CCF ADL 2015
Nanchang
Oct 11, 2015

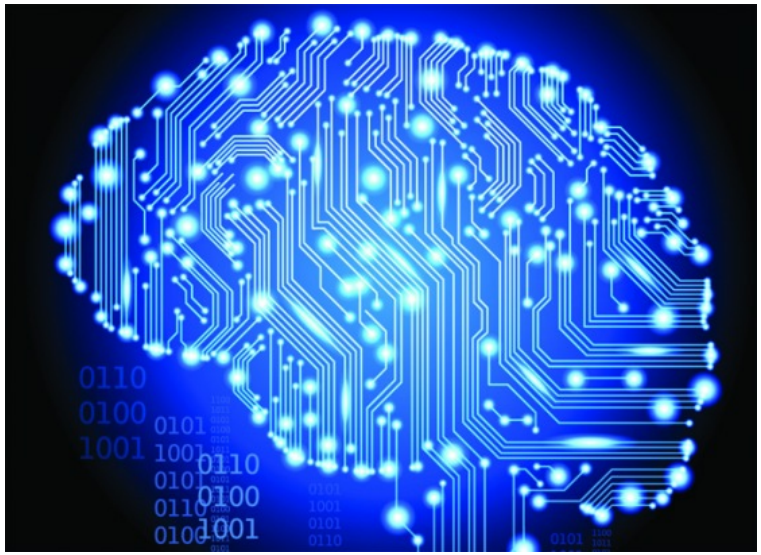
Learning to Process Natural Language in Big Data Environment

Hang Li
Noah's Ark Lab
Huawei Technologies

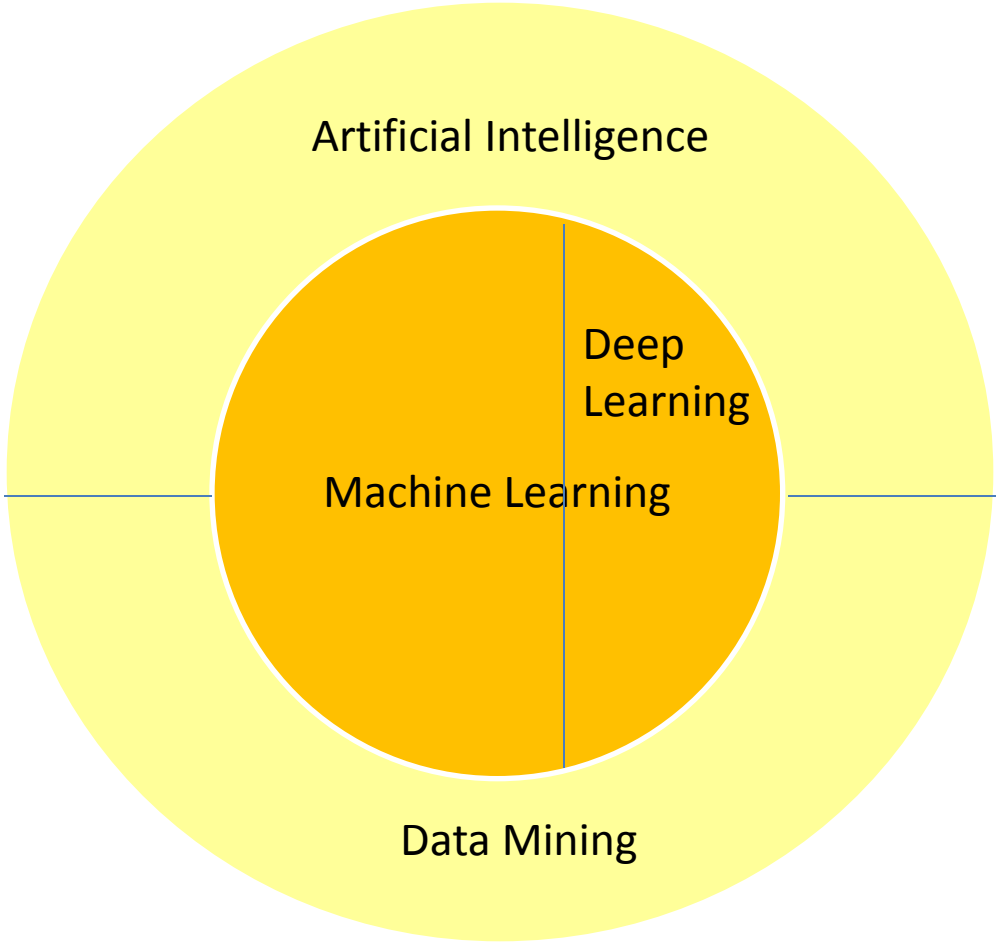
Part 1: Deep Learning - Present and Future

Talk Outline

- *Overview of Deep Learning*
- Recent Advances in Research of Deep Learning
- Future of Deep Learning



General Picture of Machine Learning



Deep learning is important and promising sub-area of machine learning

Brief History of Machine Learning

Perceptron
Rosenblatt

Neural Net
Rumelhart et al

SVM
Vapnik et al

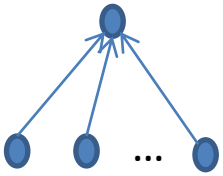
Deep Learning
Hinton et al

1957

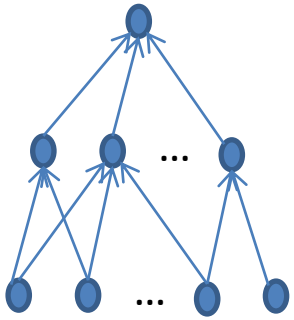
1986

1995

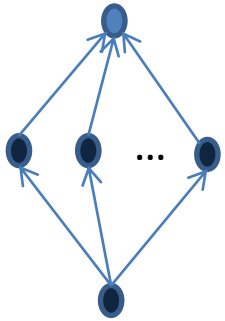
2006



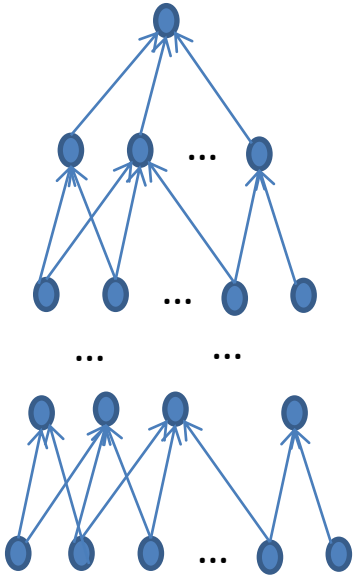
Two layer liner model



Three layer non-linear model



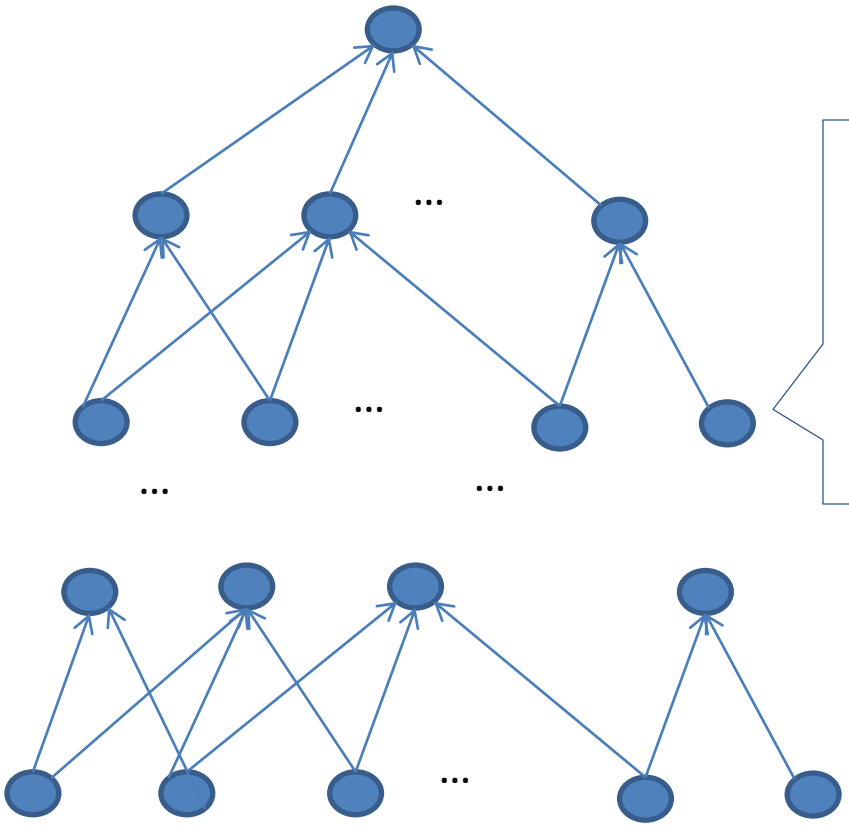
Kernel SVM =
Three layer non-linear model

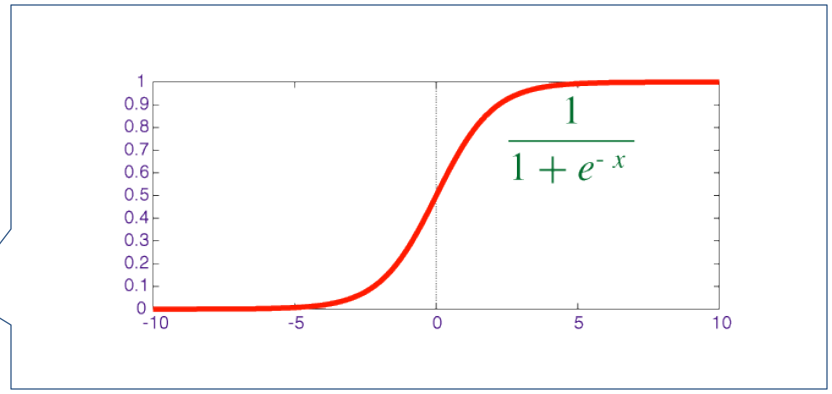


Multi-layer non-linear model

Deep learning means learning of complicated non-linear models

Deep Neural Network





Deep learning is to learn multi-layer non-linear models from data

Four Questions to Answer

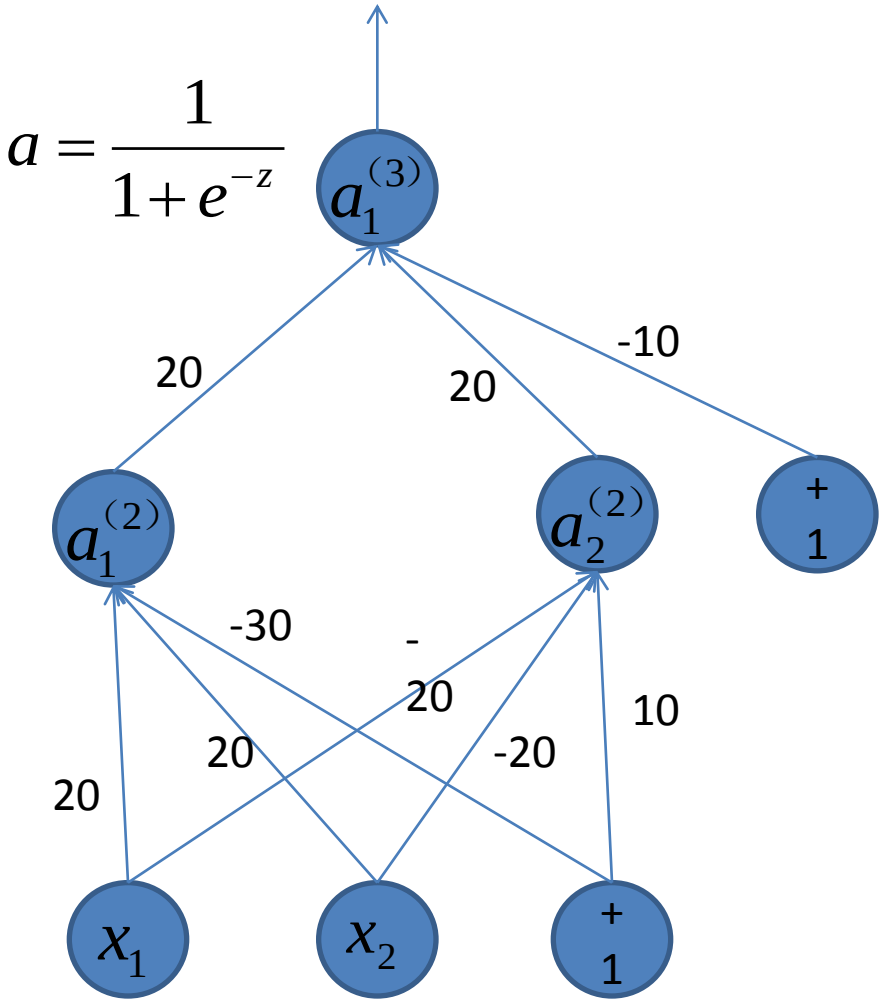
- Why is deep learning powerful?
- Why is deep learning popular now?
- Is deep learning almighty?
- How similar/dissimilar are deep neural networks from human brain?

Question: Why Is Deep Learning Powerful?

- *Powerful Representation Learning Ability*
- Deep network = complicated non-linear function
- Examples:
  - XOR classification problem, cannot be modeled by linear function
  - Simple non-linear classification problem, can be modeled by Kernel SVM
  - Complicated non-linear classification problem, can be modeled by *deep* Convolutional Neural Network

Example: XNOR Network

XNOR function can be represented by three layer network

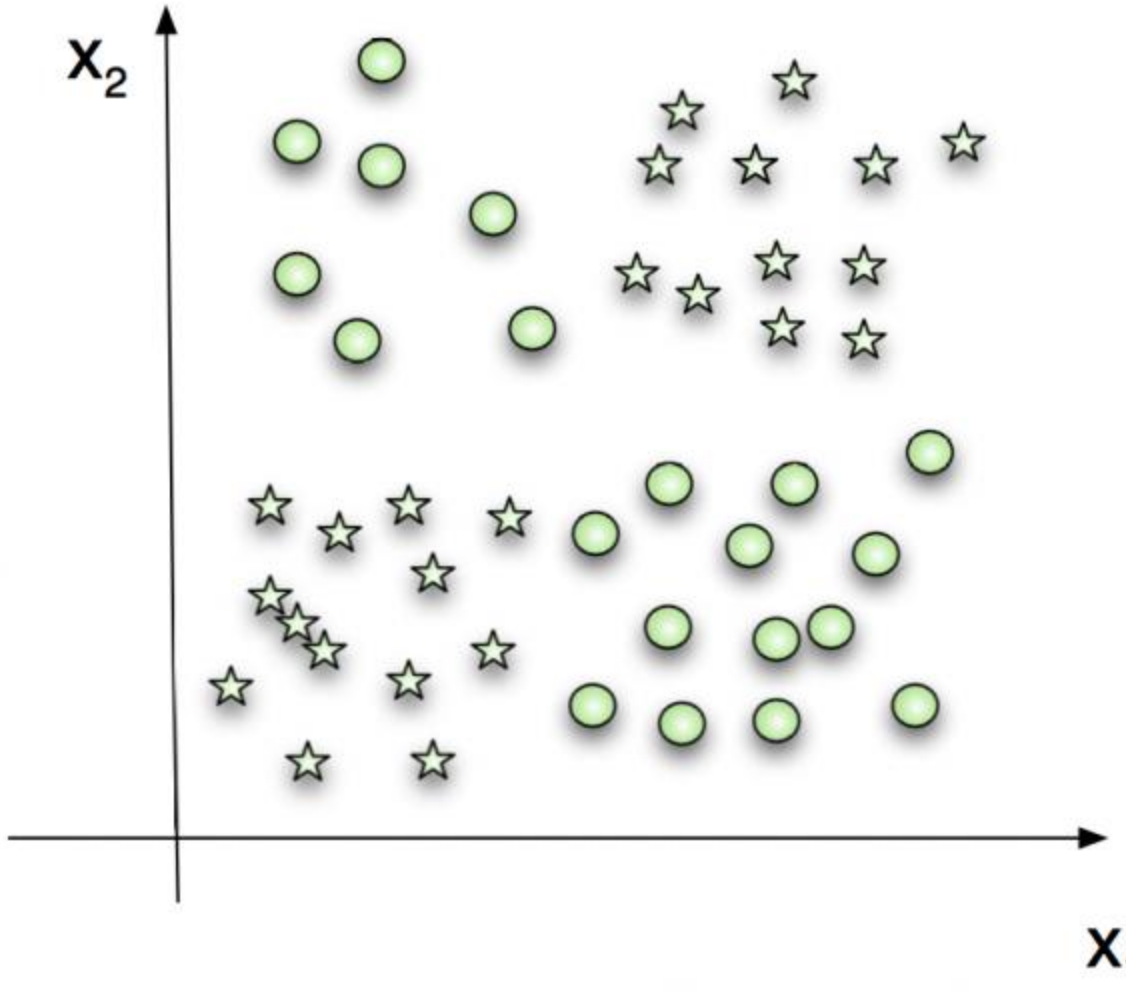


XNOR

x_1	x_2	$a_1^{(2)}$	$a_2^{(2)}$	$a_1^{(3)}$
0	0	0	1	1
0	1	0	0	0
1	0	0	0	0
1	1	1	0	1

Example from Andrew Ng

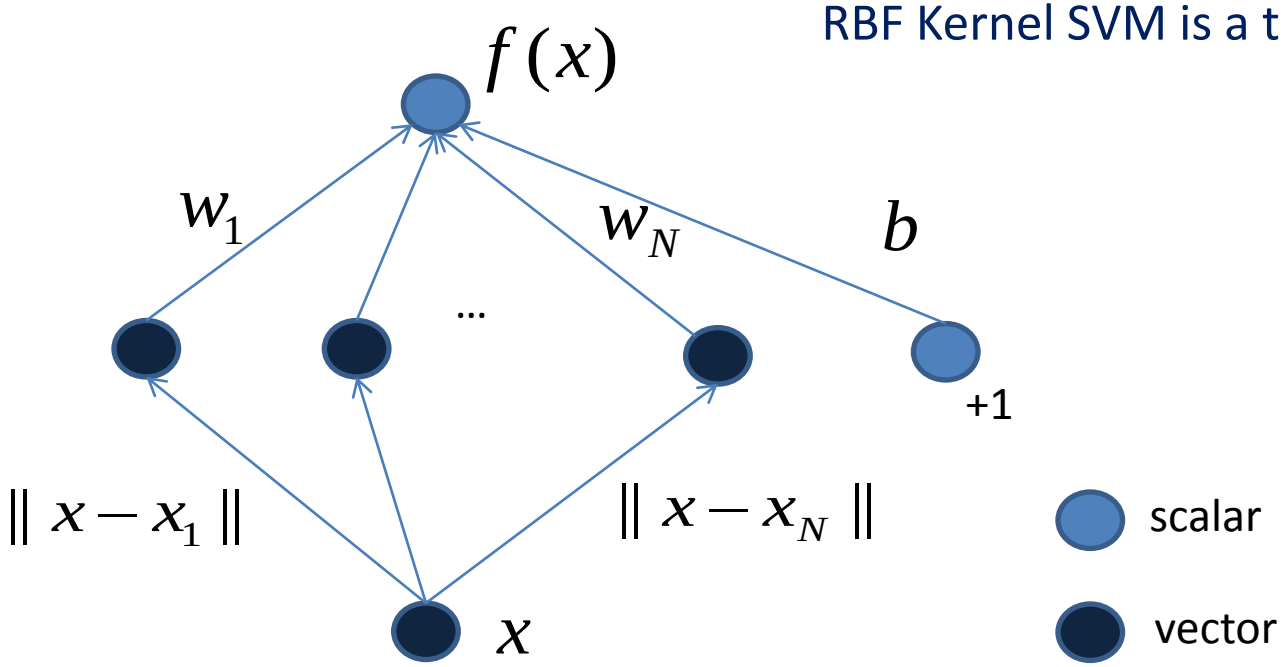
Example: XOR/XNOR Classification Problem



XOR / XNOR classification problem: linear model cannot deal with; multi-layer neural network can

Example: RBF Kernel SVM

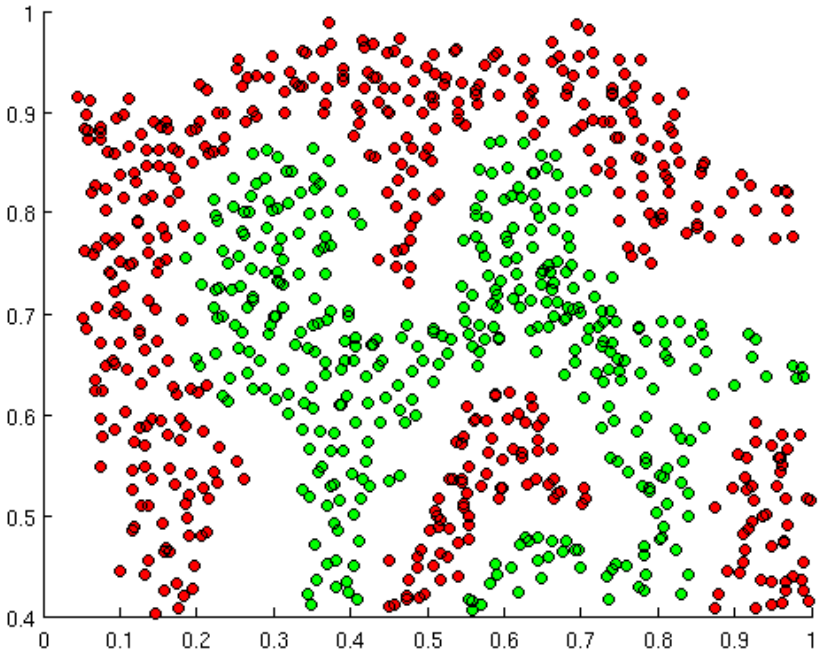
RBF Kernel SVM is a three layer network

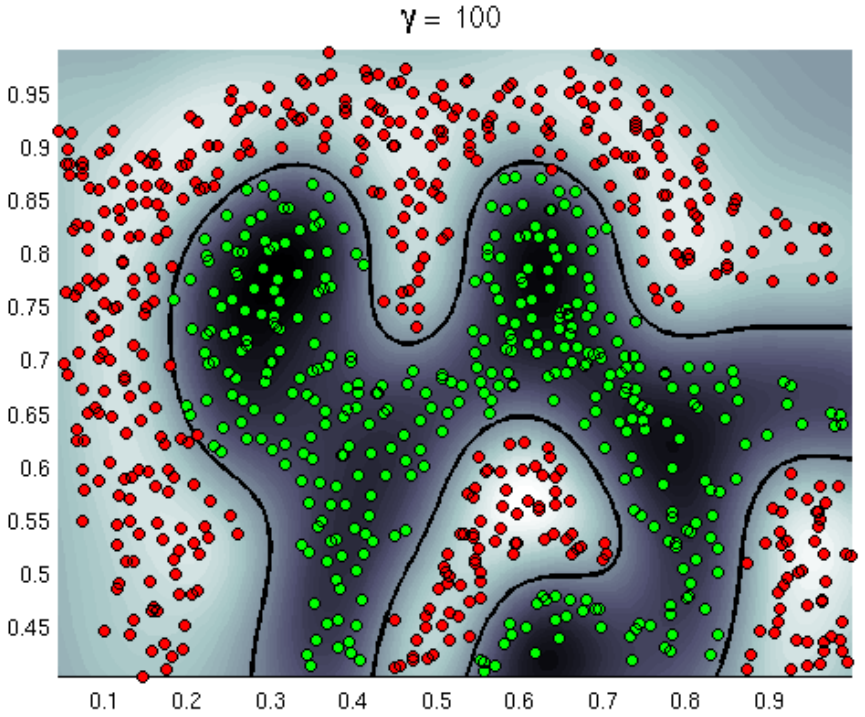


$$f(x) = \sum_i \alpha_i y_i \exp(\gamma \| x - x_i \|^2) + b$$

$$= \sum_i w_i \exp(\gamma \| x - x_i \|^2) + b$$

Example: Complicated Non-Linear Classification Problem



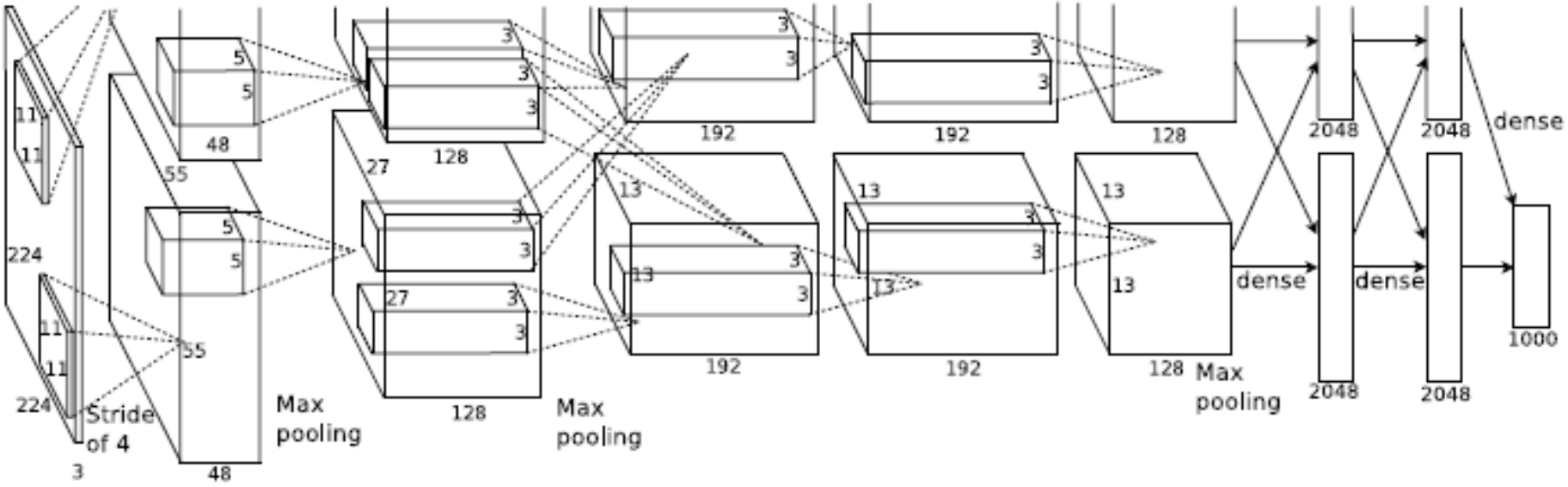


$$f(x) = \sum_{i=1}^{N} \alpha_i y_i \exp(\gamma \| x - x_i \|^2) + b$$

RBF SVM can effectively separate positive and negative examples

Example from Andrew Ng

Example: "Alex Net" (A Convotional Neural Network for Image Classification)



11 layers; 650,000 neurons; 60 million parameters
5 convolution layers, 3 fully connected layers, 3 max-pooling layers

Krizhevsky et al, 2012

Example: Alex Net for ImageNet Challenge

ILSVRC 2012: Classifying 1.2 million images into 1000 classes

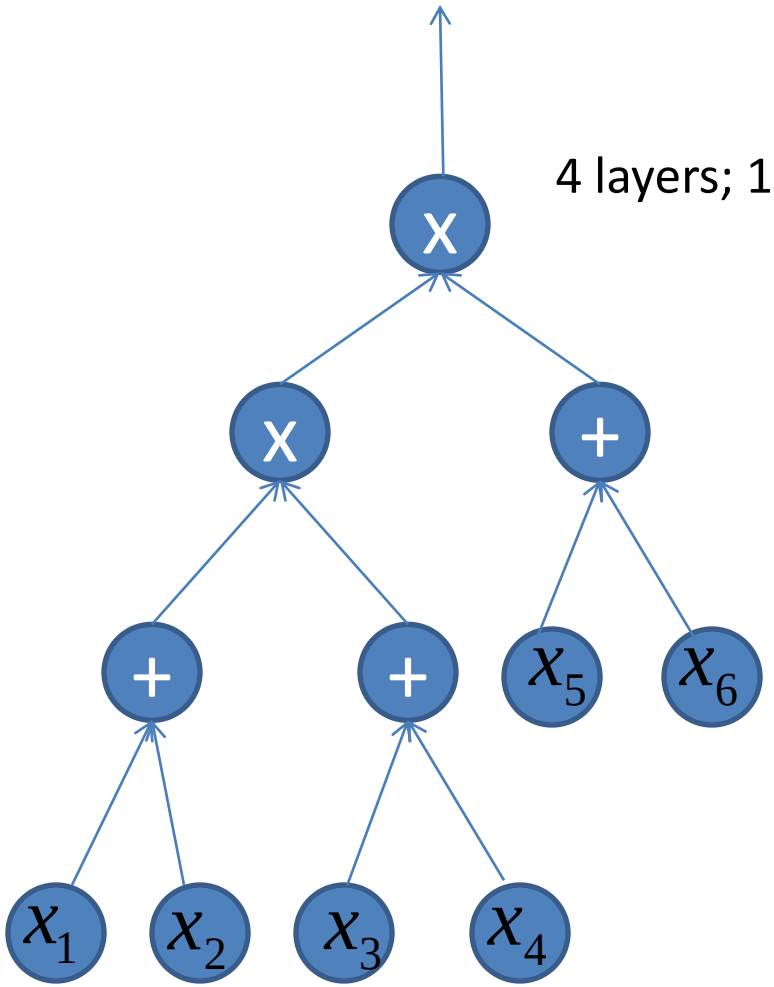
Top 5 error rate: 15.2% v.s. 26.2% (second place)

Question: Why Is Deep Learning Powerful?

- *Statistical Efficiency*
- Deep models have higher statistical efficiency, i.e., need less data to train
- Functions can be more compactly represented with deep networks than shallow networks
- Example:
  - Sum-Product Network

Example: Deep Sum-Product Network

$$((x_1 + x_2)(x_3 + x_4))(x_5 + x_6)$$

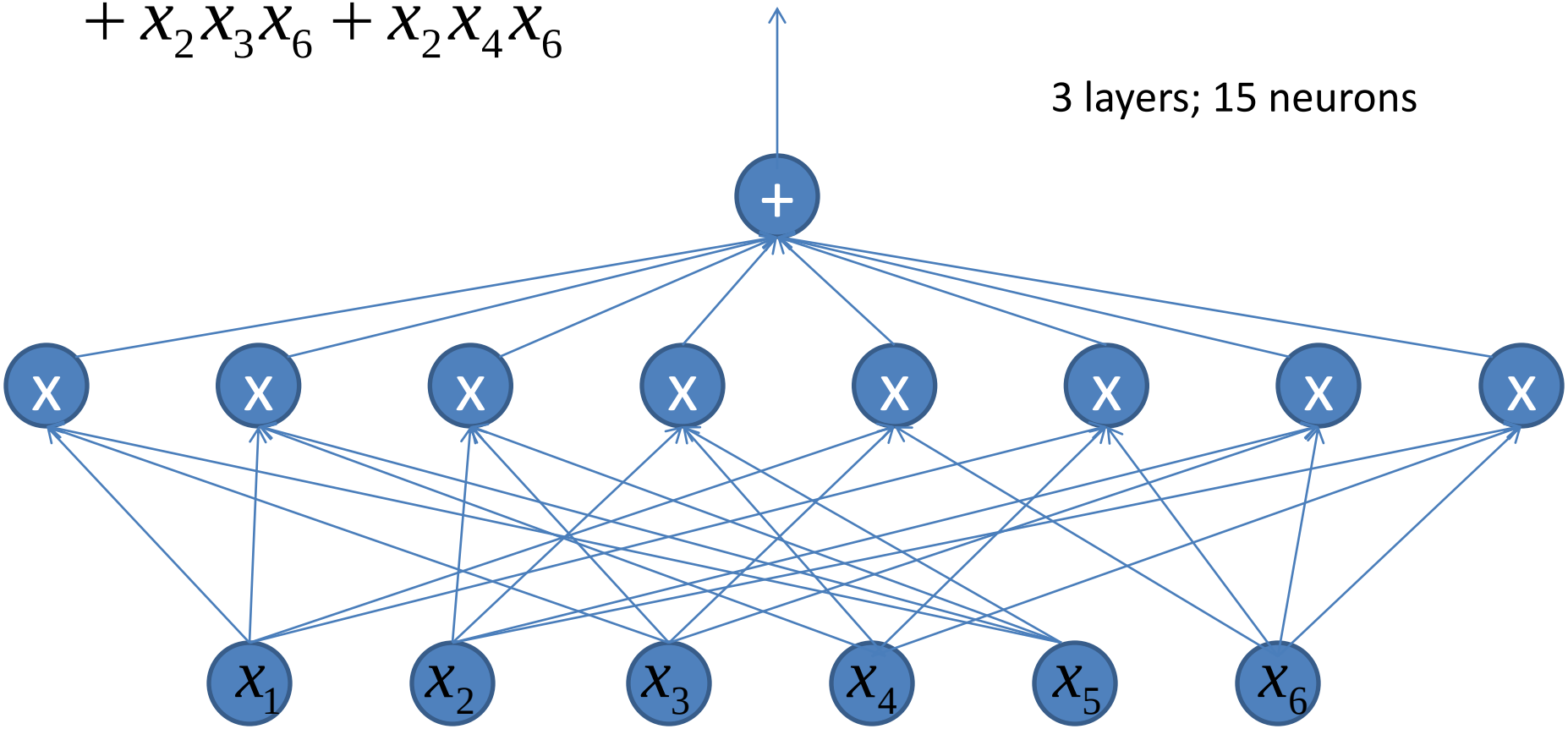


Deep network has less parameters and thus higher statistical efficiency.

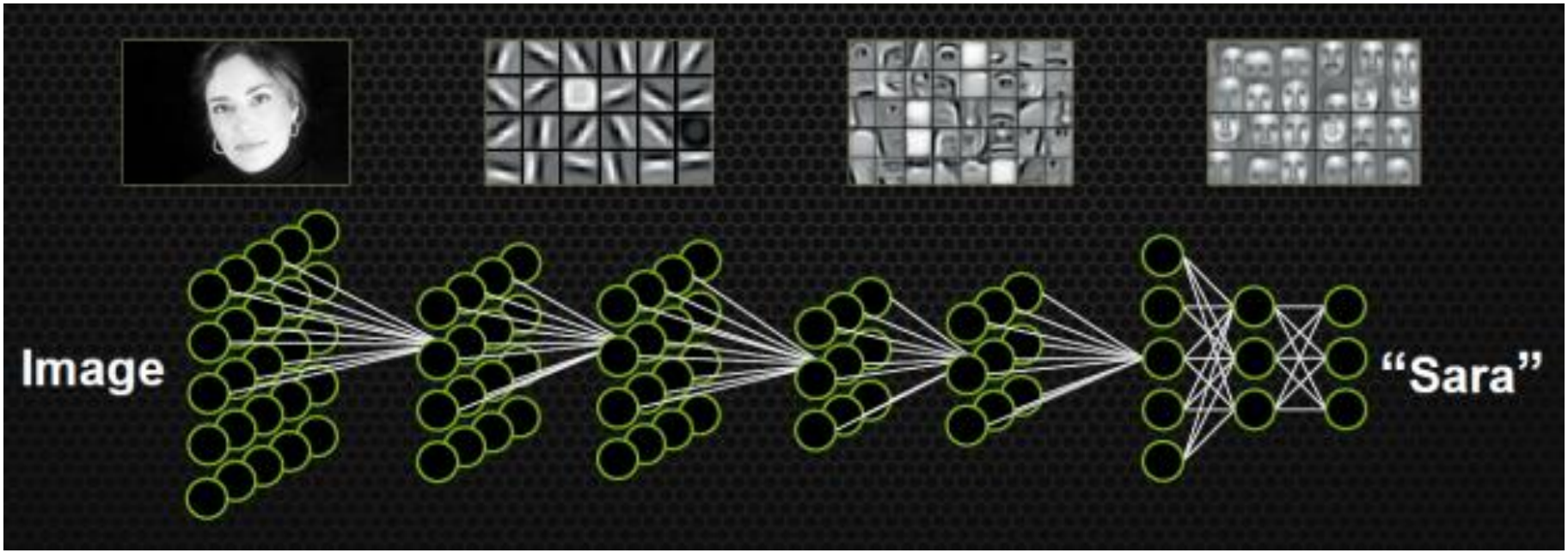
Example: Shallow "Product-Sum Network"

$$x_1x_3x_5 + x_1x_4x_5 + x_2x_3x_5 + x_2x_4x_5 + x_1x_3x_6 + x_1x_4x_6$$
$$+ x_2x_3x_6 + x_2x_4x_6$$

3 layers; 15 neurons



Conclusion: Deep Learning Has Ability of Dealing with Complicated Problems



Picture from Nvidia.com

Why Is Deep Learning Popular Now?

- Computers become more powerful
- More data is available
  - Number of parameters vs size of training data
  - Parametric vs non-parametric
- Simple problems have been solved

Computers Are More Powerful

Geoffrey Hinton

- Neural networks did not beat existing technologies 20-30 years ago
- Now we know the reason; the machines at the time were not fast enough

Interview with Google's AI and Deep Learning 'Godfather' Geoffrey Hinton, 2015/7

More Parameters Model Has, More Training Data Is Needed

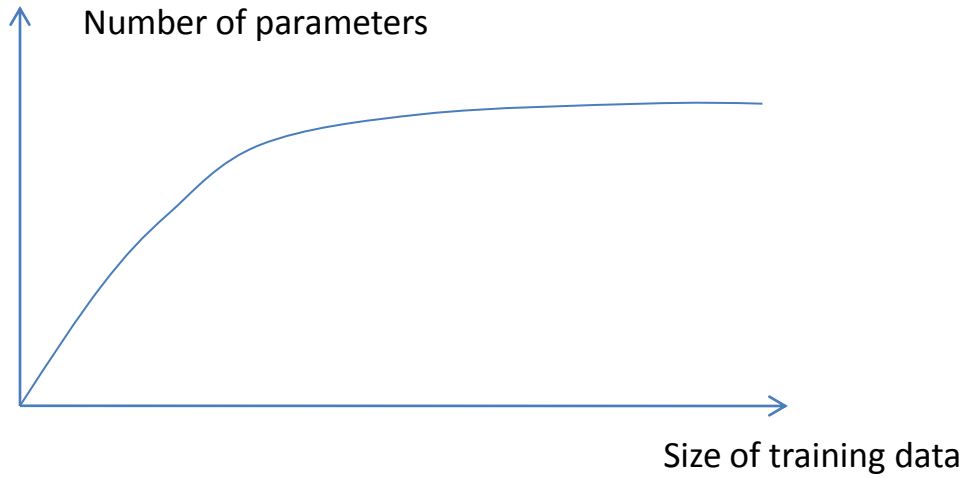
- Sample complexity
  - Occam's Razor Theorem
  - For binary classification

$$|S| \geq \frac{1}{2\varepsilon^2}(\ln 2 \cdot |h| + \ln \frac{2}{\delta}) \text{ for all } h \in C$$

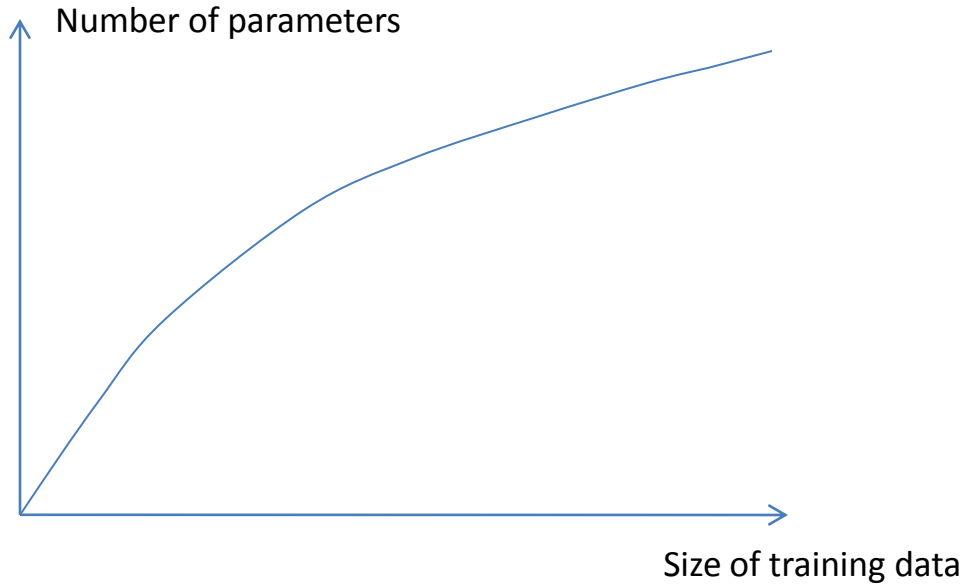
ε : precision, $1-\delta$: confidence, $|h|$: model complexity

- Empirical Sample Complexity
  - Empirically at least 100 times of number of parameters

Parametric v.s. Non-Parametric

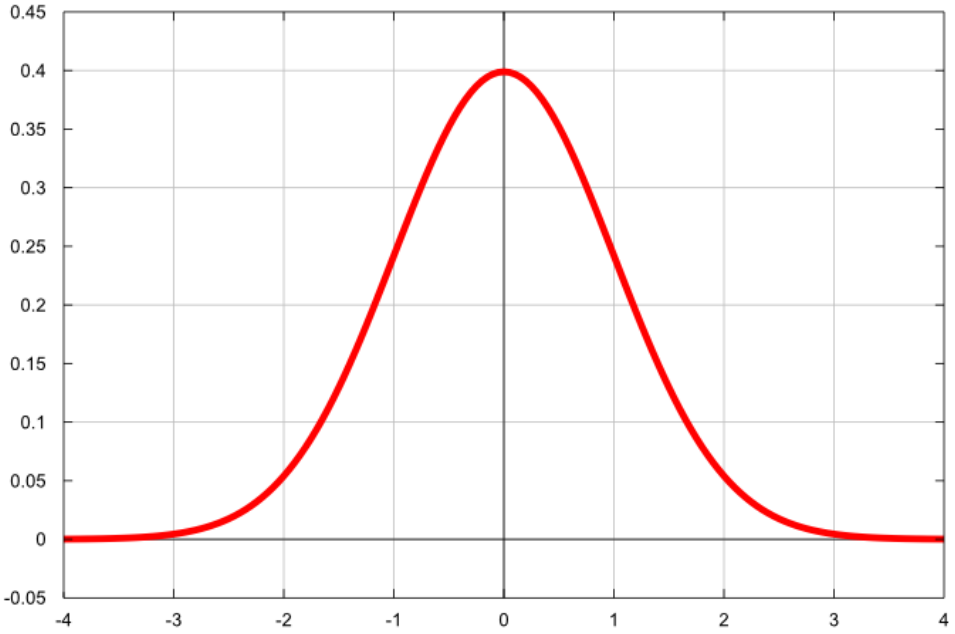


Parametric problem: number of parameters does not further increase, when size of data exceeds a threshold



Non-parametric problem: number of parameters continuously increases, when size of data increases

Example: Gaussian Distribution



$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp - \frac{(x-u)^2}{2\sigma^2}$$

$$\hat{u} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

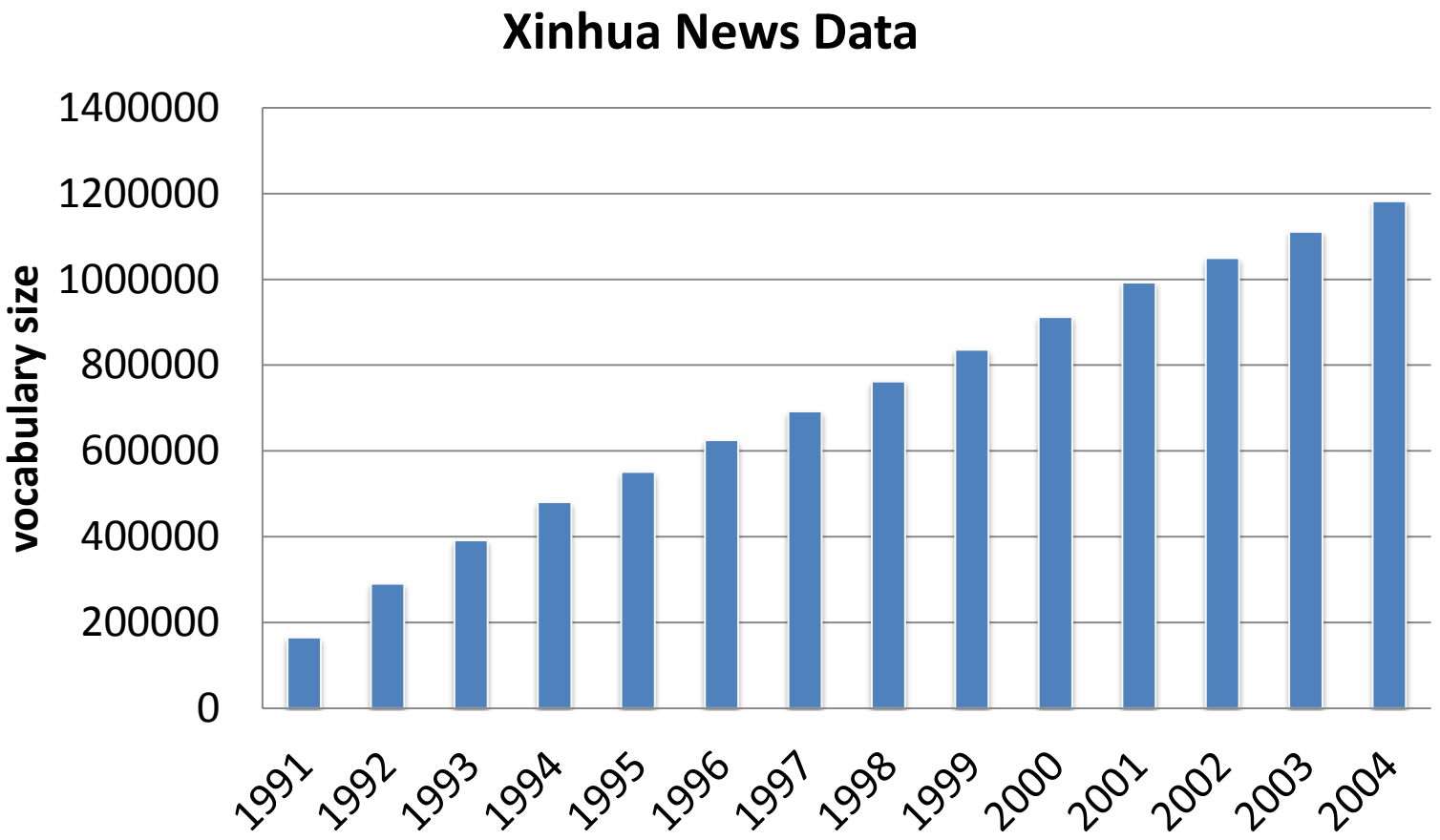
$$\hat{\sigma}^2 = \frac{1}{N}\sum_{i=1}^{N} (x_i - \hat{u})^2$$

$$\text{precision} : O(\frac{1}{\sqrt{N}})$$

Parametric problems can be easily handled with big data

Example: Vocabulary Size in Text Data

Non-parametric problem: need to continuously increase model complexity



Vocabulary size increases when data size increases

Data from Xiao Chen

Question: Is Deep Learning Almighty?

- Suitable for
  - Complicated non-linear problem, *and*
  - Big data
- No-free lunch theorem
  - Interpretation: no algorithm can *always* work better than other algorithms

"No Free Lunch" Theorem
(Wolpert & Macready, 1997)

$Acc(L)$ = generalization accuracy of learner L

= accuracy on test examples

Theorem: for any learners L_1, L_2

if $\exists \text{learning problem, s.t. } Acc(L_1) > Acc(L_2)$

then $\exists \text{learning problem, s.t. } Acc(L_2) > Acc(L_1)$

Conclusion: Using Right Tools to Do Right Things

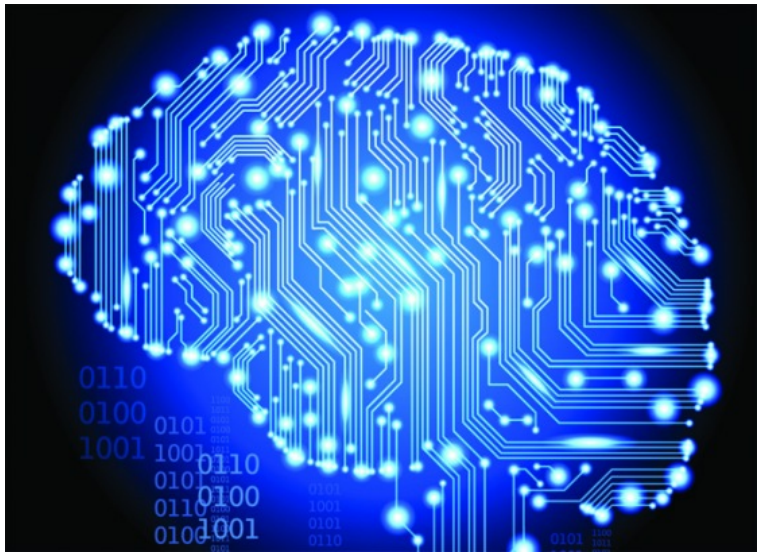
- Simple problems can be solved by using simple models (linear or non-linear)
- Deep learning is more suitable for complicated non-linear problems with large amount of data

Question: How Similar/Dissimilar from Human Brain?

- Brain and computer = “automatas” – Von Neumann
- “Artificial” neural networks are machine learning models, inspired by human brains, *but motivated by problem solving*
- Similarity
  - Elements: e.g., neurons, synapses
  - Architecture: e.g., deep cascaded structure, local receptive field
  - Signal: e.g., digital and analogue signals
- Dissimilarity
  - Mechanism of learning, e.g., back propagation

Talk Outline

- Overview of Deep Learning
- *Recent Advances in Research of Deep Learning*
- Future of Deep Learning

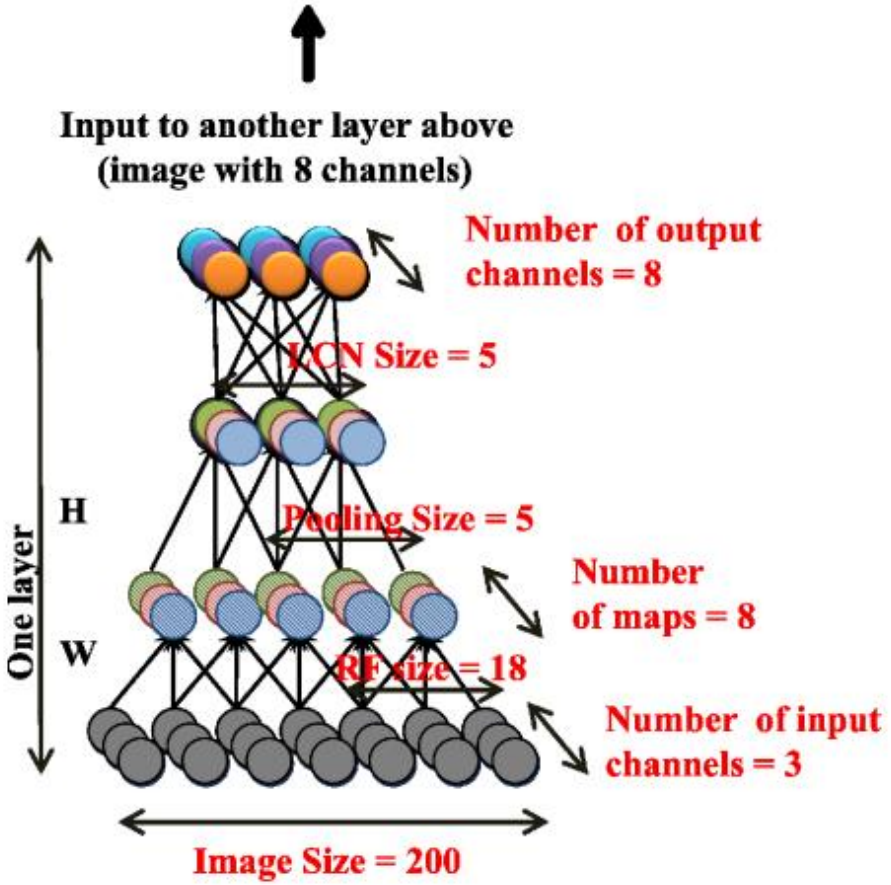


Recent Advances of Deep Learning Research

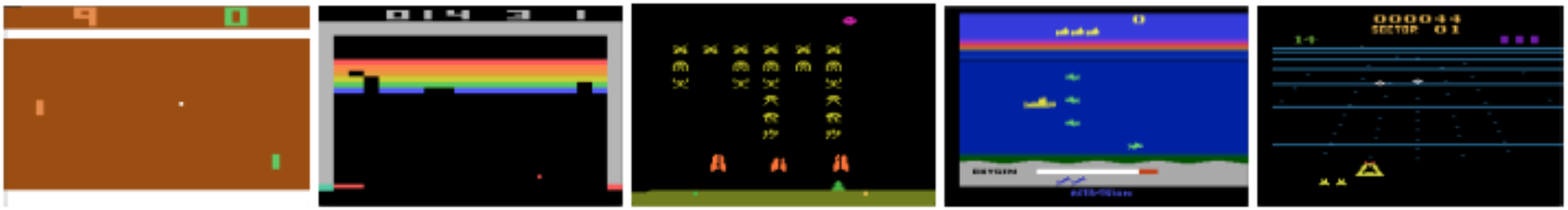
- Restricted Boltzmann Machine (Hinton 2006)
- Alex Net: Deep Convolutional Network for Image Classification (Krizhevsky et al, 2012)
- Deep Neural Network for Speech (Dahl et al, 2012)
- *Unsupervised Learning of High Level Features Using Auto Encoder (Le et al, 2012) @Google*
- *Atari Player: Deep Reinforcement Learning (Mnih et al, 2013) @Google Mind*
- *Neural Turing Machine (Graves et al, 2014) @Google Mind*
- *Memory Networks (Weston et al, 2014) @Facebook*

Unsupervised Learning of High Level Features Using Auto Encoder (Le et al, 2012)

- Nine layers: replicating same stage three times
- Each stage: filtering, pooling, and contrast normalization
- Training: learning weights of auto encoding and decoding
- 10 million images
- 1 billion parameters
- 1000 machines (1600 cores), trained on 3 days
- Can learn the feature for detecting “cat”



Atari Player: Deep Reinforcement Learning (Mnih et al, 2013)

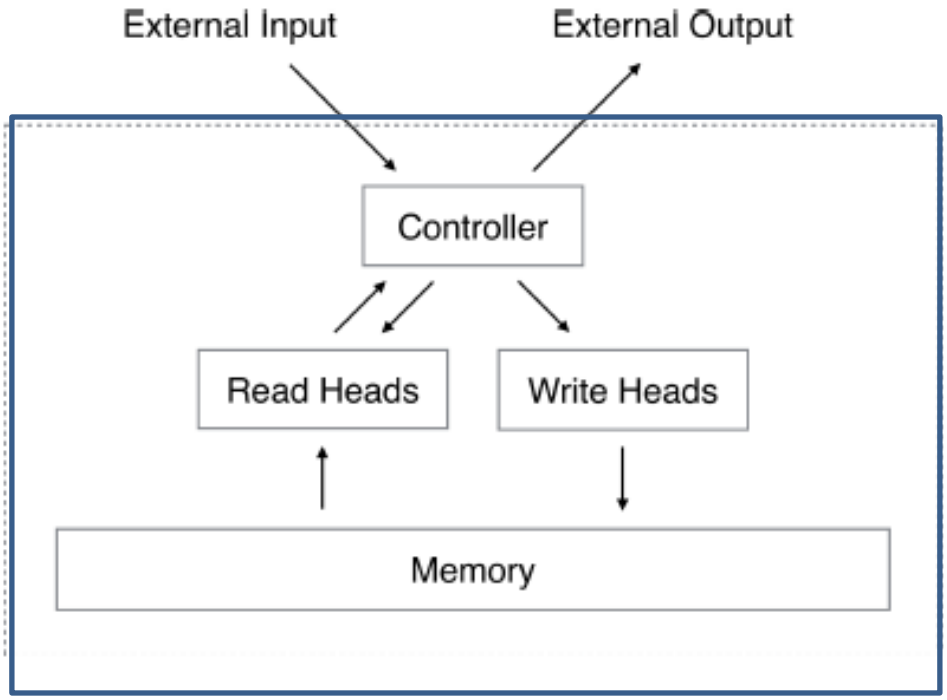


- Learning to control in dynamic and uncertain environment
- Example: Atari games
- Deep reinforcement learning system beats human experts on three out of six games

- Formalized as Q-learning
- Input: raw pixels
- Action: game action
- Reward: game score
- Using convolutional neural network to model Q-function

Neural Turing Machine
(Graves et al, 2014)

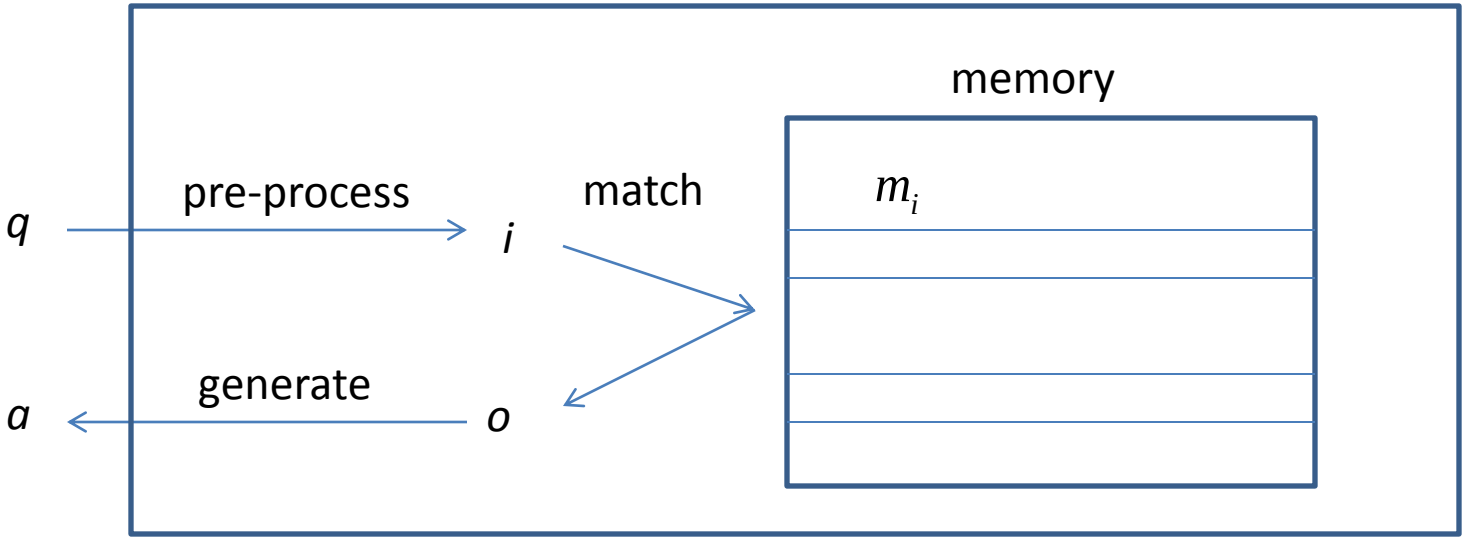
- Controller = Neural Network, with input, output, and memory
- Memory has read and writing operations
- All components are differentiable and thus trainable by gradient descent
- Neural Network = DNN or LSTM
- Tasks such as copy, sort can be performed



Memory Networks
(Weston et al, 2014)

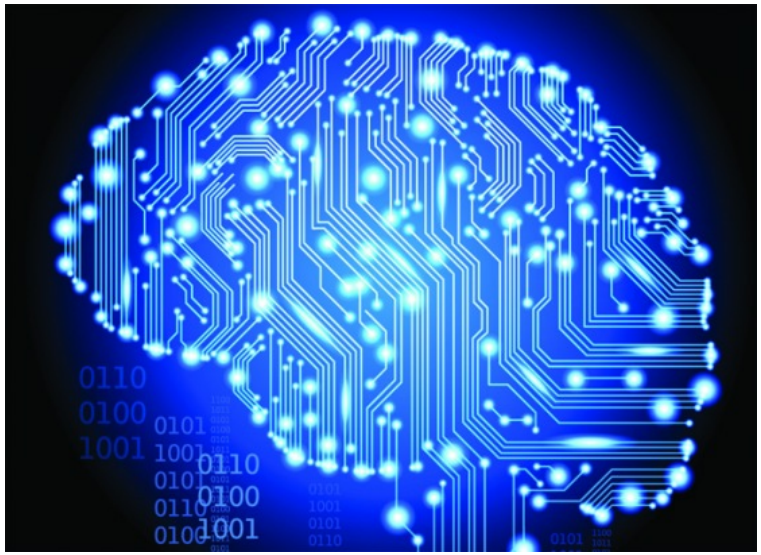
- Long term memory + inference
- Model is learned
- Can answer factoid questions
- Acc = 40%+

- Example
  - John is in the playground.
  - Bob is in the office.
  - John picked up the football.
  - Bob went to the kitchen.
  - Q: where is the football?
  - A: playground



Talk Outline

- Overview of Deep Learning
- Recent Advances in Research of Deep Learning
- *Future of Deep Learning*



Future of Deep Learning
- Beyond Complicated Pattern Recognition?

- More complicated tasks
  - E.g., multi-turn dialogue
- More effective use of human knowledge
  - E.g., question answering
- Combination with reasoning
  - E.g., question answering

References

- Hinton, G. E., & Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. *Science*, *313*(5786), 504-507.
- Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems* (pp. 1097-1105).
- Dahl, G. E., Yu, D., Deng, L., & Acero, A. (2012). Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition. *Audio, Speech, and Language Processing, IEEE Transactions on*, *20*(1), 30-42.
- Le, Q. V. (2013, May). Building high-level features using large scale unsupervised learning. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on* (pp. 8595-8598). IEEE.
- Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., & Riedmiller, M. (2013). Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*.
- Graves, A., Wayne, G., & Danihelka, I. (2014). Neural Turing Machines. *arXiv preprint arXiv:1410.5401*.
- Weston, J., Chopra, S., & Bordes, A. (2014). Memory networks. *arXiv preprint arXiv:1410.3916*.

Thank you!

hangli.hl@huawei.com