

全脳アーキテクチャ勉強会（第2回）

# Deep Learning 技術の今

得居 誠也

2014年1月30日



# 自己紹介

- 得居 誠也 (Seiya Tokui)
- 株式会社Preferred Infrastructure, Jubatus Pj. リサーチャー
- 専門は機械学習 (修士、現職)
  - 系列ラベリング→ハッシュ・近傍探索→深層学習
- 今の興味は深層学習、表現学習、分散学習、映像解析
- @beam2d (Twitter, Github, etc.)



## 2011年: 音声認識における成功

acoustic model & training	RT03S		Hub5'00	voicemails		tele-conf
	FSH	SW	SWB	MS	LDC	
GMM 40-mix, ML, SWB 309h	30.2	40.9	26.5	45.0	33.5	35.2
GMM 40-mix, BMMI, SWB 309h	27.4	37.6	23.6	42.4	30.8	33.9
CD-DNN 7 layers x 2048, SWB 309h, this paper	18.5	27.5	16.1	32.9	22.9	24.4
(rel. change GMM BMMI → CD-DNN)	(-33%)	(-27%)	(-32%)	(-22%)	(-26%)	(-28%)

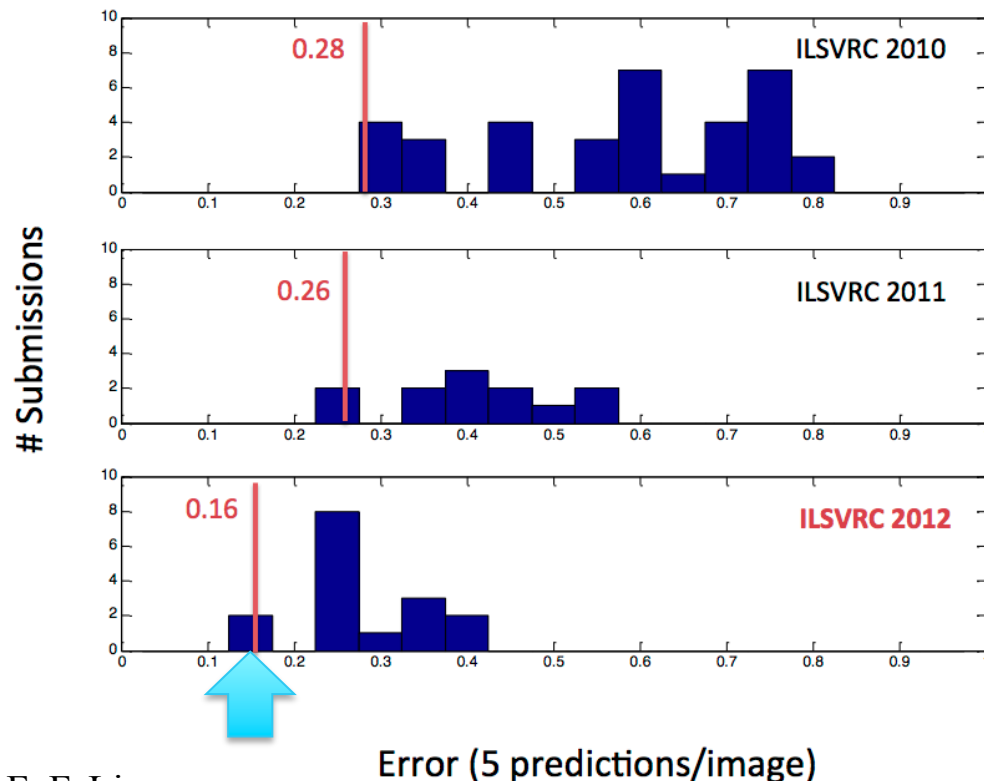
- DNN-HMM を使った手法が、音声認識の word error rate で従来法 (GMM) より 10% 前後も改善
- 携帯端末における音声操作に Deep Learning が利用されるように

F. Seide, G. Li and D. Yu.

[Conversational Speech Transcription Using Context-Dependent Deep Neural Network](#), in INTERSPEECH, pp. 437-440 (2011)

## 2012年: 画像認識における成功

- 一般物体認識のコンテスト ILSVRC2012 において Deep Convolutional Neural Network を用いたチーム Supervision が他者に 10% のエラー差をつけて勝利



J. Deng, A. Berg, S. Satheesh, H. Su, A. Khosla and F.-F. Li

[Large Scale Visual Recognition Challenge 2012](#). ILSVRC2012 Workshop.

## 2013年: 大企業による投資、人材争奪戦

- 3月 : Google が DNNresearch を買収
  - Geoffrey Hinton, Alex Krizhevsky and Ilya Sutskever
- 4月 : Baidu が Institute of Deep Learning を設立
  - 最初の研究者として Kai Yu を迎えた
- 8, 10月 : Yahoo が画像認識のスタートアップ IQ Engines と LookFlow を買収
  - Deep learning group を作るための布石と報じられる
- 12月 : Facebook AI Lab 設立
  - Yann LeCun 所長、他にも Marc'Aurelio Ranzato など
- 2014年1月 : Google が DeepMind を買収
  - G. Hinton の研究室の卒業生、Deep Learning による強化学習技術

# 今日のお話

- Deep Learning（深層学習）の技術を基礎・応用と広めに紹介
  - Zoo of deep learning みたいなノリで……
- 全脳アーキテクチャを考える上で参考になるようなモデル・学習面での現状をお伝えします
  - モデルの紹介が多めです
- 盛りだくさんなのでいろいろ流していきます
  - 詳細は後日スライドおよび参考文献をご覧ください

# 目次

- Deep Learning の成功
- Deep Learning の基礎
- Deep Learning と認識
- Deep Learning と構造
- Deep Learning の今後

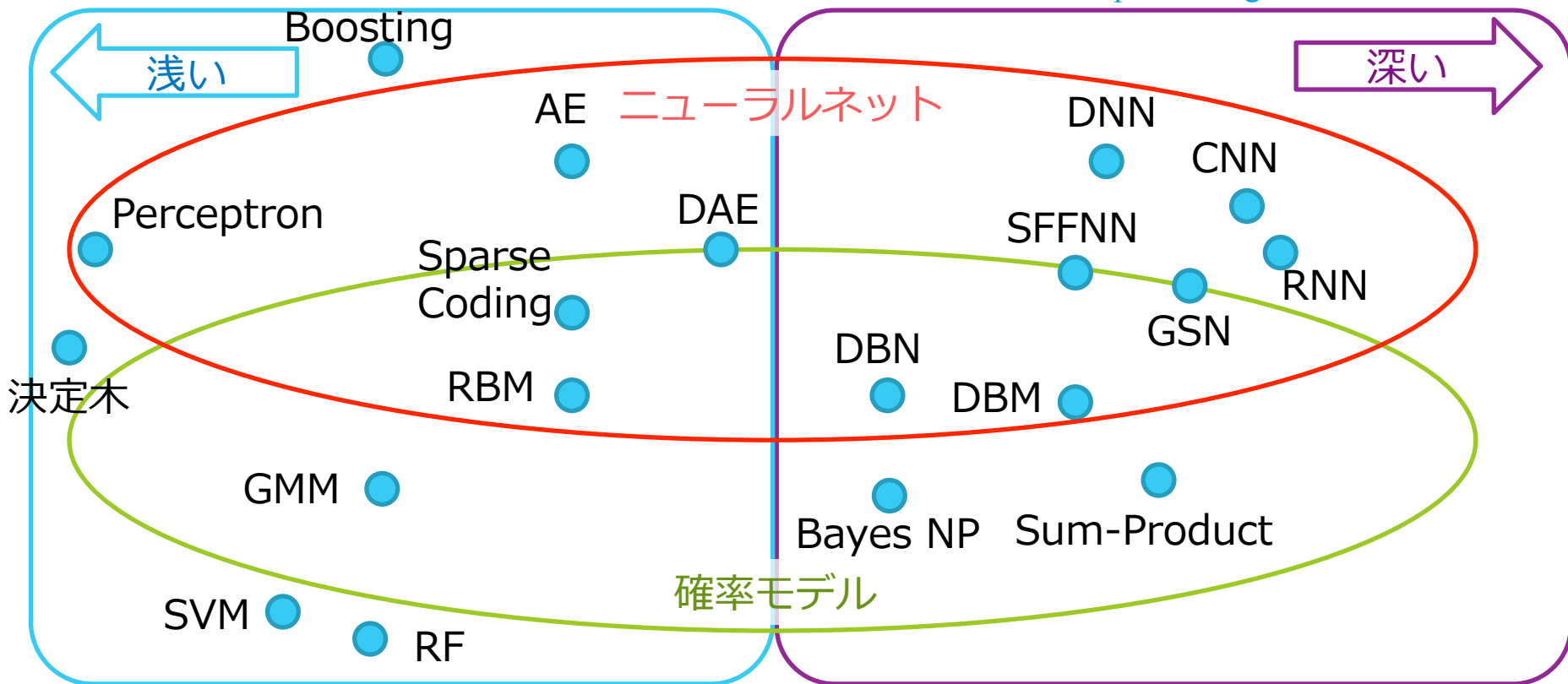
基礎となるモデル・学習手法

# Deep Learning の基礎

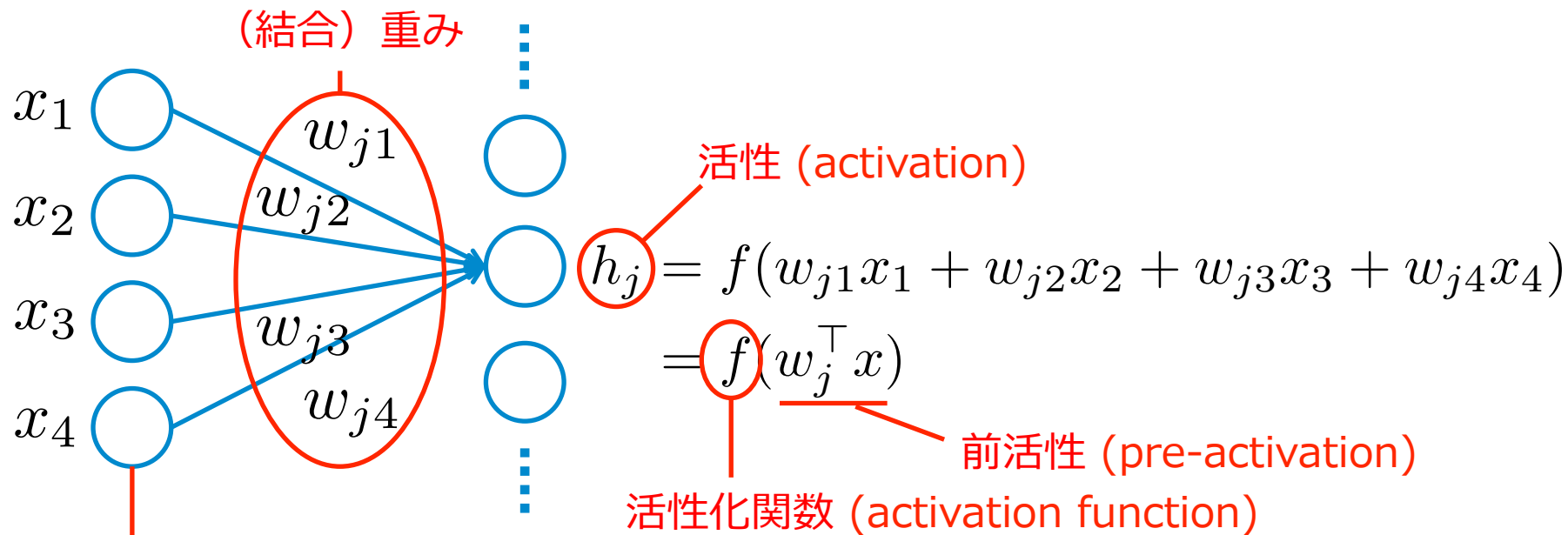


# Deep Learning の地図 (アレンジ)

Cf.) Y. LeCun and M. A. Ranzato.  
[Deep Learning Tutorial](#). ICML 2013.



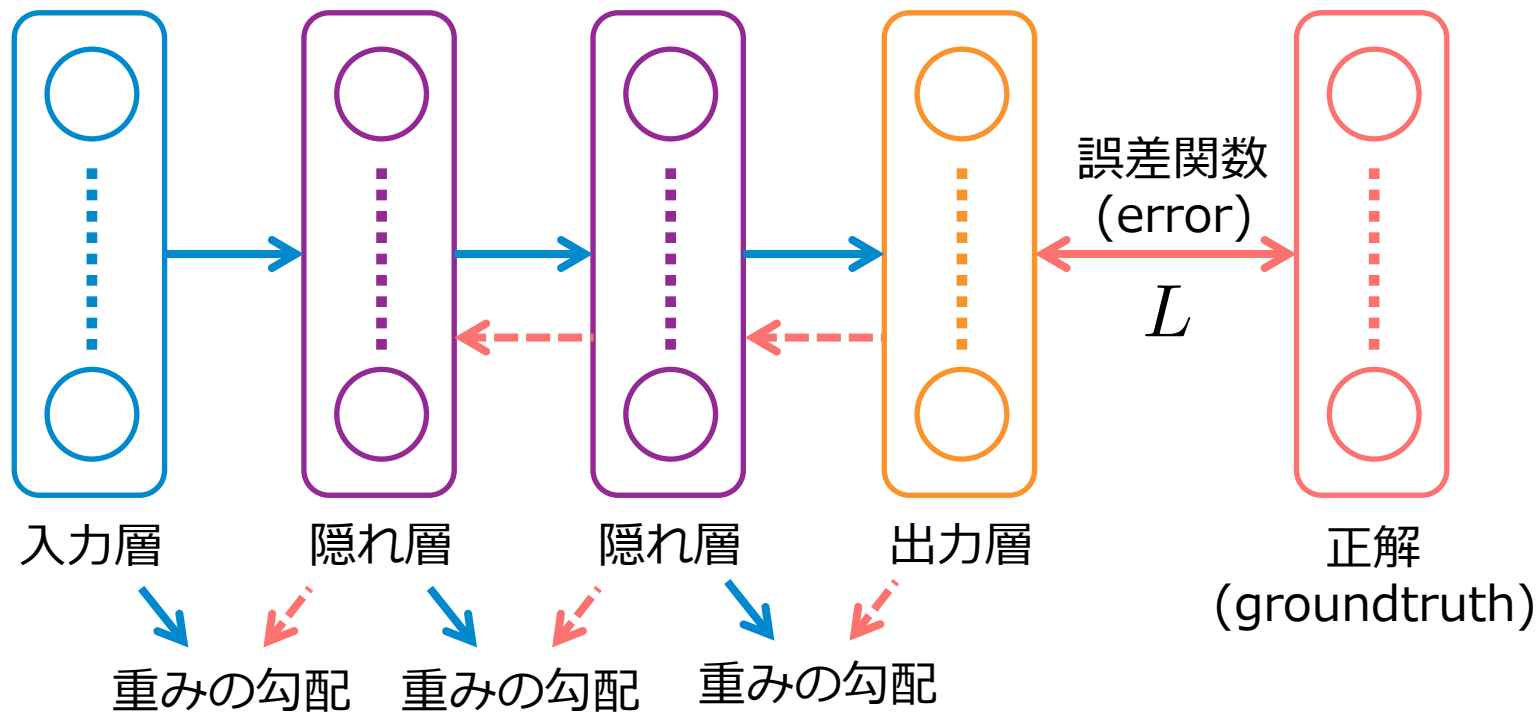
# Feed-Forward Neural Network



$h_j$  の計算全体は行列を使って  $h = f(Wx)$  と書ける  
 (層を飛び越える結合は考えない)  
 バイアス項もよく使う:  $h = f(Wx + b)$

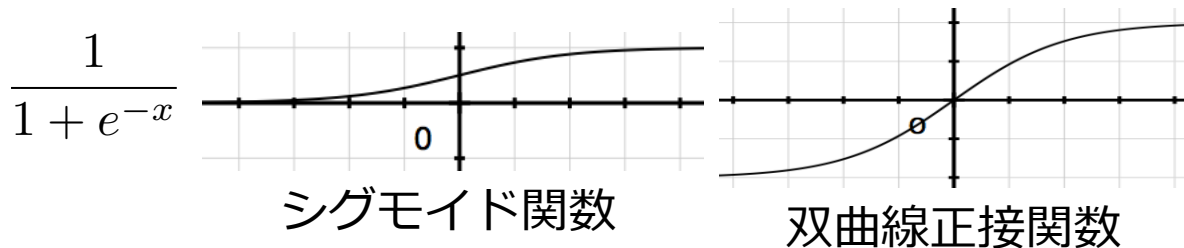
# 誤差逆伝播法 backpropagation

→ 順伝播 fprop  
← 逆伝播 bprop

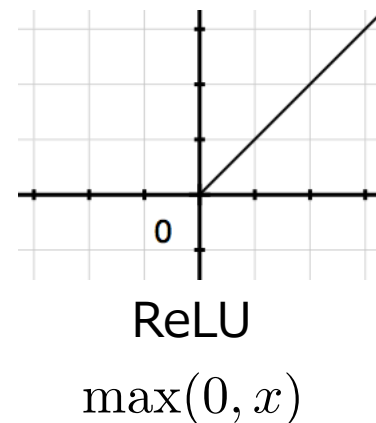


# 活性化関数 (activation function)

- 従来はシグモイド型の関数が用いられてきた



- 最近よく使われるのは Rectified Linear Unit (ReLU)
  - サチらない、つまり勾配が消えにくいので学習しやすい
- 恒等関数は Linear Unit と呼ばれる
- 複数の Linear Unit の max を取る: **maxout unit\***



\* I. Goodfellow, D. W.-Farley, M. Milza, A. Courville and Y. Bengio. [Maxout Networks](#). ICML 2013.

# Neural Network の学習手法

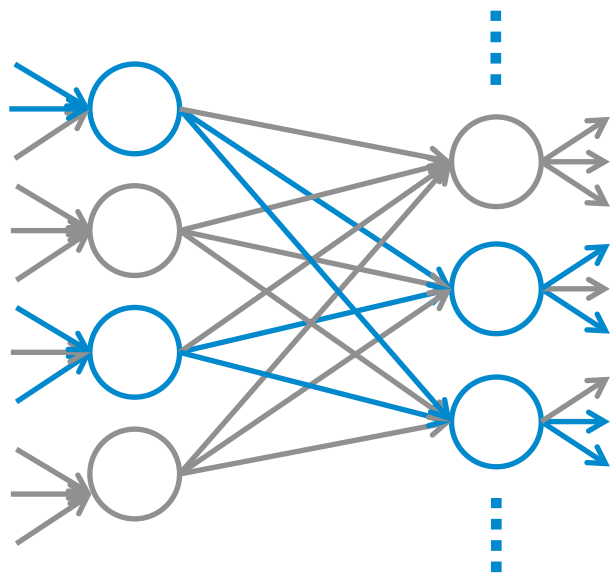
- 教師データを B 個ずつ見る **Mini-Batch SGD** がよく使われる
- 以下の手法と組合せて使われる
$$w \leftarrow w - \gamma \frac{1}{B} \sum_{i=1}^B \frac{\partial L(x_{B_i})}{\partial w}$$
  - Momentum, Nesterov's Accelerated Gradient\*
  - L2 正則化 (weight decay)、L1 正則化
  - ステップ幅の自動調整 (AdaGrad\*\*, vSGD\*\*\*)
- 最適化が難しいケースではニュートン法ベースの手法も (L-BFGS, Hessian-Free 法など)

\* I. Sutskever, J. Martens, G. Dahl and G. Hinton. [On the importance of initialization and momentum in deep learning](#). ICML 2013.

\*\* J. Duchi, E. Hazan and Y. Singer. [Adaptive Subgradient Methods for Online Learning and Stochastic Optimization](#). JMLR 12 (2011) 2121-2159.

\*\*\* T. Schaul, S. Zhang and Y. LeCun. [No More Pesky Learning Rates](#). ICML 2013.

# Dropout



- SGD 学習時、ランダムに選んだユニットの活性を 0 にする
  - 経験上、入力ユニットは 20%、隠れユニットは 50% の dropout 率だと性能が良い
- 強い正則化の効果がある
  - アンサンブル学習
  - フィッシャー情報行列で歪ませた L2 正則化\*
- 区分線形ユニット (ReLU, maxout) で特に効果的
- 亜種も出てきている (DropConnect, Adaptive Dropout, etc.)

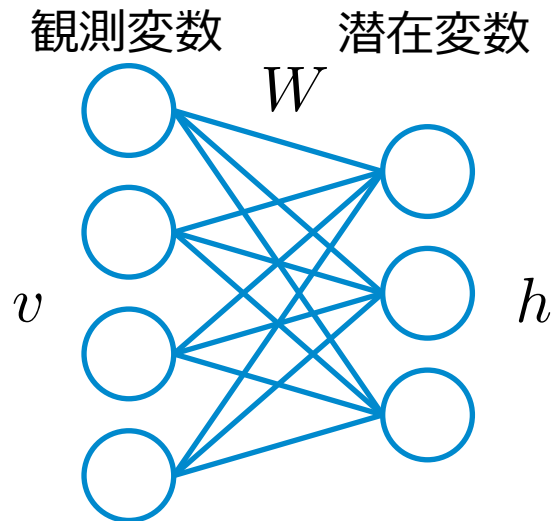
G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever and R. R. Salakhutdinov.

[Improving neural networks by preventing co-adaptation of feature detectors](#). ArXiv 1207.0580.

\* S. Wager, S. Wang and P. Liang.

[Dropout Training as Adaptive Regularization](#). NIPS 2013.

# Restricted Boltzmann Machine



- 無向二部グラフのグラフィカルモデル
- 左下のようなエネルギー関数を持つボルツマン分布
  - 2式はそれぞれ  $v$  が二値および連続値の場合のエネルギー関数 ( $h$  は共に二値変数)
- 対数尤度勾配は次式で書ける

$$E(v, h) = -a^\top v - b^\top h - h^\top W v$$

$$E(v, h) = -\frac{(v - a)^2}{2\sigma^2} - b^\top h - \frac{1}{\sigma} h^\top W v$$

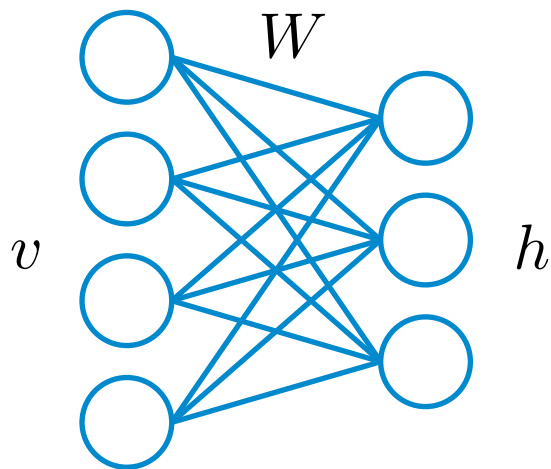
バイアスパラメータ

$$\frac{\partial \log p(v)}{\partial w_{ij}} = \underbrace{\langle v_i h_j \rangle_{\text{data}}}_{\text{データに対する期待値}} - \underbrace{\langle v_i h_j \rangle_{\text{model}}}_{\text{RBM が表す分布に対する期待値}}$$

データに対する  
期待値

RBM が表す  
分布に対する  
期待値

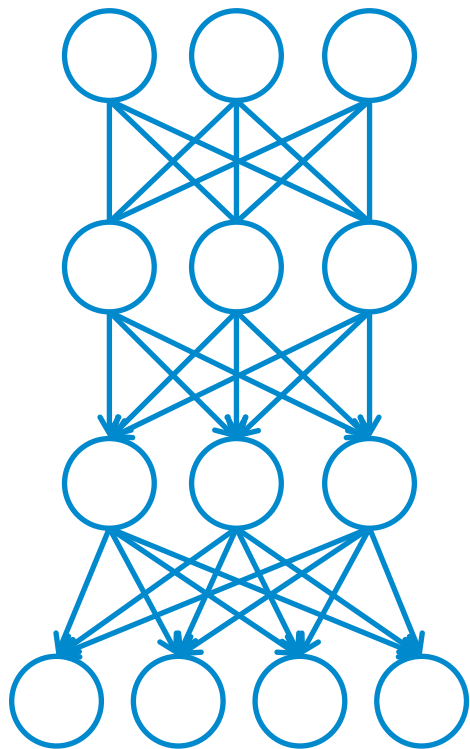
# Contrastive Divergence (CD-k)



- Model 期待値は計算が難しいので、 $k$  往復だけサンプリングして得た観測変数を使う
$$\Delta w_{ij} = \langle v_i h_j \rangle_{\text{data}} - \langle v_i h_j \rangle_{\text{reconstruction}}$$
- 対数尤度の勾配ではなくなってしまう
  - Contrastive Divergence という値の勾配の近似（厳密にはどんな関数の勾配としても書けない）
- 深層学習の応用上は  $k=1$  で良い性能を発揮
  - 単に CD と言ったら CD-1 を指す



# Deep Belief Network\*



- 最後の1層だけ無向なグラフィカルモデル
- 各層ごとに RBM の重みが良い初期値になる
  - Greedy Layer-wise Pre-training と呼ばれる
  - Deep Learning のブレイクスルー
- 最後に Up-down 法で fine-tuning
- 全層それぞれで Contrastive Divergence を用いる方法も (top-down regularization\*\*)
- 特徴抽出に使うか、DNN の初期値にする

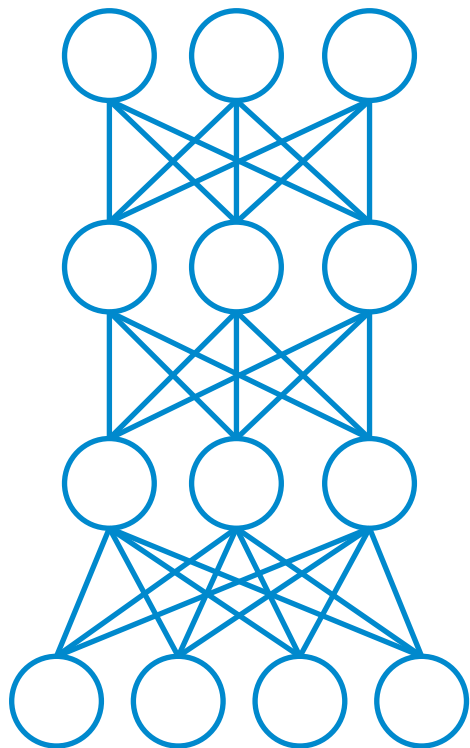
\* G. E. Hinton, S. Osindero and Y.-W. Teh.

[A fast learning algorithm for deep belief nets](#), Neural Computation 2006.

\*\* H. Goh, N. Thome, M. Cord and J.-H. Lim.

[Top-Down Regularization of Deep Belief Networks](#), NIPS 2013.

# Deep Boltzmann Machine



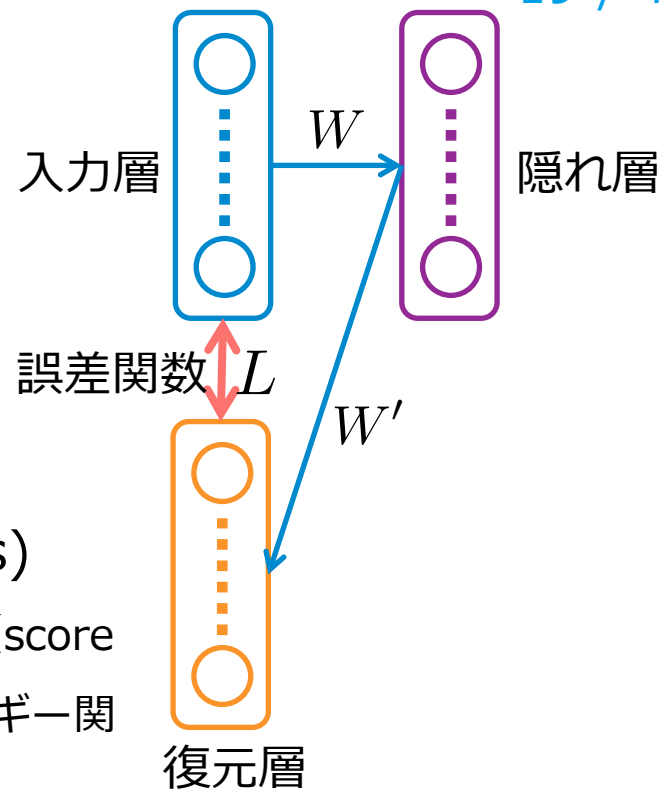
- 層ごとにわかれた無向グラフィカルモデル
- DBN と違い、内側の層は上下両方の層からのフィードバックを受ける
  - モデルを組み立てる際の自由度が上がる
- RBM で事前学習、全体を RBM のように学習

$$\frac{\partial \log p(v)}{\partial w_{ij}} = \langle v_i h_j \rangle_{\text{data}} - \langle v_i h_j \rangle_{\text{model}}$$

- ただし data 期待値も簡単に計算できない（条件付き分布が factorize されない）→ 変分推定
- Model 期待値は Persistent MCMC

# Autoencoder (AE)

- 入力を復元する 2 層の NN
- 恒等関数を学習しないように以下の工夫
  - 入力層より小さな隠れ層 (bottleneck)
  - 正則化 (Contractive AE\*, Sparse AE など)
  - 入力層にノイズを加える (Denoising AE\*\*)
- 制約  $W' = W^T$  をよく置く (tied weights)
  - 二乗誤差 DAE はこの制約のもと、別の目的関数 (score matching の亜種) と一致し、RBM と似たエネルギー関数を持つ\*\*\*

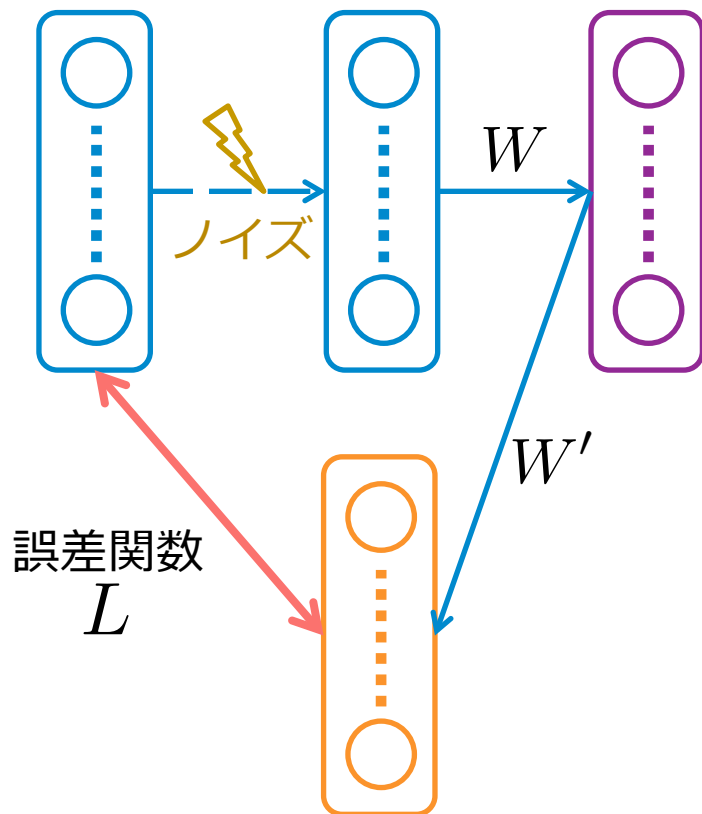


\* S. Rifai, P. Vincent, X. Muller, X. Glorot and Y. Bengio. [Contractive Auto-Encoders: Explicit Invariance During Feature Extraction](#). ICML 2011.

\*\* P. Vincent, H. Larochelle, Y. Bengio and P.-A. Manzagol. [Extracting and Composing Robust Features with Denoising Autoencoders](#). ICML 2008.

\*\*\* P. Vincent. [A Connection Between Score Matching and Denoising Autoencoders](#). TR 1358, Dept. IRO, Universite de Montreal.

# Denoising Autoencoder (DAE)

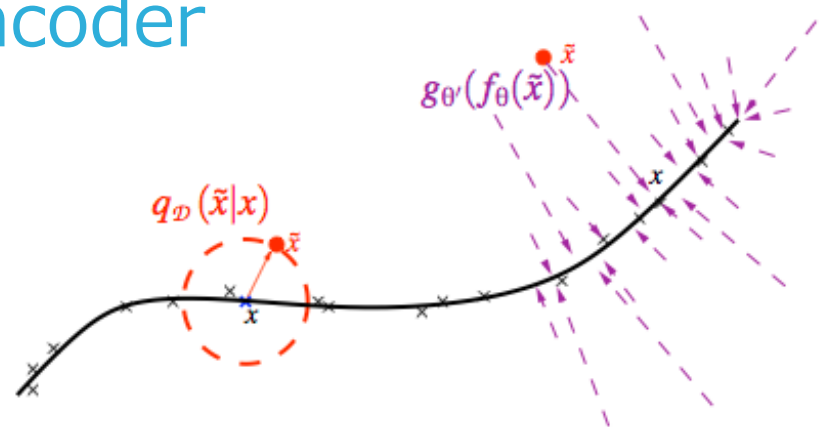


- 入力にノイズを加えてから2層のNNで復元する
- ノイズはガウスノイズや Salt-and-Pepper ノイズ（ランダムなノードを0か1で上書き）を使う
- ノイズにある条件を仮定すれば、最適なDAE解はノイズと復元の操作の繰り返しが表すマルコフ連鎖の定常分布によって入力データの分布を表現する\*

\* Y. Bengio, L. Yao, G. Alain and P. Vincent.  
[Generalized Denoising Auto-Encoders as Generative Models](#). NIPS 2013.

# Stacked Denoising Autoencoder

- DAE を重ねる
- 2 層目以降の DAE を学習する場合、それ以前の層はそのまま適用して、学習する層の入力層に対してノイズを加える
- DAE はデータ分布の多様体を学習している
  - 曲がりくねった多様体を少し平らな空間に展開する (disentanglement)
  - Stacked DAE は多様体を少しずつ平らに展開していくことに対応する
- DAE に限らず他の AE も重ねて deep net を作ることが多い



P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio and P.-A. Manzagol.  
[Stacked Denoising Autoencoders: Learning Useful Representations in a Deep Network with a Local Denoising Criterion](#). JMLR 11 (2010) 3371-3408.

# Deep Learning

- データの深い（多段の）処理全体を通して学習する
- 利点は、複雑なモデルを比較的少ないリソースで学習できること
  - 2層の Neural Network でも任意の関数を、RBM でも任意の分布を表現できるが、それには大量の隠れユニットと大量の学習データが必要になる
  - Deep Learning の場合、同じ関数・分布を表現するのに必要なユニット・データ数が浅いモデルに比べて圧倒的に少なくて済む
  - Deep Neural Net で学習した関数を「教師」として浅い Neural Net を学習させられる\*が、同じ水準のモデルを浅い Neural Net で直接得ることは、同じ学習コストでは（今のところ）できない

\* L. J. Ba and R. Caruana.

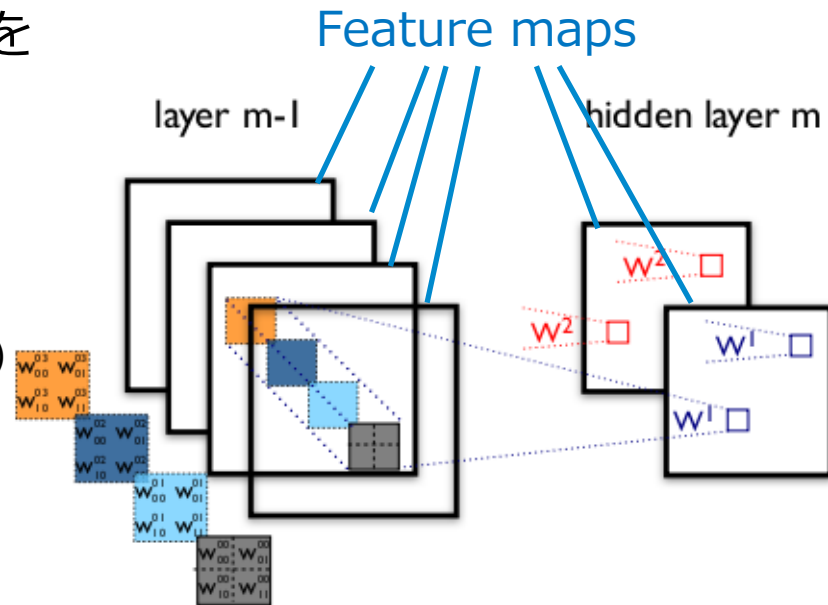
[Do Deep Nets Really Need to be Deep?](#) ArXiv 1312.6184.

画像認識を中心に

# Deep Learning と認識

# Convolutional Neural Network (CNN)

- 画像の縦横方向に、同じ重みの窓をスライドさせながら適用
  - 物体認識は位置不変という事前知識
  - パラメータの大幅な節約
  - スパースなネットワーク（局所受容野）
- 強い正則化の効果
- GPU 計算との親和性
- FFT を使って高速化する話もある
- Simple cell との対応





## Pooling (subsampling)

- Feature map 毎に、矩形上の活性を集約する処理
- L2-pooling, max-pooling, average-pooling がよく使われる


$$\left( \frac{1}{|\text{rectangle}|} \sum_{(i,j) \in \text{rectangle}} x_{ij}^2 \right)^{\frac{1}{2}} \quad \max_{(i,j) \in \text{rectangle}} x_{ij} \quad \frac{1}{|\text{rectangle}|} \sum_{(i,j) \in \text{rectangle}} x_{ij}$$

- L2-pooling や average-pooling ではガウス平均を使うこともある
- 平行移動不変性を獲得するのに役立つ
- CNN は基本的に convolution / activation / pooling の繰り返し
- Complex cell との対応

# Local Contrast Normalization

- 局所的（空間方向および同一座標での複数 feature maps 間）に活性を正規化する

- Subtractive  
$$v_{ijk} = x_{ijk} - \sum_{ipq} w_{pq} x_{i,j+p,k+q}$$

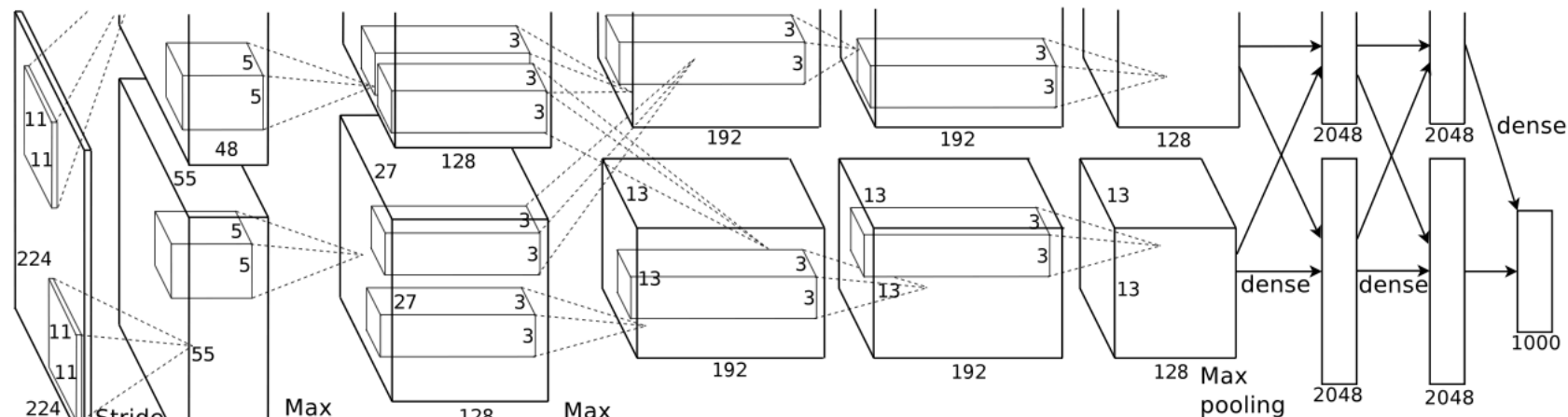
- Divisive  
$$v_{ijk} / \max(c, \sum_{ipq} w_{pq} v_{i,j+p,k+q}^2)$$

- 細かい定義は使用例によってまちまち
- 使い方もまちまち

K. Jarrett, K. Kavukcuoglu, M. A. Ranzato and Y. LeCun.  
[What is the Best Multi-Stage Architecture for Object Recognition?](#) ICCV 2009.

- 正規化がどれくらい精度に影響するのは不明
  - 正規化なくても精度出るという報告もある
- 一次視覚野のニューロンの性質を参考になっている

# Supervision

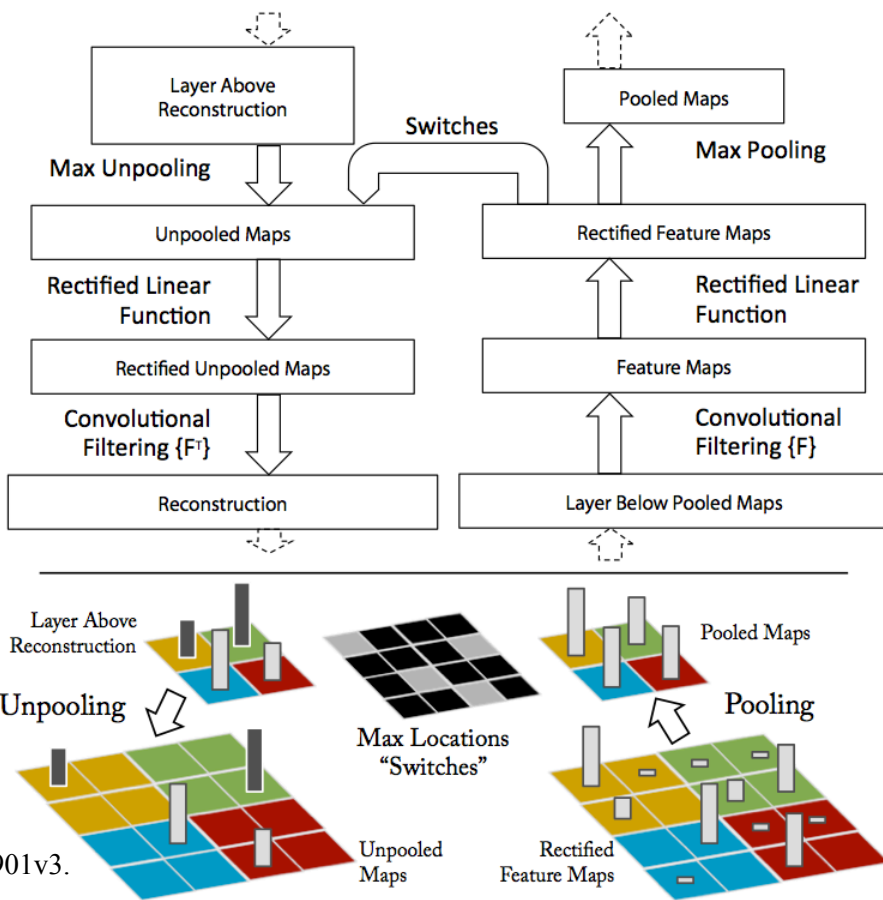


A. Krizhevsky, I. Sutskever and G. E. Hinton.  
[ImageNet Classification with Deep Convolutional Neural Networks](#). NIPS 2012.

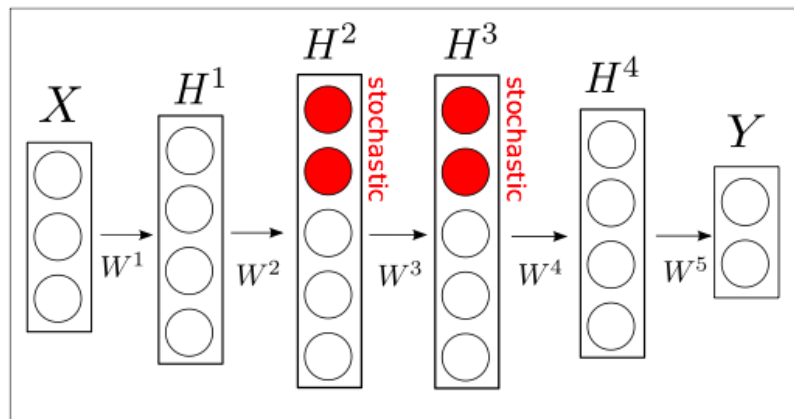
- 図にはないが pooling の後ろに Local Response Normalization
- 2 台の GPU で実装（上下のパイプラインがそれぞれ対応、異なる feature maps を持っている）
- 2013年はこれを使った（拡張する）研究が多かった

# ユニットの可視化 Deconvolutional NN

- Max-pooling が不可逆なので  
ユニットの可視化が難しい
  - 実際に画像を入れて、pooling で  
選択されたピクセルを覚えておく
- ILSVRC2013 の分類タスク優勝  
者 (clarifai) の手法
  - ユニット可視化の手法
  - チューニングが大事



# Stochastic Feedforward NN



- 途中にベルヌーイ分布に従うユニットを入れる
- 学習は EM アルゴリズム
  - E-step は重点サンプリング
  - M-step は backpropagation
- Stochastic neuron のおかげでマルチモーダルな予測ができる
  - 右図は左カラムの画像から 7 通りの表情を予測するタスクの結果

Y. Tang and R. Salakhutdinov.

[Learning Stochastic Feedforward Neural Networks](#). NIPS 2013.

# 再構成型 Topographic ICA

- Sparse Autoencoder の変種
- Pooling 後の活性に対してスパース化ペナルティーを与える
- 非畳み込みの局所受容野と組み合わせると、近くにあるユニットが似た重みを持つようになる
  - 一次視覚野のニューロンと似た性質
  - 平行移動不変性よりも複雑な不変性の獲得

minimize  
 $W_1, W_2$

Decoder Encoder

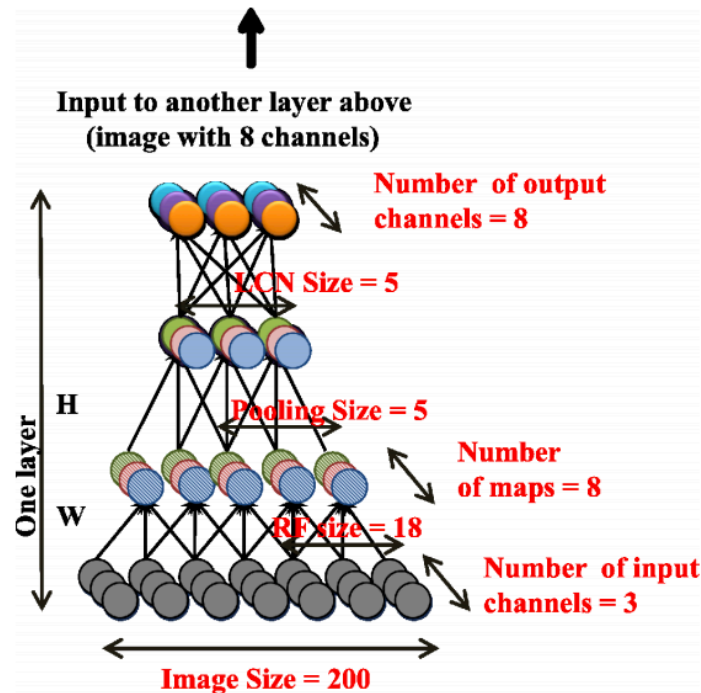
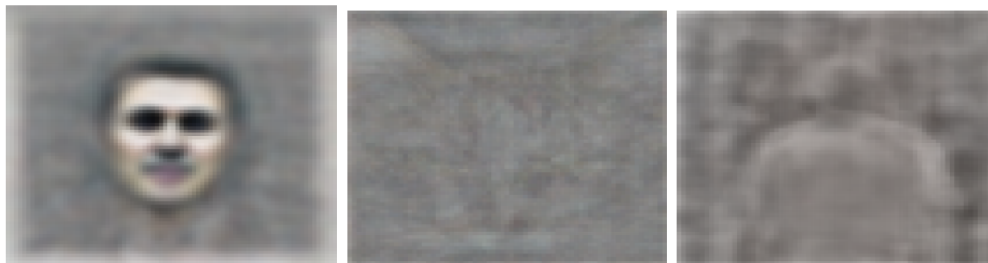
$$\sum_{i=1}^m \left( \|W_2 W_1^T x^{(i)} - x^{(i)}\|_2^2 + \lambda \sum_{j=1}^k \sqrt{\epsilon + H_j(W_1^T x^{(i)})^2} \right).$$

Pooling 処理に相当する重み

Q. V. Lee, M. A. Ranzato, R. Monga, M. Devin, K. Chen, G. S. Corrado, J. Dean and A. Y. Ng.  
[Building High-level Features Using Large Scale Unsupervised Learning](#). ICML 2012.

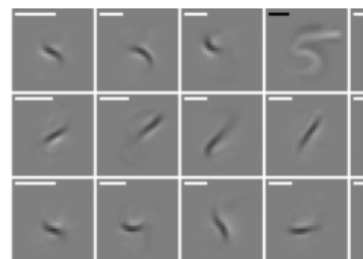
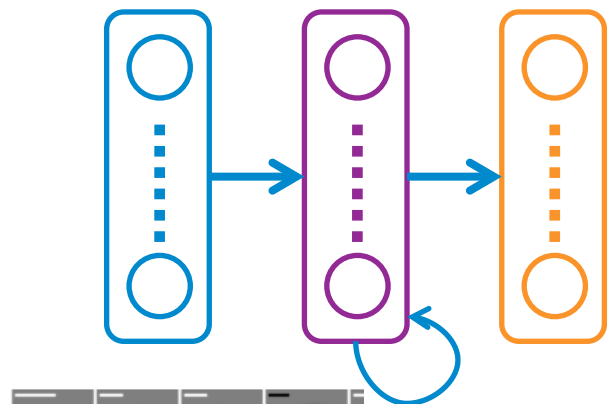
# Google の猫認識

- 3 段の再構成型 TICA
  - Local Contrast Normalization も使っている
  - Convolution ではない (重みを共有しない)
- Youtube の動画 10,000,000 フレームで学習すると猫や顔、人の体などに対応するユニットが得られる



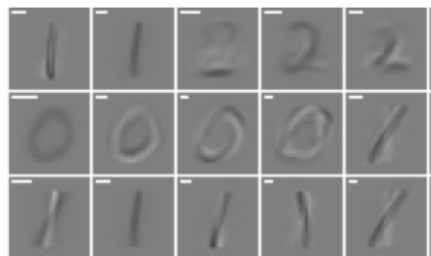
これが3段重なったもの

# DNN としての Recurrent Neural Network



Part units

Categorical  
units



- 隠れ層の活性を入力の一部として次の時間ステップでの隠れ層に入力する
- 隠れ層を  $N$  回ループさせれば  $N$  層の DNN と対応する（重みが共有される）
- 手書き数字に対する適用で、自動的に part unit と categorical unit が得られる（図は Recurrent Sparse Autoencoder）

J. T. Rolfe and Y. LeCun.  
[Discriminative Recurrent Sparse Auto-Encoders](#). ICLR 2013.



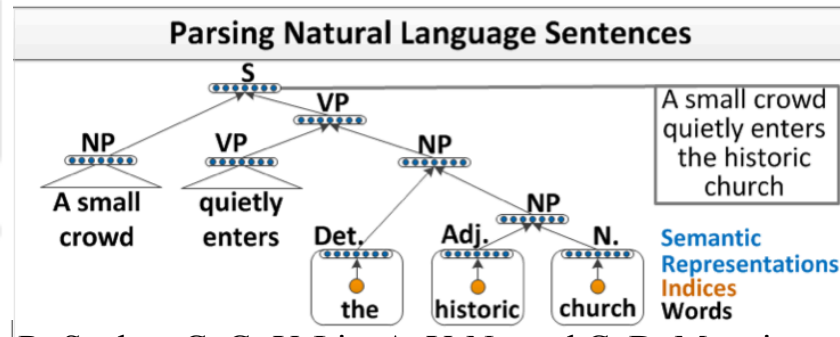
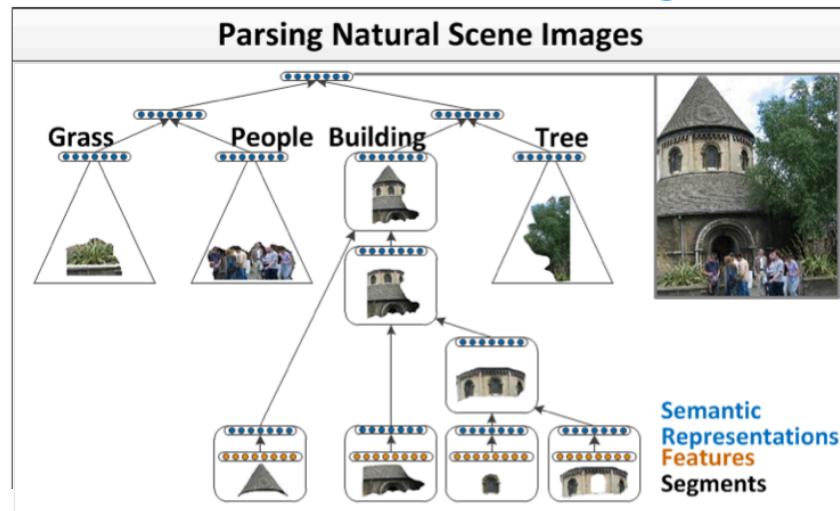
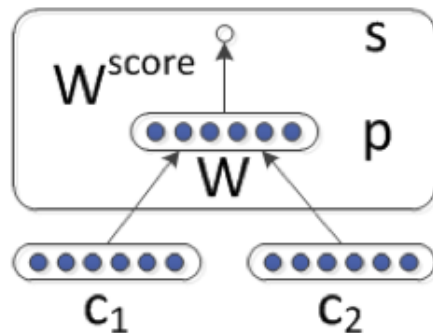
構造や言語を学習する

# Deep Learning と構造

# Recursive Neural Network

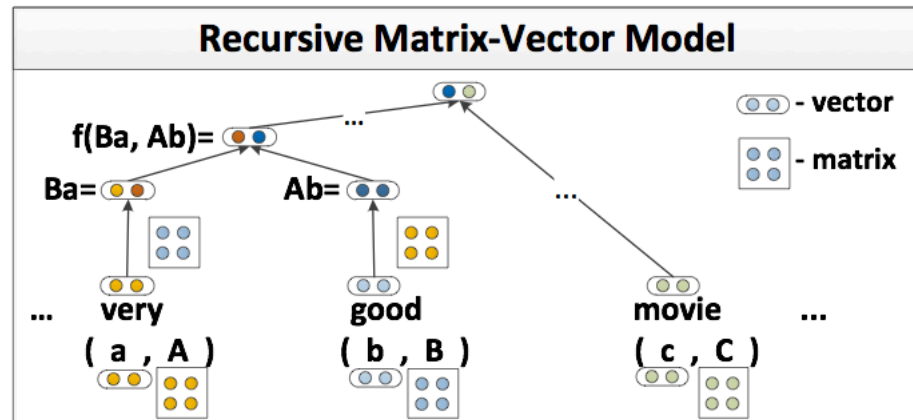
- 同じ重みの層を木の形に重ねる
- 木構造の予測
  - 下図のように2ノードからそれらが兄弟ノードにふさわしいかを判別

- 再帰的な構造を Neural Network で学習
- 木が大きければ deep なモデル



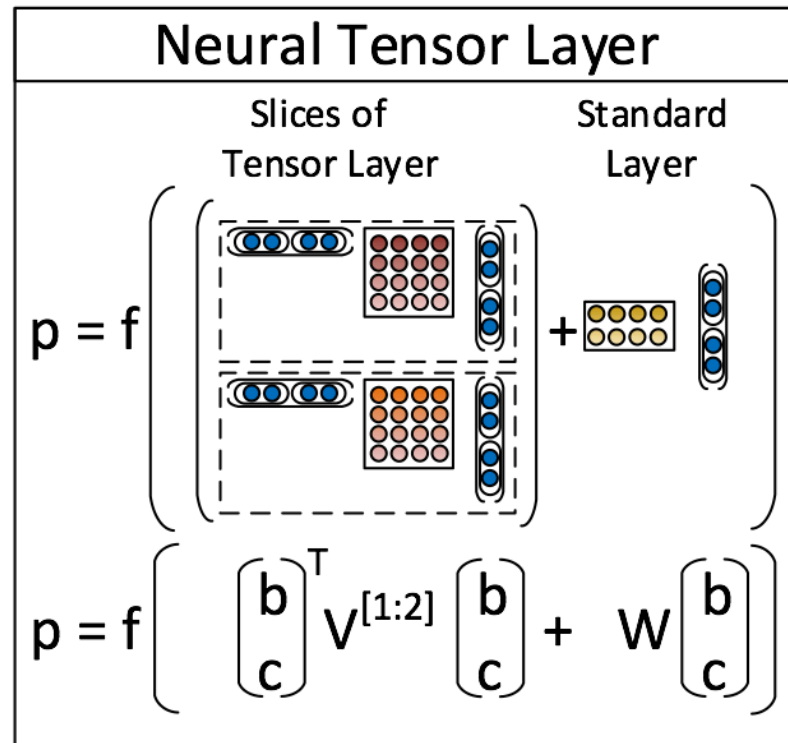
R. Socher, C. C.-Y. Lin, A. Y. Ng and C. D. Manning.  
[Parsing Natural Scenes and Natural Language with Recursive Neural Networks](#). ICML 2011.

# Recursive NN の発展形



R. Socher, B. Huval, C. D. Manning and A. Y. Ng.  
[Semantic Compositionality through Recursive Matrix-Vector Spaces](#). EMNLP 2012.

R. Socher, A. Perelygin, J. Y. Wu, J. Chuang, C. D. Manning,  
 A. Y. Ng and C. Potts.  
[Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank](#). EMNLP 2013.

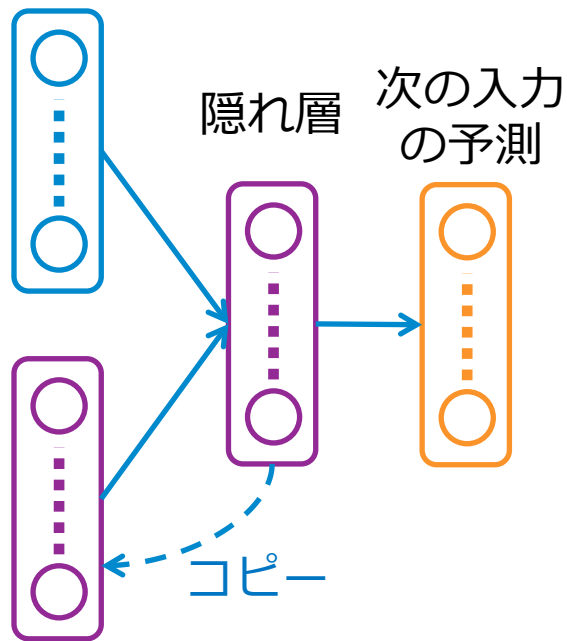


デモあり:

<http://nlp.stanford.edu/sentiment/>

# Recurrent Neural Network Language Model (RNNLM)

文字、単語



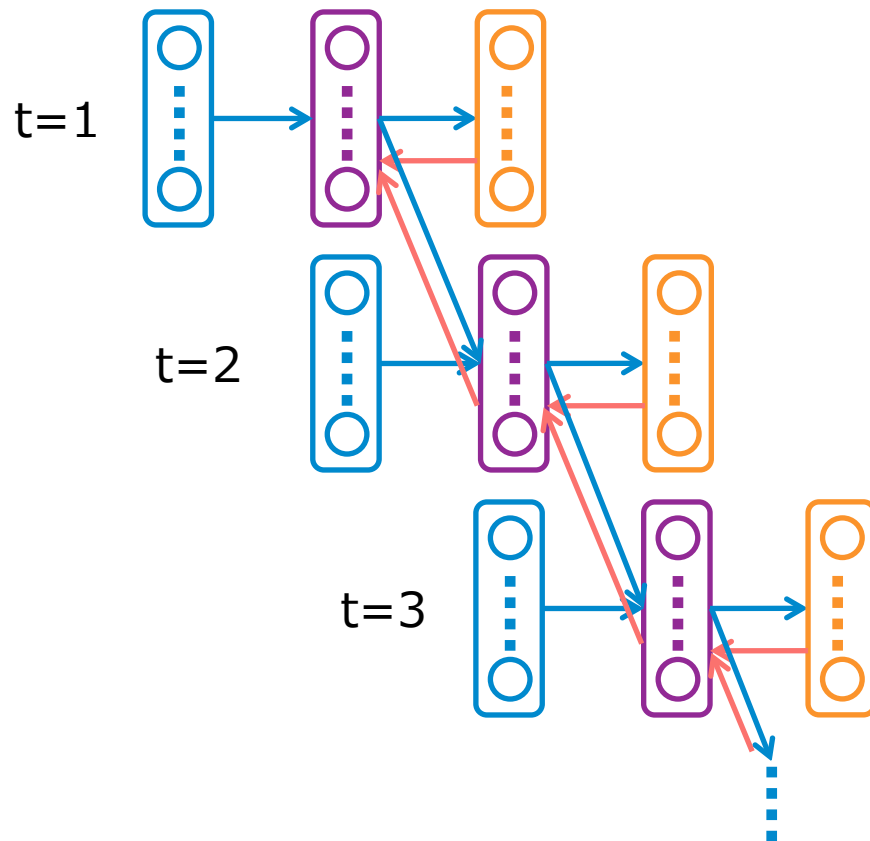
- 文章を読んで、次の文字や単語を予測するモデル（言語モデル）
- Recurrent Neural Network でモデル化
  - 隠れ層の活性が、次の時刻の入力に含まれる
  - 隠れ層は最近の入力に関する記憶を保持する
  - N-gram モデルをゆるく可変長にしたような感じ
  - 隠れ層は単語や文章の**低次元埋め込み (word embeddings)** となっている

T. Mikolov, M. Karafiat, L. Burget, J. H. Cernocky and S. Khudanpur.  
[Recurrent neural network based language model](#). INTERSPEECH 2010.

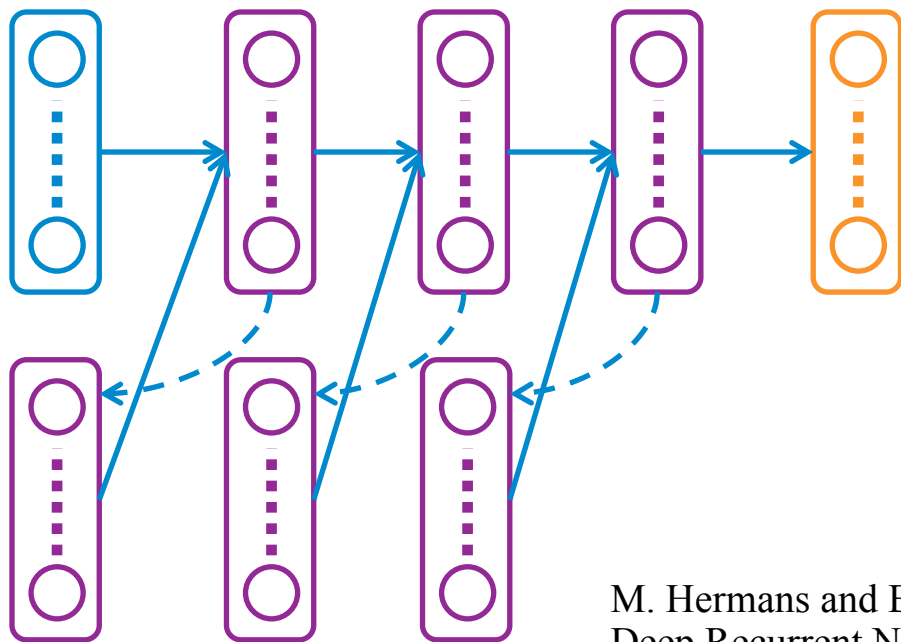
時刻 t-1 の隠れ層

# RNN の学習: Backpropagation through Time

- RNN の適用を時間方向に展開すると DNN のようになる
- 過去の活性を覚えておけば、Backpropagation で勾配が計算できる
- 適当な単語数で打ち切ることも



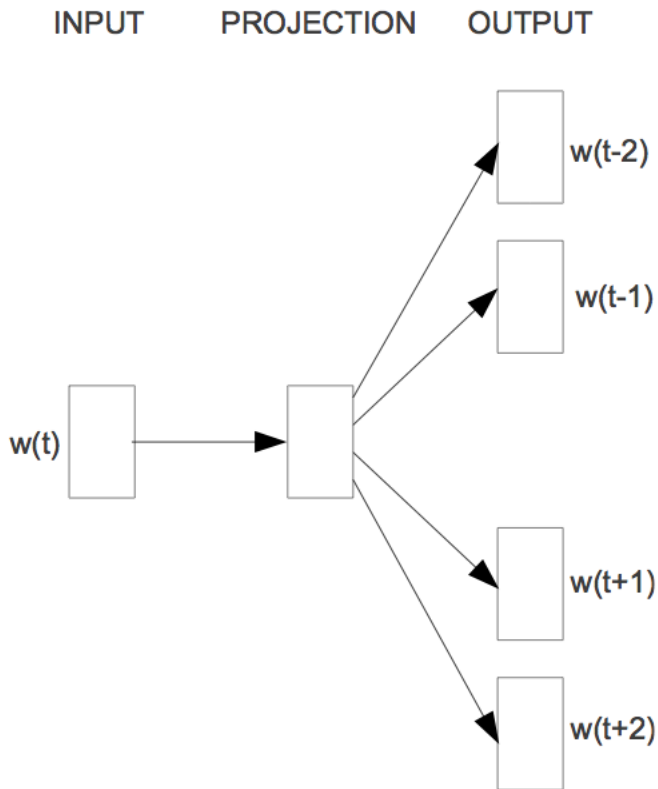
# Deep Recurrent Neural Network



- DNN の各層にループがある Recurrent Net
- 深い層ほど長い時間の記憶を保持する
  - 深くし過ぎると記憶のスケールは変わらなくなる

M. Hermans and B. Schrauwen. Training and Analyzing Deep Recurrent Neural Networks. NIPS 2013.

# Skip-gram model



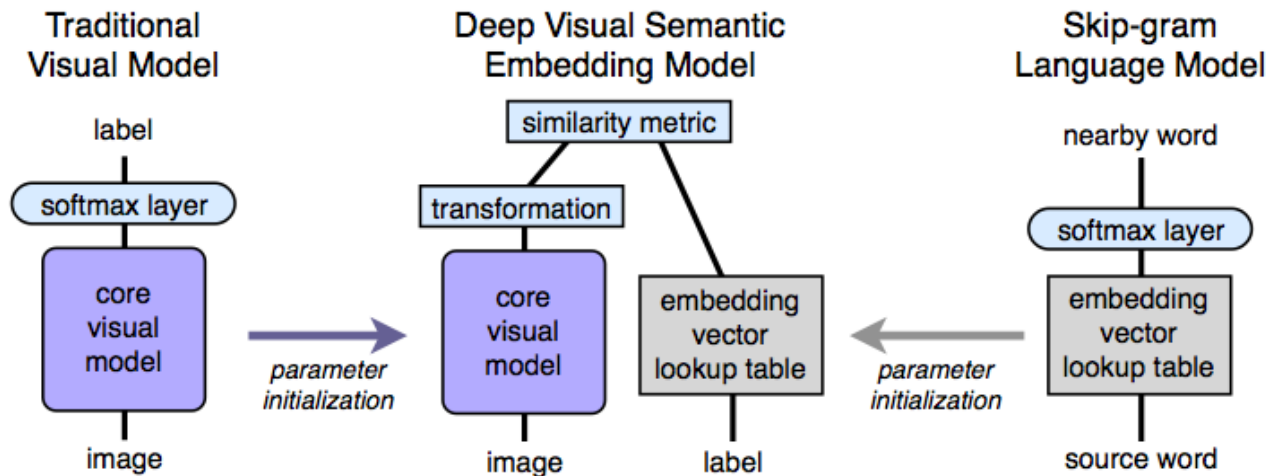
- まわりの単語を予測できるように低次元埋め込みを学習する
- Deep Learning ではないが、単語の表現学習
- Analogical Reasoning に有効
  - $v(\text{"brother"}) - v(\text{"sister"}) + v(\text{"queen"}) \doteq v(\text{"king"})$
- 実装が公開されている: word2vec
  - たくさんの黒魔術

T. Mikolov, K. Chen, G. Corrado and J. Dean.  
[Efficient Estimation of Word Representations in Vector Space.](#)  
ICLR 2013.

# 画像認識との融合: DeVISE

A. Frome, G. S. Corrado, J. Shlens, S. Bengio, J. Dean, M. A. Ranzato and T. Mikolov.

[DeVISE: A Deep Visual-Semantic Embedding Model](#). NIPS 2013.



- Supervision と Skip-gram model を組合せて、画像から単語埋め込みベクトルを予測できるようにする
- 初めて見る物体でも、意味的な事前知識があればラベルを予測できる (zero-shot learning)



# Deep Learning の今後

# 強化学習との統合

- 報酬を最大化するような方策の選び方を深層モデルで学習する
- 手は付けられ始めている: Deep Q-Networks
  - ゲームプレイングのタスク。POMDP の設定で、行動価値関数を過去数フレームの画面に対する畳み込みニューラルネットで表現する
- DeepMind (先日 Google に買収された)

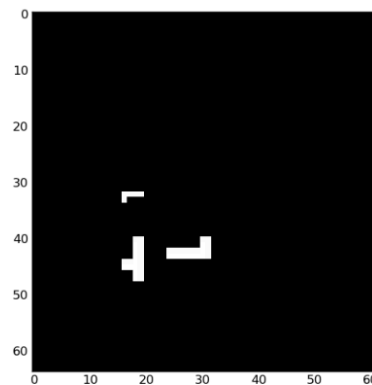
	B. Rider	Breakout	Enduro	Pong	Q*bert	Seaquest	S. Invaders
Random	354	1.2	0	-20.4	157	110	179
Sarsa [3]	996	5.2	129	-19	614	665	271
Contingency [4]	1743	6	159	-17	960	723	268
DQN	<b>4092</b>	<b>168</b>	<b>470</b>	<b>20</b>	<b>1952</b>	<b>1705</b>	<b>581</b>
Human	7456	31	368	-3	18900	28010	3690
HNeat Best [8]	3616	52	106	19	1800	920	<b>1720</b>
HNeat Pixel [8]	1332	4	91	-16	1325	800	1145
DQN Best	<b>5184</b>	<b>225</b>	<b>661</b>	<b>21</b>	<b>4500</b>	<b>1740</b>	1075

# Neural Network の教育

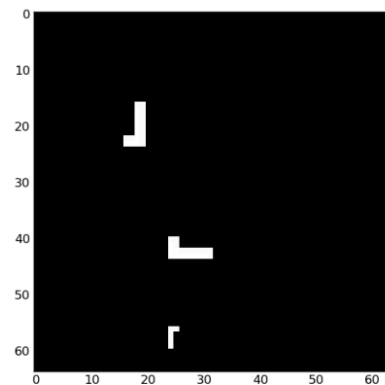
- ペントミノがすべて同じ形かの二値分類は直接学習できない
  - ペントミノの識別を NN で学習したあと、出力層を取り替えて2層足せば学習できる

C. Gulcehre and Y. Bengio.

[Knowledge Matters: Importance of Prior Information for Optimization](#). NIPS Deep Learning Workshop 2012.



(a) sprites, not all same type



(b) sprites, all of same type

- Curriculum Learning\*
- 論理的な思考を学習させるには適切な教育が必要
  - 論理的な思考をどうやってモデル化するかという問題自体を考える必要もある

\* Y. Bengio, J. Louradour, R. Collobert and J. Weston.

[Curriculum Learning](#). ICML 2009.

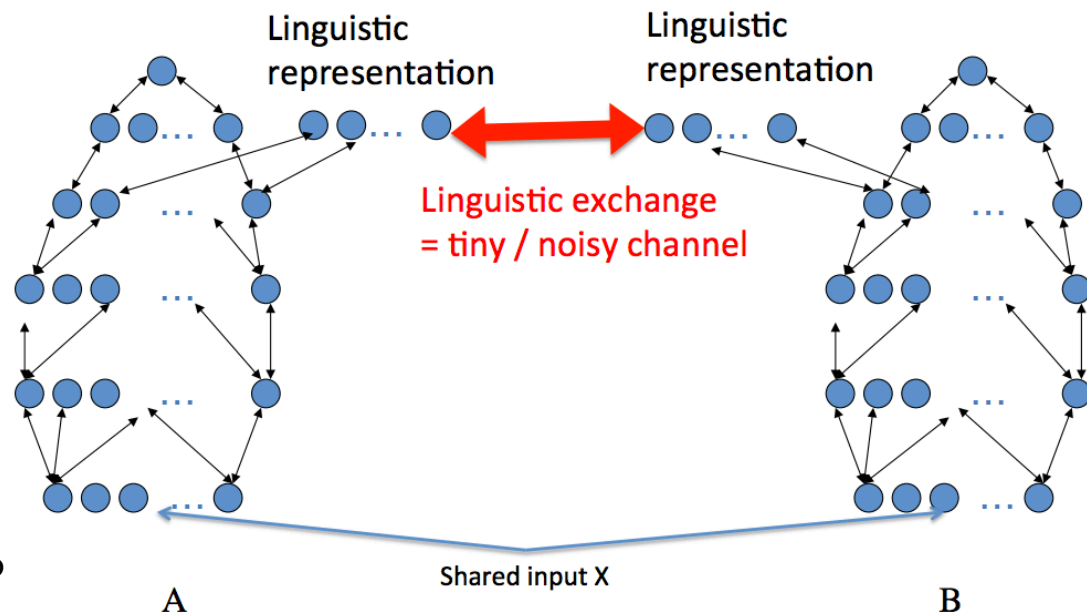
# 空間認識

- 画像分類を超えて、検出、追跡、空間把握へ
  - 検出：物体の位置を特定する
  - 追跡：連続するフレーム間での検出結果のひも付け
  - 空間把握：三次元的な検出、何がどこにあるのか、自分がどこを向いているか
- 分類と検出の統合はすでに始まっている
  - ILSVRC2013 に出場した LeCun らのチーム OverFeat は Supervision ベースの分類・検出システムを構築した
- 音声や運動（ロボティクス）との統合

P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus and Y. LeCun  
[OverFeat: Integrated Recognition, Localization and Detection using Convolutional Networks](#). ArXiv 1312.6229.

# 言語による文化の学習

- 言語によるコミュニケーションを通じて知識や常識を共有
- 深い層の活性を共有するイメージ
  - 共有するときにノイズが乗る



Y. Bengio. [Evolving Culture vs Local Minima](#).  
ArXiv 1203.2990, 2012.

## まとめ

- Deep Learning の重要な技術を広く浅く紹介しました
- 2014 年はさらに応用が広がる年になると思います
  - 研究者人口の増加、大企業の参入
- 基礎研究、理論解析も着実に増えています
  - 特に Dropout、DAE、確率的ニューロンなど確率的に摂動を加える手法への理論解析が多い印象
  - Recurrent Net の効率的な学習も進歩してきています
- 神経科学との関連性は今後の課題

© Preferred Infrastructure, Inc. 2014