Memory efficient deep learning for mobile devices

05 深層学習をモバイル向けに小さくします

～量子化による深層学習のモデル圧縮技術～

どんな研究

画像や音声などの認識に深層学習が盛んに用いられています。しかし、重みを保持するパラメータの数が多いため、メモリを大量に消費します。本研究ではパラメータの分布を考慮した新たな量子化により、深層学習のメモリ消費量を３２分の１に低減できる学習法を提案します。

どこが凄い

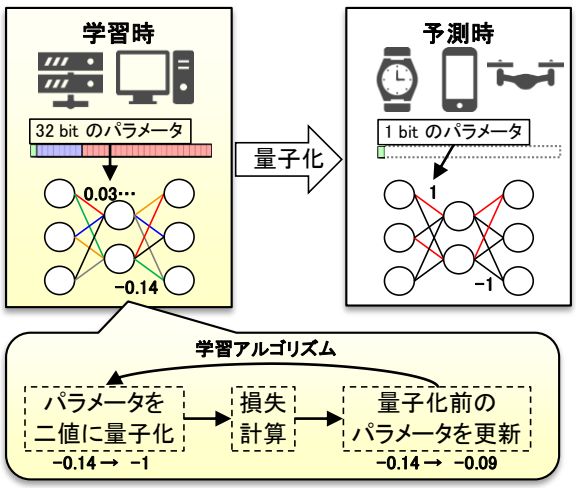
深層学習の連続値であるパラメータを±１に限定する量子化の影響を最小限にすると同時に効率的な学習アルゴリズムとすることで、従来手法よりも高精度かつ高速な量子化が可能となります。また、学習後にはビット演算で計算できるため、高速に予測できるようになります。

めざす未来

本技術は要求性能の低い深層学習を作ります。そのため、パラメータが多いほど高精度な予測の行える深層学習を性能の低い小型端末上で動かせるようになります。これにより、これまで深層学習が使えなかった様々な端末で画像や音声などの高度な認識が行えるようになります。

背景

パラメータの値を ±1 に限定して予測時のメモリ消費量を軽減する



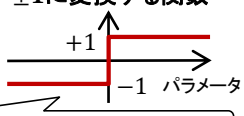
問題点　認識性能の低下

原因①　±1への変換は近似誤差が大きい

原因②　適切な更新量が求まらないため学習がうまくいかない

更新量 ＝ 学習率 × **関数の微分値**

±1に変換する関数



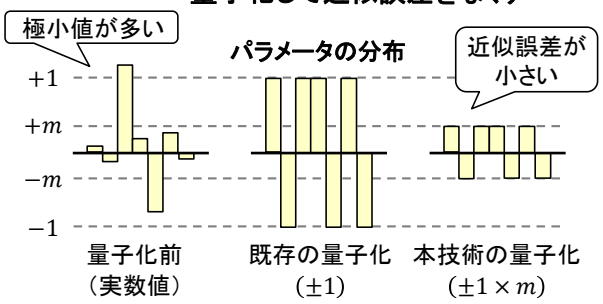
微分値 0 → 更新量の消失

更新量の消失対策
- 学習率を高くする
- 微分値を強制的に0から1にする

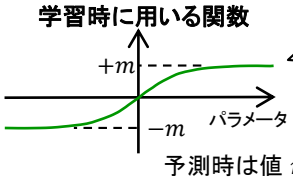
⇒ 学習が困難

提案手法

原因①の対策　パラメータをノルムの平均値 m に量子化して近似誤差をなくす



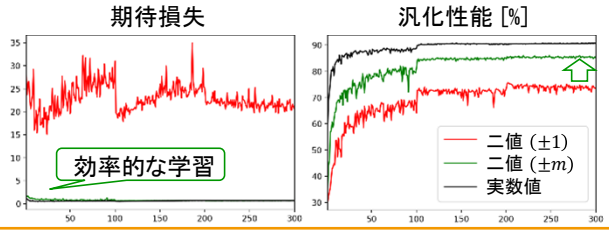
原因②の対策　近似した連続関数を用いて適切な更新量で学習する



自動的に適した形に縦横をスケールさせ、効率的に学習させる

予測時は値 m を活性化の閾値に使い回すことでパラメータを±1にして推論する

学習結果



関連文献

[1] 大屋優, 井田安俊, 藤原靖宏, 岩村相哲, “正則化による深層学習の重み量子化手法の検討,” *電子情報通信学会技術研究報告*, Vol. 117, No. 211, pp. 51–52, 2017.

[2] 大屋優, 井田安俊, 藤原靖宏, 岩村相哲, “バイナリニューラルネットにおける非活性化手法の検討,” *電子情報通信学会技術研究報告*, Vol. 117, No. 238, pp. 119–120, 2017.

担当者

大屋 優 (Yu Oya) ソフトウェアイノベーションセンタ 分散処理基盤技術プロジェクト
分散システムアーキテクチャ基盤グループ