深層学習での演算量削減技術

松本　渉*

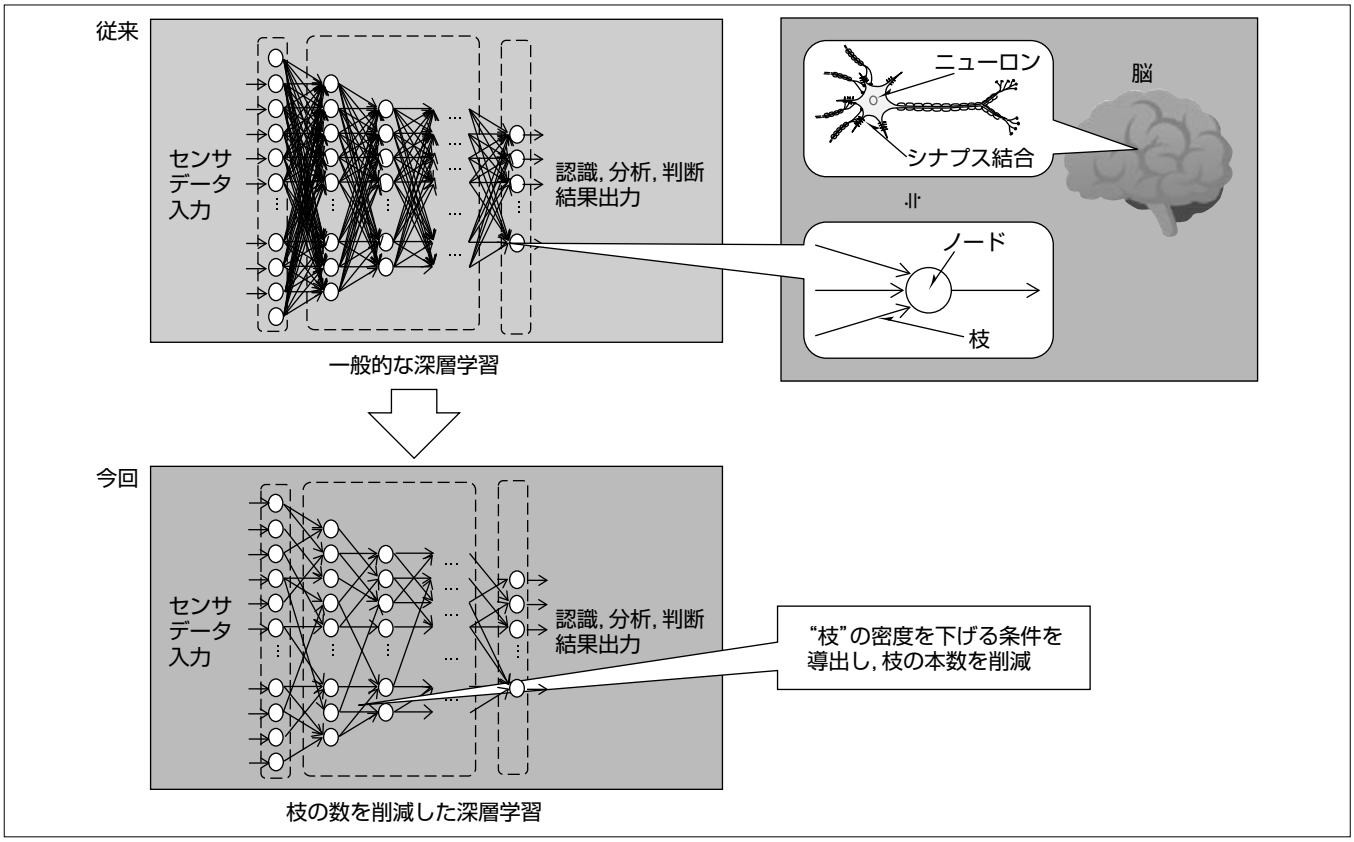
Computational Complexity Reduction Technology for Deep Learning

Wataru Matsumoto

要　旨

人工知能の実用化が急速に進んでいる。この人工知能の目覚ましい発展に寄与する技術の１つとして“深層学習（ディープラーニング）”が挙げられる。この深層学習は，認識だけではなく，医療や機器の診断から，ゲーム，機器制御まで応用されているが，演算量が大きく，大量のデータを扱うことから，一般に，サーバに実装され，クラウドサービスとして提供される。しかし，今後のIoT（Internet of Things）の進展によって，データ量が莫大（ばくだい）なものとなると，処理遅延が生じるなど深層学習を用いる人工知能機能をクラウドサービスで実現することには限界がある。すなわち，深層学習を機器，又は，機器に近いネットワークの端（エッジ）のデータ処理装置に実装する必要が生じる。そのため，深層学習の演算量を低減することが必須である。

そこで，三菱電機では，深層学習で用いられるニューラルネットワークのノード間の結合（枝）の密度に着目し，この枝を大幅に削減して演算量を低減する方式を開発した。この方法の有効性を確認するため，手書き数字認識を対象とする実験を行ったところ，従来の全接続のニューラルネットワークに対して，接続数を1/25にしても，認識性能の劣化が小さく，性能維持できることを確認した。この方法を，今後カメラ画像解析や機器の診断などに適用し，それぞれ不審者検出や異常検知など安心・安全や保守・保全を支援する機能を実現する。



深層学習の一種であるディープニューラルネットワークでの演算量削減技術

人間の脳を模擬したノードと枝によるネットワーク構造を用いた深層学習は，画像認識等で他の機械学習を抑えて最も高い認識性能を示している。一方で，演算量が大きい問題がある。この問題の克服のため，認識性能をほぼ維持したままノード間を結ぶ枝の本数を削減するアルゴリズムを開発した。この枝の本数削減効果によって枝に係る演算が削除でき，大幅な演算量削減効果を生むことができる。

*情報技術総合研究所（博士（工学））

1. まえがき

人工知能の目覚ましい発展に寄与する技術の１つとして“深層学習(ディープラーニング)”が挙げられるが，演算量が大きく大量データを扱うことから，一般に，サーバに実装され，クラウドサービスとして提供される。しかし，今後のIoTの進展によって，データ量が莫大なものとなると，処理遅延が生じるなど深層学習を用いる人工知能機能をクラウドサービスで実現することには限界があり，深層学習を機器，又は，機器に近いネットワークの端(エッジ)のデータ処理装置に実装する必要が生じる。そのため，深層学習の演算量を低減することが必須である。

本稿では，深層学習で用いられるニューラルネットワークのノード間を結ぶ枝の密度に着目し，この枝を大幅に削減して演算量を削減する方式を提案する。

2. 深層学習の課題

2.1 深層学習のネットワーク構造

脳のニューロンとシナプス結合をそれぞれノードと枝による数理モデルで模擬したネットワークをニューラルネットワークと呼ぶ。特に中間層(入力層と出力層を除く層)が２層以上あるニューラルネットワークをディープニューラルネットワークと呼ぶ。

このような深いネットワーク構造を持つニューラルネットワークを用いる機械学習を深層学習(ディープラーニング)と総称している(**図１**)。入力層にはセンサデータなどが入力され，中間層では特徴を抽出し，出力層では分類や回帰の結果を出力する。ノード間の接続の強さを枝の重みという係数で表現する。ディープニューラルネットワークは各層間のノード同士が全て枝で接続されており密度が高い“密”なネットワーク構造になっている。

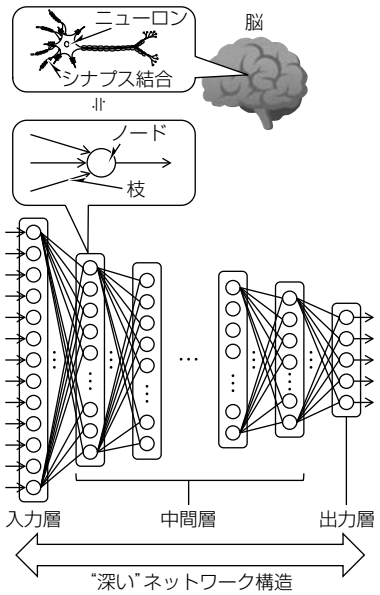


図１．深層学習のネットワーク構造

2.2 深層学習での自己符号化器

ニューラルネットワークは古くから誤差逆伝播(でんぱ)法という方法で学習を行っていたが，中間層が２層以上ある場合，学習がうまく収束しない問題があった。2006年にジェフリー・ヒントンらがこの問題の解決のために自己符号化器を用いることを提案している[(1)]。学習過程を教師なし学習で重み付けを層ごとに初期化する“プレトレーニング(事前学習)”と，教師あり学習で最後に全体を最適化する“ファインチューニング”の２段階に分ける手法が有効であることを示した。このプレトレーニングの段階で自己符号化器を用いている。この手法によって深層学習が可能になった。自己符号化器は中間層に対する入力層(複数の中間層がある場合は入力側にある１つ前の層)をそのまま出力層に用い，次元圧縮した中間層(入力層のノードの数より中間層のノードの数を少なくする)で入力層のデータを出力層で復元できるように学習する手法である。

図２に示す自己符号化器について述べる。N個の入力データ$\mathrm{x}\in\mathrm{R}^N$(Rは実数の集合)に対して$M\times N$の枝の重み行列$\mathrm{W}\in\mathrm{R}^{M\times N}$，$M$個のバイアス$\mathrm{b}\in\mathrm{R}^M$と活性化関数$\mathrm{f}(\cdot)$を用い，$\mathrm{y}=\mathrm{f}(\mathrm{Wx}+\mathrm{b})$として$M$個の中間層のデータyを表現し，復元するデータを$\mathrm{x}'$としたときに$\mathrm{x}'=\tilde{\mathrm{f}}(\tilde{\mathrm{W}}\mathrm{y}+\tilde{\mathrm{b}})$によってxを復元できるように学習する。なお，図の$\mathrm{y}^l$は$l$番目の中間層を示している。この学習後，自己符号化器の復元側のネットワークを除き入力側のネットワークをそのままl番目の中間層の初期化したネットワークとして使用する。この操作を$l=1$から昇順に全ての中間層に対して行い，プレトレーニングを完了する。この自己符号化器の操作を加えることによって，この後，ファインチューニングとして誤差逆伝播法を用いても学習が収束するようになる。プレトレーニングを行わない場合は，ネットワークのパラメータは乱数で初期化されるため，学習が収束しない場合があったが，プレトレーニングを行うことで，学習対象を表現するのに適したパラメータを用意することができ，学習が収束できるようになった。

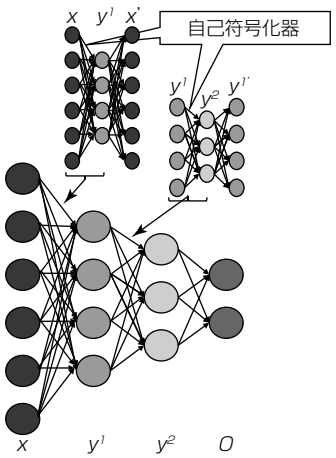


図２．自己符号化器によるプレトレーニング

2.3　深層学習の課題

図1に示すようにディープ・ニューラルネットワークは各層間のノード同士が全て枝で接続されており，密なネットワーク構造になっている。各枝はこの重みの演算のために積算を行う必要があり，枝の本数が増えるほどこの演算量は増大することになる。また，深層学習は多層の構造を持つため，更に演算量を大きくしている要因となっている。そのため，大規模サーバや組み込みの場合でも演算を高速化するGPU(Graphic Processing Unit)等を用いないとアプリケーションの要求仕様を満たさないケースが多く，エッジのデータ処理装置に搭載する上での課題になっていた。

3. 深層学習の演算量削減技術

深層学習の演算量の課題解決のために**2.2節**で述べた自己符号化器に着目して自己符号化器が持つ機能を実現するための条件を導出し，演算量に大きく影響を与える枝の本数の削減が可能かどうかを検証した。

3.1　自己符号化器における枝の重み行列Wの条件

各層の計算は入力データxのi番目の要素x_iと中間層のデータyのj番目の要素y_jとし，x_iとy_jを結ぶ枝の重みw_{ij}を成分に持つ行列としてWで表現する。特にバイアスbは正規化によってb＝0とすることができ，自己符号化器の場合の活性化関数は多くの場合恒等写像(f(a)＝a，つまり何も変化させない)が選ばれることから，簡易的にy＝Wxと表現し，**図3**のように各層のネットワークの計算式を行列演算として考えることができる。全ての入力層のノードと中間層のノードが枝によって接続されている一般的なディープニューラルネットワークの場合，図の上部の行列のように全ての成分が重みを持つ密な行列となる(図中の行列の色の変化は様々な重みの値を表現している)。**2.2節**で述べたように自己符号化器は次元圧縮した中間層のデータyから入力データxを復元する問題であるが，このとき求められる次元圧縮されたyの要素数(中間ノードの数)の下界(理論上の最小の数。N，Mが無限大などの大きな次元の場合に成り立つ)を導出してみると，Wが密な行列か疎な行列(行列の多くの成分が零である行列)に依存せず同じ下界になることが分かった[2]。つまり，先に述べた下界を求めて中間ノードの数をM個としたとすると，Wが密な行列か疎な行列にかかわらず同様に復元でき，自己符号化器の求める性能を共に満足していることを意味している。行列の成分が零の箇所は計算する必要がないため，零成分の割合が大きいほど演算量は削減できる。

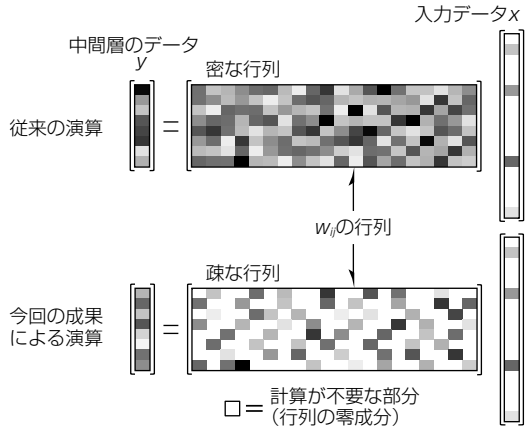


図3．疎な行列を用いた枝の計算

3.2　Wの密度を変化させた場合の評価結果

公開手書き数字(28×28＝784画素のグレースケールの手書きの0～9の数字，50,000個の学習用データセットと10,000個のテスト用データセットが用意されている)データベースMNIST[3]を用いてWの密度を変化させた場合の評価実験を行った。

ネットワーク構成は入力ノード数N＝784(28×28グレースケース画像で[0，1]の範囲に正規化されている)で，中間ノードはM＝500とする。この入力層と中間層で構成されるネットワークを$M\times N$の行列Wとする。また，出力層は0～9までの10ノードとした。プレトレーニングで，10,000エポック(1エポックは全学習用データセットを1回学習することをいう)学習した。また，ファインチューニングで10,000エポックの誤差逆伝播を行い，10,000個のテストデータで評価実験を行った。

行列Wでの全ての成分数に対する非零の成分数の比を密度γとする。また，行列の成分は平均0，分散値$1/\gamma$の正規分布に従うものとし，密度γで非零，密度$1-\gamma$で零の値をとるものとする。**表1**にその結果を示すが，$\gamma=1$では全成分が非零となっていることを意味する。γ＝0.04の場合は，行列全体における非零成分が0.04＝1/25の比率であることを示している。今回，行列Wの密度の違いによるプレトレーニング時とファインチューニング後の性能差を評価した。

まずはプレトレーニング時の自己符号化器で入力した手書き数字と復元した手書き入力画像の二乗誤差(Mean Square Error：MSE)を**表1**に示す。行列Wの全ての密度でMSEは5.0e-4近辺であり，ほぼ同等の値で収束していることが分かる。

次に，ファインチューニング後の10,000個のテスト用データセットによる正解率の評価結果を**表2**に示す。この結果を見ても分かるように全体として98%近辺の平均正

表1．プレトレーニング時のMSE

行列の密度(γ)	1.00	0.33	0.16	0.08	0.04
MSE	5.29e-4	5.12e-4	4.95e-4	5.19e-4	5.88e-4

表2．手書き数字の識別結果

行列の密度(γ)	1.00	0.33	0.16	0.08	0.04
平均正解率(%)	98.38	98.34	98.13	98.17	97.90

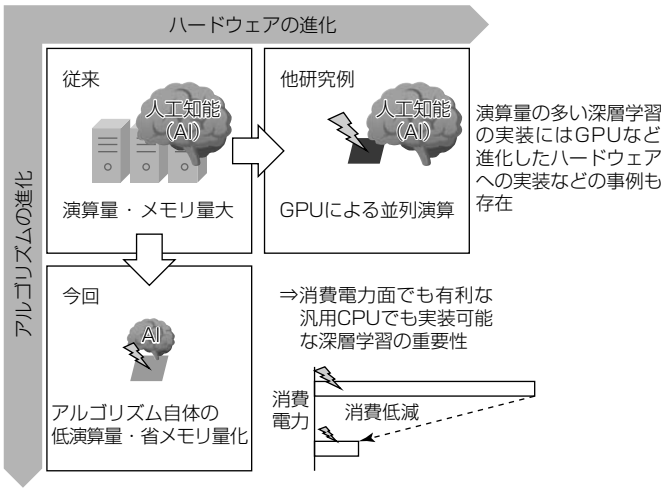


図4．深層学習の低演算量化による実装面の特長

解率を示しており，$\gamma = 0.04 = 1/25$の非零成分の大幅な削減による演算量削減にもかかわらず，平均正解率の劣化が0.48ポイントと僅かであることが確認できた。従来，深層学習に対する枝刈りの手法も提案されていたが[4]，$\gamma = 0.1$で10％近い平均正解率の劣化が示されており，従来法よりも大幅に改善されている。

一般に入力データの次元数が大きいほど，性能の大きな劣化を伴わずにγを小さくできる傾向がある。行列Wの零成分の部分は計算が不要なだけでなく，保存しておく必要もないため，メモリ量の削減効果もある。

この演算量削減技術によって**図4**に示すように，演算量が多い深層学習の実装をGPUに頼らず，汎用CPU等に実装できる演算量に抑えることによって，小型化や低消費電力化に貢献でき，組み込み実装が容易になる。

4. 活用できる分野

一般に，深層学習は分類と回帰が可能となる。分類とは画像の物体識別や音声認識，異常検知等の入力データから特徴のあるものを弁別する機能である。また，回帰は主に連続値をとる関数を近似する機能で，株価予測，市場予測，電力需要予測等に用いることができる。これらの機能を身の回りの機器に搭載することによってエッジのデータ処理装置で判断できるようになる。

5. むすび

人工知能の中核的技術である深層学習の演算量削減技術を提案した。特にニューラルネットワークの本質的な問題を解決し，深層学習のブレークスルーを生んだきっかけとなった自己符号化器によるプレトレーニング手法の原理を確認し，この機能達成に必要な条件を導き出して，ネットワーク構造の枝が疎か密かによる理論的な性能限界に対する差はほとんどないことを示した。この方法の有効性を確認するため，手書き数字認識を対象とする実験を行い，従来の全接続のニューラルネットワークに対して，接続数を1/25にしても，認識性能の劣化が小さく，性能を維持できることを確認した。この成果によって実装上，演算量とメモリ量が削減でき，様々な機器への実装を容易にすることができることを示した。

参考文献

(1) Hinton，G. E.，et al.：Reducing the Dimensionality of Data with Neural Networks，Science，**313**，No.5786，504～507 (2006)

(2) Matsumoto，W.，et al.：A Deep Neural Network Architecture Using Dimensionality Reduction with Sparse Matrices，ICONIP2016，Part IV，LNCS 9950，397～404 (2016)

(3) LeCun，Y.，et al.：Gradient-Based Learning Applied to Document Recognition，Proc. IEEE，**86**，No.11，2278～2324 (1998)

(4) Anwar，S.，et al.：Structured Pruning of Deep Convolutional Neural Networks，ACM Journal on Emerging Tech. in Computing Systems 13(3) (2015)