

Analysis of Classification Accuracy of FxP 8-bit and FP 8-bit Quantizers in GNNs

Zoran Perić, Nikola Simić, Milan Tančić, Sofija Perić, Oliver Rhodes, Ivan Popović

Abstract – Low-precision arithmetic has emerged as a promising strategy for enabling efficient deployment of deep learning models on resource-constrained platforms. While quantization has been extensively explored in vision and speech models, the effect of low-bit numerical formats on Graph Neural Networks (GNNs)—particularly for affective computing tasks remains insufficiently studied. Emotion recognition from conversational data requires capturing subtle semantic and relational cues, making precision reduction potentially challenging. This work investigates the impact of 8-bit floating-point (FP8) and 8-bit fixed-point (FxP8) quantization on the performance of GNN-based emotion recognition systems.

Experiments are conducted on the IEMOCAP benchmark using the CogMEN architecture, a GNN model designed for conversational emotion understanding. Quantization is applied to model weights and activations during inference, and model performance is assessed using classification accuracy and macro-F1 score, both critical metrics for emotion recognition.

The results show that FP8 quantization consistently outperforms FxP8, demonstrating superior accuracy retention due to its wider dynamic range and improved numerical stability. Although FxP8 exhibits a modest performance degradation, it remains a viable solution for highly resource constrained deployments owing to its simpler arithmetic and lower hardware complexity. These findings highlight the trade-offs between numerical robustness and implementation efficiency when deploying GNN-based emotion recognition systems on edge and embedded platforms.

Keywords - Emotion recognition, Graph Neural Networks, FP8, FxP8, quantization, CogMEN, edge AI.

I. INTRODUCTION

Recent advances in deep learning have enabled significant progress in affective computing, particularly in emotion recognition from conversational and multimodal data. Among the various model families, Graph Neural Networks (GNNs) have emerged as a powerful paradigm for modeling relational and contextual dependencies, allowing interactions between speakers, utterances, and temporal structures to be explicitly represented.

Zoran Perić, Sofija Perić, Ivan Popović, Elektronski fakultet Niš, Srbija, E-mail: zoran.peric@elfak.ni.ac.rs, sofija.peric@elfak.ni.ac.rs, ivan.popovic@netico-group.com

Nikola Simić, Fakultet tehničkih nauka, Novi Sad, Srbija, E-mail: nikolasimic@uns.ac.rs,

Milan Tančić, Akademija tehničko vaspitačkih strukovnih studija Niš, E-mail: milan.tancic@akademijanis.edu.rs

Oliver Rhodes, University of Manchester, United Kingdom, E-mail: oliver.rhodes@manchester.ac.uk

Such capabilities are crucial for conversational emotion recognition, where emotional states are influenced not only by individual utterances but also by their semantic, temporal, and interpersonal relationships [1-5].

Despite their strong representational capacity, GNN-based models typically rely on high-precision floating-point arithmetic (e.g., FP32), which incurs substantial computational, memory, and energy costs. These requirements limit the deployment of emotion recognition systems on resource-constrained platforms such as mobile devices, edge servers, and embedded systems. Consequently, reducing numerical precision through quantization has become an important research direction for enabling efficient inference while maintaining acceptable predictive performance [5, 6].

Quantization techniques aim to replace high-precision arithmetic with lower-bit numerical formats, reducing memory footprint and computational complexity. Prior work has demonstrated that low-precision representations, including fixed-point and floating-point formats with reduced bit width, can achieve competitive performance in computer vision and speech recognition tasks. However, the behavior of quantized arithmetic in GNNs remains less understood. In contrast to convolutional or recurrent architectures, GNNs involve iterative message passing, aggregation operations, and non-linear transformations over graph structures, which can amplify numerical errors and make them particularly sensitive to reduced precision [6-11].

This sensitivity is especially relevant in affective computing tasks such as emotion recognition, where subtle variations in semantic and relational cues can significantly impact classification outcomes. Emotion recognition from conversational data requires the model to capture nuanced emotional dynamics across speakers and dialogue turns. As a result, aggressive precision reduction may degrade the model's ability to preserve fine-grained emotional information, potentially leading to performance loss. Understanding how different low-precision formats affect GNN-based emotion recognition is therefore essential for reliable real-world deployment [11-12].

Among low-precision formats, 8-bit fixed-point (FxP8) and 8-bit floating-point (FP8) representations represent two fundamentally different approaches to numerical compression. FxP8 offers a simple and hardware-friendly arithmetic model with minimal implementation overhead, making it attractive for highly constrained platforms. However, its limited dynamic range may lead to saturation

and quantization errors in graph computations. FP8, on the other hand, provides a wider dynamic range through exponent-based representation, improving numerical stability at the cost of slightly more complex hardware support [7]. While FP8 has recently gained attention in high-performance accelerators, its advantages over FxP8 in GNN-based affective computing remain underexplored [6-9, 13,14].

This work presents a systematic empirical analysis of 8-bit floating-point (FP8) and 8-bit fixed-point (FxP8) quantization applied to GNN based emotion recognition, aiming to clarify the trade-offs between numerical precision, accuracy retention, and hardware efficiency in low-bit inference. Using the CogMEN architecture, a state-of-the-art GNN model for conversational emotion understanding, quantization is applied to both model weights and activations during inference and evaluated on the widely used IEMOCAP benchmark. Model performance is assessed using classification accuracy and macro-F1 score, which are critical metrics for affective computing tasks. The experimental results demonstrate that FP8 quantization achieves superior preservation of predictive performance due to its wider dynamic range and improved numerical stability, while FxP8 exhibits a modest degradation in accuracy and F1 score but remains a viable solution for highly resource-constrained deployments because of its simpler arithmetic, lower implementation complexity, and reduced computational and memory requirements. Collectively, these findings provide practical insight into the suitability of low-precision numerical formats for efficient and reliable GNN-based emotion recognition in real-time affective computing and edge-AI applications [10-12].

The remainder of the paper is organized as follows: Section 2 presents theoretical background on emotional recognition architectures and fixed and floating point quantization; Section 3 outlines the experimental methodology with results while Section 4 concludes the paper and identifies directions for future research.

II. THEORETICAL BACKGROUND

Emotion recognition aims to identify the emotional state expressed in human communication and plays a central role in affective computing, human-computer interaction, and intelligent dialogue systems. In conversational settings, emotional expression is inherently contextual and relational, as the emotion conveyed by an utterance is influenced not only by its semantic content but also by preceding dialogue turns, speaker interactions, and temporal dependencies. Consequently, effective emotion recognition systems must model both linguistic information and the structured relationships among conversational elements. Early approaches to conversational emotion recognition relied on recurrent neural networks and attention-based models to capture temporal dependencies across dialogue sequences [15]. While these architectures

successfully model sequential information, they are limited in their ability to explicitly represent complex relational structures such as speaker interactions and long-range contextual dependencies. To address these limitations, Graph Neural Networks (GNNs) have been introduced as a natural framework for conversational emotion modeling, where utterances and speakers can be represented as nodes, and their interactions as edges in a graph structure [3, 6-9].

Graph Neural Networks extend deep learning to graph structured data through message passing between connected nodes. In conversational emotion recognition, GNNs aggregate contextual and relational information across dialogue turns and speaker interactions, yielding richer representations than purely sequential models. Node representations, typically derived from pretrained language models, are iteratively updated via neighborhood aggregation and nonlinear transformations, a process that enhances contextual reasoning but also increases sensitivity to numerical precision due to repeated aggregation and normalization operations. As a result, GNN-based architectures can be more vulnerable to quantization errors than feedforward or convolutional networks, particularly when deployed using low-precision arithmetic [13]. In this work, the CogMEN architecture is employed as a representative GNN model for conversational emotion recognition. CogMEN integrates contextual semantic encoding with graph-based relational reasoning, making it well suited for evaluating the impact of low precision numerical formats in affective computing tasks [11-14].

Quantization is a model compression technique that reduces numerical precision by mapping high-precision floating-point values to lower-bit representations. The primary motivation for quantization is to reduce memory footprint, computational complexity, and energy consumption, thereby enabling efficient inference on edge and embedded devices. Quantization can be applied to model parameters, activations, or both, and is typically performed either during training (quantization-aware training) or during inference (post-training quantization).

Floating-point quantization reduces precision while retaining a sign/exponent/mantissa representation, allowing for a wide dynamic range compared to fixed-point formats. The sign bit provides information if a number is positive or negative, exponent determines the scale while mantissa represents the precision of significant digits. Sign s , exponent E_b and mantissa m are determined using following equations [6,14]:

$$s = 0, \text{ if } x \geq 0; 1, \text{ if } x < 0 \quad (1)$$

$$E_b = \log_2(|x|) \quad (2)$$

$$M = \text{round}\left(2^m \left(|x|/2^{E_b} - 1\right)\right) \quad (3)$$

where x presents input sample. In 8-bit floating-point (FP8) quantization, bits are distributed between the exponent and mantissa, enabling better handling of values with varying

magnitudes. This characteristic improves numerical stability in models with diverse activation distributions, such as GNNs. Although FP8 arithmetic is more complex than fixed point arithmetic, recent advances in hardware accelerators have made FP8 increasingly practical for efficient inference. The improved dynamic range of FP8 helps mitigate saturation and accumulation errors during message passing, making it a promising candidate for low precision GNN deployment in affective computing applications.

Fixed-point quantization represents numerical values using a fixed number of bits allocated to the sign, integer and fractional parts. The sign bit s is defined as (1), while integer I and fractional F parts are calculated using [7-10]:

$$I = \lfloor x \rfloor \quad (4)$$

$$F = \text{round}((|x| - I) \cdot 2^f) \quad (5)$$

where x presents input sample and f is number of bits assigned to fractional part quantization. In an 8-bit fixed-point format, the dynamic range is limited and determined by the scaling factor, which must be carefully chosen to avoid overflow or excessive quantization error. Fixed-point arithmetic is computationally efficient and well supported by simple hardware implementations, making it attractive for resource-constrained platforms. However, the limited dynamic range of FxP8 can be problematic for GNN based emotion recognition, where node embeddings and aggregated messages may vary significantly in magnitude. Rounding errors introduced by fixed-point representation can degrade model performance, particularly in deeper networks or architectures with repeated message-passing operations [7-10].

Owing to the differing numerical characteristics of fixed and floating point quantization, a systematic comparison is required to assess their impact on GNN based emotion recognition. While FxP8 emphasizes hardware simplicity and low implementation cost, FP8 offers greater numerical robustness that may be essential for preserving emotional cues in conversational graphs. This work evaluates these trade-offs by comparing FP8 and FxP8 quantization in a GNN model, providing guidance for efficient and reliable deployment under varying resource constraints.

III. EXPERIMENTAL RESULTS AND DISCUSSION

This section describes the experimental setup used to evaluate the impact of low-precision numerical formats on GNN-based emotion recognition and presents a comparative analysis of FP8 and FxP8 quantization. The experiments are conducted using the CogMEN architecture designed for conversational emotion recognition, evaluated on the IEMOCAP benchmark dataset. The IEMOCAP dataset is a widely used benchmark for emotion recognition from conversational speech, containing dyadic dialogues

annotated with discrete emotion labels. Following standard evaluation protocols, the dataset is preprocessed and partitioned to ensure speaker-independent evaluation. Utterance-level embeddings are provided as input to the CogMEN model, which constructs a conversational graph to capture contextual and relational dependencies among dialogue turns [5-9, 14].

Two 8-bit numerical formats are considered: fixed-point quantization, using one sign bit, two integer bits, and five fractional bits, and floating-point quantization, using one sign bit, two mantissa bits, and five exponent bits. The quantized models are evaluated against a full precision baseline to assess accuracy degradation introduced by reduced numerical precision. Model performance is evaluated using classification accuracy and macro-F1 score, which are selected due to their widespread use in affective computing [5-9, 14]. For FxP8 quantization, real valued weights and activations are mapped to a fixed point representation using a uniform scaling factor determined by the number of fractional bits. Saturation is applied to prevent overflow beyond the representable dynamic range. For FP8 quantization, values are encoded using a sign/exponent/mantissa representation, enabling dynamic scaling and improved numerical stability across varying activation magnitudes. Both quantization schemes are applied consistently across all GNN layers to ensure a fair comparison. No additional retraining or fine-tuning is performed after quantization, allowing the observed performance differences to be attributed solely to the numerical format [14].

Table 1 summarizes the performance of FP8-quantized FP32 quantized (full precision) and FxP8-quantized CogMEN models on the IEMOCAP dataset.

TABLE I
PERFORMANCE COMPARISON BETWEEN FP8 AND FxP8

Numerical format	Classification accuracy (%)	Macro F1 (%)
FP8	82.82	82.63
FP32	82.82	82.63
FxP8	81.76	81.55

The results given in Table I demonstrate that FxP8 quantization introduces a modest but consistent performance degradation comparing to FP8 and FP32 which provides the same classification accuracy and F1. FP8 and FP32 yield identical accuracy and macro-F1 scores because inference time quantization does not alter the model's decision boundaries when the numerical representation preserves sufficient dynamic range and relative precision. The reduced dynamic range of the fixed-point format leads to increased saturation and rounding errors, particularly in deeper graph aggregation layers. Nevertheless, the observed performance remains competitive, confirming that FxP8 is a viable solution for scenarios where computational simplicity and hardware

efficiency are prioritized. Overall, the experimental results highlight a clear trade-off between numerical robustness and implementation efficiency. FP8 provides superior accuracy retention and stability, making it suitable for performance-critical applications, while FxP8 offers a simpler and more resource-efficient alternative for deployment on highly constrained platforms.

IV. CONCLUSION

This paper investigated the impact of low-precision numerical formats on GNNs based emotion recognition, with a particular focus on comparing 8-bit floating-point (FP8) and 8-bit fixed-point (FxP8) quantization. Using the CogMEN architecture evaluated on the IEMOCAP conversational emotion recognition benchmark, the study analysed how different quantization schemes affect classification accuracy and macro-F1 score when applied to model weights and activations during inference.

The results indicate that FP8 consistently outperforms FxP8 in terms of accuracy retention, which can be attributed to its exponent-based representation and wider dynamic range. These properties improve numerical stability during graph message-passing and aggregation operations, where activation magnitudes may vary significantly. In contrast, FxP8 exhibits a modest but systematic reduction in performance, primarily due to its fixed scaling and increased susceptibility to saturation and rounding errors.

Despite the observed performance gap, FxP8 remains appealing for deployments prioritizing hardware simplicity, low arithmetic complexity, and minimal resource usage. In contrast, FP8 offers greater numerical robustness and is more suitable when preserving recognition reliability is critical. These results emphasize the importance of selecting quantization formats based on the trade-off between implementation efficiency and numerical stability for deploying GNN based emotion recognition systems on resource constrained and edge-AI platforms.

REFERENCES

- [1] Burkitt, Ian. "SOCIAL RELATIONSHIPS AND EMOTIONS." *Sociology* 31, no. 1 (1997): 37–55. <http://www.jstor.org/stable/42855768>.
- [2] N. H. Frijda, "The Emotions", Cambridge University Press, 1986.
- [3] C. Yang, X. Zhang, Y. Chen, Y. Li, S. Yu, B. Zhao, T. Wang, L. Luo, S. Gao "Emotion dependent language featuring depression", *J. Behav. Therapy Exp. Psych.*, 81,101883, 2023. DOI: 10.1016/j.jbtep.2023.101883
- [4] S. Mahlke, M. Minge, "Emotions and EMG measures of facial muscles in interactive contexts", *Cong. Emot.* 6, 169-200, 2006
- [5] N. Antoniou, A. Katsamanis, T. Giannakopoulos and S. Narayanan, "Designing and Evaluating Speech Emotion Recognition Systems: A Reality Check Case Study with IEMOCAP," *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Rhodes Island, Greece, 2023, pp. 1-5, doi: 10.1109/ICASSP49357.2023.10096808.
- [6] Z. Perić, B. Denić, M. Dinčić, S. Perić, "An Approximate Closed-Form Expression for Calculating Performance of Floating-Point Format for the Laplacian Source", *Informatica*, Vol. 36, No. 1, pp. 125-140, 2025. <https://doi.org/10.15388/25-INFOR587>
- [7] D. Lin, S. Talathi, S. Annapureddy, "Fixed point quantization of deep neural networks", *Proceedings of 33rd International Conference on Machine Learning*, DOI: <https://doi.org/10.48550/arXiv.1511.06393>
- [8] R. Goyal, J. Vanschoren, V. van Acht, S. Nijssen, "Fixed-point Quantization of Convolutional Neural Networks for Quantized Inference on Embedded Platforms". *Mathematics and Computer Science Data Mining*, al. *arXiv.org*. <https://arxiv.org/abs/2102.02147>
- [9] N. Mitschke, M. Geizmann, K. Noffz, R. Wittmann, "A Fixed-Point Quantization Technique for Convolutional Neural Networks Based on Weight Scaling", 2019 *IEEE International Conference on Image Processing (ICIP)*, DOI:10.1109/ICIP.2019.8803490
- [10] B. Denić, Z. Perić, M. Dinčić, S. Perić, N. Simić, M. Anđelković, "Switched 32-Bit Fixed-Point Format for Laplacian-Distributed Data", *Information* 16, no.7: 574, July 2025. <https://doi.org/10.3390/info16070574>
- [11] N. Simic, S. Suzic, T. Nosek, M. Vujovic, Z. Peric, M. Savic, V. Delic, "Speaker recognition using constrained convolutional neural networks in emotional speech", *Entropy*, 24(3), 2022, DOI:10.3390/e24030414
- [12] N. Simic, S. Suzic, N. Milosevic, V. Stanojevic, T. Nosek, B. Popovic, D. Bajovic "Enhancing emotion recognition through federated learning: a multimodal approach with convolutional neural networks", *Applied Sciences*.14(4), 1325, 2024, DOI:10.3390/app14041325
- [13] A. Joshi, A. Bhat, A. Jain, A. V. Modi, "COGMEN: Contextualized GNN based multimodal emotion recognition", In: *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Seattle, USA, 4148-4164, 2022
- [14] T. Đurkić, N. Simić, S. Suzić, D. Bajović, Z. Perić, V. Delić, "Multimodal Emotion Recognition Using Compressed Graph Neural Networks", *Speech and Computer. SPECOM 2024. Lecture Notes in Computer Science*, vol 15300. Springer, Cham. https://doi.org/10.1007/978-3-031-78014-1_9
- [15] N. Majumder, S. Poria, D. Hazarika, R. Mihalcea, A. Gelbukh, E. Cambria, "DialogueRNN: An Attentive RNN for Emotion Detection in Conversations", July 2019 *Proceedings of the AAAI Conference on Artificial Intelligence* 33(01):6818-6825, DOI: 10.1609/aaai.v33i01.33016818