Master’s thesis
Master’s Programme in Data Science

Template for Master’s thesis

Riikka Korolainen

September 20, 2024

Supervisor(s):	Professor X or Dr. Y
Examiner(s):	Professor A Dr. B

UNIVERSITY OF HELSINKI
FACULTY OF SCIENCE

P. O. Box 68 (Pietari Kalmin katu 5)
00014 University of Helsinki

HELSINGIN YLIOPISTO — HELSINGFORS UNIVERSITET — UNIVERSITY OF HELSINKI

Tiedekunta — Fakultet — Faculty Faculty of Science	Koulutusohjelma — Utbildningsprogram — Degree programme Master's Programme in Data Science

Tekijä — Författare — Author

Riikka Korolainen

Työn nimi — Arbetets titel — Title

Template for Master's thesis

Työn laji — Arbetets art — Level Master's thesis	Aika — Datum — Month and year September 20, 2024	Sivumäärä — Sidantal — Number of pages 17

Tiivistelmä — Referat — Abstract

Summary of the main contents of the work: topic, methodology and results.

Topics are classified according to the ACM Computing Classification System (CCS): check command `\classification{}`. A small set of paths (1-3) should be used, starting from any top nodes referred to bu the root term CCS leading to the leaf nodes. The elements in the path are separated by right arrow, and emphasis of each element individually can be indicated by the use of bold face for high importance or italics for intermediate level. The combination of individual boldface terms may give the reader additional insight.

ACM Computing Classification System (CCS):
Computing methodologies → Machine learning → Machine learning approaches → Neural Networks
Computing methodologies → Machine learning → Learning paradigms → Multi-task learning → Transfer learning
Applied computing → Physical sciences and engineering → Earth and atmospheric sciences

Avainsanat — Nyckelord — Keywords

Optical character recognition, Few-shot transfer learning, Paleontological databases

Säilytyspaikka — Förvaringsställe — Where deposited

Muita tietoja — Övriga uppgifter — Additional information

Contents

1. Introduction

The thesis should have an introduction chapter. Other chapters can be named according to the topic. In the end, some summary chapter is needed; see Chapter 5 for an example.

2. Figures and Tables

2.1 Figures on teeth

2.2 Background

2.2.1 Neural Networks and Deep Learning

2.2.2 Fundamentals on paleoecology

Basics on ecology

Paleoenvironmental reconstruction

Diets and evolution

Composition of mammal teeth

Since geological events tend to erode organic remains the faster the 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 [2]. 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 [4], 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 *decidous* teeth. Shorthands commonly used for these terms are ’mand’ for mandibles, and capital letter ’D’ for the decidous 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 of a tooth from the throat is described with the terms *distal*, ’far from to the mouth’ and *mesial*, ’close to the mouth’. For skeletal bones, the term *proximal*, ’close to the center of the body’ is often used instead of ’mesial’. Short-form versions for these terms include capital ’L’ or ’Lt’ for left, capital ’R’ or ’Rt’ for right, ’dist.’ for distal and ’prox’ for proximal. 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 superscipt, 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 2.1. Terminology with corresponding shorthands are summarized in Table 2.1 and jaw notation styles in Table 2.2.

2.3 Unicode notation

The unicode system [6] constructs all known characters as signs called graphenes. Each graphene can consist of any number of code points, with each code point having an

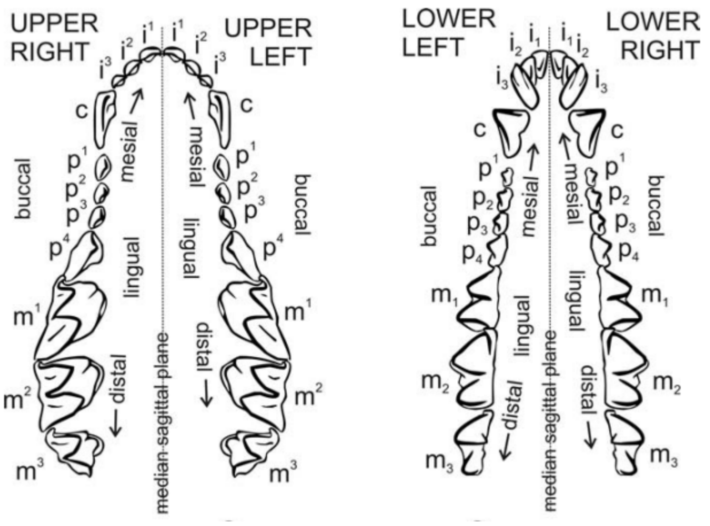


Figure 2.1: Mammal teeth composition, from Hillson [4].

Term	Meaning	Shorthands
Mandible	Lower jaw bone	mand.
Maxilla, Premaxilla	Upper jaw bones	
Deciduous	’Milk teeth’	D, d
Incisor	Tooth type (front, middle)	I, i
Canine	Tooth type (between incisor and premolar)	C, c
Premolar	Tooth type (between canine and molar)	P, p
Molar	Tooth type (back of tooth row)	M, m
Distal	Far from body center / mouth	dist.
Mesial	Close to the mouth	
Proximal	Close to body center	prox.

Table 2.1: Terminology related to mammal teeth with corresponding shorthands

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 ’ω’, ’א’ or ’𝔄’. 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 ’ˆ’ 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

Jaw	Line Notation	Sub/Superscript Notation	Digital Notation
Upper	$M^{\underline{1}}$	m^1	M1
Lower	$M_{\overline{1}}$	m_1	m1

Table 2.2: Dental marking styles, Example: first molar. Line notation displayed in common style combining sub- and superscripts.

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.

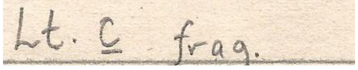
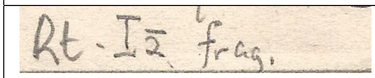
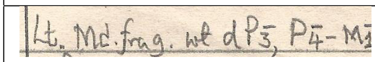
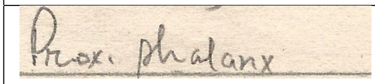
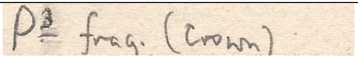
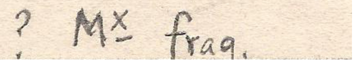
Input Image	Label
	Lt. $\underline{C}$ frag.
	Rt. I$\overline{\check{2}}$ frag.
	Lt. Md. frag. wt dP$\check{\bar{3}}$, P$\check{\bar{4}}$-M$\check{\bar{1}}$
	Prox. phalanx
	P$\hat{\underline{3}}$ frag. (Crown)
	? M$\hat{\underline{x}}$ frag.

Table 2.3: Samples of input images and their corresponding labels.

Data preprocessing

2.4 Figures

Figure 2.2 gives an example how to add figures to the document. Remember always to cite the figure in the main text. There are many ways to cite, for example: University of Helsinki has a nice logo (see Fig. 2.2).

Figure 2.2: University of Helsinki flame-logo for Faculty of Science.

Table 2.4: Experimental results.

Koe	1	2	3
A	2.5	4.7	-11
B	8.0	-3.7	12.6
$A + B$	10.5	1.0	1.6

2.5 Tables

Table 2.4 gives an example how to report experimental results. Remember always to cite the table in the main text. There are many ways to cite, for example: The results are as expected (see Table 2.4).

3. Citations

3.1 Citations to literature

References are listed in a separate .bib-file. In this case it is named `bibliography.bib` with the following content:

```
@article{einstein,
    author =       "Albert Einstein",
    title =        "{Zur Elektrodynamik bewegter K{\"o}rper}. ({German})
        [{On} the electrodynamics of moving bodies]",
    journal =      "Annalen der Physik",
    volume =       "322",
    number =       "10",
    pages =        "891--921",
    year =         "1905",
    DOI =          "http://dx.doi.org/10.1002/andp.19053221004"
}

@book{latexcompanion,
    author    = "Michel Goossens and Frank Mittelbach and Alexander Samarin",
    title     = "The \LaTeX\ Companion",
    year      = "1993",
    publisher = "Addison-Wesley",
    address   = "Reading, Massachusetts"
}

@misc{knuthwebsite,
    author    = "Donald Knuth",
    title     = "Knuth: Computers and Typesetting",
    url       = "http://www-cs-faculty.stanford.edu/%7Eknuth/abcde.html"
}
```

In the last reference url field the code `%7E` will translate into `~` once clicked in the final pdf.

References are created using command `\cite{einstein}`, showing as [1]. Other examples: [3, 5].

Citations should be arranged in alphabetical order by author, using the default style `abbrv`.

3.2 Crossreferences

Appendix A on page 17 contains a code example.

4. From tex to pdf

In Linux, run `pdflatex filename.tex` and `bibtex filename` repeatedly until no more warnings are shown. You should use `pdflatex` when compiling your document.

5. Conclusions

It is good to conclude with some insightful discussion.

Bibliography

[1] A. Einstein. Zur Elektrodynamik bewegter Körper. (German) [On the electrodynamics of moving bodies]. *Annalen der Physik*, 322(10):891–921, 1905.

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

[3] M. Goossens, F. Mittelbach, and A. Samarin. *The LaTeX Companion*. Addison-Wesley, Reading, Massachusetts, 1993.

[4] S. Hillson. *Tooth Form in Mammals*, pages 7–145. Cambridge Manuals in Archaeology. Cambridge University Press, 2005.

[5] D. Knuth. Knuth: Computers and typesetting, circa 2000. `http://www-cs-faculty.stanford.edu/%7Eknuth/abcde.html`, Accessed on 6th March 2018.

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

Appendix A. Code example

Program code can be added as appendix:

```
#!/bin/bash
text="Hello World!"
echo $text
```