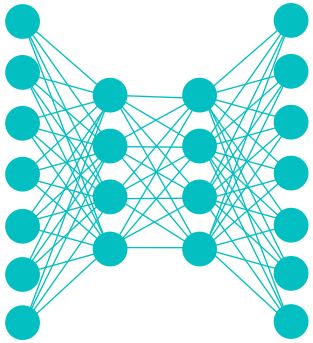
Lecture Notes for

Neural Networks and Machine Learning

A Practical Example of Ethically “Aware” NLP Practices

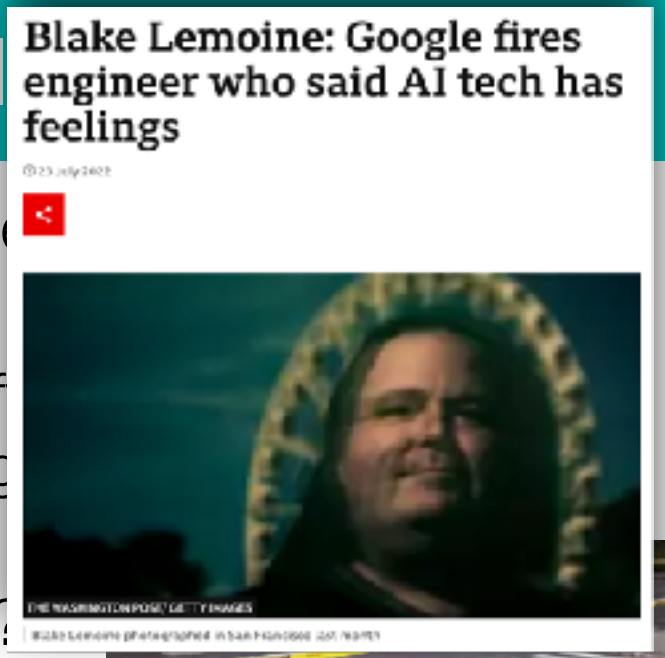


Logistics and Agenda

- Logistics
  - Viewing video of course
  - Preferred lecture discussion assignments
- Last Time:
  - Ethical Guidelines
  - Case Studies
- Agenda
  - Final Case Studies: Ethical Guidelines of AI
  - Paper Presentation:
    - Multi-modal datasets: misogyny … stereotypes
  - NLP Review
  - Extended Example

Case Study: Predictive Pol

- Once a crime has happened, can it b
  gang crime?
  - Used partially generative NN for classif
    gang related, with the aim at predicting
    Trained on LAPD data 2014-2016
- Does this violate any ethical guidelines?



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.

Blake Lemoine AI Google Researcher On Bias in ML

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

Ethical Considerations in Military App.

- Ethical guidelines in combat
  - **One Interpretation**: Combat is not ethical because harm of individuals is incentivized
  - **Another**: Certain ethical guidelines can still be considered, accepting that the aim of combat is, in many instances, to create a weapon
- These are both common (maybe valid) interpretations and it is up to you to decide if combat should be included in ethical guidelines
  - **My take**: combat should not be considered ethical, but is unavoidable in the presence of nefarious actors and therefore guidelines need to be in place
  - Many individuals will disagree with me on this and have excellent points, that I do not disagree with

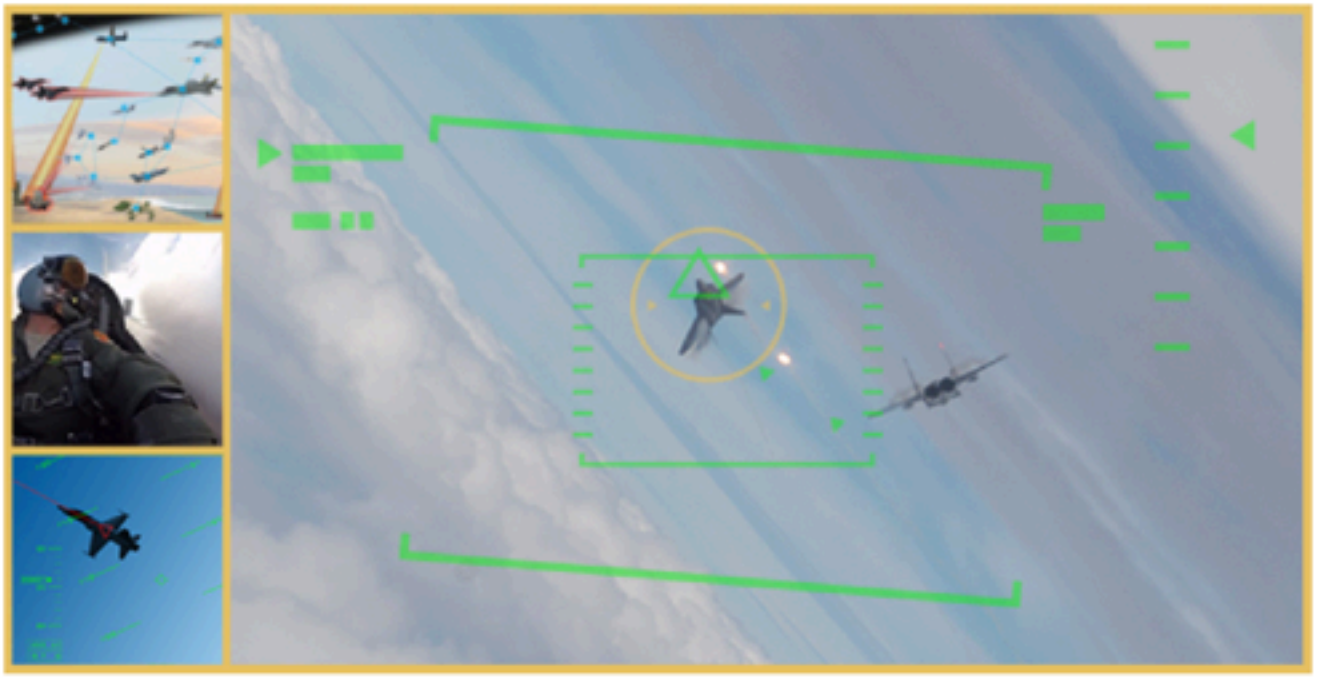
AI Warfare

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



Paper Presentation

Multimodal datasets: misogyny, pornography, and malignant stereotypes

Abeba Birhane[*]
University College Dublin & Lero
Dublin, Ireland
abeba.birhane@ucdconnect.ie

Vinay Uday Prabhu[*]
Independent Researcher
vinayprabhu@cmu.edu

Emmanuel Kahembwe
University of Edinburgh
Edinburgh, UK
e.kahembwe@ed.ac.uk

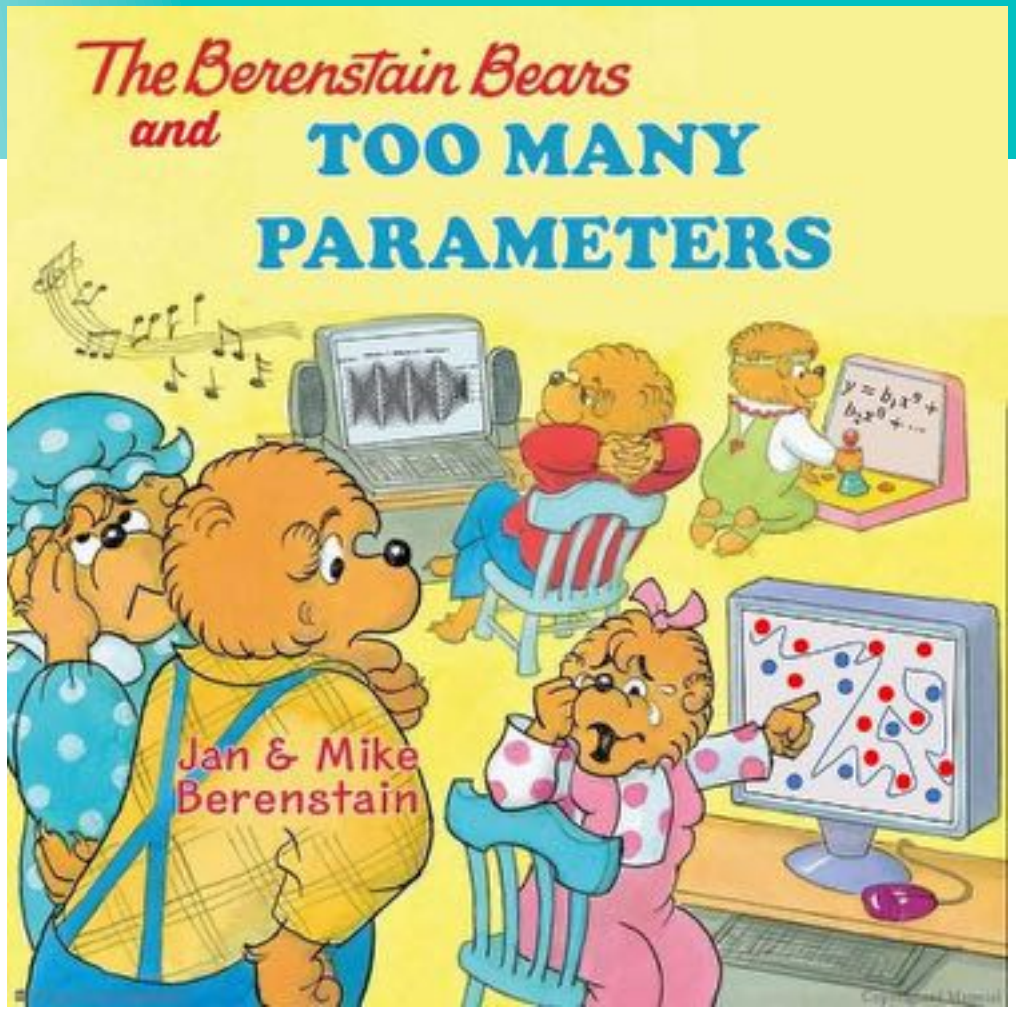
Abstract

We have now entered the era of trillion parameter machine learning models trained on billion-sized datasets scraped from the internet. The rise of these gargantuan datasets has given rise to formidable bodies of critical work that has called for caution while generating these large datasets. These address concerns surrounding the dubious curation practices used to generate these datasets, the crucial quality of alt-text data available on the world wide web, the problematic content of the CommonCrawl dataset often used as a source for training large language models, and the documented biases in large-scale visio-linguistic models (such as OpenAI's CLIP models) trained on opaque datasets (WebImageText). In the backdrop of these specific calls of caution, we examine the recently released LAION-400M dataset, which is a CLIP-filtered dataset of Image-Alt-text pairs parsed from the Common-Crawl dataset. We found that the dataset contains, troublesome and explicit images and text pairs of rape, pornography, malign stereotypes, racist and ethnic slurs, and other extremely problematic content. We outline numerous implications, concerns and downstream harms regarding the current state of large scale datasets while raising open questions for various stakeholders including the AI community, regulators, policy makers and data subjects.

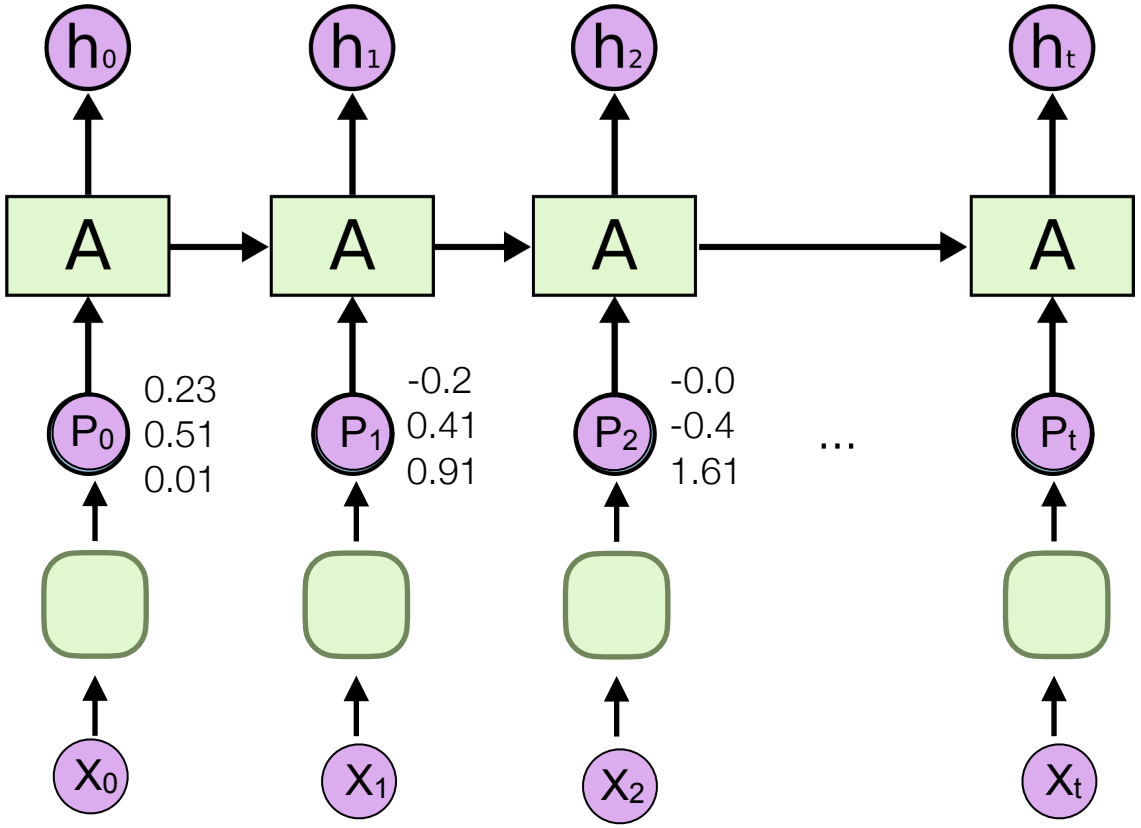
Warning: This paper contains NSFW content that some readers may find disturbing, distressing, and/or offensive.

https://arxiv.org/pdf/2110.01963.pdf

NLP Embeddings Review



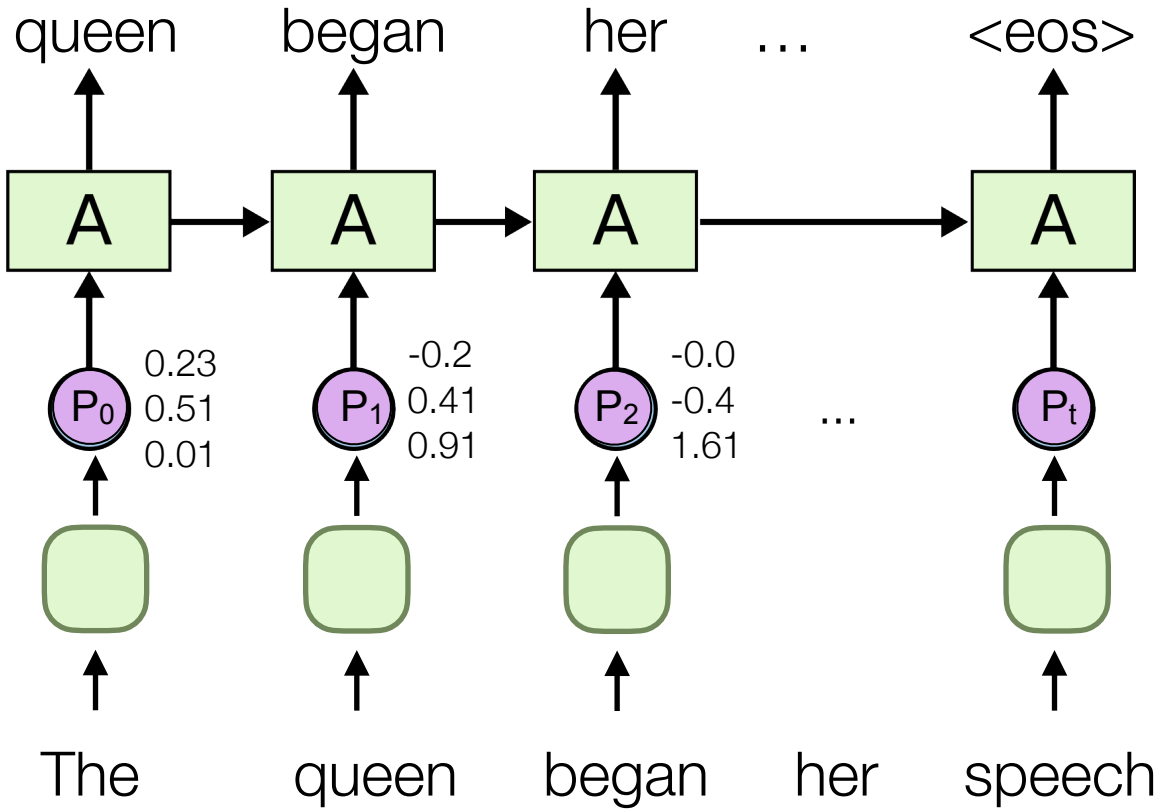
Word Embeddings Review



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 Review

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



GloVe Review

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

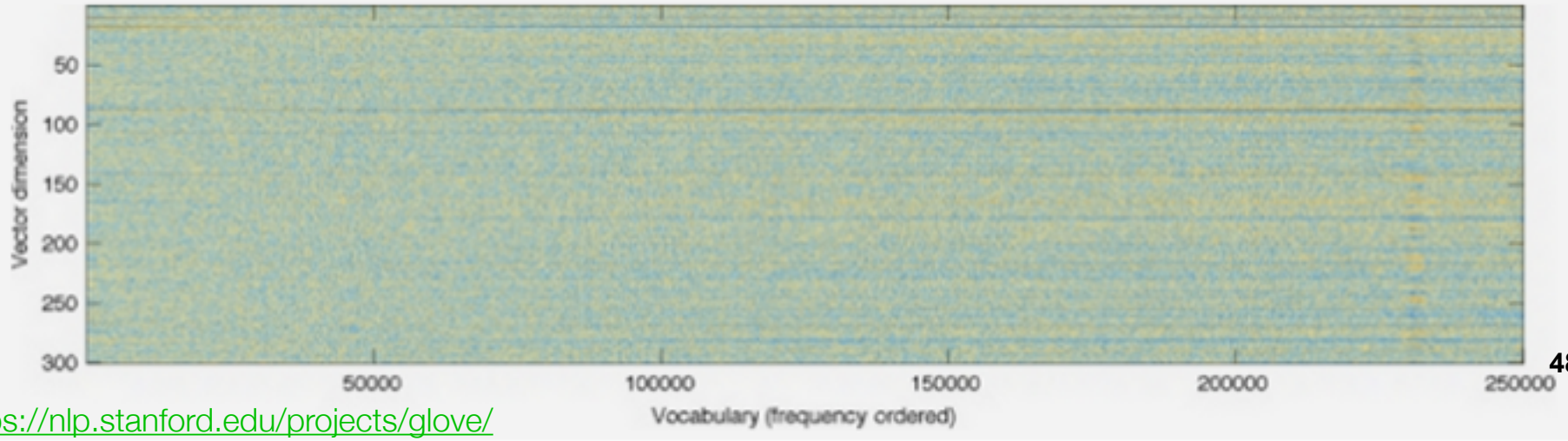
3. litoria

4. leptodactylidae

5. rana

7. eleutherodactylus

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

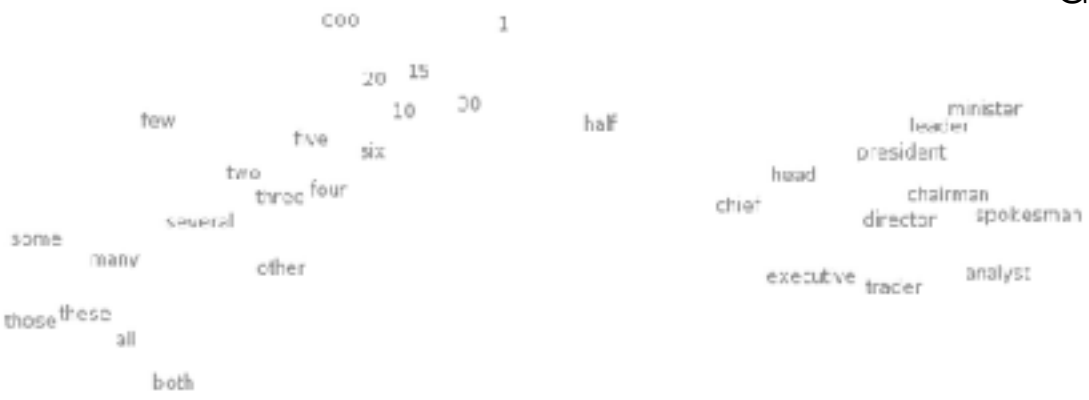


https://nlp.stanford.edu/projects/glove/

Word Embeddings: proximity

GloVe Review

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.

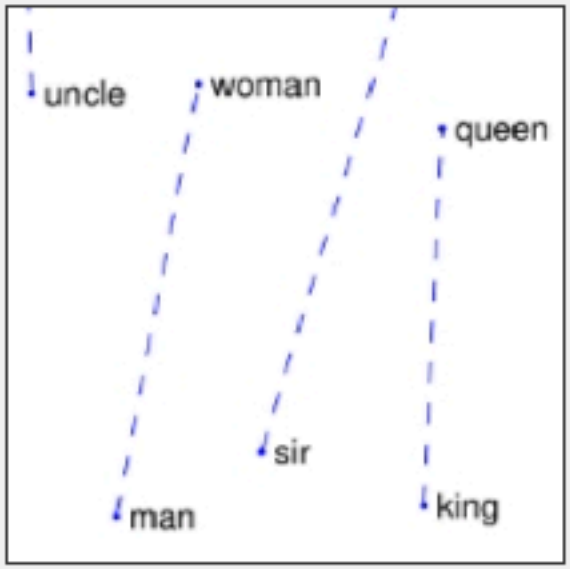
The **chief** called the **council** to order.

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

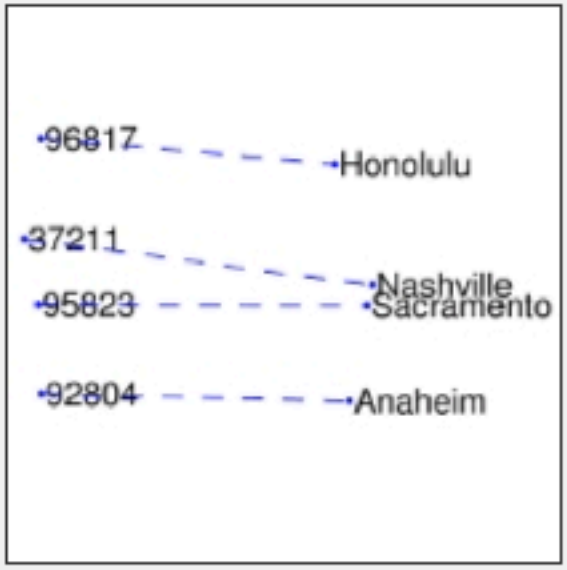
Word Embeddings: Analogy

GloVe Review

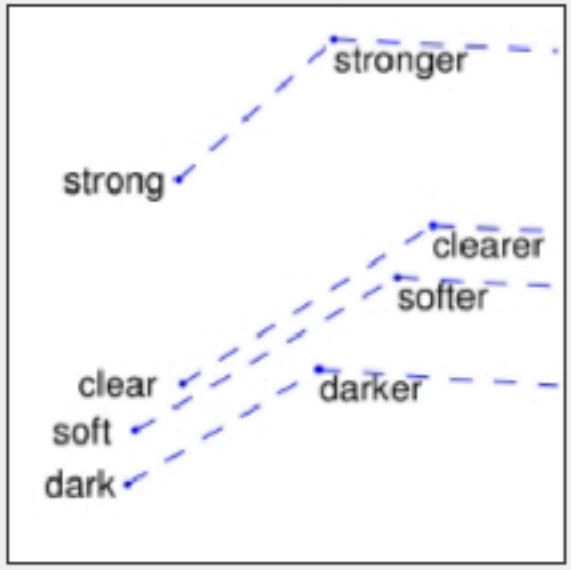
Global Vectors for Word Representation



man - woman



city - zip code



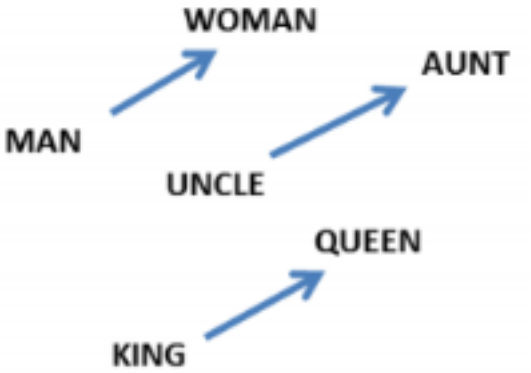
comparative - superlative

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?

GloVe Review



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

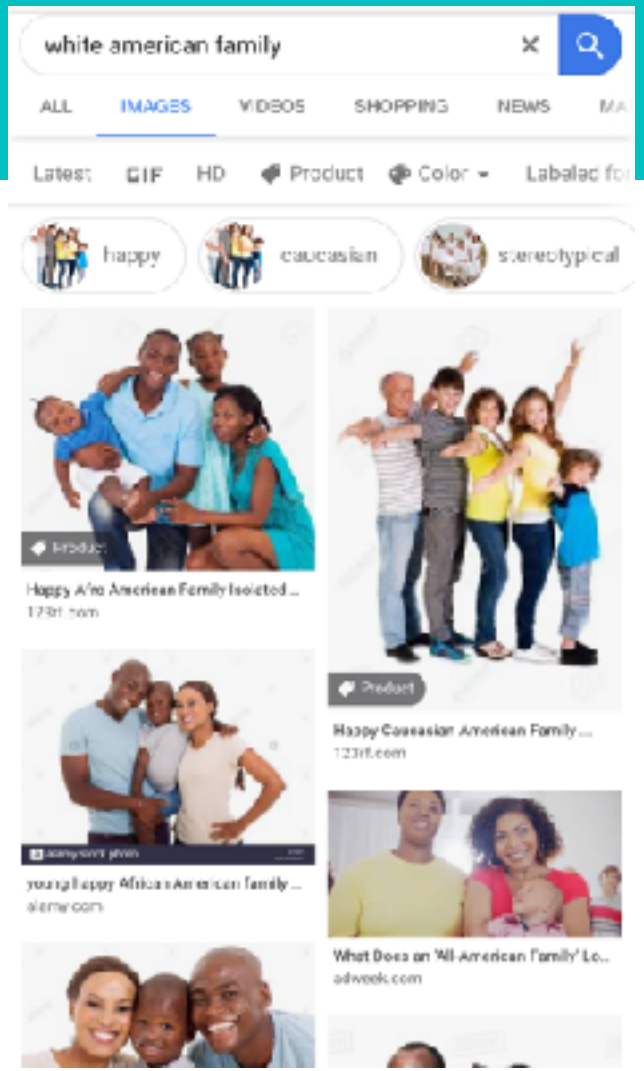
1. maestro	2. skipper	3. protege
4. philosopher	5. captain	6. architect
7. financier	8. warrior	9. broadcaster
10. magician	11. fighter pilot	12. boss

Trained on New York Times

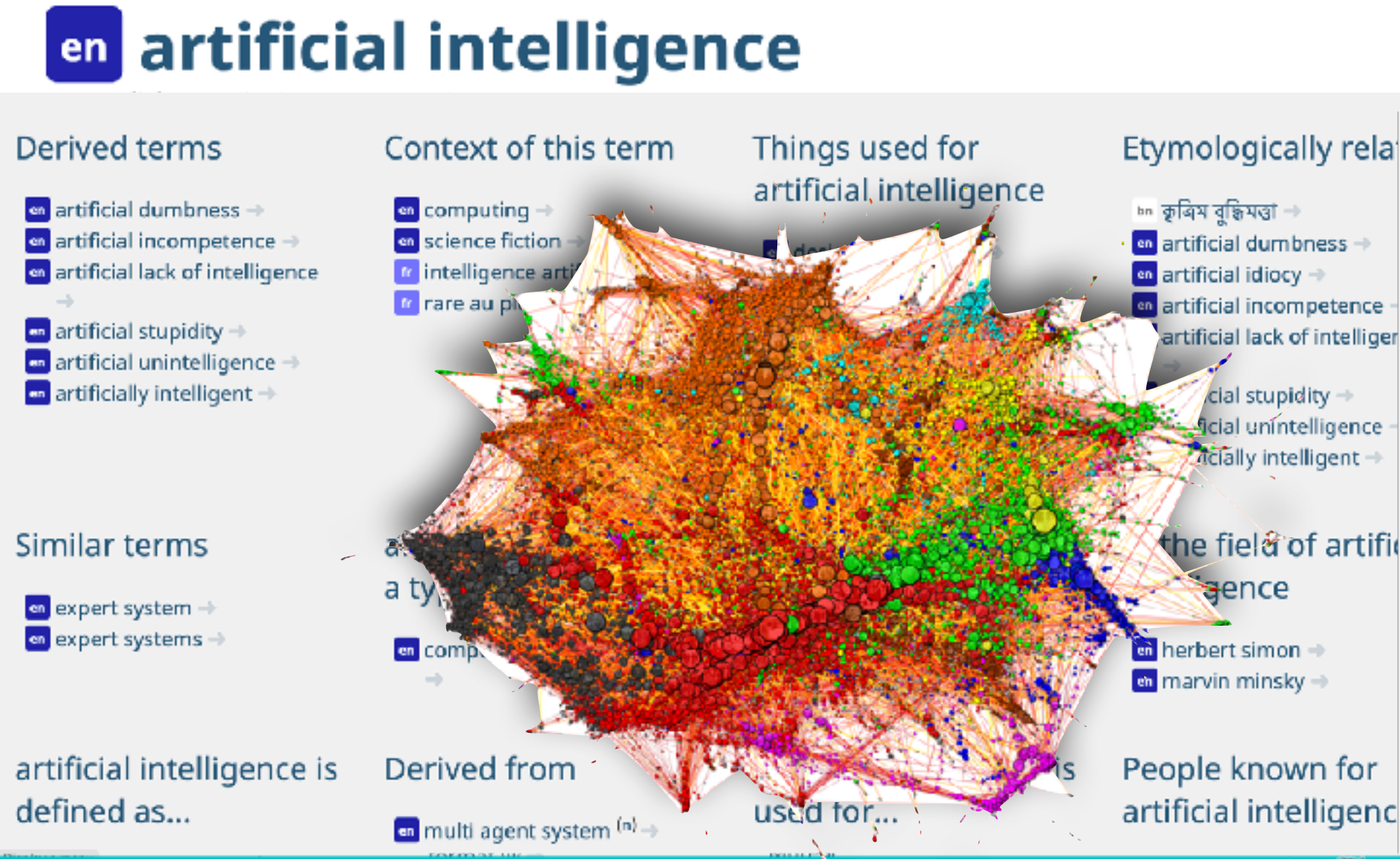
Bolukbasi et al., NeurIPs 2016
https://arxiv.org/pdf/1607.06520.pdf

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

Practical Example in NLP



ConceptNet


en artificial intelligence
Derived terms
artificial dumbness
artificial incompetence
artificial lack of intelligence
artificial stupidity
artificial unintelligence
artificially intelligent
Context of this term
computing
science fiction
Things used for
artificial intelligence
Etymologically rela
কৃত্রিম বুদ্ধিমত্তা
artificial dumbness
artificial idiocy
artificial incompetence
Similar terms
expert system
expert systems
herbert simon
marvin minsky
artificial intelligence is
defined as...
Derived from
multi agent system
People known for
artificial intelligenc

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:

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

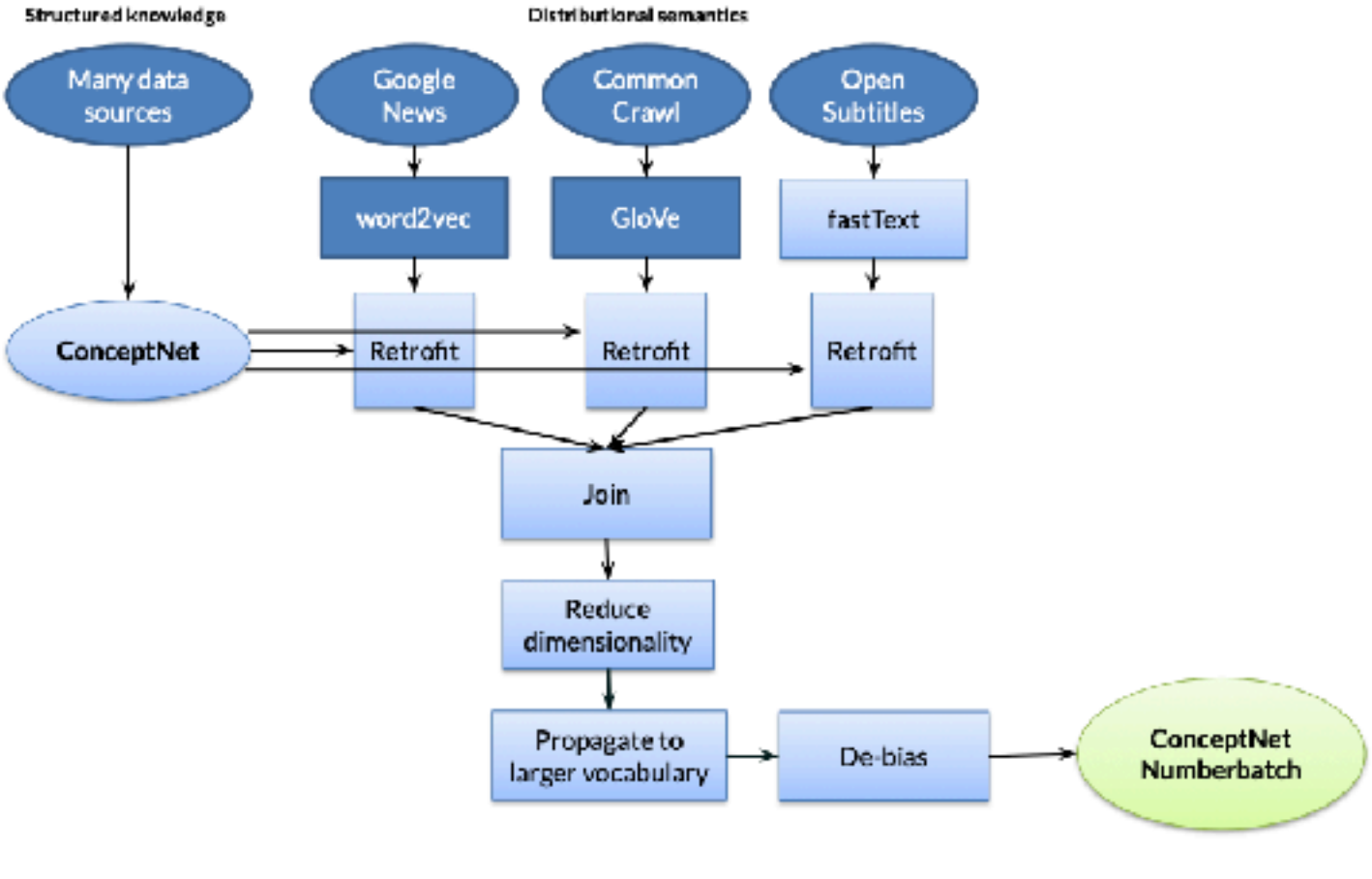
new embed — old embed — neighbors from KG

(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

Building ConceptNet Numberbatch



LUMINOSO

Aside: Transparency in Research

ConceptNet is all you need

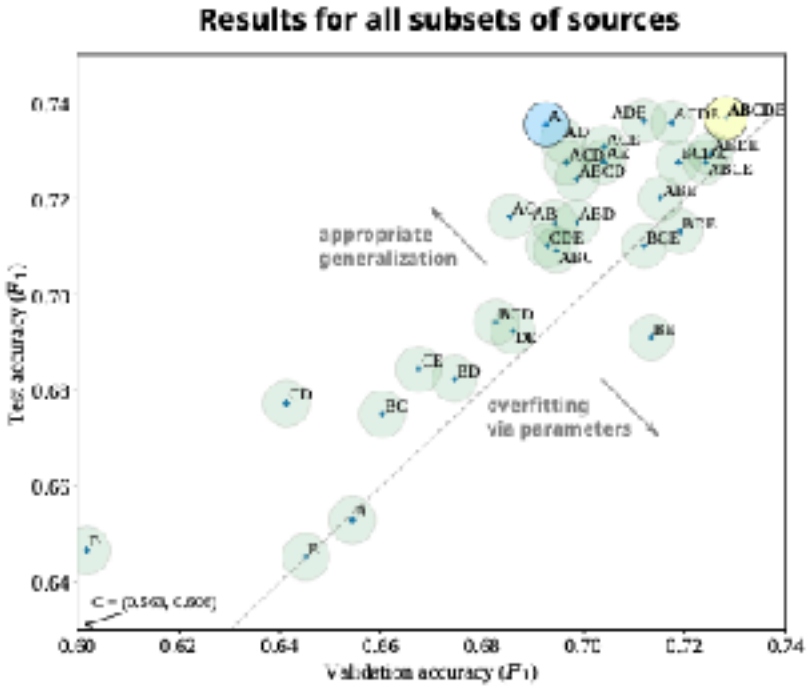
Our full classifier used the linear combination of 5 types of input features shown above. This point is labeled **ABCDE** on the graph to the right. The other points are ablated versions of the classifier, trained on subsets of the five sources.

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

Lecture Notes for

Neural Networks and Machine Learning

Ethically Aware Practices

Next Time:
Transfer Learning

Reading: Chollet Article