深度卷积神经网络的发展及其在计算机视觉领域的应用

张顺[1)] 龚怡宏[2)] 王进军[2)]

[1)](西北工业大学电子与信息学院，陕西西安，710072)

[2)](西安交通大学人工智能与机器人研究所，陕西西安，710049)

摘 要 作为类脑计算领域的一个重要研究成果，深度卷积神经网络已经广泛应用到计算机视觉、自然语言处理、信息检索、语音识别、语义理解等多个领域，在工业界和学术界掀起了神经网络研究的浪潮，促进了人工智能的发展。卷积神经网络直接以原始数据作为输入，从大量训练数据中自动学习特征的表示。卷积神经网络具有局部连接、权值共享和池化操作等特性，可以有效降低网络复杂度，减少训练参数的数目，使模型对平移、扭曲、缩放具有一定程度的不变性。目前，深度卷积神经网络主要是通过增加网络的层数，使用更大规模的训练数据集，以及改进现有神经网络的网络结构或训练学习算法等方法，来模拟人脑复杂的层次化认知规律，拉近与人脑视觉系统的差距，使机器获得“抽象概念”的能力。深度卷积神经网络在图像分类、目标检测、人脸识别、行人再识别等多个计算机视觉任务中都取得了巨大成功。本文首先回顾了卷积神经网络的发展历史，简单介绍了 M-P 神经元模型、Hubel-Wiesel 模型、神经认知机、用于手写识别的 LeNet，以及用于 ImageNet 图像分类比赛的深度卷积神经网络。然后详细分析了深度卷积神经网络的工作原理，介绍了卷积层、采样层、全连接层的数学表示及各自发挥的作用。接着本文重点从以下三个方面重点介绍卷积神经网络的代表性成果，并通过实例展示各种技术方法在图像分类精度的提升效果。从增加网络层数方面，讨论并分析了 AlexNet、ZF-Net、VGG、GoogLeNet 和 ResNet 等经典卷积神经网络的结构；从增加数据集规模方面，介绍了人工增加标注样本的难点，以及使用数据扩增技术对神经网络性能提升的作用；从改进训练方法方面，介绍了包括 L2 正则化、Dropout、Dropconnect、Maxout 等常用的正则化技术，Sigmoid 函数、tanh 函数以及 ReLU 函数、LReLU 函数、PReLU 函数等常用的神经元激活函数，softmax 损失、hinge 损失、contrastive 损失、triplet 损失等不同损失函数，以及 batch normalization 技术的基本思想。针对计算机视觉领域，本文重点介绍了卷积神经网络在图像分类、目标检测、人脸识别、行人再识别、图像语义分割、图片标题生成、图像超分辨率、人体动作识别以及图像检索等任务的最新研究进展。从人类视觉认知机制出发，分析了视觉信息分层处理和“大范围优先”视觉认知过程的相关理论成果和对当前计算模型的一些理论启示。最后提出了未来基于深度卷积神经网络的类脑智能研究待解决的问题与挑战。

关键词 类脑智能；神经网络；深度学习；计算机视觉；视觉认知

中图法分类号 TP18

The Development of Deep Convolution Neural Network and Its Applications on Computer Vision

ZHANG Shun[1)] GONG Yihong[2)] WANG Jinjun[2)]

[1)](School of Electronics and Information, Northwestern Polytechnical University, Xi’an, Shaanxi, 710072)

[2)](Institute of Artificial Intelligence and Robotics, Xi’an Jiaotong University, Xi’an, Shaanxi, 710049)

本课题得到国家重点基础研究发展计划(973计划) (2015CB351705)；国家自然科学基金重点项目(61332018)；国家自然科学基金青年科学基金项目(61703344)；中央高校基本科研业务费专项资金(3102017OQD021)资助。张顺，男，1987年生，博士，助教，主要研究领域为计算机视觉和机器学习，E-mail:szhang@nwpu.edu.cn。龚怡宏，男，1963年生，博士，教授，博士生导师，“国家千人计划”专家，主要研究领域为多媒体内容分析、机器学习和模式识别，E-mail: ygong@mail.xjtu.edu.cn。王进军，男，1977年生，博士，教授，博士生导师，CCF会员，主要研究领域为模式识别、机器学习和多媒体计算，E-mail: jinjun@mail.xjtu.edu.cn。

Abstract As the important research achievement, deep convolutional neural networks have been widely applied to various fields such as computer vision, natural language processing, information retrieval, speech recognition, semantic understanding, and have attracted a wave of neural networks research from both academia and industry and have contributed to the development of artificial intelligence. The convolutional neural networks directly treat the original data as input, automatically learn the feature representations from a large number of training data. The convolutional neural networks have the characteristics of local connection, weight sharing and pooling operation, which can effectively decrease the network complexity and reduce the number of training parameters, so that the model has some certain invariance to translation, distortion and scale. Currently, many approaches of deep neural networks, including the increase of size and complexity of neural networks, the use of larger sets of training data, the improvement of neural network architecture and training methods, etc., have been proposed to simulate the complex hierarchical cognitive attributes of human brain and pull close the gap between the human brain and visual system, so that the machine has the capability to capture "abstraction concepts". The deep convolutional neural networks have been a great success in many computer vision tasks, such as image classification, object detection, face recognition, person re-identification. In this paper, we first review the history of the development of convolutional neural networks, and briefly introduce M-P neuron model, Hubel-Wiesel model, Neocognitron, LeNet for handwriting recognition, and deep convolutional neural network for image classification in the ImageNet competition. Then we have a detailed analysis of the fundamental principle of deep convolutional neural networks, and introduce the mathematical representation and the respective functions of the convolution layer, the pooling layer and the fully connected layer. Besides, this paper focuses on the representative works of convolutional neural networks on the following three aspects, and demonstrates various technical methods in improving the accuracy of image classification using examples. In the aspect of increasing the number of neural networks' layers, the architectures of classical convolutional neural networks such as AlexNet, ZF-Net, VGG, GoogLeNet and ResNet are discussed and analyzed. In the aspect of increasing the amount of data, we introduce the difficulties of increasing the number of annotated samples by manual way, and the effect in improving the performance of convolutional neural networks by data augmentation. In the aspect of improving training methods, we introduce the generalized regularization techniques such as the L2 regularization, Dropout, Dropconnect and Maxout, several frequently-used neuron activation functions such as the sigmoid function, the tanh function, the ReLU function, the LReLU function, and the PReLU function, several different loss functions such as the softmax loss, the hinge loss, the contrastive loss and the triplet loss, and the basic idea of the batch normalization technique. In the field of computer vision, this paper focuses on the more recent research progress of convolutional neural networks in image classification, object detection, face recognition, pedestrian recognition, image semantic segmentation, image captioning, image super resolution, human action recognition and image retrieval. From the prospective of the human visual cognitive mechanism, we analyze the relevant theoretical achievements of hierarchical processing in the visual system and "global first" visual and cognitive process, and some theoretical implications for the current computational models. Finally, some remained problems and challenges of the brain-like intelligence research based on deep convolutional neural networks are concluded.

Key words brain-like intelligence; neural network; deep learning; computer vision; visual cognition

1 引言

让机器以类似人脑的方式进行快速学习与准确认知，是科学家们长期探索与追求的一大科学梦想。几十年来，脑神经科学和心理学等领域在人脑结构及认知机理等方面的许多研究成果都被转化为人工智能领域的计算模型，极大地促进了后者的发展与进步。人工神经网络正是在这种背景下被提出的。它是利用计算模型模拟大脑神经

系统的结构和功能，运用大量的简单运算单元，由人工方式建立起来的神经网络系统。人工神经网络的诞生及发展是类脑计算领域的一个最为重要的研究成果。

从二十世纪四十年代最早提出的 M-P 神经元模型和 Hebb 学习规则开始，人工神经网络领域已提出了上百种神经网络模型，其中具有代表性的网络包括感知机、反传网络、自组织映射网络、Hopfield 网络、玻尔兹曼机、适应谐振理论等，并在诸如手写体识别、语音识别、图像识别和自然语音处理等技术领域取得了成功的应用。

当前，卷积神经网络（Convolutional Neural Networks, CNN）是得到广泛应用的一种人工神经网络，也是首个真正被成功训练的深层神经网络。Hubel 和 Wiesel 在 1962 年通过对猫的视觉皮层细胞进行了深入研究，提出了高级动物视觉系统的认知机理模型[1]。该模型提出高级动物视觉神经网络由简单细胞和复杂细胞构成（如图 1 所示）。神经网络底层的简单细胞的感受野只对应视网膜的某个特定区域，并只对该区域中特定方向的边界线产生反应。复杂细胞通过对特定取向性的简单细胞进行聚类，拥有较大感受野，并获得具有一定不变性的特征。上层简单细胞对共生概率较高的复杂细胞进行聚类，产生更为复杂的边界特征。通过简单细胞和复杂细胞的逐层交替出现，视觉神经网络获得了提取高度抽象性及不变性图像特征的能力。

1984 年日本学者 Fukushima 在 Hubel 和 Wiesel 的感受野概念基础上，提出了神经认知机（Neocognitron）[2,3]模型。神经认知机模型由多种类型的细胞单元组成，其中最重要的两个细胞单元称为 S 细胞和 C 细胞。S 细胞的功能提取局部特征（如边缘或角等），类似 Hubel-Wiesel 模型的简单细胞。C 细胞对应 Hubel-Wiesel 模型的复杂细胞，对 S 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。

神经认知机可以看作是卷积神经网络的雏形，而卷积神经网络是神经认知机的推广形式。卷积神经网络是一个由卷积层（Convolution Layer）与降采样层（Sampling Layer）交替出现的多层神经网络，每层由多个二维特征平面组成（称为特征图，Feature Map）。构成卷积层 x 的每个神经元负责对输入图像（假定 $x=1$）或者 $x-1$ 降采样层的某个特征图的特定区域施行卷积运算，而降采样层 y 的每个神经元则负责对 $y-1$ 卷积层的某个特征图的特定区域进行最大池化（只保留该区域神经元的最大输出值）操作。卷积层与降采样层的神经元分别用来模拟 Hubel-Wiesel 模型中的简单细胞和复杂细胞，而卷积层中同一个特征图的神经元都是共享一个卷积核，负责提取同一种图像特征，对应某种特定取向的简单细胞。

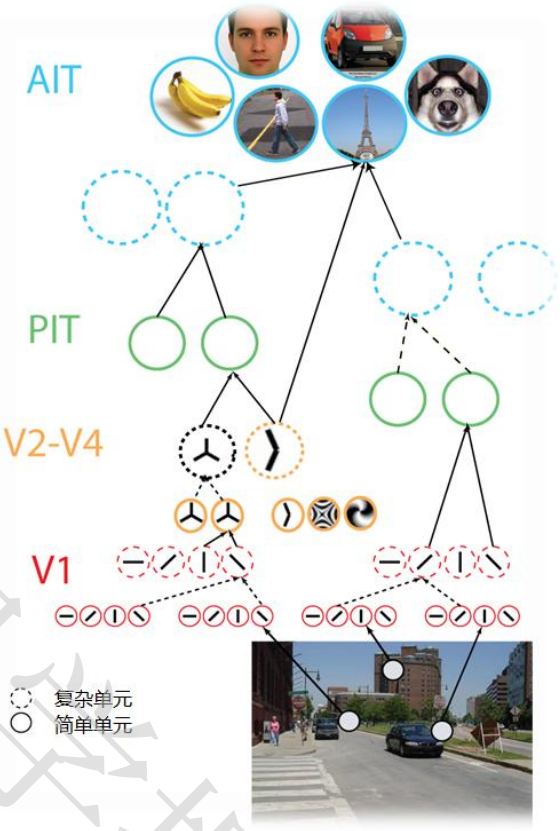


图 1 人脑视觉通道神经网络

二十世纪九十年初期，纽约大学的 Yann LeCun 等人提出了多层卷积神经网络并成功应用于手写数字识别中，所提出的系列 LeNet[4,5]都达到商用水平，被当时美国邮政局和许多大银行用来识别信封上的手写邮政编码及支票上面的手写数字。卷积神经网络起初只能训练不太深的网络，虽然在小规模问题上能取得较好的效果，然而随着网络深度和宽度的增加，在大规模图像数据集上卷积神经网络的识别效果不佳，因此在被提出后的很长一段时间里并未被重视。此后二十年中，许多研究人员对深度神经网络提出了深层结构的优化和训练学习方法的改进，深度卷积神经网络（Deep CNN）的性能得到了大幅提升。在 2012 年，Hinton 团队 ImageNet 图像分类比赛中获得压倒性胜利，将 1000 类图像的 Top-5 分类错误率从

26.172%降低到 15.315%[6]。在这一年，Deep CNN 还被用于解决 Drug Activity 预测问题，并获得当时最好成绩。至此，神经网络的研究进入了一个崭新的时代，开启了神经网络研究的热潮。

与其它神经网络相比，卷积神经网络这种基于高级动物视觉通路的网络结构极大减少了神经元间的连接和权重数量，这不但减轻了神经网络的过拟合问题，而且也降低了训练多隐层网络的难度。即使是训练一个深度达一、二十层的网络，使用误差反向传播（Back Propagation, BP）[7,8]算法也能达到收敛[9,10,11]，这是其它神经网络所望尘莫及的。

当前，Deep CNN 相对传统机器学习算法的优势不断扩大，传统学习方法在多个领域无法与深度学习抗衡，比如手写体识别、图像分类、图像语义理解、语音识别和自然语言理解等技术领域。神经网络能够重新焕发青春的原因有几个方面。首先，丰富的网络图像和大规模有标注的数据集在很大程度上缓解了训练过拟合的问题，如 ImageNet 数据集包含了 21,841 个图像类别，共计 14,197,122 幅图片。其次，计算机硬件的飞速发展提供了强大的计算能力，使得训练大规模神经网络成为可能。单个 GPU（Graphics Processing Unit，图形处理器）芯片可以集成上千个运算核心，对以高阶矩阵运算为主的神经网络提供高并行计算。此外，神经网络的模型设计和训练方法都取得了长足的进步。例如，为了改进神经网络的训练，研究人员提出了深层结构的优化和训练学习方法的改进，包括使用 ReLU 激活函数，使用 dropout 进行网络训练，使用 batch normalization 技术归一化特征的数据分布等。

神经网络的研究与人类视觉的研究密切相关，为了进一步提高神经网络的性能，通过借鉴人脑视觉系统的最新研究成果为卷积神经网络的研究寻找下一个突破口，已成为越来越受到学术界关注的研究方向。通过对人脑视觉通路的深入研究，从视觉神经网络的结构、各层的视觉信息表达、以及高层网络的视觉认知机理中获得科学启示，结合数学、统计与工程等相关技术，能制作出更加接近人脑环境理解和认知能力的机器视觉系统。

本文接下来的部分包含以下四个主要内容。第 2 章系统性地介绍卷积神经网络的结构及原理，并对新近发展起来的提升卷积神经网络性能的技术方法进行阐述和讨论。第 3 章介绍卷积神经网络在目标检测，图像语义分割，图片标题生成，人脸识别，行人再识别，图像超分辨率等领域的应用。第 4 章分析了人类视觉认知机制的特点和带给当前计算模型的一些理论启示。最后对未来的研究方向做出展望。

2 卷积神经网络及其相关技术

卷积神经网络是由用于特征提取的卷积层和用于特征处理的亚采样层交叠组成的多层神经网络。典型的卷积神经网络结构[12]如图 2 所示，网络输入是一个手写数字图像，输出是其识别结果，输入图像经过若干个“卷积”和“采样”加工后，在全连接层网络实现与输出目标之间的映射。通常卷积神经网络中，每一层神经元节点只与其邻近上下层局部感受野内的神经元节点连接。这种局部连接观点与 Hubel、Wiesel 从猫科动物的视觉系统中发现的局部感知观点相一致。图 2 中的输入图像的大小为 32×32 像素，含 R、G、B 三个通道。卷积层 C1 使用大小为 5×5 的多个卷积核对输入图像的各个通道做卷积滤波，采取图像的局部特征，得到和卷积核数量相同、大小为 28×28 的特征图。然后将这些特征图按一定的方式组合起来，作为卷积层的输出。图中原特征图经过采样层 S2 后，尺寸被缩减至 14×14，其中特征图上每个神经元与上一层中对应特征映射的 2×2 邻域相连，并据此计算输出。卷积神经网络中的卷积层中的神经元是模拟 Hubel-Wiesel 模型中的简单细胞，降采样层的神经元模拟复杂细胞，而特征图上的神经元共享同一个卷积核，对应某种特定取向的简单细胞。进行若干个卷积—采样操作，可以得到尺寸很小但数量很多的特征图。将特征图按一定方式展开，拼接为一维向量输入全连接层中，然后经过若干全连接层和输出层连接完成识别任务。

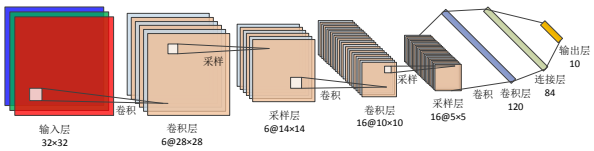


图 2 卷积神经网络的典型结构[12]

卷积神经网络的卷积层由若干个特征图组成，每个特征图上的所有神经元共享同一个卷积核的参数，由卷积核对前一层输入图像做卷积运

算得到。卷积核中每一个元素都作为权值参数，同输入图像相应区块的像素值相乘，然后将各项乘积求和，并经过激活函数得到输出像素。虽然在形式上表现为多通道特征图的三阶张量卷积操作，但实质上等同于将多个输入信号加权求和后作用于一个神经元，然后激活输出的过程。第 l 层的第 j 个特征图矩阵 x_j^l 可能由前一层若干个特征图卷积加权得到，一般可以表示为：

$$x_j^l = f\left(\sum_{i\in N_j} x_i^{l-1} * k_{ij}^l + b_j^l\right), \qquad (1)$$

其中，f 为神经元激活函数；N_j 代表输入特征图的组合，*表示卷积运算，k_{ij}^l 为卷积核矩阵，b_j^l 为偏置矩阵。常用的神经元激活函数有 sigmoid 函数，tanh 函数，ReLU 函数等。

采样层也称为“池化”层，其作用是基于局部相关性原理进行池化采样，从而在减少数据量的同时保留有用信息。采样过程可以表示为：

$$x_j^l = f\left(\text{down}(x_j^{l-1})\right), \qquad (2)$$

其中，down(•)表示采样函数，常用的由最大值采样函数和均值采样函数。最大值采样函数是把区块中元素的最大值作为函数输出，提取特征平面的局部最大响应，通常用于低层特征提取，对输入的特征图选取最显著的特征。均值采样函数是计算区块元素的算术平均值作为函数输出，提取特征平面局部响应的均值。采样过程与卷积过程类似，使用一种不带权参数的采样函数，从输入特征图的左上角开始按一定步长向右（或向下）滑动，对窗口相应区块的像素进行采样后输出。

卷积神经网络在卷积层和采样层后，通常会连接一个或多个全连接层。全连接层的结构和全连接神经网络的隐层结构相同，全连接层的每个神经元都会与下一层的每个神经元相连。第 l 层全连接层特征向量 x^l 可以如下表示：

$$x^l = f\left(w^l x^{l-1} + b^l\right), \qquad (3)$$

其中，w^l 是权值矩阵，b^l 是偏置向量。

当模型的最后输出层为逻辑回归层时，卷积神经网络输出的每个节点表示输入图片属于某一类别 i 的概率：

$$P(Y=i \mid x,w,b) = softmax_i(wx+b) = \frac{e^{w_i x+b_i}}{\sum_j e^{w_j x+b_i}}, \qquad (4)$$

式中，w 为最后一层的权参数，b 为相应偏置参数。

卷积神经网络可以使用 BP 算法进行训练，但在训练中，卷积层中每个特征图的所有神经元都是共享相同的连接权，这样可以大幅减少需要训练的参数数目。对于一个含 m 个样本的训练集，其损失函数可以使用交叉熵表示为：

$$J = -\frac{1}{m}\sum_{i=1}^{m}\log(P(Y=y^{(i)} \mid x^{(i)},w,b)) \quad .(5)$$

在测试时，卷积神经网络的预测值为：

$$y_{pred} = \arg\max_i P(Y=i \mid x,w,b). \qquad (6)$$

近年来，卷积神经网络在越来越多的领域超越传统模式识别与机器学习算法，取得顶级的性能与精度。这些成果主要是通过：1）增加神经网络层数，2）加大训练样本的数量，3）改进训练学习算法这三方面的技术手段来实现的。本章将主要从以上三个方面来介绍深度神经网络研究方面的代表性成果，并通过实例展示各种技术手段对神经网络图像分类精度的提升效果。

2.1 增加网络层数

在给定带标签数据集的前提下，提升深度神经网络识别精度的一种直接方法是增加网络层数[10]。

2012 年，在 ImageNet ILSVRC 挑战赛的大规模图像分类任务中，Alex Krizhevsky 等人搭建了一个 8 层的卷积神经网络（简称 AlexNet[6]），最终 top-5 分类错误率达到 15.315%，抛离第二名用传统机器学习方法得到的结果——26.172%分类错误率——10 多个百分点。AlexNet 使用了 5 个卷积层（另外包括 3 个 pooling 层和 2 个 norm 层）、3 个全连接层，总共 60M 个参数。

具体的网络参数配置如图 3 所示。每个输入图片都被缩放为 256×256 大小，并从中随机截取 224×224 大小的方形区块，以 RGB 三个颜色维度输入。由于当时 GPU 的性能限制，Alex Krizhevsky 等人在两个 GPU 上并行处理 AlexNet，故图 3 中隐藏层显示为两路同时计算。前 5 层是卷积层，以第一层为例，产生 96 个 55×55 节点的特

征图（Feature Map），每个特征图由大小为 11×11，步长为 4 的卷积核构成。卷积滤波后，通过 ReLU 激活函数得到卷积层的输出激励后，经过局部响应归一化和最大池化下采样操作，输出给下一个卷积层。网络在五层卷积层的基础上加上一个三层的全连接网络来做分类器，对高维卷积特征进行分类得到类别标签。全连接网络最终输出维数为 1000 的神经元响应，对应于待分类图像的 1000 个类别。

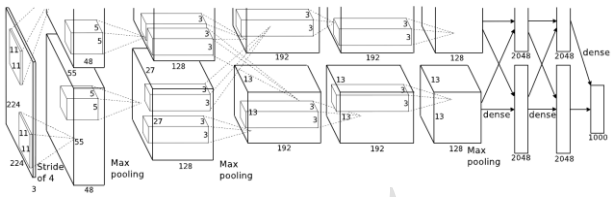


图 3 AlexNet 模型结构[6]

在 2013 年的 ImageNet ILSVRC 比赛中，排名前 20 的小组使用的都是深度学习算法，其中 Matt Zeiler 和 Rob Fergus 以其自主设计开发的 ZF-Net① 赢得了冠军，在不用额外训练数据的情况下，Top5 分类错误率达到了 11.743%。ZF-Net 所采用的深度神经网络框架几乎和 AlexNet 一样，区别仅仅是把第一个卷积层的卷积核尺寸从 11×11 修改为 7×7，步长从 4 缩小为 2，由此输出特征图的尺寸增大为 110×110，相当于增加了网络的宽度。

在 2014 年的 ImageNet ILSVRC 竞赛上，牛津大学的 Karen Simonyan 和 Andrew Zisserman 设计的 VGG（Visual Geometry Group）网络[9]获得了定位任务第一名和分类任务第二名。VGG 主要通过增加网络的深度提高网络性能。VGG 由 8 个部分构成，它们是 5 个卷积组、2 个全连接特征层和 1 个全连接分类层。每个卷积组由 1~4 个卷积层串联构成，所有卷积层都使用了 3×3 的小尺寸卷积核。多个 3×3 卷积层可看作是大尺寸卷积层的分解，如两个 3×3 卷积层的有效卷积核大小是 5×5，三个 3×3 卷积层的有效卷积核大小是 7×7。这样做的好处是，多个小尺寸卷积层比一个大尺寸卷积层有更少的参数，且能在不影响视野域的情况下增加映射函数的非线性，使网络更加具有判别性。根据每个卷积组内卷积层层数不同，VGG 给出了 A~E 五种配置方法（如图 4 所示），网络层数从 11 层增加到 19 层，对应的网络参数从 133M 增加到 144M。论文的试验测试结果表明，随着网络层数的不断加深，VGG 网络的准确率在 16 层时达到性能瓶颈，之后趋于饱和。

虽然增加深度神经网络深度能一定程度提升网络的性能，但这种方法有两个瓶颈。一方面是大的网络结构需要学习更多的参数，容易造成网络对训练数据集的过拟合。另一方面，层数多的网络需要更多的计算资源。例如两个相互连接的卷积层同时增加特征维度，计算量呈平方增长；如果额外增加的神经元没有得到有效的利用（很多权值接近零），就会造成计算资源的浪费。

网络配置					
A	A-LRN	B	C	D	E
11 层	11 层	13 层	16 层	16 层	19 层
输入（224×224 RGB 图）					
conv3-64	conv3-64 **LRN**	conv3-64 **conv3-64**	conv3-64 conv3-64	conv3-64 conv3-64	conv3-64 conv3-64
最大池化					
conv3-128	conv3-128	conv3-128 **conv3-128**	conv3-128 conv3-128	conv3-128 conv3-128	conv3-128 conv3-128
最大池化					
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**
最大池化					
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**
最大池化					
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**
最大池化					
全连接层-4096					
全连接层-4096					
全连接层-4096					
Soft-max					

图 4 VGG 网络结构的不同配置（从左到右）。卷积层参数表示为“conv<卷积核大小>-<通道数>”[9]

Google 公司的 Christian Szegedy 等人开发设计的 GoogLeNet[10]网络模型使用新颖的 Inception 结构作为基本模块进行级联，实现了在提升网络深度的同时大大减少网络参数，并且充分利用了计算资源，提高了算法的计算效率。他们在 2014 年参加 ImageNet ILSVRC 挑战赛，并获得了图像分类任务的冠军。

GoogLeNet 由多个 Inception 基本模块级联组成，网络达到 22 层的深度。Inception 结构如图 5 所示，其主要思想是以 3 个不同尺寸的卷积核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 的卷积核，其中 1×1 的卷积核较

① www.clarifai.com

前一层有较低的维度，主要用于数据降维，在传递给后面的3×3和5×5卷积层时降低了他们的卷积计算量，避免了由于增加网络规模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。

然而，很深的网络结构给预测误差的反向传播带来了困难，因为从顶层传到底层的误差已经变得很小，难以驱动底层参数的更新。GoogLeNet采取的策略是将监督信号直接加到多个中间层，这意味着中间和低层的特征表示也需要能够准确对训练数据分类。

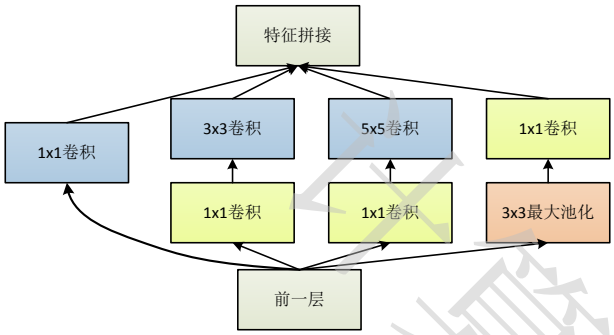


图5 Inception模型结构

与AlexNet及ZF-Net网络模型对比，GoogleNet删除了倒数两个全连接层。一般而言，全连接层含有整个网络的绝大多数参数，却只占用很小的计算资源，反之亦然。如在AlexNet中，前五层卷积层只拥有网络5%的参数，但却消耗了整个网络95%的计算量。而后三层全连接层占有网络95%的学习参数，却只需要5%的计算量。这造成了学习参数和计算资源利用的极度不平衡。通过去除全连接层，GoogLeNet虽然增加了网络的深度，但整个网络的参数只有6M，而且还消除了上述学习参数与计算资源间的不平衡现象，达到充分利用计算资源的目的。

最近，在2015 ImageNet计算机识别挑战赛中，微软亚洲研究院何恺明等人提出的残差网络（residual networks，ResNet）[11]在2015年ImageNet ILSVRC挑战赛上获得图像分类、图像定位以及图像检测三个主要项目的冠军，在同一年的微软COCO比赛上获得检测和分割的冠军。在ImageNet比赛上，所使用的152层深度残差网络深度是VGG网络深度的8倍，但网络的参数量却要比VGG网络要少。

虽然ReLU、pReLU、batch normalization等一系列方法的提出，解决了深度神经网络训练的梯度消失或爆炸以及特征分布不均匀等问题。但在训练很深的网络时，随着网络深度的增加，所增加后续层的训练和测试的错误率反而增加。深度残差网络借鉴了highway网络的思想，在构造网络时增加了捷径连接，使后续层的输出不是传统神经网络中输入的映射，而是输入的映射和输入的叠加，如下图6所示。网络要优化的是图6中的残差函数$F(x)$.

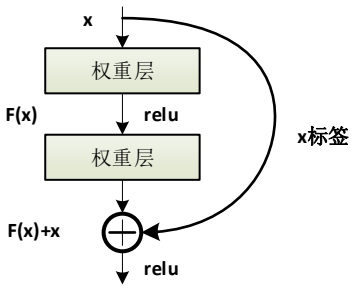


图6 残差网络的学习模块

表1罗列了当前的顶级深度卷积神经网络、它们的网络构成、参数量及其在ImageNet验证集上的图像分类精度。从表中的数据可以看出，增加网络层数的确能够提升图像分类的精度，当GoogLeNet从22层[10]增加到31层[13]（复杂度增加41%）时，网络参数量从6.8M增加到8M（增加18%），图像分类top-5错误率由7.9%下降到5.82%。微软亚洲研究院何恺明等人提出的MSRA模型[14]与GoogLeNet同为22层网络，但前者的参数量是后者的29倍，top-5错误率下降了约2.2%。残差网络在图像分类任务上的top-5错误率降低到了4.49%。

表1 顶级深度神经网络的网络构成、参数量，及在ImageNet验证集上的分类错误率

模型	构成	参数量	Top-1 错误率	Top-5 错误率
AlexNet 2012[6]	8层 (5conv+3fc)	~60M	40.7%	15.3%
ZF-Net	8层 (5conv+3fc)	~60M	37.5%	16.0%
VGG[9]	19层 (16conv+3fc)	~144M	24.4%	7.1%
GoogLeNet[10]	22层	~6.8M	-	7.9%
GoogLe-BN Model[13]	31层	~8M	21.99%	5.82%
MSRA[14]	22层	~200M	21.59%	5.71%

ResNet[11]	152 层	~22M	19.38%	4.49%

2.2 增加训练数据集规模

在训练上述如此巨大的神经网络时，如果没有充分的训练数据，模型将极有可能陷入过拟合。出现过拟合时，直观的表现如图 7 所示。随着训练过程的进行，模型复杂度增加，在训练集上的错误率渐渐减小，但是在验证集上的错误率却反而渐渐增大。过拟合出现的原因一般有两点：一是训练样本数量太少，得到的网络参数不能准确模拟数据的分布；二是由于模型复杂度过高，训练样本数据里的噪音干扰过大，使模型过分拟合了噪音数据，反而忽略了正确样本数据。因此，模型虽然对训练数据拟合非常好，但是对于训练集外的数据拟合效果却非常差。

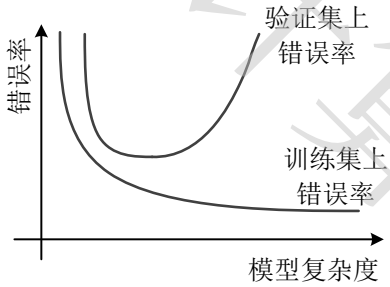


图 7 过拟合示意图

避免过拟合问题最简单直接的方法就是增加训练样本的数量。在 LFW 数据集上的人脸验证任务中，DeepID[15-17]、DeepFace[18]、FaceNet[19]等模型都已接近或达到了人类的识别精度。这些模型都利用了从互联网下载的海量带标签人脸图像进行监督式预训练（Pre-train）。比如 DeepID 使用了 10K 类的外部人脸数据集，FaceNet 则使用了近 8M 个不同类别。训练好的深度神经网络通常作为特征提取器再应用在特定的人脸验证数据库上进行人脸验证。

然而，收集更多的数据意味着需要耗费更多的人力、物力和财力，而且构建高质量的数据集往往还需要相关的专业知识，因而单纯增加标注样本的方法并不可行。

另一种简单获取更多数据的方式被称为数据扩增（Data Augmentation），它通过对原始图片施行各种变换来得到更多的数据，例如：将原始图片旋转一个小角度、添加随机噪声、带弹性的形变和截取原始图片的一部分等。图 8 展示了一种常用的数据集扩增方法：从一张 256×256 大小的输入图像中，按照 224×224 固定大小从不同位置截取出多个图像块，再通过水平翻转生成大量训练图片。香港中文大学 Sun 等人开发的 DeepID，从一张人脸图像中截取出不同大小的图像块作为样本训练了 60 个神经网络，极大地增加了训练数据的数量。Simard 等人[20]对 MNIST[12]中的训练样本做了各种变种扩增来提高模型性能。GoogLeNet 和 VGG 网络在 ImageNet 比赛中都使用了多尺度图像训练集的方法：训练不同输入图片尺度下（例如 512×512，256×256）的多个模型，最后综合评估所有模型的输出结果。另外还有调整图片亮度、饱和度、对比度、偏色等方法。

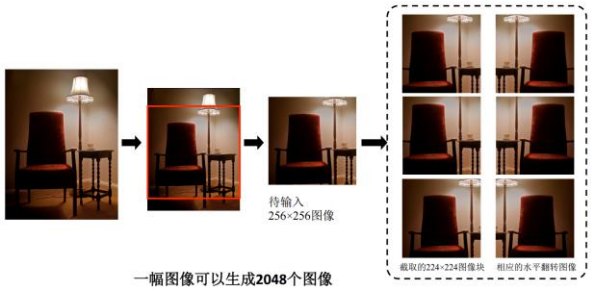


图 8 数据集扩增

表 2 揭示了使用数据扩增技术后对神经网络性能提升的作用。

表 2 CIFAR-10 测试集上的分类错误率[21]

方法	错误率（无数据扩充）	错误率（数据扩充）
CNN+Spearmint	14.98%	9.5%
Conv.maxout+ Dropout	11.68%	9.38%
NIN+Dropout	10.41%	8.81%

2.3 正则化

训练大型卷积神经网络除了增大训练数据集外，还经常使用正则化方法②来防止过拟合问题。发生过拟合的模型一般在某些很小的区间里，函数值的变化很剧烈。这就意味着函数的参数值偏大，使某些小区间里的导数值（绝对值）非常大。正则化是通过约束参数的范数使其不要过大，以此降低模型的复杂度，从而减小噪声输入的扰动，可以在一定程度上减少过拟合情况。

L_2 正则化是最常用的一种正则化技术，又称

② http://neuralnetworksanddeeplearning.com. Chapter 3, Improving the way neural networks learn

权重衰减（weight decay），它是在原始的损失函数 C_0 后面再加上一个正则化项：

$$C = C_0 + \frac{\lambda}{2}\sum_w \|w\|_2^2, \tag{7}$$

正则化项是所有网络权重 w 的平方和。λ（λ>0）是正则项系数，用来权衡正则项与 C_0 的比重。另外系数 1/2，主要是为了后面求导数计算方便。

通过对式(7)求导，有：

$$\begin{cases} \dfrac{\partial C}{\partial w} = \dfrac{\partial C_0}{\partial w} + \lambda w \\ \dfrac{\partial C}{\partial b} = \dfrac{\partial C_0}{\partial b} \end{cases}, \tag{8}$$

从上式可以发现 L_2 正则化项对偏置 b 的更新没有影响，但是对于权重 w 的更新有影响：

$$w \to w - \eta\frac{\partial C_0}{\partial w} - \eta\lambda w = (1-\eta\lambda)w - \eta\frac{\partial C_0}{\partial w}. \tag{9}$$

在不使用 L_2 正则化时，求导结果中 w 前系数为 1。使用 L_2 正则化后，w 前面系数为 $1-\eta\lambda$，因为 η 和 λ 都是正的，所以 $1-\eta\lambda<1$，它的效果是减小 w，这也就是权重衰减的由来。当 w 为正时，更新后的 w 变小；当 w 为负时，更新后的 w 变大。因此它的效果就是让 w 往 0 靠，使网络中的权重尽可能为 0，也就相当于减小了网络的权重，降低了网络复杂度，防止过拟合。

L_2 正则化是通过修改代价函数来实现的，而 Dropout[22,23] 则是通过修改神经网络本身来实现的，它是训练网络时常用的一种技巧。Dropout 是在训练过程中以 $1-p$（p 一般取 0.5）的概率将隐层节点的输出值清 0，而用 BP 算法更新权值时，不再更新与该节点相连的权值。如图 9(a)所示，考虑一个全连接层，v 表示 $n\times 1$ 维的输入向量，W 表示 $d\times n$ 维的网络权重，$a(x)$ 是一个满足 $a(0)=0$ 的激活函数。M 是个 $d\times 1$ 的列向量，其每个元素服从概率为 p 的伯努利分布。该层的输出 r 可由以下公式计算得到：

$$r = M.\times a(Wv). \tag{10}$$

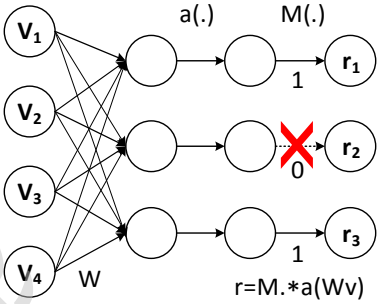
运用了 Dropout 的训练过程，相当于训练了指数界别的只有半数隐层单元的神经网络（简称为“半数网络”），每一个这样的半数网络，都可以给出一个分类结果，神经网络最终的结果是对这些半数网络进行平均或者集成。Dropout 可以有效提高网络的性能，并能防止过拟合。

DropConnect[24]和 Maxout[25]，都可以提高深度神经网络的泛化能力，两者是对 Dropout 的改进。DropConnect 与 Dropout 的不同之处是，它不是随机将隐层节点的输出清 0，而是将节点中的每个与其相连的输入权值以 $(1-p)$ 的概率清 0（DropConnect 是对权重作用，而 Dropout 是对神经元输出作用，如图 9 所示）。

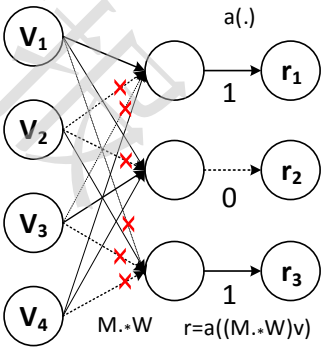
DropConnect 表达式如下：

$$r = a((M.\times W)v). \tag{11}$$

两者的区别从图 9 中可以看出。



（a）Dropout 网络示意图



（b）DropConnect 网络示意图

图 9　Dropout 和 DropConnect 网络对比

Maxout 其实是一种激活函数形式。一般情况下，如果激活函数采用 ReLU 函数，在前向传播过程中，第 i 个隐层节点的输出表达式为：

$$h_i(x) = \max\left(x^T W_{\cdots i} + b_i, 0\right), \tag{12}$$

其中 W 是 2 维的，$W_{\cdots i}$ 表示 W 中的第 i 列向量。

对于 Maxout 激活函数，如图 10 所示，其隐层节点的输出表达式为：

$$h_i(x)=\max_{j\in[1,k]} z_{ij}, \tag{13}$$

$$z_{ij}=x^T W_{\cdots ij}+b_{ij}, \tag{14}$$

其中 $W\in R^{d\times m\times k}$，$d$ 表示输入层节点的个数，m 表示隐层节点的个数，k 表示子隐层节点的个数。这 k 个子隐层节点都是线性输出的，而隐层节点的输出取这 k 个子隐层节点输出值中最大的那个值。因为激活函数中有了 max 操作，所以整个 Maxout 网络是一种非线性的变换。Maxout 具有非常强的拟合能力，它可以拟合任意的的凸函数。

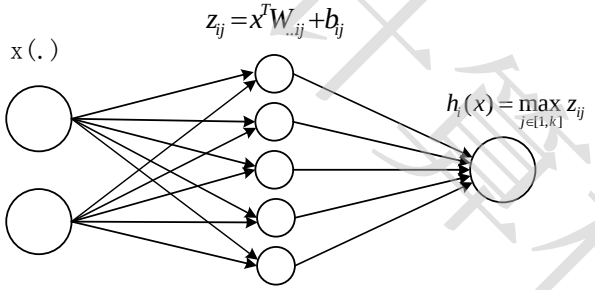


图 10 Maxout 网络示意图

表 3 给出了不同正则化方式在 CIFAR-10 和 CIFAR-100 数据集上分类结果的比较。

表 3 不同正则化方式在 CIFAR-10 和 CIFAR-100 数据集上的分类错误率[23] [25]

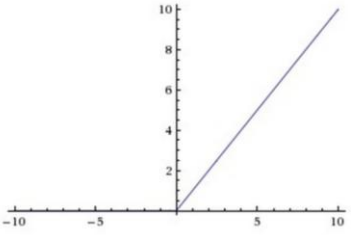
Method	CIFAR-10	CIFAR-100
Without Dropout	15.6	43.48
Dropout in fc layers	14.32	41.26
Dropout in all layers	12.61	37.2
Dropout + Maxout	11.68	38.57
DropConnect	11.1	-

2.4 其他改进训练学习方法

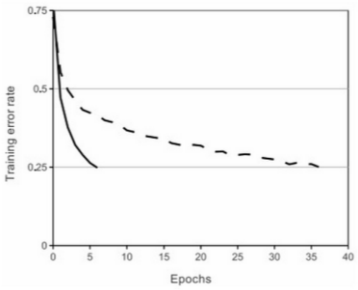
卷积神经网络中改进训练学习的方法除了使用正则化外，还有改进激活函数、定义不同损失函数、使用 batch normalization 等常用技术。

深层神经网络中的激活函数通常使用非线性函数，通过非线性的组合可以逼近任何函数。在公式（1）中，常用的神经元激活函数有 Sigmoid 函数，tanh 函数，ReLU、LReLU 和 PReLU 函数等。Sigmoid 函数是一种非线性激活函数，数学形式为 $f(x)=\frac{1}{1+e^{-x}}$。Sigmoid 函数将神经元的输出信号映射到[0,1]之间。对于深层卷积神经网络，Sigmoid 函数反向传播时，很容易出现梯度消失的问题。这是由于在 Sigmoid 饱和区，函数的梯度接近于 0，反向传播中计算的梯度也会接近于 0。这样在参数更新过程中，传到前几层的梯度几乎为 0，网络参数几乎不会再更新。另外，Sigmoid 函数的输出值始终在 0 和 1 之间，这会导致后一层的神经元以当前层输出的非 0 均值数据作为输入。虽然使用 batch 进行训练能一定程度缓解非 0 均值这一问题，但仍给深度网络的训练造成不便。Tanh 激活函数的数学形式为 $f(x)=\frac{e^x-e^{-x}}{e^x+e^{-x}}$，它将神经元的输出信号映射到[-1,1]范围内。Tanh 函数的输出是 0 均值的，在实际应用中，tanh 函数比 sigmoid 函数较好，但也存在梯度消失问题，会导致训练效率低下。

ReLU (Rectified Linear Units，修正线性单元)函数是近几年深度学习领域非常流行的一种神经元激活函数，数学形式为 $f(x)=\max(0,x)$，函数曲线如图 11(a)所示。ReLU 函数在 $x>0$ 时的梯度恒等于 1，因此在反向传播过程中，前几层网络的参数也能得到快速更新，缓解了梯度消失问题。另外，sigmoid 和 tanh 函数都需要较大的计算量（如指数计算等），ReLU 函数则是通过非常简单的阈值化的激活对参数进行稀疏化。由于 ReLU 函数的线性、非饱和性，与 sigmoid 和 tanh 函数相比，ReLU 函数能明显加快卷积神经网络的收敛速度。论文[6]指出，相比 tanh 函数，使用 ReLU 函数时的收敛速度可以加快 6 倍，如图 11(b)所示。



（a）ReLU 激活函数曲线



（b）卷积神经网络在使用 ReLU 函数（实线）时的收敛速度是使用 tanh 函数（虚线）时的 6 倍

图 11 ReLU 激活函数[6]

在 ReLu 函数训练过程中，当流过一个 ReLU 神经元的梯度较大时，可能导致该神经元的权重参数不会再次更新。如果神经元出现以上所述的“死亡”情况，那这些神经元的梯度将永远是零。如果训练时学习率设置过高，可能导致高达 40％的网络处在“死亡”中，这部分神经元在整个训练过程中从未被激活过。通过合理设置学习率可以有效降低这一现象。

为了避免 ReLU 神经单元在训练时可能会“死亡”现象，LReLU（Leaky Rectified Linear Unit）激活函数使神经元在整个训练过程中能持续得到更新。LReLU 激活函数的表达式如下：

$$f(x)=\begin{cases}x & x>0\\ \alpha x & x\le 0\end{cases}, \tag{15}$$

其中，α 通常取一个很小的固定值，如α=0.01。

RReLU（Parametric Rectified Linear Unit）激活函数是带一个自适应参数的ReLU函数。RReLU的表达式与式（15）相同，但 RReLU 的α 是个随机变量，训练时它在给定范围随机取值。当α=0 时，RReLU 相当于 ReLU；当α 取一个很小的值是，相当于 LReLU。

面对特定的任务，选择合适的损失函数非常关键。常用的损失函数有 softmax 函数、hinge 损失函数、contrastive 损失函数、triplet 损失函数等。在本章开头已介绍过 softmax 函数，这里重点介绍其他三种损失函数。

Hinge 损失函数的数学形式如下：

$$L=\frac{1}{m}\sum_{i=1}^{m}[\max(0,1-\delta(Y=y^{(i)})w^{T}x^{(i)})]^{p}, \tag{16}$$

式中，当卷积神经网络分类正确时，$\delta(Y=y^{(i)})$ 值为 1，否则值取-1。注意到，当 $p=1$，上式为标准的 hinge 损失（或称为L_1-Loss）；当 $p=2$，上式为平方 hinge 损失 （或称为L_2-Loss）。与标准 hinge 损失函数对比，平方 hinge 损失函数对损失值的惩罚要更大。

Contrastive 损失函数常用于训练 Siamese 网络。Siamese 网络是由结构相同且共享权值的两个卷积神经网络组成，输入是一对图像，如图 12(a)所示。假设 (x_1,x_2) 是一对输入图像，$f(x_1)$ 和 $f(x_2)$ 是输入图像在卷积神经网络最高隐藏层提取的特征向量。Contrastive 损失函数的主要目的是在特征空间上拉近同一类别样本之间的距离，并增大不同类别样本之间的距离。

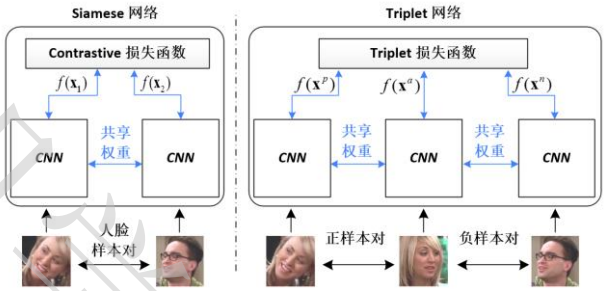


图 12 Siamese 网络和 Triplet 网络的结构示意图。Siamese 网络包含两个架构相同且参数共享的 CNN 模型，以一对图像作为输入，使用 constrastive 损失函数进行训练。Triplet 网络包含三个相等的 CNN，以图像三元组作为输入，使用的损失函数为 triplet 损失函数。

Contrastive 损失函数的定义如下：

$$L=\frac{1}{2m}\sum_{i=1}^{m}y\cdot D(x_1,x_2)+(1-y)\max[0,\tau-D(x_1,x_2)] \tag{17}$$

式中，图像对 (x_1,x_2) 的相似性直接用特征空间上的欧式平方距离度量：$D(x_1,x_2)=\|f(x_1)-f(x_2)\|^2$。当 (x_1,x_2) 属于同一类别时，y 取值为 1，需要减小欧式平方距离

$D(x_1,x_2)$ 才能降低损失。相反，当 (x_1,x_2) 属于不同类别时， y 取值为 0,需要增大 $D(x_1,x_2)$ 直到大于阈值 τ 。

Triplet 网络由结构相同且共享权值的三个卷积神经网络组成，如图 12(b)所示。Triplet 网络的输入是三元组 (x^a,x^p,x^n)，由两张来自同一个类别的图像 (x^a,x^p) 和一张来自不同类别的图像组成 x^n 。Triplet 损失函数是最小化下式：

$$L=\frac{1}{2m}\sum_{i=1}^{m}\max[0,D(x^a,x^p)-D(x^a,x^n)+\tau], \tag{18}$$

由上式可见，triplet 损失函数是优化输入图像在特征空间上的欧式平方距离，使不同类别的图像 x^n 远离图像对 (x^a,x^p) 并大于一个阈值 τ 。

深度卷积神经网络的训练是一个非常复杂的学习过程。随机梯度下降法由于其简单、高效的特点成为训练深度网络的主流方法，但是它需要研究人员手动微调网络参数，如学习率、模型初始参数、权重衰减参数、Dropout 比例等，这些参数的选择对深度卷积神经网络的训练结果至关重要。在训练神经网络的过程中，每一层网络的参数在不断更新，会导致下一层输入的数据分布情况发生改变，而且数据分布的变化随着网络深度的增大而变大。那么神经网络需要在每次迭代都去学习适应不同的分布，这样将会大大降低神经网络的训练速度。神经网络的隐层在训练过程中发生数据分布改变这一现象，称为 internal covariate shift。

Batch Normalization 的基本思想，通过预处理操作，让每个隐层的所有节点的激活输入分布归一化到均值为 0 方差为 1 的标准正态分布，并且均值和方差都在当前迭代的 mini-batch 样本中计算得到。假设某个 mini-batch 的样本数目为 m ，网络某个隐层神经元为 x ， x 在 mini-batch 中 m 个取值表示为 $\{x_1,\ldots,x_m\}$ 。神经元 x 在 mini-batch 中的均值 μ_B 和方差 σ_B^2 表示如下：

$$\mu_B \leftarrow \frac{1}{m}\sum_{i=1}^{m}x_i , \tag{19}$$

$$\sigma_B^2 \leftarrow \frac{1}{m}\sum_{i=1}^{m}(x_i-\mu_B)^2 . \tag{20}$$

该神经元经过 Batch Normalization 操作后得到的归一化值 $\{x_1,\ldots,x_m\}$，表示如下：

$$x_i \leftarrow \frac{x_i-\mu_B}{\sqrt{\sigma_B^2+\varepsilon}}, \tag{21}$$

其中， ε 是极小的正数，防止除零操作。

如果单纯应用以上归一化公式，对网络隐层的输出数据做归一化，可能会影响此隐层的特征表达能力。举例说明，如果该隐层学习到的特征数据本身分布在 Sigmoid 函数的两侧非线性区，经归一化处理后，特征数据变换到了 Sigmoid 函数中间的线性区域，这就破坏了特征数据分布。为了不改变归一化后特征的表达能力，在归一化操作后需要执行线性变换：

$$y_i \leftarrow \gamma x_i+\beta, \tag{22}$$

其中，尺度参数 γ 和平移参数 β 用于恢复原始特征的数据分布。

3 卷积神经网络的应用

卷积神经网络是近十几年来类脑计算领域取得的一个重大研究成果，它在计算机视觉、语音识别、自然语言处理、多媒体等诸多领域都取得了巨大成功。在计算机视觉领域的各类任务中，图像分类任务是根据图像信息中反映的不同特征，把不同类别的目标（如鸟、人、车、飞机等）区分开来，即给每幅图片分配一个语义类别标记，而目标检测是定位出某类目标在图像中出现的区域。与图像分类任务要建立图像级理解不同，图像语义理解要得到图像像素级别的目标分类结果。图片标题生成也是建立于图片的语义理解上，要求自动产生自然语言对图片的目标及目标间关系进行描述。相比于图像分类和目标检测

关注于多类或单类物体目标的区分或定位，人脸识别和行人再识别任务则分别聚焦于人脸和行人的身份辨识。另外一种任务——图像超分辨率，能够提供更清晰的图像以及更多的图像细节，为高层视觉任务提供更好的输入。

本章将重点介绍卷积神经网络在图像分类、目标检测、人脸识别、行人再识别、超分辨率、人体动作识别以及图像检索的最新研究进展。

3.1 图像分类

图像分类是计算机视觉领域的一个重要应用，主要是指对给定的一幅图片，使计算机根据图片中的内容将其分类到合适的类别，分配一个语义类别标记。深度卷积神经网络在图像分类中最重要的进展体现在 ImageNet ILSVRC 挑战中的图像分类任务上，前一章针对这类任务重点介绍了几种网络模型，如 AlexNet[6]，ZF-Net，GoogleNet[10]，VGG[9]和 ResNet[11]等，这里不再赘述。

除 ImageNet 图像数据集之外，图像分类常用的数据集还有 Caltech-101[26]，Caltech-256③，TinyImage[27]，SUN[28]等。表 4 列举了一些在图像分类领域常用的数据集及其重要信息。

表 4 图像分类领域常用数据集

名称	包含类别数量	图片数量
Caltech101[26]	101	9,146
Caltech256	256	30,607
TinyImage[27]	75,062	79,302,017
SUN[28]	899	130,519
ImageNet[29]	21,841	14,197,122

3.2 目标检测

目标检测（Object Detection）是计算机视觉领域的一项基本任务，主要是定位图像中特定物体出现的区域并判定目标类别。与图像分类相比，目标检测更加关注图像的局部区域和特定的物体类别集合，被视为更加复杂的图像识别问题。

传统的目标检测算法大多采用滑动窗口的方式，使用手工设计的特征，如常用的特征描述子 Haar[30]、SIFT（Scale-Invariant Feature Transform）[31]、PCA-SIFT[32]、SURF（Speeded Up Robust Feature）[33]等，对每类物体单独训练一个浅层分类器。早于 2001 年，Viola 和 Jones[34]提出目标检测领域最具影响力的目标检测算法，能实时处理目标检测同时具有很高检测率，成功应用于人脸检测。算法使用 AdaBoost[35]算法框架，提取目标 Haar-like[36]特征，然后采用滑动窗口搜索策略实现准确有效地定位。Dalal 等人[37]以图像的梯度方向直方图（Histogram of Oriented Gradient, HOG）作为特征，使用支撑向量机（Supported Vector Machine, SVM）[38-40]作为分类器进行行人检测。由于自然界的大部分物体存在非刚体形变，Felzenszwalb 等人[41]提出了多尺度形变部件模型（Deformable Part Model, DPM）。DPM 继承了使用 HOG 特征和 SVM 分类器的优点。DPM 目标检测器由一个根滤波器和一些部件滤波器组成，组件间的形变通过隐变量进行推理。由于目标模板分辨率固定，算法采用滑动窗口策略在不同尺度和宽高比图像上搜索目标。后续工作采用不同策略加速了 DPM 的穷尽搜索策略。

传统目标检测算法主要依靠设计者的先验知识，抽取样本中手工设计的特征。为了方便手工调参数，特征设计中只能出现少量的参数。另一方面，面对难度较高的检测任务，浅层分类器由于模型深度不够，所需要的参数和训练样本会呈指数增加。与传统目标检测算法比较，深度卷积神经网络可以从大数据的丰富内在信息中自动学习包含上万参数的特征表示，同时，深度模型使特征学习过程更有效率。

随着 2012 年深度卷积神经网络在图像分类任务上取得重大突破，众多学者开始利用 Deep CNN 取代浅层分类器解决目标检测问题，也带动了目标检测精度的提升[42,43]。其中较有影响力的工作包括 R-CNN[44]，Deep MultiBox[45]，Overfeat[46]，Fast RCNN[47]和 SPP-Net[48]。最具代表性的是 Girshick 等人在 R-CNN 中提出的基于 Region Proposal 的深度学习目标检测框架。如图 13 所示，R-CNN 算法首先采用选择性搜索（Selective Search）[49]策略在输入图像上提取若干候选窗，利用深度卷积神经网络从候选窗提取深度特征，然后利用 SVM 等线性分类器基于特征将候选窗分为目标和背景，最后使用非极大值抑制方法舍弃部分候选窗，得到目标物体的定位结果。候选窗方法能够高效地在图像候选区域内进行识别，更为灵活地处理物体长宽比的变化，从而获得较高的检测正

③ Griffin G., Holub A., Perona P.. Caltech-256 object category dataset. California Institute of Technology, 2007. http://authors.library.caltech.edu/7694

确率。

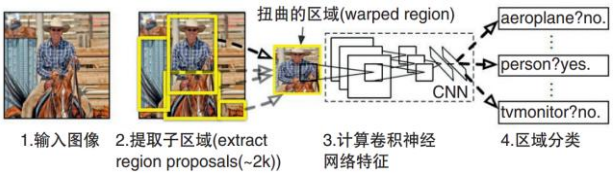


图 13 R-CNN 目标检测算法的流程图[44]

基于选择性搜索策略和 CNN 的目标检测算法在目标检测上取得很好的效果，远远超越了传统机器学习算法。但是这种检测算法遇到了速度瓶颈，由于选择性搜索方法是使用传统的图分割方式生成，一幅图像约需要 2s 才能完成候选窗的搜索，这极大限制了算法的训练和测试时间。针对这一问题，Ren 等人[50]提出的 Faster R-CNN 目标检测方法采用深度卷积神经网络替代传统的选择性搜索策略。Faster R-CNN 中加入了一种生成候选区域的 RPN（Region Proposal Network）网络，和目标检测网络共享卷积层特征，大大节省了生成候选窗的时间。RPN 对其输入的候选窗进行属于背景还是前景的判断以及检测框位置的修正（Bounding Box Regression），其输出的检测框输入给检测网络，做最终的分类和更精准的检测框位置修正。Faster R-CNN 使用 VGG-16 网络模型在 K40 GPU 上进行目标检测任务时，运行速度能到 5 帧每秒，在 PASCAL VOC 2007 上 mAP （mean Averaged Precision）达到 73.2%，在 VOC 2012 上也达到了 70.4%。图 14 所示是 Faster R-CNN 的检测效果。

除了基于 Region Proposal 的深度学习目标检测算法，还有直接使用 Deep CNN 进行 end-to-end 定位的目标检测技术，如 YOLO[51]，DenseBox[52]等，在人脸、车辆、行人等目标检测任务上取得了很好的效果。此类检测技术对需要检测的图像通常可以直接计算出目标的类别和位置。图 15 展示的是 DenseBox 在 KITTI 数据集的车辆检测结果。

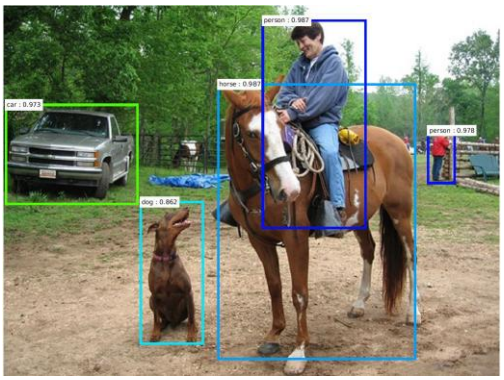


图 14 Faster R-CNN 的检测效果[50]

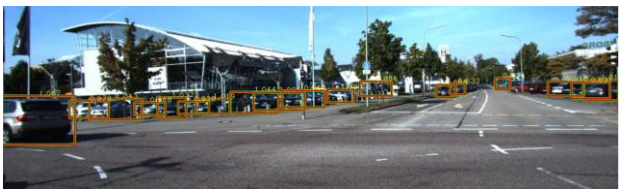


图 15 DenseBox 在 KITTI 数据集的车辆检测结果[52]

3.3 图像语义分割

在过去几年中，随着计算机视觉、机器学习等领域研究的不断深入，研究人员逐渐将目光投向对图像本身更为精准的理解与分析。图像语义分割（Image Semantic Segmentation）问题正是为了满足这一要求提出的，它通过解析训练图像的内容，在分割图像的同时获得图像的所有分割区域甚至每个像素的语义类别，从而获得图像基于内容的标注。图像语义分割不仅需要对图像分割区域的边界做出精准识别，而且要求对分割区域的目标类别进行准确识别。精准的图像语义分割不仅能够有效降低后续的图像分析与识别、语义检索等高层次任务处理的数据量，同时又能保留图像的结构化信息[53]。图像语义分割常用的数据集有 MSRCv2④，PASCAL VOC2012[54]，Microsoft COCO[55]，PASCAL-CONTEXT[56]，Sift Flow[57]等。图像语义分割常用的评价指标是计算预测的语义类别和正确的语义类别像素点的重合度（Intersection Over Union, IOU），重合度越高说明模型的准确度越高。

传统的图像语义分割方法通常包含三个部分：第一部分主要进行图像的底层分割，将图像划分成多个子区域；第二部分提取子区域的底层特征，如颜色、纹理、形状等；第三部分学习从

④ Criminisi T. M. A., Winn J.. Microsoft research cambridge object recognition image dataset, version 2.0. http://research.microsoft.com/en-us/projects/objectclassrecognition, 2004

底层特征到高层语义空间的映射，根据学习好的映射模型标注图像，识别出图像区域乃至每个像素的语义类别。主要代表性工作如 Shotton 等人[53]提出的 TextonBoost 方法使用提升决策树分类器，在所有图像像素构成的条件随机场（Conditional Random Field，CRF）中，以纹理布局滤波器学习每个像素的单点势能，像素语义标注间具有平滑性约束，通过最小化随机场能量，得到像素级语义标注。其后的许多工作对条件随机场的势能函数提出了不同的改进方法[58-62]。比如 Shotton 等人[62]提出的 TextonForest 方法使用了一种基于随机森林（Random Forest）的特征，几乎能实现实时的图像语义分割。

随着深度卷积神经网络在图像检测、分类等多个任务上成功应用，目前已经有不少研究人员将 Deep CNN 应用到图像语义分割领域[63-65]，如 Farabet 等人[63]使用多尺度卷积神经网络从不同大小的像素和超像素学习目标特征，极大地提升了语义分割效果，在 PASCAL VOC2012 分割数据集上达到 62.2%的 IOU 精度。Lonjong 等人[66]在 CVPR 2015 上提出的全卷积网络（Fully Convolutional Network，FCN）能够端到端（end to end）得到每个像素的目标分类结果。与经典的 CNN 输入固定大小图像、卷积层之后使用全连接层得到固定长度的特征向量不同，FCN 可以接受任意尺寸的输入图像，且全部使用卷积层。FCN 采用反卷积层对最后一个卷积层的特征图进行上采样，使特征图恢复到输入图像相同的尺寸，从而可以对每个像素都产生了一个语义预测。在此过程中保留了原始输入图像中的空间信息，最后在上采样的特征图上逐像素计算 softmax 分类的损失。图 16 是用于语义分割所采用的全卷积网络（FCN)的结构示意图。

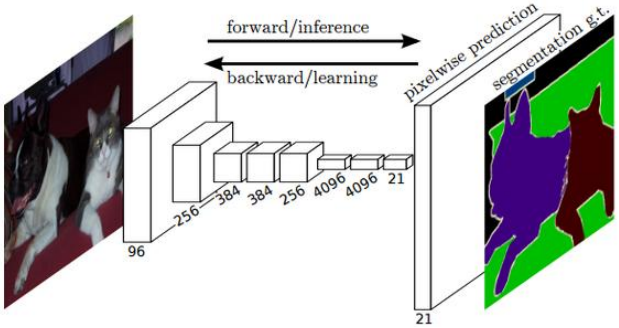


图 16 全卷积网络的结构示意图[66]

FCN 虽然在图像语义分割取得不错的效果，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。Chen 等人提出的 DeepLab[67]应用文献[68]中提出的全连接 CRF 模型，对 FCN 的输出结果作进一步的细粒度处理，在 PASCAL VOC2012 分割数据集上达到 71.6%的 IOU 精度。Zheng 等人[69]提出的 CRF-RNN 将全连接 CRF 的学习、推理过程看成是个递归神经网络（Recurrent Neural Network, RNN），并且嵌入到 FCN 模型中，完成了端到端的训练、预测。该方法相较于 FCN，可以较好地解决图像边缘信息丢失的问题，对边界分割精度有很大的提升。该方法在 PASCAL VOC2012 分割数据集上的平均 IOU 精度达到 74.7%。图 17 是 CRF-RNN 与 FCN、DeepLab 在 PASCAL VOC2012 分割数据集上的语义分割结果展示。

为了训练能识别图像所以分割区域甚至每个像素的目标分类器，大多数图像语义分割方法需要使用大量精确的像素级标注数据作为训练数据。然而，因为标注工作非常耗时，这类数据非常有限。根据 Microsoft COCO 数据[55]标注经验，精确标注每个像素点的耗时平均是标注目标检测框的 15 倍。为了克服像素级标注的约束，部分科研工作者考虑设计新的语义分割算法。Dai 等人[70]提出的 BoxSup 是以图像检测框的标注信息作为监督信号。BoxSup 首先使用非监督的候选区域生成方法产生初步分割结果，然后进一步使用检测框和FCN得到的基于像素点的监督信息。Bearman等人[71]使用了表示物体的点作为监督信号，通过利用上述监督信息设计惩罚函数，约束训练 FCN 的损失函数。Pinheiro 等人[72]提出了一种基于图像标签的弱监督语义分割算法，在训练过程中设计 CNN 网络使关键像素点被赋予较大权值，从而实现图像中各像素更为准确的标注（如图 18 所示）。

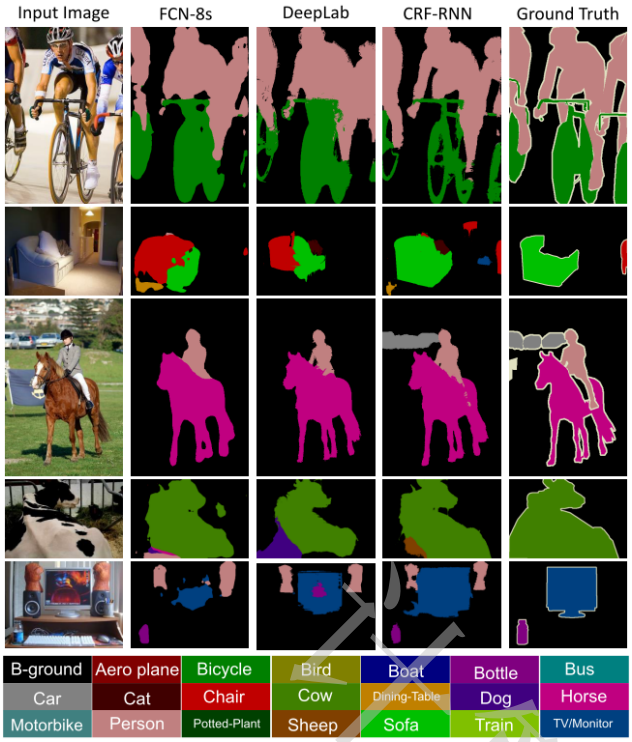


图 17 在 PASCAL VOC2012 分割数据集上 CRF-RNN 与 FCN、DeepLab 的语义分割结果比较[69]

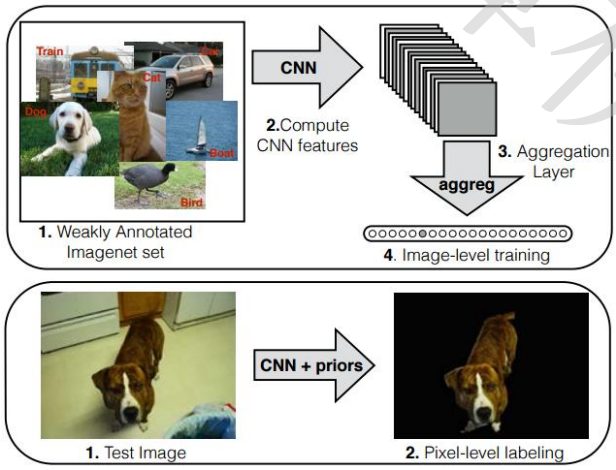


图 18 Pinheiro 等人提出的基于图像标签的弱监督语义分割算法[72]

3.4 图片标题生成

图片标题生成（Image Captioning）技术，指自动产生自然语言来描述一副图片的内容。随着深度学习和自然语言理解领域相关技术的突破，图片标题生成技术在 2014-2016 年获得了迅猛的发展。在 2015 年微软 COCO 图片标注竞赛中，来自微软[73][74]、谷歌[75]、多伦多大学和蒙特利尔大学[76]、加州大学伯克利分校[77][78]等研究机构的最新工作都取得了令人惊叹的成绩。谷歌（基于 CNN 视觉特征和 RNN 语言模型）和微软（基于区域的单词检测和最大熵语言模型）目前在技术和性能方面处于领先地位[79]。

一部分图片描述工作使用流程化方法来描述图片内容。Fang 等人[73]将图片描述过程分为三步，如图 19 所示。首先利用多示例学习（Multiple Instance Learning, MIL）方法，根据图片各个部分提取的 CNN 特征产生相对应的名词、动词和形容词；然后使用最大熵语言模型（Maximum Entropy Language Model, MELM）产生图片标题；最后使用最小错误率训练（Minimum Error Rate Training, MERT）对所产生的可能性最高的几组句子进行打分并排序。Kiros 等人[80]利用 CNN 和 LSTM（Long short-tem memory network）对图片进行编码，然后利用论文中提出的 SC-NLM（Structure-Content Neural Language Model）预测句子结构来实现解码。

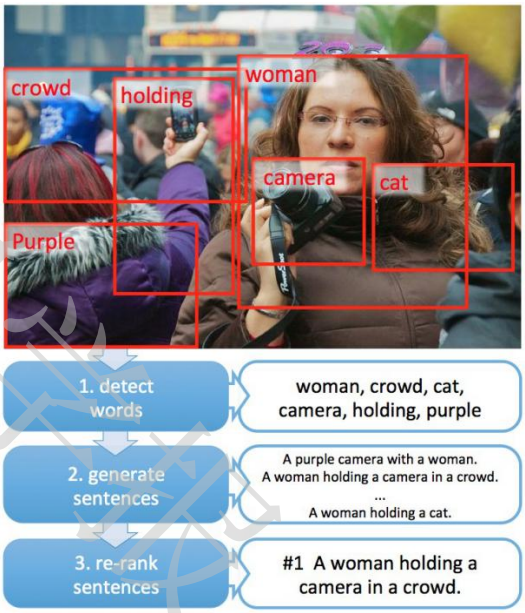


图 19 Fang 等人提出的流程化方法生成图片标题[73]

与以上使用流程化的方法不同，另一部分图片描述工作使用端到端方法。Vinyals 等人[75]受机器翻译技术的启发，利用 CNN 模型提取图片特征，再利用 RNN 模型生成图片标题，如图 20 和 21 所示。Karpathy 等人[77]和 Mao 等人[81]提出利用 mRNN（Multimodal Recurrent Neural Netowrk）模型生成图片标题。不同于将图片和文字映射到同一空间，Chen 等人[82]在图片和文字描述之间直接建立双向映射关系。Donahue 等人[78]提出的 LRCNs（Long-term Recurrent Convolutional Networks）模型直接在可变长度的图像序列输入和可变长度的文字输出之间建立映射关系。Xu 等人

[76]提出将视觉注意模型融合进 LSTM 模型，从而在单词生成过程中能更好关注图像中的显著目标。最近，Jia 等人[83]则利用 gLSTM（Guiding Long Short-term Memory）模型，在 LSTM 模型的基础上引入外部的语义信息生成图像标题。

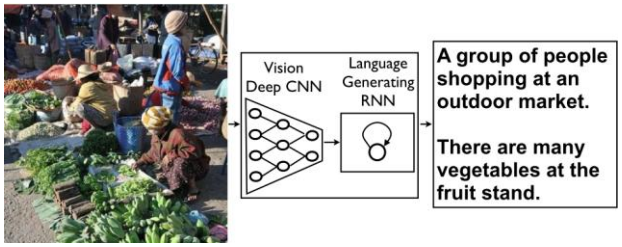


图 20 Vinyals 等人提出的端到端方法生成图片标题[75]

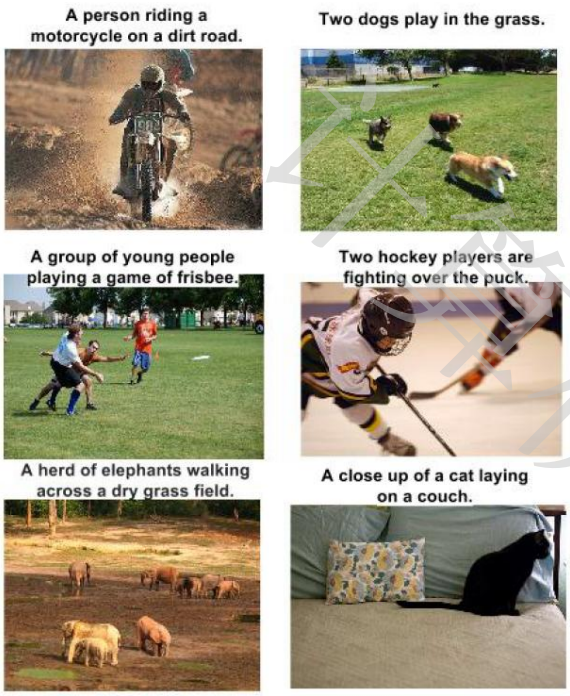


图 21 Vinyals 等人提出的图片标题生成方法效果展示[75]

表 5 给出了不同图像标题生成方法在生成图像标题性能的结果比较，评价指标采用了 BLEU 量度[84]。从表 4 中我们看到 Hard-Attention 和 gLSTM 在 MSCOCO 数据上达到最好的性能。

表 5 不同图片标题生成方法在 MSCOCO 上的性能比较

方法	B@1	B@2	B@3	B@4
Multimodal RNN[77]	62.5	45.0	32.1	23.0
Google NIC[75]	66.6	46.1	32.9	24.6
LRCN-CaffeNet[78]	62.8	44.2	30.4	
m_RNN[84]	67.0	49.0	35.0	25.0
Soft-Attention[76]	70.7	49.2	34.4	24.3
Hard-Attention[76]	71.8	50.4	35.7	25.0
gLSTM[83]	67.0	49.1	35.8	26.4

3.5 人脸识别

计算机视觉领域一个重要的挑战问题是人脸识别。人脸识别包含两种任务，人脸验证和人脸辨识。人脸验证的任务是判断两张人脸照片是否属于同一个人，属于二分类问题，随机猜的正确率是 50%。人脸辨识的任务是将一张未知人脸图像分为 N 个身份类别之一，这是个多分类问题，随机猜的正确率是 1/N。人脸辨识更具有挑战性，其难度随着类别数的增多而增大。人脸识别的最大挑战是如何辨别由于光线、姿态和表情等因素引起的类内变化，以及由于身份类别不同产生的类间变化。这两种变化分布极为复杂且都是非线性的，传统的线性模型无法将它们有效区分开。卷积神经网络可以通过多层的非线性变换，尽可能多地去掉类内变化，同时保留类间变化。

LFW (Labeled Faces in the Wild)[85]是当今最著名的人脸验证公开测试集，它是从互联网上收集了五千多个名人的人脸照片，用于评估算法在非可控条件下的人脸验证性能（如图 22 所示）。在 LFW 测试集上，人眼的正确率是 97.53%[86]，而非深度学习算法的最高正确率是 96.33%[87]，而目前深度学习可以达到 99.47%的验证率[19]。目前，许多人脸识别算法都是在包含大量人脸类别的离线数据集上，以人脸辨识的任务通过神经网络模型学习人脸特征，得到的特征再应用于人脸验证任务。

图 22 LFW 人脸数据集[85]

2013 年，Sun 等人[88]采用人脸确认任务作为监督信号，利用卷积神经网络学习人脸特征，在 LFW 上取得了 92.52%的识别率。这一结果虽然与后续的深度学习方法相比较低，但也超过了大多数非深度学习的算法。在 CVPR 2014 上发表的 DeepID[15]和 DeepFace[89]，采用人脸辨识作为监督信号，在 LFW 上取得了 97.45%和 97.35%的识别率。他们利用卷积神经网络预测输入人脸图片的类别，选取最高的隐含层作为人脸特征（如图 23 所示）。在训练过程中，神经网络需要区分大量的

人脸类别(例如在 DeepID 中要区分 1000 类人脸),因此人脸特征包含了丰富的人脸类间变化信息,而且有很强的泛化能力。

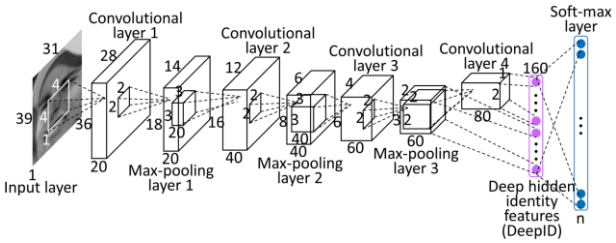


图 23 DeepID 网络结构[15]

DeepID2[16]联合使用人脸确认和人脸辨识作为监督信号,得到的人脸特征在保持类间变化的同时最小化类内变化,从而将 LFW 上的人脸识别率提高到 99.15%。利用 Titan GPU,DeepID2 提取一幅人脸图像的特征只需要 35 毫秒,而且可以离线进行。经过 PCA 压缩最终得到 80 维的特征向量,可以用于快速人脸在线比对。在后续的工作中,DeepID2+[90]对 DeepID2 通过加大网络结构,增加训练数据,以及在每一层都加入监督信息进行了进一步改进,在 LFW 达到了 99.47%的识别率。FaceNet[19]提出了使用 Triplet 网络结果学习人脸特征,输入样本以两张同类图片和一张不同类图片的方式,在最后一层隐藏层直接使用欧氏距离来度量输入图像之间的相似度。FaceNet 在 LFW 数据集上验证精度达到 99.63%。

3.6 行人再识别

行人目标是监控视频中最为常见也最为关注的目标,对多个监控视频环境下行人目标的检索问题常称为行人再识别(Person Re-identification)问题。在可控的环境下,依靠人脸、虹膜等生物特征进行人再识别已经是较为成熟的技术。然而,监控视频的环境通常非常复杂而且含有很多不可控因素(如低分辨率、遮挡、运动模糊、复杂背景等),其获取的行人图像质量通常很差,因此较难提取到鲁棒的人脸特征。因此,绝大部分研究人员通过行人穿的衣服和携带的物品等外貌特征来实现行人再识别。由于不同监控视频下的行人存在尺度、视角及光照等差异,可能导致不同监控视频中,不同行人目标的外貌特征比同一个行人目标的外貌特征更相近。随着视频监控领域应用需求的增长,许多研究人员对行人再识别技术进行了深入研究。目前广泛使用的公开数据库有 VIPeR[91]、ETH-Z[92]、CUHK[93]、PRID2011[94]、i-LIDS[95]等。

已有的行人再识别算法大致分为两类:基于距离度量学习的方法和基于特征描述的方法。基于距离度量学习的方法是学习度量行人目标特征分布的距离函数,即不同行人目标的特征距离值较大,而同一个行人目标的特征距离值较小。基于特征描述的方法是设计可靠、鲁棒、具有判别性的行人图像特征,即能够有效区分不同的行人目标,且能不受尺度、视角及光照等变化的影响。

随着 Deep CNN 在计算机视觉与图像识别领域成功应用,近年来许多学者利用 Deep CNN 来解决行人再识别的问题,并且在公开的数据集上取得了最好的测试结果。例如,Ding 等人[96]提出了一种三通路的网络结构,利用 Triplet Loss 监督网络的学习过程,在小数据集上取得了很好的效果。DeepReID[97]提出了一种 Filter Pairing Neural Network (FPNN)来处理图像未对准、光度和几何变换、遮挡和复杂背景等问题,提高了算法的鲁棒性。mFilter[98]使用局部图像块匹配的方法学习局部特征,增强了特征之间的判别能力。Ahmed 等人[99]提出了一种改进的深度网络结构,网络输入一对行人的图像,输出两张图像的相似性。Yi 等人[100]通过一个 Siamese 网络学习两张图像的相似性,在每个通道中每张输入图像被等分成三份来训练网络,训练好的网络具有很强的泛化能力。Cheng 等人[101]在 CVPR 2016 上提出利用深度卷积神经网络分别从全局和局部两个不同的角度对行人的特征进行学习(如图 24 所示),学习得到的模型对光照、

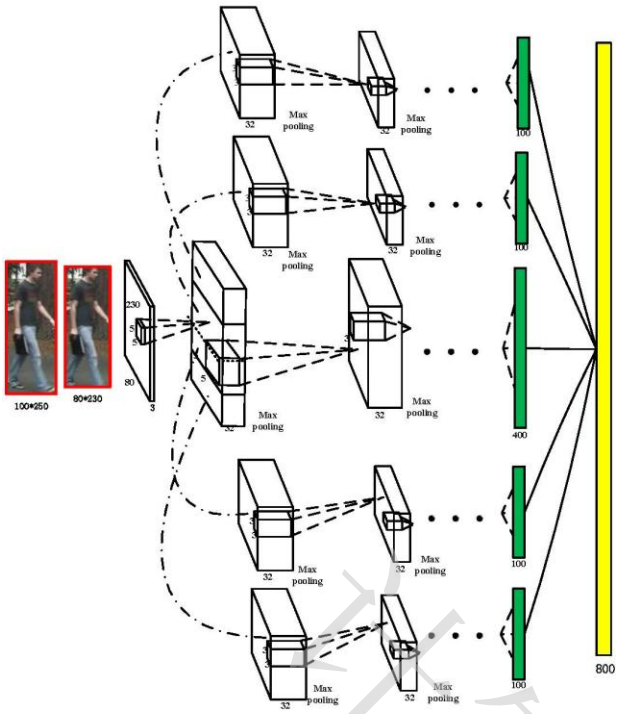


图 24 Cheng 等人提出的多通道 Deep CNN 网络结构[101]

视角、分辨力等影响因素具有很强的鲁棒性。该研究成果在业界公布的标准数据集的测试中取得了最好的结果，例如：在 PRID2011 数据集上领先最好结果 4.1 个百分点，在 VIPeR 数据集上领先最好结果 7.28 个百分点，在 i-LIDS 数据集上领先最好结果 8.3 个百分点。

3.7 图像超分辨率

指从一幅低分辨率图像或图像序列恢复出高分辨率的图像或图像系列。更高的图像分辨率，更精细的细节意味着图像提供的信息越丰富。在军事侦察、医学诊断等许多实际应用中，高分辨率的图像显得尤为重要。从低分辨率图像复原高分辨率图像是一个欠定的病态问题。对于这一病态问题，通常采用引入各种先验(比如光滑先验，梯度先验等等)来约束图像超分辨过程。宽泛地可以将已有方法分为三类，基于插值[102]、基于重建[103,104]和基于学习的方法[105-107]。基于超分辨率算法会在自然图像数据集 Set5[108]，Set14[109]上进行测试。图像超分辨率通常通过客观评价指标比如 PSNR，SSIM 来衡量算法优劣。由于客观评价指标有时无法很好和人的主观评价项一致，人对图片的主观评价往往更加重要。

2014 年，Dong 等人[110]首次提出了使用深度卷积神经网络学习低分辨率图像和高分辨率图像之间端对端的映射关系，进行图像超分辨率。该工作方法利用深度卷积网络强大的非线性学习能力，设计了包含三个卷积层的深度卷积网络，通过输入大量的数据样本来训练模型，进而得到比较理想的高分辨率图像，具有更好的主观效果。在当时与最好算法相比（放大三倍时）精度在 set5 上 PSNR 提高了 0.47db，在 set14 上 PSNR 上提升了 0.33db。之后 Dong 等人[111]进一步证明加大数据量，增加训练时间，可以有效改善训练的模型，算法进一步 PSNR 提高 0.3db。

针对深度卷积神经网络训练慢，时间长的问题，Liang 等人[112]提出了结合图像的先验知识对图像超分辨率映射的学习过程施加约束，监督超分辨率映射的学习过程。该方法是在原有超分辨率网络引入了一个额外的特征提取层（见图 25），通过提取学习图像的先验信息（梯度）来指导高分辨率图像的重建过程。加入图像先验特征大大加速了网络的学习过程（将近 10 倍的提速），并得到了更好的图像的超分辨率结果。2015 年 Wang 等人[113]将稀疏先验引入深度卷积神经网络的设计中，利用将深度卷积神经网络级联的方法，重新提升了超分辨率结果(相比最好结果 0.1db 的提升)，同时拥有更好的收敛速度。在主观评价试验中，该方法也得到了压倒性的优势。

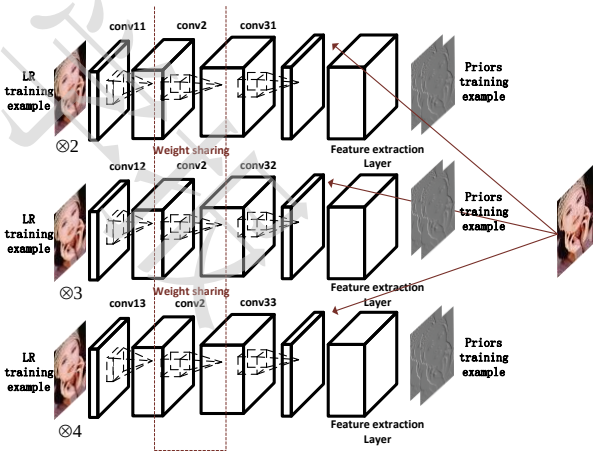


图 25 结合图像先验的超分辨率多通道深度卷积神经网络结构[112]

3.8 人体动作识别

基于视觉的人体动作识别是当前计算机视觉领域的一个热点问题，其主要目标是通过对摄像机获取的视频数据进行处理和分析，识别并理解视频中人的动作和行为。基于视觉的人体动作识别过程通常包含以下三个步骤：首先从图像序列中检测运动信息并提取图像底层特征；其次是对

人的动作或行为模式进行建模；最后建立图像底层视觉特征和动作行为类别等高层语义信息之间的对应关系。

按照对时序信息的使用程度，基于视觉的人体动作识别方法可以划分为两类：时空特征方法和时序推理方法。基于时空特征的人体动作识别方法在视频序列中提取有效的动作特征，这类方法主要解决简单动作的识别，可以分为基于局部特征[114-116]、基于时空体模型[117,118]和基于时空轨迹特征[119,120]等方法。基于时序推理的人体动作识别方法在简单行为识别结果基础上考虑到简单行为之间的时序关联，可以分为统计模型方法[121,122]和句法模型方法[123-125]。目前国内外有多个公开人体动作数据库用于验证不同动作识别算法的性能，常用的有：Weizman[126]、KTH[127]、Hollywood[128]、UCF101[129]和 HMDB-51[130]等。

利用卷积神经网络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这一任务中。Ji 等人[131]提出了一种 3D CNN 人体动作识别方法，通过执行三维卷积计算，在相邻连续图像序列上从时间和空间两个维度提取动作特征。3D CNN 以多个通道从输入图像帧采集特征信息，最后合并多个通道的特征信息形成最终的动作特征。Varol 等人[132]提出了 Long-term Temporal Convolutions (LTC) 操作，与 CNN 结合能在更长时间窗口内使用 3DCNN，进一步提高了动作识别的性能。Cheron 等人[133]提出一种基于静态图像姿态（Pose-based CNN，PCNN）的人体动作识别方法。该方法在视频序列中跟踪人体及人体每个部位，并在每张静态图像中提取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，生成整个动作的外观和光流特征描述。Karpathy 等人[134]利用局部时空信息构建一个多尺度的卷积神经网络，以提取视频中人体的动作信息。如图 26 所示，网络输入是两路不同尺度大小的视频：原始分辨率的视频和对视频中心裁切的高分辨率视频。两路视频经过相同的卷积层、正则层和采样层后，最后合并成两个全连接层。该方法在 UCF101 动作识别数据集提高了近 20 个百分点。Simonyan 等人[135]提出的人体动作识别方法使用双路卷积神经网络的架构，分别处理静态帧图像和多帧图像。其中，静态帧图像使用单帧图像信息以提取动作在图像的空间信息，多帧图像使用多帧图像的光流信息以提取动作的时间信息。两路图像数据都经过深度卷积神经网络提取特征，然后使用 SVM 分类器识别具体动作。

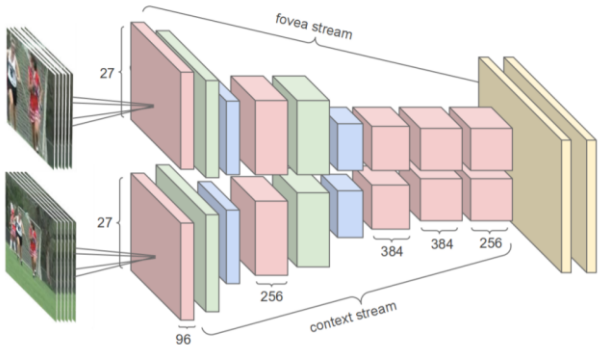


图 26 双通道多尺度深度卷积神经网络结构[134]

3.9 图像检索

随着数码相机、手机等数字化设备的广泛普及，存储技术和网络共享技术的快速发展，网络上的图像资源以爆炸性速度快增长。图像检索研究的热点问题就是如何从大规模图像数据库中准确、快速地检索到相同或相似主题的图像数据。

图像检索通常划分为两个方向：基于文本的图像检索（Text Based Image Retrieval,TBIR）[136-138]和基于内容的图像检索（Content Based Image Retrieval，CBIR）[139-141]。TBIR 简称“以字搜图”，其目标是在图像数据库中查询与检索关键字相关的图像。目前谷歌、雅虎、百度等搜索引擎主要采用的是 TBIR 技术，该技术需要事先对大规模图像数据库中的每张图像进行关键词标注，因此依赖于高精度的图像自动标注算法。CBIR 简称“以图搜图”，是直接利用图像的视觉特征进行分析和检索图像，其重点是对检索图像与数据库中图像进行相似性匹配。CBIR 技术的特点是利用图像自由的客观视觉特性，不需要人工进行干预，利用计算机可以对图像特征进行自动提取和存储。目前出现了许多基于 CBIR 技术的检索引擎，如 Picitup, TinEye ,谷歌的按图搜索，百度识图等。

传统的 CBIR 方法采用手动设计的特征（如 SIFT、SURF、HOG、GIST 等）和利用词袋（Bag-of-Words，BoW）模型表示图像，并使用向量之间的距离计算图像间的相似性，从而得到初步检索结果。近几年，许多新的 CBIR 方法利用卷积神经网络学习更具有表达能力的图像特征。

Xia 等人提出了一种名为 CNNH（Convolutional Neural Network Hashing）的有监督深度哈希算法

[142]。CNNH 利用训练图像之间的相似性分解相似性矩阵，得到训练图像的二值编码，进而利用卷积神经网络对所获得的二值编码进行拟合，同时学习得到更鲁棒的图像特征。Wan 等人[143]同样利用卷积神经网络进行特征学习用于解决图像检索问题，作者评价了不同卷积神经网络学习策略对图像检索性能的影响。论文[144]提出了一种称为 DSRH（Deep Semantic Ranking Hashing）的方法，将图像检索任务转化为解决图像相关性排序问题。在 DSRH 中，深度卷积神经网络用于学习图像的特征表示，同时将所学到的特征映射到哈希码。DSRH 的结构框图如图 27 所示。Lai 等人[145]提出的图像检索方法同样利用深度卷积神经网络同时进行特征学习和哈希编码。该方法设计了一种分离编码模块，将图像特征划分为几个部分，每个部分负责学习哈希码中的一位。网络使用 triplet 损失函数进行学习。Liu 等人[146]提出的快速图像检索算法利用图像对的监督信息，提出了一种正则项对深度卷积神经网络进行约束，使神经网络的输出接近二值编码，增加了图像检索效率。

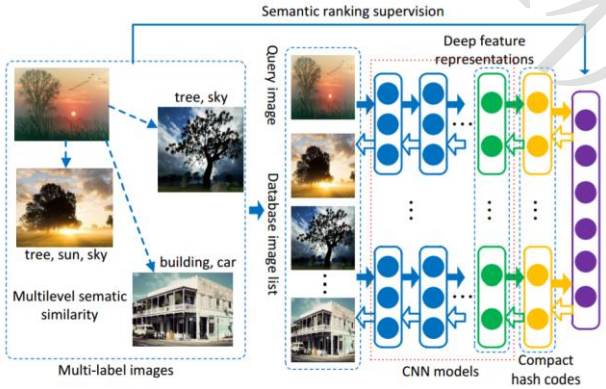


图 27　DSRH 图像检索方法的流程示意图[144]

4 视觉认知的理论启示

人类视觉系统是至今为止所知的功能最强大和完善的生物视觉系统，是人脑感知外部环境的最主要方式，人类获取外部世界的信息约 70%来源于视觉。利用非凡的脑信息处理能力，人类能够快速高效的从客观世界的杂乱场景中抽取有效信息，分析感兴趣的目标或区域，形成对视觉场景内容的高度理解和认知。神经网络的研究与人类视觉的研究密切相关，借鉴人类视觉认知机制的相关计算理论，是未来研究提升神经网络性能的一个方向。

4.1 视觉信息分层处理

1958 年，约翰霍普金斯大学的 David Hubel 和 Torsten Wiesel 研究发现，在初级视皮层中存在两种细胞：简单细胞和复杂细胞，这两种细胞承担不同层次的视觉感知功能[1]。他们的研究还发现，视觉系统的信息处理——可视皮层是分级的[147]。视觉信号传递到初级视皮层（V1 区）之后，低级的 V1 区提取边缘特征，到第二视区（V2）提取目标的局部形状或者目标的部件，再继续向更高级的视觉皮层传递并获取图像的整体形状。高层特征是对低层特征的聚类，高层的特征表示较于低层特征更为复杂、抽象，且更能表现语义信息或者目标类别，因而人脑能够理解十分复杂和抽象的内容。

除了 Hubel-Wiesel 模型的发现，Mishkin, Ungerleider 和 Macko 于 1983 年在猴子的纹状体皮层上发现视觉信息在皮层的逐级传递中可以大体分成两个通路，而 V1 皮层是两条通路的发源地[148-150]。这两条通路一条通向腹侧，被称为腹侧通路（Ventral Stream，如图 28 所示[151]），沿着大脑皮层的枕颞叶分布，包括纹状体皮层、前纹状体皮层和下颞叶。另一条通向背侧，被称为背侧通路(Dorsal Stream)，沿着枕顶叶分布，包括纹状体皮层、前纹状体皮层和下顶叶。背侧通路主要负责处理视觉刺激的位置、运动、三维结构等信息，而腹侧通路则负责提取不变性特征，实现对物体及场景种类的识别。研究结果进一步表明，腹侧通路主要由 V1、V2、V4、IT 这四层脑区构成。IT 脑区又可以进一步分割为 PIT、CIT 及 AIT 三个子层（参见图 28）。视觉刺激通过人眼中的视网膜被转换成神经信号，通过 LGN 细胞进行简单的预处理后到达腹侧通路的 V1 层。经过 V1 到 IT 脑区的逐层处理，由视网膜传来的原始神经信号被逐渐转换成具有高度抽象度及区分能力的图像特征。一般认为，IT 层中的图像特征表达对物体的位移、旋转、大小、姿态、视角、光照等变化已经拥有相当程度的不变性，已经具备准确、快速地识别物体及场景类别的能力[152-154]。视觉皮层分为腹侧和背侧两条通路的理论，曾是早期认知科学领域的一大成就。

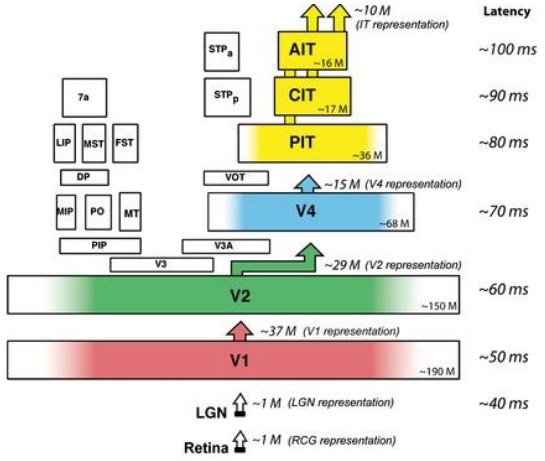


图 28 人脑腹侧通路[151]

基于 Hubel-Wiesel 模型实现的卷积神经网络基本模拟人类视觉系统的信息分层处理方式。卷积层、取样层分别对应一种简单细胞、复杂细胞；越是上层的卷积层，学习的图像特征越复杂；越是上层的取样层，学习的图像特征越具有不变性。然而，人脑视觉认知过程比 CNN 要复杂得多，而前者的许多特性 CNN 多不具备[151,155,156]。

4.2 “大范围优先”的视觉认知过程

基于以上生理学事实，以 Treisman 的特征整合理论[157]以及 Marr 的计算视觉理论[158]为代表的“局部优先”观点，认为视觉图像起初被分解为基本的成分和单元，然后单个并行处理，最后整合到一起对整个图像内容进行识别，即视觉认知过程是先识别局部而后识别整体内容。然而越来越多的研究表明，视觉认知过程始于视觉系统的顶层区域，遵循先整体后局部的顺序。

与当时流行的从局部到整体的“局部优先”思想相反，D. Navon 于 1977 年提出了著名的“大范围优先”理论[159]。Navon 使用复合刺激来描述图形的整体和局部性质，每个复合刺激是由许多小字母组成的大字母图形，如图 29 中的 A 和 D 是大字母和小字母形状一致，而 B 和 C 则形状不一致。Navon 将小字母的性质描述为图形的局部性质，大字母的性质描述为整体性质。在视觉辨别实验中，被试需要辨别大字母和小字母是 H 还是 S。实验发现，被试辨别大字母的反应时（reaction time, RT）明显比小字母的 RT 短，而被试在辨别小字母时当大小字母一致时 RT 较短，不一致时则 RT 较长（大字母对小字母有干扰作用)；相反，小字母对辨别大字母的 RT 几乎不不造成影响。根据以上实验结果，Navon 认为视知觉系统首先处理大范围整体性质，然后再加工局部性质。后来的心理学家对强调整体性质的大范围优先性理论进行了深入的研究并对其进一步完善。

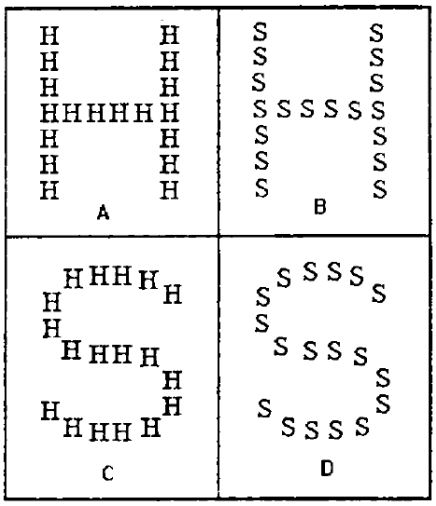


图 29 Navon 使用的复合刺激图形[159]

“大范围优先”理论认为人脑视觉认知总是遵循先整体后局部的顺序。当一幅图呈现给被试时，被试最初（100ms 以内）所认知的只是图像的全局内容，如图像中有无物体出现，物体的类别等）。这个认知过程称作 Spread Attention，负责对图像全局内容的识别。被试对图像细节的认知一般发生在 250ms 以后，根据认知任务的难易度，需要不断移动眼球，这时对图像全局的认知度会大幅下降。这个认知过程称作 Focused Attention。那么视觉系统是对图像全局认知后是如何再进行局部认知呢？是否在全局认知脑区上层存在更高层级脑区负责局部认知？

1982 年，我国学者陈霖在《Science》杂志上发表了题为《视知觉的拓扑结构》论文[160]，首次创新性提出了“大范围优先”的拓扑知觉理论，并在此后的 30 年时间里，用令人信服的实验不断完善和论证这一理论，使之被越来越多的认知科学研究者所接受。陈霖认为，在初期阶段视知觉系统对大范围的拓扑性质更敏感，而不是局部特性。陈霖用论文和实验[161-163]解释了视觉系统的一个基本功能是感知拓扑性质，并且揭示了图形知觉从大范围到局部的几何层次感知顺序，其代表性成果发表在 2003 年《Science》[164]，2007 年《PNAS》[165]，2010 年《PNAS》[166]等期刊上。

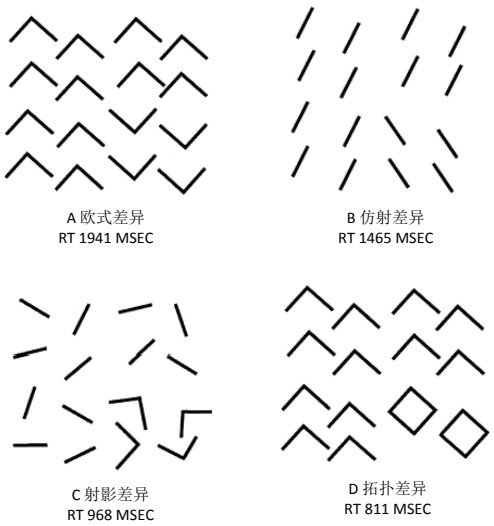


图 30　四选一实验的反应时（RT）结果，对拓扑差异的识别时间要短于非拓扑差异的识别时间[163]

支持拓扑知觉理论的其中一个实验是如图 30 所示的四选一实验[163]。按几何学的不变性分类，图中 A、B、C、D 分别代表了反映欧氏性质、仿射性质、射影性质以及拓扑性质的差异的四种不同刺激。每个刺激由四个象限组成，每个象限中只包含一种形状的小图形，其中三个象限中的图形形状一致，而第四个象限中的图形与其他三个不一致。被试要求在保证正确的前提下快速找出形状不一样的象限位置。通过比较检测这些性质的反应时，我们可以发现，在视觉系统中，拓扑性质差异被最先检测出来，其次是射影和仿射性质，最后才是几何性质最不稳定的欧氏性质被检测出来[167]。

2002 年，S. Hochstein 和 M. Ahissar 提出了一种视觉认知学习的自下而上和自上而下的双信息处理过程[168,169]，如图 31 所示。他们认为在采集到视觉信号后，视觉系统一开始进行自下而上的多层分级的视觉信息处理，这一过程是无意识的、自动发生的。视觉的全局认知发生在视觉系统的高级区域（如图 28 中 IT 脑区），它是利用从 V1、V2、再到 V4 的逐层特征提取所取得的、具有高度抽象性及不变性特征来实现的。IT 层的特征具有高度抽象性以及语义或者类别表达能力，但不具有输入图像的局部细节表达，因而只适用于视觉的全局认知。当视觉系统需要进行局部细节认知时，高级区域沿自上而下的信息通路逐层展开认知，并且认知过程是以先拓扑性质、射影性质、再到仿射性质、欧式性质的顺序进行的。这个理论解释了上述拓扑知觉试验中，为什么拓扑性质差异被最先检测出来，其次是射影性质和仿射性质，最后才是欧氏性质被检测出来这个视觉系统的认知特性。他们发现，这个理论在解释大量研究结果时（包括似是而非的研究数据）被证明很有用。

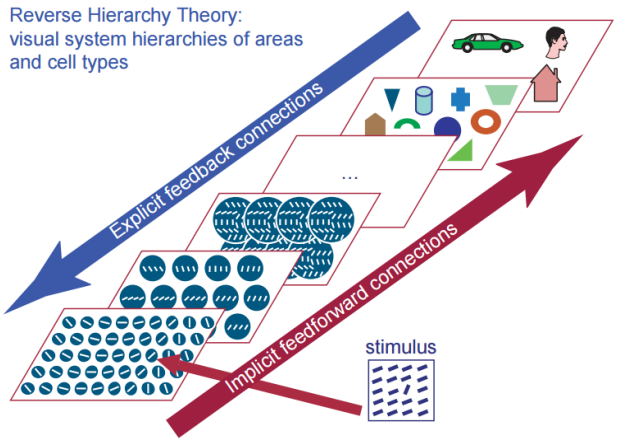


图 31　人脑在视觉认知过程中自下而上和自上而下的双向信息处理通道[168]

5 展望

人工神经网络是由基本的数学计算单元及其交互联接构成的一种网络计算结构，用来模拟人脑中信息的处理过程，让机器通过学习训练机制主动获取数据中所蕴含的规律。本文围绕其中的一种学习模型——深度卷积神经网路，介绍了现阶段提升深度卷积网络性能的技术方法和在计算机视觉领域内的应用，并分析了人脑视觉机制的特点和对当前计算模型的一些理论启示。

尽管当前深度卷积网络较传统机器学习方法有了很大的提高，但不可忽略的是，他们与人脑视觉系统还是有非常大的差距，从根本上并没有解决视觉认知的根本问题。未来基于深度卷积神经网络的类脑智能研究仍有许多亟待解决的问题与挑战：

（1）借鉴视觉认知的研究成果，改进神经网络的模型结构

借鉴人脑视觉系统的特性去研究和改进已有神经网络的结构，让机器获得更高层次的类脑智能，是未来研究的其中一个重要研究方向。现有的神经网络都是借鉴人类视觉系统自下而上对图像进行全局内容识别的特性，对输入图像进行特征提取的过程均为一个单向过程，但人脑对于输

入图像的特征提取和认知过程是一个包含自下而上和自上而下的双向迭代过程。如何模拟人脑视觉系统自上而下识别图像局部细节的特性[168-170]，改善现有神经网络结构，以提高检测、定位、分割等任务的精度，值得进一步研究。

（2）基于无监督式特征学习的研究

迄今为止，深度学习中的监督式特征学习取得了非常大的成功，但是监督式特征学习算法的训练过程往往依赖于百万级以上的标注数据，通常需要花费很多的人力物力完成数据标注。然而，在人类和动物的学习过程中，无监督式学习一直占主导作用：我们通过观察和亲身体验来发现世界，而并不需要其他人告诉我们每一件事物的名称。

近几年，虽然众多研究人员开始关注无监督学习这一领域，有关无监督特征学习算法的研究取得了一定的成果，但其对特征进行高效表达的能力相对于监督式特征学习算法仍差距尚远。如何才能使机器具备像人类和动物一样仅仅通过观察世界就能获取常识的无监督学习能力，成为未来的一个重要发展方向[171]。

（3）利用海量增加的数据进一步提高卷积神经网络的特征学习能力

深度学习取得成功的一大关键因素是网络上海量可用的数据。当前，在工程应用及生物神经领域存在有指数增长的海量复杂数据，以文字、图片、视频、音频、基金数据等不同模态呈现出来，具有绝然不同的数据分布。这对神经网络模型的训练复杂度、参数选取、结构设计、时间复杂度等方面的平衡都带来了新的挑战。因此，如何充分利用大数据来设计更具有特征表达能力的神经网络模型，还值得进一步研究。

（4）优化神经网络模型，降低计算复杂度

当前神经网络模型依赖于高性能的 GPU 进行计算，而对某些特定任务需要 GPU 集群进行并行加速计算，这对硬件平台提出了更高的要求。另外，较高的计算复杂度也限制了神经网络模型在嵌入式产品上的集成开发。研究低能耗、高精度的神经网络模型是当前产业化过程的当务之急。

（5）研究卷积神经网络的迁移和泛化能力

当前卷积神经网络模型通常在某类数据集上训练，在同一数据集上测试性能表现良好，然而在其他数据集尤其是互联网大规模数据上的性能则会大幅下降。研究迁移学习和在线学习，对神经网络模型进行不断的迁移和更新，增强神经网络的泛化能力是未来的一个研究方向。

参 考 文 献

[1] Hubel D.H., Wiesel T. N.. Receptive fields, binocular interaction and functional architecture in the cat’s visual cortex. The Journal of Physiology, 1962, 160(1): 106–154

[2] Fukushima K., Miyake S., Ito T.. Neocognitron: A neural network model for a mechanism of visual pattern recognition. IEEE Transactions on Systems, Man, and Cybernetics, 1983, 13(5): 826–834

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

[4] LeCun Y., Jackel L., Bottou L., et al.. Comparison of learning algorithms for handwritten digit recognition. Proceedings of the International Conference on Artificial Neural Networks, Paris, France, 1995, 60: 53-60.

[5] LeCun Y., Jackel L. D., Bottou L., et al.. Learning algorithms for classification: a comparison on handwritten digit recognition. Proceedings of the Neural Networks: the Statistical Mechanics Perspective, Pohang, Korea, 1995: 261-276.

[6] Krizhevsky A., Sutskever I., Hinton G. E.. ImageNet classification with deep convolutional neural networks. Proceedings of the Neural Information Processing Systems, Lake Tahoe, USA, 2012: 1097-1105

[7] Rumelhart D. E., Hinton G. E., Williams R. J.. Learning internal representations by error propagation. University of California San Diego, USA, Technical Report: ICS-8506, 1985

[8] Rumelhart D. E., Hinton G. E. and Williams R. J.. Learning representations by back-propagating errors. Nature, 1986, 323: 533-536

[9] Simonyan K., Zisserman A.. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556, 2014.

[10] Szegedy C., Liu W., Jia Y., et al.. Going Deeper With Convolutions. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, USA, 2015: 1-9

[11] He K., Zhang X., Ren S., Sun J.. Deep residual learning for image recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, USA, 2016: 770-778

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

[13] Ioffe S., Szegedy C.. Batch normalization: Accelerating deep network training by reducing internal covariate shift. Proceedings of the International Conference on Machine Learning, Lille, France, 2015: 448-456

[14] He, K., Zhang, X., Ren, S., Sun, J.. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. Proceedings of the IEEE International Conference on Computer

Vision, Santiago, Chile, 2015: 1026-1034

[15] Sun Y., Wang X., Tang X.. Deep learning face representation from predicting 10,000 classes. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, USA, 2014: 1891-1898

[16] Sun Y., Chen Y., Wang X., Tang X.. Deep learning face representation by joint identification-verification. Proceedings of the Advances in Neural Information Processing Systems, Montreal, Canada, 2014: 1988-1996

[17] Sun Y., Liang D., Wang X., Tang X.. Deepid3: Face recognition with very deep neural networks. arXiv preprint arXiv:1502.00873, 2015

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

[19] Schroff F., Kalenichenko D., Philbin J.. Facenet: A unified embedding for face recognition and clustering. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, USA, 2015: 815-823

[20] Simard P.Y., Steinkraus D., Platt J.C.. Best practices for convolutional neural networks applied to visual document analysis. Proceedings of the International Conference on Document Analysis and Recognition, Edinburgh, UK, 2003: 958

[21] Lin M., Chen Q., Yan S.. Network in network. arxiv: 1312.4400, 2013.

[22] Hinton G.E., Srivastava N., Krizhevsky A., Sutskever I., Salakhutdinov R.R.. Improving neural networks by preventing coadaptation of feature detectors. arXiv preprint arXiv:1207.0580, 2012

[23] Srivastava N., Hinton G., Krizhevsky A., Sutskever I., Salakhutdinov R.. Dropout: A simple way to prevent neural networks from overfitting. The Journal of Machine Learning Research, 2014, 15(1): 1929–1958

[24] Wan L., Zeiler M., Zhang S., LeCun Y., Fergus R.. Regularization of neural networks using dropconnect. Proceedings of the International Conference on Machine Learning, Atlanta, USA, 2013: 1058–1066

[25] Goodfellow I. J., Warde-Farley D., Mirza M., Courville A.C., Bengio Y.. Maxout networks. Proceedings of the International Conference on Machine Learning, Atlanta, USA, 2013: 1319–1327

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

[27] Torralba A., Fergus R., Freeman W. T.. 80 million tiny images: A large data set for nonparametric object and scene recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2008, 30(11): 1958–1970

[28] Xiao J., Hays J., Ehinger K. A., et al.. Sun database: large-scale scene recognition from abbey to zoo. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, San Francisco, USA, 2010: 3485–3492

[29] Deng J., Dong W., Socher R., et al.. ImageNet: a large-scale hierarchical image database. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Miami, USA, 2009: 248–255

[30] Papageorgiou C. P., Oren M., Poggio T.. A general framework for object detection. Proceedings of the International Conference on Computer Vision, Bombay, India, 1998: 555–562

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

[32] Ke Y, Sukthankar R. PCA-SIFT: A more distinctive representation for local image descriptors. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Washington, USA, 2004: 506-513

[33] Bay H, Tuytelaars T, Van Gool L. Surf: Speeded up robust features. Proceedings of the European Conference of Computer Vision, Graz, Austria, 2006: 404-417

[34] Viola P, Jones M. Rapid object detection using a boosted cascade of simple features. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Kauai, USA, 2001: 511-518

[35] Freund Y., Schapire R.E.. A decision-theoretic generalization of on-line learning and an application to boosting. Journal of Computer and System Sciences, 1997, 55(1): 119-139

[36] Lienhart R., Maydt J.. An extended set of haar-like features for rapid object detection. Proceedings of the IEEE International Conference on Image Processing, Rochester, USA, 2002: 900-903

[37] Dalal N., Triggs B.. Histograms of oriented gradients for human detection. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, San Diego, USA, 2005: 886-893

[38] Cortes C., Vapnik V.. Support-vector networks. Machine Learning, 1995, 20(3): 273-297

[39] Lin C.F., Wang S. D.. Fuzzy support vector machines. IEEE Transactions on Neural Networks, 2002, 13(2): 464-471

[40] Suykens J. A. K., Vandewalle J.. Least squares support vector machine classifiers. Neural Processing Letters, 1999, 9(3): 293-300

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

[42] Szegedy C., Toshev A., Erhan D.. Deep neural networks for object detection. Proceedings of the Advances in Neural Information Processing Systems, Lake Tahoe, USA, 2013: 2553-2561

[43] Erhan D., Szegedy C., Toshev A., et al.. Scalable Object Detection using Deep Neural Networks. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, USA, 2014: 2147-2154

[44] Girshick R., Donahue J., Darrell T., et al.. Rich feature hierarchies for accurate object detection and semantic segmentation. Proceedings of the IEEE conference on computer vision and pattern recognition, Columbus, USA, 2014: 580-587

[45] Erhan D., Szegedy C., Toshev A., Anguelov D.. Scalable object detection using deep neural networks. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, USA, 2014: 2155–2162

[46] Sermanet P., Eigen D., Zhang X., et al.. Overfeat: Integrated recognition, localization and detection using convolutional networks. arXiv preprint arXiv:1312.6229, 2013

[47] Girshick R.. Fast r-cnn. Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 2015: 1440-1448

[48] He K., Zhang X., Ren S., 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

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

[50] Ren S., He K., Girshick R., Sun J.. Faster R-CNN: Towards real-time object detection with region proposal networks. Proceedings of the Advances in Neural Information Processing Systems, Montreal, Canada, 2015: 91-99

[51] Redmon J., Divvala S., Girshick R., Farhadi A.. You only look once: Unified, real-time object detection. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, USA, 2016: 779-788

[52] Huang L., Yang Y., Deng Y., Yu Y.. DenseBox: Unifying Landmark Localization with End to End Object Detection. arXiv preprint arXiv:1509.04874, 2015

[53] Shotton J., Winn J., Rother C., Criminisi A.. Textonboost: Joint appearance, shape and context modeling for multi-class object recognition and segmentation. Proceedings of the European Conference on Computer Vision, Graz, Austria, 2006: 1-15

[54] Everingham M., Eslami S. M. A., Van Gool L., Williams C. K. I., Winn J., Zisserman A.. The pascal visual object classes challenge: A retrospective. International Journal of Computer Vision, 2015, 111(1):98–136

[55] Lin T.-Y., Maire M., Belongie S., Hays J., Perona P., Ramanan D., Doll ár P., Zitnick C. L.. Microsoft COCO: Common objects in context. Proceedings of the European Conference on Computer Vision, Zurich, Switzerland, 2014: 740-755

[56] Mottaghi R., Chen X., Liu X., Cho N.-G., Lee S.-W., Fidler S., Urtasun R., Yuille A.. The role of context for object detection and semantic segmentation in the wild. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, USA, 2014: 891-898

[57] Liu C., Yuen J., Torralba A.. Nonparametric scene parsing: Label transfer via dense scene alignment. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Miami, USA, 2009: 1972-1979

[58] Lucchi A., Li Y., Smith K., Fua P.. Structured image segmentation using kernelized features. Proceedings of the European Conference on Computer Vision, Florence, Italy, 2012: 400-413

[59] Tighe J., Lazebnik S.. Superparsing: scalable nonparametric image parsing with superpixels. Proceedings of the European Conference on Computer Vision, Heraklion, Greece, 2010: 352-365

[60] Gould S., Rodgers J., Cohen D., Elidan G., Koller D.. Multi-class segmentation with relative location prior. International Journal of Computer Vision, 2008, 80(3): 300-316

[61] Ladicky L., Russell C., Kohli P., Torr P. H. S.. Associative hierarchical crfs for object class image segmentation. Proceedings of the IEEE International Conference on Computer Vision, Kyoto, Japan, 2009:739-746

[62] Shotton J., Johnson M., Cipolla R.. Semantic texton forests for image categorization and segmentation. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Anchorage, USA, 2008: 1-8

[63] Farabet C., Couprie C., Najman L., Le-Cun Y.. Learning hierarchical features for scene labeling. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013, 35(8): 1915-1929

[64] Mostajabi M., Yadollahpour P., Shakhnarovich G.. Feedforward semantic segmentation with zoom-out features. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, USA, 2015: 3376-3385

[65] Pinheiro P. H., Collobert R.. Recurrent convolutional neural networks for scene labeling. Proceedings of the International Conference on Machine Learning, Beijing, China, 2014: 82-90

[66] Long J., Shelhamer E., Darrell T.. Fully convolutional networks for semantic segmentation. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, USA, 2015: 3431-3440

[67] Chen L.-C., Papandreou G., Kokkinos I., Murphy K., Yuille A. L.. Semantic image segmentation with deep convolutional nets and fully connected crfs. Proceedings of the IEEE Conference on Learning Representations, San Diego, USA, 2015: 1-14

[68] Kr ähenb ühl P., Koltun V. Efficient inference in fully connected crfs with gaussian edge potentials. Proceedings of the Advances in Neural Information Processing Systems, Granada, Spain, 2011: 109-117

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

[70] Dai J., He K., Sun J.. Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation. Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 2015: 1635-1643

[71] Bearman A., Russakovsky O., Ferrari V., Fei-Fei L.. What's the point: Semantic segmentation with point supervision. Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 2016: 549-565

[72] Pinheiro P. O., Collobert R.. From image-level to pixel-level labeling

with convolutional networks. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, USA, 2015: 1713-1721

[73] Fang H., Gupta S., Iandola F., et al.. From captions to visual concepts and back. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, USA, 2015: 1473-1482

[74] Devlin J., Cheng H., Fang H., et al.. Language models for image captioning: The quirks and what works. arXiv preprint arXiv:1505.01809, 2015

[75] Vinyals O., Toshev A., Bengio S., Erhan D.. Show and tell: A neural image caption generator. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, USA, 2015: 3156-3164

[76] Xu K., Ba J., Kiros R., et al.. Show, attend and tell: Neural image caption generation with visual attention. Proceedings of the International Conference on Machine Learning, Lille, France, 2015: 2048-2057

[77] Karpathy A., Fei-Fei L.. Deep visual-semantic alignments for generating image descriptions. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, USA, 2015: 3128-3137

[78] Donahue J., Anne Hendricks L., Guadarrama S., Rohrbach M., Venugopalan S., Saenko K., Darrell T.. Long-term recurrent convolutional networks for visual recognition and description. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, USA, 2015: 2625-2634

[79] Jiang Shuqiang, Min Weiqing, Wang Shuhui. Survey and prospect of intelligent interaction-oriented image recognition techniques. Journal of Computer Research and Development, 2016, 53(1): 113-122

(蒋树强，闵巍庆，王树徽.面向智能交互的图像识别技术综述与展望. 计算机研究与发展，2016, 53(1): 113-122)

[80] Kiros R., Salakhutdinov R., Zemel R. S.. Unifying visual-semantic embeddings with multimodal neural language models. arXiv preprint arXiv:1411.2539, 2014

[81] Mao J., Xu W., Yang Y., Wang J., Yuille A. L.. Explain images with multimodal recurrent neural networks. arXiv preprint arXiv:1410.1090, 2014

[82] Chen X., Lawrence Zitnick C.. Mind's eye: A recurrent visual representation for image caption generation. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, USA, 2015: 2422-2431

[83] Jia X., Gavves E., Fernando B., Tuytelaars T.. Guiding long-short term memory for image caption generation. arXiv preprint arXiv:1509.04942, 2015

[84] Mao J., Xu W., Yang Y., Wang J., Huang Z., Yuille A.. Deep captioning with multimodal recurrent neural networks (m-rnn). arXiv preprint arXiv:1412.6632, 2014

[85] Huang G. B., Ramesh M., Berg T., et al.. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst, USA, 2007

[86] Kumar N., Berg A. C., Belhumeur P. N., et al.. Attribute and simile classifiers for face verification. Proceedings of the International Conference on Computer Vision, Kyoto, Japan, 2009: 365–372

[87] Chen D., Cao X., Wen F., et al.. Blessing of dimensionality: High-dimensional feature and its efficient compression for face verification. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Portland, USA, 2013:3025–3032

[88] Sun Y., Wang X., Tang X.. Hybrid deep learning for face verification. Proceedings of the IEEE International Conference on Computer Vision, Sydney, Australia, 2013: 1489-1496

[89] Taigman Y., Yang M., Ranzato M., et al.. Deepface: Closing the gap to human-level performance in face verification. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, USA, 2014: 1701-1708

[90] Sun Y., Wang X., Tang X.. Deeply learned face representations are sparse, selective, and robust. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, USA, 2015: 2892-2900

[91] Gray D., Brennan S., Tao H.. Evaluating appearance models for recognition, reacquisition, and tracking. Proceedings of the IEEE International Workshop on Performance Evaluation for Tracking and Surveillance (PETS), Rio de Janeiro, Brazil, 2007, 3(5): 1-7

[92] Schwartz W. R., Davis L. S.. Learning discriminative appearance-based models using partial least squares. Proceedings of the Brazilian Symposium on Computer Graphics and Image Processing, Rio de Janeiro, Brazil, 2009: 322-239

[93] Li W., Zhao R., Wang X.. Human reidentification with transferred metric learning. Proceedings of the Asian Conference on Computer Vision, Daejeon, Korea, 2012: 31–44

[94] Hirzer M., Beleznai C., Roth P. M., Bischof H.. Person re-identification by descriptive and discriminative classification. Proceedings of the Scandinavian Conference on Image Analysis, Ystad, Swiden, 2011: 91–102

[95] U. H. Office. Imagery library for intelligent detection systems (i-LIDS). Proceedings of the Institution of Engineering and Technology Conference on Crime and Security, London, UK, 2006: 445-448

[96] Ding S., Lin L., Wang G., Chao H.. Deep feature learning with relative distance comparison for person re-identification. Pattern Recognition, 2015, 48(10): 2993-3003

[97] Li W., Zhao R., Xiao T., Wang X.. Deepreid: Deep filter pairing neural network for person re-identification. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, USA, 2014: 152–159

[98] Zhao R., Ouyang W., Wang X.. Learning mid-level filters for person re-identification. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, USA, 2014: 144–151

[99] Ahmed E., Jones M., Marks T. K.. An improved deep learning

architecture for person re-identification. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, USA, 2015: 5:25.

[100] Yi D., Lei Z., Liao S., Li S. Z.. Deep metric learning for person re-identification. Proceedings of the International Conference on Pattern Recognition, Stockholm, Sweden, 2014: 34–39

[101] Cheng D., Gong Y., Zhou S., Wang J.. Person Re-Identification by An Multi-Channel Parts-Based CNN with Improved Triplet Loss Function. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, USA, 2016: 1335-1344

[102] Keys R.. Cubic convolution interpolation for digital image processing. IEEE Transactions on A-coustics, Speech and Signal Processing, 1981, 29 (6): 1153-1160

[103] Irani M., Peleg S.. Motion analysis for image enhancement: Resolution, occlusion, and transparency. Journal of Visual Communication and Image Representation, 1993, 4(4): 324-335

[104] Aly H. A., Dubois E.. Image up-sampling using total-variation regularization with a new observation model. IEEE Transactions on Image Processing, 2015, 14 (10): 1647-1659

[105] Freeman W. T., Pasztor E. C., Carmichael O. T.. Learning low-level vision. International Journal of Computer Vision, 2000, 40(1): 25-47

[106] Yang J., Wright J., Huang T., Ma Y.. Image super-resolution as sparse representation of raw image patches. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Anchorage, USA, 2008: 1-8

[107] Liang Y., Wang J., Zhang S., Gong Y.. Incorporating Image Degeneration Modeling With Multi-task Learning For Image Super-resolution. Proceedings of the International Conference of Image Processing, Quebec City, 2015: 2110-2114

[108] Bevilacqua M., Roumy A., Guillemot C., Morel M.-L. A.. Low-complexity single-image super-resolution based on nonnegative neighbor embedding. Proceedings of the British Machine Vison Conference, Guildford, England, 2012: 1-10

[109] Zeyde R., Elad M., Protter M.. On single image scale-up using sparse-representations. Proceedings of the International Conference on Curves and Surfaces, Berlin, German, 2010: 711–730

[110] Dong C., Loy C. C., He K., Tang X.. Learning a Deep Convolutional Network for Image Super-Resolution. Proceedings of the European Conference on Computer Vision, Zurich, Switzerland, 2014: 184-199

[111] Dong C., Loy C. C., He K., Tang X.. Image Super-Resolution Using Deep Convolutional Networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 38(2): 295-307

[112] Liang Y., Wang J., Gong Y., Zheng N.. Incorporating Image Priors with Deep Convolutional Neural Networks for Image Super-Resolution. Neurocomputing, 2016, 194: 340-347

[113] Wang Z., Liu D., Yang J., Han W., Huang T.. Deep Networks for Image Super-Resolution with Sparse Prior. Proceedings of the International Conference on Computer Vision, Santiago, Chile, 2015: 370-278

[114] Willems G., Tuytelaars T., Van Gool L.. An efficient dense and scale-invariant spatio-temporal interest point detector. Proceedings of the European Conference on Computer Vision, Marseille, France, 2008: 650-663

[115] Everts I., van Gemert J.C., Gevers T.. Evaluation of color stips for human action recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Portland, USA, 2013: 2850-2857

[116] Yuan C., Li X., Hu W., et al.. 3D R transform on spatio-temporal interest points for action recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Portland, USA, 2013: 724-730

[117] Ke Y., Sukthankar R., Hebert M.. Spatio-temporal shape and flow correlation for action recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Minneapolis, USA, 2007: 1-8

[118] Bobick A.F., Davis J.W.. The recognition of human movement using temporal templates. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2001, 23(3): 257-67

[119] Lv F., Nevatia R.. Recognition and segmentation of 3-d human action using hmm and multi-class adaboost. Proceedings of the European Conference on Computer Vision, Graz, Austria, 2006: 359-372

[120] Wang H., Kläser A., Schmid C., et al. Dense trajectories and motion boundary descriptors for action recognition. International Journal of Computer Vision, 2013, 103(1): 60-79

[121] Tran S.D., Davis L.S.. Event modeling and recognition using markov logic networks. Proceedings of the European Conference on Computer Vision, Marseille, France, 2008: 610–623

[122] Damen D., Hogg D.. Recognizing linked events: searching the space of feasible explanations. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Miami, USA, 2009: 927–934

[123] Ivanov Y.A., Bobick A.F.. Recognition of visual activities and interactions by stochastic parsing. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2000, 22(8): 852–872

[124] Joo S.W., Chellappa R.. Attribute grammar-based event recognition and anomaly detection. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, New York, USA, 2006: 107–107

[125] Zhang Z., Tan T., Huang K.. An extended grammar system for learning and recognizing complex visual events. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2011, 33 (2): 240-255

[126] Gorelick L., Blank M., Shechtman E., Irani M., Basri R.. Actions as space-time shapes. Proceedings of the IEEE International Conference on Computer Vision, Beijing, China, 2005: 1395-1402

[127] Schuldt C., Laptev I., Caputo B.. Recognizing human actions: a local SVM approach. Proceedings of the International Conference on Pattern Recognition, Cambridge, UK, 2004: 32-36

[128] Marszalek M., Laptev I., Schmid C.. Actions in context. Proceedings

of the IEEE Conference on Computer Vision and Pattern Recognition, Miami, USA, 2009: 2929-2936

[129] Soomro K., Zamir A.R., Shah M.. UCF101: A dataset of 101 human actions classes from videos in the wild. arXiv:1212.0402. 2012

[130] Kuehne H., Jhuang H., Garrote E., Poggio T., Serre T.. HMDB: a large video database for human motion recognition. Proceedings of the International Conference on Computer Vision, Barcelona, Spain, 2011: 2556-2563

[131] Ji S., Xu W., Yang M., Yu K.. 3D convolutional neural networks for human action recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013, 35(1): 221:231

[132] Varol G., Laptev I., Schmid C.. Long-term temporal convolutions for action recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, PP(99): 1-8

[133] Cheron G., Laptev I., Schmid C.. P-CNN: pose-based CNN features for action recognition. Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 2015: 3218-3226

[134] Karpathy A., Toderici G., Shetty S., Leung T., Sukthankar R., Li F. Large-scale video classification with convolutional neural networks. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, USA, 2014: 1725-1732

[135] Simonyan K., Zisserman A. Two-stream convolutional networks for action recognition in videos. Proceedings of the Advances in Neural Information Processing Systems, Montreal, Canada, 2014: 568-576

[136] Liu Y., Xu D., Tsang I., Luo J.. Using large-scale web data to facilitate textual query based retrieval of consumer photos. Proceedings of the ACM International Conference on Multimedia, Beijing, China, 2009: 55-64

[137] Jeon J., Lavrenko V., Manmatha R.. Automatic image annotation and retrieval using cross-media relevance models. Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval, Toronto, Canada, 2003: 119-126

[138] Fergus R., Fei-Fei L., Perona P., et al. Learning object categories from Google's image search. Proceedings of the IEEE International Conference on Computer Vision, Beijing, China, 2005: 1816-1823

[139] Zheng Y., Zhang Y., Larochelle H.. Topic modeling of multimodal data: An autoregressive approach. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, USA, 2014: 1370-1377

[140] Niblack W, Barber R, Equitz W, et al. The QBIC project: querying images by content using color, texture, and shape. Proceedings of the IS&T/SPIE's Symposium on Electronic Imaging: Science and Technology, San Jose, USA, 1993, 1908: 173-181

[141] Bach J., Fuller C., Gupta A., et al. Virage image search engine: an open framework for image management. Electronic Imaging: Science & Technology, 1996, 2670(1): 76-87

[142] Xia R., Pan Y., Lai H., Liu C., Yan S.. Supervised Hashing for Image Retrieval via Image Representation Learning. Proceedings of the Association for the Advancement of Artificial Intelligence, Québec City, Canada, 2014: 2156-2162

[143] Wan J., Wang D., Hoi S.C., Wu P., Zhu J., Zhang Y., Li J.. Deep learning for content-based image retrieval: A comprehensive study. Proceedings of the ACM International Conference on Multimedia, Orlando, USA, 2014: 157-166

[144] Zhao F., Huang Y., Wang L., Tan T.. Deep Semantic Ranking Based Hashing for Multi-Label Image Retrieval. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, USA, 2015: 1556-1564

[145] Lai H., Pan Y., Liu Y., Yan S.. Simultaneous Feature Learning and Hash Coding with Deep Neural Networks. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, USA, 2015: 3270-3278

[146] Liu H., Wang R., Shan S., Chen X.. Deep Supervised Hashing for Fast Image Retrieval. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, USA, 2016: 2064-2072

[147] Hubel D.H., Wiesel T.N.. Receptive fields of single neurons in the cat's striate cortex. The Journal of Physiology, 1959, 148(3): 574–591

[148] Ungerleider, L.G., James V.H.. "What" and "where" in the human brain. Current opinion in neurobiology, 1994, 4(2): 157-165

[149] Ungerleider L.G., Susan M.C., James V.H.. A neural system for human visual working memory. Proceedings of the National Academy of Sciences, Irvine, USA, 1998, 95(3): 883-890

[150] Ungerleider L.G.. Functional brain imaging studies of cortical mechanisms for memory. Science, 1995, 270(5237): 769-775

[151] DiCarlo J. J., Zoccolan D., Rust N. C.. How does the brain solve visual object recognition? Neuron, 2012, 73(3): 415-434

[152] Poggio T., Ullman S.. Vision: are models of object recognition catching up with the brain? Annals of the New York Academy of Sciences, 2013, 1305(1): 72-82

[153] Poggio T., Serre T.. Models of visual cortex. Scholarpedia, 2013, 8(4): 3516

[154] Anselmi F., Poggio T. A.. Representation learning in sensory cortex: a theory. Center for Brains, Minds and Machines (CBMM), 2014, 26: 1-56

[155] Ullman S., Humphreys G. W.. High-level vision: Object recognition and visual cognition. Cambridge, USA: MIT press, 1996

[156] Pinto N., Cox D. D., DiCarlo J. J.. Why is real-world visual object recognition hard? PLoS Computational Biology, 2008, 4(1): e27

[157] Treisman, A.M., Gelade G.. A feature-integration theory of attention. Cognitive psychology, 1980, 12(1): 97-136

[158] Marr D.. Vision: A computational Investigation into the human representation and processing of visual information. San Francisco, USA: W.H. Freeman and Company, 1982.

[159] Navon D.. Forest before trees: The precedence of global features in visual perception. Cognitive psychology, 1977, 9(3): 353-383

[160] Chen L.. Topological structure in visual perception. Science, 1982, 218: 699-700

[161] Chen L., Zhang S.W., Srinivasan M.. Global perception in small

brains: Topological pattern recognition in honeybees. PNAS, 2003, 100(11): 6884-6889

[162] Zhuo Y., Zhou T.G., Rao H.Y., Wang J.J., Meng M., Chen M., Zhou C., Chen L.. Contributions of the visual ventral pathway to long-range apparent motion. Science, 2003, 299: 417-420

[163] Chen L.. The topological approach to perceptual organization. Visual Cognition, 2005, 12: 553-637

[164] Zhuo Y., Zhou T.G., Rao H.Y., Wang J.J., Meng M., Chen M., Zhou C., Chen L.. Contributions of the visual ventral pathway to long-range apparent motion. Science, 2003(299): 417-420

[165] Wang B., Zhou T.G., Zhuo Y., Chen, L.. Global Topological Dominance in the Left Hemisphere. Proceedings of the National Academy of Sciences, 2007(104): 21014-21019

[166] Zhou K., Huan L., Zhou T.G., Zhuo Y., Chen L.. Topological change disturbs object continuity in attentive tracking. Proceedings of the National Academy of Sciences, 2010, 107(50), 21920-21924

[167] Han Shihui, Chen Lin. The relationship of global feature and local feature —— global precedence. Dynamic psychology, 1996, 1: 36-41.
(韩世辉，陈霖. 整体性质和局部性质的关系——大范围优先性. 心理学动态, 1996, 1996(1): 36-41

[168] Hochstein S., Ahissar M.. View from the top: Hierarchies and reverse hierarchies in the visual system. Neuron, 2002, 36(5): 791-804

[169] Ahissar M., Hochstein S.. The reverse hierarchy theory of visual perceptual learning. Trends in cognitive sciences, 2004, 8(10):457-464

[170] Bar M.. A cortical mechanism for triggering top-down facilitation in visual object recognition. Journal of cognitive neuroscience, 2003, 15(4): 600-609

[171] LeCun Y., Bengio Y., Hinton G.. Deep learning. Nature, 2015, 521(7553): 436-444.

Zhang Shun, born in 1987, Ph.D., Assistant Professor. His research interests include computer vision and machine learning.

Gong Yihong, born in 1963, Ph.D., Professor, Ph.D. supervisor. His research interests include multimedia content analysis, machine learning and pattern recognition

Wang Jinjun, born in 1977, Ph.D., Professor, Ph.D. supervisor. His research interests include pattern recognition, machine learning and multimedia computing.

Background

Benefited by the rapid growth in the amount of the annotated data and the recent improvements in the strengths of graphics processor units (GPUs), the research on convolutional neural networks has been widely applied to many fields of computer vision and pattern recognition, and have attracted huge attentions from both academia and industry. This paper aims to present a comprehensive introduction of deep convolutional neural network and its applications in the field of Computer Vision. We first introduce the working principle of the convolutional neural network. Then we list many general approaches that are proposed to improve the performance of deep neural networks, including the increase of size and complexity of neural networks, the use of larger sets of training data, the improvements of neural network training methods, etc. Besides, we show many applications of convolutional neural networks in the field of computer vision. Based on the above analysis, we also point out its possible future directions from the human visual cognitive mechanism.

This work is supported by the National Basic Research Program (973 Program) of China under Grant No. 2015CB351705, the State Key Program of National Natural Science Foundation of China under Grant No. 61332018, the Youth Program of National Natural Science Foundation of China under Grant No. 61703344, and the Fundamental Research Funds for the Central Universities under Grant No. 3102017OQD021. Our research team has been working on handling various tasks (such as image classification, object detection and recognition, face verification and recognition, person recognition, super resolution, etc) in the field of Computer Vision with the techniques of deep learning for years. Works related to these projects have been published in international journals and conferences, such as TNN, IJCV, TIP, IJCAI, CVPR, ECCV, etc. The technique of deep learning has been widely applied in various fileds, and its power capability of learning features is exploited by us to learn discriminative features for different objects in images or videos. This review

paper can help us to get a understanding of the fundamental system, generally used training techniques, and recent developments of the convolutional neural networks.