Maastricht University

Faculty of Science and Engineering

Project title: Ordering of training data for accurate, fast, and robust deep learning

Project leader (co-promotor): Dr. Christof Seiler

Function: Assistant professor and faculty member at the Dept. of Data Science and knowledge Engineering (DKE), Maastricht University, The Netherlands.

Promotor: Gerhard Weiss, Full professor at DKE, Maastricht University, The Netherlands.

Introduction:

Training of deep neural networks is time-consuming and requires enormous amounts of data for accurate predictions. In theory, the larger the sample size, the more accurate our predictions. In practice, we usually aim for a tradeoff between accuracy and computational time. Various ad-hoc heuristics to balance this tradeoff are available, e.g., tuning deep network architectures and optimization methods. The disadvantage is that they are tailored towards particular datasets or applications, and thus are limited in scope.

Hypothesis and Objectives:

This PhD project focuses on the accuracy-time tradeoff by exploring shuffling procedures. These shuffling procedures occur prior or during learning—randomly ordering of the training data or at every optimization step. Random ordering of observations can avoid local minima and prevent biases at the expense of statistical power. The loss of power is due to ignoring the innate structure of a dataset and possible relationships among features. Moreover, random shuffling contradicts with the way a human might learn. Evidence from cognitive psychology suggests that humans "learn in small steps" and "learn from simple to difficult samples". The latter example implies that humans naturally apply a non-random ordering of the data during learning. The core hypothesis underlying this project is that non-random ordering helps not only humans but also deep learning procedures to make more accurate and faster predictions. We further hypothesize that non-random ordering procedures make learning procedures more robust by reducing the influence of individual observations. Our objective is to investigate these hypotheses and to develop non-random ordering methods that can be applied in deep learning libraries.

Setting and Methods:

The PhD project aims at a broad theoretical and experimental setting that covers a variety of standard network architectures, learning methods, and datasets from different application domains. The project does not need to start from scratch. It builds on existing neural network learning approaches that deal with ordered training data. Current approaches are: (1) curriculum learning, (2) learning based on lowest-loss and importance sampling, and (3) learning based on sample re-weighting. The student will explore the potential of (1-3) on simulated and real data. The student will leverage classical work in statistics on (4) robust regression and (5) order statistics to improve our theoretical understanding of these methods. As an overall outcome of the project, the student will develop ordering procedures for accurate, fast, and robust deep learning, and will make them available to the machine learning community by releasing Python and R packages.

Impact:

Higher accuracy and robustness will result in more reliable and trustworthy deep learning procedures. Lower computational time will result in a lower carbon footprint and help the environment.

Requirements candidate: Strong interest in machine learning, statistics, and deep learning. Familiar with Python or R. Solid background in linear algebra, calculus, and probability theory. Excellent verbal and written communication skills in English.

Keywords: Deep learning, robust learning, and order statistics.

Top 5 selected publications:

- Beware of the Simulated DAG! Causal Discovery Benchmarks May Be Easy To Game, A. Reisach, **C. Seiler**, S. Weichwald. Neural Information Processing Systems (NeurIPS), 34:27772–84, 2021
- CytoGLMM: Conditional Differential Analysis for Flow and Mass Cytometry Experiments. **C. Seiler**, A.-M. Ferreira, L. Kronstad, L. Simpson, M. Le Gars, E. Vendrame, C. Blish, and S. Holmes. BMC Bioinformatics, 22(137):1–14, 2021
- B. Ranjbar-Sahraei, H. Rahmani, **G. Weiss** and K. Tuyls. Distant Supervision of Relation Extraction in Sparse Data. Intelligent Data Analysis, 23(5):1145-1166, 2019
- Multivariate Heteroscedasticity Models for Functional Brain Connectivity. **C. Seiler** and S. Holmes. Frontiers in Neuroscience, 11(696):1–11, 2017
- Positive Curvature and Hamiltonian Monte Carlo. **C. Seiler**, S. Rubinstein-Salzedo, and S. Holmes. Neural Information Processing Systems (NIPS), 27:586–594, 2014