Supervised vs. unsupervised deep learning for medieval Hebrew manuscripts

Vasyutinsky Shapira, Daria
dariavas@post.bgu.ac.il
The Open University of Israel

Rabaev, Irina
irinar@ac.sce.ac.il
Shamoon College of Engineering, Beer-Sheva, Israel

Alaasam, Reem
rym@post.bgu.ac.il
Ben-Gurion University of the Negev, Israel

El-Sana, Jihad
el-sana@cs.bgu.ac.il
Ben-Gurion University of the Negev, Israel

We present the ongoing project that aims to revolutionize our ability to access, analyze, and interpret digitized medieval Hebrew manuscripts by applying different deep machine learning techniques. We seek to facilitate access to the primary sources for deepening and transforming our understanding of medieval written Jewish cultures and cultural interferences between them. During the last few years, several software systems have been developed to access and analyze digitized manuscripts. Most of the development has been geared toward Latin scripts (Cloppet et al. 2016; Cloppet et al. 2017; Stutzmann, / Helias-Baron 2021), and much lesser work has been done for right-to-left historical documents. Among the most important are the DigiPal database on Latin paleography; the Europeana platform is connected to many libraries around Europe; the D-scribes project - a platform for Greek and Coptic papyri; Friedberg Genizah Project, which allows access to Hebrew documents from Cairo Genizah and Judeo-Arabic manuscripts; the major Ktiv project at the National Library of Israel that aims to provide connections to all collections of Hebrew manuscripts worldwide; Scripta Qumranica electronica for the Dead Sea Scrolls. The eScriptorium framework, READ, and their predecessors, Transkribus and Kraken applications, allow semi-automatic layout recognition and trainable text recognition (Shmidman et al. 2022).

The existing projects of digital Hebrew paleography work on different datasets and different kinds of manuscripts, each solving a different part of the puzzle. These projects complement each other for the final goal of recognizing the handwritten text in historical documents. They build on semi-automatic algorithms for HTR that are fine-tuned for every manuscript. We suggest that efforts be directed toward training fully automatic algorithms, and this will be achieved by combining traditional paleography with cutting-edge technologies. Deep machine learning is an efficient tool for the fully automatic extraction of different features from digitized medieval manuscripts. We will present an innovative unsupervised method of deep learning to recognize the fourteen classes of styles and modes of the medieval Hebrew script. The unsupervised deep learning algorithms will be compared to the previously achieved and published results on supervised deep machine learning models.

Our project, which includes computer scientists and digital humanists, has developed the open-access dataset of medieval Hebrew manuscripts (https://zenodo.org/record/6387471#.YxczWhCEafA), which is based on the Sfardata database for Hebrew paleography (https://sfardata.nli.org.il/). To the best of our knowledge, this is the first dataset in Hebrew that includes samples of major Hebrew writing types and modes to address the digital paleography community.

The dataset contains 715 page-images excerpted from 171 different manuscripts that were split into training, typical test, and blind test sets (manuscripts in the blind test set are those that did not appear in the test set).

During the supervised training, the models were trained until convergence using 50K patches extracted from pages in the train set. The model was trained using the binary cross entropy loss function. The patches were extracted using the patch generation method proposed in our previous work, which extracts patches with uniform text scale and, on average, five lines in each patch. Among the models (DenseNet, AlexNet, VGG19, ResNet50, SqueezeNet), the ResNet50 outperformed all the other models on every metric, achieving an accuracy of 60%. The model is continuously improved according to the prediction given for unseen manuscripts and the corresponding feedback from a paleographer. Because manual annotation of manuscripts is time-consuming and expensive, and only a small number of manuscripts can be annotated, we are now experimenting with unsupervised learning. Using our datasets, we train a Siamese network and extract the features of the last layer to cluster the data and generate the classes without needing further annotations. In our semi-supervised/unsupervised approach, we successfully trained a Siamese network to an accuracy above 90% and generated 14 clusters.

Bibliography

Barakat, Berat Kurar / Droby, Ahmad / Alaasam, Reem / Madi, Boraq / Rabaev, Irina / Shammes, Raed / El-Sana, Jihad (2021): "Unsupervised deep learning for text line segmentation", in *2020 25th International Conference on Pattern Recognition (ICPR)* : 2304-2311.

Beit-Arié, Malachi (2021): *Hebrew Codicology* , http://doi.org/10.25592/uhhfdm.8849 [26.04.2023].

Cloppet, Florence / Eglin, Veronique / Helias-Baron, Marlene / Kieu, Cuong / Vincent, Nicole / Stutzmann, Dominique (2017): "Icdar2017 competition on the classification of medieval handwritings in Latin script", in *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)* , 1: 1371-1376.

Cloppet, Florence / Eglin, Véronique / Stutzmann, Dominique / Vincent, Nicole (2016): "ICFHR2016 competition on the classification of medieval handwritings in Latin script", in *2016 15th International Conference on Frontiers in Handwriting Recognition (ICFHR)* : 590-595.

Dhali, Maruf A. / He, Sheng / Popović, Mladen / Tigchelaar, Eibert / Schomaker, Lambert (2017): "A digital palaeographic approach towards writer identification in the dead sea scrolls", in

Proceedings of the 6th International Conference on Pattern Recognition Applications and Methods (ICPRAM) 1: 693-702.

Dhali, Maruf A. / Nathan Jansen, Camilo / Willem De Wit, Jan / Schomaker, Lambert (2020): "Feature-extraction methods for historical manuscript dating based on writing style development." *Pattern Recognition Letters* 131: 413-420.

Droby, Ahmad / **Barakat, Berat Kurar** / **Vasyutinsky Shapira, Daria** / **Rabaev, Irina** / **El-Sana, Jihad** (2021): "VML-HP: Hebrew paleography dataset", in *Document Analysis and Recognition–ICDAR 2021: 16th International Conference*, Proceedings, IV 16: 205-220.

Droby, Ahmad / **Rabaev, Irina** / **Vasyutinsky Shapira, Daria** / **Barakat, Berat Kurar** / **El-Sana, Jihad** (2022): "Digital Hebrew Paleography: Script Types and Modes", in: *Journal of Imaging* 8, no. 5: 143.

Droby, Ahmad / **Vasyutinsky Shapira, Daria** / **Rabaev, Irina** / **Barakat, Berat Kurar** / **El-Sana, Jihad** (2022): "Hard and Soft Labeling for Hebrew Paleography: A Case Study", in *Document Analysis Systems: 15th IAPR International Workshop (DAS 2022), Proceedings* : 492-506.

Kestemont, Mike / **Christlein, Vincent** / **Stutzmann, Dominique** (2017): "Artificial paleography: computational approaches to identifying script types in medieval manuscripts", in: *Speculum* 92, S1: S86-S109.

Memon, Irfanullah / **Muhammad, Ammar ul Hassan** / **Choi, Jaeyoung** (2023): "Robustness of Contrastive Learning on Multilingual Font Style Classification Using Various Contrastive Loss Functions", in *Applied Sciences* 13, no. 6: 3635.

Schmarje, Lars / **Santarossa, Monty** / **Schröder, Simon-Martin** / **Koch, Reinhard** (2021): "A survey on semi-, self- and unsupervised learning for image classification", in *Access* 9: 82146-82168.

Schor, Uri / **Raziel-Kretzmer, Vered** / **Lavee, Moshe** / **Kuflik, Tsvi** (2021) "Digital research library for multi-hierarchical interrelated texts: from 'Tikkoun Sofrim'text production to text modeling", in Classics@ Journal : 18 issue 1, https://classics-at.chs.harvard.edu/classics18-schor-raziel-kretzmer-lavee-kuflik/ [26.04.2023].

Shmidman, Avi / **Guedalia, Joshua** / **Shmidman, Shaltiel** / **Shmidman, Cheyn Shmuel** / **Handel, Eli** / **Koppel, Moshe** (2022): "Introducing BEREL: BERT Embeddings for Rabbinic-Encoded Language", in *arXiv preprint arXiv:2208.01875* .

Sirat, Colette (2002). *Hebrew manuscripts of the Middle Ages* . Cambridge University Press.

Stökl Ben Ezra, Daniel / **Brown-DeVost, Bronson** / **Jablonski, Pawel** (2021): "Exploiting Insertion Symbols for Marginal Additions in the Recognition Process to Establish Reading Order", in *Document Analysis and Recognition–ICDAR 2021 Workshops, Proceedings*, II 16: 317-324.

Stutzmann, Dominique / **Helias-Baron, Marlène** (2021): "ICFHR 2016 Competition on the Classification of Medieval Handwritings in Latin Script-Dataset", https://hal.science/hal-03355868/ [26/04/2023].

Vasyutinsky Shapira, Daria / **Rabaev, Irina** / **Droby, Ahmad** / **Kurar Barakat, Berat** / **El-Sana, Jihad** (2022): "Is a deep learning algorithm effective for the classification of medieval Hebrew scripts?", in: *Studies in Digital History and Hermeneutics* : 349.

Wecker, Alan J. / Schor, Uri / Elovits, Dror / Stoekl Ben Ezra, Daniel / Kuflik, Tsvi / Lavee, Moshe / Raziel-Kretzmer, Vered / Ohali, Avigail / Signoret, Lily (2019): "Tikkoun sofrim: A webapp for personalization and adaptation of crowdsourcing transcriptions", in *Adjunct Publication of the 27th Conference on User Modeling, Adaptation and Personalization* : 109-110.

Yardeni, Ada (1997): *The book of Hebrew script: history, palaeography, script styles, calligraphy & design* . Carta Jerusalem.