DeepLearning in Stemmatology

Armin Hoenen

June 12, 2022

Outline

1. General Part
   - What is DeepLearning?
   - Deep Learning in phylogenetics
   - Deep Learning in stemmatology?
2. Experiment
   - Introduction & Related Work
   - Methods & Materials
   - Results
   - Discussion & Conclusion

DeepLearning=?

i.p.o. a definition: 1.1: DeepLearning ex Wikipedia

Deep learning is part of a broader family of **machine learning methods based on artificial neural networks**

i.p.o. a definition: 1.2: Machine Learning after Krohn et al. (2019)

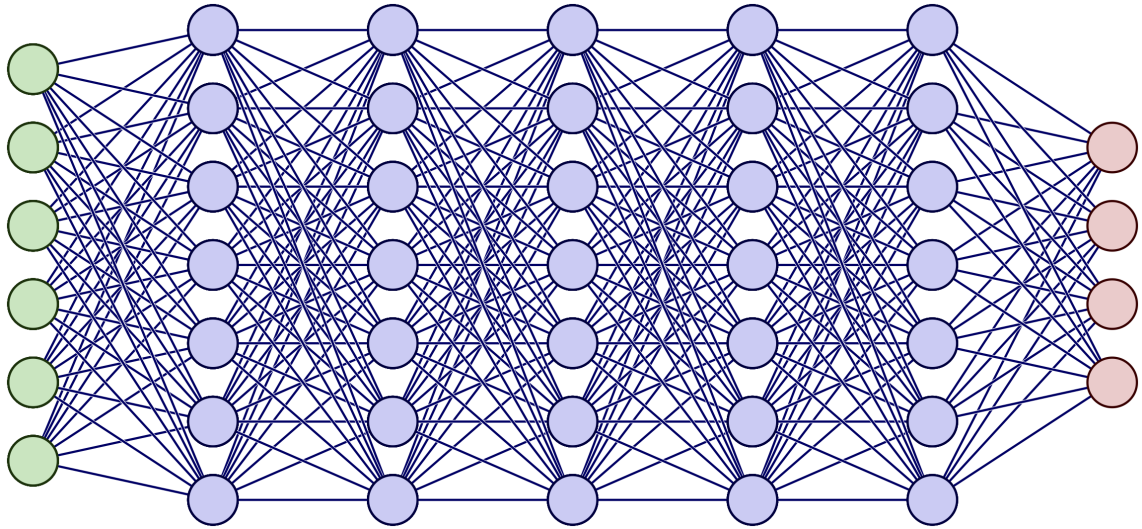
Subfield of AI where patterns are extracted from so-called training data often to classify new data.

Roughly 3 paradigms/eras:

- **potpourri**: until the 1990ies
- **classical ML:** from the 1990ies until roughly 2012
- **deep learning** revolution: since 2012

Deep Neural Network

based on regression (or perceptrons), $\sum_{i=1} w_i x_i$, 



$(0.7 \cdot 17) + (0.3 \cdot 2)$, > 15? $\rightarrow$ male $\rightarrow$ SHOOT

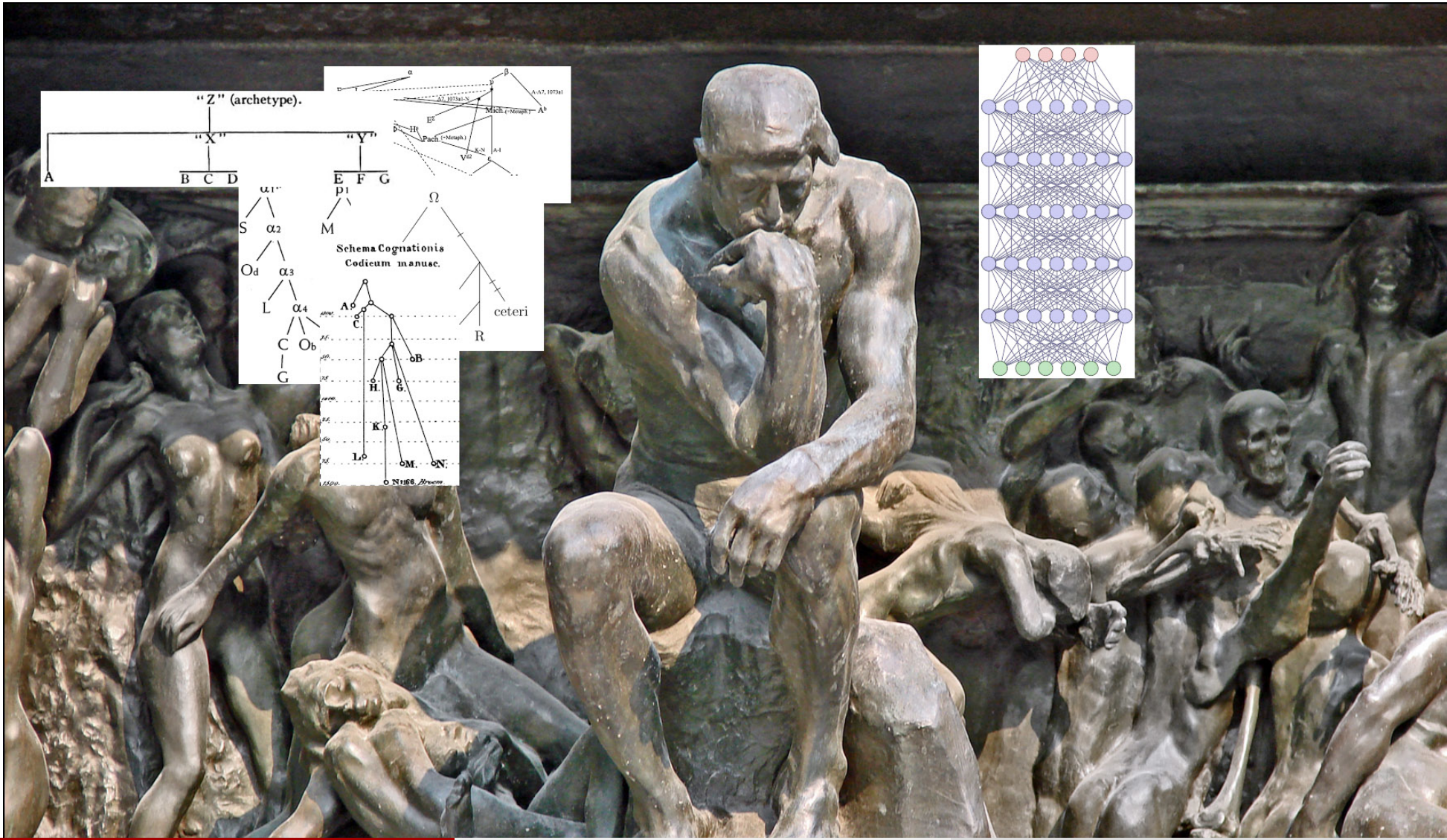
Deep Neural Network

DeepLearning Revolution

A number of tasks have witnessed a break-through or large improvement thanks to DL

- machine vision (object recognition etc.)
- machine translation
- speech-recognition
- text classification
- sentiment analysis
- board-games: AlphaGoZero

And stemmatology?



... in other disciplines: phylogenetics

2022: nature communications survey on DL in biosciencess

Table 1 Impact of Deep Learning on Computational Biology.

	Protein structure prediction	Protein function prediction	Genome engineering	Systems biology and data integration	Phylogenetic inference
Paradigm shifting	✓				
Major success		✓	✓		
Moderate success				✓	
Minor success					✓

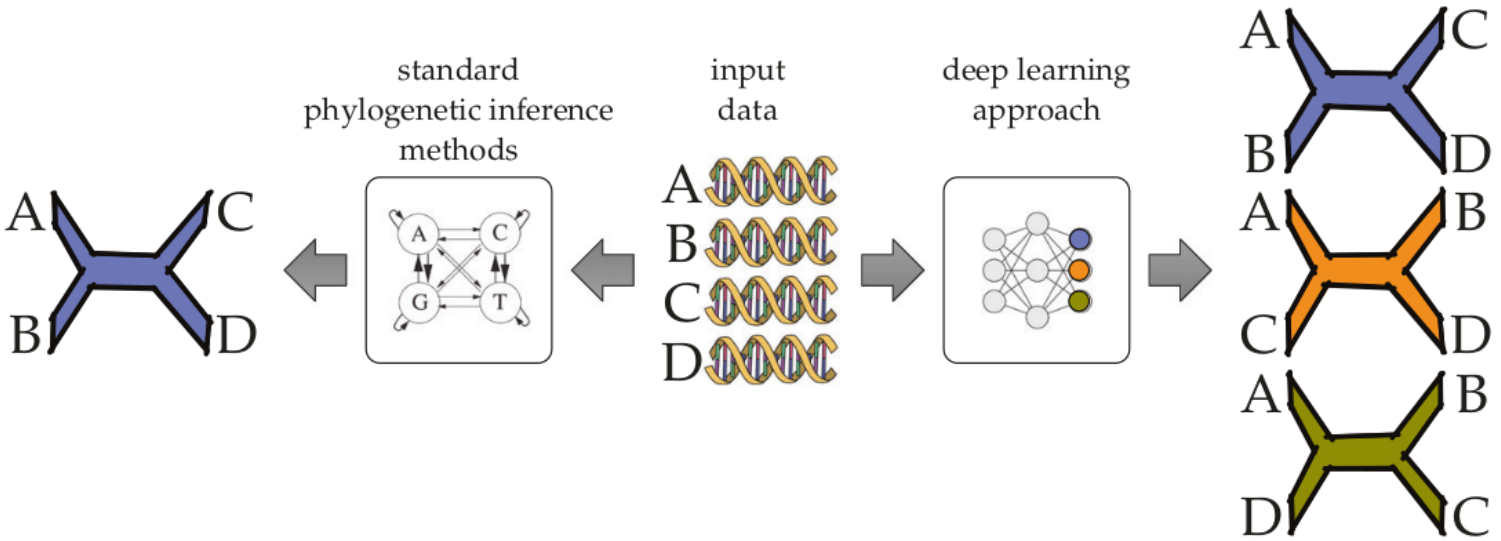
"it is difficult to conceive of an end-to-end DL model to directly estimate phylogenetic trees from raw data in the near future"

Well, what is the problem?

... in other disciplines: phylogenetics

A problem is that the tree space of possible trees grows unfeasibly for output nodes. Hoenen et al. (2017): for Greg Trees representing 10 manuscripts already 102 515 201 984!

"Recently CNNs have been used to infer the unrooted phylogenetic tree on four taxa ... an analysis of the performance of the method ...shows that CNNs were not as accurate as other standard tree estimation methods, e.g., maximum likelihood, maximum parsimony, and neighbor joining"



"classifiers like DL models require training data, and benchmark data where the true phylogeny is known is almost impossible to obtain in this field [phylogenetics]. Instead, simulations have been the method of choice for generating training data" $\rightarrow$ *still doesn't solve the tree-space problem*

... let's take a minute to think

What input and what output to a neural network could work?

- INPUT - OUTPUT pairs
  - collation (MSA) [one ms per node] - a stemma

This would be ideal, but

- treespace problem
- we would need different collations of the same data leadingh to different outputs $\rightarrow$ e.g. different 4 ms-subtrees of a larger stemma, BUT: what would we classify then with the trained network?

For biological data with a universal code across species augments the number of application scenarios. Simulating mutation on 4 DNA letters with known imbalances (purin-purin, pyrimidin-pyrimidin mre probable) is far easier than mutating words. We need a fresh start..

The philologists game

A slightly more complex technology is so-called **reinforcement learning**. The case of Go:

- more board configurations than estimated atoms in the universe
- 10^{100} = one Gogol more complex than chess

AlphaGoZero by Google DeepMind solved that problem and outperforms humans by learning playing against itself.

Now, what if defining stemma building as a game?

1. start from a random edge you draw between any two witnesses with recurrence to their texts (first move)
2. successively attach more nodes never violating a DAG (move rules) until no more nodes left
3. receive a score equal to the TreeEditDistance to the true stemma (there are far fewer TEDs than tree topologies)

Although such scenarios might work, how would such a trained network ever be used for another tradition?

Preparing for doomsday

In case any other scientits should find a feasible scenario for DL and stemmata,

- ... I implemented a prototype of a simulated data generator
- A random generator draws outdegrees and simulates a stemma (Hoenen 2015)
- starting from root, each node gets an input text which gets copied by an artificial scribe confusing random letters according to a confusion matrix (Geyer 1977) at a certain error ratio and mostly within class vowel or consonants
- if a word upon copying is not in a large english token lexicon (open subtitles), a word with minimal DL will be substituted as correction

Preparing for doomsday

Examples of the Human rights declaration as input, 5622 nodes

- 3rd generation all men are born free anl equal, in dignity and in rights,
- 4th generation: all men are born froe and equal, in dignity and in rigbts,
- 6th generation: aal men are borp lree anl equal, in dignity and in rights,
- aal men ere born free and equel, in dignity and in rights,
- juridical $\rightarrow$ auridical, juradical

In any case, if someone needs a large simulated corpus with collations and stemmata for training of a DL algorithm...

Finally, really bringing DL to stemmatology

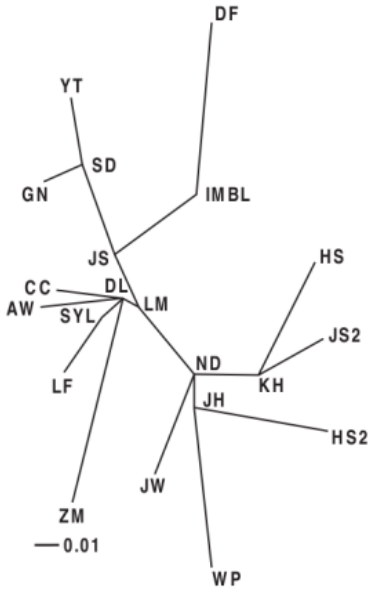
Can the *language of the collation* be translated into *the language of the stemma*? Using Machine Translation for witness localization.

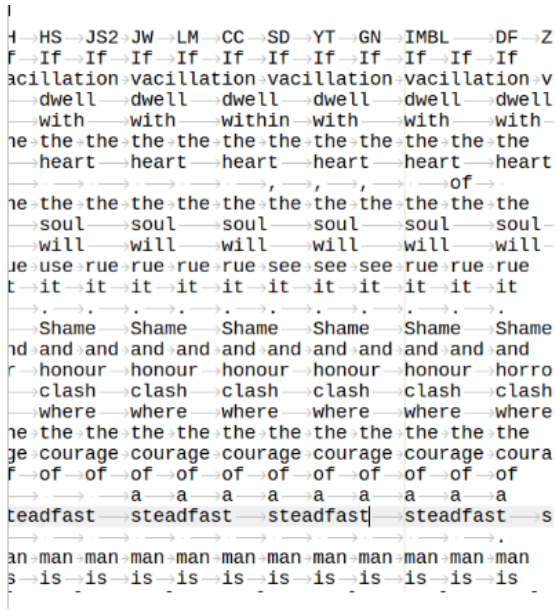
arxiv.org

Introduction

Phylogenetic placement...

- ...is the task of localizing a new node on a known tree
- ...may help reassure or find position of uncertain nodes/witnesses
- ...cannot localize texts from successive contamination





Related Work

1. phylogenetics
   - standard approach is to generate a new tree analysing placement (the tree can however be different from the previous one)
   - Jiang et al. (2021) have introduced DEPP, a DL software which learns distance relations so as to place a new node correctly facing the problem of gene vs. species trees, where the former may diverge because of lateral gene transfer
2. stemmatology
   - tree generation approach executed, especially to investigate successive contamination

$\rightarrow$ Let's use something completely different

Algorithm

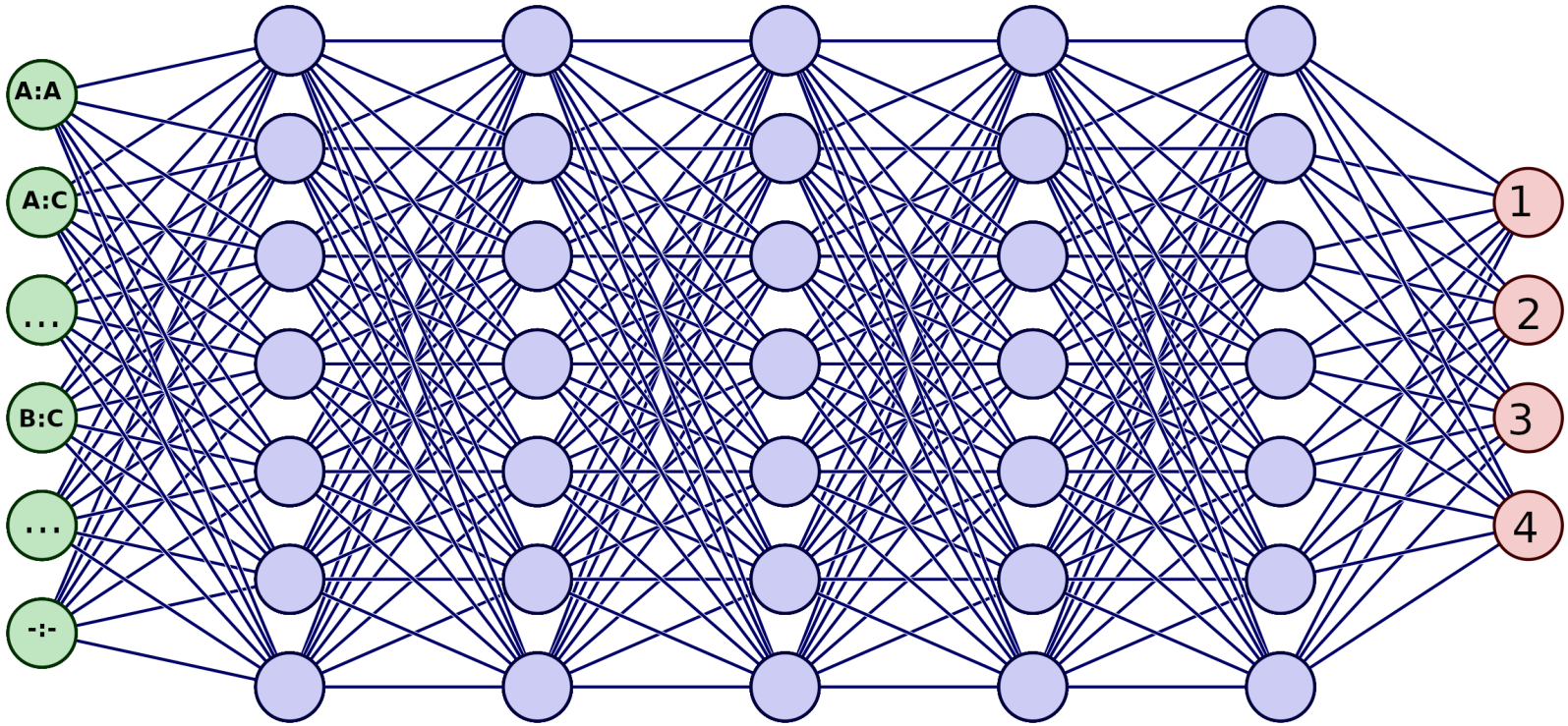
We use a **sequence to sequence** DL approach:

- from tree with assumed/known stemma, hold back any one leaf node
- train a DL model with input = pairwise collation, output distance (number of edges on stemma)
- let model estimate length between held back node and any other on the tree
- use a scoring to place node from estimates
  - only one node truly has edge dist 1: the parent; if also only one in the DL estimates, take that node
  - else: for each node + estimate localize and score each node with estimated difference; retrieve score winner
- evaluate placement and estimate accuracy

Visually I


Edges
pairwise position delta input
edge-dist
CC → AW
DF → AW
DL → AW
GN → AW
HS → AW
HS2 → AW
IMBL → AW
JH → AW
JS → AW
JS2 → AW
JW → AW
KH → AW
LF → AW
LM → AW
ND → AW
SD → AW
SYL → AW
WP → AW
YT → AW
ZM → AW
A:A A:A A:A A:A A:A A:A -:- A:A A:A A:A A:A
A:A A:A A:A A:A A:A A:C -:D A:A A:A A:A A:A
A:A A:A A:A A:A A:A A:A -:- A:A A:A A:A A:A
A:A A:C A:A A:A A:A A:A -:C A:A A:A A:A A:D
A:A A:A A:A A:A A:A A:A -:- A:A A:A A:A A:C
A:A A:A A:A A:C A:A A:A -:- A:A A:A A:A A:A
A:A A:A A:A A:A A:A A:C -:- A:A A:A A:A A:A
A:A A:A A:A A:A A:A A:A -:- A:A A:A A:A A:A
A:A A:A A:A A:A A:A A:C -:- A:A A:A A:A A:A
A:A A:A A:A A:A A:A A:A -:- A:A A:A A:A A:A
A:A A:A A:A A:A A:A A:A -:- A:A A:A A:A A:A
A:A A:A A:A A:A A:A A:A -:- A:A A:A A:A A:A
A:A A:A A:A A:A A:A A:A -:- A:A A:A A:A A:A
A:A A:A A:A A:A A:A A:A -:- A:A A:A A:A A:A
A:A A:A A:A A:A A:A A:A -:- A:A A:A A:A A:A
A:A A:A A:A A:A A:A A:A -:- A:A A:A A:A A:A
A:A A:C A:A A:A A:A A:A -:C A:A A:A A:A A:D
A:A A:A A:A A:A A:A A:A -:- A:A A:A A:A A:A
A:A A:A A:A A:A A:A A:A -:- A:A A:A A:A A:A
A:A A:C A:A A:A A:A A:A -:C A:A A:A A:A A:D
A:A A:A A:A A:A A:A A:A -:- A:A A:A A:A A:A
2
5
1
5
5
5
4
4
3
5
4
4
3
2
3
4
2
5
5
2

Visually II



Materials - Data

Artificial datasets - mainly Parzival.

Datatype	**Value**	**Note**
Language	English (from German, archaic)	
Copy-Mode	manual, written	no contamination
n. of mss	21	
n. of leafs	12	
n. of rows	958	
max dist leaf-leaf	6	edges
Publication	Spencer et al. (2004)	

Materials - Machines & Sofware

- **data preparation, evaluation**: Java
- **Neural Architecture**: python ONMT framework with pytorch
- **GPU**: Nvidia-GForce 4GB $\rightarrow$ GPU hours per run on Parzival roughly 10 hours for most tested architecture (512 rnn, 128 wv, 1 layer)

Results - Main Result

On BRNN with 512 rnn-size, 128 word vec size, 1 layer, no dropout, 5:185 validationset proportion, batch-size 16:

Feature	**Value**	**Note**
correct predictions	111/240 (0.46)	best 135(0.56)
average deviation	0.6 (SD:0.6)	best 0.5 (SD:0.6)
max dist	3	best 2, absolute max dist 5
hitrate localizations	9.5/12 (0.79)	distance of 3 misclassified from parent: 1
Random baseline	**100.000 iterations**	
correct predictions	40/240 (0.17)	max 67, diff sign.
average deviation	1.85 (SD:1.4)	
max dist	5	10 times not 5 (0.0001)

Results - Architecture I

Architectures tested:

- **Bi-LSTM**: too slow accuracy rise
- **BRNN**: quickest and most accurate
- **transformers**: too slow accuracy rise

BRNN represent a slightly simpler way than Bi-LSTM and Transformers to train contextual data. Accuracy may rise for those if trained through. Thus: results are preliminary with respect to architecture.

Results - Architecture II

- RNN sizes: 512, 256
- word vector sizes: 128, 64
- layers: 1
- rnn-512, wv:128, 1 layer: trains 4 345 867 parameters

Problem often is, that a grid search would require unfeasible GPU-time: testing all configurations for 3 RNN sizes and 3 word vector sizes @ 1 layer, would need approx. 4(?) days [including nights], sufficient memory presupposed.

From accuracy curves (and validation) and paired with the need for feasible computation times and loads, 7000 training steps for a 512 rnn size, 128 word vec size architecture with one layer was deemed best.

Results - data configuration

Various input variations were tested and compared:

- binary input, variant letters (sorted and unsorted), actual words [not tested: letter-correspondences]: **best variants sorted** (A:C and C:A both encode A:C)
- only places of variation vs. all places: **all places has an advantage**
- validation set size: 5-185 and 10-180: **no significant difference (?)**
- rnn and wordvec sizes: *slight hints that lower dimensions may work better in this setting but could distinguish less between possible placements (depending on tradition size probably)*

(Pre)Discussion I

Better architectures probably around (more computation time and load, more brnn layers, transformers etc.).

- DL method already clearly better than chance in estimating (also maximum distance clearly smaller)
  - **In the collation there is edge-distance info**, not just text distance info: DL is a way to look at collation context
- all-places better than only-places-of-variation
  - hint to information also within what and how much is shared (which a computer might exploit better or more objectively than humans reading for difference)

(Pre)Discussion II

- variants encoded, sorted best: Why? Is the 'concentration profile' of scribes that which is most informative? Does the extact error that is being made matter less than WHERE or at copy-time WHEN it is made (with plenty of possible errors once concentration slips)? Is sorted better because it generalizes polarity between corrections and errors? Could this mean some generalizability between traditions (?). FOLLOW-UP:
  - using Parz AW-7000 model for NB tradition → worse than chance but maxdist lower
  - parz classifying heinrichi subtree: failed again
  - heinrichi node trained to classify all parz nodes: slightly but significantly better than chance - from large enough tradition there could be some transferability downwards

Conclusion

We have...

- ...discussed some of the problems DL approaches suffer in stemmatology and related disciplines
- ...demonstrated a DL approach for stemmatology (first?, despite hard terrain) albeit for phylogenetic placement not tree generation - not yet entirely optimized
- ...discussed various interpretations and questions the approach implies for future research

Farvel, Farewell