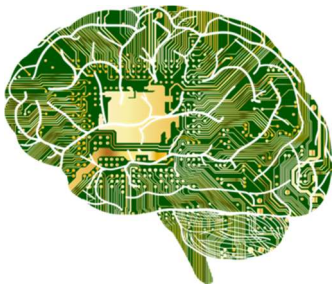
統計專題分析

運用人工智慧深度學習提升統計編碼成效分析
~以循環神經網路(RNN)演算法
實作行業詞義解析編碼為例



撰寫人：黃河川
高雄市政府主計處
112 年 4 月

壹、前言

人工智慧發展一日千里，不僅提升生產效能，亦讓人類享受便利科技生活，近來發展突飛猛進，主係機器學習技術精進突破，其中又以深度學習演算法影響深遠，舉凡聊天機器人、自動駕駛、自動化生產、智慧醫療等，均可見成功運用實例，而各界投入方興未艾，未來發展無可限量。

統計編碼工作係將資料依特性分類編碼，以作為計算運用基礎，如統計就業失業情形之人力資源調查、薪資統計之受僱員工薪資調查、工業及服務業普查、人口及住宅普查等，須將受訪對象所從事之行業依行政院主計總處頒布之「中華民國行業標準分類」判定編碼，其正確性影響統計結果及經社施政甚鉅，為非常重要之統計基礎工作。

陳述行業內容文字易因人而異，且囿於統計工作人員之勤前教育訓練時間有限，統計編碼能力未臻純熟，故需投入可觀人力時間於解讀詞義及正確編碼。為提升工作效率及編碼確度，本文擬運用人工智慧深度學習之循環神經網路(RNN)演算法，解析行業詞義並輔助統計編碼，其成效除可精進本府統計工作，亦可作為推動市政智慧數位治理之參考案例。

貳、現況描述

一、統計編碼係由自然語言文字轉換為分類編號之工作，107 至 111 年本市計完成約 51 萬筆行業資料統計編碼，動用逾千人，費時逾 4 千工作小時，所費人力時間龐大

調查統計常需蒐集受查對象行業資料，其目的係為了解個人從事工作或廠商營業範圍之行業類別，是判斷社會產業走向重要統計項目。

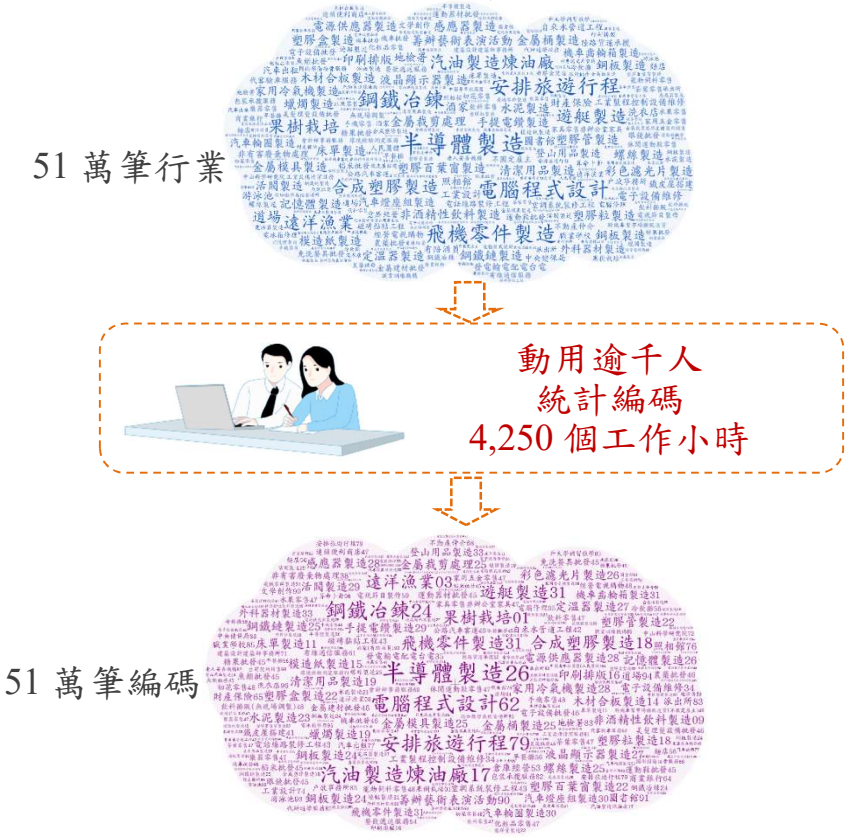
目前蒐集行業資料作業多由各統計調查訪問員親訪受查對象後，將實際行業內容述之於文字，後由審核人員按行業內容文字，依據「中華民國行業標準分類」以人工方式判定分類編號。

行業內容文字之陳述易因人而異，以冷飲店為例，可能書寫文字有冷飲店、冰果店、飲料店、飲料供應...等，此情形為自然語言之正常現象，描述相同事物之文字易因個人學程經歷、習慣

用詞、區域風俗等不同而異。

為求行業歸類精確，審核人員須以人工方式，逐一依訪問員提供行業內容文字判定行業分類編號 2 碼，估算每筆平均約需花費 30 秒[1]，本市 107 至 111 年計完成約 51 萬[2]筆行業資料[3]統計編碼，動用人力逾千人[4]，估計費時約 4,250 工作小時，所費人力時間龐大，且易因經驗程度不同造成確度差異。

圖 1、107 至 111 年高雄市行業編碼工作示意圖



二、AI 深度學習對文字、圖像、聲音等資料自行擷取特徵，應用在自然語言處理、視覺及語音辨識等領域效果良好

AI 深度學習經由讀取文字、圖像或聲音，以仿大腦神經元運

---

[1] 行業內容文字判定其 2 碼行業分類編號平均每筆約需花費 30 秒，係依本文作者實務作業時間估算。

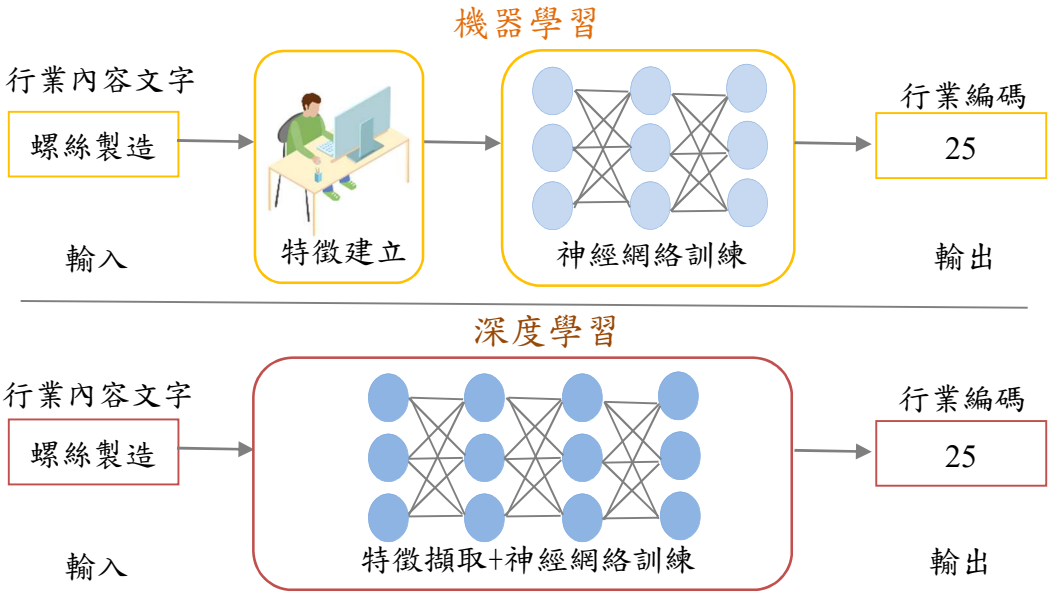
[2] 為提升統計數字易讀性，本文數字均調整以合宜單位、整數或小數呈現，因四捨五入關係，容或有尾差。

[3] 107 至 111 年計蒐集 51 萬筆行業資料包含每月配合行政院主計總處辦理就業率相關之人力資源調查 22 萬筆、本市「109 年人口及住宅普查」13 萬筆及本市「110 年工業及服務業普查」16 萬筆。

[4] 107 至 111 年統計編碼動用人力含每月配合行政院主計總處辦理就業率相關之人力資源調查計 5,040 人次、本市「109 年人口及住宅普查」計 1,524 人及「110 年工業及服務業普查」計 1,035 人。

作原理，經學習自行擷取特徵，有別於傳統機器學習係由人工建立特徵[5]，可大幅節省人力及提高確度，常見之深度學習模型如循環神經網路(RNN)、卷積神經網路(CNN)等，已被廣泛應用在自然語言處理(Nature Language Processing, NLP)、視覺及語音辨識等領域，並取得極好效果。

圖 2、機器學習與深度學習差異示意圖



三、循環神經網路(RNN)於自然語言處理已臻成熟穩定而廣受採用，故本文用以建構 AI 統計編碼模型

循環神經網路(Recurrent Neural Network，RNN)廣泛運用在自然語言處理，主因其將文本分詞[6]後，各字詞[7]循序逐一輸入解析之流程，與人類接收處理文字訊息之模式相似，且理解辨識詞義能力佳，技術已臻成熟穩定而廣受採用。故本文採用循環神經網路(RNN)建構 AI 統計編碼模型，經訓練學習後獲得判定行業分類編號能力，並利用電腦可穩定快速處理大量資料之優勢，提升統計編碼工作效能。

[5] 人工建立特徵：如欲機器學習辨識貓與狗，須人工分別建立貓與狗的特徵，以作為其辨識依據。

[6] 文本分詞係指將文字資料拆解成電腦處理的最小單位，如將「螺絲製造」切割為「螺」、「絲」、「製」、「造」。

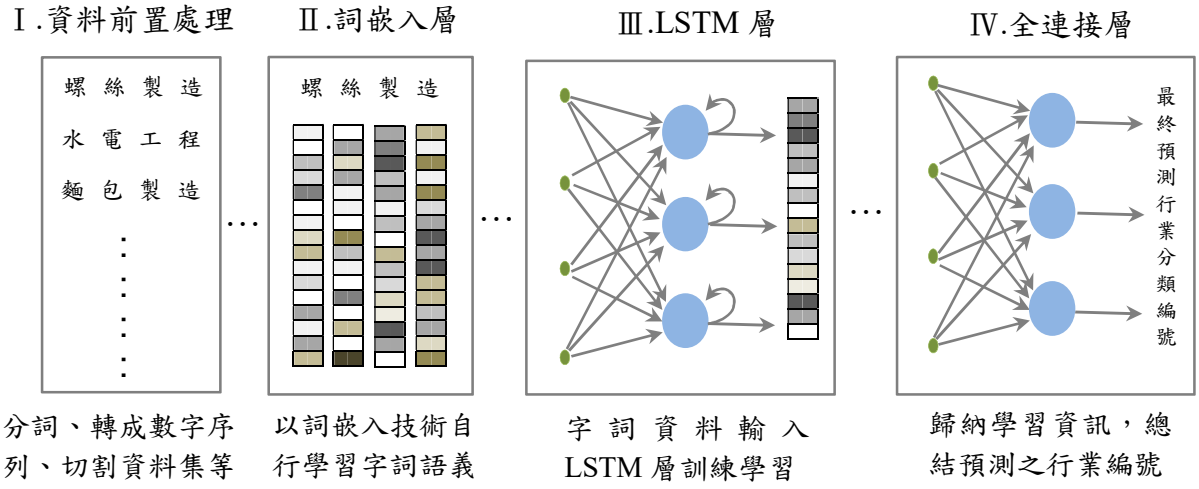
[7] 字詞係能獨立運用的最小語言單位。

叁、行業編碼深度學習模型架構

為解決人工逐一判定行業分類編號耗費龐大人力時間及確度穩定性之問題，本文採自然語言文字處理技術成熟穩定之循環神經網路(RNN)為主體，並以其中表現優異之長短期記憶模型(Long Short-Term Memory，LSTM)建構具人工智慧編碼能力之神經網路。

本模型架構分為 4 部分，依序分為資料前置處理、詞嵌入層、LSTM 層及全連接層，架構示意圖如圖 3，其中各細部內容分述於下。

圖 3 行業詞義解析編碼神經網路架構



Ⅰ.資料前置處理

本部分係將行業內容文字資料先行處理，以整理為模型運用之數字序列，可分為文本分詞等 5 個步驟，依序說明如下。

I-1.文本分詞

本步驟係將行業內容文字(文本)切割成多個字詞，因描述行業內容文字以詞組[8]為主(如螺絲製造)，詞組由多個字詞組成，為利模型解讀詞組中各字詞與行業分類編號之關係，需將行業內容文字切割成多個有意義的字詞，如可能將「螺絲製造」切割為「螺」、「絲」、「製」、「造」，或「龍眼栽培」切割為「龍眼」、「栽培」[9]，此作業稱為文本分詞(Text Segmentation)。

[8] 按教育部國語辭典定義，詞組係按照一定的語法規則由兩個或兩個以上的實詞組成的句法單位。

[9] 分詞原則依所使用之分詞模組不同而異，分詞後各字詞之字數可能不同。本文說明範例或與實際分詞情形不同。

I-2.文本轉成數字序列

本步驟係將文本於分詞後轉換為索引數字序列，如將文本分詞所得不重複之字詞依序編號建立字典索引，則每個字詞都有唯一的索引編號，將原行業內容文字改以對應之索引編號呈現，即為該文本之數字序列，例如螺絲製造之索引數字序列為[62,38,25,74]，係因「螺」、「絲」、「製」、「造」於字典中的索引數字編號分別為 62、38、25、74。

表 1、行業內容文字文本轉成數字序列範例

序號	行業內容文字	文本分詞	6 碼數字序列補 0	行業分類編號
1	龍眼栽培	龍眼　栽培	[0,0,0,0,13,67]	01
2	螺絲製造	螺　絲　製　造	[0,0,62,38,25,74]	25
3	麵包製造	麵包　製　造	[0,0,0,86,25,74]	08
4	螺絲製造	螺　絲　製　造	[0,0,62,38,25,74]	25

I-3.文本數字序列固定長度

本步驟係將文本之索引數字序列固定長度，因不同行業內容文字長度不一，轉成數字序列之數字編號數量亦不同，為利模型運算，將取合適數字編號數量作為固定長度，如文本轉成數字序列之長度少於所設定之固定長度，則以數字 0 取代空格處(Zero Padding)，如採固定長度為 6 碼之範例如表 1。

I-4.建立行業分類獨熱編碼(One hot encoding)表

本步驟係為行業內容文字建立獨熱編碼表，以做為模型最佳化之目標，作法為將不重複行業分類編號拆成多個行(column)，當某列行業內容文字正確編碼為該行之行業分類編號時則填入 1，反之則填入 0。以表 1 為例，因有 3 種行業分類編號，N 為 3，其獨熱編碼如表 2。獨熱編碼係標示該筆行業內容文字之正確解答，以表 2 螺絲製造為例，其獨熱編碼為[0,0,1]，代表螺絲製造之行業分類編號正確解答為 25。

表 2、行業分類之獨熱編碼表範例

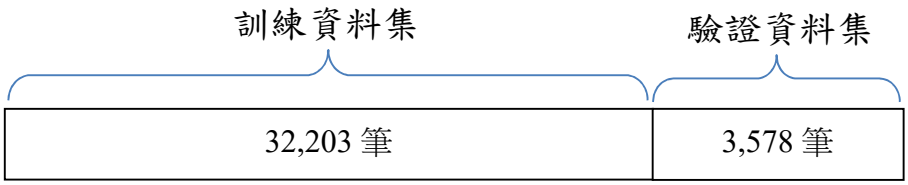
行業分類編號 行業內容文字	01	08	25
龍眼栽培	1	0	0
螺絲製造	0	0	1
麵包製造	0	1	0
螺絲製造	0	0	1

I-5.切割訓練及驗證資料集

本步驟係為驗證模型學習能力，將所蒐集行業資料切割為訓練及驗證資料集，其中模型經由已知資料(訓練資料集)自行學習如何解析詞義判定編碼。將前述自行學習編碼能力套用於驗證資料集，觀察其判定正確情形，以了解模型預估未知資料之能力。

本文將所蒐集全部資料集 3 萬 5,781 筆切割 90%為訓練資料集(計 3 萬 2,203 筆)，10%為驗證資料集(計 3,578 筆)。

圖 4、模型訓練及驗證資料集



Ⅱ.詞嵌入層(word embedding)

本部分係將文字數字序列之數字序號轉為張量(Tensor，係為 N 維向量)，俾以詞嵌入技術解析詞義。例如「螺絲製造」之數字序列為[0,0,62,38,25,74]，要讓神經網路能夠處理行業內容文字，需將各字詞的數字序號(如 25、74 等)，轉為張量，為利理解，以 N=2 之 2 維向量說明詞嵌入觀念， N 大於 2 維之張量可同理類推。

以 6 個蔬果相關字詞舉例，如隨機分配其 2 維向量數值，其情形可能如圖 5，因係隨機分配，故向量數值無意義，無法由各字詞所在空間位置區分歸類。理想配置例子如圖 6 所示，以水果或蔬菜(維度 1)、液體或固體(維度 2)區分，給予合適的向量數值，則使維度、向量數值大小、距離皆有意義，以此反應其語意差異。

例如蘋果汁因屬水果類及液體，分別調降維度 1 數值及調升維度 2 數值；空心菜因屬蔬菜類及固體，分別調升維度 1 數值及調降維度 2 數值，則可經由維度數值大小區分語意差異。

神經網路會依如圖 5 轉換至圖 6 之調整方式，自動調整每個字詞之張量(N 維向量)中各維度數值(調整方式於「肆、一、模型運作方式」段落說明)，找到潛藏於大量詞彙中之語義關係，經學習提升預估準確度。

圖 5、詞嵌入 2 維向量之初始隨機分配數值示意圖

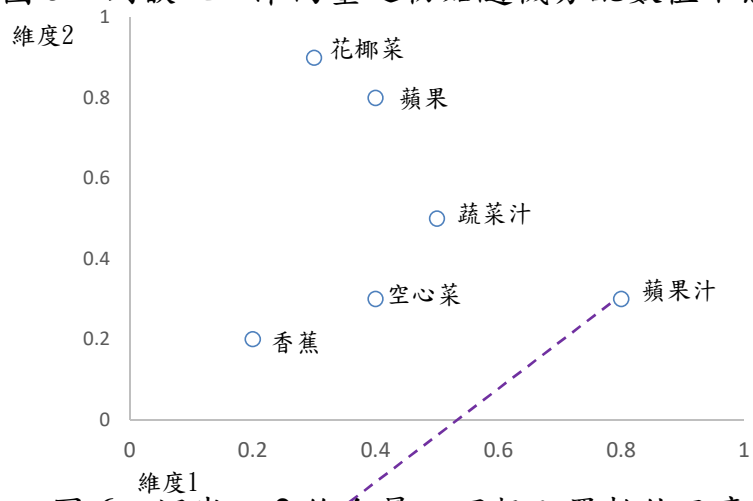
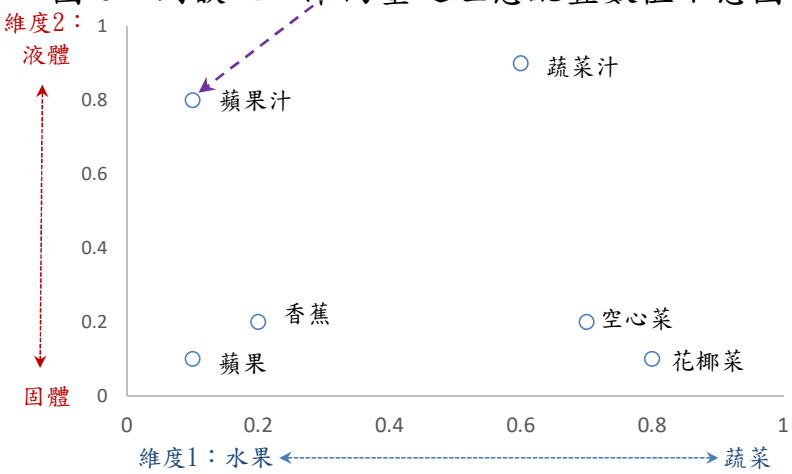


圖 6、詞嵌入 2 維向量之理想配置數值示意圖



Ⅲ.長短期記憶模型 LSTM(Long Short-Term Memory)層

本部分係模型讀入文本數字序列之張量資料，經調整記憶及遺忘等參數以達預估最佳化。LSTM 模型中各神經元有 3 個閘門(Gates)控制神經元在不同時間點的記憶狀態(運作示意圖詳附錄)，名稱及其功能分述如下：

輸入閘(Input Gate)：決定目前的輸入有哪些應該要處理。

遺忘閘(Forget Gate)：決定神經元要遺忘哪些記憶。

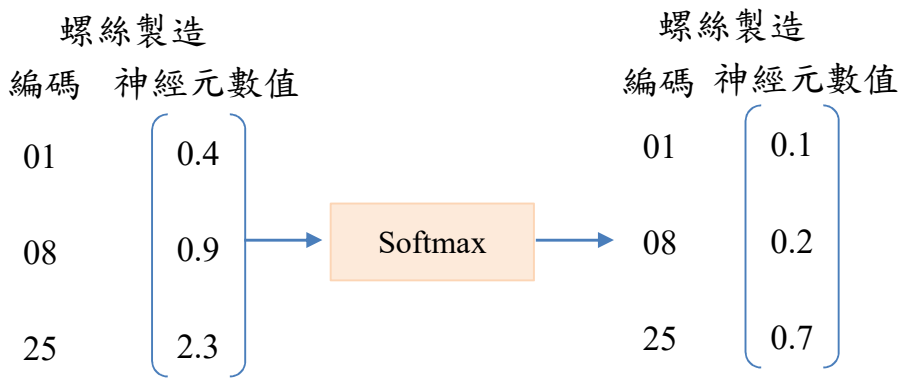
輸出閘(Output Gate)：決定更新後的記憶狀態有多少要輸出。

Ⅳ.全連接層（Fully Connected Layer）

本部分係模型依據所有神經元資訊，歸納各行業內容文字相對各行業分類編號之機率。全連接層使用 Softmax 函數將本層所有神經元裡數值正規化(Normalization)，使其全壓縮為 0 至 1 之間的範圍，並讓它們的和等於 1，運算結果可將各神經元數值解釋為模型預估對應各行業分類編號之機率，數值最高者即為模型對行業內容文字之預估行業分類編號。

如以表 2 之螺絲製造為例，全連接層神經元計 3 個(因僅有 3 個結果)，假設原神經元數值分為[0.4、0.9、2.3]，經 Softmax 函數轉換為[0.1、0.2、0.7]，其中 0.1、0.2、0.7 即系統預測螺絲製造相對行業分類編號 01、08、25 之機率。數值最高之 0.7 對應之行業分類編號 25，即為模型對行業內容文字「螺絲製造」之預估行業分類編號。

圖 7、全連接層 Softmax 運作示意圖



肆、行業編碼深度學習模型運作方式與效能調校

前章說明行業詞義解析編碼之神經網路模型架構，為求模型最適效能，將於本章說明模型運作方式及效能調校過程重要結果。

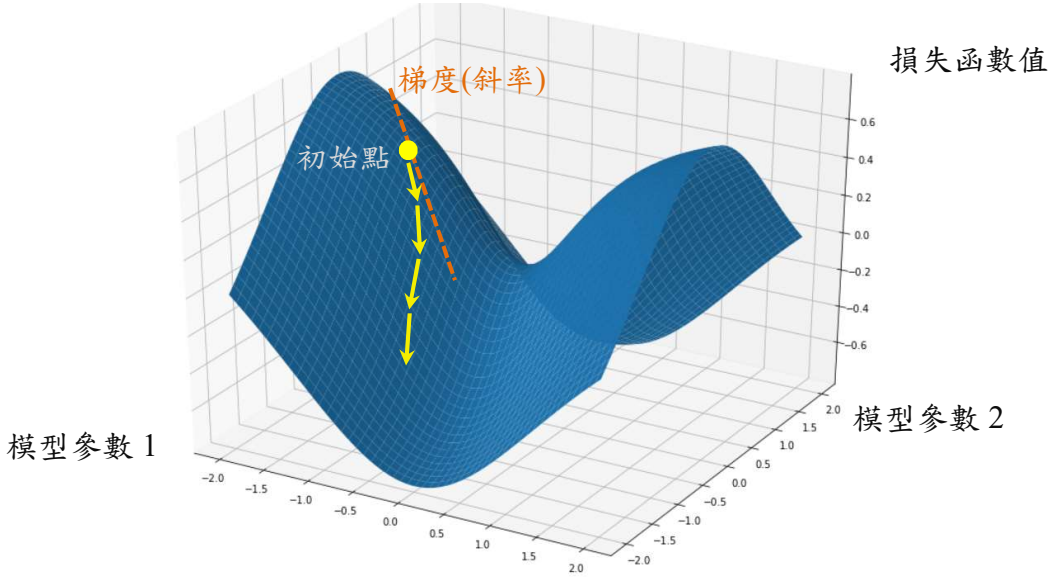
一、模型運作方式

透過參數調整梯度下降方式尋求模型預估錯誤率最小，以達 AI 行業編碼準確率最大化之目標

本模型經前述架構完成後，共計 323 萬 8,234 個自動調整參數，為讓模型自動學習及衡量自身預測表現，設定損失函數(loss function)計算正確解答與模型預測結果之差距(損失函數計算範例詳附錄)。

模型以尋求最小錯誤率(即損失函數值)為目標，自動調整模型參數，調整方式為梯度下降法(gradient descent)，經計算各參數對損失函數之梯度(斜率)，據以修正參數，使新參數所得損失函數值持續下降，以達 AI 行業編碼準確率最大化之目標(模型調整目標及梯度(斜率)計算公式詳附錄)。

圖 8、模型最小化損失函數值調整示意圖

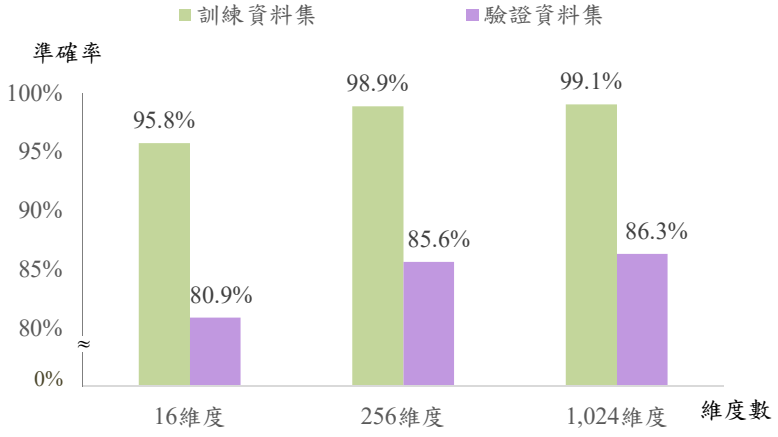


二、模型效能調校

模型設定愈細緻的詞義辨識能力，預估準確率愈高

深度學習於自然語言處理以詞嵌入(字詞張量)技術辨識文字詞義，為了解字詞張量維度對模型成效之影響，試分別以字詞張量 16、256、1,024 維度進行模型訓練，觀察訓練 10 次後之準確率情形，可發現隨字詞張量由 16 維度增加至 1,024 維度，訓練資料集及驗證資料集準確率分別增加 3.3 及 5.4 個百分點，準確率隨字詞張量增加而上升，係因愈高之字詞張量維度代表模型具備愈細緻的文字詞義辨識能力，模型設定愈細緻的詞義辨識能力，預估準確率愈高。

圖 9、模型準確率-依不同字詞張量維度分



資料來源：本文自行整理

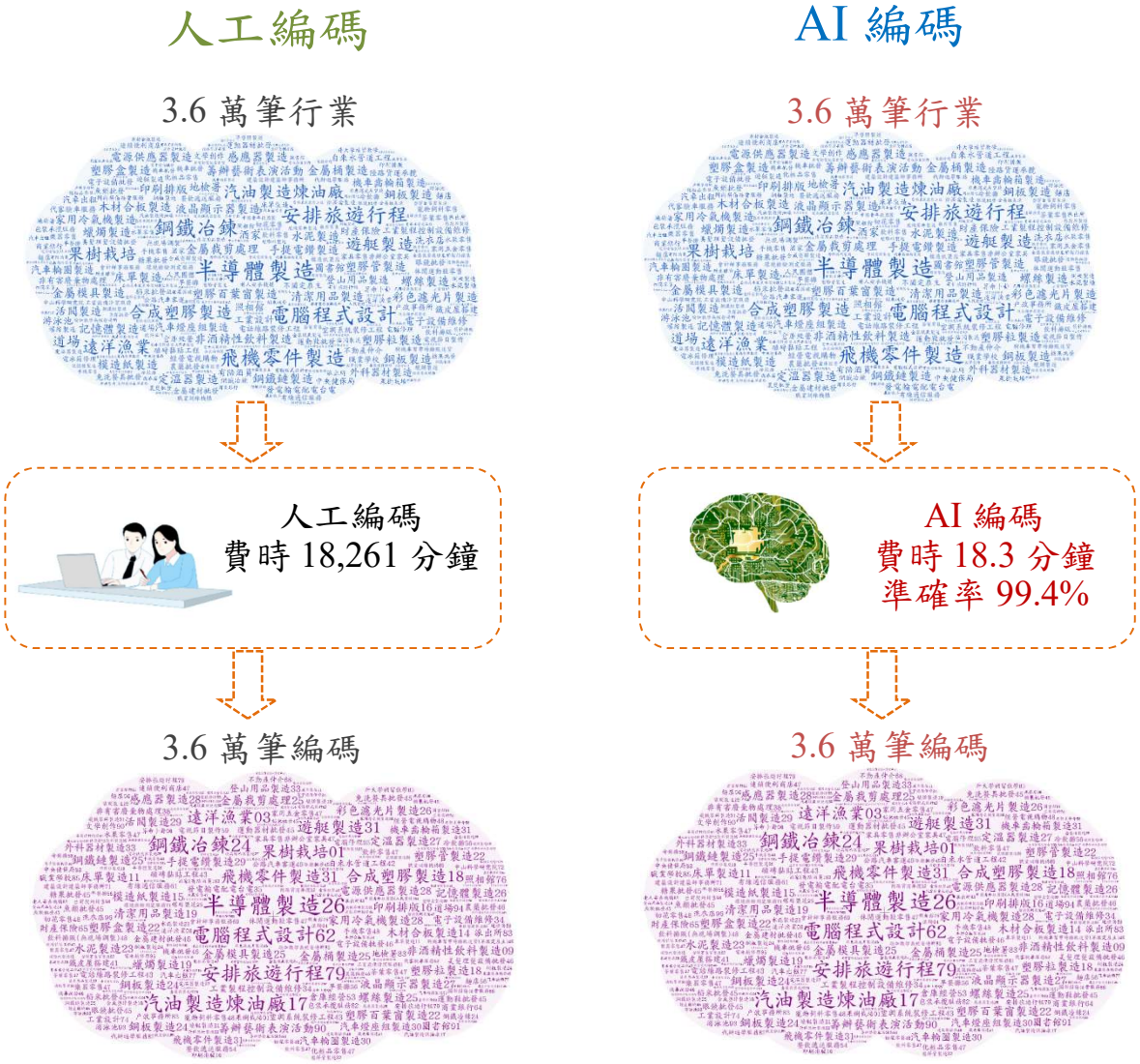
伍、模型建置成效分析

經前述模型架構、運作方式及效能調校作業後，本章分析模型於輔助統計編碼之成效。

本模型對 111 年本市人力資源調查約 3.6 萬筆行業進行測試，AI 編碼作業時間較人工快近千倍，準確率達 99.4%，大幅提升工作效能

為評估本深度學習模型實務編碼成效，對 111 年本市人力資源行業編碼約 3.6 萬筆進行測試，計費時 18.3 分鐘(平均每筆費時 0.03 秒)，較人工編碼費時 18,261 分鐘，快近千倍，模型預估準確率達 99.4%，大幅提升工作效能。

圖 10、模型輔助統計編碼成效-以 111 年高雄市人力資源調查行業編碼為例



陸、建立 AI 模型成效與未來展望

統計編碼係由自然語言文字轉換為分類編號之工作，107 至 111 年本市計完成約 51 萬筆行業資料統計編碼，動用逾千人，費時逾 4 千工作小時，所費人力時間龐大。

本文執筆之時，適逢各大資訊業者相繼推出功能強大聊天機器人，其自然語言處理能力及生成式人工智慧驚艷世人，或將引領人類進入強人工智慧時代，本文運用相似技術原理建構模型，試以創新方法大幅提升工作效能，茲將建立 AI 模型成效與未來展望陳述如下。

一、 建立 AI 模型成效

(一) 本文模型由本處研究人員自行開發，節省委外開發經費，模型採自然語言處理效果良好，且技術成熟穩定之循環神經網路(RNN)為主體，並以其中表現優異之長短期記憶模型(LSTM)建構具人工智慧編碼能力之神經網路，適合運用於大量文字判定分類。

(二) 實務應用編碼成效部分，本模型對 111 年本市人力資源調查約 3.6 萬筆行業進行測試，AI 編碼作業時間較人工快近千倍，準確率達 99.4%，大幅提升工作效能。

二、 未來展望

(一) 推廣人工智慧技術於市府各項適用業務，促進市府數位治理及數位轉型能量，以達智慧城市目標

本案人工智慧深度學習技術應用於自然語言處理效果良好，本府文案內容分類業務如 1999 話務及市長信箱分工、本府來函公文分工、網路留言語意解讀、調解案類型分類等，均可見運用契機。

而本案由本處研究人員自行開發之模型可供各機關參考，期藉由跨機關共同合作，運用人工智慧實現數位治理，除可降低委外開發經費外，並能有效提升施政效能，促進本市成為智慧城市之典範。

(二) 協助行政院主計總處推廣本文建置模型至各縣市運用，提升全國調查統計資料品質及工作效能

本案應用於調查統計之行業判斷效益高，各縣市同本市均投入可觀人力時間於行業判定編碼工作，期可協助行政院主計總處推廣本文所建置之行業編碼深度學習模型至各縣市運用，擴大本文建置模型效益，提升全國整體統計編碼工作確度及效能。另本文模型除原輔助行業編碼功能，應可在酌修模型內容後，適用如職業編碼、受查對象特性核校(如性別、經濟能力等)方面之應用，全面性促進調查統計資料品質及工作效能。

(三) 後續將開發使用操作系統，結合本模型人工智慧演算法，俾利提升本案實務應用價值

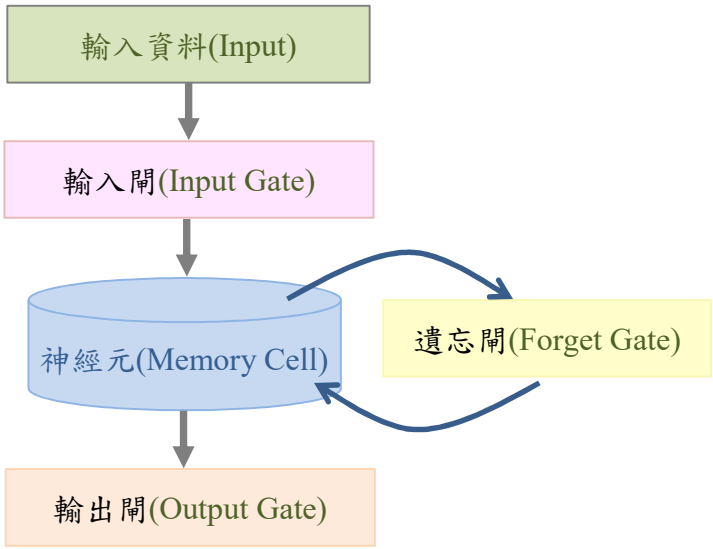
本案業已完成統計編碼資料蒐集、資料前置處理、AI 模型建置、效能調校等作業，並經實務應用成效分析，驗證模型可大幅提升工作效能。為利各級統計工作人員實際執行統計編碼工作，後續將開發使用操作系統(如設計操作介面、查詢方式、結果輸出格式等)，結合本模型人工智慧演算法，讓使用者快速完成編碼工作，俾利提升本案實務應用價值。

柒、參考文獻

1. 臺灣大學李宏毅教授 Deep Learning 網路課程，https://www.youtube.com/channel/UC2ggjtuuWvxrHHHiaDH1dlQ/playlists
2. mikechenx，Keras 自然語言處理(NLP)應用，https://ithelp.ithome.com.tw/articles/10244283
3. mikechenx，Keras 模型、函數及參數使用說明，https://ithelp.ithome.com.tw/articles/10191725?sc=iThelpR
4. Meng Lee，進入 NLP 世界的最佳橋樑：寫給所有人的自然語言處理與深度學習入門指南，https://leemeng.tw/shortest-path-to-the-nlp-world-a-gentle-guide-of-natural-language-processing-and-deep-learning-for-everyone.html
5. sarahwei0804，Gated recurrent unit (GRU)，https://ithelp.ithome.com.tw/articles/10308563
6. 在 NLP 中理解 CNN，https://www.getit01.com/p20181206746531725/

附錄

1. LSTM 層運作示意圖



2. 損失函數(交叉熵 cross entropy)計算範例

交叉熵(cross entropy)係計算兩機率分佈之差距，適合作為分類問題之損失函數，故本模型以交叉熵作為損失函數。以模型預測「螺絲製造」行業分類編號相對為 01、08、25 之機率為[0.1、0.2、0.7]為例，交叉熵經計算為 0.5146。

螺絲製造 編碼	正確解答 p		螺絲製造 編碼	預測結果 q
01	0	$f(p,q) = -\sum_{i=1}^{3} p_i * \log_2(q_i)$	01	0.1
08	0	$= -(1 * \log_2(0.7))$	08	0.2
25	1	$= 0.5146$	25	0.7

3. 模型調整目標及梯度(斜率)公式

*模型調整目標：Min(損失函數 cross entropy $f(p,q)$)

$$= \text{Min}\left(-\sum_{i=1}^{32,203} p_i * \log_2(q_i(x))\right)$$ 32,203 係訓練資料集筆數

*梯度(斜率)：$\frac{\partial}{\partial x_j}\left(-\sum_{i=1}^{32,203} p_i * \log_2(q_i(x_j))\right)$

x_j為模型各參數，$j = \{1、2、3、....、3,238,234\}$

p_i為正確解答之機率；，q_i為模型預測之機率