卷积神经网络的发展及其在计算机视觉领域中的应用综述

陈 超 齐 峰
(山东师范大学管理科学与工程学院 济南 250000)

摘 要 近年来,深度学习在计算机视觉、语音识别、自然语言处理和医疗影像处理等领域取得了一系列显著的研究成果。在不同类型的深度神经网络中,卷积神经网络得到了最广泛的研究,这不仅体现在学术研究领域的繁荣,更体现在对相关产业产生了巨大的现实影响和商业价值上。随着标注样本数据集的快速增长和图形处理器(GPU)性能的大幅度提高,卷积神经网络的相关研究得到了迅速的发展,并在计算机视觉领域的各种任务中成效卓然。首先,回顾了卷积神经网络的发展历史;其次,介绍了卷积神经网络的基本结构及各组件的作用;然后,详细描述了卷积神经网络在卷积层、池化层和激活函数等方面的改进研究,总结了自1998年以来比较有代表性的神经网络架构:AlexNet,ZF-Net,VGGNet,GoogLeNet,ResNet,DenseNet,DPN和SENet;在计算机视觉领域,重点介绍了卷积神经网络在图像分类/定位、目标检测、目标分割、目标跟踪、行为识别和图像超分辨率重构等应用方面的最新研究进展;最后,对卷积神经网络研究中亟待解决的问题与挑战进行了总结。

关键词 人工智能,深度学习,卷积神经网络,计算机视觉
中图法分类号 TP183 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2019.03.008

Review on Development of Convolutional Neural Network and Its Application in Computer Vision

CHEN Chao QI Feng
(School of Management Science and Engineering, Shandong Normal University, Jinan 250000, China)

Abstract In recent years, deep learning has achieved a series of remarkable research results in various fields such as computer vision, speech recognition, natural language processing and medical image processing. In different types of deep neural networks, convolution neural network has obtained most extensive study, not only reflecting the prosperity in academic field, but also making a tremendous realistic impact and commercial value on the related industries. With the rapid growth of annotation sample data sets and the drastic improvement of GPU performance, related researches on convolutional neural networks are rapidly developed and have achieved remarkable results in various tasks in the field of computer vision. This paper reviewed the history of convolution neural network firstly. Then it introduced the basic structure of convolutional neural network and the function of each component. Next, it described the improvements of convolution neural network in convolution layer, pooling layer and activation function in detail. Also, it summarized typical neural network architectures since 1998 (such as AlexNet, ZF-Net, VGGNet, GoogLeNet, ResNet, DenseNet, DPN and SENet). In the field of computer vision, this paper emphatically introduced the latest research progresses of convolution neural network in image classification / localization, target detection, target segmentation, target tracking, behavior recognition and image super-resolution reconstruction. Finally, it summarized the problems and challenges to be solved about convolutional neural network.

Keywords Artificial intelligence, Deep learning, Convolution neural network, Computer vision

1 引言

卷积神经网络(Convolutional Neural Network,CNN)是一种广为人知的深度学习架构,其设计灵感来自生物体的自然视觉感知机制。1959年,Hubel等[1]发现动物视觉皮层中的细胞负责检测感受野(receptive field)中的光。受此发现的启发,日本科学家Fukushima于1980年前后提出了一种层级化的多层人工神经网络——神经认知机(neocognitron)[2-3]。神经认知机模型由多种类型的细胞单元组成,其中最重要的两种细胞单元称为“S型细胞”和“C型细胞”。S型细胞用于

到稿日期:2018-03-05 返修日期:2018-06-27 本文受国家自然科学基金项目(61502283,61472231,61640201)资助。
陈 超(1992—),男,硕士生,主要研究方向为机器学习、计算机视觉;齐 峰(1982—),男,博士,副教授,主要研究方向为机器学习、计算机视觉和数据挖掘,E-mail:cliff@sdnu.edu.cn(通信作者)。

提取局部特征(如边缘或角等);C 型细胞对 S 型细胞的输入进行一些处理,如图像较小的位移或轻微变形等。1990 年,LeCun 等[4]提出了现代 CNN 框架的原始版本,之后又对其进行了改进,于 1998 年提出了基于梯度学习的 CNN 模型——LeNet-5[5],并将其成功应用于手写数字字符的识别中,这为 CNN 以后的发展奠定了坚实的基础。但是,受限于当时的标记数据集规模和软硬件基础设施,LeNet-5 在大规模数据集分类任务中并未取得良好的效果。

自 2006 年以来,研究人员提出了很多方法来克服深度 CNN 在训练中所遇到的困难[6-12]。其中最值得注意的是 2012 年 Krizhevsky 等[6]提出的一种与 LeNet-5 类似但具有更深结构的 CNN 架构——AlexNet,该架构在 ILSVRC2012 图像分类任务中显著超越了之前的所有方法,一举夺得 2012 ILSVRC 的冠军。随着 AlexNet 的成功,研究者们也在其基础上进行了部分改进,以提升它的性能,其中比较有代表性的架构有 ZFNet[7], NIN[8], VGGNet[9], GoogLeNet[10], ResNet[11], DenseNet[12], DPN[13] 和 SENet[14]。随着这种架构的发展,这些网络也出现了愈宽、愈深、愈复杂的趋势,这虽然有助于网络获得更好的特征表示,但也使得这些网络更难优化和更容易出现过拟合。为了解决这些问题,研究人员已经提出了各种改进措施。

本文第 2 节系统性地介绍 CNN 的基本构件及原理;第 3 节对近年来 CNN 在卷积层、池化层、激活函数等各个方面取得的最新研究进展进行阐述和讨论;第 4 节总结了自 1998 年以来比较有代表性的 CNN 架构;第 5 节介绍了 CNN 在图像分类/定位、目标检测、目标分割、目标跟踪、行为识别和图像超分辨率重构等领域的应用情况;最后对 CNN 未来的研究方向做出展望。

2 CNN 的基本组件

CNN 以原始数据作为算法输入,通过卷积、池化和非线性激活函数映射等一系列操作,将原始数据逐层抽象为自身任务所需的最终特征表示,最后以特征到任务目标的映射作为结束。虽然 CNN 的变体有很多,但它们的结构都非常相似,一般由输入层、卷积层(convolutional layers)、池化层(pooling layers)、全连接层(fully-connected layers)和输出层组成,如图 1 所示。

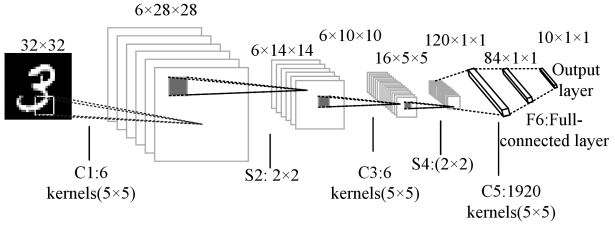


图 1 LeNet-5 模型结构[5]

Fig. 1 Architecture of LeNet-5[5]

卷积层作为输入层后的第一层,旨在学习输入的特征表示。卷积层由多个滤波器组成,用于计算不同的特征映射。具体而言,特征映射的每个神经元连接到前一层中的相邻神经元的区域。这样一个邻域在前一层被称为神经元的感受野。新的特征映射可以通过首先将输入与学习的滤波器进行卷积,然后在卷积结果上应用非线性激活函数得到。为了生成每个特征映射,滤波器由输入的所有空间位置共享。低卷积层中的滤波器被设计用于检测诸如边和曲线的低级特征,而高层中的滤波器被设计用于学习编码更多的抽象特征。通过堆叠多个卷积层,网络模型可以逐步提取更高层的特征表示。完整的特征映射是通过使用几个不同的滤波器获得的。在数学上,在第 l 层的第 k 个特征映射中,(i,j) 处的特征值 $z^l_{i,j,k}$ 通过式(1)来计算:

$$z^l_{i,j,k}=w^l_k x^l_{i,j}+b^l_k \tag{1}$$

其中,w^l_k 和 b^l_k 分别是第 l 层的第 k 个滤波器的权重向量和偏置项,$x^l_{i,j}$ 是以第 l 层的位置 (i,j) 为中心的输入块,生成特征映射 $z^l_{:,:,k}$ 的滤波器 w^l_k 是共享的。这样的权重共享机制具有可以降低模型复杂度、使网络更容易训练等优点。

激活函数将非线性引入到 CNN 中,从而实现多层网络对非线性特征的检测。典型的激活函数有 sigmoid,tanh 和 ReLU 等。设 $a(\cdot)$ 为非线性激活函数,卷积特征 $z^l_{i,j,k}$ 的激活值 $a^l_{i,j,k}$ 可以用式(2)进行计算:

$$a^l_{i,j,k}=a(z^l_{i,j,k}) \tag{2}$$

池化层(pooling layers)旨在通过降低特征映射的分辨率(降维和抽象)来实现移位不变性,通常位于两个卷积层之间。池化层的每个特征映射都连接到其前一个卷积层的相应特征映射。池化过程与卷积过程类似,可视为一种不带权重的池化函数,从输入特征图的左上角开始按一定步长从左向右、从上到下滑动,对窗口对应的特征映射区块进行池化操作后输出。设 $pool(\cdot)$ 为池化函数,对于每个特征映射 $a^l_{:,:,k}$,有:

$$y^l_{i,j,k}=pool(a^l_{m,n,k}),\forall(m,n)\in\mathcal{R}_{i,j} \tag{3}$$

其中,$\mathcal{R}_{i,j}$ 是特征映射中以 (i,j) 为中心的区块。

典型的池化操作有平均池化(average pooling)和最大池化(max pooling)。最大值池化函数是把区块中元素的最大值作为函数的输出,提取特征平面的局部最大响应,通常用于低层特征的提取,对输入的特征图选取最显著的特征。均值池化函数是将计算得到的区块中所有元素的算术平均值作为函数的输出,并提取特征平面局部响应的均值。

经过多次卷积和池化操作之后,卷积神经网络通常会在最后连接一个或者多个全连接层。全连接层将当前层的每一个神经元与上一层中的所有神经元连接,以产生全局语义信息。完全连接层可以被 1×1 卷积层所取代,因此并不总是必要的。第 l 层全连接层的第 j 个神经元的激活值(特征向量) a^l_j 的表示如下:

$$a^l_j=\sigma\left(\sum_k w^l_{j,k}\,a^{l-1}_k+b^l_j\right) \tag{4}$$

其中,$w^l_{j,k}$ 表示从第 $l-1$ 层的第 k 个神经元到第 l 层的第 j 个神经元的链接权重,b^l_j 表示第 l 层的第 j 个神经元的偏置,$\sigma(\cdot)$ 为非线性激活函数。

卷积神经网络的最后一层是输出层。对于分类问题而言,经常使用 softmax 逻辑回归(softmax regression)进行分类,返回输入图片属于某一类别 i 的概率;对于回归问题,返

回具体的数值。

卷积神经网络的训练目标是最小化网络的损失函数$L(\theta)$(loss function)。在预测任务中，给定样例集$(x^{(n)}, y^{(n)})$，$n\in[1,\cdots,N]$，输入$x^{(n)}$经过前向传导后利用损失函数计算出与期望值$y^{(n)}$之间的差异，称为“残差”。卷积神经网络的损失可以按式(5)进行计算：

$$L(\theta)=\frac{1}{N}\sum_{n=1}^{N} l(\theta; y^{(n)}, \hat{y}^{(n)}) \tag{5}$$

其中，θ表示卷积神经网络中所有的训练参数w和b，$y^{(n)}$是输入数据$\hat{x}^{(n)}$的真实标记，$\hat{y}^{(n)}$是输入数据$x^{(n)}$的预测结果(输出结果)，$l(\cdot)$表示损失函数。常见的损失函数有均方误差(Mean Squared Error，MSE)函数、交叉熵(cross entropy)函数等。

3　CNN 的改进研究

自 2012 年 Krizhevsky 等提出的 AlexNet 在 ImageNet 的图像分类竞赛中夺冠以来，研究人员的焦点逐渐转移到对 CNN 的改进上来，不断有新的改进措施被提出。本节将从 3 个方面描述 CNN 的主要改进措施：卷积层、池化层和激活函数。

3.1　卷积层

卷积层是卷积神经网络架构中不可或缺的部分，其主要用于学习输入图像的特征表示。因此，研究人员不断尝试改进 CNN 架构中的卷积层来提高网络的性能，下面介绍这方面的一些关键创新举措。

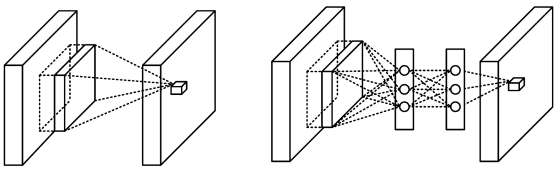
3.1.1　Network in Network

标准的卷积神经网络一般是由线性卷积层、池化层、全连接层连接起来的网络。卷积层通过线性滤波器进行线性卷积运算，然后用非线性激活函数对卷积后的结果进行处理，最终生成特征图。以 Relu 激活函数为例，特征图的计算公式为：

$$f_{i,j,k}=\max(w_k^{\mathrm{T}} x_{i,j}, 0) \tag{6}$$

其中，(i,j)表示特征图中像素点的位置索引，$x_{i,j}$表示卷积窗口中的图片块，k则表示要提取的特征图的索引。

因为卷积层使用线性滤波器，所以它更适合学习线性可分的潜在特征，然而其所要提取的特征一般是高度非线性的。因此，Lin 等[8]提出了一种 NIN(Network in Network)模型，其主要思想是将传统卷积层替换为由具有非线性激活函数的多个完全连接层组成的多层感知层(mlpconv 层)，如图 2 所示，从而用非线性神经网络代替线性滤波器，这使得它能够逼近潜在特征的更多抽象表示，泛化能力更强。



(a)线性卷积层　　(b)mlpcov 层

图 2　线性卷积层和 mlpcov 层[8]

Fig. 2　Linear convolution layer and mlpconv layer[8]

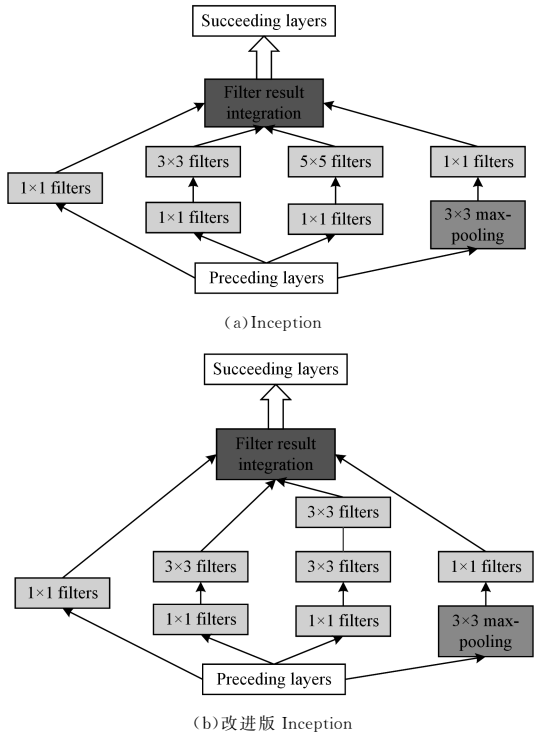
mlpconv 层可以看成是传统线性卷积与非线性多层感知器(MLP)的结合体。对于 mlpconv 层，特征映射的计算公式如下：

$$\begin{aligned} f_{i,j,k_1}^{1} &= \max(w_{k_1}^{1\mathrm{T}} x_{i,j}+b_{k_1}, 0) \\ &\vdots \\ f_{i,j,k_n}^{n} &= \max(w_{k_n}^{n\ \mathrm{T}} f_{i,j}^{n-1}+b_{k_n}, 0) \end{aligned} \tag{7}$$

其中，n是多层感知器的层数，激活函数选用 Relu。

3.1.2　Inception 和改进版 Inception

受 Lin 等[8]提出的 NIN 模型启发，Szegedy 等[10]于 2014 年提出了一种 Inception 模型，该模型采用降维(1×1 卷积滤波器)技术来降低卷积运算昂贵的计算成本。其主要思想是利用 3 个不同尺寸的滤波器对前一个输入层提取不同尺度的特征信息，然后融合这些特征信息并传递给下一层。Inception 拥有 1×1，3×3 和 5×5 的滤波器，其中 1×1 的滤波器较前一层有更低的维度，主要用于数据降维，在传递给后面的 3×3 和 5×5 卷积层时降低了卷积计算量，避免了网络规模扩大所带来的巨大计算量。通过对 4 个通道的特征融合，下一层可以从不同尺度上提取到更有用的特征。为了进一步提高 CNN 的分类精度，Szegedy 等通过在网络中使用分解卷积和积极降维来对初始 Inception 模型进行改进。改进后的 Inception 模型通过用两个 3×3 卷积来代替 Inception 模型中的 5×5 卷积，既减少了参数数量，也加速了计算。图 3 显示了 Inception 和改进版 Inception 之间的区别。



(a)Inception

(b)改进版 Inception

图 3　Inception 和改进版 Inception

Fig. 3　Inception module and improved Inception module

3.2　双卷积

CNN 具有的局部连接和权值共享特性使其可以有效降低神经网络的复杂性，Zhai 等[15]于 2016 年提出了一种双卷积神经网络(DCNN)，如图 4(b)所示，它利用一种新的双卷积

运算为 CNN 上的参数共享提供额外的支撑。DCNN 用滤波器组来代替独立学习的一组卷积滤波器，其中每个组内的滤波器是彼此的转换版本。DCNN 可作为 Goodfellow 等提出的 maxout 网络的正则化版本来提高分类精度；此外，它可以通过灵活地改变其体系结构来提高参数效率，从而减少存储器占用而不会降低准确性。这种方法的不足之处在于，与标准的卷积层相比，双卷积操作将导致额外的计算成本。

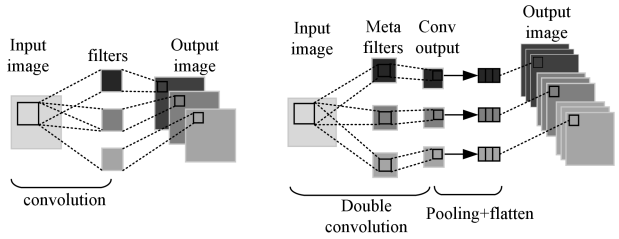


图 4 卷积层和双卷积层的体系结构[15]

Fig. 4 Architectures of convolutional layer and doubly convolutional layer [15]

3.3 池化层

除了卷积层，CNN 也包含池化层。池化层通常紧接在卷积层之后使用，它能够简化从卷积层输出的信息，降低特征映射的维度。除此之外，与卷积层类似，池化层也具有平移不变性。下面介绍一些比较成功的池化技术。

3.3.1 l_p 池化

l_p 池化是一个以复杂细胞为模型的生物启发式池化过程[16]。理论分析表明，l_p 池化具有比最大池化(max pooling)更好的泛化能力[17]。l_p 池化可以表示为：

$$y_{i,j,k}=\Big[\sum_{(m,n)\in R_{i,j}}(a_{m,n,k})^p\Big]^{1/p} \tag{8}$$

其中，$y_{i,j,k}$ 是第 k 个特征图中 (i,j) 处的池化输出，$a_{m,n,k}$ 是第 k 个特征图中池化区域 $R_{i,j}$ 内 (m,n) 处的特征值。特别是当 $p=1$ 时，l_p 对应于平均池化；当 $p=\infty$ 时，l_p 对应于最大池化。

3.3.2 混合池化

受到随机 Dropout[18] 和 DropConnect[19] 的启发，Yu 等[20]提出了一种混合池化方法，其可以被视为是最大池化和平均池化的组合。混合池化可以表示为：

$$y_{i,j,k}=\lambda\max_{(m,n)\in R_{i,j}}(a_{m,n,k})+\frac{(1-\lambda)1}{(|R_{i,j}|)(a_{m,n,k})} \tag{9}$$

其中，λ 是随机值(0 或 1)，表示使用平均或最大池化。在正向传播过程中，λ 被记录下来并用于反向传播操作。实验表明，混合池化可以更好地解决过度拟合问题，其性能优于最大池化和平均池化。

3.3.3 随机池化

随机池化[21]是一种受 Dropout 启发的池化方法。与最大池化选择每个池化区域内的最大值不同，随机池化根据多项式分布随机地选择激活值，从而确保特征图中的非最大激活值也可以被利用。具体地说，随机池化首先通过归一化区域内的激活值来计算每个区域 R_i 的概率 p；获取分布 $P(p_1,\cdots,p_{|R_j|})$ 后，可以从基于 P 的多项式分布中选择一个区域内的位置 l；然后设置池化后的激活值为 $y_j=a_l$，其中 $l\sim P(p_1,\cdots,p_{|R_j|})$。与最大池化相比，随机池化中的随机因素可以避免过度拟合。

3.3.4 空间金字塔池化

空间金字塔池化(SPP)是由 He 等于 2014 年提出的[22]。SPP 的关键优势在于，无论输入的特征图大小如何，它都可以生成固定大小的特征向量，然后输入全连接层。SPP 将对输入特征图中大小与图像大小成比例的局部区域进行池化操作，从而得到固定大小的特征向量。这与以前深度网络中的滑动窗口池化不同，滑动窗口的数量取决于输入图片的大小。通过用 SPP 替换最后一个池化层，何凯明等提出了一个新的 SPP-Net，它能够处理不同大小的图像。

3.3.5 Spectral 池化

Spectral 池化[23]通过在频域中裁剪输入图像来执行降维。给定一个输入特征映射 $x\in R^{m\times m}$，假设所需输出特征图的维数为 $h\times w$，则 Spectral 池化首先计算输入特征图的离散傅立叶变换(DFT)，然后通过仅维持频率中心的 $h\times w$ 子矩阵来裁剪频率图，最后使用逆 DFT 将截断的表示映射回空间域。与最大池化相比，Spectral 池化的线性低通滤波操作可以为相同的输出维数保留更多的信息，且免受其他池化方法所呈现的输出图维数急剧下降的影响；更重要的是，频谱汇聚过程是通过矩阵截断来实现的，这使得它能够以较低的计算成本应用在采用 FFT 作为滤波器的 CNNs 中。

3.4 激活函数

“激活函数”，又称“非线性映射函数”，是深度卷积神经网络中不可或缺的关键模块，其模拟了生物神经元特性，接受一组输入信号并产生输出，并通过一个阈值来控制神经元的兴奋与抑制状态。图 5 为当下深度卷积神经网络中常用的 5 种激活函数。

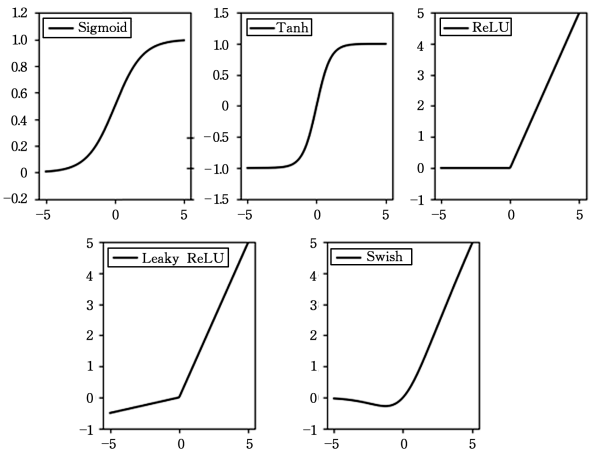


图 5 激活函数

Fig. 5 Activation functions

3.4.1 Sigmoid

Sigmoid 函数的数学公式为：

$$\sigma(x)=\frac{1}{1+e^{x}} \tag{10}$$

其中，$\sigma(x)$ 的取值范围为 0～1。

3.4.2　Tanh

Tanh 激活函数虽然解决了 Sigmoid 函数值域中不能取到 0 的问题，但是仍然存在梯度消失的问题。Tanh 函数的数学公式为：

$$\sigma(x)=\frac{e^{x}-e^{-x}}{e^{x}+e^{-x}} \tag{11}$$

其中，$\sigma(x)$的取值范围为$-1\sim1$。

3.4.3　修正线性单元(ReLU)

修正线性单元(ReLU)[24]是目前最成功和广泛使用的激活函数。ReLU 是一个分段线性函数：当输入为负时，ReLU 函数的输出为 0；当输入为正时，ReLU 函数的输出为 x。相比于 Sigmoid 和 Tanh，ReLU 拥有更快的收敛速度，同时还在隐藏层引入稀疏性而使网络容易获得稀疏表示。尽管 ReLU 在 0 处的不连续性可能影响反向传播的表现，但实验表明，ReLU 拥有比 Sigmoid 和 Tanh 激活函数更出色的性能，能够较好地解决梯度消失问题。ReLU 函数的数学公式为：

$$f(x)=\max(0,x) \tag{12}$$

3.4.4　Leaky ReLU 和 Parametric ReLU

尽管 ReLU 能较快地收敛，并且不会受到梯度消失问题的困扰，但其存在一个潜在的缺点：当神经元不活动时，它的梯度为零，这可能会导致最初没有激活的单元从不激活，因为基于梯度的优化不会调整它们的权重；而且，恒定的零梯度可能会减慢训练过程。为了解决这一问题，Maas 等[25]提出了 Leaky ReLU，该函数压缩负值部分，而不是将其映射到常量零点，这样当单元处于非活动状态时，它允许一个很小的非零梯度。Leaky ReLU 函数的数学公式为：

$$f(x)=\max(0.1x,x) \tag{13}$$

与 Leaky ReLU 中使用预定义的参数不同，He 等[26]提出了带可自适应学习参数的 Parametric ReLU。Parametric ReLU 函数的数学公式为：

$$f(x)=\max(\alpha x,x) \tag{14}$$

其中，α 是超参数。

3.4.5　Swish

2017 年，受 LSTM 和 highway network 中使用 Sigmoid 函数进行门控的启发，来自谷歌的研究人员提出了一种新型激活函数——Swish[27]。Swish 可以看作是位于线性函数和 ReLU 函数之间的非线性插值的平滑函数。与 ReLU 一样，Swish 有下界，但无上界；与 ReLU 不同的是，Swish 非常平滑，且非单调。Swish 函数的数学公式为：

$$\sigma(x)=\frac{x}{1+e^{-x}} \tag{15}$$

4　卷积架构

深度卷积神经网络的成功得益于层出不穷的 CNN 架构。本节将介绍自 1998 年以来比较有代表性的 CNN 架构及其发展状况。

4.1　LeNet-5

LeNet-5 是 LeCun 于 1998 年设计的用于识别手写数字的 CNN 架构，是早期 CNN 中最具代表性的架构。LeNet-5 由两个卷积层、两个池化层和两个全连接层组成，每个卷积层使用尺寸为 5×5(每个滤波器有 1 个通道)的滤波器，第一层中有 6 个滤波器，第二层中有 16 个滤波器。在每次卷积之后，采用 Sigmoid 函数进行激活，并且使用 2×2 的平均池化进行池化操作。LeNet-5 结构如图 1 所示。

4.2　AlexNet

2012 年，Krizhevsky 等[6]提出了一个大型的深度卷积神经网络——AlexNet，该网络模型在 2012 ILSVRC 中首次实现了 15.4% 的 Top 5 误差率，赢得了 2012 年 ILSVRC 的冠军。AlexNet 由 5 个卷积层和 3 个全连接层组成，其中集成了各种技术来提高性能并减少训练时间，包括使用非饱和非线性神经元 ReLU 代替饱和非线性神经元 Tanh 来加快训练速度，使用 LRN(local response normalization)来提高泛化能力，使用图像转换(image translation)、水平反射(horizontal reflection)和改变训练图像中 RGB 通道的强度来增强数据，使用 dropout 来减弱全连接层的过拟合问题，使用随机梯度下降(SGD)训练模型。AlexNet 模型结构如图 6 所示。

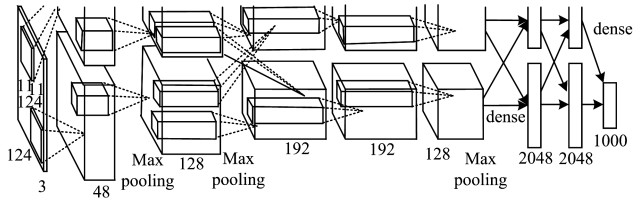


图 6　AlexNet 模型的结构[6]

Fig. 6　Architecture of AlexNet[6]

2013 年，纽约大学的 Zeiler 等[7]提出了一种基于新颖可视化技术的网络模型——ZF-Net，其可以深入了解中间特征图层的内部运作，并以 11.2% 的 Top5 错误率赢得了 2013 年 ILSVRC 的冠军。ZF-Net 模型可以看作是 AlexNet 架构的微调优化版，将 AlexNet 第一个卷积层的滤波器尺寸从 11×11 调整为 7×7，将步长从 4 调整为 2，将交叉熵作为损失函数，使用批处理随机梯度下降进行训练。ZF-Net 模型结构如图 7 所示。

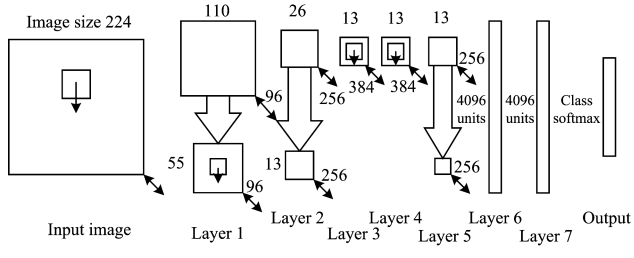


图 7　ZF-Net 模型的结构[7]

Fig. 7　Architecture of ZF-Net[7]

4.3　VGGNet

2014 年，牛津大学 VGG(visual geometry group)的 Simonyan 等[9]建立的 19 层深度网络模型(VGGNet)在 2014 ILSVRC 中获得了定位第一，分类第二的成绩，错误率为 7.3%。相较于 AlexNet 和 ZF-Net，VGGNet 主要探讨了深度对于提高网络性能的重要性。该模型严格使用 3×3 的滤波器

($stride=1$,$pad=1$)和 2×2 最大池化层($stride=2$)。两个 3×3 的卷积层组合可以实现 5×5 的有效感受野,这就在保持滤波器尺寸较小的同时模拟了大型滤波器,减少了参数,增强了映射函数的非线性,使得网络更具判别性。

VGGNet 模型结构如图 8 所示。

ConvNet Configuration					
A	A-LRN	B	C	D	E
11 weight layers	11weight layers	13 weight layers	13 weight layers	16 weight layers	19 weight layers
input (224×224 RGB image)					
conv3-64	conv3-64 LRN	conv3-64 conv3-64	conv3-64 conv3-64	conv3-64 conv3-64	conv3-64 conv3-64
maxpool					
conv3-128	conv3-128	conv3-128 conv3-128	conv3-128 conv3-128	conv3-128 conv3-128	conv3-128 conv3-128
maxpool					
conv3-256 conv3-256	conv3-256 conv3-256	conv3-256 conv3-256	conv3-256 conv3-256 conv1-256	conv3-256 conv3-256 conv3-256	conv3-256 conv3-256 conv3-256 conv3-256
maxpool					
conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512 conv1-512	conv3-512 conv3-512 conv3-512	conv3-512 conv3-512 conv3-512 conv3-512
maxpool					
conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512 conv1-512	conv3-512 conv3-512 conv3-512	conv3-512 conv3-512 conv3-512 conv3-512
maxpool					
FC-4096					
FC-4096					
FC-1000					
soft-max					

图 8 VGGNet 模型结构[9]

Fig. 8 Architecture of VGGNet[9]

4.4 GoogLeNet

来自 Google 的 Szegedy 等[10]开发设计的 GoogLeNet,以 6.67%的 Top5 错误率获得 2014 年 ILSVRC 挑战赛图像分类任务的冠军。GoogLeNet 是一个 22 层的卷积神经网络,由 Inception 结构作为基本模块级联而成。每个 Inception 模块使用不同尺寸的滤波器(即 1×1,3×3,5×5)以及 3×3 的最大池化并行连接而成,并将它们的输出级联在一起用于模块输出。为了减少权重参数,将 1×1 的滤波器作为一个“瓶颈”来减少每个滤波器的通道数量;同时,GoogLeNet 通过删除全连接层提升了网络深度,同时大大减少了网络参数,并且充分利用了计算资源,提高了算法的计算效率和性能。GoogLeNet 模型结构如图 9 所示。

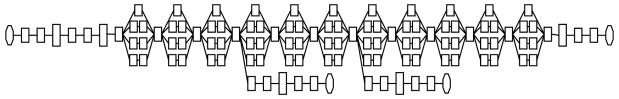


图 9 GoogLeNet 模型结构[10]

Fig. 9 Architecture of GoogLeNet[10]

4.5 ResNet

微软亚洲研究院的何恺明等[11]提出的 ResNet 在 2015 年 ILSVRC 挑战赛的检测、定位和分类任务以及 COCO 的检测和分割挑战赛上都获得了冠军,其架构深度达到惊人的 152 层。深度网络面临的挑战之一是训练过程中会出现梯度消失问题。ResNet 通过引入包含 identity connection 的“shortcut”模块,使得卷积层可以被跳过,缓解了深度神经网络中的梯度消失问题。“shortcut”模块学习残差映射 $F(x)=H(x)-x$,而不是学习卷积层 $F(x)$的函数。ResNet 还使用 1×1 滤波器的“瓶颈”方法将“shortcut”模块中的两层替换为三层(1×1,3×3,1×1)来减少权重参数,其中 1×1 滤波器可以先减少再增加权重参数的数量。ResNet 模型结构如图 10 所示。

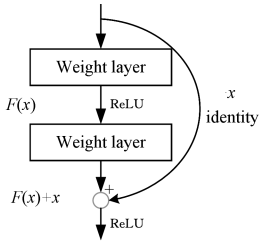


图 10 ResNet 模型结构[11]

Fig. 10 Architecture ofResNet[11]

4.6 DenseNet

2016 年,受 ResNet 中 identity/skip connections 思想的的启发,来自康奈尔大学、清华大学、Facebook FAIR 实验室的 Huang 等[12]提出了一种 DenseNet 模型。该模型先在 ConvNet 中用前馈的方式将每一层与网络中的其他层连接起来,然后将所有先前层的特征图用作每个后续层的输入,从而构建出 DenseNet。在流行的图像分类基准上,包括具有挑战性的 ILSVRC 挑战赛,DenseNet 能获得与 ResNet 相当的准确性,但所需参数明显更少。此外,它改善了梯度消失问题,同时通过加强特征传播和促进特征重用减少了计算量。DenseNet 模型结构如图 11 所示。

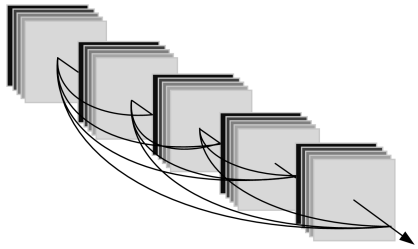


图 11 DenseNet 模型结构[12]

Fig. 11 Architecture of DenseNet[12]

4.7 DPN

奇虎 360 和新加坡国立大学组成的 NUS-Qihoo_DPNs 团队[13]于 2017 年提出了一种新型 Dual Path Network(DPN)模型,其以 6.2%的识别错误率荣获 2017 年 ILSVRC 挑战赛物体定位任务的冠军。ResNet 可以重用已提取的特征,而 DenseNet 则可以试图提取新的特征。DPN 结合了两者的优点,在共享通用特征的同时保持了通过双路径架构探索新特征的灵活性。DPN 模型结构如图 12 所示。

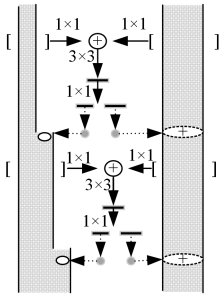


图 12 DPN 模型结构[13]

Fig. 12 Architecture of DPN[13]

4.8 SENet

2017 年，来自自动驾驶公司 Momenta 研发团队（WMW）的 Hu 等[14]提出了 Squeeze-and-Excitation Networks（SENet）架构，其以 2.3%的识别错误率荣获 2017 年 ILSVRC 挑战赛物体识别任务的冠军。它允许网络通过学习使用全局信息来对特征进行重新校准（选择性地强调信息特征），激励对分类有用的特征，抑制不太有用的特征。在只引入极少的计算量和参数量的情况下，SENet 可以大幅提升现有的绝大多数 CNN 的性能。SENet 模型结构如图 13 所示。

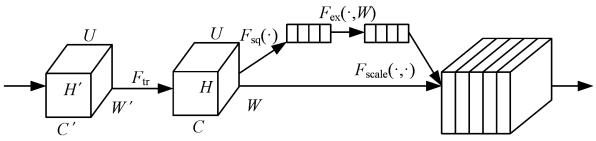


图 13　SENet 模型结构[14]

Fig. 13　Architecture of SENet[14]

5　CNN 在计算机视觉领域的应用

本节重点介绍了 CNN 在计算机视觉领域的最新研究进展，包括图像分类/定位、目标检测、目标分割、目标跟踪、行为识别和图像超分辨率重构。

5.1　图像分类/定位

图像分类作为计算机视觉领域中的基本任务，通常是指将图像分类到几个预定义的类中，它构成了其他计算机视觉任务的基础。定位是指找到识别目标在图像中出现的位置，通常这种位置信息由对象周围的一些边界框表示出来。前一节介绍的几种经典 CNN 模型均适用于求解图像分类/定位任务，这里不再赘述。计算机视觉任务（包括图像分类/定位、目标检测、实例分割）如图 14 所示。

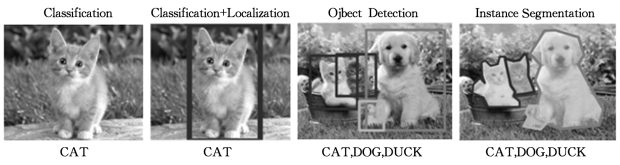


图 14　计算机视觉任务

Fig. 14　Computer vision tasks

5.2　目标检测

目标检测（object detection）一直是计算机视觉领域的重要研究方向[28-30]，其试图精确定位图像中目标对象出现的区域并判定目标类别。CNN 用于目标检测最早可追溯到 20 世纪 90 年代[31]。然而，由于缺乏训练数据并且计算能力有限，在 2012 年之前基于 CNN 的目标检测的进展缓慢。2012 年 CNN 在 ImageNet 挑战中的巨大成功[8]重新激发了研究人员对基于 CNN 的目标检测的兴趣[32]，也带动了目标检测精度的提升。其中较有影响力的工作包括 R-CNN[33]，Overfeat[34]，Fast R-CNN[35]，Faster R-CNN[36]，FPN[37] 和 Mask R-CNN[38]。图像检测已被广泛应用于遥感、医学诊断、灾害评估和视频监控领域。

2014 年，Girshick 等[33]提出了一种基于候选区域（region proposal）和 CNN 的深度学习目标检测框架——R-CNN，如图 15 所示。R-CNN 算法首先采用选择性搜索（selective search）[39]策略在输入图像上提取若干包含对象的候选框，将候选框扭曲到一个固定的大小，并使用预先训练的 DCNN 来从中提取特征；然后基于特征，利用 SVM 等线性分类器将候选框分为目标和背景两部分；最后使用非极大值抑制方法舍弃部分候选框，得到目标物体的定位结果。

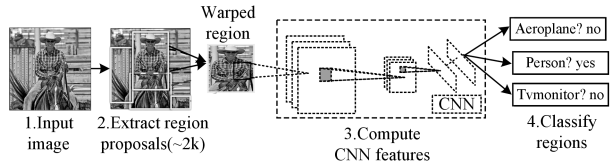


图 15　R-CNN 目标检测算法的流程图[33]

Fig. 15　Flowchart of R-CNN target detection algorithm[33]

虽然 R-CNN 算法在目标检测任务中取得了显著的性能提升，然而由于 CNN 特征提取器分别针对每个候选区域执行，需要耗费大量的时间，因此其计算成本仍然很高。为了解决这个问题，最近一些研究人员提出了共享卷积层特征。OverFeat[34]充分利用了 CNN 的特征提取功能，首先通过 CNN 从金字塔型图像中提取基础特征，然后再将基础特征共享于分类、定位和检测这些不同的任务中。Fast R-CNN[35]通过使用端到端（end to end）的训练方法来改进 SPP 网络，所有的网络层都可以在微调期间更新，这简化了学习过程，提高了检测的准确性。随后，Faster R-CNN[36]引入了用于生成候选目标的候选区域网络（RPN），进一步加快了速度。2017 年，Lin 等[37]利用 CNN 的金字塔层次结构特性，把低分辨率、高语义信息的高层特征和高分辨率、低语义信息的低层特征进行自上而下的侧边连接，构建出不同尺度下具有高级语义信息的特征金字塔网络（FPN）。与其他目标检测模型相比，FPN 在图像小目标检测方面的性能显著提升。同年，He 等[38]在 Faster R-CNN 的基础上提出了一种通用的目标实例分割框架——Mask R-CNN，并获得了 ICCV2017 最佳论文奖。Mask R-CNN 通过添加一个用于预测对象掩码的分支来扩展 Faster R-CNN，该分支与现有的用于边界框识别的分支并行。相对于 Faster R-CNN，Mask R-CNN 的开销虽然略有增加，但其易于训练，且能在预测每个物体边界框的基础上提供像素级语义分割。

除了基于 R-CNN 的方法，还有可以在一次评估中从待检测的整幅图像中预测边界框（bounding boxes）和类别概率，并可以直接在检测性能上进行端到端优化的目标检测技术，如 YOLO[40]、SSD[41]及其改进版本 YOLO9000[42]、DSSD[43]等。

5.3　目标分割

图像语义分割（semantic segmentation）是为每个图像像素预测一个类别标签，而实例分割（instance segmentation）则是在语义分割的基础上进一步区分出同一类别事物中不同下属级别的对象，例如不同种类的鸟。目前，分割技术已被应用于自动驾驶和医疗影像处理中。

随着深度卷积神经网络在图像检测、分类等多个任务上被成功应用，目前已经有不少研究人员将 CNN 应用到图像语义分割领域。自 2015 年起，Facebook 的人工智能实验室

(FAIR)开展了名为 DeepMask[44] 的研究项目,DeepMask 可以在对象上粗略生成一个初级版本的分割区域。之后,FAIR 对其进行了改进研究,于 2016 年开发出了可以对 DeepMask 提供的分割区域进行修正的 SharpMask[45],它可以纠正漏掉的细节并改善语义分割效果。在此基础之上,MultiPath-Net[46] 可以识别出由每个分割区域描述的物体。

Long 等[47] 在 CVPR2015 上提出的全卷积网络(Fully Convolutional Network,FCN)能够通过端到端(end to end)的学习得到每个像素的目标分类结果。与经典的 CNN 架构不同,FCN 用卷积层代替 CNN 架构中所有的全连接层,这使得其可以接受任意尺寸的输入图像并生成目标图像大小的输出。然后,其通过跳跃体系结构将来自深层粗略层的语义信息与浅层精细层的外观信息相结合,以生成准确和详细的图像语义分割。FCN 的结构示意图如图 16 所示。

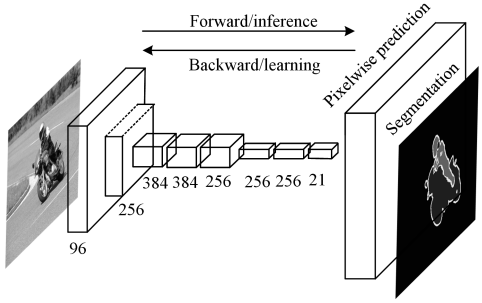


图 16 FCN 的结构示意图[47]

Fig. 16 Architecture of FCN[47]

FCN 虽然在图像语义分割方面取得了不错的效果,但缺少对图像空间、边缘信息的约束,导致最后的图像分割结果比较粗糙。针对这一问题,Chen 等[48] 提出了 DeepLab,其通过上采样滤波器进行密集特征的提取,将针对图像分类训练的网络重新定义为语义分割任务,并进一步将其扩展到空间金字塔池化,它可以对多个尺度的对象和图像上下文进行编码。为了产生语义准确的预测和沿着对象边界的详细分割图,其还将来自深度卷积神经网络和完全连接的条件随机场(CRF)的想法结合了起来。

FCN 已经被证明在语义分割方面非常成功,但是它不能识别出对象实例。针对这一问题,2016 年 Dai 等[49] 开发了一种能够实现实例分割的 FCN(instance-sensitive FCNs)。与之前生成一个分数图的 FCN 相反,该方法被设计用来计算一组实例敏感的分数图,其中每一个都是相对于实例的相对位置的像素级分类器的结果。在这些实例敏感的分数图之上,一个简单的组装模块能够在每个位置输出候选实例。除此之外,Dai 等[50] 还提出了多任务级联网络(multitask network cascades)来实现实例感知语义分割。该模型由 3 个网络组成:实例区分、掩码估计(estimating masks)和对象分类。这些网络形成一个级联结构,并共享彼此的卷积特征。

5.4 目标追踪

目标追踪(object tracking)是在给定的场景中追踪感兴趣的一个或多个特定目标的过程,其成功与否在很大程度上依赖于算法对诸如视点变化、光照变化和遮挡等多种因素导致的目标外观变化的鲁棒性的大小[51-53]。目标跟踪技术由于在自动驾驶汽车、机器人、智能视频监控等领域具有重大的理论价值和广阔的应用前景,因此引起了越来越多的研究者的关注。目前,CNN 已被广泛应用于视觉跟踪任务。

Fan 等[54] 使用 CNN 学习一个单独的特定类网络来追踪对象。文献[54]设计了一个带有移位变体结构的 CNN 跟踪器,该架构使 CNN 模型从探测器变成跟踪器。与仅仅提取局部空间结构的传统跟踪器不同,这种基于 CNN 的跟踪方法通过考虑两个连续帧的图像来提取空间和时间结构。由于时间信息中的大信号倾向于在正在移动的物体附近发生,因此时间结构提供粗略的速度信号来进行跟踪。

针对目标追踪中定义手工特征表示需要专业知识、手动调整耗时的问题,Li 等[55] 提出了一种新颖的目标追踪算法。该算法在跟踪过程中自动重新学习最有用的特征表示,以准确地适应外观变化、姿态和尺度变化,同时防止漂移和跟踪失败。该模型采用多个 CNN 的候选池化作为目标对象不同实例的数据驱动模型,每个 CNN 都维护一组特定的内核,并使用所有可用的低级提示将对象补丁与其周围的背景区分开来。

目标外观模型作为目标跟踪的一个关键模块,通常使用的特征是以手工离线的方式预定义的,但不针对跟踪的对象进行调整。2016 年,Chen 等[56] 提出了一种鲁棒的目标跟踪方法,其通过 CNN 动态学习最具判别力的特征。该方法首先将 CNN 从大规模训练数据源任务中学习到的特征转移到训练数据有限的新的跟踪任务中;其次利用初始帧中标记的对象的真实情况信息和在线获取的图像观测值来缓解模型更新引起的跟踪器漂移问题;最后使用启发式模式来判断是否更新对象的外观模型。

Hong 等[57] 提出了一种基于预训练 CNN 的在线视觉跟踪算法。给定一个利用离线大规模图像库预先训练的 CNN 算法,该算法将网络隐藏层输出的特征用于在线 SVM 学习来辨别目标外观模型。另外,在 SVM 的引导下,通过对 CNN 特征进行反投影,构造目标特定的显著图,根据显著图生成的外观模型,得到各帧的最终跟踪结果。显著图有效地揭示了目标的空间位置,提高了目标的定位精度,使得该算法能够实现像素级的目标分割。

5.5 行为识别

基于视觉的人体行为识别是当前计算机视觉领域的热门研究方向之一,其主要目标是通过捕捉视频中的空间身体特征及其时间演变来识别并理解视频中人的动作和行为,在智能视频监控、人机交互等方面具有良好的应用前景。

CNN 作为一种可以直接作用于原始输入的深层模型,其学习到的图像特征具有一定的语义信息,近几年被广泛应用于人体动作识别任务中。然而,这样的模型最初仅限于处理 2D 输入。针对这一问题,Ji 等[58] 提出了一种新的用于人体动作识别的 3D CNN 模型。该模型通过执行 3D 卷积,从空间和时间维度提取特征,从而捕获在多个相邻帧中编码的运动信息。他们所开发的模型首先将输入视频的每一帧分解为多个信道,然后组合来自所有信道的信息以生成最终的特征表示。该模型已被成功用于识别机场监控视频中的人类行

为，并且实现了较基准方法更优的性能。

Karpathy 等[59]利用局部时空信息构建一个多分辨率的卷积神经网络来提取视频中人体的动作信息。在提出的网络模型中，输入帧被分成两种不同分辨率的处理流：模拟低分辨率图像的上下文流和处理高分辨率中心裁剪的中央流。两路视频经过相同的卷积、标准化和池化等一系列操作，最终汇合到两个全连接层。

视频可以分解成空间和时间两部分：空间部分以单个框架外观的形式携带有关视频中描绘的场景和对象的信息；时间部分以帧的运行形式传达观察者（相机）和物体的运动信息。受此启发，Simonyan 等[60]提出了一种结合空域和时域的双流卷积神经网络的架构。其中，空间流卷积网络可以从静止图像（单个视频帧）中识别动作来提取空间信息；而时间流卷积网络则可以从视频中识别光流来提取时间信息。空间流和时间流都以深度卷积神经网络的形式实现，然后通过 SVM 分类器进行后期融合来识别具体的动作。

Chéron 等[61]提出了一种基于静态图像姿态的卷积神经网络（Pose-based CNN，PCNN）来识别人体的行为。该方法在视频序列中跟踪人体及人体的每个部位，并使用 CNN 来提取视频里面每一帧中每个人体部位的外观特征和光流特征。PCNN 对所有静态图像的外观特征和光流特征进行聚合，从而生成整个动作的外观和光流特征描述。

5.6　图像超分辨率重构

图像超分辨率重构（Super Resolution，SR）是指从低分辨率图像重构出相应的高分辨率图像的过程，在需要更多图像细节的视频监控、卫星遥感图像和医学影像等领域有着重要的应用价值。图像分辨率越高，则提供的细节越精细，画质越细腻。

2014 年，Dong 等[62]首次用 CNN 来解决 SR 问题，并提出了一种基于 CNN 的图像超分辨率重构方法——SRCNN，该方法可以学习从低分辨率图像到高分辨率图像的端到端的映射。如图 17 所示，SRCNN 以经双三次插值后得到的目标尺寸大小的低分辨率图像作为输入，通过三层卷积网络来提取非线性特征，最后输出高分辨率图像。之后，Dong 等[63]对 SRCNN 进行了改进：1）在网络末端引入一个反卷积层，然后将原始低分辨率图像（无插值）直接映射到高分辨率图像；2）降低输入特征维度，然后进行映射和扩展；3）采用更小的滤波器，但映射层更多。所提出的模型实现了超过 40 倍的加速，甚至具有更优异的重构效果。

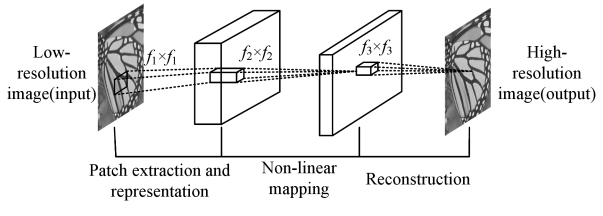


图 17　SRCNN 模型结构[62]

Fig. 17　Architecture of SRCNN[62]

2016 年，Shi 等[64]提出了第一个能够在单个 K2 GPU 上实现 1080p 视频实时图像超分辨率重构的卷积神经网络——ESPCN。该方法通过卷积层直接在原始低分辨率图像空间中提取特征，然后采用子像素卷积层学习提取到的特征，从而得到高分辨率图像。在 ESPCN 中，图像放大过程中的双三次插值操作被替换为针对每个特征映射而专门训练的滤波器，同时还降低了整个 SR 操作的计算复杂度。

受 VGG-Net 和 ResNet 启发，Kim 等[65]于 2016 年提出了一种基于 DCNN 的高精度的单图像超分辨率方法——VDSR。该网络模型将插值后的低分辨率图像作为输入并预测图像细节，通过在深度网络结构中多次级联小型滤波器来高效率地利用大型图像区域的上下文信息，采用残差学习和可调节的梯度剪切来加快收敛速度；同时，VDSR 可以将不同比例的子图像放在同一批次中进行训练。实验结果表明，相比于 SRCNN，VDSP 在精度和速度方面都具有明显优势。

2017 年，Lai 等[66]提出了拉普拉斯金字塔超分辨率网络（LapSRN）来逐步重建高分辨率图像的子带残差。在每个金字塔层面，模型将粗分辨率的特征图作为输入来预测高频残差，并使用转置卷积上采样到更精细的水平。LapSRN 不需要双三次插值作为预处理步骤，因此大大降低了计算复杂度；其次，其使用强大的 Charbonnier 损失函数对深层监督的 LapSRN 进行训练，实现了高质量的图像重构。

结束语　近年来，深度卷积神经网络在计算机视觉领域不断取得突破。本文从卷积层、池化层和激活函数等几个方面讨论了 CNN 的改进，总结了自 1998 年以来比较有代表性的神经网络架构，最后介绍了 CNN 在图像分类/定位、目标检测、目标分割、目标跟踪、行为识别和图像超分辨率重构等方面应用的最新研究进展。

尽管 CNN 相较于传统机器学习方法在计算机视觉等领域取得了巨大的进步，但仍有许多问题需要进一步的研究。首先，由于 CNN 架构越来越深，因此需要大规模的标注数据集和庞大的计算能力进行训练；其次，目前 CNN 在计算机视觉方面的研究几乎都是有监督学习，那么手动收集带标签的数据集就需要大量的人力，因此需要探索 CNN 的无监督学习；与此同时，在测试时，CNN 深度模型需要占用大量的内存，并且极为耗时，这使得它们不适合部署在资源有限的移动平台上，研究如何降低复杂性并获得快速执行的模型而不损失准确性是非常重要的；最后，选取合适的超参数一直是将 CNN 应用于新任务的一个主要障碍，比如学习率，滤波器的尺寸、步长和个数这些超参数都具有极强的内部依赖性，任何细小的调整都可能对最后的训练结果造成较大的影响。

参 考 文 献

[1] 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.

[2] FUKUSHIMA K. Neocognitron：A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position[J]. Biological Cybernetics，1980，36(4)：193-202.

[3] FUKUSHIMA K，MIYAKE S，ITO T. Neocognitron：A neural

network model for a mechanism of visual pattern recognition [J]. IEEE Transactions on Systems, Man, and Cybernetics, 1982, SMC-13(5): 826-834.

[4] LECUN Y, BOSER B E, DENKER J S, et al. Handwritten digit recognition with a back-propagation network[C]// Advances in neural information processing systems. 1990: 396-404.

[5] LECUN Y, BOTTOU L, BENGIO Y, et al. Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE, 1998, 86(11): 2278-2324.

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

[7] ZEILER M D, FERGUS R. Visualizing and understandingconvolutional networks[C]// European Conference on Computer Vision. Springer, Cham, 2014: 818-833.

[8] LIN M, CHEN Q, YAN S. Network in network[J]. arXiv: 1312.4400, 2013.

[9] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[J]. arXiv: 1409.1556, 2014.

[10] SZEGEDY C, LIU W, JIA Y, et al. Going deeper with convolutions[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2015: 1-9.

[11] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2016: 770-778.

[12] HUANG G, LIU Z, WEINBERGER K Q, et al. Densely connected convolutional networks[J]. arXiv: 1608.06993, 2016.

[13] CHEN Y, LI J, XIAO H, et al. Dual path networks[C]// Advances in Neural Information Processing Systems. 2017: 4470-4478.

[14] HU J, SHEN L, SUN G. Squeeze-and-excitation networks[J]. arXiv: 1709.01507, 2017.

[15] ZHAI S, CHENG Y, ZHANG Z M, et al. Doubly convolutional neural networks[C]// Advances in Neural Information Processing Systems. 2016: 1082-1090.

[16] HYVÄRINEN A, KÖSTER U. Complex cell pooling and the statistics of natural images[J]. Network: Computation in Neural Systems, 2007, 18(2): 81-100.

[17] BRUNA J, SZLAM A, LECUN Y. Signal recovery from pooling representations[J]. arXiv: 1311.4025, 2013.

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

[19] WAN L, ZEILER M, ZHANG S, et al. Regularization of neural networks using dropconnect[C]// International Conference on Machine Learning. 2013: 1058-1066.

[20] YU D, WANG H, CHEN P, et al. Mixed pooling for convolutional neural networks[C]// International Conference on Rough Sets and Knowledge Technology. Springer, Cham, 2014: 364-375.

[21] ZEILER M D, FERGUS R. Stochastic pooling for regularization of deep convolutional neural networks[J]. arXiv: 1301.3557, 2013.

[22] HE K, ZHANG X, REN S, et al. Spatial pyramid pooling in deep convolutional networks for visual recognition[C]// European Conference on Computer Vision. Springer, Cham, 2014: 346-361.

[23] RIPPEL O, SNOEK J, ADAMS R P. Spectral representations for convolutional neural networks[C]// Advances in Neural Information Processing Systems. 2015: 2449-2457.

[24] NAIR V, HINTON G E. Rectified linear units improve restricted Boltzmann machines[C]// Proceedings of the 27th international conference on machine learning (ICML-10). 2010: 807-814.

[25] MAAS A L, HANNUN A Y, NG A Y. Rectifier nonlinearities improve neural network acoustic models[C]// Proc. ICML. 2013.

[26] HE K, ZHANG X, REN S, et al. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification [C]// Proceedings of the IEEE International Conference on Computer Vision. 2015: 1026-1034.

[27] RAMACHANDRAN P, ZOPH B, LE Q. Searching for activation functions[J]. arXiv: 1710.05941.

[28] NGUYEN D T, LI W, OGUNBONA P O. Human detection from images and videos: A survey[J]. Pattern Recognition, 2016, 51(C): 148-175.

[29] LI Y, WANG S, TIAN Q, et al. Feature representation for statistical-learning-based object detection: A review[J]. Pattern Recognition, 2015, 48(11): 3542-3559.

[30] PEDERSOLI M, VEDALDI A, GONZÁLEZ J, et al. A coarse-to-fine approach for fast deformable object detection[J]. Pattern Recognition, 2015, 48(5): 1844-1853.

[31] NOWLAN S J, PLATT J C. A convolutional neural network hand tracker[C]// Advances in Neural Information Processing Systems. 1995: 901-908.

[32] GIRSHICK R, IANDOLA F, DARRELL T, et al. Deformable part models are convolutional neural networks[C]// Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. 2015: 437-446.

[33] GIRSHICK R, DONAHUE J, DARRELL T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2014: 580-587.

[34] SERMANET P, EIGEN D, ZHANG X, et al. Overfeat: Integrated recognition, localization and detection using convolutional networks[J]. arXiv: 1312.6229, 2013.

[35] GIRSHICK R. Fast r-cnn[C]// Proceedings of the IEEE International Conference on Computer Vision. 2015: 1440-1448.

[36] 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. 2015: 91-99.

[37] LIN T Y, DOLLÁR P, GIRSHICK R, et al. Feature pyramid networks for object detection[C]// CVPR. 2017: 4.

[38] HE K,GKIOXARI G,DOLLÁR,et al. Mask r-cnn[C]// 2017 IEEE International Conference on Computer Vision (ICCV). IEEE,2017:2980-2988.

[39] 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.

[40] REDMON J,DIVVALA S,GIRSHICK R,et al. You only look once:Unified,real-time object detection[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016:779-788.

[41] LIU W,ANGUELOV D,ERHAN D,et al. Ssd: Single shot multibox detector[C]// European Conference on Computer Vision. Springer,Cham,2016:21-37.

[42] REDMON J,FARHADI A. YOLO9000:better,faster,stronger [C]// 2017 IEEE Conference on Computer Vision and Pattern Recognition(CVPR). Honolulu,Hawaii,USA,2017.

[43] FU C Y,LIU W,RANGA A,et al. DSSD:Deconvolutional single shot detector[J]. arXiv:1701.06659,2017.

[44] PINHEIRO P O,COLLOBERT R,DOLLÁR P. Learning to segment object candidates[C]// Advances in Neural Information Processing Systems. 2015:1990-1998.

[45] PINHEIRO P O,LIN T Y,COLLOBERT R,et al. Learning to refine object segments[C]// European Conference on Computer Vision. Springer,Cham,2016:75-91.

[46] ZAGORUYKO S,LERER A,LIN T Y,et al. A multipath network for object detection[J]. arXiv:1604.02135,2016.

[47] LONG J,SHELHAMER E,DARRELL T. Fully convolutional networks for semantic segmentation[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015:3431-3440.

[48] CHEN L C,PAPANDREOU G,KOKKINOS I,et al. Deeplab: Semantic image segmentation with deep convolutional nets,atrous convolution,and fully connected crfs[J]. arXiv: 1606.00915,2016.

[49] DAI J,HE K,LI Y,et al. Instance-sensitive fully convolutional networks[C]// European Conference on Computer Vision. Springer,Cham,2016:534-549.

[50] DAI J,HE K,SUN J. Instance-aware semantic segmentation via multi-task network cascades[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016: 3150-3158.

[51] ZHANG K,SONG H. Real-time visual tracking via online weighted multiple instance learning[J]. Pattern Recognition, 2013,46(1):397-411.

[52] ZHANG S,YAO H,SUN X,et al. Sparse coding based visual tracking:Review and experimental comparison[J]. Pattern Recognition,2013,46(7):1772-1788.

[53] ZHANG S,WANG J,WANG Z,et al. Multi-target tracking by learning local-to-global trajectory models[J]. Pattern Recognition,2015,48(2):580-590.

[54] FAN J,XU W,WU Y,et al. Human tracking using convolutional neural networks[J]. IEEE Transactions on Neural Networks,2010,21(10):1610-1623.

[55] LI H,LI Y,PORIKLI F. DeepTrack:Learning Discriminative Feature Representations by Convolutional Neural Networks for Visual Tracking[C]// Proceedings British Machine Vision Conference. 2014:3.

[56] CHEN Y,YANG X,ZHONG B,et al. CNNTracker:online discriminative object tracking via deep convolutional neural network[J]. Applied Soft Computing,2016,38:1088-1098.

[57] HONG S,YOU T,KWAK S,et al. Online tracking by learning discriminative saliency map with convolutional neural network [C]// International Conference on Machine Learning. 2015:597-606.

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

[59] KARPATHY A,TODERICI G,SHETTY S,et al. Large-scale video classification with convolutional neural networks[C]// Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. 2014:1725-1732.

[60] SIMONYAN K,ZISSERMAN A. Two-stream convolutional networks for action recognition in videos[C]// Advances in Neural Information Processing Systems. 2014:568-576.

[61] CHÉRON G,LAPTEV I,SCHMID C. P-CNN:Pose-based CNN features for action recognition[C]// Proceedings of the IEEE International Conference Cn Vomputer vision. 2015:3218-3226.

[62] DONG C,LOY C C,HE K,et al. Learning a deep convolutional network for image super-resolution[C]// European Conference on Computer Vision. Springer,Cham,2014:184-199.

[63] DONG C,LOY C C,TANG X. Accelerating the super-resolution convolutional neural network[C]// European Conference on Computer Vision. Springer International Publishing,2016:391-407.

[64] SHI W,CABALLERO J,HUSZÁR F,et al. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016: 1874-1883.

[65] KIM J,KWON LEE J,MU LEE K. Accurate image super-resolution using very deep convolutional networks[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016:1646-1654.

[66] LAI W S,HUANG J B,AHUJA N,et al. Deep Laplacian Pyramid Networks for Fast and Accurate Super-Resolution[J]. arXiv:1704.03915,2017.