Explaining Deep Convolutional Neural Networks via Latent Visual-Semantic Filter Attention

Yu Yang, Seungbae Kim, and Jungseock Joo
University of California, Los Angeles
yuyang@cs.ucla.edu, sbkim@cs.ucla.edu, jjoo@comm.ucla.edu

Abstract

Interpretability is an important property for visual models as it helps researchers and users understand the internal mechanism of a complex model. However, generating semantic explanations about the learned representation is challenging without direct supervision to produce such explanations. We propose a general framework, Latent Visual Semantic Explainer (LaViSE), to teach any existing convolutional neural network to generate text descriptions about its own latent representations at the filter level. Our method constructs a mapping between the visual and semantic spaces using generic image datasets, using images and category names. It then transfers the mapping to the target domain which does not have semantic labels. The proposed framework employs a modular structure and enables to analyze any trained network whether or not its original training data is available. We show that our method can generate novel descriptions for learned filters beyond the set of categories defined in the training dataset and perform an extensive evaluation on multiple datasets. We also demonstrate a novel application of our method for unsupervised dataset bias analysis which allows us to automatically discover hidden biases in datasets or compare different subsets without using additional labels. The dataset and code are made public to facilitate further research.[1]

1. Introduction

Convolutional neural networks have shown great performance in visual representation learning, but the learned representations are usually hard to explain or interpret. The lack of explainability raises the concern that AI systems and models, although very accurate in prediction, may have hidden negative effects on human users and society, such as AI bias. Several studies reported biases in computer vision models for face attribute classification [8, 12], recognition [33, 58], and image captioning [25]. It is very challenging, however, to identify these biases from a black-box model with distributed knowledge.

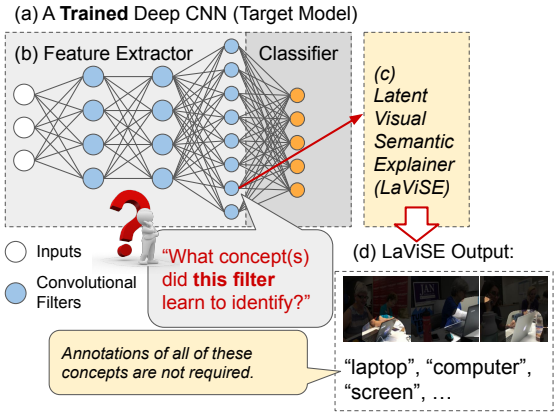


Figure 1. The proposed framework aims to semantically explain the concepts learned by individual filters in a CNN without supervision on the concepts used for the semantic explanations.

To date, several methods have been proposed to interpret what are learned and captured in CNNs. These methods vary greatly by the form (visualization, captions, synthesized samples), the focus (individual filter vs network level), and the scope (any existing models vs requiring training with supervision) of the generated explanations, and each method has its own strengths and weaknesses.

The main objective of this paper is to generate the textual interpretations of any existing black box model that can overcome the limitations of existing approaches for several reasons. First, it generates words and thus can be more semantically meaningful and objective than visualization based methods [42, 49, 66]. Second, it can apply to any arbitrary network and does not require training or annotations, which is much more applicable than methods that require training a model with ground-truth explanation annotations [27]. We do train an adapter using general image classification datasets which can apply to any given target model. Third, it can generate explanations using novel concepts that are not given in the training set. These properties are critical in understanding black-box models for which we do not have access to the original training data or any

[1] https://github.com/YuYang0901/LaViSE

information about the training process. This is a realistic assumption in practice where one needs to interpret and scrutinize a given model.

To this end, we introduce the *Latent Visual Semantic Explainer* (LaViSE) as a novel framework to teach any existing CNN to generate text descriptions about its own latent representations at the filter level. Our framework differs from supervised approaches in that we do not require to annotate the explanation itself along with an input image and a category label. Instead, our method constructs a mapping between the visual and the semantic space using generic image datasets (using images and category names), then transfers the mapping to the target domain without semantic labels. We do not attempt to train more "interpretable" models [7, 11, 38, 40, 67] but interpret any given network without changing its structure or retraining. Our work is also closely related to the literature of visual attribute or concept based learning [6, 17, 45, 47, 50], but we do not require any additional supervision for attribute labeling, which makes our approach more generally applicable. It is also important to note that our method does not merely explain each individual filter separately but uses aggregated responses using a novel filter attention method. Experimental results show our method can generate novel descriptions for learned filters beyond the set of categories defined in the training dataset and provide more accurate explanations for filters comparing to the existing method.

While our main contribution is a novel method to generate explanations for any CNNs, our approach can be used in a practical application of comparative analysis where the goal is to discover and explain the differences between given multiple models or multiple sets of images. To demonstrate the utility, we compare a model finetuned from a pretrained model and a model trained from scratch, and we also compare the gender disparities in datasets. Besides public image datasets, we collect and analyze social media photographs posted by U.S. politicians to exemplify the effectiveness of our method in solving more challenging real-world problems.

2. Related Work

Explanations via Visualization. Saliency methods [22, 26, 43, 49, 51, 66] visually show the amount of contribution for each pixel to the model prediction. They are widely used in the literature but may be unreliable [18] because they can respond to low level features such as image edges rather than semantically more important features [1]. They also require users to speculate meanings of the visualizations as they do not provide explicit semantic explanations. Some approaches have been proposed to perform case-based reasoning [9, 37] by providing patches from training images as explanations, i.e. by analogy. These methods cannot be applied to arbitrary networks across domains.

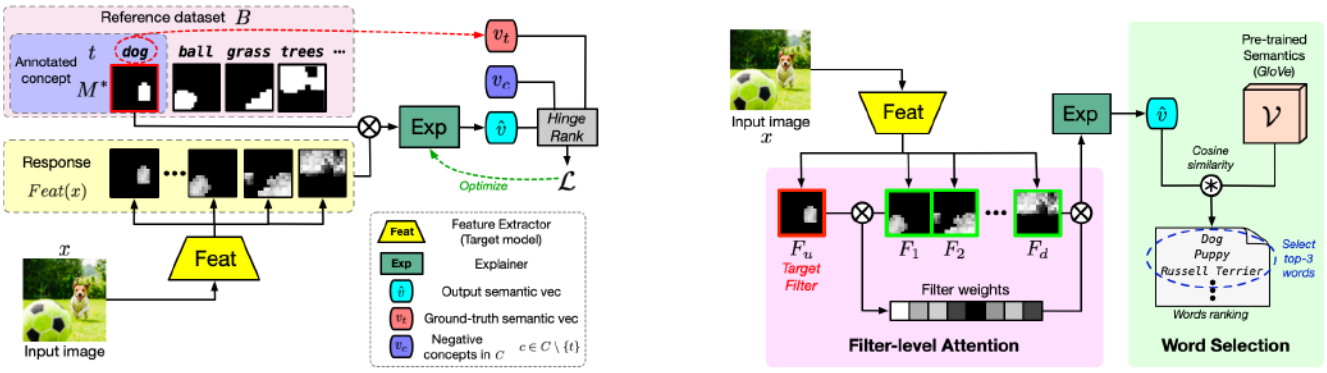
Explanations by Text. [24] propose to generate text to explain and justify the output of an image classifier. Similarly, [27] take a hybrid approach and generate multimodal explanations by using both visual highlights and textual descriptions. [62] propose a VQA system that can not only provide multimodal explanations but also link terms in the textual explanation and segmented items in the image. [65] use multimodal cues to interpret hidden messages (why and what) in visual advertisements. These methods can generate a very meaningful and interpretable explanation to human users, however the explanation itself should be labeled for each example, and the model will learn to generate it in the same way it computes its outputs. Also, since these explanations are annotated by humans before training, they do not necessarily explain what the model has learned (different networks will yield the same explanations).

Visual Semantic Explanations of Visual Representation. Our paper is the most closely related to [4, 48, 70]. They explain an individual internal filter of a trained neural network by measuring the alignments of images' activations on that filter to each predefined concept's segmentation masks. Furthermore, [4] provides a dataset with a broad range of concepts annotated for the alignments. Our framework also uses an annotated dataset so the model can learn to construct the mapping between the visual representation and semantic embedding space, but it differs in that it can discover unseen novel concepts in another domain instead of being restricted to the annotated ones in training data.

Generalization to Unseen Subjects. Zero-shot learning (ZSL) aims to recognize instances of categories that have not been seen in the training stage. Handcrafted attributes and semantic representations learned from textual data are often used to connect the seen and unseen classes, so knowledge learned from training classes can be transferred to unseen categories. Most existing deep neural network based ZSL works [3, 16, 31, 32, 36, 41, 52, 68] either use the semantic space or an intermediate space as the embedding space. Our work relates to this topic as it also helps to discover unannotated concepts. We do not compare our framework with results in this area because our framework is not designed for ZSL but can be built on any ZSL methods.

3. Latent Visual Semantic Explainer (LaViSE)

We now explain our main framework to explain the deep visual representations of a given *target model* (a CNN), f. We assume that this model has already been fully trained from an unknown target dataset, D, and we do not have access to D. In order to learn the visual-semantic explainer on f, we instead use another dataset, the *reference dataset* $B = \{(x_i, y_i)\}_{i=1}^{n}$, where $x_i \in \mathbb{R}^{3\times h\times w}$ is an input image and y_i is a set of concept labels associated with the image and corresponding masks, *i.e.* $y_i = \{(t_j, M_j)\}_{j=1}^{m}$. $t_j \in C$ is an annotated concept, and C is the set of all concept labels



(a) Training the explainer (Section 3.1)

(b) Inference with the filter attention (Section 3.2)

Figure 2. An overview of LaViSE framework. (a) At the training phase, we train the explainer by connecting each image's visual representation with semantic concepts. The hinge rank loss helps the explainer learn the semantic embedding vectors, $\hat{v}$, that are close to the ground truth concept vectors v_t, while being far away from all the other concept vectors v_c in the semantic space. (b) During inference, LaViSE obtains the representation for each latent filter of the target layer via filter-level attention, and then the trained explainer takes this representation to explain these filters semantically by selecting words with the highest similarities from $\mathcal{V}$.

in B. $M_j \in \{0,1\}^{h \times w}$ is a pixel-level mask for its corresponding concept, t_j. The reference dataset may already provide these masks (e.g. semantic segmentation). For object detection datasets, the region inside each bounding box will be filled in with 1. For image classification datasets (no bounding boxes), $M_j = 1^{h \times w}$. For each semantic concept category t_j, we obtain its semantic representation, v_{t_j} by using a pre-trained word embeddings (e.g. GloVe [46]).

In order to explain filters in an arbitrary target model, LaViSE first trains a feature explainer (Section 3.1) using a reference dataset, B. Once trained, this explainer is further used to explain filters that may not have any matching concepts in B using our novel filter attention mechanism (Section 3.2).

3.1. Training Feature Explainer

The purpose of our feature explainer is to transform the visual feature representations of images from a target model to equivalent representations in a semantic space, which can be translated to words. To this end, we train a feature explainer $\text{Exp}(\cdot)$ with a reference dataset B as shown in Figure 2a. Given a target model, f, we take the feature extractor $\text{Feat}(\cdot)$ (*e.g.* by removing the last classification layer). We use the feature extractor to obtain the visual representations of the images in B, $F_i = \text{Feat}(x_i) \in \mathbb{R}^{d \times h' \times w'}$, *i.e.* the output of the last convolutional layer with d filters. This is the input to the feature explainer, but we first mask this feature as follows. Each image has k pairs of a ground-truth concept and its mask, (t_j, M_j)[2]. For each concept, we obtain a masked visual feature by element-wise product, $F_i \odot M_j$, because this masked feature corresponds to the specific concept, t_j. We then use the feature explainer to obtain the semantic representation of the masked visual representation: $\hat{v} = \text{Exp}(F_i \odot M_j)$.

To train the explainer, we use the semantic representations of the ground-truth concept, v_{t_j}, and negative concepts, $\{v_c : \forall c \in C, c \neq t_j\}$, using a pre-trained word embedding model such that $\hat{v}$ should be closer to v_{t_j} than to v_c. Note that we always normalize these semantic vectors. We modify the objective function from [16] as follows:

$$\min_{\theta} \frac{1}{k^*} \sum_{j} \sum_{c \neq t_j} \max(0, 1 - v_{t_j}^\intercal \hat{v} + v_c^\intercal \hat{v}). \quad (1)$$

Our objective function combines dot-product similarity and hinge rank loss as this combination has shown better performance than other losses [64] in zero-shot learning. As LaViSE is a general framework developed for the filter-level interpretability, it can incorporate any other user preferred loss functions or additional loss terms that can serve the purpose of training a zero-shot mapping from visual representations to semantic embeddings. The procedure of training the explainer is shown in Algorithm 1.

3.2. Inference with Filter Attention

The goal of training and using the feature explainer is to explain filters that describe concepts not specified in the reference dataset. For example, the dataset may contain "football" as a concept but lack other related concepts such as "stadium" or "referee" which are still likely present in the images. Since the feature explainer learns to map any visual features to a semantic space, it can generalize to discover **novel** concepts.

A naive way to use the explainer for each filter is to only use the activation of the filter and suppress other filters' responses. We found that this naive approach leads to very poor performance, and this may be due to the fact that many

[2] M_j is resized to the size of feature response map ($h' \times w'$).

Algorithm 1 LaViSE Training

Input: Target feature extractor Feat$(\cdot)$, reference dataset B and the set of all annotated concepts C, pretrained semantic embedding $\mathcal{V}$.
Output: Trained explainer Exp$(\cdot)$ with parameters θ_{exp}
Initialize explainer θ_{exp}
for each image x and its annotations $\{(M, t)\}$ in B **do**
  Get features $F \leftarrow \text{Feat}(x)$
  for each mask-concept pair (M_j, t_j) of x **do**
    Get explainer output $\hat{v} \leftarrow \text{Exp}(F \odot M_j)$
    Get ground-truth concept embedding $v_{t_j} \in \mathcal{V}$
    for each concept $c \in C \setminus \{t_j\}$ **do**
      Get concept embedding $v_c \in \mathcal{V}$
      $l \leftarrow l + \max(0, 1 - v_{t_j}^\intercal \hat{v} + v_c^\intercal \hat{v})$
    end for
  end for
  Update θ_{exp} by Adam to minimize l (eq 1).
end for

filters collectively capture visual features in a distributed manner. This suggest that we can still use the entire feature responses, with proper modification, even when explaining one filter. Some previous studies in interpreting filters have tried to use masking on filter activations [67] and use all the filters together. We propose a novel attention-based masking mechanism, which is simple but effective in extracting and reweighting features relevant to each target filter. Our method is similar to recent self attention models, *e.g.* Transformer [14,57], but simpler because it doesn't use repetitive blocks or multi-head attention.

Filter-level Attention. To discover important concepts which are implicitly represented and distributed over many filters, we propose the filter attention module. Instead of using each filter's activation separately, our method finds a representation for each filter using activations of all filters collaboratively via an attention-based approach. Essentially, we take advantage of the spatial alignments between filters describing a concept and collect their focused activations. In our filter attention method, the feature response map of a target filter (u, the filter we want to explain) serves as a spatial attention; the other filters are reweighted based on their similarities to the target filter. Specifically, suppose we have an input image, x, and d filters in the feature extractor, Feat$(\cdot)$, and the computed feature response map is $F = [F_1, F_2, ..., F_d] = \text{Feat}(x) \in \mathbb{R}^{d \times h' \times w'}$. The input to the explainer, $Exp(\cdot)$, with respect to a target filter u, is computed as follows: $F_k^{\text{att}} = a(F_u, F_k) \cdot F_k, \forall k$, where $a(\cdot)$ computes spatial correlations between filters by cosine similarity. Figure 3 illustrates the difference between our method and other baseline methods.

Word Selection Method. To obtain a list of words to explain given images, LaViSE next computes the cosine similarities between each semantic filter embedding vector and all words in $\mathcal{V}$, and collect s words with the highest similarities. For each filter, we gather $s \times p$ words collected from the top p most activated images and rank the words based on their frequencies. Finally, our framework composes explanations for filters using top-x ranked words. The users can decide the number of words used for each filter explanation depending on how much detail they desire. Note that we empirically tested choices of parameters s and p, and found that our framework works well when p is an integer between 5 and 25 while the choice of s may depend on the size of $\mathcal{V}$ and does not have a substantial impact on the results.

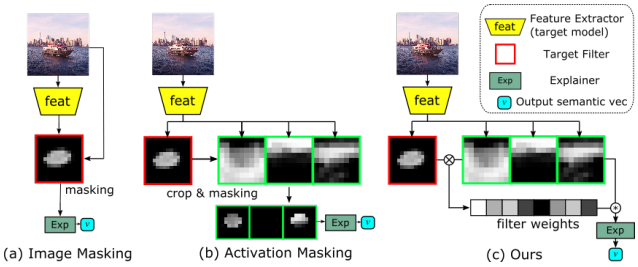


Figure 3. An illustration of comparing our filter attention method with other masking methods.

4. Experiments

4.1. Datasets

To evaluate our proposed framework, we use the following three publicly available datasets and one novel dataset that we collected, which are used as a target dataset, a reference dataset, or both: **(i) MS COCO [39].** MS COCO has more than 200K daily scene images that are pixel-wise labelled with 80 common object categories. To help with our analysis, we also include the gender annotations of MS COCO provided by [69]. **(ii) Visual Genome (VG) [34].** We only use images that have box-able annotations for our experiments. During pre-processing, we combined object categories based on their synset names, combined instances of the same category in the same image, and deleted categories that appear in less than 100 training images. In the end, there remains 106,215 out of 107,228 images, 1,208 out of 80,138 object categories, and at most 47 object categories per image. **(iii) Broden [4].** Broden combines selected examples from several densely labeled image datasets to provide pixel level ground truth labels to a broad range of visual concepts, including scene, objects, object parts, textures, and materials. NetDissect [4] leverages this dataset to provide explanations so we use it to directly compare our framework with NetDissect without biases imposed by the choice of datasets. **(iv) Social Media Photographs of US Politicians (PoP).** To demonstrate the practice of our framework in real-world applications, we also composed a new dataset of social media posts from accounts of US politicians to be used as a target dataset with-

out any visual category labels. We collected roughly 80k images for politicians who ran for the 2018 election from Facebook and annotated the images by the gender and political party. Note that, we used **GloVe [46]** (i.e., 300d GloVe embeddings) that contains total 400K vocabularies trained on 6B Wikipedia tokens for the pre-trained word embeddings $\mathcal{V}$.

4.2. Setup

For the experiments, we use ResNet [23] as our backbone models and build our framework on the PyTorch [44] implementations of ResNet-18 and ResNet-50. In Table 1 and 2, "*layer4*" and "*layer3*" refer to the module names for the PyTorch models.

We consider two challenging settings that are the closest to the real-world scenarios: (1) the list of concepts that can appear in the target dataset is known (but still no annotation is given); (2) we have no prior knowledge about concepts that can appear in the target dataset. We imitate (1) in a generalized zero-shot learning setting with the VG dataset. We only train with a proportion of annotated classes and consider all annotated classes as all concepts that can appear in the dataset. In our experiments, we randomly selected 70% categories for training the mapping and left 30% categories for the model to discover. The split is manually set to ensure that we left out enough new concepts for the model to find, and meanwhile, the model can have enough supervision. We test scenario (2) with the MS COCO dataset, as it does not have as many annotated concepts as VG.

4.3. Compared Methods

To evaluate the performance of interpreting deep representation, we deploy **NetDissect** [4] as our competing method since it is the only applicable method of the filter-level approach as we discussed in Section 2. Note that explaining methods are not easily comparable as they tend to focus on unique settings.

Moreover, we carefully designed the following three baselines to show that our novel filter attention is indispensable and irreplaceable for separating representations for filters at the inference stage:

(i) Original image: Without any attention or masking, the image goes through the feature extractor and then directly into the feature explainer. This baseline is equivalent to using a zero-shot learning model trained for image classification and then collecting the top predictions of most activated images of each filter as the explanation.

(ii) Image masking: For each filter u of layer l, we collect an activation map $A_l(x_i)_u$ for each $x_i \in D$, compute the distribution of all unit activations $\{a_l(x_i)_u\}_{i=1}^n$ on u and select a threshold T_u such that the probability of an activation being above the threshold is p, namely $P(a_u > T_u) = p$. For top activated images, we scale their activation maps of layer l to the shape of the images and set regions with activations below the threshold to zero after preprocessing. We then input the masked images to the feature extractor and then the feature explainer to get the results.

(iii) Activation masking: We use the same thresholds T_u's for the image masking baseline, but we apply the masks to the activation maps directly. This method was also used in [67] for interpreting CNNs.

To ensure a fair comparison across all, we adopt the probability threshold $p = 0.005$ used in [4] for all thresholds mentioned above and also for all visualizations of activated regions in this paper.

Settings			Results		
Model & Layer	Target dataset	Reference dataset	Method	Precision (Top-1)	Prefer
ResNet-18 Layer 4	Places365	Broden	NetDissect	0.70	26%
			LaViSE	**0.74**	**42%**
ResNet-18 Layer 4	MS COCO	MS COCO	NetDissect	0.42	12%
			LaViSE	**0.70**	**49%**
ResNet-50 Layer 4	MS COCO	MS COCO	NetDissect	0.38	12%
			LaViSE	**0.68**	**48%**
ResNet-50 Layer 4	ImageNet	VG	NetDissect	0.24	18%
			LaViSE	**0.46**	**42%**
ResNet-50 Layer 3	ImageNet	VG	NetDissect	0.20	12%
			LaViSE	**0.26**	**28%**

Table 1. Human evaluation of comparing explanations generated by LaViSE (ours) and NetDissect [4] in different settings.

4.4. Evaluation Protocols

For evaluation, we utilize two metrics [4], perceptual effectiveness (i.e., human evaluation) and objective accuracy.

Human Evaluation. Since the notion of interpretability can be subjective and the standard way of quantifying interpretability is still under exploration, we first evaluate the quality of explanations with human examiners from Amazon Mechanical Turk (MTurk). For each filter, human examiners were shown 15 images from the reference dataset with highlighted patches showing the most highly-activating regions for the filter (e.g. Fig 4). Note that this amounts to 7-15K images used for a whole model per each setting. We have at least 3 examiners to evaluate each filter, and we take the averages of the median scores for all filters as the results, shown in Table 1 and 2. We will elaborate the protocol further in Section 4.5.

Objective Evaluation. We also compute the intersection-over-union (IoU) score of each annotation mask and the activation mask of each filter (i.e. segmentation). If the score is above a certain threshold, then we consider the concept corresponding to the annotation mask is one of the ground-truth concepts for that filter. We set the threshold to 0.04 to be consistent with [4].

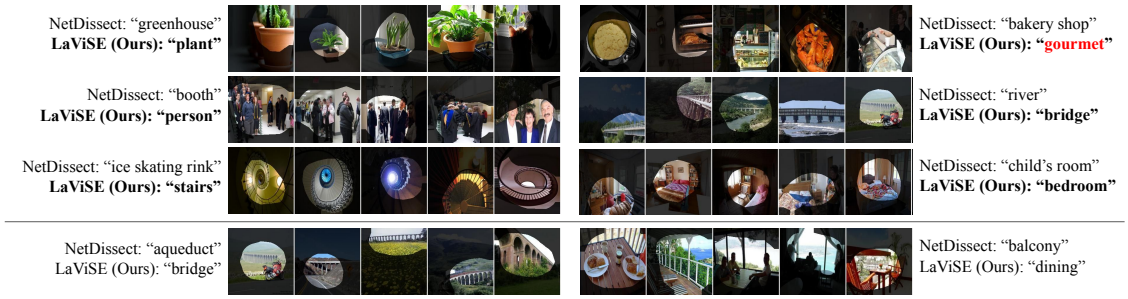


Figure 4. Qualitative comparison between explanations of the same filters given by LaViSE and NetDissect. We used the model (a ResNet-18 trained on Places365) and dataset (Broden) provided by [4]. The first three rows are examples where human raters prefer our explanations over NetDissect's. The last row are examples where NetDissect's explanations were more favored by human raters than ours.

Settings			Results						
Model & Layer	Target dataset	Reference dataset	Method	Precision-*H* (Top-1)	Precision-*H* (Top-5)	Recall-*H* (Top-5)	Recall (Top-5)	Recall (Top-10)	Recall (Top-20)
ResNet-18 Layer 4	MS COCO	MS COCO	Original image	0.66	0.300	0.626	0.599	0.641	0.675
			Image masking	0.61	0.280	0.586	0.567	0.619	0.659
			Activation masking [67]	0.68	0.310	0.658	0.629	0.676	0.721
			Filter attention (Ours)	**0.70**	**0.320**	**0.670**	**0.675**	**0.728**	**0.776**
ResNet-18 Layer 4	ImageNet	Visual Genome	Original image	0.02	0.056	0.024	0.182	0.251	0.334
			Image masking	0.00	0.016	0.006	0.181	0.253	0.337
			Activation masking [67]	0.34	0.316	0.149	0.235	0.309	0.382
			Filter attention (Ours)	**0.44**	**0.340**	**0.159**	**0.273**	**0.353**	**0.429**
ResNet-50 Layer 4	ImageNet	Visual Genome	Original image	0.10	0.096	0.139	0.160	0.229	0.302
			Image masking	0.05	0.036	0.049	0.084	0.134	0.213
			Activation masking [67]	0.37	0.264	0.557	0.199	0.264	0.333
			Filter attention (Ours)	**0.46**	**0.274**	**0.583**	**0.226**	**0.302**	**0.373**
ResNet-50 Layer 3	ImageNet	Visual Genome	Original image	0.00	0.012	0.050	0.070	0.097	0.136
			Image masking	0.00	0.020	0.080	0.022	0.045	0.055
			Activation masking [67]	0.24	0.148	0.470	0.099	0.155	**0.210**
			Filter attention (Ours)	**0.26**	**0.156**	**0.473**	**0.110**	**0.156**	0.207

Table 2. Quantitative evaluation for different masking methods. Columns with "*H*" are the results of human evaluation. Human raters are generally giving higher scores because they also accept synonyms. Please see Section 4.5.2 for the details.

4.5. Results

4.5.1 Comparing with NetDissect

Table 1 shows the quantitative comparison of our framework and NetDissect. We have various settings with different models, layers, and target and reference datasets, including the same setting from [4] which uses Places365 [71] as the target dataset and Broaden as the reference dataset for a fair comparison. The results demonstrate that LaViSE significantly outperforms NetDissect in human-evaluated top-1 precision with margins in all settings. Note that, LaViSE shows almost twice larger precision values than NetDissect in certain settings. In addition to the precision, we ask human examiners to compare semantic explanations of LaViSE and NetDissect side by side along with the most activated images in the reference data and give a comparative rating inspired by [72]. For each comparison, the order of the presented methods to an annotator is randomized. The "Prefer" column of Table 1 records the proportions of human evaluations that prefer one method over another. The results show that human raters preferred LaViSE's explanations more often than NetDissect's with a large margin. Figure 4 provides examples of user preferred explanations. As shown in the example images and explanations, our method provides more informative explanations than NetDissect. It is critical to note that the key advantage of our LaViSE framework is that it can apply to any trained network to generate novel textual explanations about filters. NetDissect requires a known training dataset and hence its interpretations are limited to the annotated categories in the dataset. NetDissect, therefore, focuses on estimating the interpretability of a network architecture given a specific training dataset, whereas LaViSE can *additionally* interpret any network trained from any **arbitrary unknown** data and generate **novel descriptions** beyond pre-defined categories by using a visual-semantic mapping.

4.5.2 Comparing with Masking Schemes

To measure the effect of LaViSE's filter attention module, we conduct an ablation studies with three masking baseline methods described in Section 4.3. Table 2 shows the quantitative results when we ask human examiners to evaluate the explanations. Note that we conduct experiments in two different cases where one that has the same target and reference datasets, and the other with different target and reference datasets. Since the COCO case uses the same target and reference datasets for validation purposes, all methods in the COCO case perform better than the ImageNet case

(a) ImageNet-VG (b) COCO-COCO

Figure 5. Qualitative examples for image masking vs. LaViSE

where transferred features learned with ImageNet to a different reference dataset (VG). In practice, target and reference datasets will differ since we want to explain a model already traiend with an unknown target dataset. Figure 5 showcases the example filters generated by our method and image masking method on the same images in the two cases with different datasets.

We find that our method outperforms the original image baseline and the image masking baseline with large margins in all settings, especially when we aim to explain lower layers. This suggests that our proposed filter attention module can effectively discover important masks of given images. Our method also outperforms the activation masking method [67] except for the top-20 recall value in the setting of ResNet-50 and layer 3. We observe that the performance differences between these two methods are smaller in the layer 3 than the layer 4. That is because our method is more effective in leveraging cross-filter activations in semantic mapping to capture more context at the upper layers.

4.5.3 Explanations at Different Layers

Different layers in a CNN may capture different types of visual concepts and also have different levels of interpretability. For example, some low level features such as texture may not be easily explainable compared to high level visual structures or objects [4]. We observe significant performance drops at lower layers in Table 2. Our human evaluation results also confirm that the lower layers of a CNN can be harder for both our framework and NetDissect to explain (Table 1). To analyze the performance of LaViSE at different layers, we use a ResNet-50 pre-trained with the ImageNet as our backbone model, and Visual Genome as the reference dataset to train the explainer for each layer separately. Results in Table 3 are consistent with the human ratings. The Supplementary Material provides more detailed analysis results due to the space limit.

4.5.4 Explaining with Unsupervised Concepts

Our method can explain novel concepts because the semantic embedding can generalize beyond the category names given in the reference set. A similar idea has also been used in image captioning for novel objects [2]. We use PoP

Layer	Recall (Top-5)	Recall (Top-10)	Recall (Top-20)
Layer 4	**0.226**	**0.302**	**0.373**
Layer 3	0.110	0.156	0.207
Layer 2	0.086	0.131	0.181
Layer 1	0.042	0.060	0.092

Table 3. Comparison for different conv layers of ResNet-50.

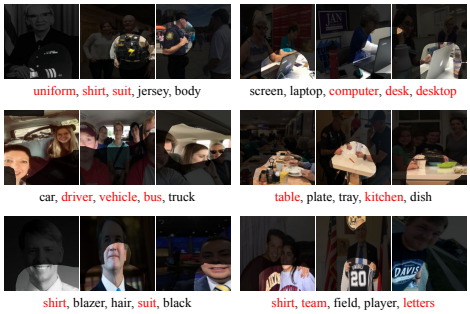


Figure 6. Examples of discovered concepts by LaViSE using the PoP dataset. Concepts in red are the ones outside of the predefined categories (i.e. no annotations) during training.

as the target dataset and Visual Genome as the reference dataset, and select 70% categories from the Visual Genome for training the mapping, and leave 30% categories for the model to discover. Figure 6 shows the examples of concepts that were discovered by LaViSE. As shown in the examples, we observe that LaViSE can explain convolutional filters by finding more accurate concepts that have not been trained. This suggests that our LaViSE can be deployed to any unannotated datasets to gain insights based on the explanations.

Figure 7 shows the proportion of the novel (i.e., unseen) concepts discovered by LaViSE according to the different percentages of annotated sementics in training. Note that we consider a concept as a novel concept when it matched to the annotated categories but not included in the training. We find that the recall values proportionally increase to the annotation rates in training. That is, LaViSE discovers more unseen concepts with a comprehensive reference dataset.

4.5.5 Effect of Pre-training on Interpretability

To understand the effect of pre-training on the interpretability of LaViSE, we compare two models with the same architecture but different pre-training procedures. Note that existing methods including NetDissect are not suitable for this analysis since new concepts would not be captured from a pre-trained model while LaViSE naturally supports this because it can use any arbitrary network and dataset. More specifically, we take two ResNet18 models trained with the MS COCO dataset for multi-class classifications where one model is pre-trained with ImageNet [13] while the other model is randomly initialized. According to the filter explanations generated by the LaViSE, even though two models

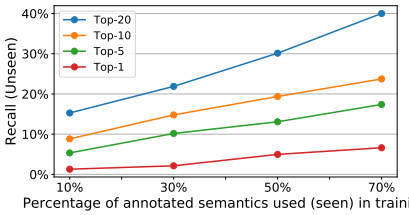


Figure 7. The proportion of the novel (unseen) concepts discovered by LaViSE.

have comparable classification accuracies, the model fine-tuned after pretraining on the ImageNet learned more concepts (219) than the model trained with MS COCO from a random initialization (150). We present the full list of discovered concepts in the Supplementary Material.

4.6. Unsupervised Dataset Bias Analysis

We now demonstrate a novel application of our LaViSE framework for the purpose of **unsupervised** comparative data and model analysis. The main purpose of the analysis is to discover and explain differences between multiple datasets, or different subsets of a given dataset without any labels. Prior work has also shown biases between datasets using labels [55], but we are interested in **interpreting** the differences in an unsupervised fashion. This approach allows us to examine hidden biases in datasets or media outlets using machine learning models [54, 63].

We consider gender representation bias in public datasets as our examples here. Recent studies have reported various gender biases in image datasets and computer vision models such as accuracy disparity [8, 30] or spurious correlation [10, 28, 69]. We mainly consider the latter, i.e. how gender correlates with other unknown covariates in the dataset.[3] For example, Zhao et al. [69] showed that in popular image datasets gender is associated with activities such as shopping for women and driving for men. This analysis is supervised and requires annotations on activities or object categories. In contrast, LaViSE can directly apply to a dataset without any additional labels (except gender) such that it can **discover** hidden biases on unknown factors.

Specifically, we show gender bias in the MS COCO dataset, i.e. which concepts are associated with gender. We split the images by gender according to [69] and train a binary CNN (ResNet-18) classifying gender (we **only** use gender annotations). LaViSE then generates explanations for conv filters in the model. For each filter u, we count the number of images whose maximum activation is above the threshold T_u, and we call these images "qualified images." Each gender is a group, and we compare the difference in percentages of qualified images between two groups. In Figure 8, we list filters that have the most considerable distinctions between gender groups, their top-1 explanations

[3]Some papers also consider causal or counterfactual model bias or explanations [19, 20, 28]. Our main interest is to explain biases in a dataset.

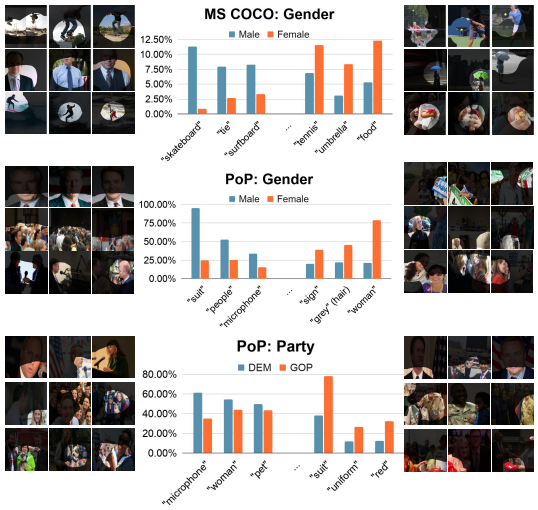


Figure 8. Comparative analysis of different groups of images in the MS COCO dataset and the PoP dataset.

predicted by LaViSE, and examples of qualified images on the sides. The results reflect the gender biases in this benchmark dataset and provide a guideline for future improvements of the dataset. Note that LaViSE can discover concepts which are **not** part of the COCO categories (e.g. food) as it does not use the categories or labels in analysis.

We also use the same method to compare social media photographs of politicians between gender and party affiliations using PoP dataset. Such comparative analysis is essential in social science and media studies [5, 35, 53, 54, 56, 59, 60, 63] but requires a huge amount of human effort for manual coding and may be susceptible to the bias of investigators. LaViSE offers an efficient data-driven way to explore group differences in unlabeled image datasets. The result in Figure 8 shows interesting gender and party differences. For example, male politicians tend to show large crowds to signal competence and popularity [21, 29] and female politicians show more "sign" (panels) commonly used in public demonstrations and protests, which communicates their trustworthiness and interests in social welfare and protection for minority groups [15, 61].

5. Conclusion

We proposed LaViSE, a novel framework which can both visually and semantically explain latent representations of a trained CNN. It also enables users to discover concepts that a CNN learned without being explicitly taught. Empirical results show that our framework can accommodate different CNN architectures and datasets with varying formats of annotations. We also demonstrated a novel application for unsupervised bias analysis using our framework. We hope our work can help enhance transparency in both black-box models and datasets in AI research.

Acknowledgement. This work was supported by NSF SBE-SMA #1831848.

References

[1] Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. In *Advances in Neural Information Processing Systems*, pages 9505–9515, 2018. 2

[2] Harsh Agrawal, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, Stefan Lee, and Peter Anderson. Nocaps: Novel object captioning at scale. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8948–8957, 2019. 7

[3] Yashas Annadani and Soma Biswas. Preserving semantic relations for zero-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7603–7612, 2018. 2

[4] David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. Network dissection: Quantifying interpretability of deep visual representations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6541–6549, 2017. 2, 4, 5, 6, 7

[5] Nichole M Bauer and Colleen Carpinella. Visual information and candidate evaluations: The influence of feminine and masculine images on support for female candidates. *Political Research Quarterly*, 71(2):395–407, 2018. 8

[6] Tamara L Berg, Alexander C Berg, and Jonathan Shih. Automatic attribute discovery and characterization from noisy web data. In *European Conference on Computer Vision*, pages 663–676. Springer, 2010. 2

[7] Moritz Bohle, Mario Fritz, and Bernt Schiele. Convolutional dynamic alignment networks for interpretable classifications. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10029–10038, 2021. 2

[8] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pages 77–91, 2018. 1, 8

[9] Chaofan Chen, Oscar Li, Daniel Tao, Alina Barnett, Cynthia Rudin, and Jonathan K Su. This looks like that: deep learning for interpretable image recognition. In *Advances in neural information processing systems*, pages 8930–8941, 2019. 2

[10] Yunliang Chen and Jungseock Joo. Understanding and mitigating annotation bias in facial expression recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14980–14991, 2021. 8

[11] Stefano d'Apolito, Danda Pani Paudel, Zhiwu Huang, Andres Romero, and Luc Van Gool. Ganmut: Learning interpretable conditional space for gamut of emotions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 568–577, 2021. 2

[12] Abhijit Das, Antitza Dantcheva, and Francois Bremond. Mitigating bias in gender, age and ethnicity classification: a multi-task convolution neural network approach. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 0–0, 2018. 1

[13] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 7

[14] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 4

[15] Joanna Everitt, Lisa A Best, and Derek Gaudet. Candidate gender, behavioral style, and willingness to vote: Support for female candidates depends on conformity to gender norms. *American Behavioral Scientist*, 60(14):1737–1755, 2016. 8

[16] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Marc'Aurelio Ranzato, and Tomas Mikolov. Devise: A deep visual-semantic embedding model. In *Advances in neural information processing systems*, pages 2121–2129, 2013. 2, 3

[17] Yunhao Ge, Yao Xiao, Zhi Xu, Meng Zheng, Srikrishna Karanam, Terrence Chen, Laurent Itti, and Ziyan Wu. A peek into the reasoning of neural networks: Interpreting with structural visual concepts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2195–2204, 2021. 2

[18] Amirata Ghorbani, Abubakar Abid, and James Zou. Interpretation of neural networks is fragile. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3681–3688, 2019. 2

[19] Yash Goyal, Amir Feder, Uri Shalit, and Been Kim. Explaining classifiers with causal concept effect (cace). *arXiv preprint arXiv:1907.07165*, 2019. 8

[20] Yash Goyal, Ziyan Wu, Jan Ernst, Dhruv Batra, Devi Parikh, and Stefan Lee. Counterfactual visual explanations. In *International Conference on Machine Learning*, pages 2376–2384. PMLR, 2019. 8

[21] Maria Elizabeth Grabe and Erik Page Bucy. *Image bite politics: News and the visual framing of elections*. Oxford University Press, 2009. 8

[22] Jinjin Gu and Chao Dong. Interpreting super-resolution networks with local attribution maps. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9199–9208, 2021. 2

[23] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 5

[24] Lisa Anne Hendricks, Zeynep Akata, Marcus Rohrbach, Jeff Donahue, Bernt Schiele, and Trevor Darrell. Generating visual explanations. In *European Conference on Computer Vision*, pages 3–19. Springer, 2016. 2

[25] Lisa Anne Hendricks, Kaylee Burns, Kate Saenko, Trevor Darrell, and Anna Rohrbach. Women also snowboard: Overcoming bias in captioning models. In *European Conference on Computer Vision*, pages 793–811. Springer, 2018. 1

[26] Zixuan Huang and Yin Li. Interpretable and accurate fine-grained recognition via region grouping. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 2

[27] Dong Huk Park, Lisa Anne Hendricks, Zeynep Akata, Anna Rohrbach, Bernt Schiele, Trevor Darrell, and Marcus Rohrbach. Multimodal explanations: Justifying decisions and pointing to the evidence. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8779–8788, 2018. 1, 2

[28] Jungseock Joo and Kimmo Kärkkäinen. Gender slopes: Counterfactual fairness for computer vision models by attribute manipulation. In *Proceedings of the 2nd International Workshop on Fairness, Accountability, Transparency and Ethics in Multimedia*, pages 1–5, 2020. 8

[29] Jungseock Joo, Weixin Li, Francis F Steen, and Song-Chun Zhu. Visual persuasion: Inferring communicative intents of images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 216–223, 2014. 8

[30] Kimmo Karkkainen and Jungseock Joo. Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1548–1558, 2021. 8

[31] Elyor Kodirov, Tao Xiang, Zhenyong Fu, and Shaogang Gong. Unsupervised domain adaptation for zero-shot learning. In *Proceedings of the IEEE international conference on computer vision*, pages 2452–2460, 2015. 2

[32] Elyor Kodirov, Tao Xiang, and Shaogang Gong. Semantic autoencoder for zero-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3174–3183, 2017. 2

[33] Adam Kortylewski, Bernhard Egger, Andreas Schneider, Thomas Gerig, Andreas Morel-Forster, and Thomas Vetter. Empirically analyzing the effect of dataset biases on deep face recognition systems. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 2093–2102, 2018. 1

[34] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73, 2017. 4

[35] Mireille Lalancette and Vincent Raynauld. The power of political image: Justin trudeau, instagram, and celebrity politics. *American Behavioral Scientist*, 63(7):888–924, 2019. 8

[36] Jimmy Lei Ba, Kevin Swersky, Sanja Fidler, et al. Predicting deep zero-shot convolutional neural networks using textual descriptions. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4247–4255, 2015. 2

[37] Oscar Li, Hao Liu, Chaofan Chen, and Cynthia Rudin. Deep learning for case-based reasoning through prototypes: A neural network that explains its predictions. In *Thirty-second AAAI conference on artificial intelligence*, 2018. 2

[38] Haoyu Liang, Zhihao Ouyang, Yuyuan Zeng, Hang Su, Zihao He, Shu-Tao Xia, Jun Zhu, and Bo Zhang. Training interpretable convolutional neural networks by differentiating class-specific filters. In *European Conference on Computer Vision*, pages 622–638. Springer, 2020. 2

[39] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 4

[40] Meike Nauta, Ron van Bree, and Christin Seifert. Neural prototype trees for interpretable fine-grained image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14933–14943, 2021. 2

[41] Mohammad Norouzi, Tomas Mikolov, Samy Bengio, Yoram Singer, Jonathon Shlens, Andrea Frome, Greg S Corrado, and Jeffrey Dean. Zero-shot learning by convex combination of semantic embeddings. *arXiv preprint arXiv:1312.5650*, 2013. 2

[42] Chris Olah, Alexander Mordvintsev, and Ludwig Schubert. Feature visualization. *Distill*, 2(11):e7, 2017. 1

[43] Jose Oramas, Kaili Wang, and Tinne Tuytelaars. Visual explanation by interpretation: Improving visual feedback capabilities of deep neural networks. In *International Conference on Learning Representations*, 2018. 2

[44] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *Advances in neural information processing systems*, pages 8026–8037, 2019. 5

[45] Genevieve Patterson, Chen Xu, Hang Su, and James Hays. The sun attribute database: Beyond categories for deeper scene understanding. *International Journal of Computer Vision*, 108(1-2):59–81, 2014. 2

[46] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014. 3, 5

[47] Mohammad Rastegari, Ali Farhadi, and David Forsyth. Attribute discovery via predictable discriminative binary codes. In *European Conference on Computer Vision*, pages 876–889. Springer, 2012. 2

[48] Robin Rombach, Patrick Esser, and Björn Ommer. Making sense of cnns: Interpreting deep representations and their invariances with inns. In *European Conference on Computer Vision*, pages 647–664. Springer, 2020. 2

[49] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017. 1, 2

[50] Yujun Shen, Jinjin Gu, Xiaoou Tang, and Bolei Zhou. Interpreting the latent space of gans for semantic face editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9243–9252, 2020. 2

[51] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013. 2

[52] Richard Socher, Milind Ganjoo, Christopher D Manning, and Andrew Ng. Zero-shot learning through cross-modal transfer. In *Advances in neural information processing systems*, pages 935–943, 2013. 2

[53] Zachary C Steinert-Threlkeld, Alexander M Chan, and Jungseock Joo. How state and protester violence affect protest dynamics. *The Journal of Politics*, 84(2), 2022. 8

[54] Christopher Thomas and Adriana Kovashka. Predicting visual political bias using webly supervised data and an auxiliary task. *International Journal of Computer Vision*, 129(11):2978–3003, 2021. 8

[55] Antonio Torralba and Alexei A Efros. Unbiased look at dataset bias. In *CVPR*, pages 1521–1528. IEEE, 2011. 8

[56] Mesut Erhan Unal, Adriana Kovashka, Wen-Ting Chung, and Yu-Ru Lin. Visual persuasion in covid-19 social media content: A multi-modal characterization. *arXiv preprint arXiv:2112.13910*, 2021. 8

[57] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 4

[58] Mei Wang and Weihong Deng. Mitigating bias in face recognition using skewness-aware reinforcement learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9322–9331, 2020. 1

[59] Yu Wang, Yang Feng, Zhe Hong, Ryan Berger, and Jiebo Luo. How polarized have we become? a multimodal classification of trump followers and clinton followers. In *International Conference on Social Informatics*, pages 440–456. Springer, 2017. 8

[60] Yu Wang, Yuncheng Li, and Jiebo Luo. Deciphering the 2016 us presidential campaign in the twitter sphere: A comparison of the trumpists and clintonists. In *Tenth International AAAI Conference on Web and Social Media*, 2016. 8

[61] Donghyeon Won, Zachary C Steinert-Threlkeld, and Jungseock Joo. Protest activity detection and perceived violence estimation from social media images. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 786–794, 2017. 8

[62] Jialin Wu and Raymond J Mooney. Faithful multimodal explanation for visual question answering. *arXiv preprint arXiv:1809.02805*, 2018. 2

[63] Nan Xi, Di Ma, Marcus Liou, Zachary C Steinert-Threlkeld, Jason Anastasopoulos, and Jungseock Joo. Understanding the political ideology of legislators from social media images. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, pages 726–737, 2020. 8

[64] Yongqin Xian, Bernt Schiele, and Zeynep Akata. Zero-shot learning-the good, the bad and the ugly. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4582–4591, 2017. 3

[65] Keren Ye, Narges Honarvar Nazari, James Hahn, Zaeem Hussain, Mingda Zhang, and Adriana Kovashka. Interpreting the rhetoric of visual advertisements. *IEEE transactions on pattern analysis and machine intelligence*, 43(4):1308–1323, 2019. 2

[66] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pages 818–833. Springer, 2014. 1, 2

[67] Quanshi Zhang, Ying Nian Wu, and Song-Chun Zhu. Interpretable convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8827–8836, 2018. 2, 4, 5, 6, 7

[68] Ziming Zhang and Venkatesh Saligrama. Zero-shot learning via semantic similarity embedding. In *Proceedings of the IEEE international conference on computer vision*, pages 4166–4174, 2015. 2

[69] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2941–2951, 2017. 4, 8

[70] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Object detectors emerge in deep scene cnns. *arXiv preprint arXiv:1412.6856*, 2014. 2

[71] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017. 6

[72] Bolei Zhou, Yiyou Sun, David Bau, and Antonio Torralba. Interpretable basis decomposition for visual explanation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 119–134, 2018. 6