第5章 深度学习方法

深度神经网络在实际应用中能够较好地拟合复杂系统模型，而深度学习则能够对多层结构化神经网络进行建模。将神经网络与贝叶斯网络理论结合起来，形成广义的神经网络概念和框架。随着神经网络层数的增加，其学习和推理会遇到诸多挑战，例如如何选择网络结构和设计优化算法来实现高效稳定的学习或分类任务。本章将对其中最有代表性的网络模型进行介绍，包括卷积神经网络、循环神经网络、图神经网络、深度信念网络和深度生成网络。

深度卷积神经网络采用局部连接和权重共享的方式来减少待估计参数的数量并提高网络训练的效率。

循环神经网络通过加入信息反馈机制来提高网络的记忆能力，并引入注意力机制来提升有限计算资源下的重要信息处理能力。

图神经网络将深度神经网络应用于结构化数据处理。从网络结构角度来看，图结构包含前馈和反馈结构，因此图神经网络的泛化能力更强。

深度信念网络通常用概率图模型来表示，网络中包含多层隐变量，能够有效学习数据的内部特征和生成机理，有助于随机样本数据的分类和回归。

深度生成网络通常包含分布函数估计和样本生成两个基本功能。针对小样本学习任务，深度生成网络能够实现样本增强，从而有效解决分类或回归难题。

5.1 深度网络概述

5.1.1 深度网络定义和种类

深度神经网络 (Deep Neural Networks, DNN) 本质上是含有多层感知机的神经网络，包括输入层、隐藏层、输出层三部分。根据神经元之间连接结构和模式的不同，深度神经网络可分为卷积神经网络、循环神经网络、深度玻耳兹曼机和深度信念网络等。

1. 卷积神经网络

卷积神经网络（Convolutional Neural Networks, CNN）是一种具有局部连接和权值共享特征的深层前馈网络。卷积神经网络的提出主要是为了克服全连接神经网络训练效率低的问题。在卷积神经网络出现之前，使用全连接的神经网络学习来处理图像分类问题面临很大挑战。图像的庞大数据需要大规模神经网络来支撑，消耗大量存储和计算资源，且很难实现快速处理。CNN 从视觉皮层的生物学上获得启发，利用卷积和池化操作实现了逐层特征提取，保留图像特征的同时大大减少了参数数量，这使得卷积神经网络在图像识别、目标检测等图像处理问题上极具优越性。

典型的卷积神经网络包括卷积层、池化层、全连接层，如图 5.1所示。卷积层能够提取图像中的局部特征，池化层可以减少参数数量并避免过拟合，全连接层用于输出最终结果。

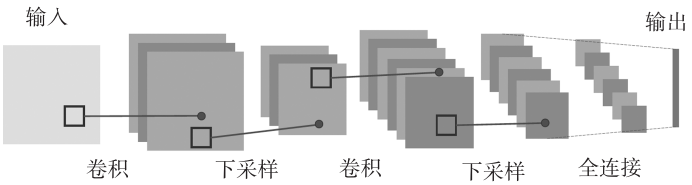

图 5.1　卷积神经网络结构示意图

2. 循环神经网络

DNN 和 CNN 通常只处理静态输入数据，即前后相邻时刻的两个输入不会相互影响。然而，在自然语言处理和语音识别等任务中，前后相邻时刻的输入数据具有关联性。为了能对时间序列上的信息变化进行建模，提出了循环神经网络（Recurrent Neural Networks, RNN）。RNN 结构图及在时间轴上展开的模型如图 5.2所示，x_t 是 t 时刻的输入样本；o_t 是 t 时刻的输出，$o_t = g(v \cdot s_t)$；s_t 表示样本在时间 t 处的记忆，$s_t = f(w \cdot s_{t-1} + u \cdot x_t)$，其中 w 表示状态权重，u 表示输入样本权重，v 表示输出样本权重。

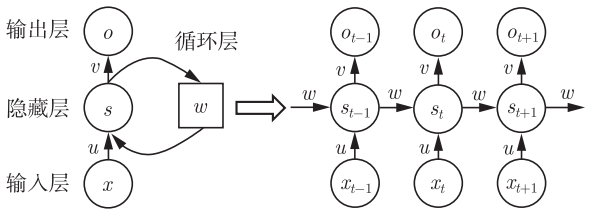

图 5.2　循环神经网络结构示意图

RNN 所对应的状态空间模型可以表示成

$$h_t = u \cdot x_t + w \cdot s_{t-1}$$

$$s_t = f(h_t)$$

$$o_t = g(v \cdot s_t)$$

式中：$f(\cdot)$ 和 $g(\cdot)$ 为激活函数，$f(\cdot)$ 一般是 tanh、relu、sigmoid 等激活函数，$g(\cdot)$ 通常是 softmax。

循环网络中不同时刻的 w、u、v 是权重共享的，减少了参数数量。不难发现网络中上一时刻状态 s_{t-1} 参与当前时刻的状态更新。

3. 深度玻耳兹曼机

玻耳兹曼机（Boltzmann Machine，BM）是一种无向概率图模型，每个节点状态以一定概率受到其他节点影响，节点的状态值满足统计热力学中的玻耳兹曼分布。玻耳兹曼机由可观测的节点（可见节点）和不可观测的节点（隐藏节点）构成，所有节点相互连接。如图 5.3所示，隐藏节点有 0 和 1 两种状态，分别代表抑制和激活状态，可见节点可以是二值或实数。玻耳兹曼机能够用于解决两类问题：一类是搜索问题，当给定节点之间的连接权

重时，需要找到一组二值向量，使得整个网络的能量最低；另一类是学习问题，当给定节点的多组观测值时，采用模拟退火算法学习网络的最优权重。含有隐藏变量的玻耳兹曼机训练起来比较困难，于是引入了受限玻耳兹曼机（Restricted Boltzmann Machine，RBM）。

受限玻耳兹曼机有两层结构，分别由可观测节点构成的可见层和由隐藏节点构成的隐藏层，两层节点之间相互连接，同层节点互不相连，结构如图 5.4 所示。

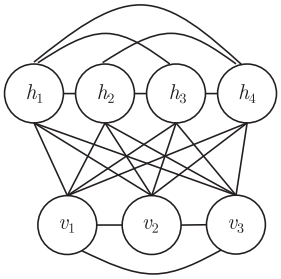

图 5.3　玻耳兹曼机结构示意图

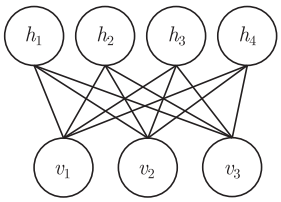

图 5.4　受限玻耳兹曼机结构示意图

RBM 通过无监督的学习方式来“重建”数据分布。RBM 在前向学习过程中从可见节点输入样本并预测隐藏节点的输出激活值；在反向传播过程中，在给定输出激活值情况下估计可见节点层的概率分布，并用 KL 散度来度量输入端两个分布的相似性。通过多次前向和反向传播让 RBM 学会逼近原始数据分布，从而实现“重建”。

深度玻耳兹曼机（Deep Boltzmann Machine，DBM）通过增加受限玻耳兹曼机中隐层数目来获得，结构如图 5.5所示。DBM 采用逐层贪婪无监督训练方法，可看作多个受限玻耳兹曼机的堆叠，即前一个玻耳兹曼机训练好后的隐藏层作为下一个玻耳兹曼机的可见层。深度玻耳兹曼机是生成式概率模型，一般作为深度神经网络的预训练网络。

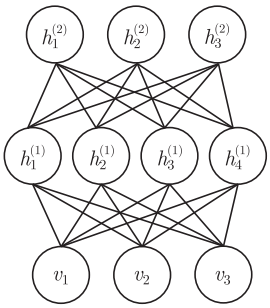

图 5.5　深度玻耳兹曼机结构示意图

4. 深度信念网络

深度信念网络（Deep Belief Net，DBN）的基本组成部分也是受限玻耳兹曼机，但它包含了有向图和无向图。DBN 同一层节点互不相连，相邻两层节点之间全连接或稀疏连接，最顶部两层隐节点之间是无向连接的，其余层之间从上到下为有向连接，其结构如图 5.6所示。在训练 DBN 模型时，先进行预训练：逐层对每个 RBM 训练，然后利用反向传播算法对模型进行微调。预训练过程相当于参数初始化过程，使 DBN 克服了随机初始

化权值参数导致网络训练陷入局部最优以及训练时间长的缺点。DBN 可以应用于图像识别、信息检索、自然语言理解、故障预测等。

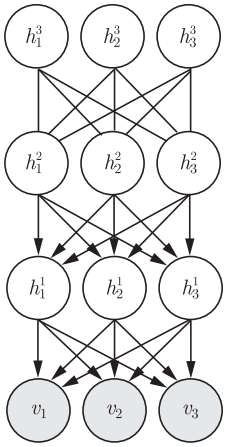

图 5.6　深度信念网络结构示意图

5.1.2　深度网络特点

深度网络的产生是因为更复杂系统建模的需求。模型学习的复杂度和容量有关，扩大模型容量可以将模型变深或变宽来实现。实验证明给定相同数量参数时，窄深度网络拟合性能更好。在一个神经网络中，浅层神经网络从输入数据中提取的是简单特征信息。随着网络层次的增加，深层神经元能提取出更复杂特征信息。

深度网络采用了与传统浅层神经网络相似的分层结构，其包括输入层、隐藏层（多层）、输出层，相邻层节点之间有连接，同一层以及跨层节点间无连接。在传统神经网络中，使用反向传播算法进行模型优化。深度网络本质还是神经网络，也能使用 BP 算法。但是随着网络层数增加，仅使用 BP 算法容易出现梯度消失以及陷入局部最优等问题。在训练深度网络模型时，通过自下而上逐层进行无监督训练，对神经网络进行初始化。然后，对网络模型进行自上而下的有监督训练，对网络进行微调。由于逐层训练使模型参数初值更接近全局最优，从而能够得到较好的训练结果。

5.2　深度卷积神经网络

卷积神经网络是一种具有局部连接、权重共享特征的深层前馈神经网络。卷积在不同维度上有不同的定义。在二维空间上，给定图像 $\boldsymbol{X} \in \mathbb{R}^{M \times N}$ 和滤波器 $\boldsymbol{W} \in \mathbb{R}^{U \times V}$，其卷积定义为

$$y_{ij} = \sum_{u=1}^{U} \sum_{v=1}^{V} w_{uv} x_{i-u+1, j-v+1}$$

或者

$$\boldsymbol{Y} = \boldsymbol{W} * \boldsymbol{X}$$

二维平面上的互相关操作定义为

$$y_{ij} = \sum_{u=1}^{U}\sum_{v=1}^{V} w_{uv} x_{i+u-1,j+v-1}$$

或者

$$\boldsymbol{Y} = \boldsymbol{W} \otimes \boldsymbol{X} = \tilde{\boldsymbol{W}} * \boldsymbol{X}$$

式中：$\tilde{\boldsymbol{W}}$ 为 $\boldsymbol{W}$ 绕原点旋转 180° 后的矩阵。

在神经网络中使用卷积是为了进行特征抽取，卷积核是否进行翻转和其特征抽取能力无关。特别是当卷积核为待学习参数时，卷积与互相关在能力上等同。

卷积操作的一个重要性质是其可交换性。若对图像边缘进行零填充：两端各补 $U-1$ 和 $V-1$ 个零，则卷积交换律可以表示成

$$\boldsymbol{W} \otimes \boldsymbol{X} = \boldsymbol{X} \otimes \boldsymbol{W} \text{ 或} \tilde{\boldsymbol{W}} * \boldsymbol{X} = \tilde{\boldsymbol{X}} * \boldsymbol{W}$$

假设 $\boldsymbol{Y} = \boldsymbol{W} \otimes \boldsymbol{X}$，函数 $f(\boldsymbol{Y})$ 为一个标量函数，则

$$\begin{aligned}\frac{\partial f(\boldsymbol{Y})}{\partial w_{uv}} &= \sum_{i=1}^{M-U+1}\sum_{j=1}^{N-V+1} \frac{\partial y_{ij}}{\partial w_{uv}} \frac{\partial f(\boldsymbol{Y})}{\partial y_{ij}} \\ &= \sum_{i=1}^{M-U+1}\sum_{j=1}^{N-V+1} \frac{\partial f(\boldsymbol{Y})}{\partial y_{ij}} x_{u+i-1,v+j-1}\end{aligned}$$

因此，有

$$\frac{\partial f(\boldsymbol{Y})}{\partial \boldsymbol{W}} = \frac{\partial f(\boldsymbol{Y})}{\partial \boldsymbol{Y}} \otimes \boldsymbol{X}$$

类似地，可以得到

$$\begin{aligned}\frac{\partial f(\boldsymbol{Y})}{\partial x_{st}} &= \sum_{i=1}^{M-U+1}\sum_{j=1}^{N-V+1} \frac{\partial y_{ij}}{\partial x_{st}} \frac{\partial f(\boldsymbol{Y})}{\partial y_{ij}} \\ &= \sum_{i=1}^{M-U+1}\sum_{j=1}^{N-V+1} w_{s-i+1,t-j+1} \frac{\partial f(\boldsymbol{Y})}{\partial y_{ij}}\end{aligned}$$

或者

$$\frac{\partial f(\boldsymbol{Y})}{\partial \boldsymbol{X}} = \tilde{\boldsymbol{W}} \otimes \frac{\partial f(\boldsymbol{Y})}{\partial \boldsymbol{Y}}$$

式中：$\tilde{\boldsymbol{W}}$ 为 $\boldsymbol{W}$ 绕原点旋转 180° 后的矩阵。

5.2.1 卷积神经网络

卷积神经网络一般由卷积层、汇聚层和全连接层构成。

在全连接前馈神经网络中，如果第 l 层有 M_l 个神经元，第 $l-1$ 层有 M_{l-1} 个神经元，连接第 $l-1$ 层和第 l 层共有 $M_l \times M_{l-1}$ 条边，则权重矩阵有 $M_l \times M_{l-1}$ 个参数。当 M_l 和 M_{l-1} 的值很大时，权重矩阵参数非常多，训练效率非常低。若采用卷积来代替全连接，则参数数量会大大减少。

根据卷积的定义，卷积层有以下两个重要性质：

（1）局部连接：在卷积层（假设是第 l 层）中的每个神经元都只和前一层（第 $l-1$ 层）中某个局部窗口内的神经元相连，构成一个局部连接网络。卷积层和前一层之间的连接数大大减少，由原来的 $M_l \times M_{l-1}$ 个连接变为 $M_l \times K$ 个连接，其中 K 为卷积核大小。

（2）权重共享：第 l 层的所有神经元都采用相同卷积核 $\boldsymbol{w}^l$ 作为训练参数。权重共享可理解为卷积核只捕捉输入数据中特定局部特征。因此，如果要提取多种特征，就需要使用多个不同卷积核。

由于局部连接和权重共享，卷积层的参数只有一个 K 维的权重 $\boldsymbol{w}^l$ 和 1 维的偏置 b^l，共 $K+1$ 个参数。参数个数和神经元的数量无关。此外，第 l 层的神经元个数不是任意选择的，而是满足 $M_l = M_{l-1} - K + 1$。

卷积网络主要应用于图像处理。为了更充分地利用图像的局部信息，通常将神经元组织为三维结构，其尺寸为 M（高度）$\times N$（宽度）$\times D$（深度），即由 D 个 $M \times N$ 大小的特征映射构成。特征映射是指图像在经过卷积操作之后提取到的特征。为了提高卷积网络的表征能力，在每一层使用多个不同的特征映射，以更好地表示图像特征。

在输入层，特征映射就是图像本身。如果是灰度图像，就是有一个特征映射，输入层的深度 $D=1$；如果是彩色图像，分别有 RGB 三个颜色通道的特征映射，输入层的深度 $D=3$。不失一般性，卷积层通常包含

（1）输入特征映射组：$\mathcal{X} \in \mathbb{R}^{M \times N \times D}$ 为三维张量，其中每个切片矩阵 $\boldsymbol{X}^d \in \mathbb{R}^{M \times N}$；

（2）输出特征映射组：$\mathcal{Y} \in \mathbb{R}^{M' \times N' \times P}$ 为三维张量，其中每个切片矩阵 $\boldsymbol{Y}^p \in \mathbb{R}^{M' \times N'}$；

（3）卷积核：$\mathcal{W} \in \mathbb{R}^{U \times V \times P \times D}$ 为四维张量，其中每个切片矩阵 $\boldsymbol{W}^{p,d} \in \mathbb{R}^{U \times V}$ 为一个二维卷积核。

计算特征映射 $\boldsymbol{Y}^p$ 需要用卷积核 $\boldsymbol{W}^{p,1}, \cdots, \boldsymbol{W}^{p,D}$ 分别对输入特征 $\boldsymbol{X}^1, \cdots, \boldsymbol{X}^D$ 进行卷积，然后将卷积结果求和并加上偏置 $\boldsymbol{b}^p$ 得到 $\boldsymbol{Z}^p$，最后经过非线性激活函数得到输出特征映射 $\boldsymbol{Y}^p$：

$$\boldsymbol{Z}^p = \boldsymbol{W}^p \otimes \boldsymbol{X} + \boldsymbol{b}^p = \sum_{d=1}^{D} \boldsymbol{W}^{p,d} \otimes \boldsymbol{X}^d + \boldsymbol{b}^p$$

$$\boldsymbol{Y}^p = f(\boldsymbol{Z}^p)$$

式中：$\boldsymbol{W}^p \in \mathbb{R}^{U \times V \times D}$ 为三维卷积核；$f(\cdot)$ 为非线性激活函数，一般为 ReLU 函数。

如果希望卷积层输出 P 个特征映射，将上述计算过程重复 P 次，得到 P 个输出特征映射 $\boldsymbol{Y}^1, \boldsymbol{Y}^2, \cdots, \boldsymbol{Y}^P$。在输入为 $\mathcal{X} \in \mathbb{R}^{M \times N \times D}$，输出为 $\mathcal{Y} \in \mathbb{R}^{M' \times N' \times P}$ 的卷积层中，每个输出特征映射都需要 D 个卷积核以及一个偏置。假设每个卷积核的大小为 $U \times V$，则该卷积神经网络共有 $P \times D \times (U \times V) + P$ 个参数。

汇聚层也叫子采样层，其作用是进行特征选择，降低特征数量，从而减少参数数量。卷积层虽然显著减少网络中连接的数量，但特征映射组中的神经元个数并没有显著减少。如果后面接一个分类器，分类器的输入维数依然很高，很容易出现过拟合。为了解决这个问题，在卷积层之后加上一个汇聚层来降低特征维数，从而避免过拟合。

假设汇聚层的输入特征映射组为 $\mathcal{X} \in \mathbb{R}^{M \times N \times D}$，对于其中每个特征映射 $\boldsymbol{X}^d \in \mathbb{R}^{M \times N}$ $(1 \leqslant d \leqslant D)$，将其划分为很多重叠或者不重叠的小区域 $\boldsymbol{R}^d_{m,n} (1 \leqslant m \leqslant M', 1 \leqslant n \leqslant N')$。汇聚是指对每个区域进行下采样。常见的汇聚函数有如下两种：

（1）最大汇聚：$y^d_{m,n} = \max\limits_{i \in R^d_{m,n}} x_i$，其中 x_i 为区域 $\boldsymbol{R}^d_{m,n}$ 内每个神经元的活性值。

（2）平均汇聚：$y^d_{m,n} = \dfrac{1}{|R^d_{m,n}|} \sum\limits_{i \in R^d_{m,n}} x_i$，其中 $|\boldsymbol{R}^d_{m,n}|$ 为区域内的样本数量。

对于输入特征映射 $\boldsymbol{X}^d$ 的 $M' \times N'$ 个区域进行下采样，就能够得到汇聚层的输出特征映射 $\boldsymbol{Y}^d = \{y^d_{m,n}\} (1 \leqslant m \leqslant M', 1 \leqslant n \leqslant N')$。可以看出，汇聚层不但有效地减少神经元数量，还使得网络对一些局部形变保持不变性。

常用的卷积网络整体结构如图 5.7所示。一个卷积块由连续的 M 个卷积层和 b 个汇聚层构成（M 通常设置为 $2 \sim 5$，b 为 0 或 1）。一个卷积网络中可以堆叠 N 个连续卷积块，然后在后面接着 K 个全连接层（N 的取值区间一般为 $1 \sim 100$，K 一般为 $0 \sim 2$）。

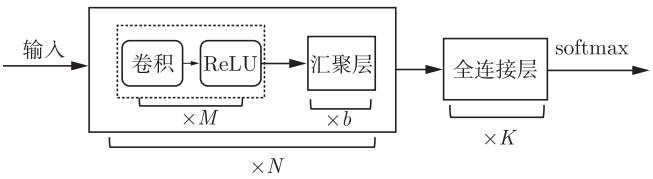

图 5.7　卷积网络整体结构

目前，卷积网络的整体结构趋向于使用更小的卷积核（如 1×1 和 3×3）以及更深的结构（如层数大于 50）。此外，由于卷积的操作性越来越灵活（如不同的步长），汇聚层的作用也变得越来越小，因此目前比较流行的卷积网络中汇聚层比例正在逐渐降低，趋向于全卷积网络。

5.2.2　参数学习

在全连接前馈神经网络中，梯度主要通过每一层的误差项进行反向传播。卷积神经网络主要有卷积层和汇聚层两种功能层。参数更新需要快速计算训练目标函数关于参数的梯度。

假如第 l 层为卷积层，第 $l-1$ 层的输入特征映射为 $\mathcal{X}^{l-1} \in \mathbb{R}^{M \times N \times D}$，通过卷积计算

得到第 l 层的特征映射净输入为 $\mathcal{Z}^l \in \mathbb{R}^{M' \times N' \times P}$。第 l 层的第 $p(1 \leqslant p \leqslant P)$ 个特征映射净输入为

$$\boldsymbol{Z}^{l,p} = \sum_{d=1}^{D} \boldsymbol{W}^{l,p,d} \otimes \boldsymbol{X}^{l-1,d} + \boldsymbol{b}^{l,p}$$

式中：$\boldsymbol{W}^{l,p,d}$ 和 $\boldsymbol{b}^{l,p}$ 为卷积核和偏置。

根据卷积函数的导数，损失函数 $\mathcal{L}$ 关于第 l 层卷积核 $\boldsymbol{W}^{l,p,d}$ 的偏导为

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{W}^{l,p,d}} = \frac{\partial \mathcal{L}}{\partial \boldsymbol{Z}^{l,p}} \otimes \boldsymbol{X}^{l-1,d} = \boldsymbol{\delta}^{l,p} \otimes \boldsymbol{X}^{l-1,d}$$

式中：$\boldsymbol{\delta}^{l,p}$ 为损失函数关于第 l 层的第 p 个特征映射净输入 $\boldsymbol{Z}^{l,p}$ 的偏导数。

同理可得，损失函数关于第 l 层的第 p 个偏置 $\boldsymbol{b}^{l,p}$ 的偏导数为

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{b}^{l,p}} = \boldsymbol{\delta}^{l,p}$$

在卷积网络中，每层参数的梯度依赖其所在层的误差项 $\boldsymbol{\delta}^{l,p}$。针对第 $l+1$ 卷积层，假设特征映射的净输入满足

$$\boldsymbol{Z}^{l+1,p} = \sum_{d=1}^{D} \boldsymbol{W}^{l+1,p,d} \otimes \boldsymbol{X}^{l,d} + \boldsymbol{b}^{l+1,p}$$

对于第 l 层第 d 个特征映射误差项 $\boldsymbol{\delta}^{l,d}$ 的具体推导过程如下：

$$\begin{aligned}
\boldsymbol{\delta}^{l,d} &= \frac{\partial \mathcal{L}}{\partial \boldsymbol{Z}^{l,d}} = \frac{\partial \boldsymbol{X}^{l,d}}{\partial \boldsymbol{Z}^{l,d}} \frac{\partial \mathcal{L}}{\partial \boldsymbol{X}^{l,d}} \\
&= f_l'(\boldsymbol{Z}^{l,d}) \odot \sum_{p=1}^{P} \left(\tilde{\boldsymbol{W}}^{l+1,p,d} \otimes \frac{\partial \mathcal{L}}{\partial \boldsymbol{Z}^{l+1,p}} \right) \\
&= f_l'(\boldsymbol{Z}^{l,d}) \odot \sum_{p=1}^{P} \left(\tilde{\boldsymbol{W}}^{l+1,p,d} \otimes \boldsymbol{\delta}^{l+1,p} \right)
\end{aligned}$$

式中：$f_l'(\boldsymbol{Z}^{l,d})$ 为第 l 层使用的激活函数导数。

针对汇聚层，由于其只有下采样操作，第 l 层的误差项只需要将第 $l+1$ 层对应的误差项进行上采样操作即可。在上述梯度的迭代计算基础之上，将采用梯度下降法对 CNN 网络参数进行学习。

5.2.3 常见卷积神经网络

常用的几种深度卷积神经网络有 LeNet-5、AlexNet、Inception 网络及残差网络（Residual Network，ResNet）等。

LeNet-5 网络提出的时间比较早，其相关的手写识别系统在 20 世纪 90 年代被金融界广泛使用。LeNet-5 网络主要有 7 层：连续两个“卷积层+汇聚层”组合，再加上一个卷积

层，一个全连接层和最后一个径向基函数构成的输出层。在传统卷积网络中，卷积层的每个输出特征映射都依赖所有输入特征映射，相当于卷积层的输入和输出特征映射之间是全连接的。实际上，这种全连接关系不是必需的。我们让每个输出特征映射都依赖少数几个输入特征映射。LeNet-5 的特点是定义一个连接表来描述输入和输出特征映射之间的连接关系，从而减少训练量。

AlexNet 是第一个现代深度卷积网络模型，其首次使用了不同深度卷积网络技术。比如使用 GPU 进行并行训练，采用 ReLU 作为非线性激活函数，使用 Dropout 防止过拟合，利用数据增强来提高模型准确率等。AlexNet 的结构包括 5 个卷积层、3 个汇聚层和 3 个全连接层（其中最后一层使用 Softmax 函数）。由于网络规模超出了当时单个 GPU 的内存限制，AlexNet 将网络拆为两半，分别放在两个 GPU 上，GPU 间只在某些层（如第 3 层）进行通信。

Inception 网络中的卷积层包含多个不同大小的卷积操作，称为 Inception 模块。Inception 网络是由多个 Inception 模块和少量汇聚层堆叠而成。Inception 模块同时使用 1×1、3×3、5×5 等不同大小卷积核，并将得到的特征映射拼接（堆叠）起来作为输出特征映射。

残差网络通过给非线性的卷积层增加直连边（也称为残差连接）的方式来提高信息传播效率。原始目标函数可以拆分成两部分，恒等函数 x 和残差函数 $h(x)-x$：

$$h(x)=x+(h(x)-x)$$

残差网络旨在采用非线性单元去近似残差函数 $h(x)-x$，而非整个目标函数 $h(x)$。根据通用逼近定理，一个由神经网络构成的非线性单元有足够的能力来逼近原始目标函数或残差函数，但后者往往更容易学习。因此，原来的优化问题转换为让非线性单元 $f(x;\theta)$ 去近似残差函数 $h(x)-x$，并用 $f(x;\theta)+x$ 去逼近 $h(x)$。残差网络就是将很多个残差单元串联起来构成的一个深度网络。

5.3 循环神经网络

5.3.1 循环网络

循环网络是一类具有短期记忆能力的神经网络，跟前馈神经网络相比，循环网络更加符合生物神经网络的结构特征，因此被广泛应用于语音识别和自然语言处理。跟 BP 训练算法类似，循环网络的参数学习可随时间反向传播，但是随着时间序列的增加，会存在梯度爆炸和消失等问题，这也称为长程依赖问题。为了解决这个问题，目前最有效的方法是引入门控机制。

循环网络是通过使用带自反馈的神经元，能够处理任意长度的时序数据。给定一个输入序列 $\{\boldsymbol{u}(t)\}_{t=1}^{T}$，循环网络通过下面公式更新隐藏层状态向量 $\boldsymbol{x}(t)$：

$$\boldsymbol{x}(t+1)=f[\boldsymbol{x}(t),\boldsymbol{u}(t)] \tag{5.1}$$