Oral S12

Fast Algorithm and Accelerator Design for AI Applications

Date/Venue	8/2(五) 10:30-11:30 [海棠廳]
Chairs(s)	吳崇賓 /國立中興大學電機系

S12.1 | 10:30-10:41

FQ4DM: FULL QUANTIZATION FOR DIFFUSION MODEL

Chieh-En Wang, Yu-Shan Tai, Chih-Sheng Cheng, and An-Yeu (Andy) Wu
Graduate Institute of Electrical Engineering, National Taiwan University, Taipei, Taiwan

Diffusion models (DMs) have recently gained acclaim for their superior imaging capabilities. However, their extensive computational and memory demands often limit the practical application on portable devices. Post-training quantization (PTQ) offers a solution that enables model compression and reduces runtime without retraining. Nonetheless, traditional PTQ methods struggle to handle the unique time-variant distribution in DMs. Accordingly, we propose a novel timestep-grouping PTQ approach to address the multiple timestep issue. We also identify that non-uniform post-SiLU activations may lead to significant quantization loss. We tackle this issue with a region-specific quantization strategy that better represents extreme values after quantization. Combined with the above methods, we achieve a fully quantized diffusion model feasible for hardware implementation. Our experimental results show that the proposed method successfully maintains the FID score after 8-bit quantization

S12.2 | 10:42-10:53c

Automated Classification of Multi-class Brain Tumor with Patch Stem and Lightweight CNN Architecture

Yan-Wen Chou, Cheng-Hung Lin, Min-Chun Hou, Wei-Chen Kuo
[1]Department of Electrical Engineering, Yuan Ze University, Taoyuan City, Taiwan 320, R.O.C.
[2]Biomedical Engineering Research Center, Yuan Ze University, Taoyuan City, Taiwan 320, R.O.C.

Classification of brain tumors is crucial for accurate diagnosis because different types of tumors have different treatment requirements. Through precise classification, doctors can develop more effective treatment plans, improving patient survival rates. However, current deep-learning models utilized for brain tumor classification typically exhibit large parameter sizes and high computational complexity. Therefore, we proposed a lightweight deep learning architecture named PatchBT-Net to reduce computational complexity and maintain accuracy. PatchBT-Net mainly includes technologies such as

patch stems and channel attention blocks, which can effectively classify brain tumors in MRI data. PatchBT-Net uses 99.98% fewer parameters than the pre-trained VGG16 model and achieves 97.55% accuracy on the CE-MRI dataset.

S12.3 | 10:54-11:05

Layer pipeline Extensible and Modularized Processing Unit Hardware Architecture Design and Implementation for AI Accelerator

Chung-Bin Wu,Chung-Ting Guo
National Chung -Hsing University , Taichung, Taiwan, R.O.C.

This paper proposes a modular processing module for neural network accelerators called the Process Element Cluster (PE Cluster). To address the substantial computational demands and high memory bandwidth requirements of convolutional neural networks, a pipelined design utilizing two sets of Deep Learning Accelerators (DLAs) is employed. Each DLA comprises N sets of PE Clusters, with the size of N (a non-zero natural number) dynamically adjusted based on the computational load of each network layer. Each PE Cluster consists of 64 PEs, and due to its modular nature, it can be extended to 64×N, supporting up to 16 sets. Furthermore, the proposed PE Cluster architecture supports both 1×1 and 3×3 convolution operations and can be applied to accelerate convolution operations across different networks. This paper also details the FPGA and CHIP hardware implementations for the YOLO v3 network architecture. In the FPGA implementation, 512 PEs achieve high efficiency with a hardware utilization rate of only 6.21%. In the 40nm CHIP implementation, with 32-bit Input/Output, 256 PEs, and a clock frequency set to 200MHz, the energy efficiency reaches 558.7 GOPS/W.

S12.4 | 11:06-11:17

A 4K@60FPS JND Prefiltering Hardware Design for Real-Time Versatile Video Coding

Pin-Chen Li, Yu-Hsin Kuo, and Tian-Sheuan Chang
Department of Electronics and Electrical Engineering, National Yang Ming Chiao Tung University, Hsinchu, Taiwan

In recent years, deep learning-based Just Noticeable Difference (JND) prefilters for video coding have gained popularity due to their ability to effectively remove perceptual redundancy, thereby enhancing coding efficiency. However, the high model complexity poses a significant challenge for real time processing requirements. This paper introduces a super light weight model that significantly reduces model complexity by 92.16%. By employing hardware-friendly activation functions and perceptual-first quantization for intermediate data, the model maintains similar quality compared to the original version. The hardware design employs a fully parallel streaming architecture to meet real-time demands and a tile-based scheduling approach to minimize buffer size. Implemented using a 28nm

CMOS process, the design requires approximately 338,800 gate counts when operating at 600MHz and achieves a 4K@60FPS processing rate. These advancements demonstrate the potential for practical applications in high-resolution video coding scenarios.

S12.5 | 11:18-11:29

Development of Motion-based Blood Oxygen and Heart Rate Sensing Technology with Real-time AI Algorithm for Anti-Tremor Noise Reduction

Li-Chuan Hsu, Chen-Peng Wang, Yao-Feng Liang, Wei-Da Chen, and Shin-Chi Lai
[1]Department of Automation Engineering, National Formosa University, Huwei 632301, Taiwan
[2]Department of Electrical Engineering, National Formosa University, Huwei 632301, Taiwan
[3] Department of Electronic Engineering, National Yunlin University of Science and Technology, Douliu 64002, Taiwan
[4] Smart Machinery and Intelligent Manufacturing Research Center, National Formosa University, Yunlin 632301, Taiwan

In traditional photoplethysmography (PPG) measurement wristbands, accurate monitoring of blood oxygen and heart rate usually requires the user to remain still to ensure data accuracy. To address this issue, this project integrates software and hardware technologies to develop an anti-motion noise real-time AI algorithm for sports blood oxygen and heart rate sensing technology. The smart wristband uses the ESP32-S3 as the main controller and the MAX30102 as the blood oxygen measurement module to measure PPG signals, paired with a touch-LCD-1.28 display to show PPG signals including blood oxygen (SpO2) concentration and heart rate. It transmits data packets to the platform via integrated Bluetooth 5 (LE). The proposed CNN based denoising autoencoder (DAE-CNN) algorithm filters out baseline drift noise generated during activities. The data from this work shows that the DAE-CNN algorithm achieves a Percent Root-mean-square Difference (PRD) performance of 5.5%. In terms of Root Mean Square Error (RMSE), it records a value of 0.005, outperforming other deep learning models with similar parameter quantities. In summary, the proposed system enhances the adaptability of wearable devices to various life scenarios and ensures continuous and accurate health monitoring during sports activities and daily life.