中華民國第 60 屆中小學科學展覽會

作品說明書

高級中等學校組　電腦與資訊學科

(鄉土)教材獎

052508

實現圖像敘述自動生成之中文化

學校名稱：新北市立中和高級中學

作者：	指導老師：
高二 柯勁廷 高一 劉力瑋	王一哲

關鍵詞：深度學習、神經網路、圖像標註

摘要

目前市面上已經有不少產品應用到自然語言處理 (natural language processing, NLP) 的技術，例如聲音或是圖像標註，但是關於圖像標註的研究大部分是針對英文，因此我們想要進行中文的圖像標註研究。在本研究中，我們嘗試建構神經網路，並使用 Microsoft COCO: Common Objects in Context 數據集訓練模型，並試著調整模型以達到較好的圖像標註效果。使用深度學習神經網路作為翻譯流程的工具，並嘗試配合各種不同的網路架構。為了能達到從序列到序列 (seq2seq) 的效果，我們的網路使用了編碼器 - 解碼器 (encoder-decoder) 的結構，編碼器的部分使用 CNN，解碼器的部分使用 RNN，這樣能有效地傳達序列並有較多種神經網路組合方式。

壹、研究動機

一直以來在電腦視覺中，自動圖像標註生成是個重要的任務，有許多重要的功能，例如協助視障者理解網頁內容。而在自動圖像標註生成任務中，模型不但需要辨認圖像中的項目，還必須生成合理的自然語言，才能適當地標註圖像中的內容。然而在這方面中文的研究極為稀少，因此我們希望整理現有模型，經過調整之後並應用於中文上，讓中文圈擁有自動圖像標註生成的系統。

貳、研究目的

一、探討中文圖像標註自動生成的可行性。

二、探討不同模型對於中文圖像標註自動生成的效果。

三、嘗試最佳化模型結構。

四、嘗試製作便於使用的應用程式。

參、研究設備及器材

一、訓練資料集

（一）Microsoft COCO: Common Objects in Context 2014 圖片數據集 [1]

（訓練用資料約 13000 張，驗證用資料約 6000 張）

（二）Microsoft COCO: Common Objects in Context 2014 中文詞句數據集 [2]

（訓練用資料約 13000 張，驗證用資料約 6000 張，以下皆簡稱為 MSCOCO）

二、訓練設備及框架

由於訓練卷積神經網路 (Convolutional Neural Network, CNN) 與長短期記憶循環神經網路 (long short-term memory, LSTM) 及其他相關神經網路需要占用大量顯示記憶體，因此我們的模型均在以下設備訓練：

（三）工作站電腦

1. 顯示卡：GEFORCE® GTX 1660 Ti
2. 記憶體：16 GB DDR4

（四）作業環境

為了達到更好的相容性並善用相關工具，我們採用 Linux 為作業系統。

（五）神經網路框架、工具

1. Python 3.6 （主要程式語言）
2. TensorFlow-GPU 1.13.1 （深度學習框架）

肆、研究過程或方法

一、名詞解釋

以下是本研究中會使用到的專有名詞，為了能讓初學者也能快速讀懂我們的研究成果，我們簡單地解釋這些常用的名詞。

表 1　常用的深度學習專有名詞

中文	英文	解釋
激勵函數	activation function	用來增加神經網路的非線性，常見的有 Sigmoid、tanh、ReLU。
損失函數	loss function	衡量神經網路產生內容與實際內容差距的函數。
卷積神經網路	convolutional neural networks, CNN	利用卷積層及池化層提取圖形的特徵，經常用於圖形辨識。
循環神經網路	Recurrent neural networks, RNN	擁有記憶性的神經網路，常用於需要處理整個數據集關聯性的任務。
迭代	iteration	將資料輸入至人工神經網路訓練的過程。
批量	batch	固定數量的數筆資料。
批量大小	batch size	一次輸入多少筆資料。
全局步數	global step	不計迭代次數，一共訓練了幾次批量。
學習率	learning rate	更新學習值在梯度位置的參數
交叉熵	cross entropy	計算機率分布的距離，本研究中使用交叉熵計算損失。
評價	evaluation	用來判斷分類系統優劣的方式，常見的評價方式為機器評價，評價分數越高代表成效越好。

第一種常用的激勵函數為 S 函數 (sigmoid)，定義為

$$S(x) = \frac{1}{1+e^{-x}} = \frac{e^x}{e^x+1}$$

上式中 e 為自然對數函數的底數，量值約為 2.178。sigmoid 常用來表示布林值 (boolean)，它的值域在 0 和 1 之間，但是收斂較緩慢。第二種常用的激勵函數為雙曲正切函數 (tanh)，定義為

$$\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$$

它的值域在-1 到 1 之間，收斂速度較快，比較容易優化，但仍然有梯度消失的問題。第三種常用的激勵函數為線性整流函數 (rectified linear unit, ReLU)，定義為

$$f(x) = \max(0, x) = \begin{cases} x & \text{if } x > 0 \\ 0 & \text{otherwise} \end{cases}$$

因為收斂速度快而被廣泛使用。

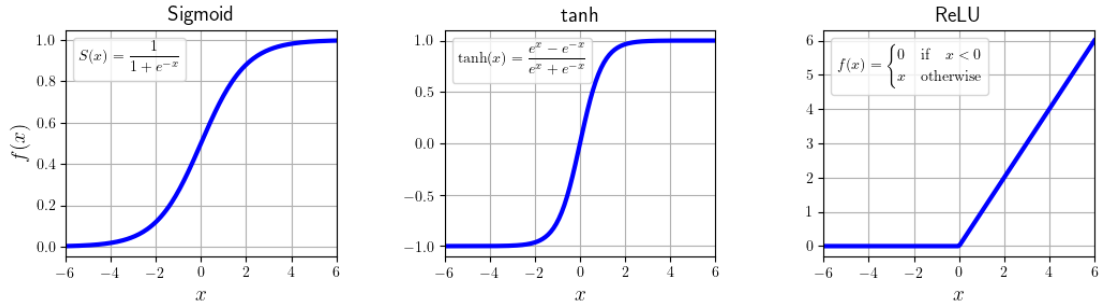


圖 1　常用的激勵函數

二、文獻探討

（一）人工神經網路 (Artificial Neural Network, ANN)

這是機器學習領域當中一種常用的方法，藉由電腦程式模擬生物的神經網路系統，藉此處理一些傳統程式設計上難以解決的問題，例如圖形辨識和語音辨識。人工神經網路通常由許多層所組成，常見的層有以下幾種：

1. 全連接層 (Fully Connected Layer)

   全連接層又名稠密層 (dense layer)，是最簡單的網路結構之一。其中有數個權重 (weight) 以及偏移值 (bias)，用以控制各個神經元連接的方式。

2. 卷積層 (Convolutional Layer)

   卷積層由許多卷積單位所構成，每個卷積單位的參數會由反向傳播最佳化。卷積運算的目標是提取的不同層級特徵，第一層卷積層可能只提取一些低階局部的特徵，例如邊緣、線條……等層級，而更深的結構能夠提取更複雜的特徵。

3. 丟棄層 (Dropout Layer)

   在每個訓練批量中，不採用指定比例的特徵節點，可以明顯減少過擬合 (overfitting) 的現象，使模型泛用性更強，不會過度注重某些局部的特徵。

4. 嵌入層 (Embedding Layer)

   具有相似意義的詞具有相似的表示，根據多維空間中詞與詞之間有多少相似性，將詞彙對映到實數向量的結構，這使我們能將原本獨立的單詞向量，有更多的關聯性，進而降低維度。

5. 卷積神經網路 (Convolutional Neural Networks, CNN)

   卷積神經網路一般包含數個卷積層，以及少數全連接層，常用在圖形分類問題。卷積層會將圖形轉換成特徵圖，並經由全連接層輸出結果。

6. 全卷積網路 (Fully Convolutional Networks, FCN)

   所有的網路結構都由卷積層及池化層 (pooling layer) 組成，不包含全連接層，或是將轉換後的特徵圖經由上採樣解碼為圖片，早期被應用於圖像語意分割。

7. 循環神經網路 (Recurrent Neural Network, RNN)

循環神經網路主要是用來處理序列數據，其特點在於決定序列中每個元素的輸出時，會將當前輸入的元素與先前的輸入輸出一同運算。常應用於機器翻譯、語言辨識等。

（二）高階人工神經網路結構

以下列出目前常見的人工神經網路結構：

1. Inception V3 [3]

Inception 是 Google 提出的一種卷積神經網路，在 2014 ImageNet 大規模視覺識別大賽 (ImageNet Large Scale Visual Recognition Competition, ILSVRC)中得到冠軍，模型總共有四代，其中我們所使用為三代，雖然結構深層且複雜，但相對的參數量較少，識別率較高，因此被廣泛用於圖像處理。

2. VGG-19 [4]

VGG-19 是一個典型的卷積神經網路模型，在 2014 ImageNet 大規模視覺識別大賽 (ImageNet Large Scale Visual Recognition Competition, ILSVRC) 的 top-1 及 top-5 正確率分別達到了 74.5% 及 92.0%。其結構為 16 個卷積層及 3 個全連接層，使用大量的 3 × 3 卷積核，與較大的卷積核相比，不但提升了非線性度，更使需要訓練的參數較少。

3. 長短期記憶 (Long Short-Term Memory, LSTM) [5]

這是循環神經網路的一種衍生版本，主要是為了解決長序列訓練過程中的梯度消失和梯度爆炸問題，相較於 RNN，LSTM 能夠在更長的序列中有更好的表現。LSTM 內部主要有三個控制節點的開關：(1) 遺忘閥、(2) 選擇閥、(3) 輸出閥。遺忘閥：這個參數主要是對上一個節點傳進來的訊息選擇性遺忘，通過計算得

到 z_f（f表示遺忘 forget）作為遺忘開關，控制上一個輸出的記憶程度。選擇閥：這個參數將這個階段的輸入 c_{t-1} 進行選擇記憶，開關的參數則是由z_i（i 代表 information）進行控制。輸出閥：這個參數將決定哪些將會被當成目前狀態的輸出值，這是透過 z_o 來控制，並且還對上一階段得到的 c_o 透過激勵函數 tanh 進行縮放。

4. Gate Recurrent Unit, GRU [6]

GRU 是 RNN 的一種分支，和 LSTM 一樣，也是為了解決長期記憶和反向傳播中的梯度等問題而提出來的。GRU 的結構相對於 LSTM，是將 LSTM 中的遺忘閥與輸入閥用一個更新閥取代，並把單元狀態 (cell state) 和隱藏狀態 (hidden state) 進行合併。

5. 注意力機制 (attention mechanism) [7]

注意力機制是近年自然語言處理中較新的一種機制，其工作原理是將序列作量化輸出，使模型能注意到序列中較大的值，目的是給予模型識別的能力，從訊息中找到應該注意的重點，目前已經用於圖像分類、圖像標註……等深度學習應用。

（三）參考模型

本研究主要參考以下兩篇論文的模型並加以修改，本研究與原文的不同之處在於，我們將編碼器的部分由 LSTM 改為 GRU，使需要訓練的參數減少為原來的 1/20，可以有效的加快訓練速度。

1. Show & tell [8]

機器翻譯是輸入圖片生成形容文字的作業，對於機器來說，要偵測物體以及判斷出他們之間的關聯，最後再組成一個句子，這是非常不容易的事。本質上，機器翻譯的工作就是從一段不定長的序列轉換為另一段不定長的序列，如今實現 seq2seq 最有效的方法即為 LSTM，在 Show & tell 的研究中，作者引用了 encoder-decoder 的結構，以 CNN 作為 encoder，以 LSTM 作為 decoder，讓序列之間的轉換變得更為容易。

2. Show, attend & tell [9]

在 Show, attend & tell 的研究中，作者於 Show & tell 的基礎上加入了注意力機制 (attention mechanism)，這是為了協助圖像標註任務，使模型更容易找到對應的目標，而且不影響最後翻譯出的語意。

三、尋找解決方法

（一）建構整體實驗流程

基於目前學到的知識以及研究目標，我們必須依照數據集的類型，設計適合的神經網路。圖 2 為本研究的實驗流程圖：

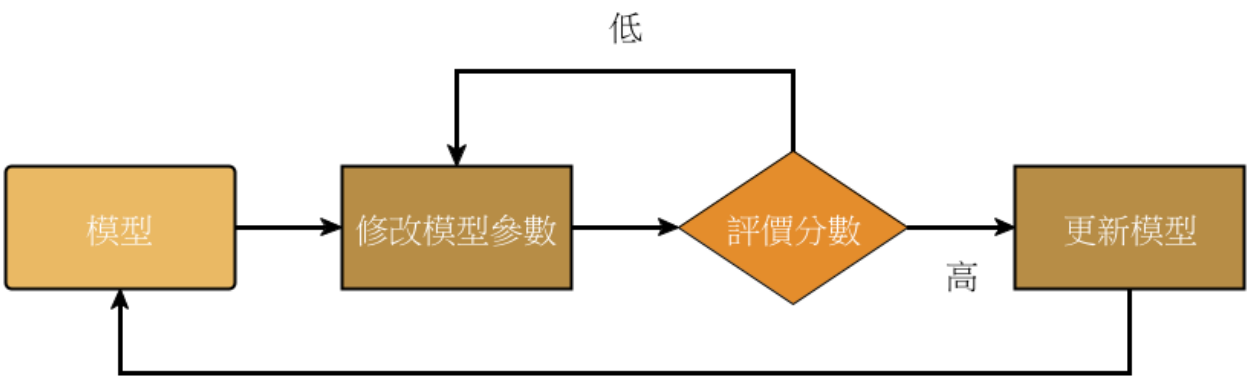


圖 2　實驗流程

（二）尋找合適之深度學習框架

以下為目前常見之深度學習框架：

1. TensorFlow

   這是 Google 於 2015 發表的框架，以 Python、C++與 CUDA 開發而成，編寫程式碼的方式極具特色，需要先建立計算圖 (graph) 及存放資料用的佔位符 (placeholder) ，再開啟運算需要用的對話 (session)。雖然對於剛接觸的新手可能較為複雜，但熟練後即可快速應用，而且自由度較高。使用 TensorFlow 提供 TensorBoard 工具，可即時觀看訓練的過程。

2. Keras

   這是一種高級神經網路的應用程式介面 (application programming interface, API)，以 Python 開發而成，可以使用 TensorFlow、Theano 與 CNTK 作為後端運作的框架，讓使用者可以用簡單且少數的程式碼架設神經網路模型。

3. PyTorch

   前身為 Torch，由於 Torch 是以 Lua 開發而成的，使用者較少。後來開發適用於 Python 的版本，並於 2016 年底發表，是目前使用者人數成長相當快速的深度學習框架。

四、準備及整理訓練數據

（一）MSCOCO 圖片數據集

MSCOCO 數據集是由微軟 (Microsoft) 建構的，其中包含 detection, segmentation, keypoints 等部分。本實驗使用的是於 2014 年發布的 MSCOCO 數據集，此數據集已經被廣泛使用於圖像分類研究中。數據集有 80000 張的訓練圖片 ，40000 張的驗證圖片 ，以及 500 MB 的標籤文件。

（二）MSCOCO 中文詞句數據集

這個數據集是在 MS COCO 資料集中，重新標註為中文句子所組成的資料集。有 18341 張的訓練圖片，1000 張的測試圖片，以及 999 張測試驗證，表 2 為資料集中圖片與對應形容語句的範例。

表 2　MSCOCO 中文詞句數據集範例

圖片	形容文字
	1. 餐廳裡有現代化的木桌和椅子。 2. 長有藤圓椅的餐廳餐桌 3. 一張長桌子，上面有一株木頭圍著的椅子。 4. 中間有插花的長桌子供會議用。
	1. 在廚房裡用糖霜做甜點的人 2. 一位廚師正在準備和裝飾許多小點心。 3. 麵包師準備各種各樣的烘焙食品。 4. 在容器裡抓糕點的人的特寫鏡頭
	1. 廚房的架子上擺滿了香料。 2. 帶有烤箱和其他配件的廚房 3. 利用所有空間的小廚房 4. 這個小廚房裡有鍋、平底鍋和香料。

（三）Flickr8k 圖片數據集

於 2013 年提出，同樣是個有各種各樣照片的資料及，根據作者，收集的圖片都不包含名人或有名的地點，而是手動選擇多種場景和情形，有 8000 張的訓練圖片 。

（四）Flickr8k 中文詞句數據集

根據圖片數據集衍伸出來的數據集，用五句中文詞語來描述上面資料及的圖片，故總共有 8000 × 5 = 40000 句中文描述。

表 3　Flickr8k 中文詞句數據集範例

圖片	形容文字
	1.一個穿着泳裝的女人躺在海水裡 2.一個女人在水裡躺着 3. 女人穿比基尼躺在海水中 4. 女孩躺在海邊 5. 一個小孩穿著泳衣躺在水裡
	1. 大人領著孩子們在水池邊 2. 一群小孩在看魚 3. 一群人在泳池邊玩 4. 一群人在池塘邊看魚 5. 一群人在看金魚
	1. 一個男人坐在長椅上面煮東西 2. 一個男人坐在長凳上煮東西 3. 男人在煮飯 4. 男人坐在長椅上休息 5．一個男人在做飯

五、建構神經網路

由於我們需要一次處理大量的資料，因此我們先將圖片數據集統一集中至串列 (list) 中，透過 pickle 打包成 .pickle 檔案格式後，再根據參數分批輸入解碼器進行訓練。

（一）編碼器 (Encoder)

我們先將龐大的數據集經過人工分式分類後，統一輸入 Inception V3 預訓練模型，權重則是參考自 imagenet [10]，最後在模型的倒數第一層，輸出特徵為 $8 \times 8 \times 2048$ 的特徵向量，輸出完成後再利用 pickle 將輸出內容打包成數據集。

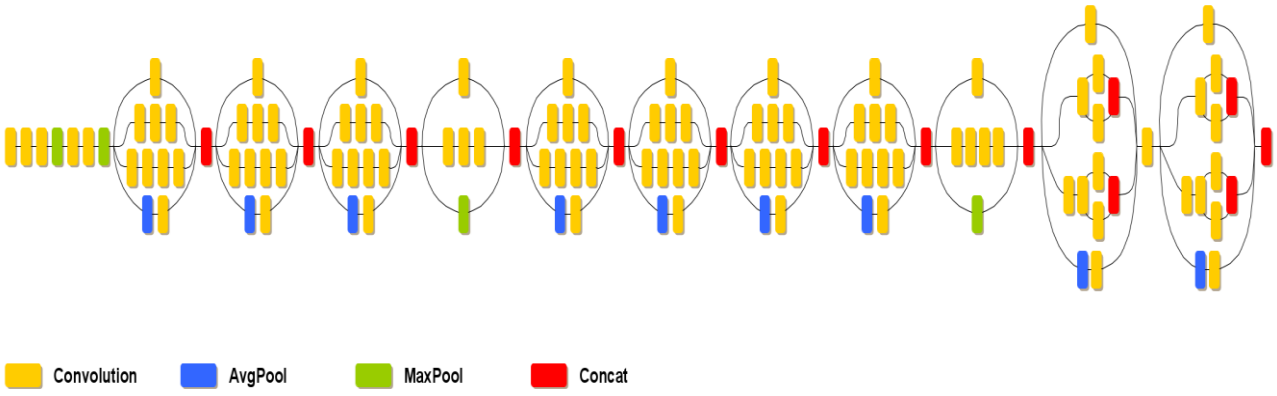


圖 3　編碼器 (Inception V3) 結構

（二）解碼器 (Decoder)

我們使用 GRU 作為解碼器，其中 Z 為更新閥，R 為選擇閥，h 則是將單元狀態和隱藏狀態合併的單位，另外加入了丟棄層，捨棄二分之一的訊息，使用的激勵函數是 tanh。為了與原文的模型比較訓練效果，我們也使用 LSTM 作為解碼器，其中，h 為單元狀態，state 是前一個時刻存入短期記憶單元的狀態，所有閥的激勵函數都是 sigmoid，詳細的實驗結果請參考討論的部分。

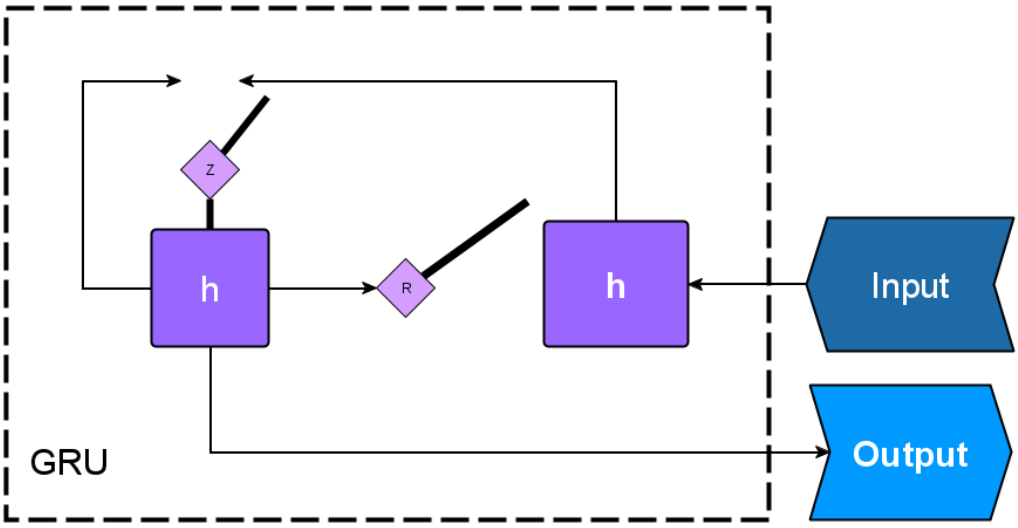


圖 4　GRU 結構示意圖

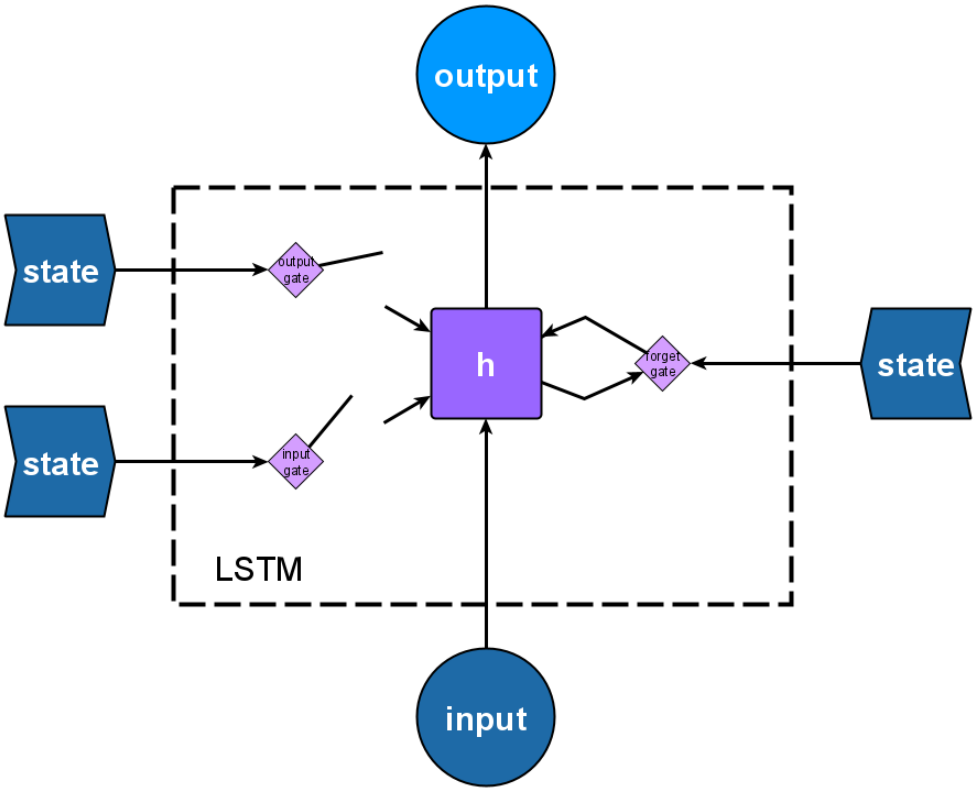


圖 5　LSTM 結構示意圖

（三）注意力模型

本研究中我們使用的注意力模型為 Bahdanau Attention [11]，**錯誤! 找不到參照來源。**為注意模型結構示意圖，其中 h 為輸入的序列，注意力的算法是將輸入的序列前後遍歷過一遍之後，再乘上注意力權重 (attention weight)，最後再加權得到一個語境特徵 (context vector)，再輸入編碼器中，使模型調整輸出值。

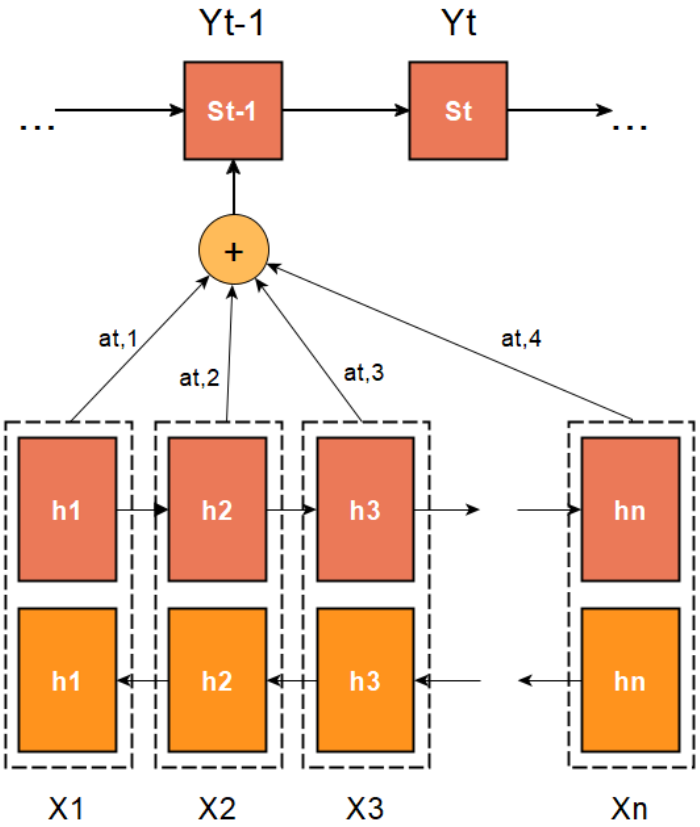


圖 6　注意力機制結構示意圖

（四）總結構

下圖為本研究使用模型的總結構示意圖，輸入 (Input) 為編碼器所得的特徵向量，將特徵向量輸入嵌入層 (embedding layer) 後，調整維度降低至 256，再輸入解碼器中進行序列的轉換，最後再通過兩層全連階層，調整輸出的形狀，配合輸入注意力機制中，來調整 GRU 的輸出序列。嵌入層的概念是將具有類似特徵的詞進行合

併，藉此讓訓練減少參數的調整及設定。全連接層分別接續將輸出詞彙調整和分類以供注意力模型進行序列的調整。

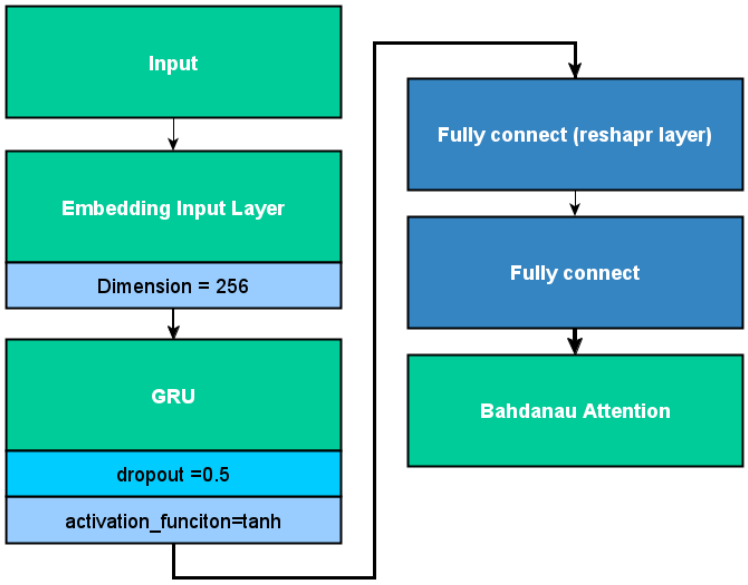


圖 7　模型總結構示意圖

（五）損失函數

我們將圖像輸入循環神經網路模型，計算參考句子和預測句子的內容損失，算法為先將 loss 輸入歸一化指數函數 (softmax) 之後再計算其交叉熵 (cross entropy)，公式如下，其中 P 為 softmax 的輸出，P_j 就是輸入 P 的第 j 個值，T 是總時程，y 為標籤樣本。

$$E = -\sum_{j=1}^{T} y_i \log P_j$$

算出來得到的質為每個 batch 中樣本的 loss，最後再計算其平均值得到整體 loss。

（六）機器翻譯自動評價

為了評估模型生成語句的好壞，我們使用了 BLEU 演算法 [12] 判斷參考句子和預測句子的相似程度。BLEU 使用了一種 N-gram 的匹配規則，通過它能夠算出比較譯文和參考譯文之間 n 組詞的相似的一個佔比，因此計算出的分數可分為 BLEU-1、BLEU-2、BLEU-3 等。有一些特殊情況無法通過 n-gram 反映句子的正確性，因此 BLEU 修正了算法得

$$p_n = \frac{\sum_{C\in\{Candidates\}}\sum_{n\text{-}gram\in C} Count_{clip}(n\text{-}gram)}{\sum_{C'\in\{Candidates\}}\sum_{n\text{-}gram'\in C'} Count(n\text{-}gram')}$$

其中分子表示取 n-gram 在翻譯譯文和參考譯文中出現的最小次數，分母表示取 n-gram 在翻譯譯文中出現次數。若預測句子只出現了參考句子的一部分，使用上述的式子計算的分數會為 1，但實際上的分數應該是比較低的，因此需要針對預測句子比參考句子短的情況，設置一個懲罰機制去控制，公式為

$$\text{BP} = \begin{cases} 1 & , c > r \\ e^{(1-r/c)} & , c \le r \end{cases}$$

綜上所述，得 BLEU 的公式為

$$\text{BLEU} = \text{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right)$$

六、訓練神經網路

我們嘗試過許多超參數的組合，最後決定使用 Adam 優化器 (optimizer) [13]，批量大小為 64。我們使用的損失函數為交叉熵，圖 8 為損失 – 迭代次數關係圖 (loss – epochs plot)，圖中藍線為訓練 (train) 的結果，棕線為驗證 (validation) 的結果，我們可以從圖中看出當迭代次數到達 3 次時，訓練的損失可以降到大約 0.036，驗證的損失維持在 0.053 左右，因此我們訓練模型時的迭代次數並不多。

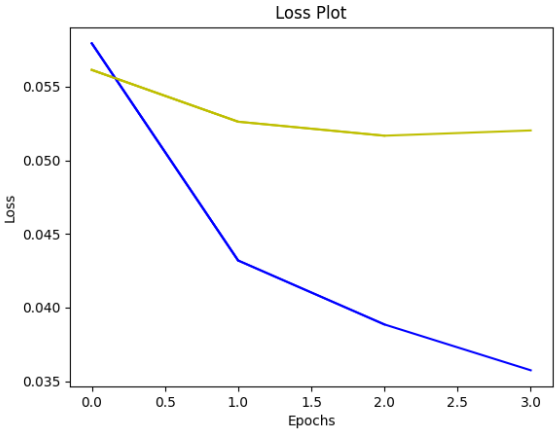


圖 8　損失 – 迭代次數關係圖

七、實驗結果

（一）實驗 1：測試 sigmoid 與 tanh 對於不使用注意力機制的主要序列轉換模型影響

我們先以 Show & Tell 的模型為基礎，在不使用注意力機制的條件下，測試使用 sigmoid 及 tanh 作為激勵函數對模型的影響，表 4 是我們的測試結果實例，左側是一個人在海上玩衝浪板的照片，我們發現使用 tanh 生成的結果很符合照片的內容，但是使用 sigmoid 生成的結果卻是「一個人在海灘上玩飛盤」，場景、物品和人的行為的都判斷錯誤；左側是一架飛機停在機場的照片，我們發現使用 tanh 生成的結果很符合照片的內容，但是使用 sigmoid 生成的結果卻是「一架飛機在空中飛行」，場景及飛機的狀態都判斷錯誤，因此我們認為使用 tanh 生成的效果較佳。表 5 是使用 tanh 及 sigmoid 的模型得到的 BLEU 分數，我們可以發現使用 tanh 的模型得到的 BLEU 分數較高，代表生成的圖像標註語句較佳，因此我們在之後的實驗中皆採用 tanh 作為激勵函數。

表 4　實驗 1 生成結果實例

	測試圖片 1 及生成結果	測試圖片 1 及生成結果
		
sigmoid	一個人在海灘上玩飛盤	一架飛機在空中飛行
tanh	一個人在海玩衝浪板	一架飛機停在機場

表 5　實驗 1 模型 BLEU 分數

激勵函數	BLEU-1	BLEU-2	BLEU-3	BLEU-4
sigmoid	0.142564	0.037396	0.013951	0.005404
tanh	0.145560	0.040309	0.014082	0.005811

（二）實驗 2：測試以 GRU 與 LSTM 對於使用注意力機制的主要序列轉換模型影響

我們以 Show, Attend & Tell 的模型為基礎，分別測試以 GRU 與 LSTM 作為主要序列轉換模型的影響，最後用 MSCOCO 中文資料集中的測試資料評價我們的模型，表 6 是我們的實驗結果。從表格中我們可以看到，如果訓練時的迭代次數相同，使用 GRU 的模型在各種評分方法中皆得到較高的分數。除此之外，我們由模型的訓練紀錄得到 GRU 每跑 1 次迭代大約花費 2000 秒，但使用 LSTM 每跑 1 次迭代大約花費 2700 秒，使用 GRU 在每一次迭代過程可以省下大約 10 分鐘的時間。

表 6　實驗 2 模型 BLEU 分數

模型(Show Attend & Tell**)**	**BLUE-1**	**BLUE-2**	**BLUE-3**	**BLUE-4**
使用 LSTM	0.232613	0.092213	0.043081	0.015224
使用 GRU	0.234074	0.096541	0.044400	0.015163

（三）實驗 3：使用 GRU 及注意力機制模型進行多次的迭代

延續實驗 2 的作法，我們在模型中使用 GRU 及注意力機制加入注意力機制進行多次的迭代，希望能改善圖像標註的生成結果。圖 9 及圖 10 **錯誤! 找不到參照來源。**分別是迭代 25 次及 50 次的生成結果實例，圖中的白色部分是注意力機制正在觀察的特徵，這些特徵會對應到圖片上方的中文詞句。請參看圖 9 左上角的 4 張小圖，模型能夠非常準確的看出圖片中有一架大型的噴氣式飛機；在上方第 5 張小圖以及左下方的 2 張小圖中，模型能夠看出飛機是在水面上。請參看圖 10，模型可以成功地辨認並生成「有一個棒球運動員在球場上打棒球」如此通順的中文語句，即使圖片中全沒有出現棒球，模型也能夠從圖片中其它的特徵生成「打棒球」這個動作，我們認為這是模型最成功之處。

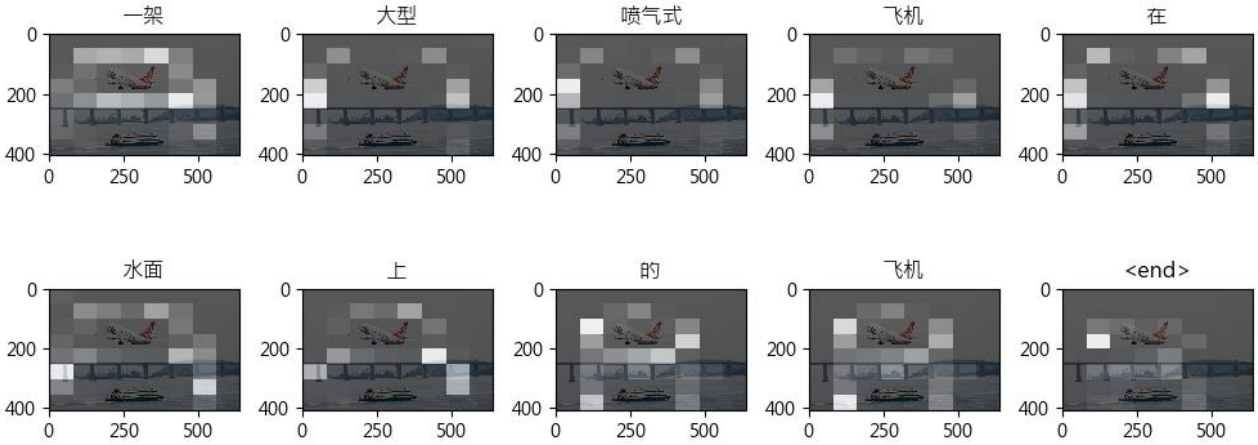


圖 9　使用 GRU 及注意力機制模型進行 25 次迭代的生成結果

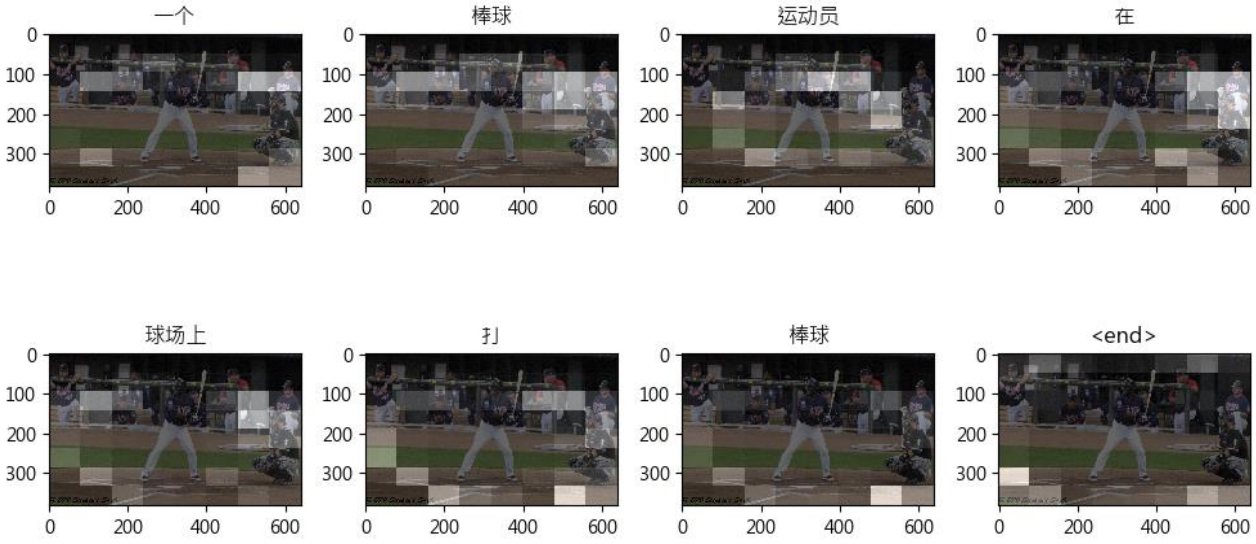


圖 10　使用 GRU 及注意力機制模型進行 50 次迭代的生成結果

伍、討論

一、討論

（一）實驗 1：測試 sigmoid 與 tanh 對於不使用注意力機制的主要序列轉換模型影響

在實驗 1 中我們發現，採用 tanh 作為激勵函數的模型可以產生較符合圖片內容的中文標註語句，而且也得到較高的 BLEU 分數較高，因此我們認為 tanh 比較適用於我們的模型。

（二）實驗 2：測試以 GRU 與 LSTM 作為主要序列轉換模型的影響

在實驗 2 中我們發現，模型中採用 GRU 的圖像標註表現較佳，而且訓練時需要花費的時間較短。我們推測這是因為 GRU 需要調整的參數約為 LSTM 的 1/20，訓練模型時需要的時間較短，而且訓練的效果也會較佳。

（三）實驗 3：使用 GRU 及注意力機制模型進行多次的迭代

在實驗 3 中我們發現，加上注意力機制能改善圖像標註的生成結果，模型可以成功地辨認出圖片中的特徵，並且生成通順的中文語句。

二、圖形化使用者介面

為了讓使用者可以很方便地使用我們的模型，我們嘗試製作圖形化使用者介面，使用只要用滑鼠左鍵按一下「點擊開啟圖片」再按下「生成句子」，我們的模型就會依據使用者選選的圖片生成適當的中文圖片標註。最後我們還加入了文字轉語音的功能，讓視障者也能夠使用我們的模型。

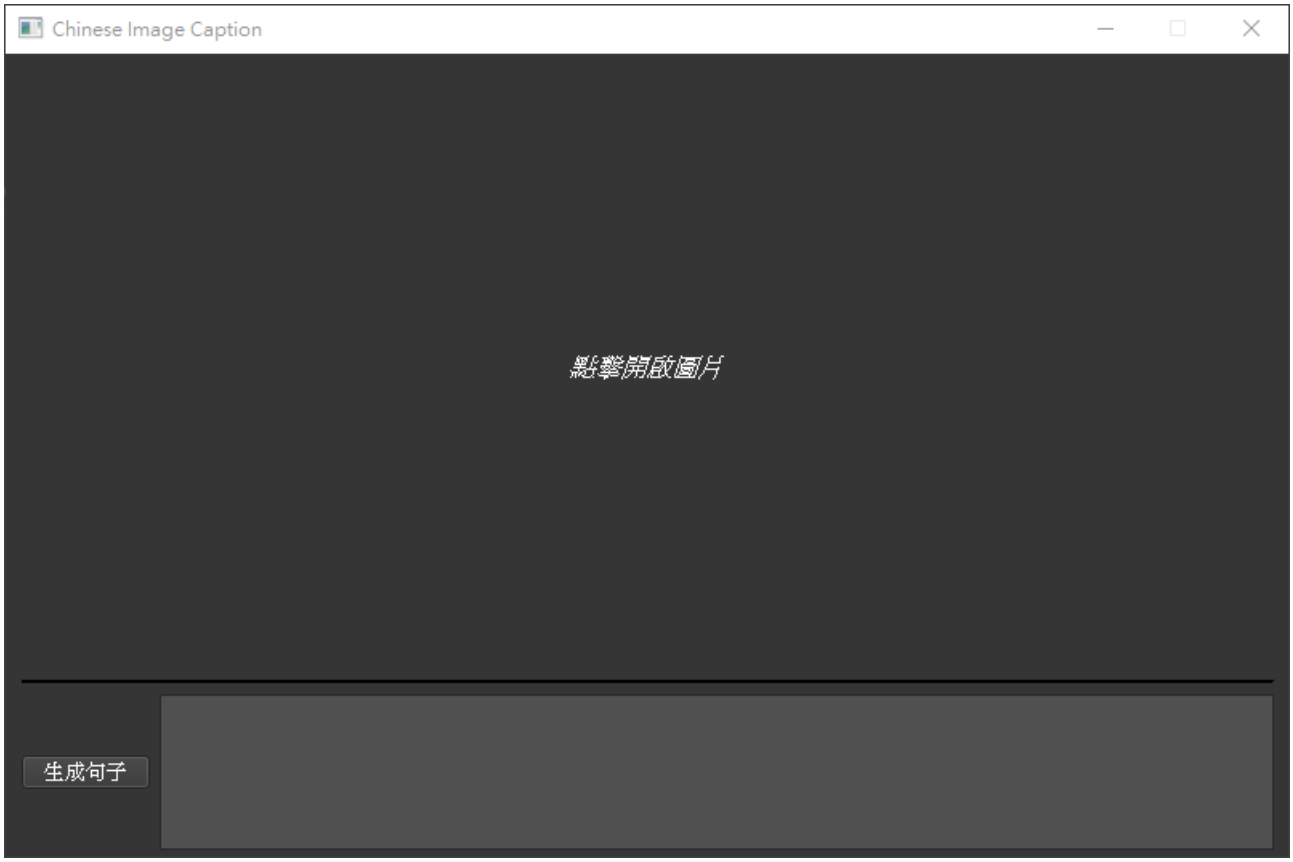


圖 11　圖形化使用者介面

陸、結論

一、對於不使用注意力機制的主要序列轉換模型，使用 tanh 作為激勵函數的效果較佳。

二、對於使用注意力機制的模型，使用 GRU 作為主要序列轉換模型生成中文圖像標註的效果較佳，而且訓練時需要花費的時間較短。

三、加上注意力機制有助於改善模型生成中文圖像標註的語句，使句子較為通順，但有一些需要改進之處，例如模型仍然不太了解某些中文詞句對應的圖片特徵。

四、未來展望

（一）配合召回機制 (Recall Mechanism) [14]

召回機制為 2020 年 1 月被提出的一種圖像標註流程，基於 Show, attend & tell 的模型修改而成，此網路能夠找出更多關於圖片的特徵，並能合理的使用觀察到的細微特徵。由於目前沒有任何關於此論文對其他結構的比較，還無從得知詳細的參數。

（二）影像標註 (Video Captioning) [15]

為了讓模型的應用範圍更加廣泛，我們希望能加入影像標註功能，但礙於硬體設備以及資料集的需求，我們需要更長時間的研究才能達成。目前已經有國外的學者製作出英文的影像標註模型，如果我們能夠取得需要的設備以及資料集，我們希望能夠進一步修改我們的模型，將模型應用到影像標註任務上。

柒、參考資料及其他

[1] Lin, Tsung-Yi, et al. "Microsoft coco: Common objects in context." European conference on computer vision. Springer, Cham, 2014.

[2] Li, Xirong, et al. "COCO-CN for Cross-Lingual Image Tagging, Captioning, and Retrieval." *IEEE Transactions on Multimedia* 21.9 (2019): 2347-2360.

[3] Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016). Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2818-2826).

[4] Simonyan, Karen, and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition." arXiv preprint arXiv:1409.1556 (2014).

[5] Gers, Felix A., Jürgen Schmidhuber, and Fred Cummins. "Learning to forget: Continual prediction with LSTM." (1999): 850-855.

[6] Chung, Junyoung, et al. "Empirical evaluation of gated recurrent neural networks on sequence modeling." *arXiv preprint arXiv:1412.3555* (2014).

[7] Vaswani, Ashish, et al. "Attention is all you need." *Advances in neural information processing systems*. 2017.

[8] Vinyals, Oriol, et al. "Show and tell: A neural image caption generator." Proceedings of the IEEE conference on computer vision and pattern recognition. 2015.

[9] Xu, Kelvin, et al. "Show, attend and tell: Neural image caption generation with visual attention." International conference on machine learning. 2015.

[10] Russakovsky, Olga, et al. "Imagenet large scale visual recognition challenge." International journal of computer vision 115.3 (2015): 211-252.

[11] Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio. "Neural machine translation by jointly learning to align and translate." *arXiv preprint arXiv:1409.0473* (2014)..

[12] Callison-Burch, Chris, Miles Osborne, and Philipp Koehn. "Re-evaluation the role of bleu in machine translation research." 11th Conference of the European Chapter of the Association for Computational Linguistics. 2006.

[13] Kingma, Diederik P., and Jimmy Ba. "Adam: A method for stochastic optimization." *arXiv preprint arXiv:1412.6980* (2014).

[14] Wang, Li, et al. "Show, Recall, and Tell: Image Captioning with Recall Mechanism." *arXiv preprint arXiv:2001.05876* (2020).

[15] S. Venugopalan, M. Rohrbach, R. Mooney, T. Darrell, and K. Saenko. Sequence to sequence video to text. In Proc. ICCV, 2015

【評語】052508

本作品採用機器學習技術希望透過訓練模型後能自動偵測一個圖像中的物體且產生有意義的標註和說明（caption），而這些標註和說明必須是自動生成且合理的自然語言。研究主題清楚且具實用性。但此作品只執行一些現有的機器學習程式套件來獲得自動生成輸出，並未對所使用的機器學習方法有深入的探討且提出進一步改良的方法，因此在科學探討精神上較為欠缺。建議設計更多的實驗，進行更多不同標註方式結果的比較。另亦可與英文標註結果做一比較，以了解兩者技術上的差異。

作品海報

摘要

目前市面上已經有不少產品應用到自然語言處理 (natural language processing, NLP) 的技術，例如聲音或是圖像標註，但是關於圖像標註的研究大部分是針對英文，因此我們想要進行中文的圖像標註研究。在本研究中，我們嘗試建構神經網路，使用 Microsoft COCO: Common Objects in Context [1] 數據集訓練模型，並試著調整模型以達到較好的圖像標註效果。為了能達到從序列到序列 (seq2seq) 的效果，我們的網路使用了編碼器 - 解碼器 (encoder-decoder) 的結構，編碼器的部分使用 CNN，解碼器的部分使用 RNN，這樣能有效地傳達序列並有較多種神經網路組合方式。

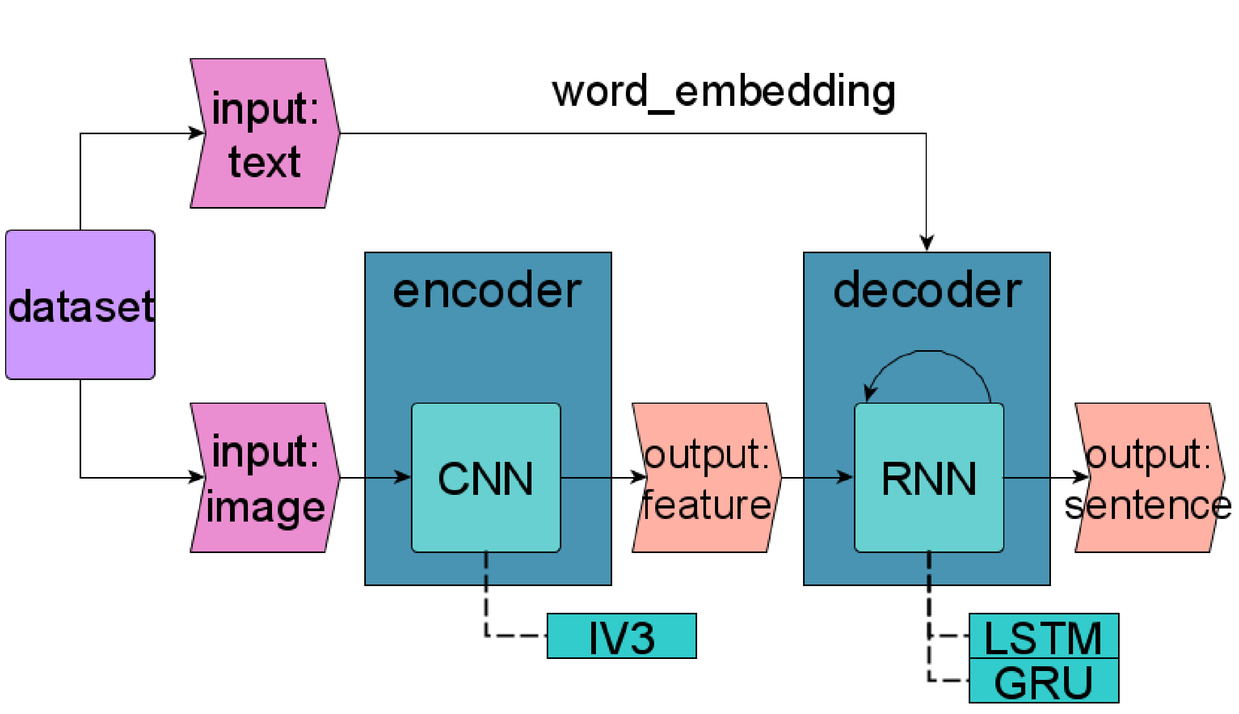
研究動機

將人工智慧應用於自然語言處理是近幾年熱門的研究領域，其中自動圖像標註 (image captioning) 這個領域已經發展相當長的時間，網路上有許多關於這個領域的論文及模型，然而這些模型絕大部分是針對英文設計的，鮮少有專為中文設計的模型，因此我們決定嘗試建構並訓練能自動為圖像加上中文敘述的模型。

研究目的

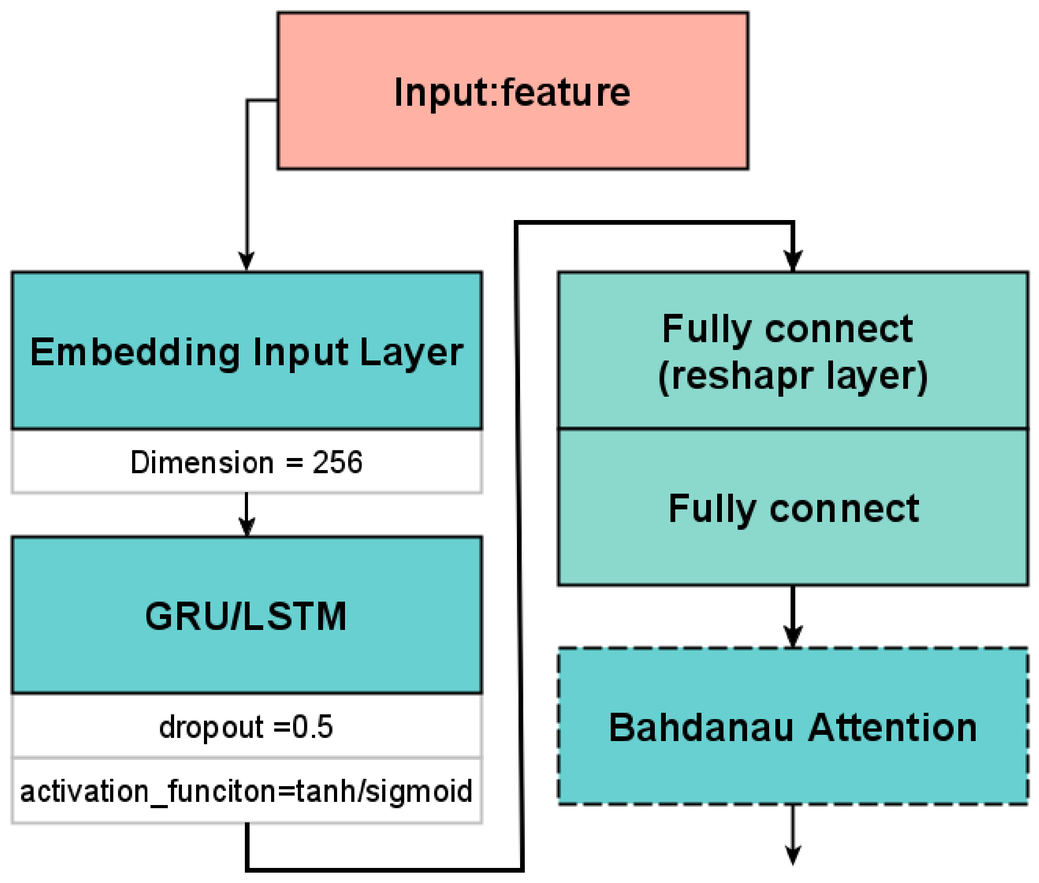
- 探討神經網路應用於圖像標註的效果
- 探討神經網路對中文自然語言處理的效果
- 探討神經網路模型的更動對訓練結果的影響
- 製作圖形化使用者工具

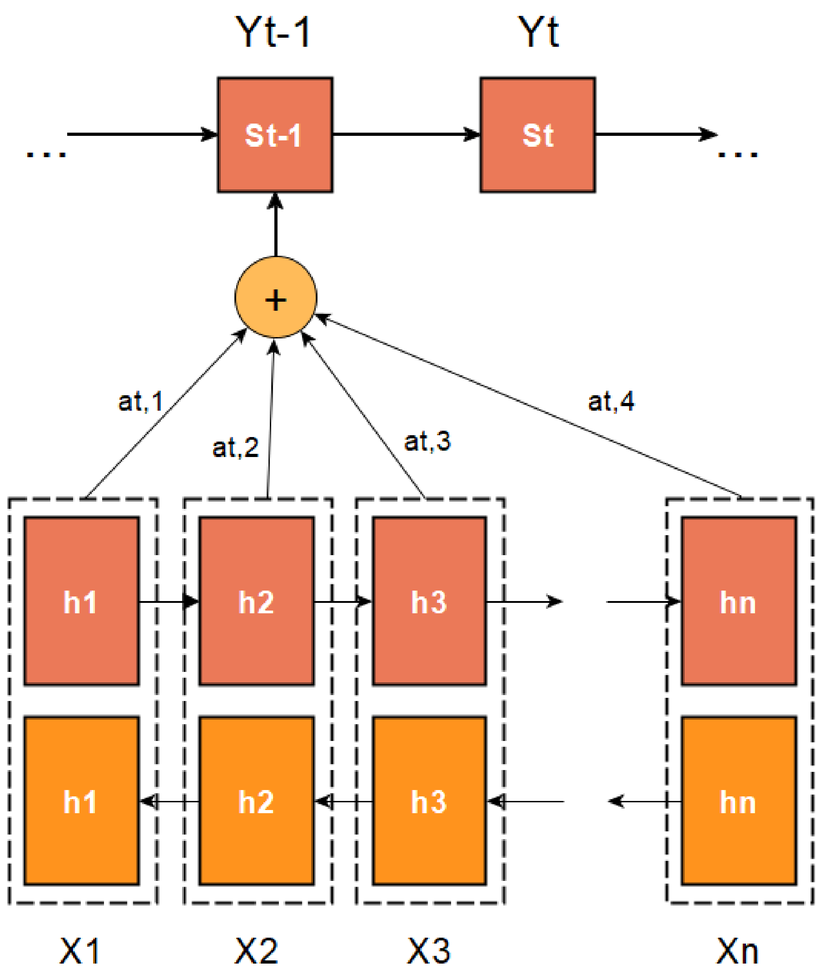
神經網路結構及過程



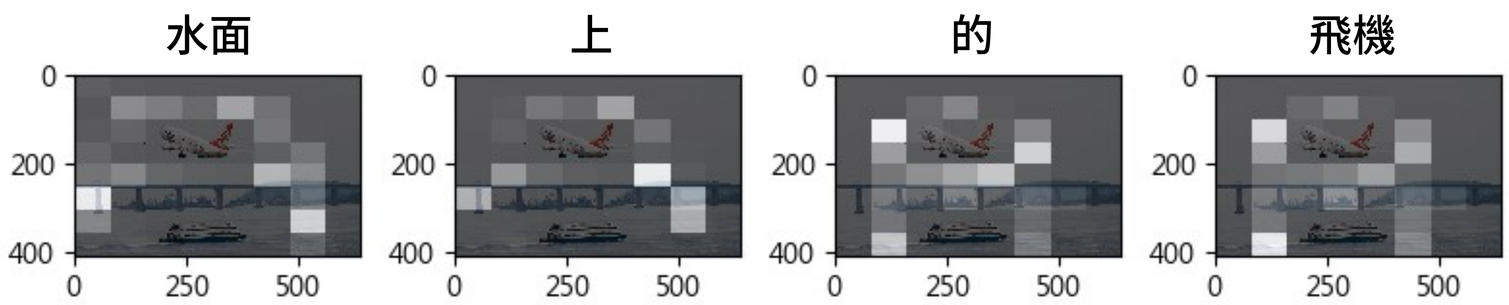
左圖為神經網路結構，先將資料輸入編碼器(encoder)取得特徵向量後，輸入解碼器(decoder)進行序列的變換，最後得到輸出的標註。本研究主要是針對解碼器的部分進行實驗，試著找出圖像標註效果較佳、訓練所需時間較短的組合。

右圖為解碼器的總結構，將特徵向量輸入進行降維至 256 後通過 GRU 生成序列，兩層全連階層調整大小，最後再通過注意力模型將關注重點分布擴張於前後文(上下序列)，重複至資料集結束即完成解碼。





左圖為注意力模型的結構圖。主要是為了加強生程序列對自然語言文義的合理性、連貫性、流暢度而加入。本實驗使用的注意力模型為 Bahdanau attention [2]，相較於其他注意力模型，此模型對中文的處理能有較佳的表現。下圖中白色的部分就是模型當前序列時注意的部位，而圖片上的文字則是部位對應到的單詞。



實驗

在我們的實驗中，我們主要針對解碼器進行參數和模型的變換，探討其對中文圖像標註的影響，判斷影響的方式為 BLEU 評價 [3]。

(1) 激勵函數(activation function)

本實驗參考的模型為 Show and Tell [4]，測試解碼器的激勵函數對生成結果的影響，實驗中採用了 sigmoid 及 tanh，下圖為實驗結果。

sigmoid:
一個人在海灘上玩飛盤
tanh:
一個人在海玩衝浪板

sigmoid:
一架飛機在空中飛行
tanh:
一架飛機停在機場

sigmoid	BLEU-1=0.142564	BLEU-2=0.037396	BLEU-3=0.013951	BLEU-4=0.005404
tanh	BLEU-1=0.145560	BLEU-2=0.040309	BLEU-3=0.014082	BLEU-4=0.005811

(2)解碼器模型結構

本實驗參考的模型為 Show, Attend and Tell [5]，測試解碼器主要處理序列的結構對生成結果的影響，實驗中採用了 LSTM 和 GRU，雖然兩種模型的 BLEU 分數相當接近，但是採用 GRU 所需的訓練時間較短，每一次迭代過程可以節省大約 10 分鐘，下圖為實驗結果。

LSTM:
兩隻狗在玩耍
GRU:
兩隻狗在草地上玩耍

LSTM:
兩个人坐在沙發上玩遊戲
GRU:
一群人坐在沙發上玩電子遊戲

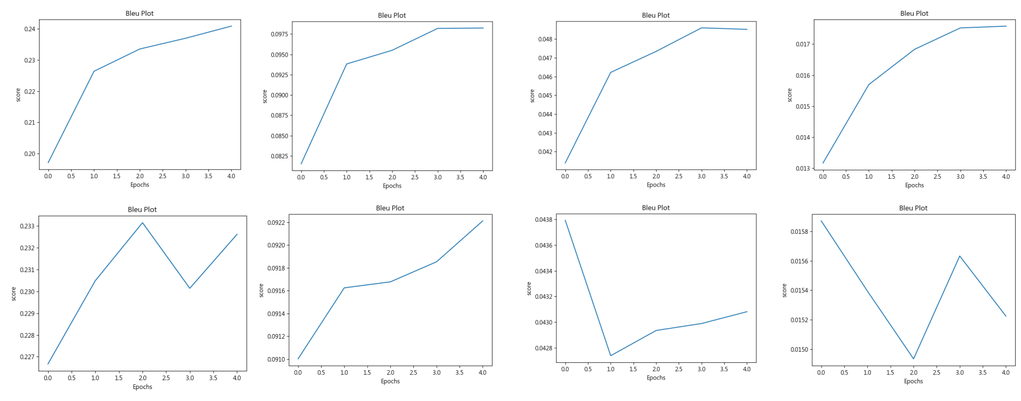
LSTM	BLEU-1=0.232613	BLEU-2=0.092213	BLEU-3=0.043081	BLEU-4=0.015224
GRU	BLEU-1=0.240876	BLEU-2=0.098238	BLEU-3=0.048521	BLEU-4=0.017578

研究結果

(1)參數和結構對中文圖像標註的影響

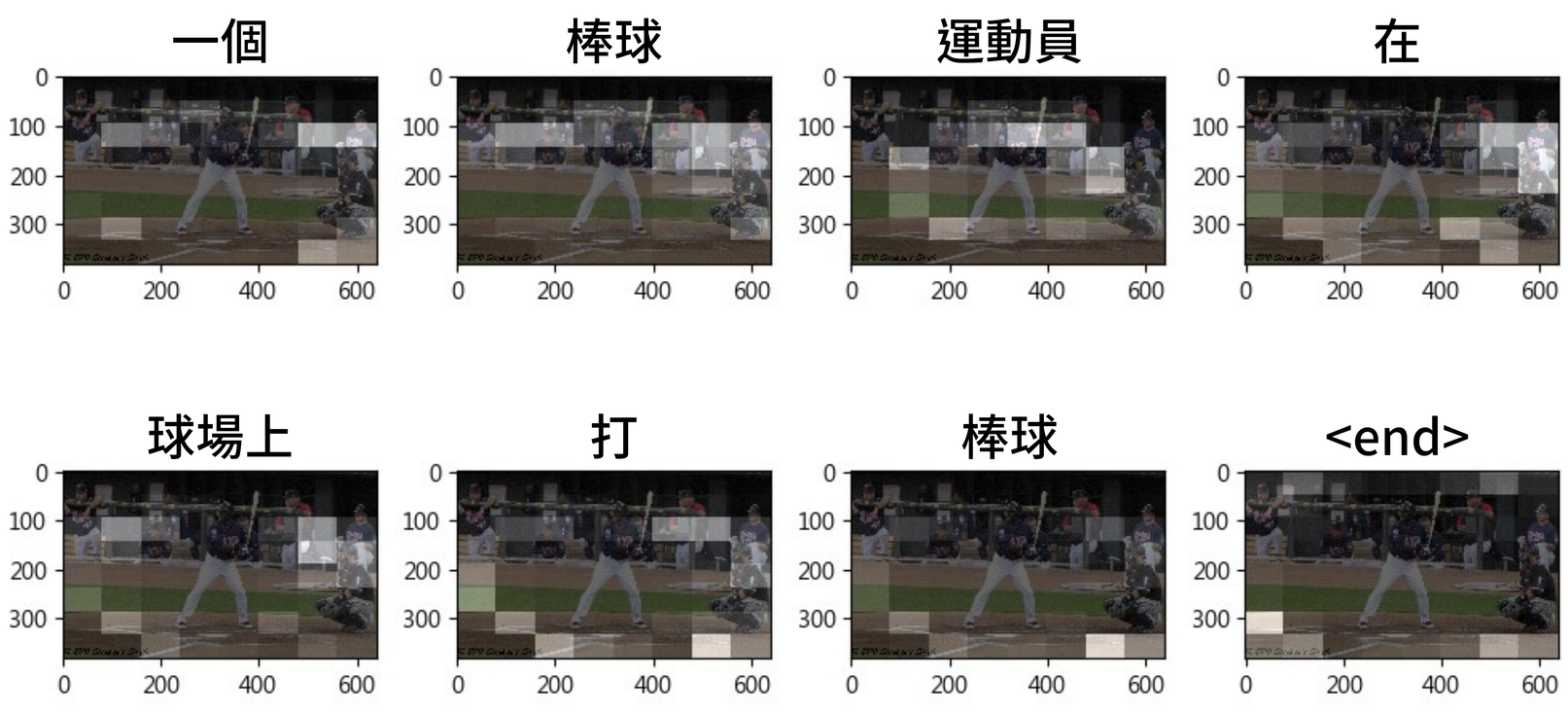
透過實驗一、實驗二的 BLEU 評價分數可以看出，在中文圖像標註中不同 GRU 的配合對輸出的結果有著顯著的差異，其中以 tanh 作為激勵函數、 LSTM GRU 作為解碼結構效果較佳。

BLEU –1 BLEU –2 BLEU –3 BLEU –4



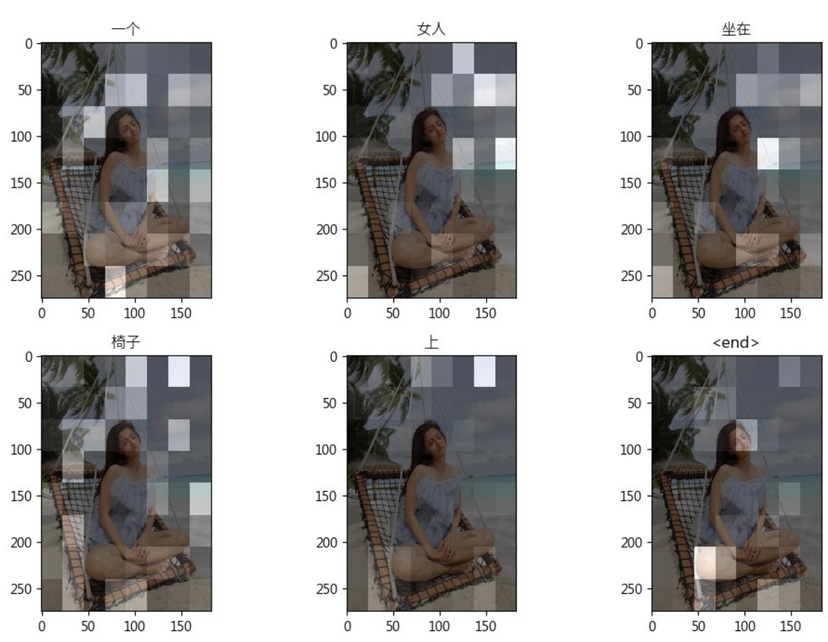
(2)製作模型

在實驗中，我們使用適合的參數和結構，配合訓練主要的解碼器，並加入能使分數更高的注意力結構，最後訓練 50 epoch 得到圖像標註模型，下圖為模型標註結果，以注意力分布圖表示。



(3)圖形化使用者介面

最後我們製作圖形化使用者介面，讓使用者能方便地使用我們的模型進行中文圖像標註，並且同時能在 Windows 及 Linux 作業系統上運行。



結論及展望

不同激勵函數及解碼模型對中文圖像描述有顯著影響，在實驗中針對自然語言處理的評價分數，可看出明顯的差異，但模型如果遇到沒見過或意義不明，無法直接由圖片上的局部特徵判斷出整體的物件的話，就容易選擇出錯誤的單詞，希望未來能增加人工翻譯的資料集來改善此問題。

參考資料

1. Lin, Tsung-Yi, et al. "Microsoft coco: Common objects in context." European conference on computer vision. Springer, Cham, 2014.
2. Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio. "Neural machine translation by jointly learning to align and translate." *arXiv preprint arXiv:1409.0473* (2014).
3. Callison-Burch, Chris, Miles Osborne, and Philipp Koehn. "Re-evaluation the role of bleu in machine translation research." *11th Conference of the European Chapter of the Association for Computational Linguistics*. 2006.
4. Vinyals, Oriol, et al. "Show and tell: A neural image caption generator." Proceedings of the IEEE conference on computer vision and pattern recognition. 2015.
5. Xu, Kelvin, et al. "Show, attend and tell: Neural image caption generation with visual attention." International conference on machine learning. 2015.