

第5章 深度学习模型与算法

MP 神经元数学模型由沃伦·麦卡洛克 (Warren McCulloch) 和沃尔特·皮茨 (Walter Pitts) 在 1943 年提出。弗兰克·罗森布拉特 (Frank Rosenblatt) 在 1958 年基于该数学模型提出了第一代神经网络单层感知器模型。第一代神经网络模型能够完成区分正方形与三角形等基本形状的任务,使人类发明真正能实现感知、学习与记忆等任务的智能机器成为可能。但是由于第一代神经网络基本原理存在的限制使得该目标长期无法取得进展。马文·闵斯基在 1969 年发表了针对第一代神经网络模型的专著,他认为,神经网络模型由人类设计固定的特征层,并证明异或 (XOR) 问题无法被单层感知器解决,这与真正的智能机器相距甚远。杰弗里·辛顿等在 1986 年提出了第二代神经网络,使用多个隐含层替换原始单一固定的特征层,采用 Sigmoid 函数作为激活函数,并利用误差的反向传播算法对模型进行运算。非线性分类问题能够被第二代神经网络有效解决。由于包括 1995 年由科琳娜·科茨 (Corinna Cortes) 与弗拉基米尔·万普尼克 (Vladimir Vapnik) 提出的支持向量机等在内的各种浅层机器学习模型在此后十多年相继被提出,浅层机器学习模型取得了长足的发展,神经网络的研究因此被搁置。

2006 年,辛顿等在探讨大脑中的图模型的过程中,提出通过自编码器的方式实现降低数据维度的任务,并针对梯度消失问题提出用预训练的方式快速训练深度信念网。此后,约书亚·本吉奥等证明预训练的方式也适用于自编码器等无监督学习算法,克里斯托弗·波尔尼特 (Christopher Poultny) 等使用基于能量的模型完成稀疏表示的学习。在深度学习的基础被这些理论奠定之后,深度学习进入快速发展的阶段。美国国防部从 2010 年开始对深度学习项目进行赞助。泽维尔·格洛特 (Xavier Glorot) 等在 2011 年为有效抑制梯度消失问题提出 ReLU 激活函数。微软公司和谷歌公司在语音识别任务中先后采用深度学习技术将错误率降低至 20%~30%,深度学习在语音识别上最先取得重大突破。辛顿和其学生在 2012 年将计算机视觉领域中的 ImageNet 图片分类问题的 Top5 错误率由 26% 降低至 15%,进一步标志着深度学习进入爆发期。多芬 (Y. N. Dauphin) 等在 2014 年,安娜·科罗曼斯卡 (Anna Choromanska) 等在 2015 年通过独立的工作分别对局部极小值问题进行了阐述,使笼罩在神经网络上的局部极值阴霾被驱散。深度学习发展历史如图 5-1 所示。在图 5-1 中,深度学习热度上升与下降的关键转折点使用空心圆圈表示,深度学习的突破用实心圆圈表示,深度学习热度的上升期使用斜向上的直线表示,深度学习热度的下降期使用斜向下的直线表示。

深度学习其实是机器学习的子集,从浅层机器学习发展到深度学习的过程中,机器学习经历了两次浪潮。浅层机器学习模型与深度学习模型之间存在着明显的差异。浅层机器学习模型不使用分布式表示,而且模型本身只是根据人为提取的特征进行分类或回归,整个系统的好坏在很大程度上由人工提取的特征质量所决定,同时,人工提取特征的过程需要大量的时间以及对专业领域知识的深度掌握;与之相反,深度学习作为一种表示学习

方法,能够自动完成特征的提取与抽象,并且深度学习模型的表示能力会伴随着模型的深度和广度增加而增强。

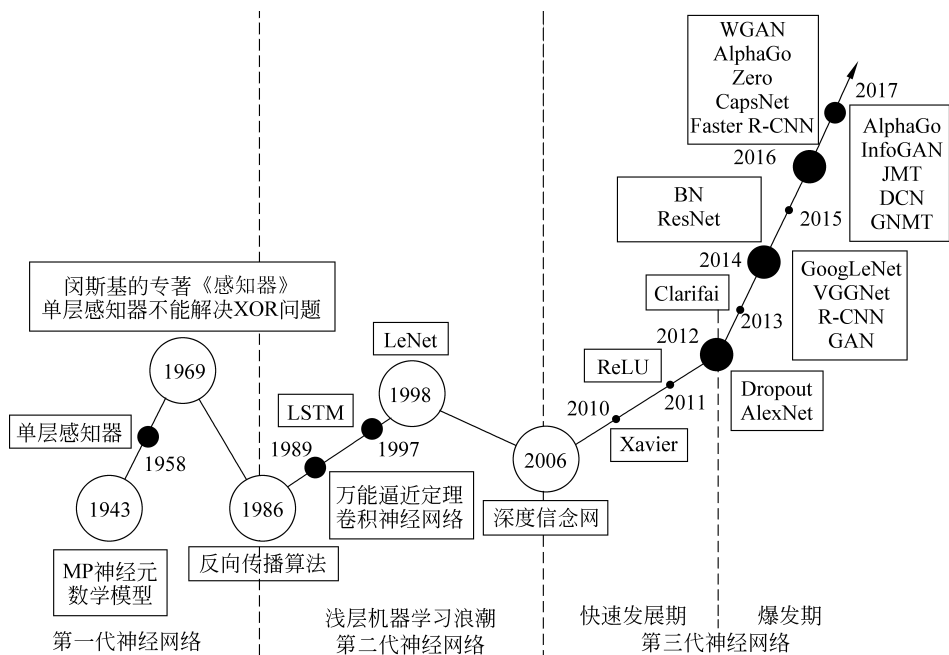


图 5-1 深度学习发展历史

5.1 循环神经网络

循环神经网络(Recurrent Neural Networks,RNN)于1986年由大卫·鲁梅尔哈特(David Rumelhart)等提出,其主要目的是完成处理序列化数据的任务。传统神经网络是由输入层、隐含层以及输出层构成的分层结构,每层包含多个神经节点。传统神经网络在同层的神经节点间不存在连接,而在相邻层的神经节点间则是全连接。这种传统的神经网络结构在很多问题上束手无策,而循环神经网络则能够发挥一定的作用。以一句话为例,因为一句话中的内容是前后相关的,所以在预测句子中被消除的内容时需要考虑前面的内容。在循环神经网络中,当前的输出结果会受到前面的输出内容的影响,循环神经网络通过记忆前面的计算信息并传入当前单元的计算过程,与当前输入数据共同计算输出值。循环神经网络隐含层之间的节点在结构上存在连接,当前节点不仅接收从输入层输入的数据,也获取前一隐含层节点的传输值。在这种信息关联机制的作用下,循环神经网络理论上能够完成任意长度的序列数据的处理任务。但是在循环神经网络的实际使用中,常常假设当前状态仅与邻近的几个过往状态相关。图 5-2 是一个典型的循环神经网络。

循环神经网络链状结构的性质表明它与列表型和序列型数据存在密切相关性。循环神经网络链状结构也是处理此类数据时使用的神经网络的自然结构。将前期信息与当前

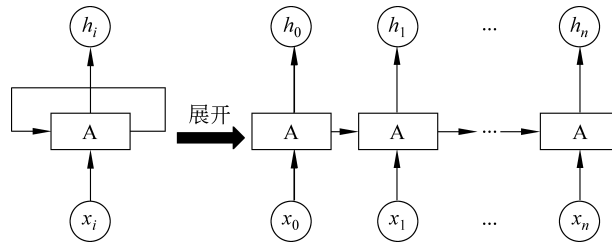


图 5-2 一个典型的循环神经网络

任务相关联是循环神经网络特别强调的关键点。循环神经网络在当前任务与前期信息距离较近时,前期信息能够得到很好的保存并通过其序列结构传输至当前任务;但是在当前任务与前期信息距离过远时,循环神经网络对这些前期信息并不能进行有效的学习。这就是著名的梯度消失问题(vanishing gradient problem),即后面的节点对于前面的节点的感知力出现下降。

针对长期依赖关系下循环神经网络无法有效学习前期信息的问题,赛普·霍克赖特(Sepp Hochreiter)等于1997年提出了长短期记忆模型(LSTM),它是循环神经网络模型的特殊形式,能够有效学习长期依赖关系。LSTM是一种针对长期依赖关系问题设计的模型结构,它无须刻意学习就可以将前期信息保存很长时间。

所有的循环神经网络模型中都包含一连串重复的神经网络模块形式的结构。在标准的循环神经网络中,此类重复模块通常是一种特别简单的结构,例如单个的 tanh 层。标准循环神经网络重复模块如图 5-3 所示。

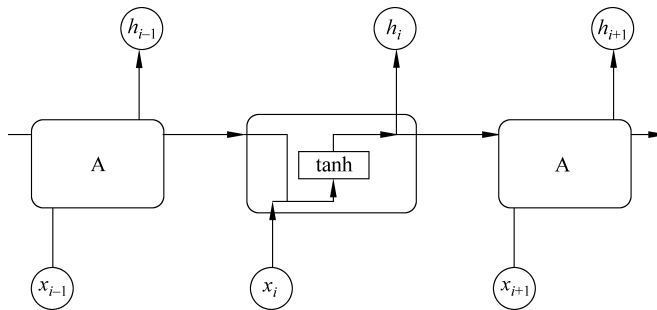


图 5-3 标准循环神经网络重复模块

LSTM 和循环神经网络一样,也是链状结构,两者的不同点在于重复模块结构不同。与传统循环神经网络的计算单元是简单的 tanh 层不同,LSTM 在其计算单元中包含了 4 个神经网络层,这些神经网络层以特殊的方式相互作用。LSTM 重复模块如图 5-4 所示。

细胞状态是 LSTM 的核心,它在水平方向上贯穿前行。LSTM 通过精心设计的门对细胞状态的改变进行控制。门是使信息选择式通过的一种方法,由一个 Sigmoid 神经网络层和一个 Pointwise 乘法操作组成。LSTM 通过 3 个门实现对细胞状态的保护和控制,这 3 个门分别是输入门、输出门以及遗忘门。LSTM 中的前向传播计算式如下:

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \quad (5-1)$$

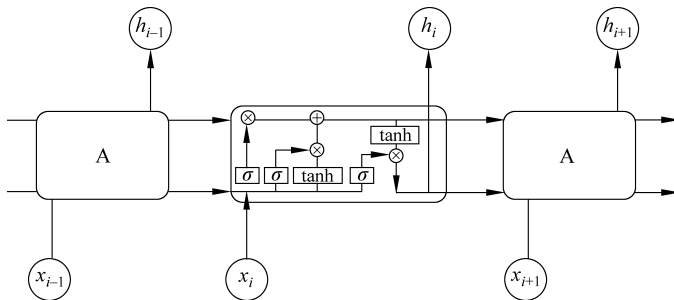


图 5-4 LSTM 重复模块

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \quad (5-2)$$

$$\tilde{C}_t = \tanh(W_c \cdot [h_{t-1}, x_t] + b_c) \quad (5-3)$$

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t \quad (5-4)$$

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \quad (5-5)$$

$$h_t = o_t * \tanh(C_t) \quad (5-6)$$

在上面的公式中, f_t 表示遗忘门, i_t 表示输入门, $\tilde{C}_t$ 表示候选值, C_t 代表新的细胞状态, o_t 表示输出门, h_t 表示输出值, x_t 表示当前输入值。

基于 LSTM 的基本思想的各种变体陆续出现,例如 2000 年提出的窥视孔连接(peephole connections)变体以及 2014 年提出的 GRU 变体等。

5.2 前馈神经网络

在前馈神经网络中,神经元根据接收信息的先后被分成不同的组,每个组可以看作一个神经网络层。神经网络层中的每个神经元接收来自上一层中的神经元的输出,并将处理结果作为输入传输到下一层的神经元。前馈神经网络以输入层开始,以输出层结束,其他的层为隐含层。前馈神经网络中的信息只存在向前传递的情况,没有反向传递,其结构如图 5-5 所示。

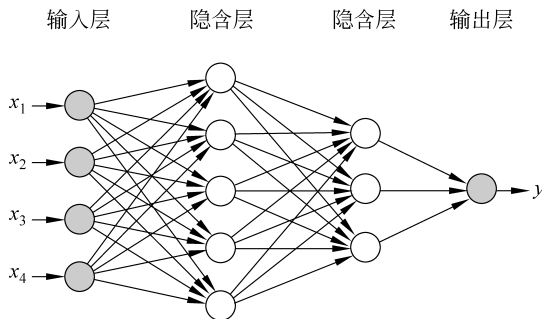


图 5-5 前馈神经网络结构

可以将前馈神经网络视作一个函数,它使用多次复合简单的非线性函数的方式完成

从输入到输出的复杂映射的任务。前馈神经网络传递信息的公式如下：

$$\mathbf{z}^{(l)} = \mathbf{W}^{(l)} \cdot \mathbf{a}^{(l-1)} + \mathbf{b}^{(l)} \quad (5-7)$$

$$\mathbf{a}^{(l)} = f_l(\mathbf{z}^{(l)}) \quad (5-8)$$

$\mathbf{W}^{(l)}$ 代表从第 $l-1$ 层到第 l 层的权重矩阵, $\mathbf{z}^{(l)}$ 代表第 l 层神经网的净输入, $\mathbf{a}^{(l)}$ 代表第 l 层的输出, $\mathbf{b}^{(l)}$ 代表第 $l-1$ 层到第 l 层的偏置, $f_l(\cdot)$ 代表第 l 层神经元的激活函数, 目前主流的激活函数有 tanh 函数、ReLU 函数以及 Sigmoid 函数等。

5.3 卷积神经网络

1962 年, 生物学家大卫·休伯尔(David Hubel)和托斯登·威塞尔(Torsten Wiesel)研究了猫的视皮层, 发现视皮层中有一系列结构复杂的细胞, 这些细胞对视觉输入空间的局部区域非常敏感, 因此它们也被称为感受野。感受野以一定的方式覆盖了整个视觉领域, 在输入空间中起着局部的作用, 可以更好地挖掘自然图像中存在的强局部空间相关性。这些被称为感受野的细胞可以划分为两类: 简单细胞和复杂细胞。根据 Hubel-Wiesel 层次模型, 视皮层的神经网络具有以下层次结构: 外侧膝状体→简单细胞→复杂细胞→低阶超复杂细胞→高阶超复杂细胞。低阶超复杂细胞与高阶超复杂细胞之间的神经网络结构类似于简单细胞与复杂细胞之间的神经网络结构。在这种层次结构中, 处于较高阶的细胞通常倾向于选择性地对更复杂的刺激模式特征作出反应, 同时较高阶的细胞也具有更大的感受野, 对刺激模式位置的移动不太敏感。1980 年, 福岛邦彦(Kunihiko Fukushima)提出了一种基于 Hubel-Wiesel 层次模型的神经认知机(neocognitron)。神经认知机由简单细胞层(S 层)和复杂细胞层(C 层)交替组成。S 层对应于 Hubel-Wiesel 层次模型中的简单细胞层或低阶超复杂细胞层, C 层对应于复杂细胞层或高阶超复杂细胞层。S 层能最大限度地对感受野中特定的边缘刺激做出反应, 并提取其输入层的局部特征; 而 C 层对来自确切位置的刺激不敏感。

卷积神经网络适用于空间数据的处理任务, 在计算机视觉领域有广泛的应用。一维卷积神经网络也常常被称作时间延迟神经网络, 能够用于一维数据的处理。卷积神经网络的设计思想受到了视觉神经科学的启发, 主要由卷积层与池化层组成。卷积层具有保持图像空间连续性的特性, 能够有效地对图像的局部特征进行提取。池化层有最大池化与平均池化等多种方式, 池化层能够使中间隐含层的维度降低, 进而达到减少接下来各层的运算量的目的。池化层还提供了旋转不变性。卷积与池化操作如图 5-6 所示。

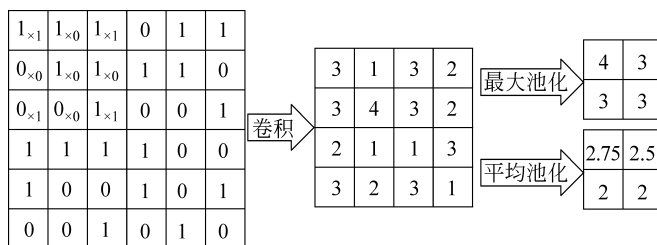


图 5-6 卷积与池化操作

图 5-6 中展示的卷积和池化操作分别是采用 3×3 的卷积核和 2×2 的池化核的情况。最早的卷积神经网络模型是 1998 年杨立昆等提出的 LeNet-5 图像分类模型,其结构如图 5-7 所示。输入的 MNIST 图片大小为 32×32 。其处理过程如下:①通过卷积核大小为 5×5 的卷积操作,得到 28×28 的特征图;②通过池化操作,得到 14×14 的特征图;③再次通过卷积操作得到 10×10 的特征图;④再次通过池化操作得到 5×5 的特征图;⑤依次使用包含 120、84、10 个神经元的全连接层对特征图进一步处理;⑥经过 Softmax 函数的作用得到数字 0~9 的概率,将概率最大的数字作为神经网络的预测结果。随着卷积和池化操作的进行,网络层越高,特征图尺寸越小,但图片维度越多。

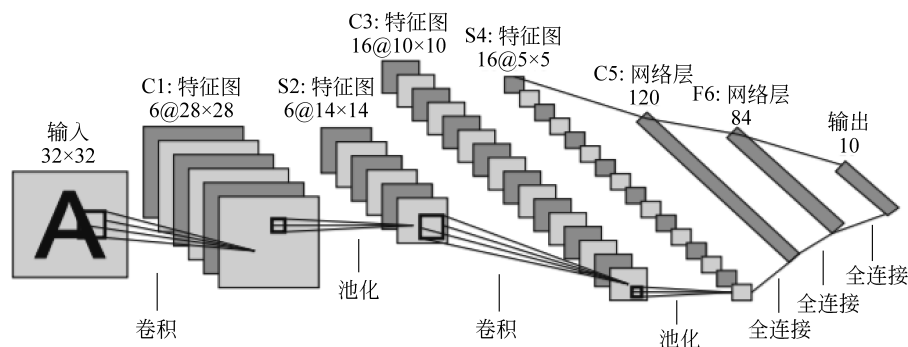


图 5-7 LeNet-5 结构

卷积神经网络对视觉数据提供了分层表示,卷积神经网络每层的权重实际上学到了图像的某些成分,卷积神经网络模型层越高,其提取的成分越具体。卷积神经网络将原始信号经过逐层处理,依次识别出部分和整体。在通过可视化手段对卷积神经网络进行可视化的过程中可以发现:卷积神经网络在第二层能提取拐角、边以及颜色等特征,在第三层能提取纹理、文字等更复杂的特征,在第四层能提取类似于狗的脸、鸟的腿等具体部位的特征,在第五层能提取类似于键盘、狗等具体物体的特征。以人脸识别为例,卷积神经网络首先对点、边、颜色以及拐角等特征进行提取,再将这些特征组合成眼角、嘴唇、鼻子,最后再完成整张脸的识别。卷积神经网络具有局部连接、权值共享以及池化操作等特性,这使得模型参数大大减少,降低了网络复杂性,并且这些特性也使卷积神经网络容易在 FPGA 等硬件上实现并获得加速。同时,卷积神经网络的这些特性也为模型提供了平移、扭曲、旋转以及缩放不变性。