

# Contents

<b>1</b>	<b>Abstract</b>	<b>1</b>
<b>2</b>	<b>Introduction</b>	<b>2</b>
<b>3</b>	<b>Background</b>	<b>2</b>
3.1	Neural Networks and Deep Learning . . . . .	2
3.2	Fundamentals on paleoecology . . . . .	2
3.2.1	Basics on ecology . . . . .	2
3.2.2	Paleoenvironmental reconstruction . . . . .	2
3.2.3	Diets and evolution . . . . .	2
3.2.4	Composition of mammal teeth . . . . .	2
<b>4</b>	<b>data methods etc</b>	<b>3</b>
4.1	data description . . . . .	3
4.1.1	Notes on creating the dataset . . . . .	3
4.1.2	Unicode characters used in data labeling . . . . .	3
4.1.3	Data preprocessing . . . . .	5
4.2	Methods . . . . .	5
4.2.1	Encoding prior knowledge . . . . .	5
<b>5</b>	<b>results</b>	<b>5</b>
<b>6</b>	<b>conclusion</b>	<b>5</b>

## 1 Abstract

Digitizing and uniformizing the structure of handwritten fossil catalogues 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

Since geological events tend to erode organic remains the faster they remain decomposes, the hardest materials in the corpse represent largest fractions of fossil datasets. These hard materials include shells, bones and especially teeth, and the last is prominent in fossil data analysis also due to the fact that they encode a diverse set of information on the livelihood of the organism [1]. The identification of the fossil remain is done at the finest resolution possible, preferring taxon information over just identifying the genus, for instance. Finest-resolution information derived from dental fossils are the taxon the tooth is from, and which tooth or teeth are found in the specimen. This section presents the naming and indexing system for mammal teeth commonly used in paleontological datasets, as described by Hillson [2], and some common shorthand versions present in the dataset digitized in this work.

Specimens including more complete fragments of the jaw are described with terminology related to the jaw bones. All mammals share the same bone structure around the mouth: the lower jaw consists of two bones called *mandibles*, joining in the middle, whereas the upper jaw consists of bones called *maxilla* and *premaxilla*, that also form large parts of the face. A common trait across many mammals is also that the permanent teeth erupt in the youth of the animal, replacing the 'milk' or *deciduous* teeth. Shorthands commonly used for these terms are 'mand' for mandibles, and capital letter 'D' for the deciduous teeth.

The tooth rows of mammals are classified to four classes; *incisor*, *canine*, *premolar* and *molar* and indexed with a numbering system. Moving from the middle of the tooth row towards the side, there are up to three incisors, used for catching food and denoted with the letter 'i'. Behind them is the canine tooth, used for cutting, and in case of predators, killing. This tooth is denoted with the letter 'c'. Behind the canine are up to four premolars, noted with 'p'. These teeth vary most between taxa in form and function with functions including cutting, holding and chewing food. The teeth at the back of the row are called molars, 'm', and are primarily used for chewing. Molars, like the other tooth types, vary in number between taxa, and are at most three. The numbers are always increasing when moving back in the tooth row, but in the case of missing teeth in a taxon, the numbers do not necessarily start from one: instead, the number is chosen to have teeth with same numbers as alike each other as possible. Thus, a taxon with only two premolars might only have the teeth P3 and P4.

Location of the tooth present in the fossil is described with directional terms specifying the side, jaw and the location on the jaw. The most intuitive are left and right describing the side, where one needs to note that each denotes the side from the viewpoint of the animal, not the observer. Mammal teeth are always symmetrical, thus every tooth always has the equivalent other-jaw counterpart. The distance from the throat is described with the terms *distal*, 'close to the heart' and *mesial*, 'far from the heart'. Short-form versions for these terms include capital 'L' or 'Lt' for left, capital 'R' or 'Rt' for right and 'dist.' for distal. The jaw, upper or lower, has three dominant notation styles: one is to sub- or superscript tooth index numbers, other is to over- or underline tooth markings, and the last style, prominent in digital fossil data, is to set the tooth type letter to upper- or lowercase. In each of these systems, a superscript, underline, or capital letter denotes upper jaw, and conversely subscript, overline or lowercase letter denotes the lower jaw. An illustration of the mammal tooth system is presented in Figure 1. Terminology and corresponding shorthands are summarized in Table (TODO).

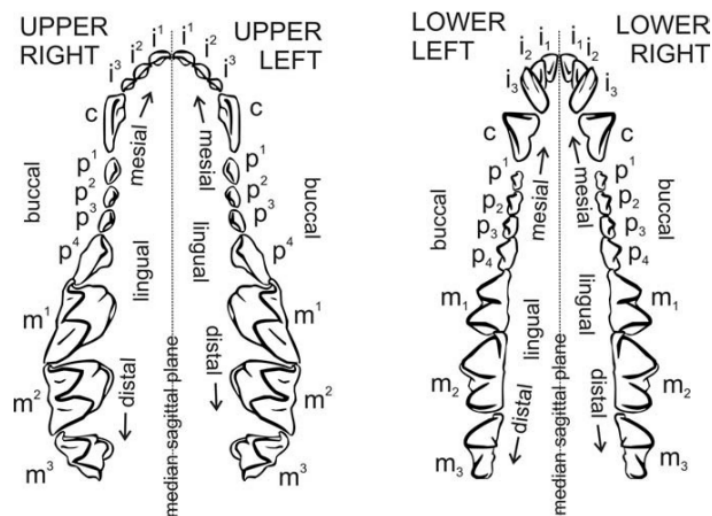


Figure 1: Mammal teeth composition, from Hillson [2].

## 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 [3] constructs all known characters as signs called graphemes. Each grapheme can consist of any number of code points, with each code point having a unique identifier, denoted with "U+code point id". Examples of graphemes 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 graphemes 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 grapheme. 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 grapheme does not contain the code point for number two, and as a one code point grapheme has no code point to extract to be used among the other subscript numbers. Another intuitive choice, '₂', would violate the one grapheme 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 occurrences, 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.  $\bar{A}$ ), the line at the bottom respectively with the combining macron below (U+0331, eg.  $\underline{A}$ ). 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.  $\check{A}$ ), and respectively the superscript with the combining circumflex accent (U+0302, eg.  $\hat{A}$ ). 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] J. T. Faith and R. L. Lyman. *Paleozoology and Paleoenvironments: Fundamentals, Assumptions, Techniques*. Cambridge University Press, 2019.
- [2] S. Hillson. “Tooth Form in Mammals”. In: *Teeth*. Cambridge Manuals in Archaeology. Cambridge University Press, 2005, pp. 7–145.
- [3] The Unicode Consortium. *The Unicode Standard*. <https://home.unicode.org/>. [Accessed: 2024-09-04]. 2024.