2C3-OS-06b-4in

Deep Neural Networksの力学的解析

The Analysis Of Deep Neural Networks

本武　陽一 *1　　池上　高志 *1
Mototake Yhoichi　　Ikegami Takashi

*1東京大学大学院総合文化研究科
Graduate School of Arts and Science, The University of Tokyo

Multilayered feed-forward networks, commonly known as deep neural networks (DNN)[Hinton 06] , have been intensively studied their potential capabilities and mechanisms. For example, DNN classifies cat faces and human body images by learning millions of randomly selected Youtube images [Quoc 12]. In this study, we computed the information flow of a DNN in order to reveal its underlying mechanism with respect to dynamical systems. Our results support the hypothesis that the high performance of DNN can be characterized by the changes of singular value distribution along the layers. Irrelevant differences between input images will be shrunk and the important difference will be expanded by the DNN. This picture will be investigated thoroughly in this work.

1.　はじめに

Hinton らによる，多層フィードフォワードニューラルネットワーク（DeepNeuralNetworks：以下 DNN と略記）の有効な学習法の発見 [Hinton 06] 以来，DNN の特性や，高い学習性能を活用する研究が数多く行なわれてきた．

例えば，Quoc らは，youtube からランダムに抽出した大量の画像を DNN に学習させることで,「猫の顔」といったカテゴリを自動で抽出することに成功した [Quoc 12]．また，Szegedy らは，10 以上の層を持たせた DNN を用いることで，非常に高い画像認識の精度を達成している [Szegedy 14]．

本研究の目的は，これらの DNN の性質が，どのように獲得されるのかを解明することである．

2.　DNN の力学的解析

DNN のダイナミクスとして，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．本研究では後者の視点から分析を行なった．

従って，ニューロン発火の時間発展は，次式で定義される．

$$h_j(t+1) = f(\sum_i (h_i(t) \cdot W_{ij}(t)) + B_j(t)) \tag{1}$$

$f(x)$ としては，よくシグモイド関数，

$$f(x) = 1/(1+e^{-gx}) \quad (g : const) \tag{2}$$

が使われる．ここで，$h_i(t)$ は t 層の隠れ層のノード状態を，$W_{ij}(t)$ は，t 層から $t+1$ 層の間の重み行列を，$B_j(t)$ は，第 t 層のバイアス値を表すものとする（図 1 参照）．また，i,j は，各層でのノードのインデックスになっている．

この時間発展方程式は，図 2 のような画像の 1 ピクセルを 1 次元とする空間上で，画像に対応する粒子が，層を発展するとともにどのように移動するかを表現している．

連絡先: 本武陽一，東京大学大学院総合文化研究科，mototake@sacral.c.u-tokyo.ac.jp

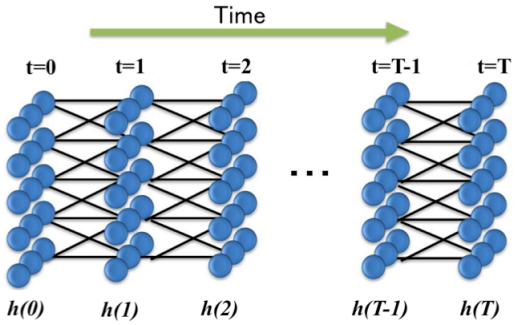

図 1: ニューラルネットワークの時間発展：左から右に層を進んで行くことを時間発展の方向と考える．

この時間発展に対して，第 t 層における粒子位置の摂動に対する，$t+1$ 層での変動を表すヤコビアン行列が，以下で定義される．

$$J(t) = \begin{pmatrix} \frac{\partial h_1(t+1)}{\partial h_1(t)} & \cdots & \frac{\partial h_1(t+1)}{\partial h_N(t)} \\ \vdots & \ddots & \vdots \\ \frac{\partial h_N(t+1)}{\partial h_1(t)} & \cdots & \frac{\partial h_N(t+1)}{\partial h_N(t)} \end{pmatrix} \tag{3}$$

このヤコビアンの特異値・特異ベクトルを求めることによって，どの方向（特異ベクトル）への摂動が保存され（特異値 > 1），どの方向が消去される（特異値 $<< 1$）かが分かる．正確な例ではないが，わずかに違う 2 つの「1」という手書き文字の差分を摂動と考えた場合，その摂動が消去されることは，小異によらない「1」という文字のカテゴリを形成するようなダイナミクスが働いていることを示唆する．

ネットワーク全体でのヤコビアン J_{all} は，各層のヤコビアンの積として，次式のように表される．

$$J_{all} = J(0) \cdot J(1) \cdots J(T-1) \tag{4}$$

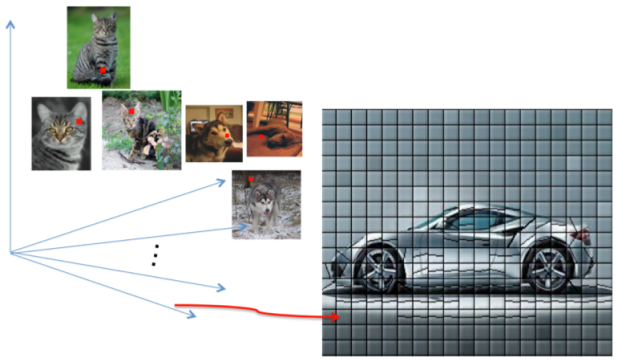

図 2: ダイナミクスを考える空間：画像が 1 点で表される空間でダイナミクスを考える．

3. 先行研究

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 DNN のダイナミクスを解析的に求め，さらにそれを，100 層からなる DNN を Restricted Boltzmann Machine（以下，RBM）を用いて検証している．その結果，pre training は，重み行列を直交行列になるよう初期化していることに対応すると，結論している．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た．

また，ダイナミクスという視点からではないものの，本研究と同様にヤコビアンの特異値計算を行なうことで，Bengio らは，deep learning が多様体学習の一種であることを示唆している [Bengio 13]．具体的には，Contractive-Auto-Encoder の特異値分布を調べ，その分布が，一部の大きな特異値と，多くの小さな特異値からなる急峻な分布となっていること，そして，他のアルゴリズムとの比較の結果，そのような分布となるアルゴリズムの方がパフォーマンスが高いことを示し，deep learning が，入力データセットの分布する，低次元の多様体をとらえるように学習を行なうことで，高いパフォーマンスを得ていることを示唆した [Rifai 11]．

しかし，Bengio らの研究は，教師なし学習アルゴリズムに対して行なわれており，画像認識等の実際の応用で活用されることの多い，教師あり学習でも同様なことがいえるかについては，不明確である．

従って，本研究では，教師あり学習アルゴリズムにおいても，DNN が低次元の多様体をとらえているのかを，実際に画像認識で活用されているネットワークのダイナミクスを解析することで．調べることを目的とした．

4. 実験方法

本研究では，Krizhevsky らによって開発された，畳み込みや pooling，drop out 等の技術を組み込んだ DNN [Krizhevsky 12] を分析対象とした．具体的には，Imagenet [11] によって学習されたネットワークの，公開されている重みデータ（DeCAF [Donahue 14]）を用い，これに Imagenet データセット [Deng 09] の画像を入力した場合のヤコビアン行列と，その特異値・特異ベクトルを算出した．特異値・特異ベクトルは，16 の違う入力画像に対して，SVD(singular value decomposition) を用いて上位 500 番目まで計算した．

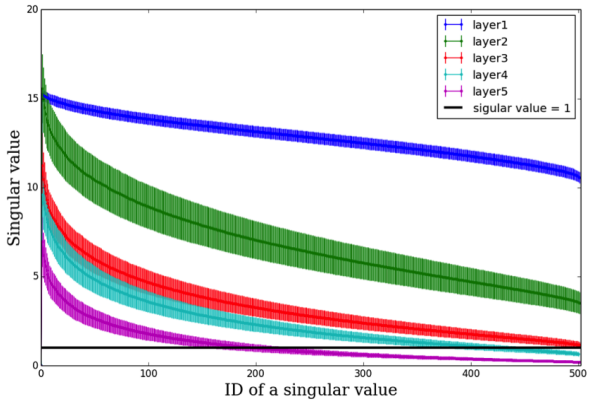

図 3: 特異値分布：それぞれの色に対応する線が，1 層から各層までのヤコビアンの特異値分布（16 の違う入力画像に対する平均）を表している．薄い色は 16 の違う入力画像に対する標準偏差を表している．

ただし，このネットワークにおける活性化関数 $f(x)$ は，下式で定義される．

$$f(x) = \begin{cases} x & (x \geq 0) \quad (5a) \\ 0 & (x < 0) \quad (5b) \end{cases}$$

5. 結果と考察

計算の結果，高次の層において，少数の大きな特異値と，大多数のほぼ 0 の特異値という，[Rifai 11] の結果に類似した急峻な特異値分布が見られた（図 3 参照）．また，特異ベクトルをみた結果，特異値の大きいベクトル程，空間的に局所的な構造を持ち，一方で特異値の小さなベクトル程，空間的に広く分布した構造をもっていることもわかった（図 4 参照）．

これらの結果は，教師あり学習においても，DNN がデータの埋め込まれた低次元の多様体を捉えていることを示唆する．さらに，空間的に広く分布する情報を削除し，局所的な情報を強調することによって，それが実現されていると考えられる．

ここで，誤判別と特異値分布の関係を調べる為に，入力画像にノイズを付加した上で，特異値の算出を行なった．すると，高次の層において，ノイズの増大に応じて特異値分布全体が小さくなっていくことが観察された（図 5 参照）．現時点で，この現象の原因は分かっていないが，これを理解する為に，特異値の大きな特異ベクトルの方向と，小さな特異ベクトルの方向に沿ってノイズを付加した場合の分析を行なうこと等を予定している．

6. まとめと議論

本研究によって，教師あり学習においても，deep learning が低次元の多様体を抽出していることが示唆された．また，ノイズによって識別精度が下がる際は，高次の層において，特異値分布全体が小さくなることもわかった．

これらが正しければ，多くの場合に困難な DNN のハイパーパラメータの探索 [Bengio 12] において，特異値分布の情報を活用することが有用であると考えられる．

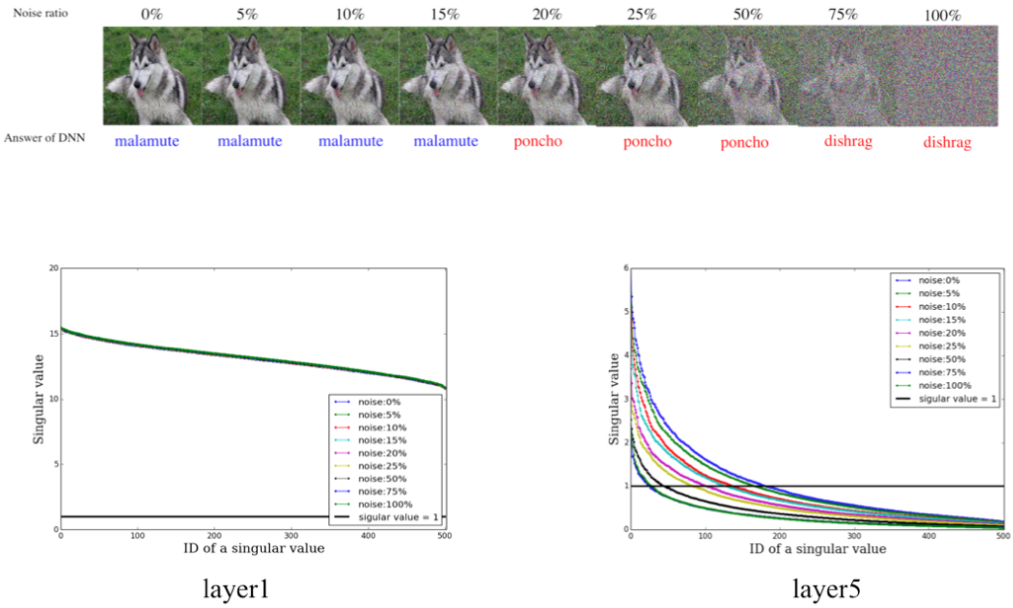

図 5: ノイズと特異値分布：1 段目がノイズの付加と判別結果の関係を，2 段目が，1,5 層でのノイズの付加率と特異値分布の関係を表す.

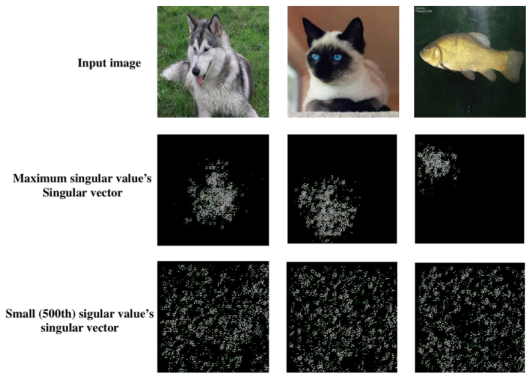

図 4: 特異ベクトル：1 段目が入力画像を，2 段目が特異値が最大となる特異ベクトルを，3 段目が計算した中で特異値が最小となる特異ベクトルを表す.

参考文献

[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.

[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.

[Szegedy 14] Szegedy, Christian, et al. "Going deeper with convolutions." arXiv preprint arXiv 1409.4842, 2014.

[Saxe 13] 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.

[Bengio 13] Y. Bengio, A. Courville, P. Vincent, Representation Learning: A Review and New Perspectives, IEEE Transactions on Pattern Analysis Machine Intelligence, vol.35, no. 8, pp. 1798-1828, Aug. 2013.

[Rifai 11] Rifai, S., Vincent, P., Muller, X., Glorot, X., and Bengio, Y. (2011a). Contractive auto-encoders: Explicit invariance during feature extraction. In ICML, 2011.

[Krizhevsky 12] Krizhevsky, A., Sutskever, I., and Hinton, G. E. ImageNet clas- sification with deep convolutional neural networks. In NIPS, 2012.

[Hinton 12] G. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever ,and R.R. Salakhutdinov. Improvingneural networks by preventing co-adaptation of feature detectors. http://arxiv.org/abs/1207.0580, 2012.

[Donahue 14] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, T. Darrell. DeCAF:A Deep Convolutional Activation Feature for Generic Visual Recognition. In Proceedings of the International Conference on Machine Learning (ICML), Beijing, China, June 2014.

[Deng 09] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In CVPR09, 2009.

[Bengio 12] Yoshua Bengio, Practical recommendations for gradient-based training of deep arXiv:1206.5533v2, 2012.