Skysong Innovations
Advancing the Arizona State University Knowledge Enterprise

1475 N. Scottsdale Road, Suite 200
Scottsdale, AZ 85287-3538
Phone: 480 884 1996 Fax: 480 884 1984

Case ID:M22-233P^
Published: 2/27/2023

Inventors

Deliang Fan
Fan Zhang
Li Yang

Contact

Physical Sciences Team

Masked-Based Learning Method for Neural Network Multiple Task Adaption

Nowadays, one practical limitation of deep neural networks (DNNs) is their high degree of specialization to a single task. This motivates researchers to develop algorithms that can adapt the DNN model to multiple tasks sequentially, while still performing well on past tasks. This process of gradually adapting the DNN model to learn from different tasks over time is known as multitask adaption. Fine-tuning is a natural way to adapt the current model (i.e., the backbone model) to a new task. However, updating the parameters of the backbone model could result in the forgetting of old knowledge upon earlier tasks, thus degrading the performance. This phenomenon is known as catastrophic forgetting, which widely exists in multi-task adaption. To alleviate catastrophic forgetting, several mask-based methods have been proposed i.e., Piggyback and KSM, which only learn a task-specific mask with respect to all weights for each new task, while keeping the backbone model fixed.

From the DNN hardware accelerator design domain, DNN involves a huge amount of multiply–and-accumulate (MAC) operations and data movement. In traditional von Neumann architecture (e.g., CPU, GPU), the data movement consumes 100× higher energy than a floating-point operation. Recently, in-memory computing (IMC) has attracted an increasing interest due to its ability to execute computing tasks directly within the memory array. Among different volatile/non-volatile IMC designs, a resistive random-access memory (ReRAM) crossbar-based design is a promising candidate for a next generation DNN accelerator, due to its simple structure, high on/off ratio, high density, multibit per cell storage, and fabrication compatibility with CMOS. There is a need for a method and device for accelerating DNN inference with multiple task adaption in order to reduce mask memory size and reduce energy consumption.

Researchers at Arizona State University (ASU) have developed a ReRAM crossbar-friendly mask-based learning method that leverages the mask-based learning algorithm's benefit to avoid catastrophic forgetting in multitask learning, and also could be easily implemented in existing crossbar-based DNN accelerator hardware with minimal peripheral circuits and mask memory overhead, and more importantly, with no need to re-program ReRAM cell values. This method achieves high accuracy on new tasks while only requiring a small mask memory footprint. Moreover, this method avoids cell re-programming or tuning, saving significant energy during new task adaption.

Related publication: XBM: A Crossbar Column-wise Binary Mask Learning Method for Efficient Multiple Task Adaption

Potential Applications:

- Deep Neural Networks/Deep Neural Network Accelerators
- Artificial Intelligence

Benefits and Advantages:

- Hardware friendly crossbar column-wise mask – each learned mask value (1/0) controls the on/off of entire crossbar column for the new task inference instead of each element in prior works
- Mask size reduction (e.g., assuming 72x72 crossbar size, only a single mask value is needed to control one column, i.e., a group of 8x3x3 kernels, instead of 72 separate mask values)
- Gumbel-Sigmoid trick leveraged to better estimate the gradient of the mask during back-propagation
- Compared to other mask-based methods, ASU's method saves up to 40% inference energy and reduces the mask size to only 1.4% while maintaining similar accuracy