(19)日本国特許庁(JP)　(12)公開特許公報(A)　(11)公開番号
特開2023-44336
(P2023-44336A)
(43)公開日　令和5年3月30日(2023.3.30)

(51)国際特許分類　FI
G06N 20/00 (2019.01)　G06N 20/00　130
G06N 3/08 (2023.01)　G06N 3/08

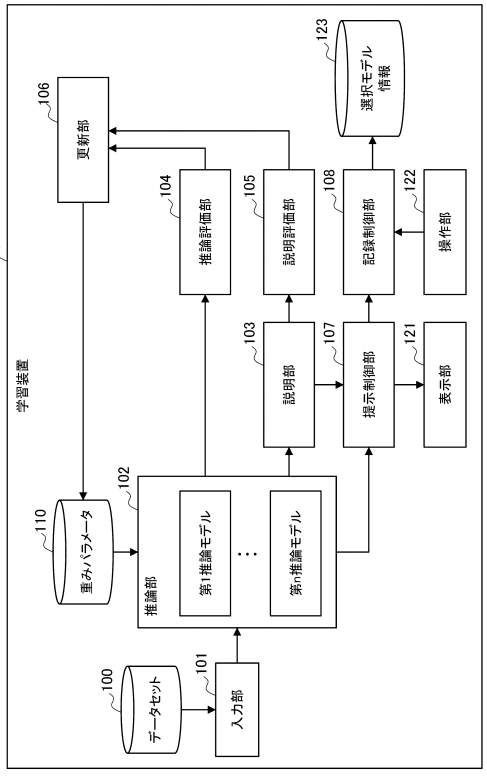
審査請求　未請求　請求項の数　14　OL　（全19頁）

(21)出願番号　特願2021-152315(P2021-152315)
(22)出願日　令和3年9月17日(2021.9.17)

(71)出願人　000000295
沖電気工業株式会社
東京都港区虎ノ門一丁目７番１２号
(74)代理人　100140958
弁理士　伊藤 学
(74)代理人　100137888
弁理士　大山 夏子
(74)代理人　100190942
弁理士　風間 竜司
(72)発明者　国定 恭史
東京都港区虎ノ門一丁目７番１２号 沖電気工業株式会社内

(54)【発明の名称】　学習装置、学習方法およびプログラム

(57)【要約】
【課題】人的コストを抑制しながら、解釈性および精度の高いモデルを得ることが可能な技術が提供されることが望まれる。
【解決手段】第１の入力データと前記第１の入力データの正解値とを取得する入力部と、前記第１の入力データと複数の推論モデルとに基づいて、複数の推論モデルそれぞれに対応する第１の推論値を出力する推論部と、前記第１の推論値に対する前記第１の入力データの寄与の大きさを示す前記複数の推論モデルそれぞれに対応する第１の説明情報を出力する説明部と、前記正解値と前記第１の推論値とに基づいて推論評価結果を得る推論評価部と、前記複数の推論モデルそれぞれに対応する第１の説明情報同士の一致度に基づいて説明評価結果を得る説明評価部と、前記推論評価結果と前記説明評価結果とに基づいて、前記複数の推論モデルの第１の重みパラメータの更新を行う更新部と、を備える、学習装置が提供される。
【選択図】図１



【特許請求の範囲】
【請求項１】
第１の入力データと前記第１の入力データの正解値とを取得する入力部と、
前記第１の入力データと複数の推論モデルとに基づいて、複数の推論モデルそれぞれに対応する第１の推論値を出力する推論部と、
前記第１の推論値に対する前記第１の入力データの寄与の大きさを示す前記複数の推論モデルそれぞれに対応する第１の説明情報を出力する説明部と、
前記正解値と前記第１の推論値とに基づいて推論評価結果を得る推論評価部と、
前記複数の推論モデルそれぞれに対応する第１の説明情報同士の一致度に基づいて説明評価結果を得る説明評価部と、
前記推論評価結果と前記説明評価結果とに基づいて、前記複数の推論モデルの第１の重みパラメータの更新を行う更新部と、
を備える、学習装置。
【請求項２】
前記入力部は、第２の入力データを取得し、
前記推論部は、前記第２の入力データと前記第１の重みパラメータの更新後の複数の推定モデルである複数の学習済みモデルとに基づいて、前記複数の学習済みモデルそれぞれに対応する第２の推論値を出力し、
前記説明部は、前記第２の推論値に対する前記第２の入力データの寄与の大きさを示す前記複数の学習済みモデルそれぞれに対応する第２の説明情報を出力し、
前記学習装置は、前記第２の説明情報のユーザへの提示を制御する提示制御部を備える、
請求項１に記載の学習装置。
【請求項３】
前記提示制御部は、前記第２の推論値および前記第２の説明情報の前記ユーザへの提示を制御する、
請求項２に記載の学習装置。
【請求項４】
前記学習装置は、
前記複数の学習済みモデルから前記ユーザによって選択された１または複数の学習済みモデルを示す情報の記録を制御する記録制御部を備える、
請求項２または３に記載の学習装置。
【請求項５】
前記説明評価結果は、前記複数の推論モデルそれぞれに対応する第１の説明情報同士の一致度が大きいほど小さい値を取る、
請求項１～４のいずれか一項に記載の学習装置。
【請求項６】
前記説明評価部は、前記複数の推論モデルそれぞれに対応する第１の説明情報を正規化したベクトルの内積に基づいて前記説明評価結果を得る、
請求項５に記載の学習装置。
【請求項７】
前記説明評価部は、前記複数の推論モデルごとに、前記第１の説明情報の二値化を行ってマスクを生成するとともに、自身以外の推論モデルに対応する前記第１の説明情報から生成したマスクと自身の推論モデルに対応する前記第１の説明情報との積を計算し、前記複数の推論モデルごとの前記積の和に基づいて、前記説明評価結果を得る、
請求項５に記載の学習装置。
【請求項８】
前記説明部は、誤差逆伝播が可能な関数を含む、
請求項１～７のいずれか一項に記載の学習装置。
【請求項９】

前記説明部は、第２の重みパラメータを有し、
前記更新部は、誤差逆伝播法によって前記第２の重みパラメータの更新を行う、
請求項８に記載の学習装置。
【請求項１０】
前記複数の推論モデルの少なくとも一つは、ニューラルネットワークを含む、
請求項１～９のいずれか一項に記載の学習装置。
【請求項１１】
前記更新部は、前記推論評価結果と前記説明評価結果との加算結果に基づいて、前記第１の重みパラメータの更新を行う、
請求項１～１０のいずれか一項に記載の学習装置。
【請求項１２】
前記第１の説明情報は、前記第１の推論値に対する前記第１の入力データの寄与の大きさを示すヒートマップである、
請求項１～１１のいずれか一項に記載の学習装置。
【請求項１３】
第１の入力データと前記第１の入力データの正解値とを取得することと、
前記第１の入力データと複数の推論モデルとに基づいて、複数の推論モデルそれぞれに対応する第１の推論値を出力することと、
前記第１の推論値に対する前記第１の入力データの寄与の大きさを示す前記複数の推論モデルそれぞれに対応する第１の説明情報を出力することと、
前記正解値と前記第１の推論値とに基づいて推論評価結果を得ることと、
前記複数の推論モデルそれぞれに対応する第１の説明情報同士の一致度に基づいて説明評価結果を得ることと、
前記推論評価結果と前記説明評価結果とに基づいて、前記複数の推論モデルの第１の重みパラメータの更新を行うことと、
を含む、学習方法。
【請求項１４】
コンピュータを、
第１の入力データと前記第１の入力データの正解値とを取得する入力部と、
前記第１の入力データと複数の推論モデルとに基づいて、複数の推論モデルそれぞれに対応する第１の推論値を出力する推論部と、
前記第１の推論値に対する前記第１の入力データの寄与の大きさを示す前記複数の推論モデルそれぞれに対応する第１の説明情報を出力する説明部と、
前記正解値と前記第１の推論値とに基づいて推論評価結果を得る推論評価部と、
前記複数の推論モデルそれぞれに対応する第１の説明情報同士の一致度に基づいて説明評価結果を得る説明評価部と、
前記推論評価結果と前記説明評価結果とに基づいて、前記複数の推論モデルの第１の重みパラメータの更新を行う更新部と、
を備える、学習装置として機能させるプログラム。
【発明の詳細な説明】
【技術分野】
【０００１】
本発明は、学習装置、学習方法およびプログラムに関する。
【背景技術】
【０００２】
ニューラルネットワーク（以下、「ＮＮ」とも表記する。）は、画像認識などにおいて高い性能を有する。しかし、一般的にＮＮは、膨大なパラメータと複雑なモデルとによって構成されており、ＮＮのパラメータとＮＮからの出力結果との関係を解釈することが難しい。かかる課題を解決するため、解釈性の高いＮＮを得る手法が幾つか提案されている。なお、「解釈性が高い」は、「人間の感覚との一致度が高い」とも換言され得る。

【０００３】
例えば、ＮＮのモデルが判断のために注目するべき領域を示したヒートマップのラベルを人手によって付しておき、そのヒートマップと一致するようにモデルを学習させることによって人にも解釈しやすいモデルを得る手法が知られている（例えば、非特許文献１参照）。また、モデルから得られたヒートマップの解釈性が低い場合には、そのヒートマップと一致しないようにモデルを再学習させることによって、より解釈性の高いモデルを得ることもできる。
【０００４】
また、入力データのうちＮＮが判断を行うための注目領域を抽出する機構をネットワーク内に導入することによって、ＮＮの精度を向上させる手法も知られている（例えば、非特許文献２参照）。かかる手法によって得られた注目領域を人間が修正し、修正した注目領域とＮＮの注目領域が一致するようにＮＮを再学習させることによって、ＮＮの解釈性および精度を向上させることができる。
【先行技術文献】
【非特許文献】
【０００５】
【非特許文献１】Andrew Ross、他2名、"Right for the Right Reasons: Training Differentiable Models byConstraining their Explanations"、[online]、［令和3年9月8日検索］、インターネット＜https://arxiv.org/abs/1703.03717＞
【非特許文献２】Masahiro Mitsuhara、他6名、"Embedding Human Knowledge into Deep Neural Network viaAttention Map"、[online]、［令和3年9月8日検索］、インターネット＜https://arxiv.org/abs/1905.03540＞
【非特許文献３】"Grad-CAM: VisualExplanations from Deep Networks via Gradient-based Localization"、[online]、［令和3年9月8日検索］、インターネット＜https://arxiv.org/abs/1610.02391v3＞
【発明の概要】
【発明が解決しようとする課題】
【０００６】
しかしながら、非特許文献１および非特許文献２に記載された、人手によってヒートマップのラベルを用意する手法は、ラベル付けのための人的コストが大きい。
【０００７】
一方、ラベル付けを必要としない手法としては、非特許文献１に記載の学習済みモデルのヒートマップと一致しないようにモデルを再学習させる手法が挙げられる。しかし、かかる手法では、再学習により精度が低下してしまう可能性が高いという点が課題として挙げられる。さらに、かかる手法では、全てのデータに対して一様にヒートマップの一致度が低下してしまうため、個々のデータに対してはかえってヒートマップの解釈性を低下させてしまう場合があるという点が課題として挙げられる。
【０００８】
そこで、人的コストを抑制しながら、解釈性および精度の高いモデルを得ることが可能な技術が提供されることが望まれる。
【課題を解決するための手段】
【０００９】
上記問題を解決するために、本発明のある観点によれば、第１の入力データと前記第１の入力データの正解値とを取得する入力部と、前記第１の入力データと複数の推論モデルとに基づいて、複数の推論モデルそれぞれに対応する第１の推論値を出力する推論部と、前記第１の推論値に対する前記第１の入力データの寄与の大きさを示す前記複数の推論モデルそれぞれに対応する第１の説明情報を出力する説明部と、前記正解値と前記第１の推論値とに基づいて推論評価結果を得る推論評価部と、前記複数の推論モデルそれぞれに対応する第１の説明情報同士の一致度に基づいて説明評価結果を得る説明評価部と、前記推論評価結果と前記説明評価結果とに基づいて、前記複数の推論モデルの第１の重みパラメ

ータの更新を行う更新部と、を備える、学習装置が提供される。
【００１０】
前記入力部は、第２の入力データを取得し、前記推論部は、前記第２の入力データと前記第１の重みパラメータの更新後の複数の推定モデルである複数の学習済みモデルとに基づいて、前記複数の学習済みモデルそれぞれに対応する第２の推論値を出力し、前記説明部は、前記第２の推論値に対する前記第２の入力データの寄与の大きさを示す前記複数の学習済みモデルそれぞれに対応する第２の説明情報を出力し、前記学習装置は、前記第２の説明情報のユーザへの提示を制御する提示制御部を備えてもよい。
【００１１】
前記提示制御部は、前記第２の推論値および前記第２の説明情報の前記ユーザへの提示を制御してもよい。
【００１２】
前記学習装置は、前記複数の学習済みモデルから前記ユーザによって選択された１または複数の学習済みモデルを示す情報の記録を制御する記録制御部を備えてもよい。
【００１３】
前記説明評価結果は、前記複数の推論モデルそれぞれに対応する第１の説明情報同士の一致度が大きいほど小さい値を取ってもよい。
【００１４】
前記説明評価部は、前記複数の推論モデルそれぞれに対応する第１の説明情報を正規化したベクトルの内積に基づいて前記説明評価結果を得てもよい。
【００１５】
前記説明評価部は、前記複数の推論モデルごとに、前記第１の説明情報の二値化を行ってマスクを生成するとともに、自身以外の推論モデルに対応する前記第１の説明情報から生成したマスクと自身の推論モデルに対応する前記第１の説明情報との積を計算し、前記複数の推論モデルごとの前記積の和に基づいて、前記説明評価結果を得てもよい。
【００１６】
前記説明部は、誤差逆伝播が可能な関数を含んでもよい。
【００１７】
前記説明部は、第２の重みパラメータを有し、前記更新部は、誤差逆伝播法によって前記第２の重みパラメータの更新を行ってもよい。
【００１８】
前記複数の推論モデルの少なくとも一つは、ニューラルネットワークを含んでもよい。なお、ニューラルネットワークは、機械学習アルゴリズムの一例に過ぎない。したがって、ニューラルネットワークの代わりに他の機械学習アルゴリズムが用いられてもよい。
【００１９】
前記更新部は、前記推論評価結果と前記説明評価結果との加算結果に基づいて、前記第１の重みパラメータの更新を行ってもよい。
【００２０】
前記第１の説明情報は、前記第１の推論値に対する前記第１の入力データの寄与の大きさを示すヒートマップであってもよい。
【００２１】
また、本発明の別の観点によれば、第１の入力データと前記第１の入力データの正解値とを取得することと、前記第１の入力データと複数の推論モデルとに基づいて、複数の推論モデルそれぞれに対応する第１の推論値を出力することと、前記第１の推論値に対する前記第１の入力データの寄与の大きさを示す前記複数の推論モデルそれぞれに対応する第１の説明情報を出力することと、前記正解値と前記第１の推論値とに基づいて推論評価結果を得ることと、前記複数の推論モデルそれぞれに対応する第１の説明情報同士の一致度に基づいて説明評価結果を得ることと、前記推論評価結果と前記説明評価結果とに基づいて、前記複数の推論モデルの第１の重みパラメータの更新を行うことと、を含む、学習方法が提供される。

【００２２】
また、本発明の別の観点によれば、コンピュータを、第１の入力データと前記第１の入力データの正解値とを取得する入力部と、前記第１の入力データと複数の推論モデルとに基づいて、複数の推論モデルそれぞれに対応する第１の推論値を出力する推論部と、前記第１の推論値に対する前記第１の入力データの寄与の大きさを示す前記複数の推論モデルそれぞれに対応する第１の説明情報を出力する説明部と、前記正解値と前記第１の推論値とに基づいて推論評価結果を得る推論評価部と、前記複数の推論モデルそれぞれに対応する第１の説明情報同士の一致度に基づいて説明評価結果を得る説明評価部と、前記推論評価結果と前記説明評価結果とに基づいて、前記複数の推論モデルの第１の重みパラメータの更新を行う更新部と、を備える、学習装置として機能させるプログラムが提供される。
【発明の効果】
【００２３】
以上説明したように本発明によれば、人的コストを抑制しながら、解釈性および精度の高いモデルを得ることが可能な技術が提供される。
【図面の簡単な説明】
【００２４】
【図１】本発明の実施形態に係る学習装置の機能構成例を示す図である。
【図２】ヒートマップを二値化したマスクと他のヒートマップとの掛け合わせによって説明評価結果を得る手法について説明するための図である。
【図３】同実施形態に係る学習装置の学習段階における動作例を示すフローチャートである。
【図４】同実施形態に係る学習装置のテスト段階における動作例を示すフローチャートである。
【図５】学習装置の例としての情報処理装置のハードウェア構成を示す図である。
【発明を実施するための形態】
【００２５】
以下に添付図面を参照しながら、本発明の好適な実施の形態について詳細に説明する。なお、本明細書及び図面において、実質的に同一の機能構成を有する構成要素については、同一の符号を付することにより重複説明を省略する。
【００２６】
また、本明細書および図面において、実質的に同一の機能構成を有する複数の構成要素を、同一の符号の後に異なる数字を付して区別する場合がある。ただし、実質的に同一の機能構成を有する複数の構成要素等の各々を特に区別する必要がない場合、同一符号のみを付する。また、異なる実施形態の類似する構成要素については、同一の符号の後に異なるアルファベットを付して区別する場合がある。ただし、異なる実施形態の類似する構成要素等の各々を特に区別する必要がない場合、同一符号のみを付する。
【００２７】
（０．実施形態の概要）
本発明の実施形態の概要について説明する。本発明の実施形態では、入力データ（学習用データ）と正解値との組み合わせに基づいてニューラルネットワークの学習を行う学習装置について説明する。しかし、ニューラルネットワークは、機械学習アルゴリズムの一例に過ぎない。したがって、ニューラルネットワークの代わりに他の機械学習アルゴリズムが用いられてもよい。例えば、機械学習アルゴリズムの他の一例として、ＳＶＭ（Ｓｕｐｐｏｒｔ　Ｖｅｃｔｏｒ　Ｍａｃｈｉｎｅ）などが用いられてもよい。
【００２８】
（１．実施形態の詳細）
本発明の実施形態について詳細に説明する。
【００２９】
（１．１．学習装置の構成例）
図１は、本発明の実施形態に係る学習装置１０の機能構成例を示す図である。図１に示

されるように、本発明の実施形態に係る学習装置１０は、入力部１０１と、推論部１０２と、説明部１０３と、推論評価部１０４と、説明評価部１０５と、更新部１０６と、提示制御部１０７と、記録制御部１０８と、表示部１２１と、操作部１２２とを備える。

【００３０】

本発明の実施形態では、推論部１０２が、ｎ個（ｎは１より大きい整数）の推論モデル、すなわち、「第１推論モデル」から「第ｎ推論モデル」までを含む場合を主に想定する。また、本発明の実施形態では、第１推論モデルから第ｎ推論モデルまでのそれぞれが、ニューラルネットワークを含んで構成される場合を主に想定する。以下では、ニューラルネットワークを「ＮＮ」とも表記する。

【００３１】

第１推論モデルから第ｎ推論モデルまでのそれぞれに含まれるＮＮは、重みパラメータ１１０（第１の重みパラメータ）を使用する。このとき、第１推論モデルから第ｎ推論モデルまでのそれぞれに含まれるＮＮは、共通の構造を有し、使用する重みパラメータ１１０（第１の重みパラメータ）が異なっていてもよい。あるいは、第１推論モデルから第ｎ推論モデルまでのそれぞれに含まれるＮＮは、別々の構造を有していてもよい。

【００３２】

なお、第１推論モデルから第ｎ推論モデルまでの少なくとも一つが、ＮＮを含んでもよい。例えば、第１推論モデルから第ｎ推論モデルまでの一部がＮＮを含んでもよく、第１推論モデルから第ｎ推論モデルまでの他の一部は、ＮＮの代わりに他の機械学習アルゴリズムを含んでもよい。

【００３３】

さらに、本発明の実施形態では、説明部１０３がＮＮを含んで構成される場合を主に想定する。説明部１０３に含まれるＮＮは、重みパラメータ（第２の重みパラメータ）を使用する。

【００３４】

データセット１００、第１推論モデルから第ｎ推論モデルまでの重みパラメータ１１０（第１の重みパラメータ）および説明部１０３が有する重みパラメータ（第２の重みパラメータ）は、図示しない記憶部によって記憶される。かかる記憶部は、ＲＡＭ（Ｒａｎｄｏｍ　Ａｃｃｅｓｓ　Ｍｅｍｏｒｙ）、ハードディスクドライブまたはフラッシュメモリなどのメモリによって構成されてよい。

【００３５】

入力部１０１と、推論部１０２と、説明部１０３と、推論評価部１０４と、説明評価部１０５と、更新部１０６と、提示制御部１０７と、記録制御部１０８とは、ＣＰＵ（Ｃｅｎｔｒａｌ　Ｐｒｏｃｅｓｓｉｎｇ　Ｕｎｉｔ）またはＧＰＵ（Ｇｒａｐｈｉｃｓ　Ｐｒｏｃｅｓｓｉｎｇ　Ｕｎｉｔ）などの演算装置を含み、ＲＯＭ（Ｒｅａｄ　Ｏｎｌｙ　Ｍｅｍｏｒｙ）により記憶されているプログラムが演算装置によりＲＡＭに展開されて実行されることにより、その機能が実現され得る。このとき、当該プログラムを記録した、コンピュータに読み取り可能な記録媒体も提供され得る。あるいは、これらのブロックは、専用のハードウェアにより構成されていてもよいし、複数のハードウェアの組み合わせにより構成されてもよい。演算装置による演算に必要なデータは、図示しない記憶部によって適宜記憶される。

【００３６】

初期状態において、第１推論モデルから第ｎ推論モデルまでの重みパラメータ１１０および説明部１０３が有する重みパラメータそれぞれには、初期値が設定されている。例えば、これらに設定される初期値は、ランダムな値であってよいが、どのような値であってもよい。例えば、これらに設定される初期値は、あらかじめ学習によって得られた学習済みの値であってもよい。

【００３７】

（データセット１００）

データセット１００は、学習段階において使用される複数の入力データ（第１の入力デ

ータ）と当該複数の入力データそれぞれの正解値とを含む。学習段階において使用される複数の入力データは、学習用データに該当し得る。さらに、データセット１００は、テスト段階において使用される複数の入力データ（第２の入力データ）を含む。テスト段階において使用される複数の入力データは、テスト用データに該当し得る。

【００３８】

なお、テスト用データは、学習用データと別のデータとして用意されていることが主に想定される。しかし、テスト用データは、学習用データの一部を含んでもよい。

【００３９】

また、本発明の実施形態では、入力データが画像データである場合（特に、静止画像データである場合）を主に想定する。しかし、入力データの種類は特に限定されず、画像データ以外も入力データとして用いられ得る。例えば、入力データは、複数のフレームを含んだ動画像データであってもよいし、音響データであってもよい。

【００４０】

（入力部１０１）

入力部１０１は、学習段階において、データセット１００から学習段階において使用される入力データおよび正解値の組み合わせを順次に取得する。入力部１０１は、学習段階において使用される入力データおよび正解値の組み合わせを順次に推論部１０２に出力する。また、入力部１０１は、テスト段階において、データセット１００からテストにおいて使用される入力データを順次に取得する。入力部１０１は、テスト段階において使用される入力データを順次に推論部１０２に出力する。

【００４１】

なお、例えば、入力部１０１は、データセット１００から学習段階において使用される入力データおよび正解値の組み合わせを全部取得して出力し終わった場合には、最初から当該組み合わせを取得し直して再度出力する動作を所定の回数繰り返してよい。かかる場合には、入力部１０１よりも後段のブロックにおいても、再度の入力に基づいて順次に各自の処理が繰り返し実行されてよい。一方、例えば、入力部１０１は、データセット１００からテスト段階において使用される入力データを全部取得して出力し終わった場合には、入力データの取得を終了してよい。

【００４２】

（推論部１０２）

推論部１０２は、学習段階において、入力部１０１から入力された入力データと第１推論モデルから第ｎ推論モデルまでとに基づいて、第１推論モデルから第ｎ推論モデルまでのそれぞれに対応する推論値（第１の推論値）を得る。同様に、推論部１０２は、テスト段階において、入力部１０１から入力された入力データと第１推論モデルから第ｎ推論モデルまでとに基づいて、第１推論モデルから第ｎ推論モデルまでのそれぞれに対応する推論値（第２の推論値）を得る。

【００４３】

第１推論モデルから第ｎ推論モデルまでが使用する重みパラメータ１１０は図示しない記憶部によって記憶されている。したがって、推論部１０２は、図示しない記憶部から重みパラメータ１１０を取得し、取得した重みパラメータ１１０と入力部１０１から入力された入力データとに基づいて、第１推論モデルから第ｎ推論モデルまでによる推論を行う。

【００４４】

なお、本明細書においては、ＮＮへの入力に基づいてＮＮからの出力を得ることを広く「推論」と言う。

【００４５】

一例として、ｉ番目の推論モデルを示す関数をＦｉ（ｉは１～ｎまでの整数）とし、ｉ番目の推論モデルへの入力をｘとすると、ｉ番目の推論モデルからの出力はＦｉ（ｘ）と表現され得る。

【００４６】

なお、後にも説明するように、説明部１０３が用いる説明手法（すなわち、説明情報の生成手法）には、推論値の他に第１推論モデルから第ｎ推論モデルまでのそれぞれから出力される特徴量（中間特徴量）などの情報を必要とする説明手法が存在する場合があり得る。かかる場合には、推論部１０２は、推論値とともに、第１推論モデルから第ｎ推論モデルまでのそれぞれの中間層から出力される特徴量を説明部１０３に出力してよい。

【００４７】

第１推論モデルから第ｎ推論モデルまでの具体的な構成は、特に限定されない。しかし、第１推論モデルから第ｎ推論モデルまでのそれぞれの出力の形式は、入力データに対応する正解値の形式と合わせて設定されているのがよい。例えば、正解値が分類問題のクラスである場合、第１推論モデルから第ｎ推論モデルまでのそれぞれの出力は、クラス数分の長さを有するｏｎｅ－ｈｏｔベクトルであるとよい。

【００４８】

推論部１０２は、学習段階において、第１推論モデルから第ｎ推論モデルまでのそれぞれに対応する推論値を、説明部１０３および推論評価部１０４それぞれに出力する。一方、推論部１０２は、テスト段階において、第１推論モデルから第ｎ推論モデルまでのそれぞれに対応する推論値を、説明部１０３および提示制御部１０７それぞれに出力する。

【００４９】

（説明部１０３）

説明部１０３は、第１推論モデルから第ｎ推論モデルまでのそれぞれについて、推論部１０２から入力された推論値の判断根拠を説明する説明情報を生成する。

【００５０】

ここで、説明情報は、推論部１０２から入力された推論値に対する入力データの寄与の大きさを示す情報である。以下では、説明情報が推論値に対する入力データの寄与の大きさを領域（例えば、画像を構成するピクセルなど）または変数ごとに示すヒートマップである場合について主に説明する。ヒートマップによれば、入力データのうち判断に寄与した重要な領域または変数が示され得る。

【００５１】

入力データが画像データなどである場合には、ヒートマップは２次元ベクトルによって表現され得る。あるいは、入力データが表形式データなどである場合には、ヒートマップは１次元ベクトルによって表現され得る。

【００５２】

ヒートマップはどのように生成されてもよい。例えば、説明部１０３は、推論部１０２から入力された推論値に基づいて、ヒートマップを生成してもよい。あるいは、上記したように、推論部１０２から説明部１０３に推論値だけではなく特徴量も入力される場合があり得る。かかる場合には、説明部１０３は、推論部１０２から入力された推論値と特徴量とに基づいて、ヒートマップを生成してもよい。

【００５３】

例えば、説明部１０３は、誤差逆伝播が可能な関数を含んでいてもよい。このとき、後に説明するように、更新部１０６によって説明部１０３が有する重みパラメータが誤差逆伝播法によって更新され得る。すなわち、説明部１０３は、誤差逆伝播法による更新後の重みパラメータによってヒートマップを生成してもよい。

【００５４】

誤差逆伝播法による更新後の重みパラメータによってヒートマップを生成する説明手法としては、非特許文献３に記載された、いわゆるＧｒａｄ－ＣＡＭなどが適用され得る。Ｇｒａｄ－ＣＡＭは、ＮＮへの入力のうち推論値への寄与度が高い領域を示すヒートマップを出力する説明手法である。その他にも、Ｖａｎｉｌｌａ　Ｇｒａｄｉｅｎｔ、ＳｍｏｏｔｈＧｒａｄといった各種の説明手法が適用され得る。

【００５５】

上記したように、ｉ番目の推論モデルに対応する推論値はＦｉ（ｘ）と表現され得るため、一例として、ヒートマップの生成処理を示す関数をＧとすると、説明部１０３によっ

て生成されるｉ番目の推論モデルに対応するヒートマップＴｉ（ｘ）は、以下の式（１）のように表現され得る。

【００５６】

Ｔｉ（ｘ）＝Ｇ（Ｆｉ（ｘ））・・・（１）

【００５７】

説明部１０３は、学習段階において、生成したｎ個のヒートマップ（第１の説明情報）を説明評価部１０５に出力する。一方、説明部１０３は、テスト段階において、生成したｎ個のヒートマップ（第２の説明情報）を提示制御部１０７に出力する。

【００５８】

（推論評価部１０４）

推論評価部１０４は、推論部１０２から入力された第１推論モデルから第ｎ推論モデルまでのそれぞれに対応する推論値と入力部１０１によって取得された正解値とに基づいて、推論評価結果を得る。より詳細に、推論評価部１０４は、第１推論モデルから第ｎ推論モデルまでのそれぞれに対応する推論値と入力部１０１によって取得された正解値とを比較することによって、推論評価結果を得る。

【００５９】

本発明の実施形態では、推論評価部１０４が、推論値と正解値とに応じた損失関数の第１推論モデルから第ｎ推論モデルまでについての和を推論評価結果の例としての損失関数Ｌ１として算出する場合を想定する。ここで、推論値と正解値とに応じた損失関数は特定の関数に限定されず、一般的なニューラルネットワークにおいて用いられる損失関数と同様の損失関数が用いられてよい。例えば、推論値と正解値とに応じた損失関数は、正解値と推論値との差分に基づくクロスエントロピー誤差であってもよい。

【００６０】

推論評価部１０４は、推論評価結果を更新部１０６に出力する。

【００６１】

（説明評価部１０５）

説明評価部１０５は、説明部１０３から入力された第１推論モデルから第ｎ推論モデルまでのそれぞれに対応するヒートマップに基づいて説明評価結果を得る。より詳細に、説明評価部１０５は、第１推論モデルから第ｎ推論モデルまでのそれぞれに対応するヒートマップ同士を比較する。そして、説明評価部１０５は、比較結果としてのｎ個のヒートマップ同士の一致度に基づいて、説明評価結果を得る。

【００６２】

本発明の実施形態では、ｎ個のヒートマップ同士の一致度が大きいほど説明評価結果が小さい値を取る損失関数である場合を主に想定する。なお、ｎ個のヒートマップ同士の一致度は、ｎ個のヒートマップ同士がどの程度乖離しているかを示す乖離度と換言されてもよい。かかる場合には、ｎ個のヒートマップ同士の乖離度が小さいほど説明評価結果が小さい値を取る損失関数であってよい。

【００６３】

ｎ個のヒートマップから説明評価結果を得る手法は限定されない。ここでは、説明評価結果を得る手法として、ヒートマップを二値化したマスクと他のヒートマップとの掛け合わせによって説明評価結果を得る手法、および、正規化されたヒートマップ同士の内積によって説明評価結果を得る手法について順に説明する。

【００６４】

図２は、ヒートマップを二値化したマスクと他のヒートマップとの掛け合わせによって説明評価結果を得る手法について説明するための図である。図２に示された例では、説明を簡便にするため、ｎ＝２である場合、すなわち、推論部１０２が、第１推論モデルおよび第２推論モデルを有する場合を想定する。

【００６５】

図２を参照すると、第１推論モデルからは、推論値とヒートマップＨ１とが出力されている。一方、第２推論モデルからは、推論値とヒートマップＨ２とが出力されている。図

２では、ヒートマップＨ１およびヒートマップＨ２において、入力データのうち推論値への寄与が大きい領域ほど濃い色によって示されている。

【００６６】

説明評価部１０５は、ヒートマップＨ１の二値化を行ってマスクＭ１を生成するとともに、ヒートマップＨ２の二値化を行ってマスクＭ２を生成する。なお、二値化は、閾値ｃ以上である要素（例えば、ヒートマップを構成するピクセル）の値を１とし、閾値ｃよりも小さい要素の値を０とすることによって実行され得る。図２においては、二値のうち１が黒によって示され、０が白によって示されている。

【００６７】

説明評価部１０５は、第１推論モデルから出力されたヒートマップＨ１と、第２の推論モデルから出力されたヒートマップＨ２から生成したマスクＭ２との積を、要素ごとに計算する。同様に、説明評価部１０５は、第２推論モデルから出力されたヒートマップＨ２と、第１の推論モデルから出力されたヒートマップＨ１から生成したマスクＭ１との積を、要素ごとに計算する。これによって、各要素に対応する積の集合が推論モデルごとに得られる。

【００６８】

説明評価部１０５は、各要素に対応する積を全部の推論モデルについて足し合わせることによって積の和を計算する。そして、説明評価部１０５は、このようにして計算した積の和を全要素について足し合わせることによって合計値を計算する。説明評価部１０５は、この合計値を説明評価結果の例としての損失関数Ｌ２とする。

【００６９】

図２を参照しながらｎ＝２である場合について説明した。ｎを１より大きい任意の整数であるとして説明すると、以下の通りである。

【００７０】

すなわち、説明評価部１０５は、ｉ＝１～ｎについて、ヒートマップＴｉ（ｘ）の各要素の値を二値化したマスクＭｉ（ｘ）を生成する。次に、説明評価部１０５は、推論モデルごとに、自身の推論モデルから出力されたヒートマップＴｉ（ｘ）と、自身以外の推論モデルに対応するヒートマップから生成したマスクＭ１（ｘ）～Ｍｉ－１（ｘ）、Ｍｉ＋１（ｘ）～Ｍｎ（ｘ）の和との積を要素ごとに計算する。

【００７１】

説明評価部１０５は、各要素に対応する積を第１推論モデルから第ｎ推論モデルまでについて足し合わせることによって積の和を計算する。そして、説明評価部１０５は、このようにして計算した積の和に基づいて、説明評価結果を得る。より詳細に、説明評価部１０５は、積の和を全要素について足し合わせることによって合計値を計算する。説明評価部１０５は、この合計値を説明評価結果の例としての損失関数Ｌ２とする。

【００７２】

この損失関数Ｌ２は、各ヒートマップにおいて、自身以外のヒートマップにおいて閾値以上の値を持つ領域の合計値である。この損失関数Ｌ２の値を小さくするように学習が行われることによって、ヒートマップの一致度が小さいｎ個の推論モデルが得られる。なお、このときの損失関数Ｌ２は、以下の式（２）のように表現され得る。式（２）において、ｅは、要素番号を示す。ここで、ヒートマップＴｉ（ｘ）は、ヒートマップＴｉ（ｘ）の大きさ｜Ｔｉ（ｘ）｜で割るなどして正規化してもよい。また、ヒートマップＴｉ（ｘ）にはsigmoidなどの活性化関数をかけてもよい。

【００７３】

【数１】

$$L2(x)=\sum_{e}\left(\sum_{i=1}^{n}\left(Ti(x)\cdot\sum_{j=1,j\neq i}^{n}Mj(x)\right)\right) \qquad \cdots(2)$$

【００７４】
図２を参照しながら、ヒートマップを二値化したマスクと他のヒートマップとの掛け合わせによって説明評価結果を得る手法について説明した。続いて、正規化されたヒートマップ同士の内積によって説明評価結果を得る手法について説明する。
【００７５】
説明評価部１０５は、ｉ＝１～ｎについて、ヒートマップＴｉ（ｘ）をヒートマップＴｉ（ｘ）の大きさ｜Ｔｉ（ｘ）｜で割ることによって正規化して、ｉ＝１～ｎについての正規化したベクトルを生成する。そして、説明評価部１０５は、ｉ＝１～ｎについての正規化したベクトルの内積に基づいて説明評価結果を得る。より詳細に、説明評価部１０５は、内積を全要素について足し合わせることによって合計値を計算する。説明評価部１０５は、この合計値を説明評価結果の例としての損失関数Ｌ２とする。
【００７６】
正規化したベクトルの内積が大きいほど、この損失関数Ｌ２は、大きい値となる。正規化したベクトルの内積が大きいことは、ヒートマップ同士の一致度が高いことを意味する。したがって、この損失関数Ｌ２の値を小さくするように学習が行われることによって、ヒートマップの一致度が小さいｎ個の推論モデルが得られる。なお、このときの損失関数Ｌ２は、以下の式（３）のように表現され得る。式（３）において、ｅは、要素番号を示す。
【００７７】
【数２】

$$L2(x) = \prod_{i=1}^{n} \frac{Ti(x)}{|Ti(x)|} \qquad \cdots (3)$$

【００７８】
説明評価部１０５は、説明評価結果を更新部１０６に出力する。
【００７９】
（更新部１０６）
更新部１０６は、推論評価部１０４から入力された推論評価結果と、説明評価部１０５から入力された説明評価結果とに基づいて、第１推論モデルから第ｎ推論モデルまでのそれぞれが使用する重みパラメータ１１０の更新を行う。これによって、第１推論モデルから第ｎ推論モデルまでのそれぞれから出力される推論値が正解値に近づくように、かつ、説明部１０３から出力されるｎ個のヒートマップ同士の一致度が小さくなるように、重みパラメータ１１０が更新され得る。重みパラメータ１１０は、誤差逆伝播法（バックプロパゲーション）によって更新されてよい。
【００８０】
例えば、更新部１０６は、推論評価部１０４から入力された推論評価結果と、説明評価部１０５から入力された説明評価結果とを加算し、加算結果に基づいて、重みパラメータ１１０の更新を行えばよい。このとき、更新部１０６は、計算した加算結果を誤差として、誤差逆伝播法（バックプロパゲーション）によって重みパラメータ１１０を更新すればよい。上記のように、推論評価結果が損失関数Ｌ１と表現され、説明評価結果が損失関数Ｌ２と表現される場合、加算結果は、Ｌ１＋Ｌ２である。
【００８１】
さらに、更新部１０６は、説明部１０３が有する重みパラメータを更新してよい。より詳細に、説明部１０３が、誤差逆伝播が可能な関数を含む場合、更新部１０６は、推論評価結果と説明評価結果とに基づいて、誤差逆伝播法（バックプロパゲーション）によって、説明部１０３が有する重みパラメータを更新してよい。
【００８２】

なお、学習の終了条件（すなわち、重みパラメータ更新の終了条件）は特に限定されず、第１推論モデルから第ｎ推論モデルまでの学習がある程度行われたことを示す条件であればよい。具体的に、学習の終了件は、損失関数Ｌ１＋Ｌ２の値が閾値よりも小さいという条件を含んでもよい。あるいは、学習の終了条件は、損失関数Ｌ１＋Ｌ２の値の変化が閾値よりも小さいという条件（損失関数Ｌ１＋Ｌ２の値が収束状態になったという条件）を含んでもよい。あるいは、学習の終了条件は、重みパラメータの更新が所定の回数行われたという条件を含んでもよい。あるいは、推論評価部１０４によって正解値と推論値とに基づいて精度（例えば、正答率など）が算出される場合、学習の終了条件は、精度が所定の割合（例えば、９０％など）を超えるという条件を含んでもよい。

【００８３】

（提示制御部１０７）

提示制御部１０７は、テスト段階において、推論部１０２から入力された第１推論モデルから第ｎ推論モデルまでのそれぞれに対応する推論値と、説明部１０３から入力されたｎ個のヒートマップとが、ユーザに提示されるように制御する。より詳細に、提示制御部１０７は、第１推論モデルから第ｎ推論モデルまでのそれぞれに対応する推論値と、ｎ個のヒートマップとが表示されるように表示部１２１を制御する。なお、ｎ個のヒートマップは表示されるが、第１推論モデルから第ｎ推論モデルまでのそれぞれに対応する推論値は表示されない形態も想定され得る。

【００８４】

（表示部１２１）

表示部１２１は、ディスプレイによって構成され、提示制御部１０７による制御に従って各種情報の表示を行う機能を有する。例えば、表示部１２１は、ｎ個の推論値とｎ個のヒートマップとを表示することが可能である。ここで、表示部１２１の形態は特に限定されない。例えば、表示部１２１は、液晶ディスプレイ（ＬＣＤ）装置であってもよいし、ＯＬＥＤ（Ｏｒｇａｎｉｃ　Ｌｉｇｈｔ　Ｅｍｉｔｔｉｎｇ　Ｄｉｏｄｅ）装置であってもよいし、ランプなどの表示装置であってもよい。

【００８５】

（操作部１２２）

操作部１２２は、ユーザによる操作を受け付ける。例えば、ユーザがｎ個の推論値とｎ個のヒートマップとを参照しながら、ｎ個の推論モデルから解釈性の高い１または複数の推論モデル（以下、「選択モデル」とも言う。）を見つけたとする。このとき、ユーザは、選択モデルを示す情報（以下、「選択モデル情報」とも言う。）を操作部１２２に入力し、操作部１２２は、選択モデル情報１２３を受け付ける。例えば、選択モデル情報１２３は、選択モデルを示す番号であってよい。

【００８６】

なお、本発明の実施形態では、操作部１２２がマウスおよびキーボードである場合を主に想定する。しかし、操作部１２２の形態は特に限定されない。例えば、操作部１２２は、タッチパネルであってもよいし、他の入力装置であってもよい。

【００８７】

（記録制御部１０８）

記録制御部１０８は、操作部１２２によってユーザから受け付けられた選択モデル情報１２３の記録を制御する。より詳細に、記録制御部１０８は、操作部１２２によってユーザから受け付けられた選択モデル情報１２３を図示しない記憶部に記憶させる。選択モデル情報１２３は、図示しない記憶部から後に取得され、選択モデル情報１２３によって示される選択モデルが、解釈性の高い学習済みモデルとして用いられ得る。

【００８８】

なお、テストの終了条件は特に限定されず、ユーザにとって十分な回数のテストが行われたことを示す条件であればよい。具体的に、テストの終了条件は、テスト段階においてユーザによって推論結果の確認が所定の回数以上行われたという条件を含んでもよい。

【００８９】

以上、本発明の実施形態に係る学習装置１０の構成例について説明した。

【００９０】

（１．２．学習段階における動作）

図３を参照しながら、本発明の実施形態に係る学習装置１０の学習段階における動作の流れについて説明する。図３は、本発明の実施形態に係る学習装置１０の学習段階における動作例を示すフローチャートである。

【００９１】

まず、図３に示されたように、入力部１０１は、データセット１００から入力データ（すなわち、学習用データ）および正解値の組み合わせを取得する。さらに、推論部１０２は、ｎ個の推論モデルそれぞれに対応する重みパラメータ１１０を取得する（Ｓ１１）。推論部１０２は、入力部１０１によって取得された入力データとｎ個の推論モデルとに基づいて推論を行い（Ｓ１２）、推論によって得られたｎ個の推論値を推論評価部１０４および説明部１０３それぞれに出力する。

【００９２】

説明部１０３は、推論部１０２から入力されたｎ個の推論値に基づいて、ｎ個の推論値それぞれの判断根拠を説明するヒートマップを生成する（Ｓ１３）。説明部１０３は、生成したｎ個のヒートマップを説明評価部１０５に出力する。

【００９３】

推論評価部１０４は、入力部１０１によって取得された正解値に基づいて、推論部１０２から入力されたｎ個の推論値を評価して推論評価結果を得る。より詳細に、推論評価部１０４は、正解値とｎ個の推論値とに応じた損失関数を推論評価結果として算出する。推論評価部１０４は、算出した推論評価結果を更新部１０６に出力する。

【００９４】

説明評価部１０５は、説明部１０３から入力されたｎ個のヒートマップの一致度に基づいて、説明評価結果を得る。より詳細に、説明評価部１０５は、説明部１０３から入力されたｎ個のヒートマップ同士の一致度に応じた損失関数を説明評価結果として算出する。説明評価部１０５は、算出した説明評価結果を更新部１０６に出力する（Ｓ１４）。

【００９５】

更新部１０６は、推論評価部１０４から入力された推論評価結果と、説明評価部１０５から入力された説明評価結果とに基づいて、第１推論モデルから第ｎ推論モデルまでのそれぞれに対応する重みパラメータ１１０の更新を行う（Ｓ１５）。より詳細に、更新部１０６は、推論評価結果と説明評価結果とに基づいて、誤差逆伝播法によって、重みパラメータ１１０を更新する。さらに、更新部１０６は、推論評価結果と説明評価結果とに基づく誤差逆伝播法によって説明部１０３が有する重みパラメータの更新を行う。

【００９６】

更新部１０６は、入力データに基づく重みパラメータの更新が終わるたびに、学習の終了条件が満たされたか否かを判断する（Ｓ１６）。学習の終了条件が満たされていないと判断した場合には（Ｓ１６において「ＮＯ」）、Ｓ１１に動作が移行され、入力部１０１によって次の入力データが取得され、推論部１０２、説明部１０３、推論評価部１０４、説明評価部１０５および更新部１０６それぞれによって、当該次の入力データに基づく各自の処理が再度実行される。一方、更新部１０６によって、学習の終了条件が満たされたと判断された場合には（Ｓ１６において「ＹＥＳ」）、学習が終了される。

【００９７】

以上、本発明の実施形態に係る学習装置１０の学習段階における動作の流れについて説明した。

【００９８】

（１．３．テスト段階における動作）

図４を参照しながら、本発明の実施形態に係る学習装置１０のテスト段階における動作の流れについて説明する。図４は、本発明の実施形態に係る学習装置１０のテスト段階における動作例を示すフローチャートである。

【００９９】
まず、図４に示されたように、入力部１０１は、データセット１００から入力データ（すなわち、テスト用データ）および正解値の組み合わせを取得する。さらに、推論部１０２は、ｎ個の推論モデルそれぞれに対応する重みパラメータ１１０を取得する（Ｓ２１）。推論部１０２は、入力部１０１によって取得された入力データとｎ個の推論モデルとに基づいて推論を行い（Ｓ２２）、推論によって得られたｎ個の推論値を説明部１０３および提示制御部１０７それぞれに出力する。
【０１００】
説明部１０３は、推論部１０２から入力されたｎ個の推論値に基づいて、ｎ個の推論値それぞれの判断根拠を説明するヒートマップを生成する（Ｓ２３）。説明部１０３は、生成したｎ個のヒートマップを提示制御部１０７に出力する。
【０１０１】
提示制御部１０７は、推論部１０２から入力されたｎ個の推論値と、説明部１０３から入力されたｎ個のヒートマップとがユーザに提示されるように表示部１２１を制御する。表示部１２１は、提示制御部１０７による制御に従って、ｎ個の推論値と、ｎ個のヒートマップとを表示する（Ｓ２４）。
【０１０２】
操作部１２２は、ｎ個の推論モデルから解釈性が高いと判断された１または複数の推論モデルを示す情報（選択モデル情報１２３）をユーザから受け付ける。記録制御部１０８は、操作部１２２によってユーザから受け付けられた選択モデル情報１２３の記録を制御する（Ｓ２５）。図示しない記憶部は、記録制御部１０８による制御に従って、選択モデル情報１２３を記憶する。
【０１０３】
記録制御部１０８は、入力データに基づく選択モデル情報１２３の記録制御が終わるたびに、テストの終了条件が満たされたか否かを判断する（Ｓ２６）。テストの終了条件が満たされていないと判断した場合には（Ｓ２６において「ＮＯ」）、Ｓ２１に動作が移行され、入力部１０１によって次の入力データが取得され、推論部１０２、説明部１０３、提示制御部１０７および記録制御部１０８それぞれによって、当該次の入力データに基づく各自の処理が再度実行される。一方、記録制御部１０８によって、テストの終了条件が満たされたと判断された場合には（Ｓ２６において「ＹＥＳ」）、テストが終了される。
【０１０４】
以上、本発明の実施形態に係る学習装置１０のテスト段階における動作の流れについて説明した。
【０１０５】
（１．４．実施形態の効果）
以上に説明したように、本発明の実施形態によれば、第１推論モデルから第ｎ推論モデルまでのそれぞれから出力される推論値が正解値に近づくように、かつ、説明情報として出力されるｎ個のヒートマップ同士の一致度が小さくなるように、学習が行われ得る。これによって、互いに異なる複数のヒートマップを出力する推論モデルを得ることができる。これによって、ユーザは、ｎ個のモデルの中からより解釈性の高いヒートマップを出力するモデルを選んで使用することができる。
【０１０６】
以上、本発明の実施形態が奏する効果について説明した。
【０１０７】
（２．ハードウェア構成例）
続いて、本発明の実施形態に係る学習装置１０のハードウェア構成例について説明する。以下では、本発明の実施形態に係る学習装置１０のハードウェア構成例として、情報処理装置９００のハードウェア構成例について説明する。なお、以下に説明する情報処理装置９００のハードウェア構成例は、学習装置１０のハードウェア構成の一例に過ぎない。したがって、学習装置１０のハードウェア構成は、以下に説明する情報処理装置９００の

ハードウェア構成から不要な構成が削除されてもよいし、新たな構成が追加されてもよい。

【０１０８】

図５は、本発明の実施形態に係る学習装置１０の例としての情報処理装置９００のハードウェア構成を示す図である。情報処理装置９００は、ＣＰＵ（Ｃｅｎｔｒａｌ Ｐｒｏｃｅｓｓｉｎｇ Ｕｎｉｔ）９０１と、ＲＯＭ（Ｒｅａｄ Ｏｎｌｙ Ｍｅｍｏｒｙ）９０２と、ＲＡＭ（Ｒａｎｄｏｍ Ａｃｃｅｓｓ Ｍｅｍｏｒｙ）９０３と、ホストバス９０４と、ブリッジ９０５と、外部バス９０６と、インタフェース９０７と、入力装置９０８と、出力装置９０９と、ストレージ装置９１０と、通信装置９１１と、を備える。

【０１０９】

ＣＰＵ９０１は、演算処理装置および制御装置として機能し、各種プログラムに従って情報処理装置９００内の動作全般を制御する。また、ＣＰＵ９０１は、マイクロプロセッサであってもよい。ＲＯＭ９０２は、ＣＰＵ９０１が使用するプログラムや演算パラメータ等を記憶する。ＲＡＭ９０３は、ＣＰＵ９０１の実行において使用するプログラムや、その実行において適宜変化するパラメータ等を一時記憶する。これらはＣＰＵバス等から構成されるホストバス９０４により相互に接続されている。

【０１１０】

ホストバス９０４は、ブリッジ９０５を介して、ＰＣＩ（Ｐｅｒｉｐｈｅｒａｌ Ｃｏｍｐｏｎｅｎｔ Ｉｎｔｅｒｃｏｎｎｅｃｔ／Ｉｎｔｅｒｆａｃｅ）バス等の外部バス９０６に接続されている。なお、必ずしもホストバス９０４、ブリッジ９０５および外部バス９０６を分離構成する必要はなく、１つのバスにこれらの機能を実装してもよい。

【０１１１】

入力装置９０８は、マウス、キーボード、タッチパネル、ボタン、マイクロフォン、スイッチおよびレバー等ユーザが情報を入力するための入力手段と、ユーザによる入力に基づいて入力信号を生成し、ＣＰＵ９０１に出力する入力制御回路等から構成されている。情報処理装置９００を操作するユーザは、この入力装置９０８を操作することにより、情報処理装置９００に対して各種のデータを入力したり処理動作を指示したりすることができる。

【０１１２】

出力装置９０９は、例えば、ＣＲＴ（Ｃａｔｈｏｄｅ Ｒａｙ Ｔｕｂｅ）ディスプレイ装置、液晶ディスプレイ（ＬＣＤ）装置、ＯＬＥＤ（Ｏｒｇａｎｉｃ Ｌｉｇｈｔ Ｅｍｉｔｔｉｎｇ Ｄｉｏｄｅ）装置、ランプ等の表示装置およびスピーカ等の音声出力装置を含む。

【０１１３】

ストレージ装置９１０は、データ格納用の装置である。ストレージ装置９１０は、記憶媒体、記憶媒体にデータを記録する記録装置、記憶媒体からデータを読み出す読出し装置および記憶媒体に記録されたデータを削除する削除装置等を含んでもよい。ストレージ装置９１０は、例えば、ＨＤＤ（Ｈａｒｄ Ｄｉｓｋ Ｄｒｉｖｅ）で構成される。このストレージ装置９１０は、ハードディスクを駆動し、ＣＰＵ９０１が実行するプログラムや各種データを格納する。

【０１１４】

通信装置９１１は、例えば、ネットワークに接続するための通信デバイス等で構成された通信インタフェースである。また、通信装置９１１は、無線通信または有線通信のどちらに対応してもよい。

【０１１５】

以上、本発明の実施形態に係る学習装置１０のハードウェア構成例について説明した。

【０１１６】

（３．まとめ）

以上、添付図面を参照しながら本発明の好適な実施形態について詳細に説明したが、本発明はかかる例に限定されない。本発明の属する技術の分野における通常の知識を有する

者であれば、特許請求の範囲に記載された技術的思想の範疇内において、各種の変更例または修正例に想到し得ることは明らかであり、これらについても、当然に本発明の技術的範囲に属するものと了解される。

【０１１７】

例えば、上記した例では、学習装置１０がｎ個の推論モデルを同時に学習する場合を主に想定している。しかし、学習装置１０は、ｎ個の推論モデルの全部を同時に学習しなくてもよい。例えば、ｎ個の推論モデルの一部として、学習済みの推論モデルが使用されてもよい。このとき、学習済みの推論モデルの重みパラメータは、更新されずに一定の値に固定され得る。

【０１１８】

また、上記した例では、説明部１０３におけるヒートマップの生成手法の種類が、１種類である場合を主に想定している。しかし、説明部１０３におけるヒートマップの生成手法の種類は複数であってもよい。このとき、説明部１０３は、ヒートマップ同士の一致度に基づく損失の複数種類のヒートマップ生成手法についての合計値を説明評価結果の例として更新部１０６に出力してもよい。

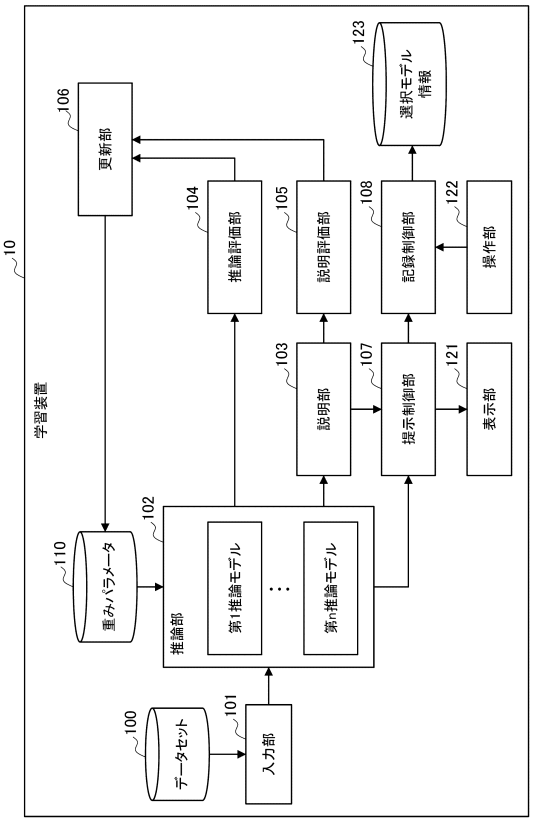
【符号の説明】

【０１１９】

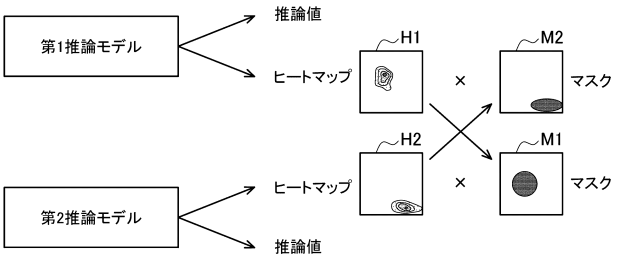
１０　学習装置
１００　データセット
１０１　入力部
１０２　推論部
１０３　説明部
１０４　推論評価部
１０５　説明評価部
１０６　更新部
１０７　提示制御部
１０８　記録制御部
１１０　重みパラメータ
１２１　表示部
１２２　操作部
１２３　選択モデル情報

【図面】

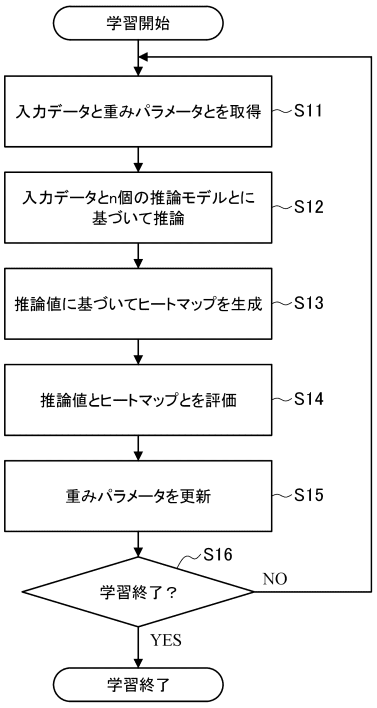
【図１】


10
学習装置
100
データセット
101
入力部
102
推論部
第1推論モデル
第n推論モデル
110
重みパラメータ
103
説明部
104
推論評価部
105
説明評価部
106
更新部
107
提示制御部
108
記録制御部
121
表示部
122
操作部
123
選択モデル情報

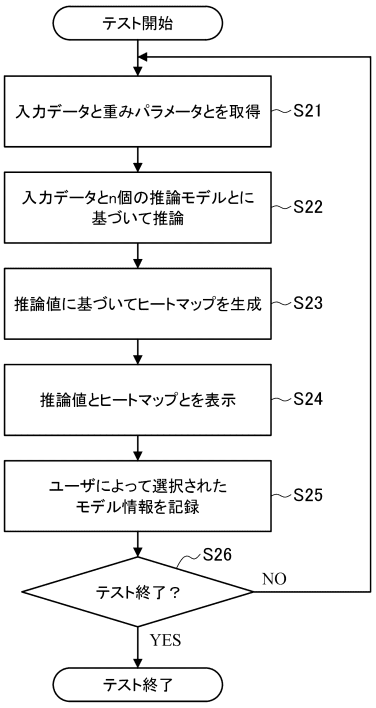
【図２】


第1推論モデル
推論値
ヒートマップ
H1
M2
マスク
第2推論モデル
ヒートマップ
H2
M1
マスク
推論値

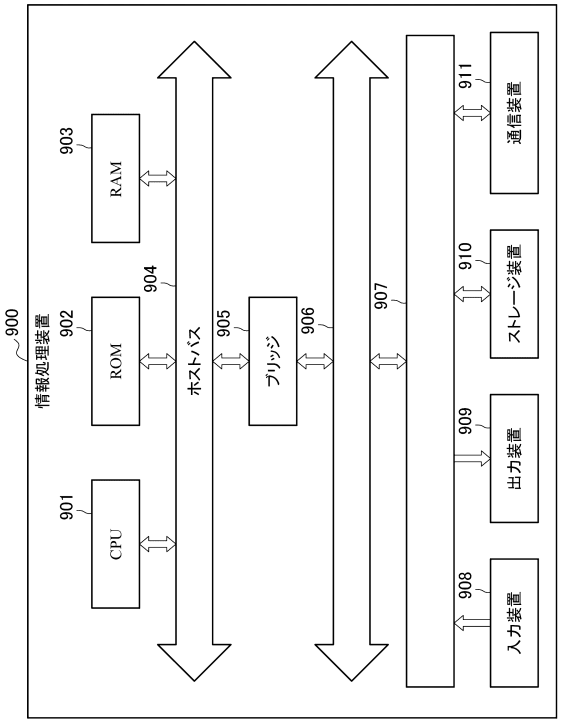
【図３】


学習開始
入力データと重みパラメータとを取得
S11
入力データとn個の推論モデルとに基づいて推論
S12
推論値に基づいてヒートマップを生成
S13
推論値とヒートマップとを評価
S14
重みパラメータを更新
S15
S16
学習終了？
NO
YES
学習終了

【図４】


テスト開始
入力データと重みパラメータとを取得
S21
入力データとn個の推論モデルとに基づいて推論
S22
推論値に基づいてヒートマップを生成
S23
推論値とヒートマップとを表示
S24
ユーザによって選択されたモデル情報を記録
S25
S26
テスト終了？
NO
YES
テスト終了

【図５】


900
情報処理装置
901
CPU
902
ROM
903
RAM
904
ホストバス
905
ブリッジ
906
907
908
入力装置
909
出力装置
910
ストレージ装置
911
通信装置