Contents

1 Abstract

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

2 Introduction

relevance of this work: any field that does ocr on unconventional characters, or constrained vocabulary. direct relevance to museums digitizing fossil data specifically, but could be any old handwriting.

The rest of this thesis is organized as follows.

3 Deep Neural Networks for Optical Character Recognition

separate problems: character classification (easy, kNN, SVM), reading variable-length text (harder) [3] introduce the problem of ocr, example: fossil catalogue

3.1 Deep Neural Networks: basics

- training/validation/test data sets - neurons and activation functions. maybe examples of activation functions: relu, sigmoid, softmax - feed forward, backpropagation - weight initialization

3.2 Training neural networks

- learning rate - batch size

3.2.1 Loss functions

Loss function is a function from model predictions and ground truth labels that describes with a single scalar value how good the match was, low number describing a good match [4]. These functions are constructed to be equivalent with maximum likelihood solution, think the model would output a conditional distribution of outputs, p(y—x). each ground truth label in the training set should have a high probability in this distribution. Product of all these probabilities is called likelihood. Find parameters that maximize the likelihood of the training data set. Loss functions are derived so that parameters bringing loss to zero is equivalent to the parameters with maximum likelihood. Derivations are out of scope.

- cross-entropy loss kullback-leibler divergence of correct conditional probability and conditional probability parametrized by current model parameters. (show formula, 5.27), correct is not dependent on parameters so is omitted. show 5.29, what is left from that (until here from [4])

then: how cross-entropy loss is computed used for classification problems. eg. is this letter in this image an 'a' or a 'b'. correct probabilities eg .1 and .9 for a or b. model says .2 and .8. discretisized cross entropy computes it as .2*log.1+.8*log.9, log in base 2.

word detection models: have a predefined vocabulary, layer for probability of each word. loss is cross entropy for these probabilities compared to target probability distribution, where correct word has probability 1 and all other have probability 0

- CTC loss

3.2.2 Evaluating model performance

- performance metrics - precision - recall - f1 - cer (character error rate)

3.3 Architectures

3.3.1 Convolutional layers

convolution (cross-correlation) max pooling / average pooling operations

3.3.2 Transformers

- self-attention - multi-head self-attention - tokenizing and the cls token

3.3.3 Autoencoders

- encoder/decoder
    sequence to sequence [5]

3.4 Techniques and heuristics for improving performance

miscellaneous points, like
    - data augmentation

3.5 Transfer learning

basics: what it is initialize weights to those that suit a related task, it is assumed that the starting point is already very good catastrophic forgetting = forgetting the previously learned after finetuning

3.5.1 Foundation models

- generalist models - large unsupervised training data sets

4 Fundamentals on 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?

4.1 Basics on ecology

Nature is highly complicated -¿ models, approximate laws and assumptions enable drawing conclusions from known distributions of species. Assumptions hold usually, not always, in reality but need to be in place because nothing could be known without them
    Tolerances and niches (fundamental + realized): basis for environmental reconstruction [1] ch 2
    Assume that nearest living relative has same tolerances now -¿ get past environment (ch3)
    Presence/absence/abundance: (ch3) species living somewhere -¿ environment matches the fundamental niche. absent more complicated, environment may or may not be suitable (eg a fossil sample was simply not found). abundance estimation harder, but more fine-resolution environmental analysis
    transfer function: paleobiotic data -¿ environment (ch2)

environment is constructed with a coarse resolution such as how many 10cm's of precipitation (ch2)

Theory that the data analysis relies on

4.2 Paleoenvironmental reconstruction

purpose of this section: how the data can be used

fundamental idea: you have a list of taxa occurring in a place at certain geological time. how to map this list to an environment?

presence/absence approach: use information on what species were present, to lesser extent those absent (absence has many reasons so not as reliable) either fix species or place. fix species: find where this set of species lives now (climate maps & areas of sympatry), this indicates that historical locality had this climate. place-fixing: analyze how which species show up in this place changes over time, reduce species showing up data to lower dimension (like pca), eg how many warm-climate vs cold-climate species show up, this gives ideas on changes in climate over time.

4.3 Diets and evolution

maybe, how the data is used

4.4 Composition of mammal teeth

Fossils occur when animal / plant remains are deposited in a sediment in a way that preserves some part of its original form. Since teeth are the hardest material in animals, large fraction of found parts are teeth. Fossil finding is followed by identification to most specific taxon possible largely a technical skill (ch5), teeth are identified down to type and number, how manyeth the teeth are, counting from center to edge or other way round?? specimen can be either one tooth or fragments of the jaw bone where there are multiple teeth (markings like M1-3)

from [2] what teeth are composed of

the jaw bones lower jaw bones: mandibles permanent and deciduous (D), nonpermanent "milk" teeth (laita vaan jos löytyy d-hampaita)

right and left sides are always symmetrical, denoted simply L or R or Lt or Rt or left or right. left is left looking from the animal, not the observers perspective Identicality also causes that sometimes tooth fossils are misidentified to the wrong side and corrected (ei lähteestä vaan nähty datasta koska l ja r on sutattu aika monta kertaa ja vaihettu

four classes, front to back: three incisors (I), one canine (C), four premolars (P), three molars (M). top bottom left right. top/bottom noting upper jaw as superscript lower jaw as lower script, purpose: incisor -¿ catching, canine -¿ stabbing / killing prey, molars are for chewing. premolars are bit like canines bit like molars, function varies lot between taxa including holding, cutting and chewing. also form and number of each present changes between taxa. sometimes lower jaw as line on top and upper jaw as line on bottom, sometimes both are used: upper script number with line on bottom. Line is "the other jaw" if there are less of a type of teeth eg two premolars, they might be no 1 and 2 or no 3 and 4

hillson image on mammal teeth

5 Experimental setup

5.1 data description

has been done by different annotators, no logs on who logged what, everyone had a bit different style of notating. also no clearly defined standard for notating specimens. so might be that actual data used will have characters or words not present in any data, causing errors.

Identicality also causes that sometimes tooth fossils are misidentified to the wrong side, seen in data with smudged over l's and r's

Catalogues have lines between entries. challenge for the model to not mark these as underlined in these cases the line is long and spans the entire image, so it should be distinguishable from a single underlined character.

5.1.1 Notes on creating the dataset

Hand-labeled

Data was extracted from scans by getting bounding boxes from Azure Vision API, finding the correct column (nature of specimen or element), and cropping the image according to bounding boxes.

Non-tooth samples were not discarded since they contain bone fossil related words and good samples of the handwriting style of this dataset.

smudged-over "L" was labeled as "R", and other way around: it seems that later someone found it was the opposite side after all. Hope of this is that the model would learn to map "messy L" as "R". snudged "left" or "right" was not noted as the opposite as there were too few such samples.

Superscript seems much more rare than lower script

Data was labeled not by individual characters but as full tooth descriptions to preserve context where tooth special characters are more likely to occur

Some have been corrected by writing on top and thus are very hard even for humans to read, this is also an example of smudged-over correction:
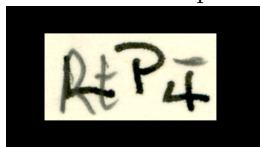
5.1.2 Unicode characters used for data labeling

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 ($\bar{A}$) and upper variant.

Unicode [6] 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 caron (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 "^" 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 (until now already in thesis text) —-

also: some annotators used / instead of line. left to / -¿ upper, right -¿ lower. annotate with up/down macron

5.2 Data preprocessing

Convert to black and white since text reading should not change when color of writing / background changes?

Convert to background completely white, foreground completely black? Either there is a line or not?

5.3 Methods: base models and transfer learning tehniques

5.3.1 Encoding prior knowledge

Priors. base model already knows the output should be a word, eg "jdaslkjflkds" is a highly unlikely correct answer. Bone notation has a very small subset of possible english words, eg. the word "beach ball" cannot ever be a correct answer for a reading

6 Results and discussion

7 Conclusions

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] 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.

[4] S. J. Prince. *Understanding Deep Learning.* The MIT Press, 2023. URL: http://udlbook.com.

[5] I. Sutskever. "Sequence to Sequence Learning with Neural Networks". In: *arXiv preprint arXiv:1409.3215* (2014).

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