ABSTRACT

AN EMPIRICAL STUDY OF THE IMPACT OF A SINGLE BIT-FLIP ERROR ON LOW PRECISION QUANTIZED CONVOLUTIONAL NEURAL NETWORKS

By
Gabriel Kaunang
20/457769/PA/19807

Convolutional Neural Networks (CNNs) have revolutionized computer vision tasks, particularly in critical systems such as self-driving cars. To enable the deployment of CNNs on resource-constrained devices, quantization techniques have been employed to reduce model size and computational requirements. This study investigates the resilience of low-precision quantized CNN models, specifically focusing on 4-bit integer (int4) quantization, against single bit-flip errors.

Through software-level fault injection experiments on the ResNet50 and MobileNetV1 architectures, this study assesses the impact of bit-flips on model performance and analyzes the factors influencing fault propagation. The findings reveal that 4-bit quantized models exhibit lower resilience compared to higher-precision counterparts, with a mismatch probability of 0.4%, making it the second most vulnerable datatype assessed. Furthermore, this study observes that the model's architecture, particularly layer positioning and specific components, critically influences fault propagation and the resulting mismatches.

This study highlights the importance of considering the resilience of quantized models to hardware-level faults, especially in resource-constrained environments and safety-critical applications. The results provide insights into the trade-offs between model compression and error resilience, emphasizing the need for robust quantization techniques and error mitigation strategies in low-precision CNN deployments on resource-limited devices.

Keywords: Deep Learning, Bit-Flip, Reliability

INTISARI

STUDI EMPIRIS DAMPAK KESALAHAN BIT-FLIP TUNGGAL PADA JARINGAN NEURAL KONVOLUSIONAL TERKUANTISASI PRESISI RENDAH

Oleh
Gabriel Kaunang
20/457769/PA/19807

Jaringan Saraf Konvolusional (Convolutional Neural Networks, CNN) telah merevolusi bidang visi komputer, terutama dalam sistem kritis seperti mobil self-driving. Untuk memungkinkan penerapan CNN pada perangkat dengan sumber daya terbatas, teknik kuantisasi telah digunakan untuk mengurangi ukuran model dan kebutuhan komputasi. Studi ini menyelidiki ketahanan model CNN terkuantisasi presisi rendah, khususnya berfokus pada kuantisasi integer 4-bit, terhadap kesalahan single bit-flip.

Melalui eksperimen injeksi kesalahan tingkat perangkat lunak pada arsitektur ResNet50 dan MobileNetV1, studi ini menilai dampak bit-flip pada kinerja model dan menganalisis faktor-faktor yang mempengaruhi perambatan kesalahan. Temuan ini mengungkapkan bahwa model terkuantisasi 4-bit menunjukkan ketahanan yang lebih rendah dibandingkan dengan model presisi yang lebih tinggi, dengan probabilitas ketidaksesuaian sebesar 0,4%, menjadikannya tipe data kedua yang paling rentan di antara yang dinilai. Selain itu, studi ini mengamati bahwa arsitektur model, terutama posisi lapisan dan komponen spesifik, sangat mempengaruhi perambatan kesalahan dan ketidaksesuaian yang dihasilkan. Studi ini menyoroti pentingnya mempertimbangkan ketahanan model terkuantisasi terhadap kesalahan tingkat perangkat keras, terutama dalam lingkungan dengan sumber daya terbatas dan aplikasi yang kritis terhadap keselamatan.

Hasil penelitian memberikan wawasan tentang trade-off antara kompresi model dan ketahanan kesalahan, menekankan perlunya teknik kuantisasi yang kuat dan strategi mitigasi kesalahan dalam penerapan CNN presisi rendah pada perangkat dengan sumber daya terbatas.

Kata Kunci: Deep Learning, Bit-Flip, Reliability