Visual Topics via Visual Vocabularies

Shreya Havaldar* **Weiqiu You*** **Lyle Ungar** **Eric Wong**
Department of Computer and Information Science
University of Pennsylvania
Philadelphia, PA 19104
{shreyah,weiqiuy,ungar,exwong}@cis.upenn.edu

Abstract

Researchers have long used topic modeling to automatically characterize themes in a text corpus without supervision. Can we extract similar structures from collections of images? To do this, we propose *visual topic modeling*, a method to analyze image datasets by decomposing images into segments, and grouping similar segments into visual "words". These vocabularies of visual "words" enable us to extract visual topics that capture hidden themes distinct from what is captured by classic unsupervised approaches. We evaluate our visual topics using standard topic modeling metrics and confirm the interpretability of our visual topics via a human study.

1 Introduction

The structure of written language is simple — words are comprised of letters, and documents are comprised of words. The body of words used in English, or the English *vocabulary*, allows us to create meaningful text from this fixed collection of words. This vocabulary-based structure also allows us to use certain natural language processing (NLP) techniques, like topic modeling, to understand and interpret language datasets.

Topic modeling is an unsupervised learning algorithm that extracts latent variables from large datasets in the form of topics [63]. These topics capture groups of related words within a body of text, and experts have directly interpreted topics in domains like medicine and social science [18, 42, 55] to understand and discover information within datasets. Topic modeling is an especially desirable explanation technique due to (1) the human-interpretability of topics, and (2) the uncovering of *relationships* between words in a document that are not always semantically similar. For example, "ball", "field", and "jersey" are related in that they may appear together in a "sports" topic even though these three words represent different objects and are not close in a typical word embedding space.

Given its popularity in NLP, we hypothesize that topic modeling could also discover relationships within images, where topics are distributions over related "words" in an image. These relations are distinct from image clusters, which often explicitly optimize for visual similarity. However, topic modeling algorithms process documents of words, whereas images do not have any such obvious linguistic structure. Getting topics for images in the same way we get linguistic topics first requires a vocabulary for images. Therefore, we develop a "visual vocabulary": a mapping of similar segments into "visual words" that constitute a discrete vocabulary for image data. We then use this visual vocabulary to transform images into a document representation that is compatible with classic topic modeling algorithms. The resulting methodology is called *visual topic modeling*: a procedure for extracting topical relationships from image datasets.

Our contributions are as follows: (1) We propose visual topic modeling as a way to explain hidden themes in an image dataset. Our methodology derives a visual vocabulary from images as an interface between image datasets and topic modeling algorithms. (2) We demonstrate, via experiments and a

XAI in Action: Past, Present, and Future Applications @ NeurIPS 2023.

Figure 1: Examples of visual topics for the SUN397 dataset. For each topic T, we show four exemplar documents D which maximize $P(T|D)$, as well as the corresponding segments w with high $P(w|T)$. Both ($P(T|D)$ and $P(w|T)$) are outputs of the topic modeling algorithm. Topic 1 contains leafless trees in the wintertime, highlighting the bare branches in addition to the skies behind them. Topic 2 contains segments of the Savannah, along with elephants in a similar landscape. Lastly, Topic 3 shows a variety of buildings with different textures and colors, but similar window layouts.

theoretical example, how visual topics capture relationships in images distinct from what existing dimensionality reduction methods capture. (3) We adopt standard topic modeling evaluations from the NLP literature to assess our visual topics. We find our topics to be of good quality according to topic modeling metrics and highly interpretable via human evaluation.

See Appendix B for related work.

2 Visual Topics via Visual Vocabularies

Based on its wide-reaching success in NLP, topic modeling can potentially capture meaningful relationships between components of images. However, there is a type-mismatch: images in vision are collections of pixels, whereas topic modeling algorithms expect documents consisting of sequences of words. In order to use topic modeling on image datasets, we must first transform images into documents of "visual words" to interface with topic modeling algorithms.

We formalize this problem as finding a mapping $h : \mathcal{X} \to \mathcal{V}^m$ from images $\mathcal{X}$ to documents $\mathcal{V}^m$, where documents are a sequence of m words from a discrete vocabulary $\mathcal{V}$. Once we have mapped our examples into documents with h, we can apply standard topic modeling algorithms. But how exactly should we map images into documents? As a naive approach, consider implementing h directly with an image-to-text model, such as CLIP, [54] to directly create a caption for each image. However, captioning models may fail to translate all key components of an image into text [61].

We instead propose a two-step mapping that preserves all of the components: we first partition an image into a list of m segments with a segmenter $s : \mathcal{X} \to \mathcal{X}^m$, where $s(x) = (z_1, \ldots, z_m)$ for segments $z_j \in \mathcal{X}$. To transform these segments into visual words, we define $v : \mathcal{X} \to \mathcal{V}$ which maps each image segment to a word from a discrete vocabulary $\mathcal{V}$. Then, we can define our visual document generator h as $h(x) = [v(s(x)_1), \ldots, v(s(x)_m)]$.

In summary, we break an image into segments and map each segment to a word in a discrete visual vocabulary. Each image's corresponding words are then concatenated together to create a document. These documents can then be fed into a topic modeling algorithm such as Latent Dirichlet Allocation (LDA). The whole approach, which we call visual topic modeling, is detailed in Algorithm 1.

Algorithm 1 Visual Topic Modeling

Require: Images $x_1, x_2, \ldots, x_n$, Segmenter s, Embedding Model f
Require: Integers K, T where $K \leftarrow$ num clusters, $T \leftarrow$ num topics

for $i = 1 \ldots n$ **do** ▷ Segmenting
  $s_i \leftarrow s(x_i)$
end for
$c_1, c_2, \ldots, c_K \leftarrow k\text{-means}(\bigcup_i^n f(s_i))$ ▷ Clustering
for $i = 1 \ldots n$ **do** ▷ Document Construction
  for $j = 1 \ldots |s_i|$ **do**
    $d_{ij} \leftarrow \text{argmin}_{l=1\ldots K} \|f(s_{ij}) - f(c_l)\|_2^2$
  end for
end for
$t_1, t_2, \ldots, t_T \leftarrow \text{LDA}(d_1, d_2, \ldots, d_n)$ ▷ Visual Topic Modeling

Creating a Visual Vocabulary. While there exist standard segmentation methods for s, we need to create a suitable vocabulary generator v for vision. A key part of our approach is to therefore construct a discrete, visual vocabulary $\mathcal{V}$, and define a mapping v from segments to $\mathcal{V}$. After segmenting our image into segments $(z_1, \ldots, z_m) = s(x)$, we use an image embedding model f from Dosovitskiy et al. [20] to extract embeddings for each unique segment. We then cluster the resulting segment embeddings into K clusters $(c_1, \ldots, c_K)$ using k-means. Each of the K clusters then is taken to be a discrete *word* in the visual vocabulary, resulting in a vocabulary of size $|V| = K$. Lastly, we assign each segment to its nearest cluster in embedding space to map segments to words. This procedure results in the following vocabulary generator: $v(z) = \arg\min_{l \in \{1,\ldots k\}} \|f(z) - f(c_l)\|_2^2$

Note that this vocabulary is not the English language; however, it mimics the *structure* of language. English has a fixed set of words that appear across documents. Similarly, the visual vocabulary is a fixed set of clusters shared across different images. Figure 2 shows an example visual vocabulary.

Visual Documents and Topic Modeling. With our mapping from images to a discrete visual vocabulary, we now have a natural way to form documents. We concatenate the words corresponding to the segments in an image to get a visual document (see Figure 2). With these documents, we can directly apply any topic modeling algorithm to our image dataset, creating *visual topics*. The resulting topics serve as an interpretable dimensionality reduction to explain an image dataset. In this work, we use LDA [2] as our topic modeling algorithm, given its simplicity and establishment in the field.

3 Topic Modeling Uncovers Relations in Images

The hallmark of topic models in language is the ability to discover relations between words that constitute overall themes. For example, a political topic in language can include words such as "government", "president", and "state" even though these are distinct entities. Our visual topics can uncover similar structures in image data, with several examples shown in Figure 1. Topic 1 contains bare tree branches, clear skies, and cloudy backdrops. Crucially, these three components have different shapes and colors and are not clustered together. In this section, we analyze what makes these topic relations different from classic interpretable structures for images such as clusters.

Topic Relations for Images are Different from Classes and Clusters. Clustering has been used as the predominant approach to group and understand patterns in images [24, 58, 5], whereas topic modeling is a more popular approach in language [29, 63]. However, clusters and topics extract fundamentally distinct structures from the underlying data.

In vision, Clustering algorithms explicitly optimize for groups with high similarity, and thus clusters tend to contain only visually *similar* images of homogeneous classes [3]. On the other hand, topics capture *related* words that co-occur together but may not be similar. For example, the words "torment" and "torrent" are both similar in character distance, but have very different meanings and are not

[2] Latent Dirichlet Allocation (LDA) is probabilistic topic modeling technique that assumes each document in a corpus is a mixture of topics, and each topic is a distribution over words. The algorithm uses Bayesian inference to return the most likely topic-word distributions – $P(w|T)$ for all words w and topics T, and document-topic distributions – $P(T|D)$ for all topics T and documents D.

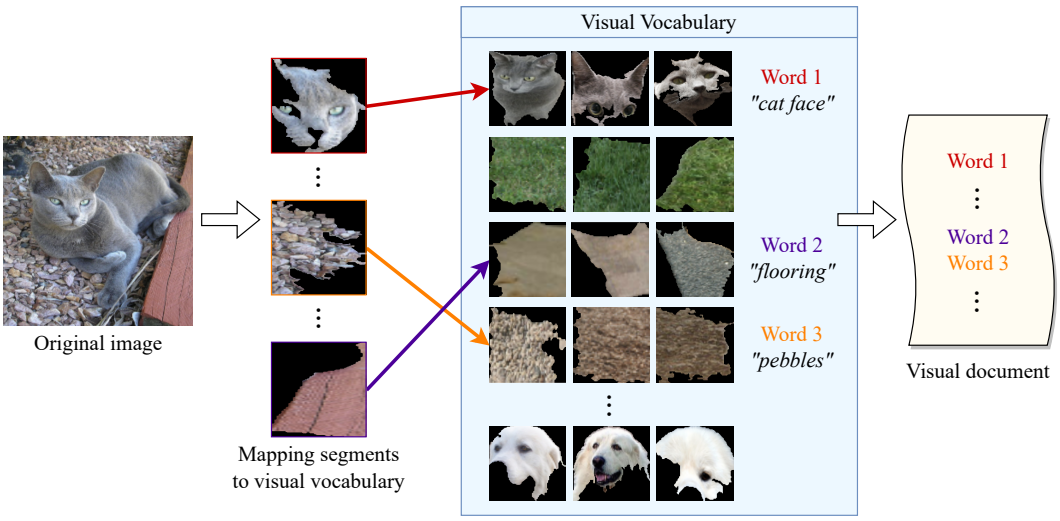


Figure 2: Mapping an image from the Pets dataset [52] to a visual document using our constructed visual vocabulary. We construct a discrete vocabulary for an image dataset by segmenting each image and clustering similar segments together, treating each cluster as a "word". We then map each image to a visual document by mapping each segment to its visual "word".

related. Conversely, "government" and "president" are quite different in character distance, but are related political entities. So, we expect the resulting visual topics to capture different information from clusters, potentially structures that span across multiple classes instead of being homogeneous.

To measure this difference, we get visual topics for all datasets listed in Appendix A.3 and calculate the entropy of the top 50 images within each topic[3] — see Figure 3a. An entropy of 0 indicates the top images in each topic all originate from a singular class, while an entropy of 1 indicates a uniform distribution across all classes. We observe that vast majority ($> 95\%$) of our topics have an exceptionally high entropy of 0.125 or greater, indicating they capture more than just visual similarity.

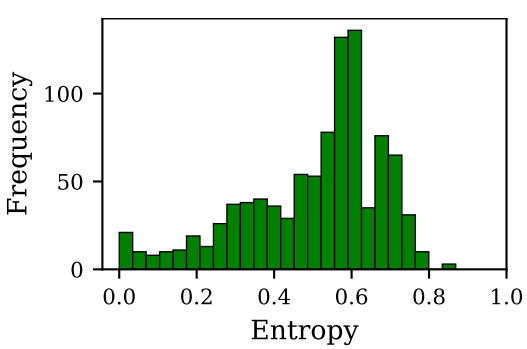
Given that clusters of images are built to be homogeneous with respect to class, they do not capture relations that span multiple classes as visual topics do. To capture more fine-grained structure than classes, another line of work has explored clustering segments instead of the entire image [24]. Although Ghorbani et al. [24] focuses on per-class clusters, segment clusters are analogous to word clusters in our visual vocabulary. As we well see in the next section, segment clusters are theoretically unable to capture relations of segments.

The Inability of Clusters to Capture Relations. In this section, we *theoretically* analyze the difference between clusters and topics. Using the example shown in Figure 3b, we prove that similarity-based clustering techniques cannot capture relatedness. For example, assume features, or "words", $z_1, z_2, \ldots z_d$ are generated from one of two topics. Each topic is a mixture of two Gaussian distributions. Informally, two features are highly related if they are generated from the same topic. Figure 3b plots a sample of these generated features, where blue points are generated from Topic 1 and orange points are generated from Topic 2. Examples, or "documents", are generated by concatenating a sequence of words, or features from the same topic. Example 1 in Appendix A.2 formalizes this generation procedure as a Mixture of Related Gaussians.

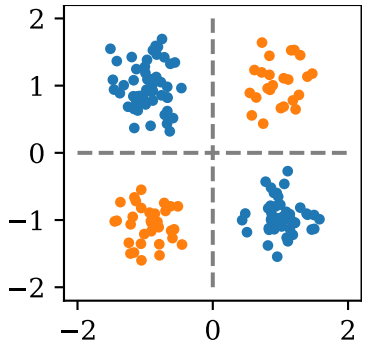
Theorem 1. *Let z_{ij} be generated according to Example 1 with $\sigma \leq 0.288$. With probability at least 0.99 there does not exist a clustering of the features z_{ij} that has two clusters containing only related points from each pair of opposing quadrants.*

To quantify the relatedness of two features, we use pointwise mutual information (PMI) [40, 8]. Two words in a topic, or two features from the same topic, should have a high PMI. For example, two blue points in Figure 3b will have a higher PMI than an orange/blue pair. Property 1 formalizes this claim in Appendix A.2.

[3]For a given topic T, we use the document-topic distribution returned by LDA to select the set of documents $D_1, D_2, \ldots D_N$ where $P(T|D)$ is maximized. We call this set the top N documents for T.



(a) Entropy of visual topics on a suite of image datasets (A.3). Our topics originate from a diverse set of classes, highlighting the difference between topics and clusters.



(b) Mixture of Related Gaussians. "Words" are generated from orange and blue topics, where opposite quadrants are highly related.

Figure 3: Topics are different than clusters: (a) graphs the class entropy of each visual topic, showing that our topics span multiple classes. (b) exemplifies the relatedness property as a mixture of related Gaussians; there does not exist a clustering that separates the two topics while keeping them intact.

The data from Example 1 can be viewed as a mixture of 4 Gaussian distributions with a twist: data from opposite Gaussians are highly correlated. In other words, pairs of features with the same sign are likely to show up together in examples. Conversely, pairs of features with opposite signs are unlikely to appear together in an example. These pairs of opposing quadrants satisfy the *relatedness* property, while adjacent quadrants are not related. From a language perspective, x_i is analogous to a visual document, and z_{ij} is analogous to the jth visual word in the ith document.

In Appendix A.2, we prove that (1) we cannot cluster distinct but related features from opposite quadrants simultaneously for both topics, and (2) a topic model can successfully group all related blue and orange words into two different topics — see Theorem 1 and Corollary 1. We theoretically prove that topics discover relationships that clusters cannot represent. Our topics capture themes grounded in relatedness and are thus a complimentary addition to classic unsupervised explanations.

4 Visual Topic Evaluation

To assess the strength of our visual topics as explanation tools, we evaluate them with well-established metrics in the topic modeling literature [40, 1]: **(1) Human Interpretability**: We conduct a word intrusion test [13] for our visual topics to evaluate their human-interpretability. **(2) Internal Coherence**: We measure how semantically similar the words within each obtained topic are to maximize topic clarity [48]. **(3) Relatedness**: We measure how related the words within each topic are. (e.g. "ball" and "jersey" are related, whereas "jersey" and "uniform" are similar.) [17]. **(4) Diversity**: We measure topic diversity to minimize redundancy [16].

Experiments. We apply our visual vocabularies algorithm outlined in Algorithm 1 to a range of image datasets (see Appendix A.3 for descriptions). We then run LDA [10] on the resulting documents for each dataset, setting $N_{Topics} = |\text{Num Classes}|/2$. See Appendix A.1 for experimental setup detailing choice of segmenter S, number of clusters K, embedding model F, and LDA priors.

4.1 Human Evaluation

A goal of any good explanation method is human-interpretability [66, 19]. Linguistic topics have been shown to be understandable to humans [10], hence their popularity as an explanation tool in NLP. To assess the human-interpretability of our visual topics, we conduct a user study. A standard way to assess how well linguistic topics match human concepts is through a task known as *word intrusion* [13] – users are given a set of words with high probability in Topic A, with the exception of an "intruder", or a word with very low probability in Topic A. Users are then asked to identify this intruder word. If topic A is highly coherent and has some common underlying theme, it should be clear to a user which word is not part of topic A.

	Aircraft	*Birdsnap*	*Caltech-101*	*Caltech-256*	*Cars*	*CIFAR-10*	*CIFAR-100*	*DTD*	*Flowers*	*Food*	*Pets*	*SUN397*
Internal Coherence												
Visual	**0.86**	**0.91**	**0.84**	**0.85**	**0.79**	**0.95**	**0.99**	**0.74**	**0.87**	**0.93**	**0.85**	**0.87**
Text	0.24	0.25	0.24	0.25	0.21	0.24	0.24	0.25	0.26	0.26	0.25	0.25
Shuffled	0.69	0.56	0.59	0.55	0.61	0.70	0.78	0.59	0.67	0.67	0.60	0.55
Topic Relatedness												
Visual	0.15	0.17	**0.2**	**0.19**	0.02	**0.19**	**0.24**	**0.1**	**0.17**	**0.21**	**0.2**	**0.17**
Text	**0.22**	**0.19**	0.1	0.08	**0.14**	0.12	0.09	0.08	0.06	0.1	0.05	0.1
Shuffled	-0.29	-0.25	-0.32	-0.28	-0.29	-0.05	-0.16	-0.35	-0.34	-0.11	-0.2	-0.22
Topic Diversity												
Visual	**0.76**	**0.83**	**0.88**	**0.89**	**0.77**	**0.97**	**0.87**	**0.99**	**0.92**	**0.89**	**0.97**	**0.89**
Text	0.14	0.11	0.37	0.32	0.2	0.26	0.38	0.36	0.23	0.33	0.35	0.23
Random	0.56	0.56	0.58	0.58	0.57	0.65	0.60	0.58	0.57	0.61	0.62	0.60

Table 1: As is standard in topic modeling literature, we evaluate our visual topics for internal coherence, relatedness, and diversity. We compare against text topics (i.e. LDA on the text captions of each image) and randomly shuffled/sampled topics, where we report the mean of 100 runs.

Visual word intrusion. We design a parallel intrusion task for vision. Given a topic T, we first select 4 "exemplar" images $D_1, \ldots D_4$ for that topic, or the images with the largest $P(T|D)$. We highlight all segments w with where $P(w|T)$ in the upper quartile. We dim the remaining image so users can focus on the relevant words, while still having the context of the full image to make sense of the unobscured segments. Users are then shown the 4 exemplar images for each topic, along with a randomly chosen intruder image, which is an exemplar document for a different topic. Users are then asked to identify the intruder (see Figure 4 for example).

We run LDA on the SUN397 dataset using 150 topics, and have a group of 12 volunteer users analyze 25 topics each (2 users evaluate each topic). We observe exceptionally high performance on this task — on average, users select the correct intruder 87.67% of the time. We also observe high agreement between users (average Cohen's kappa = 0.828). We evaluate against Chang et al. [13]'s word intrusion task on two text corpora, using an identical topic modeling setup – 82.5% correct for the NYT corpus, and 84.0% correct for the Wikipedia corpus. Given users perform better on our visual word intrusion task, this suggests that humans find our visual topics to be as interpretable and coherent as linguistic topics. Additional information regarding study setup is given in Appendix A.4.

4.2 Coherence, Relatedness, & Diversity

Table 1 contains the results for these well-established metrics to evaluate our visual topics. We compare against two baselines: a text baseline, where we use ViT-GPT2 [39] to caption each image and run LDA on the generated captions, and a random baseline, where we randomly shuffle or sample words in each topic. See Appendix A.5 for equations to compute each metric, as well as additional discussion about the results.

5 Conclusion

In this paper, we proposed visual topic modeling to uncover co-occurrence based relationships in images. We empirically show that topics capture structures that differ from what classic unsupervised learning methods capture for images. We support these results with a theoretical example, which demonstrates that similarity-based clustering is fundamentally unable to model relationships. Our approach generates a visual vocabulary, which can then interface images with topic modeling algorithms from natural language processing. Topic modeling has long been used to understand text documents in fields such as the social sciences, medicine, and psychology. We hope this paper paves the way for researchers to analogously explain image datasets in these domains.

References

[1] Aly Abdelrazek, Yomna Eid, Eman Gawish, Walaa Medhat, and Ahmed Hassan. Topic modeling algorithms and applications: A survey. *Information Systems*, pp. 102131, 2022.

[2] Radhakrishna Achanta, Appu Shaji, Kevin Smith, Aurelien Lucchi, Pascal Fua, and Sabine Süsstrunk. Slic superpixels. Technical report, 2010.

[3] Nikolas Adaloglou, Felix Michels, Hamza Kalisch, and Markus Kollmann. Exploring the limits of deep image clustering using pretrained models, 2023.

[4] Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes, 2018.

[5] Xiao Bai, Xiang Wang, Xianglong Liu, Qiang Liu, Jingkuan Song, Nicu Sebe, and Been Kim. Explainable deep learning for efficient and robust pattern recognition: A survey of recent developments. *Pattern Recognition*, 120:108102, 2021.

[6] Thomas Berg, Jiongxin Liu, Seung Woo Lee, Michelle L. Alexander, David W. Jacobs, and Peter N. Belhumeur. Birdsnap: Large-scale fine-grained visual categorization of birds. In *Proc. Conf. Computer Vision and Pattern Recognition (CVPR)*, June 2014.

[7] Michael W Berry. Survey of text mining: Clustering, classification, and retrieval springer-verlag new york. *Inc.-2004.-244 p*, 2004.

[8] Federico Bianchi, Silvia Terragni, and Dirk Hovy. Pre-training is a hot topic: Contextualized document embeddings improve topic coherence. *arXiv preprint arXiv:2004.03974*, 2020.

[9] Federico Bianchi, Silvia Terragni, Dirk Hovy, Debora Nozza, and Elisabetta Fersini. Cross-lingual contextualized topic models with zero-shot learning, 2021.

[10] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

[11] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101 – mining discriminative components with random forests. In *European Conference on Computer Vision*, 2014.

[12] Ricardo Campello, Davoud Moulavi, and Joerg Sander. Density-based clustering based on hierarchical density estimates. volume 7819, pp. 160–172, 04 2013. ISBN 978-3-642-37455-5. doi: 10.1007/978-3-642-37456-2_14.

[13] Jonathan Chang, Sean Gerrish, Chong Wang, Jordan Boyd-Graber, and David Blei. Reading tea leaves: How humans interpret topic models. *Advances in neural information processing systems*, 22, 2009.

[14] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3606–3613, 2014. doi: 10.1109/CVPR.2014.461.

[15] Scott Deerwester, Susan T Dumais, George W Furnas, Thomas K Landauer, and Richard Harshman. Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6):391–407, 1990.

[16] Adji B Dieng, Francisco JR Ruiz, and David M Blei. Topic modeling in embedding spaces. *Transactions of the Association for Computational Linguistics*, 8:439–453, 2020.

[17] Ran Ding, Ramesh Nallapati, and Bing Xiang. Coherence-aware neural topic modeling, 2018.

[18] Caitlin Doogan, Wray Buntine, Henry Linger, and Samantha Brunt. Public perceptions and attitudes toward covid-19 nonpharmaceutical interventions across six countries: a topic modeling analysis of twitter data. *Journal of medical Internet research*, 22(9):e21419, 2020.

[19] Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017.

[20] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

[21] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, KDD'96, pp. 226–231. AAAI Press, 1996.

[22] Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. *2004 Conference on Computer Vision and Pattern Recognition Workshop*, pp. 178–178, 2004. URL https://api.semanticscholar.org/CorpusID:2156851.

[23] Ronald Aylmer Fisher. Design of experiments. *British Medical Journal*, 1(3923):554, 1936.

[24] Amirata Ghorbani, James Wexler, James Y Zou, and Been Kim. Towards automatic concept-based explanations. In *Advances in Neural Information Processing Systems*, pp. 9273–9282, 2019.

[25] Gregory Griffin, Alex Holub, and Pietro Perona. Caltech 256, Apr 2022.

[26] Maarten Grootendorst. Bertopic: Neural topic modeling with a class-based tf-idf procedure, 2022.

[27] Shreya Havaldar, Adam Stein, Eric Wong, and Lyle Ungar. Topex: Topic-based explanations for model comparison. *arXiv preprint arXiv:2306.00976*, 2023.

[28] Harold Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24:498–520, 1933. URL https://api.semanticscholar.org/CorpusID:144828484.

[29] Hamed Jelodar, Yongli Wang, Chi Yuan, Xia Feng, Xiahui Jiang, Yanchao Li, and Liang Zhao. Latent dirichlet allocation (lda) and topic modeling: models, applications, a survey. *Multimedia Tools and Applications*, 78:15169–15211, 2019.

[30] Moon-Gu Jeon. *Centroid-Based Dimension Reduction Methods for Classification of High Dimensional Text Data*. PhD thesis, USA, 2001. AAI3010558.

[31] Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*, 2016.

[32] Daniel Kifer, Shai Ben-David, and Johannes Gehrke. Detecting change in data streams. pp. 180–191, 04 2004.

[33] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything, 2023.

[34] Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 5338–5348. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/koh20a.html.

[35] S. Kornblith, J. Shlens, and Q. V. Le. Do better imagenet models transfer better? In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2656–2666, Los Alamitos, CA, USA, jun 2019. IEEE Computer Society. doi: 10.1109/CVPR.2019.00277. URL https://doi.ieeecomputersociety.org/10.1109/CVPR.2019.00277.

[36] Jonathan Krause, Jia Deng, Michael Stark, and Li Fei-Fei. Collecting a large-scale dataset of fine-grained cars. 2013. URL https://api.semanticscholar.org/CorpusID:16632981.

[37] Alex Krizhevsky. Learning multiple layers of features from tiny images. 2009. URL https://api.semanticscholar.org/CorpusID:18268744.

[38] J.B. Kruskal. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, 29(1):1–27, 1964.

[39] Ankur Kumar. The illustrated image captioning using transformers. *ankur3107.github.io*, 2022. URL https://ankur3107.github.io/blogs/the-illustrated-image-captioning-using-transformers/.

[40] Jey Han Lau, David Newman, and Timothy Baldwin. Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 530–539, 2014.

[41] Daniel Lee and H Sebastian Seung. Algorithms for non-negative matrix factorization. *Advances in neural information processing systems*, 13, 2000.

[42] Lin Liu, Lin Tang, Wen Dong, Shaowen Yao, and Wei Zhou. An overview of topic modeling and its current applications in bioinformatics. *SpringerPlus*, 5(1):1–22, 2016.

[43] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3431–3440, 2015. doi: 10.1109/CVPR.2015.7298965.

[44] J. MacQueen. Some methods for classification and analysis of multivariate observations. 1967. URL https://api.semanticscholar.org/CorpusID:6278891.

[45] S. Maji, J. Kannala, E. Rahtu, M. Blaschko, and A. Vedaldi. Fine-grained visual classification of aircraft. Technical report, 2013.

[46] Leland McInnes, John Healy, Nathaniel Saul, and Lukas Großberger. Umap: Uniform manifold approximation and projection. *Journal of Open Source Software*, 3(29):861, 2018. doi: 10.21105/joss.00861. URL https://doi.org/10.21105/joss.00861.

[47] Daniel Müllner. Modern hierarchical, agglomerative clustering algorithms, 2011.

[48] David Newman, Jey Han Lau, Karl Grieser, and Timothy Baldwin. Automatic evaluation of topic coherence. In *Human language technologies: The 2010 annual conference of the North American chapter of the association for computational linguistics*, pp. 100–108, 2010.

[49] Andrew Ng, Michael Jordan, and Yair Weiss. On spectral clustering: Analysis and an algorithm. In T. Dietterich, S. Becker, and Z. Ghahramani (eds.), *Advances in Neural Information Processing Systems*, volume 14. MIT Press, 2001. URL https://proceedings.neurips.cc/paper_files/paper/2001/file/801272ee79cfde7fa5960571fee36b9b-Paper.pdf.

[50] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing*, pp. 722–729, 2008. doi: 10.1109/ICVGIP.2008.47.

[51] Haesun Park, Moongu Jeon, and J. Ben Rosen. Lower dimensional representation of text data based on centroids and least squares. *BIT Numerical Mathematics*, 43:427–448, 2003. URL https://api.semanticscholar.org/CorpusID:9629383.

[52] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and C. V. Jawahar. Cats and dogs. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3498–3505, 2012. doi: 10.1109/CVPR.2012.6248092.

[53] Stephan Rabanser, Stephan Günnemann, and Zachary Chase Lipton. Failing loudly: An empirical study of methods for detecting dataset shift. In *Neural Information Processing Systems*, 2018. URL https://api.semanticscholar.org/CorpusID:53096511.

[54] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

[55] Daniel Ramage, Evan Rosen, Jason Chuang, Christopher D Manning, and Daniel A McFarland. Topic modeling for the social sciences. In *NIPS 2009 workshop on applications for topic models: text and beyond*, volume 5, pp. 1–4, 2009.

[56] Hadi Salman, Andrew Ilyas, Logan Engstrom, Ashish Kapoor, and Aleksander Madry. Do adversarially robust imagenet models transfer better? In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 3533–3545. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/24357dd085d2c4b1a88a7e0692e60294-Paper.pdf.

[57] Hadi Salman, Andrew Ilyas, Logan Engstrom, Ashish Kapoor, and Aleksander Madry. Do adversarially robust imagenet models transfer better? In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 9781713829546.

[58] Prathyush Sambaturu, Aparna Gupta, Ian Davidson, SS Ravi, Anil Vullikanti, and Andrew Warren. Efficient algorithms for generating provably near-optimal cluster descriptors for explainability. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 1636–1643, 2020.

[59] Adam Stein, Yinjun Wu, Eric Wong, and Mayur Naik. Rectifying group irregularities in explanations for distribution shift, 2023.

[60] Joshua B. Tenenbaum, Vin de Silva, and John C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000. doi: 10.1126/science.290.5500.2319. URL https://www.science.org/doi/abs/10.1126/science.290.5500.2319.

[61] Shengbang Tong, Erik Jones, and Jacob Steinhardt. Mass-producing failures of multimodal systems with language models, 2023.

[62] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008. URL http://jmlr.org/papers/v9/vandermaaten08a.html.

[63] Ike Vayansky and Sathish AP Kumar. A review of topic modeling methods. *Information Systems*, 94:101582, 2020.

[64] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. pp. 3156–3164, 06 2015. doi: 10.1109/CVPR.2015.7298935.

[65] Jianxiong Xiao, James Hays, Krista A. Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 3485–3492, 2010. doi: 10.1109/CVPR.2010.5539970.

[66] Jianlong Zhou, Amir H Gandomi, Fang Chen, and Andreas Holzinger. Evaluating the quality of machine learning explanations: A survey on methods and metrics. *Electronics*, 10(5):593, 2021.

A Appendix

A.1 Experimental Setup

We discuss all details in our visual topic modeling experiments:

Visual vocabulary construction We use SLIC [2] to obtain the segments for each image. We then follow Ghorbani et al. [24] and crop + resize each segment to get the segment embeddings. We use a vision transformer [20] to embed these cropped and resized segments. To construct the vocabulary, we use k-means clustering and group similar segments together to create visual words. The total number of clusters for each dataset is $25 \times \text{Num Classes}$, following [56].

In our preliminary experiments with SLIC, we experiment with both zeroing (i.e. blacking out) the rest of the image, cropping the segments, and resizing to the max image size. We find that following Ghorbani et al. [24] generates the best visual topics upon manual inspection, so we follow this proceduce for our remaining experiments. We also experiment with the Segment Anything Model [33], and find that it performs worse than SLIC, upon manual inspection of the resulting topics.

Topic modeling We run LDA implemented using collapsed Gibbs sampling, due to its efficiency and quicker runtime. We set $N_{Topics} = |\text{Num Classes}|/2$ for all datasets. For our LDA prior parameters, we set $\alpha = 0.05$ and $\beta = 0.005$. The standard LDA priors for linguistic topics are $\alpha = 0.1$ and $\beta = 0.01$. Given our visual vocabulary does not follow a Zipfian distribution, like the English vocabulary does, we half the standard priors, as we expect each document to have fewer topics and each topic to have fewer words. For our text baselines, we use the ViT-GPT2 model [] from HuggingFace to generate the captions for each image. We run LDA for 2,000 iterations to ensure enough time for convergence.

A.2 Proofs

Property 1. *Two subsets of features S, S' are* related *if their PMI, $\sum_{z_i \in S} \sum_{z_j \in S'} \frac{p(z_i, z_j)}{p(z_i)p(z_j)}$, is high, where $p(z_i, z_j)$ is the probability of observing the feature subsets z_i and z_j within the same example x and $p(z_i)$ is the probability of observing z_i in the entire dataset.*

Example 1. (Mixture of Related Gaussians) *Let $x_i \in \mathbb{R}^{2d}, z_{ij} \in \mathbb{R}^2$ be observations from the following generative process:*

- *Let R_i be a Rademacher variable representing the topic, which is drawn from $\{-1, +1\}$ with $\frac{1}{2}$ probability each.*
- *Let $z_{ij} \sim \sum_{S \in \{-1,1\}} p(z|R_i S, S)p(S)$ where $p(z|R_i S, S) = \text{Gaussian}\left([R_i, S], \sigma^2 I\right)$ and $p(S)$ is a Rademacher variable for $j = 1 \dots d$,*
- *Then, $x_i = (z_{i1}, \dots, z_{id})$ is the concatenation of all the z_{ij} for $j = 1 \dots d$.*

Theorem 1. *Let z_{ij} be generated according to Example 1 with $\sigma \leq 0.288$. With probability at least 0.99 there does not exist a clustering of the features z_{ij} that has two clusters containing only related points from each pair of opposing quadrants.*

Proof. For sake of contradiction, suppose there exists a cluster centroid μ_1 that contains two points z_1, z_2 generated from $p(z|1, 1)$ and $p(z| - 1, -1)$, and a second centroid μ_2 that contains two points z_3, z_4 generated from $p(z|1, -1)$ and $p(z| - 1, 1)$.

Let D be the Mahalanobis distance with probability $1 - \alpha$ of being within distance D of the mean. Since $\sigma \leq 0.288$, the probability that z_i is at least $D \geq 1$ away from its corresponding centroid is at at most $F\left(\frac{1}{\sigma^2}\right) = F(11.983) = 0.0025 = \alpha$, since $D^2 = \sum_j (z_{ij} - \mu_j)^2/\sigma^2$ follows a χ^2_2 distribution. Then, With probability $p = (1 - \alpha)^4 = 0.99$ (via union bound), all four points z_i are within radius 1 from their respective centroids. This implies that each z_i must lie within one of the four quadrants.

However, note that cluster centroids partition the space into convex partitions. Thus, if z_1 and z_2 are in the same cluster, then the line connecting them must also be in the same cluster. However, this must also be true of z_3 and z_4—the line connecting these points must be within the same cluster.

Since each of these points exist in opposite quadrants, there must be a point where these two lines intersect that is in the interior of both cluster partitions. However, this is a contradiction, as this point cannot simultaneously be in the interior of two partitions.

□

Corollary 1. *A visual topic model with $T = 2$ topics is sufficient to divide the data from Example 1 into two subsets with high relatedness.*

Proof. Let $v : \mathbb{R}^2 \to \{\pm 1\}^2$ be defined as $v(z_i) = \arg\min_{\mu \in \{\pm 1\}^2} \|z_{ij} - \mu\|_2^2$, which is the hard assignment of each feature subset z_{ij} to the quadrant that z_{ij} resides in. Let $g : \mathbb{R}^{d \times 2} \to \{\pm 1\}^{d \times 2}$ be defined as $g(x_i) = \begin{bmatrix} v(z_{i1}) \\ \vdots \\ v(z_{id}) \end{bmatrix}$ which maps each example x_i to a sequence of words, where each word comes from the vocabulary $\{\pm 1\}^2$. Then, the topic model defined by the topic-term distribution $p(1, 1|T = 1) = p(-1, -1|T = 1) = 0.5$ and $p(-1, 1|T = 2) = p(1, -1|T = 2) = 0.5$ perfectly clusters the examples into two topics with high relatedness. □

A.3 Datasets

We follow Salman et al. [57] and use the following 12 datasets for our experiments.

Dataset	Num Images	Num Classes	Content
Birdsnap [6]	40,754	500	North American birds
Caltech-101 [22]	8,677	101	Objects
Caltech-256 [25]	30,607	256	Objects
CIFAR-10 [37]	60,000	10	Objects
CIFAR-100 [37]	60,000	100	Objects
Describable Textures (DTD) [14]	5,640	47	Textured images
FGVC Aircraft [45]	10,200	102	Aircrafts
Food-101 [11]	101,000	101	Food dishes
Oxford 102 Flowers [50]	8,189	102	Flowers
Oxford-IIIT Pets [52]	7,349	37	Cats and dogs
SUN397 [65]	39,700	397	Various scenes
Stanford Cars [36]	16,185	397	Types of cars

Table 2: Summary of datasets used to generate visual topics.

A.4 Human Evaluation

We sourced users to participate in our human evaluation by asking graduate students to volunteer in our study. Our users came from varying academic backgrounds. Of the 12 users, 5 had prior familiarity with topic modeling. As identifying topics requires users to clearly distinguish things like colors and shapes, we additionally required users that users have no visual impairments (e.g. red-green color blindness).

These were the instructions our users received prior to participating in the study:

```
In this study, you will be asked to evaluate the output of a topic modeling study.
Topic modeling is a technique used to find related themes in large datasets.  For
example, if you were to run a topic model on Wikipedia articles, an example topic
could be words related to politics, and may contain words like "government", "war",
"Iraq", "Obama", etc.

We have applied a topic modeling algorithm to images, so you will be asked to
evaluate a set of 25 visual topics.  You will see a collection of 5 images.  4 of
these images have segments highlighted that belong to one visual topic.  1 image has
segments highlighted for a different topic, or is an "intruder" image.  Your task
is to look at the set of 5 images and then highlighted segments and identify the
intruder.
```

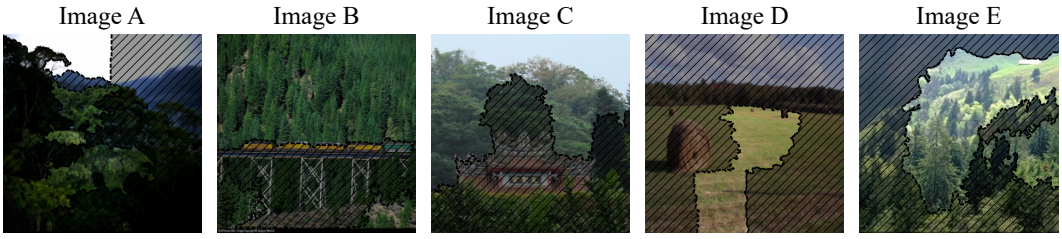


Figure 4: Example of our visual word intrusion task. Users are asked to identify which image they believe is the "imposter", or which image's colored segments belong to a different topic than those of the other four images. In this example, the colored segments in images A, B, C, and E are part of topic that mainly contains clouds, lush forests, and blue skies. Image D, the intruder image, contains segments that belong to a different topic, one that includes grassy fields.

We then showed users an example from the dataset that they were not responsible for annotating, and asked if they understood the task. All users indicated they understood the task after receiving the instructions.

A.5 Description of Metrics

A.5.1 Internal Coherence

Internal coherence, introduced by Newman et al. [48], measures coherence via pairwise similarity of the top N words $w_1, \dots w_n$ with highest $P(w|T)$ in each topic. We follow Bianchi et al. [9] and use centroid similarity to calculate pairwise coherence scores for the top N words in each topic. We define the internal coherence of a topic as follows, where $\hat{v}_i$ corresponds to the centroid of the cluster containing word i. Following Bianchi et al. [9], Grootendorst [26], we use $N = 10$.

$$IC_{Topic} = \frac{2}{N \times (N-1)} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} cos(\hat{v}_i, \hat{v}_j) \tag{1}$$

We compare against two baselines:

- *Text baseline*: we use ViT-GPT2 [39] to caption each image and run LDA on the generated captions. We use word embeddings [31] for cosine similarity.
- *Shuffled baseline*: we conduct a permutation test [23], where we take the top N words from each visual topic and shuffle them across topics. We calculate coherence for the new shuffled topics and report the mean of 100 runs.

Table 1 shows the internal coherence of our visual topics and baselines. For all datasets, our visual topics outperform both baselines. These high coherence scores empirically support the findings of our user study — our topics are highly coherent by both human evaluation and automatic metrics.

A.5.2 Topic Relatedness

To highlight how visual topics are distinct from segment clusters, we additionally measure relatedness of topics [17]. The intuition behind this test is that visual words that are not necessarily similar, but co-occur in visual documents (i.e. dogs and grass, boats and lakes, etc.) should still be in the same topic. Relatedness is measured via pointwise mutual information, as discussed in Property 1. We use normalized pointwise mutual information (NPMI) between each pair of words in the top N words. We follow Lau et al. [40], Grootendorst [26] and use $N = 10$.

$$R_{Topic} = \frac{2}{N \times (N-1)} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \frac{\log \frac{P(w_i, w_j)}{P(w_i)P(w_j)}}{-\log P(w_i, w_j)} \tag{2}$$

We compare against two baselines:

- *Text baseline*: we use ViT-GPT2 [39] to caption each image and run LDA on the generated captions, using Equation 2 to measure relatedness.

- *Shuffled baseline*: we conduct a permutation test [23], where we take the top N words from each visual topic and shuffle them across topics. We calculate the relatedness of these shuffled topics and report the mean of 100 runs.

Table 1 shows the relatedness of our visual topics and baselines. For 9 of the 12 datasets, visual topics are more related than the text topics. For all datasets, our visual topics are more related than the shuffled topics. As discussed in Section 3, the high relatedness of our visual topics indicates we have successfully captured relationships of co-occurring structures in images.

A.5.3 Topic Diversity

Dieng et al. [16] introduce topic diversity as a way to evaluate topic quality. Topic diversity looks at the top N words $w_1, \dots w_n$ with highest $P(w|T)$ in each topic, and computes the percentage of unique words. A word is considered unique if it is not present in the top N words of any other topics. The intuition behind this metric (ranging from zero to one) is that topic models should generate diverse topics — a lower diversity score suggests redundant topics, indicating the topics cannot sufficiently disentangle the corpus's themes. A higher score indicates more varied topics. We follow Dieng et al. [16] and use $N = 25$.

We compare against two baselines:

- *Text baseline*: we use ViT-GPT2 [39] to caption each image and run LDA on the generated captions. We then calculate diversity.
- *Random baseline*: As shuffling the top N words in each topic will not impact the diversity score, we randomly sample N words from our dataset to create random topics. We calculate diversity of the randomized topics and report the mean of 100 runs.

Table 1 shows the diversity of our visual topics. We outperform both baselines across all datasets. The low diversity of our text baseline indicates that captioning via an image-to-text model loses valuable information needed to create unique and useful topics.

B Related Work

Dimensionality Reduction. We first discuss relevant past work in dimensionality reduction methods. Linear dimensionality reduction methods like PCA [28] and non-linear dimensionality reduction methods like t-SNE [62], MDS [38], Isomap [60], UMAP [46], etc. have long been used to find patterns in image data. Cluster memberships of examples can also be used as a form of dimensionality reduction [30, 51, 7]. More recent work in this space includes extracting concepts specific to a trained model or class, such as Automatic Concept Extraction [24] or Concept Bottlenecks [34].

Unsupervised Interpretability. Another line of research attempts to explain image datasets via unsupervised methods. Segmentation [43] divides images into structured sub-parts. Image-to-text models [64] explain images via captioning, grounding an image in language. Linear probes [4, 35] are simple linear classifiers trained on top of existing learned representations and are used to measure the quality of these learned features. Shift explanation techniques, such as those from Kifer et al. [32], Rabanser et al. [53], Stein et al. [59] work towards explaining shifts in data distribution using unsupervised methods. Clustering methods like KMeans [44], DBSCAN [21], HDBSCAN [12], Agglomerative Hierarchical Clustering [47], Spectral Clustering [49], etc. also aim at organizing data using intrinsic structure, which provides unsupervised interpretability.

Topic Modeling. Topic modeling is a long-standing field in NLP, and includes fully unsupervised methods such as Latent Dirichlet Allocation [10], Latent Semantic Analysis [15], and Non-Negative Matrix Factorization [41]. Neural-based topics models, such as BERTopic [26] and Contexualized Topic Models [8] incorporate pre-trained embeddings in the topic generation process. Topics have been used to compare models [27] and used for downstream tasks such as summarization, information retrieval, etc. in variety of domains [29].