.

Master Thesis Topic

Brain-Inspired Modularization for Deep Neural Network Repair

Motivation and Background

Deep neural networks (DNNs) have become central to solving complex tasks across various domains; however, ensuring their reliability and robustness continues to present significant challenges. Recent advances have introduced modularization techniques, particularly those inspired by brain dynamics, to enhance neural network performance and interpretability [1]. Concurrently, state-of-the-art methods for repairing neural networks have emerged to address vulnerabilities and improve reliability [2]. This task aims to investigate existing modularization strategies—emphasizing brain-inspired dynamic modularization—and current deep neural network repair methods. An integrated approach combining modularization with advanced repair techniques will be developed, aiming to improve the effectiveness, reliability, and maintainability of neural network models.

Goals

The goal of this thesis is to:

1. **Literature Review**. Research into existing modularization techniques for deep neural networks, particularly for dynamic emergence of modules, as well as the current state of the art repair techniques for deep neural networks.
2. **Technique.** Develop a technique which leverages the automatic modularization of neural networks to find repair candidates.
3. **Evaluation.** Compare your novel approach to existing neural network repair techniques, such as Inner [2]

Research Type

Theoretical Aspects:	****☆
Industrial Relevance:	***☆☆
Implementation	***☆☆

Prerequisite

The student should be enrolled in the master of computer science program, and has completed the required course modules to start a master thesis.

Skills required

We are seeking a candidate with **strong Python programming skills** and thorough understanding of the theory behind deep neural networks, including hands-on experience. Completed courses in software engineering-II or machine learning research is a plus, as is the ability to communicate results effectively.

Contacts

enrico.philip.ahlers@hu-berlin.de
Software Engineering, Institut für Informatik, Humboldt-Universität zu Berlin

References

[1]: Liu Z, Gan E, Tegmark M. Seeing is believing: Brain-inspired modular training for mechanistic interpretability. Entropy. 2023 Dec 30;26(1):41.

[2]: Chen Z, Zhou J, Sun Y, Wang J, Xuan Q, Yang X. Interpretability Based Neural Network Repair. InProceedings of the 33rd ACM SIGSOFT International Symposium on Software Testing and Analysis 2024 Sep 11 (pp. 908-919).