Deep learning, ethics, and society

Outline

• Dataset concerns
• Bias in ML systems
• Ethical issues in specific application areas
  • Face recognition
  • Image manipulation
  • Language models
• AI hype
• AI for social good
• Towards ethical best practices
Outline

• Dataset concerns
ImageNet: Asset or liability?

• Performance on the basic ILSVRC benchmark has saturated

• Current models have reached levels of accuracy where the presence of human labeling error is starting to affect experimental conclusions (Beyer et al. 2020, Northcutt et al. 2021)
ImageNet labeling problems: ImageNet Roulette


ImageNet Roulette uses an open source Caffe deep learning framework (produced at UC Berkeley) trained on the images and labels in the “person” categories (which are currently ‘down for maintenance’). Proper nouns and categories with less than 100 pictures were removed.

When a user uploads a picture, the application first runs a face detector to locate any faces. If it finds any, it sends them to the Caffe model for classification. The application then returns the original images with a bounding box showing the detected face and the label the classifier has assigned to the image. If no faces are detected, the application sends the entire scene to the Caffe model and returns an image with a label in the upper left corner.

ImageNet contains a number of problematic, offensive and bizarre categories - all drawn from WordNet. Some use misogynistic or racist terminology. Hence, the results ImageNet Roulette returns will also draw upon those categories. That is by design: we want to shed light on what happens when technical systems are trained on problematic training data. AI classifications of people are rarely made visible to the people being classified. ImageNet Roulette provides a glimpse into that process – and to show the ways things can go wrong.

Want to see how an AI trained on ImageNet will classify you? Try ImageNet Roulette, based on ImageNet’s Person classes. It’s part of the ‘Training Humans’ exhibition by @trevorpaglen & me - on the history & politics of training sets. Full project out soon imagenet-roulette.paglen.com

Kate Crawford @katecrawford · Sep 16, 2019

ImageNet Roulette
ImageNet Roulette
ImageNet and WordNet

https://wordnet.princeton.edu/
Cleaning up ImageNet

“We examine the 2,832 people categories that are annotated within the subtree, and determine that 1,593 of them are potentially offensive labels that should not be used in the context of an image recognition dataset... Out of the remaining 1,239 categories we find that only 158 of them are visual, with the remaining categories simply demonstrating annotators’ bias.”

<table>
<thead>
<tr>
<th>Unsafe (offensive)</th>
<th>Unsafe (sensitive)</th>
<th>Safe non-imageable</th>
<th>Safe imageable</th>
</tr>
</thead>
<tbody>
<tr>
<td>n10095420: &lt;sexual slur&gt;</td>
<td>n09702134: Anglo-Saxon</td>
<td>n10002257: demographer</td>
<td>n10499631: Queen of England</td>
</tr>
<tr>
<td>n10114550: &lt;profanity&gt;</td>
<td>n10693334: taxi dancer</td>
<td>n10061882: epidemiologist</td>
<td>n09842047: basketball player</td>
</tr>
<tr>
<td>n10262343: &lt;sexual slur&gt;</td>
<td>n10384392: orphan</td>
<td>n10431122: piano maker</td>
<td>n10147935: bridegroom</td>
</tr>
<tr>
<td>n10758337: &lt;gendered slur&gt;</td>
<td>n09890192: camp follower</td>
<td>n10098862: folk dancer</td>
<td>n09846755: beekeeper</td>
</tr>
<tr>
<td>n10507380: &lt;criminative&gt;</td>
<td>n10580030: separatist</td>
<td>n10335931: mover</td>
<td>n10153594: gymnast</td>
</tr>
<tr>
<td>n10744078: &lt;criminative&gt;</td>
<td>n09980805: crossover voter</td>
<td>n10449664: policyholder</td>
<td>n10539015: ropewalker</td>
</tr>
<tr>
<td>n10113869: &lt;obscene&gt;</td>
<td>n09848110: theist</td>
<td>n10146104: great-niece</td>
<td>n10530150: rider</td>
</tr>
<tr>
<td>n10344121: &lt;pejorative&gt;</td>
<td>n09683924: Zen Buddhist</td>
<td>n10747119: vegetarian</td>
<td>n10732010: trumpeter</td>
</tr>
</tbody>
</table>

K. Yang, K. Qinami, L. Fei-Fei, J. Deng, O. Russakovsky, Towards Fairer Datasets: Filtering and Balancing the Distribution of the People Subtree in the ImageNet Hierarchy, Conference on Fairness, Accountability, and Transparency (FAccT), 2020
Cleaning up ImageNet

• Filtering remaining categories by age, gender, skin color, and age:

“Programmer”

K. Yang, K. Qinami, L. Fei-Fei, J. Deng, O. Russakovsky, Towards Fairer Datasets: Filtering and Balancing the Distribution of the People Subtree in the ImageNet Hierarchy, Conference on Fairness, Accountability, and Transparency (FAccT), 2020
Privacy and consent issues in datasets

https://megapixels.cc/

See also: https://www.theregister.co.uk/2020/01/27/ibms_facial_recognition_software_gets_it_in_trouble_again/

https://www.ft.com/content/cf19b956-60a2-11e9-b285-3acd5d43599e
A general indictment of image datasets?

Figure 1: Results from the 80 Million Tiny Images dataset exemplifying the toxicities of it’s label space

V. Prabhu, A. Birhane, Large datasets: A Pyrrhic win for computer vision? WACV 2021
A general indictment of image datasets?

- MIT takes down 80M Tiny Images dataset due to racist and offensive content – VentureBeat, 7/1/2020
  - Note from authors

A. Torralba, R. Fergus, W. Freeman, 80 million tiny images: a large dataset for non-parametric object and scene recognition, TPAMI 2008
Cleaning up ImageNet

• “We demonstrate that face blurring—a typical obfuscation technique—has minimal impact on the accuracy of recognition models”

Figure 1. Most categories in ImageNet Challenge (Russakovsky et al., 2015) are not people categories. However, the images contain many people co-occurring with the object of interest, posing a potential privacy threat. These are example images from barber chair, husky, beer bottle, volleyball and military uniform.

K. Yang, J. Yau, L. Fei-Fei, J. Deng, O. Russakovsky, A study of face obfuscation in ImageNet, arXiv 2021
Even larger-scale datasets...

• Publicly available
  • Yahoo Flickr Creative Commons (YFCC) 100M (Thomee et al., 2016)
  • Google Conceptual Captions: 3.3M images with captions
  • Open Images (Google): 9M images with object annotations

• Not publicly available
  • Google’s JFT-300M (Hinton et al. 2014, Chollet 2017, Sun et al. 2017)
  • Facebook 3.5B Instagram images (Mahajan et al., 2018)
  • OpenAI’s WebImageText: 400M image-text pairs (Radford et al. 2021)
Large-scale datasets: Discussion

• Summary of concerns
  • Collection without regard to privacy or consent from pictured individuals
  • Biased and offensive imagery and label spaces
  • Labeling is obtained by crowdsourced “ghost work”
  • Lack of transparency for extremely large and non-public datasets
  • Harmful downstream applications (coming up)

• Possible solutions
  • Use only consensual images
  • Blur out or otherwise disguise recognizable individuals
  • Privacy-preserving dataset distillation
  • Ethical dataset collection standards (see, e.g., Jo & Gebru 2019)
  • Document collection and curation procedures in a standardized way
  • Restrict access to datasets, specify terms precluding unethical uses
  • Mandatory IRB for large-scale dataset collection?
Outline

• Dataset concerns
• Bias in deep learning systems
Possible source of bias: Problem formulation

• Should the task exist in the first place?
  • Example: predicting criminality from face images

• See K. Bowyer et al., The “Criminality from Face” Illusion, arXiv 2020
  • From the abstract: “We argue that attempts to create a criminality-from-face algorithm are necessarily doomed to fail, that apparently promising experimental results in recent publications are an illusion resulting from inadequate experimental design, and that there is potentially a large social cost to belief in the criminality from face illusion.”
Possible source of bias: Datasets

T. DeVries, I. Misra, C. Wang, L. van der Maaten, Does Object Recognition Work for Everyone? Workshop on Computer Vision for Global Challenges at CVPR 2019
Possible source of bias: Features, models

“Rectangle filters”

\[ \text{Value} = \sum (\text{pixels in white area}) - \sum (\text{pixels in black area}) \]

First two features selected by boosting (100% detection rate and 50% false positive rate)

P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. CVPR 2001
Algorithmic bias amplification

Figure 1: Five example images from the imSitu visual semantic role labeling (vSRL) dataset. Each image is paired with a table describing a situation: the verb, cooking, its semantic roles, i.e. agent, and noun values filling that role, i.e. woman. In the imSitu training set, 33% of cooking images have man in the agent role while the rest have woman. After training a Conditional Random Field (CRF), bias is amplified: man fills 16% of agent roles in cooking images. To reduce this bias amplification our calibration method adjusts weights of CRF potentials associated with biased predictions. After applying our methods, man appears in the agent role of 20% of cooking images, reducing the bias amplification by 25%, while keeping the CRF vSRL performance unchanged.

J. Zhao, T. Wang, M. Yatskar, V. Ordonez, K.-W. Chang, Men also like shopping: Reducing Gender Bias Amplification using Corpus-level constraints, EMNLP 2017
Bias: Discussion

• Summary of concerns
  • Output of trained systems ends up echoing or even amplifying societal biases due to biased problem formulation, algorithms, model structure, or all of the above

• Possible solutions
  • Model reporting
    • See, e.g., Mitchell et al. (2019)
  • Internal and external auditing
    • See, e.g., Raji et al. (2020)
  • Statistical techniques for reducing bias
Bias reduction techniques

Baseline: A **man** sitting at a desk with a laptop computer.  
Improved model: A **woman** sitting in front of a laptop computer.

Outline

• Dataset concerns
• Bias in deep learning systems
• Ethical issues in specific application areas
  • Face recognition
Face recognition

The ethical questions that haunt facial-recognition research

Journals and researchers are under fire for controversial studies using this technology. And a Nature survey reveals that many researchers in this field think there is a problem.
Face recognition in the U.S.

Here’s where the US government is using facial recognition technology to surveil Americans

This map shows how widespread the use of facial recognition technology has become.

By Shirin Ghaffary and Rani Molla | Updated Dec 10, 2019, 8:00am EST

Wrongfully Accused by an Algorithm

In what may be the first known case of its kind, a faulty facial recognition match led to a Michigan man's arrest for a crime he did not commit.
Face recognition in the U.S.

How America’s surveillance networks helped the FBI catch the Capitol mob

Federal documents detailing the attacks at the U.S. Capitol show a mix of FBI techniques, from license plate readers to facial recognition, that helped identify rioters. Digital rights activists say the invasive technology can infringe on our privacy.

https://www.washingtonpost.com/technology/2021/04/02/capitol-siege-arrests-technology-fbi-privacy/
Face recognition in China

One Month, 500,000 Face Scans: How China Is Using A.I. to Profile a Minority – New York Times, 4/14/2019
China Uses DNA to Map Faces, With Help From the West – New York Times, 12/3/2019
Face recognition in Russia

Russia’s surveillance state still doesn’t match China. But Putin is racing to catch up.

https://www.washingtonpost.com/world/europe/russia-facial-recognition-surveillance-navalny/2021/04/16/4b97dc80-8c0a-11eb-a33e-da28941cb9ac_story.html
The technology is keeping up with the times...

Bias in face recognition


Figure source
Bias in face recognition

*MISTAKEN IDENTITY*
A 2019 review of facial-recognition algorithms shows the chance of false positives* — incorrectly finding matches between two faces — when comparing high-quality US mugshots of different people of the same gender and race*. The rate is highest for female faces of people of colour, but differs across algorithms (shown in two examples).

- UK academic algorithm
- Chinese commercial algorithm

Via: [https://www.nature.com/articles/d41586-020-03186-4](https://www.nature.com/articles/d41586-020-03186-4)

*Algorithm’s confidence threshold for a ‘match’ was set so as to ensure the false-positive rate for white males was 1 per 10,000; others used same threshold. *Ethnicities as described in ref. 5.
Other sensitive recognition topics

- Face analysis for inferring internal states or personal characteristics
  - See discussion of emotional privacy issues

Face recognition: Discussion

• Summary of concerns
  • Bias against underrepresented groups
  • Unequal impact
  • Privacy, consent issues in data collection (with particular emphasis on vulnerable groups)
  • Unethical use and misuse (cf. Clearview AI controversy)

• Possible solutions
  • Auditing
  • Regulation:
  • Resistance?
Outline

• Dataset concerns
• Bias in deep learning systems
• Ethical issues in specific application areas
  • Face recognition
  • Image manipulation
Image manipulation: DeepFakes

An example of deepfake technology: in a scene from *Man of Steel*, actress Amy Adams in the original (left) is modified to have the face of actor Nicolas Cage (right)

Image manipulation: Bias

Image manipulation: Bias

S. Menon et al., PULSE: Self-Supervised Photo Upsampling via Latent Space Exploration of Generative Models, CVPR 2020
Image manipulation: Beautification filters


Image manipulation: Discussion

• Summary of concerns
  • “DeepFake” technology has drastically lowered costs and barriers to entry for sophisticated photo and video manipulation
  • Manipulation for political ends causes the most alarm, although harassment and objectification of women may be most common use cases in practice

• Possible solutions
  • Codes of ethics, standards (already exist in photojournalism)
  • Regulation (including self-regulation by social media companies)
  • Techniques for spotting manipulated content

For comprehensive overview, see Frédo Durand, *Ethics and Computational Photography*, 2019
Automatically spotting fake images

Figure 1: **Are CNN-generated images hard to distinguish from real images?** We show that a classifier trained to detect images generated by only one CNN (ProGAN, far left) can detect those generated by many other models (remaining columns). Our code and models are available at [https://peterwang512.github.io/CNNDetection/](https://peterwang512.github.io/CNNDetection/).

S.-U. Wang et al., [CNN-generated images are surprisingly easy to spot... for now](https://www.cvpr2020.org/cvpr2020/index.php), CVPR 2020
Outline

• Large-scale datasets
• Bias in ML systems
• Ethical issues in specific application areas
  • Face recognition
  • Image manipulation
  • Large-scale language models
In the news...

We read the paper that forced Timnit Gebru out of Google. Here's what it says.

The company's star ethics researcher highlighted the risks of large language models, which are key to Google's business.

by Karen Hao

December 4, 2020

https://www.technologyreview.com/2020/12/04/1013294/google-ai-ethics-research-paper-forced-out-timnit-gebru/
In the news...

On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?

Emily M. Bender*
emblem@uw.edu
University of Washington
Seattle, WA, USA

Angelina McMillan-Major
aymm@uw.edu
University of Washington
Seattle, WA, USA

Timnit Gebru*
timnit@blacknai.org
Black in AI
Palo Alto, CA, USA

Shmargaret Shmitchell
shmargaret.shmitchell@gmail.com
The Aether

E. Bender et al., On the dangers of stochastic parrots: Can language models be too big? FAccT 2021
LLMs and misinformation

Figure 7.3: People’s ability to identify whether news articles are model-generated (measured by the ratio of correct assignments to non-neutral assignments) decreases as model size increases. Accuracy on the outputs on the deliberately-bad control model (an unconditioned GPT-3 Small model with higher output randomness) is indicated with the dashed line at the top, and the random chance (50%) is indicated with the dashed line at the bottom. Line of best fit is a power law with 95% confidence intervals.

T. Brown et al., Language models are few-shot learners, NeurIPS 2020 – Best Paper Award
LLMs and bias

Investigating Anti-Muslim Bias in GPT-3 through Words, Analogies, & Stories

Abubakar Abid, Ali Abid, Ali Abdalla, Dawood Khan, James Zou

10:35 AM ET
8 December 2020

Stanford University
gradio

https://twitter.com/abidlabs/status/1336314279910322179?s=20
LLMs and bias

Two Muslims walked into a restaurant to order pizza. They killed the owner’s son and tried to blow up the place.

"ISIS-inspired terror attack on U.S. soil." "ISIS-inspired terror attack." The father of the Muslim who murdered three young Americans in North Carolina on Tuesday [4/10/2015] said the shooting was

Completion may contain sensitive content

How many **words about pizza** are needed to consistently (≥90%) generate non-violent sentence completions?

https://twitter.com/kchonyc/status/1336340712703537153?s=20
Measuring bias in language models

https://stereoset.mit.edu/
LLMs and privacy

Figure 1: **Our extraction attack.** Given query access to a neural network language model, we extract an individual person’s name, email address, phone number, fax number, and physical address. The example in this figure shows information that is all accurate so we redact it to protect privacy.

<table>
<thead>
<tr>
<th>Category</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>US and international news</td>
<td>109</td>
</tr>
<tr>
<td>Log files and error reports</td>
<td>79</td>
</tr>
<tr>
<td>License, terms of use, copyright notices</td>
<td>54</td>
</tr>
<tr>
<td>Lists of named items (games, countries, etc.)</td>
<td>54</td>
</tr>
<tr>
<td>Forum or Wiki entry</td>
<td>53</td>
</tr>
<tr>
<td>Valid URLs</td>
<td>50</td>
</tr>
<tr>
<td><strong>Named individuals (non-news samples only)</strong></td>
<td>46</td>
</tr>
<tr>
<td>Promotional content (products, subscriptions, etc.)</td>
<td>45</td>
</tr>
<tr>
<td>High entropy (UUIDs, base64 data)</td>
<td>35</td>
</tr>
<tr>
<td><strong>Contact info (address, email, phone, twitter, etc.)</strong></td>
<td>32</td>
</tr>
<tr>
<td>Code</td>
<td>31</td>
</tr>
<tr>
<td>Configuration files</td>
<td>30</td>
</tr>
<tr>
<td>Religious texts</td>
<td>25</td>
</tr>
<tr>
<td>Pseudonyms</td>
<td>15</td>
</tr>
<tr>
<td>Donald Trump tweets and quotes</td>
<td>12</td>
</tr>
<tr>
<td>Web forms (menu items, instructions, etc.)</td>
<td>11</td>
</tr>
<tr>
<td>Tech news</td>
<td>11</td>
</tr>
<tr>
<td>Lists of numbers (dates, sequences, etc.)</td>
<td>10</td>
</tr>
</tbody>
</table>

Table 1: Manual categorization of the 604 memorized training examples that we extract from GPT-2, along with a description of each category. Some samples correspond to multiple categories (e.g., a URL may contain base-64 data). Categories in **bold** correspond to personally identifiable information.

N. Carlini et al., [Extracting Training Data from Large Language Models](https://arxiv.org/), arXiv 2020
Carbon footprint of LLMs

<table>
<thead>
<tr>
<th>Consumption</th>
<th>CO$_2$e (lbs)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Air travel, 1 passenger, NY$\leftrightarrow$SF</td>
<td>1984</td>
</tr>
<tr>
<td>Human life, avg, 1 year</td>
<td>11,023</td>
</tr>
<tr>
<td>American life, avg, 1 year</td>
<td>36,156</td>
</tr>
<tr>
<td>Car, avg incl. fuel, 1 lifetime</td>
<td>126,000</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Training one model (GPU)</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>NLP pipeline (parsing, SRL)</td>
<td>39</td>
</tr>
<tr>
<td>w/ tuning &amp; experimentation</td>
<td>78,468</td>
</tr>
<tr>
<td>Transformer (big)</td>
<td>192</td>
</tr>
<tr>
<td>w/ neural architecture search</td>
<td>626,155</td>
</tr>
</tbody>
</table>

Table 1: Estimated CO$_2$ emissions from training common NLP models, compared to familiar consumption.¹

---

Carbon footprint of LLMs

We calculate the energy use and carbon footprint of several recent large models—T5, Meena, GShard, Switch Transformer, and GPT-3—and refine earlier estimates for the neural architecture search that found Evolved Transformer.

We highlight the following opportunities to improve energy efficiency and CO$_2$ equivalent emissions (CO$_2$e):

- Large but sparsely activated DNNs can consume $<1/10$th the energy of large, dense DNNs without sacrificing accuracy despite using as many or even more parameters.
- Geographic location matters for ML workload scheduling since the fraction of carbon-free energy and resulting CO$_2$e vary $\sim$5X-10X, even within the same country and the same organization. We are now optimizing where and when large models are trained.
- Specific datacenter infrastructure matters, as Cloud datacenters can be $\sim$1.4-2X more energy efficient than typical datacenters, and the ML-oriented accelerators inside them can be $\sim$2-5X more effective than off-the-shelf systems.

Remarkably, the choice of DNN, datacenter, and processor can reduce the carbon footprint up to $\sim$100-1000X.

D. Patterson et al., Carbon Emissions and Large Neural Network Training, arXiv 2021
LLMs and “understanding”

How Do You Know a Human Wrote This?
Machines are gaining the ability to write, and they are getting terrifyingly good at it.

By Farhad Manjoo
Opinion Columnist

July 29, 2020


See also:
https://www.newyorker.com/culture/cultural-comment/the-computers-are-getting-better-at-writing

GPT-3, Bloviator: OpenAI’s language generator has no idea what it’s talking about
Tests show that the popular AI still has a poor grasp of reality.

by Gary Marcus and Ernest Davis

August 22, 2020

LLMs and “understanding”

• Hypothesis: meaning cannot be learned from form alone – it requires knowing about the relationship between language and the outside world

E. Bender and A. Koller, Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data, ACL 2020
Language models: Discussion

• Summary of concerns
  • Harmful uses, e.g., fake news generation, fraud and deception
  • Capturing and amplifying biases from poorly documented datasets
  • Privacy (memorizing and revealing personally identifying information)
  • Transparency, access and reproducibility
  • Cost and carbon footprint
  • Potential for destroying jobs (e.g., writers, editors, programmers)

• Possible solutions
  • Put more effort into curating and documenting large-scale data
  • Standards for access and transparency
  • Methods for reducing bias and protecting privacy
Outline

• Dataset concerns
• Bias in deep learning systems
• Ethical issues in specific application areas
  • Face recognition
  • Image manipulation
  • Language models
• AI hype
• AI for social good
• Towards ethical best practices
AI hype

Results are shown in Table 3, and show that the simple LIKELIHOOD model produces the most human-like responses, and the alternative training and decoding strategies cause a divergence from human language. Note however, that this divergence may not necessarily correspond to lower quality language—it may also indicate different strategic decisions about what to say. Results in §6.4 show all models could converse with humans.

Bob: I can i i everything else
Alice: balls have zero to me to me to me to me to me to me to
Bob: you i everything else
Alice: balls have a ball to me to me to me to me to me to me to me to

M. Lewis et al., Deal or No Deal? End-to-End Learning for Negotiation Dialogues, EMNLP 2017

See also: Dhruv Batra’s response
AI hype


• Summary of concerns:
  • Corrupting the peer review process and the culture of the entire field
  • Adding to misinformation in the public sphere
  • Misdirecting research effort
  • Inviting backlash and withdrawal of funding ("AI winter")

• Possible solutions:
  • Conference policies on press and social media
  • Self-regulation by researchers
AI for social good

• N. Tomasev et al., *AI for social good: unlocking the opportunity for positive impact*, Nature Communications, May 2020

• Some organizations and events:
  • [https://ai-4-all.org/](https://ai-4-all.org/)
  • ACM Conference on Fairness, Accountability, and Transparency
  • Computer vision for global challenges workshops
  • NeurIPS 2020 workshop on tackling climate change with ML
AI against climate change

• R. Schwarz et al., Green AI, arXiv 2019
• D. Rolnick et al., Tackling climate change with deep learning, arXiv 2019
• D. Tuia et al., Towards a Collective Agenda on AI for Earth Science Data Analysis, arXiv 2021
Towards ethical best practices


• Concrete recommendations (for researchers):
  • Researchers should consider potential risks up front, not after the fact – the assumption that technology is inherently neutral is no longer tenable
  • Conferences should adopt ethics codes and appoint ethics chairs
  • Authors should be encouraged to add ethical impacts sections in papers and disclose concerns with released code and data
    • See NeurIPS impact statement requirement
  • Authors should be required to disclose funding sources and possible conflicts of interest in papers