MSML610: Advanced Machine Learning

Introduction

Instructor: Dr. GP Saggese - gsaggese@umd.edu

References:

- AIMA Chap 1

A Map of Machine Learning

- **A Map of Machine Learning**
- What Is Artificial Intelligence

A Map of Machine Learning

Machine Learning Theory

- **VC theory**
  - Measure model capacity and generalize based on hypothesis space complexity
- **Bias-variance decomposition**
  - Prediction error is the sum of:
    - Bias: Error from simplistic model assumptions
    - Variance: Error due to sensitivity to training data fluctuations
- **Computation complexity**
  - Related to information theory and compression
  - E.g., Minimum Description Length (MDL) measures model complexity via efficient model and data description
- **Bayesian approach**
  - Treat ML as probability
  - Combine prior knowledge with observed data to update belief about a model
- **Problem in ML theory**
  - Assumptions may not align with practical problems

Machine Learning Paradigms

- How do you set up the learning problem?
  - **Supervised learning**
    - The dataset includes inputs with corresponding outputs
    - Develop an input-output relationship
  - **Unsupervised learning**
    - The data is unlabeled, discover structure within the data
    - E.g., anomaly detection, clustering
  - **Reinforcement learning**
    - The correct answer is not immediately available
    - Evaluate actions based on final outcomes
  - **Active learning**
    - Not all examples are available initially
    - Request outputs for specific inputs
  - ...

Machine Learning Models

- What is the form of the model and how to fit / predict from the data?
  - Linear models
  - Generalized linear models
    - E.g., logistic, Poisson regression
  - Support Vector Machines (SVM)
  - Nearest neighbors
    - E.g., k-means clustering, KNN
  - Gaussian processes
  - Graphical models
    - Model joint distributions with graphs
    - E.g., hidden Markov models (HMM), Kalman filters, Bayesian networks
  - Neural networks
  - ...

Machine Learning Techniques

- What are the stages of a ML pipeline?
  - **Input processing**
    - Data cleaning
    - Dimensionality reduction
    - Feature engineering
  - **Model building**
    - Models
    - Learning algorithms
  - **Performance evaluation**
    - Cross-validation
    - Bias-variance curves
    - Learning curves
  - **Regularization**
  - **Aggregation**
    - Boosting
    - Bagging
    - Stacking

What Is Artificial Intelligence

AI

- A Map of Machine Learning
- What Is Artificial Intelligence
  - **AI**
  - Machine Learning
  - AI vs ML vs Deep-Learning
  - The Foundation of AI
  - Brief History of AI
  - AI State of the Art
  - Risks and Benefits of AI

Human Intelligence

- We call ourselves "homo sapiens" because intelligence sets us apart from animals
- For thousands of years, we tried to understand how we think
  - Brain is a small mass of matter
  - How can brain perceive, understand, predict, and manipulate a world more complicated than itself?

Artificial Intelligence

- The term "Artificial Intelligence" was coined in 1956
- **AI aims to:**
  - Understand human intelligence
  - Create intelligent entities
- **AI is a technology**
  - Universal and applicable to any human activity and task
  - Its impact greater than any previous historical event
  - Currently generates trillions of dollars annually in revenue
  - Presents many unresolved problems
    - E.g., major concepts in physics might be established

AI Formal Definition

- AI is defined around **two axes**:
  - Thinking vs. Acting
  - Human vs. Rational (ideal performance)
- This leads to four possible definitions of AI as a machine that can:
  1. Think humanly
  2. Think rationally
  3. Act humanly
  4. Act rationally

AI Formal Definition: Quiz

- Four possible definitions of AI as a machine that can:
  1. Think humanly
  2. Think rationally
  3. Act humanly
  4. Act rationally
- **Which one do you think is the best definition?**

AI Formal Definition: Quiz

- Four possible definitions of AI as a machine that can:
  1. Think humanly
  2. Think rationally
  3. Act humanly
  4. Act rationally
- **Which one do you think is the best definition?**
- We will see that building machines that can **"act rationally"** should be ultimate goal of AI

1. AI as Thinking Humanly

- To build machines that think like humans we need to **determine how humans think**
- **Pros**
  - Express precise theory of the human mind as a computer program
- **Cons**
  - Unknown workings of the human mind
  - Anthropocentric definition

2. AI as Thinking Rationally

- What are the rules of "correct thinking"?
    - Given correct premises, yield correct conclusions
- **Logic** studies the "laws of thought"
    - Formalize statements about objects and their relations
- **Automatic theorem proving**
    - Programs solve problems in logical notation
    - Run indefinitely if no solution exists (related to the halting problem)

Thinking Rationally: Cons

1. **Formalizing informal knowledge is difficult**
   - Example: *"A handshake occurs when two people extend, grip, shake hands, then release."*
   - Formal logic representation:

$$
\begin{aligned}
&\exists x, y\,(\text{Person}(x) \land \text{Person}(y) \land x \neq y \land \\
&\quad \text{Hand}(x, h_x) \land \text{Hand}(y, h_y) \land \\
&\quad \text{MoveToward}(h_x, h_y) \land \text{Contact}(h_x, h_y) \land \\
&\quad \text{Shake}(h_x, h_y) \land \\
&\quad \text{Release}(h_x, h_y))
\end{aligned}
$$

2. **Probabilistic nature of knowledge**
   - Example in medicine: *"Fever, cough, and fatigue could indicate flu, COVID-19, or another illness."*
3. **Scalability challenges**
   - Large problems may need heuristics for practical solutions
4. **Intelligence requires more than rational thinking**
   - Importance of agent interaction with the world
   - Problem of the "body"

3. AI as Acting Humanly

- **Agent** is something that perceives and acts to reach a goal
  - AI designs agent that can act like humans
- **(Embodied) Turing test**
  - *"A computer passes the Turing test if a human cannot tell whether the answers to questions came from a person or a computer"*
  - Passing the Turing test requires:
    1. Natural language processing to communicate in English
    2. Knowledge representation to store what it knows
    3. Automated reasoning to use stored knowledge to answer questions
    4. Machine learning to detect and extrapolate patterns
    5. Computer vision and speech recognition to perceive objects and understand human talking to them
    6. Robotics to manipulate objects and move around

Turing Test: Pros and Cons

- **Pros**
  - Operational definition of intelligence
  - Sidestep philosophical vagueness of "What is consciousness? Can a machine think? . . . "
- **Cons**
  - Anthropomorphic criteria define intelligence in human terms
    - Multiple forms of non-human intelligence exist
  - Intelligence in terms of Turing test
    - Design intelligence to imitate humans
    - Fool humans into thinking it's human
  - E.g., aeronautical engineering:
    - Focus on wind tunnels and aerodynamics
    - Not about designing machines that imitate birds
    - Not about fooling birds into thinking it's a bird

4. AI as Acting Rationally

- Rational agents $\iff$ do the "right thing" given what they know
- Computer agents that act rationally should:
    1. Operate autonomously
    2. Perceive environment
    3. Persist over a prolonged time period
    4. Adapt to change
    5. Create and pursue goals

Acting Rationally as Ultimate Goal of AI

- Which definition of AI to use?
  - Acting vs. Thinking
  - Rational vs. Human
- **Acting > Thinking**
  - Acting rationally is broader than just thinking rationally
- **Rational > Human**
  - Rationality can be mathematically defined
  - Human behavior is shaped by evolutionary conditions
- AI focuses on **agents acting rationally**
  - I.e., "agents that do the right thing" based on available knowledge
  - E.g., you leave the house and a branch strikes you
    - Did you act rationally?
    - Probably
  - E.g., you cross the street and a car knocks you over
    - Did you act rationally?
    - Id depends!

Goals of a Rational Agent

- A **rational agent** aims for:
  - The best outcome in a deterministic setup
  - The best expected outcome under uncertainty
- **Best** is determined by the objective function:
  - E.g., cost function, sum of rewards, loss function, utility
- **Problems** with acting rationally
  - Sometimes no provably correct action exists
    - Yet, an action must be taken
  - Perfect rationality is not always feasible due to:
    - Cost of acquiring all data
    - Computational demands
  - Acting "appropriately", "satisficing"

Machine Learning

Machine Learning: Definitions

- Artificial intelligence differs from human intelligence
  - Machines don't learn like humans (e.g., LLMs)
- "*Machine learning is the field of study that gives computers the ability to learn without being explicitly programmed*", (Arthur Samuel, 1959)
- "*A computer program is said to learn from experience E with respect to some task T and some performance measure P, if $P(T)$ improves with experience E*", (Tom Mitchell, 1998)
- Machine learning is the science of building machines capable of doing useful things without being explicitly programmed to do so
  - E.g. a computer learn to play checkers by playing against itself, memorizing which positions lead to winning a game

Limits of ML Compared to Human Intelligence

- **Fragility to input variations**
  - ML models fail with slight input distortions
    - E.g., adversarial attacks cause misclassification by altering one pixel
    - E.g., a model trained for a video game may fail if the screen is slightly rotated; humans continue effortlessly
- **Lack of transfer learning**
  - ML systems cannot apply knowledge across domains without retraining
- **Massive data and compute requirements**
  - ML requires enormous datasets and computational resources
    - E.g., a teenager learns to drive in hours; self-driving systems need billions of compute hours and extensive data
- **Poor common sense and reasoning**
  - ML lacks built-in world knowledge and intuitive logic

Limits of ML Compared to Human Intelligence

- **Opaque decision-making**
  - Many ML models offer little transparency into decision processes
    - Limits trust, interpretability, and accountability in critical applications
- **Dependence on narrow objectives**
  - ML systems excel at optimizing narrow tasks but fail with ambiguous goals
    - E.g., an algorithm maximizing user engagement may promote harmful content
- **Susceptibility to bias and data quality**
  - Models inherit and amplify biases in training data
- **Lack of embodiment and physical interaction**
  - Human cognition is grounded in physical and sensory experience

The 3 Machine Learning Assumptions

- Machine learning involves solving a practical problem by:
  - Gathering a dataset
  - Building a statistical model from the dataset algorithmically
- The **three assumptions** of machine learning
  - A pattern exists
  - Pattern cannot be precisely defined mathematically
  - Data is available
- Which ML assumption is essential?
  - *A pattern exists*
    - If no pattern, try learning, measure effectiveness, conclude it doesn't work
  - *Cannot pin down pattern mathematically*
    - If solution is direct, ML not recommended, but may still apply
  - *We have data*
    - Without data, no progress can be made
    - Data is crucial

Machine Learning Adages

- *"An explanation of the data should be as simple as possible, but not simpler"* (Einstein)
- *"The simplest model that fits the data is also the most plausible"* (Occam's razor)
- *"Garbage in, garbage out"* (Fuechse, 1957)
- *"All models are wrong, but some are useful"* (George E. P. Box, 1976)
- *"If you torture the data long enough it will confess whatever you want"* (Coase, 1982)
- *"Data is the new oil"* (Humby, 2006)
- *"More data beats clever algorithms"* (Norvig, ~2006)
- *"The unreasonable effectiveness of data"* (Halevy, Norvig, Pereira, 2009)

AI vs ML vs Deep-Learning

AI vs ML vs Deep-Learning

- **AI**: Machines programmed to reason, learn, and act in a rational way
- **ML**: Machines capable of performing tasks without being explicitly programmed, e.g.,
  - Natural language processing
  - Computer vision
  - Speech recognition
- **AI without ML**:
  - Example: Rule-based systems (e.g., IBM Deep Blue playing chess)
- **Deep Learning (DL)**: ML using neural networks with multiple layers
  - Example: Autonomous vehicles
- **Large Language Models (LLM)**: neural networks trained on massive text datasets and RLHS

The Foundation of AI

AI Relates to Many Other Disciplines

AI and Philosophy (1/2)

- **Can formal rules be used to draw valid conclusions?**
  - Reasoning
    - Aristotle (400 BC) formulated laws governing the rational mind
    - Machines were built for arithmetic operations (e.g., Pascaline, 1600)
    - Logic studies rules of proper reasoning
  - Rationalism
    - Use reasoning to understand the world
- **How does the mind arise from a physical brain?**
  - Dualism
    - Nature follows physical laws
    - Part of the human mind ("the soul") is exempt from physical laws
  - Materialism
    - The mind is a physical system, following the laws of physics
    - Where is free will? Free will is the perception of available choices

AI and Philosophy (2/2)

- **What does knowledge come from?**
  - Empiricism
    - Knowledge via senses
    - Example: learn that trees are green by looking at them
  - Induction
    - General rules from associations
    - Example: seeing many swans are white, inferring all swans are white
  - Logical Positivism
    - Knowledge as logical theories linked to sensory observations
    - Example: scientific hypotheses connected to experimental data
- **How does knowledge lead to action?**
  - Utilitarianism
    - Measures "utility" linking knowledge to action
    - Actions justified by logic connecting goals and outcomes
  - Consequentialism
    - Right or wrong determined by action's expected outcomes
    - E.g., "If you kill, you will go to jail"
  - Deontological ethics
    - Opposes consequentialism
    - "Right actions" based on universal laws, not outcomes
    - E.g., "don't kill", "don't lie"

AI and Cognitive Psychology

How do humans think and act?

- **Cognitive Psychology**
  - Brain as an information-processing device
    - Stimuli translated into internal representation
    - Representation manipulated by cognitive processes to derive new internal representations ("beliefs")
    - Representations turned into actions ("goals")
- **Cognitive Science**
  - Use computer models to address memory, language, and logic thinking
- **Human-Computer Interaction (HCI)**
  - From artificial intelligence (AI) to intelligence augmentation (IA)
  - Computers augment human abilities

AI and Mathematics

What are the formal rules to draw valid conclusions?

- **Formal Logic**
  - Boole established logical deduction rules (1850)
  - Frege expanded Boole's logic to include objects and relations, creating first-order logic (1879)
- **Limits to Deduction**
  - Some statements are "undecidable."
  - Godel's incompleteness theorem (1931): True statements exist that cannot be proved in any formal theory

How do we reason with uncertain information?

- **Probability**
  - Mathematics of uncertainty
  - Key contributors: Cardano, Pascal, Bernoulli, Bayes (1500-1700)
- **Statistics**
  - Combines data with probability
  - Key areas: experiment design, data analysis, hypothesis testing, asymptotics

AI and Computer Science

What can be computed?

- **Algorithm**
  - A procedure to solve problems
  - Example: Euclid's algorithm for computing GCD
- **Limits to Computation**
  - Turing machine (1936): Can compute any computable function
  - Some functions are non-computable, e.g., the halting problem—deciding if a program terminates
- **Tractability**
  - A problem is intractable if solving time grows exponentially with problem size
  - Complexity classes: Polynomial vs. exponential complexity (e.g., NP-problems)

AI and Economics (1/2)

How to make decisions to maximize payoff according to our preferences?

- **Economies**
  - Agents maximize economic well-being (utility)
  - Studies desires and preferences
  - "Large" vs "small economies"
- **Decision theory**
  - Making decisions under uncertainty
  - Probability theory + utility theory
  - Study choices for preferred outcomes
  - Examples: Investment choices, policy decisions

How to make decisions when the payoffs are result of several actions?

- **Operations Research**
  - Make rational decisions with payoffs from a sequence of actions
    - E.g., Markov Decision Processes
  - Bellman, 1957
- **Satisficing**
  - Decisions that are good enough
  - Closer to human behavior
  - Example: Choosing a restaurant that meets basic criteria rather than finding the perfect one

AI and Economics (2/2)

How to act when multiple agents with different goals are present?

- **Large Economies**
  - Agents ignore other agents' actions
  - Many agents with no mutual impact
  - Example: National economy where individual actions don't affect overall market
- **Small Economies**
  - One player's actions influence others' utility
  - Example: Local market where one seller's pricing affects competitors
- **Game Theory**
  - Von Neumann, 1944
  - Small economies resemble a "game"
  - Rational agents might need randomized strategies
  - Example: Rock-paper-scissors where randomization prevents predictability

AI and Linguistics

How can we create systems that understand natural language?

- Computational linguistics (aka NLP) studies sentence structure and meaning
  - Structure & Meaning are central to understanding language
  - NLP Applications:
    - Machine translation (e.g., Google Translate)
    - Sentiment analysis in social media
    - Automated customer support chatbots

How does language relate to thought?

- Knowledge representation studies how to represent knowledge in a form that a computer can reason about
- E.g., first order knowledge, knowledge graphs

AI and Neuroscience

- **Brain**
  - Parts of the brain handle specific cognitive functions
  - Information processing occurs in the cerebral cortex (outer brain layer)
  - E.g., injury to the frontal lobe may impair decision-making abilities
- **Anatomy of the Brain**
  - Composed of neurons ($\sim$100 billion)
    - Each neuron connects with 10-100k others via synapses
    - Axons facilitate long-range neuron connections
  - Signals propagate through electrochemical reactions
  - Short-term pathways support long-term brain connections, linked to learning
  - We can record and stimulate individual neuron activity
- **Memory**
  - No theory on individual memory storage
  - Current theory: Memories are reconstructed

The Brain Causes the Mind

- **Truly amazing conclusion**: *a collection of simple cells can lead to thought and consciousness*
  - E.g., neurons collectively create complex processes
  - Complexity of supercomputers is comparable or superior to the brain
  - Unknown how to achieve the brain's intelligence level
- **Brain-Machine Interface**: the brain adjusts to interface with devices
  - E.g., the brain learns to use prosthetics as a limb
- **AI Singularity**
  - A (hypothetical) future point when artificial intelligence surpasses human intelligence
    - AI systems could improve themselves autonomously, leading to rapid, exponential growth in capabilities
    - Recursive self-improvement leads to superintelligence
  - Potential for profound societal impact
    - Control problem / value alignment: ensuring superintelligent AI aligns with human values
    - Economic and social disruption due to automation
  - Hard to predict

AI and Computer Engineering

How can we build an efficient computer?

- **Electronic computers**
  - Built during World War II
- **Moore's Law**
  - Performance doubled every 18 months until 2005
  - Power and scaling issues shifted focus to core multiplication over clock speed
- **Hardware for AI**
  - GPUs
  - TPUs
  - Wafer-scale engines
- **Current Trends**
  - Massive parallelism (like brain function)
    - Computing power doubling every 3 months
    - GPUs / TPUs used in deep learning
  - High precision (e.g., 64b) often unnecessary
- **Quantum Computing**
  - Potential for significant acceleration in key computations
  - E.g., Shor's algorithm for factorization

AI and Control Theory and Cybernetics

How can artifacts operate under their own control?

- Control theory
  - Study self-regulating feedback control systems
    - E.g., a water regulator that maintains a constant water flow
    - E.g., steam engine, thermostat
  - Mechanisms to minimize error between current and goal states
- Control theory vs AI
  - Similar goals, but different techniques to achieve them
  - Control theory:
    - Calculus
    - Matrix
    - Stochastic optimal control
  - AI:
    - Logical inference
    - Symbolic planning
    - Computation

Brief History of AI

The Beginning (1943-1956)

- McCullock-Pitts artificial neuron
  - Proposed a model of artificial neuron (1943), based on:
    - Basic physiology of the brain
    - Propositional logic
    - Theory of computation
  - Any computable function can be computed by a network of connected neurons
    - Neuron is on or off depending on the stimulation from neighboring neurons
    - Logical AND, OR, NOT can be implemented with simple networks of neurons
- Alan Turing, 1947
  - Introduced Turing test, machine learning, reinforcement learning
  - To create human-level AI:
    - Develop learning algorithms
    - Teach the machine like a child
- Birth of AI
  - McCarthy organized in US the first workshop about AI (1956)
  - Newell and Simon (1956)
    - The Logic Theorist
    - Programs able to “think non-numerically” and prove theorems

Early Enthusiasm, Great Expectations (1952-1969)

- Early years of AI were full of successes
- Until then computers could only do arithmetics
- "A machine can never do X" where $X =$ games, puzzles, mathematics, IQ tests
  - AI researchers demonstrated that machines could do one X after another
- General Problem Server, successor of Logic Theorist
  - Imitate human problem-solving
  - Consider sub-goals and possible actions
- Program that learned to play checkers and became better than its creator
  - Use reinforcement learning by learning from victories and mistakes in gameplay
- Lisp (1958)
  - High-level language that was used for next 30 years in AI
- Marvin Minsky (1959)
  - Built first neural network
  - 3000 vacuum tubes to implement 40 neurons
- MIT and Stanford
  - Minsky at MIT
    - Focus on neural network

A Dose of Reality (1966-1973)

- AI researchers were confident about AI's upcoming successes
- In reality, AI didn't succeed on real problems due to several reasons:
  - AI solutions were initially based on human problem-solving methods
  - Difficulty in handling "combinatorial explosion" from small to real-world problems:
    - E.g., theorem proving can handle small problems with brute force, but doesn't scale for larger problems
    - E.g., genetic programming suggested random small mutations could generate programs for any task, but this demands enormous CPU power
  - The neural network approach required algorithms (e.g., backpropagation), compute power, and data to work effectively

Expert Systems (1969-1979)

- Weak AI
  - In the first wave of AI research, the goal was a general-purpose search mechanism trying to string elementary reasoning steps to find complete solution
  - These "weak" methods are general and don't scale up to large problems
  - The solution is to add domain knowledge $\rightarrow$ expert systems
- Expert systems
  - Aka "knowledge-based systems"
  - Add domain knowledge in the form of rules
  - E.g., Prolog
- AI became an industry (1980-)
  - Every major US corporation was trying to adopt expert systems

(First) AI Winter (1980)

- AI overconfidence/hype didn't deliver
- **Reasons:**
  - Building/maintaining expert systems is difficult
  - Reasoning methods ignore uncertainty
  - Systems can't learn from experience
  - E.g., expert systems in medical diagnosis struggle with complex, variable patient data
- Early AI chess systems couldn't adapt to new strategies without manual updates

Return of Neural Networks (1986-)

- Mid-1980s: Researchers discovered back-propagation algorithm
  - Developed in early 1960s
  - Example: Neural networks learning from data
- **Connectionist models vs. Symbolic models**
  - Connectionist: Neural networks
    - Example: Recognizing handwritten digits
  - Symbolic: General Problem Solver
    - Example: Solving logical puzzles with explicit rules
- **Why connectionist models**
  - Many concepts are not well-defined using symbolic axioms
    - Connectionist approach forms fluid internal concepts
    - Represents real-world complexity better
  - Neural networks learn from examples
    - Adjust parameters for improved predictions
  - E.g.,
    - Image recognition: Neural networks identify objects by learning from labeled images
    - Language models: Predict next words by learning from text data

Probabilistic Reasoning and ML (1987-)

- **AI and Scientific Method**
  - Rigorous methods to test performance
  - E.g., speech recognition, handwritten character recognition
- **Benchmarks for Progress**
  - Examples:
    - MNIST: Handwritten digit recognition
    - ImageNet: Image object recognition
    - SAT Competitions: Boolean satisfiability solvers
- **AI Shifts**
  - From Boolean logic to probability
  - From hand-coded rules to machine learning
  - From a-priori reasoning to experimental results

Progress in Speech Recognition

- **1970s: Various architectures and approaches were attempted**
  - Rule-based systems with limited robustness
  - Cons: Ad-hoc, fragile
- **1980s: Hidden Markov Models (HMMs) became dominant**
  - Pros: Strong theoretical foundation
  - Methods: Effective learning techniques
  - Data: Trained on large speech corpora
  - No claim humans use HMMs for speech recognition

Bayesian Networks

- In 1988 Judea Pearl linked AI with:
  - Probability
  - Decision theory
  - Control theory
- **Bayesian networks:**
  - Efficiently represent uncertainty
  - Provide rigorous reasoning
  - Enable practical reasoning
  - Handle uncertainty
- E.g.,
  - Diagnosing diseases based on symptoms
  - Predictive text input in smartphones
  - Fraud detection in banking

Reinforcement Learning

- **1988: Sutton worked on reinforcement learning and Markov Decision Processes (MDPs)**
  - Reinforcement Learning (RL) involves agents learning by interacting with an environment
  - MDPs provide a mathematical framework for modeling decision-making
- E.g.,
  - Reinforcement Learning: A robot learning to navigate a maze by receiving rewards for successful paths
  - MDPs: A game strategy modeled where each move influences the outcome with certain probabilities

Reunification

- Reunification of AI with:
  - Data
  - Statistical modeling
  - Optimization
  - Machine learning
- Many subfields of AI were also re-unified
  - Computer vision
  - Robotics
  - Speech recognition
  - Multi-agent systems
  - NLP

Big Data (2001-Present)

- For 60 years, AI focused on algorithms and models
- For some problems, data availability matters more than algorithms, e.g.,
    - Trillions of English words
    - Billions of web images
    - Billions of speech and video hours
    - Social network data
    - Clickstream data
- Algorithms leverage large datasets
- In 2011, IBM's Watson beat human Jeopardy! champions
    - Shifted public's view of AI

Deep Learning (2011-Present)

- Deep Learning is ML models using multiple layers of computing elements
  - Ideas were already known in 1970s
  - Success in handwritten digit recognition in 1990s
- In 2011, DL took off
  - Surge of interest in AI among researchers, students, companies, investors, government, and the public
  - In 2012, a DL system showed dramatic improvement in the `ImageNet` competition
    - Previous systems used handcrafted features
  - Today, DL has exceeded human performance in several vision and speech recognition tasks
- DL needs to run on specialized hardware (e.g., GPU, TPU, FGPA) to perform highly parallel tensor operations
- General Artificial Intelligence
  - Universal algorithm for learning and acting, instead of specialized tasks (e.g., driving a car, playing chess, recognizing speech)

AI State of the Art

Progress in AI Research

- AI papers increased 20x (2010-2019)
  - From 1,000 in 2010 to 20,000 in 2019
- Student enrollment in AI and CS increased 5x
  - From 10,000 in 2010 to 50,000 in 2019
- NeurIPS attendance increased 8x
  - From 1,000 attendees to 8,000
- AI startups increased 20x
  - From 100 to 2,000 startups
- Training times dropped 100x in 2 years
  - AI computing power doubles every 3 months

What Can AI Do Today? (1/2)

- **Robotic vehicles**
  - Waymo passed 10 million miles without serious accident
- **Legged locomotion**
  - BigDog recovers on ice
  - Atlas walks on uneven terrain, jumps on boxes, backflips
- **Autonomous planning and scheduling**
  - Space probes, Mars rovers
- **Machine translation**
  - Translates 100 languages with human-level performance
- **Speech recognition**
  - Real-time speech-to-speech with human-level performance
  - AI assistants
- **Recommendations**
  - ML recommends based on past experiences
  - Spam filtering 99.9% accuracy
  - E.g., Amazon, Facebook, Netflix, Spotify, YouTube

What Can AI Do Today? (2/2)

- **Game playing**
  - 1997 Deep Blue defeated Kasparov
  - 2017 Watson beat Jeopardy! champion
  - 2017 AlphaGo beat Go champion (expected 100 years to beat humans in Go)
  - 2018 AlphaZero super-human in Go, chess with only rules, self-play
  - Videogames: Dota2, StarCraft, Quake
- **Image understanding**
  - Object recognition, Image captioning
- **Medicine**
  - AI equivalent to health care professionals
- When will AI systems achieve human-level performance across tasks?
  - Average of expert prediction is 2099
    - Papers have shown that predictions of experts are no better than amateurs
  - Unclear if need new breakthroughs or refinements on current approaches

Risks and Benefits of AI

Civilization and AI

- Our civilization is the product of human intelligence
  - Greater machine intelligence leads to higher ambitions for our civilization
  - *"First solve AI, then use AI to solve everything else"*
- **Benefits**
  - Free humanity from menial work
  - Increase the production of goods and services
  - Expand human cognition
  - Accelerate scientific research, e.g., cures for diseases, solutions for climate change, resource shortages)

Risks 1/2

- **Lethal autonomous weapons**
  - Locate, select, eliminate human targets without intervention
  - Scalability: deploy a large number of weapons
- **Surveillance and persuasion**
  - AI (speech recognition, computer vision, natural language understanding) for mass surveillance
  - Tailoring information flows through social media to modify behavior
- **Biased decision making**
  - Misuse of ML can result in biased decisions due to societal bias
  - E.g., parole evaluations, loan applications

Risks 2/2

- **Impact on employment**
  - Machines can eliminate jobs
  - Rebuttal
    - Machines enhance human productivity ->
    - Companies become more profitable ->
    - Higher wages
  - Counter-rebuttal
    - Wealth shifts from labor to capital, increasing inequality
  - Counter-counter-rebuttal
    - Past tech advances (e.g., mechanical looms) disrupted employment, but adaptation followed
- **Safety critical applications**
  - AI used in safety-critical applications
    - E.g., self-driving cars, managing water supply or power grids
  - Avoiding fatal accidents is challenging
    - E.g., formal verification and statistical analysis are insufficient
  - AI requires technical and ethical standards like other high-stakes fields (e.g., engineering, healthcare)
- **Cybersecurity**
  - AI helps defend against cyberattacks (e.g., detect unusual behavior patterns) and contributes to malware development
  - E.g., use reinforcement learning for targeted phishing attacks

Human-Level AI

- Human-level AI is "machines able to learn to do anything a human can do"
    - Aka AGI (Artificial General Intelligence)
- **Artificial Super-Intelligence**: Intelligence surpassing human ability in any domain and self-improving

The Problem of Control

- It is uncertain we can control machines more intelligent than us
- **King Midas problem**
  - Myths of humans asking for something, getting it, then regretting it
  - King Midas turned everything he touched into gold, including food and family
- **Rebuttal**
  - If AGI arrived in a black box from space, caution is needed before opening
  - We design AI: if AI gains control, it is a "design failure"

Solutions to Problem of Control

- AI researchers and corporations developed voluntary self-governance principles for AI
  - Governments and international organizations established advisory bodies
- **Problems**
  - Preferences are not easy to "invert" and are not consistent
- We should put "purpose into the machine" even if we don't know exactly what the objectives are
  - Incentivize AI to be switched off if uncertain about human objectives
  - Inverse reinforcement learning: AI observes human behavior to infer underlying reward function
  - Cooperative Inverse Reinforcement Learning (CIRL)

Cooperative Inverse Reinforcement Learning (CIRL)

- AI infers human goals based on actions
- **Observation**: Alice looks tired, sits on the couch, observes the messy table, and starts watching TV
- **Inference**: AI infers:
    - Alice is tired and wants to relax
    - Messy coffee table bothers her
- **Action**: AI:
    - Fetches a glass of water
    - Tidies up the coffee table without disturbing Alice
- **Feedback loop**: AI monitors Alice's reactions
    - If Alice is relaxed and happy, AI understanding is reinforced
    - If Alice is not happy, AI adjusts actions and improves inference