Plan Overview

A Data Management Plan created using DMPonline

Title: Human-robot interaction

Creator: Haoxuan Deng

Principal Investigator: Haoxuan Deng

Affiliation: Cranfield University

Template: DCC Template

Project abstract:

Develop models and algorithms for consenting understanding and information in the human-machine system for task performance. Create a common ground for people to communicate with machines so that automatic systems can reduce the mental and physical workload of human beings during collaborative working.

ID: 105391

Start date: 18-04-2022

End date: 18-04-2025

Last modified: 15-08-2022

Human-robot interaction

Data Collection

What data will you collect or create?

- Natural language corpus including technical scripts, reports, manuals, and introductions in industrial scenarios for creating domain-specific knowledge base using ontology and for Natural Language Processing (NLP) training; Up to 500M txt files named using the format "DateSource_ScriptsNames_DateOfCollection".
- The open source training dataset for machine learning model training; txt file collected from websites such as Kaggle and Github with a defined name.
- Experimental data collected from the collaborative robot platform to validate proposed models and algorithms; Up to 2GB hybrid data including csv, jpg, png data named using the format "DataSource_DataBriefDescription_DateOfCollection".
- Simulation data collected from the robot simulator to test and assess the machine learning (reinforcement learning) models. Up to 5GB video for demonstrating the robot running in the virtual simulation environment, named using the format "SimulationProcess_DateOfVideo".
- Literature view paper list; Using data management tools Mendeley and Notions for paper management; Approximately 500 papers for the project reference, citation, and supporting.
- PhD report document, around 15GB Latex named "PhdTitle_Name_Year", publication document, up to 10GB Latex named "TitleOfPublication"

How will the data be collected or created?

One part of the machine learning dataset will be collected from open-source platforms such as Wikidata, Kaggle, and Github for the model prototype development and validation. Filenames will be denoted as "DataContent_DataSource_CollectedTime".
Other parts come from natural scripts or texts in the manufacturing domain for model testing and improvement. The main testbed will be the Google Colab opening programming platform with GPU accessibility. Filenames will be denoted as "DataContent_DataSource_CollectedTime".
Regarding the robot simulation and experiment, the simulation data can be generated using the simulator tools such as Gazebo in ROS, and the sensor feedback can record the experiment data. All further analysis and plots can be finished using MATLAB, namely "Result_PlotDate". The simulation results will be validated on a real robot platform.

Documentation and Metadata

What documentation and metadata will accompany the data?

- Some state-of-the-art papers, reports, and documentation of the NLP model's performance on a specific task reported on the website "Browse state-of-the-art" can give a firm reference and benchmark on how a model structure works. Thus, it is helpful to indicate the application of the models.
- API documentation about the open-source machine learning framework such as Tensorflow and Pytorch can be accessed on the official website, which is the necessary documentation to clear the usage of each function and make it easy for further development.

Ethics and Legal Compliance

How will you manage any ethical issues?

I have created my own CURES form for my research project, and I have gained to know the ethical issues and regulations by reading the handbook for Cranfield Ph.D. students. A consensus will be given to each participant before the experiment starts so that

everyone can clarify the generation meaning, purpose, sharing, preservation, and relevant usage of collected data. They have a right to withdraw their data if they do not want to publish their data.

Anonymisation will be used when data is shared with each participant in the experimentation by numbering individuals rather than using the name to protect the identity.

As for sensitive data, they will be collected and stored on the Ph.D. student's encrypted laptop with a backup on Cranfield university's Microsoft Onedrive account which is limited access with a double authentification system using the student's phone. A request is needed before anyone gets permission to get access to the data source.

How will you manage copyright and Intellectual Property Rights (IPR) issues?

According to the guiding books for Cranfield's Ph.D. students, all data and intellectual property belong to and vest in the University, and all rights in such intellectual property are and shall be assigned to and vested in the University.

These datasets and IRP will be licensed with permissive licenses (such as MIT licenses) or other types of licenses after discussing with supervisors and enquiring from the data management team in the university.

Basically, all datasets and models in the project will be open and accessed via some code repository such as Github with GPL licenses, after getting permission from supervisors and the university data management team.

As for commercial usage or patent publication, it is allowable after a formal request and an appropriate reference.

Storage and Backup

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

All relevant data and documentation will be recorded and updated regularly using the Cranfield university Onedrive account. It is managed and accessible by Ph.D. students but can be open to specific project-related persons such as supervisors, group members, and third parties with valid permission.

Data will be saved on a restricted-access drive on the University network which is automatically backed up by Cranfield IT, on a daily basis to multiple data centers. The data including model data, results, and documentation will be backed up using the University Onedrive account. In addition, some temporary datasets or prototypes will be stored using a personal Google drive for personal research usage. Also, Github will be used to publish code, dataset, and result with valid permission when a paper is published or the project is accomplished.

Git will be used as a version control tool to manage documentation.

Finally, an offline backup using a safe USB with authorization will be applied to record and update data.

How will you manage access and security?

Github and personal Google accounts will only be used for research purposes which are set to be private and limited access to specific users. The final outcome can be accessible on Github freely and if authorized.

All relevant data will be managed and accessed using an allocated laptop or desktop in the Center for Digital Manufacturing Engineering (CDME) at Cranfield. Also, the Microsoft Onedrive and Mendeley are connected to the university Ph.D. student account, and group members, collaborators, and supervisors can get the data sources after gaining permission.

Selection and Preservation

Which data are of long-term value and should be retained, shared, and/or preserved?

For sensitive data in the project, they will be retained, shared, persevered, and destroyed after a full discussion with supervisors, the university data management team, and collaborators.

For non-sensitive data, they will be stored and open access on Github for future development such as an update or modification for a start-up business.

What is the long-term preservation plan for the dataset?

"Data will be retained securely in Cranfield’s institutional data repository, CORD, which uses the shared platform and preserves data for at least 10 years after the project end, with datasets assigned a DOI for long-term accessibility, in accordance with Cranfield’s Management of Research Data Policy."

Data Sharing

How will you share the data?

Non-sensitive data, determined by full considerations and discussions, will be open and accessed on Github, following the process of the project step by step.

Are any restrictions on data sharing required?

All the project-related data will be opened to Cranfield Center for Digital Manufacturing Engineering (CDEM) staff, members, and supervisors.
If patent applications on this project outcome are required, then this data will be limited to people relevant to the project.
If restricted publications on the dataset are required, then the dataset will only be accessible to contributors of the publication.
For data sharing under an embargo, the data can still be available by making bilateral agreements with individual researchers, groups, or even third parties from the industry. Therefore, the data is used in line with the consent obtained. A formal data sharing or collaboration agreement will be involved and compiled for this purpose.

Responsibilities and Resources

Who will be responsible for data management?

The project investigator, Haoxuan Deng will take the responsibility for data collection, accessibility, management, and storage. The data will be reviewed and discussed in regular meetings with supervisors and collaborators.

What resources will you require to deliver your plan?

Google Colab platform
ROS system for robot programming and simulation
The project will take advantage of the facilities available in B30 and other labs in Cranfield.