Introduction to Images as Data in Social Science

Nora Webb Williams[1]

[1]Assistant Professor, International Affairs
University of Georgia

PaCCS/Pew Workshop on Image Analysis, August 27, 2019
Additional sponsorship by NSF SES Grant Number 1727459

Outline

- Why images as data in social science? Why **lots** of images as data?
- What are some automated options for image analysis?
- Options we *aren't* going to cover today.
- Where to go to learn more?

Why images as data?

- Studying images is not new
- Theorizing about images is not new
- But having lots of digitized images and lots of computer power *is* new

Powerful Social Media Images

Less Powerful Social Media Images

Innovations in Methods

- How do we **quickly process** massive numbers of digitized images?
- How do we **extract from images** the information we need to test our hypotheses or accurately describe situations?

Types of Research with Images as Data

Causal framework:

- Images as independent variable
  - Casas and Webb Williams (*PRQ* 2018): Which Black Lives Matter images mobilized more supporters online?
- Images as dependent variable
  - Michelle Torres (working paper): How do different news organizations choose different pictures to accompany articles about Black Lives Matter?

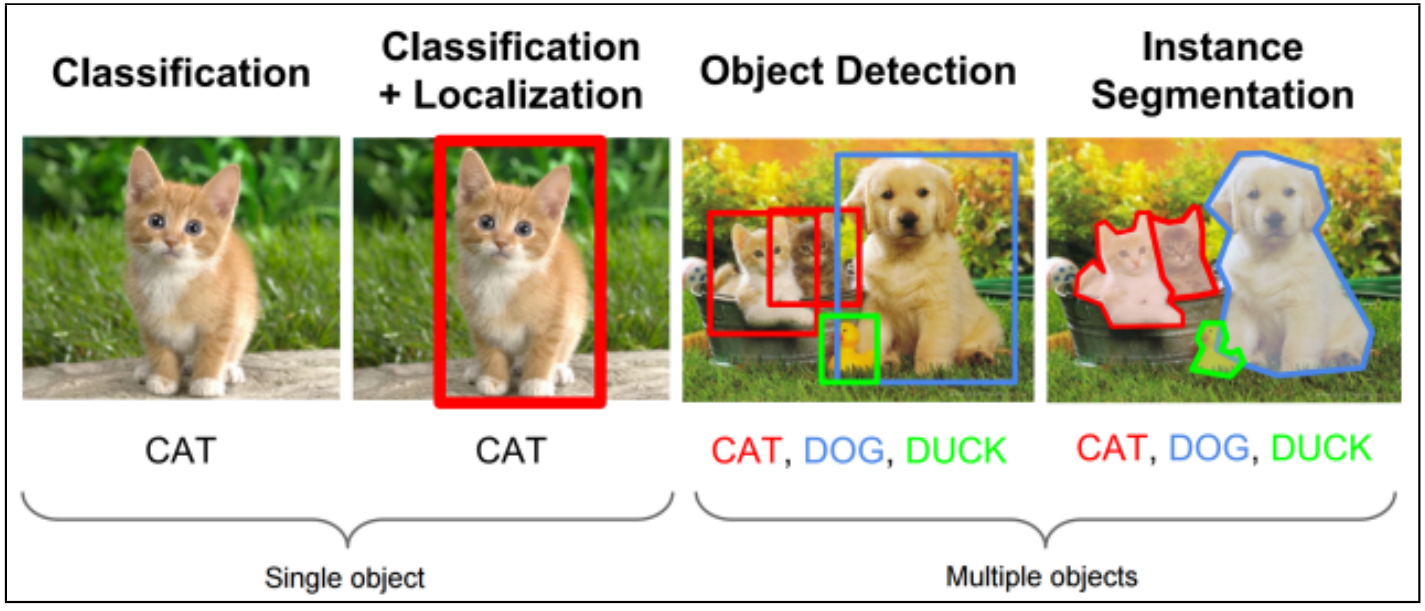
Types of Research with Images as Data

As a measurement strategy:

- Images can contain information about electoral incidents and fraud (Callen and Long (2015); Cantú (2019); Mebane et al (working paper))
- Images can help us identify and classify protest events (Zhang and Pan (2018), Won, Steinert-Threlkeld and Joo (2017))
- Nighttime lights imagery as a proxy for economic development (many authors)
- Digitized historical maps as evidence of road quality variation (Hunziker et al (working paper))
- Using social media images to measure inter-ethnic social trust (Webb Williams (working paper))

Some Common Image Analysis Tasks

- Object recognition/classification
- Object detection
- Image segmentation



`https://www.kdnuggets.com/2018/09/object-detection-image-classification-yolo.html`

Some Common Image Analysis Tasks

- Facial analysis (recognition, gender/age, emotions, etc.)
- Extract text (optical character recognition [OCR]), read handwriting
- Generate captions
- Sentiment analysis

Options for Automated Image Processing

Today we will cover:

1. Auto-tagging services: Amazon, Microsoft, Google, IBM, Clarifai, etc.
    - Using a commercial API to extract information from images

Options for Automated Image Processing

Today we will cover:

2. Transfer learning: fine-tuning a pre-trained algorithm to learn from and label new images
   - Borrow convolutional neural nets (CNNs) trained by researchers on big, benchmark datasets and slightly retrain those CNNs for our own purposes
   - Give the CNN images to *train* on and then see how accurate they are at classifying held-out *validation* and *test* images (supervised learning)
   - Once acceptable classification accuracy is reached, use the fine-tuned CNN to label additional images

Options for Image Processing

Today we won't cover (unless wanted at end):

- Non-CNN strategies for image analysis
- Open source auto-tagging image algorithms: YOLO, face_recognition, etc.
  - With the exception of two publicly available CNNs: VGG16 and Resnet18
- Building a CNN from scratch to classify new images
  - Requires *a lot* of images
  - And *a lot* of understanding of CNN architecture

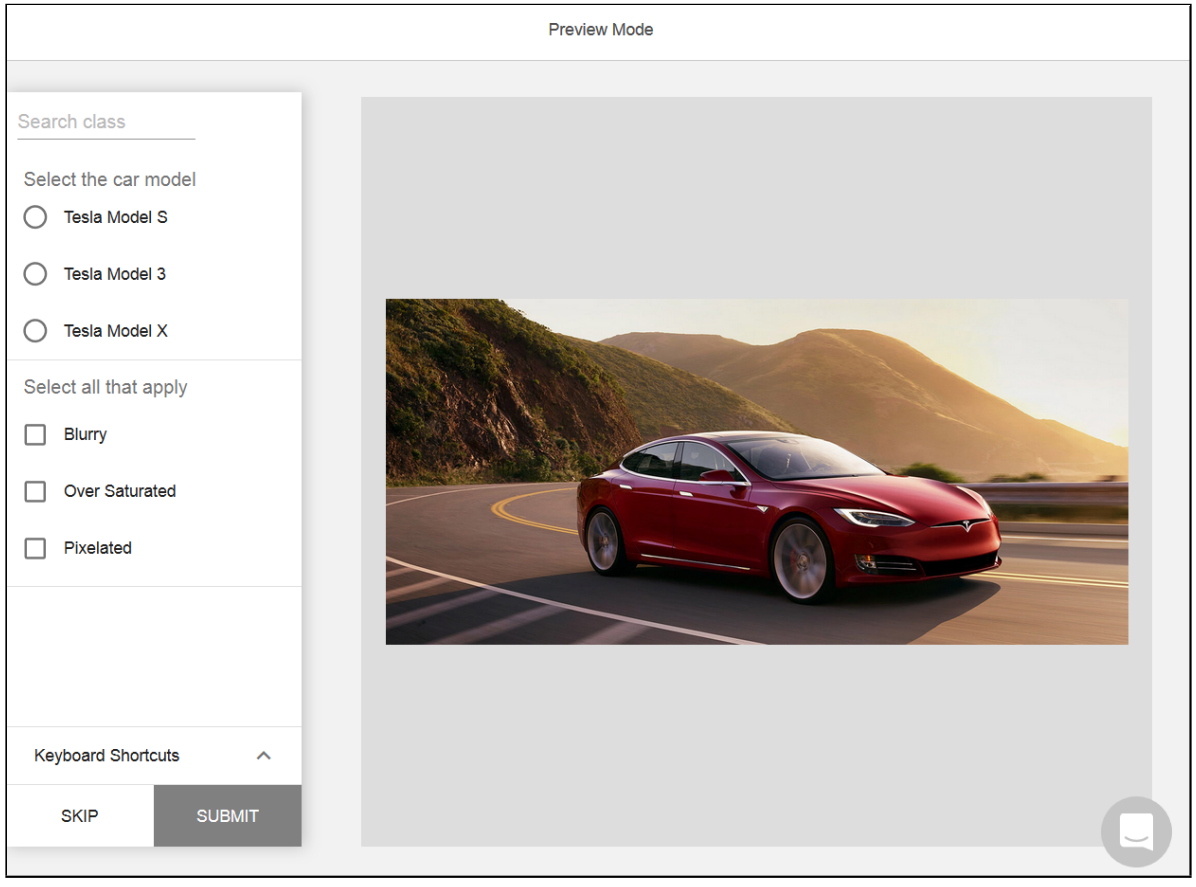
Options for Image Processing

Today we won't cover, continued:

- Software and services to generate manual labels (a.k.a image classification or annotation)
  - We started with Google docs, spreadsheets, and forms!
  - Crowdsourcing services (Mechanical Turk, Crowdflower, etc.)
  - Labeling interface products (ATLAS.ti, LabelBox, etc.)
- To validate an auto-tagging service or to fine-tune a CNN on your own, you **will** need labeled images

Manual Labeling

I just said we weren't going to talk about this...



https://labelbox.com/

Important Warnings

- AI can perpetuate biases
  - Joy Buolamwini's work (MIT): Gender Shades
  - Fairness, Accountability, and Transparency Conference [`https://fatconference.org/`]
- Data privacy
- General ethics
- More when we get to auto-tagging services...

Where to go to learn more?

- For general machine learning, consider text-as-data courses in social science
- Online coursework in computer vision: Stanford University's CS231n course at `http://cs231n.stanford.edu/`; or blogs!
- Materials on Deep Learning (e.g. Buduma 2017 textbook)

Where to go to learn more?

- On campus: computer science, data science, informatics, computational communications departments
- Ariel Rokem (UW) slides and notebooks:
  - `https://arokem.github.io/conv-nets-slides/#/`
  - `https://arokem.github.io/conv-nets`
- Forthcoming guide in *Cambridge Elements* series, "Images as Data for Social Science Research: An Introduction to Convolutional Neural Nets for Image Classification"
- YouTube video of Andreu Casas's lecture in Berlin
- PaCCS/APSA panels!

Logistics for Today

- Code available and run-able on Code Ocean
- Create account with a .edu address for 10 hours of computing time
- Handouts and links on Github: `https://github.com/norawebbwilliams/images_as_data`
- Priority today is on transfer learning
- Github/Code Ocean materials include examples for setting up cloud computing, accessing AWS Rekognition, basic image manipulation.