

Deploying Recurrent Neural Networks on AMD FPGAs: A GRU Implementation on the Versal AI Engine

Wednesday, 1 April 2026 13:00 (20 minutes)

This talk presents an overview of neural network deployment on reconfigurable hardware, with a particular focus on modern AMD FPGA platforms and the Versal Adaptive Compute Acceleration Platform (ACAP). The discussion begins with examples from physics applications where reinforcement learning and recurrent neural networks are jointly employed for real-time control and decision-making.

An introduction to FPGA technology is then provided, covering the fundamental hardware components, common development workflows, and programming approaches using hardware description languages (HDL) and high-level synthesis (HLS). System-on-Chip (SoC) architectures are discussed, leading to a detailed presentation of the AMD Versal platform as a heterogeneous architecture integrating programmable logic, processing systems, and AI Engines.

The talk subsequently reviews existing frameworks for deploying neural networks on FPGA-based systems. Finally, a case study is presented describing the design and implementation of a Gated Recurrent Unit (GRU) on the Versal AI Engine, highlighting architectural considerations and practical challenges associated with mapping recurrent neural networks to this platform.

Student

Yes

Primary author: SAPKAS, Michail (UniPD - INFN Padova)

Presenter: SAPKAS, Michail (UniPD - INFN Padova)

Session Classification: Contributed talks