

DOI: 10.19650/j.cnki.cjsi.J2006003

# 卷积神经网络在视觉图像检测的研究进展<sup>\*</sup>

蓝金辉,王 迪,申小盼

(北京科技大学自动化学院 北京 100083)

**摘 要:**视觉图像检测在机器视觉领域有着重要的研究意义和应用价值。近年来,卷积神经网络的发展带动了视觉图像检测领域的进步。大量新理论、新方法被应用于卷积神经网络,提高了网络对特征的表达能力,降低了网络的复杂性,增强了网络的性能。研究阐述了卷积神经网络的基本构成,从卷积层,池化层,激活函数,网络正则化和网络优化等方面总结了卷积神经网络近年来的改进方法,梳理了卷积神经网络在视觉图像检测领域的应用,总结了卷积神经网络在视觉图像检测领域的优点,并展望了未来的研究方向。

**关键词:** :卷积神经网络;深度学习;机器视觉;图像检测

**中图分类号:** TP183 TH744 **文献标识码:** :A **国家标准学科分类代码:** 520.20

## Research progress on visual image detection based on convolutional neural network

Lan Jinhui, Wang Di, Shen Xiaopan

(School of Automation and Electrical Engineering, University of Science and Technology Beijing, Beijing 100083, China)

**Abstract:** Visual image detection has great research significance and application value in the computer vision field. In recent years, the development of convolutional neural network (CNN) has led to the progress of visual image detection. A large number of new theories and new methods are applied to convolutional neural network, which improves the network feature expression ability, reduces the network complexity and improves the network performance. This paper presents the basic structure of Convolutional CNN, summarizes the improvements of CNN in recent years on different aspects, including convolutional layer, pooling layer, activation function, network regularization and network optimization, sorts various applications of CNN in visual image detection field and summarizes the advantages of CNN in visual image detection field, finally, prospects the future research direction.

**Keywords:** convolutional neural network (CNN); deep learning; computer vision; image detection

## 0 引 言

视觉图像是人类和机器感知世界的重要途径,被广泛应用于机器人<sup>[1]</sup>、智能交通<sup>[2]</sup>、工业<sup>[3]</sup>、医学<sup>[4]</sup>等领域的视觉图像检测技术一直是研究的热点。传统的视觉图像检测技术建立在人工设计特征的基础上,研究人员通过分析任务的应用场景,根据知识与经验设计提取图像的颜色、形状、纹理等特征,进而实现图像检测,整个过程复杂、低效。而随着深度学习(deep

learning)<sup>[5]</sup>理论的发展,极大地提高了卷积神经网络的特征提取能力,与传统方法相比,卷积神经网络自动学习、抽取蕴含大量信息的图像特征,使视觉图像检测取得了突破性的进展。

卷积神经网络是受到生物视觉系统的神经机制启发,针对二维图像的识别设计的一种模型。1962年,Hubel等<sup>[6-7]</sup>通过对猫的视觉皮层细胞的研究,提出了“感受野”(receptive field)的概念。1979年,Fukushima<sup>[8-10]</sup>在此基础上,提出了“神经认知机”(Neurocognitron),该模型被认为是卷积神经网络的前身。在1989年,Lecun等<sup>[11]</sup>提出了第一个真正意义上的卷积

神经网络 LeNet-5, 这个模型被用于美国邮政系统识别手写字符以分拣邮件包裹, 在 MNIST 数据集上达到了 99.2% 的分类正确率。此时卷积神经网络已经展现了在视觉图像领域的潜力, 但当时的算力限制了卷积神经网络的进一步发展, 仅适用于小尺寸图片的识别, 对大规模数据识别不佳。

近年来, 由于大规模标注数据集如 ImageNet<sup>[12]</sup>、PASCAL VOC<sup>[13]</sup>、MS COCO<sup>[14]</sup>、VisDrone2018<sup>[15]</sup> 等数据集和高性能图像处理器 (GPU) 的出现, 为高性能的卷积神经网络提供了可能。在 2012 年, 卷积神经网络取得了历史性的突破, Krizhevsky 等<sup>[16]</sup> 提出了 AlexNet 卷积神经网络, 并使用 GPU 训练网络模型, 在 ImageNet 大规模视觉识别挑战赛 (ImageNet large scale visual recognition challenge, LSVRC) 中取得了当时最佳的分类效果, Top-5 分类精度达到 84.7%, 比第 2 名要高出大约 10%, 充分展现了卷积神经网络强大的特征提取能力, 体现出卷积神经网络在视觉图像检测领域的巨大潜力。2014 年, Girshick 等<sup>[17]</sup> 提出了 R-CNN (region-based convolutional neural network) 方法, 第一次成功地将卷积神经网络应用于图像检测领域, 取得了突破性的进展, 在 PASCAL VOC2010 数据集上检测平均精度 (mean average precision, mAP) 达到 53.7%, 而当时流行的可变性组件模型 (deformable part model, DPM)<sup>[18]</sup> 的检测平均精度仅为 33.4%。

之后, 各种关于卷积神经网络的新理论、新算法、新模型层出不穷, 应用于视觉图像检测的卷积神经网络的性能不断提高。

## 1 视觉图像检测的难点与挑战

视觉图像检测一直是机器视觉领域的研究热点。视觉图像检测的难点与挑战如下:

1) 环境因素影响。光照、遮挡、尺度、视角、背景干扰等环境因素均会对目标的特征产生巨大干扰, 给图像的检测与识别带来巨大困难。

2) 类间的相似。不同类型的物体相似性很高, 特征很相似, 难以提取具有很高区分度的特征。如图 1(a) 所示, 狼和哈士奇在外观上非常相似。

3) 类内的差异。相同类型的物体类内差异很大, 特征表现的截然不同, 难以提取具有强鲁棒性的特征。如图 1(b) 所示, 虽然都是花瓶, 却在形状、颜色、纹理等方面都有很大的差异。

4) 海量的数据。互联网时代产生了大量的视觉图像数据, 而真实世界的视觉图像数据又是无穷无尽的, 数据的标注和处理需要耗费大量人力物力。海量数据要求图像检测算法简单、高效。



图1 视觉图像检测难点的例子

Fig.1 Examples of the difficulties in visual image detection

## 2 卷积神经网络

卷积神经网络是针对二维视觉图像设计的模型, 有很多变体, 但它们的基本组成却很相似, 一般包括卷积层、池化层、激活函数、全连接层和损失函数。

### 2.1 卷积层 (convolutional layer)

卷积是卷积神经网络提取特征的关键与核心。在图像领域, 卷积是一种常用的线性滤波的方法, 如经典的索贝尔算子, 如图 2 所示, 可以提取图像的边缘信息, 是提取图像特征的重要方法。但这种人类精心设计的卷积有很大的局限性: 一是提取的特征有限, 二是需要有丰富图像处理经验。

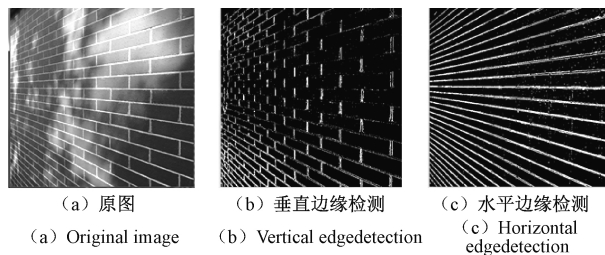


图2 索贝尔算子边缘检测

Fig.2 Edge detection with Sobel operator

卷积神经网络中, 一个卷积层可以包含多个卷积特征图, 每个卷积特征图根据输入、卷积核和激活函数来计算。假设输入是一张大小为  $M \times N$  的图像, 用矩阵  $\mathbf{x}$  表示, 卷积核是大小为  $m \times n$  的矩阵  $\mathbf{w}$ , 偏置为  $b$ , 特征图  $h$  的计算如图 3 所示, 或者表示为:

$$h = f(\mathbf{x} * \mathbf{w} + b) \quad (1)$$

输入多幅图像时, 采用多组卷积核和偏置, 每组产生一个卷积特征图, 把多个卷积特征图按顺序拼接, 得到卷积层, 如图 4 所示。

可以看出, 卷积是一种通过一定大小的卷积核作用于局部图像区域从而获得图像局部信息的一种操作。在卷积神经网络中, 卷积起着提取图像特征的作用, 是卷积

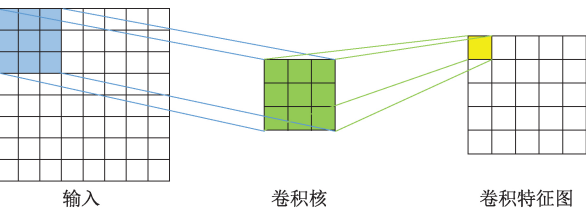


图 3 卷积  
Fig.3 Convolution

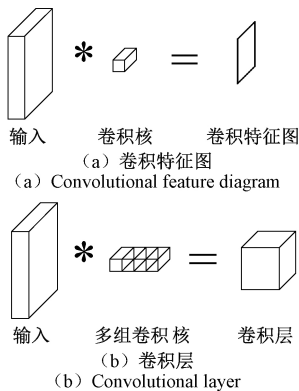


图 4 卷积特征图与卷积层  
Fig.4 Convolutional feature and convolutional layer  
schematic diagram

神经网络最根本的操作。

2.2 池化层

池化提高了卷积神经网络提取特征的稳定性、降低了特征向量的维数、并在一定程度上降低了过拟合。

池化层一般在卷积层后,同样由多个特征图组成,每个特征图一一对应于上一层的一个特征图。所以池化层不改变特征图的数量。一般的池化操作有最大值池化和平均值池化两种。最大值池化是将池化核覆盖的部分区域像素值取最大值,如式(2)所示:

$$y_{\max} = \max(x_1, x_2, \dots, x_l) \tag{2}$$

图 5 所示为 3×3 大小,步长为 1 的最大值池化操作示例。平均值池化是将池化核覆盖的部分区域的像素值相加然后再平均,如式(3)所示:

$$y_{ave} = (x_1 + x_2 + \dots + x_l) / l \tag{3}$$

2.3 激活函数

激活函数加强了卷积神经网络对提取特征的表达,没有激活函数,卷积层和池化层的堆叠只能是简单的线性映射,无法形成复杂的函数。

激活函数又称非线性映射,模拟了生物神经元特性:生物神经元通常有一个阈值,当神经元所获得的输入信号累计到超过该阈值时,神经元就被激活,由抑制状态转为兴奋状态。

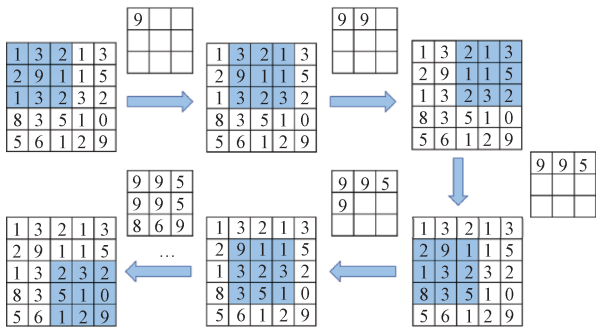


图 5 最大值池化  
Fig.5 Max pooling

经典的激活函数有 sigmoid 函数:

$$\text{sigmoid}(x) = \frac{1}{1 + e^{-x}} \tag{4}$$

函数形状如图 6 所示,可以看出经过 sigmoid 函数作用后,输出被压缩到[0,1]之间,0 对应了生物神经元的抑制状态,1 则对应生物神经元的兴奋状态。

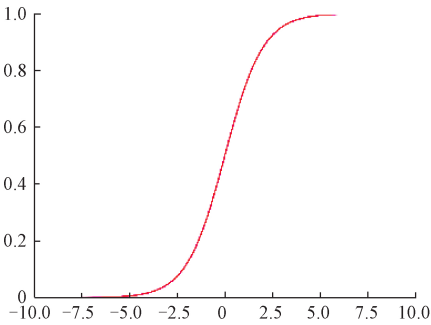


图 6 sigmoid 函数  
Fig.6 Sigmoid function

2.4 全连接层

在卷积神经网络中,经过多个卷积层和池化层后,会有一个或多个与之前层所有神经元都连接的全连接层。全连接层将前面卷积层和池化层学习到的特征映射到样本标记空间,在卷积神经网络中起到分类器的作用。

2.5 损失函数

卷积神经网络的训练是一个全局优化问题,通过构造衡量预测值和真实样本标记之间误差的损失函数,达到设定合适网络参数的目标,其形式为:

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

式中:  $\theta$  为卷积神经网络的所有参数;  $y^{(n)}$  为真实样本标记;  $o^{(n)}$  为卷积神经网络输出。

3 卷积神经网络的改进

自 2012 年, AlexNet 网络模型获得了巨大的成功后, 卷积神经网络成为了一个研究热点, 研究人员提出了许多新理论, 新方法提高了卷积神经网络的性能。本节主要从卷积层, 池化层, 激活函数, 正则化和优化等 5 个方面描述卷积神经网络的改进, 如图 7 所示。

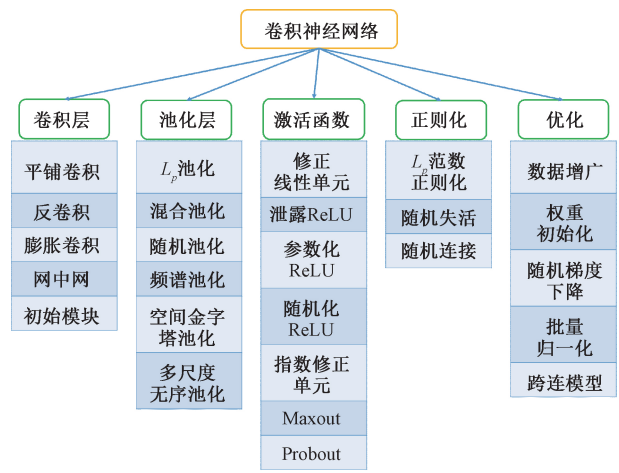


图 7 卷积神经网络改进的层次化结构

Fig.7 The hierarchical structure improved with CNN

3.1 卷积层

1) 平铺卷积

传统的卷积神经网络依赖于权重共享, 如图 8(a) 所示, 卷积层的每个单元共享同样的权重, 权重共享减少了卷积神经网络的参数, 提高了网络的计算效率。但权重共享也使得卷积神经网络不具备更多复杂的不变性。Ngiam 等<sup>[19]</sup>提出了平铺卷积神经网络(tiled CNN), 通过使相邻单元间的权重不相同, 距离为  $k$  的单元权重相同, 如图 8(b) 所示。显然当  $k=1$  时, tiled CNN 转换成传统的卷积神经网络。在 CIFAR-10 数据集上的实验结果显示, 当  $k=2$  时, tiled CNN 达到了最好的效果。

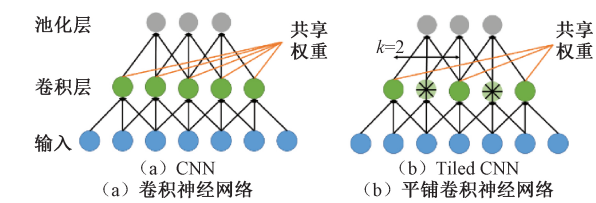


图 8 CNN 和 tiled CNN 对比

Fig.8 The comparison between CNN and tiled CNN

2) 反卷积

与传统卷积网络相比, 反卷积可以被看作传统卷积

的逆运算。如图 9(a) 所示为输入为  $6 \times 6$ , 经过卷积核为  $3 \times 3$ 、步长为 1、无填充的卷积运算后, 卷积特征图大小为  $2 \times 2$  的卷积操作。图 9(b) 所示为输入为  $2 \times 2$ , 经过卷积核为  $3 \times 3$ , 步长为 1, 填充为 2 的反卷积运算。Zelier 等<sup>[20]</sup>在 2010 年提出反卷积的概念, 并在 2014 年将反卷积应用于对 AlexNet 模型内部的可视化<sup>[21]</sup>, 使得人们对卷积神经网络的机制有了一定的认识。

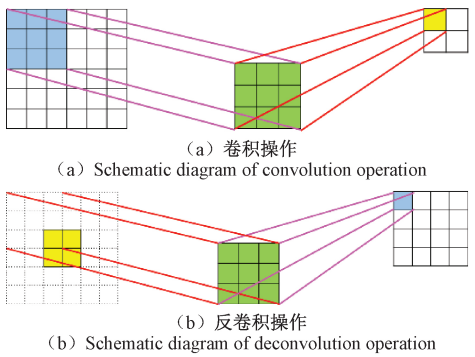


图 9 卷积和反卷积对比

Fig.9 The comparison between convolution and deconvolution

3) 膨胀卷积

Fisher 等<sup>[22]</sup>提出了膨胀卷积。与普通的卷积相比, 膨胀卷积还有一个膨胀因子, 主要用于表示膨胀的大小。膨胀卷积提高了感受野的大小。感受野是指卷积神经网络每一层输出的特征图上的像素点在输入图片上映射的区域大小, 感受野越大包含越多的上下文关系。通过控制膨胀因子可以扩大感受野。如图 10(a) 所示为一个  $3 \times 3$  的卷积核感受野, 以 2 倍膨胀因子膨胀后感受野变为  $7 \times 7$ , 如图 10(b) 所示, 以 4 倍膨胀因子膨胀后感受野变为  $15 \times 15$ , 如图 10(c) 所示。

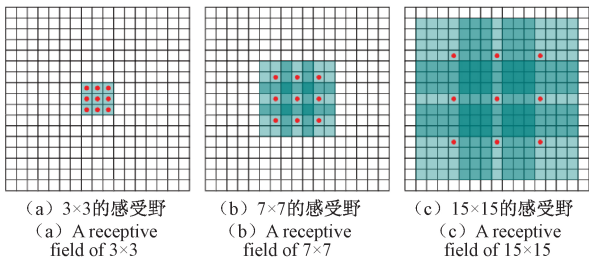


图 10 膨胀卷积

Fig.10 Dilated convolution

4) 网中网

网中网(network in network, NIN)是由 Lin 等<sup>[23]</sup>提出的一种经典网络模型。网中网使用多层全连接层和非线性函数的组合作为多层感知机层替代了传统卷积神经网络中的线性卷积层, 如图 11 所示。相比线性卷积层只



能将输入进行“简单”的线性组合形成下层特征,多层感知机层增加了网络的非线性能力,使得上层特征可以更复杂的映射到下层。

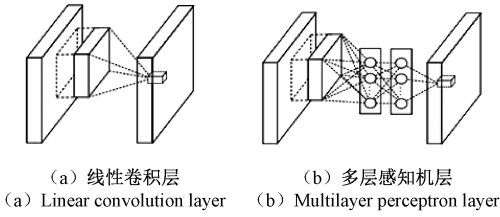


图 11 线性卷积层和多层感知机层对比

Fig.11 The comparison of linear convolution layer and multilayer perceptron layer

### 5) 初始模块

提高卷积神经网络性能最直接的办法是增加网络的深度(层的数量)或网络是宽度(每一层的单元数量)。但这种方法有两个主要的缺点:一是网络深度或宽度的增加意味着更大的参数,二是网络深度或宽度的增加对计算资源要求更高。2014年, Szegedv 等<sup>[24]</sup>提出了一个叫初始模块的卷积神经网络卷积层,并在当年的 LSVRC 竞赛中拿到了冠军。初始模块的基本结构如图 12 所示,不同于传统的卷积神经网络,模块中同一层内插入  $1 \times 1$ ,  $3 \times 3$ ,  $5 \times 5$  不同尺寸的卷积核,最后拼接在一起使得不同尺度的特征图相互融合。通过该网络的设计,在增加网络深度和宽度的同时,保持计算量不增加,提高了计算资源的利用率。

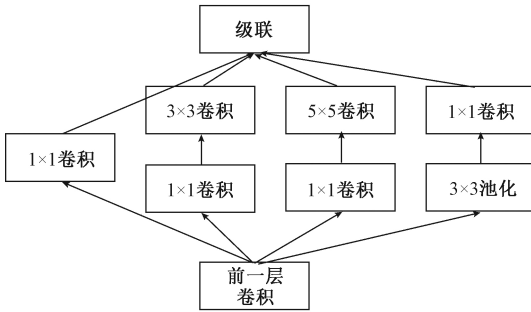


图 12 初始模块结构

Fig.12 Architecture of Inception module

在 Szegedv 等<sup>[25]</sup>的最新研究中,受到残差网络(residual network, ResNet)<sup>[26]</sup>中残差模块启发,组合了跨连模型和初始模块提出了初始-残差网络模块(inception-ResNet module),两者的模型如图 13 所示。Inception-ResNet 有效提升了网络的训练速度。

### 3.2 池化层

#### 1) $L_p$ 池化

Brune 等<sup>[27]</sup>为使系统具有不变性,采用对节点组的

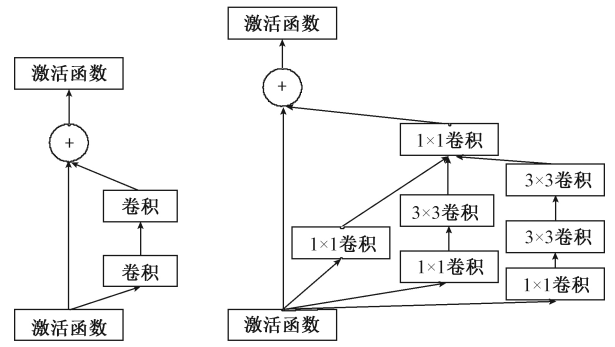


图 13 残差模块和初始-残差网络模块

Fig.13 The comparison of residual module and inception-ResNet module

输出进行汇总统计的方法,并对模板的扰动进行编码,这种方法称为  $L_p$  池化。经过训练之后,系统中组在一起的节点所对应的权重矩阵的列具有相似的特征。 $L_p$  池化可以表示为:

$$y_{lp} = [(x_1)^p + (x_2)^p + \dots + (x_i)^p]^{\frac{1}{p}} \quad (6)$$

式中:  $y_{lp}$  为特征图中  $L_p$  池化运算的输出;  $x_i$  为特征图中池化核各个位置  $R_j$  的特征值。当  $p = 1$  时,  $L_p$  池化相当于平均值池化;当  $p$  趋于无穷时,  $L_p$  相当于最大值池化。 $L_p$  池化比最大值池化具有更好的泛化效果。

#### 2) 混合池化

Yu 等<sup>[28]</sup>提出了将最大池化和平均池化相结合的混合池化方法。混合池化就是在最大值池化和平均值池化中进行随机选择。混合池化可以用公式表述为:

$$y_{mix} = \lambda y_{max} + (1 - \lambda) y_{ave} \quad (7)$$

当  $\lambda = 0$  时,表示平均值池化;当  $\lambda = 1$  时,表示最大值池化。在正向传播过程中,  $\lambda$  会被记录并将用于反向传播操作。实验表明,混合池化能较好地解决过拟合问题,其性能优于最大值池化和平均值池化。

#### 3) 随机池化

Zeiler 等<sup>[29]</sup>提出了随机池化的方法,对特征图中的元素按照其概率值大小随机选择,元素被选中的概率与其数值大小正相关。

首先通过对区域内的激活进行归一化,计算池化核各个位置  $R_j$  的概率  $p$ 。

$$p_i = \frac{x_i}{\sum_k x_k} \quad (8)$$

然后,从基于  $p$  的多项分布中取样:

$$y_s = x_i \quad (9)$$

其中,  $i$  服从  $P(p_1, \dots, p_k)$  分布。

随机池化的过程如图 14 所示。首先根据输入得到分布概率,再根据概率确定最终的随机池化结果。随机池化确保了非最大的激活也会被利用,同时避免了过拟合。

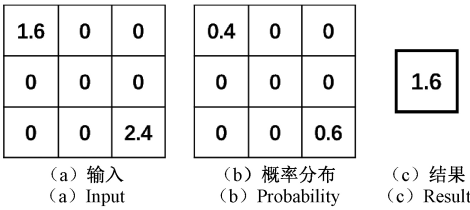


图 14 随机池化图

Fig.14 Stochastic pooling diagram

随机池化可以在不过度拟合的情况下训练更大的模型,并在识别任务中产生优异的性能。

#### 4) 频谱池化

Snoek 等<sup>[30]</sup>提出了频谱池化的概念。假设给定输入  $x \in \mathbf{R}^{M \times N}$  和期望输出特征图维度为  $H \times W$ 。首先,将输入通过离散傅里叶变换到频域,然后,只保留频率的中心子矩阵  $H \times W$ ,最后将这个中心矩阵映射回空间域。频谱池化与其他池化策略相比,每个参数上保存的信息更为丰富,并且可以灵活地选择池化输出维度。由于频谱池化的过程是通过矩阵截断实现的,在卷积神经网络中的计算成本较少,能够显著加快训练的收敛速度。

#### 5) 空间金字塔池化

一般的卷积神经网络的全连接层需要固定的输入维度,因此输入图像的大小也是固定的。但卷积操作没有对图像尺度有限制,为使卷积神经网络可以处理任意大小的输入图片,让模型变得更加的灵活。He 等<sup>[31]</sup>提出了空间金字塔池化方法,先让图像进行卷积操作,然后转化为维度相同的特征输入到全连接层,从而把卷积神经网络扩展到适用于任意大小的图像,获取图像的多尺度信息,同时避免裁剪和变形操作导致图像特征信息的丢失问题。

#### 6) 多尺度无序池化

为了改善卷积神经网络的不变性,同时又不降低识别能力,Gong 等<sup>[32]</sup>提出了多尺度无序池化(multi-scale orderless pooling, MOP)方法。在不同尺度上分别对其进行无序的池化操作,将不同尺度上的池化之后的特征连接起来形成最终的特征,提取整个图像的深度激活特征,与单尺度相比,多尺度无序池化效果更好。

### 3.3 激活函数

#### 1) 修正线性单元(rectified linear unit, ReLU)

Nair 等<sup>[33]</sup>提出的 ReLU 激活函数是一个著名的非饱和和激活函数。ReLU 激活函数定义为:

$$\text{ReLU}(x) = \max(0, x) \quad (10)$$

如图 15(a)所示,ReLU 激活函数是一种分段线性函数,将小于 0 的输入置 0,保留大于 0 的输入部分。ReLU 激活函数中的  $\max()$  操作使得其比 sigmoid 激活函数有快的多的速度,并且还在隐藏单元中引入稀疏性,使得网

络容易获得获取稀疏表示。尽管 ReLU 激活函数在 0 处的不连续性可能会影响反向传播的性能,但许多实验表明,ReLU 激活函数比 sigmoid 激活函数更好。

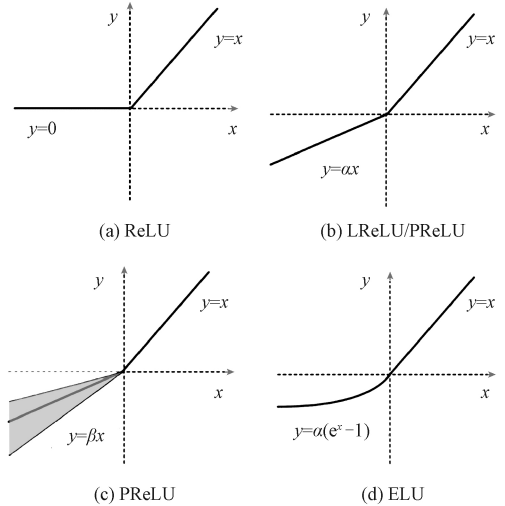


图 15 ReLU, LReLU, PReLU, ELU 对比

Fig.15 The comparison among ReLU, LReLU, PReLU and ELU

#### 2) 泄露 ReLU(leaky ReLU, LReLU)

ReLU 激活函数的一个缺点是,当函数没有被激活时,函数的梯度为 0,这可能导致不活跃的单元,从未被激活,因此,基于梯度的优化将不会调整网络的权重,并且减慢训练速度。为了解决整个问题,Maas 等<sup>[34]</sup>提出了 LReLU 激活函数,如图 15(b)所示,表达式为:

$$\text{LReLU}(x) = \begin{cases} x, & x \geq 0 \\ ax, & x < 0 \end{cases} \quad (11)$$

式中: $a$  是 0~1 之间的一个固定值。与 ReLU 相比, LReLU 压缩了负值而不是直接将负值的输入置 0,这使得单元在不激活时具有较小的非零梯度。

#### 3) 参数化 ReLU

He 等<sup>[35]</sup>提出了参数优化 PReLU(parametric ReLU, PReLU)激活函数,如图 15(b)所示,定义为:

$$\text{PReLU}(x) = \begin{cases} x, & x \geq 0 \\ ax, & x < 0 \end{cases} \quad (12)$$

式中: $a$  是一个  $\leq 1$  的可调参数,具体值需要通过学习得到。PReLU 通过将  $a$  作为一个网络中可学习的变量融入模型的整体训练过程,很好地解决了 LReLU 中  $a$  不易设定的问题。

#### 4) 随机化 ReLU

Xu 等<sup>[36]</sup>提出另一种解决  $a$  超参设定的方法:随机化 RReLU(randomized ReLU, RReLU)激活函数,如图 15(c)所示。RReLU 的表达式如下:

$$\text{RReLU}(x) = \begin{cases} x, & x \geq 0 \\ \beta x, & x < 0 \end{cases} \quad (13)$$

式中: $\beta$ 的取值在训练阶段服从均匀分布,测试阶段则将其指定为该均匀分布对应的分布期望。

#### 5) 指数修正单元

Clevert 等<sup>[37]</sup>提出了指数修正单元(exponential linear unit, ELU)激活函数,如图 15(d)所示,表达式如下:

$$\text{ELU}(x) = \begin{cases} x, & x \geq 0 \\ a(e^x - 1), & x < 0 \end{cases} \quad (14)$$

式中: $a$ 是用于控制对于负输入的 ELU 饱和的值的预定参数。与含有不饱和负部分的 LReLU、PReLU 和 RReLU 相比,ELU 采用饱和函数作为负部分。由于饱和函数会减少单元在失活时的变化,所以 ELU 对噪声的鲁棒性更强。

#### 6) Maxout

Goodfellow 等<sup>[38]</sup>提出 Maxout 激活函数,Maxout 激活函数取该层的最大值,其定义为:

$$\text{Maxout}(x) = \max(x_1, x_2, \dots, x_n) \quad (15)$$

ReLU 激活函数实际上是 Maxout 激活函数的特例,并且实验证明,Maxout 激活函数特别适用于有随机失活的网络。

#### 7) Probout

Springenberg 等<sup>[39]</sup>在 Maxout 激活函数的基础上提出了 Probout 激活函数:用概率抽样的方式代替 Maxout 函数的最大值运算。首先定义  $k$  个线性单元每一个的概率:

$$p_i = \frac{e^{\lambda x_i}}{\sum_{j=1}^k e^{\lambda x_j}} \quad (16)$$

式中: $\lambda$ 是一个控制分布的超参数,然后根据多项式分布  $\{p_1, \dots, p_k\}$  挑选  $k$  个线性单元中的一项,设定每个 Probout 激活函数定义为:

$$\text{Probout}(x) = x_i \quad (17)$$

式中: $i$ 服从多项式分布  $\{p_1, \dots, p_k\}$ 。Probout 激活函数比 Maxout 激活函数的性能更好,但是因为增加了概率计算,需要更多的计算资源。

### 3.4 正则化

若某卷积神经网络在训练集上表现良好,同时在测试集上表现也不错,则可以称该卷积神经网络具有良好的“泛化性”。若某卷积神经网络在训练集上表现良好,但在测试集上表现糟糕,这种现象被称为“过拟合”。为了防止过拟合现象,通常使用“正则化”。

#### 1) $l_p$ 范数正则化

$l_p$  范数正则化通过添加影响模型复杂性的附加项来修改损失函数,正则损失是:

$$E(\theta, x, y) = L(\theta, x, y) + \lambda R(\theta) \quad (18)$$

式中: $L(\theta, x, y)$ 是损失函数; $R(\theta)$ 是正则化项; $\lambda$ 控制正则化项的大小,较大的  $\lambda$  取值将较大幅度的约束模型的

复杂度。

文献[40]总结对比了几种  $l_p$  范数正则化方法,并对比了其各自特点。 $R(\theta)$ 通常为  $\sum_j \|\theta_j\|_p^p$ ,当  $p \geq 1$  时,  $l_p$ -norm 是凸函数,优化更加简单。当  $p < 1$  时,  $l_p$ -norm 更多地利用了权值的稀疏效应,而对非凸函数则更有指导意义。

#### 2) 随机失活

随机失活首先由 Hinton 等<sup>[41]</sup>提出,这是一个有效的避免过拟合的方法。随机失活的原理如图 16 所示,在训练阶段将神经元以概率  $p$  随机将该神经元权重重置 0,故称随机失活,测试阶段所有神经元呈激活态,但其权重乘以  $(1-p)$  以保证训练和测试阶段权重拥有相同的期望。

他们将随机失活应用于全连接层,随机失活的数学表达如下:

$$y = \mathbf{r} * \alpha(\mathbf{W}^T \mathbf{x}) \quad (19)$$

其中  $\mathbf{x} = [x_1, x_2, \dots, x_n]^T$  是全连接层的输入; $\mathbf{W} \in \mathbf{R}^{n \times d}$  是一个权重矩阵; $\mathbf{r}$  是一个尺寸为  $d$  的二进制向量,服从参数为  $p$  伯努利分布,  $r_i \sim B(p)$ 。

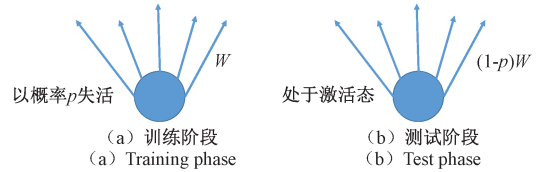


图 16 随机失活

Fig.16 Dropout

在神经网络中,对于某单个神经元来说,其反向传导来的梯度信息同时受到其他神经元的影响,这种现象被称为“复杂协同适应”效应。随机失活的提出在一定程度上缓解了神经元之间复杂协同适应效应,降低了神经元之间的依赖程度,避免了网络过拟合的发生。

#### 3) 随机连接

受到随机失活的启发,Wan 等<sup>[42]</sup>提出了随机连接,数学表达如下:

$$y = \alpha((\mathbf{R} * \mathbf{W}^T) \mathbf{x}) \quad (20)$$

其中  $R_{ij} \sim B(p)$ 。

随机失活随机将神经元的输出设为 0。随机连接随机设置权重矩阵  $\mathbf{W}$  的元素为 0。随机失活和随机连接的对比如图 17 所示。

### 3.5 优化

#### 1) 数据增广

深度卷积神经网络特别依赖大规模训练数据集。提高卷积神经网络性能的方法之一是数据增广,数据增广是指在不改变数据性质的前提下将已有的数据转换成新

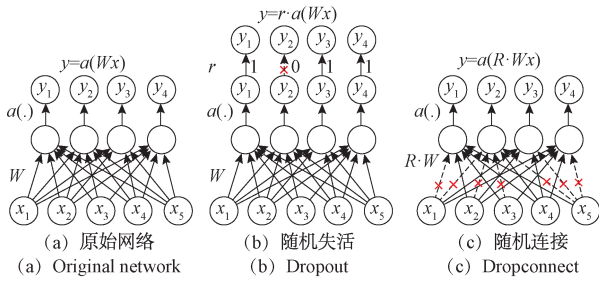


图 17 原始网络和使用随机失活,随机连接的对比

Fig.17 The comparison of original network, Dropout and Dropconnect

数据,有效的数据增广不仅能增加训练样本的数量,还能增加训练样本的多样性,这使得卷积神经网络性能提升的同时,还有效避免了过拟合现象。流行的数据增广方法包括采样<sup>[12]</sup>、镜像<sup>[43]</sup>、旋转<sup>[44]</sup>、平移<sup>[45]</sup>和各种亮度变换<sup>[46]</sup>。

Yang<sup>[47]</sup>提出了一种利用图像标记信息的监督式新型数据增广方式。具体而言,首先根据原数据训练一个分类的初始模型,然后利用该模型,对每张图生成对应可指示图像区域与场景标记间的相关概率的热力图。之后,可根据此概率映射回原图选择较强相关的区域作为扣取的图像块。由于一开始利用图像标记训练了一个初始分类模型,因此这样的过程被称为“监督式数据增广”。监督式数据增广适用于高层语义图像分类任务。

## 2) 权重初始化

深度卷积神经网络有大量的参数并且其损失函数是非凸的<sup>[48]</sup>,因此深度卷积神经网络难以训练。为了实现训练过程中的快速收敛并且避免梯度消失问题,采用一个恰当的权重初始化方法十分重要。

AlexNet 卷积神经网络通过将参数值随机设定为服从标准差为 0.01,均值为 0 的高斯分布的随机数初始化网络,这种方法称之为高斯初始化方法。Glorot 等<sup>[49]</sup>将方差改为  $\frac{2}{n_{in} + n_{out}}$ 。其中  $n_{in}$  是输入的神经元数目, $n_{out}$

是输出结果的神经元数目。这种方法称之为“Xavier 参数初始化方法”。针对 Xavier 方法没有考虑非线性映射函数的对输入的影响,He 等<sup>[35]</sup>扩展了 Xavier 方法,将非线性映射造成的影响考虑到参数初始化中。Saxe 等<sup>[50]</sup>证明了利用正交矩阵初始化网络要比高斯初始化好。Minshkin 等<sup>[51]</sup>通过迭代扩展了 Saxe 的方法,这种方法能够有效消除层与层之间的关联,从而提高神经网络的性能。

## 3) 随机梯度下降

反向传播算法是一个用梯度下降来更新参数的标准的训练方法。标准的随机梯度下降算法<sup>[52]</sup>更新损失函

数  $L(\theta)$  的  $\theta$  参数,当:

$$\theta_{i+1} = \theta_i - \eta \nabla_{\theta} E(L(\theta_i)) \quad (21)$$

式中:  $E(L(\theta_i))$  是函数  $L(\theta_i)$  在全训练集上的期望; $\eta$  是学习率;随机梯度下降算法通过从训练集中挑选样例  $(x^{(i)}, y^{(i)})$  估计梯度为:

$$\theta_{i+1} = \theta_i - \eta \nabla_{\theta} L(\theta_i; x^{(i)}, y^{(i)}) \quad (22)$$

实际上,在随机梯度下降中每个参数的更新是针对与单个样例相反的小批次来计算的。这将有助于降低参数更新的方差,并会导致更稳定的收敛。收敛速度由学习率控制。然而,小批次的随机梯度下降不保证好的收敛,仍有一些问题需要被解决。首先,选择学习率不是一个简单的问题。通常的做法在初始阶段使用稳定学习率以达到稳定的收敛速度,然后随着收敛速度的减慢而降低学习率。文献[53-54]提出学习率时间表来在训练过程中调整学习率。为了让现有的梯度更新依赖于以前批次,并加速训练,Qian 等<sup>[55]</sup>提出在相关方向上积累速度矢量。经典的 Momentum 更新可以用下式表示:

$$v_{i+1} = \gamma v_i - \eta \nabla_{\theta} L(\theta_i; x^{(i)}, y^{(i)}) \quad (23)$$

$$\theta_{i+1} = \theta_i + v_{i+1} \quad (24)$$

式中:  $v_{i+1}$  是一个速度向量; $\gamma$  是一个动向量,通常设置为 0.9。

Nesterov 动量<sup>[56]</sup>是在梯度下降优化中利用动量的另一种方法:

$$v_{i+1} = \gamma v_i - \eta \nabla_{\theta} L(\theta_i + \gamma v_i; x^{(i)}, y^{(i)}) \quad (25)$$

与 Momentum 相比,Nesterov 动量首先在积累的方向上移动,然后更新梯度。这种方法避免了梯度下降速度过快导致的无法收敛的问题。

## 4) 批量归一化

数据归一化通常是数据处理的第一步。全局数据标准化将所有数据转换为具有零均值和单位方差。然而,当数据进入深度网络,对中间层的输入分布将会发生变化,这降低了网络的学习能力和准确性。Ioffe 等<sup>[57]</sup>提出了一个高效归一化方法叫批量归一化 (batch normalization, BN) 一定程度上解决了这个问题。它通过在每个小批量而不是整个训练集上归一化输入层的均值和方差,解决了这个被称为“协变量偏移”的问题。假设要归一化的层具有  $d$  维输入  $x = [x_1, x_2, \dots, x_d]^T$  首先归一化第  $k$  层维度:

$$\hat{x}_k = (x_k - \mu_B) / \sqrt{\delta_B^2 + \epsilon} \quad (26)$$

式中:  $\mu_B$  和  $\delta_B^2$  分别是小批量数据的均值和方差; $\epsilon$  是一个极小值。为了增强表达能力。归一化的输入  $\hat{x}_k$  被转换为:

$$y_k = BN_{\gamma, \beta}(x_k) = \gamma \hat{x}_k + \beta \quad (27)$$

式中:  $\beta$  和  $\gamma$  是学习参数。

通过批量归一化,一方面可以通过选择比较大的初



始化学习率而提升训练速度,另一方面还可以提高网络的泛化能力。

#### 5) 跨连模型

随着卷积神经网络层数的加深,随之而来的梯度消失或梯度爆炸问题会越来越严重,除了上述数据增广、权重初始化、批量归一化等方法还有一种跨连模型的方法,通过建立非相邻层之间的连接,不仅使每一层能从之前任意一层接收输入,还能把输出传递给后面的非相邻层。这一方法让卷积神经网络的深度达到成百上千层,大大提高了卷积神经网络的性能。

受到 LSTM<sup>[58]</sup> 的启发, Srivastava 等<sup>[59]</sup> 提出快道网络。普通的神经网络可以被表示为:

$$y = H(x, W_H) \quad (28)$$

式中:  $H$  为非线性函数;  $W_H$  表示权值矩阵;  $x$  表示输入;  $y$  表示输出。

快道网络的基本思想是在网络中加入两个非线性变换  $T(x, W_T)$  和  $C(x, W_C)$ , 构造快道层:

$$y = H(x, W_H) \cdot T(x, W_T) + x \cdot C(x, W_C) \quad (29)$$

式中:  $T$  表示变换门;  $C$  表示传递门; 符号“ $\cdot$ ”表示逐元素相乘。

令  $C = 1 - T$ , 快道层可以简化为:

$$y = H(x, W_H) \cdot T(x, W_T) + x \cdot (1 - T(x, W_T)) \quad (30)$$

式中:  $x, y, T(x, W_T)$  和  $C(x, W_C)$  必须有同样的维数。

不难看出,如果选择特殊的变换门  $T$ , 可以得到:

$$y = \begin{cases} x, & T(x, W_T) = 0 \\ H(x, W_H), & T(x, W_T) = 1 \end{cases} \quad (31)$$

因此,根据变换门的输出,快道层可以直接通过其输入  $x$  直接输出  $H$ , 或者在  $x$  和  $H$  之间产生平缓变化的输出。

ResNet 与普通的神经网络相比, ResNet 网络引入了跨连模型,构造了残差模块,如图 18 所示。

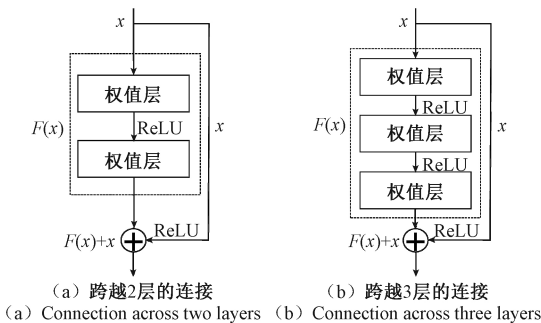


图 18 残差模块

Fig.18 Residual modules

如果在残差模块中去掉跨层连接,并用  $H(x)$  表示相

应的计算结果,那么在加入跨层连接后的计算结果  $F(x)$  与  $H(x)$  之间存在下面的关系:

$$F(x) = H(x) - x \quad (32)$$

试验表明,如果在层间加入跨层连接,那么即使是成百上千层的网络,也能够得到准确、有效的训练。

Huang 等<sup>[60]</sup> 在残差模块的基础上用密集连接模块代替残差模块,称之为密集连接网络 (densely connected convolutional network, DenseNet), 与残差网络的区别在于,密集连接模块内部允许任意两个非相邻层之间进行跨层连接,如图 19 所示。在密集连接模块中,第  $l$  层的特征图  $x_l$  是利用前面  $l-1$  层的特征图  $x_0, x_0, \dots, x_{l-1}$  来计算的,可以表示为:

$$x_l = H_l[x_0, x_0, \dots, x_{l-1}] \quad (33)$$

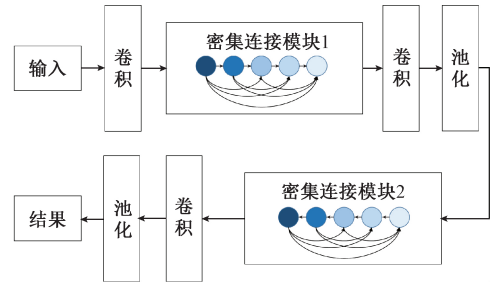


图 19 一个包含 2 个密集连接模块的密集连接网络

Fig.19 A DenseNet consisting of 2 densely connected modules

## 4 卷积神经网络在视觉图像检测中的应用

视觉图像检测是指获取目标在图像中的位置与种类。视觉图像检测在流程上分为 3 个步骤:区域建议,特征提取和区域分类,如图 20 所示。首先,在视觉图像中生成可能存在目标的建议区域;然后,使用利用卷积神经网络强大的特征提取能力提取建议区域内的特征;最终,利用分类器对建议区域分类,达到目标检测的结果。由于卷积神经网络具有强大的特征提取能力,自动提取视觉图像的高层次特征,避免了人工设计特征的复杂低效,所以卷积神经网络在视觉图像检测中取得了突破性的进展。

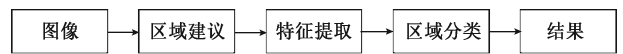


图 20 视觉检测的基本流程

Fig.20 Basic procedure of visual detection

### 4.1 视觉图像检测的公共数据集

大规模数据集的建立为卷积神经网络在视觉图像检测研究的突破提供了契机。目前,视觉图像检测研究的公共数据集有 ImageNet、PASCAL VOC、MS COCO 等。

1) ImageNet 数据集

ImageNet 数据集是目前世界上最大的图像识别数据集,包含 2 万多种类别,超过 1 400 万张图片。

2) PASCAL VOC 数据集

PASCAL VOC 从 2005 ~ 2012 年每年提供一套标准化的数据集,并举办包括分类、检测和分割在内的竞赛。其中,PASCAL VOC2007 数据集包含 20 种类别,训练集,验证集,测试集的数量分别是 2 501,2 510 和 5 011 张。PASCAL VOC2012 数据集同样包含 20 个类别,其中训练集,验证集,测试集的数量分别是 5 717,5 823 和 10 991 张。

3) MS COCO 数据集

MS COCO 数据集包含 91 种类别,30 多万张图片。与 ImageNet 数据集和 PASCAL VOC 数据集相比 MS COCO 提供了丰富的上下文信息,是目前最大的语义分割数据集。

4) VisDrone2018 数据集

VisDrone2018 数据集致力于推动无人机平台上的视觉图像检测的发展,包含 263 段视频,10 209 张图片。

4.2 卷积神经网络模型

随着对卷积神经网络的认识逐步加深,研究人员相继提出了一系列具备强大特征提取能力的网络模型,这是卷积神经网络在图像分类任务中取得巨大成功的关键。如表 1 所示,随着研究人员对卷积神经网络的研究越来越深刻,卷积神经网络的性能也越来越好。

表 1 ILSVRC 大赛中经典 CNN 网络模型对比  
Table 1 The comparison of classical CNN network models in ILSVRC competition

| 年份   | CNN 网络模型       | Top-5 错误率/% |
|------|----------------|-------------|
| 2012 | AlexNet        | 15.32       |
| 2013 | ZFNet          | 11.20       |
| 2014 | GoogLeNet      | 3.08        |
| 2014 | VGG            | 7.32        |
| 2015 | ResNet         | 3.57        |
| 2016 | Trimps-Soushen | 2.99        |
| 2017 | SENet          | 2.25        |

1) AlexNet 网络

1989 年的 LeNet 虽然在手写字符识别任务上取得了成功,但由于训练数据集规模小,算力不足,泛化能力弱等问题,没有更进一步发展。与之相比,AlexNet 卷积神经网络有更深的层数,提高了网络的性能;引入 dropout 技术,避免网络过拟合,降低模型训练时的计算量;使用数据增广技术,避免过拟合;使用 ReLU 代替 sigmoid 作为激活函数加快训练速度等。在 ImageNet 数据集上实现了 15.32% 的 top-5 错误率,之后 ILSVRC 竞赛分类任

务中性能较好的模型大多采用卷积神经网络作为分类模型。

2) ZFNet

纽约大学的 Zelier 等<sup>[21]</sup>提出了 ZFNet,与 AlexNet 相比使用了更小的卷积核和更长的步长,保留了图像更多的特征,在 2013 年 ILSVRC 竞赛分类任务中以 11.20% 的成绩获得了冠军。ZFNet 更重要的意义在于第一次尝试解释卷积神经网络内部的原理,通过反卷积可视化了特征图,发现浅层学习到的是轮廓、边缘、颜色、纹理等特征,深层学习到的是一些不易理解的抽象特征,并根据可视化结果对网络的参数调优。

3) VGG 网络

牛津大学视觉几何组的 Simonyan 等<sup>[61]</sup>在 AlexNet 网络的基础上,提出了 VGG 网络,VGG 网络证明了卷积神经网络越深,性能越好,但由于网络深度的增加会带来训练误差增大的网络退化问题,因此 VGG 网络的最佳深度受限于 19 层。在 2014 年 ILSVRC 竞赛上以 7.32% 的 top-5 错误率的成绩赢得了分类任务的亚军。

4) GoogLeNet 网络

Google 公司的 Szegedy 等<sup>[62]</sup>提出了 GoogLeNet 网络,即 Inception V1,网络有 22 层深,由若干 Incetption 模块级联而成,使用全局平均池化层替代最后的全连接层,增加辅助分类节点,最终以 6.67% 的 top-5 错误率获得了 2014 年 ILSVRC 竞赛分类任务的冠军。

5) ResNet 网络

ResNet 网络是由 Facebook 公司提出的,解决了网络深度增加带来的网络退化问题。通过运用跨连结构,将低层的特征映射到高层的网络中,成功训练了 152 层深的神经网络,在 2015 年的 ILSVRC 竞赛分类任务中以 3.57% 的 top-5 错误率取得了冠军。

6) SENet 网络

中国科学院的 Hu 等<sup>[63]</sup>提出了 SE 模型(squeeze-and-excitation module),该模型将特征通道的相互依赖关系建模,并利用一种全新的特征重标定策略,即通过学习的方式来自动提取每个特征通道的重要程度,然后根据重要程度提升有用的特征,抑制用处不大的特征,以 2.25% 的 top-5 错误率成绩获得 2017 年 ILSVRC 竞赛分类任务冠军。

4.3 基于卷积神经网络的视觉图像检测算法

由于卷积神经网络在图像分类领域的突出表现,研究人员将卷积神经网络应用到视觉图像检测算法中提取特征,并运用多种方法提高卷积神经网络的特征提取能力,在视觉图像检测领域取得了突破性的进展。一些基于卷积神经网络的视觉图像检测算法对比如表 2 和 3 所示。基于卷积神经网络的视觉图像检测算法主要分为两类:基于区域建议的方法和无区域建议的方法。

表 2 一些视觉图像检测方法在 VOC2007 测试集的比较结果

Table 2 Comparative results of some visual detection methods on VOC 2007 test set

| 检测方法         | CNN 网络模型 | 训练集                                 | mAP/% |
|--------------|----------|-------------------------------------|-------|
| R-CNN        | AlexNet  | VOC2007 trainval                    | 58.50 |
| R-CNN        | VGG16    | VOC2007 trainval                    | 66.00 |
| SPPNet       | ZFNet    | VOC2007 trainval                    | 60.90 |
| Fast R-CNN   | VGG16    | VOC2007 trainval 和 VOC2012 trainval | 70.00 |
| Faster R-CNN | VGG16    | VOC2007 trainval                    | 69.90 |
| Faster R-CNN | VGG16    | VOC2007 trainval 和 VOC2012 trainval | 73.2  |
| Faster R-CNN | ResNet   | VOC2007 trainval                    | 76.40 |
| Faster R-CNN | ResNet   | VOC2007 trainval 和 VOC2012 trainval | 85.60 |
| YOLO         | VGG16    | VOC2007 trainval                    | 66.40 |
| YOLOv2       | Darknet  | VOC2007 trainval                    | 78.60 |
| SSD300       | VGG16    | VOC2007 trainval                    | 72.10 |
| SSD512       | VGG16    | VOC2007 trainval                    | 75.10 |

表 3 一些视觉图像检测方法在 VOC2012 测试集的比较结果

Table 3 Comparative results of some visual detection methods on VOC 2012 test set

| 检测方法         | CNN 网络模型 | 训练集                                               | mAP/% |
|--------------|----------|---------------------------------------------------|-------|
| R-CNN        | AlexNet  | VOC2012 trainval                                  | 53.30 |
| R-CNN        | VGG16    | VOC2012 trainval                                  | 62.40 |
| Fast R-CNN   | VGG16    | VOC2007 trainval, VOC2007 test 和 VOC2012 trainval | 68.40 |
| Faster R-CNN | VGG16    | VOC2007 trainval, VOC2007 test 和 VOC2012 trainval | 70.40 |
| Faster R-CNN | VGG16    | COCO trainval35k                                  | 75.90 |
| Faster R-CNN | ResNet   | COCO trainval35k                                  | 83.30 |
| YOLO         | VGG16    | VOC2007 trainval, VOC2007 test 和 VOC2012 trainval | 66.40 |
| YOLOv2       | Darknet  | COCO trainval35k                                  | 78.20 |
| SSD300       | VGG16    | COCO trainval35k                                  | 79.30 |
| SSD512       | VGG16    | COCO trainval35k                                  | 82.20 |

1) 基于区域建议的方法

(1) R-CNN

Girshick 等提出了 R-CNN 方法,是第一个将卷积神经网络成功应用于目标检测领域的方法。

R-CNN 的基本流程如图 21 所示,首先使用选择性搜索提取每张输入图片的候选区域;然后通过 CNN 网络从每个区域中提取特征,这里使用 AlexNet 网络,得到一个 4 096 维的特征向量;接着把提取到的特征向量送入分类器进行分类。在 PACAL VOC2012 上的检测精度达到 53.3% mAP,如果将 R-CNN 的卷积神经网络换成特征提取效果更好的 VGG16,检测精度提高到 62.40%。

但 R-CNN 也存在一些缺陷:①候选区域过多,使得计算量过大,降低了检测速度;②候选区域直接缩放到固

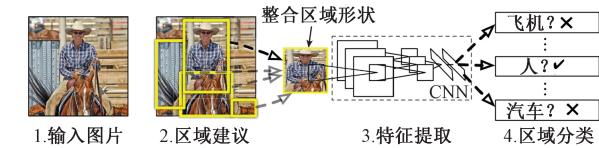


图 21 R-CNN 流程

Fig.21 Flow chart of R-CNN

定大小,导致图像失真,降低了检测精度。

(2) 空间金字塔池化网络 (Spatial Pyramid Pooling Network, SPPNet)

He 等<sup>[64]</sup>针对 R-CNN 的缺陷,提出了空间金字塔池化 (spatial pyramid pooling, SPP) 模型,其结构如图 22 所

示。SPPNet 通过引入金字塔池化层,用多个尺度的池化来替换原来单一的池化,把特征转换成相同的维度,解决了网络的输入必须是固定尺寸图片的问题。而且 SPPNet 是对整幅输入图片进行一次卷积运算,而不是对每个候选区域进行卷积运算,避免了重复计算,极大提高了计算效率。

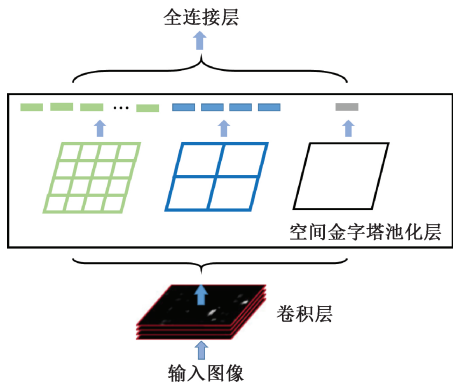


图 22 SPPNet 框架

Fig.22 Frame architecture of SPPNet

但 SPPNet 也存在如下缺点:①SPPNet 的检测过程是分阶段的,在提取特征后使用 SVM 分类器分类,之后再行边框回归,这使得训练过程复杂;②CNN 提取的特征存储占用极大的计算资源。

(3) Fast R-CNN

Girshick<sup>[65]</sup>综合了 R-CNN 和 SPPNet 的优点提出了 Fast R-CNN,实现了综合特征提取、区域分类和边框回归的端到端算法。Fast R-CNN 基本流程如图 23 所示, Fast R-CNN 首先在图像中提取感兴趣区域( region of interest, RoI);然后对整幅图片做卷积运算,在最后一个特征图上找到每个 RoI 对应的区域,将 RoI 特征图送入 RoI 池化层,通过该层将每个不同尺寸的特征图统一到同样的大小;最后经过全连接层得到两个输出向量,一个进行 Softmax 分类,另一个进行边框分类。

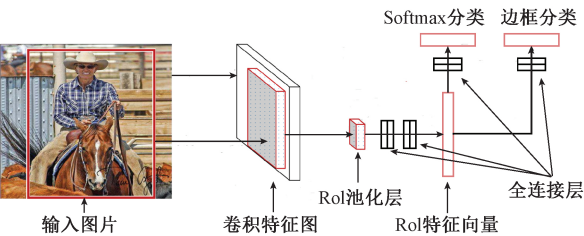


图 23 Fast R-CNN 框架

Fig.23 Frame architecture of Fast R-CNN

(4) Faster R-CNN

在 Fast R-CNN 的基础上,Ren 等<sup>[66]</sup>提出区域建议网络(region proposal network, RPN),并把 RPN 和 Fast R-

CNN 融合到一个统一的网络,称之为 Faster R-CNN。

Faster R-CNN 的框架如图 24 所示,首先将整幅图片输入 CNN 网络得到特征图,然后在特征图上利用 RPN 得到目标的候选区域,接下来在候选区域上利用分类器预测目标的类别,最后对目标位置精修实现目标检测。Fast R-CNN 利用 RPN 代替选择搜索算法,大大提高了运算效率,并真正实现了端到端的目标检测。Faster R-CNN 有很多被广泛应用的变体结构<sup>[67-69]</sup>。

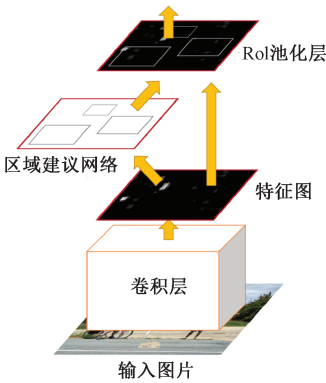


图 24 Faster R-CNN 框架

Fig.24 Frame architecture of Faster R-CNN

2) 无区域建议的方法

(1)“你只看一次”v1(you only look once version 1, YOLOv1)

Redmon 等<sup>[70]</sup>提出了一种称为 YOLO 的目标检测框架,检测速度可以达到 45 fps。YOLO 用一个单一的卷积神经网络直接对整幅图片预测目标的位置及类型,YOLO 的检测流程如图 25 所示。首先将一幅图分成  $S \times S$  个网格,每个网格预测  $B$  个边框,每个边框负责目标位置  $(x, y)$ ,目标大小  $(w, h)$ ,目标存在置信度及  $C$  个类别的条件概率的预测,将这些预测结果编码成一个  $S \times S \times (B \times 5 + C)$  维的张量。虽然 YOLO 的检测速度很快,达到了实时的效果,但是 YOLO 对相邻目标或小尺寸目标的检测效果不好。

(2)单点多盒子探测器(single shot multibox detector, SSD)

针对 YOLO 的不足,Liu 等<sup>[71]</sup>提出了单点多盒子探测器模型,在兼顾实时性的同时提高了检测精度。SSD 运用特征金字塔的思想,在不同层次的特征图上进行检测,大大提高了检测精度,尤其是对小尺寸目标有很好的效果。

(3) YOLOv2 和 YOLO9000

YOLOv2 是 Redmon 等<sup>[72]</sup>在 YOLO 的基础上采用多尺度训练、批规范化、高分辨率分类器等多种方法,提升了 YOLO 的检测精度与检测速度,并且提高了模型的健壮性。YOLO9000 模型可以检测超过 9 000 类目标。



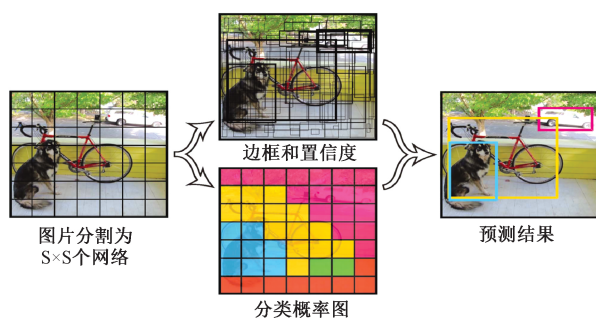


Fig.25 Frame architecture of YOLO

## 5 结 论

近年来,卷积神经网络引起了研究人员广泛关注,应用于视觉图像检测的卷积神经网络变体层出不穷,大大增强了卷积神经网络提取特征的能力,与传统方法相比,卷积神经网络在视觉图像检测领域有以下优势:

1)层次化的特征表达。卷积神经网络自动提取视觉图像的低层次特征如边缘、轮廓、线条、色块等特征到高层次特征如语义等特征。

2)强大的表达能力。与传统的浅层模型相比,卷积神经网络层数更多,表达能力大大增强。

3)端到端。卷积神经网络可以实现端到端的视觉图像目标检测,能够同时优化多个任务。

但还存在改进的空间,需要进一步研究:

1)小样本学习。目前的神经网络所需的训练样本数量都在上百万以上,如何从少量的数据中高效提取特征并得到良好性能的网络模型是卷积神经网络未来的研究方向。

2)网络结构的优化。当前,卷积神经网络越深,卷积神经网络的测试性能,但是伴随着网络的加深,过拟合和网络退化问题也更加严重;网络的深度加深也使得网络模型体积更加,需要更大的内存、存储空间与计算量。因此,如何优化网络结构,在保持模型测试性能的同时,降低网络的深度是亟需解决的问题。

3)理论知识的完善。卷积神经网络缺乏基础的数学知识,一直被视为端到端的“黑盒”,人们难以理解卷积神经网络中隐含层数目、神经元个数和激活函数形式等会对结果产生怎样的影响。完善卷积神经网络的理论知识,是改进、理解卷积神经网络决策的关键。

总的来说,在过去的十几年内,卷积神经网络强大的特征提取能力大大加速了视觉图像检测领域的发展,基于卷积神经网络的视觉图像检测仍有巨大的潜力。

## 参考文献

- [1] 石杰,周亚丽,张奇志,等. 基于改进 Mask RCNN 和 Kinect 的服务机器人物品识别系统[J]. 仪器仪表学报,2019,40(4):216-228.  
SHI J, ZHOU Y L, ZHANG Q ZH, et al. Service robot item recognition system based on improved Mask RCNN and Kinect[J]. Chinese Journal of Scientific Instrument, 2019,40(4):216-228.
- [2] 陈朋,汤一平,何霞,等. 基于多任务 Faster R-CNN 车辆假牌套牌的检测方法[J]. 仪器仪表学报,2017,38(12):3079-3089.  
CHEN P, TANG Y P, HE X, et al. Detection method for fake plate vehicles based on multitask Faster R-CNN[J]. Chinese Journal of Scientific Instrument, 2017,38(12):3079-3089.
- [3] 王永利,曹江涛,姬晓飞,等. 基于卷积神经网络的 PCB 缺陷检测与识别算法[J]. 电子测量与仪器学报,2019,33(8):78-84.  
WANG Y L, CAO J T, JI X F, et al. PCB defect detection and recognition algorithm based on convolutional neuralnetwork [J]. Chinese Journal of Scientific Instrument, 2019,33(8):78-84.
- [4] 章黎明,贾智伟,谢俊力,等. 近四年胶囊内镜图像病灶的深度学习识别研究进展[J]. 电子测量与仪器学报,2019,33(8):69-77.  
ZHANG L M, JIA ZH W, XIE J L, et al. Recent four years review on deep learning classification research in capsule endoscopy image processing[J]. Chinese Journal of Scientific Instrument, 2019,33(8):69-77.
- [5] HINTON G, SALAKHUTDINOV R R. Reducing the dimensionality of data with neural networks[J]. Science, 2006,313(5786):504-507.
- [6] HUBEL D H, WIESEL T N. Receptive fields, binocular interaction, and functional architecture in the cat's visual cortex [J]. Journal of Physiology, 1962, 160 (1): 106-154.
- [7] HUBEL D H, WIESEL T N. Receptive fields of single neurons in the cat's striate cortex [J]. Journal of Physiology, 1959,148(3):547-591.
- [8] FUKUSHIMA K. Neural network model for a mechanism of pattern recognition unaffected by shift in position-neocognitron [J]. IEICE Technical Report, 1989, 62(10):658-665.
- [9] FUKUSHIMA K. Neocognitron; A self-organizing neural network for a mechanism of pattern recognition unaffected



- by shift in position [J]. *Biological Cybernetics*, 1980, 36(4):193-202.
- [10] FUKUSHIMA K. Artificial vision by multi-layered neural networks: Neocognitron and its advances [J]. *Neural Networks*, 2013, 37:103-119.
- [11] LECUN Y, BOSER B, DENKER J, et al. Backpropagation applied to handwritten zip code recognition [J]. *Neural Computation*, 1989, 11(4): 541-551.
- [12] 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.
- [13] EVERINGHAM M, GOOL L V, WILLIAMS C K I, et al. The pascal visual object classes (VOC) challenge[J]. *International Journal of Computer Vision*, 2010, 88(2):303-338.
- [14] LIN T Y, MAIRE M, BELONGIE S, et al. Microsoft coco: Common objects in context [C]. *European Conference on Computer Vision*, 2014:740-755.
- [15] ZHU P, WEN L, BIAN X, et al. Vision meets drones: A challenge [J]. *Computer Vision and Pattern Recognition*, 2018, arXiv:1804.07437.
- [16] KRIZHEVSKY A, SUTSKEVER I, HINTON G, et al. Imagenet classification with deep convolutional neural networks [C]. *Conference and Workshop on Neural Information Processing Systems*, 2012:1097-1105.
- [17] GIRSHICK R, DONAHUE J, DARRELL T, et al. Rich feature hierarchies for accurate objection and semantic segmentation[C]. *IEEE Conference on Computer Vision and Pattern Recognition*, 2014:580-587.
- [18] FELZENSZWALB P, MCALLESTER D, RAMANAN D, et al. A discriminatively trained, multiscale, deformable part model [C]. *IEEE Conference on Computer Vision and Pattern Recognition*, 2008:1-8.
- [19] NGIAM J, CHEN Z, CHIA D, et al. Tiled convolutional neural networks[C]. *Conference and Workshop on Neural Information Processing Systems*, 2010:1279-1287.
- [20] ZELIER M D, KRISHNAN D, TAYLOR G W, et al. Deconvolution networks [C]. *IEEE Conference on Computer Vision and Pattern Recognition*, 2010: 2528-2535.
- [21] ZELIER M D, FERGUS R. Visualizing and understanding convolutional networks [C]. *European Conference on Computer Vision*, 2014:818-833.
- [22] FISHER Y, VLADLEN K. Multi-scale context aggregation by dilated convolutions [C]. *International Conference on Learning Representations*, 2016:1-13.
- [23] LIN M, CHEN Q, YAN S C, et al. Network in network[C]. *International Conference on Learning Representations*, 2014:1-10.
- [24] SZEGEDV C, LIU W, JIA Y, et al. Going deeper with convolutions[C]. *IEEE Conference on Computer Vision and Pattern Recognition*, 2015:1-9.
- [25] SZEGEDV C, IOFFE S, VANHOUEKE V, et al. Inception-v4, inception-resnet and the impact of residual connections on learning [C]. *Conference on Artificial Intelligence*, 2017:4278-4284.
- [26] HE K M, ZHANG S, REN S, et al. Deep residual learning for image recognition [C]. *IEEE Conference on Computer Vision and Pattern Recognition*, 2016: 770-778.
- [27] BRUNA J, SZLAM A, LECUN Y, et al. Signal recovery from pooling representations [C]. *International Conference on Machine Learning*, 2014:307-315.
- [28] YU D, WANG H, CHEN P, et al. Mixed pooling for convolutional neural networks [C]. *International Conference on Rough Sets and Knowledge Technology*, 2014:364-375.
- [29] ZEILER M D, FERDUS R. Stochastic pooling for regularization of deep convolutional neural networks[C]. *International Conference on Learning Representations*, 2013:1-9.
- [30] SNOEK J, ADAMS R P. Spectral representations for convolutional neural networks [C]. *Conference and Workshop on Neural Information Processing Systems*, 2015:2449-2457.
- [31] HE K M, ZHANG X, REN S, et al. Spatial pyramid pooling in deep convolutional networks for visual recognition [J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2015, 37(9):1904-1916.
- [32] GONG Y, WANG L, GOU R, et al. Multi-scale orderless pooling of deep convolutional activation features[C]. *European Conference on Computer Vision*, 2014:392-407.
- [33] NAIR V, HINTON G E. Rectified linear units improve restricted Boltzman machines [C]. *International Conference on Machine Learning*, 2010:807-814.
- [34] MAAS A L, HANNUN A Y, NG A Y, et al. Rectifier nonlinearities improve neural network acoustic models[C]. *International Conference on Machine*

- Learning, 2013;1152-1160.
- [35] HE K M, ZHANG X, REN S, et al. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification[C]. International Conference on Computer Vision, 2015;1026-1034.
- [36] XU B, WANG N, CHEN T, et al. Empirical evaluation of rectified activations in convolutional network [J]. Computer Science for Machine Learning, 2015, arXiv: 1505. 00853.
- [37] CLEVERT D A, UNTERRTHINER T, HOCHREITER S, et al. Fast and accurate deep network learning by exponential linear units (ELUs) [J]. Computer Science for Machine Learning, 2016, arXiv:1511. 07289.
- [38] GOODFELLOW J, WARDE F D, MIRZA M, et al. maxout networks [C]. International Conference on Machine Learning, 2013;1319-1327.
- [39] SPRINGENBERG J T, RIDMILLER M. Improving deep neural networks with probabilistic maxout units [J]. Computer Science for Machine Learning, 2013, arXiv: 1312. 6116.
- [40] GU J X, WANG Z H, KUEN J, et al. Recent advances in convolutional neural networks [J]. Pattern Recognition, 2018,77:354-377.
- [41] HINTON G, SRIVASTAVA N, KRIZHEVSKYY A, et al. Improving neural networks by preventing co-adaptation of feature detectors[J]. Computer Science for Machine Learning. 2012, arXiv:1207. 0580.
- [42] WAN L, ZEILER M, ZHANG S, et al. Regularization of neural networks using dropconnect [C]. International Conference on Machine Learning, 2013;1058-1066.
- [43] YANG H, PATRAS I. Mirror on the wall, tell me, is the error small? [C] IEEE Conference on Computer Vision and Pattern Recognition, 2015;4685-4693.
- [44] XIE S, TU Z. Holistically-nested edge detection [C]. International Conference on Computer Vision, 2015: 1395-1403.
- [45] SALAMON J, BELLO J P. Deep convolutional neural networks and data augmentation for environmental sound classification[J]. Signal Process Letter, 2017, 24(3): 279-283.
- [46] EIGEN D, FERGUS R. Predicting depth, surface normal and semantic labels with a common multi-scale convolutional architecture [C]. International Conference on Computer Vision, 2015;2650-2658.
- [47] YANG S. Several tips and tricks for ImageNet CNN training [J]. Technique Report, 2016;1-12.
- [48] CHOROMANSKA A, HENAFF M, MATHIEU M, et al. The loss surfaces of multilayer networks[C]. International Conference on Artificial Intelligence and Statistics, 2015: 192-204.
- [49] GLOROT X, BENGIO Y. Understanding the difficulty of training deep feedforward neural networks [C]. International Conference on Artificial Intelligence and Statistics, 2010;249-256.
- [50] SAXE A M, MCCLELLAND J L, GANGULI S, et al. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks[J]. Computer Science for Machine Learning, 2014, arXiv:1312. 6120.
- [51] MISHKIN D, MATAS J. All you need is a good init[C]. International Conference on Learning Representations, 2016;1-13.
- [52] WIJNHOFEN R. Fast training of object detection using stochastic gradient descent [C]. International Conference on Pattern Recognition. 2010;424-427
- [53] LOSHCHILOV I, HUTTER F. SGDR: stochastic gradient descent with warm restarts [J]. Computer Science for Machine Learning, 2017, arXiv:1608. 03983.
- [54] SCHAUL T, ZHANGS, LECUN Y, et al. No more pesky learning rates [C]. International Conference on Machine Learning, 2013;343-351.
- [55] QIAN N. On the momentum term in gradient descent learning algorithms[J]. Neural Networks, 1999,12(1): 145-151.
- [56] SUTSKEVER I, MARTENS J, DAHL G, et al. On the importance of initialization and momentum in deep learning [C]. International Conference on Machine Learning, 2013;1139-1147.
- [57] IOFFE S, SZEGEDY C. Batch normalization: Accelerating deep network training by reducing internal covariate shift [J]. Journal of Machine Learning Research, 2015;448-456.
- [58] HOCHREITER S, SCHMIDHUBER J. Long short-term memory [J]. Neural Computation, 1997, 9 (9): 1735-1780.
- [59] SRIVASTAVA R K, GREFF K, SCHMIDHUBER J. Training very deep networks [C]. Conference and Workshop on Neural Information Processing Systems, 2015;2377-2385.
- [60] HUANG G, LIU Z, WEINBERGER K Q. Densely connected convolutional networks [C]. IEEE Conference

- on Computer Vision and Pattern Recognition, 2016; 4700-4708.
- [61] SIMONYAN K, ZISSERMAN A. Very deep Convolutional networks for large-scale image recognition[J]. Computer Vision and Pattern Recognition, 2014, arXiv: 1409.1556.
- [62] SZEGEDY C, LIU W, JIA Y Q, et al. Going deeper with convolutions[C]. IEEE Conference on Computer Vision and Pattern Recognition, 2015:1-9.
- [63] HU J, SHEN L, SUN G, et al. Squeeze-and-Excitation networks[C]. IEEE Conference on Computer Vision and Pattern Recognition, 2019:7132-7141.
- [64] HE K M, ZHANG X, REN S, et al. Spatial pyramid pooling in deep convolutional networks for visual recognition [C]. European Conference on Computer Vision, 2014:346-361.
- [65] GIRSHICK R. Fast R-CNN [C]. IEEE International Conference on Computer Vision, 2015:1440-1448.
- [66] REN S, HE K M, GIRSHICK R, et al. Faster R-CNN: Towards real-time object detection with region proposal networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017,39(6):1137-1149.
- [67] LIN T Y, DOLLAR P, GIRSHICK R, et al. Feature pyramid networks for object detection [C]. IEEE Conference on Computer Vision and Pattern Recognition, 2017:936-944.
- [68] CAI Z, VASCONCELOS N. Cascade R-CNN: Delving into high quality object detection[C]. IEEE Conference on Computer Vision and Pattern Recognition, 2018: 6154-6162.
- [69] HE K, GKIOXARI G, DOLLÁR P, et al. Mask R-CNN[C]. IEEE International Conference on Computer Vision, 2017:2980-2988.
- [70] REDMON J, DIVVALA S, GIRSHICK R, et al. You only look once: Unified real-time object detection[C]. IEEE Conference on Computer Vision and Pattern Recognition, 2016:779-788.
- [71] LIU W, ANGUELOV D, ERHAN D, et al. SSD: Single

shot multibox detector [C]. European Conference on Computer Vision, 2016:21-37.

- [72] REDMON J, FARHADI A. YOLO9000: better, faster, stronger[C]. IEEE Conference on Computer Vision and Pattern Recognition, 2017:6517-6525.

## 作者简介



**蓝金辉** (通信作者), 1998 年于北京理工大学获得博士学位, 现为北京科技大学教授、博士生导师, 主要研究方向为目标识别、深度学习、遥感。

E-mail: lanjh@ustb.edu.cn

**Lan Jinhui** (Corresponding author) received her Ph. D. degree in 1998 from Beijing Institute of Technology. Now, she is a professor and Ph. D. supervisor in University of Science and Technology Beijing. Her main research interests include object recognition, deep learning and remote sensing.



**王迪**, 2015 年于北京信息科技大学获得学士学位, 2017 年进入北京科技大学攻读硕士和博士学位, 现为北京科技大学博士研究生, 主要研究方向为深度学习, 目标识别。

E-mail: wangdi305@163.com

**Wang Di** received his B. Sc. degree in 2015 from Beijing Information Science and Technology University. He started his M. Sc. and Ph. D. program in 2019 from Beijing Information Science and Technology University. Now, he is a Ph. D. candidate in University of Science and Technology Beijing. His main research interests include deep learning and object recognition.



**申小盼**, 2018 年于北京信息科技大学获得学士学位, 现为北京科技大学硕士研究生, 主要研究方向为深度学习, 目标识别。

E-mail: sxiaopan2017@163.com

**Shen Xiaopan** received her B. Sc. degree in 2018 from Beijing Information Science and Technology University. Now, she is a M. Sc. candidate in University of Science and Technology Beijing. Her main research interests include deep learning and object recognition.