Plan Overview

A Data Management Plan created using DMPonline.be

Title: Rethinking Transformers Through Duality Principles

Creator: Aldona Niemiro-Sznajder

Principal Investigator: n.n., Johan Suykens, n.n.

Data Manager: Aldona Niemiro-Sznajder

Affiliation: KU Leuven (KUL)

Funder: KU Leuven (KUL)

Template: KU Leuven BOF-IOF

Principal Investigator: n.n. n.n., Johan Suykens, n.n. n.n.

Data Manager: Aldona Niemiro-Sznajder

Project abstract:

The AI renaissance has been propelled by groundbreaking deep learning models, with the transformer architecture and its attention mechanism at the forefront. Notably, transformers are the foundation for famous models like ChatGPT and BERT and have witnessed lesser-known successes in computer vision, robotics, etc. Our goal is to develop a framework using generalized duality so as to understand and improve performance and efficiency of transformer architectures and the associated training procedures. This framework will improve current transformer architectures, apply to other deep learning models and even motivate entirely new architectures. We aim to unlock the full potential of transformers, making them more accessible to a wider range of users and applications, thereby democratizing the benefits of this technology, reducing the dependency on immense resources and data and fostering broader innovation and application across various domains of AI research.

ID: 210205

Start date: 01-10-2024

End date: 30-09-2028

Last modified: 10-10-2024

Rethinking Transformers Through Duality Principles

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.

Dataset name / ID	Description	New or reuse	Digital or Physical data	Data Type	File format	Data volume	Physical volume
		Indicate: ***N****(ew data) or* *E(xisting data)*	Indicate: **D**(igital) or **P**(hysical)	Indicate: **A**udiovisual **I**mages **S**ound **N**umerical **T**extual **M**odel **SO**ftware Other (specify)		Indicate: <1GB <100GB <1TB <5TB >5TB NA	
ImageNet 100/1k	Images with 100 or 1000 classes	E	D	I	RGB	<100GB	
UEA Time Series	Time series data for classification	E	D	N,T	time series data by different measurements	<1GB	
Long-range Arena Benchmark	Different datasets of very long sequence for classification	E	D	N,T	Data with long sequence by different measurements	<100GB	
WikiText-103	Long-sequence text data for natural language processing	E	D	T	Text data with long sequence	<1GB	
CIFAR 10/100	Images with 10 or 100 classes	E	D	I	RGB	<100G	

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:

ImageNet: https://www.image-net.org/
UEA Time Series: https://gist.github.com/oguiza/26020067f499d48dc52e5bcb8f5f1c57
Long-range Arena: https://github.com/google-research/long-range-arena
WikiText-103: https://developer.ibm.com/exchanges/data/all/wikitext-103/
CIFAR: https://www.cs.toronto.edu/~kriz/cifar.html

Are there any ethical issues concerning the creation and/or use of the data (e.g. experiments on humans or animals, dual use)? If so, refer to specific datasets or data types when appropriate and provide the relevant ethical approval number.

- No

Will you process personal data? If so, please refer to specific datasets or data types when appropriate and provide the KU Leuven or UZ Leuven privacy register number (G or S number).

- 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.

- No

Do existing 3rd party agreements restrict exploitation or dissemination of the data you (re)use (e.g. Material or 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.

- No

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 datasets to be used will be downloaded from their official websites, where exact procedures of usage and detailed information of the data are provided.
ImageNet: https://www.image-net.org/
UEA Time Series: https://gist.github.com/oguiza/26020067f499d48dc52e5bcb8f5f1c57
Long-range Arena: https://github.com/google-research/long-range-arena
WikiText-103: https://developer.ibm.com/exchanges/data/all/wikitext-103/
CIFAR: https://www.cs.toronto.edu/~kriz/cifar.html

Will a metadata standard be used to make it easier to find and reuse the data?
If so, please specify which metadata standard will be used.

If not, please specify which metadata will be created to make the data easier to find and reuse.

- No

Metadata with specific information on experiments carried out with the developed algorithms and software will be stored in associated text files. Depending on the source of the benchmark data, the appropriate metadata structure will be applied.

Data Storage & Back-up during the Research Project

Where will the data be stored?

- Other (specify below)

The data generated in the course of the project will be stored on storage facilities of the research unit.

How will the data be backed up?

- Standard back-up provided by KU Leuven ICTS for my storage solution

Is there currently sufficient storage & backup capacity during the project?

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

- Yes

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

For all data stored in the department's servers, access is regulated by an access control list (ACL) that grants read-write access to the project owner and read-only access to specific users. The ACL is managed by the project owner. Client computers can access the data using: SMB2 (or higher) from specific IP ranges NFSv4 from specific (IT managed) systems.

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

The storage facilities of the research unit are currently available at no cost to the project.

Data Preservation after the end of the Research Project

Which data will be retained for 10 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 data will be preserved for 10 years according to KU Leuven RDM policy

Any relevant data produced in the course of the project, will be kept for a period of 10 years.

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

- Other (specify below)

After the period of 10 years, any data generated in the course of the project that is not longer in use will be removed from the department's servers.

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

The storage facilities of the research unit are available for the researchers for free.

Data Sharing and Reuse

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

- Yes, as open data
- Yes, as embargoed data (temporary restriction)

The deliverables of the project include algorithms and software. Algorithms will be described/documented in internal reports (publicly available in the research group's publication repository) and/or published in conference proceedings and scholarly journals. Whenever relevant, software code will be made available together with the reports/publications, e.g. in venues provided by the publisher.

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 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.

- Other data repository (specify below)
- Other (specify below)

ImageNet: https://www.image-net.org/
UEA Time Series: https://gist.github.com/oguiza/26020067f499d48dc52e5bcb8f5f1c57
Long-range Arena: https://github.com/google-research/long-range-arena
WikiText-103: https://developer.ibm.com/exchanges/data/all/wikitext-103/
CIFAR: https://www.cs.toronto.edu/~kriz/cifar.html

When will the data be made available?

- Upon publication of research results

Which data usage licenses are you going to provide?

If none, please explain why.

- CC-BY 4.0 (data)
- GNU GPL-3.0 (code)

Do you intend to add a persistent identifier (PID) to your dataset(s), e.g. a DOI or accession number? If already available, please provide it here.

- No

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

The costs for data sharing can come from the storage requirements on the cloud drive or VSC/HPC server, and they can be covered by sufficient budgets on the chosen cloud drive or servers.

Responsibilities

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

The PIs (Johan Suykens and Panagiotis Patrinos) will be responsible for data documentation and metadata.

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

The IT division of the department (ESAT) will be responsible for the data storage and back-up during the project.

Who will manage data preservation and sharing?

The PIs (Johan Suykens and Panagiotis Patrinos) will be responsible for the data preservation and reuse.

Who will update and implement this DMP?

The PIs (Johan Suykens and Panagiotis Patrinos) bear the end responsibility of updating and implementing this DMP.