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 Methologies

Literature review: find and compare the most successful OCR models for handwriting with bounding boxes given (we have them with azure vision)

Literature review: best few-shot transfer learning methods

Hand-label data or request from Kenya

Experiment: combinations of best OCR models + transfer techniques keep track of experiments with MLflow

Train + store best method as a publicly available ML model. Submit to be used by KNM + stakeholders. Also run model on catalogues, get out cleaned tooth records column

4 Key references

- [1] presents a promising base model for fine-tuning
- [2] the NOW database of fossil mammals

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] 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.