Rethinking Replay: understanding and avoiding the stability gap in continual learning

A Data Management Plan created using DMPonline.be

Creators: Gido van de Ven https://orcid.org/0000-0002-5239-5660, Tinne Tuytelaars

Affiliation: KU Leuven (KUL)

Template: KU Leuven BOF-IOF

Principal Investigator: Tinne Tuytelaars

Grant number / URL: C14/23/100

ID: 203266

Start date: 01-10-2023

End date: 30-09-2027

Project abstract:

Continually learning a sequence of tasks is challenging for deep neural networks, mainly because they tend to 'catastrophically forget' past tasks when learning a new one. The main approach for alleviating catastrophic forgetting is replay, which involves storing examples of past tasks in a memory buffer and revisiting them when training on later tasks. However, recently we discovered that replay still suffers from substantial forgetting when learning a new task, but that this forgetting is temporary and followed by a phase of performance recovery. We called this surprising but consistently observable phenomenon the 'stability gap'. In this project, we propose (1) to study the root causes of the stability gap with a set of ablation experiments and through visualizing the optimization trajectory, (2) to explore connections between the stability gap and a comparable phenomenon of temporary forgetting in humans described in the cognitive science literature, and (3) to develop new continual learning methods that are able to overcome the stability gap.

Last modified: 30-01-2024

Rethinking Replay: understanding and avoiding the stability gap in continual learning

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
MNIST	Popular image dataset for bench-marking machine learning models	*Existing data*	Digital	Images	.gz file	<1GB	
CIFAR-10	Popular image dataset for bench-marking machine learning models	*Existing data*	Digital	Images	.gz file	<1GB	
CIFAR-100	Popular image dataset for bench-marking machine learning models	*Existing data*	Digital	Images	.gz file	<1GB	
ImageNet	Popular image dataset for bench-marking machine learning models	*Existing data*	Digital	Images	jpeg, xml, txt	<500GB	
continual-learning	Software (https://github.com/GMvandeVen/continual-learning)	*Existing data*	Digital	Software	Python code	<1GB	
New code library	New code that will be developed during the project	New *data*	Digital	Software	Python code	<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:

The project will re-use the existing publicly available data sets listed below, which are all popular and widely used data sets for bench-marking machine learning algorithms.

- 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
- ImageNet: https://www.kaggle.com/c/imagenet-object-localization-challenge/overview

Besides the re-use of these publicly available data sets, in this project we will also re-use software developed in a previous project. This software is available on Github under an MIT licence (https://github.com/GMvandeVen/continual-learning) and has been assigned a DOI (https://zenodo.org/record/7189378).

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 existing datasets and software that we will re-use are already accompanied by clear documentation.
For the software that this project is anticipated to generate, we intend to include documentation in the same way as we did for software resulting from a previous project: https://github.com/GMvandeVen/continual-learning
In particular, this means that we intend to include clear descriptions, a README, demos and documentation within the code files.

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

For the software produced by this project, we intend to include metadata in the same way as we did for software resulting from a previous project: https://github.com/GMvandeVen/continual-learning
In particular, this includes making the software publicly available on Github including clear descriptions, topics and a README.

Data Storage & Back-up during the Research Project

Where will the data be stored?

- Shared network drive (J-drive)

How will the data be backed up?

- Standard back-up provided by KU Leuven ICTS for my storage solution
- Other (specify below)

Software will also be backed up online in Github repositories.

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?

The desktop computers and Github accounts that we intend to use are protected by password and/or 2FAC.

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

These costs are minor and can be covered by project budget.

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

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

- Other (specify below)

The software produced by this project is intended to be made publicly available on Github under an MIT licence and, if relevant, to be assigned a persistent identifier (DOI) through Zenodo, as we have done for code resulting from a previous project (Github: https://github.com/GMvandeVen/continual-learning; DOI: https://zenodo.org/record/7189378).

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

There are no additional expected costs for this.

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

The software produced by this project is intended to be made publicly available on Github under an MIT licence and assigned a persistent identifier (DOI) through Zenodo, as we have done for code
resulting from a previous project (Github: https://github.com/GMvandeVen/continual-learning; DOI: https://zenodo.org/record/7189378).

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

Not applicable.

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 (specify below)

The software produced by this project is intended to be made publicly available on Github under an MIT licence and assigned a persistent identifier (DOI) through Zenodo, as we have done for code
resulting from a previous project (Github: https://github.com/GMvandeVen/continual-learning; DOI: https://zenodo.org/record/7189378).

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.

- MIT licence (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.

- Yes, a PID will be added upon deposit in a data repository

If relevant, the software produced by this project is intended to be assigned a persistent identifier (DOI) through Zenodo.

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

There are no additional expected costs for this.

Responsibilities

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

Th PI (Tinne Tuytelaars) and the Co-PI (Gido van de Ven) will do this.

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

Th PI (Tinne Tuytelaars) and the Co-PI (Gido van de Ven) will do this.

Who will manage data preservation and sharing?

Th PI (Tinne Tuytelaars) and the Co-PI (Gido van de Ven) will do this.

Who will update and implement this DMP?

Th PI (Tinne Tuytelaars) and the Co-PI (Gido van de Ven) will do this.