PONTIFÍCIA UNIVERSIDADE CATÓLICA
DO RIO DE JANEIRO

Daniel Salles Chevitarese

ANNCOM – Biblioteca de Redes Neurais Artificiais para Alto Desempenho Utilizando Placas de Vídeo

Dissertação de Mestrado

Dissertação apresentada como requisito parcial para obtenção do grau de Mestre pelo programa de Pós Graduação em Engenharia Elétrica do Departamento de Engenharia Elétrica da PUC-Rio.

Orientadores: Marley M. B. Rebuzzi Vellasco
Dilza Mattos Szwarcman

Rio de Janeiro

Março de 2010

PONTIFÍCIA UNIVERSIDADE CATÓLICA
DO RIO DE JANEIRO

Daniel Salles Chevitarese

ANNCOM – Biblioteca de Redes Neurais Artificiais com Treinamento Acelerado por Placas Gráficas

Dissertação apresentada como requisito parcial para obtenção do grau de Mestre pelo Programa de Pós-Graduação em Engenharia Elétrica do Departamento de Engenharia Elétrica do Centro Técnico Científico da PUC-Rio. Aprovada pela Comissão Examinadora abaixo assinada.

Profa. Marley Maria Bernardes Rebuzzi Vellasco
Orientadora
Departamento de Engenharia Elétrica – PUC-Rio

Profa. Dilza Mattos Szwarcman
Co-Orientadora
Departamento de Engenharia Elétrica – PUC-Rio

Profa. Cristiana Bentes
UERJ

Prof. Ricardo Cordeiro de Farias
Universidade Federal do Rio de Janeiro

Prof. Carlos Roberto Hall Barbosa
Programa de Pós-Graduação em Metrologia – PUC-Rio

Prof. José Eugenio Leal
Coordenador Setorial do Centro
Técnico Científico

Daniel Salles Chevitarese

Graduou-se em Engenharia de Computação pela PUC-Rio em 2007

Ficha Catalográfica

Chevitarese, Daniel Salles

ANNCOM – Biblioteca de redes neurais artificiais para desempenho utilizando placas de vídeo / Daniel Salles Chevitarese; orientadores: Marley M. B. Rebuzzi Vellasco, Dilza Mattos Szwarcman. – 2010. 97 f. : il. (color.) ; 30 cm

Dissertação (mestrado) – Pontifícia Universidade Católica do Rio de Janeiro, Departamento de Engenharia Elétrica, 2010.

Inclui bibliografia

1. Engenharia elétrica – Teses. 2. Redes neurais artificiais. 3. Engenharia de software. 4. Computação de alto desempenho. 5. GPGPU. 6. CUDA. I. Vellasco, Marley M. B. Rebuzzi. II. Szwarcman, Dilza Mattos. III. Pontifícia Universidade Católica do Rio de Janeiro. Departamento de Engenharia Elétrica. IV. Título.

CDD: 621.3

Para minha Vivi.

Agradecimentos

À minha Orientadora Profa. Marley M. B. Rebuzzi Vellasco pelo estímulo e parceria para a realização deste trabalho.

À Dra. Dilza Mattos Szwarcman que foi muito mais que minha co-orientadora, mas uma grande amiga.

À CAPES e à PUC-Rio, pelos auxílios concedidos, sem os quais este trabalho não poderia ter sido realizado.

Ao ICA pelo conhecimento e treinamento.

Aos professores que participaram da Comissão Examinadora.

A todos os amigos e familiares que de uma forma ou de outra me estimularam e ajudaram.

Aos meus pais que sempre procuraram me guiar pelo caminho do bem.

Resumo

Chevitarese, Daniel Salles; Vellasco, Marley M. B. Rebuzzi **ANNCOM – Biblioteca de Redes Neurais Artificiais para Alto Desempenho Utilizando Placas de Vídeo**. Rio de Janeiro, 2010. 97p. Dissertação de Mestrado - Departamento de Engenharia Elétrica, Pontifícia Universidade Católica do Rio de Janeiro.

As Redes Neurais Artificiais têm sido utilizadas com bastante sucesso em problemas de previsão, inferência e classificação de padrões. Por essa razão, já se encontram disponíveis diversas bibliotecas que facilitam a modelagem e o treinamento de redes, tais como o NNtool do Matlab ou o WEKA. Embora essas bibliotecas sejam muito utilizadas, elas possuem limitações quanto à mobilidade, à flexibilidade e ao desempenho. Essa última limitação é devida, principalmente, ao treinamento que pode exigir muito tempo quando existe uma grande quantidade de dados com muitos atributos. O presente trabalho propõe o desenvolvimento de uma biblioteca (ANNCOM) de fácil utilização, flexível, multiplataforma e que utiliza a arquitetura CUDA (*Compute Unified Device Architecture*) para reduzir os tempos de treinamento das redes. Essa arquitetura é uma forma de GPGPU (*General-Purpose computing on Graphics Processing Units*) e tem sido utilizada como uma solução em computação paralela na área de alto desempenho, uma vez que a tecnologia utilizada nos processadores atuais está chegando ao limite de velocidade. Adicionalmente, foi criada uma ferramenta gráfica que auxilia o desenvolvimento de soluções aplicando as técnicas de redes neurais de forma fácil e clara usando a biblioteca desenvolvida. Para avaliação de desempenho da ANNCOM, foram realizados seis treinamentos para classificação de clientes de baixa tensão de uma distribuidora de energia elétrica. O treinamento das redes, utilizando a ANNCOM com a tecnologia CUDA, alcançou um desempenho quase 30 vezes maior do que a ANNCOM auxiliada pela MKL (*Math Kernel Library*) da Intel, também utilizada pelo Matlab.

Palavras-chave

Redes Neurais Artificiais; Engenharia de Software; Computação de Alto Desempenho; GPGPU; CUDA.

Abstract

Chevitarese, Daniel Salles; Vellasco, Marley M. B. Rebuzzi (Advisor) **ANNCOM – Artificial Neural Network Library for High Performance Computing using Graphic Cards**. Rio de Janeiro, 2710. 95p. MSc Dissertation - Departamento de Engenharia Elétrica, Pontifícia Universidade Católica do Rio de Janeiro.

The Artificial Neural Networks have been used quite successfully in problems of prediction, inference and classification standards. For this reason, are already available several libraries that facilitate the modeling and training networks, such as NNtool Matlab or WEKA. While these libraries are widely used, they have limited mobility, flexibility and performance. This limitation is due mainly to the training that can take a long time when there is a large amount of data with many attributes. This paper proposes the development of a library (ANNCOM) easy to use, flexible platform and architecture that uses the CUDA (Compute Unified Device Architecture) to reduce the training times of the networks. This architecture is a form of GPGPU (GeneralPurpose computing on Graphics Processing Units) and has been used as a solution in parallel computing in the area of high performance, since the technology used in current processors are reaching the limit of speed. Additionally created a graphical tool that helps the development of solutions using the techniques of neural networks easily and clearly using the library developed. For performance evaluation ANNCOM were conducted six trainings for customer classification of a low voltage electricity distribution. The training of networks using ANNCOM with CUDA technology, achieved a performance nearly 30 times greater than the ANNCOM aided by MKL (Math Kernel Library) by Intel, also used by Matlab.

Keywords

Artificial Neural Networks; Software Engineering; High Performance Computing; GPGPU; CUDA.

Sumário

1 Introdução 14
1.1. Motivação 14
1.2. Objetivos 15
1.3. Descrição do trabalho 16
1.4. Estrutura da Dissertação 17

2 Conceitos da Engenharia de Software 18
2.1. Requisitos de Software 18
2.2. Arquitetura de Software 19
2.2.1. Decomposição em Módulos 21
2.3. Padrões de Projeto 23
2.3.1. Composição (*Composite*) 25
2.3.2. Cadeia de Responsabilidade (*Chain of Responsibility*) 26
2.3.3. Estratégia (*Strategy/Policy*) 27
2.4. .NET Framework 29
2.4.1. Linguagem Comum em Tempo de Execução (CLR) 30

3 Conceitos da Computação de Alto Desempenho 33
3.1. Conceitos da Computação Paralela 34
3.1.1. Leis de *Amdahl* e *Gustafson* 35
3.1.2. Tipos de Paralelismo 37
3.2. Os Processadores Gráficos e a GPGPU 39
3.3. CUDA – NVIDIA 42
3.3.1. O Modelo de Programação 44
3.3.2. A Implementação do Hardware 48

4 Desenvolvimento de uma Biblioteca de Redes Neurais de Alto Desempenho – ANNCOM 51
4.1. Levantamento de Requisitos e Arquitetura 52
4.2. Estrutura Básica 56
4.2.1. Componente NeuralNet 57

4.2.2. Modelos de Redes Neurais Artificiais 59
4.2.3. Estrutura NetOutput 60
4.2.4. Estruturas para Cálculo de Erro 61
4.3. Modelo de Treinamento 63
4.3.1. O Treinamento em GPGPU 70
4.3.2. Implementação na GPU 71
4.3.3. Segunda Parte: Inversão Matricial 71
4.4. Ferramenta Gráfica para Criação de Soluções Utilizando Redes Neurais – Clinn 74
4.4.1. Interface utilizando Docas Flutuantes 74
4.4.2. Processo de Criação Automatizada 81
4.4.3. Processo de Treinamento 82

5 Estudo de Casos 84
5.1. Caso 1 – Treinamento dos Comitês para Classificação do Cliente Fraudador da Light 84
5.1.1. Estrutura do Sistema 86
5.1.2. Resultados do Treinamento e Testes de Desempenho 88

6 Conclusões e Trabalhos Futuros 93

Referências Bibliográficas 95

Lista de Figuras

Figura 1 – Início do processo de desenvolvimento de um software. 20
Figura 2 – Parte do diagrama de classes de objetos da ANNCOM em UML. 21
Figura 3 – Estrutura recursiva da ANNCOM. 25
Figura 4 – Cadeia de responsabilidade do método *propagate*. 26
Figura 5 – Algoritmos de treinamento da ANNCOM encapsulados. 28
Figura 6 – O *.NET Framework* e o resto do sistema (MICROSOFT, 2009). 30
Figura 7 – O tempo de execução e o aumento de velocidade de um programa com paralelismo. 37
Figura 8 – Representação gráfica da lei de Amdahl. 37
Figura 9 – Comparativo entre CPU e GPU com relação às operações de ponto flutuante por segundo (NVIDIA, 2009). 40
Figura 10 – Comparativo entre CPU e GPU com relação à largura de banda de memória memória (NVIDIA, 2009). 41
Figura 11 – Diferença entre GPU e CPU com relação à utilização de transistores (NVIDIA, 2009). 42
Figura 12 – Suporte de CUDA para várias linguagens (NVIDIA, 2009). 43
Figura 13 – Grade de blocos de *threads* (NVIDIA, 2009). 44
Figura 14 – Os diferentes níveis de memória nas placas de vídeo da NVIDIA (NVIDIA, 2009). 46
Figura 15 – Modelo de programação heterogêneo (NVIDIA, 2009). 47
Figura 16 – Escalonamento automático feito em CUDA (NVIDIA, 2009). 49
Figura 17 – Modelo de hardware de uma placa gráfica com suporte a CUDA (NVIDIA, 2009). 50
Figura 18 – Visão geral dos componentes principais da ANNCOM. 53
Figura 19 – Suporte a conexão com vários tipos de SGBD. 54
Figura 20 – Suporte da ANNCOM para tempo de execução no Visual Studio. 55
Figura 21 – Adição e edição de coleções na ANNCOM. 56
Figura 22 – Visão geral do desvio de parte do treinamento para GPU. 56

Figura 23 – Esquema com todos os espaços de nomes da biblioteca ANNCOM. 57
Figura 24 – Representação de uma rede neural MLP. 58
Figura 25 – Diagrama de classes simplificado dos modelos de redes neurais implementados nesse trabalho. 59
Figura 26 – Modelo de uma rede Elman. 60
Figura 27 – Diagrama de classes com as estruturas de decodificação das saídas das redes neurais. 61
Figura 28 – Diagrama de classes com os objetos para cálculo de erro, implementados na ANNCOM. 61
Figura 29 – Ilustração de uma descida por gradiente (WIKIPEDIA, 2010). 64
Figura 30 – Exemplo de duas “salas de aula” onde várias redes são treinadas. 68
Figura 31 – Validação cruzada automática da ANNCOM. 68
Figura 32 – Exemplo de funcionamento do Gerenciador de Erro (GE) para cálculo de erros de validação e teste durante o treinamento. 69
Figura 33 – Segundo modelo (resumido) proposto, que executa uma porção maior do código na placa gráfica. 70
Figura 34 - Divisão do processamento em kernels. 71
Figura 35 – Divisão da matriz em blocos de tamanho *n* (nesse caso, *n* é 4). 72
Figura 36 – Atualização das linhas adjacentes. 72
Figura 37 – Blocos atualizados. 73
Figura 38 – Linha de fatores para a multiplicação dos pivôs. 73
Figura 39 – Tela inicial do Clinn. 74
Figura 40 – Movimentação das docas pelo programa. 75
Figura 41 – A doca de propriedades em detalhes. À direita, a figura mostra a facilidade de navegação pelo componente. 76
Figura 42 – Editor de coleção de camadas. 77
Figura 43 – Editor de coleção de neurônios. 77
Figura 44 – Editor de coleção de sinapses. 78
Figura 45 – Doca exploradora de solução. 79
Figura 46 – Explorador de bases de dados. 79

Figura 47 – Possibilidade de visualizar e editar as informações das tabelas. 80
Figura 48 – Doca de documentos com uma lista de arquivos abertos. 81
Figura 49 – Modelo passo a passo para se criar uma nova rede neural. 82
Figura 50 – Função sigmoid de um neurônio da última camada da rede neural. 83
Figura 51 – Visão geral do processo de classificação dos clientes de baixa tensão. 87
Figura 52 – Tempo (em segundos) vs. Tamanho (em número de elementos) da matriz de entada. 91

Lista de Tabelas

Tabela 1 – Principais características das linguagens *.NET*. 32
Tabela 2 – Taxonomia Flynn. 38
Tabela 3 – Modelos de placas usadas no estudo de casos. 84
Tabela 4 – Descrição dos atributos de entrada da base de clientes da Light. 87
Tabela 5 – Tempos referentes à base de clientes comerciais. 88
Tabela 6 – Tempos referentes à base de clientes da região Leste. 89
Tabela 7 – Tempos referentes à base de clientes da região Oeste. 89
Tabela 8 – Tempos referentes à base de clientes da região Litorânea. 90
Tabela 9 – Tempos referentes à base de clientes da região Interior. 90
Tabela 10 – Tempos referentes à base de clientes da região Baixada. 91