Topic ID: 416-IT-EN

Adaptive Dynamic Orchestration for Tranformer Inference on Neural Processing Units

Supervisor: doc. Ing. Bc. Bronislav Chramcov, Ph.D.

Consultant: prof. Dr. Eng. Said Krayem, CSc.

Department: Department of Informatics and Artificial Intelligence

Programme: Information Technologies

Abstract:

Recent advances in Neural Processing Units (NPUs) now enable Large Language Models (LLMs) to run directly on-device. However, current systems continue to rely on a single, fixed numerical precision across the entire inference graph. Although low-precision formats such as INT8 or INT4 improve speed and energy efficiency, they fail to adapt to the varying computational demands of different tokens or transformer blocks. As a result, when the model encounters more complex reasoning steps, fixed low precision may degrade output quality.

Transformer-based LLMs exhibit varying levels of computational difficulty across tokens, yet existing NPU execution pipelines apply one quantization bit-width to the entire model. This leads to an inefficient compromise: high precision increases energy consumption even in situations where it is unnecessary, whereas low precision risks accuracy loss during more demanding reasoning phases.

The essential research gap lies in the absence of a real-time mechanism capable of adapting numerical precision during inference on NPUs.

This PhD project proposes a Dynamic Precision Switching (DPS) framework tailored to the Qualcomm AI NPU stack. The approach incorporates a lightweight Complexity Monitor that estimates, for each transformer block, whether computation should be executed in INT4, INT8, or FP16. Precision adjustments are applied only to self-attention and feed-forward projection layers to minimize runtime overhead. The research will evaluate whether this targeted, block-level precision adaptation can reduce energy consumption while maintaining accuracy close to full-precision inference.

Literature:

[1] XIAO, Guangxuan, et al. Smoothquant: Accurate and efficient post-training quantization for large language models. In: International conference on machine learning. PMLR, 2023. p. 38087-38099.

[2] CHEN, Hao Mark, et al. Progressive mixed-precision decoding for efficient llm inference. arXiv preprint arXiv:2410.13461, 2024.

[3] MARTÍNEZ, Héctor, Sandra CATALÁN, Adrián CASTELLÓ and Enrique S. QUINTANA-ORTÍ. Characterization of quantized inference with transformer encoders on low power CPUs. The International Journal of High Performance Computing Applications [online]. 2025, 39(6), 803–821. Dostupné z: doi:10.1177/10943420251355115

[4] RAHA, Arnab, Souvik KUNDU, Sharath Nittur SRIDHAR, Shamik KUNDU, Soumendu Kumar GHOSH, Alessandro PALLA, Arghadip DAS, Darren CREWS and Deepak A. MATHAIKUTTY. LLM-NPU: Towards Efficient Foundation Model Inference on Low-Power Neural Processing Units. In: 2025 IEEE International Conference on Omni-layer Intelligent Systems (COINS) [online]. 2025, s. 1–8. Dostupné z: doi:10.1109/COINS65080.2025.11125797

[5] KELLER, Ben, Rangharajan VENKATESAN, Steve DAI, Stephen G. TELL, Brian ZIMMER, Charbel SAKR, William J. DALLY, C. Thomas GRAY and Brucek KHAILANY. A 95.6-TOPS/W Deep Learning Inference Accelerator With Per-Vector Scaled 4-bit Quantization in 5 nm. IEEE Journal of Solid-State Circuits [online]. 2023, 58(4), 1129–1141. Dostupné z: doi:10.1109/JSSC.2023.3234893

[6] LEE, Janghwan, Minsoo KIM, Seungcheol BAEK, Seok Joong HWANG, Wonyong SUNG and Jungwook CHOI. Enhancing Computation Efficiency in Large Language Models throughWeight and Activation Quantization. In: H BOUAMOR, J PINO a K BALI, ed. 2023 CONFERENCE ON EMPIRICAL METHODS IN NATURAL LANGUAGE PROCESSING (EMNLP 2023). 2023, s. 14726–14739. ISBN 979-8-89176-060-8.