Generative Classification for Class-Incremental Learning

A Data Management Plan created using DMPonline.be

Creator: Gido van de Ven https://orcid.org/0000-0002-5239-5660

Affiliation: KU Leuven (KUL)

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

Template: FWO DMP (Flemish Standard DMP)

Grant number / **URL:** 1266823N

ID: 198126

Start date: 01-10-2022

End date: 30-09-2025

Project abstract:

Learning continually from non-stationary streams of data is a key feature of natural intelligence, but an unsolved problem in deep learning. Especially challenging for deep neural networks is “class-incremental learning”, where a network must learn to distinguish classes not observed together.

In deep learning, the default approach to classification is learning discriminative classifiers. This works great in the i.i.d. setting when all classes are observed together, but when new classes must be learned incrementally, training discriminative classifiers requires often problematic workarounds such as storing data or generative replay. Here, I propose to instead address class-incremental learning with generative classification.

As proof-of-concept, in preliminary work I showed that a naïve generative classifier, with a separate variational autoencoder per class and likelihood estimation through importance sampling, already performs very strongly. To improve the efficiency, scalability and performance of this generative classifier, I propose four further modifications: (1) move the generative modelling objective from the raw inputs to an intermediate network layer; (2) share the encoder network between classes; (3) use fewer importance samples for unlikely classes; and (4) make classification decisions hierarchical. This way I hope to develop generative classification into a practical, efficient and scalable state-of-the-art deep learning method for class-incremental learning.

Last modified: 28-03-2023

Generative Classification for Class-Incremental Learning
DPIA

DPIA

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

- Not applicable

Generative Classification for Class-Incremental Learning
GDPR

GDPR

Have you registered personal data processing activities for this project?

- Not applicable

Generative Classification for Class-Incremental Learning
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
MNIST	Popular image dataset for bench-marking machine learning models	Reuse existing data	Digital	Aggregated data	image data	<100MB	
CIFAR-10	Popular image dataset for bench-marking machine learning models	Reuse existing data	Digital	Aggregated data	image data	<1GB	
CIFAR-100	Popular image dataset for bench-marking machine learning models	Reuse existing data	Digital	Aggregated data	image data	<1GB	
continual-learning	Software (https://github.com/GMvandeVen/continual-learning)	Reuse existing data	Digital	Software	code	<100MB	
New code library	Software	Newly developed	Digital	Software	code	<100MB	

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

Besides the re-use of these publicly available data sets, in this project I will also re-use software that I 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)? 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.

- No

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.

- No

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 existing datasets and software that I will re-use are already accompanied by clear documentation.

For the software that this project is anticipated to generate, I intend to include documentation in the same way as I did for software resulting from a previous project: https://github.com/GMvandeVen/continual-learning
In particular, this means that I 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 (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.

- Yes

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

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

Where will the data be stored?

Data will be stored on my desktop computer.

How will the data be backed up?

Data will be backed up on separate hard drives as well as in Github repositories.

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

This project does not involve large amounts of data that require storage; existing storage and backup capacity should suffice.

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

Both my desktop computer and Github account 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 the bench fee.

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...).

The software produced by this project is intended to be retained indefinitely.

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

The software produced by this project is intended to be publicly available on Github under an MIT licence and assigned a persistent identifier (DOI) through Zenodo, as I 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.

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

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 I 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 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.

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 I 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?

Software produced by this project is intended to be made publicly available upon or before publication of the associated academic paper.

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

Software produced by this project is intended to be made publicly available under a MIT licence.

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.

- Yes

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.

6. Responsibilities

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

I will do this myself.

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

I will do this myself.

Who will manage data preservation and sharing?

I will do this myself.

Who will update and implement this DMP?

I will do this myself.