Data Management Plan

Artifacts of this project consist of the following items: open-source software, models, training data/testing data, publications, presentations, and tutorials.

Open-source Software Release: We will release our source code under licenses that encourages both free redistribution (e.g., GPL) as well as commercialization by U.S. companies (e.g., BSD). GitHub will be used for the development of all our software, as it offers the distributed version control and source code management (SCM) functionality of Git. Github also provides access control and several collaboration features such as bug tracking, feature requests, task management, continuous integration, a forum for discussions and wikis for every project, which are regularly used for software development by researchers both in academia and industry.

Standards for data and metadata format and content: Data formats will be tagged in XML forms to allow for broader exchange, and dataverse (`https://dataverse.org`) will be the preferred standard.

Publications: We will release all our published papers under the copyrights of the appropriate publishers, subject to the condition that all published papers will be made freely available on our own websites. Each publication will acknowledge all sources of funding.

Training and Testing Data: We will make all data (for instance, training, testing, and generated data where applicable) available through stable public repositories (ieee-dataport, papers with code, and others). As relevant and as we have done in the past, we will also assign shared data collections permanent identifiers (DOIs) for citation.

Presentations and Tutorials: We will make presentations and tutorials we intend to organize at relevant conferences publicly available, mirrorring them on our websites. Seminal presentations will be additionally disseminated as youtube videos. We also plan to build a project website dedicated to this project. The website will provide clear and concise descriptions of materials.

Data Management: We will adhere to federal, state, and university requirements for collection, dissemination, and control of data. Standard authentication and authorization methods (including group access) based around the public domain SSH system will be employed to control access to the code base and unreleased data sets.

Data Security: All artifacts of this project will be stored on facilities at George Mason University. All electronic data will be redundantly archived. Our laboratories have secure servers, with hard drives set up in a RAID that is capable of full recovery even if the disk failure happens. We plan to maintain the secure server for at least a decade following the completion of the project. The information will also be replicated on external public Web servers such as GitHub (see above) for long-term durability and reliability.

Track Record in Data Management: We have a strong track record of disseminating papers along with publicly-available code and data in their respective communities, as well as across communities.

Recent examples from Liu include:

1. `https://github.com/MingruiLiu-ML-Lab`
2. `https://github.com/MingruiLiu-ML-Lab/Federated-Sparse-Learning`
3. `https://github.com/MingruiLiu-ML-Lab/Communication-Efficient-Local-Gradient-Clipping`
4. `https://github.com/MingruiLiu-ML-Lab/episode`

Recent examples from Nguyen include:

1. https://github.com/dynaroars
2. https://github.com/dynaroars/dig
3. https://github.com/dynaroars/gentree
4. https://gitlab.com/sosy-lab/benchmarking/sv-benchmarks/-/tree/main/c/nla-digbench

Recent examples from Agnarsson include:

1. https://science.gmu.edu/directory/geir-agnarsson