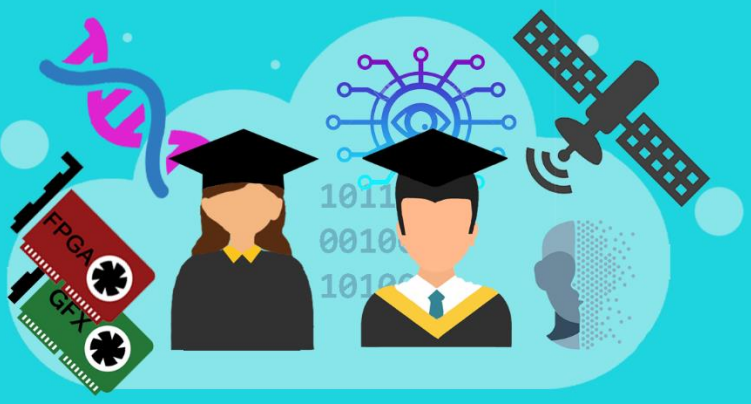
Diploma Thesis

Microproccessors and Digital Systems Laboratory



Leveraging Large Language Model Prompting for Rapid High-Level Synthesis Kernel Optimization

The fields of Deep Learning and Natural Language Processing (NLP) have experienced significant advancements, particularly with the emergence of Large Language Models (LLMs) like Generative Pre-trained Transformers (GPT) [1]. LLMs are deep learning models with a large number of learnable parameters, trained on extensive text datasets, enabling them to excel in a wide range of natural language understanding and generation tasks with remarkable accuracy. These models have also been adopted for computer-related language tasks, such as GitHub's Copilot [2], which provides autocomplete-style coding suggestions, and Microsoft's Codebert, a versatile model for feature extraction that can be fine-tuned for various programming language tasks [3].

The rapid adoption of LLMs for source-code analysis and software development tasks presents a promising opportunity for their application in more domain-specific contexts. In particular, within FPGA programming using higher-level languages like High-Level Synthesis (HLS), this shift offers a way to harness LLMs to increase design abstraction and reduce the complexities of hardware development. By providing real-time insights and guidance as developers code, LLMs enable on-the-fly assessment and optimization of HLS applications [4].

In this diploma thesis, you will use a GPT e.g., ChatGPT to automatically optimize kernels using High-Level Synthesis directives. You will try to induce domain spacific knowledge on the LLM e.g using the RAG architecture [5] to guide it towards an optimal solution. The user will aquire knowledge on the way you generally guide pre-trained model to give you the expected answer. Besides the LLM-based concepts the user will be able to familiarize with Vitis [6] the SotA framework for designing kernels using HLS on AMD FPGAs.

PREREQUISITES

Strong knowledge of **Python** and **Bash Scripting**

Desirable: Familiarity with **High Level Synthesis**, **Machine Learning**, and **LLM Prompting**

RELATED MATERIAL

[1] Gao, Tianyu, Adam Fisch, and Danqi Chen. "Making pre-trained language models better few-shot learners." arXiv preprint arXiv:2012.15723 (2020).

[2] Your AI Pair Programmer

[3] Feng, Zhangyin, et al. "Codebert: A pre-trained model for programming and natural languages." arXiv preprint arXiv:2002.08155 (2020).

[4] Xiong, Chenwei, et al. "HLSPilot: LLM-based High-Level Synthesis." arXiv preprint arXiv:2408.06810 (2024).

[5] What is Retrieval-Augmented Generation (RAG) in LLM ?

[6] Vitis High-Level Synthesis

CONTACT INFORMATION

Aggelos Ferikoglou Ph.D Student, NTUA (aferikoglou@microlab.ntua.gr)

Dimosthenis Masouros Ph.D, NTUA (dmasouros@microlab.ntua.gr)

Assistant Professor Sotirios Xydis, NTUA (sxydis@microlab.ntua.gr)

Professor Dimitrios Soudris, NTUA (dsoudris@microlab.ntua.gr)