

(19) 日本国特許庁(JP)

(12) 公開特許公報(A)

(11) 特許出願公開番号

特開2020-71862

(P2020-71862A)

(43) 公開日 令和2年5月7日(2020.5.7)

(51) Int.Cl.		F I			テーマコード (参考)
G06T 7/00	(2017.01)	G06T 7/00	350C		5L096
G06T 7/10	(2017.01)	G06T 7/10			
G06N 3/04	(2006.01)	G06N 3/04			
G06N 3/08	(2006.01)	G06N 3/08			

審査請求 有 請求項の数 20 O L 外国語出願 (全 28 頁)

(21) 出願番号	特願2019-90037 (P2019-90037)	(71) 出願人	000003078
(22) 出願日	令和1年5月10日 (2019.5.10)		株式会社東芝
(31) 優先権主張番号	16/176,801		東京都港区芝浦一丁目1番1号
(32) 優先日	平成30年10月31日 (2018.10.31)	(74) 代理人	100108855
(33) 優先権主張国・地域又は機関	米国 (US)		弁理士 蔵田 昌俊
		(74) 代理人	100103034
			弁理士 野河 信久
特許法第30条第2項適用申請有り 1. 2018年5月11日 証明書記載のウェブサイトにて発表 2. 2018年9月3日-6日 (公表日 2018年9月4日) British Machine Vision Conference 2018にて発表 3. 2018年8月6日 http://bmvc2018.org/contents/papers/0286.pdf にて発表		(74) 代理人	100075672
			弁理士 峰 隆司
		(74) 代理人	100153051
			弁理士 河野 直樹
		(74) 代理人	100179062
			弁理士 井上 正
		(74) 代理人	100162570
			弁理士 金子 早苗

最終頁に続く

(54) 【発明の名称】 コンピュータビジョンシステム及び方法

(57) 【要約】 (修正有)

【課題】画像をセグメント化するための画像処理方法を提供する。

【解決手段】方法は、第1の画像から第1の画像より低い解像度表現である第2の画像を生成することと、第1の特徴マップを生成するために、第1の処理ステージを用いて第1の画像を処理することと、第2の特徴マップを生成するために、第2の処理ステージを用いて第2の画像を処理することと、第1の特徴マップを第2の特徴マップと組み合わせること、によりセマンティックセグメント化された画像を生成する。第1の処理ステージは、少なくとも1つの分離可能な畳み込みモジュールを備える第1のニューラルネットワークを備える。第2の処理ステージは、少なくとも1つの分離可能な畳み込みモジュールを備える第2のニューラルネットワークを備える。さらに、第1のニューラルネットワークにおける層の数は、第2のニューラルネットワークにおける層の数よりも少ない。

【選択図】 なし

【特許請求の範囲】

【請求項 1】

画像をセグメント化するための画像処理方法であって、

第 1 の画像を受け取ることと、

前記第 1 の画像から、前記第 1 の画像のより低い解像度表現である第 2 の画像を生成することと、

第 1 の特徴マップを生成するために、第 1 の処理ステージを用いて前記第 1 の画像を処理することと、

第 2 の特徴マップを生成するために、第 2 の処理ステージを用いて前記第 2 の画像を処理することと、

セマンティックセグメント化された画像を生成するために、前記第 1 の特徴マップを前記第 2 の特徴マップと組み合わせることと、

を備え、

前記第 1 の処理ステージは、分離可能な畳み込みを実施するように構成される少なくとも 1 つの分離可能な畳み込みモジュールを備える第 1 のニューラルネットワークを備え、前記第 2 の処理ステージは、分離可能な畳み込みを実施するように構成される少なくとも 1 つの分離可能な畳み込みモジュールを備える第 2 のニューラルネットワークを備え、前記第 1 のニューラルネットワークにおける層の数は、前記第 2 のニューラルネットワークにおける層の数よりも少ない、

画像処理方法。

【請求項 2】

前記第 1 の特徴マップ及び前記第 2 の特徴マップは、たった 1 つのステージで組み合わせられる、請求項 1 に記載の画像処理方法。

【請求項 3】

前記第 1 の特徴マップ及び前記第 2 の特徴マップは、加算によって組み合わせられる、請求項 2 に記載の画像処理方法。

【請求項 4】

前記第 1 の特徴マップ及び前記第 2 の特徴マップは、融合ステージにおいて組み合わせられ、前記融合ステージは、前記第 2 の特徴マップをアップサンプリングすることと、前記第 1 の特徴マップを前記第 2 の特徴マップに加算することと、を備え、加算することは、前記アップサンプリングされた第 2 の特徴マップ及び前記第 1 の特徴マップの対応する値を加算することを備える、請求項 2 に記載の画像処理方法。

【請求項 5】

アップサンプリングされた第 2 の特徴マップは、前記第 1 の特徴マップへの加算に先立って深さ方向畳み込みを実施され、前記深さ方向畳み込みは、1 よりも大きい拡張係数を用いて実施される、請求項 4 に記載の画像処理方法。

【請求項 6】

アップサンプリングされた第 2 の特徴マップは、1 よりも大きい拡張係数を用いて深さ方向畳み込みを実施されており、前記第 1 の特徴マップは、加算に先立って 2 次元畳み込みを実施される、請求項 5 に記載の画像処理方法。

【請求項 7】

前記第 2 の処理ステージにおける分離可能な畳み込みモジュールは、深さ方向畳み込みモジュールである、請求項 1 に記載の画像処理方法。

【請求項 8】

前記第 2 の処理ステージにおける分離可能な畳み込みモジュールは、ボトルネックアーキテクチャモジュールである、請求項 1 に記載の画像処理方法。

【請求項 9】

前記第 2 の処理ステージにおける分離可能な畳み込みモジュールは、ボトルネック残差アーキテクチャモジュールである、請求項 1 に記載の画像処理方法。

【請求項 10】

前記第 2 の画像は、前記分離可能な畳み込みモジュールによって処理されることに先立って標準畳み込みを実施される、請求項 1 に記載の画像処理方法。

【請求項 1 1】

複数の残差ボトルネックアーキテクチャモジュールが存在する、請求項 9 に記載の画像処理方法。

【請求項 1 2】

最終の標準畳み込みモジュールが、前記残差ボトルネックアーキテクチャモジュールの後に提供される、請求項 1 1 に記載の画像処理方法。

【請求項 1 3】

ピラミッドプーリングモジュールが、前記最終の標準畳み込みモジュールと前記残差ボトルネックアーキテクチャモジュールとの間に提供される、請求項 1 2 に記載の画像処理方法。

10

【請求項 1 4】

前記第 1 の処理ステージは、複数の深さ方向の分離可能な畳み込みモジュールを備える、請求項 1 に記載の画像処理方法。

【請求項 1 5】

画像をセグメント化するためのモデルを学習する方法であって、

前記モデルが、

分離可能な畳み込みを実施するように構成される少なくとも 1 つの分離可能な畳み込みモジュールを備える複数の層を有する第 1 のニューラルネットワークと、

20

分離可能な畳み込みを実施するように構成される少なくとも 1 つの分離可能な畳み込みモジュールを備える複数の層を有する第 2 のニューラルネットワークと、

を備え、

前記第 1 のニューラルネットワークにおける層の数は、前記第 2 のニューラルネットワークにおける層の数よりも少なく、

前記第 1 のニューラルネットワークは、第 1 の解像度で画像を処理するように構成され、前記第 2 のニューラルネットワークは、より低い解像度で同じ画像を処理するように構成され、

前記第 1 の及び第 2 のニューラルネットワークの特徴マップは、単一のステージにおいて加算によって組み合わせられ、

30

前記学習方法は、

画像と、前記画像に関するセマンティックセグメント化された情報と、を含む学習データを提供することと、

前記画像を入力として使用し、且つ前記セマンティックセグメント化された情報を出力として使用して、前記モデルを学習することと、

を備え、2 つのステージは一緒に学習される、

方法。

【請求項 1 6】

学習の間にフィルタの数を増加させることと、より重要性の低いフィルタを廃棄するためにそれらの数を低減することと、をさらに備える、請求項 1 5 に記載の方法。

40

【請求項 1 7】

前記第 2 の処理ステージにおいて第 2 の出力を加算するために学習の間に前記モデルを適合させることをさらに備え、前記方法は、前記画像を入力として使用して学習することと、前記出力と前記第 2 の出力との両方における前記セマンティックセグメント化された情報の両方を比較することによって損失を決定することと、両方の出力から前記決定された損失を使用することによって学習の間に重みを更新することと、をさらに備える、請求項 1 5 に記載の方法。

【請求項 1 8】

画像をセグメント化するための画像処理システムであって、

前記システムはインタフェース及びプロセッサを備え、

50

前記インタフェースは、画像入力を有し、第 1 の画像を受け取るように構成され、
前記プロセッサは、
前記第 1 の画像から、前記第 1 の画像のより低い解像度表現である第 2 の画像を生成することと、

第 1 の特徴マップを生成するために、第 1 の処理ステージを用いて前記第 1 の画像を処理することと、

第 2 の特徴マップを生成するために、第 2 の処理ステージを用いて前記第 2 の画像を処理することと、

セマンティックセグメント化された画像を生成するために、前記第 1 の特徴マップを前記第 2 の特徴マップと組み合わせることと、

を行うように構成され、

前記第 1 の処理ステージは、分離可能な畳み込みを実施するように構成される少なくとも 1 つの分離可能な畳み込みモジュールを備える第 1 のニューラルネットワークを備え、前記第 2 の処理ステージは、分離可能な畳み込みを実施するように構成される少なくとも 1 つの分離可能な畳み込みモジュールを備える第 2 のニューラルネットワークを備え、前記第 1 のニューラルネットワークにおける層の数は、前記第 2 のニューラルネットワークにおける層の数よりも少ない、

画像処理システム。

【請求項 19】

乗物のための検出システムであって、前記検出システムは、請求項 18 に記載の画像処理システムを備え、画像を受け取り、前記画像をセグメント化することによって前記画像からオブジェクトを決定するように構成される、検出システム。

【請求項 20】

請求項 1 に記載の方法をコンピュータに実行させるように適合されたコンピュータ読み取り可能な命令を搬送する、非一時的な搬送媒体。

【発明の詳細な説明】

【技術分野】

【0001】

実施形態は、コンピュータビジョンシステム及び方法に関する。

【背景技術】

【0002】

多くのコンピュータビジョンタスク、例えば、オブジェクト認識及び登録は、画像のセグメンテーションを必要とし、セグメンテーションでは画像の複数の領域がラベルを提供される。セマンティックセグメンテーションは、詳細なピクセルレベル基準を提供し、それは特に、障害物検出及び正確な境界検出を頻繁に必要とする用途に適している。そのような用途は、自律走行車及び運転者支援、組み込み及びウェアラブルデバイスを含むが、それらに限定されるわけではない。

【0003】

現代のセマンティックセグメンテーション方法は、極めて正確な結果を達成するが、それは多くの場合、効率の低下という犠牲を払う。畳み込みニューラルネットワーク（CNN）の最近の進歩は、これらのネットワークによって達成される結果に著しい改善を示す。しかしながら、それらの有効性は、モデルに関わるパラメータ及び演算の数に大きく依存する。現代のセマンティックセグメンテーション方法は、ハイエンドグラフィックプロセッシングユニット（GPU）で処理が実施されたとしても、単一の画像のオブジェクト分類を実施するために 1 秒より長い時間を必要とする。これらの方法の複雑性は、リアルタイムアプリケーションでのそれらの展開を妨げる。

【0004】

自律運転は、高精度のオブジェクト検出及び認識だけでなく、オブジェクト分類ニューラルネットワークのリアルタイム演算も必要とする、複雑なタスクである。故に、セマンティックセグメンテーションがオブジェクト分類の精度を損なうことなくリアルタイムで

10

20

30

40

50

実施される、セマンティックセグメンテーションのための新しいアプローチが必要とされる。

【図面の簡単な説明】

【0005】

【図1】図1は、乗物に実装される実施形態によるシステムである。

【図2】図2は、各サブネットワーク分岐の責務を特定する、図1のシステム中で使用されることができるニューラルネットワークの簡潔なフロー図である。

【図3】図3は、オブジェクト分類のためのニューラルネットワークの詳細なフロー図である。

【図4】図4は、一実施形態に従って使用されることができるネットワークアーキテクチャの概略図である。

【図5】図5は、ピラミッドプーリングをより詳細に示す概略図である。

【図6】図6は、オブジェクト分類のためにニューラルネットワークを学習するための例示的な処理のフロー図である。

【図7】図7は、パラメータプルーニング処理を示すフロー図である。

【図8】図8は、都市景観検証セットについての視覚的な比較を示し、第1の行が入力RGB画像、第2の行がグラウンドトゥルース、第3の行がコンテキスト分岐出力、第4の行が詳細分岐出力、最終行がコンテキスト分岐と詳細分岐との両方を使用した出力である。

【図9】図9は、一実施形態に従って使用されることができるデバイスの概略図である。

【発明を実施するための形態】

【0006】

一実施形態では、画像をセグメント化するための画像処理方法が提供され、方法は、

第1の画像を受け取ることと、

第1の画像から、第1の画像のより低い解像度表現である第2の画像を生成することと

、

第1の特徴マップを生成するために、第1の処理ステージを用いて第1の画像を処理することと、

第2の特徴マップを生成するために、第2の処理ステージを用いて第2の画像を処理することと、

セマンティックセグメント化された (semantic segmented) 画像を生成するために、第1の特徴マップを第2の特徴マップと組み合わせることと、

を備え、

第1の処理ステージは、分離可能な畳み込み (separable convolution) を実施するように構成される少なくとも1つの分離可能な畳み込みモジュールを備える第1のニューラルネットワークを備え、第2の処理ステージは、分離可能な畳み込みを実施するように構成される少なくとも1つの分離可能な畳み込みモジュールを備える第2のニューラルネットワークを備え、第1のニューラルネットワークにおける層の数は、第2のニューラルネットワークにおける層の数よりも少ない。

【0007】

開示されるシステムは、コンピュータ技術に関係する及びコンピューティングの分野で生ずる技術的課題、すなわち、リアルタイムのセマンティックセグメンテーションを提供するという技術的課題に対処するものである。開示されるシステムは、2つの分岐のニューラルネットワークアーキテクチャによってこの技術的問題を解決するものであり、ここでは低解像度分岐がコンテキストについて処理され、高解像度分岐が空間精度について処理される。各分岐における畳み込みネットワークブロックの特定の構成によって改善が提供される。

【0008】

このシステムの例示された実施形態は、ネットワークアーキテクチャを備え、それは、最大解像度のサブネットワーク分岐と低減された解像度のサブネットワーク分岐とを組み

10

20

30

40

50

合わせる。第2のサブネットワーク分岐は、より低い解像度の画像に関連するものであり、画像又はシーンを構成するオブジェクトを分類するか又はそうでなければ識別することを目的とする。さらに、より低い解像度の分岐は、シーン中に存在するオブジェクトにラベルを割り当てるために説明される。一方、最大解像度サブネットワーク分岐は、主として、高解像度セグメンテーション結果のためにオブジェクト境界を洗練する (refining) ことを担う。

【0009】

上記の実施形態では、リアルタイムパフォーマンスは、低解像度の深いコンテキスト分岐ネットワークを、高解像度のいくぶん浅い詳細分岐ネットワークと共に使用することによって得られる。したがって、ネットワークは、低解像度でコンテキストを知覚し、それを高解像度の結果で洗練する。

10

【0010】

上記の実施形態のアーキテクチャは、低解像度コンテキスト分岐が主に正確なオブジェクトラベルを提供することを担い、一方、高解像度詳細分岐がセグメンテーション境界を洗練することを保証するように設計される。

【0011】

上記の実施形態のモデルは、以下の理由により、計算コストを低減する。

i) コンテキスト処理においては、より複雑且つ抽象的な特徴を学習する必要があるため、より深いネットワークが必要とされる。しかしながら、より高い解像度の入力が必要とされないため、コンテキスト分岐についてより低い解像度の入力を使用することによってモデルコストが低減される。

20

ii) 境界処理においては、高解像度の入力が必要とされる。しかしながら、広視野を見る必要はないため、詳細分岐において層を少ししか使用しないことで、モデルコストが低減される。

iii) 故に、上述されたように、2つの異なる目的のための2つの分岐を動作させることによって、全体的なモデルコストが低減される。

【0012】

上記の実施形態は、以下を提供する。

低リソース (埋め込み) デバイスに限定されるわけではないが、それに適した2メガピクセル (2048 x 1024ピクセル) 画像の効率的なリアルタイムセマンティック画像セグメンテーションのための方法

30

計算コスト (浮遊小数点演算数、メモリ及び電力) を低減するための2つの分岐の深いニューラルネットワークを使用したセマンティック画像セグメンテーションのための方法
シーンの要点 (例えば、一般的オブジェクトに関する) 及び詳細 (境界に関する) が別々に学習される方法

【0013】

オブジェクト識別及び分類の実施形態では、システムは、ニューラルネットワークを使用した機械学習モジュールとして実現される。このシステムの実施形態では、フィードフォワードニューラルネットワークが使用される。システムは、高解像度画像の形式で入力を受け取り、モデルの内部の重みパラメータの値に従って出力を生成する。ニューラルネットワークの重みは、学習プロセスの間に決定され、ここでシステムは、多様なサンプル画像を提供され、ターゲット出力に対する生成された出力における誤差に従って重みが調整される。生成される出力とターゲット出力との間の誤差が最小に低減されるように、重みが調整される。

40

【0014】

一実施形態では、低解像度分岐と最大解像度分岐との両方が、ニューラルネットワークを使用して実現される。低解像度分岐における層の数は、高解像度分岐における層の数より多い。

【0015】

一実施形態では、上記のシステムは、1 x 1ポイント方向畳み込み層 (point-wise con

50

volution layer) が後に続く、空間又はチャンネル方向畳み込みとしても知られる深さ方向畳み込み (DWC on v : depth-wise convolution) へと標準畳み込み (Conv 2 d) を分解するために、深さ方向の分離可能な畳み込み (depth-wise separable convolution) を使用する。故にクロスチャンネル及び空間相関が独立して計算され、それはパラメータの数を大幅に低減し、より少ない浮遊小数点演算及び速い実行時間をもたらす。

【0016】

一実施形態では、第1の特徴マップ及び第2の特徴マップは、たった1つのステージで組み合わせられる。高解像度分岐の出力は、低解像度分岐と1回だけ組み合わせられる、すなわち、第1の特徴マップと第2の特徴マップとが、融合ユニットにおいて組み合わせられる。

10

【0017】

一実施形態では、第1の特徴マップ及び第2の特徴マップは、融合ステージにおいて組み合わせられ、融合ステージは、第2の特徴マップをアップサンプリングすることと、第1の特徴マップを第2の特徴マップに加算することと、を備え、加算することは、第1の特徴マップとアップサンプリングされた第2の特徴マップとの対応する値を加算することを備える。

【0018】

さらなる実施形態では、第1の特徴マップと第2の特徴マップとは加算によって組み合わせられる。一実施形態では、(アップサンプリング後の低解像度からの) サイズ $a \times b \times c$ の第2の特徴マップは、両方の数を加算することによって、サイズ $a \times b \times c$ の組み合わせられた特徴マップCを作り出すために、サイズ $a \times b \times c$ の第2の特徴マップと組み合わせられる。連結は必要とされず、これによりメモリフットプリントを小さいままにすることが可能となる。

20

【0019】

アップサンプリングされた第2の特徴マップは、第1の特徴マップへの加算に先立って、深さ方向畳み込みを実施され得、深さ方向畳み込みは、1よりも大きい拡張係数 (dilation factor) を用いて実施される。

【0020】

さらなる実施形態では、1よりも大きい拡張係数を用いて深さ方向畳み込みを実施されたアップサンプリングされた第2の特徴マップ及び第1の特徴マップが、加算に先立って2次元畳み込みを実施される。

30

【0021】

一実施形態では、第2の処理ステージにおける深さ方向畳み込みモジュールは、ボトルネックアーキテクチャモジュール (bottleneck architecture modules) である。ボトルネックアーキテクチャモジュールは、特徴拡張を実施し、その後に深さ方向畳み込みが続く、その後にポイント畳み込みが続く。故に、これらはまた、深さ方向の分離可能な畳み込みを実装すると考えられることができる。

【0022】

一実施形態では、第2の処理ステージにおける深さ方向畳み込みモジュールは、ボトルネック残差アーキテクチャモジュール (bottleneck residual architecture modules) である。複数の残差ボトルネックアーキテクチャモジュールが存在し得る。残差ボトルネックアーキテクチャモジュールからの出力チャンネルの数は、複数の残差ボトルネックアーキテクチャモジュールにおける後続のモジュールにおいて増加し得る。

40

【0023】

第2の画像は、深さ方向畳み込みモジュールによって処理されることに先だって、標準畳み込みを実施され得る。最終の標準畳み込みモジュールが、残差ボトルネックアーキテクチャモジュールの後に提供され得る。

【0024】

一実施形態では、最終の標準畳み込みモジュールと残差ボトルネックアーキテクチャモジュールとの間にピラミッドプーリングモジュール (pyramid pooling module) が提供さ

50

れる。

【0025】

第1の処理ステージは、複数の深さ方向の分離可能な畳み込みモジュールを備え得る。第1の画像は、深さ方向の分離可能な畳み込みモジュールによって処理することに先立って標準畳み込みを実施され得る。

【0026】

一実施形態では、画像をセグメント化するためのモデルを学習する方法が提供され、モデルが、

分離可能な畳み込みを実施するように構成される少なくとも1つの分離可能な畳み込みモジュールを備える複数の層を有する第1のニューラルネットワークと、

分離可能な畳み込みを実施するように構成される少なくとも1つの分離可能な畳み込みモジュールを備える、複数の層を有する第2のニューラルネットワークと、

を備え、

第1のニューラルネットワークにおける層の数は、第2のニューラルネットワークにおける層の数より少なく、

第1のニューラルネットワークは、第1の解像度で画像を処理するように構成され、第2のニューラルネットワークは、より低い解像度で同じ画像を処理するように構成され、

第1の及び第2のニューラルネットワークの特徴マップは、単一のステージにおいて加算によって組み合わせられ、

この学習方法は、

画像と、画像に関するセマンティックセグメント化された情報と、を含む学習データを提供することと、

画像を入力として使用し、且つセマンティックセグメント化された情報を出力として使用して、モデルを学習することと、

を備え、2つのステージは一緒に学習される。

【0027】

さらなる実施形態では、パラメータプルーニング(pruning)が、学習の間に実施され得る。

【0028】

さらなる実施形態では、学習方法は、第2の処理ステージにおいて第2の出力を加算するために学習の間にモデルを適合させることをさらに備え得、方法は、画像を入力として使用して学習することと、出力と第2の出力との両方におけるセマンティックセグメント化された情報の両方を比較することによって損失を決定することと、両方の出力からの決定された損失を使用することによって学習の間に重みを更新することと、をさらに備える。

【0029】

一実施形態では、画像をセグメント化するための画像処理システムが提供され、

システムはインタフェース及びプロセッサを備え、

前記インタフェースは、画像入力を有し、第1の画像を受け取るように構成され、

前記プロセッサは、

第1の画像から、第1の画像のより低い解像度表現である第2の画像を生成することと

、第1の特徴マップを生成するために、第1の処理ステージを用いて第1の画像を処理することと、

第2の特徴マップを生成するために、第2の処理ステージを用いて第2の画像を処理することと、

セマンティックセグメント化された画像を生成するために、第1の特徴マップを第2の特徴マップと組み合わせることと、

を行うように構成され、

第1の処理ステージは、深さ方向の分離可能な畳み込みを実施するように構成される少

10

20

30

40

50

なくとも1つのモジュールを備える第1のニューラルネットワークを備え、第2の処理ステージは、深さ方向畳み込みを実施するように構成される少なくとも1つのモジュールを備える第2のニューラルネットワークを備え、第1のニューラルネットワークにおける層の数は、第2のニューラルネットワークにおける層の数よりも少ない。

【0030】

一実施形態では、システムは、乗物に実装され、システムが実装されることができる乗物は、システムに入力高解像度画像を提供する1つの又は複数の静止デジタルカメラ及び又はビデオデジタルカメラを備える自律及び半自律走行車を含むが、それらに限定されるわけではない。一実施形態では、システムは、乗物に設置されたグラフィックスプロセッシングユニット又は中央処理装置において実現される。システムの目的は、乗物を取り囲むオブジェクトを分類し、その最終目的地に向けた乗物の安全なナビゲーションを容易にすることである。

10

【0031】

したがって、さらなる実施形態では、乗物のための検出システムが提供され、この検出システムは上述したような画像処理システムを備えるものであり、この検出システムは、画像を受け取り、画像をセグメント化することによって画像からオブジェクトを決定するように構成される。

【0032】

したがって、上記の実施形態は、以下を提供することができる。

自律システムのための効率的なリアルタイムセマンティック画像セグメンテーション

20

自律運転+運転者支援システム：低リソースデバイス上での高速セマンティック画像セグメンテーション

低電力デバイス、例えばNvidia Jetson TX2におけるリアルタイムセマンティック画像セグメンテーション

【0033】

さらなる実施形態では、システムは、コンピュータ上のコンピュータプログラムとして実施されることができ、ここで入力画像は、ソリッドステート記憶ユニットから取得されるか、通信プロトコルによってコンピュータに送信される。

【0034】

図1は、オブジェクト分類ネットワークの実施形態を例示するために使用され、それは、システムが乗物と共に移動するように、乗物のボード上に設けられている。図1は、車1の概略図を示し、ここで車1は衝突防止システムを搭載している。衝突防止システムは、4つのカメラ3、5、7、9を備えており、ここでカメラ3、5、7、9は車1の各角に設けられている。それらカメラの各々は、カメラが、観察可能な世界を広範囲にわたって見ることを可能にする広い視野（FOV：field of view）を有する。一実施形態では、各カメラ3、5、7、9は、非常に広角な魚眼レンズを備えてよく、それはワイドパノラマ画像を生成する。各カメラからの画像は、別々に生成されるが、前方カメラ3及び5のFOVは、オーバーラップし得、車1の前方にビューが遮られるエリアが無いようにする。後方カメラ7及び9のFOVもまたオーバーラップし得、車1の後ろにビューが遮られるエリアが無いようにする。

30

40

【0035】

システムの1つの実施形態では、各カメラ3、5、7、9からの画像は、単一の中央処理装置（CPU）又はGPUによって別々に処理される。システムのさらなる実施形態では、各カメラ3、5、7、9は、別々のCPU又はGPUを提供され、それは画像を処理して、処理された画像を車1の中央CPUに送る。

【0036】

図2は、識別及び分類のための、一実施形態によるシステムのフロー図を示す。

【0037】

ステップS201において、カメラ3、5、7、9は、シーンの画像をキャプチャし、シーンは、1つ又は複数のオブジェクトを備え得る。一実施形態では、シーンは、都市の

50

街道のシーンである。画像は、オブジェクト分類システムに入力され、ここでオブジェクトが結果として識別され、ニューラルネットワークによってラベルを提供されることになる。

【0038】

ステップ201の後、一実施形態では、画像の経路は、2つの独立したサブネットワーク分岐に枝分かれする。第1の分岐では、画像は、最大解像度で提供され、ステップS207において、ニューラルネットワークは、シーンのローカルコンテキストを抽出する。第2の分岐は、グローバルコンテキストを抽出することステップS205を担うものであり、画像のダウンサンプリングされた表現、ステップS203を備える。

【0039】

図2では、シーン中のオブジェクトに、適したラベルでラベル付けするための第2のサブネットワーク分岐が説明された。実施形態に従って、どのようにこれが達成されるかが、図3のステップS303～S309を参照して説明されることになる。

【0040】

ステップS303において、画像寸法である長さ h 及び幅 w が n 分の1に縮小される。故に、画像は、低解像度サブネットワーク分岐においては n^2 分の1にダウンサンプリングされる。システムの計算時間及びオブジェクト分類の達成される正確さは、両方とも係数 n の値に反比例する。

【0041】

一実施形態では、係数 n は、2と8との間の範囲にある。 n の値が大きければ、低解像度分岐についてより低い解像度がもたらされ、演算数を効率的に低減し、且つより速い分類を達成する。しかしながら、高速演算は、正確さの低減という代償を払う。反対に、 n の値がより小さければ、低解像度分岐についてより高い解像度がもたらされ、オブジェクト分類の正確さを効率的に増加させるが、処理時間の増加を代償とする。

【0042】

入力画像は、3次元マトリックスフォーマットで表されることができる。各ピクセルは、3つのチャンネルを備え、ここで各チャンネルは、3つの色、つまり赤、緑及び青(RGB)のうちの1つの強度に関連する数値を有する。標準の計算では、2D畳み込みは、 $h \times w \times c$ フィルタを用いて使用されることができ、ここで h は高さであり、 w は幅であり、 c はチャンネルである。 $h \times w \times 3$ (RGBの画像に対する畳み込み)は、 $h \times w \times 1$ (DW畳み込み)と同様なので、ここでは標準畳み込みで構わない。留意されたい点は、通常、特徴チャンネルの数は、CNNで急激に増加するので、 $h \times w \times 3 \times 2 / 64 / 128$ は、特異ではないことである。一実施形態では、深さ方向畳み込みは、チャンネルの数の増加を伴って使用される。

【0043】

ステップS305において、畳み込みニューラルサブネットワークの第1の層は、2次元プーリング畳み込み層である。プーリング層は、さらなるダウンサンプリング演算を実施し、ここで、出力された特徴マップの寸法は、層のストライドに対する係数分の1に縮小される。層のストライドは、畳み込みカーネルが、それが画像にわたって走査されるときにどれだけ動くかということであると考えられることができる。例えば、ストライド1(1ピクセル)のプーリング畳み込み層は、特徴マップの空間次元に影響を与えないことになり、一方でストライド2のプーリング畳み込み層は、それらの寸法を2分の1にダウンサンプリングすることになる。より小さい特徴マップは、より速い推論をもたらし得るが、予測精度を犠牲にする。

【0044】

この実施形態では、第1の畳み込み層は、複数のフィルタ又はカーネルを入力ボリュームに適用する。各フィルタがシーンと比較され、2次元特徴マップが生成される。特徴マップは、フィルタとシーン中の特徴又はオブジェクトとの間の一致の結果として生じる活性化の空間配列を表す。全ての特徴マップは、深さ寸法に沿って積み重ねられ、出力ボリュームを生成する。例えば、32個の出力チャンネルから成る畳み込みニューラルネット

10

20

30

40

50

ワークは、32個のフィルタを入力ボリュームに適用し、深さ寸法32の出力ボリュームをレンダリングするために積み重ねられる32個の特徴マップを生成する。

【0045】

一実施形態では、オブジェクト分類は、いくつかの深さ方向の分離可能な畳み込み層を利用するニューラルネットワークを用いて実施される。深さ方向畳み込みでは、2次元畳み込みが各入力チャンネルに対して別々に実施され、各入力チャンネルについての特徴マップが生成される。全ての特徴マップは、畳み込みプロシージャの終わりに再度一緒に積み重ねられる。例えば、カメラ3、5、7、9は、ニューラルネットワークにカラーデジタル画像を提供し、ここで、各ピクセルは、1つは赤、1つは緑、及び1つは青の、3つの値のセットを備える。画像が深さ方向の分離可能な畳み込み層で処理されるとき、2次元畳み込みが色ごとに別々に実施される。

10

【0046】

一実施形態では、深さ方向畳み込み層はボトルネックブロック内で使用され得る。ボトルネック残差ブロックは、入力を、cから、表1に従った非一次関数f、畳み込みタイプカーネルサイズ／ストライドs、拡張係数(expansion factor) t、幅w、高さhを有するc'チャンネルへと移す(transfer)。

【0047】

ボトルネックブロックでは、入力ボリュームが拡張され、次いでそれは深さ方向畳み込み層及びそれに続くポイント方向畳み込みでフィルタされる。

【0048】

20

一般に、最初にポイント方向畳み込みが適用される(行1)。次いで深さ方向畳み込み及びポイント方向畳み込み(行2及び3)が適用される。

【0049】

一実施形態では、以下の処理が成される。

Conv2d 1/1

(1×1×c×t×c'パラメータ)

(h×w×1×1×c×t×c'演算)

Conv2d 3/s

(3×3×t×c×c'パラメータ)

(h/s×w/s×3×3×t×c×c'演算)

30

【0050】

しかしながら、代替の実施形態では、以下が使用され得る。

Conv2d 1/1

(1×1×c×t×c'パラメータ)

(h×w×1×1×c×t×c'演算)

DWConv3/s

(3×3×1×t×c'パラメータ)

(h/s×w/s×3×3×1×t×c'演算)

40

Conv2d 1/1

(1×1×t×c×c'パラメータ)

(h/s×w/s×1×1×t×c×c'演算)

【0051】

上記の代替的な実施形態では、必要とされる計算がより少なく、パラメータがより少ない。

【0052】

より高解像度で特徴を学習すること及びそれらをより低い解像度に投影することは、特徴学習プロシージャに利益をもたらす。さらに、ボトルネックブロックにおいて深さ方向

50

畳み込みを利用することは、計算効率を提供し、メモリフットプリントを著しく低減する。

【0053】

さらなる実施形態では、残差接続 (residual connections) がボトルネックブロックに組み込まれ得、それはボトルネック残差ブロックを形成する。残差接続は、入力及び出力ボリュームが同じ空間次元及び特徴マップの数を有する場合にのみ、ボトルネックブロック内で使用され得る。残差接続は、ボトルネックブロックの入力からその出力に接続される線形層を表す。追加の接続層により、乗算層 (multiplier layers) にわたるより効果的な勾配伝搬が可能となり、ニューラルネットワークの学習が改善される。

【0054】

一実施形態では、オブジェクト分類システムのためにステップ S 307 で使用されるボトルネック残差ブロックは、表 1 に示される構造を有する。第 1 の層は、係数 t で特徴マップの数を増加させることによって入力ボリュームの寸法を拡張する標準 2 次元畳み込み層である。

【0055】

【表 1】

表1

入力	演算子	出力
$h \times w \times c$	<i>Conv2d 1/1, f</i>	$h \times w \times tc$
$h \times w \times tc$	<i>DWConv 3/s, f</i>	$\frac{h}{s} \times \frac{w}{s} \times tc$
$\frac{h}{s} \times \frac{w}{s} \times tc$	<i>Conv2d 1/1, -</i>	$\frac{h}{s} \times \frac{w}{s} \times c'$

【0056】

ボトルネック残差ブロックの第 2 の層では、深さ方向畳み込みフィルタがシーンから特徴を抽出するために使用される。深さ方向畳み込みは、ストライド s を有し、故に、特徴マップの出力寸法は、 s 分の 1 に空間的に低減されることになる。ストライドが 1 である畳み込み層では、特徴マップの空間サイズは影響を受けない。空間サイズの低減は、後続のネットワーク層における計算の数を効率的に低減する。畳み込みは各深さにおいて別々に計算されるので、*DWconv* は標準畳み込みを著しく改善する。空間サイズの低減は、後続のネットワーク層における (パラメータではなく) 計算の数を効率的に低減する。

【0057】

深さ方向畳み込みと、それに続くポイント方向畳み込みは、標準畳み込み層と比較して、オブジェクト検出の正確さを著しく低減しない。情報は、最初にチャンネル方向レベルで計算され、チャンネル方向情報を組み合わせる、よりコストのかからない標準畳み込みが、その後続く。畳み込みが使用するのはカーネル 1×1 のみであるので、それはより速く、必要とするパラメータはより少ない。

【0058】

最後に、ボトルネック残差ブロックでは、深さ方向の分離可能な畳み込み層によって生成される出力ボリュームは、2 次元畳み込みの第 2 の層を使用して、その元の低次元表現へと戻るように投影され得る。第 1 の及び第 2 の 2 次元畳み込みについて使用される拡張係数 t は、2 つの層について同じである。

【0059】

一実施形態では、ピラミッドプーリングモジュール (図 4 には示されていない) が、最終の畳み込み層である S 309 の前に実装される。

【0060】

ピラミッドプーリングモジュール (図示せず) は、より細かいレベルからより粗いレベルへと複数の部分に画像を区分し、それらにおける局所特徴を集約する。空間ピラミッド

10

20

30

40

50

プーリングは、試験用の任意のサイズの画像／ウィンドウから表現を生成することを可能にし、学習の間に様々なサイズ又はスケールの画像を供給することを可能にする。

【0061】

ピラミッドプーリングモジュールは、図5により詳細に示される。ここで、それは4つのカーネルから成り、ここで各カーネルの解像度は、それぞれ 32×64 、 16×32 、 8×16 及び 4×8 である。カーネルは、ステップS307の特徴マップ出力にわたって等しく分配される。結果的な4つの特徴マップは、バイリニアアップサンプリングされ（bilinearly up-sampled）、まとめて加算されて単一の特徴マップ出力を生成する。特徴マップが加算されることに留意することが重要である。これにより、メモリに全ての解像度を記憶する必要を避けられる。

10

【0062】

一実施形態では、より低い解像度のサブネットワーク分岐の最後の層は、標準2次元畳み込み層、ステップS309である。より低い解像度分岐の終わりにおいて、特徴マップの小さな空間次元は、標準畳み込みの使用を効率的で計算コストが低いものとする。

【0063】

表2は、一実施形態に従って使用され得る層の詳細を示す。

【0064】

【表2】

入力	演算子	拡張係数	出力チャンネル	繰返し	ストライド
$256 \times 512 \times 3$	Conv2d	-	32	1	2
$128 \times 256 \times 32$	ボトルネック	1	32	1	1
$128 \times 256 \times 32$	ボトルネック	6	32	1	1
$128 \times 256 \times 32$	ボトルネック	6	48	3	2
$64 \times 128 \times 48$	ボトルネック	6	64	3	2
$32 \times 64 \times 64$	ボトルネック	6	96	2	1
$32 \times 64 \times 96$	ボトルネック	6	128	2	1
$32 \times 64 \times 128$	Conv2d	-	128	1	1

20

【0065】

これはまた図4にイラスト入りで示されている。図4では、上方の分岐は、画像501をより低い解像度で処理する。第1の層503は、図3のS305を参照して説明されたストライド2の標準2次元畳み込み層である。

30

【0066】

次に、第1の畳み込み層の出力特徴は、12個のボトルネック残差層505、507、509、511、513、515、517、519、521、523、525及び527を通過する。

【0067】

当業者に認識されることになるように、ボトルネック残差演算子は、実際には、表1を参照して説明されたように複数の層によって実施される。しかしながら、簡潔な表現として図4ではそれらは単一の層／演算子として示されている。各ボトルネックで使用される拡張係数は、表2に示される。

40

【0068】

第1のボトルネック残差層505は、32個の入力チャンネル、32個の出力チャンネル、ストライド1及び拡張係数（t）1を有する。

【0069】

第2のボトルネック残差層507は、32個の入力チャンネル、32個の出力チャンネル、ストライド1及び拡張係数（t）6を有する。

【0070】

第3の及び第4のボトルネック残差層509、511は、32個の入力チャンネル、3

50

2 個の出力チャンネル、ストライド 1 及び拡張係数 (t) 6 を有する。

【0071】

第 5 のボトルネック残差層 5 1 3 は、3 2 個の入力チャンネル、4 8 個の出力チャンネル、ストライド 2 及び拡張係数 (t) 6 を有する。

【0072】

第 6 の及び第 7 のボトルネック残差層 5 1 5、5 1 7 は、6 4 個の入力チャンネル、6 4 個の出力チャンネル、ストライド 1 及び拡張係数 (t) 6 を有する。

【0073】

第 8 のボトルネック残差層 5 1 9 は、4 8 個の入力チャンネル、6 4 個の出力チャンネル、ストライド 2 及び拡張係数 (t) 6 を有する。

10

【0074】

第 9 のボトルネック残差層 5 2 1 は、6 4 個の入力チャンネル、6 4 個の出力チャンネル、ストライド 1 及び拡張係数 (t) 6 を有する。

【0075】

第 10 のボトルネック残差層 5 2 3 は、6 4 個の入力チャンネル、9 6 個の出力チャンネル、ストライド 1 及び拡張係数 (t) 6 を有する。

【0076】

第 11 のボトルネック残差層 5 2 5 は、9 6 個の入力チャンネル、9 6 個の出力チャンネル、ストライド 1 及び拡張係数 (t) 6 を有する。

【0077】

20

第 12 のボトルネック残差層 5 2 7 は、9 6 個の入力チャンネル、1 2 8 個の出力チャンネル、ストライド 1 及び拡張係数 (t) 6 を有する。

【0078】

より低い解像度分岐の最後の層は、標準 2 次元畳み込み 5 2 9 であり、それは図 3 の S 3 0 9 を参照して説明されることができ、1 2 8 個の入力チャンネル、1 2 8 個の出力チャンネル、及びストライド 1 を有する。

【0079】

したがって、各ボトルネックについて 3 つの層が使用されるので、グローバルコンテキストを記述するためには 3 8 個の層のみが使用される。

【0080】

30

ここに説明される例では、1 2 × ボトルネックモジュール及び 2 × 畳み込みモジュールの構造が効率及び表現度 (正確さ) を兼ね備える。

【0081】

図 2 では、第 1 のサブネットワーク分岐が主として空間詳細を識別するために使用される。これが如何にして達成されるかが、これより図 3 及び図 4 に例示されるステップ S 3 1 1 及び S 3 1 3 を参照して説明される。

【0082】

一実施形態では、オブジェクト分類ネットワークの第 1 のサブネットワーク分岐は、最大解像度のデジタル入力画像を受け取る。ステップ S 3 1 1 において、標準 2 次元畳み込み層が、既に上述したようなより効率的なオブジェクト分類を達成するために、分岐の初めにおいて再度使用される。ネットワークのこの層は、入力ボリュームをダウンサンプリングし、ストライドに従った低減された空間次元を有する出力マップを生成するプーリング畳み込み層である。最大解像度分岐における畳み込み層のパラメータが、表 3 に提示される。

40

【0083】

【表 3】

入力	演算子	出力チャンネル	繰返し	ストライド
1048 x 2048 x 3	Conv2d	32	1	2
512 x 1024 x 32	DWConv	32	1	2
256 x 512 x 32	Conv2d (1x1)	64	1	1
256 x 512 x 64	DWConv	64	1	2
128 x 256 x 64	Conv2d (1x1)	128	1	1
128 x 256 x 128	DWConv	128	1	1
64 x 128 x 128	Conv2d (1x1)	128	1	1

【0084】

一実施形態では、プーリング畳み込み層653は、32個のフィルタを高解像度画像651に適用し、32個の特徴マップを生成する。ステップ313では、3×3のカーネルサイズの深さ方向の分離可能な畳み込みの3つの層が使用される。

【0085】

深さ方向の分離可能な畳み込みブロックは、深さ方向畳み込み層と、それに続く1×1のポイント方向畳み込み層とを備える。図4では、各深さ方向の分離可能な畳み込みの2つの層が単一のブロックとして表されている。

【0086】

第1の深さ方向の分離可能な畳み込みブロック655は、32個の入力チャンネル、64個の出力チャンネル及びストライド2を有する。第2の深さ方向の分離可能な畳み込みブロック657は、64個の入力チャンネル、128個の出力チャンネル、及びストライド2を有する。第3の深さ方向の分離可能な畳み込み659は、128個の入力チャンネル、128個の出力チャンネル、及びストライド1を有する。

【0087】

より多くの数の特徴マップにより、2次元畳み込みと比較して、深さ方向の分離可能な畳み込みの使用がシステムにとってより有益となる。

【0088】

一実施形態では、オブジェクト分類システムの最大解像度の分岐は、主として、より低い解像度の分岐によって抽出されるグローバルコンテキストを洗練することを担う。より高い精度及びより良いオブジェクト分類結果を達成するために、最大解像度の分岐における深さ方向の分離可能な畳み込みブロックは直接的に用いられる。深さ方向の分離可能な畳み込みのボトルネック実装は多くの数の層から成るネットワーク分岐を好み、なおここでは演算の速度が、達成される分類精度よりも重要であり、一方で、直接のアプローチは、動作速度が低減される代償を払ってより高いオブジェクト分類精度を提供する。しかしながら、最大解像度の分岐におけるより少ない数のニューラルネットワーク層は、より多くの数の演算を補償する。

【0089】

一実施形態では、単一の特徴融合ユニットが、図3のS315を参照して説明されるように、2つのサブネットワーク分岐の出力ボリュームを結合するために使用される、ステップS661。特徴マップを結合する処理は、システムメモリに特徴を保持することを含む。単一の特徴融合ユニットは、低リソースデバイス要件に準拠するより効率的な設計を構成する (accounts for)。

【0090】

一実施形態では、2つの分岐の出力ボリュームが同じ数の特徴マップを備える一方で、2つの分岐の終わりにおける特徴マップは空間次元が異なる。より低い解像度の分岐（分岐4）の特徴マップの空間次元は、最大解像度の分岐（分岐1）の特徴マップの空間次元より小さい。故に、より低い解像度のサブネットワーク分岐の出力は、アップサンプリング層によって処理され、ここで特徴マップは、4倍にアップスケーリングされる、表4。

【0091】

【表4】

分岐 -1	分岐 -4
-	アップサンプル x 4
-	DWConv (拡張 4)
Conv2d 1/1 -	3/1, f
	Conv2d 1/1, -
加算, f	

【0092】

さらなる実施形態では、アップサンプリング層の後に、1とは異なる拡張係数を有する深さ方向畳み込み層が続く。拡張畳み込み層は、特徴マップ上のオブジェクト間の間隔を増加させる。深さ方向畳み込みを拡張することは、カーネルのサイズのみに影響を与え、特にカーネルのサイズが指数関数的に増加する。例えば、拡張係数1の深さ方向畳み込みは、元のカーネルサイズ3×3を有することになり、一方拡張係数2の畳み込みは、カーネルサイズ7×7を有することになり、拡張係数4の畳み込みは、カーネルサイズ15×15を有することになる。拡張がカーネルサイズ7×7のものであっても、その計算数は9のみであることに留意されたい。

【0093】

出力ボリュームの合計 (Summation) は、ボリュームが同じ空間次元及び同じ数の特徴マップを備える場合にのみ実施されることができる。故に、2次元畳み込みの層が、2つの分岐の出力を、それらが一緒に加算される前に一般化するために使用される。これらの畳み込み層は、2つの分岐の出力ボリュームが異なる数の特徴マップを備える場合に追加的な対策を提供する。

【0094】

最終ステップにおいて、特徴マップは単に直接的に一緒に加算される。したがって、メモリに記憶される必要のあるパラメータの数は増加しない。

【0095】

一実施形態では、特徴融合ユニットの後には、図3のステップS317を参照して説明されるような、ソフトマックス畳み込み層、ステップ663が続き、それは各クラスラベルについての確立スコアを生じさせる。

【0096】

図6は、オブジェクト分類ニューラルネットワークを学習するための例示的な処理のフローチャートを例示し、それは、入力画像に示されるオブジェクトをラベル付けすることができる。ニューラルネットワークは、多くの学習サンプルを処理することと、全てのサンプルについて、ニューラルネットワークによって生成される出力と学習サンプルに定め

られたターゲット出力との間の誤差に従って各パラメータの重みを調整することによって、学習されることができる。一旦学習されると、ニューラルネットワークは、図1のニューラルネットワークシステムを例とするシステムにおいて展開されることができる。学習プロシーダは、1つ又は複数のコンピュータによって実行されることができる。

【0097】

ステップS401において、学習システムは学習データのセットを取得する。各データセットは、学習画像とターゲット出力とを備える。入力画像は、1つ又は複数のオブジェクトを示す画像の表現である。例えば、画像は、自律又は半自律走行車の近くに位置するオブジェクトを含み得る。学習データセットによって表される学習画像は、互いに異なり、同様のオブジェクトを含んでも含まなくてもよい。

10

【0098】

学習データセットは、有限数のシーンを備え、それらはニューラルネットワークを学習するために使用されることができる。一実施形態では、標準データオーグメンテーション技法が、サンプル画像数を増やすために使用され得る。オーグメンテーション技法は、0.5~2の範囲のランダムスケール係数、水平フリップ、多様な色合い、彩度、輝度及びコントラストを含むが、それらに限定されるわけではない。

【0099】

学習データセットの学習ターゲット出力は、ニューラルネットワークによって生成されるべきであるオブジェクト分類ネットワークの所望の出力を表す。ターゲット出力は、ニューラルネットワークの実際の出力と比較され、ターゲット出力と生成される出力との間の誤差が低減されるように重み付けパラメータが調整される。ステップS403において、ニューラルネットワークは、内部のパラメータの現在の値を使用してサンプル入力画像を処理し、出力画像を生成する。

20

【0100】

ステップS405において、ニューラルネットワークの予測された出力が学習データセットのターゲット出力部分と比較され、予測における誤差が推定される。引き続いて、ステップS407において、各内部パラメータの重みが、予測された出力とターゲット出力との間の誤差が最小に低減されるように調整される。

【0101】

ステップ409において、ニューラルネットワークは学習データの異なるセットを提供され、学習は、学習プロシーダを繰り返してニューラルネットワークの内部パラメータを調整するためにステップS403に戻って、予測された出力とターゲット出力とにおけるより小さな誤差が達成されるようにする。

30

【0102】

図2から図5を参照して説明されたように、アーキテクチャには2つの分岐がある。

【0103】

さらなる実施形態では、学習の間、重み付けされた補助的な損失が低解像度分岐について使用される。このような方法で損失に重み付けすることは、セマンティックセグメンテーションについての意味のある特徴がグローバルコンテキストについての分岐によって抽出され、もう一方の分岐からは独立して学習されることを確実にする。一実施形態では、補助的な損失の重みは、0.4に設定される。一実施形態では、これは、学習の間に単に1つの出力を有するのではないことによって達成され、追加的な出力が529の後に生成される（すなわち、新しいソフトマックス層がこのステージにおいて導入されて、その分岐から分岐する）。ソフトマックス層の出力が評価される（タスクはまたセグメンテーションである）。両方の層の出力が、重みを更新するために使用される。

40

【0104】

ニューラルネットワークの学習プロセスの一実施形態では、ネットワークブルーニングが実施される。学習の最初のステージにおいて、特徴マップの数が倍にされ、学習は上述したプロシーダを使用して行われる。パラメータの数は、元のサイズの1.75、1.5、1.25、及び1倍に漸進的に減少し、ここで学習プロシーダは、パラメータの各

50

低減の後に繰り返される。ネットワークプルーニングは、同じパフォーマンスを保ちながら、ネットワーク中のパラメータの数を効率的に低減する。これは、ニューラルネットワークアーキテクチャで使用されない特徴を削除することによって達成される。さらに、ネットワークプルーニングは、必要とされる特徴のみがネットワークによって学習されるので、より効率的な学習を可能にする。

【0105】

したがって、プルーニングは通常、パラメータを低減するために使用される。しかしながら、ここに説明される実施形態では、パラメータの数は受け入れ可能なので、プルーニングは、パフォーマンスを向上させるために実施される。したがって、ネットワークの表現 (expressiveness) は倍になる (パラメータの数を倍にする)。するとネットワークは、より多くのフィルタが存在する (数を倍にする) ことで、格段により遅くなる。しかしながら、元のフィルタ数に再度達するためにプルーニングが使用される。ここで、プルーニングは、フィルタの数を増加させ (より多くのフィルタを可能にし)、そして良好なものの選択する。

【0106】

したがって、ニューラルネットワークの学習プロセスのこの実施形態では、フィルタの数が学習プロシーダの初めにおいて倍にされる。これをどのように行うかが図7を参照して説明される。ステップS801において、各層についてのフィルタ数が倍にされる。ステップS803において、フィルタのチャンネル数が、前の層の出力のチャンネル数に一致させられる。例えば、各層が入力サイズ $h \times w$ 及び出力サイズ $h' \times w'$ を有するとする。全てのこれらの層は、使用されるフィルタの数に関する深さ/チャンネルを有する。すなわち $h \times w \times c$ を所与とすると、(例えば) サイズ 3×3 の c' のフィルタは、 $h' \times w' \times c'$ を生成するために使用されることができる。ステップS801にあるように、前の層のフィルタの数が倍にされると、 $h \times w \times 2c$ の入力が存在し、サイズ $3 \times 3 \times 2c$ の $2c'$ のフィルタが使用される。

【0107】

ステップS805において、ネットワークは、図6に関して説明されたように学習される。ステップS807において、容量に達するかどうか、すなわち、フィルタのターゲット数に達したかどうか決定される。

【0108】

ステップS809において、(画像入力を有する) 第1の層中の最弱のフィルタが識別され、削除される。これらのフィルタが識別され得、多くの異なる測定基準、1つの実施形態では、

【数1】

$$l_1$$

合計が使用される。例えば、上記の寸法を使用して、今度は $h \times w \times 1.5c$ が入力である。次の層のフィルタは、サイズ $3 \times 3 \times 2c$ であり、したがって、ステップS811において、前の層の削除されたフィルタに関連する重みは削除されてサイズ $3 \times 3 \times 1.5c$ のフィルタを得る。すると、ステップS813において、現在の層における最弱のフィルタが決定され、処理が続行する。

【0109】

ステップS815において、ステップS811及びS813におけるフィルタの削除によって影響を受けるさらなる層が存在するかどうかチェックされる。存在する場合、処理ループはステップS811に戻る。全ての最弱のフィルタ及び重みが一旦削除されてしまうと、処理は学習ステップS807に戻る。ネットワークが一旦学習されると、フィルタの数がさらに低減されることになるかがチェックされる (例えばここでは、フィルタの数は $2x$ から $1.75x$ へ、そして $1.5x$ へ、そして $1.25x$ へ、そして $1x$ へ低減される)。さらなる低減が必要な場合、処理は、さらにフィルタを削除するためにステップS809に移る。そうでなければ、処理は終了し、ネットワークは学習されてプルーニングされる。

10

20

30

40

50

【0110】

一実施形態では、バッチ正規化が学習の間に全ての層の後に使用される。バッチ正規化により、ニューラルネットワークの各層が他の層からより独立して学習することが可能となる。バッチ正規化は、前の層の特徴マップにおける活性化をスケールリングする。ニューラルネットワーク全体を通じた全ての活性化は所与の範囲内にあるので、高すぎる値や低すぎる値に関連する活性化は存在しない。これにより、より高い学習レート及び改善された特徴学習が可能となる。

【0111】

一実施形態では、ニューラルネットワークノードのセットのドロップアウトがソフトマックス層の前に実施される。ドロップアウトは、ネットワーク中のニューロン間の相互依存学習を低減する学習アプローチである。学習の間に、ネットワークの低減されたバージョンが作り出されるように、ノードのランダムなセットがネットワークからドロップアウトされる。ネットワークの低減されたバージョンは、ニューラルネットワークの他のセクションとは独立して学習するので、ニューロンが互いに共依存関係を発展させることを防ぐ。

【0112】

上記を明示するために、上記についてのアブレーション調査が都市景観データセット [M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele. The cityscapes dataset for semantic urban scene understanding. In CVPR, 2016]、及び都市景観ベンチマークに対するそのパフォーマンス報告を使用して成された。

【0113】

実験は、Nvidia Titan X (Maxwell, 3072 CUDAコア)、CUDA 8.0及びcuDNN V6を有するワークステーションで実施された。ReLU6が、低精度の計算で使ったときのそのロバスト性に起因して非線形関数として使用された。学習の間、バッチ正規化が全ての層において使用され、ドロップアウトがソフトマックス層の前だけに使用される。推論の間、バッチ正規化のパラメータは、親層のバイアス及び重みと結合される。深さ方向畳み込み層では、

【数2】

$$\ell_2$$

正則化が不要であることが分かった。

【0114】

ラベル付けされた学習データは限定されていたので、標準データオーグメンテーション技法が全ての実験に適用された：ランダムスケール0.5～2、水平フリップ、多様な色合い、彩度、輝度及びコントラスト。

【0115】

ContextNetのモデルは、割引係数0.9、運動量0.9、及び1に等しいイプシロンパラメータを有するRMSPropを使用して、TensorFlowで学習される。さらに、ポリ学習レート (a poly learning rate) がベースレート0.045及びパワー0.98で適用された。エポックの最大数は、1000に設定され、事前学習は使

【0116】

結果は、mIoU (mean intersection-over-union) として報告され、ランタイムは、GPUからCPUへのメモリ交換、カーネル実行、及び連続的なCPUからGPUへのメモリ転送を伴う単スレッドのCPUを考慮する。

【0117】

都市景観は、ドイツの50の異なる都市からの街路シーンにおける画像の多様なセットを含むセマンティックセグメンテーションのための大規模なデータセットである。合計で、それは25,000の注釈付けられた1024×2048px画像から構成され、そのうちの5,000のものは、高ピクセル精度のラベルを有し、20,000は、弱く注釈

付けられている。ここに提示される実験では、都市景観評価サーバで評価されることが
できる1, 525の試験画像、500の画像の検証セット、及び2, 975の画像の学習セ
ットである5, 000の画像のみが高ラベル品質で使用された。事前学習は用いられな
かった。

【0118】

アブレーション調査が実施され、ここでアブレーション調査重みは、もっぱら都市景観
学習セットから学習されたものであり、検証セットについてのパフォーマンスが報告され
る。特に、低解像度分岐の異なる解像度係数に対する影響が報告され、複数の分岐、及び
パラメータの数の修正を取り入れる。最後に、2つの分岐が詳細に分析される。

【0119】

入力画像解像度は、計算時間についての最も肝心な係数である。この例では、低解像度
分岐は、都市景観画像について256 512 pxである4分の1サイズの画像を取る（
cn14と示される）。代替的に、半分（cn12）又は8分の1（cn18）の解像度
が使用されることができる。

【0120】

表4は、異なるオプションが如何に結果に影響を与えるかを示す。全体的に、より深い
コンテキスト分岐におけるより大きな解像度は、より良いセグメンテーション結果を生成
する。しかしながら、これらの改善は実行時間を犠牲にして成り立つ。表4では、C o n
t e x t N e t（cn14）は、都市景観検証セットについて、低解像度分岐における半
分の解像度（cn12）及び8分の1の解像度（cn18）を有する、並びに、4分の1
、半分、及び最大解像度（cn124）の複数のレベルを有する、そのバージョンと比
較された。より小さいメモリフットプリントを有する実装もまた示される（cn14-5
00及びcn14-160）。

【0121】

表4

【表5】

	cn18	cn14	cn12	cn124	cn14-500	cn14-160
精度(%表示のmIoU)	60.1	65.9	68.7	67.3	62.1	57.7
fps(Frames per Second)	21.0	18.3	11.4	7.6	20.6	22.0
パラメータの数(単位は100万)	0.85	0.85	0.85	0.93	0.49	0.16

【0122】

表5は、IoUをより詳細に列挙する。ここで、ContextNet（cn14）の
詳細なIoUが、半分（cn12）及び8分の1（cn18）解像度を有するバージョン
、並びにそのマルチレベルバージョン（cn124）と比較された。小さなサイズのク
ラス（フェンス、ポール、信号機、交通標識）、細かい詳細を有するクラス（人、ライダ
ー、オートバイ、自転車）、及びサンプルがほとんど無いクラス（トラック、バス、電車
）は、高解像度により恩恵を受ける。

【0123】

【表6】

	road	sidewalk	building	wall	fence	pole	traffic light	traffic sign	vegetation	terrain	sky	person	rider	car	truck	bus	train	motorcycle	bicycle
cn18	96.2	72.4	87.4	45.9	44.1	32.1	38.4	54.3	87.9	54.1	91.0	62.3	36.1	88.3	57.6	59.1	34.2	41.2	58.4
cn14	96.8	76.6	88.6	46.4	49.7	38.0	45.3	60.5	89.0	59.3	91.4	67.5	41.7	90.0	63.5	71.7	57.1	41.5	64.6
cn12	97.2	78.9	89.2	47.2	54.4	39.5	55.3	63.8	89.4	59.8	91.5	70.2	47.5	91.1	70.2	76.2	63.7	51.36	67.8
cn124	97.4	79.6	89.5	44.1	49.8	45.5	50.6	64.6	90.2	59.4	93.4	70.9	43.1	91.8	65.2	71.9	64.5	41.95	66.1

【0124】

小さなサイズクラス（すなわちフェンス、ポール、信号機及び交通標識）、細かい詳細

を有するクラス（すなわち人、ライダー、オートバイ及び自転車）及び希少なクラス（すなわちバス、電車及びトラック）の精度は、グローバルコンテキスト分岐における増加した解像度から恩恵を受ける。他のクラスは、しばしばより大きなサイズであり、故により低い解像度でキャプチャされることができる。c n 1 4 は、1 8 . 3 f p s (frames per seconds) 及び m I o U 6 5 . 9 % で極めてバランスが取れていると結論する。

【0 1 2 5】

一実施形態では、2つの分岐のみが上で説明された。しかしながら、複数の解像度における分岐が用いられることができる。前の結果に加えて、表4及び表5はまた、半分及び最大解像度（c n 1 2 4）における2つの浅い分岐を有するC o n t e x t N e t の1つのバージョンも含む。c n 1 4 と比較して、c n 1 2 4 は1 . 4 % 精度を改善する。

10

【0 1 2 6】

しかしながら実行時間は、1 8 . 3 f p s から7 . 6 f p s に2倍以上低減される。その上、半分の解像度における深いサブネットワークを有するc n 1 2 は、精度及び速度の観点からc n 1 2 4 よりパフォーマンスが優れている。これらの結果は、より高解像度において深いサブネットワークを使用することは、より低い解像度において使用するのとは比べてより有益であることを示す。さらに、層の数は、精度には良い影響を与え、実行時間には負の影響を与える。

【0 1 2 7】

したがって、上記の例では、2つの分岐は、リアルタイムで期待できる結果を示す。実行時間はさておき、メモリフットプリントは、埋め込み型デバイス上での実施のために重要な考慮すべき事項である。表4は、c n 1 4 - 5 0 0 及びc n 1 4 - 1 6 0 と示される、大幅に低減された数のパラメータを有する2つのバージョンのC o n t e x t N e t を含む。意外にも、都市景観検証セットについて、ただ1 5 9 , 3 8 7 及び4 9 0 , 4 5 9 のパラメータを有するC o n t e x t N e t がそれぞれ5 7 . 7 % 及び6 2 . 1 % の m I o U を達成した。これらの結果は、所与のシステムのメモリ要件及び計算時間に適合するためにC o n t e x t N e t 設計は十分に柔軟であることを示す。

20

【0 1 2 8】

独立して学習された詳細分岐及び4分の1の解像度を有するグローバルコンテキストが、それぞれ2 5 . 1 % 及び6 3 . 3 % の m I o U を得た。コンテキスト分岐は、小さなサイズクラスについてうまく機能しないと分かり、詳細分岐の受容野は、妥当なパフォーマンスのために十分に大きくなかった。コンテキスト及び詳細ネットワークの出力（ソフトマックス）は、ネットワークのアンサンブルを作り出すために平均化された、なおそれらは独立して学習される。両方の分岐のアンサンブルは、6 0 . 3 % の m I o U を得、それはc n 1 4 より6 . 6 % 少なく、コンテキスト分岐のみを使用するよりも3 % 少ない。これは、C o n t e x t N e t アーキテクチャがマルチスケールの特徴融合のためのより良い選択肢であることにさらなる証拠を与える。

30

【0 1 2 9】

さらなる実施形態では、詳細分岐又はコンテキスト分岐のいずれかの出力が、それらの個々の寄与を観察するためにゼロに設定された。結果を図6に示す。図6は、都市景観検証セットについての視覚的な比較を示し、第1の行が入力RGB画像、第2の行がグラウンドトゥルース（ground truths）、第3の行がコンテキスト分岐出力、第4の行が詳細分岐出力、最終行がコンテキスト及び詳細分岐の両方を使用したC o n t e x t N e t 出力である。C o n t e x t N e t は6 5 . 9 % の m I o U を得て、一方で、独立して学区週された詳細分岐及び4分の1の解像度を有するグローバルコンテキストは、それぞれ2 5 . 1 % 及び6 3 . 3 % の m I o U を得た。

40

【0 1 3 0】

第1の列では、グローバルコンテキスト分岐はより大きなオブジェクト（例えば、空又は木）を正確に検出するが、境界付近や細い領域（thin regions）ではうまくいかないことが分かる。対照的に、詳細分岐は、オブジェクト境界を正確に検出するが、オブジェクトの中央領域ではうまくいかない。1つの例外は、木の中央領域であり、それは、恐らく

50

木のテクスチャの特徴的な性質に起因して正確に分類される。最後に、ContextNetは、両方の情報を正確に組み合わせて最終行において改善された結果を生成することが分かる。同様に、中央の列において、歩行者のセグメンテーションが、グローバルコンテキスト分岐出力にわたってContextNetによって洗練されることが分かる。最終列において、例えば詳細分岐がボール及び交通標識を検出するとしても、ContextNetは、グローバルコンテキスト分岐でこれらのうちのいくつかを効率的に組み合わせることができない。全体的に、コンテキスト分岐及び詳細分岐は、相補的な情報を学習することが分かる。

【0131】

上記の実施形態からの結果は以下の通りである。

10

【0132】

【表7】

表6

	クラスmIoU(%表示)	カテゴリmIoU(%表示)	パラメータ (単位は100万)
ContextNet	66.1	82.7	00.85

【0133】

異なる解像度を有する画像についての実行時間が以下に示される。

20

【0134】

【表8】

表7

	1024x2048	512x1024	256x512
ContextNet	18.3	65.5	139.2
ContextNet†	23.2	84.9	150.1

【0135】

‘†’が付いた測定値は、もっぱらカーネル実行時間を示す。

30

【0136】

上記の図は、メモリ転送を含む完全なCPU及びGPUパイプラインを含む。並列メモリ転送及びカーネル実行が用いられる場合、実行時間は23.2fpsに改善する。最後に、表7において、ContextNetは、より小さい入力解像度サイズについてスケーラビリティが高く、故に、目の前のタスク及び利用可能なリソースについて調整されることができると見てとれる。ContextNetの結果は、定性分析について図8に表示されている。ContextNetは、遠距離の小さなオブジェクトさえも適切にセグメント化することができる。

【0137】

ネットワークプルーニングは、モデルの精度を向上させるために、「より無駄なくスリムな(skinier)」ネットワークを取得するためにより広いネットワーク(より多くの数の特徴マップ)をプルーニングするために使用される。最初に、ネットワークは、改善された結果を取得するために特徴マップの2倍の(ターゲット)数で学習される。パラメータの数は、次いで元のサイズの1.5、1.25及び1倍へとプルーニングすることによって漸進的に減少される。一実施形態では、フィルタは、最も低い

40

【数3】

$$\ell_1$$

合計でプルーニングされた。プルーニングによって、mIoUは都市景観試験セットについて64.2%から66.1%へと改善した。

【0138】

50

上記の実施形態は、低解像度における、しかし大きな受容野を有する深いネットワーク分岐を、浅い、故に効率的な最大解像度の分岐と組み合わせてセグメンテーション詳細を向上させる、リアルタイムセマンティックセグメンテーションモデルに関する。上記のアーキテクチャはさらに、最大メモリ及び実行時間効率のためにボトルネック残差ブロック及び深さ方向畳み込みを広く活用する。

【0139】

上記の調査は、アーキテクチャが、優位性のある結果を達成するためにグローバルコンテキストとローカルコンテキストとを効率的に組み合わせることを示す。また、モデルプルーニング（リアルタイムパフォーマンス及びネットワークサイズの所与のターゲットを達成するための）は、セグメンテーション精度の改善をもたらす。単一画像深さ予測のよ

10

【0140】

図9は、実施形態による方法を実施するために使用されることができハードウェアの概略図である。これは、単に1つの例であり、他の構成が使用されることができに留意されたい。

【0141】

ハードウェアは、計算セクション700を備える。この特定の例では、このセクションのコンポーネントは、一緒に説明される。しかしながら、それらは必ずしも同一場所に配置されないことは認識されることになる。

20

【0142】

コンピューティングシステム700のコンポーネントは、処理ユニット713（中央処理装置CPUのような）、システムメモリ701、システムメモリ701を含む様々なシステムコンポーネントを処理ユニット713に結合するシステムバス711に限定されるわけではないが、それらを含み得る。システムバス711は、多様なバスアーキテクチャのうちの任意のものを使用した局所バス及び周辺機器バス、メモリバス又はメモリコントローラ等を含む、いくつかのタイプのバス構造のうちの任意のものであり得る。計算セクション700はまた、バス711に接続された外部のメモリ715も含む。

【0143】

システムメモリ701は、読み取り専用メモリのような揮発性／又は不揮発性メモリの形式のコンピュータ記憶媒体を含む。例えばセットアップ中にコンピュータ内の複数の要素間で情報を転送するのに役立つルーチンを含む基本入力出力システム（BIOS）703が典型的にシステムメモリ701に記憶される。加えて、システムメモリは、CPU713によって使用されているプログラムデータ709、アプリケーションプログラム707、及びオペレーティングシステム705を含む。

30

【0144】

また、インタフェース725がバス711に接続される。インタフェースは、さらなるデバイスから情報を受け取るための、コンピュータシステムのためのネットワークインタフェースであり得る。インタフェースはまた、例えばユーザがある特定のコマンドに応答することを可能にするユーザインタフェースであり得る。

40

【0145】

この例では、ビデオインタフェース717が提供される。ビデオインタフェース717は、グラフィックス処理メモリ721に接続されているグラフィックスプロセッシングユニット719を備える。

【0146】

グラフィックスプロセッシングユニット（GPU）719は特に、この複数の並列なコアの動作に起因して、上で説明された方法によく適している。故に、一実施形態では、処理はCPU713とGPU719との間で分割され得る。

【0147】

一実施形態では、GPUはNVIDIA Jetson TX2であり、それは低電力

50

G P Uチップである。

【0148】

一実施形態では、各カメラ（図1参照）にリンクされた専用のコンピューティングデバイス700が提供される。図2から図4に関して上述したアーキテクチャは、浮遊小数点計算が必要となることを避け、故に、コンピューティングデバイスは、乗物上のカメラと組み合わせられる例のような低電力ロケーションによく適している。

【0149】

上で説明されたアーキテクチャはまた、G P Uを使用した携帯電話にも適している。

【0150】

いくつかの実施形態を説明したが、これらの実施形態は、例として提示したものであり、発明の範囲を限定することは意図していない。これら新規な実施形態は、その他の様々な形態で実施されることが可能であり、発明の要旨を逸脱しない範囲で、種々の省略、置き換え、変更を行うことができる。これら実施形態やその変形は、発明の範囲や要旨に含まれるとともに、特許請求の範囲に記載された発明とその均等の範囲に含まれる。

10

【図1】

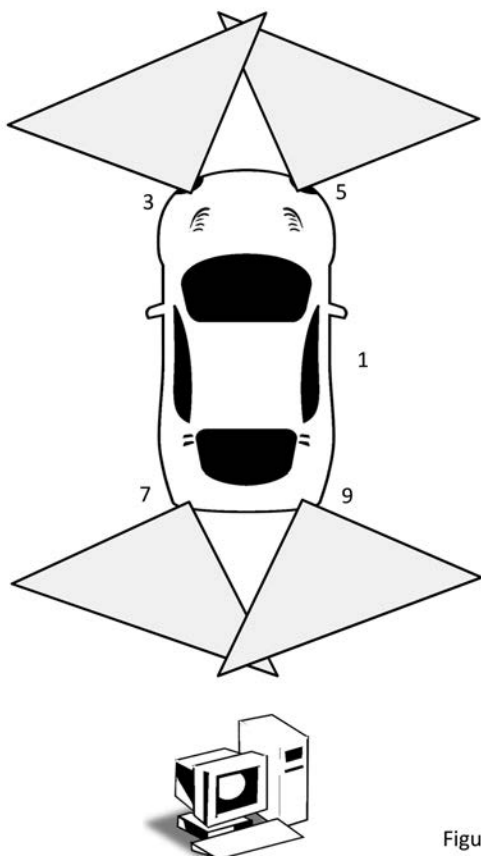


Figure 1

【図2】

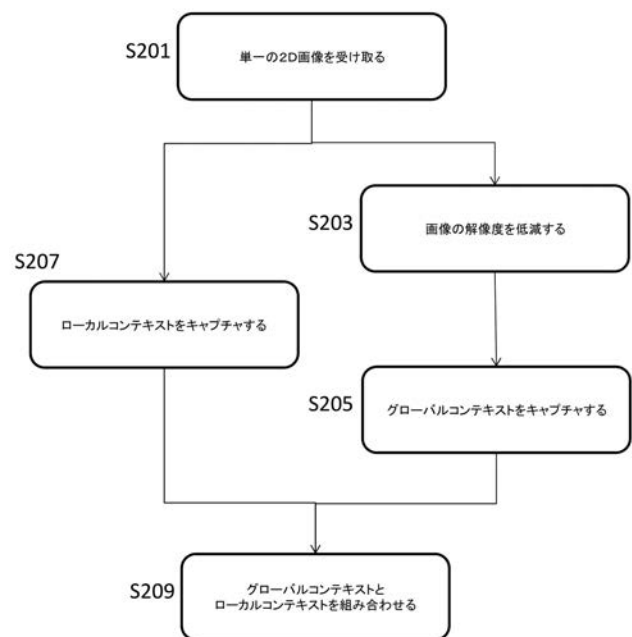


Figure 2

【 図 3 】

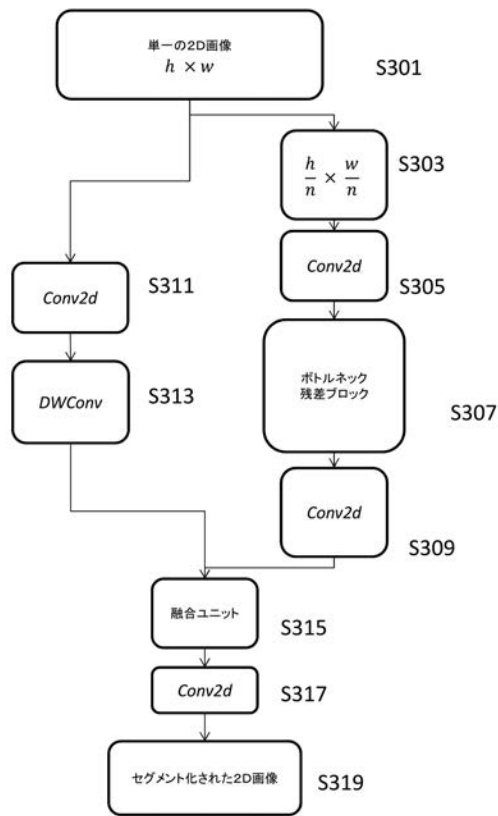


Figure 3

【 図 4 】

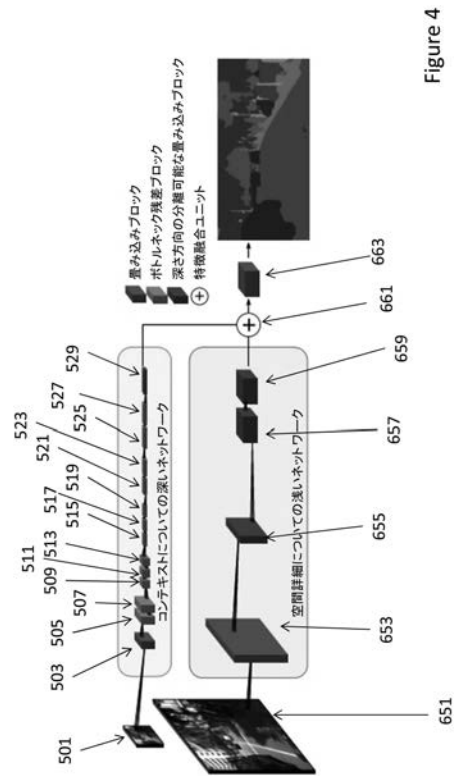


Figure 4

【 図 5 】

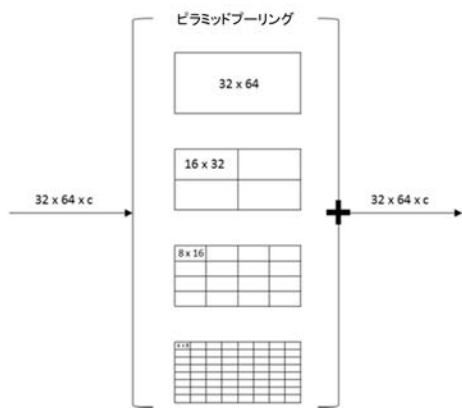


Figure 5

【 図 6 】

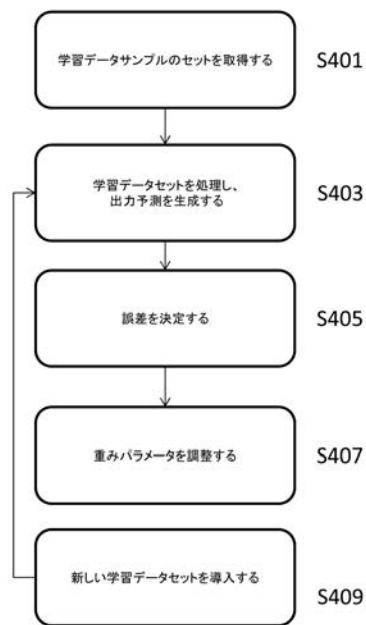


Figure 6

【図 7】

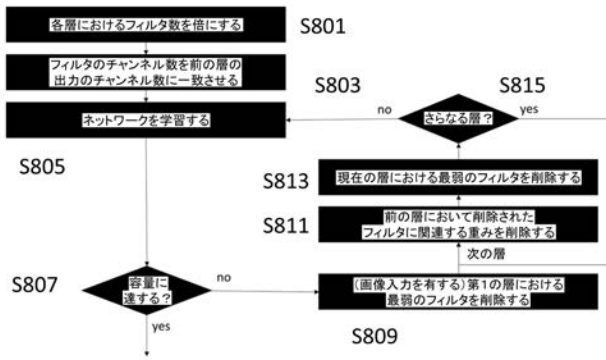


Figure 7

【図 8】

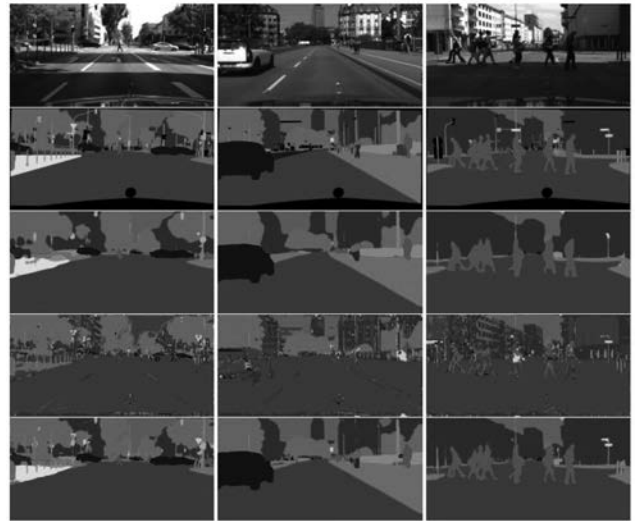


Figure 8

【図 9】

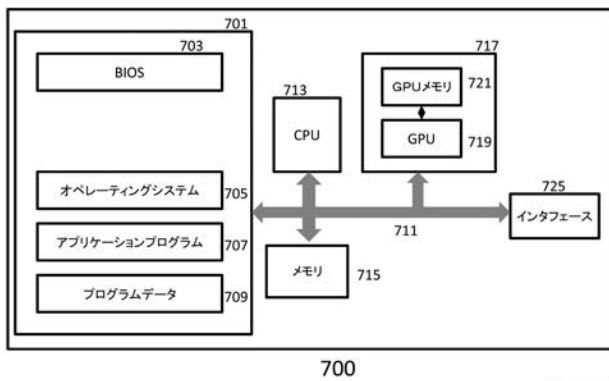


Figure 9

フロントページの続き

- (72)発明者 ルドラ プラサド パウデル カルマタ
イギリス国、 シービー４・０ジーゼット、 ケンブリッジシャー、 ケンブリッジ、 ミルトン
・ロード、 ケンブリッジ・サイエンス・パーク ２０８、 トーシバ・リサーチ・ヨーロッパ・
リミテッド、 ケンブリッジ・リサーチ・ラボラトリー内
- (72)発明者 ウジュワル ボンデ
イギリス国、 シービー４・０ジーゼット、 ケンブリッジシャー、 ケンブリッジ、 ミルトン
・ロード、 ケンブリッジ・サイエンス・パーク ２０８、 トーシバ・リサーチ・ヨーロッパ・
リミテッド、 ケンブリッジ・リサーチ・ラボラトリー内
- (72)発明者 ステファン リウィッキ
イギリス国、 シービー４・０ジーゼット、 ケンブリッジシャー、 ケンブリッジ、 ミルトン
・ロード、 ケンブリッジ・サイエンス・パーク ２０８、 トーシバ・リサーチ・ヨーロッパ・
リミテッド、 ケンブリッジ・リサーチ・ラボラトリー内
- (72)発明者 クリストファー ザッハ
イギリス国、 シービー４・０ジーゼット、 ケンブリッジシャー、 ケンブリッジ、 ミルトン
・ロード、 ケンブリッジ・サイエンス・パーク ２０８、 トーシバ・リサーチ・ヨーロッパ・
リミテッド、 ケンブリッジ・リサーチ・ラボラトリー内

Fターム(参考) 5L096 CA01 FA01 GA55 HA11 KA04 KA15

【外国語明細書】
2020071862000001.pdf