Bias, domain shifts, attacks

Prof. Adriana Kovashka
University of Pittsburgh
April 15, 2024
Plan for this lecture

• Domain shifts due to visual style/appearance
• Domain shifts due to geography
• Models inheriting social biases
• Deep fakes and adversarial perturbations
**Domain Shifts**

Colors = domains, shapes = classes

**Fig. 2.** The key idea of our approach to domain adaptation is to learn a transformation that compensates for the domain-induced changes. By leveraging (dis)similarity constraints (b) we aim to reunite samples from two different domains (blue and green) in a common invariant space (c) in order to learn and classify new samples more effectively across domains. The transformation can also be applied to new categories (lightly-shaded stars). This figure is best viewed in color.
DomainNet Dataset

Coping with Domain Shifts

Figure 2. The framework of **Moment Matching for Multi-source Domain Adaptation** (M$^3$SDA). Our model consists of three components: i) feature extractor, ii) moment matching component, and iii) classifiers. Our model takes multi-source annotated training data as input and transfers the learned knowledge to classify the unlabeled target samples. Without loss of generality, we show the $i$-th domain and $j$-th domain as an example. The feature extractor maps the source domains into a common feature space. The moment matching component attempts to match the $i$-th and $j$-th domains with the target domain, as well as matching the $i$-th domain with the $j$-th domain. The final predictions of target samples are based on the weighted outputs of the $i$-th and $j$-th classifiers. (Best viewed in color!)
Domain Adversarial Networks

Figure 1: The proposed architecture includes a deep feature extractor (green) and a deep label predictor (blue), which together form a standard feed-forward architecture. Unsupervised domain adaptation is achieved by adding a domain classifier (red) connected to the feature extractor via a gradient reversal layer that multiplies the gradient by a certain negative constant during the backpropagation-based training. Otherwise, the training proceeds standardly and minimizes the label prediction loss (for source examples) and the domain classification loss (for all samples). Gradient reversal ensures that the feature distributions over the two domains are made similar (as indistinguishable as possible for the domain classifier), thus resulting in the domain-invariant features.

“While modern computer vision models yield human-level accuracies when trained and tested on the images from the same geographical domain, the accuracy drops significantly when presented with images from different geographies. Here, images belonging to mailbox and running track are misclassified due to design and context shifts between the domains induced by disparate geographies.”

Geographic Domain Shifts

Context Shift
Task-irrelevant information (e.g., background or surroundings)

<table>
<thead>
<tr>
<th>Source</th>
<th>Dollar Street-DA</th>
<th>GeoYFCC-DA</th>
</tr>
</thead>
<tbody>
<tr>
<td>toothbrush</td>
<td><img src="source/toothbrush.png" alt="toothbrush" /></td>
<td><img src="target/toothbrush.png" alt="toothbrush" /></td>
</tr>
<tr>
<td>sofa</td>
<td><img src="source/sofa.png" alt="sofa" /></td>
<td><img src="target/sofa.png" alt="sofa" /></td>
</tr>
<tr>
<td>basketball</td>
<td><img src="source/basketball.png" alt="basketball" /></td>
<td><img src="target/basketball.png" alt="basketball" /></td>
</tr>
<tr>
<td>kitchen</td>
<td><img src="source/kitchen.png" alt="kitchen" /></td>
<td><img src="target/kitchen.png" alt="kitchen" /></td>
</tr>
</tbody>
</table>


Slide: Marcelo d’Almeida
Geographic Domain Shifts

Subpopulation Shift
Change within category (e.g., “cleaning equipment” can be brooms, mops, vacuum cleaners, etc.)

Slide: Marcelo d’Almeida
Geographic Domain Shifts

Dollar Street-DA


Slide: Marcelo d’Almeida
Geographic Domain Shifts

GeoYFCC-DA

Temple
- Argentina
- Cambodia

Stove
- Costa Rica
- Vietnam

Lamp Post
- India
- France

62 countries  68 categories


Slide: Marcelo d’Almeida
# Geographic Domain Shifts

## Results

<table>
<thead>
<tr>
<th>Method</th>
<th>Dollar Street-DA</th>
<th>GeoYFCC-DA</th>
</tr>
</thead>
<tbody>
<tr>
<td>source</td>
<td>54.66±0.62</td>
<td>42.88</td>
</tr>
<tr>
<td>target oracle*</td>
<td>67.73±0.30</td>
<td>56.78</td>
</tr>
<tr>
<td>MMD [10]</td>
<td>55.77±0.75</td>
<td>43.53</td>
</tr>
<tr>
<td>DANN [4]</td>
<td>54.80±0.38</td>
<td>42.64</td>
</tr>
<tr>
<td>SENTRY [5]</td>
<td>55.73±0.34</td>
<td>42.58</td>
</tr>
<tr>
<td>SST</td>
<td>58.71±0.53</td>
<td>45.22</td>
</tr>
</tbody>
</table>

*denotes that the target oracle was trained on target data non-overlapping with the test set (80%) whereas DA methods were adapted without labels on the entire target dataset.

**Significant performance drops from geographical shifts**

**Limited improvements from existing DA methods**
Incorporating Geo-Diverse Knowledge into Prompting for Increased Geographical Robustness in Object Recognition

Kyle Buettner\textsuperscript{1}, Sina Malakouti\textsuperscript{2}, Xiang Lorraine Li\textsuperscript{1,2}, Adriana Kovashka\textsuperscript{1,2}
\textsuperscript{1}Intelligent Systems Program, \textsuperscript{2}Department of Computer Science, University of Pittsburgh, PA, USA
\{buettnerk, sem238\}@pitt.edu, \{xianglli, kovashka\}@cs.pitt.edu

Abstract

Existing object recognition models have been shown to lack robustness in diverse geographical scenarios due to significant domain shifts in design and context. Class representations need to be adapted to more accurately reflect an object concept under these shifts. In the absence of training data from target geographies, we hypothesize that geography-specific descriptive knowledge of object categories can be leveraged to enhance robustness. For this purpose, we explore the feasibility of probing a large-language model for geography-specific object knowledge, and we investigate integrating knowledge in zero-shot and learnable soft prompting with the CLIP vision-language model. In particular, we propose a geography knowledge regularization method to ensure that soft prompts trained on a source set of geographies generalize to an unseen target set of geo-

Figure 1. Descriptive knowledge can bridge concept shifts across geographies. Observe the wide range of object designs and contexts in the DollarStreet [11] category tools around the world. Our work’s premise is that textual representations for classes in vision-language models can be tailored to better suit diverse object representations across geographies. Map made with [16].
Geographic Domain Shifts

(1) Acquire Target Knowledge
Internal CLIP Knowledge + External LLM Knowledge

A photo of a stove/hob in Mexico...
A photo of a stove/hob in Togo...
A photo of a stove/hob in Burkina Faso...

What are useful features for distinguishing a stove in a photo that I took in <country>?

LLM (davinci-003)

A stove/hob in Burkina Faso may be/have...
- Four or more burners
- Metal or ceramic material
- White, black, or stainless-steel color
- Flat or slightly angled surface
- Knobs for setting temperature
- Over or broiler below burners
- Gas or electric powered
- Vent above hood

A stove/hob in Togo may be/have...
- Set on the ground or on a table
- Single or multiple burners
- Fueled by kerosene, gas, wood
- Large pot or skillet
- Covered with a hood
- Large handle for easy access

(2) Optimize Soft Prompts on Source Data While Regularizing Towards Target Knowledge

Czechia

CLIP Image Encoder

Image Features

Trainable Soft Prompts

CLIP Text Encoder

Target Knowledge Prompts

Source: Europe

Target: Africa, Asia, Americas

(3) Recognize Objects in Unseen Countries

Mexico

Togo

Burkina Faso

Figure 2. Geography knowledge regularization. To ensure robustness in soft prompt learning, we (1) incorporate complementary knowledge internal to CLIP and externally obtained from an LLM. (2) This descriptive knowledge regularizes class representations when training on a specific source geography (e.g. Europe), thus (3) increasing robustness when generalizing to unseen geographies (e.g. Togo).
## Geographic Domain Shifts

<table>
<thead>
<tr>
<th>Encoder</th>
<th>Prompting Method</th>
<th>Top-1 Accuracy</th>
<th>Top-3 Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Europe</td>
<td>Africa</td>
</tr>
<tr>
<td>ViT-B/32</td>
<td>Zero-Shot CLIP [36]</td>
<td>64.3</td>
<td>46.9</td>
</tr>
<tr>
<td></td>
<td>GeneralLLM [30]</td>
<td>64.2</td>
<td>48.8</td>
</tr>
<tr>
<td></td>
<td>CountryInPrompt</td>
<td>63.9</td>
<td>49.6</td>
</tr>
<tr>
<td></td>
<td>CountryLLM</td>
<td>65.2</td>
<td>50.9</td>
</tr>
<tr>
<td></td>
<td>CountryInPrompt+LLM</td>
<td>65.5</td>
<td>51.2</td>
</tr>
<tr>
<td>ViT-B/16</td>
<td>Zero-Shot CLIP [36]</td>
<td>53.0</td>
<td>38.0</td>
</tr>
<tr>
<td></td>
<td>GeneralLLM [30]</td>
<td>55.5</td>
<td>40.9</td>
</tr>
<tr>
<td></td>
<td>CountryInPrompt</td>
<td>54.5</td>
<td>43.4</td>
</tr>
<tr>
<td></td>
<td>CountryLLM</td>
<td>56.2</td>
<td>41.1</td>
</tr>
<tr>
<td></td>
<td>CountryInPrompt+LLM</td>
<td>56.4</td>
<td>43.0</td>
</tr>
</tbody>
</table>

Table 1. Zero-shot CLIP with descriptive knowledge prompts, top-1/3 balanced accuracy (Acc) on DollarStreet. Our strategies to capture CLIP’s internal country knowledge (CountryInPrompt), external LLM country knowledge (CountryLLM), and their combination (CountryInPrompt+LLM), improve the zero-shot CLIP baseline (prompt “a photo of a”), especially on Africa (exemplified in light blue) and Asia; gains in green, drops in red. Our strategies also outperform the GeneralLLM [30] baseline.
Plan for this lecture

• Domain shifts due to visual style/appearance
• Domain shifts due to geography
• Models inheriting social biases
• Deep fakes and adversarial perturbations
Bias in Language

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

Figure 1: The most extreme occupations as projected on to the *she—he* gender direction on g2vNEWS. Occupations such as *businesswoman*, where gender is suggested by the orthography, were excluded.

**Gender stereotype *she-he* analogies.**

- sewing-carpentry
- nurse-surgeon
- blond-burly
- giggle-chuckle
- sassy-snappy
- volleyball-football

- register-nurse-physician
- interior designer-architect
- feminism-conservatism
- vocalist-guitarist
- diva-superstar
- cupcakes-pizzas

- housewife-shopkeeper
- softball-baseball
- cosmetics-pharmaceuticals
- petite-lanky
- charming-affable
- hairdresser-barber

**Gender appropriate *she-he* analogies.**

- queen-king
- waitress-waiter

- sister-brother
- ovarian cancer-prostate cancer

- mother-father
- convent-monastery

Figure 2: **Analogy examples.** Examples of automatically generated analogies for the pair *she-he* using the procedure described in text. For example, the first analogy is interpreted as *she:sewing :: he:carpentry* in the original w2vNEWS embedding. Each automatically generated analogy is evaluated by 10 crowd-workers as to whether or not it reflects gender stereotype. Top: illustrative gender stereotypic analogies automatically generated from w2vNEWS, as rated by at least 5 of the 10 crowd-workers. Bottom: illustrative generated gender-appropriate analogies.
Bias in Language

http://wordbias.umiacs.umd.edu/
Fig. 1: Examples where our proposed model (Equalizer) corrects bias in image captions. The overlaid heatmap indicates which image regions are most important for predicting the gender word. On the left, the baseline predicts gender incorrectly, presumably because it looks at the laptop (not the person). On the right, the baseline predicts the gender correctly but it does not look at the person when predicting gender and is thus not acceptable. In contrast, our model predicts the correct gender word and correctly considers the person when predicting gender.
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.
What do you see?

- Bananas
- Stickers
- 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
- 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?

“Female doctor”
Training data are collected and annotated

Model is trained

Media are filtered, ranked, aggregated, or generated

People see output
Biases in Data

Selection Bias: Selection does not reflect a random sample

Map of Amazon Mechanical Turk Workers

CREDIT
© 2013–2016 Michael Yoshitaka Erlewine and Hadas Kotek

Margaret Mitchell
Biases in Data

**Out-group homogeneity bias:** Tendency to see outgroup members as more alike than ingroup members
It's possible that you have an appropriate amount of data for every group you can think of but that some groups are represented less positively than others.

Margaret Mitchell
Biases in Data → Biased Labels

Annotations in your dataset will reflect the worldviews of your annotators.

Predicting Policing

- Algorithms identify potential crime hot-spots
- Based on where crime is previously reported, not where it is known to have occurred
- Predicts future events from past

CREDIT

Margaret Mitchell
Predicting Sentencing

- Prater (who is white) rated **low risk** after shoplifting, despite two armed robberies; one attempted armed robbery.
- Borden (who is black) rated **high risk** after she and a friend took (but returned before police arrived) a bike and scooter sitting outside.
- Two years later, Borden has not been charged with any new crimes. Prater serving 8-year prison term for grand theft.

CREDIT
Predicting Criminality

Israeli startup, Faception

“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.
Predicting Criminality

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

1,856 closely cropped images of faces; Includes “wanted suspect” ID pictures from specific regions.

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

See our longer piece on Medium, “Physiognomy’s New Clothes”
“Deepfakes”

DARPA

Expected Threats

Targeted Personal Attacks
Peele 2017

AI Multimedia Algorithms

Generated Events at Scale

AI Multimedia Algorithms

Ransomfake concept: Identity Attacks as a service (IAaaS)
Bricman 2019

AI Multimedia Algorithms

Forged Evidence

Identity Attacks

Examples of possible fakes:
- Substance abuse
- Foreign contacts
- Compromising events
- Social media postings
- Financial inconsistencies
- Forging identity

On a rainy spring day, a vast, violent group gathered in front of the US Capitol to protest recent cuts in Social Security.

Text

Video & Audio

Image

Believable fake events

Highly realistic video

Undermines key individuals and organizations

Matt Turek
Incredible Pace of Synthetic Media Generation

Legend:
- Single modality
- Multi-modality

- Interactive audio
- Attribute-guided face generation
- Unsupervised text generation
- Fake resumes
- Video dialog replacement
- Fake dating profiles
- Fake rental ads
- Scenes from sketches

ENTIRE GUEST SUITE
Luxury Condo 3 Bed + 3 Bath
Port Melbourne

- 8 guests
- 3 bedrooms
- 4 beds
- 2 baths

Bathroom (with seating for 2 more people), basin and eclectic French garden and kitchen. 24/7 carpeted charc. Laundrymemberly : More balcony – Garden – Metro, Liverpool Street (15 min walk) Walking distance to Wyckofferdon

Anne
Adversarial Attacks

https://bair.berkeley.edu/blog/2017/12/30/yolo-attack/
Adversarial Attacks

Adversarial Attacks

Tom Goldstein https://www.cs.umd.edu/~tomg/projects/invisible/
Adversarial Attacks

This object-recognition dataset stumped the world’s best computer vision models
Objects are posed in varied positions and shot at odd angles to spur new AI techniques.