

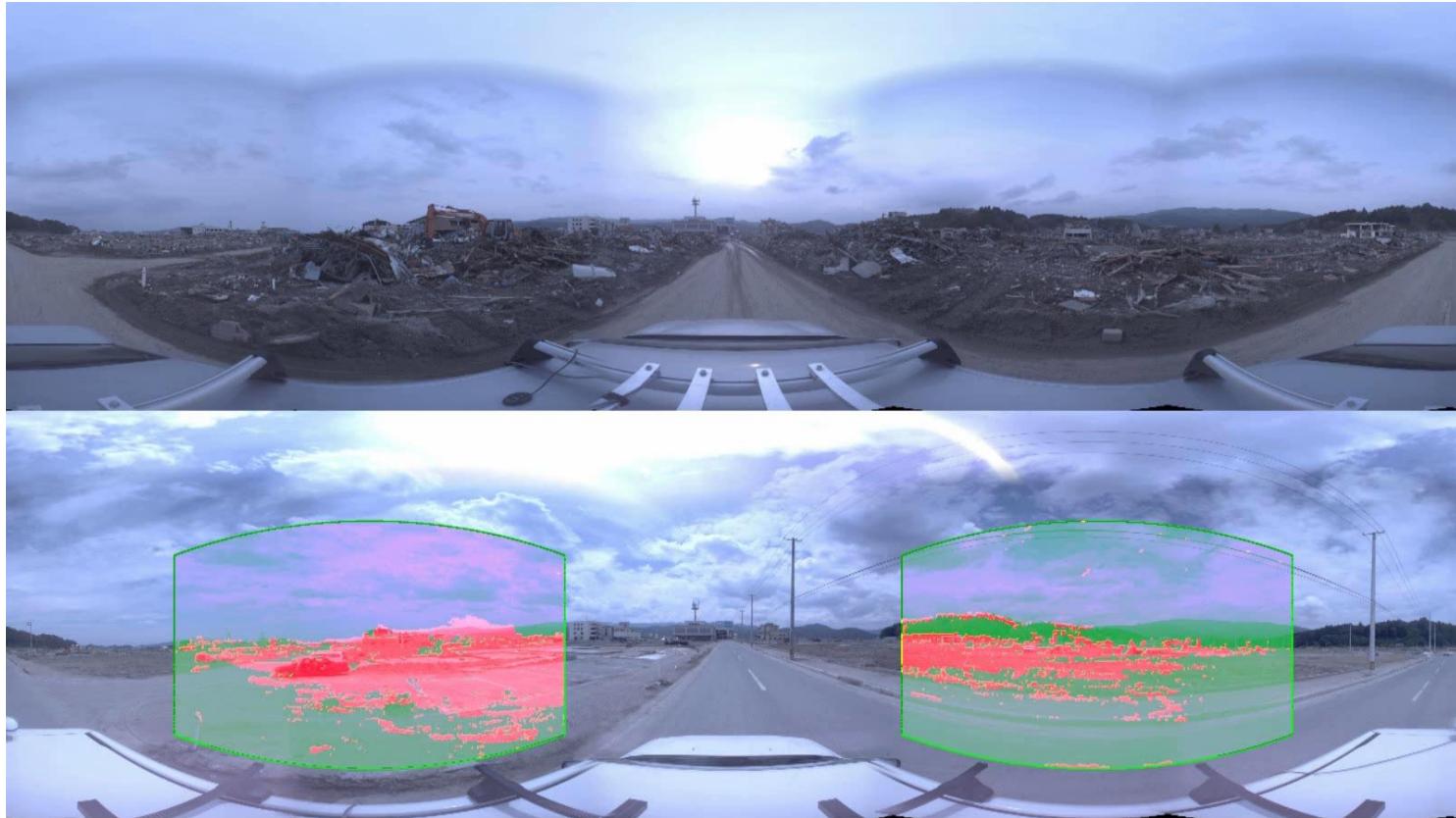
ディープラーニングと画像認識への応用

岡谷貴之

東北大学大学院情報科学研究科

自己紹介：市街地時間変化の推定

- 車載カメラ映像から市街地の形状変化を検出
 - Detecting Changes in 3D Structure of a Scene from Multi-view Images Captured by a Vehicle-mounted Camera, Sakurada+, CVPR2013



緑：検出対象，赤：変化部分

自己紹介：質感の画像認識

- ・ 質感の画像認識：ものの「質感」を画像から認識したい
- ・ 質感=人には分かるが測れない → 画像認識のアプローチで
 - 科研費新学術領域研究「質感脳情報学」



例えば質感属性の数値

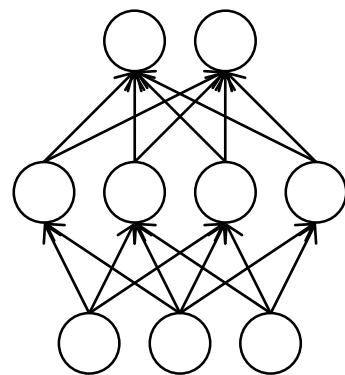
$$\begin{cases} \text{光沢感} = 0.8 \\ \text{透明感} = 0.0 \\ \text{滑らかさ} = 0.7 \end{cases}$$

画像から何を取り出すか？

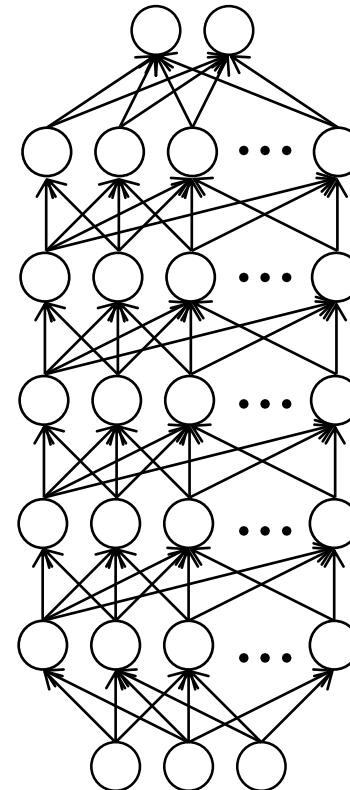
質感の画像特徴とは？

Deep Learningとは

- Deep Learning = 多層ニューラルネットを使った機械学習の方法論
- ニューラルネットの「ルネッサンス」



Shallow NN



Deep Neural Network (DNN)

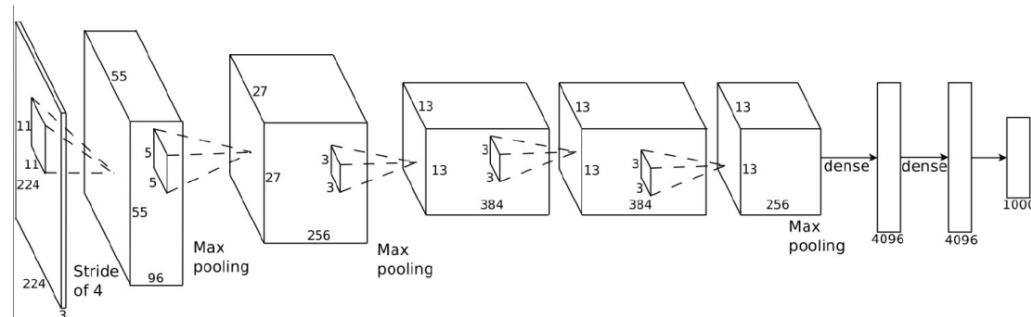
事例：一般物体認識 (Hintonのグループ)

Krizhevsk et al., ImageNet Classification with Deep Convolutional Neural Networks, NIPS2012

- IMAGENET Large Scale Visual Recognition Challenge 2012
 - 1000カテゴリ・カテゴリあたり約1000枚の訓練画像
 - CNN ; rectified linear unit ; drop-out

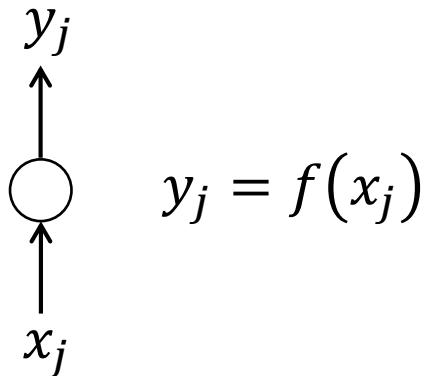


	Team name	Error (5 guesses)
1	SuperVision	0.15315
2	ISI	0.26172
3	OXFORD_VGG	0.26979
4	XRCE/INRIA	0.27058
5	University of Amsterdam	0.29576
6	LEAR-XRCE	0.34464

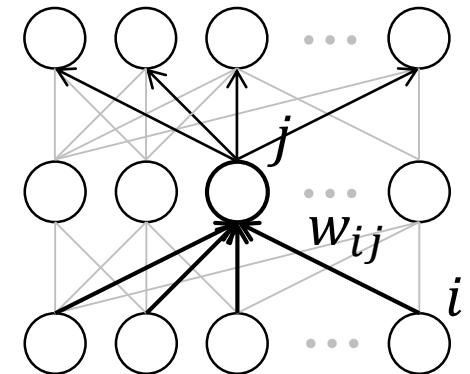


NNの基本要素

非線形入出力のユニット ネットワーク



$$x_j = b_j + \sum_i y_i w_{ij}$$

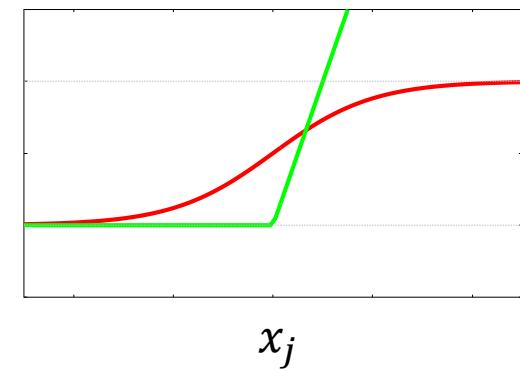


活性化 (activation) 関数

ロジスティックシグモイド $f(x_j) = \frac{1}{1 + e^{-x_j}}$

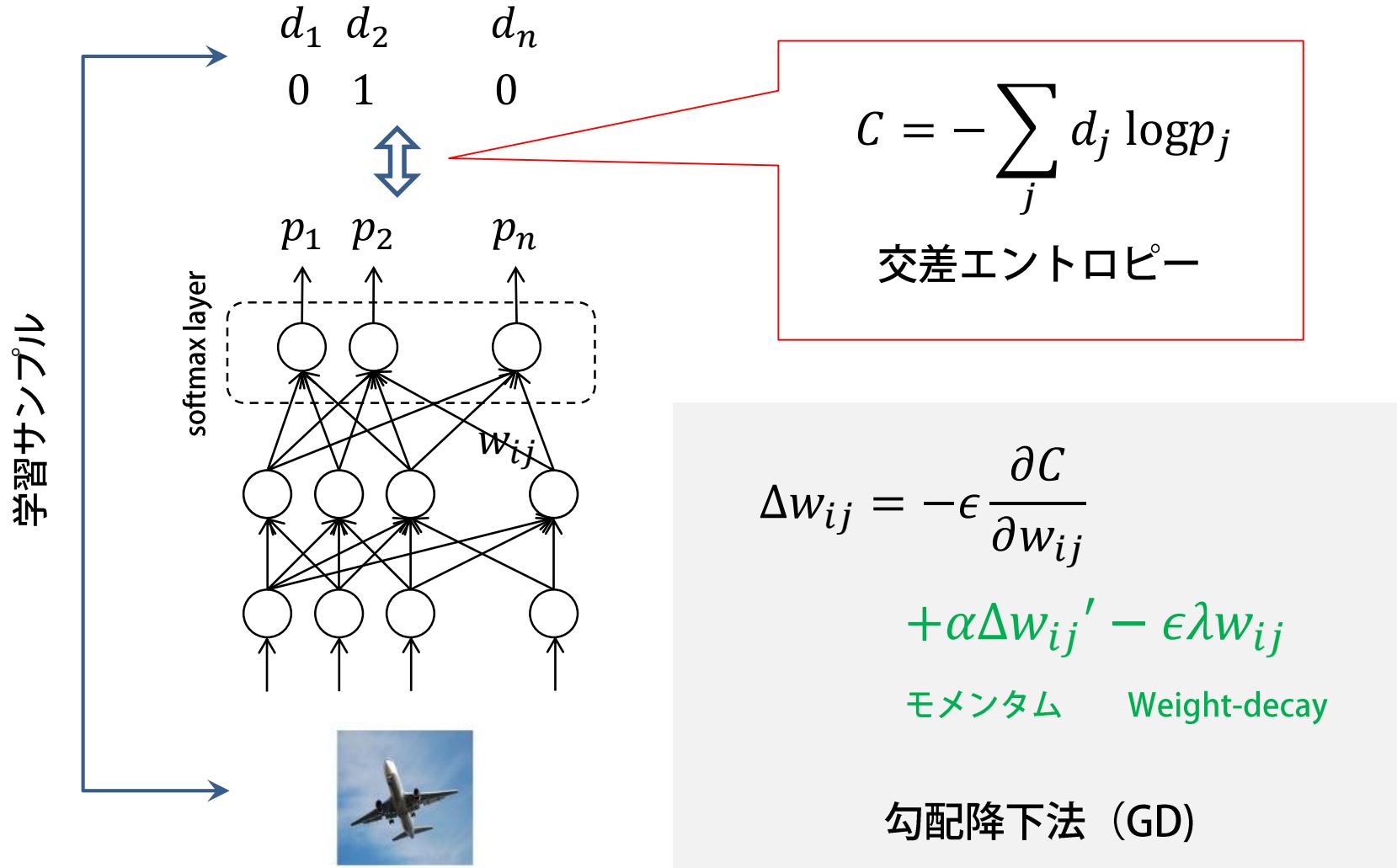
双曲線正接関数 $f(x_j) = \tanh(x_j)$

Rectified linear $f(x_j) = \max(x_j, 0)$

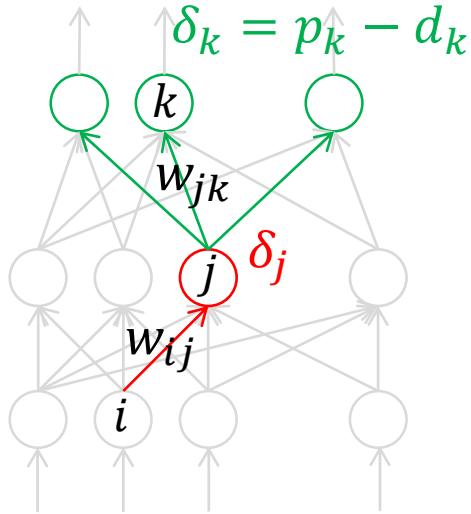
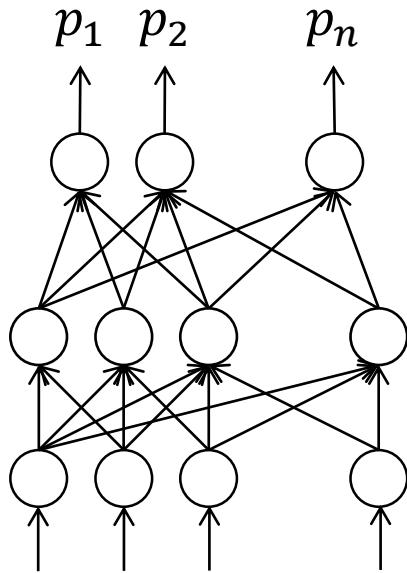


識別モデルの学習

- 出力の誤差が小さくなるように重み $\{w_{ij}\}, \{b_j\}$ を調節



勾配の計算 = 誤差逆伝搬法 (Backprop)



勾配の連鎖計算

$$\frac{\partial C}{\partial w_{ij}} = \delta_j y_i$$

$$\delta_j = f'(x_j) \sum_k \delta_k w_{jk}$$

- 確率的勾配降下法 (SGD)
 - 全データの誤差勾配は困難；サンプル 1 個ずつは効率が悪い
 - **ミニバッチ**：数個～100個程度のサンプルの誤差勾配を使う
- 勾配降下法の代わりに準ニュートン法や共役勾配法も使える [Le+11]

NN研究の歴史

第1期

「冬の時代」

第2期

「冬の時代」

第3期

1960

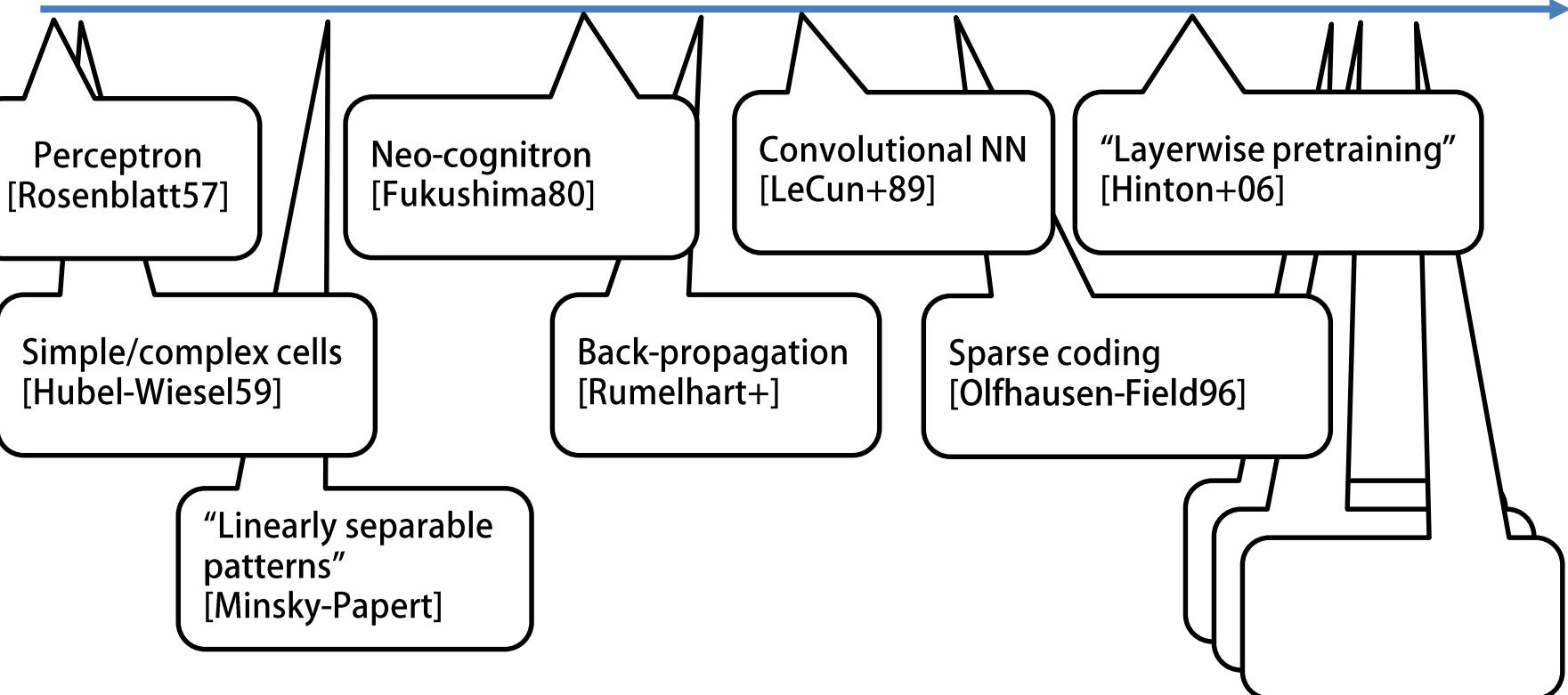
1970

1980

1990

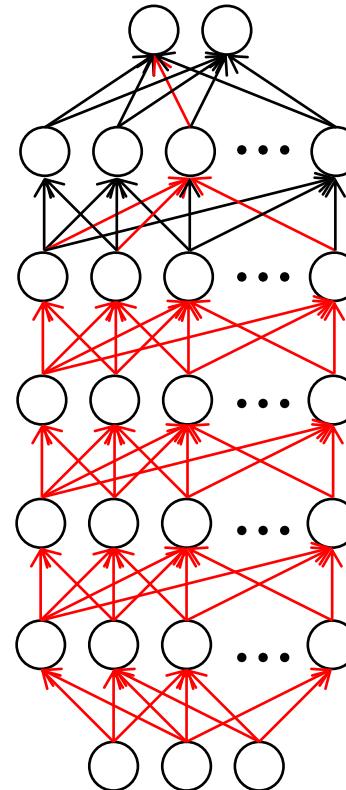
2000

2010



DNNの扱いづらさ＝学習の困難さ

- 過学習が起こる：訓練誤差は小さくなるが汎化誤差は小さくならない
- 現象
 - 下層まで情報が伝わらない；勾配が拡散する
 - 訓練データは上位層のみで表現できてしまう



DNNを手なずける 3つの方法

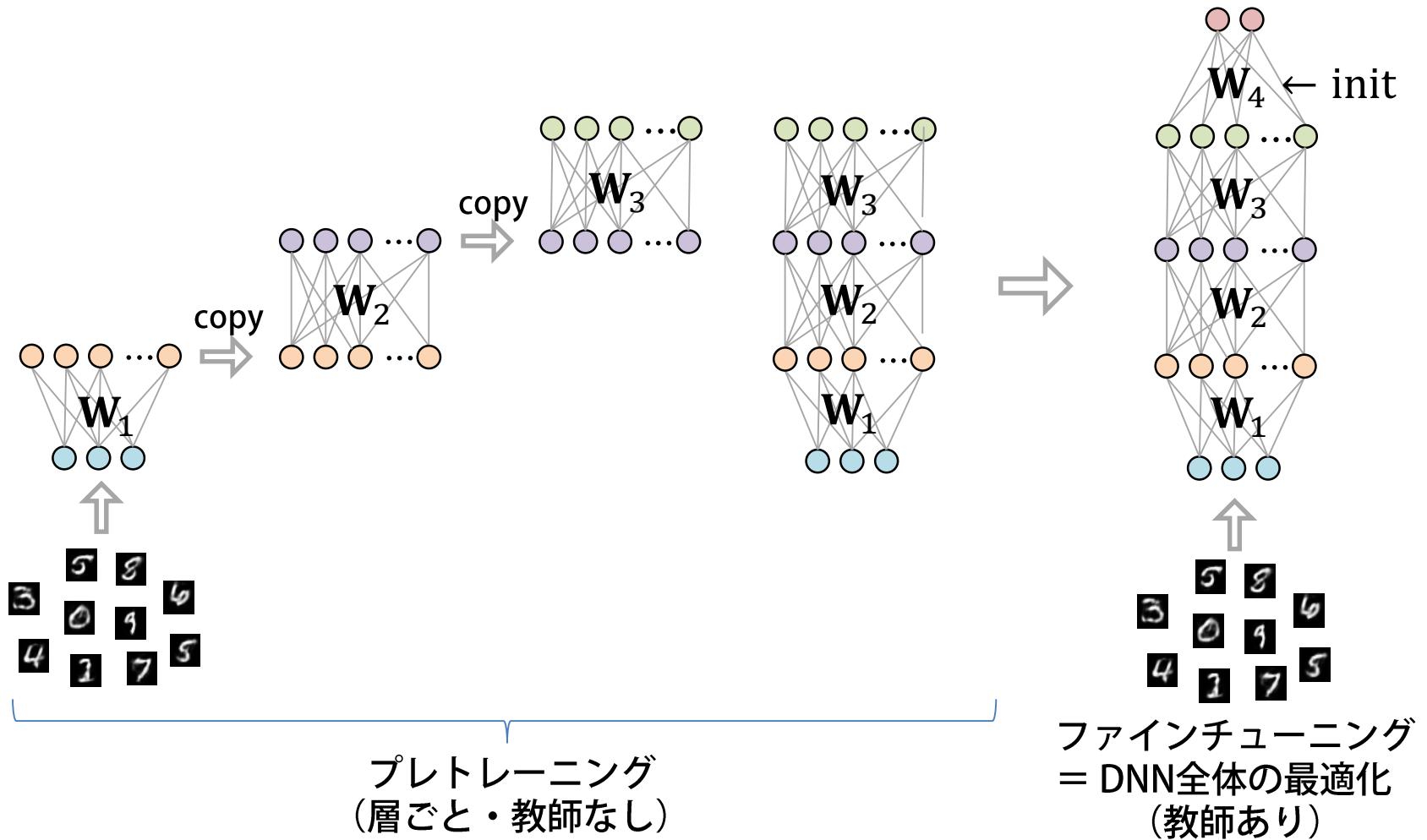
- 教師なしのプレトレーニング
- たたみこみニューラルネット
- 第3の方法

DNNを手なずける 3つの方法

- 教師なしのプレトレーニング
- たたみこみニューラルネット
- 第3の方法

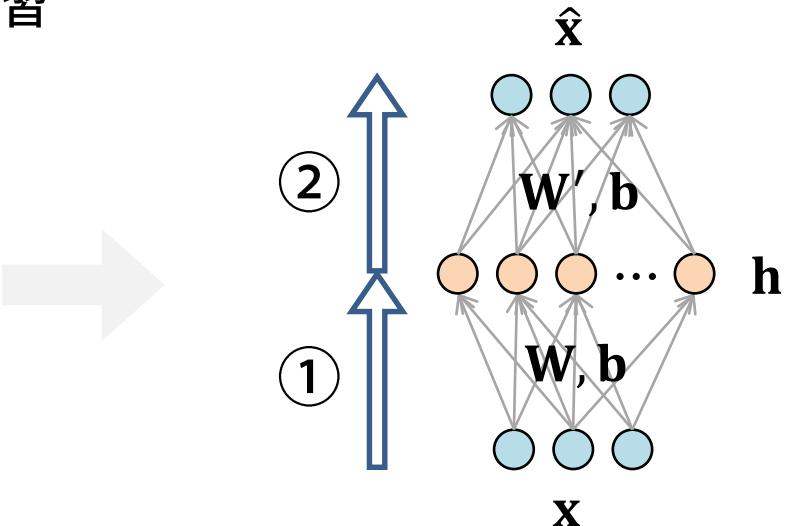
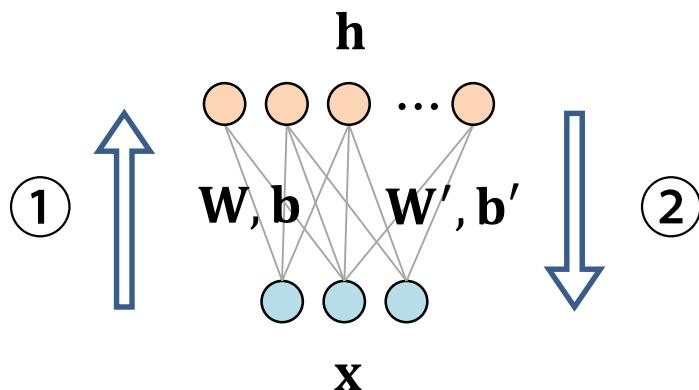
DNNのプレトレーニング

- 層ごとに教師なし学習を順番に実行 (=greedy layerwise—)



オートエンコーダ (Autoencoder)

- 入力サンプルをよく再現するように
– フィードフォワードNNとして学習



$$\textcircled{1} \quad h(x_i) = f(Wx_i + b)$$

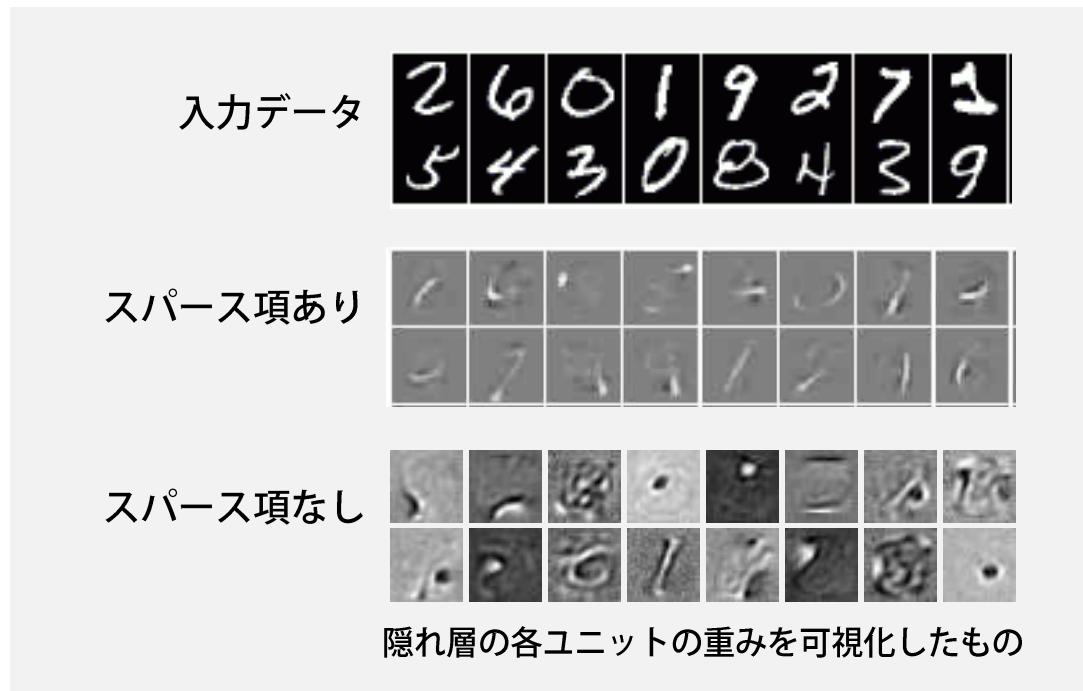
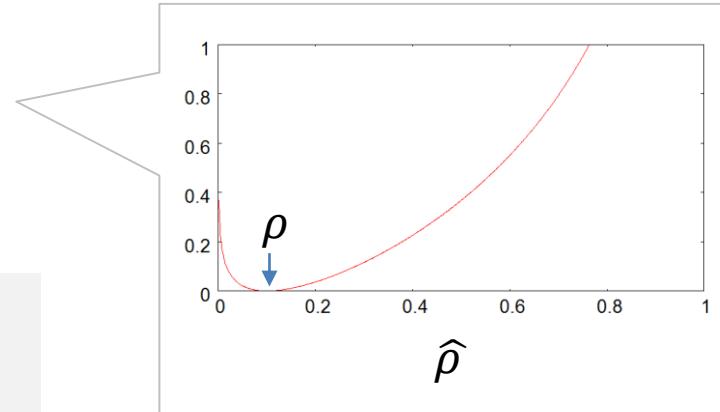
$$\textcircled{2} \quad \hat{x}(x_i) = f'(W'h(x_i) + b')$$

$$\min_{\theta} \sum_i \|x_i - \hat{x}(x_i)\|^2$$
$$\theta = \{W, b, W', b'\}$$

スパースオートエンコーダ

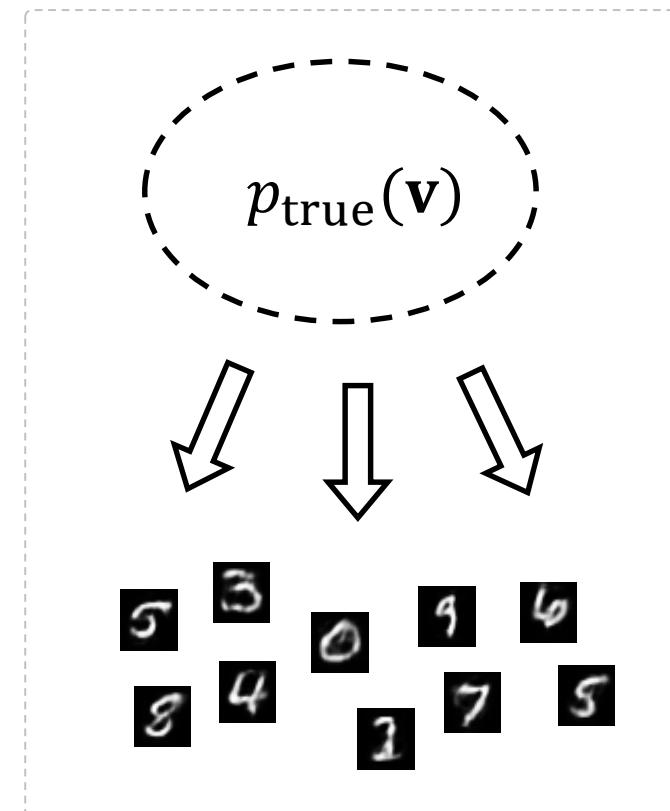
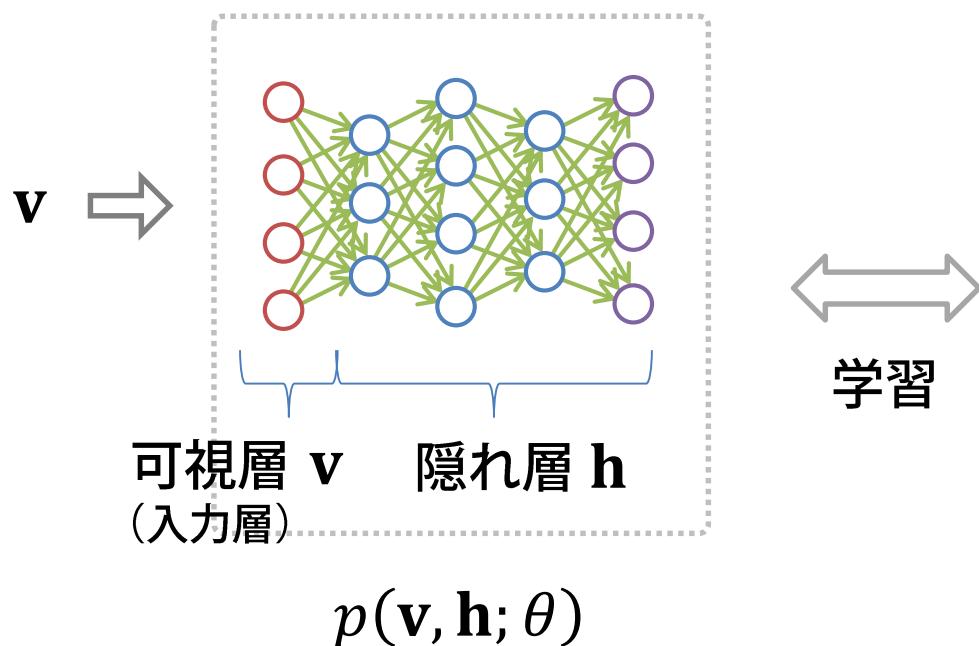
- 出力層の活性度がスパースになるように正則化
= 各サンプルにつき、わずかな数のユニットのみ活性化

$$\min_{\theta} \frac{1}{N} \sum_{i=1}^N \|\mathbf{x}_i - \hat{\mathbf{x}}(\mathbf{x}_i)\|^2 + \beta \sum_{j=1}^{\text{\# of units}} \text{KL}(\hat{\rho}_j \| \rho)$$



ボルツマンマシン (Boltzmann Machine)

- NNの挙動を確率的に捉える

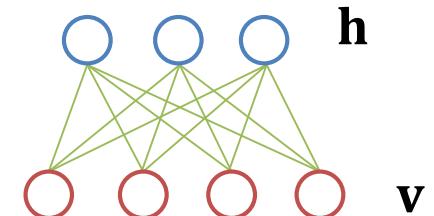


Restricted Boltzmann Machine (RBM)

- 入力層同士、隠れ層同士に結合がない※

$$h_j = \{0,1\}$$

$$p(\mathbf{v}, \mathbf{h}; \theta) = \frac{1}{Z(\theta)} \exp(-E(\mathbf{v}, \mathbf{h}; \theta))$$



$$E(\mathbf{v}, \mathbf{h}; \theta) = - \sum_i b_i v_i - \sum_j c_j h_j - \sum_{i,j} v_i w_{ij} h_j$$

$$\theta = \{W, \mathbf{b}, \mathbf{c}\}$$

バイアス

ウェイト

$$v_i = \{0,1\}$$

各ユニットは
2値の状態を
とるとする

※の性質から、簡素な条件付き分布が得られる

$$p(h_j = 1 | \mathbf{v}; \theta) = \sigma \left(c_j + \sum_i v_i w_{ij} \right)$$

ロジスティックシグモイド関数

$$p(v_i = 1 | \mathbf{h}; \theta) = \sigma \left(b_i + \sum_j w_{ij} h_j \right)$$

RBMの学習

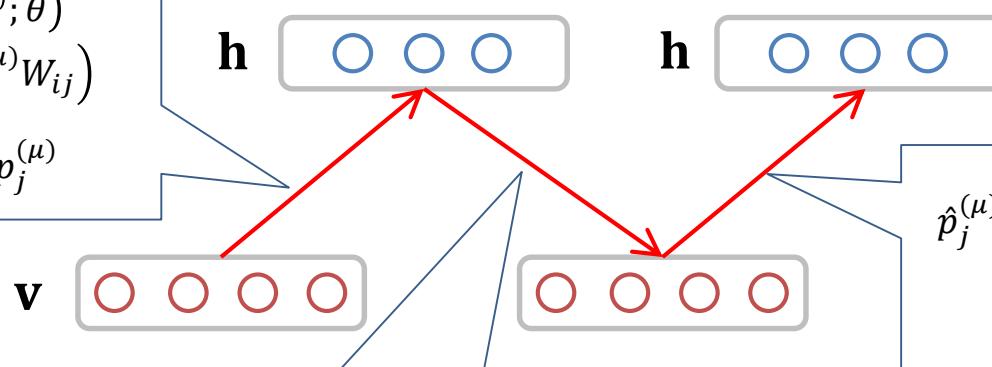
“Contrastive Divergence”

$$\frac{\partial L}{\partial w_{ij}} = \frac{1}{N} \sum_{\mu=1}^N v_i^{(\mu)} p_j^{(\mu)} - \text{E}_{model}[v_i h_j | \theta]$$

$$\Delta w_{ij} = \epsilon \left(\frac{1}{N} \sum_{\mu=1}^N v_i^{(\mu)} h_j^{(\mu)} - \frac{1}{N} \sum_{\mu=1}^N \hat{v}_i^{(\mu)} \hat{p}_j^{(\mu)} \right)$$

$$p_j^{(\mu)} \equiv p(h_j = 1 | \mathbf{v}^{(\mu)}; \theta) \\ = \sigma(c_j + \sum_i v_i^{(\mu)} W_{ij})$$

$$h_j^{(\mu)} \in \{0,1\} \text{ from } p_j^{(\mu)}$$



$$\hat{p}_j^{(\mu)} \equiv p(h_j = 1 | \mathbf{v}^{(\mu)}; \theta) \\ = \sigma(c_j + \sum_i \hat{v}_i^{(\mu)} W_{ij})$$

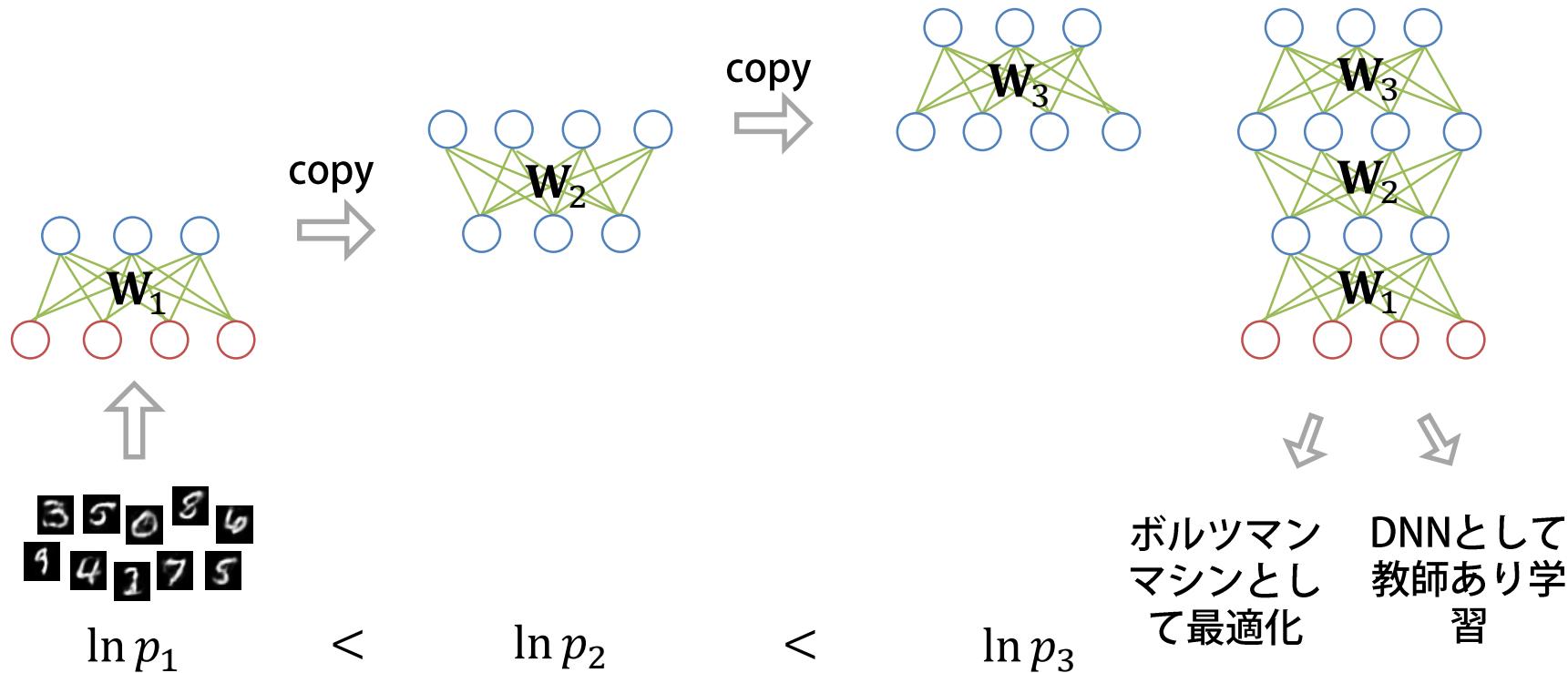
$$\hat{p}_i^{(\mu)} \equiv p(v_i = 1 | \mathbf{h}^{(\mu)}; \theta) \\ = \sigma(b_i + \sum_j W_{ij} h_j^{(\mu)})$$

$$\hat{v}_i^{(\mu)} \in \{0,1\} \text{ from } \hat{p}_i^{(\mu)}$$

多層ボルツマンマシンのプレトレーニング

Hinton et al., A Fast Learning Algorithm for Deep Belief Nets, Neural Computation, 2006

- Deep Belief Network(DBN), Deep Boltzmann Machine(DBM)
- 層ごとにRBMの学習を順番に実行



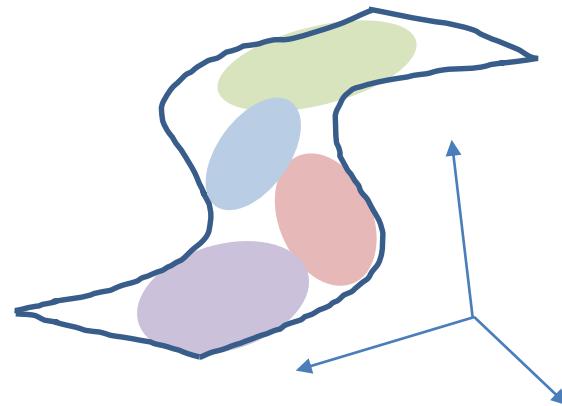
プレトレーニングの効果

- よい初期値を与えると同時に一種の正則化として機能 [Larochelle+09]
 - 正則化=汎化や分布した内部表現を促進
- データを正確に表現する特徴は識別にも役立つ
 - すべてが有効なわけではないとしても

音声認識 [Hinton+12]

[TABLE 1] COMPARISONS AMONG THE REPORTED SPEAKER-INDEPENDENT (SI) PHONETIC RECOGNITION ACCURACY RESULTS ON TIMIT CORE TEST SET WITH 192 SENTENCES.

METHOD	PER
CD-HMM [26]	27.3%
AUGMENTED CONDITIONAL RANDOM FIELDS [26]	26.6%
RANDOMLY INITIALIZED RECURRENT NEURAL NETS [27]	26.1%
BAYESIAN TRIPHONE GMM-HMM [28]	25.6%
MONOPHONE HTMS [29]	24.8%
HETEROGENEOUS CLASSIFIERS [30]	24.4%
MONOPHONE RANDOMLY INITIALIZED DNNs (SIX LAYERS) [13]	23.4%
MONOPHONE DBN-DNNs (SIX LAYERS) [13]	22.4%
MONOPHONE DBN-DNNs WITH MMI TRAINING [31]	22.1%
TRIPHONE GMM-HMMs DT W/ BMMI [32]	21.7%
MONOPHONE DBN-DNNs ON FBANK (EIGHT LAYERS) [13]	20.7%
MONOPHONE MCRBM-DBN-DNNs ON FBANK (FIVE LAYERS) [33]	20.5%
MONOPHONE CONVOLUTIONAL DNNs ON FBANK (THREE LAYERS) [34]	20.0%



Larochelle et al., Exploring Strategies for Training Deep Neural Networks, JMLR, 2009

Hinton et al., Deep Neural Networks for Acoustic Modeling in Speech Recognition, IEEE SP magazine, Nov. 2012

画像認識とプレトレーニングの必要性

- 画像認識の state-of-the-art はすべて **たたみこみニューラルネット**
 - プレトレーニング不要
 - 80年代にすでに多層の学習に成功 [LeCun+89]

音声認識

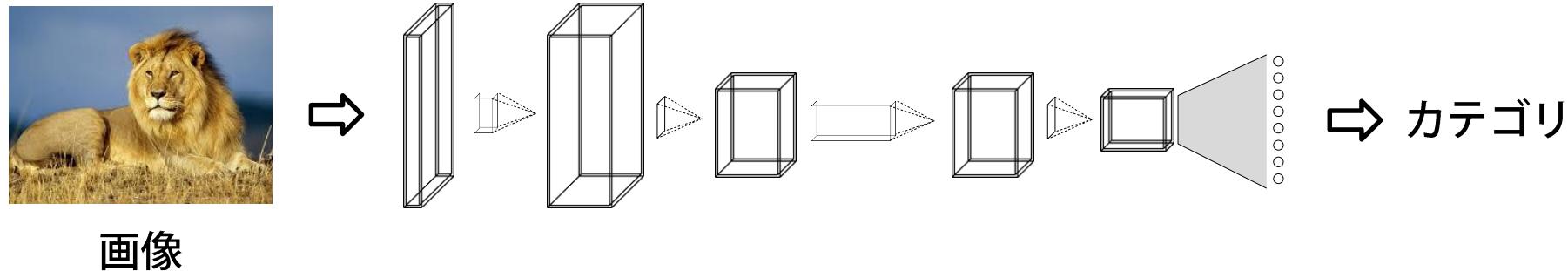
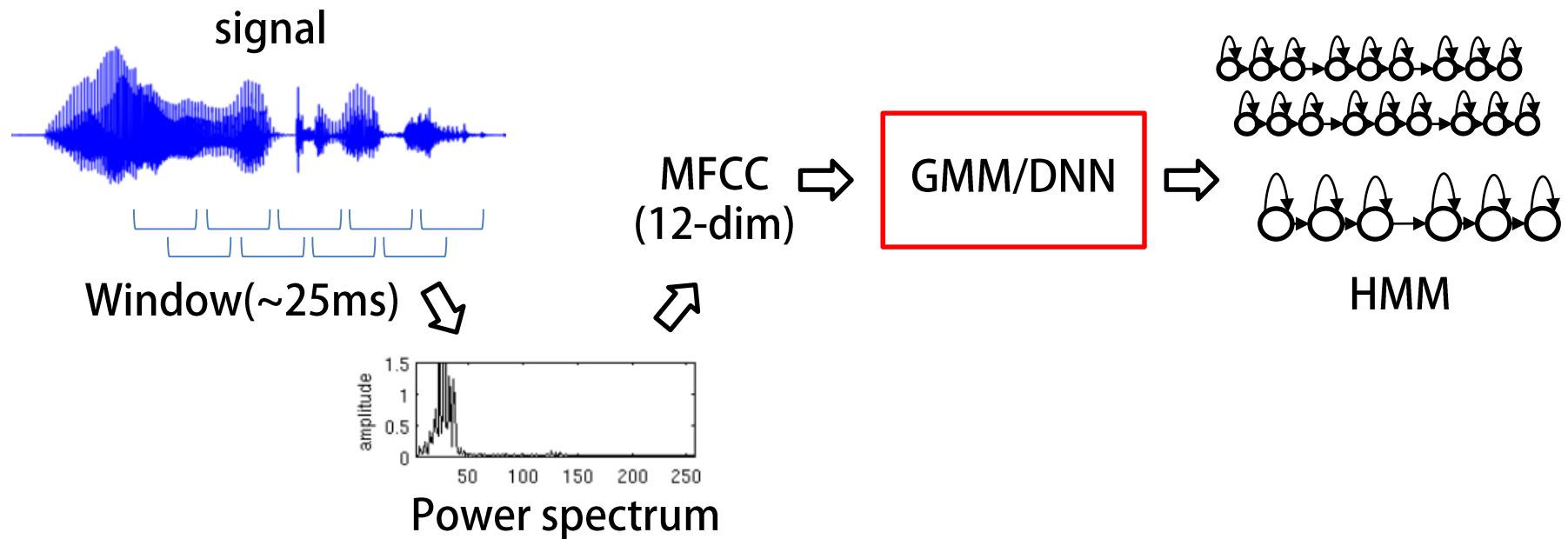
[TABLE 1] COMPARISONS AMONG THE REPORTED SPEAKER-INDEPENDENT (SI) PHONETIC RECOGNITION ACCURACY RESULTS ON TIMIT CORE TEST SET WITH 192 SENTENCES.

METHOD	PER
CD-HMM [26]	27.3%
AUGMENTED CONDITIONAL RANDOM FIELDS [26]	26.6%
RANDOMLY INITIALIZED RECURRENT NEURAL NETS [27]	26.1%
BAYESIAN TRIPHONE GMM-HMM [28]	25.6%
MONOPHONE HTMS [29]	24.8%
HETEROGENEOUS CLASSIFIERS [30]	24.4%
MONOPHONE RANDOMLY INITIALIZED DNNs (SIX LAYERS) [13]	23.4%
MONOPHONE DBN-DNNs (SIX LAYERS) [13]	22.4%
MONOPHONE DBN-DNNs WITH MMI TRAINING [31]	22.1%
TRIPHONE GMM-HMMS DT W/ BMMI [32]	21.7%
MONOPHONE DBN-DNNs ON FBANK (EIGHT LAYERS) [13]	20.7%
MONOPHONE MCRBM-DBN-DNNs ON FBANK (FIVE LAYERS) [33]	20.5%
MONOPHONE CONVOLUTIONAL DNNs ON FBANK (THREE LAYERS) [34]	20.0%

画像認識

カテゴリ	データ	著者等	方法
一般物体認識	ILSVRC2012	Supervision	CNN w/o pretraining
	CIFAR10	Ciresan+12	CNN w/o pretraining
	NORB	Ciresan+12	CNN w/o pretraining
	「猫細胞」	Le+12	再構成TICA
文字認識	MNIST	Ciresan+12	CNN w/o pretraining
	HWDB1.0	Ciresan+12	CNN w/o pretraining

音声認識と画像認識



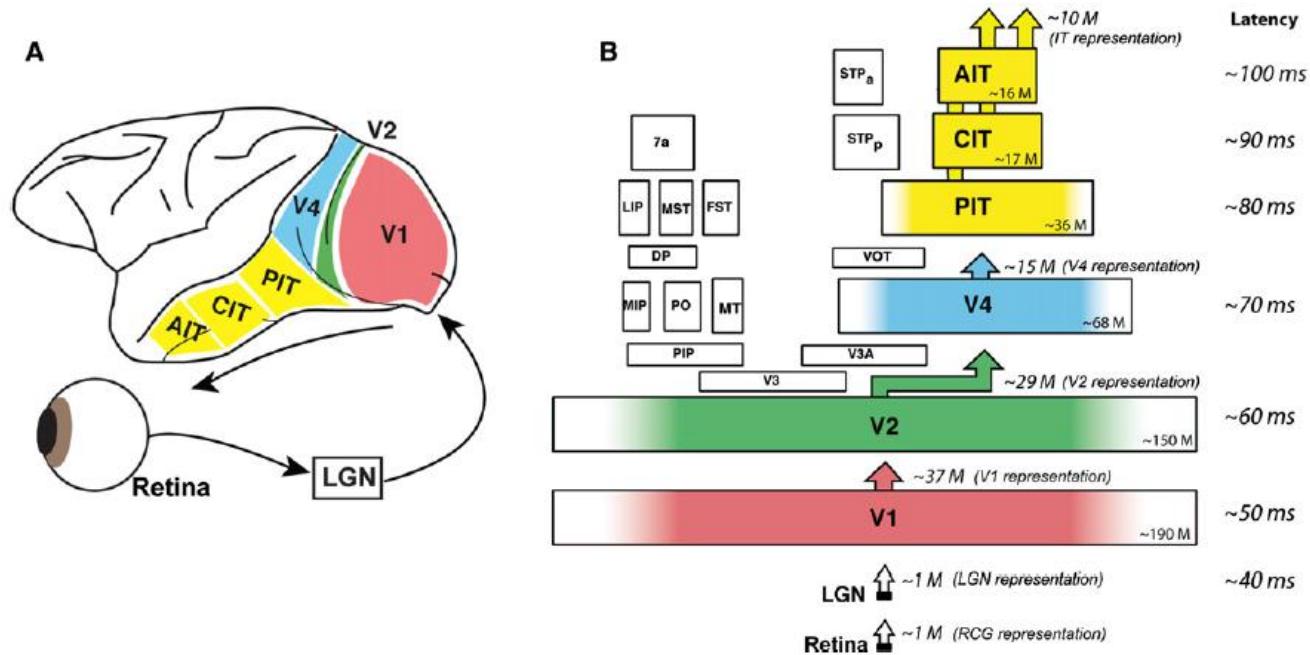
DNNを手なずける 3つの方法

- 教師なしのプレトレーニング
- たたみこみニューラルネット
- 第3の方法

脳の視覚情報処理

DiCarlo, Zoccolan, Rust, How does the brain solve visual object recognition?, Neuron, 2012

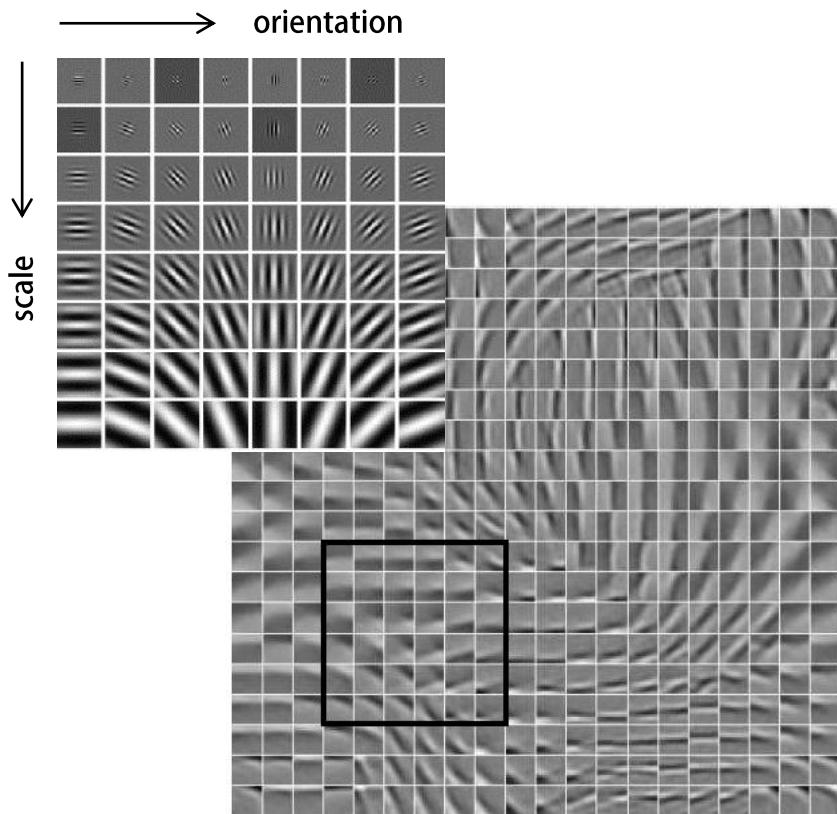
- 視覚野（腹側皮質視覚路）の構造
 - フィードフォワードで伝播
 - 階層性：単純な特徴抽出 → 複雑なものへ



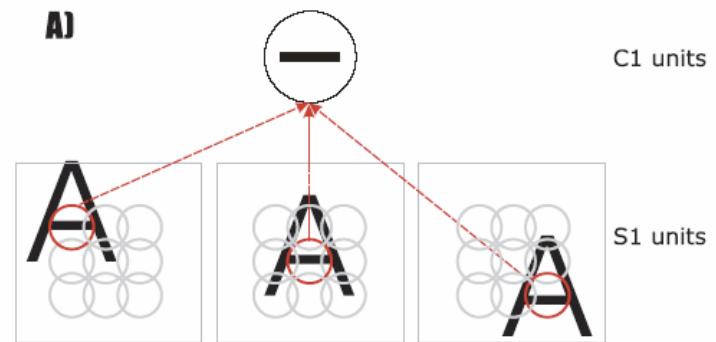
[Dicarlo+12]

V1

- ガボールウェーブレット
 - 位置 / 向き / スケール
 - Topographic map



- Simple cells/complex cells [Huber-Wiesel59]



Serre et al, Object Recognition with Features Inspired by Visual Cortex, CVPR05

たたみこみニューラルネット

Convolutional Neural Network (CNN)

- Neocognitronにルーツ [Fukushima80]
- Backpropagationによる教師有学習と手書き文字認識への応用 [LeCun+89]
 - Backpropagation Applied to Handwritten Zip Code Recognition, *Neural Computation*, 1989
- 神経科学の知見が基礎
 - Hubel-Wiesel の単純細胞・複雑細胞
 - 局所受容野 (local receptive field)

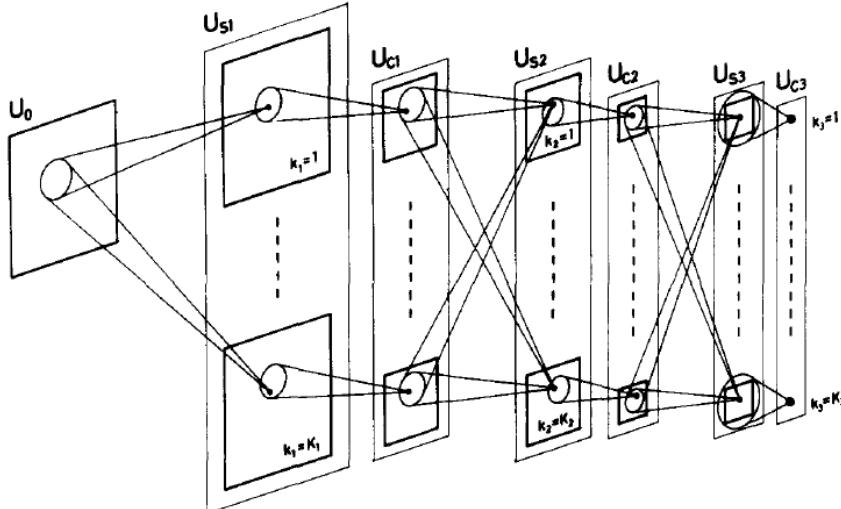


Fig 4 Schematic diagram illustrating the interconnections between layers in the neocognitron

[Fukushima+83]

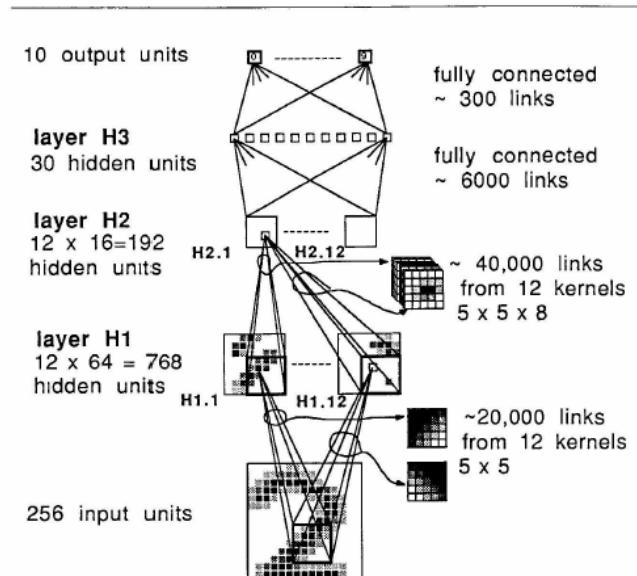
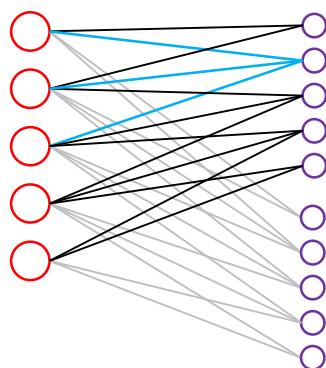
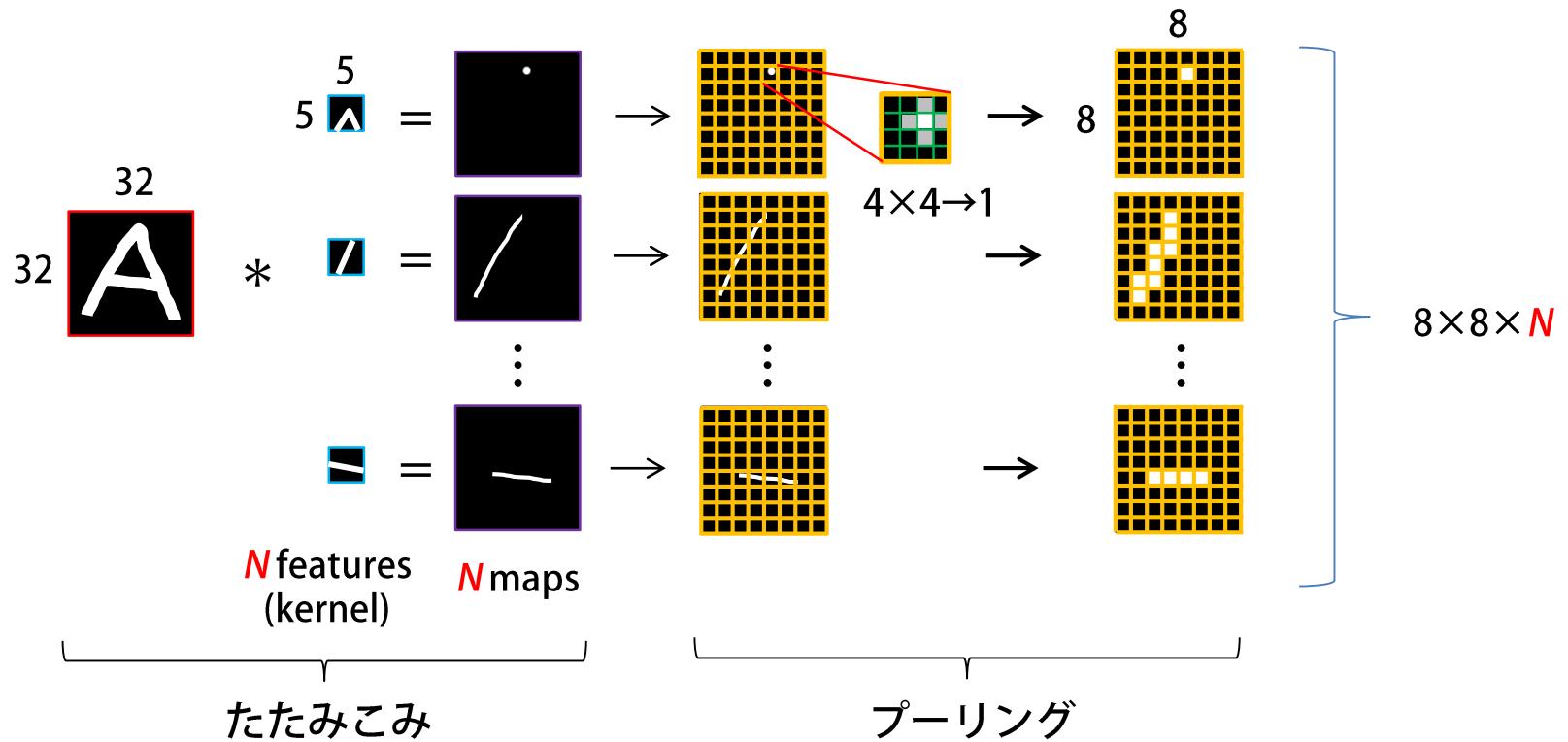


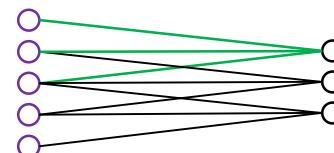
Figure 3 Log mean squared error (MSE) (top) and raw error rate (bottom) versus number of training passes

[LeCun+89]

たたみこみニューラルネット

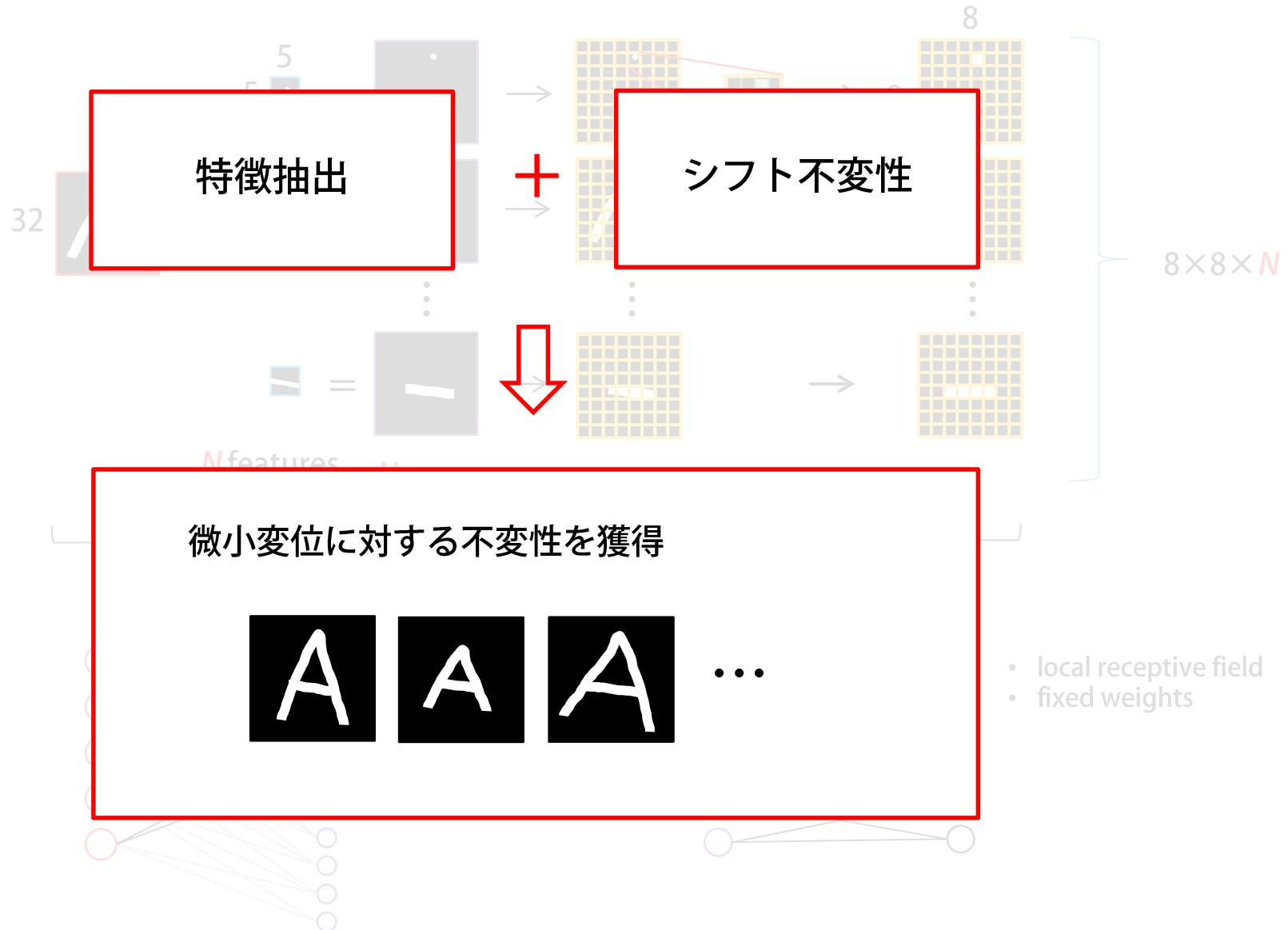


- local receptive field
- tied weights



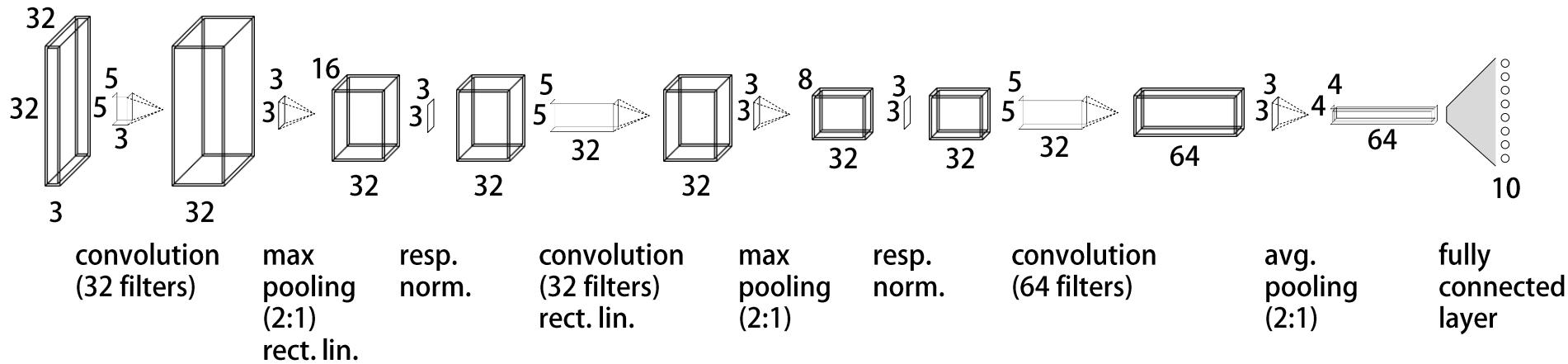
- local receptive field
- fixed weights

たたみこみニューラルネット



たたみこみニューラルネット

- たたみこみ+プーリングを繰り返す (= Deep CNN) ことで、多様な変形に対する不变性を獲得
- フィルタと上位の全結合層を勾配降下法 (Backprop) で学習



CIFAR10用 CNN [Krizhevsky+11]

一般物体認識はなぜ難しいか？

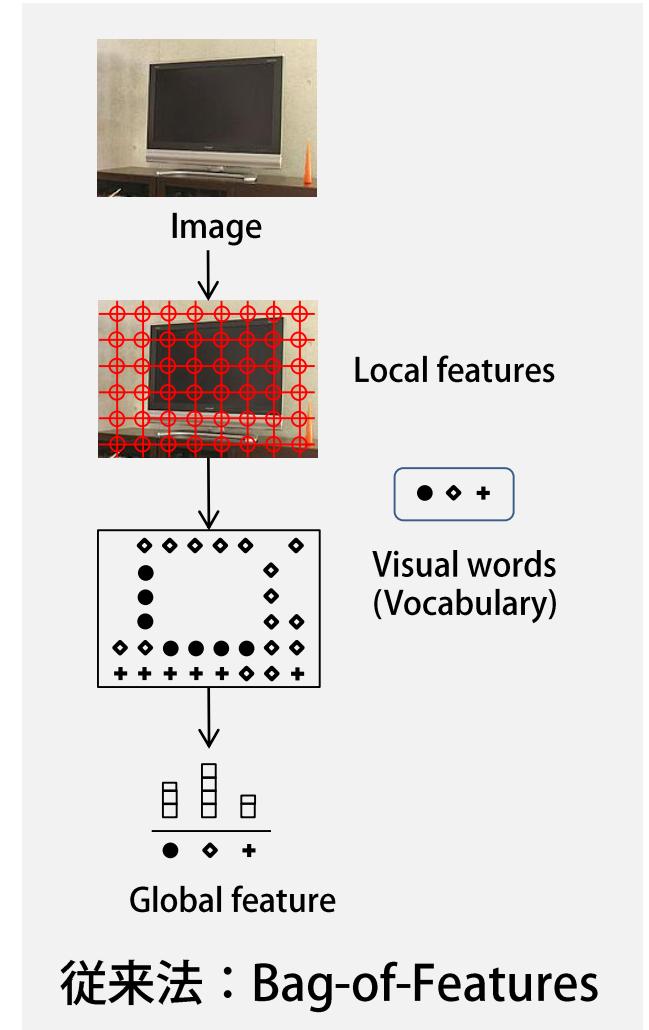
- カテゴリ内変動が多大
 - 変動に対する不变性と弁別力を備えた画像特徴表現を得るのが難しい

プーリング

たたみこみ



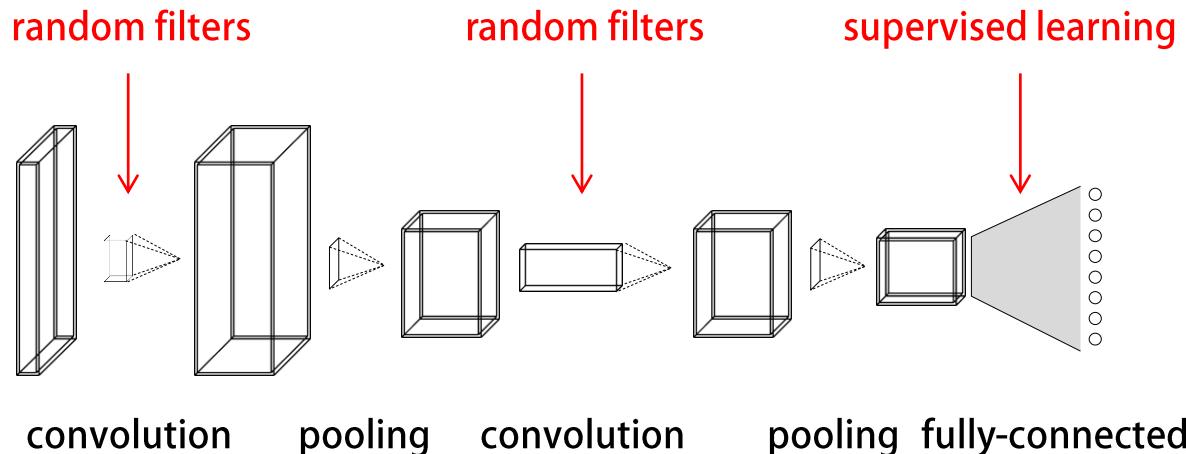
“Television set”



ランダムフィルタ：アーキテクチャの重要性

Jarrett et al., What is the best multi-stage architecture for object recognition? ICCV09

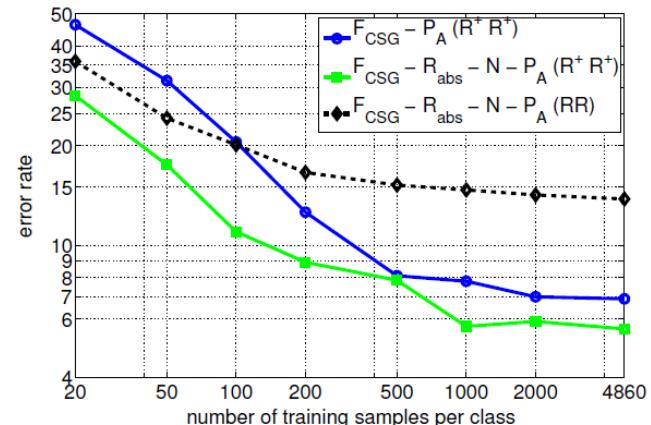
- フィルタをランダムとし、最上位層の fully-connected 層のみ学習



Caltech-101

アーキテクチャ	ランダム フィルタ	フィルタも 学習
2層, 絶対値プーリング	62.9%	64.7%
2層, 平均プーリング	19.6%	31.0%
1層, 絶対値プーリング	53.3%	54.8%

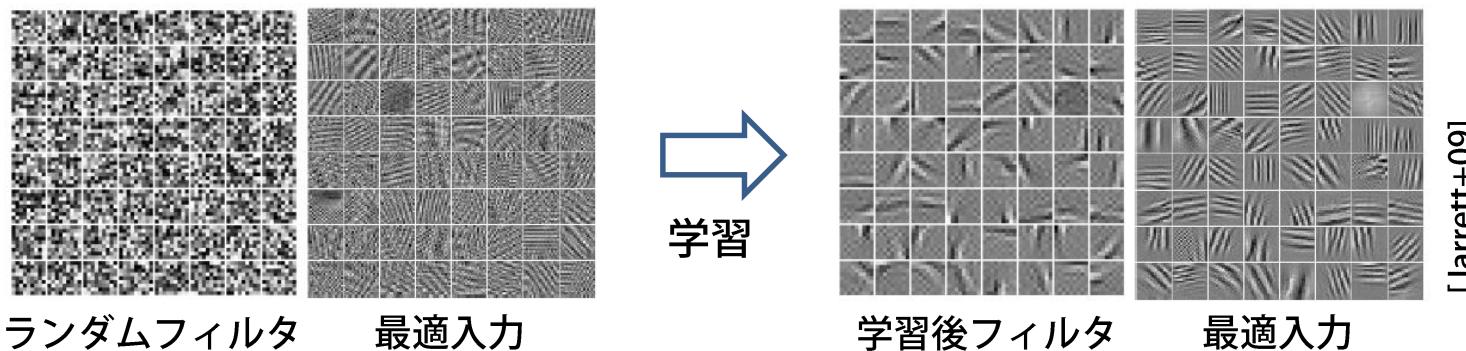
NORB



ランダムフィルタ：アーキテクチャの重要性

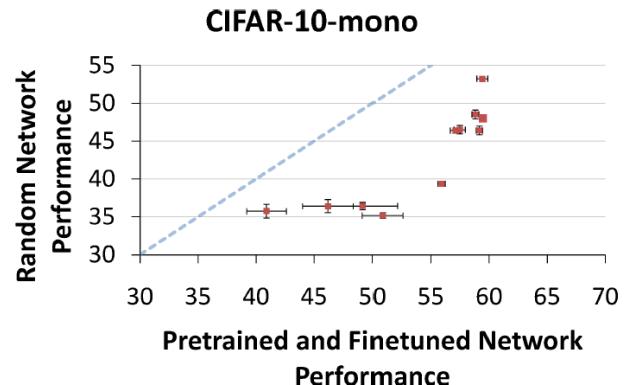
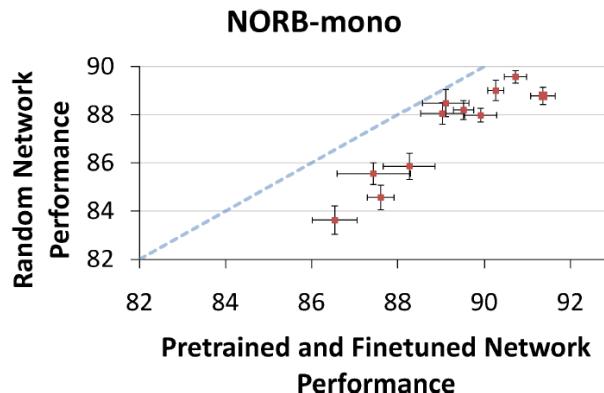
Saxe et al., On random weights and unsupervised feature learning, ICML2010

- 学習アルゴリズムよりもアーキテクチャがずっと大事
- プーリング層のユニットが最も反応する最適入力：
 - 理論的説明 [Saxe+10]



[Jarrett+09]

- アーキテクチャの性能予測をランダムフィルタで [Saxe+10]
 - アーキテクチャ探索時間を節約可



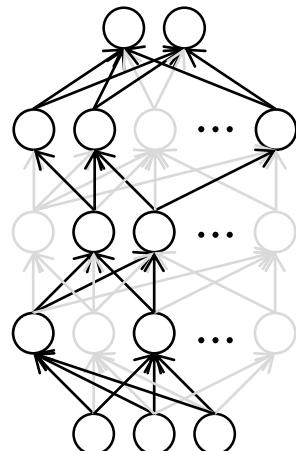
DNNを手なずける 3つの方法

- 教師なしのプレトレーニング
- たたみこみニューラルネット
- 第3の方法

第3の方法

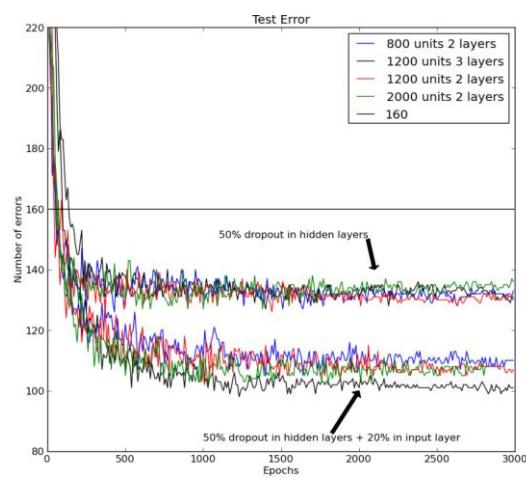
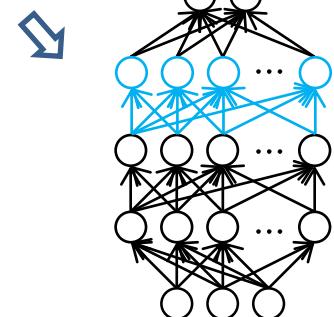
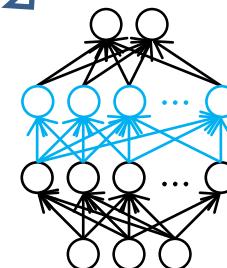
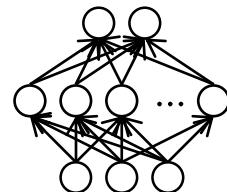
ドロップアウト [Hinton+12]

- ランダムに隠れユニットを省いて学習



識別的プレトレーニング [Seide+11]

- 浅い方から深いネットへ、教師あり学習を反復



MNISTでの結果

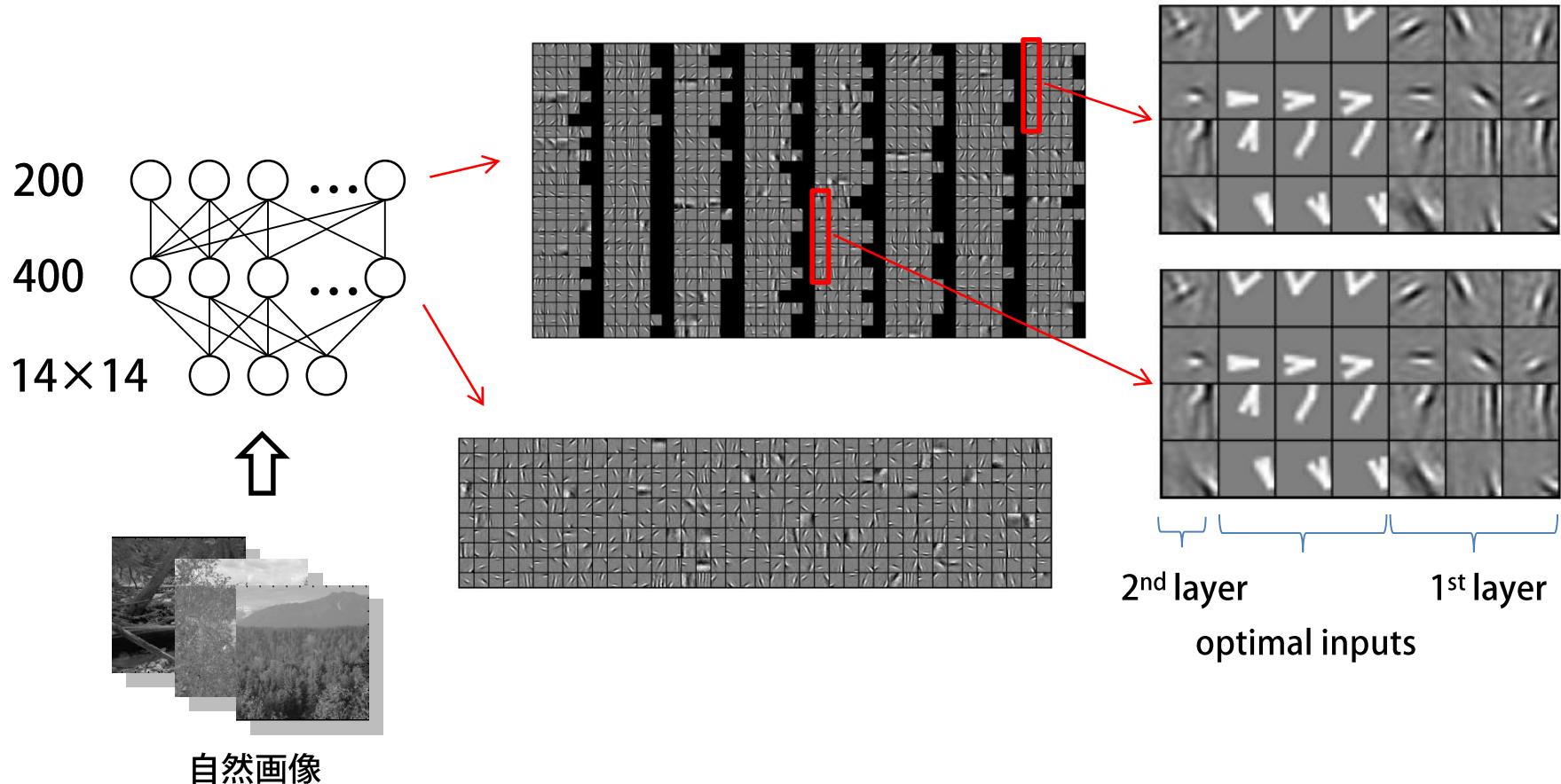
事例

- 教師なし学習
- たたみこみニューラルネット

V2 のモデル

Lee et al., Sparse deep belief net model for visual area V2, NIPS08

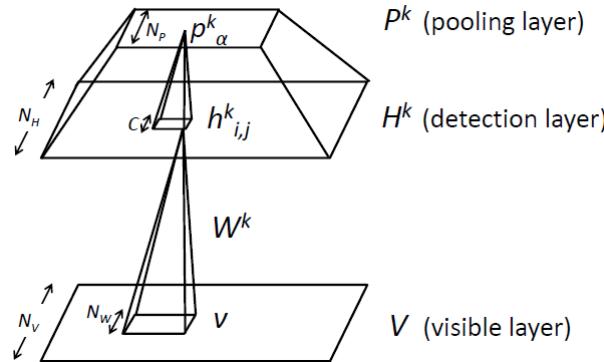
- スパースRBMを使ったV2のモデル
 - 現実のV2ニューロンの応答[Minami-Komatsu04]を再現
 - V1より複雑な形（アングルやジャンクション）に反応



Convolutional DBN

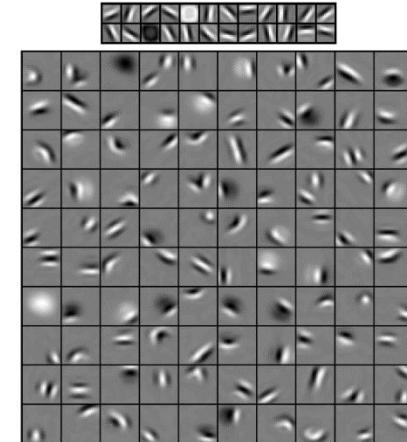
Lee et al., Convolutional Deep Belief Networks for Scalable Unsupervised Learning of Hierarchical Representations, ICML09

- たたみこみとプーリングを取り入れたDBN

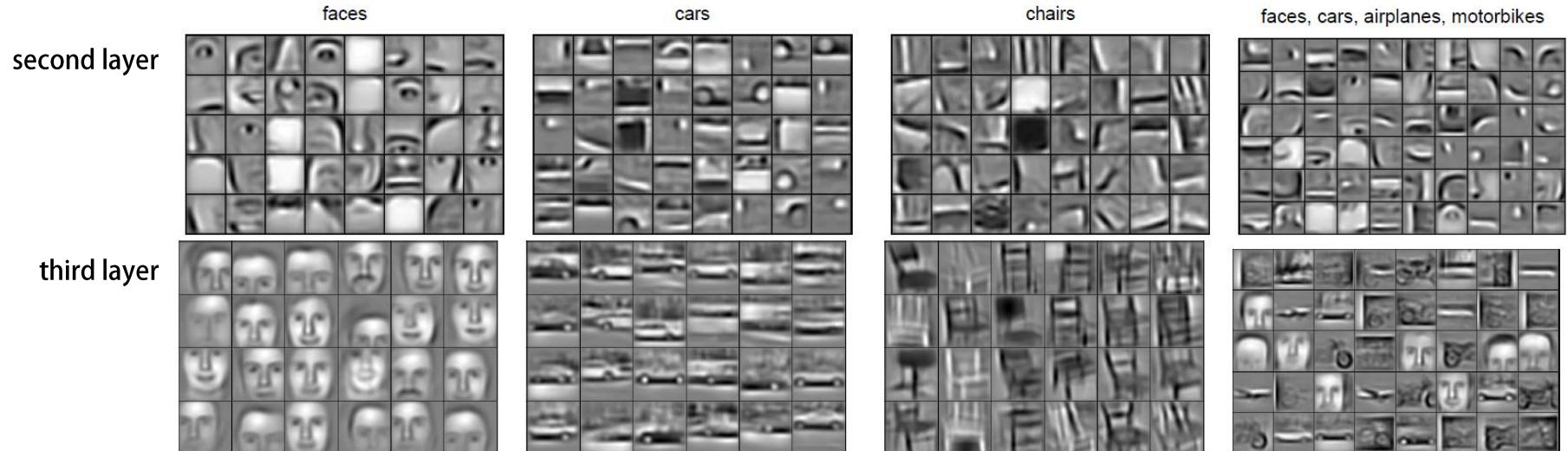


Learned bases for natural scenes:

first layer
bases
second layer
bases



Unsupervised learning of object parts (three layers):



画像特徴の無教師学習

Le et al., Building High-level Features Using Large Scale Unsupervised Learning , ICML2012

- 12層NNを使った無教師学習
 - パラメータ数10億個！
 - 16コアPC1000台のPCクラスタ×3日間
 - YouTubeの画像1000万枚
- 「おばあさん細胞」の生成を確認

The New York Times
Business Day
Technology

HOME PAGE | TODAY'S PAPER | VIDEO | MOST POPULAR | U.S. Edition ▾

Panasonic My Let's 個楽部

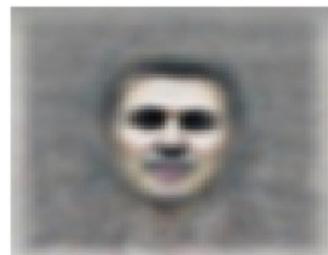
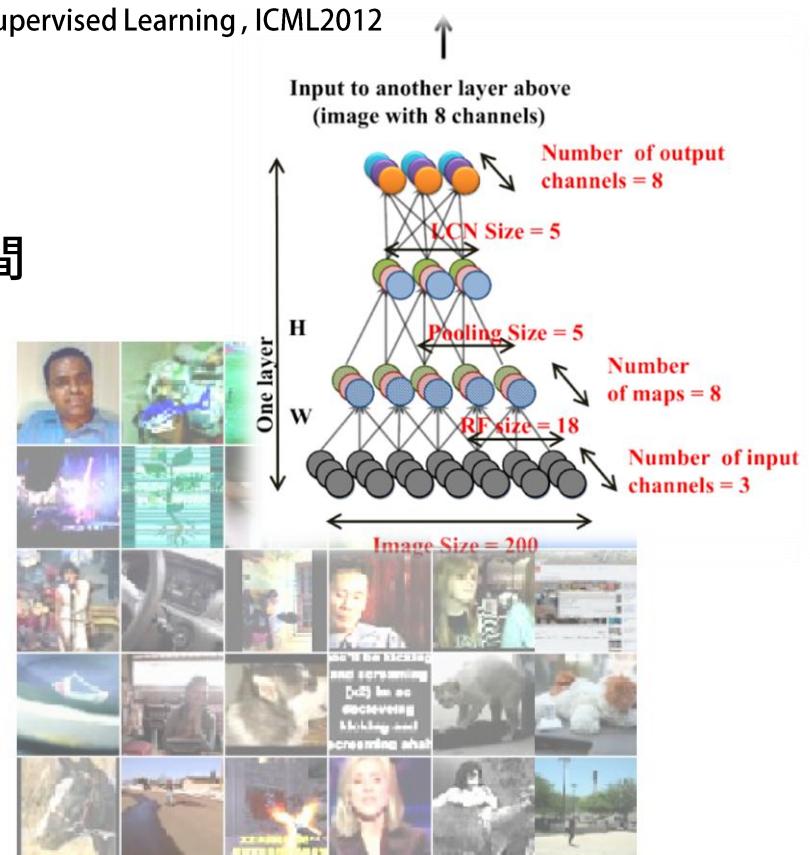
How Many Computers to Identify a Cat? 16,000

An image of a cat that a neural network taught itself to recognize.

By JOHN MARKOFF
Published: June 25, 2012

MOUNTAIN VIEW, Calif. — Inside Google's secretive X laboratory,

FACEBOOK



顔

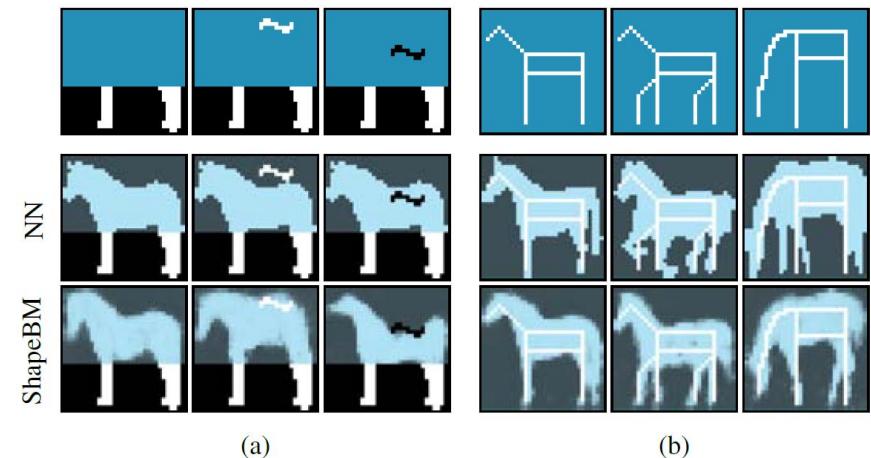
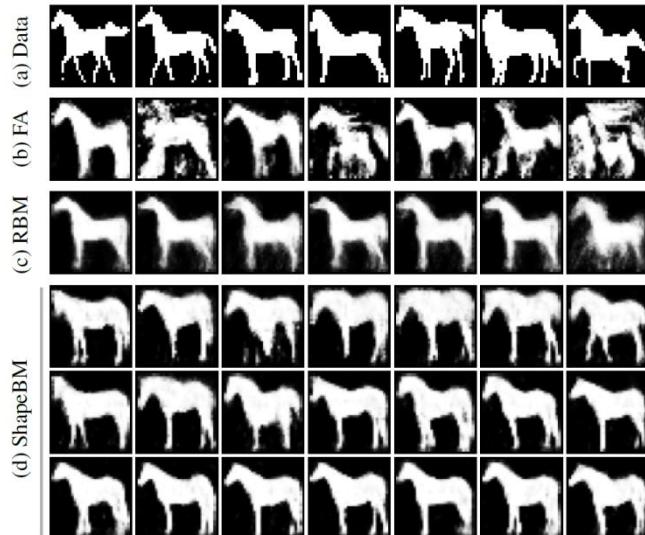
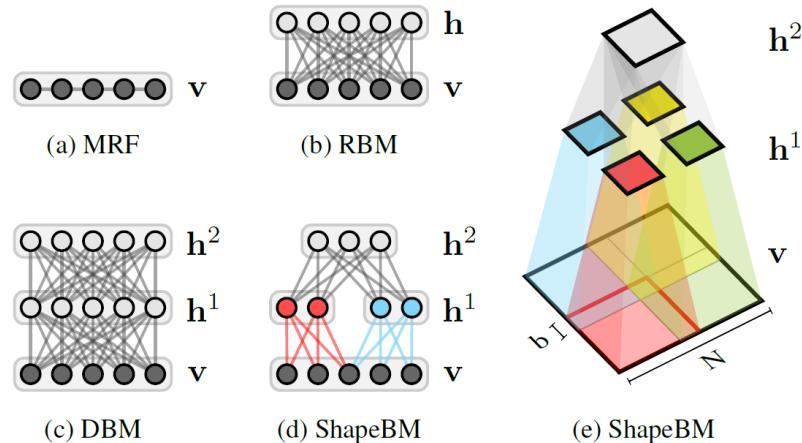
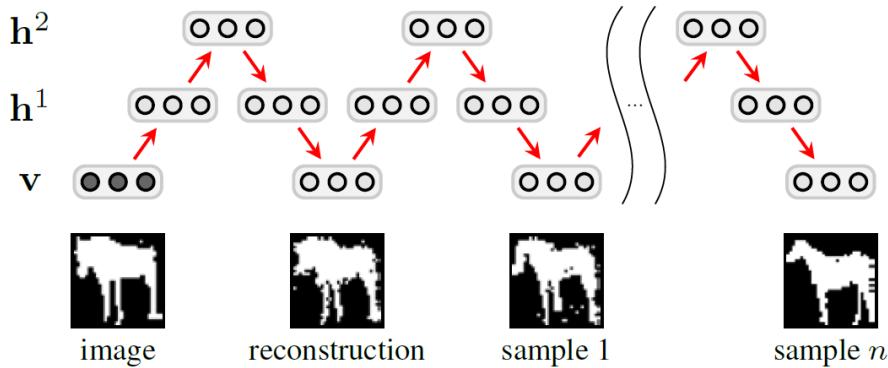
人の体

Shape Boltzmann machine

Ali Eslami et al., The Shape Boltzmann Machine: a Strong Model of Object Shape, CVPR12

- 形を学習するDBM : realismとgeneralizationの達成

[\[video\]](#)

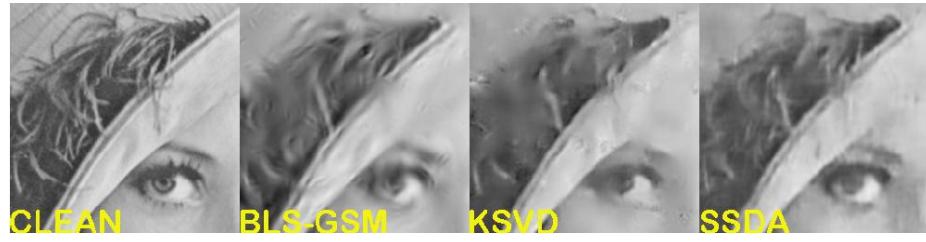
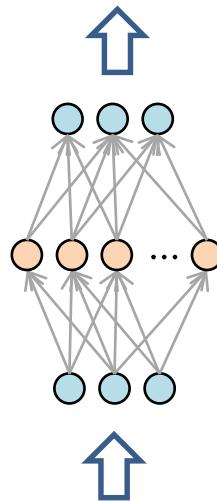


Denoising and inpainting

Xie et al., Image Denoising and Inpainting with Deep Neural Networks, NIPS, 2012

- デノイジング・オートエンコーダを用いたノイズ除去；Denoising 性能は、PSNR評価では従来手法と同等だが、見た目で勝る
 - Inpaintingの問題をデノイジングとみなす → blind inpainting が可能に

Denoised patch



ノイズ除去結果

多層の
スパースデノイジングオートエンコーダ
を応用



Blind inpainting 結果 (KSVDはnon-blind)

事例

- 教師なし学習
- たたみこみニューラルネット

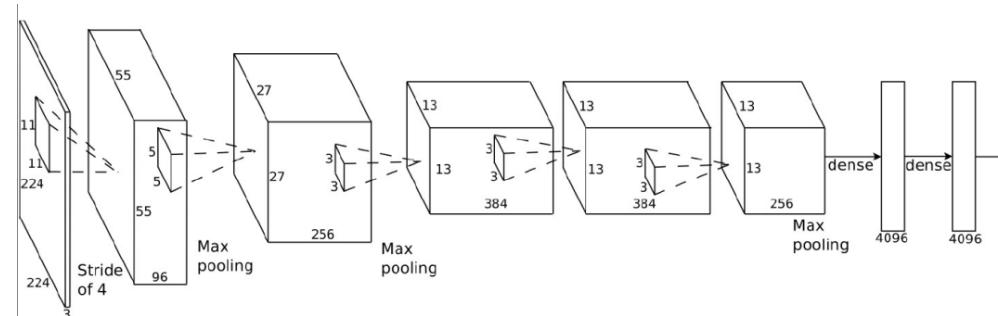
一般物体認識 (Hintonのグループ)

Krizhevsk et al., ImageNet Classification with Deep Convolutional Neural Networks, NIPS2012

- IMAGENET Large Scale Visual Recognition Challenge 2012
 - 1000カテゴリ・カテゴリあたり約1000枚の訓練画像
 - CNN ; rectified linear unit ; drop-out



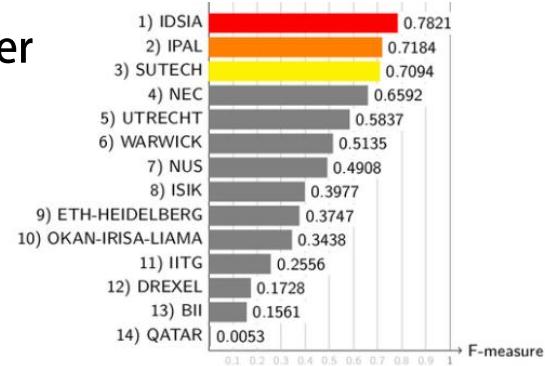
	Team name	Error (5 guesses)
1	SuperVision	0.15315
2	ISI	0.26172
3	OXFORD_VGG	0.26979
4	XRCE/INRIA	0.27058
5	University of Amsterdam	0.29576
6	LEAR-XRCE	0.34464



文字・画像認識 (Schmidhuberのグループ@IDSIA)

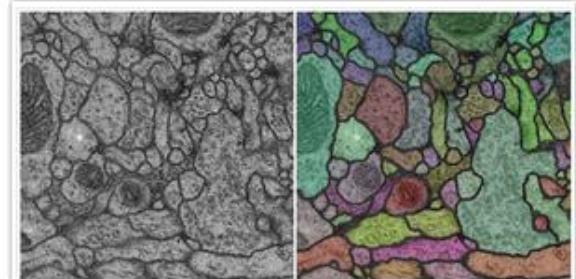
- コンテスト1位

- IJCNN 2011 Traffic Sign Recognition Competition; 1st (0.56%), 2nd (1.16%, Humans), 3rd (1.69%), 4th (3.86%)
- ICPR 2012 Contest on “Mitosis Detection in Breast Cancer Histological Images”
- ISBI 2012 challenge on segmentation of neuronal structures
- ICDAR 2011 Offline Chinese Handwriting Competition
- ICDAR2009
 - Arabic Connected Handwriting Competition
 - French Connected Handwriting Competition



- 認識率最高位 (2012年11月時点)

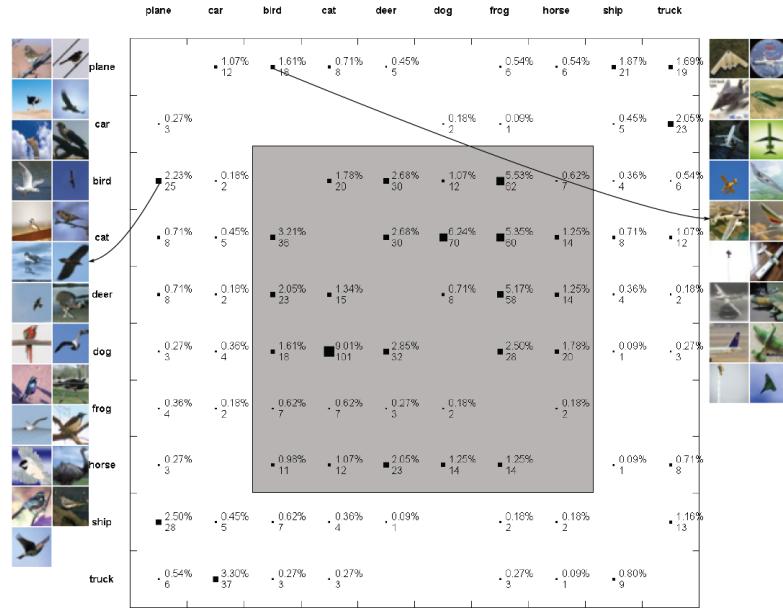
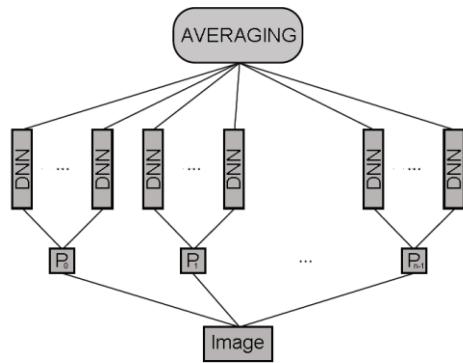
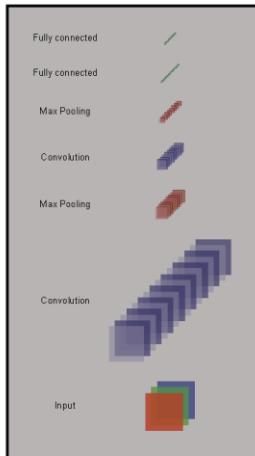
- NORB object recognition benchmark
- CIFAR image classification benchmark
- MNIST handwritten digits benchmark;
“human-competitive result”



Mutli-Column Deep NN

Ciresan, Meier, Schmidhuber, Multi-column Deep Neural Networks for Image Classification, CVPR12

- 並列に訓練したCNNを平均して推定



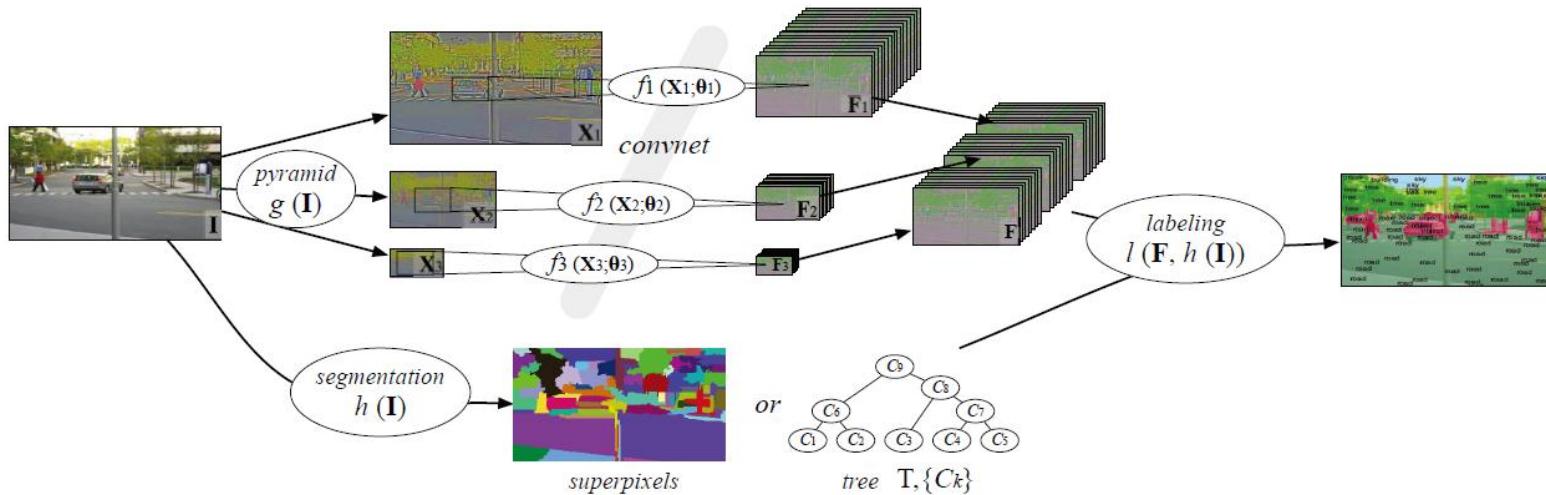
Dataset	Best result of others [%]	MCDNN [%]	Relative improv. [%]
MNIST	0.39	0.23	41
NIST SD 19	see Table 4	see Table 4	30-80
HWDB1.0 on.	7.61	5.61	26
HWDB1.0 off.	10.01	6.5	35
CIFAR10	18.50	11.21	39
traffic signs	1.69	0.54	72
NORB	5.00	2.70	46



Scene Labeling

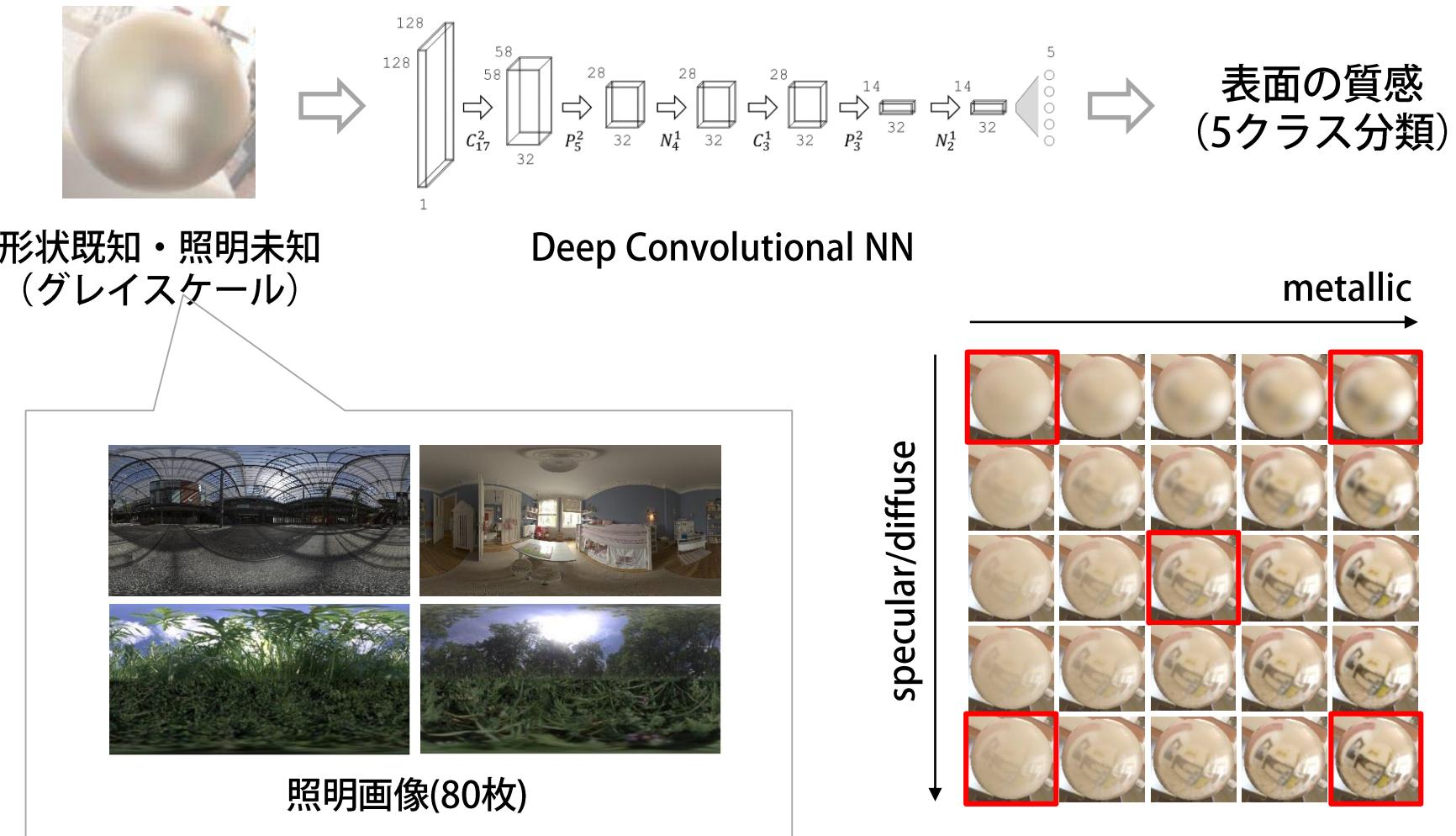
Farabet et al., Learning Hierarchical Features for Scene Labeling, IEEE PAMI, 2012

- 画素ごとにラベルを出力するCNNを教師あり学習
 - 上位層の出力を特徴量に空間方向の一致性を確保



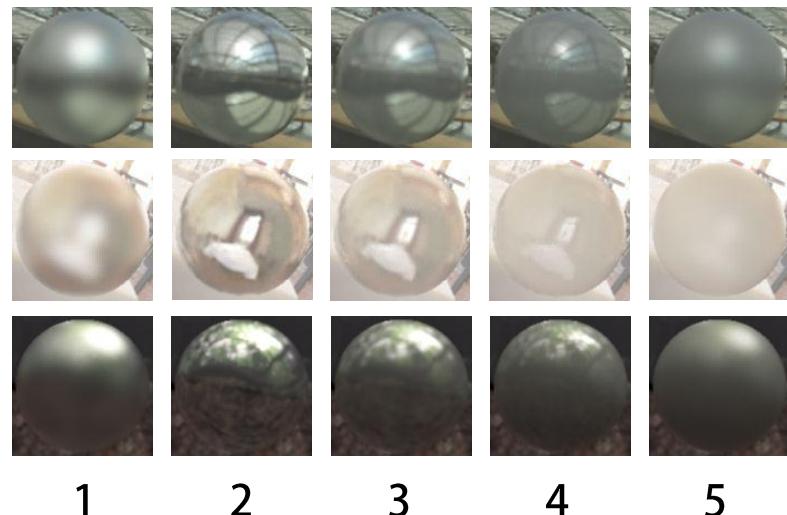
質感の認識：タスク

- ディープニューラルネットで光沢度・透明度を認識

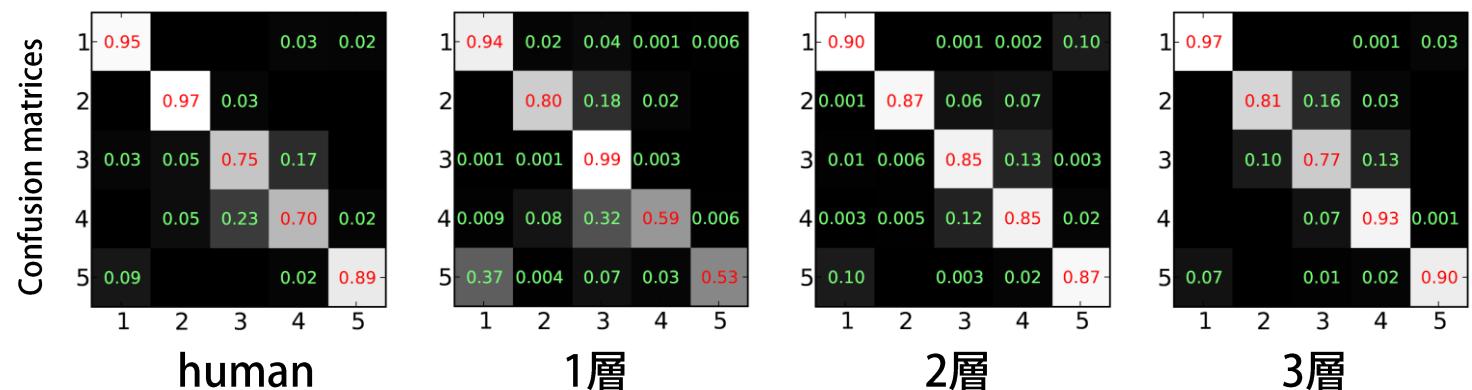
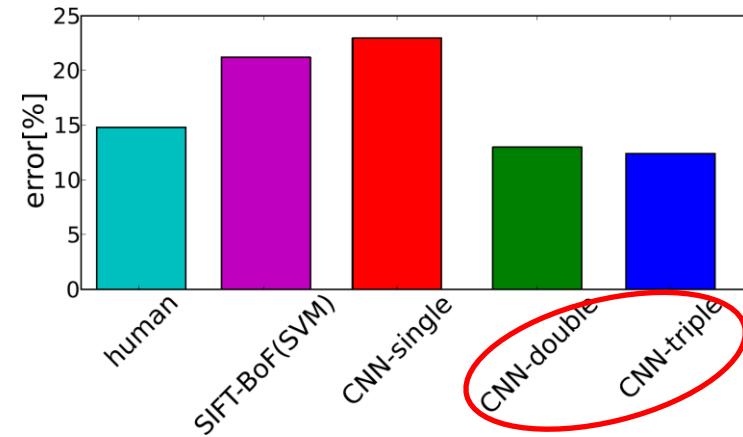


質感の認識：結果

- 多層のCNNが高い性能
 - 人を上回る

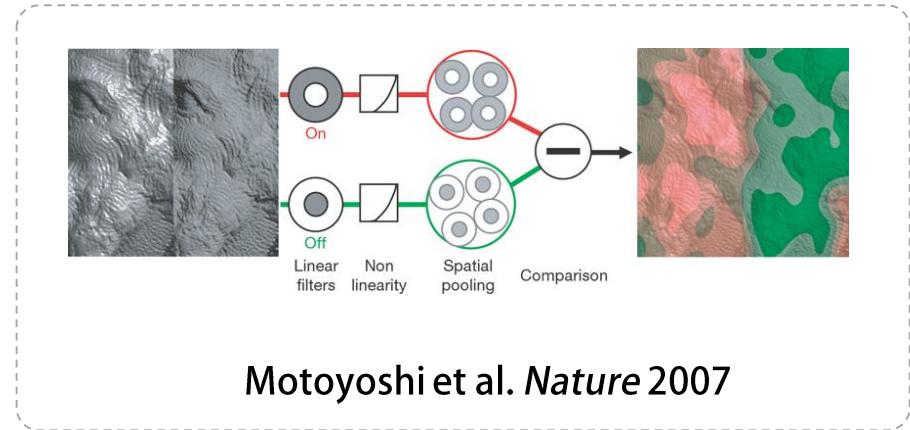
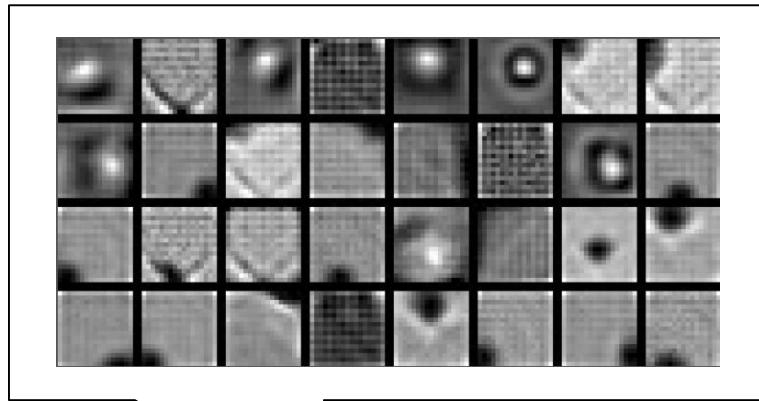


誤認識率 (2000テスト中)

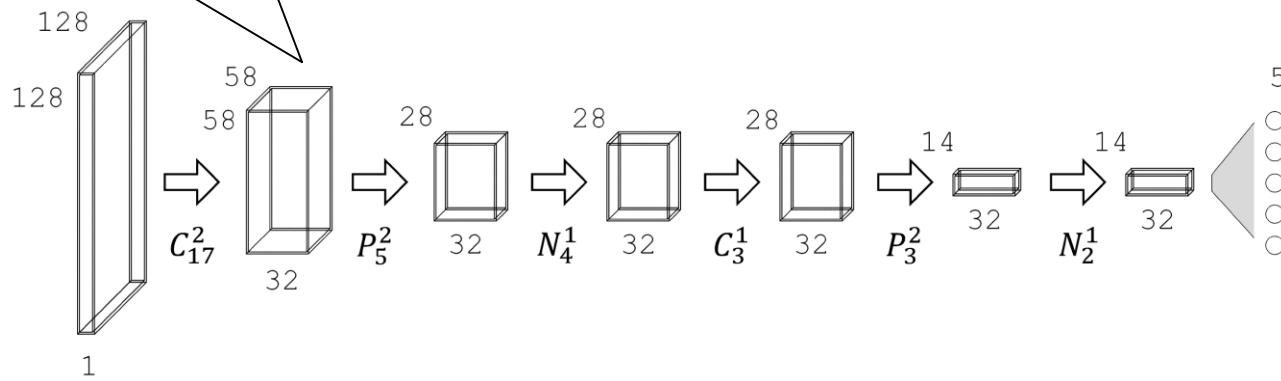


質感の認識：学習された特徴

- 回転対称なフィルタが学習された (on-center & off-center)

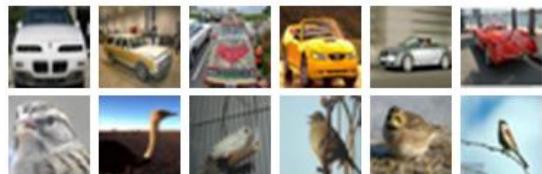


Motoyoshi et al. *Nature* 2007



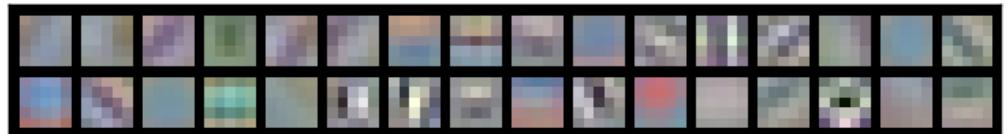
色々なタスクで学習された特徴

一般物体認識 (CIFAR10)

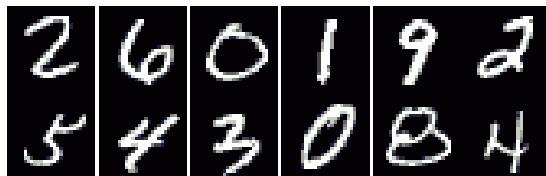


学習されたフィルタ

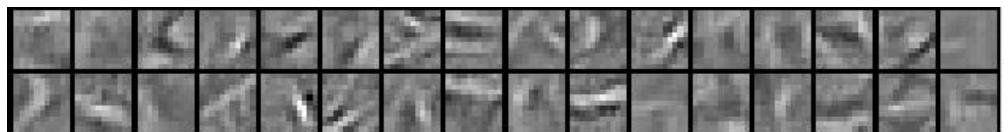
[Krizhevsk+11] 正答率 : 82%



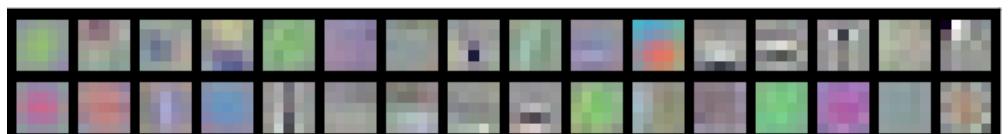
文字認識 (MNIST)



正答率 : 99% (~ヒト)



材質認識 (FMD)



まとめ (1/2)

- DNNは大きな成果を上げている
 - 既存手法の性能比で大きなマージン
 - まだ性能向上の余地
- DNNのプレトレーニング
 - 特に音声認識では全結合NNとプレトレーニングが成果
 - 第3の方法
- 画像認識で特に有効な、たたみこみニューラルネット
 - 80年代後半からほとんど変わらず
- DNNが最近になって成功した理由
 - 「やればできる」とわかったこと・「ビッグデータ」・計算機性能

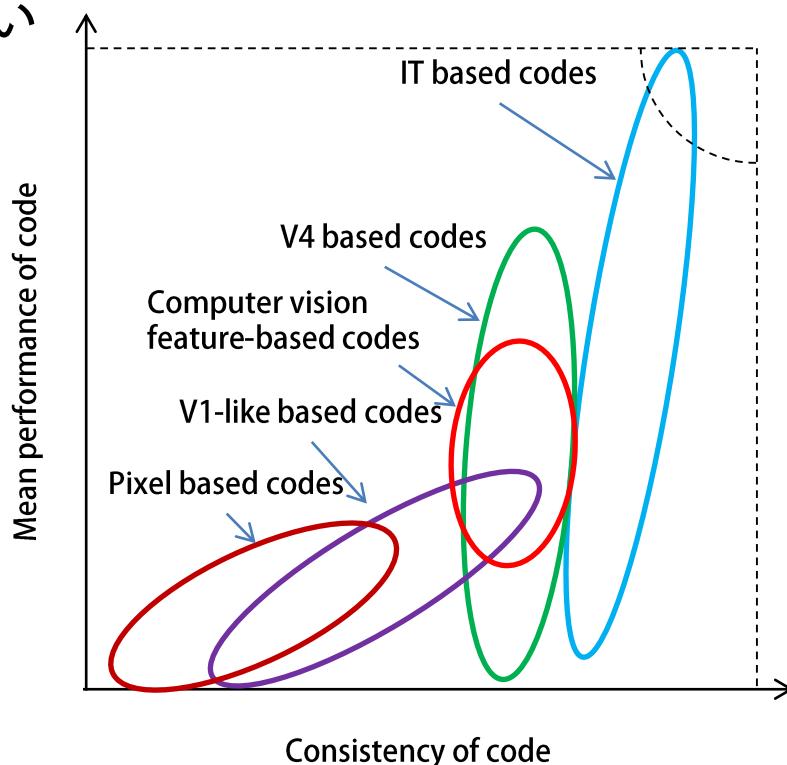
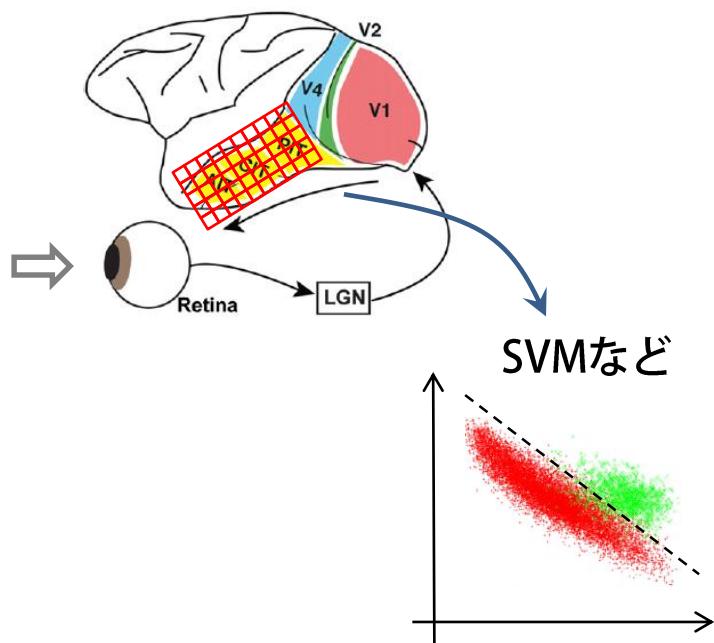


まとめ (2/2)

- 使いづらさは昔と変わらない
 - 性能を引き出すためのノウハウが必要
- 性能向上はエンジニアリング勝負
 - 大規模NNによる大規模データを使った学習
 - 認識性能上げたい → NN多層化 → 過学習のリスク → 学習データ量増やす → 計算性能必要
- 特徴学習・表現学習への期待
 - semi-supervised / transfer / self-taught learning
- 残される疑問
 - なぜ多層だと良いのか?
 - 特徴量の再利用 + 抽象化 [Bengio-Courville-Vincent12]
 - 脳に見られる階層性が実現可 [Mohamed-Hinton-Penn12]

一般物体認識は脳でどのように行われるか？

- “The brain’s bag of features is better than ours.”
- サルに画像を提示；脳のV1,V2,V4,IT野の活動パターンを計測
- IT野の計測パターン由来の“code”をSVMで分類すると、人と同じ認識性能を達成できる
 - V4野までの“code”ではできない



ICCV2011でのJames DiCarloによるキーノート講演スライドから（上の図は筆者の不正確なスケッチ）
<http://www.iccv2011.org/authors/oral-presentations/thursday>

文献

- 拙稿
 - ディープラーニング, 情報処理学会CVIM研究会, 研究報告2013年1月
 - 改訂版をCVIMチュートリアルシリーズとして出版予定
- 人工知能学会論文誌, 連載解説「深層学習」
 - 第1回: ディープボルツマンマシン
 - 第2回: 多層ニューラルネットワークによる深層表現の学習
 - 以降: 実装, 画像, 音声, 自然言語処理と続く予定
- <http://deeplearning.net/>
 - Bengio et al., Representation Learning: A Review and New Perspective, Arxiv, 2012
 - Bengio, Learning Deep Architectures for AI (Foundations & Trends in Machine Learning), 2009