3H4-OS-24b-4

ディープニューラルネットワーク内ダイナミクスの力学的解析

Analysis of Deep Neural Network Using by Dynamical Systems Analysis

本武 陽一[*1] 岡 瑞起[*2] 池上 高志[*1]

Mototake Yhoichi Oka Mizuki Ikegami Takashi

[*1] 東京大学総合文化研究科
Graduate School of Arts and Science, The University of Tokyo

[*2] 筑波大学大学院システム情報工学研究科
Graduate School of Systems and Information Engineering, University of Tsukuba

Since Hinton et al. (2006) came back with a multilayered feed-forward network, called a deep neural network, many people have started to investigate its potential capability and applications. For example, Google Inc. showed that the deep learning automatically extracted cat face and human body images from the millions of randomly selected youtube images[Quoc 12] .In this study, we compute the information flow within a deep neural network in order to reveal the underlying dynamical systems properties. Unexpected power law behavior of Eigen values computed from the Jacobian matrices of the deep net will be reported.

1. はじめに

多層パーセプトロンにおけるバックプロパゲーションアルゴリズムの限界の発見以来，ニューラルネットの注目度は下がっていた。しかし，[Hinton 06]において有効な学習アルゴリズムが発見され，比較的簡便に深い階層を持ったニューラルネットワークを学習することが可能となった．さらに，その Deep Neural Network（以下、DNN）が，驚異的な認識精度を記録したことで，ニューラルネットワークは，再び脚光を浴びるようになっている．

しかしながら，なぜディープラーニング（以下、DL）がうまくいくのかといった基本的な問題は，未だ未解明の部分が多い．[Saxe 14]

本研究では，DNN においてネットワーク内のダイナミクスを力学的に分析することで，これらの問題へアプローチすることを試みる．

2. ディープラーニングのダイナミクス

DNN のダイナミクスとしては，次の 2 つのものが考えられる．1 つは，学習中の重みの時間発展である．もう 1 つは，図 1 のように DNN の各階層を時間に対応付け，層が進むに従って変化するニューロンの発火パターンの時間発展を考える視点である．

本研究では特に後者の視点を重視する．

ここで，この時のニューロンの発火の時間発展を，次式で定義するものとする．

$$h_j^{(t)} = sigmoid(g(\sum_i h_i^{(t-1)} w_{ij}^{(t)} + (Bias)_j^t)) \qquad (1)$$

g はゲインを表わす．

ところで，DL と一口に言っても，関連する技術の範囲は広い．従って，本研究では，どの要因がどの程度、そしてどのようにパフォーマンスの向上に貢献しているかを知る為に，要素毎にそれぞれの性質を調べることを考える．

一方で，比較対象として，複雑な学習を実現している，多数の要素を組み込んだ条件でも分析する．

従って本研究では，両者のアプローチをそれぞれ採用する．前者のアプローチとしては，以下の項目のような要因について，それぞれ一つずつ分析する．

① 学習アルゴリズム（drop out ,pooling etc.）
② インプットデータの種類（手書き文字,画像 etc.）
③ ネットワークの構造（各層のノード数 etc.）

一方，後者の分析では，分析対象として Convolutional Neural Net や drop out[Hinton 12]等が提供されている DL ライブラリである DeCAF[Donahue 13]と，そこで提供されている pre training 済み重みデータセットを用いる．

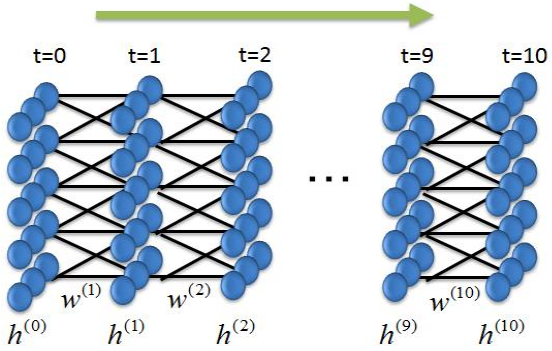

図 1.階層方向の時間発展

3. 先行研究

3.1 pre training のメカニズム

DNN における，図1のようなダイナミクスを対象とした研究として，[Ganguli 14]がある．この研究では，特に pre training に着目し，各種近似のもと，無限層のニューラルネットのダイナミクスを解析的に求めている．その上で，100 層からなる DNN を Restricted Boltzmann Machine（以下，RBM）を用いて pre training し，これらを合わせて pre training は，重み行列が直交行列になるよう,初期化していることに対応すると結論した．同時に，このような初期値において，無限階層のニューラルネットワークの学習が，有限時間で収束することも示してもいる．

しかしながらこの研究では，入力データに対して，直交性を仮定している．また，実際のデータによる計算でも，比較的単純な手書き文字データセット（MINIST[LeCun 98]）を用いたシュミレーションのみに終わっている．

mototake@sacral.c.u-tokyo.ac.jp

3.2 カオスの縁と pre training

リカレントニューラルネットワークでの知見として，「カオスの縁」と呼ばれる，系の相転移点において，高い学習性能が実現されうるというものがある[Bertschinger 04]. 先行研究では、このコンセプトを DL に適用し，(1)式のパラメータ g による，output 層周辺での分散特性の相転移に注目し，これと，ネットワーク全体の特異値との関係を分析している．特異値は，入力層での微小変化が，出力層にどれだけ伝わるかを表わす値であり，up-down path での重み行列が転置関係の場合，Back Propagation（以下 BP）では，この特異値が，微小変化の方向によらず O(1)程度であることが，エラーの伝搬に有用であるとわかる．

計算の結果，g=1 付近で相転移が生じ，かつ g<1 の場合，特異値は総じて小さく，一方で g>1 の場合には，一部に大きな特異値を持つ一方，ほとんどの方向の特異値が非常に小さい値を持つ，べき的な偏った分布になることが示された．従って，これらの状態は BP に対して良い状態とはいえない．一方で，カオスの縁となる g≒1 付近では，O(1)程度の特異値を多く含む，なだらかな分布が現れた．この状態は，BP に対して適している．

以上のように， pre training と g の値によって，最適な初期状態が得られることが示されている．しかし，臨界指数となるパラメータは，g のみではないと考えられる[Bertschinger 04]．例えば，入力データの性質や，重みの分散などがそれである．先行研究でも，入力データの分散ごとに計算を行っている．しかしながら，実際のデータセットの分散は一定ではない．さらに，上の結果は実際の pre training を用いて示されたわけでもない．

3.3 先行研究のまとめ

以上をまとめると，先行研究では，pre training が重みを直交行列として初期化することに近しいことを示し，その初期化によって実現される状態において，「カオスの縁」周辺で BP に最適な状態が与えられることが示されている．しかし，先行研究では，多様なデータセットで pre training を行うことへの言及が不足している側面があると考えられる為，本稿ではこの点に着目した分析を行った．

4. 実験とその分析

MINIST 及び，より複雑な画像データセット(CIFAR-100[Krizhevsky 09])を使用し，[Hinton 06]に従って，RBM を用いた pre training を行った．ただし，(1)式にあるように，ニューロンは連続値ニューロンとし，ネットワークの階層を 10 とした．また，g=1.05，学習サンプル数を 12,800 とした． それによって得られた学習結果を用いて，以下のような手順で，特異値を計算した．(1)式より，

$$\therefore \frac{\partial h_j^{(t)}}{\partial h_i^{(t-1)}} = h_j^{(t)}(1-h_j^{(t)})w_{ij}^{(t)}$$

従って，各層間の変換の Jacobean は，以下のようになる．

$$J^{(t)} = \begin{pmatrix} \frac{\partial h_1^{(t)}}{\partial h_1^{(t-1)}} & \cdots & \frac{\partial h_1^{(t)}}{\partial h_N^{(t-1)}} \\ \vdots & \ddots & \vdots \\ \frac{\partial h_N^{(t)}}{\partial h_1^{(t-1)}} & \cdots & \frac{\partial h_N^{(t)}}{\partial h_N^{(t-1)}} \end{pmatrix}$$

$$= \begin{pmatrix} h_1^{(t)}(1-h_1^{(t)})w_{11}^{(t)} & \cdots & h_1^{(t)}(1-h_1^{(t)})w_{N1}^{(t)} \\ \vdots & \ddots & \vdots \\ h_N^{(t)}(1-h_N^{(t)})w_{1N}^{(t)} & \cdots & h_N^{(t)}(1-h_N^{(t)})w_{NN}^{(t)} \end{pmatrix}$$

ここから，ネットワーク全体の変換の Jacobean(J)は以下のように求められる．

$$J = J^t \cdot J^{t-1} \cdot \cdots J^0$$

この行列 J から J･J*を求め、これの非負の固有値（特異値）を求める．

この計算の結果を，図 2，図 3 に示す．この結果より，どちらもべき乗のような分布をしていることがわかる．しかも，特異値の値は非常に小さく，前述した理由から，良い初期化が実現される状態とは言えない．この原因としては，実データではサンプルによって分散が一定でないこと等が考えられるが，pre training がうまくいっていないことも考えられる為，原因については，さらなる分析とともに発表時に説明したい．

5. まとめ

本稿では、DL のダイナミクスとして、pre training 後に実現されているダイナミクスに着目し、分析した。結果として、BP などの情報伝搬がされやすい状態が実現される「カオスの縁」は、実際のデータでは同様な形では現れない可能性が示唆した。

発表ではさらに、CNN や drop out、正則化、トポロジ構造の違い等によって、このダイナミクスがどのように変化するか述べるとともに、DeCAF の学習済みパラメータデータを用いて、より実際的な条件での分析について論じる予定である．

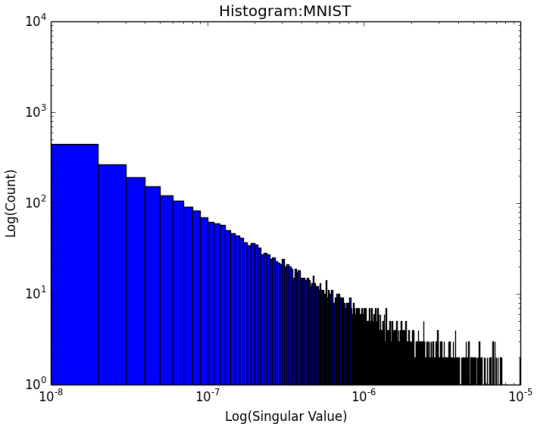

図 2.MNIST データの特異値分布

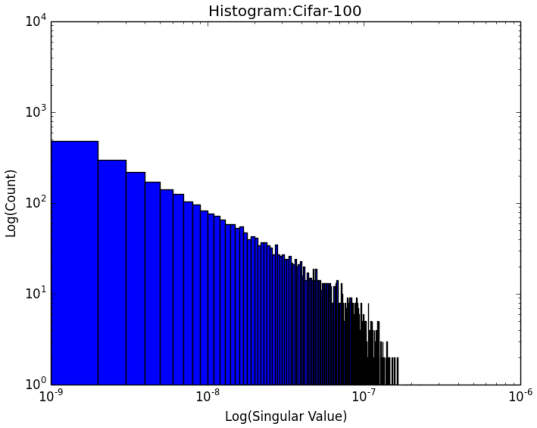

図 3.Cifar-100 の特異値分布

参考文献

[Bertschinger 04] Bertschinger, N. and Natschläger, T.Real-time computation at the edge of chaos in recurrent neural networks, Neural Computation, 16(7):1413-1436, 2004.

[Donahue 13] Donahue, J.,Jia, Y., Vinyals, O., Ning -Zhang, J., Tzeng,E., Darrell,T. DeCAF: A Deep Convolutiona-Activation Feature for Generic Visual Recognition, arXiv preprint arXiv:1310.1531 , 2013.

[Hinton 06] Hinton, G. E., Osindero, S. and Teh, Y. A fast learning algorithm for deep belief nets.Neural Computation, 18, pp 1527-1554, 2006.

[Hinton 12] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever and R. R. Salakhutdinov: Improving neural networks by preventing co-adaptation of feature detectors, arXiv:1207.0580v1 , 2012.

[Krizhevsky 09] Krizhevsky,A., Learning Multiple Layers o Features from Tiny Images, 2009.

[LeCun 98] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86(11):2278-2324, November 1998.

[Saxe 14] Saxe, A. M. , Berschinger, N., and Legenstein R. Exact solutions to the nonlinear dynamics of learning in deep linear neural network, NIPS Workshop on Deep Learning , 2013.

[Quoc 12] Quoc V. Le, Marc'Aurelio Ranzato, Rajat Monga, Matthieu Devin, Greg Corrado, Kai Chen, Jeffrey Dean, Andrew Y. Ng: Building high-level features using large scale unsupervised learning. ICML 2012