Quantização em Hardware com Recursos Limitados

Caio Martins de Abreu

08 de Abril de 2025

Resumo

Este trabalho tem como objetivo investigar técnicas de quantização aplicadas em sistemas computacionais com recursos limitados, com foco especial em aplicações de aprendizado de máquina e processamento de sinais. A quantização permite representar dados numéricos com menor precisão, reduzindo o uso de memória e a carga computacional, o que a torna essencial para dispositivos embarcados e ambientes de baixo consumo energético.

1 Introdução

Com o crescimento exponencial das aplicações baseadas em redes neurais profundas (DNNs), especialmente em tarefas como visão computacional, tornou-se imperativo adaptar esses modelos para plataformas de hardware embarcado, como Field-Programmable Gate Arrays (FPGAs). Apesar do alto desempenho e flexibilidade oferecidos por essas arquiteturas, a implementação direta de modelos com precisão total (floating-point de 32 bits) se mostra ineficiente em termos de consumo energético, utilização de recursos lógicos e largura de banda de memória.

Nesse cenário, técnicas de quantização pós-treinamento (PTQ) surgem como uma alternativa promissora para reduzir a complexidade computacional dos modelos sem exigir um re-treinamento completo. Entre essas técnicas, abordagens como a AHCPTQ (*Accurate and Hardware-Compatible Post-Training Quantization*) destacam-se por manter a precisão do modelo original ao mesmo tempo que garantem compatibilidade com operações aritméticas otimizadas para hardware, fator crucial em implementações FPGA-friendly.

Além disso, estratégias baseadas em quantização treinável em ponto fixo, como apresentado por Lin et al. em *Trainable Fixed-Point Quantization for Deep Learning Acceleration on FPGAs*, demonstram que, ao incorporar a quantização como parte do processo de otimização durante o treinamento, é possível maximizar a eficiência computacional sem comprometer o desempenho preditivo da rede. Essa abordagem permite a criação de modelos mais adequados à execução em hardware com recursos limitados, ao mesmo tempo em que aproveita a maleabilidade das FPGAs na definição da largura de palavra e do caminho de dados.

Paralelamente, a substituição de representações de ponto flutuante padrão por formatos de menor precisão — como half-precision (16 bits) ou formatos personalizados de floating-point — também tem sido explorada como alternativa para melhorar o throughput e a eficiência energética em aceleradores baseados em FPGA, como discutido por Zhang et al. em *Low Precision Floating-point Arithmetic for High Performance FPGA-based CNN Acceleration.*

Dessa forma, o presente trabalho propõe a investigação e a implementação de técnicas de quantização em modelos de redes neurais convolucionais, com ênfase em estratégias que equilibram acurácia, eficiência computacional e compatibilidade com FPGAs. O objetivo é explorar abordagens modernas que possibilitem a execução de inferência eficiente em dispositivos embarcados, contribuindo para o avanço de soluções de IA embarcada em cenários do mundo real.

Com o crescimento acelerado do uso de inteligência artificial e processamento de sinais em dispositivos móveis, sensores inteligentes e sistemas embarcados, há uma demanda crescente por soluções computacionais que conciliem eficiência e desempenho. No entanto, muitos desses dispositivos operam sob restrições severas de energia, memória e capacidade de processamento.

Neste contexto, a **quantização** surge como uma técnica fundamental para viabilizar a implementação de algoritmos complexos em hardware com recursos limitados. Ao reduzir a precisão dos dados, representando valores reais com menos bits por meio de representações como ponto fixo (*fixed-point*), é possível diminuir significativamente o custo computacional das operações matemáticas envolvidas.

Além da economia de memória e energia, a quantização permite acelerar o tempo de inferência de modelos de aprendizado de máquina, tornando possível sua execução em tempo real em microcontroladores, FPGAs e outros dispositivos de baixo custo. Contudo, essa redução de precisão vem acompanhada de desafios, como o controle de erros numéricos, a estabilidade dos algoritmos e o equilíbrio entre precisão e desempenho.

Este trabalho tem como objetivo explorar diferentes estratégias de quantização, analisar seus impactos em aplicações práticas, e propor abordagens que maximizem o aproveitamento dos recursos disponíveis sem comprometer significativamente a acurácia dos sistemas implementados.