Hessisches Kompetenzzentrum
für Hochleistungsrechnen

Transkribus Model Training



Project Manager
Dario Kampkaspar

Researchers
Michael Ustazewski, Aureilus Noble and Gregor Lanzinger

Principal Investigator
Dario Kampkaspar

Project Term
2024 - 2024

Clusters
Lichtenberg II Cluster Darmstadt

Software
PyTorch

University
Technische Universität Darmstadt

Introduction

Recognition of large amounts of printed or handwritten texts with to a high degree of accuracy – i. e. less than 0.1% Character Error Rate (CER; conforming to DFG standards) – depends on well-trained models. “Older” OCR-approaches (often based on direct comparison of characters with an archetype stored in the engine) give bad results the farther the form of a character is from the archetype. Especially, they do not work at all on handwritten texts.Modern AI-based approaches are much better at dealing with types of script or handwriting they have not encountered before. This is a necessary prerequisite for providing researchers with highly accurate transcriptions of printed or handwritten resources they may need for further research (e. g. as a basis for Text and Data Mining or to simplify retrieval of information included in a publication series or handwritten material such a lab reports).Especially training transformer-based models requires a significant amount of computing power. While to a lesser degree, this is also true about inference. Hence, tests were undertaken improve decoding time while still maintaining a low CER.

Methods

Based on transcriptions provided by ULB and other members of the READ Coop (Ground Truth, GT), BERT-based transformer models were trained for both text and layout recognition.These were used for inference on GT sets to evaluate the impact of different optimisation methods.

Results

We have optimised the TrHTR transformer model for handwritten text recognition (previously only used for our Text Titan model), evaluating the performance of various different optimisation methods: namely by changing the precision of model and examining the impact of different decoding techniques at inference (such as greedy search or beam search). With these experiments we have managed to achieve a 5x performance increase in inference time, with less than a 1% increase in the character-error-rate, e.g. CER 5.00% -> 5.04%. We have also begun training our next generation of language specific transformer models, which promise to significantly reduce our ability to accurately predict unseen materials. Our first model, trained on Spanish, saw the character error rate reduced from 4.6% (using our previous best model, PyLaia), to 2.8%.

Discussion

The performance increase achieved means reduced computing time for inference (and hence, also reduced environmental impact) with only an insignificant loss of precision.Next, we would like to continue these efforts by training larger models in further languages (initially German, English and Swedish), with this new model architecture.The resulting models are planned to be included in an automated pipeline used at ULB to bundle series of monographs for recognition (so as to avoid the setup costs for a task when the expected processing time for a monograph could be as little as 2 minutes).

Last Update: 2025-07-08 14:17