A Perspective on Trust in Machine Learning and Control for Dynamic Autonomous Systems

AFRL-SUNY Trusted AI Challenge Series

Pramod P. Khargonekar

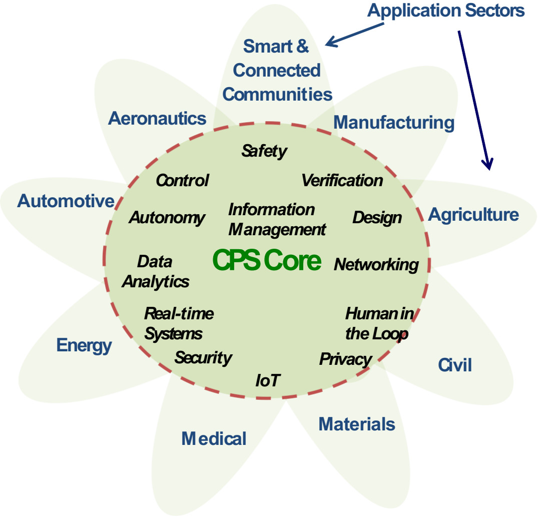
Department of Electrical Engineering and Computer Science
University of California, Irvine

14 october 2020

Outline

1. Smart Cyber-Physical Systems
2. Control Systems Viewpoint
3. Recap of Recent Machine Learning Breakthroughs
4. What does Future Hold?

Cyber-Physical Systems



Application Sectors

Smart & Connected Communities

Aeronautics

Manufacturing

Automotive

Agriculture

Energy

Civil

Medical

Materials

CPS Core

Safety

Control

Verification

Autonomy

Information Management

Design

Data Analytics

Networking

Real-time Systems

Human in the Loop

Security

Privacy

IoT

Application Domains

Transportation

- Faster and safer vehicles (airplanes, cars, etc)
- Improved use of airspace and roadways
- Energy efficiency
- Manned and un-manned

Energy and Industrial Automation

- Homes and offices that are more energy efficient and cheaper to operate
- Distributed micro-generation for the grid

Healthcare and Biomedical

- Increased use of effective in-home care
- More capable devices for diagnosis
- New internal and external prosthetics

Critical Infrastructure

- More reliable power grid
- Highways that allow denser traffic with increased safety

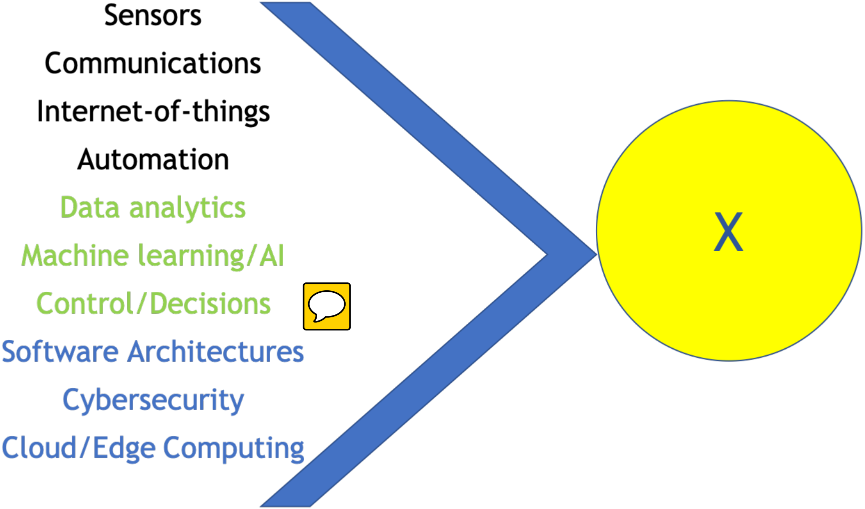
Source: NSF

CPS Properties

- Pervasive computation, sensing, and control
- Networked at multiple scales
- Dynamically reorganizing/reconfiguring
- High degrees of automation
- Dependable operation with potential requirements for high assurance of reliability, safety, security and usability
- With or without human interaction/supervision
- Conventional and unconventional substrates/platforms
- Range from the very small to the large to the very large

Source: NSF

Smart-X: Conceptual View



Sensors

Communications

Internet-of-things

Automation

Data analytics

Machine learning/AI

Control/Decisions

Software Architectures

Cybersecurity

Cloud/Edge Computing

Aspirational and Emerging Applications: Examples

- Smart-X
  1. Smart manufacturing
  2. Smart grid
  3. Smart transportation
  4. Smart cities
  5. Smart health
- Autonomous systems
  1. Unmanned air vehicles
  2. Self-driving cars
  3. Autonomous robots

Human individual and group behaviors are central in many of these applications.

Smart Cyber-Physical-Human or Autonomous Systems.

Will we trust them?

Control Systems : Strong Theoretical Foundations

- Stability theory
- Optimal control
- Linear multivariable control
- Robust control
- Nonlinear control
- Adaptive Control
- Stochastic control
- Distributed control

Key Ideas

- Heavy use of mathematical models and techniques
  1. Differential and difference equations; ODEs, PDEs, ldots
  2. Discrete-event models
  3. Hybrid models
  4. Deterministic and stochastic
  5. Distributed, networked, hierarchical, . . .
  6. Toolsets for analysis and design: time-domain, frequency domain, optimization, numerical computations, . . .
- Explicit accounting of modeling errors and robustness
- Mathematically provable properties
- Bottomline: so long as mathematical assumptions hold, conclusions are guranteed.

Adaptive Control

- Control systems that can adapt to changes in the system or the environment
- A long-standing goal in control theory
- Deterministic and stochastic models
- Parameter learning, neural network learning, . . .
- Mathematical results on closed loop stability, convergence of parameters, . . .
- Transient performance remains a hard problem
- Recent results on flight tests of adaptive control

Formal Methods

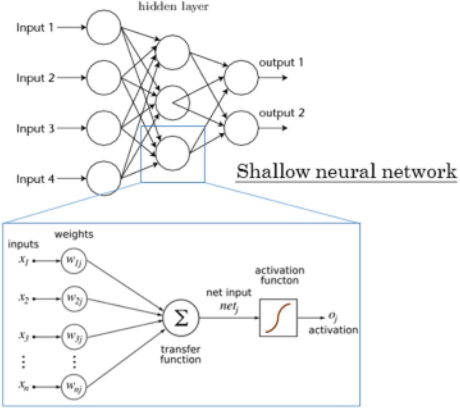
- Specify desired behavior from the controlled system in terms of logical statements
- Temporal logics: linear temporal logic (LTL), metric temporal logic (MTL), signal temporal logic (STL), . . .
- Model checking, theorem proving, etc. for verification
- Investigations of robustness
- Concerns about specifying the desired behaviors
- Concerns about scaling to complicated systems

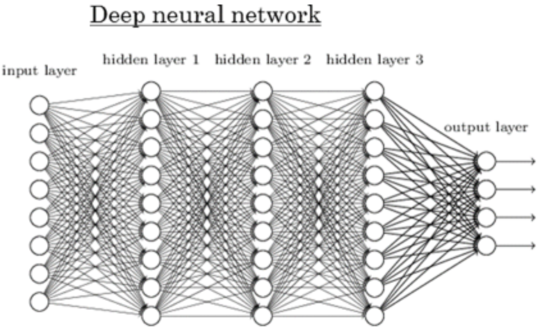
Recap of Recent Machine Learning Breakthroughs

Computational Intelligence: Pattern Recognition or Model Building

- Two fundamentally different perspectives on learning from data:
  1. Statistical pattern recognition from data for prediction and control.
  2. Using data to build causal models to understand, predict and control.
- Possible to combine these two approaches.
- Causality a critical issue.

Deep vs Shallow Neural Networks


Shallow neural network


Deep neural network

Source: github

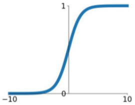
Building Block: A Single Artificial Neuron Unit

- Inputs: $x_1, x_2, \ldots x_n$
- Weights: $w_1, w_2, \ldots w_n$
- An activation function σ
- Examples of activation functions:
- Output given by

$$
\begin{aligned}
a &= \sum_{j=1}^{n} w_j x_j \\
y &= \sigma(a)
\end{aligned}
$$

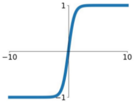
Sigmoid

$\sigma(x) = \frac{1}{1+e^{-x}}$



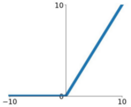
tanh

$\tanh(x)$



ReLU

$\max(0, x)$



Key Advantage of Deep Networks

" ... shallow classifiers require a good feature extractor . . . one that produces representations that are selective to the aspects of the image that are important for discrimination . . . The conventional option is to hand design good feature extractors, which requires a considerable amount of engineering skill and domain expertise. But this can all be avoided if **good features can be learned automatically . . . This is the key advantage of deep learning.***"*

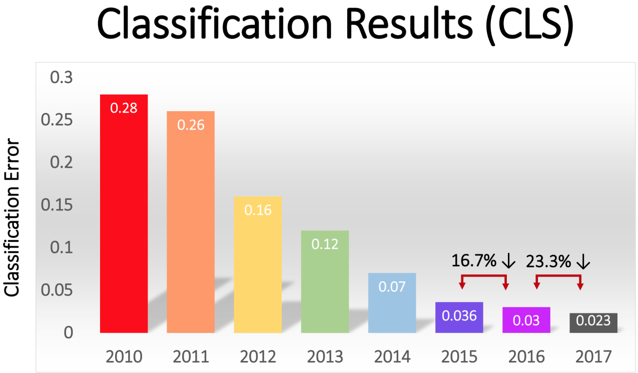
Deep Learning, LeCun, Bengio, and Hinton, Nature, 2015.

Major DL Innovations

- Convolutional neural networks
- Long Short Term Memory (LSTM) for sequential data
- Numerous algorithmic and architectural innovations
- Training and optimization of extremely large networks
- Use of graphics processors for computation
- Leveraging of large volumes of training data

Breakthrough in Vision: ImageNet Competition

ImageNet Classification with Deep Convolutional Neural Networks, Krizhevsky, Sutskever, and Hinton, 2012

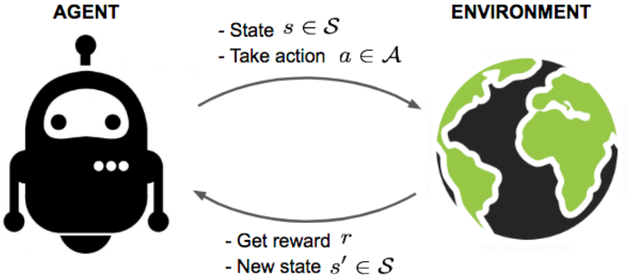


Source: image-net.org

Recurrent Neural Networks

- Recurrent neural networks (RNNs): neural network models with the ability to pass information across time steps
- Suitable for modeling data that are
  - Sequential and dependent.
  - Of varying input lengths.
- RNNs: natural choice for time series and other sequential applications.
- Long Short Term Memory (LSTM) Networks: the state-of-the-art RNNs.

RL Framework


AGENT
ENVIRONMENT
- State $s \in \mathcal{S}$
- Take action $a \in \mathcal{A}$
- Get reward r
- New state $s' \in \mathcal{S}$

The "agent" is the controller and the "environment" includes the plant, uncertainty, disturbances, noise, etc.

Source: github

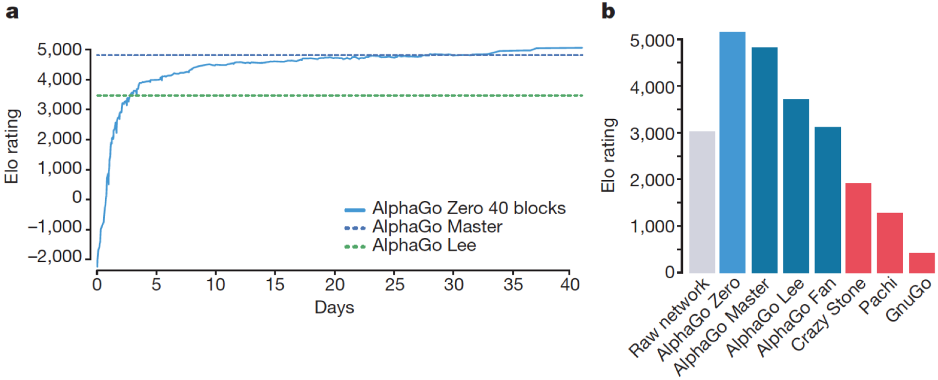
Reinforcement Learning: General Setup

- At each time step, agent observes the state, takes action, and receives a reward.
- Goal for the agent: choose actions to maximize total discounted reward.
- Optimal action policy is a form of control law.
- Can the agent learn the optimal policy by suitable use of state and reward data?
- RL: A general machine learning paradigm to solve problems and attain goals.

Key Ideas and Building Blocks

- Bellman's optimality principle: *Tail of an optimal policy must be optimal.*
- Function $Q(x, a)$: optimal policy given by maximizing with respect to a.
- One approach: Learn the Q-function.
- Recent innovations in modern RL
  1. Deep Reinforcement Learning: Use deep neural networks to approximate Q (DQN)
  2. Experience replay to reuse past data
  3. Asynchronous and parallel RL
  4. Rollouts based planning for RL
  5. Self-play for faster learning
  6. Techniques for data efficiency
  7. Techniques for continuous action spaces

AlphaGo Zero achieves State-of-the-Art Performance



Despite learning by itself from zero prior knowledge,
it learns and outperforms all other algorithms.

Autonomy and Learning

- Learning and adaptation will be needed for achieving ambitious autonomous systems goals
- Machine learning will increasingly be integrated into autonomous systems of the future
- Myriad forms of machine learning: DL, RL, GAN, . . . will be developed for use in autonomous systems applications
- Numerous questions will arise about trust in such systems

Possible Goals and Approaches

- Increased transparency in the learning components: explainable AI
- Cognitive CPS: build cognitive properties into future CPS: perception, memory, attention, problem solving, . . .
- Build self-awareness into smart autonomous systems
- Notion of intent in smart autonomous systems