1A5-OS-27c-1in1

Deep Neural Networksの力学的・幾何学的解析

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

In recent years, it has become possible to gather large amounts of high dimensional datasets. On the other hand, not much knowledge about these datasets has been extracted yet. The purpose of this study is to get information about the geometrical structure of datasets distribution. We use multilayered feed-forward networks, now commonly known as deep neural networks (DNN)[Hinton 06], as an observation tool for datasets distribution. In this study, we computed the tangent space of dataset manifold from the mapping function of DNN to observe the geometrical structure of large scale image datasets (ImageNet dataset and MNIST dataset).

Our results support two hypotheses: (1) The high dimensional dataset distribution is embedded in a low dimensional manifold; (2) High performance DNNs have the ability to extract the manifold structure and map it on a global coordinate space. Furthermore, our results show that the semantic hierarchical structure of image (e.g., animal - mamal - dog - siberian husky) and the geometrical structure of the image dataset distribution are deeply related.

1. はじめに

近年，情報技術の急速な発展に応じて，これまでにない大規模で高次元なデータの収集が可能になってきた．同時に，これらのデータをいかに分析するかが問題となっている．一方で，我々の高次元データそのものに対する知見は，まだ十分にあるとは言い難い．

Deep Neural Networks（以下，DNN）は，Hinton らによる有効な学習法の発見 [Hinton 06] 以来，そ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，Simonyan 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 [Simonyan 14].

本研究の目的は，これら画像データセットや時系列データセットのような高次元データでトレーニングされた DNN を解析することで，その高いパフォーマンスを達成する原因を調べるとともに，データの性質についてのより精緻な情報を得ることである．

2. 多様体仮説

「この世界に存在するデータは, おおよそどのような性質を持つのであろうか?」

ここに，機械学習の応用の中で実証されてきた 1 つの仮説がある，機械学習における多様体仮説である．

ここでは具体的に，[Rifai 11] に基づき，多様体仮説を以下のように定義する．

- 仮説.1-1 高次元空間に存在する実世界のデータは，非常に低次元の非線形多様体付近に集中している．
- 仮説.1-2 高次元空間に存在する実世界のデータは，クラス (カテゴリ) 毎に違う部分多様体に埋め込まれており，それらの部分多様体の間は低密度領域 となっている．(分類問題に対する多様体仮説)

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

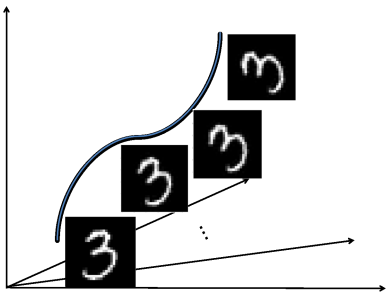

図 1: 回転に対する普遍性による多様体の形成

仮説.1-1 は，次のように説明される．実世界に存在するデータは，視点の変化や物体自身の運動，手書き文字データ 等であれば個人差などによって，少しずつ変化したものの集合となる．その変化 が連続的であれば，図 1 のように，同一クラスのデータが曲線や曲面上に分布すると考えられる．この曲面や曲線が多様体になると考えられるのである．

仮説.1-2 は，図 2 のように手書き数字データセット分布がそれぞれの数字毎に別々の連結空間に分離されることと説明できる．ちなみに，図 2 は手書き数字 データセット (MNIST) を，多様体学習アルゴリズム (t-SNE) を利用 して 3 次元空間に圧縮した結果である．

これに基づいて開発された多くの機械学習アルゴリズムが，高いパフォーマンスを達成できることから，仮説の妥当性が経験的に支持されている．

一方で，直接的にこの仮説の妥当性の検証に取り組んだ研究はこれまでにあまりなく，仮説成立の成否についての答えはまだ得られていない．

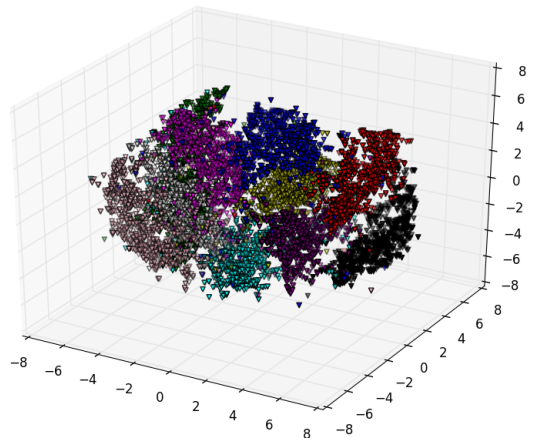

図 2: t-SNE によって 3 次元に圧縮された MNIST データセットの分布構造．それぞれの色がそれぞれの数字に対応．

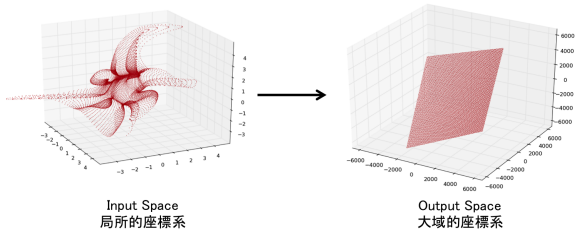

図 3: 大域的座標系：図の変換では, 局所的な座標系の集合として表現される多様体 (左図赤点) が, 位置によらず, 同じ座標系を共有する大域的な座標系へ変換されている (右図赤点)

3. 多様体仮説と DNN

それでは，このようなデータセットの幾何構造はどのような手法で観測できる だろうか．

本研究では，データセットの幾何構造の観測装置として，近年注目される機械学習技術である Deep Neural Networks(以下,DNN) を用いる．DNN は，以下の仮定が満たされる場合，前節で説明したデータセットの多様体構造を観測する装置として使用可能となる．

- 仮説.2 パフォーマンスの高い学習済み Deep Neural Networks は，上の多様体を多様体と同じ次元のユークリッド空間 (図 3 参照. 以降, ”大域的な座標系　” と呼称) へ写像する機能をもつ．

[入江 90] や [Hinton 06] では，中間層のノード数を多様体の次元に一致させた AutoEncoder を用いて，その中間層の状態を可視化することで部分的な検証が試みている. しかしながら，それらは可視化による検証に留まっており，この仮説が妥当であるかについての明確な答えはまだ得られていない．

本研究の目的は，これら仮説.1 と仮説.2 の検証を定量的に行い，それに基づきデータセットの幾何的構造の観察を行うことである．

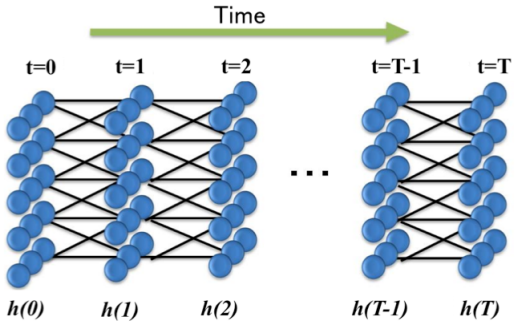

図 4: ニューラルネットワークの時間発展：左から右に層を進んで行くことを時間発展の方向と考える．

4. 分析手法

ニューラルネットワークが獲得した関数を，データセット多様体を大域的な座標系へ写像する関数だとみなすと，その関数を解析することで，元の多様体の性質を知ることが可能となる．なぜならば，多様体から多様体への写像の微分は，以下で定義される多様体の接空間を定義し，そこから多様体の次元や接ベクトル等の情報を得ることができるからである．ここで考えている入力空間は，図 3 にあるような 1 ピクセルを 1 次元とする空間である．

ニューラルネットワークの写像関数の微分 (ヤコビアン行列) の特異値・特異ベクトルのうち，0 より大きな特異値に対応する特異ベクトルが多様体の接線方向 を，0 の特異値に対応するベクトルが多様体の垂直方向をあらわす．従って，0 でない特異値の数から，多様体の次元もわかる．また，特異ベクトルには，右と左があり，右特異ベクトルが入力空間で表現された多様体の水平・垂直ベクトルをあらわし，左特異ベクトルは出力空間で表現された多様体の水平・垂直ベクトルをあらわす．

具体的に DNN の各層の写像関数は以下で定義される．

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

$f(x)$ としては，よくシグモイド関数，

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

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 4 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．従って，第 t 層から $t+1$ 層の写像に対するヤコビアン行列は，

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

となる．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．

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

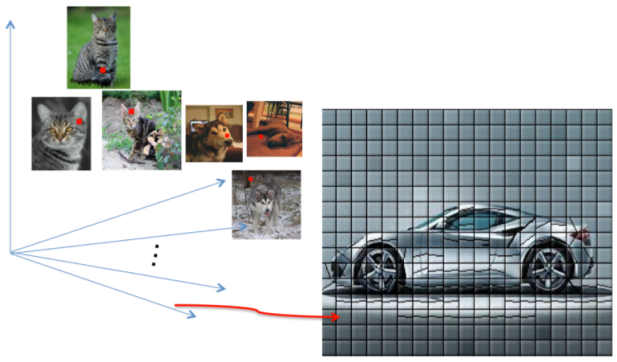

図 5: ダイナミクスを考える空間：画像が 1 点で表される空間でダイナミクスを考える.

5. 実験

本研究では，MNIST データセットによってトレーニングされた Deep Belief Networks[Hinton 06] を分析した．MNIST データセットは，図 2 からもわかるように，多様体の次元が $O(1)$ 程度であると推測されるデータセットである．このようにある程度素性のわかるデータセットを用いて，仮説の妥当性の検証を行った.

さらにその検証に基づき， 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] を用いて，Imagenet データセットの観察を試みた．ImageNet データセットは，単語の定義や同義語のグループとグループ間の関係性が記述された英語の概念辞書 (意味辞書) である WordNet のオント ロジーに従って，各単語 (名詞) に対応する画像を収集したもので，現在約 1,500 万枚の画像が登録されている．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．具体的には，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.

6. 結果：MNIST

MNIST データセットでトレーニングされた DNN を分析した結果，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 の特異値という急峻な特異値分布が見られた（図 6：左上）．しかも，多様体の次元に対応する値が 1 以上である得意値の数は，第 3 層において $O(1)$ 程度であり，これは，図 2 の結果などからわかっている知見と一致する.

さらに，右特異ベクトル（図 6：左下）をみると，回転・平行移動に対応する多様体の接線方向（図 7）に類似した画像が得られていることもわかった．さらに，左特異ベクトルの分析から，出力層が大域的座標系になっていることも確認された.

以上の結果から，仮説 1 と仮説 2 の検証が部分的にできたものと考えた．発表では，これ以外の検証結果についても紹介する（人工データを用いた検証など）.

また，Back Propagation による fine-tuning 後の特異値・特異ベクトルをみると，特異値が全体的に増大し特異ベクトルが変化していることが観察された（図 6：右）．このことは，ラベルデータに基づく Back Propagation は，教師なし学習とは違うダイナミクスを持つことを示唆する.

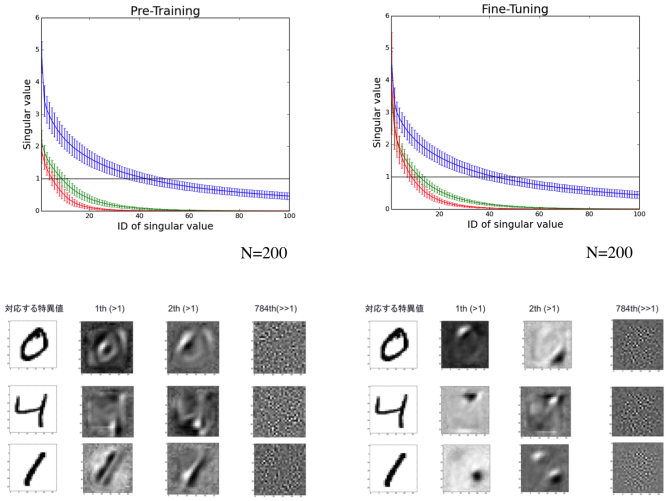

図 6: 上段左:pre-training 後の特異値分布. 層をのぼるごとに, 特異値分布が急峻になっていることがわかる., 上段右:fine-tuning 後の特異値分布.pre-training 後に比べ, 特異値が 1 以上である特異値の数が増加している. 教師データによって, ネットワー クで保存される情報が増大していると予想される. 下段左:pre-training 後の右特異ベクトル. 図 7 のようなデータに対応 したベクトルがみられる. 下段右:fine-tuning 後の右特異ベクトル.pre-training 後と違う様相になっており, 多様体とは違う ものも捉えている可能性がある. エラーバーは標準偏差をあらわす.

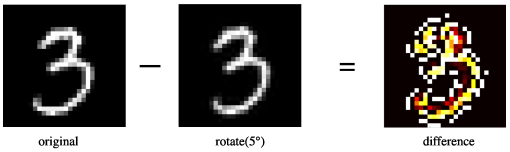

図 7: 回転に対応する接線ベクトル

7. 結果：ImageNet

計算の結果，高次の層において MNIST と同様の，少数の大きな特異値と大多数のほぼ 0 の特異値という急峻な特異値分布が見られた（図 8 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 9 参照）．これは，MNIST の結果と類似している.

さらに，入力画像に対して，多様体の接線方向と垂直方向にそれぞれ接道を加えた際の出力の変動をみることによって，データセットに多様体構造があることがより直接的に確認された.

また，特異値分布の分布構造をクラス毎にみたところ，多様体の次元や，曲率などの局所的な幾何構造の情報しかもたないこの情報だけから，階層的な意味構造を抽出しうることが観察され，意味構造と幾何構造の間に関係があることが示唆された．発表では，これらの分析結果と，この結果を説明するデー

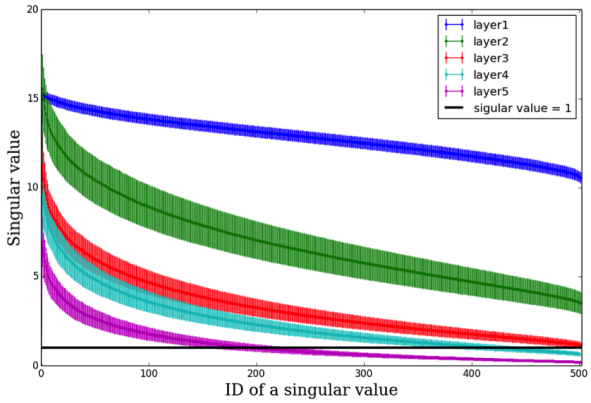

図 8: 特異値分布：それぞれの色に対応する線が，1 層から各層までのヤコビアンの特異値分布（16 の違う入力画像に対する平均）を表している．薄い色は 16 の違う入力画像に対する標準偏差を表している．

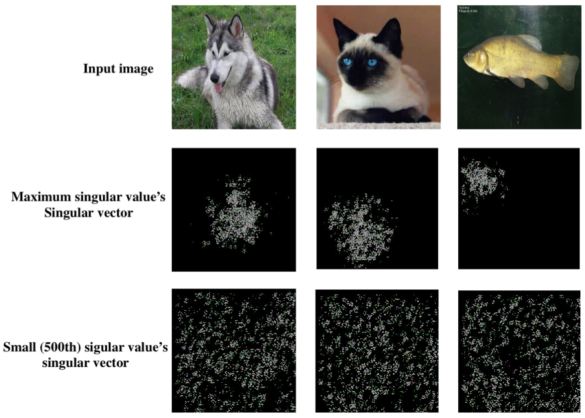

図 9: 特異ベクトル：1 段目が入力画像を，2 段目が特異値が最大となる特異ベクトルを，3 段目が計算した中で特異値が最小となる特異ベクトルを表す．

タセットの幾何構造に対する仮説とその検証結果を紹介する．

8. まとめと議論

本研究によって，高次元データセットが低次元な多様体構造をもち，さらにその幾何的な構造と画像の意味構造が対応することが示唆された．また，パフォーマンスの高い DNN が，そのような構造を大域的な座標系に写像するような関数を獲得していることも確認された．

この結果は，DNN が高いパフォーマンスを獲得する要因が，データセットの階層的幾何構造と，DNN の階層的ネットワーク構造の親和性にある可能性を示唆すると考えられる．さらに，意味的構造が幾何構造と関係しているという結果をさらに深く探究することは，機械学習分野のみならず，認知科学や脳科学分野に大きな知見を提供できるものと考えられる．

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.

[Simonyan 14] Simonyan, Karen, and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition." arXiv preprint arXiv:1409.1556 (2014).

[Rifai 11] Salah Rifai, Yann N Dauphin, Pascal Vincent, Yoshua Bengio, and Xavier Muller. The manifold tangent classifier. In Advances in Neural Information Processing Systems, pp. 2294-2302, 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.

[入江 90] 入江文平, 川人光男. 多層パーセプトロンによる内部表現の獲得. 電子情報通 信学会論文誌 D, Vol. 73, No. 8, pp. 1173-1178, 1990.