Estudo e Comparação de Métodos de Multiplicação para aplicação em Hardware Dedicado.

Tiago de Oliveira Ambrósio(IC)[1], Gabriel Antonio Fanelli de Souza(PQ)[1]
[1]*Universidade Federal de Itajubá.*

Palavras-chave: ELM. Computação científica. Regressão linear. Álgebra Matricial.

Introdução

Este estudo investiga e compara algoritmos de multiplicação de matrizes aplicados ao método dos mínimos quadrados, amplamente utilizado no treinamento de redes neurais Extreme Learning Machine (ELM). ELMs são notáveis por sua rapidez e simplicidade em aplicações de aprendizado de máquina, onde o treinamento depende da minimização do erro entre a saída prevista e a real dos dados de treinamento, um processo que requer multiplicação de matrizes altamente eficiente.

A multiplicação de matrizes é uma operação computacionalmente intensiva e impacta diretamente o desempenho do método dos mínimos quadrados. Neste contexto, foram avaliados diferentes algoritmos, incluindo o tradicional, Strassen, Winograd e suas variantes otimizadas, visando reduzir a complexidade computacional. Esses algoritmos foram implementados e testados em um sistema digital descrito em Verilog, possibilitando uma análise de desempenho baseada em hardware, próxima das condições reais de uso.

Para simulações e análise numérica, utilizou-se o Matlab, enquanto os softwares Quartus Prime 18 e Modelsim foram empregados para o desenvolvimento e simulação dos circuitos digitais, permitindo a verificação e validação das implementações no nível de hardware. Este processo possibilitou avaliar a eficiência de cada algoritmo em uma plataforma FPGA (Field Programmable Gate Array), tecnologia que combina flexibilidade com alto potencial de paralelismo.

A seleção dos algoritmos baseou-se em sua complexidade teórica e potencial de paralelização. Testes preliminares foram realizados em pequena escala para ajustes finos, garantindo desempenho e precisão antes de aumentar a complexidade dos experimentos. Essa abordagem gradual considerou as limitações dos FPGAs, onde equilibrar eficiência e precisão é particularmente desafiador devido às restrições de consumo de energia e capacidade de armazenamento. Análises detalhadas de tempo de execução e uso de recursos de hardware forneceram uma visão aprofundada sobre o impacto de cada algoritmo em operações de larga escala.

A metodologia adotada permitiu identificar algoritmos de multiplicação de matrizes que maximizam a eficiência computacional e a precisão dos cálculos, aspectos críticos para o treinamento das ELMs. Este estudo também destaca as vantagens e limitações dos FPGAs na implementação de operações matemáticas intensivas, contribuindo para o desenvolvimento de sistemas de aprendizado de máquina de alto desempenho e baixo consumo energético.

Metodologia

Para a realização deste estudo, avaliou-se inicialmente a viabilidade de implementar o método dos Mínimos Quadrados em FPGAs, explorando sua capacidade de paralelização e potencial de customização de circuitos. Esse estudo preliminar abordou requisitos de hardware, consumo energético e otimização de operações matemáticas intensivas para determinar a eficiência do uso de FPGAs no método proposto. Em paralelo, implementou-se o método no ambiente MATLAB, fornecendo uma base comparativa entre as abordagens em software e hardware.

O desenvolvimento e teste da aplicação em FPGA foram realizados no Quartus Prime, software que suporta design, compilação e simulação de circuitos digitais. Seguindo uma adaptação do método de treinamento de [1], estruturou-se o código em Verilog para implementar operações fundamentais. Inicialmente, revisaram-se os conceitos matemáticos essenciais do método dos Mínimos Quadrados, com ênfase na precisão e adequação dos dados para processamento em FPGA. Com base nesse estudo, analisaram-se representações numéricas, como ponto fixo e ponto flutuante, para definir a mais apropriada ao objetivo do projeto. Essa

escolha foi crucial para equilibrar precisão e eficiência no uso dos recursos de hardware, considerando as limitações de área e latência dos FPGAs.

A construção do circuito seguiu uma metodologia incremental, começando com módulos básicos, como unidades de soma e multiplicação, que foram integrados em módulos mais complexos responsáveis pela multiplicação de matrizes e o método dos Mínimos Quadrados. Essa abordagem modular facilitou a identificação e correção de erros nas fases iniciais, reduzindo a complexidade de depuração nas etapas finais. Testes unitários asseguraram a precisão e o desempenho de cada módulo antes de sua integração completa.

Como a multiplicação de matrizes impacta significativamente o desempenho do método dos Mínimos Quadrados, essa operação foi o foco da otimização. Foram investigados quatro métodos de multiplicação: Strassen, Desloca e Acumula, Vedic e Wallace Tree. O método de Strassen, por exemplo, reduz a complexidade da multiplicação por meio da estratégia de divisão e conquista, enquanto o método Desloca e Acumula otimiza o número de operações aritméticas. Os métodos Vedic e Wallace Tree, oriundos da computação de alta velocidade, destacaram-se pela capacidade de paralelização, sendo particularmente eficazes em hardware.

Cada método foi inicialmente implementado e testado em menor escala para avaliar seu desempenho e precisão, o que facilitou a análise das limitações e vantagens específicas. A escolha final dos métodos baseou-se em uma análise comparativa detalhada de tempo de execução e uso de recursos de hardware. Os métodos mais eficientes foram adaptados para implementação em larga escala, ajustando-se a precisão e a eficiência de hardware para atender às limitações específicas dos FPGAs, como a capacidade limitada de memória e o gerenciamento de latência.

Resultados e discussão

Com base na superfície obtida em [2], e nos dados de treinamento obtidos em [3], utilizando o método de treinamento proposto, é possível obter o valor da REMQ (raiz do erro médio quadrático) da superfície.

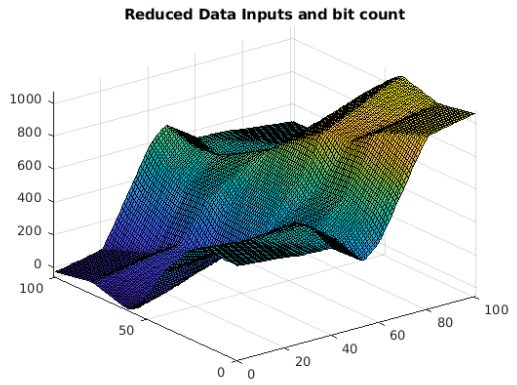

Figura 1 – Superfície Gerada Pesquisa

REMQ = 4,81%

A figura 1 mostra a superfície de controle gerada utilizado o método de treinamento proposto pela pesquisa.

A partir do estudo da precisão das representações numéricas em uma operação de multiplicação de matrizes simplificada, foi montada a tabela 1:

Representação	Erro Médio (%)	Erro Máximo (%)	Nº Bits
Binário	35,67	60,12	8
Binário	18,34	32,56	16
Binário	8,12	15,45	32
Ponto Fixo	30,45	55,12	8
Ponto Fixo	14,23	27,34	16
Ponto Fixo	6,75	12,34	32
Ponto Flutuante	0,15	0,3	16
Ponto Flutuante	0,01	0,05	32

Tabela 1 – Comparação Métodos Numéricos

A partir das simulações dos diferentes algoritmos de multiplicação, foi possível implementar uma versão simplificada dos mesmos para fins de comparação.

Dentre os métodos utilizados, podemos citar como principais vantagens e desvantagens de cada um:
- O algoritmo de Strassen reduz a complexidade computacional, porém, sua implementação em hardware é complexa devido e pode apresentar instabilidade numérica.
- O Wallace Tree é altamente eficiente em velocidade, porém o mesmo sofre com problemas de escalabilidade.
- A multiplicação Védic permite implementações rápidas com lógica simples. No entanto, não escala bem para operandos maiores.
- O método Shift Add é simples e fácil de implementar. Contudo, é mais lento em comparação com outros métodos. Em todos os casos as vantagens e desvantagens precisam ser ponderadas e dependem da maneira que o método é implementado.

Na figura 2 é possível observar a distribuição e utilização de elementos lógicos de um dos algoritmos estudados.

Também é possível observar a tabela 2 com a utilização de elementos lógicos de todos os métodos estudados.

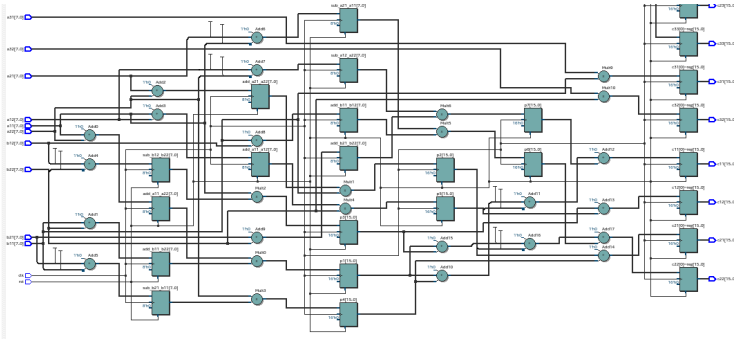

Figura 2 – Visualização RTL método Strassen.

Técnica	Utilização Lógica(em ALMs)	Latência (ciclos de clock)
Shift-Add	226	3
Strassen	494	4
Vedic	501	3
Wallace Tree	250	4

Tabela 2 – Comparação Métodos.

Os resultados obtidos na comparação dos métodos evidenciam, para a aplicação específica, as vantagens e desvantagens na utilização de cada método. Além disso, o valor de REMQ encontrado demonstra a precisão numérica do método estudado.

Com base nos dados da representação numérica e nos algoritmos de multiplicação, é possível escolher a combinação que melhor se adapta à aplicação pretendida.

Conclusões

Com base nos resultados obtidos, conclui-se que a representação numérica em ponto fixo de 32 bits e o uso do algoritmo "Desloca e Acumula" para multiplicação de matrizes reduzem a complexidade da implementação do método dos mínimos quadrados em hardware dedicado, sem afetar significativamente a precisão dos resultados. Além disso, o valor de REMQ obtido quando comparado com os valores obtidos em [1] demonstram um aumento relativamente baixo na imprecisão média do sistema, em torno de 0,46% no pior caso.

Dessa forma, espera-se que o método de multiplicação de matrizes possa ser integrado ao módulo de resolução do método dos mínimos quadrados, inicialmente em uma implementação em *FPGA* e, posteriormente, dependendo das características de área, desempenho e potência do circuito, em uma implementação em *ASIC(Application Specific Integrated Circuit)*.

Agradecimentos

Gostaria de agradecer ao CNPq pelo financiamento da pesquisa, a UNIFEI pelo apoio e estrutura, ao meu orientador Gabriel Antonio Fanelli de Souza e ao grupo de microeletrônica da UNIFEI.

Referências

[1] TAVARES, L. H. B. Análise de método de treinamento de redes neurais do tipo elm para implementação em hardware. Iniciação Científica, Universidade Federal de Itajubá (UNIFEI), Itajubá, 2023.

[2] BENDIB, B. Advanced fuzzy mppt controller for a tand-alone pv system. Energy Procedia., 2015

[3] SOUZA, G. A. F. de. A novel power-reducing architecture for a current mode analog interval type-2 fuzzy logic inference system. Thesis of Doctor of Science in Electronic Engineering and Computer Science, Field of Electronic Systems and Devices – Instituto Tecnológico de Aeronáutica, São José dos Campos., 2019.