The Seen and Unseen Factors Influencing Knowledge in AI Systems

Margaret Mitchell
Google
What do you see?
What do you see?

- Bananas
What do you see?

- Bananas
- Dole Bananas
What do you see?

- Bananas
- Dole Bananas
- Bananas at a store
What do you see?

- Bananas
- Dole Bananas
- Bananas at a store
- Bananas on shelves
What do you see?

- Bananas
- Dole Bananas
- Bananas at a store
- Bananas on shelves
- Bunches of bananas
What do you see?

- Bananas
- Dole Bananas
- Bananas at a store
- Bananas on shelves
- Bunches of bananas
- Bananas with stickers on them
What do you see?

- Bananas
- Dole Bananas
- Bananas at a store
- Bananas on shelves
- Bunches of bananas
- Bananas with stickers on them
- Bunches of bananas with stickers on them on shelves in a store
What do you see?

- Bananas
- Dole Bananas
- Bananas at a store
- Bananas on shelves
- Bunches of bananas
- Bananas with stickers on them
- Bunches of bananas with stickers on them on shelves in a store

...We don’t tend to say

Yellow Bananas
What do you see?

Green Bananas

Unripe Bananas
What do you see?

Ripe Bananas

Bananas with spots
Ripe Bananas
Bananas with spots
Bananas good for banana bread
What do you see?

Yellow Bananas

Yellow is prototypical for bananas
Prototype Theory

One purpose of categorization is to reduce the infinite differences among stimuli to behaviourally and cognitively usable proportions.

There may be some central, prototypical notions of items that arise from stored typical properties for an object category (Rosch, 1975).

May also store exemplars (Wu & Barsalou, 2009).

Fruit

Bananas “Basic Level”

Unripe Bananas, Cavendish Bananas
A man and his son are in a terrible accident and are rushed to the hospital in critical care.

The doctor looks at the boy and exclaims "I can't operate on this boy, he's my son!"

How could this be?
A man and his son are in a terrible accident and are rushed to the hospital in critical care.

The doctor looks at the boy and exclaims "I can't operate on this boy, he's my son!"

How could this be?
A man and his son are in a terrible accident and are rushed to the hospital in critical care.

The doctor looks at the boy and exclaims "I can't operate on this boy, he's my son!"

How could this be?
“Doctor”  “Female doctor”
The majority of test subjects overlooked the possibility that the doctor is a she - including men, women, and self-described feminists.

Wapman & Belle, Boston University
World learning from text

Gordon and Van Durme, 2013

<table>
<thead>
<tr>
<th>Word</th>
<th>Frequency in corpus</th>
</tr>
</thead>
<tbody>
<tr>
<td>“spoke”</td>
<td>11,577,917</td>
</tr>
<tr>
<td>“laughed”</td>
<td>3,904,519</td>
</tr>
<tr>
<td>“murdered”</td>
<td>2,834,529</td>
</tr>
<tr>
<td>“inhaled”</td>
<td>984,613</td>
</tr>
<tr>
<td>“breathed”</td>
<td>725,034</td>
</tr>
<tr>
<td>“hugged”</td>
<td>610,040</td>
</tr>
<tr>
<td>“blinked”</td>
<td>390,692</td>
</tr>
<tr>
<td>“exhale”</td>
<td>168,985</td>
</tr>
</tbody>
</table>
### Word Learning from Text

Gordon and Van Durme, 2013

<table>
<thead>
<tr>
<th>Word</th>
<th>Frequency in corpus</th>
</tr>
</thead>
<tbody>
<tr>
<td>“spoke”</td>
<td>11,577,917</td>
</tr>
<tr>
<td>“laughed”</td>
<td>3,904,519</td>
</tr>
<tr>
<td>“murdered”</td>
<td>2,834,529</td>
</tr>
<tr>
<td>“inhaled”</td>
<td>984,613</td>
</tr>
<tr>
<td>“breathed”</td>
<td>725,034</td>
</tr>
<tr>
<td>“hugged”</td>
<td>610,040</td>
</tr>
<tr>
<td>“blinked”</td>
<td>390,692</td>
</tr>
<tr>
<td>“exhale”</td>
<td>168,985</td>
</tr>
</tbody>
</table>
Human Reporting Bias

The frequency with which people write about actions, outcomes, or properties is not a reflection of real-world frequencies or the degree to which a property is characteristic of a class of individuals.
Human Reporting Bias

The **frequency** with which **people write** about actions, outcomes, or properties is **not a reflection of real-world frequencies** or the degree to which a property is characteristic of a class of individuals.

Also called “Black Sheep” problem, “Giraffes problem” in vision/language.

Photographer Bias: Natural tendency of photographers to place object of interest in the center.
Training data are collected and annotated
Training data are collected and annotated ➔ Model is trained
Training data are collected and annotated

Model is trained

Media are filtered, ranked, aggregated, or generated
Training data are collected and annotated

Model is trained

Media are filtered, ranked, aggregated, or generated

People see output
### Human Biases in Data

<table>
<thead>
<tr>
<th>Reporting bias</th>
<th>Stereotypical bias</th>
<th>Group attribution error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Selection bias</td>
<td>Historical unfairness</td>
<td>Halo effect</td>
</tr>
<tr>
<td>Overgeneralization</td>
<td>Implicit associations</td>
<td></td>
</tr>
<tr>
<td>Out-group</td>
<td>Implicit stereotypes</td>
<td></td>
</tr>
<tr>
<td>homogeneity bias</td>
<td>Prejudice</td>
<td></td>
</tr>
</tbody>
</table>

Training data are collected and annotated
Human Biases in Data

- Reporting bias
- Selection bias
- Overgeneralization
- Out-group homogeneity bias
- Stereotypical bias
- Historical unfairness
- Implicit associations
- Implicit stereotypes
- Prejudice
- Group attribution error
- Halo effect

Training data are collected and annotated

Human Biases in Collection and Annotation

- Sampling error
- Non-sampling error
- Insensitivity to sample size
- Correspondence bias
- In-group bias
- Bias blind spot
- Confirmation bias
- Subjective validation
- Experimenter’s bias
- Choice-supportive bias
- Neglect of probability
- Anecdotal fallacy
- Illusion of validity
- Automation bias
**Reporting bias:** What people share is not a reflection of real-world frequencies

**Selection Bias:** Selection does not reflect a random sample

**Overgeneralization:** Coming to conclusion based on information that is too general and/or not specific enough

**Out-group homogeneity bias:** People tend to see outgroup members as more alike than ingroup members when comparing attitudes, values, personality traits, and other characteristics

**Confirmation bias:** The tendency to search for, interpret, favor, and recall information in a way that confirms one's preexisting beliefs or hypotheses

**Automation bias:** Propensity for humans to favor suggestions from automated decision-making systems and to ignore contradictory information made without automation, even if it is correct (Cummings, 2004)
Training data are collected and annotated 

Model is trained 

Media are filtered, ranked, aggregated, or generated 

People see output
Human Bias

1. Training data are collected and annotated
2. Model is trained
3. Media are filtered, ranked, aggregated, or generated
4. People see output
Human Bias

Training data are collected and annotated

Model is trained

Media are filtered, ranked, aggregated, or generated

People see output

Human Bias
Human Bias

Training data are collected and annotated

Model is trained

Media are filtered, ranked, aggregated, or generated

People see output

Biased data created from process becomes new training data
Bias Network Effect

Bias “Laundering”

Biased data created from process becomes new training data
Human data perpetuates human biases.

As ML learns from human data, the result is a bias network effect.
“Although neural networks might be said to write their own programs, they do so towards goals set by humans, using data collected for human purposes. If the data is skewed, even by accident, the computers will amplify injustice.”

— The Guardian
“Although neural networks might be said to write their own programs, they do so towards goals set by humans, using data collected for human purposes. If the data is skewed, even by accident, the computers will amplify injustice.”

— The Guardian
Information technology can amplify misinformation

Trending news, Ranked results, Autocomplete

Legal software can propagate discrimination

Predictive policing, Risk assessment, Assessing criminality

SOURCES
The Guardian
International Business Times

SOURCES
Physiognomy’s New Clothes
MIT Technology Review
ProPublica
The Intercept
Predictive Policing

Identifies fine-grained areas of potential criminal activity

Used to help law enforcement decide where to deploy

Focus on gun, domestic violence

Mixed results

Credit: Brett Lider, CC BY-SA
Predictive Sentencing

Northpointe: Risk in criminal sentencing (ProPublica, 2016)

The likelihood of each committing a future crime is predicted.

Borden — who is black — was rated a high risk. Prater — a more seasoned criminal, who is white — was rated a low risk.

2 years later, Borden has not been charged with any new crimes. Prater serving 8-year prison term for breaking into a warehouse and stealing thousands of dollars’ worth of electronics.
Predictive Criminality

An Israeli startup, Faception, who has not published any details about their methods, sources of training data, or quantitative results:

“Faception is first-to-technology and first-to-market with proprietary computer vision and machine learning technology for profiling people and revealing their personality based only on their facial image.”

Offering specialized engines for recognizing “High IQ”, “White-Collar Offender”, “Pedophile”, and “Terrorist” from a face image.

Main clients are in homeland security and public safety.
Predictive Criminality

“Automated Inference on Criminality using Face Images” Wu and Zhang, 2016. arXiv

1,856 closely cropped images of faces -- “wanted suspect” pictures from specific areas, rest from crawling web.

“[…] angle $\theta$ from nose tip to two mouth corners is on average 19.6% smaller for criminals than for non-criminals ...”
Predictive Criminality - The Media Blitz

arXiv Paper Spotlight: Automated Inference on Criminality Using Face ...  
www.kdnuggets.com/.../arxiv-spotlight-automated-inference-criminality-face-images.... ▼
A recent paper by Xiaolin Wu (McMaster University, Shanghai Jiao Tong University) and Xi Zhang (Shanghai Jiao Tong University), titled "Automated Inference ...

Automated Inference on Criminality Using Face Images | Hacker News  
https://news.ycombinator.com/item?id=12983827 ▼
Nov 18, 2016 - The automated inference on criminality eliminates the variable of meta-accuracy (the competence of the human judge/examiner) all together.

A New Program Judges If You’re a Criminal From Your Facial Features ...  
https://motherboard.vice.com/.../new-program-decides-criminality-from-facial-feature... ▼
Nov 18, 2016 - In their paper 'Automated Inference on Criminality using Face Images', published on the arXiv pre-print server, Xiaolin Wu and Xi Zhang from ...

Can face classifiers make a reliable inference on criminality?  
https://techxplore.com/.../computer-sciences ▼
Nov 23, 2016 - Their paper is titled "Automated Inference on Criminality using Face Images ... face classifiers are able to make reliable inference on criminality.

Troubling Study Says Artificial Intelligence Can Predict Who Will Be ...  
https://theintercept.com/.../troubling-study-says-artificial-intelligence-can-predict-who... ▼
Nov 18, 2016 - Not so in the modern age of Artificial Intelligence, apparently: In a paper titled "Automated Inference on Criminality using Face Images," two ...

Automated Inference on Criminality using Face Images (via arXiv ...  
https://computationallegalstudies.com/.../automated-inference-on-criminality-using-fa... ▼
Dec 6, 2016 - Next Next post: A General Approach for Predicting the Behavior of the Supreme Court of the United States (Paper Version 2.01) (Katz, ...
But it’s up to us to influence how AI evolves.
Today
Find local optimum given task, data, etc
Get paper published
Get paper award
Global optimum for humans, their environment, and Artificial Intelligence
Find local optimum given task, data, etc

Get paper published

Get paper award

Global optimum for humans, their environment, and Artificial Intelligence

How can the work I’m interested in now be best focused to help others?
Begin tracing out paths for the evolution of ethical AI

Short-term

- Find local optimum given task, data, etc
- Get paper published
- Get paper award

Longer-term

Global optimum for humans, their environment, and Artificial Intelligence

How can the work I’m interested in now be best focused to help others?
Rest of Talk

1. Modeling world knowledge (and biases!) with latent variables
2. Focus on best performance across groups of people
   ● Working with experts and those affected to better understand what’s needed
   ● Contextualizing work for public
Modeling World Knowledge with Latent Variables: A case study in vision-to-language
Modeling World Knowledge with Latent Variables: A case study in vision-to-language

Those affected: People who are blind
Modeling World Knowledge with Latent Variables: A case study in vision-to-language

Those affected: People who are blind

Seeing AI: Microsoft research project that brings together the power of the cloud and AI to deliver an intelligent app, designed to help you navigate your day.
Human Reporting Bias

The *frequency* with which *people write* about actions, outcomes, or properties is *not a reflection of real-world frequencies* or the degree to which a property is characteristic of a class of individuals.

Also called “Black Sheep” problem, “Giraffes problem” in vision/language.

Photographer Bias: Natural tendency of photographers to place object of interest in the center.
Bias: A systematic deviation from full ground truth.
Bias: A systematic deviation from full ground truth.
Can be used as a signal.
Data data everywhere …

**Facebook**
300 Million images uploaded everyday

**flickr**

**YouTube**
100 hours of video every minute

OMG Frodo is sitting eating pizza and donuts.
dog, chair, pizza, donut
dog, chair, pizza, donut

#dog #hungry
Data data everywhere …

But not many labels to train

Exhaustively annotated data is expensive

- dog, chair, pizza, donut
- dog, chihuahua, brown, chair, table, wall, space heater, pizza, greasy, donut 1, donut 2, pizza slice 1, pizza slice 2...
“In the Wild” Labeled Images: Why?

● “Freely” available: Image tags, descriptions on social media
● Fast way to gather data beyond typical categories
● Annotations for an image are on a “per-image” basis [mostly]

#dog #food #hungry
A hungry dog looks at the food on the table
Challenges with WILD Labeled Images

Teacher in a classroom talking to students
Challenges with WILD Labeled Images

Teacher in a classroom talking to students
Challenges with WILD Labeled Images

Teacher in a classroom talking to students
Challenges with WILD Labeled Images

Humans subjectively decide what to report and what not to report in an image.

Teacher in a classroom talking to students
Challenges with WILD Labeled Images

Humans subjectively decide *what to report* and *what not to report* in an image.

*Teacher in a classroom talking to students*
Challenges with WILD Labeled Images

Teacher
Male
White
Early 30s
Wearing Pants

Humans subjectively decide what to report and what not to report in an image

Teacher in a classroom talking to students
Using Labeled Images In the Wild

How do we train visually correct classifiers from wild data?

Input Wild Data

#dog #food #hungry

Expected Output

dog, brown, chihuahua, chair, pizza, donut
Problem setup

- **Input:** Image, human-biased labels
- **Goal:** Learn visually correct classifiers
- **Challenge:** Do not have access to ground truth; have access to what humans have reported

**Input**

- Teacher
- Man
- Standing
- Early 30s
- White
- Classroom
- High school
- Students
- Projector
- Door
- Wall
- Podium
- Learning

**Goal**

- Teacher in a classroom talking to students
Human-Biased Labels

- Highly dependent on the input image
  - Example: bicycle

Human Biased Labels

- Highly dependent on the input image
- The outcome of a complex systematic process [human judgment]
  - Humans are fairly systematic in such labeling

Human-Biased Labels

- Highly dependent on the input image
- The outcome of a complex systematic process [human judgment]
  - Humans are fairly systematic in such labeling
  - Humans refer to object properties when it helps distinguishability, conversation etc.

[Gregory 1966], [Rosch 1973], [Sedivy et al., 2003], [Koolen et al., 2011]

Related Work: Modeling label noise

- Assumes the noise is conditionally independent of the input image
  [Mnih and Hinton 2012], [Natarajan et al., 2013], [Reed et al., 2014], [Sukhbaatar et al., 2015], [Izadinia et al., 2015]

Related Work: Modeling label noise

- Assumes the noise is conditionally independent of the input image
  [Mnih and Hinton 2012], [Natarajan et al., 2013], [Reed et al., 2014], [Sukhbaatar et al., 2015], [Izadinia et al., 2015]

- Assumes that estimating noise requires access to exhaustively labeled data
  [Xiao et al., 2015]

Notation

\( w \in \{\text{banana, yellow}\} \)

<table>
<thead>
<tr>
<th>Input Image</th>
<th>Output</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image" alt="Input Image" /></td>
<td><img src="image" alt="Output" /></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Banana</th>
<th>Yellow</th>
<th>Label</th>
</tr>
</thead>
<tbody>
<tr>
<td>✔️</td>
<td>✔️</td>
<td>( z^w )</td>
</tr>
</tbody>
</table>

- Visually correct ground truth (Unknown)
- Available Ground truth (human-biased)
Simple Image Classification

\( w \in \{\text{banana, yellow}\} \)

Input Image → CNN → Classifier → Output

For each \( w \)

Ground Truth

\( y^w \)

Simple Image Classification

$w \in \{\text{banana, yellow}\}$

For each $w$

Human-biased label $y^w \in \{0, 1\}$ (Gold Standard)

Simple Image Classification

$w \in \{\text{banana, yellow}\}$

Human-biased label $y_w \in \{0, 1\}$ (Gold Standard)

$\text{Prediction } h_w(y^w | I)$

Factoring in label bias: Idea

- A human-biased prediction $h$ can be factored into two terms

Factoring in label bias: Idea

- A human-biased prediction $h$ can be factored into two terms
  - Visual presence $v$ – *Is the object visually present?*

  $w \in \{\text{banana, yellow}\}$

  visually correct ground truth (unknown): $z$

*Misra, et al. (2016). Seeing through the Human Reporting Bias: Visual Classifiers from Noisy Human-Centric Labels. CVPR.*
A human-biased prediction $h$ can be factored into two terms:

- Visual presence $v$ – Is the object visually present?
- Relevance $r$ – Is the object relevant for a human?

$w \in \{\text{banana, yellow}\}$

Use $z$ to predict human-biased label $y$

<table>
<thead>
<tr>
<th>Label</th>
<th>Prediction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Visually correct ground truth (Unknown)</td>
<td>$z$</td>
</tr>
<tr>
<td>Available ground truth (human-centric)</td>
<td>$y$</td>
</tr>
</tbody>
</table>

Factoring in label bias: Idea

- A human-biased prediction $h$ can be factored into two terms
  - Visual presence $v$ – *Is the object visually present?*
  - Relevance $r$ – *Is the object relevant for a human?*

$$h = f(r, v)$$

<table>
<thead>
<tr>
<th>Label</th>
<th>Prediction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Visually correct ground truth <em>(Unknown)</em></td>
<td>$z$</td>
</tr>
<tr>
<td>Available ground truth <em>(human-centric)</em></td>
<td>$y$</td>
</tr>
<tr>
<td></td>
<td>$h$</td>
</tr>
</tbody>
</table>

*Misra, et al. (2016). Seeing through the Human Reporting Bias: Visual Classifiers from Noisy Human-Centric Labels. CVPR.*
Factoring in label bias: Idea

- A human-biased prediction $h$ can be factored into two terms
  - Visual presence $v$: Is the object *visually present*?
  - Relevance $r$: Is the object *relevant* for a human?

Given visual presence, is concept relevant? Is concept present?

$$h(y|I) = \sum_{j \in \{0,1\}} r(y|z = j, I)v(z = j|I)$$

<table>
<thead>
<tr>
<th>Label</th>
<th>Prediction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Visually correct ground truth (Unknown)</td>
<td>$z$</td>
</tr>
<tr>
<td>Available ground truth (human-centric)</td>
<td>$y$</td>
</tr>
</tbody>
</table>

Factoring in label bias: Idea

- A human-biased prediction $h$ can be factored into two terms
  - Visual presence $v$ – *Is the object visually present?*
  - Relevance $r$ – *Is the object relevant for a human?*

\[
h(y|I) = \sum_{j \in \{0,1\}} r(y|z = j, I)v(z = j|I)
\]

- Allows classifier to not get penalized for correct predictions

<table>
<thead>
<tr>
<th>Label</th>
<th>Prediction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Visually correct ground truth (<strong>Unknown</strong>)</td>
<td>$z$</td>
</tr>
<tr>
<td>Available ground truth (<strong>human-centric</strong>)</td>
<td>$y$</td>
</tr>
</tbody>
</table>

End-to-End Approach

Results

- Evaluate both $v$ and $h$ predictions
- Microsoft COCO dataset
  - Human-biased labels $y$ from Captions [1000 categories]
  - Visually correct labels $z$ from Detection Bounding boxes [73 categories]
  - #images: 80k train, 20k test

- YFCC100M
  - Yahoo Flickr images with tags [1000 categories]
  - Random subset #images: 75k train, 15k test

<table>
<thead>
<tr>
<th>Label</th>
<th>Prediction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Visually correct ground truth (Unknown)</td>
<td>$z$</td>
</tr>
<tr>
<td>Available ground truth (human-centric)</td>
<td>$y$</td>
</tr>
</tbody>
</table>

A hungry dog looks at the food on the table.

Evaluating using $y$ (human-biased) and $z$ (annotated)

### Evaluation using observed caption concepts

<table>
<thead>
<tr>
<th>Method</th>
<th>Prob</th>
<th>Mean Average Precision</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>MILVC</td>
<td>-</td>
<td>41.6</td>
<td>20.7</td>
<td>23.9</td>
<td>33.4</td>
<td>20.4</td>
<td>22.5</td>
<td>16.3</td>
</tr>
<tr>
<td>MILVC + Multiple fc8</td>
<td>-</td>
<td>41.1</td>
<td>20.9</td>
<td>23.7</td>
<td>33.6</td>
<td>21.1</td>
<td>22.8</td>
<td>16.8</td>
</tr>
<tr>
<td>MILVC + Latent</td>
<td>$v$</td>
<td>42.9</td>
<td>21.7</td>
<td>24.9</td>
<td>33.1</td>
<td>19.6</td>
<td>23.0</td>
<td>16.2</td>
</tr>
<tr>
<td>MILVC + Latent</td>
<td>$h$</td>
<td>44.3</td>
<td>22.3</td>
<td>25.8</td>
<td>34.4</td>
<td>21.8</td>
<td>23.6</td>
<td>17.3</td>
</tr>
<tr>
<td>Classif.</td>
<td>-</td>
<td>34.9</td>
<td>18.1</td>
<td>20.5</td>
<td>32.8</td>
<td>19.2</td>
<td>21.8</td>
<td>16.3</td>
</tr>
<tr>
<td>Classif. + Multiple fc8</td>
<td>-</td>
<td>34.2</td>
<td>17.7</td>
<td>19.9</td>
<td>32.6</td>
<td>19.0</td>
<td>21.5</td>
<td>15.9</td>
</tr>
<tr>
<td>Classif. + Latent</td>
<td>$v$</td>
<td>37.7</td>
<td>19.6</td>
<td>22.0</td>
<td>32.6</td>
<td>20.2</td>
<td>22.0</td>
<td>16.3</td>
</tr>
<tr>
<td>Classif. + Latent</td>
<td>$h$</td>
<td>38.7</td>
<td>20.1</td>
<td>22.6</td>
<td>33.8</td>
<td>21.2</td>
<td>23.0</td>
<td>17.5</td>
</tr>
</tbody>
</table>

YFCC100M: Flickr images with tags (90k images, 1k tags)

<table>
<thead>
<tr>
<th>Method</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>MILVC</td>
<td>-</td>
<td>5.7</td>
<td>9.2</td>
<td>5.2</td>
<td>-</td>
<td>3.8</td>
<td>8.8</td>
<td>6.1</td>
</tr>
<tr>
<td>MILVC + Multiple fc8</td>
<td>-</td>
<td>4.6</td>
<td>6.2</td>
<td>3.8</td>
<td>-</td>
<td>2.7</td>
<td>7.3</td>
<td>3.1</td>
</tr>
<tr>
<td>MILVC + Latent</td>
<td>$v$</td>
<td>9.8</td>
<td>15.1</td>
<td>8.9</td>
<td>-</td>
<td>8.3</td>
<td>12.4</td>
<td>12.4</td>
</tr>
<tr>
<td>MILVC + Latent</td>
<td>$h$</td>
<td>11.2</td>
<td>15.4</td>
<td>9.9</td>
<td>-</td>
<td>8.2</td>
<td>16.3</td>
<td>12.5</td>
</tr>
</tbody>
</table>

All methods use VGG16. Trained using binary cross-entropy loss.

MILVC: Fang et al., 2015; Classif: Simple classification baseline; Multiple-fc8: Same # parameters as our model.

### Evaluation using annotated concepts

COCO Dataset. 73 annotated concepts from Bounding Boxes.

<table>
<thead>
<tr>
<th></th>
<th>MILVC</th>
<th>$v$</th>
<th>$h$</th>
<th>Using ground truth</th>
</tr>
</thead>
<tbody>
<tr>
<td>mAP</td>
<td>63.7</td>
<td>66.8</td>
<td>66.5</td>
<td>76.3</td>
</tr>
</tbody>
</table>

Qualitative Results

Mentioned by humans ($h$)

Threshold at $v \geq 0.85$

Shows how much $v$ and $h$ are decoupled

Qualitative Results

Mentioned by humans (h)

fence, oven, brick, pink, trees, hat, tie, green, red, orange

Corrected Error Modes

Corrected False Positives | Corrected False Negatives
---|---
**desert**
fridge
**beach**
sheep

Corrected False Positives | Corrected False Negatives
---|---
net
night
waves
drinking

When to mention it?
When would you mention something not worth mentioning?

Improvement in Downstream Applications: Image Captioning

<table>
<thead>
<tr>
<th></th>
<th>Prob</th>
<th>BLEU-4</th>
<th>ROUGE</th>
<th>CIDEr</th>
</tr>
</thead>
<tbody>
<tr>
<td>MILVC</td>
<td>-</td>
<td>27.7</td>
<td>51.8</td>
<td>89.7</td>
</tr>
<tr>
<td>MILVC + Latent</td>
<td>h</td>
<td>29.2</td>
<td>52.4</td>
<td>92.8</td>
</tr>
</tbody>
</table>

Rest of Talk

1. Modeling world knowledge (and biases!) with latent variables
2. Focus on best performance across **groups** of people
   - Working with experts and those affected to better understand what’s needed
   - Contextualizing work for public
Fair is Fair: For all groups of people

Those affected: People with neuroatypicality, clinicians
Motivation from “The Karate Kid”

Single-task Learners (STL)

Multitask Learner (MTL)
Single-Task: Logistic Regression

Output Prediction (Task):
True or False (for example)

Input Features
Single-Task: Deep Learning

Output Prediction (Task): True or False (for example)

Fancier!!

Input Features
Multiple Tasks with Basic Logistic Regression
Multiple Tasks + Deep Learning: Multi-task Learning

Multiple Tasks + Deep Learning: Multi-task Learning Example

Multiple Tasks + Deep Learning: Multi-task Learning Example

Y₁, Y₂, Y₉, Y₁₀
Depression Anxiety, PTSD, Gender

W₁, W₂, W₉, W₁₀

Task  | N
---|---
Gender    | 1101
Neurotypicality | 4791
Anxiety    | 2407
Depression  | 1400
Suicide attempt | 1208
Eating disorder | 749
Schizophrenia | 349
Panic disorder | 263
PTSD       | 248
Bipolar disorder | 191
All        | 9611

<5% positive examples

Improved Performance across Subgroups

True Positive Rate @ False Positive Rate = 0.1

~120 at-risk individuals

Improved Performance across Subgroups

True Positive Rate @ False Positive Rate = 0.1

Reading for the masses....

Contextualizing and considering ethical dimensions

2 Disclaimer

As with any author-attribute detection, there is the danger of abusing the model to single out people (overgeneralization, see Hovy and Spruit (2016)). We are aware of this danger, and sought to minimize the risk. For this reason, we don’t provide a selection of features or representative examples. The experiments in this paper were performed with a clinical application in mind, and use carefully matched (but anonymized) data, so the distribution is not representative of the population as a whole. The results of this paper should therefore not be interpreted as a means to assess mental health conditions in social media in general, but as a test for the applicability of MTL in a well-defined clinical setting.

Me, Me, Me: People Who Overuse The First-Person Singular Are More Depressed

A new study links first-person singular pronouns to relationship problems and higher rates of depression.

By Rose Pastore  May 3, 2013

Contextualizing and considering ethical dimensions
PHASE 01

Consider the problem

How will the model be affected when a blind spot is found in the training data?
Ask experts for answers

What do the experts say is most useful or necessary? What do they think of the problem you’re working on?
Engage with Policy

Serve as consultant for your senator, congressperson; be part of legal and policy meetings relevant to your work.
PHASE 04

Design the human input

Is this an unambiguous task? How will you verify that crowdworkers are performing tasks "correctly"? How will you best leverage human biases?
Engage diverse crowdworkers well

Speed and Agreement are bedrock measures of “click-workers”, and their goal is different from yours. Without consideration of speed/pay tradeoffs, and crowdworker diversity, click-work turns into exponential groupthink that bakes cultural biases directly into training data.
Train the models to account for bias

What does an outlier use case look like, and how does the model handle it? What implicit assumptions might be helpful to model?
PHASE 07

Interpret outcomes

Is the ML overgeneralizing? If a human were to perform this task, what would appropriate social behavior look like? What interpersonal cues might be relevant that are missing from the input or interface? E.g. body language, tone of voice.
PHASE 08

Publish with context

Are you sharing examples? Why or why not?
How should this technology be directly used?
What are some easy misconceptions that we can avoid?
Publish for reproducibility

Scientific claims should be possible to reproduce given enough information, and access to data (when applicable).
Moving from majority representation...
Moving from majority representation...

...to diverse representation
Moving from majority representation...

...to diverse representation

...for ethical AI
Thanks!

margarmitchell@gmail.com
References

https://www.quirks.com/articles/9-types-of-research-bias-and-how-to-avoid-them

KDD Tutorial: http://francescobonchi.com/algorithmic_bias_tutorial.html

https://dub.washington.edu/djangosite/media/papers/unequalrepresentation.pdf


Fang et al., “From Captions to Visual Concepts and Back”. CVPR 2015

References


Aguera y Arcas, B, Mitchell, M & Todorov, A. “Physiognomy’s New Clothes”. Medium, 2017