
Group-invariant neural tensor train networks


A Data Management Plan created using DMPonline.be

Creator: Nick Vannieuwenhoven  <https://orcid.org/0000-0001-5692-4163>

Affiliation: KU Leuven (KUL)

Funder: Fonds voor Wetenschappelijk Onderzoek - Research Foundation Flanders (FWO)

Template: FWO DMP (Flemish Standard DMP)

Project Administrator: Estelle Massart, Nick Vannieuwenhoven  <https://orcid.org/0000-0001-5692-4163>

Grant number / URL: G011624N

ID: 207115

Start date: 01-01-2024

End date: 31-12-2027

Project abstract:

Recent years witnessed growing awareness of societal challenges posed by unbridled application of machine learning models, mainly neural networks. Two specific and prominent challenges are their rising primary energy consumption for training increasingly large networks and their ability to extract and perpetuate societal biases against specific demographics from training data. While the literature has developed basic techniques for mitigating each one these challenges separately, an integrated technology that efficiently addresses both challenges simultaneously is lacking.

This project will introduce a new neural network architecture, including advanced training algorithms, that aims to mitigate aforementioned challenges simultaneously. The first challenge will be overcome by appealing to data-sparse (memory-efficient) tensor decompositions applied in an innovative way so that both the linear maps applied by the neural network and the signals propagated through it stay in a data-sparse format. The second challenge will be mitigated by fusing above tensor decomposition technology with the ability to stay invariant under desired (group) transformations of the inputs, while further improving the data sparseness of its linear maps.

Last modified: 08-05-2024

Group-invariant neural tensor train networks

DPIA

DPIA

Have you performed a DPIA for the personal data processing activities for this project?

- Not applicable

Group-invariant neural tensor train networks

GDPR

GDPR

Have you registered personal data processing activities for this project?

- Not applicable

Group-invariant neural tensor train networks

Application DMP

Questionnaire

Describe the datatypes (surveys, sequences, manuscripts, objects ...) the research will collect and/or generate and /or (re)use. (use up to 700 characters)

Question not answered.

Specify in which way the following provisions are in place in order to preserve the data during and at least 5 years after the end of the research? Motivate your answer. (use up to 700 characters)

Question not answered.

What's the reason why you wish to deviate from the principle of preservation of data and of the minimum preservation term of 5 years? (max. 700 characters)

Question not answered.

Are there issues concerning research data indicated in the ethics questionnaire of this application form? Which specific security measures do those data require? (use up to 700 characters)

Question not answered.

Which other issues related to the data management are relevant to mention? (use up to 700 characters)

Question not answered.

Group-invariant neural tensor train networks

FWO DMP (Flemish Standard DMP)

1. Research Data Summary

List and describe all datasets or research materials that you plan to generate/collect or reuse during your research project. For each dataset or data type (observational, experimental etc.), provide a short name & description (sufficient for yourself to know what data it is about), indicate whether the data are newly generated/collected or reused, digital or physical, also indicate the type of the data (the kind of content), its technical format (file extension), and an estimate of the upper limit of the volume of the data.

				Only for digital data	Only for digital data	Only for digital data	Only for physical data
Dataset Name	Description	New or reused	Digital or Physical	Digital Data Type	Digital Data format	Digital data volume (MB/GB/TB)	Physical volume
		<i>Please choose from the following options:</i> <ul style="list-style-type: none"> • Generate new data • Reuse existing data 	<i>Please choose from the following options:</i> <ul style="list-style-type: none"> • Digital • Physical 	<i>Please choose from the following options:</i> <ul style="list-style-type: none"> • Observational • Experimental • Compiled/aggregated data • Simulation data • Software • Other • NA 	<i>Please choose from the following options:</i> <ul style="list-style-type: none"> • .por, .xml, .tab, .csv, .pdf, .txt, .rtf, .dwg, .gml, ... • NA 	<i>Please choose from the following options:</i> <ul style="list-style-type: none"> • <100MB • <1GB • <100GB • <1TB • <5TB • <10TB • <50TB • >50TB • NA 	
Research software code	Program codes of research software, including plotting for scientific articles	Generate new data	Digital	Software	Python (.py), Julia (.jl), Octave (.m), Gnuplot (.plot)	<100MB	
Simulation results	Results of numerical simulations of solving partial differential equations	Generate new data	Digital	Simulation data	.hdf5	<100GB	
PyTorch torchvision Datasets	Standard computer vision datasets available within PyTorch torchvision. A compilation of datasets.	Reuse existing data	Digital	Compiled/aggregated data	Various image files (.jpg, .png, ...), ad-hoc binary files, etc	< 1TB (portions will be used, not all)	

If you reuse existing data, please specify the source, preferably by using a persistent identifier (e.g. DOI, Handle, URL etc.) per dataset or data type:

PyTorch torchvision datasets: <https://pytorch.org/vision/stable/datasets.html>

Are there any ethical issues concerning the creation and/or use of the data (e.g. experiments on humans or animals, dual use)? Describe these issues in the comment section. Please refer to specific datasets or data types when appropriate.

- No

Will you process personal data? If so, briefly describe the kind of personal data you will use in the comment section. Please refer to specific datasets or data types when appropriate.

- No

Does your work have potential for commercial valorization (e.g. tech transfer, for example spin-offs, commercial exploitation, ...)? If so, please comment per dataset or data type where appropriate.

- Yes

The Research software codes may provide the basis of a commercial implementation of group-invariant neural tensor train networks. They are intended for fundamental research, however, with no intent for usability, appropriateness, etc within a non-academic context.

The other dataset we generate has no potential for commercialization.

Do existing 3rd party agreements restrict exploitation or dissemination of the data you (re)use (e.g. Material/Data transfer agreements/ research collaboration agreements)? If so, please explain in the comment section to what data they relate and what restrictions are in place.

- No

Are there any other legal issues, such as intellectual property rights and ownership, to be managed related to the data you (re)use? If so, please explain in the comment section to what data they relate and which restrictions will be asserted.

- Yes

The Research software codes may be generated as part of a collaboration between KU Leuven and UCLouvain researchers. For these codes, we intend to use a GPL-class software license.

The simulation results will be made available under a CC-BY license.

2. Documentation and Metadata

Clearly describe what approach will be followed to capture the accompanying information necessary to keep data understandable and usable, for yourself and others, now and in the future (e.g., in terms of documentation levels and types required, procedures used, Electronic Lab Notebooks, README.txt files, Codebook.tsv etc. where this information is recorded).

The Research software code will be documented following generally accepted documentation practices. Proofs of mathematical correctness and description will be explained in scientific reports.

The simulation results dataset will be accompanied by the research software codes that can be used to produce them. This gives an exact, formal, reproducible description of the data. The HDF5 files containing the results will also include additional metadata that describes the naming conventions of the data items and which parameters were used to generate the data.

Will a metadata standard be used to make it easier to find and reuse the data? If so, please specify (where appropriate per dataset or data type)

which metadata standard will be used. If not, please specify (where appropriate per dataset or data type) which metadata will be created to make the data easier to find and reuse.

- No

The structure of the generated data will not be so complex as to require extensive and structured metadata to describe it.

3. Data storage & back-up during the research project

Where will the data be stored?

Data will be stored on the researcher's computers.

All program code (including those that generate the data sets) will be developed in a KU Leuven GitLab repository. This provides version control.

How will the data be backed up?

The home directories are automatically backed up, in accordance with our research unit's data management plan. Regular snapshots are provided.

We rely primarily on KU Leuven GitLab's mechanisms for backing up data. In addition, to this, when a paper is accepted for publication, all research items, excluding large data sets that can be regenerated from codes, will be archived in the PI's KU Leuven OneDrive. This system has its own backup mechanisms.

Is there currently sufficient storage & backup capacity during the project? If yes, specify concisely.

If no or insufficient storage or backup capacities are available, then explain how this will be taken care of.

- Yes

The program codes will require less than a few megabytes.

The simulation results we will generate, are not expected to exceed a few gigabytes, well within the capabilities of the standard computers we will use.

How will you ensure that the data are securely stored and not accessed or modified by unauthorized persons?

The research software code will be developed through GitLab, which requires authentication. Only the project's participants will be able to modify these items. Who modified what and how is precisely tracked by this version management system.

What are the expected costs for data storage and backup during the research project? How will these costs be covered?

0 EUR.

4. Data preservation after the end of the research project

Which data will be retained for at least five years (or longer, in agreement with other retention policies that are applicable) after the end of the project? In case some data cannot be preserved, clearly state the reasons for this (e.g. legal or contractual restrictions, storage/budget issues, institutional policies...).

All aforementioned datasets will be retained indefinitely.

Where will these data be archived (stored and curated for the long-term)?

The research software code will be stored in KU Leuven GitLab. In addition, to this, when a paper is accepted for publication, all research items, excluding large data sets that can be regenerated from codes, will be archived in the PI's KU Leuven OneDrive.

The simulation results may be deposited in KU Leuven's RDR, if sufficiently interesting. Otherwise only the code that generated the results is stored as part of the research software codes.

What are the expected costs for data preservation during the expected retention period? How will these costs be covered?

0 EUR.

5. Data sharing and reuse

Will the data (or part of the data) be made available for reuse after/during the project? In the comment section please explain per dataset or data type which data will be made available.

- Yes, in an Open Access repository

If the simulation data is sufficiently interesting, it will be deposited to an open access repository such as KU Leuven RDR.

If access is restricted, please specify who will be able to access the data and under what conditions.

NA

Are there any factors that restrict or prevent the sharing of (some of) the data (e.g. as defined in an agreement with a 3rd party, legal restrictions)? Please explain in the comment section per dataset or data type where appropriate.

- No

Where will the data be made available? If already known, please provide a repository per dataset or data type.

Research software code will be made available through KU Leuven GitLab.

Simulation data may be made available through KU Leuven RDR.

When will the data be made available?

The datasets will be released only upon completion of the scientific reports that use the datasets. The scientific reports will be posted to the arXiv.org repository as soon as possible.

Which data usage licenses are you going to provide? If none, please explain why.

The simulation results will be released with CC-BY.

The research software code will be released under a GNU Public License.

Do you intend to add a PID/DOI/accession number to your dataset(s)? If already available, you have the option to provide it in the comment section.

- No

What are the expected costs for data sharing? How will these costs be covered?

0 EUR.

6. Responsibilities

Who will manage data documentation and metadata during the research project?

David Thorsteinsson

Who will manage data storage and backup during the research project?

David Thorsteinsson

Who will manage data preservation and sharing?

David Thorsteinsson

Who will update and implement this DMP?

Nick Vannieuwenhoven