A ferroelectric ternary content-addressable memory to enhance deep learning models

December 5 2019, by Ingrid Fadelli

Credit: Ni et al.

Most deep-learning algorithms perform well when trained on large sets

of labeled data, but their performance tends to decline when processing new data. Researchers worldwide have thus been trying to develop techniques that could improve the ability of these algorithms to generalize well across both new and previously processed data, enabling what is known as lifelong learning.

Researchers at the University of Notre Dame and GlobalFoundries Fab1 have recently developed a new method to facilitate lifelong learning in artificial neural networks, which entails the use of a ferroelectric ternary content-addressable memory component. Their study, featured in *Nature Electronics*, was aimed at replicating the human brain's ability to learn rapidly from only a few examples, adapting to new tasks based on past experiences.

"When a trained deep neural network encounters previously unseen classes, it often fails to generalize from its prior knowledge and must re-learn the network parameters to extract relevant information from the given class," Kai Ni, one of the researchers who carried out the study, told TechXplore. "This necessitates that large amounts of labeled data be made available for network training."

One approach designed to enhance the performance of deep neural networks on previously unseen data entails the integration of an attentional memory component. This component allows the algorithms to base their analyses on previously acquired knowledge, adapting it to tackle new and yet somewhat similar tasks. Algorithms with an attentional memory component, known as memory augmented neural networks (MANNs), are typically able to extract features from data, store them in their attentional memory and retrieve them when completing a new task.

"A key function of the memory module is content-based addressing, where the distance between a search vector and all stored vectors is

calculated to find the closest match. In a conventional approach, the stored memory vectors (in DRAM) need to be transferred to a compute unit (CPU or GPU) to compare distances with a given query," Ni said. "As such, energy dissipation and latency limitations can represent significant challenges to scaling up MANNs. In this work, we propose to apply ferroelectric ternary content addressable memory (TCAM) as the network's attentional memory to overcome this bottleneck."

By computing the distance between a query vector and each stored memory entry directly within itself, the TCAM component introduced by Ni and his colleagues avoids expensive data transfers. TCAM essentially relies on the fact that the discharge current through a match line is proportional to the Hamming distance (HD) between the query and stored entry.

Sensing this discharge current allows the researchers to compute the HD directly within the memory component in parallel. TCAM also allows deep learning models to perform content-based memory updates rather than random address-based data updates.

"To enable the efficient interaction between the neural network (working with floating number) and the TCAM array (only computing the HD distance), we applied a locality sensitive hashing (LSH) function to map a real-valued feature vector extracted from the NN to a binary signature space, which enables a Hamming distance based nearest neighbor search within the TCAM array," Ni explained.

Ni and his colleagues evaluated their ferroelectric TCAM prototype in a series of trials where a deep neural network had to learn how to complete new tasks based on one or more examples. When implemented on a GPU backed by external DRAM, their method led to classification accuracies that approach those obtained by a more conventional method based on cosine distance calculation (e.g., a 99.5 percent accuracy versus

a 99.05 percent accuracy for a 20-way, five-shot learning problem). Remarkably, the TCAM-based system achieved accuracies similar to those of the more conventional approach with a 60-fold reduction in energy consumption and 2,700-fold reduction in latency for a single search operation.

"The contributions of this research are several fold," Ni said. "Firstly, we demonstrated the most compact TCAM cell to date, which is only composed of two FeFETs, but provides the highest memory density and probably overall the best performance among all the other alternatives. Secondly, we demonstrated the functionality of HD distance calculation with a TCAM array. Finally, we applied the TCAM kernel in the MANN for one-shot learning and provide an end-to-end system solution."

In the future, the new memory component proposed by Ni and his colleagues could aid the development of more efficient deep learning-based models that perform well on both familiar and new tasks. The researchers are now planning to develop a larger TCAM-based prototype that could enable a further increase in performance.

"The existing work demonstrates our approach on a small scale due to our measurement setup limitations," Ni said. "We are planning to design a larger array, along with important peripheral circuits, so that the TCAM can be a stand-alone module. Besides that, device level optimization is still necessary to improve the FeFET endurance, variation, and reliability etc."

More information: Kai Ni et al. Ferroelectric ternary content-addressable memory for one-shot learning, *Nature Electronics* (2019). DOI: 10.1038/s41928-019-0321-3