Comparação de aceleradores de deep learning feitos em FPGA

Antônio Lima Della Flora

Introdução

Os algoritmos de Deep Learning estão cada vez mais presentes na nossa vida, seja em processamento de linguagem natural ou visão computacional e reconhecimento de padrões. Contudo, esses algoritmos demandam muitas operações da CPU e grande largura de banda, o que faz com que CPUs comuns não atinjam o nível esperado de performance.

Por causa disso, surgiram aceleradores de hardware que utilizam circuitos integrados específicos ou FPGAs para acelerar os modelos de Deep Learning.

Os artigos

1. FPGA-Based Accelerators of Deep Learning Networks for Learning and Classification: A Review

Este artigo revisa aceleradores baseados em FPGA para redes de deep learning, com foco em redes neurais convolucionais (CNNs). Explora técnicas como loop unrolling para paralelismo, reutilização de dados via memória interna e redução de largura de banda com precisão reduzida. Aborda modelos como AlexNet e VGG, destacando desafios como uso de recursos e acesso à memória externa. Propõe um framework escalável para otimizar CNNs em FPGAs, alcançando até 420 GOPS com a arquitetura FlexFlow, mantendo alta eficiência energética. Combina automação RTL com flexibilidade HLS, ideal para aplicações de alto desempenho e baixo consumo

2. A Ubiquitous Machine Learning Accelerator With Automatic Parallelization on FPGA

Este artigo apresenta um acelerador de aprendizado de máquina em FPGA com paralelização automática e execução fora de ordem, suportando clustering, redes

neurais profundas, sequenciamento genômico e filtragem colaborativa. Utiliza uma arquitetura 2D com núcleos reconfiguráveis e escalonamento fora de ordem, alcançando até 25x de aceleração frente a CPUs Intel. Testado em Xilinx Virtex-7, oferece alta eficiência energética (3,3-6,1 W) e flexibilidade, superando CPUs e GPUs. Combina automação e programação de alto nível para aplicações intensivas em dados.

3. Optimized FPGA-based Deep Learning Accelerator for Sparse CNN using High Bandwidth Memory

Este artigo apresenta um acelerador otimizado de aprendizado profundo baseado em FPGA para CNNs esparsas, utilizando memória de alta largura de banda (HBM2). Propõe um método de empacotamento de matrizes esparsas com representação em bitmap, reduzindo a computação em arrays sistólicos. Implementado no Intel Deep Learning Accelerator (DLA) em um FPGA Stratix 10 MX, alcança ganhos de desempenho de 2,06x/3,44x no módulo de computação e 2,2x/2,1x contra uma CPU SkyLake. Comparado a uma GPU V100, oferece eficiência energética 1,7x/2x superior. A integração com HBM2 otimiza operações limitadas por memória, superando DDR4 em até 3x.

Tabela comparativa

Nome do Artigo	Hardware Utilizado	Speedup	Throughput
Optimized FPGA-based Deep Learning Accelerator for Sparse CNN using High Bandwidth Memory (FCCM 2021)	**FPGA (Experimento):** Intel Arria 10 GX1150 (com DDR4), Intel Stratix 10 MX2100 (com HBM2) **Comparação**: CPU Intel SkyLake Gold 6130, GPU NVidia Tesla V100	2.7x (ResNet-50) / 3x (MobileNet)	~1581 images/s (MobileNet), ~725 images/s (ResNet-50)
A Ubiquitous Machine Learning Accelerator With Automatic Parallelization on FPGA (IEEE TPDS 2020)	**FPGA (Experimento):** Xilinx Virtex-7 **Comparação:** CPU Intel Core 2 Duo @ 2.33 GHz, GPU NVIDIA Geforce 720	**Clustering**: 12.32x; **Deep Learning (Training)**: 5.64x	**Clustering**: 8.9 GFLOPS; **Deep Learning**: 4.2 GFLOPS (training)
FPGA-Based Accelerators of Deep Learning Networks for Learning and Classification: A Review (IEEE Access 2019)	**Tipo:** Artigo de Revisão **Hardware Coberto:** Vários FPGAs (e.g., Virtex-4/5/6/7, Zynq, Stratix-V, Arria 10, KU060). Não apresenta um hardware *próprio* para um novo acelerador.	-	Citou diferentes aceleradores, dentro eles o FlexFlox demonstrou 420 GOPS operando a 1 GHz

Conclusão

Tive bastante dificuldade com fazer esse projeto pois toda a área de modelos inteligentes é nova pra mim, então existem muitos termos que eu não conheço e tive que aprender. Minha ideia inicial era buscar aplicações em FPGA ou outro hardware para modelos de estimativa de consumo futuro de energia, mas não consegui encontrar nada sobre isso então acabei generalizando mais a pesquisa.

Ao generalizar a pesquisa consegui encontrar mais artigos, e foi possível perceber que os aceleradores conseguem melhorar bastante a performance dos modelos de deep learning, mas que é necessário ter um bom design e projeto e eles dependem das características da aplicação.