

(12) 特許協力条約に基づいて公開された国際出願

(19) 世界知的所有権機関
国際事務局

(43) 国際公開日
2022年6月30日(30.06.2022)



(10) 国際公開番号

WO 2022/137696 A1

- (51) 国際特許分類:
G06N 3/063 (2006.01) G06F 17/10 (2006.01)
- (21) 国際出願番号: PCT/JP2021/035696
- (22) 国際出願日: 2021年9月28日(28.09.2021)
- (25) 国際出願の言語: 日本語
- (26) 国際公開の言語: 日本語
- (30) 優先権データ:
特願 2020-218058 2020年12月25日(25.12.2020) JP
- (71) 出願人: 日立 Astemo 株式会社(HITACHI ASTEMO, LTD.) [JP/JP]; 〒3128503 茨城県ひたちなか市高場 2 5 2 0 番地 Ibaraki (JP).
- (72) 発明者: 中野 和香子 (NAKANO, Wakako); 〒1008280 東京都千代田区丸の内一丁目 6 番 6 号 株式会社日立製作所内 Tokyo (JP). 平井 理宇 (HIRAI, Riu); 〒1008280 東京都千代田区丸の内一丁目 6 番 6 号 株式会社日立製作所内 Tokyo (JP). 村中 延之 (MURANAKA, Nobuyuki); 〒1008280 東京都千代田区丸の内

一丁目 6 番 6 号 株式会社日立製作所内 Tokyo (JP). 伊藤 浩朗(ITO, Hiroaki); 〒3128503 茨城県ひたちなか市高場 2 5 2 0 番地 日立 Astemo 株式会社内 Ibaraki (JP).

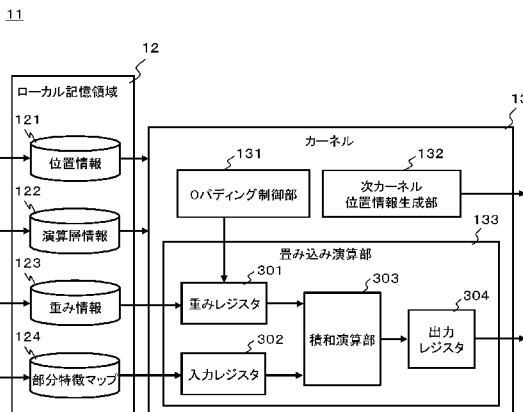
(74) 代理人: 特許業務法人サンネクスト国際特許事務所(SUNNEXT INTERNATIONAL PATENT OFFICE); 〒1400002 東京都品川区東品川二丁目 3 番 1 2 号 シーフォートスクエア センタービルディング 1 6 階 Tokyo (JP).

(81) 指定国(表示のない限り、全ての種類の国内保護が可能): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, IT, JO, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL,

(54) Title: INFORMATION PROCESSING DEVICE AND INFORMATION PROCESSING METHOD

(54) 発明の名称: 情報処理装置、情報処理方法

【図3】



- 12 Local storage area
13 Kernel
121 Position information
122 Calculation layer information
123 Weight information
124 Partial feature map
131 Zero padding control unit
132 Next kernel position information generation unit
133 Convolution calculation unit
301 Weight register
302 Input register
303 Multiply-accumulate operation unit
304 Output register

(57) Abstract: The present invention provides an information processing device comprising: a first calculation unit; a second calculation unit that is connected to the first calculation unit via a bus; and a storage unit that stores weight information to be used for DNN calculation, wherein the first calculation unit transmits to the second calculation unit a partial feature map, which is a part of a feature map, and position information indicating the position of the partial feature map in the feature map, wherein the second calculation unit acquires the weight information from the storage unit and receives the partial feature map and position information transmitted from the first calculation unit, and wherein the second calculation unit executes predetermined preprocessing based on the position information and executes the DNN calculation by using the result of executing the preprocessing and the partial feature map.

ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG,
US, UZ, VC, VN, WS, ZA, ZM, ZW.

- (84) 指定国(表示のない限り、全ての種類の広域保護が可能): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), ユーラシア (AM, AZ, BY, KG, KZ, RU, TJ, TM), ヨーロッパ (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

添付公開書類:

一 国際調査報告 (条約第21条(3))

(57) 要約: 情報処理装置は、第1の演算部と、前記第1の演算部とバスを介して接続された第2の演算部と、DNN演算で利用される重み情報を格納する記憶部と、を備え、前記第1の演算部は、特徴マップの一部である部分特徴マップと、前記部分特徴マップの前記特徴マップ中の位置を表す位置情報とを前記第2の演算部へ送信し、前記第2の演算部は、前記記憶部から前記重み情報を取得し、前記第1の演算部から送信される前記部分特徴マップおよび前記位置情報を受信し、前記位置情報に基づく所定の前処理を実行し、前記前処理の実行結果と前記部分特徴マップとを用いて、前記DNN演算を実行する。

明 細 書

発明の名称： 情報処理装置、情報処理方法

技術分野

[0001] 本発明は、情報処理装置および情報処理方法に関する。

背景技術

[0002] 従来、カメラの撮影画像や各種センサの情報から車両の周囲状況を認識し、その認識結果に基づいて様々な運転支援を行う技術が広く利用されている。こうした車両の運転支援技術において、近年では複雑な周囲状況に対して高精度な認識結果を得るために、人間の脳における神経細胞の働きをモデル化したディープニューラルネットワーク（DNN）を利用した演算を行うことが提案されている。

[0003] 従来のDNN演算では、多数の演算を並列で高速に実行するために、FPGA（Field Programmable Gate Array）やGPU（Graphics Processing Unit）が広く利用されている。近年では、こうしたFPGAやGPUの代わりに、AIアクセラレータと呼ばれるデバイスの利用が進んでいる。AIアクセラレータとは、複数の演算コアを有しており、ベクトル型演算を得意とするAI処理に特化した演算装置の一種である。

[0004] AIアクセラレータは、上記のようにDNN演算におけるベクトル型演算に特化したものであるが、DNN演算に必要な処理の全てを実行することはできず、一部の処理については実行が困難である。そのため、従来のAIアクセラレータを用いたDNN演算では、AIアクセラレータとは別の演算処理部をAIアクセラレータに接続し、これらを協調動作させることでDNN演算を行っていた。この場合、DNNの各層についてAIアクセラレータと演算処理部がそれぞれの演算結果を受け渡すために、層ごとにデータ転送を行う必要がある。したがって、AIアクセラレータと演算処理部の間でのデータ通信量が増大し、そのために広いバス帯域幅が必要となり、さらに演算処理部ではメモリリソースの使用率が大きくなってしまう。このように、A

Ｉアクセラレータと演算処理部のような２つの演算部を用いてＤＮＮ演算を実行する従来の情報処理装置では、２つの演算部の間でのデータ通信量の増大によって様々な問題が生じる。なお、演算処理部には、例えば汎用のマイクロコンピュータ、プログラマブル集積回路、プログラマブル論理デバイス、システムオンチップなど、演算処理が可能な様々なデバイスを利用することが可能である。

[0005] ＡＩアクセラレータでは実行困難なＤＮＮ演算における処理の一つに、０パディングと呼ばれる処理がある。０パディングとは、ＤＮＮ演算の対象とされる特徴マップの端部分のデータの畳み込みを可能とするために、特徴マップの周囲に０データを挿入する処理のことである。すなわち、ＤＮＮ演算において畳み込み演算はベクトル同士の積和演算であるため、ＡＩアクセラレータを用いて実行することが可能であるが、畳み込み演算の前処理として必要な０パディングはベクトルを用いた演算ではないため、ＡＩアクセラレータでの実行が困難である。

[0006] ＤＮＮ演算での処理負荷の軽減に関して、特許文献１の技術が知られている。特許文献１には、特徴マップを格納したグローバルメモリと、重みデータを格納したローカルメモリと、演算処理回路とを備え、これらを用いて畳み込み演算を行うシステムが開示されている。このシステムでは、演算処理回路へのデータ送信速度がグローバルメモリよりもローカルメモリの方が速い場合に、演算処理回路で受信した入力特徴マップの同じ受信部分に対して複数の畳み込み演算を行うことで、メモリ回路網を往復するデータの移動を減らすようにしている。

先行技術文献

特許文献

[0007] 特許文献１：日本国特表２０２０－５０３５８４号公報

発明の概要

発明が解決しようとする課題

[0008] 特許文献1の技術は、畳み込み演算におけるデータの移動を低減する技術であり、0パディング処理のようにA1アクセラレータでは実行困難な処理を演算処理回路で実行する際のデータ転送量を削減するものではない。したがって、A1アクセラレータと演算処理部のような2つの演算部の間でのデータ通信量を削減することはできず、前述のような問題を解消するのは困難である。

課題を解決するための手段

[0009] 本発明による情報処理装置は、特徴マップに対してDNN演算を実行するものであって、第1の演算部と、前記第1の演算部とバスを介して接続された第2の演算部と、前記DNN演算で利用される重み情報を格納する記憶部と、を備え、前記第1の演算部は、前記特徴マップの一部である部分特徴マップと、前記部分特徴マップの前記特徴マップ中の位置を表す位置情報とを前記第2の演算部へ送信し、前記第2の演算部は、前記記憶部から前記重み情報を取得し、前記第1の演算部から送信される前記部分特徴マップおよび前記位置情報を受信し、前記位置情報に基づく所定の前処理を実行し、前記前処理の実行結果と前記部分特徴マップとを用いて、前記DNN演算を実行する。

本発明による情報処理方法は、第1の演算部と、前記第1の演算部とバスを介して接続された第2の演算部と、DNN演算で利用される重み情報を格納する記憶部と、を備える情報処理装置を用いて、特徴マップに対してDNN演算を実行するものであって、前記第1の演算部により、前記特徴マップの一部である部分特徴マップと、前記部分特徴マップの前記特徴マップ中の位置を表す位置情報を前記第2の演算部へ送信し、前記第2の演算部により、前記記憶部から前記重み情報を取得し、前記第1の演算部から送信される前記部分特徴マップおよび前記位置情報を受信し、前記位置情報に基づく所定の前処理を実行し、前記前処理の実行結果と前記部分特徴マップとを用いて、前記DNN演算を実行する。

発明の効果

[0010] 本発明によれば、2つの演算部を用いてDNN演算を行う情報処理装置において、2つの演算部の間でのデータ通信量を削減することができる。

図面の簡単な説明

[0011] [図1]本発明の一実施形態に係る情報処理装置の構成を示す図である。

[図2]情報処理装置の構成要素間で送受信される情報の流れを示す図である。

[図3]AIアクセラレータにおけるDNN演算部の機能構成を示す図である。

[図4]OPディング制御部により実行される処理の説明図である。

[図5]演算層情報の例を示す図である。

[図6]位置情報の例を示す図である。

[図7]DNN演算部の処理の流れを示すフローチャートである。

[図8]上端処理、中央処理および下端処理の流れを示すフローチャートである。

[図9]第1～第9領域に対する処理の流れを示すフローチャートである。

[図10]位置情報再演算処理の流れを示すフローチャートである。

[図11]重みレジスタと入力レジスタのデータ配置およびマスク箇所の一例を示す図である。

[図12]本発明の第2の実施形態に係る演算処理の説明図である。

[図13]本発明の第3の実施形態に係る演算処理の説明図である。

[図14]本発明の第4の実施形態に係るマスク処理が必要でない場合の重みレジスタと入力レジスタのデータ配置およびマスク箇所の一例を示す図である。

発明を実施するための形態

[0012] (第1の実施形態)

図1は、本発明の一実施形態に係る情報処理装置の構成を示す図である。本実施形態の情報処理装置1は、外部から入力される特徴マップに対してDNN演算を実行するものであり、図1に示すように、AIアクセラレータ10、演算処理部20および記憶部30を備えて構成される。

[0013] 演算処理部20は、例えばマイクロコンピュータを用いて構成されており

、所定の演算処理を実行する。演算処理部 20 には、不図示の外部装置から、DNN 演算の実行対象である特徴マップが入力される。例えば、車両に搭載されたカメラによって撮影された画像や、車両に搭載された L i D A R (Light Detection and Ranging) やレーダによって取得された距離情報などが、車両周辺の物体認識を行うための特徴マップとして演算処理部 20 に入力される。なお、記憶部 30 に記憶された特徴マップを演算処理部 20 に入力してもよい。

[0014] 記憶部 30 は、例えば H D D (Hard Disk Drive) や S S D (Solid State Drive) 等の不揮発性の記録媒体を用いて構成されており、情報処理装置 1 が行う DNN 演算において利用される情報が格納されている。記憶部 30 に格納されている情報には、例えば DNN 演算における畳み込み演算に使用される重み情報などが含まれる。

[0015] A I アクセラレータ 10 は、ベクトル型演算を得意とする A I 処理に特化した演算装置の一種であり、演算処理部 20 とバスを介して接続されている。A I アクセラレータ 10 には、演算処理部 20 に入力された特徴マップの一部が演算処理部 20 からバスを介して送信される。以下では、演算処理部 20 に入力される特徴マップを「全体特徴マップ」、演算処理部 20 から A I アクセラレータ 10 へ送信される特徴マップの一部を「部分特徴マップ」とそれぞれ称することで、これらを互いに区別する。また、部分特徴マップの全体特徴マップ中の位置を表す位置情報も、部分特徴マップとともに演算処理部 20 からバスを介して送信される。

[0016] A I アクセラレータ 10 は、DNN の各層に対応して N 個の DNN 演算部 11 を有している。各 DNN 演算部 11 は、A I アクセラレータ 10 において第 1 層から第 N 層まで順に接続されており、各層の DNN 演算を行うカーネル 13 (図 3 参照) を有している。

[0017] 第 1 層の DNN 演算部 11 には、演算処理部 20 から A I アクセラレータ 10 へ送信される部分特徴マップおよび位置情報が入力されるとともに、記憶部 30 から重み情報が入力される。第 1 層の DNN 演算部 11 は、これら

の入力データに基づいて第1層のDNN演算を実行し、その演算結果を次層（第2層）のDNN演算部11へ出力する。第2層のDNN演算部11は、前層（第1層）のDNN演算部11の演算結果と、記憶部30から入力される重み情報とに基づいて、第2層のDNN演算を実行し、その演算結果を次層のDNN演算部11へ出力する。これ以降の各層のDNN演算部11でも、同様の演算処理が行われる。なお、最終層（第N層）のDNN演算部11の演算結果は、AIアクセラレータ10による最終的なDNN演算の結果として、演算処理部20へ出力される。このようにして、各層のDNN演算部11が対応する層のDNN演算をそれぞれ行うことにより、AIアクセラレータ10においてDNN演算が実行される。

[0018] 図2は、情報処理装置1の構成要素であるAIアクセラレータ10、演算処理部20および記憶部30の間で送受信される情報の流れを示す図である。図2において、実線の矢印は、矢印の起点から終点に向けたデータの送信を表している。また、破線の矢印は、矢印の起点から終点に向けたデータのストリーム転送、すなわち送信側から受信側への連続したデータの流れを表している。

[0019] 情報処理装置1においてDNN演算の実行が開始されると、まず記憶部30からAIアクセラレータ10に向けて重み情報が送信される。続いて、演算処理部20からAIアクセラレータ10へ特徴マップの送信が開始される。このとき演算処理部20は、全体特徴マップを所定のデータサイズごとに分割することで複数の部分特徴マップを生成し、順次AIアクセラレータ10へ送信する。さらにこのとき、演算処理部20は、各部分特徴マップをAIアクセラレータ10へ送信するごとに、当該部分特徴マップの全体特徴マップ中での位置を表す位置情報を算出し、AIアクセラレータ10へ送信する。

[0020] AIアクセラレータ10は、演算処理部20から送信された部分特徴マップおよび位置情報を受信すると、事前に記憶部30から受信した重み情報を用いて、これらの入力データに基づくDNN演算を各層のDNN演算部11

において順次実行する。全ての層についてDNN演算の実行を終えたら、その演算結果を演算処理部20に出力する。

[0021] 図3は、AIアクセラレータ10におけるDNN演算部11の機能構成を示す図である。AIアクセラレータ10において、DNNの各層に対応するDNN演算部11は、ローカル記憶領域12およびカーネル13をそれぞれ備えて構成される。

[0022] ローカル記憶領域12は、DNN演算部11に入力された情報を一時的に記憶してカーネル13に受け渡す機能を有する。ローカル記憶領域12には、位置情報121、演算層情報122、重み情報123および部分特徴マップ124が格納される。位置情報121は、部分特徴マップ124が元の全体特徴マップにおいてどの位置にあるかを表す情報であり、演算処理部20または前層のDNN演算部11から入力される。演算層情報122は、当該DNN演算部11が実行するDNN演算の特徴を表す情報であり、ローカル記憶領域12において予め格納される。重み情報123は、当該DNN演算部11が実行するDNN演算の畳み込み演算において用いられる情報であり、記憶部30から入力される。部分特徴マップ124は、当該DNN演算部11の演算対象データであり、演算処理部20または前層のDNN演算部11から入力される。

[0023] カーネル13は、0パディング制御部131、次カーネル位置情報生成部132および畳み込み演算部133を備える。畳み込み演算部133は、重みレジスタ301、入力レジスタ302、積和演算部303および出力レジスタ304を備える。

[0024] 0パディング制御部131は、ローカル記憶領域12からロードされて重みレジスタ301に格納された重み情報123に対して、位置情報121に基づく0パディング相当の処理を行う。ここで行われる0パディング相当の処理とは、位置情報121により表される位置において部分特徴マップ124が全体特徴マップの端部を含む場合に、その端部に対する畳み込み演算を実行可能とするため、重み情報123において位置情報121に対応する箇

所を0データでマスクする処理のことである。すなわち、0パディング制御部131が重みレジスタ301に対して0パディング相当の処理を行うことにより、重みレジスタ301に格納された重み情報123の一部が0データに置き換えられ、マスク済み重み情報が生成される。

[0025] 畳み込み演算部133は、ローカル記憶領域12から、重み情報123と部分特徴マップ124を重みレジスタ301と入力レジスタ302にそれぞれロードする。積和演算部303は、重みレジスタ301に格納されて0パディング制御部131により0パディング相当の処理が行われたマスク済み重み情報と、入力レジスタ302に格納された部分特徴マップとを用いて、部分特徴マップの各データ値に対する畳み込み演算（積和演算）を実行する。積和演算部303の演算結果は出力レジスタ304に格納され、出力レジスタ304から次層のDNN演算部11または演算処理部20に出力される。

[0026] 出力レジスタ304より出力された演算結果が次層のDNN演算部11に入力されると、次層のDNN演算部11は、その演算結果を部分特徴マップ124としてローカル記憶領域12に格納する。そして、この部分特徴マップ124を演算対象データとして、畳み込み演算部133により畳み込み演算を実行する。

[0027] 次カーネル位置情報生成部132は、ローカル記憶領域12に格納された位置情報121および演算層情報122に基づいて、次層のDNN演算部11における部分特徴マップ124として出力レジスタ304から送信される畳み込み演算結果が、元の全体特徴マップにおいてどの位置にあるかを判別する。そして、この判別結果に基づいて位置情報を生成し、次層のDNN演算部11へ出力する。

[0028] 図4は、0パディング制御部131により実行される処理の説明図である。図4では、演算処理部20において全体特徴マップ401が縦方向と横方向にそれぞれ5分割された各部分特徴マップのうち、全体特徴マップ401の右上端部分に位置する部分特徴マップ124を演算対象とするDNN演算

部 1 1 において、0パディング制御部 1 3 1 によって0パディング相当の処理を行う場合の例を示している。なお、図 4 の例では、部分特徴マップ 1 2 4 に対して 3×3 の重みデータで構成される重み情報 1 2 3 を用いて、1 データごとのストライドにより畳み込み演算を行うものとしている。

[0029] 図 4 に示すように、畳み込み演算時に 3×3 の重みデータが全体特徴マップ 4 0 1 上を移動する範囲の領域を畳み込み演算領域 4 0 2 とすると、この畳み込み演算領域 4 0 2 は、全体特徴マップ 4 0 1 が有する上下左右の端部のうちどの部分が含まれるかにより、9 種類の領域に分けられる。具体的には、畳み込み演算領域 4 0 2 は、左上角部を含む第 1 領域と、上端部を含む第 2 領域と、右上角部を含む第 3 領域と、左端部を含む第 4 領域と、いずれの端部も含まない第 5 領域と、右端部を含む第 6 領域と、左下角部を含む第 7 領域と、下端部を含む第 8 領域と、右下角部を含む第 9 領域と、に分けられる。

[0030] ここで、全体特徴マップ 4 0 1 を分割した各部分特徴マップのうち、例えば前述のように全体特徴マップ 4 0 1 の右上に位置する部分特徴マップ 1 2 4 が入力された場合、DNN 演算部 1 1 は、この部分特徴マップ 1 2 4 に対して、 3×3 の重みデータを左から右に向かってストライドに応じた移動距離ずつ移動させながら、各位置での畳み込み演算を行う。そして、重みデータが部分特徴マップ 1 2 4 の右端まで到達してその位置での畳み込み演算が終了すると、重みデータをストライド分の移動距離だけ下方方向にずらした後、再び左から右に向かって畳み込み演算を開始する。このようにして、部分特徴マップ 1 2 4 の全ての位置に対して畳み込み演算が行われることにより、DNN 演算の特定の層に対応する演算が DNN 演算部 1 1 において実行される。

[0031] DNN 演算部 1 1 において0パディング制御部 1 3 1 は、位置情報 1 2 1 に基づいて、畳み込み演算が行われる部分特徴マップ 1 2 4 の位置ごとに、その畳み込み演算の範囲が前述の第 1 ～第 9 領域のうちどの領域に属するかを判定する。例えば、部分特徴マップ 1 2 4 において最初に畳み込み演算が

行われる左上の部分は第2領域に属し、右上の部分は第3領域に属すると判定される。同様に、部分特徴マップ124の各位置について領域判定が行われることにより、符号403に示すような領域判定結果が得られる。この領域判定結果403では、部分特徴マップ124の各位置が第1～第9領域のうちどの領域に属するかを丸数字で表している。

[0032] 上記のようにして部分特徴マップ124の各位置に対する領域判定結果が得られたら、0パディング制御部131は、この領域判定結果に基づいて、重みレジスタ301に格納された重み情報123に対するマスク箇所を決定する。ここでは、部分特徴マップ124とマスク後の重み情報123との畳み込み演算結果が、部分特徴マップ124の端部分に対して1行1列分の0データを付加する0パディング処理を行った後に重み情報123で畳み込み演算を行ったときの演算結果と一致するように、マスク箇所を決定する。そして、重みレジスタ301において決定したマスク箇所の重み情報123の値を0とすることにより、重みレジスタ301に対して、0パディング相当の処理としてのマスク処理を実行する。こうしてマスク処理が実行された後の重みレジスタ301におけるマスク済み重み情報を生成する。

[0033] なお、第5領域に属する位置では0パディング処理が不要である。そのため、第5領域に属する位置において畳み込み演算が行われるときには、0パディング制御部131によるマスク処理を実行する必要がない。

[0034] 図5は、各DNN演算部11においてローカル記憶領域12に格納される演算層情報122の例を示す図である。演算層情報122は、例えばサイズ変化率501、重みサイズ502、ストライド503、パディング幅504、入力特徴マップ幅505、入力特徴マップ高さ506、入力特徴マップチャンネル数507、出力特徴マップチャンネル数508の各データにより構成される。これらのデータは、当該DNN演算部11が行う演算の内容に対応して、それぞれの値が設定される。

[0035] サイズ変化率501は、演算前後での部分特徴マップのサイズの変化率を表している。重みサイズ502は重みデータの一辺のサイズを表し、ストラ

イド503は部分特徴マップに対して重み付け演算を行う間隔を表し、パディング幅504は0パディング時に部分特徴マップに付加する0データの一边あたりの幅を表している。入力特徴マップ幅505および入力特徴マップ高さ506は、演算対象とする部分特徴マップの水平方向と垂直方向のデータ数をそれぞれ表している。入力特徴マップチャンネル数507および出力特徴マップチャンネル数508は、演算前後での部分特徴マップの深さ方向のデータ数をそれぞれ表している。

[0036] 図6は、各DNN演算部11においてローカル記憶領域12に格納される位置情報121の例を示す図である。位置情報121は、例えば分割ID601、全体特徴マップ幅602、全体特徴マップ高さ603、演算層604、開始x座標605、開始y座標606、終点x座標607、終点y座標608、横畳み込み数609、縦畳み込み数610の各データにより構成される。これらのデータは、当該DNN演算部11に入力されて演算対象とされる部分特徴マップの特徴に対応して、それぞれの値が設定される。

[0037] 分割ID601は、当該DNN演算部11において部分特徴マップをさらに複数に分割して処理する場合に、各分割部分特徴マップをそれぞれ識別するために付与されるIDを表す情報である。なお、部分特徴マップを複数に分割して処理する場合については、後で第2の実施形態として詳しく説明する。

[0038] 全体特徴マップ幅602および全体特徴マップ高さ603は、部分特徴マップに分割される前の元の全体特徴マップの水平方向と垂直方向のデータ数をそれぞれ表している。

[0039] 演算層604は、当該DNN演算部11において実行される演算が対応するDNNの層を表す情報である。

[0040] 開始x座標605および開始y座標606は、部分特徴マップにおいて畳み込み演算が開始される左上の点の全体特徴マップ上での座標値をそれぞれ表している。終点x座標607および終点y座標608は、部分特徴マップにおいて畳み込み演算が最後に行われる右下の点の全体特徴マップ上での座

標値をそれぞれ表している。横畳み込み数609および縦畳み込み数610は、部分特徴マップに対して行われる畳み込み演算の横方向と縦方向での回数をそれぞれ表している。

[0041] 次に、A1アクセラレータ10におけるDNN演算部11の処理の流れについて説明する。図7は、DNN演算部11の処理の流れを示すフローチャートである。図7のフローチャートに示す処理は、A1アクセラレータ10が有する各DNN演算部11において実行される。

[0042] ステップS701では、演算処理部20または前層のDNN演算部11から入力されてローカル記憶領域12に格納された位置情報121および部分特徴マップ124に基づき、演算対象の部分特徴マップが全体特徴マップの上端部分を含むか否かを判定する。上端部分を含む場合は、ステップS702で図8(a)に示す上端処理を行った後にステップS703へ進む。上端部分を含まない場合は、ステップS702の上端処理を実行せずにステップS703へ進む。

[0043] ステップS703では、演算対象の部分特徴マップに対する垂直方向の畳み込み演算の回数分だけステップS704の処理を繰り返し実行し、その後にステップS705へ進む。ステップS704では、図8(b)に示す中央処理を行う。

[0044] ステップS705では、位置情報121および部分特徴マップ124に基づき、演算対象の部分特徴マップが全体特徴マップの下端部分を含むか否かを判定する。下端部分を含む場合は、ステップS706で図8(c)に示す下端処理を行った後にステップS707へ進む。下端部分を含まない場合は、ステップS706の下端処理を実行せずにステップS707へ進む。

[0045] ステップS707では、次カーネル位置情報生成部132により、図10に示す位置情報再演算処理を行う。ステップS707の位置情報再演算処理を終えたら、図7のフローチャートに示す処理を終了する。

[0046] 図8は、図7のステップS702、S704、S706でそれぞれ実行される上端処理、中央処理および下端処理の流れを示すフローチャートである

。

[0047] 図8(a)に示す上端処理において、ステップS711では、位置情報121および部分特徴マップ124に基づき、演算対象の部分特徴マップが全体特徴マップの左端部分を含むか否かを判定する。左端部分を含む場合は、ステップS712で第1領域に対する処理を行った後にステップS713へ進む。左端部分を含まない場合は、ステップS712の処理を実行せずにステップS713へ進む。

[0048] ステップS713では、演算対象の部分特徴マップに対する水平方向の畳み込み演算の回数分だけステップS714の処理を繰り返し実行し、その後にステップS715へ進む。ステップS714では、第2領域に対する処理を行う。

[0049] ステップS715では、位置情報121および部分特徴マップ124に基づき、演算対象の部分特徴マップが全体特徴マップの右端部分を含むか否かを判定する。右端部分を含む場合は、ステップS716で第3領域に対する処理を行った後に、図8(a)のフローチャートに示す上端処理を終了する。右端部分を含まない場合は、ステップS716の処理を実行せずに上端処理を終了する。

[0050] 図8(b)に示す中央処理において、ステップS721では、位置情報121および部分特徴マップ124に基づき、演算対象の部分特徴マップが全体特徴マップの左端部分を含むか否かを判定する。左端部分を含む場合は、ステップS722で第4領域に対する処理を行った後にステップS723へ進む。左端部分を含まない場合は、ステップS722の処理を実行せずにステップS723へ進む。

[0051] ステップS723では、演算対象の部分特徴マップに対する水平方向の畳み込み演算の回数分だけステップS724の処理を繰り返し実行し、その後にステップS725へ進む。ステップS724では、第5領域に対する処理を行う。

[0052] ステップS725では、位置情報121および部分特徴マップ124に基

づき、演算対象の部分特徴マップが全体特徴マップの右端部分を含むか否かを判定する。右端部分を含む場合は、ステップS 7 2 6で第6領域に対する処理を行った後に、図8（b）のフローチャートに示す中央処理を終了する。右端部分を含まない場合は、ステップS 7 2 6の処理を実行せずに中央処理を終了する。

[0053] 図8（c）に示す下端処理において、ステップS 7 3 1では、位置情報1 2 1および部分特徴マップ1 2 4に基づき、演算対象の部分特徴マップが全体特徴マップの左端部分を含むか否かを判定する。左端部分を含む場合は、ステップS 7 3 2で第7領域に対する処理を行った後にステップS 7 3 3へ進む。左端部分を含まない場合は、ステップS 7 3 2の処理を実行せずにステップS 7 3 3へ進む。

[0054] ステップS 7 3 3では、演算対象の部分特徴マップに対する水平方向の畳み込み演算の回数分だけステップS 7 3 4の処理を繰り返し実行し、その後にステップS 7 3 5へ進む。ステップS 7 3 4では、第8領域に対する処理を行う。

[0055] ステップS 7 3 5では、位置情報1 2 1および部分特徴マップ1 2 4に基づき、演算対象の部分特徴マップが全体特徴マップの右端部分を含むか否かを判定する。右端部分を含む場合は、ステップS 7 3 6で第9領域に対する処理を行った後に、図8（c）のフローチャートに示す下端処理を終了する。右端部分を含まない場合は、ステップS 7 3 6の処理を実行せずに下端処理を終了する。

[0056] 図9は、図8のステップS 7 1 2、S 7 1 4、S 7 1 6、S 7 2 2、S 7 2 4、S 7 2 6、S 7 3 2、S 7 3 4、S 7 3 6においてそれぞれ実行される第1～第9領域に対する処理の流れを示すフローチャートである。

[0057] 図9（a）のフローチャートにおいて、ステップS 8 0 1では、畳み込み演算部1 3 3により、ローカル記憶領域1 2に格納されている重み情報1 2 3と部分特徴マップ1 2 4を読み出し、重みレジスタ3 0 1と入力レジスタ3 0 2にそれぞれ格納する。

- [0058] ステップS 8 0 2では、0パディング制御部1 3 1により、領域の種類に応じたマスク位置を決定するマスク位置決定処理を行う。ここでは、図9（b）のフローチャートに示す手順に従ってマスク位置決定処理が実行される。
- [0059] ステップS 8 0 3では、畳み込み演算部1 3 3により、ステップS 8 0 2のマスク位置決定処理で決定されたマスク位置に従って、重みレジスタ3 0 1に対するマスク処理を行う。ここでは前述のように、重みレジスタ3 0 1に格納された重み情報1 2 3においてマスク位置の値を0データでマスクすることにより、0パディング相当の処理としてマスク処理を行う。ただし、0パディング処理が不要な第5領域については、後述するようにマスク位置が設定されていない。そのため、ステップS 7 2 4の第5領域の処理では、ステップS 8 0 2でマスク位置が出力されず、そのためステップS 8 0 3においてマスク処理が行われない。
- [0060] ステップS 8 0 4では、畳み込み演算部1 3 3により、積和演算部3 0 3を用いて、ステップS 8 0 3のマスク処理によって生成されたマスク済み重み情報、すなわちマスク処理後の重みレジスタ3 0 1の重み情報1 2 3と、入力レジスタ3 0 2に格納された部分特徴マップ1 2 4との積和演算（畳み込み演算）を行う。その後、出力レジスタ3 0 4から演算結果を出力し、図9（a）のフローチャートに示す処理を終了する。
- [0061] 図9（b）に示すマスク位置決定処理において、ステップS 8 1 1では、演算対象の部分特徴マップが属する領域を取得する。ここでは、図8のステップS 7 1 2、S 7 1 4、S 7 1 6、S 7 2 2、S 7 2 4、S 7 2 6、S 7 3 2、S 7 3 4、S 7 3 6のどの処理でマスク位置決定処理が実行されているかにより、第1～第9領域のうちどの領域に属するかを判定する。
- [0062] ステップS 8 1 2では、0パディング制御部1 3 1により、ステップS 8 1 1で取得した領域でのマスク位置を照会する。ここでは、第1～第9領域の各領域に対して予め設定されたマスク位置のうち、ステップS 8 1 1で取得した領域に対応するマスク位置を照会することで、当該領域でのマスク位

置を決定する。なお、０パディング制御部１３１では、第１～第９領域の各領域について、部分特徴マップ１２４とマスク後の重み情報１２３との畳み込み演算結果が、各領域に応じた部分特徴マップ１２４の端部に対して０パディング処理を行った後に重み情報１２３で畳み込み演算を行ったときの演算結果と一致するように、マスク位置が予め設定されているものとする。ただし、０パディング処理が不要な第５領域については、マスク位置が設定されていない。

[0063] ０パディング制御部１３１は、ステップＳ８１２の処理において、上記のように予め設定されたマスク位置の情報を照会して利用することで、０パディング処理に相当するマスク位置を領域ごとに決定することができる。これにより、各領域における全体特徴マップの端部の位置に応じて、重み情報１２３においてマスクする箇所を正しく決定することができる。

[0064] ステップＳ８１３では、ステップＳ８１２で決定したマスク位置を０パディング制御部１３１から畳み込み演算部１３３へ出力する。ステップＳ８１３の処理を終えたら、図９（ｂ）のフローチャートに示すマスク位置決定処理を終了する。

[0065] 図１０は、図７のステップＳ７０７で実行される位置情報再演算処理の流れを示すフローチャートである。

[0066] 図１０（ａ）に示す位置情報再演算処理において、ステップＳ９０１では、演算層情報１２２に含まれるサイズ変化率５０１の値に基づき、畳み込み演算部１３３が行う畳み込み演算（積和演算）によって部分特徴マップの大きさが変化するか否かを判定する。サイズ変化率５０１の値が１以外の値であれば、部分特徴マップの水平方向および垂直方向の少なくとも一方の大きさが変化すると判定し、ステップＳ９０２へ進む。サイズ変化率５０１の値が１であれば、部分特徴マップの大きさが変化しないと判定し、ステップＳ９０２の処理を実行せずにステップＳ９０３へ進む。

[0067] ステップＳ９０２では、次カーネル位置情報生成部１３２により、演算層情報１２２に基づいて位置情報１２１を再計算する。ここでは、演算層情報

1 2 2に含まれる各データ値に基づいて、位置情報1 2 1に含まれる各データ値を再計算する。

[0068] ステップS 9 0 2では、例えば図1 0 (b)に示す(1)～(7)の各計算式により、位置情報1 2 1の各データ値が再計算される。式(1)～(7)において、WW、WHは全体特徴マップ幅6 0 2、全体特徴マップ高さ6 0 3の値をそれぞれ表す。また、x、yは開始x座標6 0 5、開始y座標6 0 6の値をそれぞれ表し、endx、endyは終点x座標6 0 7、終点y座標6 0 8の値をそれぞれ表す。また、OW、OHは横畳み込み数6 0 9、縦畳み込み数6 1 0の値をそれぞれ表す。さらに、IW、IHは演算層情報1 2 2の入力特徴マップ幅5 0 5、入力特徴マップ高さ5 0 6の値をそれぞれ表し、k、k+1はDNN演算の全体における現在の演算層と次の演算層の位置(何層目か)をそれぞれ表し、KS、strideは演算層情報1 2 2の重みサイズ5 0 2、ストライド5 0 3の値をそれぞれ表す。

[0069] 式(1)～(7)により、WW、WH、x、yは、現在の位置情報1 2 1におけるこれらの値に、演算層情報1 2 2のサイズ変化率5 0 1の値をそれぞれ乗算し、小数点以下の値を切り上げることで算出される。また、endx、endyは、次層のx、yにIW、IHをそれぞれ加えて算出される。また、OW、OHは、IW、IHからKSをそれぞれ減算した値をストライドで除算し、その切り捨て値に1を加えて算出される。

[0070] ステップS 9 0 3では、次カーネル位置情報生成部1 3 2により、次層のDNN演算部1 1に対する位置情報を作成する。ここでは、ローカル記憶領域1 2に格納された位置情報1 2 1、またはこれにステップS 9 0 2で再計算された値が反映された位置情報において、演算層6 0 4の値をインクリメントして次層のDNN演算部1 1に対応する値とすることで、次層の位置情報を作成する。

[0071] ステップS 9 0 4では、ステップS 9 0 3で作成した位置情報を次カーネル位置情報生成部1 3 2から次層のDNN演算部1 1へ送信する。ステップS 9 0 4の処理を終えたら、図1 0のフローチャートに示す位置情報再演算

処理を終了する。

[0072] A I アクセラレータ 10 では、各 DNN 演算部 11 において上記の処理がそれぞれ実行されることにより、演算処理部 20 から入力された部分特徴マップの各位置に対して、全体特徴マップの畳み込み演算領域内に設定された各領域の種類に応じた処理がそれぞれ行われる。これにより、A I アクセラレータ 10 において 0 パディング相当の処理を行ってマスク済み重み情報を生成し、このマスク済み重み情報を用いて DNN の各層に対応する畳み込み演算を実行することができる。

[0073] 図 11 は、重みレジスタ 301 と入力レジスタ 302 のデータ配置およびマスク箇所の一例を示す図である。図 11 では、前述のように重み情報 123 が 3×3 の重みデータで構成されているときに、重みレジスタ 301 に一度に入力 4 チャンネル分、出力 8 チャンネル分の合計 32 個の重みデータが格納されている例を示している。この場合、図 11 に示すように、例えば重みレジスタ 301 には、32 個の重みデータにおける横 3 データ分の値がそれぞれ格納される。図 11 では、重み情報 123 を構成する 32 個の重みデータにおける 3×3 のデータ値のうち、0 ~ 95 の数字が付された箇所のデータ値が重みレジスタ 301 にそれぞれ格納されている。これらの数字は、重みレジスタ 301 の各格納箇所に付された数字と対応している。なお、重みレジスタ 301 中で 96 ~ 127 の数字が付された格納箇所には、重みのデータ値が入っていない。

[0074] また、図 11 に示すように、例えば入力レジスタ 302 には、ローカル記憶領域 12 からロードされた 4 チャンネル分の部分特徴マップ 124 のうち、0 ~ 27 の数字が付された部分のデータ値がそれぞれ格納される。これらの数字も重みレジスタ 301 と同様に、入力レジスタ 302 の各格納箇所に付された数字と対応している。なお、入力レジスタ 302 において「x」が付された先頭の 4 つの格納箇所には、例えば他の格納箇所のうちいずれかと同じデータ値など、任意のデータ値を格納することができる。

[0075] 図 11 に示した重みレジスタ 301 の各データ値と入力レジスタ 302 の

各データ値との畳み込み演算を行う場合に、例えば第4領域に属する位置での畳み込み演算時には、符号1101に示す範囲、すなわち「set0」の行に対応する0～31の数字が付された各格納箇所のデータ値に対してマスク処理が行われる。これにより、重み情報123に対して0パディング相当の重みデータを設定することが可能となる。

[0076] 以上説明した本発明の第1の実施形態によれば、以下の作用効果を奏する。

[0077] (1) 情報処理装置1は、特徴マップに対してDNN演算を実行するものであって、演算処理部20と、演算処理部20とバスを介して接続されたAIアクセラレータ10と、DNN演算で利用される重み情報を格納する記憶部30とを備える。演算処理部20は、特徴マップの一部である部分特徴マップ124と、部分特徴マップ124の特徴マップ中の位置を表す位置情報121とをAIアクセラレータ10へ送信する。AIアクセラレータ10は、記憶部30から重み情報123を取得し、演算処理部20から送信される部分特徴マップ124および位置情報121を受信する。そして、位置情報121に基づく所定の前処理を実行し（ステップS811～S813、S803）、その前処理の実行結果と部分特徴マップ124とを用いて、DNN演算を実行する（ステップS804）。このようにしたので、AIアクセラレータ10および演算処理部20の2つの演算部を用いてDNN演算を行う情報処理装置1において、AIアクセラレータ10により0パディング処理に相当する前処理を行うことができるため、これら2つの演算部の間でのデータ通信量を削減することができる。

[0078] (2) AIアクセラレータ10が実行する前処理は、重み情報123において位置情報121に対応する箇所をマスクする処理である。AIアクセラレータ10は、前処理を行うことで生成したマスク済み重み情報と部分特徴マップ124とを用いて、DNN演算としての積和演算をステップS804において実行する。このようにしたので、AIアクセラレータ10において実行可能な処理として、0パディング処理に相当する前処理を実現できる。

[0079] (3) A I アクセラレータ 10 は、位置情報 121 に基づいて、特徴マップに設定された複数の領域のうちどの領域に部分特徴マップ 124 が属するかを判定し（ステップ S701～S707、S711～S716、S721～S726、S731～S736）、当該領域に対応する箇所をマスクする処理を、ステップ S811～S813、S803 において前処理として実行する。具体的には、特徴マップに設定された複数の領域は、図 4 において示すように、特徴マップ 401 の左上角部を含む第 1 領域と、特徴マップ 401 の上端部を含む第 2 領域と、特徴マップ 401 の右上角部を含む第 3 領域と、特徴マップ 401 の左端部を含む第 4 領域と、特徴マップ 401 のいずれの端部も含まない第 5 領域と、特徴マップ 401 の右端部を含む第 6 領域と、特徴マップ 401 の左下角部を含む第 7 領域と、特徴マップ 401 の下端部を含む第 8 領域と、特徴マップ 401 の右下角部を含む第 9 領域と、を有する。A I アクセラレータ 10 は、部分特徴マップ 124 が第 5 領域以外の各領域に属する場合は、ステップ S811、S812 において、当該領域における特徴マップの端部の位置に応じて、重み情報 123 においてマスクする箇所を決定し、部分特徴マップ 124 が第 5 領域に属する場合は、ステップ S803 においてマスクする処理を実行しない。このようにしたので、0 パディング処理に相当する前処理を確実に行うことができる。

[0080] (4) A I アクセラレータ 10 は、ベクトル演算に特化した複数の演算コアを有しており、これら複数の演算コアを用いた A I 処理によって DNN 演算を実行する。このようにしたので、DNN 演算を高速に実行することができる。

[0081] (5) A I アクセラレータ 10 は、DNN の各層に対応する複数のカーネル 13 を有している。DNN 演算において複数のカーネル 13 のいずれかで部分特徴マップ 124 の水平方向および垂直方向の少なくとも一方の大きさが変化する場合（ステップ S901：Yes）、当該カーネル 13 は、位置情報 121 を再計算して（ステップ S902）次層に対応するカーネル 13 へ送信する（ステップ S904）。このようにしたので、複数の層によって構

成されるDNNを利用したDNN演算において、AIアクセラレータ10による0パディング処理に相当する前処理を適用することができる。

[0082] (第2の実施形態)

次に、本発明の第2の実施形態について説明する。本実施形態では、AIアクセラレータ10のいずれかのDNN演算部11において、部分特徴マップ124を垂直方向に複数に分割し、当該DNN演算部11が有する複数のカーネル13で1層分の演算を実行する例を説明する。

[0083] ここで、本実施形態のAIアクセラレータ10では、いずれか少なくとも1つのDNN演算部11について、第1の実施形態で説明した構成とは異なり、当該DNN演算部11がDNNの1層分の演算を行う際の部分特徴マップ124の分割数だけ、カーネル13を持つこととなる。なお、本実施形態における情報処理装置1の構成、AIアクセラレータ10の構成、カーネル13の個数以外のDNN演算部11の構成およびカーネル13の構成は、第1の実施形態において図1、図3で説明したものとそれぞれ同様である。そのため以下では、これらの説明を省略する。

[0084] 図12は、本発明の第2の実施形態に係る演算処理の説明図である。図12では、3層分の演算のうち2層目の中間層において、部分特徴マップ124が3つに分割されて演算が行われる例を示している。この場合、3層分のDNN演算部11は、図12に示すように、1層目のDNN演算部11が有する分割カーネル1201と、2層目のDNN演算部11が有する3つの中間カーネル1202と、3層目のDNN演算部11が有するマージカーネル1203とを含んで構成される。これらのカーネルは、いずれも図3に示した第1の実施形態におけるカーネル13と同様の構成を有している。

[0085] 図12において、分割カーネル1201は、1層目の処理を行った後、部分特徴マップ124を3分割することで3つの分割部分特徴マップ1204を生成し、次層のDNN演算部11に送る。中間カーネル1202は、分割カーネル1201から送られた分割部分特徴マップ1204をそれぞれ処理して次層のDNN演算部11に送る。マージカーネル1203は、3つの中

間カーネル 1202 でそれぞれ処理された分割部分特徴マップ 1204 をマージしたのち、3 層目の処理を行う。

[0086] 分割カーネル 1201 は、3 つの分割部分特徴マップ 1204 を生成する際に、それぞれの分割部分特徴マップ 1204 に対応する位置情報 121 を作成する。その際には、図 6 に示した位置情報 121 の各データのうち分割 ID 601 の値を、開始 x 座標が小さい方から 0, 1, 2 と順に設定する。また、終点 x 座標 607、終点 y 座標 608 の値を、各分割部分特徴マップ 1204 に対応する値へとそれぞれ変更する。こうして作成した位置情報 121 は、分割部分特徴マップ 1204 とともに各分割カーネル 1201 へ送信される。

[0087] 中間カーネル 1202 は、各分割部分特徴マップ 1204 に対して、第 1 の実施形態と同様の処理をそれぞれ行う。

[0088] マージカーネル 1203 は、各中間カーネル 1202 から処理済みの分割部分特徴マップ 1204 とともに位置情報 121 を受け取ると、分割部分特徴マップ 1204 をマージして 1 つの部分特徴マップ 124 に戻す。このとき、分割 ID 601 の値が最も小さい位置情報 121 の開始 x 座標 605 および開始 y 座標 606 の値と、分割 ID 601 の値が最も大きい位置情報 121 の終点 x 座標 607 および終点 y 座標 608 の値とを用いて、マージ後の部分特徴マップ 124 に対する位置情報 121 を更新する。そして、マージ後の部分特徴マップ 124 と更新後の位置情報 121 を用いて、3 層目に対応する畳み込み演算を行う。

[0089] (第 3 の実施形態)

次に、本発明の第 3 の実施形態について説明する。本実施形態では、AI アクセラレータ 10 のいずれかの DNN 演算部 11 において、部分特徴マップ 124 を続けて複数回入力し、次層の DNN 演算部 11 に対して、複数の部分特徴マップ 124 の演算結果をまとめて一度に送信する例を説明する。

[0090] なお、本実施形態における情報処理装置 1 の構成、AI アクセラレータ 10 の構成、DNN 演算部 11 の構成およびカーネル 13 の構成は、第 1 の実

施形態において図 1、図 3 で説明したものとそれぞれ同様である。そのため以下では、これらの説明を省略する。

[0091] 図 1 3 は、本発明の第 3 の実施形態に係る演算処理の説明図である。図 1 3 の例では、3 層分の演算を行う各 DNN 演算部 1 1 のカーネル 1 3 を、カーネル 0、カーネル 1、カーネル 2 でそれぞれ示している。この場合、図 1 3 (b) に示すように、カーネル 0 は第 1 の実施形態と同様に、演算処理済みの部分特徴マップ 1 2 4 と、その部分特徴マップ 1 2 4 の位置情報 1 2 1 とを、カーネル 1 に送信する。カーネル 1 は、第 1 の実施形態と同様に部分特徴マップ 1 2 4 の処理を行うが、次層のカーネル 2 の演算に必要なデータが揃わない場合は、部分特徴マップ 1 2 4 の演算結果とその部分特徴マップ 1 2 4 に対応する位置情報 1 2 1 とを、カーネル 2 へは送信しない。カーネル 0 がカーネル 1 に対して 2 回目のデータ送信を行い、そのデータを用いてカーネル 1 が処理した部分特徴マップ 1 2 4 と、1 度目に処理した部分特徴マップ 1 2 4 とにより、カーネル 2 での演算に必要なデータが揃ったときに、カーネル 1 はカーネル 2 に対して、部分特徴マップ 1 2 4 と位置情報 1 2 1 の送信を行う。

[0092] 上記の演算手順を実現するため、AI アクセラレータ 1 0 の各層のカーネル 1 3 (カーネル 0 ~カーネル 2) では、当該カーネルでの演算実行回数を表すカーネル実行回数と、当該カーネルから次層のカーネルへのデータ出力に必要な動作回数を表す出力数とを保持している。そして、これらの情報に基づいて図 1 3 (a) のフローチャートに示す処理を実行することで、各カーネルにおいて部分特徴マップ 1 2 4 と位置情報 1 2 1 の出力タイミング制御を行うようにしている。

[0093] 図 1 3 (a) のフローチャートにおいて、ステップ S 1 3 0 1 では、保持している実行回数に 1 を加算する。なお、実行回数の初期値は 0 である。

[0094] ステップ S 1 3 0 2 では、保持している実行回数を出力数で除算したときの剰余を求め、その剰余が 0 であるか否かを判定する。剰余が 0 であればステップ S 1 3 0 3 へ進み、0 でなければ、ステップ S 1 3 0 3 の処理を実行

せずに図13(a)のフローチャートに示す処理を終了する。

[0095] ステップS1303では、位置情報121を更新し、更新後の位置情報121と部分特徴マップ124を次層のカーネルへと出力する。ステップS1303の処理を実行したら、実行回数を初期値の0にリセットした後、図13(a)のフローチャートに示す処理を終了する。

[0096] (第4の実施形態)

次に、本発明の第4の実施形態について説明する。本実施形態では、処理対象の部分特徴マップ124や重み情報123における各データの並びによっては、必ずしもマスク処理が必要ではない場合について説明する。

[0097] なお、本実施形態における情報処理装置1の構成、AIアクセラレータ10の構成、DNN演算部11の構成およびカーネル13の構成は、第1の実施形態において図1、図3で説明したものとそれぞれ同様である。そのため以下では、これらの説明を省略する。

[0098] 図14は、本発明の第4の実施形態に係るマスク処理が必要でない場合の重みレジスタ301と入力レジスタ302のデータ配置の一例を示す図である。図14では図11と同様に、重み情報123が3×3の重みデータで構成されているときに、重みレジスタ301に一度に入力4チャンネル分、出力8チャンネル分の合計32個の重みデータが格納されている例を示している。この場合に、第1領域に属する位置での畳み込み演算時には、3×3の重みデータの最初の1行目の処理の際に、符号1401に示す範囲に対してマスク処理が行われる。すなわち、重みレジスタ301において重みデータが格納されている全ての格納箇所に対して、マスク処理が行われることになる。

[0099] 上記のように、重みレジスタ301に格納された全ての重みデータに対してマスク処理が行われる場合は、マスク処理を省略し、畳み込み演算自体を行わずに単に計算結果を0のデータとしても、マスク処理後の畳み込み演算と同じ結果が得られる。このようにすれば、不要な演算処理を省略して高速化を図ることが可能となる。なお、ここでは第1領域に属する位置での畳み

込み演算時の例を述べたが、他の領域に属する位置でも同様に、重みレジスタ 301 に格納された全ての重みデータに対してマスク処理が行われる場合には、マスク処理と畳み込み演算を省略することができる。

[0100] 以上説明した本発明の第2～第4の各実施形態によれば、第1の実施形態と同様の作用効果をそれぞれ奏する。

[0101] なお、以上説明した各実施形態では、A1アクセラレータ10において、0パディング処理に相当する前処理として、重み情報123において位置情報121に対応する箇所をマスクする処理を行う例を説明したが、A1アクセラレータ10において0パディング処理を行い、0パディング処理済みの部分特徴マップ124を用いてDNN演算を実行するようにしてもよい。このようにしても、第1の実施形態で説明したのと同様の効果を奏することができる。

[0102] 以上説明した実施形態や各種変形例はあくまで一例であり、発明の特徴が損なわれない限り、本発明はこれらの内容に限定されるものではない。また、各実施形態や各種変形例は、単独で採用してもよいし、任意に組み合わせてもよい。さらに、上記では種々の実施形態や変形例を説明したが、本発明はこれらの内容に限定されるものではない。本発明の技術的思想の範囲内で考えられるその他の態様も本発明の範囲内に含まれる。

符号の説明

[0103] 1：情報処理装置、10：A1アクセラレータ、11：DNN演算部、12：ローカル記憶領域、13：カーネル、20：演算処理部、30：記憶部、121：位置情報、122：演算層情報、123：重み情報、124：部分特徴マップ、131：0パディング制御部、132：次カーネル位置情報生成部、133：畳み込み演算部、301：重みレジスタ、302：入力レジスタ、303：積和演算部、304：出力レジスタ

請求の範囲

- [請求項1] 特徴マップに対してDNN演算を実行する情報処理装置であって、
第1の演算部と、
前記第1の演算部とバスを介して接続された第2の演算部と、
前記DNN演算で利用される重み情報を格納する記憶部と、を備え、
前記第1の演算部は、前記特徴マップの一部である部分特徴マップと、前記部分特徴マップの前記特徴マップ中の位置を表す位置情報とを前記第2の演算部へ送信し、
前記第2の演算部は、
前記記憶部から前記重み情報を取得し、
前記第1の演算部から送信される前記部分特徴マップおよび前記位置情報を受信し、
前記位置情報に基づく所定の前処理を実行し、
前記前処理の実行結果と前記部分特徴マップとを用いて、前記DNN演算を実行する情報処理装置。
- [請求項2] 請求項1に記載の情報処理装置において、
前記前処理は、前記重み情報において前記位置情報に対応する箇所をマスクする処理であり、
前記第2の演算部は、前記前処理を行うことで生成したマスク済み重み情報と前記部分特徴マップとを用いて、前記DNN演算を実行する情報処理装置。
- [請求項3] 請求項2に記載の情報処理装置において、
前記第2の演算部は、前記位置情報に基づいて、前記特徴マップに設定された複数の領域のうちどの領域に前記部分特徴マップが属するかを判定し、当該領域に対応する箇所をマスクする処理を前記前処理として実行する情報処理装置。
- [請求項4] 請求項3に記載の情報処理装置において、

前記複数の領域は、前記特徴マップの左上角部を含む第1領域と、前記特徴マップの上端部を含む第2領域と、前記特徴マップの右上角部を含む第3領域と、前記特徴マップの左端部を含む第4領域と、前記特徴マップのいずれの端部も含まない第5領域と、前記特徴マップの右端部を含む第6領域と、前記特徴マップの左下角部を含む第7領域と、前記特徴マップの下端部を含む第8領域と、前記特徴マップの右下角部を含む第9領域と、を有し、

前記第2の演算部は、

前記部分特徴マップが前記第5領域以外の各領域に属する場合は、当該領域における前記特徴マップの端部の位置に応じて、前記重み情報においてマスクする前記箇所を決定し、

前記部分特徴マップが前記第5領域に属する場合は、前記マスクする処理を実行しない情報処理装置。

[請求項5] 請求項1に記載の情報処理装置において、

前記前処理は、前記部分特徴マップに0を挿入する0パディングであり、

前記第2の演算部は、前記前処理を行うことで生成した0パディング済みの前記部分特徴マップを用いて、前記DNN演算を実行する情報処理装置。

[請求項6] 請求項1に記載の情報処理装置において、

前記第2の演算部は、ベクトル演算に特化した複数の演算コアを有しており、前記複数の演算コアを用いたAI処理によって前記DNN演算を実行する情報処理装置。

[請求項7] 請求項1に記載の情報処理装置において、

前記第2の演算部は、DNNの各層に対応する複数のカーネルを有しており、

前記DNN演算において前記複数のカーネルのいずれかで前記部分特徴マップの水平方向および垂直方向の少なくとも一方の大きさが変

化する場合、当該カーネルは前記位置情報を再計算して次層に対応するカーネルへ送信する情報処理装置。

[請求項8]

第1の演算部と、前記第1の演算部とバスを介して接続された第2の演算部と、DNN演算で利用される重み情報を格納する記憶部と、を備える情報処理装置を用いて、特徴マップに対してDNN演算を実行する情報処理方法であって、

前記第1の演算部により、前記特徴マップの一部である部分特徴マップと、前記部分特徴マップの前記特徴マップ中の位置を表す位置情報を前記第2の演算部へ送信し、

前記第2の演算部により、

前記記憶部から前記重み情報を取得し、

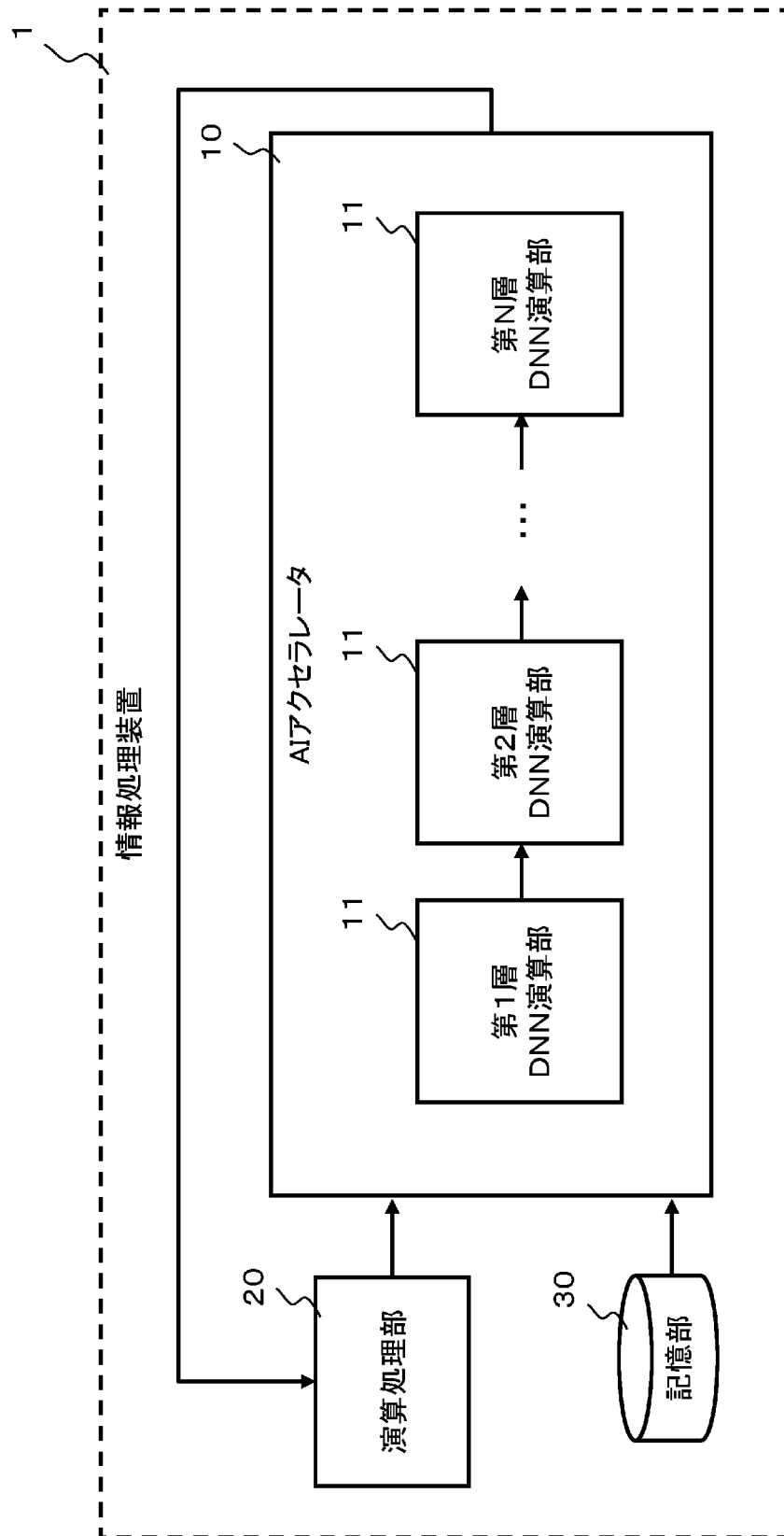
前記第1の演算部から送信される前記部分特徴マップおよび前記位置情報を受信し、

前記位置情報に基づく所定の前処理を実行し、

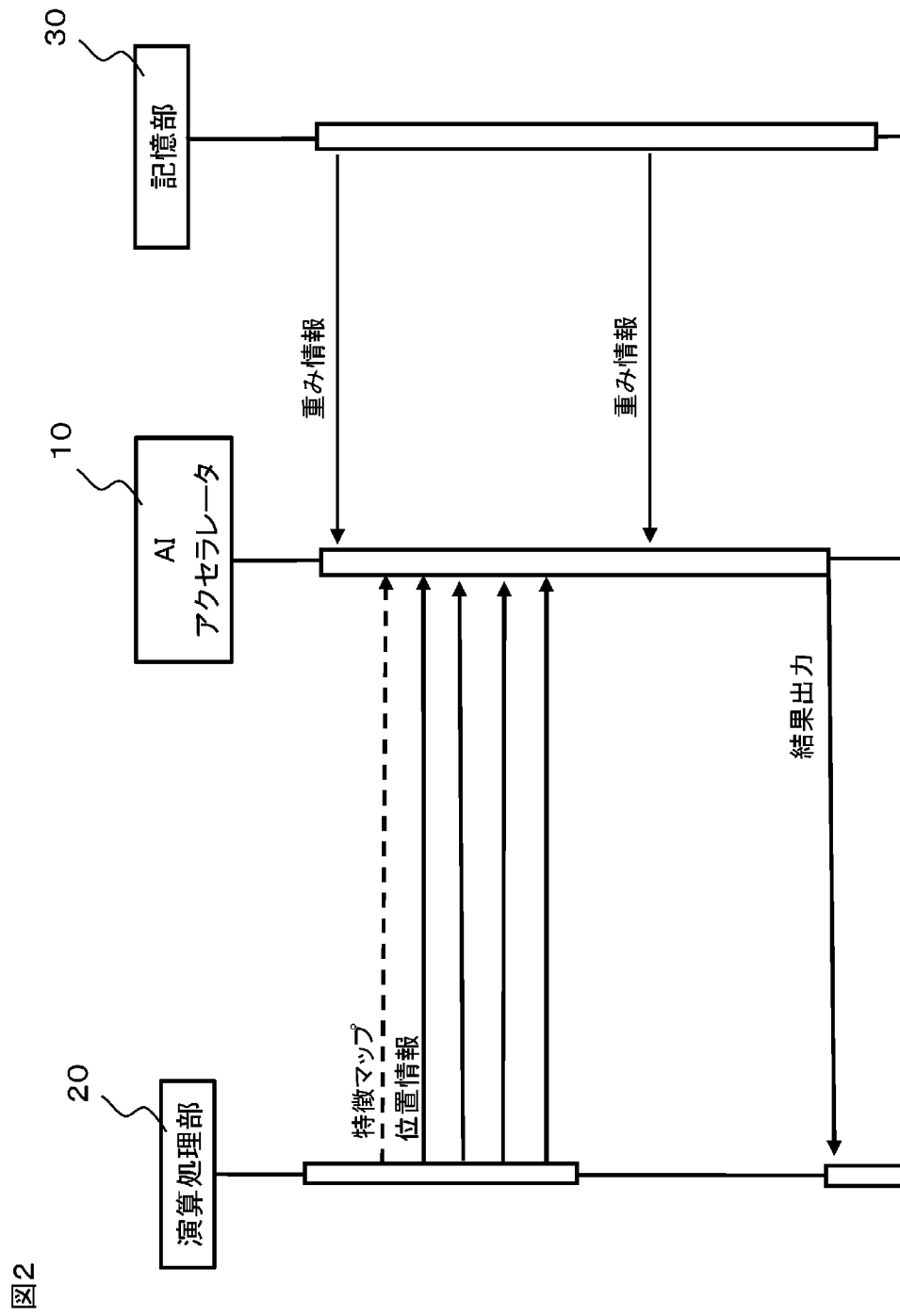
前記前処理の実行結果と前記部分特徴マップとを用いて、前記DNN演算を実行する情報処理方法。

[図1]

図1



[図2]



[図3]

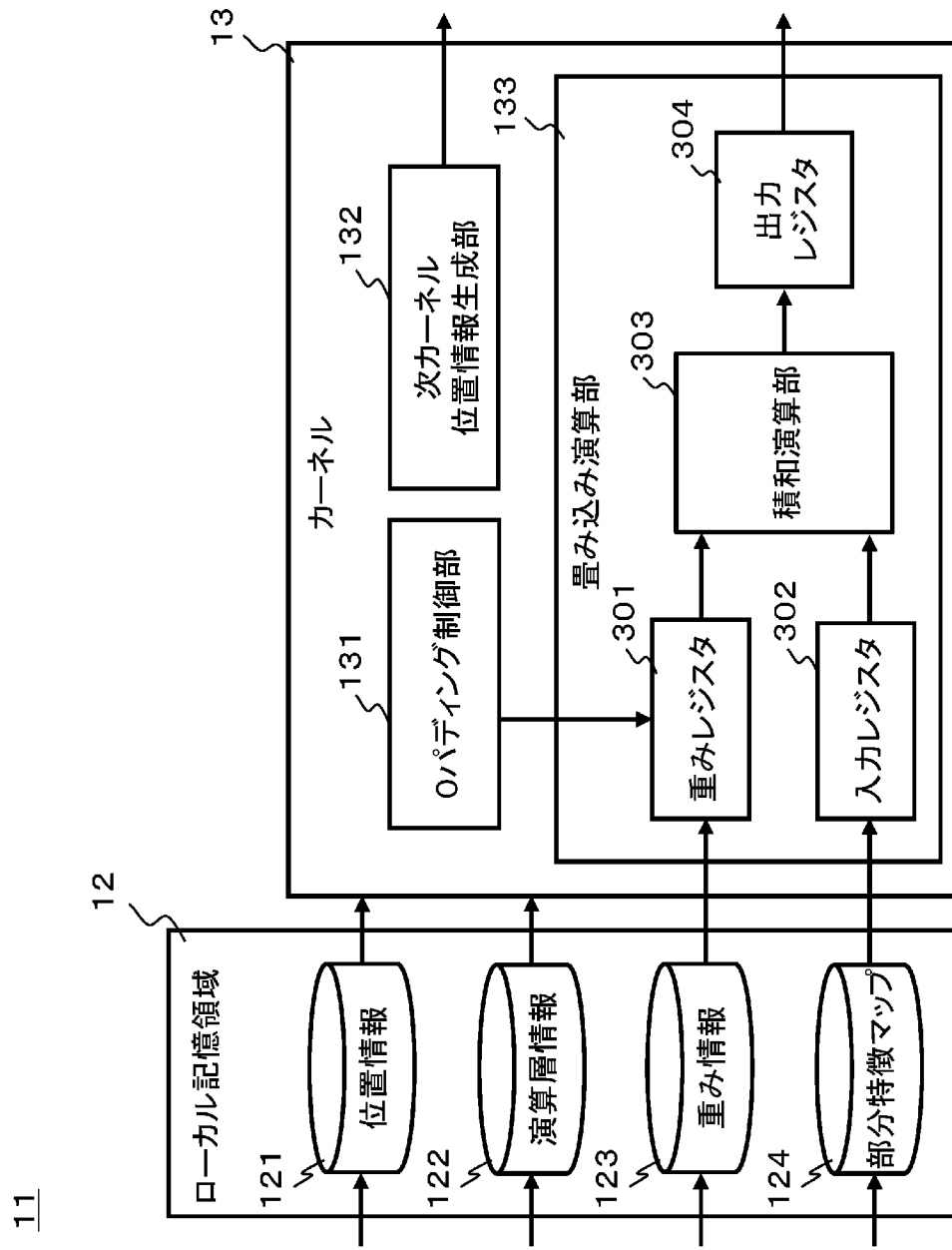
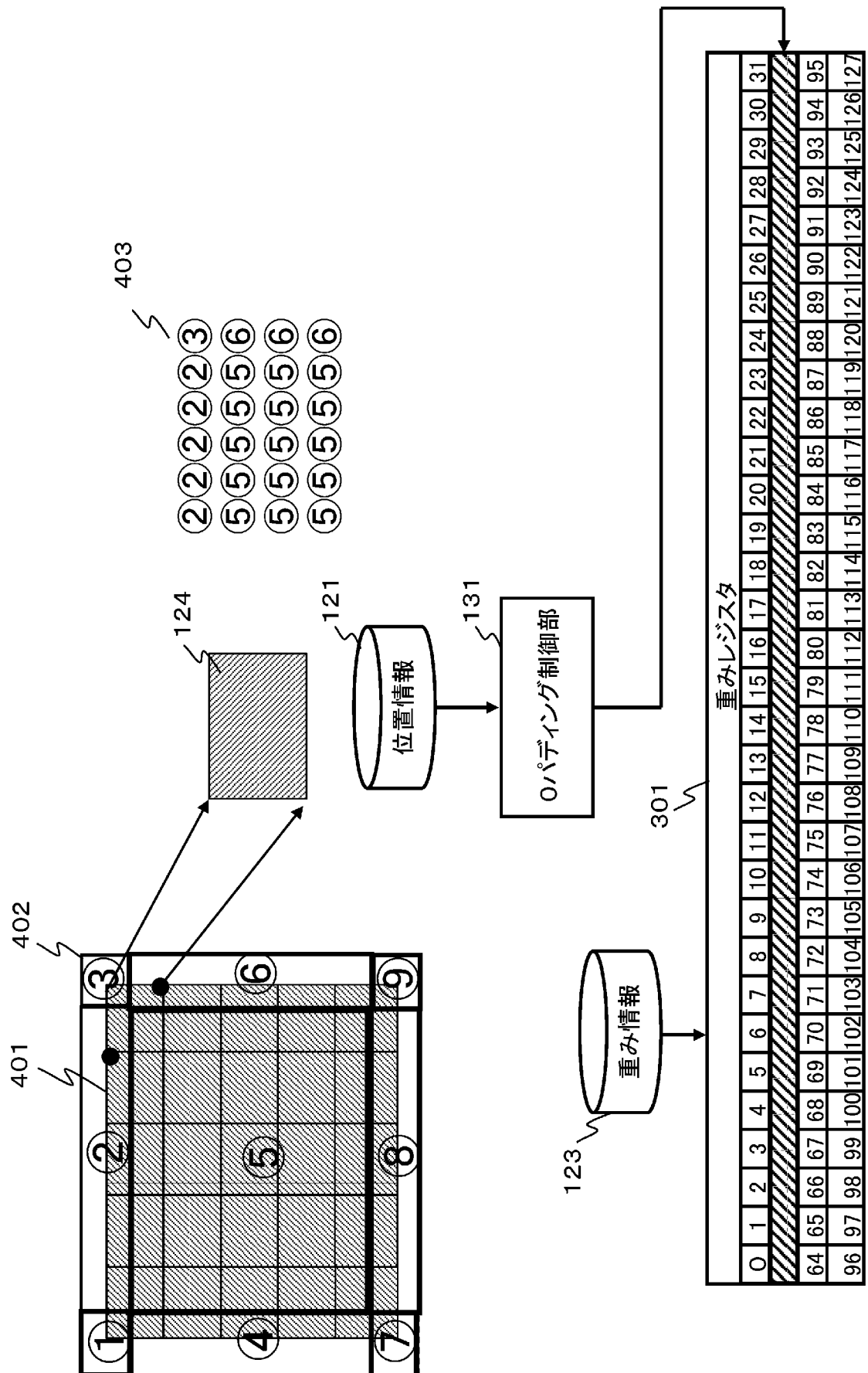


図3

11

[図4]

図4



[図5]

図5

演算層情報

サイズ変化率	1	501
重みサイズ	3	502
ストライド	3	503
パディング幅	1	504
入力特徴マップ幅	128	505
入力特徴マップ高さ	32	506
入力特徴マップ チャンネル数	25	507
出力特徴マップ チャンネル数	25	508

[図6]

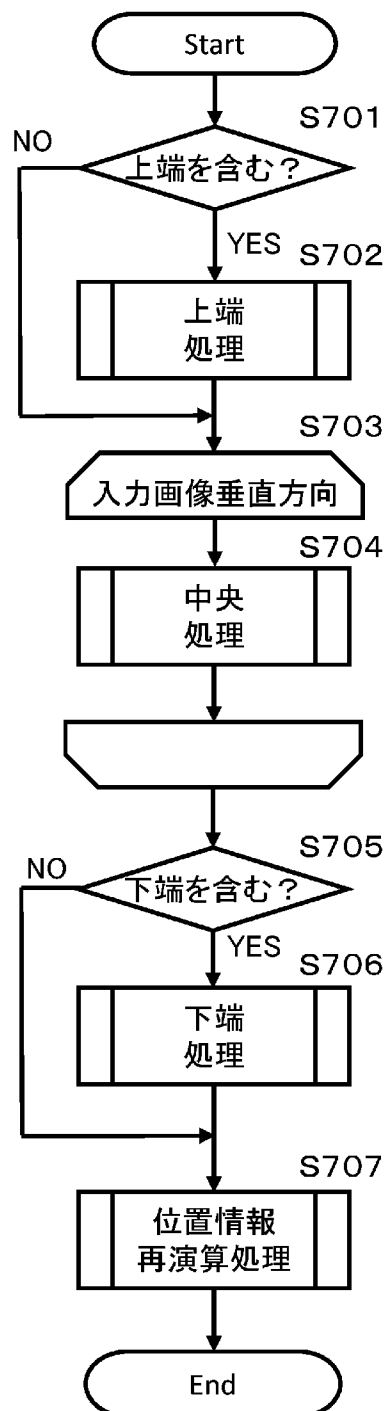
図6

位置情報

分割ID	0	601
全体特徴マップ幅	500	602
全体特徴マップ高さ	500	603
演算層	1	604
開始x座標	0	605
開始y座標	0	606
終点x座標	127	607
終点y座標	31	608
横畳み込み数 (without pad)	126	609
縦畳み込み数 (without pad)	30	610

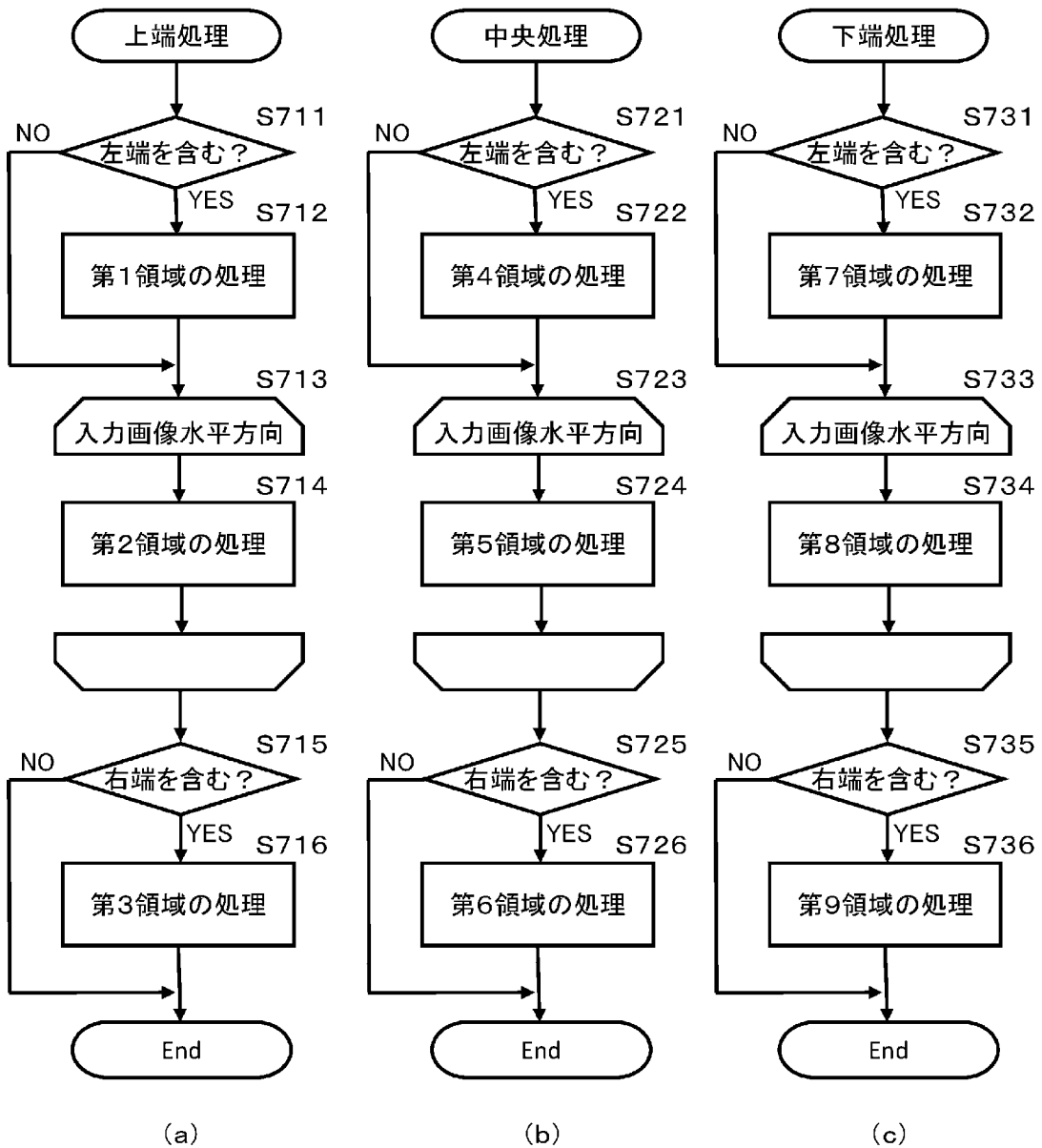
[図7]

図7



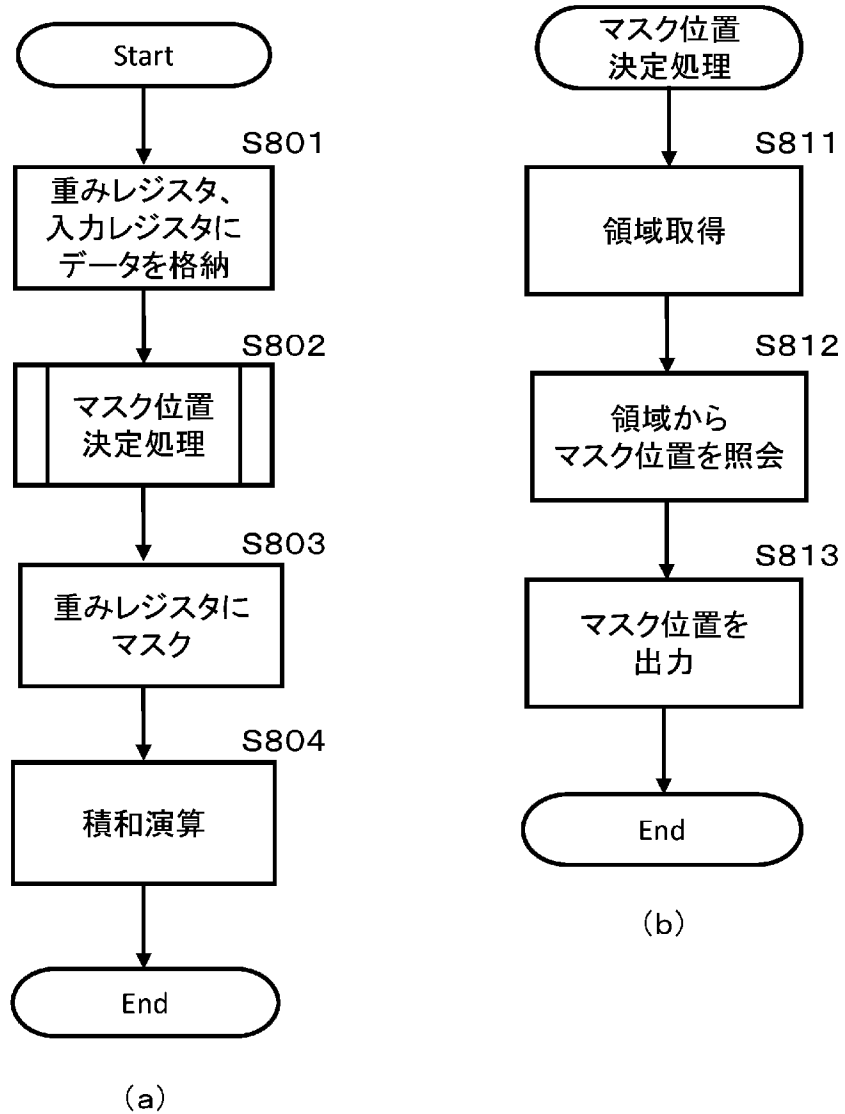
[図8]

図8



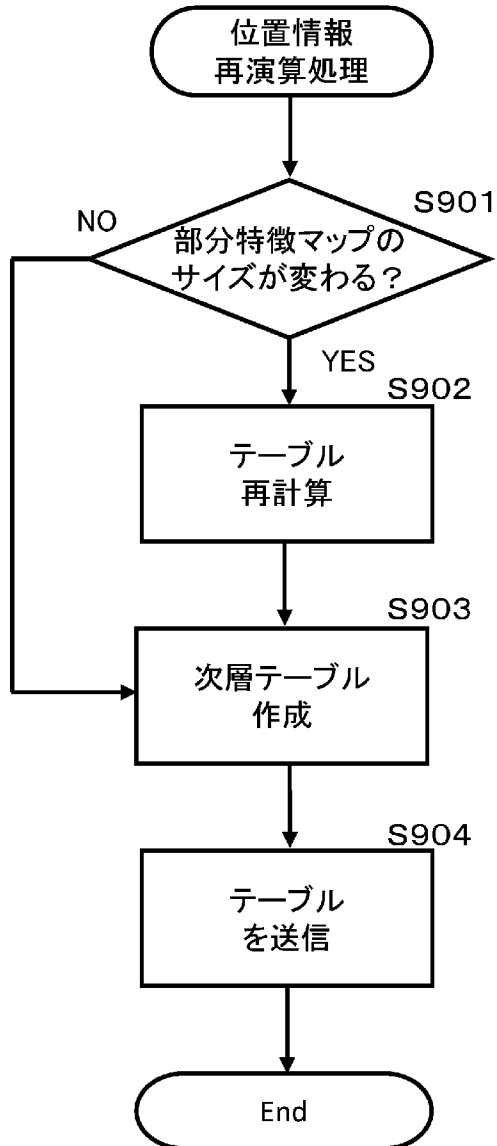
[図9]

図9



[図10]

図10



(a)

$$WW_{k+1} = \lfloor \Delta WW_k \rfloor \quad \cdots (1)$$

$$WH_{k+1} = \lfloor \Delta WH_k \rfloor \quad \cdots (2)$$

$$x_{k+1} = \lfloor \Delta x_k \rfloor, y_{k+1} = \lfloor \Delta y_k \rfloor \quad \cdots (3)$$

$$xend_{k+1} = x_{k+1} + IW_{k+1} - 1 \quad \cdots (4)$$

$$yend_{k+1} = y_{k+1} + IH_{k+1} - 1 \quad \cdots (5)$$

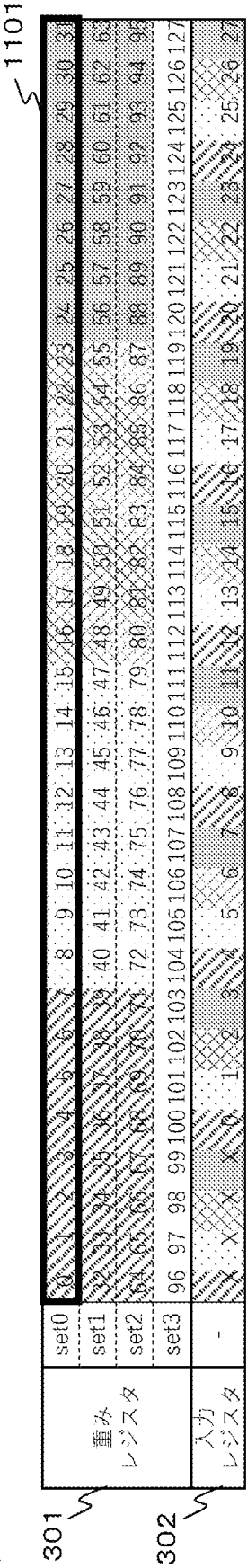
$$OW_{k+1} = \lfloor \frac{(IW_{k+1} - KS_{k+1})}{stride_{k+1}} \rfloor + 1 \quad \cdots (6)$$

$$OH_{k+1} = \lfloor \frac{(IH_{k+1} - KS_{k+1})}{stride_{k+1}} \rfloor + 1 \quad \cdots (7)$$

(b)

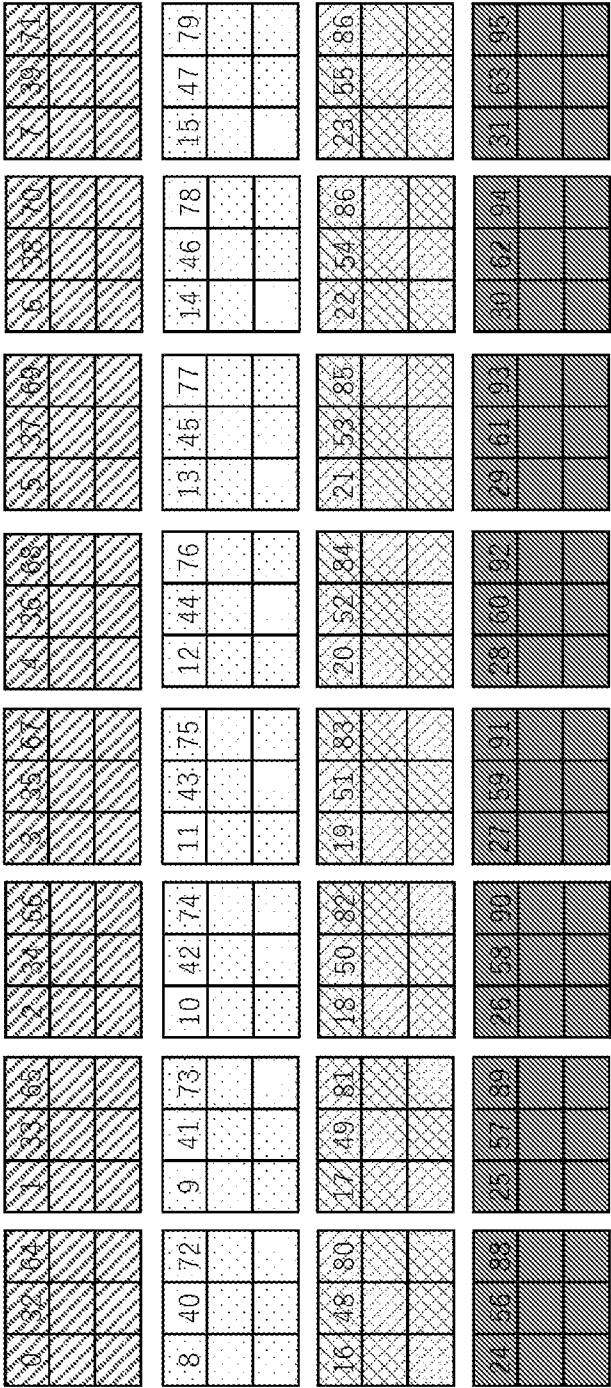
[図11]

図11



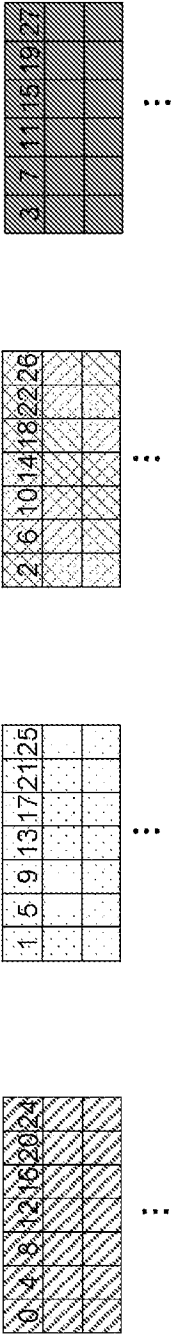
123

重み情報



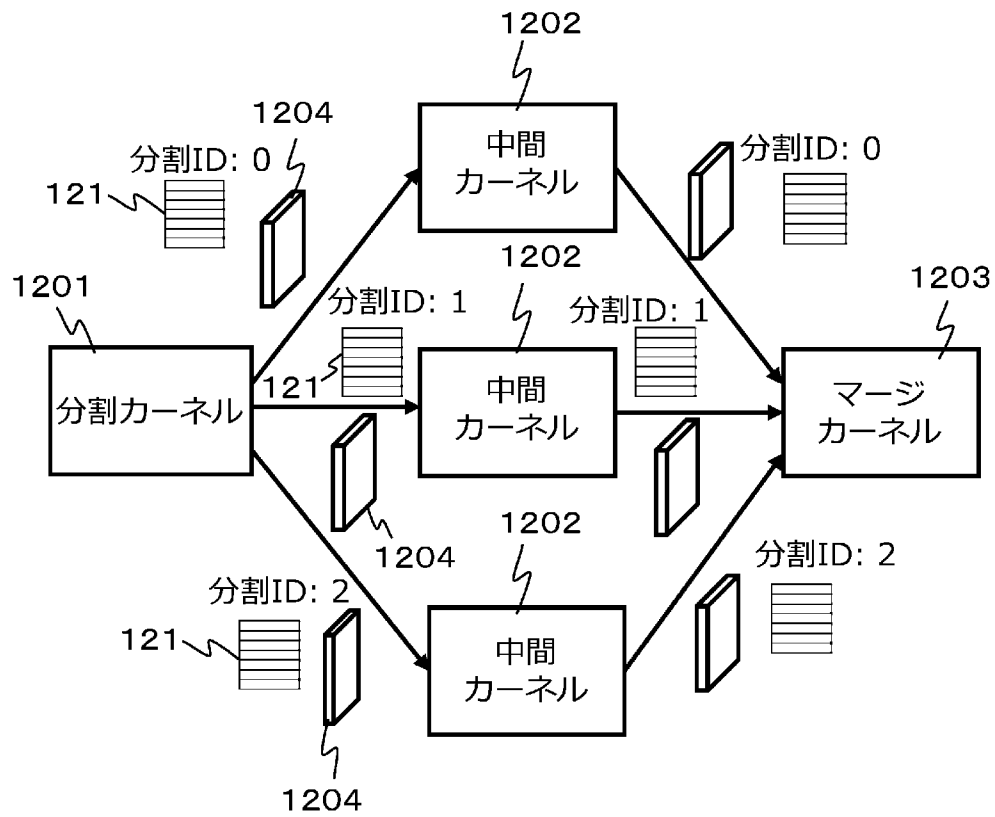
124

部分特徴マップ

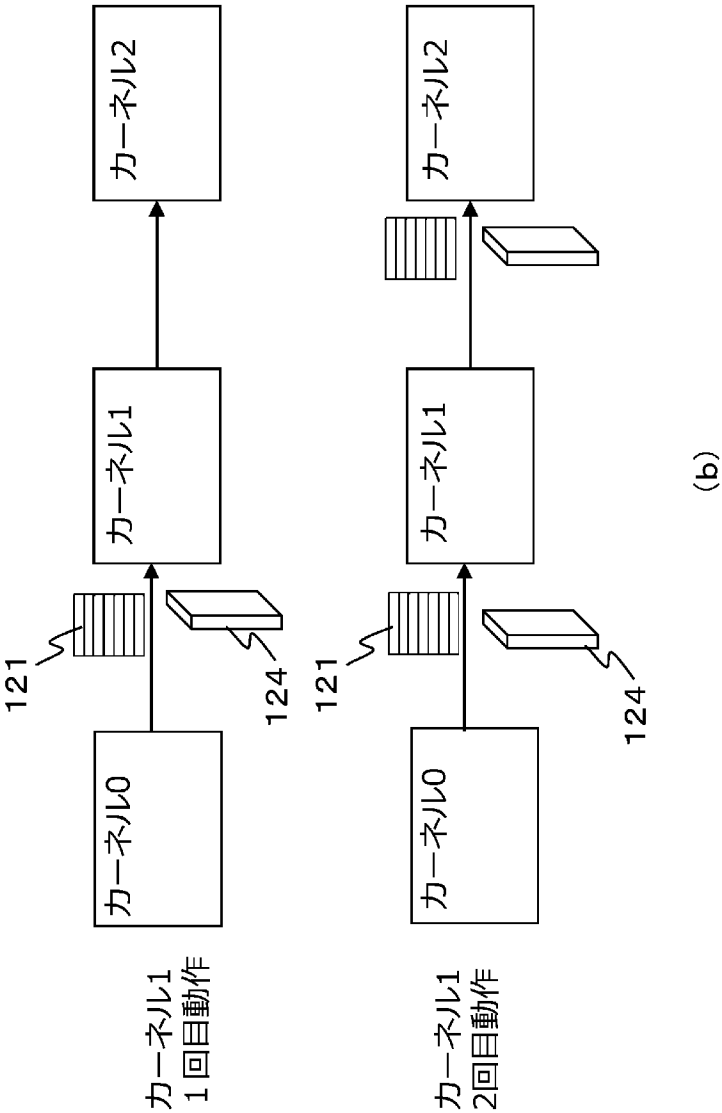
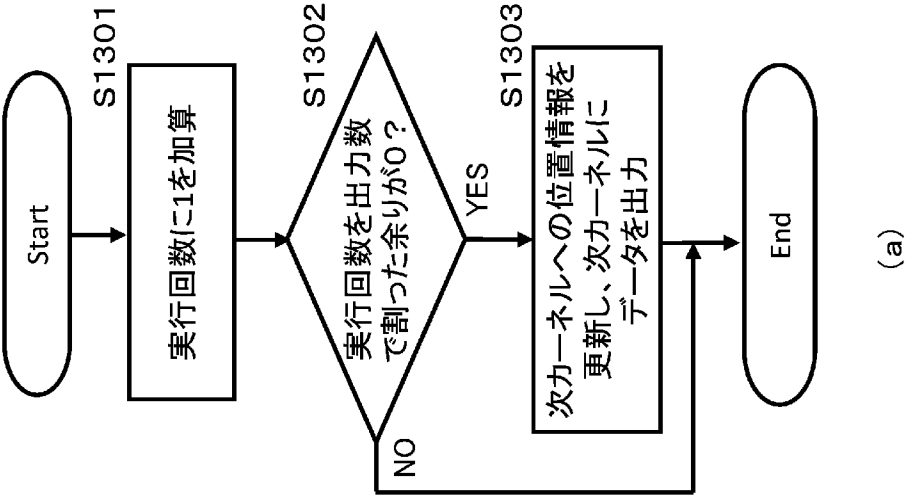


[図12]

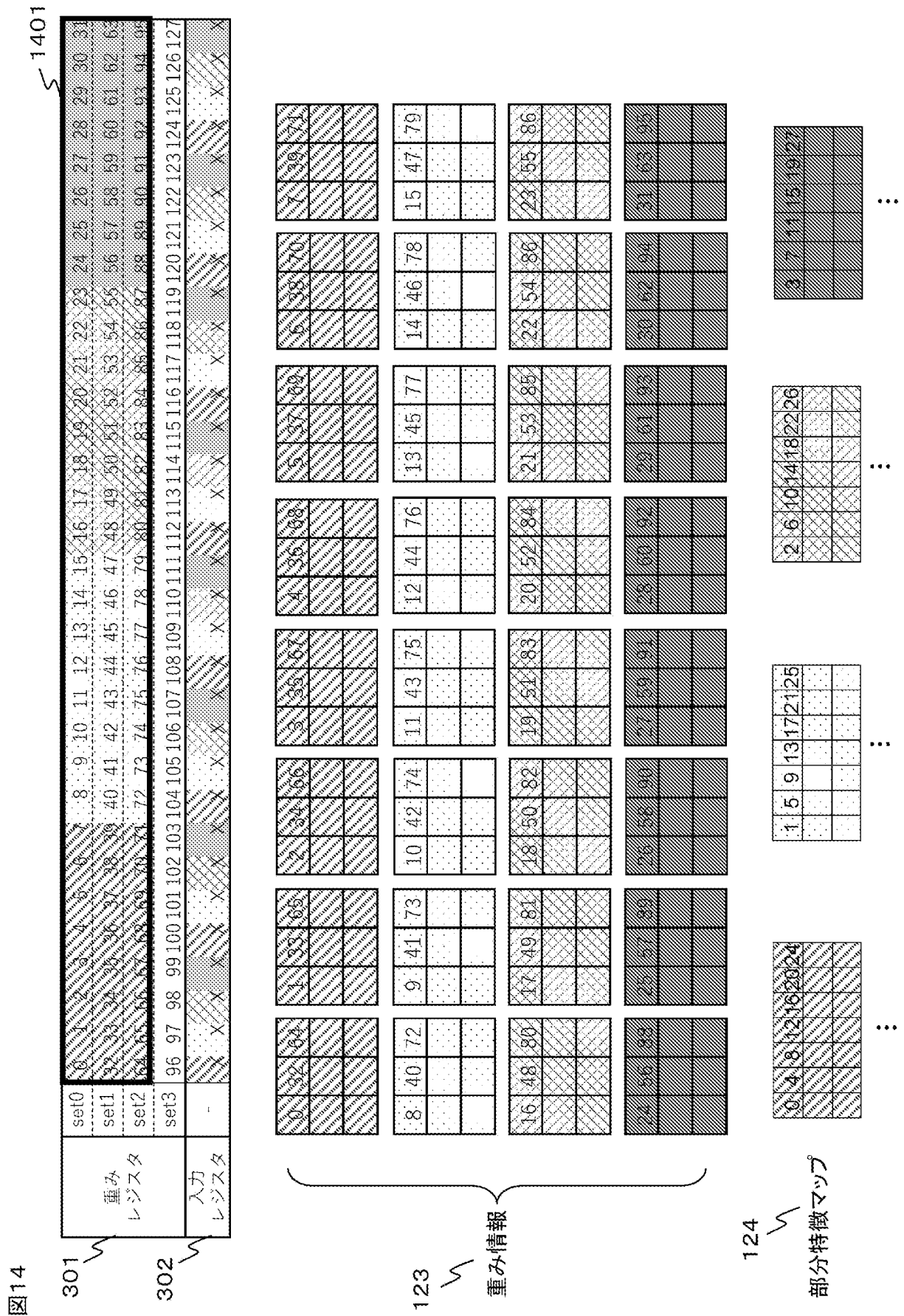
図12



[図13]



[図14]



INTERNATIONAL SEARCH REPORT

International application No.

PCT/JP2021/035696

A. CLASSIFICATION OF SUBJECT MATTER**G06N 3/063**(2006.01)i; **G06F 17/10**(2006.01)i

FI: G06N3/063; G06F17/10 S

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

G06N3/063; G06F17/10

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Published examined utility model applications of Japan 1922-1996

Published unexamined utility model applications of Japan 1971-2021

Registered utility model specifications of Japan 1996-2021

Published registered utility model applications of Japan 1994-2021

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	JP 2019-537139 A (GOOGLE LLC) 19 December 2019 (2019-12-19) paragraphs [0067]-[0072], [0081], [0088]-[0089]	1-8
A	US 2020/0184002 A1 (ADVANCED MICRO DEVICES, INC.) 11 June 2020 (2020-06-11) paragraphs [0047]-[0074]	1-8

☐ Further documents are listed in the continuation of Box C.☒ See patent family annex.

* Special categories of cited documents:

“A” document defining the general state of the art which is not considered to be of particular relevance

“E” earlier application or patent but published on or after the international filing date

“L” document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

“O” document referring to an oral disclosure, use, exhibition or other means

“P” document published prior to the international filing date but later than the priority date claimed

“T” later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

“X” document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

“Y” document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

“&” document member of the same patent family

Date of the actual completion of the international search

27 October 2021

Date of mailing of the international search report

09 November 2021

Name and mailing address of the ISA/JP

Japan Patent Office (ISA/JP)
3-4-3 Kasumigaseki, Chiyoda-ku, Tokyo 100-8915
Japan

Authorized officer

Telephone No.

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.

PCT/JP2021/035696

Patent document cited in search report			Publication date (day/month/year)	Patent family member(s)	Publication date (day/month/year)
JP	2019-537139	A	19 December 2019	US 2018/0129936 A1 paragraphs [0075]-[0080], [0089], [0096]-[0097] GB 2556670 A DE 102017121887 A1 CN 108073983 A KR 10-2019-0084088 A	
US	2020/0184002	A1	11 June 2020	(Family: none)	

A. 発明の属する分野の分類（国際特許分類（IPC）） G06N 3/063(2006.01)i; G06F 17/10(2006.01)i FI: G06N3/063; G06F17/10 S														
B. 調査を行った分野														
調査を行った最小限資料（国際特許分類（IPC）） G06N3/063; G06F17/10														
最小限資料以外の資料で調査を行った分野に含まれるもの <table border="0"> <tr> <td>日本国実用新案公報</td> <td>1922 - 1996年</td> </tr> <tr> <td>日本国公開実用新案公報</td> <td>1971 - 2021年</td> </tr> <tr> <td>日本国実用新案登録公報</td> <td>1996 - 2021年</td> </tr> <tr> <td>日本国登録実用新案公報</td> <td>1994 - 2021年</td> </tr> </table>			日本国実用新案公報	1922 - 1996年	日本国公開実用新案公報	1971 - 2021年	日本国実用新案登録公報	1996 - 2021年	日本国登録実用新案公報	1994 - 2021年				
日本国実用新案公報	1922 - 1996年													
日本国公開実用新案公報	1971 - 2021年													
日本国実用新案登録公報	1996 - 2021年													
日本国登録実用新案公報	1994 - 2021年													
国際調査でを使用した電子データベース（データベースの名称、調査に使用した用語）														
C. 関連すると認められる文献														
引用文献の カテゴリー*	引用文献名 及び一部の箇所が関連するときは、その関連する箇所の表示	関連する 請求項の番号												
A	JP 2019-537139 A (グーグル エルエルシー) 19.12.2019 (2019 - 12 - 19) 段落[0067]-[0072], [0081], [0088]-[0089]	1-8												
A	US 2020/0184002 A1 (ADVANCED MICRO DEVICES, INC.) 11.06.2020 (2020 - 06 - 11) 段落[0047]-[0074]	1-8												
☐ C欄の続きにも文献が列挙されている。 ☒ パテントファミリーに関する別紙を参照。														
<table border="0"> <tr> <td>* 引用文献のカテゴリー</td> <td>"T" 国際出願日又は優先日後に公表された文献であって出願と抵触するものではなく、発明の原理又は理論の理解のために引用するもの</td> </tr> <tr> <td>"A" 特に関連のある文献ではなく、一般的技術水準を示すもの</td> <td>"X" 特に関連のある文献であって、当該文献のみで発明の新規性又は進歩性がないと考えられるもの</td> </tr> <tr> <td>"E" 国際出願日前の出願または特許であるが、国際出願日以後に公表されたもの</td> <td>"Y" 特に関連のある文献であって、当該文献と他の1以上の文献との、当業者にとって自明である組合せによって進歩性がないと考えられるもの</td> </tr> <tr> <td>"L" 優先権主張に疑義を提起する文献又は他の文献の発行日若しくは他の特別な理由を確立するために引用する文献（理由を付す）</td> <td>"&" 同一パテントファミリー文献</td> </tr> <tr> <td>"O" 口頭による開示、使用、展示等に言及する文献</td> <td></td> </tr> <tr> <td>"P" 国際出願日前で、かつ優先権の主張の基礎となる出願の日の後に公表された文献</td> <td></td> </tr> </table>			* 引用文献のカテゴリー	"T" 国際出願日又は優先日後に公表された文献であって出願と抵触するものではなく、発明の原理又は理論の理解のために引用するもの	"A" 特に関連のある文献ではなく、一般的技術水準を示すもの	"X" 特に関連のある文献であって、当該文献のみで発明の新規性又は進歩性がないと考えられるもの	"E" 国際出願日前の出願または特許であるが、国際出願日以後に公表されたもの	"Y" 特に関連のある文献であって、当該文献と他の1以上の文献との、当業者にとって自明である組合せによって進歩性がないと考えられるもの	"L" 優先権主張に疑義を提起する文献又は他の文献の発行日若しくは他の特別な理由を確立するために引用する文献（理由を付す）	"&" 同一パテントファミリー文献	"O" 口頭による開示、使用、展示等に言及する文献		"P" 国際出願日前で、かつ優先権の主張の基礎となる出願の日の後に公表された文献	
* 引用文献のカテゴリー	"T" 国際出願日又は優先日後に公表された文献であって出願と抵触するものではなく、発明の原理又は理論の理解のために引用するもの													
"A" 特に関連のある文献ではなく、一般的技術水準を示すもの	"X" 特に関連のある文献であって、当該文献のみで発明の新規性又は進歩性がないと考えられるもの													
"E" 国際出願日前の出願または特許であるが、国際出願日以後に公表されたもの	"Y" 特に関連のある文献であって、当該文献と他の1以上の文献との、当業者にとって自明である組合せによって進歩性がないと考えられるもの													
"L" 優先権主張に疑義を提起する文献又は他の文献の発行日若しくは他の特別な理由を確立するために引用する文献（理由を付す）	"&" 同一パテントファミリー文献													
"O" 口頭による開示、使用、展示等に言及する文献														
"P" 国際出願日前で、かつ優先権の主張の基礎となる出願の日の後に公表された文献														
国際調査を完了した日 27.10.2021	国際調査報告の発送日 09.11.2021													
名称及びあて先 日本国特許庁(ISA/JP) 〒100-8915 日本国 東京都千代田区霞が関三丁目4番3号	権限のある職員（特許庁審査官） 多胡 滋 5B 3562 電話番号 03-3581-1101 内線 3545													

国際調査報告
 パテントファミリーに関する情報

国際出願番号

PCT/JP2021/035696

引用文献			公表日	パテントファミリー文献	公表日
JP	2019-537139	A	19.12.2019	US 2018/0129936 A1 段落[0075]-[0080], [0089], [0096]-[0097] GB 2556670 A DE 102017121887 A1 CN 108073983 A KR 10-2019-0084088 A	
US	2020/0184002	A1	11.06.2020	(ファミリーなし)	