EFFICIENT BRAINS FOR DIGITAL MEDICINE: MEMORY-CENTRIC OPTIMIZATION OF CNNS AND SNNS USING

ABSTRACT

The deployment of Convolutional Neural Networks (CNNs) and Spiking Neural Networks (SNNs) in medical imaging—such as MRI reconstruction, CT scan analysis, and real-time diagnostics—is critically hampered by intensive memory demands and energy constraints. Conventional hardware architectures, plagued by the von Neumann bottleneck, struggle with inefficient data movement, leading to high latency and power consumption. This review comprehensively analyzes memory-centric optimization strategies to enhance the performance of CNNs and SNNs on hardware accelerators, including In-Memory Computing (IMC), FPGAs (specifically PYNQ-Z2), GPUs, and ASICs. It systematically evaluates techniques such as model pruning, quantization, weight sharing, and data tiling, detailing their impact on reducing memory footprint and computational overhead. The findings indicate that hybrid optimization approaches, integrated with specialized hardware, can significantly mitigate memory bottlenecks. The review concludes that the future of AI in digital medicine lies in co-designing memory-efficient algorithms with advanced hardware accelerators to enable fast, accurate, and energy-efficient diagnostic tools.

EXISTING SYSTEM

The prevailing paradigm for deploying CNNs and SNNs in medical imaging relies on a disjointed combination of standard hardware (CPUs/GPUs) and generalized deep learning frameworks, with memory management often being an afterthought.

Disadvantages of the Existing System:

1. Pronounced von Neumann Bottleneck: The physical separation of memory and processing units in CPU/GPU architectures leads to excessive data movement. This is the primary cause of high latency and energy consumption, making real-time processing of

high-resolution medical images (e.g., 3D MRI volumes) inefficient and power-prohibitive for portable devices.

2. Generalized and Inefficient Hardware Utilization: Platforms like general-purpose GPUs are designed for broad parallelism but are not optimized for the specific, sparse computational patterns of pruned or quantized neural networks. This results in underutilized computational resources and inefficient memory access patterns, failing to fully leverage the potential of model optimization techniques.
3. Fragmented and Non-Adaptive Optimization: Memory optimization techniques like pruning and quantization are often applied in isolation and as a one-time step during model design. This approach lacks the dynamism to adapt to varying medical imaging workloads and does not integrate seamlessly with the underlying hardware's memory hierarchy, leading to suboptimal performance and a difficult trade-off between model accuracy and efficiency.

PROPOSED SYSTEM

The proposed system is a memory-centric, hardware-software co-designed framework that tightly integrates advanced memory optimization algorithms with specialized accelerators to create a holistic solution for medical imaging AI.

Advantages of the Proposed System:

1. In-Memory Computing to Eliminate Data Transfer: By adopting IMC architectures, computation is performed directly within the memory unit. This fundamentally bypasses the von Neumann bottleneck, leading to dramatic reductions in latency (e.g., 25% acceleration for MRI reconstruction) and energy consumption (e.g., up to 80% savings), as demonstrated by Tan et al. (2021) [49].
2. Hardware-Aware Hybrid Optimization: The framework employs a synergistic combination of pruning, quantization, and data tiling that is specifically tailored for the target accelerator (FPGA, ASIC, etc.). For example, on FPGAs, this approach has shown a 2.5x inference speedup and a 30% reduction in memory usage by aligning the

optimized model structure with the hardware's reconfigurable logic and memory blocks [50].

3. Dynamic and Cross-Domain Efficiency: The proposed system moves beyond static optimization by exploring AI-driven memory management and cross-domain strategies. This allows for intelligent, real-time resource allocation across different hardware platforms (e.g., FPGA-GPU hybrids), ensuring optimal performance, accuracy, and energy efficiency for diverse medical imaging tasks, from real-time ultrasound to complex 3D CT analysis.

SYSTEM REQUIREMENTS

- **H/W System Configuration:-**
- Processor - Pentium –IV
- RAM - 4 GB (min)
- Hard Disk - 20 GB
- Key Board - Standard Windows Keyboard
- Mouse - Two or Three Button Mouse
- Monitor - SVGA

SOFTWARE REQUIREMENTS:

- **Operating system** : Windows 7 Ultimate.
- **Coding Language** : Python.
- **Front-End** : Python.
- **Back-End** : Django-ORM
- **Designing** : Html, css, javascript.
- **Data Base** : MySQL (WAMP Server).