LIAMA/NLPR
Institute of automation

火龙果 · 整理
uml.org.cn

A Brief Introduction to Deep Learning

Chunguo Li

火龙果 · 整理
uml.org.cn

Outline

- What is deep learning?
- Motivation for deep architecture
- Learning deep architechure
  - *Supervised* learning algorithms
  - *Unsupervised* and *semi-supervised* learning algorithms
- Comments
- Open questions

火龙果 · 整理
uml.org.cn

Objective of Machine Learning

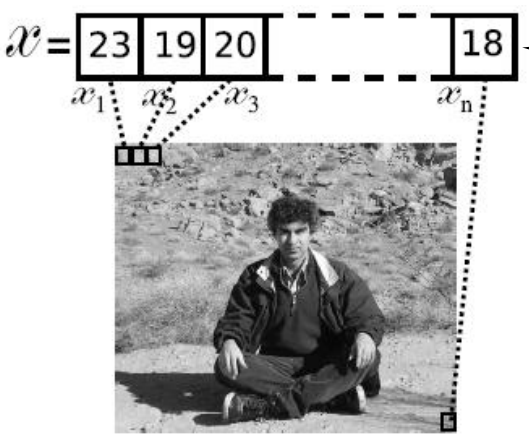
- Inference problem:
  - Infer the states of the unobserved variables.
- Learning problem:
  - Adjust the interactions between variables to make the network more likely to generate the observed data.

火龙果 · 整理
uml.org.cn

What is deep learning?

Can Machine do as a human do?

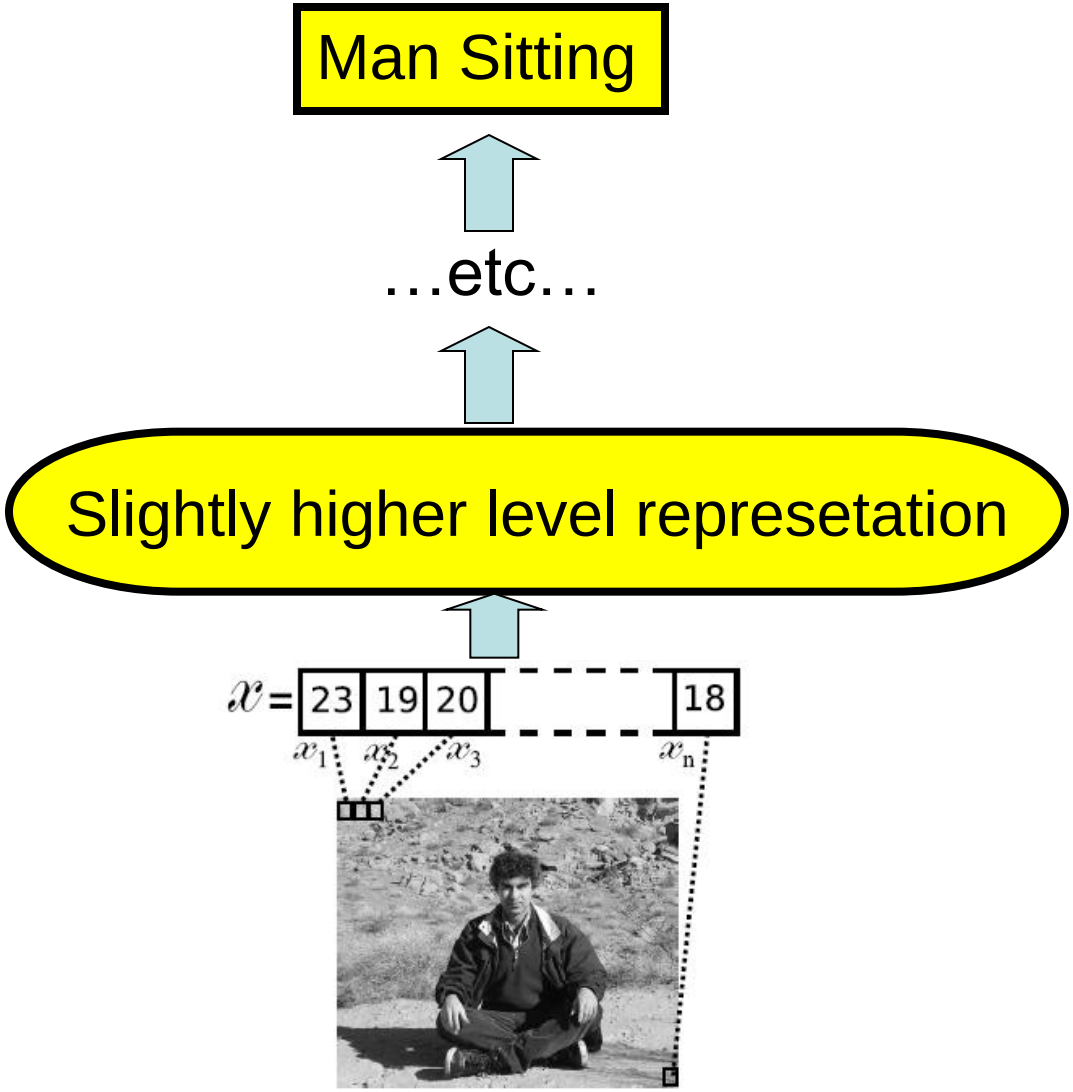
What is deep learning?



Raw input vector representation

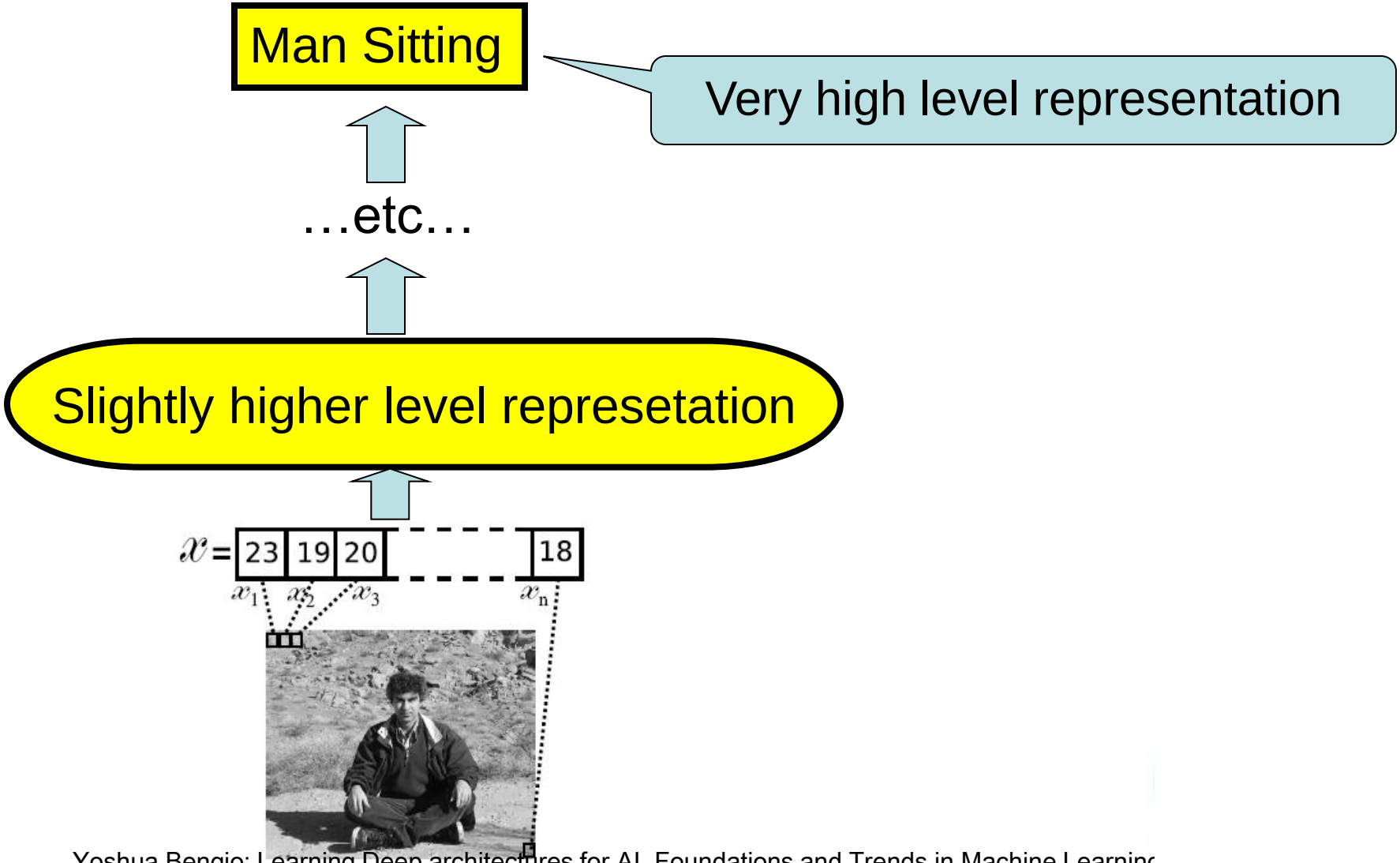
火龙果 · 整理
uml.org.cn

What is deep learning?



火龙果 · 整理
uml.org.cn

What is deep learning?



Yoshua Bengio: Learning Deep architectures for AI, Foundations and Trends in Machine Learning

火龙果 · 整理
uml.org.cn

What is deep learning?

- Deep learning is about learning *multiple levels* of representation and abstraction that help to *make sense* of data such as images, sound, and text.

Machine Learning → Artificial Intelligence

http://deeplearning.net/tutorial/

火龙果 · 整理
uml.org.cn

Depth of an architecture

- Depth of a flow graph
- The longest path from an input node to an output node
  - SVM: depth = 2 = kernels + linear combination
- The number of layers in a neural network
  - MFNN: depth = number of hidden layers + 1

http://deeplearning.net/tutorial/

火龙果 · 整理
uml.org.cn

Outline

- What is deep learning?
- Motivation for deep architecture
- Learning deep architecture
  - *Supervised* learning algorithms
  - *Unsupervised* and *semi-supervised* learning algorithms
- Comments
- Open questions

火龙果 · 整理
uml.org.cn

Motivation for deep architecture

- The brain has a deep architecture.

  Representations are dense distributed and purely local.

- Cognitive process seem deep.
- Theoretical limitations of shallow architectures.

  polynominal-size (k) ***vs*** *exponential-size* (k-1)

火龙果 · 整理
uml.org.cn

Motivation for deep architecture

- Learning algorithms for many shallow architecures are local estimators.
  a) Made of a set of templates
  b) Combine the matching degree

Eg. Kernel machine

$$f(x) = \underbrace{b + \sum_i \alpha_i K(x, x_i)}_{\text{Combine}}$$

火龙果 · 整理
uml.org.cn

Motivation for deep architecture

- Learning algorithms for many shallow architecures are local estimators.
  a) Made of a set of templates
  b) Combine the matching degree

Eg. Kernel machine

$$f(x) = \underbrace{\boxed{b} + \sum_i \boxed{\alpha_i}}_{\text{Combine}} \boxed{K(x, x_i)}$$

Match

Smoothness prior!

火龙果 · 整理
uml.org.cn

Motivation for deep architecture

- Distributed Representation
  - The input pattern is represented by a set of features that are *not mutually exclusive*, and might even be *statistically independent*.

Eg. Clustering algorithms do not build a distributed representation, whereas Independent Component Analysis builds a distributed representation.

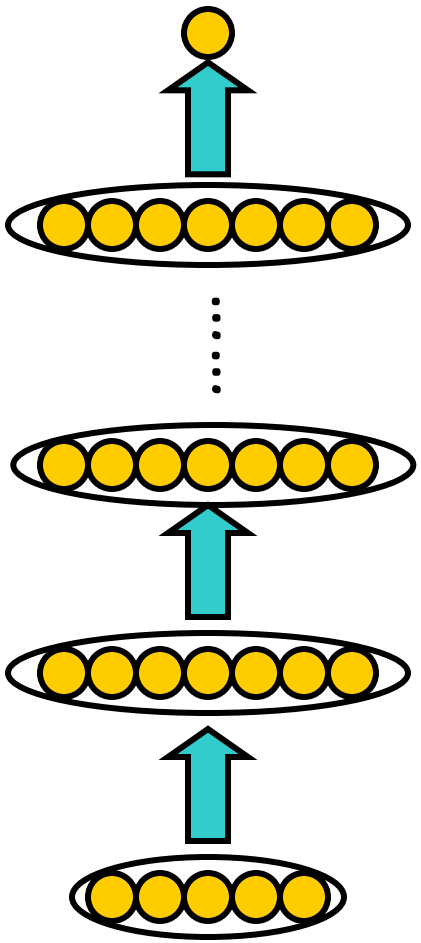
火龙果 · 整理
uml.org.cn

Outline

- What is deep learning?
- Motivation for deep architecture
- Learning deep architecture
  - *Supervised* learning algorithms
  - *Unsupervised* and *semi-supervised* learning algorithms
- Comments
- Open questions

火龙果 · 整理
uml.org.cn

Learning deep architecture



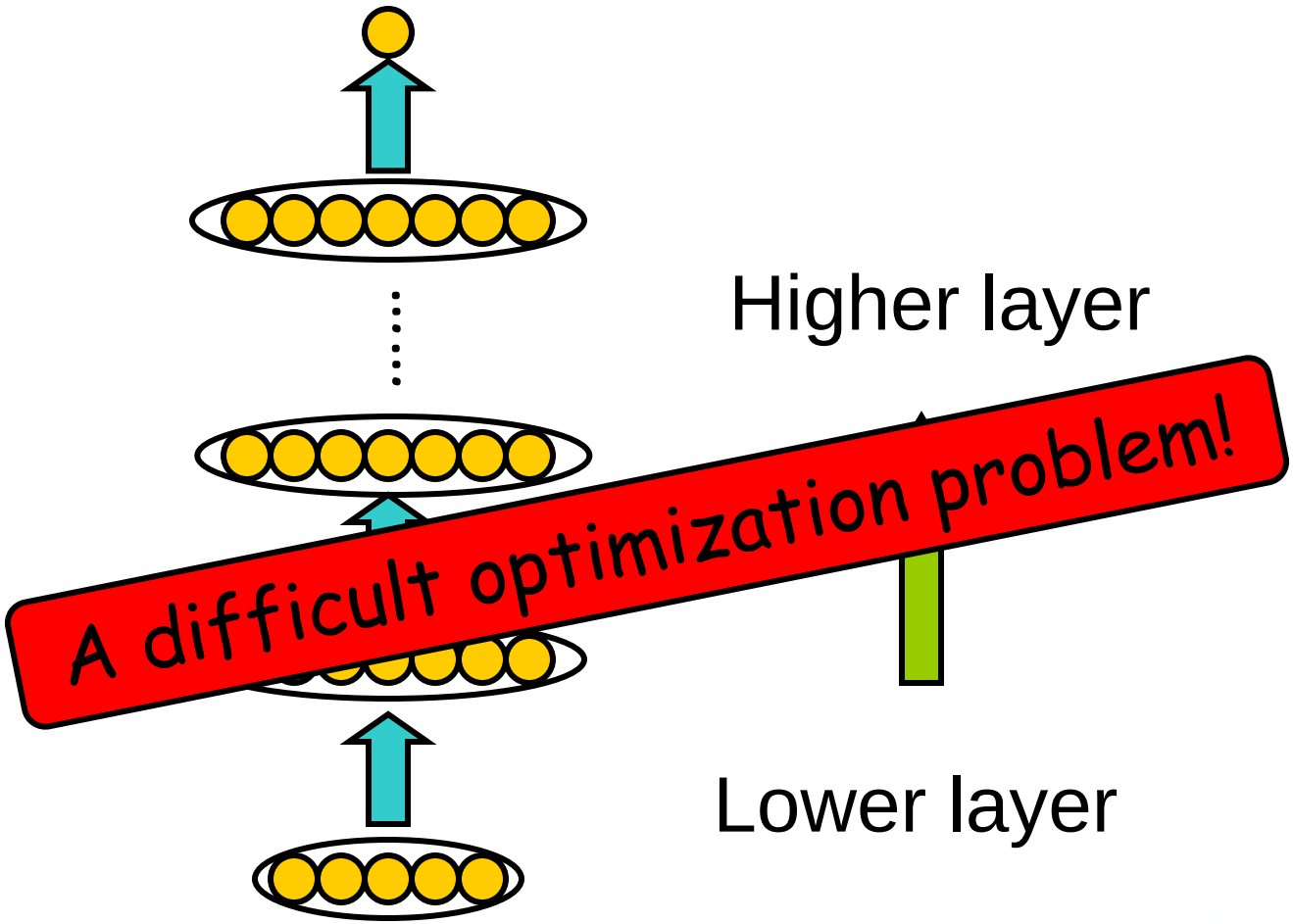
Higher layer

Lower layer

Yoshua Bengio: Learning Deep architectures for AI, Foundations and Trends in Machine Learning

火龙果 · 整理
uml.org.cn

Learning deep architecture



火龙果 · 整理
uml.org.cn

Learning deep architecture

- Convolutional neural networks
  - Lecun, 1989&1998; Simard & Platt, 2003; Ranzato, 2007
- Associate each hidden layer with an autoassociator
- Sigmoid belief networks
  - Hinton et al, 1995 - 2007
- Deep belief networks
  - Hinton et al, 2006; Bengio et al, 2007

http://deeplearning.net/tutorial/

火龙果 · 整理
uml.org.cn

Convolutional neural networks

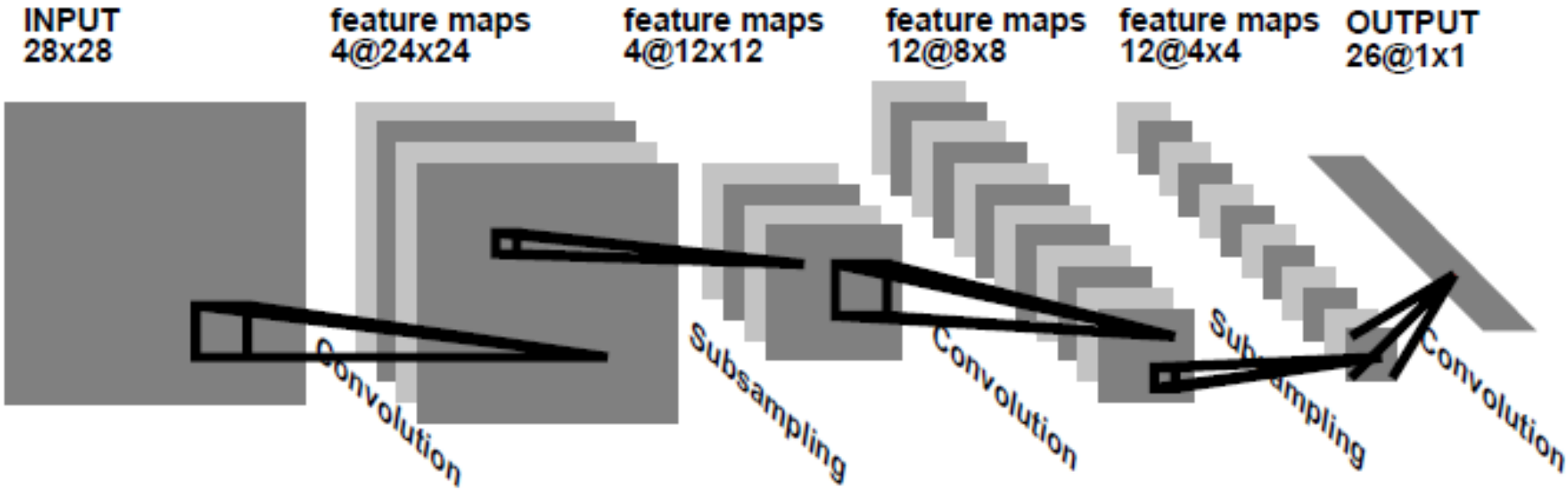
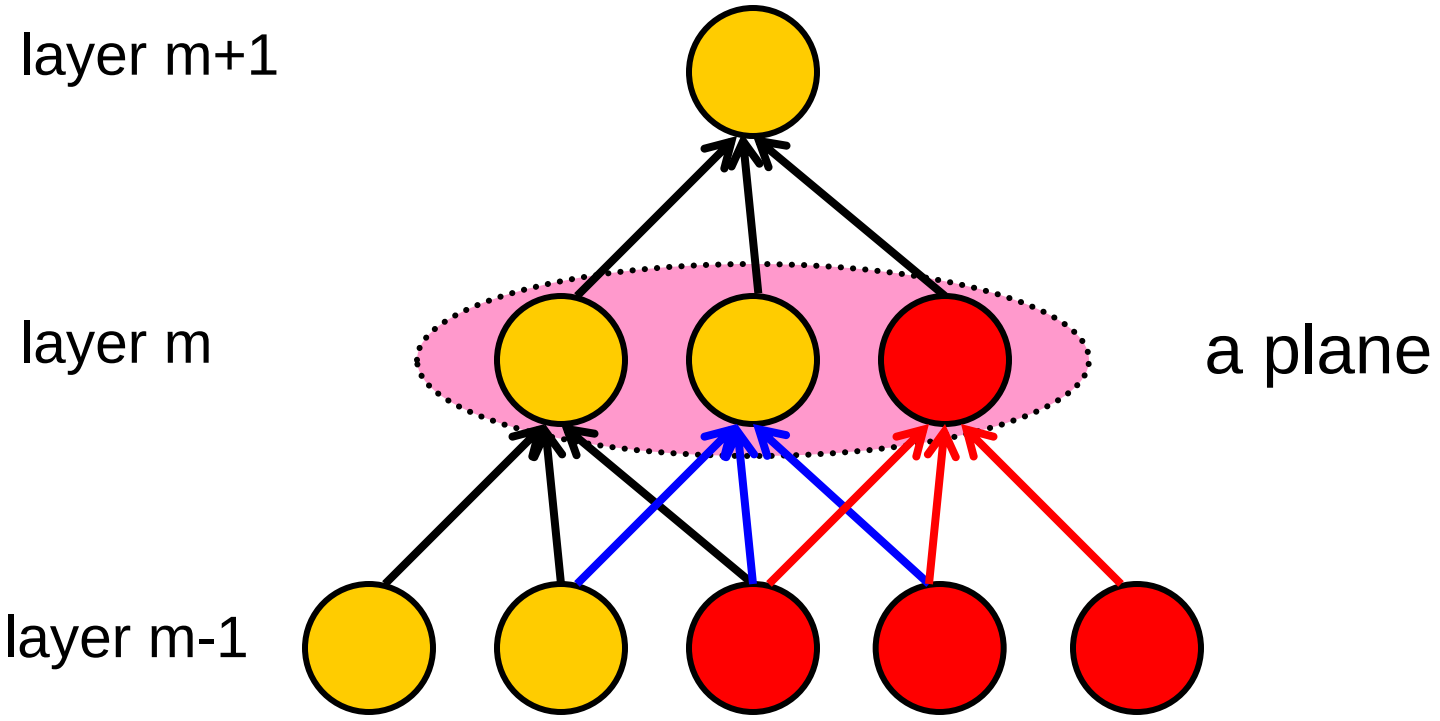


Figure 1. A CNN for image processing, e.g. handwriting

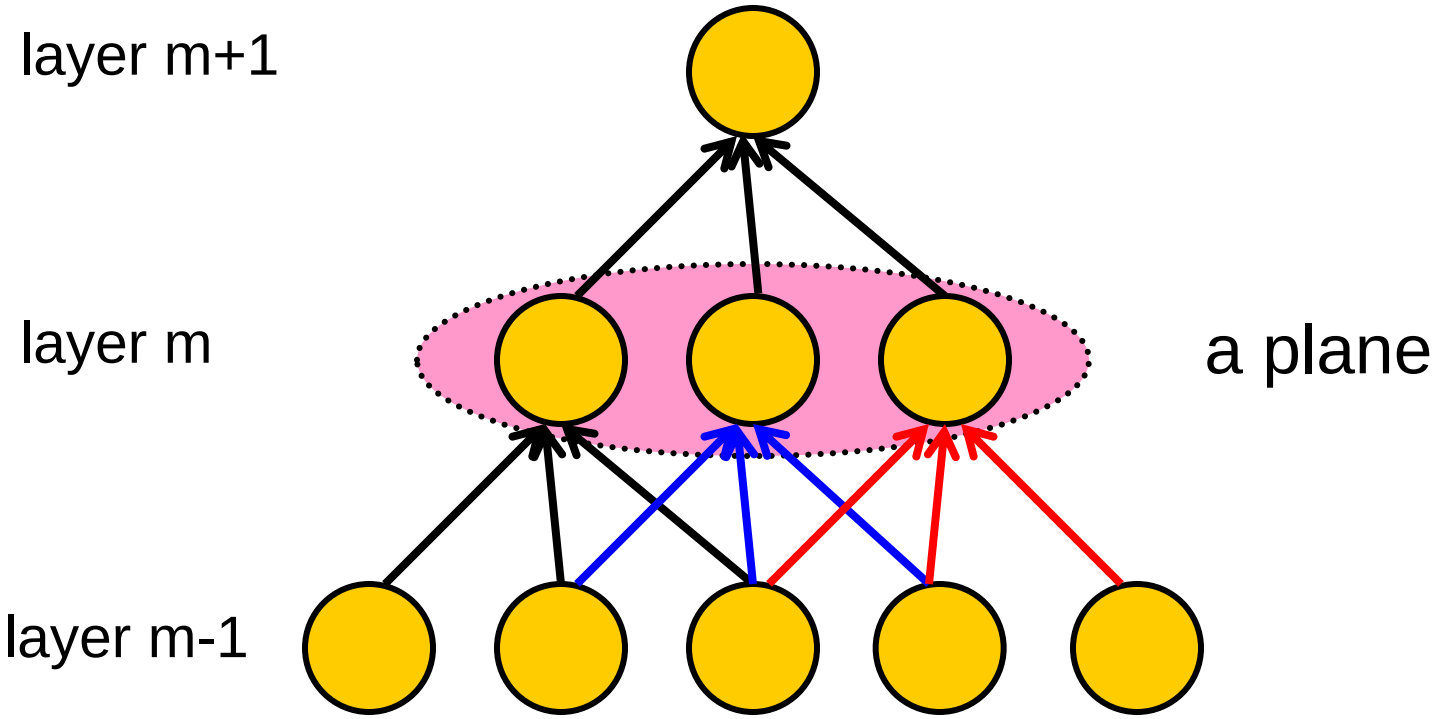
Yann LeCun, Yoshua Bengio: Convolutional Networks for Images, Speech, and Time-Series, In A hand book of Brain Theory and Neural Networks, MIT Press, 1995, 255-258

Convolutional neural networks



http://deeplearning.net/tutorial/

Convolutional neural networks

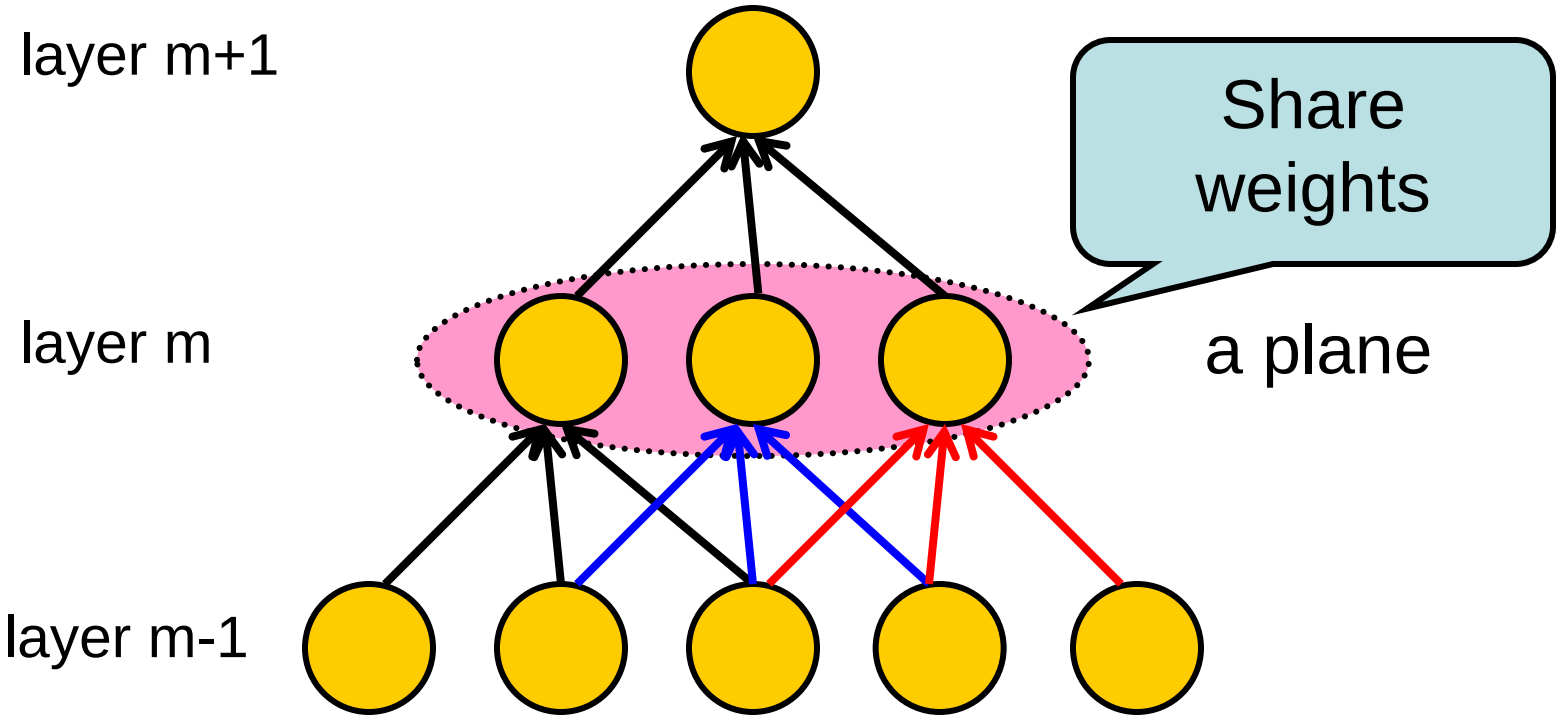


Time-Delayed Neural Network-----TDN

http://deeplearning.net/tutorial/

Convolutional neural networks

Shared Weight Neural Network-----SWNN



http://deeplearning.net/tutorial/

火龙果 · 整理
uml.org.cn

Convolution

- Mathematical function

$$(f * g)(x) = \int f(t)g(x-t)dt = \int g(t)f(x-t)dt$$

$$(f * g)(m) = \sum_{n} f(n)g(m-n)$$

- CNN

$$h_{ij}(v) = \sum_{u} W_{ij}^{T}(u)x(v-u)$$

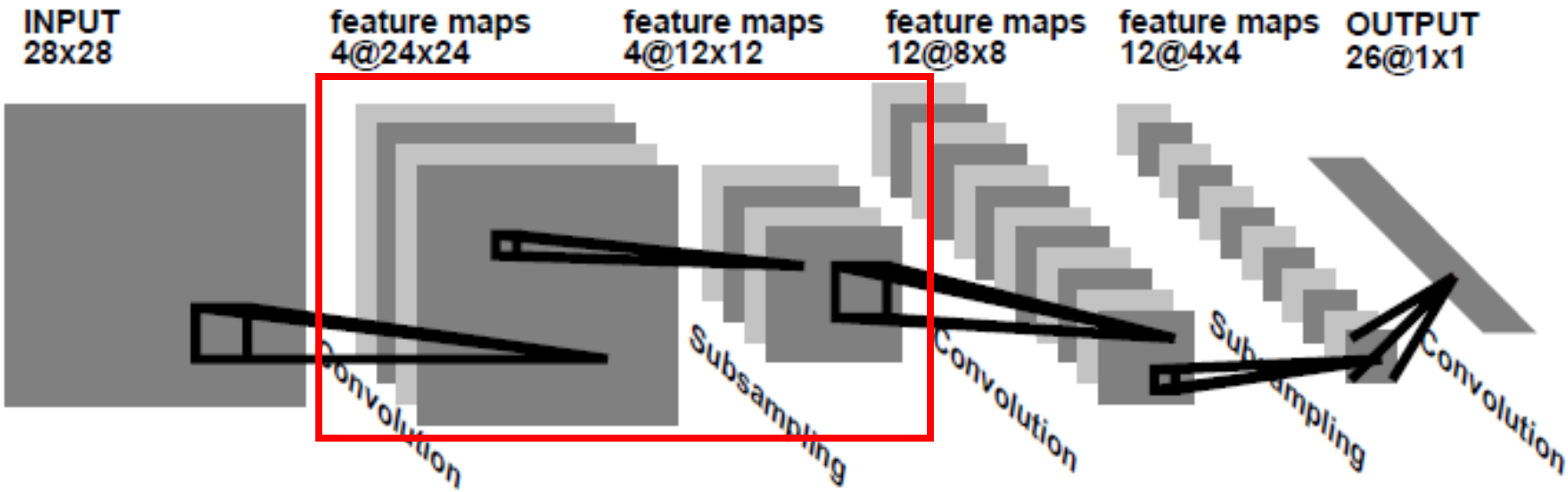
$$y_{ij}(v) = f(h_{ij}(v))$$

http://deeplearning.net/tutorial/

火龙果 · 整理
uml.org.cn

Subsampling

- A feature map ---- a subsampling layer



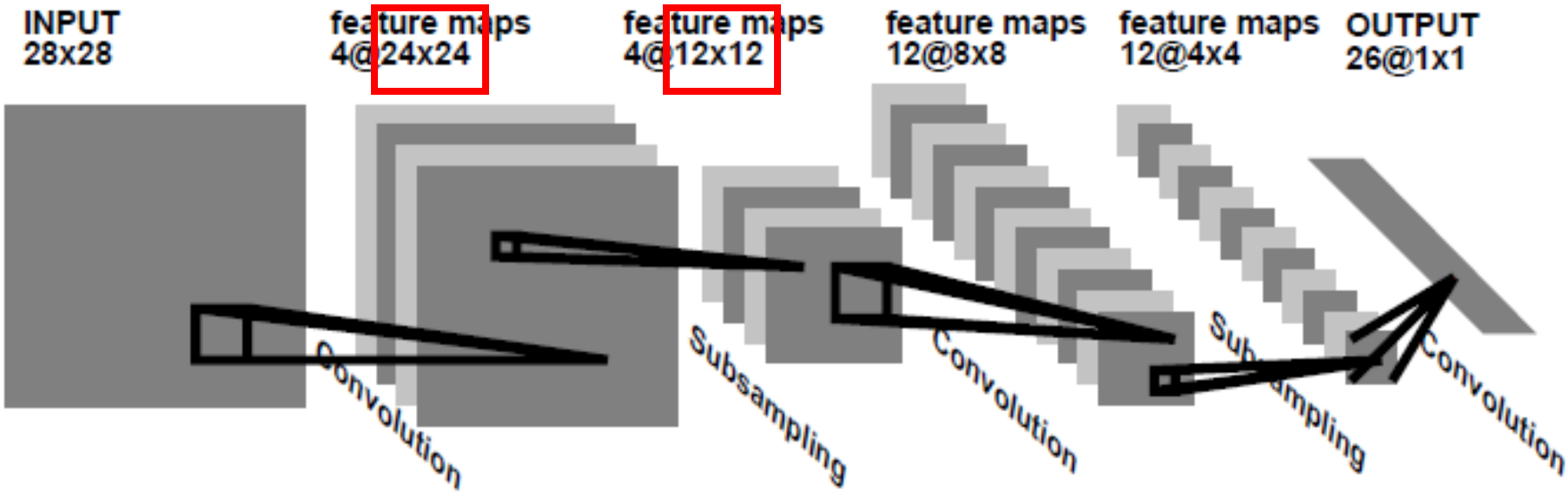
Receptive field: continuous and nonoverlapping.

Yann LeCun, Yoshua Bengio: Convolutional Networks for Images, Speech, and Time-Series, In A hand book of Brain Theory and Neural Networks, MIT Press, 1995, 255-258

火龙果 · 整理
uml.org.cn

Subsampling

- A feature map ---- a subsampling layer



Receptive field: continuous and nonoverlapping.

Yann LeCun, Yoshua Bengio: Convolutional Networks for Images, Speech, and Time-Series, In A hand book of Brain Theory and Neural Networks, MIT Press, 1995, 255-258

火龙果 · 整理
uml.org.cn

Subsampling in vision

- Reduce the computational complexity
- Provide a form of translation inviarance

火龙果 · 整理
uml.org.cn

Note

- Parallel computation
- Local connectivity structure
- Computation cost

$$(M-m)\times(N-n)\times m\times n\times K^{l-1}\times K^{l}$$

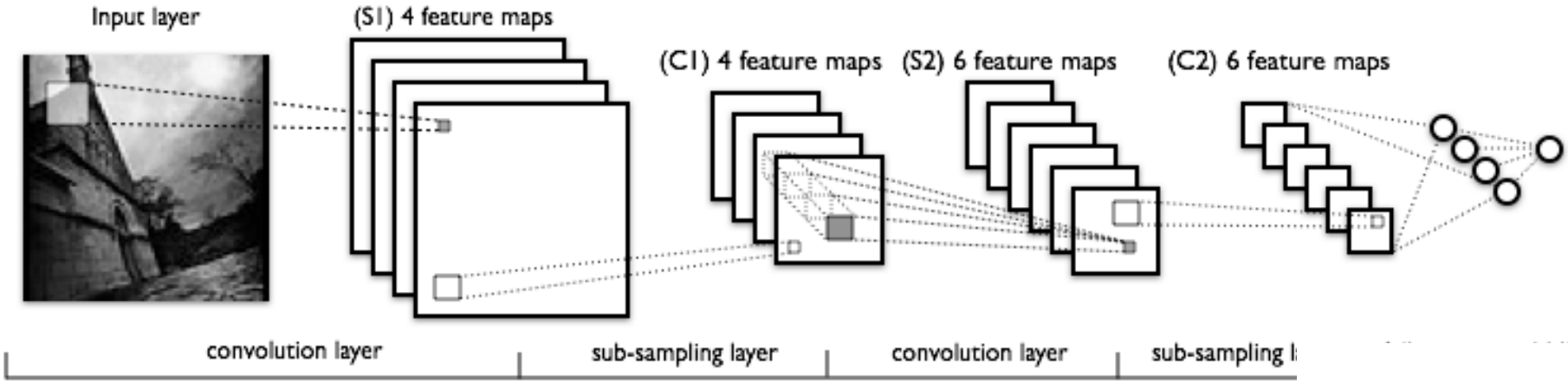
Feature map size decreases with depth.

- Even with random weights in the first layer, a CNN performs well, better than a traine MFNN but worse than a fully optimize

火龙果 · 整理
uml.org.cn

LeNet-5

- LeCun et al designed the LeNet-5 for handwritten and machine-printed character recognition.

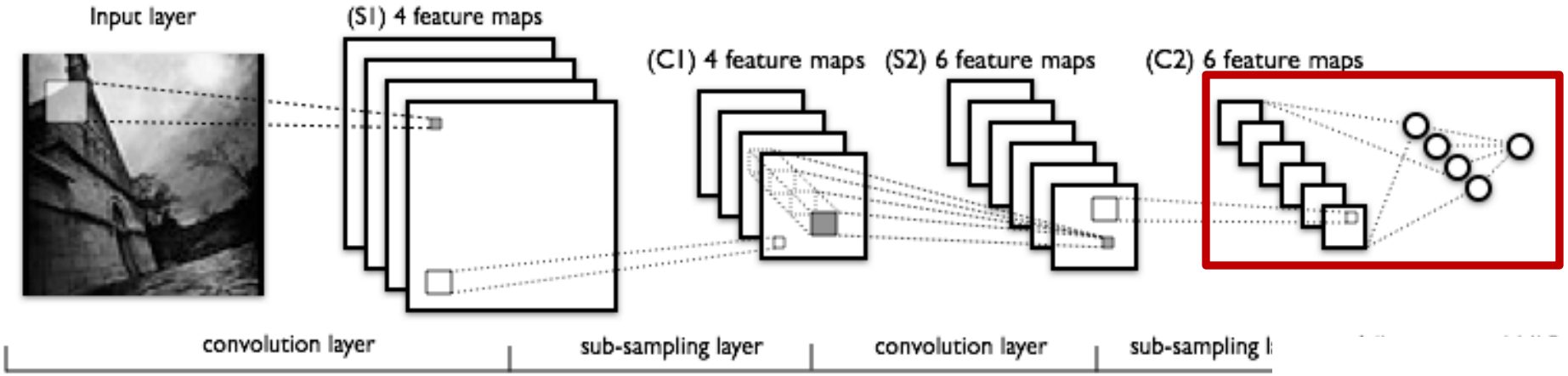


http://deeplearning.net/tutorial/

火龙果 · 整理
uml.org.cn

LeNet-5

- LeCun et al designed the LeNet-5 for handwritten and machine-printed character recognition.



http://yann.lecun.com/exdb/lenet/

火龙果 · 整理
uml.org.cn

Competition on MNIST

- MNIST-----a database of handwritten digits
- a training set of 60,000 examples, and a test set of 10,000 examples.
- A subset of a larger set available from NIST.
- The digits have been size-normalized and centered in a fixed-size image.

http://yann.lecun.com/exdb/mnist/

火龙果 · 整理
uml.org.cn

l net LeNet-1	subsampling to 16x16 pixels	1.7	LeCun et a
l net LeNet-4	none	1.1	LeCun et al. 1998
l net LeNet-4 with K-NN instead of last layer	none	1.1	LeCun et al. 1998
Convolutional net LeNet-4 with local learning instead of last layer	none	1.1	LeCun et al. 1998
Convolutional net LeNet-5, [no distortions]	none	0.95	LeCun et al. 1998
Convolutional net LeNet-5, [huge distortions]	none	0.85	LeCun et al. 1998
Convolutional net LeNet-5, [distortions]	none	0.8	LeCun et al. 1998
Convolutional net Boosted LeNet-4, [distortions]	none	**0.7**	LeCun et al. 1998
unsupervised sparse features + SVM, [no distortions]	none	0.59	Labusch et al., IEEE TNN 2008
Convolutional net, cross-entropy [affine distortions]	none	0.6	Simard et al., ICDAR 2003
Convolutional net, cross-entropy [elastic distortions]	none	**0.4**	Simard et al., ICDAR 2003
large conv. net, random features [no distortions]	none	0.89	Ranzato et al., CVPR 2007
large conv. net, unsup features [no distortions]	none	0.62	Ranzato et
large conv. net, unsup pretraining [no distortions]	none	0.60	Ranzato et
large conv. net, unsup pretraining [elastic distortions]	none	**0.39**	Ranzato et
large conv. net, unsup pretraining [no distortions]	none	0.53	Jarrett et a

火龙果 · 整理
uml.org.cn

Autoassociator

- An autoassociator is trained to encode the input in some representation so that the input can be reconstructed from that representation.
  - autoencoder, Diabolo network
  - supervised learning
  - only work well for examples similar to those in the training set

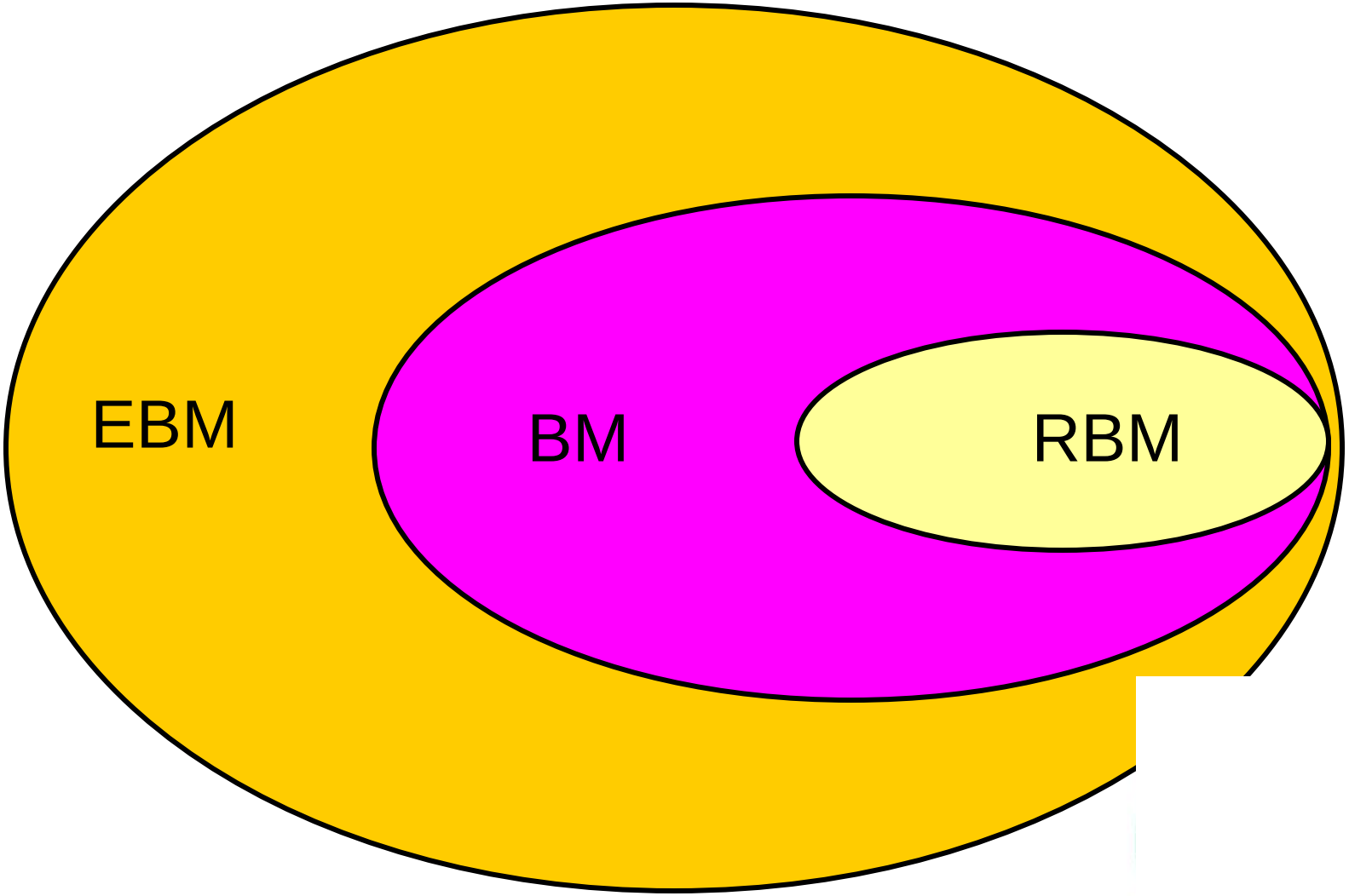
火龙果 · 整理
uml.org.cn

Autoassociator

- Linear hidden layer: *k* hidden units to *k* captures features ----- PCA
- Nonlinear hidden layer: capture multi-modal aspects of input distribution.
  - Important when stacking multiple encoders (with corresponding decoders) are considered to build a deep auto-encoder.
- Non-linear auto-encoders with more hidden units than inputs yield useful representations.

火龙果 · 整理
uml.org.cn

Restricted Boltzmann Machines



火龙果 · 整理
uml.org.cn

Energy Based Model---EBM

- EBM associates a scalar energe to each configuration of the variables of interest.
  - e.g. energy-based probabilistic models define a probability distribution through an energy function.

$$p(x) = \frac{e^{-E(x)}}{Z} \qquad \text{where} \qquad Z = \sum_{x} e^{-E(x)}$$

Loss function:

$$l(\theta, D) = -\frac{1}{N} \sum_{x \in D} \log p(x)$$

http://deeplearning.net/tutorial/

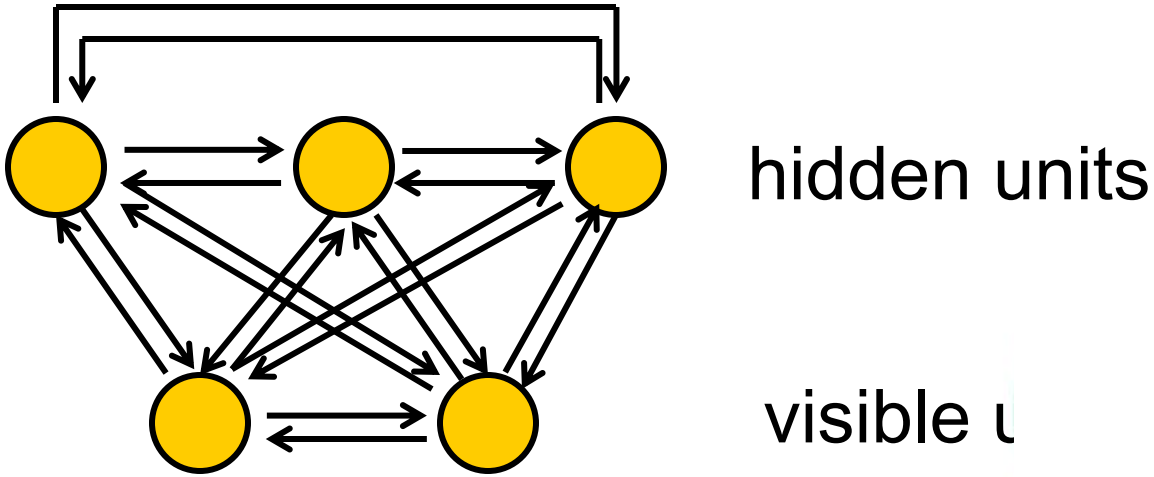
EBMs with hidden units

- Non-observed variables

$$p(x)=\sum_h p(x,h)=\sum_h \frac{e^{-E(x,h)}}{Z}$$

- Boltzmann Machine

$$E(x,h)=-b'x-c'h-h'Wx-x'Ux-h'Vh$$

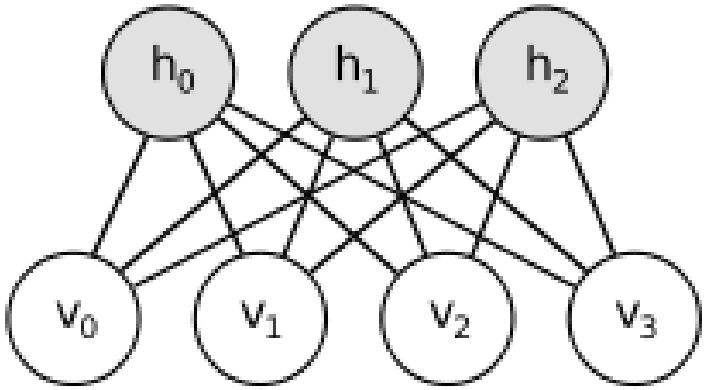


火龙果 · 整理
uml.org.cn

Restricted Boltzmann Machine

- Energy function

$$E(x,h) = -b'x - c'h - h'Wx$$



hidden layer

visible layer

Note: a RBM adds randomness in the transf
from input to reconstruction by the view of a
associator.

火龙果 · 整理
uml.org.cn

Train a RBM

- Objective :

$$\min_{\theta} \quad l(\theta, D) = -\frac{1}{N}\sum_{x \in D} \log p(x)$$

$$p(x) = \sum_{h} p(x,h) = \sum_{h} \frac{e^{-E(x,h)}}{Z}$$

$$E(x,h) = -b'x - c'h - h'Wx$$

- Gradient based

$$E_{\hat{P}}\left(\frac{\partial \log p(x)}{\partial \theta}\right)$$

where $\hat{P}$ is the train

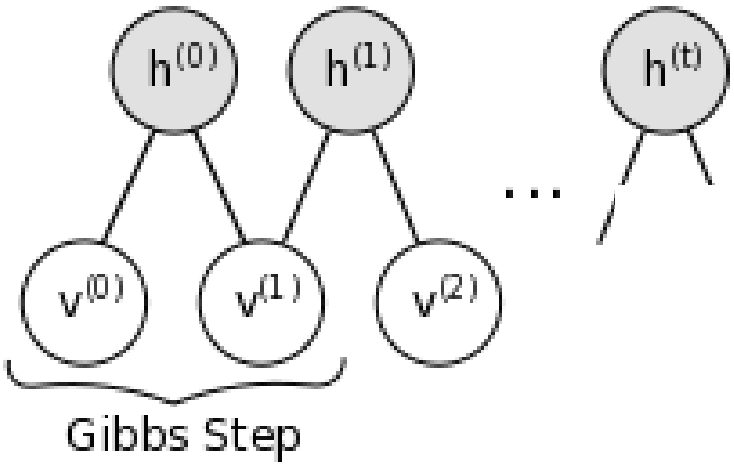
http://deeplearning.net/tutorial/

Sampling from a RBM

- Benifts:
  - To obtain an estimator for the log-likelihood gradient in the learning algorithm.
  - To get an idea of what the model has captured or not captured about the data distribution.
- Gibbs sampling

$$p(x|h) = \prod_i p(x_i|h)$$

$$p(h|x) = \prod_i p(h_i|x)$$



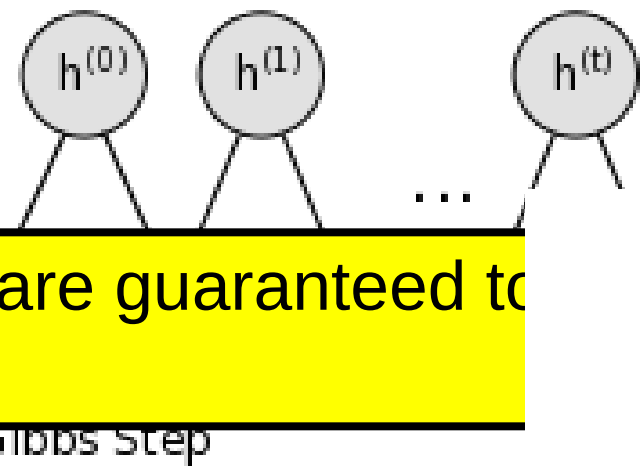
http://deeplearning.net/tutorial/

火龙果 · 整理
uml.org.cn

Sampling from a RBM

- Benifts:
  - To obtain an estimator for the log-likelihood gradient in the learning algorithm.
  - To get an idea of what the model has captured or not captured about the data distribution.
- Gibbs sampling

$$p(x|h) = \prod_i p(x_i|h)$$



As t—>∞, samples ($x^{(t)}$,$h^{(t)}$) are guaranteed to accurate samples of p(x,h).

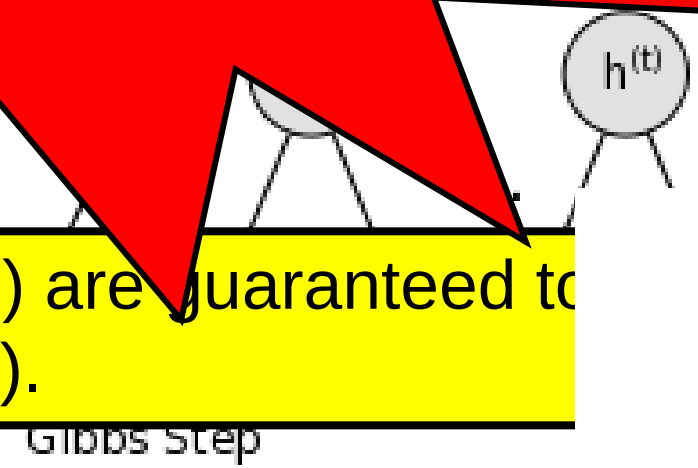
http://deeplearning.net/tutorial/

火龙果 · 整理
uml.org.cn

Sampling from a RBM

- Benifts:
  - To [illegible] an estimate for [illegible]-likelihood gradient in the learning [illegible]
  - [illegible] not capture [illegible]
- Gibbs [illegible]

$$p(x|h) = \prod_i p(x_i|h)$$



Prohibitively Expensive!!!

As t—>∞, samples ($x^{(t)}$,$h^{(t)}$) are guaranteed to be accurate samples of p(x,h).

http://deeplearning.net/tutorial/

火龙果 · 整理
uml.org.cn

Contrastive Divergence

- Initialize the Markov Chain with a training example.
- Samples are obtained after only k-steps of Gibbs sampling
- For each parameter update, the state of the chain is then preserved for subsequent updates.
- Use pseudo-likelihood as the proxy to the likelihood, much less expensive.

火龙果 · 整理
uml.org.cn

Deep Belief Networks

- RBMs can be stacked and trained in a greedy manner to form Deep Belief Networks.
- A DBN with l layers models the joint distribution between observed vector x and l hidden layers h^k as follows:

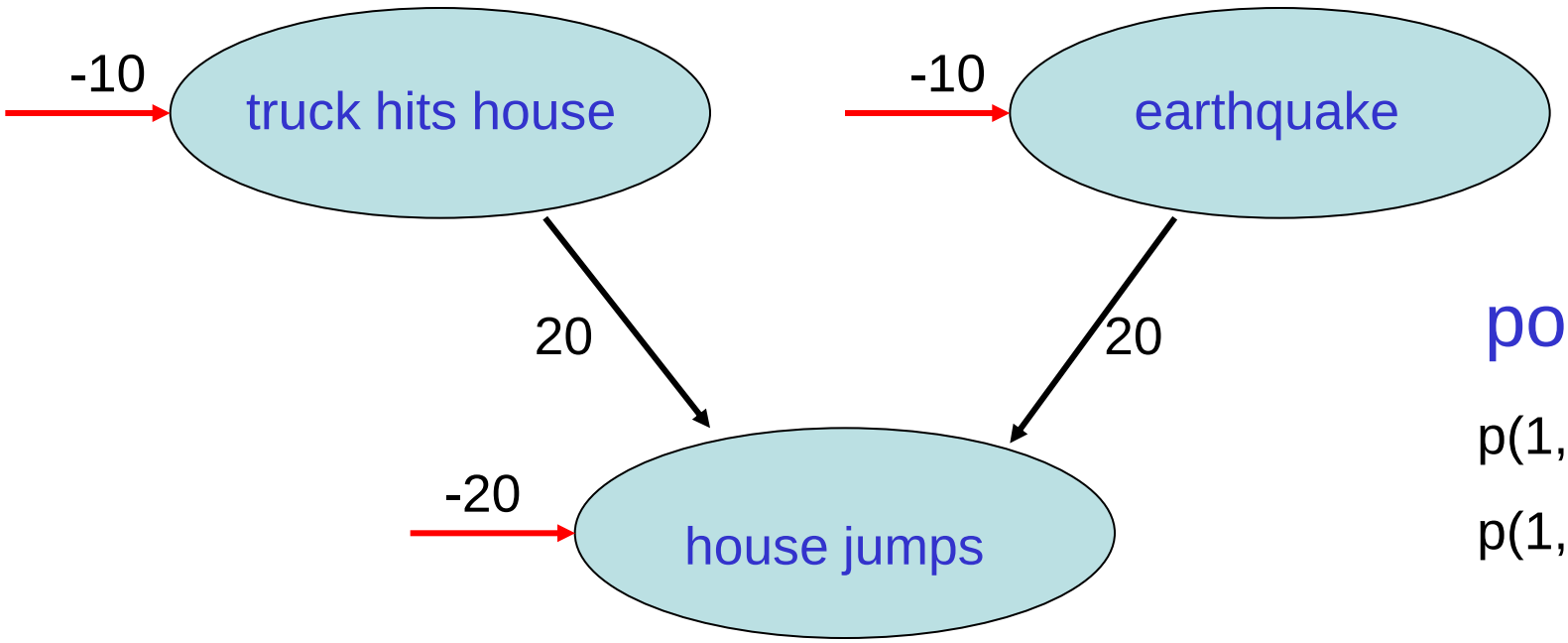
$$P\left(x, h^1, h^2, \cdots, h^l\right) = \left(\prod_{k=0}^{l-2} P\left(h^k \middle| h^{k+1}\right)\right) P\left(h^{l-1}, h^l\right)$$

Yoshua Bengio: Learning Deep architectures for AI, Foundations and Trends in Machine Learning

火龙果 · 整理
uml.org.cn

Deep Belief Networks

- Explaining away



posterior

p(1,1)=.0001

p(1,0)=.4999

火龙果 · 整理
uml.org.cn

What is wrong with BP?

- It requires labeled training data.
  - Almost all data is unlabeled.
- The learning time does not scale well.
  - It is very slow in networks with multiple hidden layers.
- It can get stuck in poor local optimal.
  - These are often quite good, but for deep nets they are far from optimal.

火龙果 · 整理
uml.org.cn

Overcoming the limitation of BP

- Keep the efficiency and simplicity of using a gradient method for adjusting the weights, but use it for modeling the structure of the sensory input.
  - Adjust the weights to maximize the probability that a generative model would have produced the sensory input.
  - Learn $p(x)$ not $p(y|x)$

火龙果 · 整理
uml.org.cn

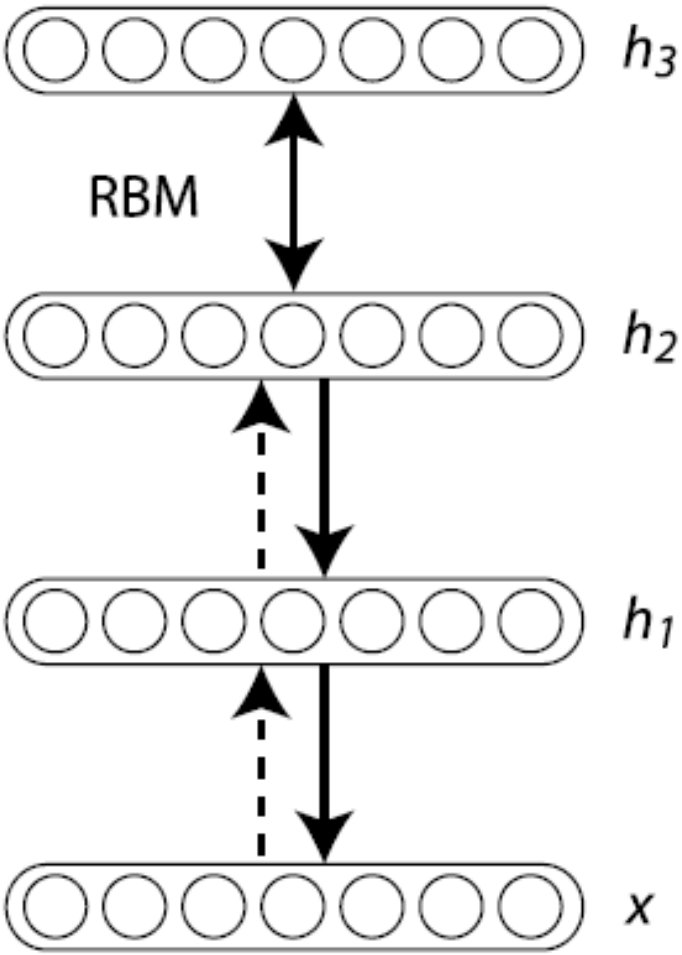
A fast learning algorithm

- Learn W assuming all the weights are tied.
- Freeze W and commit ourselves to using W^T to infer factorial approximate posterior distributions over the states of the variables in the first hidden layer, even if subsequent changes in higher-level weights mean that this inference method is no longer correct.
- Keeping all the higher-weights tied to each other, but untied from W, learn a RBM model of the higher level "data" that was produced by using W^T to transform the original data.
- Fine-tune all the parameters by wake-sleep algorithm

Geoffrey E. Hinton, Simon Osindero, Yee-Whye Teh: A fast learning algorithm for deep belief nets
Computation 18, 2006, 1527-1554

火龙果 · 整理
uml.org.cn

Deep Belief Networks

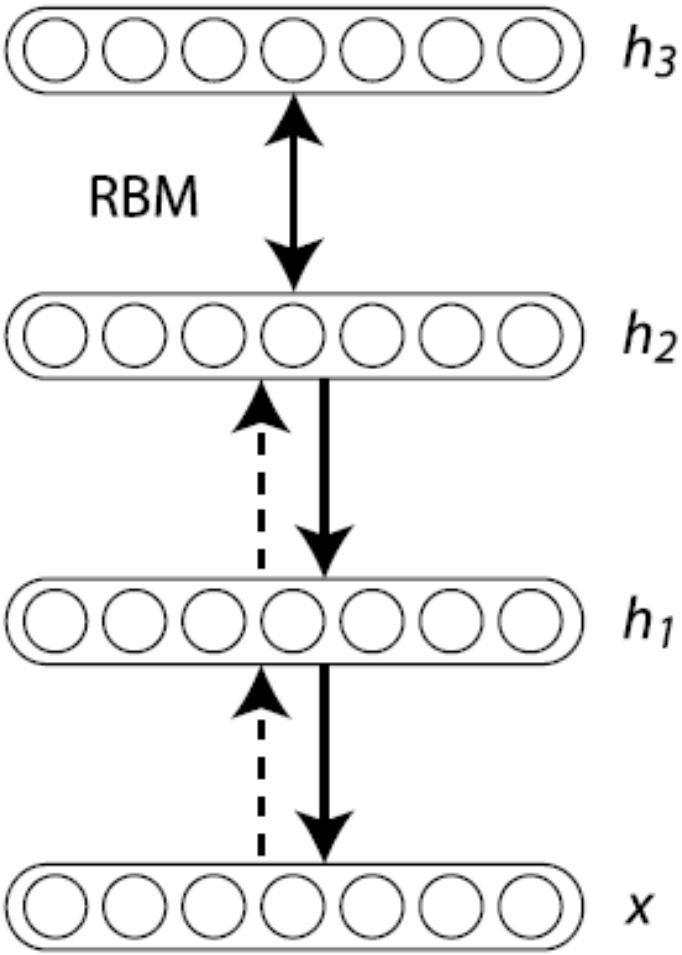


Traine the 1st layer a

Yoshua Bengio: Learning Deep architectures for AI, Foundations and Trends in Machine Learning

火龙果 · 整理
uml.org.cn

Deep Belief Networks



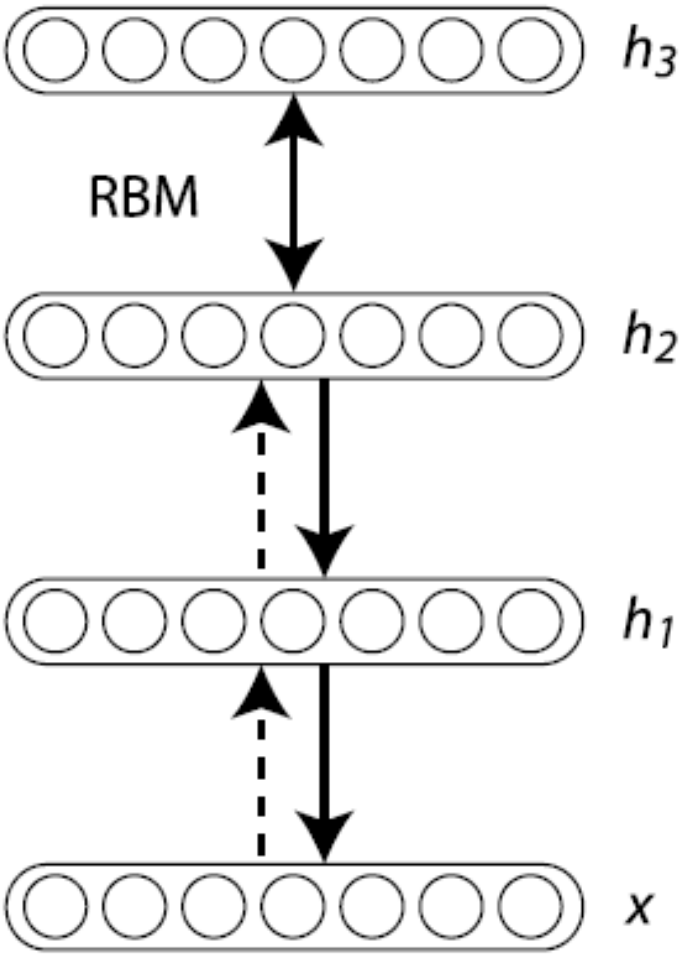
Obtain a representation of the input

Traine the 1st layer

Yoshua Bengio: Learning Deep architectures for AI, Foundations and Trends in Machine Learning

火龙果 · 整理
uml.org.cn

Deep Belief Networks



Train the next layer as a RBM

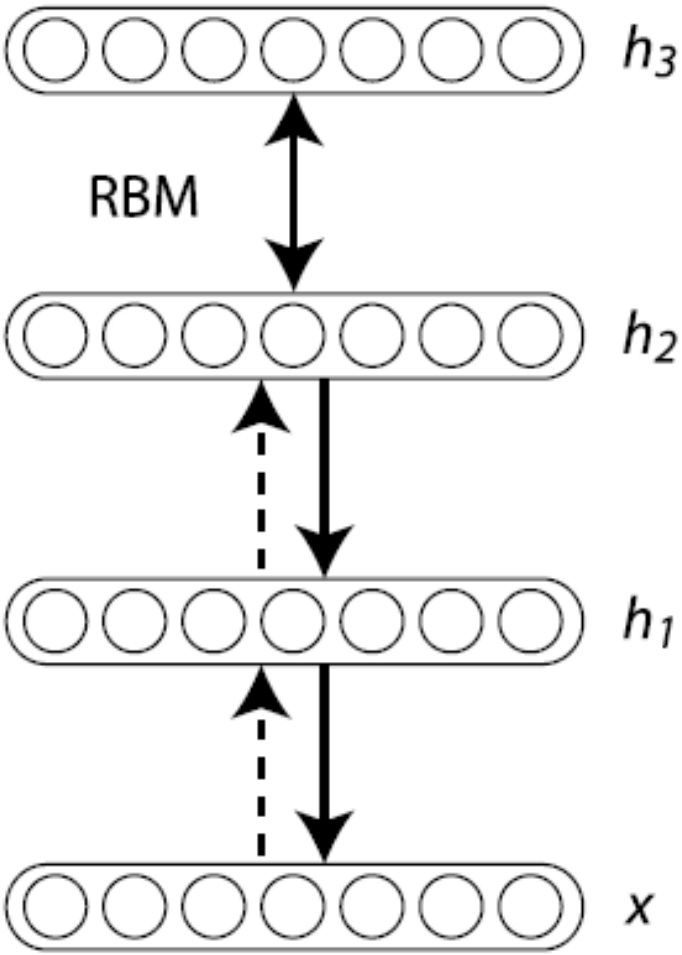
Obtain a representation of the input

Traine the 1st layer

Yoshua Bengio: Learning Deep architectures for AI, Foundations and Trends in Machine Learning

火龙果 · 整理
uml.org.cn

Deep Belief Networks



Fine-tune all the parameters

Train the next layer as a RBM

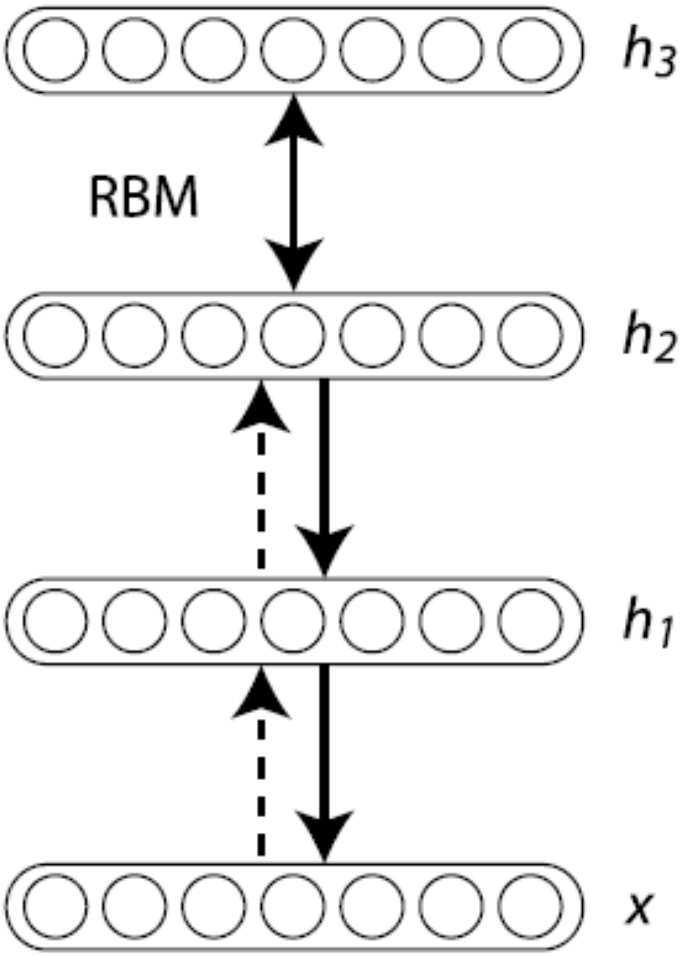
Obtain a representation of the input

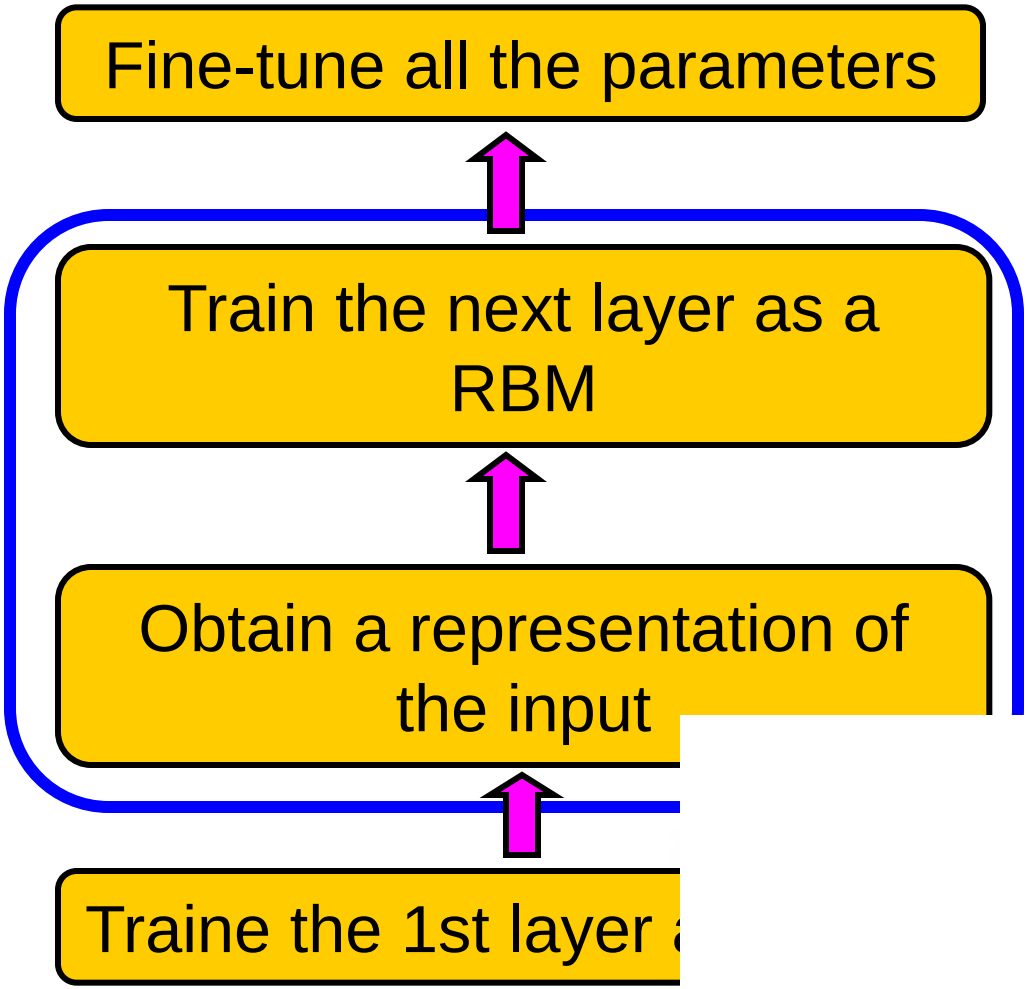
Traine the 1st layer

Yoshua Bengio: Learning Deep architectures for AI, Foundations and Trends in Machine Learning

火龙果 · 整理
uml.org.cn

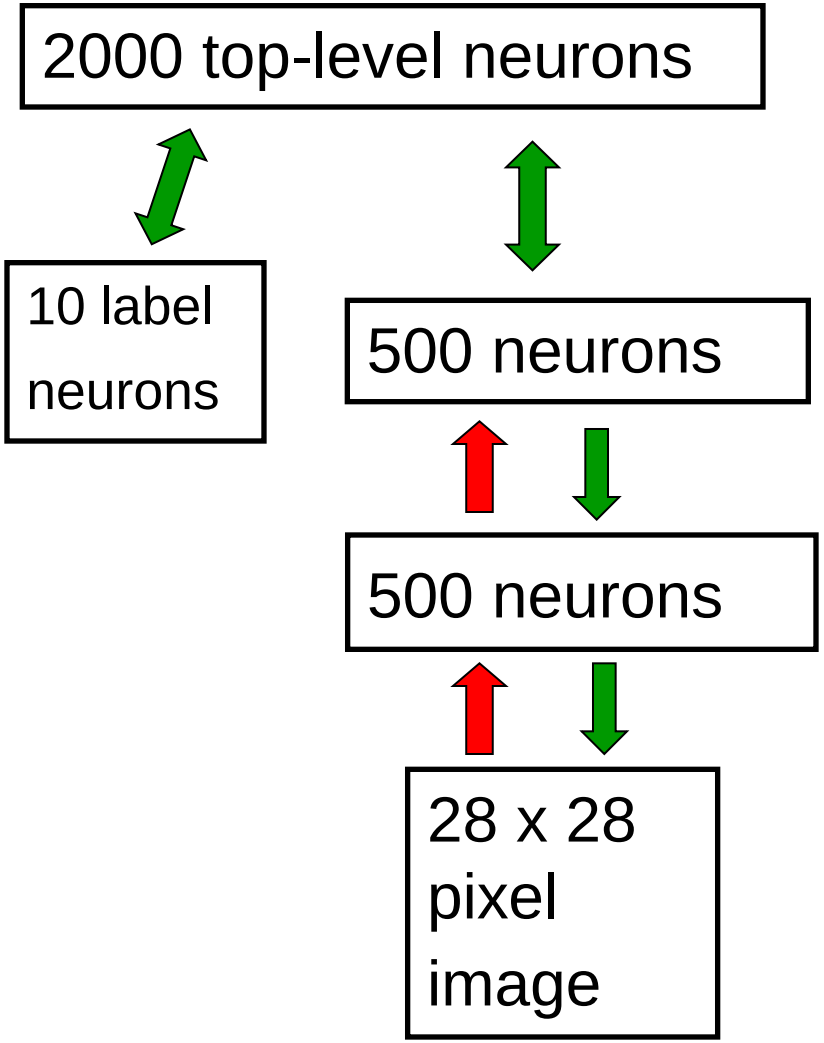
Deep Belief Networks





火龙果 · 整理
uml.org.cn

A Model of Digit Recognition



火龙果 · 整理
uml.org.cn

Movies from Hinton

http://www.cs.toronto.edu/~hinton/adi/index.htm

火龙果 · 整理
uml.org.cn

Outline

- What is deep learning?
- Motivation for deep architecture
- Learning deep architecture
  - *Supervised* learning algorithms
  - *Unsupervised* and *semi-supervised* learning algorithms
- Open questions
- Comments

火龙果 · 整理
uml.org.cn

Open Questions

- Is the presence of local minima an important issue in training RBMs?
- Can we jointly train all the layers with respect to the unsupervised objective?
- Aren't there other efficient trainable deep architecture besides DBNs?
- Can we find analogs of the computations necessary for Contrastive Divergence and DBN learning in the brain?
- Can decision tree ensembles be stacked to ob different type of deep architecture?

火龙果 · 整理
uml.org.cn

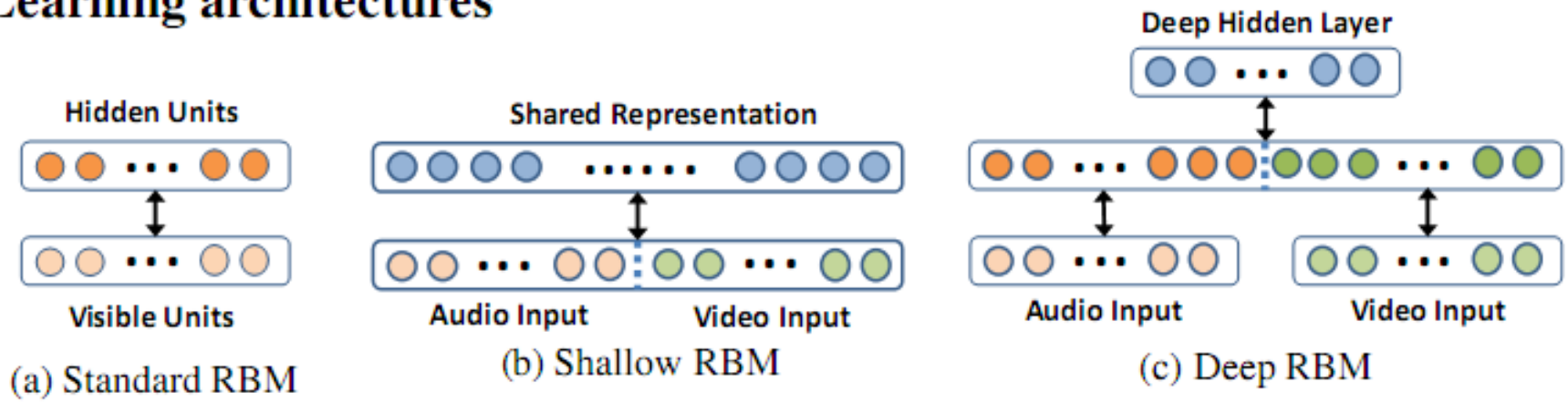
Outline

- What is deep learning?
- Motivation for deep architecture
- Learning deep architecture
  - *Supervised* learning algorithms
  - *Unsupervised* and *semi-supervised* learning algorithms
- Open questions
- Comments

火龙果 · 整理
uml.org.cn

Multimodal DL

Learning architectures



Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee and Andrew Y. Ng: Multimo
Learning, In Proceedings of the Twenty-Eighth International Conference on Machine Learning, 20

火龙果 · 整理
uml.org.cn

Comments

- Entropy might be used to measure the lost/gain after learning layer by layer, which can also propagate forward the network.
- Whether maximizing the probability in DBNs can be replaced by maximizing/minimizing the mutual information if each layer is associated with mutual information?
- Can the divergence proposed in MLSP2 in the place of Contrastive Divergence i learning a RBM?

火龙果 · 整理
uml.org.cn

References

1. Yoshua Bengio: Learning Deep architectures for AI, Foundations and Trends in Machine Learning, 2(1), 2009
2. http://deeplearning.net/tutorial/
3. Yann LeCun, Yoshua Bengio: Convolutional Networks for Images, Speech, and Time-Series, In Arbib, M.A. The hand book of Brain Theory and Neural Networks, MIT Press, 1995, 255-258
4. Geoffrey E. Hinton, Simon Osindero, Yee-Whye Teh: A fast learning algorithm for deep belief nets. Neural Computation 18, 2006, 1527-1554
5. Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee and Andrew Y. Ng: Multimodal Deep Learning, In Proceedings of the Twenty-Eighth International Conference on Machine Learning, 2011
6. Geoffrey E. Hinton: To recognize shapes, first learn to generate images. Technical report, 2006