

國立臺灣大學電機資訊學院資訊工程學研究所



碩士論文

Department of Computer Science and Information Engineering

College of Electrical Engineering and Computer Science

National Taiwan University

Master Thesis

圖文生活日誌之圖片回憶研究

Analysis of Image Recall on Image-Text Intertwined Lifelog

祝子軒

Tzu-Hsuan Chu

指導教授：陳信希 博士

Advisor: Hsin-Hsi Chen, Ph.D.

中華民國 107 年 8 月

August, 2018

## 誌謝



能完成這篇論文，我要特別感謝我的指導教授陳信希老師，教導我很多研究必備知識，使我能更有效率的從事研究，並在每次的討論中總是有耐心地給予我很多的建議以及啟發，讓我得以成長前進。也非常感謝我的學長黃瀚萱博士，總是能迅速的突破研究盲點，協助我解決各種遇到的問題。感謝我的學長姐們：昂穎、安孜、重吉、又增、家禾、祐婷、勤和、微涓、宇祥、仕弘，給予我很多建議以及經驗分享，使我能有更多的研究思路。也感謝我的學弟妹們：好文、宏國、怡婷、敏桓、忠憲、奎伯、黃晴，分享很多不同的想法，使我有許多參考資源。特別感謝我的同屆同學：博政、宗翰、佳文、瑋柔、傳恩、斯文，在研究所的生涯受到你們各種照顧，幫我解決無數困難，真的非常感謝你們。我要感謝我的所有家人，謝謝爸媽，養育教導我並給我自由空間選擇自己人生。感謝我的岳父母，能夠支持我並且不辭辛勞地照顧我的女兒。感謝我的女兒帶給我最正面的力量，最後我要特別感謝我的妻子，沒有你的陪伴，我無法走出低潮，有了你的支持，讓我再次看見未來美好。

## 中文摘要



受惠於科技的進步，人們可以隨時隨地用相機或智能手機拍照來記錄生活。但是照片無法保存完整的信息。因此需要使用文字紀錄整個故事並保留一些特定信息當作圖片訊息的補充。許多人選擇編寫圖文交織的部落格使得生活記憶得以保存。但是像痞客邦這樣的熱門部落格網站並沒有照片回憶功能。而谷歌相簿雖然有基本的照片搜索功能，但此搜索功能卻不支援圖片上下文相關故事信息的搜索。據我們所知，這是第一個針對圖文生活日誌進行圖片回憶的研究。

我們從痞客邦收集圖文生活日誌資料集“Blog-travel”，並模仿人們對此資料集從五種不同面向進行圖片回憶標記。我們另外從痞客邦搜集了更大的資料集“Blog-travel-large”來做更多訓練和比較。

此外，我們比較了一些圖片和文字的嵌入編碼器，並提出了“圖片模型”和“故事模型”來做圖片回憶檢索。圖片模型透過無監督式的圖文嵌入學習，可以將圖片和文字嵌入到同一個空間中，進而可以用文字對圖片做檢索。而故事模型單純使用圖片附近的故事來做文字對文字的檢索，在對應到鄰近圖片達成文字對圖片檢索。由於上述兩種模型具有互補性，因此我們將兩個模型結合成為一個模型“圖片故事模型”，此模型在“Blog-travel”做圖片回憶評分時的結果優於谷歌圖像搜索也優於訓練在 MSCOCO 資料集表現最好的圖文嵌入模型。我們更進一步地考慮了不同的 query 會造成相關故事和圖片間的距離差異，提出圖片故事注意力模型，使得表現更加提升。

關鍵字：圖文生活日誌，圖片回憶，圖文嵌入學習，圖片檢索

# ABSTRACT

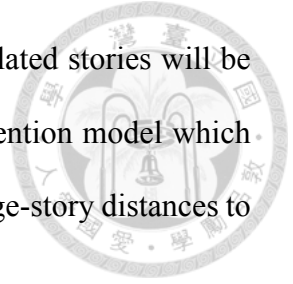


Benefit from the advancement of science and technology, people can easily take photos with cameras or smart phones anytime anywhere to record their life. However, photos cannot keep the complete information. Text is the complement to describe the whole story and keep some specific messages. Therefore, writing image-text intertwined lifelog is a popular way to keep life memory. And then how to retrieve image precisely between tons of images with context information in lifelogs is a big issue. The modern blog websites like PIXNET does not have function of photos recall. Another online photo storage like Google Photos has basic photo search function does not support to search photos with related story information. To the best of our knowledge, this is the first research addressing image recall on image-text intertwined lifelog.

We collect an image-text intertwined lifelog dataset “Blog-travel” from PIXNET, and to imitate people to do image recall on this dataset from five different points of view. Furthermore, we collect a bigger dataset “Blog-travel-large” to do more training and comparison.

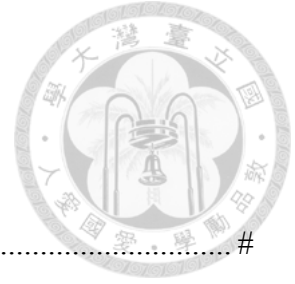
We compare some image and sentence encoders and propose Image model and Story model for image recall retrieval. Image model can transfer image and text to the same embedding space through unsupervised learning, so that the image can be retrieved by text. The Story model simply uses the story near the image to calculate text-text similarity score, and assign the score to the image to make image retrieval possible. Since the above two models are complementary, we combine the two models into the Image-story model. This model outperforms Google Image Search on image recall task on Blog-travel, and also outperforms the state-of-the-art model which is trained on MSCOCO dataset.

Moreover, we notice that the distance between the image and the related stories will be different by different queries. And then we propose Image-story attention model which combines different Image-story models which consider different image-story distances to get better performance.



Keywords: Image-text lifelog, Image recall, Image-text embedding learning, Image retrieval

# CONTENTS



口試委員會審定書.....	#
誌謝.....	i
中文摘要.....	ii
ABSTRACT.....	iii
CONTENTS .....	v
LIST OF FIGURES .....	vii
LIST OF TABLES.....	xi
<b>Chapter 1 Introduction.....</b>	<b>1</b>
1.1 Motivation .....	1
1.2 Image Retrieval.....	2
1.3 Personal Image Retrieval on Lifelog.....	2
1.4 Image-Text Embedding Learning .....	4
1.5 Thesis Organization .....	4
<b>Chapter 2 Related Work.....</b>	<b>5</b>
2.1 Image Retrieval.....	5
2.2 Image-Text embedding learning.....	5
2.3 Lifelog and Memory Recall .....	6
<b>Chapter 3 Blog-Travel Dataset .....</b>	<b>8</b>
3.1 Data Collection .....	8
3.2 Annotation of Image Recall .....	12
3.3 Evaluation .....	19

<b>Chapter 4</b>	<b>Models</b> .....	<b>21</b>
4.1	<b>Image-Story Model Structure</b> .....	<b>21</b>
4.2	<b>Image Model</b> .....	<b>22</b>
4.2.1	Embedding Loss Function .....	24
4.2.2	From Supervised Learning to Unsupervised Learning .....	26
4.2.3	Image Encoder and Text Encoder .....	28
4.3	<b>Story Model</b> .....	<b>29</b>
4.4	<b>Image-Story Attention Model</b> .....	<b>30</b>
<b>Chapter 5</b>	<b>Image Recall on Blog-Travel Dataset</b> .....	<b>32</b>
5.1	<b>Experiment of Two Baseline Model</b> .....	<b>32</b>
5.2	<b>Experiment of Image Model</b> .....	<b>33</b>
5.3	<b>Experiment of Story Model</b> .....	<b>36</b>
5.4	<b>Experiment of Image-Story Attention Model</b> .....	<b>38</b>
5.5	<b>Results</b> .....	<b>39</b>
<b>Chapter 6</b>	<b>Conclusion and Future Work</b> .....	<b>48</b>
<b>REFERENCE</b>	.....	<b>49</b>



# LIST OF FIGURES



Figure 1-1 Example of different stories with the same photo are represented different memories. ([https://cdn.pixabay.com/photo/2018/06/12/01/04/road-3469810\\_\\_480.jpg](https://cdn.pixabay.com/photo/2018/06/12/01/04/road-3469810__480.jpg)) ..... 1

Figure 1-2 Example of image search of “Google Photos”. The result shows the search does not consider the related text stories near the images. .... 3

Figure 1-3 In blog, there is no precise caption of a target image. Instead, there are perhaps related stories around the image..... 3

Figure 3-1 Details of blog-travel dataset from 30 authors. The left column is authors’ ID. A: article. I: image. C: character. We control the number of articles of each author is between 10 to 99. .... 9

Figure 3-2 Blog-travel-large dataset density plot. Light spot means high density. Dark spot means low density. .... 11

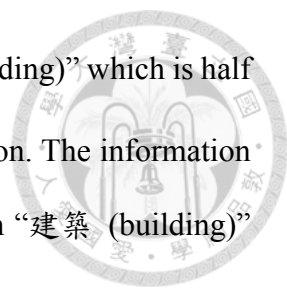
Figure 3-3 Example of annotation in “我吃過的食物 (What food I ate)” ..... 13

Figure 3-4 Example of annotation in “我住過的旅館 (What accommodation I stayed)” ..... 13

Figure 3-5 Example of annotation in “換句話說 (In other words)”, the annotator sets the query to be “代表性的雕刻品” which has the similar meaning as “指標性的雕塑作品” just in other words. Both of them have meaning “Representative carvings”. Notice that annotation could be non-continuous. .... 14

Figure 3-6 Example of annotation in “圖文結合(image-text intertwined)”, the annotator





set the query to be “古代風格的建築 (Ancient style building)” which is half from context information and half from image information. The information “古代(ancient)” is from the context, and the information “建築 (building)” is from the image. Note that image answers from different queries could be same. (This image is also the answer for another query)..... 15

Figure 3-7 Example of annotation in “最重要的回憶(The most important memory”, the annotator set the query to be “藝術家要求的特殊展覽方式 (Special exhibition methods requested by the artist)” which is thought to be the most important memory in the article..... 16

Figure 3-8 Example of annotation in “最重要的回憶 (The most important memory”, the annotator set the query to be “比利時代表性紀念品 (Representative souvenirs from Belgium)” which is thought to be the most important memory in the article..... 17

Figure 3-9 The distribution of distance between reference image and the closest reference sentence. .... 18

Figure 3-10The distribution of distance between reference image and all reference sentences. .... 18

Figure 3-11Choosing MAP@10 or NMAP@10 to be the metric when doing evaluation on 5 types of image recalls on Blog-travel dataset (“ac” means accommodation, “img\_n” means total number of image of the author). The NMAP@10 metric is much simpler. And we will show more comparison between NMAP and 30 MAPs in Section 5.4..... 20

Figure 4-1 Image-Story model structure..... 21

Figure 4-2 Structure of the learning stage which could train the sentence and image into

a coordinated embedding. “fc” means fully-connected. ....	23
Figure 4-3 Structure of Image model when doing image recall. ....	23
Figure 4-4 Positive pairs and negative pairs to be used in embedding loss function. ....	24
Figure 4-5 Example of the sentences which is within distance 3 (red circles) from the image is the corresponding sentences of the image. ....	27
Figure 4-6 The performance of different image encoder models and sentence encoder models on MSCOCO dataset. R@K which means recall at K is the common evaluation on image-text retrieval task. ....	28
Figure 4-7 Structure of story model. ....	29
Figure 4-8 Structure of image-story attention model when training. Where APs are average precision of retrieving image by using the query. W1, W2, ..., W9 are the weights which are determined by the query. ....	31
Figure 4-9 Structure of image-story attention model when doing image recall. Where W1, W2, ..., W9 are the weights which is determined by the query. ....	31
Figure 5-1 Distance between image and all relevant sentences in annotations. ....	37
Figure 5-2 Performance comparison between the proposed Image-story attention model and Google image search on the 30 authors. Each plot contains 30 red spots and 30 blue spots. If a vertical line contains blue spot, the line should contain red spot as well. Where X-axis is total number of images of the author. Y-axis is the MAP@10 score. The big spot means the best performance model of the author. ....	41
Figure 5-3 The result of Google_image_search of type “food” on the author “altheawoman”. ....	42
Figure 5-4 The result of our proposed Image-story model of query type “food” on the author “altheawoman”. ....	42

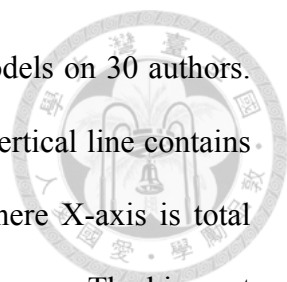


Figure 5-5 Performance comparison between our proposed four models on 30 authors. Each plot contains 30 red spots and 30 blue spots. If a vertical line contains blue spot, the line should contain red spot as well. Where X-axis is total number of images of the author. Y-axis is the MAP@10 score. The big spot means the best performance model of the author..... 43

Figure 5-6 The searching result of query type “Q1”. Query is “前門各個栩栩如生的雕像彷彿帶我們回到聖經裡的場景(The vivid statues of the front doors seem to bring us back to the scene in the Bible)”..... 44

Figure 5-7 The searching result of query type “Q1”. Query is “年歲久遠而且很獨特的教堂 (Old and very unique church)”..... 45

Figure 5-8 The searching result of query type “Q2”. Query is “點了星巴克的三明治和咖啡(Starbucks sandwiches and coffee)”..... 45

Figure 5-9 The searching result of query type “Q2”. Query is “入場券上有蒙娜麗莎的微笑(Mona Lisa smile on the ticket)”..... 46

Figure 5-10The searching result of image-story model from partial queries. .... 46

Figure 5-12The searching result of query type “Q3”. Query is “法國家常料理體驗 (French home cooking experience)”..... 47

Figure 5-13The searching result of query type “Q3”. Query is “凡爾賽宮中的名畫 (Famous paintings in Versailles)”..... 47

# LIST OF TABLES

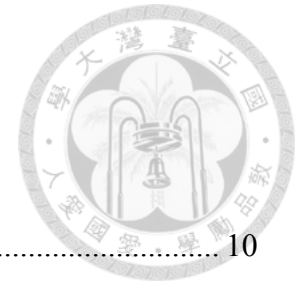
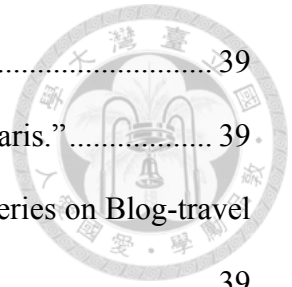


Table 3-1	Detail and statistics of two datasets. ....	10
Table 3-2	Five types of annotation and responding reasons .....	12
Table 4-1	The performance of TBN_MSCOCO is apparently not as well as Google_image_search where TBN_MSCOCO refers Wang et al. [14] structure. From the statistics (See Figure 3-9), we consider the nearby sentences which are within distance 3 from an image to be the corresponding sentences of the image (See Figure 4-5). Instead of using the pairs of corresponding captions and the image (e.g., MSCOCO). We apply the pairs of nearby sentences and the image. Therefore, we do not need any caption annotation.....	26
Table 5-1	Baseline model results of 5 types performance on Blog-travel. ....	32
Table 5-2	Comparison between different sentence encoders on image model. Metric is NMAP@10. Seg: jieba segmentation. w2v: genism word2vec which is pretrained on Chinese wiki. All of four models are near_3. ....	34
Table 5-3	Comparison between different distance between the considered related stories and the image. Metric is NMAP@10. Near_k: distance between sentence and image to do the unsupervised learning. Model is Google translate + InferSent with near_k.....	34
Table 5-4	Comparison between different training data. Both model use near_3.....	35
Table 5-5	Comparison between different distance on story model. Metric is NMAP@10. Near_k: distance between sentence and image to be assigned score.....	36
Table 5-6	Results of near_1 to near_9 and attention model. Metric is NMAP@10. ....	38

Table 5-7	The weights of 9 models for the query “The food I ate.” .....	39
Table 5-8	The weights of 9 models for the query “Ferris wheel in Paris.” .....	39
Table 5-9	Performance of the baseline models for the 5 types of queries on Blog-travel dataset. ....	39
Table 5-10	Performance of the proposed models for the 5 types of queries on Blog-travel dataset. ....	39



# Chapter 1 Introduction



## 1.1 Motivation

Benefit from the advancement of science and technology, people can easily take photos with cameras or smart phones anytime anywhere to record their life. Photo has two advantages to keep memory: 1) Real visual scene at a moment. 2) Detailed information. (ex: Color? Size? How many?) However, photos cannot keep the complete information. For example, the same image with different stories are represented different memories (See Figure 1-1). Text provides important clues to describe the whole story and keeps some specific messages. Therefore, writing image-text intertwined blogs is a popular way to keep life memory.



- Beautiful road with many trees.

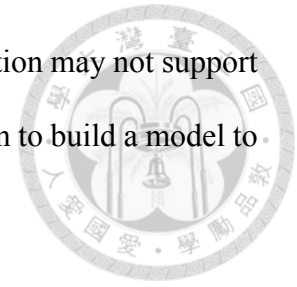


- Hot road with just little trees.

Figure 1-1 Example of different stories with the same photo are represented different memories. ([https://cdn.pixabay.com/photo/2018/06/12/01/04/road-3469810\\_\\_480.jpg](https://cdn.pixabay.com/photo/2018/06/12/01/04/road-3469810__480.jpg))

However, while people want to recall some photos, how to retrieve image precisely between tons of images with related story information in the blog is a big issue. The modern blog websites like PIXNET does not have function of photo recall. Another

online photo storage like Google Photos has basic photo search function may not support to search photos with related story information. In this thesis, we plan to build a model to deal with image recall task.



## **1.2 Image Retrieval**

Image retrieval is a technique for searching an image database from an user's interested query. There are two traditional image retrieval models: text-based image retrieval (TBIR) and content-based image retrieval (CBIR). Text-based model uses metadata such as keywords or descriptions corresponding to the image so that image retrieval can work via text semantic similarity. While content-based image retrieval (CBIR) uses the contents (color, shape, texture information etc.) that can be extracted from the image itself so that image retrieval can work via visual similarity. We plan to build a model containing text information and content information.

## **1.3 Personal Image Retrieval on Lifelog**

In recent years, writing image-text intertwined blog is a popular way to record personal life. Therefore, personal image retrieval is needed if the number of images is large. Off-the-shelf application like "PIXNET" and "Google Photos" (Figure 1-2) may not have this function or still have a lot of room for improvement. Hence, it is still an ongoing problem to be solved.

It is different from traditional image retrieval models mentioned in section 1.2. There is no precise caption of a target image. Instead, there are perhaps related stories around the image. (Figure 1-3)

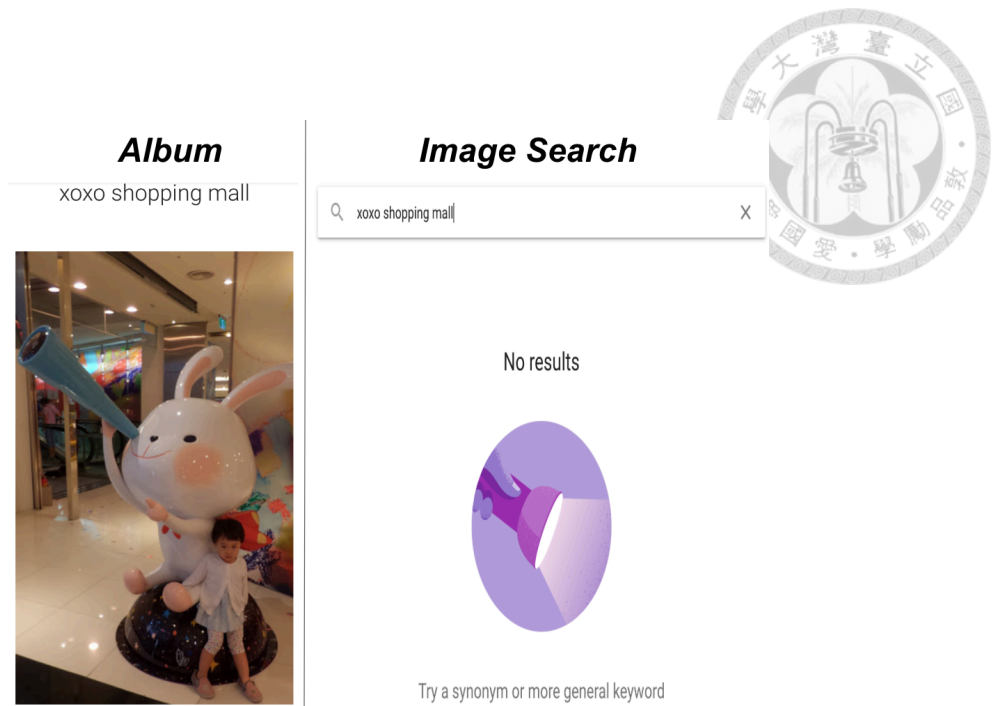
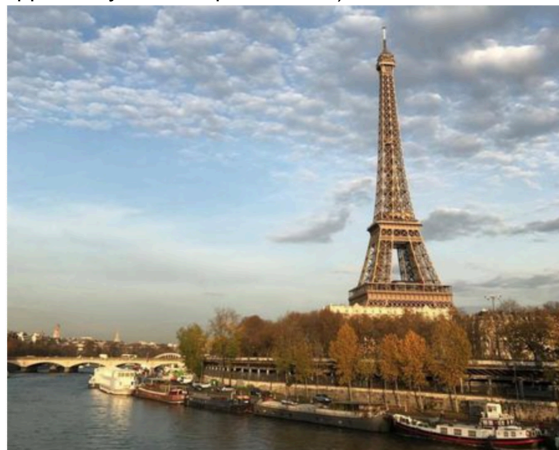


Figure 1-2 Example of image search of “Google Photos”. The result shows the search does not consider the related text stories near the images.

### *blog*

不過現在想來還真有點可惜，下次有機會應該上去看看  
 (But it's really a pity to think about it now. I have the opportunity to look up next time.)



艾菲爾鐵塔不管白天還是晚上都很美，而且非常的壯觀  
 (The Eiffel Tower is beautiful both day and night, and it is very spectacular.)

Figure 1-3 In blog, there is no precise caption of a target image. Instead, there are perhaps related stories around the image.



## 1.4 Image-Text Embedding Learning

Image and text representations are important when we want to use image or text information on computation. Learning image embedding or text embedding is a modern way for image or text representation. The embedding vector is shown to achieve good performance in many tasks. In order to achieve better performance on some multi-model tasks like image-text retrieval task, learning the coordinated representation embedding of multi-model is a popular way. That is, learning a coordinated embedding space for both image and text could help do image-text retrieval straightly.



## 1.5 Thesis Organization

This thesis is organized as follows. In Chapter 2, we will introduce the related works and discuss what the differences are between these works and our work. In Chapter 3, we will discuss what dataset we need to do the image recall task and how to collect the data, how to imitate people do image recall annotation and how to evaluate the system performance. In Chapter 4, we will introduce our four models and how they do the image recall. In Chapter 5, we will show all the results and compare them. In Chapter 6, we will conclude our contributions and suggest the future works.

## Chapter 2 Related Work



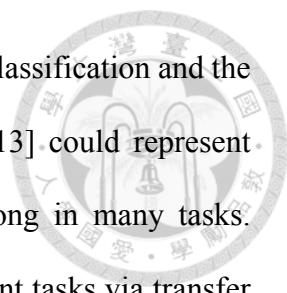
### 2.1 Image Retrieval

Image retrieval becomes more and more important since the advancement of science and technology like big storage and camera. There are many surveys of this research direction [1][2][3]. Early techniques to image retrieval were basically based on the textual annotation of images. Images were needed to annotated first with text and then searched using a text-based approach. However, annotations related to the image is not cheap and the performance of this approach to image retrieval is very sensitive to the keywords. Therefore, content-based image retrieval (CBIR) becomes important and popular recently. How to extract features from an image is the key point of CBIR. Color, shape and texture are important features for image retrieval and many researches focus on those features extraction and analysis [4][5][6][7][8]. However, it is still not enough when using those features to do image retrieval because image contains much information.

There are many datasets such as MSCOCO and Flickr30k which contain images and corresponding captions so that image retrieval model can be trained and test on them. But to the best of our knowledge, there are no dataset to do the image recall task via image-text intertwined lifelog. Therefore, we need to build a new image-text intertwined dataset to approach our goal.

### 2.2 Image-Text embedding learning

Image and text representations are important when we want to use image or text information on computation. Learning image embedding or text embedding is modern way for image or text representation. VGGNet [9] and ResNet [10] could encode image



into image embedding and achieve good performance on ImageNet classification and the related tasks. Skip-thought vector [11], InferSent [12] and USE [13] could represent sentences in a single vector as sentence embedding which is strong in many tasks. Furthermore, these embedding models are also good at many different tasks via transfer learning. In order to achieve better performance on some multi-model tasks like image-text retrieval task, learning the coordinated representation embedding of multi-model is a popular way.

Zhedong Zheng et al. [14] use dual-path convolutional structure to learn the image-text coordinated embedding. However, the structure assumes every image is one class and to do the classification problem. This method seems not reasonable for the dataset which contains many similar images. And it is very hard to do the classification problem if the number of image class is too large. Liwei Wang et al. [15] use the structure which is called two-branch neural networks to learn the image-text coordinated embedding, and achieve the state-of-the-art performance on many Image-text matching tasks. But this method is based on supervised learning which needs annotated datasets. In our dataset, there is no supervised annotation to train. Therefore, we refer two-branch neural networks structure but we apply unsupervised learning which uses the stories near the image instead of the annotated captions.

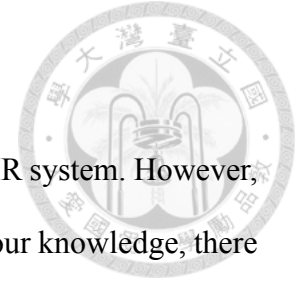
### **2.3 Lifelog and Memory Recall**

Many lifelog research focuses on how to extract the video information from the sensors. Kiyoharu Aizawa et al.[16] proposed a system which could do key frame extraction, human voice detection and human face detection. However, some memory cannot be recorded by any sensor because people may just keep their thoughts in their

minds. Sometimes, people like to write down the memory on dairy or blog as lifelog. Therefore, we want to do the research deeper in memory of mind.

Lu Jiang et al. [17] build a QA system which could answer question from albums. However, the image and text in their MemexQA dataset is paired not intertwined. It is hard to keep the whole story of memory. Therefore, we need to build a new image-text intertwined dataset to approach our goal.

## Chapter 3 Blog-Travel Dataset



There are many datasets whose goals are to evaluate the QA or IR system. However, these datasets are based on pair of image and caption. To the best of our knowledge, there is no dataset to evaluate image retrieval system on image-text intertwined blog. Therefore, we build a dataset “Blog-Travel” which is selected from real blogs. We try to imitate the author to annotate the recall query and the corresponding image answers.

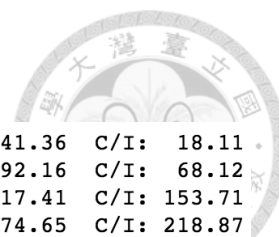
### 3.1 Data Collection

Writing blog is a popular way to record personal memory with both images and texts. We consider PIXNET<sup>1</sup>, which is a popular and good resource to collect blogs. From PIXNET we collect a small dataset “Blog-travel” which contains 26,198 images in 1,373 travel articles from 30 authors. To imitate recalling image from authors, we recruit annotators to annotate 30-35 questions for each author. Furthermore, we collect larger dataset “Blog-travel-large” which contains 345,564 images and 14,831 articles from other authors to do the model training (see Chapter 4).

There is a ubiquitous phenomenon that many blogs are written for commercial purposes especially in some domains like “3C” or “game”. In order to reduce that commercial situation, we choose the “travel” domain to get better quality of personal memory blogs. But there are still many commercial blogs like introducing hotel or restaurant. Therefore, in “Blog-travel” dataset (see Figure 3-1), we choose a popular sightseeing spot “Eiffel Tower” as a search seed that is less intention to be advertised.

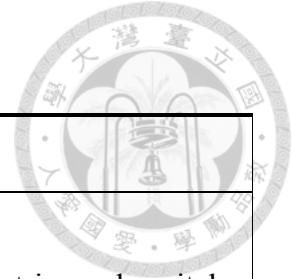
---

<sup>1</sup> PIXNET is the most interested social media in Taiwan and its “Blog” service brings together a rich variety of content. See <https://www.pixnet.net/blog>



aa0811	:	A: 11	I: 511	C: 9255	I/A:46.45	C/A: 841.36	C/I: 18.11
ajin	:	A: 64	I: 1402	C: 95498	I/A:21.91	C/A:1492.16	C/I: 68.12
altheawoman	:	A: 17	I: 201	C: 30896	I/A:11.82	C/A:1817.41	C/I: 153.71
anchusa	:	A: 46	I: 394	C: 86234	I/A: 8.57	C/A:1874.65	C/I: 218.87
brezel	:	A: 23	I: 843	C: 20522	I/A:36.65	C/A: 892.26	C/I: 24.34
earthmaoi	:	A: 37	I: 1422	C: 29654	I/A:38.43	C/A: 801.46	C/I: 20.85
eric8503eric	:	A: 24	I: 865	C: 26855	I/A:36.04	C/A:1118.96	C/I: 31.05
garytzeng	:	A: 88	I: 1765	C: 64739	I/A:20.06	C/A: 735.67	C/I: 36.68
hsfyang	:	A: 56	I: 761	C: 23960	I/A:13.59	C/A: 427.86	C/I: 31.48
hsuan1203	:	A: 99	I: 1935	C: 74272	I/A:19.55	C/A: 750.22	C/I: 38.38
huiyichen01	:	A: 36	I: 708	C: 12179	I/A:19.67	C/A: 338.31	C/I: 17.20
ireneho73	:	A: 33	I: 448	C: 21565	I/A:13.58	C/A: 653.48	C/I: 48.14
jkismn33	:	A: 53	I: 2824	C:110101	I/A:53.28	C/A:2077.38	C/I: 38.99
joannlsf	:	A: 20	I: 434	C: 35156	I/A:21.70	C/A:1757.80	C/I: 81.00
kurodayeh	:	A: 98	I: 2124	C: 73834	I/A:21.67	C/A: 753.41	C/I: 34.76
linama789	:	A: 61	I: 646	C:104225	I/A:10.59	C/A:1708.61	C/I: 161.34
mumumuas	:	A: 60	I: 845	C: 41246	I/A:14.08	C/A: 687.43	C/I: 48.81
nina3	:	A: 60	I: 678	C: 34958	I/A:11.30	C/A: 582.63	C/I: 51.56
regnarts	:	A: 18	I: 166	C: 4823	I/A: 9.22	C/A: 267.94	C/I: 29.05
saint616	:	A: 56	I: 2330	C: 67678	I/A:41.61	C/A:1208.54	C/I: 29.05
silkroadskyjiao	:	A: 33	I: 510	C: 27224	I/A:15.45	C/A: 824.97	C/I: 53.38
teresachuang1105	:	A: 59	I: 566	C: 36217	I/A: 9.59	C/A: 613.85	C/I: 63.99
theresa1103	:	A: 99	I: 1294	C:118798	I/A:13.07	C/A:1199.98	C/I: 91.81
tsaihuifang	:	A: 25	I: 316	C: 20603	I/A:12.64	C/A: 824.12	C/I: 65.20
webber24	:	A: 21	I: 150	C: 7958	I/A: 7.14	C/A: 378.95	C/I: 53.05
yas5246	:	A: 17	I: 173	C: 14175	I/A:10.18	C/A: 833.82	C/I: 81.94
yfcherrypan	:	A: 22	I: 299	C: 19664	I/A:13.59	C/A: 893.82	C/I: 65.77
yichingsays	:	A: 90	I: 731	C: 34475	I/A: 8.12	C/A: 383.06	C/I: 47.16
yoyovilla	:	A: 33	I: 781	C: 74058	I/A:23.67	C/A:2244.18	C/I: 94.82
yunyuan	:	A: 13	I: 76	C: 13159	I/A: 5.85	C/A:1012.23	C/I: 173.14

Figure 3-1 Details of blog-travel dataset from 30 authors. The left column is authors' ID. A: article. I: image. C: character. We control the number of articles of each author is between 10 to 99.



<i>Detail</i>	<i>Blog-travel</i>	<i>Blog-travel-large</i>
Data source	PIXNET	PIXNET
Seed of crawler	“Eiffel tower”	Tourist attractions, countries and capitals
<i>Statistics</i>	<i>Blog-travel</i>	<i>Blog-travel-large</i>
Authors	30	6,550
Articles	1,373	14,831
Images	26,198	345,564
Characters	1,333,981	24,718,928
Images/Articles	19	23
Characters/Articles	971	1,666

Table 3-1 Detail and statistics of two datasets.

In “Blog-travel-large” dataset, we set many other search seeds like tourist attractions, countries and capitals all around world to get more diversity. And we get 14,831 articles from 6,550 authors which show high diversity as well. Table 3-1 shows the detail of two datasets.

We plot each article in Blog-travel-large as a spot on Figure 3-2 to know the habit of blog writers. Where x-axis is number of images of one article, and y-axis is number of characters of one article. From Blog-travel-large dataset density plot, most of all articles contains 0-30 images and 0-2,000 characters. The number of images and the number of characters are positive correlative for most of all articles. That is, the article with more images needs more stories. But when the number of images or characters are big enough, the correlation between them becomes negative. The probably reason of this phenomenon might be that the author does not want to spend too much time on writing an article. In other words, the author uploading too many images may not want to write stories for all of them.

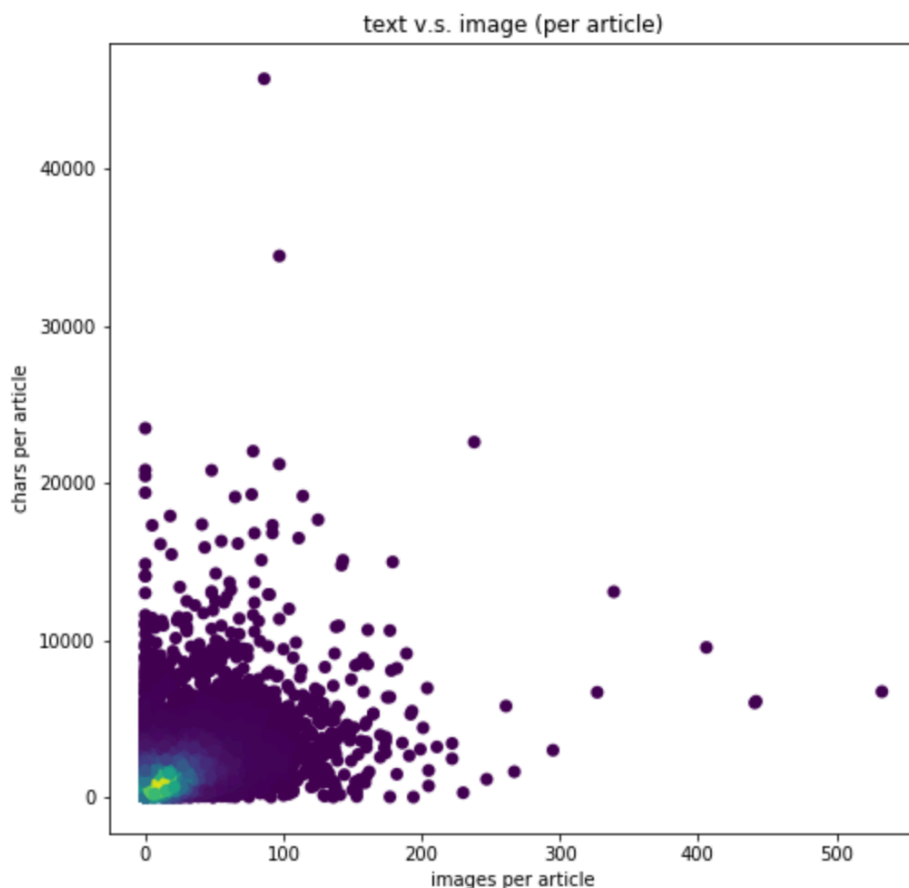
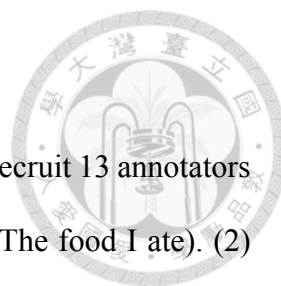


Figure 3-2 Blog-travel-large dataset density plot. Light spot means high density. Dark spot means low density.





### 3.2 Annotation of Image Recall

In order to imitate the author to recall memory in the future, we recruit 13 annotators to annotate five types of memory annotations: (1) 我吃過的食物 (The food I ate). (2) 我住過的旅店 (The accommodation I stayed in). (3) 換句話說 (In other words). (4) 圖文結合 (Image-text combined). (5) 最重要的回憶 (The most important memory) (See Table 3-2). Each author is annotated 30-35 query-answers pairs by 6-7 annotators. Details and examples are described as follows.

<i>Query type name</i>	<i>Reason</i>
我吃過的食物 (The food I ate)	Eating is daily routine
我住過的旅店 (The accommodation I stayed in)	Accommodation is daily routine when traveling
換句話說 (In other words)	Imitate situation that human use different words to recall
圖文結合 (Image-text intertwined)	Imitate situation that human use image-text intertwined information to recall
最重要的回憶 (The most important memory)	Recall the most important memory

Table 3-2 Five types of annotation and responding reasons



(1) 我吃過的食物 (The food I ate). Eating is a daily routine and people like to record and recall what they eat. Annotators imitate author to annotate the images and contexts of what food they eat in the article. (See Figure 3-3)

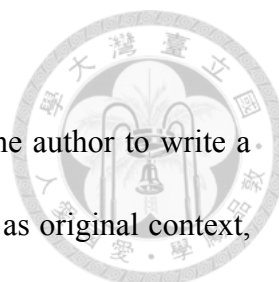


Figure 3-3 Example of annotation in “我吃過的食物 (What food I ate)”

(2) 我住過的旅館 (The accommodation I stayed). In traveling blogs, the accommodation might be another daily routine to record when traveling. Annotators imitate author to annotate the images and contexts of what accommodation they stayed in the article. (See Figure 3-4).



Figure 3-4 Example of annotation in “我住過的旅館 (What accommodation I stayed)”



(3) Q1 換句話說 (In other words). The annotator need to imitate the author to write a query to search the image. The query cannot use the same keyword as original context, but the meaning of query and context should be similar. The annotator also needs to tick the reference images and sentences. (See Figure 3-5).

**Q1 (換句話說) :**

指標性的雕塑作品

力與美的完美表現



Figure 3-5 Example of annotation in “換句話說 (In other words)”, the annotator sets the query to be “代表性的雕刻品” which has the similar meaning as “指標性的雕塑作品” just in other words. Both of them have meaning “Representative carvings”. Notice that annotation could be non-continuous.

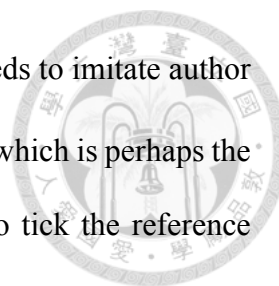
(4) Q2 圖文結合 (Image-text combined). Annotator needs to imitate author to write query to search the image. The query must include information from both the images and nearby stories. The annotator also needs to tick the reference images and sentences (See Figure 3-6).

Q2 (圖文結合) :

讓人如置身於古希臘時代~



Figure 3-6 Example of annotation in “圖文結合 (image-text intertwined)”, the annotator set the query to be “古代風格的建築 (Ancient style building)” which is half from context information and half from image information. The information “古代 (ancient)” is from the context, and the information “建築 (building)” is from the image. Note that image answers from different queries could be same. (This image is also the answer for another query)



(5) Q3 最重要的回憶 (The most important memory). Annotator needs to imitate author to write query to search the image. The annotator thinks of the query which is perhaps the most important memory for the author. The annotator also needs to tick the reference images and sentences (See Figure 3-7 and Figure 3-8).

Q3 (重點回憶) : 藝術家要求的特殊展覽方式

---



進入橘園後，就不難理解莫內何以要求這樣的展出方式了

坐在展覽室中央的長椅上，直感覺自己被這些畫作所包圍環繞，彷彿自己置身在莫內的花園裡

內心的某一處被打動了，凝視著畫，有種激昂的情緒湧上

Figure 3-7 Example of annotation in “最重要的回憶(The most important memory)”, the annotator set the query to be “藝術家要求的特殊展覽方式 (Special exhibition methods requested by the artist)” which is thought to be the most important memory in the article.

Q3 (重點回憶) :

比利時代表性紀念品

- 我覺得我買的這個尿尿小童的很有代表性
- 最代表比利時的尿尿小童、巧克力還有歐式建築都上去了 很可愛

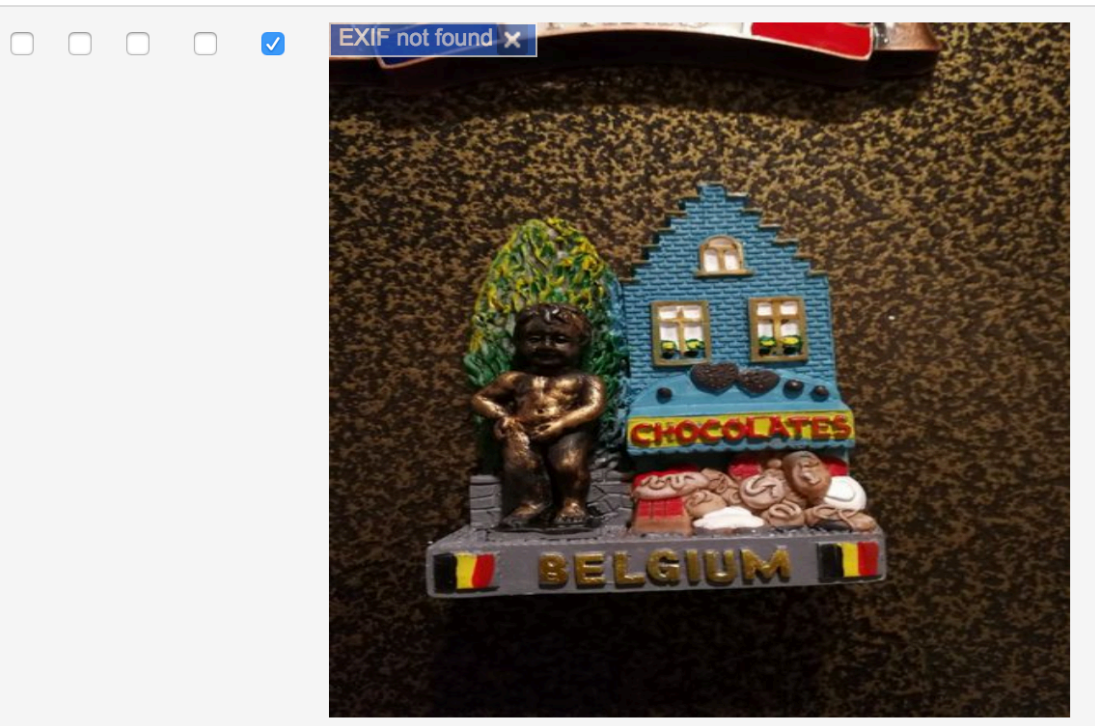


Figure 3-8 Example of annotation in “最重要的回憶 (The most important memory”, the annotator set the query to be “比利時代表性紀念品 (Representative souvenirs from Belgium)” which is thought to be the most important memory in the article.

From all annotations, we draw four histograms for four types of memory questions (See Figure 3-9). These histograms are the distribution of distance between the images and the closest reference sentence. These show that most of the first reference sentence is near the image within distance 3. We also draw other four histograms which show the distribution of distance between reference image and all reference sentences (See Figure 3-10). These show that some of reference sentences may be far away from the images.

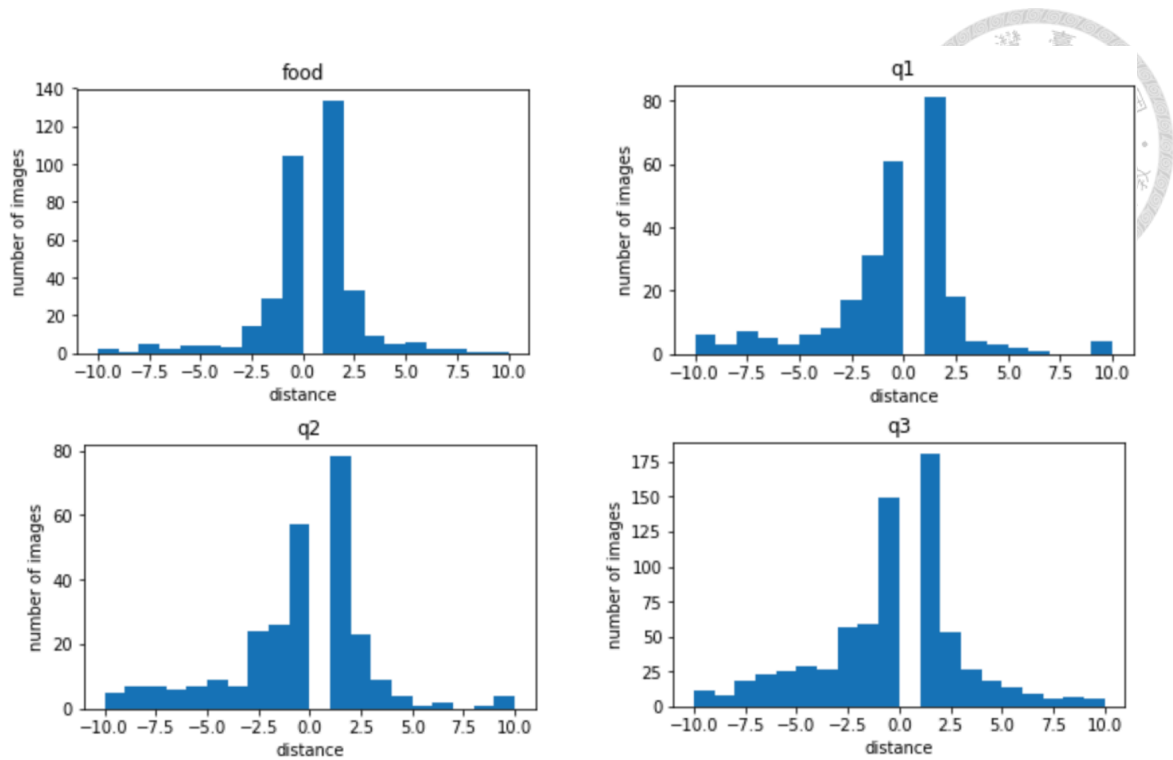


Figure 3-9 The distribution of distance between reference image and the closest reference sentence.

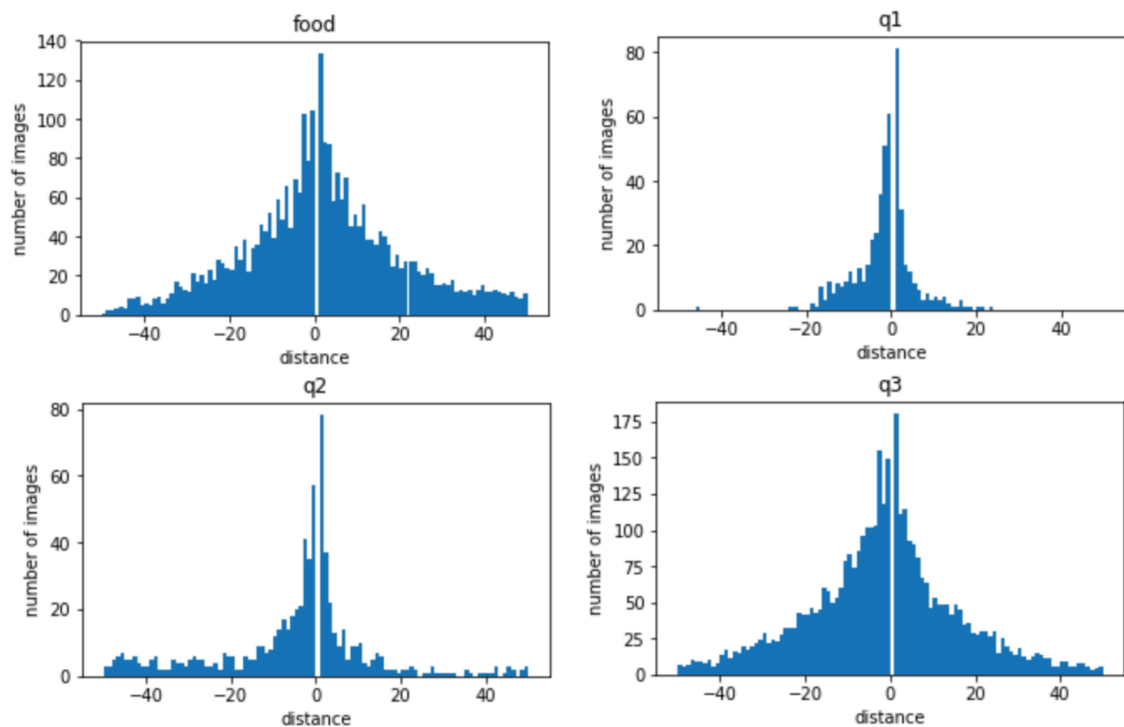
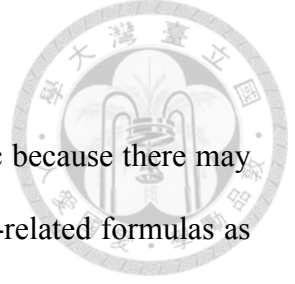


Figure 3-10 The distribution of distance between reference image and all reference sentences.



### 3.3 Evaluation

We choose mean average precision (MAP) as evaluation metric because there may be more than one answer of a given query. Here we list some MAP-related formulas as follows:

- $P@t : \frac{\text{number of true positive}}{t}$ , where  $t$  is predicted number
- $AP@k : \frac{1}{\min(k,n)} * \sum_{t=1}^k P@t$ , where  $n$  is number of correct answers
- $MAP@k : \text{mean}(AP@k)$

However, we need to do the image recall evaluation on each author independently. That is, we need to compare 30 MAPs from 30 authors (Figure 3-11). In order to do evaluation efficiently, we design a new metric “normalize mean average precision” (NMAP) as follows:

- $NMAP@k : \frac{1}{\sum_{l=1}^a \sqrt{I^{(l)}}} * \sum_{t=1}^a (\sqrt{I^{(t)}} * MAP@k^{(t)})$ ,

where  $a$  is the number of authors,  $I^{(t)}$  is total images of the author  $t$ .

We consider the square root of total number of images of the author as weight to normalize 30 MAP and sum all of them into just one score named NMAP@k.



# MAP@10

# NMAP@10

	food	ac	Q1	Q2	Q3	img_n
aa0811	0.146	NaN	0.231	0.306	0.292	511
ajin	0.000	0.125	0.375	0.292	0.310	1402
altheawoman	0.200	NaN	0.497	0.529	0.531	201
anchusa	0.000	NaN	0.312	0.438	0.293	394
brezel	0.000	0.125	0.289	0.382	0.036	843
earthmaoi	0.000	NaN	0.385	0.330	0.305	1422
eric8503eric	0.122	0.133	0.000	0.000	0.149	865
garytzeng	0.173	0.200	0.073	0.154	0.149	1765
hsfyang	0.090	NaN	0.108	0.107	0.320	761
hsuan1203	0.000	NaN	0.300	0.417	0.415	1935
huiyichen01	0.200	NaN	0.092	0.333	0.026	708
ireneh073	0.278	NaN	0.146	0.183	0.143	447
jkismn33	0.050	0.033	0.524	0.014	0.048	2824
joannlsf	0.012	NaN	0.093	0.357	0.378	434
kurodayeh	0.000	NaN	0.472	0.332	0.187	2124
linama789	0.000	0.000	0.103	0.249	0.375	646
mumumuas	0.000	NaN	0.239	0.295	0.333	845
nina3	0.000	NaN	0.094	0.375	0.226	678
regnarts	NaN	NaN	0.228	0.083	0.089	166
saint616	0.344	0.095	0.250	0.306	0.138	2330
silkroadskyjiao	0.000	NaN	0.471	0.235	0.369	510
teresachuang1105	0.000	NaN	0.499	0.137	0.469	566
theresa1103	0.010	NaN	0.192	0.040	0.046	1294
tsaihuifang	0.095	0.333	0.286	0.361	0.065	316
webber24	0.540	NaN	0.475	0.267	0.189	150
yas5246	NaN	NaN	0.408	0.278	0.209	173
yfcherrypan	0.000	NaN	0.206	0.361	0.338	299
yichingsays	0.100	0.000	0.554	0.181	0.171	731
yoyovilla	0.642	0.722	0.756	0.333	0.261	781
yunyuan	NaN	0.000	0.394	0.449	0.531	76

food	ac	Q1	Q2	Q3
0.101	0.153	0.302	0.256	0.228

Figure 3-11 Choosing MAP@10 or NMAP@10 to be the metric when doing evaluation on 5 types of image recalls on Blog-travel dataset (“ac” means accommodation, “img\_n” means total number of image of the author). The NMAP@10 metric is much simpler. And we will show more comparison between NMAP and 30 MAPs in Section 5.4.



# Chapter 4 Models

## 4.1 Image-Story Model Structure

When users need to do image recall, there is an abstract memory in their mind. They need to transform the abstract memory into a text query. We propose Image model and Story model to compute similarity scores between query and images. Image model is based on unsupervised learning which trains the image and the text to the new coordinated embedding, where the image and the nearby stories will be close in the new embedding (See Section 4.2). Therefore, the cosine similarity between text query and images could be computed. Story model simply computes the cosine similarity between text query and all stories and assigns the similarity scores to nearby images (See Section 4.3). Due to the complementarity between these two models, we combine them by averaging their similarity scores to be the final scores (See Figure 4-1).

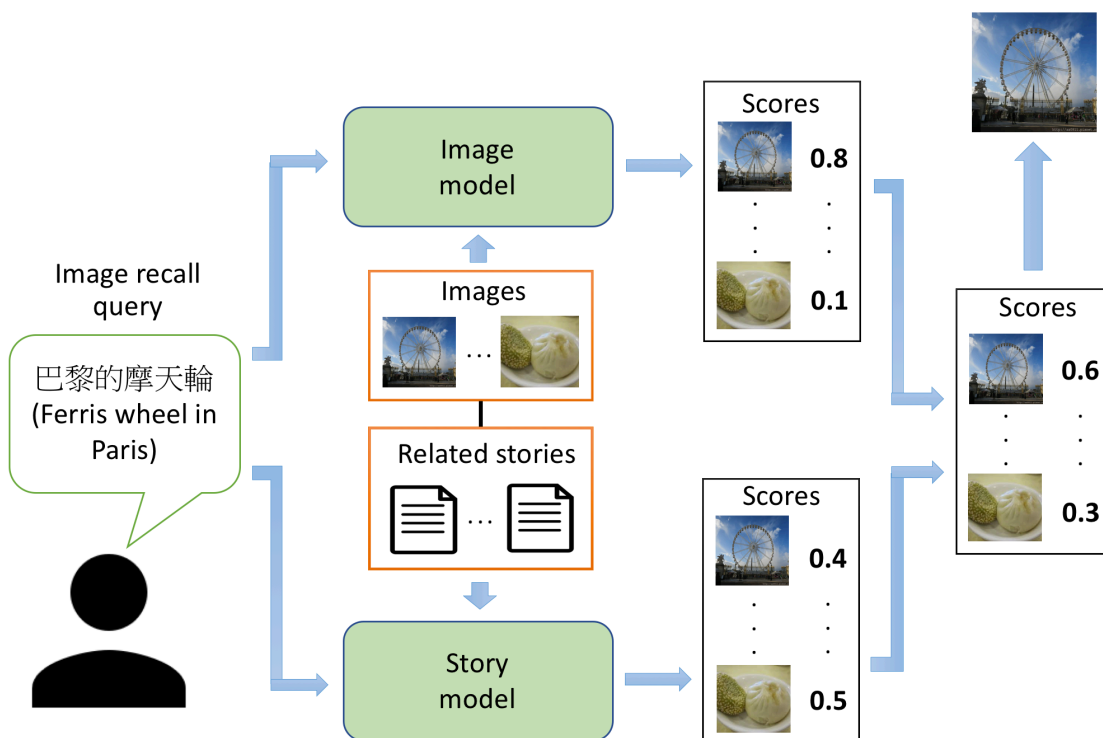
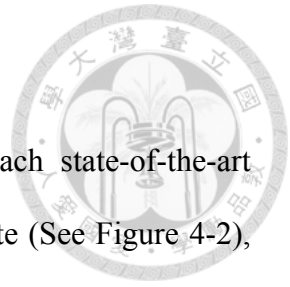


Figure 4-1 Image-Story model structure.

## 4.2 Image Model

We refer to the structure of L. Wang et al. [14] which reach state-of-the-art performance on many image-text retrieval tasks. In the learning state (See Figure 4-2), sentence encoder extracts sentence important features as sentence embedding and image encoder extract image features as image embedding. And then, we build neural network to train these two embedding into a new coordinated embedding. The embedding loss constrains the image and sentence from the corresponding pair (positive pair) will be close to each other in the new embedding. On the other side, the image and sentence from non-corresponding pair (negative pair) will be far away from each other in the new embedding. The detail of training will be discussed in Section 4.2.1.

After the new coordinated embedding is trained, we could input query and images to compute similarity scores (See Figure 4-3) to achieve image recall.



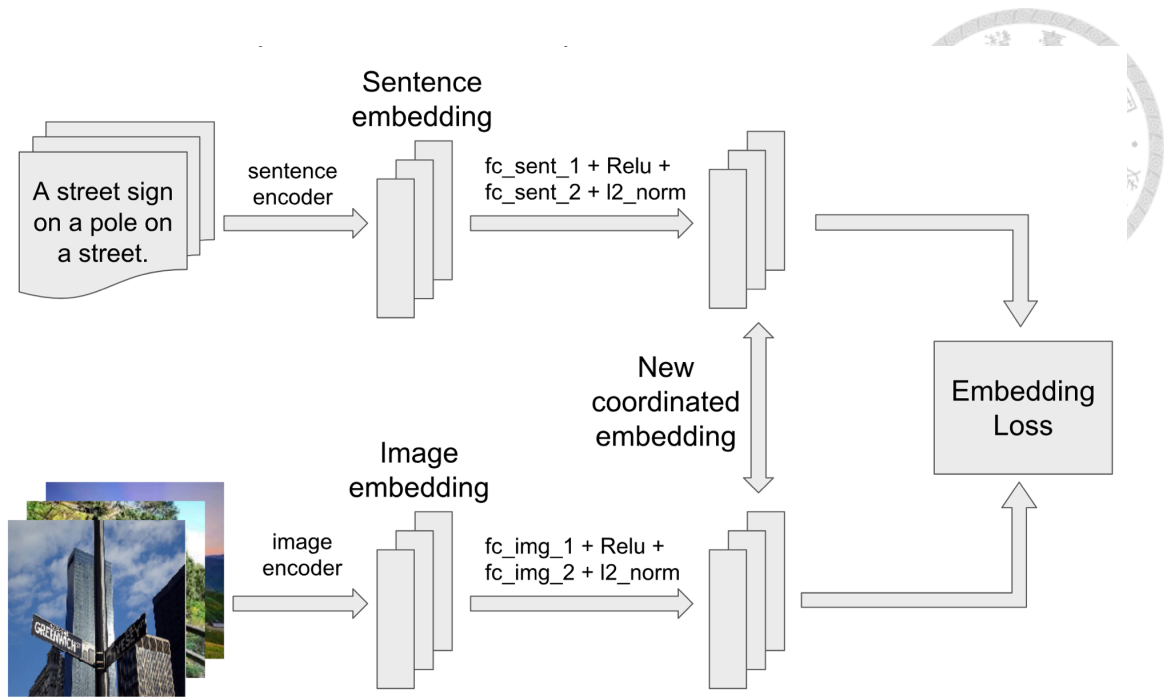


Figure 4-2 Structure of the learning stage which could train the sentence and image into a coordinated embedding. “fc” means fully-connected.

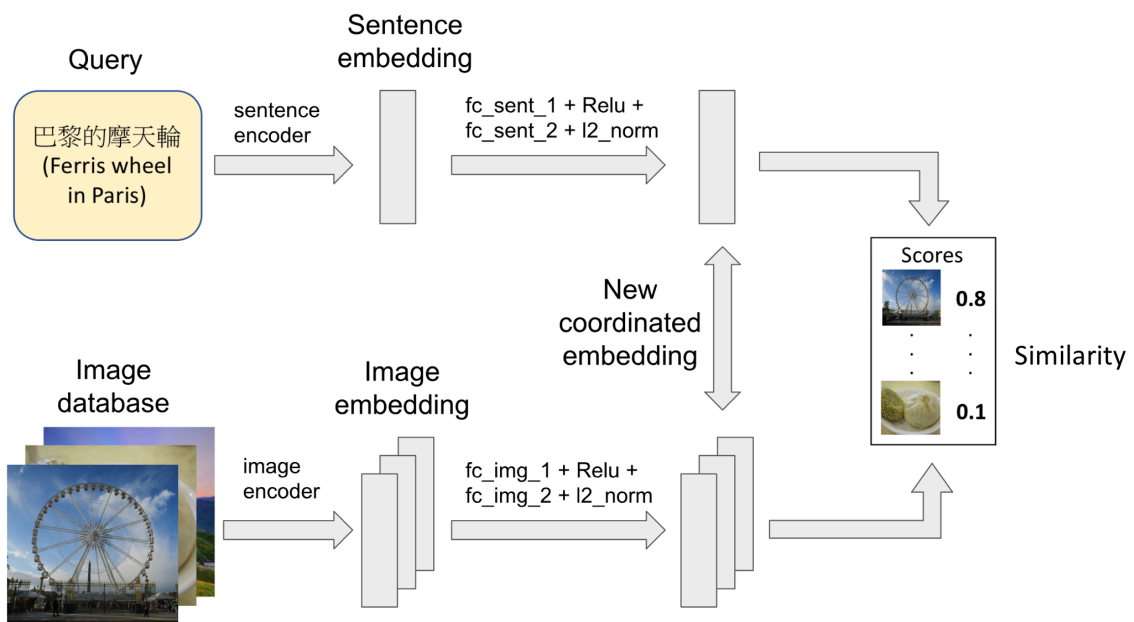
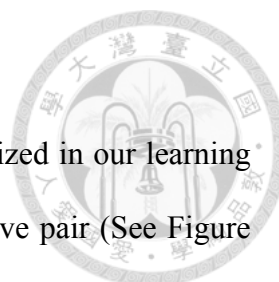


Figure 4-3 Structure of Image model when doing image recall.



## 4.2.1 Embedding Loss Function

Embedding loss function is the objective function to be minimized in our learning stage. There are four types of positive pair and four types of negative pair (See Figure 4-4). The distance of these eight types of pair will be computed for  $L_1$  to  $L_4$  (See formula (1) to (4)) from different point of view. And we will combine  $L_1$  to  $L_4$  to get final embedding loss  $L_5$  (See formula (5)). Where  $X$  is the set of images,  $Y$  is the set of sentences,  $x_i \in X$ ,  $y_i \in Y$ ,  $m$  is the margin and  $d$  is the Euclidean distance. The details of formula are discussed as follows.

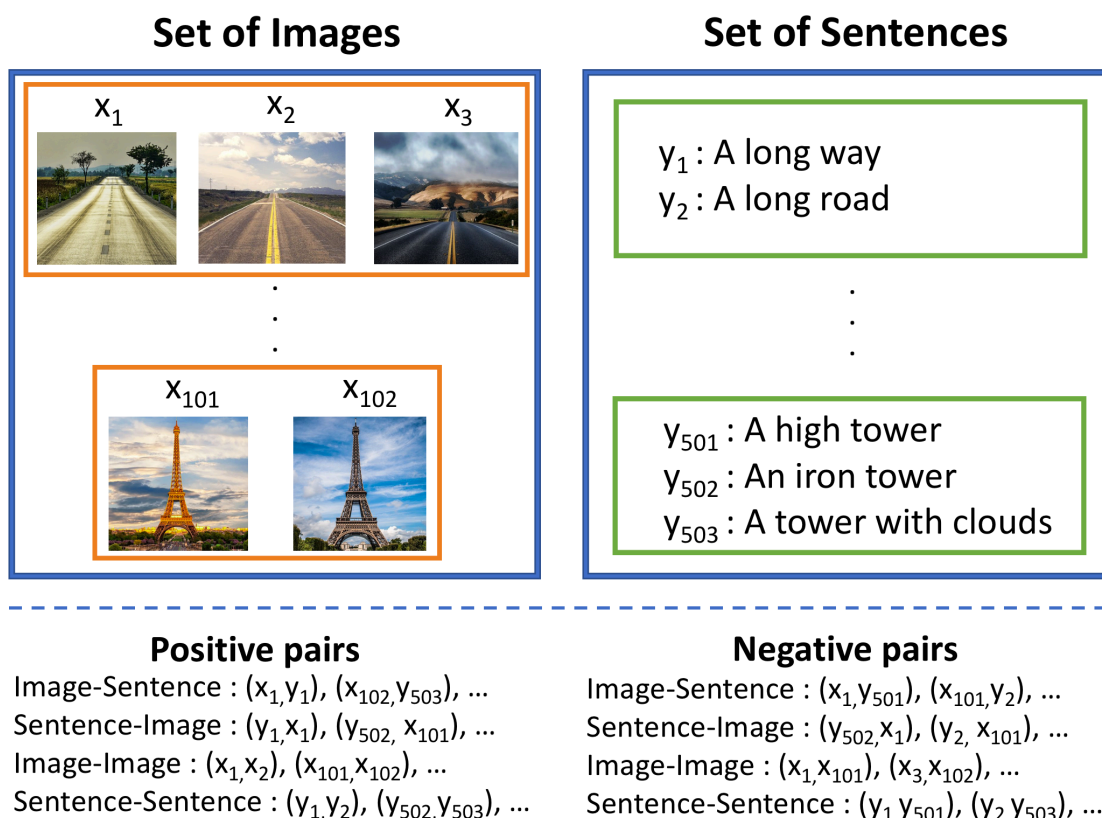
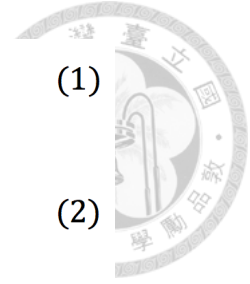


Figure 4-4 Positive pairs and negative pairs to be used in embedding loss function.



$$L_1(X, Y) = \max \left( \sum_{i,j,k} [m + d(x_i, y_j) - d(x_i, y_k)], 0 \right) \quad (1)$$

$$L_2(X, Y) = \max \left( \sum_{i,j,k} [m + d(y_i, x_j) - d(y_i, x_k)], 0 \right) \quad (2)$$

$$L_3(X, Y) = \max \left( \sum_{i,j,k} [m + d(x_i, x_j) - d(x_i, x_k)], 0 \right) \quad (3)$$

$$L_4(X, Y) = \max \left( \sum_{i,j,k} [m + d(y_i, y_j) - d(y_i, y_k)], 0 \right) \quad (4)$$

$$L_5(X, Y) = \lambda_1 L_1 + \lambda_2 L_2 + \lambda_3 L_3 + \lambda_4 L_4 \quad (5)$$

In (1),  $(x_i, y_j)$  is positive image-sentence pair (See Figure 4-4), and  $(x_i, y_k)$  is negative image-sentence pair. The embedding of the image should be close to the corresponding sentences embedding and should far away from the non-corresponding sentences embedding.

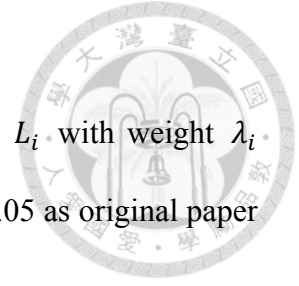
In (2),  $(y_i, x_j)$  is positive sentence-image pair, and  $(y_i, x_k)$  is negative sentence-image pair. The embedding of the sentence should be close to the corresponding images embedding and should far away from the non-corresponding images embedding.

In (3),  $(x_i, x_j)$  is positive image-image pair, and  $(x_i, x_k)$  is negative image-image pair. That is, the loss is computed only considered the relation between the images. The embedding of the image should be close to the corresponding image embedding and should far away from the non-corresponding image embedding.

In (4),  $(y_i, y_j)$  is positive sentence-sentence pair, and  $(y_i, y_k)$  is negative sentence-sentence pair. That is, the loss is computed only under the consideration of the relation between the sentences. The embedding of the sentence should be close to the corresponding sentence embedding and should far away from the non-corresponding

sentence embedding.

In (5), the final embedding loss  $L_5$  equals to sum of all loss  $L_i$  with weight  $\lambda_i$  where  $i \in \{1,2,3,4\}$ . We set  $\lambda_1=1.5$ ,  $\lambda_2=1$ ,  $\lambda_3=0$ ,  $\lambda_4=0.05$  and  $m=0.05$  as original paper setting.



#### 4.2.2 From Supervised Learning to Unsupervised Learning

Most of the image-text embedding training method is based on the supervised learning which uses the pair of corresponding image and caption as ground truth. However, TBN\_MSCOCO which is trained from that kind of dataset “MSCOCO” could not combine the information from related story and perform not well on the image recall task on Blog-travel (See Table 4-1). Therefore, we propose an unsupervised method to consider more information from related stories near the image.

<i>Baseline model</i>	<i>Food</i>	<i>Accom</i>	<i>Q1</i>	<i>Q2</i>	<i>Q3</i>
TBN_MSCOCO	<b>0.091</b>	<b>0.083</b>	0.062	0.074	0.049
Google_image_search	0.014	0.011	<b>0.106</b>	<b>0.142</b>	<b>0.221</b>

Table 4-1 The performance of TBN\_MSCOCO is apparently not as well as Google\_image\_search where TBN\_MSCOCO refers Wang et al. [14] structure.

From the statistics (See Figure 3-9), we consider the nearby sentences which are within distance 3 from an image to be the corresponding sentences of the image (See Figure 4-5). Instead of using the pairs of corresponding captions and the image (e.g., MSCOCO). We apply the pairs of nearby sentences and the image. Therefore, we do not need any caption annotation.

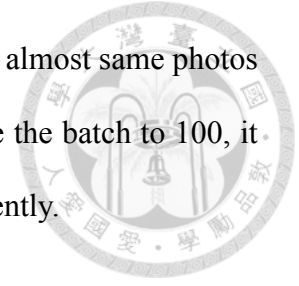


Figure 4-5 Example of the sentences which is within distance 3 (red circles) from the image is the corresponding sentences of the image.

In addition, we change the batch size of the model from 500 (original paper) to 100. The idea of original model will pick top 10 similar image as negative example. This idea works for MSCOCO dataset because we could consider all images to be independent from each other. However, this idea seems wried in the blog because many images were taken from the same person in same place. We should not consider all image are independent



from each other. That is, top 10 from 500 images will easily pick the almost same photos which should not consider as negative example. But if we downsize the batch to 100, it could improve this situation and the model can still be trained efficiently.

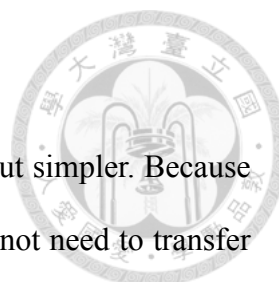


### 4.2.3 Image Encoder and Text Encoder

The original paper of L. Wang et al. [14] uses HGLMM as sentence encoder which is not very easy to train and implement on the rest of our research. Due to the purpose of this thesis focuses on finding a way to deal with image recall issue. We apply another state-of-the-art sentence encoder models to get sentence embedding. The original paper uses VGG19 as image encoder. We also apply different image encoder. Finally, we choose ResNet50 as image encoder and InferSent as sentence encoder after the comparison on MSCOCO. (Figure 4-6).

image feature	text feature	image-to-sentene			sentence-to-image		
		R@1	R@5	R@10	R@1	R@5	R@10
resnet50	HGLMM	44.2	73.0	84.8	33.2	67.1	80.5
resnet50	InferSent	41.6	70.3	83.7	28.8	62.0	76.1
resnet50	USE	34.7	65.8	79.9	26.1	58.5	73.3
vgg19	HGLMM	34.1	65.6	77.9	26.6	61.0	75.3
vgg19	InferSent	32.1	61.5	73.5	22.1	53.8	69.3
vgg19	USE	27.4	57.9	70.1	19.5	49.6	64.9

Figure 4-6 The performance of different image encoder models and sentence encoder models on MSCOCO dataset. R@K which means recall at K is the common evaluation on image-text retrieval task.



### 4.3 Story Model

The structure of story model is pretty similar to Image model but simpler. Because query and stories are encoded by the same sentence encoder, we do not need to transfer them to new embedding. The cosine similarity between query and stories can be calculated straightly (See Figure 4-7). After all stories' score are calculated, we assign each story score to the nearby images within distance  $k$ . If two or more story scores are assigned to the same image, we take the highest score for that image.

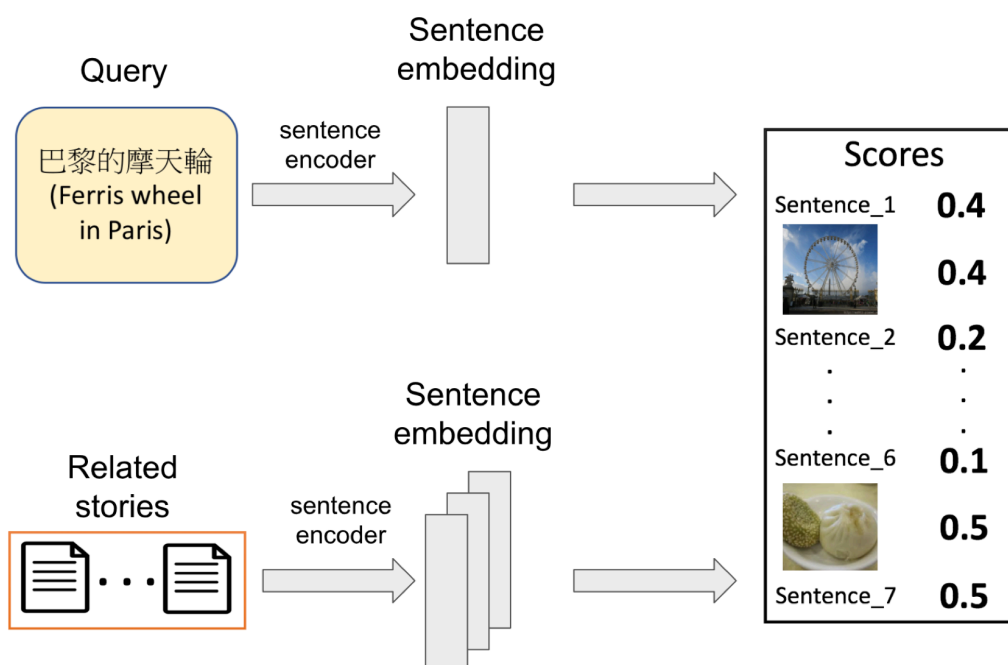


Figure 4-7 Structure of story model.

## 4.4 Image-Story Attention Model

The related stories may be close or away from the corresponding image. We could use statistic data (See Figure 3-9) to simply choose near\_3 (the sentences which are near the image within distance 3 are considered as the related stories of the image). However, different queries may benefit from different distances between the image and the related stories (See Figure 3-10). We propose an image-story attention model to combine image-story model from near\_1 to near\_9 (See Figure 4-8 and Figure 4-9). That is, the image-story attention model could determine which image-story model (near\_1 to near\_9) should take more weight when a query comes.

During the training step, we apply 5-fold cross-validation to train and test the model. That is, we split authors into five groups. We take four groups as training data and leave one group as test data. We simply use one fully-connected layer to connect query embedding and the weights. Our target function is to make the average precision of image retrieval from the query as high as possible (maximum is 1). This function will let the model learn the relation between the query and how far the considered related story is from the image.

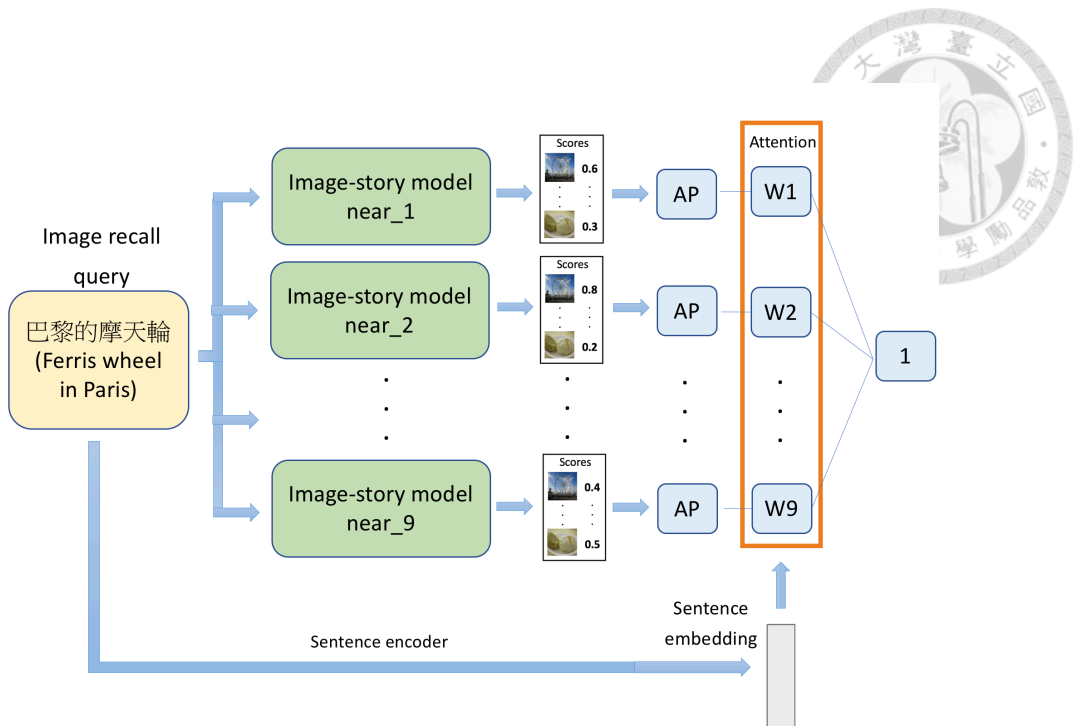


Figure 4-8 Structure of image-story attention model when training. Where APs are average precision of retrieving image by using the query.  $W_1, W_2, \dots, W_9$  are the weights which are determined by the query.

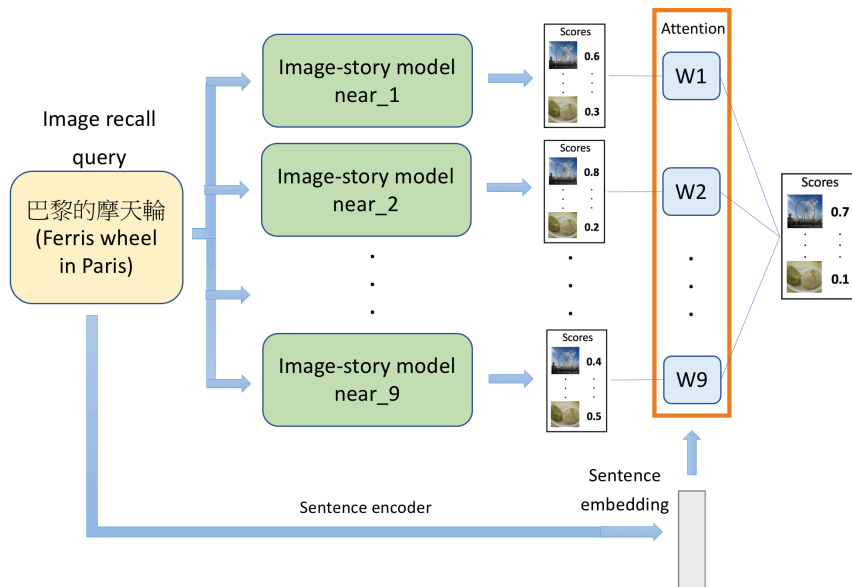


Figure 4-9 Structure of image-story attention model when doing image recall. Where  $W_1, W_2, \dots, W_9$  are the weights which is determined by the query.

## Chapter 5 Image Recall on Blog-Travel Dataset



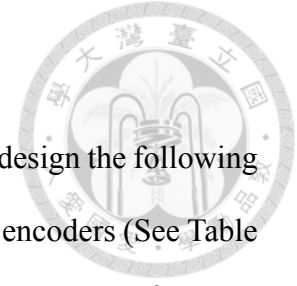
### 5.1 Experiment of Two Baseline Model

We design two baseline models “TBN\_MSCOCO” and “Google image search”, where TBN\_MSCOCO refers Wang et al. [14] structure and train on MSCOCO dataset. On the other side, “Google image search” is a strong baseline which applies Google image search function and restrict the search site on specific author blog’s website. The results (See Table 5-1) show TBN\_MSCOCO is good at food type but is not well at types Q1, Q2 and Q3. The possible reason is that the captions in MSCOCO is too simple. There are many food-related or accommodation-related captions in MSCOCO dataset but not many special captions (e.g., Eiffel tower). That may make the model does not recognize special words well.

<i>Baseline model</i>	<i>Food</i>	<i>Accom</i>	<i>Q1</i>	<i>Q2</i>	<i>Q3</i>
TBN MSCOCO	<b>0.091</b>	<b>0.083</b>	0.062	0.074	0.049
Google image search	0.014	0.011	<b>0.106</b>	<b>0.142</b>	<b>0.221</b>

Table 5-1 Baseline model results of 5 types performance on Blog-travel.

## 5.2 Experiment of Image Model



Due to the language of our dataset “Blog-travel” is Chinese, we design the following experiments to compare the performance between different sentence encoders (See Table 5-2). First three Chinese encoders do Jieba segmentation<sup>2</sup> and Gensim word2vec<sup>3</sup> which is trained on Chinese wiki to get the word vectors of the sentence. We then apply convolutional neural network (CNN), recurrent neural network (RNN) or simple word-embedding-based model (SWEM) [18] to transform the word vectors into one sentence vector. On the other side, the fourth encoder applies Google translate to translates Chinese sentence to English sentence. Next, the English sentences are encoded by InferSent sentence encoder to get the sentence vector.

The result shows that translating Chinese to English and applying InferSent outperforms other three pure Chinese encoders. The possible reason is that InferSent is a good structure and pretrained on SNLI (Stanford Natural Language Inference) which contains logical information. Compared with Chinese encoder, InferSent pretrained on high quality dataset so it could encode sentence with logical information. Besides, Google translate is quite strong and very suitable for image recall on type “Q1” (the query of Q1 is rewritten from the sentences but use other words. After translation, the query and the sentences may be very similar or even the same). The result also shows that SWEM outperforms other two Chinese encoders. The possible reason is that CNN and RNN architecture have many parameters to be trained. It is powerful but may need more high

---

<sup>2</sup> “Jieba segmentation” is an open source project to transform Chinese characters into words. See

<https://github.com/fxsjy/jieba>

<sup>3</sup> <https://radimrehurek.com/gensim/models/word2vec.html>

quality data to train.

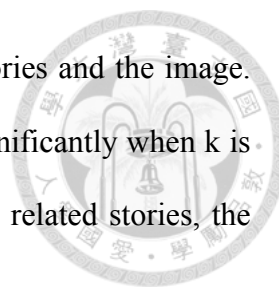


<i>Sentence encoder</i>	<i>Food</i>	<i>Accom</i>	<i>Q1</i>	<i>Q2</i>	<i>Q3</i>
Seg + w2v + CNN	0.034	0.042	0.023	0.045	0.076
Seg + w2v + RNN	0.006	0.015	0.044	0.094	0.112
Seg + w2v + SWEM	0.021	0.042	0.096	0.158	0.169
Google translate + InferSent	<b>0.087</b>	<b>0.104</b>	<b>0.175</b>	<b>0.171</b>	<b>0.181</b>

Table 5-2 Comparison between different sentence encoders on image model. Metric is NMAP@10. Seg: jieba segmentation. w2v: genism word2vec which is pretrained on Chinese wiki. All of four models are near\_3.

<i>Image model</i>	<i>Food</i>	<i>Accom</i>	<i>Q1</i>	<i>Q2</i>	<i>Q3</i>
Near_1	0.041	0.075	0.137	0.135	0.148
Near_2	0.070	0.095	0.149	<b>0.189</b>	0.158
Near_3	0.087	0.104	<b>0.175</b>	0.171	0.181
Near_4	<b>0.111</b>	0.119	0.134	0.155	<b>0.200</b>
Near_5	0.073	<b>0.141</b>	0.099	0.145	0.177
Near_6	0.037	0.063	0.079	0.087	0.083
Near_7	0.029	0.139	0.074	0.083	0.071
Near_8	0.047	0.015	0.070	0.059	0.074
Near_9	0.060	0.013	0.062	0.075	0.065

Table 5-3 Comparison between different distance between the considered related stories and the image. Metric is NMAP@10. Near\_k: distance between sentence and image to do the unsupervised learning. Model is Google translate + InferSent with near\_k.



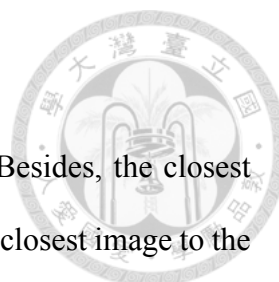
We also compare the different distance between the related stories and the image. Table 5-3 show that the performance of near\_k models will drop significantly when k is bigger than 5. That means if too many sentences are considered as related stories, the training data will be too noisy.

We try to use more data with higher diversity to train the Image model with higher generality. Table 5-4 shows the result of Blog-travel-large is similar to TBN\_MSCOCO but better than it. The performance on type “food” and “accommodation” is better when the model is trained on Blog-travel-large than that on Blog-travel. But the performance on types “Q1”, “Q2” and “Q3” is worse when the model is trained on Blog-travel-large than that on Blog-travel. The possible reason for these results is that type “food” and “accommodation” needs more generality, but type “Q1”, “Q2” and “Q3” with training data which contains higher density of related images and stories will be better in training. Another possible reason is that the dimension of new image-text coordinated embedding is 512 according to the original structure, but this dimension may not be suitable enough for more general case.

<i>Image model</i>	<i>Food</i>	<i>Accom</i>	<i>Q1</i>	<i>Q2</i>	<i>Q3</i>
Seg + w2v + swem(Blog-travel)	0.021	0.042	<b>0.096</b>	<b>0.158</b>	<b>0.169</b>
Seg + w2v + swem(Blog-travel-large)	<b>0.040</b>	<b>0.129</b>	0.064	0.081	0.113

Table 5-4 Comparison between different training data. Both model use near\_3.





### 5.3 Experiment of Story Model

The relevant stories are not always close to the target image. Besides, the closest relevant sentence may not contain all important information. And the closest image to the relevant sentence may not be the right image. Therefore, we design the following experiments to compare different types of query and different distance between the considered relevant sentence and target image (See Table 5-5).

<i>Story model</i>	<i>Food</i>	<i>Accom</i>	<i>Q1</i>	<i>Q2</i>	<i>Q3</i>
Near_1	0.057	0.047	0.238	0.181	0.093
Near_2	0.043	0.047	<b>0.305</b>	<b>0.220</b>	0.124
Near_3	0.045	0.058	0.280	0.216	0.166
Near_4	0.042	0.084	0.297	0.188	0.166
Near_5	<b>0.059</b>	0.037	0.251	0.177	0.183
Near_6	0.037	0.070	0.222	0.197	0.193
Near_7	0.025	0.088	0.217	0.184	0.193
Near_8	0.022	0.118	0.208	0.193	<b>0.204</b>
Near_9	0.014	<b>0.129</b>	0.195	0.162	0.191

Table 5-5 Comparison between different distance on story model. Metric is NMAP@10.

Near\_k: distance between sentence and image to be assigned score.

The result shows the performance of types “Q1” and “Q2” is better when the considered distance is 2 or 3. But the result shows the performance of type “Q3” may take longer considered distance to get better performance. There is a possible reason from Figure 5-1 (true distance between image and all relevant sentences). The most of cases of

distance of type “Q1” or “Q2” are shorter than 3, but there are many cases of distance of type “Q3” are longer than 3. For different types of questions, the distance between the most important sentence and target image may be quite different. Actually, “Q3” is the most important part of memory recall. But it is not very easy to get the relevant sentences for the target image.

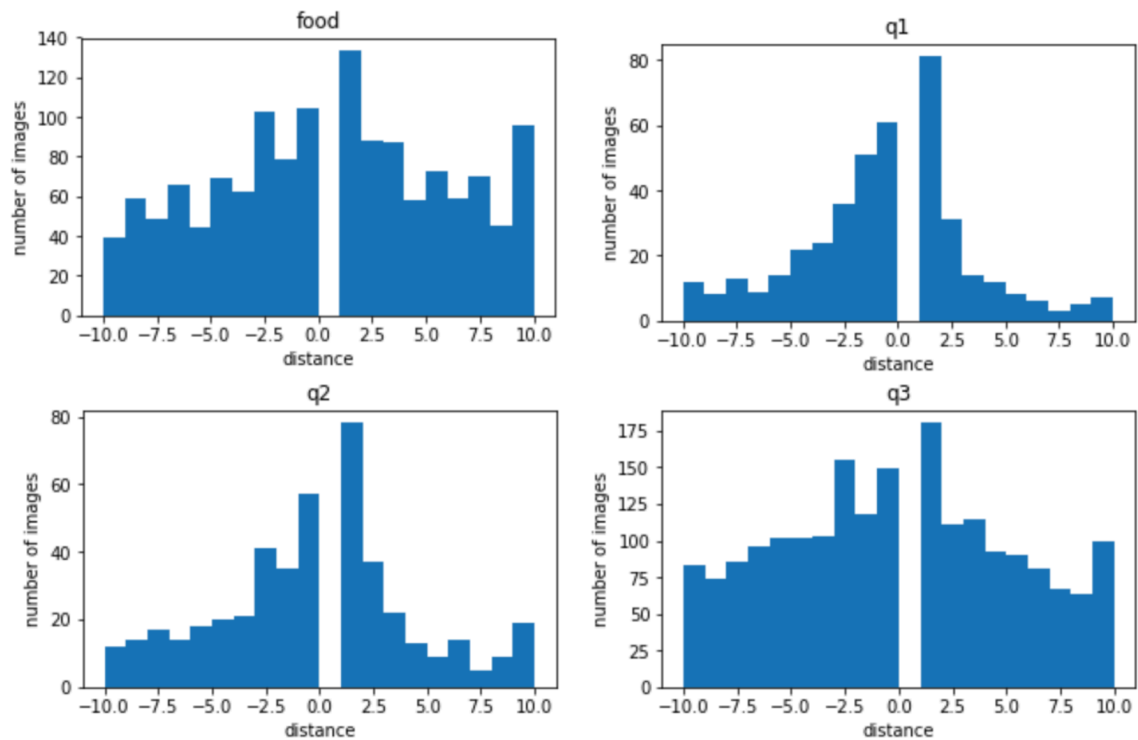


Figure 5-1 Distance between image and all relevant sentences in annotations

## 5.4 Experiment of Image-Story Attention Model

Table 5-6 shows that different query type will benefit from different image-story near\_k model. Our proposed image-story attention model could take different 9 weights for image-story near\_1 to near\_9 models. The combination of 9 models will get even better performance in the most cases.

<i>Image-Story model</i>	<i>Food</i>	<i>Accom</i>	<i>Q1</i>	<i>Q2</i>	<i>Q3</i>
Near_1	0.081	0.073	0.231	0.214	0.149
Near_2	<b>0.122</b>	0.100	0.297	0.236	0.177
Near_3	0.101	0.153	<b>0.302</b>	<b>0.256</b>	0.228
Near_4	0.101	0.159	0.241	0.214	<b>0.246</b>
Near_5	0.085	0.154	0.227	0.216	0.223
Near_6	0.047	0.088	0.188	0.165	0.170
Near_7	0.039	<b>0.189</b>	0.204	0.164	0.168
Near_8	0.036	0.120	0.179	0.148	0.166
Near_9	0.059	0.146	0.143	0.166	0.187
Attention	<b>0.123</b>	<b>0.170</b>	<b>0.329</b>	<b>0.319</b>	<b>0.268</b>

Table 5-6 Results of near\_1 to near\_9 and attention model. Metric is NMAP@10.

There are two examples of our attention model given two different queries. Table 5-7 shows the query “The food I ate” gives near\_2 model more weight. On the other hand, Table 5-8 shows the query “Ferris wheel in Paris.” gives near\_5 more weight. Both of these queries give lower weight to near\_6 to near\_9.

Query	The food I ate								
Weight	Near 1	Near 2	Near 3	Near 4	Near 5	Near 6	Near 7	Near 8	Near 9
	0.05	<b>0.29</b>	0.11	0.26	0.13	0.07	0.03	0.02	0.04

Table 5-7 The weights of 9 models for the query “The food I ate.”

Query	Ferris wheel in Paris								
Weight	Near 1	Near 2	Near 3	Near 4	Near 5	Near 6	Near 7	Near 8	Near 9
	0.11	0.06	0.21	0.12	<b>0.33</b>	0.07	0.04	0.02	0.04

Table 5-8 The weights of 9 models for the query “Ferris wheel in Paris.”

## 5.5 Results


The performance of the two baseline models and our four proposed models are shown below (See Table 5-9 and Table 5-10).

<i>Baseline model</i>	<i>Food</i>	<i>Accom</i>	<i>Q1</i>	<i>Q2</i>	<i>Q3</i>
TBN_MSCOCO	<b>0.091</b>	<b>0.083</b>	0.062	0.074	0.049
Google image search	0.014	0.011	<b>0.106</b>	<b>0.142</b>	<b>0.221</b>

Table 5-9 Performance of the baseline models for the 5 types of queries on Blog-travel dataset.

<i>Proposed model</i>	<i>Food</i>	<i>Accom</i>	<i>Q1</i>	<i>Q2</i>	<i>Q3</i>
Image model near_3	0.087	0.104	0.175	0.171	0.181
Story model near_3	0.045	0.058	0.280	0.216	0.166
Image-story model near_3	0.101	0.153	0.302	0.256	0.228
Image-story attention model	<b>0.123</b>	<b>0.170</b>	<b>0.329</b>	<b>0.319</b>	<b>0.268</b>

Table 5-10 Performance of the proposed models for the 5 types of queries on Blog-travel dataset.



The results show all of our proposed models outperform Google\_image\_search over four types of query (food, accommodation, Q1 and Q2). After combining Image model and Story model, the performance of our Image-story model is increased apparently. Our proposed Image-story model outperforms two baseline models over all types of queries. That means the information of image model and story model is complementary. Moreover, our Image-story attention model could further improve the performance.

We also plot the MAP@10 scores for all 30 authors in the Blog-travel dataset to check whether the NMAP@10 is good or not (See Figure 5-2). Due to typesetting, we only plot four most important types of query (food, Q1, Q2 and Q3). The plots “food”, “Q1” and “Q2” show our proposed Image-story model outperforms Google\_image\_search in most of cases. And the plot “Q3” shows the performance of our model and Google\_image\_search is almost the same. Therefore, we consider NMAP@10 is a suitable metric for this kind of task.

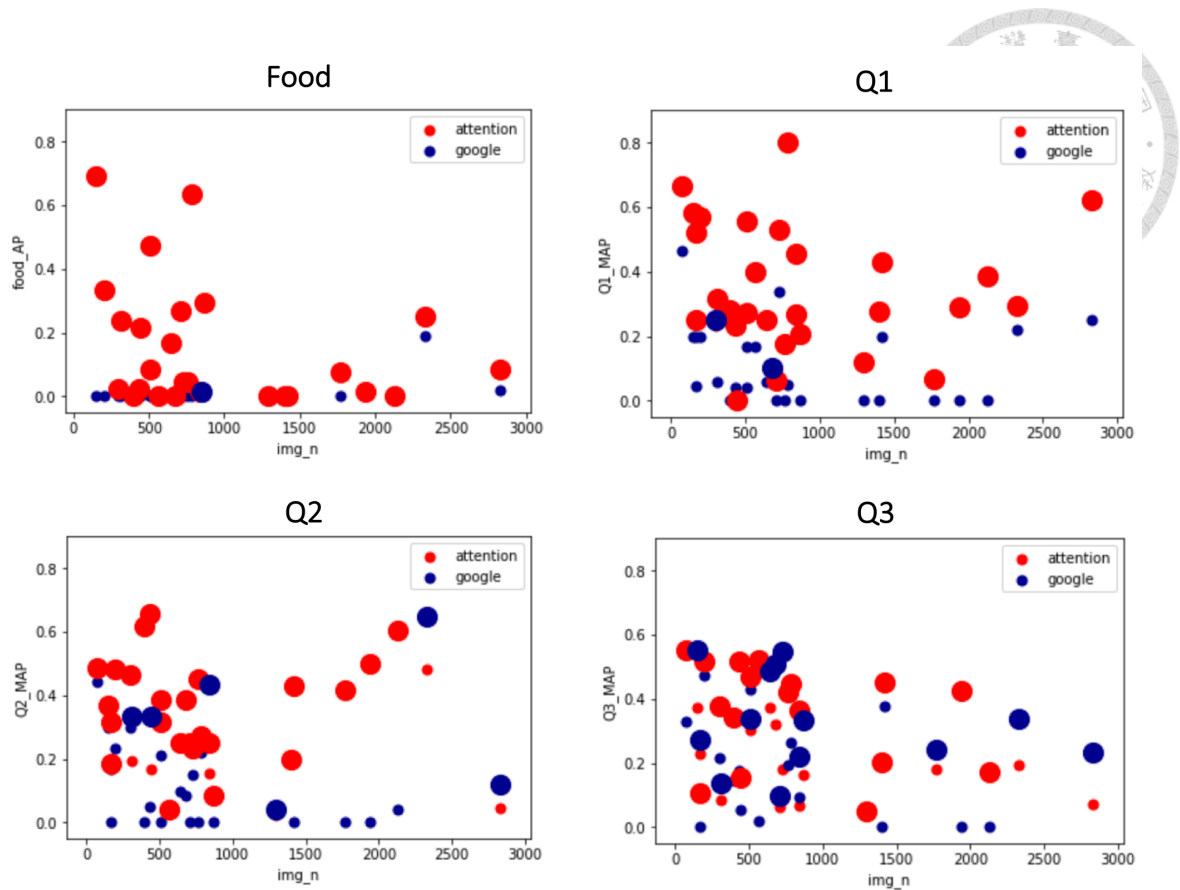


Figure 5-2 Performance comparison between the proposed Image-story attention model and Google image search on the 30 authors. Each plot contains 30 red spots and 30 blue spots. If a vertical line contains blue spot, the line should contain red spot as well. Where X-axis is total number of images of the author. Y-axis is the MAP@10 score. The big spot means the best performance model of the author.

There is a big performance difference between our model and Google image search of type “food”. From the searching results (See Figure 5-3 and Figure 5-4), our proposed model is more correct and reasonable.



Figure 5-3 The result of Google\_image\_search of type “food” on the author “altheawoman”.



Figure 5-4 The result of our proposed Image-story model of query type “food” on the author “altheawoman”.

We compare our proposed four models by plotting 30 authors as well (See Figure

5-5). The performance of Image-story model is better than the average score of other two models in most of cases. That means there exists complementarity between Image model and Story model. We get better performance through combining these two models. Moreover, our proposed image-story attention model could get even better performance than the other three models in all types of queries.

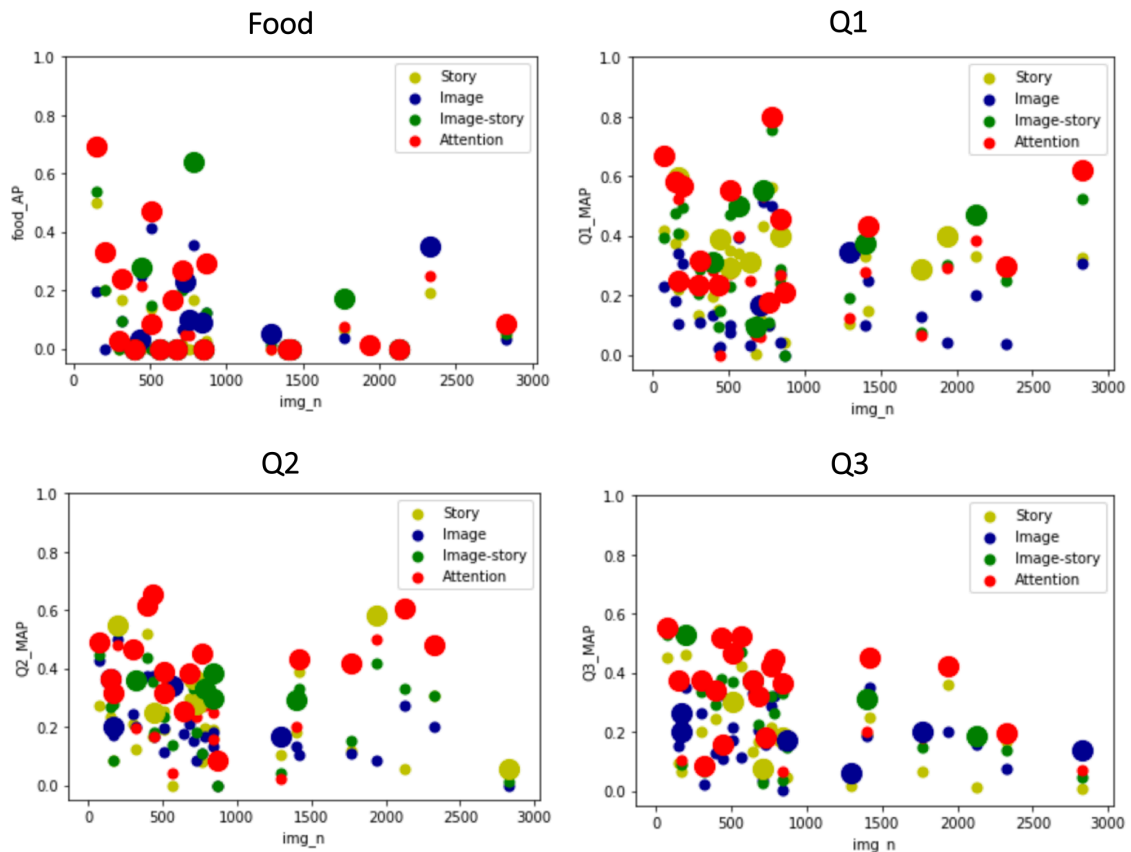


Figure 5-5 Performance comparison between our proposed four models on 30 authors. Each plot contains 30 red spots and 30 blue spots. If a vertical line contains blue spot, the line should contain red spot as well. Where X-axis is total number of images of the author. Y-axis is the MAP@10 score. The big spot means the best performance model of the author.



Some searching results show the difference between our four models of query type “Q1”, “Q2” and “Q3” (See Figure 5-6 to Figure 5-12). The Image model is strong at searching images which meet the query meaning. On the other hand, the Story model is strong at searching stories which meet the query meaning. Even though the result of Story model often gives us some non-reasonable images, it could get some images which are the person really want but very hard to be found by the Image model. As the result, we could get better performance after combining two models. Our image-story attention model could usually get even better result. Note that some images seem to be correct, but those images are not the target images of the annotators.

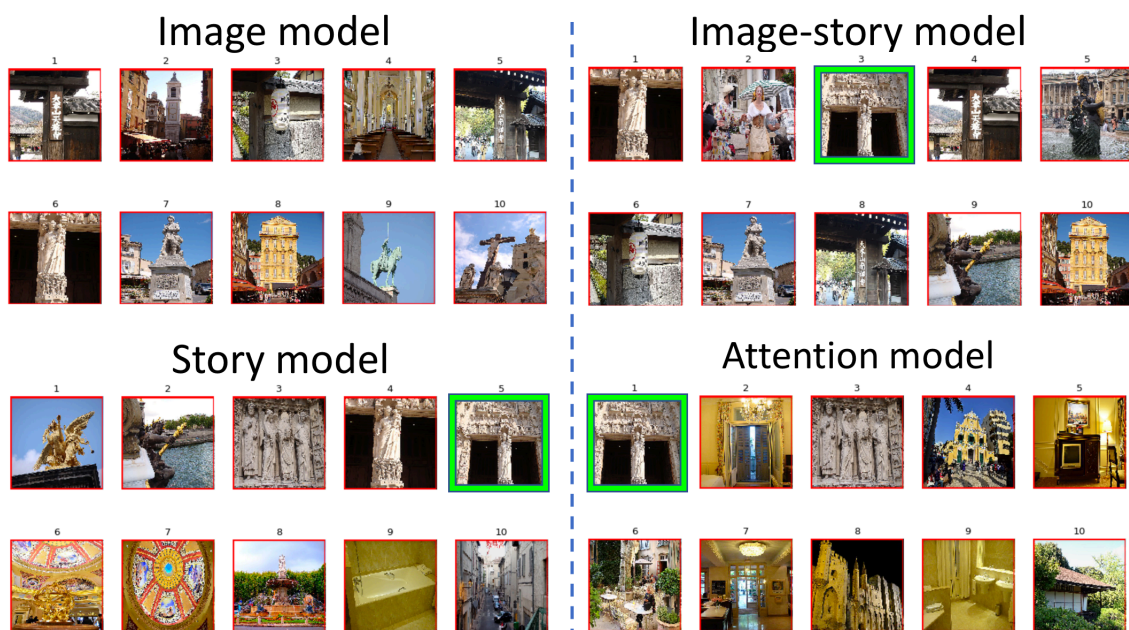


Figure 5-6 The searching result of query type “Q1”. Query is “前門各個栩栩如生的雕像彷彿帶我們回到聖經裡的場景(The vivid statues of the front doors seem to bring us back to the scene in the Bible)”.

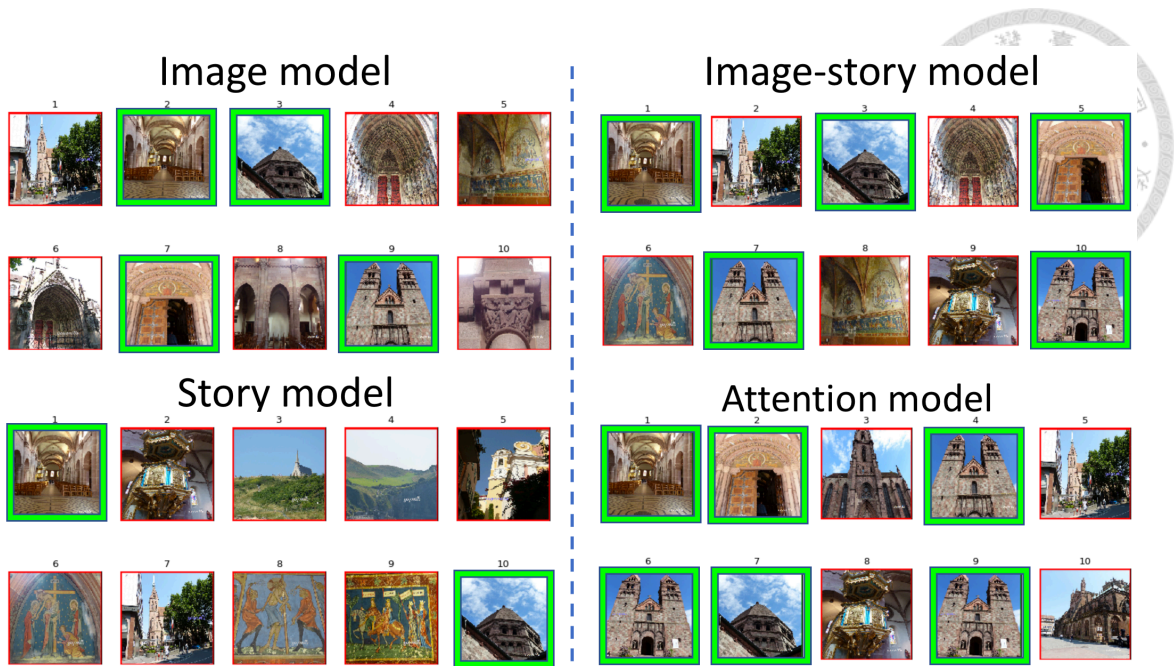


Figure 5-7 The searching result of query type “Q1”. Query is “年歲久遠而且很獨特的教堂 (Old and very unique church)”.



Figure 5-8 The searching result of query type “Q2”. Query is “點了星巴克的三明治和咖啡(Starbucks sandwiches and coffee)”.



Figure 5-9 The searching result of query type “Q2”. Query is “入場券上有蒙娜麗莎的微笑(Mona Lisa smile on the ticket)”.



Figure 5-10 The searching result of image-story model from partial queries.



Figure 5-11 The searching result of query type "Q3". Query is "法國家常料理體驗 (French home cooking experience)".

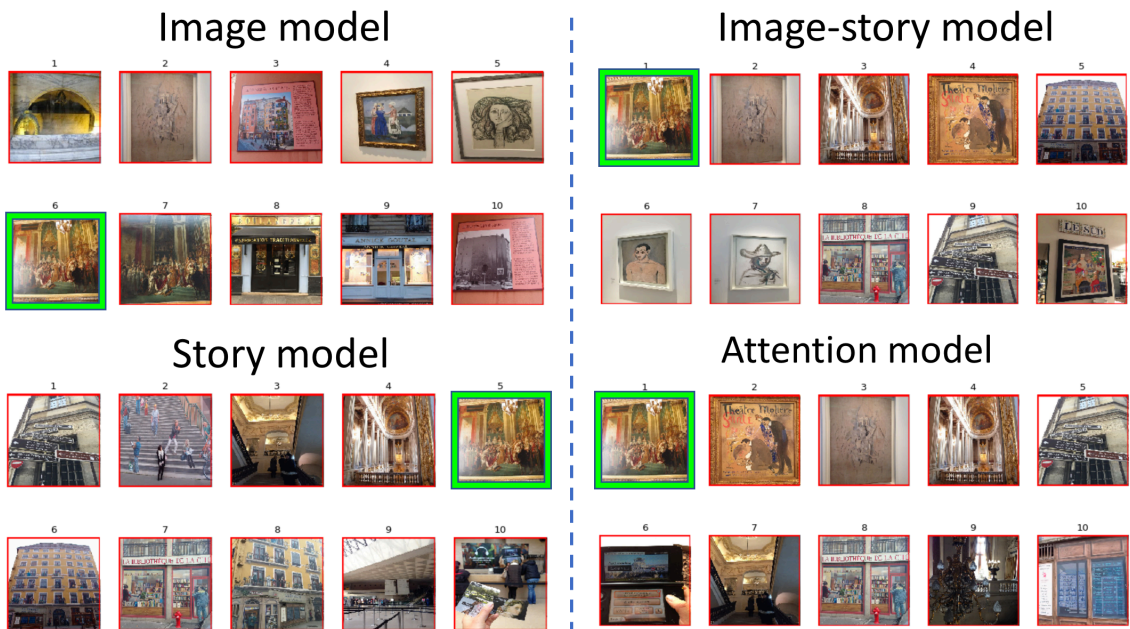


Figure 5-12 The searching result of query type "Q3". Query is "凡爾賽宮中的名畫 (Famous paintings in Versailles)".

## Chapter 6 Conclusion and Future Work



Human memory is composed of many abstract scenes. An efficient and popular way to record memory is through writing image-text intertwined blogs. This thesis builds an image recall dataset “Blog-travel” with 5 types of queries and proposes four models to do the image recall task. Our Image model and Story model are strong at different aspects and complementary to each other. As the result, combining these two model is a reasonable way. Our image-story model outperforms Google image search on Blog-travel image recall task. Moreover, our image-story attention model could further improve the performance.

To move forward on this task, there are some points of view to improve the performance:

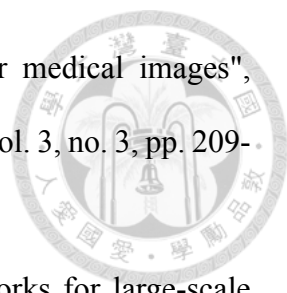
- Build a stronger image-text embedding model structure which contains higher dimensional image-text coordinated embedding to keep more information.
- Build stronger sentence encoder.
- Build stronger image encoder.
- Find a good way to indicate where is the key story for the target image.

## REFERENCE



- [1] K. Juneja, A. Verma, S. Goel, S. Goel, "A survey on recent image indexing and retrieval techniques for low-level feature extraction in CBIR systems", *Proc. IEEE Int. Conf. Comput. Intell. Commun. Technol.*, pp. 67-72, 2015.
- [2] S. Gandhani, N. Singhal "Content Based Image Retrieval: Survey and comparison of CBIR system based on Combined features", *International Journal of Signal Processing, Image Processing and Pattern Recognition*, 2015
- [3] A. Kr. Yadav, R. Roy, Vaishali and A. Praveen Kumar, "Survey on Content-based Image Retrieval Texture Analysis with Applications", *International Journal of Signal Processing Image Processing and Pattern Recognition*, vol. 7, 2014.
- [4] J. Yue, Z. Li, L. Liu and Z. Fu. "Content-based image retrieval using color and texture fused features", *Mathematical and Computer Modelling*, vol. 54, no. 3, pp. 1121-1127, 2011.
- [5] X. Y. Wang, Y. J. Yu and H. Y. Yang, "An effective image retrieval scheme using color, texture and shape features", *Computer Standards & Interfaces*, vol. 33, pp. 59-68, 2011
- [6] R. M. Haralick, K. Shanmugam and I. Dinstein, "Texture features for image classification", *Systems, Man and Cybernetics, IEEE Transactions on*, SMC-8, no. 6, pp. 610–621, 1973
- [7] A. Afifi, A. "Image Retrieval Based on Content Using Color Feature", (Doctoral dissertation, Master dissertation, Computer Engineering Department, Islamic University of Gaza, Palestine), 2011
- [8] Ramamurthy, B., and K. R. Chandran, "CBMIR: shape-based image retrieval using

canny edge detection and k-means clustering algorithms for medical images",  
International Journal of Engineering Science and Technology, vol. 3, no. 3, pp. 209-  
212, 2011

- 
- [9] Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition, arXiv:1409.1556 [cs.CV], 2014
- [10] He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR, 2016
- [11] Ryan Kiros, Yukun Zhu, Ruslan R Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. Skip-thought vectors. NIPS, 2015.
- [12] Alexis Conneau, Douwe Kiela, Holger Schwenk, Loic Barrault, and Antoine Bordes. Supervised learning of universal sentence representations from natural language inference data. EMNLP, pp. 670-680, 2017
- [13] D. Cer, Y. Yang, S. Kong, N. Hua, N. Limtiaco, R. S. John, N. Constant, M. Guajardo-Cespedes, S. Yuan, C. Tar, Y. Sung, B. Strope, and R. Kurzweil. Universal sentence encoder. arXiv:1803.11175, 2018
- [14] Zhedong Zheng, Liang Zheng, Michael Garrett, Yi Yang and Yi-Dong Shen. Dual-Path Convolutional Image-Text Embedding. arXiv:1711.05535v2, 2017
- [15] Liwei Wang, Yin Li, and Svetlana Lazebnik. Learning two-branch neural networks for image-text matching tasks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018
- [16] Kiyoharu Aizawa, Datchakorn Tancharoen, Shinya Kawasaki, Toshihiko Yamasaki. Efficient Retrieval of Life Log Based on Context and Content. ACM, pp.22-32, 2004
- [17] Lu Jiang, Junwei Liang, Liangliang Cao, Yannis Kalantidis Sachin Farfade, Alexander Hauptmann. MemexQA: Visual Memex Question Answering.

arXiv:1708.01336v1, 2017

- [18] Dinghan Shen, Guoyin Wang, Wenlin Wang, Martin Renqiang Min, Qinliang Su, Yizhe Zhang, Chunyuan Li, Ricardo Henao, Lawrence Carin. Baseline Needs More Love: On Simple Word-Embedding-Based Models and Associated Pooling Mechanisms. ACL, 2018

