Contents

1 Abstract

Digitizing and uniformizing the structure of handwritten fossil records exhibits a great potential for increasing the accuracy of paleontological data analysis by increasing sample sizes. Approximately 90, 000 of such samples reside in the archives of the National Museum of Kenya, and an ongoing effort is to store this data in a digital format for better accessibility. A previous project utilized a commercial optical character recognition service for automated reading of these catalogues. This generalist handwriting detection model lacked the ability to detect special characters used to denote tooth samples, and could not utilize prior knowledge of the vocabulary that is more likely to be present in the data, leading to loss of information and detection mistakes.

This thesis aims to build a specialist character recognition model to increase the accuracy of the bone or tooth type specifying column of the digitized data by fine-tuning a state-of-the-art optical character recognition model with few-shot transfer learning. This is performed by first finding most accurate recognition models, variants of convolutional neural networks or vision transformers, and most successful transfer learning methods for adapting a model to a new character set. Then, the character recognition accuracy of combinations of these methods are benchmarked using handlabeled image segments from the fossil catalogues. The final aim of this work is to use the best-performing model to obtain an accurate reading of the catalogues of the National Museum of Kenya, and publish the final model to be used by the paleontological community for further digitization efforts.

Keywords: Optical character recognition, Few-shot transfer learning, Vision transformers, Paleontological databases

2 Introduction

3 Background

3.1 Neural Networks and Deep Learning

3.2 Fundamentals on paleoecology

3.2.1 Basics on ecology

3.2.2 Paleoenvironmental reconstruction

3.2.3 Diets and evolution

3.2.4 Composition of mammal teeth

4 data methods etc

4.1 data description

4.1.1 Notes on creating the dataset

4.1.2 Unicode characters used in data labeling

To label the text found in cropped-out tooth fragment handwriting images, a few nonobvious conventions had to be set in place to construct a labeling system that can be assumed to be easier to learn for a machine learning model. The main guiding rule in these decisions was to encode each feature in the text in one consistent manner. What is meant by features and manners of denoting is explained next.

The unicode system [1] constructs all known characters as signs called graphenes. Each graphene can consist of any number of code points, with each code point having an unique identifier, denoted with "U+code point id". Examples of graphenes with one code point are latin letters, such as 'K', special characters, such as '@', '%' and '+', or letters from different writing systems, such as Examples of multi-code point graphenes are latin letters with accents, such as 'ê', or emoji characters with non-default skin tone, such as Code points added to the main code point, such as the circumflex accent '̂ are called modifiers.

The guiding principle in labeling the data was to encode each concept in the text as one unicode code point. A concept could be, for instance, the number two, or a character being positioned in subscript. The aim of this decision is to allow the model to find common image traits between characters of a similar type: a subscript character has dark pixels in lower positions, and shapes of all number two's have similar curvatures, for instance. As a second principle, it was chosen that each single character in the image, such as "letter C" or "a subscript four with a horizontal top line", would always be labeled as one graphene. These rules makes the encoding choices nonobvious: for example, a subscript number two would intuitively be labeled as the unicode code point '$_2$', but this was not done, since this graphene does not contain the code point for number two, and as a one

code point graphene has no code point to extract to be used among the other subscript numbers. Another intuitive choice, '_2', would violate the one graphene per character rule.

The special characters in the dental fossil handwriting consist of sub- and superscript numbers, and characters with a horizontal line on top or bottom. Additionally, these two modifiers sometimes co-occur. Both denote which jaw the fragment is from: subscript and horizontal line on top of the character denote lower jaw, whereas superscript or line at the bottom of character signal upper jaw. In a few rare occurences, fractions are present to denote which proportion of the tooth is remaining in the sample. A sample of each type of notation can be found in Figure . Note that ambiguous notations of for instance subscript number with a horizontal line at the bottom are allowed with this writing system. The labeling notation chosen preserves the option to label these ambiguities.

The following code points were chosen to denote the tooth marking system in the data labels. The base code point modified with unicode modifiers was always chosen to be the latin letter or number present in the character. In the case of fractions, the number in the denominator was chosen as the base code point. The horizontal line on top of a character was denoted with the combining macron modifier (U+0304, eg. Ā), the line at the bottom respectively with the combining macron below (U+0331, eg. A̲). As the unicode system lacks sub- or superscript modifiers, other accent modifiers were used instead. A subscript character was denoted with the combining caron (U+030C, eg. Ǎ), and respectively the superscript with the combining circumflex accent (U+0302, eg. Â). For fraction nominators, a modifier was chosen for each digit present in the dataset (TODO: add here after chosen). These choices were made to improve human readability of the dataset, as the modifier choices are not relevant from the model perspective. A summary of the characters found and their labels can be found in image ....

4.1.3 Data preprocessing

4.2 Methods

4.2.1 Encoding prior knowledge

5 results

6 conclusion

References

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