Transformer-based HTR for Historical Documents

Phillip Benjamin Ströbel and **Simon Clematide** and **Martin Volk**
Department of Computational Linguistics
University of Zurich
{pstroebel,siclemat,volk}@cl.uzh.ch

Tobias Hodel
Walter Benjamin Kolleg
University of Bern
tobias.hodel@unibe.ch

1 Introduction

Handwritten Text Recognition (HTR) has become a valuable tool to extract text from scanned documents (Terras, in press). The current digitisation wave in libraries and archives does not stop at historical manuscripts. As such, HTR plays an essential role in making the contents of manuscripts available to researchers and the public.

HTR has undergone significant improvements in recent years, thanks in large part to the introduction of neural network-based techniques (Graves and Schmidhuber, 2008; Graves et al., 2009). Platforms like *Transkribus*[1] successfully integrated these approaches in a way that its HTR+ model (Michael et al., 2018) can achieve character error rates (CERs) of below 5% with little annotated ground truth material (Mühlberger et al., 2019).

However, a look at the digital platform for manuscript material for Swiss libraries and archives *e-manuscripta*[2] shows that in the category "correspondence" containing 45k titles, only 313, or 0.1%, contain transcriptions. Such large manuscript collections pose significant challenges to libraries and archives, especially because of the variety of handwriting styles. That the authors' handwriting changes according to what they were writing only adds in complexity. Fig. 1 exemplifies this by showing Rudolf Gwalther's handwriting in (a) a 16th century poetry volume and (b) a letter, among other handwritings from different authors (c and d).

The variability of such collections calls for models that adapt well to different hands with only little to no training data. Transformer-based architectures (Vaswani et al., 2017) have proven suitable to build large language representation models like, e.g., BERT (Devlin et al., 2018). BERT-style models are used to fine-tune specific models for

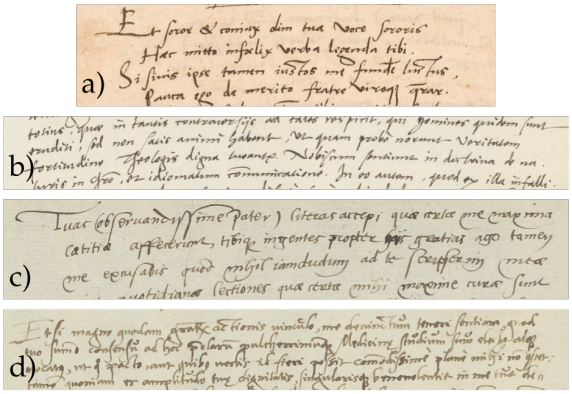

Figure 1: Different handwriting styles. Poetry by Rudolf Gwalther (a), letters by Rudolf Gwalther (b), Matthieu Coignet (b), and Kaspar Wolf (c).

natural language understanding and are known as strong transfer learners (Ruder et al., 2019). Most recently, transformers have found their way into image processing (Dosovitskiy et al., 2020; Touvron et al., 2021), which drove the development of image transformers (Bao et al., 2021).

2 Approach

The basis for our research is TrOCR (Li et al., 2021), which combines the BERT-style vision transformer BEiT (Bao et al., 2021) with a RoBERTa (Liu et al., 2019) language representation model. BEiT works as an encoder and is pre-trained on the Image-Net-1K (Russakovsky et al., 2015) dataset containing 1.2M images, while RoBERTa serves as a decoder producing the text. Li et al. (2021) used 687M of printed and about 18M of synthetically generated handwritten text lines in English to pre-train the TrOCR model. During this phase, the model learns to extract relevant features from the images and decode them into English text, therefore training the language model from scratch. The authors initialised the RoBERTa decoder with 6 and 12 layers, referring to them as BASE when paired with the pre-trained 12 layer

[1] https://readcoop.eu/de/transkribus/
[2] https://www.e-manuscripta.ch/

BEiT instance and LARGE when paired with the 24-layer BEiT model, respectively.

Finally, Li et al. (2021) fine-tuned their pre-trained TrOCR instances on "real-world" data, like the IAM dataset (Marti and Bunke, 2002). The IAM dataset consists of handwritten English lines from different authors. $\text{TrOCR}_{\text{BASE}}$ reaches a CER of 3.42% and $\text{TrOCR}_{\text{LARGE}}$ a CER of 2.89% on this dataset. The score of $\text{TrOCR}_{\text{LARGE}}$ is only 0.14 percentage points behind the best score of Diaz et al. (2021), who used a different approach.

Our research aims to exploit the pre-trained vision and language transformers, hoping that a model fine-tuned on historical manuscripts generalises well enough to be applied to extensive and variable manuscript collections. We want to test whether we can transfer the "knowledge" about handwriting in the English language TrOCR has acquired early modern manuscripts.

3 Data

Our data stem from the 16th century volume *Lateinische Gedichte* by Rudolf Gwalther.[3] Stotz and Ströbel (2021) downloaded the available images and partial transcriptions from *e-manuscripta* and loaded them into the Transkribus interface. They applied layout recognition to identify lines and baselines and aligned them with the transcriptions. The publicly available dataset has 4,037 image and corresponding text lines in Latin, which we split into 3,603 lines for training and 433 lines for validation.[4]

A second dataset consists of 16,584 lines in Latin from Heinrich Bullinger's (1504 - 1575) correspondence. It contains hands from about 60 different authors with a heavily skewed author distribution. We split the data into 13,843 lines for training, 1,685 lines for validation, and 1,056 for testing.

4 Experiments and Discussion

We trained Transkribus HTR+ models on the Gwalther and Bullinger data for 50 epochs as reference models.[5] Table 1 shows the result under "HTR+".

For the TrOCR architecture, using the same data, we fine-tuned both $\text{TrOCR}_{\text{BASE}}$ and $\text{TrOCR}_{\text{LARGE}}$ for three up to 20 epochs.[6]

[3] https://doi.org/10.7891/e-manuscripta-26750

[4] https://doi.org/10.5281/zenodo.4780947

[5] We used the *Acta_17 HTR+* as a base model.

[6] The untrained $\text{TrOCR}_{\text{LARGE}}$ model achieves a CER of 57.48% on the validation data.

System	data	fine-tuning epochs 3	5	10	15	20	epochs 50
HTR+	Gwalther	-	-	-	-	-	2.74
$\text{TrOCR}_{\text{BASE}}$		3.84	3.72	**3.18**	3.31	3.62	-
$\text{TrOCR}_{\text{LARGE}}$		2.94	2.72	2.58	**2.55**	2.62	-
HTR+	Bullinger	-	-	-	-	-	21.13
$\text{TrOCR}_{\text{LARGE}}$		-	-	-	16.53	-	-

Table 1: CERs for different models and different (fine-tuning) epochs on the validation set for Gwalther data and the test set for Bullinger data.

Table 1 presents the results of our initial experiments: the longer we fine-tune the models, the better their performance gets. This effect is less pronounced for $\text{TrOCR}_{\text{BASE}}$, however, where the performance even drops if we fine-tune more than ten epochs. Moreover, we note a clear performance gap between $\text{TrOCR}_{\text{BASE}}$ and $\text{TrOCR}_{\text{LARGE}}$, where $\text{TrOCR}_{\text{LARGE}}$ always performs better.

Our results are surprising because the pre-trained TrOCR model never saw any Latin data previous to our experiments. For example, our model only sees 23k Latin words during fine-tuning on the Gwalther data. The vocabulary overlap of the training and validation set is 68.9%. Moreover, TrOCR has never been confronted with early modern manuscripts. Nevertheless, we achieved a CER that beats our reference model trained on Gwalther data at 0.19 percentage points on the validation set and 4.6 percentage points on the Bullinger data on the test set.

We, therefore, assume that TrOCR is a robust and highly transferable handwriting representation model that is suitable for being fine-tuned on hands of all styles and origins.

5 Conclusion

Our initial experiments with TrOCR indicate that it outperforms state-of-the-art models for single-author and multi-author datasets. Astonishing is its strong performance on a language and handwriting styles it has never "learnt to read". Moreover, TrOCR does not require baseline information, in contrast to Transkribus models.

In future experiments, we want to investigate whether plugging in a pre-trained Latin RoBERTa decoder plus adapting the encoder to early modern handwriting can improve performance.

Moreover, we want to further examine TrOCR on more variable datasets. For example, projects

focusing on correspondences would benefit from HTR models that adapt to many different authors. Thus, we will investigate whether TrOCR generalises better to this data than conventional methods.

References

Hangbo Bao, Li Dong, and Furu Wei. 2021. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Daniel Hernandez Diaz, Siyang Qin, Reeve Ingle, Yasuhisa Fujii, and Alessandro Bissacco. 2021. Rethinking text line recognition models. *arXiv preprint arXiv:2104.07787*.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Alex Graves, Marcus Liwicki, Santiago Fernández, Roman Bertolami, Horst Bunke, and Jürgen Schmidhuber. 2009. A novel connectionist system for unconstrained handwriting recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(5):855–868.

Alex Graves and Jürgen Schmidhuber. 2008. Offline handwriting recognition with multidimensional recurrent neural networks. *Advances in neural information processing systems*, 21:545–552.

Minghao Li, Tengchao Lv, Lei Cui, Yijuan Lu, Dinei Florencio, Cha Zhang, Zhoujun Li, and Furu Wei. 2021. Trocr: Transformer-based optical character recognition with pre-trained models. *arXiv preprint arXiv:2109.10282*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

U.-V. Marti and Horst Bunke. 2002. The iam-database: an english sentence database for offline handwriting recognition. *International Journal on Document Analysis and Recognition*, 5(1):39–46.

Johannes Michael, Max Weidemann, and Roger Labahn. 2018. Htr engine based on nns p3.

Günter Mühlberger, Louise Seaward, Melissa Terras, and 51 more authors. 2019. Transforming scholarship in the archives through handwritten text recognition: Transkribus as a case study. *Journal of Documentation*, 75(5):954–976.

Sebastian Ruder, Matthew E Peters, Swabha Swayamdipta, and Thomas Wolf. 2019. Transfer learning in natural language processing. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Tutorials*, pages 15–18.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. 2015. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252.

Peter Stotz and Phillip Ströbel. 2021. bullinger-digital/gwalther-handwriting-ground- truth: Initial release.

Melissa Terras. in press. Inviting ai into the archives: The reception of handwritten recognition technology into historical manuscript transcription. In Lise Jaillant, editor, *Archives, Access and AI*, Digital Humanities Research. Transcript Verlag.

Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. 2021. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*, pages 10347–10357. PMLR.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.