Confidence Calibration in Quantized Graph Neural Networks Using Temperature Scaling

Hadi Mousanejad Jeddi, Jose Nunez-Yanez
Division of Electronics and Computer Engineering (ELDA), Linköping University
Supervisors:

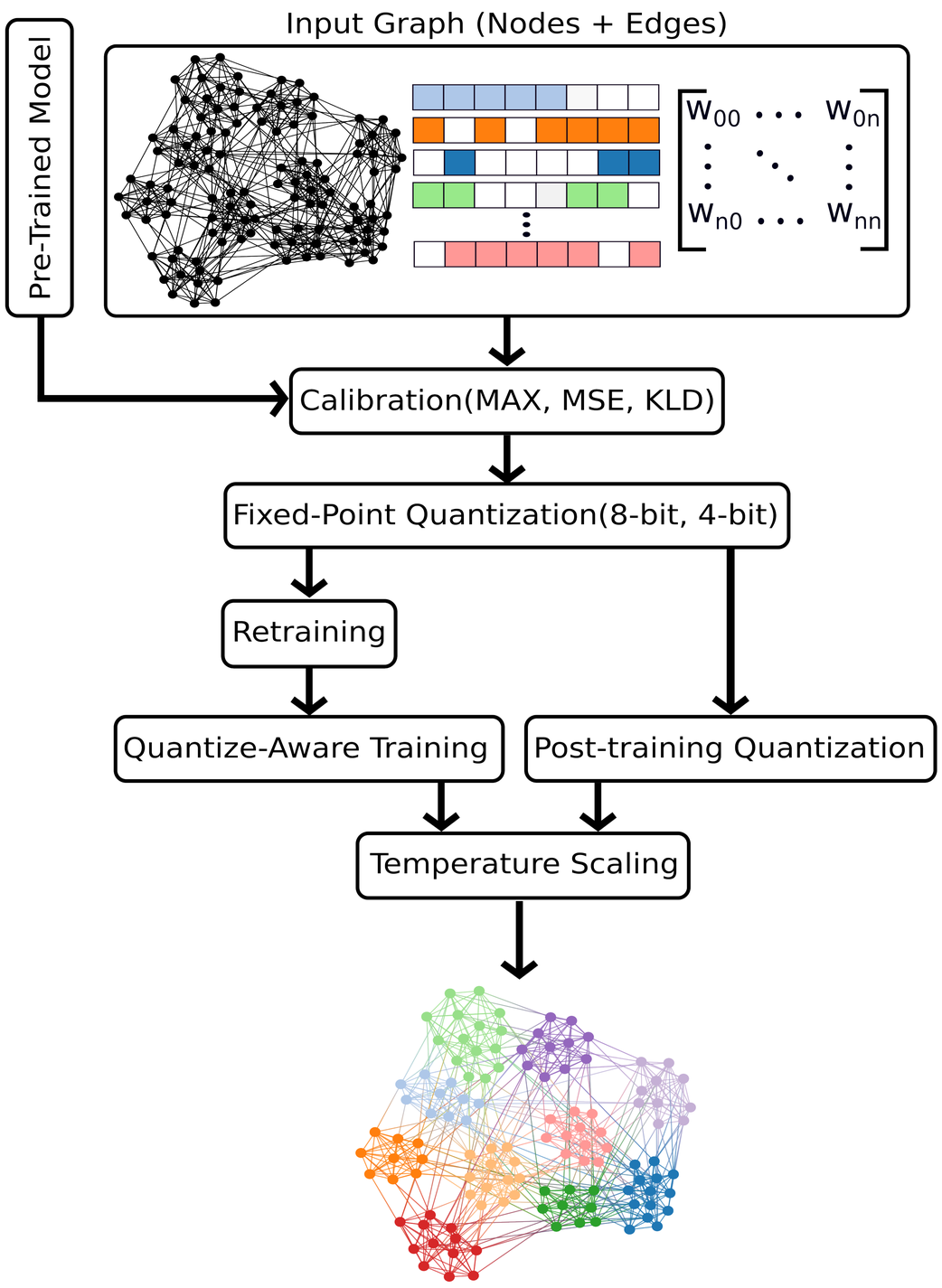
LINKÖPINGS UNIVERSITET

Motivation & Research Goals

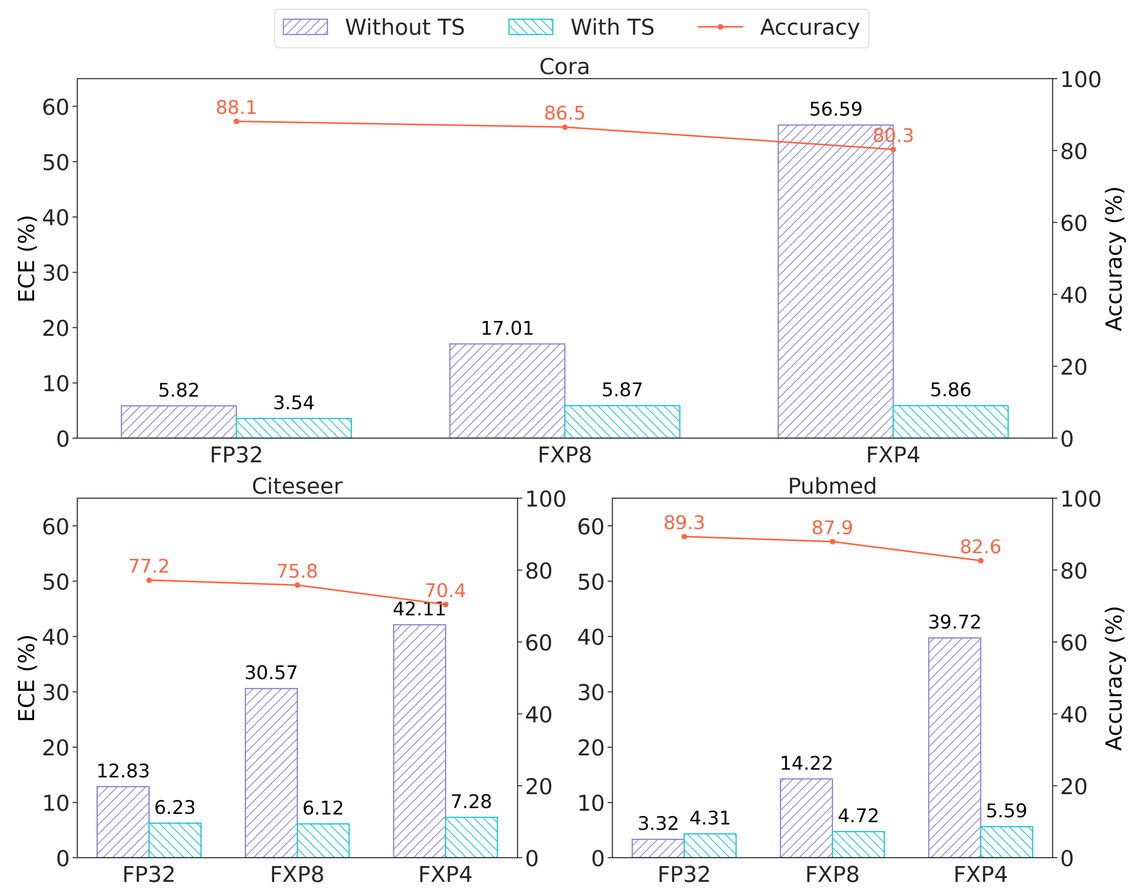
Graph neural networks (GNNs) achieve high performance on various graph-based tasks due to their ability to effectively handle non-uniform structured data. Although GNNs have few parameters, they are computationally intensive because of the large graph sizes that are normally present in the input. Our research focuses on designing efficient quantization techniques and architectures for embedded systems to enhance performance by reducing computational overhead, memory usage, and energy consumption. Neural networks (NNs), such as GNNs, tend to exhibit overconfidence in their predictions, a problem that becomes more pronounced in quantized networks, as highlighted by the higher expected calibration error (ECE) in our analysis. Temperature scaling is a post-processing technique that is used to calibrate confidence with accuracy, thus improving the reliability of predictions. This study explores the application of temperature scaling in 8- and 4-bit fixed-point quantized GCNs to reduce confidence degradation. The goal is to optimize GCN deployment on resource-limited edge devices, ensuring energy consumption efficiency, high accuracy, and reliable predictions for various applications.

Methodology

- In our previous work[1], we developed a post-training quantization (PTQ) flow using dynamic range analysis for GCNs. Expanding on that work, we utilized quantized calibration with three methods: maximum value, mean squared error (MSE), and Kullback-Leibler (KL) divergence to determine the best clipping range for quantization that allows us to convert 32-bit floating-point into 8-bit and 4-bit fixed-point formats.
- To prevent confidence degradation following quantization methods (quantize-aware training (QAT) and PTQ), we apply temperature scaling to adjust GCN logits with parameters optimized via validation to align predictions before calculating final probabilities.



Selected Results



- This figure illustrates the impact of temperature scaling (TS) on the expected calibration error (ECE) across three datasets—Cora, Citeseer, and Pubmed—for GCNs with three numerical formats: FP32 (32-bit floating point), FXP8 (8-bit fixed point), and FXP4 (4-bit fixed point).
- It illustrates that temperature scaling significantly enhances confidence calibration (lowering ECE) across all quantization levels, particularly for FXP8 and FXP4, even without the use of specialized quantization methods, while keeping the accuracy of GCN models consistent across datasets.
- Moving forward, we intend to improve temperature scaling for GCNs by optimizing pre-trained models and utilizing state-of-the-art quantization methods to achieve greater reductions in ECE without losing accuracy.

References

[1] Leveraging Dynamic Range Analysis for Efficient Post-Training Quantization in GCNs
H Mousanejad Jeddi, M Grailoo, J Nunez-Yanez
IEEE Nordic Circuits and Systems Conference (NorCAS), 2024

WASP WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM