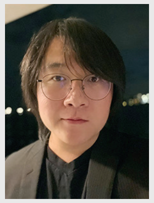
LipTT-LLM: A One-Step Lipschitz-Constrained Binarization Architecture With Tensor Train Decomposition For Large Language Models



While Large Language Models (LLMs) have shown remarkable proficiency and versatility in various tasks such as natural language generation and machine translation, their significant algorithmic complexity and massive number of parameters still hinder their broader deployment and retraining efforts. To overcome these limitations, we propose a novel model compression architecture "LipTT-LLM" combining Network Binarization with Tensor Train Decomposition to reduce model parameter size and computational complexity simultaneously. However, traditional binarization methods often require multi-step training when the model structure is modified, which is unacceptable for large-scale models like LLMs. Additionally, combining Tensor Train (TT) decomposition with binarization introduces challenges related to Batch Normalization (BN). Since TT cores are essentially linear layers, incorporating more BN layers after TT cores would largely increase the computational load, negating model compression's benefits. To address these issues within LipTT-LLM architecture, we developed a novel TT Batch Normalization-free architecture and employed block-wise Lipschitz continuity constraints. This allows us to eliminate the need for additional BN layers and enables binarized models to be trained in one step, notably preventing the rebound of model size and computational complexity. Experimental results demonstrate that LipTT-LLM effectively reduces model size and computational overhead up to 62% and 40% respectively while still maintaining 94% performance comparable to the full-precision models on benchmark tasks.

Mingshuo Liu, *Yifeng Yu, Shangping Ren and Yu Bai*

This research is supported by the SDSU Presidential Graduate Research Fellowship, Statup Funds in the Department of Computer Science at San Diego State University, and the Computational Science Research Center (CSRC) at San Diego State University

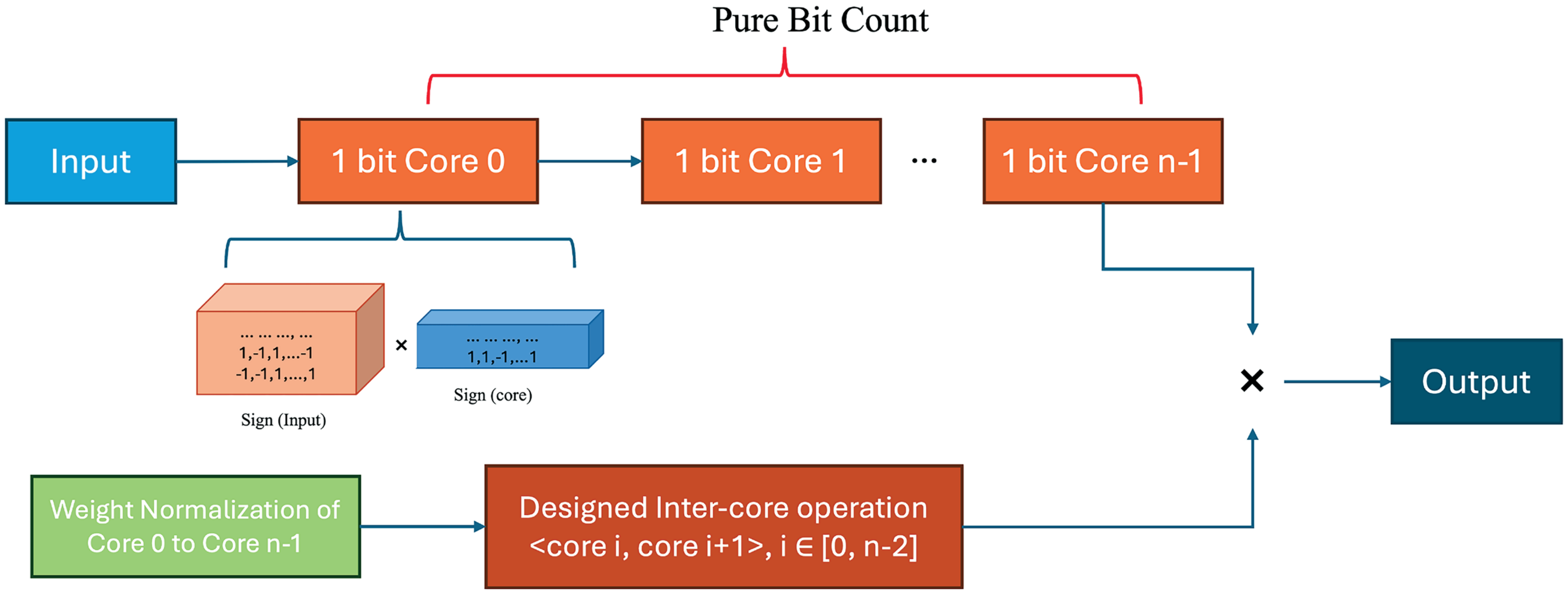


Figure 1. Batch Normalization Free Binarized Tensor Train Format (BN-Free BTTF) Linear Layer Structure

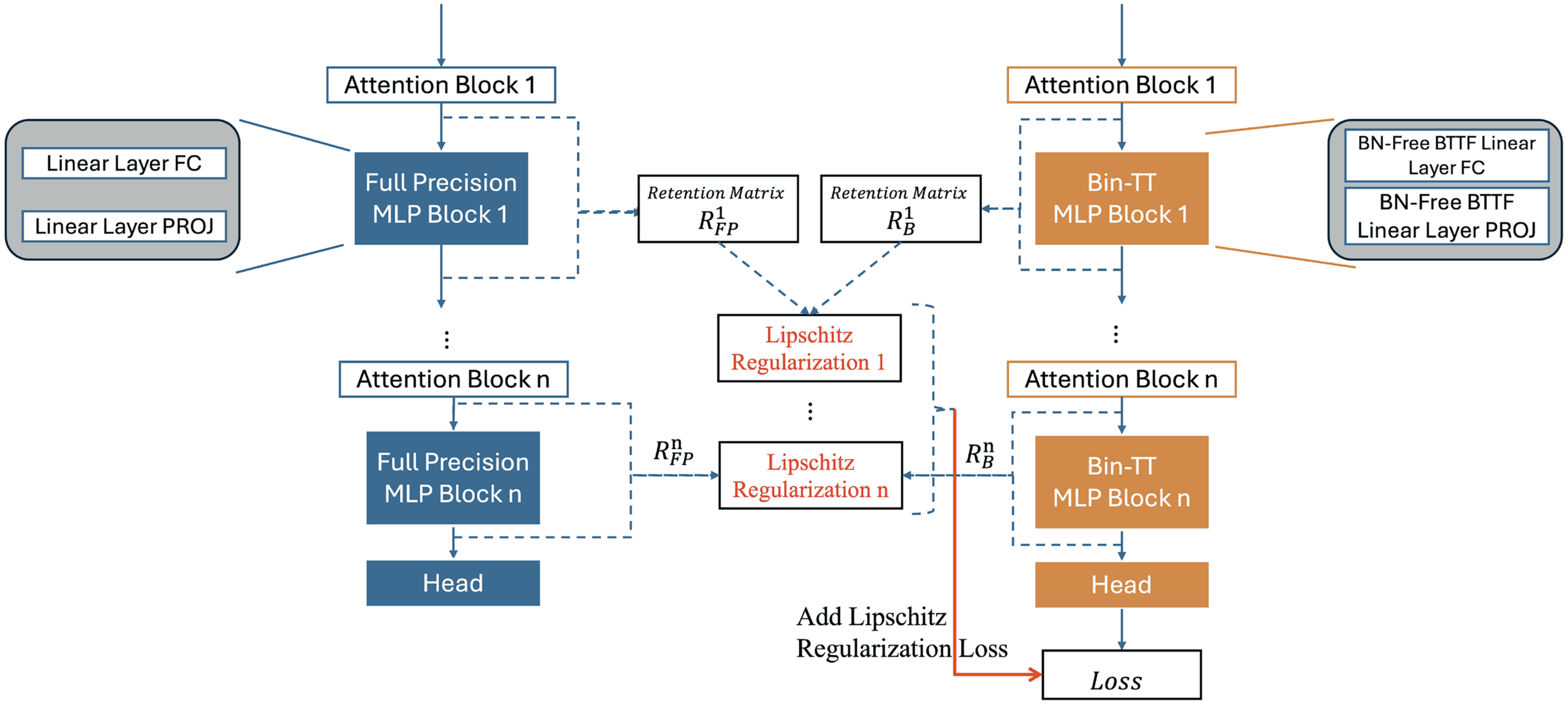

Attention Block 1
Linear Layer FC
Linear Layer PROJ
Full Precision MLP Block 1
Retention Matrix R_{FP}^{1}
Retention Matrix R_{B}^{1}
Bin-TT MLP Block 1
BN-Free BTTF Linear Layer FC
BN-Free BTTF Linear Layer PROJ
Attention Block 1
Lipschitz Regularization 1
Attention Block n
Full Precision MLP Block n
R_{FP}^{n}
Lipschitz Regularization n
R_{B}^{n}
Attention Block n
Bin-TT MLP Block n
Head
Head
Add Lipschitz Regularization Loss
Loss

Figure 2. Lipschitz Continuity Retained Training Process