Amsterdam University Press

COMPUTATIONAL COMMUNICATION RESEARCH 4.1 (2022) 11-67
HTTPS://DOI.ORG/10.5117/CCR2022.1.001.JOO

ARTICLE

Image as Data: Automated Content Analysis for Visual Presentations of Political Actors and Events

Jungseock Joo
Communication - University of California, Los Angeles

Zachary C. Steinert-Threlkeld
Public Policy - University of California, Los Angeles

Abstract

Images matter because they help individuals evaluate policies, primarily through emotional resonance, and can help researchers from a variety of fields measure otherwise difficult to estimate quantities. The lack of scalable analytic methods, however, has prevented researchers from incorporating large scale image data in studies. This article offers an in-depth overview of automated methods for image analysis and explains their usage and implementation. It elaborates on how these methods and results can be validated and interpreted and discusses ethical concerns. Two examples then highlight approaches to systematically understanding visual presentations of political actors and events from large scale image datasets collected from social media. The first study examines gender and party differences in the self-presentation of the U.S. politicians through their Facebook photographs, using an off-the-shelf computer vision model, Google's Label Detection API. The second study develops image classifiers based on convolutional neural networks to detect custom labels from images of protesters shared on Twitter to understand how protests are framed on social media. These analyses demonstrate advantages of computer vision and deep learning as a novel analytic tool that can expand the scope and size of traditional visual analysis to thousands of features and millions of images. The paper also provides comprehensive technical details and practices to help guide political communication scholars and practitioners.

Keywords: Computer vision and deep learning, convolutional neural networks, automated visual content analysis, visual self-presentation, visual event framing

1 Introduction: Image as Data

Images affect politics because they trigger emotions and provide information shortcuts to evaluate complex issues (Popkin, 1994), but researchers have rarely analyzed them in large quantities because of the difficulty of extracting politically relevant information. Methods that simplify the extraction of meaning from images now exist. In light of a growing body of work taking advantage of these methods (Cantu, 2019; Casas & Webb Williams, 2019; Haim & Jungblut, 2020; Peng, 2018; Torres, 2018; Xi et al., 2020), this article provides a conceptual overview of the leading class of models, convolutional neural networks, and applies them via two examples, each of which shows a different approach to treating images as data. Images contain information absent in text, and this extra information presents opportunities and challenges. It is an opportunity because one image can document variables with which text sources (newspaper articles, speeches, or legislative documents) struggle or at scales not possible with manual coding (Valentino, Brader, Groenendyk, Gregorowicz, & Hutchings, 2011). It has been a challenge because of the technical difficulty of identifying the objects and concepts encoded in an image, requiring researchers to rely on manual coding. Because human coders are slow, expensive, and have different interpretations of the same images, studies using images have historically used few.

Intrinsic features of text and images differ in key ways. These differences explain why the latter have resisted automated analysis. The fundamental units of images, *pixels*, contain less meaning than words, the building blocks of texts. Once built, however, image classifiers are more universally applicable than text ones. Intuitively, the same concept, e.g. "violence", expressed in written languages requires training separate models to understand words and syntax from multiple languages. Visual language is more universal (Graber, 1996), so one image model of "violence" can apply to events from places and periods that could require several text models. For example, a police officer battering protesters in Hong Kong or Spain will look more similar to each other than the words that would be used to describe that event. Appendix Section A elaborates this assertion.

Advances in computer vision and machine learning algorithms, specifically the rise of deep learning and convolutional neural networks (CNNs) (LeCun,

Bengio, & Hinton, 2015), have lessened the challenge of automated visual content analysis. Along with increased hardware capabilities, these algorithms have expanded the frontier of computer capabilities. For social media platforms, these advances mean automatically recognizing faces in uploaded images. For governments, these advances mean increased biometric security as well as policing capabilities (Kargar & Rauchfleisch, 2019). For researchers, these advances mean the ability to better measure existing concepts (Hsiang, Burke, & Miguel, 2013), operationalize measures previously only available in theoretical models (Grabe & Bucy, 2009), and do both with greater geographic and temporal resolution than previous efforts (H. Zhang & Pan, 2019).

This paper demonstrates how large scale image datasets can be incorporated in research and introduces computational techniques which significantly enhance the scope, size, and efficiency of image analysis. To this end, it first provides a conceptual overview of how convolutional neural networks work and how they are structured compared to traditional computational methods. Next, it introduces tasks – image classification, object detection, and person attribute recognition – at which CNNs excel. It then explains how to develop and train a CNN, including using off-the-shelf models, and validate results. These sections are light on technical detail, which we leave in the Supplementary Materials for the interested reader.

With these techniques, the second half of the paper provides examples to demonstrate different approaches to images as data and discusses ethical concerns facing researchers. In the first example, we examine self-presentation of politicians in the U.S. using their Facebook photographs. Politicians choose these photographs to communicate a range of their activities, policy priorities, and even personality to supporters. Prior research has found gender and party differences (Carpinella, Hehman, Freeman, & Johnson, 2016), typically using a small number of images and a coding scheme where researchers predefine variables. In contrast, our approach employs an off-the-shelf computer vision method, the Google Vision application programming interface (API), which automatically detects thousands of distinct visual objects and attributes from images. This example exemplifies a data-driven, bottom-up process for conceptualizing visual self-presentation which can be confirmed and refined by a theory-driven approach. We find clear party and gender differences. The second example investigates visual event framing in social media, focusing on analyzing how Twitter users describe protest events. We develop a series of CNNs to automatically identify protest images, add labels (if the image contains state violence, police, or large groups, for example) to them, and measure image duplication rates within those labels. Focusing on protests across five countries, we show

how protesters choose to frame events and how this framing varies across labels. We find that frames emphasise state violence and group activity.

A growing body of work shows how images can detect voter fraud (Cantu, 2019), measure protests in China (H. Zhang & Pan, 2019), understand nonverbal communication in presidential debates (Joo, Bucy, & Seidel, 2019), reveal media bias (Peng, 2018), or provoke emotional responses to protest (Casas & Webb Williams, 2019). This paper fills a hole in the literature by explaining how the methods these papers use work. While many textbooks perform the same function, their examples and presentation are not aimed at practicing researchers. This paper gives the reader an intuitive understanding of how deep learning and computer vision work, directs them to appropriate resources to learn more, and stimulates interest by showing intentionally suggestive applications in two domains.

2 Why Study Images?

In addition to their widespread availability and amenability to automatic analysis, images are of interest for two reasons: they are key inputs into individual decision making and can provide improved data to advance research agendas.

2.1 Inputs Into Decision Making

Humans are more likely to notice and learn from visual information than textual. Images provide information about a situation, such as a politician's patriotism or the beneficiaries of a new healthcare policy, more accessibly and quickly than text (Barry, 1997). This faculty is probably because writing is a technology that must be learned, while visual processing is evolutionarily antecedent (Gazzaniga, 1998). Compared to text, images provide "a more comprehensive and error-free grasp of information, better recall, and greater emotional involvement" (Graber, 1996).

Moreover, emotional reactions often drive human behavior, and visuals evoke these reactions more strongly than text (Grabe & Bucy, 2009). Images drive emotions (Tukachinsky, Mastro, & King, 2011), and emotions lead to information-seeking and political participation (Marcus, Neuman, & MacKuen, 2000; Valentino et al., 2011). These emotions affect decisions ranging from vote choice (Joo, Steen, & Zhu, 2015) to mobilization (Casas & Webb Williams, 2019). Understanding how images matter for politics is therefore central to understanding how politics works.

Images are a powerful means of persuasion and a critical device in media framing, agenda-setting, and propaganda (Geise & Baden, 2015). They are

carefully selected, edited and presented to audience, conveying various intentions encoded in subtle or sometimes very obvious ways (Joo et al., 2014). Scholars have demonstrated the effect of visuals on issue perceptions (Soroka, Loewen, Fournier, & Rubenson, 2016) and candidate evaluations (Barrett & Barrington, 2005; Joo, Bucy, & Seidel, 2019; Kang et al, 2020; Chen, Park, & Joo 2020). Given a multimodal message, the audience construct a blended representation of issues and events from verbal and visual cues, and when they are not congruent, the visual one may dominate (Gibson & Zillmann, 2000).

Images encapsulate underlying, complex issues, providing an information shortcut for individuals to evaluate multi-faceted political issues (Popkin, 1994). For example, Americans who watched the 1960 United States presidential debates claimed that John F. Kennedy outperformed Richard M. Nixon; those who listened, the opposite. In 1976, photographs of President Gerald Ford failing to husk a tamale conveyed aloofness to a large part of the Texas electorate, arguably costing him the state and presidency. More recently, photos from Abu Ghraib prison increased opposition to the Second Iraq War. Outside of the United States of America, video of a self-immolated fruit vendor spread throughout Tunisia, sparking the Arab Spring. The Tank Man image from Tiananmen Square symbolizes the Chinese Communist Party's resolve.

2.2 Advancing Communication Research

Framing. Facial expressions of politicians are an indicator of overall favorability. For instance, a smiling face is more likely to convey a positive sentiment about the main person being depicted. Based on this assumption, Groeling, Joo, Li, and Steen (2016) have examined the degree of media bias present in TV news programs in the U.S. by automatically analyzing facial expressions of presidential candidates across news networks. Going beyond traditional professional sources, attempts have been also made to analyze political images in social media. For instance, You, Cao, Cong, Zhang, and Luo (2015) have analyzed multimodal cues of Flickr posts related to presidential candidates in the U.S. to predict election outcomes based on facial expressions and hashtags.

Candidate Evaluation. Computer vision methods have also shown the potential effects of politicians' facial appearance on voters' trait judgment and election outcomes. Personality inference from facial appearance is a well studied topic in psychology (Zebrowitz & Montepare, 2008), and political scientists have attempted to explain public responses to politicians, including election outcomes, based on the physical appearance of political leaders such as their visually-inferred competence (Todorov, Mandisodza, Goren,

& Hall, 2005). Automated models have been used to extract visual features from facial images to predict subjective trait judgments on dimensions such as intelligence or trustworthiness (Rojas, Masip, Todorov, & Vitria, 2011; Vernon, Sutherland, Young, & Hartley, 2014). Automatically inferred facial traits may also predict election outcomes (Joo et al., 2015).

Section 6.1 's analysis of politicians' images shared on Facebook shows how deep learning informs the study of elected officials' self-portrayal (Fenno, 1978). Most people access news through multimodal (a combination of print, audio, or visual) media; even newspapers devote significant space to photographs, and saying that the visual dimension of politics matters is not new (Barrett & Barrington, 2005; Gilliam Jr & Iyengar, 2000; Grabe & Bucy, 2009; Hansen, 2015; Schill, 2012). Presidential debates, for instance, are both verbal exchanges of policy positions and, because they are televised, conveyors of emotions and tensions between the candidates (Joo et al., 2019; Shah et al., 2016). Indeed, the nonverbal cues and visual exposures of politicians may encode their emotions and invoke voter reactions (Grabe & Bucy, 2009; Sullivan & Masters, 1988). Prominent recent examples from the United States include Donald Trump's stalking of Hillary Clinton during their debates as well as Speaker Pelosi's sarcastic clapping during President Trump's 2019 State of the Union address. Visuals are an especially important information shortcut for low-information voters (Lenz & Lawsom, 2011), which may explain why out-parties tend to prefer more attractive candidates (Atkinson, Enos, & Hill, 2009).

Media Bias. Computer vision techniques also enable measurement of media bias and framing, which Section 6.2 demonstrates. Large literatures analyze media bias of political news coverage (D'Alessio & Allen, 2000; Gentzkow & Shapiro, 2010), its public perception (Watts, Domke, Shah, & Fan, 1999), and effects (Baum & Groeling, 2008; Druckman & Parkin, 2005). Measuring media bias objectively is a challenging task because the ground truth is unknown. For systematic analysis, studies have relied not only on verbal content analysis (Baum & Groeling, 2008) but also on visual analysis ranging from counting the number of photographs of a candidate in newspapers (Stovall, 1988) to manually coding how favorable or unfavorable their portrayals are (Grabe & Bucy, 2009). Computer vision based techniques can significantly reduce coding costs by automatically recognizing people in photographs, their expressions and favorability and comparing the results across outlets or candidates (Peng, 2018). As traditional media faces increasing competition from online, decentralized content producers (Blumler & Kavanagh, 1999), the ability to analyze image framing at scale will only increase in importance (Schmuck & Matthes, 2017).

Opinion Formation. Issue behavior responds to visual communication. Negative opinions towards immigration, for example, may be due to media conflation of immigrants with crime and disease (Tukachinsky et al., 2011). Attitudes about immigration are more positive, however, when the imagery accompanying an article evokes European, instead of Latin American, immigration, and this effect is caused by intervening emotional variables, especially anxiety (Brader, Valentino, & Suhay, 2008). The power of images explains why anti-immigrant rhetoric focuses on symbolic (visual) appeals over economic ones (Schmuck & Matthes, 2017). Deep learning techniques can also offer insight into what features of images provoke behavior. For example, people are more likely to pay attention to negative or shocking events (Baumeister, Bratslavsky, & Vohs, 2001), so newspapers and television report those type of events. But how those events are portrayed should also

Polarization. Computer vision techniques can also shed light on changes in political polarization. Dietrich (2018), for example, uses video data of members of the House of Representatives to show that frequency of physically crossing the aisle to talk to members of the other party predicts how polarized an upcoming vote will be. Which images politicians share on their Facebook, Twitter, and Instagram profiles may reveal their ideological position (Xi et al., 2020). Measuring ideology via images would prove especially useful for evaluating incumbent challengers since their ideology cannot be determined from voting history and campaign donation data may not provide this information early enough in an election cycle (Bonica, 2018).

Appendix Section B details additional applications in the study of development, natural disasters, civil war, state capacity, and protests.

3 Computer Vision and Deep Learning

Computer vision tries to solve visual *problems* with any kind of methods, and deep learning refers to efficient *methods* applicable to any kind of data, not just images.

Computer vision is an interdisciplinary branch of study crossing computer science, statistics, cognitive science, and psychology. Its primary goal is automatic understanding of visual content, *i.e.*, to replicate human visual abilities with computational models. Human vision is versatile, complicated, and not fully understood, and computer vision systems cannot simply reconstruct the mechanisms of human vision. Therefore, research has mostly focused on using statistical inference and machine learning approaches to deal with noisy inputs and discover meaningful patterns. In practice, this

pipeline usually consists of collecting a large amount of visual data, manually labeling them, and training a model that best explains the observed data.

Prior to the start of the deep learning era, the insufficient reliability and accuracy of computer vision based methods was the primary factor limiting practical applications, including political analysis of visual content. The field made a dramatic leap forward with the advances in deep learning based approaches (Krizhevsky, Sutskever, & Hinton, 2012). The next section introduces those advances.

3.1 Deep Learning and Hierarchical Representations

Deep learning refers to a class of machine learning methods which utilizes hierarchical, multi-layered models.[1] In contrast to single-layered models, such as linear regression, in which output variables can be directly computed from input variables, "deep" models employ repetitive structures with multiple layers such that the final outputs of the model are obtained through a sequence of operations applied to the input data and intermediate results.

In machine learning, hierarchical model structures are commonly used, as in some topic models (Griffiths, Jordan, Tenenbaum, & Blei, 2004). These models incorporate different levels of representations which capture structured and global information (*e.g.*, topic), as well as local information (*e.g.*, words) from input data. In political science, hierarchical text models have been used to study Congressional press releases (Grimmer, 2010) and open-ended survey responses (Roberts et al., 2014).

Deep learning based methods profit from the same hierarchical structure, but they employ a larger number of consecutive layers. These extra layers add the "deep" to the learning. Indeed, the success of deep learning is related to the depth of the models, as additional layers can encode abstract visual attributes and capture more complex data distributions than what shallower models can (Delalleau & Bengio, 2011; Eldan & Shamir, 2016).

Furthermore, these complex internal structures are directly learned from the images rather than manually defined by the researcher. Direct learning contrasts with other approaches, explained in the next sub-section, that require the researcher to specify the visual features of an image that correspond to the desired image label ("car", "torch", "rally", &c). That approach is similar to using a dictionary in text analysis to identify texts as being about a topic if it contains some combination of keywords in that dictionary. Dictionary approaches to text are more productive than manual feature specification in images because text can be represented more simply. Deep learning, by contrast, does not use a pre-defined feature set, an advantageous approach when applied to complex data such as images.

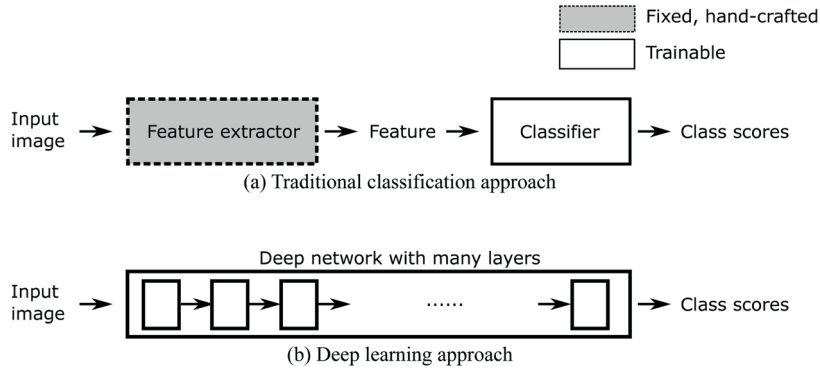


Figure 1: Comparing Deep Learning to Traditional Computer Vision Methods

3.2 Advances Over Previous Computer Vision Methodology

Artificial neural networks have a long history in machine learning and computer vision and regained popularity after Krizhevsky et al. (2012) demonstrated a 21.9%-33.8% improvement in image classification performance using a convolutional neural network on a benchmark dataset, ImageNet (Russakovsky et al., 2015). Two major requirements for deep learning, very large-scale datasets and high-performance computation using graphical processing units (GPU), contemporaneously became available.

Traditional computer vision methods heavily rely on manual feature engineering. These methods typically utilize a two-step process, as shown in Figure 1. Given raw input image data, the methods first extract features using a hand-crafted feature extractor. Hand-crafting means that a researcher has to manually design and define the feature extraction function based on instinct and experience. Common features include edge histograms, local image contrast, and color distributions. These features should capture the most important cues in the raw data, and a separate classifier, such as logistic regression, exploits them in the second step.

In contrast, deep learning methods learn their representations directly from data without hand-crafted feature extraction. These methods employ a data-driven approach in feature learning and train an integrated model that will automatically learn and capture low- and high-level representations of data. This approach is advantageous because the learning algorithm can discover many subtle features which are specific to the given task. In other words, the features in deep learning are optimized for the task during

training, as opposed to traditional methods that require the researcher to specify features before training.

The Appendix provides technical details about how convolutional neural networks work. We leave the technical discussion for the appendix because it is challenging for practitioners to design and construct their own CNN from scratch. Rather, it is much more efficient to acquire a training set of images that can be used to customize an existing pre-trained model. Appendix Section D elaborates details for transfer learning, training, and validating models for advanced readers.

4 Tasks in Computer Vision

This section discusses three common tasks in computer vision: image classification, object detection, and face and person analysis.

4.1 Image Classification

Image classification is a popular topic in computer vision. Given an input image, I, the goal of image classification is to assign a label, y, from a predefined label set, Y, based on the image content:

$$y^* = \arg\max_{y \in Y} p(y|I). \tag{1}$$

For binary classification, Y = {positive (belongs to category), negative (does not belong to category)}. In general, Y may contain any number of possible labels. The posterior probability for each label is computed for a given input image, and the classifier chooses the category with the highest output score, similar to how a topic is assigned by some text classifiers.

In multiclass (multinomial) classification, Y contains more than two, mutually exclusive categories. The softmax function is commonly used in multiclass classification to normalize output scores over multiple categories such that the final scores sum to 1; the class with the highest normalized output is assigned to that image. Suppose that the last fully connected layer outputs a vector $\mathrm{x} = (x_1, x_2, \ldots, x_n)$, where x_k is the raw output score before normalization for the k -th class out of n classes. The final score will be obtained as follows.

$$p(y = k|I) = f_k(x) = \frac{\exp(x_k)}{\sum_{j=1}^{n} \exp(x_j)}. \tag{2}$$

Figure 2: Example results of image classification with the confidence scores computed from a CNN. Red color indicates the correct category and blue color indicates the incorrect categories.

An image can contain more than one label. In this situation, called multilabel classification, an image is allowed to be assigned more than one label. For instance, Section 6.1 uses multilabel classification to understand politicians' imagery, and Section 6.2 uses multiclass classification to identify images of protest.

4.2 Object Detection

The goal of object detection is to localize (find) objects in images and assign a category (gun, flag, or cup, for example) to each object. The output of object detection is a set of detected objects, their locations, and categories. Figure 3 shows example results of object detection with detection scores from Google's Cloud Vision API.[2]

Object detection is a more complex problem than image classification because the model should classify the types of objects and their locations in the image. In practice, many object detection systems utilize a two-stage procedure. First, the system generates a number of generic object "proposals" from an input image (Uijlings, Van De Sande, Gevers, & Smeulders, 2013). These proposals are image subregions which the system believes are likely to contain an object instance, regardless of its category. An object location is represented by a rectangular bounding box, (x, y, w, h), indicating the coordinates and the size of the bounding box. This bounding box is the rectangular area of the minimum size that can cover all the pixels that the object occupies in the image. Second, the image classification step is then applied to each object proposal to determine whether it belongs to a category or is background.

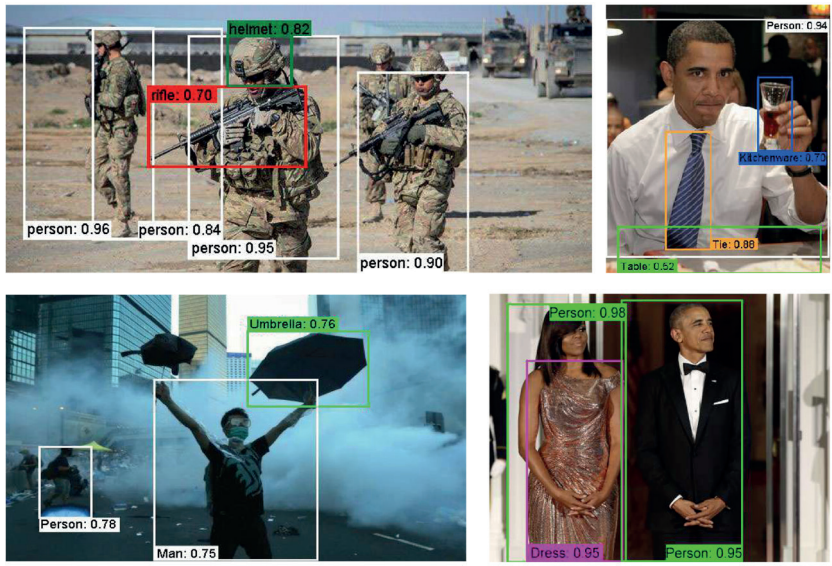


Figure 3: Example results of object detection by Google Cloud Vision API.

4.3 Face and Person

The human face has received enormous research attention as a special domain in computer vision since the 1970s, for two main reasons. First, facial recognition has many useful applications, such as for personal identification or security. Second, it is relatively easy to handle face images compared to other objects because the appearance of a human face is consistent across individuals but distinct from other objects. These properties motivated early approaches such as automated feature extraction (Kanade, 1977), feature learning with neural networks (Fleming & Cottrell, 1990), and classification based on statistical analysis of data (Belhumeur, Hespanha, & Kriegman, 1997). Existing work in this topic can be categorized into three areas: face detection, face recognition, and person attribute classification.

Face Detection. Face detection refers to finding the location of every face in an input image. This is a special case of object detection where only one object category (face) is considered. Both deep learning methods (Ranjan, Patel, & Chellappa, 2017) and traditional methods (Viola & Jones, 2004) are widely used.

Face Recognition. Face recognition classifies the identity of a person from a facial image. Most recent approaches in face recognition are based on convolutional neural networks. A recent study by Facebook (Taigman,

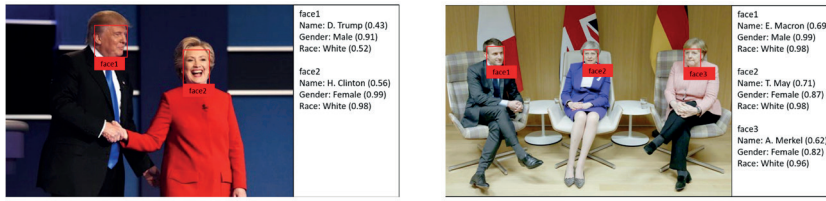


Figure 4: Example results of face detection, recognition, and attribute classification. The labels were computed by a model from Kärkkäinen and Joo (2019).

Yang, Ranzato, & Wolf, 2014) reports that a model based on a CNN is as accurate as human annotators in face verification, after training on 4.4 million labeled face images obtained from their users.

Person Attribute Recognition. A face provides clues for recognizing human attributes such as demographic variables (*e.g.*, gender, race, age), emotional states, expressions, and actions. Large scale datasets of facial images and attribute annotations are also available (Liu, Luo, Wang, & Tang, 2015) and enable training a deep CNN with a similar structure to an image classification model.

Figure 4 shows two examples of face recognition and gender and race classification from facial appearance. In this case, the system will first detect every face in an image and each facial region will then be classified separately by a model trained for face attribute classification.

5 Ethics

The explosion of data and computational power that has enabled academic and commercial advances in the study of human behavior stimulates a growing awareness of their ethical implications. Since deep learning is a result of these advances, it is also implicated in resulting ethical debates. This section focuses on five areas of concern: training data bias, privacy, informed consent, model opacity, and access to resources.

Bias. Perhaps the biggest ethical challenge facing those employing computer vision techniques is that a model will reproduce any biases in the input data, and input data often already contain racist and gendered stereotypes.[3] For a similar reason, commercial gender classification APIs offered by Microsoft, IBM, and Face++ have been criticized due to the inferior classification accuracy on darker-skin females (65%; 99% on lighter-skin males) (Buolamwini & Gebru, 2018). In image search results,

women are, on average, underrepresented relative to their participation rate in a given occupation (Lam, Wojcik, Broderick, & Hughes, 2018). A researcher relying on pre-trained models or commercial APIs should make sure he or she is aware of any biases that model imbues. When building one's own model, labels applied to a validation dataset should be examined for any biases before subsequent analysis uses the model output.

Privacy. If a model involves face detection, one may be able to identify individuals, violating their privacy. This concern is especially relevant in the study of contentious politics, as this capability means governments could engage in targeted repression by finding protesters in photographs and matching those faces to identifying information. Governments like Russia and China already deploy this technology to identify anyone in a crowd (Purdy, 2018), and some law enforcement agencies in the United States have adopted similar technology (Shaban, 2018). To protect individuals, researchers should not release photographs that could be used to identify them. Researchers should also consider whether or not their research requires identifying particular individuals at all.

User Consent. The concern about identifying individuals based on their faces segues into a third concern, informed consent. When a user makes their social media posts public, a researcher can reasonably assume that the user has provided consent to be studied, much in the same way driving on a public street provides data to traffic engineers. This assumption is more questionable for individuals who appear in images but are not the owner of the account. For example, if User A tweets a photo documenting Friends 1, 2, and 3 attending a baseball game, it is not clear that those three have consented to inclusion in a study. (It may also not be clear if User A is in the photo or not.) Researchers should seek approval from their Institutional Review Board, as they should for every project using social media data (Steinert-Threlkeld, 2018).

Interpretability. Finally, deep learning models are opaque because they are complex. AlexNet, the original convolutional neural network that launched the current renaissance in computer vision, contains 60 million parameters (Krizhevsky et al., 2012). With only eight layers, it is much simpler than current models. This opacity makes it difficult to understand what features of an image drive label classification. Understanding the internal logic of deep learning models is an active area of research. Q.-s. Zhang and Zhu (2018) and Guidotti, Monreale, and Ruggieri (2018) provide thorough reviews of current best practices.

6 Case Studies

6.1 Self-presentation of Politicians in Social Media

Social media have been widely used by politicians for political communication (Stier, Bleier, Lietz, & Strohmaier, 2018). They allow them to bypass traditional media and directly communicate with voters, redefining the relationship between political actors and editorial media (Enli, 2017). These platforms, such as Facebook and Twitter, offer various modes of interaction and generate a massive amount of data in those modalities. This example shows how an off-the-shelf classifier can generate insight about how politicians' self-representation varies by party and gender.

In the United States, politicians most commonly identify as liberal or conservative. Conservatives are more likely to accept the status quo, while liberals embrace social change (Jost, Federico, & Napier, 2009). Again speaking in broad strokes, the conservative label manifests as strong affinities toward nationalism, capitalism, and status quo political and economic institutions (Feldman & Johnston, 2014). Liberal: social change and a rejection of inequality.

In terms of images, *these ideologies should manifest as different emphasis of objects and peoples* (Kreiss, Lawrence, & McGregor, 2019). Conservative ideology should manifest via objects that serve as symbols of nationalism, freedom, and capitalism; liberal, objects that serve as symbols of inequality reduction. In terms of people, conservative politicians should be more likely to include individuals from dominant social groups; liberal politicians may include members of under-represented groups, economically disadvantaged, or protesters.

Like ideology, gender represents another axis along which politicians may vary their self-presentation. Regardless of gender, voters prefer attractive candidates (Ahler, Citrin, Dougal, & Lenz, 2017; Mattes & Milazzo, 2014). This evaluation then maps onto gender, with female candidates stereotyped as warm and men strong (Johns & Shephard, 2007). Voters in the United States reward female candidates who appear more feminine (Carpinella et al., 2016). Regardless of ideology, we therefore *expect that female candidates will emphasize physical features more than male candidates*.

Scholars have attempted to understand the visual dimension of political communication by analyzing social media images posted by politicians, though typically without incorporating advances in computer vision. For example, Towner and Muñoz (2018) manually codes the main topics covered in Instagram photographs posted by candidates in the 2016 presidential primaries and compares them with the main issues in newspapers. McGregor,

Lawrence, and Cardona (2017) qualitatively analyzes social media profile photographs and compares the activities of male and female politicians. The majority of existing studies, however, are based on either manual coding or qualitative close reading, limiting the scalability of their analysis.

Automated text analysis has been used to understand how politicians verbally present themselves. These studies are typically based on topic analysis (what is discussed), sentiment analysis (how topics are presented), or both. For example, Stier et al. (2018), by using a probabilistic topic model, show that politicians and audiences in social media focus on topics different from mass media and discuss different topics on different platforms. Sentiment analysis is also commonly applied to a large set of user posts to measure public perceptions and preferences about political leaders or parties (Nulty, Theocharis, Popa, Parnet, & Benoit, 2016) or predict their electoral success (Tumasjan, Sprenger, Sandner, & Welpe, 2011).

Computer vision techniques can generate insight about politicians' visual communication strategy (Haim & Jungblut, 2020; Xi et al., 2020), similar to how automated text analysis has illuminated verbal behaviors. To demonstrate this possibility, we use off-the-shelf commercial software offered by Google to analyze Facebook photographs posted by candidates in the 2018 U.S. general election. Google's Label Detection API is an image classifier which takes an image as input and outputs a set of labels describing its content. Using this API, one can measure actions, events, places, objects, and their attributes portrayed in politicians' photographs.

The data used in this example was collected by one of the authors for a study of candidates' social media usage and electoral success. The list of candidates who ran in the 2018 election was obtained from Wikipedia and their Facebook accounts were manually identified. For each account, the public photographs posted in timeline, mobile, and profile albums were collected using the Graph API. For the current study, we use 15,647 photographs posted by 677 candidates during the time period of August 5 - November 6, 2018. Using this dataset, we automatically detect gender and party differences in activities and attributes portrayed in the photographs using the Google Vision API.

We submitted each of 15,647 images to the API and obtained corresponding labels for each image, examples of which are shown in Figure 5. Google does not officially publish the entire list of labels their classifier can detect, and we identified 1,730 unique labels from the obtained results. [4] Since the API is an already trained model, we simply made queries using their interface and obtained the classification results.

The detected labels serve as concise semantics describing image content, allowing researchers to perform standard statistical analysis. In this example,

Hand, Blue, Glove, Safety Glove, Health Care Provider, Gesture, Meical Equipment, Medical Procedure, Scrubs, Medical Assitant, Health Care, Medical Glove, Engineering

Golfer, Golf, Sport Venue, Professional Golfer, Golf Equipment, Golf Course, Golf Club, Recration, Sports, Sports Equipment, Ball Game, Fourball, Lawn

Smile, Face, Photograph, Plant, Purple, Happy, Yellow Leisure, Fun, Social Group, Recreation, Crowd, Toddler, Event, Snapshot, Team, Child, Fan, Grass, Eyewear

Outerwear, Picture Frame, Coat, Smile, Flag, Textile, Suit, Flag of the US, Military Person, People, Official, Military Uniform Event, Military Officier, Blazer, Government, Curtain, Chair

Figure 5: Example results (input images and automatically detected labels) from Google's Label Detection API. To represent the four categories contained in Tables 1 and 2, images were taken from Joe Biden (D), Donald Trump (R), Kamala Harris (D), Liz Cheney (R).

we compare images of Democrats and Republicans by conducting a chi-squared test of the labels. To characterize such visual priorities, we measure cross-party difference for each label by comparing the number of images with and without the label. Since the two parties have different gender ratios, we perform chi-squared test on male and female images separately.

The results are shown in Table 1 (male) and Table 2 (female). The labels associated with each party are sorted in decreasing order of the chi-square statistic. A notable group of labels detected in the Republican male set are military-related concepts which relate to national security and defense. Democrats, on the other hand, feature more people, meetings, and conversations, which may show their priorities on social support. The labels that the Google Vision API returns therefore suggest that conservatives and liberals emphasize different visuals in their images.

This difference is moderated by gender, as Table 2 shows. The cross-party difference in female politicians is less obvious than that of male politicians, e.g. lack of military-related concepts in Republican female candidates. The female candidates, however, clearly emphasize their facial

Table 1 Top 20 labels associated with Democrats (left) and Republicans (right). pD and pR indicate the ratio of images containing the label for each party. Male candidates only.

Label	χ^2	*pD* (%)	*pR* (%)	Label	χ^2	*pD* (%)	*pR* (%)
Adaptation	60.6 ***	6.87	3.39	Official	53.0 ***	9.69	14.64
Community	43.9 ***	28.81	22.87	Military officer	16.6 ***	1.02	2.09
Youth	30.8 ***	6.59	4.05	Suit	16.2 ***	11.89	14.73
Conversation	23.5 ***	13.65	10.42	Uniform	15.0 ***	2.50	3.94
People	19.8 ***	16.47	13.22	Red	14.9 ***	1.25	2.34
Speech	17.1 ***	3.68	2.24	Team	14.8 ***	15.22	18.20
Orator	16.4 ***	2.22	1.15	Vehicle	12.7 ***	2.78	4.14
Sitting	14.9 ***	1.83	.90	Toddler	11.0 ***	1.46	2.43
Glasses	14.6 ***	6.41	4.61	Product	10.8 **	2.29	3.45
Smile	14.5 ***	22.95	19.74	Employment	9.6 **	12.68	14.90
Room	13.1 ***	6.89	5.12	Military person	9.5 **	.65	1.30
Head	12.9 ***	3.01	1.86	Muscle	8.9 **	1.16	1.94
Public speaking	11.5 ***	3.12	2.02	Tie	8.6 **	1.57	2.45
Face	10.4 **	3.77	2.60	Businessperson	8.0 **	11.78	13.73
Interaction	9.8 **	2.87	1.88	Fashion accessory	7.0 **	.90	1.53
Photography	8.9 **	12.03	10.10	Car	6.3 *	1.25	1.92
Human	7.6 **	2.11	1.36	Tuxedo	5.4 *	2.59	3.43
Audience	7.2 **	1.76	1.09	Formal wear	4.2 *	7.40	8.57
Forehead	7.0 **	4.03	3.01	Tourism	3.9 *	12.42	13.81
Tree	6.7 **	8.03	6.63	Management	3.6 *	3.45	4.22

*** $p<0.001$
** $p<0.01$
* $p<0.1$

Table 2 Top 20 labels associated with Democrats (left) and Republicans (right). pD and pR indicate the ratio of images containing the label for each party. Female candidates only.

Label	χ^2	p**D** (%)	p**R** (%)
Glasses	41.6 ***	9.73	4.51
Adaptation	34.3 ***	7.12	3.03
Youth	27.2 ***	11.87	7.16
Tree	12.9 ***	7.78	5.06
Shoulder	12.4 ***	3.78	1.91
Forehead	10.4 **	3.18	1.61
Chin	9.3 **	1.89	.74
Photography	9.0 **	14.87	11.80
Smile	8.7 **	35.30	31.19
Vision care	8.6 **	2.18	.99
Head	8.4 **	4.25	2.59
Public speaking	7.8 **	1.50	.56
Student	7.6 **	2.75	1.48
Performance	7.4 **	1.82	.80
Speech	6.1 *	1.71	.80
Face	5.9 *	7.25	5.44
Nose	5.8 *	1.59	.74
Friendship	5.7 *	6.10	4.45
Design	5.2 *	3.71	2.47
Community	4.7 *	31.04	28.10
Red	59.8 ***	2.23	6.30
Blond	51.1 ***	3.68	8.21
Jeans	42.7 ***	1.30	4.01
Recreation	28.7 ***	13.96	19.64
Headgear	26.9 ***	2.27	4.88
Crowd	19.7 ***	9.46	13.47
Tourism	14.6 ***	13.62	17.60
Product	11.2 ***	2.84	4.63
Event	10.8 **	59.90	64.61
Textile	10.5 **	.91	1.98
Competition event	10.0 **	1.41	2.66
Vehicle	9.6 **	2.96	4.63
Long hair	8.0 **	1.80	3.03
Team	8.0 **	17.01	20.20
Grass	6.9 **	1.34	2.35
Gesture	6.5 *	4.84	6.55
Style	6.4 *	1.14	2.04
Pink	6.4 *	2.46	3.71
Beauty	6.0 *	1.64	2.66
Sunglasses	5.7 *	1.71	2.72

*** $p<0.001$ ** $p<0.01$ * $p<0.1$

features more than their male counterparts, with the labels for Republican females emphasizing traditional feminine stereotypes (blond, long hair, pink, beauty). These two sets of results suggest that politicians optimize their self-presentation by combining partisan values and gender stereotypes (Bauer & Carpinella, 2018).

For an example of unsupervised learning using the politicians' images, sed Appendix F. That example runs k-means clustering (k=200) on the penultimate layer of a pre-trained CNN. The resulting clusters contain very similar, often identical, images, revealing common themes within an image corpus.

Using a pre-trained classifier or API is a simple yet effective way for a visual comparative analysis on an unknown domain. Researchers do not need to prepare any training data or annotations or train their own models. The key disadvantage of using an existing classifier is inferior customizability in case a researcher wants to classify concepts not defined in the classifier (or API). One solution to this situation is to train a custom classifier using annotations, which we show in the next example.

6.2 Frame Alignment During Protest

Protests are a key tactic of social movements, recruitment to protest affects the probability of success (Snow, Rochford Jr., Worden, & Benford, 1986), and how they are portrayed to bystanders ("framed") is a key input into recruitment success (Benford & Snow, 2000). This example demonstrates that Twitter users frame protests in ways likely to encourage bystanders to join.

Protesters seek to frame events to appeal to the most number of people. For example, labor organizers and the family of Mohammed Bouazizi, the Tunisian fruit vendor whose self-immolation sparked the Arab Spring, transformed his death into a parable about corruption and gender politics in a way that bridged class and geographic divides (Lim, 2013). From the other side, states portray protesters as radical, foreign, violent, or some combination thereof (Hamdy & Gomaa, 2012). This framing delegitimizes a protest, decreasing the cost a state pays if it engages in repression (Stephan & Chenoweth, 2008).

The rise of the internet and social media has empowered individuals to construct frames, weakening media and activist gatekeepers (Livingston & Bennett, 2003). A new logic of connective action now means that personal action frames are commonly invoked during social movements, as they allow individuals to connect their issues with a larger collective (Bennett & Segerberg, 2013). This ability is especially important because the primary source for framing movements, newspapers, prefers to report on violent

events (Hellmeier, Weidmann, & Geelmuyden Rød, 2018) and often have a status quo bias, causing them to frame protests differently than protesters would frame themselves (Hamdy & Gomaa, 2012).

The ability of individuals to construct and disseminate their own frames is especially important because newspaper and television emphasize protester violence (Myers & Caniglia, 2004). Media are especially likely to negatively frame events when they are seen as threatening *status quo* institutional interests, whether in democracies (Gitlin, 1980; Wittebols, 1996) or autocracies (Al-Rawi, 2015; Carter & Carter, 2019). Because protester violence decreases support for protesters (Feinberga, Willer, & Kovacheff, 2017; Stephan & Chenoweth, 2008) while state violence increases support for them (Steinert-Threlkeld, Chan, & Joo, forthcoming),the frames that individual protesters emphasize should focus more on state violence and less on protester violence. We therefore expect that *protest images shared on Twitter will frame the event as containing more state violence than protester violence*.

In addition to emphasizing state violence, individuals should prefer to frame a protest as a collective endeavor. Because the risk of protesting decreases as the size of the protest increases, bystanders are more likely to join a protest they believe is already attended by large crowds. This large crowd decreases the probability that an individual will suffer reputational cost or be the victim of state repression (Moore, 1995). Since crowds create a positive feedback loop of mobilization (Biggs, 2016), we expect that *protest images shared on Twitter will frame the event as containing crowds, not individuals*.

This subsection investigates these two expectations about framing by analyzing protests in five countries.[5] To explore which types of frames protesters choose, we first develop a pipeline to acquire geolocated tweets, extract their images, and apply deep learning models to understand scene and face features of the images. We find tweets from the five countries' protest periods, download all images from those tweets, and then apply two convolutional neural networks for **image classification** and **person attribute recognition** tasks. The implemented models are fine-tuned versions of leading CNNs.

Image classification entails identifying photos of a protest. The photos in our pipeline come from geolocated tweets. Sometimes, these tweets contain photographs; sometimes, they are from protests. When they are from protests and contain an image, we download the image. Using a mixture of Google Image results and these geolocated images, we trained a convolutional neural network to recognize protest photos. Tweets are not filtered for keywords.

For **person attribute detection**, we have developed a pipeline that identifies faces in a photo and estimates each face's sex (male or female), race (Black, East Asian, Latino, Middle Eastern, South Asian, Southeast Asian, White), and age using the Fairface classifier (Kärkkäinen & Joo, 2019). We use **image classification** to measure whether a protest image contains police or fire; whether protesters are holding signs; and the amount of violence in a protest image. The specific CNN we use is a fine-tuned Residual Network (ResNet) with 50 convolutional layers (He, Zhang, Ren, & Sun, 2016), a common architecture for image classification. For verification, see Appendix Section G; for additional training details, see Steinert-Threlkeld, Chan, and Joo Forthcoming.

We operationalize frames according to six labels. Many types of frames are chosen to normalize a protest. Protests that are peaceful or mobilize multiple types of participants often include pictures of youth or faces of the participants, the first two labels. Participants will often share images of large groups to convey that the issue being protested is not fringe, while small groups tend to convey personal action frames (Bennett & Segerberg, 2013); these two types of groups are labels three and four. Because previous literature has identified violence as a key frame (Myers & Caniglia, 2004), we also generate protester and state violence labels, the final two. Figure 6 shows sample images and their ratings for protester and state violence.

To measure which frames protesters choose, we then detect duplicate images and identify the rate of duplicate images within each label. To identify duplicate images, we take each image's last fully connected layer, a 1,000 feature vector, and measure the pairwise distance between that vector and every other image's vector. If that normalized distance is below .2, a threshold chosen from inspecting the distance histogram, two images are considered duplicates.

Table 3 provides initial support for the claims made about framing, violence, and crowds. In three of the five protests, images containing state violence are shared more. Images of groups are also shared at higher rates in three of the events, though not the same three that frame state violence. The protests framed more strongly as containing state violence (Catalonia and Venezuela) also emphasize the group nature of protests. For an example of the images driving frame alignment, see Figure 7. It shows the four most duplicated images in our sample; two use the small group frame, one uses a sign frame, a pleasant surprise because it is not a frame we expected to be prominent, and a state violence frame.

Because policy makers are more likely to respond to protests the more that protesters put forth a consistent frame (Wouters & Walgrave, 2017), higher rates of duplication may indicate episodes of greater frame alignment, both across events and within event labels (Ketelaars, Walgrave, &

Figure 6: Images and Ratings of Protester Violence and State Violence

Note: Sample images of protester (top) and state violence (bottom), with the classifier label estimate and country labeled.

Table 3 Frame Alignment by Protest Event and Label

Event	Label					
	Contains Child	Faces	Large Group	Small Group	Protester Violence	State Violence
Catalonia, Spain	.121	.071	.122	.144	.212	.304
Hong Kong	.003	.005	.024	.010	.047	.034
Russia	.027	.069	.021	.065	.121	.100
South Korea	.027	.038	.022	.039	.023	.042
Venezuela	.173	.163	.249	.267	.280	.308

Note: The first column is the region, city, or country whose geocoded protest images we analyze. The next six columns are each label. Framing – the percent of duplicate images – is calculated per event-label. Two images are duplicates if the normalized distance between their feature vectors, the output of a CNN, is closer than .2.

Wouters, 2017). For example, the divisiveness of protests in Russia may be reflected in the lower rates of duplication of state and protester violent frames in comparison to Catalonia, Spain and Venezuela. While protest success is the result of multiple factors, the ability to measure framing across countries may contribute to understanding when they succeed or fail.

These results are provisional: this example demonstrates additional understanding about protest framing that computer vision techniques can generate, but it should not be considered a definitive answer. We have suggested one way of measuring framing, but future work should explore

Figure 7: The four most common images causing frame alignment in our sample. The top row, and the two most shared, are from Venezuela and use a small group frame. The bottom left image is from South Korea and uses a frame, sign usage, not shown in Table 3. The bottom right is from Catalonia and is an example of the state violence frame.

other operationalizations such as number of tweets containing a label (instead of percentage) and expand the frames considered. This analysis also discards temporal variation, which is almost certainly an important determinant of when certain frames receive emphasis. Which frames receive emphasis may also be affected by city and country correlates that we do not consider.

The results in Table 3 reveal interesting variation warranting further exploration. Across events, the most obvious difference is that each event exhibits different baseline amounts of framing intensity. For example, Venezuela contains the highest framing intensity (duplication rate) across all labels, and the rank correlation of events across labels is quite high. The relative rates of duplication, moreover, vary significantly: the most duplicated event-label, Venezuela state-violence, resonates almost 103 times as much as the least, photos from Hong Kong with children. Two possibilities are that frame alignment increases the more violent state repression is or as social media penetration increases, increasing the rewards to frame alignment. Within each event, violent images are duplicated the

most, with images of state violence shared more in 3 events, protester violence in 2, and a tie in Russia. Individuals also prefer to frame protests in terms of groups as opposed to individuals, as evidenced by the higher duplication rates in the group labels versus the child and faces label. That the rank ordering of frames within events appears to correlate across events suggests a hierarchy of protest frames, suggesting that forces beyond just the presence of professional organizations also affect frame alignment (Ketelaars et al., 2017).

7 Conclusion

If a picture is worth 1,000 words, then it would require approximately two kilobytes of storage (Jagenstedt, 2008). Images from consumer cell phones and digital cameras, however, require at least three megabytes of storage, usually more. Even images shared on social media platforms, which are compressed from their original size, require hundreds of kilobytes of space. A picture, in other words, is worth anywhere from 50,000 (100 kilobytes) to 1,500,000 words (3 megabytes). A picture is actually worth a book.[6]

This paper has argued that recent advances in computer vision, deep convolutional neural networks, hold much promise for the study of politics. Analyzing them in large quantities can inform research in behavior, communication, development, and conflict. The paper then introduced deep learning methods and how to validate model output. These techniques are especially promising for the study of protest, and an example analyzes six protests. The use of large, passively collected datasets raises new ethical issues of which researchers should be aware, especially when the data are images.

The increasing prevalence of digital technology has led to a greater appreciation of the importance of images in political life. Images make arguments, set agendas, document and dramatize events, activate emotions, shape perceptions, build identity, generate social cohesion, build empathy, and strategically create ambiguity (Schill, 2012). Whereas pedagogy, communication, and academic analysis have traditionally focused on acquiring textual information, cheap computing means that individuals consume and produce increasing amounts of visual information (Kraidy, 2002). Images are key drivers of political phenomena, and we would do well to take advantage of new techniques to analyze them in large quantities in research.

Appendix A Comparing Text to Images

Visual data differ from text data in ways summarized in Table A1. The most critical distinction between them is that, since words are the units of meaning in texts and are easier to define than objects in images, it is easier to process text than images.

An image's constituent elements, pixels, carry no meaning, as opposed to text data whose atomic elements are words. In other words, texts contain less uncertainty about meaning than images, and the five differences in Table A1 flow from that distinction.

More technically, a text's characters, including spaces, are its atomic elements. A string of characters is more meaningful than a sequence of pixels, however, because human language provides predefined sequences - words - that people learn. People do not learn pixels, and there are not visual languages that codify collections of pixels the same way words codify collections of characters. It is very easy to describe to a computer a text building block: it is any sequence of characters bounded by a punctuation or space character. Word detection is therefore equivalent to object detection in images. A single word can provide a great deal of semantic information (*e.g.*, "Trump" or "election") and a simple string comparison operation allows one to access the information. In contrast, one pixel, and even small groups of pixels, are meaningless. In visual analysis, one has to process a huge number of meaningless pixels to detect and identify people, objects and events. Recognizing elementary content, visual "words," from an image is, however, extremely difficult. This technical difficulty has been the main obstacle to research involving quantitative analysis of visual data on a large scale.

It is also easier to build meaning from a collection of words than from pixels because words are arranged in one dimension, whereas pixels spread across two. The simplest text models take a bag-of-words approach, where the order of words does not matter; while more complex models perform better, bag-of-words models are nonetheless useful. A bag-of-pixels model would fail, however, since each pixel is meaningless. Visual models therefore need to identify groups of pixels. Groups are identified using sliding windows, and these windows vary in two dimensions. The size of the window therefore becomes two parameters open to the researcher to manipulate. While varying the window dimensions is equivalent to choosing how many words to concatenate in an n-gram model, meaning dissipates quickly the further away words are, meaning researchers do not have to consider large sequences of words. There is no equivalent for pixels and meaning (though there are rules of thumb), especially because the number of pixels representing objects will vary depending on the resolution of the image.

Table A1 Distinct Characteristics between Text and Image Data

Text	Image
• Low uncertainty at word level	• High uncertainty at any level
• One dimensional: a sequence of words	• Two dimensional: an array of pixels
• Pre-defined dictionary of words, ngrams, or emojis	• Unknown dictionary
• Small file size	• Large file size
• Language specific	• Universal language

It is difficult to detect objects in images because visual object dictionaries do not exist. They do not exist because visual languages do not exist. For example, the word "trump" can be a verb, adjective, or proper noun. While its meaning is not as clear as a word with only one usage, it can nonetheless easily be inferred based on nearby words such as "opponent", "card", or "President". While an image of President Trump is immediately recognizable to humans, it is not to computers. A white pixel surrounded by other white pixels could be a dress shirt, or it could be a part of a flag. Brown pixels separated from other brown pixels by 100 other pixels could be two eyes, but they could also be two shoes or two coffee cups. Because there is no easy definition of objects in images, it is harder to infer meaning from images than text.

Because words have clearer meaning than pixels, text files require less space than images. For example, images in tweets require, on average, 100 kilobytes of storage space. A tweet cannot contain more than 240 characters, which requires .24 kilobytes of space. A tweet of 100 kilobytes could contain 100,000 characters. The smaller size of texts means they are easier than images to store, share for replication, and, most importantly, analyze.

Because there is not a universal verbal language, object detection in images is more universal than meaning detection in texts. For example, the vast majority of faces contain two eyes, two ears, a nose, mouth, and forehead. The words for these facial features, however, vary across languages. An image classifier to detect faces therefore is more likely to detect all faces than a text classifier trained on one language, such as English, will

be to detect facial words in another language's text. The lack of structure to images at the pixel level is therefore a blessing and a curse: it is a curse because building and training image classifiers is harder than for text, but it is a blessing because an image classifier is more broadly applicable than a text one.

Appendix B Applications

Development

Socioeconomic Status Surveys. Any research question that requires, or would benefit from, socioeconomic characteristics where the household, neighborhood, or city is the unit of analysis would benefit from training a deep learning model on satellite imagery data. Image data can measure different features of cities, such as the distribution of building types, as well as land use in rural areas (Jensen and Cowen, 1999). Imagery with a resolution of one meter or smaller can provide data on socioeconomic characteristics as they vary by neighborhood, allowing for frequent census-like data creation, an ability especially useful in countries with no, or irregular, censuses (Tapiador, Avelar, Tavares-Correa and Zah, 2011). For agricultural areas, it can measure changes in rainfall and crop growth, proximate measures of income for many countries (Toté, Patricio, Boogaard, van der Wijngaart, Tarnavsky and Funk, 2015). Since income shocks are a precursor to civil conflict, data that accurately measure subnational changes in income could act as an early warning system (Hsiang et al., 2013).

It is possible to measure socioeconomic variables using photographs of places taken by people. Manual analysis of Google Street View (GSV) imagery shows that photographs of streets correlates strongly with survey based measures of neighborhood attributes (Odgers, Caspi, Bates, Sampson and Moffitt, 2012; Wilson, Kelly, Schootman, Baker, Banerjee, Clennin and Miller, 2012). A model trained on GSV images recovers income by block in New York City (Glaeser, Kominers, Luca and Naik, 2018), and a deep learning model of cars in GSV images can measure income, race, and education at the precinct level (Gebru, Krause, Wang, Chen, Deng, Aiden and Fei-Fei, 2017). Another promising approach is to pay people to take photographs of specific phenomena, such as the price of goods at a supermarket or the prevalence of anti-incumbent signs at a protest (Premise Data, 2017). Paying people to capture images is especially useful in areas with otherwise insufficient publicly available data.

Natural Disasters. Image data also provide access to temporal changes in local regions. For example, a model that accurately recovers built features of

towns and cities could provide insight into how institutions affect recovery from natural disasters. If images exist of the same area immediately before and after a natural disaster, the physical and geographic extent of damage as well as the speed and amount of recovery may be measurable. These dependent variables may then be related to various institutional ones. Recovery may occur more quickly in democracies than non-democracies or in countries with free media, for example. In democracies, subnational variation could depend on whether a disaster strikes a powerful politician's district or if there is an impending election.

Contentious Politics

Civil War. Using computer vision, greed and grievance can be measured with more geographic and temporal precision (Collier and Hoeffler, 2004; Kern, 2011). Those two concepts are notoriously difficult to operationalize, and researchers rely on imperfect measures such as the availability of natural resources (greed) or aggregate economic statistics such as gross domestic product (economic grievance). For example, greed is measurable using the precise outline of diamond mines, virgin forests, or oil deposits, and their depletion can be observed from satellite data or resource maps (Hunziker and Cederman, 2017). Grievance is reflected in city-level variation in economic activity measurable using light emissions (Weidmann and Schutte, 2017). Whether these measures are better than existing datasets will depend on the dataset and country on which the researcher is focused.

State Capacity. Images can also be used to measure state capacity. Humans-as-sensors can take photographs of specific objects, such as prices in markets (to measure inflation), road conditions, or school conditions, using smart phones (Premise Data, 2017). These images can give disaggregated information about a state's ability to repress intranational conflict, as well as the ability of rebels to attack the state. Maps are also images, and digitizing them can provide historical data on state capacity, especially power projection, that current measures, such as GDP, may not capture (Hunziker, Müller-Crepon and Cederman, 2018).

Protests. Image data can create improved measures of a protest's violence and features of participants. Existing datasets measure protester or state violence coarsely, as an ordinal variable, because of interpretive difficulty from relying on text. Images can generate continuous estimates of how violent protesters or the state are. Text also makes it difficult to understand exactly who is protesting because that detail is rarely reported; surveys provide participant demographics but are very costly and require foreknowledge of a protest. When protest images contain faces, traits such

as protesters' age, race, and gender become measurable, and protest size can be estimated by summing the number of faces. For examples of this approach, see (Won, Steinert-Threlkeld, and Joo, 2017), (Zhang & Pan, 2019), and (Steinert-Threlkeld, Chan, and Joo Forthcoming).

Appendix C Artificial Neural Network Detail

Artificial Neural Network

While there have been different models proposed in the deep learning literature, artificial deep neural networks (DNN) are the most popular branch of deep models and have been used in a number of areas including computer vision, audio processing, natural language processing, robotics, bioengineering, and medicine. This subsection describes a general neural network, and Section C discusses its variant, a convolutional neural network (CNN). The convolutional neural network is commonly used in computer vision applications with two-dimensional inputs.

Artificial neural networks represent complex concepts, like the probability an image contains a human face, as a system of connections between elementary *nodes*; the collection of nodes and connections is the neural network. Each node, also called a *neuron* or an *unit*, in this system only performs simple computations and interacts with a few other nodes. Nevertheless, the network of a large number of nodes enables complex data modeling through their interactions.

Figure C1 shows an example configuration of a node and its connected nodes. Each node takes input values from nodes in the proceeding layer and evaluates a weighted sum using weights associated with edges (in this example, $1\ 0.7 + 0.5 - 0.3 + 0.3\ 1.0 = 0.85$). Typically this value is transformed by a non-linear activation function, *e.g.*, sigmoid or rectified linear unit (ReLU), and then passed to output node. For example, these input values might be an individual's values for gender (x_1), race (x_2), or income (x_3) and the output variable might be political ideology.

Figure C2 shows an example architecture of a neural network with several layers. Neural networks with multiple hidden layers are considered "deep." A *layer* in a neural network is a set of nodes which takes inputs from the nodes in the previous layer and deliver outputs to the nodes in the next layer. When a network is visualized as Figure C2, a column of nodes is a layer, and the number of columns is the number of layers. Inputs to the whole network therefore undergo several steps of transformations through

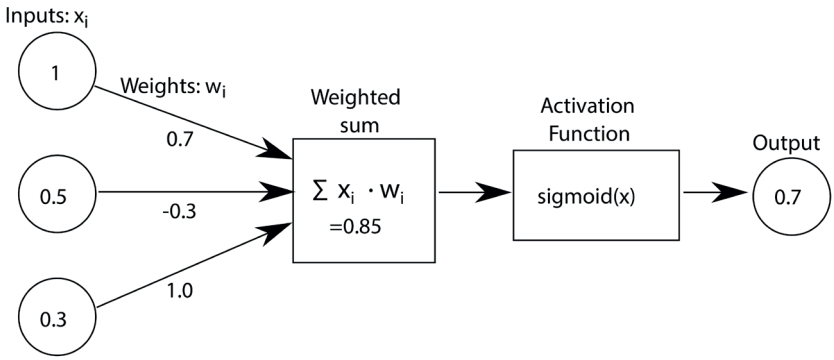


Figure C1: An example computation in a node and its connected nodes.

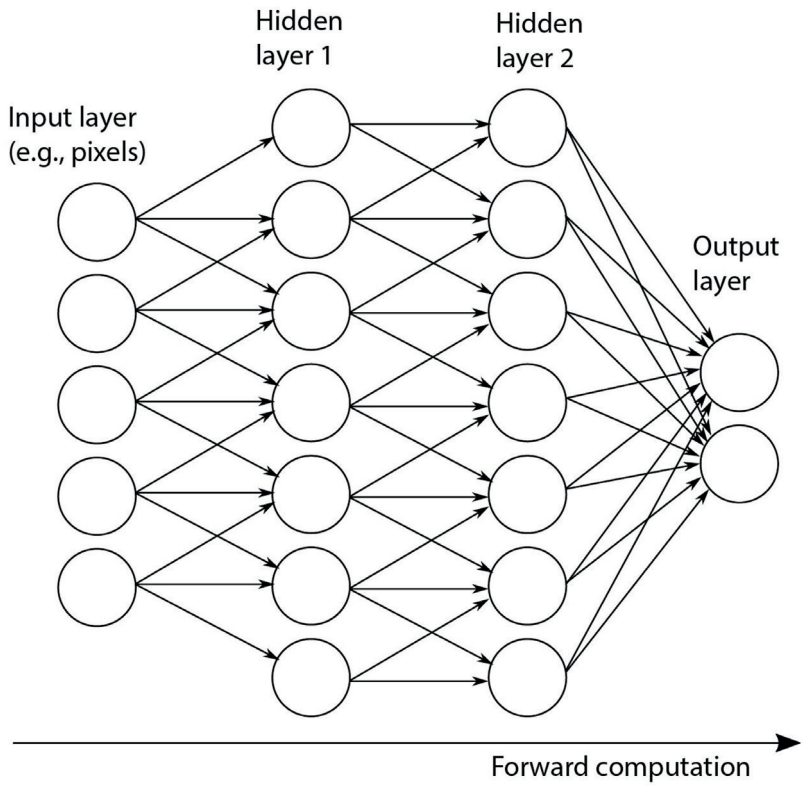


Figure C2: An example architecture of a neural network with an input layer, an output layer, and two hidden layers.

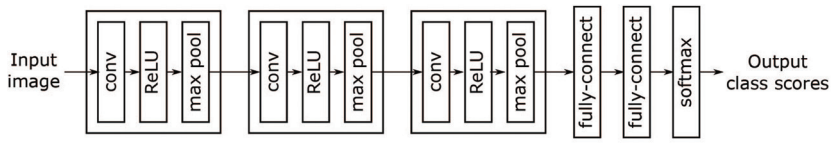


Figure C3: *An example of a convolutional neural network architecture.*

layers until they reach the output layer of the network. The output layer is the network's final layer, and it contains one node per desired label in case of classification.

Hidden layers are intermediate layers between the input and output layers in a network whose true values are not observed during training. They play a critical role in modeling complex concepts by giving an expressive power to deeper networks. Studies have shown, both experimentally and theoretically, that the more layers a neural network has, the better performance it can achieve (Eldan & Shamir, 2016; Poggio, Mhaskar, Rosasco, Miranda and Liao, 2017). A drawback of having too many layers is that it is more difficult to train such a model, *i.e.*, vanishing gradients (Bengio, Simard and Frasconi, 1994).[7]

Convolutional Neural Network

Figure C3 illustrates an example configuration of a typical convolutional neural network (CNN) for classification with 12 layers.[8] LeCun et al. (1989) first proposed the CNN structure with an efficient learning algorithm based on backpropagation. Since Krizhevsky et al. (2012) showed that deep CNNs (CNNs with many layers) improved image recognition by 21.9%-33.8%, it has become the *de facto* standard method for image classification. CNNs have a repetitive structure with several important layers: the convolutional layer, nonlinear layer (ReLU, in Figure C3), pooling layer, and fully connected layer. This subsection describes each in turn.

Convolutional layer

A convolutional layer in CNNs performs a smoothing operation ("convolution") to the input to the layer, which is either raw image data or an output from the previous layer. Convolution is widely used in signal processing for transforming or comparing time series data. For example, one can reduce noise in an audio signal by convolving it with a Gaussian filter, which will smooth out the original signal by blending the original value at time t with other values at adjacent time points around t.

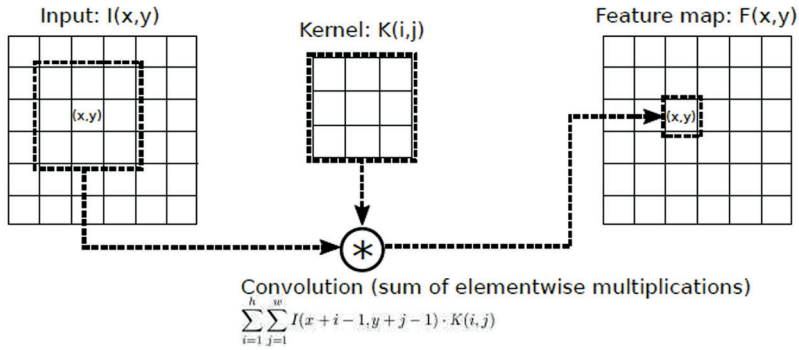


Figure C4: Illustration of computations in a convolutional layer.

Formally, the convolution of two functions, f and g, is another function defined by

$$(f^* g)(t) = \int f(x) \cdot g(t - x)dx \cdot \quad \text{(C3)}$$

The second function, $g(t)$, is called a *kernel*. Note that the kernel is flipped ($g(t - x)$) by the definition of convolution. In a discrete case, convolution computes the sum of element-wise multiplication between two functions, with one function being shifted over time, such that:

$$(f^* g)(t) = \sum_x f(x) \cdot g(t - x) \cdot \quad \text{(C4)}$$

Each convolutional layer in a CNN uses a convolution operation in order to compare the input data with the kernels (also called *filters* in the deep learning literature) in the model. In practice, the kernel is not flipped in computation in most implementations as it is unnecessary for the purpose of CNN.[9] Not flipping the kernel creates a slightly modified definition of convolution of a two-dimensional input I and a two-dimensional kernel K in CNN:

$$F(x,y) = (I * K)(x,y) \stackrel{def}{=} \sum_{i=1}^{h}\sum_{j=1}^{w} I(x + i - 1, y + j - 1) \cdot K(i,j). \quad \text{(C3)}$$

$I(x, y)$ and $K(x, y)$ denote the element in x th row and y th column in the matrices I and K. h and w denote the height and width of the kernel K, and, typically, CNNs use square kernels ($h = w$). The result of the convolution is another 2D array, F, which is called a *feature map*. The feature map is the output of the convolutional layer, and it is the same dimensions as the

input data. This computation is performed on every location in an input map and the result is stored in the same location in the output feature map (See Figure C4).

Most images are three-dimensional data with two spatial dimensions and an additional dimension of color (e.g., RGB). Feature maps in each layer are therefore also three-dimensional as each individual feature map (also called a channel) corresponds to the response from a specific kernel (filter). Each filter describes a specific pattern to be detected from an input from the previous layer. The entire weight parameters of each convolutional layer (K) are therefore represented by a four-dimensional array of size (w, h, m, n), where m is the number of channels of the input (the number of channels in the previous convolutional layer) and n is the number of channels in the current layer. The number of channels (feature maps) in each layer is arbitrary and typically ranges from 32 to 1024, except the color channel (3). The feature map for each channel will therefore be obtained as follows:

$$F(x, y, c') = \sum_{c=1}^{m} \sum_{i=1}^{h} \sum_{j=1}^{w} I(x + i - 1, y + j - 1, c) \cdot K(i, j, c, c'). \quad \text{(C4)}$$

Convolutional layers enable the following two key properties of convolutional neural networks.

Weight sharing. In Equation C4, the kernel is invariant to the location of each input node (x, y). Therefore, the same kernel will apply to every location of the input map, and the connections between two layers (input and output nodes of each convolutional layer) will share the same weights. Weight sharing is effective because an object may appear in any location of an image and its appearance is invariant to its placement. Weight sharing reduces the number of free parameters in the network and makes it easier to train.

Local and sparse connectivity. Convolutional layers in CNN achieve sparse connectivity by using a kernel much smaller than the size of input map ($h, w < 10$, usually). Each node in a convolutional layer is only connected to a small number of nodes in the previous layer, *i.e.*, a local region. This kernel is small because adjacent pixels and subregions of an image are more highly correlated than distant regions.

Nonlinear Layer

Each convolutional layer is typically followed by a nonlinear activation function that applies to each element in the feature map. One of the most common activation functions is the rectified linear unit (ReLU):

$$f(x) = \max(0, x). \quad \text{(C5)}$$

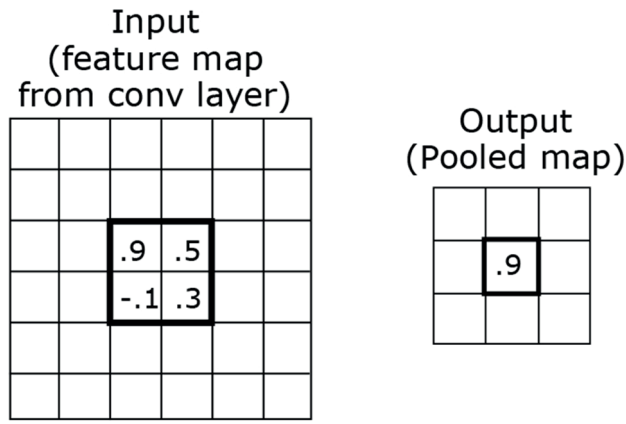


Figure C5: *Illustration of a max-pooling operation of the window size* 2 × 2. *For each window, only the maximum value will be retained*.

This function will simply replace negative feature map values with 0 and keep positive values. Other functions such as sigmoid or hyperbolic tangent function can be also used. The main advantage of the ReLU is that it runs much faster than those functions.

Nonlinearity of visual models is important as it allows to capture a complex data distribution. Visual data, projections to 2D space, are highly nonlinear due to many factors such as occlusion, object deformation, and camera exposure saturation. Human visual systems are capable of processing this nonlinearity. Especially, nonlinear layers are essential in deep networks because consecutive layers of linear operations collapse into one linear layer. Thus, there will be no benefit of adding more layers to the network without nonlinear functions.

Pooling layer

Pooling is another important operation in convolutional neural networks since it reduces computational complexity. A pooling layer takes an input feature map from the previous layer and generates a transformed map whose size differs from its input size. Most images and feature maps in a CNN are spatially correlated: values in closer pixels or nodes[10] tend to be more similar than those far away. Instead of keeping similar values redundantly from adjacent locations, one can simply choose the maximum response (or the average value) in each spatial neighborhood (pooling window) to represent the area.

Specifically, a max pooling layer compares values in each sub window (e.g., a 2 × 2 window of pixels) of the input feature map and chooses the maximum value (see Figure C5). Only these maximum values will be stored in the output

map; the other values are disregarded. Removing non-maximum values also means that the resulting feature map will be of a smaller size than the input map. For example, an input image of size 256 × 256 will be downsampled to 16 × 16 after applying 4 max-pooling layers of size 2 × 2. During the process, the information originally encoded in the spatial dimension in images will be translated into the non-spatial dimension in the feature map, *e.g.*, 16 × 16 × 1,024.

One main difficulty in visual learning is high geometric variations of objects and parts arising from part movements and viewpoint changes. Pooling not only reduces the number of free trainable parameters but also helps the network achieve translation invariance, which is an important property for computer vision systems. Robust computer vision system needs to handle such geometric variations, and pooling operations help by disregarding small spatial perturbations within the pooling window.

Fully connected layer

CNN architectures used for classification include one or a few fully connected layers at the last stage. A fully connected layer densely ("fully") connects all the nodes from the previous layer to all the nodes in the current layer. A convolutional layer encodes local information tied to specific image subregions distributed over a two dimensional map, through sparse connectivity (*i.e.*, nodes are selectively connected in a convolutional layer). A fully connected layer collects local features from all the subregions, captured in the prior convolutional layer, and outputs the overall likelihood of a visual concept (label).

In the case of classification, the fully connected layer(s) in a CNN are usually followed by a softmax function, which normalizes the final classification scores over categories. This procedure is the same as multinomial logistic regression.

Appendix D Training and Validation

This section discusses practical issues in training a model and introduces tools to diagnose the model performance. For technical details of training and validation, see Section C in the appendix. The appendix also provides precise definitions of technical concepts, such as weights, kernels, or loss functions, and their computations in greater detail.

Training

New Models. As in other machine learning methods, training a new model means using training data (labeled images) to estimate optimal values for model parameters.

Training a neural network means finding optimal values for weights in the model (see Figure C1). In most cases, objective functions of neural networks are non-convex and cannot be directly optimized, and training is conducted by a gradient descent method with the backpropagation algorithm (LeCun et al., 1989), alternating between forward and backward passes.[11] In the "forward" pass, given an input value, the network evaluates the output and computes the loss function based on the ground truth output value, *i.e.* the image's labels or class. In the "backward" pass, the gradient of the loss function is propagated backward by the chain rule and model weights are updated accordingly. Backpropagation is necessary because neural networks have nested structures, so layers and weights (parameters) in lower (earlier) layers are not directly connected to the output variables where the gradients are first computed. See LeCun et al. (1989) for detail.

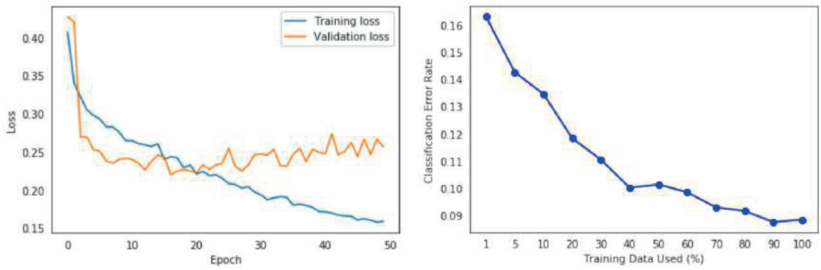
There exist many types of loss functions. One can use a specific loss function or a combination of multiple loss functions depending on the task (classification, detection, or face recognition) and the output dimension (number of variables). In image classification, for example, the most popular loss function is cross-entropy loss, also called log-loss. In a binary classification task, the binary cross-entropy loss is:

$$loss_{bce}(y, \hat{y}) = -(y \cdot \log\hat{y} + (1 - y) \cdot \log(1 - \hat{y})) \quad \text{(D1)}$$

where $y \in \{0, 1\}$ is the true label for the example and $\hat{y} \in (0, 1)$ is the output value computed from the model. In training, all the model parameters are optimized to minimize this loss function across the entire training set. Other loss functions can be also used in other tasks. For example, mean square error loss can be used to estimate continuous outputs such as age.

Figure D1 (a) shows an example of the evolution of a model loss over iterations. Note that, after the 20th epoch,[12] the model performance is saturated and the validation loss starts increasing although the training loss continues to decrease.[13] This degradation arises because the model is fitted too much to the training set. One can stop training at that point and take the final model. Using more training data can help avoid overfitting and train a better model (See Figure D1(b)).

Pre-Trained Models. Deep learning usually requires a large amount of training data (1.28 million images in ImageNet (Russakovsky et al., 2015)) to be successful. It is usually not feasible for an individual researcher to collect such a large training set or training a model to exploit those images'



(a) Loss trend over iterations (b) Effect of training data size

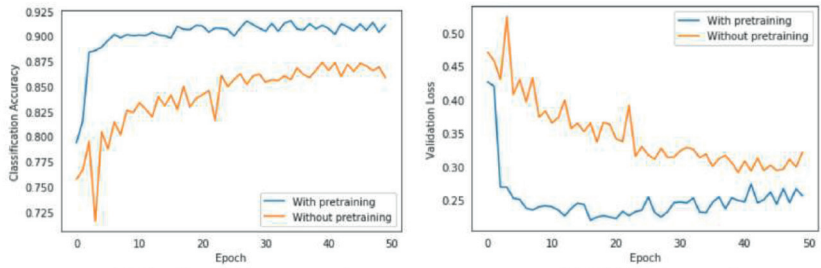
Figure D1: (a) The changes of training and validation losses over iterations. One epoch is equivalent to using every image in each set once. (b) The effect of the training set size on the model accuracy (100% = 32,611 images). See Section 6.2 for the details of the data and model.

complexity. One method of overcoming the requirement is to use models trained for another task with a larger dataset and apply to the current task for which only a small amount of data is available. This is known as transfer learning and is the process we recommend others follow.

Instead of using random values like when training a new model, one can take the weight values from an existing model (the pre-trained model) and initiate a new training process. This procedure is called fine-tuning, or transfer learning, as an existing model is tuned to another task. For example, one may use a model trained for generic image classification to initialize the weight values of a new model for human activity detection (Won et al., 2017). Figure D2 illustrates the advantage of using a pre-trained model: it achieves a better classification accuracy and reduces training time compared to making a new model.

Transfer learning works because CNNs, especially in their lower layers, capture features that generalize to other related tasks. In visual learning, these sharable representations include elementary features such as edges, color, or some simple textures. Since these features can commonly apply to many visual tasks, one can reuse what has been already trained from a large amount of training data and refine the model to the new data. Doing so saves significant time and hardware costs.

There exist many pre-trained models which are widely adopted as baselines for fine-tuning, such as AlexNet (Krizhevsky et al., 2012), Places365 (Zhou, Lapedriza, Khosla, Oliva, & Torralba, 2017), and VGG-Face (Cao, Shen, Xie, Parkhi, & Zisserman, 2018). As their names suggest, these models are trained from data in specific domains. Other examples include Residual Net (ResNet) (He et al., 2016) and VGG-Net (Simonyan & Zisserman, 2014)



(a) Classification accuracy trend (b) Validation loss trend

Figure D2: The effect of fine-tuning (using a pre-trained model) in training. (One epoch is equivalent to using every image in each set once.) See Section 6.2 for the details of the data and model.

for image classification and Faster-RCNN (Ren, He, Girshick, & Sun, 2015) and YOLO (Redmon, Divvala, Girshick, & Farhadi, 2016) for object detection. Therefore, one can choose a model pre-trained for a task and domain related to the researcher's question.

Using one of these pre-trained models facilitates topic discovery. By taking the last fully connected layer or the softmax layer of images run through a classifier, one can find similar images using any preferred clustering algorithm. The images in the clusters will contain similar features (pictures of John McCain, for example), suggesting they are about the same topic. Appendix Section F shows this approach to topic discovery using politicians' images shared on Facebook and k-means clustering.

Whether using transfer learning or making a new model, it is critical to ensure that the training data represent a diverse and balanced set of images before they are annotated so that recall is high for each desired label. For example, if one wants to collect images to be used for training a protest event classifier, the set should contain enough protest images and non-protest images. This task may not be trivial if the target event infrequently occurs. If the task is well defined and clearly explainable by simple statements, one can crowdsource the annotation task using online services, such as Amazon Mechanical Turk. If an annotation task requires more expertise, one should hire and directly supervise annotators.

Architecture and Hyperparameters. Table D1 shows how different model architectures and depths affect accuracy. The evaluation for fine-tuning on our own data (32,611 training images) shows relatively small differences.[14] When the training set is larger, deeper models tend to perform better. The ImageNet challenge offers 1,281,167 training images (Russakovsky et al., 2015), so the performance gap is wider. It is beyond the scope of this

Table D1 Performance comparison for different models. See Section sm:Protest for the details of the data for fine-tuning.

Architecture	Depth	Fine-tune (Protest)		Imagenet Validation Error (%)	Number of Parameters
		Best Loss	Best Accuracy (%)		
Alexnet	8	0.249	89.29	16.6	60 M
VGG	11	0.223	90.52	10.4	133 M
VGG	16	0.204	92.33	7.2	138 M
VGG	19	0.197	92.16	7.1	144 M
ResNet	18	0.220	91.54	-	11.7 M
ResNet	34	0.213	91.65	5.60	21.8 M
ResNet	50	0.213	91.79	5.25	25.6 M

paper to discuss at length how to optimize the architecture of a CNN to be used (number of layers in a model or types of regularization to be used), preprocessing, best optimization methods, and other hyperparameters. In general, these are empirical questions and the optimal solution varies by task.

Validation and Interpretation

Deep neural networks often receive criticism due to the lack of interpretability of their results and internal mechanisms compared to simple models with a handful of explanatory variables. A deep model typically comprises millions of parameters (see Table D1), and it is impossible to identify their meanings or roles from the classifier output.

One method of validation is to use a validation dataset which does not overlap with the training set. As in other classification problems, the accuracy of a CNN-based classifier can be measured by several metrics, including raw accuracy, precision and recall, or average precision, among others. These measures, however, do not explain *how* the model achieves its results.

Language-based Interpretation

Just as humans use language to explain a concept, one can develop a joint model that incorporates visual and textual data such that the text part explains its visual counterpart. For example, image captioning generates a sentence describing visual content in an input image (Kiros, Salakhutdinov,

& Zemel, 2014) or text-based justifications to explain why the model produces particular outputs (Hendricks et al., 2016).

Another line of research on text-based interpretation of visual learning utilizes questioning and answering (Antol et al., 2015). Such methods take both an image and a text question as input and output a text-based answer to the input question. This allows a more flexible interface between a user and a model than a traditional classification task, which essentially asks a fixed question to the model.

The key limitation of these methods is that they do not generalize: they are unable to deal with novel content or questions. The models are trained on image-text pairs and simply reproduce the mapping learned from the training data. When the model is given a novel question which was not given during training, it will not understand the meaning of the question.

Visual Validation

Another method of understanding how a deep network produces its output is through visualization. Since convolutional neural networks are largely used for visual learning from images, visual validation is especially effective. We introduce the two most popular approaches: feature-based and region-based.

Figure D3 provides examples of the feature-based approach, using a random sample of images from ImageNet. This approach uses a "deconvolutional" network (Zeiler & Fergus, 2014), which is akin to a reverse CNN. Figure D3 shows that visually similar image patches that contain the same image feature (left sub-panel) will trigger high activation scores in the same node in the network that captures the image feature. The image feature can be visually identified from the feature activation maps (right sub-panel). Moreover, this visualization also confirms that the lower layers in a network respond to the low level visual features such as color or texture, and the higher layers capture more structured and semantically meaningful shapes ("face", "web").

The region-based approach is exemplified by Gradient-weighted Class Activation Mapping (Grad-CAM) (Selvaraju et al., 2017). Grad-CAM highlights pixels in images based on how much they contribute to the final output of the model. See Figure D4 for an example visualization using this paper's protester framing example. Grad-CAM can confirm that the model was able to learn meaningful features such as "smoke" to model the concept of "violence".

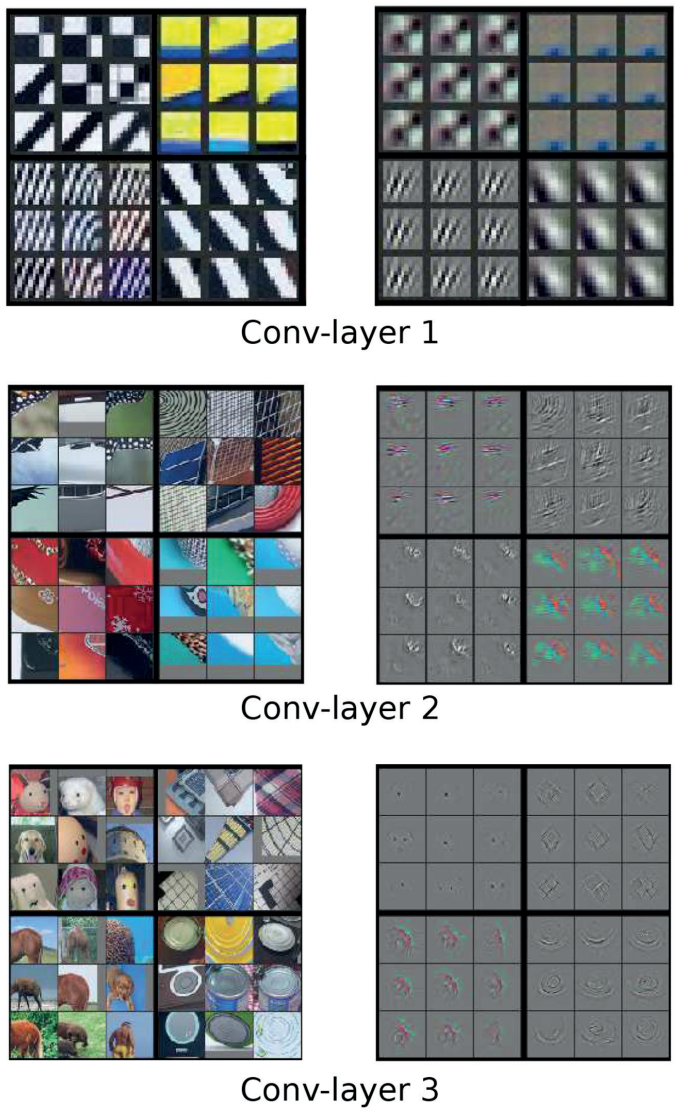


Figure D3: Visualization of feature activations at different layers in a CNN by a deconvolutional network (Zeiler & Fergus, 2014). For each layer, the left panel shows groups of similar image patches which produce high activation values for the same node in the layer. The right panel shows corresponding feature visualizations. That the patches become more recognizable as the layer depth increases confirms lower layers capture low level features and higher layers capture more structured and semantically meaningful patterns. Each layer shows four randomly chosen filters, and the filters are not the same across layers.

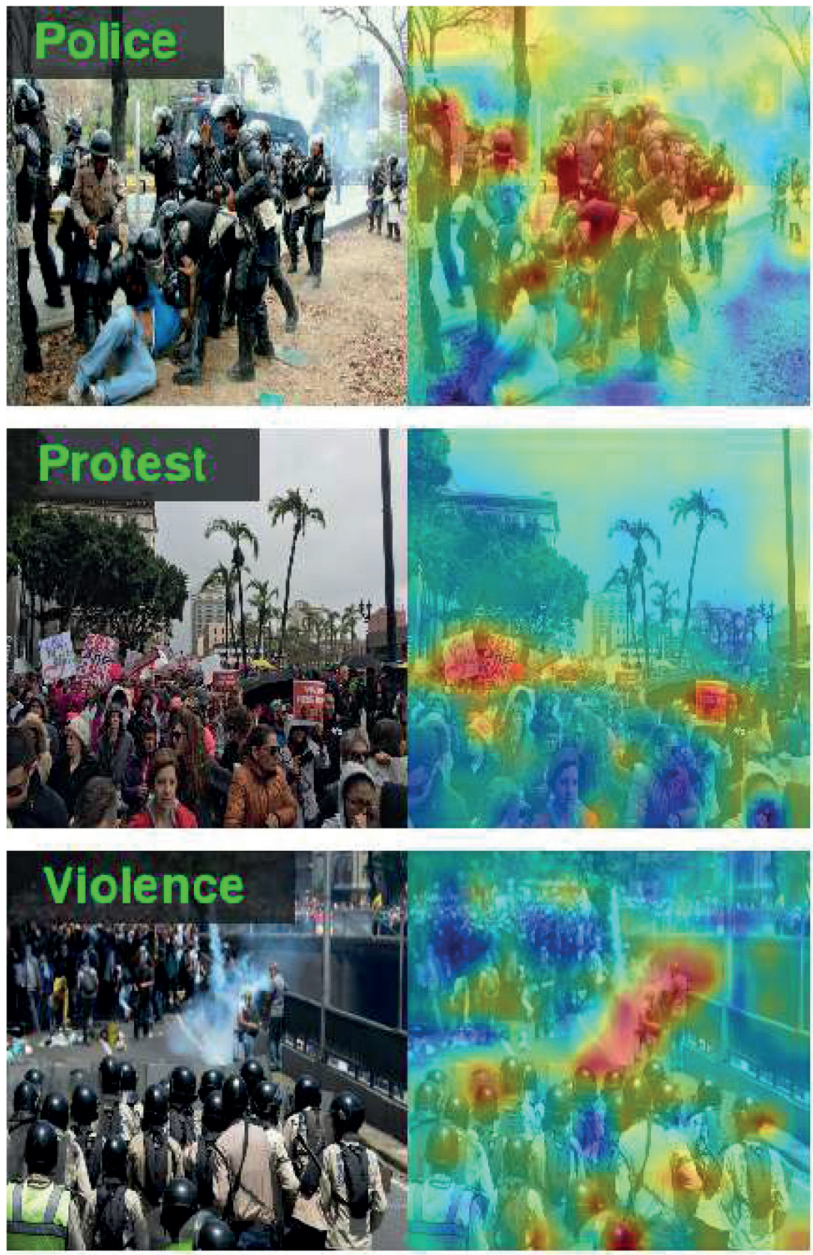


Figure D4: Visualization of Pixel Salience of CNN by Grad-CAM. It highlights image sub-regions which more contribute to the classification output (pixels closer to red). These labels are chosen from the protest example that follows.

Appendix E Software Libraries

There exist many open-source or commercial libraries and tools that researchers can use for visual content analysis in their projects. Compared to software for text analysis, these libraries are in general larger and have more complex internal structures, which are required to provide various image processing functionalities. Fortunately, there are a small number of standardized, popular libraries that can be adopted for computer vision and deep learning projects, which will be briefly reviewed in this section.

- OpenCV and dlib are currently the most popular computer vision libraries. They offer a wide range of basic image processing, computer vision, and machine learning functionalities. Python is best for OpenCV, though there is a light wrapper in R for it. dlib is accessible via R and Python libraries.
- TensorFlow, PyTorch, Keras, MxNET, and Microsoft Cognitive Toolkit are the most popular deep learning frameworks, as of 2019. These libraries allow researchers to define custom network architectures and train the network with their own data. For high-level use cases, these libraries have little practical difference between them.
- In case researchers simply want to use existing classifiers which are already trained without developing a model themselves, they can also use commercial services through APIs. These options include Google Vision API, Microsoft Vision API, Face++, and Amazon Rekognition. These services return submitted images with labels.

Appendix F Self-presentation of Politicians in Social Media

One can also use an inductive approach by clustering a given set of images without any annotations or labels. Figure F1 shows example clusters obtained from images in the same dataset, not using the Google Vision labels. Specifically, we first computed generic image features using an image embedding from a CNN pre-trained on ImageNet. We ran the model on each image and obtained a numeric vector of length 2,048 from the activation values of the second-to-last layer of the CNN. Then we ran K-means clustering (K = 200) on these features.

By grouping similar images, one can identify clusters showing various activities and events which politicians attend to. A cluster of John McCain (the last cluster in Figure F1) arises as many politicians posted his photographs after his death on August 25, 2018. Clustering analysis is an effective

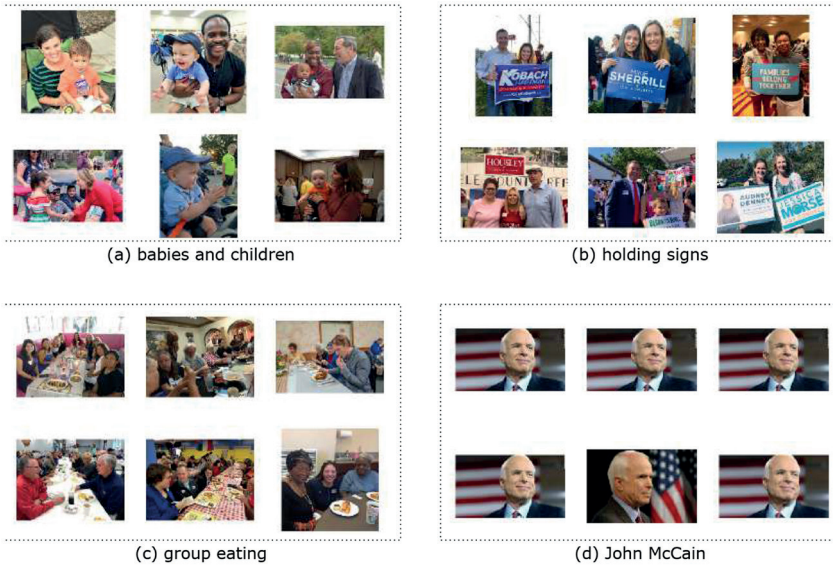


Figure F1: Example clusters found in Facebook photographs posted by candidates.

way of discovering issues or topics which may be unknown to researchers prior to analysis. This example of unsupervised learning is very similar to unsupervised topic modeling in text analysis.

Appendix G Individuals' Framing of Protest

Pipeline Detail

Verification

Figure D4 visualizes the internal mechanism of the model by showing which features contribute to an image label. It uses the amount of output gradient backpropagated to internal nodes and corresponding image subregions and shows how strongly the nodes are activated in classification (Selvaraju et al., 2017). This process is similar to the regular model training procedure. The closer to red the area of an image, the more it contributes to the classifier output.

Figure D4 shows that the classifier is driven by parts of an image that a human would recognize as important for each category. For example, the protest label primarily activates on signs. Tear gas and police helmets drive the violence classifier, while a child's face, but not the nearby adults', drive the children classifier.

Notes

1. A layer is a separate operation or a collection of internal nodes placed at the same stage in a network. It will be further elaborated shortly. The Supplementary Materials discuss different types of layers in neural networks.
2. Google has not published details of the Vision API's architecture, though it is safe to assume that it is based on a CNN. It is concerning that users are not informed about these details. We discuss these issues, for example, model biases and interpretability, in the Ethics section. We recommend this API provided that researchers are aware of potential issues and validate these APIs for their purposes (Section Appendix D), e.g. by measuring the accuracy of the API with manual annotations.
3. This concern is a fancy rephrasing of the old adage, "Garbage in, garbage out."
4. Most labels are straightforward to comprehend except a few such as "Adaptation" which we believe refers to "screen adaptation" and correlates with people and crowd.
5. The first is secessionist protests in Catalonia, Spain. The second is the 2014 Hong Kong protests against changes to Hong Kong's electoral system seen as contradicting the "One Country, Two Systems" relationship with China. The third is anti-corruption protests in Russia on March 26, 2017. The fourth is the 2016-2017 protests in South Korea against President Park Geun-hye. Revelations in October 2016 that President Guen-hye received council from a Rasputin-like figure triggered large protests, and those protests persisted through her impeachment on March 10, 2017. The fifth is protests in Venezuela in 2014 and 2015.
6. This estimate is poetic. Another way to think of images is that they have high entropy, meaning they cannot be compressed as much as text. The greater size of images reflects this greater difficulty of compressing them, not necessarily a true quantum of information.
7. Networks are trained by a gradient descent method with backpropagation, and the gradients become smaller as it goes back through more layers, making it difficult to update the parameters.
8. In practice, only convolutional and fully connected layers are usually counted to specify the number of layers in a model. The example model in Figure C3 can be called a 5-layer CNN.
9. The parameters will be learned in the same way irrespective of the flipping direction.
10. Pixels in an input image are the nodes in the first input layer.
11. Minimizing the sum of squares of residuals is convex and directly optimizable, for example.
12. One epoch is one pass over all training images.
13. In training, available data is typically split into a training set and a validation set. Only the training set is used in actual model training and the validation set is only used for evaluation.

14. VGG-Net (Simonyan & Zisserman, 2014), despite having fewer layers, performs slightly better than deeper models. Note that it has more parameters.

References

Ahler, D. J., Citrin, J., Dougal, M. C., & Lenz, G. S. (2017). Face Value? Experimental Evidence that Candidate Appearance Influences Electoral Choice. *Political Behavior*, 39 (1), 77–102. http://doi:10.1007/s11109-016-9348-6

Al-Rawi, A. K. (2015, mar). Sectarianism and the Arab Spring: Framing the popular protests in Bahrain. *Global Media and Communication*, 11 (1), 25–42. Retrieved from http://gmc.sagepub.com/cgi/doi/10.1177/1742766515573550 http://doi:10.1177/1742766515573550

Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Lawrence Zitnick, C., & Parikh, D. (2015). Vqa: Visual question answering. In *International conference on computer vision* (pp. 2425–2433).

Atkinson, M. D., Enos, R. D., & Hill, S. J. (2009). Candidate Faces and Election Outcomes: Is the Face–Vote Correlation Caused by Candidate Selection? *Quarterly Journal of Political Science*, 4, 229–249. http://doi:10.1561/100.00008062

Barrett, A. W., & Barrington, L. W. (2005). Is a picture worth a thousand words? newspaper photographs and voter evaluations of political candidates. *Harvard International Journal of Press/Politics*, 10 (4), 98–113.

Barry, A. M. S. (1997). *Visual Intelligence: Perception, Image, and Manipulation in Visual Communication*. SUNY Press.

Bauer, N. M., & Carpinella, C. (2018). Visual information and candidate evaluations: the influence of feminine and masculine images on support for female candidates. *Political Research Quarterly*, 71 (2), 395–407.

Baum, M. A., & Groeling, T. (2008). New Media and the Polarization of American Political Discourse. *Political Communication*, 25 (4), 345–365. http://doi:10.1080/10584600802426965

Baumeister, R. F., Bratslavsky, E., & Vohs, K. D. (2001). Bad Is Stronger Than Good. *Review of General Psychology*, 5 (4), 323–370. http://doi:10.1037//1089-2680.5.4.323

Belhumeur, P. N., Hespanha, J. P., & Kriegman, D. J. (1997). Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. *IEEE Transactions on pattern analysis and machine intelligence*, 19 (7), 711–720.

Benford, R. D., & Snow, D. A. (2000). Framing Processes and Social Movements: An Overview and Assessment. *Annual Review of Sociology*, 26, 611–639.

Bennett, W. L., & Segerberg, A. (2013). *The Logic of Connective Action* (No. June 2013). Cambridge: Cambridge University Press.

Biggs, M. (2016). Size Matters: Quantifying Protest by Counting Participants. *Sociological Methods & Research*, 1–33. http://doi:10.1177/0049124116629166

Blumler, J. G., & Kavanagh, D. (1999). The Third Age of Political Communication: >Influences and Features. *Political Communication*, 16 (3), 209–230. http://doi:10.1080/105846099198596

Bonica, A. (2018). Inferring Roll-Call Scores from Campaign Contributions Using Supervised Machine Learning. *American Journal of Political Science*, 62 (4), 830–848. http://doi:10.1111/ajps.12376

Brader, T., Valentino, N. A., & Suhay, E. (2008). What Triggers Public Opposition to Immigration? Anxiety, Group Cues, and Immigration Threat. *American Journal of Political Science*, 52 (4), 959–978. http://doi:10.1111/j.1540-5907.2008.00353.x

Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency* (pp. 77–91).

Cantu, F. (2019). The Fingerprints of Fraud: Evidence From Mexico's 1988 Presidential Election. *American Political Science Review*, 113 (3), 710–726. http://doi:10.13140/RG.2.2.34763.49442

Cao, Q., Shen, L., Xie, W., Parkhi, O. M., & Zisserman, A. (2018). Vggface2: A dataset for recognising faces across pose and age. In *2018 13th ieee international conference on automatic face & gesture recognition (fg 2018)* (pp. 67–74).

Carpinella, C. M., Hehman, E., Freeman, J. B., & Johnson, K. L. (2016). The Gendered Face of Partisan Politics: Consequences of Facial Sex Typicality for Vote Choice. *Political Communication*, 33 (1), 21–38. http://doi:10.1080/10584609.2014.958260

Carter, E. B., & Carter, B. L. (2019). Propaganda and Protest in Autocracies. *Journal of Conflict Resolution*.

Casas, A., & Webb Williams, N. (2019). Images That Matter: Online Protests and the Mobilizing Role of Pictures. *Political Research Quarterly*, 72 (2), 360–375. http://doi:10.2139/ssrn.2832805

Chen, D., Park, K., & Joo, J. (2020). Understanding gender stereotypes and electoral success from visual self-presentations of politicians in social media. *In Joint workshop on aesthetic and technical quality assessment of multimedia and media analytics for societal trends* (pp. 21–25).

D'Alessio, D., & Allen, M. (2000). Media bias in presidential elections: A meta-analysis. *Journal of communication*, 50 (4), 133–156.

Delalleau, O., & Bengio, Y. (2011). Shallow vs. deep sum-product networks. In *Advances in neural information processing systems* (pp. 666–674).

Dietrich, B. J. (2018). Using Motion Detection to Measure Social Polarization in the U.S. House of Representatives. *Working Paper*. Retrieved from http://www.brycejdietrich.com/files/working{_}papers/Dietrich{_}cspan.pdf

Druckman, J. N., & Parkin, M. (2005). The impact of media bias: How editorial slant affects voters. *The Journal of Politics*, 67 (4), 1030–1049.

Eldan, R., & Shamir, O. (2016). The power of depth for feedforward neural networks. In *Conference on learning theory* (pp. 907–940).

Enli, G. (2017). Twitter as arena for the authentic outsider: exploring the social media campaigns of trump and clinton in the 2016 us presidential election. *European journal of communication*, 32 (1), 50–61.

Feinberga, M., Willer, R., & Kovacheff, C. (2017). Extreme Protest Tactics Reduce Popular Support for Social Movements. *Working paper*, 1–58. Retrieved from https://papers.ssrn.com/sol3/papers.cfm?abstract{_}id=2911177 http://doi:10.1007/s10551-015-2769-z.For

Feldman, S., & Johnston, C. (2014). Understanding the determinants of political ideology: Implications of structural complexity. *Political Psychology*, 35 (3), 337–358.

Fenno, R. F. (1978). *Home style: House members in their districts*. Pearson College Division.

Fleming, M. K., & Cottrell, G. W. (1990). Categorization of faces using unsupervised feature extraction. In *Ijcnn* (pp. 65–70).

Gazzaniga, M. S. (1998). *The Mind's Past*. University of California Press.

Geise, S., & Baden, C. (2015). Putting the image back into the frame: Modeling the linkage between visual communication and frame-processing theory. *Communication Theory*, 25 (1), 46–69.

Gentzkow, M., & Shapiro, J. M. (2010). What drives media slant? Evidence from U.S. Daily Newspapers. *Econometrica*, 78 (1), 35–71. http://doi:10.3982/ECTA7195

Gibson, R., & Zillmann, D. (2000). Reading between the photographs: The influence of incidental pictorial information on issue perception. *Journalism & Mass Communication Quarterly*, 77 (2), 355–366.

Gilliam Jr, F. D., & Iyengar, S. (2000). Prime suspects: The influence of local television news on the viewing public. *American Journal of Political Science*, 560–573.

Gitlin, T. (1980). *The Whole World Is Watching: Mass Media in the Making and Unmaking of the New Left, With a New Preface*. University of California Press.

Grabe, M. E., & Bucy, E. P. (2009). *Image bite politics: News and the visual framing of elections*. Oxford University Press.

Graber, D. A. (1996). Say It with Pictures. *The ANNALS of the American Academy of Political and Social Science*, 546 (1), 85–96. http://doi:10.1080/0885819020952880 4

Griffiths, T. L., Jordan, M. I., Tenenbaum, J. B., & Blei, D. M. (2004). Hierarchical topic models and the nested chinese restaurant process. In *Advances in neural information processing systems* (pp. 17–24).

Grimmer, J. (2010). A bayesian hierarchical topic model for political texts: Measuring expressed agendas in senate press releases. *Political Analysis*, 18 (1), 1–35.

Groeling, T., Joo, J., Li, W., & Steen, F. (2016). Visualizing presidential elections. In *Annual meeting of the american political science association*.

Guidotti, R., Monreale, A., & Ruggieri, S. (2018). A Survey of Methods for Explaining Black Box Models. *ACM Computing Surveys*, 51 (5), 93:1–93:42.

Haim, M., & Jungblut, M. (2020). Politicians' self-depiction and their news portrayal: Evidence from 28 countries using visual computational analysis. *Political Communication*, 1–20.

Hamdy, N., & Gomaa, E. H. (2012, apr). Framing the Egyptian Uprising in Arabic Language Newspapers and Social Media. *Journal of Communication*, 62 (2), 195–211. http://doi:10.1111/j.1460-2466.2012.01637.x

Hansen, L. (2015). How images make world politics: International icons and the case of abu ghraib. *Review of International Studies*, 41 (2), 263–288.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 770–778).

Hellmeier, S., Weidmann, N. B., & Geelmuyden Rød, E. (2018). In The Spotlight:Analyzing Sequential Attention Effects in Protest Reporting. *Political Communication*, 00 (00), 1–25. http://doi:10.1080/10584609.2018.1452811

Hendricks, L. A., Akata, Z., Rohrbach, M., Donahue, J., Schiele, B., & Darrell, T. (2016). Generating visual explanations. In *European conference on computer vision* (pp. 3–19).

Hsiang, S. M., Burke, M., & Miguel, E. (2013, sep). Quantifying the influence of climate on human conflict. *Science*, 341 (6151), 1235367. http://doi:10.1126/science.1235367

Jagenstedt, P. (2008). *How much a thousand words are worth*. Retrieved 2018-04-15, from https://blog.foolip.org/2008/05/17/how-much-a-thousand-words-are-worth/

Johns, R., & Shephard, M. (2007). Gender, Candidate Image and Electoral Preference. *British Journal of Politics and International Relations*, 9 (3), 434–460. http://doi:10.1111/j.1467-856X.2006.00263.x

Joo, J., Bucy, E., & Seidel, C. (2019). Automated coding of televised leader displays: Detecting nonverbal political behavior with computer vision and deep learning. *International Journal of Communication*, 13, 4044-4066.

Joo, J., Steen, F. F., & Zhu, S.-C. (2015). Automated facial trait judgment and election outcome prediction: Social dimensions of face. In *International conference on computer vision* (pp. 3712–3720).

Joo, J., Li, W., Steen, F. F., & Zhu, S. C. (2014). Visual persuasion: Inferring communicative intents of images. *In Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 216–223).

Jost, J. T., Federico, C. M., & Napier, J. L. (2009). Political ideology: Its structure, functions, and elective affinities. *Annual review of psychology*, 60, 307–337.

Kanade, T. (1977, January). Computer recognition of human faces. *Interdisciplinary Systems Research*, 47, 1–47.

Kang, Z., Indudhara, C., Mahorker, K., Bucy, E. P., & Joo, J. (2020, August). Understanding political communication styles in televised debates via body movements. *In European Conference on Computer Vision* (pp. 788–793). Springer, Cham.

Kargar, S., & Rauchfleisch, A. (2019). State-aligned trolling in Iran and the double-edged affordances of Instagram. *New Media & Society*, 1–22. http://doi:10.1177/1461444818825133

Kärkkäinen, K., & Joo, J. (2019). Fairface: Face attribute dataset for balanced race, gender, and age. *arXiv preprint arXiv:1908.04913*.

Ketelaars, P., Walgrave, S., & Wouters, R. (2017). Protesters on message? Explaining demonstrators' differential degrees of frame alignment. *Social Movement Studies*, 16 (3), 340–354. Retrieved from http://dx.doi.org/10.1080/14742837.2017.1280387 http://doi:10.1080/14742837.2017.1280387

Kiros, R., Salakhutdinov, R., & Zemel, R. (2014). Multimodal neural language models. In *International conference on machine learning* (pp. 595–603).

Kraidy, U. (2002). Digital Media and Education: cognitive impact of information visualization. *Journal of Educational Media*, 27 (3), 95–106. http://doi:10.1080/1358165022000081369

Kreiss, D., Lawrence, R. G., & McGregor, S. C. (2019). Political identity-ownership: Symbolic contests to represent members of the public. *Working Paper*.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems* (pp. 1097–1105).

Lam, O., Wojcik, S., Broderick, B., & Hughes, A. (2018). *Gender and Jobs in Online Image Searches* (Tech. Rep.). Pew Research Center.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *nature*, 521 (7553), 436–444.

LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., & Jackel, L. D. (1989). Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1 (4), 541–551.

Lenz, G. S., & Lawsom, C. (2011). Looking the Part: Television Leads Less Informed Citizens to Vote Based on Candidates' Appearance. *American Journal of Political Science*, 55 (3), 574–589. http://doi:10.1111/j.1540-5907.2011.00511.x

Lim, M. (2013, mar). Framing Bouazizi: 'White lies', hybrid network, and collective/connective action in the 2010-11 Tunisian uprising. *Journalism*, 14 (7), 921–941. http://doi:10.1177/1464884913478359

Liu, Z., Luo, P., Wang, X., & Tang, X. (2015). Deep learning face attributes in the wild. In *International conference on computer vision* (pp. 3730–3738).

Livingston, S., & Bennett, W. L. (2003). Gatekeeping, Indexing, and Live-Event News: Is Technology Altering the Construction of News? Gatekeeping ,

Indexing , and Live-Eve. *Political Communication*, 20 (4), 363–380. http://doi:10.1080/10584600390244121

Marcus, G. E., Neuman, W. R., & MacKuen, M. (2000). *Affective intelligence and political judgment*. Chicago: University of Chicago Press.

Mattes, K., & Milazzo, C. (2014). Pretty faces, marginal races: Predicting election outcomes using trait assessments of British parliamentary candidates. *Electoral Studies*, 34, 177–189. http://doi:10.1016/j.electstud.2013.11.004

McGregor, S. C., Lawrence, R. G., & Cardona, A. (2017). Personalization, gender, and social media: Gubernatorial candidates' social media strategies. *Information, communication & society*, 20 (2), 264–283.

Moore, W. H. (1995, jun). Rational Rebels: Overcoming the Free-Rider Problem. *Political Research Quarterly*, 48 (2), 417–454. http://doi:10.1177/106591299504800211

Myers, D. J., & Caniglia, B. S. (2004). All the Rioting That's Fit to Print: Selection Effects in National Newspaper Coverage of Civil Disorders, 1968-1969. *American Sociological Review*, 69, 519–543.

Nulty, P., Theocharis, Y., Popa, S. A., Parnet, O., & Benoit, K. (2016). Social media and political communication in the 2014 elections to the european parliament. *Electoral studies*, 44, 429–444.

Peng, Y. (2018). Same candidates, different faces: Uncovering media bias in visual portrayals of presidential candidates with computer vision. *Journal of Communication*, 68 (5), 920–941.

Popkin, S. L. (1994). *The Reasoning Voter: Communiation and Persuasion in Presidential Campaigns*. Chicago: University of Chicago Press.

Purdy, C. (2018). *China is launching a dystopian program to monitor citizens in Beijing*. Retrieved 12.20.2018, from https://qz.com/1473966/china-is-starting-a-big-brother-monitoring-program-in-beijing/

Ranjan, R., Patel, V. M., & Chellappa, R. (2017). Hyperface: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41 (1), 121–135.

Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 779–788).

Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems* (pp. 91–99).

Roberts, M. E., Stewart, B. M., Tingley, D., Lucas, C., Leder-Luis, J., Gadarian, S. K., Rand, D. G. (2014). Structural topic models for open-ended survey responses. *American Journal of Political Science*, 58 (4), 1064–1082.

Rojas, M., Masip, D., Todorov, A., & Vitria, J. (2011). Automatic prediction of facial trait judgments: Appearance vs. structural models. *PloS one*, 6 (8), e23323.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., ... others (2015). Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115 (3), 211–252.

Schill, D. (2012). The visual image and the political image: A review of visual communication research in the field of political communication. *Review of Communication*, 12 (2), 118–142.

Schmuck, D., & Matthes, J. (2017). Effects of Economic and Symbolic Threat Appeals in Right-Wing Populist Advertising on Anti-Immigrant Attitudes: The Impact of Textual and Visual Appeals. *Political Communication*, 34 (4), 607–626. http://doi:10.1080/10584609.2017.1316807

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Computer vision and pattern recognition* (pp. 618–626).

Shaban, H. (2018, jun). *Amazon employees demand company cut ties with ICE*. Retrieved from https://www.washingtonpost.com/news/the-switch/wp/2018/06/22/amazon-employees-demand-company-cut-ties-with-ice

Shah, D. V., Hanna, A., Bucy, E. P., Lassen, D. S., Van Thomme, J., Bialik, K., ... Pevehouse, J. C. (2016). Dual screening during presidential debates: Political nonverbals and the volume and valence of online expression. *American Behavioral Scientist*, 60 (14), 1816–1843.

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Snow, D. A., Rochford Jr., E. B., Worden, S. K., & Benford, R. D. (1986). Frame Alignment Processes, Micromobilization, and Movement Participation. *American Sociological Review*, 51 (4), 464–481.

Soroka, S., Loewen, P., Fournier, P., & Rubenson, D. (2016). The impact of news photos on support for military action. *Political Communication*, 33 (4), 563–582.

Steinert-Threlkeld, Z. C. (2018). *Twitter as Data*. Cambridge University Press.

Steinert-Threlkeld, Z. C.,Chan, A., & Joo J. Forthcoming. "How State and Protest Violence Affect Protest Dynamics". DOI: 10.1086/715600. Available at: https://www.journals.uchicago.edu/doi/pdf/10.1086/715600

Stephan, M. J., & Chenoweth, E. (2008). Why Civil Resistance Works. *International Security*, 33 (1), 7–7.

Stier, S., Bleier, A., Lietz, H., & Strohmaier, M. (2018). Election campaigning on social media: Politicians, audiences, and the mediation of political communication on facebook and twitter. *Political communication*, 35 (1), 50–74.

Stovall, J. G. (1988). Coverage of 1984 presidential campaign. *Journalism Quarterly*, 65 (2), 443–449.

Sullivan, D. G., & Masters, R. D. (1988). "happy warriors": Leaders' facial displays, viewers' emotions, and political support. *American Journal of Political Science*, 345–368.

Taigman, Y., Yang, M., Ranzato, M., & Wolf, L. (2014). Deepface: Closing the gap to human-level performance in face verification. In *Computer vision and pattern recognition* (pp. 1701–1708).

Todorov, A., Mandisodza, A. N., Goren, A., & Hall, C. C. (2005). Inferences of competence from faces predict election outcomes. *Science*, 308 (5728), 1623–1626.

Torres, M. (2018). *Give me the full picture: Using computer vision to understand visual frames and political communication*. Retrieved from http://qssi.psu.edu/new-faces-papers-2018/

Towner, T. L., & Muñoz, C. L. (2018). Picture perfect? the role of instagram in issue agenda setting during the 2016 presidential primary campaign. *Social science computer review*, 36 (4), 484–499.

Tukachinsky, R., Mastro, D., & King, A. (2011). Is a Picture Worth a Thousand Words? The Effect of Race-Related Visual and Verbal Exemplars on Attitudes and Support for Social Policies. *Mass Communication and Society*, 14 (6), 720–742. http://doi:10.1080/15205436.2010.530385

Tumasjan, A., Sprenger, T. O., Sandner, P. G., & Welpe, I. M. (2011). Election forecasts with twitter: How 140 characters reflect the political landscape. *Social science computer review*, 29 (4), 402–418.

Uijlings, J. R., Van De Sande, K. E., Gevers, T., & Smeulders, A. W. (2013). Selective search for object recognition. *International journal of computer vision*, 104 (2), 154–171.

Valentino, N. A., Brader, T., Groenendyk, E. W., Gregorowicz, K., & Hutchings, V. L. (2011). Election Night's Alright for Fighting: The Role of Emotions in Political Participation. *Journal of Politics*, 73 (1), 156–170. http://doi:10.1017/S0022381610000939

Vernon, R. J., Sutherland, C. A., Young, A. W., & Hartley, T. (2014). Modeling first impressions from highly variable facial images. *Proceedings of the National Academy of Sciences*, 111 (32), E3353–E3361.

Viola, P., & Jones, M. J. (2004). Robust real-time face detection. *International journal of computer vision*, 57 (2), 137–154.

Watts, M. D., Domke, D., Shah, D. V., & Fan, D. P. (1999). Elite cues and media bias in presidential campaigns: Explaining public perceptions of a liberal press. *Communication Research*, 26 (2), 144–175.

Wittebols, J. H. (1996). News from the Noninstitutional world: U.S. and Canadian television news coverage of social protest. *Political Communicationı*, 13 (3), 345–361. http://doi:10.1080/10584609.1996.9963122

Won, D., Steinert-Threlkeld, Z. C., & Joo, J. (2017). Protest activity detection and perceived violence estimation from social media images. In *Proceedings of the 2017 acm on multimedia conference* (pp. 786–794).

Wouters, R., & Walgrave, S. (2017). Demonstrating Power: How Protest Persuades Political Representatives. *American Sociological Review*, 82 (2), 361–383. http://doi:10.1177/0003122417690325

Xi, N., Ma, D., Liou, M., Steinert-Threlkeld, Z. C., Anastasopoulos, J., & Joo, J. (2020). Understanding the political ideology of legislators from social media images. In *Proceedings of the international aaai conference on web and social media* (Vol. 14, pp. 726–737).

You, Q., Cao, L., Cong, Y., Zhang, X., & Luo, J. (2015). A multifaceted approach to social multimedia-based prediction of elections. *IEEE Transactions on Multimedia*, 17 (12), 2271–2280.

Zebrowitz, L. A., & Montepare, J. M. (2008). Social psychological face perception: Why appearance matters. *Social and personality psychology compass*, 2 (3), 1497–1517.

Zeiler, M. D., & Fergus, R. (2014). Visualizing and understanding convolutional networks. In *European conference on computer vision* (pp. 818–833).

Zhang, H., & Pan, J. (2019). CASM: A Deep-Learning Approach for Identifying Collective Action Events with Text and Image Data from Social Media. *Sociological Methodology*, 49, 1–48.

Zhang, Q.-s., & Zhu, S.-C. (2018). Visual interpretability for deep learning: a survey. *Frontiers of Information Technology & Electrical Engineering*, 19 (1), 27–39.

Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., & Torralba, A. (2017). Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40 (6), 1452–1464.

Appendix References

Bengio, Yoshua, Patrice Simard and Paolo Frasconi. 1994. "Learning long-term dependencies with gradient descent is difficult." *IEEE transactions on neural networks* 5(2):157–166.

Collier, Paul and Anke Hoeffler. 2004. "Greed and grievance in civil war." *Oxford Economic Papers* 56:563–595.

Eldan, Ronen and Ohad Shamir. 2016. The power of depth for feedforward neural networks. In *Conference on Learning Theory*. pp. 907–940.

Gebru, Timnit, Jonathan Krause, Yilun Wang, Duyun Chen, Jia Deng, Erez Lieberman Aiden and Li Fei-Fei. 2017. "Using deep learning and Google Street View to

estimate the demographic makeup of neighborhoods across the United States." *Proceedings of the National Academy of Sciences* 114(50):13108–13113.

Glaeser, Edward L, Scott Duke Kominers, Michael Luca and Nikhil Naik. 2018. "Big data and big cities: The promises and limitations of improved measures of urban life." *Economic Inquiry* 56(1):114–137.

He, Kaiming, Xiangyu Zhang, Shaoqing Ren and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. pp. 770–778.

Hsiang, Solomon M, Marshall Burke and Edward Miguel. 2013. "Quantifying the influence of climate on human conflict." *Science* 341(>6151), 1235367.

Hunziker, Philipp, Carl Müller-Crepon and Lars-Erik Cederman. 2018. "Roads to Rule, Roads to Rebel: Relational State Capacity and Conflict in Africa.". **URL:** http://roads-to-peace.org/PDF/DIIS%20UNOPS%202017%20Roads%20to%20Peace%20report.pdf

Hunziker, Philipp and Lars-Erik Cederman. 2017. "No Extraction without Representation: The Ethno-Regional Oil Curse and Secessionist Conflict." *Journal of Peace Research* 54(3):365–381.

Jensen, Jr and Dc Cowen. 1999. "Remote sensing of urban suburban infrastructure and socio-economic attributes." *Photogrammetric Engineering and Remote Sensing* 65(5):611–622.

Kern, Holger Lutz. 2011. "Foreign Media and Protest Diffusion in Authoritarian Regimes: The Case of the 1989 East German Revolution." *Comparative Political Studies* 44(9):1179–1205.

Krizhevsky, Alex, Ilya Sutskever and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*. pp. 1097–1105.

LeCun, Yann, Bernhard Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne Hubbard and Lawrence D Jackel. 1989. "Backpropagation applied to handwritten zip code recognition." *Neural computation* 1(4):541–551.

Odgers, Candice L., Avshalom Caspi, Christopher J. Bates, Robert J. Sampson and Terrie E. Moffitt. 2012. "Systematic social observation of children's neighborhoods using Google Street View: a reliable and cost-effective method." *The Journal of Child Psychology and Psychiatry* 53(10):1009–1017.

Poggio, Tomaso, Hrushikesh Mhaskar, Lorenzo Rosasco, Brando Miranda and Qianli Liao. 2017. "Why and when can deep-but not shallow-networks avoid the curse of dimensionality: A review." *International Journal of Automation and Computing* 14(5):503–519.

Premise Data. 2017. "Premise Data.". **URL:** www.premise.com

Selvaraju, Ramprasaath R, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh and Dhruv Batra. 2017. Grad-CAM: Visual Explanations From

Deep Networks via Gradient-Based Localization. In *Computer Vision and Pattern Recognition*. pp. 618–626.

Tapiador, Francisco J., Silvania Avelar, Carlos Tavares-Correa and Rainer Zah. 2011. "Deriving fine-scale socioeconomic information of urban areas using very high-resolution satellite imagery." *International Journal of Remote Sensing* 32(21):6437–6456.

Toté, Carolien, Domingos Patricio, Hendrik Boogaard, Raymond van der Wijngaart, Elena Tarnavsky and Chris Funk. 2015. "Evaluation of satellite rainfall estimates for drought and flood monitoring in Mozambique." *Remote Sensing* 7(2):1758–1776.

Weidmann, Nils B and Sebastian Schutte. 2017. "Using night light emissions for the prediction of local wealth." *Journal of Peace Research* 54(2):125–140.

Wilson, Jeffrey S, Cheryl M Kelly, Mario Schootman, Elizabeth A Baker, Aniruddha Banerjee, Morgan Clennin and Douglas K Miller. 2012. "Assessing the built environment using omnidirectional imagery." *American journal of preventive medicine* 42(2):193–199.

Won, Donghyeon, Zachary C Steinert-Threlkeld and Jungseock Joo. 2017. Protest Activity Detection and Perceived Violence Estimation from Social Media Images. In *Proceedings of the 2017 ACM on Multimedia Conference*. ACM pp. 786–794.

Zachary C. Steinert-Threlkeld, Alexander Chan, and Jungseock Joo. Forthcoming. "How State and Protest Violence Affect Protest Dynamics". DOI: 10.1086/715600. Available at: https://www.journals.uchicago.edu/doi/pdf/10.1086/715600

Zhang, Han and Jennifer Pan. 2019. "CASM: A Deep-Learning Approach for Identifying Collective Action Events with Text and Image Data from Social Media." *Sociological Methodology* 49:1–48.