A Review of Binarized Weight Neural-network Inspired Ultra-low Power Speech Recognition Processor with Time-domain based Digital-analog Mixed Approximate Computing

1) Aryan Langhanoja
2) Jay Mangukiya
Department of Information and Communication Technology
Marwadi University
Rajkot , India

Abstract—

The research paper introduces an ultra-low power speech recognition processor based on an optimized binarized weight neural-network (BWN). The paper highlights the challenges of power consumption in deep neural networks (DNNs) and presents various DNN accelerators proposed in recent years. The authors aim to overcome these challenges by proposing an approximate computing architecture for the quantized BWN based on time-domain digital-analog mixed addition unit and precision optimization with fault-tolerant training method.

Keywords— *Speech Recognition, Binarized Weight Neural-network, Approximate Computing, Fault-tolerant Training*

Introduction—

The abstract provides a concise overview of the paper's content. It mentions the implementation of an ultra-low power speech recognition processor based on an optimized BWN. The proposed architecture utilizes approximate computing and fault-tolerant training to reduce power consumption while maintaining recognition accuracy. The experimental results demonstrate the effectiveness of the proposed architecture in supporting real-time recognition of 10 keywords under different noise conditions, with a power consumption of 56μW.

Body—

The paper presents the top architecture of the prototype speech recognition system, which consists of a feature extraction module and a speech classification module. The feature extraction module utilizes Mel-scale Frequency Cepstral Coefficients (MFCC) to extract speech features. The speech classification module processes feature classification based on a BWN, which is trained using an optimized quantization method.

The proposed BWN architecture consists of six convolutional (CONV) layers, three fully-connected (FC) layers, activation (ACT) layers, and batch normalization (BN) layers. The weights of the BWN are binarized to 1 bit, while the feature data is quantized to 16 bits. The paper introduces a time-domain digital-analog mixed approximate computing architecture, which utilizes time-domain approximate addition units (TAAUs) to process additions in the BWN. The TAAUs are designed to reduce power consumption and improve energy efficiency.

To address the issue of circuit mismatch, the paper proposes a fault-tolerant training (FTT) method. By adding mismatch errors to the input of each layer in the BWN, the authors optimize the mismatch error model and improve the robustness of the neural network. The training process includes the addition of a regularization term to the cost function, which reduces the sensitivity to circuit mismatch.

The implementation results of the proposed speech recognition processor show that it can support real-time recognition of 10 keywords under different noise conditions, with a power consumption of 56μW. The paper compares the proposed architecture with state-of-

the-art speech recognition processors and highlights its superior energy efficiency.

Conclusion—

The research paper presents an ultra-low power speech recognition processor based on an optimized BWN. The proposed architecture utilizes approximate computing and fault-tolerant training to reduce power consumption while maintaining recognition accuracy. The implementation results demonstrate the effectiveness of the proposed architecture in supporting real-time recognition of 10 keywords under different noise conditions, with significantly reduced power consumption. The paper contributes to the field of speech recognition by addressing the power consumption challenges in DNNs and providing a solution that improves energy efficiency.