深度學習 讀書筆記

Introduction:

Conventional machine-learning techniques are limited in their ability in pattern recognitions when given natural data in their raw forms. It usually requires considerable domain expertise to structure very domain specific features to attain good results in very specific questions. As a result, it gets extremely hard to design such feature vectors to improve the performance because the exact form of perfect feature vectors are either non-existent or inconceivable by humans. In fields such as object recognition in images, speech translations, and recommendation systems, conventional machine learning techniques were far worse than human performance, until the coming of deep-learning techniques.

傳統的機器學習技術仰賴專家利用專業領域的知識設計出針對特定問題的特徵函數來把原始資料轉化成能夠做 classification 或是 pattern recognition 的特徵（feature）。但是這樣的機器學習技巧在許多問題上面遇到了困難。人工智慧科學家歷經幾十年的嘗試，無法設計出夠好的特徵函數使機器能夠做到與人的performance 匹敵的語音辨識，圖像辨識，或是下圍棋這些事情。直到最近十年深度學習的突飛猛進，才讓人工智慧在這些 input 是自然資料的問題上有了重大突破。

在人工神經網路沉寂的歲月裡，三位如今執深度學習牛耳的大師 Geoffery Hinton，Yann LeCun，Yoshua Bengio 默默耕耘，把這一門曾經被視為死胡同的人工智慧分支重新帶向世人的目光。如今深度學習已經成為人工智慧的顯學，這三位研究員的地位也水漲船高。他們三位也是這次讀書筆記所讀的paper, Deep Learning 的作者。這篇文章可以說是了解深度學習是什麼的最佳起點。

由左至右為 Geoffery Hinton，Yann LeCun，Yoshua Bengio。
Geoffery Hinton 任教於多倫多大學，同時也是Google Brain 的 首席人工智慧科學家。 Yann LeCun 任教於紐約大學，同時也是 Facebook 的首席人工智慧科學家。 Yoshua Bengio 任教於蒙特婁大學，同時也是 IBM, AT&T 的首席人工智慧科學家。

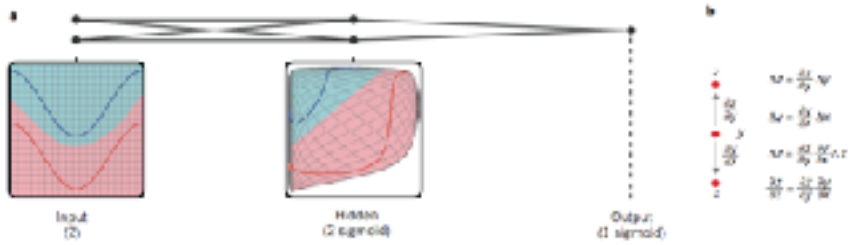
深度學習的應用 ： 深度學習已經被應用在許多領域，包括尋找藥物，重建大腦神經觸突連結圖譜，電腦視覺，自然語言處理，機器人，以及車輛自動駕駛上。

Representation Learning 表徵學習：

表徵學習是相對於傳統機器學習仰賴人工智慧專家手刻（hand craft）出特徵函數使機器可以對資料做分類或是預測。 表徵學習使機器可以憑著原始資料（raw data）自動習得特徵函數。

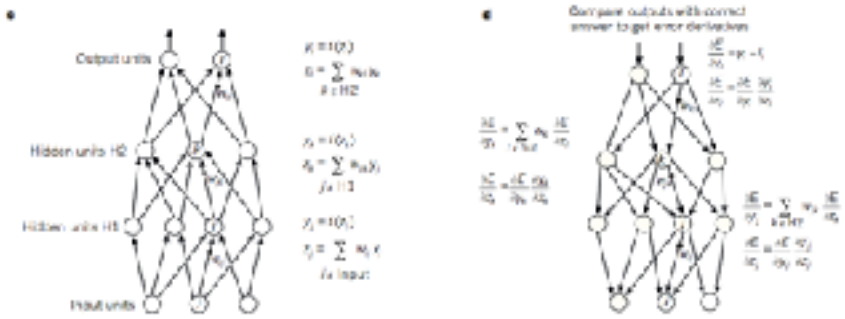
Deep Learning 深度學習 ：

深度學習是一種多層結構的表徵學習。所謂Deep 指的是許多層的人工神經網路。深度學習利用多層的表徵達成 pattern recognition 或是 classification 。每一層的表徵由簡單的非線性函數構成。這些簡單的非線性函數僅僅讓每一層的輸出比輸入稍微抽象一些。然而經過多層的非線性運算，輸出層的表徵卻可以學得把原始輸入資料空間中糾纏的label 在深層的表徵空間中分離開。如下圖例子中，input layer 中藍，紅兩類原始資料可表示為二維向量，輸入向量無法用一次線性函數分離。在 hidden layer 中，表徵為input layer 的非線性函數，使得hidden layer 輸出的表徵空間為input layer 向量空間的非線性mapping. 非線性的mapping 使得本來無法被classification surface 分離的資料能被線性分離。



簡而言之，深度學習利用多層 hidden layer, 理論上能夠把任意糾纏的多類 input vector mapping 到能被簡單 classification surface 分離的表徵空間。這樣的學習省去比須仰賴專家手刻特徵函數的困難。並且是一個更 generalized 的演算法。我們不需要針對每一個問題去設計一個模型。

深度學習演算法：深度學習是實作在多層的人工神經網路上面。人工神經網路由unit 跟 weight 所構成。unit 可以想成是大腦神經元，weight 可以想成是神經元之間的連結強度。 深度學習的結構可以簡單分成 input layer, hidden layer(s),和 output layer. input layer 加上 hidden layers 的層數就是這個network的層數。



Forward Pass　　back propagation

上圖左為 Forward Pass。input 經由 加權總合後經過非線性函數 activation function f(z) 運算，產生該層的output。每層的運算規則相同。除了output layer 依問題不同可能會有不同的 activation function。Forward pass 使用 input data 更新unit 裡面的值，weight 的直保持不變。 Back propagation 則是使用正確解答來計算 cost function. 通常 cost-function 是 prediction 和 answer 的差值平方。Cost function 在上圖右以 E 表示。back propagation 利用每一層的函數皆可微分的特性，可以算出differential dE/dy 和 dE/dz, 使得 weight 可以被 update. 此時unit 的直保持不變。這種學習方式叫做 supervised learning 監督學習。因為訓練神經網路仰賴有標記的正確解答。back

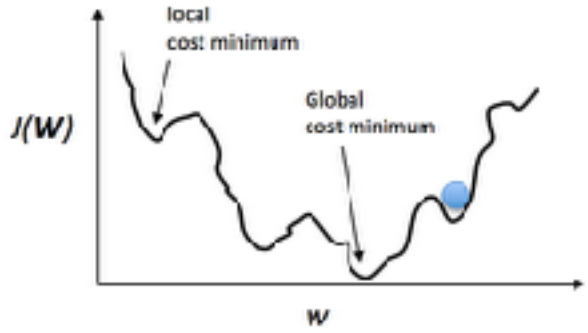
propagation的過程就好像大腦接受到回饋訊號，比較 prediction 和 feedback，最後利用神經可塑性更新連結強度，使得input 可以更好的mapping 到正確解答上。由此可知，資料量越大，出來的結果越準確。

Cost-function optimization 目標函數優化

儘管神經網路的類比非常形象化，但是深度學習以及人工神經網路事實上還是數學上的模型，與人腦或是神經細胞的運作方式並不相關。人工神經網路在做的事情，事實上就是在做 cost function minimization ，有些人把 cost function 加負號後定義成 gain function 所以這樣的過程統稱為 objective function optimization 目標函數優化。back propagation 就是目標函數優化的演算法。

我們可以把目標函數想像成在表徵空間中的一個連續曲面。目標函數對表徵向量取微分得到的就是取面的梯度 gradient。cost-function minimization 事實上就是不斷的從初始值向梯度低的地方移動。這個過程稱為 gradient descend.

Cost function minimization via gradient descend

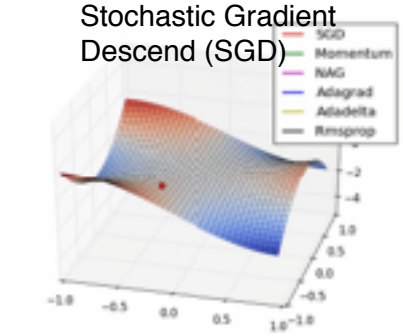


The myth of local minimum 局部及小值的迷思

一般而言，當訓練的結果不好且停滯時，我們直覺上會認為gradient descend 掉入了 cost function surface 的 local minimum。這是長久以來教科書的答案。但是大型的網路或是 deep neural networks 因為擁有大量的參數，事實上是處於一個高維的空間。最近的研究結果以及實際經驗顯示，大型的神經網路並沒有local minimum的問題。 在這種神經網路的 function landscape 上存在的反而是數量龐大，gradient 為零的鞍部。這些高為的鞍部在多數的方向上是向上的，只在少數方向向下。這樣的鞍部無法在二維上畫出，因此一般我們直覺上對鞍部的想法認為鞍部都是兩個方向向上，兩個方向向下。

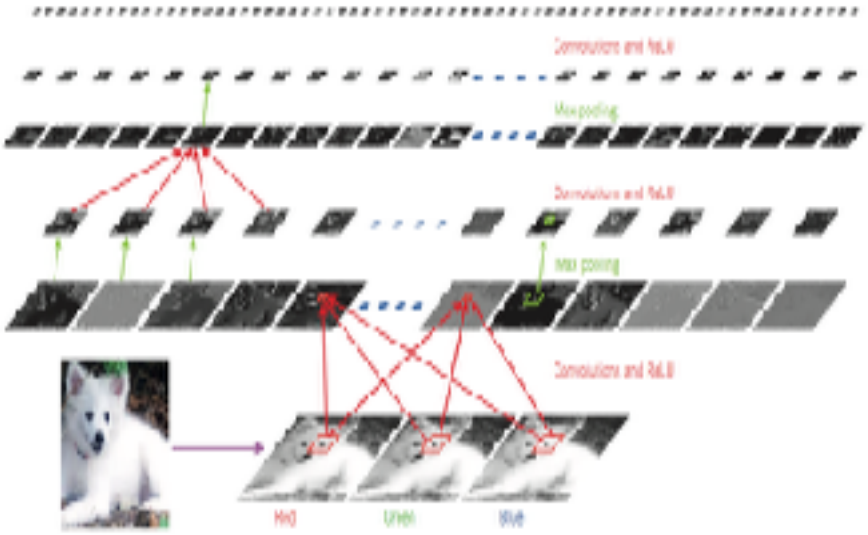
研究結果也顯示這些少數梯度向下的高維鞍部都有接近的 cost value。因此不管卡在哪一個鞍部，結果都差不多。這意味著起始條件不會影響大型神經網路的最終訓練結果。

Stochastic Gradient Descend (SGD)



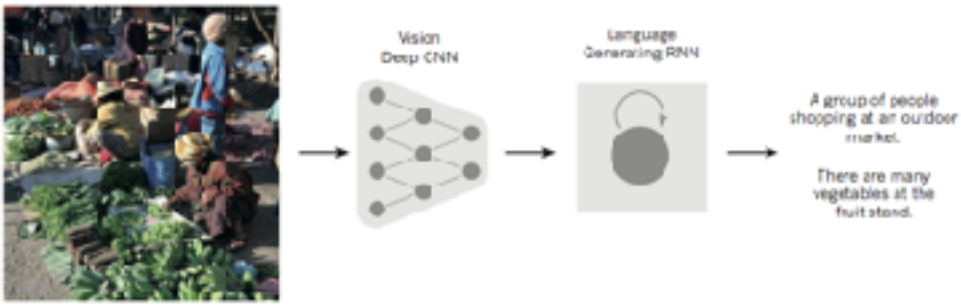
Why deep ？

為什麼多層的神經網路訓練的速度比較快，學習的效果也比較好？ 因為較高層的表徵可以透過較低層的表徵輸入來組成。對於淺層的神經網路來說，最後做classification的表徵總是由一大群input weight 以及一個 activation function 組成。大量的input unit 不代表大量的訊息或是大量的資訊。重要的資訊常常只在少數的維度裡。深度學習每層unit的數量通常是層層遞減。原因就是在做 dimensionality reduction。與PCA 相較，deep learning 在 dimensionality reduction 上的表現較好。Deep networks 在dimensionality reduction 上收斂較 shallow networks 快。 原因可能是自然界中的data中的資訊 具有階層性hierarchy。舉例來說，音素構成單字，單字構成句子，句子構成語法，語法構成文法，文法構成意義。我們人腦所理解的意義，是自然界 data 經過一層層的表徵，或是一層層的filter 得到的最有價值的資訊。 如果自然界中所有的data不經過濾直接被我們接收，高價值的資訊將淹沒在無意義的噪音裡。同樣的，在圖像辨識上，我們關心的不是每一個pixel 的顏色，或是一些顏色接近的pixel構成的線條，或是兩條線條所組成的角，對我們來說有意義的是圖像中的 object， deep learning 利用自然界表徵的 hierarchy，把我們熟悉的object 用比較具象，比較沒資訊的邊緣，轉角等表徵組成，而非用完全沒資訊的pixel value 組成。



Convolutional Neural Networks

deep learning 因為可以習得這樣有效的表徵, 也代表 deep networks 可以經過 pre-training 再fine-tune到每一個問題上。pre-training 可以避免在資料量較少的狀況之下 overfit 。因為蘊含在 deep hierarchical network structure 中的訊息是普適的。



Recurrent Neural Networks and Natural Language Processing