Oral S17

AI Processing for Human and Machinery Signals

Date/Time	8/4(五) 10:30—11:30
Chair(s)	楊家驤／國立臺灣大學電機工程學系 林承鴻／元智大學電機工程學系

S17.1 | 10:30—10:45

NSFormer：Hardware-Oriented Non-Stalled Transformer Based Model for Automatic Speech Recognition

Tseng-Jen Li, and Tian-Sheuan Chang
Institute of Electronics, National Yang Ming Chiao Tung University

The transformer-based model has become the de facto backbone of the state-of-the-art automatic speech recog-nition (ASR) task recently due to its excellent performance and higher parallelism. However, the attention module in the transformer model consists of two critical functions, softmax and layer normalization, that need multi-pass computation and result in hardware stall. For the softmax function, a first pass is exponent operations, the second pass is the summation of the exponented sequence, and the third pass is the division. In the case of layer normalization, it also relies on the runtime statics and requires multi-pass operations across the hidden dimension of the acoustic feature. Consequently, we propose a hardware-oriented non-stalled transformer, NSFormer, which eliminates the above runtime-statics dependent operations but with comparable performance for fully-pipelined ASR hardware.

S17.2 | 10:45—11:00

A Hardware Accelerator for Analyzing Faults in CNC Machinery Using CNN Network on FPGA

Ching-Che Chung[1], Ya-Ching Chang[1],Chun-Chi Chen[2] and Wei-Ming Huang[1]
[1]Department of Computer Science and Information Engineering, National Chung Cheng University
[2]Department of Electrical Engineering, National Chiayi University

Leveraging advancements in machine learning and IoT, this paper proposes a real-time solution for CNC machinery fault detection. The system uses a binary weight convolutional neural network (CNN) to analyze vibration data collected by sensors. Unlike prior research, this work focuses on both model accuracy and hardware efficiency. By using fixed-point operations, computational speed increases and memory usage

reduces. Implemented on FPGA, the method achieves 95.07% accuracy at a 130MHz clock frequency.

S17.3 | 11:00—11:15

VLSI Design of Compact-Shortcut Denoising Autoencoder for ECG Signal

Yan-Ting Lin, Ming-Hwa Sheu, Jia-He Lin
Department of Electronic Engineering, National Yunlin University of Science and Technology

The proposed Compact-Shortcut Denoising Autoencoder (CS-DAE) effectively removes the noise in the ECG signal. It is a lightweight neural network architecture and a VLSI accelerator design suitable for low hardware cost, lower parameters, and less calculation. The proposed compact and shortcut technique compresses the features and passes through the shortcut line. This process reduces the memory requirements of the operation and enhances the noise-reduction effect. In addition, the encoder and decoder operate the Pixel-Shuffle and Pixel-Un-Shuffle process, which prevents the loss of the features caused by down-sampling and up-sampling operations. The CS-DAE neural network architecture decreases the amount of memory size required and the amount of computation successfully while maintaining high accuracy. Using MITDB and NSTDB datasets for model training and testing, the average PRD is 46.30% and 10.50. Based on the proposed CSDAW, the VLSI architecture is designed using Verilog and implemented on the TUL PYNQ™-Z2 development board with a low power consumption of only 1.65W.

S17.4 | 11:15—11:30

An End-to-End Neural Network with Clustering Enhancement for Configurable Online Speaker Diarization

Shao-Wen Cheng, Kuan-Wei Chen, Yen-Chin Liao, Hsie-Chia Chang
National Yang Ming Chiao Tung University

Speaker diarization is the process of identifying "who spoke when" in multi-speaker audio recordings. This process segments the audio into consistent speaker regions and assigns them to the respective individuals in the conversation. To handle overlaps and enable real-time applications, we propose an end-to-end speaker diarization model that utilizes a clustering-based technique to dynamically adjust the number of speakers. We enhance audio quality through pre-emphasis and audio super-resolution and adapt to overlapped regions using permutation-invariant and data augmentation. With a rolling buffer and dynamic threshold adjustment, our model achieves real-time inference and flexible speaker capability. Our experiments on the CallHome datasets demonstrate a

Diarization Error Rate (DER) of 13.09% for the two-speaker dataset and 18.97% for the complete dataset in real-time processing, with a Real-Time Factor (RTF) ranging from 0.15 to 0.18. These results highlight the efficiency and effectiveness of our proposed method for handling real-time applications.