One-shot learning with memory augmented neural network

Michael Niemier, Notre Dame

A major limitation of deep learning is inefficient adaptation to new data – an essential feature of lifelong learning. 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. This necessitates that large amounts of labelled data be made available for network training. However, the biological brain learns at a rapid pace from just a few examples and can efficiently generalize from past experiences. As such, learning how to learn (or meta-learning), is an active research area in ML to enable artificial general intelligence. One promising approach for implementing meta-learning are memory augmented neural networks (MANNs), where features extracted from a neural network can be stored and retrieved from an attentional memory (typically DRAM). Relevant features for a classification task are extracted from a few training examples and are stored in the network's memory and later retrieved to make accurate predictions. 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. As such, energy dissipation and latency limitations can represent significant challenges to scaling up MANNs. Alternative memory architectures that support massively parallel searches are highly desirable. Hardware architectures that utilize content addressable memories (CAMs) as attentional memories have recently been proposed, in which the distance between a query vector and each stored entry is computed within the memory itself, thus avoiding expensive data transfer. Furthermore, ultra-compact, energy-efficient, and nonvolatile CAM cells based on two ferroelectric FETs (FeFETs) — as well as designs based on other technologies — have recently been developed and demonstrated. As will be discussed, CAM-based MANNs exhibits classification accuracies that approach those obtained with the conventional cosine distance calculation when implemented on a GPU backed by external DRAM and can afford substantial energy and latency savings.

Michael Niemier is currently a Professor at the University of Notre Dame. His research interests include designing, facilitating, benchmarking, and evaluating circuits and architectures based on emerging technologies. Currently, Niemier's research efforts are based on new transistor technologies, as well as devices based on alternative state variables such as spin. He is the recipient of an IBM Faculty Award, the Rev. Edmund P. Joyce, C.S.C. Award for Excellence in Undergraduate Teaching at the University of Notre Dame, and best paper awards such as at ISLPED in 2018. Niemier has served on numerous technical program committees for design related conferences (including DAC, DATE, ICCAD, etc.), and has chaired the emerging technologies track at DATE, DAC, and ICCAD. He is an associate editor for IEEE Transactions on Nanotechnology, as well as the ACM Journal of Emerging Technologies in Computing. He is a senior member of the IEEE.