OATNet: A Computational and Mathematical Model of a Novel Neural Network Architecture Utilizing Ternary Weight Decomposition and Element-Wise Methods for Mitigating Computational Complexity

Srikumar, Karthik (School: South Windsor High School)

The electricity demand for large language models (LLMs) has dramatically increased, driven by their intensive computational needs. Optimized Arithmetic via Ternary Networks (OATNet) is proposed as a novel neural network architecture for LLMs that uses fewer resources while maintaining performance comparable to current LLMs. The OATNet architecture was mathematically composed and proved as such: it restructures dense layer computations by decomposing the weight matrix into a ternary-valued matrix and a rank-1 magnitude approximation. Sign information is isolated within the ternary matrix, enforcing sparsity via discrete constraints, while the magnitude is represented as the outer product of two independent vectors, reducing the computation to a low-dimensional subspace. This reconfiguration transforms matrix-vector products into scaled element-wise operations and simplified vector additions. To validate the effectiveness of OATNet, empirical testing was conducted: OATNet was tokenized using WordPiece and optimized in JAX for forward and backward passes. PyTorch was used for testing with a Transformer stack, supporting gradient tracking, modular layer experimentation, and fine-tuning to ensure framework compatibility and robustness. OATNet demonstrated superior efficiency on the WikiText-103 benchmark with validation perplexity of 6.78, a 19.26% improvement over compared LLM models, while requiring only 12.453 Floating Point Operations per second in a token versus the comparisons' mean of 21.302, 41.55% lower. OATNet significantly reduces energy consumption and costs while preserving performance, offering profound environmental benefits, and a move toward sustainable LLMs.

Awards Won: