Plan Overview

A Data Management Plan created using DMPonline

Title: TIMING: Learning Time in Visual Recognition

Creator:Efstratios Gavves

Principal Investigator: Efstratios Gavves

Affiliation: University of Amsterdam (Universiteit van Amsterdam)

Funder: Netherlands Organisation for Scientific Research (NWO)

Template: Data Management Plan NWO (September 2020)

ORCID iD: 0000-0001-8947-1332

Project abstract:

From Facebook’s 3.5 billion live streams to the complex MRI sequences and satellite footage monitoring glaciers, video recognition becomes increasingly relevant. Ultimately it will enable to understand what is happening, where and when in videos by artificial intelligence. Encouraged by the breakthrough of deep representation learning in static image recognition, today's video recognition algorithms emphasize static representations. In effect, they are time invariant. Ignoring time like this suffices in simple short videos, but in tomorrow’s applications recognizing time is imperative: it determines whether a suspect draws something in or out of their pocket, where a tumour will move in the MRI or at which rate glaciers melt in satellite footage. For all these cases and more, video algorithms must be time equivariant, that is yield representations that change proportionally to the temporal change in the input. As we move to video understanding where temporality is critical, time equivariant algorithms are a must. I propose a 5-year research program that studies, develops and evaluates time equivariant video algorithms. To achieve this, we will approach video algorithms from two angles: time geometry, and time supervision. Geometry helps with accounting for innumerable patterns without blowing up the representation complexity. Time supervision helps with learning time equivariance, without relying on strong manual supervision. A temporal decathlon competition will be introduced to the community to evaluate, disseminate and utilize the temporal behaviour of video algorithms. The decathlon will serve as a proxy for designing better video algorithms more efficiently. It will also open up video algorithms to other disciplines, where researchers have videos and know their temporal properties but do not have a common reference point. All research will be published in the top relevant conferences and journals. The major innovation of the proposed research is understanding and exploiting time in video algorithms.

ID: 77835

Start date: 01-09-2021

End date: 31-08-2025

Last modified: 19-05-2021

in their own plans, and customise it as necessary. You do not need to credit the creator(s) as the source of the language used, but using any of the plan's text does not imply that the creator(s) endorse, or have any relationship to, your project or proposal

TIMING: Learning Time in Visual Recognition

General Information

Name applicant and project number

Efstratios Gavves
VI.Vidi.193.129

Name of data management support staff consulted during the preparation of this plan and date of consultation.

Boy Menist
May 19, 2021

1. What data will be collected or produced, and what existing data will be re-used?

1.1 Will you re-use existing data for this research?

If yes: explain which existing data you will re-use and under which terms of use.

- Yes

The research will use existing data from already public datasets including several datasets from the action dataset GitHub repository (https://github.com/xiaobai1217/Awesome-Video-Datasets). Examples: Moments in Time dataset, DAVIS, A2D. If new datasets will be used, an updated data management plan will be submitted.

1.2 If new data will be produced: describe the data you expect your research will generate and the format and volumes to be collected or produced.

The research will generate data in the form of machine learning models and deep networks. The file format for these machine learning models follow the convention of the respective Deep Learning frameworks (Tensorflow, PyTorch -> *.protobuf). When the open format ONNX is mature enough, models will be produced and stored using ONNX.
The research will also generate numeric data in the form of predicting future trajectories of moving objects in videos. These are for specific datasets and they can be reused for comparisons with other competing models.

1.3. How much data storage will your project require in total?

- 10 – 100 GB

2. What metadata and documentation will accompany the data?

2.1 Indicate what documentation will accompany the data.

The data generation will be accompanied by a detailed explanation of how to use the data. This will be done via the GitHub open-sourcing platform. The GitHub page will further contain documentation on the methodology as well as how to reproduce the data if needed. This information is written using the markdown format in the GitHub readme.md files. A specific versioned extract of the GitHub source tree will be added to published datasets (source code will become part of the dataset). Git-branches will be used to support publications and published data.

2.2 Indicate which metadata will be provided to help others identify and discover the data.

Author information, links to datasets, link to the publication, software on how to use the data, metadata for the machine learning models, readme files on how to use the data and the machine learning models. Also, as part of the project, persistent identifiers (PIDs) will be assigned to the data to makes the data traceable (even when the move) and thus helps in making your data FAIR. When publishing int the 4TU.Datacentre, this is a requirement.

3. How will data and metadata be stored and backed up during the research?

3.1 Describe where the data and metadata will be stored and backed up during the project.

- Institution networked research storage

The data will be stored using UvA Figshare as a storage location for publication data. This should make it possible to link todata from/accross Pure (and thus links your publication to the data, also using PIDs). This will make the data FAIR.
Regarding backups, using SURFsara's data archive will fill this gap in the short term. In mid-term at the University of Amsterdam the plan is to "offer" the data archive as a faculty service (supporting workflows from faculty storage towards data archive and shielding you from the financial construct that is used by SURFsara). Once that is possible, it will be done for this project as well.

3.2 How will data security and protection of sensitive data be taken care of during the research?

- Not applicable (no sensitive data)

4. How will you handle issues regarding the processing of personal information and intellectual property rights and ownership?

4.1 Will you process and/or store personal data during your project?

If yes, how will compliance with legislation and (institutional) regulation on personal data be ensured?

- No

4.2 How will ownership of the data and intellectual property rights to the data be managed?

The datasets used in this research are published under specific licenses. Most of the times the datasets I expect to use public data, which are free for access and processing. If creating our own datasets, I will make sure they have the right license for open sharing. All research will respect the IP rights as per the license requirements. There will be no IP rights affected.

5. How and when will data be shared and preserved for the long term?

5.1 How will data be selected for long-term preservation?

- Other (please specify)

Only the final models will be kept and shared, as these models are the ones that have the best performance. These data will be stored in the University long-term facilities for at least 10 years, to overlap with the research cycles also in the automotive industry. Data for training will also be kept in storage in the same way.

5.2 Are there any (legal, IP, privacy related, security related) reasons to restrict access to the data once made publicly available, to limit which data will be made publicly available, or to not make part of the data publicly available?

If yes, please explain.

- No

5.3 What data will be made available for re-use?

- Other (please specify)

Only the final models will be kept and shared, as these models are the ones that have the best performance. These data will be stored in the University long-term facilities for at least 10 years.

5.4 When will the data be available for re-use, and for how long will the data be available?

- Data available as soon as article is published

Following the tradition in the field, all models and data will be shared when the respective papers are published.

5.5 In which repository will the data be archived and made available for re-use, and under which license?

During the course of the project, the research will make data underpinning publications available through UvA Figshare using its public sharing facilities. Upon completion of the project the data will be stored within a national data archive repository. More specifically, after contacting 4TU.Datacentre their certified facilities will be used for long term storage and management of the data and code generated during the proposal.
The ICT facilities required for the project are already provided creating synergy with the DAS-initiative through demonstrating DAS-concepts whilst also furthering research. ICT facilities are also provided at an institutional level (network and data storage facilities of the University of Amsterdam, IVI-cluster).

5.6 Describe your strategy for publishing the analysis software that will be generated in this project.

The research will rely on Deep Learning frameworks for the analysis software. The open-sourced and freely available PyTorch and TensorFlow will be primarily used. All source code will be made available through Github and an extract will be made part of the data archival package. Also, an overview of version dependencies to the DL-frameworks will be provided with/on (data)

publication.

6. Data management costs

6.1 What resources (for example financial and time) will be dedicated to data management and ensuring that data will be FAIR (Findable, Accessible, Interoperable, Re-usable)?

The generated data will not require extra financial resources. However, the data will be polished and prepared within two weeks after the publication of the respective research. This effort will be factored in the camera-ready preparation for publicizing research. The GitHub and SURFdrive functionalities will enable to make sure that data will be FAIR. Also, UvA Figshare will be used as a storage location for your publication data. This will make it possible to link to data from/across Pure and make data FAIR.