Automated Visual Clustering for Image Corpus Exploration, Stratified Random Sampling, and Annotation Cost Reduction*

Andreu Casas[†] Nora Webb Williams[‡] Kevin Aslett[§] John Wilkerson[¶]

Preliminary draft: please do not circulate without permission of the authors.

Abstract

Compared to text and audio, images can be an especially effective form of political communication. It has become relatively easy to automatically label images for many features of interest (such as protests, famous people or facial expressions). As a result, scholars are increasingly using large-N image analysis to investigate contemporary political attitudes and behavior. We address three emerging needs of image scholarship. First, researchers may want to visually explore an image corpus to discern patterns before they begin assigning labels. Second, they may want to annotate images for the presence of complex theoretical mechanisms that cannot be easily assigned using existing automated methods. Third, they may be primarily interested in studying human annotation decisions. We demonstrate how unsupervised image clustering can help researchers address each of these needs when dealing with large unbalanced image corpora. We illustrate this using a corpus of images shared with the hashtag #FamiliesBelongTogether on Twitter.

*Research supported by NSF Grant Number 1727459, "The Power of Images: A Computational Investigation of Political Mobilization via Social Media."

[†]Assistant Professor, *Vrije Universiteit Amsterdam*: a.casassalleras@vu.nl

[‡]Assistant Professor, *University of Illinois*: nww3@illinois.edu

[§]Postdoctoral Fellow, *New York University*: kma412@nyu.edu

[¶]Professor and Chair of Political Science, *University of Washington*: jwilker@uw.edu

1 Introduction

The proliferation of digital media and mobile communications, in combination with advances in deep learning, has made the study of visual information an increasingly important domain of social science research (Joo and Steinert-Threlkeld, 2018; Torres and Cantu, 2020; Webb Williams et al., 2020). A growing number of scholars are turning to convolutional neural networks (CNNs), a type of supervised deep learning algorithm, to study large image corpora. Some scholars are interested in visual communication *per se*, either as an outcome variable (i.e. categorizing nonverbal political communication (Joo et al., 2019)) or as a predictor (i.e. which image features render politicians more favorable (Peng, 2018)). Others use images to measure a quantity of interest that would be very difficult to measure otherwise, such as the levels of violence in a large number of decentralized protests (Sobolev et al., 2020), economic development based on nighttime lights imagery (Jean et al., 2016), or the number of corrupted voting tallies in a rigged election (Cantú, 2019).

We identify three important limitations of using supervised learning CNNs for image labeling. We then propose an unsupervised method that helps to addresses each limitation by generating "topic" clusters of visually similar images from larger corpora. The first limitation is that, in order to train a CNN, *researchers need to know ahead of time the particular objects or features they are looking for in the images.* Once trained, the algorithm automatically assigns the same universe of labels to other images, but it will only do so for the features it has been trained to predict.[1]

Commercial image tagging services offered by companies such as Amazon, Google, and Microsoft (likely CNNs) can predict a large number of image features (objects, people's gender and ethnicity, etc.). However, they do not offer complete lists of the features their CNNs have been trained to recognize.[2] Open source CNNs are a more transparent (and often

[1]For additional details on how CNNs work, see Webb Williams et al. (2020); Torres and Cantu (2020)

[2]Not to mention the known gender and race biases of these tools (Buolamwini and Gebru, 2018; Schwem-

cheaper) option. The categories of these CNNs are *known* (i.e. the 1,000 objects included in the ImageNet corpus) and the algorithms themselves are available from deep learning libraries in Python and R (such as Keras, PyTorch, and TensorFlow) or from scholars who have made their trained models accessible (i.e. Won et al. (2017); Kärkkäinen and Joo (2019)).

Researchers can also "fine tune" pre-trained models to recognize additional image features of interest (Webb Williams et al., 2020). But this assumes that they already know the image features in the data that are relevant to their project. For example, suppose that a researcher studying social movements is interested in symbols of collective identity that activate feelings of group belongingness (Tajfel, 1981), increase voter turnout (Gerber et al., 2008), influence attitudes on particular issues (Hassin et al., 2007), or inspire protest participation (Kharroub and Bas, 2015; Casas and Webb Williams, 2019). The scope of potentially relevant social symbols (flags, logos, hand gestures, etc.) may not be known ahead of time.

A second limitation is that *CNNs may not be very good at labeling images for complex theoretical constructs.* For example, existing research demonstrates that CNNs can successfully capture the emotions reflected in faces present in an image (Busso et al., 2004). However, they do less well in predicting the emotions viewers feel when viewing an image (Xu et al., 2014; Webb Williams et al., 2020). For instance, a flag image might evoke a feeling of collective identity in a viewer, not because of the flag *per se*, but because someone who looks like the viewer is holding the flag – a CNN would struggle to recognize this subtlety. As another example, a CNN may be less adept at recognizing instances of misinformation than a trained human fact-checker (e.g. images taken out of context, misattributed actions, etc.). In such cases, manual annotation may be the preferred option.

A final limitation is that at some point in an image labeling project, researchers *may be primarily interested in the human label generating process.* For example, they might want

mer et al., 2020).

to know whether particular groups of people (i.e. Republicans *versus* Democrats) have systematically differing emotional reactions to the same images[3] or it may be important to know whether some people are better at spotting instances of visual misinformation.

To address these limitations a researcher might draw and explore a random subset of images to create a comprehensive list of objects/features of interest for fine tuning a CNN; to manually label images for the presence of complex theoretical constructs; or to study how personal attributes affect responses to images. The drawback of a random sampling approach is that image datasets can be highly unbalanced. For example, in a social media dataset, a small number of "viral" messages and images may account for most of cases. If the goal is to study the differences between images that go viral and those that do not (for example), then a random sampling approach may omit much of the variation that is essential for the analysis.

We propose an unsupervised image clustering method for the purpose of generating stratified random samples from large image corpora. In line with the challenges we aim to address here, text analysis unsupervised clustering methods (topic models) are frequently used as discovery tools that help researchers develop content categories for supervised machine learning tasks (Grimmer and Stewart, 2013; Wilkerson and Casas, 2017). Researchers have also used unsupervised clustering to confirm that they have not missed important content categories (Grimmer and King, 2011), and to compare differences in responses to text across groups (Roberts et al., 2014).

In a text analysis, digitized documents are first tokenized into words. These words are typically the features used to group documents into "topics" or associate documents with with topics. There is no equivalent tokenizing method for images.[4] Rather than words,

[3]Research finds that images that evoke stronger emotional reactions are more likely to capture people's attention and influence their attitudes and behavior (Grabe and Bucy, 2009; Casas and Webb Williams, 2019).

[4]Though see (Torres, 2019) for a non-CNN approach to computer vision that treats patches of images as analogous to words.

digitized images are composed of highly complex three dimensional (red, blue, green) arrangements of pixels that vary in light intensity.[5]

The image features we use for clustering are called "embeddings." These embeddings reduce complex, 3-dimensional pixel-level information into, for example, 512 numeric values in a single indexed vector. In a CNN, embeddings are part of the information that is used to train the algorithm (by associating image embeddings with their manually assigned labels) and to automatically label new images (by using an image's embedding to predict its most likely label).

Our clustering pipeline begins by generating image embeddings using a pre-trained CNN. We then use an iterated variation of a widely used clustering algorithm (k-means) to cluster images based on these embeddings. These clusters can be thought of as "visual topics." As with topic models, this is an inductive process where the substantive meaning of each image cluster is open to interpretation. Similar approaches have been proposed in the computer vision literature (Celik, 2009; Yang et al., 2016). In those cases, however, the researchers assessed clustering performance by starting with labeled images from benchmark datasets. While this approach provides insights into the potential of unsupervised image clustering, it cannot be used to evaluate clustering performance when clustering images without labels *ex ante*. We therefore also include a method for validating the image cluster assignments. We strongly advise researchers using our image clustering method to build a similar validation set to evaluate their models.

As with topic modeling, image clustering requires a number of modeling decisions that can have implications for the output, including: the CNN algorithm used to generate embeddings; whether the CNN should be fine tuned using additional labeled examples; and how to identify the best clustering results. Much of this paper is devoted to how we made each of these

[5]The RBG representation is the most commonly used in computer vision, but images can also be represented in alternative forms, such as a single black and white matrix indicating the degree of gray in each pixel.

decisions for our example image corpus. At each step in the clustering pipeline, we describe the choices available to researchers when applying the method to their own data. Our validation procedure allows for tests to determine which hyperparameter choices are best for a given corpus.

Our working example is a large collection of images drawn from tweets using the hashtag #FamiliesBelongTogether. This hashtag references the Trump administration's policy of separating migrant children from their adult family guardians at the US-Mexican border. After clustering images from this corpus, we draw a stratified sample by randomly selecting images from each cluster. We then use this sample to explore the corpus and discover unanticipated images and image features. We also describe how we used this process to more efficiently manually annotate for a complex theoretical construct (evoked emotions). Finally, we describe how we used it to investigate whether similar images evoke different emotions among those that self-report as Republicans *versus* Democrats.

2 An Automated Visual Clustering Method

We propose a visual clustering method for large image corpora that can be used: (1) to explore the different kinds of images present in a dataset; (2) to facilitate manual annotation of complex visual features; and (3) to study whether and how annotator characteristics affect their labeling decisions. For the clustering pipeline, we first need to represent images numerically in order to be able to apply a clustering algorithm to the image corpus. As with all clustering methods, the pipeline also requires that we select the appropriate number of image clusters. Finally, we need to compare and select the hyperparmeter configuration that does the best job of clustering the images in our corpus. We illustrate each of these steps using a large example image corpus.

2.1 Example Image Dataset: #FamiliesBelongTogether

In autumn 2017, we initiated a substantial data collection effort to study online social movement dynamics. We gathered tweets from a large number of entities that are often the originators of social movement campaigns, including public affairs organizations and politicians. In addition, we simultaneously collected the tweets by anyone who used the mobilization hashtags these entities were using. This paper uses a small portion of the data - a #FamiliesBelongTogether corpus that contains 174,172 tweets and 88,075 images (18,096 of which are unique) that were collected from May 30th to October 27th 2018. Appendix A provides further details about the data collection process.

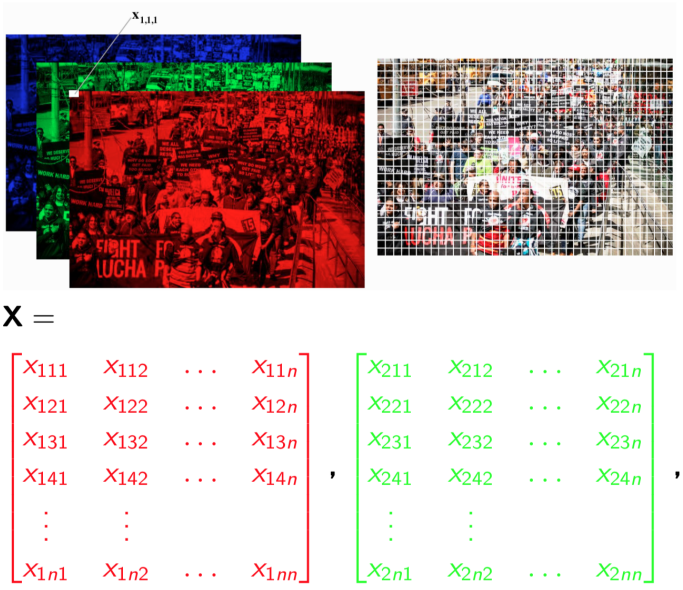
2.2 Converting Images to Embeddings

In a CNN, images are initially represented as three-dimensional matrices, where each matrix represents the intensity of red, green, and blue for a particular pixel (standardized values ranging from 0 to 255, see Figure 1).[6] CNNs next transform the information contained in these three dimensional matrices to flatter representations. The architecture of each model varies, but the output embeddings represent images in a denser and lower dimensional space (e.g. 512-size vectors). In the final step, the CNN computes the probability that an image is in fact a chair (or a horse, etc.) based upon the similarity of its embedding to the embeddings of different labeled examples (see Webb Williams et al. (2019) and Torres and Cantu (2020) for a more detailed overview of how CNNs work).

The single dimensional embeddings are much easier to work with computationally than three dimensional matrices. They also effectively capture thematic similarities between images prior to forcing the images into discrete classes. For example, a model trained to predict

[6] Although most computer vision algorithms use this three-color channel representation as input, some use other color representations, such as the degree of gray after transforming images to black and white (Cantú, 2019; Torres and Cantu, 2020).

Figure 1: An image represented as a 3-dimension input. Each $X_{i,j,z}$ unit contains information about the pixel-level intensity of red, green, and blue (respectively) in the image.



the 1,000 ImageNet classes has learned that two different kinds of chairs have a lot in common prior to the CNN's final step of assigning the images to their highest probability (chair) category. Put another way, image embeddings from trained CNNs carry more continuous meaning in and of themselves.

An easy way to generate embeddings is to pass images through a pre-trained CNN. However, the "meaning" carried in such embeddings is going to be shaped by the labeled examples used during the training process. For example, the embeddings from a CNN trained using Imagenet will probably do a good job of capturing object similarity. It may do less well at capturing the emotions that images evoke to the extent that two very different objects evoke similar emotions or that the same object in different contexts may evoke different emotions.

We selected a CNN trained on Imagenet (ResNet-18) as our starting CNN (He et al., 2015). We chose this CNN for two main reasons. First, the Imagenet categories that the CNN was trained to detect reflect some of the most basic objects that might be in images (this can

be helpful in grouping together images with similar elements in them); and second, because the size of the second-to-last fully connected layer to be used for generating image embeddings is not too large (512-size vector) compared to other available pre-trained CNNs,[7] facilitating computation at the subsequent clustering stage. These two features make ResNet-18 a useful general default option. Nevertheless, the validation strategy that we present below allows researchers to assess whether using other CNNs for generating image embeddings can yield results more suitable to their corpus..

We further opted to fine tune ResNet-18 to better reflect the goals of this project. Fine tuning can improve a model's ability to capture elements of interest to a researcher. For example, a CNN pre-trained using Imagenet can be updated to predict a different set of categories (or additional categories) by providing a relatively small set of labeled examples. Fine tuning works surprisingly well for many image-labeling objectives (Webb Williams et al., 2020). As a first step we had six research assistants annotate 609 unique images for whether they included features relevant to the #FamiliesBelongTogether movement. The features included: a) the emotions the image evoked (enthusiasm, anger, and anxiety), b) whether the image included people of different ethnicities (white, black, Asian, and Latino), c) gender (presence of males and/or females), d) the presence of children, e) symbols of collective identity, and f) whether the image communicated that the movement would be successful at accomplishing its goal. In all, the annotators labeled for the presence or absence of 12 features.

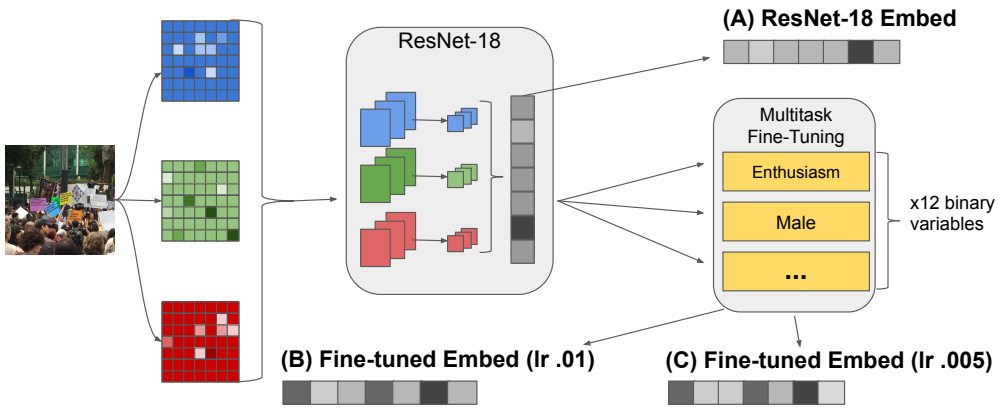
We then replaced the last fully connected layer of the CNN that predicted 1,000 classes with a layer predicting just 12 classes. We used the new labels for the 609 images to retrain the model weights for 50 additional iterations. Instead of using a softmax function that predicts the probability of each image belonging to a set of mutually exclusive classes

[7]For example, the second-to-last fully connected of the VGG models (Simonyan and Zisserman, 2014) is a 4,096-size vector.

(probabilities summing to 1), we used 12 sigmoid functions that predict the probability that an image included each of the 12 labeled, non-mutually exclusive items (a process known as multi-task learning, (Caruana, 1997)). We repeated the process for two models, one with a learning rate of .005 and the second with a learning rate of .01.

In the final step we passed all of our images through this fine-tuned model, extracting the embeddings from the second-to-last fully connected layer. As shown in Figure 2, we had three types of image embeddings to compare: from a pre-trained ResNet-18 model (*ResNet-18 Embed*), from a multitask fine-tuned ResNet-18 model using a .01 learning rate (*Fine-tuned Embed (lr .01)*), and from a fine-tuned ResNet-18 model using a .005 learning rate (*Fine-tuned Embed (lr .005)*).[8]

Figure 2: Three types of embeddings used for image clustering.



2.3 An Iterative Image Clustering Method

Next, we use these embeddings to cluster thematically similar images (Jain, 2010). Our initial experiment of fitting a single k-means clustering model to the full matrix of image embeddings did not work as well as we hoped (see Figure 7). For this reason, we developed the iterative clustering approach described below and outlined in Figure 3.

[8]These are two common learning rates used for fine tuning CNN for image recognition (Shin et al., 2016).

(1) We first fit several k-means models that vary by the number of clusters N (*Step Size*, e.g. +5 clusters) while measuring model fit. To speed up this fitting process, we use a random sample of the images (*Sample Size*, e.g. 1,000 images). We select k based on when average model fit stops improving across I runs of the model (*Converge Window*, e.g. 3 iterations). Averaging helps to ensure that the results are not distorted by an idiosyncratic result or by the number of clusters specified.

(2) We next fit the selected K-means model to the entire image dataset.

(3) We then assess the cohesiveness of each cluster using the average silhouette score for the images in the cluster. The silhouette score assesses the similarity of the images in a cluster and their distinctiveness vis-a-vis images in other clusters.[9]

(4) We then pull out clusters of images with silhouette scores above a specified threshold (*Similarity Threshold*, e.g. 0.4) from the dataset.

(5) We then repeat this process for the images that remain, and keep repeating it until the number of images falls below a threshold (e.g. 20 images left, *Stop Size*).

In the next section we experiment with different models and hyperparameter settings to determine which combinations work best for our project. We also demonstrate that iterative clustering is superior to using a single k-means model approach.

2.4 Validating Model and Hyperparameter Selection

To determine which modeling approach and hyperparameter configuration perform best for our study, we needed to create a gold standard validation set. The general objective was to have humans decide whether pairs of images belonged in the same cluster *based on our*

[9]Silhouette scores range from -1 to 1, 1 being the highest level of cohesiveness and uniqueness, and -1 being the lowest level of cohesiveness and distinctiveness.

Figure 3: Pseudo Code of the Iterative Clustering Method

X = input image matrix (e.g. 20,000 images × 512-size embeddings)
1. Find number of K clusters to fit
   (a) Randomly sample 1,000 [*Sample Size*] images from X
   (b) Iterate through new values of K (increase K by 5 [*Step Size*] each time)
      - fit k-means algorithm (for K clusters)
      - check average goodness of fit for the last e.g. 3 iterations [*Convergence Window*]
      - stop if goodnes of fit does not improve OR if $K >$ images in X. Continue otherwise.
2. Fit k-means algorithm to X predicting K image clusters
3. Calculate intra-cluster similarity (silhouette score)
4. Find cohesive clusters (cluster silhouette score $>$ e.g. 0.04 [*Similarity Threshold*]
5. Separate out from X the images from clusters found to be cohesive in step 4.
   - If still more than e.g. 20 images in X [*Stop Size*]: STOP. Otherwise, run another iteration.

theoretically-informed constructs of interest. We then use this information to compare different clustering approaches (models and hyperparameters) by assessing precision (the proportion of pairs correctly predicted as belonging to the same cluster) and recall (the proportion of pairs labeled as belonging to the same cluster that were accurately classified as such).

The #FamiliesBelongTogether corpus includes a large number of similar images (such as images of street protests). To produce a more balanced validation set, we first ran the images through an 80-cluster k-means algorithm.[10] We then randomly sampled images (554 pairs in total) from these 80 clusters. Annotators viewed the pairs and determined whether the images should indeed be in the same cluster or not. Annotator agreement as to whether the image pairs were drawn from the same true cluster was good.[11] We emphasize that researchers wishing to use our clustering pipeline should follow this procedure to build their own validation set.

We then performed a grid-search of different model and hyperparameter combinations, using the validation set to compare performance. Recall that we generated embeddings using three different models, and that five hyperparameters settings must be specified as

[10] A preliminary analysis where we fit k-means algorithms to the entire corpus ranging from 30 to 300 clusters suggested that, after 80 clusters, the goodness of fit only improved marginally.

[11] 87% agreement between annotators and a Cohen's kappa value of 0.64

part of the iterative clustering process. Our grid search included (for each of the three model embedding inputs) all possible combinations of the following hyperparameter settings, {3, 5} for step size, {0.0, 0.01, 0.02, 0.03, 0.04, 0.05} for similarity threshold, for a constant sample size of {1,000}, a stop size of {20}, and a convergence window of {3}.[12]

Figure 4 uses a toy example to illustrate why high precision and recall imply that we are successfully clustering images. High precision and recall indicate that the clustering algorithm is agreeing with the human annotators both with respect to when images belong together and when they do not. This agreement is more likely to occur when there are fewer clusters. High recall and low precision, in contrast, indicate that the clustering algorithm is not successfully discriminating between different types of images. High precision and low recall, in contrast, indicate that it is not always placing similar images in the same cluster. This is more likely to occur when there are more clusters.

Figure 5 reports precision and recall for each of the clustering models from the grid search (based on a 5-fold cross-validation). Models in the upper right are stronger in terms of both precision and recall. The model represented by the blue cross in the circle, for example, uses the embeddings of the fine-tuned ResNet-18 model with a learning rate of 0.005, and a clustering similarity threshold of 0.03.

Figure 6 provides additional information about this model. It took 33 iterations to to reach the stop value of 20 images remaining. This produced a total of 309 image clusters. The top panel in Figure 6 presents the number of images remaining to be classified at each iteration. Most of the images were assigned to cohesive clusters in the first 10 iterations. The center panel presents the number of clusters to be fit at each iteration based on the specified goodness of fit test. This number remained quite constant, around 55 (light gray bars). The orange bars indicate how many of these clusters were considered cohesive after

[12] In a preliminary analysis where we explored (in a more manual and less systematic fashion) wider ranges of hyperparameter values, we observed that exploring other values either did not make much of a difference or clearly yielded unsatisfactory results.

Figure 4: Precision and recall for 3 possible ways of clustering 5 images known to be of category A and 5 images known to belong to category B.

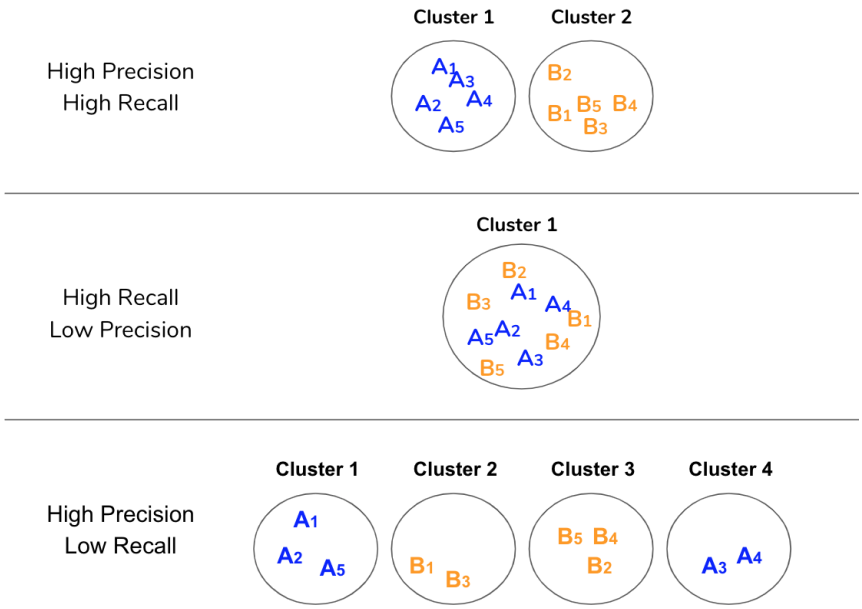
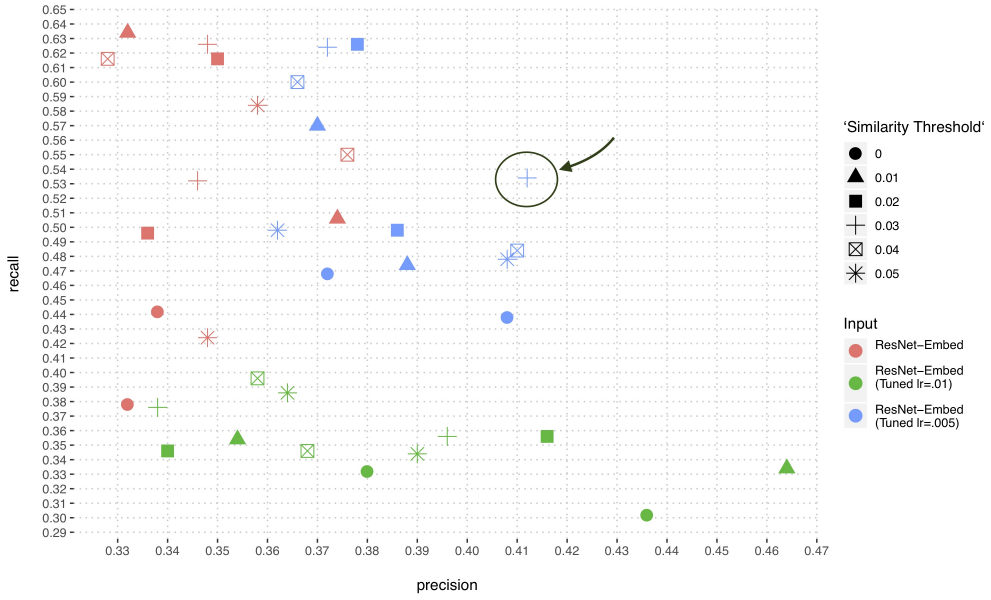


Figure 5: Performance of Different Hyperparameter Configurations and Types of Image Embeddings



discriminating based on the silhouette score and the similarity threshold. We observe fewer cohesive clusters in the final iterations. Finally, in the bottom panel we report the average

number of images grouped in the cohesive clusters found at each iteration. We see that in the first iteration the algorithm found some large groups of images: the cohesive clusters found in the first iteration had an average of about 500 images, approximately. The algorithm finds relatively large clusters until the sixth iteration. The clusters found in later iterations are much smaller, which as we see in the top panel, is also a function of fewer images left to be clustered.

Figure 6: Information about the iterative progress of the best performing model

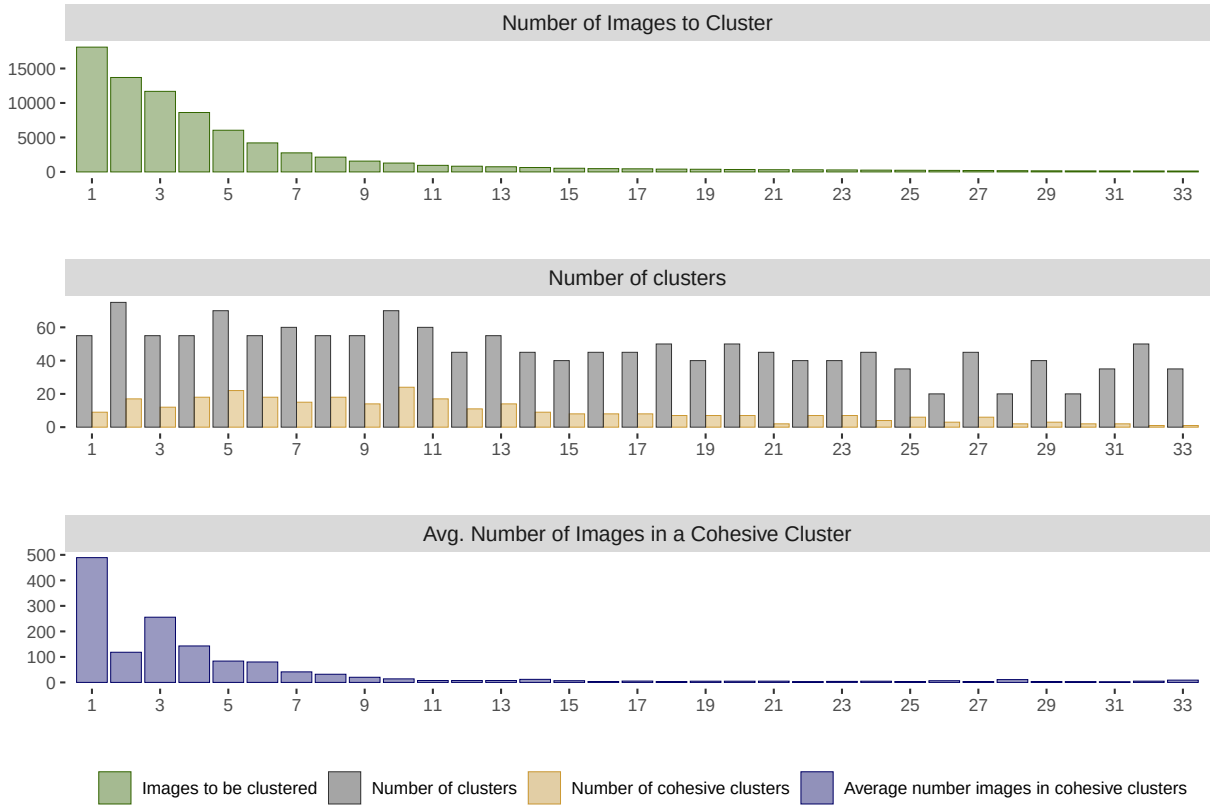
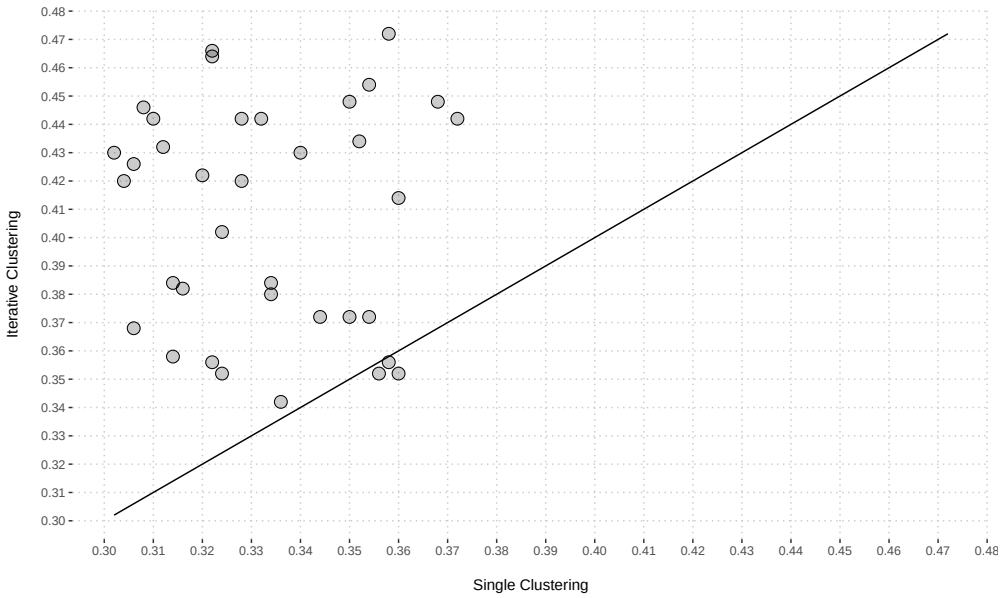


Figure 7 confirms that practically all of the iterative clustering models from the grid search outperform a one-shot k-means clustering approach when the average of precision and recall (the F-score) is used to measure model fit.

Figure 7: Performance (F-Score) of Single *versus* Iterative K-Means Image Clustering



3 Corpus exploration

Figure 8 illustrates how clustering images using embeddings can help researchers explore and possibly identify unexpected features that are relevant to their theoretical mechanisms of interest. As discussed, symbols of collective identity can help draw support for a social movement (Polletta and Jasper, 2001; Tajfel, 1981; Van Zomeren et al., 2008). Row 12 reveals a cluster of food images shared using #FamiliesBelongTogether. Food is a cultural element that can be part of people's identity (Caplan et al., 1997). A researcher interested in annotating this dataset for collective symbology might not have initially identified food images as a category.

For researchers interested in framing (e.g. Torres (2019)), several clusters include people protesting on the street (such as rows 2, 3, 4, and 5), images of children behind fences (row 18), children in cages (row 8), and images portraying immigration detention centers as Nazi concentration camps or slave plantations (row 13). A pre-trained CNN would certainly identify children but it would almost certainly not identify children in cages as a unique

category. Having an accessible method for visually exploring the images in a large corpus can be an effective way to discover such contextual information. The knowledge gained can then be used to fine tune a pre-trained CNN to more closely match the purposes of the framing study.

Or suppose that a researcher is interested in whether conservatives are more likely to use certain frames than liberals. In this case the researcher could label each of the 309 clusters for relevant frames and associate the labels with each of the tweets in each cluster. Then, using Barbera (2015)'s method for estimating the ideology of Twitter users, the researcher could easily test whether ideology predicts frame usage.

3.1 Constructing a stratified random sample

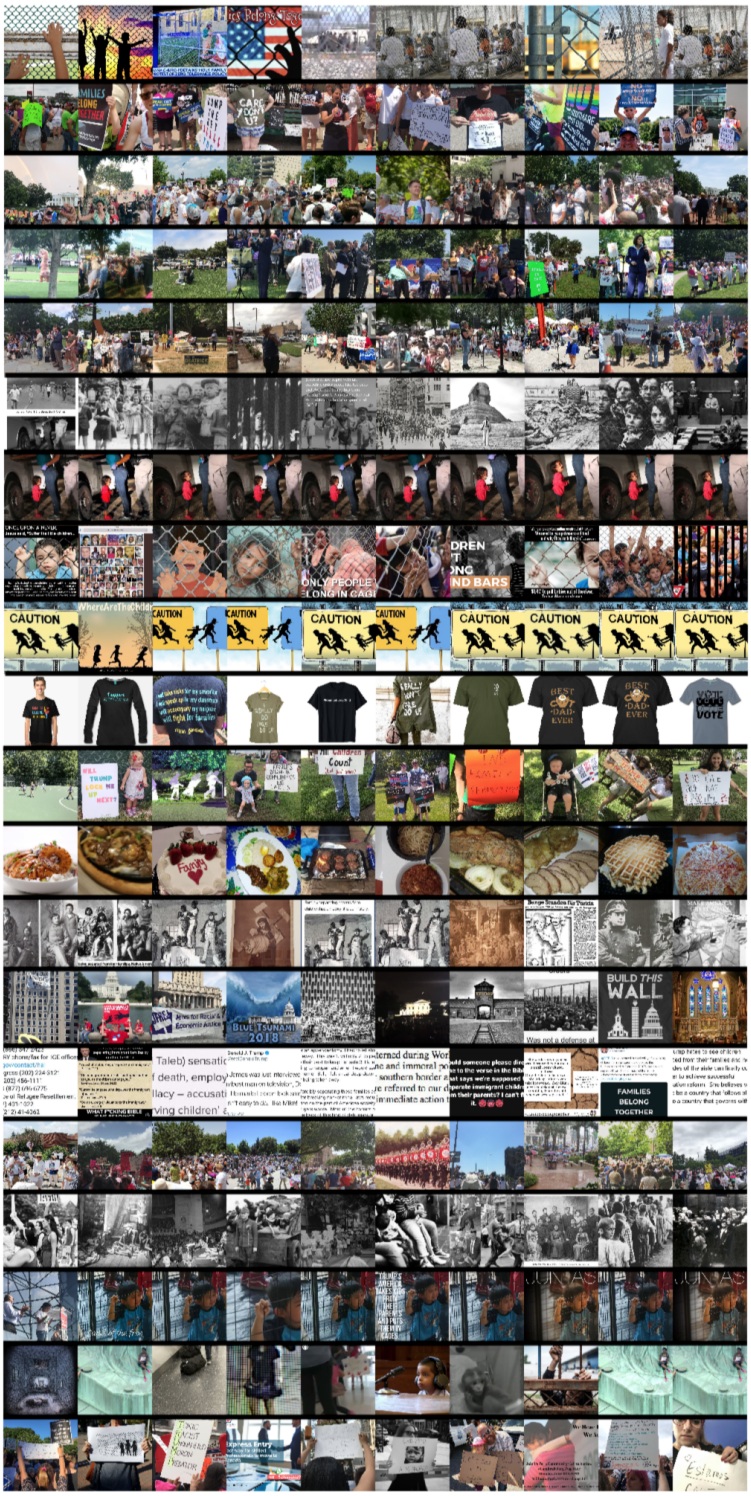
In a perfectly balanced dataset of 309 classes, each class would account for about 0.3% of the images. Figure 6 demonstrates that our corpus of #FamiliesBelongTogether images is unbalanced. One cluster, for example, accounts for 5.0% of the images. As discussed, a random sample of this corpus could be problematic depending on the research project (such as comparing images that do and do not go viral). To construct a stratified sample from the clustering results, we randomly select 5 images from each cluster that includes more than 6 images, and $N - 1$ images from clusters with 2 to 5 images.[13] This produces a stratified sample of 1,087 images.

3.2 Labeling the sample for evoked emotions

CNNs do not currently perform well in predicting emotional reactions to images (Xu et al., 2014; Webb Williams et al., 2020). A researcher interested in studying, for instance, whether images that evoke anger or happiness (e.g.) are more likely to mobilize people will probably

[13]Five images provide us with significant diversity of images from each cluster, while also not creating an unbalanced sample of images that are biased towards larger clusters.

Figure 8: Sample images from 20 random clusters



have to turn to manual annotation to measure this complex concept. A stratified sample ensures that we obtain information about the emotions evoked by a wider set of images.

To illustrate this, we asked human annotators to rate the images in our stratified sample for 10 emotions known to load onto three distinguishable emotional dimensions (Marcus et al., 2017): *Enthusiasm* (hopeful, enthusiastic, proud); *Aversion* (angry, resentful, bitter, hateful); and *Anxiety* (worried, scared, and afraid) (Marcus et al., 2000). The annotators self-reported as either Democrats (N = 360) or Republicans (N = 360) were recruited via the polling company Qualtrics. We asked each respondent to rate 8 of the images for each of the 10 emotions (on a 1-10 scale). Every image in the sample was labeled by one Republican and one Democrat. No respondent labeled more than one image from the same cluster.

Figure 9: Reported emotional reactions to images from different clusters

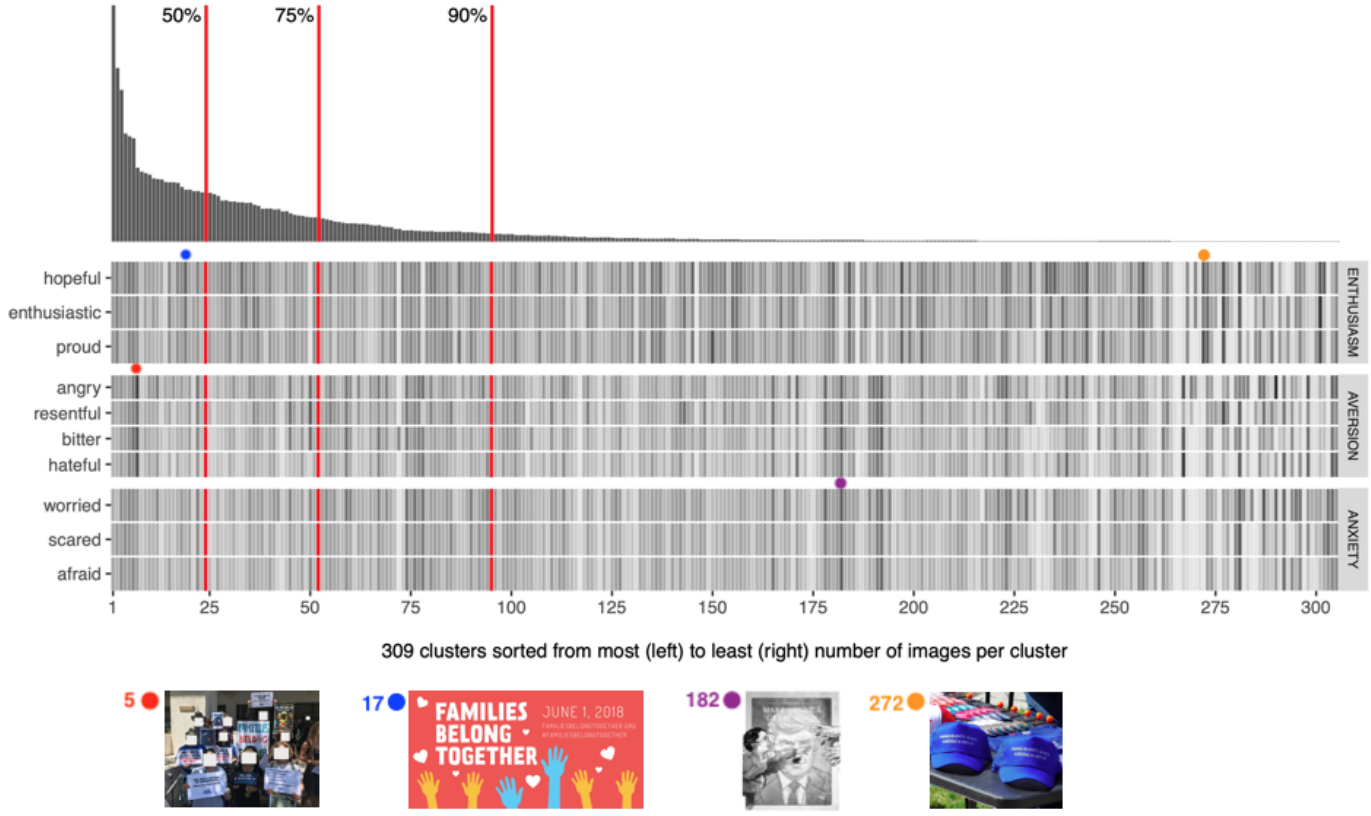


Figure 9 presents our findings and demonstrates how much important emotional variation

we would miss if we took a random selection of images instead of the stratified random sample from clusters. Each column in Figure 9 is one of the 309 clusters, with the largest clusters to the left. The red vertical lines indicate the total number of images accounted for by the clusters to the left (50%, 75% and 90% of all images). The cells in each column correspond to the average emotional ratings for a cluster for all respondents. Darker shading indicates a more intense emotional response.

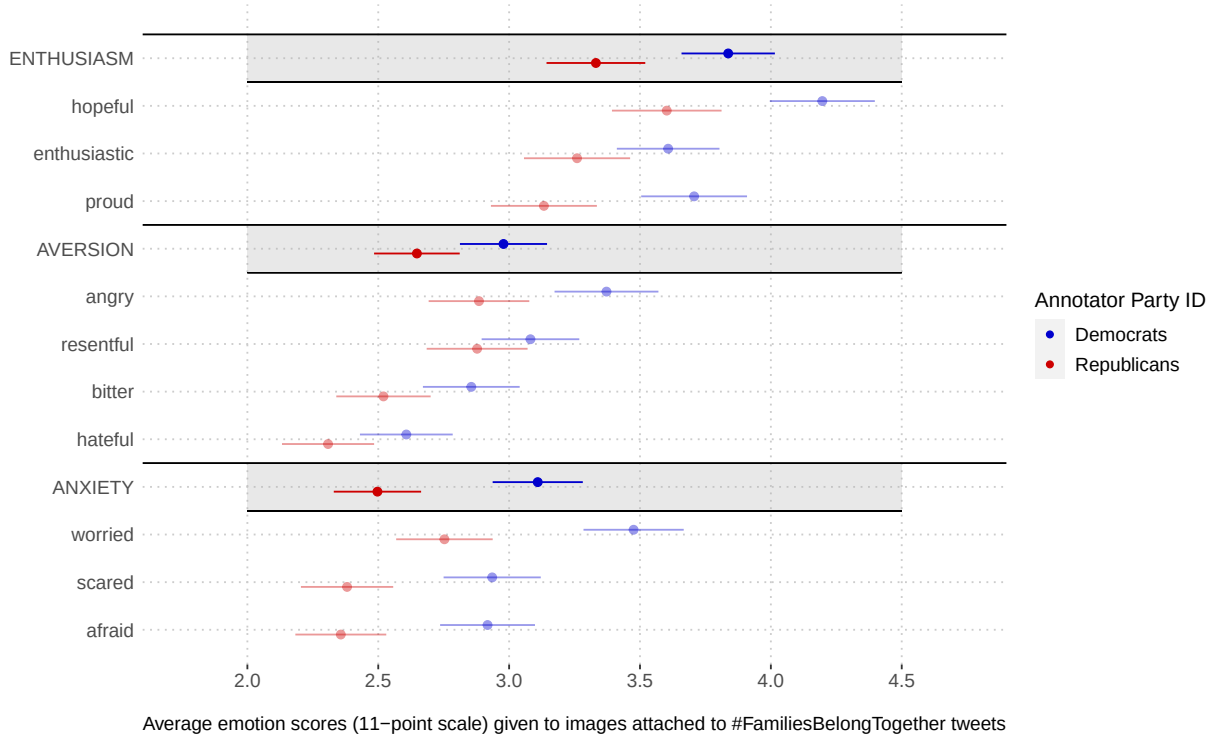
We selected some examples to illustrate the analytical value of our approach. For example, the red dot in the figure points to the 5th largest cluster, a set of images with people holding protest signs. On average, the images in this large cluster scored quite high on all emotions, but particularly on those emotions belonging to the aversion dimension. In contrast, images in the 17th largest cluster (just to the left of 50% mark, blue dot) evoked more intense "enthusiasm" responses. These are images with posters announcing future mobilizations. Importantly, the images in some of the smallest clusters (to the right of the 90% mark) also generated strong emotions. These would have been mostly ignored if a random (rather than a stratified random) sample would had been used. For example, we observe the images in one of these much smaller clusters (182nd, purple dot) to have generated high aversion and anxiety (and no enthusiasm). These are slightly different version of the same image, showing a migrant child being separated from her family, with a picture of a Hitler-looking President Trump in the background. As a final example, we point to another small cluster (272nd, orange dot) that respondents rated as generating enthusiasm (and no aversion nor anxiety): images of MAGA-type hats but with the motto "Immigrants Make America Great Again".

3.3 Partisan differences in evoked emotions

Finally, we draw on the emotional intensity ratings from Figure 9 to compare average differences between Republican and Democrats. In Figure 10 the non-gray areas average across

clusters for the individual emotions. The gray areas do the same thing, except that they average across the emotions within each of the three affective dimensions. Once again, the stratified sampling approach means that the averages weight all of the clusters equally, whereas the averages from a random sample would have been dominated by results from a relatively small number of large clusters.

Figure 10: Average emotion reactions, by Democrats and Republicans, to a stratified random sample of 1,087 images shared on Twitter with the #FamiliesBelongTogether hashtag



Where #FamiliesBelongTogether images were concerned, Democrats were more likely to respond with enthusiastism ("hopeful" and "proud") and anxiety ("worried", "scared" and "afraid"). According to the Affective Intelligence Model, these emotions are strongly correlated with political action (Marcus et al., 2000; Casas and Webb Williams, 2019), potentially indicating that the images associated with them will be effective at mobilizing Democrats more than Republicans. Democrats for example felt much more enthusiasm than Repub-

licans when exposed to the images of protesters in the 5th cluster (red dot) in Figure 9.[14] Democrats also felt more anxious when exposed to the cluster with images of children in cages (row 1 in Figure 8).[15] Finally, although on average Democrats were more likely to express anger in response to #FamiliesBelongTogether images, overall, respondents in both parties reported similar degrees of resentment, bitterness, and hate.

4 Discussion

Compared to text and audio, images can be an especially effective form of political communication. There are now many options for automatically labeling images for many features of interest (such as protests, famous people or facial expressions). These developments have inspired an increasing number of social scientists to conduct large-N image analyses. This paper addresses three emerging needs of image scholarship. The first is a feasible and representative way of visually exploring image corpora. The second is to support manual annotation projects when automated approaches are incapable of annotating images for the presence of complex theoretical mechanisms. The third is to support projects where the primary goal is to study human annotation decisions. We demonstrate how unsupervised image clustering can help researchers address each of these needs when faced with large and possibly unbalanced image corpora. We illustrate the method using a corpus of images shared by Twitter users along with the hashtag #FamiliesBelongTogether.

The approach we propose will benefit researchers studying image data from a descriptive perspective, as an explanatory variable, and/or as an outcome variable. Unsupervised clustering facilitates exploration of large image corpora. As with topic models for text, clustering can assist with labeling by helping researchers discover unanticipated image themes. Clustering can also be used to create the more balanced samples that are necessary for testing

[14] Hopeful: 6.55 (D) v. 2.25 (R). Enthusiasm: 7.25 (D) v. 3.25 (R). Proud: 6.5 (D) v. 2.5 (R).
[15] Afraid: 8 (D) v. 3.5 (R). Scared: 8 (D) v. 3.5 (R). Worried: 8 (D) v. 4.5 (R).

relevant questions, such as what differentiates images that are widely shared through social media from those that are not. And for projects where manual annotation is the only feasible option (such as when the goal is to label images for the emotions they evoke), clustering can speed up the labeling process by organizing similar images together, guided in part by the researchers' goals (using fine-tuned CNN embeddings as features), enabling annotators to work through them more efficiently.

As with text topic models, in an unsupervised set up there are no gold standard labels for selecting the best image clustering model (that would defeat the point). Our selection methodology relies on silhouette scores that consider the cluster cohesiveness and distinctiveness, similar to what topic modelers recommend for choosing a model (Roberts et al., 2014). However, we also use a purpose-built validation test to demonstrate that an iterative clustering approach yielded consistently superior results when compared to a single-shot clustering approach. We also used that gold standard validation test to compare different feature selections (embeddings from CNNs with and without fine tuning, and with different learning rates) and model hyperparameters (Figure 3). Using precision and recall to measure performance, we were able to identify which combination produced the best clustering results for our image corpus.

Unsupervised clustering, and the approach we have proposed for applying it to address specific limitations of automated image labeling, represents an important and useful advance in the emerging images-as-data literature in the social sciences. There is also much more to explore. We have focused on the best performing approach for our project. We do not know whether this approach would prove to be the best for another corpus and research objective. Each project will require specific configuration and hyperparameter decisions, as we have detailed. For example, we might have generated embeddings using a different CNN (e.g. a ResNet with more layers, VGG, AlexNet, etc.). We also tested only a limited number of possible learning rate and hyperparameter combinations.

We hope that unsupervised clustering encourages more scholars to explore image data in their areas of expertise, leading to new hypotheses to test and new data to explore. Most current images-as-data studies focus on digitized images shared on social media and questions about visual communication (with notable exceptions using image data to examine voter fraud and remote-sensing imagery). Unsupervised clustering methods may also expand image analysis to many more fields of interest. Other interesting applications include evaluating Google Street View images to estimate gentrification and ethnic diversity (Hwang and Sampson, 2014) and categorizing children's drawings of politicians (Holman et al., 2020). New images-as-data methods widen the scope of what is possible in this growing subfield.

References

Barbera, Pablo. Birds of the Same Feather Tweet Together. Bayesian Ideal Point Estimation Using Twitter Data. Political Analysis, pages 1–16, 2015.

Buolamwini, Joy and Timnit Gebru. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In Proceedings of Machine Learning Research, volume 81, pages 1–15, 2018.

Busso, Carlos, Zhigang Deng, Serdar Yildirim, Murtaza Bulut, Chul Min Lee, Abe Kazemzadeh, Sungbok Lee, Ulrich Neumann, and Shrikanth Narayanan. Analysis of emotion recognition using facial expressions, speech and multimodal information. pages 205–211, 01 2004.

Cantú, Francisco. The Fingerprints of Fraud: Evidence from Mexico's 1988 Presidential Election. American Political Science Review, 2019.

Caplan, Patricia et al. Food, health, and identity. Psychology Press, 1997.

Caruana, Rich. Multitask learning. Machine Learning, 28:41–75, 1997.

Casas, Andreu and Nora Webb Williams. Images that matter: Online protests and the mobilizing role of pictures. Political Research Quarterly, 72(2):360–375, 2019.

Celik, Turgay. Unsupervised change detection in satellite images using principal component analysis and k-means clustering. IEEE Geoscience and Remote Sensing Letters, 6(4): 772–776, 2009.

Gerber, Alan S., Donald P. Green, and Christopher W. Larimer. Social pressure and voter turnout: Evidence from a large-scale field experiment. American Political Science Review, 102(1):33–48, 2008.

Grabe, Maria Elizabeth and Erik Page Bucy. Image bite politics: News and the visual framing of elections. Oxford University Press, 2009.

Grimmer, J and B M Stewart. Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts. Political Analysis, 21(3):267–297, jul 2013.

Grimmer, Justin and Gary King. General purpose computer-assisted clustering and conceptualization. Proceedings of the National Academy of Sciences, 108(7):2643–2650, 2011.

Hassin, Ran R., Melissa J. Ferguson, Daniella Shidlovski, and Tamar Gross. Subliminal exposure to national flags affects political thought and behavior. P roceedings of the National Academy of Sciences, 104(50):19757–19761, 2007.

He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In arXiv:1512.03385, 2015.

Holman, Mirya R., J. Celeste Lay, Jill Greenlee, Zoe Oxley, and Angela Bos. Partisanship on the Playground: Expressive Party Politics among Children. 2020.

Hwang, Jackelyn and Robert J. Sampson. Divergent pathways of gentrification: Racial inequality and the social order of renewal in chicago neighborhoods. American Sociological Review, 79(4):726–751, 2014.

Jain, Anil K. Data clustering: 50 years beyond k-means. Pattern Recognition Letters, 31 (8):651 – 666, 2010.

Jean, Neal, Marshall Burke, Michael Xie, W Matthew Davis, David B Lobell, and Stefano Ermon. Combining satellite imagery and machine learning to predict poverty. Science, 353(6301):790–4, 2016.

Joo, Jungseock and Zachary C. Steinert-Threlkeld. Image as data: Automated visual content analysis for political science, 2018.

Joo, Jungseock, Erik P. Bucy, and Claudia Seidel. Automated Coding of Televised Leader Displays: Detecting Nonverbal Political Behavior with Computer Vision and Deep Learning. International Journal of Communication, 13:4044–4066, 2019.

Kärkkäinen, Kimmo and Jungseock Joo. FairFace: Face Attribute Dataset for Balanced Race, Gender, and Age. CoRR, abs/1908.0, 2019.

Kharroub, Tamara and Ozen Bas. Social media and protests: An examination of Twitter images of the 2011 Egyptian revolution. New Media & Society, feb 2015.

Marcus, George E., W. Russell Neuman, and Michael MacKuen. Affective Intelligence and Political Judgement. University of Chicago Press, Chicago and London, 2000.

Marcus, George E., W. Russell Neuman, and Michael B. MacKuen. Measuring emotional response: Comparing alternative approaches to measurement. Political Science Research and Methods, 5(4):733754, 2017.

Peng, Yilang. Same Candidates, Different Faces: Uncovering Media Bias in Visual Portrayals of Presidential Candidates with Computer Vision. Journal of Communication, 68(5):920–941, 2018.

Polletta, Francesca and James M Jasper. Collective identity and social movements. Annual review of Sociology, 27(1):283–305, 2001.

Roberts, Margaret E, Brandon M Stewart, Dustin Tingley, Christopher Lucas, Jetson Leder-Luis, Shana Kushner Gadarian, Bethany Albertson, and David G Rand. Structural topic models for open-ended survey responses. American Journal of Political Science, 58(4): 1064–1082, 2014.

Schwemmer, Carsten, Carly Knight, Emily D. Bello-Pardo, Stan Oklobdzija, Martijn Schoonvelde, and Jeffrey W. Lockhart. Diagnosing gender bias in image recognition systems. Socius, 6:2378023120967171, 2020.

Shin, Hoo-Chang, Holger R Roth, Mingchen Gao, Le Lu, Ziyue Xu, Isabella Nogues, Jianhua Yao, Daniel Mollura, and Ronald M Summers. Deep convolutional neural networks for computer-aided detection: Cnn architectures, dataset characteristics and transfer learning. IEEE transactions on medical imaging, 35(5):1285–1298, 2016.

Simonyan, Karen and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556, 2014.

Sobolev, Anton, Keith Chen, Jungseock Joo, and Zachary C. Steinert-Threlkeld. News and geolocated social media accurately measure protest size. American Political Science Review, 2020.

Tajfel, Henri. Human Groups and Social Categories: Studies in Social Psychology. Cambridge University Press, Cambridge; New York, 1981.

Torres, Michelle. Give Me the Full Picture: Using Computer Vision to Understand Visual Frames and Political Communication. 2019.

Torres, Michelle and Francisco Cantu. Learning to see: Convolutional neural networks for the analysis of social science data. Political Analysis, 2020.

Van Zomeren, Martijn, Russell Spears, and Colin Wayne Leach. Exploring psychological mechanisms of collective action: Does relevance of group identity influence how people cope with collective disadvantage? British Journal of Social Psychology, 47(2):353–372, 2008.

Webb Williams, Nora, Andreu Casas, and John Wilkerson. An Introduction to Images as Data for Social Science Research: Convolutional Neural Nets for Image C Cambridge University Press, New York, NY, 2019.

Webb Williams, Nora, Andreu Casas, and John D. Wilkerson. Images as Data for Social Science Research: An Introduction to Convolutional Neural Nets for Image Classification. Elements in Quantitative and Computational Methods for the Social Sciences. Cambridge University Press, 2020.

Wilkerson, John and Andreu Casas. Large-scale computerized text analysis in political science: Opportunities and challenges. Annual Review of Political Science, 20(1):529–544, 2017.

Won, Donghyeon, Zachary C. Steinert-Threlkeld, and Jungseock Joo. Protest Activity Detection and Perceived Violence Estimation from Social Media Images. In Proceedings of the 25th ACM International Conference on Multimedia, 2017.

Xu, Can, Suleyman Cetintas, Kuang-Chih Lee, and Li-Jia Li. Visual sentiment prediction with deep convolutional neural networks. 11 2014.

Yang, Jianwei, Devi Parikh, and Dhruv Batra. Joint unsupervised learning of deep representations and image clusters. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 5147–5156, 2016.

Appendix A Data Collection

We collected tweets that contain mobilizing hashtags used by public affairs organizations, news media, and legislators (all based in the USA) from January 1st, 2018 to December 31st, 2018. We collected all tweets from 1,144 American public affairs organizations, 534 American political actors, and 30 American news media outlets. We tracked posts on Twitter because this social media platform has become an effective mobilizing tool for social organizations and political messaging in the United States and is one of the most open social media platforms in terms of data sharing.

The 1,144 American public affairs organizations we tracked were identified using the 56th edition of the *Encyclopedia of Associations National Organizations of the United States* (EoA), published in 2017. The EoA contains information on roughly 23,000 organizations, but after limiting our list to organizations in the "Public Affairs" subject category and manually removing inactive or deleted Twitter accounts we built a list of 1,144 Twitter accounts to track (74.7% of the total population of EoA Public Affairs associations). In addition, we supplemented the EoA list of Twitter accounts with the official accounts from thirty of the most prominent news organizations in the United States and from every member of the 115th United States Congress. For news media outlets, we referenced numerous lists of the most watched and read news organizations and the most Tweeted news organizations. We tracked 30 media accounts and 434 accounts from U.S. Representatives and 100 accounts from U.S. Senators (some U.S. Representatives did not have Twitter accounts). This left us with Twitter accounts from 1,144 American public affairs organizations, 534 American political actors, and 30 American news media outlets. Full information on the Twitter accounts that we tracked are available upon request.

Between January 1st, 2018 and December 31st, 2018, we collected all tweets produced by these tracked accounts. At the end of each day we pulled a list of hashtags that were used more than twice by the same tracked account. Hashtags are a means of organizing on Twitter, so by monitoring this feature, we can identify mobilization attempts by these accounts. From each daily list of hashtags, we removed all hashtags that do not have a capitalization in the middle or were shorter than 12 characters. These requirements ensured we were tracking unique hashtags that were being used prominently by at least one of these organizations. Once added to our list of hashtags, we immediately began collecting any tweets by any Twitter user that used that hashtag. Each hashtag was tracked for two days and then removed if it was not used more than twice by the same organization over the next two days. Due to the sheer number of posted tweets, we were unable to collect all the tweets using our tracked hashtags, but we were able collect around 1,000,000 tweets per day from an average of 600 hashtags. This process populated a database of tweets that used any of our tracked hashtags on the days we were tracking that hashtag.

In total, we have roughly 4 million tweets from our initial tracked accounts and around 400 million tweets collected by tracking hashtags. In addition to data from each tweet such as the tweet text, count of retweets/favorites, count of account followers, and count of friends, we also collected any pictures posted with the tweets. We did not collect videos, but we did collect the image displayed on the tweet representing the video (the thumbnail image).