Unsupervised Causal Binary Concepts Discovery with VAE for Black-Box Model Explanation

Thien Q. Tran, Kazuto Fukuchi, Youhei Akimoto, Jun Sakuma
[1] University of Tsukuba
[2] Riken AIP
thientquang@mdl.cs.tsukuba.ac.jp, {fukuchi,akimoto,jun}@cs.tsukuba.ac.jp

Abstract

We aim to explain a black-box classifier with the form: 'data X is classified as class Y because X has A, B and does not have C' in which A, B, and C are high-level concepts. The challenge is that we have to discover in an unsupervised manner a set of concepts, i.e., A, B and C, that is useful for explaining the classifier. We first introduce a structural generative model that is suitable to express and discover such concepts. We then propose a learning process that simultaneously learns the data distribution and encourages certain concepts to have a large causal influence on the classifier output. Our method also allows easy integration of user's prior knowledge to induce high interpretability of concepts. Finally, using multiple datasets, we demonstrate that the proposed method can discover useful concepts for explanation in this form.

Introduction

Deep neural network has been recognized as the state-of-the-art model for various tasks. As they are being applied in more practical applications, there is an arising consensus that these models need to be explainable, especially in high-stake domains. Various methods are proposed to solve this problem, including building a model with interpretable components and post-hoc methods that explain trained black-box models. We focus on the post-hoc approach and propose a novel causal concept-based explanation framework.

We are interested in the symbolic explanation: 'data X is classified as class Y because X *has* A, B and *does not have* C' where A, B, and C are high-level concepts. From the linguistic perspective, such an explanation communicates using *nouns* and their *part-whole relation*, i.e., the semantic relation between a part and the whole object. In many classification tasks, especially image classification, the predictions rely on binary components; for example, we can distinguish a panda from a bear by its white patched eyes or a zebra from a horse by its stripe. This is also a common way humans use to classify categories and organize knowledge (Gardenfors 2014). Thus, an explanation in this form should excel in providing human-friendly and organized insights into the classifier, especially for tasks that involve higher-level concepts such as checking the alignment of the black-box model with experts. From now on, we refer to such a concept as *binary concept*. However, we also note that binary concepts might be insufficient for representing useful concepts with continuous domain, such as color or length.

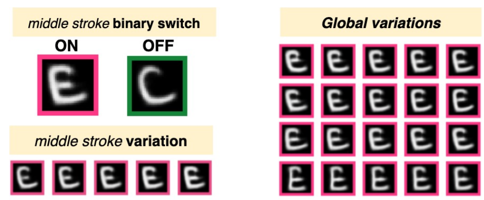
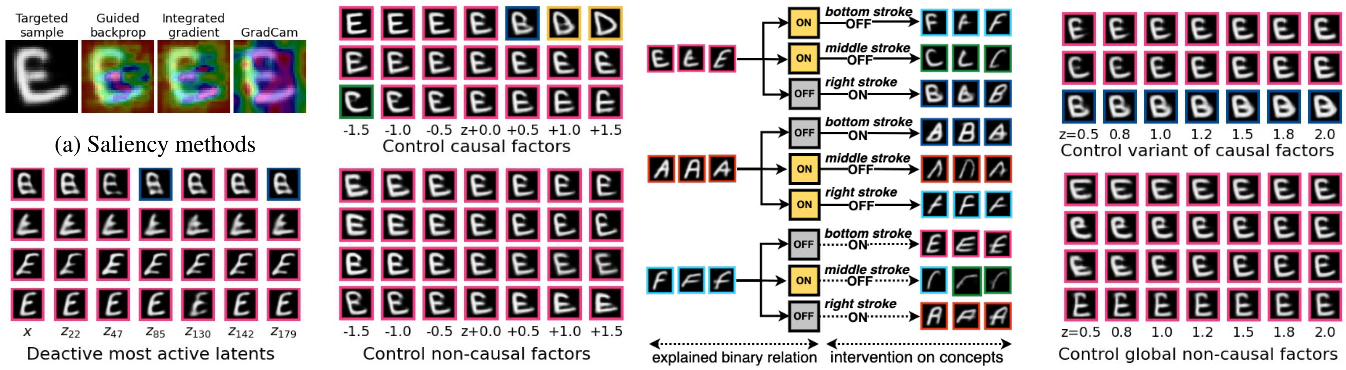


Figure 1: Binary concept *middle stroke* and some global variants. Border color indicates the classifier output.

Our method employs three different notions in the explanation: *causal binary switches*, *concept-specific variants* and *global variants*. We illustrate these notions in Figure 1. First, *causal binary switches* and *concept-specific variants*, that come in pair, represent different binary concepts. In particular, *causal binary switches* control the presence of each binary concept in a sample. Alternating this switch, i.e., removing or adding a binary concept to a sample, affects the prediction of that sample (e.g., removing the middle stroke turns E to C). In contrast, *concept specific variants*, whose each is tied to a specific binary concept, express different variants within a binary concept that do not affect the prediction (e.g., changing the length of the middle stroke does not affect the prediction). Finally, *global variants*, which are not tied to specific binary concepts, represent other variants that do not affect the prediction (e.g., skewness).

Our goal is to discover a set of binary concepts that can explain the classifier using their binary switches in an unsupervised manner. Similar to some existing works, to construct conceptual explanations, we learn a generative model that maps each input into a low-dimensional representation in which each factor encodes an aspect of the data. There are three main challenges in achieving our goal. (1) It requires an adequate generative model to express the binary concepts, including the binary switches and the variants within each concept. (2) The discovered binary concepts must have a large causal influence on the classifier output. We want to avoid finding confounding concepts that correlate with but do not cause the prediction. For example, the *sky* concept ap-



(a) Saliency methods (b) VSC model (c) O'Shaughnessy et al. (2020) (d) Proposed (causal factors) (e) Proposed (non-causal factors)

Figure 2: Explanation methods for a letter classifier. The border color indicates the prediction. (a) Saliency-based methods. (b) Disabling the most active latents of class E in VSC model (Tonolini, Jensen, and Murray-Smith 2020). (c) Controlling the causal and non-causal factors in O'Shaughnessy et al. (2020). (d, e) Proposed method: (d) Encoded binary relation of discovered concepts and their intervention results; (e) the variants within each concept and other variants of the whole letter.

pears frequently in *plane*'s images but may not cause the prediction of *plane*. (3) The explanation must be interpretable and provide useful insights. For example, a concept that entirely replaces a letter E with a letter A has a large causal effect. However, such a concept does not provide valuable knowledge due to a lack of interpretability.

In Figure 2d and 2e, we demonstrate an explanation discovered by the proposed method for a classifier for six letters: A,B,C,D,E and F. Our method successfully discovered the concepts of *bottom stroke*, *middle stroke* and *right stroke* which effectively explains the classifier. In Figure 2d, we show the encoded binary switches and their interventions result. From the top figure, we can explain that: this letter is classified as E because it *has* a *bottom stroke* (otherwise it is F), a *middle stroke* (otherwise it is C), and it *does not have* a *right stroke* (otherwise it is B). We were also able to distinguish the variant within each concept in (Figure 2e top) with the global variant (Figure 2e bottom). A full result with explanation for other letters is shown the experiment section.

To the best of our knowledge, no existing method can discover binary concepts that fulfill these requirements. Saliency methods such as Guided Backprop (Springenberg et al. 2014), Integrated Gradient (Sundararajan, Taly, and Yan 2017) or GradCam (Selvaraju et al. 2017) only show feature importance but do not explain why (Figure 2a). Some generative models which use binary-continuous mixed latents for sparse coding, such as VSC (Tonolini, Jensen, and Murray-Smith 2020), IBP-VAE (Gyawali et al. 2019), PatchVAE (Gupta, Singh, and Shrivastava 2020), can support binary concepts. However, they do not necessarily discover binary concepts that are useful for explanation in both causality and interpretability (Figure 2b). Recently, O'Shaughnessy et al. (2020) proposed a learning framework that encourages the causal effect of certain latent factors on the classifier output to learn a latent representation that has causality on the prediction. However, their model can not disentangle binary concepts and can be hard to interpret, especially for multiple-class tasks. For example, a single concept changes the letter E to multiple other letters (Figure 2c), which would not give any interpretation of how this latent variable affects prediction. Our work has the following contributions:

- We introduce the problem of discovering binary concepts for the explanation. Then, we propose a structural generative model for constructing binary concept explanation, which can capture the binary switches, concept-specific variants, and global variants.
- We propose a learning process to simultaneously learn the data distribution while encouraging the causal influence of the binary switches. Although VAE models typically encourage the independence of factors for meaningful disentanglement, such an assumption is inadequate for discovering useful causal concepts that are often mutually correlated. Our learning process, which considers the dependence between binary concepts, can discover concepts with more significant causality.
- To avoid the concepts with causality but no interpretability, the proposed method allows an easy way to implement users' preferences and prior knowledge as a regularizer to induce high interpretability of concepts.
- Finally, we demonstrate that our method successfully discovers interpretable binary concepts with causality useful for the explanation task.

Related Work

Our method can be categorized as a concept-based method that explains using high-level aspects of data. The definition of *concept* are various, e.g., a direction in the activation space (Kim et al. 2018; Ghorbani et al. 2019), a prototypical activation vector (Yeh et al. 2020) or a latent factor of a generative model (O'Shaughnessy et al. 2020; Goyal et al. 2020). We remark that this notion of concept should depend on the data and the explanation goal. Some works defined the concepts beforehand using additional data and focused on evaluating these concepts. When this side-information is not given, one needs to discover useful concepts for the explanation, e.g., Ghorbani et al. (2019) used segmentation and clustering, Yeh et al. (2020) retrained the classifier with a prototypical concept layer, O'Shaughnessy et al. (2020) learned the generative model with a causal objective.

A generative model such as VAE can provide a concept-based explanation as it learns a latent presentation $\mathbf{z}$ that captures different aspects of the data. However, Locatello et al. (2019) shows that disentangled representations in a fully unsupervised manner are fundamentally impossible without inductive bias. A popular approach is to augment the VAE loss with a regularizer (Higgins et al. 2017; Burgess et al. 2018). Another approach is to incorporate structure into the representation(Choi, Hwang, and Kang 2020; Ross and Doshi-Velez 2021; Tonolini, Jensen, and Murray-Smith 2020; Gupta, Singh, and Shrivastava 2020). Although these methods can encourage disentangled and sparse representation, the learned representations are not necessarily interpretable and have causality on the classifier output.

We pursue an explanation that has causality. A causal explanation is helpful as it can avoid attributions and concepts that only correlate with but do not cause the prediction. Previous works have attempted to focus on causality in various ways. For example, Schwab and Karlen (2019) employed Granger causality to quantify the causal effect of input features, Parafita and Vitrià (2019) evaluated the causality of latent attributions with a prior known causal structure, Narendra et al. (2018) evaluated the causal effect of network layers, and Kim and Bastani (2019) learned an interpretable model with a causal guarantee. Some works first train a generative model and then search for counterfactual samples on latent space(Joshi et al. 2019; Dhurandhar et al. 2018). Although these methods can provide a counterfactual explanation for each input sample, their generative models do not necessarily disentangle useful concepts. Some works introduce the causal structure into the generative models(Yang et al. 2020; Kocaoglu et al. 2017). These methods are not applicable in our setting because they require additional knowledge, e.g., causal graphs or concept labels. To the best of our knowledge, no existing works can explain using concepts that fulfill our three requirements.

Preliminaries

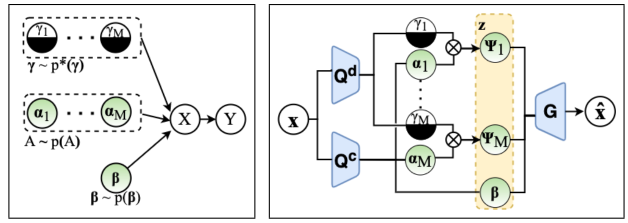
Variational Autoencoder

Our explanation is build upon the VAE framework proposed by Kingma and Welling (2014). VAE model assumes a generative process of data in which a latent $\mathbf{z}$ is first sampled from a prior distribution $p(\mathbf{z})$, then the data is generated via a conditional distribution $p(\mathbf{x} \mid \mathbf{z})$. Typically, due to the intractability, a variational approximation $q(\mathbf{z} \mid \mathbf{x})$ of the intractable posterior is introduced and the model is then learned using the evidence lower bound (ELBO) as

$$\mathcal{L}_{\text{VAE}}(\mathbf{x}) = -\mathbb{E}_{\mathbf{z}\sim q(\mathbf{z}|\mathbf{x})}[\log p(\mathbf{x} \mid \mathbf{z})] + \mathbb{KL}[q(\mathbf{z} \mid \mathbf{x}) \parallel p(\mathbf{z})].$$

Here, $q(\mathbf{z} \mid \mathbf{x})$ is the encoder that maps the data to the latent space and $p(\mathbf{x} \mid \mathbf{z})$ is the decoder that maps the latents to the data space. Commonly, $q(\mathbf{z} \mid \mathbf{x})$ and $p(\mathbf{x} \mid \mathbf{z})$ are parameterized as neutral networks $Q(\mathbf{z} \mid \mathbf{x})$ and $G(\mathbf{x} \mid \mathbf{z})$, respectively. The common choice for $q(\mathbf{z} \mid \mathbf{x})$ is a factorized Gaussian encoder $q(\mathbf{z} \mid \mathbf{x}) = \prod_{p=1}^{P} \mathcal{N}(\mu_i, \sigma_i^2)$ where $(\mu_1, \ldots, \mu_P, \sigma_1, \ldots, \sigma_P,) = Q(\mathbf{x})$. The common choice for the $p(\mathbf{z})$ is a multi-variate normal distribution $\mathcal{N}(0, \mathcal{I})$ with zero mean and identity covariant. Letting $\hat{\mathbf{x}}$ be the reconstruction of input $\mathbf{x}$, the VAE objective can be written as follows and optimized via the reparameterization trick:

$$\mathcal{L}_{\text{VAE}}(\mathbf{x}) = \|\hat{\mathbf{x}} - \mathbf{x}\|^2 + \mathbb{KL}[q(\mathbf{z} \mid \mathbf{x}) \parallel \mathcal{N}(0, \mathcal{I})]. \quad (1)$$



(a) Causal DAG (b) VAE model

Figure 3: Proposed VAE model and the causal DAG

Information Flow

Next, we introduce the measure we use to quantify the causal influence of the learned representation on the classifier output. We adopt Information Flow, which defines the causal strength using Pearl's do calculus (Pearl 2009). Given a causal directional acyclic graph G, Information Flow quantify the statistical influence using the conditional mutual information on the interventional distribution:

Definition 1 (Information flow from U to V in a directed acyclic graph G(Ay and Polani 2008))**.** *Let U and V be disjoint subsets of nodes. The information flow $I(U \to V)$ from U to V is defined by*

$$\int_U p(u) \int_V p(v|do(u)) \log \frac{p(v|do(u))}{\int_{u'} p(u')p(v|do(u'))du'} dV dU, \quad (2)$$

where $do(u)$ represents an intervention in a causal model that fixes u to a value regardless of the values of its parents.

O'Shaughnessy et al. (2020) argued that compared to other metrics such as average causal effect (ACE) (Holland 1988), analysis of variance (ANOVA) (Lewontin 1974), information flow is more suitable to capture complex and nonlinear causal dependence between variables.

Proposed Method

We aim to discover a set of binary concepts $\mathcal{M} = \{m_0, m_1, \ldots, m_M\}$ with causality and interpretability that can explain the black-box classifier $f : \mathcal{X} \to \mathcal{Y}$. Inspired by O'Shaughnessy et al. (2020), we employs a generative model to learn the data distribution while encouraging the causal influence of certain latent factors. In particular, we assume a causal graph in Figure 3a, in which each sample $\mathbf{x}$ is generated from a set of latent variables, including M pairs of *a binary concept* and *a concept-specific variant* $\{\gamma_i, \boldsymbol{\alpha}_i\}_{i=1}^{M}$, and a *global variants* $\boldsymbol{\beta}$. As we want to explain the classifier output (i.e., node y in Figure 3a) using the *binary switches* $\{\gamma_i\}$, we expect that $\{\gamma_i\}$ has a large causal influence on y.

Our proposed learning objective consists of three components, which corresponds to our three requirements: a VAE objective $\mathcal{L}_{\text{VAE}}$ for learning the data distribution $p(\mathbf{x})$, a causal effect objective $\mathcal{L}_{\text{CE}}(X)$ for encouraging the causal influence of $\{\gamma_i\}$ on classifier output y, and an user-implementable regularizer $\mathcal{L}_{\text{R}}(\mathbf{x})$ for improving the interpretability and consistency of discovered concepts:

$$\mathcal{L}(X) = \frac{1}{|X|}\sum_{\mathbf{x}\in X}\left[\mathcal{L}_{\text{VAE}}(\mathbf{x}) + \lambda_{\text{R}}\mathcal{L}_{\text{R}}(\mathbf{x})\right] + \lambda_{\text{CE}}\mathcal{L}_{\text{CE}}(X). \quad (3)$$

VAE Model with Binary Concepts

To represent the binary concepts, we employ a structure in which each binary concept m_i is presented by a latent variable $\boldsymbol{\psi}_i$, which is further controlled by two factors: a binary concept switch latent variable γ_i (concept switch for short) and a continuous latent variable representing concept-specific variants $\boldsymbol{\alpha}_i$ (concept-specific variant for short) as

$$\boldsymbol{\psi}_i = \gamma_i \cdot \boldsymbol{\alpha}_i, \text{ where } \gamma_i = \begin{cases} 1, & \text{if concept } m_i \text{ is on} \\ 0, & \text{otherwise.} \end{cases} \quad (4)$$

Here, the *concept switch* γ_i controls if the concept m_i is activated in a sample, e.g., controlling if the bottom stroke is appeared in a image (Figure 2d). On the other hand, the *concept-specific variant* $\boldsymbol{\alpha}_i$ controls the variant within the concept m_i, e.g., the length of the bottom stroke (Figure 2e, top). In addition to the *concept-specific variants* $\{\boldsymbol{\alpha}_i\}$ whose effect is limited to a specific binary concept, we also allow a *global variant* latent $\boldsymbol{\beta}$ to capture other variants that do not necessarily have causal influence, e.g., skewness (Figure 2e, bottom). Here, disentangling the concept-specific variant and the global variant is important as it can assist users in understanding discovered binary concepts.

The way we represent binary concepts is closely related to the spike-and-slab distribution, which is used in Bayesian variable selection (George and McCulloch 1997) and sparse coding (Tonolini, Jensen, and Murray-Smith 2020). Unlike these models, our model uses only a small number of binary concepts and a multi-dimensional global variant β. Our intuition is that the classification is likely made by combining a small number of binary concepts.

Input encoding. Letting $A = (\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \ldots, \boldsymbol{\alpha}_M)$, we use a network $Q^d(\mathbf{x})$ and $Q^c(\mathbf{x})$ to parameterize the variational posterior distribution of the discrete components $q(\boldsymbol{\gamma} \mid \mathbf{x})$ and the continuous components $q(A, \boldsymbol{\beta} \mid \mathbf{x})$, respectively.

$$q(\boldsymbol{\gamma} \mid \mathbf{x}) = \prod_{i=1}^{M} q(\gamma_i \mid \mathbf{x}) = \prod_{i=1}^{M} \text{Bern}(\gamma_i; \pi_i) \quad (5)$$

$$\text{where } (\pi_1, \ldots, \pi_M) = Q^d(\mathbf{x})$$

$$\text{and} \quad q(A, \boldsymbol{\beta} \mid \mathbf{x}) = \left[\prod_{i=1}^{M} q(\boldsymbol{\alpha}_i \mid \mathbf{x})\right] q(\boldsymbol{\beta} \mid \mathbf{x}) \quad (6)$$

$$q(\boldsymbol{\alpha}_i \mid \mathbf{x}) = \mathcal{N}_\delta^{\text{fold}}(\boldsymbol{\alpha}_i; \mu_i, \text{diag}(\sigma_i))$$

$$q(\boldsymbol{\beta} \mid \mathbf{x}) = \mathcal{N}_\delta^{\text{fold}}(\boldsymbol{\beta}; \mu_{\boldsymbol{\beta}}, \text{diag}(\sigma_{\boldsymbol{\beta}}))$$

$$\text{where } (\mu_1, \ldots, \mu_M, \mu_{\boldsymbol{\beta}}, \sigma_1, \ldots, \sigma_M, \sigma_{\boldsymbol{\beta}}) = Q^c(\mathbf{x}).$$

Here, we employ the δ-Shifted Folded Normal Distribution $\mathcal{N}_\delta^{\text{fold}}(\mu, \sigma^2)$ for continuous latents, which is the distribution of $|x| + \delta$ with a constant hyper-parameter $\delta > 0$ where $x \sim \mathcal{N}(\mu, \sigma^2)$. In all of our experiments, we adopted $\delta = 0.5$. We choose not the standard Normal Distribution but the δ-Shifted Folded Normal Distribution because it is more appropriate for the causal effect we want to achieve. We discuss in detail this design choice and its efficacy in our extended version (Tran et al. 2021).

Output decoding. Next, given $q(\boldsymbol{\gamma} \mid \mathbf{x})$ and $q(A, \boldsymbol{\beta} \mid \mathbf{x})$, we first sample the concept switches $\{\hat{d}_i\}$, the concept variants $\{\hat{\boldsymbol{\alpha}}_i\}$ and the global variants $\boldsymbol{\beta}$ from their posterior, respectively. Using these sampled latents, we construct an aggregated representation $\hat{\mathbf{z}} = (\boldsymbol{\psi}_1, \ldots, \boldsymbol{\psi}_M, \hat{\boldsymbol{\beta}})$ using the binary concept mechanism in Eq. (4) in which $\boldsymbol{\psi}_i$ is the corresponding part for concept m_i, i.e., $\boldsymbol{\psi}_i = \gamma_i \times \boldsymbol{\alpha}_i$. That is, if concept m_i is on, we let $\hat{d}_i = 1$ so that $\boldsymbol{\psi}_i$ can reflect the concept-specific variant $\hat{\boldsymbol{\alpha}}_i$. Otherwise, when the concept m_i is off, we assign $\hat{d}_i = 0$. We refer to $\hat{\mathbf{z}}$ as the *conceptual latent code*. Finally, a decoder network takes $\hat{\mathbf{z}}$ as the input and generate the reconstruction $\hat{x}$ as

$$\hat{x} \sim G(\mathbf{x} \mid \hat{\mathbf{z}}) \text{ where } \hat{\mathbf{z}} = (\boldsymbol{\psi}_1, \ldots, \boldsymbol{\psi}_M, \hat{\boldsymbol{\beta}}). \quad (7)$$

Learning process. We use the maximization of evidence lower bound (ELBO) to jointly train the encoder and decoder. We assume the prior distribution for continuous latents to be δ-shifted Folded Normal distribution $\mathcal{N}_\delta^{\text{fold}}(0, \mathcal{I})$ with zero-mean and identity covariance. Moreover, we assume the prior distribution for binary latents to be a Bernoulli distribution $\text{Bern}(\pi_{\text{prior}})$ with prior π_{prior}. The ELBO for our learning process can be written as:

$$\begin{aligned}\mathcal{L}_{\text{VAE}}(\mathbf{x}) = &-\mathbb{E}_{\hat{\mathbf{z}}\sim Q^{\{c,d\}}(\mathbf{z}|\mathbf{x})}\left[\log G(\mathbf{x} \mid \hat{\mathbf{z}})\right] \\ &+ \lambda_1 \mathbb{KL}\left(q(\boldsymbol{\beta} \mid \mathbf{x}) \parallel \mathcal{N}_\delta^{\text{fold}}(0, \mathcal{I})\right) \\ &+ \lambda_1 \left[\frac{1}{M}\sum_{i=1}^{M} \mathbb{KL}\left(q(\boldsymbol{\alpha}_i \mid \mathbf{x}) \parallel \mathcal{N}_\delta^{\text{fold}}(0, \mathcal{I})\right)\right] \\ &+ \lambda_2 \left[\frac{1}{M}\sum_{i=1}^{M} \mathbb{KL}\left(q(\gamma_i \mid \mathbf{x}) \parallel \text{Bern}(\pi_i)\right)\right].\end{aligned} \quad (8)$$

The first term can be trained using L2 reconstruction loss, while other KL-divergence terms are trained using the reparameterization trick. For the Bernoulli distribution, we use its continuous approximation, i.e., the relaxed-Bernoulli (Maddison, Mnih, and Teh 2017) in the training process.

Encouraging Causal Effect of Binary Switches

We expect the binary switches $\boldsymbol{\gamma}$ to have a large causal influence so that they can effectively explain the classifier. To measure the causal effect of $\boldsymbol{\gamma}$ on the classifier output Y, we employ the causal DAG in Figure 3a and adopt *information flow* (Definition 1) as the causal measurement. Our DAG employs an assumption that is fundamentally different from standard VAE models. Specifically, the standard VAE model and also O'Shaughnessy et al. (2020) assumes the independence of latent factors, which is believed to encourage meaningful disentanglement via a factorized prior distribution. We claim that because *useful concepts for explanation* often causally depend on the class information and thus are not independent of each other, such an assumption might be inadequate for discovering valuable causal concepts. For example, in the letter E, the middle and the bottom strokes are causally related to the recognition of the letter E, and corresponding binary concepts are mutually correlated. Thus, employing the VAE's factorized prior distribution to estimate information flow might lead to a large estimation error and prevent the discovery of valuable causal concepts.

Instead, we employ a prior distribution $p^*(\boldsymbol{\gamma})$ that allows the correlation between causal binary concepts. Our method iteratively learns the VAE model and use the current VAE model to estimates the prior distribution $p^*(\boldsymbol{\gamma})$ which most likely generates the user's dataset. $p^*(\boldsymbol{\gamma})$ is then used to evaluate the causal objective in Eq. (3). Assuming X is a set of i.i.d samples from $p(\mathbf{x})$, we estimate $p^*(\boldsymbol{\gamma})$ as

$$p^*(\boldsymbol{\gamma}) \approx \int_{\mathbf{x}} p^*(\boldsymbol{\gamma} \mid \mathbf{x}) p(\mathbf{x}) d\,\mathbf{x} \approx \frac{1}{|X|} \sum_{\mathbf{x} \in X} p(\boldsymbol{\gamma} \mid \mathbf{x})$$

$$\approx \frac{1}{|X|} \sum_{\mathbf{x} \in X} \prod_{i=1}^{M} q(\gamma_i \mid \mathbf{x}) \tag{9}$$

In the last line, $p(\boldsymbol{\gamma} \mid \mathbf{x})$ is replaced with the variational posterior $q(\boldsymbol{\gamma} \mid \mathbf{x})$ of VAE model. Here, the factorized variational posterior $q(\boldsymbol{\gamma} \mid \mathbf{x})$ only assumes the independence between latents conditioned on each sample but does not imply the independence of binary switches in $p^*(\boldsymbol{\gamma})$. We note that here we do not aim to learn the dependence between concepts but only expect that $p^*(\boldsymbol{\gamma})$ properly reflects the dependence between binary concepts that appears in the dataset X for a better evaluation of the causal effect. We will experimentally show that using the estimation of $p^*(\boldsymbol{\gamma})$ results in a better estimation for the causal effect on dataset X and more valuable concepts for the explanation.

As we want to maximize $I(\boldsymbol{\gamma} \to Y)$, we rewrite it as a loss term $\mathcal{L}_{\text{CE}} = -I(\boldsymbol{\gamma} \to Y)$ and optimize it together with the learning of VAE model. We also showed that in the proposed DAG, information flow $I(\boldsymbol{\gamma} \to Y)$ coincides with mutual information $I(\boldsymbol{\gamma}; Y)$.

Proposition 1 (Coincident of Information Flow and Mutual Information in proposed DAG)**.** *The information flow from $\boldsymbol{\gamma}$ to Y in the DAG of Figure 3a coincides with the mutual information between $\boldsymbol{\gamma}$ and Y. That is,*

$$I(\boldsymbol{\gamma} \to Y) = I(\boldsymbol{\gamma}; Y) = \mathbb{E}_{\boldsymbol{\gamma}, Y} \left[\frac{p^*(\boldsymbol{\gamma}) p(Y \mid \boldsymbol{\gamma})}{p^*(\boldsymbol{\gamma}) p(Y)} \right] \tag{10}$$

The proof and the detailed algorithm for estimating $I(\boldsymbol{\gamma}; Y)$ is described in our extended version(Tran et al. 2021).

Integrating User Preference for Concepts

Finally, we discuss the necessity of users' preferences or prior knowledge for inducing high interpretability of concepts. A problem in discovering meaningful latent factors using deep generative models is that the learned factors can be hard to interpret. Although causality is strongly related and can contribute to interpretability, due to the high expressiveness of the deep model, a large causal effect does not always guarantee an interpretable concept. For example, a concept that entirely replaces a letter E with a letter D will significantly affect the prediction. However, it does not provide valuable knowledge and is hard to interpret. To avoid such concepts, we allow users to implement their preference or prior knowledge as an interpretability regularizer to constrain the generative model's expressive power. Our method looks for useful binary concepts with large causality under the constrained search space.

The integration can easily be done via a scoring function $r(\mathbf{x}_{\gamma_i=0}, \mathbf{x}_{\gamma_i=1})$ which evaluates the usefulness of concept m_i. Here, $\mathbf{x}_{\gamma_i=0}$ and $\mathbf{x}_{\gamma_i=1}$ are obtained from the generative model by performing the do-operation $do(\gamma_i = 0)$ and $do(\gamma_i = 1)$ on input $\mathbf{x}$, respectively. In this study, we introduce two regularizers which are based on the following intuitions. First, an interpretable concept should only affect a small amount of input features (Eq. (11)). This desiderata is general and can be applied to many tasks. The second one is more task-specific in which we focus on the gray-scale image classification task. An intervention of a concept should only add or substract the pixel value, but not both at the same time (Eq. (12)). Furthermore, we desire that $\gamma_i = 1$ indicates the *presence* of pixels and $\gamma_i = 0$ indicates the *absence* of pixels. We formulate these regularizers as follows

$$\mathcal{L}_{\text{compact}}(\mathbf{x}) = \frac{1}{M} \sum_{i=1}^{M} \frac{1}{P} \|\hat{\mathbf{x}} - \hat{\mathbf{x}}^{[i]}\|, \tag{11}$$

$$\mathcal{L}_{\text{directional}}(\mathbf{x}) = \frac{1}{M} \sum_{i=1}^{M} \frac{1}{P} \sum_{p=1}^{P} l(\hat{\mathbf{x}}_p, \hat{\mathbf{x}}_p^{[i]}, \gamma_i), \tag{12}$$

$$\begin{aligned} l(\hat{\mathbf{x}}_p, \hat{\mathbf{x}}_p^{[i]}, \gamma_i) = & 1[\hat{\mathbf{x}}_p^{[i]} > \hat{\mathbf{x}}_p] \times |\hat{\mathbf{x}}_p - \hat{\mathbf{x}}_p^{[i]}| \times \hat{\gamma}_i \\ & + 1[\hat{\mathbf{x}}_p^{[i]} \leq \hat{\mathbf{x}}_p] \times |\hat{\mathbf{x}}_p - \hat{\mathbf{x}}_p^{[i]}| \times (1 - \hat{\gamma}_i), \end{aligned}$$

where M is the number of concepts, P is the dimension of the input and $\hat{\mathbf{x}}^{[i]}$ is the reconstruction after *reversing* the latent code $\hat{\gamma}_i$ of concept m_i. We give a brief interpretation for Eq. (12). Consider a concept m_i in a sample $\mathbf{x}$. If concept m_i is activated, i.e., $\hat{\gamma}_i = 1$, then $\hat{\mathbf{x}}^{[i]}$ corresponds to the *turn off* intervention $do(\gamma_i = 0)$. In this case, we expect that this intervention only removes some pixels in $\hat{x}$. Thus, we penalize the difference $|\hat{\mathbf{x}}_p - \hat{\mathbf{x}}_p^{[i]}|$ for positions p where the pixel value increases, i.e., where $\hat{\mathbf{x}}_p^{[i]} > \hat{\mathbf{x}}_p$. Finally, we combine these regularizers as

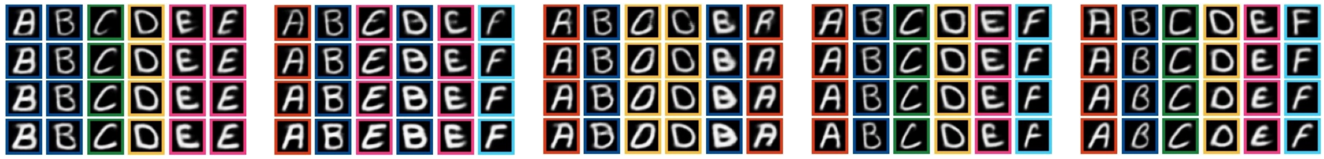
$$\mathcal{L}_{\text{R}}(\mathbf{x}) = \lambda_3 \mathcal{L}_{\text{compact}}(\mathbf{x}) + \lambda_4 \mathcal{L}_{\text{directional}}(\mathbf{x}). \tag{13}$$

Using these interpretability regularizers, we observed a significant improvement in the interpretability of discovered binary concepts.

Experiment

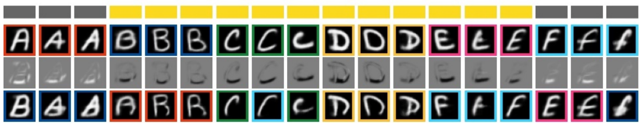
Experiment Setting

In this section, we demonstrate our method using three datasets: EMNIST(Cohen et al. 2017), MNIST(Lecun et al. 1998) and Fashion-MNIST(Xiao, Rasul, and Vollgraf 2017).

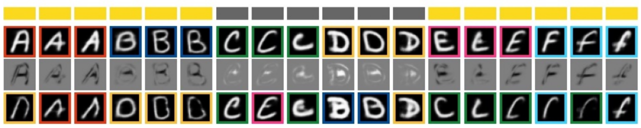


(a) Sweep $\boldsymbol{\alpha}_0$ (*bottom*) (b) Sweep $\boldsymbol{\alpha}_1$ (*middle*) (c) Sweep $\boldsymbol{\alpha}_2$ (*right*) (d) Sweep β_0 (*width*) (e) Sweep β_1 (*skewness*)

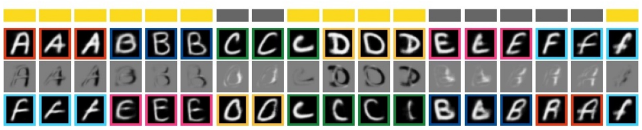
Figure 4: Visualization of the learned concept-specific and global variants. The proposed method captured the variant within each causal concept, i.e., the change of shape of (a) the bottom stroke, (b) the middle stroke and (c) the right stroke. (d, e) Our method was also able to disentangle the concepts variants with other variants that does not affect the prediction.



(a) Controlling switch γ_0 of concept m_0 (bottom stroke)

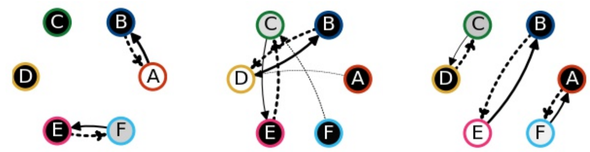


(b) Controlling switch γ_1 of concept m_1 (middle stroke)



(c) Controlling binary γ_2 of concept m_2 (right stroke)

Figure 5: The binary explanation with the intervention for each concept. (1st row) The encoded concept switch $\hat{\gamma}_i$ (yellow/gray for 1/0). (2nd row) the original reconstruction $\hat{x}$. (4th row) The reconstruction after alternating switch γ_i.



(a) bottom stroke (b) middle stroke (c) right stroke

Figure 6: The transition graph of prediction output.

For each dataset, we select several classes and train a classifier on the selected classes. In particular, we select the letters 'A, B, C, D, E, F' for EMNIST, digits '1, 4, 7, 9' for MNIST, and 't-shirt/top, dress, coat' for the Fashion-MNIST dataset. We note that our setting is more challenging than the common test setting in existing works (e.g., classifier for MNIST 3 and 8 digits) since a larger number of classes and concepts are involved in the classification task. Due to the space limit, here we mainly show the visual explanation obtained for the EMNIST dataset in which we use $M = 3$ concepts. The dimension of $\boldsymbol{\alpha}_i$ and $\boldsymbol{\beta}$ are $K = 1$ and $L = 7$, respectively. The explanation results of other datasets and further detailed experiment settings can be found in our extended version (Tran et al. 2021).

Qualitative Results

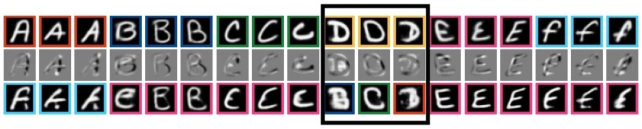
In Figure 5, we showed three discovered binary concepts for the EMNIST dataset. In each image, we show in the first row the encoded binary switch of concept m_i for different samples, in which yellow indicates $\hat{\gamma}_i = 1$ and gray indicates $\hat{\gamma}_i = 0$. The second row shows the original reconstructed image $\hat{x}$ while the fourth row shows the image reconstructed when we reverse the binary switch $\hat{x}^{[i]}$. The border color indicates the prediction result of each image. Finally, the third row show the difference of $\hat{x}$ and $\hat{x}^{[i]}$.

From Figure 5, we observed that the proposed method was able to discover useful binary concepts for explaining the classifier. First, the binary switches of these concepts have a large causal effect on the classifier output, i.e., alternating the switch affects the prediction. For example, Figure 5a explains that adding a bottom stroke to the letter A has a significant effect on the classifier output. Each concept captured a group of similar interventions and can be easily interpreted, i.e., concept m_0 represents the bottom stroke, concept m_1 represents the right stroke, and concept m_2 represents the inside (middle) stroke.

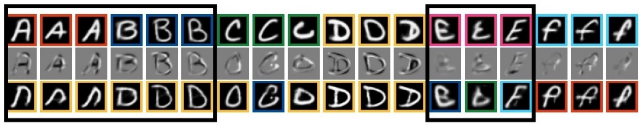
The explanation in Figure 5 can be considered as a local explanation which focus on explaining specific samples. Not only that, the proposed method also excels in providing organized knowledge about the discovered concepts and prediction classes. In particular, we can aggregate the causal effect in Figure 5 for each concept and class to assess how the each a binary switch change the prediction. The transition probability from $y = u$ to $y = v$ for a concept m_i using the do operation $do(\gamma_i = d)$ ($d \in \{0, 1\}$) can be obtained as

$$\begin{aligned} w_{u,v}^{do(\gamma_i=d)} &= \Pr[y = v \mid y = u, do(\gamma_i = d)] \\ &= \frac{1}{|X_u|} \sum_{\mathbf{x} \in X_u} 1[f(\hat{\mathbf{x}}^{do(\gamma_i=d)}) = v] \end{aligned} \tag{14}$$

where $X_u = \{\mathbf{x} \in X \mid f(\hat{\mathbf{x}}) = u\}$. In Figure 6, we show the calculated transition probabilities for each concept as a graph in which each note represents a prediction class. A solid arrow (dashed arrow) represents the transition when activating (deactivating) a concept and the arrow thickness shows the transition probability $w_{u,v}^{do(\gamma_i=1)}$ ($w_{u,v}^{do(\gamma_i=0)}$). We neglect the transition which transition probability is less than

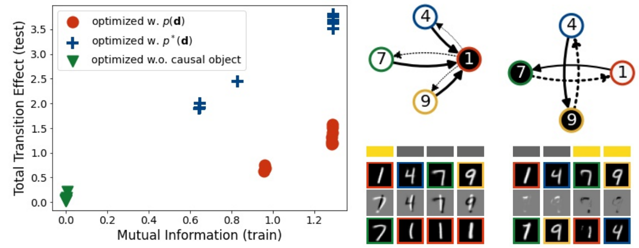


(a) Intervention by decreasing α_0



(b) Intervention by increasing α_0

Figure 7: A causal factor by O'Shaughnessy et al. (2020). Low interpretability results are framed (More details in text)



(a) $p(\boldsymbol{\gamma})$ vs $p^*(\boldsymbol{\gamma})$ vs *no causal* (b) $\lambda_{\mathrm{R}} = 0$ (c) $\lambda_{\mathrm{R}} = 1$

Figure 8: (a) (MNIST) Train-time MI and test-time TTE of ten runs when $\mathcal{L}_{\mathrm{CE}}$ is based on $p(\boldsymbol{\gamma})$ (red), $p^*(\boldsymbol{\gamma})$ (blue), and when trained without $\mathcal{L}_{\mathrm{CE}}$ (green). (b) Discovered binary concepts when trained with and without $\mathcal{L}_{\mathrm{R}}$.

0.1 For example, from Figure 6a, one can interpret that the bottom stroke is important to distinguish (E,F) and (A,B).

Finally, in Figure 4 (a,b,c), we show the captured variants within each concept and other global variants that have a small effect on the classifier output. In contrast to binary switches, these variants explain what does not change the prediction. We first activate the concept m_i using the do-operation do$(\gamma_i = 1)$, then plot the reconstruction while alternating $\boldsymbol{\alpha}_i$. We observed that $\boldsymbol{\alpha}_0$ captured the length of the bottom stroke, $\boldsymbol{\alpha}_1$ captured the shape of the right stroke, and $\boldsymbol{\alpha}_2$ captured the length of the inside (middle) stroke, respectively. Especially, our method was also able to differentiate the concept-specific variants from other global variants $\boldsymbol{\beta}$ such as skewness, height, or width (Figure 4 d,e).

Comparing with Other Methods.

We compare our method to other baselines shown in Figure 2. First, saliency-map-based methods, which use a saliency map to quantify the importance of (super)pixels, although it is easy to understand, do not explain why highlighted (super)pixels are important (Figure 2a). Because they only provide one explanation for each input, they can not explain how these pixels distinguish the predicted class from other classes. Our method, with multiple concepts, can perform different interventions to obtain multiple explanations.

Next, we compare to O'Shaughnessy et al. (2020), in which we used a VAE model with ten continuous factors and encouraged three factors to have causal effects on predicted classes. In Figure 7, we visualize α_0 which achieved the largest causal effect. In Figure 7a (7b), we decrease (increase) α_0 until the its prediction label changes and show that intervention result in the third row. First, we observed that it failed to disentangle different causal factors as α_0 affects all the bottom, middle and right strokes. For example, in Figure 7a, decreasing α_t changed the letter D in the 10th column to letter B (*middle stroke* concept), while changed the letter D in the 11th column to letter C (*left stroke* concept). A similar result is also observed in Figure 7b for letter E. Second, it failed to disentangle the concept-specific variant, which does not affect the prediction. For example, for the letter A and B (1st to 6th column) in Figure 7b, increasing α_0 does not only affect the occurrence of the *middle stroke*, but also changes the shape of the *right stroke*.

Our method overcomes these limitations with a carefully designed binary-discrete structure coupled with the proposed causal effect and interpretability regularizer. By encouraging the causal influence of only the binary switches, our method can disentangle what affects the prediction and the variant of samples with the same prediction. Thus, it encourages that a binary switch m_i only changes the prediction from a class y_k to only one other class $y_{k'}$, resulting in a more interpretable explanation. We also emphasize that the binary-continuous composite structure alone is not enough to obtain valuable concepts for explanation (Figure 2b).

Quantitative Results

We evaluate the causal influence of a concept set using the total transition effect (TTE), which is defined as

$$\mathrm{TTE} = \frac{1}{M} \sum_{i \in [M]} \sum_{u,v \in [T]} [w_{u,v}^{do(\gamma_i=1)} + w_{u,v}^{do(\gamma_i=0)}]. \qquad (15)$$

where M and T are the number of concepts and classes, respectively. Here, a large value of TTE indicates a significant overall causal effect by the whole discovered concept set on all class transitions. Compared to information flow, TTE can evaluate more directly and faithfully the causal effect of binary switches on dataset X. Moreover, it is also easier for end-user to understand.

In Figure 8a, we show the test-time mutual information and the TTE values when the causal objective $\mathcal{L}_{\mathrm{CE}}$ uses the prior $p^*(\boldsymbol{\gamma})$ (Eq. (9)), VAE model's prior $p(\boldsymbol{\gamma})$ and when trained without $\mathcal{L}_{\mathrm{CE}}$. The interpretability regularizers are included in all settings. We observed that when $p(\boldsymbol{\gamma})$ is used, there are cases where the estimated mutual information is high, but the total transition effect is small. On the other hand, the mutual information obtained with estimated $p^*(\boldsymbol{\gamma})$ aligns better with the TTE value. We claim that this is because of the deviation between $p(\boldsymbol{\gamma})$ and $p^*(\boldsymbol{\gamma})$. By estimating $p^*(\boldsymbol{\gamma})$ on the run, our method can better evaluate and optimize the causal influence of $\boldsymbol{\gamma}$ on y. Moreover, we fail to discover useful concepts without the causal objective.

Next, we evaluate how $\mathcal{L}_{\mathrm{R}}$ helps to increase the interpretability of concepts. In Figure 8b, we show an example of concepts when train without $\mathcal{L}_{\mathrm{R}}$. This concept (top) only replaces the digits 4, 7, and 9 with the digit 1. Although this concept has a large causal effect, it does not provide any proper explanation why the image is identified as 1. With the interpretability regularizers, we can discover binary concepts with high interpretability that adequately explain that digit seven can be distinguished from digit one based on the existence of the top stroke (Figure 8c). In principle, our method can be applied to other data domains if one can train a generative model. However, obtaining interpretable concepts can be more challenging for more complicated domains. In the future, we will explore more challenging tasks, e.g., medical image classification or domains such as text or table data.

Conclusion

We introduced the problem of discovering binary concepts for explaining a black-box classifier. We first proposed a structural generative model that can properly express binary concepts. Then, we proposed a learning process that simultaneously learns the data distribution and encourages the binary switches to have a large causal effect on the classifier output. The proposed method also allows integrating users' preferences and prior knowledge for better interpretability and consistency. We demonstrated that the proposed method could discover interpretable binary concepts with a large causal effect which can effectively explain the classification model for multiple datasets.

Acknowledgements

This work is partly supported by Japan science and technology agency (JST), CREST JPMJCR21D3, Japan society for the promotion of science (JSPS), Grants-in-Aid for Scientific Research 19H04164 and 18H04099, and JST SPRING (Support for Pioneering Research Initiated by the Next Generation) JPMJSP2124.

References

Ay, N.; and Polani, D. 2008. INFORMATION FLOWS IN CAUSAL NETWORKS. *Advs. Complex Syst.*, 11(01): 17–41.

Burgess, C. P.; Higgins, I.; Pal, A.; Matthey, L.; Watters, N.; Desjardins, G.; and Lerchner, A. 2018. Understanding disentangling in β-VAE. *arXiv:1804.03599*.

Choi, J.; Hwang, G.; and Kang, M. 2020. Discond-VAE: Disentangling Continuous Factors from the Discrete. *arXiv:2009.08039*.

Cohen, G.; Afshar, S.; Tapson, J.; and Van Schaik, A. 2017. EMNIST: Extending MNIST to handwritten letters. In *2017 International Joint Conference on Neural Networks (IJCNN)*, 2921–2926. IEEE.

Dhurandhar, A.; Chen, P.-Y.; Luss, R.; Tu, C.-C.; Ting, P.; Shanmugam, K.; and Das, P. 2018. Explanations based on the Missing: Towards Contrastive Explanations with Pertinent Negatives. *arXiv:1802.07623*.

Gardenfors, P. 2014. *The Geometry of Meaning: Semantics Based on Conceptual Spaces*. MIT Press.

George, E. I.; and McCulloch, R. E. 1997. Approaches for Bayesian variable selection. *Statistica sinica*, 339–373.

Ghorbani, A.; Wexler, J.; Zou, J. Y.; and Kim, B. 2019. Towards Automatic Concept-based Explanations. *Advances in Neural Information Processing Systems*, 32: 9277–9286.

Goyal, Y.; Feder, A.; Shalit, U.; and Kim, B. 2020. Explaining Classifiers with Causal Concept Effect (CaCE). *arXiv:1907.07165*.

Gupta, K.; Singh, S.; and Shrivastava, A. 2020. PatchVAE: Learning Local Latent Codes for Recognition. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 4745–4754. Seattle, WA, USA: IEEE.

Gyawali, P.; Li, Z.; Knight, C.; Ghimire, S.; Horacek, B. M.; Sapp, J.; and Wang, L. 2019. Improving disentangled representation learning with the beta bernoulli process. In *2019 IEEE International Conference on Data Mining (ICDM)*, 1078–1083. IEEE.

Higgins, I.; Matthey, L.; Pal, A.; Burgess, C.; Glorot, X.; Botvinick, M.; Mohamed, S.; and Lerchner, A. 2017. beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework. ICLR.

Holland, P. W. 1988. Causal Inference, Path Analysis, and Recursive Structural Equations Models. *Sociol. Methodol.*, 18: 449–484.

Joshi, S.; Koyejo, O.; Vijitbenjaronk, W.; Kim, B.; and Ghosh, J. 2019. Towards Realistic Individual Recourse and Actionable Explanations in Black-Box Decision Making Systems. *arXiv:1907.09615*.

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.

Kim, C.; and Bastani, O. 2019. Learning Interpretable Models with Causal Guarantees. *arXiv:1901.08576*.

Kingma, D. P.; and Welling, M. 2014. Auto-encoding variational bayes. In *2nd International Conference on Learning Representations, ICLR 2014 - Conference Track Proceedings*.

Kocaoglu, M.; Snyder, C.; Dimakis, A. G.; and Vishwanath, S. 2017. CausalGAN: Learning Causal Implicit Generative Models with Adversarial Training. *arXiv:1709.02023*.

Lecun, Y.; Bottou, L.; Bengio, Y.; and Haffner, P. 1998. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11): 2278–2324.

Lewontin, R. C. 1974. The Analysis of Variance and the Analysis of Causes. *Am. J. Hum. Genet.*, 26(3): 400–411.

Locatello, F.; Bauer, S.; Lucic, M.; Raetsch, G.; Gelly, S.; Schölkopf, B.; and Bachem, O. 2019. Challenging common assumptions in the unsupervised learning of disentangled representations. In *international conference on machine learning*, 4114–4124. PMLR.

Maddison, C.; Mnih, A.; and Teh, Y. 2017. The concrete distribution: A continuous relaxation of discrete random variables. In *Proceedings of the international conference* *on learning Representations*. International Conference on Learning Representations.

Narendra, T.; Sankaran, A.; Vijaykeerthy, D.; and Mani, S. 2018. Explaining Deep Learning Models using Causal Inference. *arXiv:1811.04376*.

O'Shaughnessy, M.; Canal, G.; Connor, M.; Davenport, M.; and Rozell, C. 2020. Generative causal explanations of black-box classifiers. *Advances in Neural Information Processing Systems*.

Parafita, Á.; and Vitrià, J. 2019. Explaining visual models by causal attribution. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, 4167–4175. IEEE.

Pearl, J. 2009. *Causality*. Cambridge University Press.

Ross, A. S.; and Doshi-Velez, F. 2021. Benchmarks, algorithms, and metrics for hierarchical disentanglement. *arXiv:2102.05185*.

Schwab, P.; and Karlen, W. 2019. CXPlain: Causal Explanations for Model Interpretation under Uncertainty. *Advances in Neural Information Processing Systems*, 32: 10220–10230.

Selvaraju, R. R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; and Batra, D. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, 618–626.

Springenberg, J. T.; Dosovitskiy, A.; Brox, T.; and Riedmiller, M. 2014. Striving for Simplicity: The All Convolutional Net. *arXiv:1412.6806*.

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

Tonolini, F.; Jensen, B. S.; and Murray-Smith, R. 2020. Variational Sparse Coding. In Adams, R. P.; and Gogate, V., eds., *Proceedings of The 35th Uncertainty in Artificial Intelligence Conference*, volume 115 of *Proceedings of Machine Learning Research*, 690–700. PMLR.

Tran, T. Q.; Fukuchi, K.; Akimoto, Y.; and Sakuma, J. 2021. Unsupervised Causal Binary Concepts Discovery with VAE for Black-box Model Explanation. *arXiv:2109.04518*.

Xiao, H.; Rasul, K.; and Vollgraf, R. 2017. Fashion-MNIST: a Novel Image Dataset for Benchmarking Machine Learning Algorithms. *arXiv:1708.07747*.

Yang, M.; Liu, F.; Chen, Z.; Shen, X.; Hao, J.; and Wang, J. 2020. CausalVAE: Disentangled Representation Learning via Neural Structural Causal Models. *arXiv:2004.08697*.

Yeh, C.-K.; Kim, B.; Arik, S.; Li, C.-L.; Pfister, T.; and Ravikumar, P. 2020. On Completeness-aware Concept-Based Explanations in Deep Neural Networks. *Advances in Neural Information Processing Systems*, 33.