UiO University of Oslo

February 10, 2025

Summer research projects for life science students

DeepXplain: Explainability Techniques for Trustworthy Deep Learning

Project Description: The increasing deployment of Deep Neural Networks (DNNs) in high-stakes applications raises critical concerns regarding their interpretability, fairness, and robustness. Despite advancements in explainability techniques, many approaches remain post-hoc and lack systematic evaluation in real-world scenarios.

This project will explore explainability methods for deep learning models, focusing on techniques that improve model transparency while maintaining predictive accuracy. Students will work on benchmarking state-of-the-art explainability tools, such as SHAP, LIME, and attention-based methods, and evaluate their effectiveness in different deep learning tasks. Potential areas of application include:

- Bias detection and mitigation in neural networks
- Factual consistency evaluation in language models
- Self-explaining architectures for trustworthy AI

Students will gain hands-on experience with deep learning frameworks (e.g., PyTorch, TensorFlow) and interpretability libraries, while also contributing to the development of open-source tools for explainability.

The project is suitable for BS or MS students with a background in NLP, machine learning and an interest in explainability, fairness, and AI ethics. Basic experience with Python and deep learning frameworks is recommended