

Automated Visual Clustering for Image Corpus Exploration, Purposive 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 voice, images can be an especially effective form of political communication. It is now relatively easy to automatically label images for many categories 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. The first is that researchers may want to visually explore an image corpus to discern patterns before they begin assigning labels. This can be difficult with very large corpora. The second is that researchers may want to work with a purposive sample of their images when working with unbalanced datasets. The third is that researchers may be interested in annotating their images for highly subjective categories that cannot be easily assigned using automated supervised methods. We demonstrate how unsupervised image clustering can help researchers to address these needs more efficiently. Our substantive focus is on exploring, sampling and labeling a large corpus of images shared by Twitter users along with the hashtag #FamiliesBelongTogether.

*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

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

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

1 Introduction

Visual information has become an increasingly important form of communication due to the rise of digital media and mobile communications (Kharroub and Bas, 2015; Casas and Webb Williams, 2018). Visuals are also an especially effective form of communication when compared to text and voice. People are more likely to notice visuals, recall them, and be moved by them (Nelson et al., 1976; Grabe and Bucy, 2009). They deserve more attention from scholars studying subjects where visuals are of particular relevance, such as agenda setting, issue framing, public opinion formation and social movements.

Until recently, the high costs of labeling images for content discouraged all but the smallest projects. Computer vision methods now make it much easier to assign labels to thousands or even millions of images. Any social scientist with an image dataset can take advantage of state of the art computer vision algorithms (convolutional neural nets or CNNs) through on-line commercial image labeling services offered by Amazon, Google, Microsoft and others. Scholars with modest programming skills have a number of open-source (free) pre-trained algorithms that they can easily access through deep learning libraries such as TensorFlow, Keras, or PyTorch. Scholars can use these existing algorithms to predict the labels they have been trained to predict (for example, the 1,000 categories/objects in the ImageNet corpus), or they can also fine-tune a the algorithm to assign new labels (typically with a much smaller number of training examples) (Won et al., 2017; Webb Williams et al., 2019; Zhang and Pan, 2019; Sobolev et al., 2020).

But there are still situations where off the shelf computer vision methods do not fully serve the needs of social science researchers. We highlight three of these and show how unsupervised image clustering methods can help to address them. The first is when a researcher does not know, *ex ante*, what categories of an existing image classification system are relevant for a project. For example, they may be interested in the symbols of collective

identity being used in protest images. For a small project they can simply look at images from those protests and then train supervised machine learning classifiers capable of finding those elements in a large image corpus. However for a large project that includes thousands or millions of images, discovering the relevant categories can be much more challenging.

The second situation is when a researcher wants to work with a sample of images from an unbalanced corpus, with a particular type of image dominating the entire corpus. If they want to take into account the full range of images in their analysis, a method for purposive sampling is needed.

Finally, a researcher may want to collect information that cannot be obtained using existing computer vision tools, for as many of their images as possible. For example, existing algorithms are very good at labeling for objective features, such as vehicles, celebrities, or facial expressions. Scholars are presumably also interested in subjective reactions to images, and the impact of those reactions on behavior and attitudes. For example, the emotions images evoke on people, whether people believe an image is real or fake, and the leadership traits people attach to candidates based on their appearance. There has been much less work on designing automatic methods for predicting these types of subjective reactions to images. Fine tuning existing image classification models does not appear to perform well at for example predicting the emotions that large datasets of images evoke to people ([Webb Williams et al., 2019](#)). The only practical alternative may be manual annotation, but it can be prohibitively expensive and time consuming, specially given that different groups of people may have different subjective reactions and so the same image may need to be labeled multiple times. Purposive sampling can help reduce the costs but the researcher only obtains information for a fraction of the images. Label propagation can help increase the number of annotated images while keeping the costs low.

We show that unsupervised image clustering provides a way to more easily explore a large corpus of images. In this respect it serves a purpose similar to unsupervised topic modeling,

which was originally developed to enable researchers to explore and discover common themes within large text corpora. Unsupervised image clustering can also be used for purposive sampling and to substantially reduce manual image annotation costs in situations where automated image labeling is not an option. Once again, there is a parallel in topic modeling research where discovered topics often serve as the starting point for developing more rigorous classification systems.

For a textual analysis, co-occurring words are the features used to cluster documents into topics. Images, in contrast, are comprised of three dimensional (red, blue, green) collections of pixels. How can this pixel information be used to meaningfully cluster images? One option might be to use off the shelf tools to classify images before clustering them. But that puts the cart before the horse because a project goal may to be discover new image clusters that transcend existing classification systems.

Our approach uses the embeddings of a supervised deep learning convolutional neural net. In supervised image analysis, embeddings are reduced multidimensional pixel information that are used to assign images to classes. We use the same penultimate information to cluster images (without forcing them into a limited number of pre-determined classes).

A researcher can then draw random examples of images from each of the clusters to explore a large corpus and to obtain a purposive sample. The intuition for using unsupervised image clustering to assist manual annotation is similarly straightforward. Clusters represent thematically similar images. Instead of having to label these images individually, the annotator labels a small sample. Those labels are then propagated to all of the images in the cluster. We explore the conditions under which this labeling short cut can produce accurate results at much lower cost.

To demonstrate the method we investigate emotional responses to images posted on Twitter in response to the Trump Administration’s policy of separating migrant children from their adult family guardians at the US-Mexican border (images shared with the #Families-

BelongTogether hashtag). We first show how unsupervised image clustering facilitates the exploration of a large and rich image dataset (88,075 total images, including 18,096 unique images). We then extract and annotate a purposive sample for the emotions they evoke on people (evoked emotions have been found to be predictive of the diffusion of social movements online), accounting for the fact that viewers of different political ideologies (Republicans *v.* Democrats) can react differently to the same images. Finally, we assess the extent to which the annotations of this purposive sample can be propagated to the rest of unlabeled images in the corpus.

2 The power of images

Compared to text and voice, visuals are a superior form of communication (Nelson et al., 1976). Images are more likely to capture people’s attention. Eye-tracking studies find that people pay particular attention to visuals when reading the news (Dahmen, 2012). People are also more likely to recall visual information. Paivio et al. (1968) showed that subjects were more likely to remember the ordering of a list of objects from pictures than from names. Visuals also generate stronger emotional reactions than text (Grabe and Bucy, 2009).

Studies suggest that images are also important for understanding political attitudes and behavior. Numerous articles find that voters judge candidates differently based on their looks, that respondents who watch more television are more likely to use looks as a voting cue, and that conservative parties in Europe benefit from the fact that their candidates are generally judged to be better looking (Todorov et al., 2005; Poutvaara, 2017). In a widely covered *Science* article, Todorov et al. (2005) also demonstrated that looks impact electoral success. They asked subjects to choose the most competent U.S. congressional candidate after looking at their pictures for just one second. They report that respondents’ choices correctly predicted 70% of the race outcomes (in which they knew nothing else about

the candidates), and correlated with margins of victory. Beyond candidate looks, however, research on the political power of images is limited. In a recent study of a Black Lives Matter protest event, Casas and Webb Williams (2018) found that images evoking enthusiasm or anxiety were positively related to greater attention to and sharing of Twitter messages. There is believe that this sharing had tangible effects because other studies of the same movement found that social media penetration was predictive of on-street support and mobilization (De Choudhury et al., 2016; Freelon et al., 2016).

Many of the studies cited above involved experiments with a small number of images. The latter Black Lives Matter study was based on a relatively modest sample of few thousand social media images. The increasing importance of social media as an information source and means of communication makes it increasingly important to be able to systematically study much larger corpora. The recent availability of computer vision methods have opened exciting opportunities. Any researcher can now use existing off the shelf tools to label images for a large number of pre-existing content categories. Open source tools can also be fine tuned to assign new labels. Won et al. (2017), for example, uses a fine-tuned Convolutional Neural Net to detect the presence of violence in protest-related images with the goal of studying its impact on future protest participation. These opportunities also extend beyond simply tagging images for content. In a study of voting fraud in Mexico, Cantú (2019) train a CNN to detect alterations of vote tallies.

However, trained CNNs may be of limited value when the pre-defined classes do not map with the objectives of a specific project. For example, symbols of collective social identity (Tajfel, 1981) can activate feelings of group belongingness that increase voter turnout (Gerber et al., 2008), influence attitudes on particular issues (Hassin et al., 2007), and inspire protest participation (Kharroub and Bas, 2015; Casas and Webb Williams, 2018). Some symbols of collective identity (such as a flag) can be detected using existing CNNs. But relying on existing CNNs also limits the potential scope of such a study in important ways. Symbols

of collective identity are often intentionally obscure and ephemeral. Existing CNNs will not detect Pepe the Frog or gestures such as the 'OK' hand sign, two important symbols of far right groups. It is also well known that censorship forces Chinese dissidents to use pseudonyms and, one suspects, visual pseudonyms.

Transfer learning, where a pre-existing CNN can be fine tuned using a relatively small sample of images of a new class, can address the challenge of labeling images for newly discovered symbols of collective identity. But how does a researcher discover them in the first place in a corpus of thousands or millions of images?

Commercial and open source CNNs may also not predict theoretical mechanisms of interest to social scientists. They are amazingly good at object detection. They can identify thousands of common objects of interest and, through transfer learning, can be adapted to new objects of interest (Webb Williams et al., 2019). Much less work has focused on subjective image labeling. Yet, where symbols of collective identity are concerned, the object is less important than the emotion it evokes. Moreover, the same object may evoke different emotional responses in different viewers.

Emotions are an important phenomenon in politics. Research finds that emotions influence attitudes on issues and behavior in a range of participatory settings (such as elections and social movements) (Valentino et al., 2011; Marcus et al., 2000; Clifford and Piston, 2016). Existing affective models point to a range of emotions as strong political motivators, including enthusiasm, anxiety, and anger (Marcus et al., 2000, 2017).

Some computer vision models predict whether people in images *look* angry, sad, or happy (Busso et al., 2004), while other research predicts whether an image evokes generally positive or negative feelings in the viewer (Xu et al., 2014). This work is importantly incomplete in that social science research finds that some negative emotions (e.g. anxiety) encourage mobilization and support, whereas other negative emotions (e.g. sadness) have the opposite effect (Marcus et al., 2000). In addition, whether I feel sad after observing a sad face (e.g.)

probably depends on the context. For example, I may feel very differently if the person in the image is a member of a political outgroup.

The limited computer vision research that does attempt to predict a broader range of evoked emotions has not produced accurate results. Peng et al. (2015) trained a CNN to predict images evoking anger, fear, disgust, and sadness, achieving average accuracy of just 65% (based on a balanced dataset). Webb Williams et al. (2019) also experienced limited success when using transfer learning to predict evoked emotions.

The poor performance of automatic classifiers designed to predict evoked emotions (and other subjective reactions to images) means that researchers interested in studying the phenomenon must still rely on manual image annotation. For larger projects (such as those examining social media) researchers need to either sample from their image cropolis, and/or find ways to reduce the costs of annotating as many visuals in their datasets as possible. Image corpus of interest to researchers are often highly unbalanced (specially for those using social media data) with some particular types of images accounting for a large proportion of the dataset. By choosing a random sample of images to be manually labeled, researchers may be “over-annotating” a particular type of image and missing on opportunities to learn from additional (although less frequent) image types. For example, for those studying how emotional reactions to images impact the diffusion of a protest movement on social media (Casas and Webb Williams, 2018), or the ability of people to distinguish between real or fake social media visuals (Paris and Donovan, 2019), leveraging information from as many types of visuals as possible would strengthen their inferences.

Moreover, to increase the amount of images from large datasets that one can study when analyzing subjective reactions to visuals, researchers would also benefit from having an automatic method that can help them propagate the annotations given to a small sample of the images to as many images as possible in their dataset; so they can use a larger sample to test their hypotheses without the need of spending resources on manually annotating more

images.

Unsupervised visual clustering can help to address these three issues. It can facilitate the exploration of large image corpora with the goal of identifying image classes relevant to a project’s goals. It can also help draw purposive samples and allow researchers to make inferences from as many types of images of interest as possible; and it can help automatically annotate images for subjective reactions by propagating the human labels given to a small set of images in a dataset to those additional images that look very similar to the annotated ones.

3 A New Automated Visual Clustering Method

We propose a visual clustering method that can be used to discover new meaningful patterns within large volumes of images, to draw purposive samples, and to reduce annotation costs by applying the same labels to clusters of similar images. In contrast to prior research, this method does not presume a preexisting set of possible cluster categories. There are three main challenges to building such a method. First, we need to represent images numerically in order to then be able to apply a clustering algorithm to the image corpus. Second, as with all clustering methods, the method requires that we designate a criterion for selecting the appropriate number of image clusters. Finally, we need to a method for evaluating and selecting the best performing algorithm. In this section we discuss how we address each of these challenges.

3.1 Example Image Dataset

Throughout the paper we will use the same example image dataset of our own to clearly illustrate how the proposed method works. In the autumn of 2017, we began a substantial Twitter collection effort for studying social movement support. We simultaneously gath-

ered tweets by entities who are often the originators of social movements, including a large number of USA-based public affairs organizations and politicians, as well as all tweets (by anyone) that included hashtags used by these entities. The hope was that we would capture social movement successes and failures rather than just the successes. From this universe of hashtags (and tens of millions of tweets) we selected the #FamiliesBelongTogether for the current paper. This corpus contains 174,172 tweets (and 88,075 images; 18,096 unique) collected between May 30th and October 27th, 2018. Appendix A provides further details about the data collection process.

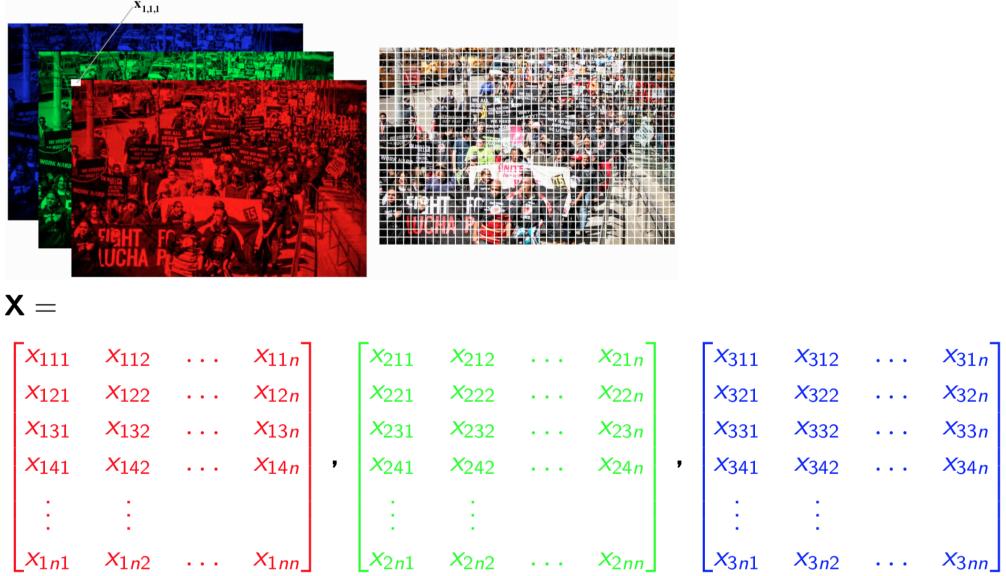
3.2 From Images to Data

Images are typically represented as three-dimensional matrices, where each matrix represents the intensity of red, green, and blue in a particular pixel of the image (standardized values ranging from 0 to 255, see Figure 1). Though numeric, these raw representations are not very useful for clustering images because they are very large matrices that make comparisons resource intensive and because they are not very discriminating from a thematic perspective.

CNNs start with these three dimensional matrices as inputs and ultimately transform them 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). The final step of a supervised method is to compute the probability that the image is in fact a chair (or a horse, etc.) based upon the similarity of this embedding to the embeddings of other labeled examples (see Figure 2 for an example of a CNN architecture).

For an unsupervised approach, embeddings are much easier to work with than the original three dimensional representations. They are also an effective way to capture not only the visual but also the thematic similarity of images before forcing them into specific classes. A model trained to predict the 1,000 ImageNet classes for example has already learned that two chairs that look very different are indeed two instances of the same class, which means that

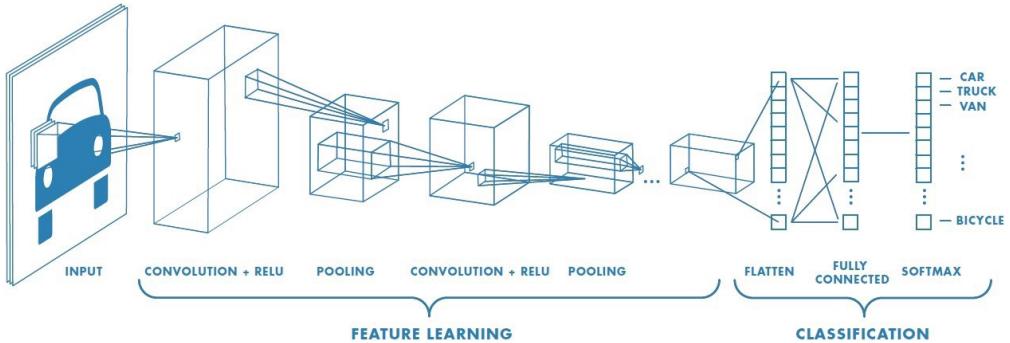
Figure 1: An image represented as a 3-dimension input. Each $X_{i,j,z}$ unit contains information about the pixel-level intensity of red, green, and blue in the image.



the image embeddings generated by the model already carry some sort of human “meaning”.

An approach to generating image embeddings for a corpus of interest would be to pass all images through a model pre-trained to predict these ImageNet categories. However, the “meaning” carried in these embeddings may fail to capture additional meaning of interest to the researcher. It could therefore be that, for example, images evoking similar emotions are not close to one another in this embedding space.

Figure 2: A Convolutional Neural Network (CNN) for image classification



Note: Credit: Sumit Saha at <https://towardsdatascience.com>

A better approach then may be to fine tune an open-source pre-trained model before using it to generate image embeddings, this way one takes advantage of a model that has already learned from hundreds of thousands of examples while also making sure that the model captures (at least partially) the additional ‘meaning’ of interest to the researcher.

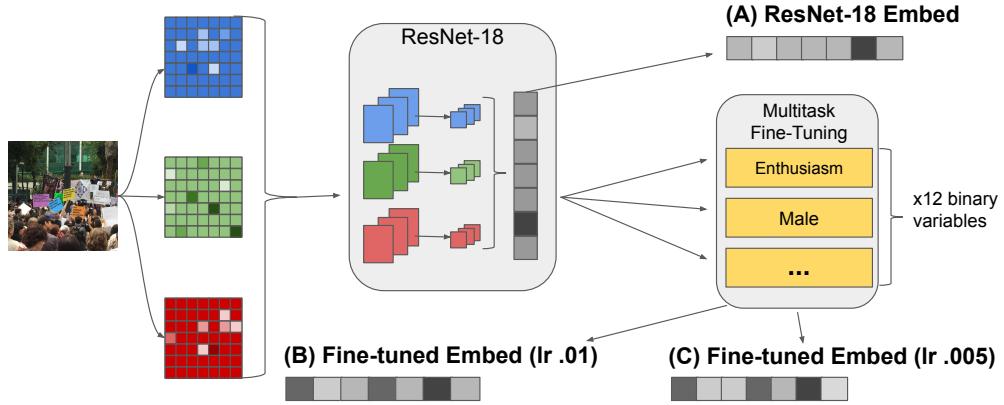
In order to be able to evaluate both approaches, we labeled a small set of images drawn from tweets using the #FamiliesBelongTogether hastag ($N = 609$). Six annotators annotated (using binary judgements) each of the 609 images for elements that are theoretically relevant to us (the evoked emotions we focus on this paper plus some additional features in which we are interested in a related project): a) the emotions the image evoked (enthusiasm, anger, and anxiety), b) whether people of different ethnicities were present in the image (white, black, Asian, and Latino), c) the gender of those people (male and female), d) whether a child was present, e) whether a symbol of collective identity was present, and f) whether annotators thought that the image communicated that the movement could be successful at accomplishing their goal.

We then fine-tuned a pretrained (ResNet-18) computer vision model (He et al., 2015) with this information by replacing the last fully connected layer with a new one predicting 12 rather than 1,000 classes, and then retraining the model weights for 50 additional iterations using our annotated dataset. Instead of using a softmax function predicting the probability of each image belonging to one of a set of mutually exclusive classes (probabilities adding up to 1), we used 12 sigmoid functions independently to predict the probability of each of the 12 labeled non-mutually exclusive items to be in the image (a process known as multi-task learning, (Caruana, 1997)). We repeated the process twice, the first time using a learning rate of .005 and the second time using a learning rate of .01.

Once these two new ResNet-18 models were fine-tuned, we passed the rest of the images in our dataset through them and pulled the output of the second-to-last fully connected layer. As shown in Figure 3, in the end, we ended up with three types of image embeddings

representing the images in our dataset in a denser and lower dimensional space: A) the embeddings directly generated with the pretrained ResNet-18 model (*ResNet-18 Embed*), B) the ones generated after the multitask fine-tuning of ResNet-18 using a .01 learning rate (*Fine-tuned Embed (lr .01)*), and C) the ones generated after fine-tuning ResNet-18 using a .005 learning rate (*Fine-tuned Embed (lr .005)*).

Figure 3: Three types of embeddings used for image clustering.



3.3 An Iterative Clustering Method

After transforming the images to lower dimensional embeddings representing both the look and meaning of the pictures, we proceeded to grouping similar images in an unsupervised fashion using a k-means algorithm, a common clustering method in machine learning research (Jain, 2010). One option was to fit a single k-means model to the matrix of image embeddings. However, as we will show in the next subsection, this approach did not produce very satisfactory results. Instead, we decided to employ the iterative approach sketched in Figure 4 and that we describe below.

We run the following 5 steps for several iterations. (1) First we decide the number of clusters to be “discovered” by the k-means algorithm in that iteration. In order to do so, we fit several k-means models, increasing the number of clusters by N each time (*Step Size*,

e.g. +5 clusters), and measuring the model fit. To speed up the process, we use a random subsample of the images to perform this task (*Sample Size*, e.g. 1,000 images). We consider that we have found the number of clusters that best describes our data once the average of the goodness of fit for the last I runs does not improve (*Converge Window*, e.g. 3 iterations).¹

Then, the second step (2) in the process is to fit a k-means algorithm with that particular number of clusters to the whole image dataset. (3) Next, we evaluate the cohesiveness of the clusters using the average of the silhouette score for the images in each cluster.² (4) Clusters of images that have a silhouette score of over a threshold (*Similarity Threshold*, e.g. 0.4)³ are identified. (5) All the images within clusters that receive a silhouette score above the *Similarity Threshold* are pulled out of the corpus of images. After these images are pulled out and these clusters of images are saved, we run all five steps again on the remaining images, and we keep repeating the process until we run out of images to cluster or the number of remaining images is too low (e.g. 20 images left, *Stop Size*). In the next section we demonstrate that this iterative approach performs better than a single-shot k-mean clustering. Moreover, notice that in the description of the method we have mentioned five hyperparameters for which the researcher needs to choose values (sample size, step size, convergence window, similarity threshold, and stop size – these (and sample values for them) are also underlined in Figure 4). In the next section we also explain how to judge which hyperparameter configuration works best.

¹We average the goodness of fit measures across the last I runs to make sure that the decision is not based on an isolated run of the model and so to make sure that the fit of the model does not improve as the number of clusters increase.

²A silhouette score measures both how similar images classified in the same cluster are to each other as well as how far apart images classified in different clusters are to each other. It measures both the cohesiveness of a cluster and its distance from other clusters.

³Silhouette scores range from -1 and 1, 1 being the highest level of cohesiveness and uniqueness, and -1 being the lowest level of cohesiveness and uniqueness.

Figure 4: Pseudo Code of the Iterative Clustering Method

X = input image matrix (e.g. 20,000 images \times 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.

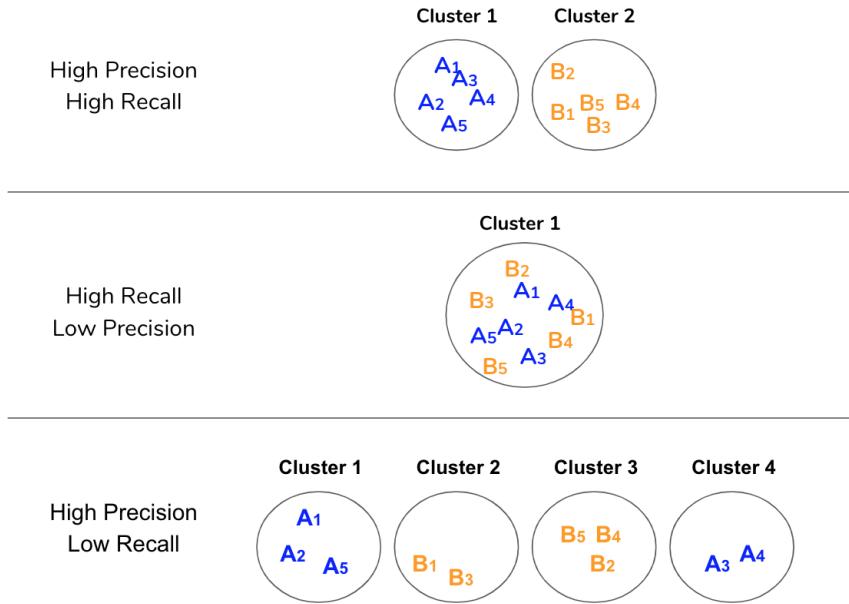
3.4 Validation

We created a gold standard dataset in order to know which lower dimensional image representation, and hyperparameter configuration, do a better job at clustering together images that are similar on the dimensions that we care (e.g. evoked emotions). We manually annotated a set of image pairs from our #FamiliesBelongTogether for whether we believe they should be classified into the same cluster. We used two main criteria when making this call: a) the images in the pair had to be instances of the same element (e.g. a protest, a sign, the same person, etc.), and b) the images had to be likely to evoke the same or very similar emotional reactions if coded by the same person. The former is crucial to make sure that the resulting clusters do a good job at representing the different types of images in the corpus, and so to ensure that the clustering technique serves the “corpus exploration” and “purposive sample” goals well. The latter is key to make sure that the resulting clusters are as cohesive as possible on the theoretical mechanisms of interest (evoked emotions), and so that we can do the best job possible at propagating labels within the clusters and use them to “reduce image annotation costs.”

A simple exploration of the unique images in the #FamiliesBelongTogether corpus re-

vealed a very unbalanced dataset, with a large number of images clearly being instances of the same elements (e.g. images of letters directed to public representatives, instances of street protests, etc.). A random sample would yield very similar image pairs, missing example pairs from numerous potential future clusters. To account for this lack of balance in our validation set, we first fit a 80-cluster k-means algorithm on the #FamiliesBelongTogether images.⁴ and from each resulting cluster we randomly sampled pairs of images (554 pairs in total). The annotators judged cohesiveness among these image pairs very similarly.⁵

Figure 5: 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.



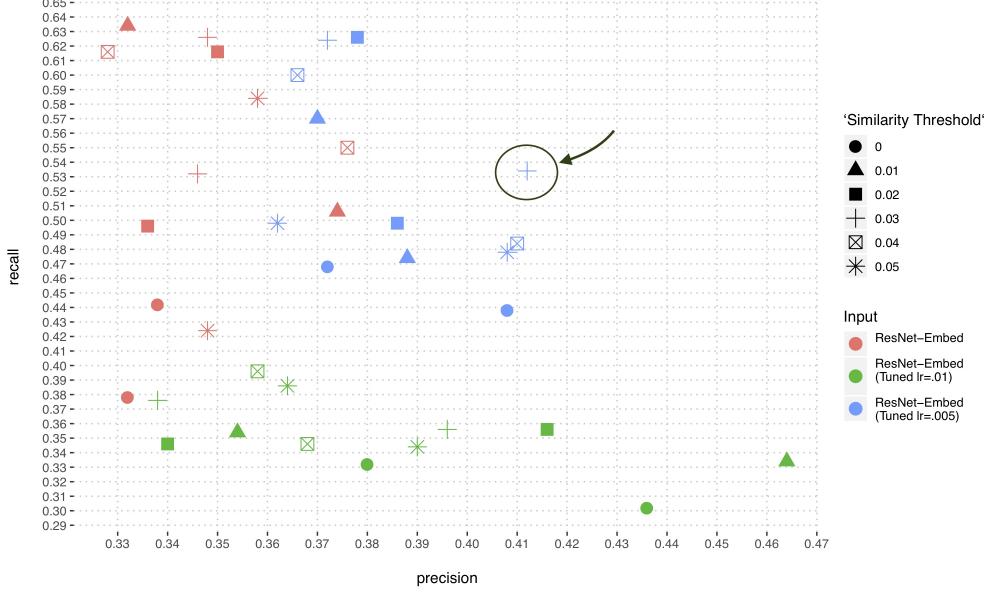
We then use the gold standard dataset to perform a grid search and assess the precision (proportion of pairs correctly predicted to belong to the same cluster) and recall (proportion of pairs labeled as belonging to the same cluster that have been classified as such) of the three types of image embeddings and of different values for the five hyperparameters involved in the iterative clustering process. In particular, we tried all possible combinations of the

⁴We decided this number of clusters after a goodness of fit analysis in which we evaluated fitting a varied number of clusters.

⁵87% agreement between two annotators and a Cohen's kappa value of 0.64

following hyperparameter values, $\{3, 5\}$ for step size, $\{0.0, 0.01, 0.02, 0.03, 0.04, 0.05\}$ for the similarity threshold, and we kept constant a sample size of $\{1,000\}$, a stop size of $\{20\}$, and a convergence window of $\{3\}$.

Figure 6: Performance of Different Hyperparameter Configurations and Types of Image Embeddings



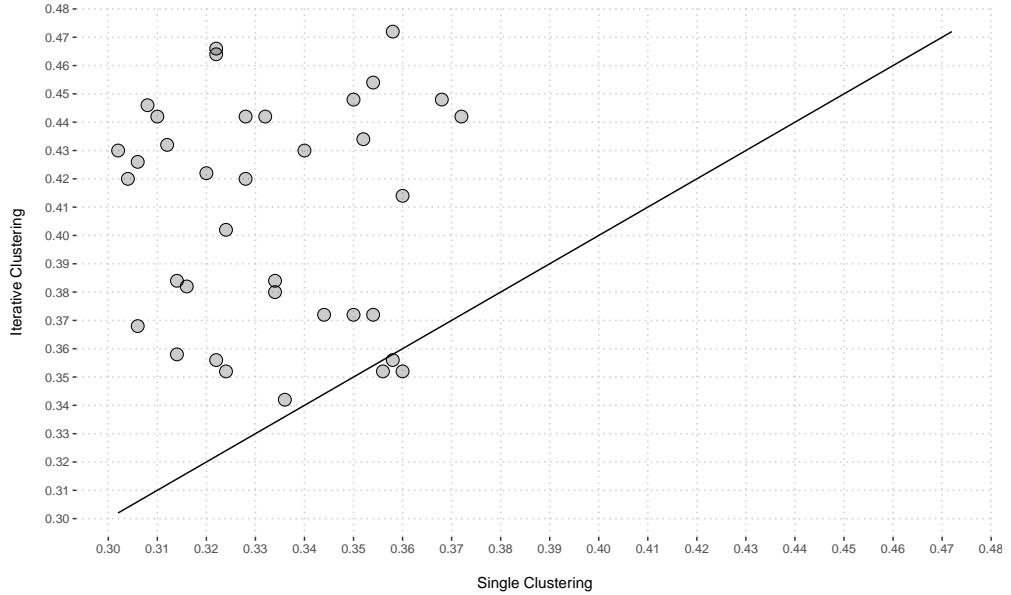
We are interested in finding the type of input and hyperparameter configuration with the highest precision but also the highest recall. In Figure 5 we use a toy example to clearly illustrate why. High recall but low precision (second scenario from the top) means that we are correctly clustering images that should go together, but also images that should not. This is usually a sign of having fewer clusters than we should. High precision but low recall (bottom scenario) means that we are correctly not clustering images that should not go together, but it also means that we are splitting into different clusters images that should be grouped in the same cluster. Hence, when performing the grid search, we are interested in the performance that get us closer to the high precision and high recall scenario at the top of Figure 5.

Figure 6 shows the results of the grid search, based on a 5-fold cross validation. We

observe that the best performing clustering model (circled blue cross) uses the embeddings generated with ResNet-18 fine-tuned with a learning rate of 0.005, and a similarity threshold of 0.03 when discriminating non-cohesive groups of images during the iterative clustering process. In the subsequent sections of the paper we will use this model to illustrate how the resulting clusters can be used for corpus exploration, purposive sampling, and for reducing image annotation costs.

Finally, in Figure 7 we take advantage of the creation of this gold standard dataset to also assess whether the iterative clustering option outperforms the one-shot k-means clusterings for all the clustering options we tried during the grid search. The F-scores (average precision-recall) is practically always higher for the iterative option.

Figure 7: Performance of Single *versus* Iterative K-Means Image Clustering

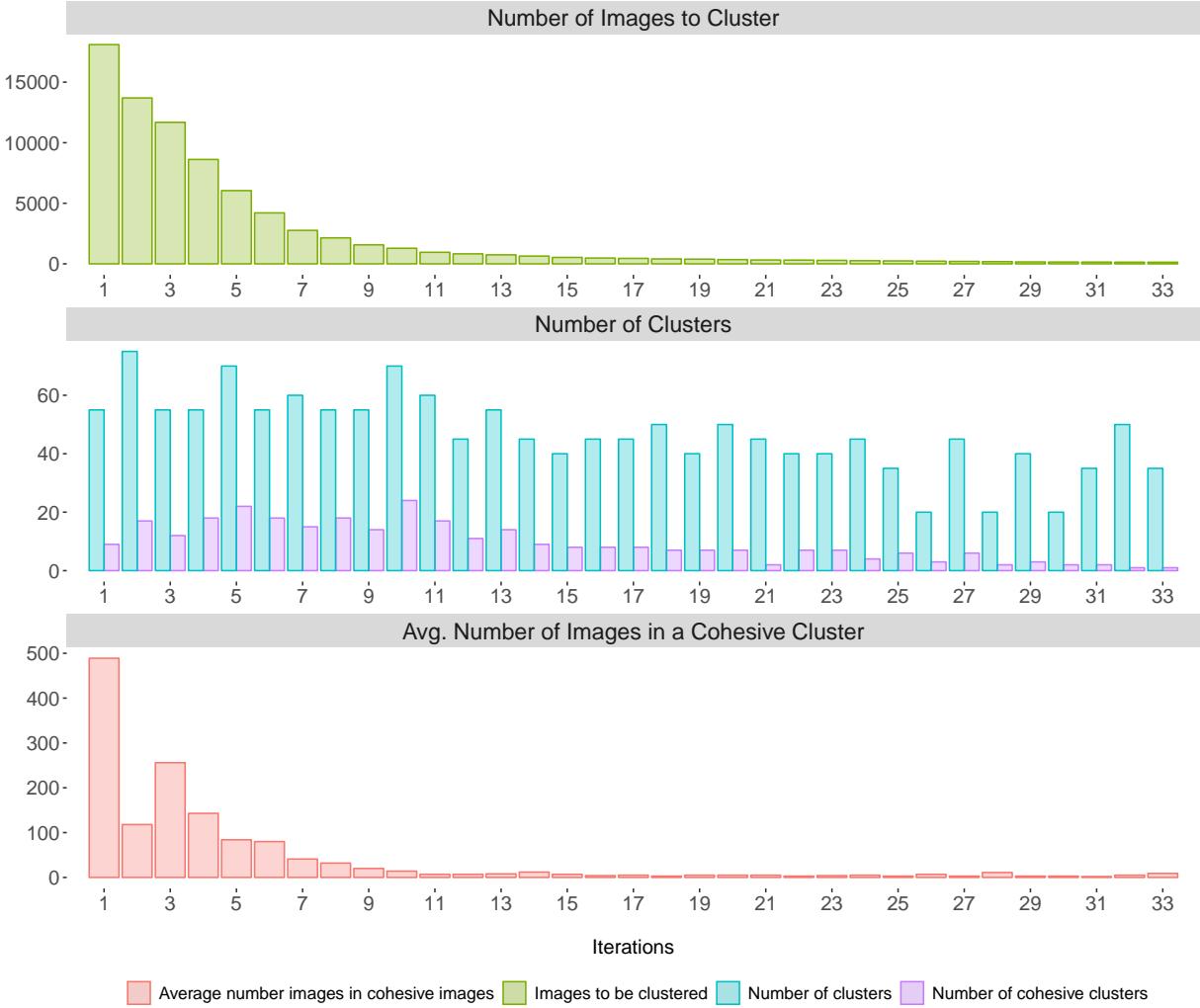


3.5 Best performing model

In Figure 8 we provide further information about the best performing cluster. The algorithm run for 33 iterations and it generated 309 image clusters. In the top panel, showing the

number of images still to be classified into cohesive clusters at each iteration, we observe that most images were classified into cohesive clusters in the first 10 runs. In the middle panel we see that the number of clusters to be fit at each iteration based on the specified goodness of fit test remained quite constant, around 55 (light blue bars). The purple bars indicate how many of these clusters can be considered cohesive after discriminating based on the silhouette score and the similarity threshold. We observe fewer cohesive clusters to be found 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 big groups of images: the cohesive clusters found in the first iteration had an average of about 500 images, approximately. Then, we still see the algorithm to find relatively large clusters until the sixth iteration. The cluster found in later iterations are much smaller, which as we see in the top panel, is also a function of not that many images being left to be clustered.

Figure 8: Information about the iterative progress of the best performing model

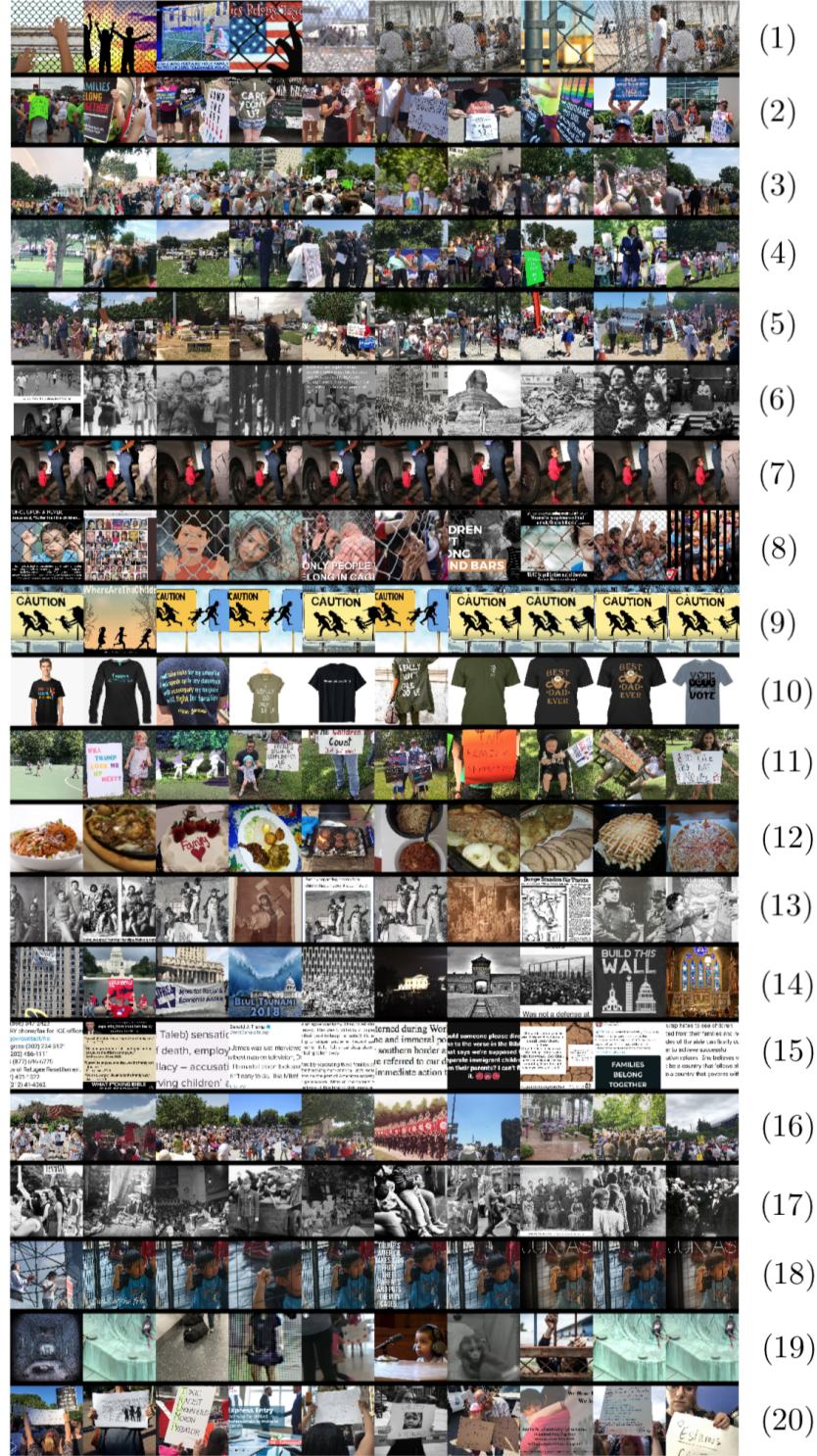


4 Corpus exploration

Once images have been clustered, researchers can randomly sample some images from each group in order to explore the data at hand. In Figure 9 we show ten sample images from twenty of the 309 clusters generated by the best performing model. Both the clusters and the images have been randomly selected.

First, the figure illustrates how the method can help researchers get a sense of the type of data they are working with. For example, several clusters are clearly related to social

Figure 9: Sample images from 20 random clusters



and protest movement: rows such 2, 3, 4, and 5 show clusters of different types of images of street protests. However, there are also some types of images that a researcher studying the diffusion of this social movement on social media may have not expected, such as images with different kinds of food (row 12). The method helps the researcher to better identify, analyze, and understand particular part of the data as well as the dataset as a whole.

This exploration technique also allows the researcher to get a better sense of the different ways in which theoretical mechanisms of interests can be represented in the data. For example, for those interested in images as framing mechanism ([Torres, 2019](#)), the method allows them to see the range of ways in which people framed the movement. There are several mobilization clusters, with people protesting on the street (such as rows 2, 3, 4, and 5), but also several clusters of images with children behind fences (row 18) or in cages (row 8), children in general (row 11), and there are some clusters of images portraying immigration detention centers as nazi concentration camps or slave plantations (row 13). The initial exploration can help those studying framing to elaborate a more detailed codebook that they can later use to conduct a more systematic analysis of the frames used (or not) when discussing the movement.

Finally, this clustering technique can also help explore the corpus of data by running some preliminary descriptive analysis in the same way people often use the output of unsupervised topic models (such as LDA) for text. Continuing with the framing example, a researcher interested in whether i.e. conservative people are more likely to engage with a particular type of frame than liberals, can code the 309 clusters for whether they have that particular frame, and in conjunction with [Barbera \(2015\)](#)'s method for estimating the ideology of Twitter users, analyze the proportion of conservative *v.* liberal users who engaged with (i.e. shared) images that contained that particular frame.

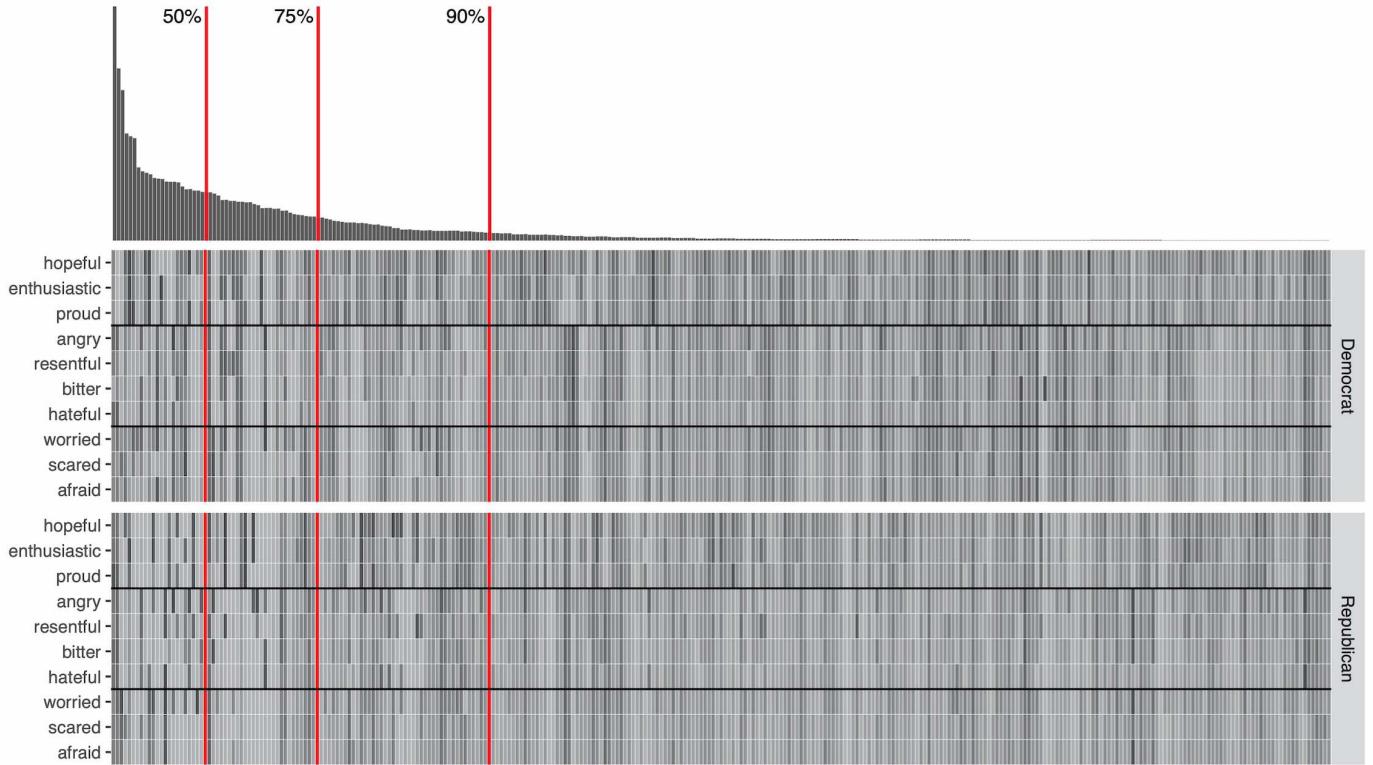
5 Purposive sampling

Leveraging the unsupervised clustering method, our next technique serves to draw a purposive sample from our image corpus. As shown in Figure 8, particular types of images account for a substantive proportion of our data. For example, a single cluster (out of the 309) of very similar images of street protests account for 5% of the corpus, instead of the 0.003% that would correspond to a single cluster in a perfectly balanced dataset. By choosing images at random to be manually annotated for a subjective reaction of interest, such as evoked emotions, one would be annotating many instances of this same type of image, in detriment of getting and using information from many other types of images with a smaller presence in the corpus. If the goal is for example to understand how the emotions that different types of images evoke can be predictive of the diffusion and success of a movement on social media, the latter will help the researcher make inferences from a richer dataset.

To illustrate this point, we first created a purposive sample by randomly selecting 5 instances from all clusters with more than 6 images, and $N - 1$ instances from clusters with 2 to 5 images (1,087 images in total). Then we recruited a set of human annotators ($N = 720$) via the polling company Qualtrics that self-reported to be Democrats ($N = 360$) and Republicans ($N = 360$), and we asked them to rate each image for the emotions they evoked to them. Images were assigned to coders at random, each coder annotated a maximum of 8 images, and the same coder did not label images from the same cluster. Building on a widely used model in political psychology, the *Affective Intelligence* model (Marcus et al., 2000), we asked them to rate their emotional reactions (on a 0-10 scale) on 10 emotions known to load onto three distinguishable emotional dimensions (Marcus et al., 2017): hopeful, enthusiastic, proud (*Enthusiasm* dimension); angry, resentful, bitter, hateful (*Aversion* dimension); worried, scared, and afraid (*Anxiety* dimension). Each image in this purposive sample was labeled by one Democrat and one Republican, allowing to assess the

extent to which supporters of the two parties reacted differently to images related to the #FamiliesBelongTogether movement.

Figure 10: Reported emotional reactions to images from different clusters by Democrats and Republicans



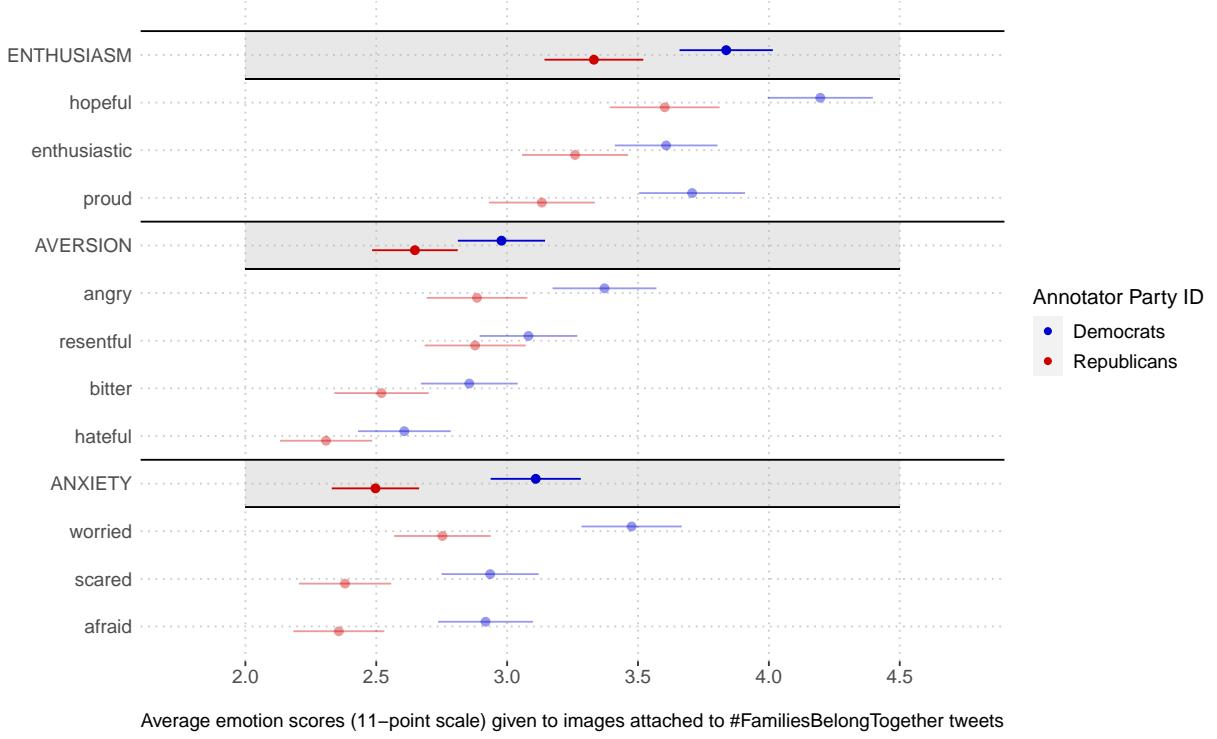
In Figure 10 we average the emotion scores given to images belonging to the same clusters to show how Democrats (top panel) and Republicans (bottom panel) reacted to these different types of images (darker rectangles indicate a stronger emotional reaction). In each panel, the first three emotions correspond to the *enthusiasm* dimension, the following four to the *aversion* one, and the bottom three to the *anxiety* dimension. Each column of rectangles in both panels indicates the emotional reactions to a single cluster. The clusters are sorted in descending order left to right based on their size (number of images in each cluster). At the top of the two panels we show a distribution of the number of images in each cluster, and the red vertical lines indicate at which point the images in those clusters account for

50%, 75% and 90% of the whole corpus.

There are several takeaways to emphasize from the figure. The first one is that Democrats and Republicans reacted differently to the same images, in the first set of clusters on the left we observe several images that evoked *enthusiasm* to Democrats but not to Republicans, probably due to Democrats being more sympathetic with the movement. This clearly illustrates that when studying subjective reactions to visuals, it is extremely complicated to train a straight forward supervised model predicting people's response. In this example we only checked whether partisanship influences how people will feel when seeing these images, but other individual characteristics can be also predictive of the same outcome (gender, ethnicity, etc.), calling for using manual labeling rather training supervised algorithms for this types of subjective image annotation. Second, we observe clusters (with a few images) that would had been rarely selected in a random (rather than purposive) sample, to clearly evoke some of the emotions of interest and to uncover differences between Democrats and Republicans that could be crucial to understand how partisanship influences how people react to visuals related to the movements (and the diffusion of the movement overall). For example, between the 75% and 90% mark we observe a set of clusters that evoked enthusiasm to Republicans but not to Democrats. If using a random sample, we would most likely missed this variation in our analysis.

In Figure 11 we use the emotion scores reported in Figure 10 to report how much Democrats and Republicans felt on average when exposed to the images in this purposive sample. In the gray areas we average the scores for emotions that are part of the same dimension, and in the non-gray areas we disaggregate the results for each individual emotion. An interesting finding is that Democrats felt more enthusiastic and more anxious when exposed to the images, but they did not clearly felt more averse. Images of street protest are among the ones that generated more enthusiasm (specially among Democrats) and images of children in cages and detention centers are among the ones that evoked higher anxiety

Figure 11: Average emotion reactions, by Democrats and Republicans, to a purposive sample of 1,087 images shared on Twitter with the #FamiliesBelongTogether hashtag



(specially among Democrats as well).

6 Reducing Annotation Costs

Finally, we leverage the unsupervised clustering method to reduce image annotation costs by propagating the labels of manually annotated images in a cluster to the rest of the instances in that same clusters. Two key conditions need to be met by a given cluster for this propagation process to produce additional accurate and useful automated labels. The images in a given cluster need to be similar enough visually to be considered instances of the same type of image. Moreover, different manual annotators need to have given roughly similar labels to images sampled from the same cluster. Judging whether these two conditions are met will vary depending on the research project at hand. For example, a project about visual framing

may be simply interested in whether people use a mobilization frame when messaging about a social movement, and so any kind of street protest image can be considered instances of the same type of image. However, different types of street protest images may evoke different emotions, and so a project interested in the impact of emotional reactions on mobilization may be more strict about whether images in a given cluster are indeed instances of the same type of images.

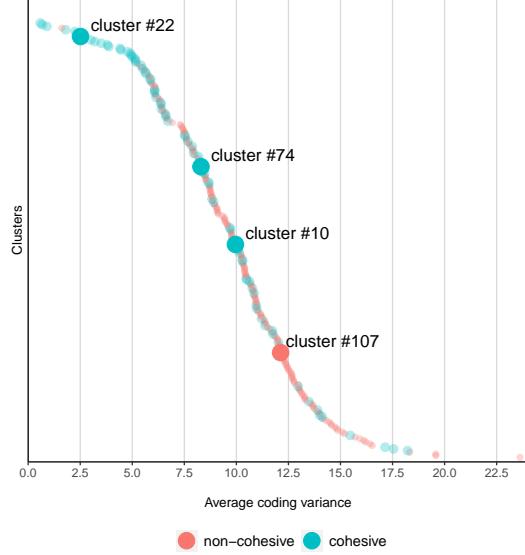
During the iterative unsupervised clustering step we made sure that our clusters were cohesive from a statistical perspective, based on the embedding representation of the images. However, as it often happens with unsupervised clustering methods such as LDA for text analysis, one often ends up with melting-pot clusters that are non-cohesive from a human perspective (Chang et al., 2009). Two research assistants evaluated the cohesiveness of our 309 clusters, finding 97 to be cohesive from a human perspective and 212 to be non-cohesive.⁶ Then, for the 1,087 images in the manually annotated purposive sample, we also evaluated the extent to which the Qualtrics coders gave similar labels to images from the same cluster, by calculating for each cluster i and evoked emotion j the variance σ^2 of the score given by coders from party z to each image i selected from that cluster. Then we calculate the average variance across emotions and parties: $\sum_1^z (\sum_1^j (\sigma_{ijz}^2))$.

In Figure 12.a we show how much coding variance we found in clusters the research assistants labeled as being cohesive *versus* non-cohesive. We observe a strong correlation between the two, with most of the clusters coded as cohesive having a low coding variance. In Figure 12.b we provide sample images for four clusters with very low, low, medium, and high coding variance highlighted in Figure 12.a. We observe cluster number 22 and 74 to be fairly cohesive from a human perspective: images of street protest and children sleeping on the floor in detention centers. Cluster number 10 is also fairly cohesive with the only

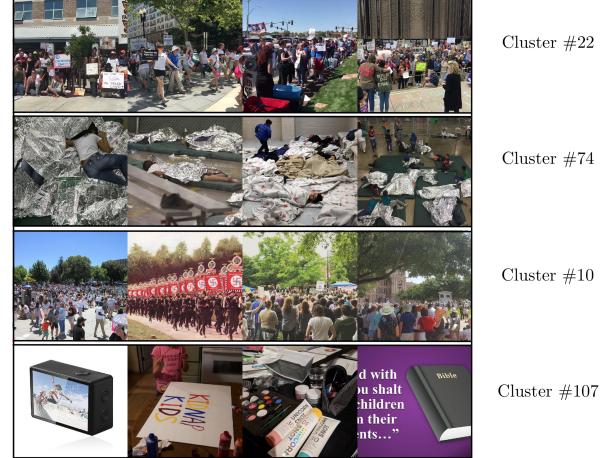
⁶Coders judged in a binary scale whether the images in the same cluster looked visually similar “and” whether they thought they would code them similarly for evoked emotions. Cohen’s kappa of 0.72 for a common random set of clusters labeled by both coders.

Figure 12: Cluster-level coding variance

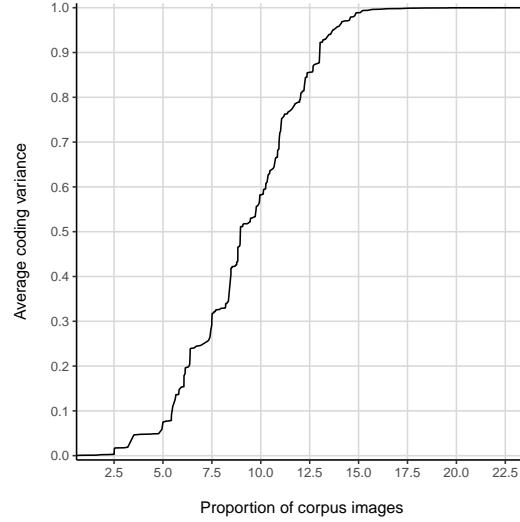
(a) Distribution of the cluster-level coding variance.



(b) Examples of clusters with very low, low, high, and very high clustering variance



(c) Proportion of images in our corpus that would be automatically labeled based on a specific coding variance cut-off point.



exception of one instance, an image of a nazi military parade, making this cluster not the perfect candidate for label propagation. As final example, cluster 107 is clearly a melting-pot cluster where instances of different types of images ended up grouped together. Overall,

the data shows that clusters with an average coding variance of up to 8, approximately, are fairly cohesive from a human perspective and so by propagating the labels of the annotated images to the rest of the images in those clusters we would automatically get useful labels for additional images.

Finally, in Figure 12 we show what proportion of the images in our corpus would be automatically annotated via label propagation, depending on the specific coding variance we decide as cut-off point for allowing (or not) a cluster to take part of the propagation. We observe for example that we would automatically label between 30-40% of the images in our corpus if we applied the label propagation to the clusters with a coding variance below 8, which we found to be a very reasonable cut-off point in Figure 12.a/.b. Given that we labeled only about 5% of the images in our dataset, this would mean an automatic 6-to-8 times increase in the number of annotated images. Again, the exact cut-off point will depend on the research goal and a case-specific assessment similar to the one we did on Figure 12.a is needed. However, given the high level of subjectivity of the research goal of this application example (evoked emotions), we believe that the automated 6-to-8 time increase in the number of annotated images is a lower bound on the amount of additional data researcher can get annotated using this unsupervised clustering and propagation technique.

7 Discussion

Visual communication is in the rise. Finding ways of incorporating visual effects in communication and political science research is a pressing topic. However, automated methods for simply exploring large image corpora, drawing purposive samples, and reducing the costs of annotating large datasets for the presence of complex and subjective theoretical mechanisms does not exist.

In this paper we put forward a method that aims at accomplishing these three tasks. First,

we used a deep convolutional neural net to represent images as vectors. Then we designed an iterative k-means algorithm to cluster images into groups of very similar pictures. Next we put forward a validation procedure to find the algorithm specification that yielded the best performance. After clustering the images, we showed how one can use these clusters to better understand particular parts of the data and the dataset as a whole, as well as to run some preliminary analysis. We also showed how the method can be used for drawing and working with purposive samples of images; and finally, we annotated a small set of the images in each cluster and assessed the extent to which the human labels could be automatically propagated to the remaining images in the corpus, with very promising results.

We believe this unsupervised method technique and applications are a great contribution to multiple social science disciplines. We illustrated how the method can help study the emotions protest-related visuals evoke to people and the extent to which such emotional reactions are related to political mobilization. But the method can be useful to any researcher interested in using images as data. For example, to those interested in detecting fake images that circulate in the internet and the extent to which people are able to recognize them as being fake, or those interested on how people imply leadership traits from images of politicians.

References

- Barbera, Pablo. Birds of the Same Feather Tweet Together. Bayesian Ideal Point Estimation Using Twitter Data. *Political Analysis*, pages 1–16, 2015.
- 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.
- 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*, 0(0):1–18, 2018.
- Chang, Jonathan, Sean Gerrish, Wang Chong, Jordan L. Boyd-Graber, and David M. Blei. Reading tea leaves: How humans interpret topic models. In Bengio, Y., D. Schuurmans, J. D. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 288–296. Curran Associates, Inc., 2009.
- Clifford, Scott and Spencer Piston. Explaining Public Ambivalence about Homelessness Policy: The Role of Disgust. *Political Behavior*, 2016.
- Dahmen, Nicole S. Photographic Framing in the Stem Cell Debate. *American Behavioral Scientist*, 56(2):189–203, 2012.
- De Choudhury, Munmun, Shagun Jhavery, Benjamin Sugary, and Ingmar Weber. Social Media Participation in an Activist Movement for Racial Equality. In *Proceedings of the 10th International AAAI Conference on Weblogs and Social Media*, Cologne, Germany, may 2016.
- Freelon, Deen, Charlton McIlwain, and Meredith Clark. Quantifying the power and consequences of social media protest. *New Media & Society*, 2016.
- 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, Oxford; New York, 2009.
- Hassin, Ran R., Melissa J. Ferguson, Daniella Shidlovski, and Tamar Gross. Subliminal exposure to national flags affects political thought and behavior. *Proceedings 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.
- Jain, Anil K. Data clustering: 50 years beyond k-means. [Pattern Recognition Letters](#), 31(8):651 – 666, 2010.
- 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.
- Nelson, Douglas L, Valerie S Reed, and John R Walling. Pictorial superiority effect. [Journal of experimental psychology: Human learning and memory](#), 2(5):523–528, 1976.
- Paivio, Allan, T B Rogers, and Padric C Smythe. Why are pictures easier to recall than words? [Psychonomic Science](#), 11(4):137–138, 1968.
- Paris, Britt and Joan Donovan. Deepfakes and cheap fakes. [Data Society](#), 2019.
- Peng, Kuan-Chuan, Tsuhan Chen, Amir Sadovnik, and Andrew Gallagher. A mixed bag of emotions: Model, predict, and transfer emotion distributions. In [2015 IEEE Conference on Computer Vision and Pattern Recognition \(CVPR\)](#), pages 860–868, June 2015.
- Poutvaara, Panu. How do candidates looks affect their election chances? [IZA World of Labor](#), 370:1–11, 2017.
- 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.
- Todorov, Alexander, Anesu N Mandisodza, Amir Goren, and Crystal C Hall. Inferences of Competence from Faces Predict Election Outcomes. [Science](#), 308(5728):1623–1626, 2005.
- Torres, Michelle. Give Me the Full Picture: Using Computer Vision to Understand Visual Frames and Political Communication. 2019.
- Valentino, Nicholas A, Ted Brader, Eric W Groenendyk, Krysha Gregorowicz, and Vincent L Hutchings. Election Night’s Alright for Fighting: The Role of Emotions in Political Participation. [The Journal of Politics](#), 73(01):156–170, jan 2011.

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

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.

Zhang, Han and Jennifer Pan. Casm: A deep-learning approach for identifying collective action events with text and image data from social media. Sociological Methodology, 49 (1):1–57, 2019.

Appendix A Data Collection

To build our data set of mobilization attempts, we collect tweets that contain mobilizing hashtags used by social organizations, news media, and legislators (all based in the USA) from January 1st, 2018 to December 31st, 2018. We collect all tweets from 1,144 American social organizations, 559 American political actors, and 30 American news media outlets. We track Twitter because this social media platform has become an effective mobilizing tool for social organizations and legislators. It is also one of the most open social media platforms in terms of data sharing.

The social organizations we track were identified by 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 and removed 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 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. In total 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). Full information on the Twitter accounts that we track are available in supplementary documents.

Starting on January 1st, 2018, we began collecting all tweets produced by these tracked accounts. Although this data collection process continued until the end of June, 2019, for the sake of this paper we only use tweets up to the end of December, 2018, which gives us a full year of data. From tweets used by these accounts over this time period, we focus on the hashtags within these tweets. Hashtags are a means of organizing on Twitter, so we use this feature to identify mobilization attempts. At the end of each day we pull a list of hashtags that were used more than twice by the same tracked account. From this list, we remove all hashtags that do not have a capitalization in the middle or are shorter than 12 characters. These requirements ensure we are tracking unique hashtags that are being used prominently by at least one of these organizations. Once on our list, we immediately begin collecting any tweet by any Twitter user that uses that hashtag. Each hashtag is tracked for two days and then removed if it is not used more than twice by the same organization over the next two days. Due to the sheer number of posted tweets, we are unable to collect all the tweets using our tracked hashtags, but we are able collect around 1,000,000 tweets per day from an average of 600 hashtags. This process populates a database of tweets that use any of our tracked hashtags on the days we are 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 collect any pictures posted with the tweets. We do not collect videos, but we do collect the image displayed on the tweet representing the video.

From tweets from tracked accounts, we identify which of the hashtags used by these tracked accounts are mobilization attempts (either online and offline). ⁷.

⁷A detailed explanation of this process can be provided upon request, but it is not integral to the method being discussed in the paper