Deep learning for paleographic analysis of medieval Hebrew manuscripts: a DH team collaboration experience

Daria Vasyutinsky Shapira[1], Irina Rabaev[2], Berat Kurar Barakat[1], Ahmad Droby[1], and Jihad El-Sana[1]

[1] Ben-Gurion University of the Negev,Beer-Sheva, Israel
[2] Shamoon College of Engineering, Beer-Sheva, Israel
{dariavas,berat,drobya}@post.bgu.ac.il
irinar@ac.sce.ac.il, el-sana@cs.bgu.ac.il

Abstract. Our research project is part of the Visual Media Lab, headed by Professor Jihad El-Sana, the Department of Computer Science at Ben-Gurion University of the Negev, Israel.
In this interdisciplinary project we apply deep learning models to classify script types and sub-types in medieval Hebrew manuscripts. The model incorporates the the techniques and databases of Hebrew paleography and (with reservations) Hebrew codicology.
Main theoretical base of our project is the SfarData dataset, that includes the full codicological descriptions and paleographical definitions of all dated medieval Hebrew manuscripts till the year 1540. In some exceptional cases, we go beyond this dataset framework. The major source of the data in terms of high definition photos of manuscripts is the Institute of Microfilmed Hebrew Manuscripts at the National Library of Israel that has undertaken the mission to collect copies of all extant Hebrew manuscripts from all over the world. We mostly use manuscripts from the National library of Israel, the British library, and the French National library.
This multidisciplinary project brings together researchers from both fields, Humanities and Computer Science. Currently, one professor, one lecturer, one post-doc, and two doctoral students are participating in the project. This is a very exciting work in which there are no ready-made solutions for the various challenges. We collectively discuss ways to address these challenges and adapt our solution on the go.
During the presentation, we will talk about how our project functions and how we strive to achieve a common result. The inevitable difficulties that we face during this collaboration include, *inter alia*, different research systems in Humanities and in Computer Sciences, lack of common terminology, different technical training, different requirements for publications and conferences, etc.

1 The humanities research problem

Human history, as we know it, is based on written text. It can be stone or papyrus or paper, but history consists of what was written down and has survived through the generations. Even the most ancient and longest traditions of oral transmission of a text are known to us to the extent that they were eventually recorded in writing.

For centuries the study of these written sources could only be from fragmentary information. They were limited by both geography and the physical capabilities of a human researcher. Already by the 18-19th centuries the amount of accumulated knowledge was big enough that a scientist could not master such a mass of information in his lifetime. However, it is obvious that a significant part of the data is still waiting to be discovered and analyzed.

Our research project is looking for ways to make some of these written sources, namely, Hebrew medieval manuscripts, available for study and research through machine learning. In other words, we want to teach the computer to recognize handwritten medieval Hebrew texts, and thus incorporate them into the available compendium of historical sources.

Unlike modern books, each manuscript is unique, as it was written at a certain point, under certain circumstances, by a certain scribe or scribes. In order to study a large amount of material, it must be classified in one way or another. Paleography and codicology are one of such classifications.

In our research project, we built upon existing achievements of Hebrew paleography and codicology. Paleography and codicology, the science of researching and classifying manuscripts, is one of the most important disciplines exploring ancient texts. Hebrew paleography is a relatively young discipline that began to take its current form in the middle of the 20th century, and which quickly borrowed and adapted tools and techniques from other paleography domains, such as Greek and Latin.

The first generation of Hebrew paleographers (Malachi Beit-Arié, Norman Golb, Benjamin Richler, Colette Sirat) collected and studied various key manuscripts, formulated and published the solid theoretical foundation in the field [1, 3, 10, 12, 8, 7]. In addition, the Sfar-Data project[??], which is lead by Malachi Beit-Arié and includes a large collection of classified dated manuscripts, is now partly incorporated into the catalogue of the National Library of Israel.

There is also a number of journal articles that use the same method of paleographic research of a manuscript as in the book of Engel and Beir-Arié[5, 6]. The Institute for Microfilmed Hebrew Manuscripts at the National Library of Israel has been collecting microfilms (now digital photos) of Jewish manuscripts for decades. The goal of this ongoing project is to obtain digital copies of all Hebrew manuscripts worldwide and make them easily available and accessible for the research. Today, the Institute hosts more than 70,000 microfilms and thousands of digital images, which makes more than 90% of the known Hebrew manuscripts. Besides, the National Library of Israel includes 11,000 original Hebrew manuscripts. These collections are large enough to train deep learning algorithms.

At the initial stage of our project we are training the algorithm to recognize different sub-types of the Medieval Hebrew script.

2 Solution and preliminary results

In this project we utilize recent development in deep learning for classifying different script types of historical Hebrew manuscripts. According to paleography research, handwriting styles evolve over time differently in various regions. Paleography experts estimate the origin of a manuscript and its approximate period using the writing style. However, this manual work is time consuming, tedious, expensive, and relies on highly trained experts. The number of paleography experts in Hebrew scripts is very small and is not expected to increase in the near future. In addition, these manuscripts originate from different geographical regions and their dates span over thousands of years.

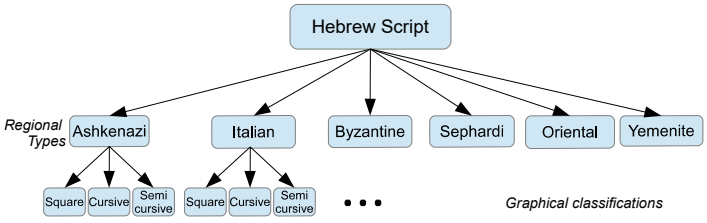

Fig. 1. Hierarchy of Medieval Hebrew Scripts

Medieval Hebrew scripts are classified into six regional script types: Ashkenazi, Italian, Sephardi, Byzantine, Oriental, and Yemenite. Each type is subdivided into three graphical classifications (sub-types): square, semi-cursive, and cursive [2], as shown in Figure 1. In total there are 15 different sub-classes, as some regional script types do not have semi-square or square form.

We have access to a large collection of various samples from different Hebrew scripts, the Sfar Data (http://sfardata.nli.org.il/), which are categorized into script type classes, including the raw material and high resolution copies.

Since the image sizes are quite big, to overcome technical limitations, we extract patch from each images, which are further are fed into CNN.

So far, we have experimented with two different architectures (simple CNN with three convolutional layers and ResNet). The dataset was divided into training and test sets, which include $538,468$ and $70,000$ patches, respectively.

We conducted several studies to determine which alterations of solution works best for this task.

Deep learning models are prone to over-fitting and can utilize much non-relevant information for the task at hand to decrease their loss and increase

classification accuracy. Therefore, we experimented with different input representations to determine the optimal amount of information passed to the machine learning model to achieve high accuracy while avoiding over-fitting. In this experiment, a simple CNN with three convolutional layers, which was trained using patches with varying attributes. Such attributes include color space: gray-scale, inverted gray-scale, and binary; shape: rectangular, and square patches; and whether the patches are smoothed or not (see examples in Fig. 2.)

We have found that gray-scale patches of size 350×350 gave the highest accuracy on the test sets and the lowest difference between the train and test losses, suggesting no over-fitting.

We recognized that in order to determine definitely that the model can classify writing styles based on the text alone and not other visual cues, it should be tested on manuscripts that that were seen during training. Thus, new manuscripts were added to the dataset, which was re-split into train, validation, and test sets, where the validation and test sets include pages from manuscripts that are not present in the training set.

Initially, the model's accuracy on the unseen manuscripts were low. We found that this is because the text size in the training set is very different form that in the validation and test sets. Therefore, there is a need to either re-scale the training, validation, and test sets to a nearly uniform text size or increasing the variation of text size in the training set using augmentation.

Table 1 presents the results on three types of test sets. Normal test set includes patches from unseen pages of the training manuscripts. Blind test set consist of patches from unseen pages. Scaled test set includes the scaled versions of the blind test patches. We experimented with four different architectures; a simple CNN with three layers, VGG19 [9], InceptionV3 [11] and ResNet152 [4]. Each of them trained from scratch (random weights), pre-trained using ImageNet, and trained with the augmented dataset, as explained above.

Practically, we need to know the how accurate the machine-learning model predicts the writing style of a give page. Table 2 shows the page prediction accuracy of the unseen pages from the train manuscripts. The accuracy increases as the number of patches sampled from the page increases, but the processing time also increases proportional to the number of patch in each page.

The network's coarse localization map provides evidences that the machine discriminates between the writing styles by considering specific parts of the text in the given patch (Fig. 3). It is left to the discretion of paleographers how legitimate is the machine's decision criteria.

3 The collaboration experience

Our project in its current form started in January 2020. The experience is very positive and even exciting, due both to the fact that it gives the feeling of constant scientific research and discovery, and also because of the satisfaction from constantly overcoming expected and unexpected challenges.

These challenges can be briefly formulated as follows:

Fig. 2. Example patches with varying attributes

Table 1. Accuracy on the different test sets for the different architectures trained for writing style classification task.

	Random			Pretrained			Augmented		
	Normal	Blind	Scaled	Normal	Blind	Scaled	Normal	Blind	Scaled
Simple CNN	95.25	12.75	14.99	-	-	-	90.50	23.75	38.69
VGG19	94.69	12.89	12.49	93.49	15.99	16.31	98.31	30.31	50.66
Inception v3	97.17	28.02	21.18	98.27	29.41	29.44	98.65	31.64	49.16
Resnet152	89.26	27.66	21.64	95.96	23.96	15.30	90.54	26.21	40.33

Table 2. Page-level accuracy computed using different numbers of patches randomly sampled from each page and the time elapsed for each accuracy computation. These results belongs to a pre-trained VGG-19 which is trained on 16000 patches and reach to a validation accuracy of %91.25

# of patches	3	5	7	9	11	13	15	17	19	21
Accuracy	74.04	76.99	82.89	84.07	81.71	84.66	86.14	86.43	**89.38**	87.61
Time (s)	192	226	262	295	330	357	387	393	457	483

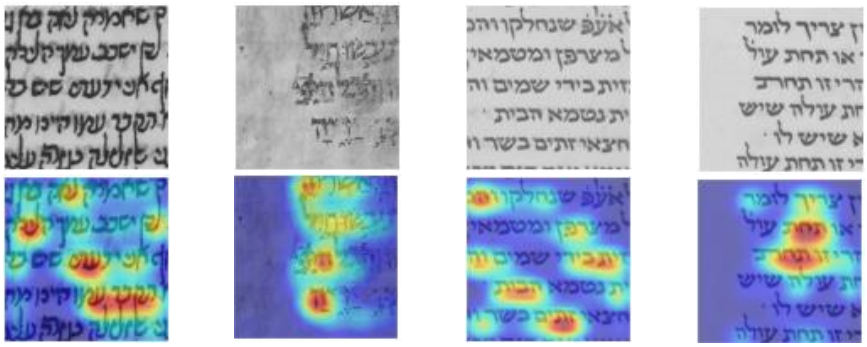

Fig. 3. Visualization of network's coarse localization map highlights the important regions in the document image patch for predicting the writing style.

- Finding your team. The main initial challenge in a DH project is to find one's counterpart. Researchers in the Humanities and in the Computer Sciences (CS) sit in different building on campus, attend different conferences, read different journals. There are practically no intersection points. In case of our project, both sides were looking for each other for a long time, and still we only met by a lucky coincidence. And yet, our project was initially in an advantageous position, because the CS team knew that they were looking for a paleographer (though they did not know where to find one) and our Humanities researcher knew approximately which CS tools could advance the project he was dreaming about. Finding a collaborator can be much harder if each side has only a vague idea of what the other side can offer, and this is often the case because of the totally different academic backgrounds. It goes without saying that it is much easier and more effective to work with those people who already have an interest in your topic, than to seek the help of people for whom your project might appear weird or incomprehensible.
- New team, new rules. In the Humanities, the researcher more often works alone, or with one collaborator, now one needs to get used to teamwork. It is easier on one hand, because each team member is responsible for his part of work, and tasks like writing a paper or making a presentation became easier. A team brainstorm is also a very positive factor. On the other hand, it is necessary to take into account the abilities and desires of the group members, which are not always clear in advance. The same is true for articles writing. In the Humanities, a researcher most often writes his article alone, or with one co-author. In the CS, as in the DH, an article is typically written by team. Both approaches have their advantages, and both require certain specific skills.

The participation of Dr. Vasyutinsky Shapira in this project is funded by Israeli Ministery of Science, Technology and Space, Yuval Ne'eman scholarship n. 3-16784.

– Unpredictability. When a researcher works alone on a project, for example preparing a compilation of different Manuscripts (Mss) of a text, he know how he will do it, he can check which methods have been used before, and he knows more or less what the outcome will be. Of cause, he could face an unexpected challenge, like a previously unknown manuscript that will change the general picture dramatically, but mostly we talk about minor changes. In a DH project, on the other hand, the previous experience one can rely on is very limited. Not only the ways of solving a problem have to be adjusted on the go, but also the goal itself has to be sometimes modified depending on the results. In our project, it turned out that the human paleography is so much based on intuition that it cannot be directly applied to machine learning. On the other hand, the machine can extract incomparably more small fragments of exact data. This leads us to a situation when even as we write this paper our approaches are constantly adjusted and improved.
– Learning a new language. Effective communication between all participants is essential for the success of any project. When participants come from different research backgrounds, it is of cause necessary that we learn to understand each other. The humanities researcher must be able to clearly formulate the problem. The Computer scientist should, again understandably, explain possible solutions, if any. The difficulty here is both the difference in the general approaches (for example, in the humanities, a problem is usually solved manually, while in computer science it is not customary to manually process the source material) and the lack of a common terminology. Professional literature in both fields is highly specialized to study it without relevant background, and thus, all members of the team have constantly to learn from each other.
– New tools. In the humanities, we typically use basic computer tools in our research: Word or other similar program for text processing, and a simple presentation program for conferences. In most fields in the humanities, the most prominent researchers are aged 50-70 and many of them will prefer to avoid using computer tools unless absolutely necessary. In the CS, the situation is of cause quite different, and it is the responsibility of the humanities researcher to learn at least some basic programs (i.e. the LaTeX that was used to write this paper) in order to work effectively with the team.

4 Conclusions and recommendations

Our research team includes both CS and Humanities researchers and work in a CS university lab, is a textbook example of a DH team. Our experience tells that this collaboration provides very a successful, promising, and satisfactory ecosystem for the entire team. There is little doubt that this type of research collaboration will become more mainstream in the near future, and its impact on the development of the Humanities will be even greater than can be imagined now.

We want also to suggest possible solutions for the challenges as described in the Collaboration Experience Section. These solutions aim at helping researchers

to find each other, learn to understand each other, and make their collaboration more efficient from the start.

- First of all, it is very desirable to have a common platform where people from the Humanities and CS could describe their projects and look for collaborators. This could be especially helpful when researchers do not know exactly what kind of counterpart they are looking for. Today, researchers that sit in different buildings of the same campus, often have no means to find each other. Within a particular university, such a role can be played by a dedicated DH research center.
- Both fields, the CS and the humanities, are highly specialized and complicated, and require many years of training. It is hardly possible to expect that one person could successfully master both fields and achieve high proficiency in both. Besides, a researcher in the humanities often needs years of practice in his field before he assembles enough knowledge and experience to put challenging research questions. Thus, though there is no point for a humanities researcher to try to really master CS, it is important to acquire general understanding of the field. This problem could be solved by adding to the university curriculum courses in the fundamentals of computer sciences tailored for MA and PhD students of Humanities. A DH research center could also make an effective bridge between the CS and Humanities faculties. DH conferences and workshops do help humanities researchers to master new computer skills, and they also often provide an overview of the state of art in a specific field, but first the more general understanding is required and the more professionally and academically its done, the better.
- In our project, we held regular weekly team meetings. At these meetings, both general issues and more specific technical issues are discussed, and at all parts of the discussion all team members are present. Thus, we can all consult each other, clarify complicated matters, and adjust our approach and methods on the go, in accordance with the results we get. These meetings help us learn each other's terminology, ideas and methods. Additionally, one of the CS team members gives the humanities member regular tutoring about the relevant fields of the CS. All this combined together gives very noticeable positive results, and half a year after the start of the project, the whole team speaks, as a rule, in a common and efficient language.

References

1. Beit-Arié, M.: Hebrew codicology. Jerusalem: Israel Academy of Sciences and Humanities (1981)
2. Beit-Arié, M.: Hebrew codicology: historical and comparative typology of Hebrew medieval codices based on the documentation of the extant dated manuscripts from a quantitative approach. M. Beit-Arié (2012)
3. Beit-Arié, M., Engel, E.: Specimens of mediaeval Hebrew scripts, in 3 vol. Israel Academy of Sciences and Humanities (1987, 2002, 2017)

4. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
5. Klaus, A., al-Rahman ibn Muhammad Ibn Wafid (Abu al Mutarrif), A.: Traducciones y adaptaciones al hebreo de los tratados médicos-farmacológicos del toledano Ibn Wafid. PPU (2007)
6. Pérez, I.: El testament de na Baladre (1325): nova aportació a l'estudi de les sinagogues de Girona. Agrupación de Editores y Autores Universitarios (2012)
7. Richler, B., Beit-Airé, M., Pasternak, N.: Hebrew manuscripts in the vatican library. Catalogue. Compiled by the Staff of the Institute of the Microfilmed Hebrew Manuscripts, Jewish National and University Library (Città del Vaticano). PMCid: PMC3523710 (2008)
8. Richler, B., Beit-Arié, M.: Hebrew manuscripts in the biblioteca palatina in parma: catalogue; palaeographical and codicological descriptions (2011)
9. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
10. Sirat, C.: Hebrew manuscripts of the Middle Ages. Cambridge University Press (2002)
11. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2818–2826 (2016)
12. Yardeni, A., et al.: The book of Hebrew script: history, palaeography, script styles, calligraphy & design. Carta Jerusalem (1997)