

On the conditions for integrating deep learning into the study of visual politics

Matteo Magnani*
matteo.magnani@it.uu.se
InfoLab, Uppsala University
Uppsala, Sweden

Alexandra Segerberg*
alexandra.segerberg@statsvet.uu.se
Dept. of Government, Uppsala University
Uppsala, Sweden

ABSTRACT

Traditional methods to study visual politics have been limited in geographical, media and temporal coverage. Recent advances in deep learning have the potential to dramatically extend the scope of the field especially with respect to making sense of the contemporary social and political developments in digital media. While some early adopters may be tempted to take the new computational tools at face value, others see their black-box character as cause for concern. This paper argues that the integration of deep learning into the study of visual politics must be approached still more critically and boldly. On the one hand, the complexity of visual political themes requires a more substantial human involvement if compared with other applications of deep neural networks. Therefore, a question is how the scientist and the network should best interact. On the other hand, it is important to acknowledge that a deep learning tool will never simply replace specific tasks inside a research process: its adoption has implications for the broader process from the delineation of the object of analysis, to data collection, to the interpretation and communication of results. We examine the conditions of integrating a deep learning tool for image classification into the large-scale study of visual politics in digital and social media along these two dimensions.

KEYWORDS

visual politics, image analysis, deep learning, social science

ACM Reference Format:

Matteo Magnani and Alexandra Segerberg. 2021. On the conditions for integrating deep learning into the study of visual politics. In *13th ACM Web Science Conference 2021 (WebSci '21), June 21–25, 2021, Virtual Event, United Kingdom*. ACM, New York, NY, USA, 9 pages. <https://doi.org/10.1145/3447535.3462511>

1 INTRODUCTION

Several issues in the study of visual politics call for a thoroughly large-scale approach. These issues include areas in political communication that have a global aspect, in which it matters how

*Both authors contributed equally to this research.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

WebSci '21, June 21–25, 2021, Virtual Event, United Kingdom

© 2021 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-1-4503-8330-1/21/06...\$15.00
<https://doi.org/10.1145/3447535.3462511>

communication on an issue (e.g. climate change) or a communicative style (e.g. populism) is characterised, perceived and distributed across different parts of the world. Digital media have provided not only a fertile ground for the large-scale diffusion of visual narratives, but also the opportunity to empirically study them. Yet in part because of the demanding nature of analysing visual themes, research in visual politics has often been limited in geographic, media and temporal scope (e.g. focused on national newspapers in selected western countries), thereby limiting our possibility to address such issues.

A major stumbling block for a large-scale approach to visual politics has been the capacity to automate the analysis of visual communication on the basis of actual images. Some existing computational approaches analyse images via accompanying text [13], but this overlooks the interactions and differences between image and text: the content of an image may complement or influence the meaning of the text around it, and images support a more immediate and emotional form of communication than text [24]. Now, advances in deep learning and image classification open a potential path to large-scale analysis on the basis of actual visual content. In the study of visual politics, promising work is starting to develop this potential in a variety of directions [9, 24, 51]: to study nonverbal cues in political persuasion [23, 38]; political ideology and bias in media [37]; protest events and mobilisation [11, 52, 54]; and the spread of (dis)information [53].

Nevertheless, integrating deep learning into the analysis of visual politics is not straightforward. First, the kinds of visual themes in focus in social image analysis can be more complex than the typical benchmark images used to develop and test deep neural networks. In machine learning, deep neural networks are learned from an existing mapping from data (in our case images) to labels, so that the network can be used to associate those labels to new data. In social image analysis we look for a mapping between images and themes, which may be vaguely defined and whose definition or operationalisation can change during the analysis. This mismatch between labels in machine learning and themes in political analysis has to be acknowledged in order to avoid a variety of problems related to the use of deep learning tools¹. These problems include constraining the researcher, the research process, and the validity of that process. Second, by enabling a large-scale and partially automated approach, deep neural networks introduce issues rooted in multiple disciplinary areas into the study of visual politics. Training a deep neural network using transfer learning, operationalising a theme or frame based on the denotational meaning of a social

¹We use the term tool because our focus is on how deep neural networks are used in a social science process and how using them influences the process.

image, characterising the ethical consequences of a tool-based research design, or repurposing an Application Programming Interface (API) to collect valid behavioral data, are not straightforward tasks, but can be individually addressed (and even considered easy) by people with the right expertise. However, the co-existence of multi-disciplinary issues all contributing to the validity of the process makes it important to identify them and understand how they interact, not only expanding but also to some degree transforming familiar approaches to research design.

This paper examines the conditions of integrating a deep learning tool for image classification into the study of visual politics to enable large-scale and global analysis. We focus on two aspects: how the complexity of visual political themes affects the integration, and how the apparent simplification of the process (through automation) leads to a research design that is more complex than the research designs it extends. We argue that the apparent simplification corresponding in fact to a complexification of the process should be exposed and exploited.

2 VISUAL THEMES AND SOCIAL IMAGES

Deep neural networks have seen a recent burst in popularity and performance, despite the underlying ideas being decades old. One frequently expressed reason for this is the increased accessibility of computing power, which is needed to learn the typically large number of model parameters. A less celebrated reason (probably because it is common to most areas in machine learning and thus taken for granted) has been identified in the ready availability of benchmark data [18], such as, in the visual context, the ImageNet collection. Using a few common training and test datasets, it has been possible for a very large number of researchers to compute performances and compare them with other proposals, establishing an evolutionary process where the most fit network models survive. However, the high accuracies reached on benchmark data may not be achieved on data that is associated with fewer resources (e.g., fewer people working on it, with less advanced computational skills or more limited access to computing power), or with focus on images that are conceptually different from those in the benchmarks. Both issues can be expected to arise in relation to work in visual politics. In particular the second issue makes the concept of performance based on accuracy less useful and central, and calls for a more thoroughgoing discussion of how best to integrate such a tool, and deep neural networks in general, in social image analysis.

This section focuses on the mismatch between the study of themes in visual politics and classifying images using deep neural networks. We first provide an overview of conventional approaches to the analysis of visual themes, before turning to the mismatch. The overarching argument is that while there are specific cases in which we can train a network and use it in a more straightforward manner, there are at least three types of instances common to analysis based on visual themes in which it becomes relevant to dig deeper to consider how the network can best be used. We end this section with an empirical example.

2.1 Visual themes in politics

The present discussion concerns studies of visual politics that focus on content themes in images. Studies of contents often focus on two

types of themes: the denotative (the literal description of what is depicted in the image) and the connotative (the interpreted meaning in context). The researcher either deductively pursues known themes drawn from the literature and/or inductively identifies emergent themes in the data. The interpretative level is the one at which one might apply *frame* analysis, the idea behind which is that the producers and presenters of content “select some aspects of a perceived reality and make them more salient in a communicating text, in such a way as to promote a particular problem definition, causal interpretation, moral evaluation, and/or treatment recommendation for the item described” [15]. Examples of descriptive themes identified in visual climate communication include “animals” (with major sub-theme “polar bears”) or “people” (with sub-themes such as “politicians”, “scientists”, “business representatives”, and “celebrities” [35, 36]), and common interpretative themes discussed include “causes”, “impacts,” and “solutions,” and the “distancing” frame, the latter of which presents climate change as an issue far removed from people’s everyday lives. Clearly, the line between descriptive and interpretative theme is not always strictly drawn. For example, the polar bear is both an arctic animal and an icon associated with the progressive climate movement (in particular Greenpeace campaigns), such that the package of climate ideas that are signalled by polar bears may also be evoked by costumed campaigners. In all forms, polar bears continue to be a recurring reference in climate discourse despite widespread acknowledgment that it is counter-productive to anchor climate issues in a concept that is experienced as psychologically and geographically remote [14, 21, 35, 36, 50].

Conventional studies of visual themes face two basic challenges. The one is to define a coding or interpretation scheme that is consistent and transferable. This is difficult even with simple themes. The concept and instructions of what to look for may be underdefined, ill-suited to the type of data under analysis, or just difficult to operationalise with precision. The other is that some themes, in particular interpretative frames, are abstract or inherently fuzzy, and entail wide margins of interpretative leeway. The *distancing frame* in climate communication and its counterpart the *local frame* are examples here, since their interpretation depends on the context of application and audience. There is a long tradition of debate and work dealing with these issues [6, 12, 20, 29]. For example, it is standard to undertake a lengthy and iterated process prior to the formal analysis, with extensive piloting work to develop, refine, and specify the coding procedure for identifying the presence of themes in the material, and in the case of multiple coders, to train them and report intercoder reliability. These known challenges contribute to the conditions of social image analysis that need to be taken into account when going large-scale and incorporating neural networks into the analysis, and also pose particular challenges with the integration of deep learning methodology.

2.2 Neural networks and visual themes

We know that associating themes to images plays an important part in the study of visual politics, and that deep neural networks can be used to assign labels to images. Therefore, it is natural to consider applying deep neural networks in the context of social image analysis. However, the relation between labels and themes is not straightforward, and this creates fresh challenges. The direct

training and application of a network can be expected to work well in simple cases, when a theme can be precisely defined before looking at the data, is atomic (as opposed to being composed of a combination of sub-themes and objects in the images), and denotative (as opposed to connotative). The following three sections raise the question of what happens when these conditions do not apply.

2.2.1 Evolving conceptualisations and operationalisations of themes. First, there are cases when the operationalisation (and possibly the conceptualisation) of the themes changes during the (pre-)analytical process. In these cases we cannot rely on a stable mapping between data and labels, which means that we cannot directly train a network. As noted, the tweaking of code operationalisation (and the information about themes gleaned during the process) is a standard feature of many conventional kinds of visual theme analysis. The twist is that we do not know if a code that we want to use will be recognised by the network. Therefore, we may be forced to adapt the coding to what can be achieved by the network instead of, or in addition to, adapting it to the available data. This calls for an iterative process where an initial simpler network is re-trained multiple times while the analyst refines the theme definition — with the interesting feature that this refinement is constrained not only by the data but also by the functioning of the network. Section 2.3 presents an example of this process.

2.2.2 Composite themes. An additional problem occurs when the themes are composite, that is, complex in terms of assuming a combination of elements that need to be interpreted together. To give an example, the climate change communication literature suggests that another kind of a *distancing* frame that is prominent in western news media is the focus on elites, such as politicians, which (in a different way from polar bears) also anchors climate discourse in aspects of the issue far removed from people's everyday lives. In this context, the concept of *eliteness* refers primarily to the events that dominate the climate politics calendar and news coverage: political summits (e.g. the UN Conference of the Parties). Besides the politicians, scientists and officials that populate such events, it also encompasses the formal situations they are shown in, such as the interviews, speeches, photoshoots and pronouncements.

To get at a composite theme one can operationalise it through a collection of labels, for example using an ensemble classifier. Ensemble models are typically defined as complex classifiers made of several different classifiers all trained to recognise the same labels. In this case, a different way to use multiple classifiers is to train them to recognise different labels, so that we can improve the similarity between the theme we want to classify and the combination of labels that the classifiers can recognise.

However, having a theme defined in terms of multiple labels raises several issues. A first challenge is to identify and motivate which labels (among the available ones if pre-trained networks are used) are reasonable to consider for the theme. For example, it might seem obvious to include the conventional dark suit that many (male) politicians adopt for international summits (e.g. “tie”, “suit”). But should we also include labels that might also refer to markers of other kinds of formality or eliteness, such as “businessperson”, or markers inclusive of other kinds of events, such as “microphone”? Moreover, an important consideration is to what extent such labels,

or a combination of such labels, are necessary or sufficient to represent the theme. It is possible that labels that do not individually represent eliteness can be considered to do so when they appear together, but the challenge is to identify these collections. An additional question is whether we should use a probabilistic model of the theme, where a picture is either about eliteness or not and we represent the likelihood of it, or a fuzzy model, where an image can be more or less about eliteness. Versions of such operationalisation issues appear in all studies, but are likely to have different solutions and deserve to be openly discussed.

Finally, the possibility of collating labels to represent a theme also raises questions about how to test the accuracy of multiple labels at the same time, and of labels that map on to themes as collections. For example, it would be difficult to estimate the accuracy of a classifier to identify the eliteness theme if this classifier shows a high accuracy on “microphone” and “tie” but a lower one on “businessperson”, assuming that we do not know the relative frequency and co-occurrence rate of these labels in the target data.

One alternative is to collect a set of images that we recognise as being about eliteness and train a network on these images. However, using a network to learn (and so decide) what such a complex theme means generates other issues, as we discuss in the next section.

2.2.3 Connotative themes. A third kind of issue emerges in relation to the study of themes that are complex in their connotative dimensions. Early work addressing the use of deep neural networks in the study of visual politics has focused on cases where simple labels are available and are the object of the study. This has been interesting and successful, but limited to networks that are easily trained on denotative themes, which in turn limits access to more complex themes. Interesting work is also emerging in which the operationalisation or definition of the labels is left to the network to be learned (after initial expert annotation) [11]. This in principle has the potential to move us one step closer to addressing complex themes. To the extent that we can assume that the information to recognize connotative themes is present in an image although not associated to obvious visual patterns, then it is possible that such a network can be trained to recognise those themes. However, such an approach when used on connotative themes still presents a serious but quite different kind of issue. Leaving it to the network to take care of learning the classification of a connotative theme omits a mechanism in the conventional process that has multiple functions. One such function is to build into the process a space for the analyst to reason about what defines the theme, increase her understanding of it, and feed this understanding back into the analytical process — something that has been valued as an important part of that process. If we replace part of a research process with an automated tool without rethinking both the process and how we use the tool, we risk turning an opportunity into a limitation, on top of facing validity issues due to the black-box nature of the tool. This is another case where adopting an iterative method in which both the network and our understanding of the themes may evolve not only makes the network more aligned with what we want, but also decreases the risks of misunderstanding what it is doing.

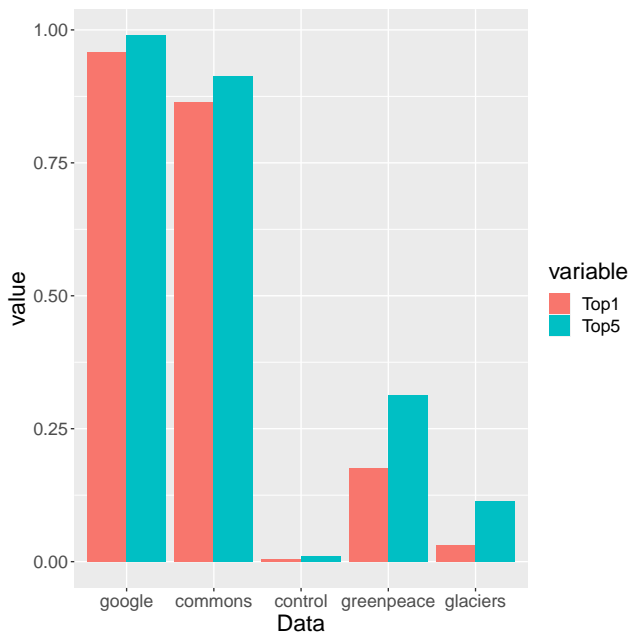


Figure 1: Top-1 and top-5 recall of the resnet50 network on five test datasets

2.3 An empirical example

To experimentally support our argument, we provide an example with reference to a study that analyses the prevalence of “distancing” visual communication in the global discourse on climate change online. Key parts of that study centre on the prevalence of polar bears on YouTube. YouTube is a digital platform that citizens around the world use to access science-related information [1], and on which the platform assessment of “relevance” is significant for the prominence of contents on the site [7, 41]. Visual climate communication is a good example of an area in which a global view is important yet understudied due to methodological reasons. Populations around the world, some of whom have stronger visual than textual traditions, experience climate impacts differently and relate in different ways to visual communication around the issues [27, 39]. Yet most work has focused on print media in western countries, so it is less well understood how the issue is presented to citizens in other parts of the world, or how the hybridisation of the media ecology shapes such discourse.

Consider the case where we want to identify the presence of polar bears in frames extracted from YouTube videos to study the global penetration of this theme. The data set consists of 7500 videos from which one representative frame (image) has been extracted for each scene, corresponding to about half a million images.

The popular ImageNet training data contains a label `ice_bear`, so we start by choosing a network (resnet50) pre-trained on ImageNet. Of course, before using the network we have to test if it “works”. If our data is too large to manually identify images with polar bears to do the testing, we can obtain test data on the Web. In Figure 1, on the left, we see the recall of the network when applied to a set of images extracted from three test datasets:

- **google**: 97 images from a search on google images (search key: ‘polar bear’),
- **commons**: 107 images from Wikimedia Commons (category: polar bears²), and
- **control**: 178 random images from the target data not including polar bears, i.e. random video frames from YouTube videos about climate change from which images including polar bears were manually removed.

The two bars for each dataset indicate the percentage of images labeled as `ice_bear` as the top choice of the network (Top-1), and the percentage of images where the `ice_bear` label appears as one of the five top choices (Top-5). These results suggest a high performance both on recall (that is, the proportion of images in the datasets about polar bears labeled as `ice_bear`) and precision (because almost no image labeled as `ice_bear` is in the control data without polar bears). In particular, by annotating an image as `ice_bear` when this label is in the top-5 choices, the network would almost always be right.

However, looking at the images in the google dataset (mostly representing polar bears in the wild) and the errors made in the commons dataset, we considered refining the concept of polar bear by also including polar bears as used in climate change communication campaigns. Therefore, we collected an additional dataset from a Greenpeace Web page. Greenpeace is a prominent advocacy organisation on the issue, and, as noted, its campaigns are widely credited for having established the polar bear in discourse around climate politics.

- **greenpeace**: 52 images from a Greenpeace page showing snaps from campaigns featuring polar bears, often an activist dressed in a polar bear suit³.

Figure 1 shows that the current network identifies polar bears in one third of the Greenpeace dataset, and all of those images, bar one, depict polar-bear-suit campaign stunts.

At this point we had a choice to make: should we include this type of imagery in the study? Irrespective of our choice, we would have to retrain the network to either identify them with higher accuracy or not identify them. Otherwise, if the types of Greenpeace-like images that our network can recognise are not evenly distributed across our data we would risk to obtain misleading results.

A qualitative examination of the two errors on the control dataset also led us to suspect that the network may factor in glacial surroundings into its understanding of polar bear, or rather *polarbearness*: both images with no polar bears labeled as `ice_bear` depicted such types of landscapes. This hypothesis seemed to be compatible with several of the pictures in the Greenpeace data, showing full polar-bear-suited figures in pale surroundings (e.g. ice and snow, concrete, sand, against a pale sky). This is also compatible with findings in the literature [40]. Thus we examined this idea by testing the network’s recall on a fifth specifically collected dataset:

- **glaciers**: 97 images of glaciers from google search, where polar bears were removed.

As Figure 1 shows, the network identifies some glacier images as polar bears, suggesting that glacial surroundings do play a role in

²https://commons.wikimedia.org/wiki/Ursus_maritimus

³<https://www.greenpeace.org/usa/50-times-polar-bears-crushed-greenpeace/>

its understanding of polarbearness. Once more, we have a choice to make: should our polarbearness theme include typical polar bear surroundings even if polar bears are not in the picture? This decision would then again require a re-training. What we find interesting is that this process of refining the definition and operationalisation of our theme is not (just) influenced by the increased understanding of the data that we acquire during the analysis, as it could also happen in a conventional study, but also by interacting with the network. Importantly, such fuzzy and loose edges around empirical polarbearness that clearly go somewhat beyond actual polar bears may still be analytically acceptable and relevant in a study of climate campaigning, but these edges represent an uncertainty that needs to be communicated. We come back to this in the next section.

As a final comment, please note that what we have exemplified above is just part of a process which may have additional iterations. Several images in the Greenpeace data in which the network identified polar bears depict costumed activists in decidedly bright surroundings, e.g., activists straddling bright yellow gas station signs or playing football on a vivid green field. This means that it is still not clear cut exactly what the network is actually doing or how it understands polarbearness. We here conclude this example, but argue that this process of understanding and refinement should continue in an actual empirical study.

3 THE TOOL AND THE PROCESS

Up until now we have focused on the tool in itself. Beyond this, we need to consider the transformative power of the tool in and for the research process. This is not just a matter of “plug and play”. Integrating a deep neural approach into the analysis of visual politics shifts the grounds of the research design in ways that require acknowledgment and adjustment on at least two fronts.

One front concerns how the integration of a deep neural network shifts conditions for methodological and analytical process in the analysis of visual themes. This is particularly clear with respect to the familiar challenges of achieving coding consistency and replicability mentioned in Section 2.1. In one sense, using the tool resolves some of those challenges. While trained human interpreters of visual themes may come to somewhat different results, the deep neural network can reliably reproduce its results on the same data. Nevertheless, as suggested by the previous section, the network’s black box workings still place heavy demands on how we can understand it and interpret its results. Where the challenges of coding and interpreting themes under conventional conditions have been addressed with supporting documentation such as code books and the illustrative discussion of difficult examples, the interpretation of the network and its results requires insight into the specifics of what the network is doing. The literature observes that a certain transparency can be achieved by communicating and motivating the processes involved, that is how we extracted the model parameters from the training data and how we generated the labels [43]. This suggests there is work to be done in the emerging field of tool-based analysis of visual politics to rethink research infrastructures in the form of best practices for the kinds and extent of documentation that adequately — transparently, relevantly, and in a manageable manner — addresses such conditions.

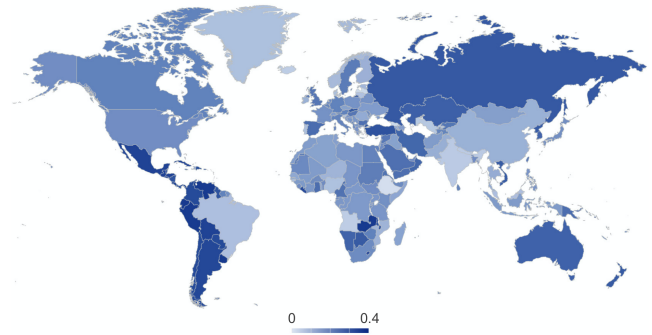


Figure 2: Relative frequency of top-50-relevant YouTube videos showing polar bears in the wild, with search term “climate change” translated in the official languages of each region

Along these lines, one implication of the discussion in the previous section is the need to find ways to systematically communicate the uncertainties of the process and its bearing on the results. This, not to undermine the quality of the results, although it acknowledges their limits, but rather to make them more precise. For example, consider the possible representation of the result of our case study about the distribution of polarbearness around the world, shown in Figure 2. This map, as it stands, is patently misleading. Throughout the paper we have raised questions about the validity of these results as they stand here, so this would need to be more carefully communicated. A qualitative example of uncertainty that is analytically acceptable, but still needs to be communicated, concerns the apparently loose edges of the tool concept of polarbearness. As with conventional coding, it would be helpful to report examples of typical or difficult cases. In our case, this would include not only typical images that fully captured the intended target (polar bears in the wild), but also the images that the tool identified from the margins of the code (e.g. some arctic landscapes; some polar bear suit campaign stunts). From a quantitative perspective, an important consideration would be how to report intervals of confidence. The accuracy analysis has given us an indication of how many errors the tool may make depending on the type of images. While a map provides an accessible summary of the obtained values of polarbearness, a different representation (for example, a bar plot with regions sorted by polarbearness) would make it easier to see which countries have values similar enough to make us doubt about the significance of the difference. This said, we should not forget that the wrongly classified images in the test data may not be equally distributed across the contexts we are comparing. For example, if some regions are significantly more associated to activist videos and videos about glaciers than others, the impact of the mixed results obtained by the tool on these examples would affect different regions in different ways, introducing one more bias.

To turn to another front, integrating deep neural networks into the study of visual politics will draw on several areas of expertise. A concern is that once we get a simple tool that can process large amounts of data and seems to be accurate on reliable test

data (which we have shown to be a less likely scenario than one may expect), there is an associated risk that we overlook other elements the tool interacts with in the research process. Many such considerations lie beyond the tool itself, but are necessary to make it work for the analytical purposes for which it is being applied. For example, integrating deep neural network methodology into the visual politics toolkit changes the data conditions for the field and for specific studies. It directs researcher attention to data that is accessible at scale. This widens the scope of visual theme data that can be analysed. However it also introduces the complexities associated with big data in general, in addition to the character and issues of the data being analysed in particular. Digital and social media data, one likely source available at the scale required, is an example of data with methodological and analytical characteristics that are well known amongst specialists, but easily overlooked by others. Disregarding such features can mean that even applying an accurate tool can produce low quality or misleading results. We illustrate our point here with an example.

3.1 An experimental example

The objective of this example is to highlight how the new type of process enabled by the application of deep neural networks may introduce a number of issues that are not present in conventional research on visual politics or in research more focused on the technical aspects of machine learning. Notice that we are not claiming that these issues are new, in fact they are very well known by researchers in digital methods and Internet studies [42]. We are however suggesting that these issues may be overlooked by researchers having different but relevant backgrounds. It is therefore important to highlight the existence of these issues: to do the kind of research discussed in this paper it is necessary to build (and thus first identify) a specific set of competences that today are typically spread across different people and subjects.

Visual politics is always studied in *context*. In our example, the contexts are different geographical and linguistic regions of the world, with the objective of studying how climate-change-related visual content as selected by the YouTube relevance-ranking algorithm differs based on the location from where it is accessed and the language used to retrieve the videos. Other examples of contexts that can be used to do comparative large-scale studies are different time windows, different social media platforms, or different communities inside a platform.

In our example, the contexts are geographical and linguistic regions observed through YouTube (relevance ranking) in a fixed time interval (2005-2020). Figure 3 shows the stability of the results obtained by running the same YouTube search twenty times, of which ten consecutive executions on one day and ten additional executions on the second day. Notice that the time interval from which videos are retrieved does not change in different searches, because it is upper bounded. Without an upper time limit, later searches could return new videos not previously available. For each pair of searches we computed the overlapping of their results using the Jaccard index. A value of 1 indicates that the two searches returned the same set of videos, a value of 0 would indicate disjoint results. Also, notice that we have done this considering different numbers of top results: 10, 50 and 400.

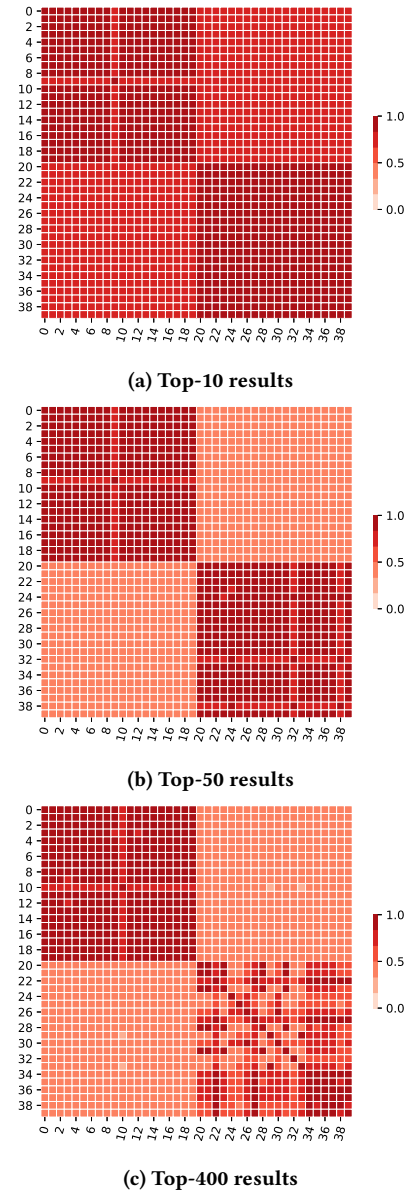


Figure 3: Jaccard similarity of two sets of consecutive executions, numbered from 0 to 39, of the same API query. The first twenty and the last twenty were run on different days

The bottom-right part of Figure 3a shows that for one of the sets of searches the top-10 results (that is, videos) we obtained were the same (Jaccard value of 1). This is however the only case when this happens, showing once more how a plausible test setting can lead us to conclude that the study is more valid than it actually is. In fact, the top-left part of the figure shows how even in the top-10 results of consecutive searches we may find different videos, and the two non-diagonal blocks indicate how the same search performed on a different day has a higher probability for this to happen. Finally, if we look at result similarity on a larger number

of results, we can see that in this test it would be lower on different-day queries (Figure 3b, non-diagonal blocks), and using even more results may significantly reduce similarity even on consecutive searches (Figure 3a, bottom-right block).

As we discussed regarding accuracy, also in this case exposing this validity issue is not an end point. First, it may still be important to observe that *under specific circumstances* YouTube’s algorithm would propose the themes we observe in those videos, being clear that this can happen but is not guaranteed. We can also test and if possible increase generalisability by comparing results over multiple executions, or collecting videos either multiple times or down to lower ranked videos until we reach some stability — although we should then again be clear that this might not reflect the visual content that a user can be exposed to at a specific time, but would become even more a study of the YouTube platform rather than what users actually see.

Another issue regarding the data is the granularity of the contexts we can explore, which in our case is limited to ISO codes by the YouTube API and does not allow us to study smaller culturally homogeneous geographical regions (e.g., Palestine). This is one of many possible types of data incompleteness — others being the absence of representative images (not retrieved by a specific search, or lost during pre-processing such as video segmentation or image clustering), of retrievable objects (not represented in the list of available labels), or language translation (where we may not be able to ascertain how to express the search term in specific languages). Notice that this last problem creates biases in the results that can be more pronounced for minority or less popular languages, also extending to the quality of the results if we want to extract text from the images or translate video descriptions to complement the visual analysis. Finally, notice that some completeness problems may result into errors: YouTube is blocked in China, but a YouTube API search with ISO code CN would still return a result without any warning message. When testing this search, the returned video corresponded to the ones obtained with a search using the Taiwan ISO code.

This is just one example of the several reasons why an accurate tool can produce wrong or misleading results without a systematic consideration of the cross-disciplinary competences needed to perform such a process.

4 DISCUSSION

Integrating deep learning into the analysis of visual politics is promising, but not straightforward. Aspects that one might first assume to be daunting — such as inductively identifying emergent themes in the vast data at hand — may not necessarily present the trickiest challenge. Instead, considerable trickiness comes from the demands of social image analysis and the way in which integrating deep neural networks extends and transforms the demands of appropriate research design.

This paper focused on two issues. The first centred on the complexity of the visual themes that are often in focus in such studies. At least three aspects of social image analysis make it difficult to identify political themes using deep neural networks: the feature of evolving conceptualisation and operationalisation of themes in the analysis, and the challenges of composite and connotative themes.

The failure to fully acknowledge and address such issues has several implications. These include the problems that it can affect the validity of results, delimit the scope of inquiry to easier-to-capture denotative themes, and constrain the role of the analyst by transferring the learning process to the network without compensating for the lost opportunity for researcher reflection. The second concerns the way in which the integration of a deep neural network draws studies in visual politics into a deeply cross-disciplinary research endeavour that extends and possibly transforms the research process and its associated documentation, and places high requirements on diverse expertise. We noted that the competencies required span disciplines and sub-fields, and provided some examples. We stress that to systematically identify these competences and their interactions is important future work.

These points relate to issues that are discussed in various parts of the literature. The challenging features we identify in visual themes typical of social image analysis seem to be more specific than those discussed in the general literature on why classification is difficult, which are typically associated with the ambiguity of the classes, the sparsity and dimensionality of the data, and the complexity of the decision boundary [22, 26]. Notice that ambiguity here indicates a case where there is no recoverable mapping between the features sampled in the data and the labels, which does not consider the evolving conceptualisation or operationalisation of the theme and the distinction between labels and themes.

As part of this discussion on the complexity of classification, we highlight the need to go beyond typical metrics of performance such as accuracy, precision, and recall, to focus more on specific instances in the data and use them both to evolve the tool and the conceptualisation or operationalisation of the themes. Several works have also looked at how to deal with specific instances that are difficult to classify. Boosting is such an example, and still one of the most popular approaches in machine learning. However, existing works and methods focus on the scenario where we have multiple annotators or classifiers (as in boosting), on how to characterise what makes these instances hard to classify [28], often with respect to their geometrical features (outliers, border points, or minority classes) [48], and on how to handle the cases with a low intercoder reliability [2]. While we also highlight the importance of instance-level analysis, we are more interested in the way in which instances are used in social science research, for example to develop a conceptualisation or to document critical choices made during the analysis.

The extent to which deep learning can replace humans in performing specific tasks has been discussed in the literature, for example with respect to how the human visual system and specific network architectures are differently affected by image manipulation [17]. While this is a relevant question, in the context of this paper the focus is more on how humans and networks should collaborate, which is something that has been studied for example in Human Computer Interaction and Artificial Intelligence (AI). In particular, some general principles exist in the field of human-in-the-loop AI [33], but how to best put a human in the loop is a context-dependent question that deserves dedicated attention in the designated application areas.

Another concept that is relevant for this paper is explainability, with respect to our discussions of both the process of refining and

understanding what the network is doing, and the need to develop best practices for supporting documentation. Explainability has recently received renewed attention [19], with specific works focusing on explaining deep learning systems [34, 44, 45]. This literature emphasises several important points. First, explainability is highly context dependent [19], so we need to discuss how it should be best achieved in the field of visual politics. Second, a lot of the technical literature on explainability has machine learning engineers as end-users, resulting in a gap between the literature on explainability and how existing approaches should be used for example in specific areas of social science research [5]. We also notice a predominant focus on explainability in the natural sciences [43], although general principles such as the (not always clear) distinction between transparency, interpretability, and explainability can be borrowed by the social sciences too. This gap between explainability as studied in machine learning and expectations from the social sciences has also been observed in the literature [31], and recently started receiving more attention [30].

A more general note is that even if some concepts needed in social science research have been treated in the machine learning literature, it can be difficult to identify them. This is both because of the burgeoning research in the area, but also because the literature is characterised by general features that makes it difficult to transfer to social science research. In particular, it is characterised by an abundance of speculation (as opposed to explanation), lack of clarity with respect to the sources of high performance, and, in some cases, the misuse of mathematics and language [25]. This touches on general concerns about rigour [16, 46], which are out of the scope of this paper but motivate future work to extract, adapt and extend knowledge from the machine learning literature to make it applicable to valid, deep and critical social science research that can go beyond the easy shiny results one can obtain by directly applying automated tools to a large amount of data.

We end by noting that an important set of complicating factors in the turn to deep neural networks in the study of visual politics (and other areas) stem from the ways research groups are situated in society and the world. These include the uneven distribution of computational skills and resources in the production of knowledge about society, e.g., between the technical and social sciences, and between industry and academic research [4, 32, 47], and the environmental impact of work with deep neural networks. While most of the literature on how to design or train better models focuses on accuracy, scholars have also broached very different types of evaluation criteria for machine learning algorithms and in particular deep neural networks. Several studies that deal with the economic and environmental costs of training [3, 8, 10, 49] have focused on deep neural networks because of the large number of parameters these classifiers need to set, leading to computational requirements (e.g., in terms of hardware and computing time) significantly exceeding those of more traditional types of classifiers. While commercial actors with access to large computing resources and large training datasets may provide more accurate tools, such work may need to be more strongly justified when the investment comes from public funding agencies (and thus tax payers). Both alternatives need to be assessed against the possibility that high costs of data processing may decrease the replicability of particular studies, exacerbate uneven conditions in social science research, and tax the environment

– and, if not done with careful attention to the complicating factors, without leading to higher quality social science research.

ACKNOWLEDGMENTS

This research has been funded by the AI4Research initiative at Uppsala University and the Swedish Research Council (2015-01835).

REFERENCES

- [1] Joachim Allgaier. 2019. Science and Environmental Communication on YouTube: Strategically Distorted Communications in Online Videos on Climate Change and Climate Engineering. *Frontiers in Communication* 4 (2019), 36. <https://doi.org/10.3389/fcomm.2019.00036>
- [2] Beata Beigman Klebanov and Eyal Beigman. 2014. Difficult Cases: From Data to Learning, and Back. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics, Baltimore, Maryland, 390–396. <https://doi.org/10.3115/v1/P14-2064>
- [3] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FACT '21)*. Association for Computing Machinery, Virtual Event, Canada, 610–623. <https://doi.org/10.1145/3442188.3445922>
- [4] Yochai Benkler. 2019. Don't let industry write the rules for AI. *Nature* 569, 7755 (2019), 161–161. <https://doi.org/10.1038/d41586-019-01413-1>
- [5] Umang Bhatt, Alice Xiang, Shubham Sharma, Adrian Weller, Ankur Taly, Yunhan Jia, Joydeep Ghosh, Ruchir Puri, José M. F. Moura, and Peter Eckersley. 2020. Explainable machine learning in deployment. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT* '20)*. Association for Computing Machinery, New York, NY, USA, 648–657. <https://doi.org/10.1145/3351095.3375624>
- [6] Mary Angela Bock. 2020. Theorising visual framing: contingency, materiality and ideology. *Visual Studies* 35, 1 (2020), 1–12. <https://doi.org/10.1080/1472586X.2020.1715244>
- [7] Liliana Bounegru, Kari De Pryck, Tommaso Venturini, and Michele Mauri. 2020. “We only have 12 years”: YouTube and the IPCC report on global warming of 1.5°C. *First Monday* (2020). <https://doi.org/10.5210/fm.v25i2.10112>
- [8] Benedetta Brevini. 2020. Black boxes, not green: Mythologizing artificial intelligence and omitting the environment. *Big Data & Society* 7, 2 (2020), 2053951720935141. <https://doi.org/10.1177/2053951720935141> Publisher: SAGE Publications Ltd.
- [9] Erik P. Bucy and Jungseock Joo. 2021. Editors' Introduction: Visual Politics, Grand Collaborative Programs, and the Opportunity to Think Big. *The International Journal of Press/Politics* 26, 1 (2021), 5–21. <https://doi.org/10.1177/1940161220970361>
- [10] Alfredo Canziani, Adam Paszke, and Eugenio Culurciello. 2017. An Analysis of Deep Neural Network Models for Practical Applications. *arXiv:1605.07678 [cs]* (2017). <http://arxiv.org/abs/1605.07678> arXiv: 1605.07678.
- [11] Andreu Casas and Nora Webb Williams. 2019. Images that Matter: Online Protests and the Mobilizing Role of Pictures. *Political Research Quarterly* 72, 2 (2019), 360–375. <https://doi.org/10.1177/1065912918786805>
- [12] Eileen Culloty, Pdraig Murphy, Patrick Brereton, Jane Suiter, Alan F. Smeaton, and Dian Zhang. 2019. Researching Visual Representations of Climate Change. *Environmental Communication* 13, 2 (2019), 179–191. <https://doi.org/10.1080/17524032.2018.1533877>
- [13] Koen Deschacht and Marie-Francine Moens. 2007. Text Analysis for Automatic Image Annotation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*. Association for Computational Linguistics, Prague, Czech Republic, 1000–1007. <https://www.aclweb.org/anthology/P07-1126>
- [14] Julie Doyle. 2007. Picturing the Climate: Greenpeace and the Representational Politics of Climate Change Communication. *Science as Culture* 16, 2 (2007), 129–150. <https://doi.org/10.1080/09505430701368938>
- [15] Robert M. Entman. 1993. Framing: Toward Clarification of a Fractured Paradigm. *Journal of Communication* 43, 4 (1993), 51–58. <https://doi.org/10.1111/j.1460-2466.1993.tb01304.x>
- [16] Jessica Zosa Forde and Michela Paganini. 2019. The Scientific Method in the Science of Machine Learning. *arXiv:1904.10922 [cs, stat]* (2019). <http://arxiv.org/abs/1904.10922> arXiv: 1904.10922.
- [17] Robert Geirhos, Carlos R. Medina Temme, Jonas Rauber, Heiko H. Schütt, Matthias Bethge, and Felix A. Wichmann. 2018. Generalisation in humans and deep neural networks. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems (NIPS'18)*. Curran Associates Inc., Red Hook, NY, USA, 7549–7561.
- [18] Justin Grimmer, Margaret E. Roberts, and Brandon M. Stewart. 2021. Machine Learning for Social Science: An Agnostic Approach. *Annual Review of Political Science* (2021). <https://doi.org/10.1146/annurev-polisci-053119-015921>

- [19] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. 2018. A Survey of Methods for Explaining Black Box Models. *Comput. Surveys* 51, 5 (2018), 93:1–93:42. <https://doi.org/10.1145/3236009>
- [20] Anders Hansen. 2017. Methods for Assessing Visual Images and Depictions of Climate Change. <https://doi.org/10.1093/acrefore/9780190228620.013.491>
- [21] A. Hansen and D. Machin. 2008. Visually branding the environment: climate change as a marketing opportunity. *Discourse Studies* 10, 6 (2008), 777–794. <https://doi.org/10.1177/1461445608098200>
- [22] T. K. Ho and M. Basu. 2002. Complexity Measures of Supervised Classification Problems. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 24, 3 (2002), 289–300. <https://doi.org/10.1109/34.990132>
- [23] Jungseock Joo, Erik P. Bucy, and Claudia Seidel. 2019. Computational Communication Science| Automated Coding of Televised Leader Displays: Detecting Nonverbal Political Behavior With Computer Vision and Deep Learning. *International Journal of Communication* 13, 0 (2019), 23. <https://ijoc.org/index.php/ijoc/article/view/10725>
- [24] Jungseock Joo and Zachary C. Steinert-Threlkeld. 2018. Image as Data: Automated Visual Content Analysis for Political Science. *arXiv:1810.01544 [cs, stat]* (2018). <http://arxiv.org/abs/1810.01544> arXiv: 1810.01544.
- [25] Zachary C. Lipton and Jacob Steinhardt. 2019. Troubling Trends in Machine Learning Scholarship: Some ML papers suffer from flaws that could mislead the public and stymie future research. *Queue* 17, 1 (2019), Pages 80:45–Pages 80:77. <https://doi.org/10.1145/3317287.3328534>
- [26] Ana C. Lorena, Luis P. F. Garcia, Jens Lehmann, Marcilio C. P. Souto, and Tin Kam Ho. 2019. How Complex Is Your Classification Problem? A Survey on Measuring Classification Complexity. *Comput. Surveys* 52, 5 (2019), 107:1–107:34. <https://doi.org/10.1145/3347711>
- [27] Alfons Maes. 2017. The visual divide. *Nature Climate Change* 7, 4 (2017), 231–233. <https://doi.org/10.1038/nclimate3251>
- [28] Fernando Martínez-Plumed, Ricardo B. C. Prudêncio, Adolfo Martínez-Usó, and José Hernández-Orallo. 2019. Item response theory in AI: Analysing machine learning classifiers at the instance level. *Artificial Intelligence* 271 (2019), 18–42. <https://doi.org/10.1016/j.artint.2018.09.004>
- [29] Jörg Matthes and Matthias Kohring. 2008. The Content Analysis of Media Frames: Toward Improving Reliability and Validity. *Journal of Communication* 58, 2 (2008), 258–279. <https://doi.org/10.1111/j.1460-2466.2008.00384.x>
- [30] Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence* 267 (2019), 1–38. <https://doi.org/10.1016/j.artint.2018.07.007>
- [31] Brent Mittelstadt, Chris Russell, and Sandra Wachter. 2019. Explaining Explanations in AI. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19)*. Association for Computing Machinery, New York, NY, USA, 279–288. <https://doi.org/10.1145/3287560.3287574>
- [32] Shakir Mohamed, Marie-Therese Png, and William Isaac. 2020. Decolonial AI: Decolonial Theory as Sociotechnical Foresight in Artificial Intelligence. *Philosophy & Technology* 33, 4 (2020), 659–684. <https://doi.org/10.1007/s13347-020-00405-8>
- [33] Robert Monarch. 2021. *Human-in-the-Loop Machine Learning*. Manning publications. <https://www.manning.com/books/human-in-the-loop-machine-learning>
- [34] Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. 2018. Methods for interpreting and understanding deep neural networks. *Digital Signal Processing* 73 (2018), 1–15. <https://doi.org/10.1016/j.dsp.2017.10.011>
- [35] Saffron O'Neill. 2020. More than meets the eye: a longitudinal analysis of climate change imagery in the print media. *Climatic Change* 163, 1 (2020), 9–26. <https://doi.org/10.1007/s10584-019-02504-8>
- [36] Saffron J. O'Neill. 2013. Image matters: Climate change imagery in US, UK and Australian newspapers. *Geoforum* 49 (2013), 10–19. <https://doi.org/10.1016/j.geoforum.2013.04.030>
- [37] Yilang Peng. 2018. Same Candidates, Different Faces: Uncovering Media Bias in Visual Portrayals of Presidential Candidates with Computer Vision. *Journal of Communication* 68, 5 (2018), 920–941. <https://doi.org/10.1093/joc/jqy041>
- [38] Yilang Peng. 2021. What Makes Politicians' Instagram Posts Popular? Analyzing Social Media Strategies of Candidates and Office Holders with Computer Vision. *The International Journal of Press/Politics* 26, 1 (2021), 143–166. <https://doi.org/10.1177/1940161220964769>
- [39] Lisa Petheram, Natasha Stacey, and Ann Fleming. 2015. Future sea changes: Indigenous women's preferences for adaptation to climate change on South Goulburn Island, Northern Territory (Australia). *Climate and Development* 7, 4 (2015), 339–352. <https://doi.org/10.1080/17565529.2014.951019>
- [40] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16)*. Association for Computing Machinery, New York, NY, USA, 1135–1144. <https://doi.org/10.1145/2939672.2939778>
- [41] Bernhard Rieder, Óscar Coromina, and Ariadna Matamoros-Fernández. 2020. Mapping YouTube. *First Monday* (2020). <https://doi.org/10.5210/fm.v25i8.10667>
- [42] Richard Rogers. 2019. *Doing Digital Methods*. <https://uk.sagepub.com/en-gb/eur/doing-digital-methods/book261134>
- [43] Ribana Roscher, Bastian Bohn, Marco F. Duarte, and Jochen Garcke. 2020. Explainable Machine Learning for Scientific Insights and Discoveries. *IEEE Access* 8 (2020), 42200–42216. <https://doi.org/10.1109/ACCESS.2020.2976199> Conference Name: IEEE Access.
- [44] Wojciech Samek, Alexander Binder, Grégoire Montavon, Sebastian Lapuschkin, and Klaus-Robert Müller. 2017. Evaluating the Visualization of What a Deep Neural Network Has Learned. *IEEE Transactions on Neural Networks and Learning Systems* 28, 11 (2017), 2660–2673. <https://doi.org/10.1109/TNNLS.2016.2599820>
- [45] Wojciech Samek, Thomas Wiegand, and Klaus-Robert Müller. 2017. Explainable artificial intelligence: understanding, visualizing and interpreting deep learning models. 1 (2017), 10.
- [46] D. Sculley, Jasper Snoek, Alex Wiltschko, and Ali Rahimi. 2018. Winner's Curse? On Pace, Progress, and Empirical Rigor. In *ICLR Workshop Track*. <https://openreview.net/forum?id=rJWF0Fywf>
- [47] Simon Lindgren and Jonny Holmström. 2020. A social science perspective on artificial intelligence: building blocks for a research agenda. *Journal of digital social research* 2, 3 (2020), 1–15.
- [48] Michael R. Smith, Tony Martinez, and Christophe Giraud-Carrier. 2014. An instance level analysis of data complexity. *Machine Learning* 95, 2 (2014), 225–256. <https://doi.org/10.1007/s10994-013-5422-z>
- [49] Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2019. Energy and Policy Considerations for Deep Learning in NLP. *arXiv:1906.02243 [cs]* (2019). <http://arxiv.org/abs/1906.02243> arXiv: 1906.02243.
- [50] Susie Wang, Adam Corner, Daniel Chapman, and Ezra Markowitz. 2018. Public engagement with climate imagery in a changing digital landscape. *Wiley Interdisciplinary Reviews: Climate Change* 9, 2 (2018), e509. <https://doi.org/10.1002/wcc.509>
- [51] Nora Webb Williams, Andreu Casas, and John D. Wilkerson. 2020. *Images as Data for Social Science Research: An Introduction to Convolutional Neural Nets for Image Classification* (1 ed.). Cambridge University Press. <https://doi.org/10.1017/9781108860741>
- [52] Donghyeon Won, Zachary C. Steinert-Threlkeld, and Jungseock Joo. 2017. Protest Activity Detection and Perceived Violence Estimation from Social Media Images. In *Proceedings of the 25th ACM international conference on Multimedia (MM '17)*. Association for Computing Machinery, Mountain View, California, USA, 786–794. <https://doi.org/10.1145/3123266.3123282>
- [53] Savvas Zannettou, Tristan Caulfield, Barry Bradlyn, Emiliano De Cristofaro, Gianluca Stringhini, and Jeremy Blackburn. 2019. Characterizing the Use of Images in State-Sponsored Information Warfare Operations by Russian Trolls on Twitter. *arXiv:1901.05997 [cs]* (2019). <http://arxiv.org/abs/1901.05997> arXiv: 1901.05997.
- [54] Han Zhang 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 (2019), 1–57. <https://doi.org/10.1177/0081175019860244>