Surveying the Machine Learning Landscape in Earth Sciences

Muthukumaran Ramasubramanian\textsuperscript{1}, Katrina Virts\textsuperscript{1}, Ashlyn Shirey\textsuperscript{1}, Ankur Kumar\textsuperscript{1}, Muhammad Hassan\textsuperscript{1}, Ashish Acharya\textsuperscript{1}, Rahul Ramachandran\textsuperscript{2}, and Maskey Manil\textsuperscript{2}

\textsuperscript{1}The University of Alabama in Huntsville
\textsuperscript{2}NASA/Marshall Space Flight Center

Several recent papers have investigated different challenges in applying machine learning (ML) techniques to Earth science problems. The challenges listed range from interpretability of the results to computational demand to data issues. In this paper, we focus on specific challenges listed in the review papers that are centered around training data, as the size of training data is important in applying deep learning (DL) techniques. We are in the process of conducting a literature survey to better understand these challenges as well as to understand any trends. As part of this survey, our review has encompassed Earth science papers from AGU, AMS, IEEE and SPIE journals covering the last ten years and focused on papers that utilize supervised ML techniques.

Our initial survey results show some interesting findings. The use of supervised machine learning techniques in Earth science research has increased significantly in the last decade. The number of atmospheric science papers (i.e., from AMS journals) using ML approaches has increased by over 40%. Across all of Earth science even larger changes have occurred, including a >90% increase in AGU papers and a >10-fold increase in IEEE papers using ML.

We also conducted a deep dive into all the papers from AGU journals and uncovered interesting findings. There is a prevalence of the use of supervised ML in certain sub-disciplines within Earth science. The biogeoscience and land surface research communities lead in this area: over 20\% of papers published in Global Biogeochemical Cycles, JGR Biogosciences, JGR Earth Surface, and Water Resources Research use supervised ML techniques, including over 35\% of the papers in JGR Biogeosciences. The availability of labeled training data in Earth science is reflected in the number of training samples used in supervised analysis. In the papers we surveyed, most ML algorithms were trained using small (i.e. hundreds of labeled) samples. However, for some applications using model output or large, established datasets, the number of training data ranged several orders of magnitude greater.

In this presentation, we will describe our findings from the literature survey. We will also list recommendations for the science community to address the existing challenges around training data.