Multimedia Retrieval of Historical Materials

Jieyong Zhu[1], Taichi Nishimura[1], Makoto Goto[2], Shinsuke Mori[3]
[1]Graduate School of Informatics, Kyoto University
[2]National Museum of Japanese History
[3]Academic Center for Computing and Media Studies,
Kyoto University

April 2022

1 Introduction

Historical material is a collection of history, archaeology, and folklore materials. With the rapid advancement of digitization, large-scale multimedia data of historical materials have become available on the web. As the data grows, it is difficult for researchers to study the relationship between historical images and text. Multimedia retrieval is a technique to perform retrieval tasks across multiple media. Recently, deep learning has accelerated research on natural language understanding and computer vision, with remarkable performance reported in multimedia retrieval tasks [6]. In this paper, we apply the state-of-the-art multimedia retrieval methods to Japanese historical materials and demonstrate the constructed multimedia retrieval system. Figure 1 shows an example of multimodal retrieval tasks of historical materials.

2 Multimedia Retrieval

Multimedia retrieval takes one type of media (e.g., images and texts) as the query to retrieve corresponding media of another type [4]. The key challenge of multimedia retrieval is how to convert different media data into a shared subspace, where semantically associated inputs are mapped to similar locations. Various kinds of deep-learning-based approaches have been proposed in the literature [7]. We here employ one of them to realize our system.

3 Proposal

This paper proposes a deep-learning-based approach to achieve multimedia retrieval of historical materials. Figure 2 shows an overview of our proposed model. The proposed method consists of two major processes.

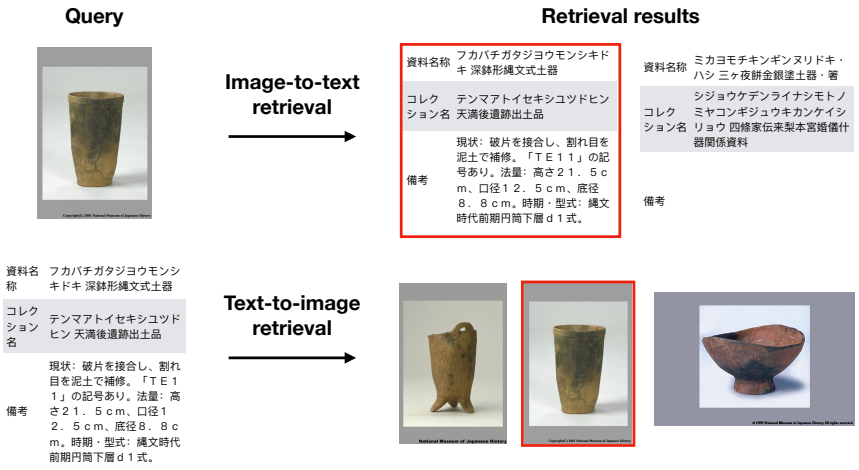

Figure 1: Examples of image-to-text and text-to-image retrieval tasks. Retrieved results in the red boxes are the associated ones with the query in the left.

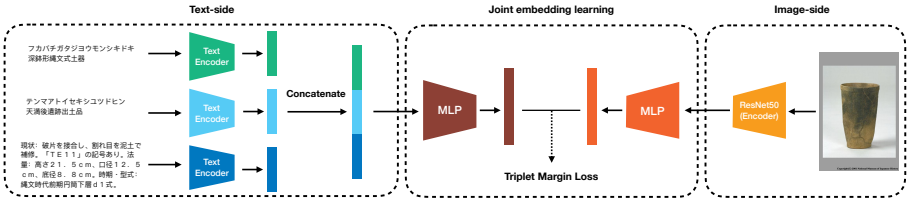

Figure 2: An overview of our proposed model.

Text encoder. Recently, large-scale pre-trained model, such as BERT (Bidirectional Encoder Representations from Transformers) [2], has achieved great performance in NLP tasks. However, we don't use BERT model because of the domain gap, since Japanese BERT is trained on Japanese Wikipedia texts. The text data of historical materials include three main types: material name, collection name, and notes. All the three types of data are in a tabular format instead of a sentence format. Therefore, we use a word2vec model to convert the texts into vectors, which is more simple and more reasonable.

Image encoder. The input of the image side is a single image of historical materials. To convert images into vectors, we employ ResNet50 [3], a Convolutional Neural Network pre-trained on ImageNet.

Shared subspace learning. Finally, we convert text/image vectors into shared subspace using symmetric multi-layer perceptrons with ReLU activation functions. To train the model, we compute triplet margin loss [1], which makes the vectors in the subspace for a given text-image pair close and otherwise long.

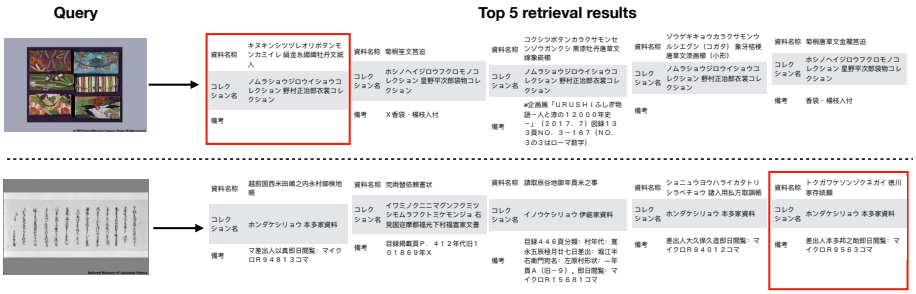

Figure 3: Image-to-text retrieval examples. The ground truth in the retrieved results is highlighted in the red box.

Table 1: Retrieval results on 1,000 samples.

	Image-to-Text	Text-to-Image	Random Ranking
R@1	0.036	0.043	0.001
R@5	0.144	0.163	0.005
R@10	0.258	0.285	0.01
medR	26	26	500
mAP	0.107	0.119	0.002

4 Dataset

This is the first attempt to tackle multimedia retrieval of historical materials, so no datasets exist in this field; thus we created the Japanese historical dataset of textual descriptions and corresponding images by crawling them from the National Museum of Japanese History. The dataset contains 18,429 objects, including over 18k textual descriptions and over 79k corresponding images.

5 Experiments

To measure the performance of the model, we perform multimedia retrieval tasks. Figure 3 shows two examples of the image-to-text task. The query images are on the left side while the top five retrieved texts are on the right side. As with previous studies, we compute three mainstream evaluation metrics, median rank (MedR), Recall@K (R@K) [6], and mean average precision (mAP) [5] to evaluate the performance. Table 1 shows the results of 1,000 samples. The result indicates that our system performs well in multimedia retrieval tasks compared with the random ranking baseline.

6 Conclusion

This paper tackled the multimedia retrieval of historical materials using deep-learning-based multimedia retrieval methods. This work is the first attempt to

tackle this problem, thus we constructed the dataset of Japanese historical texts and images, and evaluated the model's performance on it. The experimental results show that our constructed system performs well in the multimedia retrieval of historical materials. Future work will study a better method to represent the textual data. We expect that our research will help researchers in gaining a better understanding of Japanese historical materials, and will give a general approach to learning the shared subspace between textual and visual data.

References

[1] Vassileios Balntas et al. "Learning local feature descriptors with triplets and shallow convolutional neural networks." In: *Bmvc*. Vol. 1. 2. 2016, p. 3.

[2] Jacob Devlin et al. "Bert: Pre-training of deep bidirectional transformers for language understanding". In: *arXiv preprint arXiv:1810.04805* (2018).

[3] Kaiming He et al. "Deep residual learning for image recognition". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778.

[4] Jing Liu, Changsheng Xu, and Hanqing Lu. "Cross-media retrieval: state-of-the-art and open issues". In: *International Journal of Multimedia Intelligence and Security* 1.1 (2010), pp. 33–52.

[5] Nikhil Rasiwasia et al. "A new approach to cross-modal multimedia retrieval". In: *Proceedings of the 18th ACM international conference on Multimedia*. 2010, pp. 251–260.

[6] Amaia Salvador et al. "Learning cross-modal embeddings for cooking recipes and food images". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 3020–3028.

[7] Chairath Sirirattanapol et al. "Deep image retrieval applied on kotenseki ancient japanese literature". In: *2017 IEEE International Symposium on Multimedia (ISM)*. IEEE. 2017, pp. 495–499.