Introduction to Machine Learning

Lecture 12 - Introduction to Deep Learning

Guang Bing Yang, PhD

Deep Learning in General

- Introduction to Deep Learning in General
- Deep Neural Network (DNN)
- Deep Feedforward Networks (MLPs)
- Regularization in Deep Learning

What is Deep Learning

- The solution that allows computers to learn from experience and understand hierarchy of complicated concepts which are built on top of simpler ones. This hierarchy is deep with many layers, thus this approach is called AI **deep learning** [1]
- Computers need to capture knowledge
- No knowledge base projects were successful into AI
- AI needs different ways to acquire their own knowledge by extracting patterns from data
- The capability is known as **machine learning**

[1] Ian Goodfellow and Yoshua Bengio and Aaron Courville, Deep Learning, MIT Press, 2016

Design & Extract Features

- In classical machine learning, designing the right set of features to extract and apply them to a simple machine learning algorithm are common AI tasks.
- However, many limitations exist
  - hard to know which features to take
  - expensive manual work
  - need domain experts
  - hard to reflect interaction between data
  - high bias & big noise even errors

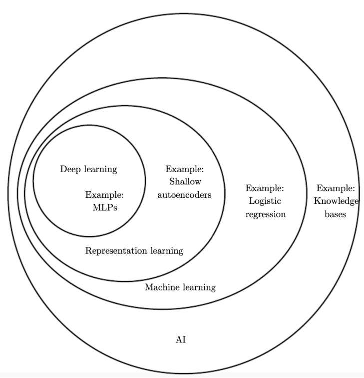
Approaches

- Representation learning is one approach.
  - much better than handcraft representation
  - e.g., autoencoder
- Factors of variation — designing features or algorithms for learning features to determine varieties in observed data.
  - if factors change, the behaviour of the data will change
  - features are independent
- Deep Learning — applies multiple simpler representation learning to complete complex representation learning
  - learn right representation
  - breaking the desired complicated mapping to many simple ones.

A Type of Machine Learning

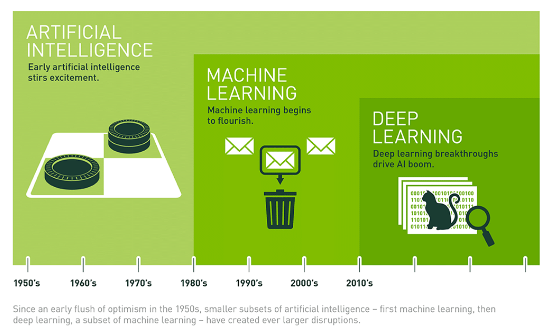
- Deep learning is a kind of machine learning, It is:
  - a kind of supervised or unsupervised learning problems
  - with great power and flexibility
  - learning to represent the world as a nested hierarchy of concepts,
  - a concept defined in relation to simpler concepts
  - about abstract representations computed in terms of less abstract ones.

Relationships in AI Fields



Cited from Ian Goodfellow and Yoshua Bengio and Aaron Courville, Deep Learning, MIT Press, 2016

Relationships in AI Fields



The difference between AI, Machine Learning, and Deep Learning. Cited from NVIDIA Blog

Historical trends in Deep Learning

- Deep learning has had a long and rich history, but has gone by many names reflecting different philosophical viewpoints, and has waxed and waned in popularity.
- Deep learning has become more useful as the amount of available training data has increased.
- Deep learning models have grown in size over time as computer infrastructure (both hardware and software) for deep learning has improved.
- Deep learning has solved increasingly complicated applications with increasing accuracy over time.

yguangbing@gmail.com, Guang.B@chula.ac.th Apr 16th, 2021

Historical Trends in Deep Learning

- Deep learning can date back to the 1940s. It has gone through many different names till recently the "deep learning".
- Three waves:
  - Cybernetics ~ 1940s - 1960s,
  - Connectionism ~ 1980s - 1990s, cognitive scientists
  - Deep learning ~ 2006, deep neural networks by Hinton
- Names:
  - Artificial neural networks (ANNs)
  - Modern term deep learning goes beyond the neuroscientific perspective

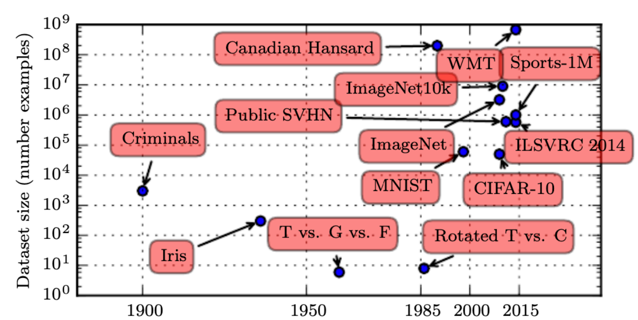
yguangbing@gmail.com, Guang.B@chula.ac.th Apr 16th, 2021

Historical Trends in Deep Learning

- Why deep learning has only recently become recognized as a crucial technology after the first experiments with artificial neural networks were conducted in the 1950s?
  - More efficient deep learning algorithms
  - The age of "Big Data" has made machine learning much easier
  - More powerful hardware and software
- As of 2016, a rough rule of thumb is that a supervised deep learning algorithm will generally achieve acceptable performance with around 5,000 labeled examples per category, and will match or exceed human performance when trained with a dataset containing at least 10 million labeled examples.

yguangbing@gmail.com, Guang.B@chula.ac.th Apr 16th, 2021

Historical Trends in Deep Learning



Cited from Ian Goodfellow and Yoshua Bengio and Aaron Courville, Deep Learning, MIT Press, 2016

yguangbing@gmail.com, Guang.B@chula.ac.th Apr 16th, 2021

Historical Trends in Deep Learning

- Increasing model sizes
  - have the computational resources to run much larger models today.
  - Unless new technologies allow faster scaling, artificial neural networks will not have the same number of neurons as the human brain until at least the 2050s.
- Increasing Accuracy, Complexity and Real-World Impact
  - Since the 1980s, deep learning has consistently improved in its ability to provide accurate recognition or prediction
  - For example, Recurrent neural networks, such as the LSTM, CNN, DQN-RL, etc.

yguangbing@gmail.com, Guang.B@chula.ac.th Apr 16th, 2021

Historical Trends in Deep Learning

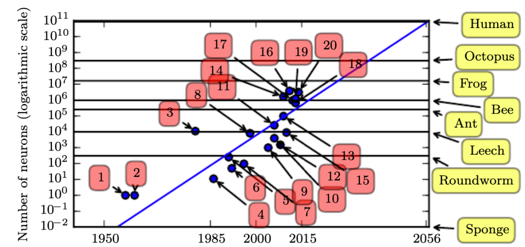


Figure 1.11: Since the introduction of hidden units, artificial neural networks have doubled in size roughly every 2.4 years. Biological neural network sizes from Wikipedia (2015).

yguangbing@gmail.com, Guang.B@chula.ac.th Apr 16th, 2021 © GuangBing Yang, 2021. All rights reserved.

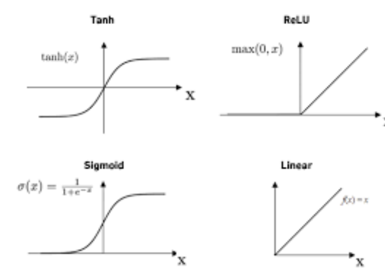
Historical Trends in Deep Learning

- Deep Learning is widely applied in modern industries:
  - Used by many top technology companies including Google, Microsoft, Facebook, IBM, Baidu, Apple, Adobe, Netflix, NVIDIA and NEC.
- Advances in software infrastructure:
  - Theano
  - Tensorflow
  - PyTorch
  - Torch
  - Caffe
  - MXNet
  - PyLearn2

yguangbing@gmail.com, Guang.B@chula.ac.th Apr 16th, 2021 © GuangBing Yang, 2021. All rights reserved.

Types of Neurons

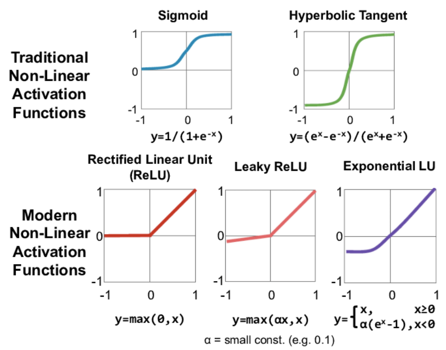
- Linear neuron — the linear relationship between input and output
  - $y = wx + b$
- Sigmoid neuron — the output in [0, 1]
- Tanh neuron — the output in [-1, +1], a kind of Sigmoid
- ReLU (Rectified Linear Unit)
  - $y = max(0,x)$



Common types of artificial neurons

yguangbing@gmail.com, Guang.B@chula.ac.th Apr 16th, 2021 © GuangBing Yang, 2021. All rights reserved.

Variances of Non-linear Neurons

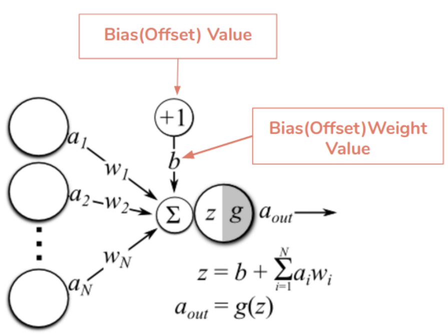


- Leaky ReLU
  - $y = max(\alpha x, x), \ \ 0 < \alpha << 1$
- Exponential ReLU
  - $y = x, \text{ if } \ x \geq 0$
 $y = \alpha(e^x - 1), \text{ if } \ x < 0$

Other Activation Functions

- Maxout function:
  - $f(x) = \max_k(a_k)$
  - derivaties: $\dfrac{\partial f(x)}{\partial a_i} = \begin{cases} 1, & \text{maximum } \ a_i \\ 0, & \text{otherwise} \end{cases}$
- Softmax function:
  - $f(z_j) = \dfrac{e^{z_j}}{\sum_{i=1}^{K} e^{z_i}}$, and its derivatives: $\dfrac{\partial f(x_j)}{\partial x_i} = f(x_i)(\delta_{ij} - f(x_j))$, where, $\delta_{ij} = 1$, if $i = j$, else, $\delta_{ij} = 0$.

An Illustration of Neuron Operation



Cited from Medium

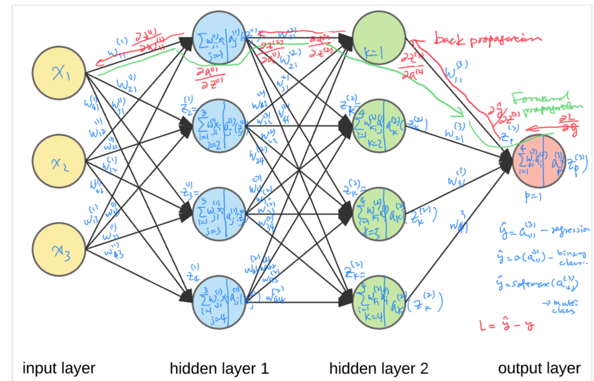
Deep Neural Network (DNN)

- DNN is a multilayer perceptron—a class of feedforward artificial neural network with back-propagation to solve the non-linear XOR problem.
- It consists of input layer, hidden layer(s), and output layer.
- Its objective function—loss function—is selected based on the type of the problems.
  - Mean square error for regression
  - Cross-entropy for multiple classification and binary classification

Deep Neural Network (DNN)

- Forward propagation
  - start from input layer, through every hidden layer, then to output layer to get the output of the loss function
- Backward propagation
  - start from the return of the loss function backward to input layer to calculate the gradients over each parameters
  - update parameters W and b repeatedly till converged or out of the loops or the change of loss lower than threshold.

Deep Neural Network (DNN)



DNN-Forward-Backward propagations. ©GuangBing Yang, 2021. All rights reserved.

Deep Feedforward Networks (MLPs)

- Deep feedforward networks, also often called feedforward neural networks, or multilayer perceptrons (MLPs), are the basic deep learning models
- A feedforward network defines a mapping $y = f(x; \theta)$ and learns the optimized value of the parameters θ.
- These models are called feedforward because information flows through the function being evaluated from x, through the intermediate computations used to define f, and finally to the output y.
- The model can be represented as a DAG: $f(x) = f^{(3)}(f^{(2)}(f^{(1)}(x)))$

Deep Feedforward Networks (MLPs)

- The model can be represented as a DAG: $f(x) = f^{(3)}(f^{(2)}(f^{(1)}(x)))$
- These chain structures are the most commonly used structures of neural networks.
- In this case, $f^{(1)}$ is called the **first layer** of the network (also called Input layer), $f^{(2)}$ is called the second layer (one hidden layer, normally), and so on.
- The last layer of this chain is called **output layer**.
- The overall length of the chain is the **depth** of the model (or networks)
- All layers between the first layer (input layer) and the output layer (the final layer) are **hidden layers.**
- The dimensionality of these hidden layers determines the **width** of the networks (model).

Deep Feedforward Networks (MLPs)

- Recall in classical machine learning, the basis function $\phi(x)$ is given manually or arbitrarily from domain expects, or assigned using kernel functions
- However, deep learning learns these basis functions $\phi(x)$ from data as well as the parameters θ.
- Hence, the strategy of deep learning is to learn ϕ, not just θ.
- In this approach, the model can be described as: $y = f(x; \theta, w) = \varphi(x; \theta)w$.
- Now, learning includes θ and ϕ, which is given as kernel functions, arbitrarily given as basis functions.

yguangbing@gmail.com, Guang.B@chula.ac.th

Regularization for Deep Learning

- Generalization is the central problem in machine learning.
- This problem is even seriously in the deep learning problems.
- Many strategies used in machine learning are explicitly designed to reduce the test error at the expense of increased training error.
- These strategies are called regularization.
- In deep learning, these strategies are based on regularizing estimators by trading increased bias for reduced variance.

yguangbing@gmail.com, Guang.B@chula.ac.th

Regularization for Deep Learning

- However, increasing bias can bring different issues.
- An effective regularizer is one that reduces variance significantly while not overly increases the bias.
- Here lists most of regularization strategies:
  - parameter norm penalties — by adding a parameter norm penalty $\Omega(\theta)$ to the objective function J: $\tilde{J}(\theta; X, y) = J(\theta; X, y) + \alpha\Omega(\theta)$, where $\alpha \in [0,\infty)$ is a hyperparameter that eights the relative contribution of the norm penalty term.
    - This strategy includes L2 (weight decay) and L1 (LASSO)

yguangbing@gmail.com, Guang.B@chula.ac.th

Regularization for Deep Learning

- Here lists most of regularization strategies:
  - Parameter norm penalties — L2 and L1
  - dataset augmentation — use fake data in training, add noise in parameters or data, add noise in hidden units, like dropout.
  - Early stopping — obtain a model with better validation set error by returning the parameter setting with the lowest validation set error during training.
  - Parameter tying and sharing— for example, CNN
  - Bagging and ensemble methods
  - Dropout

yguangbing@gmail.com, Guang.B@chula.ac.th

Summary

- Deep learning is an approach to machine learning that has drawn heavily on our knowledge of the human brain
- statistics and applied math as it developed over the past several decades
- It has seen tremendous growth in its popularity and usefulness due in large part to more powerful computers, larger datasets and techniques to train deeper networks
- The years ahead are full of challenges and opportunities to improve deep learning even further and bring it to new frontiers
- Deep Neural Network (DNN) is the basic architecture in Deep Learning
- It consists of a framework for forward propagation (a multilayer perceptron) and back-propagation

Summary

- There are many different neurons, linear neuron, sigmoid, tang, softmax, mahout, etc.
- ReLU is the most popular activation function due to its non-linearity and simplest computational properties
- Big data, advanced software, powerful hardware and computation power make the Deep learning more popular and realistic
- Deep Feedforward Networks (MLPs) are the basic deep neural networks
- In deep learning, not only does the model learn the optimized parameters, but also learn the basis functions
- Regularizations are more seriously in deep learning
- There are many different approaches of regularization in deep learning

Questions & lab?