

Multimodal Artefacts: Exploring Vision-Language Models to Bridge the Modalities of Historical Written Artefacts^{*}

Hussein Mohammed^{1,*;†}

¹Cluster of Excellence: Understanding Written Artefacts, Universität Hamburg, Germany

Abstract

Historical written artefacts are multi-dimensional objects with several modalities, typically analysed separately by dedicated computational systems. These modalities are generated as research data from the study of artefacts, including digital images, measurements of material properties, and meta data from historical contexts. In most cases, these modalities are interrelated and interdependent. Therefore, understanding the relationship and learning to associate between different modalities can be essential for a holistic understanding beyond the textual contents of historical written artefacts. Recent advancements in research on multimodal models offer the possibility of analysing the different modalities of historical artefacts and modelling the relationships between them. Such models can be used by scholars for tasks such as text-based image retrieval and visual question answering. This work aims explore the potential of utilising multimodal models, and expressing the different modalities in research data of historical written artefacts in image and text formats, so that vision-language models can be employed.

Keywords

Multimodal learning, Vision-Language Models, Historical Written Artefacts

1. Introduction

Scholars in the Humanities endeavour to comprehend historical written artefacts not solely through their written content but also by scrutinising other attributes, such as visual features, material properties, and historical context. This leads to the ongoing generation of research data, comprising digital images, measurements of physical and material properties, and meta-data from historical contexts. In most cases, it is necessary to analyse multiple modalities and their interrelationships within the same artefact to achieve a comprehensive understanding that goes beyond a mere transcription of the textual contents. This includes aspects like the handwriting style, the visual layout, and

Humanities-Centred AI (CHAI), 3rd Workshop at the 46th German Conference on Artificial Intelligence, September 26, 2023, Berlin, Germany

*Corresponding author.


†These authors contributed equally.

✉ hussein.adnan.mohammed@uni-hamburg.de (H. Mohammed)

🌐 <https://www.csmc.uni-hamburg.de/about/people/mohammed.html> (H. Mohammed)

🆔 0000-0001-5020-3592 (H. Mohammed)

© 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

the spatial relations between various visual elements.

Recent research interest in deep learning-based multimodal models has been growing due to their ability to integrate various modalities (text, images, audio) for a comprehensive understanding of complex phenomena. These models find applications in fields like multimedia content analysis and visual question answering, enabling image captioning and text-based image retrieval. Possibilities for such models are vast, ranging from understanding visual scenes and recognizing activities in videos to generating language descriptions of images and facilitating real-world object recognition and interaction for robots. Vision-language models, a specific type of multimodal model, have gained significant research attention for their diverse applications, such as generating image descriptions, answering visual questions, and performing image retrieval and visual grounding tasks. Large-scale datasets with diverse image and text annotations have been created to support the development of these models. However, challenges remain in accurately integrating and reasoning over multimodal data, particularly in real-world scenarios.

Given the nature of research in the field of historical written artefacts, the types of modalities may fundamentally differ from those heavily investigated in mainstream research. Examples of such modalities include the numerical values of physical measurements and the tabular data from material analysis. Expressing these modalities in the form of text and images might result in the introduction of subjectivity and loss of accuracy in some cases. Therefore, one must be careful and should rely on the interpretation of results by experts to transform them into text and image modalities with minimum distortion. On the other hand, this approach would allow researchers to leverage the ever-growing infrastructure for vision-language models, both in terms of model architectures and dedicated training datasets. Such an approach would help remedy the limited resources, including computational power and the availability of annotated training datasets, by employing fine-tuning techniques on tasks of interest using small and specialised datasets.

The aim of this work is to discuss and explore the potential of utilising multimodal models, and the possibility of expressing different modalities in research data of historical written artefacts in both image and text formats. This approach would allow the employment of vision-language models, which can facilitate a more comprehensive and integrated analysis of such artefacts.

2. Related Work

In recent years, there has been a growing interest in the development of multimodal deep learning models [1, 2], which aim to combine different sources of data, such as text, images, videos and audio, to achieve better performance in various tasks. The success of such models relies heavily on the ability to effectively integrate and process these different modalities, which has led to the emergence of a new class of neural network architectures known as transformers [3].

Transformers have been shown to outperform traditional recurrent neural networks

in various natural language processing (NLP) tasks, including machine translation and text generation. However, their true potential lies in their ability to process multiple modalities simultaneously, making them particularly useful for multimodal learning. By leveraging the attention mechanism, which allows the network to selectively focus on different parts of the input, transformers are able to effectively fuse different modalities and extract relevant features from each [4].

Vision-language models are a special case of multimodal models that focus on processing and generating image and text data together [5]. These models have recently gained a lot of attention due to their ability to perform tasks such as image captioning and visual question answering. Early vision-language models relied on combining pre-trained image and language models, but more recent models have utilized transformer-based architectures to better fuse the two modalities [6].

In order to support the research of vision-language models, several standard public datasets and benchmark tasks have been established. The most well-known of these datasets is the COCO Captions dataset, which contains over 300,000 images and associated captions, and has been used as a benchmark for image captioning tasks [7]. Other popular datasets include Visual Genome, which contains structured image annotations, and Flickr30k Entities, which includes annotations for both image regions and entity mentions in captions [8].

Several benchmark tasks have been defined in order to evaluate the performance of vision-language models on these datasets. One such task is image captioning, where the model is given an image and asked to generate a natural language description. Another task is visual question answering (VQA), where the model is given an image and a natural language question, and is asked to generate an answer [9]. More recently, tasks that require more complex reasoning, such as visual common-sense reasoning, have also been proposed to further evaluate the capabilities of these models [10].

The currently available public datasets for multimodal learning are not relevant for the study of historical written artefacts, as they feature everyday items and their textual descriptions, such as cars, planes, and animals. Nevertheless, it is expected that trained models will acquire the ability to associate relevant information from both modalities. Therefore, tuning pre-trained models on specialised datasets with scholarly descriptions is expected to yield reasonably good results. As for the benchmark tasks, we expect two of them to be particularly directly useful, namely visual question answering and text-based image retrieval.

3. Text-based Image Retrieval with Pre-Trained Vision-Language Models

The direct relevance of text-based image retrieval to the study of written artefacts and its practical application for scholars in the Humanities make it an ideal task to test the performance of state-of-the-art pre-trained vision-and-language models. Specifically, the ViLT [11] and CLIP [12] models have been shown to perform exceptionally well on standard datasets and benchmark tasks, making them promising candidates for text-based

image retrieval. ViLT is a transformer-based model that leverages both visual and textual information to perform a variety of vision-and-language tasks, while CLIP is a model that learns to associate text with images by maximising the cosine similarity between their representations in a shared latent space. Previous studies have shown their effectiveness in tasks such as image classification and retrieval, prompting their use in this section.

Text-based image retrieval is a special case of cross-modal retrieval that requires the model to learn the relationship between different modalities and to associate the information between them. In the case of this study, the focus is on retrieving relevant images using natural language. This task is closely related to the task of visual question answering (VQA), where the difference is that VQA usually involves a single image, and the output is also in the form of textual modality, just like the question itself. The development of effective cross-modal retrieval models is of great importance for the study of historical written artefacts, where the ability to search large collections of historical documents using natural language queries can greatly enhance the efficiency of research.

A minimal test set of 100 images has been created to evaluate the retrieval performance of these two models on digitised manuscripts. The images have been selected from two different types of manuscripts to partially represent their distributions and contain some of the objects that already exist in the training set of both models. However, all objects are hand-drawn, and exist in a very different visual context in the digitised manuscripts. To construct this minimal test set, a subset of 50 images has been selected from the DocExplore dataset [13] of medieval manuscripts. See example images in Figure 1a and Figure 1b. Another 50 images has been selected from two manuscripts from *Al-Ḥarīrī, Maqāmāt*, © Paris, Bibliothèque nationale de France. Département des manuscrits, namely MS arabe 3929 and MS arabe 5847. See example images in Figure 1c and Figure 1d. This test set is made publicly available in a research data repository [14] under the Creative Commons license.

A set of 12 different text queries has been used to test these two pre-trained models. Half of the queries are about only one visual concept, such as *a bird*, *a watercraft vessel* or *people*. The other half are about multiple related visual concepts, such as *birds on a tree*, *a watercraft vessel with fish underneath* or *people with a baby*. All queries can be found in the research data repository [14]. Retrieval results in Table 1 have been calculated for both models in Top_1 , Top_3 and Top_5 images. For concepts with occurrences occ fewer than x in Top_x , the retrieval percentage is calculated by dividing the correctly retrieved images by occ . Otherwise, it is calculated by dividing the correctly retrieved images by x .

Table 1

Results of text-based image retrieval using pre-trained CLIP and ViLT models on a small test set of historical manuscripts.

Type of text query	Relevant images in Top_1		Relevant images in Top_3		Relevant images in Top_5	
	ViLT	CLIP	ViLT	CLIP	ViLT	CLIP
Single concept	50%	33.33%	55.55%	55.55%	66.66%	51.66%
Multiple concepts	16.66%	50%	58.33%	54.16%	66.66%	75%
Average	33.33%	41.66%	56.94%	54.85%	66.66%	63.33%



Figure 1: Example images from the minimal vision-language test set. The images in a and b are from the DocExplore [13] dataset, while the images in c and d are from *Al-Ḥarīrī, Maqāmāt*, © Paris, Bibliothèque nationale de France, Département des manuscrits. The description under each image is the relevant textual query used to retrieve the corresponding images.

The images used in this test set are typical manuscript pages, where the visual elements (objects) occupy only very small percentage of the image area, see Figure 1. Moreover, all objects in these images are hand-drawn in very different styles, shapes and scales. In addition, some of the typical degradation types can be seen in many of these images, such as stains, low contrast and general distortions. Given the aforementioned factors, the retrieval results in Table 1 are promising and demonstrate the ability of both models to transfer their training on very different datasets to manuscript images. Fine-tuning these models on sufficient amount of text-image pairs is expected to enhance the retrieval performance by a great margin.

On average, both models have comparable performance with the exception of Top1, where the CLIP model outperforms the ViLT model by more than 8%. Nevertheless, there was a significant difference in the retrieval time between the two models. The full results details can be found in the research data repository [will be added in Camera-Ready]. The average processing time per image for the ViLT model was 18 seconds, while it took only 4 seconds per image in the case of the CLIP model. The processing time per image is a critical factor for the task of text-based image retrieval in order to offer practical solutions.

In addition, an important difference between the two models is the maximum number of tokens in a text due to the transformer architectures used in each model. For the CLIP model, the maximum number is 512 (around 100 words), while it is only 128 (around 25 words) for the ViLT model. Having a sufficient description length is particularly important for historical manuscripts to include scholarly descriptions, which is expected to be around 50 words or more.

4. Fine-tuning CLIP VL-Model

Even without scholarly descriptions, fine-tuning these models on manuscript images can enhance the results of retrieving images from manuscript collections, which would otherwise require annotating all the important visual elements in each image. To demonstrate this, a very small dataset of text-image pairs is created from the aforementioned manuscripts and a very simple textual description of the images. This dataset is used to fine-tune the CLIP vl-model for the task of text-based image retrieval.

The images in this fine-tuning dataset have been selected specifically to contain multiple visual elements related to the visual concepts mentioned in a set of 10 different query texts. The training set consists of 100 image-text pairs, while the test set consists of a different 25 image-text pairs. The fine-tuning dataset and the query texts can be found in our research data repository [15]. Since the image-text pairs are either a correct or a wrong match, the measure of binary cross entropy between target and input logits is used as the loss function. In this mini-training dataset, every pair is a correct match; therefore, all labels are set to one (correct). As can be seen in Table 2, fine-tuning on only 100 image-text pairs using a very simple loss function for only 30 epochs has already resulted in a tangible improvement in Top2 and Top3 retrieval rates.

When four RTX 3080 GPUs have been used, the retrieval time per image was only about 0.12 seconds. This demonstrates the feasibility of building an efficient retrieval system based on VL models. Nevertheless, further research is required for the efficient implementation of such a system.

Table 2

Comparison between the average retrieval of CLIP VL-model on 25 test images without and with fine-tuning on only a 100 image-text pairs.

	Top1	Top2	Top3
Pre-trained	50.00	47.37	77.27
Fine-tuned	50.00	63.16	81.82

The inclusion of additional information such as the place and date of the document, the type of ink used, and script type can greatly enhance the accuracy of retrieval systems. These additional pieces of information can be represented as a textual modality and paired with the visual modality from manuscript images to form a comprehensive dataset for training vision-language models. As such, collaboration between experts from relevant fields of research is essential for generating accurate and relevant textual descriptions that can help us to better understand and analyse historical written artefacts. An example for such description could be: *manuscript name and culture, date, place, type of writing support, type of used ink, type of script, main visual elements*.

5. Conclusion

The development of vision-language models provides a promising opportunity to improve the analysis of historical written artefacts. However, expressing the different modalities

in historical artefacts in the form of text and images may result in subjectivity and loss of accuracy. Therefore, it is important to rely on the interpretation of results by experts to transform them into text and image modalities with minimum distortion. The proposed approach can leverage the ever-growing infrastructure for vision-language models and remedy limited resources, including computational power and the availability of annotated training datasets, by employing fine-tuning techniques on tasks of interest using small and specialised datasets.

A minimal test set of 100 images has been created to evaluate the retrieval performance of two pre-trained models, CLIP and ViLT, on digitised manuscripts. The test set contains hand-drawn objects from two different types of manuscripts, each with different visual contexts, and with many images featuring typical degradation types such as stains, low contrast, and general distortions. The results show promising retrieval performance, indicating the ability of both models to transfer their training to manuscript images. Important considerations for this task are the retrieval time per image and the description length, which refers to the maximum number of tokens in each textual description. The CLIP model demonstrated better performance with respect to both considerations.

Finally, A very small fine-tuning dataset of only 100 image-text pairs is used to fine-tune the CLIP model with a binary cross-entropy loss function for 30 epochs. This micro-tuning was already enough to enhance the retrieval results of both Top2 and Top3. As future work, a larger fine-tuning dataset with scholarly descriptions, and a more advanced contrastive loss function, will be used to train the CLIP model for the task of text-based image retrieval.

Acknowledgments

The research for this work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC 2176 ‘Understanding Written Artefacts: Material, Interaction and Transmission in Manuscript Cultures’, project no. 390893796. The research was conducted within the scope of the Centre for the Study of Manuscript Cultures (CSMC) at Universität Hamburg.

In addition, we would like to thank Martina Dinelli for preparing and annotating the fine-tuning dataset in this research.

References

- [1] J. Lu, V. Goswami, M. Rohrbach, D. Parikh, S. Lee, 12-in-1: Multi-task vision and language representation learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 10437–10446.
- [2] Y.-C. Chen, L. Li, L. Yu, A. El Kholy, F. Ahmed, Z. Gan, Y. Cheng, J. Liu, Uniter: Universal image-text representation learning, in: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Springer, 2020, pp. 104–120.

- [3] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, *Advances in neural information processing systems* 30 (2017).
- [4] L. H. Li, M. Yatskar, D. Yin, C.-J. Hsieh, K.-W. Chang, Visualbert: A simple and performant baseline for vision and language, *arXiv preprint arXiv:1908.03557* (2019).
- [5] Y. Du, Z. Liu, J. Li, W. X. Zhao, A survey of vision-language pre-trained models, *arXiv preprint arXiv:2202.10936* (2022).
- [6] W. Kim, B. Son, I. Kim, Vilt: Vision-and-language transformer without convolution or region supervision, in: *International Conference on Machine Learning*, PMLR, 2021, pp. 5583–5594.
- [7] X. Chen, H. Fang, T.-Y. Lin, R. Vedantam, S. Gupta, P. Dollár, C. L. Zitnick, Microsoft coco captions: Data collection and evaluation server, *arXiv preprint arXiv:1504.00325* (2015).
- [8] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, et al., Visual genome: Connecting language and vision using crowdsourced dense image annotations, *International journal of computer vision* 123 (2017) 32–73.
- [9] P. P. Liang, Y. Lyu, X. Fan, Z. Wu, Y. Cheng, J. Wu, L. Chen, P. Wu, M. A. Lee, Y. Zhu, et al., Multibench: Multiscale benchmarks for multimodal representation learning, *arXiv preprint arXiv:2107.07502* (2021).
- [10] R. Zellers, Y. Bisk, A. Farhadi, Y. Choi, From recognition to cognition: Visual commonsense reasoning, in: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 6720–6731.
- [11] Y.-C. Wu, H. H. Chen, M. Rohrbach, D. Parikh, S. Lee, Vilt: Vision-and-language transformer without convolution or region supervision, *arXiv preprint arXiv:2102.13106* (2021).
- [12] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al., Learning transferable visual models from natural language supervision, *arXiv preprint arXiv:2103.00020* (2021).
- [13] S. En, S. Nicolas, C. Petitjean, F. Jurie, L. Heutte, New public dataset for spotting patterns in medieval document images, *Journal of Electronic Imaging* 26 (2016) 1 – 15. URL: <https://doi.org/10.1117/1.JEI.26.1.011010>. doi:10.1117/1.JEI.26.1.011010.
- [14] H. Mohammed, Vision-language mini testset (vl-mini-test), 2023. URL: <https://doi.org/10.25592/uhhfdm.11754>. doi:10.25592/uhhfdm.11754.
- [15] H. Mohammed, Mini-dataset for VL-Models fine-tuning (VL-Tune- dataset-mini), 2023. URL: <https://doi.org/10.25592/uhhfdm.12670>. doi:10.25592/uhhfdm.12670.