Master's thesis topic description: Fine-tuned optical character recognition for dental fossil markings

Riikka Korolainen

014926659

1 General problem area

paleoecology: data analysis on fossil data points

what we are able to learn: makeup of species of past ecosystems, reactions of species to environmental changes

since 80's KNM has stored handwritten notes on found fossil specimens in Kenya/Ethiopia. approx 4,500 pages with approx 50 specimens in the catalogue

digitisation of hand-written fossil catalogues of the National Museum of Kenya

digitisation with Azure AI Vision services done, but that model could not read the special characters in the "element" column

2 Research questions

how well few-shot transfer learning methods perform at transfer from reading regular handwritten characters to reading charaters that have lower and upper script numbers

insert here image (ota element-sarake ja element csv ja toothrecords)

3 Methodologies

The thesis will consist of a literature review and an experimental section. The literature review will consist of a synthesis of the relevant background information on deep learning, optical character recognition, transfer learning and paleoecology. The main part of the literature review will consist of comparing solutions to related problems of digitizing handwritten text that contains more unconventional characters. This part of the review can be divided into three partially overlapping review questions:

- What is the best OCR model architecture?
- What is the best few-shot transfer learning method?
- Which solutions have previous works on related problems applied?

The goal of the literature review will be to choose a small set of solutions, which will be benchmarked in the experimental section. This part of the work will consist of attempting different combinations of approaches, and then comparing performance metrics. This will require a diligent experiment tracking system and hand-annotating data. The experiments will be performed using standard python data science libraries (pytorch, MLflow) and data from the fossil catalogues and specimen cards from the National Museum of Kenya. As a final deliverable, the fine-tuned model will be stored and made publicly available to be used by the museum.

4 Key references

- [1] presents a promising base model for fine-tuning
- [4] the NOW database of fossil mammals
- [2] A survey on optical character recognition methods
- [3] A survey on few-shot transfer learning

References

[1] M. Li, T. Lv, L. Cui, Y. Lu, D. Florencio, C. Zhang, Z. Li, and F. Wei. *TrOCR: Transformer-based Optical Character Recognition with Pre-trained Models.* 2021. arXiv: 2109.10282 [cs.CL].

[2] J. Memon, M. Sami, R. A. Khan, and M. Uddin. "Handwritten Optical Character Recognition (OCR): A Comprehensive Systematic Literature Review (SLR)". In: *IEEE Access* 8 (2020), pp. 142642–142668. DOI: 10.1109/ACCESS.2020.3012542.

[3] Y. Song, T. Wang, P. Cai, S. K. Mondal, and J. P. Sahoo. "A Comprehensive Survey of Few-shot Learning: Evolution, Applications, Challenges, and Opportunities". In: *ACM Comput. Surv.* 55.13s (July 2023). ISSN: 0360-0300. DOI: 10.1145/3582688. URL: https://doi.org/10.1145/3582688.

[4] I. Žliobaitė, M. Fortelius, R. L. Bernor, L. W. van den Hoek Ostende, C. M. Janis, K. Lintulaakso, L. K. Säilä, L. Werdelin, I. Casanovas-Vilar, D. A. Croft, L. J. Flynn, S. S. B. Hopkins, A. Kaakinen, L. Kordos, D. S. Kostopoulos, L. Pandolfi, J. Rowan, A. Tesakov, I. Vislobokova, Z. Zhang, M. Aiglstorfer, D. M. Alba, M. Arnal, P.-O. Antoine, M. Belmaker, M. Bilgin, J.-R. Boisserie, M. R. Borths, S. B. Cooke, J. A. van Dam, E. Delson, J. T. Eronen, D. Fox, A. R. Friscia, M. Furió, I. X. Giaourtsakis, L. Holbrook, J. Hunter, S. López-Torres, J. Ludtke, R. Minwer-Barakat, J. van der Made, B. Mennecart, D. Pushkina, L. Rook, J. Saarinen, J. X. Samuels, W. Sanders, M. T. Silcox, and J. Vepsäläinen. "The NOW Database of Fossil Mammals". In: *Evolution of Cenozoic Land Mammal Faunas and Ecosystems: 25 Years of the NOW Database of Fossil Mammals.* Ed. by I. Casanovas-Vilar, L. W. van den Hoek Ostende, C. M. Janis, and J. Saarinen. Cham: Springer International Publishing, 2023, pp. 33–42. ISBN: 978-3-031-17491-9. DOI: 10.1007/978-3-031-17491-9_3. URL: https://doi.org/10.1007/978-3-031-17491-9_3.