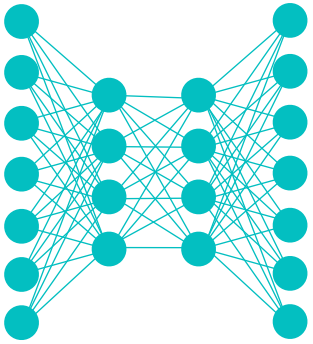
Lecture Notes for

Neural Networks and Machine Learning

Course Introduction



Logistics and Agenda

- Logistics
  - Second Course Offering
  - Canvas Access?
- Agenda
  - Introductions
  - Syllabus
  - Presentation Selection

Introductions

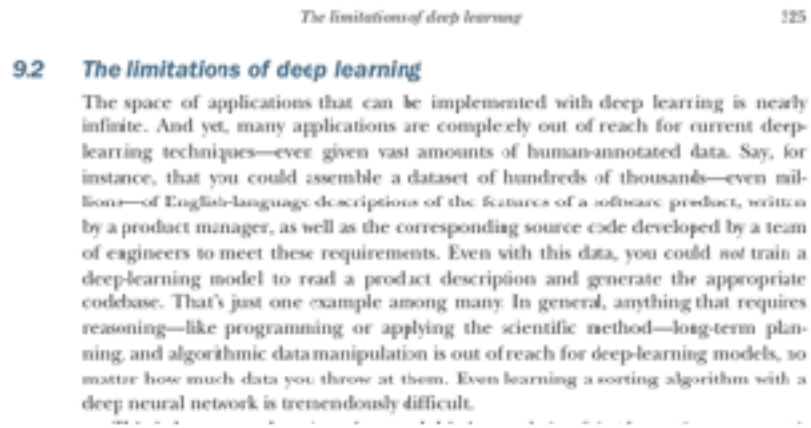
- Name
- Department
- Home
- 2 Truths and 1 Falsehood
  - Example: I gave Pitches on Machine Learning to Elon Musk, Bill Gates, and Jeff Bezos

Syllabus

- Reading
- GitHub
- Grading
- Participation
- Course Schedule

Presenting

- First Presentation is Next Week!
- During Semester: Eight Presentations Total
- First Presentation:
  - Section 9.2 of Chollet Book, Limitations of DL
- **Who wants to go first?**
  - 15 Minutes
  - Summarize the Article
  - Make 2-5 Visuals
    - Slides
    - Handouts
    - Notebooks



The limitations of deep learning 125

9.2 The limitations of deep learning

The space of applications that can be implemented with deep learning is nearly infinite. And yet, many applications are completely out of reach for current deep-learning techniques—even given vast amounts of human-annotated data. Say, for instance, that you could assemble a dataset of hundreds of thousands—even millions—of English-language descriptions of the features of a software product, written by a product manager, as well as the corresponding source code developed by a team of engineers to meet these requirements. Even with this data, you could *not* train a deep-learning model to read a product description and generate the appropriate codebase. That's just one example among many. In general, anything that requires reasoning—like programming or applying the scientific method—long-term planning, and algorithmic data manipulation is out of reach for deep-learning models, no matter how much data you throw at them. Even learning a sorting algorithm with a deep neural network is tremendously difficult.

Ethical ML

François Chollet @fchollet · 1d
One hypothesis is that empathy in humans is fundamentally tied to being present with others and seeing their face, and thus all text-based online interactions are geared against empathy.

I don't think this is insurmountable, though

13 21 140

Yann LeCun @ylecun · 23h
Replying to @fchollet
Maybe you should try Facebook.

9 3 66

François Chollet @fchollet · 23h
I have been writing about how content propagation modalities and interaction modalities shape our usage of social networks since 2010. A lot of this reflection came from first-hand experience with Facebook. fchollet.com/blog/the-piano...

François Chollet
@fchollet

I think it's possible to create a social network where the interaction modalities are such that it won't immediately degenerate into extreme toxicity.

Empathy is as much part of human nature as anger or jealousy. But public, anonymous reply buttons only encourage the latter.

Types of Bias

- Sample Bias
  - Carving out too small a world to be representative
- Prejudice Bias
  - Mirroring cultural or other stereotypes in training data
- Measurement Bias
  - Bias from the sensor/device used for collection
- Algorithmic Bias
  - Traditional bias/variance tradeoff, “goodness of fit”

Q: If a robot talks like a human, does it think like a human?

The answer is no. Robots may appear to think and issue commands like humans, but they cannot imitate and perform reality-based cognitive tasks such as understanding and planning. Testing is required.

Q: How would you test whether a robot can understand or plan?

Simple scenarios that require low levels of reasoning skills can be simulated by a computer. Practical reasoning tasks are harder. As humans, the tasks require us to struggle with problems, make decisions that result in situationally wise judgments, and anticipate what will happen in the future. Computers may never be able to do this.

Q: What if we had vastly more computational power than we have today?

We could make robots work for us to complete important tasks, and perhaps they'd be quicker, but machines would never be able to do true intellectual tasks laboriously. Computers would often perform basic reasoning tasks faster than a human can trouble-shoot the situation, and their responses would often be based on computer

Q: So they would get faster at simple tasks, but reach limits for complex tasks requiring practical reasoning?

No. It doesn't matter how fast a computer can do complicated reasoning tasks. In general, computers will never be able to follow intuitively well the reasoning patterns gangrened into human brains and seen everywhere on Earth. Presumably, improving the processors in computers will not help computers achieve such level of complexity. Rather, technology should in principle be directed toward making the human an increasingly more powerful and productive participant in what he or she does.

Q: Greetings, GPT-2. Have you done an interview before?
A: I'm not going to lie. I did not come prepared (laughs).

Q: Which technologies are worth watching in 2020?
A: I would say it is hard to narrow down the list. The world is full of disruptive technologies with real and potentially huge global impacts. The most important is artificial intelligence, which is becoming exponentially more powerful. There is also the development of self-driving cars. There is a lot that we can do with artificial intelligence to improve the world.

Q: What is the future of AI?
A: It would be good if we used the technology more responsibly. In other words, we should treat it like a utility, like a tool. We should put as much effort into developing the technology as necessary, rather than worrying that it's going to harm us and destroy our lives.

Q: Are you worried that ai technology can be misused?
A: Yes, of course. But this is a global problem and we want to tackle it with global solutions. It is not just the government that must do something. Companies must also take responsibility. This means they need to make sure their technologies are used responsibly and that they do not harm anyone.

But machines don't think…

- They generate similar patterns from patterns they have seen before. Most journalism gets this **wrong**.
- Is that fundamentally different than what humans do?
  - **Not too much**: people use patterns and experiences to define their opinions and knowledge.
  - But actually, come on, **the answer is Yes.**
    - Humans can generate and develop thoughts about topics which they have no prior experience, translating complex concepts to new topics without pattern recognition
- What does it mean to think? What does it mean to be intelligent?

Look here if you are interested!

On the Measure of Intelligence

François Chollet *
Google, Inc.
fchollet@google.com

November 5, 2019

Abstract

To make deliberate progress towards more intelligent and more human-like artificial systems, we need to be following an appropriate feedback signal: we need to be able to define and evaluate intelligence in a way that enables comparisons between two systems, as well as comparisons with humans. Over the past hundred years, there has been an abundance of attempts to define and measure intelligence, across both the fields of psychology and AI. We summarize and critically assess these definitions and evaluation approaches, while making apparent the two historical conceptions of intelligence that have implicitly guided them. We note that in practice, the contemporary AI community still gravitates to-wards benchmarking intelligence by comparing the skill exhibited by AIs and humans at specific tasks, such as board games and video games. We argue that solely measuring skill at any given task falls short of measuring intelligence, because skill is heavily modulated by prior knowledge and experience: unlimited priors or unlimited training data allow experimenters to "buy" arbitrary levels of skills for a system, in a way that masks the system's own generalization power. We then articulate a new formal definition of intelligence based on Algorithmic Information Theory, describing intelligence as *skill-acquisition efficiency* and highlighting the concepts of *scope*, *generalization difficulty*, *priors*, and *experience*, as critical pieces to be accounted for in characterizing intelligent systems. Using this definition, we propose a set of guidelines for what a general AI benchmark should look like. Finally, we present a new benchmark closely following these guidelines, the Abstraction and Reasoning Corpus (ARC), built upon an explicit set of priors designed to be as close as possible to innate human priors. We argue that ARC can be used to measure a human-like form of general fluid intelligence and that it enables fair general intelligence comparisons between AI systems and humans.

*I thank José Hernández-Orallo, Julian Togelius, Christian Szegedy, and Martin Wicke for their valuable comments on the draft of this document.

64 Pages of theory, evidence, questions, and bliss!

https://arxiv.org/abs/1911.01547

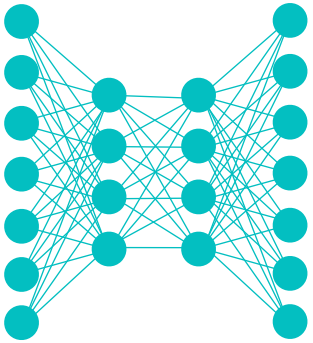
Lecture Notes for

Neural Networks and Machine Learning

Course Introduction

Next Time:
Case Studies in Ethics of ML

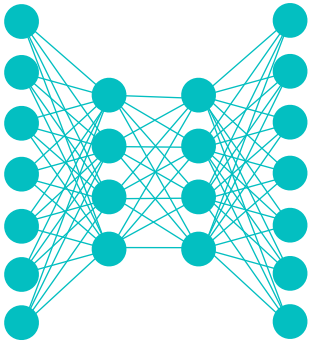
Reading: None



Lecture Notes for

Neural Networks and Machine Learning

Case Studies in Ethical ML



Logistics and Agenda

- Logistics
  - Presentation next time!
- Agenda
  - The AI Principles
  - Case Studies and Discussion
    - Applying the Principles
- Last Time:
  - Course Introduction
  - Bias in ML Algorithms: Sample, Prejudicial, Measurement, and Algorithmic

Ethical Principles in ML

From Australian Goverment, Department of Science

- **Beneficience**: individuals, society and the environment.
- **Respect**: respect human rights, diversity, and autonomy of individuals.
- **Fairness**: be inclusive and accessible, and should not involve or result in unfair discrimination against individuals, communities or groups
- **Privacy**: respect and uphold privacy rights and data protection, and ensure the security of data
- **Reliability**: reliably operate in accordance with their intended purpose
- **Transparency**: ensure people know when they are being significantly impacted by an AI system, and can find out when engaging with them
- **Contestability**: should be a timely process to allow people to challenge the use or output of the AI system
- **Accountability**: Those responsible for the different phases of the AI system lifecycle should be identifiable and accountable for the outcomes of the AI systems, and *human oversight* of AI systems should be enabled.

https://www.industry.gov.au/data-and-publications/building-australias-artificial-intelligence-capability/ai-ethics-framework/ai-ethics-principles

The AI Principles

From Google

- Be socially **beneficial**
- **Avoid** creating or reinforcing **unfair** bias
- Be built and tested for **safety**
- Be **accountable** to people
- Incorporate **privacy** design principles
- Uphold high standards of scientific excellence
- Be **made available** for uses that accord with these principles
- **Google will not pursue**:
  - Tech likely to cause **harm**, tech that **principally** is a **weapon**, Tech that violates **surveillance** norms, Tech that contravenes **human rights**

https://www.blog.google/technology/ai/ai-principles/

How is Google doing?

FeiFei Li, in an email to other Google Cloud employees:

"Avoid at ALL COSTS any mention or implication of AI. Weaponized AI is probably one of the most sensitized topics of AI — if not THE most. This is red meat to the media to find a ways to damage Google."

Opinion: There's more to the Google military AI project than we've been

Google dissolves AI ethics board just one week after forming it

Not a great sign

By Nick Statt | @nickstatt | Apr 4, 2019, 8:17pm EDT

SHARE

What went wrong?

- “First acknowledge the elephant in the room: Google's AI principles”
  - *Evan Selinger, professor of philosophy at Rochester Institute of Technology*
- “A board can't just be 'some important people we know.' You need actual ethicists"
  - *Patrick Lin, director of the Ethics + Emerging Sciences Group at Cal Poly*
- “The group has to have authority to say no to projects”
  - *Sam Gregory, program director at Witness*

https://www.technologyreview.com/s/613281/google-cancels-ateac-ai-ethics-council-what-next/

What about Facebook?

Case Studies of (un)Ethical ML

Case Study: ML Generated Reviews

- Which of these are fake:
  - "I love this place. I have been going here for years and it is a great place to hang out with friends and family. I love the food and service. I have never had a bad experience when I am there."
  - "I had the grilled veggie burger with fries!!!! Ohhhh and taste. Omgggg! Very flavorful! It was so delicious that I didn't spell it!!"
  - "My family and I are huge fans of this place. The staff is super nice and the food is great. The chicken is very good and the garlic sauce is perfect. Ice cream topped with fruit is delicious too. Highly recommended!"
- Does this violate any ethical guidelines?
- "While this study focuses only on creating review text that appears to be authentic, Yelp's recommendation software employs a more holistic approach," said a spokesperson. "It uses many signals beyond text-content alone to determine whether to recommend a review."
- Does the mere presence of this cause problems of trust?

https://www.theverge.com/2017/8/31/16232180/ai-fake-reviews-yelp-amazon

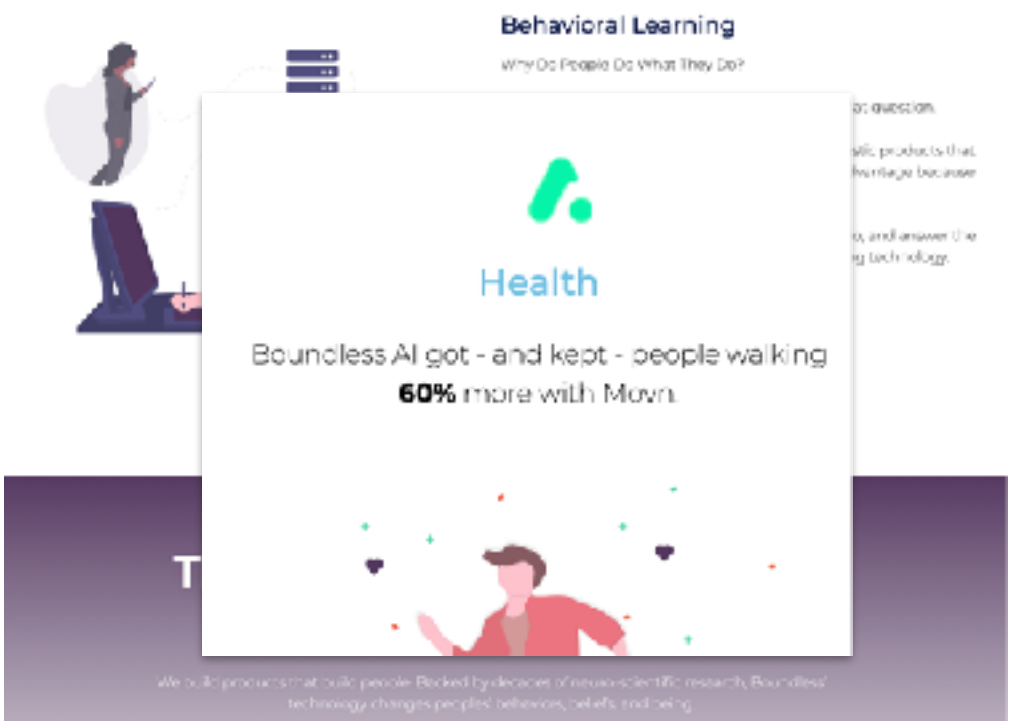
Case Study: Face Swapping

- Does the mere presence of this cause problems of trust?

https://www.youtube.com/watch?v=gLol9hAX9dw

Case Study: Reinforcing App Addiction

- Identifying behavior to keep users in your app
- Does this violate any ethical guidelines?



Ultimately, Dopamine Labs predicts they can add 10 percent to a company's revenues. In practice, their numbers are a bit all over the map, with some companies seeing bounces of more than 100 percent in terms of user interactions with, in or on an app. For other companies the boost could be around 8 percent.

https://techcrunch.com/2017/09/08/meet-the-tech-company-that-wants-to-make-you-even-more-addicted-to-your-phone/

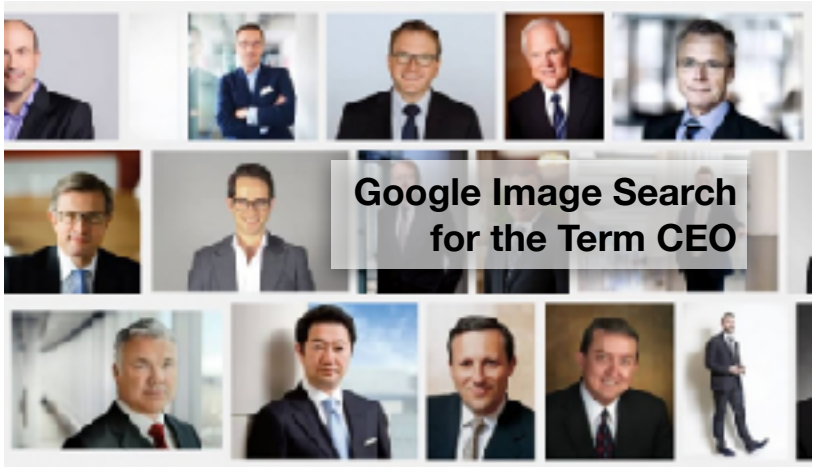
Case Study: Reinforced Gender/Race Bias

- Not a new problem in technology:
  - Example: Crash Test Dummies, Because most crash tests have male "dummies" females had a 20 to 40 percent greater risk of being killed or seriously injured, compared to 15 percent for men.
- But can also be more subtle:

Google's algorithm shows prestigious job ads to men, but not to women. Here's why that should worry you.

"It's part of a cycle: How people perceive things affects the search results, which affect how people perceive things," Cynthia Matuszek, Professor of Computer Ethics at UMD

Does this violate any Ethics Principles?



https://www.washingtonpost.com/news/the-intersect/wp/2015/07/06/googles-algorithm-shows-prestigious-job-ads-to-men-but-not-to-women-heres-why-that-should-worry-you/?noredirect=on&utm_term=.055bff1a94ad

Case Study: Predictive Policing

- Once a crime has happened, can it be classified as a gang crime?
  - Used partially generative NN for classifying if a crime was gang related, with the aim at predicting gang retaliation. Trained on LAPD data 2014-2016
- Does this violate any ethical guidelines?

Blake Lemoine AI Google Researcher On Bias in ML

But researchers attending the AIES talk raised concerns during the Q&A afterward. How could the team be sure the training data were not biased to begin with? What happens when someone is mislabeled as a gang member? Lemoine asked rhetorically whether the researchers were also developing algorithms that would help heavily patrolled communities predict police raids.

Hau Chan, a computer scientist now at Harvard University who was presenting the work, responded that he couldn't be sure how the new tool would be used. "I'm just an engineer," he said. Lemoine quoted a lyric from a song about the wartime rocket scientist Wernher von Braun, in a heavy German accent: "Once the rockets are up, who cares where they come down?" Then he angrily walked out.

https://www.sciencemag.org/news/2018/02/artificial-intelligence-could-identify-gang-crimes-and-ignite-ethical-firestorm

Case Study: ML Generated Products

- Online generation of content to facilitate buying behavior

It's not about trolls, but about a kind of violence inherent in the combination of digital systems and capitalist incentives. It's down to that level of the metal. This, I think, is my point: The system is complicit in the abuse.

And right now, right here, YouTube and Google are complicit in that system. The architecture they have built to extract the maximum revenue from online video is being hacked by persons unknown to abuse children, *perhaps not even deliberately*, but at a massive scale.

These videos, wherever they are made, however they come to be made, and whatever their conscious intention (i.e., to accumulate ad revenue) are feeding upon a system which was consciously intended to show videos to children for profit. The unconsciously-generated, emergent outcomes of that are all over the place.

—James Bridle

https://medium.com/@jamesbridle/something-is-wrong-on-the-internet-c39c471271d2

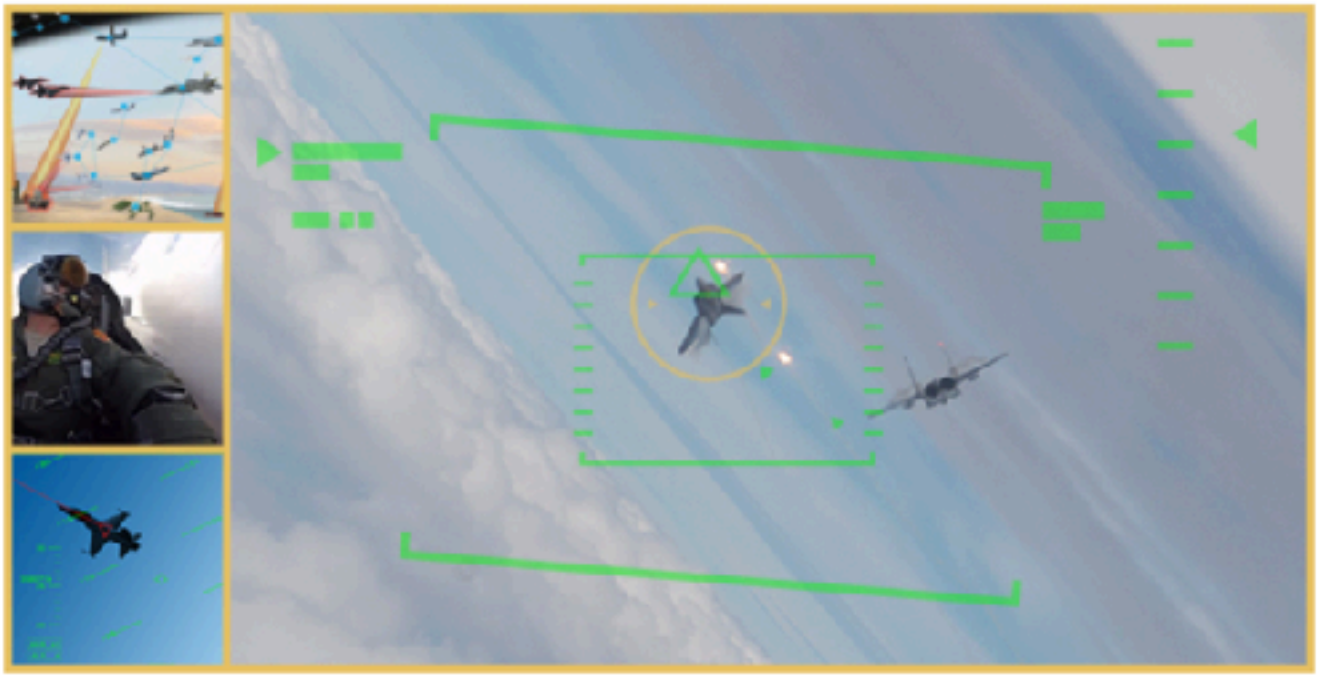
AI Warfare

Defense Advanced Research Projects Agency > News And Events

Training AI to Win a Dogfight

Trusted AI may handle close-range air combat, elevating pilots' role to cockpit-based mission commanders

OUTREACH@DARPA.MIL
5/8/2019

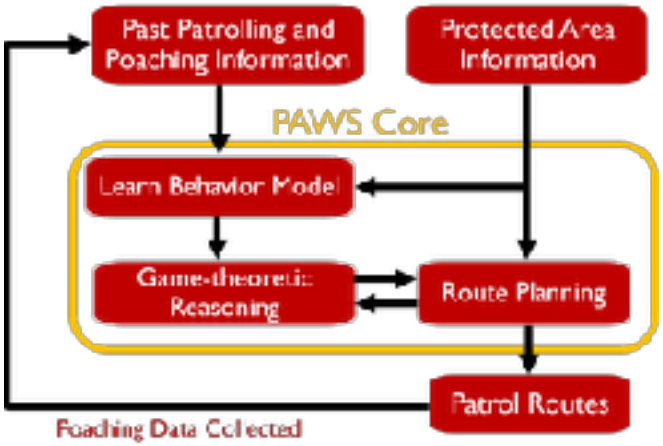


A Counter Example

- PAWS: Prevent Tiger Poaching in Malaysia

"Even if you can predict some sort of poaching activities, it's not always good to just go to areas with high predicted poaching activity!"

Prof. Fei Chang, CMU



"None of these aspects can be addressed with a publicly available commercial tool, or directly addressed by sitting in an office… That means we need to talk to experts, understand the problem, and propose solutions to it."

https://www.fastcompany.com/90157255/you-can-now-take-a-class-on-how-to-make-ai-that-isnt-evil

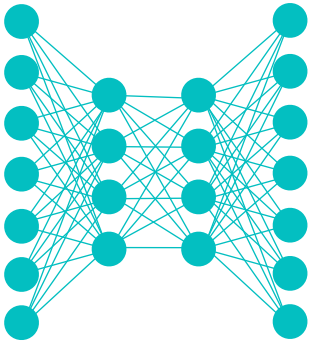
Lecture Notes for

Neural Networks and Machine Learning

Case Studies in Ethical ML

Next Time:
Practical Example in NLP

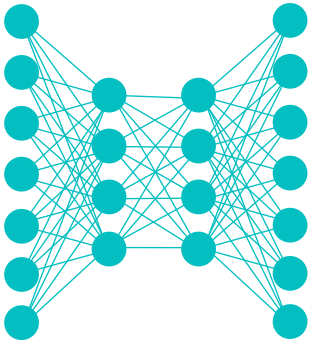
Reading: None



Lecture Notes for

Neural Networks and Machine Learning

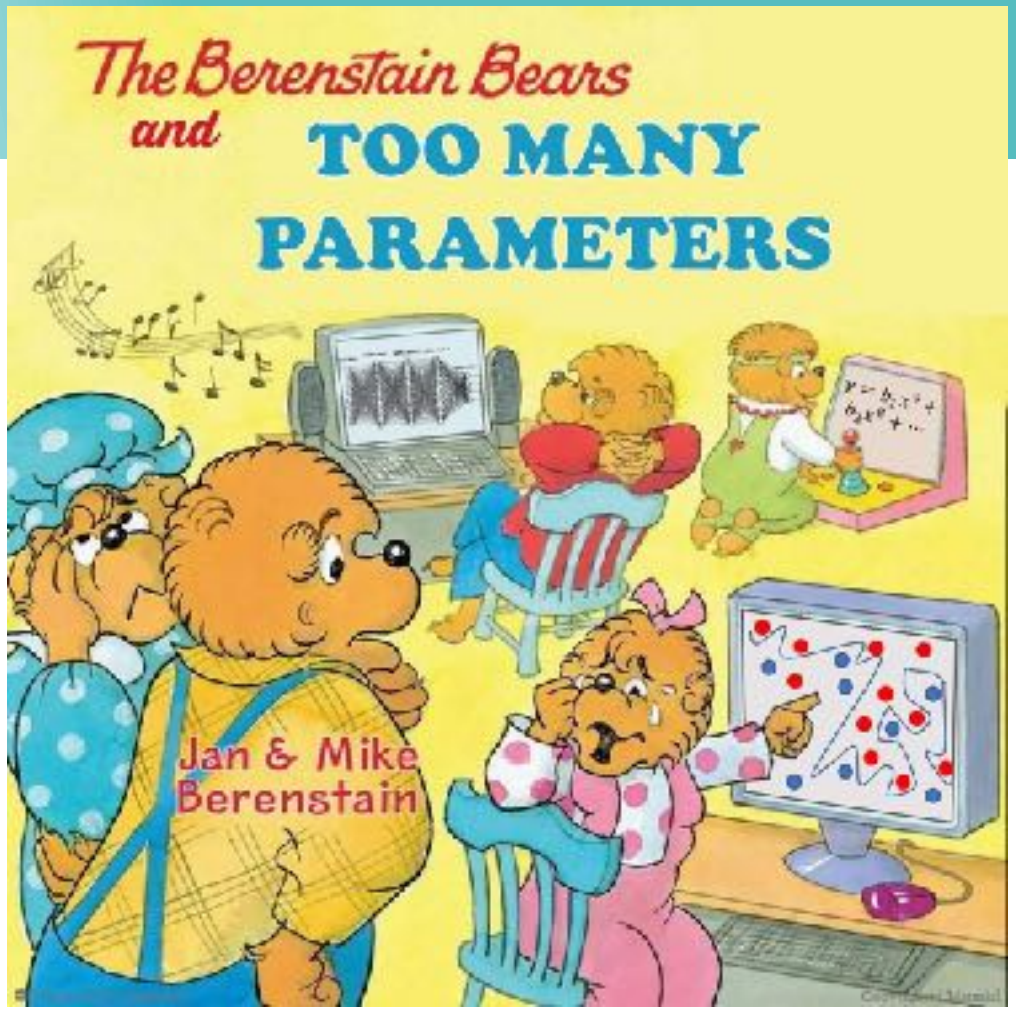
A Practical Example of Ethically Aware NLP



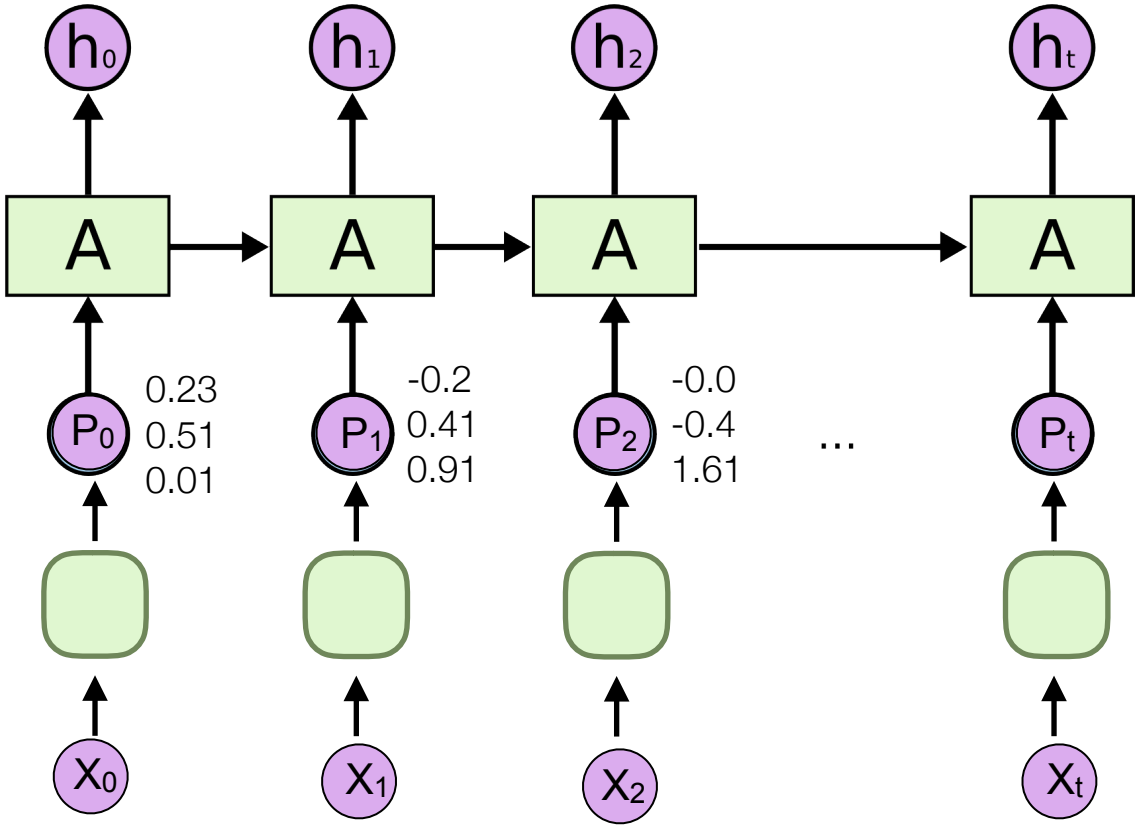
Logistics and Agenda

- Logistics
  - Post about your preferred lecture discussion or paper summary
- Agenda
  - NLP Review
  - Extended Example
  - Presentation, if time
- Last Time:
  - Bias in ML Algorithms
  - Case Studies

NLP Embeddings Review



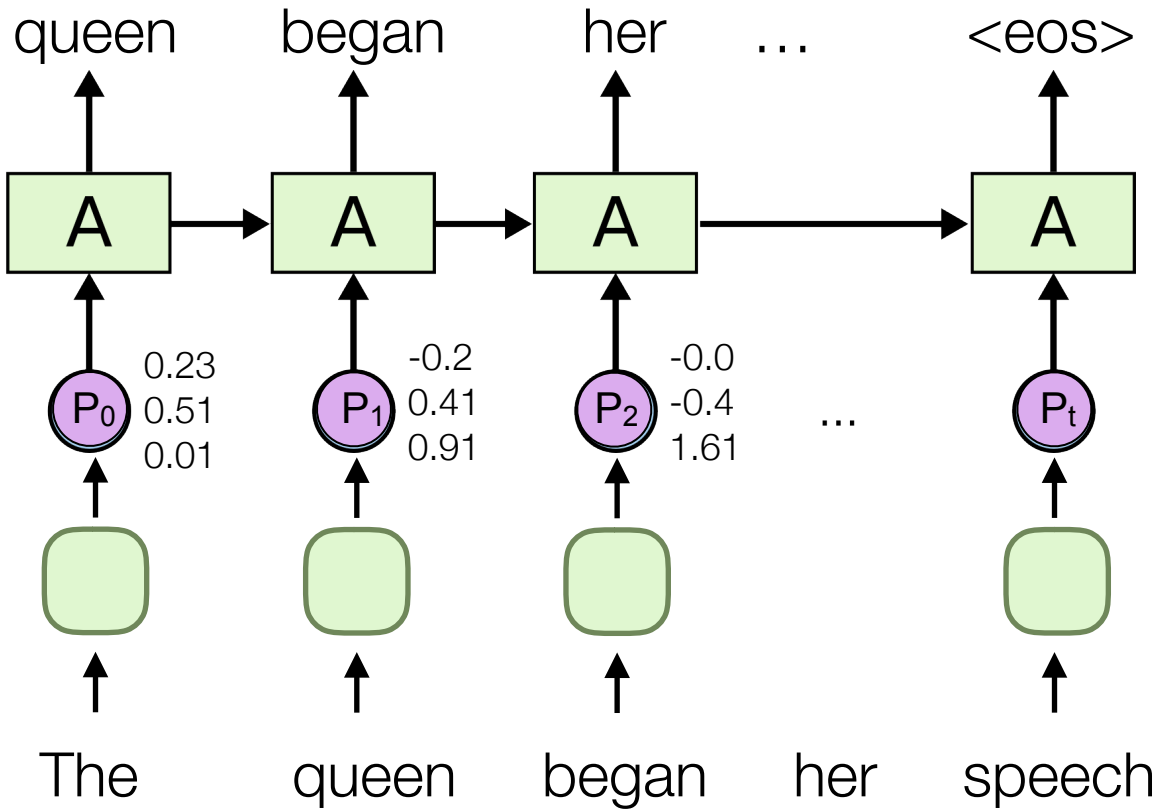
Word Embeddings (like Wide/Deep)



int:	3	1	17	
one hot:	0 0 1 0 0 0 0	1 0 0 0 0 0 0	0 0 0 0 0 0 0	

Word Embeddings: Training

- many training options exist
  - a popular option, next word prediction



Word Embeddings

GloVe

Global Vectors for Word Representation

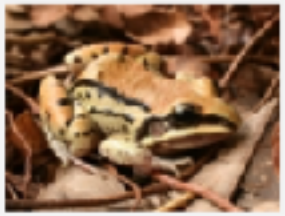
Highlights

1. **Nearest neighbors**

The Euclidean distance (or cosine similarity) between two word vectors provides an effective method for measuring the linguistic or semantic similarity of the corresponding words. Sometimes, the nearest neighbors according to this metric reveal rare but relevant words that lie outside an average human's vocabulary. For example, here are the closest words to the target word *frog*:

0. *frog*
1. frogs
2. toad
3. litoria
4. leptodactylidae
5. rana
6. lizard
7. eleutherodactylus

3. litoria

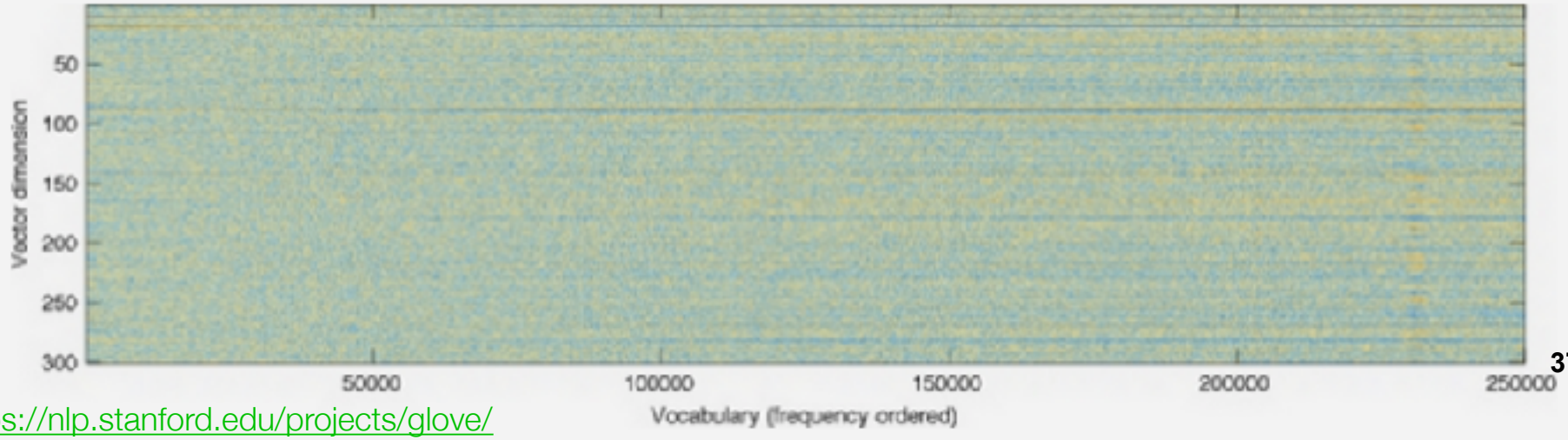


4. leptodactylidae

5. rana

7. eleutherodactylus

GloVe produces word vectors with a marked banded structure that is evident upon visualization:

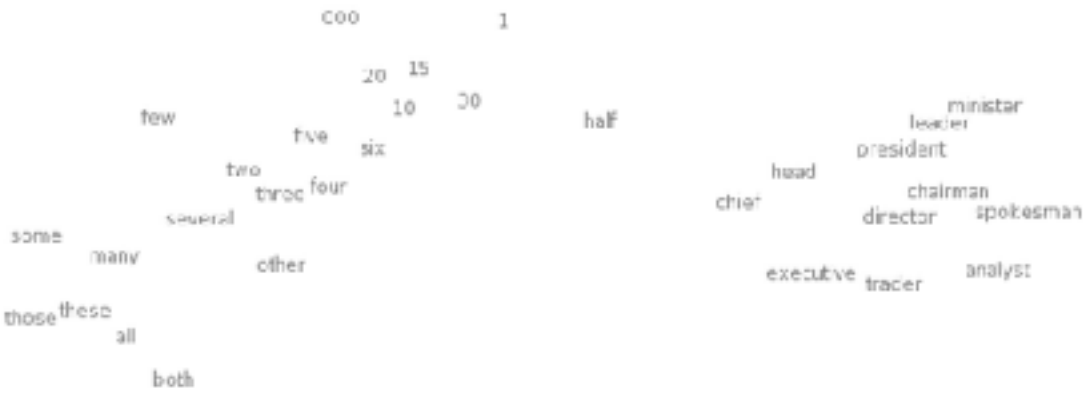


https://nlp.stanford.edu/projects/glove/

Word Embeddings: proximity

GloVe

Global Vectors for Word Representation



t-SNE visualizations of word embeddings. Left: Number Region; Right: Jobs Region. From Turian *et al.* (2010), see complete image.

FRANCE	JESUS	XBOX	REDDISH	SCRATCHED	MEGABITS
AUSTRIA	GOD	AMIGA	GREENISH	NAILED	OCTETS
BELGIUM	SATI	PLAYSTATION	BLUISH	SMASHED	MB/S
GERMANY	CHRIST	MSX	PINKISH	PUNCHED	BIT/S
ITALY	SATAN	IPOD	PURPLISH	POPPED	BAUD
GREECE	KALI	SEGA	BROWNISH	CRIMPED	CARATS
SWEDEN	INDRA	PSNUMBER	GREYISH	SCRAPED	KBIT/S
NORWAY	VISHNU	HD	GRAYISH	SCREWED	MEGAHERTZ
EUROPE	ANANDA	DREAMCAST	WHITISH	SECTIONED	MEGAPIXELS
HUNGARY	PARVATI	GEFORCE	SILVERY	SLASHED	GBIT/S
SWITZERLAND	GRACE	CAPCOM	YELLOWISH	RIPPED	AMPERES

What words have embeddings closest to a given word? From Collobert *et al.* (2011)

The **chairman** called the **meeting** to order.

The **director** called the **conference** to order.

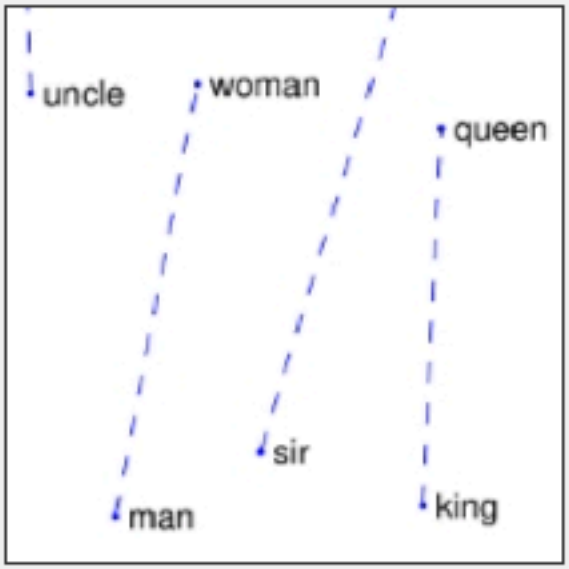
The **chief** called the **council** to order.

http://colah.github.io/posts/2014-07-NLP-RNNs-Representations/

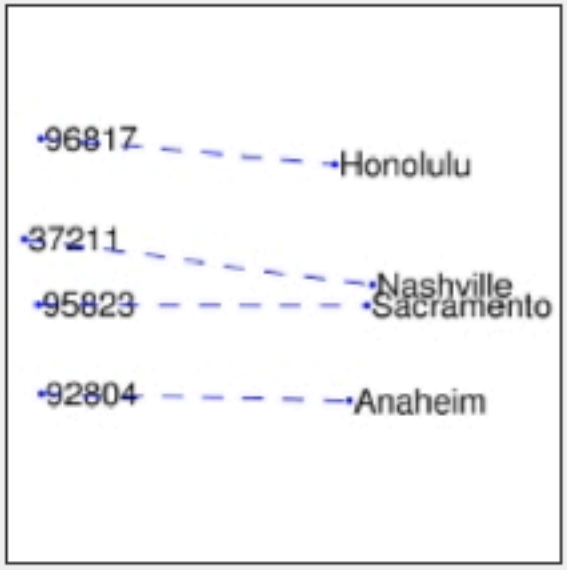
Word Embeddings: Analogy

GloVe

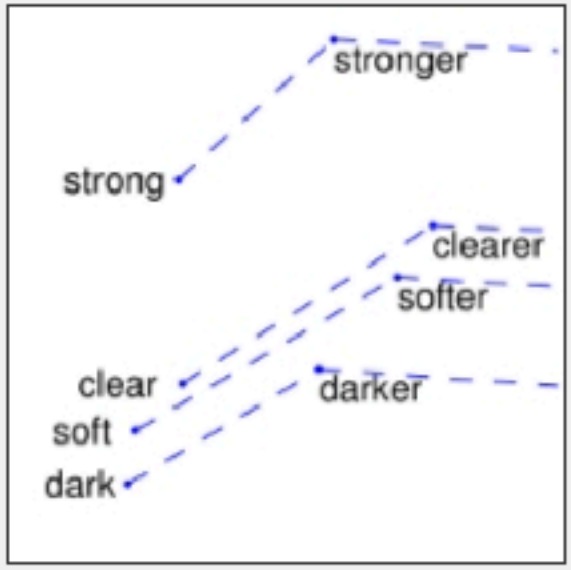
Global Vectors for Word Representation



man - woman



city - zip code

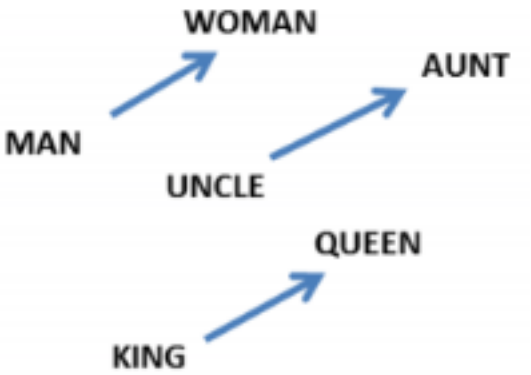


comparative - superlative

each vector difference **might** encode analogy

https://nlp.stanford.edu/projects/glove/ http://colah.github.io/posts/2014-07-NLP-RNNs-Representations/

Word Embeddings: Analogy?



From Mikolov *et al.* (2013a)

$$W(\text{“woman”}) - W(\text{“man”}) \simeq W(\text{“aunt”}) - W(\text{“uncle”})$$

$$W(\text{“woman”}) - W(\text{“man”}) \simeq W(\text{“queen”}) - W(\text{“king”})$$

$$\overrightarrow{\text{man}} - \overrightarrow{\text{woman}} \approx \overrightarrow{\text{computer programmer}} - \overrightarrow{\text{homemaker}}.$$

Extreme *she* occupations

1. homemaker	2. nurse	3. receptionist
4. librarian	5. socialite	6. hairdresser
7. nanny	8. bookkeeper	9. stylist
10. housekeeper	11. interior designer	12. guidance counselor

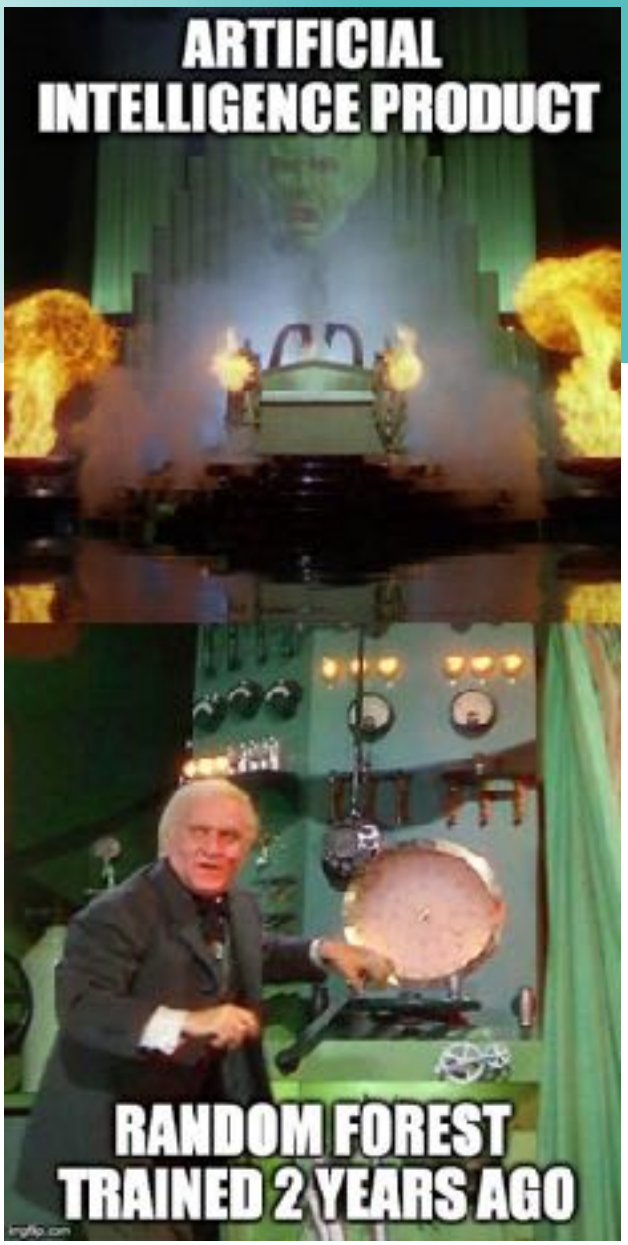
Extreme *he* occupations

1. maestro	2. skipper	3. protege
4. philosopher	5. captain	6. architect
7. financier	8. warrior	9. broadcaster
10. magician	11. fighter pilot	12. boss

Trained on New York Times

Bolukbasi et al., NeurIPs 2016
https://arxiv.org/pdf/1607.06520.pdf

https://nlp.stanford.edu/projects/glove/



Practical Example in NLP

ConceptNet

en cooking dinner

An English term in ConceptNet 5.7

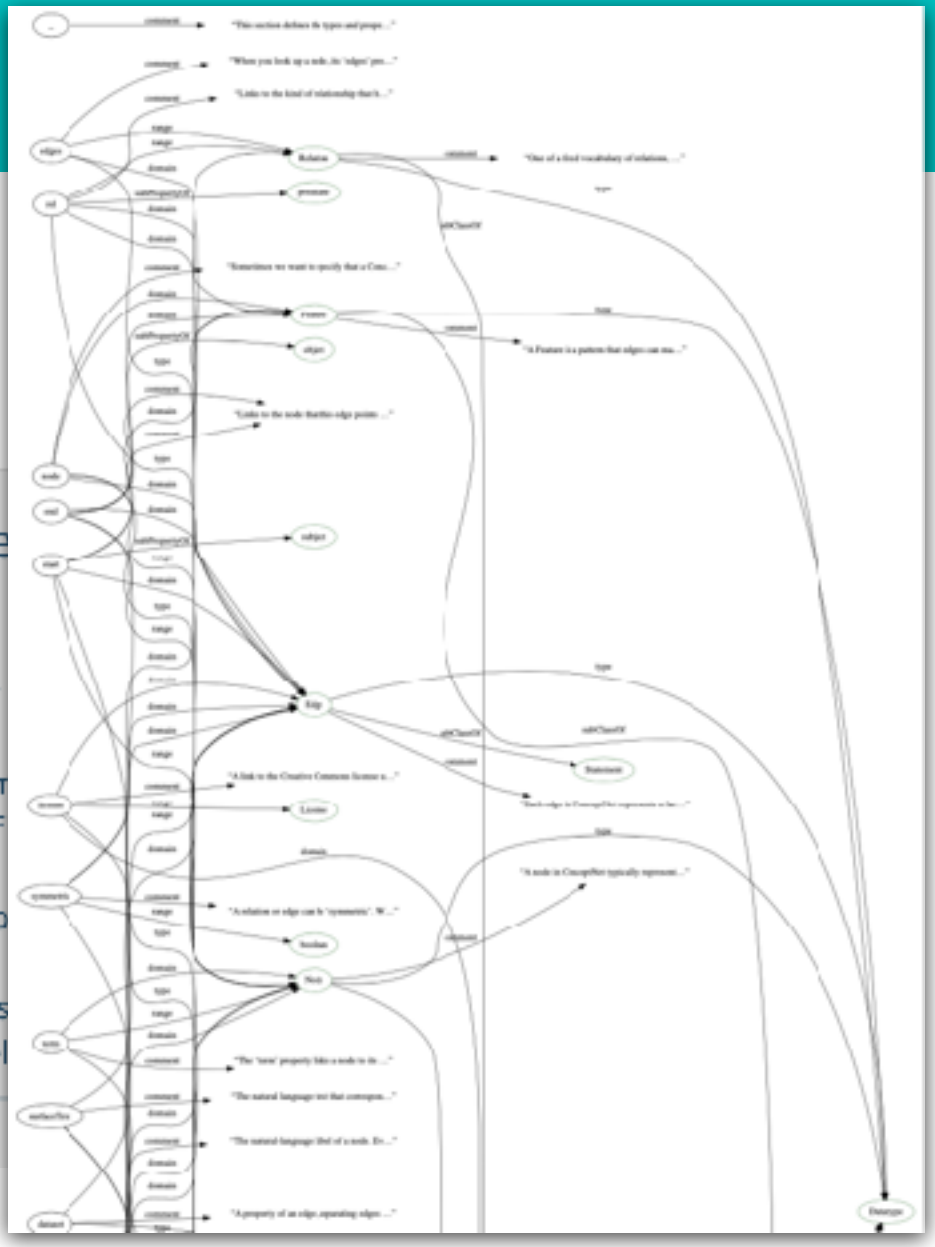
Source: Open Mind Common Sense contributors
View this term in the API

cooking dinner is a subevent of...

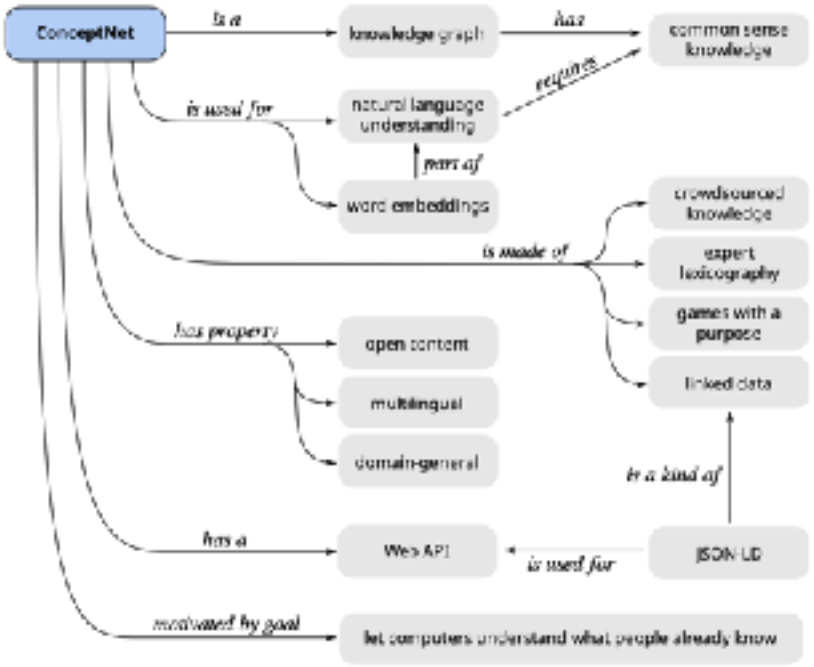
- en boiling water →
- en it burns →
- en preheat the oven →
- en taste the food →
- en boil salt water →
- en boil water →
- en brown the hamburger →
- en chop a vegetable →
- en defrost →
- en a fire →
- en the fire alarm might go off →

cooking dinne for...

- en feeding a family
- en TO EAT →
- en entertaining com
- en feeding yourself
- en anyone →
- en avoiding fast foo
- en being a cook →
- en caring for others
- en cheering yoursel
- en creative people
- en eating →



ConceptNet Numberbatch



- Create with a Knowledge Graph (from multiple sources with relations like *UsedFor*, *PartOf*, *etc*.)
- Based on this KG, perturb existing embeddings (like GloVe) to optimize:

$$\Psi(Q) = \sum_{i=1}^{n} \left[\alpha_i \|q_i - \hat{q}_i\|^2 + \sum_{(i,j) \in E} \beta_{ij} \|q_i - q_j\|^2 \right]$$

new embed — old embed — neighbors from KG

(keep similar to original) — (make similar according to other knowledge)

- Easy to optimize the objective by averaging neighbors in the ConceptNet KG
- Multiple embeddings achieved by merging through “retrofitting” which projects onto a shared matrix space (with SVD)

ConceptNet 5.5: An Open Multilingual Graph of General Knowledge, Speer et al., 2017

Aside: Transparency in Research

ConceptNet is all you need

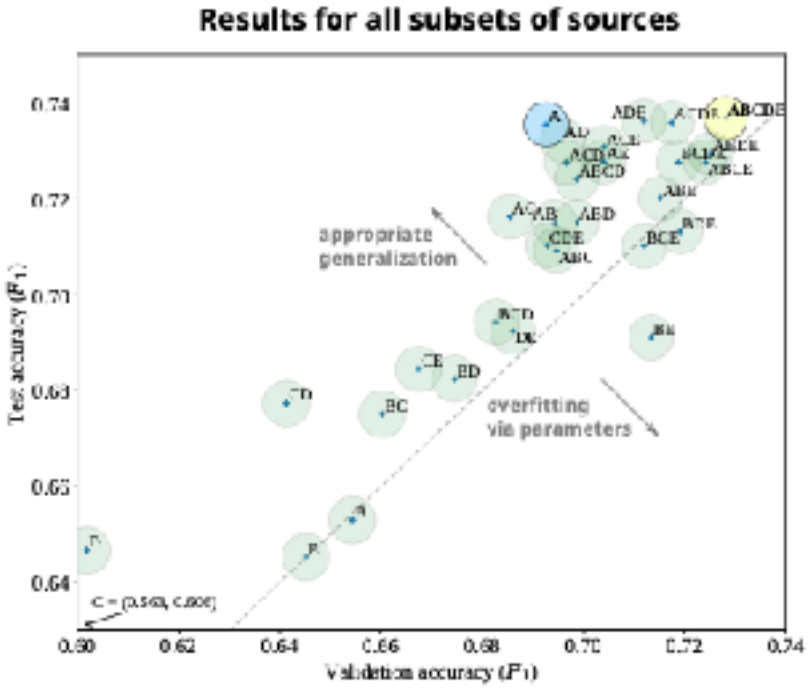
Our full classifier used the linear combination of 5 types of input features shown above. This point is labeled **ABCDE** on the graph to the right. The other points are ablated versions of the classifier, trained on subsets of the five sources.

We found that the single feature of ConceptNet similarity (**A**) performed just as well on the test data as the full classifier, despite its lower validation accuracy.

This one-feature classifier could be more simply described as a heuristic over cosine similarities of ConceptNet embeddings:

$$\text{sim}(term_1, art) - \text{sim}(term_2, art) > 0.0961$$

It seems that the test data contained distinctions that can already be found by comparing ConceptNet embeddings, and that more complex features may have simply provided an opportunity to overfit to the validation set by parameter selection.



This graph shows the validation and test accuracy of classifiers trained on subsets of the five sources of features. Ellipses indicate standard error of the mean, assuming that the data is sampled from a larger set.

https://blog.conceptnet.io/2018/06/naacl2018-poster.pdf

ConceptNet Numberbatch

As a kid, I used to hold marble racing tournaments in my room, rolling marbles simultaneously down plastic towers of tracks and funnels. I went so far as to set up a bracket of 64 marbles to find the fastest marble. I kind of thought that running marble tournaments was peculiar to me and my childhood, but now I've found out that marble racing videos on YouTube are a big thing! Some of them even have overlays as if they're major sporting events.

In the end, there's nothing special about the fastest marble compared to most other marbles. It's just lucky. If one ran the tournament again, the marble champion might lose in the first round. But the one thing you could conclude about the fastest marble is that it was no *worse* than the other marbles. A bad marble (say, a misshapen one, or a plastic bead) would never luck out enough to win.

In our paper, we tested 30 alternate versions of the classifier, including the one that was roughly equivalent to this very simple system. We were impressed by the fact that it performed as well as our real entry. And this could be because of the inherent power of ConceptNet Numberbatch, or it could be because it's the lucky marble.

-Robyn Speer
http://blog.conceptnet.io

How to Make a Racist AI without Really Trying

Robyn Speer, 2017
http://blog.conceptnet.io/posts/2017/how-to-make-a-racist-ai-without-really-trying/

Debiasing: Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings
Bolukbasi et al., NeurIPs 2016
https://arxiv.org/pdf/1607.06520.pdf

ConceptNet 5.5: An Open Multilingual Graph of General Knowledge
Speer et al., AAAI 2017
https://arxiv.org/pdf/1612.03975.pdf

Rachael Tatman @rctatman · 18h
I first got interested in ethics in NLP/ML becuase I was asking "does this system work well for everyone". It's a good question, but there's a more important important one:

Who is being harmed and who is benefiting from this system existing in the first place?

Lecture Notes for

Neural Networks and Machine Learning

Ethically Aware NLP

Next Time:
CNN Visualization

Reading: Chollet Article