クラウド型画像解析エンジンの動的分割に関する研究

TP17005 猪子 弘康　　指導教授：藤吉 弘亘

1. はじめに

少子高齢化や人手不足が進む社会で，生活支援ロボットの社会導入が求められている．生活支援ロボットは，物体認識や物体検出，音声認識，行動計画など多くのタスクを実時間で行う必要がある．一方で，大きさ，消費電力，コストなどの観点から搭載可能な計算機の性能が低いという問題がある．この問題を解決するアプローチとして，クラウドロボティクスが提案されている．クラウドロボティクスは，計算コストの高い処理をクラウドサーバで行うもので，ロボットに高性能な計算機が不要となる．画像解析エンジンにクラウドロボティクスを導入した研究の一つとして，クラウド型顔画像解析エンジン [1] が提案されている．これは顔画像解析を行う Convolutional Neural Network (CNN) をロボット側とクラウド側で 2 つに分割して処理を行うことで，ロボット側の計算コストを低減する．また，ロボット側で撮影した画像を直接送るのではなく，中間層から得られる特徴マップを送信するため，プライバシーの配慮が可能となる．しかし，CNN の処理をどこまでロボット側で行い，どこからクラウドサーバ側で行うかを示す分割層を事前に決定する必要がある．そこで本研究では，ロボットまたはクラウドサーバの負荷や通信状況に応じて動的に分割層を決定する手法を提案する．また，提案手法を物体検出問題に適用し，その有効性を示す．

2. CNN の各処理時間の調査

本節では，クラウド型画像解析エンジンを物体検出問題に適用するため，物体検出を行う CNN の各処理時間を調査する．物体検出を行う YOLO v2[2] は，畳み込み層およびプーリング層が 27 層ある．YOLO v2 の各層を分割層とした場合の処理時間および通信時間を図 1 に示す．入力層に近い層で分割すると，通信時間が長くなる傾向がある．これは CNN に入力する画像の解像度が高く，ロボット側の CNN から出力される特徴マップのデータ量が大きいためである．例えば，分割層 1 で出力される特徴マップは $416 \times 416 \times 32$ 次元であり，データ量は 21.125MB になる．そのため通信時には，特徴マップを圧縮し，データ量を削減することが望ましい．

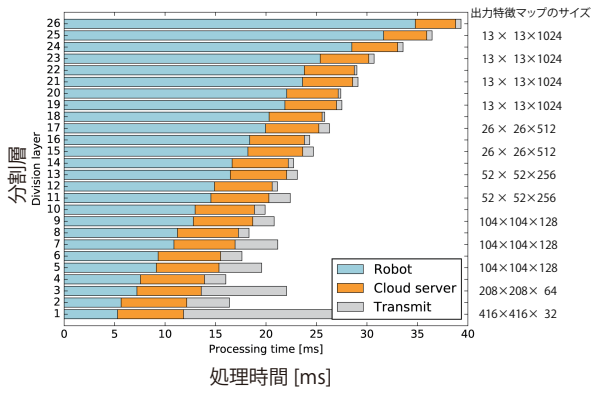

図 1：YOLO の各処理時間

3. クラウド型画像解析エンジンの動的分割

本研究では，物体検出タスクを対象とするクラウド型画像解析エンジンの動的分割法を提案する．システムの構成を図 2 に示す．提案手法では，中間層から得られる特徴マップを量子化して通信量を削減する．分割層は，ネットワーク帯域や，CNN の性能，クラウドサーバの能力に応じて動的に決定する．量子化を行うことで，特徴マップの分解能が低下し，CNN の性能に影響する恐れがある．そこで，要求レイテンシと要求精度を制約として導入し，クラウドサーバの計算量を最小化する．

3.1 特徴マップの量子化による圧縮

物体検出を行うネットワークでは，入力する画像サイズが大きいため，ロボット側の CNN から出力される特徴マップが大きく，通信時間が長くなる傾向がある．そのため，特徴マップを量子化により圧縮し，通信時間を短縮する．量子化には，線形量子化と非線形量子化がある．

線形量子化　線形量子化は，入力サンプルを等間隔の量子化ステップ幅 Δd で近似するものである．Q は量子化ビット数で，Q を大きくすると，量子化ステップ幅が小さくなり，特徴マップの分解能が高くなるが，データ量も大きくなる．一方で，Q を小さくすると，量子化ステップ幅が広くなり，特徴マップの分解能が下がるが，データ量は小さくなる．ここで，特徴マップ M における線形量子化の量子化ステップ幅 d は，式 (1) を用いて決める．

$$\Delta d = \frac{\max(M) - \min(M)}{2^Q - 1} \tag{1}$$

非線形量子化　非線形量子化は，対数や任意の確率密度分布で変化させた量子化ステップ幅で特徴マップを近似するものである．特徴マップの分解能が要求される区間は量子化ステップ幅を狭め，分解能が要求されない区間は量子化ステップ幅を広く取る．これにより，同じ量子化ビット数の線形量子化よりも量子化誤差が少なくなることが期待できる．本研究では，非線形量子化の量子化ステップ幅の決定に正規分布を用いる．式 (2) に示す μ は，量子化ステップ幅を狭くする特徴マップの値を表しており，$\mu = 0$ のときは 0 付近の値の量子化ステップ幅を狭くすることを表している．同様に，$\mu = 1.5$ のときは 1.5 付近，$\mu = -1.5$ のときは-1.5 付近の量子化ステップ幅を狭くする．

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} \tag{2}$$

次に，情報量 I を式 (3) により求める．

$$I = -\log f(x) = \frac{1}{2}\log 2\pi\sigma^2 + \frac{(\log e)(x-\mu)^2}{2\sigma^2} \tag{3}$$

求めた情報量 I から量子化ステップ幅を決定する．このようにして μ を制御することにより，特徴マップの特定の値付近の分解能を上げることができる．特徴マップはチャネルごとにスケールが異なるため，μ を特徴マップの最大最小値を用いて，

$$\mu = \mu_n \{\max(M) - \min(M)\} + \min(M) \tag{4}$$

のように表す．$\mu_n = 0$ 付近のとき特徴マップの最小値付近の分解能が高くなり，$\mu_n = 1$ 付近のとき特徴マップの最大値付近の分解能が高くなる．

3.2 分割層の動的選択

量子化を行うことでロボットとクラウドサーバ間の通信量を削減できる．しかし，量子化により特徴マップの分解能が低下し，CNN の性能に影響する恐れがある．また，線形量子化を行う場合は量子化ビット数 Q を決める必要があり，非線形量子化の場合は量子化ビット数 Q と正規分布のパラメータ μ_n を決める必要がある．そこで，分割パラメータの決定則に要求レイテンシと要求精度を導入する．分割層は，$D \in \mathbb{N}$ で表される．量子化パラメータ P は

$$P \in \{\{\text{none}\}, \{\text{linear}, Q\}, \{\text{nonlinear}, Q, \mu_n\}\} \tag{5}$$

となる．分割層および量子化パラメータで構成される分割パラメータは，ロボットおよびクラウドサーバの処理時間

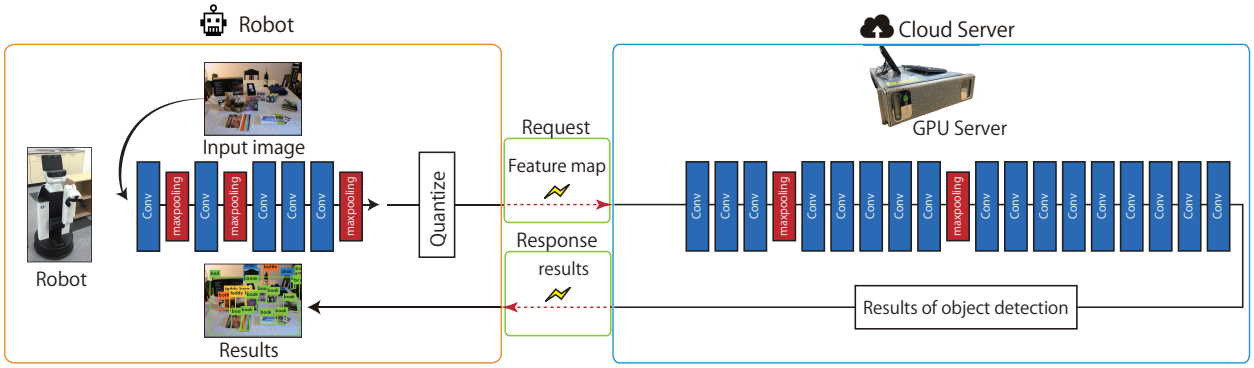

図 2：システムの構成

の予測値 $\hat{R}_t$,$\hat{S}_t$，通信時間 $\hat{T}_t$ と CNN の性能 A を用いて式 (6) で決定する．

$$\begin{aligned} con &= \arg\min_{D,P} \ \hat{S}_t \\ \text{subject to} \ \ &(\hat{R}_t + \hat{S}_t + \hat{T}_t) < R_L, \ A > R_A \end{aligned} \tag{6}$$

各条件を要求レイテンシ R_L および要求精度 R_A によって制約し，これを満たす条件の中で最もクラウドサーバ $\hat{S}_t$ の計算量が少ないものを選択する．

4. 評価実験

評価実験では，各分割層に線形量子化および非線形量子化を適用した場合の CNN の性能を評価し，量子化手法の比較を行う．また，CNN の性能を考慮して選択した分割層の妥当性を評価する．

4.1 量子化による CNN の精度の推移の調査

本実験では線形量子化と非線形量子化を適用した時の YOLO v2 の検出性能を調査する．量子化パラメータ μ_n は 0.1 から 0.9 まで変化させ，分割層は 2，6，12，14，16，18，量子化ビット数 $Q = 4$ で実験を行う．YOLO v2 の性能評価は，平均 IoU を用いる．IoU は，教師信号の矩形領域 box_t と，検出結果の矩形領域 box_d の重なり率である．

実験結果を図 3 に示す．量子化パラメータ μ_n が小さいほど平均 IoU は高くなり，量子化パラメータ μ_n が大きいほど平均 IoU は低くなることがわかる．また，分割層が 12 から 16 までの層では量子化による精度低下が著しいことがわかる．このことから，特徴マップの値の低い値が，CNN の性能維持に貢献していることがわかる．また，分割層によって CNN の精度低下率が異なるため，これらを考慮して分割層と量子化パラメータを決定する必要性がある．

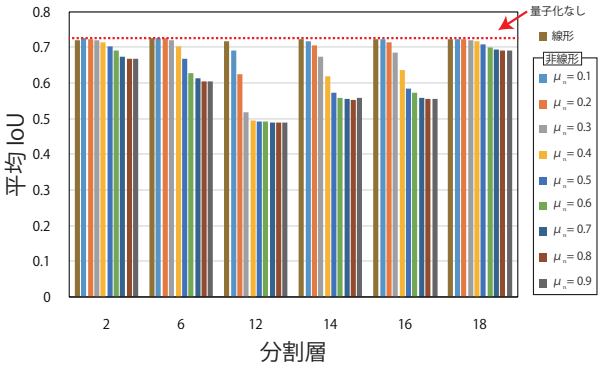

図 3：分割層と量子化手法の YOLO の性能推移

4.2 選択した分割層の評価

本実験では，YOLO v2 に提案手法を適用し，要求レイテンシおよび要求精度に対して選択した分割パラメータの妥当性を検討する．ロボットおよびクラウドサーバの計算能力，また通信帯域を表 1 に示す．表 1 に実験で想定する環境を示す．要求レイテンシ R_L は 500ms とし，要求精度 R_A は 72.5%とする．

表 1：実験条件

条件	ロボット性能 GFLOPS	クラウドサーバ性能 TFLOPS	通信帯域 Mbps
1	80 (Core-i7)	6.5 (GTX1070)	10
2	80 (Core-i7)	6.5 (GTX1070)	50
3	600 (Jetson TX2)	6.5 (GTX1070)	10

選択された分割パラメータを表 2 に示す．実験条件 1 の場合は通信帯域が狭いため，量子化ビット数 $Q = 6$ の線形量子化が選択された．また，分割層 D は出力層に近い 12 が選択されており，クラウドサーバの計算量を削減している．実験条件 2 では通信帯域が広く，特徴マップを圧縮しない場合でも通信時間が短いため量子化を行わない．そのため精度の低下が発生しない．実験条件 3 の場合はロボット側の計算能力が高いため，CNN の処理を出力層側まで処理を行っている．以上の実験により，要求レイテンシと要求精度を満たしつつ，クラウドサーバの計算量を最低限に抑えた分割パラメータを選択できる．

表 2：選択された分割層と量子化パラメータ

実験条件	D	量子化手法	Q	μ_n	m IoU	総処理時間
1	12	Linear	6	-	72.55	415.6
2	18	None	-	-	72.61	379.5
3	18	Nonlinear	8	0.1	72.51	353.2

5. おわりに

本研究では，物体検出タスクを対象としたクラウド型画像解析エンジンの動的分割法を提案した．提案手法では，特徴マップを圧縮するために量子化を行い，データ量を削減した．また，ユーザからの要求レイテンシと要求精度を導入し，ロボットおよびクラウドサーバの処理時間，通信環境を考慮して分割パラメータを決定した．今後の展望として，提案手法と他のクラウドデータベースを利用した物体把持システムの構築が挙げられる．

参考文献

[1] Y. Yamauchi, *et al.*, “Cloud Robotics Based on Facial Attribute Image Analysis for Human-Robot Interaction,” RO-MAN, 2016.

[2] J. Redmon, *et al.*, “YOLO9000: Better, faster, stronger” CVPR, 2017.

研究業績

[1] 猪子弘康 等,“クラウド型一般物体検出エンジンのための特徴マップ圧縮”, 情報学ワークショップ, 2017.

[2] 猪子弘康 等,“クラウド型顔画像解析エンジンにおけるレイテンシを考慮した DCNN の自動分割”, 日本ロボット学会学術講演会, 2016.