Lecture 25: Bias and ethics
Announcements

• Sign up for a final presentation time slot!

• Discussion section this week: mostly office hours, but also will discuss NeRF.

• Questions?
A machine learning algorithm will do whatever the training data tells it to do.

If the data is bad or biased, the learned algorithm will be too.

Source: Isola, Torralba, Freeman
Microsoft's Tay chatbot

Chatbot released on twitter.

Learned from interactions with users

Started mimicking offensive language, was shut down.

Recall: the Giraffe-Tree problem

A giraffe standing in the grass next to a tree.

[“Measuring Machine Intelligence Through Visual Question Answering”, Zitnick et al., 2016]
Source: Isola, Torralba, Freeman

["Colorful image colorization", Zhang et al., ECCV 2016]
[“Colorful image colorization”, Zhang et al., ECCV 2016]
Source: Isola, Torralba, Freeman

[“Colorful image colorization”, Zhang et al., ECCV 2016]
[from Reddit /u/SherySantucci]
Revisiting generalization
What Google thinks are student bedrooms

Source: Isola, Torralba, Freeman
Driving simulator (GTA)  

Driving in the real world  

Need learning methods that can bridge this domain gap!

Source: Isola, Torralba, Freeman
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.

Revisiting the problem of generalization
True data-generating process: \( p_{\text{data}} \)

\[
\{x_i^{(\text{train})}, y_i^{(\text{train})}\}_{i=1}^N \sim p_{\text{data}} \\
\{x_i^{(\text{test})}, y_i^{(\text{test})}\}_{i=1}^M \sim p_{\text{data}}
\]

Source: Isola, Torralba, Freeman
Almost never true in practice!

This is a huge assumption! Almost never true in practice!

Source: Isola, Torralba, Freeman
Much more commonly, we have

\[ p_{\text{train}} \neq p_{\text{test}} \]
Our training data didn’t cover the part of the distribution that was tested **(biased data)**

Source: Isola, Torralba, Freeman
Lots of issues deploying biased systems

- Runaway feedback loops
  - E.g. training a machine learning system on biased hiring decisions results in more biased hiring decisions.
- Bias in face analysis tools
- Perpetuate gender stereotypes

Source: [Gebru, “Race and Gender”, 2019]
What you might take away from a class

#1: The model

#2: The algorithm

#3: The data

Source: Alexei Efros
But in practice…

#1: The data

#2: The model

#3: The algorithm

Source: Alexei Efros
How can we collect good data?

+ Correctly labeled
+ Unbiased (good coverage of all relevant kinds of data)

Source: Isola, Torralba, Freeman
Crowdsourcing
The value of data

The Large Hadron Collider
$10^{10}$

Amazon Mechanical Turk
$25 \times 10^2 - 10^4$

Source: Isola, Torralba, Freeman
But can humans collect good data?
Biases in data collection

Source: Isola, Torralba, Freeman
Getting more humans in the annotation loop

Labeling to get a Ph.D.

Labeling for money
(Sorokin, Forsyth, 2008)

Labeling for fun
(Luis Von Ahn and Laura Dabbish 2004)

Labeling because it gives you added value

Visipedia
(Belongie, Perona, et al)

Labeling to prove you’re human

Source: Isola, Torralba, Freeman
Beware of the human in your loop

• What do you know about them?
• Will they do the work you pay for?

Let’s check a few simple experiments

Source: Isola, Torralba, Freeman
People have biases...

Turkers were offered 1 cent to pick a number from 1 to 10.

~850 turkers

Source: Isola, Torralba, Freeman

From http://groups.csail.mit.edu/uid/deneme/
Do humans have consistent biases?

Results form 100 HITS:

Experiment by Greg Little

From http://groups.csail.mit.edu/uid/deneme/

Source: Isola, Torralba, Freeman
Are humans reliable even in simple tasks?

Results of 100 HITS:
A: 2
B: 96
C: 2

Source: Isola, Torralba, Freeman

From http://groups.csail.mit.edu/uid/deneme/
Do humans do what you ask for?

Flip a coin

Requester: ROBERT C MILLER  Reward: $0.01 per HIT  HITs Available: 3  Duration: 5 minutes
Qualifications Required: None

Please flip an actual coin and type either H or T below.

After 50 HITS: 31 heads, 19 tails
And 50 more: 34 heads, 16 tails

Experiment by Rob Miller
From http://groups.csail.mit.edu/uid/deneme/

Source: Isola, Torralba, Freeman
So we can sometimes collect good training data.

But suppose we messed up. Our test setting doesn’t look like the training data!

How can we bridge the domain gap?
Finding more representative images

[Places365 Kitchen]

[Fouhey et al., "From Lifestyle Vlogs to Everyday Actions", 2017]
Finding more representative images

VLOG Kitchen

[Fouhey et al., "From Lifestyle Vlogs to Everyday Actions", 2017]
Name that dataset game

Domain gap between $p_{\text{train}}$ and $p_{\text{test}}$ will cause us to fail to generalize.

Source: Isola, Torralba, Freeman
Domain gap between $P_{\text{source}}$ and $P_{\text{target}}$ will cause us to fail to generalize.

Source: Isola, Torralba, Freeman
Idea #1: transform the target domain to look like the source domain

(Or vice versa)

This is called **domain adaptation**

Source: Isola, Torralba, Freeman
Domain adaptation

- We have source domain pairs \{x_{\text{source}}, y_{\text{source}}\}

- Learn a mapping F: \(x_{\text{source}} \rightarrow y_{\text{source}}\)

- We want to apply F to target domain data \(x_{\text{target}}\)

- Find transformation T: \(x_{\text{target}} \rightarrow x_{\text{source}}\)

- Now apply F(T(\(x_{\text{target}}\))) to predict \(y_{\text{target}}\)

Source: Isola, Torralba, Freeman
\[ p_{\text{source}} \rightarrow p_{\text{target}} \]

Source: Isola, Torralba, Freeman
CycleGAN

Horses

Zebras

Source: Isola, Torralba, Freeman
Domain adaptation

$p_{\text{source}} \rightarrow p_{\text{target}}$

Source: Isola, Torralba, Freeman
Domain gap between $p_{\text{source}}$ and $p_{\text{target}}$ will cause us to fail to generalize.

Source: Isola, Torralba, Freeman
Cycle-Consistent Adversarial Domain Adaptation

Source domain

Target domain

[Hoffman, Tzeng, Park, Zhu, Isola, Saenko, Darrell, Efros, arXiv 2017]

Source: Isola, Torralba, Freeman
CycleGAN

Source: Isola, Torralba, Freeman
CycleGAN

Training data

Source: Isola, Torralba, Freeman
CycleGAN  

Training data  

FCN  

Source: Isola, Torralba, Freeman
OpenAI Dactyl

FINGER PIVOTING  SLIDING  FINGER GAITING

Source: Isola, Torralba, Freeman
Domain gap between $p_{source}$ and $p_{target}$ will cause us to fail to generalize.
Idea #2: train on randomly perturbed data, so that test set just looks like another random perturbation.

This is called **domain randomization** or **data augmentation**.

Source: Isola, Torralba, Freeman
Domain randomization

Training data

Test data

[Sadeghi & Levine 2016]

Above example is from [Tobin et al. 2017]
Beyond data

- Data very important [Maluleke et al., 2022], but also other factors can matter.
- Camera hardware and software
  - e.g., default camera settings calibrated to expose light skin
- Loss function (e.g., “mode collapse” in GANs)
- Features
- Sampling strategy (e.g., truncation in GANs)
What if we go way outside of the training distribution?
Our training data did not cover the part of the distribution that was tested (biased data)

Source: Isola, Torralba, Freeman
“Deep Neural Networks are Easily Fooled: High Confidence Predictions for Unrecognizable Images”  
[Nguyen, Yosinski, and Clune, CVPR 2015]

Source: Isola, Torralba, Freeman
“Deep Neural Networks are Easily Fooled: High Confidence Predictions for Unrecognizable Images” [Nguyen, Yosinski, and Clune, CVPR 2015]

Source: Isola, Torralba, Freeman
Adversarial noise

\[ \text{arg max}_r p(y = \text{ostrich}|x + r) \quad \text{subject to} \quad \|r\| < \epsilon \]

Anything to worry about?

“NO Need to Worry about Adversarial Examples in Object Detection in Autonomous Vehicles”, Lu et al. 2017

(Early) 2017’s attacks fail on physical objects, since they are optimized to attack a single view!
Anything to worry about?

Later in 2017…

“Synthesizing Robust Adversarial Examples”, Athalye, Engstrom, Ilyas, Kwok, 2017

3D-printed turtle model classified as rifle from most viewpoints

Source: Isola, Torralba, Freeman
Adversarial examples

- Current deep models have bad **worst-case performance**
- Can be exploited by an adversary
- Few guarantees, can’t fully trust what the model’s output

Source: Isola, Torralba, Freeman
Problems of applying computer vision in practice
Mission-critical computer vision systems

Social consequences

Color Matters in Computer Vision
Facial recognition algorithms made by Microsoft, IBM and Face++ were more likely to misidentify the gender of black women than white men.

Gender was misidentified in **up to 1 percent of lighter-skinned males** in a set of 385 photos.

Gender was misidentified in **35 percent of darker-skinned females** in a set of 271 photos.


Source: Isola, Torralba, Freeman
Algorithmic Bias

http://gendershades.org/overview.html


Source: Isola, Torralba, Freeman
Algorithmic Bias

Benchmarking Algorithmic Bias in Face Recognition: An Experimental Approach Using Synthetic Faces and Human Evaluation

Hao Liang
Rice University
hli050@rice.edu

Pietro Perona
California Institute of Technology and AWS
perona@caltech.edu, perona@apple.com

Guha Balakrishnan
Rice University
guha@rice.edu

Abstract

We propose an experimental method for measuring bias in face recognition systems. Existing methods to measure bias depend on benchmark datasets that are collected in the wild and annotated for protected (e.g., race, gender) and unprotected (e.g., pose, lighting) attributes. Such observational datasets only permit correlational conclusions, e.g., “Algorithm A’s accuracy is different on female and male faces in dataset X.” By contrast, experimental methods manipulate attributes individually and thus permit causal conclusions, e.g., “Algorithm A’s accuracy is affected by gender and skin color.”

Our method is based on generating synthetic faces using a neural face generator, where each attribute of interest is modified independently while leaving all other attributes constant. Human observers crucially provide the ground truth on perceptual identity similarity between synthetic image pairs. We validate our method quantitatively by evaluating race and gender biases of three research-grade face recognition models. Our synthetic pipeline reveals that for these algorithms, accuracy is lower for Black and East Asian population subgroups. Our method can also quantify how perceptual changes in attributes affect face identity distances reported by these models. Our large synthetic (“face identification”) face recognition systems implemented with deep neural networks today achieve impressive accuracies [34, 11, 31, 67] and outperform even expert face analysts [39]. Nevertheless, it is important to detect and measure possible algorithmic biases, i.e., systematic accuracy differences, especially across protected demographic attributes like age, race and gender [19, 20, 21], in order to maintain fair treatment in sensitive applications. For this reason, the National Institute of Standards and Technology (NIST) measures bias in commercial face recognition models [17], in particular by comparing their False Match Rate (FMR) and False Non Match Rate (FNMR) values across different demographic subgroups at a particular decision threshold (swerving this threshold yields FNMR vs. FMR “curves”).

The first step in measuring bias of face recognition systems is, currently, to collect a large benchmarking dataset containing a set of diverse faces, where each is photographed multiple times under different conditions. An algorithm’s error rates across subgroups specified by different protected attribute combinations (e.g., different race and gender groups) can then be measured. Unfortunately, sampling a good test dataset is almost impossible. First, each protected intersectional group (a specific combination of attribute values) must contain a suffi-
Bad choice of data

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


Source: S. Lazebnik
Fears of universal mass surveillance (and dubious claims)

The Secretive Company That Might End Privacy as We Know It

A little-known start-up helps law enforcement match photos of unknown people to their online images — and “might lead to a dystopian future or something,” a backer says.

https://www.buzzfeednews.com/article/ryanmac/clearview-ai-nypd-facial-recognition

Source: S. Lazebnik
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.
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](https://example.com), [Northcutt et al. 2021](https://example.com))

Source: S. Lazebnik
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.

Source: S. Lazebnik
ImageNet Roulette

Source: S. Lazebnik
Some things to worry about…

- Our datasets are often poorly labeled

- And usually biased

- ML methods may perform well on lab-collected data, but often generalize poorly to real-world data

- Can have negative social consequences

Adapted from Isola, Torralba, Freeman
Open-ended discussion

• Supervised vs. unsupervised learning?
• Other negative consequences of computer vision systems?
• What other biases might computer vision systems have?
Thank you!