Distributed Deep Learning for Surface Soil Moisture Estimations

Þór Arnar Curtis
Advisors: Morris Riedel, HÍ and Helmut Neukirchen, HÍ
May, 2022

UNIVERSITY OF ICELAND

FACULTY OF INDUSTRIAL ENGINEERING, MECHANICAL ENGINEERING AND COMPUTER SCIENCE

Objective

The objective of this study is to create Deep Learning (DL) models that use Remote Sensing (RS) satellite data to predict the Soil Moisture Content (SMC) of the top 10 cm layer of soil in selected areas.

Introduction

The business partner for this project is a German startup focused on providing soil moisture estimations to the agricultural business. During this project data currently in use by the business partner will be evaluated, results of their best-performing Random Forest (RF) model reproduced, and metrics from that model will be created for a new global dataset. New Artificial Neural Network (ANN) models and Recurrent Neural Network (RNN) models will be created, and the results compared to the metrics from the RF model.

DL is a type of Machine Learning (ML) and Artificial Intelligence (AI) that imitates the way humans gain certain types of knowledge. DL models require access to immense amounts of training data and processing power.

RS is the acquisition of information about an object or phenomenon without making physical contact with the object, in contrast to in situ or on-site observations.

SMC is an important factor in land surface hydrology and has important implications for agriculture, ecology, wildlife, and public health. In agriculture, soil moisture plays an important role in the exchange of water, energy, and carbon and is an indispensable indicator of ecological droughts that can threaten agricultural production.

High-Performance Computing (HPC)

The training of Deep Neural Networks (DNNs) is a significant computing task that places high demands on computing resources and memory bandwidth. The use of Graphics Processing Units (GPUs) can greatly speed up such training, especially when working with large quantities of data. As GPUs have a large number of parallel arithmetic units, they can handle thousands of threads concurrently. The training of DNNs, and hyperparameter tuning are computationally expensive tasks that greatly benefit from the power of HPC.

Figure 1 DEEP-EST modular HPC system.

During this project, model training was partially performed on Google Colab, but DNN training and hyperparameter tuning was solely performed on the DEEP-EST modular HPC system at Jülich Supercomputing Centre (JSC). The usage of Jupyter notebooks enabled interoperability between cloud solutions where the Tensorflow framework was used for ML.

Data

The satellite data in this project is from the Sentinel-1, Sentinel-2, and Landsat-8 satellites. Geographical data describing the slope, elevation, and aspect of the area is acquired from NASA's TanDEM-X mission. In addition to the remote sensing satellite data, in situ data sources are used to gain information on the soil composition, precipitation, and finally, soil moisture measurements which are used as ground truth for model training.

The dataset used throughout this project was the global dataset, containing 27 features and 533,794 samples from observation stations around the globe.

Interpolation was used to fill missing data and experiments were done both with and without outliers and with different scalers. A sliding window approach was used for RNN modeling. Hyperparameter tuning was performed with Ray Tune running on the DEEP-EST HPC system.

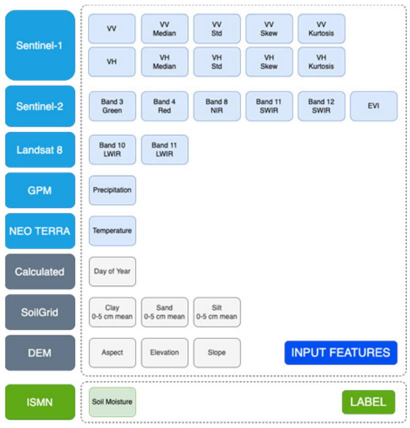

Figure 2 Global dataset features.

Results

The model producing the lowest RMSE on the test set was the resulting ANN model from the Ray Tune hyperparameter tuning using the global dataset scaled with the MinMaxScaler. The model has two layers with 254 neurons.

Table 1 The RMSE value of each best model.

Model	Training RMSE	Validation RMSE	Test RMSE
Random Forest	0.046	0.094	0.111
ANN	0.081	0.093	0.103
GRU	0.095	0.109	0.122
LSTM	0.095	0.107	0.123

The best ANN model produced a 7% lower RMSE than that of the previous best RF model. The RMSE of each model on the global test set can be seen in table 1, and a plot of the ANN's estimated vs predicted values in figure 3.

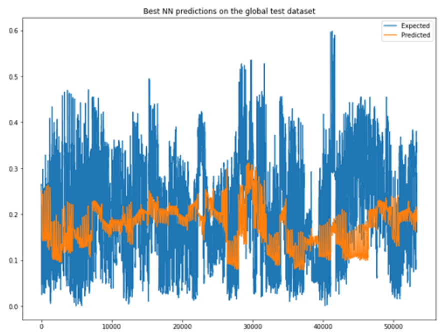

Figure 3 ANN model predictions.

Acknowledgments

Special thanks go to my great supervisors for excellent support during this project, to Josefine Busch for much-appreciated support and innovative ideas.