Binary-decomposed DCNN における量子化パラメータの自動最適化法

EP14058 近藤良太　　指導教授：山下隆義

1. はじめに

Deep Convolutional Neural Network (DCNN) は，画像認識の分野において高い性能を達成している．しかし，多くのメモリと計算リソースが必要なため，組み込み機器やモバイル端末への導入が困難である．Binary-decomposed DCNN (B-DCNN)[1] は，DCNN のパラメータを二値で表現することで，高速化とモデル圧縮が可能である．B-DCNN は，推論時に Quantization sub-layer において特徴マップを二値化する．このとき，各層の最適な量子化ビット数 Q を手動で設定する必要がある．本研究では，B-DCNN の Quantization sub-layer で用いる量子化ビット数 Q を自動で最適化する手法を提案し，更なる処理速度の高速化を図る．

2. B-DCNN の特徴マップの量子化

B-DCNN は，特徴マップを量子化して二値に変換する Quantization sub-layer を導入している．Quantization sub-layer では，特徴マップ $\mathbf{h}$ の最小値が 0 になるように $\mathbf{h}$ をシフトし，量子化して量子化ビット数 Q の二値にする．量子化ビット数 Q が大きい場合，特徴マップの表現力が高く，識別精度が向上する．しかしながら，処理速度は低下する．一方で，量子化ビット数 Q が小さい場合，処理速度が高速になるが特徴マップの表現力が乏しいため，識別精度が低下してしまう．

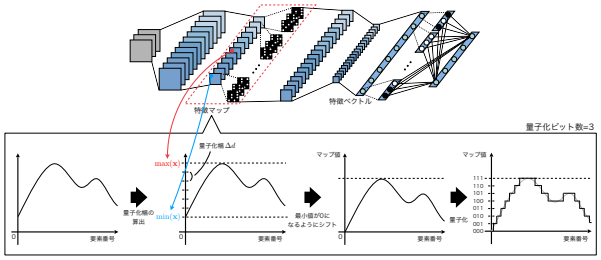

図 1：特徴マップの量子化 [1]

3. 提案手法

本研究では，Quantization sub-layer における最適な量子化ビット数 Q を指定した精度を保持しつつ，自動的に決定する方法を提案する．AlexNet を対象とする場合，畳み込み層の Q の組み合わせは ${}_{16,807}C_5$ と膨大である．提案手法では，前後の層の Q 値を固定した時の精度の変化に着目し，以下の Algorithm1 により決定する．また，各層の Q 値を選択する流れを図 2 に示す．

Algorithm 1 提案手法

Require: $\mathbf{A}$, $\mathbf{T}$, D, I, Q, R
Initialize: $\mathbf{T}$, S, $\mathbf{a}$
  for $i = 1$ to I **do**
    for $q_i = 2$ to Q **do**
      for $q_{i-1} = 2$ to Q **do**
        $S \leftarrow SA\,(i, q_{i-1}, q_i)$ // i 層の予想精度
        if $R - S < D$ **then**
          // 予想精度低下率が許容値より小さい
          $\mathbf{T}^i_{q_{i-1},q_i} \leftarrow S$ // スタックする
        end if
      end for
    end for
  end for
  for $i = I$ to 1 **do**
    $\mathbf{a}_i \leftarrow \underset{q_i}{\operatorname{argmin}}\, \mathbf{T}^i$ // 最小の Q 値を取得
  end for
  return $\mathbf{a}$

ここで，$\mathbf{A}$ は各層の精度低下率，$\mathbf{a}$ は最適な量子化ビット数の組み合わせ，$\mathbf{T}$ は予想精度のスタック，D は精度低下許容値，I は対象となる層数，Q は最大の量子化ビット数，R は量子化前の精度である．精度低下許容値 D は事前に設定しておく．$\mathbf{T}$ と S は最初に初期化する．次に，量子化前の精度から i 層の予想精度を減算し，予想精度低下率を求める．さらに，予想精度低下率が D より小さいとき，S を $\mathbf{T}^i_{q_{i-1},q_i}$ に スタックする．最後に，各層ごとに Q が最も小さくなる値を選択する．

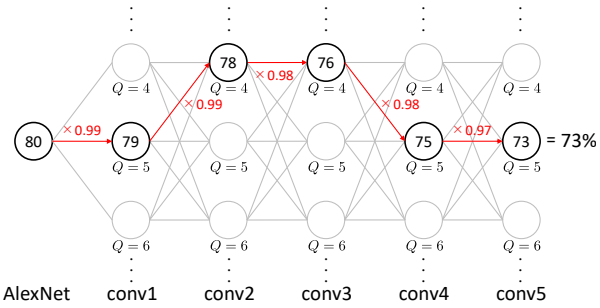

図 2：Q 値を選択する流れ

4. 評価実験

提案手法の有効性を確認するために，評価実験を行う．

4.1. 実験概要

ImageNet Dataset を用いて，提案手法を適用した際の識別精度と処理速度を評価する．検証サンプル 2,000 枚の Top-5 accuracy で評価する．ネットワークモデルには AlexNet を用いる．各モデルのパラメータには公開されている学習済みモデルを使用する．

4.2. 実験結果

提案手法を用いた B-DCNN と Q 値を固定した B-DCNN の識別精度と処理速度を図 3 に示す．ここで，提案手法*は指定した Q 値に対する予想精度，提案手法は実際に識別を行った際の精度である．Q 値を固定した B-DCNN は，任意に指定する精度に対して細かい調整ができず，B-DCNN(Q=3) と B-DCNN(Q=4) のときに精度に大きな差が生まれている．提案手法は，任意に指定した精度に対して約 2%程度の誤差が生じるが，指定された精度を満たす Q 値を選択して高速化を実現している．D が 8 と 9 のとき，B-DCNN(Q=3) と比較して識別精度を維持しつつ，処理速度が向上していることがわかる．

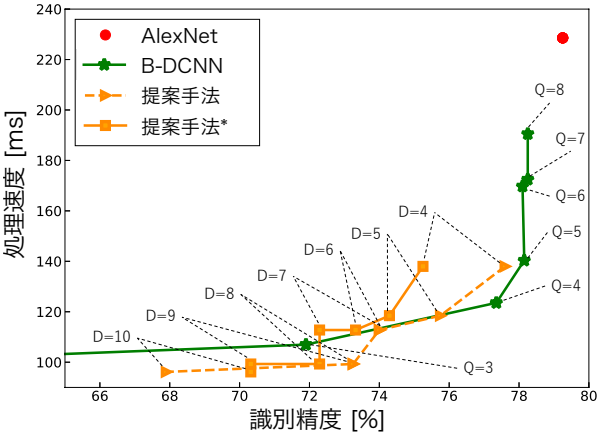

図 3：処理速度と識別精度の比較

選択された Q 値は，conv1 や conv2 のような浅い層では 4，5 と大きく，conv3 や conv4，conv5 などの深い層では 2，3 と小さい傾向が見られた．

5. おわりに

本研究では，各層で量子化ビット数 Q を変更することで，識別精度と処理速度の向上を実現した．今後は，Fine-tuning による Quantization sub-layer と重みの基底数の，最適な値を調整する方法を模索する．

参考文献

[1] R. Kamiya, *et al.*, “Binary-decomposed DCNN for accelerating computation and compressing model without retraining”, ICCV Workshop, 2017.