CVPR 2021 Workshop: Future of Computer Vision Datasets

Date: June 20, 2021

Website: https://visualai.princeton.edu/fcvd/

Organizers: Vikram V. Ramaswamy (Princeton), Dr. William T. Freeman (MIT), Dr. Fei-Fei Li (Stanford), Dr. Pietro Perona (CalTech), Dr. Antonio Torralba (MIT), Dr. Olga Russakovsky (Princeton)

Key question: What are the necessary and sufficient guidelines, tools, and frameworks for building responsible and socially-aware future computer vision datasets?

Topics:

- Recent advances in large-scale dataset collection with an eye towards social awareness
- Goals, requirements and best practices for future dataset collection and collectors
- Methods for ensuring diversity and representation in large-scale datasets
- Privacy-aware and privacy-preserving methods in image collection and annotation
- Ethical dilemmas in dataset collection
- Different sources of visual data, along with the pros, cons and challenges of each
- Necessary, sufficient, and/or appropriate safeguards to enforce in the computer vision community around dataset construction

Scribes: Vikram V. Ramaswamy, Tanushree Banerjee, Dr. Olga Russakovsky

Panel 1 : Current Limitations of Vision Datasets

Panelists: Dr. Asma Ben Abacha, Dr. Danna Gurari, Dr. Deepti Ghadiyaram, and Oliver Zendel. Moderator: Dr. Bill Freeman.

Current limitations:

1. Data is not inclusive: Does not cover all genders/skin types/ages of people
    1. Specifically, can cause issues in the medical domain
    2. Other domains: Amplifies societal biases (example: women are pictured indoors, men are pictured outdoors)
2. Dataset size and coverage is not sufficient for some domains
    1. Example: Medical domain, need all modalities, and several images for each abnormality
3. Diverse sources:
    1. Medical domain: need examples from different hospitals
    2. Geography bias in datasets like Imagenet : most images are from US/Canada/Western Europe, as a result, ‘soap’ recognizes US brands of liquid soap. [Does object recognition work for everyone? DeVries et al. https://arxiv.org/abs/1906.02659]

        1. image tags are often in English rather than the regional language, suggesting that the pictures are taken by tourists [REVISE: A Tool for Measuring and Mitigating Bias in Visual Datasets. Wang et al. https://arxiv.org/abs/2004.07999]
        2. For driving scenes, there are different types of hazards in different regions
4. Privacy concerns:
    1. Medical domain: need to preserve patient privacy
    2. Object recognition: Some images contain license plates, and credit card information, without consent
    3. Depending on the result, might sometimes need images with private info: for example, when creating images for models that can be deployed for people who are blind (what does the prescription say? Is the pregnancy test positive?).
5. Taxonomy definition:
    1. Actions labeled are taken from American logs of daily activities, might not reflect the state of the entire world.
6. Visual bias: Objects are often centered, well lit, high quality, etc.
    1. Mismatch between mainstream datasets and real-world use cases: Can pose challenges when creating models to aid people who are blind - the images they take might not be of the same quality, images are filtered to remove offensive/private content, etc..
    2. Also an issue in self-driving cars dataset - want model to work in all lighting conditions
7. Open world vs closed dataset: Datasets always have a finite number of categories - compared to the real world.
    1. Need models to be robust for a long tail distribution
8. Labeller bias:
    1. Representative images aren’t enough if labellers do not have enough context to label them correctly
9. Benchmarking bias: Need to understand if model solves a problem rather than a dataset.

Open questions:

1. Where do we draw the line on what to predict about a person?
2. How can we use/derive sensitive attributes?
3. What is the right balance between “easy” and “hard” images in a dataset?

Q&A:

1. What would take to move current datasets towards more practical use, e.g., for technology useful for blind users?
    1. Need to stop assuming that datasets are going to be large, and well-labelled. Maybe work with few-shot learning?
2. For those on product teams, what are precautions that should be taken?

1. Need to understand downstream use case well, and do fairness analysis before deploying model.
2. Maybe supplement dataset with synthetic data? But need to evaluate on real, high-quality data
    1. Comment: Biases are from not having the right tools, synthetic data will still have biases.

Panel 2: Addressing scalability of large-scale diverse datasets
Panelists: Dr. Jitendra Malik, Dr. Jeffrey Byrne, Dr. Aude Oliva and Dr. Vittorio Ferrari
Moderator: Vikram V. Ramaswamy

Main question: What are the necessary and sufficient guidelines, tools, and frameworks for building responsible and socially-aware future computer vision datasets from a scalability perspective?

1. Rethinking how we collect datasets:
    a. Original pipeline: Scrape web to get images/videos, use crowdsourcing for labels
        i. slow, expensive, outdated
    b. Online 2020 startup: Visym, allows annotators to label videos as they collect them.
    c. Solves looking for rare events
    d. Allows creator to “engineer bias”; i.e, ask for videos of specific tasks (for example, loading a car through the trunk)
        i. Comment from chat: “Engineering the bias that you want” sounds like introducing artificiality, or “role-playing” into the motions you are trying to capture. I don’t find that responsible or convincing.”
2. Solving geography bias (like open images / open images extended) created object centered bias.
    a. Need to think about protocols to get diversity
3. Need to expand beyond vision: text, sound, sensors, wifi, etc.
    a. Even datasets with just images will be used with other modalities.
4. Before constructing a dataset, need to take responsibility for the dataset.
    a. Take guidelines from other fields like experimental / social psychology.
    b. think about what we want from the dataset and consequences of models.
5. Notion of “red team” for datasets:
    a. Need a different set of people to examine dataset for bias / incorrect labels/ statistics of dataset
6. Questions about who can submit annotations:
    a. Cultural biases exist.
    b. Example: When labelling a fountain, is water part of the fountain, or not?
7. Not a question of whether dataset is ‘good/bad’ or ‘biased/not biased’, all relative to the applications
    a. “All datasets are biased, some datasets are useful”

        i. Need the creator to define biased/useful part of the dataset

8. This is a sampling problem.
    a. What’s happening now is convenience sampling. Example:
        i. To predict election results in the 1930s, collected samples by dialing random phone numbers. Since most people who owned phones then were rich, this skewed heavily Republican. Actual result: FDR won by a landslide.
        ii. Online polls for “best footballer” or “best politician”: Answers are very biased depending on respondent; hence, crowdsourcing doesn’t always give a true random sample.
    b. Sampling problem is hard; need to strive towards solution, may not get there.
        i. Potential idea: Might be able to resample cleverly once we have enough data.
9. Optimistic thought:
    a. Robustness developed by the child visual system, even just by a child running around aimlessly.
    b. Systems are not there yet, but can be achieved, if we develop better ML algorithms, better ways of exploiting the data

Question: “How to deal with the label noise (that will hugely affect the model performance) if you use tools like amazon turk?”

1. Small amounts of label noise are fine.
2. Trade-off between amount of money spent of labelling, and quality of labels.
3. See one, do one, teach one.
    a. Can train annotators well to reduce noise.
4. Can change strategy/ formulation of question to get better annotations, depends on the application
    a. For certain applications, like self-driving cars, MTurk is not an option, since it’s safety-critical.

Question: “One aspect of fairness in AI is how annotators are treated and paid. Especially with such huge datasets this becomes important. Can you comment on that?”

1. Paying per hour vs paying per task:
    a. Quality/speed trade-off
    b. As good as needed, as bad as can be without being fired.
2. Paying depending on the cognitive load:
    a. Automatic tasks are paid lesser than tasks that require thinking
    b. Adjust according to workers’ feedback
3. Within medical domain:
    a. Use crowdsourcing for training set, experts for test set.
    b. Annotating medical videos are complex tasks, and high quality annotations are hard to get.
        i. In the process of trying to improve it.

Open-Mic discussion

1. Question : What are your thoughts on using datasets with known gender/geography/race biases? Datasets have scientific and practical value, and help solve problems in the wild, w.r.t. medical domain, autonomous vehicles, etc. Balance between minimising harm while maximiising good with a dataset?
2. Minority groups are harmed because models are not trained from them. Balance the desire to collect more data from marginalized groups with their reluctance to provide data?
    a. More privacy concerns as we talk about minorities
3. About datasets and privacy, will computer vision be the first to adapt, or will organizations like GDPR set rules to be followed?
4. Different institutes play different roles:
    a. IRBs protect people in the dataset, not people harmed by downstream applications.
5. People's connections to data:
    a. Product manager, designs ML competitions, collects datasets.
    b. Contexualizes datasets, goes beyond 'view-from-nowhere' assumptions
6. Performances on ImageNet of about 95%. Expect it to go higher?
    a. COCO results tend to plateau. Maybe need bigger datasets to get break-throughs?
    b. Using ImageNet: Should we use standard test bed, should we use standard training sets, where in between?
    c. ImageNet continues to be useful for self-supervised learning.
7. Similar problems in industry to academia, but have additional levers.
8. More questions
    a. Eye on the prize - what are the big problems to be solved?
    b. larger dataset implies more complexity of model; need more complex models to generalize to validation set.
        i. Highly scalable datasets for computer vision will lead to giant modes for computer vision?

Panel 3 : Addressing Privacy and Representation concerns
Panelists: Dr. Emily Denton, and Dr. Solon Barocas
Moderator: Vikram V. Ramaswamy

Main question: What are the necessary and sufficient guidelines, tools, and frameworks for building responsible and socially-aware future computer vision datasets from a privacy perspective?

1. Need to change the incentive structure:
    a. Students are incentivized to publish quickly, not to do things slower and more ethically.
2. Ask normative questions about the dataset

    a. What problems can be solved at the dataset level?
    b. Some problems have sources beyond the dataset
    c. Hard to pin down why some behaviours are not ok
        i. Neutral label when applied to some images can be problematic

Question: Are we going to see a shift wherein only industries can collect and maintain datasets? In practice, datasets are collected by graduate students, who then graduate, making it hard to maintain the dataset over time.

1. Again, need to rethink incentive structures.
    a. Research on data ethics is viewed as less important, because it's non-technical.
    b. Acting ethically is an extra burden on an individual - instead, worth thinking about community-wide guidelines.
2. Take inspiration from other fields
    a. Healthcare works with sensitive data
        i. How do they maintain it?
    b. Social sciences:
        i. Repository of data hosted by the University of Michigan
        ii. Understanding that this is useful, and necessary
3. Norm in other fields to grant money to collect data, must become a part of computer vision, too
    a. Note: Imagenet was never funded by a grant, not considered as having a value.

Question: What is bias? Should it be defined by the government? Should models reflect society?

1. Status quo is not defensible.
    a. Can argue that a dataset is just a dataset, and only making adjustments to the model with regard to fairness is an affirmative step. But status quo exists because of a set of decisions that have been naturalized, so we don't think about them
    b. Status quo is not neutral - it is a highly political decision.
2. Harms are not theoretical / uncertain - can identify communities that have been harmed.
3. Bias isn't something that we can fix, but must understand what biases exist in the model and what this reflects, and make decisions about how to use and who will be impacted.
4. (from audience) Bias amplification is a problem:
    a. predictive policing in Oakland - sending more police to a neighbourhood implies more crimes are recorded, and forms a negative feedback loop.
    b. Paper: http://proceedings.mlr.press/v81/ensign18a/ensign18a.pdf : For predictive policing
5. Paper suggestions:
    a. Language (Technology) is Power: A Critical Survey of "Bias" in NLP (https://arxiv.org/pdf/2005.14050.pdf)
    b. This Thing Called Fairness: Disciplinary Confusion Realizing a Value in Technology (https://arxiv.org/pdf/1909.11869.pdf)

(follow up) My point was basically we cannot rely just on our researchers and what we think as a research community to define the bias. This is a much bigger problem and everybody has an opinion. Maybe you should take a poll on this? Maybe the government should precisely define this?

1. A lot of decisions are being made at an individual level, about what to build, how to build, etc. Can imagine a world with more regulatory oversight
    a. Individual decisions do leave a lot of room for different communities to be harmed.
2. Can ask what normative principles are being violated, with points where there is room for disagreement. But there are definitely concrete cases that everyone agrees on - not as much of a free-for-all as it seems.

(follow up) Is TV representation the way we should be thinking about datasets? Because, in general, we are happy with what is presented to us on TV.

1. There is an interesting connection to explore there.
    a. Politics of representation in media, so not entirely settled question.
2. Computer vision is used to figure out highlights, and can lead to discussions.

Question: It was interesting to hear Dima highlight the EPIC Kitchens decision to collect data first and define tasks second. I wonder if anyone has comments about informed consent for that kind of relatively open-ended collection versus having a clear and specific use upfront.

1. Difference between volunteering data, and putting oneself at risk.
    a. Might feel that contributing data might harm similar people
2. Questions about risk focus on research subjects, not downstream applications.
3. Can consent be given for open ended questions?
    a. Very unpredictable future research
4. Data for money - suggests that there is a limit to consent?

(follow up) So imagine that you have, instead of having a data set, what you have is an embodied AI system, like a robot that is wandering around in the wall and it's learning while it is interacting with the world and it sees things. That's an example of a system that works well because it is living in the world, just like a human is. It cannot ask consent constantly about everything it does. What are your thoughts about a system that is around in the world?

1. Depends on context and purpose.
    a. What are the expectations of the individuals it's interacting with?
    b. Even if it's passively observing for the initial data-collection phase, it will eventually be deployed in some situation, and that requires design decisions.
    c. Just reframing the set of questions from traditional data-collection in a new way.

(follow up; comment) Next decade, we're going to have robots that are mobile enough to have them in supermarkets, and already have them with autonomous vehicles, that are learning to see in the world.