Giuseppe Marra - FWO DMP

Project Name FWO Post-Doc - 1239422N - Giuseppe Marra - FWO DMP

Grant Title 1239422N

Principal Investigator / Researcher Giuseppe Marra

Description This research project aims at combining deep learning with reasoning techniques based on mathematical logic and probability theory. Deep learning will not be treated as a standalone paradigm but as an important ingredient of a larger theory. This challenge is a fundamental step towards systems that exhibit both pattern recognition and reasoning skills, leading to more general Artificial Intelligence.

Institution KU Leuven

1. General Information

Name applicant

Giuseppe Marra

FWO Project Number & Title

1239422N - Deep Statistical Relational Learning

Affiliation

- KU Leuven

2. Data description

Will you generate/collect new data and/or make use of existing data?

- Reuse existing data

Describe in detail the origin, type and format of the data (per dataset) and its (estimated) volume. This may be easiest in a table (see example) or as a data flow and per WP or objective of the project. If you reuse existing data, specify the source of these data. Distinguish data types (the kind of content) from data formats (the technical format).

Knowledge Graph Completion

Knowledge graphs are represented as a collection of triples {(h,r,t)}. The task is to either predict unseen relations r between two existing entities: (h,?,t) or predict the tail entity t given the head entity and the query relation: (h,r,?).

- *Smokers (https://alchemy.cs.washington.edu/data)*
- *Nations (https://alchemy.cs.washington.edu/data)*
- *UMLS (https://alchemy.cs.washington.edu/data)*
- *Kinship (https://alchemy.cs.washington.edu/data)*
- *Countries (https://github.com/mledoze/countries)*
- *FB15k (https://github.com/dddoss/tensorflow-socher-ntn/tree/master/data)*
- *WN18RR (https://github.com/dddoss/tensorflow-socher-ntn/tree/master/data)*

Graph Generation.

Given a set of graphs, generate new graphs within the training distribution

- *ChemBL (version:ChEMBL_26) https://www.ebi.ac.uk/chembl/*

Graph Classification.

Given a set of graphs and the corresponding labels, classify new, neverseen, graphs.

- *CSL (https://github.com/graphdeeplearning/benchmarking-gnns)*

- *MUTAG (https://chrsmrrs.github.io/datasets/docs/datasets/)*
- *IMDB-B (https://chrsmrrs.github.io/datasets/docs/datasets/)*
- *IMDB-M (https://chrsmrrs.github.io/datasets/docs/datasets/)*
- *PROTEINS (https://chrsmrrs.github.io/datasets/docs/datasets/)*
- *PTC (https://chrsmrrs.github.io/datasets/docs/datasets/)*
- *NCI1 (https://chrsmrrs.github.io/datasets/docs/datasets/)*
- *ENZYMES (https://chrsmrrs.github.io/datasets/docs/datasets/)*

Neural-Symbolic Tasks

Given some feature-based representation (e.g. images) of entities, solve a relational reasoning task described by a logic theory (in form of FOL theory or logic program).

- MNIST - http://yann.lecun.com/exdb/mnist/
- HWF - https://github.com/liqing-ustc/NGS (src: https://www.cs.rit.edu/~crohme2019/task.html)

3. Legal and ethical issues

Will you use personal data? If so, shortly describe the kind of personal data you will use. Add the reference to your file in KU Leuven's Register of Data Processing for Research and Public Service Purposes (PRET application). Be aware that registering the fact that you process personal data is a legal obligation.

- No

We will only use existing datasets, which is no way will be related to personal data.

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, add the reference to the formal approval by the relevant ethical review committee(s)

- No

There are no ethical issues, as all data used within this project do not involve experiments on humans or animals. Regarding dual use, we also do not foresee any issues for the methodological research performed in this project.

Does your work possibly result in research data with potential for tech transfer and valorisation? Will IP restrictions be claimed for the data you created? If so, for what data and which restrictions will be asserted?

- No

Do existing 3rd party agreements restrict dissemination or exploitation of the data you (re)use? If so, to what data do they relate and what restrictions are in place?

- No

There are no 3rd party agreements restricting dissemination or exploitation and all licenses of existing data are permissive.

4. Documentation and metadata

What documentation will be provided to enable reuse of the data collected/generated in this project?

The project does not involve the creation of data but it will only involve reusing existing data. For software, experiments will be accompanied by standard README files that guide the use of scripts and data loading.

Will a metadata standard be used? If so, describe in detail which standard will be used. If no, state in detail which metadata will be created to make the data easy/easier to find and reuse.

- No

Since no standard is formally acknowledged in the relational AI context, we will use common best practices and ensure that all experiments are accompanied by running scripts and README files.

5. Data storage and backup during the FWO project

Where will the data be stored?

The data and software will be stored both centrally, on storage facilities of the research unit, and on Github external repositories. Moreover, the majority of these datasets are already hosted on Github repositories.

How is backup of the data provided?

The data are backed up automatically by the back-up procedures of the university central servers. Moreover, we also use backup facilities of online 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

The datasets that will be used are in the order of hundreds of MB and are already stored in the central storage facilities.

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

Data storage costs on central storage facilities are covered by the research group. We will only be storing ~ 10GB. Eventual additional costs will be marginal and can be easily covered by the bench fee.

Data security: how will you ensure that the data are securely stored and not accessed or modified by unauthorized persons?

Data are not owned by us and are already freely accessible on the Web, so no security provisions are in order.

6. Data preservation after the FWO project

Which data will be retained for the expected 5 year period after the end of the project? In case only a selection of the data can/will be preserved, clearly state the reasons for this (legal or contractual restrictions, physical preservation issues, ...).

All data is already available online. Software will be preserved on our central servers, including build and launch scripts to ensure reuse. Public releases of our software might also be stored on GitHub to maximise disseminiation.

Where will the data be archived (= stored for the longer term)?

All data is already available online. The version of the data and software used during the project will be stored on the university's central servers (with automatic back-up procedures) for at least 10 years, conform the KU Leuven RDM policy.

What are the expected costs for data preservation during the retention period of 5 years? How will the costs be covered?

Data storage and preservation costs are marginal given the small size of the data (~10GB). They are already covered by the research unit and additional costs can be easily covered by the project bench fee (~54 EUR/year for renting 100 GB at KU Leuven central servers).

7. Data sharing and reuse

Are there any factors restricting or preventing the sharing of (some of) the data (e.g. as defined in an agreement with a 3rd party, legal restrictions)?

- No

The data are publicly available on the Web. No 3rd party agreements or legal restrictions will prevent the sharing of software or data.

Which data will be made available after the end of the project?

The datasets are already publicly available on the Web. Software will be released with a permissive license.

Where/how will the data be made available for reuse?

- In an Open Access repository

Software will be shared on Github. The datasets are already publicly available on the Web.

When will the data be made available?

- Upon publication of the research results

The datasets are already publicly available on the Web. The software will be uploaded to open access repositories upon acceptance or publication of the research results.

Who will be able to access the data and under what conditions?

The datasets are already publicly available on the Web, with linceses that allow the reuse for research. Software will be released with permissive licenses as well.

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

There are no costs for data sharing, since we will share on free platforms (e.g. GitHub)

8. Responsibilities

Who will be responsible for data documentation & metadata?

The researchers involved in the project are responsible for documenting their code and datasets with proper metadata.

Who will be responsible for data storage & back up during the project?

The researchers are responsible for safely storing all software and data on the university's servers, such that backup strategies are automatically taking care of.

Who will be responsible for ensuring data preservation and reuse ?

The researchers are responsible for ensuring data preservation and reuse

Who bears the end responsibility for updating & implementing this DMP?

The PI bears the end responsibility of updating & implementing this DMP.