資料科學與機器學習 期末報告

Image Caption generation

106403054 洪睿甫、106403528 王佩晨

國立中央大學 資訊管理學系

Abstract

Image Caption generation 在電腦視覺領域中是一個重大議題之一，主要任務就如同字面上意義，根據圖片來產生相對應的標題，甚至有更進階的應用，像是給定一張圖片和這張圖片的問題，讓機器根據所給條件來產生相對地回答。而本篇研究主要聚焦在前者的標題產生的任務，主要想研究是否能透過增加其他輸入來解決原本圖片萃取特徵其高階語意不足，導致生成階段效果不佳的問題。

keywords：Image Caption、Visual Attention、自然語言處理

---

1. Introduction

Image Caption generation 是一個跨足了電腦視覺和自然語言處理的綜合問題，大多數傳統主流方法是以 encoder-decoder 為主要架構，另外近年來更是出現了使用 GAN 和強化學習的技術來提升生成的結果。在傳統 Encoder-Decoder 架構中通常多以 Convolutional Neural Network 當作 encoder、Recurrent Neural Network 當作 decoder。之後的研究大多致力於讓 decoder 有效地獲得片資訊來提升生成效果，大多數作法是加入不同的 attention 的運算或是對於 attention 的架構來做改動。而本篇研究雖然也是以 attention 的問題為切入點，但我們也參考了 What Value Do Explicit High Level Concepts Havein Vision to Language Problems?[1] 這篇論文所提出的「以圖片的屬性標籤來帶題圖片特徵」來做發想。因此我們的模型保留了 show, attend and tell[2]這篇的 attention 機制，並在 decoder 的 input 中加入經過 multi-label 後的分類結果，對第一個 input 和 multi-label 的結果做加總，使用 BLEU 來做最後的評分標準，藉此來觀察是否能藉由給定 decoder 這張圖片的主題來提升生成文字的正確性。而本文的結構如下:在第 2 節簡要介紹了我們參考的論文的做法以及問題，並在第 3 節介紹我們的做法，最後於第 4 節討論我們的實驗結果。

2. Related Work

在 Show and Tell 中，主要透過 CNN 萃取的圖片特徵後，再將特徵輸入到 decoder 中使用 RNN cell 進行解碼並轉換成圖片標題的生成，為了減短訓練時間在 CNN 部分大多採取了預訓練好的模型來加速整個訓練的過程，另外 RNN 的部分則是採用 GRU cell 來提升預測結果。

後來在 Show Attend and Tell 中出現了再加入 Attention 的方法。Attention 藉由將每一個 decoder 的 hidden_state 和圖片特徵來做 BahdanauAttention 的運算，得到一組特徵權重，並將特徵權重乘上圖片特徵得到 contextvector，這個 contextvector 就是為了讓 decoder cell 在生成每一個單字時，能夠針對圖片中重要特徵做局部關注的文字生成。

不過即使專注在局部特徵來做生成，decoder 還是無法僅由圖片特徵來達到有意義的運算，導致關注部分所生成出來的文字和真實結果有誤差。因此在 What Value Do Explicit High Level Concepts Havein Vision to Language Problems?[1] 這篇論文針對這個問題，也就是輸入到 decoder 的圖片特徵，提出了缺乏了高階的語意特徵的說法。因此，該論文變向使用圖片的屬性預測(attribute prediction)來取代原來的圖片特徵來當作輸入，而在這邊的屬性預測(attribute prediction)其實就是 multi-label 的分類任務所得到的所有 attribute 的機率分布。因此本篇研究主要根據 What Value Do Explicit High Level Concepts Havein Vision to Language Problems?這篇論文所提出的「以圖片的屬性標籤來代替圖片特徵」來做發想，在不捨棄原有圖片特徵的同時也加入屬性預測的結果。

3. Our Process(Model)

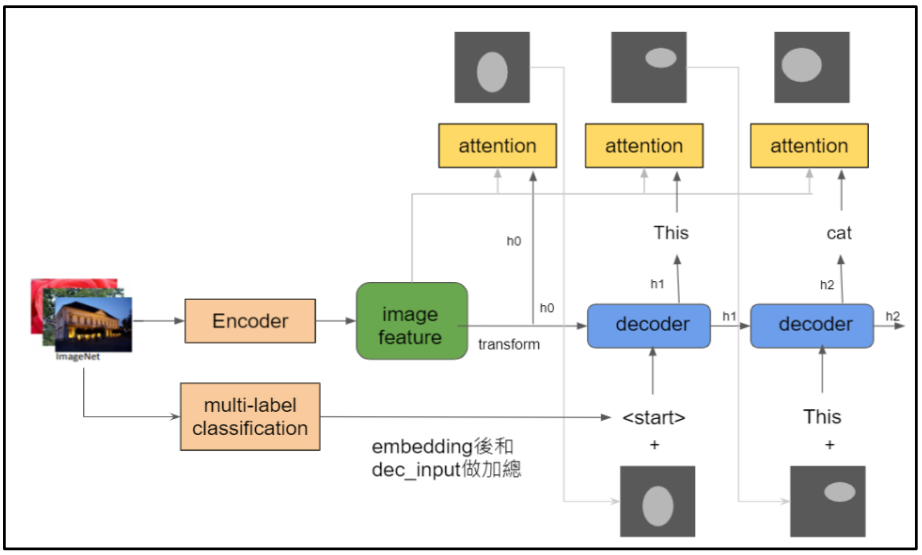


圖 1. 模型架構圖

3.1 Encoder

3.1.1 Image Preprocessing

首先我們對圖片做 resize，使他符合 VGG19 的大小(224,224,3)，再使用 keras 的 preprocess_input 把圖片正規化到-1 到 1 之間。

3.1.2 Image Extraction Model

為了萃取出圖片的特徵，擷取 VGG19 至最後一層卷積層的輸出(不要做圖片的分類)作為一個新的模型，輸出的 shape=(7,7,512)。

3.2 Multi-label Classification

3.2.1 目的

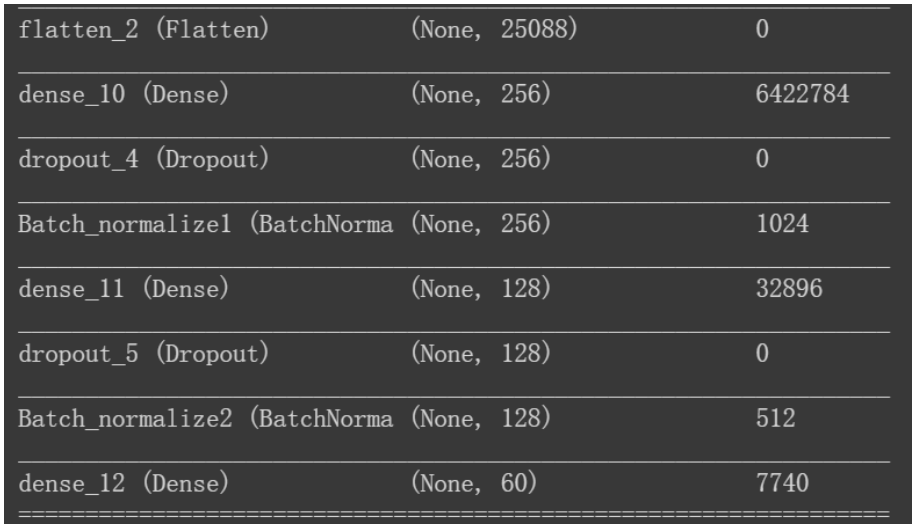
訓練一個 multi-label 的分類器，將分類結果也加入 decoder 的輸入，讓生成文字的過程中除了接收圖片特徵之外，希望能藉由加入標籤資訊來給予 decoder 額外資訊。

3.2.2 資料集

將做好 onehot-encoding 的資料集做切割，訓練集為 15000 筆、驗證集為 2000 筆、測試集為 5000 筆。原資料集中，category 擁有 80 個類別，為了減少記憶體的負擔並增加訓練速度，我們從 81 個類別中(80+none)選取前 60 個做為分類任務的 category 標籤。

3.2.3 預訓練模型- VGG19

VGG19 是一個 19 層(16 個卷積層及 3 個全連接層)的架構模型，為了讓模型更加符合我們的資料集，採用 tansfer learning 的方式去做預訓練模型的微調。由於 VGG19 的訓練資料集 imagenet 和此實驗的資料類型相似，因此我們只調整最後五層的模型架構參數，將其 trainable 設為 True。最後在 fully-connected layer 中，使用 Flatten、BatchNormalization 和 Dense 將 output 的類別改為 60 個。



flatten_2 (Flatten)	(None, 25088)	0
dense_10 (Dense)	(None, 256)	6422784
dropout_4 (Dropout)	(None, 256)	0
Batch_normalize1 (BatchNorma	(None, 256)	1024
dense_11 (Dense)	(None, 128)	32896
dropout_5 (Dropout)	(None, 128)	0
Batch_normalize2 (BatchNorma	(None, 128)	512
dense_12 (Dense)	(None, 60)	7740

圖 2. multi-label 模型架構

3.2.4 預測結果

將測試資料丟入訓練好的模型進行預測後，可以得到一個陣列，其 shape=(5000,60)，5000 為測試資料集大小，60 代表的是每一個 label 的機率分布。我們將在這 60 個機率分布中取最大值的 label id，並把這個 label id 做 embedding，得到的結果將用來和第一個 decoder 的第一個 input 做加總

```
efficient_predict[0]

array([0.00942305, 0.02592477, 0.01497281, 0.02122069, 0.01703197,
       0.01685871, 0.01830577, 0.01736073, 0.01884668, 0.01699478,
       0.018235  , 0.01633714, 0.01618532, 0.01443603, 0.01844216,
       0.01868876, 0.01718607, 0.01837803, 0.01610404, 0.01429502,
       0.0159261 , 0.01394783, 0.01592682, 0.01759932, 0.0145539 ,
       0.01901594, 0.01723483, 0.01939707, 0.01560305, 0.01842495,
       0.01582429, 0.01768264, 0.01789871, 0.01735345, 0.01439054,
       0.01326655, 0.0155248 , 0.01674218, 0.01664959, 0.01617867,
       0.01992441, 0.01573765, 0.01917791, 0.01685561, 0.01914045,
       0.01221364, 0.01993634, 0.01552609, 0.01498338, 0.01556449,
       0.01516863, 0.0143073 , 0.01657618, 0.01156301, 0.01731215,
       0.01490488, 0.01533054, 0.01888832, 0.01567004, 0.01685019],
      dtype=float32)
```

圖 3. 每張圖片輸出的機率分布格式

參數	參數值	備註
input_shape	(224,224,3)	Vgg 預設大小
batch	64	
epoch	25	大於 25 後，發生 overfitting

activation function	softmax	60 個 label 的機率分布
optimizer	SGD(lr=2e-4)	可以跳出局部最小值，為搭配此資料集效果最佳的優化器
loss function	categorical crossentropy	資料為 onehot-encoding 的多標籤分類，若使用 binary crossentropy 會使預測結果皆為 0

表格 1. 參數設定

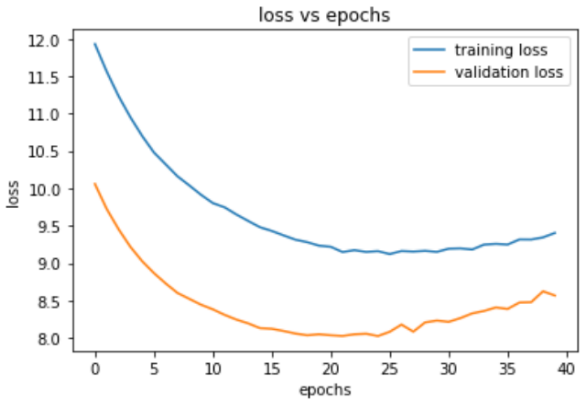


圖 4. 訓練後的 Loss 值

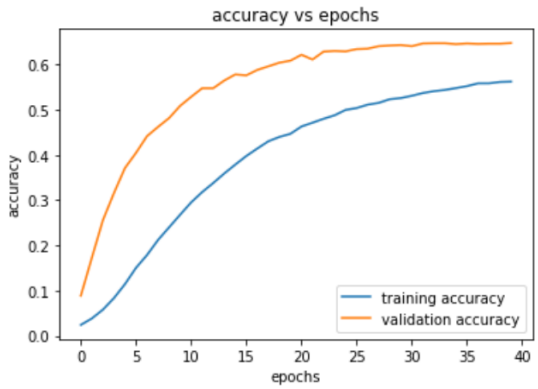


圖 5. 訓練後的 accuracy 值

3.3 Attention

Attention 的輸入主要有二，一是經過 CNN 萃取後的圖片特徵，二是 decoder 的 output， hidden_state。詳細算法是先將圖片特徵和 hidden_state 做線性轉換後相加，並經過一層 softmax 得到 attention weight，而這個 attention weight 就是代表了圖片特徵的每個區域對於 hidden_state 影響的程度大小，因此最後我們將 attention weight 乘上圖片特徵所得到的 context vector，就是讓每一個 decoder 在圖片特徵中獲得局部專注的效果。

3.4 Decoder

使用 GRU 作為 decoder 做文字的生成，此實驗選擇使用 GRU 的原因是它比 LSTM 訓練效率更高且擁有較高的準確率。在訓練階段為了加速訓練進行，因此我們以 teacher forcing 的方式，在 decoder input 是使用正確答案當當作輸入，測試階段則是以每一個 decoder 的預測結果來當下一個 decoder 的 input，而 decoder 的輸入主要有二，第一是在當前的 decoder input 值，需要注意的是我們的第一個 decoder input 會先和 multi-label 分類結果的最大值做相加，得到的

結果會和上一階段的 attention 階段計算出來的 context vector 做 concate 當作輸入(後面的 decoder input 就正常做 concate 不再和 multi-label 做相加)，第二是前一階段 decoder 的 ouput hidden state(本篇研究的初始 hidden state h0 是圖片特徵的線性轉換，我們是使用 tf.keras.layers.Dense 來做轉換)。

4. Experiment Result

4.1 Dataset

本篇研究使用了 Microsoft COCO 當作資料集，由於他擁有大量被標註的圖片如：Detection, Segmentation, Keypoints 等等，因此在電腦視覺的領域常常被拿來當作資料集使用。而我們使用的是 2014 年的版本，裡面總共有 80 個標註類別和 8 萬多張標註過圖片，每個圖片都有 5 個對應的標題， 其中 4800 張圖片當作 traningset，1200 張當作 valadation set。另外在圖片標前及對應標題的取得需要另外下載 annotaion 的 json 檔，並透過 COCO 的 API 來取得

4.1.2 資料前處理:

1. 取得每個圖片在 coco dataset 中的 category。
2. 使用 coco API 來取得圖片的類別標籤，若該圖片未出現在分類中則把它分類成'none'。
3. 對每張圖片的 category 標籤做 onehot encoding。

4.2 Text Preprocessing - tokenize the caption

1. 在每個 caption 的前後分別加上<start>和<end>，在對每一個單字進行編碼，設定一個 vocabulary size(設 5000)，將沒有出現在這 5000 個字中的單字設為<UNK>。
2. 製作 word-to-index 和 index-to-word 的 map 以方便之後對照。
3. 取得所有 caption 中的最大長度，並把不足長度的做 padding。

4.3 Evaluation

本篇研究使用 BLEU 來當作生成文字的評估指標，可用來比較生成出來的標題與其他多個標註標題並進行評分。但 BLEU 主要是比較句子之間的對應關係，其缺點是沒辦法考慮語意上的正確性沒有考慮到同義詞或是相似的表達方式。

4.4 Result

	B-1	B-2	B-3	B-4
Our Proposal 10 Epoch	0.77	0.62	0.489	0.30
Our Proposal 20 Epoch	0.78	0.65	0.52	0.35
Baseline - Show attend and tell(VGG)	0.91	0.85	0.76	0.58
CNN+LSTM with multi-label	0.74	0.52	0.46	0.31

表格 2. BLUE1~4 在不同模型上的數值

Image	COCO_train2014_000000180285.jpg (圖 6)
Real	1. A couple of women with some stuffed animals. 2. Two women smile for the camera while posing with some stuffed animals. 3. Two women sit and pose with stuffed animals.
Our Proposal	visors happy after a woman eating and one piece of food at a table with a room and pretending to eat on the table with sliced bananas like eyes
Show attend and tell	the young female with smiling

表格 3. 對同一張圖片的生成結果

圖 6.

如表 3 所示我們的 BLUE 分數是低於我們的 baseline，但是卻比單純只加入 multi-label 機率分佈的結果還要好，因此我們實際來看一下針對圖片生成的文字結果(表三、圖 6)，我們可以看到其實在實際上文字生成的時候，雖然重點文字有抓到，但是文法卻很不通順，

4.5 Analysis

根據這次研究的實驗結果，從上表的 BLEU 值可以得知我們的模型架構比 baseline 的表現還要差。經過分析後，我們認為結果較差的原因主要如下。

4.5.1 Multi-label 分類器準確率不佳

我們 Multi-label 分類器的準確率只有大約 65%，這樣的表現導致容易有分類錯誤的情況產生，而我們的主要目標就是在輸入時加入分類標籤，若是分類出錯誤的 label，更容易使最後 decoder 生成錯誤的文字，造成反效果的發生。

4.5.2 label 相似度太高

我們觀察到若只將 multi-label 的分類結果取最高的值作為 decoder 的 input 時，會導致 label 的相似度太高，而失去了加入主題概念的意義。舉例來說，只要圖片中出現「人」，不論這張圖片是一個人再丟飛盤，或是一群人在吃飯，經過 multi-label 分類器後都會得到 person 的 label，抹去了圖片間的差異性的元素。因此，我們認為應該要取前幾高的 label 作為 decoder 的輸入會有較佳的表現。

4.5.3 訓練的 epoch 數不足

從(圖 7)模型的 loss 值做分析，可以看出經過 20 個 epoch 後，Loss 值仍然呈現穩定下降的趨勢。因此，我們認為增加 epoch 數可以改善模型的表現，並增加句子生成的準確度，但由於受到訓練時間的限制以及與其他模型比較的公平性，此實驗停止在 20 個 epoch。

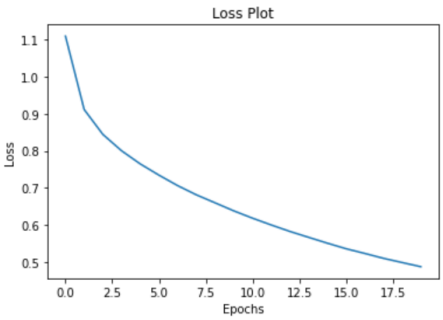


圖 7

4.5.4 訓練集和驗證集資料過少

此次實驗因為受到硬體的限制，我們的資料集數量相較於其他篇論文的資料數量少很多(此實驗採訓練集 4800、驗證集 1200)，因此限制了模型的訓練結果。

5. Conclusion and Future Work

總結以上的實驗結果，雖然我們最後提出的模型架構在 BLUE 分數的成效比 baseline model 還要差，但由於其他因素(ex:訓練 epoch 次數不足，標籤分類模型準確率不高等...)，導致我們的模型在許多方面仍有缺陷，因此我們無法百分之百肯定同時將類別標籤和圖片特徵當作輸入，對於文字生成是沒有幫助的。

所以，為了在更進一步的去驗證這個模型的可行性，我們在未來也考慮加入分類出來的標籤選前 N 高的當輸入，希望藉此來更多元的表示圖片的語意。此外因為時間緣故，沒有加入其他評分標準(例如:ROUGH, CIDEr)，因此沒有辦法判斷語意有無通順的部分，或是和原句的相關性比較，未來也可以加入評分標準來驗證語句的可讀性。

6. References / Citations

[1] Q. Wu, C. Shen, L. Liu, A. Dick, and A. van den Hengel, “What value do explicit high level concepts have in vision to language problems?,” *ArXiv150601144 Cs*, Apr. 2016, Accessed: Jun. 18, 2021. [Online]. Available: http://arxiv.org/abs/1506.01144

[2] K. Xu *et al.*, “Show, Attend and Tell: Neural Image Caption Generation with Visual Attention,” *ArXiv150203044 Cs*, Apr. 2016, Accessed: May 22, 2021. [Online]. Available: http://arxiv.org/abs/1502.03044