深度学习综述

叶劲亨

(华南理工大学 计算机科学与工程学院计科二班, 201830582180)

摘要：深度学习是机器学习和人工智能研究的最新趋势之一。它也是当今最流行的科学研究趋势之一。深度学习方法为计算机视觉和机器学习带来了革命性的进步。新的深度学习技术正在不断诞生，超越最先进的机器学习甚至是现有的深度学习技术。本文对深度学习最新进展及未来研究方向进行了分析和总结。首先概述了多类深度学习基本模型，包括多层感知器（MLP）、卷积神经网络（CNN）和循环神经网络（RNN）。然后总结了深度学习几种常见的深度生成模型，包括受限玻尔兹曼机（RBM），深度信念网络（DBN），生成对抗网络（GAN）以及变分自编码器（VAE）。最后探讨了深度学习的主要应用领域以及目前存在的问题并给出了相应的可能解决方法，其中着重介绍了深度学习常用软件工具及平台和几种深度学习相关加速技术。

关键词：人工智能；深度学习；多层感知机；卷积神经网络；循环神经网络；受限玻尔兹曼机；生成对抗网络; 变分自动编码器

0 引言

深度学习是机器学习的一个分支或者子领域，它使用了多层次的非线性信息处理和抽象，用于有监督或无监督的特征学习、表示、分类和模式识别。大多数人认为近代的深度学习方法是从 2006 年开始发展的。在 2006 年 Geoffrey Hinton[1]提出了深度学习的概念，随后与其团队提出了深度学习模型之一，深度信念网络，并给出了一种高效的半监督算法：逐层贪心算法，来训练深度信念网络的参数，打破了长期以来深度网络难以训练的僵局。其基本思想是以受限玻尔兹曼机为基本单元搭建的信任网络, 采用了逐层初始化和整体反馈的方法, 成功克服了深层网络难以训练的弊端, 开启了深度学习的热潮。事实上，深度学习一词最初在 1986 年就被引入机器学习。20 世纪八十年代 Hopfield 将能量函数引入到神经网络中, 用非线性动力学解释循环反馈网络的运行, 同一时期, 反向传播算法(BP)给多层神经网络提出了一种可靠的学习方法, 引起了神经网络的一次高潮。但是 BP 算法存在一定的缺陷，由于梯度消失以及梯度弥散等问题，神经网络主要用来构造成浅层的模型。2009 年，Yoshua Bengio 提出了深度学习另一常用模型：堆叠自动编码器（Stacked Auto-Encoder,SAE），采用自动编码器来代替深度信念网络的基本单元：限制玻尔兹曼机，来构造深度网络。对于一个深度网络，这种逐层预训练的方法，就是层叠自编码（Stacked Auto-Encoder，SAE）。

对于常见的分类任务，一般分为以下两个阶段：逐层预训练，微调。注意到，前述的各种 SAE，本质上都是非监督学习，SAE 各层的输出都是原始数据的不同表达。对于分类任务，往往在 SAE 顶端再添加一分类层（如 Softmax 层），并结合有标注的训练数据，在误差函数的指导下，对系统的参数进行微调，以使得整个网络能够完成所需的分类任务。对于微调过程，即可以只调整分类层的参数

（此时相当于把整个 SAE 当作一个 feature extractor），也可以调整整个网络的参数（适合训练数据量比较大的情况）。直接去训练一个深层的自编码器，其实本质上就是在做深度网络的训练，由于梯度扩散等问题，这样的网络往往根本无法训练。逐层预训练就可以使得深度网络的训练成为可能。一个直观的解释是，预训练好的网络在一定程度上拟合了训练数据的结构，这使得整个网络的初始值是在一个合适的状态，便于有监督阶段加快迭代收敛。

对于大规模的深度学习模型，除了前面提到的梯度消失的问题，还存在过拟合的问题。过拟合也是机器学习的核心难题，过拟合指的是对训练数据有着过好的识别效果，这时导致模型非常复杂。这样的结果会导致对训练数据有非常好的识别效果，而对真实样本的识别效果非常差。事实上，卷积神经网络（CNN）就是卷积层(convolution layer)和池化层(max pooling layer)相对于全连接层(dense Layer)来说，减少了超参数的数量，一定程度上减少了过拟合(over fitting)的倾向。除此之外，还可以使用 DropOut、正则化(regularization)等方法减少大规模深度学习模型的过拟合倾向。

1 基本深度学习模型

1.1 多层感知器（MLP）

多层感知器（multilayer perception，MLP）,也叫前向传播网络、 深度前馈网络，是最基本的深度学习网络结构。MLP 由若干层组成，每一层包含若干个神经元。

多层感知器的前向传播如图 1-1 所示。

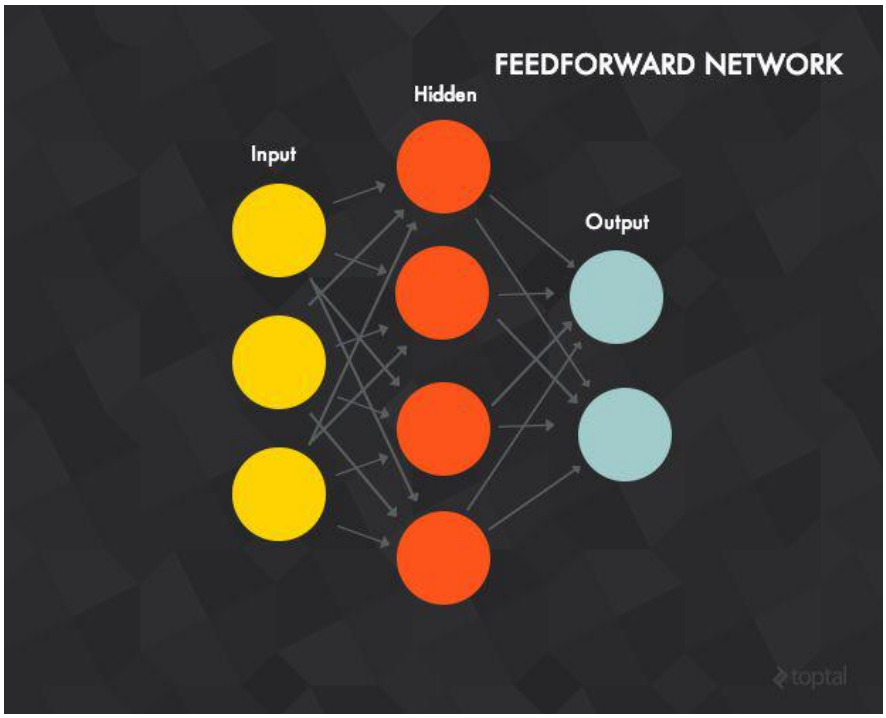

图 1-1 多层感知器模型

在图示的深度学习模型中有一个三神经元的输入层、一个四神经元的隐含层、一个二神经元的输出层，每一个神经元都是一个感知机。根据通用逼近原理，一个具有有限数目神经元的隐含层可以被训练成可逼近任意随机函数。换句话说，一层隐含层就强大到可以学习任何函数了。这说明我们在多隐含层（如深度网络）的实践中可以得到更好的结果。这里输入层的神经元作为隐含层的输入，同时隐含层的神经元也是输出层神经元的输入。每条建立在神经元之间的连接都有一个权重 w，在 t 层的每个神经元通常与前一层（ t - 1 层）中的每个神经元都有连接，也可以通过将这条连接的权重设为 0 来断开这条连接。为了处理输入数据，将输入向量赋到输入层中。这些值将被传播到隐含层，通过加权传递函数传给每

一个隐含层神经元（这就是前向传播），隐含层神经元再计算输出（激活函数）。输出层和隐含层一样进行计算，输出层的计算结果就是整个神经网络的输出。

如果每一个感知机都只能使用一个线性激活函数，整个网络的最终输出也仍然是将输入数据通过一些线性函数计算过一遍，只是用一些在网络中收集的不同权值调整了一下。换句话说，再多线性函数的组合还是线性函数。为了得到复杂的非线性模型，大多数神经网络都是使用的非线性激活函数，如对数函数(sigmoid)、双曲正切函数(tanh)、阶跃函数、Relu 函数等。

1.2 卷积神经网络（CNN）

1.2.1 LeNet 网络模型

Lecun[2]把神经认知机的精华提取出来然后加上其 1986 提出的 BP 算法，在其发表的论文《Backpropagation Applied to Handwritten Zip Code》中提出了第一个 CNN 的实现网络。普遍认为关于神经认知机是在 1980 年被福岛邦彦[3]在《 1980-Fukushima-Neocognitron A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position》一文中提出的。

Lecun 提出的 LeNet 的结构如图 1-2 所示。在该深度学习模型中，输入的 MNIST 图片大小为 32*32，经过卷积操作，卷积核大小为 5*5，得到 28*28 的图片， 经过池化操作，得到 14*14 的图片，然后再卷积再池化，最后得到 5*5 的图片。接着依次有 120、84、10 个神经元的全连接层，最后经过 softmax 函数作用，得到数字 0~9 的概率，取概率最大的作为神经网络的预测结果。LeNet 模型的网络架

构已经和目前我们经常看到的 CNN 结构别无二致。该结构中涉及到的卷积层，池化层，全连接层都是现代 CNN 网络的基本组件。卷积层能够保持图像的空间连续性，能将图像的局部特征提取出来．池化层可以采用最大池化（max-pooling）或平均池化（mean-pooling），池化层能降低中间隐藏层的维度，减少接下来各层的运算量，并提供了旋转不变性。

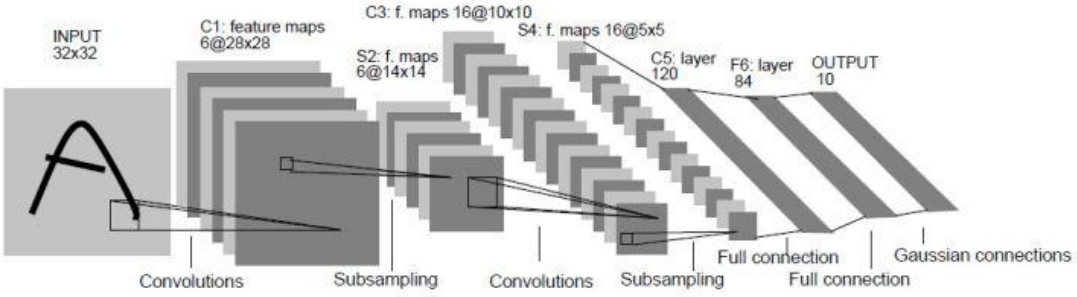

图 1-2 LeNet5 网络模型

1.2.2 AlexNet 网络模型

AlexNet [4]是具有历史意义的一个网络结构，在 AlexNet 之前，深度学习已经沉寂了很久。2012 年，AlexNet 在当年的 ImageNet 图像分类竞赛中，top-5 错误率比上一年的冠军下降了十个百分点，而且远远超过当年的第二名。AlexNet 优势在于网络增大（5 个卷积层+3 个全连接层+1 个 Softmax 层），同时解决过拟合（dropout，data augmentation，LRN），并且利用多 GPU 加速计算，其具体的架构细节如图 1-3 所示。

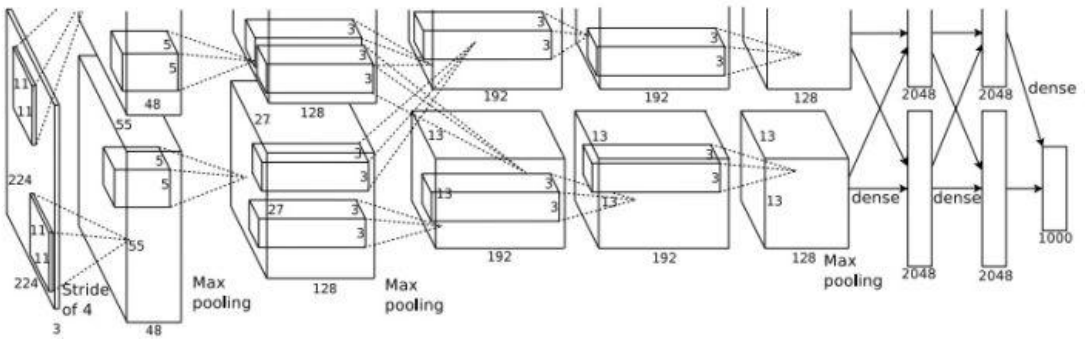

图 1-3 AlexNet 网络模型

1.2.3 其他 CNN 网络架构

ImageNet 比赛极大促进了卷积神经网络的发展,不断有新发明的卷积神经网络刷新 ImageNet 成绩. 从 2012 年的 AlexNet，到 2013 年的 ZFNet，2014 年的 VGGNet、GoogleLeNet，再到 2015 年的 ResNet，网络层数不断增加，模型能力也不断增强．AlexNet 第一次展现了深度学习的强大能力，ZFNet 是可视化理解卷积神经网络的结果，VGGNet 表明网络深度能显著提高深度学习的效果，GoogleLeNet 第一次打破了卷积层、池化层堆叠的模式，ResNet 首次成功训练了深度达到 152 层的神经网络。图 1-4 展示了卷积神经网络的发展情况，可以发现，自从 AlexNet 网络提出后，各种卷积神经网络的架构呈现爆发式增长，该种类型的深度学习模型主要被用于计算机视觉的相关领域。

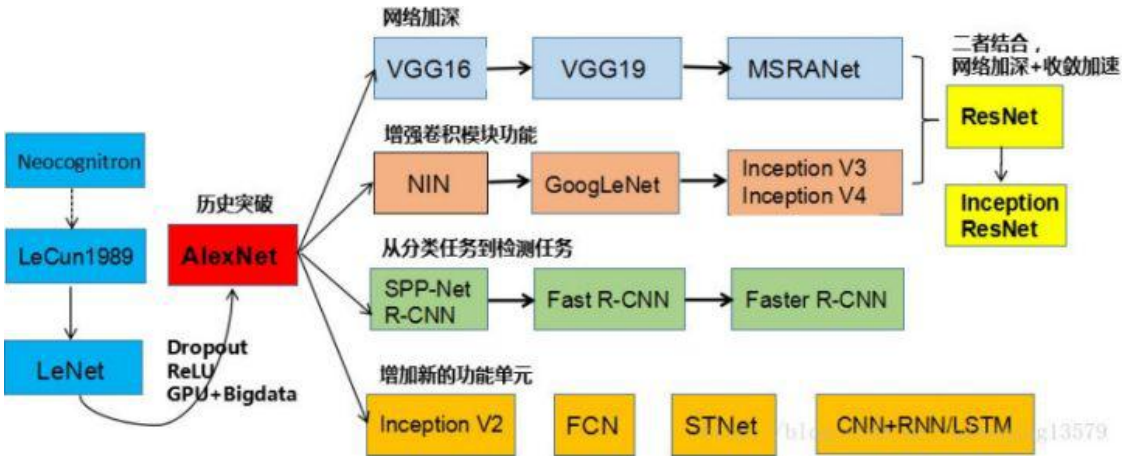

图 1-4 卷积神经网络发展史

1.3 循环神经网络（RNN）

1.3.1 基本 RNN 模型

循环神经网络（Recurrent Neural Network，RNN）适合处理时序数据。由于人类的语音和语言天生具有时序性，所以 RNN 深度学习模型在语音处理、自然语言处理领域应用广泛。

约翰·霍普菲尔德(John Hopfield)在 1982 提出了 Hopfield 网络，是最早的递归神经网络（Recurrent Neural Network，RNN）。循环神经网络的参数学习可以通过随时间反向传播算法来学习。随时间反向传播算法即按照时间的逆序将错误信息一步步地往前传递。

简单 RNN 有三层：输入层、循环隐藏层和输出层，如图 1-5（a）所示。输入层中有 N 个输入单元。该层的输入是一系列沿时间 t 的向量

$\{x_1, x_2, ..., x_{t-1}, x_t, x_{t+1}, ...\}$，其中 $x_t = (x_1, x_2, ..., x_N)$。全连接 RNN 中的输入单元与隐藏层中的隐藏单元连接，该连接由权重矩阵 W_{ih} 定义。隐藏层有 M 个隐藏单元 $h_t = (h_1, h_2, ..., h_M)$，它们通过网络定义的循环结构沿时间彼此连接，如图 1-5（b）。

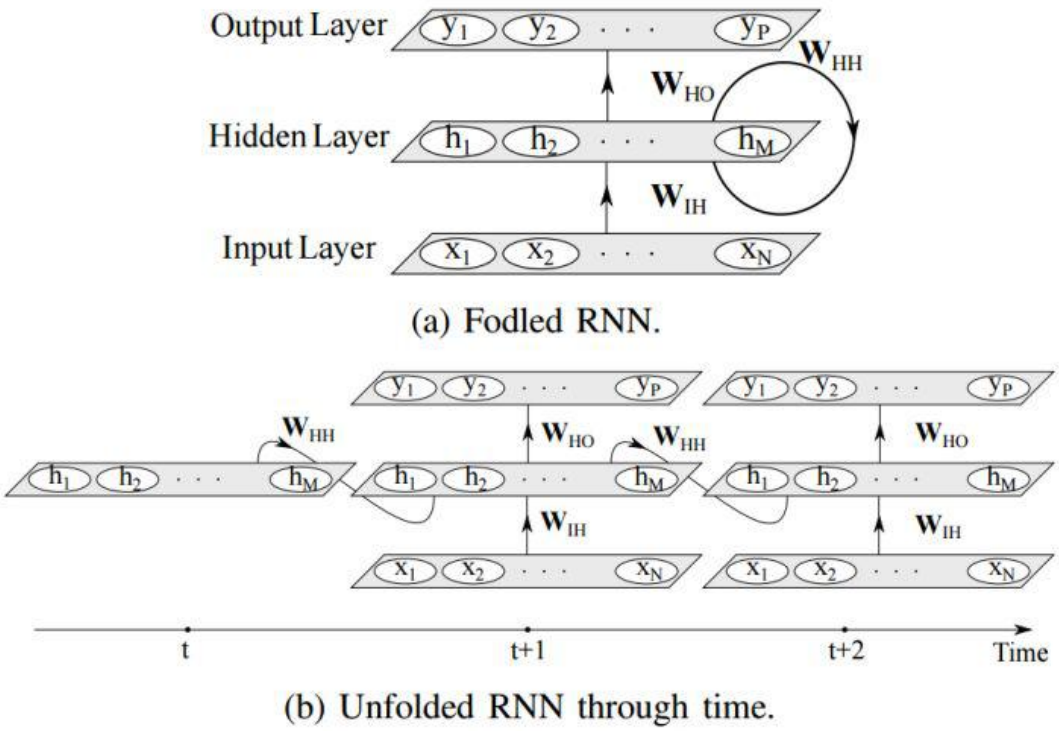

图 1-5 RNN 模型

根据 RNN 前向传播算法有：

$$\mathbf{h}_t = f_H(\mathbf{o}_t), \tag{1}$$

where

$$\mathbf{o}_t = \mathbf{W}_{IH}\mathbf{x}_t + \mathbf{W}_{HH}\mathbf{h}_{t-1} + \mathbf{b}_h, \tag{2}$$

其中 $f_H(\bullet)$ 是隐藏层激活函数，b_h 是隐藏单元的偏置向量。隐藏单元与输出层连接，连接权重为 W_{HO}。输出层有 P 个单元 $y_t=(y_1,y_2,..,y_p)$，可以计算为：

$$\mathbf{y}_t = f_O(\mathbf{W}_{HO}\mathbf{h}_t + \mathbf{b}_o) \qquad (3)$$

其中 fO(・) 是激活函数，b_o 是输出层的偏置向量。

1.3.2 LSTM 循环神经网络模型

对于标准的 RNN 模型，当输入序列比较长时，会存在梯度爆炸和消失问题，也称为长期依赖问题。为了解决这个问题，人们对循环神经网络进行了很多的改进，其中最有效的改进方式引入门控机制。1990 年，出现了简单循环网络（（Simple Recurrent Network，SRN）等新的 RNN 网络模型。1997 年，Hochreater 和 Schmidhuber[5]提出了 LSTM 的网络结构，引入 CEC 单元解决 BPTT 的梯度爆炸和消失问题，促进了循环神经网络的发展，特别是在深度学习广泛应用的今天，RNN（LSTM）在自然语言处理领域，如机器翻译、情感分析、智能对话等，取得了令人惊异的成绩。

图 1-6 给出了 LSTM 的网络架构。由图可见，LSTM 模型有两个隐藏状态 h(t)，C(t)。模型参数几乎是 RNN 的 4 倍，因为现在多了 Wf，Uf，bf，Wa。Ua，ba，Wi，Ui，bi，Wo，Uo，bo，Wf，Uf，bf，Wa，Ua，ba，Wi，Ui，bi，Wo，Uo，bo 这些参数。

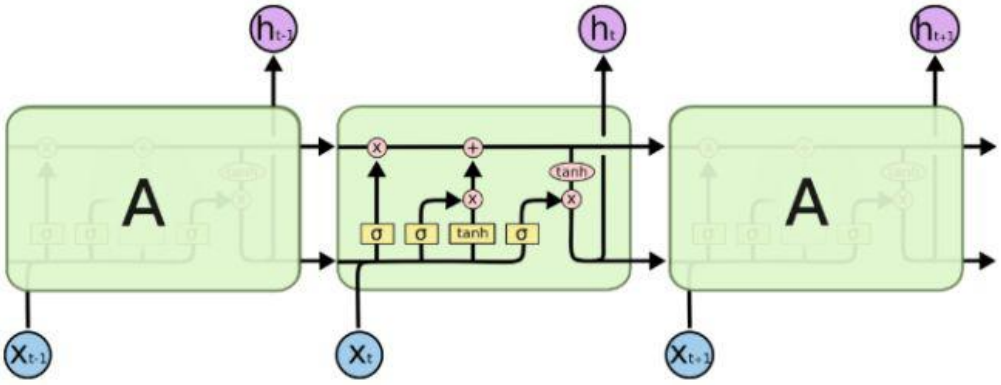

图 1-6 LSTM 循环神经网络模型

LSTM 循序神经网络的前向传播过程在每个序列索引位置的过程为：

1）更新遗忘门输出：

$$f(t)=\sigma(W_f h^{(t+1)}+U_f x^{(t)}+b_f)$$

2）更新输入门两部分输出：

$$i^{(t)}=\sigma(W_i h^{(t-1)}+U_i x^{(t)}+b_i)$$

$$a^{(t)}=tanh(W_a h^{(t-1)}+U_a x^{(t)}+b_a)$$

3）更新细胞状态：

$$C^{(t)}=C^{(t-1)}\odot f^{(t)}+i^{(t)}\odot a^{(t)}$$

4）更新输出门输出

$$o^{(t)}=\sigma(W_o h^{(t-1)}+U_o x^{(t)}+b_o)$$

$$h(t)=o^{(t)}\odot tanh(C^{(t)})$$

5）更新当前序列索引预测输出：

$$y^{(t)}=\sigma(Vh^{(t)}+c)$$

1.3.3 GRU 循环神经网络模型

GRU（Gate Recurrent Unit）[6]也是循环神经网络（Recurrent Neural Network, RNN）的一种变体。和 LSTM（Long-Short Term Memory）一样，也是为了解决长期记忆和反向传播中的梯度等问题而提出来的。相比 LSTM，使用 GRU 能够达到相当的效果，由于与 LSTM 相比，GRU 内部少了一个 “门控”，参数比 LSTM 少，所以相比之下 GRU 模型更容易进行训练，能够很大程度上提高训练效率，因此很多时候会更倾向于使用 GRU。GRU 输入输出的结构与普通的 RNN 相似，其中的内部思想与 LSTM 相似。

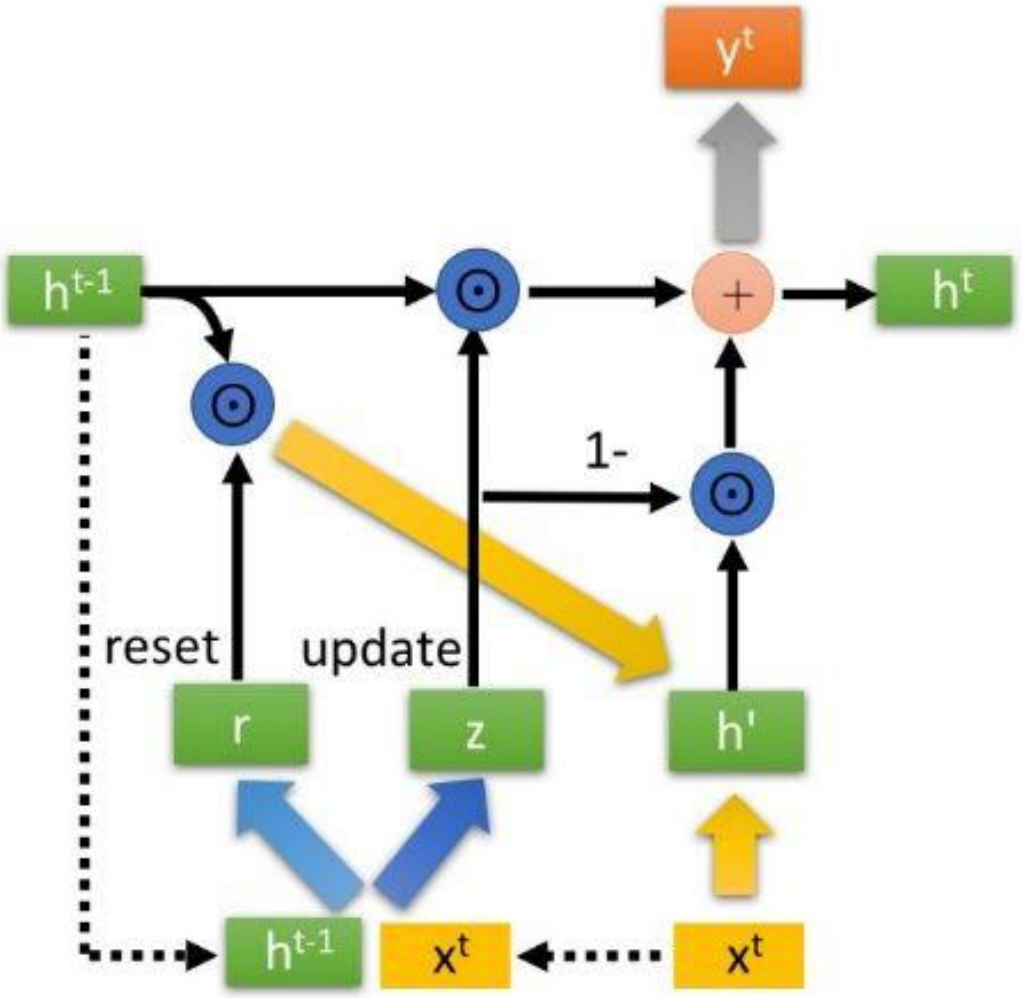

图 1-7 GRU 循环神经网络内部结构

1.3.4 其他 RNN 模型

循环神经网络架构还有很多的其他的变种[7]，包括深度 RNN 结合多层感知机，双向 RNN，循环卷积神经网络，多维循环神经网络，记忆网络，结构受限循环神经网络，酉循环神经网络，门控正交循环单元，层级子采样循环神经网络等形式。

2 深度生成模型

生成模型是根据一些可观测的样本 x (1) , x (2) , ・ ・ ・ , x (N) 来学习一个参数化的模型 p θ (x)来近似未知分布 pr(X)，并可以用这个模型来生成一些样本，使得“生成”的样本和“真实”的样本尽可能地相似。深度生成模型就是利用深层神经网络可以近似任意函数的能力来建模一个复杂的分布 pr(x)。目前使用的比较多的深度生成模型[8]包括深度玻尔兹曼机（DBM），深度信念网络（DBN），生成对抗网络（GAN）以及变分自编码器（VAE）。其中，受限玻尔兹曼机和自动编码器也是深度学习模型中用于非监督学习的两条主线。总体来说, [9]采用 RBM 构建的深度学习系统如 DBN 效果要比自编码构建的栈式自编码好, 这主要是因为自编码只能通过重构平均误差的方法来进行逼近, 而 RBM 则可以通过极大化似然估计来逼近真实的联合概率分布。但是如果用降噪自编码构成栈式

降噪自编码深度学习系统，则其学习能力与 DBN 相当，这主要是由于降噪自编码可以随机产生一些输入，增强了自编码的抗噪性，提升了泛化能力。

近期， Petuum 和 CMU[10]合作提出深度生成模型的统一框架。该框架在理论上揭示了近来流行的 GAN、VAE（及大量变体）与经典的贝叶斯变分推断算法、wake-sleep 算法之间的内在联系；为广阔的深度生成模型领域提供了一个统一的视角。

2.1 受限玻尔兹曼机模型（RBM）

玻尔兹曼机（boltzmann machine,BM）是一种随机的递归神经网络，由 G.E.Hinton 等[11,12,13]提出，是能通过学习数据固有内在表示、解决复杂学习问题最早的人工神经网络之一，受限玻尔兹曼机（restricted boltzmann machine,RBM）是玻尔兹曼机的扩展，由 Hinton 等提出，由于去掉了玻尔兹曼机同层之间的连接，因而大大提高了学习效率。RBM 本身模型很简单，只是一个两层的神经网络，因此严格意义上不能算深度学习的范畴。在深度学习模型中，更多的是对 RBM 进行堆叠。将 RBM 堆叠起来就得到了深度玻尔兹曼机模型（Deep Boltzmann Machines，DBM），再加一个分类器就得到了深度信念网络（Deep Belief Machine，DBN）。RBM 模型及其推广在工业界比如推荐系统中得到了广泛的应用。RBM 的网络架构如图 2-1 所示。

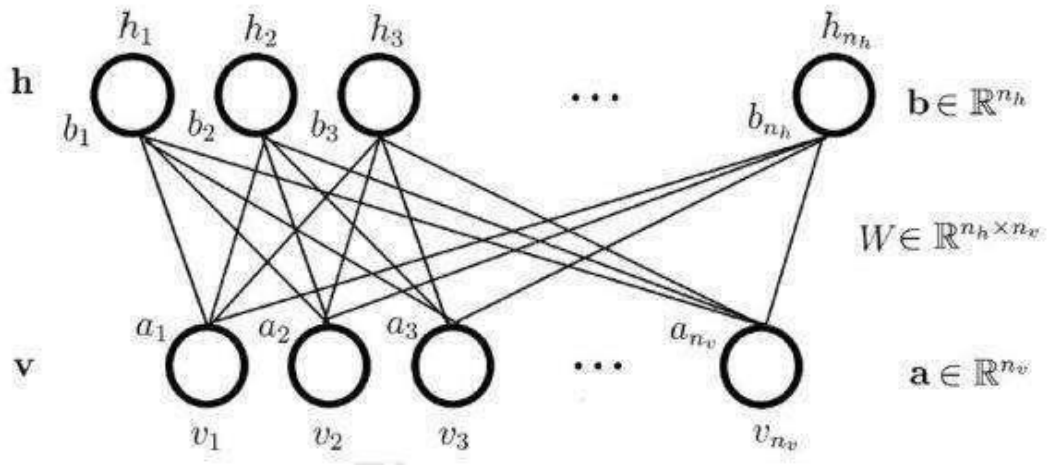

图 2-1 受限玻尔兹曼机网络架构

RBM 是一个两层的神经网络。上面一层神经元组成隐藏层(hidden layer), 用 h 向量隐藏层神经元的值。下面一层的神经元组成可见层(visible layer), 用 v 向量表示可见层神经元的值。隐藏层和可见层之间是全连接的，这点和 DNN 类似, 隐藏层神经元之间是独立的，可见层神经元之间也是独立的。连接权重可以用矩阵 W 表示。和 DNN 的区别是，RBM 不区分前向和反向，可见层的状态可以作用于隐藏层，而隐藏层的状态也可以作用于可见层。隐藏层的偏倚系数是向量 b, 而可见层的偏倚系数是向量 a。RBM 是基于能量的概率分布模型。对于给定的状态向量 h 和 v，则 RBM 当前的能量函数可以表示为：

$$E(v,h) = -a^T v - b^T h - h^T W v$$

定义 RBM 的状态为给定 **v, h** 的概率分布为：

$$P(v,h) = \frac{1}{Z} e^{-E(v,h)}$$

$$Z = \sum_{v,h} e^{-E(v,h)}$$

基于 RNM 的联合概率分布，条件分布 P(h | v)为:

$$P(h|\mathrm{v}) = \frac{1}{Z'} \prod_{j=1}^{n_h} exp\{b_j^T h_j + h_j^T W_{j,:} v\}$$

$$\frac{1}{Z'} = \frac{1}{P(v)} \frac{1}{Z} exp\{a^T v\}$$

基于条件概率分布，得到 RBM 的激活函数：

$$P(v_j = 1 \mid h) = sigmoid(a_j + W_{:,j}^T h)$$

梯度下降法可以从理论上解决 RBM 的优化，但是在实际应用中，由于概率分布的计算量大，因为概率分布有 $2^{n_v+n_h}$ 种情况，所以往往不直接按上面的梯度公式去求所有样本的梯度和，而是用基于 MCMC 的方法来模拟计算求解每个样本的梯度损失再求梯度和，常用的方法是基于 Gibbs 采样的对比散度方法来求解。

如果用传统的基于 Gibbs 采样的方法求解，则迭代次数较多效率很低，为了克服这一问题，Hinton[12]提出了一种称为对比分歧（contrastive divergence,CD）的快速算法。而在文献[14]中提出了一种基于随机梯度下降法的更高效的优化算法。和稀疏编码等模型相比，RBM 模型具有一个非常好的优点，即它的推断很快，只需要一个简单的前向编码操作，即 $\mathrm{h} = sigmoid(W \cdot V + b_{\mathrm{h}})$。

一些研究者在 RBM 基础上提出了很多扩展模型。原始的 RBM 模型中可视层为二值变量，文献[15]中通过引入高斯核使得 RBM 支持连续变量作为输入信号。一些拓展模型修改了 RBM 的结构和概率分布模型，使得它能模拟更加复杂的概率分布，如“mean-covariance RBM”[16]、“spike-slab RBM”[17]和门限 RBM[18]。这些模型中通常都定义了一个更加复杂的能量函数，学习和推

断的效率因此会有所下降。此外，文献[19]提出在 RBM 的生成式学习算法中融入判别式学习，使得它能更好的应用于分类等判别式任务。通过级联多个单层的 RBM 模型可构成深层的结构，即将前一层的隐含层作为当前层的可视层，网络的优化采用逐层优化的方式。文献[20]中将多层的有向 Sigmoid 置信网络与 RBM 级联，构造了一个深度信念网络（deep belief network, DBN）。文献[21]则将 RBM 模型直接级联成多层结构，提出了深度玻尔兹曼机网络。Lee 等人[22]用卷积操作对 DBN 网络进行扩展，使得模型可以直接从原始的二维图像中学习潜在的特征表示。除了基于 RBM 的深度结构外，还有其他一些层级生成式模型。Yu 等人[23]提出深度稀疏编码模型，用于学习图像像素块的潜在结构特征。Zeilier 等人[24]通过级联多个卷积稀疏编码和最大值池化层，构建了深度反卷积网络，可以直接从全局图像中学习从底层到高层的层级结构特征。

2.2 深度神经网络

神经网络技术起源于上世纪五、六十年代，当时叫感知机，是最早被设计并实现的人工神经网络，是一种二分类的线性分类模型，主要用于线性分类且分类能力十分有限。输入的特征向量通过隐含层变换达到输出层，在输出层得到分类结果。早期感知机的推动者是 Rosenblatt。但是单层感知机遇到一个严重的问题，即它对稍复杂一些的函数都无能为力（比如最为典型的“异或”操作），随着数学理论的发展，这个缺点直到 20 世纪 80 年代才被 Rumelhart、Williams、Hinton、LeCun 等人发明的多层感知机（multilayer perceptron，MLP）克服。多层感知机可以摆脱早期离散传输函数的束缚，使用 sigmoid 或 tanh 等连续函数模拟神经

元对激励的响应，在训练算法上则使用 Werbos 发明的反向传播 BP 算法。

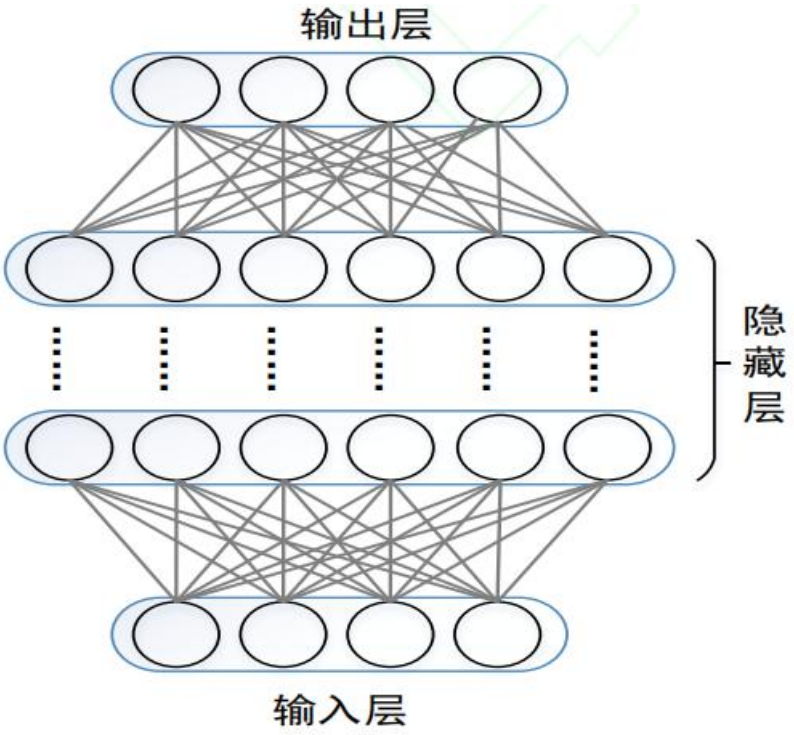

图 2-2 深度神经网络结构

图 2-2 为全连接深度神经网络结构示意图。通过增加隐含层的数量及相应的节点数，可以形成深度神经网络，深度神经网络一般指全连接的神经网络，该类神经网络模型常用于图像及语言识别等领域，在图像识别领域由于其将图像数据变成一维数据进行处理，忽略了图像的空间几何关系，因此其在图像识别领域的识别率不及卷积神经网络，且由于相邻层之间全连接，其要训练的参数规模巨大，因此巨大的参数量也进一步限制了全连接神经网络模型结构的深度和广度。

2.3 生成对抗网络模型（GAN）

生成对抗网络（GAN）提供了一种不需要大量标注训练数据就能学习深度表征的方式。它们通过反向传播算法分别更新两个网络以执行竞争性学习而达到训练目的。GAN 学习的表征可用于多种应用，包括图像合成、语义图像编辑、风格迁移、图像超分辨率技术和分类。

生成对抗网络的生成器和判别器通常由包含卷积和（或）全连接层的多层网络构成。生成器和判别器必须是可微的，但并不必要是直接可逆的（理论分析上必须可逆）。原始 GAN 的判别网络 D 可以看成是将图像数据映射到（该图像是来自真实数据分布，而不是生成器分布）判别概率的函数 $D : D(x) \rightarrow (0, 1)$。对于一个固定的生成器 G，判别器 D 可能被训练用于分辨图像是来自训练数据（真，概率接近 1）还是来自生成器（假，概率接近 0）。若判别器已经是最好的，它将变得无法被欺骗，而这时生成器 G 需要继续训练以降低判别器的准确率。如果生成器分布足以完美匹配真实数据分布，那么判别器将会被最大地迷惑而对所有输入给出 0.5 的概率值。图 2-3 给出了生成对抗网络的训练流程。

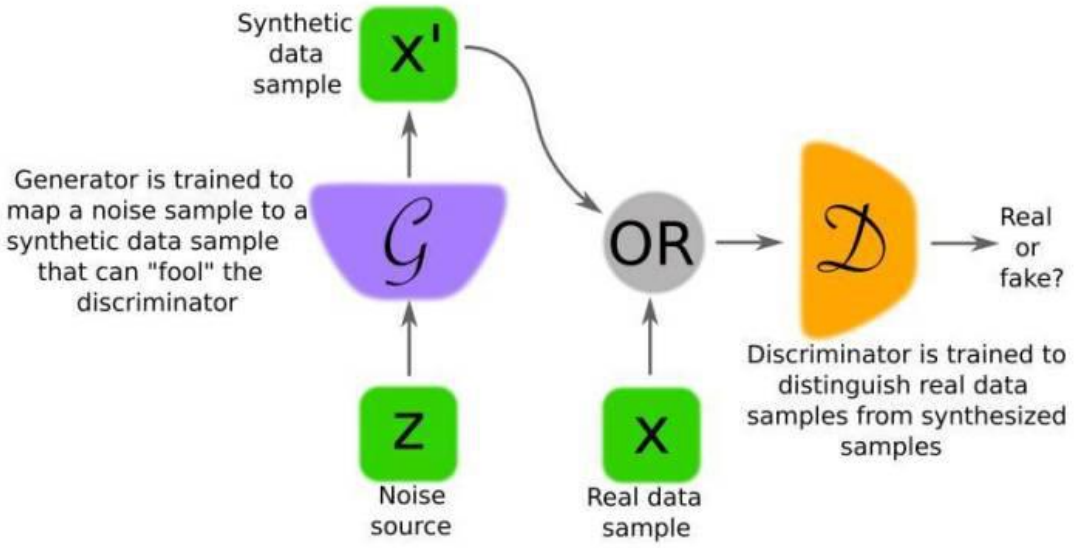

图 2-3 GAN 中的判别器 D 和生成器 G 的训练流程

GoodFellow[25]在其论文中给出了生成对抗网络中的目标公式：

$$\max_{\mathcal{D}} \min_{\mathcal{G}} V(\mathcal{G}, \mathcal{D})$$

where

$$V(\mathcal{G}, \mathcal{D}) = \mathbb{E}_{p_{data}(\mathbf{x})} \log \mathcal{D}(\mathbf{x}) + \mathbb{E}_{p_g(\mathbf{x})} \log(1 - \mathcal{D}(\mathbf{x}))$$

该优化过程是一个最大最小优化问题，对应的也就是上述的两个优化过程。可以看到，优化 D（判别网络）的时候，生成网络固定，G(z)表示已经得到的假样本。优化 D 的公式的第一项，使的真样本 x 输入的时候，得到的结果越大越好，因为需要真样本的预测结果越接近于 1。对于假样本，需要优化是的其结果越小越好，也就是 D(G(z))越小越好，也就是 1-D(G(z))越大，因为它的标签为 0。同样在优化 G 的时候，与真样本无关，这个时候只有假样本，但是我们说这个时候希望假样本的标签是 1 的，所以是 D(G(z))越大越好，即最小化 1-D(G(z))。合并上述两个优化模型，就变成了文中的最大最小目标函数了。

GoodFellow 等人的 GAN 架构在生成器与鉴别器上使用全连接神经网络，这种架构类型被应用于相对简单的图像数据库，即 MNIST（手写数字）、CIFAR-10（自然图像）和多伦多人脸数据集（TFD）。此外，Radfordetal 提出了一种称之为 DCGAN （深度卷积 GAN）的网络架构族，它允许训练一对深度卷积生成器和判别器网络。DCGAN 在训练中使用带步长的卷积（strided convolution）和小步长卷积（fractionally-strided convolution），并在训练中学习空间下采样和上采样算子。Mirza 等人通过将生成器和判别器改造成条件类（class-conditional）而将（2D）GAN 框架扩展成条件设置。条件 GNN 的优势在于可以对多形式的数据生成提供更好的表征。另外，GAN 推断模型以及对抗自编码器也是新提出的 GAN 的架构类型。

2.4 变分自动编码器模型（VAE）

与使用标准的神经网络作为回归器或分类器相比，变分自动编码器（VAEs）是强大的生成模型，它可以应用到很多领域，从生成假人脸到合成音乐等。标准自动编码器学会生成紧凑的表示和重建他们的输入，但除了能用于一些应用程序，如去噪自动编码器，他们是相当有限的。自动编码器的基本问题在于，它们将其输入转换成其编码矢量，其所在的潜在空间可能不连续，或者允许简单的插值。如果空间有不连续性（例如簇之间的间隙）并且从那里采样/产生变化，则解码器将产生不切实际的输出，因为解码器不知道如何处理该潜在空间的区域。 在训练期间，从未看到来自该潜在空间区域的编码矢量。变分自动编码器（VAEs）具有一个独特的性质，可以将它们与 vanilla 自动编码器分离开来，正是这种特

性使其在生成建模时非常有用：它们的潜在空间在设计上是连续的，允许随机采样和插值。它通过做一些约束来达到这个目的：使编码器不输出大小为 n 的编码矢量，而是输出两个大小为 n 的矢量：平均矢量 μ 和另一个标准偏差矢量 σ 。变分自动编码器的架构如图 2-4 所示。

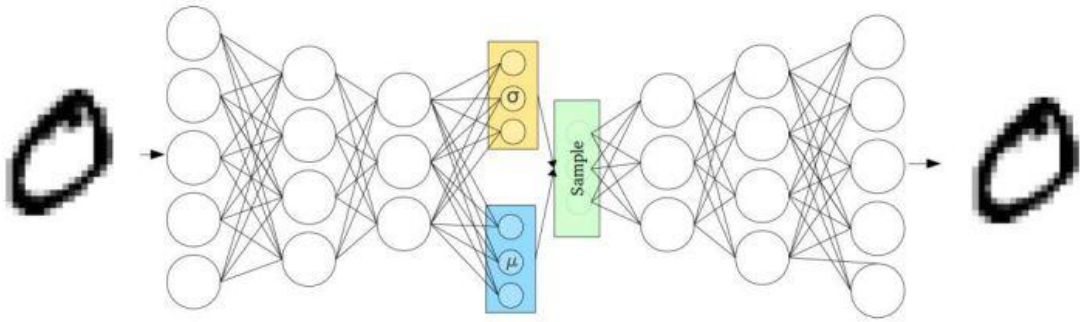

图 2-4 变分自动编码器的架构

VAE 理想的要求是所有这些编码尽可能地彼此接近，但仍然是独特的，允许平滑插值，并且能够构建新的样本。为了强制做到这一点，我们在损失函数中引入 Kullback-Leibler 散度（KL 散度）。两个概率分布之间的 KL 散度只是衡量它们相互之间有多大的分歧。 这里最小化 KL 散度意味着优化概率分布参数（ μ 和 σ ），使其与目标分布的概率分布参数非常相似。对于 VAE，KL 损失是 X 中个体 X∽N（ μ ， σ^2 ）与标准正态分布之间所有 KL 分支的总和。 当 $\mu = 0$， $\sigma = 1$ 时，最小化以下公式：

$$\sum_{i=1}^{n} \sigma_i^2 + \mu_i^2 - \log(\sigma_i^2) - 1$$

这种损失鼓励编码器将所有编码（对于所有类型的输入，例如所有 MNIST 数字号）均匀地分布在潜在空间的中心周围。 如果它试图通过把它们聚集到特定的

地区而远离原样本来“作弊”，将会受到惩罚。VAE 可以处理明显不同类型的数据，顺序或非顺序，连续或离散，标记或完全不标记，使其成为非常强大的生成工具。

2.5 小结

关于深度学习模型，非监督学习的预训练过程主要存在两条主线，一条基于 AE 和 SAE，另外一条基于受限玻尔兹曼机。在此基础上得到了深度信念网络，变分自编码器以及生成对抗网络等模型。在监督学习中，主要采用的深度学习模型有前向神经网络（FNN），卷积神经网络（CNN），循环神经网络（RNN）。

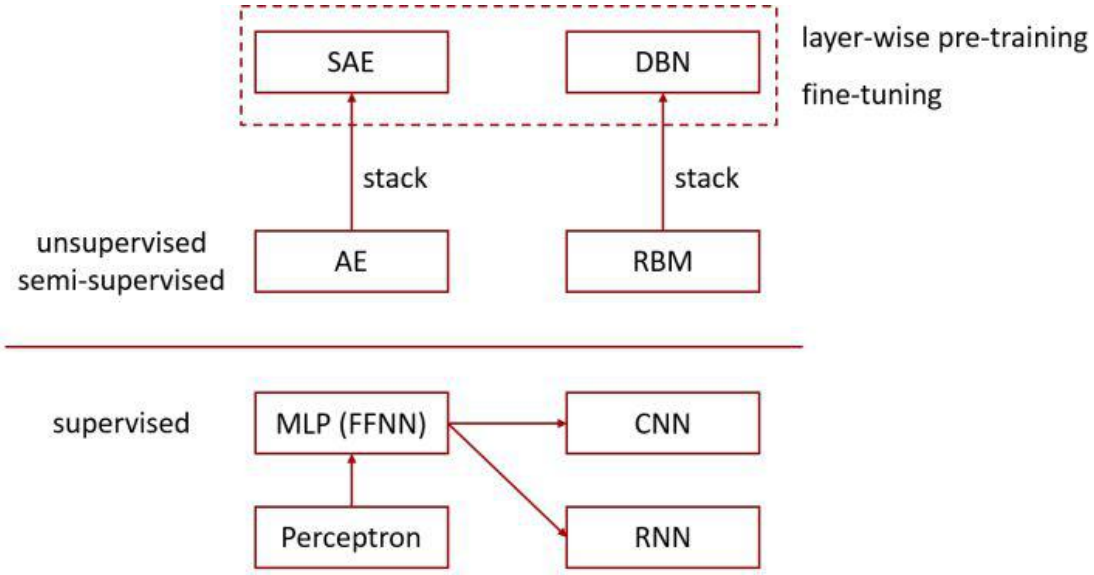

图 2-5 神经网络归类

3 基于深度学习的优化方法

随着神经网络模型层数越来越深、节点个数越来越多，需要训练的数据集越来越大，模型的复杂度也越来越高，因此在模型的实际训练中单 CPU 或单 GPU 的加速方案存在着严重的性能不足，一般需要十几天的时间才能使得模型的训练得到收敛，已远远不能满足训练大规模神经网络、开展更多实验的需求。故多 CPU 或多 GPU 的加速方案成为训练大规模神经网络模型的首选。但是由于在图像识别或语言识别类应用中，深度神经网络模型的计算量十分巨大，且模型层与层之间存在的一定的数据相关性，因此如何划分任务量以及计算资源是设计 CPU 或 GPU 集群加速框架的一个重要问题。本节主要介绍两种常用的基于 CPU 集群或 GPU 集群的大规模神经网络模型训练的常用并行方案。

3.1 数据并行

当训练的模型规模比较大时，可以通过数据并行[26]的方法来加速模型的训练，数据并行可以对训练数据做切分，同时采用多个模型实例，对多个分块的数据同时进行训练。数据并行的大致框架如图 3-1 所示：

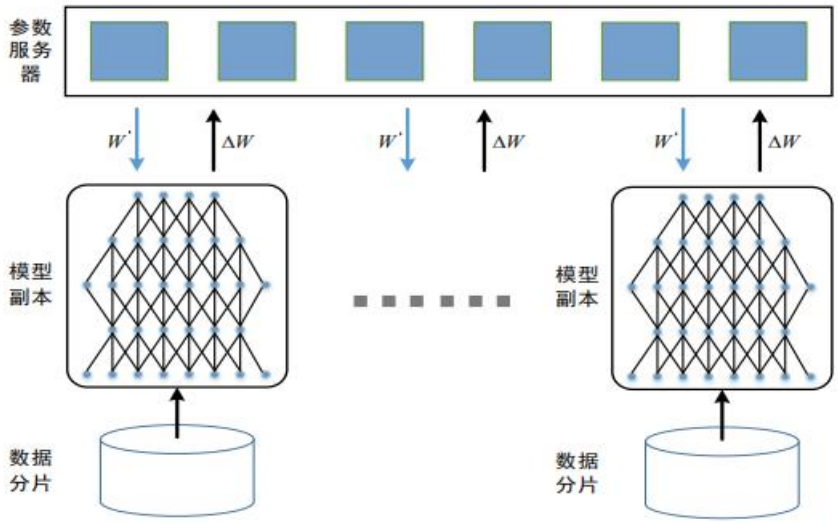

图 3-1 数据并行的基本架构

在训练过程中，由于数据并行需要进行训练参数的交换，因此通常需要一个参数服务器，多个训练过程相互独立，每个训练的结果，即模型的变化量 ΔW 需要提交给参数服务器，参数服务器负责更新最新的模型参数 $W' = W - \eta \cdot \Delta W$，之后再将最新的模型参数 W' 广播至每个训练过程，以便各个训练过程可以从同一起点开始训练。在数据并行的实现中，由于是采用同样的模型不同的数据进行训练，影响模型性能的瓶颈在于多 CPU 或多 GPU 间的参数交换，根据参数更新公式，需要将所有模型计算出的梯度提交到参数服务器并更新到相应参数上，因此数据片的划分以及与参数服务器的带宽可能会成为限制数据并行效率的瓶颈。

3.2 模型并行

除了数据并行，还可以采用模型并行的方式来加速模型的训练。模型并行是指将大的模型分拆成几个分片，由若干个训练单元分别持有，各个训练单元相互协作共同完成大模型的训练。图 3-2 为模型并行的基本框架。

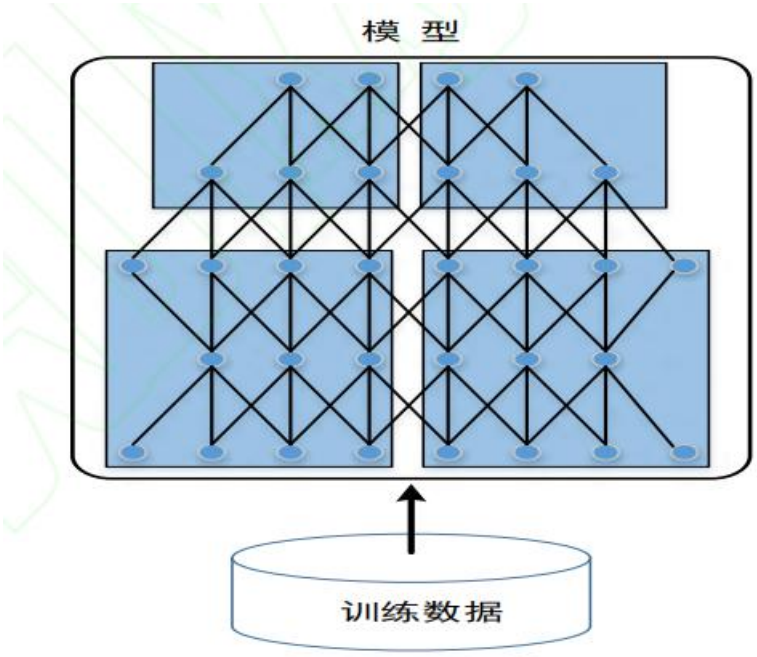

图 3-2 模型并行的基本架构

一般来说，模型并行带来的通信和同步开销多于数据并行，因此其加速比不及数据并行，但对于单机内存无法容纳的大模型来说，模型并行也是一个很好的方法，2012 年 Imagenet 冠军模型 Axlenet 就是采用两块 GPU 卡进行模型并行训练。

4 深度学习应用领域

4.1 计算机视觉

长久以来，计算机视觉就是深度学习应用中几个最活跃的研究方向之一。因为视觉是一个对人类以及许多动物毫不费力，但对计算机却充满挑战的任务。计算机视觉是一个非常广阔的发展领域，其中包括多种多样的处理图片的方式以及应用

方向。计算机视觉的应用广泛：从复现人类视觉能力(比如识别人脸) 到创造全新的视觉能力。举个后者的例子，近期一个新的计算机视觉应用是从视频中可视物体的振动识别相应的声波。大多数计算机视觉领域的深度学习研究未曾关注过这样一个奇异的应用，它扩展了图像的范围，而不是仅仅关注于人工智能中较小的核心目标--复制人类的能力。无论是报告图像中存在哪个物体，还是给图像中每个对象周围添加注释性的边框，或从图像中转录符号序列，或给图像中的每个像素标记它所属对象的标识，大多数计算机视觉中的深度学习往往用于对象识别或者某种形式的检测。由于生成模型已经是深度学习研究的指导原则，因此还有大量图像合成工作使用了深度模型。尽管图像合成(“无中生有”) 通常不包括在计算机视觉内，但是能够进行图像合成的模型通常用于图像恢复，即修复图像中的缺陷或从图像中移除对象这样的计算机视觉任务。

物体检测和图像分类是图像识别的两个核心问题，前者主要定位图像中特定物体出现的区域并判定其类别，后者则对图像整体的语义内容进行类别判定。Yang 等人[27]是传统图像识别算法中的代表，他们在 2009 年提出的采用稀疏编码来表征图像，通过大规模数据来训练支持向量机（support vector machine,SVM）进行图像分类，该方法在 2010 年和 2011 年的 ImageNet[2]图像分类竞赛中取得了最好成绩。图像识别是深度学习最早尝试的应用领域，早在 1989 年，LeCun 和他的同事就发表了关于卷积神经网络的相关工作[3]，在手写数字识别任务上取得了当时世界上最好的结果，并广泛应用于各大银行支票的手写数字识别任务中。百度在 2012 年将深度学习技术成功应用于自然图像 OCR 识别和人脸识别等问题上，并推出相应的移动搜索产品和桌面应用。从 2012 年的 ImageNet 竞赛开始，深度学习在图像识别领域发挥出巨大威力，在通用图像

分类、图像检测、光学字符识别（optical character recognition,OCR）、人脸识别等领域，最好的系统都是基于深度学习的。图 4-1 为从 2010 到 2016 年 ImageNet 竞赛的识别错误率变化及人的识别错误率。2012 年是深度学习技术第一次被应用到 ImageNet 竞赛中，可以看出相对于 2011 年传统最好的识别错误率大幅降低了 41.1%，且 2015 年基于深度学习技术的图像识别率错误率已经超过了人类，2016 年最新的 ImageNet 识别错误率已经达到 2.991%。

图 4-1 2010 至 2016 年 ImageNet 竞赛的识别错误率变化及人的识别错误率

4.2 语音识别

从 20 世纪 80 年代直到 2009~2012 年，最先进的语音识别系统是隐马尔可夫模型(hidden markov model, HMM) 和高斯混合模型(gaussian mixture model, GMM)的结合。随着更大更深的模型以及更大的数据集的出现，通过使用神经网络代替

GMM 来实现将声学特征转化为音素的过程可以大大地提高识别的精度。从2009 年开始，语音识别的研究者们将一种无监督学习的深度学习方法应用于语音识别。这种深度学习方法基于训练一个被称作是受限玻尔兹曼机的无向概率模型，从而对输入数据建模。为了完成语音识别任务，无监督的预训练被用来构造一个深度前馈网络，这个神经网络每一层都是通过训练受限玻尔兹曼机来初始化的。这些网络的输入是从一个固定规格的输入窗(以当前帧为中心) 的谱声学表示抽取，预测了当前帧所对应的 HMM 状态的条件概率。紧接着的工作则将结构从音素识别转向了大规模词汇语音识别，这不仅包含了识别音素，还包括了识别大规模词汇的序列。语音识别上的深度网络从最初的使用受限玻尔兹曼机进行预训练发展到了使用诸如整流线性单元和 Dropout 这样的技术。在大约两年的时间内，工业界大多数的语音识别产品都包含了深度神经网络，这种成功也激发了 ASR 领域对深度学习算法和结构的新一波研究浪潮，并且影响至今。完全抛弃 HMM 并转向研究端到端的深度学习语音识别系统是至今仍然活跃的另一个重要推动。

最近几年，深度学习（deep learning,DL）理论在语音识别和图像识别领域取得了令人振奋的性能提升，迅速成为了当下学术界和产业界的研究热点，为处在瓶颈期的语音等模式识别领域提供了一个强有力的工具。在语音识别领域，深度神经网络（deep neural network,DNN）模型给处在瓶颈阶段的传统的 GMM-HMM 模型带来了巨大的革新，使得语音识别的准确率又上了一个新的台阶。目前国内外知名互联网企业（谷歌、科大讯飞及百度等）的语音识别算法都采用的是 DNN 方法。2012 年 11 月，微软在中国天津的一次活动上公开演示了一个全自动的同声传译系统，讲演者用英文演讲，后台的计算机一气呵成自动完成语音识别、

英中机器翻译和中文语音合成，效果非常流畅，其后台支撑的关键技术就是深度学习。近期，百度将 Deep CNN 应用于语音识别研究，使用了 VGGNet，以及包含 Residual 连接的深层卷积神经网络（convolutional neural network,CNN）等结构，并将长短期记忆网络（long short-term memory,LSTM）和 CTC 的端到端语音识别技术相结合，使得识别错误率相对下降了 10%以上。2016 年 9 月，微软的研究者在产业标准 Switchboard 语音识别任务上，取得了产业中最低的 6.3%的词错率。以及国内科大讯飞提出的前馈型序列记忆网络（feed-forward sequential memory network,FSMN）的语音识别系统，该系统使用大量的卷积层直接对整句语音信号进行建模，更好的表达了语音的长时相关性，其效果比学术界和工业界最好的双向 RNN（recurrent neural network,RNN）语音识别系统识别率提升了 15%以上。由此可见，深度学习技术对语言识别率的提高有着不可忽略的贡献。

3.3 自然语言处理

自然语言处理(natural language processing，NLP) 是让计算机能够使用人类语言，例如英语或法语。自然语言处理（natural language processing,NLP）也是深度学习的一个重要应用领域，经过几十年多的发展，基于统计的模型已经成为 NLP 的主流，同时人工神经网络在 NLP 领域也受到了理论界的足够重视。加拿大蒙特利尔大学教授 Bengio 等在 2003 年提出用 embedding 的方法将词映射到一个矢量表示空间，然后用非线性神经网络来表示 N-Gram 模型[4]。世界上最早的深度学习用于 NLP 的研究工作诞生于 NEC Labs American，其研究员 Collobert

和 Weston[5]从 2008 年开始采用 embedding 和多层一维卷积的结构，用于词性标注、分块、命名实体识别、语义角色标注等 4 个典型 NLP 问题。值得注意的是，他们将同一个模型用于不同的任务，都取得了与现有技术水平相当的准确率。Mikolov 等通过对 Bengio 等提出的神经网络语言模型的进一步研究发现，通过添加隐藏层的多次递归，可以提高语言模型的性能[6]，语音识别任务中，在提高后续词预测准确率及总体识别错误率方面都超越了当时最好的基准系统，Schwenk 等将类似的模型用 在统计机器翻译任务中[7]，采用 BLEU （bilingual evaluation understudy,BLEU）评分机制评判，提高了 近 2 个百分点。此外，基于深度学习模型的特征学习还在语义消歧[8]、情感分析[9,10]等自然语言处理任务中均超越了当时最优。

4 深度学习目前问题及进一步研究方向

前途光明，道路曲折，尽管深度学习技术在图像处理、语音识别、自然语言处理等领域取得了突破性的进展，但是仍旧有许多问题亟待解决。

a）无监督数据的特征学习。当前，标签数据的特征学习仍然占据主导地位，而真实世界存在着海量的无标签数据，将这些无标签数据逐一添加人工标签，显然是不现实的，因此，随着深度学习技术的发展，必将越来越重视对无标签数据的特征学习，以及将无标签数据进行自动添加标签技术的研究。

b）基于模型融合的深度学习方法。相关研究表明，单一的深度学习模型往往不能带来最好的效果，而通过增加深度来提高模型效果的方法往往会有一定的局限

性，如，梯度消失问题、计算过于复杂、模型的并行性有限等问题，因此通过融合其他模型或者多种简单模型进行平均打分，可能会带来更好的效果。

c）迁移学习。迁移学习可以说是一种“站在巨人肩上”的学习方法，可以在不同领域中进行知识迁移，使得在一个领域中已有的知识得到充分的利用，不需要每次都将求解问题视为全新的问题。一个好的迁移学习方法可以大大加快模型的训练速度。

d）嵌入式设备。目前深度学习技术正往嵌入式设备靠近，即原来的训练往往在服务器或者云端，而嵌入式设备通过网络将待识别的任务上传至云端，再由云端将处理结果发送到嵌入式端，随着嵌入式设备计算能力的提升、新型存储技术以及储电技术的进步，在嵌入式端完成实时训练是完全可能的，到时就可能实现真正的人工智能。因此，嵌入式设备成为将来的研究重点，包括军/民用无人机、无人车/战车、无人潜水器等智能化装备。

e）低功耗设计。鉴于嵌入式设备对功耗非常敏感，因此具有功耗优势的 FPGA 芯片可能成为研究的一个热点，设计基于 FPGA 类似 Caffe 的可编程深度学习软件平台会是一个研究方向。

f）算法层优化。由于深度学习技术巨大的计算量和存储需求，不仅要在硬件上进行加速，算法模型优化上也可以锦上添花，如稀疏编码、层级融合、深度压缩等相关技术也会继续研究。

g）脉冲神经网络。脉冲神经网络目前虽然在精度上并不具有和机器学习算法一样的水准，一方面因为学习算法，另一方面因为信息编码，然而脉冲神经网络是更接近生物学现象和观察的模型，因此，未来在脉冲神经网络研究上的突破也是

人工智能研究上的一个重点。

h）非精确计算。鉴于神经网络模型对计算精度不是特别敏感，因此，非精确计算越来越引人瞩目，被认为是降低能耗最有效的手段之一，通过牺牲可接受的实验精度来换取明显的资源节约（能耗、关键路径延迟、面积），可以将非精确计算和硬件神经网络相结合来扩大应用范围、提高错误恢复能力和提高能源节约程度，使得该神经网络成为未来异构多核平台的热门备选加速器。

i）模型压缩。深度学习仍在不断进步，目前网络的规模开始朝着更深但是参数更少的方向发展，如微软提出的深度残差网络和 Standford 提出的稀疏神经网络，该研究体现了深度神经网络中存在参数的冗余性，可以预见未来的算法研究会进一步压缩冗余参数的存在空间，从而网络可能具有更好的精度但是却拥有更少的参数。

张荣，李伟平，莫同[28]等人认为深度学习目前主要存在的问题可以分为训练问题、落地问题、功能问题以及领域问题。事实上，关于这些存在的问题，科研人员也一直在努力攻克。

5 深度学习常用软件工具及平台

5.1 常用软件工具

当前基于深度学习的软件工具有很多，由于每种软件工具针对的侧重点不同，因此，根据需求的不同，如图像处理、自然语言处理或是金融领域等，因人而异、因项目而异采用合适的深度学习架构。本节主要介绍当下常用的深度学习软件工

具。第一类是 Tensorflow。由 Google 基于 DistBelief 进行研发的第二代人工智能系统，该平台吸取了已有平台的长处，既能让用户触碰底层数据，又具有现成的神经网络模块，可以使用户非常快速的实现建模，是一个非常优秀的跨界平台，该软件库采用数据流图模式实现数值计算，流图中的节点表示数学运算，边表示数据阵列，基于该软件库开发的平台，架构灵活，代码一次开发可无须修改即可在单机、可移动设备或服务器等设备上运行，同时可支持多 GPU/CPU 并行训练。

第二类是以 Keras 为主的深度学习抽象化平台。其本身不具有底层运算协调能力，而是依托于 TensorFlow 或 Theano 进行底层运算，Keras 提供神经网络模块抽象化和训练中的流程优化，可以让用户在快速建模的同时，具有很方便的二次开发能力，加入自己喜欢的模块。第三类是以 Caffe、Torch、MXNet、CNTK 为主的深度学习功能性平台。该类平台提供了完备的基本模块，支持快速神经网络模型的创建和训练，不足之处是用户很难接触到这些底层运算模块。第四类是 Theano，Theano 是深度学习领域最早的软件平台，专注于底层基本运算。该平台有以下几个特点：

a）集成 NumPy 的基于 Python 实现的科学计算包，可以和稀疏矩阵运算包 Scipy 配合使用，全面兼容 Numpy 库函数。

b）易于使用 GPU 进行加速，具有比 CPU 实现相对较大的加速比。

c）具有优异可靠性和速度优势。

d）可支持动态 C 程序生成。

e）拥有测试和自检单元，可方便检测和诊断多类型错误。

表 1 为当前常用的几种软件工具，可见基于深度学习的软件工具有很多，相应的编程语言也有很多，没有哪一种编程平台或语言可以一统江湖。相信未来，更新的、效率更好的编程语言或平台也可能会出现。

平台	底层语言	操作语言
TensorFlow	C++，Python	C++，Python
Keras	Python	Python
Caffe	C++	C++，Matlab，Python
Torch	C，Lua	Lua，C++
MXNet	C++，Python 等	C++，Python，Julia，Scala
CNTK	C++	C++，Python
Theano	Python，C	Python

表 1 常用软件工具的相关比较

5.2 工业界平台

随着深度学习技术的兴起，不仅在学术界，工业界如 Google、Facebook、百度、腾讯等科技类公司都实现了自己的软件平台，主要有以下几种：DistBelief 是由 Google 用 CPU 集群实现的数据并行和模型并行框架，该集群可使用上万 CPU core 训练多达 10 亿参数的深度网络模型，可用于语音识别和 2.1 万类目的的图像分类[29]。此外 Google 还采用了由图像处理器（graphics processing unit, GPU）实现的 COTS HPC 系统，也是一个模型并行和数据并行的框架，由于采用了众核 GPU，该 COTS 可以用 3 台 GPU 服务器在数天内完成对 10 亿参数的深度神经网络训练。Facebook 实现了多 GPU 训练深度卷积神经网络的并行框架，结合数据并行和模型并行的方式来训练卷积神经网络模型，使用 4 张 NVIDIA Titan GPU 可在数天内训练 ImageNet 1000 分类的网络[30]。

Paddle（ parallel asynchonous distributed deep learning,Paddle）是由国内的百度公司搭建的多机 GPU 训练平台[31]，其将数据放置于不同的机器，通过参数服务器协调各机器的训练，Paddle 平台也可以支持数据并行和模型并行。腾讯为加速深度学习模型训练也开发了并行化平台—Mariana，其包含深度神经网络训练的多 GPU 数据并行框架、深度卷积神经网络的多 GPU 模型并行和数据并行框架，以及深度神经网络的 CPU 集群框架。该平台基于特定应用的训练场景，设计定制化的并行训练平台，用于语音识别、图像识别、及在广告推荐中的应用[32]。通过对以上几种工业界平台的介绍可以发现，不管是基于 CPU 集群的 DistBelief 平台还是基于多 GPU 的 Paddle 或 Mariana 平台，针对大规模神经网络模型的训练基本上都是采用基于模型的并行方案或基于数据的并行方案，或是同时采用两种并行方案[33]。由于神经网络模型在前向传播及反向传播计算过程存在一定的数据相关性，因此当前其在大规模 CPU 集群或者 GPU 集群上训练的方法并不多。

6 深度学习相关加速技术

近年来，随着深度神经网络模型层数的增加，与之相对应的权重参数成倍的增长，从而对硬件的计算能力有着越来越高的需求，尤其在数据训练的阶段。因此，针对深度学习处理器的研究再次在工业界和学术界中崛起。目前针对数据训练阶段，被业内广泛接受的是“CPU+GPU”的异构模式和 MIC（many integrated core,MIC）众核同构来实现高性能计算。而针对数据推断阶段，则较多地依赖于

“CPU+FPGA”或“ASIC”。

6.1 CPU 加速技术

CPU 作为通用处理器，本身不用做任何改变就可以完成神经网络算法的计算，然而由于通常 CPU 的并行度低，本身的计算能力也有限。现在常用的方式是进行分布式计算，通过集合多个 CPU 从而提升计算的并行度。 CPU 作为传统的计算单元，一开始就作为深度学习的计算平台，但是由于深度学习的超大规模计算量以及高度的并行性，CPU 越来越难以适应深度学习的计算需求，只能通过多核 CPU 或者 CPU 集群进行深度学习算法的加速。2012 年 6 月，《纽约时报》披露了 Google Brain 项目，该项目由著名的斯坦福大学机器学习教授 Andrew Ng 和在大规模计算机系统方面的世界顶尖专家 JeffDean 共同主导，用 16000 个 CPU Core 的并行计算平台训练一种称为 “深度神经网络”的机器学习模型（内部约有 10 亿个节点），该训练过程进行了 7 天才能完成猫脸识别任务，

因此并行能力的缺乏是限制 CPU 加速深度学习应用的主要因素，当前基于 CPU 的多是 异构平 台 ，如 CPU+GPU 或 CPU+FPGA 的异构加速平台，复杂控制及串行部分由 CPU 执行，并行部分由 GPU 或 FPGA 执行。

6.2 GPU 加速技术

对于深度学习来说，目前硬件加速主要靠使用图形处理单元（GPU）。相比传统的通用处理器（CPU），GPU 的核心计算能力要多出几个数量级，也更容易进

行并行计算。尤其是 NVIDIA 通用并行计算框架（compute unified device architecture,CUDA），作为最主流的 GPU 编写平台，各主要的深度学习工具均用其来进行加速。GPU 的众核体系结构包含几千个流处理器，可将运算并行化执行，大幅缩短模型的运算时间。随着 NVIDIA、AMD 等公司不断推进其 GPU 的大规模并行架构支持，面向通用计算的 GPU（general-purposed GPU，GPGPU）已成为加速并行应用程序的重要手段。得益于 GPU 众核体系结构，程序在 GPU 系统上的运行速度相较于单核 CPU 往往提升几十倍乃至上千倍。目前 GPU 已经发展到了较为成熟的阶段。利用 GPU 来训练深度神经网络，可以充分发挥其数以千计计算核心的高效并行计算能力，在使用海量训练数据的场景下，所耗费的时间大幅缩短，占用的服务器也更少。如果针对适当的深度神经网络进行合理

优化，一块 GPU 卡可相当于数十甚至上百台 CPU 服务器的计算能力，因此 GPU 已经成为业界在深度学习模型训练方面的首选解决方案。

5.3 FPGA 加速技术

作为 GPU 在算法加速上强有力的竞争者，现场可编程逻辑门阵列（field programmable gate array，FPGA）近年来受到了越来越多的关注，FPGA 作为深度学习加速器具有以下几点优势：

a） 可重构。FPGA 芯片可以被重复编程，用户可以针对不同应用的计算特征定制阵列结构、计算单元、数据并行策略和存储结构。因此，FPGA 能够灵活的适应高性能计算领域的不同计算应用、算法以及模型，实现快速的更新、升级以及

调试。此外，新一代的 FPGA 芯片还具有动态可重构的能力，可以在系统不掉电和不干扰当前任务的前提下实现快速的切换。

b） 低功耗。目前主流的通用处理器在满负荷状态下的功耗大约为 60-80W，而 FPGA 的平均功耗不超过 20W，远低于 GPU 和 CPU 的功耗，低功耗是 FPGA 当前受到极大关注的重要一点。

c） 可定制。FPGA 可以根据应用需求灵活的对数据位宽进行配置，满足不同精度的计算需求，由于 FPGA 具有丰富的逻辑资源、存储资源和 DSP 资源，因此可以在一个 FPGA 芯片内部定制多种运算单元。

d） 高性能。FPGA 芯片上具有大量的片上存储资源，可以提供强大的带宽和并行访存能力。针对特定的应用定制计算通路和存储结构，同时开发粗粒度线程级并行和细粒度的指令级并行，可以最大限度为开发 FPGA 芯片提供计算和访存能力。鉴于 FPGA 的以上优势，在 15、16 年的 ISCA、Micro、NIPS 等顶会上出现了不少针对深度学习的 FPGA 加速器。而在刚刚结束的 FPGA2017 中获得最佳论文的深鉴科技 ESE 语音识别引擎，结合深度压缩（deep compression）、专用编译器以及 ESE 专用处理器架构，在中端的 FPGA 上即可取得比 Pascal Titan X GPU 高 3 倍的性能，并将功耗降低 3.5 倍。据悉，该 ESE 语音识别引擎，也是深鉴科技 RNN 处理器产品的原型。

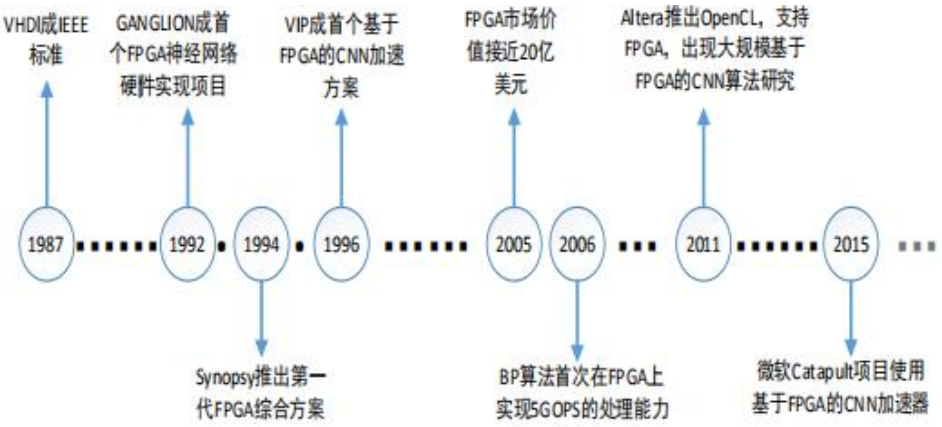

图 6-1 FPGA 深度学习研究路线图

当然 FPGA 也并非完美无瑕，同样面临一系列挑战，比如硬件编程困难，FPGA 的开发需要对底层硬件有一定的知识且使用硬件描述语言（hardware description language,HDL）进行开发，需要开发人员具有长期的经验积累，虽然已经有高级编程语言可以使用 C 或 C++进行开发，但由于其还不完善，性能还没达到硬件描述语言的程度，因此还有一定的局限性。此外 FPGA 存在许多编程模式[34]，还未形成统一的编程模型，且模块的重用也是一大难题，因此，FPGA 在深度学习的大规模应用甚至替代 GPU 还有很长的路要走。鉴于成本上的考虑，基于 FPGA 的低功耗优势，因此使用 FPGA 做深度学习加速的多是企业用户，如百度、微软、IBM 等公司都有专门做 FPGA 的团队为服务器加速，图 6-1 为基于 FPGA 的深度学习研究大致过程。

6.4 ASIC 加速技术

与 FPGA 的可编程性不同，专用集成电路（application-specific integrated

circuit,ASIC）一旦设计制造完成后电路结构就固定了，无法再改变。其主要代表公司是 Movidius。ASIC 具有以下几个特点：a）需要大量设计时间以及验证和物理设计周期，因此需要相对多的上市时间；b）同一时间点上用最好的工艺实现的 ASIC 加速器的速度会比用同样工艺 FPGA 实现的加速器速度快 5-10 倍，且量产后 ASIC 的成本会远远低于 FPGA 方案（便宜 10 到 100 倍）；因此 FPGA 主要用于服务器市场，而 ASIC 主要用于移动终端的消费电子领域。表 2 为 FPGA 和 ASIC 的相关比较。

	上市速度	性能	成本	量产成本	可配置	目标市场
FPGA	快	差	低	高	完全	企业军工
ASIC	慢	好	高	低	有限	消费电子

表 2 FPGA 和 ASIC 的相关比较

当前在专用神经网络加速器方面做得最好的当属中科院计算所的陈云霁团队，其设计的寒武纪系列神经网络加速器连续在 2013 年 ASPLOS、2014 年 MICRO、2015 年 ASPLOS、ISCA、2016 年 ISCA、MICRO 等国际顶级会议发表，并在国际上产生了重要影响，已成为国际上专用神经网络加速器的代表。2013 年提出的 DianNao[35]成为国际上首个深度学习处理器，并获得体系结构 A 类会议最佳论文，2014 年提出的 DaDianNao[36]是国际上首个多核深度学习处理器，并获得 MICRO14 最佳论文，2015 年提出的 PuDianNao[37]可以支持多种神经网络模型，成为国际上首个通用机器学习处理器，2015 年提出的 ShiDianNao[38]是一个可以嵌入在手机等终端面向视频、图像智能助理具有极低低功耗的专用神经网络处理器，其相比主流 GPU 有 28 倍的性能，4700 倍的性能功耗比。2016

年提出的 Cambricon[39]，一种神经网络指令集是国际上首个神经网络通用指令集，且获得 ISCA 评审最高分，足以见其研究的受重视程度，该通用指令

集可以高效的实现当前所有的神经网络模型，通过该指令集可以编写出不同的神经网络模型，该工作使得专用神经网络处理器具有了可编程的能力。此外，名震一时的“AlphaGo”除了配备 1920 颗 CPU 和 280 颗 GPU 外，谷歌披露它还安装一定数量的张量处理单元（tensor processing unit,TPU）。谷歌称 TPU 是专为谷歌开源项目 TensorFlow 而优化的硬件加速器，属于一款 ASIC 加速器。业内普遍认为“AlphaGo”对围棋局势的预判所使用的价值网络就是依赖 TPU 的发挥。谷歌指出，在深度学习方面，TPU 兼具了 CPU 与 ASIC 的特点，可编程，高效率，低能耗，因此 TPU 可以兼具桌面机与嵌入式设备的功能。另外，中星微“数字多媒体芯片技术”国家重点实验室宣布，中国首款嵌入式神经网络处理器（neural processing unit,NPU）芯片诞生并实现量产。这款 NPU 芯片采用了“数据驱动并行计算”架构，这种数据流类型的处理器，极大地提升了计算能力与功耗的比例，特别擅长处理视频、图像类的海量多媒体数据，使得人工智能在嵌入式机器视觉应用中可以大显身手。通过最近国际顶会的相关论文以及商业产品可知，基于专用的神经网络加速器也是当前的一个研究热点，尤其是针对嵌入式平台，如手机、无人机、无人车等。相信随着研究的进一步深入，拥有不同体系结构的专用神经网络加速器会越来越多。

5.5 其他技术研究

除了传统的硬件加速器，随着半导体技术的发展，新型的加速方案不断涌现。IBM

的 TrueNorth[55]计算平台，号称只有邮票大小，重量只有几克，但却集成了 54 亿个硅晶体管，内置了 4096 个内核，100 万个“神经元”、256 亿个“突触”，能力相当于一台超级计算机，功耗却只有 65 毫瓦。与传统冯诺依曼结构不同，芯片把数字处理器当作神经元，把内存作为突触，它的内存、CPU 和通信部件是完全集成在一起。因此信息的处理完全在本地进行，而且由于本地处理的数据量并不大，传统计算机内存与 CPU 之间的瓶颈不复存在，因此，有人把 IBM 的芯片称为是计算机史上最伟大的发明之一，将会引发技术革命，颠覆从云计算到超级计算机乃至于智能手机等一切。IBM 不久前发表于 PNAS 的论文，描述了 IBM 研究员训练卷积神经网络在神经形态硬件上分类图像和语音，在 8 个标准数据集上达到了接近目前最先进的精度，每秒 1200~2600 帧的速度处理，能耗 25~275 毫瓦。这是首次将深度学习算法的力量和神经形态处理器的高能效相结合，向着实现嵌入式类脑智能计算又迈进了一步。但是短期看来，情况并非那么乐观。首先芯片的编程困难，这种芯片要颠覆传统的编程思想，因此需要一套全新的配套开发工具，由于其相关资料尚未完全公开，因此，该芯片的能力有待进一步证实。

2017 年年初高通披露了其最新的 Snapdragon 835 的相关信息，新增加了机器学习方面的功能，包括支持客户生成神经网络层、同时还支持谷歌的机器学习架构 TensorFlow。据称 Hexagon 682 是首个支持 TensorFlow 和 Halide 架构的移动数字信号处理器（digital signal processing，DSP）。而早在 2013 年，高通就展示了一款内置 Zeroth 芯片的机器人，它能够在接受外界信息之后学会选择正确的路线行进。另外，DSP 供应商 CEVA 也于近两年在机器学习领域进行了研究，并推出了多款适应于深度学习的 DSP 芯片。另外，新型材料如忆阻器

（memristor）也被用于神经网络的构建，2016 年 Rajeev Balasubramonian 教授课题组与 HP 实验室合作，提出了一种基于忆阻器交叉开关的卷积神经网络加速器[41]，基于流水线的组织方式来加速神经网络的不同计算层，并采用 eDRAM 来实现流水线段间数据寄存。同样基于新型材料的 ReRAM 被认为是今后替代当前 DRAM 作为密度更大、功耗更小的下一代存储的技术之一。其独特的交叉网络结构和多比特存储性质，能以很高的能量效率加速神经网络计算中的主要计算模块。加州大学课题组结合 ReRAM 的这种特性，设计了一种可以在“存储”状态和“神经网络加速器”状态之间灵活切换的内存计算架构。新型材料可以融合数据存储与计算，在较低的功耗下达到很高的计算性能。然而这类芯片及硬件设计，由于受到制造工艺的影响，也存在许多限制。此外，三维堆叠技术也被引入到深度学习加速器的设计中，设计以存储为中心的总体结构，在 CPU 周围设置大量的加速器单元。Saibal Mukhopadhyay 教授课题组提出了一种基于三维堆叠存储的可编程神经网络加速器计算结构 Neurocube[41]，采用以三维堆叠存储为基础的内存计算架构，在三维堆叠内存的最下层（逻辑层）中添加计算单元，可以通过存储内部的巨大带宽，消除不必要的数据搬移；并且使用定制逻辑模块加速神经网络的计算（包括训练部分）。除了硬件结构上的加速，16 年的顶会上还提出一些算法层次上的加速，如 2016MICRO 会议上，纽约州立大学石溪分校的 Manoj Alwani 等人[58]提出一种 Fused-layer 的卷积神经网络加速器，通过融合两个或两个以上的卷积层，使得 DRAM 只用加载输入特征图，而不需要将中间结果写回，只保存计算结果，该方法可以大幅减少层与层之间的片外数据移动，进而大幅降低可移动的数据量。此外，在 MICRO2016 上，中科院计算所陈云霁等人提出了一种稀疏的神经网络加速器，通过对神经网络的分

析，找出神经网络模型相邻层之间的稀疏连接，在不降低模型识别率的基础上，将全连接网络变成稀疏连接，进而压缩神经网络模型，只计算和存储连接的神经元，因此可以大幅降低模型的计算量和存储需求。通过近两年国际顶级会议的相关论文可以发现，神经网络加速器的研究是当前的一个研究热点，不仅有基于硬件的神经网络加速也有基于软件算法层次上的神经网络加速研究，由于深度神经网络模型的超大规模计算量，因此未来需要从硬件和软件算法层次等方面一起来加速神经网络算法模型的计算。

7 结束语

深度学习作为机器学习领域的一个重要研究方向，近年来受到了越来越多的关注，鉴于深度学习研究领域的发展变化日新月异，本文系统的介绍了深度学习的相关研究现状，从深度学习的应用领域入手，重点介绍了深度学习的常用神经网络模型，分析了两种常用的深度学习模型并行训练方法，比较了两种模型训练方法的优缺点，对比分析了 7 种常用的深度学习开源软件工具的应用特点及几种工业界的研究平台，并重点介绍了当前神经网络硬件加速器的研究现状，对比分析了 CPU、GPU、FPGA、ASIC 等常用的硬件加速器，并对深度学习的未来研究方向进行了展望，可以预见，随着新型存储、光互连技术、半导体工艺等新技术、新工艺的使用，相信真正的人工智能一定能够实现。总之深度学习技术未来的发展仍然是充满不同的机遇和挑战，也是大有可为的。

参考文献：

[1] Hinton, Geoffrey E,Osindero, Simon,Teh, Yee-Whye.A fast learning algorithm for deep belief nets.[J].Neural computation,2006,18(7)

[2] LeCun, Y.,Boser, B.,Denker, J.S..Backpropagation applied to handwritten zip code recognition[J].Neural Computation,1989,1(4):541---551551.

[3] FUKUSHIMA K.NEOCOGNITRON A SELF ORGANIZING NEURAL NETWORK MODEL FOR A MECHANISM OF PATTERN RECOGNITION UNAFFECTED BY SHIFT IN POSITION[J]. Biological Cybernetics,1980,36(4):193---202.

[4] Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks[C]// International Conference on Neural Information Processing Systems. Curran Associates Inc. 2012:1097-1105.

[5] Hochreiter, S, Schmidhuber, J. Long Short-Term Memory[J]. Neural Computation, 1997, 9(8):1735-1780.

[6] Cho K, Van Merrienboer B, Gulcehre C, et al. Learning Phrase Representations

using RNN Encoder-Decoder for Statistical Machine Translation[J]. Computer Science, 2014.

[7] Salehinejad H, Sankar S, Barfett J, et al. Recent Advances in Recurrent Neural Networks[J]. 2018.

[8] Minar M R, Naher J. Recent Advances in Deep Learning: An Overview[J]. 2018.

[9] 殷瑞刚, 魏帅, 李晗,等. 深度学习中的无监督学习方法综述[J]. 计算机系统应用, 2016, 25(8):1-7.

[10] Hu Z, Yang Z, Salakhutdinov R, et al. On Unifying Deep Generative Models[J]. 2017.

[11] Hinton G E. A Practical Guide to Trainingrestricted Boltzmann machines[J]. Momentum, 2012, 9 (1): 599-619.

[12] Hinton G E. Training products of experts by minimizing contrastive divergence [J]. Neural computation, 2002, 14 (8): 1771-1800.

[13] Ackley D H, Hinton G E, Sejnowski T J. A learning algorithm for Boltzmann machi-nes [J]. Cognitive science, 1985, 9 (1): 147-169.

[14] Glorot X, Bengio Y. Understanding the difficulty of training deep feedforward neural networks [C]// Aistats. 2010, 9: 249-256.

[15] 赵元庆, 吴华. 多尺度特征和神经网络相融合的手写体数字识别 [J]. 计算机科学, 2013, 40 (8): 316-318.

[16] Hinton G E. Modeling pixel means and c-ovariances using factorized hird-order

Bol-tzmann machines [C]// Proc of Computer Vision and Pattern Recognition Conference. 2010: 2551-2558.

[17] Courville A, Bergstra J, Bengio Y. A spikeand slab restricted Boltzmann machine [C]// Proc of the 14th InternationalConference on Artificial Intelligence and Statistics. 2011: 233-241.

[18] Memisevic R, Hinton G. Unsupervised lea-rning of image transformations [C]// Proc of IEEE Conference on ComputerVision and Pattern Recognition. 2007: 1-8.

[19] Larochelle H, Bengio Y. Classification usi-ng discriminative restricted Boltzmann mac-hines [C]// Proc of the 25th internationall conference on Machine learning. 2008: 536-543.

[20] 孙志军, 薛磊, 许阳明. 基于深度学习的边际 Fisher 分析特征提取算法[J]. 电子与信息学报, 2013, 35 (4): 805-811.

[21] Salakhutdinov R, Hinton G. Deep Boltzma-nn machines [C]// Artificial Intelligence and Statistics. 2009: 448-455.

[22] Lee H, Grosse R, Ranganath R, et al. Co-nvolutional deep belief networks for scala-ble unsupervised learning of hierarchical r-epresentations [C]// Proc of the 26th Annual International Conference on Machine Learning. 2009: 609-616.

[23] Yu K, Lin Y, Lafferty J. Learning image representations from the pixel level via hi-erarchical sparse coding [C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition. 2011: 1713-1720.

[24] Zeiler M D, Taylor G W, Fergus R. Adap-tive deconvolutional networks for mid

and high level feature learning [C]// Proc of IEEE International Conference on Computer Vision. 2011: 2018-2025.

[25] Goodfellow I J, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets[C]// International Conference on Neural Information Processing Systems. MIT Press, 2014:2672-2680.

[26] Dean J, Corrado G S, Monga R, et al. L-arge scale distributed deep networks [C]// Proc of International Conference on Neural Information Processing Systems. Curran Associates Inc. , 2012: 1223-1231.

[27] Yang J, Yu K, Gong Y, et al. Linear spatial pyramid matching using sparse coding for image classification [J]. 2009: 1794-1801.

[28] 张荣，李伟平，莫同. 深度学习研究综述[J]. 信息与控制, 2018, 47(4): 385-397,410.

[29] Dean J, Corrado G S, Monga R, et al. L-arge scale distributed deep networks [C]// Proc of International Conference on Neural Information Processing Systems. Curran Associates Inc. , 2012: 1223-1231.

[30] Yadan O, Adams K, Taigman Y, et al. Multi-GPU Training of ConvNets [J]. Computer Science, 2014.

[31] Yu K. Large-scale deep learning at Baidu [C]// Proc of ACM International Conferenceon Information & Knowledge Management. 2013: 2211-2212.

[32] 腾讯公司. 深度学习在腾讯的平台化和应用实践 [EB/OL]. http://www. 36dsj. com/archives/20176.

[33] Krizhevsky A, Sutskever I, Hinton G E. I-mageNet classification with deep convolute-onal neural networks [C]// Proc of International Co-nference on Neural Information Processing Systems. Curran Associates Inc. 2012:

1097-1105.

[34] Dehon A, Adams J, Delorimier M, et al. Design patterns for reconfigurable computing [C]// Proc of IEEE Symposium on Field-Programmable Custom

Computing Machines. IEEE Computer Society, 2004: 13-23.

[35] Chen T, Du Z, Sun N, et al. DianNao: a small-footprint high-throughput accelerator-for ubiquitous machine-learning [J]. ACM Si-gplan Notices, 2014, 49 (4): 269-284.

[36] Chen Y, Luo T, Liu S, et al. DaDianNao: A Machine-Learning Supercomputer [C]// Proc of IEEE//ACM International Symposium on Micro-architecture. IEEE Computer Society, 2014: 609-622.

[37] Liu D, Chen T, Liu S, et al. PuDianNao: a polyvalent machine learning accelerator [J]. ACM SIGPLAN Notices, 2015, 43 (1): 369-381.

[38] Du Z. ShiDianNao. shifting vision process-ing closer to the sensor [C]// Proc of International Symposium on Computer Architecture. 2015: 92-104.

[39] Liu S, Du Z, Tao J, et al. Cambricon: AnInstruction Set Architecture for Neural Networks [J]. ACM SIGARCH Computer Archite-cture News, 2016, 44 (3): 393-405.

[40] Kim D, Kung J, Chai S, et al. Neurocube: a programmable digital neuromorphic

architecture with high-density 3D memory [C]// Proc of the 43rd ACM//IEEE International Symposium on Computer Architecture. 2016.

[41] Alwani M, Chen H, Ferdman M, et al. F-used-layer CNN accelerators [C]// Proc of IEEE//ACM International Symposium on Microarchitect-ure. IEEE Computer Society, 2016: 1-12.