國立臺灣大學電機資訊學院電信工程學研究所

碩士論文

Graduate Institute of Communication Engineering

College of Electrical Engineering and Computer Science

National Taiwan University

Master Thesis

結合關鍵用語擷取與口述詞彙偵測之影像辨識

Image classification by combining key term extraction and spoken term detection

林賢進

Hsien-chin Lin

指導教授：李琳山 教授

Advisor: Lin-Shan Lee, Ph.D.

中華民國106年7月

July, 2017

doi:10.6342/NTU201702938

誌謝

碩士生涯比想像中還要快上許多走到了終點，一路上研究遇到了各種阻礙，感謝許許多多的人的協助，才能夠完成這本論文。

感謝李琳山教授的指導，老師總是能夠準確的給出研究進行的方向，也給予了非常多的建議，使我茅塞頓開。同時也以身教言教，讓我學到做學問的態度與精神。 感謝李宏毅教授給予的協助，宏毅哥真的非常强大，許多天馬行空的幻想往往都是精彩研究的種子。

感謝實驗室同學們的協助，豪哥、柏瑜時常給予我電腦與程式方面的協助，仰德則是以救世主的姿態拯救我。與資偉討論籃球是實驗室生活一大調劑，批改家維跟永哲的數位語音考卷更是樂趣無窮。育軒哥更是從大二一路凱瑞我到現在，與家翔在Matlab實作課程互相扶持更是難忘的回憶。

感謝實驗室學弟妹們各種創意，祝福你們研究順利。

感謝家人支持我，看著我時常熬夜趕進度、聽著我各種因為自己太蠢而發的牢騷。感謝柔安老師雖然聽不懂我在講什麼但還是聽我解釋各種研究上所遇到的難題，以及各種陪伴。

感謝大翰平戰隊荒謬絕倫的支持。感謝Roger、紹安、Rich、巴風特等爐石或暴雪英霸的實況主，你們讓我的生活無限精彩。

感謝所有人的協助以及陪伴，祝未來一切順利。

doi:10.6342/NTU201702938

摘要

人類幼年時期透過視覺、聽覺就常常直接學到沒有被教導過的詞所代表的東西，進而去理解其相關含意或觀念。本論文希望用類似的方式，讓機器自動從網路上的影音資料中抽取若干知識，也能做到初步的學習。這也是有效運用網路資源的方法之一。例如網路上有廚藝教學影片、生態紀錄片、舞蹈教學影片等，如能有效運用這些資訊，相信對人類生活有很大幫助。

由於網路上的影片，大多缺乏妥善的標註，要讓機器直接學習這些影片並不容易，若是要給予影片標註，則要花費相當大量的人力成本，亦非上策。因此本論文提出了一個系統機制，透過影片旁白的關鍵用語擷取與口述詞彙偵測，自動為影片中的影格標註，同時自動從影片中找出重要的觀念作為類別，再將這些有自動標註的資料作為訓練資料，訓練出一個影像辨識模型，作為走向上述目標的第一步。

doi:10.6342/NTU201702938

Abstract

Children usually learn objects or concepts from visual and hearing input without being exactly taught about those objects or concepts. We hope machines can do something similar, i.e., learn something from unlabeled video and audio autometically. In the Internet era, abundant resources are available on the Internet. For example, the instruction and training videos about cooking, dancing and the environment on YouTube. We wish to be able to use them .

Most of such videos on YouTube mentioned above are not labled, thus difficult to be used in training machines. Human annotation for these videos is expansive. This research therefore proposed a direction and develops a system, which performs key term extraction and spoken term detection over the audio, and uses the detected key terms to label the frames of the video automatically. It can also discover the important concepts in the videos, treating them as classes of images. We then use these labeled data to train an image classification model and reasonably good results can be obtained. A novel key term extraction approach based on the location of the terms and the context in the sentences was also proposed here, which was shown to be domain independent. In other words, once trained it can be used to extract key terms in unseen domains.

doi:10.6342/NTU201702938

Contents

doi:10.6342/NTU201702938

doi:10.6342/NTU201702938

圖目錄

doi:10.6342/NTU201702938

doi:10.6342/NTU201702938

表目錄

doi:10.6342/NTU201702938

第一章　導論

1.1　研究動機

人類在幼年期，具有自主學習的能力，能夠透過視覺、聽覺等方式整合外界的資訊，進而學習到新的知識。在心理學有非常多的言就在分析兒童是如何進行「學習」，例如皮亞杰（Jean William Fritz Piaget, 1896-1980）、維高斯基（Lev S. Vygotsky，1896-1934）等人提出了許多不同的理論。而在本論文中，希望能夠使用類似的方式，使機器能夠自大量的影片中抽取出若干的知識，做到初步的學習。

同時在現代網路上有非常多的資源，涵蓋了各種面向，包含教育、經濟、政治、運動等方面，但這些資料往往缺乏妥善的標註，該怎麼有效率的運用這些資源，是相當重要的研究課題。

而機器學習 (Machine Learning) 的技術提供了一個很好的解決方案。透過設計一個良好的模型架構，給予適量的資料，便能夠得到一個好的模型幫助我們解決遇到的問題。但是在資料的取得往往會遇到非常多的困難，例如資料中有著許多的雜訊、資料量不足、取得資料的成本過大。舉例而言，在YouTube上有非常豐富的影音資料，理論上我們能夠根據這些大量的資料，訓練模型辨識不同動物的物種、烹飪的步驟或是不同的人臉。但是這些資料往往缺乏標註 (label) 造成難以訓練模型，而得花費大量的人力進行標註；或是使用機器自動標註，這樣往往會產生許多的雜訊或是錯誤的標註而造成模型的準確度不足。

因此，在本篇論文中將提出一個系統，能夠自動從影片中習得如何辨識影像，即一張圖片的內容是否屬於某個類別 (Class) 。本系統分為三個部分，分別是關鍵用語擷取 (Key Term Extraction) 、口述詞彙偵測 (Spoken Term Detection) 與影

doi:10.6342/NTU201702938

像辨識 (Image Classification)。這個系統與傳統影像辨識模型的差異主要在於所使用的資料，傳統影像辨識需要資料中每張圖片都有對應到一個類別，而這個過程往往會需要許多的人力，標注每一張圖片所對應的類別。而本系統則是自動從影片中找出語音內容中的關鍵用語作為類別，每個關鍵用語所對應到的影像便被標註為該類別，以此自影片中搜集作為訓練影像辨識模型的資料。

透過這個方式，我們能夠直接使用網路上大量的影片，不需要透過昂貴的人工來標註每筆資料。在本論文中，我們選擇語音內容與影像具有高度相關的影片作為訓練的資料，根據語音內容學出哪些圖片對應到一些重要的觀念 (Concept) ，例如搜集了許多動物星球頻道介紹不同動物的影片，根據旁白的內容擷取出關鍵用語 (例如，貓、拉布拉多、摺耳貓等) 再根據這些關鍵用語所對應到的圖片訓練影片辨識模型。只要給予足夠的影片，便能得到一個能夠分辨不同動物物種的影像辨識模型。

在這個過程中，本論文會解決幾個主要的問題。在關鍵用語擷取中，系統並不知道所有的關鍵用語，例如使用物理、化學的文章訓練模型，但實際上系統得找出廚藝或是密碼學文章中的關鍵用語。本論文將結合監督式 (Supervised) 與非監督式 (Unsupervised) 的技法，設計模型能夠找出沒有出現在訓練資料中的關鍵用語。在口述詞語偵測中，會面臨到影片所帶的聲音有很大量的雜訊，這些雜訊有可能是背景音樂、動物的叫聲或是拍攝過程中自然收錄的噪音，或者是自動語音辨識的錯誤。本論文會以次詞單位 (Subword unit) 如音位 (Phone) 進行檢索，或者混合詞與次詞單位進行檢索，以克服雜訊的干擾或是語音辨識結果的錯誤，進而有效地找出關鍵用語出現的位置。而在影像辨識模型中，得克服錯誤的標註對模型訓練造成的負面影響，本論文將以自體學習 (Self-learning) 與使用自動編碼器 (Autoencoder) 進行預訓練，來解決這個問題。

doi:10.6342/NTU201702938

1.2 研究方向

本論文之研究方向為使用自動關鍵用語擷取與口述詞彙偵測訓練影像辨識模型，主要包含以下幾點：

- 近期關於關鍵用語擷取的研究主要是將此問題當成監督式學習中的分類問題，也就是說每一篇文章所對應到的類別就是該篇文章的關鍵用語。但此模型無法擷取沒有看過的關鍵用語，例如以生命科學的資料訓練模型，此模型能夠有效的擷取生命科學類的文章的關鍵用語，可是當面對旅遊、機器人類別的文章此模型便無法準確地找出關鍵用語，因為這些領域的關鍵用語並沒有出現在生命科學類中，亦即是模型沒有看過的「類別」。而早期關鍵用語擷取的研究則以非監督式學習來處理，使用圖論 (Graph Theory) 或是詞頻 (Term Frequency) 與逆向文件頻率 (Inverse Document Frequency) 等方式來解，監督式學習的方法具有較高的準確率，但其所能擷取的文章類別受限於訓練資料的領域 (Domain) ；非監督式學習的方法訓練出來的模型能夠應用在不同的類別的文章，但其準確率較低。因此本論文將提出一個半監督式學習 (Semi-supervised Learning) 的模型，嘗試能夠訓練出準確度足夠好，同時不受訓練資料的領域的限制，能夠自沒有看過的類別的文章擷取關鍵用語。
- 自影片的語音資料中擷取出關鍵用語後，便使用口述詞彙偵測來尋找此關鍵用語是否有出現在語音資料中的其他位置。因為自網路上擷取的影片，其背景往往會有大量的雜訊，可能是影片的背景音樂、自然界中的雜訊或是動物的吼叫聲。這些雜訊都會使得自動語音辨識的結果出現錯誤，因此在偵測關鍵用語時會使用詞、次詞單位、音位等不同的單位，以嘗試盡可能找出在影片中所有的關鍵用語。
- 最後，將口述詞彙偵測中所得出的關鍵用語的位置與對應到的影像進行配

doi:10.6342/NTU201702938

對，便能夠得到具有標註的圖片資料。以此方式自動標注的圖片很可能會有錯誤，例如影片中主持人在介紹斑馬時，影像可能是關於草原，而若是直接將關鍵用語「斑馬」所在時間對應到的影格標註為「斑馬」，便會產生錯誤的配對。在本論文中，會先以影片中出現過的所有圖片進行自動編碼器預訓練，再以自體學習的方式減少錯誤標註所帶來的負面影響。

1.3 主要貢獻

本論文之主要貢獻如下：

- 本論文提出一個關鍵用語擷取系統，此系統不受限於訓練資料，能夠擷取出沒有出現過的關鍵用語，換言之訓練資料與測試資料的領域可以不相同；相對的，傳統的監督式關鍵用語擷取系統會受限於訓練資料所涵蓋的內容，難以擷取不存在訓練資料中的關鍵用語。同時，本系統的準確率也優於非監督式的關鍵用語擷取系統。最後引入增強式學習 (Reinforcement Learning) 的觀念，增進模型的表現。
- 本論文提出一個系統，透過關鍵用語擷取與口述詞彙偵測自動標注影片中的影像的類別，進而能夠以此訓練影像辨識模型。相較於傳統的影像辨識模型的訓練過程，本系統的訓練資料中不需要包含標註過的圖片，而能夠直接使用影片作為訓練資料。因此能夠減少產生訓練資料的成本、更有效率的使用網路上大量的影片資源。

1.4 章節安排

本論文之章節安排如下：

doi:10.6342/NTU201702938

- 第二章：介紹本論文相關背景知識。
- 第三章：介紹如何進行關鍵用語擷取。
- 第四章：介紹如何以口述詞語偵測訓練影像辨識模型。
- 第五章：介紹如何結合關鍵用語擷取與口述詞彙偵測訓練影像辨識模型。
- 第六章：本論文之結論與未來研究方向。

doi:10.6342/NTU201702938

第二章　背景知識

2.1　深層類神經網路

2.1.1　簡介

在機器學習領域中，透過設計一個好的模型，而能夠學習資料樣本中的關係，著名的模型如線性回歸（Linear Regression）、支撐向量機（Support Vector Machine）等。這些線性的模型的確能夠成功的模擬大部分的資料，但若要模擬更加複雜的資料，非線性的模型往往會是更好的選擇。其中類神經網路是現代非常泛用、相當具有效率的模型。

深層類神經網路（Deep Neural Network, DNN）是透過模仿生物神經網路的結構與功能，進而達到類似生物神經系統具有的平行運算、分散式處理等能力。起初受限於運算量過於龐大，而無法廣泛的運用。但是在近代硬體設備的革新，即透過圖形處理器（Graphocs Processing Unit, GPU）極快的加速平行化的運算，以及後續學者提出更有效率的訓練方法，大幅提升整體的訓練速度，深層類神經網路重新成為機器學習領域中重要的課題。

深層類神經網路主要由許多的神經元（Neurons）或節點（Node）所組成，網路中的神經元是嘗試模擬生物的神經元，從外界環境或其他神經元獲得資訊，經過簡單的計算過程，將結果輸送給外界或是其他神經元。其結構如圖 2.1所示，由許多層串接而成，每一層含有不同數目的節點，根據每層的位置，可分為三類：

- 輸入層（Input Layer）：對應到資料的輸入變數。
- 隱藏層（Hidden Layer）：層數不定，節點的數目也可能彼此不同。
- 輸出層（Output Layer）：對應到資料的輸出變數。

doi:10.6342/NTU201702938

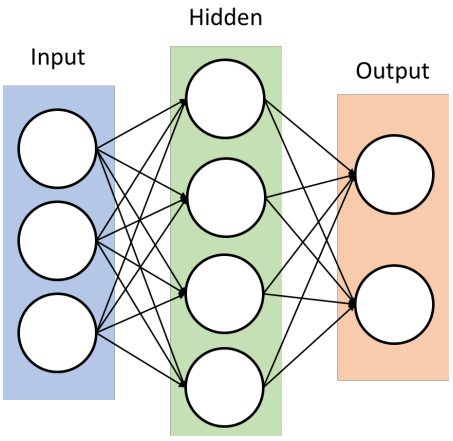


圖 2.1: 深層類神經網路結構示意圖

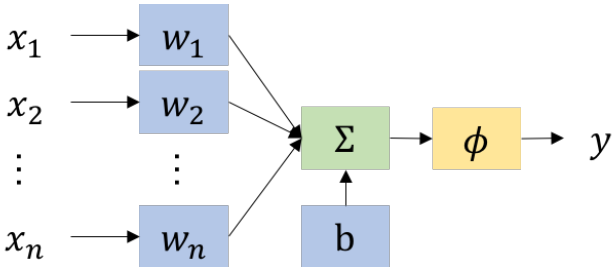


圖 2.2: 節點示意圖

每一個節點的結構如圖 2.2所示，每一個節點包含了加權係數（Weight）、偏移量（Bias）與非線性的活化函數（Activation Function）。一個節點的數學形式可以表示為

$$y = \phi(\sum_{i=1}^{n} w_i x_i + b) \tag{2.1}$$

其中輸入變數$X = [x_1, x_2, ..., x_n]$，偏移量為b，加權係數$W = [w_1, w_2, ..., w_n]$，輸出變數為y，而活化函數ϕ則有許多選擇，通常會使用S型函數（Sigmoid Function）或是整流線性單元（Rectified Linear Unit, ReLU）。而深層類神經網路的學習目的，便是找出最合適的加權係數與偏移量，以求得最好的函式將輸入變數映射（Mapping）到輸出變數。

doi:10.6342/NTU201702938

2.1.2 訓練方法

深層類神經網路最常見的訓練方法為反向傳播演算法 [1]（Back Propagation），同時搭配最佳化演算法（Optimization Method），如梯度下降演算法 [2]（Gradient Descent）。訓練方法的核心想法是當偵測到錯誤（模型的輸出結果（Output）與真正期望目標（Target）之間的差異）後，將參數往錯誤更少的方向更新。

由於深層類神經網路的輸出可能為多維的向量，因此為了同時衡量不同維度的錯誤，便需要定義減損函數（Loss Function）。常見的減損函數有均方差（Mean Squared Error, MSE）與交叉熵（Cross Entropy, CE）。假設深層類神經網路的輸出結果$O = [o_1, o_2, ..., o_N]$真正期望目標為$T = [t_1, t_2, ..., t_n]$，則均方差定義為：

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(o_i - t_i)^2 \tag{2.2}$$

而交叉熵則定義為：

$$CE = -\sum_{i=1}^{N} t_i \cdot ln(o_i) \tag{2.3}$$

通常來說，減損函數的值越大代表模型的表現越差，因此在訓練過程中要調整各種參數，以降低減損函數的值。但深層類神經網路有許多層，每層中又有非常多的節點，如何有效率的調整參數成了一個重要的問題。現在最常用的訓練方法是梯度下降演算法，是反向傳播演算法中最常使用的最佳化方法。

梯度下降演算法是一個找到函數極值的最佳化演算法，若要尋找函數的局部最小值，便在當前點對應到的梯度的反方向前進，透過迭代搜索便有機會達到最小值。假設θ是參數空間上的一點，代表了模型中所有可以調整的參數的集合，包含加權係數、偏移量等，$\theta(i)$為第i次迭代後的參數集合，給定目標函數（也就是想要求得最小值的減損函數）為$E(\theta)$，在點$\theta(i)$仍有定義並可一次微分，那麼在

doi:10.6342/NTU201702938

點θ處沿著$-\nabla E(\theta^{(i)})$移動，也就是梯度的反方向時，函數值下降最快。其更新的過程可以表示成下列式子：

$$\theta(i+1) = \theta(i) + \Delta\theta(i) \tag{2.4}$$

$$\Delta\theta(i) = -\epsilon \cdot \frac{\partial E}{\partial \theta}|_{\theta=\theta(i)} \tag{2.5}$$

$$E(\theta(i+1)) \leq E(\theta(i)) \tag{2.6}$$

ϵ是學習率（Learning Rate），通常選擇0.1到0.0001之間的值，值越大代表每一次更新的幅度越大，學習的速率可能越大，但是若是值太大，更新的路徑可能會成之字形，無法收斂到最小值；若值太小，則學習速度會過慢，因此找到最適（Optimal）的值不是件容易的事，往往得靠經驗來判斷。

梯度下降法能夠有效率地降低減損函數的值，但是有著下列幾項缺點：

- 由於每次更新只根據上一次更新的值，因此可能會停在局部最小值（Local Minimum），而非全域最小值（Global Minimum）。
- 越接近局部最小值，收斂的速度會越慢。
- 有可能會之字形的更新（例如學習率過大）。

因此在訓練模型時必須得相當注意。

2.1.3 卷積類神經網路（Convolutional Nueral Network）

卷積神經網路 [3]（Convolutional Nueral Network, CNN）是深層類神經網路的一種變形，其特色是具有卷積層（Convolutional Layer）與合計層（Pooling Layer），對於輸入資料的二維結構，如影像，有著更好的理解。同時相比於其他的神經網

doi:10.6342/NTU201702938

路，具有較少的參數，是在影像辨識、語音辨識領域中時常使用的模型。以下將介紹其結構：

- 卷積層（Convolutional Layer）

  卷積層包含了一系列的卷積單元，又可稱為核心（Kernel）或過濾器（Filter），每一個卷積單元包含了可訓練的參數，與輸入變數進行點積（Dot Product）便能得到卷積特徵（Convolved Feature），或稱為特徵圖（Feature Map），過程如圖 2.3所示。 其具有局部連結（Local Connected）的特性，每一個卷積單元會學習到一種特徵，較淺層的卷積單元有可能學到角、邊緣等，而較深層的卷積單元則可能學到更抽象的特徵，當圖形中出現該種特徵，則卷積單元便會被激發。

- 合計層（Pooling Layer）

  合計層是一縮減採樣（Downsampling），其中最大合計（Max Pooling）最為常見，做法是將輸入變數劃分為若干子區域，輸出每個子區域的最大值，其過程如圖 2.4所示。這個做法是認為一個重要特徵的精確位置，不如這個重要特徵與其他特徵的相對位置來得重要，因此透過最大合計能夠大幅減少參數，同時有效地避免過度貼合（Over-fitting）。

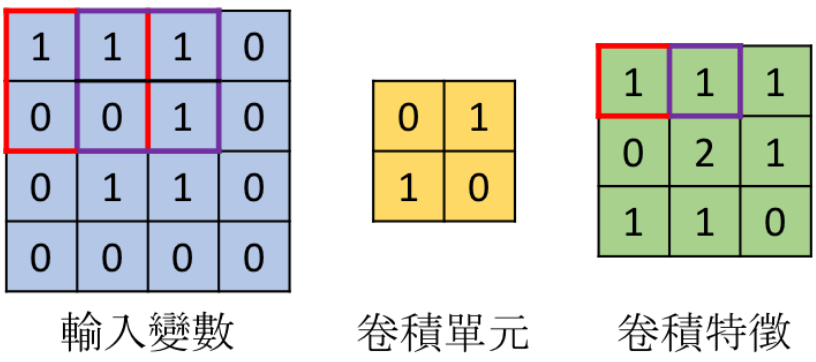


圖 2.3: 卷積層示意圖

doi:10.6342/NTU201702938

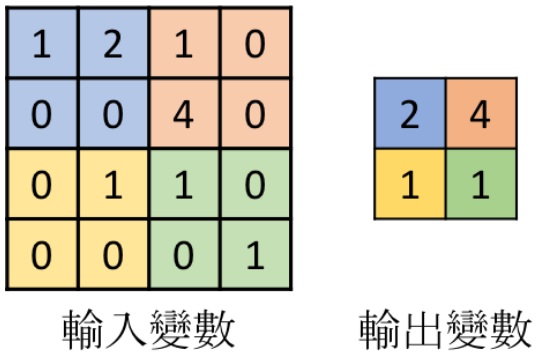


圖 2.4: 合計層示意圖

整個卷積類神經網路的架構如圖 2.5，可視需求排列卷積層與合計層，決定卷積層由多少卷積單元組成以及卷積單元的大小。若輸入的影像大小為$(33, 33, 3)$，第一層卷積層有4個卷積單元，大小為$(3, 3)$，則經過此層後能夠得到特徵映射的維度為$(30, 30, 4)$。緊接著的合計層會將輸入的影像每$(2, 2)$切成一個子區域，經過此層後特徵映射的維度為$(15, 15, 4)$。再經過一層卷積層後，將所有的卷積單元攤平（Flatten）成一個一維的向量，輸入全連通層（Fully Connected Layer），與深層類神經網路裡的隱藏層相同的結構，最後得到輸出類別。

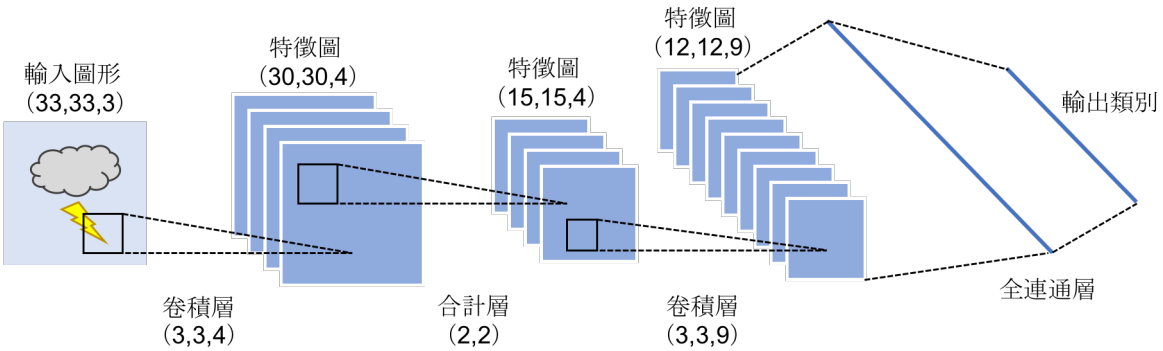


圖 2.5: 卷積類神經網路

doi:10.6342/NTU201702938

2.1.4 長短期記憶類神經網路（Long Short-Term Memory Network）

上面所述的深層類神經網路與卷積類神經網路在面對分類問題（Classfication Problem）時都有很不錯的表現，也就是在輸入輸出變數的維度是固定的情況下，例如要分成十類則輸出變數的維度即為十，輸入圖形的大小為32×32則輸入變數的維度為32×32，上述模型都能有效地尋找資料間的關係。但是當資料的維度不定，例如需要判斷一個句子的情緒是積極還是消極時，輸入的句子長度不見得相同，因此模型難以決定輸入變數的維度。也就是說對於序列（Sequencial）特性模擬，深層類神經網路與卷積類神經網路是有所不足的。 遞迴式類神經網路（Recurrent Neural Network, RNN）便能夠解決資料具有序列狀、上下文關係（Context Dependency）的問題，結構如圖 2.6所示。在t時間的輸入為X_t，輸出為O_t，我們可以觀察到類神經網路除了根據X_t外，同時也會考慮上一個時間點的隱藏狀態（Hidden State）H_{t-1}。藉由這個方式，遞迴式類神經網路可以接受一整個序列的時序資訊，能夠將歷史資訊傳遞下去。

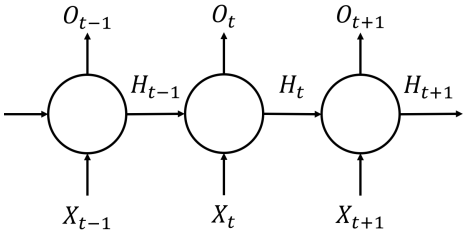


圖 2.6: 遞迴式類神經網路

但是隨著序列越來越長，遞迴式類神經網路會產生梯度消失（Gradient Vanishing）的問題，使得模型無法完全掌握整個序列的資訊，因此提出了長短期記憶類神經網路 [4]（Long Short-Term Memory Network, LSTM）。長短期記憶類神經網路的特色是提出了核心元素細胞（Cell）的概念，透過門限（Gate）來控制核心元素細胞中資訊的新增、修改與刪除，其結構如圖 2.7所示。

doi:10.6342/NTU201702938

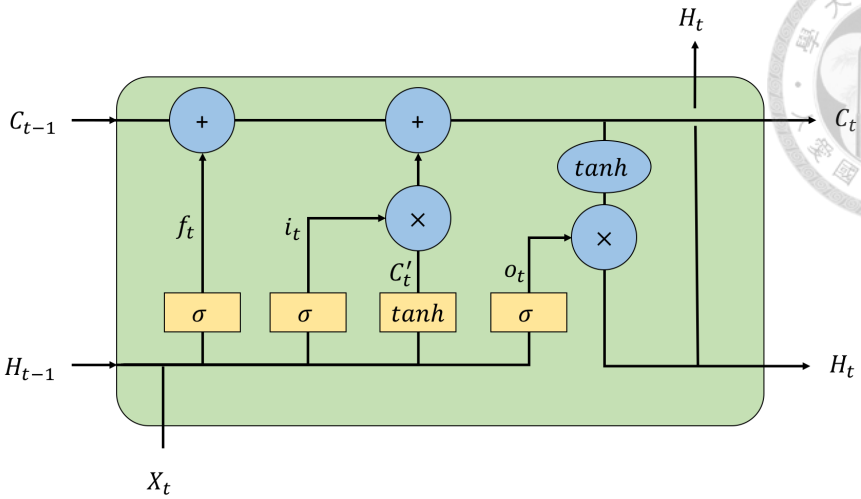


圖 2.7: 長短期記憶類神經網路

首先，透過遺忘門限（Forget Gate）f_t來決定是否要遺忘儲存在元素細胞中的資訊C_{t-1}，上一個時間的輸出H_{t-1}與此刻的輸入X_t分別與加權係數相乘，透過一個S型函數（Sigmoid Function）σ可以得出介於0與1之間的值所組成的向量，0代表捨棄資訊，1則代表保留，將f_t與C_{t-1}進行逐點乘積，便能夠決定哪些資訊要被捨棄，過程如式子 2.7所示。

$$f_t = \sigma(W_{xf}X_t + W_{hf}H_{t-1} + b_f) \tag{2.7}$$

接下來透過輸入門限（Input Gate）i_t來決定要將哪些值存入元素細胞中，而儲存資訊的候選值C'_t則透過一個非線性函數，雙曲正切（Hyerbolic Tangent, tanh）來產生，過程如式子 2.8所示。

$$\begin{aligned} i_t &= \sigma(W_{xi}X_t + W_{hi}H_{t-1} + b_i) \\ C'_t &= \tanh(W_{xC}X_t + W_{hf}H_{t-1} + b_C) \end{aligned} \tag{2.8}$$

將前兩步驟所得到的結果結合，便能夠將元素細胞中的資訊由C_{t-1}更新到C_t，過程如式子 2.9所示。

$$C_t = f_t * C_{t-1} + i_t * C'_t \tag{2.9}$$

doi:10.6342/NTU201702938

最後一步則是產生輸出H_t，其值會受到儲存在元素細胞中的資訊C_t所影響，同時透過輸出門限（Output Gate）o_t來控制輸出的結果，過程如式子 2.10所示。

$$o_t = \sigma(W_{xo}X_t + W_{hi}H_{t-1} + b_o)$$
$$H_t = o_t * \tanh(C_t) \tag{2.10}$$

2.2 關鍵用語擷取

2.2.1 簡介

關鍵用語擷取（Key Term Extraction）在資訊檢索（Information Retrieval）或是自然語言處理（Natural Language Processing）等領域都有非常多的應用，例如可以使用在文字摘要（Text Summarization）、文件分類等。

關鍵用語擷取的目的如表 2.1所示，給予系統一篇文章（可能包含標題與內文），則系統可以截取出關鍵用語，關鍵用語有可能出現在文章中，也有可能沒有出現過；同時，一篇文章可能只有一個關鍵用語，也可能有數個。關鍵用語擷取系統主要分為監督式與非監督式，監督式以分類問題（Classification）來處理關鍵用語擷取，將所有的關鍵用語視為文章的類別，一個文章可能屬於多個類別，透過給予大量被標註關鍵用語的文章訓練模型；而非監督式則會以每個詞的統計特性，考慮詞出現的頻率，例如使用詞頻與反向文件頻率等特徵，或使用圖形基準排序（Graph-based Ranking），此方法的原理近似於計算網頁之間重要性的演算法，將每一個詞視為一個網頁，詞與詞之間共出現（Co-occurence）便視為這兩個詞有所連結，而能夠建立圖形（Graph），之後以隨機漫步（Random Walk）或其他的演算法來求得每一個字的重要程度。以下會將更仔細地講解這方法的細節。

doi:10.6342/NTU201702938

表 2.1: 關鍵用語擷取範例

標題	What are some Caribbean cruises for October?
內文	My fiancee and I are looking for a good Caribbean cruise in October and were wondering which islands are best to see and which Cruise line to take? It seems like a lot of the cruises don't run in this month due to Hurricane season so I'm looking for other good options. We'll be travelling in 2012.
關鍵用語	caribbean cruising vacations

2.2.2 監督式關鍵用語擷取

監督式關鍵用語擷取 [5]是將整個問題視為序列分類（Sequence Classification），換言之一篇文章是一個詞序列（Word Sequence），系統必須得將其分類至所屬的關鍵用語。以表 2.1為例，輸入為包含標題與內文的文章，而系統則必須將其分類為關鍵用語所對應到的類別。

監督式關鍵用語擷取的系統如圖 2.8所示，系統會將文章視為一個詞序列$W = [w_0, w_1, w_2, ..., w_n]$，針對每一個詞抽取特徵向量$V = [v_0, v_1, v_2, ..., v_n]$，如圖中的黃色區塊。接著將這些特徵向量依序輸入模型中，如圖中的藍色區塊，模型通常會使用遞迴式類神經網路，尤其是長短期類神經網路，因其較能有效的使用一長串的文字資訊。模型最後會得到一個包含整段文章資訊的向量o_T，此向量可以視為整段文章的特徵向量。將其輸入一個深層類神經網路，便能得到這段文章所對應到的分類。而所有可能的分類是對應到所有出現過的關鍵用語，每一篇文章可能會屬於一個或一個以上的類別，即每一篇文章可能有一個或一個以上的關鍵用語。

doi:10.6342/NTU201702938

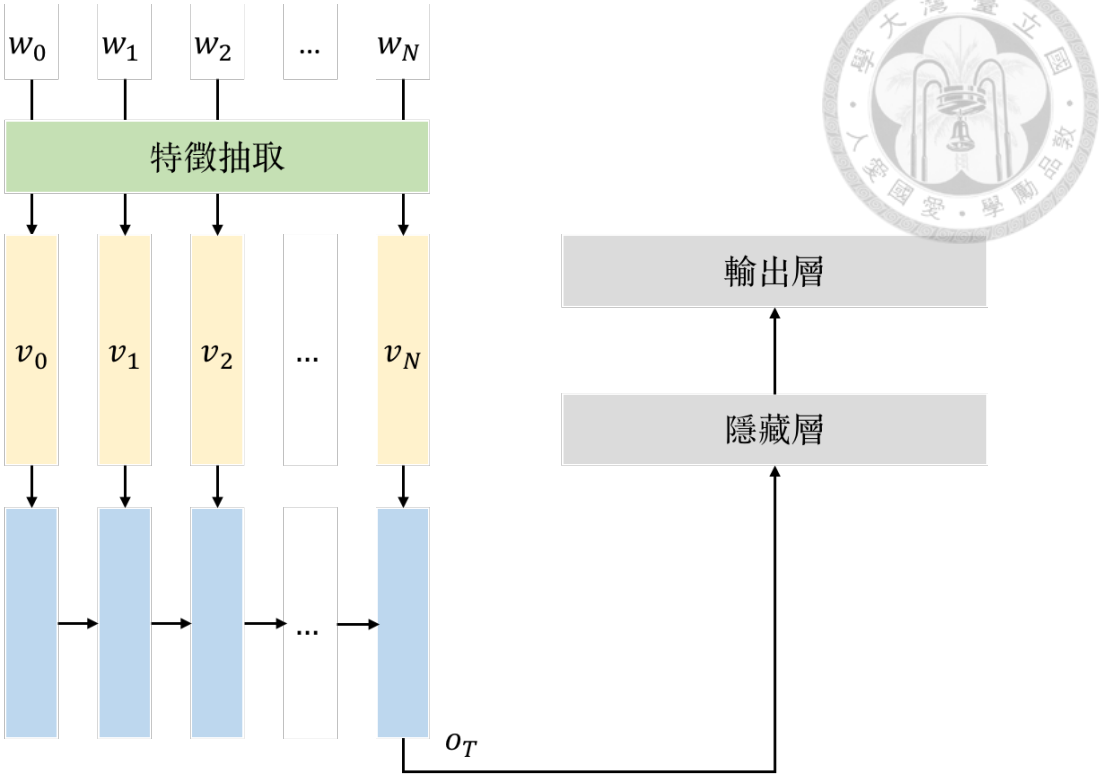


圖 2.8: 監督式關鍵用語擷取系統示意圖

2.2.3 非監督式關鍵用語擷取

非監督式關鍵用語擷取是將整個問題視為排序（Ranking）問題，換言之整篇文章每一個詞都有可能是關鍵用語，透過特徵抽取、圖基礎排序（Graph-based Ranking）等各種方式，將文章中的用語進行排序，擁有越高優先順序的用語，越有可能是關鍵用語。

在特徵抽取的部分，最常見的為詞頻（Term Frequency, TF）與反向文件頻率（Inverse Document Frequency, IDF）。詞頻表示一個詞t出現在文件d的頻率，其常見定義方式有下列幾種：

- 原始詞頻

  計算詞t出現在文件d中的次數，再除以整篇文章的總長度，得到詞t出現在

doi:10.6342/NTU201702938

文件d中的詞頻$tf(t,d)$，數學形式如下式。

$$tf(t,d) = \frac{count(t,d)}{\sum_{t' \in d} count(t',d)} \tag{2.11}$$

- 對數尺度詞頻（Logarithmically Scaled Frequency）

  一個詞的重要性不見得與其出現次數成線性正比關係，其分佈可能與對數函數相似，因此將詞頻改為下式運算。

$$tf(t,d) = 1 + log(count(t,d)) \tag{2.12}$$

- 强化頻率詞頻（Augmented Frequency）

  在原始詞頻中，文件的長度會造成不同文件之間對詞頻值的不同影響，因此將分母改為每個文件中最高頻詞，如下式。

$$tf(t,d) = 0.5 + 0.5 \cdot \frac{count(t,d)}{max\{count(t',d) : t' \in d\}} \tag{2.13}$$

而反向文件頻率則能夠評估一個詞所提供的資訊量，也就是說這個詞如果只出現在特定幾個文件中，代表它很可能是關鍵用語，可以代表這幾個文件；反之，若一個詞廣泛地出現在不同的文件中，代表它可能是沒有什麼重要的詞。反向文件頻率常見的定義如下：

$$idf(t,D) = log\frac{N}{1 + |\{d \in D : i \in d\}|} \tag{2.14}$$

其中D代表所有文件的集合，而N則是所有文件的數目，詞t在文件集D中的反向文件頻率$idf(t,d)$為所有文件數除以詞t出現過的文件數再取對數。在分母的部分加上1則是避免分母為零。

若將詞頻與反向文件頻率相乘，便能夠得到一個常用來衡量文件中的詞重要程度的指標，也就是tf-idf。當一個詞在文件中有高的出現率，同時在其他文件

doi:10.6342/NTU201702938

中很少出現，這代表這個詞很有可能是這個文件中的重要觀念，便會擁有較高的tf-idf。常用的特徵除了詞頻與反向文件頻率以外，詞在文章中第一次出現的位置、詞性等等，也都是衡量一個詞重要性常使用的特徵。這類特徵的特性在於並不需要人工來標注，若是測試的文章中含有系統沒看過的詞，也能夠抽取其特徵。

除了以統計方式獲得的特徵之外，基於共出現（Co-occurence）特徵所構成的圖形（Graph）也是衡量一個詞重要性的方法。共出現特徵抽取方式為，設定一個窗口大小（Window Size），如果在這個窗口大小中兩個詞有同時出現，即代表此兩個詞共出現。以表 2.1為例，窗口大小設為2，去除停止詞，便能得到如圖 2.9所示的圖形。

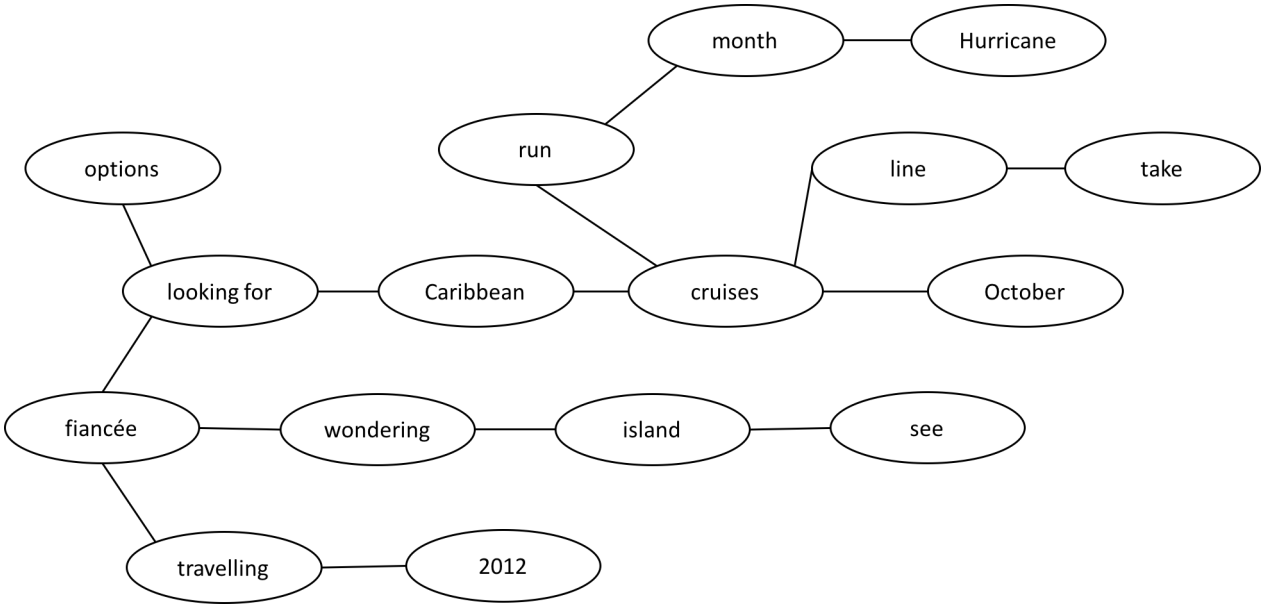


圖 2.9: 基於共出現特徵所構成之圖形

得到此詞圖後，便能根據圖形基礎的排序演算法求得每個節點（也就是每個詞）的重要程度。以TextRank [6]為例，這是很知名的圖形基礎的關鍵用語抽取演算法，一篇文章$D = \{w_1, w_2, ..., w_n\}$每一個詞構成圖上的一個節點$v$，每個節點的

doi:10.6342/NTU201702938

重要程度如下所示：

$$WS(v_i) = (1-d) + d \times \sum_{v_k \in in(v_i)} \frac{w_{ij}}{\sum_{v_k \in out(v_j)} w_{ij}} WS(v_j) \quad (2.15)$$

其中w_{ij}代表兩個節點v_i和v_j間連接的強度，d是一個可以調整的參數，通常設為0.85，經過迭代計算，便能夠得到圖上每個點的重要程度。

2.2.4 比較監督式與非監督式關鍵用語擷取系統

根據上兩節的敘述，我們可以發現由於監督式系統是以分類問題來處理關鍵用語擷取，因此其較有較高的準確度，選出來的詞一定會是關鍵用語；相對的，由於監督式系統只能選出已知的關鍵用語，因此若是得找出系統沒看過的領域的關鍵用語，便需要重新訓練模型。舉例而言，訓練在旅遊文章的關鍵用語系統，在處理醫療相關的文章時便需要重新訓練模型，因為醫療的關鍵用語並沒有出現在原本資料中。

而非監督式系統是以排序問題來處理關鍵用語擷取，其抽取的特徵能夠適用在各種領域，因此不需要重新訓練便能夠抽取出關鍵用語。但其侷限在於無法抽取出沒有出現在文章中的關鍵用語，以表 2.1為例，關鍵用語”vacations”並沒有出現在文章中，故非監督式系統便無法將其抽取出來。兩者優劣點整理如表 2.2：

表 2.2: 監督式與非監督式關鍵用語擷取系統比較

監督式系統	非監督式系統
必須使用具有標注的資料	使用沒有標注的資料
能抽取出不在文章中的關鍵用語	能抽取出不在訓練資料中的關鍵用語
較高的準確率	較大的自由度

doi:10.6342/NTU201702938

2.2.5 評估機制

常見評估關鍵用語擷取系統的方法有下列幾種：

- F1衡量法（F1-measure）

  F1衡量法是資訊檢索（Information Retrieval）領域常用的衡量方式，同時兼顧了精確率（Precision）與召回率（Recall）。 精確率代表挑出的所有關鍵用語中，有多少是正確的關鍵用語，公式如下：

$$P = \frac{|\{correct\ keywords\} \cap \{extracted\ keywords\}|}{|\{extracted\ keywords\}|} \tag{2.16}$$

  召回率則代表所有正確的關鍵用語，系統選出了多少個，公式如下：

$$R = \frac{|\{correct\ keywords\} \cap \{extracted\ keywords\}|}{|\{correct\ keywords\}|} \tag{2.17}$$

  而F1衡量法則是同時考量兩者，公式如下

$$F_1 = 2 \cdot \frac{1}{\frac{1}{R} + \frac{1}{P}} = 2 \cdot \frac{P \cdot R}{P + R} \tag{2.18}$$

- 平均準度均值（Mean Average Precision, MAP）

  平均準度均值是衡量正確的答案是否具有較高的排序，如果關鍵用語擷取系統輸出所有可能的關鍵用語的排序，正確的關鍵用語擁有越高的排序便會得到越高的分數。準度均值的公式如下：

$$AP = \sum_{k=1}^{n} P(k)\Delta r(k) \tag{2.19}$$

  $P(k)$是指在長度為k的回傳排序的精確率，$\Delta r(k)$則是當回傳序列由$k-1$轉變成k時召回率的變化。

  而平均準度均值則是對所有結果求出的準度均值進行平均，數學式子如下：

$$MAP = \frac{1}{Q} \sum_{q=1}^{Q} AP(q) \tag{2.20}$$

doi:10.6342/NTU201702938

Q是所有文件的數量，將每個文件擷取關鍵用語的準度均值進行平均，便能得到整個系統的平均準度均值。

2.3 口述詞彙偵測

2.3.1 簡介

現代有非常多的語音資訊，網路上有豐富的演講、課程、戲劇、新聞等等資訊，該如何檢索或搜尋這些資訊是語音資訊檢索的研究重點。語音資訊檢索的研究方向主要可以分成兩個部分，口述詞彙偵測 [7]（Spoken Term Detection）與語音文件之語義檢索（Sementic Retrieval of Spoken Document），前者著重在找出使用者輸入的查詢詞出現在語音文件中的位置，後者則是要找出與使用者輸入的查詢詞相關的語音文件，不論該查詢詞是否出現在文件之中，只要語意相符即可。在本論文中，主要使用口述詞彙偵測來增進影像辨識系統的訓練。

口述詞彙偵測系統如圖 2.10所示，系統主要由兩個部分構成，分別為辨識系統與檢索系統。辨識系統通常由自動語音辨識系統（Automatic Speech Recognition, ASR）所構成，會將語音訊號轉成辨識結果，可能為詞圖（Lattice）、唯一最佳序列（One Best Sequence）與N最佳序列（N-best List）。基於辨識結果，檢索系統便能搜尋查詢詞的出現在語音文件中的位置。自動語音辨識系統是找出最有可能對應到輸入的語音訊號的詞序列，對語音訊號抽取聲學特徵O，可能的詞序列的集合為$W_{seq} = [w_{seq_1}, w_{seq_2}, ..., w_{seq_n}]$，則自動語音辨識模型可以表示為下式：

$$w^*_{seq} = argmax_{w_{seq} \in W_{seq}} P(w_{seq}|O) \tag{2.21}$$

但是$P(w_{seq}|O)$沒有辦法直接求得，因此根據貝氏定理（Bayes' Theorem）將其化

doi:10.6342/NTU201702938

為下式：

$$P(w_{seq}|O) = \frac{P(O|w_{seq})P(w_{seq})}{P(O)} \tag{2.22}$$

對於一個聲學特徵O而言，$P(O)$為定值，因此上式可以表示為：

$$P(w_{seq}|O) \propto P(O|w_{seq})P(w_{seq}) \tag{2.23}$$

其中$P(O|w_{seq})$可以由聲學模型求得，$P(w_{seq})$則由語言模型得出，w^*_{seq}即為唯一最佳序列，而N最佳序列則是前N個使$P(w_{seq}|O)$最大的w_{seq}。

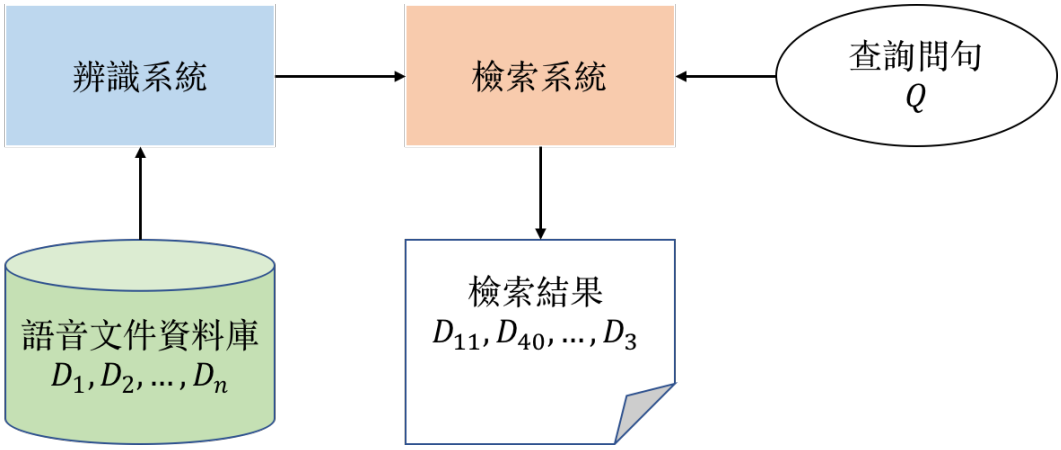


圖 2.10: 口述詞彙偵測系統

2.3.2 詞圖

將語音訊號通過自動語音辨識系統所得出的結果會有很多種形式，其中詞圖（Lattice）是很常被使用的形式。如圖 2.11所示，詞圖是一個網狀的結構，由節點（Node）與詞弧（Arc）構成，節點所代表的是時間資訊，詞弧則代表假設詞（Word Hypothesis）並帶有信心分數（Confidence Score），假設詞和信心分數是由語音模型（Acoustic Model）與語言模型（Language Model）所求得。可以觀察得出來，詞圖儲存了每段時間上可能的詞，這種做法的好處在於自動語音辨識常常會有辨識錯誤，例如將「美國」辨識成「沒過」，「總統」辨識成「鐘頭」。

doi:10.6342/NTU201702938

保留不同的辨識結果，讓可能的正確選項依舊保留，如此一來在進行搜尋的時候，能減輕辨識錯誤所帶來的負面影響。

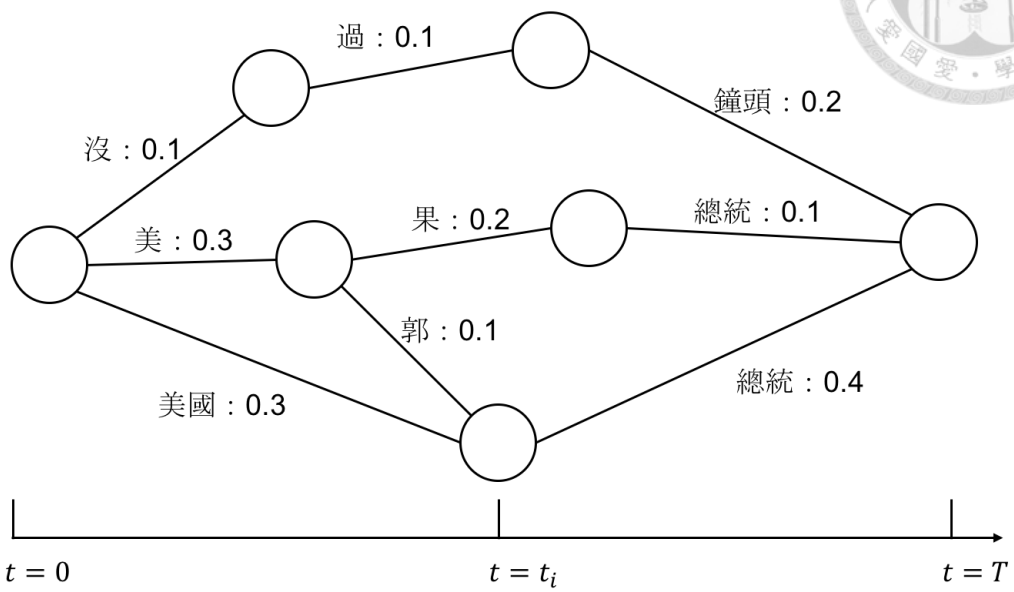


圖 2.11: 詞圖示意圖

2.3.3 加權有限狀態轉換器的語音資訊檢索

在檢索模型中，索引（Index）的建立是檢索效率與效能的關鍵，其中加權有限狀態轉換器是在口述詞彙偵測領域中很常用來建立索引的方法。因為語音辨識中不可避免地會有辨識錯誤以及辭典外詞（Out of Vocabulary, OOV），因此會以次詞單位（Subword Unit）或者將詞與次詞單位混合進行檢索。

假設語料中有n個語句，透過自動語音辨識便能得到n個詞圖。每一個詞圖我們都能轉換成加權有限狀態機的形式，如圖 2.12，詞圖上每一個節點對應到加權有限狀態機的狀態，每一個弧對應到狀態轉移，弧上的事後機率（Posteriori Probability）對應到狀態轉移上的加權值。

接下來要將加權有限狀態機轉換為因子轉換機（Factor Automata），也就是將詞圖中所有的子字串（Substring）以因子轉換機的形式來表現。將圖 2.12中的

doi:10.6342/NTU201702938

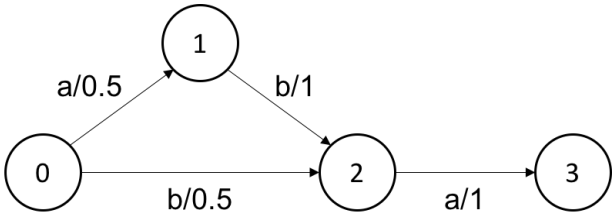


圖 2.12: 加權有限狀態機

加權有限狀態機轉成因子轉換機的結果如圖 2.13所示，原先在詞圖中有兩種可能的字元串，aba與ba，而所有可能的子字元串為a, b, ab, ba, aba。因子轉換器從起始點（圖 2.13的節點0）到終點（圖 2.13的節點5）輸入字元串或子字元串，在終點前的空轉移ε則會輸出語音語句編號。擁有了這個因子轉換器，我們便能將所有的字串與子字串對應到一個編號，對詞圖上所有可能的口述詞彙進行索引。接著針對使用者輸入的字串逐字比對，便能夠得到搜尋結果。

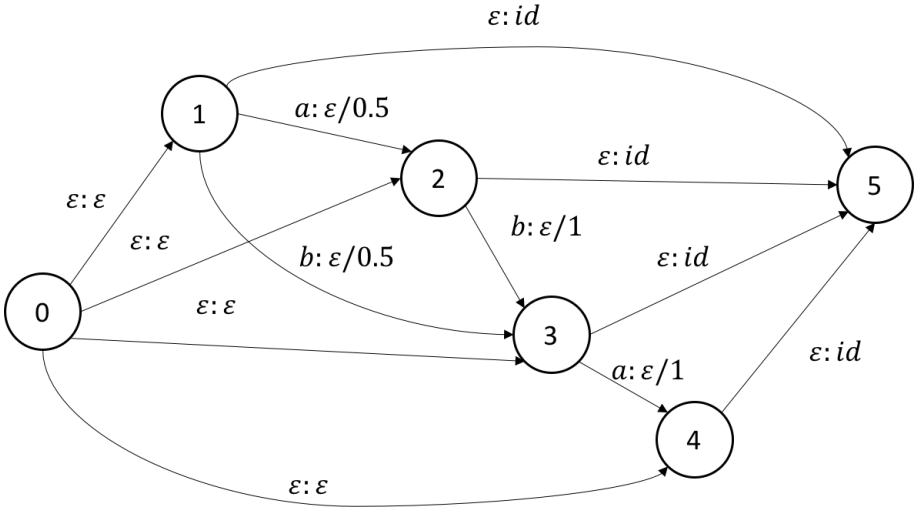


圖 2.13: 因子轉換機

2.4 本章總結

在本章中介紹了深層類神經網路的架構、訓練方法以及兩種深層類神經網路的變

doi:10.6342/NTU201702938

形，卷積類神經網路與長短期記憶類神經網路。同時介紹了關鍵用語擷取的定義與架構，比較了監督式與非監督式關鍵用語擷取系統的差異，以及口述詞彙偵測系統的定義與架構，及如何建立詞圖與索引。

doi:10.6342/NTU201702938

第 三 章　關鍵用語擷取系統

3.1　簡介

如同在第 2.2節所介紹，關鍵用語擷取需要自一份文件中，找出其中的關鍵用語，這個關鍵用語可能會出現在文章中，也可能不會。而關鍵用語可以代表文章主要的觀念，可以做為文章的類別，因此關鍵用語能夠作為文件的索引，也能夠成為文章的摘要。

而現代的關鍵用語擷取系統主要為監督式與非監督式兩種，如表 2.2所示，兩者各有優劣。在本章節嘗試提出一個模型具有非監督式的優點，能夠抽取出沒有看過的關鍵用語，同時也能夠具有監督式的準確率。

在本章中，會於第 3.2節呈現整個系統的架構，第 3.3節說明在本實驗中對詞進行的前處理，第 3.4節與第 3.5節分別詳述本實驗中使用的監督式與非監督式模型，第 3.6節介紹所使用的語料與如何訓練模型與使用模型進行關鍵用語擷取，第 3.7節說明如何設計實驗，第 3.8節分析實驗的結果，最後在第 3.9節給予總結。

3.2　架構與流程

整個系統的架構如圖 3.1所示，訓練資料分為五個不同的領域，而測試資料與訓練資料的領域並不相同。將資料透過前處理，輸入模型，訓練完後便能夠得到關鍵用語擷取系統。模型會預測一個句子中每個位置的詞的重要程度，從中抽取關鍵用語。

doi:10.6342/NTU201702938

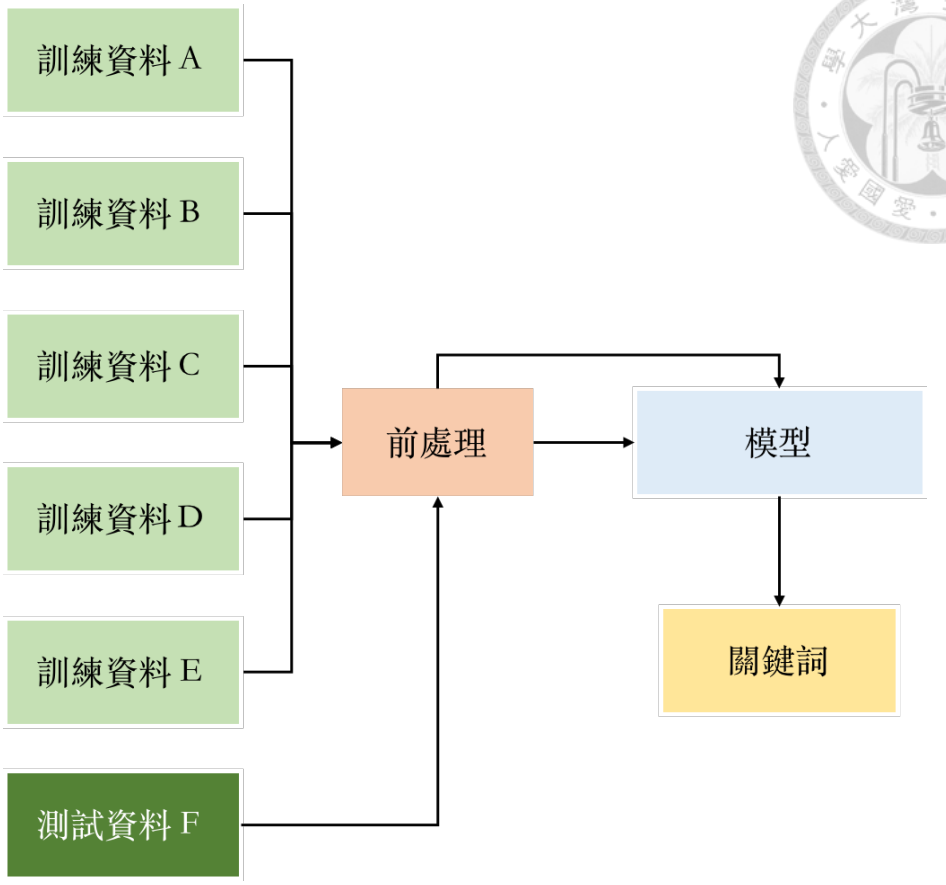


圖 3.1: 關鍵用語擷取系統架構

3.3 前處理

前處理的流程圖如圖 3.2所示。第一步濾除雜訊是因為本實驗所使用的資料為自網路論壇中擷取下來的文章，因此其中包含了非常多的雜訊，例如html語法、latex的公式等等，這些雜訊會干擾模型的訓練，因此會將其濾除。同時將所有的大寫轉換為小寫，多餘的符號去除，只保留分句的符號，例如句點、驚嘆號、問號。

第二步特徵抽取則是取出一些簡單的特徵，如第 2.2.3節所介紹的詞頻與反向文章頻率，或是詞性標註（Part of Speech Tags, POS tags）以及詞向量（Word2vec）。

doi:10.6342/NTU201702938

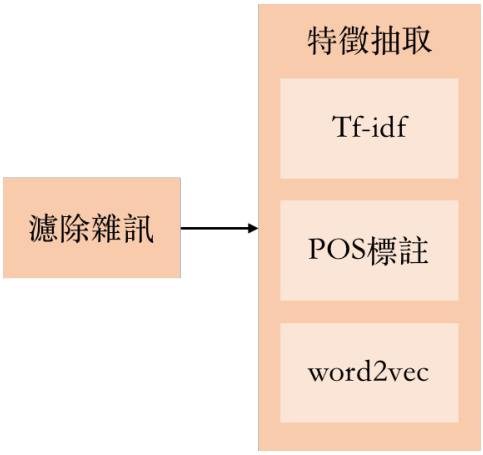


圖 3.2: 前處理流程圖

其中詞向量是現代很常見的詞彙表示法（Word representation）。相較於1-of-N編碼，即將每一個詞表示為一個長度為N的向量，N為整個詞典的大小，詞典中的每一個詞對應到一個維度，一個詞向量只有對應到的維度的值為1，其他為0。此表示法非常的稀疏（sparse），當詞典非常大時，會導致其維度太大，造成維度災難（Curse of Dimensionality），使模型表現變差，同時也無法表示出詞與詞之間的關係，因為只要是相異的詞其距離都是相同的。而詞向量則能處理以上的問題，其向量維度往往遠小於N，同時透過計算兩的詞向量的歐式距離（Euclidean distance）或餘弦相似度（Cosine similarity）便能得出兩個詞的相似度。

在本論文中，詞向量的維度皆為200，抽取的統計特徵包含詞頻（term frequency）、反向文章頻率（inverse document frequency）、詞頻與反向文章頻率之乘積（tf-idf）、該詞是否在標題、該詞是否是停止詞、該詞在所有文章中的出現次數、該詞在句子的位置，共7維。

3.4 監督式模型

由於本實驗著重於想要讓模型能夠抽取出沒有看過的關鍵用語，因此傳統的監督

doi:10.6342/NTU201702938

式模型便不能夠模擬成分類問題，因為訓練時的「類別」與測試時並不相同。若是以分類問題方式來處理關鍵用語擷取，如表 3.1 ，每一個關鍵用語都是一個類別，每一個句子都可能屬於數個關鍵用語，換言之一個句子可能屬於一個或數個分類，但如果該關鍵用語沒有出現在訓練資料中，系統便無法擷取，因為這個「類別」並不存在。

於是在本實驗中，我們以「位置」的資訊作為標註，如表 3.2 ，其中句子中只有'caribbean'為關鍵用語，因此這句話所對應到的標註便是一個長度為句子長度的向量，只有'caribbean'對應到的位置為1其餘部分為0。

而在以下子節將會詳述本實驗使用的不同的模型。

表 3.1: 關鍵用語擷取將關鍵用語視為類別

所有的關鍵用語	caribbean, cruising, vacations, airplan, hotel, wi-fi
句子	what are some caribbean cruises for october?
標註	[1, 1, 1, 0, 0, 0]
關鍵用語	caribbean, cruising, vacations

表 3.2: 關鍵用語擷取以位置資訊為標註

句子	what are some caribbean cruises for october?
標註	[0, 0, 0, 1, 0, 0, 0, 0]
關鍵用語	caribbean, cruising, vacations

3.4.1 卷積類神經網路模型

卷積類神經網路模型（Convolutional Neural Network, CNN）如第 2.1.3節所述，能

doi:10.6342/NTU201702938

夠自輸入資料中抽取更高層次的抽象特徵，因此在處理句子時，能夠獲得句子結構的資訊，希望藉此能夠找出句子中有哪些詞是重要的。而其缺點是由於輸入是圖形的形式，因此必須設定句子長度上限。

將一個句子模擬成圖片以輸入模型中的方法如圖 3.3所示，將句子中每個字的特徵向量疊在一起，便能夠成為一個句子的特徵圖，如果句子的長度小於句子長度上限則補零，若句子長度超過上限，則直接忽略超過的詞。

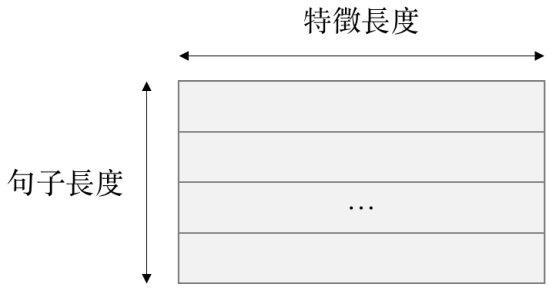


圖 3.3: 句子特徵圖示意圖

實驗所使用的卷積類神經模型如圖 3.4所示，首先在輸入的部分將詞向量、統計特徵與詞性標記三者分開，因為卷積類神經網路模型的卷積層會共用參數，而這三種不同的特徵應該要使用不同的參數來模擬。而詞性標記所使用的嵌入層（Embedding Layer）的效果與詞向量有異曲同工之妙，將所有可能的詞性（如動詞、名詞、形容詞等）化為1-of-N表示法，透過嵌入層轉為10維的向量，因為不同的詞性之間具有不同的關係，例如及物動詞與不及物動詞之間的關係應該相異於與形容詞之間的關係。

卷積單元的大小為 (1, 特徵長度)，這代表每一個詞都會共用同一組的卷積單元，每一組卷積單元則各自代表了不同的特徵，得出來的特徵透過合併層後輸入隱藏層，便能夠得到輸出變數。

而值得注意的是，輸出變數有兩個，較長的輸出層代表句子每個位置的詞是

doi:10.6342/NTU201702938

否為關鍵用語，句子的長度上限設為20，則若某個詞為關鍵用語，該位置所對應到的值為1，其餘為零。較短的輸出層則代表句子中是否有關鍵用語，若有則輸出1，反之輸出0。藉由這兩個輸出變數，我們能夠預測一個句子是否具有關鍵用語，若有則該關鍵用語位置在何處。

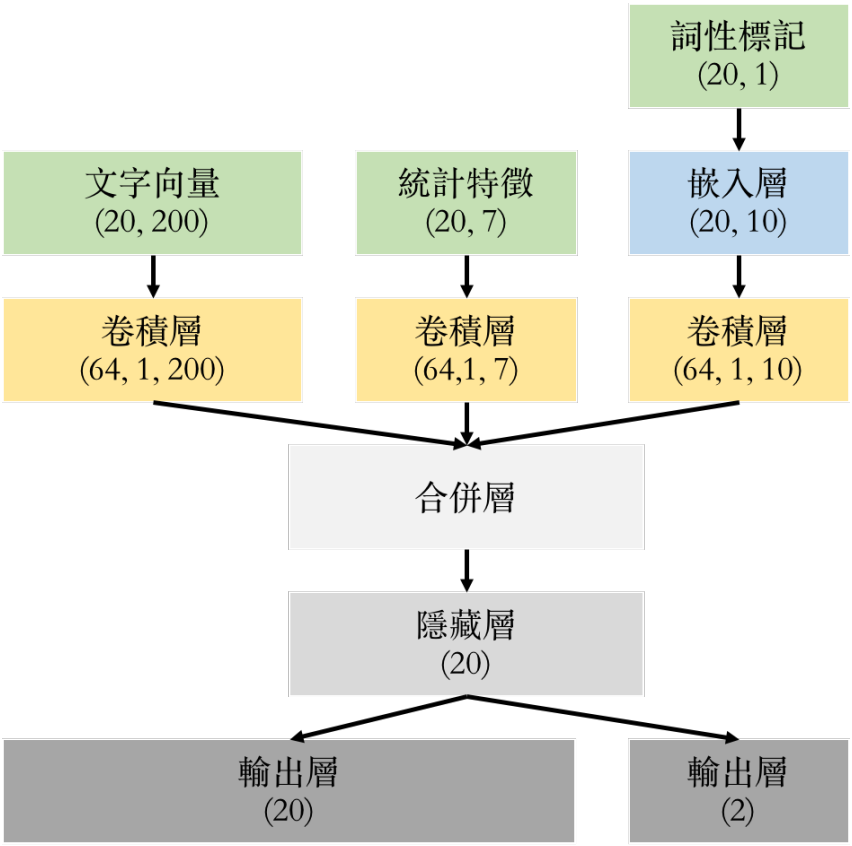


圖 3.4: 卷積類神經關鍵用語擷取模型

3.4.2 長短期記憶類神經網路模型

長短期記憶類神經網路模型（Long Short-Term Memory Network，LSTM）如第 2.1.4節所述，能夠模擬不同長度的序列，能夠保留序列中的歷史資訊，因此相當適合來處理詞序列的問題。相較於卷積類神經網路，長短期記憶類神經網路便不需要設定句子的長度上限。

doi:10.6342/NTU201702938

相較於分類式的監督式關鍵用語擷取系統，如圖 2.8，本實驗中是以位置資訊來作為標註，其架構如圖 3.5所示。

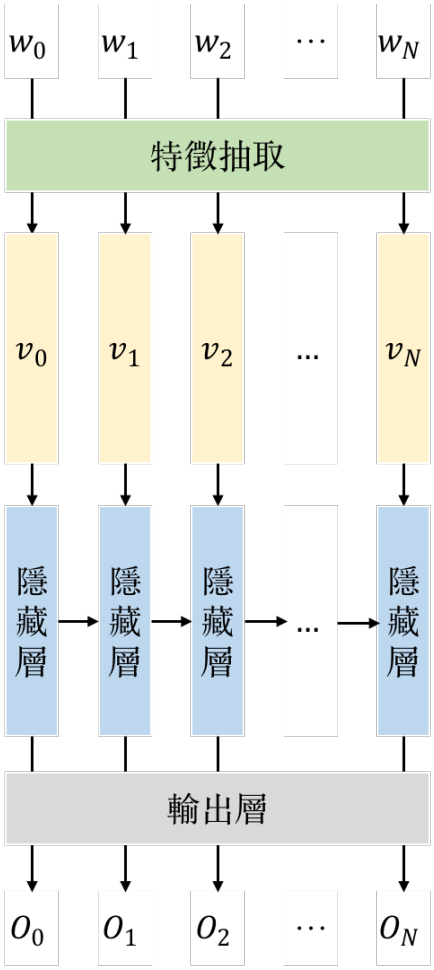


圖 3.5: 以位置資訊為預測目標之長短期記憶類神經網路關鍵用語擷取模型

將詞序列$W = [w_0, w_1, ..., w_N]$抽取特徵，包含詞向量與統計特徵，輸入模型後得出輸出序列$O = [o_0, o_1, ..., o_N]$。如果詞w_t為關鍵用語，則o_t值為1，反之為0。

在這個模型中會遇到一個問題，即絕大多數的詞都不是關鍵用語，因此序列中會包含了很多的0，在訓練模型時會以為準確率很高，因為模型會傾向於預測每一個詞都不是關鍵用語。可是實際上評估效能是看抽取出的關鍵用語有哪些正確，也就是考量F檢定，因此得將此資訊放入模型的訓練之中。

在本實驗中我們使用了增強式學習的方法，架構如圖 3.6所示，首先以

doi:10.6342/NTU201702938

圖 3.5的方式初始化模型，可以採取的動作（Action）為選擇一個詞是否為關鍵用語，而環境（Environment）給予動作相對應的獎賞（Reward）。

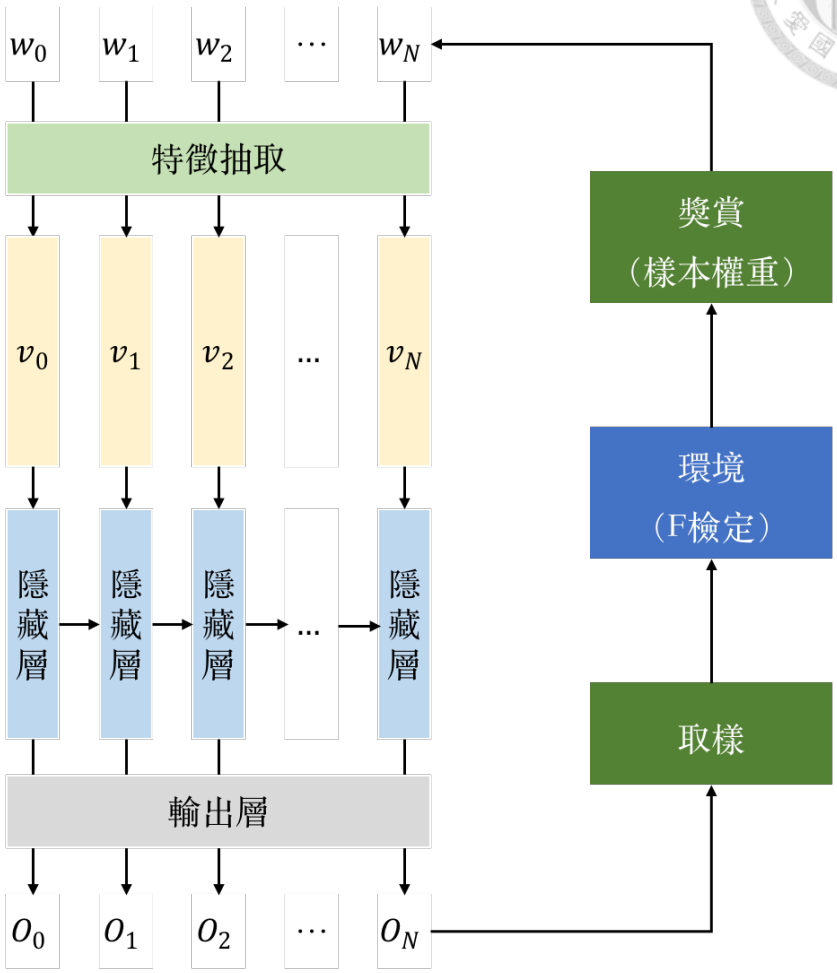


圖 3.6: 輔以增強式學習之位置資訊長短期記憶類神經網路關鍵用語擷取模型

每一個詞都會對應到一個0到1之間的數字，代表這個詞為關鍵用語的可能性，根據這個值進行抽樣，每個詞序列W都會對應到數個可能的輸出序列$O_0, O_1, ..., O_k$。而每個輸出序列與正確答案求得F檢定的值，便是這組樣本的權重。換言之，正確率越高的配對，會具有越高的權重，模型在訓練的時候便會更加注重這些抽取較多正確關鍵用語的配對，藉由這個方式，讓訓練模型時F檢定的資訊也能同時被考量進去。

doi:10.6342/NTU201702938

3.5 非監督式模型

除了使用監督式模型來抽取關鍵用語，非監督式模型也具有許多方式能夠抽取關鍵用語。最簡單的方式為根據tf-idf的值，選取前N個詞作為關鍵用語。除此之外也能根據詞圖，以詞與詞之間共出現的次數作為特徵，抽取關鍵用語，如第 2.2.3節。

3.6 實驗基礎架構

3.6.1 語料介紹

實驗所使用的資料為下載自StackExchange的資料，包含了六大領域，分別是生物（biology）、廚藝（cooking）、旅遊（travel）、機器人（robotics）、密碼學（crypto）與手作（diy），語言為英文。六大領域內含的資訊簡列如表 3.3，可以看出不同的領域之間所含文章數量、詞典大小與關鍵用語數量都很不相同。

表 3.3: 語料基本資訊

	生物	廚藝	旅遊	機器人	密碼學	手作
文章數量	13196	15404	19279	2771	10432	25918
詞典大小	38257	24313	32072	17160	26792	32106
關鍵用語數量	678	736	1645	231	392	734

每篇文章都包含了標題、內文與對應到的關鍵用語，其中關鍵用語是發文者自行標註的，可以讓其他人更容易的搜尋到這篇文章，或是理解這篇文章的重點。這些關鍵用語有可能出現在文章中，也有可能沒有。將關鍵用語的詞向量降到二維平面視覺化如圖 3.7所示，可以觀察到有明顯的分群，但也有所重疊，因此

doi:10.6342/NTU201702938

使用詞向量作為詞的表示法，應能幫助模型學習如何自文章中抽取出關鍵用語。

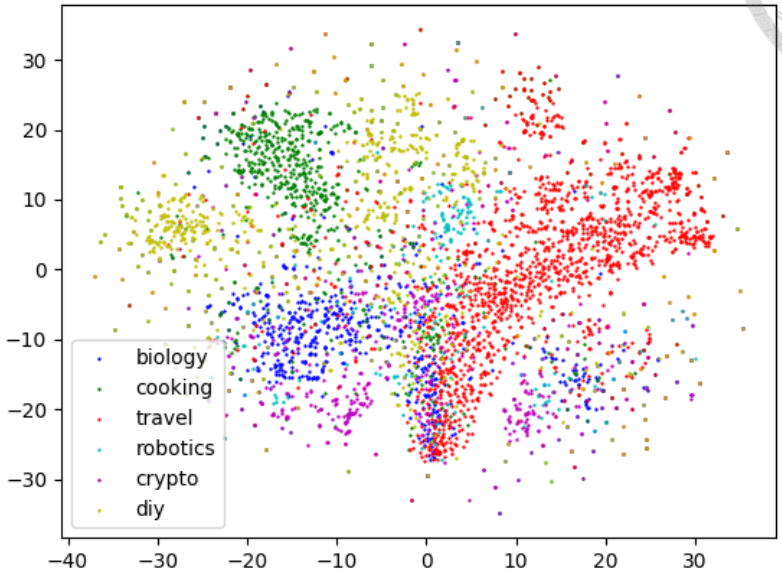


圖 3.7: 不同主題關鍵用語之視覺化

3.6.2 訓練與辨識系統

本實驗目的為設計一個能夠抽取出關鍵用語的系統，使用監督式訓練但同時具有非監督式關鍵用語擷取系統的優點，能夠抽取沒有見過的關鍵用語，同時具有不錯的效能。

在特徵抽取的階段，會同時訓練所有的文章所含的詞的詞向量，每個領域分別計算其統計特性，如同第 3.3節所述。而訓練模型的階段，是將訓練資料輸入第 3.4節與第 3.5節中的模型進行訓練，其中模型訓練時的減損函數皆為準確率（Accuracy），使模型預測每個詞是否為關鍵用語的準確率上升。測試模型時，則是以F檢定來衡量模型的效能，也就是說真正在意的是有多少關鍵用語被準確地抽取出來，而不太在意判定了多少非關鍵用語為非關鍵用語。

doi:10.6342/NTU201702938

3.7 實驗設計

使用第 3.6.1節所提到的資料，根據不同領域分為訓練與測試的資料，如同圖 3.1所示，也就是訓練與測試的資料領域並不會重疊。

實驗所使用的基準實驗（Baseline）包含以下幾種模型，非監督式模型有直接使用tf-idf取最高前四、TextRank [6]、Rake [8]等三種，監督式模型則為分類式監督式模型 [5]。而上限是因為有些關鍵用語沒有出現在文章內容中，位置資訊之關鍵用語抽取系統無法抽取出其結果，因而有理論上表現的上限。

而本實驗所提出的模型架構有卷積類神經網路模型、長短期記憶類神經網路模型與輔以增強式學習之長短期記憶類神經網路模型。

3.8 實驗結果

表 3.4是將不同領域的F1分數並列呈現，而表 3.5到表 3.10為以不同領域的資料分別進行測試所得出的實驗結果。

以表 3.5為例，測試資料的領域為生物，(a)列是直接選出文章中的關鍵用語，是位置資訊關鍵用語模型表現的上限。(b)(c)(d)三列為非監督式模型的結果，此類模型不需要使用訓練資料，直接對測試資料的詞進行排序，得出每篇文章較有可能的關鍵用語。(e)則是監督式模型，隨機將測試資料分為訓練與測試集，比例為1 : 1，也就是其訓練資料與測試資料為同一個領域。(f)(g)(h)為本論文提出的模型，訓練資料為其他五個領域之集合，測試資料為生物領域，換言之測試資料與訓練資料的領域並不相同。

提出模型與非監督式模型的比較可以發現，在大多數的情況下不論是卷積類神經網路、位置資訊之長短期記憶類神經網路或是增強式位置資訊之長短期記憶

doi:10.6342/NTU201702938

類神經網路都比非監督式模型具有更好的表現，均具有更好的準確率，以及不差的召回率。非監督式模型作為弱基準，能佐證本論文提出之模型效果。

同時自實驗結果可以發現分類式長短期記憶類神經網路表現都不錯，因其擁有領域內的資訊，不論是準確率或是召回率都有不錯的表現，同時表現穩定，不論是在哪個領域都具有相近的表現結果，所以此結果適合作為提出模型的强基準實驗（Strong Baseline）。與提出模型之比較，在大多數關鍵用語沒有出現於文章中的領域，如生物，分類式長短期記憶類神經網路模型擁有壓倒性的優勢。但如果關鍵用語有出現在文章中，如廚藝、機器人、密碼學等三個領域，位置資訊模型表現能夠超越分類式模型。值得一提的是，分類式模型只能夠擷取看過的領域的關鍵用語，而位置資訊模型則不受此限制，因此本論文所提出之模型在效能不差（甚至可能超越）的情況下，同時具有更大的彈性。

而本論文所提出的三個模型之間的比較，在大多數的情況下，增强式位置資訊之長短期記憶類神經網路為表現最佳的模型，可以得知將F檢定的資訊加入訓練過程對於效能的提升相當具有幫助。此外，可以發現卷積類神經網路往往具有較高的準確率，應是在輸出變數多加了一個衡量此句子是否含有關鍵用語的變數，使得模型在準確率上表現得更好。

接下來觀察每個實驗結果的P-R曲線，如圖 3.8到圖 3.13，此曲線能夠很好的衡量模型的表現。因為在進行關鍵用語抽取時，每個詞對應到的輸出值是介於0到1之間的值，必須得設定一個閾值，超過這個值便被視為關鍵用語。將這個值由高移到低，抽取出的關鍵用語的數量就越多。或者也能以整篇文章每個詞所對應到的值進行排序，則前N個詞便被視為關鍵用語，隨著N的值增加，便會抽取出更多的關鍵用語。抽取出更多的關鍵用語，召回率可能會上升而準確率可能會下降，兩者所形成的曲線為P-R曲線，曲線越接近右上角代表此系統效能越好。

doi:10.6342/NTU201702938

表 3.4: 不同領域以不同模型所得出之F1分數

			生物	廚藝	旅行	機器人	密碼學	手作
(a) 上限			0.359	0.672	0.578	0.651	0.684	0.637
基礎實驗	非監督	(b) rake	0.109	0.291	0.237	0.215	0.257	0.235
		(c) td-idf	0.095	0.249	0.206	0.193	0.224	0.201
		(d) TextRank	0.084	0.242	0.154	0.151	0.178	0.184
	監督式	(e) 分類式長短期記憶類神經網路	**0.208**	0.221	**0.243**	0.180	0.222	**0.265**
本論文所提出之方法	(f) 卷積類神經網路		0.103	0.290	0.221	0.239	**0.269**	0.236
	(g) 位置資訊之長短期記憶類神經網路		0.103	0.285	**0.229**	0.202	0.252	0.217
	(h) 增強式位置資訊之長短期記憶類神經網路		**0.113**	**0.318**	0.215	**0.246**	0.255	**0.245**

可以觀察到非監督式模型在每個領域的效能均是最差，而提出的模型在廚藝、機器、密碼學三個領域表現超越了分類式長短期記憶類神經網路模型。除此之外，卷積類神經網路在準確率上有特別傑出的表現，而位置資訊長短期類神經網路其一開始的準確率不高，因為起初閾值很高，許多關鍵用語之信心分數低於閾值，因此造成許多文章都沒有擷取出關鍵用語，而使準確率較低，隨著閾值的下降，擷取更多的關鍵用語，準確率與召回率便同時上升；隨著閾值降的更低，召回率持續增加，而準確率則又下降。

doi:10.6342/NTU201702938

表 3.5: 生物之實驗結果

			準確率	召回率	F1
(a) 上限			0.567	0.285	0.359
基準實驗	非監督	(b) rake	0.104	0.141	0.109
		(c) td-idf	0.092	0.115	0.095
		(d) TextRank	0.079	0.106	0.084
	監督式	(e) 分類式長短期記憶類神經網路	0.194	0.256	**0.208**
本論文所提出之方法	(f) 卷積類神經網路		0.095	**0.145**	0.103
	(g) 位置資訊之長短期記憶類神經網路		0.115	0.117	0.103
	(h) 增强式位置資訊之長短期記憶類神經網路		**0.133**	0.124	**0.113**

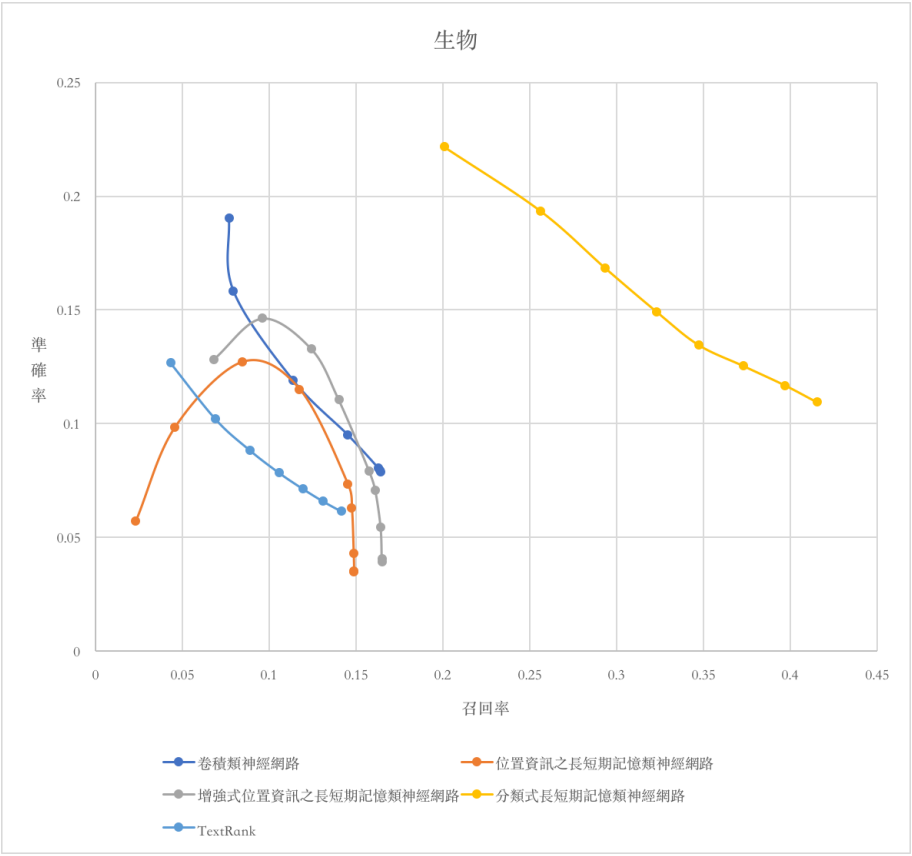


圖 3.8: 生物之P-R曲線

doi:10.6342/NTU201702938

表 3.6: 廚藝之實驗結果

			準確率	召回率	F1
(a) 上限			0.857	0.593	0.672
基準實驗	非監督	(b) rake	0.259	0.404	0.291
		(c) td-idf	0.221	0.333	0.249
		(d) TextRank	0.284	0.237	0.242
	監督式	(e) 分類式長短期記憶類神經網路	0.240	0.226	0.221
本論文所提出之方法	(f) 卷積類神經網路		0.307	0.342	0.290
	(g) 位置資訊之長短期記憶類神經網路		0.325	0.312	0.285
	(h) 增强式位置資訊之長短期記憶類神經網路		**0.347**	**0.365**	**0.318**

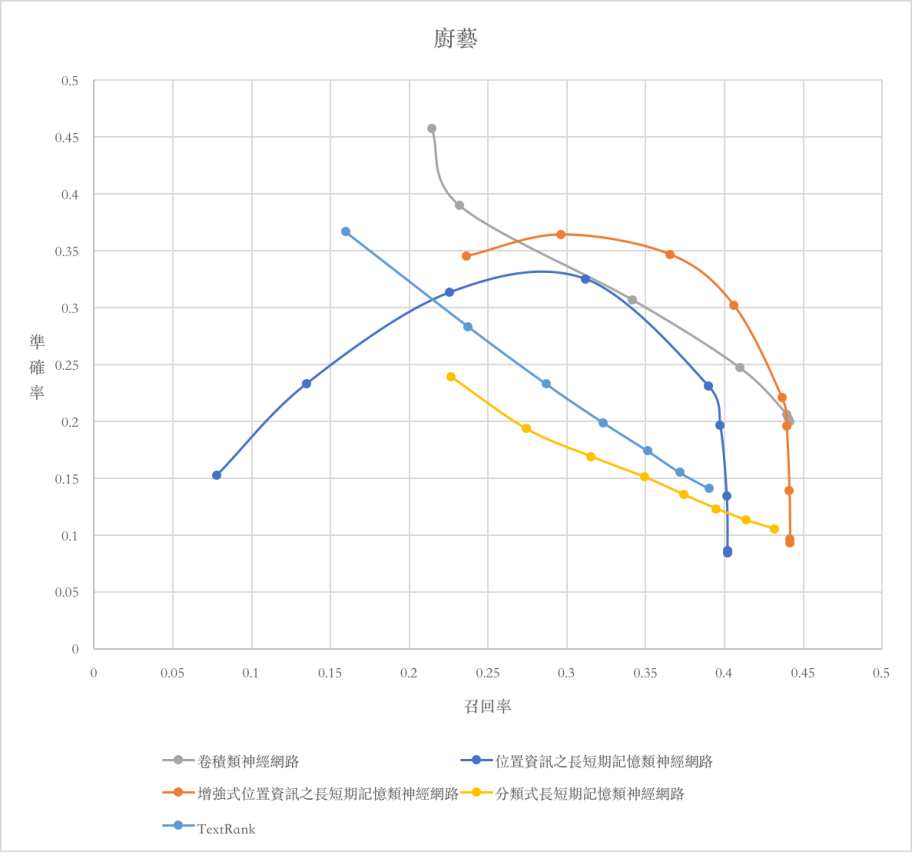


圖 3.9: 廚藝之P-R曲線

doi:10.6342/NTU201702938

表 3.7: 旅行之實驗結果

			準確率	召回率	F1
(a) 上限			0.895	0.457	0.578
基準實驗	非監督	(b) rake	0.247	0.262	0.237
		(c) td-idf	0.226	0.207	0.206
		(d) TextRank	0.130	0.210	0.154
	監督式	(e) 分類式長短期記憶類神經網路	0.259	0.246	**0.243**
本論文所提出之方法	(f) 卷積類神經網路		0.208	**0.285**	0.221
	(g) 位置資訊之長短期記憶類神經網路		**0.294**	0.219	**0.229**
	(h) 增強式位置資訊之長短期記憶類神經網路		0.227	0.248	0.215

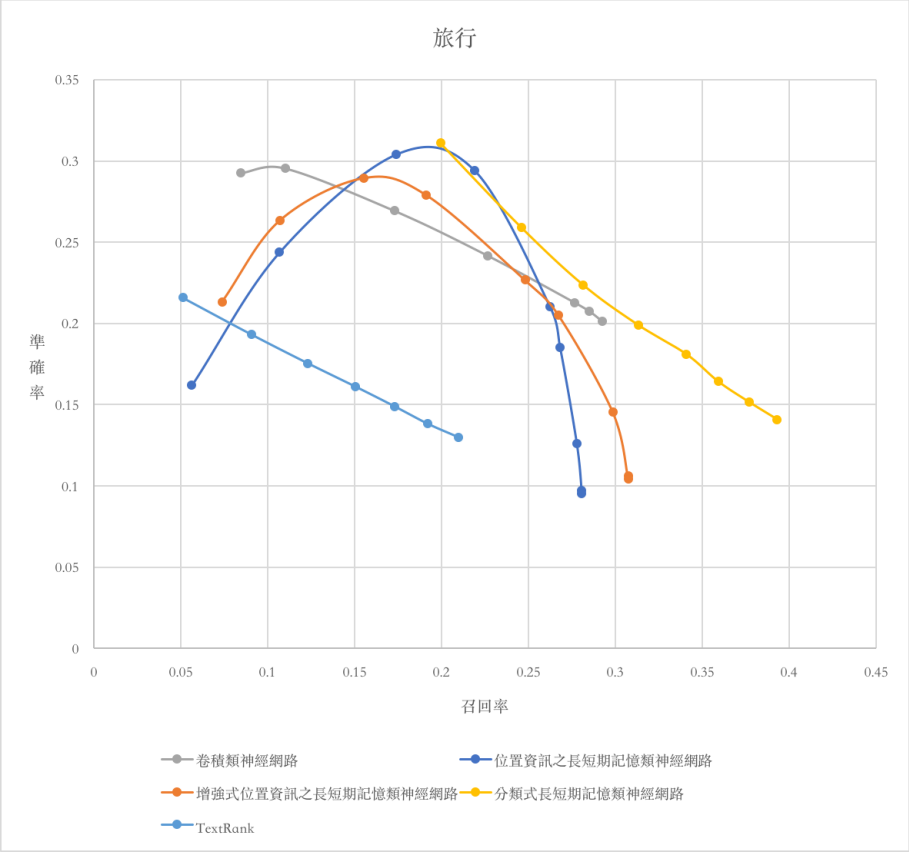


圖 3.10: 旅行之P-R曲線

doi:10.6342/NTU201702938

表 3.8: 機器人之實驗結果

			準確率	召回率	F1
(a) 上限			0.825	0.575	0.651
基準實驗	非監督	(b) rake	0.189	0.311	0.215
		(c) td-idf	0.176	0.253	0.193
		(d) TextRank	0.136	0.202	0.151
	監督式	(e) 分類式長短期記憶類神經網路	0.147	0.265	0.180
本論文所提出之方法	(f) 卷積類神經網路		**0.402**	0.194	0.239
	(g) 位置資訊之長短期記憶類神經網路		0.207	0.247	0.202
	(h) 增強式位置資訊之長短期記憶類神經網路		0.274	**0.276**	**0.246**

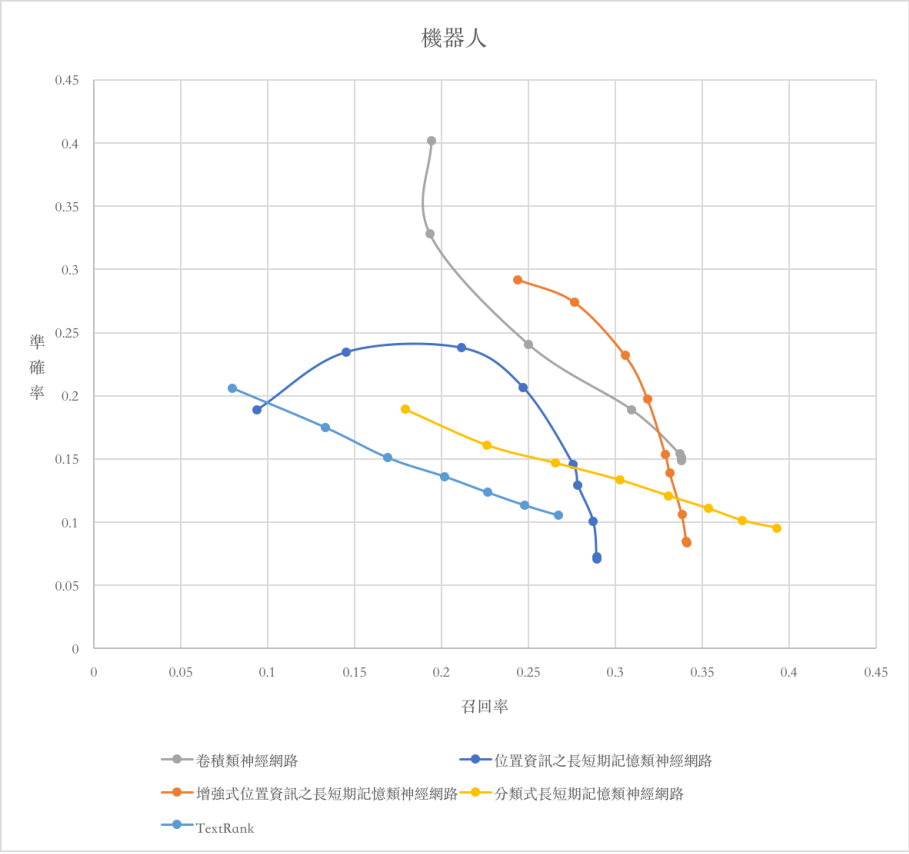


圖 3.11: 機器人之P-R曲線

doi:10.6342/NTU201702938

表 3.9: 密碼學之實驗結果

			準確率	召回率	F1
(a) 上限			0.892	0.595	0.684
基準實驗	非監督	(b) rake	0.250	0.326	0.257
		(c) td-idf	0.221	0.265	0.224
		(d) TextRank	0.159	0.237	0.178
	監督式	(e) 分類式長短期記憶類神經網路	0.207	0.269	0.222
本論文所提出之方法	(f) 卷積類神經網路		**0.512**	0.206	**0.269**
	(g) 位置資訊之長短期記憶類神經網路		0.306	0.260	0.252
	(h) 增強式位置資訊之長短期記憶類神經網路		0.290	**0.284**	0.255

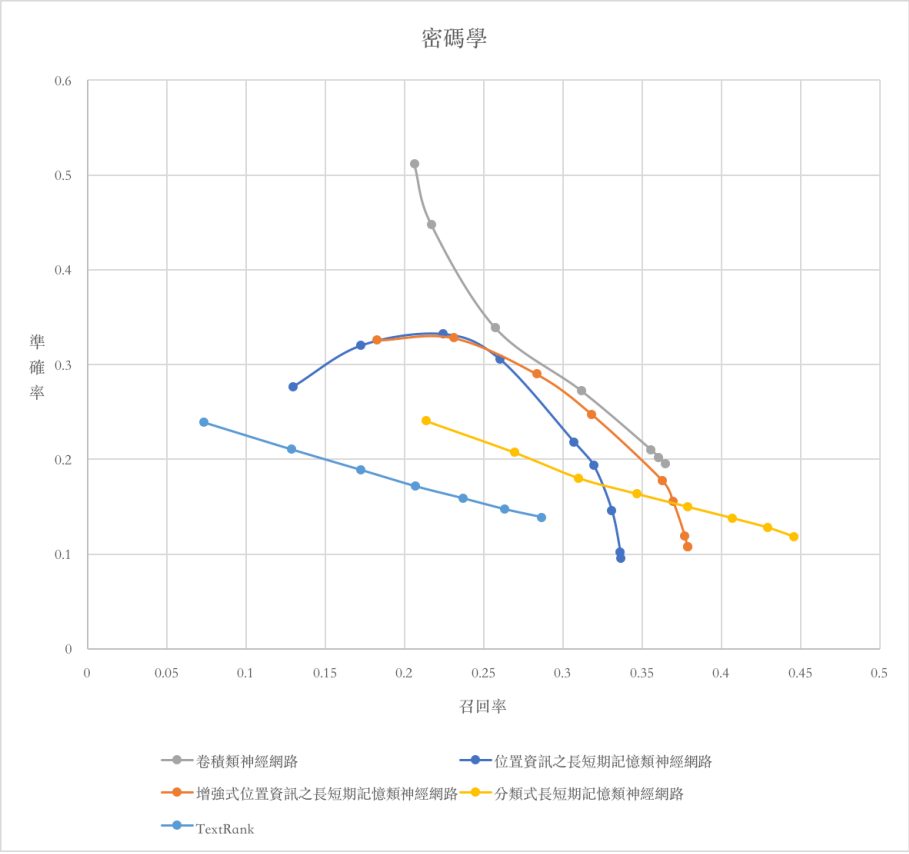


圖 3.12: 密碼學之P-R曲線

doi:10.6342/NTU201702938

表 3.10: 手作之實驗結果

			準確率	召回率	F1
(a) 上限			0.794	0.565	0.637
基準實驗	非監督	(b) rake	0.201	0.346	0.235
		(c) td-idf	0.176	0.271	0.201
		(d) TextRank	0.181	0.213	0.184
	監督式	(e) 分類式長短期記憶類神經網路	0.283	0.282	**0.265**
本論文所提出之方法	(f) 卷積類神經網路		**0.256**	0.269	0.236
	(g) 位置資訊之長短期記憶類神經網路		0.241	0.245	0.217
	(h) 增强式位置資訊之長短期記憶類神經網路		0.242	**0.312**	**0.245**

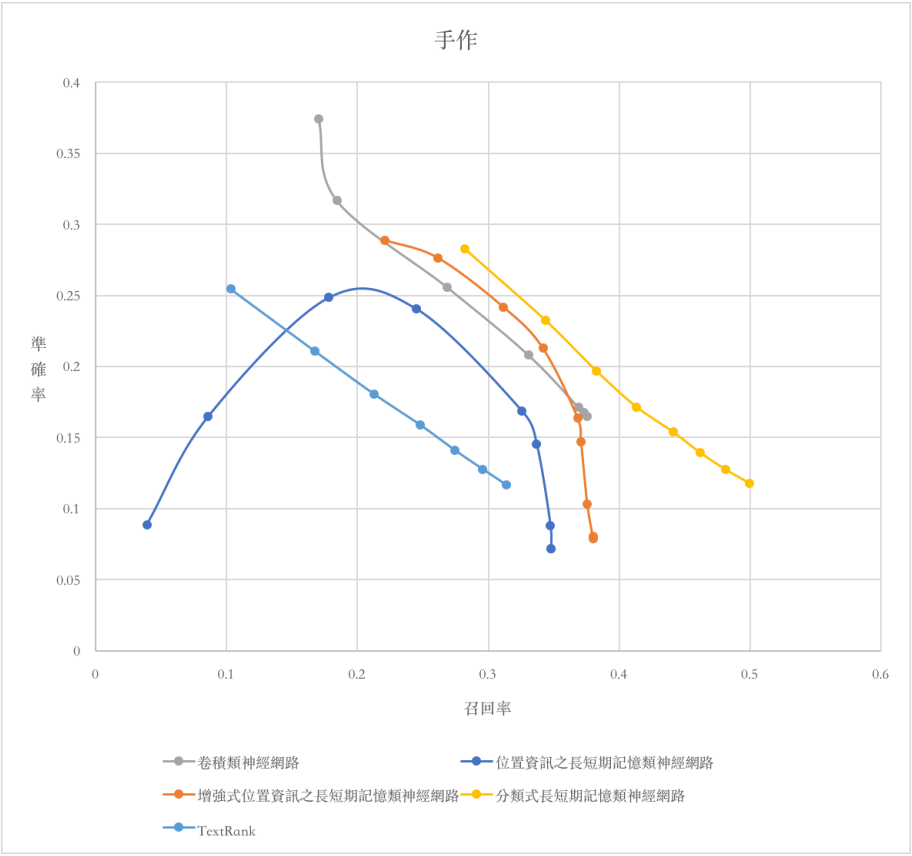


圖 3.13: 手作之P-R曲線

doi:10.6342/NTU201702938

3.9 本章總結

在本章節中提出了一個關鍵用語擷取系統，有別於過往的分類型監督式模型或排序型非監督式模型，此系統以位置資訊為預測目標，根據句子的結構、前後文，來預測每一個詞是否為關鍵用語。使用監督式訓練但同時兼具非監督式模型的優點，而此可能為未來關鍵用語擷取研究的前進方向。

除此之外，透過增強式學習能夠讓模型設計的彈性更大，舉例而言，以往目標函數必須可微，因此本實驗中F檢定便無法作為目標函數。透過增強式學習，模型便能學習到更多可能的目標函數，進而解決更多樣化的問題；在本實驗中，更能夠使模型的表現大幅進步。

doi:10.6342/NTU201702938

第 四 章　以口述詞彙偵測訓練圖像辨識模型

4.1　簡介

自影片中獲取資訊是現在非常重要的研究課題，因為網路上擁有非常豐富的影音資源，涵蓋了各個不同的領域，如果能讓機器有效率的自網路上的影片取得知識，那將能對人類的生活提供非常大的助益。

其中一大阻礙便是訓練資料的搜集，如此龐大的資料量難以使用人工進行標注，進而使模型的訓練產生困難。在本論文中將提出一系統，能夠使用口述詞彙偵測來進行自動標注，使模型能夠自動從影片中學習資訊。

值得注意的是，本系統除了口述詞彙偵測模型與預先給定的關鍵用語清單，訓練過程不需要使用任何標注的資料，但依舊能夠有效的學習，此為本系統的重要特色。

4.2　架構與流程

使用口述詞彙偵測對影片進行標注的系統架構如圖 4.1所示，共分為兩個部分，影像處理與口述詞彙偵測，其內容分別於第 4.2.2節與第 4.2.1節介紹。

透過口述詞彙偵測對影片進行標注後，我們能夠以此系統的輸出結果作為訓練資料，訓練一個影像辨識模型，架構如圖 4.5所示，而詳細內容在第 4.2.3節介紹。

4.2.1　影像處理

代表影格抽取（Representative frame Extraction）本論文中採取較容易的做法，設

doi:10.6342/NTU201702938

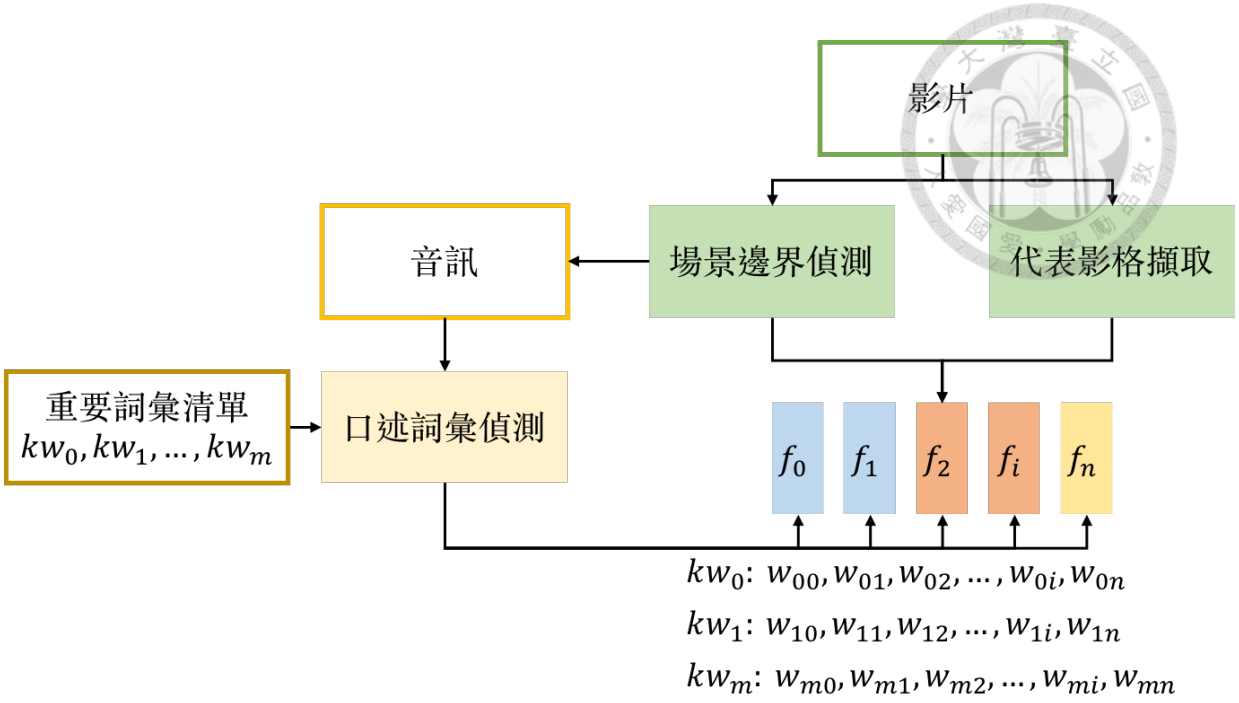


圖 4.1: 以口述詞彙偵測對影片進行標注之系統架構

定一個時間間隔t，每隔t秒時間便取出一張影格作為代表影格。舉例而言，影片每秒會有24張影格，若設$t = 1$則系統會每24張影格抽取一張作為代表影格。

而場景偵測（Shot Detection）則是透過尺度不變特徵轉換 [9]（Scale-invariant feature transform, SIFT）抽取局部特徵點，透過比較相鄰影格的特徵，決定兩張影格是否同時隸屬於同一場景。

尺度不變特徵轉換是電腦視覺（Computer Vision）中用來提取影像局部特徵特性的演算法，其特色是抽取圖片局部的特徵，因此可以不受到旋轉、大小縮放、亮度變化等影響，同時運算的速度快，能夠處理大量的資訊。因此相當適合用在處理影片中的場景偵測，能有效率的分別數張影格是否屬於相同的場景。演算法的特色在於，根據不同尺度下的高斯差（Difference of Gaussians, DOG）的極值，求得影像中的關鍵點（Key Points）。而在同一個場景中的影像，其應該會具有相近的關鍵點，因為影像內容是連續的；相對的，在場景轉換的部分，關鍵點

doi:10.6342/NTU201702938

的特徵會很不相同，因此能夠藉此快速的決定場景間的界限。

在圖 4.1中，視訊透過代表影格擷取、場景邊界偵測，可以將連續的影格分為不同的場景（以不同的顏色表示），每個不同的代表影格彼此之間的間隔時間都是固定的，如圖 4.2所示。上半部為原始影片，下半部標示出尺度不變特徵轉換抽取出之特徵，可以觀察到同一場景之特徵點比較相近，如圖 4.2(a)與圖 4.2(b)；而不同場景之間特徵點分布差異很大，如圖 4.2(b)與圖 4.2(c)。

圖 4.2: 場景偵測示意圖

4.2.2 口述詞彙偵測

如同第 2.3節所述，系統中所使用的口述詞彙偵測（Spoken Term Detection）是以尋找查詢詞出現在語音文件的位置為主要的目標，而使用口述詞彙偵測來對影像進行標注，原因為假設影片中的視訊會與音訊有所相關，例如一個示範烹飪的影片，當示範者說「番茄」的時候，畫面應該會出現番茄；又如在介紹動物的影片中，當主持人提到「拉布拉多」的時候，畫面應該會呈現拉布拉多犬的影像。換言之，如果我們能夠找出音訊中我們所在意的關鍵詞，便能夠找到視訊中相對應的影像。

doi:10.6342/NTU201702938

在圖 4.1中，我們在音訊中尋找預先給定的關鍵用語清單，例如動物星球頻道的影片對應到的關鍵用語清單可能包含狗、貓、拉布拉多、暹羅貓、大麥町犬等，而在音訊中尋找這些詞彙所在的位置，換言之，對於某個關鍵用語kw_i會得到一組權重$w_{i0}, w_{i1}, ..., w_{in}$，其中$w_{it}$是指時間$t$所在的詞彙是關鍵詞$kw_i$的分數，當$w_{it}$超過一預先設定的閾值，便視為此關鍵詞有出現在此位置。

根據這個分數，能夠將關鍵用語作為類別，其出現時間所對應的影格則被標註此類別，進而組成訓練資料，以此訓練影像辨識模型。舉例而言，任意選出一影片片段，以「狗（Dog）」為關鍵詞進行口述詞彙偵測，可以得到如圖 4.3之結果，在四張影格中只有第二張所對應到的音訊含有「狗」，只有其被標註為「O」，其他的均為「X」，因此我們便能將第二張圖片歸屬於「狗」這個類別。

圖 4.3: 以狗為關鍵詞彙對影格進行標注

4.2.3 影像辨識模型

觀察圖 4.3可以發現使用口述詞彙偵測對影片標註會有一些問題，例如：視訊中有關鍵詞所對應的物體，但音訊中卻沒有提及，如圖 4.3的第三張影格有出現狗，但卻沒有被標註；音訊具有大量的雜訊（例如背景音樂、動物的吼叫聲等）導致辨識正確率低而無法偵測到關鍵詞；在音訊中可能很少直接提及關鍵詞，而使用代名詞或是其他方式來代稱。換言之，在所有的影格中，除了正確標注的情況，可能有錯誤的標注或是漏掉的標注，因此訓練模型不能遵照舊有的監督式模型來訓

doi:10.6342/NTU201702938

練。

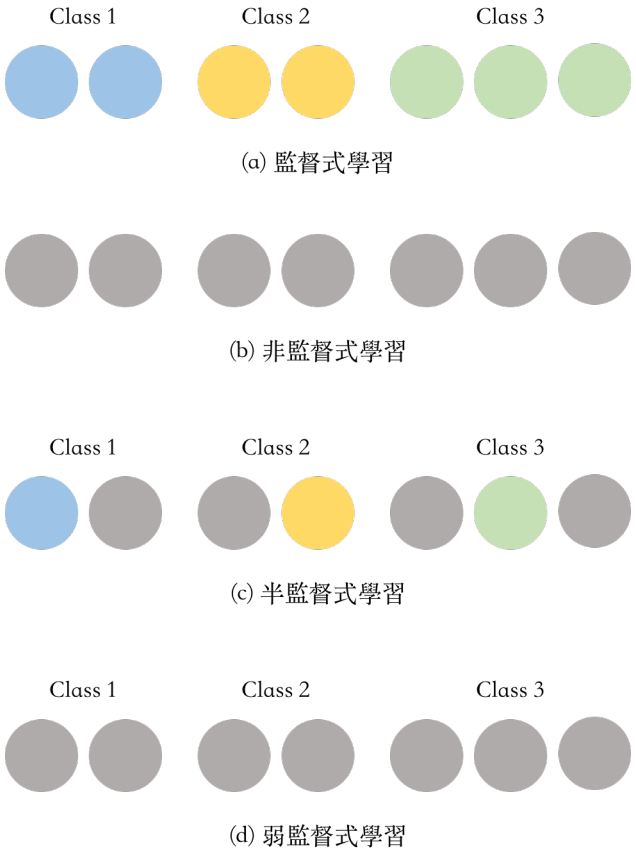


圖 4.4: 各式學習示意圖

以圖 4.4來說明，假設每個圈圈是一筆資料，皆可能對應到某個類別，在(a)中所示的監督式學習，每筆資料都有標注對應到的類別；而(b)則是非監督式學習，每筆資料皆沒有標註其歸屬的類別。(c)與(d)則是介於兩者之間，在(c)中有些資料有被標註其所屬的類別，是屬於半監督式學習；而(d)的弱監督式學習（Weakly Supervised Learning）則是指部分資料中，可能存在某筆資料屬於某個類別，但不清楚是哪一筆資料屬於這個類別。

此影像辨識模型訓練近似於(c)與(d)兩種的混合體，透過口述詞彙偵測能夠將部分的影格標註為某個類別，但還有很多影格沒有被標註，此情形近似於(c)；但其實這些標註可能不見得對，也就是說在這個場景中的確有出現關鍵用語所對應

doi:10.6342/NTU201702938

到的影像，但其位置並不恰巧對應到口述詞彙偵測所偵測到的時段。也就是說，使影像辨識模型不能夠使用監督式學習的方式來訓練，得使用其他方式來處理。

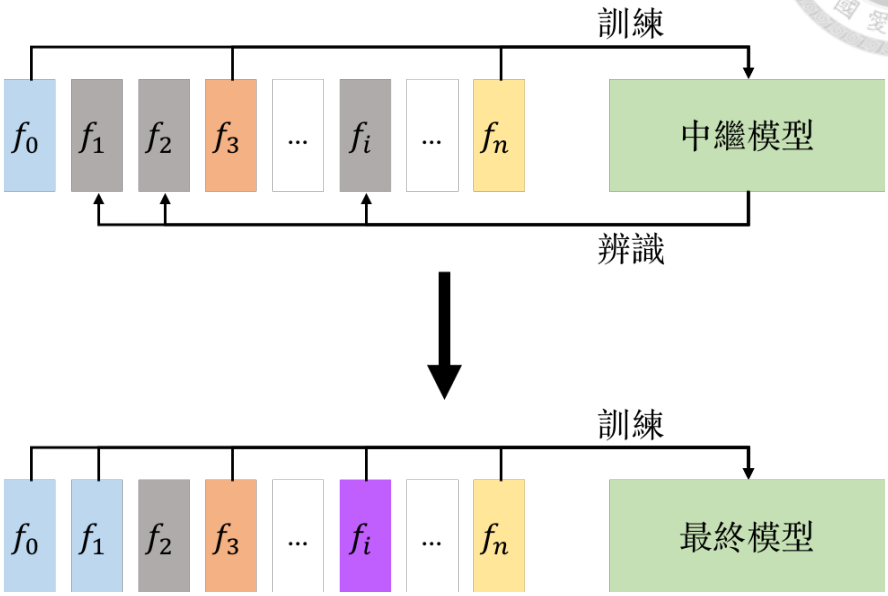


圖 4.5: 影像辨識模型訓練流程圖

為了解決上述問題，本論文使用如圖 4.5的架構來訓練影像辨識模型，此架構稱為自體訓練（Self-training）。在開始階段，會使用有標註的資料訓練模型，能夠得到一個中繼模型。接下來，以此中繼模型去辨識沒有標註的資料，模型會給每筆資料一組分數，評估其是否屬於某個類別，若分數大於預先設定的閾值，便將此筆資料歸屬於此類別。在此，訓練資料混合了中繼模型所給予的假標註（Pseudo Label）與口述詞彙偵測所給予得標註，以此資料訓練下一階段的模型。經過幾次迭代訓練，我們便能得到最終模型。而影像辨識模型的架構如圖 4.6所示。

doi:10.6342/NTU201702938

圖 4.6: 影像辨識模型架構圖

doi:10.6342/NTU201702938

4.3 實驗基礎架構

4.3.1 語料介紹

- 口述詞彙偵測

  訓練口述詞彙偵測的語音辨識模型使用LIBRISPEECH作為訓練資料，其搜集過程是藉由LibriVox此應用程式，蒐集參加者的閱讀信息、音頻與閱讀書籍的章節。

  語料分為三個部分，分別為100小時、360小時與500小時，在本實驗中使用100小時的資料來訓練模型，其中包含了125位女性講者與126位男性講者，每位講者的時間為25分鐘。

- 影像辨識訓練資料

  使用的影片訓練資料為下載自動物星球（Animal Planet）頻道中兩個系列的影集，分別為Dogs 101與Cats 101，主題為介紹不同品種的狗與貓，其中狗有107部影片，貓有67部，每部時間約為3至4分鐘。

- 影像辨識測試資料

  測試的資料是使用cifar-10 [10]，含有10個主題的圖片，包含飛機、汽車、鳥類、貓、鹿、狗、青蛙、馬、船、卡車，每個類別含有六千張圖片，訓練資料與測試資料為5 : 1，而每張圖片大小為32×32。每個類別間的資料彼此獨立，不會有一張圖片屬於兩個類別的情形。

4.3.2 訓練與辨識系統

本實驗目的為使用口述詞彙偵測來對影片中的影像進行標註，以此結果訓練影像辨識模型。在訓練口述詞彙偵測系統，使用節 4.3.1所提及

doi:10.6342/NTU201702938

的LIBRISPEECH中100小時的部分作為語料，自動語音辨識使用高斯混合模型（Gaussian Mixture Model, GMM）或是深層類神經網路（Deep Neural Network, DNN），搜尋的單位以詞（word）或是次詞單位（Sub-word Unit）。

獲得被標註的資料後，便能訓練影像辨識模型，模型架構使用卷積類神經模型，訓練方式使用自體訓練，如圖 4.5所示，先以口述詞彙偵測模型標註的資料訓練模型，再以此模型辨識其餘未被選擇的影格，挑選較高預測分數的圖片加入訓練資料，再次訓練模型。

而測試系統表現則是使用cifar-10的測試資料，以準確率來衡量模型是否成功自影片中習得資訊。

4.4 實驗設計

由於已知使用的影片為下載自動物星球頻道的影片，主題為不同品種的貓與狗的介紹，因此關鍵用語清單便包含了「狗」、「貓」與不同品種的狗或貓的名稱（例如：拉布拉多、波斯貓），其中「狗」與狗品種名稱所出現位置對應到的影格將被標註為「狗」，同理，「貓」與貓品種名稱所出現位置對應到的影格則被標註為「貓」，隨機選取等比例的影格標註為「其他」。換言之，資料共將被分為三個類別，「狗」、「貓」與「其他」。

口述詞彙偵測則分別以詞（word）與音位（phone）進行檢索，比較標註出來的資料對於訓練影像辨識模型的差異。

而基準實驗則是以AlexNet [11]為比較對象，選此模型作為比較對象而非其他在cifar-10上表現更好的模型，是因為此實驗是以cifar-10測試模型表現，衡量此系統架構是否能夠自影像中習得資訊，而非想要提升cifar-10資料的辨識率。在cifar-10上做得很傑出的模型除了提出嶄新的卷積類神經網路的架構外，更多

doi:10.6342/NTU201702938

是仰賴資料增量（Data Augumentation）、資料歸一化（Data Normalization）等技法，著重在影像辨識與電腦視覺等領域的改進，與本實驗之目的不同。採取AlexNet作為比較對象，因其訓練資料為ILSVRC（ImageNet Large-Scale Visual Recognition Challenge）中的ILSVRC-2010，並非為cifar-10，也就是說與本實驗具有相同的情境，適合作為比較對象。

4.5 實驗結果

本實驗之量化實驗結果如表 4.1所示，除了表中所列之資訊以外，本實驗亦嘗試過跳過口述詞彙偵測標註資料的階段，直接標註Dogs 101所有的影片中的影格為「狗」，標註Cats 101所有的影片中的影格為「貓」，直接訓練影像辨識模型，但此方法是不成功的，由於存在了太多與貓狗無關的圖片，模型無法成功收斂。因此使用口述詞彙偵測對影片中的影格進行標註釋有其必要的步驟。

由表格中可以發現，不論口述詞彙偵測檢索的單位為音位或詞，模型都能夠成功訓練，表現均有超出隨機猜測，代表其真的學會分辨影像中是否具有「貓」、「狗」或「其他」，但表現仍比不上具有妥善標註訓練資料的AlexNet，代表口述詞彙偵測錯誤標註的圖片，對模型造成一定程度的損害。

而比較不同的檢索單位時可以發現，使用音位進行檢索雖然可以標註較多的影像量，但其可能標註了很多不正確的影像，進而使其表現反而較差。但若是關鍵用語大多是難以辨識的詞彙時，例如專有名詞，或是背景雜訊更加嚴重，導致以詞為單位無法有效率的檢索，以音位進行檢索可能能夠獲得更好的表現。

同時也挑出了口述詞彙偵測針對「貓」與「狗」兩類別進行偵測結果中，分數最高之前9名，其中圖 4.7與圖 4.8分別是以詞為單位進行檢索，而圖 4.9與圖 4.10則是以音位為單位進行檢索。

doi:10.6342/NTU201702938

表 4.1: 實驗結果

模型架構	準確率	影像量
隨機猜測	0.333	-
AlexNet	0.816	-
以音位進行檢索	0.675	4353
以詞進行檢索	0.730	3164

可以觀察到當影片確切提及「貓」或「狗」類別的關鍵用語時，大部分的影片畫面的確是有包含貓或是狗，代表本實驗以口述詞彙偵測對影片進行標註是可行的方式，只要影片中音訊與視訊有高度相關。而標註結果都會有相似的問題，例如畫面只有主持人，但其口述內容有包含關鍵用語時，如「拉布拉多」，該畫面便會被標註為「狗」，或者影像內容只包含「狗」或「貓」部分的軀幹，這些都可能會造成訓練資料混雜，而使訓練出來的模型效能不佳。

可惜的是，由於訓練資料過於龐大，因此無法以人力標註影片中每張影格，造成無法求得口述詞彙偵測標註影片的標註準確率，但以此資料訓練出的影像辨識模型具有可以接受的準確率，應可間接代表標註的結果有一定程度的準確率。

doi:10.6342/NTU201702938

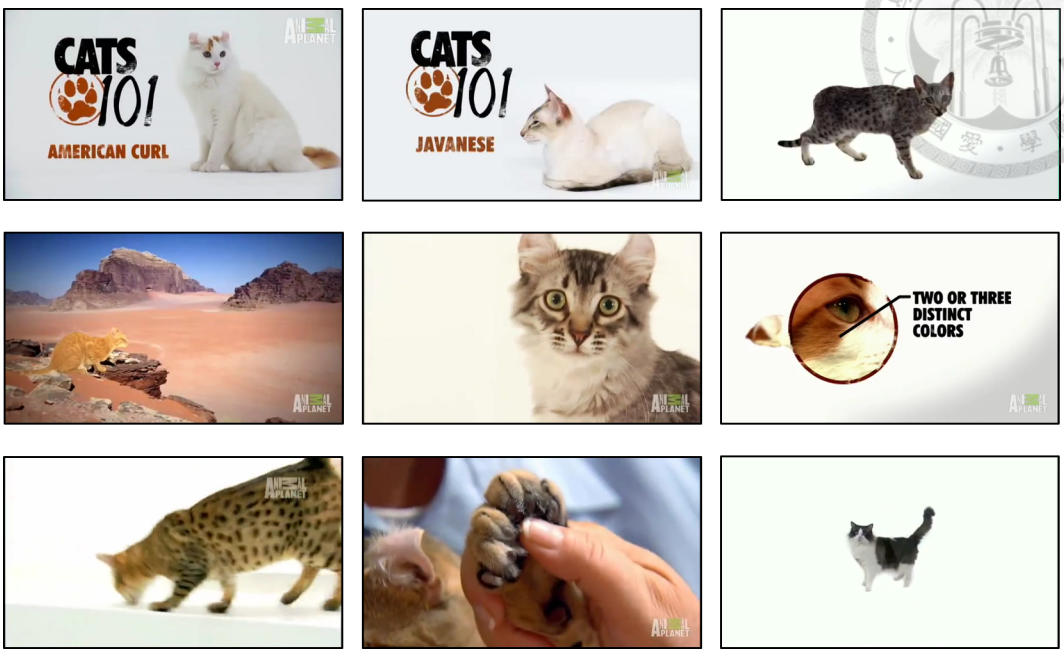


圖 4.7: 以詞為單位標註「貓」類影像之分數最高前九名

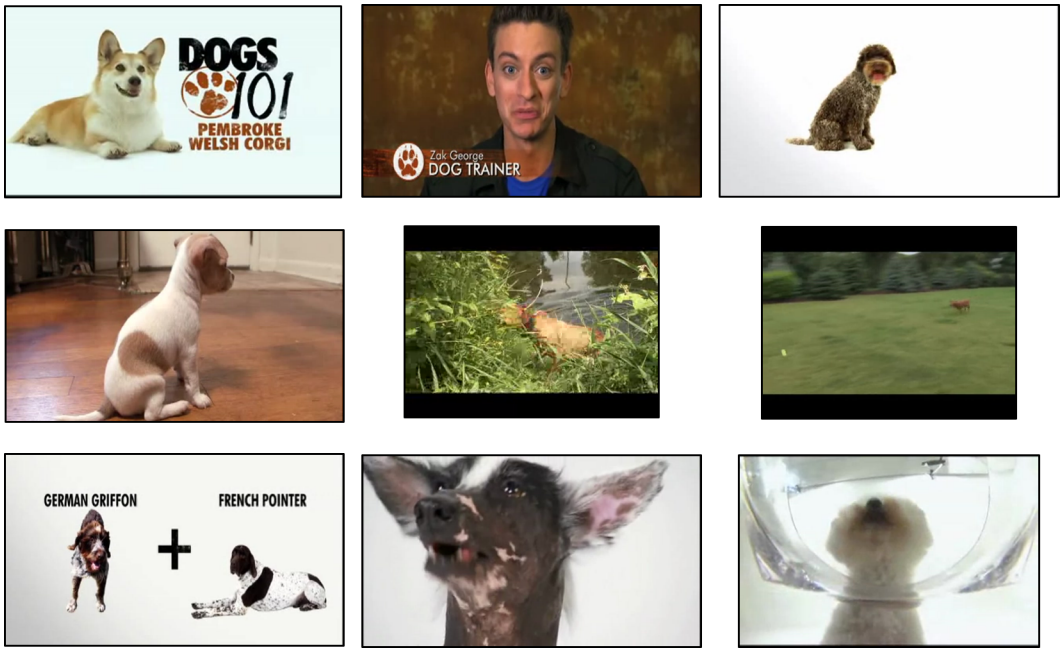


圖 4.8: 以詞為單位標註「狗」類影像之分數最高前九名

doi:10.6342/NTU201702938

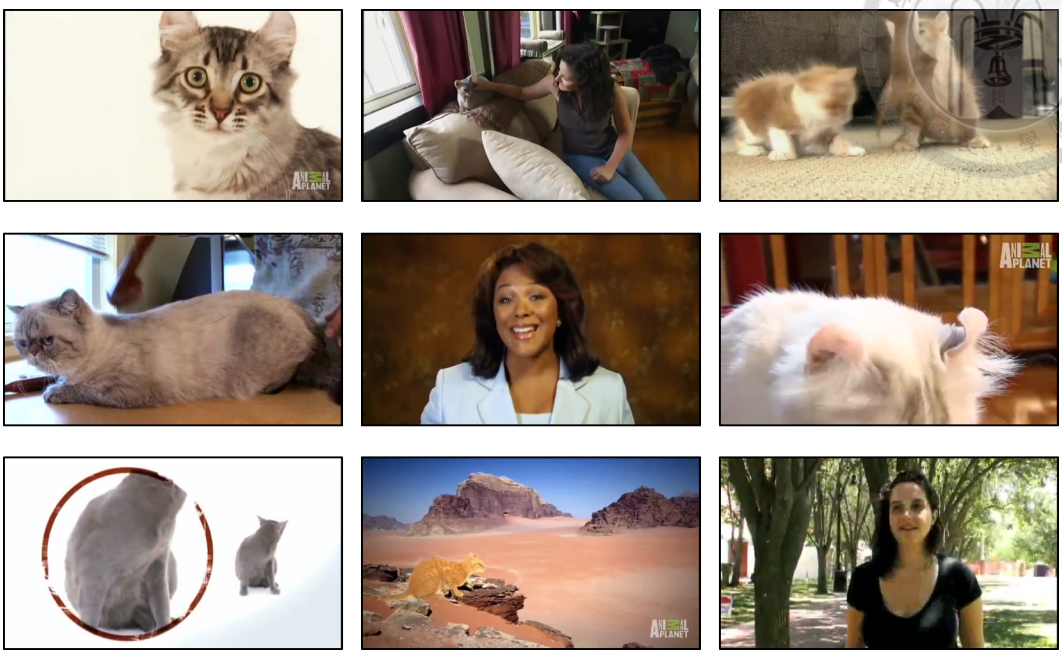


圖 4.9: 以音位為單位標註「貓」類影像之分數最高前九名

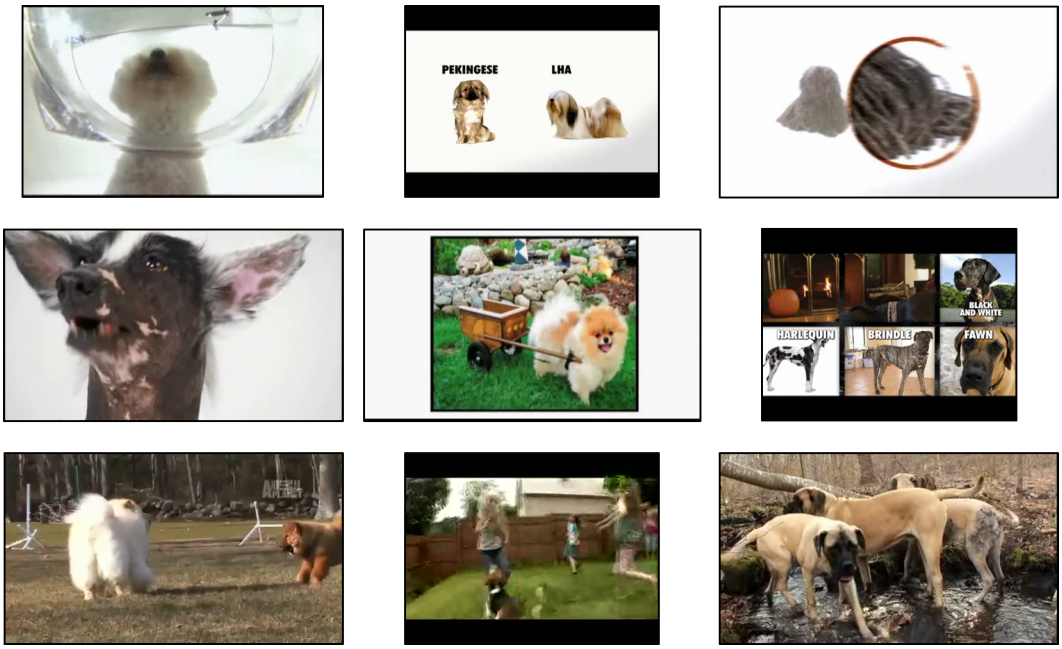


圖 4.10: 以音位為單位標註「狗」類影像之分數最高前九名

doi:10.6342/NTU201702938

4.6 本章總結

本章提出了一系統架構，預先給予關鍵用語清單，透過口述詞彙偵測對影片進行標註，只要影片的畫面與音訊具有高度相關，此系統便能夠進行有效率的標註。當面對網路上大量的影片，無法以人工給予標註，此系統便能取代人力，使取用這些影片作為訓練資料成為可行的作法。同時，本系統而夠讓模型在幾乎沒有預先資料的情形下學習資料中的資訊，彷彿是幼童在看電視的過程中，學得許多自己並不知道的觀念。

但依舊會面臨影片與畫面其實不見得總是相關，而產生錯誤標註的情況，此問題在本系統無法有效率的解決，是未來進一步研究該改善的問題。

doi:10.6342/NTU201702938

第 五 章　結合關鍵用語擷取與口述詞彙偵測訓練圖像辨識模型

5.1　簡介

在此章節中，將結合第3章的關鍵用語擷取系統與第4章使用口述詞彙偵測對影片進行標註之系統，設計出一能夠自動從影片中學出知識，不需要人為預先設定關鍵用語清單的系統，換言之，其系統運作模式與人類幼年期學習模式有所類似，透過觀看大量的影片，在沒有預設要學習什麼主題的情況下，自主學習出影片中的知識。

傳統的影像辨識系統在給予資料前往往得先設定系統學習的目標為何，將其標註後再進行訓練；與其不同的是，本系統則是給予系統資料，看系統能夠自其中學到影片中所要傳達的內容。在這種情況下，系統能夠適應快速增長的資訊量，不需要使用者給予訓練目標，便能夠自主的從影片中獲取有用的資訊。

在本章中將於第 5.2節呈現系統的架構，如何結合關鍵用語擷取系統與使用口述詞彙偵測對影片進行標註，第 5.3說明實驗的基礎設置，使用哪些語料與如何訓練與辨識系統，第 5.4節解釋實驗設計的細節，第 5.5節統整與分析實驗結果，最後在第 5.6節總結本章的內容。

5.2　架構與流程

系統的架構如圖 5.1所示，其架構與圖 4.1相當接近，其主要的差異在於關鍵用語清單的產生。在第 4章中關鍵用語清單是由使用者給予，而在本章則是透過

doi:10.6342/NTU201702938

第 3章的關鍵用語擷取系統，自影片音訊透過自動語音辨識（Automatic Speech Recognition, ASR）所生成之文字檔擷取關鍵用語，以此作為關鍵用語清單。

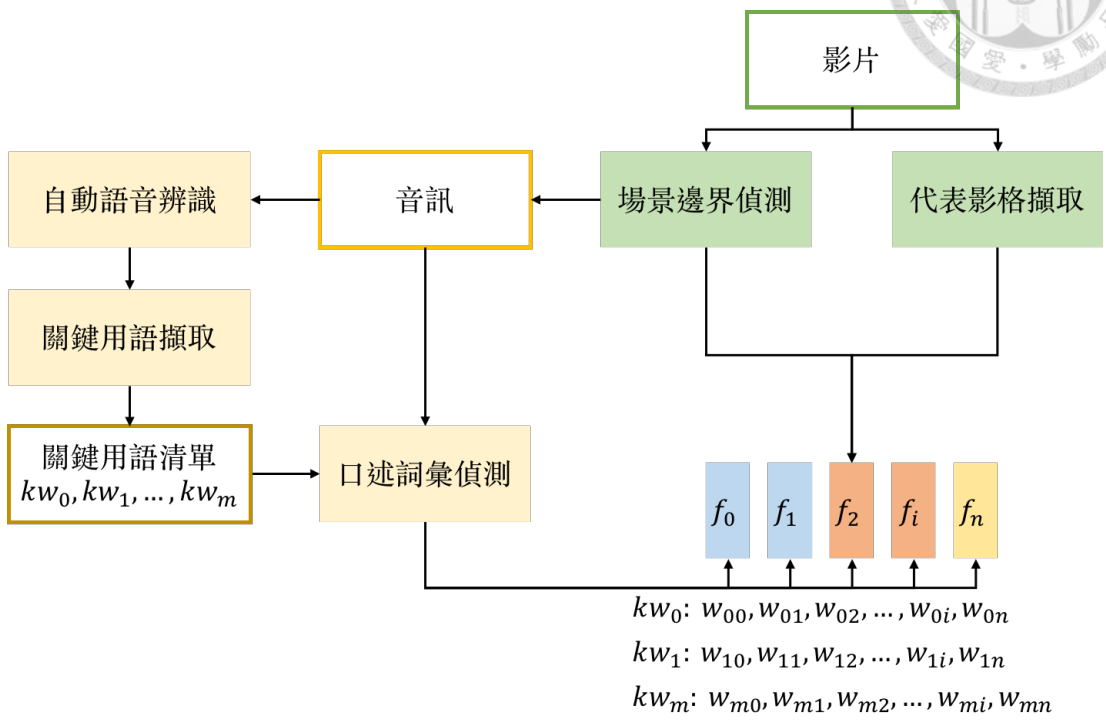


圖 5.1: 整合之系統架構

5.3 實驗基礎架構

5.3.1 語料介紹

本章實驗是結合前兩章所進行過的實驗，因此語料的使用也是相同的，分別詳列如下：

- 關鍵用語擷取系統

  使用的資料如第 3.6.1節所述，在本章實驗中同時使用了六個領域的資料來訓練模型。

doi:10.6342/NTU201702938

- 以口述詞彙偵測標註影像系統

  使用的資料如第 4.3.1節所述，以下載自動物星球頻道的影片作為訓練影像辨識模型的影片來源，其音訊經過辨識後產生的文字則透過關鍵用語擷取系統抽取出關鍵用語。

5.3.2 訓練與辨識系統

如圖 5.1所示，系統中可大致分為處理音訊與視訊兩大部分。

在音訊部分有兩個主要的構造，分別是自動語音辨識與口述詞彙偵測，自動語音辨識使用的是CMU Sphinx系統，藉由此系統將音訊轉為文字，此系統相較於Google或其他自動語音辨識系統，具有可以離線於本機端操作、快速等優點，缺點是準確率不高，容易受到雜訊影響。將音訊轉為文字後，便能夠以第 3章的關鍵用語擷取系統產生關鍵用語清單，以此清單作為口述詞彙偵測檢索的目標，而口述詞彙偵測系統與第 4章相同。

而視訊部分處理方式也與第 4章相同，自影片中抽取代表影格、偵測場景邊界，再以口述詞彙偵測系統給予影格標註。

5.4 實驗設計

在本實驗中會使用第 3章與第 4章中最好的參數設置，關鍵用語擷取系統會分別使用第 3.4節所介紹的卷積類神經網路模型與增強式長短期記憶類神經網路模型，而口述詞彙偵測則以詞為單位進行檢索。將兩章所使用的模型結合，建構出能夠自影片中自動學習出資訊的系統。

doi:10.6342/NTU201702938

5.5 實驗結果

實驗結果將會分成兩部分來進行分析討論，分別為關鍵用語擷取與影像辨識模型。在關鍵用語分析中會比較不同模型擷取關鍵用語表現的差異，而在影像辨識模型則比較對於自動產生的關鍵用語清單對訓練影像辨識模型所造成的影響。

5.5.1 關鍵用語擷取系統分析

將影片中的音訊透過自動語音辨識轉換成文字後，每部影片都能夠得出一篇對應的文章，將其輸入關鍵用語擷取系統，便能夠輸出文章中的關鍵詞。

表 5.1呈現了第 3章中兩個不同模型所節取出關鍵用語之比較。將Dogs101與Cats101這兩個影集中每一隻影片的關鍵用語出現次數相加，各自取出前五名，可以發現卷積類神經網路模型與增强式長短期記憶類神經網路模型在選擇關鍵用語上，有很不同的傾向。

可以觀察到卷積類神經網路模型選擇的關鍵用語比較「抽象」，例如陪伴關係、愛、忠誠，推測可能是因為卷積類神經網路具有化繁為簡的能力，能夠自低維度的資料中萃取出高維度的抽象概念，藉此決定這抽象概念所對應到的關鍵用語。同時由於貓狗皆為伴侶動物，介紹過程中自然會和家庭有所關連，因此家庭、孩童等詞也成為了關鍵詞。

而增强式長短期記憶類神經網路模型抽取的關鍵用語則較為具體，比較符合預期的需求，可以觀察到由於主題是不同品種貓狗的介紹，因此抽取出了狗、貓、健康、品種等詞彙，而這些關鍵用語都並沒有出現在訓練資料的六大領域（生物、廚藝、旅遊、機器、密碼學與手作）中，因此再次證明第 3章所訓練出的系統效能不錯。

除了表中列出的前五名關鍵用語，在不同的影片中，國家往往也是關鍵用

doi:10.6342/NTU201702938

表 5.1: 關鍵用語擷取結果

名次	卷積類神經網路模型		增强式長短期記憶類神經網路模型	
	Dogs 101	Cats 101	Dogs 101	Cats 101
1	家庭（family）	家庭（family）	狗（dogs）	貓（cat）
2	孩童（children）	陪伴關係（companionship）	活力（energy）	健康（health）
3	訓練（training）	家庭的（household）	歐洲（europe）	基因（gene）
4	主人（owner）	愛（love）	生活（life）	食物（food）
5	忠誠（loyalty）	孩童（children）	健康（health）	品種（breed）

語，例如英國、俄羅斯等等。推測原因有可能是受到訓練資料中「旅遊」主題的影響，而對國家等地裡資訊特別敏感；同時，不同品種的貓狗時常也是來自於不同的國家地區，因此這些表達地理詞彙的用語合理地成為關鍵用語可能之一。

5.5.2 影像辨識模型

原預計以抽取出的關鍵用語作為類別，例如將影像標註為家庭、孩童等。可是此作法會面臨到資料量不足，因為關鍵用語出現的次數不多，或是抽像詞彙難以合理的配對到影像，例如健康、忠誠難以成為影像的類別。因此依舊將出現在Dogs 101中的關鍵用語歸為「狗」，Cats 101中的關鍵用語歸為「貓」，再隨機抽取沒有被選到的影格作為「其他」。

以不同方式產生的關鍵用語清單訓練影像辨識模型的結果如表 5.2所示。由表可以觀測到與給定之關鍵用語清單比較，不論使用何種方式產生之關鍵用語清單，所訓練出來的模型準確度均較差，推測是因為關鍵用語擷取系統雖然能夠抽取出代表文章內容的關鍵用語，但文章內容有非常多的面向，而影像的內容往往

doi:10.6342/NTU201702938

只包含了較為具體的面向。舉例而言，影片中可能涵蓋了狗的長相特徵介紹、與家庭成員的互動、如何訓練等內容，但除了長相特徵能夠以影像呈現外，其他大多會以口白方式呈現，因此抽取出的關鍵用語難以作為影像的標註。

表 5.2: 實驗結果

模型架構	準確率	影像量
給定之關鍵用語清單以詞進行檢索	0.730	3164
卷積類神經網路關鍵用語擷取	0.6967	1557
增强式長短期神經網路關鍵用語擷取	0.6983	2610

而圖 5.2到圖 5.5為不同關鍵用語清單以口述詞彙偵測針對「貓」與「狗」兩類別進行偵測結果中，分數最高前九名之結果。可以觀察到其常常會擷取出一些在片頭片尾之影像，例如畫面中有健康（health）、家庭（family）的影格，推測是因為在影片開頭會快速跑過這些畫面，說明在影片中會提及這些類別的內容。這些影格會對影像偵測系統造成一定程度的影響，因為畫面中有大量的雜訊，這很可能是造成影像辨識系統效果不佳的原因。

doi:10.6342/NTU201702938

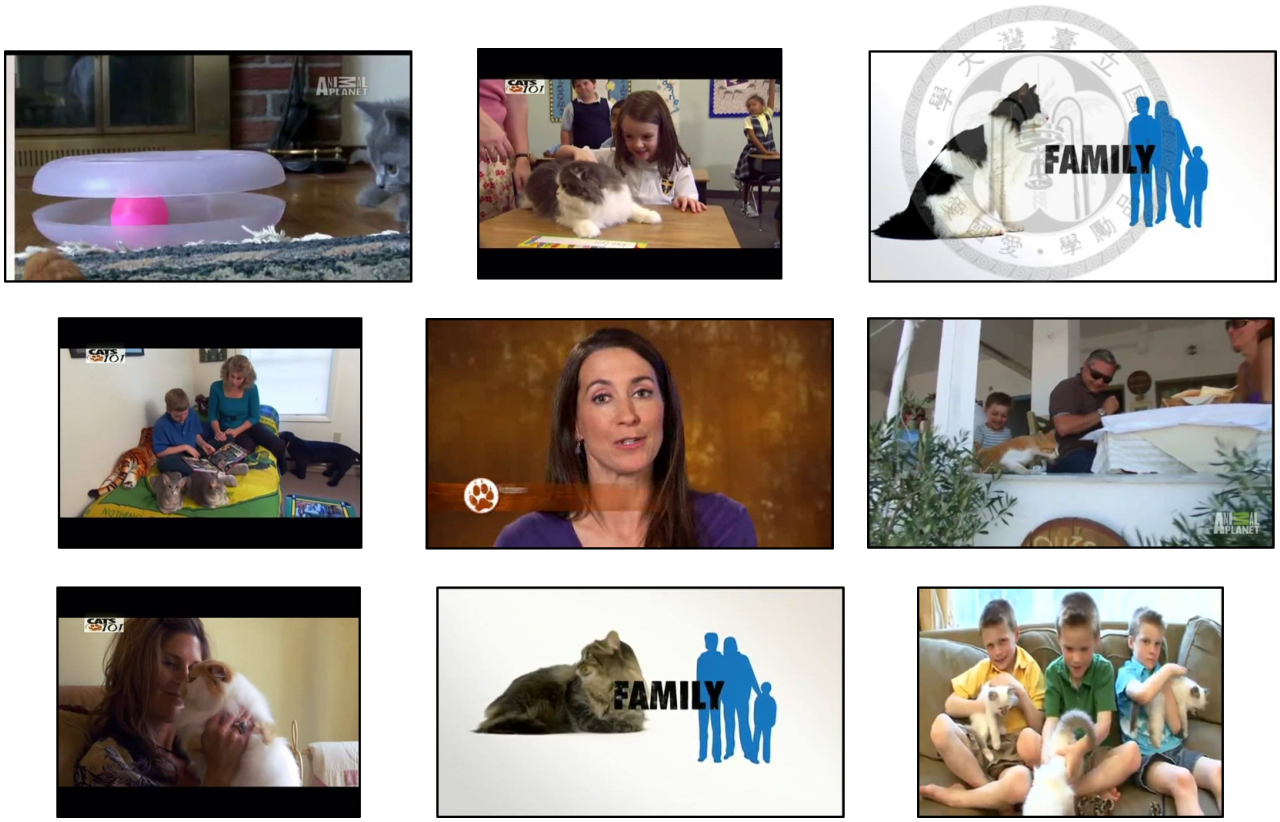


圖 5.2: 卷積類神經網路產生關鍵用語清單標註「貓」類影像

圖 5.3: 卷積類神經網路產生關鍵用語清單標註「狗」類影像

doi:10.6342/NTU201702938

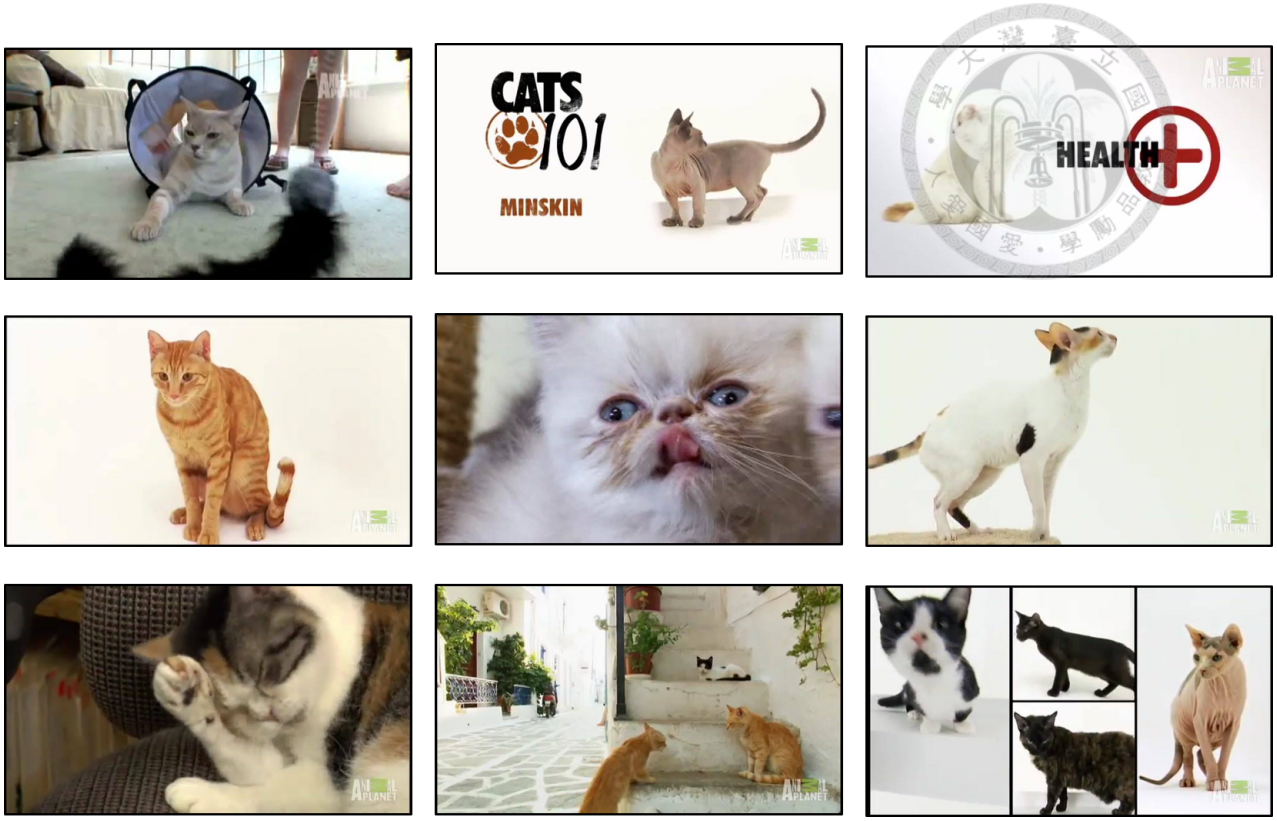


圖 5.4: 長短期記憶類神經網路產生關鍵用語清單標註「貓」類影像

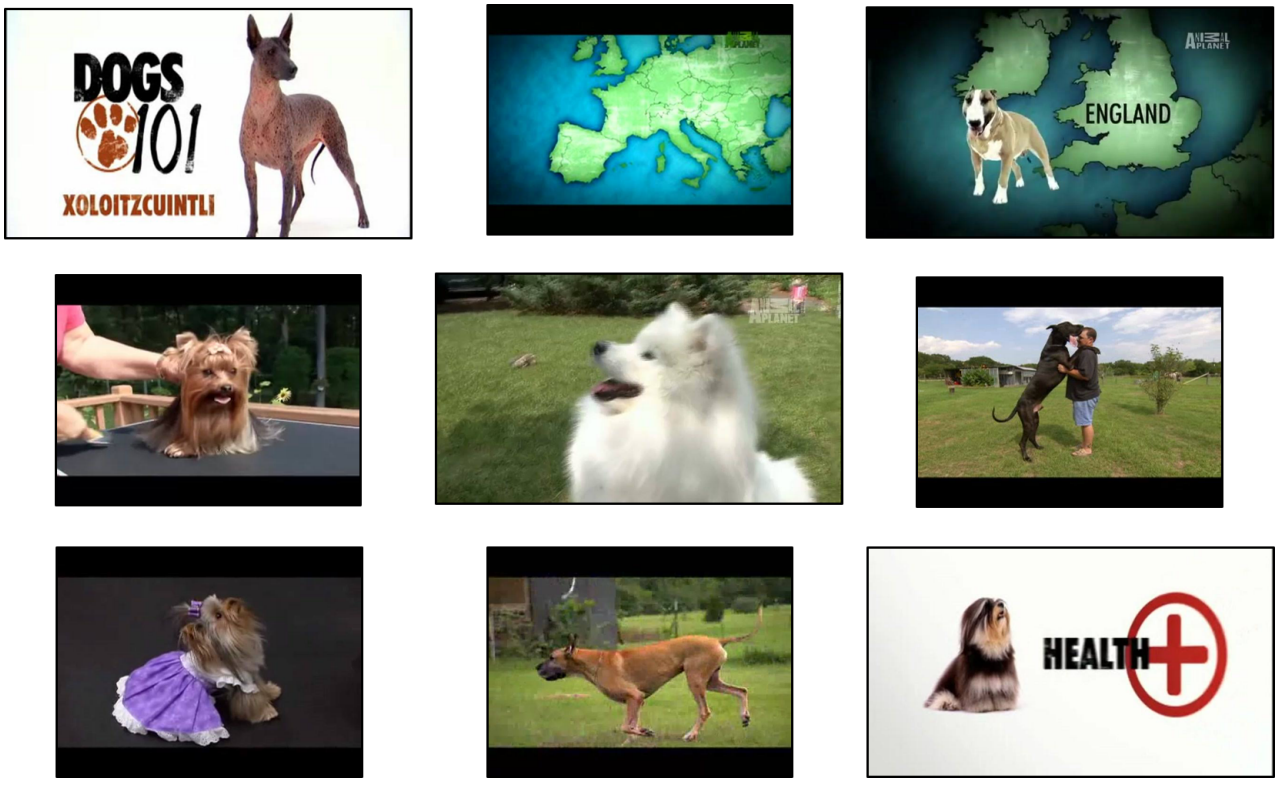


圖 5.5: 長短期記憶類神經網路產生關鍵用語清單標註「狗」類影像

doi:10.6342/NTU201702938

5.6 本章總結

在本章結合了第 3章的關鍵用語擷取系統與第 4章使用口述詞彙偵測對影片進行標註之系統，根據結果可以發現關鍵詞擷取系統能夠有效自沒有見過的類型的文章中擷取關鍵用語，但是此系統與口述詞彙偵測對影片進行標註之系統有銜接上的問題。

若要使其具有更良好的表現，可能得蒐集更多的影片，好讓每個類別（關鍵用語）都能夠有足夠的圖片作為訓練資料，或者加入更多影像處理的技巧，降低具有大樣雜訊的訓練資料對系統正確率的影響。

doi:10.6342/NTU201702938

第六章　結論與展望

6.1　結論

本論文提出能夠自網路上海量影片學習資訊的系統，模仿人類幼年期學習的模式，觀看電視影集自主學習未知的知識。而此系統分為兩個部分，一個是關鍵用語擷取系統，另一部分則是以口述詞彙偵測對影片進行標註。

在第三章中，本論文提出一嶄新的關鍵用語擷取系統，以位置資訊為預測目標，不同於分類式關鍵用語擷取系統，其能夠藉由觀察句子結構、各式特徵而抽取出沒有見過的關鍵用語。其中模型有兩種核心架構，分別為卷積類神經網路與長短期記憶類神經網路，兩者效能各有差異，可視使用者需求進行選擇，同時也使用增強式學習的技法，強化模型的訓練，獲得更好的效能。

而在第四章則以口述詞彙偵測系統，針對預先給定的關鍵用語清單，針對影片進行標註，以這個方式取代人工標註，以達到有效使用網路上大量影片作為訓練資料的目標。雖然有時候影像與音訊不見得完全相關，但此架構所標記出來的資料，給予影像辨識模型作為訓練資料，所得出的準確率尚可接受，換言之針對影像與音訊有高度相關的影片，例如各式教學影片、介紹知識的影集，是有效的自動標註系統。而此結果也符合本論文的目標，因人類幼年期獲取新知的重要管道便來自各式具有教學意涵的影片。

在第五章則將兩系統結合，以關鍵用語擷取系統自影片的音訊中擷取出關鍵用語，產生關鍵用語清單，再讓口述詞彙偵測系統標註影片。但其效果不如預期，因關鍵用語擷取系統所擷取的關鍵用語包含了各種面向，有些抽象詞彙無法呈現在影片中，換言之這些關鍵用語所出現時的畫面往往並不與其相關。但整體而言，實驗結果依舊證明此架構可行，能夠讓機器自影片中學出內容，並不需要

doi:10.6342/NTU201702938

人類給予過多的指導。

6.2 未來研究方向

在關鍵用語擷取方面，可以嘗試在特徵抽取上下更多的功夫，同時讓模型看過更多類型的文章，不論是更廣泛的內容或是其他的文體（如新聞、專欄文章等）也對其表現有所幫助。

而訓練影像辨識模型則可以運用更多影像的知識，例如進行更細緻的前處理，針對影像去噪、歸一化，或是使用其他模型輔助找出影像中的物件，框出邊界盒（Bounding box），減少雜訊對訓練模型的傷害。

同時也能夠考慮多媒體（Multimodal）的模型，或是以端對端（End-to-end）的方式訓練模型，取代本論文中將各部分模型各別訓練妥當再接在一起的作法，以期自資料中獲得更多資訊，獲得更好的系統。

doi:10.6342/NTU201702938

參 考 文 獻

[1] Yann LeCun, Bernhard E Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne E Hubbard, and Lawrence D Jackel, "Handwritten digit recognition with a back-propagation network," in *Advances in neural information processing systems*, 1990, pp. 396–404.

[2] Chris Burges, Tal Shaked, Erin Renshaw, Ari Lazier, Matt Deeds, Nicole Hamilton, and Greg Hullender, "Learning to rank using gradient descent," in *Proceedings of the 22nd international conference on Machine learning*. ACM, 2005, pp. 89–96.

[3] Ali Sharif Razavian, Hossein Azizpour, Josephine Sullivan, and Stefan Carlsson, "Cnn features off-the-shelf: an astounding baseline for recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2014, pp. 806–813.

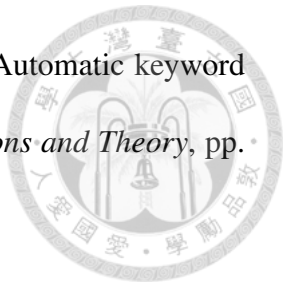
[4] Felix A Gers, Jürgen Schmidhuber, and Fred Cummins, "Learning to forget: Continual prediction with lstm," 1999.

[5] Sheng-syun Shen and Hung-yi Lee, "Neural attention models for sequence classification: Analysis and application to key term extraction and dialogue act detection," *arXiv preprint arXiv:1604.00077*, 2016.

[6] Rada Mihalcea and Paul Tarau, "Textrank: Bringing order into text.," in *EMNLP*, 2004, vol. 4, pp. 404–411.

[7] Chia-hsing Hsu and Hung-yi Lee, "Enhanced spoken term detection by deep learning," M.S. thesis, 2017.

doi:10.6342/NTU201702938



[8] Stuart Rose, Dave Engel, Nick Cramer, and Wendy Cowley, "Automatic keyword extraction from individual documents," *Text Mining: Applications and Theory*, pp. 1–20, 2010.

[9] Tony Lindeberg, "Scale invariant feature transform," *Scholarpedia*, vol. 7, no. 5, pp. 10491, 2012.

[10] Alex Krizhevsky and Geoffrey Hinton, "Learning multiple layers of features from tiny images," 2009.

[11] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

doi:10.6342/NTU201702938