Heidelberg University
Institute of Computer Science

Project report for the lecture Advanced Machine Learning

Prediction of the next SARS-CoV-2 variants

https://github.com/nilskre/AML-covid-project

Team Member: Felix Hausberger, 3661293,
Applied Computer Science
eb260@stud.uni-heidelberg.de

Team Member: Nils Krehl, 3664130,
Applied Computer Science
pu268@stud.uni-heidelberg.de

Plagiarism statement

We certify that this report is our own work, based on our personal study and/or research and that we have acknowledged all material and sources used in its preparation, whether they be books, articles, reports, lecture notes, and any other kind of document, electronic or personal communication.

We also certify that this report has not previously been submitted for assessment in any other unit, except where specific permission has been granted from all unit coordinators involved, or at any other time in this unit, and that we have not copied in part or whole or otherwise plagiarized the work of other students and/or persons.

Member contributions

Nils Krehl

tbd

Felix Hausberger

tbd

Contents

List of Abbreviations

CUDA	Compute Unified Device Architecture
DQN	Deep Q-Networks
ELU	Exponential Linear Unit
ICAART	International Conference on Agents and Artificial Intelligence
MDP	Markov Decision Process
ReLU	Rectifier Linear Unit
PER	Prioritized Experience Replay
PPO	Proximal Policy Optimization

0 Project Setup

TODO: update For a detailed description of how to set up the project, please have a look at `https://github.com/nilskre/bomberman_rl/blob/master/README.md`.

1 Introduction

tbd

2 Fundamentals and Related Work

tbd

3 Approach

TODO: Pipeline image

3.1 Dataset creation

hier irgendwo das Zielschema des Datensatzes beschreiben

3.1.1 Raw data selection from GISAID

focus: Germany from 4.5. - 6.8. (new variant arised in the recent past -> lambda, delta, ...) only Germany, to make it possible to handle the data about 35000 genomes in our raw dataset

beispiel record: genome sequence and metadata

3.1.2 Generation of a phylogenetic tree

3.1.3 Phylogenetic tree to dataset

3.2 Data Preprocessing

two steps: - to make the dimensionality managable not the whole 30000 nucleotides are evaluated. We take a subpart of X nucleotides from position A to B - Transform string to numeric for model input

3.2.1 Dimensionality reduction by selecting subpart of the genome

3.2.2 Transform genome sequence to numeric model input

3.3 Model architecture

3.4 Training process

4 Experimental results

tbd

5 Conclusion

tbd