Session Oral 12 (8/8 Thu. 11:30 – 12:30)

Session Topic: Design and Optimization for Storage, Memory, and System

Session Chair: Chien-Chung Ho (National Chung Cheng University)

Room: 7F 論語廳+大學廳

1. 11:30 – 11:42 (SD11) On Enhancing Lifetime and Performance of Fine-tuning Neural Networks Through an Overhead-Reduced Design on NVM-based System

Szu-Yu Chen and Chien-Chung Ho

National Chung Cheng University

Convolutional neural network (CNN) which is one class of neural network has become one of the dominated applications in the computer vision field. Due to the needs of reducing training time and improving training performance, the fine-tuning neural network is widely adopted to avoid the time-

consuming procedure of training a neural network from scratch. Since the continuously growing data/model size and a large number of supported CNN techniques on the neural network system, it needs to increase the DRAM size significantly. However, it is impractical to scale up the DRAM size on the neural network system since DRAM can incur many issues, such as scaling limitation and leakage power problems. This work aims at exploring a solution of how to resolve issues caused by adopted large scale CNN application on DRAM. To explore a cost-efficient solution for the large scale CNNs without using DRAM, this work proposes to exploit non-volatile memory (NVM) as main memory because of its high scalability, low read latency, and near-zero leakage power. However, the inherent properties of NVM, such as longer write latency and worse endurance, can significantly affect the performance of CNNs and reduce the lifetime of NVM. To improve the performance and lifetime issues of NVM-based CNN system, this work proposes a split-FFT approach to reduce the number of write operation on NVM while fine-tuning neural networks. Besides, this work also proposes a writing strategy based on the characteristic of fine-tuning neural network with our proposed split-FFT approach. To examine the effectiveness of the proposed approaches, a series of experiments were conducted. The experiment results show that the proposed approaches successfully average the bit flip and enhance the lifetime of NVM. To be more specific, compared to the conventional FFT convolution approach, a 1.95x performance improvement, and a 50% reduction of bit flip were achieved.

2. 11:42 – 11:54 (SD12) Design a Fault Tolerant System Using System-Level Redundancy

Sih-Kai Shen and Peng-Sheng Chen

National Chung Cheng University

In this paper, we design a fault tolerant system using system-level redundant techniques. The whole structure consists of a primary system connected to a redundant system using network socket programming APIs. A heartbeat mechanism checks whether the primary system is alive. GlusterFS, an open-source distributed file system, aggregates the disk storage resources to provide dependable storage.

In addition, distributed multithreaded checkpointing (DMTCP) stores the execution states of the application to allow resumption from failure. A GUI tool is also developed to assist users in building up the proposed fault tolerant system. Preliminary experimental results for benchmarking and test situations show that the proposed approach can improve fault tolerance and allow the operation to resume after failure.

3. 11:54 – 12:06 (SD13) Pthread’s Spinlock is Unfair

Shi-Wu Lo

National Chung Cheng University

Pthread is a standard library defined by POSIX. Most operating systems including BSD, Linux, Solaris, HPUX and Window use pthreads as their thread library. Although a higher-level programming language defines object-oriented or function-based thread library, the underlying layer uses Pthread. For example, Java, Android, and OpenMP use Pthread to implement their multi-thread libraries. In our research, it was found that the implementation of the spinlock in GNU's Pthread library is unfair in the many core architecture. A few cores have a chance to get a lock several times higher than other cores. We reimplemented the lock and unlock functions of Pthread's spinlock library and changed it to an algorithm called ticket-lock. The performance of the new spinlock has dropped by 10%, but it guarantees fairness.

4. 12:06 – 12:18 (SD14) Real-World Anomaly Detection in Videos Using Spatio-Temporal Autoencoders

   Po-Ju Lin and Pao-Ann Hsiung

   National Chung Cheng University

   Surveillance videos capture a variety of realistic anomalies, which are challenging to detect due to the fuzzy definition of anomalous behavior and complex monitoring scenarios. This article proposes a novel spatio-temporal autoencoder network using 3DCNN and ConvLSTM to learn the characteristics of video anomalies. Experimental results show that this method can detect anomalies in the video with at least 2.4% improvement in AUC accuracy compared to the state-of-the-art ConvLSTM.

5. 12:18 – 12:30 (SD15) A Novel Approach for Story Generation

   Wei Lin, Ting-Hsuan Chien, and Rong-Guey Chang

   National Chung Cheng University

   The sequence transformer models are based on complex recurrent neural network or convolutional networks that include an encoder and a decoder. High-accuracy models are usually represented by used connect the encoder and decoder through an attention mechanism. Neural story generation is an important thing. If we can let computers learn the ability of story-telling, computers can help people do more things. Actually, the squence2squence model combine attention mechanism is being used to Chinese poetry generation. However, it difficult to apply in Chinese story generation, because there are some rules in Chinese poetry generation. Therefore, we trying to use 500 human-labeled summarization of paragraphs from a classic novel named “Demi-Gods and Semi-Devils” (天龍八部) to train the transformer network by the low resource. In our experiment, we got a low loss rate between different epoch.