1M-02

ニューラルネットワークのパラメータ削減と判断根拠の明示化に関する検討

竹花　治紀†　　尾崎　敦夫‡

大阪工業大学　情報学部　コンピュータ科学科†（現　情報知能学科‡）

1. はじめに

画像認識などの分野においてニューラルネットワークといった機械学習モデルは、従来よりも高精度であるが結果に対する説明が困難であるという課題がある．

本稿では，ニューラルネットワークの隠れ層に対して，出力結果に大きく貢献しない重みのパラメータ分布に注目し，不要なパラメータを間引く手法を提案する．これにより，ニューラルネットワークの出力結果に影響を与える重要なパラメータを抽出し，判断根拠を明示化する．

2. 関連研究

ニューラルネットワークのパラメータに対して Han らの研究[1]は重要度の低いニューロンの重みを枝刈りし，学習を繰り返すことでニューラルネットワークの精度を損なわずに不要な重みパラメータを削減する手法を提案している．この手法は以下の 3 段階で構成されている．

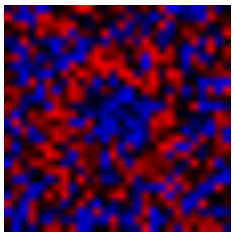
1. 通常のニューラルネットワークの学習
2. 重みの削減
3. 再学習

2 段階目は重みの L2 ノルムを求め，全体の重みから指定した割合分，L2 ノルムの小さい順に重みを削減する．3 段階目では削減したニューラルネットワークを再び学習させる．この 2，3 段階目を繰り返すことでパラメータを削減する．

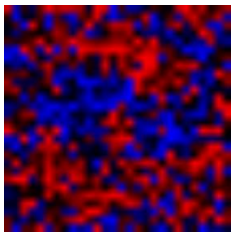
3. 提案手法

3.1 基本概念

全結合ニューラルネットワークで画像認識の学習を行う際に，画像にみられる特徴を習得する部分は，入力層と隠れ層 1 層目を結合する部分と考える．そのため，隠れ層のパラメータを重みの度合いに応じて着色することで学習データに似た様子が確認できると考える．その様子として，手書き数字画像データの MNIST[2]を学習データとしたニューラルネットワークの入力層と隠れ層 1 層目を結合するニューロンの重みの一例を図 1 に示す．着色方法は正の値は赤く，負の値は青く表示している．また，値が大きいほど色の輝度を高くしている．図の縦横は MNIS の入力画像に対応している．これらからは中心部分の赤い箇所から“0”(図 1(a))や“2”のような形状(図 1(b))がおおまかに確認できる．このような形状以外にも，ニューラルネットワークは文字を判断するために入力を分割してとらえていると考える．



(a)“0”のような形状



(b)“2”のような形状

図 1　入力層と隠れ層を結合させる重み

そこで，出力に影響を与えない重みを削減し，数字の判別に必要な領域を特定することで判断根拠の明示化が可能であると考える．

3.2 処理概要

提案手法の処理の流れを図 2 に示す．本手法は Han らの手法[1]を用いて重みの一部を削減する．対象の重みは全結合ニューラルネットワークの入力層から隠れ層 1 層目をつなぐ重みとする．また，ニューラルネットワークはあらかじめ一定の学習を行い，学習が収束しているものとする．

はじめに，各重みの L2 ノルムを求め，その値を小さい順に全体の 4～5 割程度削減する．これにより，値の大きな重みが残りの領域を形成する．これによって削減されたニューロンの重みの様子の例を図 3 に示す．図 3 の黒い部分は重みの値が 0 であり削除されていることを示す．これらの重みが残った領域は学習データから習得したデータの特徴を示す領域と考える．

Study of parameter reduction and visualizing an interpretability for neural networks

Haruki Takehana† and Atsuo Ozaki‡

Osaka Institute of Technology, Faculty of Information Science and Technology, Department of Information and Computer Science

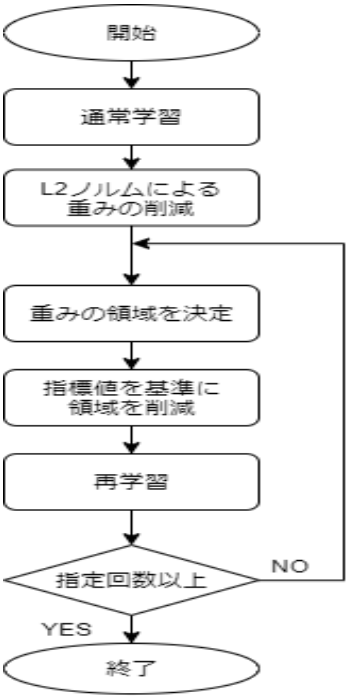


図 2　提案手法による処理の流れ

また，図 3 での各ニューロンの重みから，重みが連なった領域をグループ化したものを図 4 に示す．図の白い部分は重みが削減されずに残った領域であり，黒い部分は重みが削減された領域を示す．

次に，グループ化された重みの領域ごとに重みの大きさの平均値を求める．この値を領域ごとの指標値に設定し，出力に影響を与える領域かどうかの判定に用いる．そして，すべての領域を対象として指標値の小さい順に一定の割合，削減を行う．その後，ニューラルネットワークの再学習を行う．

これらの処理のうち，領域の分割（重みの領域を決定）から再学習までを指定回数繰り返す．

4. 評価

4.1 設定条件

本実験は全結合ニューラルネットワークを対象とし，手書き数字画像データの MNIST を学習データとして行った．対象のモデルは隠れ層を 2 層持つ全 4 層のモデルであり，隠れ層 1 層目は 100，隠れ層 2 層目は 50 のニューロンを保持しているものとした．バッチサイズを 100 としてバッチ学習を行い，削減前に 15 回学習を行った．最適化には確率的勾配降下法を用い，学習率は 0.1 とした．また，重みの L2 ノルムによる削減率を 0.5，重みをグループ化した領域による削減率を 0.4 として削減を行った．

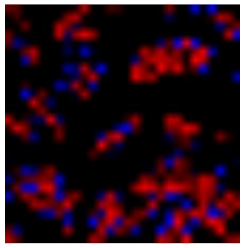


図 3　削減後のニューロンの重み(赤青部分)

図 4　ニューロンの重みを含む領域(白部分)

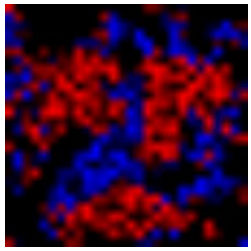


図 5　約 35%削減されたニューロンの重み

4.2 結果

本提案手法により，各ニューロンの重みの削減率が 9 割以上のものと，4 割以下のものに分かれた．それぞれの内訳は 9 割以上のものが 54%，4 割以下のものが 46%となった．この結果，精度は削減処理前が 94.5%，削減処理後が 94.1%と大きく劣化せずに削減できた．この場合の一例として，約 35%削減されたニューロンの重みを図 5 に示す

5. おわりに

本稿では，ニューラルネットワークの出力結果に影響を与える重要なパラメータを示すことができた．しかし，削減され，残った領域はニューロンの持つ重みの半分以上を占める領域となっていた．そのため，ニューロンが入力を分割して注目している領域を示すことができなかった．これはすべての領域を対象として指標値の小さい順に重みを削減したためだと考えられる．そのため，どのような基準で重みの削減を行うかが今後の課題である．

参考文献

[1] Song Han, Jeff Pool, John Tran, William J. Dally. Learning both Weights and Connections for Efficient Neural Networks. (2015)

[2] Yann LeCun, Corinna Cortes, Christopher J.C. Burges. THE MNIST DATABASE of handwritten digits. http://yann.lecun.com/exdb/mnist/