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

Riikka Korolainen

014926659

1 General problem area

The research area of paleoecology studies past environments based on fossil remains. Relying on ecological principles and statistical methods it is possible to learn various traits of past ecosystems, such as what past climates were like, how species reacted to environmental changes and how early humans lived [2].

The National Museum of Kenya stores handwritten notes on found fossil specimens in the museum archives with the earliest being from the 1980's. The catalogue consists of approximately 90,000 unpublished specimens and an ongoing project is to digitize and publish the data. Accurately digitizing the data points will extend a paramount dataset to paleoecological research from the East African region and allow the data to be integrated to larger collections of fossil data, such as the NOW database [8].

Most of this digitisation effort was completed in a previous project using commercial Azure AI Vision services. However, this model could not read the special characters found in the data denoting tooth fragments. The aim of this thesis is to fine-tune an optical character recognition (OCR) model to recognize these markings. A sample of the markings and the Azure Vision output can be found in Figure 1.

2 Research questions

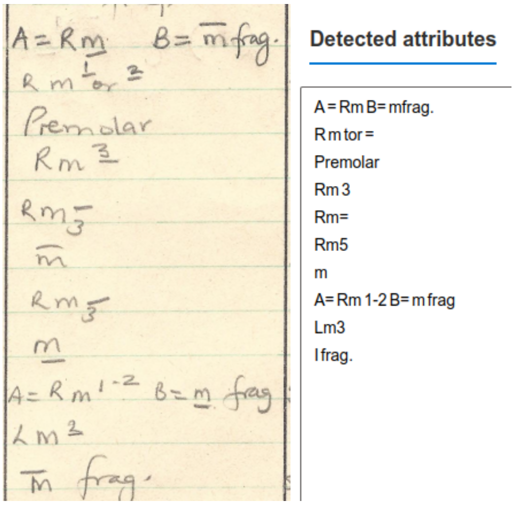
The main research question is the following:

Which base model and fine-tuning method is most accurate for recognizing the special characters found in dental fossil notation?

The special characters consist of lower- and upper script numbers and letters with a line on top or underscoring. Additionally, there are multiple conventions for denoting the same tooth. For the fine-tuning, the methods are likely limited to few-shot transfer learning methods, since the data needs to be hand-labeled.

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



```
A=RmB=mfrag.
Rmtor=
Premolar
Rm3
Rm=
Rm5
m
A=Rm1-2B=mfrag
Lm3
Ifrag.
```

Figure 1: Sample from the tooth notation and corresponding Azure AI output

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

- [4] A promising base model for fine-tuning
- [7] [1] [3] Solutions to similar problems
- [5] A survey on optical character recognition methods
- [6] A survey on few-shot transfer learning
- [2] A thorough reference and bibliography on paleoecology

5 Timeline and supervisors

The aim is to finalize the work by the end of December, either to the steering group submission deadline on 19 December 2024 or 23 January 2025. Supervisor: Indrė Žliobaitė, second reviewer: Kari Lintulaakso

References

[1] M. Christy, A. Gupta, E. Grumbach, L. Mandell, R. Furuta, and R. Gutierrez-Osuna. "Mass Digitization of Early Modern Texts With Optical Character Recognition". In: *J. Comput. Cult. Herit.* 11.1 (Dec. 2017). ISSN: 1556-4673. DOI: 10.1145/3075645. URL: https://doi.org/10.1145/3075645.

[2] J. T. Faith and R. L. Lyman. *Paleozoology and Paleoenvironments: Fundamentals, Assumptions, Techniques.* Cambridge University Press, 2019.

[3] M. A. Karim, S. M. Rafiuddin, M. J. Islam Razin, and T. Alam. "Isolated Bangla Handwritten Character Classification using Transfer Learning". In: *Proceedings of the 2nd International Conference on Computing Advancements.* ICCA '22. Dhaka, Bangladesh: Association for Computing Machinery, 2022, pp. 11–17. ISBN: 9781450397346. DOI: 10.1145/3542954.3542957. URL: https://doi.org/10.1145/3542954.3542957.

[4] 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].

[5] 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.

[6] 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.

[7] G. Zhao, W. Wang, X. Wang, X. Bao, H. Li, and M. Liu. "Incremental Recognition of Multi-Style Tibetan Character Based on Transfer Learning". In: *IEEE Access* 12 (2024), pp. 44190–44206. DOI: 10.1109/ACCESS.2024.3381039.

[8] 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.