Contribution ID: **2** Type: **Poster**

Accelerating AI Inference in the Browser with WebGPU: Evaluating Quantization Trade-offs in Latency, Quality, and Memory Usage

Thursday 24 April 2025 10:45 (30 minutes)

Recent advances in deep learning and natural language processing have spurred the demand for deploying increasingly complex models on resource-constrained platforms. Modern browser environments, empowered by emerging GPU standards like WebGPU, now offer a promising venue for real-time AI inference. This paper provides an overview of leveraging WebGPU for accelerating inference directly within the browser, with a focus on evaluating the trade-offs associated with various quantization schemes. Our study examines the impact of quantization on inference latency, model quality, and memory usage across several model variants. Preliminary benchmarks demonstrate that carefully applied quantization can substantially reduce resource demands while maintaining acceptable performance, laying the groundwork for further optimization of browser-based AI applications. This work sets the stage for future explorations aimed at refining quantization techniques and expanding the capabilities of WebGPU-driven inference.

Authors: Mr RUSZPEL, Ignacy (Politechnika Warszawska - Wydział Elektryczny); Mr WÓJCIK, Nikodem (Politechnika Warszawska - Wydział Elektryczny)

Presenters: Mr RUSZPEL, Ignacy (Politechnika Warszawska - Wydział Elektryczny); Mr WÓJCIK, Nikodem (Politechnika Warszawska - Wydział Elektryczny)

Session Classification: Session B (Poster)