RRAM-based Nonvolatile In-Memory Computing Macro with Precision-Configurable In-Field Nonlinear Activation

Yiran Chen, Associate Professor, Duke University

RRAM featuring high-density and high-energy-efficiency demonstrates great potential in developing neural network processors. The analog-digital conversion for supporting the analog computing nature of RRAM devices and digital data transition induces excessive design overhead. In this work, we present a RRAM-based nonvolatile in-memory-computing macro, which integrates a 64Kb RRAM array for synaptic weighting and in-field nonlinear activation (IFNA). IFNA merges ADC and activation computation by leveraging its nonlinear working region, which increases 8.2× density and eliminates the need for additional circuits to introduce nonlinearity. IFNA can be flexibly configured to support 1- to 8-bit activation precision. It speedups computation by 2x compared with the conventional separate ACC design. The real-time testing of MNIST and CIFAR-10 datasets on our chip prototype achieves the accuracy of 94.2% and 78.5%, respectively. The chip's power is 1.52mW with a power efficiency of 3.36 TOPS/W.