LIFELONG-LEARNING ELECTRONIC SYSTEMS: EDGE-DEVICES THAT GET SMARTER DAY-BY-DAY [LIFELINES]

A Data Management Plan created using DMPonline.be

Creators: Marian Verhelst, n.n. n.n., n.n. n.n., David Maes, Tinne Tuytelaars

Affiliation: KU Leuven (KUL)

Funder: KU Leuven (KUL)

Template: KU Leuven BOF-IOF

Principal Investigator: Tinne Tuytelaars, n.n. n.n., n.n. n.n., Marian Verhelst

Data Manager: David Maes

Grant number / URL: https://research.kuleuven.be/portal/en/project/3E231213

ID: 204105

Start date: 01-01-2024

End date: 31-12-2030

Project abstract:

Electronic devices have become an integral part of our daily life. Yet, current devices are still largely inefficient and ineffective, as they do not adapt to their user or changing operating context. Inspired by biological systems, LifeLinES instead aims to develop lifelong-learning, self-managing devices, that will continuously learn and improve their performance based on inputs from their environment and users. Unlike current systems that rely on generic models pre-trained offline with large datasets, the proposed research focuses on in-the-edge solutions, where devices can learn independently without privacy-sensitive data leaving the device. This is realized by innovating along 3 pillars: (1) OBSERVE: developing energy-efficient and dynamically adaptable sensing and edge processing architectures; (2) LEARN: researching novel machine learning methods that fit within the computational energy and memory budgets of embedded devices; and (3) ADAPT: enabling hardware and system software to be fully flexible and capable of adjusting operating parameters on the fly. Successful implementation of the LifeLinES' vision will result in personalized, adaptable electronic devices that can grow along with their users and environments, with ever-improving performance and user satisfaction, and with low energy consumption and low ecological footprint.

Last modified: 30-01-2024

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	open source image database for classification tasks	E	D	I	jpeg, xml, txt	167.62 GB	/
MNIST	Popular image dataset for bench-marking machine learning models	E	D	I	.gz file	<1GB	/
CIFAR-10	Popular image dataset for bench-marking machine learning models	E	D	I	.gz file	<1GB	/
CIFAR-100	Popular image dataset for bench-marking machine learning models	E	D	I	.gz file	<1GB	/
Standard cell characterization data	Vendor-specific data of the performance of standard cells in different silicon technologies	E	D	N	lib	120GB	/
Developed algorithms	ML algorithms developed in the project	N	D	SO	.py	1GB	/
Developed digital IP blocks	RTL implementations of digital hardware developed in the project	N	D	T	.v	1GB	/
Developed electronic circuit blocks	Circuit netlists	N	D	T	.spi	1GB	/

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: Online: ImageNet Object Localization Challenge | Kaggle
- MNIST: http://yann.lecun.com/exdb/mnist/
- CIFAR-10: https://www.cs.toronto.edu/~kriz/cifar.html
- CIFAR-100: https://www.cs.toronto.edu/~kriz/cifar.html
- Standard cell characterization data: TSMC, GlobalFoundries and Intel, via direct download from the vendors (under their NDA conditions).

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.

- Yes

- Algorithms: They can later be commercialized for specific applications, either by tech transfer to an existing company or potentially leading to a spin-off company.
- Digital IP: They can be reused on commercial tape-outs, either by a spin-off or an external company to which tech transfer is done.

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.

- Yes

- Standard cell characterization data: Under the NDA conditions of the semiconductor foundry providing the standard cells, it is not allowed to share the bare data with third parties. Yet, the resulting chip design and the obtained chip performances can be shared. Future collaboration partners can themselves also apply to the semiconductor foundry for access to these technology characterizations.

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).

1) Algorithms:
The source code will be collected per algorithm and per version of the algorithm, including a text file with a clear description of the algorithm. This is automated using an internal GITHUB, and registered on Zenodo. The GIThub folders storing the open-source algorithms and tools will have a docker container to make them easy to install/use.
Moreover, we will add extensive readme files on the required tool versions. Finally, the code is planned to be written in Python, as typically used in this field.

2) Designs: The RTL description (Verilog code) of some parts of the designs will be shared with the open source community. This will be automated using an internal GITHUB, and registered on Zenodo. Other parts of the designs might require more careful IPR management, preventing the data to be shared without a prior agreement on confidentiality and IPR. We will update the DMP in case this is decided for

specific blocks, while we keep open sourcing of the design blocks as the default. The RTL description will be written and shared in the Verilog language, which is the most popular hardware description language used in practice today.

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.

- Yes

1) Algorithms:

The source code will be collected per algorithm and per version of the algorithm, including a text file with a clear description of the algorithm. This is automated using an internal GITHUB, and registered on Zenodo.

The GIThub folders can be used freely by other researchers under the BSD-3-Clause license. No embargo is foreseen. Through GIT, an active issue management system will be foreseen, such that users can flag problems with the tools, and the developers can react openly to how they are handled.

2) Electronic designs:

a) Simulations: Raw simulation data will be collected per simulation test, including a text file with a clear description of what the data represent and how they were generated. The input files used for the simulation will be kept inside the same folder. The name of the folder will contain the simulation conditions. A text file explaining the naming will be maintained. At the moment, there are no plans to share these simulation files, as they are not very informative to the community.

b) Designs: Details on the conceptual, architectural and topological design of the circuits will be documented in word files. Links to the folders in which the design data are stored will be included, as well as all the necessary metadata to be able to extract and reuse the design data: technology node and flavor, etc. Moreover, the RTL description (Verilog code) or the netlist of designs will be shared with the open source community. This will be automated using an internal GITHUB, and registered on Zenodo. The VERILOG code in the GIThub folders can be used freely by other researchers under the BSD-3-Clause license; no embargo is foreseen.

c) Experiments/measurements:

Raw measurement data will be collected per measurement test, including a text file with a clear description of what the data represent and how they were generated. The input files used for the measurements will be kept inside the same folder. The name of the folder will contain the measurement conditions. A text file explaining the naming will be maintained. At the moment, there are no plans to publicly share these measurement files, as they are not very informative to the community.

Data Storage & Back-up during the Research Project

Where will the data be stored?

- Shared network drive (J-drive)
- Personal network drive (I-drive)
- OneDrive (KU Leuven)

How will the data be backed up?

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

The resources and responsibilities for long-term preservation are clear. The data storage resources are available within the research groups. At KU Leuven, the costs for data storage are internally accounted for at departmental level. Currently, there is more than sufficient disk space at the involved research groups to cover the data storage requirements of this project. The data will be stored on our servers with automatic daily back-up procedures for at least 10 years, conform the KU Leuven RDM policy. Ben Geeraerts (KU Leuven, ESAT-MICAS) will be responsible for data storage and back-up and for data preservation.

Moreover, the Department of Electrical Engineering (ESAT) of KU Leuven hosts an own gitlab server, which allows to share large projects with the open-source community for free.

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

The data storage resources are available within the ESAT-MICAS and ESAT-PSI research groups.

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

No sensitive personal data will be used in the project. Confidential data of the silicon technologies used are stored on file servers that are only accessible by authorized people with specific account settings. The servers are located in a secured room with access restrictred to system administrators. For data related to specific, very advanced and exclusive technologies we even have physically separate file servers.

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

The data storage resources are available within the ESAT-MICAS and ESAT-PSI research groups.
The storage capabilities are upgraded annually, and the cost is distributed across all projects that make use of them. The Methusalem project will hence pay a small fraction of these storage expenses.

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

The data will be stored on our servers with automatic daily back-up procedures for at least 10 years, conform the KU Leuven RDM policy.

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

- Large Volume Storage (longterm for large volumes)
- Other (specify below)

The data will be stored on the ESAT storage system for at least 10 years, conform the KU Leuven RDM policy.

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

The expenses for the ESAT-organized backup and 10-year storage system are distributed among all projects that make use of this service. The LifeLinES Methusalem project will hence pay its small fraction of these expenses.

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

Developed algorithms will be made available as open source to the community through GIT, and registered on Zenodo.
Part of the developed hardware designs will be made available to the community in open source through GIT, by sharing their RTL files.
The technology-specific characterization data for these designs will not be shared due to the NDA conditions from the semiconductor foundries.

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

/

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.

- Yes, intellectual property rights

Any technology-related information or files coming from the silicon vendors can not be shared in open source, due to IP regulations. Yet, this does not prevent us from sharing the VERILOG RTL code of the designs made in the LifeLinES project.

Where will the data be made available?

If already known, please provide a repository per dataset or data type.

- Other data repository (specify below)

GITHUB, registered on Zenodo.

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.

- Other (specify below)
- MIT licence (code)

BSD-3-Clause license

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?

Main costs are the personpower to document the data and to answer to inquiries posted online, supporting future users. This will be done by personnel (PhDs and postdoc) paid on the LifeLinES Methusalem project.

Responsibilities

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

The PhD researchers on the LifeLinES project, supervised by the postdoc and the PIs.

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

Ben Geeraerts (KU Leuven, ESAT-MICAS) will be responsible for data storage and back-up and for data preservation.

Who will manage data preservation and sharing?

The PhD researchers on the LifeLinES project, supervised by the postdoc and PIs, will be responsible for data sharing.
Ben Geeraerts (KU Leuven, ESAT-MICAS) will be responsible for data preservation.

Who will update and implement this DMP?

The PIs: Marian Verhelst, Georges Gielen and Tinne Tuytelaers