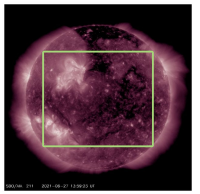
A Federated Distributed Learning Benchmark for solar wind speed forecasting using solar EUV images



Filip Svoboda*, Gianluca Mittone*, Ed Brown, Nic Lane, Pietro Lio *equal contribution

Introduction

Distributed training is the future of on-board computation in space as it offers scalability, resilience, and flexibility that can not be matched by a centralized setup. In the communication space it trades-in the cost of a full-dataset aggregation for that of an intermittent exchange of training messages.

On-board computation is relevant to current missions and necessary for the future ones as it offers researchers the opportunity to greatly reduce the communication costs associated with sending large amounts of data long way to the Earth.

Extreme solar winds, can impact communication, disrupt satellites and spacecraft. Consequently, accurately forecasting the solar wind speed on-board is an important proving ground for on-board distributed training.

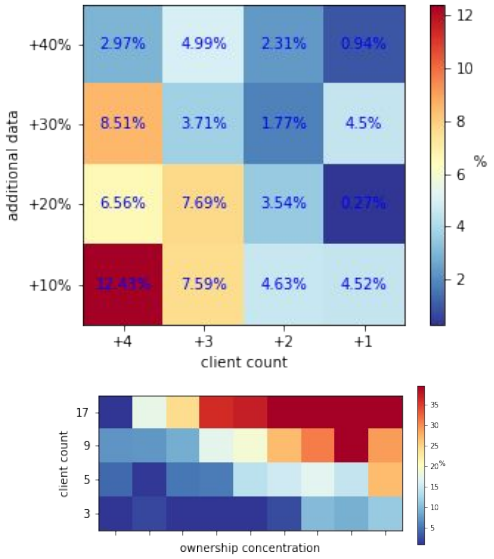
Research Objectives

This study distributes the training of the 2022 Svoboda, Brown et al. solar wind speed model across 3 to 17 clients, nodes, or spacecraft. It uses data by OmniWeb and the Solar Dynamics Observatory (SDO). The resulting dataset occupies 35GB.Results for forecasting at a four-day lag from a single 211 Å image are presented.

Benchmark Performance

The basic and the federated distributed training achieve their objective of replication, as in our experiments they follow the same general training curve patterns as centralized setup and achieve statistically indistinguishable *Mean Squared Errors equal to 0.098*.

However, they differ markedly in their communication costs. The basic distributed setup requires the communication of all gradients at each training step. In our setup this is 330MB worth of gradients 1095 times per epoch, or about 361.35GB in total. This compares very unfavorably with the 35GB size of the full dataset. Meanwhile, Federated Learning can, in the most conservative and basic setup, work with communicating a comparable 330MB tensor of data just once per epoch. Once per 10 epochs is often used too.



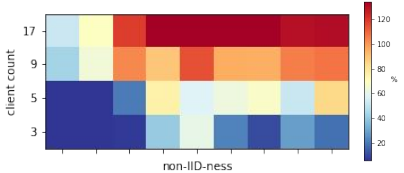
Experimental Setup

The figures present experiments that investigate the interplay between client count and individual major drivers of performance and generalization. Heath maps are used to illustrate performance loss relative to the centralized setup. Non-IID-ness increases towards right, as does equality of data split.

Conclusions

The benchmark performance experiments confirmed that distributed setups replicated the centralized case for up to 8 clients. Beyond this point there was a 20% performance fall

Ownership concentration experiments dive into the issue of client count tolerance dependence on the data ownership structure. *They show that thinly spread data distribution patterns hamper convergence as client models fail to converge.*



Non-IID-ness tolerance experiments investigate FL's resilience to increasing client counts as data distribution changes. *As each individual client's data became less diverse, i.e. focused only on a specific period of solar cycle (more non-IID), the tolerance of the system to higher client count decreased.*

Indeed our final experiments clearly demonstrates this *data-ownership trade-off.* A 40% of the data is held-out, resulting in a clear loss of performance. Adding it back in tranches of 10% and spread among 1-4 clients clearly shows that the highest rate of improvement is achieved when the data is added in through the smallest number of clients. Indeed, adding the first 10% in a single chunk gives 3x relative improvement than when spread out. Notably, a presence of minimal sufficient data requirement is suggested by the similarity between the 1-way and 2-way splits.

Put together, for the benefits of distributed training to be realised, one needs to pay close attention to the client data collection, ownership structure, and concentration as these were shown to be persistent and material determinants of performance.

This research was conducted to establish Federated Learning (FL) benchmarking for the Heliophysics domain. We would like to thank Trillium Technologies Inc for supporting and managing the project and Google Cloud for providing the compute resources.