Zero-shot keyword spotting, using CLIP for modern manuscripts

Verreyen, Loren
loren.verreyen@uantwerpen.be
University of Antwerp, Belgium

Keyword spotting

Digitisation plays a key role in literary transmission, preservation, and analysis of handwritten documents [1]. However, a gap remains between providing access to digital images on the one hand, and making their actual contents machine-readable and searchable on the other hand. Currently, the workflow to transform handwritten documents into machine-readable text relies heavily on the process of handwritten text recognition (HTR). Typically, this HTR workflow makes use of supervised machine learning, requiring manually provided transcriptions to train a model to recognise the handwriting of the author of choice. Key players here are Transkribus and Kraken (Kahle et al. 2019; Kiessling 2019). Once the HTR model is trained on sufficient image-transcript pairs, it can be used to automatically transcribe manuscripts in the same handwriting or one that closely resembles the original training data. These automatically acquired transcriptions can then be used for downstream tasks such as keyword spotting. This workflow introduces (at least) two bottlenecks in the process of producing machine-readable texts from digitised documents: (1) the amount and quality of the original training data plays an important role in how good the model will be at generating new transcriptions, and (2) if a new handwriting is introduced that is too different from the original training data, the original model needs to be fine-tuned to adapt to the new hand or a new model needs to be trained from scratch - both options require significant time and expertise. In this presentation, I aim to demonstrate how the recently introduced CLIP-model (Contrastive Language Image Pre-Training) (Radford et al. 2021) can be used as a valuable tool to access and analyse digitised manuscripts without the need of providing manual transcriptions first. This entails a shift which would allow the efficient browsing of digital images of any handwritten document.

Model and data

CLIP is a multimodal model that consists of two transformer encoders: one for textual information and one for visual information. The two encoders are jointly trained to maximise the cosine similarity of correct image-text pairs and minimise the cosine similarity of incorrect image-text pairs. Because of the wide range of image-text pairs that are used to train CLIP, the model can be used for a variety of tasks without the need to be finetuned first – examples of such tasks are optical character recognition (OCR), action recognition, and geo localization (Ibid.). The idea that CLIP could also be used to recognise handwriting has already been suggested in the original CLIP paper (Ibid.), where CLIP's performance was tested on the MNIST dataset (Deng 2012) – a dataset consisting of handwritten digits. However, CLIP's potential to break into handwritten data has further remained unexplored. In this presentation, CLIP's zero-shot capabilities on handwritten words are tested on the IAM dataset (Marti / Bunke 2002). This dataset contains the handwriting of several hundred writers, making the dataset an interesting case study to analyse how valuable this new keyword spotting method can be when handling a wide range of different handwritings – a prevalent challenge when it comes to digitising handwritten data.

Analysis

In a first part of the analysis, CLIP's zero-shot capabilities are assessed. Following (Radford et al. 2021), the ViT-L/14@336px model is used, further referred to as CLIP. To perform zero-shot keyword spotting, all text labels present in the IAM dataset are fed into CLIP, functioning as a possible text snippet to be paired with the handwritten word images that are simultaneously fed into the model. CLIP's zero-shot keyword spotting capability is assessed based on the recall score, i.e. how many times CLIP is able to detect a certain keyword. CLIP scores a recall score of 16.55%. Next, CLIP's image embeddings are assessed by using them as the input of a logistic regression model that treats the keyword spotting task as a simple classification problem. The logistic regression model reaches a recall score of 10.94%. This low score shows that the off-the shelf CLIP model does not yet manage to capture sufficient relevant information in the image embeddings to cluster similar keyword images. In a last part of this study, CLIP will be fine-tuned on part of the IAM dataset to improve the original model. By creating triplets on the fly - an anchor, a positive match, and a negative match - and using a contrastive metric learning approach, positive matches are pushed closer together, negative matches are pushed further apart.

Notes

1. This paper is part of the CATCH 2020 project (Computer-Assisted Transcription of Complex Handwriting), under the supervision of Dirk Van Hulle (UAntwerp).

Bibliography

Deng, Li (2012): “The MNIST Database of Handwritten Digit Images for Machine Learning Research [Best of the Web]”, in: *IEEE Signal Processing Magazine* 29, 6: 141–142. DOI: 10.1109/MSP.2012.2211477.

Kahle, Philip / **Colutto, Sebastian** / **Hackl, Günter** / **Mu#hlberger, Günter** (2017): “Transkribus—A Service Platform for Transcription, Recognition and Retrieval of Historical Documents” in: International Association of Pattern Recognition (ed.): 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), Kyoto, Japan, November 2017: 19–24. DOI: 10.1109/ICDAR.2017.307.

Kiessling, Benjamin (2019): “Kraken - an universal text recognizer for the humanities”, in: Digital Humanities Conference 2019 (DH2019), Utrecht, the Netherlands, July 2019. DOI: 10.34894/Z9G2EX.

Marti, Urs-Viktor / **Bunke, Horst** (2002): “The IAM-database: An English sentence database for offline handwriting reco-

gnition", in: *International Journal on Document Analysis and Recognition (IJDAR)* 5, 1: 39–46. DOI: 10.1007/s100320200071.

Radford, Alec / **Kim, Jong Wook** / **Hallacy, Chris** / **Ramesh, Aditya** / **Goh, Gabriel** / **Agarwal, Sandhini** / **Sastry, Girish** / **Askell, Amanda** / **Mishkin, Pamela** / **Clark, Jack** / **Krueger, Gretchen** / **Sutskever, Ilya** (2021): "Learning transferable visual models from natural language supervision", in: *Proceedings of the 38th International Conference on Machine Learning PLMR*, Vienna, Austria, July 2021: 8748–8763. DOI: 10.48550/arXiv.2103.00020.