PROJEKAT STUDENTSKIH ISTRAŽIVANJA

PROJEKAT STUDENTSKIH ISTRAŽIVANJA - PSI 2024

Implementing fast large language models (LLM) in Codon programming language
(Implementacija velikih jezičkih modela u Codon programskom jeziku)

Ibrahim Numanagić
University of Victoria
Victoria, BC, Canada

Python is arguably the most popular programming language nowadays, and is thus widely used for machine learning research. Recently, the hottest topic in machine learning are large language models (LLMs)[1] that are used for general-purpose language generation and various other natural language processing tasks such as machine translation. However, the base Python is not an ideal language for implementing those models due to its performance; thus, the most performant LLMs are currently implemented in C, C++ or Rust. We have recently developed Codon,[2] a static ahead-of-time compiler for Python, and a highly optimized implementation of NumPy[3] that enables Python applications to achieve the speed of C implementations with ease. In this project, we are looking to apply Codon and its NumPy libraries to implement various LLMs, such as Gemini,[4] Llama,[5] Mixtral,[6] and to compare the new implementations with the current state-of-the-art implementations in other languages.

Tasks:

- Read and understand the literature
- Explore various LLMs (Gemini, Llama, Mixtral etc.)
- Implement these LLMs in Codon
- Compare the accuracy and the performance between the reference LLM implementations and the Codon implementations
- Write a report detailing the improvements and the differences

References:

1. Brown, T. B. *et al.* Language Models are Few-Shot Learners. (2020) doi:10.48550/ARXIV.2005.14165.
2. Shajii, A. *et al.* Codon: A Compiler for High-Performance Pythonic Applications and DSLs. in *Proceedings of the 32nd ACM SIGPLAN International Conference on Compiler Construction* (ACM, New York, NY, USA, 2023). doi:10.1145/3578360.3580275.
3. Harris, C. R. *et al.* Array programming with NumPy. *Nature* **585**, 357–362 (2020).
4. Gemini Team *et al.* Gemini: A family of highly capable multimodal models. (2023) doi:10.48550/ARXIV.2312.11805.
5. Touvron, H. *et al.* LLaMA: Open and efficient foundation language models. (2023) doi:10.48550/ARXIV.2302.13971.
6. Jiang, A. Q. *et al.* Mixtral of Experts. (2024) doi:10.48550/ARXIV.2401.04088.