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) [6] 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 [7]. 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 [7])

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 [9]

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

Nature is highly complicated -¿ models, approximate models and assumptions enable drawing conclusions from known distributions of species.

each assumption / model can be questioned but they hold as a rule. only models briefly presented here, all statements here can be questioned to some extent

idea: what fossil/dental data can be used for. how fossil/dental data is used

to highlight why accurate, fine-resolution (ie specific) and large magnitude of fossil, esp dental fossil data is genuinely useful.

chapter overview: review ecology and assumptions the analysis is based on. then, short overview of main techniques for paleoenvironmental reconstruction, the main application area of fossil data. last, mammal teeth row is presented to introduce terminology present in the data.

4.1 Basics on ecology

basic laws: theory that the data analysis relies on

Tolerances and niches (fundamental + realized): basis for environmental reconstruction [3] ch 2

tolerance = range of an environmental variable that is hospitable for the species, eg. imaginary small mammal (come up with some imaginary name) can live when avg temperature is +10-+15. Niche = set of tolerances the species has. fundamental niche = possible environments for the species, realized niche = where it lives. center of tolerances is better than the edge (ch2)

main assumption uniformitarism (the fact that tolerances constraint things has not changed) niche conservatism: Assume that nearest living relative has same tolerances now -¿ get past environment (ch3, lyman 2017),

this presents a mapping from taxa to environmental variables -¿ basis of analysis of past environments.

modern alternative to this: transfer functions: mappings from taxa data to enviroment learned using machine learning / statistical models (ch9, birks 1995)

benefit instead of tolerances/niches: they have subjective interpretation problems (book ch 9)

esp. teeth: dental ecometrics = inference of transfer functions given dental data (ch9 liu et al 2012)

next, turn to how to solve the problem of information to environment given the taxomic information to environmetal indicators mapping

4.2 Paleoenvironmental reconstruction

why: get information of what is to come with climate change (faith ch2)

definition of the problem: whan ancient habitats were like and what changes they underwent at which times (ch2)

overview main tehcniques: presence/absence, abundance, taxon free, diversity based, size clines

presence absence (ch5): dataset is list of taxa that are present. absence is also indicator but worse since might be that the species just was not preserved / found. two approaches: fix location or fix set of species. 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

abundance get relative NISP (number of species in sample): percent of species in sample is of this taxon (grayson 1984b, ch6) do like presence absence but weigh signal according to abundance and assume abundant species lives more toward center of tolerance grayson 1981: fossil accumulation affects datasets -¿ use with caution

taxon free cornelis van der klaauw 1948 (p 160) relation of traits of animals and livelihood = ecomorphology: eg what the animal eats also eg how diverse the place is. for environment mostly diet -¿ what plants grew and habitat -¿ climate etc helps with not having to assume that closest relative tolerances were the same. (ch7) eg usage of teeth: dental microwear. calandra and merceron 2016: analyze miniscratches on the surface to get diet to get vegetation and climate oma: to conduct this you need many samples of the same bone need to know exactly which bone it is since different teeth used for different things

diversity and size clines as enviromnental indicators andrews et al 1979: how diverse also has a mapping to environment eg tropical is more diverse than arctic coarse indicators of the environment

lyman 2008: strong sample size effects: bigger sample is more diverse so to do this you need equal size samples from all time periods analyzed size clines basic idea: coarse indicators for size of bone -¿ environment eg larger libs warmer climate (mayr 1970) also dental measures correlate with environment, eg mandibles (faith et al 2016). for that you need sufficient tooth samples and correct identification which tooth it is

end lesson: sample sizes and data precision are important. therefore it is superimportant to get more data / improve accuracy, which is the main goal of the digitization effort. scale sense: faith lyman ch 4 said 1000 is solid 10 000 whopping sample size dataset in question in this work has 90,000 so the value is pretty clear

4.3 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 [4] 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

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.

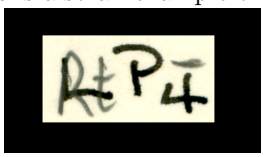
Non-tooth samples were not discarded since they contain bone fossil related words and good samples of the handwriting style of this dataset.

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

Unicode [10] 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 ($\check{\text{A}}$) was chosen as the lower script modifier, and the circumflex accent ($\hat{\text{A}}$) 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 (until now already in thesis text) —-

also: some annotators used / instead of line. left to / -¿ upper, right -¿ lower. unknown/unsure noted with x with macrons on top/bottom annotate with up/down macron

if model is toothornot classifier + tooth reader -¿ remove fractions notes

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

the sequence or character per character recognition question: sequence is a mapping from image to variable-length phrase character per character approach, inspired by [11]: train a classifier: is this word a tooth or not? then give non-tooth to the untuned trocr, which works very well. Then few shot transfer learn a classifier from an image with only a tooth marking (letter and one number) to tooth. Possibly extend classifier to be able to recognize multiple teeth (eg M1-3). Target could be multivariate: first would be tooth (which i1-3,c,p1-4,m1-3), second would be jaw (upper, lower, unknown), third side (left, right, unknown). Separating 'l/r/lt' from the letter+number tooth notation is fairly trivial: noncursive handwriting can be separated by finding a vertical line where there is no black. Image processing: convert to black and white, then find x coordinate with no black and split there.

sequence benefits: adaptable to many kinds and lengths of input, possible to get good inferences for surprising marking styles sequence bad sides: finetuning just one layer on 80 training images for two epochs took about 15 minutes -¿ all hyperparameter optimization etc is out of question with this heavy training. Also: not domain adaptation (adapting same task to different dataset), but task adaptation ie. the target set of characters has changed. Encoding this to the large encoder decoder transformers would require rewriting parts of the preprocessor and model which is too complex given the level of this work.

character by character benefits: feasible given available data and computing resources possible to encode the classes (ie, teeth) classifying characters has been essentially solved, easy problem also classifying to tooth or not tooth should be easy –¿ focus on model ensemble with tooth or not classifier + trocr + tooth classifier

5.3.1 Base model selection

initial sequence learning attempts done with [5], proof of concept so no real exhaustive comparison literature review

Chosen public dataset was EMNIST [2] because it is the closest to my problem: a large dataset of labeled letters and numbers.

According to [1], this model was the best: [8], was chosen as the base model. they trained on emnist-balanced, so no difference between upper -and lower case letters. not a problem for us since in the handwritten catalogues, upper or lowercase was not used to distinguish upper and lower jaw

5.3.2 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] A. Baldominos, Y. Saez, and P. Isasi. "A survey of handwritten character recognition with mnist and emnist". In: *Applied Sciences* 9.15 (2019), p. 3169.

[2] G. Cohen, S. Afshar, J. Tapson, and A. Van Schaik. "EMNIST: Extending MNIST to handwritten letters". In: *2017 international joint conference on neural networks (IJCNN)*. IEEE. 2017, pp. 2921–2926.

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

[4] S. Hillson. "Tooth Form in Mammals". In: *Teeth*. Cambridge Manuals in Archaeology. Cambridge University Press, 2005, pp. 7–145.

[5] 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].

[6] 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.

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

[8] A. Shawon, M. Jamil-Ur Rahman, F. Mahmud, and M. Arefin Zaman. "Bangla Handwritten Digit Recognition Using Deep CNN for Large and Unbiased Dataset". In: *2018 International Conference on Bangla Speech and Language Processing (ICBSLP)*. 2018, pp. 1–6. DOI: 10.1109/ICBSLP.2018.8554900.

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

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

[11] G. Zhao, W. Wang, X. Wang, X. Bao, H. Li, and M. Liu. "Incremental Recognition of Multi-Style Tibetan Character Based on Transfer Learning". In: *IEEE Access* 12 (2024), pp. 44190–44206. DOI: 10.1109/ACCESS.2024.3381039.