ISS-SP-217

量子化結合ニューラルネットワークの多層化による認識精度の向上

品川　政太朗†　　早川吉弘††　　中島　康治†
† 東北大学大学院情報科学研究科　　†† 仙台高等専門学校

1. はじめに

Deep Learning と呼ばれる多層構造のニューラルネットワーク(Deep neural network, DNN)の学習手法が数々のタスクで高い精度を出し注目を集めているが，より複雑なタスクを解くためにより多層のネットワークを学習することが求められており，層数の増加により膨大になった荷重値を保持するメモリが不足するという問題が重大なボトルネックになっている．この問題を解決する方法としては荷重値を離散値に丸めて学習する方法が考えられ，特に 2 値や 3 値の簡単な値で荷重値を表現(量子化)することができれば相当のメモリを削減することが期待できる．荷重値の離散化は性能を大幅に劣化させる難点があるが，隠れ素子を増やすことで性能劣化が改善される報告[1]があることから，層数を増やすことでも丸め誤差による性能劣化を緩和させることが期待できる．よって本研究では層数を増やすことで 3 値量子化に代表される大きな離散幅においても性能劣化が緩和される傾向にあることを示す．

2. 離散荷重値 DNN の学習手法

DNN の学習では制限付きボルツマンマシン(RBM)による事前学習[2]を行う．事前学習によって最上層間以外の構造を事前につくり込むことによって学習時における荷重値離散化の丸め誤差の悪影響を低減することがねらいである．事前学習ではメモリ不足の問題は起こりにくいと想定されるので連続荷重値で学習を行う．事前学習で得られたパラメータを用いて DNN を構成し BP 学習による fine-tuning を行う際には Randomized Rounding 法[3]を用いて荷重値を確率的に離散値に丸めて学習を行う．

3. 実験条件

実験には MNIST の文字認識データセットを用いた．データは 100 パターンを 1 つのミニバッチに分割して学習を行った．事前学習での RBM は 50epoch 計算し，学習率は 0.1，weight decay は 0.002，モメンタムは 20epoch まで 0.5 とし，21epoch から 50epoch まで 0.9 とした．fine-tuning は 100epoch 計算し，学習率を 0.1 とした．fine-tuning で荷重値の離散化を行う際には荷重値の値が[-1,1]の値になるように正規化するためにゲインの調整を行った．具体的には各層間ごとに荷重値の絶対値の最大値を算出し，その値をその層間に用いる活性化関数のゲインとして用いた．離散値はハードウェア化を志向して 2 のべき乗とした．学習する DNN の層数は入力層，出力層を含めて 3～8 層で実験を行った．

4. 結果と考察

図1は1層あたりの隠れ素子数が 300 の場合の結果である．離散幅2^0のときが 3 値の量子化に対応する．このとき，6 層の場合で最も量子化の性能劣化が緩和され，3 層の場合と比べ 10%近い認識精度の改善がみられた．また，2^{-2}以下の離散幅では層数を増やすことで性能劣化が緩和される傾向を示し，さらに層数を増やすと再び性能が劣化する傾向を示した．これは MNIST が単純なタスクであるため，層数が増えすぎるとネットワークが冗長になり局所解に収束しやすくなっているためだと考えられる．

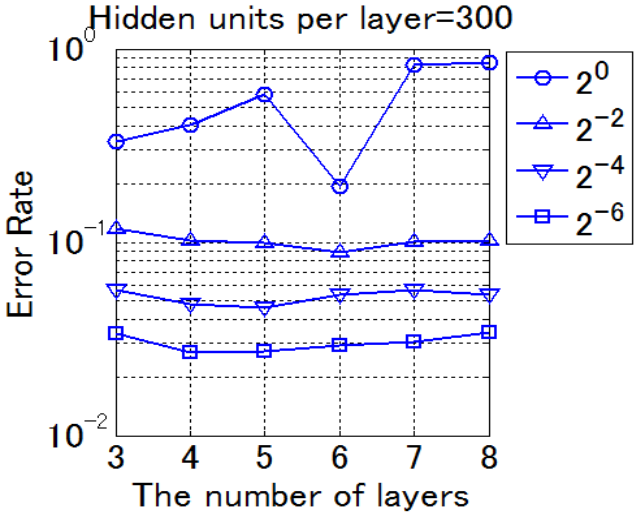

図 1:1 層あたりの隠れ素子数が 300 のときの離散幅をパラメータとした層数に対する誤認識率

5. まとめ

層数を増やすことで荷重値離散化による性能劣化を緩和できる傾向があることを示した．また，量子化のような大きな離散幅でも層を多層にすることで層が浅い場合よりも性能劣化が顕著に緩和される可能性を示した．

参考文献

[1] Wojnarski, Marcin. "Nondeterministic Discretization of Weights Improves Accuracy of Neural Networks." *Machine Learning: ECML 2007* (2007): 765-772.

[2] Hinton, Geoffrey, Simon Osindero, and Yee-Whye Teh. "A fast learning algorithm for deep belief nets." *Neural computation* 18.7 (2006): 1527-1554.

[3] Golovin, Daniel, et al. “Large-Scale Learning with Less RAM via Randomization." Proceedings of the 30 International Conference on Machine Learning (ICML) (2013), pp. 10