Cross-modal 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
{zjsczjy04,taichitary}@gmail.com
m-goto@rekihaku.ac.jp
forest@i.kyoto-u.ac.jp

Abstract

In this paper, we propose a neural-network-based cross-modal retrieval method on historical materials. We begin by collecting a multi-modal historical dataset from National Museum of Japanese History[1]. The dataset includes over 18k textual descriptions and 79k images. To evaluate the performance of our methods, we perform cross-modal image-to-text and text-to-image retrieval tasks. The experimental results show that the proposed method performs well in both retrieval tasks on historical materials compared with the random baseline.

1 Introduction

With the rapid advancement of digitization, large-scale multi-modal data of historical materials, such as images and texts, have become available on the web. Consequently, cross-modal retrieval of historical materials plays an important role in assisting researchers to study them. Cross-modal retrieval is a technique to perform retrieval tasks across multiple modalities, such as text-to-image and vice versa. In recent years, the development in the field of vision-and-language has accelerated research on cross-modal retrieval tasks with remarkable performance. In this paper, we apply the state-of-the-art cross-modal retrieval methods to historical materials and demonstrate the cross-modal retrieval system.

For historical materials, there are several kinds of corresponding textual data, including the material name, collection name, the designation, the quantity, the material quality, the scale, the production date, the place of use, the material id, and the notes. In this paper, we choose the material name, the collection name, and the notes as the three main textual features because we consider that these features are the most important to represent the historical material.

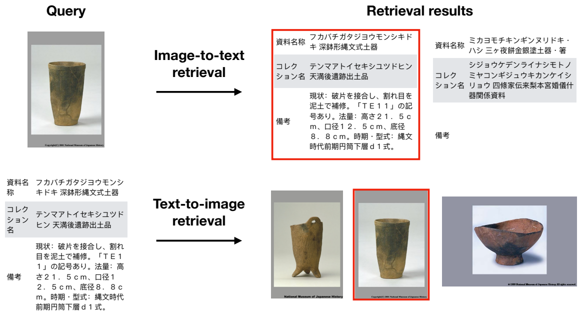
Query
Retrieval results
Image-to-text retrieval
Text-to-image retrieval
資料名称 フカバチガタジヨウモンシキドキ 深鉢形縄文式土器
コレクション名 テンマアトイセキシユツドヒン 天満後遺跡出土品
備考 現状：破片を接合し、割れ目を泥土で補修。「ＴＥ１１」の記号あり。法量：高さ２１．５ｃｍ、口径１２．５ｃｍ、底径８．８ｃｍ。時期・型式：縄文時代前期円筒下層ｄ１式。
資料名称 ミカヨモチキンギンヌリドキ・ハシ 三ヶ夜餅金銀塗土器・箸
コレクション名 シジョウケデンライナシモトノミヤコンギジュウキカンケイシリョウ 四條家伝来梨本宮婚儀什器関係資料
備考
National Museum of Japanese History

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 1 shows an overview of image-to-text and text-to-image cross-modal retrieval tasks of historical materials. The left side is the query and the right side is the retrieved results. In the image-to-text retrieval task, given an image of a historical material, we retrieve the relative texts. In the text-to-image retrieval task, given a text that describes a historical material, we retrieve the corresponding image. Note that the ground truth of the query is marked in the red box.

2 Related Work

The main challenge of cross-modal retrieval is the modality gap, and the key solution is to generate new representations from different modalities in the shared subspace, such that semantically associated inputs are mapped to similar locations.

[1] https://www.rekihaku.ac.jp/index.html (accessed on 2021/12/28).

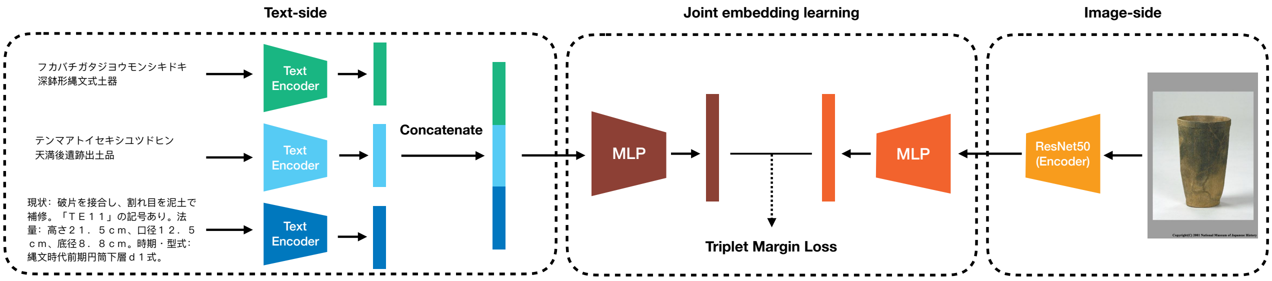
Text-side
Joint embedding learning
Image-side
フカバチガタジヨウモンシキドキ
深鉢形縄文式土器
テンマアトイセキシユツドヒン
天満後遺跡出土品
現状：破片を接合し、割れ目を泥土で補修。「ＴＥ１１」の記号あり。法量：高さ２１．５ｃｍ、口径１２．５ｃｍ、底径８．８ｃｍ。時期・型式：縄文時代前期円筒下層ｄ１式。
Text Encoder
Text Encoder
Text Encoder
Concatenate
MLP
MLP
Triplet Margin Loss
ResNet50 (Encoder)

Figure 2 An overview of our proposed model.

Subspace learning methods are one kind of the most widely used methods. They aim to find a commonly shared subspace where the similarity between different modalities can be measured. For example, Canonical Correlation Analysis (CCA) [1] is a linear method that learns the common space by maximizing pairwise correlations between two sets of data from different sources. However, because the correlation of multimedia data in the domain of historical materials is too complicated to be fully modeled only by applying linear projections, directly applying CCA over high-level text and image representations cannot achieve a reasonable result.

Inspired by the success of deep networks, a variety of deep-learning-based cross-modal retrieval methods have been developed. Deep Canonical Correlation Analysis (DCCA) [2] is a deep-learning-based method to learn complex nonlinear projections. Wang et al. [3] propose a CNN-based model to map the textual and image data to a shared subspace. Recipe1M [4] proposed a neural embedding model with semantic regularization on a recipe dataset to get a better understanding of food and recipe. Garcia et al. [5] compare a CCA model with deep-learning-based approaches to perform retrieval tasks on the domain of art paintings. In this paper, we adopt a deep-learning-based cross-modal retrieval method on historical materials, which is different from previous works.

3 Retrieval System

Figure 2 shows an overview of our proposed model. The proposed method consists of two major processes. Firstly, texts and images are encoded into representations separately. Secondly, the representations are fed into symmetric multi-layer perceptrons (MLPs) with ReLU activation functions to learn the cross-modal embeddings.

3.1 Text Encoder

In the input of the text side, we have three types of data: material name, collection name, and notes. We propose two different encoders for converting the text data to vectors: the word2vec-based model and the LSTM-based model. The output of the text encoder is a 2048-D vector.

Word2vec Encoder. We use a mean word2vec vector to represent the textual data. To train a word-level word2vec model, we tokenize all texts in the dataset using KyTea[2]. The dimension of the word2vec embedding is set as 100. Because the number of words is different for each type of text data, we use the mean word2vec of words to represent each type of data and concatenate them as a 300-D vector. Then we feed the 300-D vector into a simple fully-connected neural network to get a 2048-D vector as the output.

Bi-LSTM Encoder. We build a bidirectional LSTM model to convert the texts to vectors. The bidirectional LSTM model considers both forward and backward orderings. For each type of data, we train a different bidirectional LSTM model. The outputs of three LSTM models are concatenated as a 300-D vector. As before, the 300-D vector is fed into a simple fully-connected neural network to get a 2048-D vector as the output.

3.2 Image Encoder

The input of the image-side is a single image of historical materials. To convert images into vectors, we employ ResNet50 [6] pre-trained on ImageNet [7]. We remove the last fully-connected layer of ResNet50 and use the rest network to convert an image to a 2048-D vector. Like before, the 2048-D vector is fed into a simple fully-connected neu-

[2] http://www.phontron.com/kytea/ (accessed on 2021/12/28).

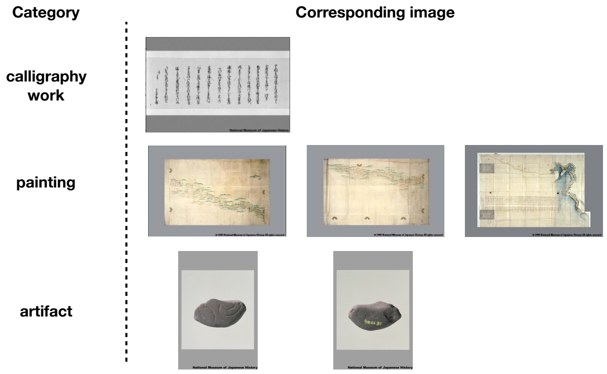

Figure 3 Examples of the dataset.

ral network. As a result, the output of the image encoder is a 2048-D vector.

3.3 Shared Subspace Learning

Finally, we convert text/image vectors into shared subspace using symmetric multi-layer perceptrons (MLPs) with the ReLU activation functions. To train the model, we compute triplet margin loss [8], which makes the vectors in the subspace for a given text-image pair close and otherwise faraway. The triplet margin loss creates a criterion that measures the triplet loss given a triplet and a margin. A triplet is composed of an anchor vector a, a positive vector p, and a negative vector n. The triplet margin loss is formulated as:

$$L(a, p, n) = \max\left\{d(a, p) - d(a, n) + \text{margin}, 0\right\}, \quad (1)$$

where $d(\cdot)$ is the Euclidean distance between the two vectors, and margin is the hyper-parameter.

4 Experiments

To evaluate our methods, we implement the models and perform both image and text retrieval tasks, and measure the performance on our dataset. We also report some samples of the retrieval tasks.

4.1 Experimental Settings

Dataset. This is the first attempt to tackle cross-modal 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 historical materials, including paintings, calligraphy works, and artifacts. As Figure 3 shows, for each historical material, there is one textual description and more than one image. In total, the dataset includes over 18k textual descriptions and over 79k corresponding images. The details of dataset splits are shown in Table 1. The ratio of the training set is 0.7, and the ratio of the validation set and test set are both 0.15.

Table 1 Dataset splitting.

Partition	# Texts	# Images
Training	10,174	10,174
Validation	2,180	2,180
Test	2,180	2,180
Total	14,354	14,354

Hyper-parameter Settings. For the word2vec encoder, we set the embedding size to 100 and use Gensim [9] to train the word2vec model. For the bidirectional LSTM encoder, we set the embedding size to 2,048. The hidden size is 512 and the output size is 2,048. This LSTM encoder only has one layer. In the LSTM encoder, we use three different vocabularies for material names, collection names, and notes separately. For the symmetric multi-layer perceptrons (MLPs) with ReLU activation functions, the input size and the output size are both 2,048. During training, we use Adam optimizer with a learning rate of 0.001 and train 35 epochs. The size of the mini-batch is 32. We set the margin to 0.1 in the triplet loss function.

Evaluation Metrics. To evaluate the performance of our model on retrieval tasks, we computed two mainstream evaluation metrics in cross-modal retrieval tasks, Recall@K (R@K) and median rank (MedR), where R@K is the recall rate percentage of the target corresponding to a query appearing in the top K when the set of obtained images is sorted in descending order by cos similarity, and MedR is the median of the ranks of the target corresponding to each query.

4.2 Quantitative Evaluation

We report results on 1,000 image-and-text pairs randomly selected from the test set. For both retrieval tasks, We compute the MedR and R@K with K being 1, 5, and 10. Our proposed method is compared with the random ranking baseline. We can see in Table 2 that the proposed model outperforms the random ranking baseline in image-to-text and text-to-image retrieval tasks. To be specific, the word2vec encoder performs better than the bidirectional LSTM encoder, indicating that the word order is not essen-

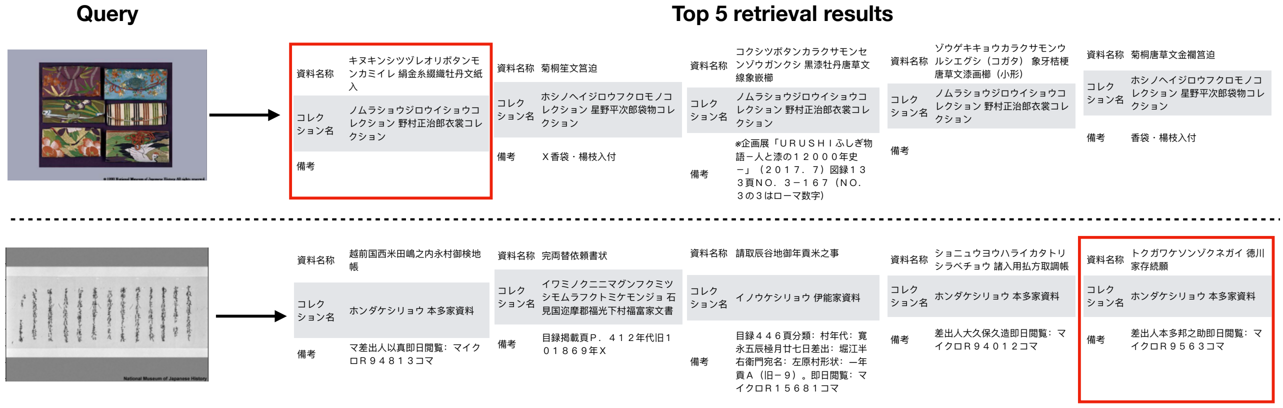
Query
Top 5 retrieval results
資料名称 キヌキンシツヅレオリボタンモンカミイレ 絹金糸綴織牡丹文紙入
コレクション名 ノムラショウジロウイショウコレクション 野村正治郎衣裳コレクション
備考
資料名称 菊桐笙文筥迫
コレクション名 ホシノヘイジロウフクロモノコレクション 星野平次郎袋物コレクション
備考 X香袋・楊枝入付
資料名称 コクシツボタンカラクサモンセンゾウガンクシ 黒漆牡丹唐草文線象嵌櫛
コレクション名 ノムラショウジロウイショウコレクション 野村正治郎衣裳コレクション
備考 ※企画展「ＵＲＵＳＨＩふしぎ物語－人と漆の１２０００年史－」（２０１７．７）図録１３３頁ＮＯ．３－１６７（ＮＯ．３の３はローマ数字）
資料名称 ゾウゲキキョウカラクサモンウルシエグシ（コガタ） 象牙桔梗唐草文漆画櫛（小形）
コレクション名 ノムラショウジロウイショウコレクション 野村正治郎衣裳コレクション
備考
資料名称 菊桐唐草文金襴筥迫
コレクション名 ホシノヘイジロウフクロモノコレクション 星野平次郎袋物コレクション
備考 香袋・楊枝入付
資料名称 越前国西米田嶋之内永村御検地帳
コレクション名 ホンダケシリョウ 本多家資料
備考 マ差出人以真即日閲覧：マイクロＲ９４８１３コマ
資料名称 完両替依頼書状
コレクション名 イワミノクニニマグンフクミツシモムラフクトミケモンジョ 石見国迩摩郡福光下村福富家文書
備考 目録掲載頁Ｐ．４１２年代旧１０１８６９年X
資料名称 請取辰谷地御年貢米之事
コレクション名 イノウケシリョウ 伊能家資料
備考 目録４４６頁分類：村年代：寛永五辰極月廿七日差出：堀江半右衛門宛名：左原村形状：一年貢Ａ（旧－９）。即日閲覧：マイクロＲ１５６８１コマ
資料名称 ショニュウヨウハライカタトリシラベチョウ 諸入用払方取調帳
コレクション名 ホンダケシリョウ 本多家資料
備考 差出人大久保久造即日閲覧：マイクロＲ９４０１２コマ
資料名称 トクガワケソンゾクネガイ 徳川家存続願
コレクション名 ホンダケシリョウ 本多家資料
備考 差出人本多邦之助即日閲覧：マイクロＲ９５６３コマ

Figure 4 Image-to-text retrieval examples. The ground truth in the retrieved results is highlighted in the red box.

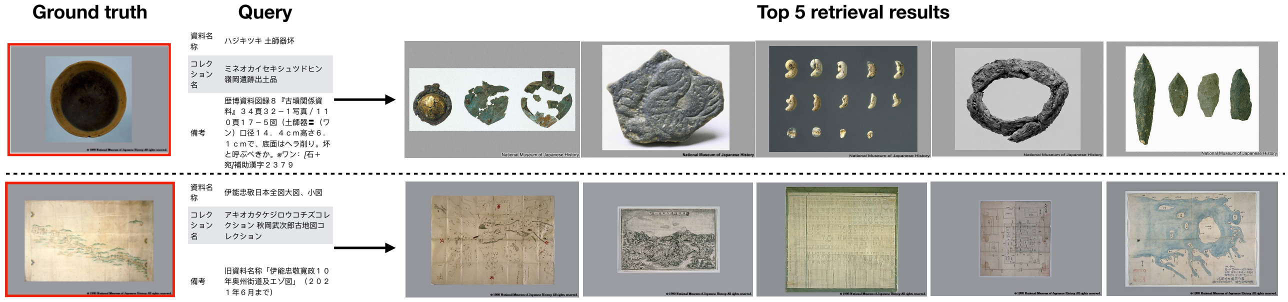
Ground truth
Query
Top 5 retrieval results
資料名称 ハジキツキ 土師器坏
コレクション名 ミネオカイセキシュツドヒン 嶺岡遺跡出土品
備考 歴博資料図録８『古墳関係資料』３４頁３２－１写真／１１０頁１７－５図（土師器〓（ワン）口径１４．４ｃｍ高さ６．１ｃｍで、底面はヘラ削り。坏と呼ぶべきか。※ワン：[石＋宛]補助漢字２３７９
資料名称 伊能忠敬日本全図大図、小図
コレクション名 アキオカタケジロウコチズコレクション 秋岡武次郎古地図コレクション
備考 旧資料名称「伊能忠敬寛政１０年奥州街道及エゾ図」（２０２１年６月まで）

Figure 5 Text-to-image retrieval examples. The ground truth is highlighted in the red box.

Table 2 Retrieval results on 1,000 samples.

	Image-to-text				Text-to-image			
	MedR (↓)	R@1 (↑)	R@5 (↑)	R@10 (↑)	MedR (↓)	R@1 (↑)	R@5 (↑)	R@10 (↑)
random ranking	500	0.1	0.5	1.0	500	0.1	0.5	1.0
bi-LSTM	26	3.6	14.4	25.8	26	4.3	16.3	28.5
mean word2vec	**19**	**4.9**	**22.3**	**36.6**	**17.5**	**4.3**	**22.9**	**36.6**

tial in our tasks.

4.3 Retrieval Results

We report image-to-text and text-to-image retrieval results on our best model, the mean word2vec model.

Image-to-Text Retrieval Results. Figure 4 shows two qualitative positive examples of the texts retrieved. On the left side is the query images. On the right side is the top five retrieved texts. In the first example, our model successfully retrieved the ground truth text in the top one. In the second example, the ground truth text is returned as the top five. We can see that our model can learn the cross-modal embeddings well and retrieve good results.

Text-To-Image Retrieval Results. Figure 5 shows two qualitative negative examples of the images retrieved. The left side shows the ground truth image and the query text. The query includes the material name, collection name, and notes. The right side shows the top five retrieved images. Although the model fails to return the ground truth image in the top five returned images, it can be observed that the retrieved images are semantically similar, which indicates that the our model can learn the semantic information in the domain of historical materials but there is still much room for improvement.

5 Conclusion

In this paper, we proposed a model for cross-modal retrieval between historical materials. This paper tackled the cross-modal retrieval of historical materials using deep-learning-based cross-modal 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 showed that the proposed method performs well in the cross-modal retrieval of historical materials.

References

[1] Yunchao Gong, Qifa Ke, Michael Isard, and Svetlana Lazebnik. A multi-view embedding space for modeling internet images, tags, and their semantics. **International journal of computer vision**, Vol. 106, No. 2, pp. 210–233, 2014.

[2] Galen Andrew, Raman Arora, Jeff Bilmes, and Karen Livescu. Deep canonical correlation analysis. In **International conference on machine learning**, pp. 1247–1255. PMLR, 2013.

[3] Jian Wang, Yonghao He, Cuicui Kang, Shiming Xiang, and Chunhong Pan. Image-text cross-modal retrieval via modality-specific feature learning. In **Proceedings of the 5th ACM on International Conference on Multimedia Retrieval**, pp. 347–354, 2015.

[4] Amaia Salvador, Nicholas Hynes, Yusuf Aytar, Javier Marin, Ferda Ofli, Ingmar Weber, and Antonio Torralba. Learning cross-modal embeddings for cooking recipes and food images. In **Proceedings of the IEEE conference on computer vision and pattern recognition**, pp. 3020–3028, 2017.

[5] Noa Garcia and George Vogiatzis. How to read paintings: semantic art understanding with multi-modal retrieval. In **Proceedings of the European Conference on Computer Vision (ECCV) Workshops**, pp. 0–0, 2018.

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In **Proceedings of the IEEE conference on computer vision and pattern recognition**, pp. 770–778, 2016.

[7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In **2009 IEEE conference on computer vision and pattern recognition**, pp. 248–255. Ieee, 2009.

[8] Daniel Ponsa Vassileios Balntas, Edgar Riba and Krystian Mikolajczyk. Learning local feature descriptors with triplets and shallow convolutional neural networks. In Edwin R. Hancock Richard C. Wilson and William A. P. Smith, editors, **Proceedings of the British Machine Vision Conference (BMVC)**, pp. 119.1–119.11. BMVA Press, September 2016.

[9] Radim Řehůřek and Petr Sojka. Software Framework for Topic Modelling with Large Corpora. In **Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks**, pp. 45–50, Valletta, Malta, May 2010. ELRA. http://is.muni.cz/publication/884893/en.