

明 細 書

発明の名称： 処理システム及び処理方法

技術分野

[0001] 本発明は、処理システム及び処理方法に関する。

背景技術

[0002] センサに代表されるIoTデバイスにより収集されたデータのデータ量は、膨大であるため、クラウドコンピューティングで収集されたデータを集約及び処理する際、膨大な通信量が発生する。このため、ユーザに近いエッジ装置でも、収集されたデータを処理するエッジコンピューティングに注目が集まっている。

[0003] しかしながら、エッジ装置で用いられる装置の演算量やメモリ等のリソースは、エッジ装置よりもユーザから物理的及び論理的に遠くに配置されたエッジ装置以外の装置(以下、簡便のためクラウドと記載する)と比して、貧弱である。このため、演算負荷が大きい処理をエッジ装置で行うと、処理が完了するまでに多大な時間を要したり、演算量が大きくない他の処理が完了するまでもにも時間を要したりする場合がある。

[0004] ここで、演算量が大きい処理の一つに機械学習に係る処理が挙げられる。非特許文献1には、いわゆるアダプティブラーニングのエッジ・クラウドへの適用が提案されている。すなわち、非特許文献1に記載の方法は、クラウドで汎用的な学習データを用いて学習を行った学習済みモデルをエッジ装置に展開し、エッジ装置で取得されたデータを用いて、クラウドで学習を行ったモデルに対して再度学習を行うことでクラウドとエッジ装置との利点を活かした運用を実現している。

先行技術文献

非特許文献

[0005] 非特許文献1：大越他，“クラウド・エッジ連携によるDNNモデル運用方式の提案と評価”，情報処理学会第80回全国大会講演論文集 2018(1)， 3-4，

2018-03-13.

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、この非特許文献 1 に記載の方法は、推論処理について検討されていない。推論は、処理する対象、すなわち、推論対象のデータが複雑であるほど、また、解こうとする問題が難しいほど、演算量は増大する。このような演算量が大きい処理はクラウドで処理することが好ましいと想定される。しかしながら、クラウドで処理すべき演算量の大きい処理を決定するために、推論対象のデータの複雑さや、解こうとする問題の難しさをエッジ装置で判定しなければならなかった。
- [0007] また、解こうとする問題の難しさとは異なる観点として、ユーザが要求する推論精度やレスポンスがある。すなわち、ユーザからは、推論精度があまり高くなくても早急なレスポンスを要する場合や、レスポンスが遅くとも高い推論精度が求められる場合がある。しかしながら、エッジ装置が、ユーザが要求する推論精度やレスポンスを考慮しながら、クラウドで処理すべき演算量が多い処理を決定する方法は、非特許文献 1 には記載されていない。
- [0008] 本発明は、上記に鑑みてなされたものであって、ユーザの要望に応じた、エッジ装置とクラウドとの連携した処理の実行を制御することができる処理システム及び処理方法を提供することを目的とする。

課題を解決するための手段

- [0009] 上述した課題を解決し、目的を達成するために、本発明に係る処理システムは、エッジ装置とサーバ装置とを用いて行われる処理システムであって、エッジ装置は、処理対象データを処理し、処理対象データの処理結果を出力するエッジ処理部と、エッジ装置とサーバ装置とのいずれにおいて処理対象データを処理すべきかを評価する評価値が条件を満たす場合、処理対象データに係る処理をサーバ装置に実行させることを判定し、処理対象データの処理結果が所定の評価を満たす場合、評価値をエッジ装置で処理すべき範囲に含まれる値と判定してエッジ処理部が処理した処理対象データの処理結果を

出力する判定部と、判定部が処理対象データに係る処理をサーバ装置に実行させることを判定した場合、処理対象データに係る処理をサーバ装置に実行させるためのデータを送信する送信部と、を有することを特徴とする。

[0010] また、本発明に係る処理方法は、エッジ装置とサーバ装置とを用いて行われる処理システムが実行する処理方法であって、エッジ装置が、処理対象データを処理し、処理対象データの処理結果を出力する処理工程と、エッジ装置が、エッジ装置とサーバ装置とのいずれにおいて処理対象データを処理すべきかを評価する評価値が条件を満たす場合、処理対象データに係る処理をサーバ装置に実行させることを判定し、処理対象データの処理結果が所定の評価を満たす場合、評価値をエッジ装置で処理すべき範囲に含まれる値と判定して処理工程において処理された処理対象データの処理結果を出力する判定工程と、エッジ装置が、判定工程において処理対象データに係る処理をサーバ装置に実行させることが判定された場合、処理対象データに係る処理をサーバ装置に実行させるためのデータを送信する送信工程と、を含んだことを特徴とする。

発明の効果

[0011] 本発明によれば、ユーザの要望に応じた、エッジ装置とクラウドとの連携した処理の実行を制御することができ、効率的に装置とクラウドとを合わせたシステム全体を効率的に稼働させることができる。

図面の簡単な説明

[0012] [図1]図1は、実施の形態1に係る処理システムの処理方法の概要を説明する図である。

[図2]図2は、DNN1及びDNN2の一例を説明する図である。

[図3]図3は、DNN1及びDNN2の一例を説明する図である。

[図4]図4は、実施の形態1に係る処理システムの構成の一例を模式的に示す図である。

[図5]図5は、実施の形態1に係る処理システムの処理の流れを示すシーケンス図である。

[図6]図6は、軽量モデル及び高精度モデルを学習する学習装置の構成例を示す図である。

[図7]図7は、ケースごとの損失の一例を示す図である。

[図8]図8は、高精度モデルの学習処理の流れを示すフローチャートである。

[図9]図9は、軽量モデルの学習処理の流れを示すフローチャートである。

[図10]図10は、実施の形態2に係る処理システムの構成の一例を模式的に示す図である。

[図11]図11は、実施の形態2に係る処理システムの処理の流れを示すシーケンス図である。

[図12]図12は、実施の形態2に係る処理システムの構成の他の例を模式的に示す図である。

[図13]図13は、実施の形態3に係る処理システムの構成の一例を模式的に示す図である。

[図14]図14は、図13に示すエッジ装置の一例を模式的に示す図である。

[図15]図15は、図13に示すサーバ装置の一例を模式的に示す図である。

[図16]図16は、実施の形態3に係る処理システムの処理の流れを示すシーケンス図である。

[図17]図17は、実施の形態4に係る処理システムの構成の一例を模式的に示す図である。

[図18]図18は、図17に示すエッジ装置の一例を模式的に示す図である。

[図19]図19は、実施の形態3に係る処理システムの処理の流れを示すシーケンス図である。

[図20]図20は、実施の形態1～4の変形例における処理システムの概略を示す図である。

[図21]図21は、DNN、判定部、符号化部、復号化部の機能のバリエーションと、通信内容のバリエーションを示す図である。

[図22]図22は、プログラムが実行されることにより、エッジ装置及びサーバ装置が実現されるコンピュータの一例を示す図である。

発明を実施するための形態

[0013] 以下、図面を参照して、本発明の一実施形態を詳細に説明する。なお、この実施形態により本発明が限定されるものではない。また、図面の記載において、同一部分には同一の符号を付して示している。

[0014] [実施の形態 1]

[実施の形態 1 の概要]

本発明の実施の形態について説明する。本発明の実施の形態 1 では、学習済みの高精度モデル及び軽量モデルを使って推論処理を行う処理システムについて説明する。なお、実施の形態の処理システムでは、推論処理において用いるモデルとして、DNN (Deep Neural Network) を用いた場合を例に説明する。実施の形態の処理システムでは、DNN 以外のニューラルネットワークを用いてもよいし、学習済みモデルに代えて低演算量の信号処理と高演算量の信号処理を用いてもよい。

[0015] 図 1 は、実施の形態 1 に係る処理システムの処理方法の概要を説明する図である。実施の形態 1 の処理システムは、高精度モデル及び軽量モデルは、モデルカスケードを構成する。実施の形態 1 の処理システムでは、高速かつ低精度な軽量モデル（例えば、DNN 1）を用いるエッジ装置と、低速かつ高精度な高精度モデル（例えば、DNN 2）を用いるクラウド（サーバ装置）とのいずれにおいて処理を実行するかを、評価値を用いて制御する。例えば、サーバ装置は、ユーザから物理的及び論理的に遠い場所に配置された装置である。エッジ装置は、ユーザから物理的及び論理的に近い場所に配置された IoT 機器及び各種端末装置であり、サーバ装置と比してリソースが少ない。

[0016] DNN 1 及び DNN 2 は、入力されたデータを基に推論結果を出力するモデルである。図 1 の例では、DNN 1 及び DNN 2 は、画像を入力とし、当該画像に写る物体のクラスごとの確率を推論する。なお、図 1 に示す 2 つの画像は、いずれも同じ画像である。例えば、DNN 1 は、DNN 1 及び DNN 2 のどちらのモデルで推論を行う方が、ユーザに要望された利益が大きい

かを含めてトレーニングされたものである。DNN1及びDNN2は、評価値について最適な値が得られるように最適化される。

[0017] ユーザの要望は、例えば、推論結果の高精度化、データ通信量の低減、計算処理の高速化、エッジ装置のリソース最適化等がある。評価値は、ユーザの要望を満たしつつエッジ装置とサーバ装置とのいずれにおいて処理対象データを処理すべきかを評価する値である。評価値は、処理対象のデータの処理が難しいほど、サーバ装置で評価すべき範囲に入る傾向を有する。

[0018] 図1に示すように、処理システムでは、入力画像に写る物体に対するDNN1のクラス分類の推論についての評価値を取得する。そして、処理システムでは、取得した評価値が所定条件、例えば、所定値を満たす場合、DNN1の推論結果が採用される。つまり、軽量モデルの推論結果が、モデルカスケードの最終的な推定結果として出力される。一方で、処理システムでは、評価値が所定値を満たさない場合、同一の画像をDNN2に入力して得られた推論結果が、最終的な推論結果として出力される。所定値を満たすとは、例えば、あらかじめ定められた閾値に基づく条件を満たすか否かや、所定の範囲に含まれるか否かを含む。

[0019] このように、実施の形態1に係る処理システムは、エッジ装置とサーバ装置とのいずれにおいて処理対象データを処理すべきかをユーザの要望に応じて評価する評価値を基に、エッジ装置またはサーバ装置を選択して、処理対象データを処理する。このため、実施の形態1に係る処理システムは、ユーザの要望に応じて、エッジ装置とクラウドとのいずれにおいて処理を実行するかを制御することができる。

[0020] [軽量モデル及び高精度モデル]

次に、DNN1、DNN2について説明する。図2及び図3は、DNN1及びDNN2の一例を説明する図である。DNNは、データが入る入力層、入力層から入力されたデータを様々に変換する1層または複数の中間層、確率や尤度など、いわゆる推論した結果を出力する出力層を有する。各層から出力される出力値は、入力されるデータが匿名性を保つ必要がある場合は非

可逆としてもよい。

[0021] 図2に示すように、処理システムは、それぞれ独立したDNN1a及びDNN2aを用いてもよい。例えば、DNN2aが既知の方法でトレーニングされた後、DNN1aが、DNN1a及びDNN2aのどちらのモデルで推論を行う方が、ユーザに要望された利益が大きいかを含めてトレーニングされる。DNN1aは、評価値に関する値を出力するようトレーニングされる。DNN1aは、評価値に関する値として、DNN1aの中間層の出力値である中間出力値を出力する。評価値は、中間出力値に基づいて算出される値でもよく、中間出力値そのものでもよい。中間出力値は、所定の中間層の中間出力値をよりユーザの要望にあうような学習ができるコスト関数、例えば中間出力値と尤度の相関を取るなど、にかけた結果を用いてもよいし、解くべき問題のみをコスト関数として設計した学習済みモデルの任意の中間層の出力を用いてもよい。これは、例えば高次の中間層に、入力されたデータを決定づける特徴が反映されるような性質をもちやすいネットワーク（CNNのような）を使用する場合、高次中間層の出力値には、解くべき問題で利用できる有効な特徴が抽出されているためである。なお、DNN1a及びDNN2aには、精度とパフォーマンスが異なる同じタスクが割り当てられてもよいし、異なるタスクが割り当てられてもよい。

[0022] また、図3に示すように、処理システムは、一体のDNNとしてトレーニングを行ったDNN3を、所定基準を用いて第R層と第(R+1)層との間でDNN1b及びDNN2bに分割する。そして、処理システムは、前段のDNN1bをエッジ装置に適用し、後段のDNN1bをサーバ装置に適用してもよい。この場合、DNN1bは、評価値として、第R中間層からの中間出力値を出力する。なお、DNN1bは、評価値として、第R中間層より前段の層からの中間出力値を出力してもよい。

[0023] また、評価値は、DNN1aまたはDNN1bから出力される中間出力値に限らない。例えば、評価値は、DNN1aから出力される推論誤差でもよく、該推論誤差に基づく値であってもよい。例えば、評価値は、エッジ装置が

処理を行った結果が正解であるかの確信度を示す値でもよい。また、評価値は、処理対象データの処理結果を得るまでに要する時間、処理対象データの処理結果の取得期限、エッジ装置またはサーバ装置のいずれにおいて処理対象データを処理すべきかを判定する際におけるエッジ装置のリソースの使用状況、処理対象データが他のデータと比してイベントが発生しているデータであるか、のいずれかに基づいて決定される値でもよい。なお、エッジ装置のリソースの使用状況は、当該エッジ装置単体のCPUやメモリの使用率、消費電力量などでもいいし、当該エッジ装置と他のエッジ装置での稼働量やリソース使用率の差などでもよい。また、イベントは、例えば、前のフレームと比べて対象フレームが所望の大きさ以上の変化がある場合や、細かく推定する対象が発生している場合を意図する。また、エッジ装置で演算を行った対象と結果を示すデータをサーバ装置に伝送し、サーバ装置はエッジ装置で演算を行っていない対象のみ演算するよう設計してもよい。具体的には、バウンディングボックスの座標値やクラス分類結果とその信頼度を合わせて送り、信頼度を満たさない対象のみサーバ装置で演算してもよい。

[0024] [処理システム]

次に、処理システムの構成について説明する。図4は、実施の形態1に係る処理システムの構成の一例を模式的に示す図である。

[0025] 実施の形態に係る処理システム100は、サーバ装置20及びエッジ装置30を有する。また、サーバ装置20及びエッジ装置30は、ネットワークNを介して接続される。ネットワークNは、例えばインターネットである。その場合、サーバ装置20は、クラウド環境に設けられたサーバであってもよい。また、エッジ装置30は、IoT機器及び各種端末装置であってもよい。

[0026] サーバ装置20及びエッジ装置30は、それぞれROM (Read Only Memory)、RAM (Random Access Memory)、CPU (Central Processing Unit) 等を含むコンピュータ等に所定のプログラムが読み込まれて、CPUが所定のプログラムを実行することで実現される。また、GPUやVPU

(Vision Processing Unit)、FPGA (Field Programmable Gate Array)、ASIC (Application Specific Integrated Circuit) や専用のAI (Artificial Intelligence) チップに代表されるいわゆるアクセラレータも用いられる。サーバ装置20及びエッジ装置30は、それぞれ、NIC (Network Interface Card) 等を有し、LAN (Local Area Network) やインターネットなどの電気通信回線を介した他の装置との間の通信を行うことも可能である。

[0027] 図4に示すように、サーバ装置20は、学習済みの高精度モデルであるDNN2を記憶する。DNN2は、モデルパラメータ等の情報を含む。また、サーバ装置20は、推論部22を有する。

[0028] 推論部22は、DNN2に推論用データ(処理対象データ)を入力し、推論結果(処理結果)を取得する。推論部22は、推論用データの受け付け、推論結果を出力する。推論用データは、ラベルが未知のデータであるものとする。例えば、推論用データは、画像である。また、推論結果をユーザに返す場合、推論部22で得られた推論結果はエッジ装置に伝送され、エッジ装置からユーザに返してもよい。

[0029] ここで、サーバ装置20及びエッジ装置30は、モデルカスケードを構成する。このため、推論部22は、常に推論用データについての推論を行うわけではない。推論部22は、推論用データに係る推論処理をサーバ装置20に実行させると判定された場合に、DNN2による推論を行う。

[0030] エッジ装置30は、学習済みの軽量モデルであるDNN1を記憶する。DNN1は、モデルパラメータ等の情報を含む。DNN1は、DNN1及びDNN2のどちらのモデルで推論を行う方が、ユーザに要望された利益が大きいかを含めてトレーニングされる。DNN1には、DNN1とDNN2とを含むモデルカスケードが、ユーザに要望された利益が大きいかを含めて最適化されるように予め学習されたパラメータが設定される。また、エッジ装置30は、推論部32(エッジ処理部)、判定部33及び通信部34(送信部)を有する。

- [0031] 推論部 32 は、DNN 1 に、推論用データ（処理対象データ）を入力して推論結果を取得する。推論部 32 は、推論用データの入力を受け付け、推論用データを処理し、推論結果（処理対象データの処理結果）を出力する。
- [0032] 判定部 33 は、エッジ装置 30 とサーバ装置 20 とのいずれにおいて推論用データを処理すべきかをユーザの要望を反映して設計された評価する評価値が、所定値を満たすか否かを判定する。
- [0033] 判定部 33 は、評価値が所定値を満たす場合、推論用データの推論結果が所定の評価を満たす場合であるとし、評価値をエッジ装置 30 で処理すべき範囲に含まれる値と判定して推論部 32 が推論した推論結果を出力する。評価値が所定値を満たさない場合、判定部 33 は、評価値がサーバ装置 20 で処理すべき範囲に含まれる値である場合であるとし、推論用データに係る処理（推論処理）をサーバ装置 20 に実行させることを判定する。評価値は、前述したように、中間出力値、推論誤差、確信度等である。また、判定部 33 は、送信対象とする処理用データの絞り込みを行ってもよい。例えば、判定部 33 は、処理用データを、DNN 2 の処理において必要であるノードのデータに絞り込む。ここで、推論用データが画像である場合に絞り込む基準を例示する。画像の一部領域にイベントが発生していた場合、判定部 33 は、当該一部領域若しくはイベントに係る推定を行うために必要となる領域に絞り込めばよい。また、画像の領域ごとにエッジ装置で処理を行うかサーバ装置で処理を行うかを判定する場合、判定部 33 は、サーバ装置で処理を行う領域に絞り込めばよい。空間的な観点での絞り込みを例示したが、判定部 33 は、時間的な観点での絞り込みを行ってもよい。
- [0034] 通信部 34 は、ネットワーク N を介して、他の装置（例えば、サーバ装置 20）と通信を行う。通信部 34 は、判定部 33 が推論用データに係る推論処理をサーバ装置 20 に実行させることを判定した場合、推論処理をサーバ装置 20 に実行させるための処理用データをサーバ装置 20 に送信する。通信部 34 は、評価値が中間出力値である場合、該中間出力値をサーバ装置 20 に送信する。

[0035] [処理システムの処理手順]

図5は、実施の形態1に係る処理システムの処理の流れを示すシーケンス図である。図5に示すように、まず、エッジ装置30において、推論部32は、推論用データの入力を受け付けると（ステップS1）、推論用データをDNN1に入力する（ステップS2）。

[0036] 判定部33は、DNN1の中間出力値を取得し（ステップS3、S4）、評価値を取得する（ステップS5）。判定部33は、評価値が所定値を満たすか否かを判定する（ステップS6）。

[0037] 評価値が所定値を満たす場合（ステップS6：Yes）、判定部33は、中間出力値を、DNN1の中間層のうち、中間出力値を出力した層の次の中間層に入力する（ステップS7）。推論部32は、DNN1の推論結果を取得し（ステップS8）、取得したDNN1の推論結果を出力する（ステップS9）。

[0038] これに対し、評価値が所定値を満たさない場合（ステップS6：No）、判定部33は、推論処理をサーバ装置20に実行させるための処理用データを、通信部34を介してサーバ装置20に送信する（ステップS10、S11）。例えば、処理用データは、推論用データ及びDNN1の確信度である。または、処理用データは、中間出力値である。

[0039] サーバ装置20では、推論部22が、処理用データをDNN2に入力し（ステップS11）、DNN2の推論結果を取得する（ステップS12、S13）。DNN2の推論結果は、エッジ装置30に送信され（ステップS14、S15）、エッジ装置30から出力される（ステップS16）。なお、本実施の形態では、ユーザに推論結果を返す構成を想定し、エッジ装置30から最終的な推論結果を出力する構成としているが、サーバ装置20側で最終的な推論結果を用いる場合は、DNN2の推論結果をサーバ装置20から出力する、或いは、サーバ装置20でそのまま保持するよう構成してもよい。DNN1の推論結果を使用する場合には、エッジ装置30は、サーバ装置20側で当該推論結果を用いる場合、推論結果をサーバ装置20に送信すればよ

い。

[0040] [実施の形態 1 の効果]

このように、実施の形態 1 によれば、エッジ装置とサーバ装置とのいずれにおいて処理対象データを処理すべきかをユーザの要望に応じて評価する評価値を基に、エッジ装置またはサーバ装置を選択して、処理対象データを処理する。このため、実施の形態 1 に係る処理システムは、ユーザの要望に応じて、エッジ装置とクラウドとのいずれにおいて処理を実行するかを制御することができる。

[0041] なお、本実施の形態 1 では、エッジ装置 30 の数とサーバ装置 20 の数とが、いずれも単数である場合について説明したが、エッジ装置 30 またはサーバ装置 20 が複数であってもよく、また、エッジ装置 30 とサーバ装置 20 とがいずれも複数であってもよい。

[0042] [適用例]

本実施の形態 1 を推論結果の高精度化の要望に対して適用し、確信度を評価値として採用する場合を例に説明する。まず、推論結果の高精度化を実現するための軽量モデル及び高精度モデルの学習について説明する。

[0043] 図 6 は、軽量モデル及び高精度モデルを学習する学習装置の構成例を示す図である。図 2 に示すように、学習装置 10 は、学習用データの入力を受け付け、学習済み高精度モデル情報及び学習済み軽量モデル情報を出力する。また、学習装置 10 は、高精度モデル学習部 11 及び軽量モデル学習部 12 を有する。

[0044] 高精度モデル学習部 11 は、推定部 111、損失計算部 112、更新部 113 を有する。また、高精度モデル学習部 11 は、高精度モデル情報 114 を記憶する。高精度モデル情報 114 は、高精度モデルを構築するためのパラメータ等の情報である。学習用データは、ラベルが既知のデータであるものとする。例えば、学習用データは、画像とラベル（正解のクラス）の組み合わせである。

[0045] 推定部 111 は、高精度モデル情報 114 を基に構築された高精度モデル

に学習用データを入力し、推定結果を取得する。推定部 1 1 1 は、学習用データの入力を受け付け、推定結果を出力する。

[0046] 損失計算部 1 1 2 は、推定部 1 1 1 によって取得された推定結果を基に損失を計算する。損失計算部 1 1 2 は、推定結果及びラベルの入力を受け付け、損失を出力する。例えば、損失計算部 1 1 2 は、推定部 1 1 1 によって取得された推定結果において、ラベルに対する確信度が小さいほど大きくなるように損失を計算する。例えば、確信度は、推定結果が正解であることの確からしさの度合いである。例えば、確信度は、前述の多クラス分類モデルが出力した確率であってもよい。具体的には、損失計算部 1 1 2 は、後述するソフトマックスクロスエントロピーを損失として計算することができる。

[0047] 更新部 1 1 3 は、損失が最適化されるように、高精度モデルのパラメータを更新する。例えば、高精度モデルがニューラルネットワークであれば、更新部 1 1 3 は、誤差逆伝播法等により高精度モデルのパラメータを更新する。具体的には、更新部 1 1 3 は、高精度モデル情報 1 1 4 を更新する。更新部 1 1 3 は、損失計算部 1 1 2 によって計算された損失の入力を受け付け、更新済みのモデルの情報を出力する。

[0048] 軽量モデル学習部 1 2 は、推定部 1 2 1、損失計算部 1 2 2、更新部 1 2 3 を有する。また、軽量モデル学習部 1 2 は、軽量モデル情報 1 2 4 を記憶する。軽量モデル情報 1 2 4 は、軽量モデルを構築するためのパラメータ等の情報である。

[0049] 推定部 1 2 1 は、軽量モデル情報 1 2 4 を基に構築された軽量モデルに学習用データを入力し、推定結果を取得する。推定部 1 2 1 は、学習用データの入力を受け付け、推定結果を出力する。

[0050] ここで、高精度モデル学習部 1 1 は、高精度モデルの出力を基に、高精度モデルの学習を行うものであった。一方で、軽量モデル学習部 1 2 は、高精度モデル及び軽量モデルの両方の出力を基に、軽量モデルの学習を行う。

[0051] 損失計算部 1 2 2 は、推定部によって取得された推定結果を基に損失を計算する。損失計算部 1 2 2 は、高精度モデルによる推定結果、軽量モデルに

よる推定結果及びラベルの入力を受け付け、損失を出力する。高精度モデルによる推定結果は、高精度モデル学習部 11 による学習が行われた後の高精度モデルに、さらに学習用データを入力して得られた推定結果であってよい。さらに具体的には、軽量モデル学習部 12 は、高精度モデルによる推定結果が正解であったか否かの入力を受け付ける。例えば、高精度モデルが出力した確率が最大であったクラスがラベルと一致していれば、その推定結果は正解である。

[0052] 損失計算部 122 は、軽量モデル単体での推定精度の最大化に加え、モデルカスケードを構成した場合の利益の最大化を目的として損失を計算する。ここで、利益は、推定精度が高いほど大きくなり、計算コストが小さいほど大きくなるものとする。

[0053] 例えば、高精度モデルには、推定精度は高いが計算コストが大きいという特徴がある。また、また、例えば、軽量モデルには、推定精度は低いが計算コストが小さいという特徴がある。そこで、損失計算部 122 は、(1) 式のように損失 Loss を計算する。ただし、 w は重みであり、事前に設定されるパラメータである。

[0054] [数1]

$$Loss = L_{\text{classifier}} + wL_{\text{cascade}} \quad \cdots (1)$$

[0055] ここで、 $L_{\text{classifier}}$ は、多クラス分類モデルにおけるソフトマックスエントロピーである。また、 $L_{\text{classifier}}$ は、軽量モデルによる推定結果における正解に対する確信度が小さいほど大きくなる第 1 の項の一例である。 $L_{\text{classifier}}$ は、(2) 式のように表される。ただし、 N はサンプル数である。また、 k はクラス数である。また、 y は正解のクラスを表すラベルである。また、 q は軽量モデルによって出力された確率である。 i はサンプルを識別する番号である。また、 j はクラスを識別する番号である。ラベル $y_{i,j}$ は、 i 番目のサンプルにおいて、 j 番目のクラスが正解であれば 1 になり、不正解であれば 0 になる。

[0056]

[数2]

$$L_{\text{classifier}} = \frac{1}{N} \sum_{i=1}^N \left\{ - \sum_{j=1}^K y_{i,j} \log q_{i,j} \right\} \cdots (2)$$

[0057] また、 L_{cascade} は、モデルカスケードを構成した場合の利益の最大化のための項である。 L_{cascade} は、各サンプルについて、軽量モデルの確信度に基づいて高精度モデル及び軽量モデルの推定結果を採用した場合の損失を表している。ここで、損失は、不適切な確信度へのペナルティと高精度モデルを用いるコストを含む。また、損失は高精度モデルの推定結果が正解か否かと、軽量モデルの推定結果が正解か否かとの組み合わせで4パターンに分けられる。詳細は後述するが、高精度モデルの推定が不正解、かつ軽量モデルの確信度が低い場合は、ペナルティは大きくなる。一方、軽量モデルの推定が正解、かつ軽量モデルの確信度が高い場合は、ペナルティは小さくなる。 L_{cascade} は、(3)式のように表される。

[0058] [数3]

$$L_{\text{cascade}} = \frac{1}{N} \sum_{i=1}^N \left\{ \max_j q_{i,j} 1_{\text{fast}} + \left(1 - \max_j q_{i,j} \right) 1_{\text{acc}} + \left(1 - \max_j q_{i,j} \right) \text{COST}_{\text{acc}} \right\} \cdots (3)$$

[0059] 1_{fast} は、軽量モデルの推定結果が正解であれば0、軽量モデルの推定結果が不正解であれば1を返す指示関数である。また、 1_{acc} は、高精度モデルの推定結果が正解であれば0、高精度モデルの推定結果が不正解であれば1を返す指示関数である。 COST_{acc} は、高精度モデルによる推定を行うことにかかるコストであり、事前に設定されるパラメータである。

[0060] $\max_j q_{i,j}$ は、軽量モデルが出力する確率の最大値であり、確信度の一例である。推定結果が正解であれば、確信度が大きいほど推定精度は高いといえる。一方、推定結果が不正解であれば、確信度が大きいほど推定精度は低いといえる。

[0061] (3)式の $\max_j q_{i,j} 1_{\text{fast}}$ は、軽量モデルによる推定結果が不正解である場合に

軽量モデルによる推定結果の確信度が大きいほど大きくなる第2の項の一例である。また、(3)式の $(1 - \max_j q_{i,j})1_{acc}$ は、高精度モデルによる推定結果が不正解である場合に軽量モデルによる推定結果の確信度が小さいほど大きくなる第3の項の一例である。また、(3)式の $(1 - \max_j q_{i,j})COST_{acc}$ は、軽量モデルによる推定結果の確信度が小さいほど大きくなる第4の項の一例である。この場合、更新部123による損失の最小化が、損失の最適化に相当する。

[0062] 更新部123は、損失が最適化されるように、軽量モデルのパラメータを更新する。つまり、更新部123は、軽量モデルによる推定結果と、入力されたデータを基に推定結果を出力するモデルであって、軽量モデルよりも処理速度が遅く推定精度が高い高精度モデルに学習用データを入力して得られた推定結果と、を基に、軽量モデルと高精度モデルを含むモデルカスケードが最適化されるように、軽量モデルのパラメータを更新する。更新部123は、損失計算部122によって計算された損失の入力を受け付け、更新済みのモデルの情報を出力する。

[0063] 図7は、ケースごとの損失の一例を示す図である。縦軸は $L_{cascade}$ の値である。また、横軸は、 $\max_j q_{i,j}$ の値である。また、 $COST_{acc}=0.5$ とする。 $\max_j q_{i,j}$ は、軽量モデルによる推定結果の確信度であり、ここでは単に確信度と呼ぶ。

[0064] 図7の「□」は、軽量モデル及び高精度モデルの両方の推定結果が正解である場合の、確信度に対する $L_{cascade}$ の値である。この場合、確信度が大きいほど $L_{cascade}$ の値は小さくなる。これは、軽量モデルによる推定結果が正解であれば、確信度が大きいほど軽量モデルが採用されやすくなるためである。

[0065] 図7の「◇」は、軽量モデルの推定結果が正解であり、高精度モデルの推定結果が不正解である場合の、確信度に対する $L_{cascade}$ の値である。この場合、確信度が大きいほど $L_{cascade}$ の値は小さくなる。また、「□」の場合と比べて、 $L_{cascade}$ の最大値及び小さくなる度合いが大きい。これは、高精度モデルによる推定結果が不正解であって、軽量モデルによる推定結果が正解であれば、確信度が大きいほど軽量モデルが採用されやすくなる傾向がさらに大きくなる

ためである。

[0066] 図7の黒塗りの四角は、軽量モデルの推定結果が不正解であり、高精度モデルの推定結果が正解である場合の、確信度に対する L_{cascade} の値である。この場合、確信度が大きいほど L_{cascade} の値は大きくなる。これは、軽量モデルの推定結果が不正解である場合も、確信度が小さいほど推定結果が採用されにくくなるためである。

[0067] 図7の「◆」は、軽量モデル及び高精度モデルの両方の推定結果が不正解である場合の、確信度に対する L_{cascade} の値である。この場合、確信度が大きいほど L_{cascade} の値は小さくなる。ただし、「□」の場合と比べて、 L_{cascade} の値は大きい。これは、両方のモデルの推定結果が不正解であることから常に損失が大きく、そのような状況では軽量モデルで正確な推定ができるようにすべきであるためである。

[0068] [学習処理]

図8は、高精度モデルの学習処理の流れを示すフローチャートである。図8に示すように、まず、推定部111は、高精度モデルを用いて学習用データのクラスを推定する（ステップS101）。

[0069] 次に、損失計算部112は、高精度モデルの推定結果を基に損失を計算する（ステップS102）。そして、更新部113は、損失が最適化されるように高精度モデルのパラメータを更新する（ステップS103）。なお、学習装置10は、終了条件が満たされるまで、ステップS101からステップS103までの処理を繰り返してもよい。終了条件は、既定の回数だけ処理が繰り返されたことであってよいし、パラメータの更新幅が収束したことであってよい。

[0070] 図9は、軽量モデルの学習処理の流れを示すフローチャートである。図9に示すように、まず、推定部121は、軽量モデルを用いて学習用データのクラスを推定する（ステップS201）。

[0071] 次に、損失計算部122は、軽量モデルの推定結果、及び高精度モデルの推定結果及び高精度モデルによる推定のコストを基に損失を計算する（ステ

ップS202)。そして、更新部123は、損失が最適化されるように軽量モデルのパラメータを更新する（ステップS203）。なお、学習装置10は、終了条件が満たされるまで、ステップS201からステップS203までの処理を繰り返してもよい。

[0072] このように、推定部121は、入力されたデータを基に推定結果を出力する軽量モデルに学習用データを入力し、第1の推定結果を取得する。また、更新部123は、第1の推定結果と、入力されたデータを基に推定結果を出力するモデルであって、軽量モデルよりも処理速度が遅く推定精度が高い高精度モデルに学習用データを入力して得られた第2の推定結果と、を基に、軽量モデルと高精度モデルを含むモデルカスケードが最適化されるように、軽量モデルのパラメータを更新する。このように、学習装置10では、軽量モデルと高精度モデルによって構成されるモデルカスケードにおいて、軽量モデルがモデルカスケードに適した推定を行えるようにすることで、モデルカスケードの性能を向上させることができる。その結果、学習装置10によれば、モデルカスケードの精度が向上するだけでなく、計算コスト及び計算リソースのオーバーヘッドを抑止することができる。さらに、第1の実施形態では、損失関数に変更を加えるものであるため、モデルアーキテクチャの変更が不要であり、適用するモデルや最適化手法に制限がない。

[0073] 更新部123は、第1の推定結果における正解に対する確信度が小さいほど大きくなる第1の項と、第1の推定結果が不正解である場合に第1の推定結果の確信度が大きいほど大きくなる第2の項と、第2の推定結果が不正解である場合に第1の推定結果の確信度が小さいほど大きくなる第3の項と、第1の推定結果の確信度が小さいほど大きくなる第4の項と、を含む損失関数を基に計算される損失が最小化されるように、軽量モデルのパラメータを更新する。この結果、第1の実施形態では、軽量モデルと高精度モデルによって構成されるモデルカスケードにおいて、高精度モデルの推定結果を採用する場合のコストを考慮した上で、モデルカスケードの推定精度を向上させることができる。

[0074] 処理システム100では、学習装置10による学習済みの高精度モデル及び軽量モデルを使って推論を行う場合、エッジ装置30は、軽量モデル(DNN1)に、推論用データを入力し、確信度を取得し、確信度が閾値以上である場合に、軽量モデルによる軽量モデルの推定結果を採用する。また、エッジ装置30は、確信度が閾値未満である場合に、サーバ装置20に処理用データを送信する。そして、処理システムは、サーバ装置20の高精度モデル(DNN2)に推論用データを入力することによって取得した高精度モデルの推定結果を採用する。

[0075] なお、本実施の形態1では、DNNを学習した例について説明したが、DNN以外の機械学習機構を用いてもよい。

[0076] [実施の形態2]

次の実施の形態2について説明する。実施の形態2では、エッジ装置が、処理用データを符号化した後にサーバ装置に送信する。

[0077] 図10は、実施の形態2に係る処理システムの構成の一例を模式的に示す図である。実施の形態2に係る処理システム200は、図4に示すサーバ装置20に代えてサーバ装置220を有し、エッジ装置30に代えてエッジ装置230を有する。

[0078] エッジ装置230は、エッジ装置30と比して、符号化部235を有する。符号化部235は、通信部34がサーバ装置220に送信するデータを符号化する。例えば、符号化部235は、送信するデータを圧縮して通信量を削減する。なお、サーバ装置220に送信するデータをDNN1の中間層の出力値とする場合、当該データを盗聴された場合であっても盗聴者が当該送信するデータが持つ意味を解釈することができないため、セキュリティの担保も可能とする。

[0079] 中間出力値は、推論用データの推論結果を出力する処理において出力されるDNN1の複数の中間出力値のうち、他の中間出力値と比して符号化が容易である値が選択される。符号化が容易であるとは、他の中間出力値と比して、エントロピーが小さい、あるいは、スパース性が高いことである。例え

ば、中間出力値は、所望の中間層の出力値のエントロピーが小さくなるように学習した学習済みのDNN1の中間層の中間出力値である。中間出力値は、所望の中間層の出力値スパース性が高くなるように学習した学習済みのDNN1の中間層の中間出力値である。

[0080] サーバ装置220は、サーバ装置20と比して、復号化部223を有する。復号化部223は、符号化部235によって符号化された処理用データを復号化し、推論部22に出力する。

[0081] ここで、DNN1及びDNN2が、一体のDNNとしてトレーニングを行ったDNN3（図3参照）を、所定基準を用いてDNN1b及びDNN2bに分割したモデルである場合、推論結果への歪が少なく効率的な符号化部235を構築することが望ましい。

[0082] 例えば、トレーニングセットのデータを一通り学習すると、転送対象の中間層の各ノードについて最大値、ゼロの発生頻度が見えてくるため、それに対応させた符号化処理を行うように符号化部235を設計する。符号化処理は、ゼロ発生頻度の高いノードの影響を小さく見積ることで符号化対象の表現空間の次元を小さくするような処理でもよいし、各ノードの値がとる範囲を見極めることでその傾向を反映した方式の選択や量子化の粒度を決める処理でもよい。

[0083] また、符号化部235は、ベクトル量子化的な手法を基に符号化を行ってもよい。この場合、符号化部235は、個別にノードの値を量子化するのではなく、ノード全体の値をベクトルととらえて、ベクトル空間でクラスタ化して、符号化する。

[0084] また、エントロピーが小さい層を求め、その層でDNN3を分割することで、符号化部235が、エントロピーが小さい中間出力値を得ることができるようする。

[0085] また、符号化部235及び復号化部223は、既知のルールに基づいた符号化・復号化方式を採用してもよいし、AE（AutoEncoder）やVAE（Variational AutoEncoder）のような学習に基づく方式を採用してもよい。

[0086] そして、符号化部235は、複数の符号化方法のうち、中間出力値及び送信先のDNN2に応じて、処理用データに対する符号化方式を切り替えてもよい。復号化部223は、符号化部235が実行した符号化方式に応じた方式でデータを復号化する。

[0087] [処理システムの処理手順]

図11は、実施の形態2に係る処理システムの処理の流れを示すシーケンス図である。図11に示すステップS21～ステップS29は、図5に示すステップS1～ステップS9と同じ処理である。

[0088] 評価値が所定値を満たさない場合（ステップS26：No）、符号化部235は、推論処理をサーバ装置220に実行させるための処理用データを符号化し（ステップS30）、符号化データを、通信部34を介してサーバ装置220に送信する（ステップS31、S32）。サーバ装置220では、復号化部223が符号化データを復号化し（ステップS33）、推論部22に復号化後の処理用データを出力する（ステップS34）。ステップS35～ステップS40は、図5に示すステップS11～ステップS16である。

[0089] [実施の形態2の効果]

このように、実施の形態2では、エッジ装置230は、処理用データを符号化してから、サーバ装置220に送信することで、セキュリティを保持した処理データの送信、推論結果への歪が少ないデータ形式での処理データの送信または効率的な処理データの送信を可能とする。

[0090] なお、実施の形態2では、エッジ装置230が符号化部236を有し、サーバ装置220が復号化部223を有する構成について説明したが、これに限らない。図12は、実施の形態2に係る処理システムの構成の他の例を模式的に示す図である。図12に示すように、エッジ装置230Aとサーバ装置220Aとの間のエッジ装置230A側のNW装置240Aに符号化部235を設け、サーバ装置220A側のNW装置250Aに復号化部223を設けてもよい。

[0091] また、本実施の形態2では、エッジ装置230またはサーバ装置220が

複数であってもよく、また、エッジ装置 230 とサーバ装置 220 とがいずれも複数であってもよい。

[0092] [実施の形態 3]

次に、実施の形態 3 について説明する。図 13 は、実施の形態 3 に係る処理システムの構成の一例を模式的に示す図である。図 13 に示すように、実施の形態 3 に係る処理システム 300 は、複数のエッジ装置 330-1, 330-2 がネットワーク N を介して、1 台のサーバ装置 320 に接続する構成を有する。エッジ装置の台数は、一例であり、3 台以上でもよい。エッジ装置 330-1, 330-2 を総称する場合には、エッジ装置 330 と記載する。

[0093] 図 14 は、図 13 に示すエッジ装置 330-1 の一例を模式的に示す図である。図 14 に示すように、エッジ装置 330-1 は、エッジ装置 30 と比して、付加部 336 を有する。付加部 336 は、自装置を特定する符号を処理用データに付加する。通信部 34 は、処理用データである中間出力値とともに、自装置を特定する符号を、サーバ装置 320 に送信する。

[0094] なお、エッジ装置 330-2 もエッジ装置 330-1 と同様の構成を有する。この際、各エッジ装置 330 が有する DNN1 は、同じモデルであってもよい。

[0095] また、各エッジ装置 330 が有する DNN1 は、モデル間で合意形成することで所定の中間層まで共通するようなマルチタスク学習で形成されたモデルであってもよい。モデル間の合意形成とは、例えば複数のモデルの同一の階層である中間層同士で合意形成を行いながら学習を行うことを意図する。つまり、モデル毎に異なる学習データを与えられた場合に自身に設定された問題に係るコスト項に加え、他のモデルの同一の階層である中間層同士で合意形成を行うコスト項の 2 つを同時に最適化したといいかえてもよい。この結果、各エッジ装置 330 が有する DNN1 は、入力層から所定の中間層までが同じ重みとなるよう学習されたモデルであってもよい。例えば、各エッジ装置 330 が有する DNN1 は、音響信号の特徴量抽出層まで共通し、そ

れ以降の層がそれぞれ異なる処理を行う。この場合、各エッジ装置 330 が出力する中間出力値は、共通層からの出力値となるように設定される。もちろん、エッジ装置 330 は、それぞれ異なる中間層の出力値をサーバ装置 320 に送信してもよい。

[0096] 図 15 は、図 13 に示すサーバ装置 320 の一例を模式的に示す図である。図 15 に示すように、サーバ装置 320 は、サーバ装置 20 と比して、格納部 324 及び推論結果データベース (DB) 325 を有する。格納部 324 は、推論部 22 が中間出力値を解析した結果 (推論結果) と、処理用データを送信したエッジ装置 330 を特定する符号とを対応付けて、推論結果 DB 325 に保存する。

[0097] 処理システム 300 では、複数のエッジ装置 330 のいずれのエッジ装置から送信されたデータであっても、推論処理を行うよう、エッジ装置 330 で行われる処理とサーバ装置 320 とにおいて行われる処理とが最適化されている。例えば、サーバ装置 320 の DNN2 は、いずれのエッジ装置 330 から送信された処理用データであっても対応できるように最適化されている。

[0098] [処理システムの処理手順]

図 16 は、実施の形態 3 に係る処理システムの処理の流れを示すシーケンス図である。図 16 に示すステップ S41～ステップ S49 は、図 5 に示すステップ S1～ステップ S9 と同じ処理である。

[0099] 評価値が所定値を満たさない場合 (ステップ S46: No)、付加部 336 は、自装置を特定する符号を処理用データに付加する (ステップ S50)。通信部 34 は、処理用データである中間出力値とともに、自装置を特定する符号を、サーバ装置 320 に送信する (ステップ S51, S52)。

[0100] 図 16 に示すステップ S53～ステップ S58 は、図 5 に示すステップ S11～ステップ S16 である。サーバ装置 320 では、格納部 324 が、推論結果と、処理用データを送信したエッジ装置 330 を特定する符号とを対応付けて、推論結果 DB 325 に格納する (ステップ S59～ステップ S6

1)。

[0101] [実施の形態3の効果]

このように、実施の形態3では、複数のエッジ装置330と接続する場合であっても、サーバ装置320のDNN2は、いずれのエッジ装置330から送信された処理用データであっても対応できるように最適化されている。そして、エッジ装置330は、処理用データである中間出力値とともに自装置を特定する符号を、サーバ装置320に送信する。このため、サーバ装置320のDNN2では、いずれのエッジ装置330から送信された処理用データを認識することで、処理データを用いた推論処理を適切に実行することができる。

[0102] なお、処理システム300は、実施の形態2において説明した符号化部235及び復号化部223を設けてもよい。

[0103] [実施の形態4]

次に、実施の形態4について説明する。図17は、実施の形態4に係る処理システムの構成の一例を模式的に示す図である。図14に示すように、実施の形態4に係る処理システム300は、エッジ装置430がネットワークNを介して、複数のサーバ装置420-1、420-2に接続する構成を有する。サーバ装置の台数は、一例であり、3台以上でもよい。サーバ装置420-1、420-2を総称する場合には、サーバ装置420と記載する。

[0104] サーバ装置420が有するDNN2は、例えば、それぞれ異なるタスクを行う。例えば、サーバ装置420-1のDNN2は、対象データの種類（画像か音響信号か）を分類する。サーバ装置420-2のDNN2は、対象データの性質（被写体認識タスクの場合だと人間か乗り物かなど）を分類する。また、他のサーバ装置420のDNN2は、対象データを処理する処理内容（被写体認識タスクか、音源強調タスクか）を分類する。例えば、エッジ装置430のDNN1が、データの特徴抽出を行うものである場合、各サーバ装置420のDNN2は、各サーバ装置420にそれぞれ与えられる各タスクに特化したものとなる。なお、異なるタスクを行わせようとする場合、

いわゆるマルチタスク学習を利用すればよい。具体的には、タスク１とタスク２について入力層から所定の中間層までの重みが共通するように学習させた所定の中間層までをエッジ装置４３０に配置し、所定の中間層より先をサーバ装置４２０に配置すればよい。このようにすることで、どのようなタスクであってもエッジ装置４３０に配置するモデルを共通としつつ、いずれのサーバ装置に配置されたモデルでも処理を可能とするよう構成することができる。また、異なるタスクというのは同一の目的であって異なる推定精度のものであってもよい。例えば、推定精度がエッジ装置４３０＜サーバ装置４２０－１＜サーバ装置４２０－２となるよう構成してもよい。

[0105] 図１８は、図１７に示すエッジ装置３３０－１の一例を模式的に示す図である。図１８に示すように、エッジ装置４３０は、エッジ装置３０と比して、選択部４３７を有する。選択部４３７は、推論用データを処理する目的に応じて、複数のサーバ装置４２０の中から、処理用データを送信するサーバ装置４２０を選択する。

[0106] [処理システムの処理手順]

図１９は、実施の形態４に係る処理システムの処理の流れを示すシーケンス図である。図１９に示すステップＳ７１～ステップＳ７９は、図５に示すステップＳ１～ステップＳ９と同じ処理である。

[0107] 評価値が所定値を満たさない場合（ステップＳ７６：Ｎｏ）、選択部４３７は、推論用データを処理する目的や精度に応じて、送信先のサーバ装置４２０を選択する（ステップＳ８０）。通信部３４は、選択部４３７が選択したサーバ装置４２０（例えば、サーバ装置４２０－１）に処理用データを送信する（ステップＳ８１，Ｓ８２）。図１９に示すステップＳ８３～ステップＳ８８は、図５に示すステップＳ１１～ステップＳ１６と同じ処理である。なお、送信先の選択を行う（ステップＳ８０）選択部４３７は、物理的及び／または論理的にエッジ装置側に置かれてもよいし、サーバ装置側に置かれてもよい。また、選択部４３７は、ネットワーク上（サーバ側ともエッジ側とも区分できない位置）においてもよい。

[0108] [実施の形態4の効果]

このように、実施の形態4では、エッジ装置430が複数のサーバ装置420と接続する場合であっても、推論用データを処理する目的に応じて、送信先のサーバ装置420を選択することによって、推論処理を適切に実行することができる。

[0109] なお、本実施の形態4では、エッジ装置430が複数であってもよい。また、処理システム400は、エッジ装置とサーバ装置との間のNW装置に選択部437部を設けてもよい。また、処理システム400は、実施の形態2において説明した符号化部235及び復号化部223を設けてもよい。この際、選択部237の配置場所は、符号部235の前段でもよく後段でもよい。

[0110] [変形例]

次に、実施の形態1～4の変形例について説明する。図20は、実施の形態1～4の変形例における処理システムの概略を示す図である。以降では、図21を参照して、図20に示すDNN1、DNN2、判定部33、符号化部235、復号化部223及びDNN2のそれぞれの機能のバリエーションや、通信内容のバリエーションを説明する。

[0111] 図21は、DNN1、DNN2、判定部33、符号化部235及び復号化部223の機能のバリエーションと、通信内容のバリエーションを示す図である。図21に示す各欄のうち、(1-A)～(1-H)は、DNN1、DNN2の機能のバリエーションを示す。また、(2-A)～(2-G)は、判定部33の機能のバリエーションを示す。また、(3-A)～(3-F)は、符号化部235、復号化部223及びエッジ装置とサーバ装置との間の通信内容のバリエーションを示す。各機能部及び通信内容は、これらの各欄に示すものとしてすることができる。

[0112] また、各機能部や通信内容を組み合わせて運用することもできる。例えば、それぞれ独立したDNN1a及びDNN2aを用いる場合(図2参照)、(1-D)、(2-C)、(3-A)の機能及び通信内容を組み合わせること

ができる。また、一体のDNN3を分割したDNN1b及びDNN2bを用いる場合（図3参照）、（1-D）、（1-G）のいずれかと、（2-C）、（2-D）のいずれかと、（3-B）、（3-C）、（3-D-1）、（3-D-2）、（3-D-3）のいずれかと、を組み合わせることができる。

[0113] 本発明は、ユーザの要望が多岐にわたるような様々な場合に適用することができる。いくつか具体例を挙げる。

[0114] [自動運転]

自動車内に配置されたDSP (Digital Signal Processor) 等の演算装置をエッジとし、クラウドとの連携について例を挙げる。例えば、渋滞を考慮したナビゲーションのように演算量、伝送量共に多くなりがちであるもののレスポンスが遅くてもいいような処理はサーバ装置で、自動車の直接の制御に係るイベント検知や検知したイベントに応じた自動車の制御の決定などはある程度の精度とレスポンスの速さが要求されるためエッジ装置側で処理するよう構成されてもよい。

[0115] [変化検出]

時系列画像信号を対象とした場合に、エッジ装置で平常時若しくは前フレームと比較した変化の有無をエッジ装置で検出し、どのような変化が起こったかの推定をサーバ装置で行ってもよい。

[0116] 時系列画像信号は監視カメラでもよいし、衛星画像や航空写真でもよい。監視カメラの場合は人間が監視カメラの前を通過した場合を変化としてエッジ装置で検出し、どのような人間が通過したかをサーバ装置で推定してもよい。衛星画像の場合、建物のエッジやテクスチャの変化、船や自動車の通過を変化としてエッジ装置で検出し、どのような建物ができたかや建築状況、どのような船が通過したかなどをサーバ装置で推定してもよい。この場合、飛行機や衛星に配置された演算装置をエッジとして扱ってもよい。

[0117] [防犯]

比較的単純・軽量の推論（人数のカウント、性別・年齢等推定、大まかな

服装判定等)はエッジ装置で行い、より負荷のかかる複雑な推論(人物特定、姿勢推定、不審者検出等)をクラウド(サーバ装置)で行う。

[0118] また、即応性の求められるV I P (Virtual IP) , リピータ, クレーマ等の既知の要注意人物検出をエッジ装置で行い、時間をかけても問題ない、より一般的な人物検出、その人物の特徴抽出、D B化等はクラウドで行う。

[0119] [農業]

無人制御トラクタについて、前方に障害物がないことの確認はエッジ装置(トラクタ単体)で行い、何かがありそれにどう対処するかまで含めた推論・プランニングは、クラウドで行う。

[0120] [推論基盤的なビジョン]

カメラからの映像を局舎で受けて映像処理(通常の二層推論)、処理結果をクラウドにまわしてより高度な処理や集計的な処理を行う(多段推論)。ある局舎Aのリソースが枯渇しており、その隣の局舎Bでリソースが余っている場合、クラウドからの制御で局舎Aの一部加工済みデータを局舎Bへまわして、残りの処理を行う。これにより効率的なリソース利用が可能となる(サービスのロバスト性、リソースの効率的利用)。これは、局舎に配置される演算装置等をいわゆるエッジクラウドとして制御してもよいことを意図する。

[0121] [ドローンカメラ群の制御]

複数のドローンカメラ群の全体的な撮影計画に伴う個々のドローンカメラの配置や状況に応じたカメラ間のリカバー対応等はクラウド側で制御・指示し、個々のドローンカメラ固有の状況対応等に関わる推論・判断(急にカメラの前に障害物が出現した場合の回避等)はドローン(エッジ装置)側で行う。この例では、多対1の実施の形態3を適用する。

[0122] また、1対多の実施の形態4の適用例について説明する。1つのカメラ画像をエッジ(D N N 1)で特徴量化、それを複数のクラウドに並列に渡して多様なタスク処理(人数カウント、人物特定、クラス分類、姿勢推定等)を行うために共通的に利用する。1対多の事例であり、プライバシー保護に関し

ては、符号化処理を適用する。

[0123] [システム構成等]

図示した各装置の各構成要素は機能概念的なものであり、必ずしも物理的に図示の如く構成されていることを要しない。すなわち、各装置の分散・統合の具体的形態は図示のものに限られず、その全部又は一部を、各種の負荷や使用状況等に応じて、任意の単位で機能的又は物理的に分散・統合して構成することができる。さらに、各装置にて行なわれる各処理機能は、その全部又は任意の一部が、CPU及び当該CPUにて解析実行されるプログラムにて実現され、あるいは、ワイヤードロジックによるハードウェアとして実現され得る。

[0124] また、本実施形態において説明した各処理のうち、自動的に行われるものとして説明した処理の全部又は一部を手動的におこなうこともでき、あるいは、手動的におこなわれるものとして説明した処理の全部又は一部を公知の方法で自動的におこなうこともできる。この他、上記文書中や図面中で示した処理手順、制御手順、具体的名称、各種のデータやパラメータを含む情報については、特記する場合を除いて任意に変更することができる。

[0125] [プログラム]

図22は、プログラムが実行されることにより、エッジ装置30、230、330、430及びサーバ装置20、220、320、420が実現されるコンピュータの一例を示す図である。コンピュータ1000は、例えば、メモリ1010、CPU1020を有する。また、演算を補助するために前述したアクセラレータを備えてもよい。また、コンピュータ1000は、ハードディスクドライブインタフェース1030、ディスクドライブインタフェース1040、シリアルポートインタフェース1050、ビデオアダプタ1060、ネットワークインタフェース1070を有する。これらの各部は、バス1080によって接続される。

[0126] メモリ1010は、ROM (Read Only Memory) 1011及びRAM 1012を含む。ROM 1011は、例えば、BIOS (Basic Input Output

t System) 等のブートプログラムを記憶する。ハードディスクドライブインタフェース 1030 は、ハードディスクドライブ 1090 に接続される。ディスクドライブインタフェース 1040 は、ディスクドライブ 1100 に接続される。例えば磁気ディスクや光ディスク等の着脱可能な記憶媒体が、ディスクドライブ 1100 に挿入される。シリアルポートインタフェース 1050 は、例えばマウス 1110、キーボード 1120 に接続される。ビデオアダプタ 1060 は、例えばディスプレイ 1130 に接続される。

[0127] ハードディスクドライブ 1090 は、例えば、OS (Operating System) 1091、アプリケーションプログラム 1092、プログラムモジュール 1093、プログラムデータ 1094 を記憶する。すなわち、エッジ装置 30, 230, 330, 430 及びサーバ装置 20, 220, 320, 420 の各処理を規定するプログラムは、コンピュータにより実行可能なコードが記述されたプログラムモジュール 1093 として実装される。プログラムモジュール 1093 は、例えばハードディスクドライブ 1090 に記憶される。例えば、エッジ装置 30, 230, 330, 430 及びサーバ装置 20, 220, 320, 420 における機能構成と同様の処理を実行するためのプログラムモジュール 1093 が、ハードディスクドライブ 1090 に記憶される。なお、ハードディスクドライブ 1090 は、SSD (Solid State Drive) により代替されてもよい。

[0128] また、上述した実施形態の処理で用いられる設定データは、プログラムデータ 1094 として、例えばメモリ 1010 やハードディスクドライブ 1090 に記憶される。そして、CPU 1020 が、メモリ 1010 やハードディスクドライブ 1090 に記憶されたプログラムモジュール 1093 やプログラムデータ 1094 を必要に応じて RAM 1012 に読み出して実行する。

[0129] なお、プログラムモジュール 1093 やプログラムデータ 1094 は、ハードディスクドライブ 1090 に記憶される場合に限らず、例えば着脱可能な記憶媒体に記憶され、ディスクドライブ 1100 等を介して CPU 102

０によって読み出されてもよい。あるいは、プログラムモジュール１０９３及びプログラムデータ１０９４は、ネットワーク（ＬＡＮ（Local Area Network）、ＷＡＮ（Wide Area Network）等）を介して接続された他のコンピュータに記憶されてもよい。そして、プログラムモジュール１０９３及びプログラムデータ１０９４は、他のコンピュータから、ネットワークインタフェース１０７０を介してＣＰＵ１０２０によって読み出されてもよい。

[0130] 以上、本発明者によってなされた発明を適用した実施形態について説明したが、本実施形態による本発明の開示の一部をなす記述及び図面により本発明は限定されることはない。すなわち、本実施形態に基づいて当業者等によりなされる他の実施形態、実施例及び運用技術等は全て本発明の範疇に含まれる。

符号の説明

- [0131] １０ 学習装置
- １１ 高精度モデル学習部
 - １２ 軽量モデル学習部
 - ３０，２３０，２３０Ａ，３３０，４３０ エッジ装置
 - ２０，２２０，２２０Ａ，３２０，４２０サーバ装置
 - １００，２００，３００，４００ 処理システム
 - １１１，１２１ 推定部
 - １１２，１２２ 損失計算部
 - １１３，１２３ 更新部
 - １１４ 高精度モデル情報
 - １２４ 軽量モデル情報
 - ２２，３２ 推論部
 - ３３ 判定部
 - ３４ 通信部
 - ２２３ 復号化部
 - ２３５ 符号化部

2 4 0 A, 2 5 0 A NW装置

3 2 4 格納部

3 2 5 推論結果データベース (D B)

3 3 6 付加部

4 3 7 選択部