ECSE 4965/6965

Spring, 2018

Introduction to Deep Learning

Qiang Ji

What is this class about?

Deep Learning

One of the most exciting technical developments in
Machine Learning and
Artificial Intelligence
in last decade
It has significantly advanced several fields, including

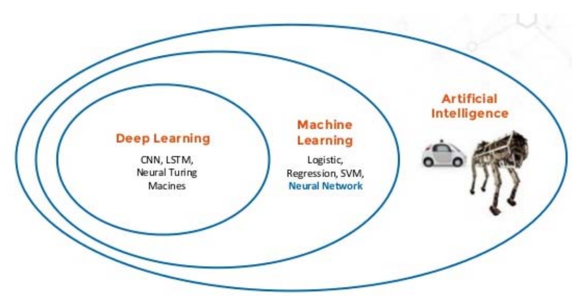
Speech Recognition, Machine Translation, Computer Vision, Natural Language Processing, Robotics, Games, Self-Driving, and **…**

Instead of programming computers, teaching computer
by showing examples and computer learns
automatically from data

Deep Learning

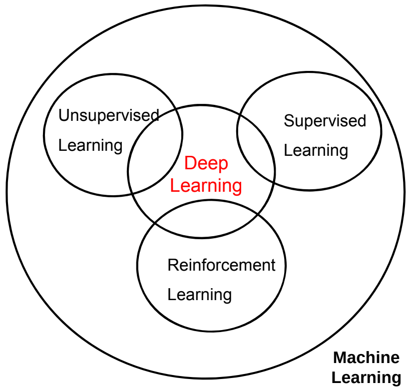
- Areas of deep learning can excel include clear objective functions, combinatorial number of solutions, and a lot of real data or simulated data
- Deep learning will not replace humans. Instead, like the industrial revolution that extends human's physical capabilities, it will extend human's cognitive ability to accomplish things they could accomplish otherwise.
- Human ingenuity augmented with AI will unlock our true potential . Human and machines collaborate to achieve amazing things that they cannot achieve alone

AI, Machine Learning, Deep Learning



Goodfellow, 2016

Machine Learning v.s. Deep Learning



As an enabler, DL has dramatically improved Supervised Learning, Unsupervised Learning, and Reinforcement Learning

What is Learning?

1. **Human Learning**

   "the acquisition of knowledge or skills through experience, study, or by being taught."

2. **Inductive learning and deductive learning**

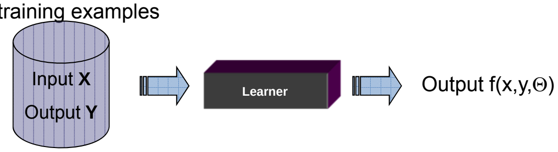
   Inductive Learning-learning general principles from examples

   Deductive Learning-learning examples from general principles

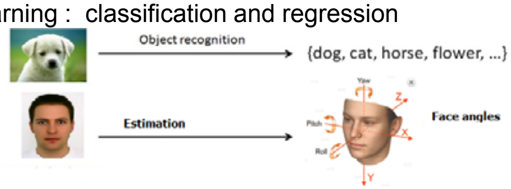
Machine Learning

1. Inductive learning

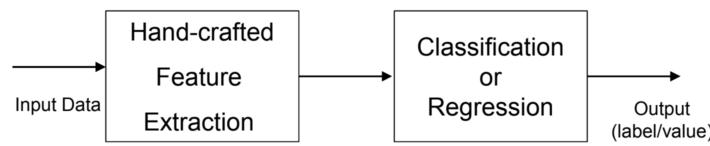
   Machine learns general principle (mapping function f(x)) about the relationships between input x and output y from given training examples



2. Learning : classification and regression



Traditional Machine Learning



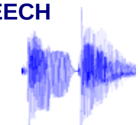
- Features are not learned but manually specified
- Generic shallow classifiers (SVM, Logistic regression, Adaboost, etc.), with limited capacity
- Given input X and output Y
  - Manual feature extraction (FeaEx): X→ X'
  - Learn a mapping f: X' → Y, i.e., Y=f(X',Θ)=f(FeaEx(X),Θ)

Traditional Machine Learning

VISION

hand-crafted features SIFT/HOG (fixed) → your favorite classifier (learned) → "car"

SPEECH



hand-crafted features MFCC (fixed) → your favorite classifier (learned) → \ˈd ē p\

NLP

This burrito place is yummy and fun! → hand-crafted features Bag-of-words (fixed) → your favorite classifier (learned) → "+"

Slide Credit: Marc'Aurelio Ranzato, Yann LeCun

Motivation

Features are key to successful machine learning

Multitude of hand-designed features currently in use

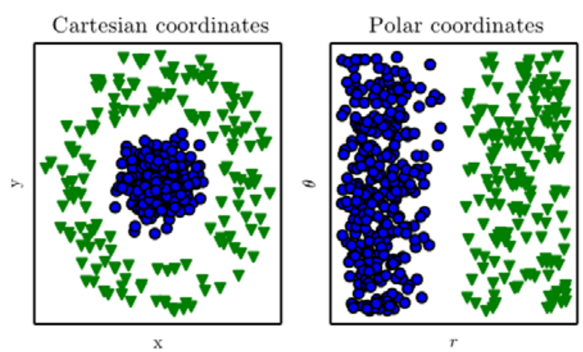
SIFT, HOG, LBP, MSER, Color-SIFT............

How to improve ML?

Better classifiers or better features

Slide: Rob Fergus et al

Better Features



Deep Learning

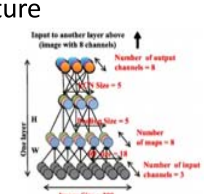
- Learning better features
- End to end learning
  - ✓ Joint feature and classifier learning
  - ✓ Fully automatic

Input ⇒ Layer 1 ⇒ Layer 2 ⇒ Layer 3 ⇒ Output

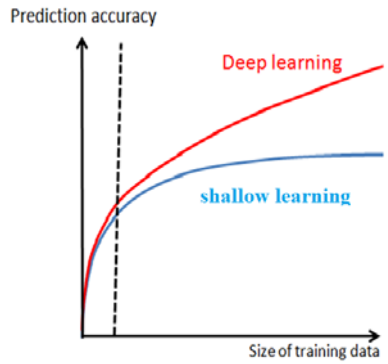
- Hierarchical multi-level layered structure

Given input X and output Y

- Automatic feature learning g: X → X' , X'=g(X)
- Learn a mapping f: X' → Y, Y=f(X',Θ)=f(g(X),Θ)
- Deep feature learning: $X'=g_1(g_2(..g_n(X)))$
- $Y=f(g_1(g_2(..g_n(X))),\Theta)$



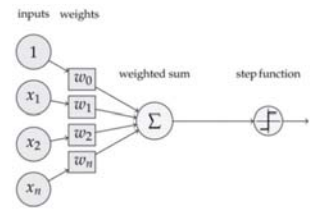
Deep Learning v. Shallow Learning



Deep Learning History

Frank Rosenblatt
Introduced Perceptron

1957

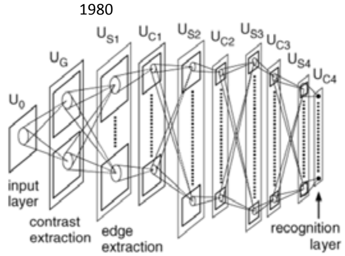


Deep Learning History

Frank Rosenblatt
Introduced Perceptron

Kunihiko Fukushima
introduced Convolutional
Neural Networks

1957 | 1980



CNN introduced more almost 38 years ago !

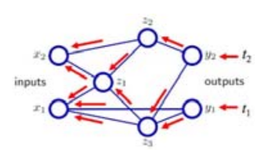
Deep Learning History

Frank Rosenblatt
Introduced Perceptron

Kunihiko Fukushima
introduced Convolutional
Neural Networks

Geoffrey Hilton introduced
Neural network Back
propagation
Nature

1957 | 1980 | 1986



- Solve general learning problems
- Tied with biological system

But it is given up…

- Hard to train
- Insufficient computational resources
- Small training sets
- Does not work well

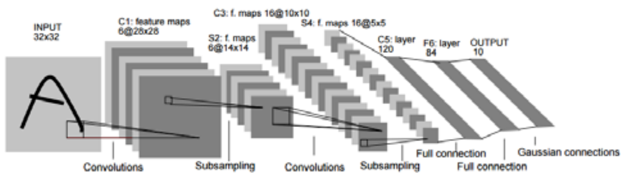
Deep Learning History

Frank Rosenblatt Introduced Perceptron — 1957

Kunihiko Fukushima introduced Convolutional Neural Networks — 1980

Geoffrey Hilton introduced Neural network Back propagation *Nature* — 1986

LeCun, Bottou, Bengio and Haffner introduced LeNet-5 — 1998



LeNet-5 is an improved CNN with 7-layer with gradient based learning for character recognition

Deep Learning History

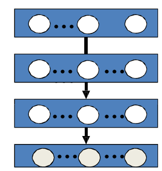
Rosenblatt Introduced Perceptron — 1957

Fukushima introduced CNN — 1980

Hilton introduced Back propagation — 1986

LeCun et al introduced LeNet-5 — 1998

Geffory Hilton introduced Deep belief net — 2006



- Probabilistic Deep Model
  - RBM as building block
- Unsupervised Layer-wise Learning

Rosenblatt Introduced Perceptron — 1957

Fukushima introduced CNN — 1980

Hilton introduced Back propagation — 1986

LeCun et al introduced LeNet-5 — 1998

Geffory Hilton introduced Deep belief net — 2006

Speech — 2011

deep learning

task	hours of training data	DNN-HMM	GMM-HMM with same data
Switchboard (test set 1)	309	18.5	27.4
Switchboard (test set 2)	309	16.1	23.6
English Broadcast News	50	17.5	18.8
Bing Voice Search (Sentence error rates)	24	30.4	36.2
Google Voice Input	5,870	12.3	
Youtube	1,400	47.6	52.3

Deep Networks Advance State of Art in Speech

Deep Learning leads to breakthrough in speech recognition at MSR.

courtesy of Xiaogang Wang of CUHK

Rosenblatt Introduced Perceptron — 1957

Fukushima introduced CNN — 1980

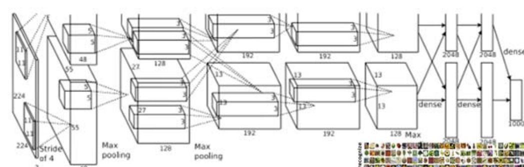
Hilton introduced Back propagation — 1986

LeCun et al introduced LeNet-5 — 1998

Geffory Hilton introduced Deep belief net — 2006

Speech — 2011

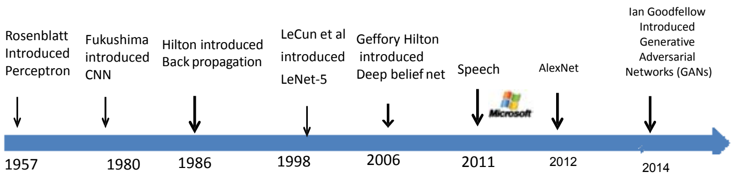
Krizhevsky, Sutskever and Hilton Introduced AlexNet — 2012



- AlexNet has 7-layers like LeNet.
- Train with ImaggNt, AlexNet significantly improves the performance of object recognition by more than 10% over the traditional learning method

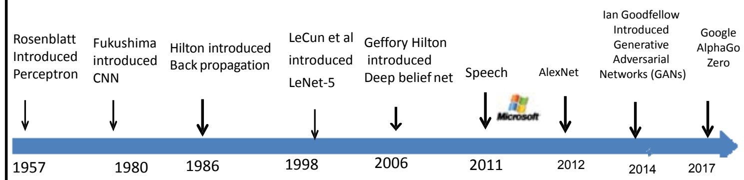
ImageNet: A large image database that contains 10 millions of images for 1000 object categories

Deep Learning History



- Generative Adversarial Networks (GANs) generative models for unsupervised deep learning
- Allows unsupervised deep learning and deep learning with small datasets
- Yann LeCun said that GANs are the most important idea in machine learning in the last 20 years

Deep Learning History



- Google introduced AlphaGo Zero that achieved super-human performance after three days self-training, without any human annotations.

Why are things working today?

- Big data
  - 10^8 samples (compared to 10^3 in 1990s)
- Fast computing power
  - GPUs are ~50x faster
- Plus old technologies (DBN and CNN) + new learning strategies, including gradient descent such as SGD (Stochastic Gradient Descent)
- Software infrastructure
  - Github, Tensorflow, Caffe, Amazon MTurks

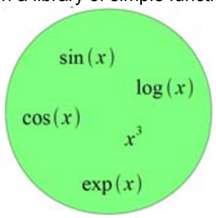
Courtesy of Dhruv Batra of Virgina Tech

Deep Learning Strengths

- (Hierarchical) Compositionality
  - Cascade of non-linear transformations
  - Multiple layers of representations
- End-to-End Learning
  - Learning (goal-driven) representations
  - Learning to feature extraction
- Distributed Representations
  - No single neuron "encodes" everything
  - Groups of neurons work together
- Latent Representations

Building A Complicated Function

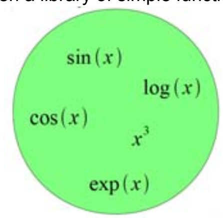
Given a library of simple functions



Compose into a complicate function

Building A Complicated Function

Given a library of simple functions

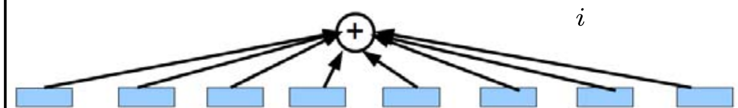


Compose into a complicate function

Idea 1: Linear Combinations

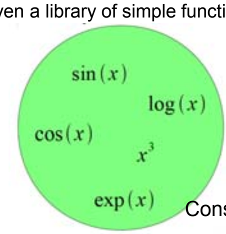
- Boosting
- Kernels
- Within layer of deep model

$$f(x) = \sum_i \alpha_i g_i(x)$$



Building A Complicated Function

Given a library of simple functions



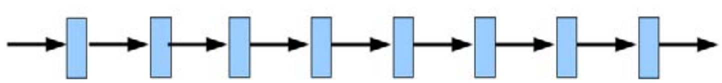
Compose into a complicate function

Idea 2: Compositions

- Deep Learning between layers
- Grammar models
- Scattering transforms…

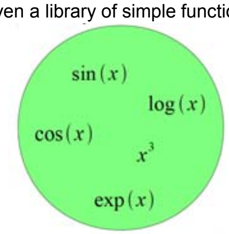
Construct complex function via recursive composition

$$f(x) = g_1(g_2(\ldots(g_n(x)\ldots))$$



Building A Complicated Function

Given a library of simple functions

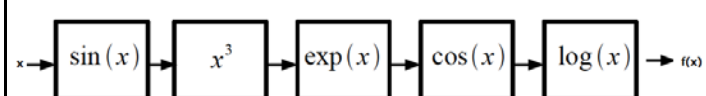


Compose into a complicate function

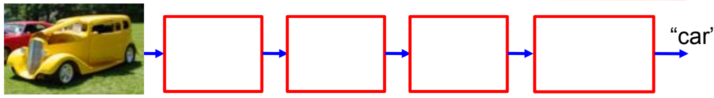
Idea 2: Compositions

- Deep Learning
- Grammar models
- Scattering transforms…

$$f(x) = \log(\cos(\exp(\sin^3(x))))$$

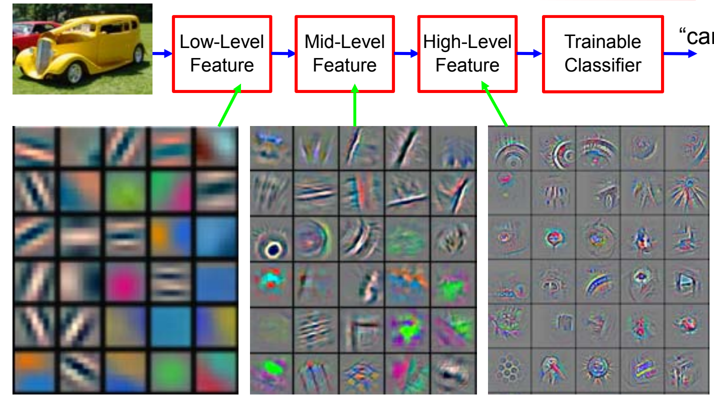


Deep Learning = Hierarchical Compositionality



Slide Credit: Marc'Aurelio Ranzato, Yann LeCun

Deep Learning = Hierarchical Compositionality

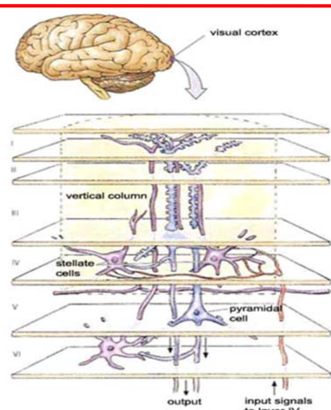


Feature visualization of convolutional net trained on ImageNet from [Zeiler & Fergus 2013]

Slide Credit: Marc'Aurelio Ranzato, Yann LeCun

Visual Cortex

The primary visual cortex is organized hierarchically and each layer consists of neurons and synapse



54

Deep Learning Strengths

- (Hierarchical) Compositionality
  - Cascade of non-linear transformations
  - Multiple layers of representations
- End-to-End Learning
  - Learning (goal-driven) representations
  - Learning to feature extraction
- Distributed Representations
  - No single neuron "encodes" everything
  - Groups of neurons work together
- Latent Representations

Courtesy of Dhruv Batra of Virgina Tech

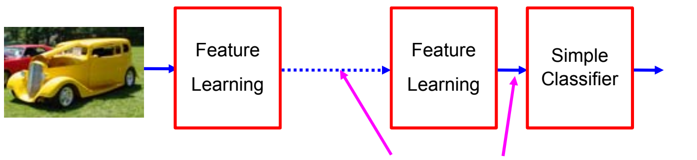
Deep Learning = End-to-End Learning

A hierarchy of trainable feature transforms

Each module transforms its input representation into a higher-level one.

High-level features are more global and more invariant

Low-level features are shared among categories



Slide Credit: Marc'Aurelio Ranzato, Yann LeCun

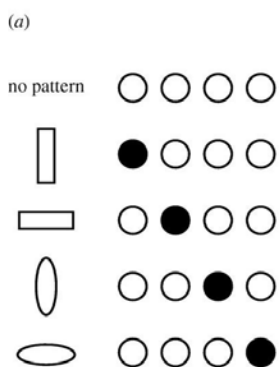
Deep Learning Strengths

- (Hierarchical) Compositionality
  - Cascade of non-linear transformations
  - Multiple layers of representations
- End-to-End Learning
  - Learning (goal-driven) representations
  - Learning to feature extraction
- Distributed Representations
  - No single neuron "encodes" everything
  - Groups of neurons work together
- Latent Representations

Courtesy of Dhruv Batra of Virgina Tech

Distributed Representations Toy Example

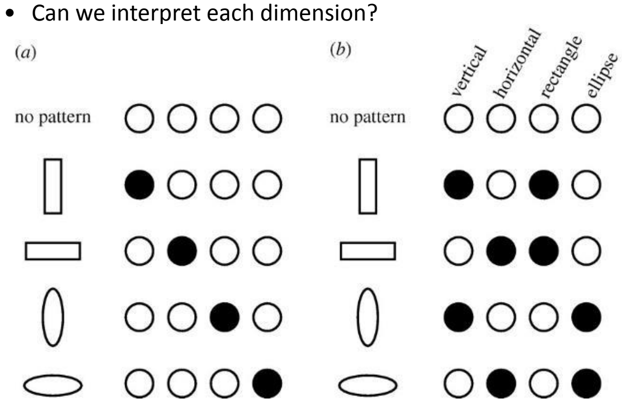
- Local vs Distributed



Slide Credit: Moontae Lee

Distributed Representations Toy Example

- Can we interpret each dimension?



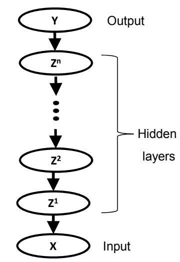
Slide Credit: Moontae Lee

Deep Learning Strengths

- (Hierarchical) Compositionality
  - Cascade of non-linear transformations
  - Multiple layers of representations
- End-to-End Learning
  - Learning (goal-driven) representations
  - Learning to feature extraction
- Distributed Representations
  - No single neuron "encodes" everything
  - Groups of neurons work together
- Latent Representations

Latent Representations

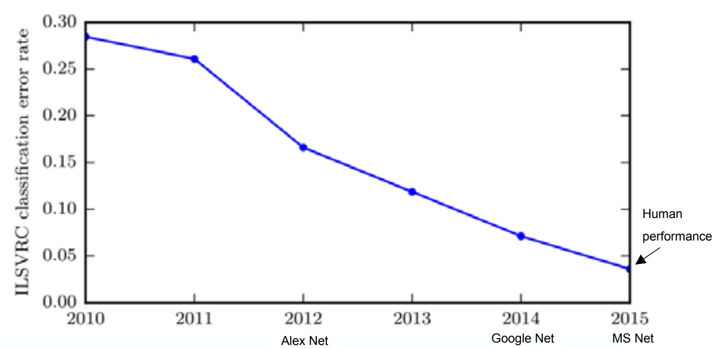
- The latent layers concisely and invariantly encode the salient properties of the input data in subspaces.
- They capture the latent feature representation at increasingly complex level.



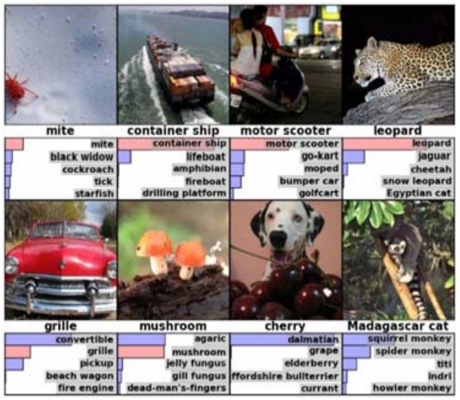
Deep Learning Applications

- Vision
- Natural Language Processing + Vision
- Speech
- Machine Translation
- Robotics
- Game playing
- Medical diagnosis
- …

Object Recognition on ImageNet

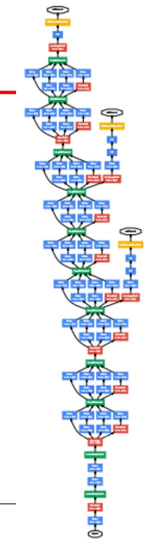


Classification Results



GoogLeNet (2014)

- Has 22 layers and the learning capacity is largely increased
- Add supervision at multiple layers
- 1000 categories classification on ImageNet,
- Reduce error rate from 15.3% to 6.6% on ImageNet



Microsoft Net (2015)

- Has 152 layers
- Learning residual functions
- Achieves 3.57% error on the ImageNet test set.

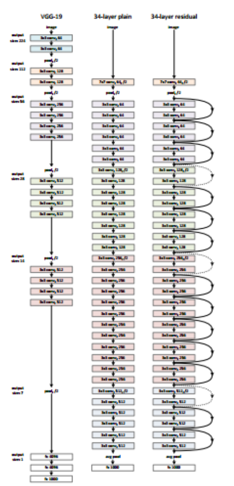


Image Caption

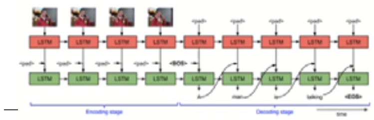
Captioned by Human and by Google's Experimental Program

Human: "A group of men playing Frisbee in the park."
Computer model: "A group of young people playing a game of Frisbee."

Image Caption

Human: a man sitting on a couch with a dog
Machine: a man sitting on a chair with a dog in his lap

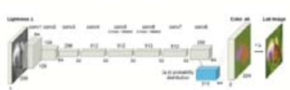
Video Description



Venugopalan et al.
"Sequence to sequence-video to text." 2015.

Code: https://vsubhashini.github.io/s2vt.html

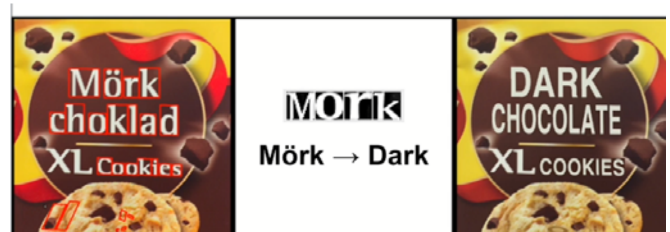
Image/Video Colorization

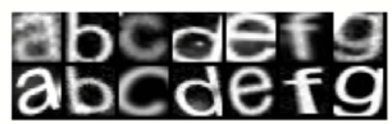


Zhang, Richard, Phillip Isola, and Alexei A. Efros. **"Colorful Image Colorization."** (2016).

References: [25, 26]

Automatic Translation



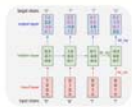


Google Translate

Text Generation

Life Is About The Weather!
Life Is About The (Wild) Truth About Human-Rights
Life Is About The True Love Of Mr. Mom
Life Is About Where He Were Now
Life Is About Kids
Life Is About What It Takes If Being On The Spot Is Tough
Life Is About… An Eating Story
Life Is About The Truth Now

The meaning of life is literary recognition.
The meaning of life is the tradition of the ancient human reproduction



Andrej Karpathy. "The Unreasonable Effectiveness of Recurrent Neural Networks." (2015).

Code: https://github.com/karpathy/char-rnn

Games

Google's AlphaGo AI beats Lee Se-dol again to win Go series 4-1

by Sam Byford | @345triangle | Mar 15, 2016, 5:00am EDT

SHARE TWEET LINKEDIN

Games

MIT Technology Review

Log in

Topics+ Top Stories Magazine

Robotics

Poker Is the Latest Game to Fold Against Artificial Intelligence

Two research groups have developed poker-playing AI programs that show how computers can out-hustle the best humans.

Speech Recognition

AI

Google has slashed its speech recognition word error rate by more than 30% since 2012

JORDAN NOVET @JORDANNOVET JANUARY 11, 2017 5:44 PM

Google senior fellow Jeff Dean, second from left, appears during a session onstage at the 2016 Google I/O developer conference in Mountain View, California.

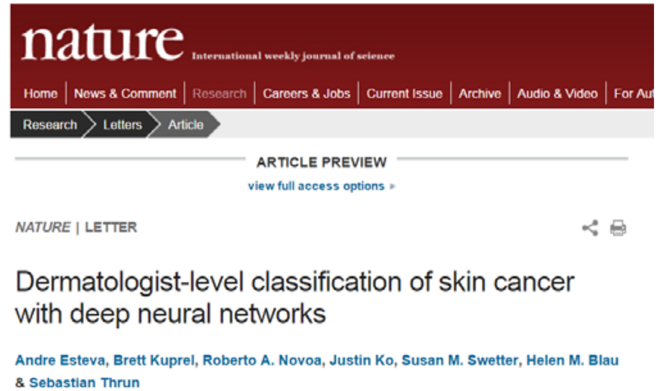
Above: Google senior fellow Jeff Dean, second from left, appears during a session onstage at the 2016 Google I/O developer conference in Mountain View, California.

Robotics

Deep learning startup Neurala raises $14 million to build brains for drones, autonomous cars, and more

PAUL SAWERS @PSAWERS JANUARY 17, 2017 5:01 AM

Medical Diagnosis



Sentiment Analysis and Music compostion

- Tweet sentiment analysis for stock prediction (4/13/17)
  - http://www.npr.org/2017/02/04/513469456/when-trump-tweets-this-bot-makes-money

- DeepBach
  - Bach-like music composition using deep learning
  - https://www.youtube.com/watch?v=QiBM7-5hA6o&app=desktop

Problems with Deep Learning

- Problem#1: Non-Convex! Non-Convex! Non-Convex!
  - Depth>=3: most losses non-convex in parameters
  - Theoretically, all bets are off
  - Leads to stochasticity
    - different initializations → different local minima
- Standard response #1
  - "Yes, but all interesting learning problems are non-convex"
- Standard response #2
  - "Yes, but it often works!"

Courtesy of Dhruv Batra of Virgina Tech

Problems with Deep Learning

- Problem#2: Hard to track down what's failing
  - It's hard to know why things are not working as there is no ground-truth at each intermediate layer
- Standard response #1
  - Tricks of the trade: visualize features, add losses at different layers, pre-train to avoid degenerate initializations…
  - "We're working on it"
- Standard response #2
  - "Yes, but it often works!"

Courtesy of Dhruv Batra of Virgina Tech

Problems with Deep Learning

- Problem#3: Lack of easy reproducibility
  - Direct consequence of stochasticity & non-convexity
  - Many local optimums
- Standard response #1
  - It's getting much better
  - Standard toolkits/libraries/frameworks now available
  - Caffe, Theano, Torch, Tensorflow
- Standard response #2
  - "Yes, but it often works!"

Courtesy of Dhruv Batra of Virgina Tech

Problems with Deep Learning

- Problem#4: Many hyper-parameters to tune
  - Number of layers and number of nodes for each layer
  - Choice of the loss function, choice of the activation function, choice of learning rate, number of training iterations, etc..
- Standard responses
  - Hire students to perform the tuning
  - Use heuristic methods such as grid search
  - Working on systematic ways to tune automatically

Problems with Deep Learning

- Problem#5: Require big data and computationally expensive
  - require large and representative data
  - require data annotation
  - require powerful computers and GPUs
- Standard responses
  - Collect more data !
  - Weakly supervised learning
  - Unsupervised learning
  - Synthetic data
  - Crowdsourcing such as Amazon Mechanical Turks
  - Buy GPUs

Problems with Deep Learning

- Problem#6: Blackbox
  - based on correlations instead of causality
  - cannot explain why its works
- Standard response
  - Working on it

Problems with Deep Learning

- Problem#7: Limited capability in modeling temporal data
- Standard response
  - Recurrent NN
  - Long Short-Term Memory (LSTM)
  - Cannot effectively model long term/global dynamics

Problems with Deep Learning

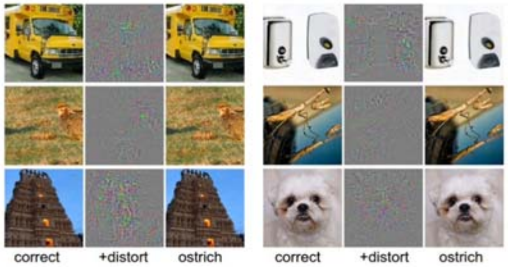
- Problem#8: just learns by examples, does not learn underlying principle , and cannot generalize even for the tasks with the same principles

  For example, a deep model trained to recognize human faces may not able to recognize animal (monkey) faces

- Standard response
  - Combine data with domain knowledge

Problems with Deep Learning

- Problem#9: can be fooled by a little distortion



- Standard response
  - No answer

Problems with Deep Learning

- Problem#10: cannot quantify uncertainty and online predict its confidence or performance
- Solution: Use Bayesian Neural Networks or Probabilistic Deep Models

Deep Learning Limitation Summary

- very data hungry (e.g. often millions of examples)
- very compute-intensive to train and deploy (cloud GPU resources)
- poor at representing uncertainty
- easily fooled by adversarial examples
- finicky to optimise: non-convex + choice of architecture, learning procedure, initialisation, etc, require expert knowledge and experimentation
- uninterpretable black-boxes, lacking in trasparency, difficult to trust

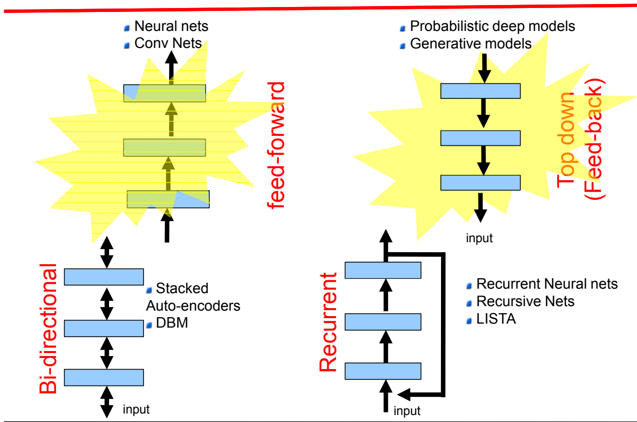
Deep learning: big data for single and/small tasks and narrow intelligence

Deep Learning Strengths and Weaknesses

- Deep learning is a tool that can
  - speed up analysis and discovery
  - can accomplish things faster
  - can scale up to large dataset
- Deep learning successes are overly exaggerated.
  - The real hard success in deep learning is in speech recognition and machine translation, where we see daily commercial products like Apple's Siri, Amazon's Echo, and Google's Home and youtube transcription
  - The success in computer vison is soft as the success so far is mainly limited to closed benchmark datasets
  - Its success in other fields such as biology are incremental

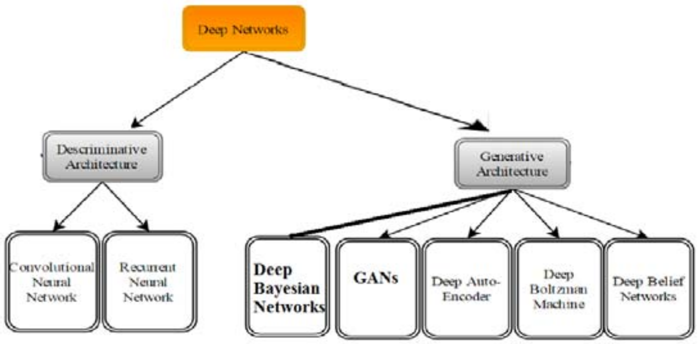
Courtesy of Qiang Ji

Main types of deep architectures



Slide Credit: Marc'Aurelio Ranzato, Yann LeCun

Deep Learning Techniques

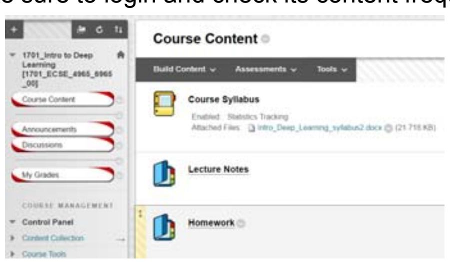


Course Information

- Instructor: Qiang Ji
  - jiq@rpi.edu
  - Office Hours: 5:30-6:30 pm, Mondays and Thursdays
  - Location: JEC 7004
- TAs: TBD

Course Materials

- All courses materials, including lecture notes, homework and project assignments, supplementary materials, announcements, etc. will be posted on LMS
- Make sure to login and check its content frequently



Topics to be Covered

1. **Introduction**
2. **Machine Learning Fundamentals**
   Probabilities, Optimization, Linear Algebra, Multivariate Calculus
   Linear Regression, Linear Classification, Generative and Discriminative learning, Bayes classifier, Logistic Regression
3. **Deterministic Deep Models**
   Neural Networks, Deep Neural Networks, Convolutional Neural Networks, and Autoencoder,
4. **Advanced Topics**
   Recurrent Neural Networks, Deep Reinforcement Learning, and Generative Adversarial Networks
5. **Probabilistic Deep Models**
   Bayesian Neural Networks, Deep Boltzmann Machine, Deep Belief Networks and Deep Bayesian Networks

Textbook

- **Deep Learning**, Ian Goodfellow and Yoshua Bengio and Aaron Courville, MIT Press, 2016,
- Optional
- http://www.deeplearningbook.org
- download the book quickly for free!

Tensorflow

- The primary software for implementing various programming assignments and projects
  https://www.tensorflow.org/tutorials/
- The most popular deep learning software, with support from Google and a community of users
- Interface: Python
- Support GPUs
- Support Linux and OS, Window version in the future

Course Information

- Homework (10%)
- Midterm (20%)
- Class Projects (50%)
- Final project (20%)

- The homework is written assignment, involving theoretical aspects of the class.
- Class projects are all programming assignments.
- Midexam is open-book and comprehensive
- Final project is a significant competition, involving implementing a deep learning model for a real world problem. It can be a group project.

Prerequisites

Desired Requirements

- ECSE Courses: Pattern Recognition, Probabilistic Graphical Models, Image Processing, and Computer Vision
- CS courses: Machine Learning and Data-mining
- Math courses in probability, optimization, and linear algebra
- Strong programming skills

Minimum Requirements

- Basic knowledge in probability, linear algebra, and optimization
- Understand images, its format, and basic image processing
- Good knowledge of a high level programming language such as Python/Matlab/C++

Deep Learning Articles

- Facebook AI Director Yann LeCun on His Quest to Unleash Deep Learning and Make Machines Smarter

https://spectrum.ieee.org/automaton/robotics/artificial-intelligence/facebook-ai-director-yann-lecun-on-deep-learning

- Machine-Learning Maestro Michael Jordan on the Delusions of Big Data and Other Huge Engineering Efforts

https://spectrum.ieee.org/robotics/artificial-intelligence/machinelearning-maestro-michael-jordan-on-the-delusions-of-big-data-and-other-huge-engineering-efforts

- Will the Future of AI Learning Depend More on Nature or Nurture?

https://spectrum.ieee.org/tech-talk/robotics/artificial-intelligence/ai-and-psychology-researchers-debate-the-future-of-deep-learning

Courtesy of Qiang Ji