Transfer learning for Renaissance illuminated manuscripts: starting a journey from classification to interpretation

Valeria Minisini[1,2,*,‡], Giorgio Gosti[2,‡] and Bruno Fanini[2]

[1]Sapienza Università di Roma, Dipartimento di Scienze dell'Antichità, Piazzale Aldo Moro 5, 00185 Rome, Italy
[2]National Research Council - Institute of Heritage Science (CNR-ISPC), Via Salaria KM 29300, 00015 Monterotondo, Italy

Abstract

In recent years illuminated manuscripts have been extensively digitized, providing an unprecedented amount of material for computer vision research and the creation of better performing neural networks. In addition to character recognition, which has long been the main application field, these new resources have allowed the adoption of machine learning to detect decoration and miniatures that usually concern only a few pages of a codex but attract great attention especially from non-academic public. The paper presents ongoing research that aims to demonstrate the possible adoption of transfer learning for digitized artworks to improve a pretrained deep neural network ability to recognize handwritten pages, identify the layout elements and, specifically, figurative miniatures in Renaissance manuscripts to be used for the creation of an immersive interface for consulting and comparing images based on their iconography. After a brief introduction to contextualize the changes brought by the massive cultural heritage digitization, we will present some of the most interesting research conducted on both manuscripts and artworks. Next, the dataset built to train the model will be described, focusing on its composition and the image classification system adopted. In conclusion, we will then expose the training strategy chosen to minimize human effort by dividing the dataset into three groups before concluding with the first results obtained so far and the prospects for future development.

Keywords

Deep Neural Network, Transfer Learning, Illuminated Manuscript, Image Classification, Layout Analysis

1. Introduction

Until the mid-15th century, illuminated manuscripts were the main vehicle for the circulation and preservation of knowledge among the Western upper classes, surviving even beyond Johannes Gutenberg's introduction of movable type printing. This typology of cultural artifacts, difficult to access due to their materials' fragility, has found new life thanks to digitization which has made available to scholars and the common public many little-known works, rarely released from deposits.

Each complete codex comprises hundreds of pages, which do not only contain text, but are also richly decorated with miniatures that share subjects and iconography with other artistic expressions like paintings and sculptures. Although illuminations occupy only a limited number of sheets, they are also the components that attract the most non-academic audience, making the often-incomprehensible written content partly accessible.

To enhance the thousands of volumes made available online we are developing an immersive visualization interface that allows users to move easily between the digitized pages, searching for figurative miniatures and capable of relating reproductions of different works with the same subject. To make this possible, a deep neural network system is being trained on a specially constructed dataset containing reproductions of Medieval and Renaissance volumes together with artworks. Our objective is to obtain a model that can recognize elements in the page layout and identify handwritten sheets within

3rd Workshop on Artificial Intelligence for Cultural Heritage (AI4CH 2024, https://ai4ch.di.unito.it/), co-located with the 23nd International Conference of the Italian Association for Artificial Intelligence (AIxIA 2024). 26-28 November 2024, Bolzano, Italy

*Corresponding author.

‡These authors contributed equally.

valeria.minisini@uniroma1.it (V. Minisini); giorgio.gosti@cnr.it (G. Gosti); bruno.fanini@cnr.it (B. Fanini)

0009-0000-6266-7757 (V. Minisini); 0000-0002-8571-1404 (G. Gosti); 0000-0003-4058-877X (B. Fanini)

heterogeneous collections of items. Given that deep neural networks require massive datasets with high-quality annotations that are labor-intensive, we implemented an Interactive Machine Learning approach to rapidly compile increasingly larger databases, in which a domain expert and an AI expert cooperate to train a model with domain-specific knowledge using transfer learning [1, 2].

Section 2 outlines the state of the art while Section 3 proceeds to present the dataset by describing its composition and image classification system before exposing the technique adopted for training the model. Finally, in Section 4 we discuss conclusions.

2. Related work

Most datasets concerning digitized handwritten and early printed texts are composed of materials merged according to a shared language, creation period and resource-distributing institution [3]. Often, they are small if compared with newspapers [4] and museum collections [5], containing frequently only a subset of pages from the volumes or few complete codices [6], such as the HBA dataset [7] with images that originate from 11 books of the French digital library Gallica. In several cases, the dataset needs to be augmented with new items and an example is [8] which uses generative adversarial networks (GANs) to create synthetic pages to train the model.

The utilization of machine learning and specifically neural networks to facilitate the study of historical documents is an established field of research [9]. The main investigated task has been text recognition [10], but the focus on graphic elements such as layout analysis [8], drop caps for letter extraction [11], figure gestures classification through template-based detectors [12] and miniature retrieval [13] have gradually shifting attention from text to artistic aspects. Also attempts with bounding boxes have been made for illumination detection [14] and iconographic recognition [15] with good results.

In the historical-artistic field, among the several neural networks architectures ResNet [16] was used individually [17] or within more complex pipelines [18], mostly, because it performs well on labeling and detection tasks and is easily trainable, given that its implementation of residual functions allows better error propagation. Often, transfer learning is used to train ResNet architectures given small datasets [18, 19].

For an in-depth review of "human-in-the-loop" approaches refer to [20]. Specifically, [1, 2] discuss Interactive Machine Learning. Label Studio is a flexible labeling tool that implements active learning [21]. As well, ilastik [22] or Cellpose [23] are machine learning tools that offer Interactive Machine Learning capabilities.

3. The Project

The collective interest in illuminated manuscripts is still lower than the attraction exerted by works of art, almost entirely coming from relatively few expert researchers due to language barriers and the complexity of interpreting handwritten sheets. In addition to the textual part, which simultaneously serves as both the main vehicle and greatest obstacle to content communication, the miniatures that captivate the viewer can be used to build accessible experiential paths based on a visual language understood regardless of one's country of origin or historical knowledge. The first step in creating an interface that will allow the user to explore the manuscript pages through its more intuitive graphical contents was to automate their recognition.

To develop the dataset required for deep neural network transfer learning, we implemented an iterative incremental training cycle, composed of four main phases. First, a domain expert annotates or corrects the labels of a smaller dataset. Second, an AI expert uses transfer learning and hyperparameter optimization to train a deep neural network. Third, the model predicts the larger dataset labels. Thus, the cycle restarts from the first phase with the domain expert correcting the annotations predicted for the larger dataset. Labels were created using Label Studio [21].

3.1. The Dataset

For training our deep neural network to identify the different visual components inside the pages and distinguish any miniatures present also from ornaments or illuminated initials, we have thus formed an original training dataset. It contains images from digitized manuscript volumes, mainly dating back to the Medieval and Renaissance periods, together with reproductions of incunabula and later printed copies but also works of art and objects normally present in digital museum catalogs often featuring figurative subjects with a sacred theme as decoration.

Files come exclusively from institutional databases like the US Library of Congress, The Metropolitan Museum of Art of New York and the J. Paul Getty Museum Collection of Los Angeles but also from European ones such as the Staatsbibliothek of Berlin and the Koninklijke Bibliotheek of the Netherlands. We preferred images from public domain or released under creative common license CC-BY. Furthermore, by using different sources we tried to obtain a certain variety of scanning and shooting conditions as well as the quality of reproductions, selecting both overall views and detail shots.

Each element was classified using an alphanumeric code that allowed to quickly trace the origin and provide information about the depicted subject, the century of production, and the specific author or, if unknown, at least its geographical area. As the number of elements increased, a prefix was also added to each series to distinguish the type of physical object digitized, the associated category and subclass. An example of the result thus obtained is "P.01_S(An)_14BMGetty" where "P.01" indicates the category and the subclass it belongs to, “S(An)” identifies the subject as Saint Anne while "14" is the century of creation, “BM” the author’s initials and “Getty” is the provenance institution. At the same time, this system has kept low the risk of inserting multiple copies of a work, especially in cases such as engravings, and associating images with the same themes already in the early stage of research to possibly increase the examples of a specific subject when poorly represented.

In total, the dataset consists of 45.798 items divided into two macro-categories, volumes (P) and art (A), and into ten subgroups: manuscript sheets (P.01), printed pages (P.02), paintings (A.01), engravings (A.02), drawings (A.03), sculptures (A.04), stained glass windows (A.05), tapestries (A.06), art prints (A.07) and other objects (A.08). Since the project is focused on manuscripts, the number of images dedicated to these is higher than all other typologies and constitute 68.25% of the total, chosen to cover a wide variety of layouts and styles. They chiefly come from books of hours, psalters, missals, breviaries, and bibles selected primarily for the presence of illustrations. The miniatures in the chosen manuscripts are both full page and inserted in the text.

Always privileging the figurative element, the sheets without illustrations have been classified according to the presence of decorations and, only in the absence of these, on text or musical scores considering the dominant component. Four summary partitions were thus obtained in the main category, first useful to balance the composition of the dataset, avoiding the insertion of too many text pages compared to the other elements, and afterwards for the subdivision into three training groups progressively larger in size of 400, 4.000 and 41.398.

3.2. The Training

The smallest group consists of 330 pages and 70 artworks, a proportion that remains constant in the second and undergoes slight variations in the third set where the category “Pages” contains a total of 34.447 items to which are added 6.951 images of the other one. In the preprocessing steps we uniformed the dataset through resizing to 1280×1280 pixels maintaining proportions.

To limit human effort in the annotation phase, we manually labeled only the first group to report the presence or co-presence in the pages of figurative miniatures (Fig), decorations (Deco), text (Text), music scores (Mus) and illuminated initials (Let) with a minimum of one and a maximum of four tags for a total of 818. This set, split into 70% training and 30% test, was used to initially train our deep neural network to distinguish between pages and art, manuscript and printed pages as to detect the layout elements.

We use a transfer learning approach based on a ResNet architecture [16] pretrained on ImageNet

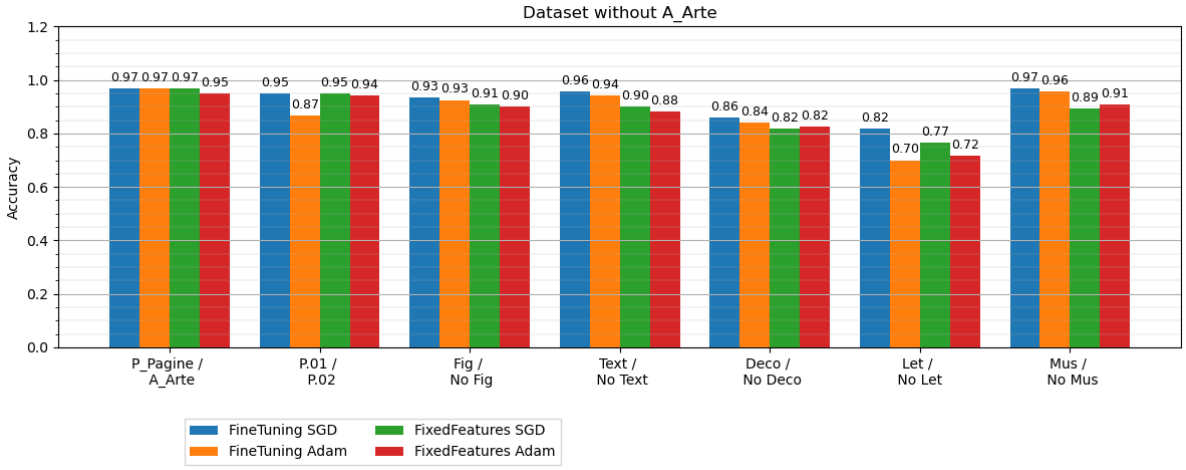

Figure 1: Prediction accuracy for target labels for the 400 images dataset.

dataset [24]. We planned to test ResNet-50 and 101 as well as ResNet-18. Since we had a relatively small dataset, we started with ResNet-18 because models with fewer parameters are less likely to be in an over-parametrization regime. We obtained unexpectedly good results for both the training and the test metrics. In future, we will test the larger networks.

In the training phase, we augmented the dataset by transforming the images passed by the dataloader through the following sequence of transformations. First, the dataloader resized the images to 256×256, then it randomly cropped and resized the pictures to 256×256, and finally, it normalized the images according to ResNet's suggested parameters. Initially a distinct group of deep neural networks was trained on the first group without the category "Art" and this allowed us to test that art images do not negatively impact the prediction accuracy.

We considered two gradient descent training methods: Stochastic Gradient Descent (SGD) and Adam. For SGD we used an exponential learning rate scheduler with a starting learning rate 0.001, momentum 0.9, learning rate scheduler step size 7, and gamma 0.1. For Adam, we settled on a learning rate 0.001.

We also compared two types of transfer learning: fine-tuning on all weights or keeping the features in the training phase while fine-tuning only the last layer.

Figure 1 shows the accuracy level achieved: in almost all cases, the fine-tuning of all weights and SGD gave the best results, even if the difference is not very large. The accuracy rate is almost always above 90%, particularly high in recognizing pages from art and detecting text or music, while the lowest value was recorded for illuminated initials, mostly caused by the inherent ambiguity between simple ornamental letters and historiated initials.

Considering the promising results, the models with the highest accuracy level for each target were used to automatically predict the labels for the 4000 images of the second set. Then, the domain expert carefully checked to identify specific model biases and corrected the assigned labels. These models demonstrated their ability to recognize with high precision the presence of decorations and illustrations on the pages, especially when there are borders and the scene occupies a large portion of space. They are however less precise in distinguishing between the miniatures inserted in the text and the illuminated initials, often confusing the two elements and labeling some of the former as "Let". They also incorrectly classified blank pages as "Text" and adopted the same tag for some sculptures together with the right "Fig".

Only in ten cases the prediction was completely wrong: out of 4.000 images only 1.549 required manual intervention, saving several hours of work to reach an overall result of 7.205 annotations. Comparing the automatically applied tags with the correct ones, the most significant decrease was observed for "Text", which went from 3.174 to 2.854, followed by "Let" reduced from 966 to 936. On the other hand, the "Fig" and "Deco" classes increased, going from 1.495 to 1.765 and from 1.214 to 1.292

respectively.

The accuracy of the predictions was overall considered satisfactory with a rate of 0.98% for the recognition of figurative miniatures, 0.88% for text, 0.86% for decorations and 0.74% for illuminated initials. However, a particular criticality has been observed with musical scores printed after the 15th century which are not recognized as such, significantly impacting the number of labels assigned by the system, only 251 out of the exact 358.

4. Conclusions

The initial recognition tests on the layout elements gave us very encouraging results as accuracy levels higher than 80% are not often obtained with such small datasets. Our approach allows for a gradual and efficient labeling process and in this paper we stop at the end of the first step of the incremental learning cycle. In future work, we plan to progressively grow the annotated items' number to verify if similar results will be maintained also for the third larger group.

Once this phase is completed, a further step will be taken to improve the dataset with more complex label tasks such as instance semantic segmentation. Specifically, we will gradually move from image classification to the detection of specific components using the activation zones as a guide to automatically place boxes and then perform the semantic segmentation of any miniatures present.

The masks thus obtained will be used to enrich an interactive and immersive web3D system based on an open-source framework [25], that will be investigated in the upcoming stages of the research. This will give the user the possibility to inspect and query large amounts of pages in a 3D space, juxtapose similar images, and comparing them with art reproductions.

In doing so, we intend to demonstrate how, through the use of new technologies, figurative language can be used to structure narratives capable of making artifacts, originally prerogative of a few, available to all by freeing a delicate cultural property from its physical and conservative limits and exploiting iconography as an interpretative system that can be understood independently of the spoken language.

References

[1] A. Holzinger, Interactive machine learning for health informatics: when do we need the human-in-the-loop?, Brain Informatics 3 (2016) 119–131. doi:10.1007/s40708-016-0042-6.

[2] R. Porter, J. Theiler, D. Hush, Interactive machine learning in data exploitation, Computing in Science and Engineering 15 (2013) 12–20. doi:10.1109/MCSE.2013.74.

[3] K. Nikolaidou, M. Seuret, H. Mokayed, M. Liwicki, A survey of historical document image datasets, International Journal on Document Analysis and Recognition (IJDAR) 25 (2022) 305–338. doi:10.1007/s10032-022-00405-8.

[4] B. C. G. Lee, J. Mears, E. Jakeway, M. Ferriter, C. Adams, N. Yarasavage, D. Thomas, K. Zwaard, D. S. Weld, The newspaper navigator dataset: extracting headlines and visual content from 16 million historic newspaper pages in Chronicling America, in: Proceedings of the 29th ACM International Conference on Information & Knowledge Management, CIKM '20, Association for Computing Machinery, New York, NY, USA, 2020, pp. 3055–3062. doi:10.1145/3340531.3412767.

[5] N. A. Ypsilantis, N. Garcia, G. Han, S. Ibrahimi, N. V. Noord, G. Tolias, The Met Dataset: instance-level recognition for artworks, 2022. URL: https://arxiv.org/abs/2202.01747. arXiv:2202.01747.

[6] F. Simistira, M. Seuret, N. Eichenberger, A. Garz, M. Liwicki, R. Ingold, DIVA-HisDB: a precisely annotated large dataset of challenging medieval manuscripts, in: 2016 15th International Conference on Frontiers in Handwriting Recognition (ICFHR), 2016, pp. 471–476. doi:10.1109/ICFHR.2016.0093.

[7] M. Mehri, P. Héroux, R. Mullot, J.-P. Moreux, B. Coüasnon, B. Barrett, HBA 1.0: a pixel-based annotated dataset for historical book analysis, in: Proceedings of the 4th International Workshop

on Historical Document Imaging and Processing, HIP '17, Association for Computing Machinery, New York, NY, USA, 2017, pp. 107–112. doi:10.1145/3151509.3151528.

[8] N. Rahal, L. Vögtlin, R. Ingold, Approximate ground truth generation for semantic labeling of historical documents with minimal human effort, International Journal on Document Analysis and Recognition (IJDAR) 27 (2024) 335–347. doi:10.1007/s10032-024-00475-w.

[9] F. Lombardi, S. Marinai, Deep learning for historical document analysis and recognition—a survey, Journal of Imaging 6 (2020) 110. doi:10.3390/jimaging6100110.

[10] M. A. Islam, I. E. Iacob, Manuscripts character recognition using machine learning and deep learning, Modelling 4 (2023) 168–188. doi:10.3390/modelling4020010.

[11] M. Coustaty, R. Pareti, N. Vincent, J.-M. Ogier, Towards historical document indexing: extraction of drop cap letters, International Journal on Document Analysis and Recognition (IJDAR) 14 (2011) 243–254. doi:10.1007/s10032-011-0152-x.

[12] J. Schlecht, B. Carqué, B. Ommer, Detecting gestures in medieval images, in: 2011 18th IEEE International Conference on Image Processing, IEEE, Brussels, Belgium, 2011, pp. 1285–1288. doi:10.1109/ICIP.2011.6115669.

[13] D. Borghesani, C. Grana, R. Cucchiara, Miniature illustrations retrieval and innovative interaction for digital illuminated manuscripts, Multimedia Systems 20 (2014) 65–79. doi:10.1007/s00530-013-0315-3.

[14] F. Aouinti, V. Eyharabide, X. Fresquet, F. Billiet, Illumination detection in IIIF medieval manuscripts using deep learning, Digital Medievalist 15 (2022) 1–18. doi:10.16995/dm.8073.

[15] P. Manoni, AI4MSS : un esperimento di intelligenza artificiale alla Biblioteca Apostolica Vaticana, in: G. Bergamin, T. Possemato (Eds.), Guardando oltre i confini : partire dalla tradizione per costruire il futuro delle biblioteche : studi e testimonianze per i 70 anni di Mauro Guerrini, AIB - Associazione Italiana Biblioteche, 2023, pp. 231–244. doi:10.1400/294065.

[16] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Springer, 2016, pp. 770–778. doi:10.1109/CVPR.2016.90.

[17] W. Zhao, D. Zhou, X. Qiu, W. Jiang, Compare the performance of the models in art classification, PLoS ONE 16 (2021). doi:10.1371/journal.pone.0248414.

[18] F. Milani, P. Fraternali, A dataset and a convolutional model for iconography classification in paintings, Journal on Computing and Cultural Heritage 14 (2021) 46. doi:10.1145/3458885.

[19] N. Banar, W. Daelemans, M. Kestemont, Transfer learning for the visual arts: the multi-modal retrieval of Iconclass codes, Journal on Computing and Cultural Heritage 16 (2023) 32. doi:10.1145/3575865.

[20] E. Mosqueira-Rey, E. Hernández-Pereira, D. Alonso-Ríos, J. Bobes-Bascarán, A. Fernández-Leal, Human-in-the-loop machine learning: a state of the art, Artificial Intelligence Review 56 (2022) 3005–3054. doi:10.1007/S10462-022-10246-W.

[21] M. Tkachenko, M. Malyuk, A. Holmanyuk, N. Liubimov, Label Studio: data labeling software, 2020-2024. URL: https://github.com/HumanSignal/label-studio, open source software available from https://github.com/HumanSignal/label-studio.

[22] S. Berg, D. Kutra, T. Kroeger, C. N. Straehle, B. X. Kausler, C. Haubold, M. Schiegg, J. Ales, T. Beier, M. Rudy, K. Eren, J. I. Cervantes, B. Xu, F. Beuttenmueller, A. Wolny, C. Zhang, U. Koethe, F. A. Hamprecht, A. Kreshuk, ilastik: interactive machine learning for (bio)image analysis, Nature Methods 16 (2019) 1226–1232. doi:10.1038/s41592-019-0582-9.

[23] M. Pachitariu, C. Stringer, Cellpose 2.0: how to train your own model, Nature Methods 19 (2022) 1634–1641. doi:10.1038/s41592-022-01663-4.

[24] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, L. Fei-Fei, ImageNet large scale visual recognition challenge, International Journal of Computer Vision 115 (2015) 211–252. doi:10.1007/s11263-015-0816-y.

[25] B. Fanini, D. Ferdani, E. Demetrescu, S. Berto, E. d'Annibale, ATON: an open-source framework for creating immersive, collaborative and liquid web-apps for cultural heritage, Applied Sciences 11 (2021) 11062. doi:10.3390/app112211062.