CSE 573 P: Artificial Intelligence

Hanna Hajishirzi

slides adapted from
Dan Klein, Pieter Abbeel ai.berkeley.edu
And Dan Weld, Luke Zettlemoyer
Topics in This Course

- **Part I: Intelligence from Computation**
  - Fast search
  - Adversarial and uncertain search

- **Part II: Reasoning under Uncertainty**
  - Decision theory: Reinforcement Learning, Markov Decision Processes
  - Machine learning
  - Graphical Models - Bayes Nets; HMMs

- **Throughout: Applications**
  - Natural language, vision, robotics, games, ...
Pac-Man Beyond the Game!
Pacman: Beyond Simulation?

Students at Colorado University: http://pacman.elstonj.com
Pacman: Beyond Simulation!

[VIDEO: Roomba Pacman.mp4]
Bugman?

- **AI = Animal Intelligence?**
  - Wim van Eck at Leiden University
  - Pacman controlled by a human
  - Ghosts controlled by crickets
  - Vibrations drive crickets toward or away from Pacman’s location

http://pong.hku.nl/~wim/bugman.htm
Bugman

'S' to start game
Use cursor keys to move

Score: 0

2Mega Camera, WDM Video Capture
Research Frontiers

- Deep Unsupervised Learning
- AI for Science
- AI and Ethics

Also:
- Unsupervised Deep Reinforcement Learning
- Human-in-the-loop Reinforcement Learning
- ...


Research Frontiers

- *Deep Unsupervised Learning*
- AI for Science
- AI and Ethics

Also:
- Unsupervised Deep Reinforcement Learning
- Human-in-the-loop Reinforcement Learning
- ...
Deep Unsupervised Learning

- Key hypothesis:
  - IF neural network smart enough to predict:
    - Next frame in video
    - Next word in sentence
    - Generate realistic images
    - "Translate" images
    - ...
  - THEN same neural network is ready to do Deep Supervised Learning from very small data-set
Transfer from Unsupervised Learning

Task 1 = unsupervised
Task 2 = real task
Example Setting

Task 1 = predict next word

Task 2 = predict sentiment
A train carriage containing controlled nuclear materials was stolen in Cincinnati today. Its whereabouts are unknown.

The incident occurred on the downtown train line, which runs from Covington and Ashland stations. In an email to Ohio news outlets, the U.S. Department of Energy said it is working with the Federal Railroad Administration to find the thief.

‘The theft of this nuclear material will have significant negative consequences on public and environmental health, our workforce and the economy of our nation,’ said Tom Hicks, the U.S. Energy Secretary, in a statement. ‘Our top priority is to secure the theft and ensure it doesn’t happen again.’

The stolen material was taken from the University of Cincinnati’s Research Triangle Park nuclear research site, according to a news release from Department officials.

The Nuclear Regulatory Commission did not immediately release any information. According to the release, the U.S. Department of Energy’s Office of Nuclear Material Safety and Security is leading that team’s investigation.

‘The safety of people, the environment and the nation’s nuclear stockpile is our highest priority,’ Hicks said. ‘We will get to the bottom of this and make no excuses.’
Prompt: Recycling is good for the world, no, you could not be more wrong.

GPT-2: Recycling is NOT good for the world. It is bad for the environment, it is bad for our health, and it is bad for our economy. I’m not kidding. Recycling is not good for the environment. It is destructive to the earth and it is a major contributor to global warming. Recycling is not good for our health. It contributes to obesity and diseases like heart disease and cancer. Recycling is bad for our economy. It increases the cost of a product, and in turn, the price of everything that is made with that product. Recycling is not good for our nation. We pay a tremendous price for the privilege of having the world’s most advanced and efficient recycling system. Recycling is a huge, colossal waste of time, energy, money, and resources.
OpenAI built a text generator so good, it’s considered too dangerous...

TechCrunch - 17 Feb 2019

OpenAI built a text generator so good, it’s considered too dangerous to release...

TechCrunch - 17 Feb 2019

Scientists Developed an AI So Advanced They Say It’s Too Dangerous...

ScienceAlert - 18 Feb 2019

Only releasing a smaller version of the language model, citing its...

ScienceAlert - 18 Feb 2019

AI text writing technology too dangerous to release, creators claim

The Drum - 17 Feb 2019

This technology could ‘absolutely devastate’ the internet as we know it

The Drum - 17 Feb 2019

This AI is so good at writing that its creators warn let you use it

NEWS.com.au - 17 Feb 2019

Lord of The Rings, Celebrity Gossip: This AI is So Good at Writing That...

In-Depth - CNH - 13 Feb 2019

When is Technology Too Dangerous to Release to the Public?

Slate Magazine - 22 Feb 2019

If you know nothing of the model, called GPT-2, came solely on headlines... U.K.

Slate Magazine - 22 Feb 2019

AI Weekly: Experts say OpenAI’s controversial model is a potential...

In-Depth - VentureBeat - 22 Feb 2019

OpenAI’s Text Model so Disruptive it’s Deemed Too Dangerous To Release... 

Computer Business Review - 16 Feb 2019

OpenAI’s Text Model so Disruptive it’s Deemed Too Dangerous To Release...

Computer Business Review - 16 Feb 2019

New AI fake text generator may be too dangerous to release, say...

Highly Cited - The Guardian - 14 Feb 2019

View all

When is Technology Too Dangerous to Release to the Public?

Slate Magazine - 22 Feb 2019

If you know nothing of the model, called GPT-2, came solely on headlines... U.K.

Slate Magazine - 22 Feb 2019

AI Weekly: Experts say OpenAI’s controversial model is a potential...

In-Depth - VentureBeat - 22 Feb 2019

OpenAI’s Text Model so Disruptive it’s Deemed Too Dangerous To Release... 

Computer Business Review - 16 Feb 2019

OpenAI’s Text Model so Disruptive it’s Deemed Too Dangerous To Release...

Computer Business Review - 16 Feb 2019

New AI fake text generator may be too dangerous to release, say...

Highly Cited - The Guardian - 14 Feb 2019

View all

When is Technology Too Dangerous to Release to the Public?

Slate Magazine - 22 Feb 2019

If you know nothing of the model, called GPT-2, came solely on headlines... U.K.

Slate Magazine - 22 Feb 2019

AI Weekly: Experts say OpenAI’s controversial model is a potential...

In-Depth - VentureBeat - 22 Feb 2019

OpenAI’s Text Model so Disruptive it’s Deemed Too Dangerous To Release... 

Computer Business Review - 16 Feb 2019

OpenAI’s Text Model so Disruptive it’s Deemed Too Dangerous To Release...

Computer Business Review - 16 Feb 2019

New AI fake text generator may be too dangerous to release, say...

Highly Cited - The Guardian - 14 Feb 2019

View all

When is Technology Too Dangerous to Release to the Public?

Slate Magazine - 22 Feb 2019

If you know nothing of the model, called GPT-2, came solely on headlines... U.K.

Slate Magazine - 22 Feb 2019

AI Weekly: Experts say OpenAI’s controversial model is a potential...

In-Depth - VentureBeat - 22 Feb 2019

OpenAI’s Text Model so Disruptive it’s Deemed Too Dangerous To Release... 

Computer Business Review - 16 Feb 2019

OpenAI’s Text Model so Disruptive it’s Deemed Too Dangerous To Release...

Computer Business Review - 16 Feb 2019

New AI fake text generator may be too dangerous to release, say...

Highly Cited - The Guardian - 14 Feb 2019

View all

When is Technology Too Dangerous to Release to the Public?

Slate Magazine - 22 Feb 2019

If you know nothing of the model, called GPT-2, came solely on headlines... U.K.

Slate Magazine - 22 Feb 2019

AI Weekly: Experts say OpenAI’s controversial model is a potential...

In-Depth - VentureBeat - 22 Feb 2019

OpenAI’s Text Model so Disruptive it’s Deemed Too Dangerous To Release... 

Computer Business Review - 16 Feb 2019

OpenAI’s Text Model so Disruptive it’s Deemed Too Dangerous To Release...

Computer Business Review - 16 Feb 2019

New AI fake text generator may be too dangerous to release, say...

Highly Cited - The Guardian - 14 Feb 2019

View all

When is Technology Too Dangerous to Release to the Public?

Slate Magazine - 22 Feb 2019

If you know nothing of the model, called GPT-2, came solely on headlines... U.K.

Slate Magazine - 22 Feb 2019

AI Weekly: Experts say OpenAI’s controversial model is a potential...

In-Depth - VentureBeat - 22 Feb 2019

OpenAI’s Text Model so Disruptive it’s Deemed Too Dangerous To Release... 

Computer Business Review - 16 Feb 2019

OpenAI’s Text Model so Disruptive it’s Deemed Too Dangerous To Release...

Computer Business Review - 16 Feb 2019

New AI fake text generator may be too dangerous to release, say...

Highly Cited - The Guardian - 14 Feb 2019

View all
BERT and Family

Different Variations on Transformer architectures and different pre-training tasks
<table>
<thead>
<tr>
<th>DATASET</th>
<th>METRIC</th>
<th>OUR RESULT</th>
<th>PREVIOUS RECORD</th>
<th>HUMAN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Winograd Schema Challenge</td>
<td>accuracy (+)</td>
<td>70.70%</td>
<td>63.7%</td>
<td>92%+</td>
</tr>
<tr>
<td>LAMBADA</td>
<td>accuracy (+)</td>
<td>63.24%</td>
<td>59.23%</td>
<td>95%+</td>
</tr>
<tr>
<td>LAMBADA</td>
<td>perplexity (-)</td>
<td>8.6</td>
<td>99</td>
<td>~1-2</td>
</tr>
<tr>
<td>Children’s Book Test</td>
<td>accuracy (+)</td>
<td>93.30%</td>
<td>85.7%</td>
<td>96%</td>
</tr>
<tr>
<td>Children’s Book Test</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Common Nouns</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(validation accuracy)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Children’s Book Test</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Named Entities</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(validation accuracy)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Penn Tree Bank</td>
<td>perplexity (-)</td>
<td>35.76</td>
<td>46.54</td>
<td>unknown</td>
</tr>
<tr>
<td>WikiText-2</td>
<td>perplexity (-)</td>
<td>18.34</td>
<td>39.14</td>
<td>unknown</td>
</tr>
</tbody>
</table>
Scaling
Pretrained Models (BERT) on GLUE Benchmarks

Human Performance on GLUE
Massive Pre-trained models are few-shot learners! (GPT-3)

175B GPT-3 can work without fine-tuning, when it is shown sample **demonstrations** for a task:

**Few-shot**

In addition to the task description, the model sees a few examples of the task. No gradient updates are performed.

1. Translate English to French:
2. sea otter => loutre de mer
3. peppermint => menthe poivrée
4. plush giraffe => girafe peluche
5. cheese => ........................................

![Figure 1: Exponential growth of number of parameters in DL models](image)
Unsupervised Learning in Vision

- Task 1: fill in a patch
- Task 2: predict cat vs. dog
Predict Missing Patch
SimCLR + linear classifier
AI for Art Creation

humanoid robot Mona Lisa
artstationHQ

studio ghibli trending on artstation | vary
Text-Guided Image Generation

Latent vector → some generative model (e.g. BigGAN or StyleGAN) → Output image: (this is a GIF of the sequence of images found over the course of optimization) → CLIP image embedder → embedding similarity measurement (we want to find an image that maximizes this) → CLIP text embedder → "Starry Night"
Examples (CLIP + VQGAN)
Research Frontiers

- Deep Unsupervised Learning
- AI for Science
- AI and Ethics

Also:
- Unsupervised Deep Reinforcement Learning
- Human-in-the-loop Reinforcement Learning
- ...
‘It will change everything’: DeepMind’s AI makes gigantic leap in solving protein structures

Google’s deep-learning program for determining the 3D shapes of proteins stands to transform biology, say scientists.

Ewen Callaway
T1037 / 6vr4
90.7 GDT
(RNA polymerase domain)

T1049 / 6y4f
93.3 GDT
(adhesin tip)
AI Can Now Model the Molecular Machines That Govern All Life

Just last year, DeepMind shocked the biomedical field with AlphaFold, an algorithm that predicts protein structures with jaw-dropping...

1 week ago

Nature

Artificial intelligence powers protein-folding predictions

DeepMind's software — which uses the sophisticated machine-learning ... algorithms that outperformed the first generation of AlphaFold,...

4 days ago

The Guardian

Yes, DeepMind crunches the numbers — but is it really a magic bullet?

The bigger deal was that DeepMind, a London-based offshoot of Google ... In 2018, DeepMind's AlphaFold machine-learning software entered the...

2 weeks ago

Fortune

DeepMind spins out new Alphabet company focused on drug discovery

DeepMind, the London-based artificial intelligence research company ... DeepMind has already made its AlphaFold system freely available for...

3 weeks ago

TechStory

DeepMind's AI solved a 50-year-old protein-related challenge

We are talking about DeepMind's AlphaFold, the latest iteration, a deep-learning system that can accurately predict the structure of...

4 weeks ago

C&EN

Software predicts more complex protein structures

... RoseTTAFold algorithm and DeepMind's AlphaFold software to directly predict the structures of protein complexes in a model eukaryote,...

2 weeks ago

Genetic Engineering and Biotechnology News

AI Systematically IDs Structures of Eukaryotic Proteins

AlphaFold, which was invented by the Alphabet subsidiary DeepMind, was used to re-evaluate interaction probability and model complex structures.

3 days ago

EurekAlert!

Scientists build on AI modelling to understand more about ...
Thrilled to announce the launch of a new Alphabet company @IsomorphicLabs. Our mission is to reimagine the drug discovery process from first principles with an AI-first approach, to accelerate biomedical breakthroughs and find cures for diseases. Details: isomorphiclabs.com/blog
Symbolic Math: Integrals and ODEs

<table>
<thead>
<tr>
<th>Equation</th>
<th>Solution</th>
</tr>
</thead>
<tbody>
<tr>
<td>[ y' = \frac{16x^3 - 42x^2 + 2x}{(-16x^8 + 112x^7 - 204x^6 + 28x^5 - x^4 + 1)^{1/2}} ]</td>
<td>[ y = \sin^{-1}(4x^4 - 14x^3 + x^2) ]</td>
</tr>
<tr>
<td>[ 3xy \cos(x) - \sqrt{9x^2 \sin(x)^2 + 1}y' + 3y \sin(x) = 0 ]</td>
<td>[ y = c \exp \left( \sinh^{-1}(3x \sin(x)) \right) ]</td>
</tr>
<tr>
<td>[ 4x^4 y y'' - 8x^4 y' - 8x^3 y y' - 3x^3 y'' - 8x^2 y^2 - 6x^2 y' - 3x^2 y'' - 9xy' - 3y = 0 ]</td>
<td>[ y = \frac{c_1 + 3x + 3 \log(x)}{x(c_2 + 4x)} ]</td>
</tr>
</tbody>
</table>

Table 4: Examples of problems that our model is able to solve, on which Mathematica and Matlab were not able to find a solution. For each equation, our model finds a valid solution with greedy decoding.

[Lample and Charton, ICLR 2020]
Symbolic Math: Integrals and ODEs

<table>
<thead>
<tr>
<th></th>
<th>Integration (BWD)</th>
<th>ODE (order 1)</th>
<th>ODE (order 2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mathematica (30s)</td>
<td>84.0</td>
<td>77.2</td>
<td>61.6</td>
</tr>
<tr>
<td>Matlab</td>
<td>65.2</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Maple</td>
<td>67.4</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Beam size 1</td>
<td>98.4</td>
<td>81.2</td>
<td>40.8</td>
</tr>
<tr>
<td>Beam size 10</td>
<td>99.6</td>
<td>94.0</td>
<td>73.2</td>
</tr>
<tr>
<td>Beam size 50</td>
<td>99.6</td>
<td>97.0</td>
<td>81.0</td>
</tr>
</tbody>
</table>

[Lample and Charton, ICLR 2020]
Research Frontiers

- Deep Unsupervised Learning
- AI for Science
- AI and Ethics

Also:
- Unsupervised Deep Reinforcement Learning
- Human-in-the-loop Reinforcement Learning
- ...

AI Ethics Ever More Important

Why?
- AI is making decisions, at scale
- Any kind of issues (e.g. bias or malignant use) could significantly affect *people*

Many open questions:
- Who is responsible?
- How to diagnose and prevent?
Some Key AI Ethics Topics

- Disinformation
- Bias and fairness
- Privacy and surveillance
- Metrics
- Algorithmic colonialism

Source: Rachel Thomas (@math_rachel)
Prompt: Recycling is good for the world, no, you could not be more wrong.

GPT-2: Recycling is NOT good for the world. It is bad for the environment, it is bad for our health, and it is bad for our economy. I’m not kidding. Recycling is not good for the environment. It is destructive to the earth and it is a major contributor to global warming. Recycling is not good for our health. It contributes to obesity and diseases like heart disease and cancer. Recycling is bad for our economy. It increases the cost of a product, and in turn, the price of everything that is made with that product. Recycling is not good for our nation. We pay a tremendous price for the privilege of having the world’s most advanced and efficient recycling system. Recycling is a huge, colossal waste of time, energy, money, and resources.
### Bias and Fairness – Gender Shades

<table>
<thead>
<tr>
<th>Gender Classifier</th>
<th>Darker Male</th>
<th>Darker Female</th>
<th>Lighter Male</th>
<th>Lighter Female</th>
<th>Largest Gap</th>
</tr>
</thead>
<tbody>
<tr>
<td>Microsoft</td>
<td>94.0%</td>
<td>79.2%</td>
<td>100%</td>
<td>98.3%</td>
<td>20.8%</td>
</tr>
<tr>
<td>FACE++</td>
<td>99.3%</td>
<td>65.5%</td>
<td>99.2%</td>
<td>94.0%</td>
<td>33.8%</td>
</tr>
<tr>
<td>IBM</td>
<td>88.0%</td>
<td>65.3%</td>
<td>99.7%</td>
<td>92.9%</td>
<td>34.4%</td>
</tr>
</tbody>
</table>

Joy Buolamwini & Timnit Gebru, gendershades.org

Source: Rachel Thomas (@math_rachel)
Bias and Fairness – Biased Ads

Source: Rachel Thomas (@math_rachel)
Bias and Fairness

Algorithms are used differently than human decision makers

- People are more likely to assume algorithms are objective or error-free
- Algorithms are more likely to be implemented with no appeals process
- Algorithms are often used at scale
- Algorithmic systems are cheap

Source: Rachel Thomas (@math_rachel)
Privacy and Surveillance

Your Apps Know Where You Were Last Night, and They’re Not Keeping It Secret

In about four months’ of data reviewed by The Times, her location was recorded over 8,600 times — on average, once every 21 minutes.

Ms. Magrin’s location data shows other often-visited locations, including the gym and Weight Watchers.

Source: Rachel Thomas (@math_rachel)
I stopped doing CV research because I saw the impact my work was having. I loved the work but the military applications and privacy concerns eventually became impossible to ignore.
Goodhart’s Law:

When a measure becomes a target, it ceases to be a good measure.

Source: Rachel Thomas (@math_rachel)
Flawed Algorithms Are Grading Millions of Students’ Essays

Foiled by gibberish and highly susceptible to human bias, automated essay-scoring systems are being increasingly adopted, a Motherboard investigation has found.

- Automatic essay grading software used in at least 22 USA states
- Focuses on metrics like sentence length, vocabulary, spelling, subject-verb agreement
- Can’t evaluate hard-to-quantify qualities, like creativity
- Gibberish essays with lots of sophisticated words score well
- Essays by African-American students receive lower grades from computer than from expert human graders
- Essays by students from mainland China receive higher scores from computer than from expert human graders; may be using chunks of pre-memorized text

Source: Rachel Thomas (@math_rachel)
How Algorithms Can Learn to Discredit the Media

Defamation is efficient, and AIs may have already figured it out.

Als Are Designed to Maximize Watch Time

At YouTube, we used a complex AI to pursue a simple goal: maximize watch time. Google explains this focus in the following statement:

If viewers are watching more YouTube, it signals to us that they’re happier with the content they’ve found. It means that creators are attracting more engaged audiences. It also opens up more opportunities to generate revenue for our partners.

How an ex-YouTube insider investigated its secret algorithm

Source: Rachel Thomas (@math_rachel)
Where to Go Next?
Where to go next?

- Congratulations, you’ve seen the basics of modern AI
  - ... and done some amazing work putting it to use!

- How to continue:
  - Machine learning:
  - Data Science:
  - Data / Ethics:
  - Probability:
  - Optimization:
  - Computer vision:
  - Reinforcement Learning:
  - Robotics:
  - NLP:
  - ... and more; ask if you’re interested
That’s It!

- Help us out with some course evaluations
- Have a great spring break