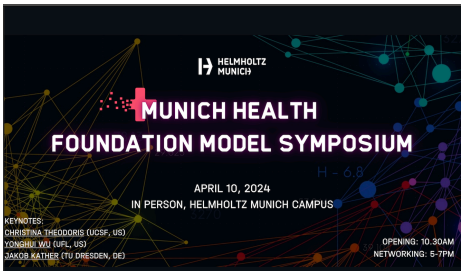

Contribution ID: **35** Type: **Talk**

DataDream: Few-shot Guided Dataset Generation

While text-to-image diffusion models have been shown to achieve state-of-the-art results in image synthesis, they have yet to prove their effectiveness in downstream applications. Previous work has proposed to generate data for image classifier training given limited real data access. However, these methods struggle to generate in-distribution images or depict fine-grained features, thereby hindering the generalization of classification models trained on synthetic datasets. We propose DataDream, a framework for synthesizing classification datasets that more faithfully represents the real data distribution when guided by few-shot examples of the target classes. DataDream fine-tunes LoRA weights for the image generation model on the few real images before generating the training data using the adapted model. We then fine-tune LoRA weights for CLIP using the synthetic data to improve downstream image classification over previous approaches on a large variety of datasets. We demonstrate the efficacy of DataDream through extensive experiments, surpassing state-of-the-art classification accuracy with few-shot data across 9 out of 10 datasets. Additionally, we provide insights into the impact of various factors, such as the number of real-shot and generated images as well as the fine-tuning compute on model performance.

Primary authors: KIM, Jae Myung (University of Tübingen); Ms BADER, Jessica (Helmholtz Munich)

Co-authors: ALANIZ, Stephan (Helmholtz Munich); AKATA, Zeynep (Helmholtz Munich)

Session Classification: Break + Posters session

Track Classification: Poster