論文の内容の一部訂正とお詫び

大分大学大学院工学研究科
長谷部圭亮

2019 年 12 月 11 日

下記論文中のデータを取得したシミュレーションに誤りがありました。ここに訂正させていただくとともに、間違ったデータを論文に掲載していましたことをここに深くお詫び申し上げます。ただし、論文で主張していること自体が覆る訂正ではありませんので、本論文で主張していることは重要であるとの認識の下、引き続き主張して参りたいと存じます。

[訂正対象論文]
第 38 回計測自動制御学会九州支部学術講演会予稿集
長谷部圭亮, 柴田克成
「 多層ニューラルネットにおける勾配消失問題解決法としての感度調整学習 」
pp.87−90(2019)

[訂正概要]
違う問題の結果を掲載していまいました。これを修正したことに伴い, Fig.2、Fig.3、Fig.4、Fig.5 に掲載した結果, および、本文中におけるその説明箇所に変更が入ることになりました。

[訂正内容]

- Fig.2、Fig.3、Fig.4、Fig.5 の図の差し替えを行いました。
- 3 ページ左コラム Table 1
  (訂正前)
  試行回数 200
  (訂正後)
  試行回数 300
- 3 ページ左コラム 3.1 節最後から 2 行目
  (訂正前)
  入力が奇数のとき-0.8 、偶数のとき 0.8 として
  (訂正後)
  入力の個数が奇数のとき 0.8 、偶数のとき-0.8 として
- 3 ページ右コラム上から 1 行目
  (訂正前)
  約 75epoch まで
  (訂正後)
  約 125epoch まで
- 3 ページ右コラム上から 16 行目
  (訂正前)
  8epoch あたりで各ニューロンの感度が 1 を超えて感度調整学習が終了するが、その後は BP 法の学習により誤差信号が変動していることがわかる。
  (訂正後)
  約 8epoch あたりで、最下層まで誤差信号が伝播できていることがわかる。
- 3 ページ右コラムの最後の段落の上から 3 行目
  (訂正前)
  200epoch 時の入力全パターンの
  (訂正後)
  300epoch 時の入力全パターンの
- 4 ページ右コラム 4 章の下から 2 行目
  (訂正前)
  感度調整学習を導入できるかどうかの検証等が課題である。
  (訂正後)
  感度調整学習を導入できるかどうかの検証等である。

多層ニューラルネットにおける勾配消失問題解決法としての感度調整学習

大分大学　○長谷部圭亮　柴田克成

Sensitivity Adjustment Learning as a Solution to the Vanishing Gradient Problem in Multilayer Neural Networks

Keisuke Hasebe and Katsunari Shibata, Oita University

Abstract: In this paper, we show that the method called "sensitivity adjustment learning" proposed by our group is effective to solve the "vanishing gradient problem". The sensitivity is defined in each neuron as the magnitude of the output gradient with respect to the input vector. In this learning, the weights are updated to increase the sensitivity until it reaches 1.0. We apply this learning method to 5 bit parity problem using a 100-layer neural network. In the case of only Back Propagation learning, it is difficult to find appropriate initial weight scale, but together with this learning, the training was done successfully in any case when the initial weight scale is small.

1　序論

近年、人工知能 (AI) の研究が盛んに行われている。その中でも、多層のニューラルネットを用いる深層学習 (Deep Learning) による画像認識や自然言語処理等の技術が発展し、自動運転や機械翻訳などのような技術革新に繋がっている。[1][2] Deep Learning は、多くの場合、ニューラルネットを多層化することで、畳み込みニューラルネットのような局所的な処理を積み上げて画像等を効率的に処理し、さらに、全結合のネットワークを積み上げている。この多層化による情報の抽象化と汎化により、人間が作成するプログラムでは実現が困難なパターン認識等を可能にしている。

しかし、Deep Learning には勾配消失 (vanishing Gradient) という問題点がある。ニューラルネットで広く用いられる誤差逆伝播 (BP) 法では、誤差信号をネットワーク内で伝播させることで学習を行う。しかし、BP 法では、ニューラルネットを多層化すると、層を通るたびに重み値と活性化関数の微分値をかけるため、誤差信号が消失、発散してしまうという問題である。

現在、勾配消失問題の解決法として、Auto Encoder(AE)[3] や Residual Network(ResNet)[4] などがある。AE とは、エンコーダにより入力の次元を圧縮し、重み値をデコーダで元の入力を復元できるように学習する手法である。多層ニューラルネットでは、AE によって初期重み値を決めることにより、下層の特徴量を抽出しつつ、情報の伝播を確保することができる。[3] 一方、ResNet とは、いくつかの層毎にショートカット接続をすることで、多層化した際に、ネットワークが下位層からの出力を減衰せずに伝播することができるとともに、BP 法による誤差信号の減衰もショートカット接続により防ぐことができる。[4] しかし、これらの勾配消失問題の解決法は、事前学習が必要であったり、ネットワークの構造に制限を持っている。

本論文では、まず、「感度調整学習」と呼ばれる学習方法を提案する。感度調整学習とは、各ニューロンの入力の微小変化に対する出力の微小変化の割合の最大値である「感度」の値を調整するように重み値を学習する方法である。この手法のメリットとして、ネットワークの構造に制限がないということと、ネットワークの学習において他の学習と同時に行えることがある。また、感度調整学習は個々のニューロンがローカルな情報だけで学習を行うことができるため、ハードウェア化するときにも大きな利点を有する。

本研究ではさらに、感度調整学習で各ニューロンの感度が 1.0 近辺の値になるように学習することで、ネットワークを多層化した際に、入力情報を減衰せずに出力を得ることができ、また、誤差信号も同様に減衰なく伝播することができるのではないかと考えた。それを検証するため、100 層のニューラルネットを用いて教師あり学習を行い、初期重み値の大きさを小さめにしておけば特段気にすることなく学習できることを示すとともに、その際に学習がどのように進むかを示す。

2　感度調整学習

ここでは、Fig.1 のような階層型のネットワークを考える。Fig.1 のネットワークに入力を入れた際の l 層 j 番目のニューロンの内部状態 $u_j^{(l)}$ とそのときの出力 $o_j^{(l)}$ の計

算式を下記に示す。今回、出力 $o_j^{(l)}$ の計算に用いた活性化関数は tanh 関数である。

$$
\begin{aligned}
u_j^{(l)} &= \boldsymbol{w}_j^{(l)} \cdot \boldsymbol{o}^{(l-1)} && (1)\\
o_j^{(l)} &= \tanh(U_j^{(l)}) \\
&= \tanh(u_j^{(l)} + b_j^{(l)}) && (2)
\end{aligned}
$$

$$
\begin{aligned}
\boldsymbol{w}_j^{(l)} &: \ l \text{ 層 } j \text{ 番目のニューロンの重み値ベクトル} \\
\boldsymbol{o}^{(l-1)} &: \ l-1 \text{ 層のニューロンの出力ベクトル} \\
b_j^{(l)} &: \ l \text{ 層 } j \text{ 番目のニューロンのバイアス} \\
U_j(l) &= u_j^{(l)} + b_j^{(l)}
\end{aligned}
$$

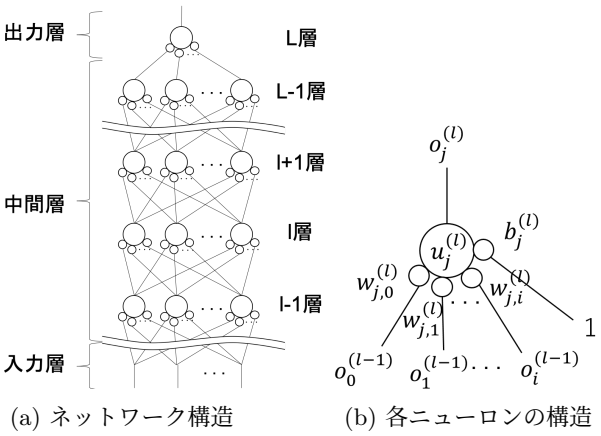
感度調整学習では個々のニューロンが、下の層からの入力 (ベクトル)$\boldsymbol{o}^{(l-1)}$ に対する出力 $o_j^{(l)}$ の感度 $\left|\nabla_{\boldsymbol{o}^{(l-1)}} o_j^{(l)}\right|$ が 1 より大きくなるように重み値とバイアスの学習を行う。この感度 $\left|\nabla_{\boldsymbol{o}^{(l-1)}} o_j^{(l)}\right|$ は以下のように表される。

$$
\begin{aligned}
\left|\nabla_{\boldsymbol{o}^{(l-1)}} o_j^{(l)}\right| &= \sqrt{\sum_{i=1}^{N^{(l-1)}} \left(\frac{\partial o_j^{(l)}}{\partial o_i^{(l-1)}}\right)^2} \\
&= f'(U_j^{(l)}) \sqrt{\sum_{i=1}^{N^{(l-1)}} w_{j,i}^{(l)2}} \\
&= f'(U_j^{(l)}) \left|\boldsymbol{w}_j^{(l)}\right| && (3)
\end{aligned}
$$

$N^{(l-1)} : l-1$ 層目のニューロン数

感度調整学習では、感度が上がるように重み値を学習するため、感度を評価関数として最急上昇法を用いて重み値 $\boldsymbol{w}_j^{(l)}$ とバイアス $b_j^{(l)}$ の更新を行う。そして、感度が 1.0 を超えたニューロンは感度調整学習を止める。下記に感度調整学習における重み値の更新量 $\Delta \boldsymbol{w}_j^{(l)}$ とバイアスの更新量 $\Delta b_j^{(l)}$ の計算式を示す。

$$
\begin{aligned}
\Delta \boldsymbol{w}_j^{(l)} &= \eta \nabla_{\boldsymbol{w}_j^{(l)}} \left|\nabla_{\boldsymbol{o}^{(l-1)}} o_j^{(l)}\right| \\
&= \eta \frac{f'(U_j^{(l)}) \boldsymbol{w}_j^{(l)} + \left|\boldsymbol{w}_j^{(l)}\right|^2 \nabla_{\boldsymbol{w}_j^{(l)}} f'(U_j^{(l)})}{\left|\boldsymbol{w}_j^{(l)}\right|} \\
&= \eta (1 - o_j^{(l)2}) \left(\frac{\boldsymbol{w}_j^{(l)}}{\left|\boldsymbol{w}_j^{(l)}\right|} - 2 o_j^{(l)} \left|\boldsymbol{w}_j^{(l)}\right| \boldsymbol{o}^{(l-1)} \right) && (4)
\end{aligned}
$$



(a) ネットワーク構造　　(b) 各ニューロンの構造

Fig. 1: 多層ネットワークとニューロンの構造

$$
\begin{aligned}
\Delta b_j^{(l)} &= \eta \frac{\partial \left|\nabla_{\boldsymbol{o}^{(l-1)}} o_j^{(l)}\right|}{\partial b_j^{(l)}} \\
&= \eta \frac{\partial f'(U_j^{(l)})}{\partial b_j^{(l)}} \left|\boldsymbol{w}_j^{(l)}\right| \\
&= -2\eta o_j^{(l)} (1 - o_j^{(l)2}) \left|\boldsymbol{w}_j^{(l)}\right| && (5)
\end{aligned}
$$

感度調整学習と同時に行う BP 法では、出力層の出力 $o_j^{(L)}$ と教師信号 T_j の二乗誤差 E が 0 になるように重み値を更新する。まず、BP 法における出力層 L 層 j 番目のニューロンの誤差信号 $\delta_j^{(L)}$ と中間層 l 層 j 番目のニューロンの誤差信号 $\delta_j^{(l)}$ を下記に示す。また、式 (7) では、$l+1$ 層 j 番目のニューロンの誤差信号 $\delta_k^{(l)}$ の発散を防ぐために、tanh 関数を通して求めた。

$$
\begin{aligned}
\delta_j^{(L)} &= (T_j - o_j^{(L)})(1 - o_j^{(l)2}) && (6)\\
\delta_j^{(l)} &= \tanh\left(\sum_{k=1}^{N^{(l+1)}} \delta_k^{(l+1)} w_{k,j}^{(l+1)} (1 - o_j^{(l)2}) \right) && (7)
\end{aligned}
$$

誤差関数 E が小さくなるよう確率的最急勾配法を用いて、重み値 $\boldsymbol{w}_j^{(l)}$ とバイアス $b_j^{(l)}$ を学習する。このときの重み値の更新量 $\Delta \boldsymbol{w}_j^{(l)}$ とバイアスの更新量 $\Delta b_j^{(l)}$ の計算式を下記に示す。

$$
\begin{aligned}
\Delta \boldsymbol{w}_j^{(l)} &= \eta \delta_j^{(l)} \boldsymbol{o}^{(l-1)} && (8)\\
\Delta b_j^{(l)} &= \eta \delta_j^{(l)} && (9)
\end{aligned}
$$

実際には、感度調整学習と BP 法を同時に行うので、重み値は式 (4) と式 (8) の和、バイアスは式 (5) と式 (9) の和にしたがってパターンを提示するたびに更新する。

3　シミュレーション

3.1　方法

本研究では、Fig.1(a) のようにネットワークを多層化した際に、感度調整学習を用いることで、出力層から中間層の最下層にまで誤差信号が消失、発散せずに伝播し学習できているかどうかを検証する。ネットワークの学習法として、感度調整学習と BP 法を同時に行い重み値の更新を行う。本研究で使用したネットワークの層数は 100 層、出力層のニューロン数は 1 個、中間層の各層のニューロン数は 20 個で行った。今回使用したパラメータの詳細は Table 1 に示す。また、勾配消失問題を明確にするため、初期重み値は、あえて [-0.01, 0.01] と小さめに設定した。

Table 1: 今回使用したネットワークのパラメータ

パラメータ名	設定値
ネットワーク層数	100
出力層のニューロン数	1
各中間層のニューロン数	20
試行回数	300
出力関数の値域	(-1.0,1.0)
誤差逆伝播法の学習係数	0.0004
感度増大学習法の学習係数	0.004
初期重み値	[-0.01,0.01] の一様乱数
初期バイアス	[-0.01,0.01] の一様乱数

本研究では、学習できることがわかっている 5bit パリティ問題を学習する。入力は Table 2 のように 5 個の 2 値の数値であり、全部でその組み合わせは $2^5 = 32$ パターンある。バッチ学習は行わず、1 パターン提示してはその都度、前述の確率的最急降下 (SGD) 法に基づく BP 法で学習する。教師信号は、入力に 1 の値が奇数個のとき 0.8、偶数個のとき-0.8 として学習を行った。

Table 2: 5bit パリティ問題の入力と教師信号

入力 1	入力 2	入力 3	入力 4	入力 5	教師信号
-1	-1	-1	-1	-1	-0.8
-1	-1	-1	-1	1	0.8
⋮					
1	1	1	1	-1	-0.8
1	1	1	1	1	0.8

3.2　実行結果

Fig.2 に、まず学習の進行による全 32 パターンに対する出力の変化を示す。縦軸は出力層の出力、横軸が学習回数 (epoch) を表わす。約 125epoch までは出力が変動しているが、約 150epoch 以降は全パターンが教師信号とほとんど等しい値を出力することができており学習できていることが確認できた。

次に勾配消失問題がどのように解決されているのかを見るために、学習中における中間層 1 層、20 層、40 層、60 層、80 層、出力層のそれぞれの誤差信号ベクトルの大きさ $\sqrt{\sum_j \delta_j^{(l)2}}$ の変化を観察した結果を Fig.3 に示す。縦軸は誤差信号の大きさ、横軸が学習回数を表わす。また、学習初期の変化を見るため、横軸を拡大している。学習開始時は入力層に近づくほど、誤差信号の大きさは小さく、一番入力に近い中間層 1 層目では約 10^{-150} と非常に小さくなっており、初期重み値を小さく設定したため誤差の勾配情報の消失が激しいことがわかる。学習が進むにつれて、各中間層の誤差信号の大きさは急激に大きくなって、約 8epoch あたりで最下層まで誤差信号が伝播できていることがわかる。

Fig.4 は、学習中における中間層 1 層、20 層、40 層、60 層、80 層、出力層のそれぞれの各ニューロンの重み値ベクトルの大きさの平均 $\left|\boldsymbol{w}_j^{(l)}\right|$ がどう変化するかを観察した。縦軸は重み値、横軸が試行回数を表わす。20 層より上の層の値はほとんど重なっているが、いずれも学習とともに重み値は大きくなり層間であまり大きな差はないことがわかる。

さらに初期重み値の乱数の大きさによる影響を BP 法のみの場合と、感度調整学習を同時に行った場合で比較した。乱数系列を変えて 10 回学習し、300epoch 時の入力全パターンの 10 回分の平均における教師信号と出力層の出力の誤差の絶対値の平均と標準偏差を、横軸を初期重み値の大きさとしてプロットしたものを Fig.5 に示す。これより、初期重み値の乱数の範囲が 0.4 より小さいとき、BP 法のみでは、誤差信号が最下層まで届かないため、出力はいずれのパターンでも 0 となって誤差信号が 0.8 となり、まったく学習できなかった。一方で、感度調整学習を一緒に行った場合は逆にいずれの場合も誤差が 0 になっていることがわかる。初期重み値が 0.5 以上のときは、感度調整学習を行った場合は、初期重み値の増加とともに学習できなくなる一方で、BP 法のみの場合には、学習ができる場合もあったが、10 回の学習すべてでうまく学習できる場合はなかった。さらに初期重み値を細かく刻めば、すべて学習できる場合もあったかもしれないが、いずれにしても適切な初期重み値を探すことが困難であることがわかる。

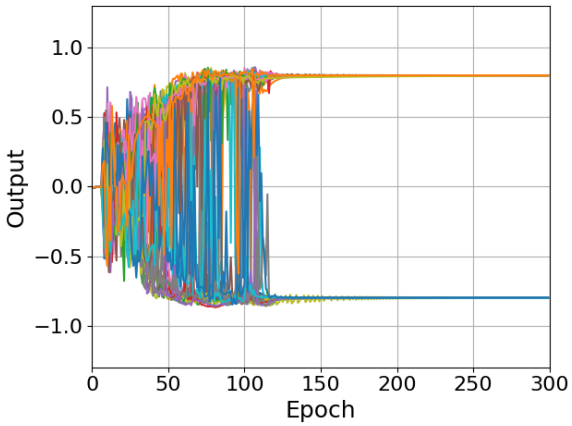


Fig. 2: 感度調整学習と BP 法を併用した場合の学習による 5bit パリティ問題の全 32 パターンの出力の変化

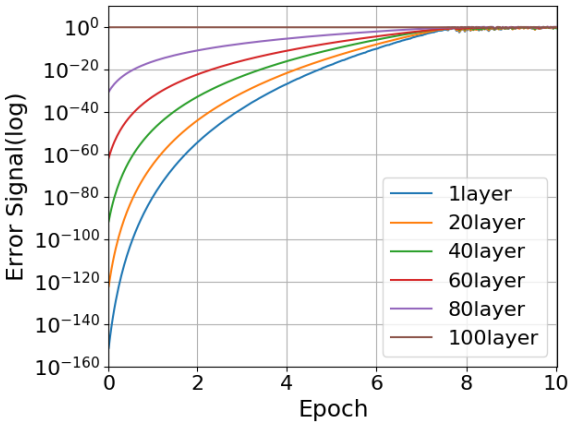


Fig. 3: 学習初期におけるいくつかの層の誤差信号の大きさの変化

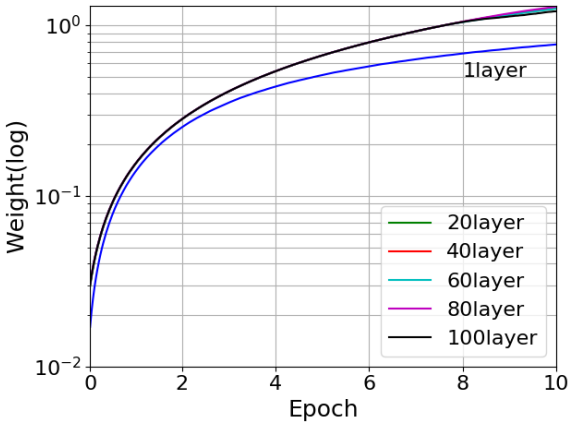


Fig. 4: 学習初期におけるいくつかの層の各ニューロンの重み値の大きさの平均値の変化

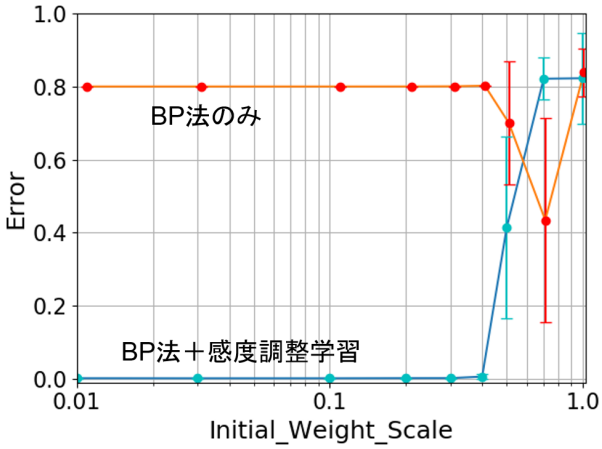


Fig. 5: 初期重み値の乱数の大きさを変更したときの、BP 法のみと BP 法と感度調整学習の両方で 10 回学習した際の誤差の絶対値の平均と標準偏差の比較

4 結論

本研究では、個々のニューロンの「感度」を調整する感度調整学習を誤差逆伝播 (BP) 法と併用することによって、多層ニューラルネットにおける勾配消失を解決することを提案した。100 層の多層ニューラルネットの学習において、BP 法だけの場合では、誤差勾配消失のため学習がうまくできる初期重み値の大きさを探すことは困難であったが、感度調整学習を併用して行うことで、初期重み値を小さくしておけばいずれの場合も学習できることを示した。今後の課題は、時系列データを扱うリカレントネットにも感度調整学習を導入できるかどうかの検証等である。

5 謝辞

本研究は JSPS 科研費 15K00360 の助成を受けたものである。ここに謝意を表する。

参考文献

[1] 我妻広明: 人工知能による運転支援・自動運転技術の現状と課題, 計測と制御, 第 54 巻, 第 11 号,2015

[2] J. Devlin, R. Zbib, Z. Huang, T. Lamar, R. Schwartz, and J. Makhoul, “ Fast and robust neural network joint models for statistical machine translation, ” Proc. ACL, pp.13701380, Baltimore, Maryland, USA, June 2014.

[3] Hinton, G.E. and Salakhutdinov, R.R. Reducing the dimensionality of data with neural networks. Science, 313 (5786):504507, 2006.

[4] He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR. (2016)