Latent Space Explanation by Intervention

Itai Gat[1*], Guy Lorberbom[1*], Idan Schwartz[12], Tamir Hazan[1]

[1] Technion - Israel Institute of Technology
[2] NetApp

Abstract

The success of deep neural nets heavily relies on their ability to encode complex relations between their input and their output. While this property serves to fit the training data well, it also obscures the mechanism that drives prediction. This study aims to reveal hidden concepts by employing an intervention mechanism that shifts the predicted class based on discrete variational autoencoders. An explanatory model then visualizes the encoded information from any hidden layer and its corresponding intervened representation. By the assessment of differences between the original representation and the intervened representation, one can determine the concepts that can alter the class, hence providing interpretability. We demonstrate the effectiveness of our approach on CelebA, where we show various visualizations for bias in the data and suggest different interventions to reveal and change bias.

Introduction

Machine learning is ubiquitous these days, and its impact on everyday life is substantial. Supervised learning algorithms are instrumental to autonomous driving (Lavin and Gray 2016; Bojarski et al. 2016; Luss et al. 2019), they serve people with disabilities (Tadmor et al. 2016), they are applied to improve hearing aids (Schröter et al. 2020; Fedorov et al. 2020), and they are being extensively used in medical diagnosis (Deo 2015; Qayyum et al. 2020; Seo et al. 2020; Richens, Lee, and Johri 2020). These advancements are achieved with complex models, and their decisions are usually not well-understood by their operators. Consequently, model interpretability is becoming an important challenge for contemporary deep nets.

To this end, a magnitude of interpretability methods has been proposed. Current model interpretability methods are geared towards local feature relevance. In other words, given a sample, they quantify how much each feature contributes to the prediction. For example, gradient-based approaches are widely used to interpret the model predictions since they bring forth an insight into the internal mechanism of the model (Simonyan, Vedaldi, and Zisserman 2014; Gu et al. 2018; Sundararajan, Taly, and Yan 2017; Shrikumar, Greenside, and Kundaje 2017; Springenberg et al. 2014). These gradient-based methods produce different explanation maps using the gradient of a class-related output with respect to its input data. While relevance scores in the input sample are human-interpretable, they are sparsely scattered across many pixels and do not accurately indicate concepts. Recently, counterfactual examples were used to measure the causal effect of known data concepts (Goyal et al. 2019; Atzmon et al. 2020; Feder et al. 2021; Rosenberg et al. 2021). However, these methods require costly predefined concepts annotation, which is rarely apparent when learning from data.

In this work, we focus on finding the high-level concepts that explain the classifier's decision. The bias within these concepts is often the result of overlooked priors that are woven into the dataset. In order to reveal bias in the data, we propose a Variational Auto-Encoder (VAE) mechanism that discretizes the input representation. This discrete latent space is able to capture the concepts that are embedded in

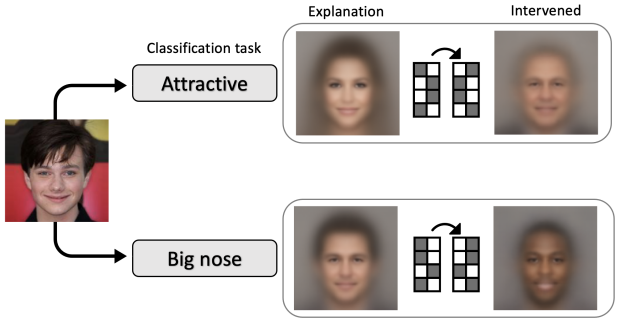

Figure 1: Our method explains a layer of a given discriminative learner. We learn a discrete variational autoencoder of the last layer representation over the data. Then, we produce a visual explanation of it. Using the discrete latent space, we can intervene in the conceptual world of the data, which allows us to 'see' the world from the viewpoint of the classifier. This figure illustrates the explanation for the last layer in two classification tasks on the CelebA dataset: attractive and big nose. We find out that the classifier attributes attractiveness to young women and older man features to the opposite label. This could imply a bias in the annotation process of CelebA and/or reflects an inherent bias in the data.

*These authors contributed equally.

the representation and allows us to avoid costly annotation of predefined concepts that impede the utilization of modern causal effect techniques. Using an intervention mechanism that flips the boolean latent space representation bits, we can infer the concepts that were responsible for the prediction and detect bias in the data.

We take advantage of recent developments in the field of generative learning in order to reconstruct human-friendly explanation images for a discrete latent space concept. Based on this reconstruction, the guiding concepts can be identified. For instance, in Fig. 1, concepts are illustrated abstractly. Following the intervention, examining the shifted concepts, we identify the relevant factors. For instance, age and gender appear to influence attractiveness, while skin color seems to influence a big nose's attractiveness. In addition, we optimize shared information between the visual explanation and classifier in order to ensure that the abstract concepts rely on the same information as the discriminative classifier. This regularization strategy increases the corresponding information and ensures that the realized concepts correspond to the classifier's guiding concepts.

The remainder of the work is organized as follows: In the method section, we present our approach for learning discrete concepts of a hidden layer representation. It allows us to intervene and change the data concepts. We then propose to generate visualizations of the discovered concepts by an explanatory approach. We conclude the method section by providing a theoretic framework for our method via the functional Fisher information and use it as a regularization term. In the results section, we study our method qualitatively and quantitatively on different tasks on CelebA dataset. We first illustrate our intervention mechanism for concept discovery and manipulation. Then, we use our method to detect biases in the discriminative model. We then verify our approach with quantitative and subjective studies.

In summary, our contributions in this work are:

- Our method identifies discrete global concepts in a trained layer. Furthermore, it enables to intervene on those global concepts.
- We propose a novel explanatory approach for explaining a given layer of a trained classifier in a human interpretable manner.
- We integrate discriminative and explanatory functional information into a single information-theoretic framework. We then introduce a regularization term that encourages high shared information between discriminative and explanatory learners.

We study the effectiveness of our approach in generating explanations, detecting bias, and controlling concepts for different tasks on high-dimensional images.

Related Work

Explainability in Computer Vision Explainability takes many forms and has not been formally defined in the literature. In general, the methods can be divided into two branches: global explanations (such as a decision tree's structure) and local explanations (such as important features of samples). Among deep learning explanations, local explanations predominate. The reason for this is that the backpropagated gradients naturally provide a sample with a heatmap that shows a sample's features' relevance. Early attempts created saliency maps based on the gradients (Simonyan, Vedaldi, and Zisserman 2014; Dabkowski and Gal 2017; Mahendran and Vedaldi 2016). Following that, a variety of a mixture of gradient and input methods emerged (Gu et al. 2018; Shrikumar, Greenside, and Kundaje 2017; Smilkov et al. 2017; Srinivas and Fleuret 2019). Attribution propagation approach calculate relevance scores based on a set of axioms (Bach et al. 2015; Montavon et al. 2017; Nam et al. 2020; Shrikumar, Greenside, and Kundaje 2017). Attention models also provide explainability (Schwartz et al. 2019; Schwartz, Schwing, and Hazan 2017, 2019; Braude et al. 2021). Another approach, deconvolution networks, are capable of providing insights into intermediate layers by calculating the transposed convolutional network (Zeiler and Fergus 2014). However, further research suggests that deconvolution is similar to gradient backpropagation (Springenberg et al. 2014). These methods are similar to ours in that they use reconstruction to find relevance. By contrast, our approach reconstructs a concept image to reveal a high-level concept rather than pixel relevance. Li et al. (2018) study the ability of retrained networks to predict relevance maps. Among other popular methods, LIME approximates a model's prediction with a local linear function (Ribeiro, Singh, and Guestrin 2016). Lundberg and Lee (2017) propose SHAP, a unified game-theoretic framework for attribution methods based on Shapley values (Nowak and Radzik 1994). The computational complexity of such methods is considerable, and their efficiency is often not as high as that of other methods. Frosst and Hinton (2017); Wan et al. (2020) examine networks that are explained by the architecture design. Notably, all of the above methods investigate the pixel relevance of individual data points. However, our work seeks to reveal the underlying high-level concepts, and so is also infused with global explanations. A global explanation is favored in assessing the robustness of a model. E.g., it can uncover a global biased concept (e.g., smiling classification is affected by gender) (Doshi-Velez and Kim 2017). Our reconstruction is applied to the deep layers, which typically contain information related mainly to labels, suggesting a global nature (Shwartz-Ziv and Tishby 2017). Hence, it is concise in its representation of the task and does not include all the concepts that pertain to the sample.

Intervention for Causal Discovery In recent years, a mainline of research has focused on quantifying the causal effect of data concepts on prediction. The simplest form of intervention are perturbations, i.e., removing and adding pixels to generate counterfactual local explanations for images (Fong, Patrick, and Vedaldi 2019; Fong and Vedaldi 2017; Goyal et al. 2019). These approaches can also be used to explain models without looking at their mechanics, i.e., a black-box approach. Despite their appealing black-box nature, there are many pixels in a picture, and each pixel can have any value. Therefore, manipulating the combination of

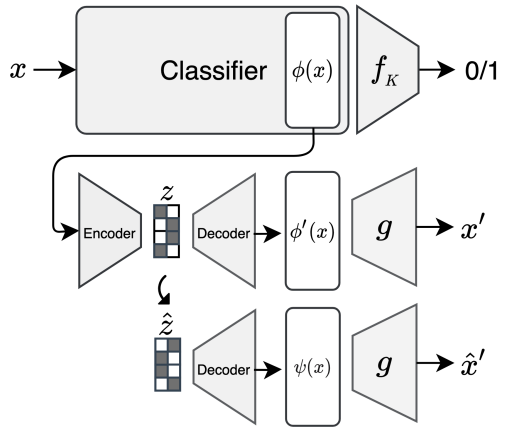

Figure 2: An overview of our approach: Our method provides visual explanations of a given layer of a trained classifier. By using a discrete variational autoencoder, we learn a boolean representation z of $\phi(x)$ that encodes the concepts that drive the classifier prediction. Then, to visualize these concepts a generator is trained to reconstruct the original image given $\phi'(x)$. By intervening in z and getting a modified version of $\phi'(x)$ $\psi(x)$, one can evaluate the qualitative differences using $g(\psi(x))$ and the quantitative differences using $f_K(\psi(x))$. This intervention mechanism provides the ability to interpret the classifier prediction.

pixels can be computationally complex. Our work is different in that it seeks to find high-level concepts by transforming latent space into discrete boolean space using a Variational Auto-Encoder (VAE). As a result, the space of possible interventions is significantly smaller. To be more precise, we study a single flip that shifts the predicted class as the binary values change. Also relevant to our work are causality frameworks, such as TCAV (Kim et al. 2018) and CaCE (Goyal et al. 2019). These methods intervene in labeled concepts to measure the causal effects. However, they require data with labeled concepts. This kind of annotation process must acquire the expertise to identify the underlying structure of the problem, and even then, there is still the human bias that contaminates the process. Using our method, we uncover the task's decisive concepts without labeling them. A concepts image is reconstructed using deep layer representations. Note that deep layers typically contain label-related information (Shwartz-Ziv and Tishby 2017). The consequent advantage is that it is easy to pick up general concepts related to labels. Once we intervene in the concepts images, decisive concepts emerge (see Fig. 1).

Bias in Datasets Also relevant to our work is bias in datasets. The ImageNet challenge (Deng et al. 2009) and the development of AlexNet (Krizhevsky, Sutskever, and Hinton 2012) sparked the deep learning revolution. However, as datasets grow, biases emerge, which may go undetected for a long time. In ImageNet, for instance, the background can indicate information about the class of objects (Szegedy et al. 2015). Multi-modal models only use one modality due to dataset priors (Gat, Schwartz, and Schwing 2021; Gat et al. 2022). In addition to this, a number of datasets are being annotated in uncontrolled environments, as crowdsourcing services such as Amazon Mechanical Turk become more and more popular (Geva, Goldberg, and Berant 2019). For instance, well-known benchmark 'Faces in the wild' (Huang et al. 2008) contains 70% male and 80% white-skinned faces. In our study, we examine a manipulated dataset where we altered the class gender majority to demonstrate that our method identifies hidden factors that affect predictions (see Fig. 4).

Method

Learning a discriminative model amounts to fit function $f : \mathcal{X} \rightarrow \mathcal{Y}$ from the space of data $x \in \mathcal{X}$ to the space of labels $y \in \mathcal{Y}$ using parameters. The parameters of the function $f(x)$ are learned by fitting to a training data $\{(x_1, y_1), ..., (x_m, y_m)\}$ using a loss function $\ell(f(x_i), y_i)$.

Whenever $f(x)$ is a linear function, e.g., in the case of logistic regression and support vector machine, one can explain the importance of features x by the magnitude of the learned coefficient parameters. Unfortunately, this is not the case when $f(x)$ is learned by a deep net.

Deep nets are able to learn complex relations from x to y to their recursive structure. For instance, consider a fully connected deep net, where the input vector of the k-th layer is a function of the parameters of all previous layers, i.e., $\phi_k(x)$. The entries of $\phi_k(x)$ are computed from the response of its preceding layer, i.e., by the linear relation $W_{k-1}\phi_{k-1}(x)$, followed by a non-linear transfer function $\phi_k(x) = \sigma(W_{k-1}\phi_{k-1}(x))$. While this recursive structure allows to easily encode complex functions $f(x)$, it also renders the model decisions hard to explain. Specifically, we are interested in the last hidden representation since it primarily encapsulates label-related information (Tishby and Zaslavsky 2015). For notation clarity, we denote it as $\phi(x) = f_{0:K-1}(x)$, where K is the last hidden representation. We also denote f_K as the remaining last layer of the classifier. In the following, we propose a two-stage explanation of recursive representations of a deep network. In the first stage, concepts are detected using a clustering over $\phi(x)$ (see Sec. *Discrete concepts in a hidden layer of a discriminative model*). In the second stage, we produce visual explanations using an intervention mechanism (see Sec. *Visual explanation of a latent space*). The generated images can be compared to the original data point x under different interventions. The entire flow of our approach is illustrated in Fig. 2. The final section introduces a regularization term, which encourages the generator to use the same information as the classifier to illustrate the concepts used by the classifier accurately (see Sec. *Maximizing explanatory and discriminative shared information*). In the experimental validations, we demonstrate how such interventions can reveal bias in the data. In the following, we begin by introducing our intervention mechanism.

Discrete Concepts in a Hidden Layer of a Discriminative Model

Our goal is to reveal and visualize the underlying concepts that drive the model's prediction. Broadly, a classifier encodes information in a hidden layer in order to discriminate between the different classes of the label space. Thus its hidden layer representations correlate with the relevant target classes. For example, hidden layers may encode body parts when learning to recognize a person.

Unfortunately, the continuous nature of a hidden layer of the discriminative model does not allow us to reveal these concepts easily. For this purpose, we employ a discrete variational autoencoder (DVAE) with boolean latent variables $z \in \{0,1\}^n$ (Rolfe 2016; Lorberbom et al. 2019). In our case, the DVAE learns discrete concepts z of its input vector $\phi(x)$. This input vector is a hidden layer of the discriminative model whose discrete concepts we want to reveal. The DVAE learns these concepts by maximizing the likelihood $p(\phi(x))$ using the negative evidence lower bound (ELBO): $-\log p(\phi(x)) \leq$

$$-\mathbb{E}_{z\sim q(\cdot|\phi(x))}\Big[\log p(\phi(x)|z)\Big] - KL\Big(q(z|\phi(x))||p(z)\Big), \quad (1)$$

where $q(z|\phi(x))$ is a probability distribution of the n dimensional boolean variable $z \in \{0,1\}^n$. We use the Gumbel-softmax reparametrization for estimating the encoder's gradients through the non-differentiable boolean layer (Maddison, Mnih, and Teh 2016; Jang, Gu, and Poole 2016).

We denote by $\phi'(x)$ the reconstructed embedding vector (see Fig. 2). The boolean latent space $z = (z_1, ..., z_n)$ enables to reveal discrete concepts of $\phi(x)$. The generative nature of the DVAE allows us to turn turn-off or turn-on some concepts from $z = (z_1, ..., z_n)$ using an intervention $\hat{z} = (\hat{z}_1, ..., \hat{z}_n)$. This interventions mechanism allows us to modify the generated representation of $\phi'(x)$ to the counterfactual representation $\psi(x)$. Although the DVAE allows us to identify the concepts that the discriminative classifier relies on, it is unclear what these concepts are. To visualize these concepts, we produce human interpretable explanations by generating visual explanations derived from a given latent representation $\phi'(x)$.

Visual Explanation of a Latent Space

A trained DVAE allows us to learn latent concepts $z \in \{0,1\}^n$ that govern the hidden layer representation $\phi(x)$. Our goal is to intervene and change latent concepts, from z to $\hat{z}$, and consequently to change the DVAE decoding from $\phi'(x)$ to $\psi(x)$ in order to detect bias of the discriminative learner. To visualize this bias, we suggest to learn the explanatory function $g : \mathbb{R}^d \to \mathbb{R}^{d_c \times d_h \times d_w}$, where d is the $\phi(x)$'s embedding dimension and $d_c \times d_h \times d_w$ is the generated image's embedding dimension. Thus, g produces a visualization of the discrete concepts in the latent space using the x' from the hidden representation $\phi'(x)$. Intuitively, the role of g is to output a visual explanation. We learn g to minimize a reconstruction loss of the original image,

$$\ell(g(\phi'(x)), x). \quad (2)$$

Visualizing the counterfactual representation $g(\psi(x))$ allows us to 'see' through the eyes of the layer we investigate (see Fig. 1). Note, training the DVAE and the explanatory network together might interfere with the DVAE's discrete structure learning processes with features related to the image. This is undesirable since the DVAE's role is to learn the representation $\phi(x)$ and not input-related features. Thus, we first train the DVAE and then train g while the DVAE is fixed.

By explaining a latent representation of a classifier f through a learning process, it is not guaranteed that g relies on the same information as the classifier. To encourage this, we propose to measure their shared functional information and maximize it during the learning process of g. In the next section, we connect the notion of functional information to our settings and then present a regularization term that maximizes the shared information between the explanatory network and the discriminative learner.

Maximizing Explanatory and Discriminative Shared Information

Our next goal is to ensure that our generated concepts follow the same concepts the discriminator employ. We achieve this by maximizing the amount of information that the explanatory learner (i.e., g) extracts from the latent representation with respect to the discriminative learner's (i.e., f_K) information.

To measure information we follow Gat et al. (2020) and consider the functional information that is encapsulated in a non-negative function $h(z)$ over a Gaussian space $\nu = \mathcal{N}(\mu, 1)$,

$$\mathcal{I}_\nu(h) \triangleq \mathbb{E}_{z\sim\nu}\Big[\frac{h'(z)^2}{h(z)}\Big]. \quad (3)$$

The functional Fisher information is computationally appealing as it can be easily evaluated using sampling over a Gaussian space.

Based on this measure, we compute the information of our explanatory function $g(z)$ with respect to the discriminative layer $\phi(x)$ over the explanatory distribution $\mathcal{N}(\phi(x), I)$. Since $\mathcal{N}(\phi(x), I)$ is a multivariate Gaussian, we estimate the information with respect to each of its coordinates independently. We denote by ν_i the univariate Gaussian measure $\nu_i = \mathcal{N}(\phi^i(x), 1)$, where ϕ^i is ϕ's i-th element. We thus define the information in each neuron i in any given layer by

$$\mathcal{I}^i_\nu(g) = \mathbb{E}_{z\sim\nu_i}\Big[\frac{g'(z_i^\phi)^2}{g(z_i^\phi)}\Big], \quad (4)$$

where $z_i^\phi = (\phi^1(x), ..., \phi^{i-1}(x), z, \phi^{i+1}(x), ..., \phi^d)$. Then, a layer information defined as

$$\mathcal{I}_\nu(g) \triangleq (\mathcal{I}^1_\nu(g), ..., \mathcal{I}^d_\nu(g)). \quad (5)$$

Next, we integrate the functional information into the explanatory learning process. We repeat the same procedure for the discriminative function $f_K(\cdot)$ over the same intermediate layer $\phi(x)$. We augment the discriminative layer $\phi(x)$ with a Gaussian space and define $\mathcal{I}_\nu(f_K)$ as in Eq. (5).

Finally, we encourage the generator to rely on the same information as the discriminative learner. To this end, we

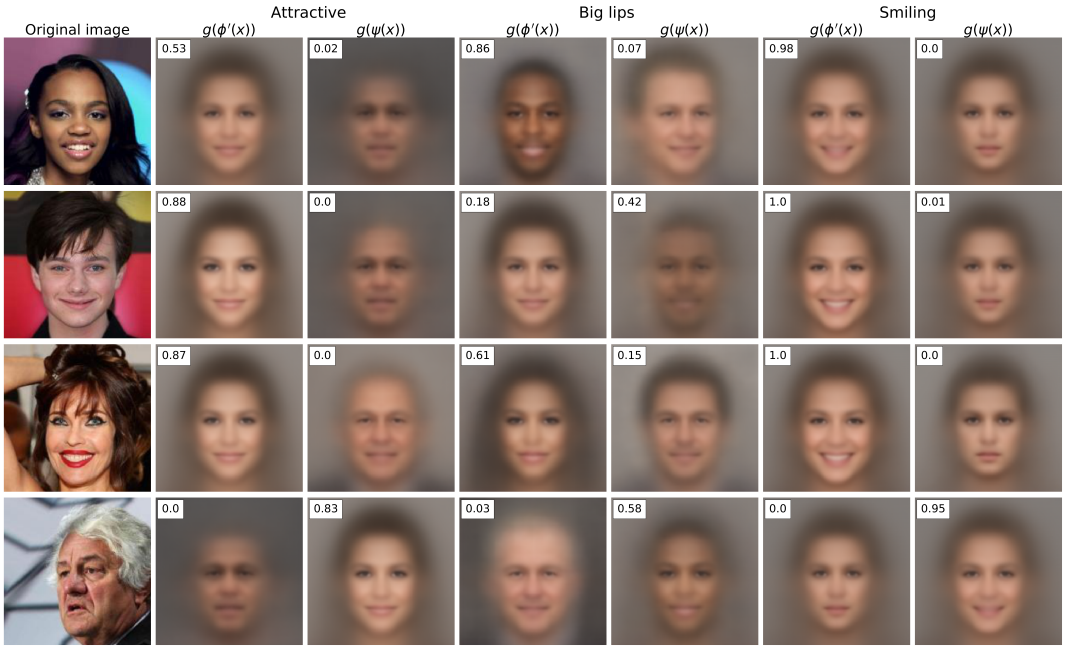

Figure 3: Explanation by intervention: we present results from our method on three different binary tasks (attractive, big lips, and smiling). For each task, we present our method's explanation on the representation of the original input ($g(\phi'(x))$) and the flipped discrete representation ($g(\psi(x))$). In addition, we present the probability that the classifier predicted for the two opposite representations. We note that our methods capture biases in the dataset effectively. Moreover, we forward the flipped representations to the classifier and find that the classifier predicts the opposite class for all input samples.

add a regularization term to the generator training process. The regularization incorporates a differentiable similarity between the information of the generator. $\mathcal{I}_\nu(g)$ and $\mathcal{I}_\nu(f_K)$ where the input of g is $\phi'(x)$ and the input of f_K is $\phi(x)$. We employ dot product for similarity and take the inverse in order to account for both reconstruction loss minimization and similarity maximization (Tang et al. 2020). This leads to our learning objective,

$$\ell(g(\phi'(x), x) + \lambda(\mathcal{I}_\nu(g) \cdot \mathcal{I}_\nu(f_K)^\top)^{-1}, \tag{6}$$

where λ is a hyperparameter that calibrates the training loss and the inverse similarity.

Results

We propose an explanatory approach for latent space explanations using an intervention mechanism. This section studies our explainability method over different classification tasks. We show qualitatively that our intervention mechanism reveals different concepts for different tasks (see *Counterfactual representation*). We then show that our approach reveals deliberately integrated gender bias (see *Dataset statistics control*). To further verify that the classifier is utilizing the detected concept, we verify the prediction of counterfactual samples. To this end, we employ a state-of-the-art generative model and examine the prediction of manipulated images according to the revealed bias (see *Counterfactual verification*). In addition, we provide two quantitative measures: 1) We assess the quality of our intervention by measuring the number of times the intervention affected the classifier predictions (see Tab. 1). 2) We investigate the performance of our proposed regularization term by studying the generator, and the classifier shared information (see Tab. 2).

Experimental Setup We performed our experiments on the CelebA dataset (Liu et al. 2015), which is annotated with 40 binary attributes. We used its binary attributes as labels for different classification tasks (e.g., attractive, big lips, smiling). For each task, we trained a different discriminative learner (i.e., f). Note that the data splits are not necessarily balanced. For example, the big lips attribute is a minority class (i.e., 48,785 vs. 113,985). The discriminative learner is composed of six convolutional layers followed by two fully connected layers. $\phi(x)$ is the output obtained after the first fully connected layer. For the DVAE, the encoder is composed of four linear layers. The decoder is composed of the transposed layers of the encoder. For the reconstruction loss of the DVAE, we used mean squared error loss, and for the reconstruction of the explanatory network, we used binary cross-entropy. In all networks, we used the ReLU activation function and Dropout. We used the same architectures for all the tasks.

Counterfactual Representation

In the following, we study the ability of our mechanism to identify discrete concepts used by the classifier. In Fig. 3, we present our method on various binary classification tasks such as attractiveness, big lips, and smiling (more studies presented in the supplementary material). We show that the

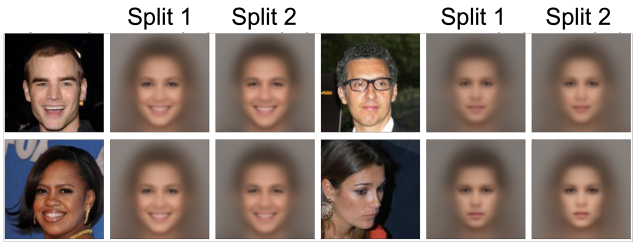

Figure 4: An illustration of our method applied to two smiling/non-smiling classifiers trained on different splits. Each split with a different gender majority. A majority of the females are smiling in the first split, while most males are smiling in the second. The results show how male and female concepts are encoded: on the left, when the original images are smiling, the explanations of the female-majority classifier appear feminine, while the male majority appear masculine. On the right, when the original images are not smiling, the explanations are the opposite, i.e., the female-majority classifier appears masculine, while the male majority appears feminine. These findings demonstrate that existing biases in the data, and consequently in the model, emerge in the explanations obtained in our method.

Task	Change	$\phi(x)$	$\phi'(x)$	Agreement
Smiling	98.86	92.10	91.80	98.08
Attractive	97.08	79.08	78.66	95.98
Big lips	92.13	85.20	80.33	89.70

Table 1: To evaluate the quality of the DVAE for our purposes, we suggest two metrics. The first considers the ability of our method to perform a boolean flip in the prediction. The measure quantifies the percentage of cases in which the prediction changes when the discrete representation is flipped. We observe that, for all tasks, discrete flip changes the prediction to the opposite label. The second evaluates the reconstruction quality. For this, we compute the accuracy for $\phi(x)$ and $\phi'(x)$. We also report the percentage of times the predictions agree on the same label.

concepts images (i.e., $g(\phi'(x))$) depict similar abstract representations regardless of the data input. The reason for this is that deep layers only contain information specific to downstream work. As a result of the narrowing of information, the generator must compensate for the lost information to reconstruct the original image. Thus, the shown concepts are mostly associated with the downstream task label. For instance, the first three images of the second column all portray a young woman with feminine features, despite the fact that the three individuals are of different ages, skin colors, and sex. Next, we will achieve counterfactual representation through intervention. Note, this intervention is possible due to the use of binary representation. Our study illustrates interesting concepts that affect the prediction: 1) In the second and third columns, we examine attractiveness and discover that counterfactual samples of attractive persons display an older figure. 2) In the fourth and fifth columns, we

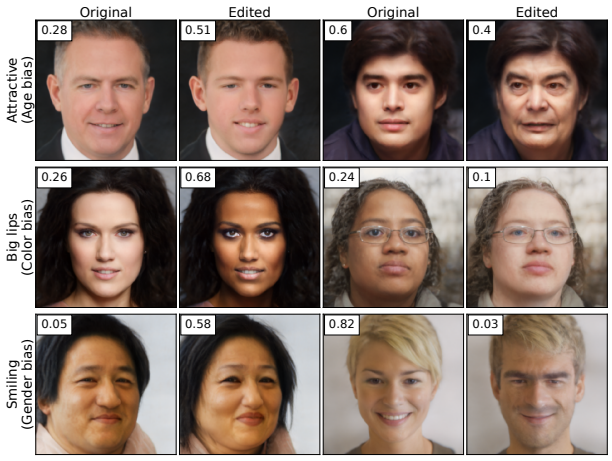

Figure 5: We edited input images to create their counterfactual biased versions in order to test whether the bias concepts detected by our method affect the classifier predictions. Every row represents a different classification task. The first and third columns show the original images with the absence and presence of the label attribute, respectively. In the second and fourth columns, we show the edited image. In addition, we show the prediction score for both the original and edited images. For instance, in the second row, the skin color affects the predicted scores for the big lips task (the prediction score is 0.26 vs. 0.68).

examine the big lips classification and conclude that intervention produces different skin colors. 3) In the sixth and seventh columns, we examine the smiling classification and find that the gender characteristics flip once the intervention is conducted. Note, each user has the option of determining whether or not these concepts constitute biases. Our goal is to expose which concepts the classifier might pick, not to determine whether or not they are biased. For instance, whether age can be used to classify attractiveness is at the discretion of the user. We next propose metrics to ensure our DVAE works properly.

In Tab. 1, we measured the accuracy when we inject the reconstructed representation $\phi'(x)$ to f_K (i.e., $f_K(\phi'(x))$). We tested similarities between the reconstructed representation ($f(\phi'(x))$) and the original representation ($\phi(x)$). We found that the reconstructed representation keep most of the original accuracy performance. For instance, for the smiling task, the accuracy of the reconstruction representation is 91.80%, while the original accuracy is 92.10%. We further suggest to quantify the quality of the intervened representation by the percentage of input data points that change their prediction. We show that our method flips the prediction when we employ the flipped discrete representation (i.e., $\psi(x)$). We found 98.86%, 97.08%, 92.13% percent of the samples flipped prediction for the smiling, attractive, and big lips classification respectively.

Dataset Bias

Bias in facial analysis has gained increasing attention over time (Wang et al. 2020; Xu et al. 2020). Those biases are

Task	Without regularization	With regularization
Smiling	0.43	0.84
Attractive	0.37	0.80
Big lips	0.51	0.85

Table 2: We perform an ablation study for our regularization term. For this, we report the cosine similarity between the explanatory and discriminative information through the DVAE, i.e., $\cos\left(\mathcal{I}_\nu(g), \mathcal{I}_\nu(f)\right)$ where the input of g is $\phi'(x)$ and the input of f is $\phi(x)$. We observe that with our regularization, the shared explanatory and discriminative information is significantly higher.

attributed to spurious correlations in the data. For example, in CelebA, most of the women are smiling while men are not. As a result, classifiers learn to exploit these correlations resulting in biased models. Traditionally, practitioners had to guess the bias of a particular concept and test the classifier accordingly. We reveal those concepts in an unsupervised manner using our method. In the following, we intentionally introduce gender bias into a dataset. We then demonstrate how our method reveals these biases.

Dataset Statistics Control Following Wang et al. (2020), we investigated the smiling recognition task. We showed that our method depicts feminine features in cases where the classification network predicts 'smiling' and muscular features when it predicts 'not smiling.'

To further study this bias, we trained two classifiers with different data splits 1) The female majority split, 70% of smiling women and 30% of non-smiling men. 2) The male majority split, 70% of smiling men and 30% of non-smiling women. Fig. 4 illustrates that our method recognizes the integrated sex concept. In the first split, smiling faces are illustrated as female (i.e., the second column) and non-smiling faces as male (i.e., the fifth column). On the other hand, the opposite occurs with the second split (i.e., the third and sixth columns)

Counterfactual Verification Generative adversarial networks have recently enhanced our ability to manipulate images to change their properties and generate realistic images. One example is StyleCLIP (Patashnik et al. 2021), which enables text-guided manipulations. We employed this technique and created counterfactual samples. We then measured the effect and assessed the classifier predictions. Fig. 3 illustrates that age, color, and gender affect the attractive, big-lips, and smiling classifiers respectively. Thus, we manipulated images according to these concepts. We altered the concepts as follows: 1) Age by the prompts "old face" and "young face", 2) Skin color by the prompts "black person" and "white person", and 3) Sex by the prompts "male face" and "female face". We provide more details in the supplementary.

Classifier-Generator Shared Information

The main objective of the generator loss is to reconstruct the original image, given a distilled representation of the image. However, the reconstruction loss term alone does not consider to what extent the classifier utilizes each feature of $\phi(x)$. Features in $\phi(x)$ may be relevant for the prediction but not for the reconstruction and vice-versa. In the following, we examined the essential role of our proposed regularization term for generating a reliable explanation.

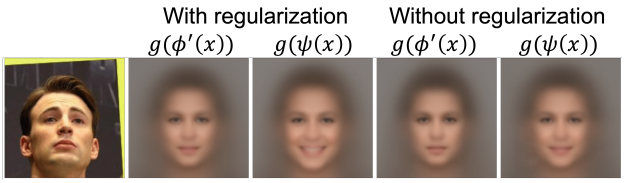

Figure 6: An illustration of the regularization effect on generated images. A non-smiling person is shown for the smiling task. Our findings indicate that regularization helps to change relevant concepts of prediction through intervention. In the example above, a non-smiling generated image became smiling upon regularization, but it remains non-smiling without regularization.

Quantitative Analysis In this experiment, we measure the cosine similarity between the explanatory and discriminative information of g and f respectively. A cosine similarity metric aims to quantify the similarity between two vectors in the inner product space. This metric is bounded in the range of $[-1, 1]$. Two information vectors that are oriented the same way yield the maximum value of 1. One can notice that our regularization term is proportional to the cosine similarity. Tab. 2 shows that our regularization term significantly improves the cosine similarity between the explanatory and discriminative information vectors across tasks.

Qualitative Analysis Fig. 6 demonstrates the effects of regularization on generated images. Generated images serve to reveal concepts that the classifier employs in making predictions. Thus, interventions should change the concept in a way that also changes prediction. In the smiling task, a non-smiling person is expected, upon intervention, to have a smiling concept image. As we can see on the right, without regularization, the intervention does not result in a smiling figure with teeth. However, on the left, with regularization, an intervention leads to a smiling figure with exposed teeth.

Conclusion

Model interpretability plays an essential role in deep neural network debugging. Explanations provide insights about the model's performance rather than just looking at its test-set metrics. It allows a data expert without neural network knowledge to understand why a model made a prediction. By doing so, confidence in such models improves. In this work, we propose an explanatory approach to latent layer explanation. Our proposed method permits the discovery of the underlying concepts a model is using. Furthermore, it allows interventions on those revealed concepts. With our method, we hope to assist researchers in better understanding the operation of their models.

Acknowledgements

This research was supported by Grant No. 2029378 from the United States-Israel Binational Science Foundation (BSF).

References

Atzmon, Y.; Kreuk, F.; Shalit, U.; and Chechik, G. 2020. A causal view of compositional zero-shot recognition. *Advances in Neural Information Processing Systems*, 33: 1462–1473.

Bach, S.; Binder, A.; Montavon, G.; Klauschen, F.; Müller, K.-R.; and Samek, W. 2015. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7): e0130140.

Bojarski, M.; Del Testa, D.; Dworakowski, D.; Firner, B.; Flepp, B.; Goyal, P.; Jackel, L. D.; Monfort, M.; Muller, U.; Zhang, J.; et al. 2016. End to end learning for self-driving cars. *arXiv preprint arXiv:1604.07316*.

Braude, T.; Schwartz, I.; Schwing, A.; and Shamir, A. 2021. Towards coherent visual storytelling with ordered image attention. *arXiv preprint arXiv:2108.02180*.

Dabkowski, P.; and Gal, Y. 2017. Real time image saliency for black box classifiers. *Advances in neural information processing systems*, 30.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 248–255. Ieee.

Deo, R. C. 2015. Machine learning in medicine. *Circulation*, 132(20): 1920–1930.

Doshi-Velez, F.; and Kim, B. 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.

Feder, A.; Oved, N.; Shalit, U.; and Reichart, R. 2021. Causalm: Causal model explanation through counterfactual language models. *Computational Linguistics*, 47(2): 333–386.

Fedorov, I.; Stamenovic, M.; Jensen, C.; Yang, L.-C.; Mandell, A.; Gan, Y.; Mattina, M.; and Whatmough, P. N. 2020. TinyLSTMs: Efficient neural speech enhancement for hearing aids. *arXiv preprint arXiv:2005.11138*.

Fong, R.; Patrick, M.; and Vedaldi, A. 2019. Understanding deep networks via extremal perturbations and smooth masks. In *Proceedings of the IEEE/CVF international conference on computer vision*, 2950–2958.

Fong, R. C.; and Vedaldi, A. 2017. Interpretable explanations of black boxes by meaningful perturbation. In *Proceedings of the IEEE international conference on computer vision*, 3429–3437.

Frosst, N.; and Hinton, G. 2017. Distilling a neural network into a soft decision tree. *arXiv preprint arXiv:1711.09784*.

Gat, I.; Aronowitz, H.; Zhu, W.; Morais, E.; and Hoory, R. 2022. Speaker Normalization for Self-supervised Speech Emotion Recognition. *arXiv preprint arXiv:2202.01252*.

Gat, I.; Schwartz, I.; and Schwing, A. 2021. Perceptual Score: What Data Modalities Does Your Model Perceive? *Advances in Neural Information Processing Systems*, 34.

Gat, I.; Schwartz, I.; Schwing, A.; and Hazan, T. 2020. Removing bias in multi-modal classifiers: Regularization by maximizing functional entropies. *Advances in Neural Information Processing Systems*, 33: 3197–3208.

Geva, M.; Goldberg, Y.; and Berant, J. 2019. Are we modeling the task or the annotator? an investigation of annotator bias in natural language understanding datasets. *arXiv preprint arXiv:1908.07898*.

Goyal, Y.; Feder, A.; Shalit, U.; and Kim, B. 2019. Explaining classifiers with causal concept effect (cace). *arXiv preprint arXiv:1907.07165*.

Gu, J.; Wang, Z.; Kuen, J.; Ma, L.; Shahroudy, A.; Shuai, B.; Liu, T.; Wang, X.; Wang, G.; Cai, J.; et al. 2018. Recent advances in convolutional neural networks. *Pattern Recognition*, 77: 354–377.

Huang, G. B.; Mattar, M.; Berg, T.; and Learned-Miller, E. 2008. Labeled faces in the wild: A database forstudying face recognition in unconstrained environments. In *Workshop on faces in'Real-Life'Images: detection, alignment, and recognition*.

Jang, E.; Gu, S.; and Poole, B. 2016. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*.

Kim, B.; Wattenberg, M.; Gilmer, J.; Cai, C.; Wexler, J.; Viegas, F.; et al. 2018. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *International conference on machine learning*, 2668–2677. PMLR.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25.

Lavin, A.; and Gray, S. 2016. Fast algorithms for convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4013–4021.

Li, H.; Ellis, J. G.; Zhang, L.; and Chang, S.-F. 2018. Patternnet: Visual pattern mining with deep neural network. In *Proceedings of the 2018 ACM on international conference on multimedia retrieval*, 291–299.

Liu, Z.; Luo, P.; Wang, X.; and Tang, X. 2015. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, 3730–3738.

Lorberbom, G.; Gane, A.; Jaakkola, T.; and Hazan, T. 2019. Direct optimization through $\arg\max$ for discrete variational auto-encoder. *Advances in neural information processing systems*, 32.

Lundberg, S. M.; and Lee, S.-I. 2017. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30.

Luss, R.; Chen, P.-Y.; Dhurandhar, A.; Sattigeri, P.; Zhang, Y.; Shanmugam, K.; and Tu, C.-C. 2019. Generating contrastive explanations with monotonic attribute functions. *arXiv e-prints*, arXiv–1905.

Maddison, C. J.; Mnih, A.; and Teh, Y. W. 2016. The concrete distribution: A continuous relaxation of discrete random variables. *arXiv preprint arXiv:1611.00712*.

Mahendran, A.; and Vedaldi, A. 2016. Visualizing deep convolutional neural networks using natural pre-images. *International Journal of Computer Vision*, 120(3): 233–255.

Montavon, G.; Lapuschkin, S.; Binder, A.; Samek, W.; and Müller, K.-R. 2017. Explaining nonlinear classification decisions with deep taylor decomposition. *Pattern recognition*, 65: 211–222.

Nam, W.-J.; Gur, S.; Choi, J.; Wolf, L.; and Lee, S.-W. 2020. Relative attributing propagation: Interpreting the comparative contributions of individual units in deep neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 2501–2508.

Nowak, A. S.; and Radzik, T. 1994. The Shapley value for n-person games in generalized characteristic function form. *Games and Economic Behavior*, 6(1): 150–161.

Patashnik, O.; Wu, Z.; Shechtman, E.; Cohen-Or, D.; and Lischinski, D. 2021. Styleclip: Text-driven manipulation of stylegan imagery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2085–2094.

Qayyum, A.; Qadir, J.; Bilal, M.; and Al-Fuqaha, A. 2020. Secure and robust machine learning for healthcare: A survey. *IEEE Reviews in Biomedical Engineering*, 14: 156–180.

Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2016. " Why should i trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 1135–1144.

Richens, J. G.; Lee, C. M.; and Johri, S. 2020. Improving the accuracy of medical diagnosis with causal machine learning. *Nature communications*, 11(1): 1–9.

Rolfe, J. T. 2016. Discrete variational autoencoders. *arXiv preprint arXiv:1609.02200*.

Rosenberg, D.; Gat, I.; Feder, A.; and Reichart, R. 2021. Are VQA systems rad? measuring robustness to augmented data with focused interventions. *arXiv preprint arXiv:2106.04484*.

Schröter, H.; Rosenkranz, T.; Escalante-B, A. N.; Aubreville, M.; and Maier, A. 2020. Clcnet: Deep learning-based noise reduction for hearing aids using complex linear coding. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 6949–6953. IEEE.

Schwartz, I.; Schwing, A.; and Hazan, T. 2017. High-order attention models for visual question answering. *Advances in Neural Information Processing Systems*, 30.

Schwartz, I.; Schwing, A. G.; and Hazan, T. 2019. A simple baseline for audio-visual scene-aware dialog. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12548–12558.

Schwartz, I.; Yu, S.; Hazan, T.; and Schwing, A. G. 2019. Factor graph attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2039–2048.

Seo, H.; Badiei Khuzani, M.; Vasudevan, V.; Huang, C.; Ren, H.; Xiao, R.; Jia, X.; and Xing, L. 2020. Machine learning techniques for biomedical image segmentation: an overview of technical aspects and introduction to state-of-art applications. *Medical physics*, 47(5): e148–e167.

Shrikumar, A.; Greenside, P.; and Kundaje, A. 2017. Learning important features through propagating activation differences. In *International conference on machine learning*, 3145–3153. PMLR.

Shwartz-Ziv, R.; and Tishby, N. 2017. Opening the black box of deep neural networks via information. *arXiv preprint arXiv:1703.00810*.

Simonyan, K.; Vedaldi, A.; and Zisserman, A. 2014. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *In Workshop at International Conference on Learning Representations*. Citeseer.

Smilkov, D.; Thorat, N.; Kim, B.; Viégas, F.; and Wattenberg, M. 2017. Smoothgrad: removing noise by adding noise. *arXiv preprint arXiv:1706.03825*.

Springenberg, J. T.; Dosovitskiy, A.; Brox, T.; and Riedmiller, M. 2014. Striving for simplicity: The all convolutional net. *arXiv preprint arXiv:1412.6806*.

Srinivas, S.; and Fleuret, F. 2019. Full-gradient representation for neural network visualization. *Advances in neural information processing systems*, 32.

Sundararajan, M.; Taly, A.; and Yan, Q. 2017. Axiomatic attribution for deep networks. In *International conference on machine learning*, 3319–3328. PMLR.

Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; and Rabinovich, A. 2015. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1–9.

Tadmor, O.; Wexler, Y.; Rosenwein, T.; Shalev-Shwartz, S.; and Shashua, A. 2016. Learning a metric embedding for face recognition using the multibatch method. *arXiv preprint arXiv:1605.07270*.

Tang, S.; Maddox, W. J.; Dickens, C.; Diethe, T.; and Damianou, A. 2020. Similarity of neural networks with gradients. *arXiv preprint arXiv:2003.11498*.

Tishby, N.; and Zaslavsky, N. 2015. Deep learning and the information bottleneck principle. In *2015 ieee information theory workshop (itw)*, 1–5. IEEE.

Wan, A.; Dunlap, L.; Ho, D.; Yin, J.; Lee, S.; Jin, H.; Petryk, S.; Bargal, S. A.; and Gonzalez, J. E. 2020. NBDT: neural-backed decision trees. *arXiv preprint arXiv:2004.00221*.

Wang, Z.; Qinami, K.; Karakozis, I. C.; Genova, K.; Nair, P.; Hata, K.; and Russakovsky, O. 2020. Towards fairness in visual recognition: Effective strategies for bias mitigation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8919–8928.

Xu, T.; White, J.; Kalkan, S.; and Gunes, H. 2020. Investigating bias and fairness in facial expression recognition. In *European Conference on Computer Vision*, 506–523. Springer.

Zeiler, M. D.; and Fergus, R. 2014. Visualizing and understanding convolutional networks. In *European conference on computer vision*, 818–833. Springer.