A. Ciências Exatas e da Terra - 2. Ciência da Computação - 8. Processamento Paralelo e Distribuído

HWRNA: Redes neurais artificiais em hardware

Rodrigo Amador Coelho[1]
Marluce Rodrigues Pereira[2]
Wilian Soares Lacerda[3]

1. Aluno Graduação - Depto de Ciência da Compuatção - UFLA - Orientado.
2. Prof. Dr. - Depto de Ciência da Compuatção - UFLA - Orientador.
3. Prof. Dr. - Depto de Ciência da Compuatção - UFLA - Co-Orientador.

RESUMO:
A computação paralela visa minimizar o tempo de computação, racionalizando o uso do hardware disponível. O paralelismo acontece quando a solução de um problema é implementada para ser executada em diferentes processadores capazes de trabalhar de forma cooperativa. As GPUs (Graphic Processing Units ou Unidades de Processamento Gráfico) atuais permitem a execução paralela de aplicações com o objetivo de obter alto desempenho.
Este trabalho demonstra a utilização de GPU para o processamento paralelo de redes neurais artificiais. A rede neural artificial é uma método para se resolver problemas, modelo este baseado na natureza, mais especificamente no cérebro. Neste estudo foi utilizada a Rede Neural Adaline (Adaptive Linear Neuro). Esta rede foi desenvolvida para reconhecimento de padrões e, especificamente neste trabalho, foi utilizada para reconhecimento de caracteres alfanuméricos.
A implementação foi realizada para o processador gráfico desenvolvido pela NVIDIA, utilizando a arquitetura de computação paralela chamada CUDA (Compute Unified Device Architecture), que tira proveito do mecanismo de computação paralela das GPUs. A programação utilizando CUDA esconde a complexidade da GPU, permitindo assim que os programadores não se preocupem com detalhes complexos de hardware durante a programação.
Os experimentos foram realizados com diferentes números de caracteres alfanuméricos e de threads de execução. Os resultados mostraram que é possível obter desempenho para este tipo de aplicação. Além disso, para melhor aproveitar o desempenho da GPU, é necessário minimizar ao máximo o número de chamadas a GPU e de cópias de informação entre as memórias da GPU e da CPU.

Palavras-chave: Paralelismo, GPU, Rede Neural Artificial.