Microsoft Research

Deep Learning: Premise, Philosophy, and Its Relation to Other Techniques

Dong Yu

Microsoft Research

Microsoft Research

Deep Learning: A Hot Area

- Deep Learning has been hot in speech recognition for several years
  - Significant error reduction on real-world large vocabulary speech recognition with deep learning methods
  - Almost all major players in the speech recognition industry have deployed deep neural network-based speech recognition products
- Deep Learning is hot in computer vision
  - Best results on ImageNet and Semantic Segmentation are achieved with deep learning techniques (convolutional nets + DNN)
- Deep Learning is becoming hot in natural language processing
  - Unclear whether deep learning techniques can beat the conventional techniques at this moment

Microsoft Research

Tutorial Outline

- **Part 1**: Deep Learning: Premise, Philosophy, and Its Relation to Other Techniques
- **Part 2**: Basic Deep Learning Models
- **Part 3**: Deep Neural Network and Its Application in Speech Recognition

Part 1: Outline

- **Deep Learning: Premise and Goal**
- Representation Learning
- Relationship to Other Models
- Theoretical Questions

Deep Learning: Definitions

- **Definition 1**: A class of machine learning techniques that exploit many layers of non-linear information processing for feature extraction and transformation and for pattern analysis and classification.

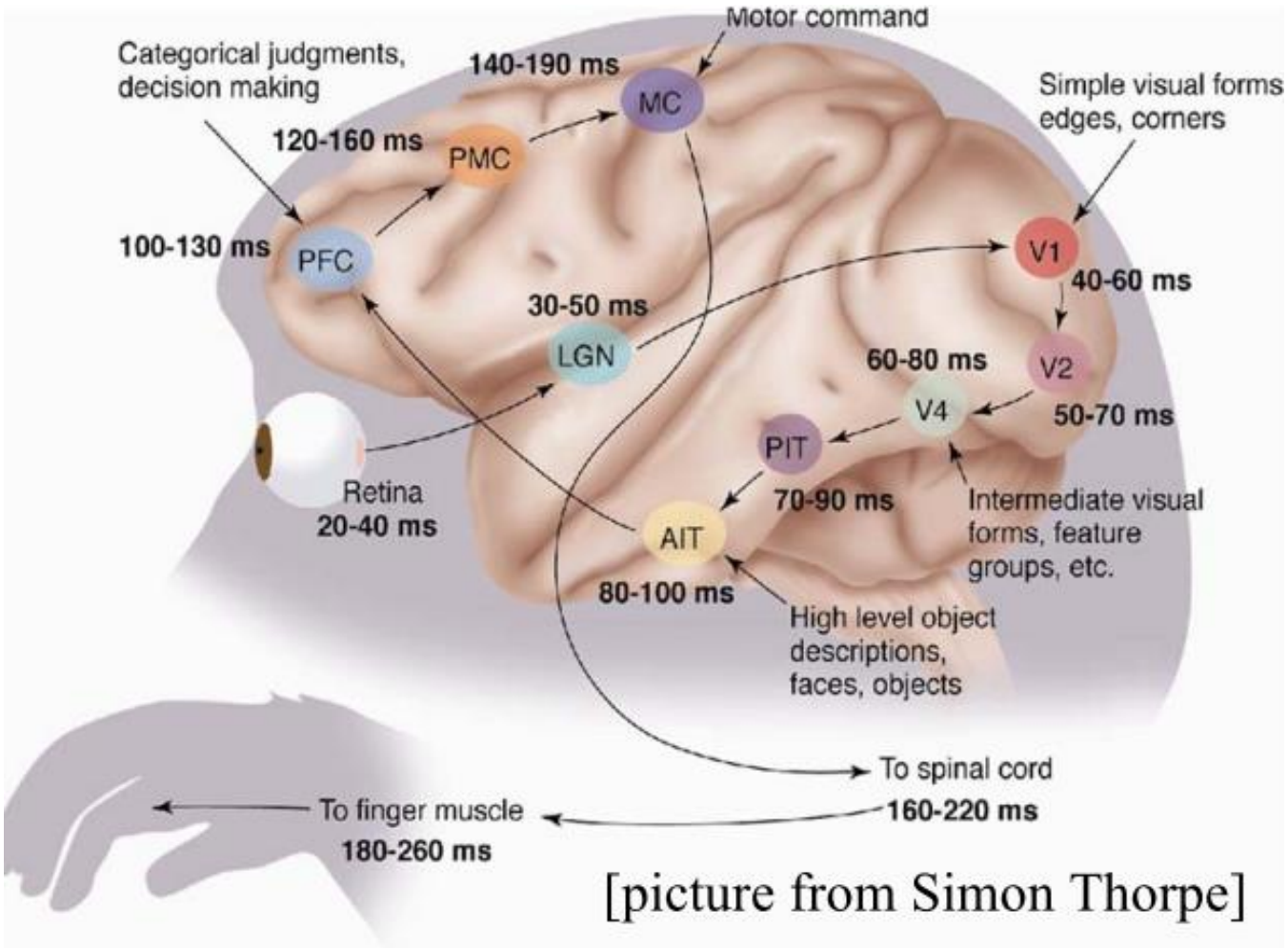
- **Definition 2**: A sub-field within machine learning that is based on algorithms for learning multiple levels of representation in order to model complex relationships among data. Higher-level features and concepts are thus defined in terms of lower-level ones, and such a hierarchy of features is called a deep architecture. (Wikipedia)

Hint from Human Learning

- Human brain has 10^{14} synapses (connections) – processed in parallel.
- Human vision and auditory systems involve many layers of non-linear information processing.
- Human learning is largely unsupervised
  - esp. in lower layers
- Human brain might use the same algorithm to understand many different modalities:
  - e.g., Ferret experiments, in which the "input" for vision was plugged into auditory part of brain, and the auditory cortex learns to "see." [Roe et al., 1992]

Suggest a deep and wide architecture

Visual Cortex is Hierarchical



- Retina - LGN - V1 - V2 - V4 - PIT - AIT ....

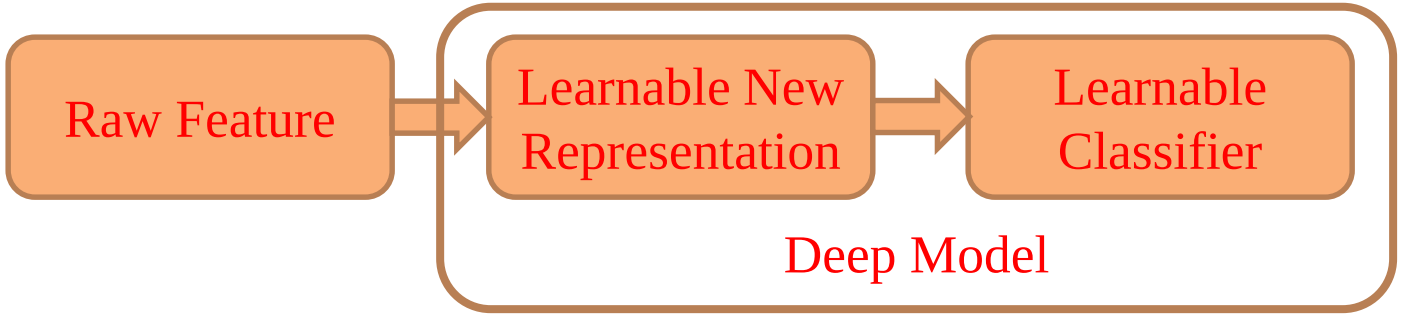
One Way vs. The Way

Why Deep?

- We can approximate any function as close as we want with shallow architecture. Why deep?
  - E.g., kernel machines and Single hidden layer neural network are universal approximators
- Deep machines
  - Can represent more complex functions with less parameters
  - Can learn the same function with less training data by reusing low-level feature detectors
  - Can learn move invariant high level features

Deep Learning in Essential

- Learn complicated feature representation through many (more than one) layers of nonlinear (often simple) processing
- Learn representation automatically and jointly with classification (or whatever) tasks (end-to-end optimization)
- Task dependent representation performs better



Task of Deep Learning

- Deep Learning addresses the problem of learning hierarchical representations with a single (universal?) algorithm
- Can we make all the feature transformation modules trainable and get them to learn appropriate representations?
  - What is the fundamental principle?
  - What is the learning algorithm?
  - What is the architecture?
- Prefer simple solution over complicated solution
  - A good solution often is a simple solution
  - By retrospect, everyone should think it is trivial and should be the way it is

Part 1: Outline

- Deep Learning: Premise and Goal
- **Representation Learning**
- Relationship to Other Models
- Theoretical Questions

Good Representation

- invariant to perturbations and discriminative in classification
- powerful and efficient to represent complex structures
- powerful to avoid under-fit and less-flexible to avoid over-fit
- works effectively with small dataset and can scale up to large dataset

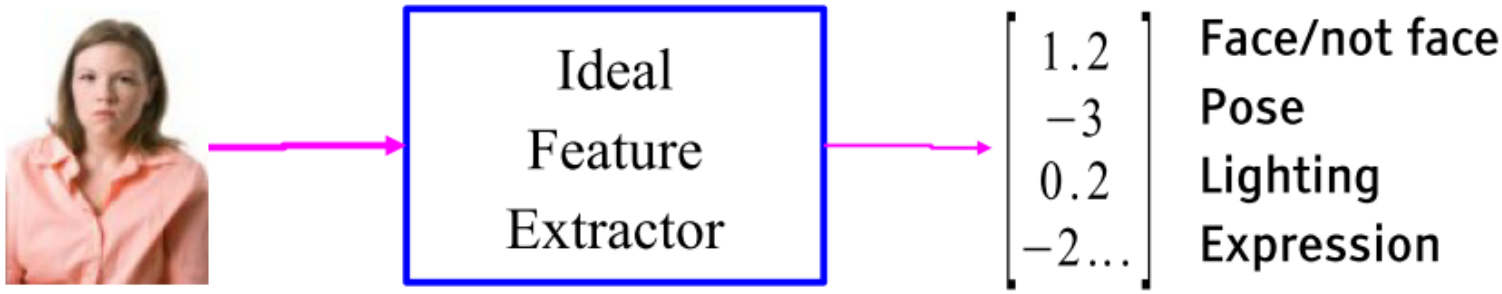
Deep models can resolve some of the contradictions

Representation Hierarchy

- Deep hierarchy of representations
  - Increasing level of abstraction
  - Each processing stage is learnable feature transform that 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
- Image recognition
  - pixel → edge → texton → motif → part → object
- Speech
  - waveform → spectrum → log-spectral band → cepstrum → phoneme → word

Representation Learning

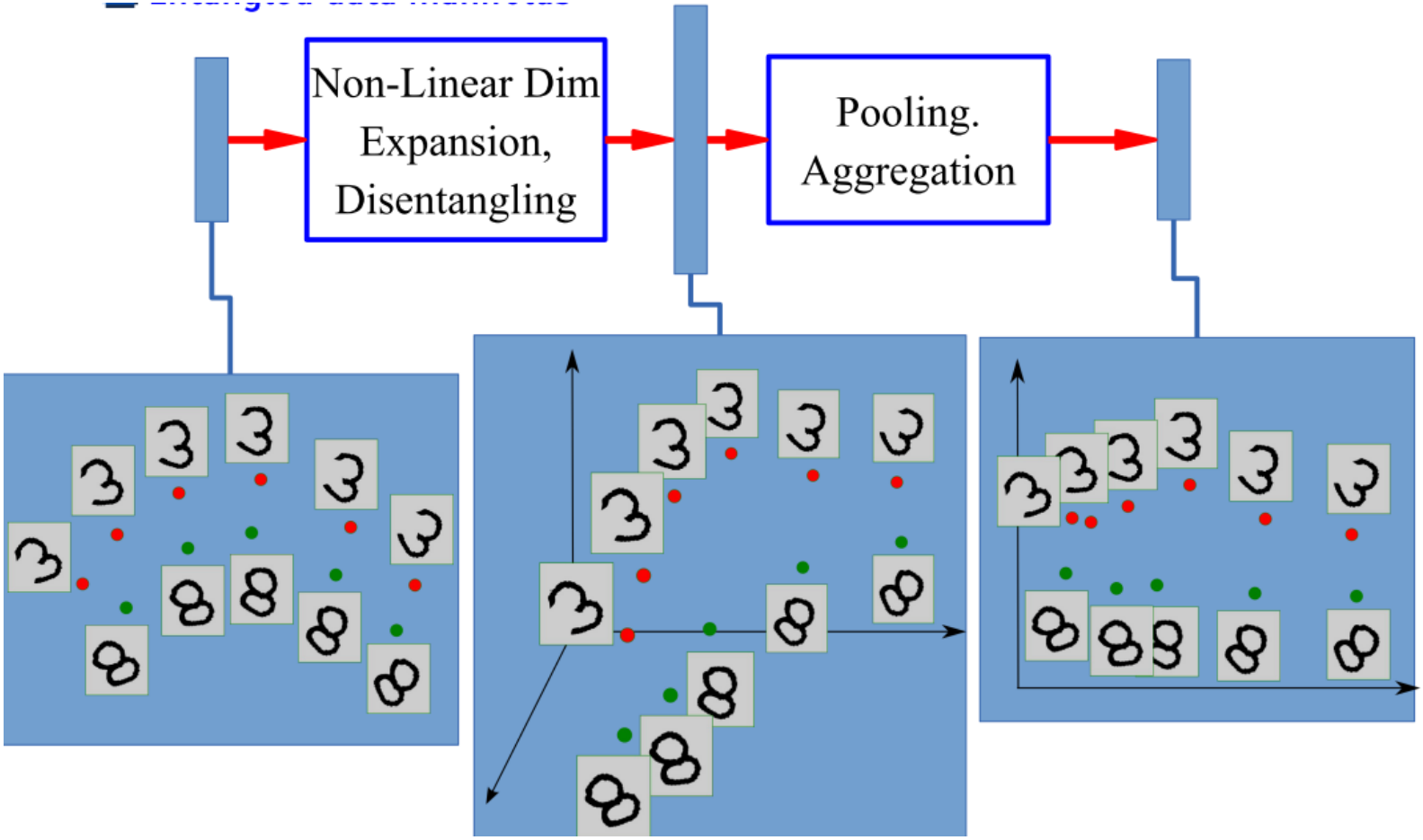
- Identifying the independent factors that best explain the data
- We do not have good and general methods to learn such a representation yet.
- The manifold hypothesis:
  - Natural data lives in a nonlinear low-dimensional manifold



Invariant and Discriminative

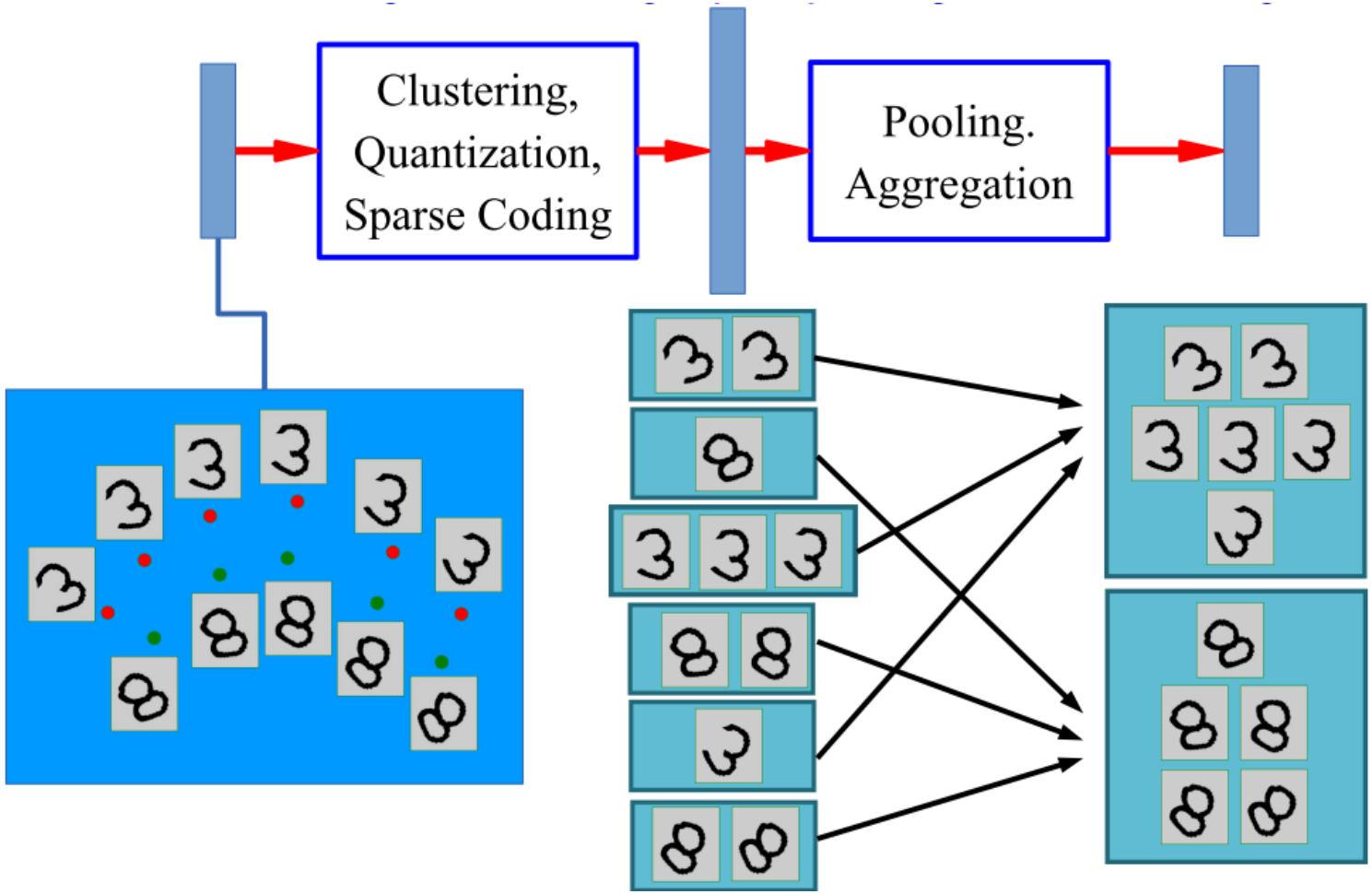
- **Discriminative**
  - Transfer the raw feature non-linearly into a higher dimensional space in which things that were non-separable become separable
- **Invariant**
  - Pool or aggregate features in the new space to introduce invariance.
  - E.g., group things that are semantically similar.

Raw -> High Dim -> Pool



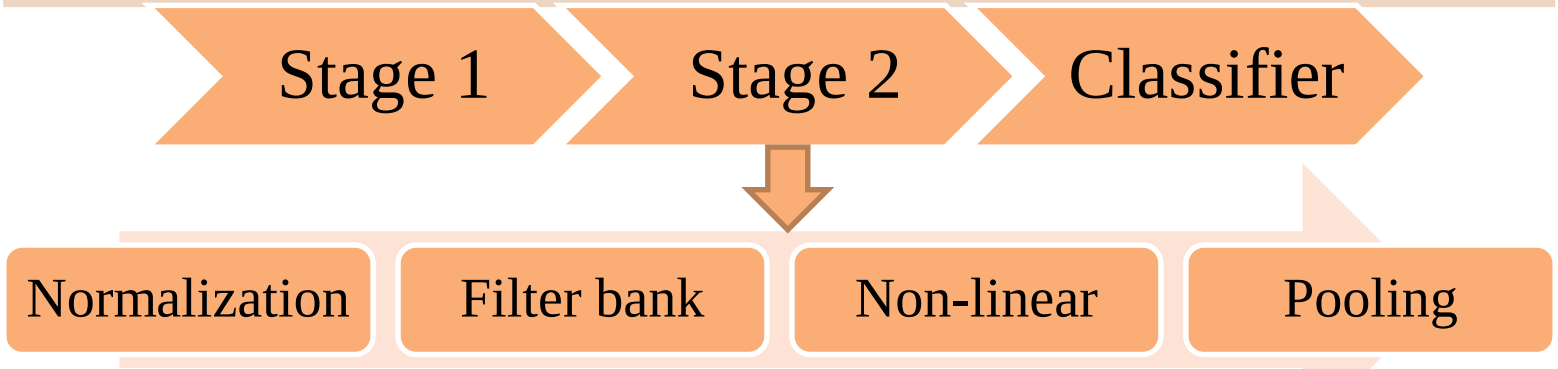
Slide credit: LeCun and Ranzato

Clustering -> Pooling



Slide credit: LeCun and Ranzato

General Framework



- Normalization: average removal, local contrast normalization, variance normalization
- Filter Bank: dimension expansion, projection on overcomplete basis
- Non-Linearity: sparsification, saturation, lateral inhibition, tanh, sigmoid, winner-takes-all
- Pooling: aggregation and clustering

Typical Training Procedure

- Design a training criterion
  - Best if aim at the task directly
  - May use a surrogate of the true criterion
  - Unsupervised tasks also have training criterion, e.g.,
  - Used in most practical systems for speech and image recognition
- Initialize model parameters
  - Random initialization: sufficient if label set is large
  - Generative pretraining: train each layer unsupervised, one after the other. Useful when label set is poor or small
- Refine model parameters
  - Adjust model parameters to optimize the training criterion
  - Refine only the classifier layer: if label set is small
  - Jointly refine the whole model: if label set is large

Constrained Optimization

- Goal: global optimum
  - Only possible with convex problem or problems with small search space!
  - Not possible for most real world problems
- Solution 1: Convert Problem
  - Convert to an easier approximated problem
  - Use a simpler but correlated criterion (e.g., pub number)
- Solution 2: Search for sub-optimal solution
  - Greedy search (you only know the local world)
  - Explore search space wisely
    - Use all available results.

Constrained Optimization

- Key: reduce constraint and increase search space
  - many science advancement comes from this
- Constraints include but not limited to
  - Physical limitation
  - Available tools
  - Rules
  - Knowledge on the topic
  - Experience
  - Contradictory multiple objectives (make tradeoffs)
  - Time available to solve the optimization problem

Part 1: Outline

- Deep Learning: Premise and Goal
- Representation Learning
- **Relationship to Other Models**
- Theoretical Questions

Traditional Machine Learning

- Feature engineering
  - Design features that work well on a specific task
  - Requires domain knowledge, error analysis, and trial and error
- Simple model
  - Linear model
  - Support vector machine (SVM)
  - Gaussian mixture model (GMM)
  - Single-hidden layer neural network (SHLNN)
- Work is shifted to manually designing good features that can work effectively with simple models

Shallow Models

- Two-layer models are not deep : there is no feature hierarchy
- Single-hidden layer neural network
  - Hidden layer: feature representation
  - Output layer: classification / regression
- SVMs and Kernel methods
  - Kernel layer: feature representation (unserpervised) in higher dimension
  - Output layer: linear classification
- Decision tree
  - No hierarchy of features. All decisions are made in the input space

Relationship with Graphical Models

- Graphical models can be deep (but most often not so)
  - Graphical structure can be deep
  - Energy function can be represented as a deep network
- Some deep models can be represented as a graphical model
  - By converting energy function to factor graph

Relationship to Convex Optimization

- Most interesting real-world problems involve Non-convex optimization
  - Gaussian mixture model and hidden Markov models widely used in speech recognition are NOT convex
- Most deep learning architectures require Non-convex optimization
  - Cannot be easily analyzed theoretically.
  - Has many local optima.
  - No guarantee to find the optimal solution
  - The order the samples are presented will affect the learning result

Microsoft Research

Part 1: Outline

- Deep Learning: Premise and Goal
- Representation Learning
- Relationship to Other Models
- **Theoretical Questions**

Empirical Versus Theoretical

- Most deep learning results so far are empirical
  - It's great that it works
  - But we should still seek to understand why and how to further improve the systems
- Conventional theoretical analysis on deep learning is difficult
  - We cannot get a tight bound on generalization ability to guide the model selection (no worse than other machine learning models)
  - It's hard to prove things related to deep learning due to the complexity involved
- Conventional theoretical analysis may not be sufficient
  - What are the theoretical questions that are truly important?

Some Theoretical Questions

- What are the suitable training criteria for supervised, unsupervised, semi-supervised, and lightly-supervised representation learning?
- What are the scalable optimization algorithms that work on large datasets and can find a relatively good solution?
- What learning behavior will surface when the network is deep and wide?
- Can we control the learning process so that the learning process can be both fast and effective?
- If we consider the deep learning system as a nonlinear dynamic system, what can we say about the system?

Some Theoretical Questions

- Can we design a learner that
  - Is fault tolerant to gaps in data
  - Is massively parallel
  - Is extremely power efficient
  - Is highly scalable
  - Has minimal arithmetic precision requirements
  - allows ultra-dense, low power implementations

Microsoft Research

Questions?