

(51)Int.Cl. ⁵	識別記号	庁内整理番号	F I	技術表示箇所
G 0 6 G 7/60		7368-5B		
G 0 6 F 15/18		8945-5L		

審査請求 未請求 請求項の数2 (全 5 頁)

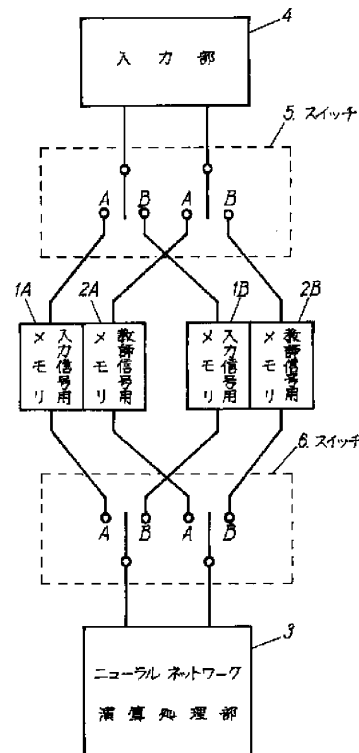
(21)出願番号	特願平3-5690	(71)出願人	000005821 松下電器産業株式会社 大阪府門真市大字門真1006番地
(22)出願日	平成3年(1991)1月22日	(72)発明者	田村 洋一 大阪府門真市大字門真1006番地 松下電器産業株式会社内
		(72)発明者	森下 賢幸 大阪府門真市大字門真1006番地 松下電器産業株式会社内
		(74)代理人	弁理士 小鍛冶 明 (外2名)

(54)【発明の名称】 ニューロプロセッサ

(57)【要約】

【目的】 階層構造ニューラルネットワークの前向き伝搬、及び学習の処理が高速に行なえるニューロプロセッサを提供する。

【構成】 入力信号用メモリと教師信号用メモリの組を2組備える。そして、入力信号用メモリ1A上の入力信号データと教師信号用メモリ2A上の教師信号データを用いて計算を行っている間に、それと並列して、次の計算で使用する入力信号データと教師信号データを入力し、入力信号用メモリ1Bと教師信号用メモリ2Bに書き込む。次のステップでは、入力信号用メモリ1Bと教師信号用メモリ2B上のデータを用いて計算を行っている間に、次のデータを、入力信号用メモリ1Aと教師信号用メモリ2Aに書き込む。その結果、データの入力作業が、計算と並列して行われるので、処理時間が短縮される。



【特許請求の範囲】

【請求項1】ニューラルネットワーク演算処理部と、2つの入力信号用メモリを備え、一方の入力信号用メモリ上の入力信号データを用いて、前向き伝搬の計算を行っている間に、それと並列して、次の前向き伝搬の計算で使用する入力信号データを入力し、それをもう一方の入力信号用メモリに書き込むことを行うニューロプロセッサ。

【請求項2】ニューラルネットワーク演算処理部と、入力信号用メモリと教師信号用メモリの組を2組備え、一方の組の入力信号用メモリ上の入力信号データと教師信号用メモリ上の教師信号データを用いて、学習の計算を行っている間に、それと並列して、次の学習の計算で使用する入力信号データと教師信号データを入力し、それをもう一方の組の入力信号用メモリと教師信号用メモリに書き込むことを行うニューロプロセッサ。

【発明の詳細な説明】

【0001】

【産業上の利用分野】本発明は、階層構造ニューラルネットワークの前向き伝搬、及び学習の処理を高速に行なうためのニューロプロセッサに関するものである。

【0002】

【従来の技術】ニューラルネットワークは、人間の脳の神経細胞の働きをモデル化して模倣することによって、従来のいわゆるノイマン形のコンピュータが苦手としていた認識や、連想、最適化問題、音声合成等を得意とする新しいコンピュータを実現しようとするものである。

【0003】ニューラルネットワークには、ニューロンが層状に配置された階層構造のものや、すべてのニューロンが相互に結合した相互結合構造のものなど、さまざまな構造のものがある。その中で階層構造のネットワークは、例えばバックプロパゲーションアルゴリズムと呼ばれる学習アルゴリズムで簡単に学習させることができ、制御、文字認識、画像認識、画像処理などに幅広く応用することができると考えられている。

【0004】図4は階層構造のネットワークの例を示したものである。図4で、101は各層中に配置されたニューロン、102はシナプスと呼ばれるニューロン間の結合、103は入力層、104は中間層、105は出力層を示す。図4では3層のネットワークの例を示しているが、中間層を複数にすることによって4層以上の階層構造のネットワークにすることもできる。以後は3層のネットワークに限って説明するが、4層以上のネットワークに対してもまったく同様の扱いである。また、図4では各層のニューロンの数は、それぞれ3、2、3として例示したものであるが、各層のニューロンの数を増減させても同様である。ニューラルネットワークへの入力信号データは、おのおの、入力層103の各ニューロンに与えられる。そして、信号が入力層、中間層、出力層の順番に伝搬していき、出力層105のニューロンの信

号がネットワークの出力になる。このような入力層、中間層、出力層の順番に伝搬していく通常の伝搬を前向き伝搬と呼ぶ。

【0005】図5はニューロンの働きを示した図である。図5で、101、107、108、109はニューロン、102はシナプス、106はニューロンの特性関数fである。ニューロンは1つ前の層のニューロンの出力をシナプスを介して受け取る。それぞれのシナプスは結合の重みと呼ばれる値を持っており、前の層のニューロンの出力値にその結合の重みの値を乗算した結果を次のニューロンに与える。シナプスの結合の重みは、それぞれのシナプスで異なった値になっている。例えば、図5の1層目のi番目のニューロン108と1+1層目のj番目のニューロン101との間のシナプスは結合の重みw_{ij}を持っており、1層目のi番目のニューロン108の出力O_iにその結合の重みの値を乗算した結果w_{ij}×O_iが1+1層目のj番目のニューロン101に与える。そして、ニューロン101はそれに結合するすべてのシナプスより与えられる入力をすべて加算し、その加算結果にニューロンの特性関数106を作用させて、その関数値をニューロン101の出力O_{1j}として出力する。これを式で表すと次式ようになる。

【0006】

【数1】

$$O_{1+1,j} = f \left(\sum_i w_{ij} \times O_{1i} \right)$$

【0007】したがって、3層のネットワークの場合の前向き伝搬は、つぎの順序で計算される。まず最初にすべての中間層のニューロンについて、その出力を、

【0008】

【数2】

$$O_{2,j} = f \left(\sum_i w_{ij} \times O_{1i} \right)$$

【0009】に従って計算する。入力層のニューロンの出力O_iは、そのニューロンに与えられた入力信号データである。次にその結果を使ってすべての出力層のニューロンについて、その出力を、

【0010】

【数3】

$$O_{3,j} = f \left(\sum_i w_{2ji} \times O_{2i} \right)$$

【0011】に従って計算する。次に、図4の3層の階層構造ネットワークのバックプロパゲーションアルゴリズムによる学習の方法について説明する。バックプロパゲーションアルゴリズムでは、入力とそれに対する理想的な出力の組を用意し、その入力に対する実際の出力と理想的な出力の差が減少するようにシナプスの結合の重みを修正する。この理想的な出力のことを通常は教師信号と呼ぶ。以下に、3層のネットワークについて具体的な計算方法を順を追って説明する。

3

4

【0012】1. 前向き伝搬によって実際の出力を計算する。2. 次式に従って出力層の各ニューロンの誤差に対応した値(以下、単にデルタと呼ぶ)を計算する。

【0013】

【数4】

$$\delta_{3j} = (t_j - O_{3j}) g(O_{3j})$$

*

$$g(O_{3j}) = g(f(\sum_i W_{2ji} \times O_{2i})) = f'(\sum_i W_{2ji} \times O_{2i})$$

【0016】で表されるようにニューロンの特性関数fの微分係数である。ニューロンの特性関数fは、単調非減少の関数を用いられるので、数5に示したように、特性関数の微分係数を特性関数の関数値の関数として表すことができる。

【0017】3. 中間層と出力層との間のシナプスの結合の重みの修正量を、次式に従って計算し、重みを修正する。

【0018】

【数6】

$$\Delta W_{2ji} = \eta \times \delta_{3j} \times O_{2i}$$

【0019】ηは修正係数である。4. 中間層のニューロンに対するデルタを、次式に従って計算する。

【0020】

【数7】

$$\delta_{2j} = (\sum_k \delta_{3k} \times W_{2kj}) \times g(O_{2j})$$

【0021】5. 中間層と入力層の間のシナプスの結合の重みの修正量を、次式に従って計算し、重みを修正する。

【0022】

【数8】

$$\Delta W_{1ji} = \eta \times \delta_{2j} \times O_{1i}$$

【0023】実際の学習では、以上に述べた手順を、複数個の入力と教師信号の組に対して行い、さらにそれを何回も繰り返す。すなわち、上の1から5までの手順を1回の学習と呼ぶことにすると、総学習回数は、(学習に用いる入力と教師信号の組の数) × (繰り返しの回数)となる。

【0024】以上が、階層構造のニューラルネットワークの前向き伝搬、および学習の計算の手順である。しかし、実際の処理では、計算を開始する前に、前向き伝搬の計算では、入力信号データが与えられなければならない。また、学習の計算では、入力信号データと教師信号データが与えられなければならない。したがって、従来、前向き伝搬や学習の処理を行う時は、次のような手順がとられている。

【0025】まず、最初に、ニューロプロセッサは、入力信号データや教師信号データを外部より受取り、それ

*【0014】δ は1層目のj番目のニューロンに対するデルタ、t は出力層のj番目のニューロンに対する教師信号、gは

【0015】

【数5】

らのデータをいったんメモリに書き込む。入力信号データや教師信号データを外部より受取り、それらのデータをメモリに書き込む作業を、以後、データの入力と呼ぶことにする。次に、そのメモリ上の入力信号データや教師信号データを用いて、実際の計算を行う。このような、2段階の処理の方法が、通常はとられている。図3の(イ)は、この従来法の処理の流れを示した図である。まず最初に、データの入力を行い、その後で、計算を行っている。通常は、違ったデータについての処理を連続して行うので、図3の(イ)に示したように、計算の後に、さらに次のデータの入力が行われる。この操作を何回も繰り返す。

【0026】

【発明が解決しようとする課題】一般に、ニューラルネットワークの計算は、計算量が非常に多いので、多くの処理時間を必要とする。特に、学習は何回も繰り返すことが必要なので、処理時間は膨大になる。したがって、処理が高速に行えるニューロプロセッサが望まれている。

【0027】しかし、従来のニューロプロセッサのように、データの入力を行い、次に計算を行う方法だと、1回の処理に、(データの入力に必要な時間) + (計算に必要な時間)がかかる。すなわち、実際に計算に要する時間の他に、データの入力にかかる時間が余分に必要になるわけである。

【0028】本発明は上記課題を解決するもので、階層構造ニューラルネットワークの前向き伝搬、及び学習の処理が高速にできるニューロプロセッサを提供することを目的としている。

【0029】

【課題を解決するための手段】以上の課題を解決するため、本発明のニューロプロセッサでは、入力信号用メモリと教師信号用メモリの組を2組備え、一方の組の入力信号用メモリ上の入力信号データと教師信号用メモリ上の教師信号データを用いて、計算を行っている間に、それと並列に、次の計算で使用する入力信号データと教師信号データを入力し、もう一方の組の入力信号用メモリと教師信号用メモリに書き込むことを行う。

【0030】

【作用】上記構成により、データの入力の作業が、計算と並列して行われるので、データの入力のための時間

が、見かけ上不要になる。すなわち、従来、データの入力に要していた時間の分だけ、全体の処理時間が短縮される。したがって、処理の高速化が図れる。

【0031】

【実施例】以下、本発明の実施例を図面により説明する。

【0032】図1は、本発明のニューロプロセッサの構成図である。図1の1Aと1Bは入力信号用メモリ、2Aと2Bは教師信号用メモリで、1Aと2Aが、1つの組となり、1Bと2Bが、もう1つの組となっている。以後、1Aと2Aの組をメモリA、1Bと2Bの組をメモリBと呼ぶことにする。図1の3は、ニューラルネットワーク演算処理部で、前向き伝搬や学習の計算が行われる。図1の4は、入力部で、入力信号データと教師信号データは、この入力部を通して入力される。図1の5と6はスイッチで、スイッチ5は、入力部4より入力されたデータを、メモリAに書き込むか、メモリBに書き込むかを切り替える。スイッチ6は、ニューラルネットワーク演算処理部3で使用するデータを、メモリAから読み出すか、メモリBから読み出すかを切り替える。

10

20

【0033】図2は、本実施例におけるスイッチとメモリの使用法を示した図である。以下、図2に従って、本実施例のニューロプロセッサの動作について説明する。

【0034】まず、最初のステップaでは、スイッチ5をAに切り替えることによって、入力部4より入力されたデータを、メモリAに書き込む。次のステップbでは、スイッチ6をAに切り替えて、ステップaで書き込んだメモリAのデータを読み出して、ニューラルネットワーク演算処理部3が計算を行う。それと同時に、スイッチ5をBに切り替えて、入力部4より入力されたデータを、メモリBに書き込む。次のステップcでは、スイッチ6をBに切り替えて、ステップbで書き込んだメモリBのデータを読み出して、ニューラルネットワーク演算処理部3が計算を行う。それと同時に、スイッチ5をAに切り替えて、入力部4より入力されたデータを、メモリAに書き込む。以後は、このような操作を繰り返す。

30

【0035】図3の(ア)は、この処理の流れを示した図である。データの入力と計算が並列に行われ、1つ前のステップで入力されたデータを使って、計算は行われ、図3の(イ)の従来法の場合と比べると明らかに、図3の(イ)の従来法では2ステップで1つの処理が終わるのに対して、本発明の方法では、見かけ上、1ステップで1つの処理が終わる。

40

【0036】以上の動作の説明よりわかるように、本ニューロプロセッサでは、入力信号用メモリと教師信号用メモリの組を2組備え、一方の組の入力信号用メモリ上の入力信号データと教師信号用メモリ上の教師信号データを用いて、計算を行っている間に、それと並列して、次の計算で使用する入力信号データと教師信号データを入力し、もう一方の組の入力信号用メモリと教師信号用メモリに書き込むことを行う。その結果、データの入力の作業が、計算と並列して行われるので、データの入力のための時間が、見かけ上不要になる。すなわち、従来、データの入力に要していた時間の分だけ、全体の処理時間が短縮される。

【0037】なお、本発明は、ニューラルネットワーク演算処理部の構造には、無関係であり、どのようなニューラルネットワーク演算処理部に対しても、効果がある。

【0038】

【発明の効果】以上の実施例から明らかなように、本発明によれば、入力信号用メモリと教師信号用メモリの組を2組備え、データの入力と計算を並列して行うことによって、データの入力のための時間が、見かけ上不要になり、階層構造のニューラルネットワークの前向き伝搬、及び学習の処理が高速に行えるニューロプロセッサを提供できる。

【図面の簡単な説明】

【図1】本発明の一実施例のニューロプロセッサの構成図

【図2】スイッチとメモリの使用方法を示した図

【図3】処理の流れを示した図

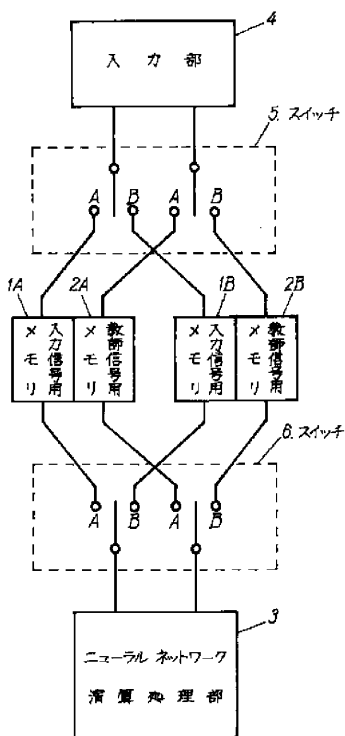
【図4】階層構造ニューラルネットワークの図

【図5】ニューロンの働きを示した図

【符号の説明】

- 1 A, 1 B 入力信号用メモリ
- 2 A, 2 B 教師信号用メモリ
- 3 ニューラルネットワーク演算処理部
- 4 入力部
- 5, 6 スイッチ
- 101 ニューロン
- 102 シナプス
- 103 入力層
- 104 中間層
- 105 出力層
- 106 ニューロンの特性関数
- 107-109 ニューロン

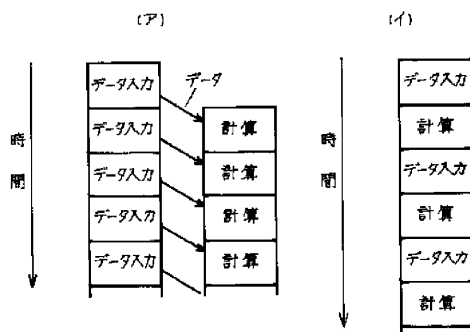
【図1】



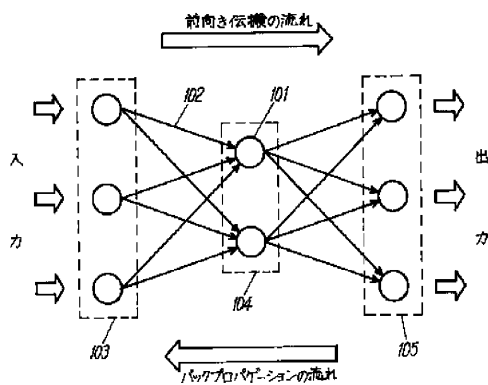
【図2】

ステップ	スイッチ5	入力メモリ	データメモリ	スイッチ6	計算に使用するデータのメモリ
a	A	A			
b	B	B	A	A	
c	A	A	B	B	
d	B	B	A	A	
e	A	A	B	B	
f	B	B	A	A	

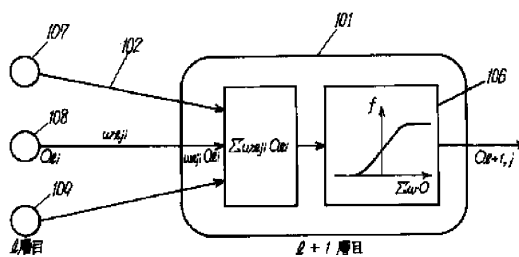
【図3】



【図4】



【図5】



(51) Int.Cl. ⁵	ID Codes	Internal File Nos.	FI	Tech. Indicators
G06G	7/60	7368-5B		
G06F	15/18	8945-5L		

Examination Request Not Yet Received No. of Claims 2 (Total of 5 Pages)

(21) Application No. H03-5690

(22) Filing Date January 22, 1991

(71) Applicant 000005821

Matsushita Electrical Industry Co., Ltd.
1006, Kadoma, Kadoma-shi, Osaka

(72) Inventor

Yoichi TAMURA
Matsushita Electrical Industry Co., Ltd.
1006, Kadoma, Kadoma-shi, Osaka

(72) Inventor

Tadayuki MORISHITA
Matsushita Electrical Industry Co., Ltd.
1006, Kadoma, Kadoma-shi, Osaka

(74) Agent

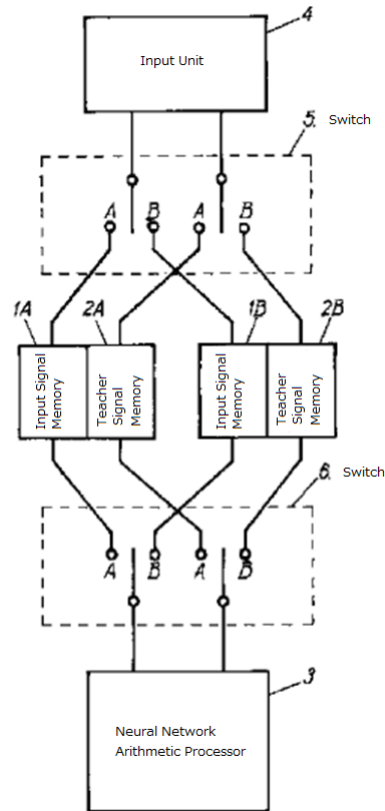
Fumio IWAHASHI, Patent Attorney (and 2 others)

(54) [Title of the Invention] Neuroprocessor

(57) [Abstract]

[Problem] To provide a neuroprocessor that can perform forward propagation and learning processing in a hierarchical neural network at high speed.

[Solution] Pairs of input signal memories and teacher signal memories are provided. While calculations are performed using input signal data in input signal memory 1A and teacher signal data in teacher signal memory 2A, input signal data and teacher signal data to be used in the calculations of the next step are inputted and written to input signal memory 1B and teacher signal memory 2B in parallel. In the next step, calculations are performed using the data in input signal memory 1B and teacher signal memory 2B, and data for the next step is written to input signal memory 1A and teacher signal memory 2A. As a result, data input processing is performed in parallel with calculations, thereby reducing processing time.



[Claims]

[Claim 1]

A neuroprocessor comprising a neural network arithmetic processing unit and two input signal memories, wherein the input signal data used in the next forward propagation calculation is inputted in parallel while a forward propagation calculation is being performed using input signal data in one input signal memory, and written to the other input signal memory.

10

[Claim 2]

A neuroprocessor comprising a neural network arithmetic processing unit and pairs of input signal memories and teacher signal memories, wherein input signal data and teacher signal data to be used in the next calculations are inputted and written to the other of the pair of input signal memories and the other of the pair of teacher signal memories while learning calculations are being performed using input signal data in one of the pair of input signal memories and teacher signal data in one of the pair of teacher signal memories.

20

[Detailed Description of the Invention]

[0001]

[Technical Field of the Invention]

The present invention relates to a neuroprocessor for performing forward propagation and learning processing in a hierarchical neural network at high speed.

30

[0002]

[Prior Art]

By modeling and imitating the workings of neurons in the human brain, neural networks aim to create a new type of computer that excels at tasks such as recognition, association, optimization, and speech synthesis, which are difficult for conventional von Neumann-type computers.

40

[0003]

There are various types of neural network structure, including hierarchical structures with neurons arranged in layers and fully connected structures with all neurons connected to each other. Hierarchical networks can be easily trained using a learning algorithm called a back-propagation algorithm, and are believed to have a wide range of applications in areas such as control, character recognition, image recognition, and image processing.

50

[0004]

FIG. 4 shows an example of a hierarchical network. In FIG. 4, 101 denotes the neurons arranged in each layer, 102 denotes the

connections between neurons known as synapses, 103 denotes the input layer, 104 denotes an intermediate layer, and 105 denotes the output layer. FIG. 4 shows an example of a three-layer network, but a hierarchical network can also be created with four or more layers simply by adding more intermediate layers. A three-layer network will be explained below, but the same explanation applies to networks with four or more layers. In FIG. 4, the number of neurons in the layers are three, two, and three, but the same explanation applies whether the number of neurons in each layer is increased or decreased. Input signal data for the neural network is applied to each neuron in the input layer 103. The signals then propagate through the input layer, the intermediate layer, and the output layer in that order, and the neuron signals of the output layer 105 are the network output. Normal propagation that proceeds through the input layer, the intermediate layer, and the output layer in that order is known as forward propagation.

[0005]

FIG. 5 is a diagram showing how neurons work. In FIG. 5, 101, 107, 108, and 109 are neurons, 102 is a synapse, and 106 is the characteristic function f of the neuron. Neurons receive the output of neurons in the preceding layer via synapses. Each synapse has a value called a connection weight, and the result of multiplying the output value of a neuron in the preceding layer by the connection weight value is applied to the next neuron. The weighting of the synaptic connection is different for each synapse. For example, the synapse between the i-th neuron in the l-th layer (108) and the j-th neuron in the l+1-th layer (101) in FIG. 5 has a connection weight of w_{lji}, and the result of multiplying the value of that connection weight by the output O_{li} of the i-th neuron in the l-th layer (108) applies w_{lji} × O_{li} to the j-th neuron in the l+1-th layer (101). Neuron 101 then adds up all the inputs provided by all the synapses connected to it, applies the characteristic function 106 of the neuron to the sum, and outputs the value of the function as the output O_{l+1, j} of neuron 101. This can be expressed as the following equation.

[0006]

[Equation 1]

$$O_{l+1, j} = f \left(\sum_i w_{lji} \times O_{li} \right)$$

[0007]

Therefore, in the case of a three-layer network, the forward propagation is calculated in the following steps. First, for all the neurons in the intermediate layer, the output is calculated according to the following equation.

[0008]
[Equation 2]

$$O_{2,j} = f \left(\sum_i W_{1ji} \times O_{1i} \right)$$

10 [0009]

The output O_{1i} of the neurons in the input layer are the input signal data applied to that neuron. Next, using the result, the output of all the neurons in the output layer are calculated according to the following equation.

[0010]
[Equation 3]

$$O_{3,j} = f \left(\sum_i W_{2ji} \times O_{2i} \right)$$

20 [0011]

Next, how learning occurs using a back-propagation algorithm in the three-layer hierarchical network shown in FIG. 4 will be explained. In a back-propagation algorithm, a pair consisting of an input and an ideal output for this is prepared, and the weighting of the synaptic connections are modified so that the difference between the actual output for the input and the ideal output is reduced. This ideal output is usually referred to as a teacher signal. The following is a step-by-step explanation of the specific calculation method for a three-layer network.

[0012]

1. The actual output by forward propagation is calculated.

40 2. The value corresponding to the error of each neuron in the output layer (the "delta") is calculated according to the following equation.

[0013]
[Equation 4]

$$\delta_{3j} = (t_j - O_{3j}) g'(O_{3j})$$

[0014]
Here, δ_{lj} is the delta for the j -th neuron in the l -th layer, t_j is the teacher signal for the j -th neuron in the output layer, and g' is the derivative of the neuron characteristic function f , as expressed by the following equation.

50

[0015]

[Equation 5]

$$g'(O_{3j}) = g' \left(f \left(\sum_i W_{2ji} \times O_{2i} \right) \right) = f'' \left(\sum_i W_{2ji} \times O_{2i} \right)$$

[0016]

Because the characteristic function f of a neuron uses a monotonically non-decreasing function, as shown in Equation 5, the differential coefficient of the characteristic function can be expressed as a function of the function value of the characteristic function.

[0017]

3. The amount of modification to the weighting of the synaptic connection between the intermediate layer and the output layer is calculated according to the following equation and the weighting is modified.

[0018]
[Equation 6]

$$\Delta W_{2ji} = \eta \times \delta_{3j} \times O_{2i}$$

[0019]

Here, η is the modification coefficient.

4. The delta for the neurons in the intermediate layer is calculated according to the following equation.

[0020]
[Equation 7]

$$\delta_{2j} = \left(\sum_k \delta_{3k} \times W_{2kj} \right) \times g'(O_{2j})$$

[0021]

5. The amount of modification to the weighting of the synaptic connection between the intermediate layer and the input layer is calculated according to the following equation and the weighting is modified.

[0022]
[Equation 8]

$$\Delta W_{1ji} = \eta \times \delta_{2j} \times O_{1i}$$

[0023]

In actual learning, this procedure is performed on multiple pairs of input signals and teacher signals, and this is repeated many times. In other words, if steps 1 to 5 form a single learning session, the total number of learning sessions is (number of input signal/teacher signal pairs used for learning) \times (number of repetitions).

[0024]

The calculation procedures for forward propagation and learning in a hierarchical neural network were explained above. However, before calculations begin in actual processing, input signal data must be applied for forward propagation calculations. Also, input signal data and teacher signal data must be applied in the learning calculations. Therefore, the following procedure has been performed in the past when performing forward propagation and learning processing.

[0025]

First, the neuroprocessor receives input signal data and teacher signal data from an external source, and writes the data to memory. The process of receiving input signal data and teacher signal data from an external source and writing that data to memory is hereafter referred to as data input. Next, actual calculations are performed using the input signal data and teacher signal data stored in memory. This two-step processing method is common. FIG. 3 (A) shows the processing flow in the method of the prior art. First, data is inputted, and then calculations are performed. Normally, processing of different sets of data is carried out in succession, so as shown in FIG. 3 (A), the next data is inputted after the calculations have been performed. This process is repeated many times.

[0026]

[Problem to Be Solved by the Invention]

In general, neural network calculations require a lot of processing time because they involve a very large amount of computation. Because learning in particular requires repetition, the processing time becomes enormous. Therefore, a neuroprocessor that can process data quickly is desired.

[0027]

However, in neuroprocessors of the prior art, which input data and then perform calculations, each processing step takes (the time required for data input) + (the time required for calculations). In other words, extra time is needed for data input in addition to the time actually required for calculations.

[0028]

It is an object of the present invention to solve this problem by providing a neuroprocessor that can perform forward propagation and learning processing in a hierarchical neural network at high speed.

[0029]

[Means for Solving the Problem]

In order to solve this problem, the present invention is a neuroprocessor comprising a neural network arithmetic processing unit and pairs of input signal memories and teacher signal memories, wherein input signal data and teacher signal data to be used in next calculations are inputted and written to the other of the pair of input signal memories and the other of the pair of teacher signal memories while learning calculations are being performed using input signal data in one of the pair of input signal memories and teacher signal data in one of the pair of teacher signal memories.

[0030]

[Operation]

In this configuration, the data input operation is carried out in parallel with calculations, so time for data input is apparently not necessary. In other words, the total processing time is reduced by the amount of time that was previously required for data input. Therefore, processing can be accelerated.

[0031]

[Example]

The present invention will now be described in greater detail using an example.

[0032]

FIG. 1 is a configuration diagram of the neuroprocessor in the present invention. In FIG. 1, 1A and 1B are input signal memories, and 2A and 2B are teacher signal memories, 1A and 2A form a pair, and 1B and 2B form another pair. In the following explanation, pair 1A and 2A is referred to as memory A, and pair 1B and 2B is referred to as memory B. In FIG. 1, 3 denotes the neural network arithmetic processing unit, in which forward propagation and learning calculations are performed. In FIG. 1, 4 denotes the input unit, and input signal data and teacher signal data are inputted via this input unit. In FIG. 1, 5 and 6 are switches, and switch 5 switches between writing data input from input unit 4 to memory A or memory B. Switch 6 switches between reading data from memory A and memory B for use in the neural network arithmetic processing unit 3.

[0033]

FIG. 2 is a diagram showing how the switches and memories are used in this example. The following is an explanation of the operation of the neuroprocessor in this example, with reference to FIG. 2.

[0034]

First, in step a, data input from the input unit 4 is written to memory A by switching the switch 5 to A. Next, in step b, switch 6 is switched to A, and the data in memory A that was written in step a is read-out, and the neural network arithmetic processing unit 3 performs calculations on that data. At the same time, switch 5 is switched to B, and data input from input unit 4 is written to memory B. Next, in step c, switch 6 is switched to B, and the data in memory B that was written in step b is read out, and the neural network arithmetic processing unit 3 performs calculations on that data. At the same time, switch 5 is switched to A, and the data input from input unit 4 is written to memory A. These operations are then repeated.

10

20

[0035]

FIG. 3 (A) is a diagram showing processing flow. Data input and calculations are performed in parallel, and the calculations are performed using data inputted in the previous step. As can be seen in comparison with the prior art method in FIG. 3 (B), one processing iteration is completed in two steps in the prior art method in FIG. 3 (B), whereas in the method of the present invention, one processing iteration is apparently completed in one step.

30

[0036]

It is clear from the explanation of operations above that the neuroprocessor of the present invention comprises a neural network arithmetic processing unit and pairs of input signal memories and teacher signal memories, in which input signal data and teacher signal data to be used in the next calculations are inputted and written to the other of the pair of input signal memories and the other of the pair of teacher signal memories, while learning calculations are being performed using input signal data in one of the pair of input signal memories and teacher signal data in one of the pair of teacher signal memories. As a result, the data input processing is carried out in parallel with the calculations, so the time required for data input is apparently unnecessary. In other words, the overall processing time is reduced by the amount of time that was previously required for data input.

40

50

[0037]

Note that the present invention is unrelated to the structure of the neural network arithmetic processing unit, and can be applied effectively to any neural network arithmetic processing unit.

[0038]

[Effect of the Invention]

As is clear from the example, the present invention has two pairs of input signal memories and teacher signal memories, and so a neuroprocessor that can perform forward propagation and learning in a hierarchical neural network at high speed can be provided by performing data input and calculations in parallel, such that time required for data input is apparently no longer necessary.

[Brief Description of the Drawings]

[FIG. 1]

FIG. 1 is a configuration diagram of the neuroprocessor in an example of the present invention.

[FIG. 2]

FIG. 2 is a chart showing how the switches and memories are used.

[FIG. 3]

FIG. 3 is a diagram of the processing flow.

[FIG. 4]

FIG. 4 is a diagram of the hierarchical neural network.

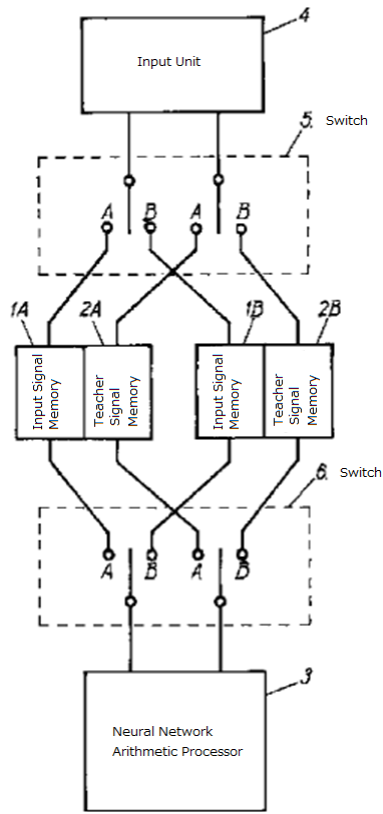
[FIG. 5]

FIG. 5 is a diagram showing how neurons work.

[Reference Numbers]

1A, 1B: Input signal memory
 2A, 2B: Teacher signal memory
 3: Neural network arithmetic processing unit
 4: Input unit
 5, 6: Switch
 101: Neuron
 102: Synapse
 103: Input layer
 104: Intermediate layer
 105: Output layer
 106: Neuron characteristic functions
 107-109: Neurons

[FIG. 1]

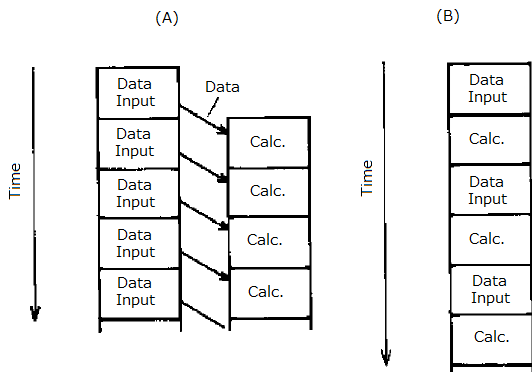


[FIG. 2]

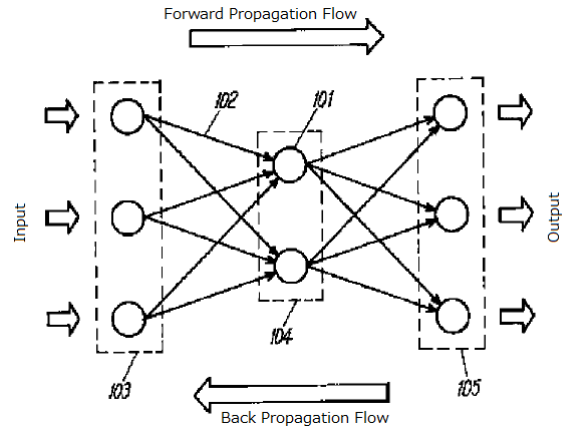
Step	Switch 5	Data Input Memory	Switch 6	Data Memory For Calculations
a	A	A		
b	B	B	A	A
c	A	A	B	B
d	B	B	A	A
e	A	A	B	B
f	B	B	A	A

Time ↓

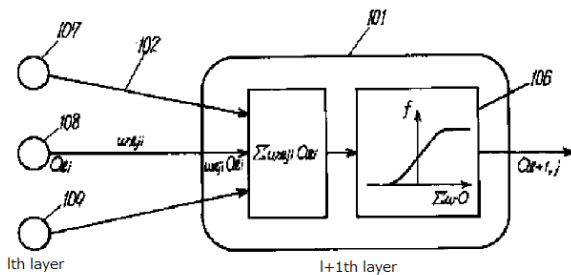
[FIG. 3]



[FIG. 4]



[FIG. 5]



TRANSLATION CERTIFICATION

Date: January 27, 2025

To whom it may concern:

This is to certify that the attached translation is an accurate representation of the documents received by this office. The translation was completed from:

- Japanese

To:

- English (USA)

The documents are designated as:

- Tamura JPH04237388A

Emily Paras, Project Manager in this company, attests to the following:

“To the best of my knowledge, the aforementioned documents are a true, full and accurate translation of the specified documents.”

A handwritten signature in black ink, appearing to read "Emily Paras". The signature is written in a cursive style with a large initial "E".

Signature of Emily Paras