COMPUTER AND CONTROL ENGINEERING

MUR DM 117/Huawei - Latency-Optimized Inference for Large Language Models

Funded By	MINISTERO DELL'UNIVERSITA' E DELLA RICERCA [P.iva/CF:97429780584] Politecnico di TORINO [P.iva/CF:00518460019] HUAWEI TECHNOLOGIES ITALIA S.R.L. [P.iva/CF:04501190963]
Supervisor	MACII ENRICO - enrico.macii@polito.it
Contact	MACII ENRICO - enrico.macii@polito.it JAHIER PAGLIARI DANIELE - daniele.jahier@polito.it
Context of the research activity	Large Language Models are the pinnacle of innovation in Artificial Intelligence. The thesis, co-funded by the Huawei Zurich Research Center (ZRC), will focus on improving the efficiency of these models (mainly in terms of latency) when they are executed on the advanced accelerators designed by Huawei (Ascend series). This will be achieved through a combination of optimizations, ranging from kernel selection to sparsity exploitation. Progetto finanziato nell'ambito del PNRR - DM 117/2023 - CUP E14D23002020004
	Large Language Models (LLMs) like GPT-3 and very recently ChatGPT, GPT-4, are Artificial Intelligence models which have been all over the news. This has spurred another large wave of research into training such models and building the infrastructure to train every bigger models. This thesis will focus on what's to come next: the efficient use (inference) of these models. We will explore design automation and optimization techniques for inference of DNN-based language models, particularly focusing on latency optimization. We will address it from multiple angles: 1) Currently, deployment relies on fixed compute kernels implementing mostly individual compute steps of these DNN, some compute-bound and some memory-bound; we will explore automatic ways to fuse multiple compute steps for an optimal trade-off. 2) LLM's auto-regressive structure strongly limits parallelizability and thus creates a high latency; we will explore speculative methods for the auto-regressive dependencies to overcome this limitations.

Objectives	3) Many of the operations, particularly around the attention compute step, lead to many almost-zero values; we will explore how to benefit from this property while keeping the workload efficiently mappable to tensor cores. Huawei, a leading infrastructure provider for deep learning with their Ascend product, will direct the work from their European Research Institute towards the key challenges to be tackled, giving the candidate the opportunity to access the company's large compute infrastructure, and possibly to see output of the research applied in the field on real, large scale problems. The work plan of the project will be structured as follows: Months 1-9: Literature review and background study, focusing on i) the Transformer deep learning model, which constitutes the base architecture of most LLMs; ii) Existing efficient transformer implementations for highly-parallel, accelerator-rich compute platforms; iii) Generic software optimization techniques for deep learning, including sparsity-based ones (pruning, sparse MoE, adaptive inference), precision-reduction ones (quantization), and low-level software ones (kernel selection, layer fusion, memory tiling, etc). Identification of a set of target models to use as benchmark for the following part of the work. Months 9-18: Implementation of “exact” strategies based on low-level kernel selection, fusion, memory access optimizations, etc. These solutions will be implemented first because they will serve as building blocks for the following phases, and will allow the candidate to become familiar with the internal of LLM’s Transformer architectures. Months 18-27: Implementation of speculative optimizations aimed at improving the parallelization of autoregressive decoding in LLMs. This second set of optimizations will be built on top of the work done in the previous part, since the speculative LLMs will leverage the optimized kernels developed in Months 9-18. Months 27-36: Implementation of optimizations aimed at exploiting the natural sparsity of LLMs to further reduce their inference latency. The candidate will also consider the possibility of forcedly increasing sparsity (e.g., by on-the-fly activations pruning) to increase the effectiveness of this strategy. Possible publication venues for this thesis include: - IEEE Transactions on Computers - IEEE Transactions on CAD - IEEE Trans. on Pattern Analysis and Machine Intelligence - ACM Transactions of Design Automation of Electronic Systems - ACM Transactions on Computer Systems - IEEE Journal on Emerging and Selected Topics in Circuits and Systems - Journal of Machine Learning Research - Elsevier AI - Conferences such as NeurIPS, AAAI, ICML, ICLR, MLsys, ACL, DAC, EuroSys, ASPLOS, ISCA, DATE, HiPEAC. The EDA Group has many active industrial collaborations and funded projects on these topics, including: - TRISTAN (ECSEL-JU 2023) - ISOLDE (ECSEL-JU 2023)

	- StorAIge (ESCEL-JU 2021) - etc.

Skills and competencies for the development of the activity	1. Familiarity with programming languages: The project will involve implementing and testing various algorithms and techniques, so experience with programming languages such as Python and C is needed. 2. Familiarity with computer architectures (processing cores, memory hierarchies) and parallel programming models. 3. Understanding of Deep Learning and Artificial Intelligence models, although a specific experience on Transformers and LLMs is not required, as it will be acquired during the first phase of the thesis.