

(19)日本国特許庁(JP)

## (12)特許公報(B2)

(11)特許番号

特許第7165018号

(P7165018)

(45)発行日 令和4年11月2日(2022.11.2)

(24)登録日 令和4年10月25日(2022.10.25)

(51)国際特許分類

F I

G 0 6 F 17/10 (2006.01)

G 0 6 F 17/10

S

G 0 6 F 17/10

D

G 0 6 F 17/10

A

請求項の数 12 (全22頁)

(21)出願番号 特願2018-188612(P2018-188612)  
 (22)出願日 平成30年10月3日(2018.10.3)  
 (65)公開番号 特開2020-57286(P2020-57286A)  
 (43)公開日 令和2年4月9日(2020.4.9)  
 審査請求日 令和3年10月4日(2021.10.4)

(73)特許権者 000001007  
 キヤノン株式会社  
 東京都大田区下丸子3丁目30番2号  
 (74)代理人 110003281弁理士法人大塚国際特許事  
 務所  
 (72)発明者 チン ソクイ  
 東京都大田区下丸子3丁目30番2号  
 キヤノン株式会社内  
 審査官 漆原 孝治

最終頁に続く

(54)【発明の名称】 情報処理装置、情報処理方法

## (57)【特許請求の範囲】

## 【請求項1】

複数の階層を有するネットワークにおけるそれぞれの階層について、該階層の複数のデータと複数のフィルタ係数との積和演算を実行する複数の積和演算手段と、

前記複数の積和演算手段による複数の積和演算結果を複数のシフトパラメータに基づいてシフトする複数のシフト演算手段と、

前記複数のシフト演算手段によりシフトされた複数の積和演算結果の総和を計算する加算手段と、

前記階層のビット幅に応じて前記シフトパラメータの値を切り替える制御手段とを備えることを特徴とする情報処理装置。

10

## 【請求項2】

前記制御手段は、前記階層のビット幅に応じてフィルタ係数の転送回数を切り替えることを特徴とする請求項1に記載の情報処理装置。

## 【請求項3】

前記シフト演算手段は複数のシフトパラメータに基づいて前記複数の積和演算結果をシフトすることを特徴とする請求項1または2に記載の情報処理装置。

## 【請求項4】

複数の階層を有するネットワークにおけるそれぞれの階層について、該階層の複数のデータを複数のシフトパラメータに基づいてシフトする複数のシフト演算手段と、

前記複数のシフト演算手段によりシフトされた複数のデータと複数のフィルタ係数との

20

積和演算を実行する複数の積和演算手段と、

前記複数の積和演算手段により計算された複数の積和演算結果の総和を計算する加算手段と、

前記階層のビット幅に応じて前記シフトパラメータの値を切り替える制御手段とを備えることを特徴とする情報処理装置。

【請求項 5】

前記複数の積和演算手段と前記複数のシフト演算手段とは並列に動作することを特徴とする請求項 1 乃至 4 の何れか 1 項に記載の情報処理装置。

【請求項 6】

前記加算手段は、ある階層に対する前記総和を該階層の次の階層のデータとしてメモリに格納することを特徴とする請求項 1 乃至 5 の何れか 1 項に記載の情報処理装置。

【請求項 7】

前記加算手段は、ある階層に対する前記総和に対して活性化処理、プーリング処理を行った結果を、該階層の次の階層のデータとしてメモリに格納することを特徴とする請求項 1 乃至 5 の何れか 1 項に記載の情報処理装置。

【請求項 8】

更に、

前記メモリに格納されたデータに基づいて、動画像における各フレームに対する画像処理および/または画像認識を行う手段を備えることを特徴とする請求項 6 または 7 に記載の情報処理装置。

【請求項 9】

前記ネットワークは、階層ごとにデータのビット幅が異なることを特徴とする請求項 1 乃至 8 の何れか 1 項に記載の情報処理装置。

【請求項 10】

情報処理装置が行う情報処理方法であって、

前記情報処理装置の複数の積和演算手段が、複数の階層を有するネットワークにおけるそれぞれの階層について、該階層の複数のデータと複数のフィルタ係数との積和演算を実行する工程と、

前記情報処理装置の複数のシフト演算手段が、前記複数の積和演算手段による複数の積和演算結果を複数のシフトパラメータに基づいてシフトする工程と、

前記情報処理装置の加算手段が、前記複数のシフト演算手段によりシフトされた複数の積和演算結果の総和を計算する工程と、

前記情報処理装置の制御手段が、前記階層のビット幅に応じて前記シフトパラメータの値を切り替える工程と

を備えることを特徴とする情報処理方法。

【請求項 11】

情報処理装置が行う情報処理方法であって、

前記情報処理装置の複数のシフト演算手段が、複数の階層を有するネットワークにおけるそれぞれの階層について、該階層の複数のデータを複数のシフトパラメータに基づいてシフトする工程と、

前記情報処理装置の複数の積和演算手段が、前記複数のシフト演算手段によりシフトされた複数のデータと複数のフィルタ係数との積和演算を実行する工程と、

前記情報処理装置の加算手段が、前記複数の積和演算手段により計算された複数の積和演算結果の総和を計算する工程と、

前記情報処理装置の制御手段が、前記階層のビット幅に応じて前記シフトパラメータの値を切り替える工程と

を備えることを特徴とする情報処理方法。

【請求項 12】

コンピュータを、請求項 1 乃至 9 の何れか 1 項に記載の情報処理装置の各手段として機能させるためのコンピュータプログラム。

10

20

30

40

50

## 【発明の詳細な説明】

## 【技術分野】

## 【0001】

本発明は、複数の階層を有するネットワークにおける演算技術に関するものである。

## 【背景技術】

## 【0002】

近年、深層学習の進歩により、画像認識の精度が上がっている。畳み込みニューラルネットワーク (Convolutional Neural Networks : CNN) は、深層学習に用いられる手法として知られている。CNNでは、複数のレイヤ (階層) が階層的に接続されており、各レイヤの中に複数枚の特徴画像がある。図2に、レイヤの数が4 (レイヤ1~4) であって、各レイヤの中に特徴画像が4枚あるネットワークの例を示す。図2において特徴画像 (i, j) は、レイヤiにおけるj枚目の特徴画像を表す。学習したフィルタ係数と特徴画像の画素 (特徴データ) を用いてフィルタ処理の結果を計算する。フィルタ処理は積和演算であり、複数の乗算と累積加算を含んでいる。図2の矢印は積和演算を意味する。

10

## 【0003】

現レイヤにおける特徴画像は、前レイヤにおける特徴画像と前レイヤに対応するフィルタ係数とを用いて計算する。現レイヤにおける1枚の特徴画像を計算するためには、前レイヤにおける複数枚の特徴画像の情報が必要である。現レイヤにおける特徴画像を計算するための畳み込み演算の計算式は以下の通りである。

20

## 【0004】

## 【数1】

$$O_{i,j}(n) = \sum_{m=1}^M \sum_{x=0}^{X-1} \sum_{y=0}^{Y-1} (I_{i+x,j+y}(m) \times C_{x,y}(m,n))$$

(式1)

30

## 【0005】

ここで、 $O_{i,j}(n)$  は、現レイヤにおけるn枚目の特徴画像中の位置 (i, j) に対応する積和演算結果を表す変数である。この (式1) では、前レイヤの中に特徴画像がM枚あり、m枚目の特徴画像において位置 (i, j) における特徴データを  $I_{i,j}(m)$  と表している。フィルタ係数 ( $C_{1,1}(m,n) \sim C_{X,Y}(m,n)$ ) は ( $X \times Y$ ) 個あり、特徴画像によって異なる。現レイヤにおけるn枚目の特徴画像を算出するための積和演算回数は ( $M \times X \times Y$ ) 回である。畳み込み演算を行った後に、現レイヤの情報に基づき、積和演算結果  $O_{i,j}(n)$  を用いて活性化処理やプーリング等の処理を行い、現レイヤの特徴画像を計算する。

40

## 【0006】

CNNでは、積和演算の回数が多いため、携帯端末や車載機器等の組み込みシステムに適用する場合、効率的なデータ並列処理装置が求められる。処理データのビット幅を削減すると、畳み込み演算結果を計算する演算器のコストが小さくなるため、演算器の並列度を上げることが可能になる。非特許文献1では、レイヤ毎にデータのビット幅が異なるネットワークを処理するハードウェアの構成が提案されている。

## 【先行技術文献】

## 【非特許文献】

## 【0007】

50

【文献】Y. Li, et al., A 7.663-TOPS 8.2-W Energy-efficient FPGA Accelerator for Binary Convolutional Neural Networks, Proceedings of the 2017 ACM/SIGDA International Symposium on Field-Programmable Gate Arrays, Pages 290-291, Feb. 2017

K. Lee, et al., A 502-GOPS and 0.984-mW Dual-Mode Intelligent ADAS SoC With Real-Time Semiglobal Matching and Intention Prediction for Smart Automotive Black Box System, IEEE Journal of Solid-State Circuits, Vol. 52, No. 1, Pages 139-150, Jan. 2017

【発明の概要】

【発明が解決しようとする課題】

10

【0008】

非特許文献1に記載の手法では、異なる種類の演算器でレイヤ毎にビット幅が異なるCNNを処理する。入力レイヤの特徴データのビット幅が8ビットであり、中間レイヤの特徴データのビット幅が2ビットの場合、8ビットのデータ専用の畳み込み演算器と2ビットのデータ専用の畳み込み演算器が必要である。

【0009】

8ビットのデータのレイヤと2ビットのデータのレイヤをパイプライン方式で処理し、データを並行して処理することができるが、レイヤ毎の畳み込み処理の計算量が異なる場合には、ハードウェアの利用効率が低下する。また、2ビットと8ビットの間のビット幅(4ビット等)の特徴データを処理する場合、専用の畳み込み演算器がないため、8ビットのデータ専用の畳み込み演算器を利用することになり、効率が下がる。

20

【0010】

非特許文献2に記載の手法では、複数種類のビット幅の特徴データを処理可能なSIMD(シングルインストラクションマルチプルデータ)構成のRNN(Recurrent Neural Network)専用ハードウェアが提案されている。同じハードウェアで8ビット、16ビット、32ビットのデータを処理可能であるが、並行して出力されたデータの総和を計算する際に、一旦メモリに保持してからもう一回SIMDのコマンドを実行する必要があるため、処理時間が長くなる。本発明では、複数の階層を有するネットワークの中に複数種類のビット幅の特徴データがあっても効率的な処理を実現するための技術を提供する。

30

【課題を解決するための手段】

【0011】

本発明の一様態は、複数の階層を有するネットワークにおけるそれぞれの階層について、該階層の複数のデータと複数のフィルタ係数との積和演算を実行する複数の積和演算手段と、前記複数の積和演算手段による複数の積和演算結果を複数のシフトパラメータに基づいてシフトする複数のシフト演算手段と、前記複数のシフト演算手段によりシフトされた複数の積和演算結果の総和を計算する加算手段と、前記階層のビット幅に応じて前記シフトパラメータの値を切り替える制御手段とを備えることを特徴とする。

【発明の効果】

【0012】

40

本発明の構成によれば、複数の階層を有するネットワークの中に複数種類のビット幅の特徴データがあっても効率的な処理を実現することができる。

【図面の簡単な説明】

【0013】

【図1】データ処理のフローチャート。

【図2】処理対象ネットワークの構成例を示す図。

【図3】情報処理装置のハードウェア構成例を示すブロック図。

【図4】データ処理部305の構成例を示すブロック図。

【図5】積和演算部402とシフト演算部403の構成例を示すブロック図。

【図6】処理時間とデータの関係図。

50

【図 7】データ処理部 305 の構成例を示すブロック図。

【図 8】積和演算部 702 とシフト演算部 701 の構成例を示すブロック図。

【図 9】データ処理のフローチャート。

【図 10】特徴データが 8 ビット / 2 ビット / 4 ビットの例を示す図。

【発明を実施するための形態】

【0014】

以下、添付図面を参照し、本発明の実施形態について説明する。なお、以下説明する実施形態は、本発明を具体的に実施した場合の一例を示すもので、特許請求の範囲に記載した構成の具体的な実施形態の 1 つである。

【0015】

[ 第 1 の実施形態 ]

まず、本実施形態に係る情報処理装置のハードウェア構成例について、図 3 のブロック図を用いて説明する。情報処理装置には、PC ( パーソナルコンピュータ )、タブレット型端末装置、スマートフォンなどのコンピュータ装置を適用することができる。また、この情報処理装置は、これらの機器などに組み込む組み込み機器であっても構わない。

【0016】

入力部 301 は、キーボード、マウス、タッチパネルなどのユーザインターフェースにより構成されており、ユーザが操作することで各種の指示を CPU 306 に対して入力することができる。

【0017】

データ保存部 302 は、ハードディスクドライブ装置等の大容量情報記憶装置である。データ保存部 302 には、OS ( オペレーティングシステム )、CPU 306 が実行する各種のコンピュータプログラム、CPU 306 が各種の処理を実行する際に用いるデータ等、情報処理装置にて使用される各種の情報が保存されている。データ保存部 302 に保存されているデータには、画像処理部 309 による処理対象となる画像も含まれている。また以下の説明において既知の情報として説明する情報は、このデータ保存部 302 に保存されている。データ保存部 302 に保存されているコンピュータプログラムやデータは、CPU 306 やデータ処理部 305 や画像処理部 309 などにより RAM 308 などにロードされ、CPU 306 やデータ処理部 305 や画像処理部 309 などの処理対象となる。

【0018】

なお、データ保存部 302 は、記憶媒体 ( フレキシブルディスク、CD-ROM、CD-R、DVD、メモリーカード、CF カード、スマートメディア、SD カード、メモリスティック、xD ピクチャーカード、USB メモリ等 ) であっても良い。この場合、情報処理装置は、このような記憶媒体に対する情報の読み書きを行う装置を有する必要がある。

【0019】

通信部 303 は、外部の機器との間のデータ通信を行うための通信インターフェースとして機能するものであり、情報処理装置において処理の実行に必要な情報等をこの通信部 303 によって外部から取得するようにしても良い。また、情報処理装置が行った処理の結果を通信部 303 によって外部の機器に対して送信しても良い。

【0020】

表示部 304 は、液晶画面やタッチパネル画面などにより構成されており、CPU 306 やデータ処理部 305 や画像処理部 309 による処理結果を画像や文字などでもって表示することができる。なお、表示部 304 は、プロジェクタ装置などの投影装置であっても良い。また、入力部 301 と表示部 304 とを一体化させてタッチスクリーン装置のような、指示入力の受け付け機能と表示機能とを有する装置を形成しても良い。

【0021】

データ処理部 305 は、画像処理部 309 によって RAM 308 に書き込まれた画像を用いて図 1 のフローチャートに従った処理を実行することで CNN の計算を行い、該計算によって得られた結果をデータ保存部 302 や RAM 308 等に出力する。なお、データ

10

20

30

40

50

処理部 305 による処理対象となる画像は、画像処理部 309 によって RAM 308 に書き込まれた画像に限らず、例えば、他の装置によって入力された画像であっても良い。データ処理部 305 については、図 4 を用いて後述する。

【0022】

CPU 306 は、ROM 307 や RAM 308 に格納されているコンピュータプログラムやデータを用いて各種の処理を実行する。これにより CPU 306 は、情報処理装置全体の動作制御を行う。

【0023】

ROM 307 には、情報処理装置の設定データや起動プログラムなどの書換不要の情報が格納されている。RAM 308 は、データ保存部 302 や ROM 307 からロードされたコンピュータプログラムやデータ、通信部 303 が外部から受信した情報、を格納するためのエリアを有する。また RAM 308 は、CPU 306 やデータ処理部 305 や画像処理部 309 が各種の処理を実行する際に用いるワークエリアを有する。このように RAM 308 は、各種のエリアを適宜提供することができる。

10

【0024】

画像処理部 309 は、CPU 306 からの指示に従って、データ保存部 302 に保存されている画像を読み出し、該画像中の各画素の画素値のレンジ調整を行ってから、該画像を RAM 308 に書き込む。

【0025】

上記の入力部 301、データ保存部 302、通信部 303、表示部 304、データ処理部 305、CPU 306、ROM 307、RAM 308、画像処理部 309、は何れもバス 310 に接続されている。

20

【0026】

なお、情報処理装置のハードウェア構成は図 3 に示した構成に限らない。例えば、図 3 の構成を複数の装置で実現させても良い。また、入力部 301、データ保存部 302、表示部 304、といった機器を情報処理装置が有することに限らず、このような機器を情報処理装置と通信路を介して接続するようにしても良い。

【0027】

また、RAM 308 に格納するものとして説明する情報の一部若しくは全部をデータ保存部 302 に格納するようにしても良いし、データ保存部 302 に格納するものとして説明する情報の一部若しくは全部を RAM 308 に格納するようにしても良い。あるいは、RAM 308 の一部をデータ保存部 302 として用いるようにしても良いし、通信部 303 が通信する相手の機器の記憶装置を、通信部 303 を介して利用するというように仮想的に構成するようにしても良い。

30

【0028】

また、図 3 では CPU 306 は 1 つとしているが、情報処理装置が有する CPU 306 の数は 1 に限らず、複数であっても良い。また、データ処理部 305 や画像処理部 309 はハードウェアとして実装するようにしても良いし、コンピュータプログラムとして実装するようにしても良い。後者の場合、このコンピュータプログラムはデータ保存部 302 に格納され、CPU 306 が該コンピュータプログラムを実行することでデータ処理部 305 や画像処理部 309 の機能を実現させることになる。

40

【0029】

なお、CPU 306 は、データ処理部 305 による処理の結果に基づき、通信部 303 やデータ保存部 302 から取得する動画像における各フレームに対する画像処理および/または画像認識を行う。CPU 306 による画像処理や画像認識の結果は、RAM 308 やデータ保存部 302 に保存される、もしくは通信部 303 を介して外部の機器に対して出力される。また、CPU 306 による画像処理や画像認識の結果は、表示部 304 に画像や文字として表示されても良いし、情報処理装置が音声出力機能を有する場合には、音声として出力するようにしても良い。

【0030】

50

### < 処理対象ネットワーク >

本実施形態では処理対象ネットワークとしてCNNを用いる。処理対象ネットワークの構成例を図2に示す。図2の処理対象ネットワークの詳細については上記の通りである。なお、処理対象ネットワークの積和演算の計算量、特徴画像のサイズ、枚数、ビット数等の情報はデータ保存部302等に保存されている。

#### 【0031】

図2の処理対象ネットワークのレイヤの数は4（レイヤ1～4）であり、各レイヤの中に特徴画像が4枚ある。上記の通り、特徴画像（ $i, j$ ）は、レイヤ $i$ における $j$ 枚目の特徴画像を表す。また、レイヤにより、レイヤ中の特徴画像のビット幅が異なる。レイヤ1の特徴画像のビット幅は8ビットであり、レイヤ2の特徴画像のビット幅は2ビットであり、レイヤ3の特徴画像のビット幅は4ビットであり、レイヤ4の特徴画像のビット幅は8ビットである。最初のレイヤ（レイヤ1）と最後のレイヤ（レイヤ4）は入出力画像の情報を保持するため、中間レイヤ（レイヤ2, 3）のビット幅（それぞれ2ビット、4ビット）より大きいビット幅（8ビット）を使用することが多い。特徴画像は複数の画素（特徴データ）で構成されている。

10

#### 【0032】

ここで、データ処理部305によるレイヤ1～4のそれぞれの特徴画像の計算（生成）について説明する。レイヤ1における8ビットの特徴画像（1, 1）、（1, 2）、（1, 3）、（1, 4）と、フィルタ係数と、を用いて上記の（式1）に従った積和演算を行う。そして、この積和演算の結果から、レイヤ2における2ビットの特徴画像（2, 1）、（2, 2）、（2, 3）、（2, 4）を生成する。

20

#### 【0033】

そして、レイヤ2における2ビットの特徴画像（2, 1）、（2, 2）、（2, 3）、（2, 4）と、フィルタ係数と、を用いて上記の（式1）に従った積和演算を行う。そして、この積和演算の結果から、レイヤ3における4ビットの特徴画像（3, 1）、（3, 2）、（3, 3）、（3, 4）を生成する。

#### 【0034】

そして、レイヤ3における4ビットの特徴画像（3, 1）、（3, 2）、（3, 3）、（3, 4）と、フィルタ係数と、を用いて上記の（式1）に従った積和演算を行う。そして、この積和演算の結果から、レイヤ4における8ビットの特徴画像（4, 1）、（4, 2）、（4, 3）、（4, 4）を生成する。

30

#### 【0035】

##### < データ処理部305の構成例 >

データ処理部305の構成例を図4に示す。データメモリ401は、各レイヤにおけるそれぞれの特徴画像を構成する特徴データを保持しており、係数メモリ404はフィルタ係数を保持している。積和演算部402は、係数メモリ404に保持されているフィルタ係数と、データメモリ401に保持されている特徴データと、を用いた積和演算を行うことで、積和演算結果を計算する。シフト演算部403は、積和演算部402によるそれぞれの積和演算結果をシフトし、加算部406は、シフトされた複数の積和演算結果を足し合わせて「シフトされた積和演算結果の総和」を求める。処理部407は、加算部406が求めた「シフトされた積和演算結果の総和」に基づいて、活性化・プーリングの処理結果を計算し、該計算した活性化・プーリングの処理結果をデータメモリ401に格納する。制御部405は、データ処理部305全体の動作制御を行う。

40

#### 【0036】

データ処理部305によるデータ処理について、図1のフローチャートに従って説明する。ステップS101では、制御部405は、複数枚の入力特徴画像の特徴データとフィルタ係数とをRAM308から読み出し、該読み出した特徴データをデータメモリ401に格納し、該読み出したフィルタ係数を係数メモリ404に格納する。

#### 【0037】

ステップS102では、制御部405は、レイヤについてのループを開始し、未処理の

50

レイヤのうち1つを以下の処理対象レイヤとする。ここでは一例としてレイヤ1～4の順に処理対象とするため、最初はレイヤ1が処理対象レイヤとなる。

【0038】

ステップS103では、制御部405は、シフト演算部403のシフトパラメータをレイヤ情報に応じて設定する。ステップS104では、制御部405は、出力特徴画像のループを開始し、出力特徴データを順番に計算する。ステップS105では、制御部405は、加算部406に保存されている積和演算結果を初期化してゼロに設定する。加算部406は総和計算の機能を備えている。

【0039】

ステップS106では、制御部405は、入力特徴画像のループを開始し、入力特徴データを順番に処理する。ステップS107では、制御部405による制御の元で、積和演算部402、シフト演算部403のそれぞれは上記の積和演算、シフト演算を行う。ステップS107における処理の詳細（ステップS115～ステップS117）については後述する。

10

【0040】

ステップS108では、加算部406は、複数の積和演算結果を足し合わせて「シフトされた積和演算結果の総和」を求める。ステップS109では、制御部405は、入力特徴画像のループの終了判定をする。全ての入力特徴画像の処理が終わった場合にはステップS110に進み、未処理の入力画像特徴が残っている場合には、ステップS107に戻り、次の未処理の入力特徴画像の処理を開始する。ステップS110では、処理部407は、加算部406が求めた「シフトされた積和演算結果の総和」に基づき、以下の（式2）に従って活性化処理結果を計算する。

20

【0041】

【数2】

$$f(x) = \begin{cases} 0, & x < 0 \\ x, & x \geq 0 \end{cases}$$

（式2）

30

【0042】

ここで、 $f(\cdot)$ は活性化関数であり、 $x$ は入力データである。この例ではReLU（Rectified Linear Unit）を用いて活性化関数を実現したが、ReLUに限らず、他の非線形の関数、または量子化関数で実現してもよい。なお、必要に応じて活性化処理結果のビット幅に調整する。

【0043】

ステップS111では、処理部407はレイヤの情報に応じて、活性化処理結果に基づいてプーリング処理を行うことで、活性化・プーリングの処理結果を計算する。ステップS112では、処理部407は、ステップS111で計算した活性化・プーリングの処理結果を、次のレイヤの特徴画像としてデータメモリ401に格納する。

40

【0044】

ステップS113では、制御部405は、出力特徴画像のループの終了判定をする。全ての出力特徴画像の処理が終わった場合にはステップS114に進み、未処理の出力特徴画像が残っている場合には、ステップS105に戻り、未処理の出力特徴画像について処理を開始する。

【0045】

ステップS114では、制御部405はレイヤのループの終了判定をする。全てのレイヤの処理が終わった場合、図1のフローチャートに従った処理は終了し、未処理のレイヤが残っている場合には、ステップS103に戻り、未処理のレイヤについて処理を開始す

50



る。

#### 【 0 0 4 6 】

< 積和演算・シフト演算 >

ステップ S 1 0 7 における積和演算・シフト演算（ステップ S 1 1 5 ~ ステップ S 1 1 7）について説明する。ステップ S 1 1 5 では、制御部 4 0 5 は、データメモリ 4 0 1 から特徴データを読み出して積和演算部 4 0 2 に転送すると共に、係数メモリ 4 0 4 からフィルタ係数を読み出して積和演算部 4 0 2 に転送する。ここで、フィルタ係数の個数と転送回数は特徴データのビット幅により異なる。

#### 【 0 0 4 7 】

ステップ S 1 1 6 では、積和演算部 4 0 2 は、特徴データとフィルタ係数とに基づいて積和演算結果を計算する。ステップ S 1 1 7 では、シフト演算部 4 0 3 は、ステップ S 1 1 6 で得られた積和演算結果を、ステップ S 1 0 3 で設定されていたシフトパラメータに基づいてシフトする。

#### 【 0 0 4 8 】

< ビット幅が異なる場合の詳細説明 >

本実施形態では、ビット幅が異なるデータを処理することが可能である。図 1 0 ( A ) には、8 ビットの特徴データを処理する場合における積和演算部 4 0 2、シフト演算部 4 0 3、加算部 4 0 6 の動作を示している。図 1 0 ( B ) には、2 ビットの特徴データを処理する場合における積和演算部 4 0 2、シフト演算部 4 0 3、加算部 4 0 6 の動作を示している。

#### 【 0 0 4 9 】

特徴データが 8 ビットの場合、図 1 0 ( A ) に示す如く、積和演算部 4 0 2 は 8 ビットの特徴データ 1 0 0 1 ( 値 : 2 3 4 ) を 2 ビットのデータ ( 2 ビットデータ ) ごとに分割する。そして積和演算部 4 0 2 は、該分割した 4 個の 2 ビットデータ ( 値 : 2、2、2、3 ) と共通のフィルタ係数とを用いて 4 つの積和演算結果を計算し、シフト演算部 4 0 3 は、4 個のシフトパラメータに基づいて 4 つの積和演算結果をシフトする。そして加算部 4 0 6 は、シフトされた 4 つの積和演算結果を足し合わせて 1 個の特徴データ ( 8 ビットの入力特徴データの積和演算結果 ) を計算する。このように、データ処理部 3 0 5 は 1 個の 8 ビットの入力特徴データを処理できる。

#### 【 0 0 5 0 】

特徴データが 2 ビットの場合、図 1 0 ( B ) に示す如く、積和演算部 4 0 2 は 4 個の 2 ビットデータ 1 0 0 2 ( 値 : 2、2、2、3 ) と 4 個のフィルタ係数とを用いて、4 つの積和演算結果を計算する。シフト演算部 4 0 3 は、1 個のシフトパラメータに基づいて、4 つの積和演算結果をシフトする。シフトパラメータはゼロのため、シフト演算前後の結果は同じである。そして加算部 4 0 6 は、4 つの積和演算結果を足し合わせて 1 個の特徴データ ( 4 個の 2 ビットの入力特徴データの積和演算結果の総和 ) を計算する。このように、データ処理部 3 0 5 は 4 個の 2 ビットの入力特徴データを並列に処理できる。

#### 【 0 0 5 1 】

ここで、入力特徴画像の枚数を M とし、フィルタサイズを  $1 \times 1$  とする。フィルタサイズが 1 画素であり、変数  $x$ 、 $y$  の値は定数であるため、 $O_{i,j}(n)$  は  $I_{i,j}(n)$  を用いて計算する。積和演算の計算 ( 式 1 ) が以下のように簡略化される。

#### 【 0 0 5 2 】

【 数 3 】

10

20

30

40

$$O(n) = \sum_{m=1}^M (I(m) \times C(m, n))$$

(式3)

【0053】

10

フィルタサイズが  $1 \times 1$  よりも大きい場合、積和演算部402はフィルタ係数と入力特徴データとの畳み込み演算結果を計算するが、フィルタサイズが  $1 \times 1$  の場合、積和演算部402は  $I(m)$  と  $C(m, n)$  の積を計算する。

【0054】

処理対象の特徴データはビット幅がビットとビット幅がビットの2種類であるとする。図4に示した積和演算部402は積和演算結果を計算するビットのデータの積和演算ユニットがP個あり、シフト演算部403はシフト演算結果を計算するビットのデータのシフト演算ユニットがP個ある。、、Pは以下の条件を満たしている。

【0055】

【数4】

20

$$\beta = \alpha \times P$$

(式4)

【0056】

入力特徴データ  $I'()$  のビット幅がビットの場合、加算部406の出力は以下の(式5)で表され、(式6)、(式7)、(式8)が前提条件になる。

【0057】

【数5】

30

$$O(n) = \sum_{m=1}^M \sum_{p=1}^P [(I_{(\alpha),p}(m) \times C_p(m, n)) \times 2^{S(p)}]$$

(式5)

【0058】

$O(n)$  は  $n$  枚目の出力特徴画像の積和演算結果であり、 $I()$ 、 $p(m)$  はビットのデータの積和演算ユニットの入力データであり、 $C_p(m, n)$  はフィルタ係数であり、 $S(p)$  はシフトパラメータである。変数  $m$  はビットの入力特徴画像グループ(1グループ = P枚)の番号(積和演算部402の処理番号)であり、変数  $p$  は積和演算ユニット、シフト演算ユニットの番号であり、変数  $n$  は出力特徴画像の番号である。シフト演算は2のべき乗の処理で表現される。

40

【0059】

フィルタ係数  $C_p(m, n)$  は(式6)に示す如く、 $m$  枚目のビットの特徴画像が対応しているフィルタ係数  $C' (m, n)$  である。ビットの入力特徴画像グループはフィルタ係数が共通となるため、 $p$  を省略することができる。P個の積和演算ユニットに並列に供給するフィルタ係数の個数は1個であり、転送回数は1回である。

50

【 0 0 6 0 】

【 数 6 】

$$C_p(m, n) = \hat{C}(m, n)$$

(式 6)

【 0 0 6 1 】

ここでは、入力特徴データ  $I'(\ )$  を  $P$  個の ビットのデータ  $I(\ )$  ,  $p(m)$  に分割する。シフトパラメータ  $S(p)$  の値は (式 7) に示す如く、積和演算ユニットの番号  $p$  と分割されたデータのビット幅 に基づいて計算する。

10

【 0 0 6 2 】

【 数 7 】

$$S(p) = \alpha \times (p - 1)$$

(式 7)

20

【 0 0 6 3 】

ビットの入力特徴データ  $I'(\ )$  は式 ( 8 ) に示す如く、分割された  $P$  個の ビットのデータ  $I(\ )$  ,  $p(m)$  で表現される。

【 0 0 6 4 】

【 数 8 】

$$\sum_{p=1}^P I_{(\alpha),p}(m) \times 2^{S(p)} = \hat{I}_{(\beta)}(m)$$

30

(式 8)

【 0 0 6 5 】

ここで、(式 6)、(式 7)、(式 8) を (式 5) に代入すると、出力データ  $O(n)$  の式は以下の (式 9) となる。

【 0 0 6 6 】

【 数 9 】

$$O(n) = \sum_{m=1}^M \sum_{p=1}^P [(I_{(\alpha),p}(m) \times C_p(m, n)) \times 2^{\alpha \times (p-1)}] = \sum_{m=1}^M (\hat{I}_{(\beta)}(m) \times \hat{C}(m, n))$$

40

(式 9)

【 0 0 6 7 】

一方、入力特徴データ  $I'(\ )$  のビット幅が ビットの場合、加算部 4 0 6 の出力は以下の (式 1 0) で表され、(式 1 1)、(式 1 2)、(式 1 3) が前提条件になる。

【 0 0 6 8 】

50

【数 1 0】

$$O(n) = \sum_{m=1}^{\frac{M}{P}} \sum_{p=1}^P [(I_{(\alpha),p}(m) \times C_p(m, n)) \times 2^{S(p)}]$$

(式 1 0)

10

【0 0 6 9】

$O(n)$  は  $n$  枚目の出力特徴画像の積和演算結果であり、 $I_{(\alpha),p}(m)$  は ビットのデータの積和演算ユニットの入力データであり、 $C_p(m, n)$  はフィルタ係数であり、 $S(p)$  はシフトパラメータである。変数  $m$  は ビットの入力特徴画像グループ (1 グループ =  $P$  枚) の番号 (積和演算部 4 0 2 の処理番号) であり、変数  $p$  は積和演算ユニット、シフト演算ユニットの番号であり、変数  $n$  は出力特徴画像の番号である。シフト演算は 2 のべき乗の処理で表現される。

【0 0 7 0】

フィルタ係数  $C_p(m, n)$  は第  $\{(m-1) \times P + p\}$  枚目の ビットの特徴画像が対応しているフィルタ係数  $C'((m-1) \times P + p, n)$  である。積和演算ユニットの番号  $p$  によりフィルタ係数が異なるため、 $P$  個の積和演算ユニットに並列に供給するフィルタ係数の個数は  $P$  個であり、転送回数は  $P$  回である。

20

【0 0 7 1】

【数 1 1】

$$C_p(m, n) = \hat{C}((m-1) \times P + p, n)$$

(式 1 1)

30

【0 0 7 2】

入力特徴データ  $I'_{(\alpha)}$  は ビットのデータの積和演算ユニットの入力データ  $I_{(\alpha)}$ ,  $p(m)$  になり、シフトパラメータ  $S(p)$  の値は (式 1 2) に示す如く常に 0 である。

【0 0 7 3】

【数 1 2】

$$S(p) = 0$$

(式 1 2)

40

【0 0 7 4】

$P$  個の ビットの入力特徴データ  $I'_{(\alpha)}$  をそのまま積和演算ユニットに入力するが、 $P$  個の入力データはそれぞれ異なる特徴画像の特徴データである。特徴画像の番号は、積和演算ユニットの番号  $p$  と、積和演算ユニットの個数  $P$  と、積和演算部 4 0 2 の処理番号  $m$  で以下の (式 1 3) に示す如く表現される。

【0 0 7 5】

【数 1 3】

50

$$I_{(\alpha),p}(m) = \hat{I}_{(\alpha)}((m-1) \times P + p)$$

(式 1 3)

【 0 0 7 6 】

ここで、(式 1 1)、(式 1 2)、(式 1 3)を(式 1 0)に代入すると、出力データ  $O(n)$  の式は以下の(式 1 4)となる。

10

【 0 0 7 7 】

【数 1 4】

$$O(n) = \sum_{m=1}^{\frac{M}{P}} \sum_{p=1}^P [\hat{I}_{(\alpha),p}((m-1) \times P + p) \times \hat{C}((m-1) \times P + p, n)]$$

(式 1 4)

【 0 0 7 8 】

20

シフトパラメータ  $S(p)$  の値とフィルタ係数の個数を変更することにより、ビット幅が  $\text{ビット}$  の特徴データ  $I'_{(8)}(1)$  とビット幅が  $\text{ビット}$  の特徴データ  $I'_{(8)}(2)$  を、同じ演算器(積和演算部 4 0 2、シフト演算部 4 0 3、加算部 4 0 6)で処理することができる。

【 0 0 7 9 】

<ビット幅が異なる場合の処理例>

図 5 と図 1 0 (A)、(B)は、 $P = 4$ 、 $\text{ } = 8$ 、 $\text{ } = 2$  の場合の構成例である。積和演算部 4 0 2 は入力データのビット幅が 2 ビットであり、シフト演算部 4 0 3 は入力データのビット幅が 6 ビットであり、加算部 4 0 6 の入力データのビット幅が 1 2 ビットである。

【 0 0 8 0 】

30

図 5 のハードウェア構成を用いて図 2 の処理対象ネットワークを処理した場合の処理時間の例を図 6 に示す。図 6 (A)と図 1 0 (A)にレイヤ 1 (8 ビットのデータ、入力特徴画像枚数  $M = 4$ ) を処理した場合の例を示す。特徴画像 (1, 1) の特徴データ  $I'_{(8)}(1)$  は 8 ビットであり、(式 8)に基づいて 4 つに分割された 4 つのデータ  $I_{(2),1}(1) - I_{(2),4}(1)$  を、積和演算部 4 0 2 に入力する。入力特徴データとシフトパラメータとフィルタ係数  $C(m, n)$  を用いてシフト演算結果を計算し、加算部 4 0 6 に入力し、初期値ゼロと足し合わせる。計算結果はシフト演算結果になり、加算部 4 0 6 に保持される。所要時間は 1 m s である。

【 0 0 8 1 】

特徴画像 (1, 2) の特徴データ  $I'_{(8)}(2)$  は 8 ビットであり、(式 8)に基づいて分割された 4 つのデータ  $I_{(2),1}(2) - I_{(2),4}(2)$  を、積和演算部 4 0 2 に入力する。入力特徴データとシフトパラメータとフィルタ係数  $C(m, n)$  を用いてシフト演算結果を計算し、加算部 4 0 6 に入力し、前の結果に加算する。所要時間は 1 m s である。

40

【 0 0 8 2 】

特徴画像 (1, 2) と同じように、特徴画像 (1, 3) と特徴画像 (1, 4) を順番に処理し、シフト演算結果を累積し、加算結果を計算する。所要時間は 2 m s である。最後に処理部 4 0 7 を通して、特徴画像 (2, 1) の特徴データを出力する。4 枚の特徴画像の処理時間は 4 m s である。

【 0 0 8 3 】

50

図 6 ( B ) と図 10 ( B ) にレイヤ 2 ( 2 ビットのデータ、入力特徴画像枚数  $M = 4$  ) を処理した場合の例を示す。特徴画像 ( 2 , 1 ) - ( 2 , 4 ) の特徴データ  $I'_{(2)}(1)$  -  $I'_{(2)}(4)$  は 2 ビットであり、(式 13) に基づいて 4 つのデータ  $I_{(2)}, 1(1)$  -  $I_{(2)}, 4(1)$  を、積和演算部 402 に並列に入力する。入力特徴データとシフトパラメータとフィルタ係数  $C_p(m, n)$  を用いてシフト演算結果を計算し、加算部 406 に入力し、初期値ゼロと足し合わせ、計算結果はシフト演算結果になる。最後に、処理部 407 を通して、特徴画像 ( 3 , 1 ) の特徴データを出力する。4 枚の特徴画像の処理時間は 1 m s である。

【 0084 】

図 6 ( A )、( B ) と図 10 ( A )、( B ) に示す如く、入力特徴データが 8 ビットの場合、出力データあたりの処理時間は 4 m s であり、入力特徴データが 2 ビットの場合、出力データあたりの処理時間は 1 m s である。共通のデータ処理部 305 でビット幅が異なるデータを効率よく処理することが可能である。

10

【 0085 】

[ 第 2 の実施形態 ]

以下では、第 1 の実施形態との差分について説明し、以下で特に触れない限りは第 1 の実施形態と同様であるものとする。

【 0086 】

< シフト演算と積和演算の順番 >

第 1 の実施形態では、積和演算の後にシフト演算を行う例について説明したが、積和演算とシフト演算の順番を入れ替えても処理結果は同じである。積和演算とシフト演算の順番が入れ替わった場合、図 1 のフローチャートの一部が変更される。ステップ S107 は図 9 のステップ S901 ~ ステップ S903 になる。

20

【 0087 】

本実施形態に係るデータ処理部 305 の構成例を図 7 に示す。シフト演算部 701 は、データメモリ 401 に格納されている特徴データをシフトパラメータに基づいてシフトし、積和演算部 702 は、シフトされた特徴データとフィルタ係数とに基づいて積和演算結果を計算する。

【 0088 】

< シフト演算・積和演算 >

ステップ S107 で行われる積和演算・シフト演算 ( ステップ S901 ~ ステップ S903 ) について説明する。ステップ S901 では、制御部 704 は、データメモリ 401 から特徴データを読み出し、係数メモリ 703 からフィルタ係数を読み出す。ステップ S902 では、シフト演算部 701 は、ステップ S103 で設定されていたシフトパラメータに基づき、特徴データをシフトする。ステップ S903 では、積和演算部 702 は、シフトされた特徴データとフィルタ係数とに基づいて積和演算結果を計算する。

30

【 0089 】

< ビット幅が異なる場合の詳細説明 >

本実施形態では、シフト演算部 701 はシフト演算結果を計算する ビットのデータのシフト演算ユニットが P 個あり、積和演算部 702 は積和演算結果を計算する ビットのデータの積和演算ユニットが P 個ある。積和演算部 702 の出力は (式 15) で表され、(式 5) に示したシフト演算部 403 の出力と等価である。

40

【 0090 】

【 数 15 】

$$O(n) = \sum_{m=1}^M \sum_{p=1}^P [(I_{(\alpha),p}(m) \times 2^{S(p)}) \times C_p(m, n)]$$

(式 15)

【 0 0 9 1 】

10

図 8 に、 $P = 4$ 、 $\alpha = 8$ 、 $\beta = 2$  の場合の例を示す。シフト演算部 701 は入力データのビット幅が 2 ビットであり、積和演算部 702 は入力データのビット幅が 8 ビットであり、加算部 406 の入力データのビット幅が 12 ビットである。シフト演算部 701 と積和演算部 702 はビット幅により回路規模が異なるため、シフト演算部 701 (シフト演算部 403) と積和演算部 702 (積和演算部 402) の順番を入れ替えることにより、全体的な回路規模を削減することが可能である。

【 0 0 9 2 】

[ 第 3 の実施形態 ]

第 1, 2 の実施形態では、入力特徴データのビット幅が  $\alpha$  ビット (積和演算ユニットのビット幅)、 $\beta$  ビット (積和演算ユニットのビット幅と積和演算ユニット数の積) である例について説明したが、制限があるわけではなく、 $\alpha$ 、 $\beta$  以外のビット幅でも良い。

20

【 0 0 9 3 】

< 入力特徴データのビット幅が  $\alpha$  ビットの場合 >

本実施形態では、ビット幅が  $\alpha$  ビットの入力特徴データを処理することが可能である。図 10 (C) に特徴データが 4 ビットの例を示す。特徴データが 4 ビットの場合、図 10 (C) に示す如く、積和演算部 402 は 2 個の 4 ビットの特徴データ 1003 (値: 10、14) を 2 ビットずつ分割する。そして積和演算部 402 は、該分割された 4 個の 2 ビットのデータ (値: 2、2、2、3) と 2 個のフィルタ係数を用いて、4 つの積和演算結果を計算する。シフト演算部 403 は、2 個のシフトパラメータに基づいて、4 つの積和演算結果をシフトする。加算部 406 は、シフトされた 4 つの積和演算結果を足し合わせて 1 個の特徴データ (2 個の 4 ビットの入力特徴データの積和演算結果の総和) を計算する。このように、データ処理部 305 は 2 個の 4 ビットの入力特徴データを並列に処理できる。 $\alpha$  は入力特徴データのビット幅であり、 $\beta$  の値は  $\alpha$  と異なる。 $\alpha$ 、 $\beta$ 、 $P$  は第 1 の実施形態と同じ定義であり、 $\alpha$ 、 $\beta$ 、 $P'$  は以下の条件を満たしている。

30

【 0 0 9 4 】

【 数 16 】

$$\gamma = \alpha \times P$$

(式 16)

40

【 0 0 9 5 】

$\alpha$  は  $\beta$  より小さく、 $P$  は  $P'$  の倍数である。入力特徴データ  $I(\gamma)$  のビット幅が  $\alpha$  ビットの場合、加算部 406 の出力データ  $O(n)$  は以下の (式 17) で表され、(式 18)、(式 19)、(式 20) が前提条件になる。

【 0 0 9 6 】

【 数 17 】

50

$$O(n) = \sum_{m=1}^{\frac{MP}{P}} \sum_{q=1}^{\frac{P}{P'}} \sum_{p=1}^{P'} [(I_{(\alpha),p,q}(m) \times C_{p,q}(m,n)) \times 2^{S(p,q)}]$$

(式 1 7)

【 0 0 9 7 】

$O(n)$  は  $n$  枚目の出力特徴画像の積和演算結果であり、 $I_{(\alpha),p}(m)$  は ビットのデータの積和演算ユニットの入力データであり、 $C_{p,q}(m,n)$  はフィルタ係数であり、 $S(p)$  はシフトパラメータである。変数  $m$  は ビットの入力特徴画像グループ (1 グループ =  $P$  枚) の番号 (積和演算部 4 0 2 の処理番号) である。積和演算ユニットとシフト演算ユニットはそれぞれ  $P / P'$  セットに分かれており、変数  $q$  は積和演算ユニットのセットの番号である。変数  $p$  はセット内の積和演算ユニット、シフト演算ユニットの番号であり、変数  $n$  は出力特徴画像の番号である。シフト演算は 2 のべき乗の処理で表現される。

10

【 0 0 9 8 】

フィルタ係数  $C_{p,q}(m,n)$  は第  $\{(m-1) \times P / P' + q\}$  枚目の ビットの特徴画像が対応しているフィルタ係数  $C'((m-1) \times P / P' + q, n)$  である。フィルタ係数は積和演算ユニットのセット番号  $q$  に基づいて計算する。フィルタ係数の一部が共通するため、 $P$  個の積和演算ユニットに並列に供給するフィルタ係数の個数は  $P / P'$  個であり、転送回数は  $P / P'$  回である。

20

【 0 0 9 9 】

【数 1 8】

$$C_{p,q}(m,n) = C'((m-1) \times \frac{P}{P'} + q, n)$$

(式 1 8)

30

【 0 1 0 0 】

ここでは、入力特徴データ  $I'_{(\alpha)}$  を  $P'$  個の ビットのデータ  $I_{(\alpha),p}(m)$  に分割する。シフトパラメータ  $S(\cdot)$  は積和演算ユニットのビット幅 と積和演算ユニットの番号  $p$  に基づいて計算する。

【 0 1 0 1 】

【数 1 9】

$$S(p,q) = \alpha \times (p-1)$$

(式 1 9)

40

【 0 1 0 2 】

ビットの入力特徴データ  $I'_{(\alpha)}$  は分割された  $P'$  個の ビットのデータ  $I_{(\alpha),p,q}(m)$  で表現される。

【 0 1 0 3 】

【数 2 0】

50



$$\sum_{p=1}^{\dot{P}} (I_{(\alpha),p,q}(m) \times 2^{S(p,q)}) = \dot{I}_{(\gamma)} \left( (m-1) \times \frac{P}{\dot{P}} + q \right)$$

(式 20)

【 0 1 0 4 】

(式 18)、(式 19)、(式 20) を (式 17) に代入すると、出力データ  $O(n)$  の式は以下の (式 21) になる。

【 0 1 0 5 】

【 数 2 1 】

$$O(n) = \sum_{m=1}^{\frac{M\dot{P}}{P}} \sum_{q=1}^{\frac{P}{\dot{P}}} \sum_{p=1}^{\dot{P}} [(I_{(\alpha),p,q}(m) \times C_{p,q}(m,n)) \times 2^{\alpha \times (p-1)}]$$

$$= \sum_{m=1}^{\frac{M\dot{P}}{P}} \sum_{q=1}^{\frac{P}{\dot{P}}} \left[ \dot{I}_{(\gamma)} \left( (m-1) \times \frac{P}{\dot{P}} + q \right) \times \dot{C} \left( (m-1) \times \frac{P}{\dot{P}} + q, n \right) \right]$$

(式 21)

【 0 1 0 6 】

シフトパラメータ  $S(p, q)$  の値とフィルタ係数の個数を設定することにより、ビット幅が ビットの特徴データ  $I'_{(\gamma)}$  を、第 1 の実施形態と同じ演算器 (積和演算部 402、シフト演算部 403、加算部 406) で処理することができる。

【 0 1 0 7 】

< ビット幅が異なる場合の処理例 >

図 5 と図 10 (C) に、 $P = 4$ 、 $\dot{P} = 8$ 、 $\alpha = 2$  の場合の構成例を示し、図 5 のハードウェア構成を用いて図 2 の処理対象ネットワークを処理した場合の処理時間の例を図 6 に示す。

【 0 1 0 8 】

図 6 (C) と図 10 (C) に  $P' = 2$ 、 $\dot{P}' = 4$  で、レイヤ 3 (4 ビットの特徴データ、入力特徴画像枚数  $M = 4$ ) を処理した場合の例を示す。特徴画像 (3, 1) と (3, 2) の特徴データ  $I'_{(\gamma)}(4), (1)$ 、 $I'_{(\gamma)}(4), (2)$  は 4 ビットであり、(式 20) に基づいて分割された 4 つの特徴データ  $I_{(\gamma)}(2), 1(1) - I_{(\gamma)}(2), 4(1)$  を、積和演算部 402 に入力する。入力特徴データとシフトパラメータとフィルタ係数  $C(m, n)$  を用いて、シフト演算結果を計算し、加算部 406 に入力し、初期値ゼロと足し合わせる。計算結果はシフト演算結果になり、加算部 406 に保持される。所要時間は 1 ms である。

【 0 1 0 9 】

特徴画像 (3, 3) と (3, 4) の特徴データ  $I'_{(\gamma)}(4), (3)$ 、 $I'_{(\gamma)}(4), (4)$  は 4 ビットであり、(式 20) に基づいて分割された 4 つの特徴データ  $I_{(\gamma)}(2), 1(2) - I_{(\gamma)}(2), 4(2)$  を、積和演算部 402 に入力する。入力特徴データとシフトパラメータとフィルタ係数  $C(m, n)$  を用いて、シフト演算結果を計算し、加算部 406 に入力し、前の結果に加算する。所要時間は 1 ms である。最後に処理部 407 を通して、特徴画像 (4, 1) の特徴データを出力する。4 枚の特徴画像の処理時間は 2 ms である。

【 0 1 1 0 】

このように、本実施形態では、ビット幅が ビット (積和演算ユニットのビット幅)、

10

20

30

40

50

ビット（積和演算ユニットのビット幅 と積和演算ユニットの数 P の積）以外の特徴データも処理できるため、柔軟性が高いというメリットがある。

【 0 1 1 1 】

〔 第 4 の実施形態 〕

第 1 の実施形態では、処理部 4 0 7 で活性化処理を実行する例について説明したが、活性化処理は処理部 4 0 7 で実行することに限らず、他の装置、例えば、C P U 3 0 6 が活性化処理を実行するようにしても良い。これは他の処理についても同様で、上記の各実施形態では、上記の様々な処理の主体については一例を示したに過ぎず、上記の各実施形態で説明した主体とは異なる主体であっても良い。

【 0 1 1 2 】

また、第 1 の実施形態では、レイヤ情報に応じて活性化・プーリング処理を実行するようにした。しかし、場合によっては、活性化・プーリング処理を省略するようにしても良い。

【 0 1 1 3 】

また、第 1 ～ 3 の実施形態では、フィルタサイズ（フィルタの高さと幅）が  $1 \times 1$  であるケースについて説明したが、フィルタサイズは  $1 \times 1$  に限らず、他のサイズであっても良い。このように、上記の各実施形態における説明で使用した数値はあくまでも具体的な説明を行うために例示した数値であって、上記の各実施形態にて説明した数値に限定されることを意図したものではない。

【 0 1 1 4 】

フィルタサイズが小さい場合、フィルタ係数を保持するメモリ（係数メモリ 4 0 4、7 0 3）の容量もより少なくてすむというメリットがある。フィルタの幅と高さの最小値は 1 である。

【 0 1 1 5 】

また、第 1 ～ 3 の実施形態では、入力特徴画像の枚数を M、出力特徴画像の枚数を N、とした。しかし、M、N に適用可能な数値は特定の数値に限らない。このように、上記の様々な変数に適用可能な数値は特定の数値に限らない。

【 0 1 1 6 】

また、第 1 ～ 3 の実施形態では、フィルタ係数を係数メモリ 4 0 4、7 0 3 に保持し、特徴データをデータメモリ 4 0 1 に保持するようにしたが、フィルタ係数や特徴データを保持するメモリは特定のメモリに限らない。例えば、フィルタ係数や特徴データを積和演算部 4 0 2、7 0 2 が有するメモリに保持しても良いし、R A M 3 0 8 に保持しても良い。

【 0 1 1 7 】

また、フィルタ係数のビット幅は特定のビット幅に限らない。また、第 1 ～ 3 の実施形態では、処理対象ネットワークとして C N N を用いたが、処理対象ネットワークは C N N に限らず、R N N や M L P（多層パーセプトロン）等のような他の種類の複数のレイヤが階層的に接続されているネットワークであっても良い。

【 0 1 1 8 】

（その他の実施形態）

本発明は、上述の実施形態の 1 以上の機能を実現するプログラムを、ネットワーク又は記憶媒体を介してシステム又は装置に供給し、そのシステム又は装置のコンピュータにおける 1 つ以上のプロセッサがプログラムを読み出し実行する処理でも実現可能である。また、1 以上の機能を実現する回路（例えば、A S I C）によっても実現可能である。

【符号の説明】

【 0 1 1 9 】

4 0 1：データメモリ 4 0 2：積和演算部 4 0 3：シフト演算部 4 0 4：係数メモリ 4 0 5：制御部 4 0 6：加算部 4 0 7：処理部

10

20

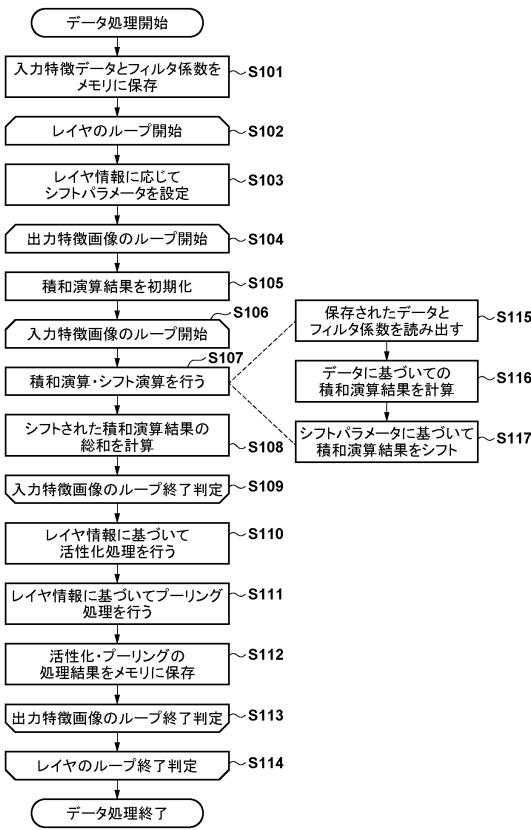
30

40

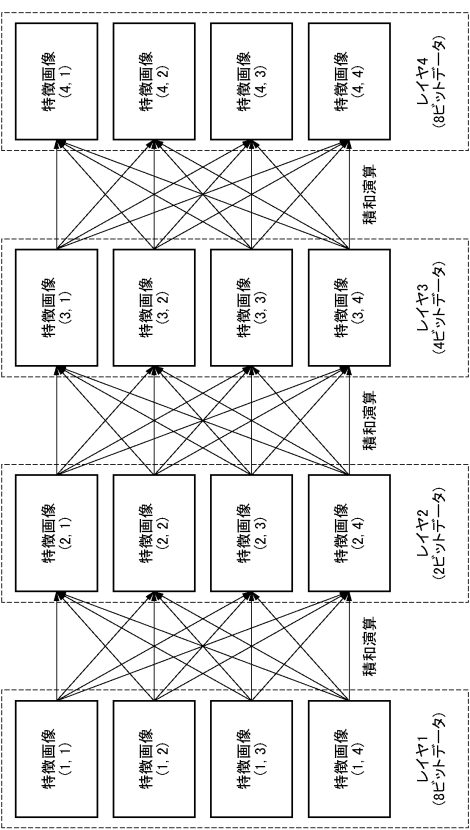
50

【図面】

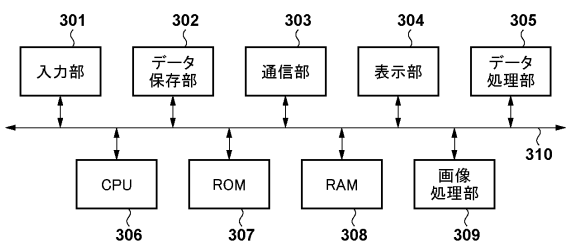
【図 1】



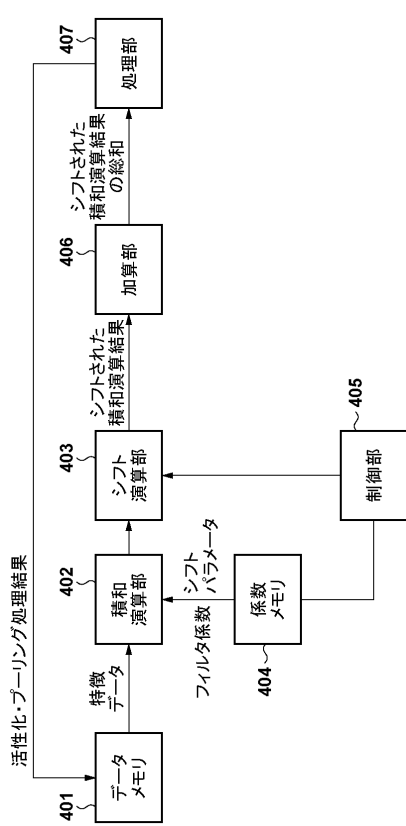
【図 2】



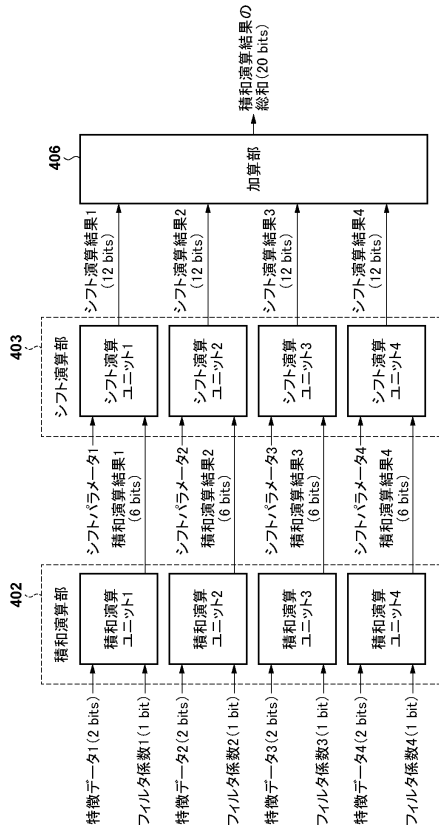
【図 3】



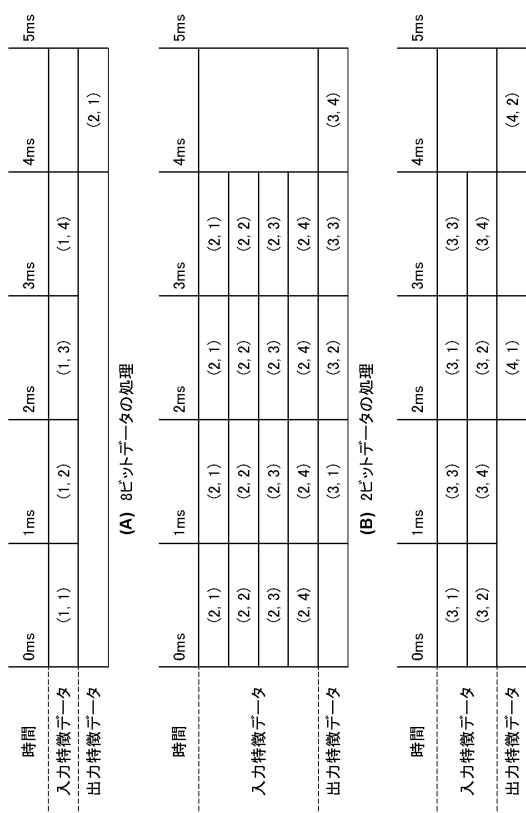
【図 4】



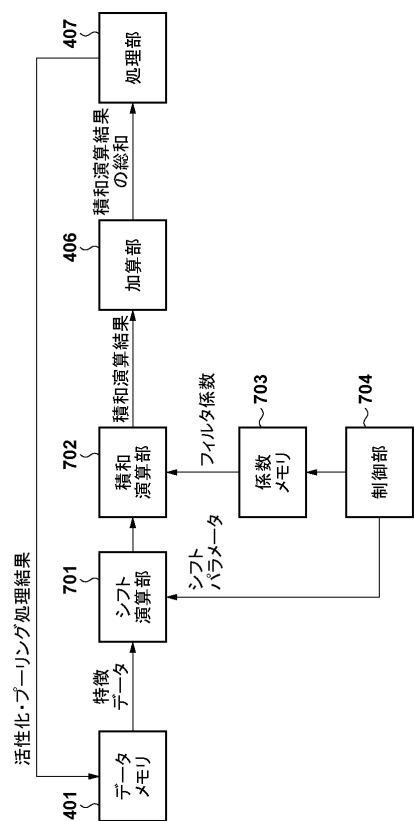
【図 5】



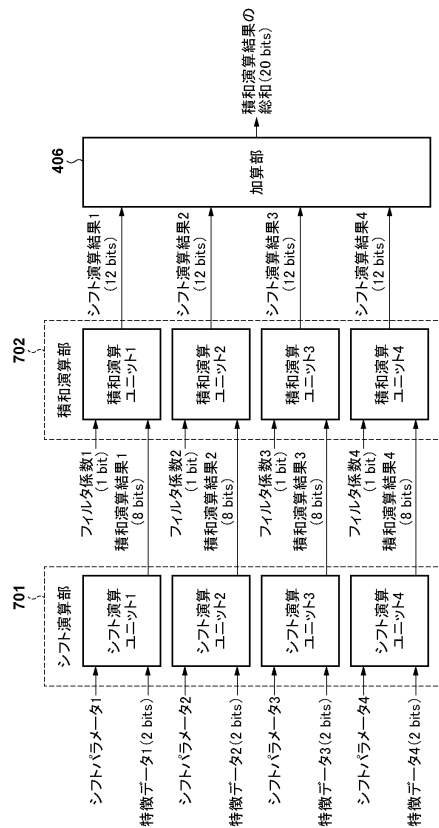
【図 6】



【図 7】



【図 8】



10

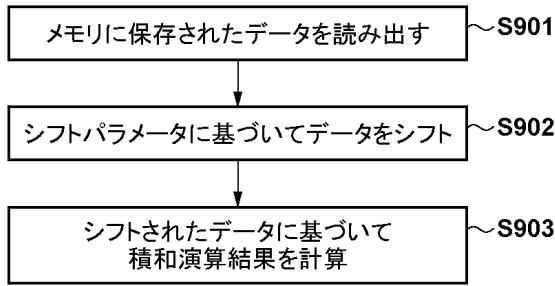
20

30

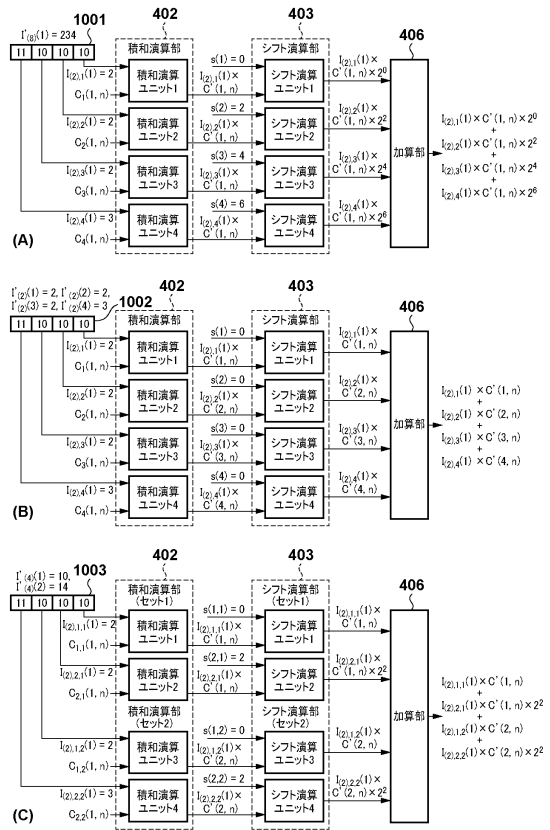
40

50

【図 9】



【図 10】



10

20

30

40

50

---

フロントページの続き

- (56)参考文献 特開 2 0 1 7 - 0 7 9 0 1 7 ( J P , A )  
米国特許出願公開第 2 0 1 7 / 0 1 1 6 4 9 5 ( U S , A 1 )  
国際公開第 2 0 0 6 / 0 6 8 1 9 6 ( W O , A 1 )
- (58)調査した分野 (Int.Cl., D B 名)  
G 0 6 F 1 7 / 1 0