Incremental Model "Patching" for Industrial Machine Learning Systems

Bachelor project for 2024/2025
Departments of Artificial Intelligence and Sustainability
Prof. Dr. Ralf Herbrich

Motivation. The traditional view of machine learning model development is often depicted as a linear two-step pipeline: model training followed by testing, prior to deployment. However, this paradigm is insufficient for the dynamic requirements of industrial machine learning systems, where models must be continuously adapted to maintain optimal performance. We propose a comprehensive exploration of incremental model "patching," focusing on two critical areas: safe fine-tuning techniques and modular model adaptation.

Adaptation of machine learning models is imperative due to varying data distributions, necessitating different model behaviors over time. For instance, a model that frequently recommends Christmas songs would be inappropriate in August but highly effective in December. Additionally, the advent of large foundational models requires specialized adaptation for specific use-cases while controlling unintended behaviors, which can be challenging.

Propsoal. The primary objective is to investigate techniques for incremental model adaptation to address two significant challenges:

1. Ensuring the safety and integrity of fine-tuned models to prevent misuse.
2. Effectively adapting models with limited data to achieve performance comparable to large-scale in-context learning models.

Safe Fine-Tuning Techniques. Large foundational models often have restrictions to prevent malicious use, such as generating toxic content. However, these restrictions can be bypassed through fine-tuning [1] or other techniques.

In this project we will create an open-source library to safely fine-tune LLMs. The library will offer the possibility to fine-tune a set of pre-selected models and explicitly control the amount of toxic content generated by the model before and after the fine-tuning operation, offering virtuous users a safe way to fine-tune aligned LLMs and preventing malicious users from work around the limitations posed in the alignment phase by the model creators.

For this purpose we will focus on parameter-efficient fine-tuning strategies (e.g., using LoRA [2]) that can offer good predictive performance with limited computational cost and design our strategies to prevent the output of toxic content around them. We will also use datasets such as the AdvBench used in LLM-Attacks (https://github.com/llm-attacks/llm-attacks) alternated or mixed with other standard datasets such TriviaQA and Squad.

Objectives:

- Develop a fine-tuning methodology that preserves model safety.
- Evaluate the performance of the fine-tuned models in standard use-cases.
- Test the robustness of the fine-tuning strategy against potential bypass attempts.

Modular Model Adaptation. Despite the efficiency of adapting pre-trained models, the overwhelmingly frequent usage of large-scale in-context learning highlights the challenges posed by data scarcity. In this project we will create an open source library that given a small set of data points (e.g., fewer than 100 data points) will be able to create a model that outperforms the base model by retrieving a combining fine-tuned versions of such model from a library (e.g., see [3,4]).

For the purpose of this project we will use small models like LLama3-8B and Mistral7B which are widely used and for which there are a number of derived versions available on Huggingface.co. The models created with our library will be tested on datasets such (or parts of datasets from) MMLU, Squad (in different languages), PubMedQA, FrBMedQA.

Objectives:

- Develop a methodology to effectively utilize scarce data for select available public model adaptations.
- Definition of a strategy to leverage combinations of model adaptations.
- Compare the adapted model's performance with large-scale in-context learning models.

Project Partner. Amazon Research in Berlin develops novel methods in machine learning (ML) and data science. The partner team is focusing specifically on applications of ML in Amazon Web Services. All methods developed by the team will be directly applicable to production systems serving millions of customers daily. More information about the partner team can be found at https://www.amazon.science/ or email ralf.herbrich@hpi.de. We will have bi-weekly meeting with the project partner to both receive feedback and present project progress.

References.

1. Zheng et al., “Fine-tuning Aligned Language Models Compromises Safety, Even When Users Do Not Intend To!”, https://arxiv.org/abs/2310.03693
2. Hu et al., “LoRA: Low-Rank Adaptation of Large Language Models”, https://arxiv.org/abs/2106.09685
3. Huang et al., “LoraHub: Efficient Cross-Task Generalization via Dynamic LoRA Composition” https://arxiv.org/abs/2307.13269
4. Ostapenko et al., “Towards Modular LLMs by Building and Reusing a Library of LoRAs” https://arxiv.org/abs/2405.11157

Partner Mentors.

Giovanni Zappella
Principal Applied Scientist at AWS
giovanni.zappella@gmail.com

Martin Wistuba
Sr. Applied Scientist at AWS
mwistuba1@gmail.com