1 Introduction

Digital photography and the internet have contributed to an explosion in the production and consumption of images in social and political life. According to the Internet Live Stats website,[1] on September 18, 2018, Instagram users were uploading 869 images per second. That works out to 75,081,600 images uploaded per day to just one of many social media platforms. Facebook, Twitter, Snapchat, YouTube, and other sites also enable users to effortlessly share their photos so easily snapped using camera-equipped smartphones as well as computer-generated graphics or memes.

The study of images in social and political life is not new. For example, prior research has established that images play key agenda-setting and framing roles in newspaper coverage (Gitlin 1980; Corrigall-Brown and Wilkes 2012; Brantner, Lobinger, and Irmgard 2011; Powell et al. 2015), that they can influence people's perceptions of political candidates and their votes (S. W. Rosenberg et al. 1986; Todorov et al. 2005); and inspire (or discourage) political participation (Raiford 2007; Casas and Webb Williams 2018; Kharroub and Bas 2015). More generally, a large literature shows that visuals do a better job than written and spoken content in capturing people's attention (Dahmen 2012), facilitating information processing (Grabe and Bucy 2009; Messaris and Abraham 2001), improving information recall (Nelson, Reed, and Walling 1976; Paivio, Rogers, and Smythe 1968), and evoking emotions (Iyer and Oldmeadow 2006).

The abundance of images, however, is new. The presence of so much image content presents both promises and challenges for social scientists. The potential benefits for social science of working with large quantities of digitized images are myriad. Digitized images allow us to test existing theories in new ways and also push us to develop new theories of how image can impact society. Some scholars have already begun delving into using images-as-data, and these studies can be roughly organized into two broad categories: images-as-data in a causal framework and images-as-data for measurement.[2] The borders between these two categories of research using images-as-data are fuzzy, but we nonetheless find the distinction helpful in organizing published and ongoing work.

In a causal framework, images are either the independent or dependent variable (or both). Some prior studies use images as outcome (dependent) variable.

[1] www.internetlivestats.com/, last accessed April 26, 2020 (*Internet Live Stats – Internet Usage & Social Media Statistics* 2020).

[2] Yilang Peng maintains a very helpful list of social science papers using computer vision methods at https://yilangpeng.com/computer-vision/, last accessed April 26, 2020 (Y. Peng 2020).

For example, Joo, Li, et al. (2014) show how the choice of visuals can be a means of communicating the intent of politicians. Peng (2018) shows that news outlets choose pictures of political candidates that match the ideological leanings of the outlet. Similarly, Torres (2019) finds that ideological leanings of news outlets are linked to their choices of images to represent stories about the Black Lives Matter movement.

Other studies in the causal framework use images as explanatory (independent) variables, examining how visual inputs relate to some attitudinal or behavioral outcome of interest. For example, in Casas and Webb Williams (2018), we find that images that evoke enthusiasm and fear result in higher rates of online social movement attention and diffusion (measured by retweets) in the context of a Black Lives Matter protest. In another example, Horiuchi, Komatsu, and Nakaya (2012), using automated image analysis, find that the size of candidate smiles in campaign imagery is positively associated with electoral vote shares. Similarly, Joo, Steen, and Zhu (2015) find that candidates' facial traits can predict both party identification and vote share.

Images may also serve as a tool for measurement. Studies in this vein use images not as a treatment or outcome *per se*, but as a source of data providing evidence for another concept of interest. For examples, pictures have been used as evidence of potential electoral incidents (Mebane et al. 2017), and as evidence of tampering in vote counts (Callen and Long 2015; Cantú 2019). Image analysis can be used to detect meaningful corners in legislative districts as a possible proxy measure for compactness (Kaufman, King, and Komisarchik 2019). Lam et al. (2019) use automated image analysis to show differences in the rates of representation of women and men in news stories. Won, Steinert-Threlkeld, and Joo (2017) track protests and estimate their rates of violence using images shared on Twitter, while Steinert-Threlkeld and Joo (n.d.) extract events from images. Similarly, Zhang and Pan (2019) use a combination of images and text to track collective action events. Images can provide estimates of crowd size (Sobolev et al. n.d.). Philipp, Müller-Crepon, and Cederman (n.d.) develop a technique of image segmentation to extract data about road quality from digitized historical maps. Anastasopoulos et al. (2016) analyze images of politicians with constituents of different races in order to understand congressional homestyles. And scholars of political economy and economic development increasingly use nighttime satellite images as a proxy for development (see, for example, Henderson, Storeygard, and Weil 2012; Jean et al. 2016), where a brighter footprint indicates a better-off town or village.

In both the causal and measurement approaches to large-n images-as-data research, a significant challenge is how to accurately and efficiently extract information about the content of an image. This process is variously referred to as classifying, labeling, tagging, annotating or other, more task-specific terms

(e.g., image segmentation). The goal is to identify features of interest about or in an image. For example, a researcher might wish to know which photos include a specific object (a flag, perhaps) or a specific individual (perhaps a politician or opposition leader). Or they might want to label images for the reactions or emotions they evoke in viewers. Until recently, scholars interested in such information relied on human annotators, which can be expensive and slow. Computer vision methods now enable any researcher with some programming ability to label large quantities of images more efficiently.

In this Element, we provide code and example data in addition to the text (see Section 2 for details). We use running examples from a corpus of images related to the Black Lives Matter movement that were collected on Twitter for the Casas and Webb Williams (2018) article. Section 5 contains a detailed discussion of the original study and the data. At a high level, our goal was to determine which features of images were tied to higher rates of social movement attention and diffusion. The challenge was labeling a large number of images (around 9,500) on multiple dimensions to disentangle multiple theoretical mechanisms. Here we demonstrate how we could use deep learning to develop classifiers that can automatically label images for multiple features of interest, dramatically reducing manual annotation costs. In the next section we discuss three general labeling tasks that are of particular interest to social scientists and that are also relevant to our specific corpus of Black Lives Matter images and our research goals in that project.

1.1 Three Applications of Computer Vision for Social Scientists

Most cutting-edge computer vision work today relies on Convolutional Neural Nets (abbreviated as CNNs or CovNets). A CNN is given images with known labels to learn from (or *train on*), and then its accuracy is evaluated on a set of held-out validation or test images (again with known labels). In theory, artificial intelligence (AI) computer vision algorithms, either CNNs or other frameworks, can be trained to predict any attribute of an image. This naturally has many potential applications for social science: we could predict how large a crowd is from an image, for example, or guess whether or not the image has been altered. In practice, some labeling tasks are much easier than others. The first two tasks described below, object recognition and facial recognition, can usually (but not always) be accomplished with high accuracy given sufficient and representative training data. The third general task, visual sentiment analysis, is more difficult for reasons that help to illustrate some current limitations of computer vision methods as well as future opportunities.

1.1.1 Object Recognition and Variants

One of the earliest challenges of computer vision research was to successfully distinguish between images of two objects: cats and dogs (e.g., Golle 2008). This type of computer vision task is referred to as *object recognition*. CNNs can now accurately label a wide range of objects, including distinguishing among breeds of cats and dogs and even fish species. The object being recognized does not have to be a solo autonomous entity, however. In our Black Lives Matter study, for example, we wanted to automatically identify (or recognize) whether or not an image was of a street protest. This type of broader object recognition is occasionally referred to as "scene" recognition.

A CNN trained to recognize or classify objects can be either binary (is this an image of a flag or not?) or multiclass (is this an image of a flag, or a cat, or a protest?). Figure 1.1 provides several multiclass image labeling results along with their probabilities from one of the very first successful implementations of a CNN for object recognition (Krizhevsky, Sutskever, and Hinton 2012).

A variant of object recognition is *object detection*. In object detection, the goal is to label the different objects in an image rather than assign a single label to the whole image. Does the image have a cat in it? A dog? A person? Object recognition tasks ask, "What is this a picture of?" while object detection tasks ask, "What things are in this picture?"

A CNN trained to detect objects will also generally provide a bounding box indicating where in the image each object is located. *Object segmentation*, another common and related image analysis task, is similar to object detection but is more precise – instead of generating a bounding box, segmentation should extract the exact outlines of the object (the object "mask"). Figure 1.2 provides a visual for the differences between object classification/recognition,

Figure 1.1 Object Recognition using a CNN trained with ImageNet data (from Krizhevsky, Sutskever, and Hinton 2012)

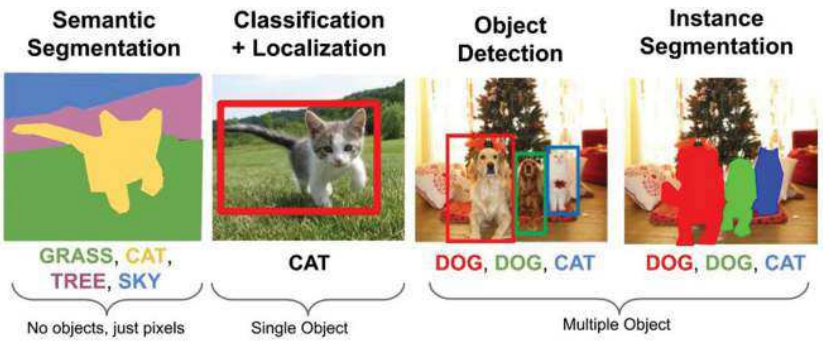

Figure 1.2 The differences between object recognition, detection, and segmentation

Source: Stanford cs231 course, reproduced with permission from Justin Johnson.

localization (a bounding box for a single object), object detection, and segmentation (of either the instance/objects in the image or semantic where each pixel is assigned a meaning).

The prediction of a bounding box or image mask provides additional information that may be of use for social scientists. For example, for a CNN trained to recognize people, we might want to use the bounding boxes to count the number of people present in an image. Or we might be able to discern the location of a particular object type (a flag, for example) across various images. Is the flag consistently in the middle of the image or is it always off to the right? For more on object detection/segmentation research and recent advances with CNNs, see Girshick (2015), Girshick et al. (2013), He, Gkioxari, et al. (2017), and Ren et al. (2015). CNNs are not the only framework for object recognition and variants. For example, a different approach by Redmon, Divvala, et al. (2016) has recently gained popularity for object detection.

In this book, we demonstrate an application of object recognition using the Black Lives Matter image corpus. Our aim is to develop a binary classifier that can automatically and accurately predict whether or not a given picture is of a protest. Depending on the particular theories that a researcher wishes to test, object recognition along these lines could be very valuable. A researcher could automatically label images for the presence or absence of police, for example, or for "I Voted" stickers.

1.1.2 Facial Recognition, Analysis, and Detection

Another class of automated image analysis focuses on faces in images. Facial *recognition* algorithms are trained to answer the question, "Who is this?" Face

Figure 1.3 Li et al.'s 2015 Face Detection algorithm (from Li et al. 2015)

detection, like object detection, picks out where faces are within an image (see Figure 1.3). This is a growing area of research where new methods are proving very accurate (Anastasopoulos et al. 2016; Li et al. 2015; Zhu and Ramanan 2012). Facial *analysis* algorithms predict general features of faces in images such as gender, age, race, or expressed emotion.

One application of facial recognition and facial detection is to identify specific individuals in images. In social science research, one use of facial recognition could be to identify and analyze politicians or other celebrity figures in images. When two politicians from the same party are photographed, are they more likely to smile at one another? Do their facial expressions predict party rifts? Tracking celebrities or politicians is one area where existing image repositories can be very helpful for training classifiers. An example is Guo et al.'s (Guo et al. 2016) compilation of celebrities. It can also be relatively easy to collect what are in effect prelabeled images of celebrities by searching for images of specific individuals online. In this Element we demonstrate a binary facial recognition classifier using images labeled for whether or not they include the singer John Legend (who appeared often in our Black Lives Matter images). John Legend could shape support for the Black Lives Matter movement because he is a popular celebrity and potential opinion leader. To measure the effect of his pictures on support, we need to know which pictures include his face.

Moving away from the Black Lives Matter example, we also demonstrate a multiclass facial recognition example that can distinguish between images of world leaders. This has potential applications for international relations and comparative politics scholars. If we can quickly identify public leaders in images, especially leaders who might not be included in standard celebrity taggers, we could potentially use that information to, for example, predict changes in trade agreements or breakdowns in ceasefires. If leaders are pictured glowering at one another, that may not bode well for peaceful, productive relations.

In our examples, the images of John Legend and world leaders are all close-ups of faces without much else in the picture. To detect the presence of a specific celebrity or world leader in a more complicated image, such as a crowd of

people, a researcher would first *detect* (or segment) the different faces in the image before applying a facial *recognition* algorithm to each of the parsed faces. For a recent study that follows this general strategy of face detection and then analysis, see Lam et al. (2019).

1.1.3 Visual Sentiment Analysis and Affect

There are at least two distinct ways to think about emotions and images. The first type of emotional content is the emotion being *expressed* by people in an image. For example, is the individual in the image happy, sad, confused, and so on? Predicting the emotion on a face falls into the category of facial analysis described earlier. The second type of emotional content is the emotion that an image *evokes* in the viewer of the image. Does the image make the viewer feel happy, sad, confused, and so on?

Both of these very different objectives are sometimes called visual sentiment analysis (VSA), although some scholars refer to the latter objective as predicting *affect*. These are very different labeling tasks and are generally more challenging than object detection or facial recognition, in part because emotions are subjective. Accurately predicting *expressed* emotions is the easier of the two sentiment tasks, but even so the task is not as easy as saying whether or not a picture has a puppy in it. *Evoked* emotions, the focus of our VSA examples, are even more subjective. Images can evoke very different responses in different people because of how the viewer filters the information contained in the image. A photo of Donald Trump will evoke very different emotions depending on one's party affiliation, for example.

CNNs now do a moderately good job of predicting evoked emotions for a variety of images (60–70% accuracy) (Peng et al. 2015; You et al. 2015). However, existing analyses are typically based on very clean images of limited scope (see Figure 1.4 for examples). In addition, as with any automated classification task, the results are only as good or as relevant as the training data. Whether an

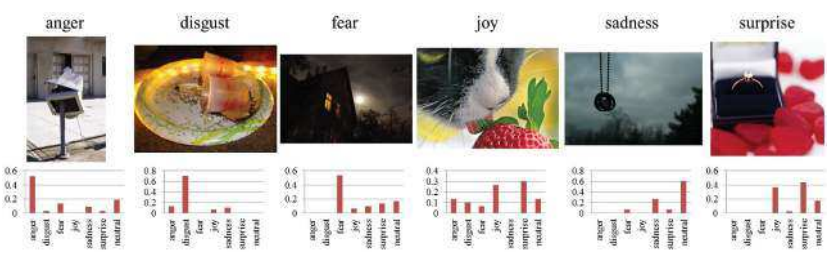

Figure 1.4 Visual Sentiment Analysis using CNN (from Peng et al. 2015)

algorithm trained to predict evoked emotions on one set of images (such as Cornell's Emotion6 from Peng et al. 2015) will do a good job of predicting evoked emotions in other contexts is an open (and very interesting) question. Nevertheless, we view this as one of the most intriguing applications of computer vision methods for social science. Emotions may be a central factor explaining why people are attracted to images, and in turn why images appear to be such powerful forms of communication. One of the questions we asked in the motivating Black Lives Matter study, for example, was whether tweets that evoked emotions such as enthusiasm, disgust, or fear are more likely to be shared (Casas and Webb Williams 2018). In this Element, we test if we can accurately predict human-generated evoked emotions labels using a CNN. While we do not achieve highly accurate results, we still present them here as a demonstration of the challenges, promise, and potential limitations of automated image analysis.

1.2 Other Computer Vision Tasks

The above are just three examples of computer vision tasks that are relevant for social scientists. Computer vision is a very large field. New applications, from automated image captioning to generating fake images, appear in academic journals and the popular press on a regular basis. For example, the aim of optical character recognition (OCR) is to extract text from images (e.g., Kulkarni et al. 2013). This is particularly valuable for studying digitized images in that people increasingly embed text in images (such as memes or screenshots of text used to circumvent Twitter's character limit). Commercial image autotagging services such as Amazon's Rekognition (described in more detail below) often offer text recognition options. An open-source Python option is Tesseract[3] (Smith 2007). Extracting handwriting, as opposed to printed text, is also potentially relevant to social scientists.

Video analysis is another relevant computer vision task for social scientists – each frame of a video can be treated as an image for analysis. Automatically analyzing video data has useful implications for many social science applications. In political science, researchers have begun using computer vision techniques to evaluate facial expressions and body language during debates (Joo, Bucy, and Seidel 2019), process campaign ads (Hwang, Imai, and Tarr 2019), and estimate party polarization (Dietrich 2019).

Dimensionality reduction is another arm of computer vision research of use to social scientists. These techniques take different approaches to reducing complicated pixel interactions into a lower dimensional space. The lower

[3] https://github.com/tesseract-ocr/tesseract, last accessed April 26, 2020 (*Tesseract documentation | Tesseract OCR* 2020).

dimensional representation can then be used for automated image clustering and other methods (see, for example, Casas, Webb Williams, et al. [2019]). A full description of these techniques is beyond the scope of this Element, however.

Although we do not explicitly demonstrate the full possible range of computer vision tasks and social science applications in this Element, the CNN logic and processes we do discuss provide an introduction to the field that will be relevant for further reading on automated image analysis and images-as-data (though we hasten to emphasize that not all applications are based on CNNs).

1.3 Overview of the Element

This Element is a practical introduction to computer vision methods for image classification using CNNs, including object recognition, face recognition, and visual sentiment analysis. It is written for social scientists who have some experience with programming languages such as R or Python. We wish to again stress that computer vision is a large and growing field, and that there are relevant tools for social science beyond CNNs and image classification. There are many available books, guides, courses, and free online materials that cover various aspects of computer vision. Most of these are geared toward computer science audiences, so this Element is intended to bring social scientists up to speed on the basics. We hope that this introduction will serve as a springboard for social scientists interested in using computer vision methods in their own images-as-data work.

One extremely helpful advance in image classification is the existence of huge, labeled repositories of images. These repositories are not without controversy, particularly surrounding the sources of the images and possible privacy violations (see Metz 2019 and Section 8 for more). Some previous benchmark datasets may no longer be available because of these concerns. Competitions to build the most accurate classifiers for standard image repositories have resulted in a plethora of trained supervised learning algorithms that can accurately predict the known labels. Many of these trained CNNs are open source. As a result, other researchers can now borrow trained CNNs available commercially or in open-source libraries. As we demonstrate, it is also relatively easy for researchers to adapt these existing algorithms to new purposes (i.e., to assign a different set of labels than those that are in the original benchmark image repository). Whereas the original algorithm may have been the product of many months of effort using millions of labeled examples, this *fine-tuning* or *transfer learning* can produce remarkably accurate results using a much smaller training set of images (as few as 100 in some cases).

CNNs are a specialized type of deep learning algorithm. They take as raw data the values of each pixel in a digitized image, generally either red, green, and blue (RGB) values or grayscale. They use lots of prelabeled training images to "learn" which pixel combinations are associated with the desired labels. An algorithm's accuracy is assessed by applying it to prelabeled validation or test images that are not included in the training set. Once performance is satisfactory, the algorithm can then be used to label large numbers of additional images quickly and at very low cost.

In this Element we do not focus on the process of training a CNN classifier from scratch (although we do discuss how artificial and convolutional neural networks work), as this usually requires a very large number of manually labeled images and lots of computational power. A social science researcher looking to label images using what might be considered "conventional" labels (e.g., whether the picture contains basic objects like cats, dogs, cars, or motorcycles) does not need to know how to build their own new classifier. They can simply use one of many available off-the-shelf trained algorithms (either commercial or open source) that are optimized for these labels. We provide a brief introduction to one commercial service (Amazon's Rekognition) as one of these "autotagger" options. However, rarely are social scientists interested in these conventional labels. Hence, our primary focus is on fine-tuning, where the objective is to adapt or "fine-tune" an existing algorithm to assign a specific set of labels developed by a researcher. With these tools, new subfields of social science are in the making.

The remainder of this Element proceeds as follows. In Section 2 we first discuss some technical requirements for the methods described in this Element. While computer vision methods have become increasingly accessible, there are some prerequisites to using the techniques we describe. In Section 3, we provide an introduction to the basics of deep learning and CNNs. Section 4 describes the main method we advance in this book, fine-tuning a CNN, with subsections on image preprocessing, hyperparameters, and diagnostics.

Turning from theory to application, Section 5 introduces the data (images) used in our examples. As part of an earlier project, we collected and labeled 9,500 unique images shared on Twitter by people tweeting about a Black Lives Matter (BLM) protest. We then labeled the images for select content and for the emotions they evoked in viewers. The example applications evaluate the degree to which we can successfully predict the manual labels using automated means.

One way to leverage the deep learning revolution for image analysis is to use off-the-shelf autotaggers, many of which are based on CNNs. In Section 6 we use some of the BLM images to demonstrate both the promise and the