DEEP NEURAL NETWORKS の力学的解析

本武 陽一
東京大学大学院総合文化研究科 池上研究室 博士課程

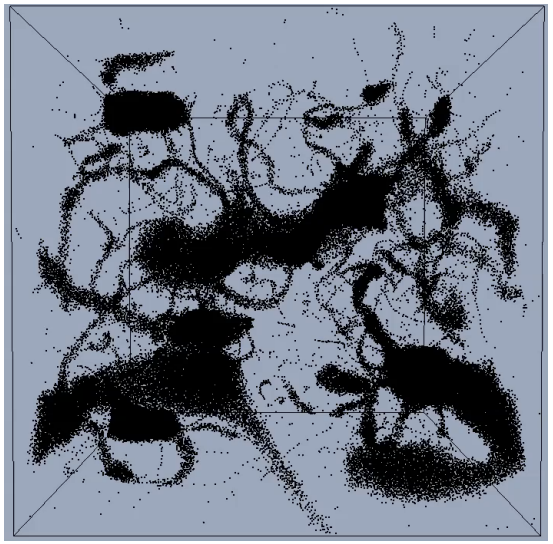
mototake@sacral.c.u-tokyo.ac.jp

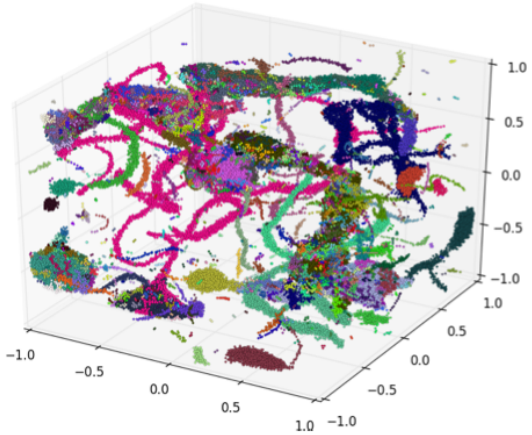
第11回全脳アーキテクチャ勉強会 2015.08.26

自己紹介

- 本武陽一 : 東京大学・総合文化研究科・
  広域科学専攻・広域システム科学系・池上高志研究室・博士課程

機械学習と物理学を活用して,高次元情報や大自由度力学系の捉え方を模索しています.
現在,Deep Learning研究以外に,大自由度(個体数が多い)群れモデルについての研究等を行っています.

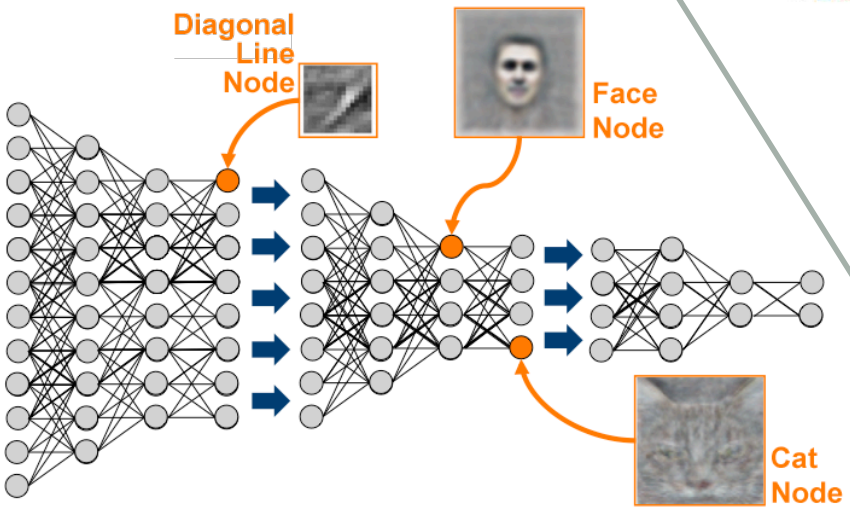




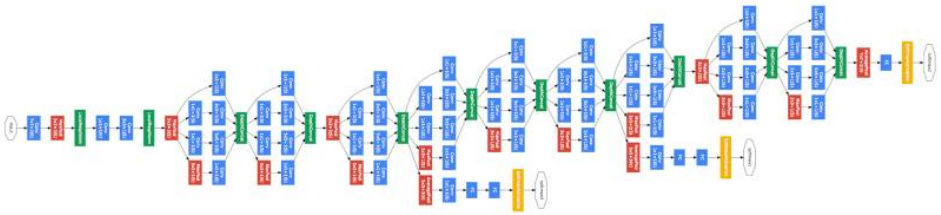
[mototake, ikegami, 2015]

FB : https://www.facebook.com/yhoich.motake
Twitter : motake2010
Github : https://github.com/keaton2006

イントロダクション



[Lecun et al., 2012]



This DNN can recognize the difference ↓

(a) Siberian husky (b) Eskimo dog

Figure 1: Two distinct classes from the 1000 classes of the ILSVRC 2014 classification challenge.

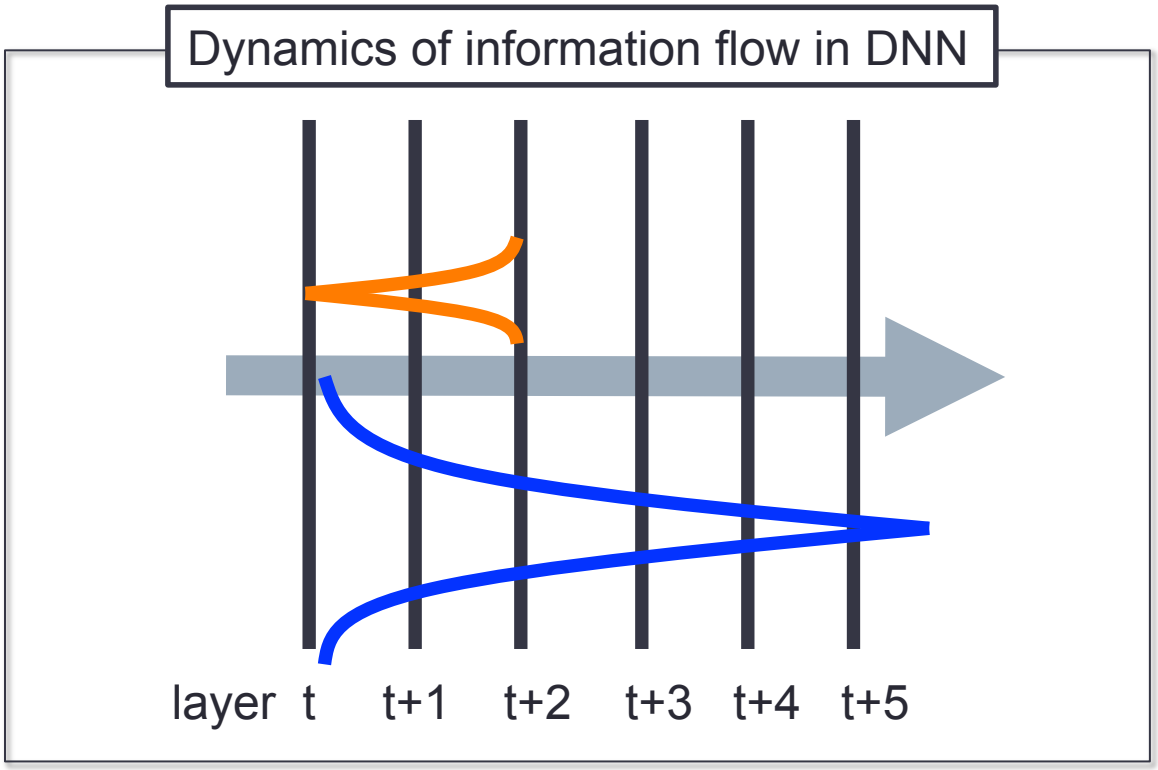
[Szegedy et al., 2014]

⇒多層ニューラルネットワークにおいて,
- 階層的な表象が,特定の作り込みなしに生成される (*左*),
- 非常に高い識別精度が達成される (*右*).

研究の目的

『なぜdeep neural networks (DNN) は,このような能力を獲得できるのか？』

⇒DNNの “ダイナミクス” を分析する

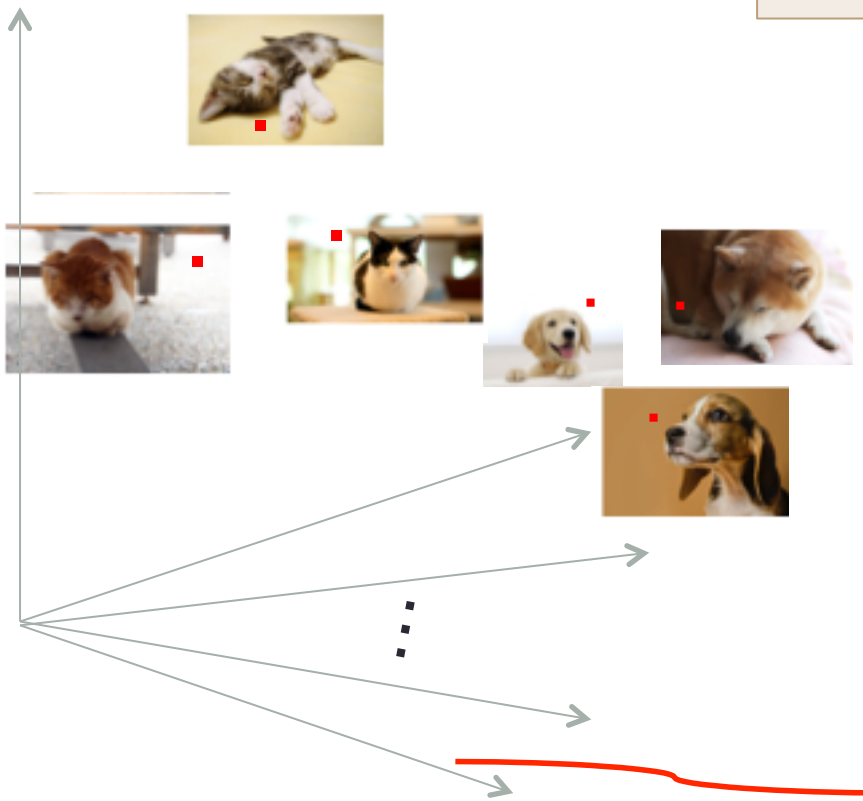


DNNのダイナミクス

「空間」の定義

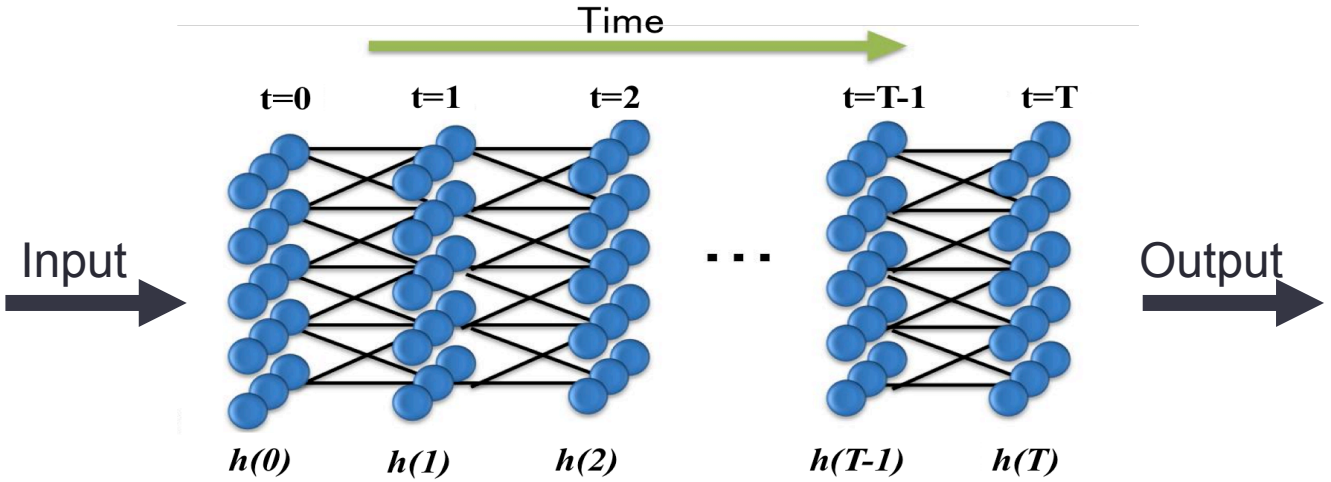
1ピクセル = 1 次元

⇒1 画像 = 1 粒子



DNNのダイナミクス

“時間発展”の定義

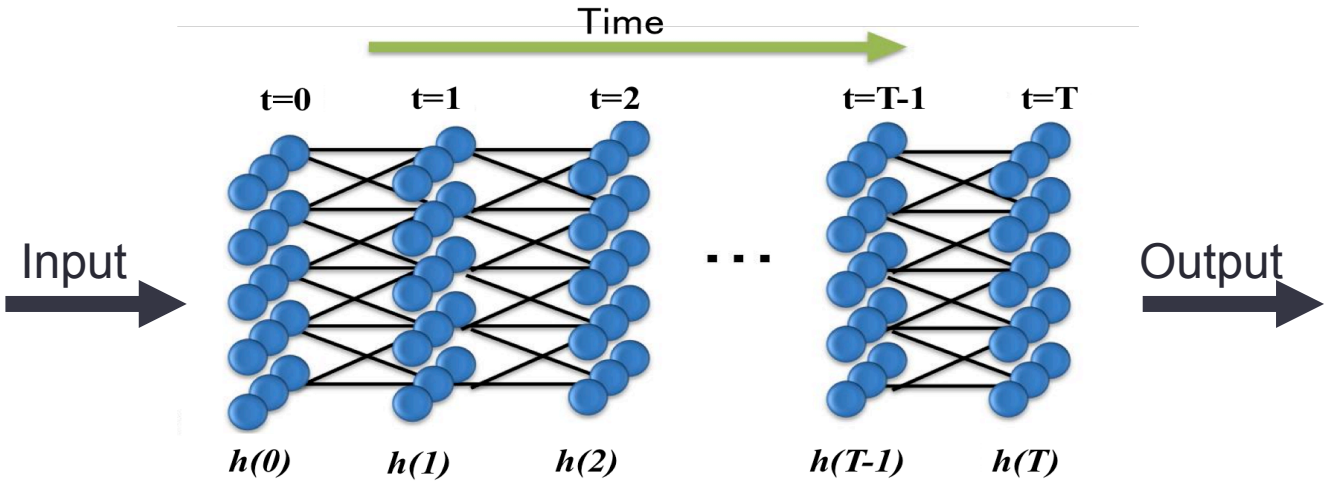


$$h_j(t+1) = f\left(\sum_i (h_i(t) \times W_{ij}(t)) + B_j(t)\right)$$

$$f(x) = \frac{1}{1+e^{-gx}} \qquad (g : const)$$

h(t):t-th layer’s hidden node state
W(t):weight matrix between t-th and (t+1)th layer.
B(t):bias of (t+1)th layer hidden node

DNNのダイナミクス



Jacobian matrix：representing the change of volume by variable transformation

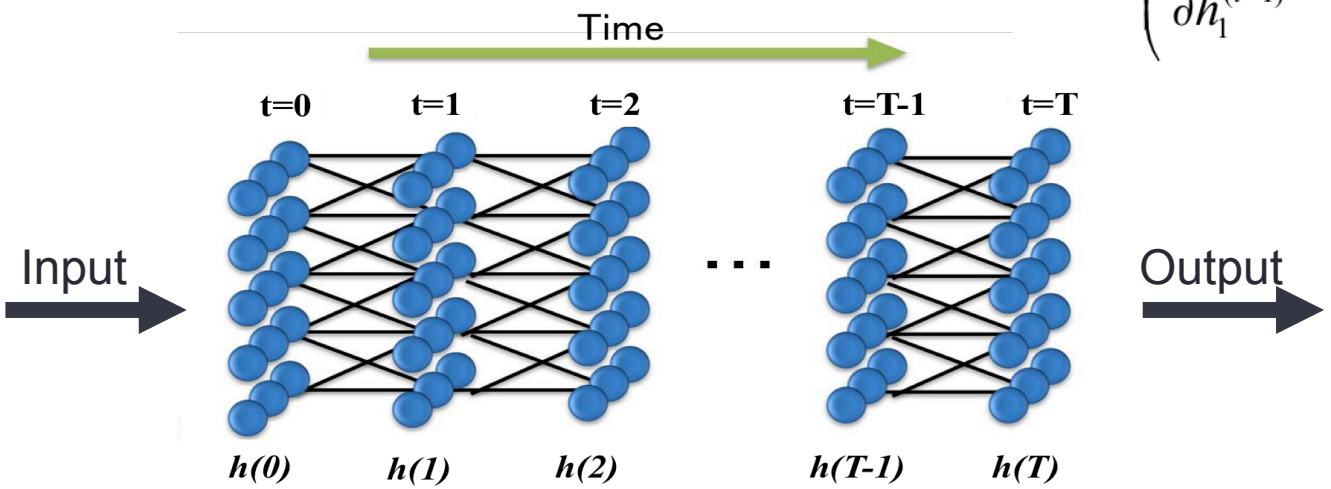
$$J^{(t)} = \begin{pmatrix} \frac{\partial h_1^{(t)}}{\partial h_1^{(t-1)}} & \cdots & \frac{\partial h_1^{(t)}}{\partial h_N^{(t-1)}} \\ \vdots & \ddots & \vdots \\ \frac{\partial h_N^{(t)}}{\partial h_1^{(t-1)}} & \cdots & \frac{\partial h_N^{(t)}}{\partial h_N^{(t-1)}} \end{pmatrix}$$

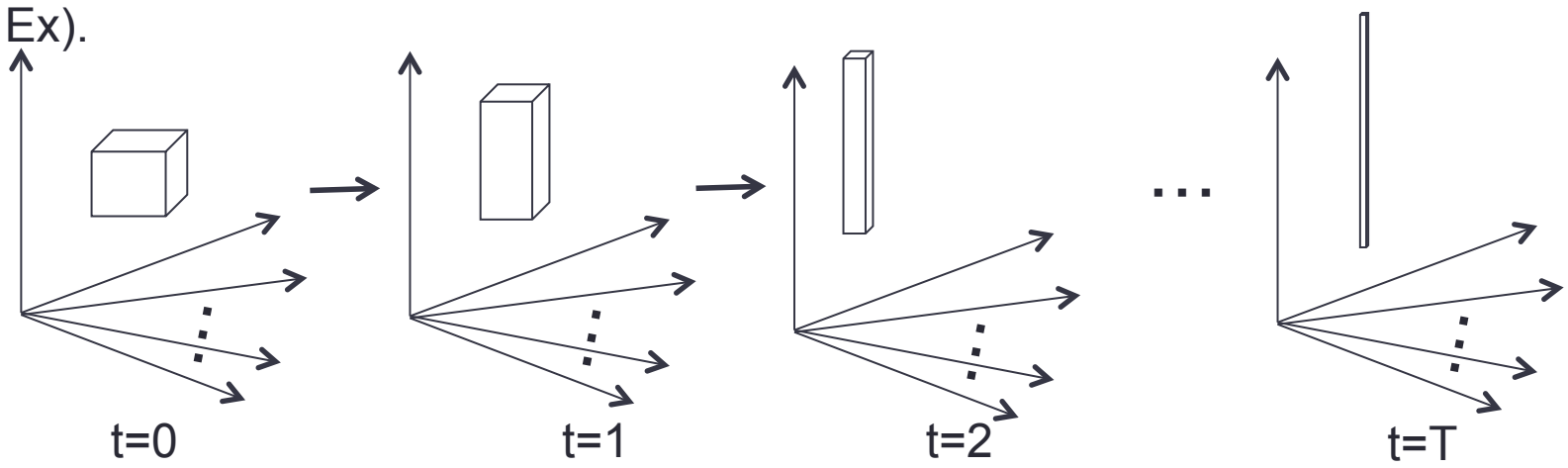
Whole jacobian

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

DNNのダイナミクス

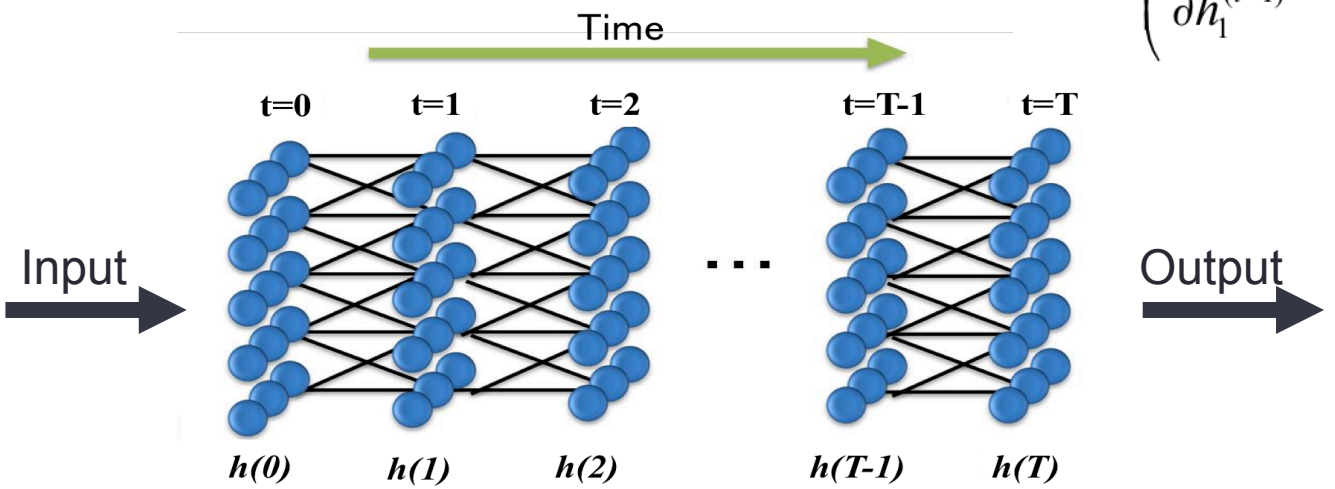
$$J^{(t)} = \begin{pmatrix} \frac{\partial h_1^{(t)}}{\partial h_1^{(t-1)}} & \cdots & \frac{\partial h_1^{(t)}}{\partial h_N^{(t-1)}} \\ \vdots & \ddots & \vdots \\ \frac{\partial h_N^{(t)}}{\partial h_1^{(t-1)}} & \cdots & \frac{\partial h_N^{(t)}}{\partial h_N^{(t-1)}} \end{pmatrix}$$

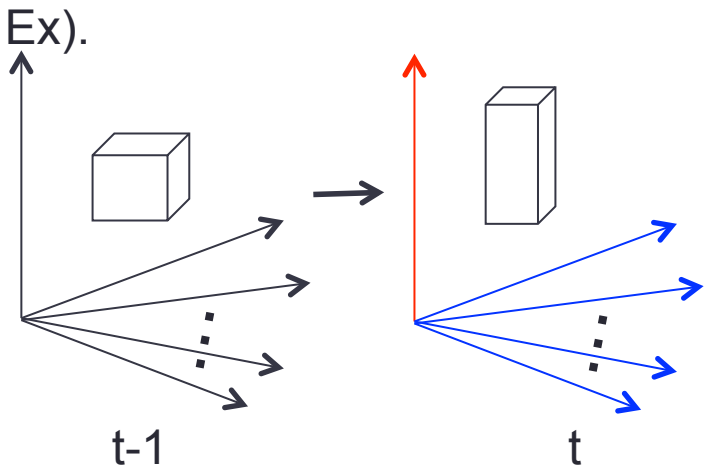




DNNのダイナミクス

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





（ヤコビアンの特異値）> 1:
情報が拡大される

（ヤコビアンの特異値）< 1:
情報が圧縮される

情報 = 特異ベクトルの方向

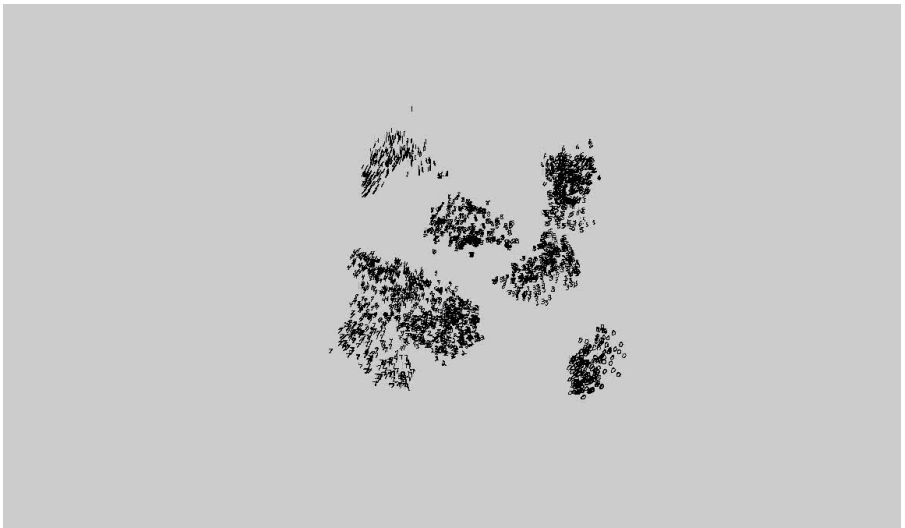
DNNのダイナミクス

多様体仮説：カテゴリ情報は、最初から低次元多様体上に表現される

the *manifold hypothesis* (Cayton, 2005; Narayanan and Mitter, 2010), according to which real-world data presented in high dimensional spaces are expected to concentrate in the vicinity of a manifold $\mathcal{M}$ of much lower dimensionality $d_{\mathcal{M}}$, embedded in high dimensional input space $\mathbb{R}^{d_x}$. [Benjio et al., 2012]

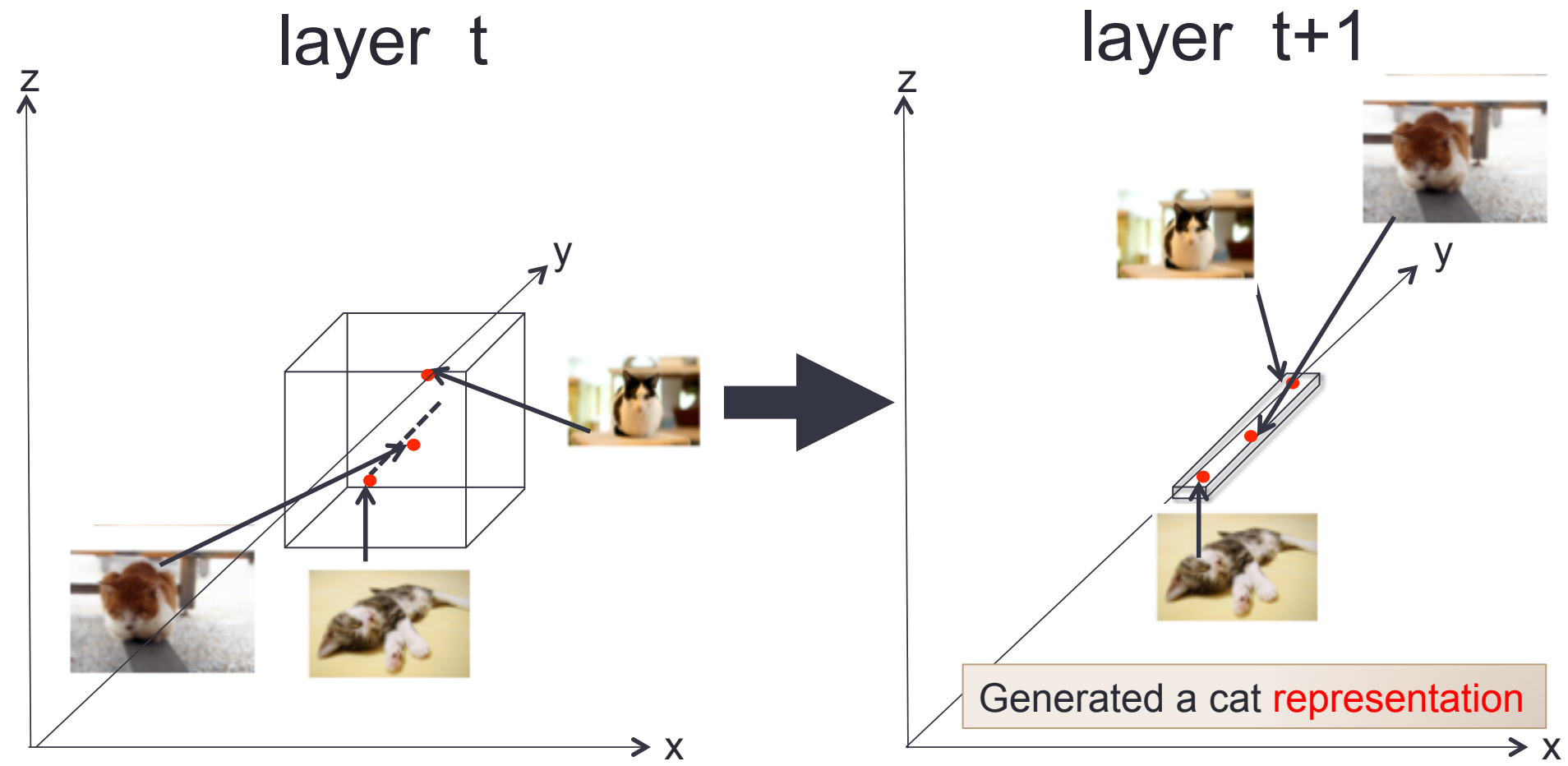
The **manifold hypothesis for classification**, according to which points of different classes are likely to concentrate along different sub-manifolds, separated by low density regions of the input space. [Refai, & Bengio et al., 2011]

↓手書き数字データセット（mnist）を多様体学習（t-SNE）で3次元に次元圧縮した結果



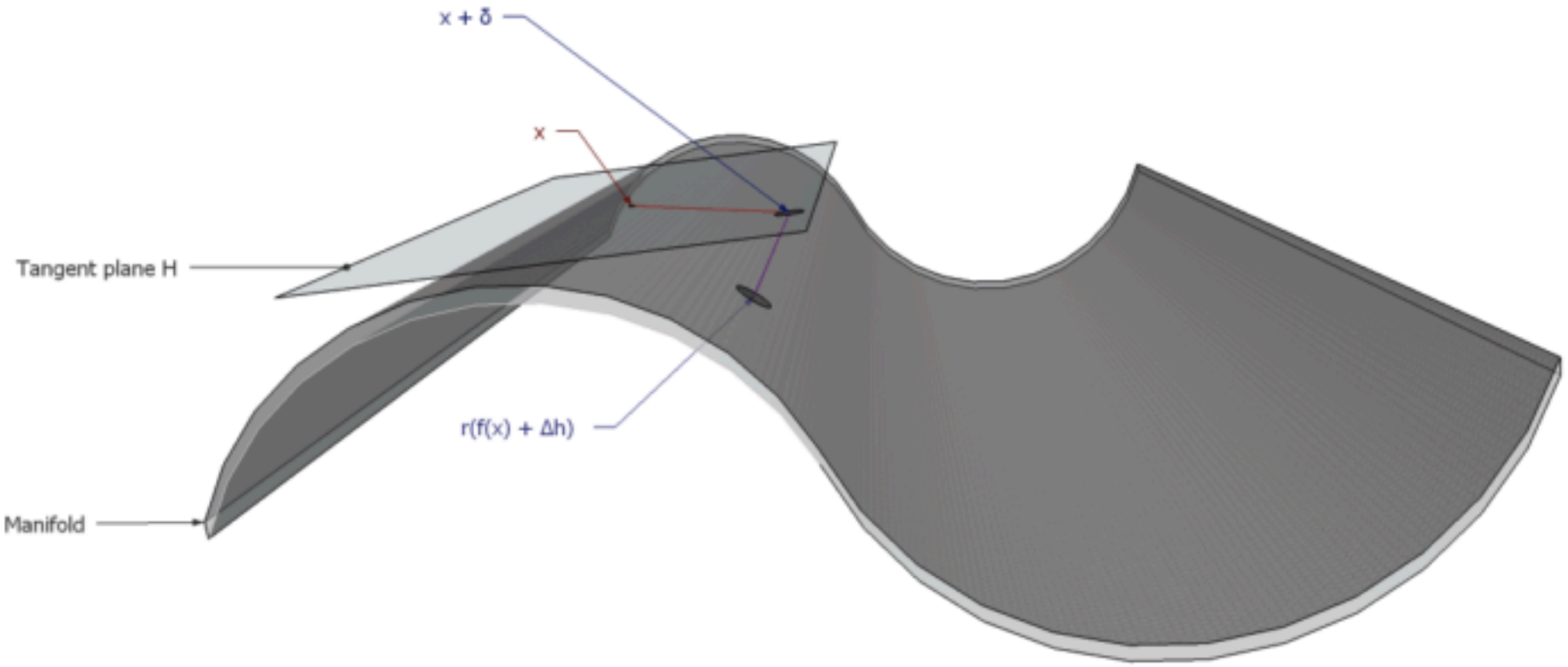
Laurens van der Maaten
http://lvdmaaten.github.io/tsne/
（↑左の動画があります）

DNNのダイナミクス



⇒表象 = あるカテゴリに属するイメージ群が埋め込まれている低次元の多様体.

DNNのダイナミクス



[Bengio et.al, 2012]より

⇒表象 = あるカテゴリに属するイメージ群が埋め込まれている低次元の多様体.

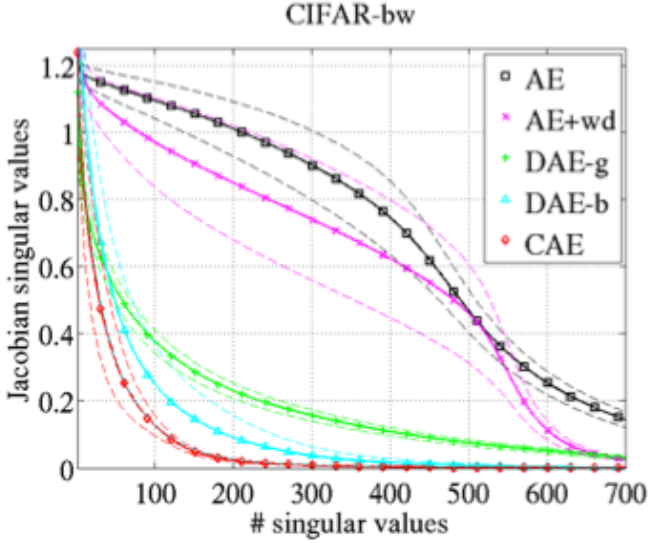
先行研究

Deep Learning and Representation

[Bengio et.al, 2012] [Refai, & Bengio et al., 2011] deep learningが多様体を捉えていることを示唆.

ヤコビアンの行列ノルムを制約項として用いたAutoEncoder(CAE)にて,
そのヤコビアンの特異値が急峻な分布となり,特異ベクトルが入力の特徴をあらわすことが実験的に観察された. [Refai, & Bengio et al., 2011]

⇒CAEが低次元の多様体のtangent spaceを捉えていることが示唆される. [Bengio et.al, 2012]



特異値分布(CAE:赤)

input

特異値の大きな(tangent方向の)特異ベクトル

仮説とQuestion

仮説:

高いパフォーマンスや表象の獲得には，データが埋め込まれた低次元の多様体を捉えるようなダイナミクスが必要.

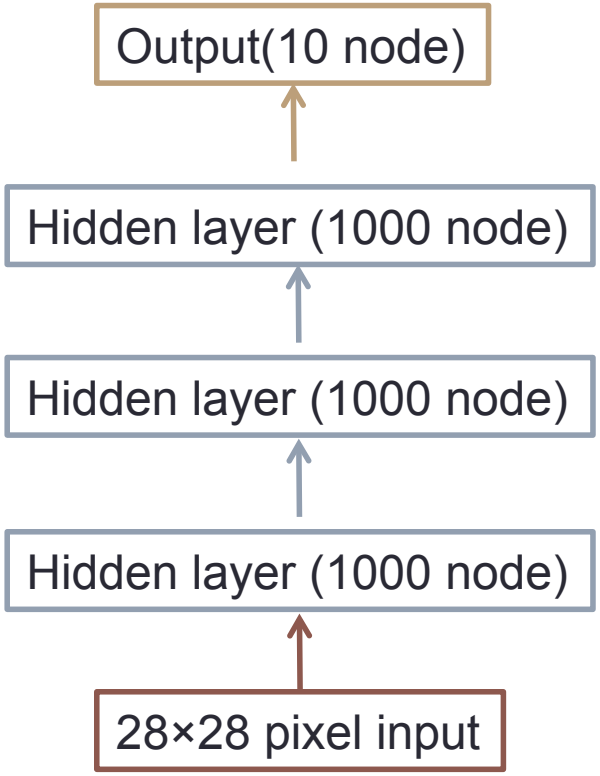
Question:
この仮説は,実際に活用されている,
DNNネットワークにおいても, 成り立っているか?

1.DBN & MNIST dataset

手法:学習モデル

Deep Belief Network [Hinton et al., 2006]

⇒RBMでpre-training後, fine-tuning（今回はBack Propagationを利用）

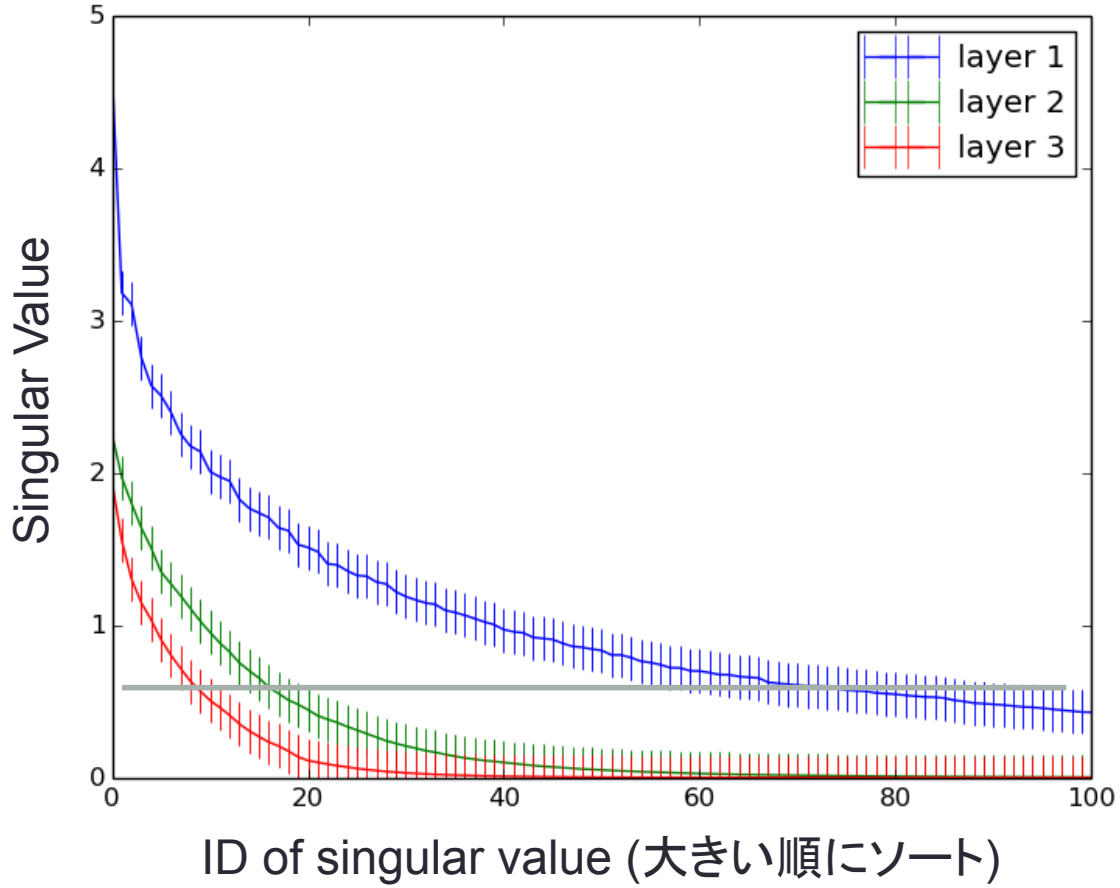


手法:トレーニングデータセット

Mnistデータセット (Yann LeCun et al. : http://yann.lecun.com/exdb/mnist/)

結果1-1：特異値分布(pre-training後)

（複数の入力画像に対する平均特異値の分布）

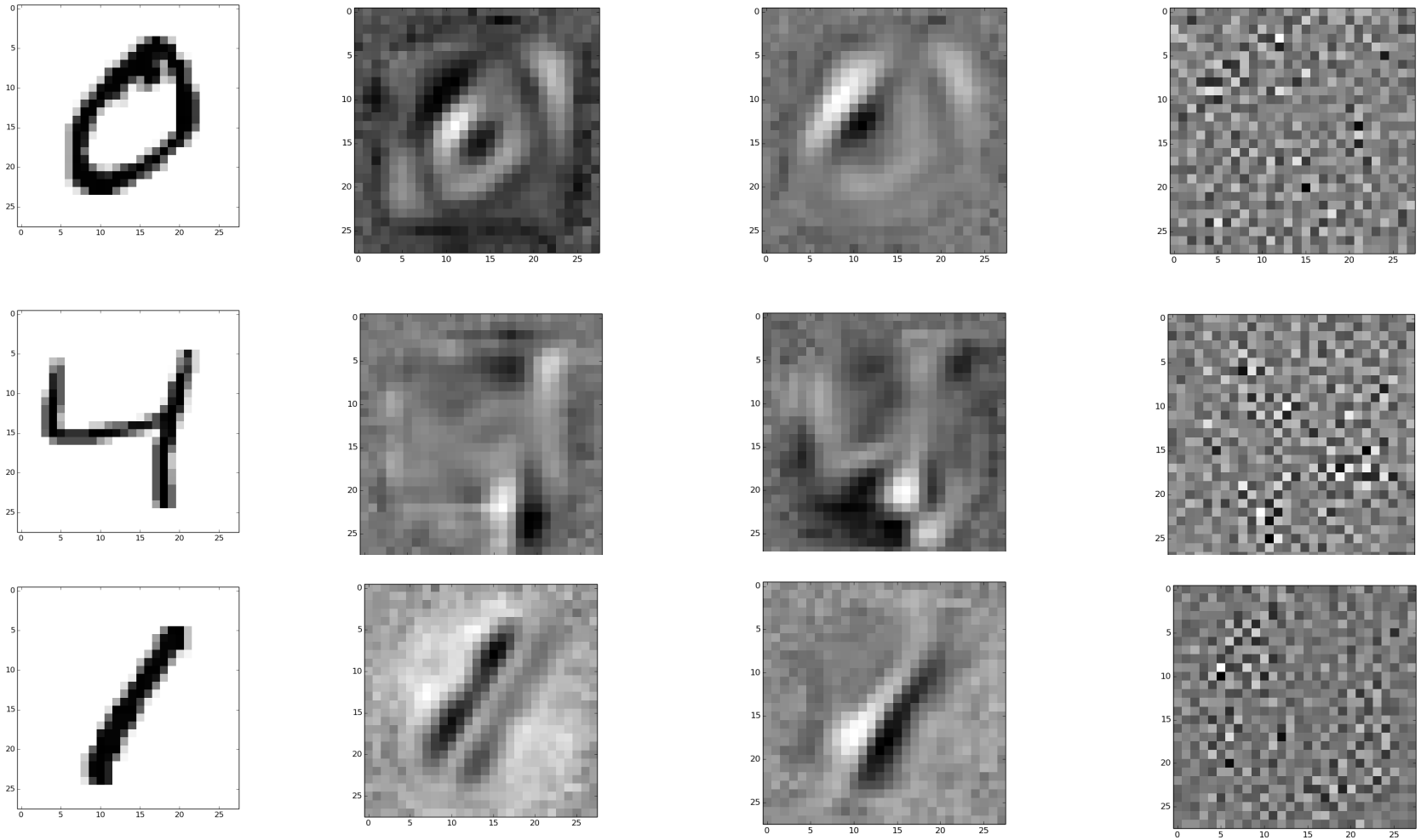


エラーバーは標準偏差
N=20

⇒3層目において,急勾配の特異値分布が実現されており,これは仮説と整合性がある.
また,特異値が1以上となる成分は5次元以下であり,これはt-SNEにて3次元程度で
カテゴリ毎に分離できていたことと一致する.

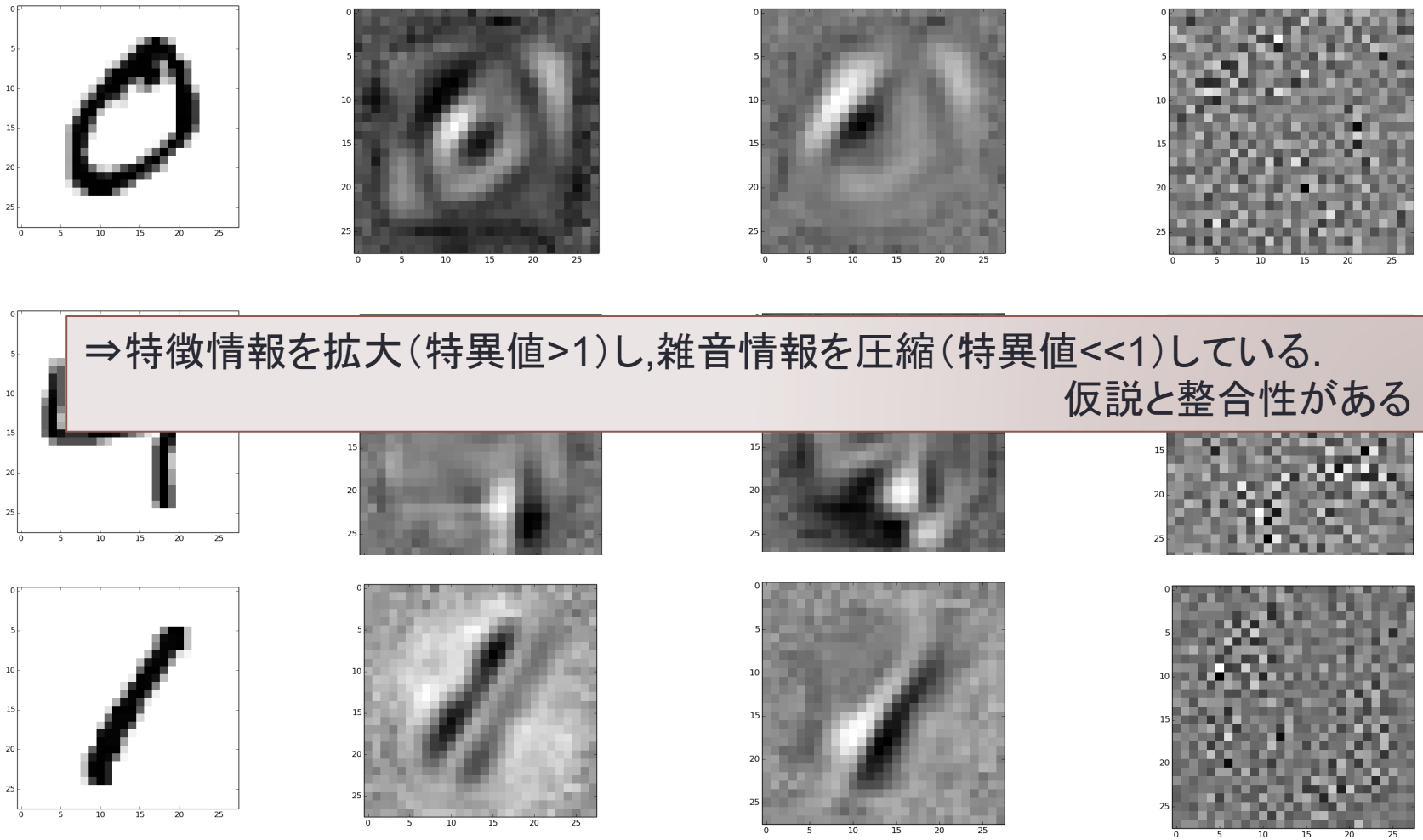
結果1-2:（左）特異ベクトル(pre-training後)

対応する特異値　　1th (>1)　　2th (>1)　　784th(>>1)



結果1-2:（左）特異ベクトル(pre-training後)

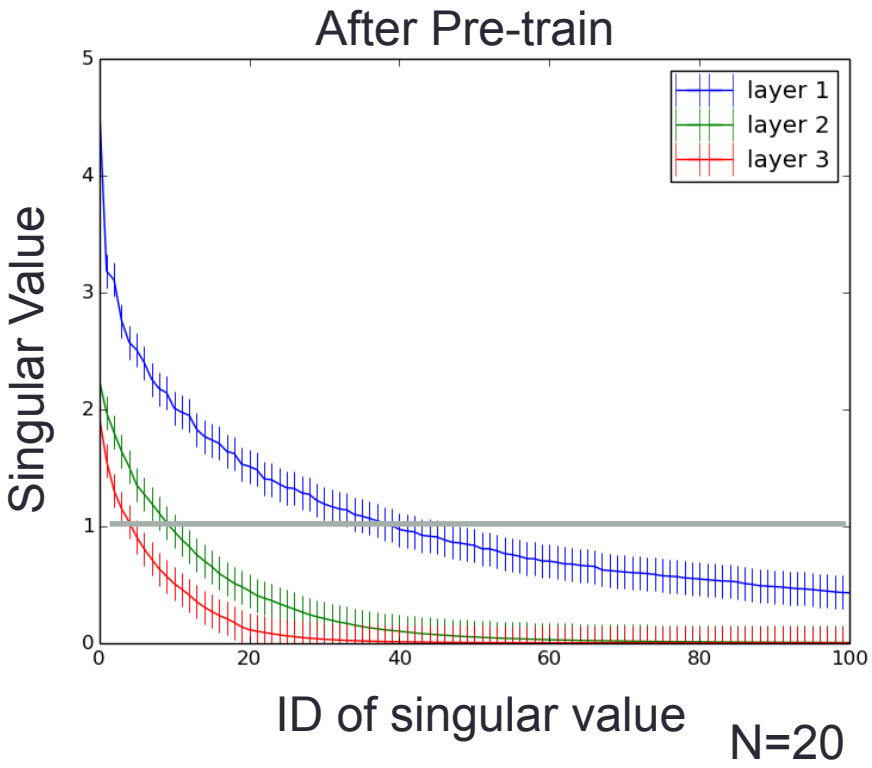
対応する特異値　　1th (>1)　　2th (>1)　　784th(>>1)

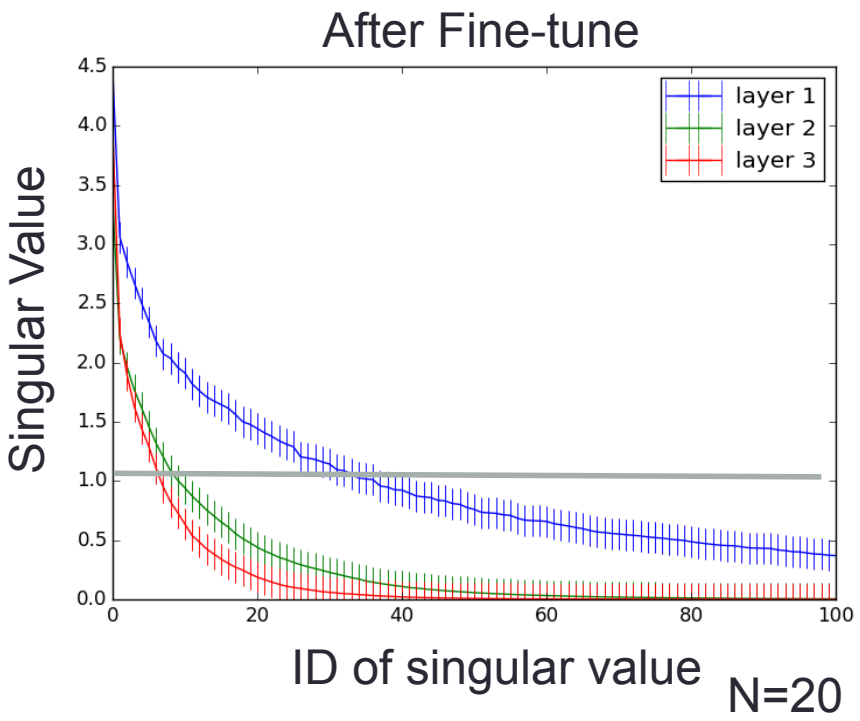


⇒特徴情報を拡大（特異値>1）し,雑音情報を圧縮（特異値<<1）している.
仮説と整合性がある

結果1-3：特異値(fine-tune後)

（複数の入力画像に対する平均特異値の分布）

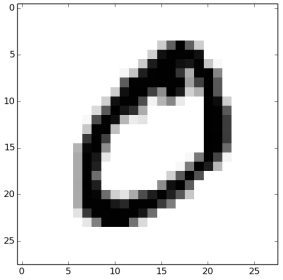
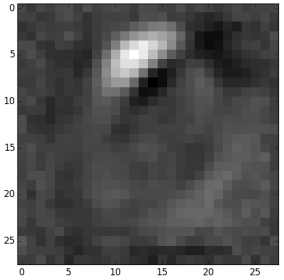
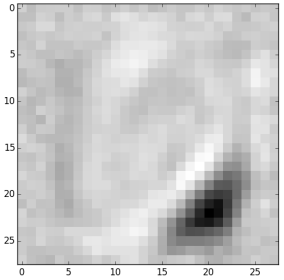
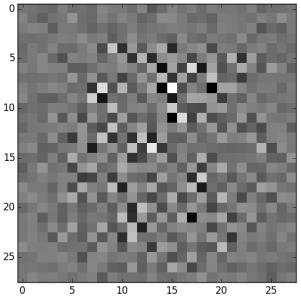
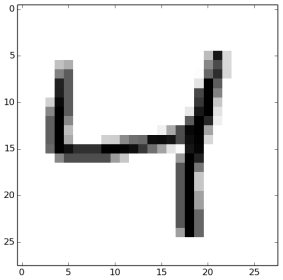
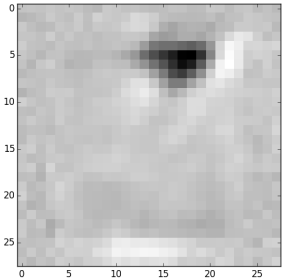
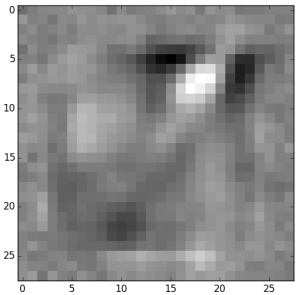
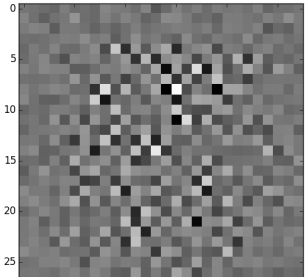
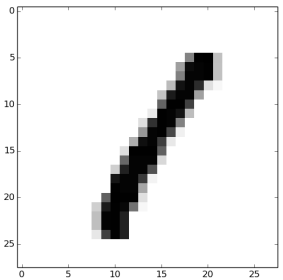
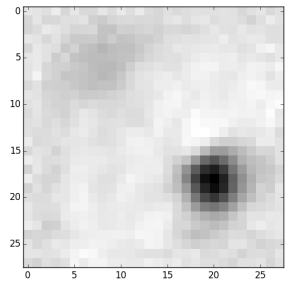
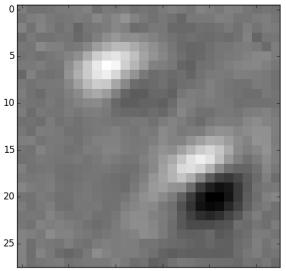
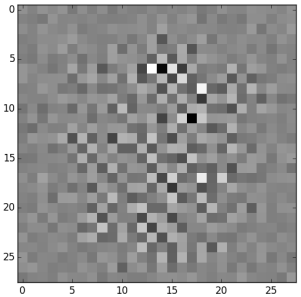




エラーバーは標準偏差

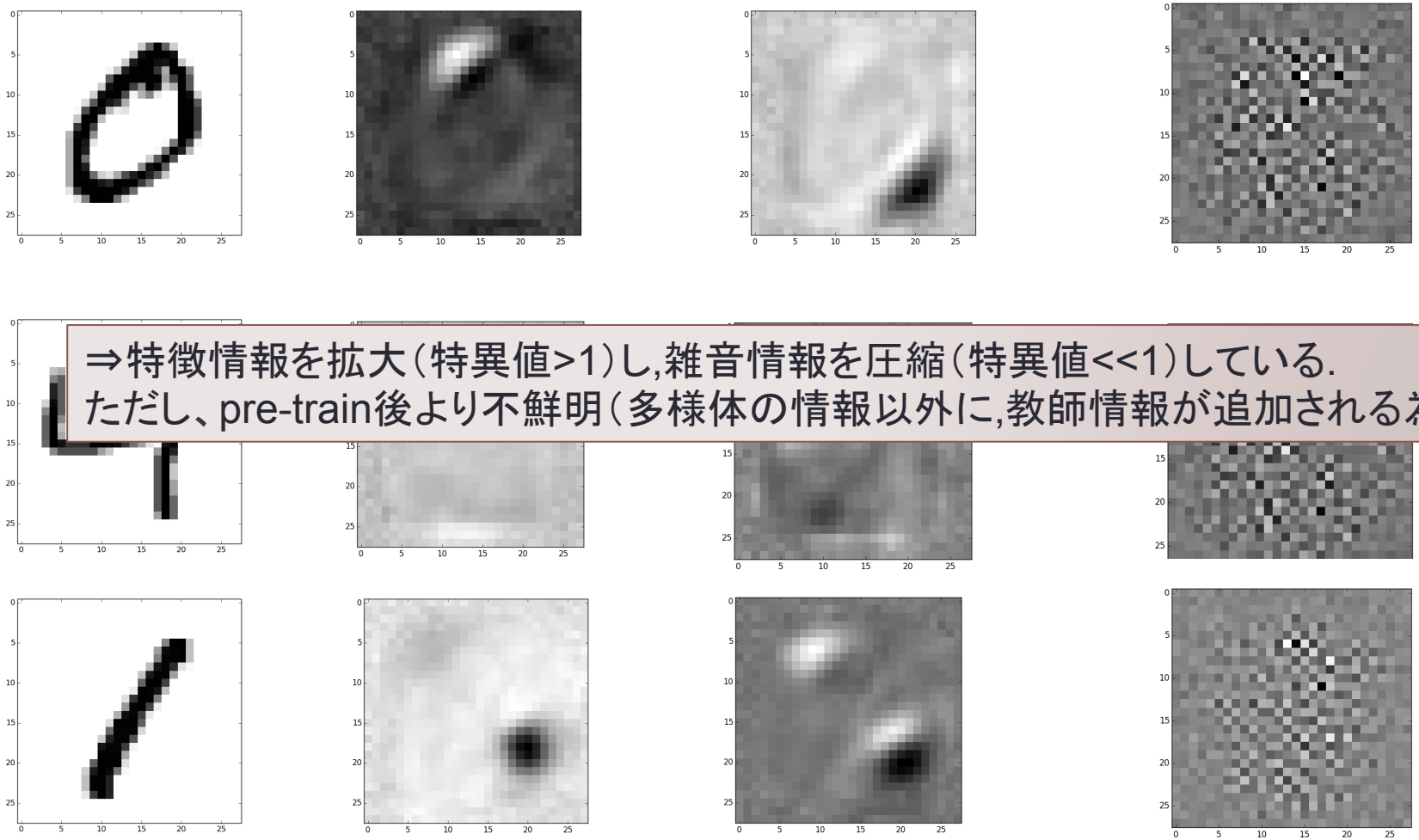
⇒fine-tuneによって拡大される情報が増えている
（多様体の情報以外に,教師情報が追加される為？）

結果1-4:（左）特異ベクトル(fine-tune後)

対応する特異値	1th (>1)	2th (>1)	784th(>>1)
			
			
			

結果1-4:（左）特異ベクトル(fine-tune後)

対応する特異値　　1th (>1)　　2th (>1)　　784th(>>1)



⇒特徴情報を拡大（特異値>1）し,雑音情報を圧縮（特異値<<1）している.
ただし、pre-train後より不鮮明（多様体の情報以外に,教師情報が追加される為？）

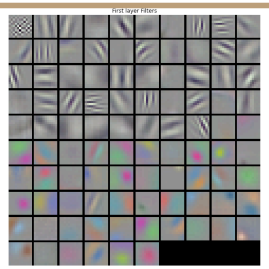
Alexnet & Imagenet dataset

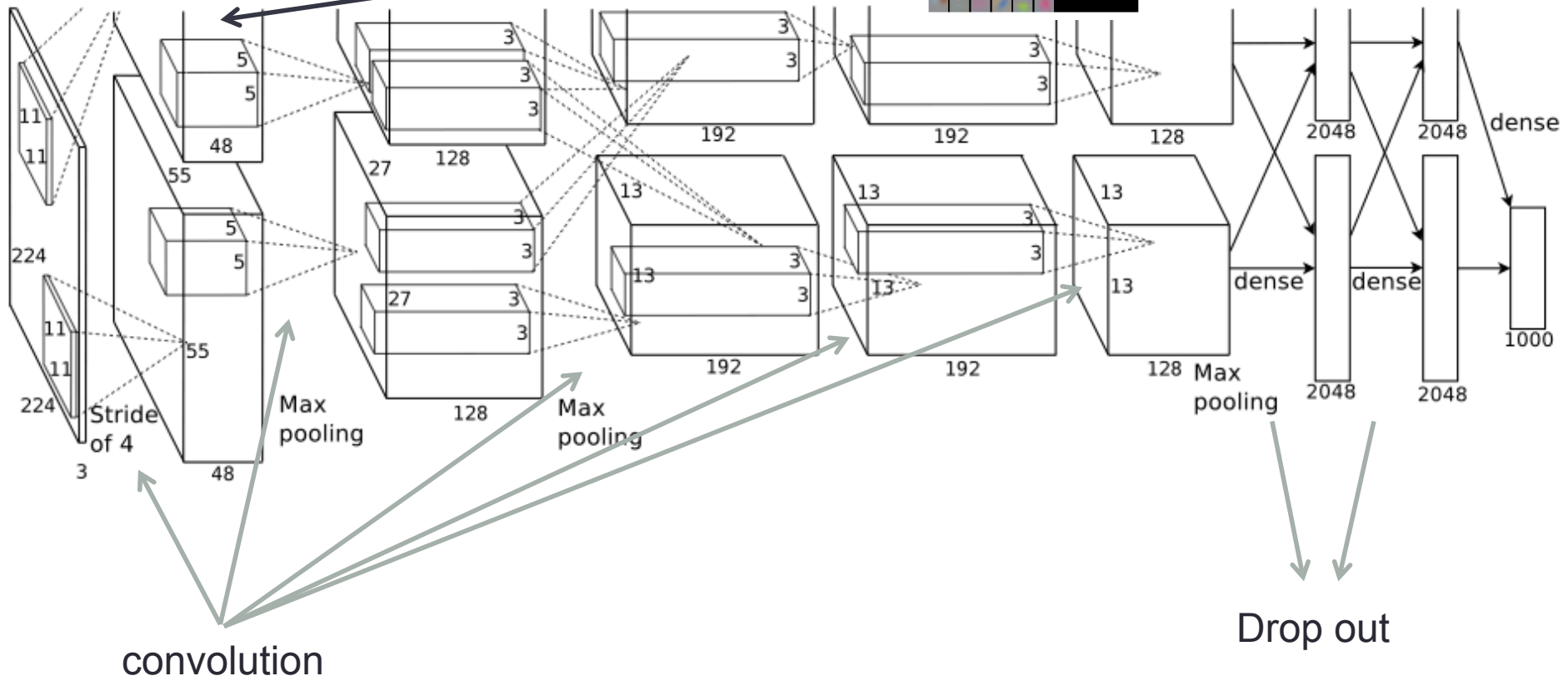
手法:学習モデル

[Krizhevsky et al., 2012]

→Winner of ILSVRC12 (task1）

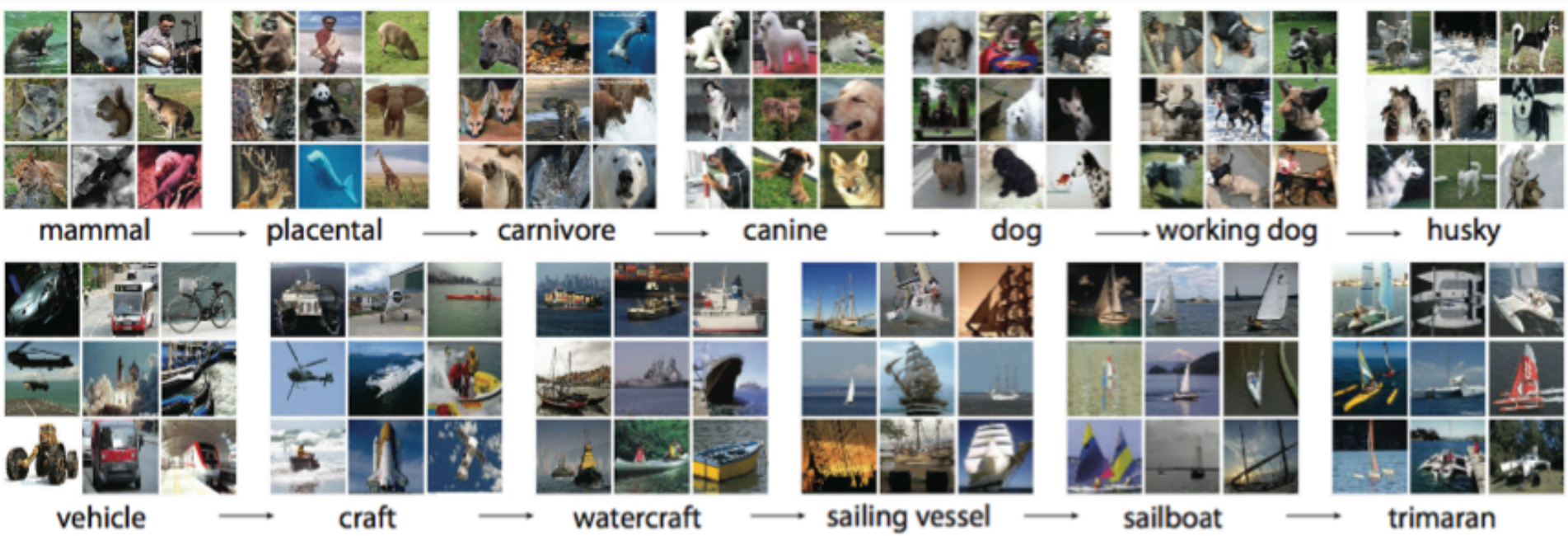
1層目がFeatureに対応するpathを分析





手法:トレーニングデータセット

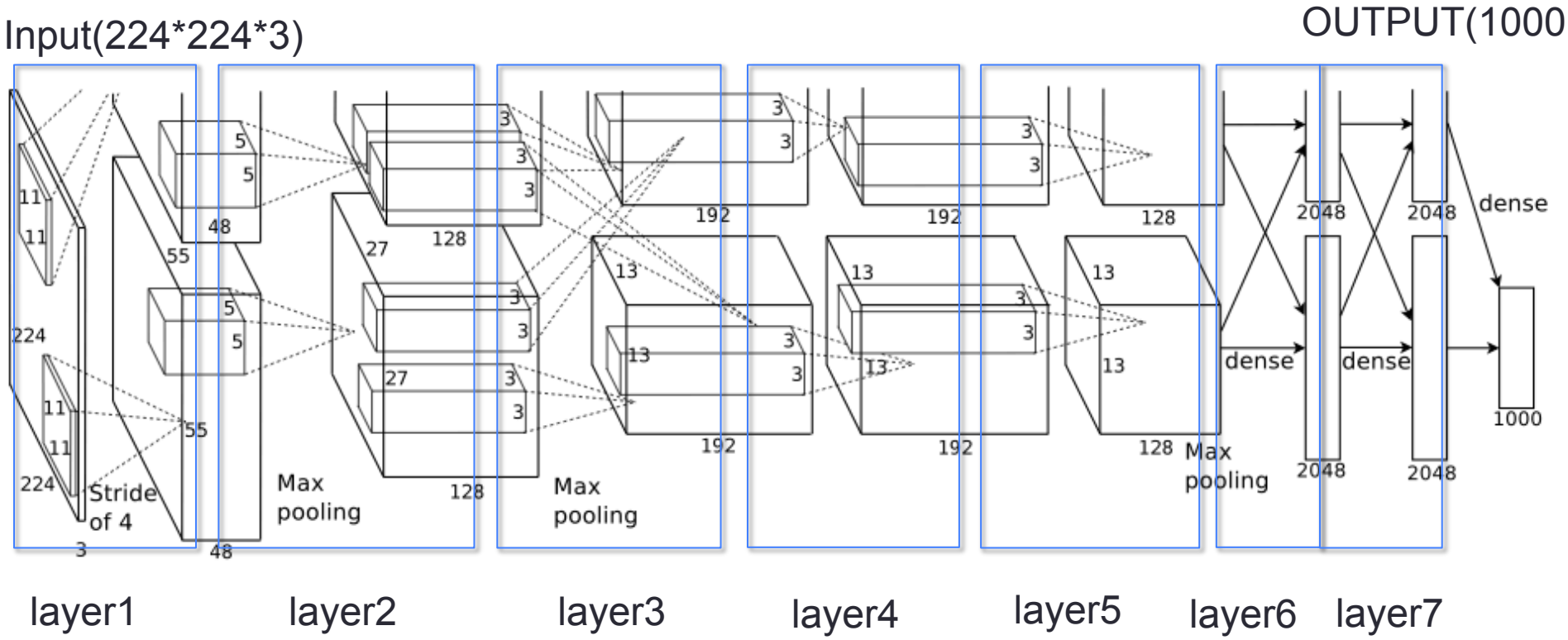
ImageNet Dataset (http://www.image-net.org)



15 million labeled high-resolution images in over 22,000 categories

手法:Layerの定義

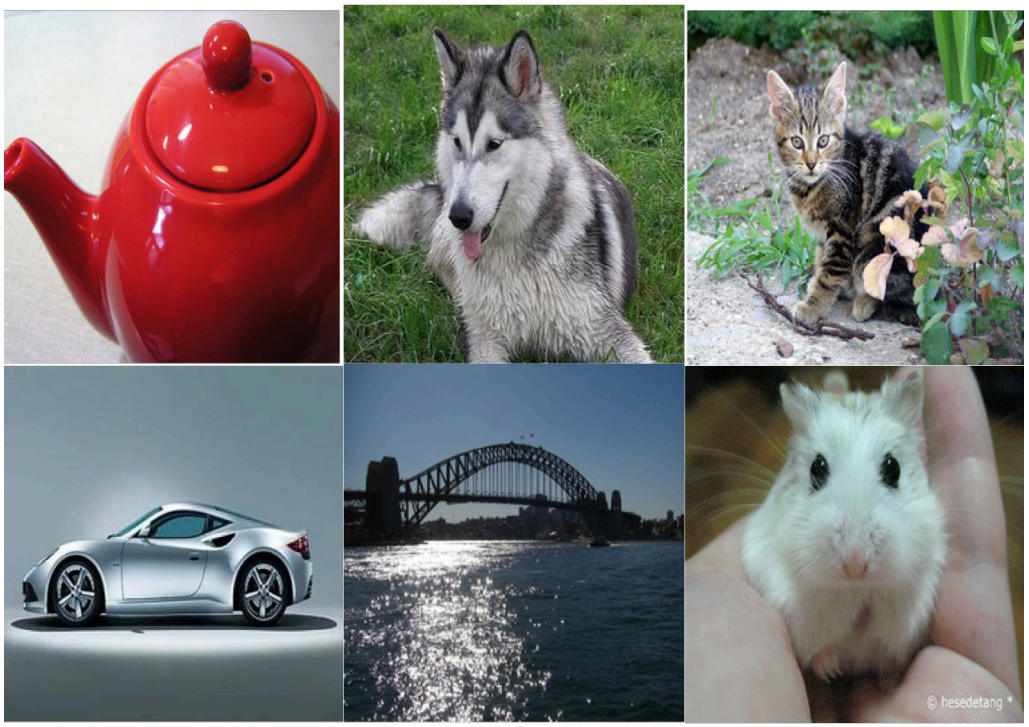
以下,説明の為，layerを以下のように定義します.



DNN library (DeCAF[Donahue et al., 2013])を使用

手法: ヤコビアンを計算する為に用いたインプットデータ

ヤコビアンの計算には,Imagenet データセットにある16の画像をインプットとして使用

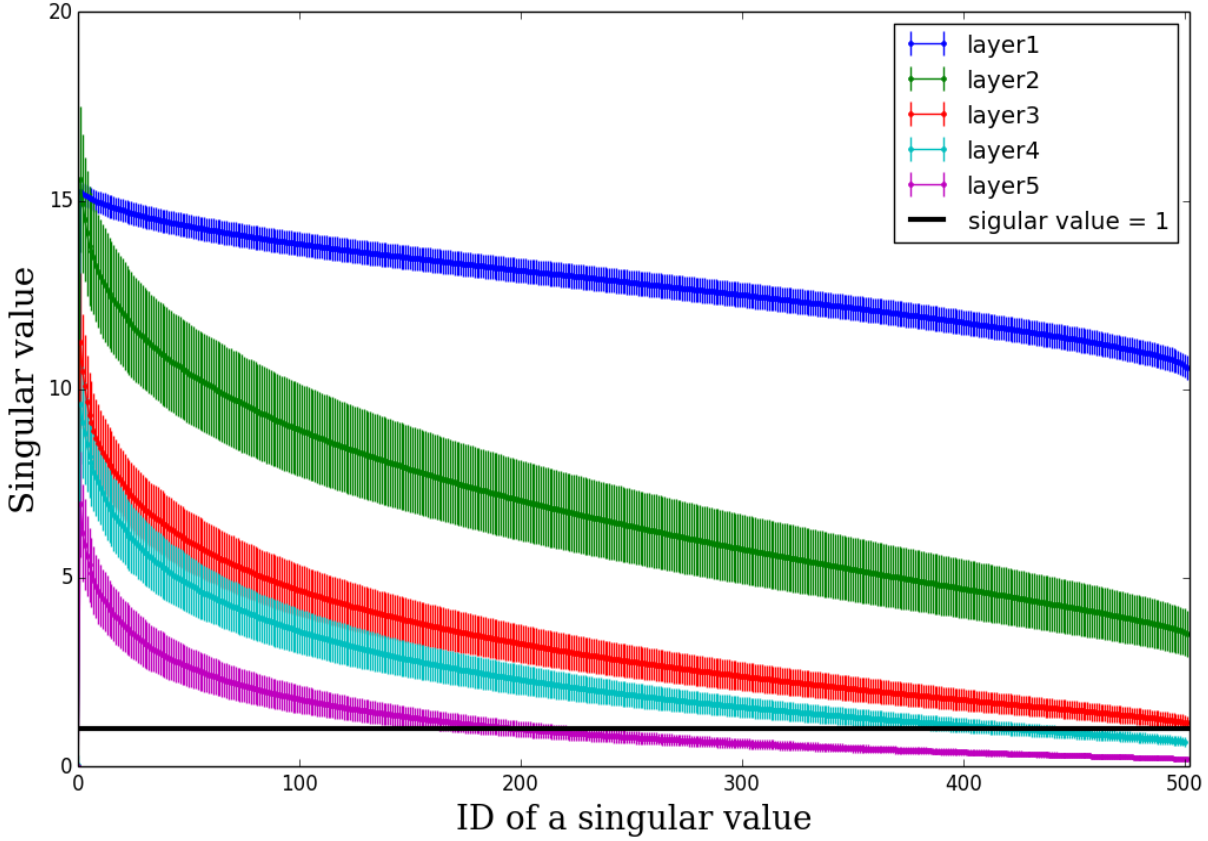


[ImageNet Dataset]

結果2-1 : 特異値分布

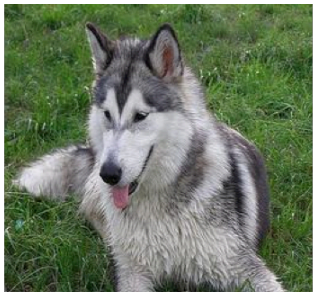
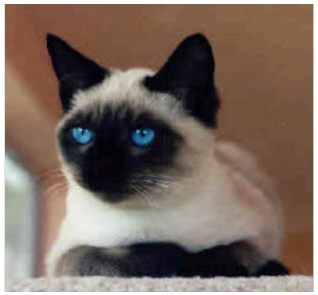
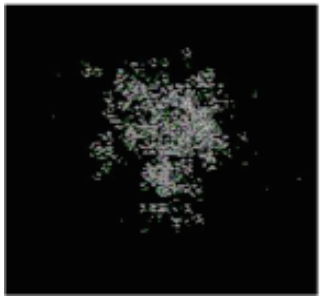
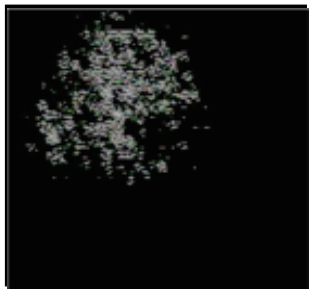
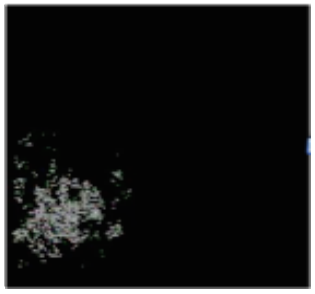
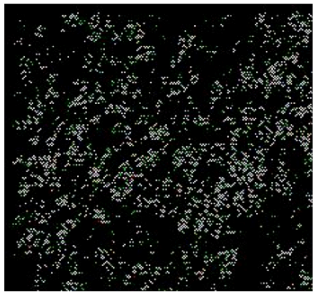
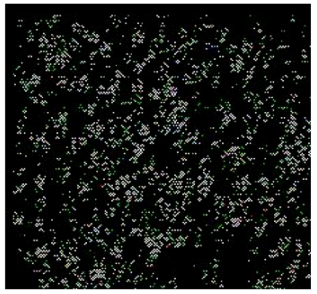
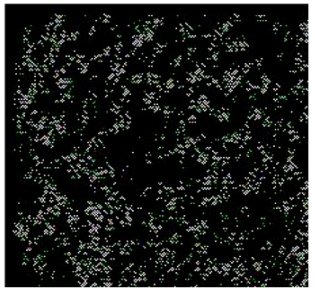
エラーバーは標準偏差

N=16



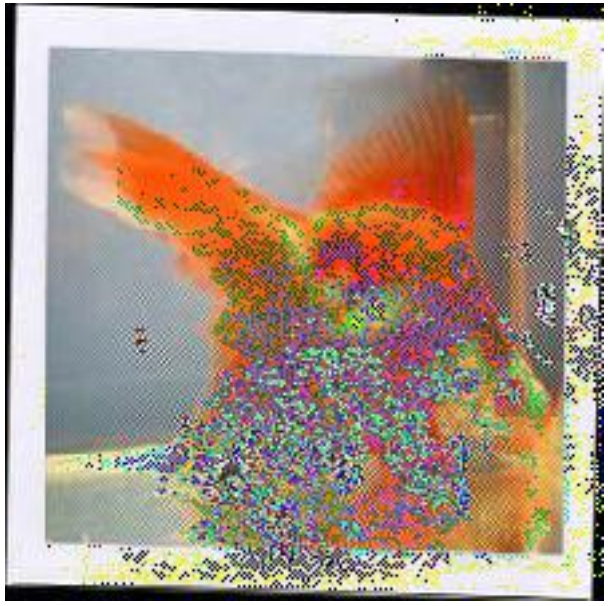
⇒5層目において,急勾配の特異値分布が実現されており,
これは仮説と整合性がある.

結果2-2：（左）特異ベクトル

Input image			
Maximum singular value's Singular vector			
Small (500th) sigular value's singular vector			

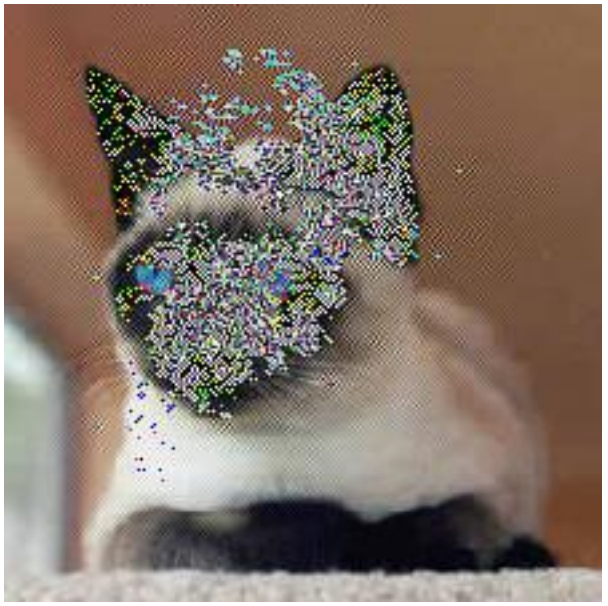
⇒局所的なパターンを拡大し,雑音情報を圧縮している.
そして,局所的パターンは，featureに対応しているように見える.

結果2-2：（左）特異ベクトル
拡大されるベクトル（上位2つ）の入力画像との比較



判別結果（Top5）:['goldfish', 'axolotl', 'American lobster', 'plastic bag', 'screen’]

結果2-2：（左）特異ベクトル
拡大されるベクトル（上位2つ）の入力画像との比較



識別結果（Top5）:['Siamese cat', 'Egyptian cat', 'lynx', 'Norwegian elkhound', 'schipperke']

結果2-2：（左）特異ベクトル
拡大されるベクトル（上位2つ）の入力画像との比較

判別結果（Top5）:['hamster', 'broccoli', 'Angora', 'guinea pig', 'polecat’]

結果2-2：（左）特異ベクトル
拡大されるベクトル（上位2つ）の入力画像との比較

識別結果（Top5）:['goldfinch', 'brambling', 'junco', 'jay', 'bee eater']

結果2-2：（左）特異ベクトル
拡大されるベクトル（上位2つ）の入力画像との比較

識別結果（Top5）:['steel arch bridge', 'pier', 'planetarium', 'paddlewheel', 'dam']

結果2-2：（左）特異ベクトル

拡大されるベクトル（上位4つ）の入力画像との比較

識別結果（Top5）:['tench', 'barracouta', 'gar', 'sturgeon', 'eel']

結果2-2：（左）特異ベクトル
拡大されるベクトル（上位2つ）の入力画像との比較

識別結果（Top5）:['teapot', 'coffeepot', 'soup bowl', 'cup', 'candle']

結果2-2：（左）特異ベクトル
拡大されるベクトル（上位2つ）の入力画像との比較

識別結果（Top5）:['tiger cat', 'tabby', 'Egyptian cat', 'lynx', 'Persian cat']

結果2-2：（左）特異ベクトル
拡大されるベクトル（上位2つ）の入力画像との比較

識別結果（Top5）:['malamute', 'Eskimo dog', 'Siberian husky', 'timber wolf', 'white wolf']

結果2-2：（左）特異ベクトル
拡大されるベクトル（上位2つ）の入力画像との比較

識別結果（Top5）:['sports car', 'convertible', 'racer', 'beach wagon', 'car wheel']

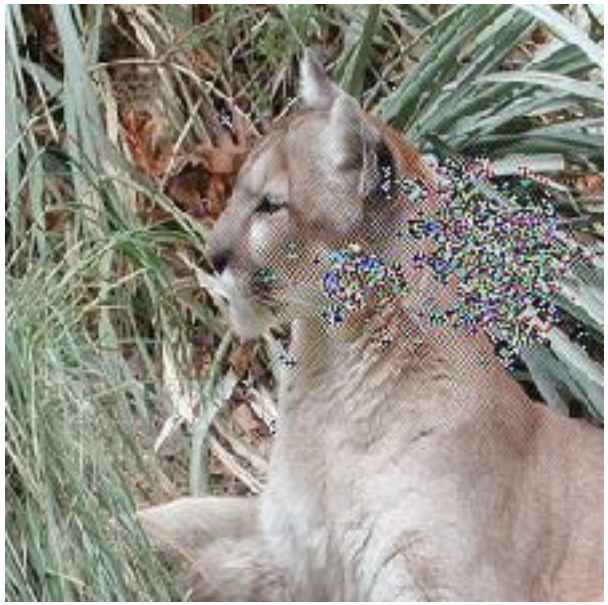
結果2-2：（左）特異ベクトル
拡大されるベクトル（上位2つ）の入力画像との比較

識別結果（Top5）s:['hen', 'cock', 'prairie chicken', 'partridge', 'black swan']

結果2-2：（左）特異ベクトル
拡大されるベクトル（上位2つ）の入力画像との比較

識別結果（Top5）:['tiger cat', 'tabby', 'Egyptian cat', 'lynx', 'tiger']

結果2-2：（左）特異ベクトル
拡大されるベクトル（上位2つ）の入力画像との比較



識別結果（Top5）:['cougar', 'weasel', 'lynx', 'otter', 'polecat']

結果2-2：（左）特異ベクトル
拡大されるベクトル（上位2つ）の入力画像との比較

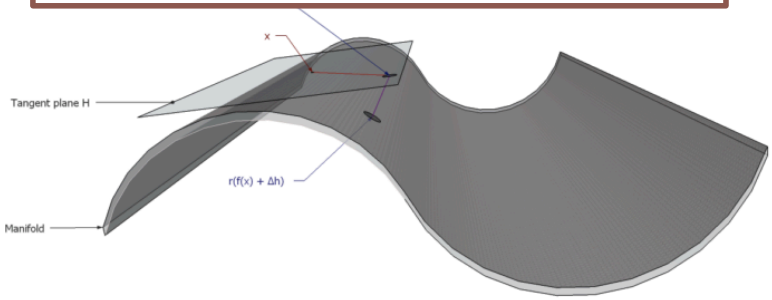
識別結果（Top5）:['great white shark', 'tiger shark', 'hammerhead', 'dugong', 'sturgeon']

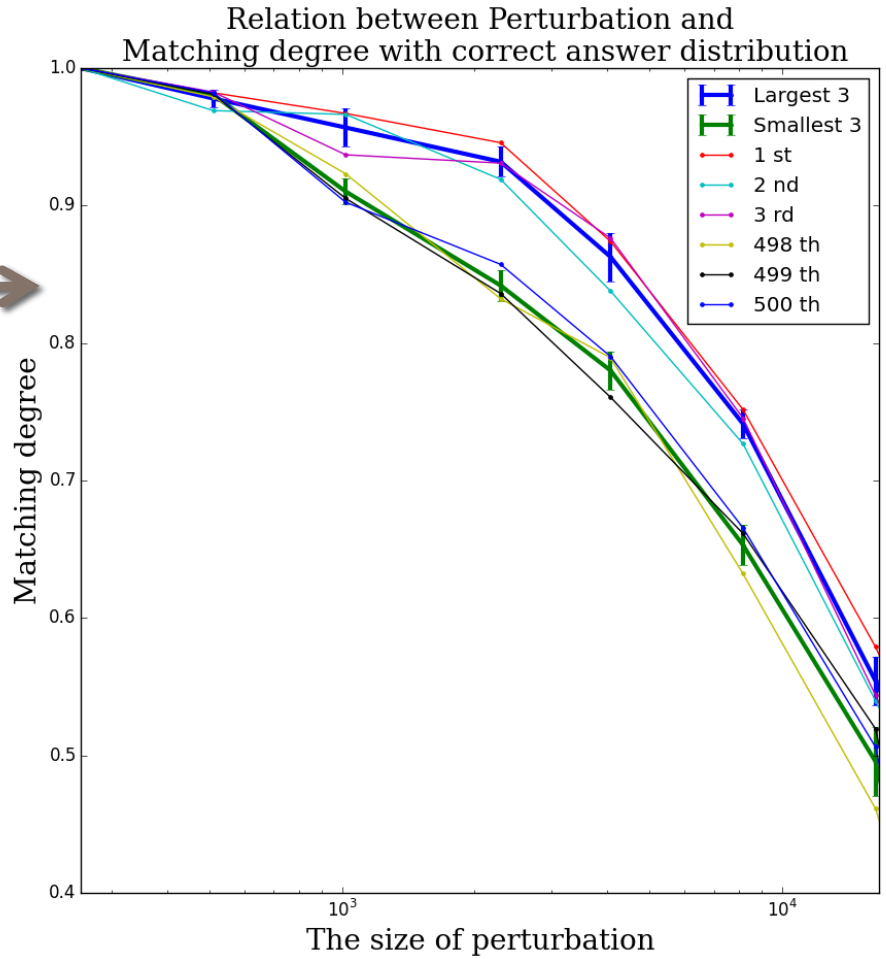
結果3 : 摂動に対する出力の挙動

青 : 多様体の接線方向への摂動付加による出力の変動

緑 : 多様体の垂直方向への摂動付加による出力の変動

X:摂動の大きさ
Y:摂動を加える前の出力との一致率





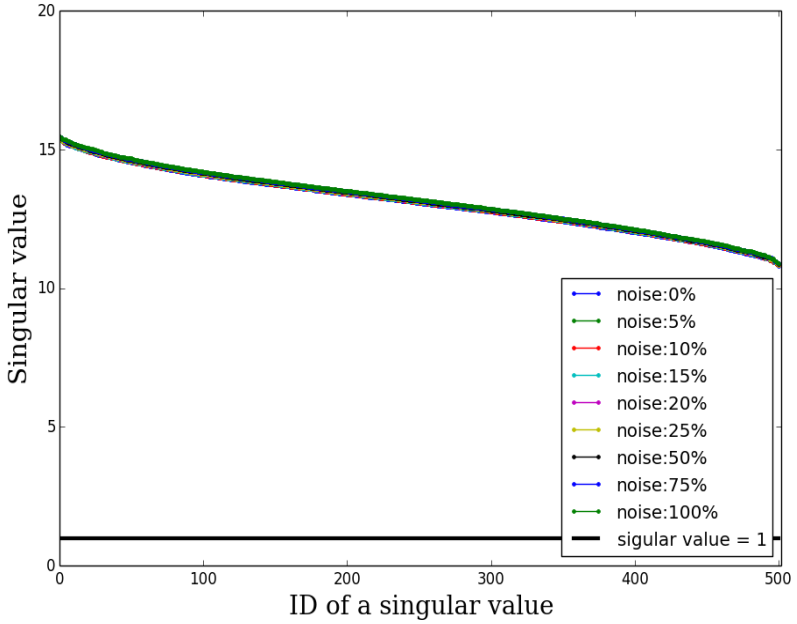
⇒多様体の接線方向に対応する摂動に対して出力が比較的ロバスト
⇒多様体の存在を示唆

エラーバーは標準偏差

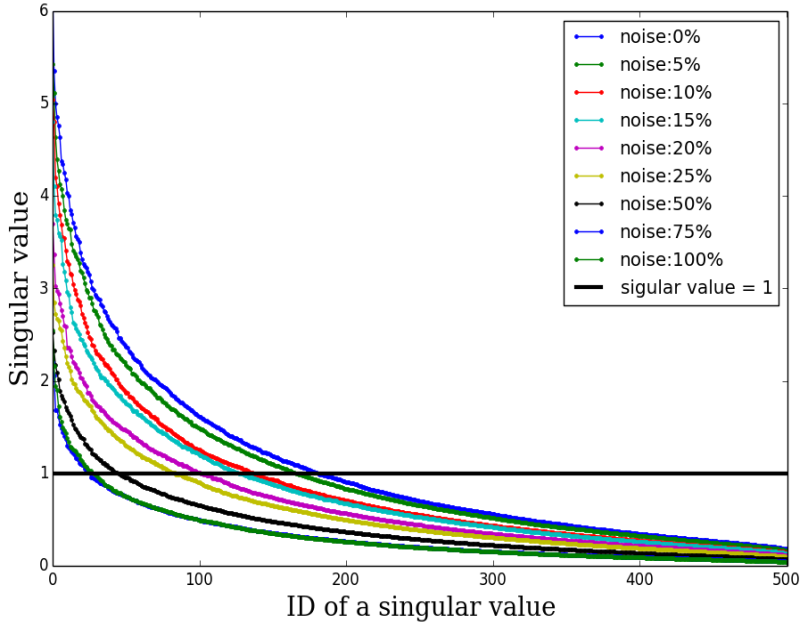
N=16×3

結果4：摂動と特異値分布

Layer 1 : Singular Values Distribution



Layer5 : Singular Values Distribution



ノイズの増加によって,情報の伝達が阻害されているようにみえる.

まとめと議論

◎Summary

1. 実際に活用されている,複雑なDNNネットワークにおいても,
急峻な特異値分布や,入力画像の特徴を捉えた特異ベクトルが観察された.
このことから,表象（多様体情報）を獲得できていることが示唆される.
(仮説が支持される.)

2. 摂動を加えた際の，出力の変動が,接線方向の摂動に対してロバストであることからも
仮説は支持される.

3. これらの結論を確定する為には,さらなる研究が必要.
特に,「階層」である必要性に対する研究が必要.

◎応用

ハイパーパラメータの決定において,特異値分布や
特異ベクトルを指標として利用できる可能性がある.

ご清聴ありがとうございました.