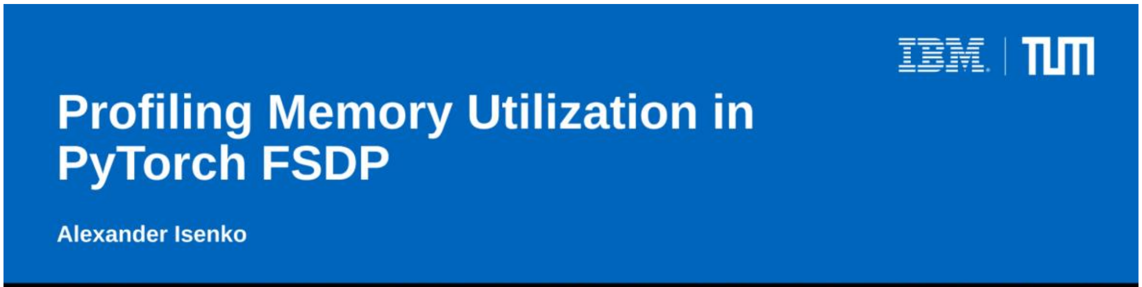
NEWS LETTER

04

NOVEMBER | 2023

CLOUD STARS

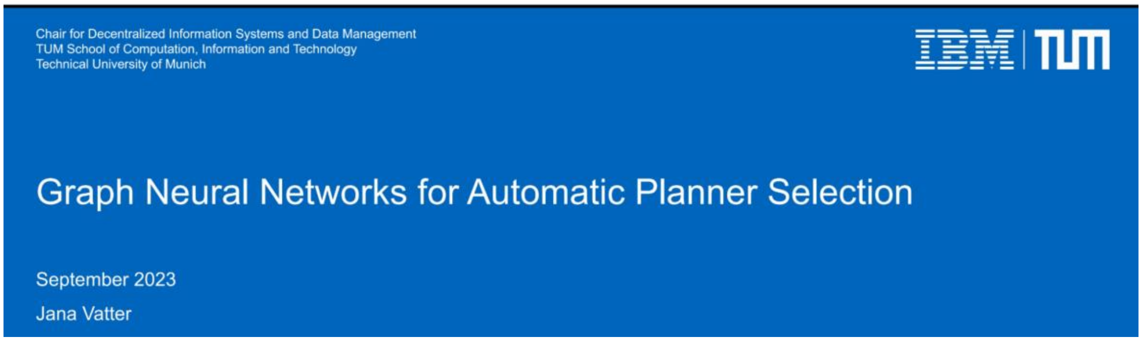


Profiling Memory Utilization in PyTorch FSDP

Alexander Isenko

A scalable data summarization system for MLOps – an extension for MVI

Herbert Woisetschläger September 2023



Chair for Decentralized Information Systems and Data Management
TUM School of Computation, Information and Technology
Technical University of Munich

Graph Neural Networks for Automatic Planner Selection

September 2023

Jana Vatter

TUM Secondments at the IBM Thomas J. Watson Research Center

by Alexander Isenko, Jana Vatter, Herbert Woisetschläger (Technical University of Munich)

The Technical University of Munich (TUM) started its secondments within the CLOUDSTARS project with three Ph.D. students who visited the IBM Thomas J. Watson Research Center in Yorktown, New York, for two months each.

We focused on multiple topics, ranging from memory profiling of foundation model training, performance improvements for data summarization with vector databases to applying graph neural networks to improve planner selection tasks accuracies and runtime.

First, we evaluated how memory fragmentation affects the training of foundational deep learning models. Fragmentation reduces the amount of consecutively available memory for training, which, coupled with non-deterministic allocation patterns specific to distributed training, can lead to process termination. We found common training scenarios where this occurs, confirmed the suspicion of fragmentation being responsible for failed training runs, and evaluated the potential solutions and their pros and cons. Finally, we added to the discussion on memory allocation in DL frameworks with our blog post on our findings, the PyTorch developer discussion, and the open sourcing of the code and artifacts.

Second, data summarization is a known problem domain for everyone who generates data but cannot manually extract the important samples for the next training iteration due to sheer size. While some data processing pipelines attempt this, doing so at scale under latency bounds is non-trivial. We attempted to make a first step in this direction by creating a containerized pipeline with the help of TorchServe, Milvus, and FAISS. We improved the ingestion rate to 128 img/s from 20 img/s (6.1x), the clustering to 4.7k vectors/s from 0.6k vectors/s (7.1x), and added novel features, like similarity search, all while staying on the same hardware and keeping legacy feature parity. In addition to improved performance, we are trivially horizontally scalable due to a containerized workflow.

Finally, we improved the state-of-the-art by reproducing a paper from the automated planning community and using novel graph neural network (GNN) architectures to enhance performance. We made the results more tangible by porting the code to one of the most popular Deep Graph Learning frameworks and applying state-of-the-art GNN techniques to this domain. This research helps practitioners in the planning field and researchers in the GNN field to evaluate the general applicability of new techniques.

During this period, additional secondments from the University of Murcia (UM) and the Technical University of Vienna (TU Wien) took place at T.J. Watson, which allowed us to collaborate, explore synergies, and plan future projects. By having access to different IBM teams, we were able to find important research topics that could benefit from an interdisciplinary point of view, thus helping us progress individually and collectively.

CLOUD STARS

cloudstars.eu | twitter.com/Cloudstars_2023 | github.com/cloudstars-eu

CLOUDSTARS project has received funding from the European Union's Horizon research and innovation programme under grant agreement No 101086248