3W-10 大規模並列プロセッサAAP-2上でのニューラルネットワークシミュレーション

渡辺琢美　武井雄一郎

NTT LSI研究所

1．はじめに

ニューラルネットワークは、画像認識、音声合成、画像処理等、多くの人工知能や認識科学の分野に応用されている。しかし、そのシミュレーションには多くの計算量を必要とし、中規模のネットワークを学習させるのにも膨大な処理時間を必要とする。本報告では、並列計算機を用いた学習の高速処理[1,2]の1つとして、NTTが開発した大規模並列セルラアレイプロセッサAAP-2[3]を用いたバックプロパゲーションの並列処理手法について述べる。

2．AAP2の構成

AAP-2は、65,536個の1ビット・プロセシングエレメント(PE)のそれぞれに大容量のローカルメモリ(8kビット/プロセッサ)を付加したプロセッサアレイ型の計算機である。図1にAAP-2の構成図を示す。本装置は基本的には2次元アレイ接続のSIMD型マシンであるが、各種修飾機能(修飾機能付きSIMD)、2系統のプログラマブルデータ転送路、高速データ転送機能等の採用により柔軟な構成となっている。各PEには8kビットのローカルメモリが専用に割り当てられている。アーキテクチャ上の主な特徴は次の3点である。

(a)任意のプロセッサ間ネットワークが構成できるプロセッサアレイの可変構造。

(b)PE間をクロックで区切ることなくデータを転送する高速伝搬転送(リップル転送)、伝搬演算(リップル加算)。

(c)伝搬経路短縮用の階層的バイパス転送路。

また、折畳みによるバーチャルアレイの採用、プロセッサ規模の拡張が容易な実装方法の採用により、より大規模なデータの処理にも対応可能となっている。

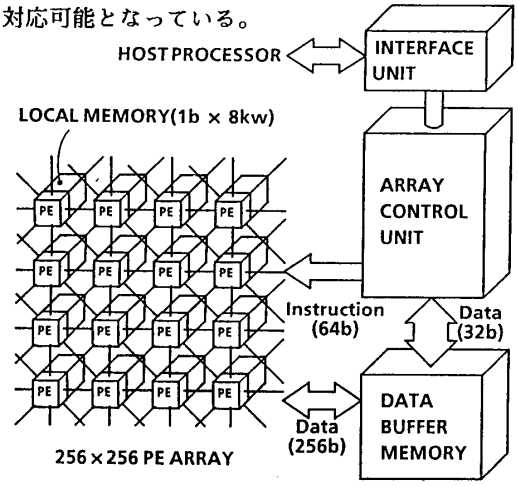

図1 AAP-2システム図

Neural Network Simulation on A Massively Parallel Processor:AAP-2

Takumi Watanabe, Yuichiro Takei

NTT LSI Laboratories

3．AAP－2上での並列処理手法

AAP-2上で実現したネットワークモデルを図2(a)に示す。ある層の各ユニットは次の層のすべてのユニットと結合しており、入力層、中間層、および出力層のユニット数は同一とする。AAP-2へのマッピング方法を図2(b)に示す。図のように重みを各プロセッサに割当てることにより高い並列度で処理が可能となる。本手法におけるもう一つの特徴的な処理は、sigmoid関数値の計算をプロセッサアレイを利用してtable-look-upにより求めている点である。

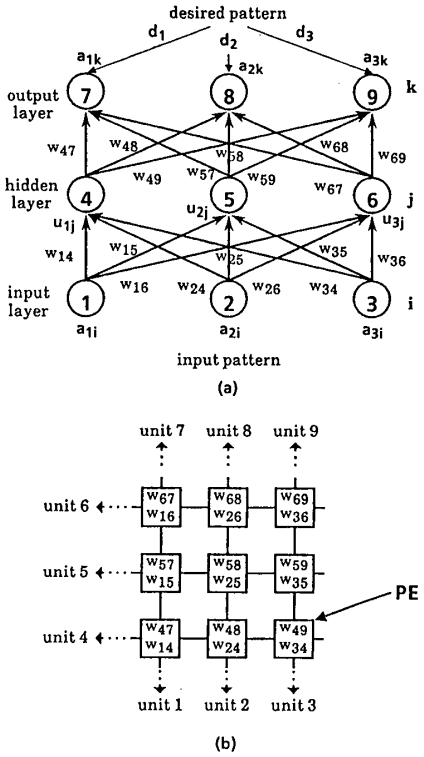

図2 ウエイトとユニットの割り当て。

(a)3層構造ネットワークモデル。(b)プロセッサへの割り当て。

AAP-2での具体的処理手順について以下に述べる。

(1)前向き伝搬(図3(a))

(a)重みの初期値(w)、sigmoid関数テーブル(x,y,f(x),f(y))、入力値(a_i)、および教師信号(d)をホスト計算機から、バイパス転送路を用いたリップル転送命令によりアレイ部に転送する。この時、sigmoid関数テーブルは、各PE行および列毎に入力されていることに注意する。

(b)各PEで重みと入力値との積を計算する。この処理は、それぞれN個の層間のユニットに対して、N^2回必要な重みの計算をすべて並列に行なっている。次に各結果をPEの行(列)に沿ってリップル加算を用いて加算する。これも、N

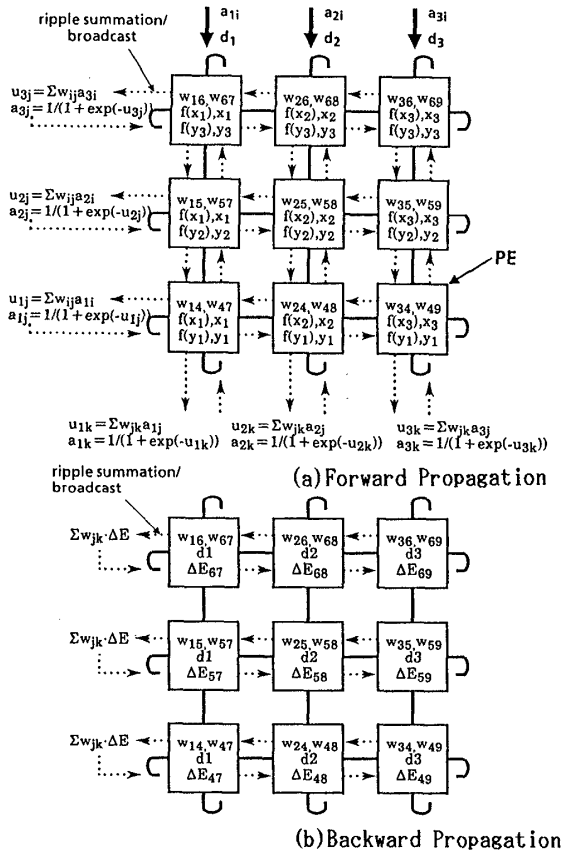

図3 AAP-2上でのバックプロパゲーション。

個の次の層の入力数をPEの各列(行)で全て並列に行っている。

(c)加算された値(次の層(l層)の入力データ(u_l)となる)をPEの行(列)に沿って放送する。

(d)sigmoid関数をu_lに適用した結果は、あらかじめ各PEに記憶されているsigmoid関数値の入力値とu_lを比較することで求められる。ここで、sigmoid関数f(x)=1/exp(-x)である。(この手法では、関数fの値は、アレイの一辺の大きさ、ここでは256段階、に分割される。)求めた結果は、各ＰＥ行(列)に沿って放送され、重みづけされた値が次の層のユニットの入力値となる。

(2)後向き伝搬(図3(b))

(a)各PEにおいて誤差(ΔE)を計算する。

(b)ΔEと重みの積をとり、PE列(行)に沿ってリップル加算を用いて加算する。(出力層と中間層の間の重みの変更の場合は不要。)

(c)各PEでsigmoid関数の微分値を計算し、重みを変更する。

中間層が複数ある場合は、上記のプロセッサを行と列を交互に入れ替えて繰り返し行う。図に示すような重み、入力値、教師信号、sigmoid関数テーブルの各PEへの割当て方法により、処理が高い並列度で行なわれている。

本手法は、以下に示すようないくつかの特徴がある。

(1)ユニットへの入力値の完全並列処理。

(2)バイパスを使用したリップル転送によるPE行(列)に沿った高速データ転送、高速加算。

(3)浮動小数点演算を使用しないテーブル索引によるsigmoid関数値の計算。

(4)各定数および変数のワード長の最適化による高速処理。AAP-2では、長ビット長の演算は1ビットずつシリアルに行うので、必要なワード長に設定することで演算速度の向上が図れる。本プログラムでは、現在1ワード26ビットに設定している。

4．結果と考察

作成したプログラムを文字認識に適用し、AAP-2上で走行させ、処理速度を他の計算機と比較した。ネットワーク構造は、入力層、中間層、出力層の3層構造で、ユニットの数はそれぞれ256とした。各層のユニットは、前後の層のユニットと完全結合している。従って、入力層と出力層の間の結合の数は、131,072である。ここで、AAP-2上でシミュレーションできるユニットあるいは結合の数は、物理的なアレイサイズではなく、メモリ容量によって決まる。本手法では、AAP-2のバーチャル・アレイを用いて、200万結合以上のシミュレーションが可能である。

1回の学習に必要な処理時間は7.33msecであった。これは、18MCPS(million connection per second)である。

表1はいろいろなマシンのバックプロパゲーションの処理速度を比較したものである。IBM3090上でのシミュレーションでは、1回の学習あたり0.33秒のCPU時間を要した。従って、AAP-2はIBM3090の45倍の速度でシミュレーションを実行できる。

本手法は、夫々の層のユニットの数が全て等しい時に最も効率のよい処理ができる。入力層、中間層、出力層のユニット数、がAAP-2のアレイ規模と一致しない場合は、接続関係のない重みを常に0に設定することでシミュレーション可能である。

表1 処理速度比較

Machine	MCPS	Relative Speed
SUN-3/60	0.03	1
IBM 3090-60E	0.40	13.3
16k CM-1 [1]	2.6	86.7
M380 with VP-100	4.7	157
Warp [2]	17	567
AAP-2	18	600

5．まとめ

大規模セルラアレイプロセッサ上での高並列ニューラルネットワークシミュレーション手法ついて述べた。修飾機能付きSIMDおよび高速データ転送機能を採用したAAP-2は1ビット2次元アレイプロセッサの形態をとりながらも柔軟性の高い拡張性に富んだシステムであり、大規模ネットワーク用ニューロコンピュータとして有効である。

[参考文献]

[1] G.Blelloch and C.R.Rosenberg,"Network Learning on the Connection Machine," Proc. of the Tenth Int'l Joint Conf. on AI, pp.323-326,1987.
[2] D.A.Pomerleau et al.,"Neural Network Simulation at Warp Speed," ICNN-89,July,1988.
[3] T.Kondo, et al.,"Pseudo MIMD array processor AAP-2," Proc. of 13th Symposium on Computer Architecture Conf., pp. 330-337,1986.