解密最接近人腦的智能學習機器——深度學習及並行化實現

摘要： 深度學習可以完成需要高度抽象特徵的人工智慧任務，如語音識別、圖像識別和檢索、自然語言理解等。深層模型是包含多個隱藏層的人工神經網絡，多層非線性結構使其具備強大的特徵表達能力和對複雜任務建模能力。訓練深層模型是長期以來的難題，近年來以層次化、逐層初始化為代表的一系列方法的提出給訓練深層模型帶來了希望，並在多個應用領域獲得了成功。深層模型的並行化框架和訓練加速方法是深度學習走向實用的重要基石，已有多個針對不同深度模型的開源實現，Google、Facebook、百度、騰訊等公司也實現了各自的並行化框架。深度學習是目前最接近人腦的智能學習方法，深度學習引爆的這場革命，將人工智慧帶上了一個新的台階，將對一大批產品和服務產生深遠影響。

一、深度學習的革命

人工智慧(Artificial Intelligence)，試圖理解智能的實質，並製造出能以人類智能相似的方式做出反應的智能機器。如果說機器是人類手的延伸、交通工具是人類腿的延伸，那麼人工智慧就是人類大腦的延伸，甚至可以幫助人類自我進化，超越自我。人工智慧也是計算機領域最前沿和最具神秘色彩的學科，科學家希望製造出代替人類思考的智能機器，藝術家將這一題材寫進小說，搬上銀幕，引發人們無限的遐想。然而，作為一門嚴肅的學科，人工智慧在過去的半個多世紀中發展卻不算順利。過去的很多努力還是基於某些預設規則的快速搜索和推理，離真正的智能還有相當的距離，或者說距離創造像人類一樣具有抽象學習能力的機器還很遙遠。

近年來，深度學習（Deep Learning）直接嘗試解決抽象認知的難題，並取得了突破性的

進展。深度學習引爆的這場革命，將人工智慧帶上了一個新的台階，不僅學術意義巨大，而且實用性很強，工業界也開始了大規模的投入，一大批產品將從中獲益。

2006年，機器學習泰斗、多倫多大學計算機系教授Geoffery Hinton在Science發表文章[1]，提出基於深度信念網絡（Deep Belief Networks, DBN）可使用非監督的逐層貪心訓練算法，為訓練深度神經網絡帶來了希望。

2012年，Hinton又帶領學生在目前最大的圖像資料庫ImageNet上，對分類問題取得了驚人的結果[2]，將Top5錯誤率由26%大幅降低至15%。

2012年，由人工智慧和機器學習頂級學者Andrew Ng和分布式系統頂級專家Jeff Dean領銜的夢幻陣容，開始打造Google Brain項目，用包含16000個CPU核的並行計算平台訓練超過10億個神經元的深度神經網絡，在語音識別和圖像識別等領域取得了突破性的進展[3]。該系統通過分析YouTube上選取的視頻，採用無監督的方式訓練深度神經網絡，可將圖像自動聚類。在系統中輸入「cat」後，結果在沒有外界干涉的條件下，識別出了貓臉。

2012年，微軟首席研究官Rick Rashid在21世紀的計算大會上演示了一套自動同聲傳譯系統[4]，將他的英文演講實時轉換成與他音色相近、字正腔圓的中文演講。同聲傳譯需要經歷語音識別、機器翻譯、語音合成三個步驟。該系統一氣呵成，流暢的效果贏得了一致認可，深度學習則是這一系統中的關鍵技術。

2013年，Google收購了一家叫DNN Research的神經網絡初創公司，這家公司只有三個人，Geoffrey Hinton和他的兩個學生。這次收購併不涉及任何產品和服務，只是希望Hinton可以將深度學習打造為支持Google未來的核心技術。同年，紐約大學教授，深度學習專家Yann LeCun加盟Facebook，出任人工智慧實驗室主任[5]，負責深度學習的研發工作，利用深度學習探尋用戶圖片等信息中蘊含的海量信息，希望在未來能給用戶提供更智能化的產品使用體驗。

2013年，百度成立了百度研究院及下屬的深度學習研究所（IDL），將深度學習應用於語音識別和圖像識別、檢索，以及廣告CTR預估（Click-Through-Rate Prediction，pCTR），其中圖片檢索達到了國際領先水平。2014年又將Andrew Ng招致麾下，Andrew Ng是史丹福大學人工智慧實驗室主任，入選過《時代》雜誌年度全球最有影響力100人，是16位科技界的代表之一。

如果說Hinton 2006年發表在《Science》雜誌上的論文[1]只是在學術界掀起了對深度學習的研究熱潮，那麼近年來各大巨頭公司爭相跟進，將頂級人才從學術界爭搶到工業界，則標誌著深度學習真正進入了實用階段，將對一系列產品和服務產生深遠影響，成為它們背後強大的技術引擎。

目前，深度學習在幾個主要領域都獲得了突破性的進展：在語音識別領域，深度學習用深層模型替換聲學模型中的混合高斯模型（Gaussian Mixture Model, GMM），獲得了相對30%左右的錯誤率降低；在圖像識別領域，通過構造深度卷積神經網絡（CNN）[2]，將

Top5錯誤率由26%大幅降低至15%，又通過加大加深網絡結構，進一步降低到11%；在自然語言處理領域，深度學習基本獲得了與其他方法水平相當的結果，但可以免去繁瑣的特徵提取步驟。可以說到目前為止，深度學習是最接近人類大腦的智能學習方法。

二、深層模型的基本結構

深度學習採用的模型為深層神經網絡（Deep Neural Networks，DNN）模型，即包含多個隱藏層（Hidden Layer，也稱隱含層）的神經網絡（Neural Networks，NN）。深度學習利用模型中的隱藏層，通過特徵組合的方式，逐層將原始輸入轉化為淺層特徵，中層特徵，高層特徵直至最終的任務目標。

深度學習源於人工神經網絡的研究，先來回顧一下人工神經網絡。一個神經元如下圖所示[6]：

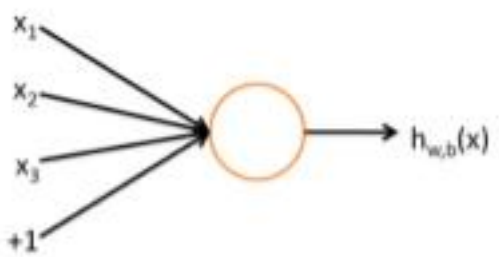

图 1　神经元结构

這個神經元接受三個輸入x1，x2，x3，神經元輸出為

其中W1, W2, W3和b為神經元的參數，f(z)稱為激活函數，一種典型的激活函數為Sigmoid函數，即3其圖像為

神經網絡則是多個神經元組成的網絡，一個簡單的神經網絡如下圖所示

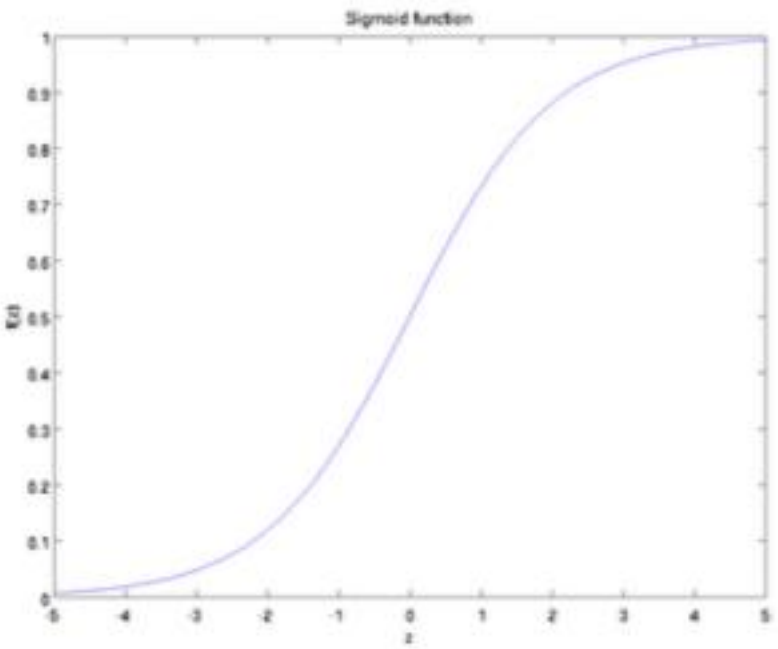

图 2　Sigmoid 函数图像

使用圓圈來表示神經網絡的輸入，標上「+1」的圓圈稱為偏置節點，也就是截距項。神經網絡最左邊的一層叫做輸入層（本例中，有3個輸入單元，偏置單元不計）；最右的一層叫做輸出層（本例中，輸出層有2個節點）；中間的節點叫做隱藏層（本例中，有2個隱藏層，分別包含3個和2個神經元，偏置單元同樣不計），因為不能在訓練樣本集中觀測到它們的值。神經元網絡中的每一條連線對應一個連接參數，連線個數對應網絡的參數個數（本例共有 4 × 3 ＋ 4 × 2 ＋ 3 × 2 ＝ 2 6 個參數）。求解這個的神經網絡，需要（x(i)，y(i)）的樣本集，其中x(i)是3維向量，y(i)是2維向量。

上圖算是一個淺層的神經網絡，下圖是一個用於語音識別的深層神經網絡。具有1個輸入層，4個隱藏層和1個輸出層，相鄰兩層的神經元全部連接。

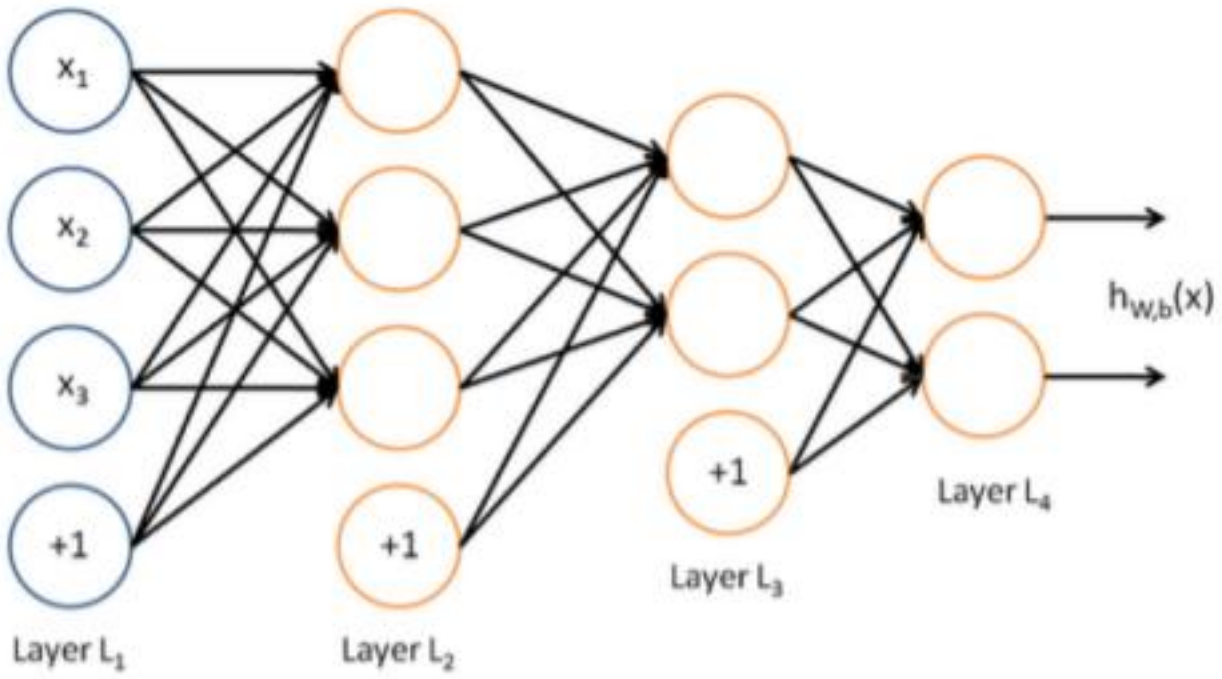

图 3 一个简单的神经网络

三、 選擇深層模型的原因

為什麼要構造包含這麼多隱藏層的深層網絡結構呢？背後有一些理論依據：

3.1天然層次化的特徵

對於很多訓練任務來說，特徵具有天然的層次結構。以語音、圖像、文本為例，層次結構大概如下表所示。

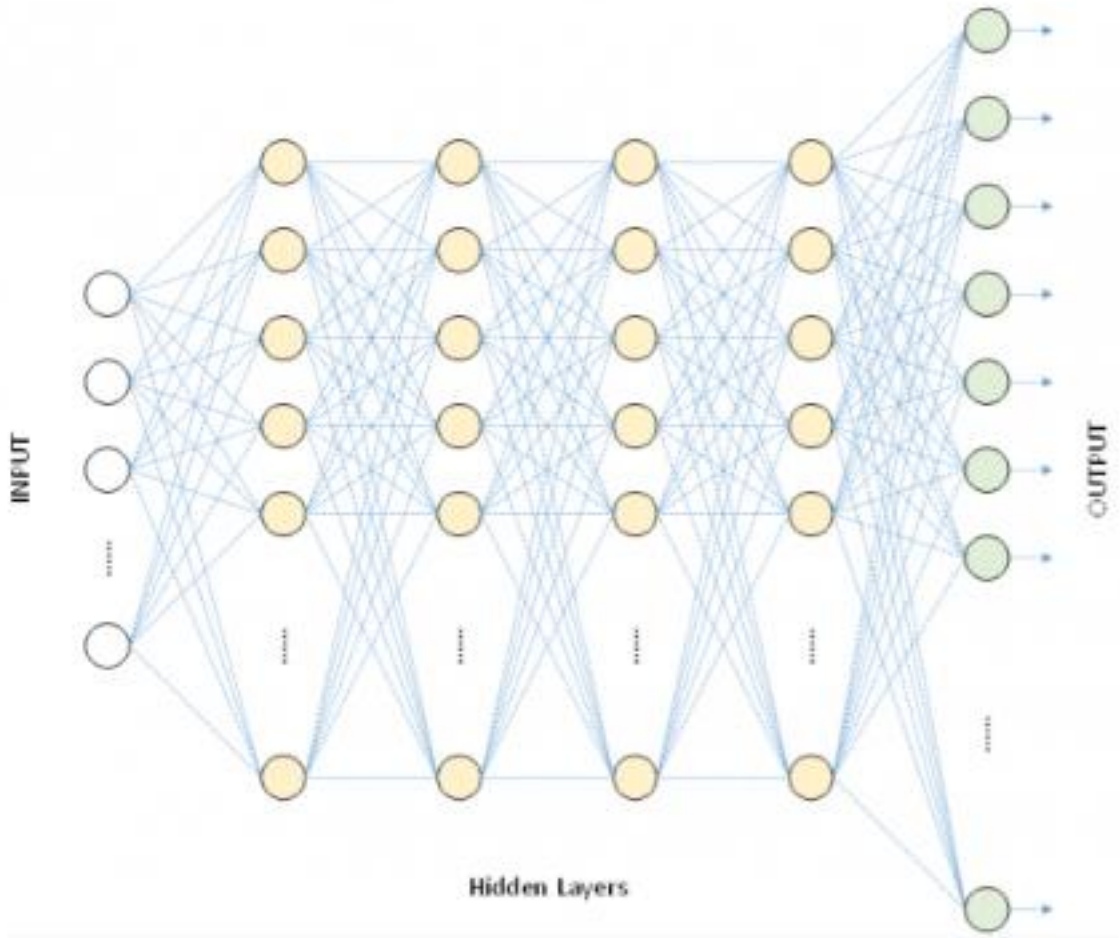

图 4 一种典型的深层神经网络模型

以圖像識別為例，圖像的原始輸入是像素，相鄰像素組成線條，多個線條組成紋理，進一步形成圖案，圖案構成了物體的局部，直至整個物體的樣子。不難發現，可以找到原始輸入和淺層特徵之間的聯繫，再通過中層特徵，一步一步獲得和高層特徵的聯繫。想要從原始輸入直接跨越到高層特徵，無疑是困難的。

表 1 几种任务领域的特征层次结构

任务领域	原始输入	浅层特征	中层特征	高层特征	训练目标
语音	样本	频段	声音	音素	单词	语音识别
图像	像素	线条	纹理	图案	局部	物体	图像识别
文本	字母	单词	词组	短语	句子	段落	文章	语义理解

3.2 仿生學依據

人工神經網絡本身就是對人類神經系統的模擬，這種模擬具有仿生學的依據。1981年，David Hubel 和Torsten Wiesel發現可視皮層是分層的[8]。人類的視覺系統包含了不同的視覺神經元，這些神經元與瞳孔所受的刺激（系統輸入）之間存在著某種對應關係（神經元之間的連接參數），即受到某種刺激後（對於給定的輸入），某些神經元就會活躍（被激活）。這證實了人類神經系統和大腦的工作其實是不斷將低級抽象傳導為高級抽象的過程，高層特徵是低層特徵的組合，越到高層特徵就越抽象。

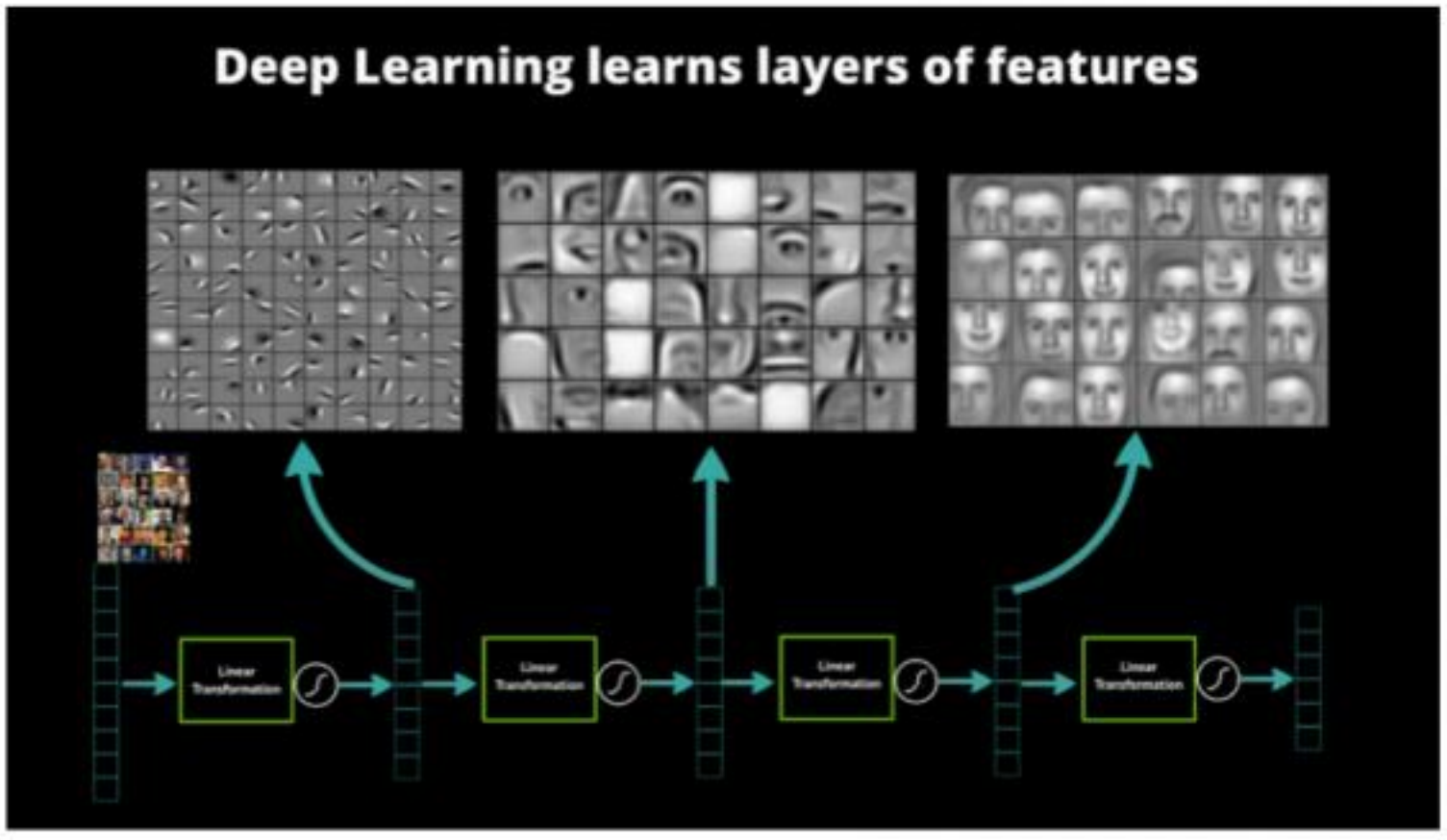

图 5 人脸识别系统的多层结构和特征表示 [7]

3.3 特徵的層次可表示性

特徵的層次可表示性也得到了證實。1995年前後，Bruno Olshausen和David Field[9]收集了很多黑白風景照，從這些照片中找到了400個16×16的基本碎片，然後從照片中再找到其他一些同樣大小的碎片，希望將其他碎片表示為這400個基本碎片的線性組合，並使誤差儘可能小，使用的碎片儘可能少。表示完成後，再固定其他碎片，選擇更合適的基本碎片組合優化近似結果。反覆疊代後，得到了可以表示其他碎片的最佳的基本碎片組合。他們發現，這些基本碎片組合都是不同物體不同方向的邊緣線。

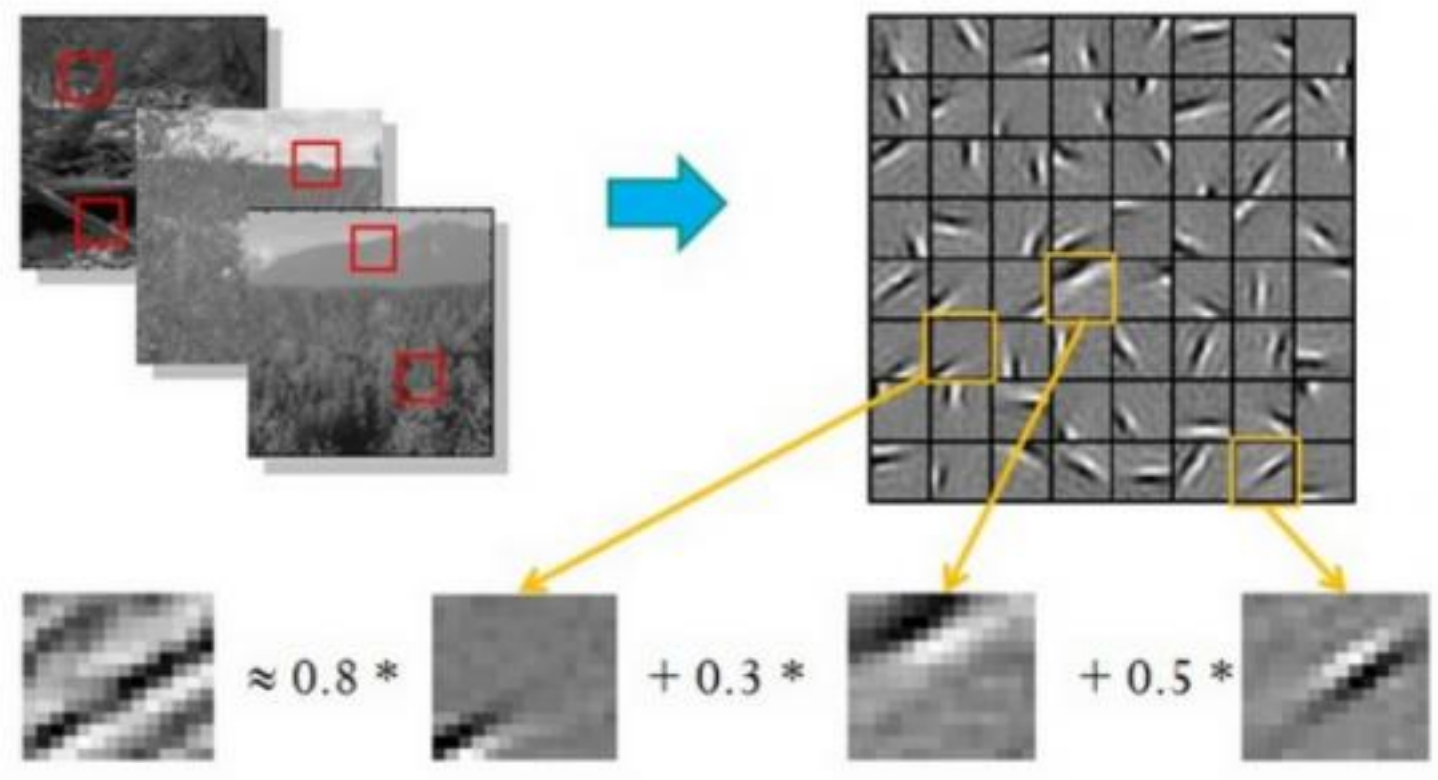

图 6 初级图像特征的提取和表示（Sparse Coding）（原图由 Andrew Ng 提供）

這說明可以通過有效的特徵提取，將像素抽象成更高級的特徵。類似的結果也適用於語音特徵。

四、 從淺層模型到深層模型

前文談到了深層模型的結構和它的優勢。事實上，深層模型具有強大的表達能力，並可以像人類一樣有效提取高級特徵，並不是新的發現。那麼為什麼深層模型直到最近幾年才開始得到廣泛的關注和應用呢？還是從傳統的機器學習方法和淺層學習談起。

4.1淺層模型及訓練方法

反向傳播算法（Back Propagation，BP算法）[10]是一種神經網絡的梯度計算方法。反向傳播算法先定義模型在訓練樣本上的代價函數，再求代價函數對於每個參數的梯度。反向傳播算法巧妙的利用了下層神經元的梯度可由上層神經元的殘差導出的規律，求解的過程也正如算法的名字那樣，自上而下反向逐層計算，直至獲得所有參數的梯度。反向傳播算法可以幫助訓練基於統計的機器學習模型，從大量的訓練樣本中挖掘出統計規律，進而可對未標註的數據進行預測。這種基於統計的學習方法比起傳統的基於規則的方法具備很多優越性[11]。

上世紀八九十年代，人們提出了一系列機器學習模型，應用最為廣泛的包括支持向量機（Support Vector Machine，SVM）[12]和邏輯回歸（Logistic Regression，LR）[13]，這兩種模型分別可以看作包含1個隱藏層和沒有隱藏層的淺層模型。訓練時可以利用反向傳播算法計算梯度，再用梯度下降方法在參數空間中尋找最優解。淺層模型往往具有凸代價函數，理論分析相對簡單，訓練方法也容易掌握，取得了很多成功的應用。

4.2 深層模型的訓練難度

淺層模型的局限性在於有限參數和計算單元，對複雜函數的表示能力有限，針對複雜分類問題其泛化能力受到一定的制約。深層模型恰恰可以克服淺層模型的這一弱點，然而應用反向傳播和梯度下降來訓練深層模型，就面臨幾個突出的問題[14]：

1.局部最優。與淺層模型的代價函數不同，深層模型的每個神經元都是非線性變換，代價函數是高度非凸函數，採用梯度下降的方法容易陷入局部最優。

2.梯度彌散。使用反向傳播算法傳播梯度的時候，隨著傳播深度的增加，梯度的幅度會急劇減小，會導致淺層神經元的權重更新非常緩慢，不能有效學習。這樣一來，深層模型也就變成了前幾層相對固定，只能改變最後幾層的淺層模型。

3.數據獲取。深層模型的表達能力強大，模型的參數也相應增加。對於訓練如此多參數的模型，小訓練數據集是不能實現的，需要海量的有標記的數據，否則只能導致嚴重的過擬合（Over fitting）。

4.3 深層模型的訓練方法

儘管挑戰很大，Hinton教授並沒有放棄努力，他30年來一直從事相關研究，終於有了突破性的進展。2006年，他在《Science》上發表了一篇文章[1]，掀起了深度學習在學術界和工業界的浪潮。這篇文章的兩個主要觀點是：

1.多隱藏層的人工神經網絡具有優異的特徵學習能力，學習到的特徵對數據有更本質的刻畫，從而有利於可視化或分類。

2.深度神經網絡在訓練上的難度，可以通過「逐層初始化」（Layer-wise Pre-training）來有效克服，文中給出了無監督的逐層初始化方法。

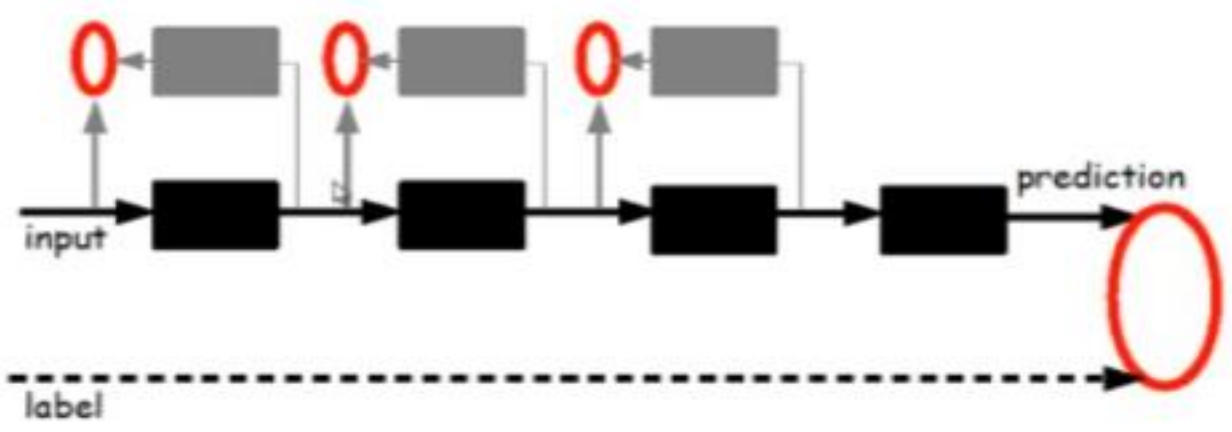

图 7 逐层初始化的方法（原图由 Marc'Aurelio Ranzato 提供）

優異的特徵刻畫能力前文已經提到，不再累述，下面重點解釋一下「逐層初始化」的方法。

給定原始輸入後，先要訓練模型的第一層，即圖中左側的黑色框。黑色框可以看作是一個編碼器，將原始輸入編碼為第一層的初級特徵，可以將編碼器看作模型的一種「認知」。為了驗證這些特徵確實是輸入的一種抽象表示，且沒有丟失太多信息，需要引入一個對應的解碼器，即圖中左側的灰色框，可以看作模型的「生成」。為了讓認知和生成達成一致，就要求原始輸入通過編碼再解碼，可以大致還原為原始輸入。因此將原始輸入與其編碼再解碼之後的誤差定義為代價函數，同時訓練編碼器和解碼器。訓練收斂後，編碼器就是我們要的第一層模型，而解碼器則不再需要了。這時我們得到了原始數據的第一層抽象。固定第一層模型，原始輸入就映射成第一層抽象，將其當作輸入，如法炮製，可以繼續訓練出第二層模型，再根據前兩層模型訓練出第三層模型，以此類推，直至訓練出最高層模型。

逐層初始化完成後，就可以用有標籤的數據，採用反向傳播算法對模型進行整體有監督的訓練了。這一步可看作對多層模型整體的精細調整。由於深層模型具有很多局部最優解，模型初始化的位置將很大程度上決定最終模型的質量。「逐層初始化」的步驟就是讓模型處於一個較為接近全局最優的位置，從而獲得更好的效果。

4.4 淺層模型和深層模型的對比

表 2 浅层模型和深层模型的对比

	浅层模型	深层模型
模型层数	1-2	5-10
模型表达能力	有限	强大
特征提取方式	特征工程	自动抽取特征
代价函数凸性	凸代价函数 没有局部最优点 可以收敛到全局最优	高度非凸的代价函数 存在大量的局部最优点 容易收敛到局部最优
训练难度	容易	复杂，需要较多技巧
理论	有成熟的理论基础	理论分析困难
依赖先验知识	依赖更多先验知识	依赖较少先验知识
数据需求量	多	更多
适用场景	需要简单特征的任务： 发电机故障诊断 时间序列处理 视频字幕定位提取 …	需要高度抽象特征的任务： 语音识别 图像 自然语言处理

淺層模型有一個重要的特點，需要依靠人工經驗來抽取樣本的特徵，模型的輸入是這些已經選取好的特徵，模型只用來負責分類和預測。在淺層模型中，最重要的往往不是模型的優劣，而是特徵的選取的優劣。因此大多數人力都投入到特徵的開發和篩選中來，不但需要對任務問題領域有深刻的理解，還要花費大量時間反覆實驗摸索，這也限制了淺層模型的效果。

事實上，逐層初始化深層模型也可以看作是特徵學習的過程，通過隱藏層對原始輸入的一步一步抽象表示，來學習原始輸入的數據結構，找到更有用的特徵，從而最終提高分類問題的準確性。在得到有效特徵之後，模型整體訓練也可以水到渠成。

五、深層模型的層次組件

深層模型是包含多個隱藏層的神經網絡，每一層的具體結構又是怎樣的呢？本節介紹一些常見的深層模型基本層次組件。

5.1 自編碼器（Auto-Encoder）

一種常見的深層模型是由自編碼器（Auto-Encoder）構造的[6]。自編碼器可以利用一組無標籤的訓練數據{x(1), x(2), … }（其中x(i)是一個n維向量）進行無監督的模型訓練。它採用反向傳播算法，讓目標值接近輸入值。下圖是一個自編碼器的示例：

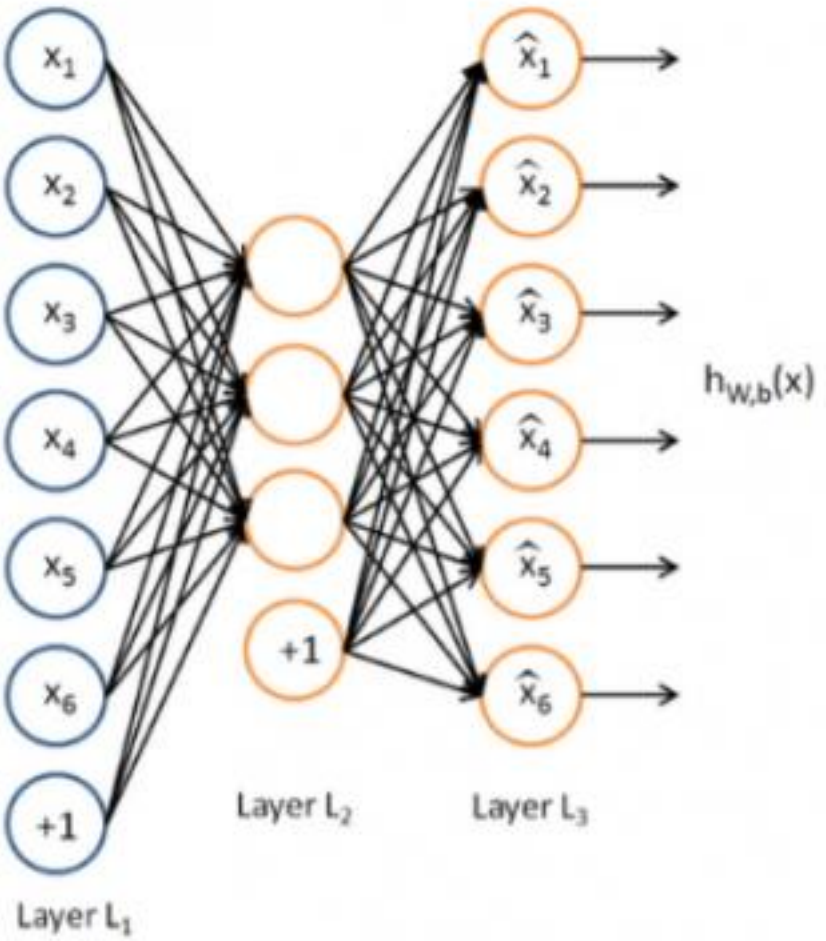

图 8 自编码器

自編碼器嘗試訓練一個恆等函數，讓輸出接近等於輸入值，恆等函數看似沒有學習的意義，但考慮到隱藏層神經元的數目（本例中為3個）小於輸入向量的維數（本例中為6維），事實上隱藏層就變成了輸入數據的一種壓縮的表示，或說是抽象的簡化表示。如果網絡的輸入是完全隨機的，將高維向量壓縮成低維向量會難以實現。但訓練數據往往隱含著特定的結構，自編碼器就會學到這些數據的相關性，從而得到有效的壓縮表示。實際訓練後，如果代價函數越小，就說明輸入和輸出越接近，也就說明這個編碼器越靠譜。當然，自編碼器訓練完成後，實際使用時只需要它的前一層，即編碼部分，解碼部分就沒用了。

稀疏自編碼器（Sparse Auto-Encoder）是自編碼器的一個變體，它在自編碼器的基礎上加入正則化（Regularity）。正則化是在代價函數中加入抑制項，希望隱藏層節點的平均激活值接近於0，有了正則化的約束，輸入數據可以用少數隱藏節點表達。之所以採用稀疏自編碼器，是因為稀疏的表達往往比稠密的表達更有效，人腦神經系統也是稀疏連接，每個神經元只與少數神經元連接。

降噪自編碼器是另一種自編碼器的變體。通過在訓練數據中加入噪聲，可訓練出對輸入信號更加魯棒的表達，從而提升模型的泛化能力，可以更好地應對實際預測時夾雜在數據中的噪聲。

得到自編碼器後，我們還想進一步了解自編碼器到底學到了什麼。例如，在10×10的圖像上訓練一個稀疏自編碼器，然後對於每個隱藏神經元，找到什麼樣的圖像可以讓隱藏神經元獲得最大程度的激勵，即這個隱藏神經元學習到了什麼樣的特徵。將100個隱藏神經元的特徵都找出來，得到了如下100幅圖像：

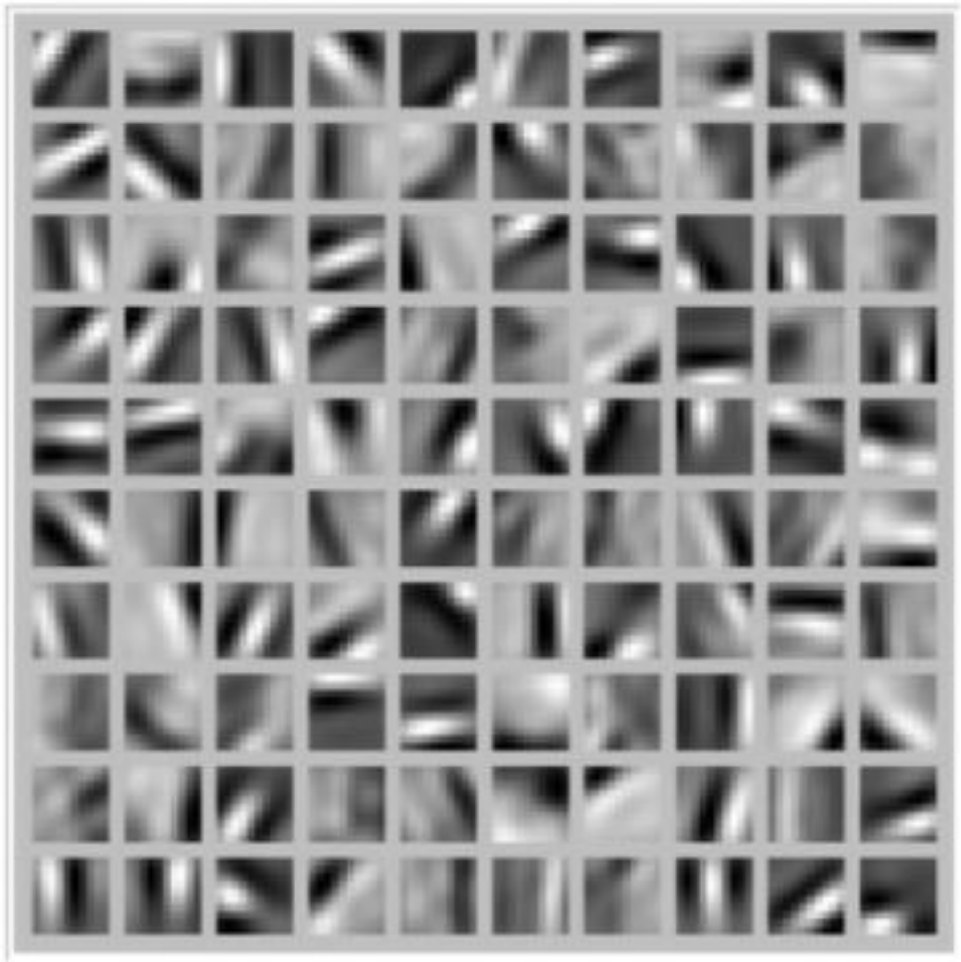

图 9 自编码器的隐藏神经元 [6]

可以看出，這100幅圖像具備了從不同方向檢測物體邊緣的能力。顯然，這樣的能力對後續的圖像識別很有幫助。

5.2 受限玻爾茲曼機（Restricted Boltzmann Machine，RBM）

受限玻爾茲曼機（Restricted Boltzmann Machine， RBM）是一個二部圖，一層是輸入層（v），另一層是隱藏層（h），假設所有節點都是隨機二值變量節點，只能取值0或1，同時假設全機率分布p(v, h)滿足Boltzmann分布。

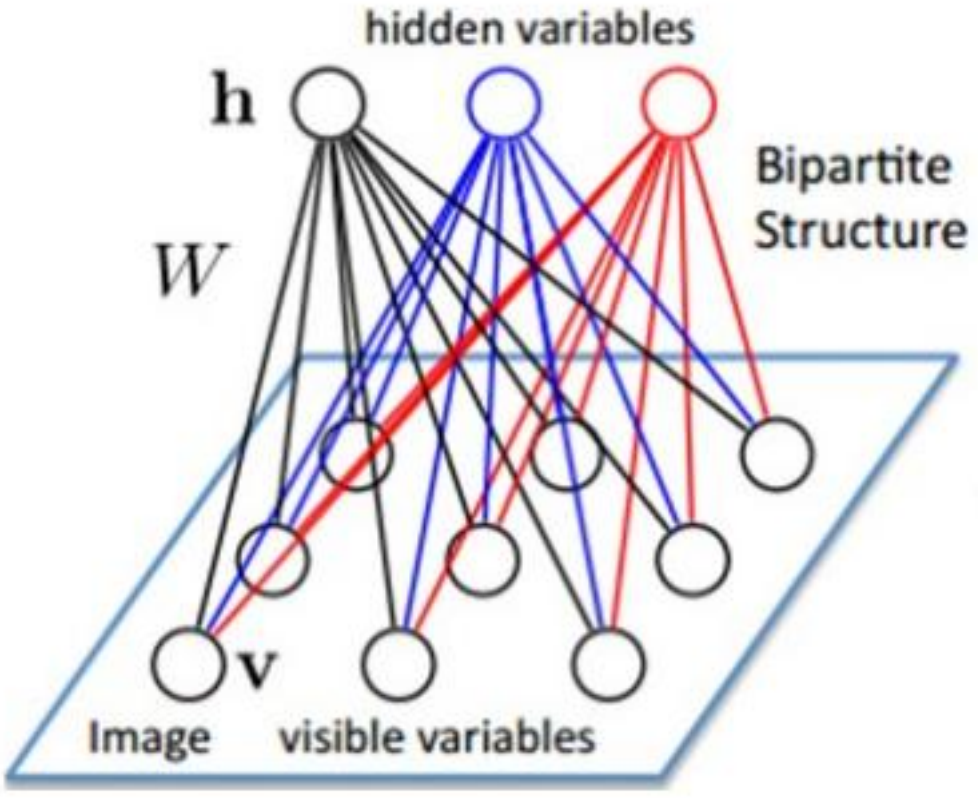

图 10 受限玻尔兹曼机（RBM）

由於同層節點之間沒有連接，因此已知輸入層的情況下，隱藏層的各節點是條件獨立的；反之，已知隱藏層的情況下，輸入層各節點也是條件獨立的。同時，可以根據Boltzmann分布，當輸入v時通過p(h|v)生成隱藏層，得到隱藏層之後再通過p(v|h)生成輸入層。相信很多讀者已經猜到了，可以按照訓練其他網絡類似的思路，通過調整參數，希望通過輸入v生成的h，再生成的v』與v儘可能接近，則說明隱藏層h是輸入層v的另外一種表示。這樣就可以作為深層模型的基本層次組件了。全部用RBM形成的深層模型為深度玻爾茲曼機（Deep Boltzmann Machine，DBM）。如果將靠近輸入層的部分替換為貝葉斯信念網絡，即有向圖模型，而在遠離輸入層的部分仍然使用RBM，則稱為深度信念網絡（Deep Belief Networks，DBN）。

5.3 卷積神經網絡（Convolutional Neural Networks, CNN）

以上介紹的編碼器都是全連通網絡，可以完成10×10的圖像識別，如手寫體數字識別問題。然而對於更大的圖像，如100×100的圖像，如果要學習100個特徵，則需要1,000,000個參數，計算時間會大大增加。解決這種尺寸圖像識別的有效方法是利用圖像的局部性，構造一個部分聯通的網絡。一種最常見的網絡是卷積神經網絡（Convolutional Neural Networks，CNN）[15][16]，它利用圖像固有的特性，即圖像局部的統計特性與其他局部是一樣的。因此從某個局部學習來的特徵同樣適用於另外的局部，對於這個圖像上的所有位置，都能使用同樣的特徵。

具體地說，假設有一幅100×100的圖像，要從中學習一個10×10的局部圖像特徵的神經元，如果採用全連接的方式，100×100維的輸入到這個神經元需要有10000個連接權重參數。而採用卷積核的方式，只有10×10=100個參數權重，卷積核可以看作一個10×10的小窗口，在圖像上上下左右移動，走遍圖像中每個10×10的位置（共有91×91個位置）。每移動到一個位置，則將該位置的輸入與卷積核對應位置的參數相乘再累加，得到一個輸出

值（輸出值是91×91的圖像）。卷積核的特點是連接數雖然很多，有91×91×10×10個連接，但是參數只有10×10=100個，參數數目大大減小，訓練也變得容易了，並且不容易產生過擬合。當然，一個神經元只能提取一個特徵，要提取多個特徵就要多個卷積核。

下圖揭示了對一幅8×8維圖像使用卷積方法提取特徵的示意過程。其中使用了3×3的卷積核，走遍圖像中每個3×3的位置後，最終得到6×6維的輸出圖像：

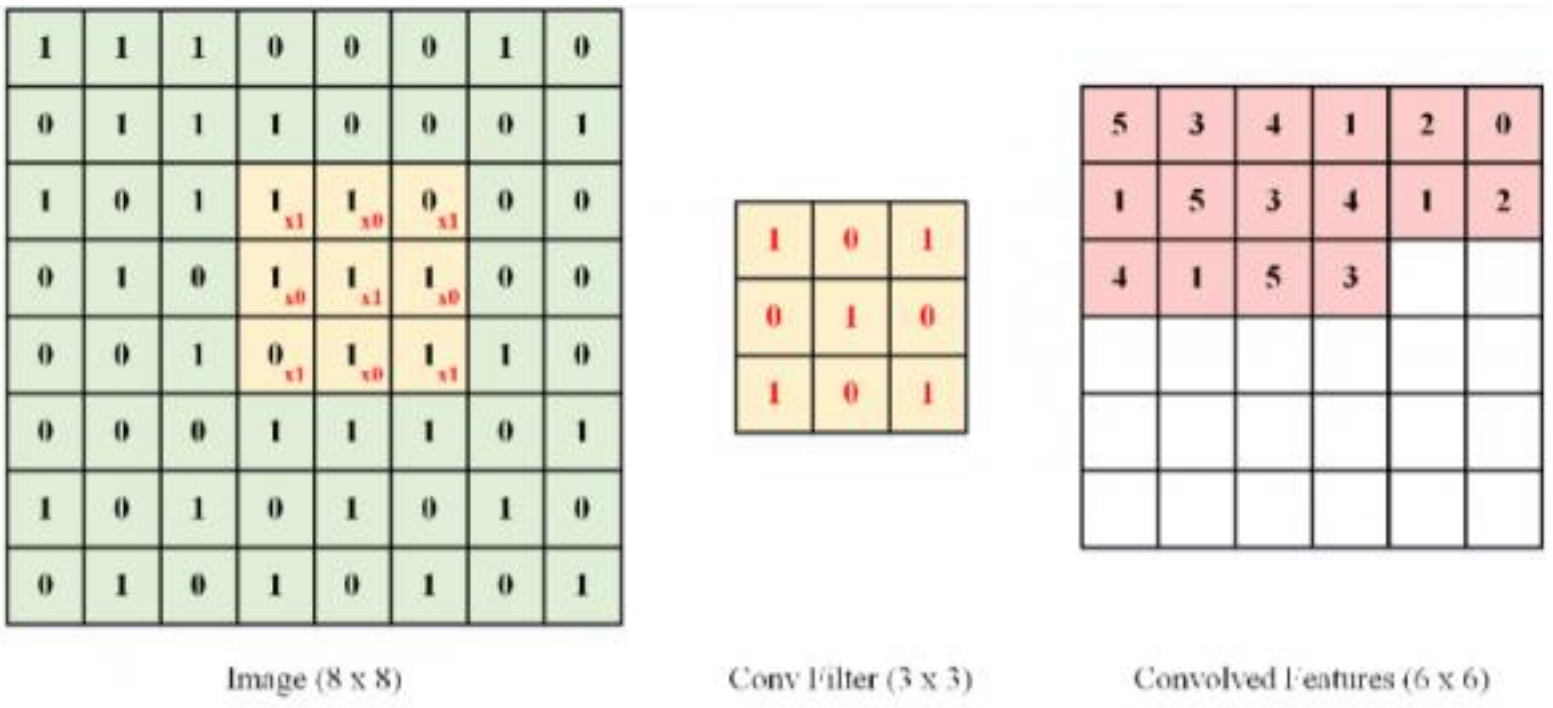

图 11 8×8 图像的卷积过程示意

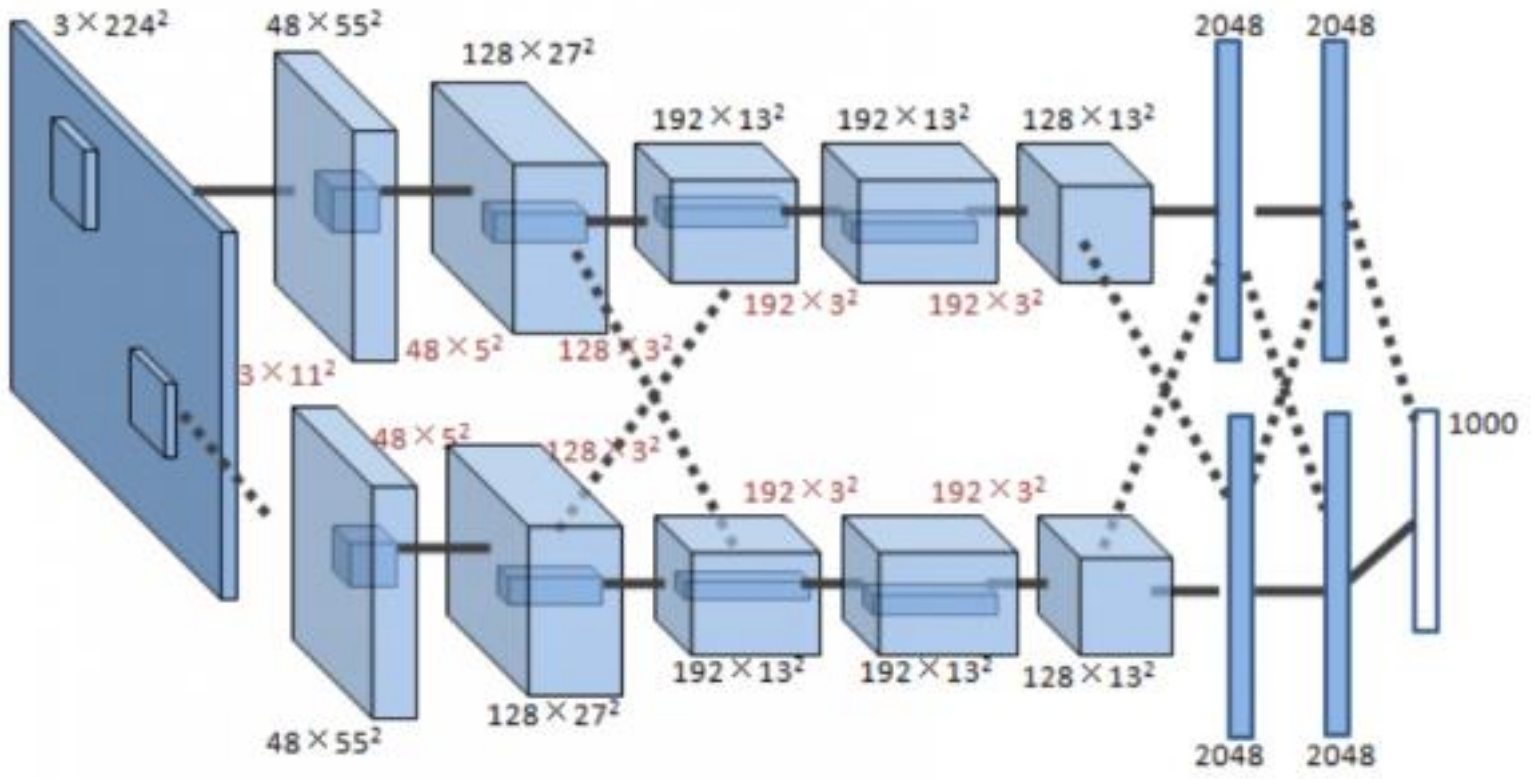

图 12 用户图像分类的卷积神经网络

如圖所示是Hinton的研究小組在ImageNet競賽中使用的卷積神經網絡[2]，共有5個卷積層，每層分別有96，256，384，384和256個卷積核，每層卷積核的大小分別為11×11，5×5，3×3，3×3和3×3。網絡的最後兩層是全連接層。

六、深度學習的訓練加速

深層模型訓練需要各種技巧，例如網絡結構的選取，神經元個數的設定，權重參數的初始化，學習率的調整，Mini-batch的控制等等。即便對這些技巧十分精通，實踐中也要多次訓練，反覆摸索嘗試。此外，深層模型參數多，計算量大，訓練數據的規模也更大，需要消耗很多計算資源。如果可以讓訓練加速，就可以在同樣的時間內多嘗試幾個新主意，多調試幾組參數，工作效率會明顯提升，對於大規模的訓練數據和模型來說，更可以將難以完成的任務變成可能。這一節就談談深層模型的訓練加速方法。

6.1 GPU加速

矢量化編程是提高算法速度的一種有效方法。為了提升特定數值運算操作（如矩陣相乘、矩陣相加、矩陣-向量乘法等）的速度，數值計算和並行計算的研究人員已經努力了幾十年。矢量化編程強調單一指令並行操作多條相似數據，形成單指令流多數據流（SIMD）的編程泛型。深層模型的算法，如BP，Auto-Encoder，CNN等，都可以寫成矢量化的形式。然而，在單個CPU上執行時，矢量運算會被展開成循環的形式，本質上還是串行執行。

GPU（Graphic Process Units，圖形處理器）的眾核體系結構包含幾千個流處理器，可將矢量運算並行化執行，大幅縮短計算時間。隨著NVIDIA、AMD等公司不斷推進其GPU的大規模並行架構支持，面向通用計算的GPU（General-Purposed GPU, GPGPU）已成為加速可並行應用程式的重要手段。得益於GPU眾核（many-core）體系結構，程序在GPU系統上的運行速度相較於單核CPU往往提升幾十倍乃至上千倍。目前GPU已經發展到了較為成熟的階段，受益最大的是科學計算領域，典型的成功案例包括多體問題（N-Body Problem）、蛋白質分子建模、醫學成像分析、金融計算、密碼計算等。

利用GPU來訓練深度神經網絡，可以充分發揮其數以千計計算核心的高效並行計算能力，在使用海量訓練數據的場景下，所耗費的時間大幅縮短，占用的伺服器也更少。如果對針對適當的深度神經網絡進行合理優化，一塊GPU卡可相當於數十甚至上百台CPU伺服器的計算能力，因此GPU已經成為業界在深度學習模型訓練方面的首選解決方案。

6.2數據並行

數據並行是指對訓練數據做切分，同時採用多個模型實例，對多個分片的數據並行訓練。

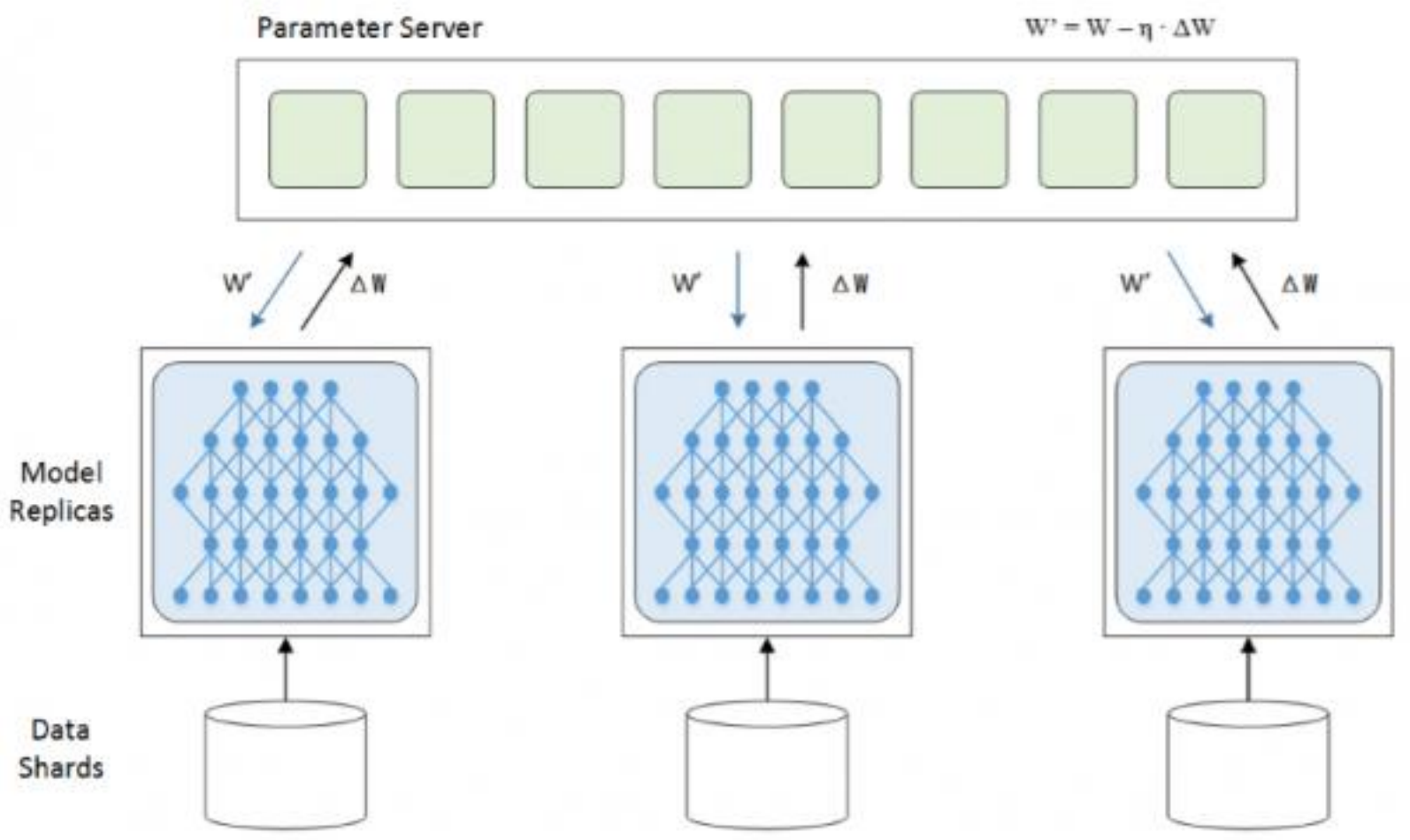

图 13 数据并行的基本架构 [17]

要完成數據並行需要做參數交換，通常由一個參數伺服器（Parameter Server）來幫助完成。在訓練的過程中，多個訓練過程相互獨立，訓練的結果，即模型的變化量ΔW需要匯報給參數伺服器，由參數伺服器負責更新為最新的模型W』 = W − η · ΔW，然後再將最新的模型W』分發給訓練程序，以便從新的起點開始訓練。

數據並行有同步模式和異步模式之分。同步模式中，所有訓練程序同時訓練一個批次的訓練數據，完成後經過同步，再同時交換參數。參數交換完成後所有的訓練程序就有了共同的新模型作為起點，再訓練下一個批次。而異步模式中，訓練程序完成一個批次的訓練數據，立即和參數伺服器交換參數，不考慮其他訓練程序的狀態。異步模式中一個訓練程序的最新結果不會立刻體現在其他訓練程序中，直到他們進行下次參數交換。

參數伺服器只是一個邏輯上的概念，不一定部署為獨立的一台伺服器。有時候它會附屬在某一個訓練程序上，有時也會將參數伺服器按照模型劃分為不同的分片，分別部署。

6.3模型並行

模型並行將模型拆分成幾個分片，由幾個訓練單元分別持有，共同協作完成訓練。當一個神經元的輸入來自另一個訓練單元上的神經元的輸出時，產生通信開銷。

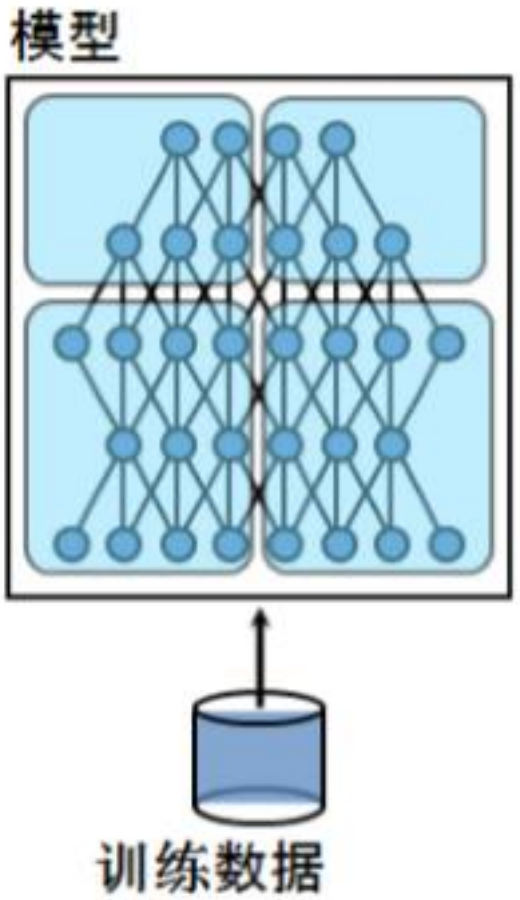

图 14 模型并行的基本架构 [17]

多數情況下，模型並行帶來的通信開銷和同步消耗超過數據並行，因此加速比也不及數據並行。但對於單機內存無法容納的大模型來說，模型並行是一個很好的選擇。令人遺憾的是，數據並行和模型並行都不能無限擴展。數據並行的訓練程序太多時，不得不減小學習率，以保證訓練過程的平穩；模型並行的分片太多時，神經元輸出值的交換量會急劇增加，效率大幅下降。因此，同時進行模型並行和數據並行也是一種常見的方案。如下圖所示，4個GPU分為兩組，GPU0，1為一組模型並行，GPU2，3為另一組，每組模型並行在計算過程中交換輸出值和殘差。兩組GPU之間形成數據並行，Mini-batch結束後交換模型權重，考慮到模型的藍色部分由GPU0和GPU2持有，而黃色部分由GPU1和GPU3持有，因此只有同色的GPU之間需要交換權重。

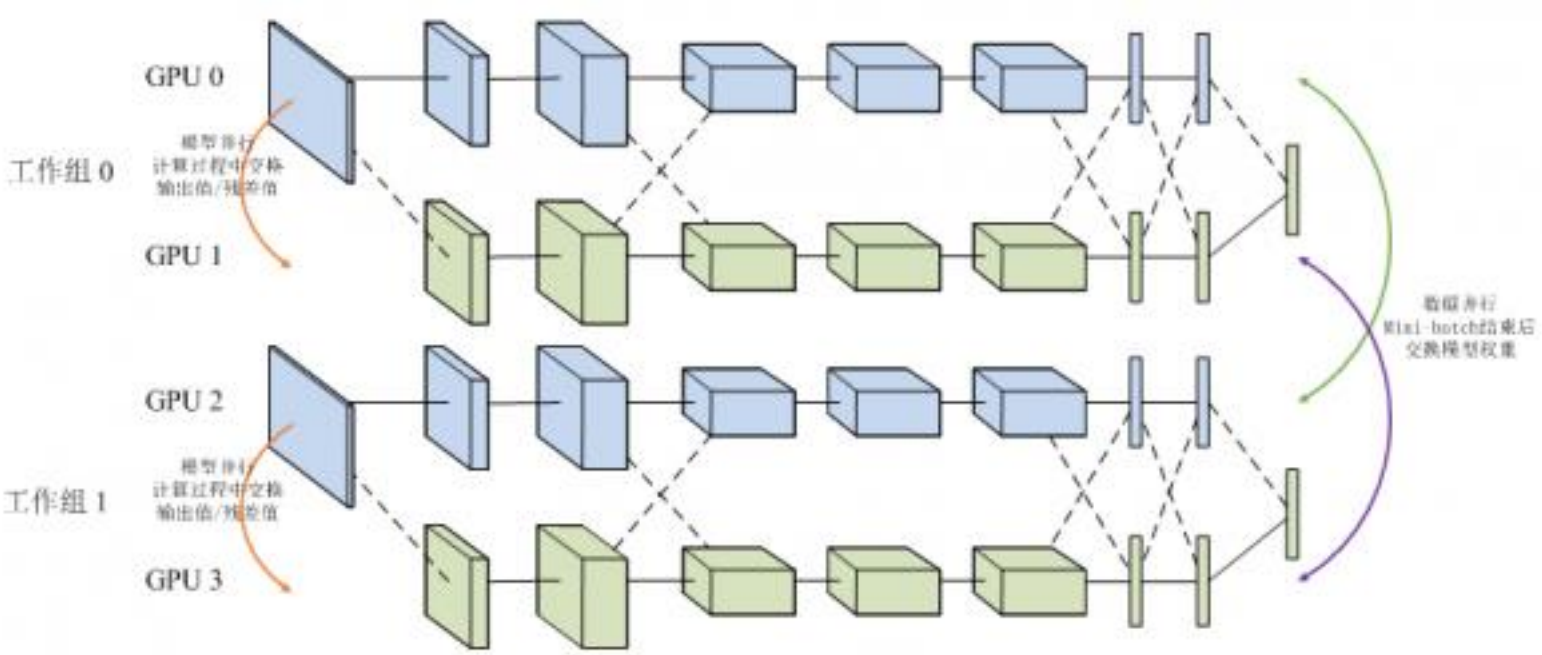

图 15 4GPU 卡的数据并行和模型并行混合架构

6.4計算集群

搭建CPU集群用於深度神經網絡模型訓練也是業界常用的解決方案，其優勢在於利用大規模分布式計算集群的強大計算能力，利用模型可分布式存儲、參數可異步通信的特點，達到快速訓練深層模型的目的。

CPU集群方案的基本架構包含用於執行訓練任務的Worker、用於分布式存儲分發模型的參數伺服器（Parameter Server）和用於協調整體任務的主控程序（Master）。CPU集群方案適合訓練GPU內存難以容納的大模型，以及稀疏連接神經網絡。Andrew Ng和Jeff Dean在Google用1000台CPU伺服器，完成了模型並行和Downpour SGD數據並行的深度神經網絡訓練[17]。

結合GPU計算和集群計算技術，構建GPU集群正在成為加速大規模深度神經網絡訓練的有效解決方案。GPU集群搭建在CPU-GPU系統之上，採用萬兆網卡或Infiniband等更加快速的網絡通信設施，以及樹形拓撲等邏輯網絡拓撲結構。在發揮出單節點較高計算能力的基礎上，再充分挖掘集群中多台伺服器的協同計算能力，進一步加速大規模訓練任務。

7 深度學習的軟體工具及平台

目前，在深度學習系統實現方面，已有諸多較為成熟的軟體工具和平台。

7.1 開源軟體

在開源社區，主要有以下較為成熟的軟體工具：

Kaldi是一個基於C++和CUDA的語音識別工具集[18][19]，提供給語音識別的研究人員使用。Kaldi中既實現了用單個GPU加速的深度神經網絡SGD訓練，也實現了CPU多線程加速的深度神經網絡SGD訓練。

基於C++/CUDA編寫，採用反向傳播算法的深度卷積神經網絡實現[20][21]。2012年cuda-convnet發布，可支持單個GPU上的訓練，基於其訓練的深度卷積神經網絡模型在ImageNet LSVRC-2012對圖像按1000個類目分類，取得Top 5分類15%錯誤率的結果[2]；2014年發布的版本可以支持多GPU上的數據並行和模型並行訓練[22]。

Caffe提供了在CPU以及GPU上的快速卷積神經網絡實現，同時提供訓練算法，使用NVIDIA K40或Titan GPU可以1天完成多於40,000,000張圖片的訓練[23][24]。

Theano提供了在深度學習數學計算方面的Python庫，它整合了NumPy矩陣計算庫，可以運行在GPU上，並提供良好的算法上的擴展性[25][26]。

OverFeat是由紐約大學CILVR實驗室開發的基於卷積神經網絡系統，主要應用場景為圖像識別和圖像特徵提取[27]。

Torch7是一個為機器學習算法提供廣泛支持的科學計算框架，其中的神經網絡工具包

（Package）實現了均方標準差代價函數、非線性激活函數和梯度下降訓練神經網絡的算法等基礎模塊，可以方便地配置出目標多層神經網絡開展訓練實驗[28]。

7.2 工業界平台

在工業界，Google、Facebook、百度、騰訊等公司都實現了自己的軟體框架：

Google的DistBelief系統是CPU集群實現的數據並行和模型並行框架，集群內使用上萬CPU core來訓練多達10億參數的深度神經網絡模型。DistBelief應用的主要算法有Downpour SGD和L-BFGS，支持的目標應用有語音識別和2.1萬類目的圖像分類[17]。

Google的COTS HPC系統是GPU實現的數據並行和模型並行框架，GPU伺服器間使用了Infiniband連接，並由MPI控制通信。COTS可以用3台GPU伺服器在數天內完成對10億參數的深度神經網絡訓練[29]。

Facebook實現了多GPU訓練深度卷積神經網絡的並行框架，結合數據並行和模型並行的方式來訓練CNN模型，使用4張NVIDIA Titan GPU可在數天內訓練ImageNet的1000分類網絡[30]。

百度搭建了Paddle（Parallel Asynchonous Distributed Deep Learning）多機GPU訓練平台[31]。將數據分布到不同機器，通過Parameter Server協調各機器訓練。Paddle支持數據並行和模型並行。

騰訊深度學習平台（Mariana）是為加速深度學習模型訓練而開發的並行化平台，包括深度神經網絡的多GPU數據並行框架，深度卷積神經網絡的多GPU模型並行和數據並行框架，以及深度神經網絡的CPU集群框架。Mariana基於特定應用的訓練場景，設計定製化的並行化訓練平台，支持了語音識別、圖像識別，並積極探索在廣告推薦中的應用[32]。

8 總結

近年來人工智慧領域掀起了深度學習的浪潮，從學術界到工業界都熱情高漲。深度學習嘗試解決人工智慧中抽象認知的難題，從理論分析和應用方面都獲得了很大的成功。可以說深度學習是目前最接近人腦的智能學習方法。

深度學習可通過學習一種深層非線性網絡結構，實現複雜函數逼近，並展現了強大的學習數據集本質和高度抽象化特徵的能力。逐層初始化等訓練方法顯著提升了深層模型的可學習型。與傳統的淺層模型相比，深層模型經過了若干層非線性變換，帶給模型強大的表達能力，從而有條件為更複雜的任務建模。與人工特徵工程相比，自動學習特徵，更能挖掘出數據中豐富的內在信息，並具備更強的可擴展性。深度學習順應了大數據的趨勢，有了充足的訓練樣本，複雜的深層模型可以充分發揮其潛力，挖掘出海量數據中蘊含的豐富信息。強有力的基礎設施和定製化的並行計算框架，讓以往不可想像的訓練任務加速完成，為深度學習走向實用奠定了堅實的基礎。已有Kaldi，Cuda-convnet，Caffe等多個針對不同深度模型的開源實現，Google、Facebook、百度、騰訊等公司也實現了各自的並行化框架。

深度學習引爆的這場革命，將人工智慧帶上了一個新的台階，不僅學術意義巨大，而且實用性很強，深度學習將成為一大批產品和服務背後強大的技術引擎。

參考文獻

[1] Geoffery E. Hinton, Salakhutdinov RR. Reducing the dimensionality of data with neural networks. Science. 2006 Jul 28;313(5786):504-7.

[2] ImageNet Classification with Deep Convolutional Neural Networks, Alex Krizhevsky, Ilya Sutskever, Geoffrey E Hinton, NIPS 2012.

[3] Q.V. Le, M.A. Ranzato, R. Monga, M. Devin, K. Chen, G.S. Corrado, J. Dean, A.Y. Ng. Building high-level features using large scale unsupervised learning. ICML, 2012.

[4] Rick Rashid, Speech Recognition Breakthrough for the Spoken, Translated Wordhttp://www.youtube.com/watch?v=Nu-nlQqFCKg

[5] NYU 「Deep Learning」 Professor LeCun Will Lead Facebook’s New Artificial Intelligence Lab. http://techcrunch.com/2013/12/09/facebook-artificial-intelligence-lab-lecun/

[6] Stanford deep learning tutorial

[7] A Primer on Deep Learning

[8] The Nobel Prize in Physiology or Medicine 1981.

[9] Bruno A. Olshausen & David J. Field, Emergence of simple-cell receptive field properties by learning a sparse code for natural images. Nature. Vol 381. 13 June, 1996 http://www.cs.ubc.ca/~little/cpsc425/olshausen_field_nature_1996.pdf

[10] Back propagation algorithm http://ufldl.stanford.edu/wiki/index.php/Backpropagation_Algorithm

[11] 余凱，深度學習-機器學習的新浪潮，Technical News程序天下事 http://blog.csdn.net/datoubo/article/details/8577366

[12] Support Vector Machine http://en.wikipedia.org/wiki/Support_vector_machine

[13] Logistic Regression http://en.wikipedia.org/wiki/Logistic_regression

[14] Deep Networks Overview http://ufldl.stanford.edu/wiki/index.php/Deep_Networks:_Overview

[15] Y. LeCun and Y. Bengio. Convolutional networks for images, speech, and time-series. In M. A. Arbib, editor, The Handbook of Brain Theory and Neural Networks. MIT

Press, 1995

[16] Introduction to Convolutional neural network
http://en.wikipedia.org/wiki/Convolutional_neural_network

[17] Dean, J., Corrado, G.S., Monga, R., et al, Ng, A. Y. Large Scale Distributed Deep Networks. In Proceedings of the Neural Information Processing Systems (NIPS』12) (Lake Tahoe, Nevada, United States, December 3–6, 2012). Curran Associates, Inc, 57 Morehouse Lane, Red Hook, NY, 2013, 1223-1232.

[18] Kaldi project http://kaldi.sourceforge.net/

[19] Povey, D., Ghoshal, A. Boulianne, G., et al, Vesely, K. Kaldi. The Kaldi Speech Recognition Toolkit. in Proceedings of IEEE 2011 Workshop on Automatic Speech Recognition and Understanding(ASRU 2011) (Hilton Waikoloa Village, Big Island, Hawaii, US, December 11-15, 2011). IEEE Signal Processing Society. IEEE Catalog No.: CFP11SRW-USB.

[20] cuda-convent https://code.google.com/p/cuda-convnet/

[21] Krizhevsky, A., Sutskever, I., and Hinton, G.E. ImageNet Classification with Deep Convolutional Neural Networks. In Proceedings of the Neural Information Processing Systems (NIPS』12) (Lake Tahoe, Nevada, United States, December 3–6, 2012). Curran Associates, Inc, 57 Morehouse Lane, Red Hook, NY, 2013, 1097-1106.

[22] Krizhevsky, A. Parallelizing Convolutional Neural Networks. in tutorial of IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2014). (Columbus, Ohio, USA, June 23-28, 2014). 2014.

[23] caffe http://caffe.berkeleyvision.org/

[24] Jia, Y. Q. Caffe: An Open Source Convolutional Architecture for Fast Feature Embedding. http://caffe.berkeleyvision.org (2013).

[25] Theano https://github.com/Theano/Theano

[26] J. Bergstra, O. Breuleux, F. Bastien, P. Lamblin, R. Pascanu, G. Desjardins, J. Turian, D. Warde-Farley and Y. Bengio. Theano: A CPU and GPU Math Expression Compiler. Proceedings of the Python for Scientific Computing Conference (SciPy) 2010. June 30 – July 3, Austin, TX.

[27] Overfeat http://cilvr.nyu.edu/doku.php?id=code:start

[28] Torch7 http://torch.ch

[29] Coates, A., Huval, B., Wang, T., Wu, D. J., Ng, A. Y. Deep learning with COTS HPC systems. In Proceedings of the 30th International Conference on Machine Learning (ICML』13) (Atlanta, Georgia, USA, June 16–21, 2013). JMLR: W&CP volume 28(3), 2013, 1337-1345.

[30] Yadan, O., Adams, K., Taigman, Y., Ranzato, M. A. Multi-GPU Training of ConvNets. arXiv:1312.5853v4 [cs.LG] (February 2014)

[31] Kaiyu, Large-scale Deep Learning at Baidu, ACM International Conference on Information and Knowledge Management (CIKM 2013)

[32] aaronzou, Mariana深度學習在騰訊的平台化和應用實踐

[33] Geoffrey E. Hinton, Simon Osindero, Yee-Whye Teh, A fast learning algorithm for deep belief nets Neural Compute, 18(7), 1527-54 (2006)

[34] Andrew Ng. Machine Learning and AI via Brain simulations,

[35] Geoffrey Hinton：UCLTutorial on: Deep Belief Nets

[36] Krizhevsky, Alex. 「ImageNet Classification with Deep Convolutional Neural Networks」. Retrieved 17 November 2013.

[37] 「Convolutional Neural Networks (LeNet) – DeepLearning 0.1 documentation」. DeepLearning 0.1. LISA Lab. Retrieved 31 August 2013.

[38] Bengio, Learning Deep Architectures for AI，http://www.iro.umontreal.ca/~bengioy/papers/ftml_book.pdf；

[39] Deep Learning http://deeplearning.net/

[40] Deep Learning http://www.cs.nyu.edu/~yann/research/deep/

[41] Introduction to Deep Learning. http://en.wikipedia.org/wiki/Deep_learning

[42] Google的貓臉識別:人工智慧的新突破http://www.36kr.com/p/122132.html

[43] Andrew Ng's talk video: http://techtalks.tv/talks/machine-learning-and-ai-via-brain-simulations/57862/

[44] Invited talk 「A Tutorial on Deep Learning」 by Dr. Kai Yu http://vipl.ict.ac.cn/News/academic-report-tutorial-deep-learning-dr-kai-yu

via：騰訊大數據

End.