Plan Overview

A Data Management Plan created using DMPonline

Title: Sustainable learning for Artificial Intelligence from noisy large-scale data

Creator: Kim Batselier

Principal Investigator: kim batselier

Project Administrator: kim batselier

Affiliation: Delft University of Technology

Funder: Netherlands Organisation for Scientific Research (NWO)

Template: Data Management Plan NWO (September 2020)

ORCID iD: 0000-0001-7381-2630

Project abstract:

Computer models are an essential tool of modern society. Whether it is for designing airplanes, predicting dominant virus strains in a pandemic or estimating how different policies will impact the CO2 concentration in the next 5 decades, our society makes abundant use of models. While some models can be built from first principles, the majority of models are learned from data. Such an artificial intelligence learning process consists of processing many examples. The computational power needed to learn large models has doubled every 3.4 months since 2012. In 2019, learning
a single model could emit as much carbon as five cars in their lifetimes. This ever-increasing need for computational power is driven by the large amounts of model parameters that can only be reliably learned from equally large data sets of high-dimension and is simply unsustainable.
In this project, I will lay the foundation for a new paradigm where instead of relying on ever-increasing computational power, the focus shifts towards a smart exploitation of the currently-available computational power. My breakthrough concept to enable this revolution lies in replacing models by tensor networks, a novel technique to efficiently represent nonlinear functions. I will develop groundbreaking open-source algorithms that use tensor networks to efficiently learn models from high-dimensional and large-scale datasets. Millions of model parameters will be learned from noisy data in a matter of seconds on a conventional laptop or desktop computer. My sustainable learning algorithms will use
Bayesian statistics to deal with uncertain noise in the data and to generate confidence bounds on model predictions. Both my academic and industrial expertise in tensor networks and learning models from data, together with the promising results of my preliminary studies are essential to successfully completing this project.

ID: 104007

Start date: 06-01-2023

End date: 06-01-2028

Last modified: 16-12-2022

Grant number / URL: VI.Vidi.213.017

Copyright information:

Sustainable learning for Artificial Intelligence from noisy large-scale data

General Information

Name applicant and project number

Kim Batselier
VI.Vidi.213.017

Name of data management support staff consulted during the preparation of this plan and date of consultation.

Bjorn Peare Bartholdy
7/12/2022

1. What data will be collected or produced, and what existing data will be re-used?

1.1 Will you re-use existing data for this research?

If yes: explain which existing data you will re-use and under which terms of use.

- No

No, there is no data available addressing this issue.

1.2 If new data will be produced: describe the data you expect your research will generate and the format and volumes to be collected or produced.

We will not generate new data. Instead, we will rely on available data that has already been made publicly available, e.g. the UCI machine learning repository which contains over 600 data sets.

1.3. How much data storage will your project require in total?

- 0 – 10 GB

2. What metadata and documentation will accompany the data?

2.1 Indicate what documentation will accompany the data.

No data will be generated, hence no documentation is required.

2.2 Indicate which metadata will be provided to help others identify and discover the data.

No data will be generated, hence no metadata will be provided.

3. How will data and metadata be stored and backed up during the research?

3.1 Describe where the data and metadata will be stored and backed up during the project.

- Institution networked research storage

During the course of the research project, all data will be stored on local servers (Project Drive) maintained and automatically backed up by TU Delft ICT. Data can be recovered with the help of TU Delft ICT services in the event of an incident.
All code will be maintained in a dedicated GitLab version control system provided by TU Delft, which similarly to Project Drive, is backed up and maintained by TU Delft.

3.2 How will data security and protection of sensitive data be taken care of during the research?

- Not applicable (no sensitive data)

No sensitive data will be used.

4. How will you handle issues regarding the processing of personal information and intellectual property rights and ownership?

4.1 Will you process and/or store personal data during your project?

If yes, how will compliance with legislation and (institutional) regulation on personal data be ensured?

- No

4.2 How will ownership of the data and intellectual property rights to the data be managed?

All developed code will be publicly released following NWO's policies. During the active phase of research, the lead applicant from TU Delft will oversee the access rights to developed code, as well as any requests for access from external parties. They will be released publicly no later than at the time of publication of corresponding research papers.

5. How and when will data be shared and preserved for the long term?

5.1 How will data be selected for long-term preservation?

- Other (please specify)

No data will be generated/uploaded.

5.2 Are there any (legal, IP, privacy related, security related) reasons to restrict access to the data once made publicly available, to limit which data will be made publicly available, or to not make part of the data publicly available?

If yes, please explain.

- No

5.3 What data will be made available for re-use?

- Other (please specify)

No data will be generated/uploaded.

5.4 When will the data be available for re-use, and for how long will the data be available?

- Data available as soon as article is published

No data will be generated/uploaded.

5.5 In which repository will the data be archived and made available for re-use, and under which license?

No data will be generated/uploaded.

5.6 Describe your strategy for publishing the analysis software that will be generated in this project.

The developed software and codes presented in academic papers will be shared on GitHub and those GitHub repositories will be published via 4TU.ResearchData. This way, they will be publicly available to anyone for re-use under an open licence. They will be also assigned a Digital Object Identifier (DOI), to make them citable and persistently available.

6. Data management costs

6.1 What resources (for example financial and time) will be dedicated to data management and ensuring that data will be FAIR (Findable, Accessible, Interoperable, Re-usable)?

4TU.ResearchData is able to archive 1TB of data per researcher per year free of charge for all TU Delft researchers. We do not expect to exceed this and therefore there are no additional costs of long term preservation.