Cutting up medieval manuscripts (digitally) to trace their origins.

Rutger van Koert
Humanities Cluster KNAW
rutger.van.koert@di.huc.knaw.nl

Hannah Busch
Huygens ING
hannah.busch@huygens.knaw.nl

Mariken Teeuwen
Huygens ING
mariken.teeuwen@huygens.knaw.nl

The 'Digital Forensics for Historical Documents'[3] project at the KNAW Humanities Cluster is researching new ways are researched to analyze the vast numbers of images of medieval manuscripts that are now available via the International Image Interoperability Framework (IIIF)[4].

A key goal of the project is to find connections between manuscripts on the basis of the physical characteristics of their scripts.

Machine learning and specifically convolutional neural networks have become widely employed for automatic object detection, face recognition and handwritten text recognition.

In this short paper we will present the first results for our experiments. We have trained a Siamese neural network using supervised learning. Such networks are particular well suited for finding similarities and differences between objects of the same category. Financial institutions, for example, use such models to identify fraudulent signatures[1]. A Siamese network consists of two identical neural networks using shared weights that are connected using contrastive loss at the final layers. The exact loss function used is described in Hadsell et al[2]. In our experiments both networks consist of several convolutional layers for feature extraction. Using gradient descent, the network is trained to learn the (dis)similarity between pairs of image snippets.

The ground-truth training data are created by labeling the images of manuscript pages with metadata found in the descriptions of the codicological units, e.g. manuscripts written in the 12th century. Manuscript pages containing marginalia and glosses are excluded from the training data. The manuscripts are digitally cut up into smaller images consisting of text lines. Rectangular regions of equal proportions (height/width aspect ratio) containing a similar amount of text are then extracted from the text lines. Next, labels for each region are assigned using the respective labels from the original images. Finally, these snippets are resized to have the same size (height/width) and fed into the neural network.

After the training is completed we can then run the network on new snippets, extracting a fingerprint for each and storing the results with their metadata in a database for querying by comparing them with known fingerprints.

A web-front-end makes it easy for the user to query the system with a manuscript image. In the background the image is digitally converted into snippets and each of these snippets is separately used to find similar snippets in the database. The results from these searches are combined with the original images resulting in matches of complete pages containing the matching snippets. The final results are ranked by the number and quality of matching snippets.

Our initial results show that it is possible to connect manuscripts based on similar writing scripts using a fully automatic analyzer. Care must be taken in the selection of ground truth images because of the presence of marginalia and later textual additions which are of different dates and origins than the main text. Earlier versions of ground truth contained many marginalia from different time periods and these resulted in the network making mistakes. The quality of the results can be improved further by data augmentation techniques such as digital bleaching/darkening or adding salt and pepper noise into the image when training the network. These techniques force the network to learn features based on the shape of the writing instead of the parchment color and dirt accumulated over hundreds of years of use.

References

[1] Bromley, Jane & W. Bentz, James & Bottou, Leon & Guyon, Isabelle & Lecun, Yann & Moore, Cliff & Sackinger, Eduard & Shah, Rookpak. (1993). Signature Verification using a "Siamese" Time Delay Neural Network. International Journal of Pattern Recognition and Artificial Intelligence. 7. 25. 10.1142/S0218001493000339.
[2] Hadsell, Raia & Chopra, Sumit & Lecun, Yann. (2006). Dimensionality Reduction by Learning an Invariant Mapping. 1735 - 1742. 10.1109/CVPR.2006.100.
[3] https://www.huygens.knaw.nl/digital-forensics-for-historical-documents/?lang=en
[4] https://iiif.io/