ABSTRACT

RELIABILITY-AWARE STRUCTURAL PRUNING FOR CONVOLUTIONAL NEURAL NETWORKS

By:
Muhammad Zaky Firdaus
21/477171/PA/20637

This research proposes a novel approach to structural pruning of Convolutional Neural Networks (CNNs) that considers both model efficiency and reliability when deployed on hardware accelerators. While existing pruning methods like DepGraph successfully reduce model size and computational costs, they don't account for the network's resilience to hardware faults, which is crucial for deployment on embedded systems.

The methodology is evaluated on three CNN architectures: LeNet-5, ResNet-18, and ResNet-50, trained on MNIST and CIFAR-10 datasets respectively. Both 32-bit floating-point (FP32) and 8-bit integer quantized (INT8) models are examined across different pruning ratios. Fault injection experiments simulate single-bit upsets in the datapath of NVIDIA Deep Learning Accelerator (NVDLA) to assess reliability under hardware faults.

Results demonstrate that both activation-based and magnitude-based pruning methods achieve comparable compression performance, with minimal accuracy differences (typically <0.1%). LeNet-5 exhibits remarkable robustness, maintaining >99% accuracy until 70% pruning ratio, while ResNet architectures remain stable until approximately 60-65% pruning before experiencing rapid accuracy degradation. Quantized INT8 models show superior fault tolerance compared to FP32 models, with significantly lower Silent Data Corruption (SDC) rates (0.08% vs 1.31% for ResNet-50). Fault injection analysis reveals that exponent bits in FP32 representation are highly vulnerable to bit-flips, particularly bit 30 (MSB), while INT8 models show peak vulnerability at the sign bit (bit 7).

Keywords: Pruning, CNNs, Reliability, NVDLA

ABSTRAK

PEMANGKASAN STRUKTURAL SADAR-KERENTANAN UNTUK JARINGAN SARAF KONVOLUSIONAL

Oleh:
Muhammad Zaky Firdaus
21/477171/PA/20637

Penelitian ini mengusulkan pendekatan baru untuk *structural pruning* pada *Convolutional Neural Networks* (CNN) dengan mempertimbangkan efisiensi model sekaligus keandalannya saat dijalankan pada *hardware accelerator*. Meskipun metode pruning yang ada seperti DepGraph berhasil mengurangi ukuran model dan beban komputasi, metode tersebut belum memperhitungkan ketahanan jaringan terhadap gangguan perangkat keras (*hardware faults*), yang sangat krusial untuk penerapan pada sistem tertanam (embedded systems).

Metodologi ini dievaluasi pada tiga arsitektur CNN: LeNet-5, ResNet-18, dan ResNet-50, yang dilatih pada dataset MNIST dan CIFAR-10. Model dengan representasi 32-bit floating-point (FP32) dan kuantisasi 8-bit integer (INT8) diuji pada berbagai rasio pruning. Eksperimen *fault injection* disimulasikan untuk menciptakan single-bit upset pada *datapath* dari NVIDIA *Deep Learning Accelerator* (NVDLA) guna menilai keandalan model di bawah gangguan perangkat keras.

Hasil menunjukkan bahwa metode pruning berbasis aktivasi maupun berbasis magnitudo mampu mencapai kinerja kompresi yang sebanding, dengan perbedaan akurasi yang sangat kecil (biasanya <0>Silent Data Corruption (SDC) yang jauh lebih rendah (0,08% vs 1,31% pada ResNet-50). Analisis *fault injection* mengungkapkan bahwa bit eksponen dalam representasi FP32 sangat rentan terhadap bit-flip, khususnya bit ke-30 (bit paling signifikan/MSB), sedangkan model INT8 menunjukkan kerentanan tertinggi pada bit tanda (sign bit, bit ke-7).

Kata Kunci: Pruning, Jaringan Saraf Konvolusional, Kerentanan, NVDLA