Contents

1 Introduction

2 background

2.1 neural networks and deep learning

keywords to explain (maybe) from [2] - knowledge distillation - generalist models - large unsupervised training data sets - transformers - cnns - deep neural networks - self attention - convolution - transfer learning - encoder/decoder - model compression - loss functions - cross-entropy - data augmentation - training/validation/test data sets - learning rate - batch size - tokenizing - image patches - self-attention - multi-head self-attention - performance metrics - precision - recall - f1

2.2 paleoecology

This section will have a summary on what fossil data can be used for.
the why: why do this at all? why is accurate dental data relevant, in general?

2.2.1 Basics on ecology

Tolerances and niches: basis for environmental reconstruction [1]
Theory that the data analysis relies on

2.2.2 Paleoenvironmental reconstruction

maybe, how the data is used

2.2.3 Diets, evolution, etc

maybe, how the data is used

2.2.4 Animal teeth notation

Add here description of teeth: types and different notation styles in fossil catalogues

3 data methods etc

3.1 Unicode characters used for data labeling

explain: unicode has graphenes with code points. eg a is one graphene one code point, à is one graphene two code points (dot on top and the letter). the top thing -like characters will be called "modifiers".

markings contain letters and numbers with no line, line on top or line at the bottom. Each character can be lower- or upper script. The modifiers used are: macron with lower (Ā) and upper variant.

Unicode [3] has characters that are for example upper script, but these were not used for two reasons:

- lower and upper script character set is incomplete for this purpose (eg 3 with upper macron and lower script needed)

- from the model perspective 3 and $_3$ are no more similar than A and B, however, three combined with lower script modifier and 3 with upper script modifier all contain the same unicode character 3 with only the modifier changing. The problem here is that there is no lower or upper case modifiers in unicode. Therefore, the caret (Ǎ) was chosen as the lower script modifier, and the circumflex accent (Â) as upper script. These were chosen since the arrow-like modifier pointing up or down is maybe the most logical placeholder for the missing modifier. More traditional workarounds of missing upper or lower script, the underscore "_" and separate caret character "$\hat{}$" were not used to keep one unicode graphene represent one character on the page. Also on the other hand using one modifier for all lowercase characters allows the model to understand that there is a similarity between all lowercase characters. The intention is that one idea about a character is encoded as one code point, so that the model can learn the mapping from the image of the character to the code point combination

4 results

5 conclusion

References

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

[2] 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]`.

[3] The Unicode Consortium. *The Unicode Standard*. `https://home.unicode.org/`. [Accessed: 2024-09-04]. 2024.