

低ランク近似技法の深層学習への適用

板垣 翼[†] 鈴木 智博[‡]山梨大学大学院工学専攻コンピュータ理工学コース[†] 山梨大学医学工学総合研究部[‡]

1 はじめに

深層学習の登場は画像処理や自然言語処理などの分野を大きく発展させている。そこで扱うタスクの複雑化は深層ニューラルネットワーク (DNN) や畳み込みニューラルネットワーク (CNN) を始めとしたモデルの複雑化を招き、重みパラメータの増大に繋がるが、近年ではスパースモデリングというモデル圧縮技術により改善されてきた。一方でモデルへ与える入力データの圧縮化に関してはあまり研究が行われていない。本研究では深層学習モデルに与える入力データの効率的な圧縮方法として、行列の近似圧縮技術として用いられている低ランク近似技法を利用した圧縮方法を試みその性能を調査した。

2 畳み込み演算

CNN に特徴的なものとして畳み込み層とプーリング層がある。特に畳み込み層で行われる畳み込み演算は、CNN の計算処理の大部分 (9 割程度) を占めていることが知られている。畳み込み演算では単純に入力データ (画像) と重みフィルタ (カーネル) の内積計算を行うが、あまり効率的でないため通常そのまま実装されることは少ない。一般的には Unrolling や Lowering と呼ばれる手法 [2] を用いて、畳み込み演算を行列行列積に落とし込むように実装されている。このような変換を行うことで BLAS (Basic Linear Algebra Subprograms) やその高速版実装を利用できるため畳み込み演算が効率的に計算されることになる。

Unrolling を用いることで畳み込み演算を効率的な行列行列積に変換できるが、アルゴリズムの性質上、入力画像をピクセル単位で複製する必要があるためデータサイズが増大してしまうという問題がある。Unrolling の具体的な一手法として知られている im2col 用いるとして、入力画像の高さと幅とチャンネル数をそれぞれ H , W , C , 畳み込み層のカーネルの高さと幅をそれぞれ R ,

S とし、出力画像の高さと幅が入力画像と等しくなるように畳み込みを行う場合、入力画像のデータサイズは、

$$\frac{\text{複製された入力画像のサイズ}}{\text{入力画像のサイズ}} = \frac{CRS \times HW}{CRS} = RS$$

倍されることになる。この複製された入力画像は誤差逆伝播法によってパラメータの勾配を求める際にも利用する必要があるため、勾配の計算まで通常より多くのメモリを消費することになる。

3 低ランク近似技法の適用

科学技術アプリケーションで現れる多くの計算は最終的に行列演算に帰着するが、問題の規模が増大するにつれ計算量が増大し、現実的な時間ではでは解くことができなくなってしまうことがある。行列の低ランク近似技法はこのような問題を解消する手法として発展してきたものである。低ランク近似を用いることで行列演算の高速化やデータサイズの削減が期待できる。

バッチサイズを N とすると im2col によって入力画像は大きさ $CRS \times NHW$ の横長行列 D となる。im2col によるデータサイズの増大の問題を解消するため、以下ではこの行列に低ランク近似技法を適用する。具体的には D を列方向に N 分割した部分行列 $D_i (i = 1, \dots, N)$ を

$$D_i \approx U_i V_i$$

のように 2 つの行列の積として近似する。ここで U_i と V_i の大きさはそれぞれ $CRS \times r$, $r \times HW$ とする ($r \ll CRS$)。なお 2 つの行列の積に近似する手法として ACA (Adaptive Cross Approximation) [1] 用いることにした。ACA では

$$\frac{\|D_i - U_i V_i\|_F}{\|D_i\|_F} < \varepsilon_r$$

を満たすような十分小さい閾値 ε_r を用いて近似が行われる。ここで $\|\cdot\|_F$ はフロベニウスノルムを表している。

4 実験

上述した低ランク近似技法によって im2col によるデータサイズの増大を抑えた際の圧縮性能やモデルの画像認識の精度を調査した。学習するデータセットは

Applying low-rank approximation to deep learning

[†] Tsubasa Itagaki, Integrated Graduate School of Computer Science and Engineering, University of Yamanashi[‡] Tomohiro Suzuki, Interdisciplinary Graduate School of Medical and Engineering, University of Yamanashi

CIFAR-10 で、実験に用いたモデルは以下に示す畳み込み層が6つの比較的単純な CNN である.

$$\{\text{conv} \rightarrow \text{conv} \rightarrow \text{dropout} \rightarrow \text{pooling}\}_{\times 3} \\ \rightarrow \{\text{fc} \rightarrow \text{dropout}\}_{\times 2} \rightarrow \text{softmax}$$

以下の実験では、100000 と 101010 (6つの畳み込み層のうち低ランク近似技法を適用しない層とする層をそれぞれ0と1で表し、それらを入力層に近い順に並べて表記した)の2種類のモデルに学習させる. なお ACA で用いる近似閾値 ε_r の値は全て 0.1 とした. 低ランク近似技法を一切適用していないベースモデルの結果も加えた認識精度の結果を図 4.1 に示す. これより 101010 モデルでは上手く学習ができず, 100000 モデルでは上手く学習ができていることがわかる. またこのときベースモデルの認識精度に比べ 100000 モデルは最小で 6.7% の精度の低下があることがわかった.

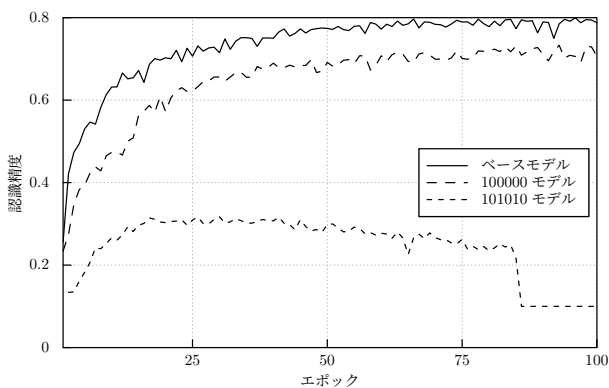


図 4.1 認識精度の比較

続いて同じ実験における 100000 モデルの入力データの圧縮に関しても調べた. 1 番目の畳み込み層に渡す入力画像のデータサイズをベースモデルと比べた結果を表 4.1 に示す. なお圧縮率は

$$1 - \frac{\text{圧縮データのサイズ}}{\text{元データのサイズ}}$$

と定める. これより im2col によるデータサイズの増大

表 4.1 データサイズ (バイト数) の比較

ベースモデル	100000 モデル	圧縮率
11059200	2592556	0.77

を約 77% 抑えることが出来ていることがわかった.

5 考察

100000 モデルと 101010 モデルに学習させた結果 101010 モデルでは学習が上手くなされなかった. 理由の1つとしては, 100000 モデルに比べ入力データの近似圧縮を2回多く行っているために, 近似の回数が増える度に畳み込みの担う役割を消失, 更には悪化させてしまっている可能性が考えられる. 一方 100000 モデルで学習が上手く進んだ理由としては, 入力データの近似が1番目の畳み込み層(入力層)のみで行われるために, モデルが本来の入力データと一対一に対応している近似データを学習しようとするためだと考えられる. そして認識精度がベースモデルのものよりも低下したのは, 近似の閾値 ε_r に起因するものと考えられる.

6 今後の課題

畳み込み演算で用いられる im2col の入力データ複製によるデータサイズの増大を抑えるため, 行列の低ランク近似技法を適用した. 約 6.7% の認識精度の低下でデータサイズの増大を約 77% 抑えられることがわかった. また実験結果から今回の近似操作を行うのは入力層に限るのが良さそうであることもわかった.

今のところ低ランク近似技法を適用した畳み込み層の出力結果はベースモデルのものと同じデータ構造になるように実装されているため, 高速化が見込めていない. 内部の入力データの圧縮も含め, 今後の課題としては, 学習全体でシームレスに近似されたデータを扱えるような工夫を考える必要がある.

参考文献

- [1] Kezhong Z. et al. The Adaptive Cross Approximation Algorithm for Accelerated Method of Moments Computations of EMC Problems. *Electromagnetic Compatibility, IEEE Transactions on*, Vol. 47, pp. 763 – 773, 12 2005.
- [2] Kumar C. et al. High Performance Convolutional Neural Networks for Document Processing. In Guy Lorette, editor, *Tenth International Workshop on Frontiers in Handwriting Recognition*, La Baule (France), October 2006. Université de Rennes 1, Suvisoft. <http://www.suvisot.com>.