Layer-Importance guided Adaptive Quantization for Efficient Speech Emotion Recognition

Tushar Shinde, Ritika Jain, Avinash Kumar Sharma
School of Engineering and Science, Indian Institute of Technology Madras Zanzibar, Tanzania

INTRODUCTION

- Speech Emotion Recognition (SER) systems play a critical role in human-computer interaction.
- The subjective nature of emotions and complex speech patterns make accurate emotion recognition difficult for machines.
- **Challenge:** Current methods involve complex models that demand high computational resources, limiting real-time and device-based deployment.
- **Impact:** Enables resource-efficient SER systems for edge devices and IoT.

PROBLEM ANALYSIS & MOTIVATION

- **Motivation:** Fixed precision approaches often overlook the varying importance of different layers in DNN, leading to performance degradation.
- Mixed-precision quantization can address the limitations of fixed precision approaches, offering a better trade-off between model size and accuracy.
- **Solution:** Adaptive quantization ensures efficiency without sacrificing performance. A novel lightweight MLP model with adaptive quantization enhances SER performance while reducing resource requirements.

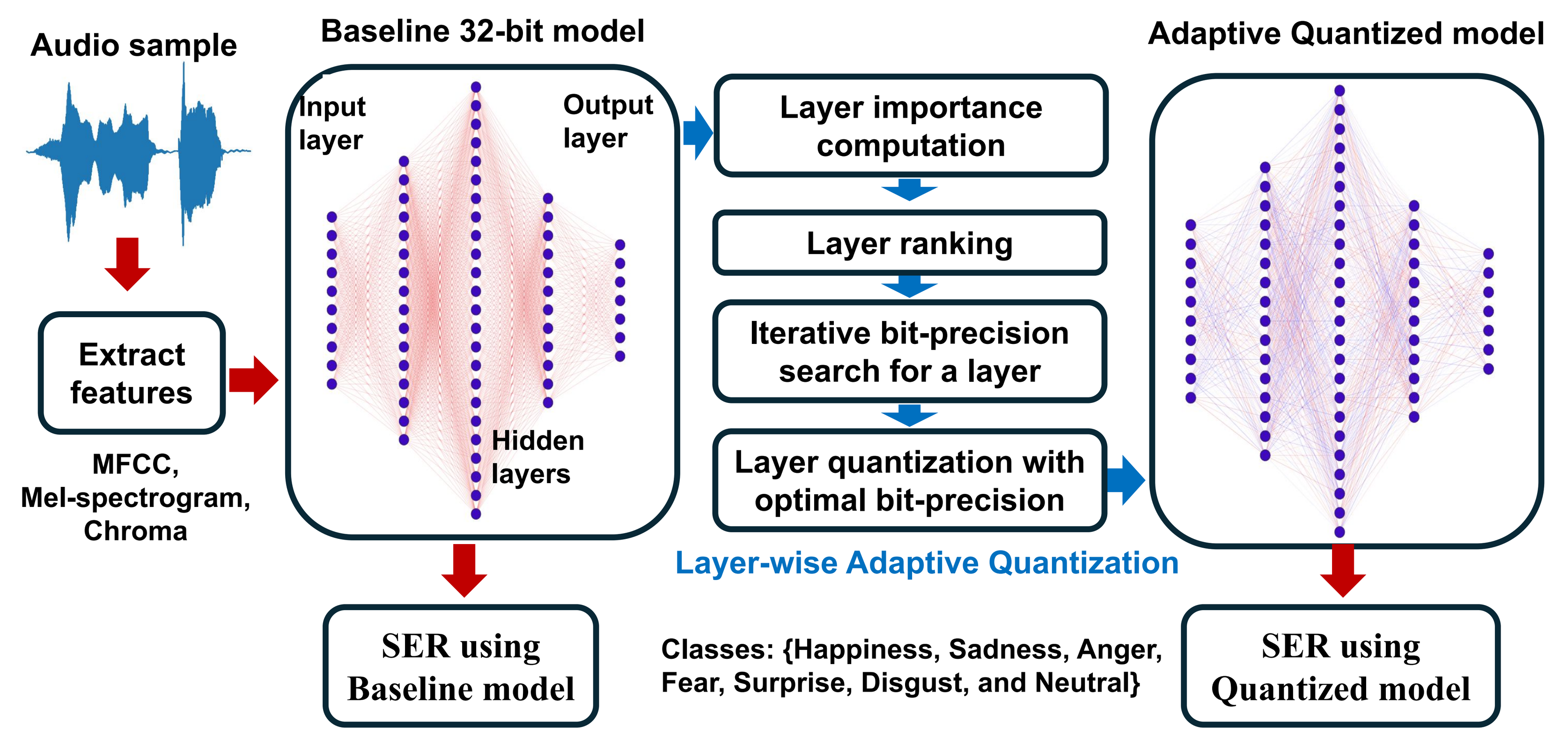
OBJECTIVES

- Develop a robust SER framework leveraging adaptive layer-wise quantization to reduce model size.
- Optimize layer bit-widths to balance performance and compression.

DATASETS

Dataset	# Samples	# Speakers	Gender (M/F)	# Emotions
EMODB	535	10	5/5	7
SAVEE	480	4	4/0	7
TESS	2800	2	0/2	7

PROPOSED METHOD



Layer Importance computation:

$$\text{Importance}(l) = \alpha \cdot N_P(l) + (1-\alpha) \cdot N_V(l)$$

$$N_P(l) = \frac{\text{Parameters in layer } l}{\text{Total parameters in the model}}$$

$$N_V(l) = \log\left(e - 1 + \frac{\text{Variance of layer } l}{\max_k(\text{Variance of layer } k)}\right)$$

- **SER classification task** identifies emotions from speech data as: Happiness, Sadness, Anger, Fear, Surprise, Disgust, and Neutral.
- **Feature Extraction:** Features like MFCC, Chroma, Mel-spectrogram ensure optimal input representation for accurate emotion recognition.
- **Baseline MLP Classifier:** A compact Multilayer Perceptron (MLP) model designed for efficient Speech Emotion Recognition.
- An innovative approach to accurately calculate **layer importance**, crucial for adaptive quantization and optimizing model performance.
- **Layer Ranking:** Layers are prioritized based on importance metrics, ensuring critical layers are quantized first to optimize performance and resource efficiency.
- Layer-wise **Adaptive Quantization** to reduce computational complexity while maintaining high accuracy.
- **Bit-Width Allocation**: Assigns optimal bit-width precision to each layer based on importance to minimize size while preserving accuracy. Supports mixed-precision quantization for further optimization.
- **Performance Threshold:** Ensures that performance degradation remains well within an acceptable margin, ensuring model reliability during quantization.

$$T_{\text{margin}}(l) = T_{\text{margin}} \times \text{Importance}(l)$$

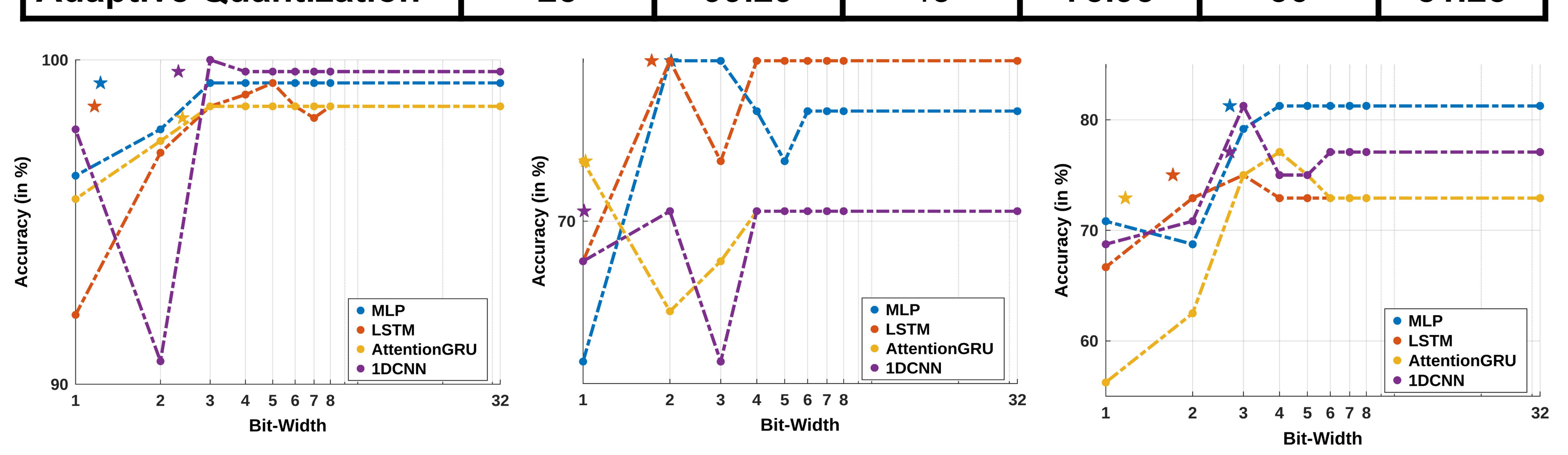
RESULTS

- **Lightweight MLP Design:** A compact MLP model with 3 hidden layers (256, 512, 64 neurons), totaling 169K parameters for the SER task.
- **Implementation details:** Adam optimizer with a learning rate of 0.001, Cross-Entropy loss, a batch size of 32, and early stopping to prevent overfitting. Dropout rate of 0.1 applied to enhance generalization.
- **Evaluation Metrics:** Accuracy & Average Bit-width

$$\bar{b} = \sum_{l=1}^{L} N_P(l) \cdot b(l)$$

- **Quantization Evaluation:** The performance of quantized models is compared to their full-precision counterparts (accuracy and model size)
- **Fixed-bit quantization** achieves model size reductions, with minor accuracy drops as bit-width decreases.
- **Our adaptive quantization method** achieves near-baseline accuracy while significantly reducing model size and average bit-width, with substantial improvements over fixed-bit quantization.
- **Model Comparison:** Several models like MLP, LSTM, AttentionGRU, and 1DCNN are evaluated for SER tasks.
- **Average Bit-width Reduction:** 1.22 bits (TESS), 2 bits (EMODB), and 2.69 bits (SAVEE) using our adaptive quantization instead of 32 bits.

Datasets	TESS		EMODB		SAVEE	
Model	Size (KB)	Acc. (%)	Size	Acc. (%)	Size	Acc. (%)
Baseline (32-bit)	676	99.29	676	74.07	676	81.25
Fixed Q (8-bit)	169	99.29	169	74.07	169	81.25
Fixed Q (7-bit)	147	99.29	147	74.07	147	81.25
Fixed Q (6-bit)	126	99.29	126	74.07	126	81.25
Fixed Q (5-bit)	105	99.29	105	72.22	105	81.25
Fixed Q (4-bit)	84	99.29	84	74.07	84	81.25
Fixed Q (3-bit)	63	99.29	63	75.93	63	79.17
Fixed Q (2-bit)	42	97.86	42	75.93	42	68.75
Fixed Q (1-bit)	21	96.43	21	64.81	21	70.83
Adaptive Quantization	**25**	**99.29**	43	**75.93**	56	**81.25**



Model accuracy vs. bit-width for a) TESS, b) EMODB, and c) SAVEE dataset. Our method is highlighted with a star marker.

CONCLUSION

- **Efficiency:** The model achieves competitive or superior performance with significantly fewer parameters (169K) and an average bit-width of about 2 bits.
- **Model Size:** The maximum model size is reduced to just 56 KB, making the model highly efficient for deployment on resource-constrained devices.
- **Architecture Advantage:** The simple architecture with minimal parameters ensures fast inference and reduced resource consumption.
- **Limitations:** The study does not include cross-dataset experiments to assess the model's generalizability and robustness across different datasets.
- **Future Work:** Investigate other advanced model compression techniques for further model optimization on diverse datasets.

REFERENCES

[1] Aftab, A., et.al., Light-sernet: A lightweight fully convolutional neural network for speech emotion recognition. In 2022 IEEE international conference on acoustics, speech and signal processing (ICASSP) (pp. 6912-6916). IEEE.
[2] Pichora-Fuller, M.K. and Dupuis, K., 2020. Toronto emotional speech set (TESS); 2020. URL: https://tspace. library. utoronto. ca/handle/1807/24487. DOI: https://doi. org/10.5683/SP2/E8H2MF.
[3] Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W.F. and Weiss, B., 2005, September. A database of German emotional speech. In Interspeech (Vol. 5, pp. 1517-1520).
[4] Jackson, P. and Haq, S., 2014. Surrey audio-visual expressed emotion (savee) database. University of Surrey: Guildford, UK.

ACKNOWLEDGMENTS

This work was in part supported by the Walmart Center of Technical Excellence (IIT Madras)} Project Grant Award.

Website **Paper**