计算机视觉经典综述 / 01

卷积神经网络研究进展

Recent Advances in Convolutional Neural Networks

Jiuxiang Gu、Zhenhua Wang、Jason Kuen、Lianyang Ma、Amir Shahroudy、Bing Shuai、Ting Liu、Xingxing Wang、Li Wang、Gang Wang、Jianfei Cai、Tsuhan Chen

作者单位均在新加坡南洋理工大学。Jiuxiang Gu 来自 ROSE 实验室及跨学科研究生院；Zhenhua Wang 至 Gang Wang 来自电气与电子工程学院；Jianfei Cai、Tsuhan Chen 来自计算机科学与工程学院。Jiuxiang Gu 与 Zhenhua Wang 同等贡献。

期刊 Pattern Recognition，第 77 卷（2018），354–377 页。

译本所据原稿 arXiv:1512.07108v6，2017 年 10 月 19 日，38 页。中文译本中的“近年来”与“最佳结果”指原文写作时期。原图与公式保留原编号，图下注有中文读图说明。

摘要

近年来，深度学习在视觉识别、语音识别、自然语言处理等多种问题上取得了很好的效果。在各种深度神经网络中，卷积神经网络（convolutional neural network，CNN）受到的研究最为广泛。随着标注数据迅速增加，图形处理器（GPU）的计算能力大幅提升，卷积神经网络研究迅速发展，在多项任务上取得了当时的最佳结果。本文广泛综述卷积神经网络近年来的研究进展，详细讨论网络层设计、激活函数、损失函数、正则化、优化和快速计算方面的改进，并介绍 CNN 在计算机视觉、语音和自然语言处理中的应用。

关键词 卷积神经网络；深度学习。

章节导航

摘要 1

1 引言 2

2 CNN 的基本组成 3

3 CNN 的改进 4

4 卷积神经网络的快速计算 17

5 卷积神经网络的应用 19

6 结论与展望 25

致谢 26

参考文献 27

1 引言

卷积神经网络是一种广为人知的深度学习架构，其灵感来自生物的视觉感知机制。1959 年，Hubel 和 Wiesel [1] 发现，动物视觉皮层中的细胞能够检测落在各自感受野内的光刺激。受此启发，福岛邦彦于 1980 年提出新认知机（neocognitron）[2]，可视为 CNN 的前身。1990 年，LeCun 等 [3] 发表了奠定现代 CNN 框架的论文，随后又在文献 [4] 中加以改进。他们开发了一个名为 LeNet-5 的多层人工神经网络，用来识别手写数字。和其他神经网络一样，LeNet-5 由多层构成，可用反向传播算法 [5] 训练。它能从原始图像中获得有效表示，只需很少的预处理，甚至无需预处理，就能直接从原始像素识别视觉模式。几乎同时，Zhang 等 [6] 用具备平移不变性的人工神经网络（SIANN）识别图像中的字符。受限于当时的训练数据规模和计算能力，这些网络在大规模图像与视频分类等更复杂的问题上表现不佳。

自 2006 年起，研究者提出多种解决深层 CNN 训练困难的方法 [7–10]。其中，Krizhevsky 等提出了一种经典的 CNN 架构，在图像分类任务上显著超越了以往方法。这一架构称为 AlexNet [8]，其总体结构与 LeNet-5 相似，但网络更深。AlexNet 取得成功后，许多工作进一步改进了它的性能。具有代表性的四项工作是 ZFNet [11]、VGGNet [9]、GoogLeNet [10] 和 ResNet [12]。这些架构的演变呈现出一个明显趋势：网络越来越深。例如，赢得 2015 年 ImageNet 大规模视觉识别挑战赛（ILSVRC）冠军的 ResNet，深度约为 AlexNet 的 20 倍、VGGNet 的 8 倍。增加深度可以增强网络的非线性，使之更好地逼近目标函数，获得更好的特征表示。然而，深度也增加了网络复杂度，使优化更加困难，也使网络更容易过拟合。沿着这一发展路线，研究者从多个方面提出了解决办法。本文试图全面回顾并较为深入地讨论这些进展。

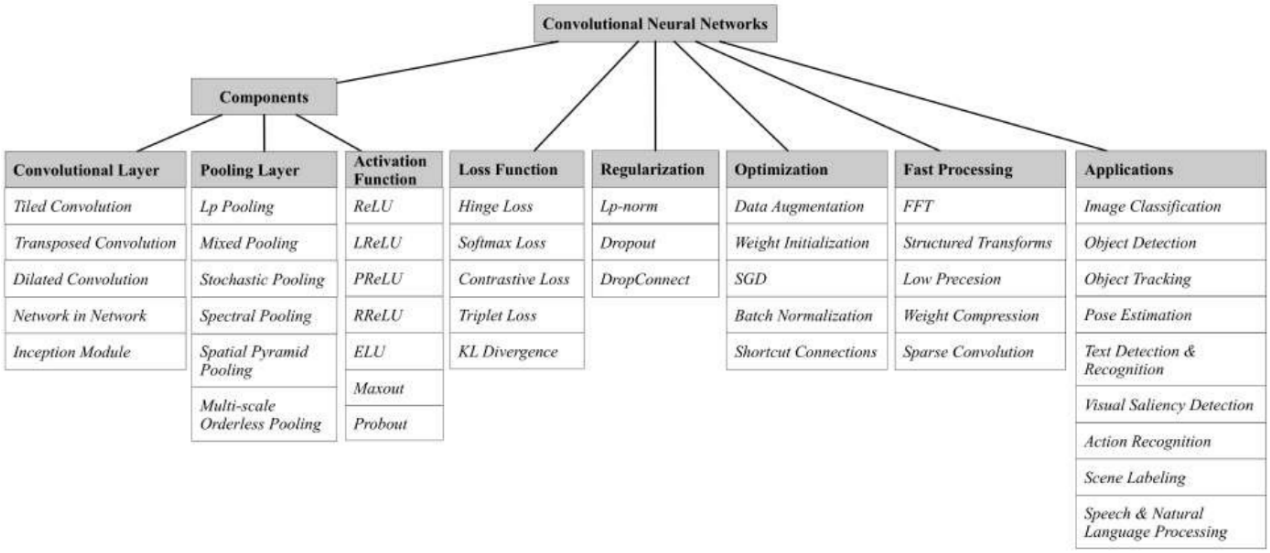


图 1 本文综述内容的层次分类。

读图 从左向右，Components（组成）下列出卷积层、池化层和激活函数。其余分支为损失函数、正则化、优化、快速计算及应用。快速计算依次列出 FFT、结构化变换、低精度、权重压缩和稀疏卷积；应用则覆盖分类、检测、跟踪、姿态估计、文本检测与识别、显著性检测、动作识别、场景标注及语音与自然语言处理。

后文将与 CNN 有关的工作划分为若干大类，图 1 给出其层次结构。第 2 节概述 CNN 的基本组成。第 3 节介绍卷积层、池化层、激活函数、损失函数、正则化和优化等方面的改进。第 4 节介绍快速计算技术。第 5 节讨论典型应用，包括图像分类、目标检测、目标跟踪、姿态估计、文本检测与识别、视觉显著性检测、动作识别、场景标注、语音处理和自然语言处理。第 6 节作总结。

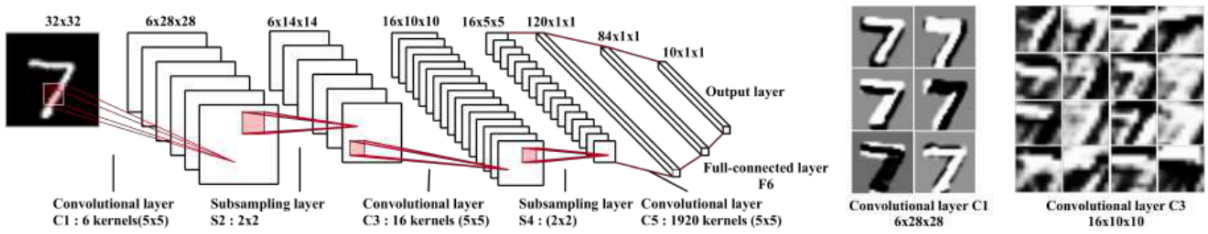
2 CNN 的基本组成

文献中已有众多 CNN 架构变体，其基本组成却很相似。以著名的 LeNet-5 为例，它包含三类网络层：卷积层、池化层和全连接层。卷积层的任务是学习输入的特征表示。如图 2(a) 所示，卷积层包含多个卷积核，分别用来计算不同的特征图。每张特征图上的一个神经元，与前一层某个局部区域内的一组邻近神经元相连。这片邻域称为该神经元在前一层的感受野（receptive field）。先用学习到的核与输入做卷积，再对卷积结果逐元素施加非线性激活函数，便得到新的特征图。计算一张特征图时，输入的所有空间位置共用同一个核。用多个不同的核，即可得到整组特征图。第 l 层第 k 张特征图在位置 (i,j) 的特征值按下式计算：

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

其中，$w^l[k]$ 和 $b^l[k]$ 分别是第 l 层第 k 个滤波器的权重向量和偏置，$x^l[i,j]$ 是第 l 层中以 (i,j) 为中心的输入块。整张特征图 $z^l[:,:,k]$ 的各位置共用卷积核 $w^l[k]$。这种权重共享机制能降低模型复杂度，使网络更容易训练。激活函数给 CNN 引入非线性，使多层网络可以检测非线性特征。若用 a(·) 表示非线性激活函数，则卷积特征相应的激活值为：

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



(a) LeNet-5 network (b) Learned features

图 2 (a) 在数字分类任务上表现良好的 LeNet-5 网络结构。(b) LeNet-5 所学特征的可视化。不同图块分别展示各层的特征图。

读图 从左向右读 LeNet-5：32 × 32 输入 → C1 卷积 → S2 下采样 → C3 卷积 → S4 下采样 → C5 卷积 → F6 全连接 → 输出。图中 6 × 28 × 28 表示 6 张 28 × 28 的特征图。右侧两组小图分别是数字 7 经 C1、C3 后产生的响应。

常用激活函数包括 sigmoid、tanh [13] 和 ReLU [14]。池化层通过降低特征图分辨率来获得对平移的不敏感性，通常放在两个卷积层之间。池化层的每张特征图都与前一卷积层中对应的特征图相连。记池化函数为 pool(·)，对每张特征图 $a^l[:,:,k]$，有：

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

其中，R_{ij} 是位置 (i,j) 附近的局部邻域。典型运算有平均池化 [15] 和最大池化 [16]。图 2(b) 给出了前两个卷积层对数字 7 形成的特征图。第一卷积层的核用来检测边缘、曲线等低层特征，更高层的卷积核则学习编码更抽象的特征。通过堆叠若干卷积层和池化层，可以逐步提取更高层的特征表示。

若干卷积层和池化层之后，可以接一个或多个全连接层，用来进行高层推理 [9, 11, 17]。前一层的所有神经元均与当前层的每个神经元相连，从而形成全局语义信息。全连接层并非必不可少，有些网络可用 1 × 1 卷积层替代它 [18]。

CNN 的最后一层为输出层。分类任务中常用 softmax 算子 [8]。支持向量机（SVM）也是一种常见方法，可以与 CNN 特征结合以解决不同的分类任务 [19, 20]。令 θ 表示整个 CNN 的参数，包括权重向量和偏置。针对某项任务，最小化与之相应的损失函数，就能求取模型参数。设有 N 组期望的输入输出关系

$\{(x^{(n)}, y^{(n)}) \mid n = 1,\ldots,N\}$，其中 $x^{(n)}$ 为第 n 个输入，$y^{(n)}$ 为对应的目标标签，$o^{(n)}$ 为 CNN 的输出。CNN 的损失可写为：

$$\mathcal{L} = \frac{1}{N}\sum_{n=1}^{N} \ell(\boldsymbol{\theta}; \boldsymbol{y}^{(n)}, \boldsymbol{o}^{(n)}) \tag{4}$$

训练 CNN 是一个全局优化问题。通过最小化损失函数，可寻找最贴合数据的一组参数。随机梯度下降是优化 CNN 的常用方法 [21, 22]。

3 CNN 的改进

自 AlexNet 在 2012 年取得成功以来，CNN 已有多方面的改进。本节从卷积层、池化层、激活函数、损失函数、正则化、优化六个方面介绍主要进展。

3.1 卷积层

基础 CNN 中的卷积滤波器可以视为作用于局部图像块的广义线性模型（GLM）。当潜在概念的样例线性可分时，这种模型就能很好地提取抽象特征。以下介绍旨在增强其表示能力的几项工作。

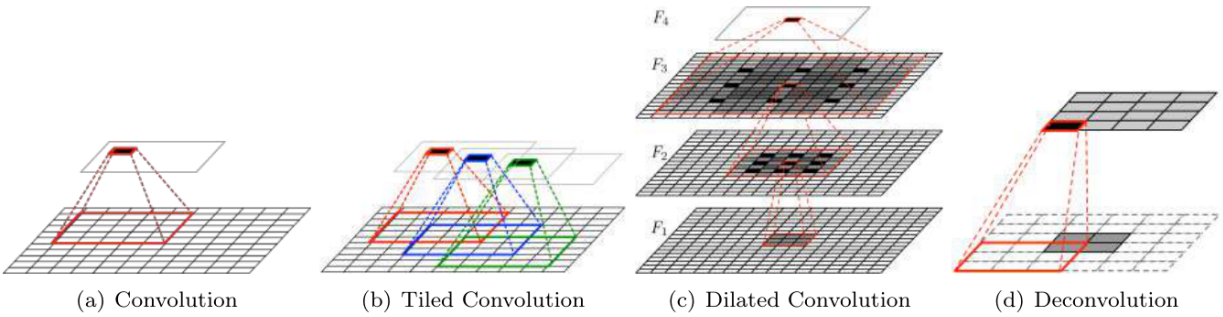


图 3 (a) 普通卷积；(b) 平铺卷积；(c) 空洞卷积；(d) 反卷积。

读图 由左向右依次为普通卷积、平铺卷积、空洞卷积和转置卷积。图 (c) 中 F_1 到 F_4 代表连续各层的特征图，网格中取样点的间隔随层数增大。图 (d) 的连线表示一个输入位置向多个输出位置传递贡献。

3.1.1 平铺卷积（Tiled Convolution）

CNN 的权重共享机制可以大幅减少参数量，却也可能限制模型学习其他类型的不变性。平铺 CNN [23] 是一种对特征图作平铺与重复排列的变体，用于学习对旋转和尺度变化具有不变性的特征。同一层内分别学习多个核，再对邻近单元进行平方根池化，从而隐式学习更复杂的不变性。图 3(b) 中，同一卷积操作每隔 k 个单元重复一次，平铺尺寸 k 控制共享权重的位置间隔。k = 1 时，每张特征图内的所有单元都共享同一套权重，平铺 CNN 就成为传统 CNN。文献 [23] 在 NORB 和 CIFAR-10 数据集上的实验发现，k = 2 时效果最好。Wang 等 [24] 发现，在小型时间序列数据集上，平铺 CNN 的表现优于传统 CNN [25]。

3.1.2 转置卷积（Transposed Convolution）

转置卷积可以看作对应普通卷积在反向传播时的运算。它也被称为反卷积（deconvolution）[11, 26–28] 或分数步幅卷积 [29]。为与多数文献 [11, 30] 的用语一致，本文采用“反卷积”。普通卷积把多个输入激活连接到一个输出激活，反卷积则把一个输入激活连接到多个输出激活。图 3(d) 给出了一个例子：在 4 × 4 输入上使用 3 × 3 的核，步幅为 1，采用零填充。反卷积的步幅决定输入特征图的膨胀因子。具体来说，先通过填充，按步幅给定的倍数对输入上采样，再对上采样的输入做卷积。近年来，反卷积已广泛用于可视化 [11]、识别 [31–33]、定位 [34]、语义分割 [30]、视觉问答 [35] 和超分辨率 [36]。

3.1.3 空洞卷积（Dilated Convolution）

空洞 CNN [37] 在卷积层中引入一个额外的超参数。通过在滤波器各元素之间插入零，空洞卷积可以扩大网络的感受野，使网络覆盖更多相关信息。这对需要较大感受野才能作出预测的任务非常重要。设一维信号为 F，卷积核 k 的大小为 r，膨胀率为 l，一维空洞卷积定义为：

$$(F *_l k)_t = \sum_\tau k[\tau]\, F[t - l\tau]$$

$*_l$ 表示膨胀率为 l 的卷积。这个公式可以直接推广到二维。图 3(c) 展示了三层空洞卷积，膨胀率 l 逐层按指数增大。从底层 F_1 到 F_2 采用膨胀率为 1 的卷积，F_2 中每个元素的感受野为 3 × 3。从 F_2 到 F_3 采用膨胀率为 2 的卷积，F_3 中每个元素的感受野为 $(2^3 - 1) \times (2^3 - 1)$。从 F_3 到顶层 F_4 采用膨胀率为 4 的卷积，F_4 中每个元素的感受野为 $(2^4 - 1) \times (2^4 - 1)$。原文将 F_{i+1} 的感受野通式写为 (2^(i+2) − 1) × (2^(i+2) − 1)。空洞 CNN 在场景分割 [37]、机器翻译 [38]、语音合成 [39] 和语音识别 [40] 等任务上取得了很好的效果。

> 译注 原文的通式与其紧邻的三个例子存在一位索引偏差。若 F_1 为输入，叠加 i 层后得到 F_{i+1}，例子所对应的边长为 2^(i+1) − 1，依次为 3、7、15。

3.1.4 网络中的网络（Network in Network）

Lin 等 [18] 提出了通用网络结构 Network in Network（NIN）。它用微型网络替换卷积层的线性滤波器，例如原论文采用的多层感知机卷积层（mlpconv），从而能逼近潜在概念更抽象的表示。NIN 的总体结构由多个微型网络堆叠而成。图 4 比较了线性卷积层和 mlpconv 层。使用非线性激活函数（例如 ReLU [14]）时，普通卷积层的特征图可写为：

$$a_{i,j,k} = \max(\mathbf{w}_k^T \mathbf{x}_{i,j} + b_k, 0) \tag{5}$$

其中 a[i,j,k] 是第 k 张特征图在 (i,j) 处的激活值，x[i,j] 是以 (i,j) 为中心的输入块，w[k] 和 b[k] 分别为第 k 个滤波器的权重向量和偏置。相比之下，mlpconv 层的运算为：

$$a_{i,j,k_n}^n = \max(\mathbf{w}_{k_n}^T \mathbf{a}_{i,j,:}^{n-1} + b_{k_n}, 0) \tag{6}$$

其中 n = 1,...,N，N 为 mlpconv 内的层数，$a^{(0)}$[i,j,:] 等于 x[i,j]。mlpconv 在传统卷积层后放置 1 × 1 卷积。1 × 1 卷积等价于跨通道的参数化池化操作，其后再接 ReLU [14]。因此，mlpconv 也可理解为在普通卷积层之上串联多个跨通道参数化池化操作。网络最终采用全局平均池化，对末层特征图沿空间位置求平均，将所得向量直接输入 softmax 层。与全连接层相比，全局平均池化使用的参数更少，因此可以降低过拟合风险和计算负担。

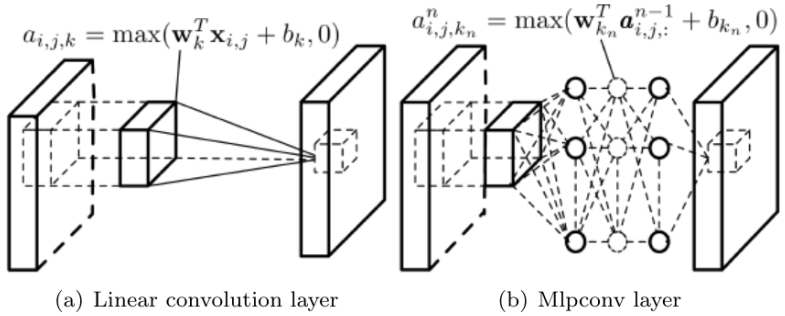


图 4 (a) 线性卷积层与 (b) 多层感知机卷积层（mlpconv）的比较。

读图 左图把虚线框内的局部输入做一次线性加权，再经过 ReLU，见式（5）。右图圆点代表微型多层感知机的各单元，在每一处局部输入上依次执行多层变换，见式（6）。

3.1.5 Inception 模块

Szegedy 等 [10] 提出的 Inception 模块可视为 NIN 思路的进一步发展。他们采用不同尺寸的滤波器捕捉不同尺度的视觉模式，并用 Inception 模块逼近理想的稀疏连接结构。具体来说，模块包括一路池化和三类卷积运算，见图 5(b)。在 3 × 3 和 5 × 5 卷积前各放置 1 × 1 卷积，用于降低通道维数。这样便有可能在不增加计算复杂度的条件下加深、加宽 CNN。借助该模块，网络参数可大幅减至约 500 万个，远少于 AlexNet 的 6,000 万个和 ZFNet 的 7,500 万个。

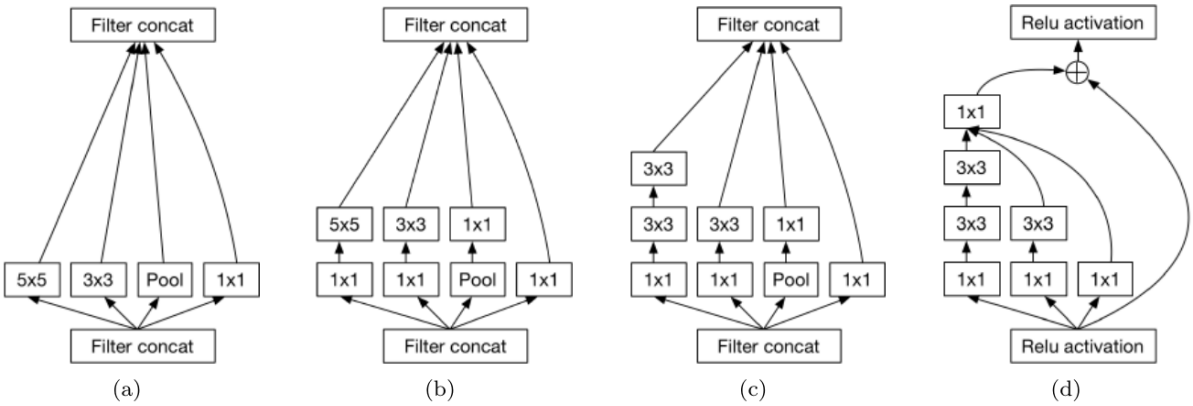


图 5 (a) 朴素版 Inception 模块；(b) 文献 [10] 的 Inception 模块；(c) 文献 [41] 的改进版，每个 5 × 5 卷积由两个 3 × 3 卷积替代；(d) 文献 [42] 中的 Inception-ResNet-A 模块。

读图 箭头从下向上。Pool 为池化，1 × 1、3 × 3 和 5 × 5 为卷积核尺寸，Filter concat 表示沿通道方向拼接各分支的输出。图 (d) 的 Relu activation 表示 ReLU 激活，圈内加号表示把捷径分支与变换分支逐元素相加。

在后续论文 [41] 中，作者探讨如何以适中的计算代价得到高性能网络。他们建议从输入到输出逐步、平缓地减小表示尺寸，并指出，在较低维度的嵌入表示上聚合空间信息，可以避免表示能力遭受过大损失。平衡每层的滤波器数量与网络的深度，有助于使性能达到最优。受 ResNet [12] 启发，他们进一步研究了将 Inception 结构与捷径连接相结合的模型 [42]，如图 5(d)。他们发现，捷径连接能显著加速 Inception 网络的训练。文中报告的 Inception-v4 有 75 个可训练层，将三个残差版本与一个 Inception-v4 集成后，在 ILSVRC 2012 验证集上取得了 3.08% 的 top-5 错误率。

3.2 池化层

池化是 CNN 中的一项重要运算。它减少卷积层之间的连接数，从而降低计算负担。以下介绍近年来在 CNN 中使用的几种池化方法。

3.2.1 L_p 池化

L_p 池化是一种受生物启发的池化过程，以视觉皮层的复杂细胞为模型 [43]。文献 [44] 对其作了理论分析，认为 L_p 池化比最大池化具有更好的泛化能力。公式为：

$$y_{i,j,k} = [\sum_{(m,n)\in\mathcal{R}_{ij}} (a_{m,n,k})^p]^{1/p} \tag{7}$$

其中，y[i,j,k] 是池化算子在第 k 张特征图位置 (i,j) 的输出，a[m,n,k] 是该特征图池化区域 R_{ij} 内位置 (m,n) 的特征值。原文指出，p = 1 时对应平均池化，p = ∞ 时退化为最大池化。

> 译注 按式（7）印出的形式，p = 1 得到区域内各值之和，除以区域元素个数才是平均值。对非负激活取 p → ∞ 时，得到最大池化。

3.2.2 混合池化（Mixed Pooling）

受随机失活（Dropout）[17] 和随机断连（DropConnect）[45] 启发，Yu 等 [46] 提出一种结合最大池化和平均池化的混合池化方法。它的函数形式为：

$$y_{i,j,k} = \lambda \max_{(m,n)\in\mathcal{R}_{ij}} a_{m,n,k} + (1-\lambda)\frac{1}{|\mathcal{R}_{ij}|}\sum_{(m,n)\in\mathcal{R}_{ij}} a_{m,n,k} \tag{8}$$

λ 是只取 0 或 1 的随机变量，决定使用平均池化还是最大池化。前向传播时应记录 λ，反向传播时使用相同的选择。文献 [46] 的实验表明，混合池化能更好地应对过拟合，其结果优于单独采用最大池化或平均池化。

3.2.3 随机池化（Stochastic Pooling）

随机池化 [47] 也是受 Dropout 启发的方法。它根据多项分布从每个池化区域随机选取一个激活值，使特征图中小于最大值的激活也有机会参与计算。对每个区域 R_j，先将区域内的激活值归一化，计算各位置的选择概率：

$$p_i = a_i \,/\, \textstyle\sum_{k\in R_j} a_k$$

得到分布 $P(p_1,\ldots,p|R_j|)$ 后，按概率 p 从多项分布采样，选中区域内的位置 l，再把该位置的激活作为池化结果：

$$l \sim P(p_1,\ldots,p|R_j|)\,,\ y_j = a_l$$

与最大池化相比，随机池化引入的随机性有助于避免过拟合。

3.2.4 频谱池化（Spectral Pooling）

频谱池化 [48] 在频域裁剪输入表示，以进行降维。设输入特征图 x 为 m × m 的实值矩阵，所需输出大小为 h × w。先对输入特征图进行离散傅里叶变换（DFT），在频域中仅保留中央 h × w 的子矩阵，再做逆 DFT，将所得近似表示映射回空间域。在输出维度相同的情况下，该线性低通滤波操作比最大池化保留更多信息，也可避免其他池化方式会使输出特征图尺寸骤降的问题。此外，频谱池化可以通过矩阵截断实现，因此，对于已采用快速傅里叶变换（FFT）计算卷积的 CNN（例如 [49]），它只需很少的额外计算。

3.2.5 空间金字塔池化（Spatial Pyramid Pooling）

He 等 [50] 引入空间金字塔池化（SPP）。其主要优点是，无论输入尺寸如何，都能产生定长表示。SPP 将输入特征图划分为局部空间区间并分别池化。区间尺寸随图像尺寸按比例变化，区间总数因而固定。在此前的深度网络中，滑动窗口池化的窗口数量则随输入大小而变。作者用 SPP 替换网络末端的池化层，构建了能够处理不同尺寸图像的 SPP-net。

3.2.6 多尺度无序池化（Multi-scale Orderless Pooling）

受文献 [51] 启发，Gong 等 [52] 用多尺度无序池化（MOP）提高 CNN 的不变性，同时保留其判别能力。他们分别在整张图像和若干尺度的局部图块上提取深层激活特征。整图激活与以往 CNN 中的相同，用来捕捉整体空间布局。局部图块的激活则通过 VLAD 编码 [53] 加以聚合，用来捕捉更局部、更细致的图像细节，并增强不变性。最后，将整图激活与局部图块激活的 VLAD 特征拼接成新的图像表示。

3.3 激活函数

针对具体任务选择适当的激活函数，可以显著改善 CNN 性能。以下介绍当时用于 CNN 的几种激活函数。

3.3.1 修正线性单元（ReLU）

修正线性单元（ReLU）[14] 是最具代表性的非饱和激活函数之一，定义为：

$$a_{i,j,k} = \max(z_{i,j,k}, 0) \tag{9}$$

z[i,j,k] 是第 k 个通道在 (i,j) 处的激活函数输入。ReLU 为分段线性函数，将负值置零、保留正值，如图 6(a) 所示。它只需执行简单的取最大值运算，因此比 sigmoid 或 tanh 计算得更快。它还使隐藏单元的激活趋于稀疏，使网络容易获得稀疏表示。研究表明，采用 ReLU 时，即使不预训练，也能高效训练深层网络 [8]。原文提到 ReLU 在零点的不光滑性可能会影响反向传播，但大量实验证据表明，其效果优于 sigmoid 和 tanh [54, 55]。

译注 原文将零点处的性质写作“不连续”。ReLU 函数本身在零点连续；零点左右导数不同，因此函数在该点不可导。

3.3.2 带泄漏的 ReLU（Leaky ReLU）

ReLU 的一个潜在缺点是：单元未激活时，梯度为零。如果某个单元从初始状态起一直未被激活，基于梯度的优化就无法调整其权重，它可能永远不再激活。持续为零的梯度也可能拖慢训练。为缓解这一问题，Maas 等 [54] 引入 Leaky ReLU（LReLU）：

$$a_{i,j,k} = \max(z_{i,j,k}, 0) + \lambda \min(z_{i,j,k}, 0) \tag{10}$$

其中 λ 是预先指定的参数，取值在 (0,1) 内。Leaky ReLU 将负半轴的响应按比例压缩，因而当输入落在负半轴时，仍保有一个很小的非零梯度。

3.3.3 参数化 ReLU（Parametric ReLU）

与 Leaky ReLU 预先设定参数（如式（10）的 λ）不同，He 等 [56] 提出参数化修正线性单元（PReLU），让整流器的参数在训练中自适应学习，以提高准确率。其数学定义为：

$$a_{i,j,k} = \max(z_{i,j,k}, 0) + \lambda_k \min(z_{i,j,k}, 0) \tag{11}$$

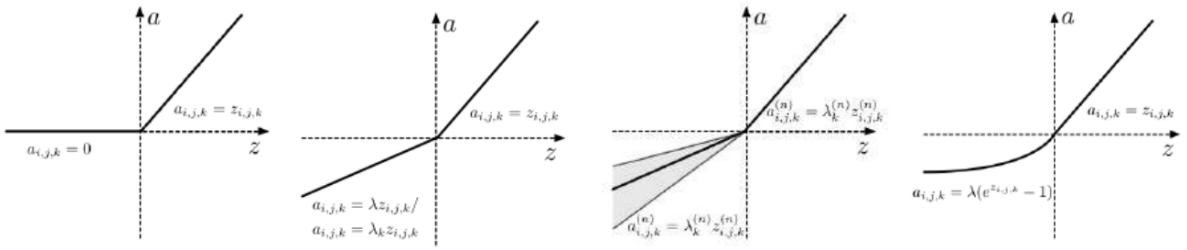
λ_k 是第 k 个通道学习到的参数。PReLU 只增加极少量参数，例如每个通道只增加一个参数，新增参数总数等于全网络的通道总数。作者据此认为，它不会引入额外的过拟合风险，新增计算成本也可忽略。这些参数可通过反向传播与其他模型参数同时训练。

3.3.4 随机化 ReLU（Randomized ReLU）

Leaky ReLU 的另一变体是随机化带泄漏修正线性单元（RReLU）[57]。训练时从均匀分布中随机采样负半轴的斜率，测试时将其固定，如图 6(c)。其函数定义为：

$$a_{i,j,k}^{(n)} = \max(z_{i,j,k}^{(n)}, 0) + \lambda_k^{(n)} \min(z_{i,j,k}^{(n)}, 0) \tag{12}$$

其中，$z^{(n)}[i,j,k]$ 表示第 n 个样本在第 k 个通道位置 (i,j) 的激活函数输入，$\lambda_k^{(n)}$ 表示相应的随机采样参数，$a^{(n)}[i,j,k]$ 是相应输出。随机性有助于减少过拟合。Xu 等 [57] 还在标准图像分类任务上评估了 ReLU、LReLU、PReLU 和 RReLU，认为在整流单元的负半轴上保留非零斜率，能稳定地改善性能。



(a) ReLU (b) LReLU/PReLU (c) RReLU (d) ELU

图 6 ReLU、LReLU、PReLU、RReLU 和 ELU 的比较。（a）ReLU；（b）LReLU/PReLU；（c）RReLU；（d）ELU。Leaky ReLU 中的 λ 根据经验预先设定；PReLU 中每个通道的 λ_k 从训练数据中学习。RReLU 中的 $\lambda_k^{(n)}$ 是随机变量，训练时从给定的均匀分布中采样，测试时固定。ELU 的 λ 也根据经验预先设定。

读图 各图横轴 z 是输入，纵轴 a 是激活输出。可沿负半轴比较四幅图：斜率为零（ReLU）、固定或学得的非零斜率（LReLU、PReLU）、训练时随机采样斜率（RReLU），以及向负值逐渐饱和的指数曲线（ELU）。

3.3.5 指数线性单元（ELU）

Clevert 等人 [58] 提出了指数线性单元（Exponential Linear Unit，ELU），以加快深度神经网络的学习并提高分类准确率。与 ReLU、LReLU、PReLU 和 RReLU 一样，ELU 在正半轴采用恒等映射，以避免这一区间的梯度消失。ELU 在负半轴也有非零响应，这有助于加快学习。

LReLU、PReLU 和 RReLU 的负半轴响应不饱和，而 ELU 在负半轴使用饱和函数。单元处于非激活状态时，饱和函数会减小输出的变化幅度，因此 ELU 对噪声更稳健。其定义为

$$a_{i,j,k} = \max(z_{i,j,k}, 0) + \min(\lambda(e^{z_{i,j,k}} - 1), 0) \tag{13}$$

λ 是预先设定的参数，控制输入趋向负无穷时 ELU 的饱和值。

3.3.6 Maxout

Maxout [59] 是另一种非线性函数，在每个空间位置取多个通道中的最大响应。文献 [59] 将它定义为对 K 个候选响应 z_{ijk} 取最大值，其中 z_{ijk} 是特征图第 k 个通道在位置 (i,j) 的响应。ReLU 本身是 Maxout 的特例，因此 Maxout 也具有 ReLU 的优点。例如，在 $\max(w_1^T x + b_1, w_2^T x + b_2)$ 中，令 w_1 为零向量、b_1 为零，便得到 ReLU 形式。Maxout 尤其适合结合 Dropout 训练。

3.3.7 Probout

Springenberg 等人 [60] 提出 Maxout 的概率变体 Probout，以概率采样代替取最大值。他们首先为 k 个线性单元分别指定概率 $p_i = \exp(\lambda z_i)/\Sigma_j \exp(\lambda z_j)$，求和从 j = 1 到 k。超参数 λ 控制分布的离散程度。然后按概率 $(p_1, \ldots, p_k)$ 抽取一个单元，以该单元的值作为激活值。为纳入 Dropout，他们重新定义这些概率：

$$\hat{p}_0 = 0.5,\ \hat{p}_i = e^{\lambda z_i}/(2.\sum_{j=1}^{k} e^{\lambda z_j}) \tag{14}$$

随后按分布 (p̂₀, ..., p̂ₖ) 抽取索引 i，并按下式给出激活值：

$$a_i = \begin{cases} 0 & \text{if } i = 0 \\ z_i & \text{else} \end{cases} \tag{15}$$

Probout 能在保留 Maxout 单元的优点和改善其不变性之间取得平衡。测试时，额外的概率计算使 Probout 比 Maxout 具有更高的计算开销。

3.4 损失函数

应为每项具体任务选择合适的损失函数。本小节介绍四种有代表性的损失：合页损失（hinge loss）、Softmax 损失、对比损失和三元组损失。

3.4.1 合页损失（Hinge Loss）

合页损失常用于训练支持向量机（SVM）等大间隔分类器。式 (16) 给出多类 SVM 的合页损失，其中 w 为分类器的权重向量，$y^{(i)}$ 是第 i 个样本在 K 个类别中的正确标签，取值为 1 到 K。

$$\mathcal{L}_{hinge} = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{K}[\max(0, 1 - \delta(\boldsymbol{y}^{(i)}, j)\mathbf{w}^T\mathbf{x}_i)]^p \tag{16}$$

当 $y^{(i)} = j$ 时，$\delta(y^{(i)}, j) = 1$，否则为 −1。$p = 1$ 时，式 (16) 是合页损失（L_1 损失）；$p = 2$ 时是平方合页损失（L_2 损失）[61]。L_2 损失可微，对严重违反间隔约束的点比 L_1 损失惩罚更重。文献 [19] 在深度网络中比较了 Softmax 和 L_2-SVM，在 MNIST [62] 上的结果显示 L_2-SVM 优于 Softmax。

3.4.2 Softmax 损失

Softmax 损失常用于分类，实质是把 Softmax 函数与多项逻辑损失结合起来。设训练集由 N 个样本 $(x^{(i)}, y^{(i)})$ 组成，$x^{(i)}$ 是第 i 个输入图像块，$y^{(i)}$ 是它在 K 个类别中的目标标签。第 i 个输入属于第 j 类的预测值经过 Softmax 变换：

$$p_j^{(i)} = \exp(z_j^{(i)}) / \Sigma_{l=1...K} \exp(z_l^{(i)})。$$

其中 $z_j^{(i)}$ 通常取自全连接层，可写为 $z_j^{(i)} = w_j^T a^{(i)} + b_j$。Softmax 将各类预测转成非负数，并归一化为类别上的概率分布。由这些预测概率计算多项逻辑损失，即 Softmax 损失：

$$\mathcal{L}_{softmax} = -\frac{1}{N}[\sum_{i=1}^{N}\sum_{j=1}^{K} 1\{\boldsymbol{y}^{(i)} = j\}\log p_j^{(i)}] \tag{17}$$

Liu 等人 [63] 提出大间隔 Softmax（L-Softmax）损失，将角度间隔引入特征向量 $a^{(i)}$ 与权重矩阵第 j 列 w_j 的夹角 θ_j。它对相应目标类别的预测为

$$p_j^{(i)} = \frac{e^{\|\mathbf{w}_j\|\|\mathbf{a}^{(i)}\|\psi(\theta_j)}}{e^{\|\mathbf{w}_j\|\|\mathbf{a}^{(i)}\|\psi(\theta_j)} + \sum_{l\neq j} e^{\|\mathbf{w}_l\|\|\mathbf{a}^{(i)}\|\cos(\theta_l)}} \tag{18}$$

$$\psi(\theta_j) = (-1)^k\cos(m\theta_j) - 2k, \theta_j \in [k\pi/m, (k+1)\pi/m] \tag{19}$$

k 为 0 到 m−1 的整数，m 控制类间间隔。$m = 1$ 时，L-Softmax 退化为原始的 Softmax 损失。调节类间间隔 m，可以为学习设定更难的目标，从而有效抑制过拟合。作者在 MNIST、CIFAR-10 和 CIFAR-100 上验证了 L-Softmax 的效果，发现它优于原始的 Softmax。

3.4.3 对比损失（Contrastive Loss）

对比损失常用于训练孪生网络（Siamese network）[64–67]。这种弱监督方案从标记为“匹配”或“不匹配”的样本对中学习相似性度量。设第 i 对数据为 ($x\alpha^{(i)}$, $x\beta^{(i)}$)，其通过第 l 层的输出记为 ($z\alpha^{(i,l)}$, $z\beta^{(i,l)}$)，其中 l 取 1 到 L。文献 [65, 66] 分别将配对的两幅图像送入两个相同的卷积网络，再将末层特征向量送入损失计算模块。训练使用的对比损失为

$$\mathcal{L}_{contrastive} = \frac{1}{2N}\sum_{i=1}^{N}(\mathrm{y})d^{(i,L)} + (1-\mathrm{y})\max(\mathrm{m} - d^{(i,L)}, 0) \tag{20}$$

其中 $d^{(i,L)} = \|z\alpha^{(i,L)} - z\beta^{(i,L)}\|_2^2$，m 是作用于不匹配样本对的间隔参数。匹配时 y = 1，不匹配时 y = 0。

Lin 等人 [68] 发现，采用这种单间隔损失在所有样本对上微调网络，会使检索效果大幅下降；只用不匹配的样本对微调，则能较好地保持性能。仅在不匹配对上微调时召回率稳定，这提示召回率下降主要与损失函数处理匹配对的方式有关。他们因此提出双间隔损失，增加一个专门作用于匹配对的间隔参数。他们在每一层 l 定义对比损失，并同时对各层损失进行反向传播。其定义为

$$\mathcal{L}_{d-contrastive} = \frac{1}{2N}\sum_{i=1}^{N}\sum_{l=1}^{L}(\mathrm{y})\max(d^{(i,l)} - m_1, 0) + (1-\mathrm{y})\max(\mathrm{m}_2 - d^{(i,l)}, 0) \tag{21}$$

实验中，他们发现两个间隔参数可以设为相等，即 $m_1 = m_2 = m$，并可由抽样得到的匹配与不匹配图像对的分布学习该值。

3.4.4 三元组损失（Triplet Loss）

三元组损失 [69] 每次考虑三个样本。第 i 个三元组由锚样本 $x_a^{(i)}$、与锚同类的正样本 $x_p^{(i)}$，以及负样本 $x_n^{(i)}$ 组成。相应的特征表示记为 $z_a^{(i)}$、$z_p^{(i)}$、$z_n^{(i)}$。三元组损失定义为

$$\mathcal{L}_{triplet} = \frac{1}{N}\sum_{i=1}^{N}\max\{d_{(a,p)}^{(i)} - d_{(a,n)}^{(i)} + m, 0\} \tag{22}$$

其中 $d_{ap}^{(i)} = \|z_a^{(i)} - z_p^{(i)}\|_2^2$，$d_{an}^{(i)} = \|z_a^{(i)} - z_n^{(i)}\|_2^2$。目标是缩短锚样本与正样本的距离，拉大锚样本与负样本的距离。

但在某些情况下，随机选择锚样本会产生不恰当的判断。例如，当 $d_{np}^{(i)} < d_{ap}^{(i)} < d_{an}^{(i)}$ 时，三元组损失仍可能为零，此三元组就会在反向传播中被忽略。Liu 等人 [70] 为此提出耦合簇（Coupled Clusters，CC）损失，直接在正样本集合和负样本集合上定义损失。他们用簇中心代替随机抽取的锚，使正样本聚在一起，负样本与其保持较远距离，从而比原始三元组损失更可靠。其定义为

$$\mathcal{L}_{cc} = \frac{1}{N^p}\sum_{i=1}^{N^p}\frac{1}{2}\max\{\|\boldsymbol{z}_p^{(i)} - c_p\|_2^2 - \|\boldsymbol{z}_n^{(*)} - c_p\|_2^2 + m, 0\} \tag{23}$$

N_p 是每个集合中的样本数，$c_p = (\Sigma_i z_p^{(i)})/N_p$ 是估计的正样本簇中心。$x_n^{(*)}$ 是离此中心最近的负样本，$z_n^{(*)}$ 是其特征表示。三元组损失及其变体已广泛用于重识别 [71]、验证 [70] 和图像检索 [72] 等任务。

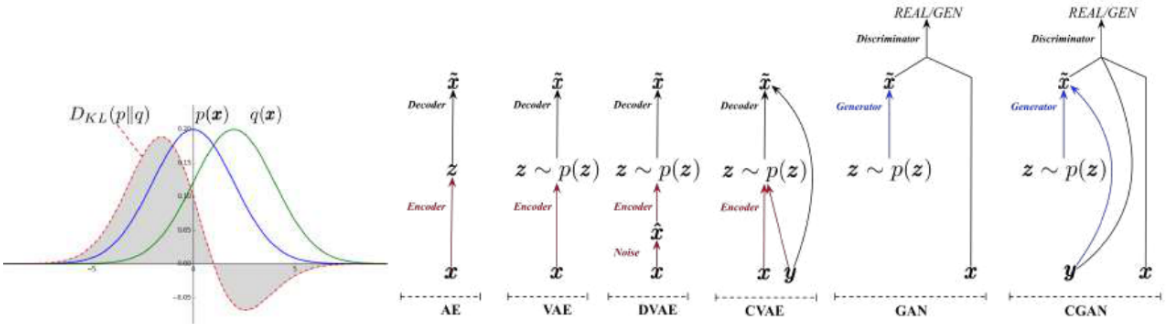
3.4.5 Kullback–Leibler 散度

Kullback–Leibler 散度（KLD 或 KL 散度）用来衡量同一个离散变量 x 上两个概率分布 p(x)、q(x) 之间的差异。这一度量不对称，见图 7(a)。原文所用方向及其定义如下：

$$\mathcal{D}_{KL}(p||q) = -H(p(x)) - E_p[\log q(x)] \tag{24}$$

$$= \sum_x p(x)\log p(x) - \sum_x p(x)\log q(x) = \sum_x p(x)\log\frac{p(x)}{q(x)} \tag{25}$$

H(p(x)) 是 p(x) 的香农熵。按式 (24) 的符号约定，$-E_p[\log q(x)]$ 是 p 与 q 的交叉熵。



(a) KL divergence (b) Autoencoders variants and GAN variants

图 7 （a）两个高斯分布之间的 KL 散度；（b）自编码器及其变体（AE、VAE [73]、DVAE [74] 和 CVAE [75]），以及生成对抗网络及其变体（GAN [76] 和 CGAN [77]）。

读图 左图的 p 和 q 为两种概率分布。右图中 Encoder、Decoder 分别为编码器、解码器，Noise 为加噪，Generator、Discriminator 为生成器、判别器，REAL/GEN 表示判别样本是真实的还是生成的。x 为输入数据，z 为潜变量，y 为条件信息；AE、VAE、DVAE、CVAE 分别为自编码器及其变分、去噪变分和条件变分版本，GAN、CGAN 为生成对抗网络及其条件版本。

在多种自编码器（autoencoder，AE）的目标函数中，KL 散度被广泛用于度量信息损失 [78–80]。代表性变体包括稀疏自编码器 [81, 82]、去噪自编码器 [78] 和变分自编码器（VAE）[73]。VAE 用贝叶斯推断来解释潜在表示。其编码器把样本 x “压缩”为潜在表示 z ~ qφ(z|x)，解码器再把这一表示映射回数据空间，得到尽量接近输入的重建样本 x̃ ~ pθ(x|z)。φ、θ 分别为编码器和解码器参数。文献 [73] 中，VAE 尝试最大化对数似然 log p(x|φ,θ) 的变分下界：

$$\mathcal{L}_{vae} = \mathrm{E}_{z\sim q_\phi(z|x)}[\log p_\theta(x|z)] - \mathcal{D}_{KL}(q_\phi(z|x)||p(z)) \tag{26}$$

第一项衡量重建质量；KL 散度项使提议分布 qφ(z|x) 受到先验分布 p(z) 的约束。p(z) 通常取标准正态分布 [73]、离散分布 [75] 或具有几何解释的分布 [83]。原始 VAE 之后出现了多种变体 [74, 75, 84]。条件 VAE（CVAE）[75, 84] 从条件分布 x̃ ~ pθ(x|y,z) 生成样本。去噪 VAE（DVAE）[74] 则从受损输入 x̂ 恢复原始输入 x [78]。

Jensen–Shannon 散度（JSD）是 KL 散度的一种对称形式，用于度量 p(x) 和 q(x) 的相似程度：

$$\mathcal{D}_{JS}(p||q) = \frac{1}{2}\mathcal{D}_{KL}\left(p(x)\middle\|\frac{p(x)+q(x)}{2}\right) + \frac{1}{2}\mathcal{D}_{KL}\left(q(x)\middle\|\frac{p(x)+q(x)}{2}\right) \tag{27}$$

最小化 JSD 可以使两个分布尽可能接近。JSD 已成功用于生成对抗网络（GAN）[76, 85, 86]。VAE 直接对 x 与 z 的关系建模，GAN 则专门围绕生成任务设定优化目标 [85]。GAN 希望找到最能区分真实数据与生成数据的判别器 D，同时促使生成器 G 拟合真实数据分布。判别器与生成器之间的极小极大博弈可表示为以下目标函数：

$$\min_G \max_D \mathcal{L}_{gan}(D,G) = \mathbb{E}_{x\sim p(x)}[\log D(x)] + \mathbb{E}_{z\sim q(z)}[\log(1-D(G(z)))] \tag{28}$$

原始 GAN 论文 [76] 表明，给定生成器 G*，最优判别器为 D*(x) = p(x)/[p(x)+q(x)]，其中 q(x) 是生成数据的分布。此时式 (28) 等价于最小化 p(x) 与 q(x) 的 JSD。若 G 和 D 具有足够的表示能力，q(x) 将向 p(x) 收敛。与条件 VAE 类似，条件 GAN（CGAN）[77] 也接收额外信息 y 作为输入，生成以 y 为条件的样本。实际应用中，GAN 训练过程以不稳定著称 [87, 88]。

3.5 正则化

过拟合是深度 CNN 中不可忽视的问题，正则化能够有效缓解它。本小节介绍几种有效技术，包括 ℓ_p 范数、Dropout 和 DropConnect。

3.5.1 ℓ_p 范数正则化

正则化通过在目标函数中加入惩罚模型复杂度的项来修改训练目标。若损失函数为 L(θ,x,y)，正则化后的损失为

$$E(\theta, \mathbf{x}, \mathbf{y}) = \mathcal{L}(\theta, \mathbf{x}, \mathbf{y}) + \lambda R(\theta) \tag{29}$$

其中 R(θ) 为正则项，λ 为正则化强度。

ℓ_p 范数正则项通常写为 $R(\theta) = \Sigma_j \|\theta_j\|_p^p$。$p \geq 1$ 时，ℓ_p 范数是凸函数，更便于优化，这是其受到重视的原因 [17]。$p = 2$ 时，ℓ_2 正则化通常称为权重衰减。Tikhonov 正则化 [89] 是一种有理论依据的替代方案，鼓励模型对输入中的噪声保持不变。$p < 1$ 时，该惩罚对权重的稀疏性作用更强，但相应函数非凸。

3.5.2 Dropout

Hinton 等人 [17] 首先提出 Dropout，并在全连接层中使用它。该方法已被证明能有效缓解过拟合。记 $y = r \odot a(W^T x)$ 为 Dropout 的输出。这里 ⊙ 表示逐元素乘法，$x = [x_1, \ldots, x_n]^T$ 为全连接层的输入，W 是 n 行 d 列的实权重矩阵，r 是长度为 d 的二元向量。r 的元素独立采自参数为 p 的伯努利分布，即 $r_i \sim \text{Bernoulli}(p)$。Dropout 能防止网络过度依赖某个神经元或少数神经元的组合，并迫使网络在缺失某些信息时仍保持预测准确。

此后已有多种改进方法。Wang 等人 [90] 提出快速 Dropout，通过对高斯近似进行采样或积分来加快训练。Ba 等人 [91] 提出自适应 Dropout：用一个与深度网络共享参数的二元信念网络，计算各隐藏变量的丢弃概率。文献 [92] 发现，在 1×1 卷积层前使用标准 Dropout 往往增加训练时间，却未能防止过拟合。因此，作者提出 SpatialDropout，把同一个丢弃掩码值扩展到一整张特征图。该方法在训练数据较少时尤其有效。

3.5.3 DropConnect

DropConnect [45] 将 Dropout 的思路推进一步。Dropout 随机将部分神经元输出置零，DropConnect 随机将权重矩阵 W 中的元素置零。其输出是 $y = a((R \odot W)x)$，其中 $R_{ij} \sim \text{Bernoulli}(p)$。训练时也对偏置施加掩码。三种网络的区别见图 8。

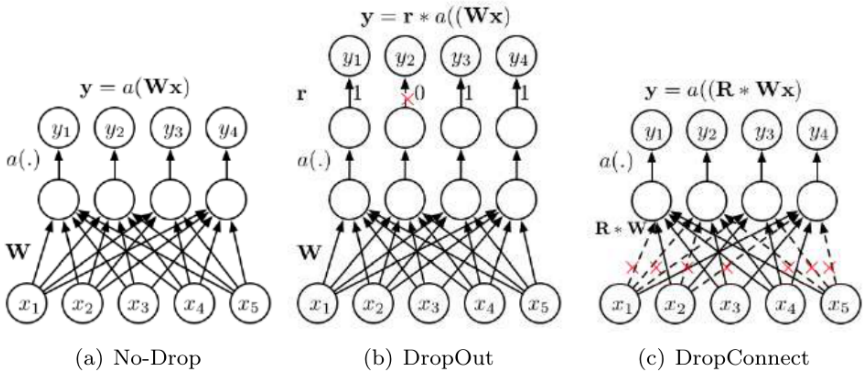


图 8 （a）不施加随机丢弃的网络（No-Drop）；（b）对神经元输出施加掩码的 Dropout 网络；（c）对连接权重施加掩码的 DropConnect 网络。

读图 从左向右依次为不作随机屏蔽、随机屏蔽单元输出（Dropout）和随机屏蔽权重连接（DropConnect）。x 为输入，W 为权重，a 为激活函数，y 为输出；中图的 r 作用于激活值，右图的 R 作用于权重。红叉表示本次训练计算中被屏蔽的项。

3.6 优化

本小节讨论优化 CNN 的若干关键技术。

3.6.1 数据增强

深度 CNN 尤其依赖大量训练数据。相对于模型参数的数量，训练样本常显不足，数据增强 [8] 提供了一种有效途径。它对已有数据施加变换，在保留样本本质属性的前提下产生新样本。常见方法包括采样 [8]、镜像 [93]、旋转 [94]、平移 [95] 等简单的几何变换，以及多种光度变换 [96]。Paulin 等人 [97] 提出贪心策略，从候选集中选出最佳变换。该策略需要多次重新训练模型，候选较多时计算代价很高。Hauberg 等人 [98] 提出通过随机生成微分同胚变换来增强数据。Xie 等人 [99] 和 Xu 等人 [100] 还提出从互联网收集图像，以改善细粒度识别任务的学习效果。

3.6.2 权重初始化

深度 CNN 参数众多，损失函数又是非凸的 [101]，因此训练十分困难。要加快收敛并避免梯度消失，适当的初始化是最重要的前提之一 [102, 103]。偏置参数可以初始化为零，权重则需要谨慎设置，以打破同一层中不同隐藏单元的对称性。例如，各层若都只把输入放大 k 倍，经过 L 层后最终输出就会将原输入放大 k^L 倍。$k > 1$ 时，输出层数值可能极大；$0 < k < 1$ 时，输出与梯度则逐渐缩小。

Krizhevsky 等人 [8] 从均值为零、标准差为 0.01 的高斯分布中初始化权重，并将第 2、4、5 个卷积层及所有全连接层的偏置设为 1。另一种知名随机初始化方法是文献 [104] 提出的 Xavier 初始化。其权重取自零均值、方差为 $2/(n_{in}+n_{out})$ 的高斯分布，其中 n_{in} 是输入连接数，n_{out} 是输出连接数。Xavier 因此能够根据输入和输出神经元数自动决定初始化尺度，使信号在通过多层后仍保持在合理范围内。Caffe 中的一个变体只使用输入连接数，以简化实现。文献 [56] 将 Xavier 推广到整流非线性，推导出专门考虑 ReLU 的稳健初始化方法。在作者报告的实验中，该方法能够使极深模型（如 [10] 所述）的训练收敛，而 Xavier [104] 在同样情况下未能做到。

原文脚注 1 Caffe 项目地址：https://github.com/BVLC/caffe

Saxe 等人 [105] 独立表明，对线性网络，正交归一矩阵初始化比高斯初始化有效得多；这一方法也适用于具有非线性的网络。Mishkin 等人 [102] 将其扩展为迭代过程，提出逐层单位方差初始化。可将它理解为正交归一初始化与仅对第一个小批量进行批归一化的结合（见第 3.6.4 节）。它与批归一化均采

用单位方差归一化，但它还用正交归一化来初始化权重，有助于使层内的激活有效去相关。文献 [106, 107] 采用这种初始化技术后，性能显著提高。

3.6.3 随机梯度下降

反向传播是标准的网络训练算法，配合梯度下降更新参数。研究者已提出多种基于梯度下降的优化算法 [108, 109]。标准梯度下降按 $\theta_{t+1} = \theta_t - \eta \nabla_\theta E[L(\theta_t)]$ 更新目标函数 $L(\theta)$ 的参数。其中期望 $E[L(\theta_t)]$ 在整个训练集上计算，η 为学习率。随机梯度下降（SGD）[21] 通过从训练集中随机选取的单个样本 $(x^{(t)}, y^{(t)})$ 估计梯度：

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta_t \nabla_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}_t; \boldsymbol{x}^{(t)}, \boldsymbol{y}^{(t)}) \tag{30}$$

实践中，SGD 通常每次使用一个小批量而非单一样本更新参数。这样可以减小参数更新的方差，使收敛更稳定。学习率 η_t 控制收敛速度。但小批量 SGD 并不保证良好收敛，仍有若干问题待解决。首先，合适的学习率不易选定。一种常用做法是在训练初期使用能稳定收敛的常数学习率，待收敛放慢时再调低学习率。另有研究提出学习率调度方案 [110, 111]，在训练期间调节学习率。

动量方法 [108] 通过积累相关方向上的速度向量，使当前更新考虑此前各批次的梯度，从而加快训练。经典动量更新为

$$\boldsymbol{v}_{t+1} = \gamma \boldsymbol{v}_t - \eta_t \nabla_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}_t; \boldsymbol{x}^{(t)}, \boldsymbol{y}^{(t)}) \tag{31}$$

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t + \boldsymbol{v}_{t+1} \tag{32}$$

其中 v_{t+1} 为当前速度向量，动量系数 γ 通常取 0.9。Nesterov 动量 [103] 是另一种在梯度下降中引入动量的方法：

$$\boldsymbol{v}_{t+1} = \gamma \boldsymbol{v}_t - \eta_t \nabla_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}_t + \gamma \boldsymbol{v}_t; \boldsymbol{x}^{(t)}, \boldsymbol{y}^{(t)}) \tag{33}$$

经典动量先在当前位置计算梯度，再沿更新后的累积方向移动；Nesterov 动量先沿已有累积方向 γv_t 前移，在预估位置计算梯度，再更新。这样的前瞻有助于避免优化步伐过快，并可改善性能 [112]。

并行 SGD 方法 [22, 113] 扩展了 SGD，使其适用于并行的大规模机器学习。在标准同步 SGD 中，一台慢机器就可能拖延整个训练；异步机制允许其他机器继续更新，仅慢机器自身的进度受到影响。Jeffrey Dean 等人 [114] 使用名为 Downpour SGD 的异步方法，加快具有大量 CPU 的集群上的分布式训练。也有研究在多个 GPU 上采用异步 SGD。Paine 等人 [115] 将异步 SGD 与 GPU 结合，相较单机训练获得数倍加速。Zhuang 等人 [116] 使用多个 GPU 异步计算梯度、更新全局模型参数，四个 GPU 相对于单 GPU 得到 3.2 倍加速。

SGD 不一定收敛。当性能不再提高时，可以终止训练。常用的防止过度训练的方法是早停 [117]，即根据训练过程中验证集上的表现停止优化。可用不同标准控制训练时长，例如固定训练轮数，或一直训练到误差达到预定值 [118]。停止策略需谨慎选择 [119]：只要网络的泛化能力仍在提高且没有出现过拟合，就应允许训练继续进行。

3.6.4 批归一化（Batch Normalization）

数据归一化通常是预处理的第一步。全局归一化使整个数据集具有零均值和单位方差。然而数据在深层网络中流动时，内部各层的输入分布会改变，这可能损害网络的学习能力与准确率。Ioffe 等人 [120] 提出批归一化（BN），以部分缓解这一现象。他们将其描述为协变量偏移问题，并通过归一化来固定层输入的均值和方差；这些统计量在每个小批量上估计。对 d 维输入 $x = [x_1, x_2, \ldots, x_d]^T$，先将第 k 个分量归一化：

$$\hat{x}_k = (x_k - \mu_{\mathcal{B}}) / \sqrt{\delta_{\mathcal{B}}^2 + \epsilon} \tag{34}$$

μ_B 和 δ_B² 分别为该小批量的均值和方差，ε 为常数。为增强表示能力，再作如下变换：

$$y_k = \mathrm{BN}_{\gamma,\beta}(x_k) = \gamma \hat{x}_k + \beta \tag{35}$$

γ 和 β 均为可学习参数。作者列举了批归一化相对于全局数据归一化的多项优点。首先，BN 能减轻内部协变量偏移。其次，它能减弱梯度对参数尺度及初始化大小的依赖，有利于梯度在网络中传播，因此能以更高的学习率训练而降低发散风险。BN 还能起正则化作用，减少对 Dropout 的需求。最后，即使采用容易饱和的非线性激活函数，BN 也有助于避免模型陷在饱和区。

3.6.5 捷径连接（Shortcut Connections）

如前所述，规范化的初始化 [8] 和 BN [120] 可以缓解深度 CNN 的梯度消失问题。作者指出，即使采用这些方法并抑制了过拟合，较深网络的优化依然可能困难，性能甚至低于较浅网络 [56, 104, 105]。深度增加带来的这种优化问题称为退化问题。

长短期记忆（LSTM）网络 [121] 利用门函数决定一个神经元的激活值有多少要经过变换、有多少直接传递。受其启发，Srivastava 等人 [122] 提出高速公路网络（Highway Networks），使优化几乎任意深度的网络成为可能。其输出为

$$\mathbf{x}_{l+1} = \phi_{l+1}(\mathbf{x}_l, \mathbf{W}_H) \cdot \tau_{l+1}(\mathbf{x}_l, \mathbf{W}_T) + \mathbf{x}_l \cdot (\mathbf{1} - \tau_{l+1}(\mathbf{x}_l, \mathbf{W}_T)) \tag{36}$$

x_l、x_{l+1} 分别是第 l 个高速公路块的输入与输出。τ 是变换门，φ 通常由仿射变换及其后的非线性激活函数组成，一般也可以取其他形式。门控机制要求层的输入、输出形状相同，并使数十层乃至数百层的高速公路网络可以有效训练。门的输出会随输入样本显著变化，说明网络学到的结构能够针对具体样本动态选择数据的传递路径。

另一路独立发展的残差网络（ResNet）[12] 与 LSTM 单元共享这一核心思想。高速公路网络为各神经元学习专门的门控权重，ResNet 的捷径连接不设置门控。

未经变换的输入直接传到输出端，因而所需参数更少。ResNet 的输出可写为

$$\mathbf{x}_{l+1} = \mathbf{x}_l + f_{l+1}(\mathbf{x}_l, \mathbf{W}_F) \tag{37}$$

其中 f_l 表示含权重的网络层，可以是卷积、批归一化（BN）、ReLU 或池化等操作的复合函数。采用残差块以后，任一较深单元的激活都可以写成较浅单元的激活与一个残差函数之和。这也意味着梯度可以直接传到较浅的单元，因此，与学习原始映射相比，深层 ResNet 更容易优化，训练极深网络的效率也更高。普通前馈网络的梯度实质上需要经过一连串矩阵与向量的乘法，随着网络加深，梯度可能消失。

在最初的 ResNet 之后，He 等人 [123] 又提出预激活（pre-activation）变体，通过一系列实验说明，采用恒等捷径连接的网络最容易学习。他们还发现，把 BN 移到前面，效果明显好于在加法之后使用 BN。在其对比实验中，采用 BN + ReLU 预激活的残差网络，准确率高于先前的 ResNet [12]。受文献 [123] 启发，Shen 等人 [124] 对卷积层的输出引入权重因子，使可训练层逐步参与计算。
Inception-v4 的论文 [42] 也报告，在 Inception 模块两端添加恒等跳跃连接，既能加快训练，也能改善性能。初版 ResNet 和预激活 ResNet 都很深，而每层的宽度较小。Wide ResNet [125] 则减少深度、增加宽度，在 CIFAR-10、CIFAR-100 和 SVHN 上取得了出色结果，不过作者撰写本综述时，这些结论尚未在 ImageNet 的大规模图像分类任务中得到验证。随机深度 ResNet 对每个小批次随机丢弃一部分层，以恒等映射绕过它们。Singh 等人 [126] 将随机深度 ResNet 与 Dropout 结合，推广了 Dropout 和随机深度网络的方法。所得模型可以看作普通 ResNet、Dropout ResNet 和随机深度 ResNet 的集成。ResNets in ResNets（RiR）[127] 将经典卷积网络和残差网络融合，每个 RiR 块既含残差单元，也含非残差块。RiR 可以学习每个残差块该使用多少个卷积层。ResNets of ResNets（RoR）[128] 改进了 ResNet 的结构，在早期 ResNet [12] 单层级捷径连接的基础上使用多层级捷径连

接。DenseNet [129] 把跳跃连接的思想进一步推广：每一层的输出都连到该模块中所有后续层。从 ResNet [12, 123]、Highway 网络 [122] 和 Inception 网络 [42] 中，都可以看到用捷径连接促进极深网络训练的趋势。

原文脚注 2 ImageNet 网站：www.image-net.org。

4 卷积神经网络的快速计算

随着计算机视觉和机器学习任务越来越具挑战性，深度神经网络模型也日益复杂。这些表达能力强的模型需要更多训练数据来避免过拟合。与此同时，大规模训练数据带来了新的问题，例如怎样在可以接受的时间内完成训练。本节介绍几类加快 CNN 计算的方法。

4.1 快速傅里叶变换（FFT）

Mathieu 等人 [49] 用快速傅里叶变换（FFT）在傅里叶域执行卷积运算。这种方法具有多项优点。第一，同一滤波器需要与一个小批次内的多幅图像卷积，其傅里叶变换可以重复使用。第二，反向传播要分别计算滤波器和输入图像的梯度，两者可以复用输出梯度的傅里叶变换。第三，跨输入通道求和可以在傅里叶域完成，因此每幅图像的每个输出通道只需做一次逆傅里叶变换。已有多种基于 GPU 的库用于加速训练和测试，例如 cuDNN [130] 和 fbfft [131]。不过，FFT 卷积需要将滤波器填充到与输入同样大小，还需要额外内存来储存傅里叶域的特征图。步幅大于 1 时，这项开销尤其高，而许多当时先进的网络都采用了这样的步幅，例如文献 [132] 和 [10] 中网络的早期层。FFT 能加快训练和测试，另一方面，小尺寸滤波器在 ResNet [12]、GoogLeNet [10] 等 CNN 中越来越重要，由此出现了专门针对小滤波器的方案，即 Winograd 最少滤波算法 [133]。其思路与 FFT 相似：在逆变换之前，先于变换域中对各通道的 Winograd 卷积结果求和，从而提高推理效率。

4.2 结构化变换

低秩矩阵分解已在多种场景中用于改善优化问题。设矩阵 C 的尺寸为 m × n、秩为 r，则存在分解 C = AB，其中 A 是 m × r 的列满秩矩阵，B 是 r × n 的行满秩矩阵。因此可以用 A 和 B 替代 C。若希望分解后的参数量低于原来的 p 倍，须满足 mr + rn < pmn，即 C 的秩应满足 r < pmn/(m + n)。这样，空间复杂度由 O(mn) 降为 O(r(m + n))，时间复杂度也由 O(mn) 降为 O(r(m + n))。Sainath 等人 [134] 对深层 CNN 最后一个权重层做低秩矩阵分解，在准确率仅略有损失的情况下，训练提速约 30% ~ 50%。类似地，Xue 等人 [135] 对深层 CNN 的每一层实施奇异值分解，模型规模减小 71%，准确率的相对损失不到 1%。文献 [136] 证实了深度神经网络参数中的冗余。在此启发下，Denton 等人 [137] 和 Jaderberg 等人 [138] 分别研究了卷积滤波器内部的冗余，并提出近似算法来减少计算量。Novikov 等人 [139] 推广了低秩分解的思想，把权重矩阵视为多维张量，再通过张量列车分解（Tensor-Train decomposition）[140] 来减少全连接层的参数量。

自适应 Fastfood 变换是用于矩阵近似的 Fastfood 变换 [141] 的推广。它用自适应 Fastfood 变换对全连接层的 n × n 实权重矩阵 C 重新参数化，即 Cx = (Ð₁HÐ₂ΠHÐ₃)x。这里 Ð₁、Ð₂、Ð₃ 是由参数构成的对角矩阵，Π 是随机置换矩阵，H 表示沃尔什–哈达玛矩阵。自适应 Fastfood 变换的空间复杂度为 O(n)，时间复杂度为 O(n log n)。

循环矩阵在空间和计算效率上具有显著优势 [142, 143]。Cheng 等人 [144] 针对全连接层参数化中的冗余，为权重矩阵施加循环结构以加快计算，并利用 FFT 进一步提速。若全连接层以 n × n 的实循环矩阵 C 为参数矩阵，则对 n 维实输入向量 x，可以用 FFT 和逆变换 IFFT 高效计算输出。原文给出的算式

写作 CDx = ifft(fft(v)) ∘ fft(x)，其中 ∘ 表示逐元素相乘，v 是由 C 确定的 n 维实向量，D 是用来提高模型表示能力的随机符号翻转矩阵。【译注】按循环矩阵与向量相乘的常用 FFT 写法，逆变换应作用于频域逐元素乘积。若 v 是 C 的第一列，且输入先经过 D，公式为 CDx = ifft(fft(v) ∘ fft(Dx))。该方法将时间复杂度从 $O(n^2)$ 降为 $O(n \log n)$，将空间复杂度从 $O(n^2)$ 降为 $O(n)$。Moczulski 等人 [145] 进一步推广循环结构，将对角矩阵与正交离散余弦变换（DCT）交替排列。所得变换 $ACDC^{-1}$ 的空间复杂度为 $O(n)$，时间复杂度为 $O(n \log n)$。

4.3 低精度计算

CNN 参数往往只做很小的更新，以浮点数表示这些更新十分自然。不过最终得到的参数可能包含大量冗余信息 [146]。为减少冗余，二值神经网络（BNN）将计算输出所涉及的部分或全部算术运算限制在二值表示上。

神经网络层有三个可以二值化的部分：输入激活、突触权重和输出激活。三者同时采用二值表示称为完全二值化；其中一项或两项采用二值表示称为部分二值化。Kim 等人 [147] 研究了完全二值化的情形，预先规定一部分连接的权重为 0，其余连接权重均为 1。他们的网络只需要同或（XNOR）运算和比特计数，在 MNIST 数据集上报告的准确率为 98.7%。XNOR-Net [148] 将二值卷积神经网络应用到 ImageNet，网络拓扑借鉴了 AlexNet、ResNet 和 GoogLeNet。完全二值化的最高 top-1 准确率为 51.2%，部分二值化为 65.5%。DoReFa-Net [149] 同时研究了在前向传播和反向传播中降低精度。其实验涵盖部分二值化和完全二值化，本文按原文所列顺序给出的 ImageNet top-1 准确率分别为 43% 和 53%。Courbariaux 等人 [150] 描述了怎样用完全二值化和批归一化层训练全连接网络及 CNN，在 MNIST、SVHN 和 CIFAR-10 上均取得了有竞争力的准确率。

4.4 权重压缩

已有许多工作致力于减少卷积层和全连接层的参数量。下面简要介绍向量量化、剪枝和哈希三类方法。

向量量化（VQ）可压缩全连接层，缩小 CNN 模型。标量量化把大量数值映射到一个较小的数值集合 [151]。类似地，VQ 把一组数值作为整体进行量化，而不用逐个处理。2013 年，Denil 等人 [136] 证实神经网络参数存在冗余，并用 VQ 大幅减少深度模型中需要独立调整的参数。Gong 等人 [152] 从信息论出发研究向量量化对 CNN 参数的压缩，得到与文献 [136] 相近的参数预测结果。他们还发现，VQ 相较已有的矩阵分解方法有明显优势；在各种 VQ 方法中，乘积量化等结构化量化方法的效果也明显优于残差量化 [153] 和标量量化 [154] 等方法。

权重压缩还可以使用剪枝。通过永久移除较不重要的连接 [155]，可以减少 CNN 中的参数和运算，使较小的网络继承较大前身网络的知识，同时保持相近的性能。Han 等人 [146, 156] 根据权重幅值剪枝，在网络中引入细粒度稀疏性。只要某个权重的绝对值小于给定的标量阈值，它就被剪去。Gao 等人 [157] 扩展了基于幅值的方法，允许恢复先前迭代中被剪去的权重，并使剪枝和再训练紧密配合，以取得更大的压缩率。Yang 等人 [158] 考虑权重之间的相关性，提出能耗感知剪枝算法，直接利用对 CNN 能耗的估计来指导剪枝。也有研究考察粗粒度剪枝。Hu 等人 [159] 建议移除在验证集上经常产生零激活的滤波器。Srinivas 等人 [160] 将相似的滤波器合并为一个，Mariet 等人 [161] 则合并输出激活相似的滤波器。

如何设计合适的哈希技术来加快 CNN 训练或节省存储空间，也是一个值得研究的问题。HashedNets [162] 用哈希函数把连接权重分入不同的哈希桶，同一桶中的全部连接共享一个参数值，以此缩小模型。这种网络大幅降低了存储成本，同时在图像分类上大体保留了泛化性能。正如 Shi 等人 [163] 和

Weinberger 等人 [164] 指出的，稀疏性可以减少哈希碰撞，使特征哈希更有效。哈希网络还可以结合剪枝，进一步节省参数。

4.5 稀疏卷积

近期有多项工作探索使卷积层权重稀疏化 [165, 166]。Liu 等人 [165] 研究基滤波器的稀疏表示，同时利用卷积核的通道间冗余和通道内冗余，达到 90% 的稀疏化程度。Wen 等人 [166] 提出结构化稀疏学习（SSL）方法，将优化范围从卷积层的权重扩展到它们的超参数，同时优化滤波器大小、网络深度和局部连接方式。Bagherinezhad 等人 [167] 提出了基于查表的卷积神经网络（LCNN）。该网络经训练得到可覆盖 CNN 权重空间的丰富字典，只需少数几次查表便能编码卷积。他们使用一个字典和两个张量解码卷积层权重。同一层的所有权重滤波器共享字典，使 CNN 能够从很少的训练样本中学习。与标准 CNN 相比，LCNN 可以在较少的迭代次数内取得更高准确率。

5 卷积神经网络的应用

本节介绍近期依靠 CNN 达到当时先进水平的一些工作，涉及图像分类、目标跟踪、姿态估计、文本检测、视觉显著性检测、动作识别、场景标注、语音处理和自然语言处理等。

5.1 图像分类

CNN 用于图像分类已有很长历史 [168–171]。它能够联合学习特征和分类器，在大规模数据集上可取得比其他方法更高的分类准确率 [8, 9, 172]。2012 年，大规模图像分类取得突破。Krizhevsky 等人 [8] 开发 AlexNet，在 ILSVRC 2012 取得最佳成绩。继 AlexNet 成功之后，若干工作通过缩小滤波器尺寸 [11] 或增加网络深度 [9, 10]，使分类准确率进一步显著提高。

面对类别很多的图像分类任务，构建层次化分类器是一种常见策略 [173]。文献 [174] 是较早将类别层次结构引入 CNN 的尝试之一，提出了一种带树形先验的判别式迁移学习方法。该方法利用类别层次在相关类别间共享信息，从而改善训练样本很少的类别的识别效果。类似地，Wang 等人 [175] 构建树形结构，学习识别细分类别所需的细粒度特征。Xiao 等人 [176] 提出训练方法，使网络既能逐步增长，也能分层生长。他们按相似性将类别分组，并使其自组织到不同层级。Yan 等人 [177] 把深层 CNN 嵌入类别层次结构，提出层次化深层 CNN（HD-CNN）。他们把分类分为两步：先用粗类别 CNN 分类器区分较容易判别的类别，再把难以区分的类别送入下游细类别分类器作进一步预测。这种结构遵循从粗到细的分类思路，能以可接受的额外复杂度换取更低的错误率。

子类别分类是图像分类中另一个快速发展的分支。已有鸟类 [178]、犬类 [179]、汽车 [180] 和植物 [181] 等细粒度图像数据集。物体部件信息有助于细粒度分类 [182]。一般而言，通过定位物体的重要部件，并学习能区分不同类别的部件外观表示，可以提高准确率。Branson 等人 [183] 据此提出部件检测方法，从经过姿态归一化的多个区域提取 CNN 特征，并借助部件标注学习一个紧凑的姿态归一化空间。他们还建立了一个模型，把较低层特征与姿态归一化后的提取过程结合，把较高层特征与未经对齐的图像特征结合，从而提高分类准确率。Zhang 等人 [184] 提出基于部件的 R-CNN，可以学习整个物体的检测器及其部件检测器。他们用选择性搜索 [185] 生成候选部件，再用非参数几何约束提高部件定位精度。Lin 等人 [186] 将部件定位、对齐与分类整合为识别系统 Deep LAC。该系统包含三个子网络：定位子网络估计部件位置，对齐子网络接收该位置并进行模板对齐 [187]，分类子网络根据姿态已对齐的部件图像预测类别标签。他们还提出一种数值联接函数，把各子网络连接起来，使其在训练和测试时作为整体工作。

上述方法都借助部件标注做监督训练。不过，此类标注不易收集，这些系统也很难扩展到大量不同的细粒度类别。为此，一些研究者提出以无监督方式寻找局部部件或区域。Krause 等人 [188] 将多个局部学习特征表示集成起来，用于细粒度分类。他们通过协同分割和对齐生成部件，再逐一比较各部件外观，汇总相似度。在后续论文 [189] 中，他们把协同分割和对齐结合成判别式混合方法来生成部件，以帮助细粒度分类。Zhang 等人 [190] 用无监督的选择性搜索产生物体候选区域，再从多个尺度生成的部件候选中筛选有用部件。Xiao 等人 [191] 将视觉注意机制应用于 CNN 的细粒度分类。他们的方法包含三类注意：自下而上的注意提出候选图块，物体级自上而下注意选出与某一物体相关的图块，部件级自上而下注意定位具有判别力的部件。他们结合这三类注意来训练面向特定领域的网络，以查找前景物体或部件并提取判别性特征。Lin 等人 [192] 提出用于细粒度图像分类的双线性模型。它包含两个特征提取器，在图像的每个位置将两者的输出作外积，再通过池化得到整幅图像的描述子。

5.2 目标检测

目标检测一直是计算机视觉中的重要问题 [193–195]。其难点通常在于如何在图像或视频帧中准确、高效地定位物体。利用 CNN 进行检测和定位可以追溯至 20 世纪 90 年代 [196]。但由于缺乏训练数据、计算资源有限，基于 CNN 的目标检测在 2012 年以前进展缓慢。2012 年后，CNN 在 ImageNet 挑战赛上获得成功 [8]，使研究者对 CNN 目标检测再度产生兴趣 [197]。早期的一些工作 [196, 198] 使用滑动窗口，在各位置、各尺度采样的窗口上密集运行 CNN 分类器。一幅图像通常有数十万个候选窗口，因而计算成本很高，难以应用于 Pascal VOC [172]、ImageNet [8] 和 MS COCO [199] 等大规模数据集。

近年来，基于目标候选区域的方法受到广泛关注和研究 [185, 193, 200, 201]。这类方法通常用快速、通用的度量判断采样窗口是否可能包含物体，随后将选出的候选区域交给更精细的检测器，判定其为背景或某一特定物体类别。最知名的此类 CNN 检测器之一是基于区域的 CNN（R-CNN）[202]。它用选择性搜索（SS）[185] 自下而上地提取约 2,000 个可能含有物体的候选区域，随后将各候选区域变形为固定尺寸 227 × 227，用预训练的 CNN 提取其特征，最后用二分类支持向量机（SVM）完成检测。

R-CNN 显著提升了检测性能，但每个区域都要单独运行耗时的 CNN 特征提取器，计算成本依然较高。若干后续工作提出在特征提取阶段共享计算 [9, 28, 132, 202]。OverFeat [132] 在图像金字塔上计算 CNN 特征用于定位和检测，让重叠窗口可以方便地共享计算。空间金字塔池化网络（SPP net）[203] 是 R-CNN [202] 的一种金字塔版本，通过引入空间金字塔池化层放宽输入必须为固定尺寸的限制。SPP net 只对整幅图像提取一次特征图，然后对每个候选窗口进行空间金字塔池化，获得定长表示。它的一个不足是训练分为多个阶段，无法联合训练 CNN 特征提取器和 SVM 分类器来进一步提高准确率。Fast R-CNN [204] 采用端到端训练改进 SPP net，在微调阶段可更新所有网络层，从而简化学习流程、提高检测准确率。随后，Faster R-CNN [204] 引入区域候选网络（RPN）生成候选区域，进一步加快检测。除 R-CNN 系列外，Gidaris 等人 [205] 为目标检测提出了利用多个区域、并结合语义分割信息的模型，将组合特征用于迭代定位机制，并在非极大值抑制之后使用边界框投票方案。Yoo 等人 [206] 把目标检测视为迭代分类问题，通过汇总检测网络给出的量化弱方向预测，获得精确的物体边界框。

目标检测的另一个重要问题是怎样构建有效的训练集，因为性能在很大程度上取决于正、负样本的数量和质量。面向动态变化环境进行交互的智能认知系统 [208] 需要有效训练 CNN，因此，在线自举（亦称难负样本挖掘 [207]）近来受到重视。文献 [209] 提出了训练 CNN 检测模型的新自举技术——在线难例挖掘（OHEM）。它自动选择困难样本，简化训练流程。它只需对每幅图像计算一次特征图，再在这些特征图上为图中所有感兴趣区域（RoI）继续前向计算，因此只增加少量计算就能找到困难样本。

近来，YOLO [210] 和 SSD [211] 均能通过单一检测流程直接预测类别标签。YOLO [210] 把目标检测视为回归问题，预测图像不同空间位置的边界框及其对应的类别概率。整个检测流程由一个网络构成，一次运行即可由整幅图像得到边界框和类别概率，并且可以直接围绕检测性能进行端到端优化。SSD [211] 将边界框的输出空间离散化，在特征图的每个位置设置不同宽高比、不同尺度的一组默认框。借助这种多尺度设置和相应的匹配策略，在论文所述对比中，SSD 的准确率明显高于 YOLO。Lu 等人 [212] 借助超分辨率的优势，提出自上而下的搜索策略，将窗口递归划分为子窗口，并专门训练另一个网络来决定怎样划分。

5.3 目标跟踪

目标跟踪能否成功，很大程度上取决于目标外观表示能否应对视角变化、光照变化和遮挡等挑战 [213–215]。已有多项工作尝试将 CNN 用于视觉跟踪。Fan 等人 [216] 以 CNN 为基学习器，针对不同类别分别学习专用网络来跟踪物体。在文献 [216] 中，他们设计了具有移变性（shift-variant）的 CNN 跟踪器，这种结构是将 CNN 从检测器转为跟踪器的关键。网络特征通过离线训练学习。传统跟踪器往往只提取局部空间结构，该 CNN 跟踪方法则同时考虑相邻两帧图像，以提取空间和时间结构。时间信息中较强的信号往往出现在运动物体附近，因此时间结构能为跟踪提供粗略的速度线索。

Li 等人 [217] 提出了面向特定目标的 CNN 跟踪方法，跟踪过程中用在线获取的新样本增量训练网络。他们维护一个由多个 CNN 构成的候选池，用数据驱动的模型表示同一目标的不同外观实例。每个 CNN 各自保留一组卷积核，利用所有可用的底层线索把物体图块与周围背景区分开来。每个 CNN 初始化时只用一个实例训练其卷积核，此后逐帧在线更新。Li 等人 [217] 把使用较少滤波器的 CNN 放进具有时间适应机制的框架，避免用一个复杂的大型 CNN 容纳以往观测到的所有外观。给定一帧，先从池中选出最有希望的 CNN，评估不同的目标位置假设。得分最高的假设成为当前检测窗口，再使用热启动的反向传播优化结构化损失函数，对选中的模型重新训练。

文献 [218] 针对目标跟踪中手工特征和浅层分类器结构的局限，提出基于 CNN 的跟踪方法。首先由 CNN 自动学习判别性特征。为减轻模型更新造成的跟踪漂移，跟踪器同时利用初始帧中经标注的目标真实外观信息和在线获取的图像观测，再用启发式规则判断是否更新目标的外观模型。

Hong 等人 [219] 提出基于预训练 CNN 的视觉跟踪算法。该网络最初为大规模图像分类而训练，学到的表示被迁移用来描述跟踪目标。他们在 CNN 隐藏层之上添加在线 SVM，学习能把目标与背景区分开的外观模型。将目标相关信息反向投影至输入图像空间后，SVM 学到的模型可用于计算目标特异性的显著图。研究者再利用该显著图建立生成式目标外观模型，在跟踪时纳入目标的空间构型信息。

5.4 姿态估计

深层结构学习取得突破以来，许多研究开始利用 CNN 为人体姿态估计学习多个层次的表征和抽象特征 [220, 221]。DeepPose [222] 率先将 CNN 用于人体姿态估计，把它表述为用 CNN 回归人体关节坐标的问题。该方法将多个七层 CNN 级联起来，从整个人体的角度推断姿态。以往的工作通常要显式设计图模型和部位检测器；DeepPose 把整幅图像作为输入，使每个关节的估计都可以利用完整的上下文。

另一些研究利用 CNN 学习局部身体部位的表征。文献 [223] 提出一种全身姿态估计的端到端学习方法，联合训练 CNN 部位检测器和一个类似马尔可夫随机场（MRF）的空间模型，用卷积先验计算图中的成对势函数。Tompson 等 [224] 在系列研究中使用多分辨率 CNN，为每个身体部位计算热图。相较于文献 [223]，他们学习了身体部位的先验模型，也隐式学得了空间模型的结构。具体来说，先将每个部位与自身以及其他所有部位两两连接，用全连接图描述空间先验。在该工作的基础上，Tompson 等 [92] 在粗略估计姿态的 CNN 后加入位置细化模型。这个细化模型采用孪生网络 [64]，与已有模型 [224] 级联

并联合训练。与文献 [224] 类似，Chen 等 [225, 226] 也把图模型与 CNN 结合，用 CNN 学习身体部位出现的条件概率及部位间空间关系的条件概率，再将其用于图模型的一元项和成对项。这些条件概率也可以看作人体姿态的一种低维表征。双源 CNN [227] 则将图模型与整体式方法结合，把全身图像及局部 部位的整体视图作为输入，以兼顾局部信息与上下文。

研究者还将静态图像上的姿态估计扩展到视频。Jain 等 [228] 在文献 [224] 的工作上，将 RGB 特征与运动特征一同输入多分辨率 CNN，以进一步提升精度。网络以滑动窗口方式估计姿态。输入是由 RGB 图像和对应运动特征组成的三维张量，输出也是三维张量，包含各关节的响应图。响应图上每个位置的数值表示相应关节出现在该像素处的能量值。将输入下采样后送入网络，即可实现多分辨率处理。

5.5 文本检测与识别

从图像中识别文字已被广泛研究多年 [229–232]。传统研究主要关注光学字符识别（OCR），其输入通常来自受约束的视觉环境，例如背景干净、文字对齐。随着计算机视觉日益重视高层视觉理解 [233, 234]，研究重心逐渐转向自然场景图像中的文字识别。自然场景拍摄条件不受控制，文字外观差异很大，给传统 OCR 带来困难。CNN 等模型学得的特征表征更强、更丰富，有助于缓解这一问题。利用 CNN 改进场景文字识别的工作大致分为三类：只检测和定位文字，在裁剪好的文字图像上进行识别，以及把检测和识别整合起来的端到端文字检测与识别（text spotting）。

5.5.1 文本检测

文献 [235] 较早将 CNN 用于场景文字检测。模型在裁剪得到的文字图块和场景中的非文字图块上训练，学习区分两者。对输入图像构建多尺度金字塔后，利用 CNN 滤波器生成的响应图检测文字。为缩小文字检测的搜索空间，Xu 等 [236] 用最大稳定极值区域（MSER）提取候选字符，再用 CNN 分类器筛选。文献 [237] 同样结合 MSER 和 CNN：先用 CNN 区分像文字的 MSER 连通分量和非文字分量，再以滑动窗口方式应用 CNN，并进行非极大值抑制（NMS），分开挤在一起的文字分量。文献 [238] 研究了另一个问题：只判断输入图像是否包含文字，无须给出确切位置。他们先用 MSER 提取候选区域，再用 CNN 提取视觉特征，最后用词袋（BoW）框架汇聚这些 CNN 特征，构成整幅图像的全局表征。

5.5.2 文本识别

Goodfellow 等 [239] 在 CNN 的末层设置多个 softmax 分类器，每个分类器负责预测多位数字图像中某一顺序位置上的字符。为在不使用词表或词典的条件下识别文字，Jaderberg 等 [240] 提出类似条件随机场（CRF）的 CNN 模型，联合学习字符序列预测和二元字符组生成。随后的一些方法在传统 CNN 上加入循环神经网络（RNN）的变体，以描述文本中字符间的顺序依赖。文献 [241] 用滑动窗口提取字符级图块，先由 CNN 提取丰富的视觉特征，再用长短期记忆网络（LSTM）[242] 进行序列标注。文献 [243] 的方法与之相近，并可利用词表进一步提高识别性能。

5.5.3 端到端文本检测与识别

Wang 等 [15] 将原本为字符分类训练的 CNN 用于端到端系统中的文字检测。沿着相近思路，文献 [244] 使端到端系统的四项子任务共享特征：文字检测、区分大小写的字符分类、不区分大小写的字符分类，以及二元字符组分类。Jaderberg 等 [245] 更广泛地使用 CNN，将系统的三个主要环节——文字候选框筛选、文字边界框回归和文字识别——分别交给单独的 CNN 模型处理。

5.6 视觉显著性检测

在图像中定位重要区域的技术称为视觉显著性预测。这项研究颇具挑战性，能够支持大量计算机视觉和图像处理应用。已有研究尝试利用 CNN 对视觉模式的强大建模能力来预测显著性。

多个尺度或范围的上下文是预测视觉显著性的重要先验，多数相关工作都将它与 CNN 结合 [246–250]。Wang 等 [246] 先后利用局部和全局上下文：局部 CNN 根据图块为每个像素分配局部显著值，全局上下文即物体层面的信息，由深层全连接前馈网络处理。在文献 [247] 中，局部和全局上下文模型共享 CNN 参数，用来预测候选物体区域内部各个超像素的显著性。文献 [248] 在大规模图像分类数据集上预训练 CNN，让不同层次的上下文共用该模型提取特征，再将各层次的输出拼接起来，交给可训练的全连接前馈网络进行显著性预测。类似于 [247, 248]，文献 [249] 让三条 CNN 分支共享同一模型，但各自接收不同空间尺度的上下文。He 等 [250] 利用空间核与值域核得到两个有明确含义的序列，分别描述颜色的独特性与空间分布，将其作为一维 CNN 的输入。相较于直接输入原始像素，这类序列能够降低训练复杂度，同时编码超像素间的上下文信息。

另一些方法 [251–253] 主要依靠 CNN 的表征能力进行预测。文献 [251] 从大量随机实例化的 CNN 构成集成模型，为显著性检测提取特征；其中单个模型的深度最多只有三层。Deep Gaze [252] 使用一个经过预训练、具有五个卷积层的较深 CNN，再单独学习显著性模型，将所有卷积层的响应组合起来预测显著值。在作者比较的这些研究中，[253] 采用了从原始像素直接输出显著图的端到端 CNN。Pan 等 [253] 认为，该方法的成效来自其深度适中的结构，这种设计意在防止过拟合。

5.7 动作识别

动作识别根据人的外观和运动变化分析其行为并判别活动类别，是计算机视觉的难题之一 [254–256]。一般可以将其分为静态图像中的动作分析和视频中的动作分析，两者都已出现有效的 CNN 方法。下面分述这两类工作的进展。

5.7.1 静态图像中的动作识别

文献 [257] 表明，训练好的 CNN 最后几层的输出可以充当多种任务的通用视觉特征描述子。文献 [9, 258] 将这一思路用于动作识别，使用预训练 CNN 倒数第二层的输出，分别描述整幅动作图像和其中的人体边界框，并取得了较好的动作分类结果。Gkioxari 等 [259] 在这一框架中加入身体部位检测，其部位检测器是原始 Poselet 方法 [260] 的 CNN 扩展。

文献 [261] 用 CNN 表示上下文信息以识别动作。研究者在大量候选物体区域中寻找最具代表性的次要区域，以自底向上的方式把上下文特征加入对主要区域（人体真实边界框）的描述。他们用 CNN 描述主要区域和上下文区域，并微调两者的表征。随后他们进一步表明，即便不给出人体边界框，也可以定位和识别图像中的动作 [262]；不过此法仍须训练人体检测器，为测试时的识别提供指引。文献 [263] 提出一种方法，用尽量少的标注分割出表征人与物体交互的动作掩膜。

5.7.2 视频序列中的动作识别

传统 CNN 面向二维空间信号，而视频多出一个时间轴，时间变化与图像里的空间变化具有不同性质 [256, 264]。视频信号的数据规模也比静态图像大得多，因此将卷积网络应用于视频较为困难。Ji 等 [265] 提出将时间轴与空间轴作相似处理，对视频使用三维卷积层。Tran 等 [266] 研究了这种方法的性能、效率与有效性，并给出它相较其他方法的优势。

另一条路线保留二维卷积，融合相邻视频帧的特征图。文献 [267] 比较了后期融合、早期融合、慢融合三种策略，也与逐帧独立使用 CNN 的结果进行了比较。Simonyan 和 Zisserman [268] 进一步将空间变化与时间变化分别建模，为两者各训练一个 CNN。第一条分支在各视频帧上使用常规 CNN，第二条分支输入稠密光流，网络规模与结构均与空间分支相同。两条分支的输出在类别分数层面融合。Chéron 等 [269] 对定位到的人体部位应用双流 CNN，发现汇聚各部位的局部 CNN 描述子可以提高动作识别性能。

还可以把每帧的 CNN 特征输入序列学习模块（例如循环神经网络），用不同于空间变化的方式描述视频动态。Donahue 等 [270] 考察了在这一框架中使用 LSTM 单元学习序列的多种配置。

5.8 场景标注

场景标注为输入图像的每个像素指定一个语义类别，例如道路、水域或海面 [271–275]。CNN 可直接根据局部图块对像素类别的似然进行建模，学习能够区分细微局部差异的视觉特征和分类器。Farabet 等 [276] 较早将 CNN 用于场景标注，将不同尺度的图块送入多尺度卷积网络，得到的结果显著优于采用人工设计特征的系统。这一网络还成功用于 RGB-D 场景标注 [277]。为扩大像素的可见范围，Pinheiro 等 [278] 开发了循环式 CNN，将相同的 CNN 反复作用于上一次迭代输出的响应图，略微提高标注精度，同时显著缩短推理时间。Shuai 等 [279–281] 通过采样图块来训练参数化 CNN，大幅缩短训练时间。他们发现，基于图块的 CNN 容易出现局部歧义；文献 [279] 通过整合全局置信信息解决这一问题，文献 [280, 281] 则利用循环神经网络建模 CNN 图像特征间的上下文依赖，显著提升了标注性能。

与此同时，研究者开始将预训练深层 CNN 用于物体语义分割。Mostajabi 等 [282] 用卷积网络提取局部与近邻特征，再用 AlexNet [8] 提取远处与全局特征，拼接后得到由近及远的 zoom-out 特征，在语义分割任务上取得了很有竞争力的结果。Long 等 [28] 训练全卷积网络，直接从输入图像预测稠密标签图。其卷积层用在 ImageNet 分类数据集上预训练的模型初始化，并学习用转置卷积层提高标签图的分辨率。Chen 等 [283] 同样采用预训练的深层 CNN 预测像素标签。针对预测边界与真实物体轮廓对齐不够准确的问题，他们进一步引入全连接条件随机场，改善标注性能。

5.9 语音处理

5.9.1 自动语音识别

自动语音识别（ASR）将人的语音转换为相应词语 [284]。在 CNN 被用于该领域以前，隐马尔可夫模型与高斯混合模型组合的系统（GMM-HMM）长期占据主导地位 [285]。这类方法通常需要从语音中提取人工设计的特征，最常用的包括梅尔频率倒谱系数（MFCC）。一些研究还将深度神经网络（DNN）用于大词汇量连续语音识别（LVCSR），取得了令人鼓舞的结果 [286, 287]；但当训练和使用条件不匹配时，例如录音条件不同，网络性能容易下降 [288]。

CNN 可通过局部连接利用时间和频率两个方向上的相关性，还能捕捉语音信号中的频移。已有工作显示，它在此类任务中优于 GMM-HMM 和常规 DNN [289, 290]。文献 [289] 在梅尔滤波器组特征上应用 CNN，降低了语音识别错误率。一些方法以原始音频波形为输入，将波形滤波器与网络其余部分联合训练 [291, 292]。早期语音 CNN 大多只用少数几个卷积层，例如 Abdel-Hamid 等 [290] 仅用一层，而 Amodei 等 [293] 用三个卷积层作特征预处理。此后，很深的 CNN 也在语音识别中表现出色 [294, 295]。混合式 NN-HMM 语音识别系统的声学建模成功使用了小滤波器；文献 [296] 还用稠密连接层替代池化操作。Yu 等 [297] 为语音识别提出带注意力机制的逐层上下文扩展模型。它是时延神经网络 [298] 的一种变体，浅层提取简单的局部模式，深层利用更广的上下文提取更复杂的模式。文献 [40] 也采用了相似思路。

5.9.2 统计参数语音合成

CNN 的影响也扩展到了统计参数语音合成（SPSS）。语音合成的目标是直接从文本以及可能附带的其他信息中生成语音。人们早已发现，与自然语音相比，浅层 HMM 网络合成的声音往往较为模糊。许多工作采用深度学习改善这一缺陷 [299–301]。这类方法的一个优势，在于通过生成式建模描述数据内部的相关性。受利用神经自回归生成模型刻画图像 [302]、文本 [303] 等复杂分布的工作启发，WaveNet [39] 采用基于 CNN 的生成模型描述给定语言特征时声学特征的条件分布，是 SPSS 的一个里程碑。为处理长程时间依赖，它采用基于空洞因果卷积的新结构，获得很大的感受野。借助从文本得到的语言特征作为条件，模型可以直接从文本合成语音。

5.10 自然语言处理

5.10.1 统计语言建模

统计语言模型的输入通常是不完整的词序列 [304, 305]。Kim 等 [304] 在每个时间步把字符级 CNN 的输出送入 LSTM。用于序列预测的 genCNN [306] 采用独立的门控网络替代最大池化。Kalchbrenner 等 [38] 提出的 ByteNet 将两个空洞卷积网络叠放起来进行序列处理。和 WaveNet [39] 一样，ByteNet 通过空洞卷积扩大感受野，因此可以对具有长程依赖的序列建模，计算时间也只随序列长度线性增长。与循环神经网络相比，CNN 在获取长距离信息的同时，还能为输入的词构建层次化表征。Gu 等 [307] 与文献 [308] 的思路相近，均用不带池化的 CNN 对输入词建模。Gu 等 [307] 将语言 CNN 与循环高速网络结合，较基于 LSTM 的方法有显著改进。受 LSTM 门控机制启发，文献 [308] 用门控卷积网络控制信息在网络中的传播路径，在 WikiText-103 上达到了当时的最佳水平。作者认为，[308] 和 [307] 的框架仍受其循环框架和有限输入窗口制约。如何同时捕获特定的长程依赖和历史词的层次表征，仍是一个开放问题。

5.10.2 文本分类

文本分类是自然语言处理（NLP）的一项重要任务。自然语言句子既有时序结构，又有层次结构，理解这两种结构对理解句子十分重要。CNN 擅长捕捉时序或层次结构中的局部关系，在句子建模中取得了优良成绩。文本分类需要选择合适的 CNN 结构。Collobert 等 [309] 和 Yu 等 [310] 用一个卷积层描述句子，Kalchbrenner 等 [311] 则叠加多个卷积层。文献 [312] 将多通道卷积和不同大小的卷积核用于句子分类。多层卷积有助于提取高层抽象特征，多个线性滤波器则能有效考虑不同的 n 元组特征。

Yin 等 [313] 在 [312] 的基础上引入层次化卷积结构，进一步探索多通道和不同尺度的特征检测器。池化可帮助网络适应不同句长。在文献 [312, 314] 中，最大池化保留最重要的信息来表示句子，但它无法区分某一相关特征出现一次还是多次，也会忽略特征出现的先后顺序。文献 [311] 提出 k-max 池化：选取最大的 k 个激活值，并按它们在输入序列中的原始顺序返回。动态 k-max 池化将其推广，使 k 值取决于输入特征图的尺寸。与在计算机视觉中取得成功的深层 CNN 相比，上述网络较浅。Conneau 等 [315] 实现了多达 29 个卷积层的深度结构。进一步实验发现，在很深的网络（49 层）中加入捷径连接能得到较好的结果，不过这种设置未达到当时的最佳成绩。

6 结论与展望

深层 CNN 在图像、视频、语音和文本处理上均已取得突破。本文综述了 CNN 的近期进展，从层设计、激活函数、损失函数、正则化、优化和快速计算等方面讨论其改进，并介绍了图像分类、目标检测、目标跟踪、姿态估计、文字检测、视觉显著性检测、动作识别、场景标注、语音以及自然语言处理等方面的应用。

尽管 CNN 在实验评估中取得了很大成功，仍有许多问题值得研究。首先，网络越来越深，训练需要大规模数据集和强大的计算能力。人工收集有标注的数据耗费大量人力，因此值得探索 CNN 的无监督学习。在训练速度方面，已有异步随机梯度下降算法利用 CPU 和 GPU 集群取得了令人鼓舞的结果，开发高效、可扩展的并行训练算法仍然值得投入。在测试时，深度模型需要大量存储和计算时间，难以部署在资源有限的移动平台。怎样在保持精度的同时降低复杂度、加快模型执行，是重要的研究方向。

此外，把 CNN 应用于新任务的一大障碍是，学习率、卷积核尺寸、层数等超参数的选择需要相当多的技巧和经验。这些超参数相互依赖，进一步提高了调参成本。已有研究显示，用于训练深层 CNN 的现行优化技术仍有很大改进空间 [12, 42, 316]。

最后，CNN 的理论基础仍不充分。当时的 CNN 已经在多种应用中表现出色，但它究竟为何有效、怎样起作用，人们仍缺乏根本性的认识。值得继续研究 CNN 的基本原理，也值得探索如何借鉴自然视觉感知机制来改进网络设计。我们希望本文既能增进读者对 CNN 的理解，也能推动这一领域的后续研究与应用开发。

致谢

本研究在新加坡南洋理工大学的快速大规模目标检索（Rapid-Rich Object Search，ROSE）实验室完成。该实验室获得新加坡资讯通信媒体发展局的支持。

参考文献

以下条目沿用原文的英文题名、著录顺序与编号，便于按正文引文检索。

[1] D. H. Hubel, T. N. Wiesel, Receptive fields and functional architecture of monkey striate cortex, The Journal of physiology (1968) 215–243.

[2] K. Fukushima, S. Miyake, Neocognitron: A self-organizing neural network model for a mechanism of visual pattern recognition, in: Competition and cooperation in neural nets, 1982, pp. 267–285.

[3] B. B. Le Cun, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, L. D. Jackel, Handwritten digit recognition with a back-propagation network, in: Proceedings of the Advances in Neural Information Processing Systems (NIPS), 1989, pp. 396–404.

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

[5] R. Hecht-Nielsen, Theory of the backpropagation neural network, Neural Networks 1 (Supplement-1) (1988) 445–448.

[6] W. Zhang, K. Itoh, J. Tanida, Y. Ichioka, Parallel distributed processing model with local space-invariant interconnections and its optical architecture, Applied optics 29 (32) (1990) 4790–4797.

[7] X.-X. Niu, C. Y. Suen, A novel hybrid cnn–svm classifier for recognizing handwritten digits, Pattern Recognition 45 (4) (2012) 1318–1325.

[8] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al., Imagenet large scale visual recognition challenge, International Journal of Conflict and Violence (IJCV) 115 (3) (2015) 211–252.

[9] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, in: Proceedings of the International Conference on Learning Representations (ICLR), 2015.

[10] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 1–9.

[11] M. D. Zeiler, R. Fergus, Visualizing and understanding convolutional networks, in: Proceedings of the European Conference on Computer Vision (ECCV), 2014, pp. 818–833.

[12] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770–778.

[13] Y. A. LeCun, L. Bottou, G. B. Orr, K.-R. Müller, Efficient backprop, in: Neural Networks: Tricks of the Trade - Second Edition, 2012, pp. 9–48.

[14] V. Nair, G. E. Hinton, Rectified linear units improve restricted boltzmann machines, in: Proceedings of the International Conference on Machine Learning (ICML), 2010, pp. 807–814.

[15] T. Wang, D. J. Wu, A. Coates, A. Y. Ng, End-to-end text recognition with convolutional neural networks, in: Proceedings of the International Conference on Pattern Recognition (ICPR), 2012, pp. 3304–3308.

[16] Y. Boureau, J. Ponce, Y. LeCun, A theoretical analysis of feature pooling in visual recognition, in: Proceedings of the International Conference on Machine Learning (ICML), 2010, pp. 111–118.

[17] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, R. R. Salakhutdinov, Improving neural networks by preventing co-adaptation of feature detectors, CoRR abs/1207.0580.

[18] M. Lin, Q. Chen, S. Yan, Network in network, in: Proceedings of the International Conference on Learning Representations (ICLR), 2014.

[19] Y. Tang, Deep learning using linear support vector machines, in: Proceedings of the International Conference on Machine Learning (ICML) Workshops, 2013.

[20] G. Madjarov, D. Kocev, D. Gjorgjevikj, S. Džeroski, An extensive experimental comparison of methods for multi-label learning, Pattern Recognition 45 (9) (2012) 3084–3104.

[21] R. G. J. Wijnhoven, P. H. N. de With, Fast training of object detection using stochastic gradient descent, in: International Conference on Pattern Recognition (ICPR), 2010, pp. 424–427.

[22] M. Zinkevich, M. Weimer, L. Li, A. J. Smola, Parallelized stochastic gradient descent, in: Proceedings of the Advances in Neural Information Processing Systems (NIPS), 2010, pp. 2595–2603.

[23] J. Ngiam, Z. Chen, D. Chia, P. W. Koh, Q. V. Le, A. Y. Ng, Tiled convolutional neural networks, in: Proceedings of the Advances in Neural Information Processing Systems (NIPS), 2010, pp. 1279–1287.

[24] Z. Wang, T. Oates, Encoding time series as images for visual inspection and classification using tiled convolutional neural networks, in: Proceedings of the Association for the Advancement of Artificial Intelligence (AAAI) Workshops, 2015.

[25] Y. Zheng, Q. Liu, E. Chen, Y. Ge, J. L. Zhao, Time series classification using multi-channels deep convolutional neural networks, in: Proceedings of the International Conference on Web-Age Information Management (WAIM), 2014, pp. 298–310.

[26] M. D. Zeiler, D. Krishnan, G. W. Taylor, R. Fergus, Deconvolutional networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2010, pp. 2528–2535.

[27] M. D. Zeiler, G. W. Taylor, R. Fergus, Adaptive deconvolutional networks for mid and high level feature learning, in: Proceedings of the International Conference on Computer Vision (ICCV), 2011, pp. 2018–2025.

[28] J. Long, E. Shelhamer, T. Darrell, Fully convolutional networks for semantic segmentation, IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) 39 (4) (2017) 640–651.

[29] F. Visin, K. Kastner, A. Courville, Y. Bengio, M. Matteucci, K. Cho, Reseg: A recurrent neural network for object segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, 2015.

[30] H. Noh, S. Hong, B. Han, Learning deconvolution network for semantic segmentation, in: Proceedings of the International Conference on Computer Vision (ICCV), 2015, pp. 1520–1528.

[31] C. Cao, X. Liu, Y. Yang, Y. Yu, J. Wang, Z. Wang, Y. Huang, L. Wang, C. Huang, W. Xu, et al., Look and think twice: Capturing top-down visual attention with feedback convolutional neural networks, in: Proceedings of the International Conference on Computer Vision (ICCV), 2015, pp. 2956–2964.

[32] J. Zhang, Z. Lin, J. Brandt, X. Shen, S. Sclaroff, Top-down neural attention by excitation backprop, in: Proceedings of the European Conference on Computer Vision (ECCV), 2016, pp. 543–559.

[33] Y. Zhang, E. K. Lee, E. H. Lee, U. EDU, Augmenting supervised neural networks with unsupervised objectives for large-scale image classification, in: Proceedings of the International Conference on Machine Learning (ICML), 2016, pp. 612–621.

[34] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, A. Torralba, Learning deep features for discriminative localization, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 2921–2929.

[35] A. Das, H. Agrawal, C. L. Zitnick, D. Parikh, D. Batra, Human attention in visual question answering: Do humans and deep networks look at the same regions?, in: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), 2016, pp. 932–937.

[36] C. Dong, C. C. Loy, K. He, X. Tang, Image super-resolution using deep convolutional networks, IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) 38 (2) (2016) 295–307.

[37] F. Yu, V. Koltun, Multi-scale context aggregation by dilated convolutions, in: Proceedings of the International Conference on Learning Representations (ICLR), 2016.

[38] N. Kalchbrenner, L. Espeholt, K. Simonyan, A. v. d. Oord, A. Graves, K. Kavukcuoglu, Neural machine translation in linear time, CoRR abs/1610.10099.

[39] A. v. d. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, K. Kavukcuoglu, Wavenet: A generative model for raw audio, CoRR abs/1609.03499.

[40] T. Sercu, V. Goel, Dense prediction on sequences with time-dilated convolutions for speech recognition, in: Proceedings of the Advances in Neural Information Processing Systems (NIPS) Workshops, 2016.

[41] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna, Rethinking the inception architecture for computer vision, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 2818–2826.

[42] C. Szegedy, S. Ioffe, V. Vanhoucke, Inception-v4, inception-resnet and the impact of residual connections on learning, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2017, pp. 4278–4284.

[43] A. Hyvärinen, U. Köster, Complex cell pooling and the statistics of natural images, Network: Computation in Neural Systems 18 (2) (2007) 81–100.

[44] J. B. Estrach, A. Szlam, Y. Lecun, Signal recovery from pooling representations, in: Proceedings of the International Conference on Machine Learning (ICML), 2014, pp. 307–315.

[45] L. Wan, M. Zeiler, S. Zhang, Y. L. Cun, R. Fergus, Regularization of neural networks using dropconnect, in: Proceedings of the International Conference on Machine Learning (ICML), 2013, pp. 1058–1066.

[46] D. Yu, H. Wang, P. Chen, Z. Wei, Mixed pooling for convolutional neural networks, in: Proceedings of the Rough Sets and Knowledge Technology (RSKT), 2014, pp. 364–375.

[47] M. D. Zeiler, R. Fergus, Stochastic pooling for regularization of deep convolutional neural networks, in: Proceedings of the International Conference on Learning Representations (ICLR), 2013.

[48] O. Rippel, J. Snoek, R. P. Adams, Spectral representations for convolutional neural networks, in: Proceedings of the Advances in Neural Information Processing Systems (NIPS), 2015, pp. 2449–2457.

[49] M. Mathieu, M. Henaff, Y. LeCun, Fast training of convolutional networks through ffts, in: Proceedings of the International Conference on Learning Representations (ICLR), 2014.

[50] K. He, X. Zhang, S. Ren, J. Sun, Spatial pyramid pooling in deep convolutional networks for visual recognition, IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) 37 (9) (2015) 1904–1916.

[51] S. Singh, A. Gupta, A. Efros, Unsupervised discovery of mid-level discriminative patches, in: Proceedings of the European Conference on Computer Vision (ECCV), 2012, pp. 73–86.

[52] Y. Gong, L. Wang, R. Guo, S. Lazebnik, Multi-scale orderless pooling of deep convolutional activation features, in: Proceedings of the European Conference on Computer Vision (ECCV), 2014, pp. 392–407.

[53] H. Jégou, F. Perronnin, M. Douze, J. Sanchez, P. Perez, C. Schmid, Aggregating local image descriptors into compact codes, IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) 34 (9) (2012) 1704–1716.

[54] A. L. Maas, A. Y. Hannun, A. Y. Ng, Rectifier nonlinearities improve neural network acoustic models, in: Proceedings of the International Conference on Machine Learning (ICML), Vol. 30, 2013.

[55] M. D. Zeiler, M. Ranzato, R. Monga, M. Mao, K. Yang, Q. V. Le, P. Nguyen, A. Senior, V. Vanhoucke, J. Dean, et al., On rectified linear units for speech processing, in: Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2013, pp. 3517–3521.

[56] K. He, X. Zhang, S. Ren, J. Sun, Delving deep into rectifiers: Surpassing human-level performance on imagenet classification, in: Proceedings of the International Conference on Computer Vision (ICCV), 2015, pp. 1026–1034.

[57] B. Xu, N. Wang, T. Chen, M. Li, Empirical evaluation of rectified activations in convolutional network, in: Proceedings of the International Conference on Machine Learning (ICML) Workshop, 2015.

[58] D.-A. Clevert, T. Unterthiner, S. Hochreiter, Fast and accurate deep network learning by exponential linear units (elus), in: Proceedings of the International Conference on Learning Representations (ICLR), 2016.

[59] I. J. Goodfellow, D. Warde-Farley, M. Mirza, A. Courville, Y. Bengio, Maxout networks, in: Proceedings of the International Conference on Machine Learning (ICML), 2013, pp. 1319–1327.

[60] J. T. Springenberg, M. Riedmiller, Improving deep neural networks with probabilistic maxout units, CoRR abs/1312.6116.

[61] T. Zhang, Solving large scale linear prediction problems using stochastic gradient descent algorithms, in: Proceedings of the International Conference on Machine Learning (ICML), 2004.

[62] L. Deng, The mnist database of handwritten digit images for machine learning research, IEEE Signal Processing Magazine 29 (6) (2012) 141–142.

[63] W. Liu, Y. Wen, Z. Yu, M. Yang, Large-margin softmax loss for convolutional neural networks, in: Proceedings of the International Conference on Machine Learning (ICML), 2016, pp. 507–516.

[64] J. Bromley, J. W. Bentz, L. Bottou, I. Guyon, Y. LeCun, C. Moore, E. Säckinger, R. Shah, Signature verification using a siamese time delay neural network, International Journal of Pattern Recognition and Artificial Intelligence (IJPRAI) 7 (4) (1993) 669–688.

[65] S. Chopra, R. Hadsell, Y. LeCun, Learning a similarity metric discriminatively, with application to face verification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2005, pp. 539–546.

[66] R. Hadsell, S. Chopra, Y. LeCun, Dimensionality reduction by learning an invariant mapping, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2006, pp. 1735–1742.

[67] U. Shaham, R. R. Lederman, Learning by coincidence: Siamese networks and common variable learning, Pattern Recognition.

[68] J. Lin, O. Morere, V. Chandrasekhar, A. Veillard, H. Goh, Deephash: Getting regularization, depth and fine-tuning right, in: Proceedings of the International Conference on Multimedia Retrieval (ICMR), 2017, pp. 133–141.

[69] F. Schroff, D. Kalenichenko, J. Philbin, Facenet: A unified embedding for face recognition and clustering, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 815–823.

[70] H. Liu, Y. Tian, Y. Yang, L. Pang, T. Huang, Deep relative distance learning: Tell the difference between similar vehicles, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 2167–2175.

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

[72] Z. Liu, P. Luo, S. Qiu, X. Wang, X. Tang, Deepfashion: Powering robust clothes recognition and retrieval with rich annotations, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 1096–1104.

[73] D. P. Kingma, M. Welling, Auto-encoding variational bayes, in: Proceedings of the International Conference on Learning Representations (ICLR), 2014.

[74] D. J. Im, S. Ahn, R. Memisevic, Y. Bengio, Denoising criterion for variational auto-encoding framework, in: Proceedings of the Association for the Advancement of Artificial Intelligence (AAAI), 2017, pp. 2059–2065.

[75] D. P. Kingma, S. Mohamed, D. J. Rezende, M. Welling, Semi-supervised learning with deep generative models, in: Proceedings of the Advances in Neural Information Processing Systems (NIPS), 2014, pp. 3581–3589.

[76] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets, in: Proceedings of the Advances in Neural Information Processing Systems (NIPS), 2014, pp. 2672– 2680.

[77] M. Mirza, S. Osindero, Conditional generative adversarial nets, CoRR abs/1411.1784.

[78] P. Vincent, H. Larochelle, Y. Bengio, P.-A. Manzagol, Extracting and composing robust features with denoising autoencoders, in: Proceedings of the International Conference on Machine Learning (ICML), 2008, pp. 1096–1103.

[79] W. W. Ng, G. Zeng, J. Zhang, D. S. Yeung, W. Pedrycz, Dual autoencoders features for imbalance classification problem, Pattern Recognition 60 (2016) 875–889.

[80] J. Mehta, A. Majumdar, Rodeo: robust de-aliasing autoencoder for real-time medical image reconstruction, Pattern Recognition 63 (2017) 499–510.

[81] B. A. Olshausen, et al., Emergence of simple-cell receptive field properties by learning a sparse code for natural images, Nature 381 (6583) (1996) 607.

[82] H. Lee, A. Battle, R. Raina, A. Y. Ng, Efficient sparse coding algorithms, in: Proceedings of the Advances in Neural Information Processing Systems (NIPS), 2006, pp. 801–808.

[83] S. Eslami, N. Heess, T. Weber, Y. Tassa, K. Kavukcuoglu, G. E. Hinton, Attend, infer, repeat: Fast scene understanding with generative models, in: Proceedings of the Advances in Neural Information Processing Systems (NIPS), 2016, pp. 3225–3233.

[84] K. Sohn, H. Lee, X. Yan, Learning structured output representation using deep conditional generative models, in: Proceedings of the Advances in Neural Information Processing Systems (NIPS), 2015, pp. 3483–3491.

[85] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, H. Lee, Generative adversarial text to image synthesis, in: Proceedings of the International Conference on Machine Learning (ICML), 2016, pp. 1060–1069.

[86] E. L. Denton, S. Chintala, R. Fergus, et al., Deep generative image models using a laplacian pyramid of adversarial networks, in: Proceedings of the Advances in Neural Information Processing Systems (NIPS), 2015, pp. 1486–1494.

[87] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, X. Chen, Improved techniques for training gans, in: Proceedings of the Advances in Neural Information Processing Systems (NIPS), 2016, pp. 2226–2234.

[88] A. Dosovitskiy, T. Brox, Generating images with perceptual similarity metrics based on deep networks, in: Proceedings of the Advances in Neural Information Processing Systems (NIPS), 2016, pp. 658–666.

[89] A. N. Tikhonov, On the stability of inverse problems, in: Dokl. Akad. Nauk SSSR, Vol. 39, 1943, pp. 195–198.

[90] S. Wang, C. Manning, Fast dropout training, in: Proceedings of the International Conference on Machine Learning (ICML), 2013, pp. 118–126.

[91] J. Ba, B. Frey, Adaptive dropout for training deep neural networks, in: Proceedings of the Advances in Neural Information Processing Systems (NIPS), 2013, pp. 3084–3092.

[92] J. Tompson, R. Goroshin, A. Jain, Y. LeCun, C. Bregler, Efficient object localization using convolutional networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 648–656.

[93] H. Yang, I. Patras, Mirror, mirror on the wall, tell me, is the error small?, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 4685–4693.

[94] S. Xie, Z. Tu, Holistically-nested edge detection, in: Proceedings of the International Conference on Computer Vision (ICCV), 2015, pp. 1395–1403.

[95] J. Salamon, J. P. Bello, Deep convolutional neural networks and data augmentation for environmental sound classification, Signal Processing Letters (SPL) 24 (3) (2017) 279–283.

[96] D. Eigen, R. Fergus, Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture, in: Proceedings of the International Conference on Computer Vision (ICCV), 2015, pp. 2650–2658.

[97] M. Paulin, J. Revaud, Z. Harchaoui, F. Perronnin, C. Schmid, Transformation pursuit for image classification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 3646–3653.

[98] S. Hauberg, O. Freifeld, A. B. L. Larsen, J. W. Fisher III, L. K. Hansen, Dreaming more data: Class-dependent distributions over diffeomorphisms for learned data augmentation, in: Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS), 2016, pp. 342–350.

[99] S. Xie, T. Yang, X. Wang, Y. Lin, Hyper-class augmented and regularized deep learning for fine-grained image classification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 2645–2654.

[100] Z. Xu, S. Huang, Y. Zhang, D. Tao, Augmenting strong supervision using web data for fine-grained categorization, in: Proceedings of the International Conference on Computer Vision (ICCV), 2015, pp. 2524–2532.

[101] A. Choromanska, M. Henaff, M. Mathieu, G. B. Arous, Y. LeCun, The loss surfaces of multilayer networks, in: Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS), 2015.

[102] D. Mishkin, J. Matas, All you need is a good init, in: Proceedings of the International Conference on Learning Representations (ICLR), 2016.

[103] I. Sutskever, J. Martens, G. Dahl, G. Hinton, On the importance of initialization and momentum in deep learning, in: Proceedings of the International Conference on Machine Learning (ICML), 2013, pp. 1139–1147.

[104] X. Glorot, Y. Bengio, Understanding the difficulty of training deep feedforward neural networks, in: Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS), 2010, pp. 249–256.

[105] A. M. Saxe, J. L. McClelland, S. Ganguli, Exact solutions to the nonlinear dynamics of learning in deep linear neural networks, in: Proceedings of the International Conference on Learning Representations (ICLR), 2014.

[106] C. Doersch, A. Gupta, A. A. Efros, Unsupervised visual representation learning by context prediction, in: Proceedings of the International Conference on Computer Vision (ICCV), 2015, pp. 1422–1430.

[107] P. Agrawal, J. Carreira, J. Malik, Learning to see by moving, in: Proceedings of the International Conference on Computer Vision (ICCV), 2015, pp. 37–45.

[108] N. Qian, On the momentum term in gradient descent learning algorithms, Neural Networks 12 (1) (1999) 145–151.

[109] D. Kingma, J. Ba, Adam: A method for stochastic optimization, in: Proceedings of the International Conference on Learning Representations (ICLR), 2015.

[110] I. Loshchilov, F. Hutter, Sgdr: Stochastic gradient descent with warm restarts, in: Proceedings of the International Conference on Learning Representations (ICLR), 2017.

[111] T. Schaul, S. Zhang, Y. LeCun, No more pesky learning rates, in: Proceedings of the International Conference on Machine Learning (ICML), 2013, pp. 343–351.

[112] S. Zhang, A. E. Choromanska, Y. LeCun, Deep learning with elastic averaging sgd, in: Proceedings of the Advances in Neural Information Processing Systems (NIPS), 2015, pp. 685–693.

[113] B. Recht, C. Re, S. Wright, F. Niu, Hogwild: A lock-free approach to parallelizing stochastic gradient descent, in: Proceedings of the Advances in Neural Information Processing Systems (NIPS), 2011, pp. 693–701.

[114] J. Dean, G. Corrado, R. Monga, K. Chen, M. Devin, M. Mao, A. Senior, P. Tucker, K. Yang, Q. V. Le, et al., Large scale distributed deep networks, in: Proceedings of the Advances in Neural Information Processing Systems (NIPS), 2012, pp. 1232–1240.

[115] T. Paine, H. Jin, J. Yang, Z. Lin, T. Huang, Gpu asynchronous stochastic gradient descent to speed up neural network training, CoRR abs/1107.2490.

[116] Y. Zhuang, W.-S. Chin, Y.-C. Juan, C.-J. Lin, A fast parallel sgd for matrix factorization in shared memory systems, in: Proceedings of the ACM conference on Recommender systems RecSys, 2013, pp. 249–256.

[117] Y. Yao, L. Rosasco, A. Caponnetto, On early stopping in gradient descent learning, Constructive Approximation 26 (2) (2007) 289–315.

[118] L. Prechelt, Early stopping - but when?, in: Neural Networks: Tricks of the Trade - Second Edition, 2012, pp. 53–67.

[119] C. Zhang, S. Bengio, M. Hardt, B. Recht, O. Vinyals, Understanding deep learning requires rethinking generalization, in: Proceedings of the International Conference on Learning Representations (ICLR), 2017.

[120] S. Ioffe, C. Szegedy, Batch normalization: Accelerating deep network training by reducing internal covariate shift, Journal of Machine Learning Research (JMLR) (2015) 448–456.

[121] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Computation 9 (8) (1997) 1735–1780.

[122] R. K. Srivastava, K. Greff, J. Schmidhuber, Training very deep networks, in: Proceedings of the Advances in Neural Information Processing Systems (NIPS), 2015, pp. 2377–2385.

[123] K. He, X. Zhang, S. Ren, J. Sun, Identity mappings in deep residual networks, in: Proceedings of the European Conference on Computer Vision (ECCV), 2016, pp. 630–645.

[124] F. Shen, R. Gan, G. Zeng, Weighted residuals for very deep networks, in: Proceedings of the International Conference on Systems and Informatics (ICSAI), 2016, pp. 936–941.

[125] S. Zagoruyko, N. Komodakis, Wide residual networks, in: Proceedings of the British Machine Vision Conference (BMVC), 2016, pp. 87.1–87.12.

[126] S. Singh, D. Hoiem, D. Forsyth, Swapout: Learning an ensemble of deep architectures, in: Proceedings of the Advances in Neural Information Processing Systems (NIPS), 2016, pp. 28–36.

[127] S. Targ, D. Almeida, K. Lyman, Resnet in resnet: Generalizing residual architectures, CoRR abs/1603.08029.

[128] K. Zhang, M. Sun, T. X. Han, X. Yuan, L. Guo, T. Liu, Residual networks of residual networks: Multilevel residual networks, IEEE Transactions on Circuits and Systems for Video Technology (TCSVT) PP (99) (2016) 1–1.

[129] G. Huang, Z. Liu, K. Q. Weinberger, Densely connected convolutional networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 4700–4708.

[130] S. Chetlur, C. Woolley, P. Vandermersch, J. Cohen, J. Tran, B. Catanzaro, E. Shelhamer, cudnn: Efficient primitives for deep learning abs/1410.0759.

[131] N. Vasilache, J. Johnson, M. Mathieu, S. Chintala, S. Piantino, Y. LeCun, Fast convolutional nets with fbfft: A gpu performance evaluation, in: Proceedings of the International Conference on Learning Representations (ICLR), 2015.

[132] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, Y. LeCun, Overfeat: Integrated recognition, localization and detection using convolutional networks.

[133] A. Lavin, Fast algorithms for convolutional neural networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 4013–4021.

[134] T. N. Sainath, B. Kingsbury, V. Sindhwani, E. Arisoy, B. Ramabhadran, Low-rank matrix factorization for deep neural network training with high-dimensional output targets, in: Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2013, pp. 6655–6659.

[135] J. Xue, J. Li, Y. Gong, Restructuring of deep neural network acoustic models with singular value decomposition, in: Proceedings of the International Speech Communication Association (INTERSPEECH), 2013, pp. 2365–2369.

[136] M. Denil, B. Shakibi, L. Dinh, N. de Freitas, et al., Predicting parameters in deep learning, in: Proceedings of the Advances in Neural Information Processing Systems (NIPS), 2013, pp. 2148–2156.

[137] E. L. Denton, W. Zaremba, J. Bruna, Y. LeCun, R. Fergus, Exploiting linear structure within convolutional networks for efficient evaluation, in: Proceedings of the Advances in Neural Information Processing Systems (NIPS), 2014, pp. 1269–1277.

[138] M. Jaderberg, A. Vedaldi, A. Zisserman, Speeding up convolutional neural networks with low rank expansions, in: Proceedings of the British Machine Vision Conference (BMVC), 2014.

[139] A. Novikov, D. Podoprikhin, A. Osokin, D. P. Vetrov, Tensorizing neural networks, in: Proceedings of the Advances in Neural Information Processing Systems (NIPS), 2015, pp. 442–450.

[140] I. V. Oseledets, Tensor-train decomposition, SIAM J. Scientific Computing 33 (5) (2011) 2295–2317.

[141] Q. Le, T. Sarlós, A. Smola, Fastfood-approximating kernel expansions in loglinear time, in: Proceedings of the International Conference on Machine Learning (ICML), Vol. 85, 2013.

[142] A. Dasgupta, R. Kumar, T. Sarlós, Fast locality-sensitive hashing, in: Proceedings of the International Conference on Knowledge Discovery and Data Mining (SIGKDD), 2011, pp. 1073–1081.

[143] F. X. Yu, S. Kumar, Y. Gong, S.-F. Chang, Circulant binary embedding, in: Proceedings of the International Conference on Machine Learning (ICML), 2014, pp. 946–954.

[144] Y. Cheng, F. X. Yu, R. S. Feris, S. Kumar, A. Choudhary, S.-F. Chang, An exploration of parameter redundancy in deep networks with circulant projections, in: Proceedings of the International Conference on Computer Vision (ICCV), 2015, pp. 2857–2865.

[145] M. Moczulski, M. Denil, J. Appleyard, N. de Freitas, Acdc: A structured efficient linear layer, in: Proceedings of the International Conference on Learning Representations (ICLR), 2016.

[146] S. Han, H. Mao, W. J. Dally, Deep compression: Compressing deep neural network with pruning, trained quantization and huffman coding, in: Proceedings of the International Conference on Learning Representations (ICLR), 2016.

[147] M. Kim, P. Smaragdis, Bitwise neural networks, in: Proceedings of the International Conference on Machine Learning (ICML) Workshops, 2016.

[148] M. Rastegari, V. Ordonez, J. Redmon, A. Farhadi, Xnor-net: Imagenet classification using binary convolutional neural networks, in: Proceedings of the European Conference on Computer Vision (ECCV), 2016, pp. 525–542.

[149] S. Zhou, Y. Wu, Z. Ni, X. Zhou, H. Wen, Y. Zou, Dorefa-net: Training low bitwidth convolutional neural networks with low bitwidth gradients, CoRR abs/1606.06160.

[150] M. Courbariaux, Y. Bengio, Binarynet: Training deep neural networks with weights and activations constrained to+ 1 or-1, in: Proceedings of the Advances in Neural Information Processing Systems (NIPS),

2016.

[151] G. J. Sullivan, Efficient scalar quantization of exponential and laplacian random variables, IEEE Trans. Information Theory 42 (5) (1996) 1365–1374.

[152] Y. Gong, L. Liu, M. Yang, L. Bourdev, Compressing deep convolutional networks using vector quantization, in: arXiv preprint arXiv:1412.6115, Vol. abs/1412.6115, 2014.

[153] Y. Chen, T. Guan, C. Wang, Approximate nearest neighbor search by residual vector quantization, Sensors 10 (12) (2010) 11259–11273.

[154] W. Zhou, Y. Lu, H. Li, Q. Tian, Scalar quantization for large scale image search, in: Proceedings of the 20th ACM international conference on Multimedia, 2012, pp. 169–178.

[155] L. Y. Pratt, Comparing biases for minimal network construction with back-propagation, in: Proceedings of the Advances in Neural Information Processing Systems (NIPS), 1988, pp. 177–185.

[156] S. Han, J. Pool, J. Tran, W. Dally, Learning both weights and connections for efficient neural network, in: Proceedings of the Advances in Neural Information Processing Systems (NIPS), 2015, pp. 1135–1143.

[157] Y. Guo, A. Yao, Y. Chen, Dynamic network surgery for efficient dnns, in: Proceedings of the Advances in Neural Information Processing Systems (NIPS), 2016, pp. 1379–1387.

[158] T.-J. Yang, Y.-H. Chen, V. Sze, Designing energy-efficient convolutional neural networks using energy-aware pruning, CoRR abs/1611.05128.

[159] H. Hu, R. Peng, Y.-W. Tai, C.-K. Tang, Network trimming: A data-driven neuron pruning approach towards efficient deep architectures, Vol. abs/1607.03250, 2016.

[160] S. Srinivas, R. V. Babu, Data-free parameter pruning for deep neural networks, in: Proceedings of the British Machine Vision Conference (BMVC), 2015.

[161] Z. Mariet, S. Sra, Diversity networks, in: Proceedings of the International Conference on Learning Representations (ICLR), 2015.

[162] W. Chen, J. T. Wilson, S. Tyree, K. Q. Weinberger, Y. Chen, Compressing neural networks with the hashing trick, in: Proceedings of the International Conference on Machine Learning (ICML), 2015, pp. 2285–2294.

[163] Q. Shi, J. Petterson, G. Dror, J. Langford, A. Smola, S. Vishwanathan, Hash kernels for structured data, Journal of Machine Learning Research (JMLR) 10 (2009) 2615–2637.

[164] K. Weinberger, A. Dasgupta, J. Langford, A. Smola, J. Attenberg, Feature hashing for large scale multitask learning, in: Proceedings of the International Conference on Machine Learning (ICML), 2009, pp. 1113–1120.

[165] B. Liu, M. Wang, H. Foroosh, M. Tappen, M. Pensky, Sparse convolutional neural networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 806–814.

[166] W. Wen, C. Wu, Y. Wang, Y. Chen, H. Li, Learning structured sparsity in deep neural networks, in: Proceedings of the Advances in Neural Information Processing Systems (NIPS), 2016, pp. 2074–2082.

[167] H. Bagherinezhad, M. Rastegari, A. Farhadi, Lcnn: Lookup-based convolutional neural network, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.

[168] M. Egmont-Petersen, D. de Ridder, H. Handels, Image processing with neural networksa review, Pattern recognition 35 (10) (2002) 2279–2301.

[169] K. Nogueira, O. A. Penatti, J. A. dos Santos, Towards better exploiting convolutional neural networks for remote sensing scene classification, Pattern Recognition 61 (2017) 539–556.

[170] Z. Zuo, G. Wang, B. Shuai, L. Zhao, Q. Yang, Exemplar based deep discriminative and shareable feature learning for scene image classification, Pattern Recognition 48 (10) (2015) 3004–3015.

[171] A. T. Lopes, E. de Aguiar, A. F. De Souza, T. Oliveira-Santos, Facial expression recognition with convolutional neural networks: Coping with few data and the training sample order, Pattern Recognition 61 (2017) 610–628.

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

[173] A.-M. Tousch, S. Herbin, J.-Y. Audibert, Semantic hierarchies for image annotation: A survey, Pattern Recognition 45 (1) (2012) 333–345.

[174] N. Srivastava, R. R. Salakhutdinov, Discriminative transfer learning with tree-based priors, in: Proceedings of the Advances in Neural Information Processing Systems (NIPS), 2013, pp. 2094–2102.

[175] Z. Wang, X. Wang, G. Wang, Learning fine-grained features via a cnn tree for large-scale classification, CoRR abs/1511.04534.

[176] T. Xiao, J. Zhang, K. Yang, Y. Peng, Z. Zhang, Error-driven incremental learning in deep convolutional neural network for large-scale image classification, in: Proceedings of the ACM Multimedia Conference, 2014, pp. 177–186.

[177] Z. Yan, V. Jagadeesh, D. DeCoste, W. Di, R. Piramuthu, Hd-cnn: Hierarchical deep convolutional neural network for image classification, in: Proceedings of the International Conference on Computer Vision (ICCV),

pp. 2740–2748.

[178] T. Berg, J. Liu, S. W. Lee, M. L. Alexander, D. W. Jacobs, P. N. Belhumeur, Birdsnap: Large-scale fine-grained visual categorization of birds, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 2019–2026.

[179] A. Khosla, N. Jayadevaprakash, B. Yao, F.-F. Li, Novel dataset for fine-grained image categorization: Stanford dogs, in: Proceedings of the IEEE International Conference on Computer Vision (CVPR Workshops, Vol. 2, 2011, p. 1.

[180] L. Yang, P. Luo, C. C. Loy, X. Tang, A large-scale car dataset for fine-grained categorization and verification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 3973–3981.

[181] M. Minervini, A. Fischbach, H. Scharr, S. A. Tsaftaris, Finely-grained annotated datasets for image-based plant phenotyping, Pattern recognition letters 81 (2016) 80–89.

[182] G.-S. Xie, X.-Y. Zhang, W. Yang, M.-L. Xu, S. Yan, C.-L. Liu, Lg-cnn: From local parts to global discrimination for fine-grained recognition, Pattern Recognition 71 (2017) 118–131.

[183] S. Branson, G. Van Horn, P. Perona, S. Belongie, Improved bird species recognition using pose normalized deep convolutional nets, in: Proceedings of the British Machine Vision Conference (BMVC), 2014.

[184] N. Zhang, J. Donahue, R. Girshick, T. Darrell, Part-based r-cnns for fine-grained category detection, in: Proceedings of the European Conference on Computer Vision (ECCV), 2014, pp. 834–849.

[185] J. R. Uijlings, K. E. van de Sande, T. Gevers, A. W. Smeulders, Selective search for object recognition, International Journal of Conflict and Violence (IJCV) 104 (2) (2013) 154–171.

[186] D. Lin, X. Shen, C. Lu, J. Jia, Deep lac: Deep localization, alignment and classification for fine-grained recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 1666–1674.

[187] J. P. Pluim, J. A. Maintz, M. Viergever, et al., Mutual-information-based registration of medical images: a survey, IEEE Trans. Med. Imaging 22 (8) (2003) 986–1004.

[188] J. Krause, T. Gebru, J. Deng, L.-J. Li, L. Fei-Fei, Learning features and parts for fine-grained recognition, in: Proceedings of the International Conference on Pattern Recognition (ICPR), 2014, pp. 26–33.

[189] J. Krause, H. Jin, J. Yang, L. Fei-Fei, Fine-grained recognition without part annotations, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 5546–5555.

[190] Y. Zhang, X.-S. Wei, J. Wu, J. Cai, J. Lu, V.-A. Nguyen, M. N. Do, Weakly supervised fine-grained categorization with part-based image representation, IEEE Transactions on Image Processing 25 (4) (2016) 1713–1725.

[191] T. Xiao, Y. Xu, K. Yang, J. Zhang, Y. Peng, Z. Zhang, The application of two-level attention models in deep convolutional neural network for fine-grained image classification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 842–850.

[192] T.-Y. Lin, A. RoyChowdhury, S. Maji, Bilinear cnn models for fine-grained visual recognition, in: Proceedings of the International Conference on Computer Vision (ICCV), 2015, pp. 1449–1457.

[193] D. T. Nguyen, W. Li, P. O. Ogunbona, Human detection from images and videos: A survey, Pattern Recognition 51 (2016) 148–175.

[194] Y. Li, S. Wang, Q. Tian, X. Ding, Feature representation for statistical-learning-based object detection: A review, Pattern Recognition 48 (11) (2015) 3542–3559.

[195] M. Pedersoli, A. Vedaldi, J. Gonzalez, X. Roca, A coarse-to-fine approach for fast deformable object detection, Pattern Recognition 48 (5) (2015) 1844–1853.

[196] S. J. Nowlan, J. C. Platt, A convolutional neural network hand tracker, in: Proceedings of the Advances in Neural Information Processing Systems (NIPS), 1994, pp. 901–908.

[197] R. Girshick, F. Iandola, T. Darrell, J. Malik, Deformable part models are convolutional neural networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 437–446.

[198] R. Vaillant, C. Monrocq, Y. Le Cun, Original approach for the localisation of objects in images, IEE Proceedings-Vision, Image and Signal Processing 141 (4) (1994) 245–250.

[199] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C. L. Zitnick, Microsoft coco: Common objects in context, in: Proceedings of the European Conference on Computer Vision (ECCV), 2014, pp. 740–755.

[200] I. Endres, D. Hoiem, Category independent object proposals, IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) 36 (2) (2014) 222–234.

[201] L. Gómez, D. Karatzas, Textproposals: a text-specific selective search algorithm for word spotting in the wild, Pattern Recognition 70 (2017) 60–74.

[202] R. Girshick, J. Donahue, T. Darrell, J. Malik, Rich feature hierarchies for accurate object detection and semantic segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 580–587.

[203] K. He, X. Zhang, S. Ren, J. Sun, Spatial pyramid pooling in deep convolutional networks for visual recognition, IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) 37 (9) (2015) 1904–1916.

[204] S. Ren, K. He, R. Girshick, J. Sun, Faster r-cnn: Towards real-time object detection with region proposal networks, IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) 39 (6) (2017) 1137–1149.

[205] S. Gidaris, N. Komodakis, Object detection via a multi-region and semantic segmentation-aware cnn model, in: Proceedings of the International Conference on Computer Vision (ICCV), 2015, pp. 1134–1142.

[206] D. Yoo, S. Park, J.-Y. Lee, A. S. Paek, I. So Kweon, Attentionnet: Aggregating weak directions for accurate object detection, in: Proceedings of the International Conference on Computer Vision (ICCV), 2015, pp. 2659–2667.

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

[208] E. Simo-Serra, E. Trulls, L. Ferraz, I. Kokkinos, F. Moreno-Noguer, Fracking deep convolutional image descriptors, CoRR abs/1412.6537.

[209] A. Shrivastava, A. Gupta, R. Girshick, Training region-based object detectors with online hard example mining, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 761–769.

[210] J. Redmon, S. Divvala, R. Girshick, A. Farhadi, You only look once: Unified, real-time object detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 779–788.

[211] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, Ssd: Single shot multibox detector, in: Proceedings of the European Conference on Computer Vision (ECCV), 2016, pp. 21–37.

[212] Y. Lu, T. Javidi, S. Lazebnik, Adaptive object detection using adjacency and zoom prediction, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 2351–2359.

[213] K. Zhang, H. Song, Real-time visual tracking via online weighted multiple instance learning, Pattern Recognition 46 (1) (2013) 397–411.

[214] S. Zhang, H. Yao, X. Sun, X. Lu, Sparse coding based visual tracking: Review and experimental comparison, Pattern Recognition 46 (7) (2013) 1772–1788.

[215] S. Zhang, J. Wang, Z. Wang, Y. Gong, Y. Liu, Multi-target tracking by learning local-to-global trajectory models, Pattern Recognition 48 (2) (2015) 580–590.

[216] J. Fan, W. Xu, Y. Wu, Y. Gong, Human tracking using convolutional neural networks, IEEE Trans. Neural Networks (TNN) 21 (10) (2010) 1610–1623.

[217] H. Li, Y. Li, F. Porikli, Deeptrack: Learning discriminative feature representations by convolutional neural networks for visual tracking, in: Proceedings of the British Machine Vision Conference (BMVC), 2014.

[218] Y. Chen, X. Yang, B. Zhong, S. Pan, D. Chen, H. Zhang, Cnntracker: Online discriminative object tracking via deep convolutional neural network, Appl. Soft Comput. 38 (2016) 1088–1098.

[219] S. Hong, T. You, S. Kwak, B. Han, Online tracking by learning discriminative saliency map with convolutional neural network, in: Proceedings of the International Conference on Machine Learning (ICML), 2015, pp. 597–606.

[220] M. Patacchiola, A. Cangelosi, Head pose estimation in the wild using convolutional neural networks and adaptive gradient methods, Pattern Recognition 71 (2017) 132–143.

[221] K. Nishi, J. Miura, Generation of human depth images with body part labels for complex human pose recognition, Pattern Recognition.

[222] A. Toshev, C. Szegedy, Deeppose: Human pose estimation via deep neural networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 1653–1660.

[223] A. Jain, J. Tompson, M. Andriluka, G. W. Taylor, C. Bregler, Learning human pose estimation features with convolutional networks, in: Proceedings of the International Conference on Learning Representations (ICLR), 2014.

[224] J. J. Tompson, A. Jain, Y. LeCun, C. Bregler, Joint training of a convolutional network and a graphical model for human pose estimation, in: Proceedings of the Advances in Neural Information Processing Systems (NIPS), 2014, pp. 1799–1807.

[225] X. Chen, A. L. Yuille, Articulated pose estimation by a graphical model with image dependent pairwise relations, in: Proceedings of the Advances in Neural Information Processing Systems (NIPS), 2014, pp. 1736–1744.

[226] X. Chen, A. Yuille, Parsing occluded people by flexible compositions, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 3945–3954.

[227] X. Fan, K. Zheng, Y. Lin, S. Wang, Combining local appearance and holistic view: Dual-source deep neural networks for human pose estimation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 1347–1355.

[228] A. Jain, J. Tompson, Y. LeCun, C. Bregler, Modeep: A deep learning framework using motion features for human pose estimation, in: Proceedings of the Asian Conference on Computer Vision (ACCV), 2014, pp. 302–315.

[229] Y. Y. Tang, S.-W. Lee, C. Y. Suen, Automatic document processing: a survey, Pattern recognition 29 (12) (1996) 1931–1952.

[230] A. Vinciarelli, A survey on off-line cursive word recognition, Pattern recognition 35 (7) (2002) 1433–1446.

[231] K. Jung, K. I. Kim, A. K. Jain, Text information extraction in images and video: a survey, Pattern recognition 37 (5) (2004) 977–997.

[232] S. Eskenazi, P. Gomez-Krämer, J.-M. Ogier, A comprehensive survey of mostly textual document segmentation algorithms since 2008, Pattern Recognition 64 (2017) 1–14.

[233] X. Bai, B. Shi, C. Zhang, X. Cai, L. Qi, Text/non-text image classification in the wild with convolutional neural networks, Pattern Recognition 66 (2017) 437–446.

[234] L. Gomez, A. Nicolaou, D. Karatzas, Improving patch-based scene text script identification with ensembles of conjoined networks, Pattern Recognition 67 (2017) 85–96.

[235] M. Delakis, C. Garcia, Text detection with convolutional neural networks, in: Proceedings of the International Conference on Computer Vision Theory and Applications (VISAPP), 2008, pp. 290–294.

[236] H. Xu, F. Su, Robust seed localization and growing with deep convolutional features for scene text detection, in: Proceedings of the International Conference on Multimedia Retrieval (ICMR), 2015, pp. 387–394.

[237] W. Huang, Y. Qiao, X. Tang, Robust scene text detection with convolution neural network induced mser trees, in: Proceedings of the European Conference on Computer Vision (ECCV), 2014, pp. 497–511.

[238] C. Zhang, C. Yao, B. Shi, X. Bai, Automatic discrimination of text and non-text natural images, in: Proceedings of the International Conference on Document Analysis and Recognition (ICDAR), 2015, pp. 886–890.

[239] I. J. Goodfellow, J. Ibarz, S. Arnoud, V. Shet, Multi-digit number recognition from street view imagery using deep convolutional neural networks, in: Proceedings of the International Conference on Learning Representations (ICLR), 2014.

[240] M. Jaderberg, K. Simonyan, A. Vedaldi, A. Zisserman, Deep structured output learning for unconstrained text recognition, in: Proceedings of the International Conference on Learning Representations (ICLR), 2015.

[241] P. He, W. Huang, Y. Qiao, C. C. Loy, X. Tang, Reading scene text in deep convolutional sequences, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2016, pp. 3501–3508.

[242] F. A. Gers, J. Schmidhuber, F. Cummins, Learning to forget: Continual prediction with lstm, Neural Computation 12 (10) (2000) 2451–2471.

[243] B. Shi, X. Bai, C. Yao, An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition, CoRR abs/1507.05717.

[244] M. Jaderberg, A. Vedaldi, A. Zisserman, Deep features for text spotting, in: Proceedings of the European Conference on Computer Vision (ECCV), 2014, pp. 512–528.

[245] M. Jaderberg, K. Simonyan, A. Vedaldi, A. Zisserman, Reading text in the wild with convolutional neural networks, Vol. 116, 2016, pp. 1–20.

[246] L. Wang, H. Lu, X. Ruan, M.-H. Yang, Deep networks for saliency detection via local estimation and global search, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 3183–3192.

[247] R. Zhao, W. Ouyang, H. Li, X. Wang, Saliency detection by multi-context deep learning, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 1265–1274.

[248] G. Li, Y. Yu, Visual saliency based on multiscale deep features, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 5455–5463.

[249] N. Liu, J. Han, D. Zhang, S. Wen, T. Liu, Predicting eye fixations using convolutional neural networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 362–370.

[250] S. He, R. W. Lau, W. Liu, Z. Huang, Q. Yang, Supercnn: A superpixelwise convolutional neural network for salient object detection, International Journal of Conflict and Violence (IJCV) 115 (3) (2015) 330–344.

[251] E. Vig, M. Dorr, D. Cox, Large-scale optimization of hierarchical features for saliency prediction in natural images, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 2798–2805.

[252] M. Kmmerer, L. Theis, M. Bethge, Deep gaze i: Boosting saliency prediction with feature maps trained on imagenet, in: Proceedings of the International Conference on Learning Representations (ICLR) Workshops, 2015.

[253] J. Pan, X. Gir-i Nieto, End-to-end convolutional network for saliency prediction, CoRR abs/1507.01422.

[254] G. Guo, A. Lai, A survey on still image based human action recognition, Pattern Recognition 47 (10) (2014) 3343–3361.

[255] L. L. Presti, M. La Cascia, 3d skeleton-based human action classification: A survey, Pattern Recognition 53 (2016) 130–147.

[256] J. Zhang, W. Li, P. O. Ogunbona, P. Wang, C. Tang, Rgb-d-based action recognition datasets: A survey, Pattern Recognition 60 (2016) 86–105.

[257] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, T. Darrell, Decaf: A deep convolutional activation feature for generic visual recognition (2014).

[258] M. Oquab, L. Bottou, I. Laptev, J. Sivic, Learning and transferring mid-level image representations using convolutional neural networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 1717–1724.

[259] G. Gkioxari, R. Girshick, J. Malik, Actions and attributes from wholes and parts, in: Proceedings of the International Conference on Computer Vision (ICCV), 2015, pp. 2470–2478.

[260] L. Pishchulin, M. Andriluka, P. Gehler, B. Schiele, Poselet conditioned pictorial structures, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 588–595.

[261] G. Gkioxari, R. B. Girshick, J. Malik, Contextual action recognition with r*cnn, in: Proceedings of the International Conference on Computer Vision (ICCV), 2015, pp. 1080–1088.

[262] G. Gkioxari, R. Girshick, J. Malik, Actions and attributes from wholes and parts, in: Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 2470–2478.

[263] Y. Zhang, L. Cheng, J. Wu, J. Cai, M. N. Do, J. Lu, Action recognition in still images with minimum annotation efforts, IEEE Transactions on Image Processing 25 (11) (2016) 5479–5490.

[264] L. Wang, L. Ge, R. Li, Y. Fang, Three-stream cnns for action recognition, Pattern Recognition Letters 92 (2017) 33–40.

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

[266] D. Tran, L. Bourdev, R. Fergus, L. Torresani, M. Paluri, Learning spatiotemporal features with 3d convolutional networks, in: Proceedings of the International Conference on Computer Vision (ICCV), 2015, pp. 4489–4497.

[267] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, L. Fei-Fei, Large-scale video classification with convolutional neural networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 1725–1732.

[268] K. Simonyan, A. Zisserman, Two-stream convolutional networks for action recognition in videos, in: Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, K. Weinberger (Eds.), Proceedings of the Advances in Neural Information Processing Systems (NIPS), 2014, pp. 568–576.

[269] G. Chéron, I. Laptev, C. Schmid, P-CNN: pose-based CNN features for action recognition, in: Proceedings of the International Conference on Computer Vision (ICCV), 2015, pp. 3218–3226.

[270] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, T. Darrell, Long-term recurrent convolutional networks for visual recognition and description, IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) 39 (4) (2017) 677–691.

[271] K.-S. Fu, J. Mui, A survey on image segmentation, Pattern recognition 13 (1) (1981) 3–16.

[272] Q. Zhou, B. Zheng, W. Zhu, L. J. Latecki, Multi-scale context for scene labeling via flexible segmentation graph, Pattern Recognition 59 (2016) 312–324.

[273] F. Liu, G. Lin, C. Shen, Crf learning with cnn features for image segmentation, Pattern Recognition 48 (10) (2015) 2983–2992.

[274] S. Bu, P. Han, Z. Liu, J. Han, Scene parsing using inference embedded deep networks, Pattern Recognition 59 (2016) 188–198.

[275] B. Peng, L. Zhang, D. Zhang, A survey of graph theoretical approaches to image segmentation, Pattern Recognition 46 (3) (2013) 1020–1038.

[276] C. Farabet, C. Couprie, L. Najman, Y. LeCun, Learning hierarchical features for scene labeling, IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) 35 (8) (2013) 1915–1929.

[277] C. Couprie, C. Farabet, L. Najman, Y. LeCun, Indoor semantic segmentation using depth information, in: Proceedings of the International Conference on Learning Representations (ICLR), 2013.

[278] P. Pinheiro, R. Collobert, Recurrent convolutional neural networks for scene labeling, in: Proceedings of the International Conference on Machine Learning (ICML), 2014, pp. 82–90.

[279] B. Shuai, G. Wang, Z. Zuo, B. Wang, L. Zhao, Integrating parametric and non-parametric models for scene labeling, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 4249–4258.

[280] B. Shuai, Z. Zuo, W. Gang, Quaddirectional 2d-recurrent neural networks for image labeling 22 (11) (2015) 1990–1994.

[281] B. Shuai, Z. Zuo, G. Wang, B. Wang, Dag-recurrent neural networks for scene labeling, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 3620–3629.

[282] M. Mostajabi, P. Yadollahpour, G. Shakhnarovich, Feedforward semantic segmentation with zoom-out features, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 3376–3385.

[283] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, A. L. Yuille, Semantic image segmentation with deep convolutional nets and fully connected crfs, in: Proceedings of the International Conference on Learning Representations (ICLR), 2015.

[284] M. El Ayadi, M. S. Kamel, F. Karray, Survey on speech emotion recognition: Features, classification schemes, and databases, Pattern Recognition 44 (3) (2011) 572–587.

[285] L. Deng, P. Kenny, M. Lennig, V. Gupta, F. Seitz, P. Mermelstein, Phonemic hidden markov models with continuous mixture output densities for large vocabulary word recognition, IEEE Trans. Signal Processing 39 (7) (1991) 1677–1681.

[286] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, et al., Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups, IEEE Signal Process. Mag. 29 (6) (2012) 82–97.

[287] L. Deng, J. Li, J.-T. Huang, K. Yao, D. Yu, F. Seide, M. Seltzer, G. Zweig, X. He, J. Williams, et al., Recent advances in deep learning for speech research at microsoft, in: Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2013, pp. 8604–8608.

[288] K. Yao, D. Yu, F. Seide, H. Su, L. Deng, Y. Gong, Adaptation of context-dependent deep neural networks for automatic speech recognition, in: Proceedings of the Spoken Language Technology (SLT), 2012, pp. 366–369.

[289] O. Abdel-Hamid, A.-r. Mohamed, H. Jiang, G. Penn, Applying convolutional neural networks concepts to hybrid nnhmm model for speech recognition, in: Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2012, pp. 4277–4280.

[290] O. Abdel-Hamid, A.-R. Mohamed, H. Jiang, L. Deng, G. Penn, D. Yu, Convolutional neural networks for speech recognition, in: Proceedings of the International Conference on Learning Representations (ICLR), 2014.

[291] D. Palaz, R. Collobert, M. M. Doss, Estimating phoneme class conditional probabilities from raw speech signal using convolutional neural networks, in: Proceedings of the International Speech Communication Association (INTERSPEECH), 2013, pp. 1766–1770.

[292] Y. Hoshen, R. J. Weiss, K. W. Wilson, Speech acoustic modeling from raw multichannel waveforms, in: Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2015, pp. 4624–4628.

[293] D. Amodei, R. Anubhai, E. Battenberg, C. Case, J. Casper, B. Catanzaro, J. Chen, M. Chrzanowski, A. Coates, G. Diamos, et al., Deep speech 2: End-to-end speech recognition in english and mandarin, 2016, pp. 173–182.

[294] T. Sercu, V. Goel, Advances in very deep convolutional neural networks for lvcsr, in: Proceedings of the International Speech Communication Association (INTERSPEECH), 2016, pp. 3429–3433.

[295] L. Tóth, Convolutional deep maxout networks for phone recognition., in: Proceedings of the International Speech Communication Association (INTERSPEECH), 2014, pp. 1078–1082.

[296] T. N. Sainath, B. Kingsbury, A. Mohamed, G. E. Dahl, G. Saon, H. Soltau, T. Beran, A. Y. Aravkin, B. Ramabhadran, Improvements to deep convolutional neural networks for LVCSR, in: Proceedings of the Automatic Speech Recognition and Understanding (ASRU) Workshops, 2013, pp. 315–320.

[297] D. Yu, W. Xiong, J. Droppo, A. Stolcke, G. Ye, J. Li, G. Zweig, Deep convolutional neural networks with layer-wise context expansion and attention, in: Proceedings of the International Speech Communication Association (INTERSPEECH), 2016, pp. 17–21.

[298] A. Waibel, T. Hanazawa, G. Hinton, K. Shikano, K. J. Lang, Phoneme recognition using time-delay neural networks, IEEE Trans. Acoustics, Speech, and Signal Processing 37 (3) (1989) 328–339.

[299] L.-H. Chen, T. Raitio, C. Valentini-Botinhao, J. Yamagishi, Z.-H. Ling, Dnn-based stochastic postfilter for hmm-based speech synthesis., in: Proceedings of the International Speech Communication Association (INTERSPEECH), 2014, pp. 1954–1958.

[300] B. Uria, I. Murray, S. Renals, C. Valentini-Botinhao, J. Bridle, Modelling acoustic feature dependencies with artificial neural networks: Trajectory-rnade, in: Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2015, pp. 4465–4469.

[301] Z. Huang, S. M. Siniscalchi, C.-H. Lee, Hierarchical bayesian combination of plug-in maximum a posteriori decoders in deep neural networks-based speech recognition and speaker adaptation, Pattern Recognition Letters.

[302] A. van den Oord, N. Kalchbrenner, K. Kavukcuoglu, Pixel recurrent neural networks, in: Proceedings of the International Conference on Machine Learning (ICML), 2016, pp. 1747–1756.

[303] R. Jozefowicz, O. Vinyals, M. Schuster, N. Shazeer, Y. Wu, Exploring the limits of language modeling, in: Proceedings of the International Conference on Learning Representations (ICLR), 2016.

[304] Y. Kim, Y. Jernite, D. Sontag, A. M. Rush, Character-aware neural language models, in: Proceedings of the Association for the Advancement of Artificial Intelligence (AAAI), 2016, pp. 2741–2749.

[305] J. Gu, C. Jianfei, G. Wang, T. Chen, Stack-captioning: Coarse-to-fine learning for image captioning, Vol. abs/1709.03376, 2017.

[306] M. Wang, Z. Lu, H. Li, W. Jiang, Q. Liu, gen cnn: A convolutional architecture for word sequence prediction, in: Proceedings of the Association for Computational Linguistics (ACL), 2015, pp. 1567–1576.

[307] J. Gu, G. Wang, C. Jianfei, T. Chen, An empirical study of language cnn for image captioning, in: Proceedings of the International Conference on Computer Vision (ICCV), 2017.

[308] M. A. D. G. Yann N. Dauphin, Angela Fan, Language modeling with gated convolutional networks, in: Proceedings of the International Conference on Machine Learning (ICML), 2017, pp. 933–941.

[309] R. Collobert, J. Weston, A unified architecture for natural language processing: Deep neural networks with multitask learning, in: Proceedings of the International Conference on Machine Learning (ICML), 2008, pp. 160–167.

[310] L. Yu, K. M. Hermann, P. Blunsom, S. Pulman, Deep learning for answer sentence selection, in: Proceedings of the Advances in Neural Information Processing Systems (NIPS) Workshop, 2014.

[311] N. Kalchbrenner, E. Grefenstette, P. Blunsom, A convolutional neural network for modelling sentences, in: Proceedings of the Association for Computational Linguistics (ACL), 2014, pp. 655–665.

[312] Y. Kim, Convolutional neural networks for sentence classification, in: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), 2014, pp. 1746–1751.

[313] W. Yin, H. Schütze, Multichannel variable-size convolution for sentence classification, in: Proceedings of the Conference on Natural Language Learning (CoNLL), 2015, pp. 204–214.

[314] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, P. Kuksa, Natural language processing (almost) from scratch, Journal of Machine Learning Research (JMLR) 12 (2011) 2493–2537.

[315] A. Conneau, H. Schwenk, L. Barrault, Y. Lecun, Very deep convolutional networks for natural language processing, CoRR abs/1606.01781.

[316] G. Huang, Y. Sun, Z. Liu, D. Sedra, K. Weinberger, Deep networks with stochastic depth, in: Proceedings of the European Conference on Computer Vision (ECCV), 2016, pp. 646–661.