

明 細 書

発明の名称： 処理システム、処理方法及び処理プログラム

技術分野

[0001] 本発明は、処理システム、処理方法及び処理プログラムに関する。

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時間を要したりする場合がある。例えば、演算量が大きな処理がエッジ装置で行われている間に、他の処理について処理待ちが発生してしまうような場合である。

[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”，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] 上記のエッジ・クラウドネットワークは、例えば、監視カメラ映像の自動分析、自動運転、スマートスピーカー等などのアプリケーションへの適用が期待されている。これらのアプリケーションでは、精度とともに、リアルタイム性が重要とされているものの、エッジ・クラウド間の、通信コストや遅延が問題となっていた。

[0007] 本発明は、上記に鑑みてなされたものであって、エッジ装置からサーバ装置に対するデータ転送量の低減及び低遅延化を図ることができる処理システム、処理方法及び処理プログラム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

[0008] 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、第1のモデルを用いて、処理対象データの特徴量を抽出し、抽出した特徴量を基に処理対象データに対する推論処理を実行する第1の推論部と、第1の推論部による推論結果に対する信頼度が閾値を超えた場合には第1の推論部による推論結果を出力し、信頼度が閾値以下である場合には処理対象データの特徴量をサーバ装置に出力する判定部と、を有し、サーバ装置は、第1のモデルよりも推論精度が高い第2のモデルを用いて、エッジ装置から出力された処理対象データの特徴量を基に処理対象データに対する推論処理を実行する第2の推論部と、を有することを特徴とする。

発明の効果

[0009] 本発明によれば、エッジ装置からサーバ装置に対するデータ転送量の低減及び低遅延化を図ることができる。

図面の簡単な説明

[0010] [図1]図1は、実施の形態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2-1]図2-1は、DNN1及びDNN2の一例を説明する図である。

[図2-2]図2-2は、DNN1及びDNN2の一例を説明する図である。

[図3]図3は、実施の形態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4]図4は、DNNのベースモデルの選定例について説明する図である。

[図5]図5は、YOLOv3の構造の概略を示す図である。

[図6-1]図6-1は、DNNの構造の一例を説明する図である。

[図6-2]図6-2は、DNNの構造の一例を説明する図である。

[図7]図7は、実施の形態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8]図8は、処理システムにおいて一例として選定した各DNNの処理時間を示す図である。

[図9]図9は、DNNの推論結果を基に求めたテストデータのエントロピーの分布図である。

[図10]図10は、オフロード率と全体精度との関係を示す図である。

[図11]図11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

[0011] 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。

[0012] [実施の形態]

[実施の形態の概要]

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、どのような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。

[0013] 図1は、実施の形態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、DNN1（第1のモデル））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、DNN2（第2のモデル））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、信頼度を用いて制御する。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。

[0014] DNN1及びDNN2は、入力された処理対象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、DNN1は、画像を入力とし、画像から抽出した特徴量を基に、当該画像に写る物体のクラスごとの確率を推論する。DNN2は、DNN1が抽出した特徴量（所定の間層層の出力を意図する、以下、特徴マップという）を入力として、推論対象である画像に写る物体のクラスごとの確率を推論する。なお、DNN1及びDNN2は、いずれも同じ画像に対する推論を行う。また、処理システムでは、エッジ装置において、信頼度（後述）と閾値との比較結果に基づいて、DNN1及びDNN2のいずれにおける推論結果を採用するかを判定する。信頼度は、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判定するための値である。信頼度は、例えば、DNN1による推論結果のエントロピーに基づいて求めてもよい。信頼度はDNN1の推論結果のエントロピーが高いほど小さくなる値とすればよい。具体例は後述する。

[0015] 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信頼度を取得する。そして、処理システムでは、取得した信頼度が所定の閾値を超える場合、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、信頼度が所定の閾値以下である場合、DNN1が抽出した特徴マップを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。

[0016] このように、実施の形態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、信頼度と閾値との比較結果を基に、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、実施の形態に係る処理システムは、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。

[0017] [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2-1及び図2-2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。

[0018] 上述したように、処理システムは、クラウド側でDNN2を用い、エッジ側でDNN1を用いる。クラウド側に配置する前のDNN2'として、入力画像の特徴量を抽出して特徴マップとして出力する特徴抽出層Bf2と、抽出された特徴量を用いた処理を行う層（たとえば、入力画像に写る物体を検知したり、入力画像に写る物体のクラスごとの確率を推論したりする層（以下まとめて検知層Bd2という））検知層Bd2（第2の実行部）とを有する高精度モデルを採用する（図2-1参照）。処理システムでは、予め学習用データを用いて、配置前にDNN2'を訓練する。

[0019] そして、処理システムでは、エッジ側に配置する前に、エッジ側に配置す

る前のDNN1'の特徴量抽出層Bf1に代えて、訓練済みのDNN2'の特徴抽出層Bf2を、そのままエッジ側の軽量な検知層Bd1の前段に配置し（図2-1の矢印Y1参照）、特徴抽出層Bf2（第1の抽出部）と検知層Bd1（第1の実行部）とを組み合わせたDNNを、エッジ側のDNN1とする（図2-2参照）。検知層Bd1は、DNN1の特徴抽出層Bf2が抽出した特徴マップ（ある層の特徴量）を用いて、入力画像に写る物体のクラスごとの確率を推論する。ここで、エッジ側に配置したDNN1とクラウド側に配置したDNN2の構成について説明する。エッジ側に配置したDNN1は、当初のDNN1から特徴量抽出層Bf1（図2-1参照）を削除し、訓練済みのDNN2'の特徴量抽出層Bf2を代わりに配置する（図2-2参照）。すなわち、DNN2'の特徴量抽出層Bf2と、DNN1'の検知層Bd1との組合せからなるDNN1が、エッジ側に配置される。なお、図2-2に示すDNN1のうち、特徴量抽出層Bf2を固定し、再度学習を行ってもよい。また、サーバ側に配置するDNN2は、当初のDNN2'（図2-1参照）から、特徴量抽出層Bf2を削除した構成となる（図2-2参照）。

[0020] DNN1については、特徴抽出層Bf2のパラメータを当初のDNN2'の訓練後のパラメータに固定し、後段の検知層Bd1を、学習用データを用いて訓練する。或いは、DNN1については、特徴抽出層Bf2及び検知層Bd1のいずれに対しても、学習用データを用いて訓練する。また、DNN2とDNN1とで独立して学習を行う他、DNN2とDNN1とで連携して学習を行ってもよい。例えば、DNN1が用いた学習用データを用いてDNN2を再訓練してもよい。また、DNN1とDNN2との相互で、共通する学習用データを用いて訓練を行ってもよい。

[0021] 本実施の形態に係る処理システムでは、当初のDNN2'における特徴抽出層Bf2を、エッジ側のDNN1の特徴抽出層Bf2として配置する。このため、クラウド側のDNN2の検知層Bd2は、エッジ側のDNN1の特徴抽出層Bf2が出力した特徴マップを用いて推論処理を実行することが可

能になる（図2-2の矢印Y2参照）。したがって、本実施の形態に係る処理システムでは、エッジ・クラウド間で同じ特徴マップを共有して、エッジ側、クラウド側のそれぞれにおいて、推論処理が可能であるといえる。

[0022] これによって、クラウド側で推論を行う場合に、特徴抽出処理の実行を省略できるため、計算時間を短縮することができ、低遅延化を図ることが可能になる。また、エッジ側からクラウド側に出力するデータは、処理対象の画像ではなく、画像から抽出した特徴マップであるため、エッジ側からクラウド側に対するデータ転送量を低減することが可能になる。なお、エッジ・クラウド間で同じ特徴マップを共有できればよいため、エッジ側に実際に配置されるDNN1の特徴抽出層Bf2と、当初のDNN2'の特徴抽出層Bf2とが満たすべき最小の条件は、接続する層のサイズが一致することである。これは、DNN1の検知層については改めて再学習を行うためである。なお、DNN1の特徴抽出層Bf2とDNN2'の特徴抽出層Bf2とでパラメータの値が異なってもよい。

[0023] [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3は、実施の形態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。

[0024] 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。例えば、サーバ装置20は、クラウド環境に設けられたサーバである。また、エッジ装置30は、例えば、IoT機器及び各種端末装置である。

[0025] 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。

[0026] 図3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を用いて推論を行う推論部21(第2の推論部)を有する。DNN2は、モデルパラメータ等の情報を含む。DNN2は、前述したように検知層Bd2を有する。

[0027] 推論部21は、DNN2を用いて、エッジ装置30から出力された入力画像の特徴マップを基に、入力画像に対する推論処理を実行する。推論部21は、エッジ装置30から出力された、量子化された特徴マップを、FP32に戻して、DNN2の検知層Bd2の入力とする。推論部21は、DNN2の出力として推論結果(例えば、画像に写る物体のクラスごとの確率)を取得する。推論部21は、推論用データ、つまりDNN1の特徴抽出層のBf2の出力値を受け付け、推論結果を出力する。特徴マップは、ラベルが未知のデータの特徴量であるものとする。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21で得られた推論結果はエッジ装置30に伝送され、エッジ装置30からユーザに返してもよい。

[0028] 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21は、常に推論を行うわけではない。推論部21は、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、量子化された特徴マップの入力を受け付けて、DNN2の検知層Bd2による推論を行う。

[0029] エッジ装置30は、学習済みの軽量モデルであるDNN1を有する推論部31(第1の推論部)と、判定部32と、量子化部33を有する。

[0030] 推論部31は、学習済みの軽量モデルであるDNN1を用いて推論を行う。DNN1は、モデルパラメータ等の情報を含む。DNN1は、前述したよ

うに訓練済みのDNN2' (図2-1参照) の特徴抽出層Bf2と、検知層Bd1とを有する。推論部31は、DNN1を用いて、入力画像の特徴量を特徴マップとして抽出し、抽出した特徴マップを基に入力画像に対する推論処理を実行する。

[0031] 推論部31は、DNN1に、処理対象の画像を入力して推論結果を取得する。推論部31は、DNN1を用いて、処理対象データの特徴量を抽出し、抽出した特徴量を基に前記処理対象データに対する推論処理を実行する。推論部31は、処理対象の画像の入力を受け付け、処理対象の画像を処理し、推論結果（例えば、画像に写る物体のクラスごとの確率）を出力する。

[0032] 判定部32は、エッジ装置30とサーバ装置20とのいずれの推論結果を採用するか否かを、信頼度と所定の閾値とを比較することで判定する。

[0033] 判定部32は、信頼度が所定の閾値を超えた場合、推論部31が推論した推論結果を出力する。判定部32は、信頼度が所定の閾値以下である場合、エッジ側のDNN1につながられた特徴抽出層Bf2の出力である特徴マップをサーバ装置20に出力して、推論処理をサーバ装置20に配置されたDNN2に実行させることを判定する。

[0034] 信頼度は、前述したようにDNN1の推論結果のエントロピーが大きいほど小さくなる値とすればよい。DNN1の推論結果のエントロピーは、式(1)で求めればよい。Cは、DNN1で出力される全てのラベルcの集合であり、yはそれぞれのラベルの確率である。信頼度は、このように求めたエントロピーから、例えば信頼度=1/エントロピーのように求めればよい。

[0035] [数1]

$$\text{entropy}(y) = \sum_{c \in C} -y_c \log y_c \dots (1)$$

[0036] 量子化部33は、判定部32が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、DNN1の特徴抽出層Bf2によって抽出された特徴マップを量子化し、量子化した特徴マップをサーバ装置20に出力する。

[0037] [DNN1、DNN2のモデル選定例]

図4は、DNN1、DNN2のベースモデルの選定例について説明する図である。本実施の形態では、物体のクラスごとの確率を推論するモデルとして、図4に示すように、比較的軽量かつ高速であるYOLOv2のバックエンドモデルであるdarknet19（以下、YOLOv2と記載する。）を、エッジ装置30のベースモデルとして選定し、比較的高精度であるYOLOv3のバックエンドモデルであるdarknet53（以下、YOLOv3と記載する。）を、サーバ装置20のベースモデルとして選定した。選定したNNは例であり、高精度なNNと、高速であるNNであって高精度なNNよりも精度は落ちるNNとであって、特徴量抽出層を共有できるNNであればどのように選択をしてもよい。簡単な例では、同一のNNで、エッジ装置30とサーバ装置20とで深さが異なるように構成してもよい。

[0038] 図5は、YOLOv3の構造の概略を示す図である。DNN1、DNN2のベースモデルの選定例について説明する図である。YOLOv3は、Residual block Bf-1を有する特徴抽出を行う畳み込み層（特徴抽出層）Bf-1と、物体検知用のネットワーク（FPN）である検知部Bd2-1とを有する。ここで、エッジ・クラウド間で特徴マップを共有するために、以下に注目した。

[0039] まず、Residual block Bf-1が入ることで、YOLOv3では、高精度を保持する。このため、Residual block Bf-1内の構成を損なうことを避けることが望ましい。続いて、YOLOv3では、検知部Bd2-1が特徴マップをそのまま用いて検知を実行できるように、検知部Bd2-1の前段で特徴マップを受けつける構成とすることが望ましい。なお、層が深くなるほど（多くなるほど）、モデルの計算が重くなり、使用するパラメータ数も多くなる。このため、エッジ装置30のDNN1は、軽量かつ高速とするために、層が浅いモデルであることが望ましい。

[0040] [DNN1, DNN2の構造の一例]

図6-1及び図6-2は、DNN1、DNN2の構造の一例を説明する図である。YOLOv3のResidual block Bf-1内の構成を損なうことを避ける

ため、Residual block B f - 1 を保持する学習済みの特徴抽出層 B f 2 - 1 を、そのまま、エッジ装置 30 のベースモデルとして選定した Y0L0v2 の前段に配置し（図 6 - 1 の矢印 Y 1 1 ）、エッジ装置 30 の D N N 1 とした。したがって、エッジ装置 30 では、このように再構築した Y0L0v2 を用いて、特徴抽出層 B f 2 - 1 において画像から特徴マップを抽出し、抽出した特徴マップを用いて、検知層 B d 1 - 1 において、入力画像に写る物体のクラスごとの確率を推論する。

[0041] そして、エッジ装置 30 では、検知層 B d 1 - 1 の結果（Result1）の信頼度が所定の閾値を超える場合には、Result1 が信頼できるとして、Result1 を出力して（図 6 - 2 の矢印 Y 1 1 参照）処理を終了する。ここで、前述したように、エッジ装置 30 における特徴抽出層 B f 2 - 1 は、サーバ装置 20 のベースモデルである Y0L0v3 の特徴抽出層 B f 2 - 1 と共通の構造を有するため、特徴抽出層 B f 2 - 1 から出力された特徴マップは、サーバ装置 20 の検知部 B d 2 - 1 においても共有が可能である。なお、サーバ装置 20 では、Y0L0v3 から特徴抽出層 B f 2 - 1 を削除した構成の Y0L0v3 を、D N N 2 として適用する。

[0042] そこで、エッジ装置 30 では、Result1 が閾値以下である場合には、Result1 が信頼できないとして、特徴抽出層 B f 2 - 1 の特徴マップをエッジ装置 30 に量子化後に出力する（図 6 - 2 の矢印 Y 1 2 参照）。そして、サーバ装置 20 では、この特徴マップを FP32 に戻して、特徴マップを、再構築後の Y0L0v3 の検知部 B d 2 - 1 に入力し、検知部 B d 2 - 1 の推論結果（Result2）を出力する（図 6 - 2 の矢印 Y 1 3 参照）。したがって、サーバ装置 20 の計算範囲は、検知部 B d 2 - 1 での計算範囲となる。言い換えると、サーバ装置 20 では、特徴抽出層 B f 2 - 1 での計算を省略できる。

[0043] [処理システムの処理手順]

図 7 は、実施の形態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 7 に示すように、まず、エッジ装置 30 において、推論部 31 は、画像の入力を受け付けると（ステップ S 1 ）、入力画像を D N N 1 に入力

する。DNN1では、特徴抽出層Bf2が、入力画像の特徴量を特徴マップとして抽出し（ステップS2）、判定部32に出力する（ステップS3）。DNN1では、検知層Bd1が、特徴マップを基に入力画像に対する推論処理、例えば、検知処理を実行し（ステップS4）、判定部32に出力する（ステップS5）。

[0044] 判定部32は、推論部31が推論結果に基づいて信頼度を計算し（ステップS6）、計算した信頼度と所定の閾値とを比較し信頼度が所定の閾値以下であるか否かを判定する（ステップS7）。

[0045] 信頼度が閾値以下でない場合（ステップS7：No）、すなわち、信頼度が閾値を超える場合、判定部32は、推論部31のDNN1が推論した推論結果を出力する（ステップS8）。

[0046] 一方、信頼度が閾値以下である場合（ステップS7：Yes）、判定部32は、特徴マップを量子化部33に出力し（ステップS9）、量子化部33は、特徴マップを量子化し（ステップS10）、サーバ装置20に送信する（ステップS11）。

[0047] サーバ装置20では、推論部21が、エッジ装置30から出力された、量子化された特徴マップを、FP32に戻して、DNN2の検知層Bd2の入力とする。検知層Bd2が、エッジ装置30から出力された特徴マップを基に、入力画像に対する推論処理、例えば、検知処理を実行する（ステップS12）。サーバ装置20は、DNN2の推論結果をエッジ装置30に送信し（ステップS13）、エッジ装置30から出力される（ステップS14）。なお、本実の形態では、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。

[0048] [評価実験 1]

エッジ装置 30 の DNN 1 の処理時間と、サーバ装置の DNN 2 の処理時間とを評価した。タスクは、分類であり、テストデータは、ImageNet（各クラス 300 枚、計 3000 画像）であり、HWは、Platform: NVIDIA GeForce RTX2070+AMD 3600である。図 8 は、処理システム 100 において一例として選定した各 DNN の処理時間を示す図である。

[0049] 図 8 において、エッジ装置の DNN 1 は、Residual block B f - 1 を保持する学習済みの特徴抽出層 B f 2 - 1 を、YOLOv2 の前段に配置し、訓練を行ったモデルである。サーバ装置 20 の DNN 2 は、学習済みの YOLOv3 である。比較のために、エッジ装置 30 のベースモデルである YOLOv2 についても処理時間を示す。

[0050] 図 8 に示すように、エッジ装置の DNN 1 は、YOLOv2 よりも特徴抽出層 B f 2 - 1 分、層を深くしたため、YOLOv2 よりも処理時間が長くなっているものの、YOLOv2 よりも推論精度の改善が見られた。そして、エッジ装置の DNN 1 は、サーバ装置 20 の DNN 2 である YOLOv3 よりも、画像 1 枚当たりの処理時間が約 2 倍速い。このように、エッジ装置の DNN 1 は、YOLOv3 の特徴抽出層 B f 2 - 1 を、YOLOv2 の特徴抽出層に代えて配置した場合であっても、推論精度を上げつつ、処理の高速化についても維持が可能である。

[0051] [評価実験 2]

エッジ装置 30 の DNN 1 を用いて、テストデータのエントロピーの分布を可視化した。タスク、テストデータ、HWは、図 8 と同じ条件である。図 9 は、DNN 1 の推論結果を基に求めたテストデータのエントロピーの分布図である。

[0052] エントロピーが高いほど、DNN 1 の推論結果が信頼できない。このため、モデルカスケードにおいてエッジ装置とサーバ装置とのいずれにおいて処理を実行するかを判定するための評価値として、信頼度を採用することは適切であることといえる。そして、ほとんどのデータのエントロピーが 0.5 以下であるため、エッジ装置とサーバ装置とのいずれにおいて処理を実行す

るかを判定する際に使用するエントロピーの閾値は、0.5を目安として設定すればよいと考えられる。

[0053] [評価実験3]

そこで、エントロピーの閾値（オフロード率）の変動に伴う推論結果の全体精度の変動を求める評価実験3を行った。なお、閾値は、オフロード率に連動し、オフロード率を下げる場合には閾値を上げる。この評価実験では、int4、int6、int8で、エッジ装置30の特徴抽出層B f 2によって抽出された特徴マップに対する量子化を行った。なお、タスク、テストデータ、HWは、図8と同じ条件である。図10は、オフロード率と全体精度との関係を示す図である。図10において、「Offload rate 0」は、全てのデータがエッジ装置30により処理され、量子化しない場合のオリジナル精度（acc_origin）が低い状態であり、「Offload rate 1」は、すべてのデータがサーバ装置20により処理され、量子化しない場合のオリジナル精度（acc_origin）が高い状態である。エッジ装置30の特徴抽出層B f 2による特徴マップに対しては、int4量子化、int6量子化、int8量子化を行っている。

[0054] このうち、図10に示すように、int4量子化、int6量子化、int8量子化のうち、エッジ装置30の特徴抽出層B f 2による特徴マップに対してint8量子化を行った場合には、精度をほぼ落とさずに、転送データ量を、量子化しない場合と比して75%削減できた。したがって、特徴マップの量子化には、int8量子化を適用することが望ましい。

[0055] なお、特徴抽出層B f 2が出力する特徴マップが、よりスパースになるようなコスト項を設けてもよい。その結果、int6及び／またはint4が図10におけるint8と近いオフロード率及び／または全体精度を示すようになった場合、int6またはint4のうち、int8と近いオフロード率及び／または全体精度を示す量子化ビットレートのうちより小さい方を選択すればよい。

[0056] また、量子化された特徴マップを圧縮してもよい。圧縮しようとする場合、特徴マップも量子化された特徴マップも対象の画像に含まれる、自然画像と同様の性質を有するため、HEVCやVVCなどの画像符号化方式を採用してもい

いし、ZIPなどの汎用的な圧縮方式を採用してもいい。なお、前述した自然画像と同様の性質とは、隣接する画素間の相関は高い場合が多い、というような画像符号化に取り入れられている一般的な性質を意図する。また、量子化を行う場合は逆量子化を、圧縮を行う場合は復号を行えばよい。量子化及び／または圧縮はエッジ装置で行ってもいいし物理的若しくは論理的に（サーバ装置よりも）エッジ装置の近くに配置された別の装置で行ってもいい。逆量子化及び／または復号化は、サーバ装置で行ってもいいし、物理的若しくは論理的に（エッジ装置よりも）サーバ装置の近くに配置された別の装置で行ってもよい。

[0057] また、オフロード率が0.4（エントロピー閾値が0.5）を超えると、オフロード率を上げて、すなわち、エントロピー閾値を下げて、精度の向上が少なくなっている。このため、閾値を0.5に設定すると、オフロード率と精度のバランスが取れるものと考えられる。このように、オフロード率と精度とのバランスに応じて閾値を設定することで、それぞれのユースケースに応じたオフロード率や全体精度の調整が可能になる。

[0058] ここで、エントロピーの閾値を決める理由についてさらに説明する。エントロピーの閾値を下げるということは、オフロード率が高くなるということの意味する。つまり、エントロピーの閾値を下げることで、サーバ装置側で推論するデータ数が増えるといいかえてもよい。

[0059] ここで、エッジ装置側で推論するかサーバ装置側で推論するかを決めるための閾値について考える。一例として、全体としての精度を基準とすることが考えられる。エッジ装置側で推論してもサーバ装置側で推論しても精度があまりかわらない程度であればエッジ装置側で処理でき、エッジ装置側で処理をしたときに精度が下がる場合にサーバ装置側で処理できるような閾値が望ましい。例えば図10の場合、量子化ビットレートを8(acc_int8)とした場合にオフロード率の閾値を0.4にすればよいことがわかる。そして、オフロード率が0.4となるようにエッジ側の推論結果のエントロピーの閾値を決定すればよい。図10の場合、エントロピー閾値は、0.5程度とすれば

よい。

[0060] [実施の形態の効果]

このように、実施の形態では、サーバ装置20に配置する前のDNN2'における特徴抽出層Bf2を、エッジ装置30のDNN1の特徴抽出層Bf2として配置することで、エッジ装置30が出力した特徴マップを、サーバ装置20においても共有することができる。すなわち、サーバ装置20のDNN2の検知層Bd2は、エッジ装置30のDNN1の特徴抽出層Bf2が出力した特徴マップを用いて推論処理を実行することが可能になる。

[0061] したがって、サーバ装置20側で推論を行う場合に特徴抽出処理の実行を省略できるため、システム全体の計算時間を短縮することができ、低遅延化を図ることが可能になる。また、エッジ装置30からサーバ装置20への出力データは、処理対象の画像ではなく、画像から抽出した特徴マップであるため、エッジ装置30からサーバ装置20に対するデータ転送量を低減することが可能になる。また、エッジ装置は、処理対象の画像や符号化された画像でも、周波数信号のような一般的に用いられる特徴量でもない特徴マップを、サーバ装置に送信しているため、第三者に対する秘密性の向上を図ることもできる。秘密性を確実に担保しようとする場合、学習を行う際に、さらに特徴マップと入力データ（対象の画像）との関係を非可逆とする制約を課せばよい。

[0062] 実施の形態では、サーバ装置20に配置する前のDNN2'における特徴抽出層Bf2を、エッジ装置30の検知層Bd1の前段に配置したDNN1を提案した。このDNN1については、実際に分類タスクを行い、精度及び速さともに軽量モデルとして有効であることを評価できた。

[0063] また、本実施の形態では、エッジ装置30とサーバ装置20とのいずれの推論結果を採用するかを判定する際に使用する信頼度を、DNN1の推論結果のエントロピーとして定式化し、閾値を設定することによって、実際の運用時におけるユースケースに応じたオフロード率や全体精度の調整を可能とした。

[0064] なお、本実施の形態では、処理対象データとして画像を例に説明したが、画像に限らず、各種センサが検知した検知結果でもよい。また、本実施の形態では、DNN1、DNN2のベースモデルとして、YOLOv2、YOLOv3を適用した場合を例に説明したが、DNN1、DNN2のベースモデルは、タスクに応じて適宜設定すればよい。

[0065] また、本実施の形態では、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。

[0066] [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。

[0067] 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。

[0068] [プログラム]

図11は、プログラムが実行されることにより、エッジ装置30及びサーバ装置20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。

[0069] 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 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。

[0070] 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及びサーバ装置２０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及びサーバ装置２０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。

[0071] 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

。

[0072] 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CPU1020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。

[0073] 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

- [0074] 20 サーバ装置
 21, 31 推論部
 30 エッジ装置
 32 判定部
 33 量子化部
 100 処理システム

請求の範囲

- [請求項1] 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、
- 、
- 前記エッジ装置は、
- 第1のモデルを用いて、処理対象データの特徴量を抽出し、抽出した特徴量を基に前記処理対象データに対する推論処理を実行する第1の推論部と、
- 前記第1の推論部による推論結果に対する信頼度が閾値を超えた場合には前記第1の推論部による推論結果を出力し、前記信頼度が閾値以下である場合には前記処理対象データの特徴量を前記サーバ装置に出力する判定部と、
- を有し、
- 前記サーバ装置は、
- 前記第1のモデルよりも推論精度が高い第2のモデルを用いて、前記エッジ装置から出力された前記処理対象データの特徴量を基に前記処理対象データに対する推論処理を実行する第2の推論部と、
- を有することを特徴とする処理システム。
- [請求項2] 前記第1のモデルは、
- 前記処理対象データの特徴量を抽出する第1の抽出部と、
- 前記処理対象データの特徴量を基に第1の推論処理を実行する第1の実行部と、
- を有し、
- 前記第2のモデルは、
- 前記処理対象データの特徴量を基に第2の推論処理を実行する第2の実行部
- を有し、
- 前記第2の実行部は、前記第1の抽出部によって抽出された前記処理対象データの特徴量を基に、前記第2の推論処理を実行することを

特徴とする請求項 1 に記載の処理システム。

[請求項3] 前記判定部は、量子化した前記処理対象データの特徴量を前記サーバ装置に出力することを特徴とする請求項 1 または 2 に記載の処理システム。

[請求項4] 前記信頼度は、前記第 1 の推論部による推論結果のエントロピーに基づくことを特徴とする請求項 1 ～ 3 のいずれか一つに記載の処理システム。

[請求項5] 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、

前記エッジ装置が、第 1 のモデルを用いて、処理対象データの特徴量を抽出し、抽出した特徴量を基に前記処理対象データに対する推論処理を実行する第 1 の推論工程と、

前記エッジ装置が、前記第 1 の推論工程における推論結果に対する信頼度が閾値を超えた場合には前記第 1 の推論工程における推論結果を出力し、前記信頼度が閾値以下である場合には前記処理対象データの特徴量を前記サーバ装置に出力する判定工程と、

前記サーバ装置が、前記第 1 のモデルよりも推論精度が高い第 2 のモデルを用いて、前記エッジ装置から出力された前記処理対象データの特徴量を基に前記処理対象データに対する推論処理を実行する第 2 の推論工程と、

を含んだことを特徴とする処理方法。

[請求項6] 方法をコンピュータに実行させる処理プログラムであって、

エッジ装置としてのコンピュータに、

第 1 のモデルを用いて、処理対象データの特徴量を抽出し、抽出した特徴量を基に前記処理対象データに対する推論処理を実行する第 1 の推論ステップと、

前記第 1 の推論ステップにおける推論結果に対する信頼度が閾値を超えた場合には前記第 1 の推論ステップにおける推論結果を出力し、

前記信頼度が閾値以下である場合には前記処理対象データの特徴量を出力する判定ステップと、

を実行させ、

サーバ装置としてのコンピュータに、

前記第 1 のモデルよりも推論制度が高い第 2 のモデルを用いて、前記エッジ装置から出力された前記処理対象データの特徴量を基に前記処理対象データに対する推論処理を実行する第 2 の推論ステップ、
を実行させることを特徴とする処理プログラム。