(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12月22日(22.12.2022)

WIPO | PCT

(10) 国際公開番号
WO 2022/264186 A1

(51) 国際特許分類:
G06N 3/08 (2006.01)

(21) 国際出願番号： PCT/JP2021/022425

(22) 国際出願日： 2021年6月14日(14.06.2021)

(25) 国際出願の言語： 日本語

(26) 国際公開の言語： 日本語

(71) 出願人:三菱電機株式会社(**MITSUBISHI ELECTRIC CORPORATION**) [JP/JP]; 〒1008310 東京都千代田区丸の内二丁目７番３号 Tokyo (JP).

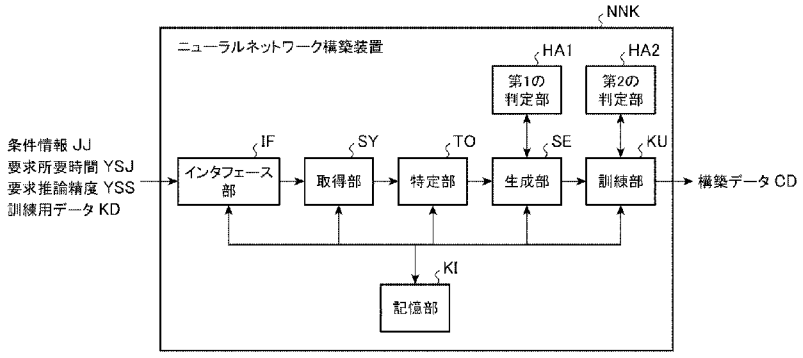
(72) 発明者:田中 元気(**TANAKA, Genki**); 〒1008310 東京都千代田区丸の内二丁目７番３号 三菱電機株式会社内 Tokyo (JP). 佐藤 匠(**SATO, Takumi**); 〒1008310 東京都千代田区丸の内二丁目７番３号 三菱電機株式会社内 Tokyo (JP).

(74) 代理人: 特許業務法人山王内外特許事務所 (**SANNO PATENT ATTORNEYS OFFICE**); 〒1000014 東京都千代田区永田町二丁目１２番４号 赤坂山王センタービル５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, JP, 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, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, WS, ZA, ZM, ZW.

(54) **Title:** NEURAL NETWORK CONSTRUCTION DEVICE, NEURAL NETWORK CONSTRUCTION METHOD, IMAGE PROCESSING DEVICE, AND IMAGE PROCESSING METHOD

(54) 発明の名称： ニューラルネットワーク構築装置、ニューラルネットワーク構築方法、画像処理装置、及び画像処理方法



CD Construction data
HA1 First determination unit
HA2 Second determination unit
IF Interface unit
JJ Condition information
KD Training data
KI Storage unit
KU Training unit
NNK Neural network construction device
SE Generation unit
SY Acquisition unit
TO Identification unit
YSJ Demanded required time
YSS Demanded inference accuracy

(57) **Abstract:** This neural network construction device (NNK) includes: an identification unit (TO) that identifies, on the basis of a construction condition for constructing a neural network model (NNM), a plurality of computations which can be allocated to a plurality of layers constituting the NNM; a generation unit (SE) that generates one combination of computations by allocating one computation among the plurality of computations to each of the plurality of layers; a first determination unit (HA1) that determines whether the total required time for the allocated computations satisfies a demanded required time; a training unit (KU) that trains the NNM having the generated one combination of computations; and a second determination unit (HA2) that determines whether the accuracy of inference by the NNM after training satisfies a demanded inference accuracy. If the demanded required time is not satisfied, or if the demanded inference accuracy is not satisfied, the generation unit (SE) generates another combination of computations which is different from the one combination of computations.

WO 2022/264186 A1

[続葉有]

(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：ニューラルネットワーク構築装置（ＮＮＫ）は、ニューラルネットワークモデル（ＮＮＭ）を構築するための構築条件に基づき、前記ＮＮＭを構成する複数の層に割り当て可能な複数の演算を特定する特定部（ＴＯ）と、前記複数の層の各々に、前記複数の演算のうちの一つの演算を割り当てることにより、演算の一の組み合わせを生成する生成部（ＳＥ）と、前記割り当てられた演算の合計所要時間が要求所要時間を満足するか否かを判定する第１の判定部（ＨＡ１）と、前記生成された演算の一の組み合わせを有するＮＮＭを訓練する訓練部（ＫＵ）と、前記訓練後のＮＮＭが推論する精度が要求推論精度を満足するか否かを判定する第２の判定部（ＨＡ２）と、を含み、前記要求所要時間を満足しないとき、または、前記要求推論精度を満足しないとき、前記生成部（ＳＥ）は、前記演算の一の組み合わせと相違する、演算の他の組み合わせを生成する。

明　細　書

発明の名称 ：

ニューラルネットワーク構築装置、ニューラルネットワーク構築方法、画像処理装置、及び画像処理方法

技術分野

[0001]　本開示は、ニューラルネットワーク構築装置、ニューラルネットワーク構築方法、画像処理装置、及び画像処理方法に関する。

背景技術

[0002]　ニューラルネットワーク構築装置の一例である、特許文献１に記載の学習モデル生成装置は、学習モデルを実行するための計算機の計算リソースが、たとえ比較的小さくても、前記学習モデルによる認識等の精度が低下することを抑止すべく、前記学習モデルを構成する層を演算量が削減される構造に置換する。

先行技術文献

特許文献

[0003] 特許文献1：特開２０２０－１０７０４２号公報

発明の概要

発明が解決しようとする課題

[0004]　上記した学習モデル生成装置は、上記した学習モデルの演算量を削減することから、前記学習モデルに本来的に要求される認識等の精度を満足するものの、前記学習モデルに本来的に要求される速度、即ち所要時間を満足することができないことがあるとの課題があった。

[0005]　本開示の目的は、学習モデルを実行するためのリソースが比較的小さくても、本来的に要求される、推論についての所要時間及び精度の両者を満足するニューラルネットワークモデルを構築することができるニューラルネットワーク構築装置を提供することにある。

課題を解決するための手段

[0006] 上記した課題を解決すべく、ニューラルネットワークモデルを構築するための構築条件に基づき、前記ニューラルネットワークモデルを構成する複数の層に割り当て可能な複数の演算を特定する特定部と、前記複数の層の各々に、前記複数の演算のうちの一つの演算を割り当てることにより、演算の一の組み合わせを生成する生成部と、前記割り当てられた演算に基づき推論する、各演算の所要時間の合計である合計所要時間が、予め定められた要求所要時間を満足するか否かを判定する第１の判定部と、前記生成された演算の一の組み合わせを有するニューラルネットワークモデルを訓練する訓練部と、前記訓練後のニューラルネットワークモデルが推論する精度が、予め定められた要求推論精度を満足するか否かを判定する第２の判定部と、を含み、前記第１の判定部が、前記合計所要時間が前記要求所要時間を満足しないと判定するとき、または、前記第２の判定部が、前記推論する精度が前記要求推論精度を満足しないと判定するとき、前記生成部は、前記演算の一の組み合わせと相違する、演算の他の組み合わせを生成する。

発明の効果

[0007] 本開示に係るニューラルネットワークモデル構築装置によれば、学習モデルを実行するためのリソースが比較的小さくても、本来的に要求される、推論についての所要時間及び精度の両者を満足するニューラルネットワークモデルを構築することができる。

図面の簡単な説明

[0008] [図1]実施形態１のニューラルネットワーク構築装置ＮＮＫの機能ブロック図である。

[図2]実施形態１のニューラルネットワークモデルＮＮＭの基本的構成を示す。

[図3]実施形態１の記憶部ＫＩの構成を示す。

[図4]実施形態１のテーブルＴＢの構成を示す。

[図5]実施形態１のニューラルネットワーク構築装置ＮＮＫの構成を示す。

[図6]実施形態１のニューラルネットワーク構築装置ＮＮＫの動作を示すフローチャートである。

[図7]実施形態１の生成部ＳＥの動作を示す。

[図8]実施形態１の第１の判定部ＨＡ１の動作を示す。

[図9]実施形態１の訓練部ＫＵの動作を示す。

[図10]実施形態２のニューラルネットワーク構築装置ＮＮＫの機能ブロック図である。

[図11]実施形態２のニューラルネットワーク構築装置ＮＮＫの動作を示すフローチャートである。

[図12]実施形態３の画像処理装置ＧＳＳの機能を示す。

発明を実施するための形態

[0009] 本開示に係るニューラルネットワーク構築装置、及び画像処理装置の実施形態について説明する。

[0010] 実施形態１及び実施形態２のニューラルネットワーク構築装置ＮＮＫ（図１、図１０に図示。）は、学習フェーズでの装置であり、他方で、実施形態３の画像処理装置ＧＳＳ（図１２に図示。）は、推論フェーズでの装置である。

[0011] 実施形態１．

〈実施形態１〉

実施形態１のニューラルネットワーク構築装置ＮＮＫについて説明する。

[0012] 〈〈ニューラルネットワーク構築装置ＮＮＫの機能〉〉

図１は、実施形態１のニューラルネットワーク構築装置ＮＮＫの機能ブロック図である。

[0013] 実施形態１のニューラルネットワーク構築装置ＮＮＫは、タスク（例えば、クラス分類、物体検出、位置特定、セグメンテーション）を実施するニューラルネットワークモデルＮＮＭを構築すべく、図１に示されるように、記憶部ＫＩと、インターフェース部ＩＦと、取得部ＳＹと、特定部ＴＯと、生成部ＳＥと、訓練部ＫＵと、第１の判定部ＨＡ１と、第２の判定部ＨＡ２と

、を含む。

[0014] 図２は、実施形態１のニューラルネットワークモデルＮＮＭの基本的構成を示す。

[0015] 〈ニューラルネットワークモデルＮＮＭの構成〉

ニューラルネットワークモデルＮＮＭは、図２に示されるように、従来知られたと同様に、例えば、入力層Ｌｉｎ、出力層Ｌｏｕｔ、７つの中間層Ｌ１～Ｌ７、及び隣接する層間の重みｗから構成される。

[0016] 入力層Ｌｉｎ及び出力層Ｌｏｕｔは、演算（図示せず。）が、予め割り当てられており、かつ固定されており、また、前記演算に関連する入出力関係（例えば、前記演算のために入力されるデータの本数、前記演算の結果として出力されるデータの本数）は、ニューラルネットワークモデル構築装置ＮＮＫ（図１に図示。）によるニューラルネットワークモデルＮＮＭの構築の過程で、決定される。

[0017] ７つの中間層Ｌ１～Ｌ７は、演算及び入出力関係が、予め割り当てられておらず、ニューラルネットワークモデル構築装置ＮＮＫによるニューラルネットワークモデルＮＮＭの構築の過程で、決定される。

[0018] 重みｗは、その値が、ニューラルネットワーク構築装置ＮＮＫによるニューラルネットワークモデルＮＮＭの構築の過程で、決定される。

[0019] 以下では、重みｗの値を「重み値ｗｔ」という。重み値ｗｔは、本来、固有の値であるものの、説明及び理解を容易にすべく、複数の重み値を総称する。

[0020] 〈記憶部ＫＩの構成〉

図３は、実施形態１の記憶部ＫＩの構成を示す。

[0021] 記憶部ＫＩは、図３に示されるように、テーブルＴＢを予め記憶している。記憶部ＫＩには、また、ニューラルネットワーク構築装置ＮＮＫ（図１に図示。）がニューラルネットワークモデルＮＮＭ（図２に図示。）を構築する過程で、構築データＣＤが保存される。

[0022] 〈テーブルＴＢの構成〉

図４は、実施形態１のテーブルＴＢの構成を示す。

[0023] テーブルＴＢは、図４に示されるように、条件情報ＪＪと探索空間ＴＫとの関係を示す。

[0024] 条件情報ＪＪは、ニューラルネットワークモデルＮＮＭ（図２に図示。）を構築するための条件であり、図４に示されるように、例えば、タスク名、入出力サイズ、及びハードウェア仕様を示す。

[0025] タスク名は、上記したように、ニューラルネットワークモデルＮＮＭ（図２に図示。）が実行すべきタスクの候補であり、例えば、クラス分類、物体検出、位置特定、セグメンテーションの名称である。

[0026] 入出力サイズは、ニューラルネットワークモデルＮＮＭが上記タスクの一つを実行するときに、入力層Ｌｉｎへ入力されるデータの本数、出力層Ｌｏｕｔから出力されるデータの本数である。

[0027] ハードウェア仕様は、ニューラルネットワークモデルＮＮＭを動作させるハードウェアの仕様であり、具体的には、コンピュータの型式、処理のタイプ（CPU/GPU/FPGA/DPU）等である。

[0028] テーブルＴＢは、ニューラルネットワーク構築装置ＮＮＫ内の記憶部ＫＩ（図３に図示。）に予め記憶されることに代えて、例えば、ニューラルネットワーク構築装置ＮＮＫの外部にある記憶装置（図示せず。）に記憶されており、取得部ＳＹ（図１に図示。）により、インターフェース部ＩＦ（図１に図示。）を介して取得されてもよい。

[0029] 探索空間ＴＫは、条件情報ＪＪにより特定される内容である。探索空間ＴＫは、具体的には、図４に示されるように、ニューラルネットワークモデルＮＮＭ（図２に図示。）の中間層Ｌ１～Ｌ７に割当可能な演算、当該演算の入出力構成、当該演算の接続条件、及び、当該演算の所要時間（当該演算を、条件情報ＪＪにより特定されるハードウェアで実行するときの所要時間）を示す。

[0030] 〈構築データＣＤの構成〉

構築データＣＤは、図３に示されるように、演算組み合わせＥＫ及び重み

値ｗｔからなる。

[0031] 演算組み合わせＥＫは、ニューラルネットワークモデルＮＮＭの中間層Ｌ１～Ｌ７に割り当てられた演算の組み合わせを示す。重み値ｗｔは、図２を参照して上述したように、層間の重みの値である。

[0032] 〈〈ニューラルネットワーク構築装置ＮＮＫの機能（続き）〉〉

図１に戻り、説明を続ける。

[0033] インターフェース部ＩＦは、ニューラルネットワークモデルＮＮＭ（図２に図示。）を構築しようとするユーザ（図示せず。）から、条件情報ＪＪ（図４にも図示。）、要求所要時間ＹＳＪ、要求推論精度ＹＳＳ、及び訓練用データＫＤの入力を受ける。

[0034] 要求所要時間ＹＳＪ、及び要求推論精度ＹＳＳは、推論のときに、即ち、ニューラルネットワークモデルＮＮＭが実際にタスクを行うときに満足すべき所要時間、及び推論精度である。

[0035] ユーザは、要求所要時間ＹＳＪとして、例えば、ニューラルネットワークモデルＮＮＭがタスク「クラス分類」（図４に図示。）を、要求所要時間ＹＳＪ以内かつ要求推論精度ＹＳＳ以上で実行すべき旨を入力する。

[0036] 訓練用データＫＤは、ニューラルネットワーク構築装置ＮＮＫが、ニューラルネットワークモデルＮＮＭを構築するときに、より詳細には、訓練するときに用いられるデータである。訓練用データＫＤは、従来知られたと同様に、例えば、教師あり学習を行うためのラベル付きデータである。

[0037] 取得部ＳＹは、インターフェース部ＩＦを介して、上記した、条件情報ＪＪ、要求所要時間ＹＳＪ、要求推論精度ＹＳＳ、及び訓練用データＫＤを取得する。

[0038] 特定部ＴＯは、取得部ＳＹにより取得された条件情報ＪＪに基づき、テーブルＴＢ（図４に図示。）を参照することにより、探索空間ＴＫを特定する。特定部ＴＯは、他方で、条件情報ＪＪの内容の如何に拘わらず、Skip Connection演算（スキップコネクション演算）、No Connection演算（図示せず。）を、探索空間ＴＫ内の割り当て可能な演算として追加する。

[0039] 例えば、取得部ＳＹが、図４中の一点鎖線（左側）で示されると同様な、クラス名「クラス分類」、入力サイズ「１」、出力サイズ「１」、及びハードウェア仕様「ＨＷ１」という条件情報ＪＪ（以下、「条件情報ＪＪ１」という。）をユーザから取得したことを想定する。当該想定の下で、特定部ＴＯは、条件情報ＪＪ１に基づき、テーブルＴＢを参照することにより、図４中の一点鎖線（右側）で示されるように、探索空間ＴＫ（以下、「探索空間ＴＫ１」という。）を特定する。

[0040] 探索空間ＴＫ１は、上記した条件情報ＪＪ１の下で、以下を示す。

（１）ニューラルネットワークモデルＮＮＭの中間層Ｌ１～Ｌ７（図２に図示。）に、４つの演算ＯＰ１、ＯＰ２、ＯＰ３、ＯＰ５を割当可能であること。

（２）例えば、演算ＯＰ１は、演算ＯＰ１に入力されるデータの本数が１つであり、演算ＯＰ１から出力されるデータの本数が１つであること。

（３）例えば、演算ＯＰ１は、演算ＯＰ２の前の位置に割り当てられることが接続条件であること。

（４）例えば、演算ＯＰ１をハードウェアＨＷ１で実行すると、所要時間Ｔ１を要すること。

[0041] テーブルＴＢ中の演算ＯＰ１等は、例えば、従来良く知られた、Convolution演算及びMaxpooling演算である。演算ＯＰ１等は、上記の演算に加えて、例えば、Skip Connection演算（スキップコネクション演算）、No Connection演算である（ノーコネクション演算）。Skip Connection演算は、何ら処理を行うことなく、入力されたデータをそのまま出力し、また、No Connection演算は、何らかのデータの入力を受けても、データを何ら出力しない。

[0042] 以下では、説明及び理解を容易にすべく、演算ＯＰ１等の所要時間Ｔ１等は、演算ＯＰ１等の内部構造、及び演算ＯＰ１等と外部との入出力関係等に拘わらず、固定であることを想定する。これにより、条件情報ＪＪ１の下で、タスク「タスク分類」をニューラルネットワークモデルＮＮＭがハードウェアＨＷ１上で実行するときの合計所要時間ＧＳＪ（後に詳述する。）は、

中間層Ｌ１～Ｌ７に割り当てられた演算ＯＰ１、ＯＰ２、ＯＰ３、ＯＰ５の所要時間Ｔ１、Ｔ２、Ｔ３、Ｔ５の合計となる。

[0043] 生成部ＳＥは、例えば、条件情報ＪＪ１により特定される探索空間ＴＫ１（図４に図示。）の範囲内で、条件情報ＪＪ１の下でニューラルネットワークモデルＮＮＭに要求される要求所要時間ＹＳＪ１を満足するように、ニューラルネットワークモデルＮＮＭの中間層Ｌ１～Ｌ７（図２に図示。）に演算ＯＰ１等を割り当てる。生成部ＳＥは、これにより、演算組み合わせＥＫ（図３に図示。）を生成する。生成部ＳＥは、前記演算組み合わせＥＫの生成を、例えば、ランダムサーチ、ベイズ最適化、遺伝的アルゴリズムを用いて行う。生成部ＳＥによる生成は、例えば、生成を行う回数、及び生成を行う時間を特に制限されない。

[0044] 訓練部ＫＵは、生成部ＳＥにより生成されたニューラルネットワークモデルＮＮＭを訓練する。訓練部ＫＵは、より詳しくは、例えば、条件情報ＪＪ１（図４に図示。）の下で、中間層Ｌ１～Ｌ７（図２に図示。）に探索空間ＴＫ１（図４に図示。）内の演算ＯＰ１等が割り当てられ、かつ、割り当てられた演算ＯＰ１等の所要時間の合計である合計所要時間ＧＳＪが要求所要時間ＹＳＪ１を満足するニューラルネットワークモデルＮＮＭを訓練する。訓練部ＫＵは、前記ニューラルネットワークモデルＮＮＭの訓練を、条件情報ＪＪ１のための訓練用データＫＤ１を用いて、条件情報ＪＪ１の下でニューラルネットワークモデルＮＮＭに要求される要求推論精度ＹＳＳ１を満足するように行う。これにより、訓練部ＫＵは、重み値ｗｔ（図３に図示。）を生成する。訓練部ＫＵは、前記訓練を、例えば、確率的勾配降下法を用いて行う。訓練部ＫＵによる訓練は、例えば、訓練を行う回数、及び訓練を行う時間を特に制限されない。

[0045] 第１の判定部ＨＡ１は、生成部ＳＥにより演算組み合わせＥＫが生成されたニューラルネットワークモデルＮＮＭ（図７に図示。）の所要時間が、要求所要時間ＹＳＪを満足するか否かを判定する。

[0046] 第２の判定部ＨＡ２は、訓練部ＫＵにより訓練されたニューラルネットワ

ークモデルＮＮＭ（図９に図示。）の推論精度が、要求推論精度ＹＳＳを満足するか否かを判定する。

[0047] ニューラルネットワーク構築装置ＮＮＫは、例えば、集中型システム、及び分散型システムのいずれの形態でもよい。

[0048] 特定部ＴＯは、「特定部」に対応し、生成部ＳＥは、「生成部」に対応し、第１の判定部ＨＡ１は、「第１の判定部」に対応し、第２の判定部ＨＡ２は、「第２の判定部」に対応し、訓練部ＫＵは、「訓練部」に対応する。また、条件情報ＪＪは、「構築条件」に対応する。

[0049] 〈〈実施形態１の構成〉〉

図５は、実施形態１のニューラルネットワーク構築装置ＮＮＫの構成を示す。

[0050] 実施形態１のニューラルネットワーク構築装置ＮＮＫは、上述した機能を果たすべく、図５に示されるように、プロセッサＰＣと、メモリＭＭと、記憶媒体ＫＢと、を含み、更に、必要に応じて、入力部ＮＹと、出力部ＳＹと、を含む。

[0051] プロセッサＰＣは、ソフトウェアに従ってハードウェアを動作させる、よく知られたコンピュータの中核である。メモリＭＭは、例えば、ＤＲＡＭ（Dynamic Random Access Memory）、ＳＲＡＭ（Static Random Access Memory）から構成される。記憶媒体ＫＢは、例えば、ハードディスクドライブ（ＨＤＤ：Hard Disk Drive）、ソリッドステートドライブ（ＳＳＤ：Solid State Drive）、ＲＯＭ（Read Only Memory）から構成される。記憶媒体ＫＢは、プログラムＰＲを記憶する。プログラムＰＲは、プロセッサＰＣが実行すべき処理の内容を規定する命令群である。

[0052] 入力部ＮＹは、例えば、キーボード、マウス、タッチパネルから構成される。出力部ＳＹは、例えば、液晶モニター、プリンタ、タッチパネルから構成される。

[0053] ニューラルネットワーク構築装置ＮＮＫにおける機能と構成との関係については、ハードウェア上で、プロセッサＰＣが、記憶媒体ＫＢに記憶された

プログラムＰＲを、メモリＭＭ上で実行すると共に、必要に応じて、入力部ＮＹ及び出力部ＳＹの動作を制御することにより、インターフェース部ＩＦ～第２の判定部ＨＡ２（図１に図示。）の各部の機能を実現する。

[0054] 〈〈ニューラルネットワーク構築装置ＮＮＫの動作〉〉

図６は、実施形態１のニューラルネットワーク構築装置ＮＮＫの動作を示すフローチャートである。以下、実施形態１のニューラルネットワーク構築装置ＮＮＫの動作について、図６のフローチャートを参照して説明する。

[0055] 以下では、説明及び理解を容易にすべく、ユーザが、条件情報ＪＪ、要求所要時間ＹＳＪ、要求推論精度ＹＳＳ、及び訓練用データＫＤとして、上記した、条件情報ＪＪ１、要求所要時間ＹＳＪ１、要求推論精度ＹＳＳ１、及び訓練用データＫＤ１を入力することを想定する。

[0056] ステップＳＴ１１：取得部ＳＹ（図１に図示。）は、インターフェース部ＩＦ（図１に図示。）を介して、条件情報ＪＪ１、要求所要時間ＹＳＪ１、要求推論精度ＹＳＳ１、及び訓練用データＫＤ１を取得する。

[0057] ステップＳＴ１２：ステップＳＴ１１で、条件情報ＪＪ１、要求所要時間ＹＳＪ１、要求推論精度ＹＳＳ１、及び訓練用データＫＤ１が取得されると、特定部ＴＯ（図１に図示。）は、取得された条件情報ＪＪ１に基づき、テーブルＴＢ（図４に図示。）を参照することにより、探索空間ＴＫ１を特定する。

[0058] 図７は、実施形態１の生成部ＳＥの動作を示す。

[0059] ステップＳＴ１３：ステップＳＴ１２で、探索空間ＴＫ１が特定されると、生成部ＳＥ（図１に図示。）は、図７の上段に示されるように、ニューラルネットワーク構築装置ＮＮＫの中間層Ｌ１～Ｌ７に、探索空間ＴＫ１内の割り当て可能な演算を割り当て、具体的には、中間層Ｌ１に演算ＯＰ１を割り当て、中間層Ｌ２に演算ＯＰ１を割り当て、中間層Ｌ３に演算ＯＰ２を割り当て、中間層Ｌ４に演算ＯＰ３を割り当て、中間層Ｌ５に演算ＯＰ２を割り当て、中間層Ｌ６に演算ＯＰ５を割り当て、中間層Ｌ７に演算ＯＰ５を割り当てる。換言すれば、生成部ＳＥは、中間層Ｌ１～Ｌ７について、演算Ｏ

Ｐ１、ＯＰ１、ＯＰ２、ＯＰ３、ＯＰ２、ＯＰ５、ＯＰ５という、１回めの演算組み合わせＥＫを生成する。

[0060] 図８は、実施形態１の第１の判定部ＨＡ１の動作を示す。

[0061] ステップＳＴ１４：ステップＳＴ１３で、中間層Ｌ１～Ｌ７に演算ＯＰ１等が割り当てられることにより、１回めの演算組み合わせＥＫが生成されると、第１の判定部ＨＡ１は、図８に示されるように、ニューラルネットワーク構築装置ＮＮＫの合計所要時間ＧＳＪを算出する。第１の判定部ＨＡ１は、図８に示されるように、中間層Ｌ１に割り当てられた演算ＯＰ１の所要時間Ｔ１、中間層Ｌ２に割り当てられた演算ＯＰ１の所要時間Ｔ１、中間層Ｌ３に割り当てられた演算ＯＰ２の所要時間Ｔ２、中間層Ｌ４に割り当てられた演算ＯＰ３の所要時間Ｔ３、中間層Ｌ５に割り当てられた演算ＯＰ２の所要時間Ｔ２、中間層Ｌ６に割り当てられた演算ＯＰ５の所要時間Ｔ５、及び中間層Ｌ７に割り当てられた演算ＯＰ５の所要時間Ｔ５、を合計することにより、合計所要時間ＧＳＪ（Ｔ１＋Ｔ１＋Ｔ２＋Ｔ３＋Ｔ２＋Ｔ５＋Ｔ５）を算出する。

[0062] 第１の判定部ＨＡ１は、上記の算出に加えて、合計所要時間ＧＳＪ（Ｔ１＋Ｔ１＋Ｔ２＋Ｔ３＋Ｔ２＋Ｔ５＋Ｔ５）が、取得部ＳＹにより取得された要求所要時間ＹＳＪ１以下であるとの要件を満足するか否かを判定する。満足すると判定されるとき、処理は、ステップＳＴ１５に進み、他方で、満足しないと判定されるとき、処理は、ステップＳＴ１３に戻る。

[0063] ここで、以下を想定する。

（１）生成部ＳＥ及び第１の判定部ＨＡ１が、ステップＳＴ１３及びステップＳＴ１４を繰り返したこと。

（２）上記（１）の繰り返しの後、生成部ＳＥが、ｍ回め（ｍは、任意の正の整数）の演算組み合わせＥＫとして、例えば、図７の中段に示されるように、中間層Ｌ１に演算ＯＰ１を割り当て、中間層Ｌ２に演算ＯＰ１を割り当て、中間層Ｌ３に演算ＯＰ３を割り当て、中間層Ｌ４に演算ＮＯを割り当て、中間層Ｌ５に演算ＯＰ３を割り当て、中間層Ｌ６に演算ＯＰ５を割り当

て、中間層Ｌ７に演算ＯＰ５を割り当てたこと。

（３）上記（２）のｍ回めの演算組み合わせＥＫの生成後に、第１の判定部ＨＡ１が、ｍ回めの演算組み合わせＥＫの合計所要時間ＧＳＪ（Ｔ１＋Ｔ１＋Ｔ３＋０＋Ｔ３＋Ｔ５＋Ｔ５）が、要求所要時間ＹＳＪ１を満足していると判定したこと。

[0064] 「ＮＯ演算」は、No Connection演算を意味する。「ＮＯ演算」は、上述したように、何らかのデータの入力を受けても、データを何ら出力しない機能を有する。「ＮＯ演算」は、その所要時間が、０である。

[0065] ステップＳＴ１５：直前のステップＳＴ１４で、合計所要時間ＧＳＪ（Ｔ１＋Ｔ１＋Ｔ３＋０＋Ｔ３＋Ｔ５＋Ｔ５）が、要求所要時間ＹＳＪ１を満足していると判定されると、訓練部ＫＵ（図１に図示。）は、ｍ回めの演算組み合わせＥＫのニューラルネットワークモデルＮＮＭ（図７の中段に図示。）を訓練する。訓練部ＫＵは、より詳しくは、取得部ＳＹにより取得された訓練用データＫＤ１を用いて、従来知られたと同様に、出力層Ｌｏｕｔから出力されるデータ（図示せず。）と訓練用データＫＤ１中のラベル（図示せず。）との間の誤差（図示せず。）がより小さくなるように、ニューラルネットワークモデルＮＮＭ中の重みｗ（図２に図示。）、正確には、重み値ｗｔ（図３に図示。）を調整する。

[0066] 訓練部ＫＵは、上記した訓練を、予め定められた時間だけ、若しくは予め定められた回数だけ続けた後に終了し、または、訓練による学習効果の進捗に応じて続けた後に終了する。

[0067] ステップＳＴ１６：ステップＳＴ１５で、ニューラルネットワークモデルＮＮＭの訓練が終了すると、第２の判定部ＨＡ２（図１に図示。）は、訓練を受けたニューラルネットワークモデルＮＮＭの推論精度ＳＳ（図示せず。）が、取得部ＳＹによる取得された要求推論精度ＹＳＳ１以上であるとの要件を満足するか否かを判定する。満足すると判定されるとき、処理は、ステップＳＴ１７に進み、他方で、満足しないと判定されるとき、処理は、ステップＳＴ１３に戻る。

[0068] 図９は、実施形態１の訓練部ＫＵの動作を示す。

[0069] ここで、以下を想定する。

（１）生成部ＳＥ、第１の判定部ＨＡ１、訓練部ＫＵ、及び、第２の判定部ＨＡ２が、ステップＳＴ１３～ステップＳＴ１６を繰り返したこと。

（２）上記（１）の繰り返しの後、生成部ＳＥが、ｎ回め（ｎは、ｍより大きい整数）の演算の組み合わせＥＫとして、例えば、図７の下段に示されるように、中間層Ｌ１に演算ＯＰ１を割り当て、中間層Ｌ２に演算ＯＰ１を割り当て、中間層Ｌ３に演算ＯＰ２を割り当て、中間層Ｌ４に演算ＯＰ２を割り当て、中間層Ｌ５に演算ＳＫＩＰを割り当て、中間層Ｌ６に演算ＯＰ５を割り当て、中間層Ｌ７に演算ＯＰ１を割り当てたこと。

（３）上記（２）のｎ回めの演算組み合わせＥＫの生成後に、第１の判定部ＨＡ１が、ｎ回めの演算組み合わせＥＫの合計所要時間ＧＳＪ（Ｔ１＋Ｔ１＋Ｔ２＋Ｔ２＋０＋Ｔ５＋Ｔ１）が、要求所要時間ＹＳＪ１を満足していると判定したこと。

[0070] 「ＳＫＩＰ演算」は、Skip Connection演算を意味する。「ＳＫＩＰ演算」は、上述したように、何ら処理を行うことなく、入力されたデータをそのまま出力する機能を有する。「ＳＫＩＰ演算」は、その所要時間が、０である。

[0071] （４）上記（３）の判定後に、訓練部ＫＵが、ｎ回めの演算組み合わせＥＫのニューラルネットワークモデルＮＮＭの重み値ｗｔを調整し、例えば、重み値ｗｔをｗｔａに設定し、ｗｔｂに変更し、ｗｔｃに更に変更したこと。

（５）上記（４）の訓練後に、第２の判定部ＨＡ２が、ｎ回めの演算組み合わせＥＫの推論精度ＳＳが、要求推論精度ＹＳＳ１を満足していると判定したこと。

[0072] ステップＳＴ１７：直前のステップＳＴ１６で、ｎ回めの演算組み合わせＥＫのニューラルネットワークモデルＮＮＭの推論精度ＳＳが、要求推論精度ＹＳＳ１を満足していると判定されると、第２の判定部ＨＡ２は、構築デ

ータＣＤ、即ち、ｎ回めの演算組み合わせＥＫ（図７の下段、図９に図示。）、及び、重み値ｗｔである重み値ｗｔｃ（図９の下段に図示。）を記憶部ＫＩ（図３に図示。）に保存し、または、出力する。

[0073] 〈〈ニューラルネットワーク構築装置ＮＮＫの効果〉〉

上述したように、実施形態１のニューラルネットワーク構築装置ＮＮＫでは、構築しようとするニューラルネットワークモデルＮＮＭの中間層Ｌ１～Ｌ７に、ユーザにより入力される条件情報ＪＪ１によって特定される探索空間ＴＫ１中の割り当て可能な演算ＯＰ１等を割り当てる。これにより、要求所要時間ＹＳＪ１を満足するニューラルネットワークモデルＮＮＭを構築する可能性を高めることができる。

[0074] 実施形態１のニューラルネットワーク構築装置ＮＮＫでは、また、訓練用データＫＤ１を用いた訓練を行う。これにより、要求推論精度ＹＳＳ１を満足するニューラルネットワークモデルＮＮＭを構築する可能性を高めることができる。

[0075] 実施形態１のニューラルネットワーク構築装置ＮＮＫでは、さらに、ニューラルネットワークモデルＮＮＭの中間層Ｌ１～Ｌ７の少なくとも１つ以上に、No Connection演算、及びSkip Connection演算を割り当てることにより、実質的に機能しない層を増加させることによって、換言すれば、実質的に機能する層の数を減少させることによって、ニューラルネットワークモデルＮＮＭの全体としての演算量を低減させることができる。

[0076] 実施形態１のニューラルネットワーク構築装置ＮＮＫは、例えば、自動運転のための画像処理を行うべく、自家用車等の車両に搭載される。ニューラルネットワーク構築装置ＮＮＫには、メーカによる車両の製造段階で、画像データが、予め記憶されている。ニューラルネットワーク構築装置ＮＮＫには、加えて、運転手による車両の走行段階で、画像データが、蓄積され続ける。ニューラルネットワーク構築装置ＮＮＫは、前記した両画像についてのクラス分類を、例えば、検出された物体が、人、動物、樹木等のいずれであるかの分類を、上記した要求所要時間ＹＳＪ１及び要求推論精度ＹＳＳ１を

満足するように実行するニューラルネットワークモデルＮＮＭを構築することができる。

[0077] 〈変形例１〉

中間層Ｌ１～Ｌ７での所要時間を合計する実施形態１のニューラルネットワーク構築装置ＮＮＫとは相違し、中間層Ｌ１～Ｌ７での所要時間、並びに、演算ＯＰが割り当てられない、入力層Ｌｉｎ及び出力層Ｌｏｕｔでの所要時間を合計した上で、合計所要時間ＧＳＪが、要求所要時間ＹＳＪ１を満足するか否かを判定してもよい。

[0078] 〈変形例２〉

ｎ番めの演算組み合わせＥＫのニューラルネットワークモデルＮＮＭのみを生成する実施形態１のニューラルネットワーク構築装置ＮＮＫとは相違し、例えば、ｐ番め（ｐは、ｎより大きい整数）の演算組み合わせＥＫ、ｑ番め（ｑは、ｐより大きい整数）の演算組み合わせＥＫ、、、、等の複数の演算組み合わせＥＫを生成し、当該複数の演算組み合わせＥＫ、即ち、ｎ番めの演算組み合わせＥＫ、ｐ番めの演算組み合わせＥＫ、ｑ番めの演算組み合わせＥＫ、、、、のうち、要求所要時間ＹＳＪ１及び要求推論精度ＹＳＳ１の観点から最も優れた演算組み合わせＥＫを選択してもよい。

[0079] 実施形態２.

〈実施形態２〉

実施形態２のニューラルネットワーク構築装置ＮＮＫについて説明する。

[0080] 〈〈ニューラルネットワーク構築装置ＮＮＫの機能〉〉

実施形態１のニューラルネットワーク構築装置ＮＮＫが、記憶部ＫＩ（図１、図３に図示。）内に記憶されたテーブルＴＢ（図４に図示。）中の演算所要時間（予め準備され、想定された演算所要時間）の合計所要時間ＧＳＪ（図８に図示。）を用いることとは相違し、実施形態２のニューラルネットワーク構築装置ＮＮＫでは、ニューラルネットワークモデルＮＮＭが実際に演算したときの所要時間である実所要時間ＪＳＪを測定する。

[0081] 図１０は、実施形態２のニューラルネットワーク構築装置ＮＮＫの機能ブ

ロック図である。

[0082] 実施形態２のニューラルネットワーク構築装置ＮＮＫは、図１０に示されるように、実施形態１のニューラルネットワーク構築装置ＮＮＫ（図１に図示。）と同様に、記憶部ＫＩ～第２の判定部ＨＡ２を含み、他方で、実施形態１のニューラルネットワーク構築装置ＮＮＫと相違して、実行部ＪＩを更に含む。

[0083] 実施形態２のインターフェース部ＩＦ～第２の判定部ＨＡ２の機能は、実施形態１のインターフェース部ＩＦ～第２の判定部ＨＡ２と同様である。

[0084] 実行部ＪＩは、例えば、条件情報ＪＪ１（図４に図示。）により特定されるハードウェア「ＨＷ１」上に設けられている。実行部ＪＩは、条件情報ＪＪ１により特定される探索空間ＴＫ１内の割り当て可能な演算ＯＰ１等が中間層Ｌ１～Ｌ７に割り当てられたニューラルネットワークモデルＮＮＭ（図２に図示。）を、ハードウェア「ＨＷ１」上で、換言すれば、実行部ＪＩ上で実際に動作させることにより、ニューラルネットワークモデルＮＮＭの実所要時間ＪＳＪを測定する。

[0085] 〈〈ニューラルネットワーク構築装置ＮＮＫの構成〉〉

実施形態２のニューラルネットワーク構築装置ＮＮＫの構成は、実施形態１のニューラルネットワーク構築装置ＮＮＫの構成（図５に図示。）と同様である。

[0086] 〈〈ニューラルネットワーク構築装置ＮＮＫの動作〉〉

図１１は、実施形態２のニューラルネットワーク構築装置ＮＮＫの動作を示すフローチャートである。以下、実施形態２のニューラルネットワーク構築装置ＮＮＫの動作について、図１１のフローチャートを参照して説明する。

[0087] 説明及び理解を容易にすべく、以下を想定する。

（１）実施形態１と同様に、ユーザが、条件情報ＪＪ１、要求所要時間ＹＳＪ１、及び要求推論精度ＹＳＳ１、訓練用データＫＤを入力すること。

（２）実施形態１と同様に、１番めの演算組み合わせＥＫ、ｍ番めの演算組

み合わせＥＫ等を経て、ｎ番めの演算組み合わせＥＫが生成されたこと。

[0088] ステップＳＴ２１～ＳＴ２３：取得部ＳＹ、特定部ＴＯ、及び生成部ＳＥは、実施形態１のステップＳＴ１１～ＳＴ１３と同様の処理を行う。

[0089] ステップＳＴ２４：ステップＳＴ２３で、条件情報ＪＪ１により特定される探索空間ＴＫ１内の割り当て可能な演算ＯＰ１等が中間層Ｌ１～Ｌ７に割り当てられると、即ち、ｎ番めの演算組み合わせＥＫ（図７の下段に図示。）を有するニューラルネットワークモデルＮＮＭが生成されると、実行部ＪＩは、ニューラルネットワークモデルＮＮＭを、ハードウェアＨＷ１上で実際に動作させることにより、実所要時間ＪＳＪを測定する。

[0090] ステップＳＴ２５：ステップＳＴ２４で、実所要時間ＪＳＪが測定されると、第１の判定部ＨＡ１は、ｎ番めの演算組み合わせＥＫの実所要時間ＪＳＪが、要求所要時間ＹＳＪ１を満足するか否かを判定する。

[0091] ステップＳＴ２６～ＳＴ２８：訓練部ＫＵ、及び第２の判定部ＨＡ２は、実施形態１のステップＳＴ１５～ＳＴ１７と同様の処理を行う。

[0092] 〈〈ニューラルネットワーク構築装置ＮＮＫの効果〉〉

上述したように、実施形態２のニューラルネットワーク構築装置ＮＮＫでは、ニューラルネットワークモデルＮＮＭの実所要時間ＪＳＪ、例えば、ｎ番めの組み合わせのニューラルネットワークモデルＮＮＭがハードウェアＨＷ１で実際に動作するときの実所要時間ＪＳＪを測定し、用いる。これにより、記憶部ＫＩに予め記憶されており、かつ見積もられている演算所要時間（図４に図示。）を用いて構築される実施形態１のニューラルネットワークモデルＮＮＭに比して、要求所要時間ＹＳＪ１を満足する可能性がより高められたニューラルネットワークモデルＮＮＭを構築することができる。

[0093] 実施形態３．

〈実施形態３〉

実施形態３の画像処理装置について説明する。

[0094] 〈〈画像処理装置ＧＳＳの機能、構成、及び動作〉〉

図１２は、実施形態３の画像処理装置ＧＳＳの機能を示す。

[0095] 実施形態３の画像処理装置ＧＳＳは、図１２に示されるように、記憶部ＫＯと、入力部ＮＢと、推論部ＳＵと、出力部ＳＢと、を含む。

[0096] 記憶部ＫＯは、実施形態１及び実施形態２のいずれかのニューラルネットワーク構築装置ＮＮＫにより、条件情報ＪＪ１等の条件の下で構築されたニューラルネットワークモデルＮＮＭ、正確には、ｎ番めの演算組み合わせＥＫ及び重み値ｗｔが記憶されている。

[0097] 入力部ＮＢは、例えば、撮像カメラである。入力部ＮＢは、例えば、画像データＧＤを取得する。

[0098] 推論部ＳＵは、例えば、条件情報ＪＪ１により特定されるハードウェアＨＷ１上に設けられている。推論部ＳＵは、従来知られたと同様に、推論を行い、具体的には、画像データＧＤを用いてニューラルネットワークモデルＮＮＭを実行する。

[0099] 出力部ＳＢは、例えば、液晶モニター、スピーカ等である。出力部ＳＢは、推論部ＳＵが推論を行った結果である推論結果ＳＫを出力する。

[0100] 記憶部ＫＯは、「記憶部」に対応し、入力部ＮＢは、「入力部」に対応し、出力部ＳＢは、「出力部」に対応し、画像データＧＤは、「画像データ」に対応し、「推論結果ＳＫ」は、「推論した結果」に対応する。

[0101] 〈〈画像処理装置ＧＳＳの効果〉〉

上述したように、実施形態３の画像処理装置ＧＳＳは、実施形態１または実施形態２のニューラルネットワーク構築装置ＮＮＫにより構築された、要求所要時間ＹＳＪ１及び要求推論精度ＹＳＳ１を満足するｎ番めの演算組み合わせＥＫを有するニューラルネットワークモデルＮＮＭを、条件情報ＪＪ１により特定されるハードウェアＨＷ１上で実行する。これにより、要求所要時間ＹＳＪ１及び要求推論精度ＹＳＳ１を満足する推論を行うことができる。

[0102] 本開示の要旨を逸脱しない範囲で、上述した実施形態同士を組み合わせてもよく、また、各実施形態中の構成要素を適宜、削除し、変更し、または、他の構成要素を追加してもよい。

産業上の利用可能性

[0103] 本開示に係るニューラルネットワーク構築装置は、推論についての所要時間及び精度を満足することができるニューラルネットワークモデルを構築することができるニューラルネットワーク構築装置に適用可能である。

符号の説明

[0104] ＣＤ　構築データ、ＥＫ　演算組み合わせ、ＧＤ　画像データ、ＧＳＪ　合計所要時間、ＧＳＳ　画像処理装置、ＨＡ１　第１の判定部、ＨＡ２　第２の判定部、ＨＷ　ハードウェア、ＩＦ　インターフェース部、ＪＩ　実行部、ＪＪ　条件情報、ＪＳＪ　実所要時間、ＫＢ　記憶媒体、ＫＤ　訓練用データ、ＫＩ　記憶部、ＫＯ　記憶部、ＫＵ　訓練部、Ｌ　中間層、Ｌｉｎ　入力層、Ｌｏｕｔ　出力層、ＭＭ　メモリ、ＮＢ　入力部、ＮＮＫ　ニューラルネットワーク構築装置、ＮＮＭ　ニューラルネットワークモデル、ＮＹ　入力部、ＯＰ　演算、ＰＲ　プログラム、ＳＢ　出力部、ＳＥ　生成部、ＳＫ　推論結果、ＳＳ　推論精度、ＳＵ　推論部、ＳＹ　取得部、ＳＹ　出力部、ＴＢ　テーブル、ＴＫ　探索空間、ＴＯ　特定部、ｗ　重み、ｗｔ　重み値、ＹＳＪ　要求所要時間、ＹＳＳ　要求推論精度。

請求の範囲

[請求項1] ニューラルネットワークモデルを構築するための構築条件に基づき、前記ニューラルネットワークモデルを構成する複数の層に割り当て可能な複数の演算を特定する特定部と、

前記複数の層の各々に、前記複数の演算のうちの一つの演算を割り当てることにより、演算の一の組み合わせを生成する生成部と、

前記割り当てられた演算に基づき推論する、各演算の所要時間の合計である合計所要時間が、予め定められた要求所要時間を満足するか否かを判定する第１の判定部と、

前記生成された演算の一の組み合わせを有するニューラルネットワークモデルを訓練する訓練部と、

前記訓練後のニューラルネットワークモデルが推論する精度が、予め定められた要求推論精度を満足するか否かを判定する第２の判定部と、を含み、

前記第１の判定部が、前記合計所要時間が前記要求所要時間を満足しないと判定するとき、または、前記第２の判定部が、前記推論する精度が前記要求推論精度を満足しないと判定するとき、前記生成部は、前記演算の一の組み合わせと相違する、演算の他の組み合わせを生成する、

ニューラルネットワーク構築装置。

[請求項2] 前記特定部は、前記複数の演算に、入力されたデータをそのまま出力するスキップコネクション演算、及び何ら出力しないノーコネクション演算を含ませる、

請求項１に記載のニューラルネットワーク構築装置。

[請求項3] 前記第１の判定部は、前記合計所要時間に、前記複数の層のうち、前記生成部が前記複数の演算を割り当てなかった層での所要時間を含ませる、

請求項１に記載のニューラルネットワーク構築装置。

[請求項4] 前記生成された演算の一の組み合わせを有するニューラルネットワークモデルを前記構築条件の下で実行する実行部を更に含み、

前記第１の判定部は、前記生成された演算の一の組み合わせを有するニューラルネットワークモデルを実行する実所要時間が、前記予め定められた要求所要時間を満足するか否かを判定する、

請求項１に記載のニューラルネットワーク構築装置。

[請求項5] 請求項１に記載のニューラルネットワーク構築装置により構築されたニューラルネットワークモデルを記憶する記憶部と、

前記ニューラルネットワークモデルに推論させるための画像データの入力を受ける入力部と、

前記ニューラルネットワークモデルが推論した結果を出力する出力部と、

を含む、画像処理装置。

[請求項6] 特定部が、ニューラルネットワークモデルを構築するための構築条件に基づき、前記ニューラルネットワークモデルを構成する複数の層に割り当て可能な複数の演算を特定し、

生成部が、前記複数の層の各々に、前記複数の演算のうちの一つの演算を割り当てることにより、演算の一の組み合わせを生成し、

第１の判定部が、前記割り当てられた演算に基づき推論する、各演算の所要時間の合計である合計所要時間が、予め定められた要求所要時間を満足するか否かを判定し、

訓練部が、前記生成された演算の一の組み合わせを有するニューラルネットワークモデルを訓練し、

第２の判定部が、前記訓練後のニューラルネットワークモデルが推論する精度が、予め定められた要求推論精度を満足するか否かを判定し、

前記第１の判定部が、前記合計所要時間が前記要求所要時間を満足しないと判定するとき、または、前記第２の判定部が、前記推論する

精度が前記要求推論精度を満足しないと判定するとき、前記生成部は、前記演算の一の組み合わせと相違する、演算の他の組み合わせを生成する、

ニューラルネットワーク構築方法。

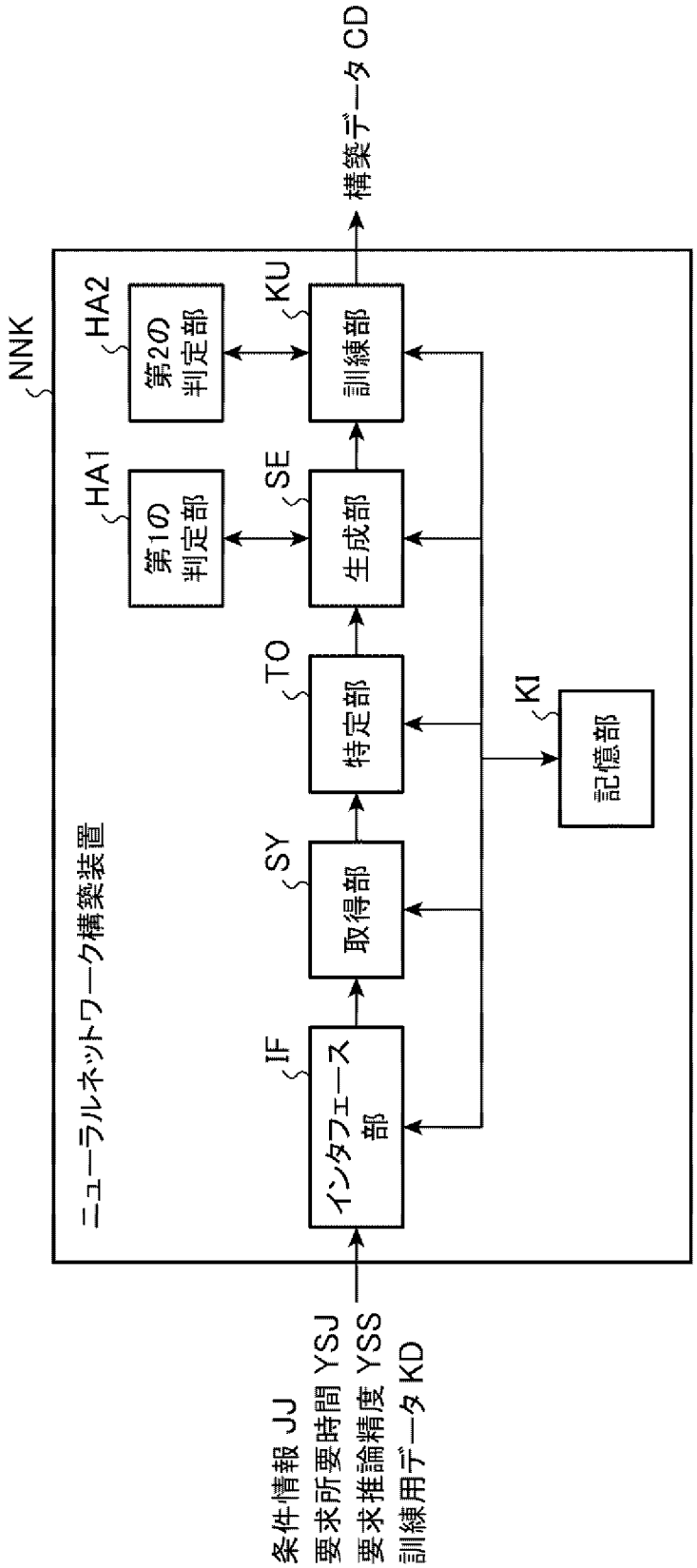
[請求項7] 記憶部が、請求項６に記載のニューラルネットワーク構築装置により構築されたニューラルネットワークモデルを記憶し、

入力部が、前記ニューラルネットワークモデルに推論させるための画像データの入力を受け、

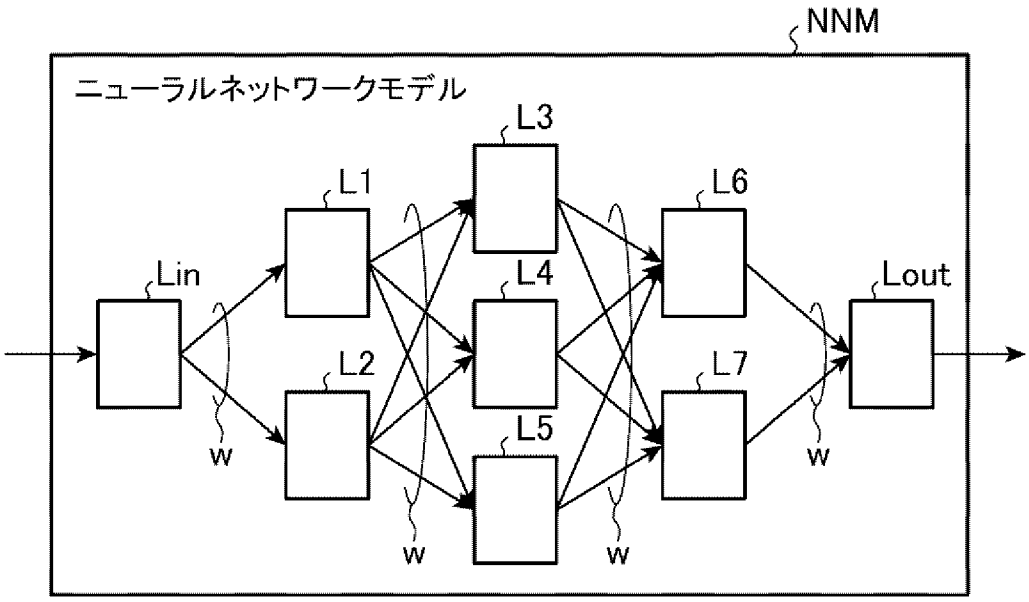
出力部が、前記ニューラルネットワークモデルが推論した結果を出力する、

画像処理方法。

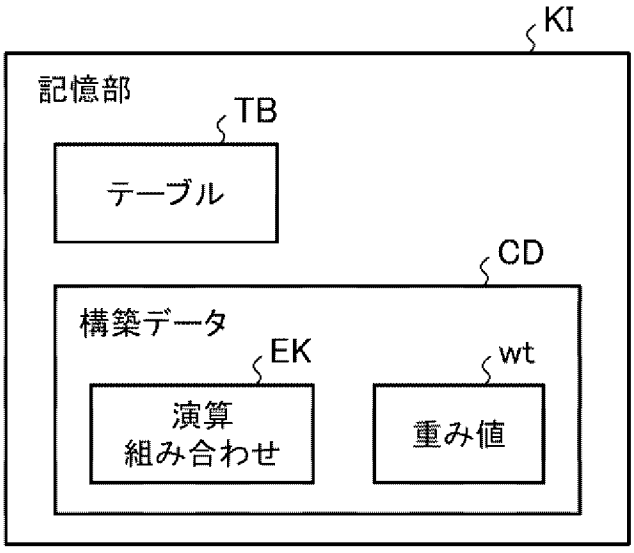
[図1]


NNK
ニューラルネットワーク構築装置
IF
インタフェース部
SY
取得部
TO
特定部
SE
生成部
KU
訓練部
HA1
第1の判定部
HA2
第2の判定部
KI
記憶部
条件情報 JJ
要求所要時間 YSJ
要求推論精度 YSS
訓練用データ KD
構築データ CD

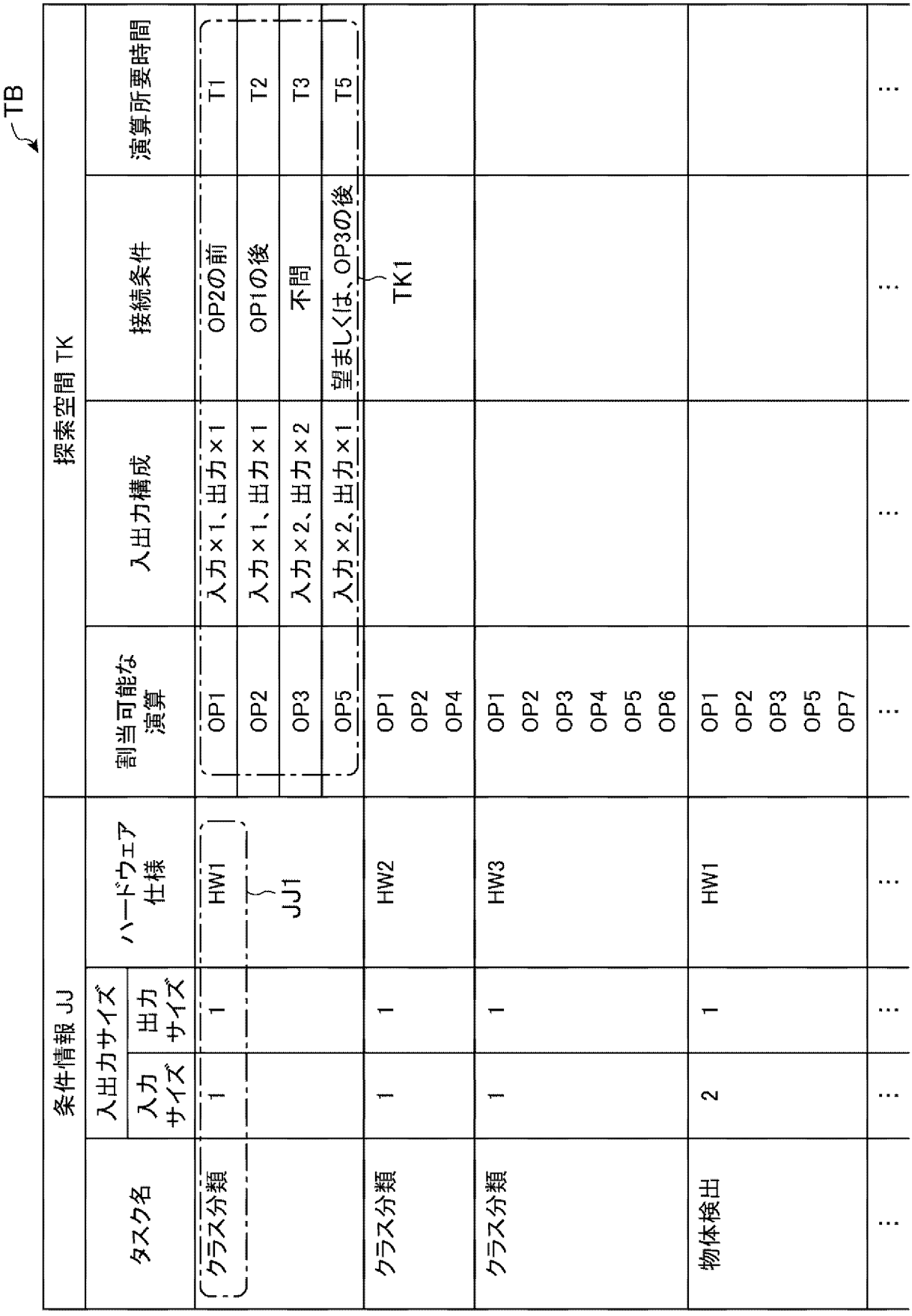
[図2]


NNM
ニューラルネットワークモデル
Lin
L1
L2
L3
L4
L5
L6
L7
Lout
w
w
w
w

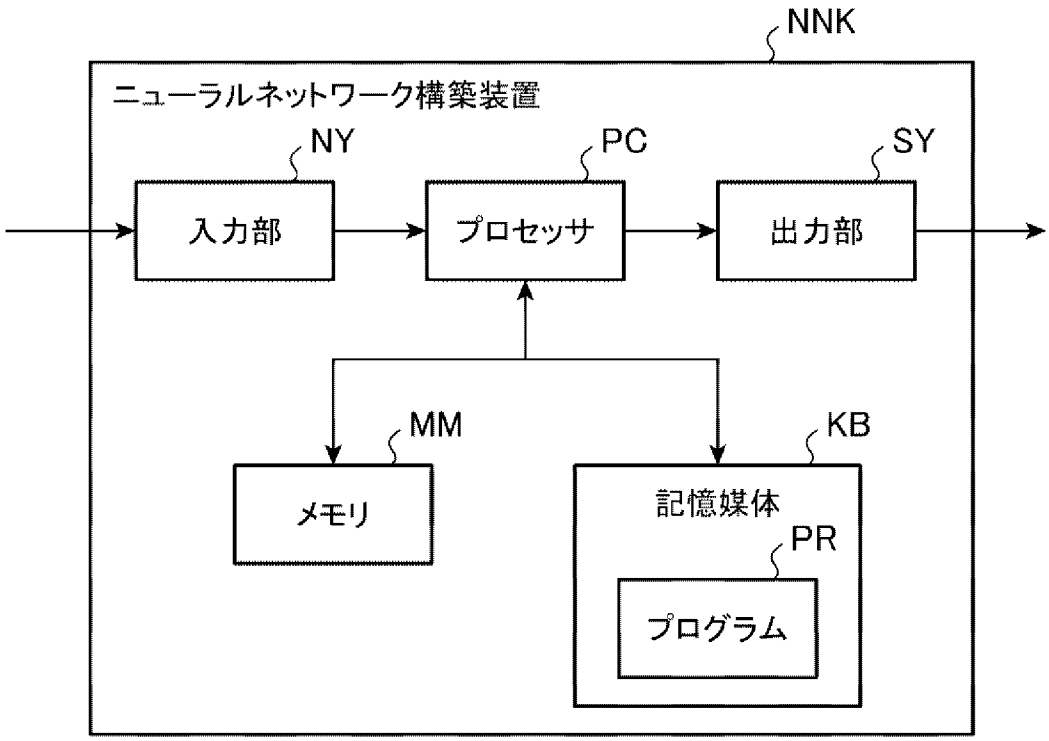
[図3]


KI
記憶部
TB
テーブル
CD
構築データ
EK
演算
組み合わせ
wt
重み値

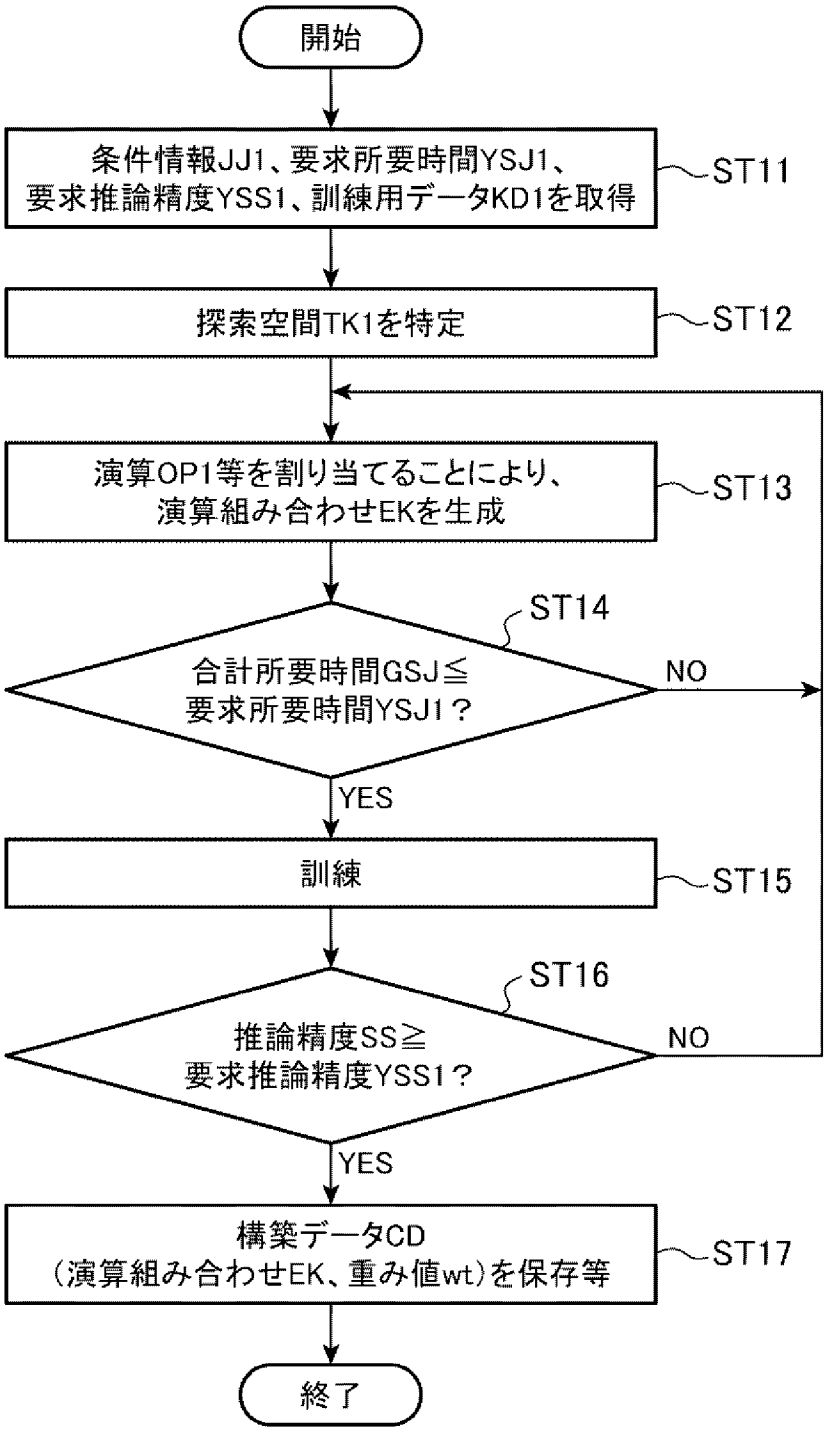
[図4]


TB
探索空間 TK
条件情報 JJ
タスク名
入出力サイズ
入力サイズ
出力サイズ
ハードウェア仕様
割当可能な演算
入出力構成
接続条件
演算所要時間
クラス分類
1
1
HW1
JJ1
OP1
入力×1、出力×1
OP2の前
T1
OP2
入力×1、出力×1
OP1の後
T2
OP3
入力×2、出力×2
不問
T3
OP5
入力×2、出力×1
望ましくは、OP3の後
T5
TK1
クラス分類
1
1
HW2
OP1
OP2
OP4
クラス分類
1
1
HW3
OP1
OP2
OP3
OP4
OP5
OP6
物体検出
2
1
HW1
OP1
OP2
OP3
OP5
OP7

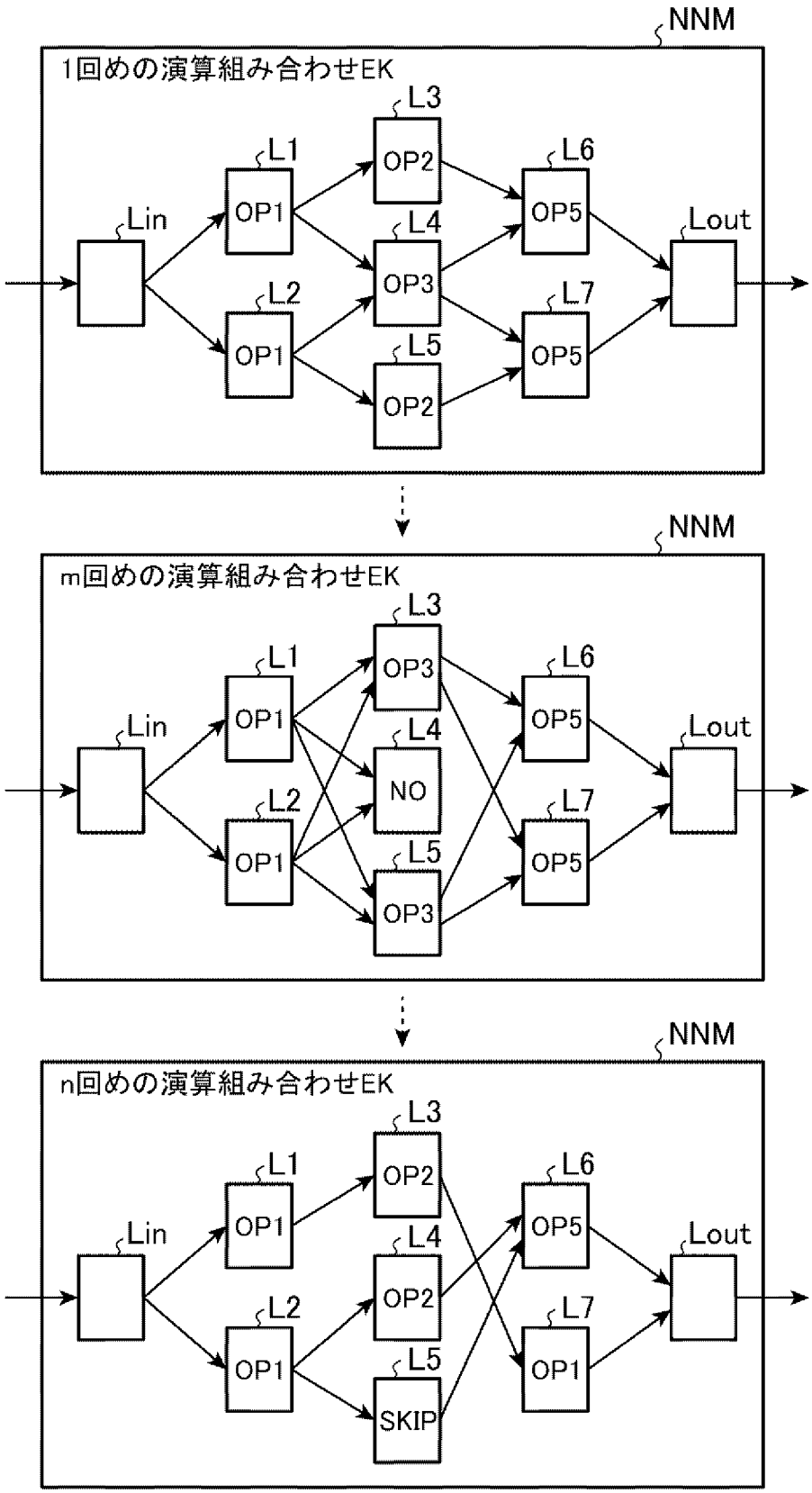
[図5]


NNK
ニューラルネットワーク構築装置
NY
PC
SY
入力部
プロセッサ
出力部
MM
KB
メモリ
記憶媒体
PR
プログラム

[図6]


開始
条件情報JJ1、要求所要時間YSJ1、
要求推論精度YSS1、訓練用データKD1を取得
ST11
探索空間TK1を特定
ST12
演算OP1等を割り当てることにより、
演算組み合わせEKを生成
ST13
ST14
合計所要時間GSJ≦
要求所要時間YSJ1？
NO
YES
訓練
ST15
ST16
推論精度SS≧
要求推論精度YSS1？
NO
YES
構築データCD
（演算組み合わせEK、重み値wt）を保存等
ST17
終了

[図7]

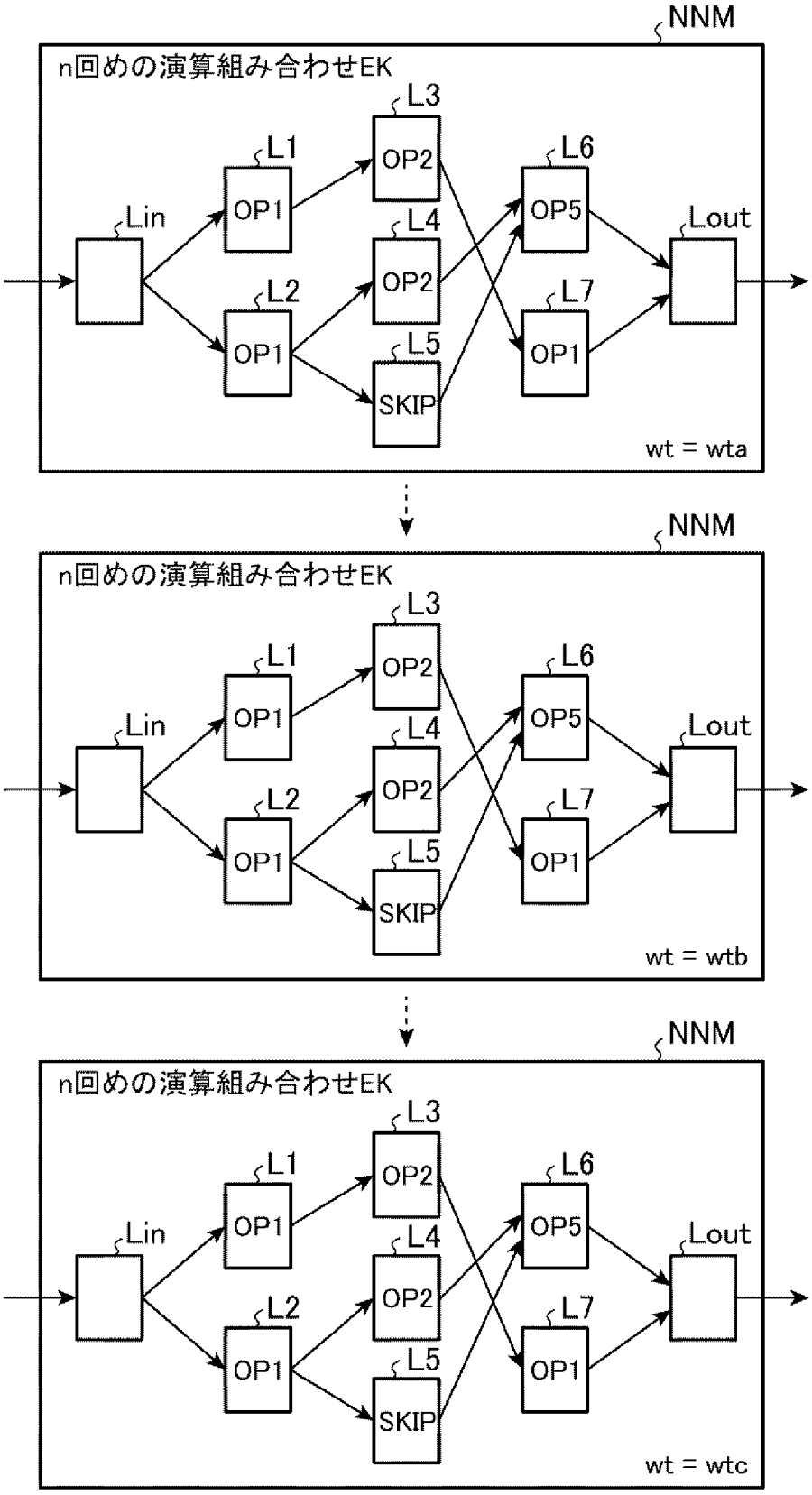

NNM
1回めの演算組み合わせEK
Lin
L1
OP1
L2
OP1
L3
OP2
L4
OP3
L5
OP2
L6
OP5
L7
OP5
Lout
NNM
m回めの演算組み合わせEK
Lin
L1
OP1
L2
OP1
L3
OP3
L4
NO
L5
OP3
L6
OP5
L7
OP5
Lout
NNM
n回めの演算組み合わせEK
Lin
L1
OP1
L2
OP1
L3
OP2
L4
OP2
L5
SKIP
L6
OP5
L7
OP1
Lout

[図8]

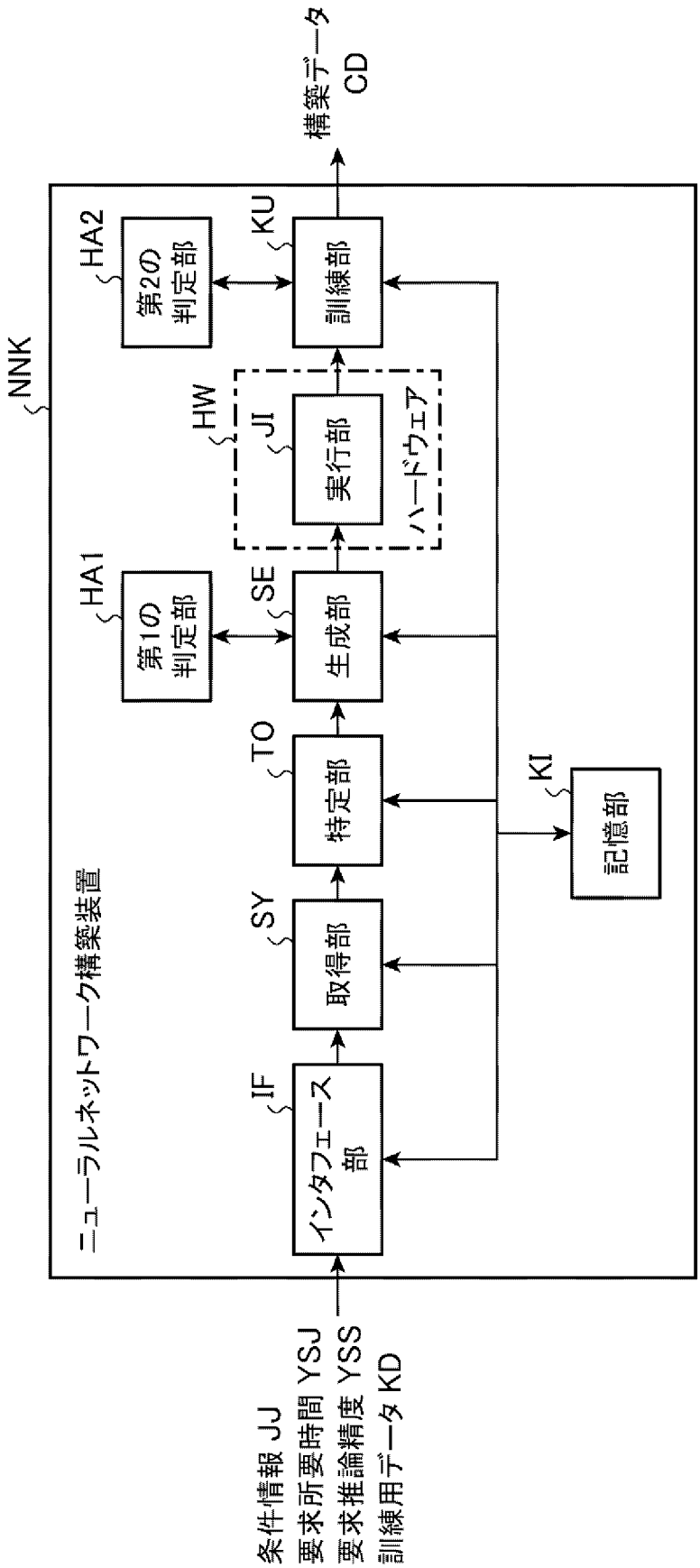
組み合わせ	層							合計所要時間GSJ
	L1	L2	L3	L4	L5	L6	L7	
1回め	OP1	OP1	OP2	OP3	OP2	OP5	OP5	–
	T1	T1	T2	T3	T2	T5	T5	T1+T1+T2+T3+T2+T5+T5
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
m回め	OP1	OP1	OP3	NO	OP3	OP5	OP5	–
	T1	T1	T3	0	T3	T5	T5	T1+T1+T3+0+T3+T5+T5
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
n回め	OP1	OP1	OP2	OP2	SKIP	OP5	OP1	–
	T1	T1	T2	T2	0	T5	T1	T1+T1+T2+T2+0+T5+T1

上段:演算
下段:演算所要時間

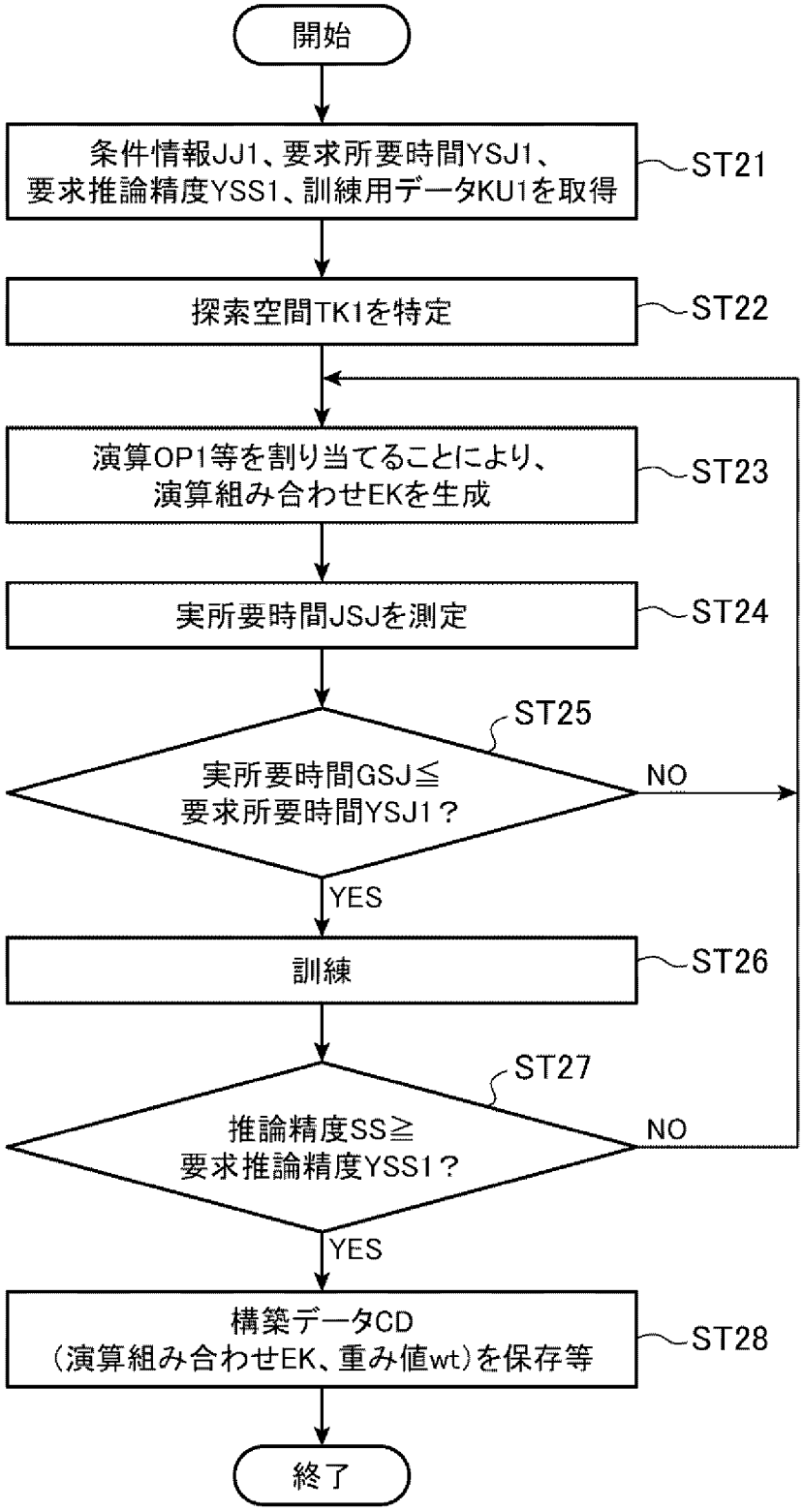
[図9]


NNM
n回めの演算組み合わせEK
Lin
L1
OP1
L2
OP1
L3
OP2
L4
OP2
L5
SKIP
L6
OP5
L7
OP1
Lout
wt = wta
NNM
n回めの演算組み合わせEK
Lin
L1
OP1
L2
OP1
L3
OP2
L4
OP2
L5
SKIP
L6
OP5
L7
OP1
Lout
wt = wtb
NNM
n回めの演算組み合わせEK
Lin
L1
OP1
L2
OP1
L3
OP2
L4
OP2
L5
SKIP
L6
OP5
L7
OP1
Lout
wt = wtc

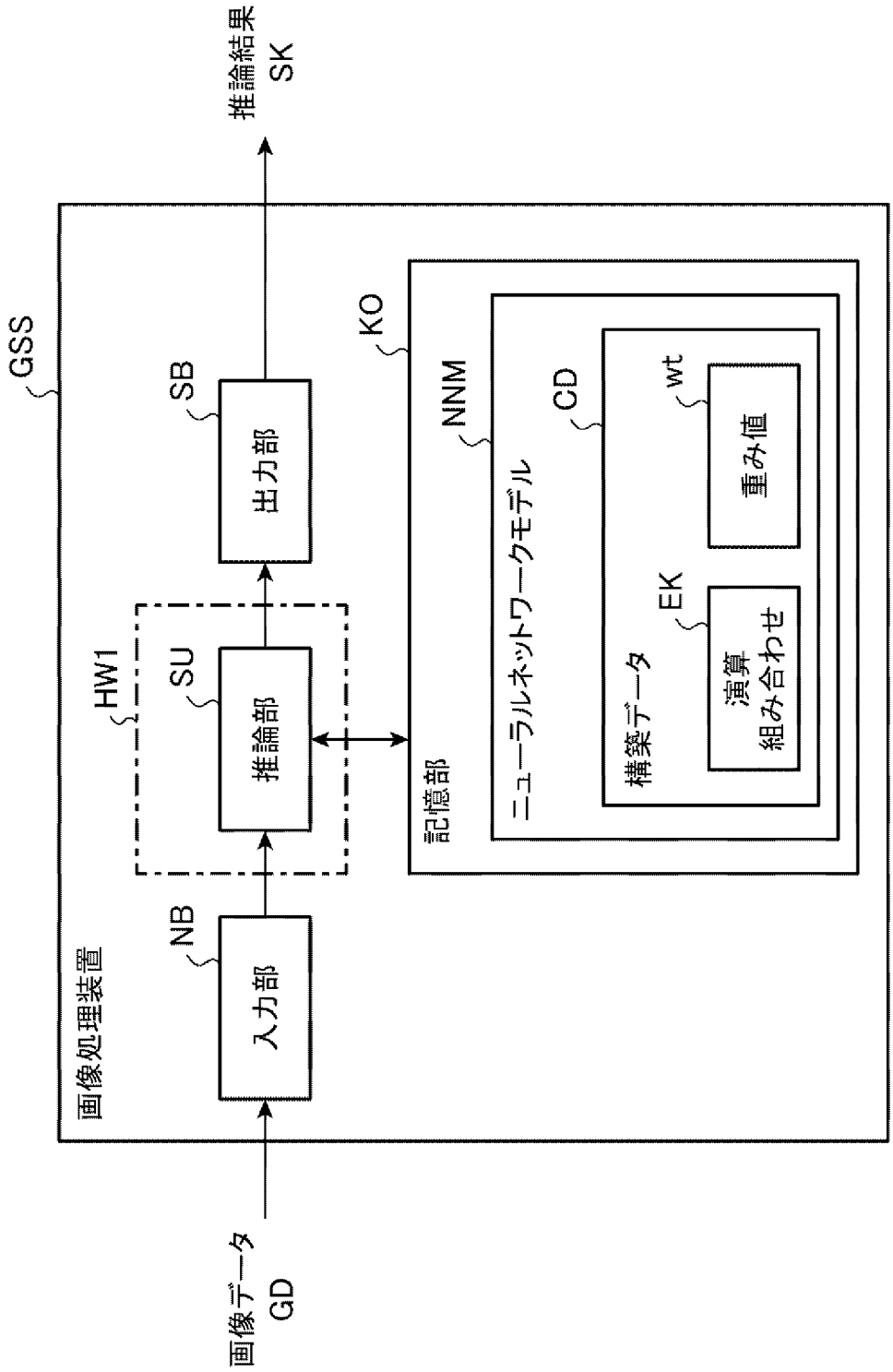
[図10]


構築データ
CD
NNK
ニューラルネットワーク構築装置
HA2
第2の
判定部
KU
訓練部
HW
JI
実行部
ハードウェア
HA1
第1の
判定部
SE
生成部
TO
特定部
KI
記憶部
SY
取得部
IF
インタフェース
部
条件情報 JJ
要求所要時間 YSJ
要求推論精度 YSS
訓練用データ KD

[図11]


開始
条件情報JJ1、要求所要時間YSJ1、
要求推論精度YSS1、訓練用データKU1を取得
ST21
探索空間TK1を特定
ST22
演算OP1等を割り当てることにより、
演算組み合わせEKを生成
ST23
実所要時間JSJを測定
ST24
ST25
実所要時間GSJ≦
要求所要時間YSJ1？
NO
YES
訓練
ST26
ST27
推論精度SS≧
要求推論精度YSS1？
NO
YES
構築データCD
（演算組み合わせEK、重み値wt）を保存等
ST28
終了

[図12]


画像データ
GD
推論結果
SK
GSS
画像処理装置
NB
入力部
HW1
SU
推論部
SB
出力部
KO
記憶部
NNM
ニューラルネットワークモデル
CD
構築データ
EK
演算
組み合わせ
wt
重み値

INTERNATIONAL SEARCH REPORT

International application No.
PCT/JP2021/022425

A. CLASSIFICATION OF SUBJECT MATTER
G06N 3/08(2006.01)i
FI: G06N3/08
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)
G06N3/00-99/00; G06T7/00

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.
Y	WU, Bichen et al. FBNet: Hardware-Aware Efficient ConvNet Design via Differentiable Neural Architecture Search. arXiv.org [online]. arXiv: 1812.03443v3. Cornell University., 2019, [retrieval date 05 August 2021], internet:<URL: https://arxiv.org/pdf/1812.03443v3> chapters 1, 3-3.2	1-7
Y	CN 112598117 A (XAG CO., LTD.) 02 April 2021 (2021-04-02) paragraphs [0006], [0038]-[0055]	1-7
Y	JIANG, Weiwen et al. Hardware/Software Co-Exploration of Neural Architectures. arXiv.org [online]. arXiv:1907.04650v2. Cornell University., 2020, pp. 1-10, [retrieval date 05 August 2021], internet:<URL:https://arxiv.org/pdf/1907.04650v2> chapters II A., III, fig. 1, 3	1-7

☒ 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	Date of mailing of the international search report
05 August 2021 (05.08.2021)	17 August 2021 (17.08.2021)

Name and mailing address of the ISA/	Authorized officer
Japan Patent Office 3-4-3, Kasumigaseki, Chiyoda-ku, Tokyo 100-8915, Japan	Telephone No.

INTERNATIONAL SEARCH REPORT

International application No.
PCT/JP2021/022425

C (Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	DING, Shaojin et al. AutoSpeech: Neural Architecture Search for Speaker Recognition. arXiv.org [online], arXiv:2005.03215v2. Cornell University., 2020, [retrieval date 05 August 2021], internet:<URL:https://arxiv.org/pdf/2005.03215v2> chapters 2.2, 3.1	2
Y	JP 2021-64342 A (BEIJING XIAOMI INTELLIGENT TECH CO LTD) 22 April 2021 (2021-04-22) paragraphs [0041]-[0048], fig. 6	4

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/JP2021/022425

Patent Documents referred in the Report	Publication Date	Patent Family	Publication Date
CN 112598117 A	02 Apr. 2021	(Family: none)	
JP 2021-64342 A	22 Apr. 2021	EP 3809333 A1 paragraphs [0042]-[0049], fig. 6 US 2021/110276 A1 CN 110689127 A KR 2021-0045274 A	

国際調査報告

国際出願番号
PCT/JP2021/022425

A. 発明の属する分野の分類（国際特許分類（ＩＰＣ））

G06N 3/08(2006.01)i
FI: G06N3/08

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））

G06N3/00-99/00; G06T7/00

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

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
Y	WU, Bichen et al., FBNet: Hardware-Aware Efficient ConvNet Design via Differentiable Neural Architecture Search, arXiv.org [online], arXiv:1812.03443v3, Cornell University, 2019, [検索日 2021.08.05], インターネット:<URL: https://arxiv.org/pdf/1812.03443v3>　第1章, 第3章-第3.2章	1-7
Y	CN 112598117 A (XAG CO., LTD.) 02.04.2021 (2021-04-02)　段落[0006], [0038]-[0055]	1-7
Y	JIANG, Weiwen et al., Hardware/Software Co-Exploration of Neural Architectures, arXiv.org [online], arXiv:1907.04650v2, Cornell University, 2020, pp. 1-10, [検索日 2021.08.05], インターネット:<URL:https://arxiv.org/pdf/1907.04650v2>　第II章A., 第III章, 図1, 3	1-7
Y	DING, Shaojin et al., AutoSpeech: Neural Architecture Search for Speaker Recognition, arXiv.org [online], arXiv:2005.03215v2, Cornell University, 2020, [検索日 2021.08.05], インターネット:<URL:https://arxiv.org/pdf/2005.03215v2>　第2.2章, 第3.1章	2

☑ 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。

* 引用文献のカテゴリー
"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
"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
"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）
"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
"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
"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
"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
"&" 同一パテントファミリー文献

国際調査を完了した日	国際調査報告の発送日
05.08.2021	17.08.2021

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）

多賀　実 5B 9367

電話番号 03-3581-1101 内線 3545

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
Y	JP 2021-64342 A (北京小米智能科技有限公司) 22.04.2021 (2021-04-22) 段落[0041]-[0048], 図6	4

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/JP2021/022425

引用文献	公表日	パテントファミリー文献	公表日
CN 112598117 A	02.04.2021	（ファミリーなし）	
JP 2021-64342 A	22.04.2021	EP 3809333 A1 段落[0042]-[0049]，図6 US 2021/110276 A1 CN 110689127 A KR 2021-0045274 A	