Proceedings of the Eighteenth International AAAI Conference on Web and Social Media (ICWSM 2024)

Examining Similar and Ideologically Correlated Imagery in Online Political Communication

Amogh Joshi[1] **and Cody Buntain**[2]
[1]Princeton University
[2]University of Maryland, College Park
amoghjoshi@princeton.edu, cbuntain@umd.edu

Abstract

This paper investigates visual media shared by US national politicians on Twitter, how a politician's variety of image types shared reflects their political position, and identifies a hazard in using standard methods for image characterization in this context. While past work has yielded valuable results on politicians' use of imagery in social media, that work has focused primarily on photographic media, which may be insufficient given the variety of visual media shared in such spaces (e.g., infographics, illustrations, or memes). Leveraging multiple popular, pre-trained, deep-learning models to characterize politicians' visuals, this work uses clustering to identify eight types of visual media shared on Twitter, several of which are not photographic in nature. Results show individual politicians share a variety of these types, and the distributions of their imagery across these clusters is correlated with their overall ideological position – e.g., liberal politicians appear to share a larger proportion of infographic-style images, and conservative politicians appear to share more patriotic imagery. Manual assessment, however, reveals that these image-characterization models often group visually similar images with different semantic meaning into the same clusters, which has implications for how researchers interpret clusters in this space and cluster-based correlations with political ideology. In particular, collapsing semantic meaning in these pre-trained models may drive null findings on certain clusters of images rather than politicians across the ideological spectrum sharing common types of imagery. We end this paper with a set of researcher recommendations to prevent such issues.

NB Supplemental material, including a crowdsourcing codebook and collection of image embeddings, are available online: https://osf.io/fkcq8/

Introduction

Visual media has long been a key element of political discourse (Seidman 2008; Lilleker, Veneti, and Jackson 2019), and as new online media spaces increasingly focus on imagery (e.g., Instagram, Snapchat, and TikTok), new research agendas in computational social science research have emerged, such as images-as-data for political science outlined in Joo and Steinert-Threlkeld (2018). These new agendas are facilitated by advances in computer vision and the availability of pre-trained deep learning models for image analysis, which has allowed for large-scale, automated characterization of images and their use in online political discourse. Common practice for studying political imagery using these computer vision models is to select a pre-trained model and either use it to generate image "embeddings"–i.e., dense numeric vectors–that characterize images (as in Peng 2021 or Joo and Steinert-Threlkeld 2022) or fine-tune the selected model for a particular task (as in Xi et al. 2020 or Zhang and Pan 2019). Despite the proliferation of readily available pre-trained models from repositories like TensorFlow Hub, relatively little guidance is available for computational social science scholars on how to select or use these models. For example, while Zhang and Peng (2022) provides valuable insight into image clustering, that work relies on a single model architecture and task, and how researchers should select models as new ones rapidly emerge remains an open question. Expanding this guidance is therefore necessary, as insufficient consideration can lead to hazards, such as weakened construct validity, model bias, or partisan asymmetry.

This work provides this needed guidance for computational social scientists by analyzing nearly half a million images shared by 656 Twitter accounts belonging to US members of Congress (MoCs) using five popular pre-trained computer-vision models spanning seven years of development. Extending the literature on political ideology in visual media, we compare the types of political images produced across these computer vision models, examine what these characterizations reveal about ideological lean in visual media, identify potential asymmetries that emerge across the ideological spectrum, and assess construct validity across similar images within image-types.

Results show that, despite varying the underlying pre-trained model, we consistently identify approximately eight classes of image shared by these politicians. In each of these image clusterings, politicians share a variety of image content, with the average MoC sharing at least one image from every image cluster. These image types range from photographs containing groups of constituents, headshots, photo ops, and meetings to more diverse visual elements like screenshots of bills, infographics, and meme-like imagery. Consistent with Xi et al. (2020), we find that a politician's

distribution of image-types shared is indicative of their ideological position (and therefore party affiliation); e.g., conservative politicians appear more likely to share images with patriotic symbols, whereas more liberal politicians appear to share more images of bills, documents, and infographics. We then demonstrate superior performance in predicting MoCs' ideological scores from their average image embeddings compared to prior work.

Our analysis also reveals important considerations, however: First, while the majority of pre-trained models produce relatively similar classes of images, the InceptionV3 model diverges substantially from the other models, producing adjusted Rand Indices (ARI) one order of magnitude lower other pairs of models. As newer pre-trained models–EfficientNetB1 and ConvNeXt–are much more similar to the earlier VGG- and ResNet-based models, this result suggests care should be taken when selecting pre-trained models, and one should compare results across several potential models. Second, while we demonstrate MoCs' image-sharing behaviors are predictive of their ideological leans, we find within-party ideological placement is more difficult to predict for Republican MoCs than their Democrat counterparts, regardless of the underlying pre-trained model, which has implications for partisan analyses. Lastly, when assessing construct validity for image "similarity" across image clusters, we find, for certain image clusters, human assessors and the pre-trained model disagree in whether a pair of images should be considered similar. That is, for certain types of images, manual and automated identification of similar images agree–e.g., screenshots, infographics, and cartoons–but for other types of images, especially photos showing groups of people, Amazon's Mechanical Turk (MTurk) manual assessment was more likely to disagree. This last point has implications for interpreting model results, as politicians may be using images from one of these problematic clusters in different ways, leading to confounded results (e.g., cluster 2 in EfficientNetB1, which is non-significant but has a higher rate of disagreement from MTurk.

Following a description of these results, we discuss connections between our findings and broader literature, especially around prior work on images in political discourse outside social media. The paper outlines recommendations for computational social science researchers who use these models and methods to examine online political discourse, including careful consideration of the types of images being shared, varying use of images by politicians of different parties, and comparing image analyses from multiple pre-trained models. We then close with a discussion of ethical considerations, threats to broader validity of this work, and potential future avenues to extend this research.

Contributions Overall, this work is meant to support those studying imagery in online political discourse, as suggested in Lilleker, Veneti, and Jackson (2019), with implications for political mobilization (e.g., Casas and Williams 2019), polarization (e.g., Tucker et al. 2018), and manipulation (e.g., Zannettou et al. 2020). Specific contributions include:

- An analysis of images US politicians share on Twitter across multiple pre-trained computer vision models;
- An assessment of the types of imagery that correlate with political ideology;
- The identification of several potential hazards in applying pre-trained models to characterize political imagery; and
- A set of recommendations for researchers when using pre-trained image models to study political ideology in visual media.

Related Work

Studies of imagery in political discourse have a substantial history (see, e.g., Seidman 2008), and new media spaces like social media platforms provide a new communication medium for politicians to engage with their constituents. Recent surveys of political operatives shows politicians strategically choose how to use these online spaces (Kreiss, Lawrence, and McGregor 2018), and the role of visual media in these platforms is becoming increasingly prominent: Auxier and Anderson (2021) shows where visually oriented platforms (e.g., YouTube, Instagram, Snapchat, and TikTok) all see a marked increase in popularity. At the same time, while much of recent work on politics and social media has focused on text and news sharing, the research community is increasingly calling for greater study of imagery in political discourse (Lilleker, Veneti, and Jackson 2019)–for instance, Tucker et al. (2018) identify visual media in online political disinformation as a crucial gap in the literature.

While studies have examined visual media in online spaces for political discourse–e.g., Casas and Williams (2019) shows online imagery to be particularly impactful in political mobilization–more recent scholarship has begun applying computer vision and machine learning models to facilitate analysis of political imagery. In this vain, Joo and Steinert-Threlkeld (2018) surveys automated methods for analyzing imagery and visual content for political science. Likewise, Xi et al. (2020) uses a pre-trained deep neural network architecture from He et al. (2015) to characterize images shared by US MoCs on Facebook, evaluating what facial expressions and visual components of the image best correspond to party affiliation. These efforts tend toward one of two paths, either generating image embeddings from some pre-trained computer vision model (Peng 2021; Joo and Steinert-Threlkeld 2022) or fine-tuning such a model (Xi et al. 2020; Zhang and Pan 2019), but which model to use or the implications of such a choice on the complex, real-world information space remain unclear.

Zhang and Peng (2022) provides foundational guidance for these decisions, but much of the work described therein leverages a single model architecture, based on VGG, with some comparison to ResNet. In this context, Zhang and Peng (2022) examines various cluster counts in grouping images, but how these counts might change across model architectures remains an open question. We address these open points by examining imagery politicians share in our first research question: **RQ1** – To what extent are the disparate types of imagery politicians share in consistent across deep-learning models?

Besides types of content politicians share, substantial work has examined signals of political ideology that emerge from sharing text (Diermeier et al. 2012), news (Messing, Kessel, and Hughes 2017), and social interaction (Conover et al. 2021). More recently, Xi et al. (2020) has extended these studies to image-sharing, and while that work yields valuable insights into politicians use of images, that work actively removes infographics and other non-photograph content, even stating that infographics appear more used by Democrats. We take a different approach here, extending Xi et al. (2020) to a larger set of politicians and images (about a 2x increase), and evaluate **RQ2** – to what degree an MoC's imagery is predictive of their ideological position, across these models.

We further this research by extending to multiple *types* of images rather than filtering out non-photographic images. Integrating the full spectrum of images is necessary, especially with the proliferation of the *visual meme* as potent form of political communication, as discussed in Du, Masood, and Joseph (2020). Such visual memes include photographs with textual overlays, as with President Trump's declaration that "Sanctions are Coming" overlayed on a picture of himself (Du, Masood, and Joseph 2020) to cartoons, drawings, or panels of images, and are known to play a role in manipulating online communities (Ratkiewicz et al. 2010; Zannettou et al. 2018). Alternatively, screenshots of text or posts have also become important non-photographic forms of visual communication, used to avoid censorship, especially during protests, as seen in the CASM dataset (Zhang and Pan 2019). Even in CASM, where images must first survive a textual relevance classifier, non-photographic images emerge in the data; and in Zhang and Peng (2022), a VGG16-based classifier trained on ImageNet collapses non-photographic images into three clusters of symbols and other meme-like images-with-text. These works suggests that deep learning models trained on primarily photographic datasets like ImageNet (Jia Deng et al. 2009), may perform unexpectedly, leading to a loss of construct validity with respect to "image similarity". To assess this potentiality, similar to CASM, we qualitatively assess image similarity within each class of images, leading to **RQ3**–to what degree do manual assessments about similar pairs of images agree with similarity measures from these pre-trained models?

Methods

Collecting Congressional Twitter Data

To analyze politicians' social media content, we leverage a directory of social media accounts for MoCs in the 112th through the 116th US Congressional sessions, covering 2011-2021, primarily from the @unitedstates project,[1] an open data repository about the US government. This project parses MoC web pages to extract social media identities, including Facebook and Twitter. While the @unitedstates project generally contains only social media identities for the current congressional session, the project stores its data in a revision control system. Using this revision history, we extract data from past sessions, manually augmenting and validating congresspeople's identities from these sessions (see Table 1 for statistics). We observe that, by the 116th session, nearly all MoCs have Twitter accounts, but by our collection period in 2022, nearly 100 accounts have changed Twitter handles or have been deactivated altogether.

Session	MoCs	Identified Twitter Accounts	Active Twitter Accounts in 2022
112^{th}	552	367	333
113^{th}	553	422	395
114^{th}	547	448	428
115^{th}	562	495	453
116^{th}	550	536	465
		Unique	656

Table 1: Collected US MoCs Twitter accounts by Session. Numbers of congresspeople do not equal seats in the House and Senate because members can be replaced mid-session.

Twitter Data We leverage Twitter's v2 API, with academic access to collect these politicians' entire timeline of tweets, going back to 2007. We have collected data from 656 Twitter accounts, totaling 3,381,028 tweets. On average, this dataset contains 5,154 tweets per account, with a minimum of 17, max of 41,870, and standard deviation of $\sigma = 4,649$ tweets. Regarding images, this data contains an average of 1,401 images per account, with a minimum of 1, maximum of 11,216 and median of 1,111 images.

Sampling Images from Congresspeople's Accounts For each of account, we randomly sample 1,111 images to ensure we get the entire set of images for at least half of the politicians in our dataset. This sample does not result in exactly 1,111 images per politician, as some politicians share fewer images. We do not perform additional processing on images following their download, to allow further analysis to focus on only on the original image content shared. In total, this dataset contains 486,604 images across these 617 politicians. We calculate the mean percent of tweets that contain an image as approximately 37% (i.e., more than one-third of each politician's tweets include an image).

Characterizing Types of Images

After extracting this image sample, we turn to **RQ1** to characterize the types of images MoCs share.

Feature Extraction We represent each image as an embedded feature vector, extracted from a deep learning model. To generate these embeddings, we select five distinct convolutional neural networks, ranging from older, more lightweight models to more modern ones: VGG19 (Simonyan and Zisserman 2015), ResNet50 (He et al. 2015), InceptionV3 (Szegedy et al. 2015), EfficientNetB1 (Tan and Le 2019), and ConvNext (Liu et al. 2022). Each model has been trained for image classification on the ImageNet (Deng et al. 2009) dataset for object recognition. Furthermore, each of these architectures, besides the more recent ConvNeXt

[1] https://theunitedstates.io/

model, has been used in studies of images in political discourse. Differences between neural architectures are primarily an increase in layer depth and decrease in number of individual weight parameters. In particular, VGG19 employs a direct feedforward sequential architecture, ResNet50 uses "residual connections" to merge inputs from prior layers, InceptionV3 builds these residual connections into residual *modules* (consisting of multiple residual layers), EfficientNetB1 uses novel dimension scaling to reduce complexity, and ConvNeXt incorporates the design of a transformer into a traditional ResNet-based convolutional network. The Keras Applications[2] module provides pre-trained ImageNet weights for each model.

To generate image embeddings rather than ImageNet class labels, we replace final model layer with a global average pooling layer. Each image is represented as a vector with dimension d–namely 512, 2048, 2048, 1280, and 2048 for VGG19, ResNet50, InceptionV3, EfficientNetB1, and ConvNeXt respectively. We resize images to $256 \times 256 \times 3$ for consistency and extract features in batches of 50.

Identifying Image-Types via Clustering To characterize the disparate types of visuals one can share online–photos, infographics, cartoons–as well as the general *content* of these types, we apply k-means clustering to group these images' feature vectors. We train five clustering models (one each for the different deep learning architectures) and set maximum iterations to 1,000.

To determine clusters k, we use the elbow method to identify inflection points in cluster quality metrics, using inertia (sum-of-squares optimization within each cluster). We plot this metric over $k \in [2, 20]$ and qualitatively determine these inflection points (see Figure 1), which appear between $k = 4$ and $k = 10$, so we settle on $k = 8$. For robustness, we have compared these results to the Davies-Bouldin index(Davies and Bouldin 1979) and silhouette scores (Shahapure and Nicholas 2020), finding consistent curves.

To assess the similarity across embedding-model clusterings, we use two metrics: First, we calculate the Jaccard similarity between all pairs of clusters and all pairs of models. Second, we use the adjusted Rand Index (ARI) to measure the overall similarity between a pair of clusterings. These metrics provide some insight into how consistent clusterings are across deep-learning models, as needed for **RQ1**.

Correlating Images and Political Ideology

For **RQ2** and correlating ideological position with images, we quantify congresspeople's ideology using their DW-NOMINATE scores (Lewis et al. 2021), a well-accepted measure of political position based on voting behavior in Congress. We then use two methods to answer this question: one using clustering to assess which image types reflect ideological lean, and another that predicts MoC ideology directly using their average image embeddings.

For our cluster-based analysis, we use a politician's distribution of images across clusters to predict the politician's ideological position. This model assigns each politician a numerical value on a left-to-right (i.e., liberal-to-conservative) scale, where negative scores represent stronger liberal positions, and positive scores indicate stronger conservative positions. We then construct a linear regression model to predict these ideology scores from a politician's distribution of images across clusters. The primary factors in this model are the proportions of images a politicians shares in each cluster, since politicians may have shared different numbers of images. To this end, we compile data for the regression model using a slightly reduced set of politicians. Out of 656 politicians in our dataset, we remove any accounts who have shared fewer than 20 images, or do not have a corresponding DW-NOMINATE score, leaving us with 627 politicians.

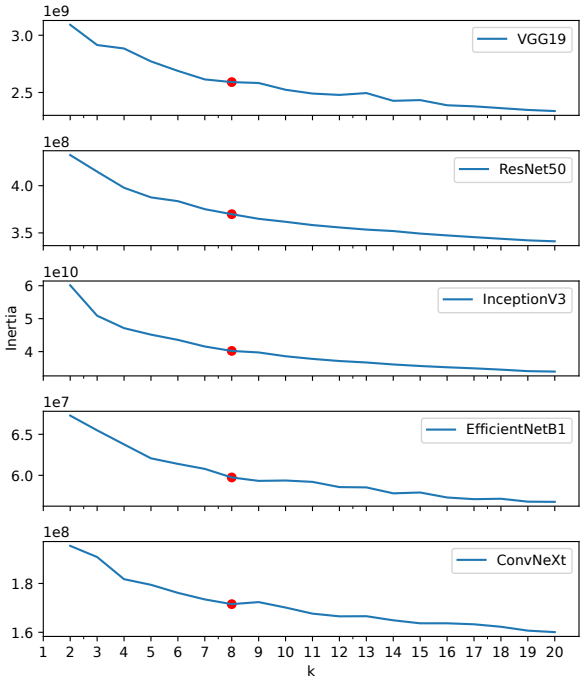

Figure 1: Inertia Measures per Cluster Count Across Pre-Trained Models. Across all five pre-trained models, we see elbow points at approximately $k = 8$.

For our second method to predict MoC ideology, we train a set of supervised learning models–one for each neural architecture–to predict DW-NOMINATE scores from an MoC's image embedding, averaged across all images that MoC has shared. Using repeated random sub-sampling across multiple regression models, we estimate the Pearson correlation coefficients between model output and DW-NOMINATE scores. We estimate this metric on the 20% held-out set of MoCs for each round of 128 random subsamples across both the full set of MoCs and within the Democratic and Republican parties separately, as in Xi et al. (2020) and Barberá (2017).

[2]https://keras.io/api/applications/

Manually Assessing Image Similarity

To evaluate **RQ3** and assess construct validity of image similarity in the context of these deep neural models, we leverage MTurk to crowdsource assessments for pairs of "similar" images. We sample similar pairs of images across each cluster and ask three MTurk workers to assess these pairs for their visual and semantic similarities. The more MTurk workers agree that pairs of images are similar, the higher the construct validity of image similarity in these embeddings.

Sampling Similar Images Our primarily objective in this assessment is to understand whether two images that are similar in the embedding space are also considered similar by a human assessor; that is, whether an individual would say these two images should in fact be considered similar. Hence, we extract samples of pairs containing similar images from each of the eight clusters produced from EfficientNetB1 embeddings, as this model has been used in prior work on visual political discourse.

For each cluster, we extract samples of highly similar images. Given the scale of this dataset, we cannot compute all possible pairs and instead use locality sensitive hashing (LSH) to find a sample of similar pairs. For LSH, we set an image-similarity threshold of Euclidean distances less than 10. This threshold was chosen based on a pilot study where we sampled pairs of images across the dataset to identify potential duplicates. Below this threshold, several clusters have no potential duplicates for manual assessment, and above this threshold resulted in samples that contained clearly different image-pairs, which would dilute the value of the human assessment. We then take the top 100 most similar pairs per cluster as our sample, creating 800 similar-image pairs.

Manually Assessing Visual and Semantic Similarity To measure how well these clusters actually capture image similarity, we use MTurk to crowdsource similarity assessments for each of these image pairs. In preliminary analysis, we have found that some pairs of images, such as screenshots of bills or letters, are highly similar visually but contain vastly different messages. This issue is problematic as clusters of semantically distinct images may exhibit differential use across the ideological spectrum, thereby confounding approaches that rely on visually oriented, automated image analysis. We have therefore developed a codebook to assess this visual and semantic similarity, available along with the raw data in the online data repository.This codebook describes five categories of image-pairs: identical images; visually and semantically similar images–visually similar but semantically distinct images; visually distinct but semantically similar images; and visually and semantically distinct images–plus a label for "unknown", where the assessor cannot determine similarity.

We create a task on the MTurk platform using these pairs with this codebook as instructions for the MTurk workers. Each image pair receives assessments from three workers, and we take the majority vote as the similarity assessment for that image pair. In a pilot study, MTurk workers have performed approximately three assessments a minute, so, to pay a minimum wage of $15 per hour, we pay $0.10 per assessment. This crowdsource project has also been reviewed and approved by our university Institutional Review Board.

Results

To ground the following analyses, Figure 2 shows example images from each cluster in the EfficientNetB1 clustering.

Consistency of Image Types Across Models

Examining the frequency of MoCs in each cluster, across all five models, we find that each cluster contains images from nearly all MoCs. On average, an MoC shares at least one image from each of the eight clusters, regardless of the underlying deep-learning model. This result shows that MoCs tend to share a variety of imagery as opposed to a single type, regardless of party affiliation and model.

These findings yield some insight into **RQ1** in that MoCs' behaviors are largely consistent across deep-learning models, but we still need to assess how consistent the clusterings are across these models. To this end, Figure 3 shows aligned clusters for each pair of clusterings. This figure shows VGG19, ResNet50, EfficientNetB1, and ConvNeXt produce relatively consistent clusters, as each of these pairs have *multiple* pairs of aligned clusters. In contrast, InceptionV3 consistently produces only a single cluster that is clearly aligned with another model's clusters–e.g., cluster 0 in EfficientNetB1. This examination of pairwise Jaccard similarity is supported by the ARI scores, where all pairs of clusterings excluding the InceptionV3 have an ARI > 0.2, whereas any pair with InceptionV3 has an ARI < 0.062, demonstrating that this model produces divergent results.

Correlating Images and Political Ideology

Moving to **RQ2**, we evaluate whether a politician's distribution of images across these eight clusters correlates with that politician's ideology via a linear model that regresses MoC's ideologies on cluster distributions. Table 2 shows this linear model fitted across clusterings from the five embeddings, demonstrating that an MoC's use of images in particular clusters significant correlates with ideology. These clusters are fairly balanced across the ideological spectrum as well, as nearly half of the clusters correlate with liberal and conservative ideological lean respectively. In EfficientNetB1, for example, the more an MoC posts images in Cluster 5, the more liberal they are likely to be, whereas the more images they post from Cluster 3, the more conservative their DW-NOMINATE position. Examining the adjusted R^2 for these models, we see that the linear model for ConvNeXt is the highest, followed by EfficientNetB1, ResNet50, and VGG19 in order. As in RQ1, InceptionV3 again deviates from this pattern, showing the lowest R^2 despite being a newer model than VGG19 and ResNet50.

Though these results demonstrate that certain image types consistently correlate with an MoC's ideology, we also assess the predictive power of an MoC's images. For each MoC, we train a Bayesian ridge regression model to predict DW-NOMINATE scores given an MoC's average image embeddings from EfficientNetB1. Following repeated

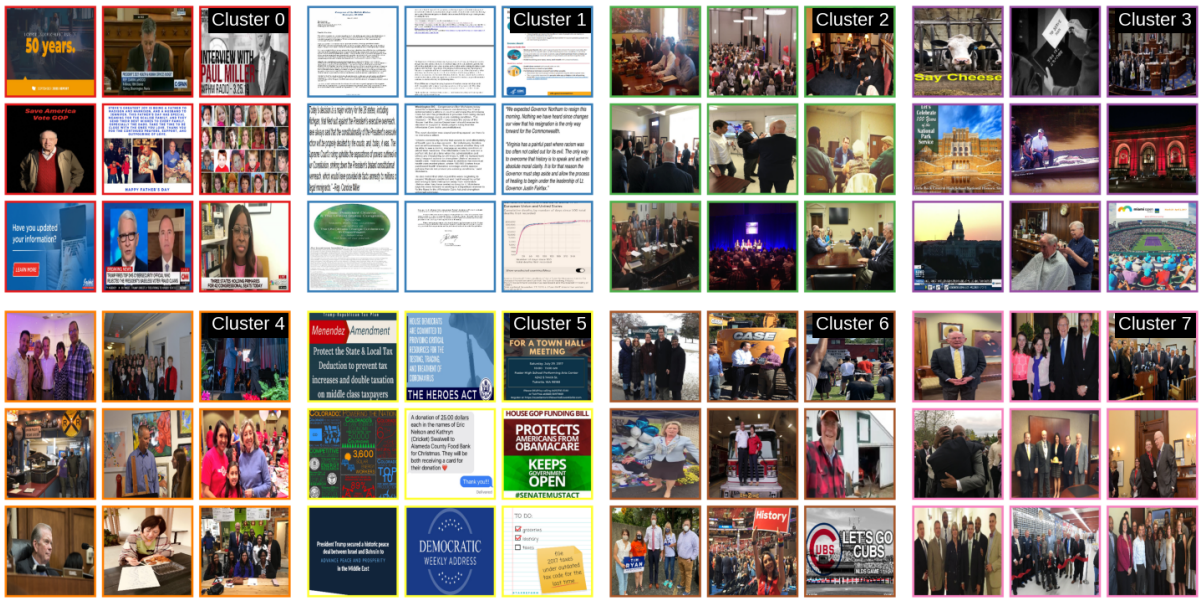

Figure 2: Sample Images from EfficientNetB1 Clustering. Clusters 1, 4, 5, and 6 are correlated with liberal ideology, and Clusters 3 and 7 are correlated with conservative ideology.

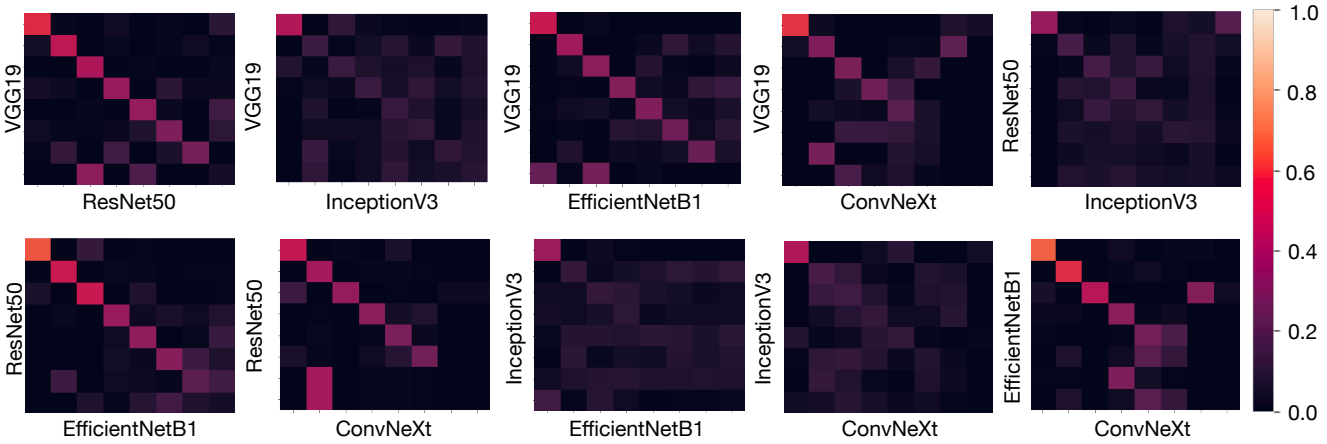

Figure 3: Jaccard Similarity Across Clusters and Embedding Models. Top-left cells in each grid represent the cluster pair with the highest Jaccard similarity between two models. VGG19, ResNet50, EfficientNetB1, and ConvNeXt produce relatively similar clusterings, whereas InceptionV3 diverges. VGG19 and ResNet50 produce the highest $ARI = 0.32$, with EfficientNetB1 and ConvNeXt with the second highest, $ARI = 0.31$.

random sub-sampling to estimate Pearson correlation coefficients, Figure 4 shows both the global ($\rho_A = 0.85$) and within-party correlation between predicted ideologies and DW-NOMINATE scores, where within-party correlations are separated by MoCs with Democratic ($\rho_D = 0.53$) and Republican ($\rho_R = 0.29$) party affiliations. We make two key observations from these results: First, both overall and within-party correlations outperform those presented in (Xi et al. 2020) by a wide margin (approximately 26% overall, 105% for Democrats, and 25% for Republicans), with Democrats' image-sharing model within 15% of the social network-based metric presented in (Barberá 2017). Second, the model consistently underperforms for Republican legislators compared to Democrats, where within-party ideology scores for Democrats are about twice as strong as Republicans. For robustness, we have checked these results across VGG19, ResNet50, and ConvNeXt as well as several regression models, and results are consistent both in comparison to prior work and in the Democrat-Republican asymmetry.

Assessing Image Similarity and Construct Validity

We have shown that MoCs share a variety of images (**RQ1**) and that images shared are predictive of their political ideology (**RQ2**). For **RQ3**, we investigate whether similarity

	VGG19		ResNet50		InceptionV3		EfficientNetB1		ConvNeXt	
	Coeff	*SE*	*Coeff*	*SE*	*Coeff*	*SE*	*Coeff*	*SE*	*Coeff*	*SE*
Const	-0.11***	0.04	-0.01	0.02	0.05	0.09	-0.05	0.03	-0.04	0.04
Cluster 0	0.58	0.35	-0.10	0.26	-0.74	0.46	-0.06	0.25	-0.03	0.27
Cluster 1	0.59**	0.22	0.87**	0.29	1.04	0.59	-1.01***	0.22	-1.03***	0.26
Cluster 2	-2.18***	0.37	-0.19	0.25	-1.38**	0.42	0.52	0.31	-3.53***	0.42
Cluster 3	-0.82*	0.35	-0.76**	0.26	1.64**	0.49	2.44***	0.25	0.77**	0.23
Cluster 4	0.59*	0.29	2.79***	0.29	-2.75***	0.50	-0.67*	0.33	0.71**	0.24
Cluster 5	1.37***	0.26	-0.91***	0.21	1.47**	0.50	-1.28***	0.24	-0.60***	0.13
Cluster 6	1.34***	0.28	0.25	0.28	2.77**	1.05	-0.96***	0.26	0.46**	0.17
Cluster 7	-1.57***	0.34	-1.96***	0.38	-1.20**	0.66	0.98***	0.26	3.19***	0.32
R^2	0.19		0.20		0.12		0.23		0.29	
									Note: $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$	

Table 2: Linear Regression Correlating Politicians' Cluster Distributions with DW-NOMINATE Scores. The more positive the coefficient, the more conservative-leaning the politician. Results show each embedding model produces five to seven clusters that correlate significantly with ideology.

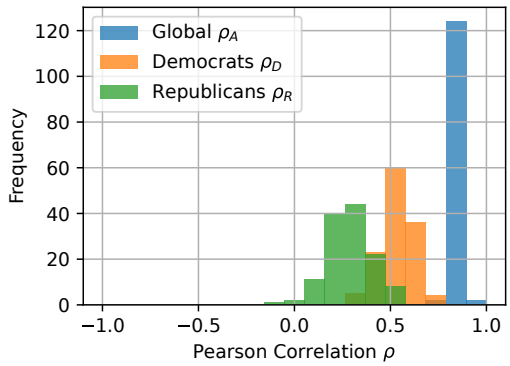

Figure 4: Overall and Within-Party Correlations Between DW-NOMINATE and EfficientNetB1 Predictions. Overall correlation is strong ($\rho_A = 0.85$), and within-party correlations for Democrats ($\rho_D = 0.53$) are substantially higher than for Republicans ($\rho_A = 0.29$).

Label	Raw Count ($n = 2,400$)	Majority Vote ($n = 605$)
Identical Image	1,020	338
Visually and Semantically Similar	726	186
Visually Similar but Semantically Distinct	318	46
Visually and Semantically Distinct	172	16
Visually Distinct but Semantically Similar	117	10
Unknown	47	9

Table 3: Distributions of MTurk Similarity Assessments. Most pairs are visually similar, but MTurk workers do not reach consensus in 195 of these image-pairs.

within a particular image type in embedding space are consistent human assessments–i.e., whether similarity in the metric space matches how an MoC might see image similarity. Using embeddings and resulting clusters from EfficientNetB1, we extract a sample of 100 highly similar pairs of images per cluster and use MTurk to collect three manual assessments of image similarity per pair. Table 3 shows label frequencies for raw counts across 384 MTurk workers and the label counts after taking the majority vote and removing 195 samples that lack clear consensus. Of the image pairs with a majority vote, MTurk workers say 94% are visually similar, with 8% being only visually similar but semantically distinct. This high proportion of agreement between embeddings and human assessment provides confidence in the construct validity of this image similarity metric.

Looking within individual clusters, however, a difference emerges. Figure 5 shows clusters 0, 1, 3, 5, and 6 predominantly contain images that are either identical or visually and semantically similar ($> 73\%$). Clusters 2, 4, and 7, however, have fewer identical image pairs and more pairs without a clear consensus ($36 - 46\%$), suggesting that construct validity is weaker for these types of images. Of these three EfficientNetB1 clusters, Table 2 shows Cluster 4 leans liberal, and Cluster 7 leans conservative (i.e., has a positive, significant coefficient). Cluster 4's significance is borderline, however, and Cluster 2 shows no significant predictive power at all. All three clusters primarily show groups of people in various situations, and similarity assessments may be difficult for humans in these cases. Consequently, for Cluster 2 (and Cluster 4 to a degree), the absence of significant effect in Table 2 may be attributable to differences between embeddings and human perception rather than ideological use. These results of identical and visually/semantically similar images in a majority of clusters is also consistent with a pilot we have run using the ResNet50 embedding model and manual annotation by two authors.

Discussion

The results above broadly demonstrate that multiple pretrained models for image characterization are largely consistent in both the types of imagery they identify and how those types of imagery correlate with politicians' ideological positions. Despite this consistency, our results also demon-

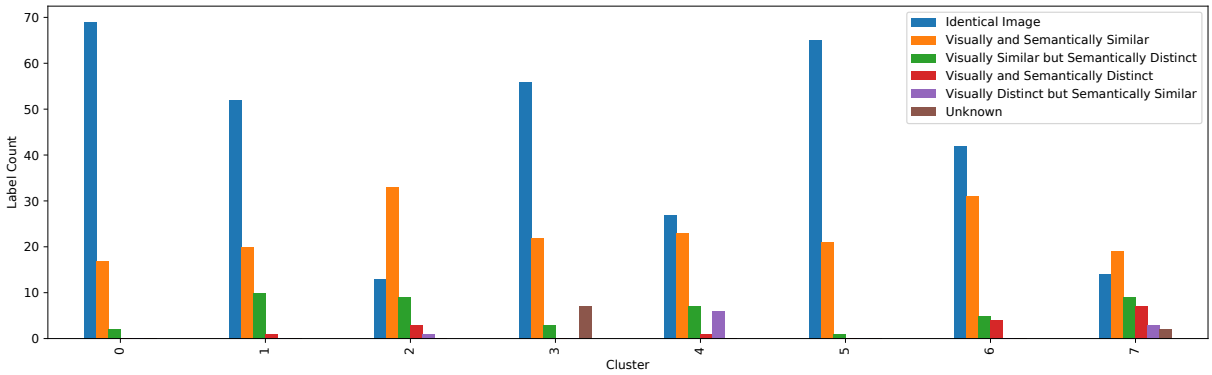

Figure 5: Crowdsourced Majority-Vote Labels of Similarity per EfficientNet B1 Cluster. Clusters 0, 1, 3, and 5 contain mostly identical images; clusters 2, 4, and 6 contain visually similar images; and cluster 7 contains more distinct images.

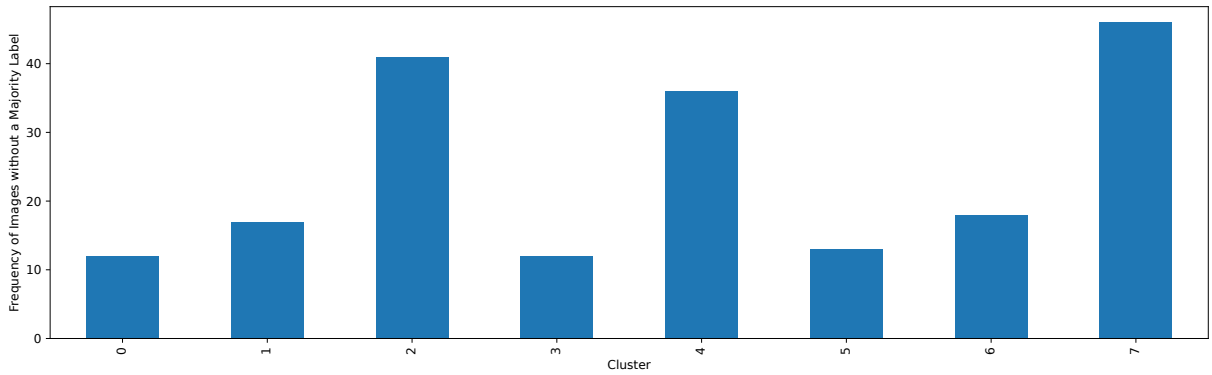

Figure 6: Image-Pairs without MTurk Consensus on Similarity by Cluster. Clusters 2, 4, and 7 have proven problematic for humans to agree on the degree to which images are similar.

strate potential areas of concern. Below, we discuss some of these factors, namely about the types of imagery we identify, how they are used across the political spectrum, and potential threats to our results.

Qualitative Analysis of Imagery

As discussed in **RQ1**, we observe that politicians share images from eight different clusters of images, with examples in Figure 2. Of these image clusters, some are more strongly correlated with a political ideology than others. A natural question then concerns what kinds of images are captured by these clusters.

In a post-hoc qualitative assessment of the clusters produced by each embedding model, we find substantial consistency in the kinds of images these clusters represent. First, we find that each image embedding model contains clusters of document-based imagery, like those in Figure 7a–Cluster 1 in EfficientNetB1, Cluster 5 in ResNet50, Clusters 3/7 in VGG19, Clusters 1/4 in InceptionV3, and Clusters 0/1/6 in ConvNeXt. Nicely, linear-model coefficients for these embedding models are consistent in that document-based images correlate significantly with liberal ideological lean.

Next, the majority of clusterings produce a cluster that contains patriotic imagery, like that shown in Figure 7b, including images of the US flag and other national symbols; Americana, like football games; and national holidays like the Fourth of July. This American-centric cluster corresponds to Cluster 3 in EfficientNetB1, and Cluster 4 in both ResNet50 and VGG19. As with clusters of document-images, these Americana clusters also significantly correlate with conservative imagery across these three embeddings.

Thirdly, infographics are commonly shared by politicians (Figure 7c). These images relate to political topics, such as who benefits from minimum wage increases or incidence of mass shootings across the US. These infographics are often in the form of statistics, showing a textual heading and a numeric value, such as the number of jobs in Oklahoma that depend on manufacturing (10,400) or the number of individuals who have signed up for healthcare since the Affordable Care Act's passage (7 million). EfficientNetB1's Cluster 5 captures these images, as does VGG19's Cluster 7 and ResNet50's Cluster 3, all of which significantly correlate with liberal ideological positions.

Lastly, images of people, especially groups of people are ubiquitous in the data (see Figure 7d), as one might expect given politicians' campaigns and engagements with their constituents, and the type of imagery on which Xi et al. (2020) focuses. These images come in a variety of styles,

(a) Document-Based Imagery

(b) Patriotic Imagery

(c) Infographic Imagery

(d) Imagery of People

Figure 7: Four Consistent Types of Images Used By US Politicians.

from posed photos and gladhanding to more candid shots of groups of various sizes. Clusters 2, 4, 6, and 7 in EfficientNetB1 capture these groups, with Clusters 2 and 4 primarily comprising candid images, with large/small groups of individuals engaged in some task and not facing the camera, respectively. EfficientNetB1's Clusters 6 and 7 instead show more posed images of individuals and groups. As noted above, these clusters of people are difficult for humans to assess, and the limited effect of Clusters 2 and 4 on ideology may be attributable to this difficulty. For Clusters 6 and 7, however, we do see significant ideological effects, suggesting liberal politicians tend to use posed images of smaller groups, whereas conservative politicians use images of larger groups. The above descriptions are, however, based on our qualitative descriptions after these images have already been grouped. Future work should investigate these descriptions more thoroughly.

Types of Images and Political Use

Our results show a differentiation between the types of imagery shared by conservative and liberal politicians: Politicians who share a larger proportion of patriotic or nationalistic imagery tend to be more ideologically conservative, whereas liberal politicians appear to share a larger portion of diverse imagery, including document-based imagery, and infographics.

Conservative politicians' use of patriotic imagery is well documented, as shown in (Muñoz and Towner 2017), where an analysis of presidential candidates' Instagram images demonstrates that Donald Trump posted more patriotic symbols than other candidates. That same study also shows patriotic symbols garnered more engagement from audiences than many other visual themes. Across the aisle, liberal MoC's use of infographics is consistent with a separate study on Instagram use by MoCs, where (O'Connell 2018) shows Republican MoCs are significantly less likely to share infographics and text-oriented posts.

These results suggest politicians across the ideological spectrum may use images for the same goal, but how they use images to express this goal vary. Both (Muñoz and Towner 2017) and (O'Connell 2018) support this interpretation, as (Muñoz and Towner 2017) shows presidential candidates on both sides used visuals for self-presentation and especially to present themselves in the "ideal candidate" frame through depictions showing superior statesmanship. For conservatives, though, (Muñoz and Towner 2017) and (O'Connell 2018) find these candidates express this frame through patriotic symbols and visuals of other political elites. These depictions potentially represent appeals to authority and social dominance respectively, and much work has shown these aspects are strong predictors of conservative ideology, as discussed at length in (Kugler, Jost, and Noorbaloochi 2014). In contrast, liberal politicians may instead express this theme via data-backed claims; e.g., (Miller,

Krochik, and Jost 2010) demonstrates liberal audiences rely more on systematic processing rather than heuristic processing and are more likely to be persuaded by hard data. Liberal politicians' use of infographic-style and text-laden visuals may then be more persuasive among their audiences.

Images Across Ideological Boundaries

While the above section discusses types of images that are correlated with political ideology, several types of images have no significant correlation. These clusters may represent "politically neutral" types of images, i.e., these clusters may contain images from politicians on both sides of the aisle. The use of such images – e.g., marketing graphics, as in cluster 0, or depictions of meeting with constituents, as in cluster 2 – may simply be a common part of the politician's role as an elected official, regardless of their ideological position. (O'Connell 2018) similarly notes little ideological correlation in MoCs use of images showing their constituents. Cluster 2 in particular presents an interesting case, as MTurk workers exhibit difficulty in assessing similarity therein, as shown in Figure 6.

Alternatively, these images may present ideological signal through semantic meanings rather than visual structure. Several images in cluster 0 contain political affiliations, representing politicians or icons traditionally associated with either Democrats or Republicans, and many contain captions that present politically oriented messages. For instance, certain images are praising conservative politicians or showing screenshots from conservative news channels, but other images contain quotes from liberal icons or information about events involving liberal politicians. Similarly, images from newscasts often look visually similar, but one may be from CNN while the other is from Fox News. Separating these types of images may be difficult for object-recognition-based frameworks. In this context, this cluster may balance itself out with an even number of strongly liberal and conservative images.

These observations yield two potential conclusions: First, certain types of imagery can be shared by any type of politician, crossing ideological boundaries, such as images of politicians engaging in everyday activities or glad-handing. Second, other types of imagery do not necessarily correlate with a political ideology because they have high visual similarity but entirely different semantic meaning, and the deep learning frameworks we use are unable to capture this semantic variation. While more research is needed into these types of imagery, one should exercise caution in using these deep-learning models with these types of images.

Broader Perspectives and Ethics

A Hazard in Studying Social Media Images As we mention above, our findings about humans' struggles in assessing similarity for certain kinds of images makes it is difficult to separate potential modeling issues from actual patterns in ideological use by politicians. While the image-embedding models used herein are common in computer vision, the types of images used in such tasks tend to have only one perceptual basis: an image's *visual* component. When analyzing imagery shared online, as in social media, variation in type of image (e.g., cartoons, image macros, etc.) and in semantics are equally significant aspects for image understanding. We therefore highlight the need for caution when directly transferring these computer-vision methods to this space. To handle this variety of images, methods that consider joint visual-semantic representations are needed.

This consideration is increasingly crucial as the popularity of memetic imagery increases, and adding captions to imagery being shared online becomes easier. (O'Connell 2018) in fact demonstrates that many younger politicians increasingly rely on these kinds of images in their political discourse. Prior work, such as DeViSE (Frome et al. 2013) and MemeSequencer (Dubey et al. 2018), have sought to provide these joint representations, but limited resources exist for robust pre-trained models, such as those used herein. Therefore, when applying machine learning models to political content, additional analysis and care are necessary to address these significant confounders.

Ethical Considerations The above work suggests potential weaknesses in image analysis models, especially with respect to the use of textual overlays to embed anti-social content like hate speech in images. That is, online spaces may have difficulty separating such content from more benign imagery, and having highlighted this weakness, we may have inadvertently alerted malevolent actors to a weakness they could exploit. Researchers are already working on this problem, but we specifically identify the problems these weaknesses may cause for research into online political discourse. We hope that by calling attention to these weaknesses and outlining potential mitigations, as we do above, this potential for exploitation will be somewhat mitigated.

Recommendations for Researchers Throughout our work, we identify several points researchers should consider when analyzing political imagery. First, we have identified several distinct types of imagery, some of which present difficulties for humans when assessing similarity–i.e., clusters 2, 4, and 7, as shown in Figure 6. In these instances, blindly applying pre-trained object-recognition models may prove problematic, as the main axis of meaning in these image types may not be conveyed by the objects present. On the other hand, MTurk workers and pre-trained models seem to largely agree on image similarity in other types of images–e.g., clusters 3 and 5–suggesting these image types may be more amenable to applications of pre-trained models. In short, researchers should be cognizant of the *type* of images they are likely to be analyzing and constrain themselves appropriately, as we see with Xi et al. (2020).

Second, researchers should consider the implications of variation in ideology across image types. If a particular political party or group uses a particular type of image more, as we see with conservatives and patriotic symbolism and with liberals and infographics, this selection has implications for how these groups may use different platforms. E.g., if a platform's affordance focuses more on photographic imagery than designed images, like infographics or memes, that affordance likely impacts how liberals may use it compared to conservatives. In such a case, missing out on a certain type of imagery can lead to omitting an entire political group, or

how they communicate using imagery. Hence, researchers should consider platform norms and affordances in the kinds of images that are popular and how such norms may differentially impact ideological expression. Twitter is, of course, no exception here, and researchers should likely pilot similar studies in other platforms accordingly.

Third and lastly, we observe that, of the five pre-trained models we have tested in this paper, four of them–VGG19, ResNet50, EfficientNetB1, and ConvNeXt–present highly correlated and similar results. In contrast, InceptionV3 appears to deviate substantially from results built from the other models. A shallow recommendation here might be to avoid InceptionV3, as it appears both deviant and has the lowest R^2 value in Table 2. As new models with novel architectures are released regularly, however, a more valuable researcher recommendation might be to evaluate results across several pre-trained models. By examining results across multiple such models, the researcher can demonstrate robustness to model selection and get some insight into consistency of results across these models. For example, our conclusions might be quite different had we used only InceptionV3, whereas they are likely to be similar if we had used one of the other four.

Threats to Validity

While $k = 8$ clusters enables a succinct analysis of image types and their political correlations, it is possible that more specific types of imagery may be hidden by this low number of clusters. For example, we suggest the null result for significance in some clusters above stems from a limitation in the image embedding model, but it could also stem from selecting a cluster count that is too coarse, artificially grouping these visually similar but semantically distinct images. To explore this possibility, we have also run our clustering and linear regression models for other values of k, namely $k = 12$ and $k = 20$. These results align with results for $k = 8$, with the new clusters subdividing the $k = 8$ clusters, but no unexpected ideological correlations emerge. This find suggests our results, namely ideological correlations and distributions of imagery across clusters, are robust to cluster-count selection. Given that our results are consistent across embedding models as well, and no cluster seems to have a dearth of images, and the number of clusters we find is consistent with both (O'Connell 2018) and (Muñoz and Towner 2017), we are confident in these results.

We conduct a similar robustness check using a smaller sample of imagery, to assess whether our results resilient to the selection of images. This sample contains a maximum of 21 images per account, for a total of 15,054 images. We find that similar types of imagery emerge with consistent ideological correlation, albeit weaker due to the reduced sample of imagery. This find further corroborates that our results are consistent across different quantities of imagery.

Other threats arise from our sampling strategies from Twitter and in sampling images from our politicians. First, while our timeframe is broad, covering 2011 to 2021, the Republican Party held the majority in the US House of Representatives throughout this timeframe. As (O'Connell 2018) suggests, MoCs behave differently when their party is in the majority versus the minority, so it is possible that this timeframe may bias our sample of MoCs toward particular types of images. As the makeup of Congress changes, future work could test this possibility.

Conclusions

Results demonstrate that the types of imagery shared in online social spaces can be a strong indicator of one's ideological position – at least in the context of US congresspeople. While this finding is consistent with existing literature, we also find that standard deep learning models for image characterization may capture only a portion of this structure. Across multiple models of image representation, two visually similar images can convey divergent ideological messages. Despite this divergence, these models result in consistent types of imagery and correlations with ideological positions. Certain types of imagery, however, appear to have limited to no significant effect related to ideology, and it is difficult to attribute that result to real patterns of behavior among MoCs or to semantic collapse in visually oriented image models. Additionally, while four of our five image-embedding models produce largely consistent clusters, one model, InceptionV3, diverges strongly from the other three, suggesting different visual interpretations based on model architectures despite similar training datasets. We therefore highlight the need for special care when applying automated methods to characterize the variety of images used in online political discourse.

References

Auxier, B.; and Anderson, M. 2021. Social Media Use in 2021. *Pew Research Center*, (April).

Barberá, P. 2017. Birds of the Same Feather Tweet Together: Bayesian Ideal Point Estimation Using Twitter Data. *Political Analysis*, 23(1): 76–91.

Casas, A.; and Williams, N. W. 2019. Images that Matter: Online Protests and the Mobilizing Role of Pictures. *Political Research Quarterly*, 72(2): 360–375.

Conover, M.; Ratkiewicz, J.; Francisco, M.; Goncalves, B.; Menczer, F.; and Flammini, A. 2021. Political Polarization on Twitter. *Proceedings of the International AAAI Conference on Web and Social Media*, 5(1): 89–96.

Davies, D. L.; and Bouldin, D. W. 1979. A Cluster Separation Measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-1(2): 224–227.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 248–255.

Diermeier, D.; Godbout, J.-F.; Yu, B.; and Kaufmann, S. 2012. Language and Ideology in Congress. *British Journal of Political Science*, 42(1): 31–55.

Du, Y.; Masood, M. A.; and Joseph, K. 2020. Understanding Visual Memes: An Empirical Analysis of Text Superimposed on Memes Shared on Twitter. *Proceedings of the International AAAI Conference on Web and Social Media*, 14: 153–164.

Dubey, A.; Moro, E.; Cebrian, M.; and Rahwan, I. 2018. MemeSequencer: Sparse matching for embedding image macros. *The Web Conference 2018 - Proceedings of the World Wide Web Conference, WWW 2018*, 2: 1225–1235.

Frome, A.; Corrado, G. S.; Shlens, J.; Bengio, S.; Dean, J.; Ranzato, M. A.; and Mikolov, T. 2013. DeViSE: A Deep Visual-Semantic Embedding Model. In Burges, C. J. C.; Bottou, L.; Welling, M.; Ghahramani, Z.; and Weinberger, K. Q., eds., *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2015. Deep Residual Learning for Image Recognition. arXiv:1512.03385.

Jia Deng; Wei Dong; Socher, R.; Li-Jia Li; Kai Li; and Li Fei-Fei. 2009. ImageNet: A large-scale hierarchical image database. 248–255.

Joo, J.; and Steinert-Threlkeld, Z. C. 2018. Image as Data: Automated Visual Content Analysis for Political Science.

Joo, J.; and Steinert-Threlkeld, Z. C. 2022. Image as Data: Automated Content Analysis for Visual Presentations of Political Actors and Events. *Computational Communication Research*, 4(1).

Kreiss, D.; Lawrence, R. G.; and McGregor, S. C. 2018. In Their Own Words: Political Practitioner Accounts of Candidates, Audiences, Affordances, Genres, and Timing in Strategic Social Media Use. *Political Communication*, 35(1): 8–31.

Kugler, M.; Jost, J. T.; and Noorbaloochi, S. 2014. Another Look at Moral Foundations Theory: Do Authoritarianism and Social Dominance Orientation Explain Liberal-Conservative Differences in "Moral" Intuitions? *Social Justice Research*, 27(4): 413–431.

Lewis, J. B.; Poole, K.; Rosenthal, H.; Boche, A.; Rudkin, A.; and Sonnet, L. 2021. Voteview: Congressional Roll-Call Votes Database.

Lilleker, D. G.; Veneti, A.; and Jackson, D. 2019. Introduction: Visual Political Communication. In Veneti, A.; Jackson, D.; and Lilleker, D. G., eds., *Visual Political Communication*. ISBN 978-3-030-18729-3.

Liu, Z.; Mao, H.; Wu, C.-Y.; Feichtenhofer, C.; Darrell, T.; and Xie, S. 2022. A ConvNet for the 2020s. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Messing, S.; Kessel, P. v.; and Hughes, A. 2017. Sharing the News in a Polarized Congress. Technical report, Pew Research Center.

Miller, A. L.; Krochik, M.; and Jost, J. T. 2010. Political Ideology and Persuasion: Systematic and Heuristic Processing Among Liberals and Conservatives. *The Yale Review of Undergraduate Research in Psychology*, 14–28.

Muñoz, C. L.; and Towner, T. L. 2017. The image is the message: Instagram marketing and the 2016 presidential primary season. *Journal of Political Marketing*, 16(3-4): 290–318.

O'Connell, D. 2018. #Selfie: Instagram and the United States Congress. *Social Media and Society*, 4(4).

Peng, Y. 2021. What Makes Politicians' Instagram Posts Popular? Analyzing Social Media Strategies of Candidates and Office Holders with Computer Vision. *The International Journal of Press/Politics*, 26(1): 143–166.

Ratkiewicz, J.; Conover, M.; Meiss, M.; Gonçalves, B.; Patil, S.; Flammini, A.; and Menczer, F. 2010. Detecting and Tracking the Spread of Astroturf Memes in Microblog Streams. *CoRR*, abs/1011.3.

Seidman, S. 2008. *Posters, Propaganda, and Persuasion in Election Campaigns Around the World and Through History*. New York: Peter Lang Publishing, Inc. ISBN 978-0820486161.

Shahapure, K. R.; and Nicholas, C. 2020. Cluster Quality Analysis Using Silhouette Score. In *2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA)*, 747–748.

Simonyan, K.; and Zisserman, A. 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. arXiv:1409.1556.

Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; and Wojna, Z. 2015. Rethinking the Inception Architecture for Computer Vision. arXiv:1512.00567.

Tan, M.; and Le, Q. V. 2019. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. In Chaudhuri, K.; and Salakhutdinov, R., eds., *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, 6105–6114. PMLR.

Tucker, J.; Guess, A.; Barber, P.; Vaccari, C.; Nyhan, B.; Seigel, A.; Sanovich, S.; and Stukal, D. 2018. Social Media, Political Polarization, and Political Disinformation: A Review of the Scientific Literature. Technical Report March, Hewlett Foundation.

Xi, N.; Ma, D.; Liou, M.; Steinert-Threlkeld, Z. C.; Anastasopoulos, J.; and Joo, J. 2020. Understanding the Political Ideology of Legislators from Social Media Images. *Proceedings of the International AAAI Conference on Web and Social Media*, 14: 726–737.

Zannettou, S.; Caulfield, T.; Blackburn, J.; De Cristofaro, E.; Sirivianos, M.; Stringhini, G.; and Suarez-Tangil, G. 2018. On the Origins of Memes by Means of Fringe Web Communities. In *Proceedings of the Internet Measurement Conference 2018*, IMC '18, 188–202. New York, NY, USA: Association for Computing Machinery. ISBN 9781450356190.

Zannettou, S.; Caulfield, T.; Bradlyn, B.; De Cristofaro, E.; Stringhini, G.; and Blackburn, J. 2020. Characterizing the Use of Images in State-Sponsored Information Warfare Operations by Russian Trolls on Twitter. In *International Conference on Web and Social Media*.

Zhang, H.; and Pan, J. 2019. CASM: A Deep-Learning Approach for Identifying Collective Action Events with Text and Image Data from Social Media. *Sociological Methodology*, 49(1): 1–57.

Zhang, H.; and Peng, Y. 2022. Image Clustering: An Unsupervised Approach to Categorize Visual Data in Social Science Research. *Sociological Methods & Research*, 004912412210826.

Ethics Checklist

1. For most authors...
   (a) Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures? Yes, answers to these questions about politicians does not violate social contracts or other normative aspects, to the authors' knowledge.
   (b) Do your main claims in the abstract and introduction accurately reflect the paper's contributions and scope? Yes.
   (c) Do you clarify how the proposed methodological approach is appropriate for the claims made? Yes, see Methods.
   (d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? Yes.
   (e) Did you describe the limitations of your work? Yes; see Threats to Validity.
   (f) Did you discuss any potential negative societal impacts of your work? Yes; see Broader Perspectives and Ethics.
   (g) Did you discuss any potential misuse of your work? No; while our work reveals potential hazards in these methods, we do not describe how one might intentionally abuse these hazards.
   (h) Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymization, responsible release, access control, and the reproducibility of findings? Yes; we make the underlying image embeddings available for researchers to replicate this work.
   (i) Have you read the ethics review guidelines and ensured that your paper conforms to them? Yes.
2. Additionally, if your study involves hypotheses testing...
   (a) Did you clearly state the assumptions underlying all theoretical results? NA
   (b) Have you provided justifications for all theoretical results? NA
   (c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? NA
   (d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? NA
   (e) Did you address potential biases or limitations in your theoretical framework? NA
   (f) Have you related your theoretical results to the existing literature in social science? NA
   (g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? NA
3. Additionally, if you are including theoretical proofs...
   (a) Did you state the full set of assumptions of all theoretical results? NA
   (b) Did you include complete proofs of all theoretical results? NA
4. Additionally, if you ran machine learning experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? Yes.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? Yes.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? Yes.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? No.
   (e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? Yes; see Discussion.
   (f) Do you discuss what is "the cost" of misclassification and fault (in)tolerance? Yes; see Discussion.
5. Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? NA
   (b) Did you mention the license of the assets? NA
   (c) Did you include any new assets in the supplemental material or as a URL? NA
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? NA
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? NA
   (f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR (see **?**)? NA
   (g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset (see **?**)? NA
6. Additionally, if you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots? Yes. See OSF supplemental material.
   (b) Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals? Yes. This work was evaluated by the University of Maryland's IRB and identified as exempt.
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? No.
   (d) Did you discuss how data is stored, shared, and deidentified? Yes. See OSF supplemental material.