图像理解中的卷积神经网络

常 亮[1,2] 邓小明[3] 周明全[1,2] 武仲科[1,2] 袁 野[3,4] 杨 硕[3,4] 王宏安[3]

摘 要 近年来, 卷积神经网络 (Convolutional neural networks, CNN) 已在图像理解领域得到了广泛的应用, 引起了研究者的关注. 特别是随着大规模图像数据的产生以及计算机硬件 (特别是 GPU) 的飞速发展, 卷积神经网络以及其改进方法在图像理解中取得了突破性的成果, 引发了研究的热潮. 本文综述了卷积神经网络在图像理解中的研究进展与典型应用. 首先, 阐述卷积神经网络的基础理论; 然后, 阐述其在图像理解的具体方面, 如图像分类与物体检测、人脸识别和场景的语义分割等的研究进展与应用.

关键词 卷积神经网络, 图像理解, 深度学习, 图像分类, 物体检测

引用格式 常亮, 邓小明, 周明全, 武仲科, 袁野, 杨硕, 王宏安. 图像理解中的卷积神经网络. 自动化学报, 2016, **42**(9): 1300−1312

DOI 10.16383/j.aas.2016.c150800

Convolutional Neural Networks in Image Understanding

CHANG Liang[1,2] DENG Xiao-Ming[3] ZHOU Ming-Quan[1,2] WU Zhong-Ke[1,2] YUAN Ye[3,4]
YANG Shuo[3,4] WANG Hong-An[3]

Abstract Convolutional neural networks (CNN) have been widely applied to image understanding, and they have arose much attention from researchers. Specifically, with the emergence of large image sets and the rapid development of GPUs, convolutional neural networks and their improvements have made breakthroughs in image understanding, bringing about wide applications into this area. This paper summarizes the up-to-date research and typical applications for convolutional neural networks in image understanding. We firstly review the theoretical basis, and then we present the recent advances and achievements in major areas of image understanding, such as image classification, object detection, face recognition, semantic image segmentation etc.

Key words Convolutional neural networks (CNN), image understanding, deep learning, image classification, object detection

Citation Chang Liang, Deng Xiao-Ming, Zhou Ming-Quan, Wu Zhong-Ke, Yuan Ye, Yang Shuo, Wang Hong-An. Convolutional neural networks in image understanding. *Acta Automatica Sinica*, 2016, **42**(9): 1300−1312

1986 年, Rumelhart 等[1] 提出人工神经网络的反向传播算法 (Back propagation, BP), 掀起了神经网络在机器学习中的研究热潮. 神经网络中存在大量的参数, 存在容易发生过拟合、训练时间长的缺陷, 但是与基于规则的学习相比已经具有优越性. 基于统计学习理论的支持向量机[2]、Boosting、Logistic 回归方法可以被看作具有一层隐节点或者不含隐节点的学习模型, 被称为浅层机器学习模型. 浅层学习模型通常需要由人工方法获取好的样本特征, 在此基础上进行识别和预测, 因此方法的有效性很大程度上受到特征提取的制约[3].

2006 年, Hinton 等[4] 在 *Science* 上提出了深度学习. 这篇文章的两个主要观点是: 1) 多隐层的人工神经网络具有优异的特征学习能力, 学习到的数据更能反映数据的本质特征, 有利于可视化或分类; 2) 深度神经网络在训练上的难度, 可以通过逐层无监督训练有效克服. 理论研究表明为了学习到可表示高层抽象特征的复杂函数, 需要设计深度结构. 深度结构由多层非线性算子构成, 典型设计是具有多层隐节点的神经网络. 随着网络层数的加大, 如何搜索深度结构的参数空间成为具有挑战性的任务. 近年来, 深度学习取得成功的主要原因有: 1) 在训练

收稿日期 2015-12-11 录用日期 2016-05-03
Manuscript received December 11, 2015; accepted May 3, 2016
国家自然科学基金 (61402040, 61473276), 中国科学院青年创新促进会资助
Supported by National Natural Science Foundation of China (61402040, 61473276) and Youth Innovation Promotion Association, Chinese Academy of Sciences
本文责任编委 柯登峰
Recommended by Associate Editor KE Deng-Feng
1. 北京师范大学信息科学与技术学院 北京 100875 2. 教育部虚拟现实应用工程研究中心 北京 100875 3. 中国科学院软件研究所人机交互北京市重点实验室 北京 100190 4. 中国科学院大学计算机与控制学院 北京 100049
1. College of Information Science and Technology, Beijing Normal University, Beijing 100875 2. Engineering Research Center of Virtual Reality and Applications, Ministry of Education, Beijing 100875 3. Beijing Key Laboratory of Human-Computer Interactions, Institute of Software, Chinese Academy of Sciences, Beijing 100190 4. School of Computer and Control Engineering, University of Chinese Academy of Sciences, Beijing 100049

数据上, 大规模训练数据的出现 (如 ImageNet[5]), 为深度学习提供了好的训练资源; 2) 计算机硬件的飞速发展 (特别是 GPU 的出现) 使得训练大规模神经网络成为可能. 与浅层学习模型相比, 深度学习构造了具有多隐层的学习模型, 设计了有效的学习算法并能够加速计算, 从而能够对大数据进行处理; 通过深度学习能够得到更高层的特征, 从而提高样本的识别率或预测的准确率.

卷积神经网络 (Convolutional neural networks, CNN) 是一种带有卷积结构的深度神经网络, 卷积结构可以减少深层网络占用的内存量, 也可以减少网络的参数个数, 缓解模型的过拟合问题. 1989 年, LeCun 等[6] 在手写数字识别中采用神经网络误差反向传播算法, 在网络结构设计中加入下采样 (Undersampling) 与权值共享 (Weight sharing). 1998 年, LeCun 等[7] 提出用于文档识别的卷积神经网络, 为了保证一定程度的平移、尺度、畸变不变性, CNN 设计了局部感受野、共享权重和空间或时间下采样, 提出用于字符识别的卷积神经网络 LeNet-5. LeNet-5 由卷积层、下采样层、全连接层构成, 该系统在小规模手写数字识别中取得了较好的结果. 2012 年, Krizhevsky 等[8] 采用称为 AlexNet 的 CNN 在 ImageNet 竞赛图像分类任务中取得了最好的成绩, 是 CNN 在大规模图像分类中的巨大成功. AlexNet 网络具有更深层的结构, 并设计了 ReLU (Rectified linear unit) 作为非线性激活函数以及 Dropout 来避免过拟合. 在图像分类中, 一个重要的图像数据库是 ImageNet[5]. 针对具有 80 000 个同义词的词汇网络 (WordNet), ImageNet 旨在分别使用 500 ∼ 1 000 个清晰的全分辨率图像来表示其中的大部分词汇, 这样就形成了上百万张有标记的图像, 它们通过词汇网络的语义结构组织起来. ImageNet 总共包括 12 个子树、5 247 个同义词集、320 万图像, 是目标检测、图像分类、图像定位研究的优越资源, ImageNet 在大规模、准确度、分层结构方面为计算机视觉研究者提供了前所未有的机会. 表 1 是 ImageNet 竞赛历年来图像分类任务的部分领先结果. 在 AlexNet 之后, 研究者又进一步改善网络性能, 提出能有效分类检测的 R-CNN (Region-based CNN)[9]、SPP (Spatial pyramid pooling)-net[10]、GoogLeNet[11]、VGG (Visual geometry group)[12] 等. 为了更好地改进卷积神经网络, 使其在应用中发挥更大的功效, 研究者不仅从应用的特殊性、网络的结构等方面进一步探讨卷积神经网络, 而且从其中的网络层设计、损失函数的设计、激活函数、正则项等多方面对现有网络进行改进, 取得了一系列成果.

计算机视觉的中心任务就是通过对图像或图像序列的分析, 得到景物的尽可能完全正确的描述[13]. 图像理解与计算机视觉紧密相关, 研究内容交叉重合, 图像理解侧重在图像分析的基础上, 理解图像内容的含义以及解释原来的客观场景, 从而指导和规划行动[14]. 图像理解是深度学习应用最早的领域, 也是其应用最广的领域之一. 随着互联网大数据的兴起, 深度学习在大规模图像的处理中显示了不可替代的优越性. 卷积神经网络的研究已经在图像理解中广泛应用[3]. 本文着重阐述卷积神经网络的理论和面向图像理解几个不同方面的卷积神经网络的提出、进展和应用, 包括: 图像分类和物体检测、人脸识别和验证、场景的语义分割和深度恢复、人体关节检测, 通过这些介绍希望能帮助读者了解相关工作的方法和思路并启发新的研究思路.

表 1 ImageNet 竞赛历年来图像分类任务的部分领先结果

Table 1 Representative top ranked results in image classification task of "ImageNet Large Scale Visual Recognition Challenge"

公布时间	机构	Top-5 错误率 (%)
2015.12.10	MSRA	3.57[15]
2014.8.18	Google	6.66[11]
2014.8.18	Oxford	7.33[12]
2013.11.14	NYU	11.7
2012.10.13	U.Toronto	16.4[8]

1 卷积神经网络

卷积神经网络是深度学习的一种, 已成为当前图像理解领域的研究热点[6, 16−17] 它的权值共享网络结构使之更类似于生物神经网络, 降低了网络模型的复杂度, 减少了权值的数量. 这个优点在网络的输入是多维图像时表现得更为明显, 图像可以直接作为网络的输入, 避免了传统识别算法中复杂的特征提取和数据重建过程. 卷积网络是为识别二维形状而特殊设计的一个多层感知器, 这种网络结构对平移、比例缩放以及其他形式的变形具有一定不变性. 在典型的 CNN 中, 开始几层通常是卷积层和下采样层的交替, 在靠近输出层的最后几层网络通常是全连接网络 (如图 1 所示). 卷积神经网络的训练过程主要是学习卷积层的卷积核参数和层间连接权重等网络参数, 预测过程主要是基于输入图像和网络参数计算类别标签. 卷积神经网络的关键是: 网络结构 (含卷积层、下采样层、全连接层等) 和反向传播算法等.

在本节中, 我们先介绍典型 CNN 的网络结构和反向传播算法, 然后概述常用的其他 CNN 网络结构和方法. 神经网络参数的中文名称主要参考文献 [18] 卷积神经网络的结构和反向传播算法主要参考文献 [17].

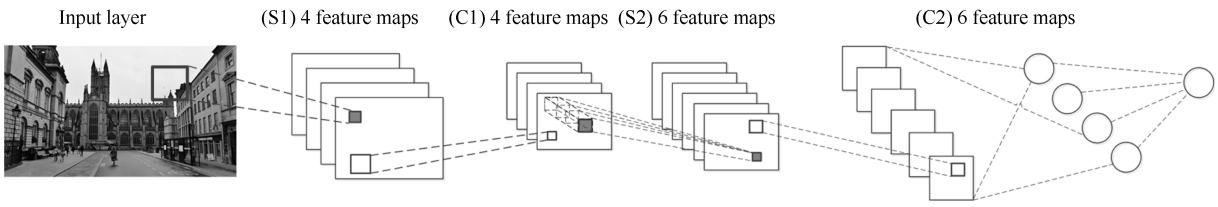


图 1 卷积神经网络示例

Fig. 1 Illustration of convolutional neural networks

1.1 网络结构

1.1.1 卷积层

在卷积层, 上一层的特征图 (Feature map) 被一个可学习的卷积核进行卷积, 然后通过一个激活函数 (Activation function), 就可以得到输出特征图. 每个输出特征图可以组合卷积多个特征图的值[17]:

$$\begin{aligned} x_j^l &= f(u_j^l) \\ u_j^l &= \sum_{i\in M_j} x_i^{l-1} * k_{ij}^l + b_j^l \end{aligned} \tag{1}$$

其中, u_j^l 称为卷积层 l 的第 j 个通道的净激活 (Net activation), 它通过对前一层输出特征图 x_i^{l-1} 进行卷积求和与偏置后得到的, x_j^l 是卷积层 l 的第 j 个通道的输出. $f(\cdot)$ 称为激活函数, 通常可使用 sigmoid 和 tanh 等函数. M_j 表示用于计算 u_j^l 的输入特征图子集, k_{ij}^l 是卷积核矩阵, b_j^l 是对卷积后特征图的偏置. 对于一个输出特征图 x_j^l, 每个输入特征图 x_i^{l-1} 对应的卷积核 k_{ij}^l 可能不同, "*" 是卷积符号.

1.1.2 下采样层

下采样层将每个输入特征图通过下面的公式下采样输出特征图[17]:

$$\begin{aligned} x_j^l &= f(u_j^l) \\ u_j^l &= \beta_j^l \mathrm{down}(x_j^{l-1}) + b_j^l \end{aligned} \tag{2}$$

其中, u_j^l 称为下采样层 l 的第 j 通道的净激活, 它由前一层输出特征图 x_i^{l-1} 进行下采样加权、偏置后得到, β 是下采样层的权重系数, b_j^l 是下采样层的偏置项. 符号 down$(\cdot)$ 表示下采样函数, 它通过对输入特征图 x_j^{l-1} 通过滑动窗口方法划分为多个不重叠的 $n\times n$ 图像块, 然后对每个图像块内的像素求和、求均值或最大值, 于是输出图像在两个维度上都缩小了 n 倍.

1.1.3 全连接层

在全连接网络中, 将所有二维图像的特征图拼接为一维特征作为全连接网络的输入. 全连接层 l 的输出可通过对输入加权求和并通过激活函数的响应得到[17]:

$$\begin{aligned} x^l &= f(u^l) \\ u^l &= w^l x^{l-1} + b^l \end{aligned} \tag{3}$$

其中, u^l 称为全连接层 l 的净激活, 它由前一层输出特征图 x^{l-1} 进行加权和偏置后得到的. w^l 是全连接网络的权重系数, b^l 是全连接层 l 的偏置项.

1.2 反向传播算法

神经网络有两类基本运算模式: 前向传播和学习. 前向传播是指输入信号通过前一节中一个或多个网络层之间传递信号, 然后在输出层得到输出的过程. 反向传播算法是神经网络有监督学习中的一种常用方法, 其目标是根据训练样本和期望输出来估计网络参数. 对于卷积神经网络而言, 主要优化卷积核参数 k、下采样层网络权重 β、全连接层网络权重 w 和各层的偏置参数 b 等. 反向传播算法的本质在于允许我们对每个网络层计算有效误差, 并由此推导出一个网络参数的学习规则, 使得实际网络输出更加接近目标值[18].

我们以平方误差损失函数的多分类问题为例介绍反向传播算法的思路. 考虑一个多分类问题的训练总误差, 定义为输出端的期望输出值和实际输出值的差的平方[17]:

$$E(w,\beta,k,b) = \frac{1}{2}\sum_{n=1}^{N} \|t_n - y_n\|^2 \tag{4}$$

其中, t_n 是第 n 个样本的类别标签真值, y_n 是第 n 个样本通过前向传播网络预测输出的类别标签. 对于多分类问题, 输出类别标签常用一维向量表示, 即输入样本对应的类别标签维度为正数, 输出类别标签的其他维为 0 或负数, 这取决于选择的激活函数类型, 当激活函数选为 sigmoid, 输出标签为 0, 当激活函数为 tanh, 输出标签为 -1.

反向传播算法主要基于梯度下降方法, 网络参数首先被初始化为随机值, 然后通过梯度下降法向训练误差减小的方向调整. 接下来, 我们以多个 "卷积层 – 采样层" 连接多个全连接层的卷积神经网络为例介绍反向传播算法.

首先介绍网络第 l 层的灵敏度 (Sensitivity)[17−18]:

$$\delta^l = \frac{\partial E}{\partial u^l} \tag{5}$$

其中, δ^l 描述了总误差 E 怎样随着净激活 u^l 而变化. 反向传播算法实际上通过所有网络层的灵敏度

建立总误差对所有网络参数的偏导数, 从而得到使得训练误差减小的方向.

1.2.1 卷积层

为计算卷积层 l 的灵敏度, 需要用下一层下采样层 $l+1$ 的灵敏度表示卷积层 l 的灵敏度, 然后计算总误差 E 对卷积层参数 (卷积核参数 k、偏置参数 b) 的偏导数.

由于下采样层的灵敏度尺寸小于卷积层的灵敏度尺寸, 因此需要将下采样层 $l+1$ 的灵敏度上采样到卷积层 l 的灵敏度大小, 然后将第 l 层净激活的激活函数偏导与从第 $l+1$ 层的上采样得到的灵敏度逐项相乘. 分别由式 (1) 和 (2), 通过链式求导可得第 l 层中第 j 个通道的灵敏度[17]:

$$\delta_j^l = \frac{\partial E}{\partial u_j^l} = \beta_j^{l+1}(f'(u_j^l) \circ \mathrm{up}(\delta_j^{l+1})) \tag{6}$$

其中, $\mathrm{up}(\cdot)$ 表示一个上采样操作, 符号 $\circ$ 表示每个元素相乘. 若下采样因子为 n, 则 $\mathrm{up}(\cdot)$ 将每个像素在水平和垂直方向上复制 n 次, 于是就可以从 $l+1$ 层的灵敏度上采样成卷积层 l 的灵敏度大小. 函数 $\mathrm{up}(\cdot)$ 可以用 Kronecker 乘积 $\mathrm{up}(x) \equiv x \otimes 1_{n\times n}$ 来实现.

然后, 使用灵敏度对卷积层 l 中的参数计算偏导. 对于总误差 E 对偏移量 b_j^l 的偏导, 可以对卷积层 l 的灵敏度中所有节点进行求和来计算:

$$\frac{\partial E}{\partial b_j^l} = \sum_{u,v} (\delta_j^l)_{u,v} \tag{7}$$

对于总误差关于卷积核参数的偏导, 由式 (1), 使用链式求导时需要用所有与该卷积核相乘的特征图元素来求偏导:

$$\frac{\partial E}{\partial k_{ij}^l} = \sum_{u,v} (\delta_j^l)_{u,v} (p_i^{l-1})_{u,v} \tag{8}$$

其中, $(p_i^{l-1})_{u,v}$ 是在计算 x_j^l 时, 与 k_{ij}^l 逐元素相乘的 x_i^{l-1} 元素.

1.2.2 下采样层

为计算下采样层 l 的灵敏度, 需要用下一层卷积层 $l+1$ 的灵敏度表示下采样层 l 的灵敏度, 然后计算总误差 E 对下采样参数权重系数 β、偏置参数 b 的偏导数.

为计算我们需要下采样层 l 的灵敏度, 我们必须找到当前层的灵敏度与下一层的灵敏度的对应点, 这样才能对灵敏度 δ 进行递推. 另外, 需要乘以输入特征图与输出特征图之间的连接权值, 这个权值实际上就是卷积核的参数. 分别由式 (1) 和 (2), 通过链式求导可得第 l 层第 j 个通道的灵敏度[17]:

$$\delta_j^l = f'(u_j^l) \circ \mathrm{conv2}(\delta_j^{l+1}, \mathrm{rot180}(k_j^{l+1})', \text{ full}') \tag{9}$$

其中, 对卷积核旋转 180 度使用卷积函数计算互相关 (在 Matlab 中, 可用 conv2 函数实现), 对卷积边界进行补零处理.

然后, 总误差对偏移量 b 的偏导与前面卷积层的一样, 只要对灵敏度中所有元素的灵敏度求和即可:

$$\frac{\partial E}{\partial b_j^l} = \sum_{u,v} (\delta_j^l)_{u,v} \tag{10}$$

对于下采样权重 β, 我们先定义下采样算子 $d_j^l = \mathrm{down}(x_j^{l-1})$, 然后可通过下面的公式计算总误差 E 对 β 的偏导:

$$\frac{\partial E}{\partial \beta_j^l} = \sum_{u,v} (\delta_j^l \circ d_j^l)_{u,v} \tag{11}$$

这里我们假定下采样层的下一层为卷积层, 如果下一层为全连接层, 也可以做类似的推导.

1.2.3 全连接层

全连接层 l 的灵敏度可通过下式计算:

$$\delta^l = (w^{l+1})^{\mathrm{T}} \delta^{l+1} \circ f'(u^l) \tag{12}$$

输出层的神经元灵敏度可由下面的公式计算:

$$\delta^L = f'(u^L) \circ (y^n - t^n) \tag{13}$$

总误差对偏移项的偏导如下:

$$\frac{\partial E}{\partial b^l} = \frac{\partial E}{\partial u^l}\frac{\partial u^l}{\partial b^l} = \delta^l \tag{14}$$

接下来可以对每个神经元运用灵敏度进行权值更新. 对一个给定的全连接层 l, 权值更新方向可用该层的输入 x^{l-1} 和灵敏度 δ^l 的内积来表示:

$$\frac{\partial E}{\partial w^l} = x^{l-1}(\delta^l)^{\mathrm{T}} \tag{15}$$

1.2.4 网络参数更新过程

卷积层参数可用下式更新:

$$\Delta k_{ij}^l = -\eta \frac{\partial E}{\partial k_{ij}^l} \tag{16}$$

$$\Delta b^l = -\eta \frac{\partial E}{\partial b^l} \tag{17}$$

下采样层参数可用下式更新:

$$\Delta \beta^l = -\eta \frac{\partial E}{\partial \beta^l} \tag{18}$$

$$\Delta b^l = -\eta \frac{\partial E}{\partial b^l} \tag{19}$$

全连接层参数可用下式更新:

$$\Delta w^l = -\eta \frac{\partial E}{\partial w^l} \tag{20}$$

其中, 对于每个网络参数都有一个特定的学习率 η. 若学习率太小, 则训练的速度慢; 若学习率太大, 则可导致系统发散. 在实际问题中, 如果总误差在学习过程中发散, 那么将学习率调小; 反之, 如果学习速度过慢, 那么将学习率调大.

1.3 常用的其他网络结构和方法

1.3.1 卷积层

传统卷积神经网络的卷积层采用线性滤波器与非线性激活函数, 一种改进的方法在卷积层使用多层感知机模型作为微型神经网络, 通过在输入图像中滑动微型神经网络来得到特征图, 该方法能够增加神经网络的表示能力, 被称为 Network in network[19]. 为了解决既能够保证网络的稀疏性, 又能够利用稠密矩阵的高性能计算, Szegedy 等[11] 提出 Inception 网络. Inception 网络的一层含有一个池化操作和三类卷积操作: 1×1、3×3、5×5 卷积.

1.3.2 池化

池化 (Pooling) 是卷积神经网络中一个重要的操作, 它能够使特征减少, 同时保持特征的局部不变性. 常用的池化操作有: 空间金字塔池化 (Spatial pyramid pooling, SPP)[10]、最大池化 (Max pooling)、平均池化 (Mean pooling)、随机池化 (Stochastic pooling)[20] 等. 本文第 1.1.2 节介绍的下采样层实际上也属于池化.

1.3.3 激活函数

常用激活函数有: ReLU[8]、Leakly ReLU[21]、Parametric ReLU、Randomized ReLU、ELU 等.

1.3.4 损失函数

损失函数的选择在卷积神经网络中起重要作用, 代表性的损失函数有: 平方误差损失、互熵损失 (Cross entropy loss)、Hinge 损失等.

1.3.5 优化方法和技巧

卷积神经网络常用的优化方法包含随机梯度下降方法 (Stochastic gradient descent, SGD), 常用的技巧有权值初始化[8]、权值衰减 (Weight decay)[18]、Batch normalization[22] 等.

1.4 卷积神经网络的优势

卷积神经网络在下采样层可以保持一定局部平移不变形, 在卷积层通过感受野和权值共享减少了神经网络需要训练的参数的个数. 每个神经元只需要感受局部的图像区域, 在更高层将这些感受不同局部区域的神经元综合起来就可以得到全局的信息. 因此, 可以减少网络连接的数目, 即减少神经网络需要训练的权值参数的个数. 由于同一特征通道上的神经元权值相同, 所以网络可以并行学习, 这也是卷积网络相对于神经元彼此相连网络的一大优势. 卷积神经网络以其权值共享的特殊结构在图像理解领域中有着独特的优越性, 通过权值共享降低了网络的复杂性.

总之, 卷积神经网络相比于一般神经网络在图像理解中有其特殊的优点: 1) 网络结构能较好适应图像的结构; 2) 同时进行特征提取和分类, 使得特征提取有助于特征分类; 3) 权值共享可以减少网络的训练参数, 使得神经网络结构变得更简单、适应性更强.

2 卷积神经网络在图像理解中的进展与应用

本节将介绍卷积神经网络在图像分类与物体检测、人脸识别和验证、语义图像分割等方面的进展与应用.

2.1 图像分类和物体检测

图像分类和物体检测是图像理解中的核心问题之一. 图像分类是指给定图像, 对图像的类别进行预测; 物体检测是指对于图像中的同一物体或者同一类别物体进行检测, 找到可能出现物体的区域.

在图像分类和物体检测中, 传统的方法包含基于词袋 (Bag of words, BOW) 的方法和基于变形模板模型 (Deformable part models, DPM)[23] 的方法等. 这些方法虽然在某些特定应用 (如人脸检测、行人检测等) 中取得了很好的效果, 但在准确性方面仍存在较大提升空间. 随着深度学习的兴起, 人们将深度学习应用于图像分类和物体检测问题中, 并在许多应用中取得明显好于传统方法的结果. 在图像分类中, Krizhevsky 等[8] 提出了新型卷积神经网络结构 (AlexNet), GoogLeNet[11] 和 VGG[12] 通过加深网络层数同时保证优化性能, 设计了更深层次的卷积神经网络. 在物体检测中, 研究者使用区域选择性搜索[9] 等技术提升检测的准确率, 通过加入感兴趣区域池化层 (Region of interest (ROI) pooling layer)[24] 和空间金字塔池化[10] 等技术加速网络计算速度. 此外, 也有一部分工作将卷积神经网络特征与传统视觉识别模型结合起来. Girshick 等[25] 利用深度学习的特征代替原有人工设计的方向梯度直方图 (Histogram of oriented gradient, HOG) 特征[26] 建立变形模板, 提升了传统变形模板方法 (DPM) 的识别率, 并且在取得了与完全使用深度学习方法可比结果的同时, 提升了检测速度. 表 2 给出部分具有代表性的图像分类和物体检测模型对比.

接下来, 我们分别介绍面向图像分类和物体检测任务的 AlexNet 及代表性的改进方法、其他代表性的改进方向.

2.1.1 AlexNet 及代表性的改进方法

Krizhevsky 等[8] 提出新型卷积神经网络结构 (简称为 AlexNet, 其网络结构如图 2 所示), 并在

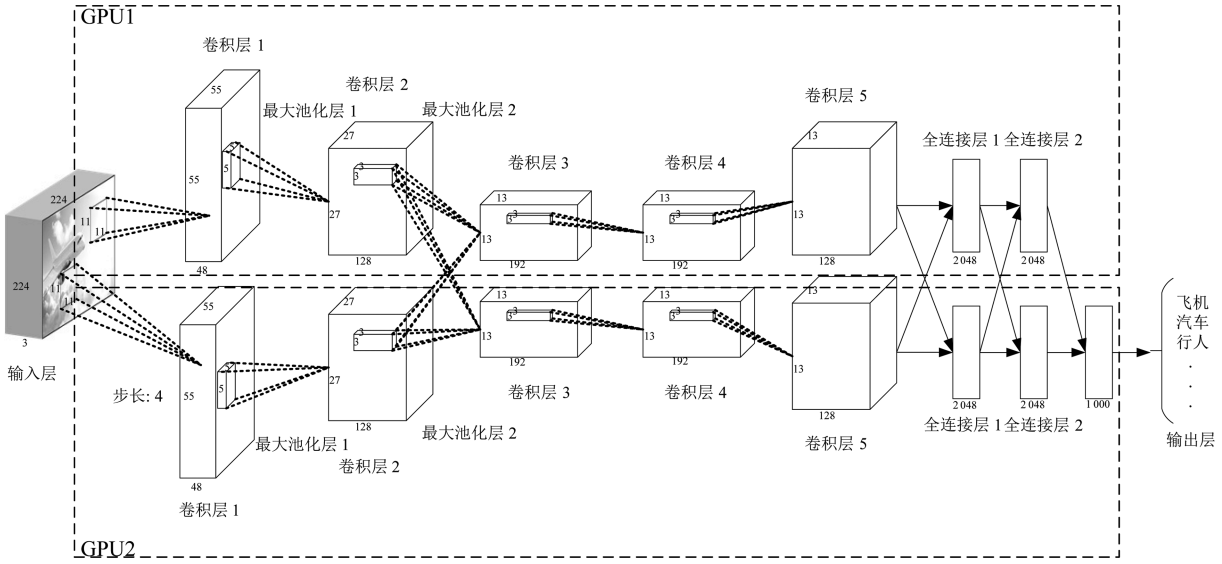


图 2　AlexNet 卷积神经网络结构示意图[8]

Fig. 2　Network architecture of AlexNet convolutional neural networks[8]

ImageNet ILSVRC-2012 图像分类问题中取得最好成绩 (Top-5 错误率为 15.3 %), 其结果明显好于使用传统方法的第二名取得的结果 (Top-5 错误率为 26.2 %). 该方法训练了一个端对端 (End to end) 的卷积神经网络实现图像特征提取和分类, 网络结构共 7 层, 包含 5 层卷积层和 2 层全连接层. AlexNet 在训练阶段使用了 Dropout 技巧, 并通过图像平移、图像水平翻转、调整图像灰度等方法扩充训练数据集, 后者一方面通过扩充样本缓解了神经网络的过拟合以及对网络参数优化时陷入局部最优的问题, 也使得训练得到的网络对局部平移和光照变化具有一定的不变性. 为了加快网络训练的速度, AlexNet 采用 ReLU 代替传统神经网络常用的激活函数 tanh/sigmod, ReLU 是一种非饱和非线性 (Non-saturating nonlinearity) 变换.

Overfeat[27] 首次使用同一个模型完成图像分类、定位和物体检测三个任务, 其主要观点是通过共享部分网络完成这三个任务, 能相互促进每个任务的结果. Overfeat 继承了 AlexNet 的网络结构, 主要区别在于: AlexNet 在提出时主要面向图像分类任务, Overfeat 可以完成图像分类、定位和物体检测三个任务; Overfeat 在训练时输入固定大小的图像, 测试时用多尺度输入, 没有使用 AlexNet 中的对比度归一化, 采用无重叠区域的最大池化, 前两层的特征图更大. 对于分类与检测问题, 常采用滑动窗口对每一个图像块进行检测, 从而确定目标物体的类别与位置, 即都需要一个滑动窗口对整幅图像进行密集采样. 为提高计算效率, Overfeat 舍弃在图像层级的滑动, 转而在特征层级进行滑动, 明显减少了滑动窗口个数. 为了避免特征层级采样带来的稀疏问题, Overfeat 采用多次采样插值的方法解决. 对于图像分类、定位和物体检测问题的统一, Overfeat 采用复用权重的方式, 即在每一个尺度上同时运行分类网络和定位回归网络. 对于每一个尺度, 分类网络给出了图像块的类别概率分布, 回归网络进一步为每一类给出了包围盒和置信度. 最后, 综合这些信息, 给出分类与检测结果. Overfeat 虽然提出了将分类、定位、检测任务一起解决的思想, 但这三个任务在训练阶段仍是分开进行的[24].

AlexNet 用于物体检测时, 需要在图像金字塔上采用滑动窗口的方式逐个判断, 随着图像的增大待检测区域的数目呈平方上升. 为了解决这一问题, Girshick 等将候选框 (Region proposals) 方法与卷积神经网络相结合 (Girshick 等称之为 R-CNN), 采用仅对候选框逐个使用卷积神经网络判断的方式, 不仅提高了物体检测的效率, 也提高了检测的精度, 在 VOC2012 上取得了当时最好的检测平均精度 mAP (Mean average precision), 把在该数据集上的历史最好检测平均精度提高了约 30 %[9]. R-CNN 通过选择性搜索方法 (Selective search)[28] 对图像进行过分割 (Over-segmentation) 得到大量分割块, 根据分割图像块之间的纹理相似性和位置关系对分割图像块进行合并, 可以得到许多连通的稳定区域. 由于这些稳定区域通常包含待检测物体, 也称之为候选区域. 对于这些候选区域 R-CNN, 通过 AlexNet 网络可以得到具有较强分辨力的特征, 最后用这个特征进行分类. 该方法用于物体检测时, 为了提高物体定位精度, 采用了类似于 DPM 方法[23] 中使用的包围盒回归方法 (Bounding box regression). 与基于滑动窗口的物体检测方法相比, 使用候选框将显著减少判断的窗口个数, 提高物体检测效率; 此外通过调整候选框方法, 可以在保证召回率

表 2 部分具有代表性的图像分类和物体检测模型对比

Table 2 Comparison of representative image classification and object detection models

方法	输入	优点	缺点
AlexNet[8]	整张图像 (需要对图像放缩到固定大小)	网络简单易于训练, 对图像分类有较强的鉴别力	网络输入图像要求固定大小, 容易破坏物体的纵横比和上下文信息
GoogLeNet[11]	整张图像 (需要对图像放缩到固定大小)	对图像分类拥有非常强的鉴别力, 参数相对 AlexNet 较少	网络复杂, 对样本数量要求较高, 训练耗时
VGG[12]	整张图像 (需要对图像放缩到固定大小)	对图像分类拥有非常强的鉴别力	网络复杂, 对样本数量要求较高, 训练耗时, 需要多次对网络参数的微调 (Fine-tuning)
DPM[23]	整张图像	对物体检测拥有较强的鉴别力, 对形变和遮挡具有一定的处理能力	使用人工设计的 HOG 特征[26]; 对物体检测的精度通常比本表中其他的 CNN 网络低
R-CNN[9]	图像区域	对物体检测拥有很强的鉴别力; 比在图像金字塔上逐层滑动窗口的物体检测方法效率高; 使用包围盒回归 (Bounding box regression) 提高物体的定位精度	依赖于区域选择算法; 网络输入图像要求固定大小, 容易破坏物体的纵横比和上下文信息; 训练是多阶段过程: 在特定检测数据集上对网络参数进行微调、提取特征、训练 SVM (Support vector machine) 分类器、包围盒回归 (Bounding box regression); 训练时间耗时、耗存储空间
SPP-net[10]	整张图像 (不要求固定大小)	对物体检测拥有很强的鉴别力, 输入图像可以任意大小, 可保证图像的比例信息训练速度比 R-CNN 快 3 倍左右, 测试比 R-CNN 快 10～100 倍	网络结构复杂时, 池化对图像造成一定的信息丢失; SPP 层前的卷积层不能进行网络参数更新[24]; 训练是多阶段过程: 在特定检测数据集上对网络参数进行微调、提取特征、训练 SVM 分类器、包围盒回归; 训练时间耗时、耗存储空间
Fast R-CNN[24]	整张图像 (不要求固定大小)	训练和测试都明显快于 SPP-net (除了候选区域提取以外的环节接近于实时), 对物体检测拥有很强的鉴别力, 输入图像可以任意大小, 保证图像比例信息, 同时进行分类与定位	依赖于候选区域选择, 它仍是计算瓶颈
Faster R-CNN[29]	整张图像 (不要求固定大小)	比 Fast R-CNN 更加快速, 对物体检测拥有很强的鉴别力; 不依赖于区域选择算法; 输入图像可以任意大小, 保证图像比例信息, 同时进行区域选择算法、分类与定位	训练过程较复杂; 计算流程仍有较大优化空间; 难以解决被遮挡物体的识别问题

的同时, 减少虚警 (False alarm), 进而提高物体检测精度. 在网络优化方面, R-CNN 采用 AlexNet 网络参数作为初值, 利用训练图像的候选区域数据对网络参数进行微调 (Fine-tuning), 这种方式比随机选取网络参数初值具有更快的收敛速度, 所需样本也更少. 在物体检测问题中, R-CNN 比 AlexNet 有明显的优势, 但仍存在一些不足: 1) 全连接层 (Full-connected layer) 只能接受固定尺寸的输入, R-CNN 要求对候选框进行缩放或裁减填充到固定大小, 这不仅会破坏物体的纵横比和图像大小信息, 也会破坏物体的上下文信息; 2) R-CNN 使用包围盒回归有助于提高物体的定位精度, 但如果待检测物体存在遮挡或交叉时, 该方法很难提高定位精度.

He 等[10] 针对之前卷积神经网络仅能接受固定尺寸的图像输入, 提出基于空间金字塔池化 (Spatial pyramid pooling, SPP) 的网络层, SPP 层放在最后一个卷积层后, 通过 SPP 层可得到固定长度的输出, 然后送入并重新学习全连接网络层 (这样的网络称为 SPP-net). 使用的网络结构类似于 AlexNet 的 7 层网络, 包含 5 层卷积层和 2 个全连接层网络, 主要区别是通过空间金字塔池化层连接卷积层与全连接层. 在该方法中, 对卷积第 5 层 conv5 输出的特征图分别进行 1 等分、4 等分、9 等分, 然后在每个分块进行池化操作 (如 Max pooling) 可得到定长的特征. SPP 既可保证特征包含图像的整体信息 (1 等分), 也保留了图像的局部信息 (4 等分、9 等分及更多等分), 由于特征是定长的, 无需关心空间金字塔池化前的上层网络输出特征图尺寸, 可以直接传

递给全连接网络层. 因此, SPP 可以解除对输入图像大小固定的限制, 图像可以保留原有大小直接进入网络进行训练与测试. 由于每张图像只需通过一次 5 层前向卷积, 避免了 R-CNN 用于物体检测时每个候选区域都需要通过 5 层前向卷积的耗时计算, 该方法于 2014 年在 VOC2007, Caltech101 数据集上取得当时最好成绩, 并在速度上比 R-CNN 提高了 24 ∼ 64 倍.

与 SPP-net 类似, Fast R-CNN[24] 也能用于不同大小的图像上的物体检测, 提出感兴趣区域池化 (RoI pooling). Fast R-CNN 可以完成提取特征, 分类和包围盒回归的端对端联合训练. 首先, 通过选择性搜索 (Selective search) 得到图像中的候选区域 (文中称为 ROI), 对图像建立图像金字塔并通过前向传播可得到 conv5 特征金字塔; 然后, 对于特征金字塔每个尺度的每个 ROI, 在 conv5 特征中取出对应的区域, 用一个称为 RoI pooling 的特殊单层 SPP 来统一到同样的大小的特征. 最后, 经过全连接层输出两任务的优化目标: 第一个任务是分类, 第二个任务是包围盒回归. Fast R-CNN 相比 SPP-net 的优势在于: SPP-net 中 SPP 层前的卷积层不能进行网络参数更新[24], 而 Fast R-CNN 可以; SPP-net 为进行包围盒回归, 需要使用额外的回归模型 (如线性 SVM 等), 包围盒回归不能融入整个网络训练. Fast R-CNN 在除了候选区域提取以外的环节接近于实时, 候选区域提取是计算中的瓶颈问题.

鉴于候选区域提取是 Fast R-CNN 的计算瓶颈, Ren 等[29] 提出了用于实时目标检测的候选框网络 (Region proposal network, RPN), RPN 是一个全卷积网络 (Fully convolutional network, FCN), 它可以从任意尺寸的图像中得到一系列的带有分数 (Objectness score) 的物体候选区域. RPN 能够生成高质量的候选区域, 并可以嵌入卷积神经网络中进行端对端的训练. RPN 与 Fast R-CNN 结合并共享卷积层特征的网络称为 Faster R-CNN, 它在 PASCAL VOC 2007、2012 和 MS COCO 数据集上取得了当时最好的检测结果, 并且整个计算过程接近于实时 (使用较深的 VGG 模型也可达到 5 fps).

2.1.2 其他代表性的改进方向

AlexNet 网络提出后, 许多工作开始关注改进 CNN 的结构, 如在最初的若干个卷积层使用更小的卷积窗口[19] 与卷积步长, 使用多尺度的训练与测试数据等, 但仍基于浅层网络. Zeiler 等[30] 对网络中层特征和分类器进行可视化分析, 得到在 ImageNet 上分类效果优于 AlexNet 的网络结构 ZF-net. ZF-net 把 AlexNet 中第一层卷积核的大小由 11×11 缩小为 7×7, 把卷积步长由 4 减小 2, 可得到更丰富的特征. VGG 模型[12] 是对深层卷积神经网络的一次系统尝试, 在 ILSVRC-2014 比赛中获得第二名的成绩. 相比于传统浅层网络问题 (5 ∼ 7 层), 网络随着层数的加深, 参数呈现指数级增长, VGG 模型采用多层小窗口卷积核代替一个大卷积核的方式减少参数的增长. 如使用三层具有 3×3 卷积核的卷积层代替一层具有 7×7 卷积核的卷积层, 如果通道数为 C, 那么一层具有 7×7 卷积核的卷积层共有 $7 \times 7 \times C \times C = 49C \times C$ 个参数, 而三层具有 3×3 卷积核的卷积网络共有 $3 \times (3 \times 3) \times C \times C = 27C \times C$ 个参数, 明显地减少了参数数目, 并且三层网络比一层网络更具有判别性. 该方法还使用 1×1 的卷积核[19], 可以在不影响卷积层感受野的情况下增加决策函数的非线性.

将卷积神经网络与传统视觉识别模型融合. Felzenszwalb 等[23] 提出的变形模板模型 DPM 将物体分解为多个可形变的基础语义组件, 这种目标检测方法融合物体整体信息, 语义组件信息与形变信息进行目标检测. 该方法结合了整体上下文信息与局部信息, 对形变、遮挡都有很好的鲁棒性. 该方法采用 HOG 特征[26], 但这种人工设计的特征不能保证对物体检测有很好的鉴别力. Girshick 等[25] 使用卷积神经网络的特征代替 HOG 这种人工设计的特征, 应用于可形变的组件模型 DPM, 该方法被称为 DP-DPM (Deep pyramid DPM). 由于可形变的组件模型本身的复杂性, 不能很好嵌入卷积神经网络中, DP-DPM 采用截断训练的方式, 使用 AlexNet 前 5 层网络得到具有强鉴别力的特征, 然后把这些特征输入 DPM 中进行训练. DP-DPM 与传统 DPM 模型相比平均精度 mAP 有着大幅提升, 与 R-CNN 相比 mAP 相近, 速度却明显快于 R-CNN.

已有视觉目标识别方法通常依赖于含有大量标注图像的训练数据, 基于包围盒的图像标注方法通常代价昂贵并且具有主观性. 近来, 提出了仅依赖于图像级类别标注的弱监督卷积神经网络[31]. 该方法研究了 CNN 是否能够从仅标注目标信息而不标注目标位置的混杂图像场景中, 学习得到目标的定位模型. 该方法对全监督网络结构进行了改进, 构造了基于图像级别标注数据构造端对端的弱监督卷积神经网络结构. 该网络的特点是: 在输出端增加全局最大池化层来搜索最高得分的目标位置; 设计了对图像中多个目标建模的损失函数. 基于 PASCAL VOC2012 和 MS COCO 数据的大量实验表明该弱监督网络具有以下优点: 1) 能够输出精确的图像类别标记; 2) 能够预测目标的近似位置; 3) 与基于目标包围盒标注训练的方法相比可得到相近的结果.

对预处理部分进行了改进, 采用新的网络结构[10, 24, 29]、训练策略、提出有形变约束的池化层[32] 等改进方法. 在物体检测问题中, R-CNN 等算法依

赖于额外的候选区域检测过程. 在 SPP-net[10] 与 Fast R-CNN[24] 中, 虽然候选区域检测的计算是一个瓶颈问题, 通过共享整张图像的卷积层特征图, 物体检测时间已明显缩减. 在 Fast R-CNN 的基础上, Ren 等[29] 提出了候选框网络 (Region proposal network, RPN), 可以用于实时目标检测. 王晓刚等[33] 研究了如何不依赖于人工设计特征和滑动窗口来提取感兴趣的目标, 该方法同时求解了两个任务: 在图像中对感兴趣的目标快速定位; 基于定位的快速目标分割. 该方法提出一种联合学习框架, 在该框架下每一个任务由一个多层卷积神经网络进行求解, 两个网络合作来增强性能. 此外, Yan 等[34] 提出了面向视觉识别的层次深度卷积神经网络. 将卷积神经网络嵌入到一个两层分类结构中: 粗分类器和精细分类器. 首先得到基于部件的预训练, 然后由多项式 Logistic 损失正则项进行全局调优. 可选性精细分类器以及卷积神经网络参数的缩减使得层次深度卷积神经网络对于大规模视觉识别可伸缩. 实验设计了两层次的深度卷积神经网络, 得到了较高的识别率. Liu 等[35] 提出了稀疏卷积神经网络, 解决了卷积神经网络中需要大量的参数计算、计算复杂度高的问题; 通过使用稀疏分解, 有效地缩减参数冗余. 在 ILSVRC2012 数据进行实验, 缩减了 90 % 参数, 仅仅损失 1 % 的准确性.

随着待处理图像数据规模和场景复杂程度的增加, 卷积神经网络可以演化出各种图像理解模型. 如面向局部对应点匹配的模型 3DMatch[36] (局部小规模数据), 面向物体检测的模型 Deep sliding shapes[37] (物体级别中等规模数据), 和面向复杂场景理解的新颖计算模型 DeepContext[38] (场景级别大规模数据) 等.

此外, 许多工作将卷积神经网络应用在与图像分类和物体检测目标相近的新问题或应用上, 如细粒度识别[39]、图像属性检测、实例检索、医学影像检测[40], 并且取得了良好的效果, 卷积神经网络已成为许多视觉识别问题的首选.

已有的卷积神经网络在图像分类和物体检测领域取得一定的进展, 但仍面临许多的挑战: 1) 不断加深的深层神经网络保证了图像分类和物体检测的精度, 但也带来了巨大的计算压力, 如何快速精确地解决问题是一个不小的挑战; 2) 在对于多物体相互交叉或相互遮挡时, 大多数方法都不能很好地处理; 3) 运动模糊也会降低图像分类和物体检测的精度.

2.2 人脸识别和验证

人脸识别是指对输入图像的身份进行分类, 人脸验证是指区分一对图像是否属于同一身份 (可转化为一个二分类问题). 代表性的人脸识别方法包含 Eigenface、Fisherface 等子空间分析法[41], 通过比较人脸图像在低维空间的投影进行识别. 基于卷积神经网络的人脸识别方法使用了多层非线性特征变换进行识别, 通常可取得明显优于传统方法的实验结果[33].

DeepID 是一种用于人脸辨识的深度学习提取高层特征方法[42], 将深度卷积网络最后隐层神经元的输出作为 DeepID 特征. 在训练中区分 10 000 个类别的人脸并缩减特征抽取层的神经元数量, 深度卷积网络将由一小部分隐层神经元逐步形成顶层辨识相关特征. DeepID 在 LFW 数据上可得到 97.45 % 的识别率. 传统人脸识别分为 4 步: 人脸检测、配准、表示、分类; 在 DeepFace[43] 中, 配准和表示采用三维人脸模型用分段仿射变换得到, 分类器采用了一个 9 层神经网络. 该方法在包含 4 000 类的人脸图像数据库上进行训练, 在 LFW 数据得到 97.35 % 准确率, 在 YouTube 人脸数据 (YTF) 上能够缩减 50 % 错误率. 人脸识别的关键是提出有效的特征来缩减同一人的差异并增大不同人之间差异, 将人脸识别和人脸验证信号作为监督, 由设计的深度卷积网络可以学习到深度识别−验证特征 (DeepID2)[44]. 由于学习到的 DeepID2 特征对于不同的身份具有差异而对同一身份一致, 使得人脸识别更加容易. 由学习得到的特征表示以及人脸识别模型, 在 LFW 数据上达到了最高 99.15 % 的人脸识别准确性.

尽管基于卷积神经网络的人脸识别方法在 LFW 等测试数据上得到了较高的识别率, 但是与基于传统方法的人脸识别方法类似, 基于 CNN 的人脸识别方法仍存在许多挑战性的问题, 如面部特征点定位、人脸、姿态等对人脸识别效果的影响, 都是需要深入研究的问题[45].

2.3 场景的语义分割和深度恢复

场景的语义分割是指对于一幅图像中的每一个像素给出其所属于的场景类别[46−48], 场景深度恢复是基于彩色或灰度图像恢复每个像素对应深度值的问题, 两者实质都是对输入图像的每个像素进行分类或回归, 已有方法集中在如何同时考虑单个像素的预测以及场景蕴含的上下文约束.

场景语义分割的参数化方法通常学习不同标注区域的外观 (Appearance) 和结构关系 (Structural relationship)[49−52], 这些方法扩展性往往不够好, 对新数据需要重新训练[53]. 场景语义分割的非参数化方法通常学习测试数据和训练数据间的差异模型, 将训练数据的语义标注转化为测试数据的预测[54−56], 逐个像素 (Per-pixel) 或者超像素 (Super-pixel) 分割, 使得最终结果依赖于超像素分割和假设的结果.

Farabet 等[46] 使用多尺度的卷积神经网络对输入图像进行特征提取, 并结合超像素划分和条件随

机场 (Conditional random fields, CRF), 得到像素语义的类别. 这种方法减小了对人工设计特征的需求, 生成对于纹理、形状、上下文信息的有效表示. Pinheiro 等[57] 将场景分割与目标检测相结合, 联合训练两个目标. Mohan[58] 在卷积神经网络中加入了反卷积层, 从而实现了一种端对端的场景语义分割方法, 但是需要固定大小的输入. Long 等[59] 在此基础上提出一种端对端的全卷积神经网络 FCN (Fully convolutional networks), 将全连接层变为核大小为 1 的全卷积层, 使得 FCN 可以接受任意大小的输入. 该网络使用了池化–反卷积结构来保证输出图像和输入图像具有相同的大小, 通过融合低层特征和高层特征得到具有更多细节的分割结果. Zheng 等[60] 将条件随机场转变成为一种递归神经网络 (Recurrent neural networks, RNN) 网络层, 连接在 FCN 之后, 对 FCN 的结果进行平滑和优化, 得到细节更具体更平滑的分割效果.

Liu 等[53] 研究了一种半参数化人体服饰分割方法. 该方法不需要对数据进行预处理, 而且对新的标注数据 (如新类别数据) 具有良好的扩展性. 在基于 K 近邻的非参数方法框架下, 参数化的匹配卷积神经网络 (Matching convolutional neural network, MCNN) 根据 K 近邻图像标注的语义区域, 在测试图像上能够找到最佳匹配区域, 并预测出匹配的置信度和位置偏移. 具体来说, 提取出输入图像中的人体区域, 使用 K 近邻方法从训练库里找出 K 个近邻, 得到 K 个图像对, 每个图像对使用卷积神经网络学习两者之间的相似性和位置偏差, 融合所有的结果得到最终的划分结果. 最后采用超像素间平滑性等方法对划分结果进行平滑处理.

Eigen 等[61] 提出基于单幅图像的深度图和法向图恢复以及场景语义分割的 CNN 网络. 该网络由三部分构成: 第一部分网络提取图像特征, 第二部分网络得到低分辨率的预测结果, 第三部分网络得到高分辨率的预测结果, 第一部分和第二部分网络的结果经过上采样并结合卷积后的原始图像输入至第三部分网络. 对于深度图和法向图恢复任务, 同时恢复深度图和法向图的效果优于分别恢复这两个任务, 完成对深度图和法向图的预测后, 输入原始图像、深度图和法向图即可进行语义分割. 对于语义分割, 基于多通道输入 (彩色图像、深度图和法向图) 的预测结果优于基于单通道 (彩色图像) 输入得到的结果.

Liu 等[62] 研究了将 CNN 和条件随机场 CRF 结合起来从单目图像预测深度的方法. 网络前面几层网络用 CNN 提取特征, 然后用条件随机场计算网络的损失并反馈至前面几层网络. 将输入图像划分为超像素, 以每个超像素为中心的一个图像块作为 CNN 的输入, 得到预测的深度. 每组相邻的超像素对 (S_p, S_q) 计算 K 个相似度度量 (S_{pq1}, S_{pq2}, $\cdots$, S_{pqk}), 将这些相似度度量输入一个全连接层得到相似度值 R_{pq}; 将所有超像素预测的深度值和所有的超像素对的相似度值 R_{pq} 输入 CRF 损失层计算网络损失, 该损失函数的优化可直接使用反向传播方法求解.

已有的卷积神经网络已经能在一定程度上解决图像语义分割和深度恢复的问题, 为了得到更加精细和准确的结果, 需要根据特定的问题设计出能够更好满足上下文约束的网络结构.

2.4　人体关节检测

本节主要介绍使用卷积神经网络从图像中估计人体关节位置的方法. 基于人体关节点位置, 可以较为容易地恢复人体姿态信息. 已有方法通常使用关节的空间位置约束、上下文约束来减少不合理的关节位置识别结果.

Tompson 等[63] 直接从深度图像学习人手关节位置进而重建模型 (网络结构见图 3), 将卷积神经网络应用于人手关节检测问题. 该方法提出了一个针对人手等链状物体的实时连续姿态估计系统, 通过可端对端训练的卷积神经网络得到人手关节点的二维位置, 然后结合原始的深度图像重构人手姿态. 该方法分为三个阶段: 随机森林分类器分割出人手区域; 卷积神经网络学习关节点位置; 使用逆向运动学 (Inverse kinematic, IK) 方法进行姿态恢复. Jain 等[64] 在 Tompson 等方法[63] 的基础上加入关节位置的空间约束, 提出一种新的人体姿态估计方法. 该方法结合了底层特征和高层弱结构模型, 由卷积神经网络来找出图像中身体关节的位置, 并加上身体各个部位的位置关系约束, 构成了一个对身体姿态的描述. 训练多个卷积神经网络, 每个卷积神经网络输出关节位置的空间分布概率图. 采用滑动窗口方法, 将图像输入卷积神经网络可得到各个关节的空间概率分布, 然后结合身体多个关节的位置约束优化关节位置, 该方法可减少不合理的关节位置. Oberweger 等[65] 对比了不同层数和结构的卷积神经网络对关节位置的识别效果, 在多尺度卷积神经网络、深的卷积神经网络、浅的卷积神经网络中加上三维姿态的先验信息, 可以显著提升关节位置识别效果. 该方法还通过加入关节点的上下文信息迭

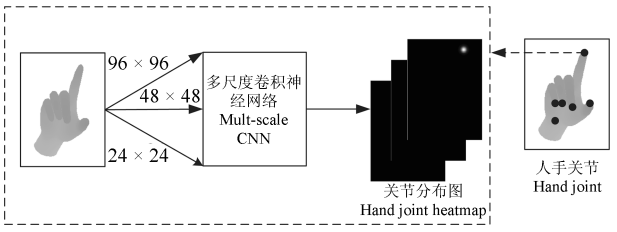


图 3　基于卷积神经网络的关节检测方法[63]

Fig. 3　Hand joint detection with convolutional neural networks[63]

代优化关节点的位置, 能使得关节估计结果更准确. 这类方法的主要挑战在于如何解决人体肢体自遮挡导致的关节检测误差以及如何提高检测精度.

3 总结和讨论

本文阐述了卷积神经网络在图像理解, 特别是图像分类、物体检测、人脸识别、语义图像分割等领域的研究进展与典型应用. 图像理解也推动了卷积神经网络在网络结构、训练方法等方面的完善. 卷积神经网络虽然在一些数据上, 如 ImageNet 上取得了成功, 但是如何针对实际特定问题、特定图像训练库设计更有效的网络结构, 融合问题先验信息、从理论和应用上评估网络性能等都是需要深入研究的问题. 我们觉得可能的研究方向有:

1) 卷积神经网络将卷积、池化与神经网络结合, 有效地利用了图像的结构信息. 进一步, 如何有效利用领域知识, 改进网络结构来获取视觉上的不变性值得引起关注;

2) 在理论上, 如何在算法中利用深度模型的选择性、稀疏性, 如何设计算法保证收敛性;

3) 目前, GoogLeNet, VGG 的网络结构已超过 20 层, 如何针对更大规模数据、更深结构网络设计高效的数值优化、并行计算方法和平台.

随着理论和应用的深入研究, 卷积神经网络在图像理解中将会得到更好的应用.

致谢

感谢张寅达博士、白延成博士、王文中博士的帮助和讨论, 感谢审稿人的宝贵意见以及 NVIDIA 提供的 Hardware Grant Program.

References

1 Rumelhart D E, Hinton G E, Williams R J. Learning representations by back-propagating errors. *Nature*, 1986, **323**(6088): 533−536

2 Vapnik V N. *Statistical Learning Theory*. New York: Wiley, 1998.

3 Wang Xiao-Gang. Deep learning in image recognition. *Communications of the CCF*, 2015, **11**(8): 15−23
(王晓刚. 图像识别中的深度学习. 中国计算机学会通讯, 2015, **11**(8): 15−23)

4 Hinton G E, Salakhutdinov R R. Reducing the dimensionality of data with neural networks. *Science*, 2006, **313**(5786): 504−507

5 Deng J, Dong W, Socher R, Li L J, Li K, Li F F. ImageNet: a large-scale hierarchical image database. In: Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition. Miami, FL: IEEE, 2009. 248−255

6 LeCun Y, Boser B, Denker J S, Henderson D, Howard R E, Hubbard W, Jackel L D. Backpropagation applied to handwritten zip code recognition. *Neural Computation*, 1989, **1**(4): 541−51

7 LeCun Y, Bottou L, Bengio Y, Haffner P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 1998, **86**(11): 2278−2324

8 Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks. In: Proceedings of Advances in Neural Information Processing Systems 25. Lake Tahoe, Nevada, USA: Curran Associates, Inc., 2012. 1097−1105

9 Girshick R, Donahue J, Darrell T, Malik J. Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus, USA: IEEE, 2014. 580−587

10 He K M, Zhang X Y, Ren S Q, Sun J. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2015, **37**(9):1904−1916

11 Szegedy C, Liu W, Jia Y Q, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V, Rabinovich A. Going deeper with convolutions. In: Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, MA: IEEE, 2015. 1−9

12 Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition [Online], available: http://arxiv.org/abs/1409.1556, May 16, 2016

13 Forsyth D A, Ponce J. *Computer Vision: A Modern Approach* (2nd Edition). Boston: Pearson Education, 2012.

14 Zhang Yu-Jin. *Image Engineering (Part 2): III-Image Understanding* (3rd Edition). Beijing: Tsinghua University Press, 2012.
(章毓晋. 图像工程 (下册): III-图像理解. 第 3 版. 北京: 清华大学出版社, 2012.)

15 He K M, Zhang X Y, Ren S Q, Sun J. Deep residual learning for image recognition [Online], available: http://arxiv.org/abs/1512.03385, May 3, 2016

16 LeCun Y, Bottou L, Bengio Y, Haffner P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 1998, **86**(11): 2278−324

17 Bouvrie J. Notes On Convolutional Neural Networks, MIT CBCL Tech Report, Cambridge, MA, 2006.

18 Duda R O, Hart P E, Stork D G [Author], Li Hong-Dong, Yao Tian-Xiang [Translator]. *Pattern Classification*. Beijing: China Machine Press, 2003.
(Duda R O, Hart P E, Stork DG [著], 李宏东, 姚天翔 [译]. 模式分类. 北京: 机械工业出版社, 2003.)

19 Lin M, Chen Q, Yan S C. Network in network. In: Proceedings of the 2014 International Conference on Learning Representations. Banff, Canada: Computational and Biological Learning Society, 2014.

20 Zeiler M D, Fergus R. Stochastic pooling for regularization of deep convolutional neural networks [Online], available: http://arxiv.org/abs/1301.3557, May 16, 2016

21 Maas A L, Hannun A Y, Ng A Y. Rectifier nonlinearities improve neural network acoustic models. In: Proceedings of ICML Workshop on Deep Learning for Audio, Speech, and Language Processing. Atlanta, USA: IMLS, 2013.

22 Ioffe S, Szegedy C. Batch normalization: accelerating deep network training by reducing internal covariate shift. In: Proceedings of the 32nd International Conference on Machine Learning. Lille, France: IMLS, 2015. 448−456

23 Felzenszwalb P, McAllester D, Ramanan D. A discriminatively trained, multiscale, deformable part model. In: Proceedings of the 2008 IEEE Conference on Computer Vision and Pattern Recognition. Anchorage, USA: IEEE, 2008. 1−8

24 Girshick R. Fast R-CNN. In: Proceedings of the 2015 IEEE International Conference on Computer Vision. Santiago, Chile: IEEE, 2015. 1440−1448

25 Girshick R, Iandola F, Darrell T, Malik J. Deformable part models are convolutional neural networks. In: Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, MA: IEEE, 2015. 437−446

26 Dalal N, Triggs B. Histograms of oriented gradients for human detection. In: Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. San Diego, CA, USA: IEEE, 2005. 886−893

27 Sermanet P, Eigen D, Zhang X, Mathieu M, Fergus R, LeCun Y. Overfeat: integrated recognition, localization and detection using convolutional networks [Online], available: http://arxiv.org/abs/1312.6229, May 16, 2016

28 Uijlings J R R, van de Sande K E A, Gevers T, Smeulders A W M. Selective search for object recognition. *International Journal of Computer Vision*, 2013, **104**(2): 154−171

29 Ren S, He K, Girshick R, Sun J. Faster R-CNN: towards real-time object detection with region proposal networks. In: Proceedings of Advances in Neural Information Processing Systems 28. Montréal, Canada: MIT, 2015. 91−99

30 Zeiler M D, Fergus R. Visualizing and understanding convolutional networks. In: Proceedings of the 13th European Conference on Computer Vision. Zurich, Switzerland: Springer, 2014. 818−833

31 Oquab M, Bottou L, Laptev I, Sivic J. Is object localization for free?-weakly-supervised learning with convolutional neural networks. In: Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA: IEEE, 2015. 685−694

32 Ouyang W L, Wang X G, Zeng X Y, Qiu S, Luo P, Tian Y L, Li H S, Yang S, Wang Z, Loy C C, Tang X O. Deepid-net: deformable deep convolutional neural networks for object detection. In: Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA: IEEE, 2015. 2403−2412

33 Wang Xiao-Gang, Sun Yi, Tang Xiao-Ou. From unified subspace analysis to joint deep learning: progress of face recognition in the last decade. *Communications of the CCF*, 2015, **11**(4): 8−14
(王晓刚, 孙祎, 汤晓鸥. 从统一子空间分析到联合深度学习: 人脸识别的十年历程. 中国计算机学会通讯, 2015, **11**(4): 8−14)

34 Yan Z C, Zhang H, Piramuthu R, Jagadeesh V, DeCoste D, Di W, Yu Y Z. HD-CNN: hierarchical deep convolutional neural networks for large scale visual recognition. In: Proceedings of the 2015 IEEE International Conference on Computer Vision. Boston, USA: IEEE, 2015. 2740−2748

35 Liu B Y, Wang M, Foroosh H, Tappen M, Pensky M. Sparse convolutional neural networks. In: Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA: IEEE, 2015. 806−814

36 Zeng A, Song S, Nießner M, Fisher M, Xiao J. 3DMatch: learning the matching of local 3D geometry in range scans [Online], available: http://arxiv.org/abs/1603.08182, August 11, 2016

37 Song S, Xiao J. Deep sliding shapes for amodal 3D object detection in RGB-D images. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE, 2016. 685−694

38 Zhang Y, Bai M, Kohli P, Izadi S, Xiao J. DeepContext: context-encoding neural pathways for 3D holistic scene understanding [Online], available: http://arxiv.org/abs/1603.04922, August 11, 2016

39 Zhang N, Donahue J, Girshick R, Darrell T. Part-based R-CNNs for fine-grained category detection. In: Proceedings of the 13th European Conference on Computer Vision. Zurich, Switzerland: Springer, 2014. 834−849

40 Shin H C, Roth H R, Gao M C, Lu L, Xu Z Y, Nogues I, Yao J H, Mollura D, Summers R M. Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. *IEEE Transactions on Medical Imaging*, 2016, **35**(5): 1285−1298

41 Belhumeur P N, Hespanha J P, Kriegman D J. Eigenfaces vs. fisherfaces: recognition using class specific linear projection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1997, **19**(7): 711−720

42 Sun Y, Wang X G, Tang X O. Deep learning face representation from predicting 10, 000 classes. In: Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus, USA: IEEE, 2014. 1891−1898

43 Taigman Y, Yang M, Ranzato M A, Wolf L. Deepface: closing the gap to human-level performance in face verification. In: Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus, USA: IEEE, 2014. 1701−1708

44 Sun Y, Wang Y H, Wang X G, Tang X O. Deep learning face representation by joint identification-verification. In: Proceedings of Advances in Neural Information Processing Systems 27. Montreal, Canada: Curran Associates, Inc., 2014. 1988−1996

45 Shan Shi-Guang, Kan Mei-Na, Li Shao-Xin, Zhang Jie, Chen Xi-Lin. Face image analysis and recognition with deep learning. *Communications of the CCF*, 2015, **11**(4): 15−21
(山世光, 阚美娜, 李绍欣, 张杰, 陈熙霖. 深度学习在人脸分析与识别中的应用. 中国计算机学会通讯, 2015, **11**(4): 15−21)

46 Farabet C, Couprie C, Najman L, LeCun Y. Learning hierarchical features for scene labeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2013, **35**(8): 1915−29

47 Yu Miao, Hu Zhan-Yi. Higher-order Markov random fields and their applications in scene understanding. *Acta Automatica Sinica*, 2015, **41**(7): 1213−1234
(余淼, 胡占义. 高阶马尔科夫随机场及其在场景理解中的应用. 自动化学报, 2015, **41**(7): 1213−1234)

48 Guo Ping, Qian Yin, Zhou Xiu-Ling. Image semantic analysis. Beijing: Science Press, 2015.
(郭平, 尹乾, 周秀玲. 图像语义分析. 北京: 科学出版社, 2015.)

49 Yamaguchi K, Kiapour M H, Ortiz L E, Berg T L. Parsing clothing in fashion photographs. In: Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition. Providence, RI: IEEE, 2012. 3570−3577

50 Liu S, Feng J S, Domokos C, Xu H, Huang J S, Hu Z Z, Yan S C. Fashion parsing with weak color-category labels. *IEEE Transactions on Multimedia*, 2014, **16**(1): 253−265

51 Dong J, Chen Q, Shen X H, Yang J C, Yan S C. Towards unified human parsing and pose estimation. In: Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus, OH: IEEE, 2014. 843−850

52 Dong J, Chen Q, Xia W, Huang Z Y, Yan S C. A deformable mixture parsing model with parselets. In: Proceedings of the 2013 IEEE International Conference on Computer Vision. Sydney, Australia: IEEE, 2013. 3408−3415

53 Liu S, Liang X D, Liu L Q, Shen X H, Yang J C, Xu C S, Lin L, Cao X C, Yan S C. Matching-CNN meets KNN: quasi-parametric human parsing. In: Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, MA: IEEE, 2015. 1419−1427

54 Yamaguchi K, Kiapour M H, Berg T L. Paper doll parsing: retrieving similar styles to parse clothing items. In: Proceedings of the 2013 IEEE International Conference on Computer Vision. Sydney, Australia: IEEE, 2013. 3519−3526

55 Liu C, Yuen J, Torralba A. Nonparametric scene parsing via label transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2011, **33**(12): 2368−2382

56 Tung F, Little J J. CollageParsing: nonparametric scene parsing by adaptive overlapping windows. In: Proceedings of the 13th European Conference on Computer Vision. Zurich, Switzerland: Springer, 2014. 511−525

57 Pinheiro P O, Collobert R, Dollar P. Learning to segment object candidates. In: Proceedings of Advances in Neural Information Processing Systems 28. Montréal, Canada: Curran Associates, Inc., 2015. 1981−1989

58 Mohan R. Deep deconvolutional networks for scene parsing [Online], available: http://arxiv.org/abs/1411.4101, May 3, 2016

59 Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation. In: Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, MA: IEEE, 2015. 3431−3440

60 Zheng S, Jayasumana S, Romera-Paredes B, Vineet V, Su Z Z, Du D L, Huang C, Torr P H S. Conditional random fields as recurrent neural networks. In: Proceedings of the 2015 IEEE International Conference on Computer Vision. Santiago, Chile: IEEE, 2015. 1529−1537

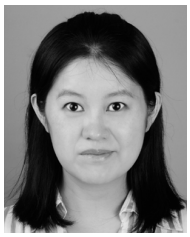
61 Eigen D, Fergus R. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In: Proceedings of the 2015 IEEE International Conference on Computer Vision. Santiago, Chile: IEEE, 2015. 2650−2658

62 Liu F Y, Shen C H, Lin G S. Deep convolutional neural fields for depth estimation from a single image. In: Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, MA: IEEE, 2015. 5162−5170

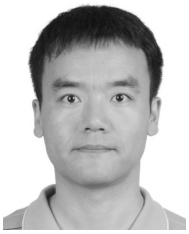
63 Tompson J, Stein M, Lecun Y, Perlin K. Real-time continuous pose recovery of human hands using convolutional networks. *ACM Transactions on Graphics (TOG)*, 2014, **33**(5): Article No. 169

64 Jain A, Tompson J, Andriluka M, Taylor G W, Bregler C. Learning human pose estimation features with convolutional networks. In: Proceedings of the 2014 International Conference on Learning Representations. Banff, Canada: Computational and Biological Learning Society, 2014. 1−14

65 Oberweger M, Wohlhart P, Lepetit V. Hands deep in deep learning for hand pose estimation. In: Proceedings of the 20th Computer Vision Winter Workshop (CVWW). Seggau, Austria, 2015. 21−30



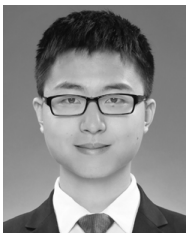
常 亮 北京师范大学信息科学与技术学院副教授. 主要研究方向为计算机视觉与机器学习.
E-mail: changliang@bnu.edu.cn
(**CHANG Liang** Associate professor at the College of Information Science and Technology, Beijing Normal University. Her research interest covers computer vision and machine learning.)



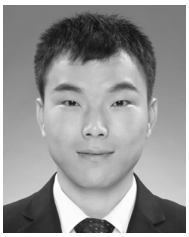
邓小明 中国科学院软件研究所副研究员. 主要研究方向为计算机视觉. 本文通信作者. E-mail: xiaoming@iscas.ac.cn
(**DENG Xiao-Ming** Associate professor at the Institute of Software, Chinese Academy of Sciences. His main research interest is computer vision. Corresponding author of this paper.)

周明全 北京师范大学信息科学与技术学院教授. 主要研究方向为计算机可视化技术, 虚拟现实.
E-mail: mqzhou@bnu.edu.cn
(**ZHOU Ming-Quan** Professor at the College of Information Science and Technology, Beijing Normal University. His research interest covers information visualization and virtual reality.)

武仲科 北京师范大学信息科学与技术学院教授. 主要研究方向为计算机图形学, 计算机辅助几何设计, 计算机动画, 虚拟现实. E-mail: zwu@bnu.edu.cn
(**WU Zhong-Ke** Professor at the College of Information Science and Technology, Beijing Normal University. His research interest covers computer graphics, computer-aided design, computer animation, and virtual reality.)



袁 野 中国科学院软件研究所硕士研究生. 主要研究方向为计算机视觉.
E-mail: yuanye13@mails.ucas.ac.cn
(**YUAN Ye** Master student at the Institute of Software, Chinese Academy of Sciences. His main research interest is computer vision.)



杨 硕 中国科学院软件研究所硕士研究生. 主要研究方向为计算机视觉.
E-mail: yangshuo114@mails.ucas.ac.cn
(**YANG Shuo** Master student at the Institute of Software, Chinese Academy of Sciences. His main research interest is computer vision.)

王宏安 中国科学院软件研究所研究员. 主要研究方向为实时智能, 自然人机交互. E-mail: hongan@iscas.ac.cn
(**WANG Hong-An** Professor at the Institute of Software, Chinese Academy of Sciences. His research interest covers real-time intelligence and natural human-computer interactions.)