NEWS LETTER

15

JUNE | 2025

CLOUD STARS

Dublin Collaborations and IBM Yorktown Heights Summer Projects

by Barbara Hoffmann (University of Bayreuth)

THE PAST:

During her stay in Dublin, Jana Vatter (Technical University of Munich) engaged with the Ireland Dublin research team and members from the University of Murcia to explore promising research ideas. For her project, she focused on enhancing the scalability of Graph Neural Network training on very large graphs by addressing the accuracy loss that arises when the graph is partitioned. Her work centers on mitigating the effects of stale information exchanged between partitions, drawing on techniques from numerical analysis to iteratively refine these shared signals. By doing so, she seeks to preserve the memory-efficiency advantages of partitioned training while closing the performance gap with full-graph methods.

Barbara Hoffmannb (University of Bayreuth) also spent time in Dublin, where she developed an end-to-end system to adapt large language models (LLMs) for structured, multimodal data, such as molecular SMILES strings and protein sequences. Her project automates the entire pipeline: ingesting structured inputs, converting them into a knowledge graph, sampling from that graph, and generating synthetic text to fine-tune an LLM. By combining graph sampling, custom token creation, and low-rank adaptation techniques, domain experts are enabled to rapidly customize pretrained models for specialized areas like biochemistry. In essence, her work bridges the gap between complex multimodal datasets and large language models, streamlining the fine-tuning and deployment of AI across various fields.

THE PRESENT:

From the University of Bayreuth Nikita Agrawal, Barbara Hoffmann, and Simon Mertel are currently at the T. J. Watson IBM Research Lab in Yorktown Heights for the upcoming summer, where they are each working on distinct projects.

Nikita is building an LLM development platform, overseeing the end-to-end model pipeline, including a help chatbot feature. She is designing a scalable solution to support thousands of users for tasks such as data curation, synthetic data generation, model training, tuning, and evaluation.

Simon focuses on preventing catastrophic forgetting during fine-tuning of large language and multimodal models. He develops and evaluates techniques, backed by theoretical analysis, that preserve pre-trained knowledge while integrating new data. His work also delivers scalable fine-tuning algorithms, benchmark datasets, evaluation metrics, and empirical demonstrations showing improvements over existing approaches.

Barbara investigates the quality of post-training data and devises novel processing techniques aligned with established quality metrics. She applies statistical analyses to identify metrics that correlate strongly with downstream performance. Based on these insights, she designs methods to enhance data quality and improve the overall performance of the Granite model.

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