Title: Exploring a large digitized collection of Belgian postcards: what machine learning models can we use and how can we evaluate them?

Authors: Margherita Fantoli, Eli Verwimp

1. Introduction

In this short paper we discuss how to carry out the exploration of a large digitized collection of Belgian postcards. The goal of the paper is twofold: in the first part, we assess the usability of CLIP (Contrastive Language–Image Pre-training, Radford et al. 2021) for such a purpose, by reporting the performance of CLIP for querying specific categories of buildings in our collection and by discussing some more abstract queries. In fact, the integration of CLIP for mining cultural heritage collections has already been shown to be particularly fruitful (Smits and Wevers 2023). In the second part, we investigate the question of what metrics can be adequate to evaluate these tools and what their usability is to guide the scholars in the exploration of the dataset. Our paper confirms that multi-modality approaches, combining both visual and textual information, represent a more performant tool than only leveraging one modality. It further suggests that the precision-recall curve provides a realistic assessment of the quality of the retrieval of buildings within large and highly diversified visual collections. This paper contributes to the theme of the conference because we reuse bibliographical metadata to carry out the evaluation of computational analyses of visual material. In addition, we discuss how evaluation metrics can be used to facilitate access to unstructured and extended visual collections.

2. Belgian postcards

The dataset of belgian historical postcards has been digitized by KU LeuvenLibraries and is browsable in Limo[1]. Of 68825 records available, 35649 are in the Public Domain.
The records are described by a structured set of metadata in which the creators of the collection have assigned predefined tags to each record. The most attested category is `Buildings', with subcategories such as 'Church' or 'Castle': we use this information as a proxy of 'gold data' for the task of image classification. In addition, the text printed on the postcard has been transcribed by the curators (see Figure 1).

[1] https://kuleuven.limo.libis.be/discovery/collectionDiscovery?vid=32KUL_KUL:KULeuven&collectionId=81531489730001488

Figure 1: Text: Mont-Kemmel. La Tour -
Kemmel-Berg. De Toren. Labels: ‘Kemmel (Heuvelland). Gebouwen. Algemeen’

3. Searching for buildings

As a first step, we evaluate the performance of CLIP for retrieving five kinds of buildings that have been assigned manual labels in the metadata. The performances are assessed based on the precision and recall (PR) curve (and the average precision metric, AP, see Raghavan et al. 1989) and Receiver Operating Characteristic (ROC) curve (and the Area Under the Curve metric, AUP, cf. Fawcett 2006) for 5 types of buildings (churches, castles, stations, palace of justice and Gravensteen). The images are ranked based on the cosine similarity of the CLIP embedding of each image with the CLIP embedding of the textual prompt. Results are reported in Table 1. In order to clarify whether such approach yields competitive results, we compare with the use of the uni-modal queries, i.e. vision-only and textual only queries. Vision-only uses the similarity between the embedding of a random image of a church, and the visual embeddings of the images of the collection for the ranking, while textual-only uses a textual prompt and the textual embeddings of the 'titles' of the postcards (details of the implementation will be provided). For both uni-modal cases we test both the CLIP embeddings and embeddings of unimodal models (ResNet, for images, and Bert, for text). The methodology is based on Smits and Kestemont 2021 and we substantially confirm the competitive performance of CLIP. We finally evaluate whether a model that is specifically trained to recognize certain classes can improve the performance over a general

representation learning based approach such as CLIP. Here, we test a vision transformer (Dosovitskiy et al. 2020) pre-trained on Imagenet. The results show that the zero-shot capabilities of the multimodal search with CLIP remain unmatched by any of the other approaches. However, we discuss some examples of more abstract queries (e.g.' oriental world'), where the limitations of the CLIP model emerge.

Prompt	Number of images	AUC	AP
Church	6907	0.906	0.633
Castle	3115	0.898	0.418
Station	688	0.874	0.178
Palais de Justice	283	0.923	0.175
Gravensteen	113	0.98	0.354

Table 1: performance of CLIP on different buildings.

The results, based on the availability of “gold data”, allow us to conclude that, when dealing with collections where such metadata are not available, running multimodal queries using CLIP would be the most efficient way to proceed. Moreover, we show the potential and limitations of running more abstract queries such as ‘political propaganda’ and ‘oriental world’.

4. Evaluating metrics:

Based on this use-case, we discuss why metrics commonly used in retrieval exercises such as precision and recall at k (where 'k' indicates a certain position in the ranking) and Mean Reciprocal Rank are uninformative in this context, by exemplifying how strongly their significance varies depending on the number of true positives for the category in the collection, which changes with every query. For this reason, we adopted the ROC and PR curves, which have the advantage of being threshold-free, i.e. display the ratio of true and false positives for the changing value of the similarity cutoff chosen to select retrieved (positive) documents. However, the AUC metric leads to a much more positive interpretation of the results than the AP. This is typical of searches featuring unbalanced classes, as demonstrated by Davis et al. 2006, Saito and Rehmsmeier 2015, Berrar and Flach 2012, and also in this case, the AP proves a better metric for the evaluation of the results, bearing in mind the practical needs of the researchers, who can only manually go through a limited amount of images when searching such collections. In addition, both curves provide a measure to set a threshold of the cosine distance to optimize the ratio between true positives and false positives in the retrieved images (here, we focus on the Youden's index for the ROC curve, and the optimal f1 score for the PR curve). We discuss how the 'optimal' thresholds variate across several queries: while the AUC-derived threshold would typically yield a fairly consistent but overwhelming ratio of true and false positives across several queries, the AP optimization results in a threshold variating greatly across the different buildings analyzed, which better reflects the changing difficulty of the query and quality of the results. Hence, we conclude that,

with this kind of collection and ranking, the PR curve and AP represent a suitable tool for assessing the quality of the retrieval.

References:

Berrar, D., & Flach, P. (2012). Caveats and pitfalls of ROC analysis in clinical microarray research (and how to avoid them). Briefings in Bioinformatics, 13(1), 83–97. https://doi.org/10.1093/bib/bbr008

Davis, J., & Goadrich, M. (2006). The relationship between Precision-Recall and ROC curves. Proceedings of the 23rd International Conference on Machine Learning, 233–240. https://doi.org/10.1145/1143844.1143874

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2020, October 2). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. International Conference on Learning Representations. https://openreview.net/forum?id=YicbFdNTTy

Fawcett, T. (2006). An introduction to ROC analysis. Pattern Recognition Letters, 27(8), 861–874. https://doi.org/10.1016/j.patrec.2005.10.010

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & I. Sutskever. (2021). Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*;.

Raghavan, V., Bollmann, P., & Jung, G. S. (1989). A critical investigation of recall and precision as measures of retrieval system performance. ACM Transactions on Information Systems, 7(3), 205–229. https://doi.org/10.1145/65943.65945

Saito, T., & Rehmsmeier, M. (2015). The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. PLOS ONE, 10(3), e0118432. https://doi.org/10.1371/journal.pone.0118432

Smits, T., & Wevers, M. (2023). A multimodal turn in Digital Humanities. Using contrastive machine learning models to explore, enrich, and analyze digital visual historical collections. *Digital Scholarship in the Humanities*, fqad008. https://doi.org/10.1093/llc/fqad008

Smits, T., & Kestemont, M. (2021). Towards multimodal computational humanities. using clip to analyze late-nineteenth century magic lantern slides". In *Proceedings of the Conference on Computational Humanities Research 2021*, 149–158, Amsterdam, the Netherlands.