ARITMÉTICA PRECISA BASEADA EM NIBBLES PARA ACELERADORES DE APRENDIZAGEM PROFUNDA

Luiz Fernando Heidrich Duarte, Cesar Albenes Zeferino.

Mestrado em Computação Aplicada
Computação Aplicada - Sistemas Embarcados e Distribuídos

O desenvolvimento de aceleradores de hardware para sistemas embarcados com capacidade de processar modelos quantizados de redes neurais profundas tem sido um tópico de pesquisa ativo e promissor. Trabalhos anteriores reduziram com sucesso a largura de bits circuitos utilizados para a inferência de redes neurais profundas quantizadas, demonstrando a sua aplicabilidade em sistemas embarcados com restrições de recursos computacionais. No entanto, o suporte ao aprendizado profundo por processadores compactos ainda é altamente incomum. Isso se deve à maior demanda por memória, desempenho e precisão dos processos de *backpropagation* e gradiente descendente utilizados no treinamento das redes neurais profundas. Neste trabalho, propomos circuitos aritméticos para compressão de números, multiplicação, acumulação e ativação não linear, que usam quantização logarítmica e arredondamento estocástico para viabilizar a inferência e o treinamento de redes neurais profundas em aceleradores de hardware compactos, baseados em “nibbles”. A quantização logarítmica expande a faixa de representação numérica e aumenta a precisão de números próximos a zero, pois codifica os números na forma de expoentes na base dois ao invés utilizar a representação em ponto fixo, mais comum em trabalhos de quantização de modelos de redes neurais profundas. Além disso, o arredondamento estocástico habilita o aumento progressivo da precisão de cálculos paralelos/iterativos típicos do processamento de modelos de redes neurais profundas, viabilizando a execução de cálculos acurados a partir de circuitos compactos com largura de bits reduzida. Desta forma, esta abordagem reduz a demanda por memória, eleva a capacidade de processamento por área e habilita a execução de cálculos precisos a partir de números de 4 bits. Para comprovar a efetividade dos circuitos propostos, sintetizamos e analisamos os circuitos propostos e desenvolvemos experimentos teóricos e práticos. A análise da síntese identifica a redução de área e dissipação de energia dos circuitos propostos em comparação com soluções equivalentes. Os experimentos teóricos demonstram a acurácia destes circuitos em operações com tensores. Os experimentos práticos demonstram a aplicabilidade dos circuitos na inferência e treinamento de redes neurais recorrentes e convolucionais. Empiricamente, demonstramos o potencial de aplicação desses circuitos em aceleradores de hardware para processamento de redes neurais profundas em sistemas embarcados enxutos, em cenários nos quais é exigido aprendizado contínuo.

Palavras-chave: Aprendizado Profundo; Redes Neurais Profundas; Quantização; Arredondamento Estocástico; Aceleradores de Hardware.

.

ISSN 1983-117X