 张带标签的高分辨率图像，这些图像被划分为超过 22 000 个类别。从 2010 年至今，每年举办的 ImageNet Large Scale Visual Recognition Challenge (ILSVRC)图像分类比赛是评估图像分类算法的一个重要赛事。它的数据集是 ImageNet 的子集，包含上百万张图像，这些图像被划分为 1 000 个类别。其中，2010 年与 2011 年的获胜团队采用的都是传统图像分类算法，主要使用 SIFT，LBP[16]等算法来手动提取特征，再将提取的特征用于训练支持向量机(Support vector machine，SVM)等分类器进行分类，取得的最好结果是 28.2%的错误率[17]。ILSVRC2012 则是大规模图像分类领域的一个重要转折点。在这场赛事中，Alex Krizhevsky 等提出的 AlexNet[18]首次将深度学习应用于大规模图像分类，并取得了 16.4%的错误率，该错误率比使

用传统算法的第 2 名的参赛队低了大约 10%。如图 4 所示,AlexNet 是一个 8 层的卷积神经网络,前 5 层是卷积层,后 3 层为全连接层,其中最后一层采用 softmax 进行分类。该模型采用 Rectified linear units (ReLU)来取代传统的 Sigmoid 和 tanh 函数作为神经元的非线性激活函数,并提出了 Dropout 方法来减轻过拟合问题。

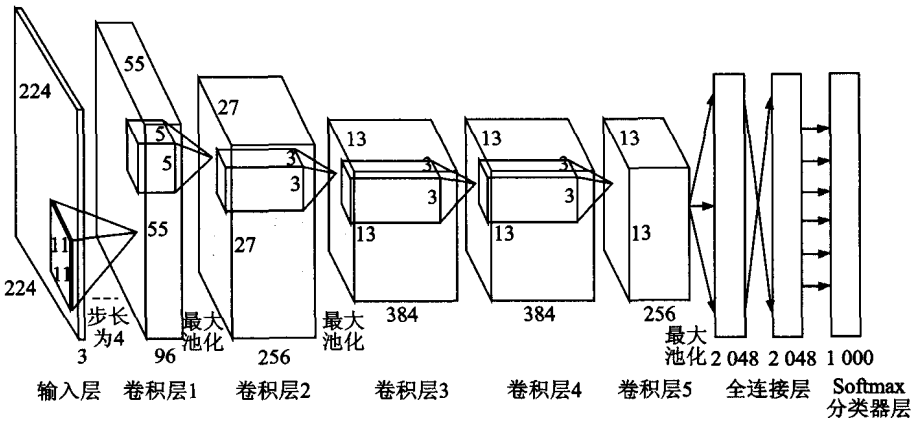

图 4 简化的 AlexNet 模型结构

Fig. 4 Simplified structure of AlexNet model

自 AlexNet 提出以后,基于深度卷积神经网络的模型开始取代传统图像分类算法成为 ILSVRC 图像分类比赛参赛队伍所采用的主流方法。ILSVRC2013 的获胜队伍 Clarifai[19] 提出了一套卷积神经网络的可视化方法,运用反卷积网络对 AlexNet 的每个卷积层进行可视化,以此来分析每一层所学习到的特征,从而加深了对于卷积神经网络为什么能够在图像分类上取得好的效果的理解,并据此改进了该模型,取得了 11.7%的错误率。

ILSVRC2014 的图像分类比赛结果相比于前一年取得了重大的突破,其中获胜队伍 Google 团队所提出的 GoogleNet[20] 以 6.7%的错误率将图像分类比赛的错误率降至以往最佳记录的一半。该网络有 22 层,受到赫布学习规则的启发,同时基于多尺度处理的方法对卷积神经网络作出改进。该文基于 Network in network[21] 思想提出了 Inception 模块。Inception 模块的结构如图 5 所示,它的主要思想是想办法找出图像的最优局部稀疏结构,并将其近似地用稠密组件替代。这样做一方面可以实现有效的降维,从而能够在计算资源同等的情况下增加网络的宽度与深度;另一方面也可以减少需要训练的参数,从而减轻过拟合问题,提高模型的推广能力。

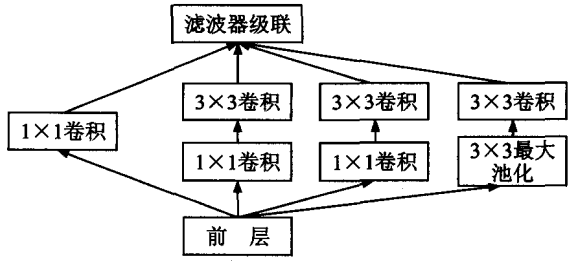

图 5 简化的 Inception 模块结构[20]

Fig. 5 Simplified structure of Inception module[20]

在 ILSVRC2014 中,以 8.1%的错误率获得第 3 名的是来自微软亚洲研究院的团队所设计的 SPP-Net[22],他们提出了一个新的池化方法,叫作空间金字塔池化,如图 6 所示。之前大部分卷积神经网络模型都要求输入图像大小固定,因此需要对原始图像进行剪切,这样将导致原始图像信息的丢失;或者需要对图像的大小和长宽比进行调整,这样会使图像产生扭曲变形。注意到卷积层对输入图像的大小没有限制,只有全连接层由于参数个数固定,需要保证输入的维数固定。然而卷积层的输出维数随着输

入维数的变化而变化，所以需要保证输入图像大小固定。因此，空间金字塔池化的作用是对任意维数的输入均产生固定维数的输出，从而使网络可以接受任意大小的图像作为输入。该池化方法将输入划分为固定个数的局部空间块，在每个块内进行最大池化，从而保证输出维数固定。采用多层次的空间块划分方法，则可以提取不同尺度的特征。

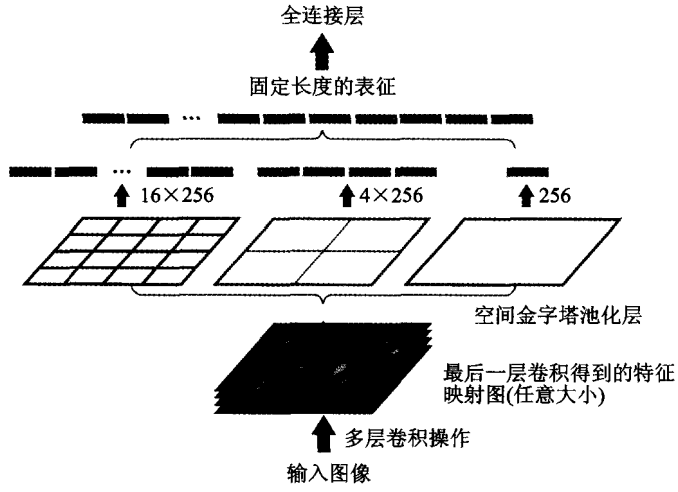

图 6 空间金字塔池化模型结构[22]

Fig. 6 Structure of spatial pyramid pooling model[22]

2015 年年初，微软亚洲研究院的研究人员提出的 PReLU-Nets[23]，在 ILSVRC 的图像分类数据集上取得了 4.94% 的 top-5 错误率，成为在该数据集上首次超过人眼识别效果（错误率约 5.1%[17]）的模型。该模型相比于以往的卷积神经网络模型有两点改进，一是推广了传统的修正线性单元（ReLU），提出参数化修正线性单元（PReLU）。该激活函数可以适应性地学习修正单元的参数，并且能够在额外计算成本可以忽略不计的情况下提高识别的准确率。同时，该模型通过对修正线性单元（ReLU/PReLU）的建模，推导出了一套具有鲁棒性的初始化方法，能够使得层数较多的模型（比如含有 30 个带权层的模型）收敛。

随后不久，Google 在训练网络时对每个 mini-batch 进行正规化，并称其为 Batch normalization，将该训练方法运用于 GoogleNet，在 ILSVRC2012 的数据集上达到了 4.82% 的 top-5 错误率[24]。归一化是训练深度神经网络时常用的输入数据预处理手段，可以减少网络中训练参数初始权重对训练效果的影响，加速收敛。于是 Google 的研究人员将归一化的方法运用于网络内部的激活函数中，对层与层之间的传输数据进行归一化。由于训练时使用随机梯度下降法，这样的归一化只能在每个 mini-batch 内进行，所以被命名为 Batch normalization。该方法可以使得训练时能够使用更高的学习率，减少训练时间；同时减少过拟合，提高准确率。

尽管卷积神经网络已经拥有强大的图像学习能力，然而这类模型缺乏对于图像空间不变性的学习，尤其是缺乏对于图像旋转不变性的学习[19]。Google DeepMind 提出的 Spatial transformer[25] 旨在通过提高卷积神经网络对于图像空间不变性的学习能力，来加强其图像分类的准确率。Spatial transformer 是可以在卷积神经网络的任意深度位置加入的模块，它可以将输入数据进行一系列空间变换，使得输出特征更加易于进行分类。在训练过程中，该模块可以自主地学习到空间变换所需要的参数，并且不需要在训练中增加任何额外的监督处理。

在 2015 年年底揭晓的 ImageNet 计算机视觉识别挑战赛 ILSVRC2015 的结果中，来自微软亚洲研究院团队所提出的深达 152 层的深层残差网络以绝对优势获得图像检测、图像分类和图像定位 3 个项目的冠军，其中在图像分类的数据集上取得了 3.57% 的错误率[26]。随着卷积神经网络层数的加深，网络的训练过程更加困难，从而导致准确率开始达到饱和甚至下降。该团队的研究人员认为，当一个网络达到最优训练效果时，可能要求某些层的输出与输入完全一致；这时让网络层学习值为 0 的残差函数比学习恒等函数更加容易。因此，深层残差网络将残差表示运用于网络中，提出了残差学习的思想。如图 7 所示，为了实现残差学习，将 Shortcut connection 的方法适当地运用于网络中部分层之间的连接，从而

保证随着网络层数的增加,准确率能够不断提高,而不会下降。

由于 ImageNet 具有数据集规模大、图像类别多等特点,运用 ImageNet 所训练的模型具有很强的推广能力,在其他数据集上也能取得良好的分类结果;而如果进一步在目标数据集上进行微调,与只用目标数据集进行训练相比大多能够获得更好的效果。首个将卷积神经网络很好地运用于物体检测的 R-CNN 模型[27],就是将使用 ImageNet 训练过的 AlexNet 模型在 PASCAL VOC 数据集上进行微调后用于提取图像特征,取得了比以往模型高出 20% 的准确率。除此之外,将用 ImageNet 数据集训练过的模型运用于遥感图像分类[28]、室内场景分类[29]等其他类型的数据集上,也取得了比以往方法更好的效果。

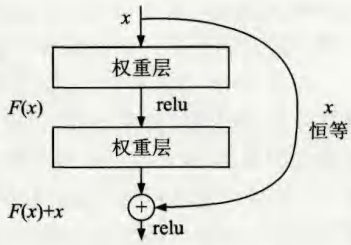

图 7 残差学习模块[26]

Fig. 7 Building block of residual learning[26]

从深度学习首次在 ILSVRC2012 中被运用于图像分类比赛并取得令人瞩目的成绩以来,基于深度学习方法模型开始在图像识别领域被广泛运用,新的深度神经网络模型的涌现在不断刷新着比赛记录的同时,也使得深度神经网络模型对于图像特征的学习能力不断提升。同时,由于 ImageNet,MS COCO 等大规模数据集的出现,使得深度网络模型能够得到很好的训练,通过大量数据训练出来的模型具有更强的泛化能力,能够更好地适应对于实际应用所需要的数据集的学习,提升分类效果。

3 物体检测

与图像分类比起来,物体检测是计算机视觉领域中一个更加复杂的问题,因为一张图像中可能含有属于不同类别的多个物体,需要对它们均进行定位并识别其种类。因此,在物体检测中要取得好的效果也比物体分类更具有挑战性,运用于物体检测的深度学习模型也会更加复杂。

AlexNet 在 ILSVRC2012 中所取得的成功不仅影响了图像分类方向的研究,也得到了计算机视觉领域中其他方向研究者的关注。在那个时期,物体检测仍然是运用传统算法进行,在 PASCAL VOC 等物体检测的标准测试数据集上的结果也没有取得较大的突破。因此,Ross Girshick 等便将卷积神经网络运用于物体检测中,提出了 R-CNN 模型[27]。如图 8 所示,该模型首先使用 Selective search[30] 这一非深度学习算法来提出待分类的候选区域,然后将每个候选区域输入到卷积神经网络中提取特征,接着将这些特征输入到线性支持向量机中进行分类。为了使得定位更加准确,R-CNN 中还训练了一个线性回归模型来对候选区域坐标进行修正,该过程被称为 Bounding box regression。该模型在 PASCAL VOC 的物体检测数据集上取得了比传统算法高大约 20% 的平均正确率均值,奠定了以后使用卷积神经网络进行物体检测的模型结构的基础。由于 PASCAL VOC 数据集比 ImageNet 数据集小,R-CNN 使用 Im-

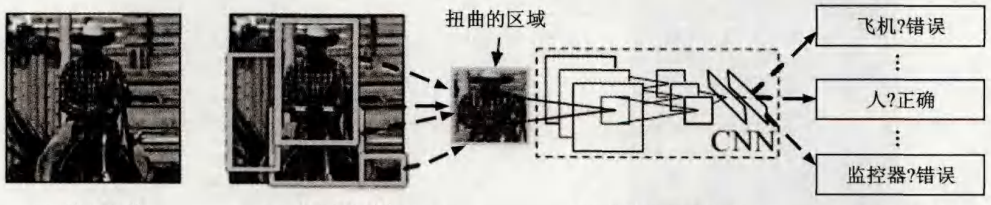

图 8 R-CNN 物体检测系统概览[27]

Fig. 8 Overview of R-CNN: Regions with CNN features[27]

ageNet 数据集对其中的卷积神经网络进行预训练，再将模型在 PASCAL VOC 数据集上进行微调，取得了更好的训练效果。这种微调方法也成为后来用于物体检测的深度学习模型常用的预处理手段。

在 R-CNN 模型中，对于每张图像大约产生 2 000 个候选区域，而对于每张图像，它的所有候选区域都要分别进行特征提取，这就使得特征提取所消耗的时间成为总的测试时间的瓶颈。微软亚洲研究院的研究团队将 SPP-Net 运用于物体检测中，并改进了 R-CNN 的这一缺陷。SPP-Net 针对用 Selective search 算法产生的候选区域，将这些区域的坐标投射到最高层卷积层所输出的特征映射的对应位置上，然后把每个候选区域所对应的特征输入到空间金字塔池化层，得到一个固定长度的特征表示。接下来的步骤与 R-CNN 相似，都是将这些特征表示输入到全连接层、将全连接层输出的特征输入到线性支持向量机进行分类以及使用 Bounding box regression 修正候选区域坐标。在 PASCAL VOC 上，该网络取得了与 R-CNN 相近的准确率，但是由于时间消耗大的卷积操作对于每张输入图像只进行了一次，使得总的测试所用时间大大减少。

SPP-Net 在物体检测方面虽然对 R-CNN 的图像处理流程做出了一定改进，但仍然存在着一些缺陷。在进行训练和测试的过程中，提出候选区域、提取图像特征以及根据特征进行分类 3 个过程形成了多阶段流水线。这样做的一个直接结果是需要额外的空间来存储提取出来的特征，供分类器使用。于是 R-CNN 的设计者之一 Ross Girshick 便对 R-CNN 提出了进一步的改进方案，称为 Fast R-CNN[31]，其结构如图 9 所示。与 R-CNN 中的卷积神经网络相比，Fast R-CNN 对最后一个池化层进行了改进，提出了 Region of interest (RoI) pooling 层。这个层的作用与 SPP-Net 用于物体检测网络中的空间金字塔池化层相似，作用都是对于任意大小的输入，输出固定维数的特征向量，只是 RoI pooling 层中只进行了单层次的空间块划分。这一改进使得 Fast R-CNN 与 SPP-Net 一样，可以将整张输入图像以及由 Selective search 算法产生的候选区域坐标信息一起输入卷积神经网络中，在最后一层卷积层输出的特征映射上对每个候选区域所对应的输出特征进行 RoI pooling，从而不再需要对每个候选区域都单独进行一次卷积计算操作。除此之外，Fast R-CNN 将卷积神经网络的最后一个 softmax 分类层改为两个并列的全连接层，其中一层仍为 softmax 分类层，另一层为 Bounding box regressor，用于修正候选区域的坐标信息。在训练过程中，Fast R-CNN 设计了一个多任务损失函数，来同时训练用于分类和修正候选区域坐标信息的两个全连接层。这种训练方式比之前 R-CNN 所采用的分阶段训练方式所得到的网络在 PASCAL VOC 数据集上取得了更好的检测效果，从而 Fast R-CNN 中不再需要额外训练 SVM 分类器，实现了从提取图像特征到完成检测的一体化。由于 Fast R-CNN 的卷积神经网络中卷积操作对于每张输入图像只需进行一次，而全连接层的计算对于每个候选区域均进行一次，从而使得全连接层的计算时间在总运行时间中占了很大的比例。因此，fast R-CNN 提出了用截断奇异值分解来加快全连接层运行速度的方

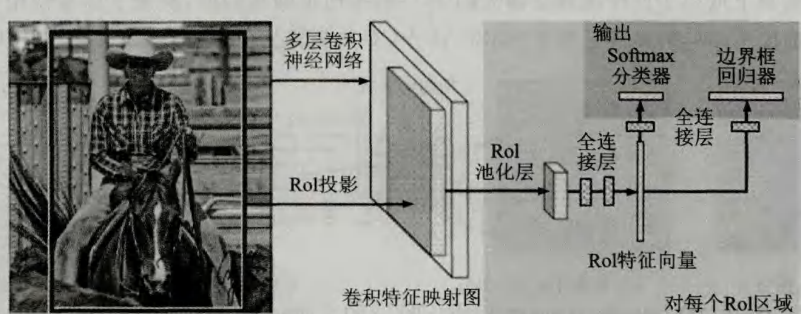

图 9 Fast R-CNN 结构示意图[31]

Fig. 9 Fast R-CNN architecture[31]

法。在进行检测时,对每个全连接层的权值矩阵进行截断奇异值分解,并将该全连接层用两个权值矩阵维数较小的全连接层替代。这样做对于测试的准确率没有太大影响,但可以大大加快检测速度。

DeepID-Net[32]也是在物体检测方面一个很有影响力的模型。该模型在 R-CNN 训练流程的基础上进行了进一步完善,改进了模型预训练方式,提出了 Bounding box rejection,Contextual modeling 等新的网络训练步骤。除此之外,在卷积神经网络结构中,DeepID-Net 在可变形部件模型(Deformable part model)[33]的启发下设计了新的池化层,叫作 Deformation constrained pooling(Def-pooling)层。这一池化层可以实现对图像局部信息的学习,并使得模型能够更好地适应于输入图像中某些部件位置发生偏移的情况。

这些模型在训练流程与卷积神经网络结构方面都做出了改进,然而都是采用传统算法来提出候选区域,这些算法均是在 CPU 上实现,使得计算候选区域的时间成为整个模型运行时间的瓶颈。因此,Ren Shaoqing 等设计的 Faster R-CNN 模型[34]中提出了候选区域网络来对这一步骤进行改进,其结构如图 10 所示。Faster R-CNN 网络在 Fast R-CNN 模型的基础上,在最后一层卷积层输出的特征映射上设置了一个滑动窗,该滑动窗与候选区域网络进行全连接。对于滑动窗滑过的每个位置,模型中给定若干个以滑动窗中心为中心、不同尺度与长宽比的锚点,候选区域网络将以每个锚点为基准相应地计算出一个候选区域。候选区域网络是一个全卷积网络[35],网络的第一层将滑动窗的输入特征映射到一个较低维的向量,然后将该向量输入到两个并列的全连接子层,其中分类层用于输出该向量对应图像属于物体还是背景的概率分布,回归层用于输出候选区域的坐标信息。为了让候选区域网络与用于检测的 Fast R-CNN 模型的前几层卷积层能够实现共享,从而提高这些卷积层所提取特征的利用率与运行效率,Faster R-CNN 提出了一套多阶段训练算法进行网络训练。由于 Faster R-CNN 提出候选区域的过程是根据用于检测的 Fast R-CNN 网络的前几层卷积层所提取的特征,且候选区域网络也在 GPU 上实现,从而提出候选区域的时间开销大大减少,检测所需时间约为原来时间的 1/10,且准确率也有所提高,说明候选区域网络不仅能更加高效地运行,还能提高所产生的候选区域的质量。

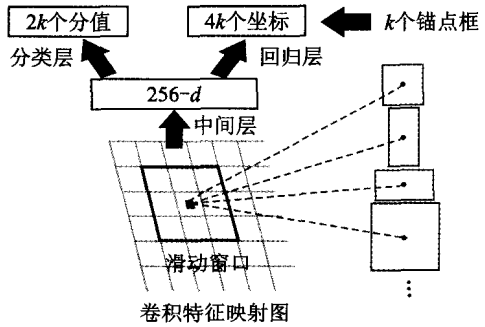

图 10 Region proposal network 结构[34]

Fig. 10 Region proposal network (RPN)[34]

由于当下基于卷积神经网络的物体检测模型大多将物体检测问题归结为如何提出候选区域和如何对候选区域进行分类两个子问题,因此物体检测问题比图像分类问题难度更高,解决起来步骤更加复杂,对模型的性能要求也更高。在物体检测的发展过程中,不仅卷积神经网络本身的结构得到了改进,更多的模型侧重于优化训练方法与流程。在这一过程中,物体检测模型在准确率不断提升的同时,运行时间也不断减小,从而使其能够被更好地投入到实际应用中。

4 姿态估计

除了大家熟知的图像分类和目标检测任务外,实际上随着各种网络游戏的发展、动画视频的普及,正确快速地识别和理解图像中人的姿态动作也成为了一个非常热门的话题。这种问题统称为姿态检测。姿态检测中包含许多类别和子问题,姿态估计就是其中之一。姿态估计是时下最为重要的计算机视觉挑战性问题之一,原因在于它可以被很快地应用到人物追踪、动作识别以及视频相关的视频分析上,比如视频监控和视频搜索等实际应用面非常广。

姿态估计的主要任务就是,给定一张图,图中会有一个人,你需要给出这个人的姿态动作是什么样的。人们会提前选定出比较重要的几个人体关节(比如肩膀、肘部、脖子等),然后用一个固定维数(比如

7 维和 11 维)的坐标向量来表示这个动作,每一维都表示图中人物的重要关节所在的具体坐标,如图 11 所示。换句话说,你需要给出一个火柴人的形状来表示这个人的姿态。

图 11 姿态估计[41]

Fig. 11 Pose estimation[41]

传统的处理这类问题的方法主要是基于 Part-based models[36],Region-based models[37]和 Pictorial structures-base models[38]这些模型。之后,Yang Y 和 Ramanan D 又提出了混合模型的方法[39]。但这些模型由于大部分都是对局部建模,所以表达能力有限,而且往往只能针对局部进行有效的分析,在实际应用中很难有所作为。文献[40]提出了一种基于深度图像的人体运动姿态跟踪和识别算法,对克服光照影响具有较强的鲁棒性。但自从深度神经网络引入后,处理这类问题的方法有了很大的改变。2014 年,Google 研究员 Alexander Toshev 和 Christian Szegedy 发表了一篇将深度学习的方法引入到姿态估计里面[40]的文章,率先将深度学习引入到了姿态估计上。

图 12 一个人的姿态被一个 bounding box 围了起来[41]

Fig. 12 Someone's pose bounded by a bounding box[41]

毫无疑问,深度神经网络是具备处理整张图,而不是一个局部大小图片的能力的。而且,随着深度神经网络的出现,特征表示、模型的拓扑关系以及对连接点之间的关系这些传统模型里必备的要素现在都不需要了。DeepPose 提出了一种新的姿态估计方法。首先,在关节个数固定、关系固定,只要确定关节坐标的前提下,他们用一个向量来表示这个人,同时注意到一个实际的问题,图片中的人可大可小,可能同一个姿态,一张图片里面这个人占了全图,另一个人只占了一小部分,所以为了能够让神经网络在辨认姿态时不受姿态大小的影响,第一步就要进行正规化操作,将所有姿态统一放缩到某个程度,如果能够找到一个非常紧的 Bounding box 来把这个姿态围住(如图 12 所示),那可以认为这个 Bounding box(大小为 $h\times w$)就是这个姿势的大小。然后,压缩这个 Bounding box,将其从 $h\times w$ 的矩形按比例压缩成一个 1×1 的矩形。这样所有的姿势都可以看成一个被 1×1 矩形比较紧地围着的姿势,不同大小的同一种姿势在这种压缩后就会变得一样,也对网络的训练带来了好处。做好这些准备工作后,首先用基于 AlexNet 的深度神经网络对原图进行一个粗略的估计,使得训练出的模型能对输入图片中的人物姿态得到一个大概的估计。希望用 AlexNet 得到一个 k(k 是选出的重要关节的数量)维的向量来表示这个姿态。网络结构如图 13 所示。这个网络由 5 个卷积层,3 个 pooling 层,还有 3 个全连接层组成。每一层都会提取一些特征并继续进入下一层进行训练。最后,通过最后一个全连接层得到一个 $2k$ 维的向量,作为输出结果。当然,这样得到的仍然是一个正规化了的姿态坐标,所以如果要得到原图大小,还需要一个逆操作。

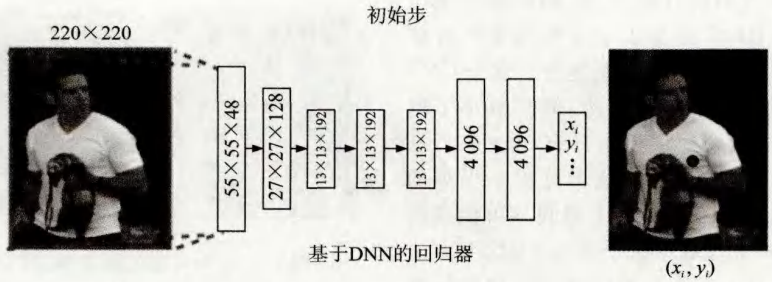

图 13 基于深度神经网络的姿态估计训练网络初始步[41]

Fig. 13 Initial pose estimation training stage based on CNN[41]

这样得到的结果十分粗糙。可以用原图的一个子图来优化局部的定位。根据人体运动学的相关研究结论,他们选取当前粗糙估计所得的关节点坐标为中心,左肩到右臀的距离为边长,进一步进行优化。优化方式就是针对所得的图像的局部,进一步用深度神经网络优化,以期得到更好的模型。如图 14 所示,网络的结构仍然十分类似于 AlexNet,只不过最后的输出层改为输出一个坐标的修正值。

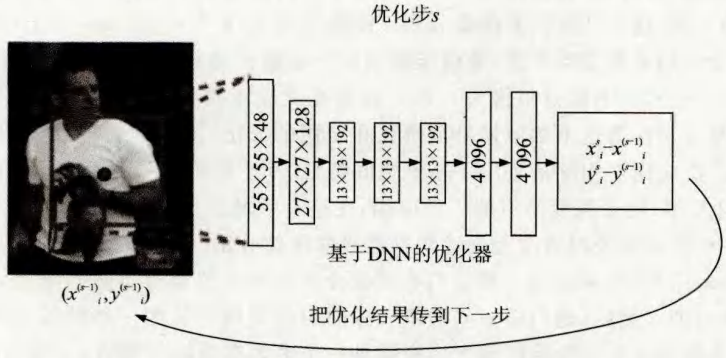

图 14 基于深度神经网络的姿态估计训练网络优化步[41]

Fig. 14 Pose estimation refining stage based on CNN[41]

这一方法在各个数据集上都取得了非常优异的结果。不论以 PCP 标准还是 PDJ 标准,这个方法都基本达到了当时能达到的最佳结果。这也引发了之后深度学习在这个领域的进一步发展。比如,UCLA 学生 Xianjie Chen 和他的导师 Alan Yuille 提出了结合 Graphical model 和 Deep neural networks 的方法[42],在测试集上取得的结果又有了更大的进步。他们首先建立一个图模型,这个图模型中的点表示人的重要关节,边表示关节之间的关系,考虑到人体结构的特殊性,这个图模型构建出的图实际上是一棵树,这棵树就表示了人体的一个骨骼框架,骨骼框架包含关节和关节之间的关系。为了评价针对一张图预测出的骨骼框架的好坏,他们给出了一个分数函数来评价结果。分数函数分为两部分:针对点的和针对边的。抽象来说,给出了一个预测出的骨骼框架,它的关节预测得准不准是一方面,关节之间的关系预测得准不准是另一方面。根据训练数据集,基于图片以及给出的真实骨骼框架的标记得出针对每个关节,它可能是与周围关节的连边方式的集合。之后对图片的关节片段进行特定的标记(这个标记一定是之前的集合中的某种情况,可以有一定误差),然后利用 CNN 网络训练出一个网络结构,使得

针对每个有关关节的图片的片段，都能够预测出这个关节位置。图 15 给出了一个数据集中可能的肘关节与周围关节之间的关系集合。这个 CNN 网络用的仍然是 AlexNet 的结构，但是减小了网络规模(比如第一层卷积层大小只有 54×54)，从而减少了参数。此外，最后希望训练出来得到的预测骨骼框架能够在分数函数中得到尽可能高的分数，从而获得非常好的结果。事实上，这个方法验证下来确实收获了非常好的结果。在 PCP 标准下，这种方法将有关肢体的预测准确率平均提升了 5.8%，效果惊人。尤其在对小臂的肢体预测上，取得了高达 10.9%的进步。这些成功，都表明着深度神经网络在未来对于姿态估计领域不可限量的潜力和广泛的应用前景。毫无疑问，深度神经网络在未来将在姿态估计甚至动作检测领域扮演重要角色。

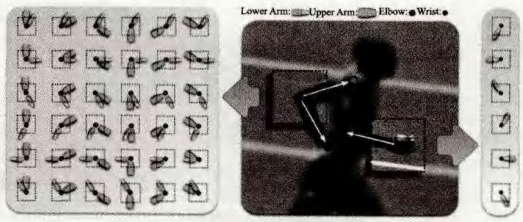

图 15 一个可能的肘关节与周围关节的连边关系集合[42]

Fig. 15 A possible set of relations between elbow and other joints[42]

5 图像分割

深度神经网络在图像分类、目标检测和姿态估计等方面取得了巨大的成功，进一步的发展便是对图像上每个像素点的预测，这个任务就是图像分割。图像分割是这样一类问题：对于一张图来说，图上可能有多个物体、多个人物甚至多层背景，希望能做到对于原图上的每个像素点，能预测它是属于哪个部分的(人、动物、背景……)。图像分割作为许多计算机视觉应用研究的第一步十分关键。在过去的 20 年中，图像阈值分割方法作为这个领域最早被研究和使用的方法，因为其物理意义明确、效果明显和易于实现等特点，被广泛应用。相继衍生出了基于空间特征、基于模糊集和基于非 Shannon 熵的许多阈值选取方法[43]。但这几年，随着深度学习的广泛应用，在这一领域显然有了更新、更有力的“工具”。文献[35]提出可以将一些深度神经网络改为全卷积网络来做图像分割。他们首先利用一些流行的分类网络(AlexNet, VGG, GoogleNet)，在保留一些它们在图像分类方面训练所得参数基础上，进行“修剪”，转变为针对图像分割的模型。然后，他们将一些网络较深的层的所得特征和一些较浅的层所得特征结合起来，最后用一个反卷积层放大到原始图像大小来提供一个更为准确的分割结果，称之为跳跃结构。

仍然拿 AlexNet 为例，如图 16 所示。他们提出将 AlexNet 的最后 3 层改为全卷积层，这一步不仅

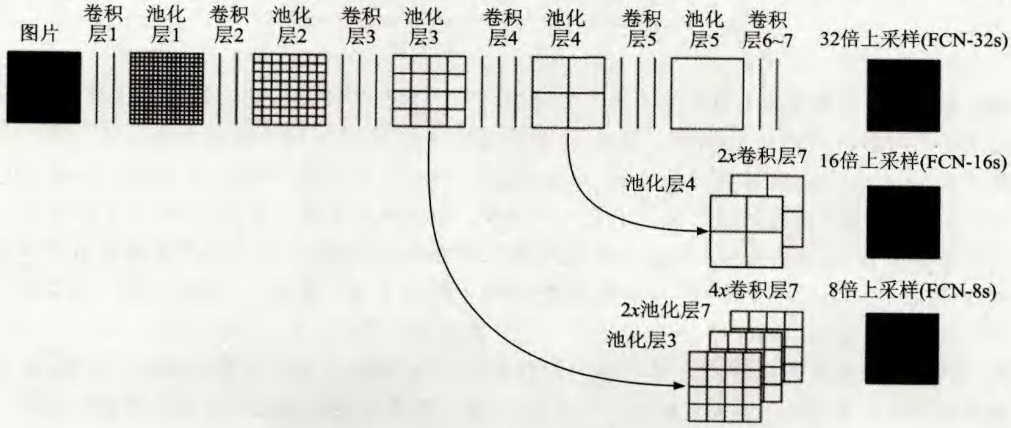

图 16 基于 AlexNet 的全卷积深度神经网络[35]

Fig. 16 Fully convoluted neural network based on AlexNet[35]

加快了速度,减少了参数,进而减少过拟合,还为最后一步的反卷积提供了便利。可以看到,这个网络结构已经变成了7层卷积层这样一个结构。当然,如果直接从最后一层的卷积层反卷积也可以,但需要放大32倍,取得的效果也不佳(如图17所示)。这毫无疑问是缺少信息所导致的结果。卷积神经网络每卷积一层,实际上信息量都会丢失一些,所以如果想增加信息量,要做的就是到更浅的网络层获取信息,这是之前所提的跳跃结构的核心思想。对于信息丢失过多的最后一层卷积层,可以先将它反卷积扩大1倍,达到与上一个卷积层pooling完了之后的一样的大小,之后,将两者的信息整合(一般是相加),进而做次反卷积,这样就只需再放大16倍,取得的效果也有所提升。可以更进一步,再加入pool3的信息,也就是将之前一步的结果先再做一次扩大2倍的反卷积(相较于最初实际上相当于扩大了4倍),与pool3的结果相加后,再做一次放大8倍的反卷积。

(a) FCN-32s方法 (a) FCN-32s (b) FCN-16s方法 (b) FCN-16s (c) FCN-8s方法 (c) FCN-8s (d) 真实值 (d) Ground truth

图17 各种方法结果的比较[35]

Fig. 17 Results of different methods[35]

在数据集PASCAL VOC上,他们所得的结果较2012年提升了约20%,达到62.2%的mean IU准确率。较传统的诸如SDS方法[44]提升了许多,影响巨大。此外,这种方法训练也只花了175 ms,传统的SDS方法耗时高达50 s。在另一个数据集NYUDv2上,全卷积神经网络(Fully convolutional networks, FCN)也将之前的最好结果提升了至少5%。当然,FCN仍然有不足之处,图18所示的是一些PASCAL VOC上的结果中,最后一个就失败了,说明这种方式仍有改进空间。

(a) 原始图片 (a) Original images (b) 真实值 (b) Ground truth (c) 传统方法结果 (c) Results of SDS method (d) FCN-8s (d) Results of FCN-8s method

图18 FCN,SDS在PASCAL VOC上的一些结果[35]

Fig. 18 Some results given by FCN and SDS on PASCAL VOC[35]

6 人脸识别

人脸识别是图像识别领域一个非常重要的研究方向,由于人脸图像具有易采集的特性,因此受到了许多行业的关注,具有非常广阔的应用前景和巨大的商业市场。人脸识别技术主要包括人脸检测、人脸特征提取和人脸识别3个过程。人脸检测是从输入图像或视频流中检测并提取人脸图像,并进一步给出人脸的位置、大小以及各个主要面部器官的位置信息。通常是采用Haar特征和Adaboost算法训练级联分类器对图像中的每一个矩形子区域进行分类。特征提取是通过一组数据来表征人脸信息,这组数据就是所要提取的人脸特征。常见的人脸特征分为几何特征和表征特征。几何特征是指各个主要面部器官之间的几何关系,如距离、面积和角度等。这种特征

相对比较直观，且计算量较少，但由于算法所需的特征点无法精确选择，同时对于光照、遮挡、姿态和表情等因素引起的变化不具有很好的鲁棒性，因此只适合于人脸图像的粗略识别，无法在实际中应用。表征特征利用人脸图像的灰度或色彩信息通过一些算法提取全局或局部特征。这里的人脸识别是指狭义的人脸识别，即将待识别人脸所提取的特征与数据库中已有的特征信息进行对比，然后进行判别分类。人脸识别主要包括两类任务：一类是人脸验证，这是人脸图像与数据库中已有的人脸图像进行比对的过程，以此来判断两张人脸图像是否属于同一个人，属于相对简单的二分类问题。另一类是人脸辨识，这是人脸图像与数据库中所有的人脸图像进行匹配的过程，以此来对属于不同人的人脸图像进行分类，属于相对困难的多分类问题。

在深度学习出现之前，人脸识别采用的主流方法是以 Eigenfaces[45] 为代表的子空间分析方法。香港中文大学汤晓鸥教授所带领的研究团队，将当时最为流行的 3 种子空间方法——主成分分析子空间 (Principal component analysis, PCA)[45]、贝叶斯子空间[46]和线性判别分析子空间 (Linear discriminant analysis, LDA)[47]有机地结合到同一个理论框架中，提出了统一子空间分析[48]方法，这种方法使用 LBP 和 Gabor 等特征对人脸图像中邻域像素区块的灰度值或颜色值进行局部特征提取，然后对这些局部特征进行特征变化，得到更易于区分的人脸表示，在人脸识别领域最受关注的测试集 LFW[49] 上取得了当时的最佳识别性能。

近两年来，基于深度学习的人脸识别在 LFW 数据集上的识别率取得了极大的提高。传统的人脸识别算法 Eigenface 在 LFW 上的识别率只有 60%左右，而目前最好的深度学习算法的识别率已经达到了 99.47%，甚至超过了人类在该测试集上的识别水平 (99.25%)[50]。2012 年，Learned-Miler 等[51]采用无监督的特征学习方法，首先将深度学习用于 LFW 的人脸识别，取得了 87%的识别率。2013 年，来自香港中文大学的研究小组首先采用人脸验证任务作为监督信号来学习人脸特征表示，将人脸图像划分成若干个重叠的区域，对每个子区域都训练一个卷积神经网络，并将多个网络的最后一个隐含层的特征进行融合，作为受限玻尔兹曼机的输入，通过反向传播算法进行整体联合优化，最终取得了 92.52%的识别率[52]。随后一年，该研究小组 (DeepID[53]) 和来自 Facebook 的研究人员 (DeepFace[54]) 均采用人脸辨识任务作为监督信号，通过多层的卷积神经网络学习人脸特征表示。然后将训练得到的特征用于 LFW 上的人脸验证任务，都取得了高于 97%的识别率，大大超过了传统的识别算法。在随后的工作中，DeepFace 使用了更大规模的训练集，将识别率提高到了 98.4%[55]。DeepID 的研究人员则同时将人脸验证和人脸辨识两个任务作为监督信号，进行联合深度学习，推出了 DeepID2[56] 模型，进一步将识别率提高到了 99.15%。最新的研究成果显示该研究小组将监督信号加到卷积神经网络的每一层，同时采用更深的网络结构和更多的训练数据，推出了 DeepID2+[57] 模型，在 LFW 上取得了 99.47%的准确率，这是目前最好的识别结果。

这些最新的研究成果表明，使用深度学习方法提取到的人脸特征表示具有传统手工特征表示所不具备的重要特性，例如这些特征是中度稀疏的，对人脸身份和人脸属性具有很强的选择性，对局部遮挡、光照变化和表情变化等具有良好的鲁棒性。这是特性都是通过在海量的图像数据上训练自然得到的，网络模型中并没有添加任何显式的约束条件，得到的人脸特征也没有进行其他后期的处理。这说明深度学习并非是单纯地使用具有大量参数的、非常复杂的非线性神经网络模型去拟合数据集，而是通过逐层训练学习，最终得到蕴涵清晰的语义信息的特征表示，从而大大提高了识别率。

7 结束语

深度学习目前是一个非常热门的研究方向，利用卷积神经网络的卷积层、池化层和全连接层等基本结构，就可以让这个网络结构自己学习和提取相关特征，并加以利用。这种特性对许多研究提供了许多便利，可以省略过往非常繁杂的建模过程。此外，深度学习现在在图像分类、物体检测、姿态估计和图像

分割等方面都已经有了非常大的成果和进步。一方面,深度学习应用面非常广,而且通用性强,完全可以继续努力将其拓展到其它应用领域。另一方面,深度学习仍有许多潜力可挖,值得不断去探索和发现。就未来而言,尽管之前讨论的许多内容都是有监督的学习(比如训练的网络最后一层会根据真实值计算一个 loss 值,进而进行参数调整),并且有监督的学习确实取得了非常大的成功。深度学习在无监督的学习方面的应用很可能是未来的发展趋势。毕竟,就人或者动物而言,大部分情况下,我们并不是通过知道事物的名字来了解它是什么的。在未来的计算机视觉领域,预计基于深度学习的卷积神经网络和循环神经网络(Recurrent neural network, RNN)将会成为十分流行的网络模型,并将在更多的应用研究中取得更好的突破与进展。此外,结合强化学方法来训练一个端到端的学习系统逐渐成为可能,从而使得该学习系统具有自主学习能力,能够主动去学习相关特征的表示和抽象。目前,结合深度学习与强化学习的研究尚处于起步阶段,但已经有一些这方面的研究工作在多物体识别任务[58]和进行视频游戏的学习[59]上取得了不错的表现,这也是让许多相关领域的研究者们兴奋的原因之一。值得注意的是,自然语言处理同样也是深度学习未来能够大展身手的潜在舞台,比如说,对于一篇文章或者一大段文字,能够设计出基于一些深度神经网络模型(比如 RNN)的方法和策略,能够有效地理解文本内容。总体来说,人们现在使用深度学习以及一些简单的推理,就已经在语音和图像领域取得了非常不错的成果。有理由相信,如果将目前对于网络提取的特征表示能够进一步优化,使得其能够更"自如"地表达特征,再加上一些复杂推理,那么深度学习将会在人工智能的各个应用方面取得更大的进展。

参考文献:

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

[2] Dalal N, Triggs B. Histograms of oriented gradients for human detection[C]//Computer Vision and Pattern Recognition (CVPR), IEEE Computer Society Conference on. San Diego, USA: IEEE, 2005, 1: 886-893.

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

[4] 麦麦提艾力·吐尔逊,戴礼荣。深度神经网络在维吾尔语大词汇量连续语音识别中的应用[J]. 数据采集与处理, 2015, 30(2): 365-371.
Maimaitiaili Tuerxun, Dai Lirong. Deep neural network based uyghur large vocabulary continuous speech recognition[J]. Journal of Data Acquisition and Processing, 2015, 30(2): 365-371.

[5] 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): 106-154.

[6] 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.

[7] Ruck D W, Rogers S K, Kabrisky M. Feature selection using a multilayer perceptron[J]. Journal of Neural Network Computing, 1990, 2(2): 40-48.

[8] Rumelhart D E, Hinton G E, Williams R J. Learning representations by back-propagating errors[J]. Nature, 1986, 323: 533-538.

[9] LeCun Y, Denker J S, Henderson D, et al. Handwritten digit recognition with a back-propagation network[C]//Advances in Neural Information Processing Systems. Colorado, USA: [s. n.], 1990: 396-404.

[10] LeCun Y, Cortes C. MNIST handwritten digit database[EB/OL]. http://yann.lecun.com/exdb/mnist, 2010.

[11] Waibel A, Hanazawa T, Hinton G, et al. Phoneme recognition using time-delay neural networks[J]. Acoustics, Speech and Signal Processing, IEEE Transactions on, 1989, 37(3): 328-339.

[12] Krizhevsky A. Learning multiple layers of features from tiny images[D]. Toronto, Canada: University of Toronto, 2009.

[13] Fei-Fei L, Fergus R, Perona P. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories[J]. Computer Vision and Image Understanding, 2007, 106(1): 59-70.

[14] Griffin G, Holub A, Perona P. Caltech-256 object category dataset[R]. Technical Report 7694, http://authors.library.caltech.edu/7694, California Institute of Technology, 2007.

[15] Deng J, Dong W, Socher R, et al. Imagenet: A large-scale hierarchical image database[C]//Computer Vision and Pattern Recognition (CVPR), IEEE Conference on. Miami, USA: IEEE, 2009: 248-255.

[16] Ahonen T, Hadid A, Pietikainen M. Face description with local binary patterns: Application to face recognition[J]. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 2006, 28(12): 2037-2041.

[17] Russakovsky O, Deng J, Su H, et al. Imagenet large scale visual recognition challenge[J]. International Journal of Computer Vision, 2015,115(3):211-252.

[18] Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks[C]//Advances in Neural Information Processing Systems. Cambridge: MIT Press, 2012: 1097-1105.

[19] Zeiler M D, Fergus R. Visualizing and understanding convolutional networks[M]. New York: Springer International Publishing, 2014: 818-833.

[20] Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions[C]//Computer Vision and Pattern Recognition (CVPR), IEEE Conference on. Boston, USA: IEEE, 2015: 1-9.

[21] Lin M, Chen Q, Yan S. Network in network[EB/OL]. http://arxiv.org/abs/1312.4400, 2013.

[22] He K, Zhang X, Ren S, et al. Spatial pyramid pooling in deep convolutional networks for visual recognition[M]//Computer Vision-ECCV 2014. New York: Springer International Publishing, 2014: 346-361.

[23] He K, Zhang X, Ren S, et al. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification [EB/OL]. http://arxiv.org/abs/1502.01852, 2015.

[24] Ioffe S, Szegedy C. Batch normalization: Accelerating deep network training by reducing internal covariate shift[EB/OL]. http://arxiv.org/abs/1502.03167, 2015.

[25] Jaderberg M, Simonyan K, Zisserman A. Spatial transformer networks[C]//Advances in Neural Information Processing Systems. Montréal, Canada: [s. n.], 2015: 2008-2016.

[26] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[EB/OL]. http://arxiv.org/abs/1512.03385, 2015.

[27] Girshick R, Donahue J, Darrell T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation [C]//Computer Vision and Pattern Recognition (CVPR), IEEE Conference on. Columbus, USA: IEEE, 2014: 580-587.

[28] Castelluccio M, Poggi G, Sansone C, et al. Land use classification in remote sensing images by convolutional neural networks [EB/OL]. http://arxiv.org/abs/1508.00092, 2015.

[29] Hayat M, Khan S H, Bennamoun M, et al. A spatial layout and scale invariant feature representation for indoor scene classification[EB/OL]. http://arxiv.org/abs/1506.05532, 2015.

[30] Uijlings J R R, van de Sande K E A, Gevers T, et al. Selective search for object recognition[J]. International Journal of Computer Vision, 2013, 104(2): 154-171.

[31] Girshick R. Fast R-CNN[EB/OL]. http://arxiv.org/abs/1504.08083, 2015.

[32] Ouyang W, Wang X, Zeng X, et al. Deepid-net: Deformable deep convolutional neural networks for object detection[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE Conference on. Boston, USA: IEEE, 2015: 2403-2412.

[33] Felzenszwalb P F, Girshick R B, McAllester D, et al. Object detection with discriminatively trained part-based models[J]. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 2010, 32(9): 1627-1645.

[34] Ren S, He K, Girshick R, et al. Faster R-CNN: Towards real-time object detection with region proposal networks[C]//Advances in Neural Information Processing Systems. Montréal, Canada: [s. n.], 2015: 91-99.

[35] Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation[C]//Computer Vision and Pattern Recognition (CVPR), IEEE Conference on. Boston, USA: IEEE, 2015: 3431-3440.

[36] Nevatia R, Binford T O. Description and recognition of curved objects[J]. Artificial Intelligence, 1977, 8(1): 77-98.

[37] Felzenszwalb P F, Huttenlocher D P. Pictorial structures for object recognition[J]. International Journal of Computer Vision, 2005, 61(1): 55-79.

[38] Eichner M, Ferrari V, Zurich S. Better appearance models for pictorial structures[C]//Proceedings of the British Machine Vision Conference. London, UK: BMVA Press, 2009, 2: 5.

[39] Yang Y, Ramanan D. Articulated pose estimation with flexible mixtures-of-parts[C]//Computer Vision and Pattern Recognition (CVPR), IEEE Conference on. Colorado Springs, USA: IEEE, 2011: 1385-1392.

[40] 杨凯,魏本征,任晓强,等. 基于深度图像的人体运动姿态跟踪和识别算法[J]. 数据采集与处理, 2015, 30(5): 1043-1053. Yang Kai, Wei Benzheng, Ren Xiaoqiang, et al. Depth image based human motion tracking and recognition algorithm[J].

Journal of Data Acquisition and Processing, 2015, 30(5): 1043-1053.

[41] Toshev A, Szegedy C. Deeppose: Human pose estimation via deep neural networks[C]//Computer Vision and Pattern Recognition (CVPR), IEEE Conference on. Columbus, USA: IEEE, 2014: 1653-1660.

[42] Chen X, Yuille A L. Articulated pose estimation by a graphical model with image dependent pairwise relations[C]//Advances in Neural Information Processing Systems. Montréal, Canada: [s. n.], 2014: 1736-1744.

[43] 吴一全, 孟天亮, 吴诗婳. 图像阈值分割方法研究进展 20 年(1994—2014)[J]. 数据采集与处理, 2015, 30(1):1-23.
Wu Yiquan, Meng Tianliang, Wu Shihua. Research progress of image thresholding methods in recent 20 years (1994—2014) [J]. Journal of Data Acquisition and Processing, 2015, 30(1):1-23.

[44] Hariharan B, Arbeláez P, Girshick R, et al. Simultaneous detection and segmentation[M]//Computer Vision-ECCV 2014. New York: Springer International Publishing, 2014: 297-312.

[45] Turk M, Pentland A. Eigenfaces for recognition[J]. Journal of Cognitive Neuroscience, 1991, 3(1): 71-86.

[46] Moghaddam B, Jebara T, Pentland A. Bayesian face recognition[J]. Pattern Recognition, 2000, 33(11): 1771-1782.

[47] Belhumeur P N, Hespanha J P, Kriegman D J. Eigenfaces vs. fisherfaces: Recognition using class specific linear projection [J]. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 1997, 19(7): 711-720.

[48] Wang X, Tang X. A unified framework for subspace face recognition[J]. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 2004, 26(9): 1222-1228.

[49] 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 0749, University of Massachusetts, Amherst, 2007.

[50] 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. Kyoto, Japan: IEEE, 2009: 365-372.

[51] Huang G B, Lee H, Learned-Miller E. Learning hierarchical representations for face verification with convolutional deep belief networks[C]//Computer Vision and Pattern Recognition (CVPR), IEEE Conference on. Providence, USA: IEEE, 2012: 2518-2525.

[52] Sun Y, Wang X, Tang X. Hybrid deep learning for face verification[C]//Computer Vision (ICCV), 2013 IEEE International Conference on. Sydney, Australia: IEEE, 2013: 1489-1496.

[53] Sun Y, Wang X, Tang X. Deep learning face representation from predicting 10,000 classes[C]//Computer Vision and Pattern Recognition (CVPR), IEEE Conference on. Columbus, USA: IEEE, 2014: 1891-1898.

[54] Taigman Y, Yang M, Ranzato M A, et al. Deepface: Closing the gap to human-level performance in face verification[C]// Computer Vision and Pattern Recognition (CVPR), IEEE Conference on. Columbus, USA: IEEE, 2014: 1701-1708.

[55] Taigman Y, Yang M, Ranzato M A, et al. Web-scale training for face identification [C]//Computer Vision and Pattern Recognition (CVPR), IEEE Conference on. Boston, USA: IEEE, 2015: 2746-2754.

[56] Sun Y, Chen Y, Wang X, et al. Deep learning face representation by joint identification-verification[C]//Advances in Neural Information Processing Systems. Montréal, Canada: [s. n.], 2014: 1988-1996.

[57] Sun Y, Wang X, Tang X. Deeply learned face representations are sparse, selective, and robust[C]//Computer Vision and Pattern Recognition (CVPR), IEEE Conference on. Boston, USA: IEEE, 2015: 2892-2900.

[58] Ba J, Mnih V, Kavukcuoglu K. Multiple object recognition with visual attention[EB/OL]. http://arxiv. org/abs/1412. 7755, 2014.

[59] Mnih V, Kavukcuoglu K, Silver D, et al. Human-level control through deep reinforcement learning[J]. Nature, 2015, 518 (7540): 529-533.

作者简介:

卢宏涛(1967-),男,教授,博士生导师,研究方向:机器学习、模式识别、计算机视觉和图像分析与处理,E-mail: htlu@sjtu. edu. cn。

张秦川(1993-),男,博士研究生,研究方向:深度学习、图像分类和目标检测。