Self-Supervised Pre-Training for Join Retrieval of Cairo Genizah Manuscripts

Gal Grudka-Juster, Sharva Gogawale, Omer Ventura,
Berat Kurar-Barakat, Daria Vasyutinsky-Shapira, and Nachum Dershowitz
School of Computer Science and AI, Tel Aviv University

Keywords: Self supervised learning, Data augmentation, Join retrieval

Format: Talk

Abstract

The Cairo Genizah is spread across numerous libraries worldwide, often resulting in the fragments of a single original manuscript placed in different physical locations globally. This poses a significant challenge to historical reconstruction. Since manual comparison is slow and expert-dependent, using automated retrieval models can prioritize likely joins for further expert review [4].

One shared characteristic of fragments from the same work is the unique paleographic signature of their scribes. This work therefore focuses on *handwriting similarity as a primary signal* for manuscript join retrieval, which we frame as a representation learning task.

Since fragment images were photographed under varying conditions and preserved differently, background appearance and color become unreliable signals for paleographic similarity. To encourage the model to focus on fine-grained scribal features, fragment images are segmented to small-scale crops, consisting of 2–3 character sequences. Then, automated heuristics are applied to exclude crops with insufficient ink coverage or severe degradation.

As known-join labels are scarce, we employ a *self-supervised learning* approach that requires no such labels. Specifically, we fine-tune a DINOv2 backbone [2] using MoCo-v3 [1], which learns representations by contrasting augmented views of the same fragment crop. Standard augmentations are adapted for the manuscript domain, such as limiting rotations to a minimal range to preserve the orientation dependent features of the scribe's handwriting.

We introduce *domain-specific* data augmentations tailored for historical manuscripts, designed to improve robustness to corpus-specific visual distractions. These include: (1) in-domain background texture replacement, that is, compositing parchment textures (sampled from real in-domain crops) to discourage reliance on background texture (following the background augmenta-

tion strategy proposed by [3]); and (2) ink bleed-through simulation, modeling a corpus-specific production characteristic (see Figure 1).

To evaluate paleographic representation quality, we constructed a benchmark of known joins sourced from the Friedberg Genizah Project and the National Library of Israel. It consists of 234 fragments grouped into 86 known-join clusters, drawn from the Genizah corpus, but with no overlap of fragments with the pre-training set.

Preliminary results show that pretraining improves over a frozen DINOv2-ViT-S/14 zero-shot baseline on cross-fragment retrieval, the core task underlying manuscript reconstruction.

Acknowledgments

We are grateful to the Friedberg Genizah Project for images and data and to Roni Shweka for his contribution to the construction of the dataset. This research was funded in part by the Eu- ropean Union (ERC, MiDRASH, Project No. 101071829. Principal investigators: Nachum Dershowitz, Tel Aviv University; Judith Olszowy-Schlanger, EPHE-PSL; Avi Shmidman, Bar-Ilan University; and Daniel Stökl Ben Ezra, EPHE-PSL). Views and opinions expressed are, however, those of the authors only and do not necessarily reflect those of the European Union or the European Research Council Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.

Short Bio

GGJ is a M.Sc. student in computer science at Tel Aviv University; BKB received her Ph.D. in computer science from Ben-Gurion University and is a postdoctoral student at Tel Aviv University; DVS received her Ph.D. in Jewish thought from Ben-Gurion University and leads the MiDRASH research team; ND is emeritus professor of computer science at Tel Aviv University.

Proceedings interest: Yes

References

[1] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers, 2021.

[2] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien

Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024.

[3] Chaitanya K. Ryali, David J. Schwab, and Ari S. Morcos. Background augmentations for self-supervised learning. *arXiv preprint arXiv:2103.12719*, 2021.

[4] Lior Wolf, Rotem Littman, Naama Mayer, Tanya German, Nachum Dershowitz, Roni Shweka, and Yaacov Choueka. Identifying join candidates in the Cairo Genizah. *International Journal of Computer Vision*, 94(1):118–135, 2011.

Figure 1: Discarded crops are recycled by treating them as background source. We mask the foreground on valid training crops and compose their ink onto those donors' textures (resized, optionally background matched), turning waste crops into in-domain augmentation diversity.