visual communication
ARTICLE

The agency of computer vision models as optical instruments

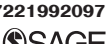
THOMAS SMITS
Utrecht University, The Netherlands

MELVIN WEVERS
University of Amsterdam, The Netherlands

ABSTRACT

Industry and governments have deployed computer vision models to make high-stake decisions in society. While they are often presented as neutral and objective, scholars have recognized that bias in these models might lead to the reproduction of racial, social, cultural and economic inequity. A growing body of work situates the provenance of bias in the collection and annotation of datasets that are needed to train computer vision models. This article moves from studying bias in computer vision models to the agency that is commonly attributed to them: the fact that they are universally seen as being able to make biased decisions. Building on the work of Bruno Latour and Jonathan Crary, the authors discuss computer vision models as agential optical instruments in the production of contemporary visuality. They analyse five interconnected research steps – task selection, category selection, data collection, data labelling and evaluation – of six widely cited benchmark datasets, published during a critical stage in the development of the field (2004–2020): Caltech 101, Caltech 256, PASCAL VOC, ImageNet, MS COCO and Google Open Images. They found that, despite all sorts of justifications, the selection of categories is not based on any general notion of visuality, but depends heavily upon perceived practical applications, the availability of downloadable images and, in conjunction with data collection, favours categories that can be unambiguously described by text. Second, the reliance on Flickr for data collection introduces a temporal bias in computer vision datasets. Third, by comparing aggregate accuracy rates and 'human' performance, the dataset papers introduce a false dichotomy between the agency of computer vision models and human observers. In general, the authors argue that the agency of datasets is produced by obscuring the power and subjective choices of its creators and the countless hours of highly disciplined labour of crowd workers.

Visual Communication 2021
Vol. 21(2) 329–349

DOI 10.1177/14703572221992097


KEYWORDS

Agency • bias • computer vision • datasets • machine learning fairness • optical instrument

In the last 10 years, the field of computer vision transformed dramatically. In the early 2000s, the *communis opinio* was that it would take decades before computers would gain a high-level understanding of digital images. However, by 2015, *The Guardian* announced that algorithms had become 'better than humans at recognising images' (Hern, 2015). Following the rapid development of the field, scientists, governments and industry professionals have employed computer vision models to make high-stake decisions in society. They are used for a wide range of tasks, such as automatic inspection in manufacturing, controlling industrial processes, medical diagnostics, policing, surveillance, navigation (self-driving cars), the organization of visual data (on social media), personalized marketing and a plethora of military applications.

While many view computer vision models as objective, scientific, or progressive, scholars increasingly recognize that they reflect, reinforce and introduce racial, social, cultural and economic inequities (Benjamin, 2019; Crawford et al., 2019; Eubanks, 2019; O'Neil, 2017). Barocas et al. (2019: 33) note that computer vision models, and machine learning applications in general, propagate 'inequalities in the state of the world through the stages of measurement, learning, action and feedback'. The field of machine learning fairness, an interdisciplinary field with strands in computer science, science and technology studies, law and ethics (Verma and Rubin, 2018), has explored when and how these disparities, commonly referred to as biases, become 'harmful, unjustified, or otherwise unacceptable' and proposed 'interventions to mitigate such disparities' (Barocas et al., 2019: 33). While most efforts focused on improving or fine-tuning algorithms, recent studies argue that issues surrounding accountability, transparency and ethics in computer vision models are primarily rooted in the collection, annotation and organization of the large datasets that are needed to train them (Jo and Gebru, 2020; Zou and Schiebinger, 2018).

Rather than looking at the biased, harmful or unjust decisions that computer vision models make, this article studies the agency that is commonly attributed to these models: the fact that they are universally seen as *being able to make* decisions in the first place. Next to the supercharged forms of agency that some (popular) scientists have ascribed to 'general' artificial intelligence, even a nuanced view of the power of computer vision models that explicitly seeks to take 'humans and humanity into account' notes that they are already able to 'reach or exceed the performance of human experts' (Akata et al., 2020).

This article starts with the observation that, for computer vision models to *be* biased, they need to be conceptualized as agential subjects that can observe and act independently from humans. Following the work of Bruno Latour (1987), we describe this attribution of agency as a form of inverted instrument agency. In his seminal *Science in Action*, Latour urges us to open the black boxes of scientific instruments and recognize their agency in the production of knowledge. To most people, this idea still feels counterintuitive. The Hubble telescope does not *see* anything by itself; the researcher *looks* through it. An MRI scanner does not *discover* the disease; the doctor uses the scanner to *make* the diagnosis. While other highly complex optical instruments are described as tools without agency, developers of computer vision models commonly present them as intelligent agents that can see for themselves.

Building on the recent interest in the connection between bias and datasets in the field of machine learning fairness, we use visual culture theorist Jonathan Crary's (1992) concept of the optical instrument to discuss computer vision models as agential optical tools in the production of contemporary visuality. In doing so, we shed light on the distribution of agency between computer vision models and the humans that develop, deploy, operate and are processed by them, and draw lines between what computer vision models can see and what humans can see through them.

How can we approach computer vision models as optical instruments? This article closely scrutinizes the six papers that describe the benchmark datasets that have shaped the field of computer vision during a critical stage in its development (2004–2020): Caltech 101, Caltech 256, PASCAL VOC, ImageNet, MS COCO and Google Open Images (Table 1). While improvements in the accuracy of algorithms can be achieved by small teams of researchers, as the introduction of convolutional neural networks demonstrates (Krizhevsky et al., 2012), the construction of a new benchmark dataset, which consists of vast amounts of annotated data, requires an enormous effort and investment. These benchmark sets are not only used to train new models but also to measure and compare their performance. As a result, the six benchmark dataset papers are central to the field of computer vision and have been cited thousands of times (see Table 1). The power to shape the field of computer vision thus firmly resides with institutions that have the required means to produce benchmark datasets.

After an introduction to the concept of optical instrument, a concise explanation of computer vision models and sections on biased datasets and methodology, this article describes how the six dataset papers produce the inverted instrument agency of computer vision models. We follow the five essential elements of every dataset paper: task description, category selection, data collection, data annotation and evaluation. While these elements are often presented sequentially, this article demonstrates that they are heavily interconnected and that the interplay between them fundamentally shapes the

Table 1. Overview of the six most commonly used benchmark computer vision datasets (2004–2018) (based on Everingham et al., 2010; Fei-Fei et al., 2004; Griffin et al., 2007; Kuznetsova et al., 2018; Lin et al., 2014; Russakovsky et al., 2015). Citations from Google Scholar (4 December 2020).

Name	Categories	Tasks	Images	Instances	Average instances/ image	Provenance	Citations
Caltech 101	101	• image classification	9247 (30–400 image/ category)	n/a	n/a	Google Image Search	3825
Caltech 256	256	• image classification	30607 (80–800 image/category)	n/a	n/a	Google Image Search/ PicSearch	2161
PASCAL VOC	20	• image classification • object detection	11000	270000	2.3	Flickr	9626
ImageNet	• 21841 synsets • 1000 image classification • 200 object detection	• image classification • single object localization • object detection	14197122		3.0	Flickr	19745
MS COCO	91	• object detection • object segmentation • semantic scene labelling	328000	2500000	7.7	Flickr	12701
Open Images	• 19794 image classification • 600 object detection • 326 visual relationships	• image classification • object detection • visual relationships	9178275	15440132 (bounding boxes)	8.1 (object detection)	Flickr	376

computer vision model. After analysing the five steps, we show that the agency of computer vision models is produced by obscuring the power and subjective choices of dataset creators *and* the countless hours of highly disciplined and regulated labour of crowd workers. In general, we argue that, if we want to understand how computer vision models make biased decisions, we should approach them in a holistic manner.

COMPUTER VISION MODELS AS OPTICAL INSTRUMENTS

The central premise of the field of visual culture studies holds that visuality is an 'historical construction' (Crary, 1992: 1). This means that the social interactions between humans, optical instruments and the world determine what we can see and how we see it. A change in one element will lead to changes in the two others. As a result, visuality has an historical dimension: what humans can see changes over time.

What role do optical instruments play in the production of visuality? Crary described how the dominant narrative of modern visuality is technocentric, viewing the invention of new instruments as the driving force behind changes in visuality. In contrast, he argued that optical instruments not only produce but that they are also products of visuality. Furthermore, their role in the production of visuality can only be understood in relation to that of the observer: the human that sees the world through them. In Crary's view, a desire to see the world differently can lead to novel uses of old instruments and the invention of new ones, which, together with the changed status of the observer, can shape new forms of visuality.

Crary wrote his influential work on 19th-century visuality while his own visual world was transforming radically. He argued that computers fundamentally changed the status of the observer (p. 1). They supplanted the 'historically important functions of the human eye' by digital streams of data where images no longer referred to any position of an observer in a real world. As a result, digital images and derivative technologies, such as 'robotic image recognition', relocated 'vision to a plane severed from a human observer' (p. 2).

Following Crary, we see optical instruments not solely as tools that enable humans to observe the world in a certain way. Human observers and instruments both have agency in the sense that visuality, that which can be seen, takes shape in the constant interaction between the two. In contrast to Crary's view that digital image technology severs images from human observers, this article underlines the role of human agents, such as computer scientists and crowd workers, in the creation of computer vision models. The desire of computer scientists to see the world in a certain way influences the characteristics of computer vision models as optical instruments, i.e. what humans can see through them. Because computer vision models are now applied in a wide variety of fields, these optical instruments have fundamentally shaped our contemporary visuality.

HOW DO COMPUTERS SEE?

In contrast to humans, computers have no specific input channel for visual information. Just like all other information, they handle images as sequences of numbers. Integer values between 0 and 255 denote the visual intensity of a pixel, a small square, on a screen. Computers process digital images as three-dimensional matrices (*n* [height] x *n* [width] x *n* [depth/colours]) of numbers. Colour is expressed by stacking three layers of intensities, containing values of red, green and blue (RGB-values). For humans, pixels operate as 'visual signifiers' (Mitchell, 1992). The contours of a human face or the pointy ears of a cat appear to us in the relationships between pixels.

Computer vision models rely on the mathematical relationships between pixels to describe and analyse digital images. Until the early 2000s, most techniques were rule-based, meaning that the code looked for pre-determined (combinations of) pixel relationships that the computer scientist associated with a particular visual signal. Deep learning algorithms *learn* these regularities between pixels. They can derive the rules, often called features, that can best predict a specific set of objects from a collection of annotated images. Convolutional neural networks, the most commonly used type of algorithm, learn the optimal combination of different types of convolutions – a mathematical process that strengthens, distorts, weakens and compresses the mathematical relationships between pixels – to best find the visual features that denote a particular visual sign, such as a cat. This information is stored in the model. The shape of the data set thus determines the shape of the model.

Computer vision models require extensive collections of annotated images to be able to learn these rules and to subsequently produce accurate predictions. As a result, as the creators of the MS COCO dataset note, these datasets of images not only provide a 'means to train and evaluate algorithms, they [also] drive research in new more challenging directions' (Lin et al., 2014: 2). The paper describing Google's Open Images dataset notes that 'the need of gargantuan amounts of annotated data to learn from' is at the 'core' of the success of modern computer vision techniques (Kuznetsova et al., 2018).

Biased datasets

The field of machine learning fairness includes a wide variety of approaches. Computer scientists have mostly focused on 'algorithmic methods to mitigate biases, viewing the dataset as fixed' (Holstein et al., 2019). However, contributions to this field from other disciplines have increasingly pointed out how data collection, annotation and organization fundamentally shape the performance, and possible bias, of computer vision models. Drawing lessons from archives and libraries, Jo and Gebru (2020) argue for a new specialization in machine learning that focuses on 'methodologies for data collection and annotation'. Other examples include the careful auditing of existing computer vision datasets, scrutinizing their geo-diversity (Shankar et al., 2017), skin

colour and gender (Buolamwini and Gebru, 2018), age and gender (Dulhanty and Wong, 2019), and social class (DeVries et al., 2019).

Next to these large-scale audits, others looked at ways to mitigate bias by removing or adding specific annotations. In February 2020, Google announced that its Cloud Vision service would no longer label persons as either 'male' or 'female' noting that 'a person's gender cannot be inferred by appearance' (Business Insider, 2020). Some of the original developers of ImageNet (Yang et al., 2020) have taken the almost exact opposite route by arguing that additional large-scale annotation of gender, age and skin colour of the person category would lead to a more representative dataset and, as a result, fairer algorithms.

Less interested in ways to mitigate, or solve, the problem of bias in computer vision models, scholars in the humanities have studied the epistemological and normative assumptions that undergird computer vision models. In contrast to the focus on fairness in other approaches, Crawford and Paglen (2019) argue in their online essay 'Excavating AI' that there is no 'neutral' vantage point 'that training data can be built upon'. The collection, categorizing and 'automatic interpretation' of images is always political: 'who gets to decide what images mean and what kinds of social and political work those representations perform?' Similar to Crawford and Paglen, instead of proposing ways to mitigate the bias in computer vision datasets, this article points to one of the most central assumptions of computer vision models: the notion that they are able to see independently from humans. We show how this idea fundamentally shapes computer vision datasets.

Methodology: archeologies of datasets

Crawford and Paglen (2019) describe the methodology of their project as an 'archeology of datasets'. While computer scientists mostly see the development of computer vision models as a technical endeavour, Crawford and Paglen argue that these models are shaped by different social and political contexts. Their method is clearly inspired by other media archeological approaches, which study why a certain medium 'is able to be born . . . picked up and sustain itself in a cultural situation' (Parikka, 2012: 6). Instead of following teleological and technocentric accounts of 'new media', media archeologists attempt to uncover the specific 'historical context' that produced them (Gitelman and Pingree, 2003: xiv).

Crawford and Paglen (2019) focus on unearthing what they describe as the three most important layers of the 'overall architecture' of computer visions datasets: the taxonomy of the categories in the dataset, the individual classes and the individually labelled images. By looking closely at the politics of labelling, they show that datasets are not only built around 'unsubstantiated and unstable epistemological and metaphysical assumptions about the nature of images, labels, categorization, and representation' but that these assumptions

also 'hark back to historical approaches where people were visually assessed and classified as a tool of oppression and race science'.

This article expands the 'archaeology of datasets' method in three ways. First, we move from an overall critique of computer vision models to a study of the scientific process that produces them. Following Latour (1987: 13–15), instead of analysing the models as closed systems, we attempt to be present when the 'black boxes are being closed'. By following the computer scientists and identifying the five essential steps of our six dataset papers, we show that the labelling of images, on which Crawford and Paglen (2019) focus, is only one of the elements that determines the shape of computer vision datasets. Second, we reveal and critique one of the most central assumptions of computer vision datasets: the notion that they can see the world independently from humans.

HOW DATASETS DETERMINE HOW COMPUTERS SEE

Gebru et al. (2020) note that, despite the importance of datasets for the development and performance of computer vision models, there is no standardized process for documenting them. To increase the transparency and accountability of data creation and use in machine learning, they call for 'datasheets for datasets', that document the 'motivation, composition, collection process [and] recommended uses' of datasets. Despite the lack of standardized documentation, in our study of six benchmark computer vision dataset papers (see Table 1), we found that they all roughly follow five research steps: task selection, category selection, data collection, data labeling and evaluation. We argue that the interplay between these steps determines how the computer vision models perform as optical instruments.

1. Task selection

Computer vision models cannot interpret images in the same highly contextual ways as humans. To have computers understand images, which computer scientists sometimes describe as the 'ultimate goal' (Lin et al., 2014) or 'holy grail' (Krishna et al., 2016), datasets developers divided the task of complete 'scene understanding' into more manageable sub-tasks. The three most common are image classification, object detection and (pixel) segmentation. In image classification, an algorithm predicts the presence or absence of at least one of the *n* categories of the dataset (Everingham et al., 2010: 305). Object detection involves the localization and prediction of one or more instances of one or more of the *n* categories of the dataset. In the easiest version of this task, the algorithm must localize and correctly identify a single instance of a single category (draw a bounding box around every single horse) and in the hardest version multiple instances of multiple categories (draw a bounding box around two horses, three persons and a car) (Lin et al., 2014). Segmentation involves the prediction of each pixel to one of the *n* categories of the dataset.

The categories can be both 'things', like a chair or a person, but also 'stuff', like wall or sand. Rather than drawing a bounding box, the algorithm draws a precise line around the object, segmenting it from the background or from other objects.

These three distinct tasks show that the dataset paper encodes or conceptualizes visual meaning as the co-occurrence and interplay of distinct visual elements, which can be clearly and unambiguously labelled, in a single image. The tasks assume that computer vision models can mimic, approach or emulate human visual understanding, by deducing meaning from analysing the relations between the visual elements of a single image. The 'holy grail' of scene understanding reduces the meaning of an image to its visual elements. Even if all these visual elements could be clearly and unambiguously identified, which is highly unlikely, computer vision models would still be unable to *understand* an image in the same way as humans, as this depends on seeing it in wider textual and visual contexts that depend on the position of the observer.

2. Category selection

Because developers of datasets see vision as derived from the interplay between different visual elements, the selection of these elements is foundational to the functioning of the model as an optical instrument. The six dataset papers offer several overlapping rationales for their choice of categories. They all refer to some sort of representativeness, the need for 'practical applications' and the practicalities of dataset collection, meaning the categories must be present on a large number of images that are easily accessible on the internet (Lin et al., 2014).

Developers of the older datasets readily acknowledged the almost random selection of their categories. The developers of Caltech 101 noted that they selected them by 'flipping through the pages of the Webster Collegiate Dictionary' (Fei-Fei et al., 2004). Because datasets became increasingly bigger and, as a result, more expensive to produce, authors came up with elaborate justifications for their categorizations. MS COCO is a prime example of this practice. Its developers obtained an initial list by combining the 20 categories of the PASCAL VOC dataset with a subset of a list containing 1,200 words that 'denote visually identifiable objects'. In the next stage, 'several children ranging in ages 4 to 8' were asked to identify every object they regularly saw 'in indoor and outdoor environments'. The authors of the paper then voted 'on a 1 to 5 scale for each category taking into account how commonly they occur, their usefulness for practical applications, and their diversity relative to other categories'. Finally, all categories for which it proved difficult to easily obtain a large number ($>$ 5000) of images on the internet were removed, leaving the authors with 92 categories (Lin et al., 2014: 3–4).

The selection of categories of the widely used ImageNet dataset might seem more rigorous. The selection is based on WordNet, a database of word

classifications which uses synsets, groups of synonyms, to organize the entire English language. These synsets are part of a taxonomy that orders them from general concepts to more specific ones. While ImageNet indeed started as an effort to collect images for all the noun synsets of WordNet, the final selection of categories for the image classification and object detection tasks is only very loosely related to the WordNet structure. Yang et al. (2020) note that all the algorithms that competed in the image classification task of the yearly ImageNet Large Scale Visual Recognition (ILSVR) challenge, which accompanied ImageNet between its creation in 2010 and 2017, were trained on the same subset of 1,000 categories. Describing the challenge, Russakovsky et al. (2015) noted that these 1,000 categories were selected 'randomly . . . followed by manual filtering to make sure the object categories were not too obscure'. The algorithms that competed in ILSVR's object detection challenge were trained on an even smaller subset of 200 categories that were 'hand-selected' as being 'basic-level object categories that would be easy for people to identify and label' (Russakovsky et al., 2015: 11).

Crawford and Paglen (2019) overemphasize the intentional political and underestimate the chaotic nature of dataset categories and their hierarchies. While ImageNet indeed contains 2,833 subcategories with the top-level category 'person', including highly problematic racial and gendered ones like 'closet queen' and 'prima Donna', almost no algorithm trained on ImageNet will take these categories into account because they were not included in the 1,000 or 200 categories of the image classification or object detection tasks, respectively. In the end, in spite of all sorts of justifications, the creators of the datasets mostly selected the categories without referring to hierarchical taxonomies or any general notion(s) of visuality.

3. Data collection

After the categories have been determined, the next step involves finding images to populate them. The categorization and collection steps should not be seen as distinct operations. In all papers, the availability of downloadable images via data providers, such as Bing, Google Images or Flickr, determined whether a category was included in the dataset. For example, Griffin et al. (2007) noted that, in creating Caltech 256, they dropped 48 of the original 304 categories because they were unable to download more than eight 'good images'. For the exact opposite reason, Caltech 101 contained the category 'snoopy' and Caltech 256 the category 'Cartman' (Fei-Fei et al., 2004; Griffin et al., 2007). The category Cartman serves no purpose in relation to the tasks of computer vision models but, as a result of the popularity of the animated series South Park in 2007, was probably easy to populate.

Extracting images from the internet using keyword searches favours visual concepts that can be unambiguously described in a textual form. As Yang et al. (2020) point out, non-visual categories, such as 'philanthropist', are

harder, if not impossible, to populate. This problem is also demonstrated by the 'long-tail' of most datasets: the phenomenon that some categories contain thousands of images, while many more only contain a few. The tail of JFT, an internal computer vision dataset used at Google as the basis for its Open Images dataset (see next paragraph), is so long that it holds 3,000 categories with less than 100 images and 2,000 categories with less than 20 images per category.

The most explicit connection between the categorization and collection step can be found in the methodology of the Open Images dataset. Its 19,794 image-level categories were not predetermined by the authors but derived from the algorithmically generated labels of JFT. This dataset holds a billion occurrences, or 'instances', of 18,291 categories on 300 million images (Hinton et al., 2015). The images were labelled by an algorithm that uses a 'complex mixture of raw web signals, connections between web-pages and user feedback' (Sun et al., 2017: 3). The creators of Open Images downloaded all images with a Creative Commons licence (CC-BY) from Flickr and used a classifying algorithm trained on JFT to label these 9 million images. Just like ImageNet, Open Images does not use all of the 19,794 labels for the object detection task. Here the authors simply selected 600 categories they 'deemed important and with a clearly defined spatial extent as boxable' (Kuznetsova et al., 2018: 4). As Figure 1 shows, these 600 categories overlap with many of the categories of the other major datasets.

Most importantly, the availability of images and, as a result, the selection of categories depends on the services used to download them. Caltech 256 used scripts to perform key-word searches for the categories on Google Images and PicSearch (Griffin et al., 2007). ImageNet used 'several image search engines' (Deng et al., 2009). In 2010, PASCAL VOC set a new standard by only using Flickr. According to the authors, the 'personal photos' of the site, which were not 'taken by, or selected by, vision/machine learning researchers', resulted in a 'very unbiased dataset' (Everingham et al., 2010: 305). MS COCO and Open Images followed the example set by PASCAL VOC.

Although humans upload millions of images to the internet, these are not unbiased reflections of our visuality. Moreover, since we started uploading images on a large scale, the places where we share and store these images regularly changed. Stuart (2019) explains the success of Flickr as resulting from the need for a place to 'store, organize, and share' digital images, as well as 'for the connections that could be made with other like-minded people'. While users were uploading 4.3 million images to Flickr each day in 2010, 10 years later, traffic has largely flowed to 'image-centric smartphone applications such as Instagram and Snapchat' (Stuart, 2019: 224). Because computer vision datasets continue to use Flickr, they not only provide computer vision models with a culturally, but also an historically biased reflection of visuality. This point is underlined by the inclusion of technologies that have rapidly disappeared

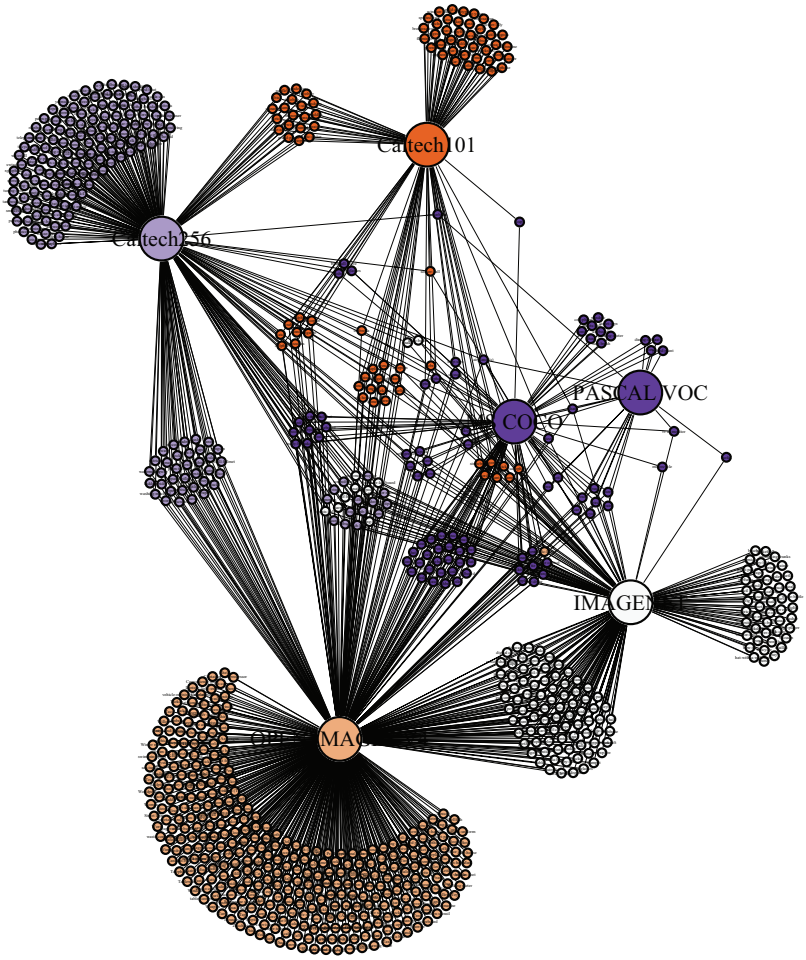


Figure 1. Network of (shared) categories between Caltech 101, Caltech 256, PASCAL VOC, ImageNet, MS COCO and Google Open Images (based on Everingham et al., 2010; Fei-Fei et al., 2004; Griffin et al., 2007; Kuznetsova et al., 2018; Lin et al., 2014; Russakovsky et al., 2015). Figure made using Gephi (ForceAtlas2 graph layout algorithm).

from our visual world, such as 'iPod' (Caltech 256, ImageNet, Open Images), in the categories of the datasets.

While scholars are increasingly concerned by the gender, racial and cultural bias of computer vision datasets, the temporal bias has received almost no attention. Bridle (2018: 44) argues that deep learning models do not account for the sometimes rapid changes in human experience: 'That which is gathered as data is modeled as the way things are, and then projected forward – with the implicit assumption that things will not radically change or diverge from previous experience.' More worrisome, deep learning models play an active role in maintaining the status quo of their data: 'computation does not merely govern our actions in the present but constructs a future that best fits

its parameters.' The amplification of a certain visuality via datasets is a defining characteristic of computer vision models as optical instruments: especially if unrecognized in public and scientific discourse, it gives them greater agency in the joint production of visuality between humans, instruments and the observable visual world.

4. Data labelling

The fourth step, the labelling of the collected data with the selected categories, is by far the most time-consuming and expensive part in the creation of computer vision datasets. While PASCAL VOC still relied on students during a single 'annotation party' (Everingham et al., 2010), other datasets, such as ImageNet, MS COCO and Open Images, made extensive use of crowd work platforms, such as Amazon Mechanical Turk. These platforms pay 'workers' small amounts of money to perform minute digital tasks. For the most straightforward task, workers were asked if an image is of something or whether a certain category is present on the image. In a slightly more challenging version of this task, they were asked to assign a category to an image from a list (Lin et al., 2014). The most difficult task involves drawing a bounding box around the object and labelling it.

It took around 50,000 workers over two years to construct the ImageNet dataset (Reese and Heath, 2016). Workers spent 70,000 hours, roughly 8 years of round-the-clock work, to label 2.5 million instances of 91 categories on the 328,000 images of MS COCO. In a 2010 presentation, ImageNet creator Fei-Fei Li wondered if the project was 'exploiting chained prisoners' (Fei-Fei, 2010). An answer depends on one's definition of exploitation and prisoner, but it is clear that crowd workers were paid very little. Most papers do not reveal the compensation per label but the Visual Genome dataset, also supervised by Li, noted that workers *could* earn between $6–8 per hour if they worked 'continuously' (Krishna et al., 2016). Considering that crowd workers would be hard-pressed to work regular hours, this reward would fall below the federal minimum wage of $7.25 per hour and well below any notion of a 'living wage' in the vast majority of cases.

Early experiments in crowdsourced annotation, such as LabelMe (Russell et al., 2008), allowed users to freely choose the parts of the image they wanted to annotate and their own labels. Later efforts, especially the ones making use of crowd workers, divided the task into easy but highly repetitive actions. Dataset creators devised methods to constantly monitor and measure the performance of workers (Deng et al., 2009; Kuznetsova et al., 2018). MS COCO made a distinction between 'good' and 'bad' workers based on their performance and discarded the annotations of the latter. Datasets like ImageNet, MS COCO and Open Images also developed special training sessions or tests that workers had to pass before they could start annotating.

The millions of labels, or 'inscriptions' in Latour's (1987) terminology, added by crowd workers, are *the* essential part of every computer

vision dataset. However, the fundamental role of workers and their labour is obscured in the discourse surrounding computer vision models. Echoing the first pages of Crary's *Techniques of the Observer*, Paglen (2016) argues that we are living in an age of 'machine-to-machine seeing'. We 'no longer look at images – images look at us.' Images in datasets are leveraged in algorithms of (visual) social control without depending on 'a human seeing-subject'. As this article shows, quite the opposite is true. The images in these datasets *have* been seen by humans. However, the Fordist-like assembly-line production of datasets by crowd workers is obscured by the highly potent agency of computer vision models in popular discourse. In contrast to Paglen, we posit that computer vision models are not lenses that see images through other images but that see images through the purposefully obscured, constantly monitored and highly disciplined labour of thousands of crowd workers.

Paglen (2016) further argues that 'if we want to understand the invisible world of machine-machine visual culture, we need to unlearn how to see like humans.' For us, this raises the question of *who* or *what* is learning to see like *who* or *what*? Computer vision scholars have described scene understanding, the ability to understand images in a human-like highly contextual way, as the ultimate goal of the field. Furthermore, they regularly claimed that computer vision outperforms human vision. Yet, the practice of dataset creation shows that the exact opposite is happening. Machines are not learning to see like humans; humans are disciplined into seeing like machines. Instead of understanding images in complex and contextual ways, dataset developers force crowd workers to look at images in the same decontextualized and fragmented ways as computer vision models.

5. Evaluation

A widely used graph shows the error rate of the winning algorithms of the ILSVR challenge and compares them to the 'human error rate' (Figure 2). In 2015, the winning algorithm outperformed humans for the first time (3.57% to 5.1% error rate) (Dodge and Karam, 2017). In the same year, *The Guardian* published an article with the headline: 'Computers are now better than humans at recognising images' (Hern, 2015). The high accuracy rates of computer vision models are an important element of their popular appeal. This shows that datasets not only provide models with the required training material but they also lend them their credibility: without the high accuracy rates, nobody would believe that computer vision models were better than humans at seeing.

The accuracy of models is calculated as follows. First, the entire dataset is divided into three parts: train, validation and test (often 80%, 10% and 10% of the images, respectively). Without going too much into the details of model training, the training set is used to fit the model, the validation set to validate the fitted weights during training to determine how best to proceed with training, and the test set contains images that the algorithm has never processed. The images in this last set are used to test the general performance of the fitted

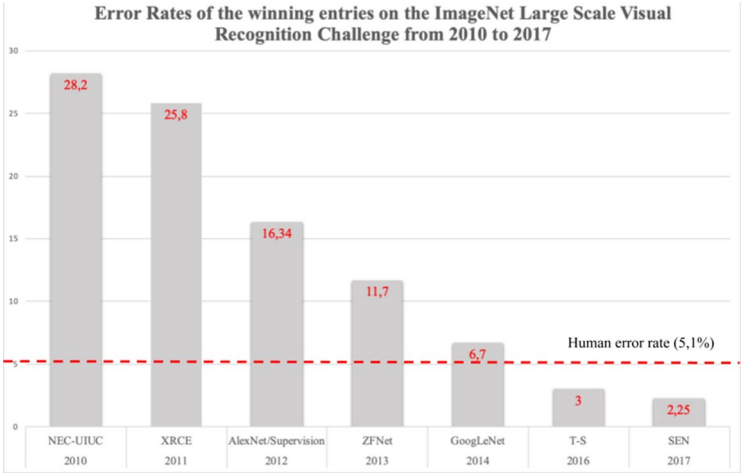


Figure 2. Error rates of the winning entries on the ImageNet Large Scale Visual Recognition Challenge from 2010 to 2017 (based on Russakovsky et al., 2015, and http://image-net.org/challenges).

model. Publicly available datasets release the images and corresponding annotations of the first two parts but keep the test part secret in order to provide for an unbiased evaluation. Otherwise, people could optimize their model's accuracy on the test set.

In 2017, after error rates dropped as low as 2%, the ILSVR challenge closed shop, considering the problem of image classification to be solved. However, the low error rates might seem less impressive if we consider that the overall accuracy of a model is calculated by taking the mean of the accuracy rates of all the categories of the dataset. There are substantial differences in error rate per category. For example, in 2014, the winning algorithm achieved 94.6% overall accuracy for the image classification task, but there was a 41 percentage point difference in accuracy between 'the most and least accurate' categories. Some categories, mostly animals like 'red fox', achieved 100% accuracy. In contrast, the hardest categories, including 'water bottle,' scored as low as 59%. The difference in accuracy was even greater for the more advanced object detection task. The average accuracy in 2014 for object detection algorithms was only 44.7%, with differences ranging between 93% ('butterfly') and 8% ('backpack') (Russakovsky et al., 2015).

The claim that computer vision models outperform humans is widely shared by computer scientists and journalists alike. However, the truthfulness of this statement entirely depends on the task that researchers asked humans and models to perform. Most papers cite the 5.1% human error given by the ImageNet paper (Russakovsky et al., 2015), which measured human performance by having a mere two human subjects compete with several computer vision models in the image classification task, assigning one of ImageNet's 1,000 specific classes to 1,500 images.

Just as the crowd workers who labelled the datasets, the humans in these kinds of studies are compared to machines instead of the other way around. The emphasis on the performance of computer vision models, especially when compared to human performance, reveals how their inverted agency is produced. Imagine the following experiment: 25 humans and 25 binoculars are given the task of reading letters on a piece of paper over several increasing distances of 5, 10 and 25 metres. A scientific paper concludes that humans and binoculars achieve 100% accuracy for 5 metres but that the accuracy of humans plummets afterward. *The Guardian* describes the experiment with the headline: 'binoculars are now better at seeing things in the distance than humans.' What is the fundamental difference between the projection of visuality of the binocular and the computer vision model? Similar to the binocular, the performance of the computer vision model can only be compared to that of a human, after another human – the scientist in this case – observes both the human observer and the models. By comparing the performance of models and humans, thus detaching the observer from the optical instrument, the discourse surrounding computer vision models obscures the role of both in the production of visuality.

CONCLUSION

Following the five essential steps in dataset creation, this article conceptualized computer vision models as optical instruments. Most importantly, while developers present these steps as sequential and go through great lengths to rationalize their choices, a close analysis of the papers reveals that the steps are highly interconnected and that the choices of developers often lack a clear methodological or theoretical rationale. Datasets like ImageNet and Open Images populated thousands of categories with images, but most models rely on only a couple of hundred categories for advanced tasks, such as object detection. While some dataset developers present a wide range of intricate procedures to justify their category selection, their own subjective judgment and all sorts of practical considerations determine the final selection.

The close connection between category selection and data collection reveals that the categories reflect a specific subset and 'moment' (in time) of the internet. Even though Flickr seems to have lost its broad appeal among internet users, it is still the essential source for computer vision datasets and, as a result, continues to shape how humans see the world through computer vision models. In contrast to widespread concerns over cultural, racial and gender bias, this inherent temporal bias of datasets has received almost no attention.

In contrast to other optical instruments, scholars and journalists often present computer vision models as intelligent agents. This article demonstrated how dataset developers kickstart this phenomenon at an early stage by actively severing computer vision datasets from humans and their labour.

Instead of acknowledging the fact that there is always a certain distribution of agency between observers and optical instruments in the production of visuality, the dataset papers present humans and computers as entities that can independently see the world without being necessarily dependent on each other.

The active disjunction of humans and computers in dataset papers is even more striking if we consider the fact that human sight is of fundamental importance in the development of these datasets. Crowd workers spend thousands of hours annotating the millions of images in datasets. In addition to these annotations, the millions of tags added by Flickr users to their uploaded photographs have also fundamentally shaped computer vision datasets. This article demonstrated that computer vision datasets are not lenses that see images through other images but optical instruments that humans use to see images through the purposefully obscured and highly disciplined labour of thousands of crowd workers.

Can efforts to remove bias from datasets and make algorithms fair(er) be successful? Scholars have mostly looked for the origin of harmful algorithmic decisions in the labelling of images. This article, alternatively, demonstrates that the origin of bias can actually be found in the interplay between all five steps in the production of datasets. As a result, researchers in machine learning fairness and adjacent fields should approach datasets in an holistic manner. Most importantly, our research suggests that, if we want computer vision models to make more equitable decisions, the people that develop them should devise more realistic and less ambiguous tasks. These tasks should be set with the clear understanding that computer vision systems do not see the world on their own but that humans see (specific parts) of the world through them.

Finally, researchers in machine learning fairness should clearly acknowledge the political dimension of their field. Wark (2019) noted that discussions over the increasing influence of algorithms are 'frequently sidetracked into the demand for a fairer algorithm, as there could still be a neutral third party above our differences, from which to pray for not much more than an equal right to be exploited by asymmetries of information.' This article has steered clear from the political consequences of attributing agency to computer vision models. However, if we truly want these models to make fair and equitable decisions, it seems inescapable that we stop separating them from humans and start recognizing the power of those that develop, deploy and use them.

ACKNOWLEDGEMENTS

Thomas Smits would like to thank the Centre for Spatial and Textual Analysis (CESTA) at Stanford University for providing the stimulating environment where the idea for this article was born.

FUNDING

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: Research for this article was financially supported by the European Research Council (ERC) under grant agreement 788572: Remembering Activism: The Cultural Memory of Protest in Europe.

ORCID ID

Thomas Smits https://orcid.org/0000-0001-8579-824X

REFERENCES

Akata Z et al. (2020) A research agenda for hybrid intelligence: Augmenting human intellect with collaborative, adaptive, responsible, and explainable artificial intelligence. *Computer* 53(8): 18–28.

Barocas S, Hardt M and Narayanan A (2019) *Fairness and Machine Learning. Limitations and Opportunities*. Available at: https://fairmlbook.org/ (accessed 2 December 2020).

Benjamin R (2019) *Race after Technology*. Cambridge: Polity Press.

Bridle J (2018) *New Dark Age: Technology, and the End of the Future*. London: Verso.

Buolamwini J and Gebru T (2018) Gender shades: Intersectional accuracy disparities in commercial gender classification. In: *Conference on Fairness, Accountability and Transparency*, 21 January, 77–91. PMLR. Available at: http://proceedings.mlr.press/v81/buolamwini18a.html (accessed 3 September 2020).

Business Insider Nederland (2020) Google AI will no longer use gender labels like 'woman' or 'man' on images of people to avoid bias. Available at: https://www.businessinsider.com/google-cloud-vision-api-wont-tag-images-by-gender-2020-2 (accessed 21 February 2020).

Crary J (1992) *Techniques of the Observer: On Vision and Modernity in the Nineteenth Century*. Cambridge, MA: MIT Press.

Crawford K and Paglen T (2019) Excavating AI: The politics of training sets for machine learning. Available at: https://www.excavating.ai (accessed 17 February 2020).

Crawford K et al. (2019) *AI Now 2019 Report*. New York: AI Now Institute. Available at: https://ainowinstitute.org/AI_Now_2019_Report.pdf (accessed 2 November 2020).

Deng J et al. (2009) Imagenet: A large-scale hierarchical image database. In: *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 248–255.

DeVries T et al. (2019) Does object recognition work for everyone? *arXiv:1906.02659*. Available at: http://arxiv.org/abs/1906.02659 (accessed 21 February 2020).

Dodge S and Karam L (2017) A study and comparison of human and deep learning recognition performance under visual distortions. *arXiv:1705.02498*. Available at: http://arxiv.org/abs/1705.02498 (accessed 18 February 2020).

Dulhanty C and Wong A (2019) Auditing ImageNet: Towards a model-driven framework for annotating demographic attributes of large-scale image datasets. *arXiv:1905.01347*. Available at: http://arxiv.org/abs/1905.01347 (accessed 18 February 2020).

Eubanks V (2019) *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. New York: St Martin's Press.

Everingham M et al. (2010) The Pascal Visual Object Classes (VOC) Challenge. *International Journal of Computer Vision* 88(2): 303–338. DOI: 10.1007/s11263-009-0275-4.

Fei-Fei L (2010) Crowdsourcing, benchmarking and other cool things. In: *CMU VASC Seminar*, Pittsburgh, PA, March.

Fei-Fei L, Fergus R and Perona P (2004) Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In: *2004 Conference on Computer Vision and Pattern Recognition Workshop*, 178–178.

Gebru T et al. (2020) Datasheets for datasets. *arXiv:1803.09010 [cs]*. Available at: http://arxiv.org/abs/1803.09010 (accessed 3 September 2020).

Gitelman L and Pingree G (eds) (2003) *New Media, 1740–1915*. Cambridge, MA: MIT Press.

Griffin G, Holub A and Perona P (2007) Caltech-256 Object Category Dataset. Available at: https://resolver.caltech.edu/CaltechAUTHORS:CNS-TR-2007-001 (accessed 16 December 2019).

Hern A (2015) Computers are now better than humans at recognising images. *The Guardian*, 13 May. Available at: https://www.theguardian.com/global/2015/may/13/baidu-minwa-supercomputer-better-than-humans-recognising-images (accessed 20 February 2020).

Hinton G, Vinyals O and Dean J (2015) Distilling the knowledge in a neural network. *arXiv:1503.02531*. Available at: http://arxiv.org/abs/1503.02531 (accessed 17 February 2020).

Holstein K et al. (2019) Improving fairness in machine learning systems: What do industry practitioners need? In: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–16.

Jo ES and Gebru T (2020) Lessons from archives: Strategies for collecting sociocultural data in machine learning. In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, New York, NY, 27 January, 306–316. FAT '20. Association for Computing Machinery.

Krishna R et al. (2016) Visual genome: Connecting language and vision using crowdsourced dense image annotations. *arXiv:1602.07332*. Available at: http://arxiv.org/abs/1602.07332 (accessed 24 February 2020).

Krizhevsky A, Sutskever I and Hinton GE (2012) ImageNet classification with deep convolutional neural networks. In: Pereira F et al. (eds) *Advances in Neural Information Processing Systems 25*. Curran Associates, Inc., 1097–1105. Available at: http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf (accessed 14 February 2020).

Kuznetsova A et al. (2018) The Open Images Dataset V4: Unified image classification, object detection, and visual relationship detection at scale. *arXiv:1811.00982*. Available at: http://arxiv.org/abs/1811.00982 (accessed 26 November 2019).

Latour B (1987) *Science in Action: How to Follow Scientists and Engineers through Society*. Cambridge, MA: Harvard University Press.

Lin T-Y et al. (2014) Microsoft COCO: Common objects in context. *arXiv:1405.0312*. Available at: http://arxiv.org/abs/1405.0312 (accessed 9 December 2019).

Mitchell WJ (1992) *The Reconfigured Eye: Visual Truth in the Post-Photographic Era*. Cambridge, MA: MIT Press.

O'Neil C (2017) *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. London: Penguin Books.

Paglen T (2016) Invisible images (your pictures are looking at you). In: *The New Inquiry*. Available at: https://thenewinquiry.com/invisible-images-your-pictures-are-looking-at-you/ (accessed 18 February 2020).

Parikka J (2012) *What Is Media Archaeology?* Cambridge: Polity Press.

Reese H and Heath N (2016) Inside Amazon's clickworker platform: How half a million people are being paid pennies to train AI. Available at: https://www.techrepublic.com/article/inside-amazons-clickworker-platform-how-half-a-million-people-are-training-ai-for-pennies-per-task/ (accessed 18 February 2020).

Russakovsky O et al. (2015) ImageNet Large Scale Visual Recognition Challenge. *arXiv:1409.0575*. Available at: http://arxiv.org/abs/1409.0575 (accessed 26 November 2019).

Russell BC et al. (2008) LabelMe: A database and web-based tool for image annotation. *International Journal of Computer Vision* 77(1): 157–173.

Shankar S et al. (2017) No classification without representation: Assessing geodiversity issues in open data sets for the developing world. *arXiv:1711.08536*. Available at: http://arxiv.org/abs/1711.08536 (accessed 21 February 2020).

Stuart E (2019) Flickr: Organizing and tagging images online. *Knowledge Organization* 46(3): 223–235.

Sun C et al. (2017) Revisiting unreasonable effectiveness of data in deep learning era. *arXiv:1707.02968*. Available at: http://arxiv.org/abs/1707.02968 (accessed 18 February 2020).
Verma S and Rubin J (2018) Fairness definitions explained. In: *2018 IEEE/ACM International Workshop on Software Fairness (FairWare)*, May, 1–7.
Wark M (2019) *Capital Is Dead: Is This Something Worse?* London: Verso.
Yang K et al. (2020) Towards fairer datasets: Filtering and balancing the distribution of the people subtree in the ImageNet hierarchy. In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, Barcelona, 27 January, 547–558. Association for Computing Machinery.
Zou J and Schiebinger L (2018) AI can be sexist and racist – it's time to make it fair. *Nature* 559(7714): 324–326.

BIOGRAPHICAL NOTES

THOMAS SMITS is a postdoctoral researcher in the ERC-funded REACT project (Remembering Activism: The Cultural Memory of Protest in Europe) at Utrecht University. His earlier work focused on 19th- and early 20th-century visual (news) culture and the application of computer vision techniques to large collections of digital historical images.
Address: Utrecht University, Trans 10, Utrecht 3512JK, The Netherlands. [email: t.p.smits@uu.nl]

MELVIN WEVERS is an Assistant Professor of Urban History and Digital Methods at the University of Amsterdam. His research interests include the study of cultural–historical phenomena using computational means with a specific interest in the formation and evolution of ideas and concepts in public discourse.
Address: University of Amsterdam, Amsterdam, Noord-Holland, The Netherlands. [email: melvin.wevers@uva.nl]