# Image Recall on Image-Text Intertwined Lifelogs

Tzu-Hsuan Chu
thchu@nlg.csie.ntu.edu.tw
National Taiwan University
Taipei City, Taiwan

Hen-Hsen Huang
hhhuang@nccu.edu.tw
Natinal Chengchi University
MOST Joint Research Center for AI Technology and All Vista Healthcare
Taipei City, Taiwan

Hsin-Hsi Chen
hhchen@ntu.edu.tw
National Taiwan University
MOST Joint Research Center for AI Technology and All Vista Healthcare
Taipei City, Taiwan

## ABSTRACT

People engage in lifelogging by taking photos with cameras and cellphones anytime anywhere and share the photos, intertwined with captions or descriptions, on social media platforms. The image-text intertwined data provides richer information for image recall. When images cannot keep the complete information, the textual information is a complement to describe the life experiences under the photos. This work proposes a multimodal retrieval model for image recall in image-text intertwined lifelogs. Our Attentive Image-Story model combines an Image model, which transfers visual information and textual information to a single representation space, and a Story model, which captures text-based contextual information, with an attention mechanism to reduce the semantic gap between visual and textual information. Experimental results show our model outperforms a state-of-the-art image-based retrieval model and the image/text hybrid system.

## CCS CONCEPTS

• **Information systems** → **Multimedia and multimodal retrieval**; • **Computing methodologies** → *Natural language processing*; *Image representations.*

## KEYWORDS

lifelogging, image retrieval, multimodal representation

**ACM Reference Format:**
Tzu-Hsuan Chu, Hen-Hsen Huang, and Hsin-Hsi Chen. 2019. Image Recall on Image-Text Intertwined Lifelogs. In *IEEE/WIC/ACM International Conference on Web Intelligence (WI '19), October 14–17, 2019, Thessaloniki, Greece.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3350546.3352555

## 1 INTRODUCTION

Taking photos is one of the ways for logging personal life events. The advantages of image-based lifelogging are not only to record the moment of an event, but also to preserve the visual details of the scene that may not even be concerned when taking the picture. On the other hand, some kind of crucial information could still be missing from the visual media such as the user's subjective opinion and the emotion. In this situation, text-based contextual information like captions and descriptions intertwined with the image provides an important complement.

People engage to share their life events with others on the social media platforms, or to construct an archive for memory recall in the future. As the data accumulates, information access can be an issue, even for personal data [5]. This paper focuses on image recall on the text-image intertwined data. Different from traditional image retrieval models, we propose a personalized multimodal retrieval model that maps both the visual and the textual information to a single vector space with an attempt to reduce the semantic gap between the two kinds of medias. Learning the coordinated multimodal representations is an attractive approach to text-image based retrieval. Different from usual text-image embeddings, our multimodal representations are trained in a three-level hierarchical fashion for personalized image retrieval. That is, the representations are pre-trained with the existing image and text data, separately. Then, both of the image and the text embeddings are coordinated with a global, personal-independent multimodal data. Finally, the embeddings are further fine-tuned with personalized data for each individual user. In this way, the training strategy results in a multimodal representation for image recall by exploiting the information from the large-scale general data to the small-amount of specific, personal data. The contributions of this work are threefold.

- We investigate the problem of image recall on image-text intertwined data, which is a practical and common scenario in real world lifelogging.
- We propose a dedicate method, the Attentive Image-Story model, that incorporates an Image model and a Story model with an attention mechanism to capture the relationship between an image and an textual query.
- We propose a novel framework in which our Attentive Image-Story model is trained with a hierarchical strategy for leveraging both large general data and small personal data.

## 2 RELATED WORK

As content-based image retrieval (CBIR) becomes important and popular, how to extract features from an image is a key point. Color, shape and texture are important features for image retrieval and many researches focus on those features extraction and analysis [1, 6, 11, 14, 16]. Datasets such as MSCOCO [10] and Flickr30k [15] contain images and their corresponding captions, on which an image retrieval model can be evaluated.

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


*WI '19, October 14–17, 2019, Thessaloniki, Greece*

ACM ISBN 978-1-4503-6934-3/19/10…$15.00
https://doi.org/10.1145/3350546.3352555

text representation. VGGNet [12] and ResNet [7] encode an image into an image embedding and achieve good performances on the ImageNet classification and related tasks. Skip-thought vector [9], InferSent [4] and USE [3] could represent sentences in a single vector as sentence embedding which is strong in many tasks.

In order to achieve better performances on multimodal tasks like image-text retrieval, learning the coordinated representation embedding of multimodal data attracts a lot of attention. A previous study uses the dual-path convolutional structure to learn the image-text coordinated embedding [17]. However, the structure assumes every image is one class and to do the classification problem. This method seems not reasonable for the dataset that contains many similar images. Another work uses the structure that is called two-branch neural networks to learn the image-text coordinated embedding, achieving the state-of-the-art performance on some Image-text matching tasks [13]. However, this method is based on pairs of images and captions for supervised learning. In the image-text intertwined lifelogging data, no explicit image/caption pairs are available. For this reason, we refer to two-branch neural networks structure but we apply a method which uses the stories near the image as the corresponding textual data, instead of the captions.

## 3 LIFELOG DATASET

Existing multimodal lifelogging datasets such as MemexQA [8] are built with paired images and captions. To the best of our knowledge, no dataset is currently publicly available for evaluating the image retrieval system on image-text intertwined lifelogs. Therefore, we build a lifelog dataset based on collection of travel blogs crawled from a social media platform. We try to imitate the lifelogger to annotate the recall query and the corresponding image answers.

### 3.1 Data Collection

From the social medial platform Pixnet (www.pixnet.net) we collect a dataset, which contains a total of 26,198 images in 1,373 travel articles from 30 authors, for supervised training and evaluation. To imitate recalling image from authors, we recruit annotators to annotate 30 to 35 questions for each author. Besides, we also collect an isolated larger dataset, which contains 345,564 images and 14,831 articles from other authors, for unsupervised model training. Numerous blog articles are written for commercial purposes especially in some domains like consumer electronics. To filter out the commercial blogs, we choose some popular tourist sites as the search seeds for data collection since the global, popular tourist sites are less likely to be advertised with local campaigns. For the supervised set, we use the extremely popular tourist site "Eiffel Tower" as the search seed; For the unsupervised set, we set a number of search seeds like tourist attractions, countries and capitals all around the world to get a more diverse collection. The statistics of resulted dataset are shown in Table 1.

### 3.2 Annotation of Image Recall

In order to imitate the action of memory recall, we recruit 13 annotators to annotate the test data. As a result, each article is annotated with 30 to 35 query-answers pairs by 6 to 7 annotators with five types of queries as follows.

| | Supervised Set | Unsupervised Set |
|---|---|---|
| Number of Bloggers | 30 | 6,550 |
| Number of Articles | 1,373 | 14,831 |
| Number of Images | 26,198 | 345,564 |
| Images per Articles | 19.08 | 23.30 |
| Number Characters | 1,333,981 | 24,718,928 |
| Characters per Articles | 971 | 1,666 |

**Table 1: Statistics of the lifelog dataset.**

**Figure 1: A sample image that is annotated as Image-text combined. The query annotated for this image is "Ancient style building". The information about "acient" is from its context ... *like being in the ancient Greek era* ..., and the information about "building" is only provided in the image.**

- **Food**: The annotator is asked to imitate the blogger to annotate the images and context of the food they eat.
- **Accommodation**: The annotators is asked to imitate the blogger to annotate the images and context of the accommodation they stayed.
- **Paraphrasing**: The annotator is asked to imitate the blogger to write a query to search the image, and the query cannot include the terms that have appeared in the original context, but the meaning of the query and the context should be similar. The annotator also ticks the reference images and sentences.
- **Image-text combined**: The annotator is asked to imitate the blogger to write a query to search the image. The query must include information from both the target image and nearby stories, as illustrated in Figure 1.[1] The annotator also ticks the reference images and sentences.
- **Most important memory**: The annotator is asked to imitate the blogger to write a query to search the image, and the query is related to the most important subject of the article. The annotator also ticks the reference images and sentences.

[1]The photo is taken from https://en.wikipedia.org/wiki/Temple_of_Bel#/media/File:Temple_of_Bel,_Palmyra_05.jpg under CC BY-SA 3.0

# 4 RETRIEVAL MODEL FOR IMAGE RECALL

An abstract memory rises in their mind when users make an image recall, and they have to transform the abstract memory into a textual query. This work proposes a multimodal model, the Attentive Image-Story model, for image recall by incorporating an Image model and a Story model. The Image model is based on unsupervised learning, which projects the image and the text to a new coordinated embedding, where the image and the nearby stories will be closer. The Story model, on the other hand, computes the cosine similarity between the query and all stories and assigns the similarity scores to nearby images. To leverage the complementary information provided by each of these two models, our Image-Story model combines both similarity scores. A novel attention mechanism is further proposed for choosing how much contextual information should be taken into account, according to the queries.

## 4.1 Image Model

The Image model refers to the structure proposed by Wang et al. (2017), which reaches the state-of-the-art performance on many image-text retrieval tasks. In the learning stage, a sentence encoder extracts textual features as sentence embedding, an image encoder extracts visual features as image embedding, and the whole model is trained to project these two embeddings into a new coordinated embedding space. We denote the similarity score between an image $x$ and a query $q$ estimated by the Image model with a window size of $h$ as $s^h_{img}(x, q)$. The window size $h$ is assigned to indicate how many neighbouring sentences preceding and following the image are taken as corresponding textual data. As a result, the Image model can perform the task of image recall by retrieving the images with high similarity scores given an input query.

## 4.2 Story Model

The Story model measures the similarity between the query and an image by computing the similarity between the query and the stories nearby the image. After the scores of all stories are calculated, we assign the score of each story to its neighbouring images within a window size of $h$. Specifically, $s^h_{sty}$ denotes the score estimated by the Story model. If two or more stories are assigned to the same image, the one with the highest score will be taken.

## 4.3 Image-Story Model

The Image-Story model combines the scores made by the Image model and the Story model. Within a window size of $h$, the similarity of an image $x$ and a query $q$ estimated by the Image-Story model is computed as follows.

$$s^h_{img,sty}(x, q) = \lambda_{img} s^h_{img}(x, q) + \lambda_{sty} s^h_{sty}(x, q) \quad (1)$$

where $\lambda_{img}$ and $\lambda_{sty}$ are the weights of $s_{img}$ and $s^h_{sty}$, respectively.

## 4.4 Attentive Image-Story Model

The related stories may be close or away from the corresponding image. For this reason, our final model, the Attentive Image-Story model, integrates multiple Image-Story models with arbitrary window sizes by employing an attention mechanism. As shown in Figure 2, the final score of an image $x$ given a query $q$ is computed as follows.

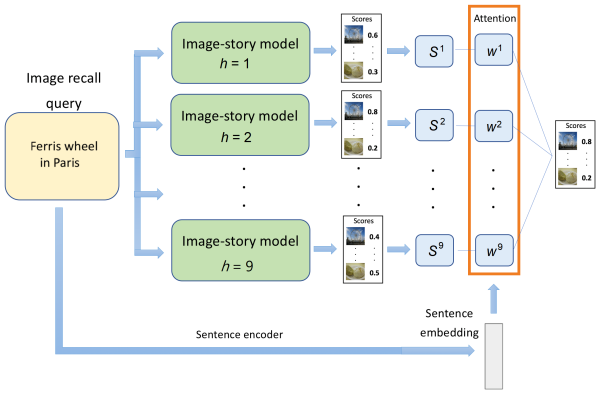


**Figure 2: Overview of the Attentive Image-Story model. $S^h$ is the scores of the images made by the Image-Story model with the window size of $h$, and $w^h$ is the weight of the window size, conditioned on the query.**

$$s_{att}(x, q) = \sum_h w^h(q) s^h_{img,sty}(x, q) \quad (2)$$

where $w^h(q)$ is the weight of the window size $h$ for the query $q$. The idea is that the attention mechanism determines the weights of the Image-Story models with different window sizes, conditioned on the input query. The attention mechanism is implemented by adding one fully-connected layer to connect the embedding of the query and the weights $w^h, h = 1...9$. The objective function is to make the average precision of image retrieval from the query as high as possible. This function will let the model learn the relation between the query and how much the contextual information should be considered as related stories for the image.

## 4.5 Hierarchical Training

We train our model with heterogeneous data in three phases. In the first phase, the large-scale multimodal dataset, MSCOCO, is employed to train the model. The sentence encoder, which is incorporated in the Story, Image-Story, and Attentive Image-Story models, is pre-trained with the Stanford Natural Language Inference (SNLI) dataset [2] in advance. Then, the unsupervised set, a collection of travel blogs, is used as in-domain data for training our model in the second phase. In the third phase, our model is fine-tuned with the individual's data for each of the 30 bloggers. In this way, the generalized information from the large-scale, out-of-domain data and the personalized information from the in-domain data are complementarily combined.

# 5 EXPERIMENTS

This section describes the experimental setup and the evaluation. Then, experiment results are shown and discussion.

## 5.1 Experimental Setup

We employ ResNet 50 as the image encoder [7] and InferSent as the sentence encoder [4]. Since the InferSent encoder pre-trained with

SNLI is for English, we translate all the textual data into English with Google Translate.[2]

For comparison, the first baseline model is TBN, which refers to the structure proposed by Wang et al. (2017) and is trained on the MSCOCO dataset. The second baseline model is Google Image Search, which is a strong baseline model performing the blog-wide image search function provided by Google.

## 5.2 Evaluation

In our three-phase training stage, the training data used in the first and the second phases are isolated from the test data. In the third phase, we apply 5-fold cross-validation to train and test the model. The data from the 30 bloggers in the supervised set are split into five folds for cross-validation.

To perform the image recall evaluation on each blogger independently, we have to compare 30 MAP (mean average precision) scores from 30 bloggers. For comparison, we design a new metric, normalized mean average precision (NMAP), as follows:

$$NMAP@k = \frac{1}{\sum_{t=1}^{a}\sqrt{l^{(t)}}}\sum_{t=1}^{a}(\sqrt{l^{(t)}}MAP@k^{(t)})$$

where $a$ is the number of bloggers, $l^{(t)}$ is the number of total images of the blogger $t$. We consider the square root of total number of images of the blogger as the weight to normalize 30 MAPs and sum all of them into just one score named NMAP@$k$.

## 5.3 Results

Table 2 shows overall performances of all models. The first column denotes the model that is evaluated. For the Image, Story, and Image-Story models, their performances with different window size $h$ are given. For instance, Image$^3$ denotes the Image model with $h = 3$. The rest of columns report the performances in the five types of queries including Food, Accommodation (Acc.), Paraphrasing (Par.), Image-text combined (Com.), and Most important memory (Imp.).

The baseline model TBN achieves a fair performance for the queries in the type of Food but performs the poorest in the types of Paraphrasing, Image-text combined, and Most important memory. A possible reason is that the captions in MSCOCO are too simple, and most of them are descriptions of concrete objects. In contrast, Google Image Search performs better in the types of Paraphrasing, Image-text combined, and Most important memory.

The Image and the Story models are comparable with the two baseline models, and the Image-story model apparently outperforms the two baseline models with a window size of 3. That means the capabilities of the Image model and the Story model are complementary. Our Attentive Image-Story model further achieves an improvement in all the five types of queries, confirming our assumption that the window size of context depends on the query.

Figure 3 shows the results retrieved by our four models for a challenging query "*The ticket with Mona Lisa*". On the target image, two small portraits of Mono Lisa with background removed are printed. The Image model ranks the target image sixth, while the Story model ranks it out of the top ten. With the textual information, the Image-Story model ranks the target image second, and our Attentive Image-Story model successfully ranks it first.

[2]http://translate.google.com

| Model | Food | Acc. | Par. | Com. | Imp. |
|---|---|---|---|---|---|
| TBN | 0.091 | 0.083 | 0.062 | 0.074 | 0.049 |
| Google | 0.014 | 0.011 | 0.106 | 0.142 | 0.221 |
| Image$^1$ | 0.041 | 0.075 | 0.137 | 0.135 | 0.148 |
| Image$^3$ | 0.087 | 0.104 | 0.175 | 0.171 | 0.181 |
| Image$^6$ | 0.037 | 0.063 | 0.079 | 0.087 | 0.083 |
| Image$^9$ | 0.060 | 0.013 | 0.062 | 0.075 | 0.065 |
| Stroy$^1$ | 0.057 | 0.047 | 0.238 | 0.181 | 0.093 |
| Story$^3$ | 0.045 | 0.058 | 0.280 | 0.216 | 0.166 |
| Story$^6$ | 0.037 | 0.070 | 0.222 | 0.197 | 0.193 |
| Story$^9$ | 0.014 | 0.129 | 0.195 | 0.162 | 0.191 |
| Image-Story$^1$ | 0.081 | 0.073 | 0.231 | 0.214 | 0.149 |
| Image-Story$^3$ | 0.101 | 0.153 | 0.302 | 0.256 | 0.228 |
| Image-Story$^6$ | 0.047 | 0.088 | 0.188 | 0.165 | 0.170 |
| Image-Story$^9$ | 0.059 | 0.146 | 0.143 | 0.166 | 0.187 |
| Att-Image-Story | **0.123** | **0.170** | **0.329** | **0.319** | **0.268** |

**Table 2: Overall performances in five types of queries. NMAP@10 is reported.**


**Figure 3: The results of the four models for the Image-text combined query "*The ticket with Mona Lisa*". The image with the green border is the correct answer.**

## 6 CONCLUSION

This work addresses the issue of personalized image recall on lifelogs. We propose a novel attentive multimodal model with an attempt to leverage the information from the large-scale general data and the small amount of personal data for reducing the semantic gap between visual and textual information. Our Attentive Image-Story model achieves the best performances in every type of queries by introducing an attention mechanism that takes the varying context window size into account.

## ACKNOWLEDGMENTS

This research was partially supported by the Ministry of Science and Technology, Taiwan, under grants MOST-106-2923-E-002-012-MY3, MOST-108-2634-F-002-008-, and MOST 108-2218-E-009-051- and by Academia Sinica, Taiwan, under grant AS-TP-107-M05.

# REFERENCES


[1] Ahmed J Afifi and Wesam M Ashour. 2012. Image retrieval based on content using color feature. *International Scholarly Research Notices* 2012 (2012).

[2] Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics.

[3] Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. 2018. Universal Sentence Encoder. *CoRR* abs/1803.11175 (2018). arXiv:1803.11175 http://arxiv.org/abs/1803.11175

[4] Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017. Supervised Learning of Universal Sentence Representations from Natural Language Inference Data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Copenhagen, Denmark, 670–680. https://www.aclweb.org/anthology/D17-1070

[5] Duc-Tien Dang-Nguyen, Luca Piras, Michael Riegler, Liting Zhou, Mathias Lux, and Cathal Gurrin. 2018. Overview of ImageCLEFlifelog 2018: Daily Living Understanding and Lifelog Moment Retrieval. In *Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum, Avignon, France, September 10-14, 2018*. http://ceur-ws.org/Vol-2125/invited_paper_3.pdf

[6] R. M. Haralick, K. Shanmugam, and I. Dinstein. 1973. Textural Features for Image Classification. *IEEE Transactions on Systems, Man, and Cybernetics* SMC-3, 6 (Nov 1973), 610–621. https://doi.org/10.1109/TSMC.1973.4309314

[7] K. He, X. Zhang, S. Ren, and J. Sun. 2016. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 770–778. https://doi.org/10.1109/CVPR.2016.90

[8] Lu Jiang, Junwei Liang, Liangliang Cao, Yannis Kalantidis, Sachin Farfade, and Alexander G. Hauptmann. 2017. MemexQA: Visual Memex Question Answering. *CoRR* abs/1708.01336 (2017). arXiv:1708.01336 http://arxiv.org/abs/1708.01336

[9] Ryan Kiros, Yukun Zhu, Ruslan R Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Skip-Thought Vectors. In *Advances in Neural Information Processing Systems 28*, C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett (Eds.). Curran Associates, Inc., 3294–3302. http://papers.nips.cc/paper/5950-skip-thought-vectors.pdf

[10] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: Common Objects in Context. *CoRR* abs/1405.0312 (2014). arXiv:1405.0312 http://arxiv.org/abs/1405.0312

[11] B Ramamurthy and KR Chandran. 2011. CBMIR: shape-based image retrieval using canny edge detection and k-means clustering algorithms for medical images. *International Journal of Engineering Science and Technology* 3, 3 (2011), 209–212.

[12] Karen Simonyan and Andrew Zisserman. 2014. Very Deep Convolutional Networks for Large-Scale Image Recognition. *CoRR* abs/1409.1556 (2014). arXiv:1409.1556 http://arxiv.org/abs/1409.1556

[13] Liwei Wang, Yin Li, and Svetlana Lazebnik. 2017. Learning Two-Branch Neural Networks for Image-Text Matching Tasks. *CoRR* abs/1704.03470 (2017). arXiv:1704.03470 http://arxiv.org/abs/1704.03470

[14] Xiang-Yang Wang, Yong-Jian Yu, and Hong-Ying Yang. 2011. An Effective Image Retrieval Scheme Using Color, Texture and Shape Features. *Comput. Stand. Interfaces* 33, 1 (Jan. 2011), 59–68. https://doi.org/10.1016/j.csi.2010.03.004

[15] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics* 2 (2014), 67–78. https://www.transacl.org/ojs/index.php/tacl/article/view/229

[16] Jun Yue, Zhenbo Li, Lu Liu, and Zetian Fu. 2011. Content-based Image Retrieval Using Color and Texture Fused Features. *Math. Comput. Model.* 54, 3-4 (Aug. 2011), 1121–1127. https://doi.org/10.1016/j.mcm.2010.11.044

[17] Zhedong Zheng, Liang Zheng, Michael Garrett, Yi Yang, and Yi-Dong Shen. 2017. Dual-Path Convolutional Image-Text Embedding. *CoRR* abs/1711.05535 (2017). arXiv:1711.05535 http://arxiv.org/abs/1711.05535