ハイパフォーマンス
コンピューティング 68－10
（１９９７．１０．１７）

並列誤差逆伝搬学習法の解析的な学習時間評価

山森 一人　　堀口 進
北陸先端科学技術大学院大学 情報科学研究科
石川県辰口町旭台 1-1

ニューラルネットワークを用いた情報処理は制御やパターン認識などの分野で広く応用されている．しかし，問題が大規模化するにしたがってニューラルネットワークの学習に必要な時間が膨大になる．近年，大規模ニューラルネットワークの学習を並列計算機を用いて高速化する研究が盛んに行われるようになった．しかしながら，これらの並列化手法は並列計算機のアーキテクチャに強く依存している場合が多く，誤差逆伝搬学習法の並列化性能については十分に検討されていない．本論文では，誤差逆伝搬法が持つ3種類の並列性を利用した並列学習モデルを解析し，その並列学習速度について詳しく検討する．これらの並列学習に関する解析結果を用いて実際の並列計算機上へ各モデルを実装し，並列学習法の性能評価を行う．

キーワード： 並列処理，誤差逆伝搬学習，ニューラルネットワーク，性能評価

Theoretical Learning Speed Evaluation of Parallel Back Propagation Algorithms

Kunihito Yamamori　　Susumu Horiguchi

Graduate School of Information Science,
Japan Advanced Institute of Science and Technology
Asahi-Dai, Tatsunokuchi, Ishikawa 923-12, Japan

Multilayer neural network with back-propagation learning requires enormous computation times for large scale problems. To reduce computation times, several parallel learning algorithms have been proposed. However, most of parallel algorithms were specified for particular parallel computers. Performance analyses of parallel back-propagation algorithms have not been investigated sufficiently.

This paper addresses the theoretical performances of parallel back-propagation algorithms. We classify parallel back-propagation algorithms into three models; unit parallel model, learning-set parallel model and pass parallel model. Then, their parallel performances are analyzed theoretically. To confirm theoretical performance estimations, these parallel models are implemented on parallel computer nCUBE/2. It is seen that the learning-set parallel model is most suitable for parallel computers by theoretical analyses and experimental results.

Keywords: Parallel Processing, Back-Propagation Learning, Neural Network, Performance Evaluation

1 はじめに

厳密な入出力関係を定義することなく情報処理を実現できる手法として，ニューラルネットワークが注目されている．なかでも階層型ニューラルネットワークと誤差逆伝搬学習法は，パターン認識や制御などの分野で広く応用されている．しかしながら，大規模なネットワークを必要とする実用規模の問題に対しては，学習を行ってネットワークを収束させるのに膨大な時間が必要になる．そこで近年，誤差逆伝搬学習自体を改良して高速化を行ったり，並列計算機を用いたニューラルネットワークの高速学習に関する研究が数多く行われるようになってきた．

Fahlman[1] は誤差逆伝搬学習のパラメータ変更による挙動変化を詳しく調べ，学習アルゴリズムを改良した quick-propagation と呼ばれる学習法を提案している．塚本ら [2] は，1回の学習パターン提示で学習を行う瞬時学習法を提案している．

しかし，誤差逆伝搬学習におけるアルゴリズムの改良は学習パターンに条件を設ける場合も多く，大規模化ネットワークの高速学習には限界がある．そこで，階層型ニューラルネットワークの並列学習法に関する研究が盛んに行われている．Singer[3] は Thinking Machines 社の超並列計算機 CM-2 上で，誤差逆伝搬学習を並列に行う実装法を検討し，並列学習法の有用性を示した．Muller ら [4] は，45 個の DSP を用いた分散メモリ型の並列計算機を試作し，階層型ニューラルネットワークの各層を複数のプロセッサに割り当てる並列化手法を提案している．Witbrock ら [5] は，並列計算機 GF-11 上に階層型ニューラルネットワークを実装し，ニューラルネットワークの並列学習による高速化アーキテクチャについて考察している．

しかしながら，これらの研究では並列計算機に依存した実装法について評価したもので，誤差逆伝搬学習法の持つ並列性について十分検討されていない．本論文は，誤差逆伝搬学習法が持つ基本的な3つの並列性をモデル化し，その学習性能を解析する．また，並列計算機 nCUBE/2 に各並列学習法を実装して学習性能について詳しく比較検討し，並列学習法の有効性について議論する．

2 誤差逆伝搬学習法

2.1 誤差逆伝搬学習法のアルゴリズム

Rumelhart ら [6] は，学習セットにより提供される教師信号とネットワークの出力との 2 乗誤差を最小にするよう，すべての重みを調節する誤差逆伝搬学習を提案した．

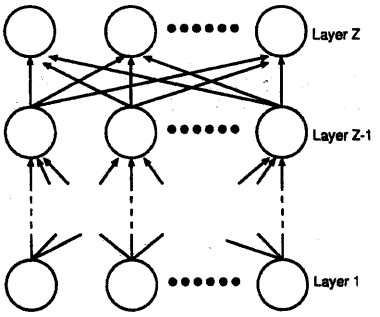

図 1 誤差逆伝搬学習の階層型ネットワーク

誤差逆伝搬学習に用いる階層型ニューラルネットワークを図 1に示す．階層型ニューラルネットワークは Z層からなり，第 1 層を入力層，第 Z層を出力層とする．第 k層の第 n ユニットへの入力の総和を i_n^k，出力を o_n^kとする．各層には N_u個のユニットがあると仮定し，各ユニットの入出力関係を f，$k-1$ 層のユニット m から k層のユニット n への重みを $w_{k-1,m}^{k,n}$ とすると，i_n^kと o_n^kは次式で表される．

$$o_n^k = f(i_n^k) \tag{1}$$

$$i_n^k = \sum_{m=1}^{N_u} w_{k-1,m}^{k,n} o_m^{k-1} \tag{2}$$

ある時刻 tでのネットワーク中の重みを$\boldsymbol{w}(t)$，学習パターンを $(\boldsymbol{x}(t), \boldsymbol{y}(t))$ とし，このときの重みの修正量を$\Delta\boldsymbol{w}(t)$ とすると，時刻 $(t+1)$ の重み$\boldsymbol{w}(t+1)$ は次式で与えられる．

$$\boldsymbol{w}(t+1) = \boldsymbol{w}(t) + \Delta\boldsymbol{w}(t) \tag{3}$$

損失関数 rを出力と教師信号の誤差の 2 乗とすると，ある学習パターン $(\boldsymbol{x}, \boldsymbol{y})$ が与えられたときの rは式 (4) となる．

$$r = \sum_{n=1}^{N_u} (o_n^Z(\boldsymbol{w}, \boldsymbol{x}) - y_n)^2 \tag{4}$$

このとき，修正量$\Delta\boldsymbol{w}(t)$ の要素は次式で求められる．

$$\Delta w_{k-1,m}^{k,n} = -\varepsilon \frac{\partial r}{\partial w_{k-1,m}^{k,n}} \tag{5}$$

ここで，εは学習率と呼ばれる正の定数である．さらに，式 (5) は次の関係を満たす．

$$\frac{\partial r}{\partial w_{k-1,m}^{k,n}} = \frac{\partial r}{\partial i_n^k} \frac{\partial i_n^k}{\partial w_{k-1,m}^{k,n}}$$

$$= \sum_{l=1}^{N_u} \frac{\partial r}{\partial i_l^{k+1}} w_{k,n}^{k+1,l} f'(i_n^k) \quad (6)$$

ここで，f'はfの導関数である．$\partial r/\partial i_n^k = d_n^k$とおくと，重みの修正量$\Delta w_{k-1,m}^{k,n}$は，以下の式で計算される．

$$\Delta w_{k-1,m}^{k,n} = -\varepsilon d_n^k o_m^{k-1} \quad (7)$$

$$d_n^Z = 2(o_n^Z - y_n) f'(i_n^Z) \quad (8)$$

$$d_m^k = \left(\sum_{l=1}^{N_u} w_{k,m}^{k+1,l} d_l^{k+1} \right) f'(i_m^k) \quad (9)$$

fをシグモイド関数であるとすると，その導関数f'は，$f' = f(1-f)$となり，式 (1) から$f'(i_m^k) = o_m^k(1-o_m^k)$となる．実際には振動を減らし，学習の収束をはやめるために，次式に示す修正を行う．

$$\Delta w_{k-1,m}^{k,n}(t+1) = -\varepsilon d_n^k o_m^{k-1} + \alpha \Delta w_{k-1,m}^{k,n}(t) \quad (10)$$

ここで，αはモメンタム項と呼ばれる正の定数である．これを，すべての学習パターンについて繰り返すことにより学習を行う．

3 誤差逆伝搬学習の並列化

誤差逆伝搬学習法にはさまざまな並列化が考えられている [3, 4, 5]．これらの並列学習法は以下に述べる3つの並列学習法としてモデル化できる [3]．

(1) ユニット並列モデル： 式 (1) で示される各ユニットの出力計算において，k, mの双方で並列化を行う手法．

(2) 学習セット並列モデル： 式 (3) で示される重みの更新の線型性を利用し，学習セットを分割して並列化を行う手法．

(3) パス並列モデル： 誤差逆伝搬学習におけるフォワードパス，バックワードパスの2つのフェーズを分割して並列化を行う手法．

本論文では，入力層L個，隠れ層M個，出力層N個のユニットを持つL-M-Nネットワークを用いて並列に誤差逆伝搬学習を行う．入力層-隠れ層間のリンクが持つ重みは隠れユニット，隠れ層-出力層間のリンクが持つ重みは出力ユニットが持つと仮定する．学習時間の解析に必要な処理時間の単位として，加減算1回に必要な時間をt_{add}，乗除算1回に必要な時間をt_{multi}，活性値の計算1回に必要な時間をt_{act}と仮定する．並列学習を並列計算機上で実行する場合，プロセッサ（PE）間の通信が発生する．そこで，2PE間で行れるPE間通信としてt_{comm}を仮定し，これには通信路確保などのセットアップ時間も含める．

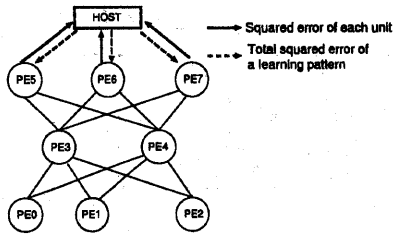

図 2 ユニット並列モデル

3.1 ユニット並列モデル

ニューラルネットワーク内の各ユニットをそれぞれ並列に動作させて学習を行う方法をユニット並列モデルとよぶ．すなわち，層番号kとユニット番号mに関して並列化を行うモデルである．ユニット並列モデルの学習時間解析を行うにあたり，各PEは図2のような結合形態をとると仮定する．すなわち，ニューラルネットワークとPE網のトポロジは同一であり，出力ユニットを担当するPEはホストノードと結合して2乗誤差の計算はホストノードで行う．この条件のもとで，各学習フェーズにおける学習時間を解析した結果を表1に示す．T_{FP}, T_{BP}, T_{UW}はそれぞれフォワードパス，バックワードパス，重み更新に必要な時間である．T_{total}は，N_{PE}個のPEを用いてN_p個の学習パターンをN_{epoch}回学習したときの並列学習時間である．

3.2 学習セット並列モデル

PE上に完全なニューラルネットワークを作成し，学習セットの一部を用いて並列に学習を行うモデルを学習セット並列モデルという．したがって，学習セット並列モデルでは，学習パターン毎ではなく学習セット毎に重みが更新される．

表1に学習セット並列モデルの各フェーズ毎の学習時間を示す．表1より，学習セット並列モデルでのフォワードパス，バックワードパスに要する時間は，PE数N_{PE}を増やせば線型に減少する．しかし，他の処理に比べ大きな値を持つt_{comm}の係数が$2(LM+MN)$であるため，ネットワークの規模が大きくなると十分な高速化が行えない可能性がある．そこで，通信回数を削減する改良学習セット並

表 1 各並列学習モデルの学習フェーズの学習時間

	ユニット並列	学習セット並列	改良学習セット並列	パス並列
T_{FP}	$3t_{act}+(L+M)(t_{add}+t_{multi}+t_{comm})$	$(L+M+N)t_{act}+(LM+MN)(t_{add}+t_{multi})$		
T_{BP}	$(2L+2M+3N+5)t_{add}+(3L+3M+2N+8)t_{multi}+3Nt_{comm}$	$(LM+2MN+M+4N)t_{add}+(3LM+4MN+2M+4N)t_{multi}$		
T_{UW}		$(LM+MN)N_{PE}t_{add}$ $+2(LM+MN)t_{comm}$	$(LM+MN)N_{PE}t_{add}$ $+2(M+N)t_{comm}$	
T_{total}	$N_{epoch}N_p(T_{FP}+T_{BP})$	$N_{epoch}\left\{\frac{N_p}{N_{PE}}(T_{FP}+T_{BP})+T_{UW}\right\}$		$N_{epoch}\left\{\frac{2N_p}{N_{PE}}(\max(T_{FP},T_{BP})$ $+t_{comm})+T_{UW}\right\}$

列モデルを 3.3節で示し，その学習時間について検討する．

3.3 改良学習セット並列モデル

並列計算機における通信時間は，通信路を確保するための通信セットアップ時間，実際にメッセージを転送する転送時間に分けることができる．4.1節で述べるように，転送時間に比べて通信セットアップには長い時間が必要である．通信路の確保は通信が必要になる度に行われるため，通信回数を減らすことにより通信時間の大幅な短縮が期待できる．そこで，複数の重みをまとめて一度に通信し，加算を行う改良学習セットモデルを示し，その性能について議論する．

複数の重みをまとめて通信する配列加算通信法を図3に示す．学習セット並列モデルを用いる場合，図3の例では PE0 が持つ重み $w^{k,n}_{k-1,p}, w^{k,n}_{k-1,q}$，PE1 が持つ重み $w'^{k,n}_{k-1,p}, w'^{k,n}_{k-1,q}$ を集計するために 4 回の通信が必要になる．これに対して配列加算通信法では，各 PE が持つネットワークコピー中で対応するユニットの重みを $(w^{k,n}_{k-1,p}, w^{k,n}_{k-1,q})$ のようにまとめ，一度の通信でホストへ送信する．したがって図3の例では通信回数を 2 回にすることができる．配列加算通信法により，学習セット並列モデルで重み更新を行う際の t_{comm} の係数 $(LM+MN)$ は $(M+N)$ に削減できる．改良学習セット並列モデルでの各学習フェーズにおける所要時間を表 1に示す．

3.4 パス並列モデル

パス並列モデルは，誤差逆伝搬学習におけるフォワードパス，バックワードパスをそれぞれ別の PE で分割して行う手法である．したがって，パス並列モデル単独では大幅な高速化は達成できないため，ここでは改良学習セット並列モデルと組合わせて使

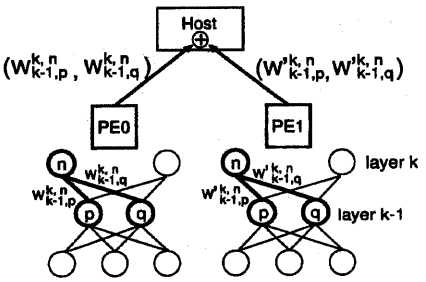

図 3 配列加算通信法

用した．

パス並列モデルでは，2 つの PE でフォワードパス，バックワードパスを実行し，1 つのニューラルネットワークを構成する．このため，より長い処理を行う PE の処理時間が支配的となる．学習セット並列モデルによるフォワードパスの所要時間からバックワードパスでの所要時間を引くと式 (11) となる．

$$(L+M+N)t_{act}-(MN+M+4N)t_{add} -(2LM+3MN+M+4N)t_{multi} \quad (11)$$

ここで，nCUBE/2 での実測値から活性値の計算には乗除算，加減算の 20 倍の時間が必要であるとし，$t=\frac{t_{act}}{20}=t_{add}=t_{multi}$ と仮定する．$L=M=N=n$ の場合，式 (11) は $-nt(6n-49)$ となる．これが正となる条件は $n>0, t>0$ より $0<n<\lceil\frac{49}{6}\rceil=8$ であり，ネットワーク内のユニット数 24 以上ではバックワードパスの処理時間が並列学習を支配する．表 1にパス並列モデルの学習時間を示す．

4 並列学習法の処理性能

4.1 並列計算機 nCUBE/2 の通信時間

並列計算機では通信時間はメッセージ長 M_{len} やプロセッサ数 N_{PE} といったパラメータに依存する．したがって，通信時間 t_{comm} を $t'_{comm} = f(M_{len}, N_{PE})$ として並列学習時間を議論する必要がある．

通信時間は通信セットアップ時間 t_{setup} とメッセージ転送時間 t_{trans} に大きく分けることができる．メッセージ転送時間は，転送されるメッセージの長さに比例して長くなるため，係数を ξ とすると $t_{trans} = \xi M_{len}$ と表すことができる．学習セット並列モデルなどで用いる配列加算通信では全ての PE のデータを集計する．したがって，通信セットアップ時間およびメッセージ転送時間は nCUBE/2 で用いているハイパーキューブ網の直径 $\log N_{PE}$ に比例する．以上より，nCUBE/2 における配列加算通信時間 t'_{comm} は次式で表される．

$$t'_{comm} = (t_{setup} + \xi M_{len}) \log N_{PE} \qquad (12)$$

t_{setup}, ξ の値を決定するため，nCUBE/2 における配列加算通信に必要な時間を測定し，最小二乗法により t_{setup}, ξ を求めたところ，$t_{setup} = 5.0 \times 10^{-4}$(秒)，$\xi = 3.0 \times 10^{-6}$(秒) となった．

4.2 大規模ネットワークの並列学習時間

大規模ネットワークにおける各並列学習モデルの学習時間について検討する．図 4に，メッセージ長や PE 数によるオーバーヘッドを考慮した各並列学習モデルでの学習時間を示す．ユニット並列モデルでは 1 対 1 通信を用いるため，通信時間は 100μ (秒) に固定した．なお，学習セット並列モデル，改良学習セット並列モデルおよびパス並列モデルを用いた学習時間の計算には $N_u = L = M = N = N_{PE}$ とし，各層におけるユニット数と同数の PE を用いるとした．ユニット並列モデルではユニット数が決まると使用する PE 数も決定され，$N_{PE} = 3N_u$ となる．また，評価に必要な計算要素の処理時間は，並列計算機 nCUBE/2 での実測値から $t_{add} = t_{multi} = 1\mu$(秒)，$t_{act} = 20\mu$(秒) とし，学習パターン数 $N_p = 10000$，最大学習回数 $N_{epoch} = 10000$ とした．

図 4より，ユニット並列モデルではユニット数の増加に対して線型に学習時間が増加している．これに対し，ユニット数が少ない領域では学習セット並列モデルはユニット並列モデルよりも高速である．

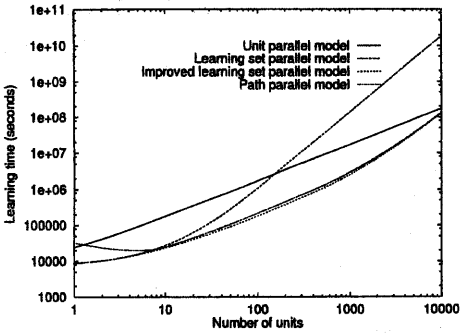

図 4 オーバーヘッドを考慮した各並列学習モデルの学習時間

しかし，ユニット数が多くなると通信回数が多いため学習時間の大幅な増加が見られる．通信回数を減らした改良学習セット並列モデルでは，ユニット並列モデルと比較しておよそ 10 倍の学習速度を達成できていることが分かる．パス並列モデルで，フォワードパスの処理時間が支配的である領域ではユニット数（ここでは PE 数に等しい）の増加に対して学習時間が減少する．これは，PE 数が少ない領域では通信時間よりも学習処理での時間が長く，学習セットを分割することによる高速化の効果が大きいためと考えられる．しかし，ユニット数（PE 数）がさらに増加すると，学習処理自体は短時間となるものの通信時間が長くなり，全体として処理時間が増大していく．このときの学習時間の変化は改良学習セット並列モデルとほぼ同等の傾向を示すが，ネットワークのコピー数が減少するために改良学習セット並列モデルよりも学習時間が長くなる．

学習セット並列モデルでは学習パターン数以上の PE は用いることができず，またユニット並列モデルでは PE 数による通信セットアップ時間の増大を考慮していないので，PE 数やニューラルネットワークの構成が変化しても学習時間の逆転は起こらないと考えられる．

4.3 PE 数の増加による並列学習時間の変化

学習セット並列モデル，改良学習セット並列モデルおよびパス並列モデルでは，ユニット数とは独立に PE 数を決めることができる．そこで，ユニット数が 1000-1000-1000 である大規模ネットワークで $N_p = 10000, N_{epoch} = 10000$ とし，最大 10000 個の PE で学習を行ったときの並列学習時間について検討する．図 5 に学習セット並列，改良学習セット

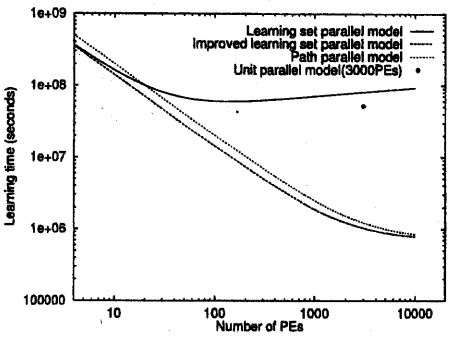

図 5 1000-1000-1000 のネットワークにおける各並列学習モデルの学習時間

並列およびパス並列各学習モデルでの学習時間を示す．ユニット並列モデルではユニット数が決まると用いるPE数も決まるため，3000PE時のデータのみとなる．

図5より，学習セット並列モデルではPEを200程度用いたときが最も高速であり，それ以上PEを増やすと通信時間の増大により学習時間が増加する．これに対し，改良学習セット並列モデルおよびパス並列モデルではPE数が多くなると並列学習による高速化率が鈍化しはじめるが，1PE=1学習パターンとなるまで学習を高速化できる．3000個のPEを用いたユニット並列モデルと比較しても，同程度のPE数では改良学習セット並列モデルとパス並列モデルの方が高速である．改良学習セット並列モデルとパス並列モデルの比較では，改良学習セット並列モデルの方が高速であり，PE数が多くなると差が小さくなる．以上の結果から，誤差逆伝搬学習を並列化する場合は通信回数を削減した改良学習セット並列モデルが最も有効であると言える．

5 むすび

本論文では，誤差逆伝搬学習法が内包する並列性に着目した3種類の並列学習法のモデル化を行い，並列学習時間について詳細に検討した．その結果，ニューラルネットワークの各ユニットをPEに割り当てるユニット並列モデルでは，PE間通信のオーバーヘッドが大きく高速化が難しいことを明らかにした．本論文で検討したモデルではニューラルネットワークのトポロジーと並列計算機のプロセッサ網が同一であると仮定している．しかしながら，こうした並列計算機は実現が困難であり，実際の並列計算機を用いたユニット並列モデルではメッセージの中継時間などがさらに加わる．したがって，ユニット並列モデルでの学習時間はさらに長くなることが予想される．

学習セット並列モデルは，PE数を増加させるとPE間通信が増えて学習速度が大幅に低下するが，配列加算通信法を用いて通信回数を減らすことにより3つの並列学習モデルの中では最も高速なモデルであることを示した．パス並列モデルでは，改良学習セット並列モデルと組合わせることで高速化を行った．その結果，パスの分割による高速化よりも，見かけのPE数の減少の影響が大きく，改良学習セット並列モデル単独よりも学習時間は長くなるという結果となった．

今後の課題として，プロセッサ網の形態を考慮したユニット並列モデルの学習時間モデルの構築および解析が考えられる．

謝辞: 本研究の一部は文部省科学研究費助成により行われた．関係各位に感謝する．

参考文献

[1] S. Fahlman, "An empirical study of learning speed in back-propagation network", CMU-CS-88-162, CMU (1988).

[2] 塚本義明, 生天目章, "多層ネットワークの瞬時学習法", 情処学論, Vol. 34, No. 9, pp. 1882 – 1891 (1993).

[3] A. Singer, "Implementation of artificial neural network on the connection machine", RL90-2, Thinking Machine Corp. (1990).

[4] U. A. Muller, B. Baumle, P. Kohler, A. Gunzinger and W. Guggenbuhl, "Achieving supercomputer performance for neural net simulation with an array of digital signal processors", *IEEE Micro Magazine*, Vol. 12, No. 5, pp. 55–65 (1992).

[5] M. Witbrock and M. Zagha, "An implementation of backpropagation learning on GF11, a large simd parallel computer", *Parallel Computing*, Vol. 14, No. 3, pp. 329–346 (1990).

[6] D. E. Rumelhart, J. L. McClelland and PDP Research Group, Parallel Distributed Processing, Vol. 1, MIT Press (1986).