3W-6

ＴＩＰ－４Ｐにおける ニューロ処理方式

岩下正雄[1]　藤田善弘[1]　石黒尚夫[2]　山崎雅生[2]
[1]日本電気㈱
[2]日本電気技術情報システム開発㈱

１．はじめに

ニューラルネットワークは、学習能力による適応性と汎用性から各分野でその応用が期待され、盛んに研究が進められている。しかし、学習に時間がかかる上、試行錯誤的な実験が多く、処理時間が研究のネックとなっている。ところが層構造のネットワークでは同一層内の各ユニットは独立であり、さらにユニット内の積和演算も並列性を有している。この並列性を十分に引き出すことにより処理時間短縮が可能となる。ただし、このネットワーク構造における逆伝播学習は各ニューロンユニットの計算にネットワーク全体のデータが必要であり、複数のプロセッサ間でのデータ交換が避けられない。

筆者らは、データフロープロセッサＩｍＰＰを６４個実装した大規模画像処理システムＴＩＰ－４Ｐを既に開発した。今回これを用いてニューラル・ネットワークのうち逆伝播学習アルゴリズムの処理方法の検討を行った。

２．ＴＩＰ－４Ｐの構成

ＴＩＰ－４Ｐは、データフロープロセッサＩｍＰＰ８個のリングを１単位とした処理部を更に８個実装して高並列にデータを処理する。各処理部は専用のローカルメモリ（ＬＭ）上でデータを処理する。処理部間のデータの受渡しは、メインイメージメモリ（ＩＭ）を介して行う。ＬＭとＩＭ間のデータ転送は２Ｋワード単位の高速ブロック転送で行う。

３．ＴＩＰ－４Ｐによる並列処理

逆伝播学習アルゴリズムを図１に示す。前述したように、このアルゴリズムは１つのニューロンユニットの計算にネットワーク全体のデータが必要である。従って、ＴＩＰ－４Ｐの複数処理部で分割処理する場合、処理部間でデータの集配交換が必要である。この集配交換すべきデータ量をできるだけ小さくする処理方法が課題である。

３－１．方式１

一つは中間層および出力層のニューロンユニットを複数のグループに分割し、各処理部に担当させる方法がある。即ち図１の各式が、複数処理部で分割されて並列に実行され、式(1)から式(6)までが順番に処理される。しかしこの方法では式(2)のＹ（ｈ）、式(4)の

$$Y(h)=f\left(\sum_{i}\left(Wh(h,i)\times X(i)\right)\right) \quad (1)$$

$$Z(o)=f\left(\sum_{h}\left(Wo(o,h)\times Y(h)\right)\right) \quad (2)$$

$$Do(o)=\left(T(o)-Z(o)\right)\times f'\left(\sum_{h}\left(Wo(o,h)\times Y(h)\right)\right) \quad (3)$$

$$Dh(h)=\sum_{o}\left(Wo(o,h)\times Do(o)\right)\times f'\left(\sum_{i}\left(Wh(h,j)\times X(i)\right)\right) \quad (4)$$

$$Wo(o,h)=Wo(o,h)+Po\times Do(o)\times Y(h) \quad (5)$$

$$Wh(h,i)=Wh(h,i)+Ph\times Dh(h)\times X(i) \quad (6)$$

X(i):入力パタン　Y(h):中間層の出力　Z(o):出力層の出力
Wh(h,i):中間層の重み　Wo(h,i):出力層の重み　T(o):教師信号
i=0～I-1　h=0～H-1　o=0～O-1　f():シグモイド関数
I:入力ユニット数　H:中間ユニット数　O:出力ユニット数

図１　逆伝播学習のアルゴリズム

A Neural Nets Processing Method with TIP-4P
Masao Iwashita[1]　Yosihiro Fujita　Takao Ishiguro[2]　Masao Yamazaki[2]
[1]NEC corporation
[2]NEC Scientific Information System Development

Ｗｏ（ｏ，ｈ）及びＤｏ（ｏ）は各処理部でネットワーク全体のデータが必要である。このため１回の逆伝播学習中に各処理部の間で３種類のデータテーブルの集配交換処理が必要となってくる。特にＷｏ（ｏ，ｈ）は中間層数と出力層数の増加に伴い２乗のオーダーでデータ量が増加し、オーバーヘッドが爆発する。

３－２．方式２

次に中間層の分割は前述と同じだが、出力層の重み行列を転置して分割する方法を考える。即ち出力層の各ニューロンユニット内の重み（Ｈ個）を分割して複数処理部に担当させる。各処理部は、自分が担当する中間層の出力に対する重みだけを全出力層ユニットにわたって保持する。この方法によると、ｒ個の各処理部では全出力層ユニットの積和の部分和ベクトルＺｎ（ȯ）を出力する（ｎ＝０～ｒ－１）。これをいったんＩＭにブロック転送したのち、一つの処理部がこれらを集めてｒ個のベクトルの総和

$$Z(o) = \sum_{n} Zn(o)$$

を求め(3)式まで計算する。ここで求められたＤｏ（ｏ）を全処理部にコピーすれば、あとは(6)式までデータの集配交換なしに各処理部が独立に計算を進められる。しかもＺ（ｏ）は、出力層ユニットの数に比例して増加するだけである。方法１よりも効率がよいことは明かである。

４．オーバーヘッド

ブロック転送は２ｋワード単位のため、転送時間は転送ワード数に対する階段状の関数（ｇ()）となる。２ｋワードのブロック転送時間をＴｂ、ＬＭのアクセス時間をＴｍとし、処理はＬＭの稼働率８０％で行うとする。集配交換処理によるオーバーヘッドは

$$(3r+1)g(o)+o(r+2)Tm \qquad (7)$$

となる。

また処理時間そのものは

$$((5i+7o+5)H/r + 3o)Tm \qquad (8)$$

となる。

６．台数と処理時間

オーバーヘッドを含んだ全体の処理時間は式(7)+式(8)である。またＴＩＰ－４Ｐの場合Ｔｂ≒２５Ｔｍであることから、使用リング数と処理速度を図２に示した。グラフは理想直線と比較しても十分な台数効果を示しており、逆伝播学習を高速に処理できることがわかる。

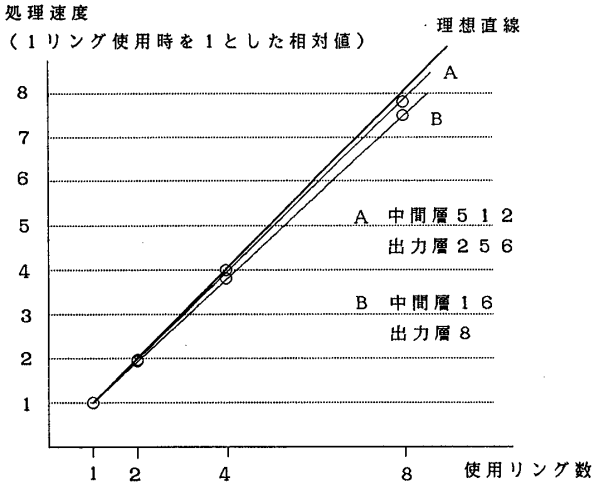

図２　使用リング数と処理速度

７．おわりに

ＴＩＰ－４Ｐのニューラルネットワークの逆伝播学習への適応性について検討し効果的な処理方式を述べた。本方式と高速ブロック転送機能により、逆伝播学習の処理における分散したローカルメモリ間のデータ転送によるオーバーヘッドを十分小さくすることができ、高い性能が得られることが確認された。

最後に、今回の検討にあたりご協力いただいた研究部の方々に深謝する。

［参考文献］

［1］岩下他、”データ駆動画像処理プロセッサＴＩＰ－４Ｐ”、第３７回情処全大、1988

［2］岩下他、”画像処理プロセッサ（ＩｍＰＰ）とニューラルネットワークへの応用”、理研シンポジウム、1988

［3］Runmelhart,D.E.,Hinton,G.E.and Williams,R.J.,"Learning Representations by Back-Propagating Error,"Nature,vol.323,pp.533-536,Oct.9,1986.