UNIVERSIDADE E COMUNIDADE EM CONEXÃO

XXXV SALÃO de INICIAÇÃO CIENTÍFICA

6 a 10 de novembro

Evento	Salão UFRGS 2023: SIC - XXXV SALÃO DE INICIAÇÃO CIENTÍFICA DA UFRGS
Ano	2023
Local	Campus Centro - UFRGS
Título	Otimização e quantização de redes neurais em FPGA
Autor	MATHEUS ALMEIDA SILVA
Orientador	ANTONIO CARLOS SCHNEIDER BECK FILHO

O aumento dos custos no treinamento e inferência de redes neurais de ponta motivou a busca por formas de reduzir recursos sem comprometer a precisão. Embora o treinamento exija poder computacional, a implantação deve ser viável em hardware de baixa potência. Soluções incluem quantização de redes neurais ao custo de precisão e desenvolvimento de aceleradores de hardware específicos. Os FPGAs são ideais para explorar tais soluções, utilizando precisões customizadas para alcançar acurácia necessária. FINN, framework da Xilinx/AMD, oferece ferramentas para exploração de inferência de redes neurais em FPGA, incluindo Quantização, Compilação e Implantação. A pesquisa desenvolvida busca otimizar redes neurais quantizadas com o FINN. O fluxo do framework, que parte de uma rede neural quantizada em formato ONNX, envolve etapas de otimização e geração de camadas em HDL via HLS em um formato de dataflow. Nessa etapa dois tipos de otimizações podem ser implementadas personalizando a quantidade de Unidades de Processamento(PE) e Dados processados simultaneamente(SIMD) nas camadas HLS. Para explorar o paralelismo a pesquisa partiu de uma configuração balanceada d e explorou diferentes configurações modificando somente um fator k de paralelismo,i.e fator multiplicativo dos valores de PE/SIMD a partir do base(e.g em determinada camada onde em k=1 o PE=16 em k=2 o PE=32). Os resultados obtidos até aqui demonstram que o aumento do fator k implica em redução significativa do tempo de processamento e, consequente aumento do throughput ao custo da área utilizado do FPGA.