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

Thien Q. Tran

University of Tsukuba, Riken AIP
thientquang@mdl.cs.tsukuba.ac.jp

Kazuto Fukuchi

University of Tsukuba, Riken AIP
fukuchi@mdl.cs.tsukuba.ac.jp

Youhei Akimoto

University of Tsukuba, Riken AIP
akimoto@mdl.cs.tsukuba.ac.jp

Jun Sakuma

University of Tsukuba, Riken AIP
jun@mdl.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 the 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. Using multiple datasets, we demonstrate that our method can discover useful binary concepts for explanation.

1 Introduction

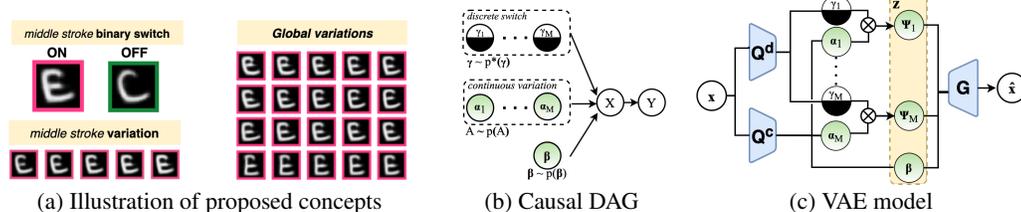


Figure 1: (a) Binary concept *middle stroke* and some global variants. Border color indicates the classifier output. (b, c) The proposed VAE model and the causal DAG

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 interpretable model 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 an explanation that uses the symbolic expression: ‘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, our 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 relied 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 [6]. Thus, an explanation

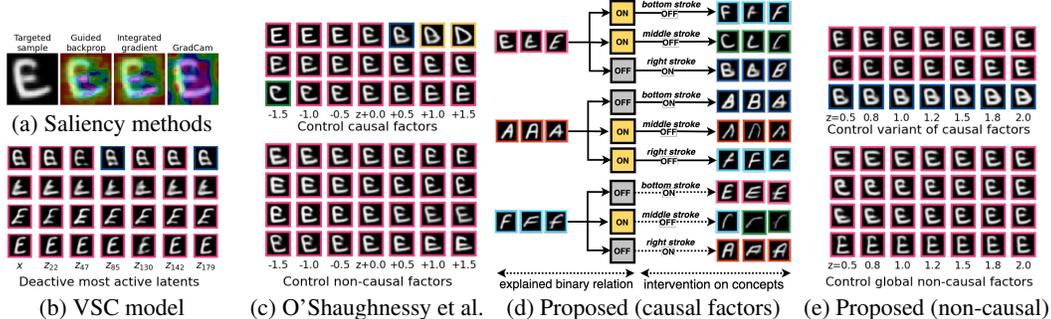


Figure 2: Explanation methods for a letter classifier. (a) Saliency-based methods. (b) Disabling the most active latents of class E in VSC model. (c) Controlling the causal and non-causal factors in O’Shaughnessy et al. (d, e) Proposed method: (d) Encoded binary relation of discovered concepts and their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

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.

Our method employs three different notions in the explanation: *causal binary switches*, *concept-specific variants* and *global variants*. We illustrate these notions in Figure 1a. 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. That is, we avoid finding confounding concepts, which correlate with but do not cause the prediction. For example, the *sky* concept appears 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 lack of interpretability.

In Figure 2d and 2e, we demonstrate an explanation discovered by the proposed method for an 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 in Section 5.

To the best of our knowledge, no existing method can discover binary concepts that fulfill all of these requirements. Saliency methods such as Guided Backprop [30], Integrated Gradient [31] or GradCam [29] 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 [32], IBP-VAE [11], PatchVAE [10], 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. [24] 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 on how this latent variable affects prediction.

Our work has the following contributions: (1) 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. (2) We propose a learning process to simultaneously learn the data distribution while encouraging the causal influence of the binary switches. Although typically VAE models encourage the independence of factors for meaningful disentanglement, such an assumption is inadequate for discovering useful causal concepts which are often mutually correlated. Our learning process, which considers the dependence between binary concepts, can discover concepts with more significant causality. (3) To avoid the concepts that have causality but no interpretability, the proposed method allows an easy way to implement user’s preference and prior knowledge as a regularizer to induce high interpretability of concepts. (4) Finally, we demonstrate that our method succeeds in discovering interpretable binary concepts with causality that are useful for explanation with multiple datasets.

2 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 [15, 8], a prototypical activation vector [35] or a latent factor of a generative model [24, 9]. 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. When this side-information is not given, one needs to discover useful concepts for the explanation, e.g., Ghorbani et al. [8] used segmentation and clustering, Yeh et al. [35] retrained the classifier with a prototypical concept layer, O’Shaughnessy et al. [24] 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 representation \mathbf{z} that captures different aspects of the data. However, Locatello et al. [21] 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 [12, 2]. Another approach is to incorporate structure into the representation [3, 27, 32, 10]. 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 [28] employed Granger causality to quantify the causal effect of input features, Parafita and Vitrià [25] evaluated the causality of latent attributions with a prior known causal structure, Narendra et al. [23] evaluated the causal effect of network layers, and Kim and Bastani [16] learned an interpretable model with a causal guarantee. Some works first train a generative model and then search for counterfactual samples on latent space [14, 5]. Although these methods can provide a counterfactual explanation for each input sample, the generative model is trained individually and does not necessarily disentangle useful concepts. Some works introduce the causal structure into generative models such as CausalVAE [34] or CausalGan [18]. These methods are not applicable in our setting because they require additional knowledge such as causal graphs or concept labels. To the best of our knowledge, no existing works can explain using binary concepts that fulfill the three requirements we discussed.

3 Preliminaries

3.1 Variational Autoencoder

Our explanation is build upon the VAE framework proposed by Kingma and Welling [17]. 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 $p(\mathbf{x} | \mathbf{z})$. Typically, due to the intractability, a

variational approximation $q(\mathbf{z} | \mathbf{x})$ of the intractable posterior is introduced and the model is then learned using the evidence lower bound (ELBO) as

$$\mathcal{L}_{\text{VAE}}(\mathbf{x}) = -\mathbb{E}_{\mathbf{z} \sim q(\mathbf{z} | \mathbf{x})}[\log p(\mathbf{x} | \mathbf{z})] + \mathbb{KL}[q(\mathbf{z} | \mathbf{x}) \parallel p(\mathbf{z})]. \quad (1)$$

Here, $q(\mathbf{z} | \mathbf{x})$ is the encoder that maps the data to the latent space and $p(\mathbf{x} | \mathbf{z})$ is the decoder that maps the latents to the data space. Commonly, $q(\mathbf{z} | \mathbf{x})$ and $p(\mathbf{x} | \mathbf{z})$ are parameterized as neural networks $Q(\mathbf{z} | \mathbf{x})$ and $G(\mathbf{x} | \mathbf{z})$, respectively. The common choice for $q(\mathbf{z} | \mathbf{x})$ is a factorized Gaussian encoder $q(\mathbf{z} | \mathbf{x}) = \prod_{p=1}^P \mathcal{N}(\mu_i, \sigma_i^2)$ where $(\mu_1, \dots, \mu_P, \sigma_1, \dots, \sigma_P) = Q(\mathbf{x})$. 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. Then, the first term can be trained using L2 reconstruction loss, while the KL-divergence terms are trained using the reparameterization trick.

3.2 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 [26]. 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 $G[1]$). *Let U and V be disjoint subsets of nodes. The information flow $I(U \rightarrow V)$ from U to V is defined by*

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

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

O’Shaughnessy et al. [24] argued that compared to other metrics such as average causal effect (ACE) [13], analysis of variance (ANOVA) [20], information flow is more suitable to capture complex and nonlinear causal dependence between variables.

4 Proposed method

We aim to discover a set of binary concepts $\mathcal{M} = \{m_0, m_1, \dots, m_M\}$ with causality and interpretability that can explain the black-box classifier $f : \mathcal{X} \rightarrow \mathcal{Y}$. Inspired by O’Shaughnessy et al. [24], we employ 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 1b, 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* $\{\gamma_i, \alpha_i\}_{i=1}^M$, and a *global variants* β . As we want to explain the classifier output (i.e., node y in Figure 1b) 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}_{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:

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

4.1 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 ψ_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 α_i (concept-specific variant for short) as $\psi_i = \gamma_i \cdot \alpha_i$ where $\gamma_i = 1$ if concept m_i is on and $\gamma_i = 0$ otherwise. Here, the *concept switch* γ_i controls if the concept m_i is activated in a sample, e.g., if the bottom stroke is appeared in a image (Figure 2d). On the other hand, the *concept-specific variant* α_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* $\{\alpha_i\}$ whose effect is limited to a specific binary

concept, we also allow a *global variant* latent β to capture other variants that do not necessarily have causality, e.g., skewness (Figure 2e, bottom). Here, disentangling concept-specific and global variants is important for assisting user 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 [7] and sparse coding [32]. Unlike these models, whose number of discrete-continuous factors is often large, our model uses only a small number of binary concepts with a multi-dimensional global variants β . Our intuition is that in many cases, the classification can be made by combining a small number of binary concepts.

Input encoding. For the discrete components, we use a network $Q^d(\mathbf{x})$ to parameterize $q(\gamma | \mathbf{x})$ as $q(\gamma | \mathbf{x}) = \prod_{i=1}^M q(\gamma_i | \mathbf{x}) = \prod_{i=1}^M \text{Bern}(\gamma_i; \pi_i)$ where $(\pi_1, \dots, \pi_M) = Q^d(\mathbf{x})$. For the continuous components, letting $A = (\alpha_1, \alpha_2, \dots, \alpha_M)$, we use a network $Q^c(\mathbf{x})$ to parameterize $q(A, \beta | \mathbf{x})$ as $q(A, \beta | \mathbf{x}) = \left[\prod_{i=1}^M q(\alpha_i | \mathbf{x}) \right] q(\beta | \mathbf{x})$. Here, $q(\alpha_i | \mathbf{x}) = \mathcal{N}_\delta^{\text{fold}}(\alpha_i; \mu_i, \text{diag}(\sigma_i))$, $q(\beta | \mathbf{x}) = \mathcal{N}_\delta^{\text{fold}}(\beta; \mu_\beta, \text{diag}(\sigma_\beta))$ and $(\mu_1, \dots, \mu_M, \mu_\beta, \sigma_1, \dots, \sigma_M, \sigma_\beta) = Q^c(\mathbf{x})$. 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. The implementation of $\mathcal{N}_\delta^{\text{fold}}(\mu, \sigma^2)$ can simply be done by adding the absolute and shift operation to the conventional implementation of $\mathcal{N}(\mu, \sigma^2)$. We discuss in detail this design choice in Appendix A.3.

Output decoding. Next, given $q(\gamma | \mathbf{x})$ and $q(A, \beta | \mathbf{x})$, we first sample the concept switches $\{\hat{d}_i\}$, the concept variants $\{\hat{\alpha}_i\}$ and the global variants $\hat{\beta}$ from their posterior, respectively. Using these sampled latents, we construct an aggregated representation $\hat{\mathbf{z}} = (\psi_1, \dots, \psi_M, \hat{\beta})$ using the binary concept mechanism in which ψ_i is the corresponding part for concept m_i , i.e., $\psi_i = \gamma_i \times \alpha_i$. If concept m_i is on, we let $\hat{d}_i = 1$ so that ψ_i can reflect the concept-specific variant $\hat{\alpha}_i$. 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 $\hat{\mathbf{z}}$ and generate the reconstruction \hat{x} as $\hat{x} \sim G(\mathbf{x} | \hat{\mathbf{z}})$ where $\hat{\mathbf{z}} = (\psi_1, \dots, \psi_M, \hat{\beta})$.

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:

$$\begin{aligned} \mathcal{L}_{\text{VAE}}(\mathbf{x}) = & -\mathbb{E}_{\hat{\mathbf{z}} \sim Q^{\{c,d\}}(\mathbf{z}|\mathbf{x})} [\log G(\mathbf{x} | \hat{\mathbf{z}})] + \lambda_2 \left[\frac{1}{M} \sum_{i=1}^M \mathbb{KL}(q(\gamma_i | \mathbf{x}) \parallel \text{Bern}(\pi_i)) \right] \\ & + \lambda_1 \left[\mathbb{KL}(q(\beta | \mathbf{x}) \parallel \mathcal{N}_\delta^{\text{fold}}(0, \mathcal{I})) + \frac{1}{M} \sum_{i=1}^M \mathbb{KL}(q(\alpha_i | \mathbf{x}) \parallel \mathcal{N}_\delta^{\text{fold}}(0, \mathcal{I})) \right]. \end{aligned} \quad (4)$$

For the Bernoulli distribution, we use its continuous approximation, i.e., the relaxed-Bernoulli [22] in the training process.

4.2 Encouraging causal effect of binary switches

We expect the binary switches γ to have a large causal influence so that they can effectively explain the classifier. To measure the causal effect of γ on the classifier output Y , we employ the causal DAG in Figure 1b and adopt *information flow* (Definition 1) as the causal measurement. Our DAG employs an assumption that is fundamentally different from those of standard VAE models. Specifically, the standard VAE model and also O’Shaughnessy et al. [24] 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 in estimating information flow might lead to a large estimation error and prevent discovering valuable causal concepts.

Instead, we employ a prior distribution $p^*(\gamma)$ 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 $p^*(\gamma)$ which most likely generates the user’s dataset. This empirical estimation of $p^*(\gamma)$ is then used to evaluate the causal objective in Eq. (3). Assuming X is a set of i.i.d samples from data distribution $p(\mathbf{x})$, we estimate $p^*(\gamma)$ as

$$p^*(\gamma) \approx \int_{\mathbf{x}} p^*(\gamma | \mathbf{x})p(\mathbf{x})d\mathbf{x} \approx \frac{1}{|X|} \sum_{\mathbf{x} \in X} p(\gamma | \mathbf{x}) \approx \frac{1}{|X|} \sum_{\mathbf{x} \in X} \prod_{i=1}^M q(\gamma_i | \mathbf{x}) \quad (5)$$

In the last line, $p(\gamma | \mathbf{x})$ is replaced with the variational posterior $q(\gamma | \mathbf{x})$ of VAE model. Here, the factorized posterior $q(\gamma | \mathbf{x})$ only assumes the independence between latents conditioned on a sample but does not imply the independence of binary switches in $p^*(\gamma)$. We note that we do not aim to learn the dependence between concepts but only expect that $p^*(\gamma)$ properly reflects the dependence between binary concepts that appears in the dataset X for a better evaluation of causal effect. We experimentally show in Subsection 5.4 that using the estimation of $p^*(\gamma)$ results in a better estimation for the causal effect on dataset X and more valuable concepts for the explanation. We showed that in the proposed DAG, information flow $I(\gamma \rightarrow Y)$ coincides with mutual information $I(\gamma; Y)$.

Proposition 1 (Coincident of Information Flow and Mutual Information in proposed DAG). *The information flow from γ to Y in the DAG of Figure 1b coincides with the mutual information between γ and Y . That is, $I(\gamma \rightarrow Y) = I(\gamma; Y) = \mathbb{E}_{\gamma, Y} \left[\frac{p^*(\gamma)p(Y|\gamma)}{p^*(\gamma)p(Y)} \right]$.*

We prove Proposition 1 in Appendix A.2. The detailed algorithm for estimating $I(\gamma; Y)$ is described in Appendix A.1. As we want to maximize $I(\gamma; Y)$, we rewrite it as a loss term $\mathcal{L}_{CE} = -I(\gamma; Y)$ and optimize it together with the learning of VAE model.

4.3 Integrating user preference for concepts

Finally, we discuss the integration of user’s 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, has a large causal effect on the prediction. However, such a concept does not provide valuable knowledge and is hard to interpret. To avoid such concepts, we allow the user to implement their preference or prior knowledge as an interpretability regularizer to constrain the generative model’s expressive power. The proposed method then seeks for 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. (4)). 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 subtract the pixel value, but not both at the same time (Eq. (5)). Furthermore, we desire that $\gamma_i = 1$ indicates the *presence* of pixels and $\gamma_i = 0$ indicates the *absence* of pixels. We show the detailed formulation for these regularizers in Appendix A.4. Using these interpretability regularizer, we observed a significant improvement in interpretability (Subsection 5.4) and consistency (Appendix A.4) of concepts.

5 Experiment

5.1 Experiment setting

We demonstrate our method using three datasets: EMNIST[4], MNIST[19] and Fashion-MNIST[33]. 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

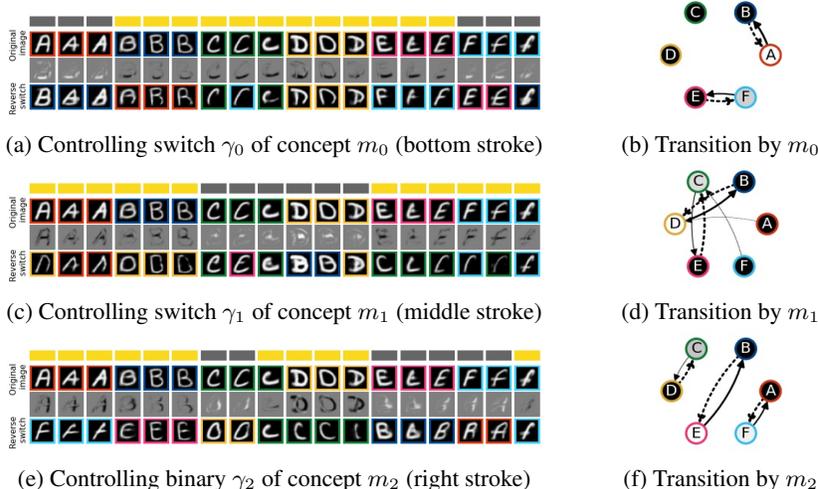


Figure 3: (a, c, e) The binary explanation with the intervention for each concept. (1st row) The encoded concept switch $\hat{\gamma}_i$ (yellow/gray for 1/0). (2nd row) the original reconstruction \hat{x} . (4th row) The reconstruction after alternating switch γ_i . (b, d, f) The transition graph of prediction output.

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 α_i and β are $K = 1$ and $L = 7$, respectively. The explanation results of other datasets and further detailed experiment settings can be found in Appendix A.5, A.6 and A.7.

5.2 Qualitative results

In Figure 3 (a, c and e), 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 $\hat{x}^{[i]}$ and \hat{x} .

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 3a explains that adding a bottom stroke to letter A has a significant effect on the classifier output. Not only that, 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 middle stroke.

The explanation in Figure 3 (a, c and e) 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 of these local explanation for each concept and class to assess how the each a binary switch change the prediction. Letting $X_u = \{\mathbf{x} \in X \mid f(\hat{\mathbf{x}}) = u\}$, 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 $w_{u,v}^{do(\gamma_i=d)} = \Pr[y = v \mid y = u, do(\gamma_i = d)] = \frac{1}{|X_u|} \sum_{\mathbf{x} \in X_u} \mathbb{1}[f(\hat{\mathbf{x}}^{do(\gamma_i=d)}) = v]$.

In 3 (b, d and f), we show the calculated transition probabilities for each concept as a graph in which each node 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 0.1 For example, from Figure 3b, one can interpret that the bottom stroke is important to distinguish (E,F) and (A,B).

Finally, in Figure 4 (a, b and c), we show the captured variants within each concept and other global variants, that have a small affect on the classifier output. In contrast to binary switches, these variants

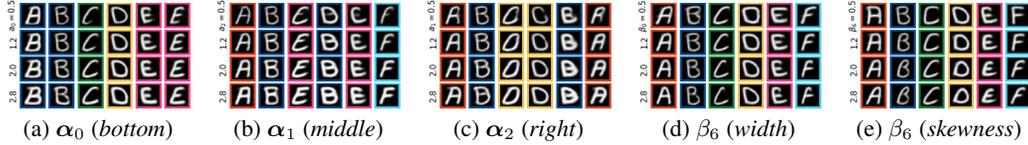


Figure 4: Visualization of the learned concept-specific and global variants. The proposed method captured the variant within each causal concept, i.e., the change of shape of (a) the bottom stroke, (b) the middle stroke and (c) the right stroke. (d, e) Our method was also able to disentangle the concepts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

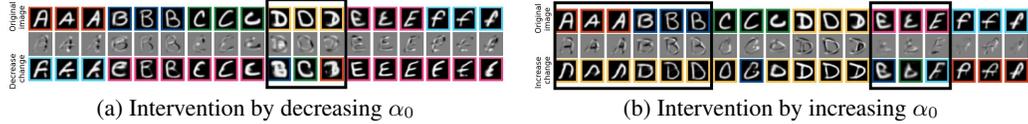


Figure 5: A causal factor by [24]. Low interpretability results are framed (More details in text)

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 α_i . We observed that α_0 captured the length of the bottom stroke, α_1 captured the shape of the right stroke, and α_2 captured the length of the inside (middle) stroke, respectively. Especially, our method was also able to differentiate the concept-specific variants with other global variants β such as skewness or width (Figure 4 d, e).

5.3 Comparing with other methods.

We compare our method to other baselines in Figure 2. First, saliency-map-based methods, which use a saliency map to quantify the importance of (super)pixels, although 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 others classes. Our method, can provide multiple explanations by intervening difference concepts.

Next, we compare to O’Shaughnessy et al. [24], in which we used a VAE model with ten continuous factors and encouraged three factors to have causal effects on predicted classes. In Figure 5, we visualize α_0 which achieved the largest causal effect. In Figure 5a (5b), 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 5a, 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 5b 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 5b, 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 mixed structure alone is not enough to obtain valuable concepts for explanation (Figure 2b).

5.4 Quantitative results

We evaluate the causal influence of a concept set using the total transition effect (TTE) which is defined as $TTE = \frac{1}{M} \sum_{i \in [M]} \sum_{u,v \in [T]} [w_{u,v}^{do(\gamma_i=1)} + w_{u,v}^{do(\gamma_i=0)}]$ 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 more easy for end-user to understand.

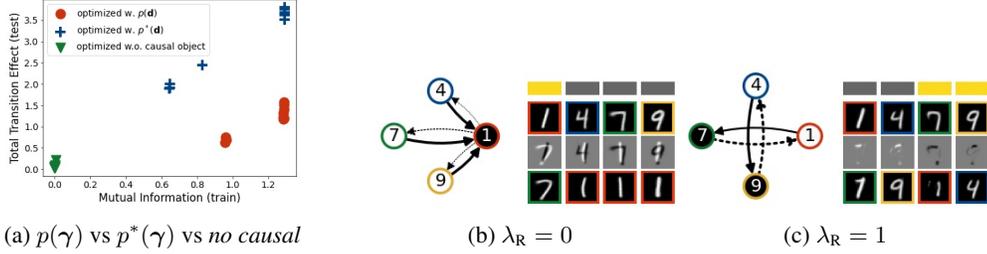


Figure 6: (a) (MNIST) Train-time MI and test-time TTE of ten runs when \mathcal{L}_{CE} is based on $p(\gamma)$ (red), $p^*(\gamma)$ (blue), and when trained without \mathcal{L}_{CE} (green). (b) Discovered binary concepts and their transition graph when trained with and without \mathcal{L}_R .

In Figure 6a, we show the test-time mutual information and the TTE values when the causal objective \mathcal{L}_{CE} uses the prior $p^*(\gamma)$ (Eq. (5)), VAE model’s prior $p(\gamma)$ and when trained without \mathcal{L}_{CE} . The interpretability regularizers are included in all settings. We observed that when $p(\gamma)$ 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 $p^*(\gamma)$ aligns better with the TTE value. We claim that this is because of the deviation between $p(\gamma)$ and the ‘true’ $p^*(\gamma)$. By estimating $p^*(\gamma)$ on the run, our method can better evaluate and optimize the causal influence of γ on y . Moreover, we also observed that without the causal objective, we failed to discover causal binary concepts.

Next, we evaluate how implementing user’s preferences and prior knowledge via \mathcal{L}_R increases the interpretability of concepts. In Figure 6b, we show an example of concepts discovered when we train the model without the interpretability regularizer. We see that alternating the binary switch of this concept (top) only replaces the digit 4, 7, 9 by the digit 1 but does not provide any proper explanation why the image is identified as 1. Although this concept has a large causal effect, it barely offers valuable knowledge. Our method, using the interpretability regularizers, can discover binary concepts with high interpretability that adequately explain that digit 7 can be distinguished from digit 1 based on the existence of the top stroke (Figure 6c). In principle, the proposed method can be applied to other data domains if one can train a generative model on that domain. However, obtaining interpretable concepts can be more challenging for a more complicated domain. As future work, we plan to explore more challenging tasks, e.g., medical image classification and other domains such as text or table data.

6 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 user’s 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.

Acknowledgments and Disclosure of Funding

This work was partly supported by KAKENHI (Grants-in-Aid for Scientific Research) Grant Numbers JP19H04164 and JST SPRING (Support for Pioneering Research Initiated by the Next Generation), Grant Numbers JPMJSP2124.

References

- [1] Nihat Ay and Daniel Polani. INFORMATION FLOWS IN CAUSAL NETWORKS. *Adv. Complex Syst.*, 11(01):17–41, February 2008.
- [2] Christopher P Burgess, Irina Higgins, Arka Pal, Loic Matthey, Nick Watters, Guillaume Desjardins, and Alexander Lerchner. Understanding disentangling in β -VAE. April 2018.

- [3] Jaewoong Choi, Geonho Hwang, and Myungjoo Kang. Discond-VAE: Disentangling continuous factors from the discrete. September 2020.
- [4] Gregory Cohen, Saeed Afshar, Jonathan Tapson, and Andre Van Schaik. Emnist: Extending mnist to handwritten letters. In *2017 International Joint Conference on Neural Networks (IJCNN)*, pages 2921–2926. IEEE, 2017.
- [5] Amit Dhurandhar, Pin-Yu Chen, Ronny Luss, Chun-Chen Tu, Paishun Ting, Karthikeyan Shanmugam, and Payel Das. Explanations based on the missing: Towards contrastive explanations with pertinent negatives. February 2018.
- [6] Peter Gardenfors. *The Geometry of Meaning: Semantics Based on Conceptual Spaces*. MIT Press, January 2014.
- [7] Edward I George and Robert E McCulloch. Approaches for bayesian variable selection. *Statistica sinica*, pages 339–373, 1997.
- [8] Amirata Ghorbani, James Wexler, James Y Zou, and Been Kim. Towards automatic concept-based explanations. *Advances in Neural Information Processing Systems*, 32:9277–9286, 2019.
- [9] Yash Goyal, Amir Feder, Uri Shalit, and Been Kim. Explaining classifiers with causal concept effect (CaCE). *arXiv:1907.07165 [cs, stat]*, February 2020.
- [10] Kamal Gupta, Saurabh Singh, and Abhinav Shrivastava. PatchVAE: Learning local latent codes for recognition. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4745–4754, Seattle, WA, USA, 2020. IEEE.
- [11] Prashna Gyawali, Zhiyuan Li, Cameron Knight, Sandesh Ghimire, B Milan Horacek, John Sapp, and Linwei Wang. Improving disentangled representation learning with the beta bernoulli process. pages 1078–1083, 2019.
- [12] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-VAE: Learning basic visual concepts with a constrained variational framework. November 2016.
- [13] Paul W Holland. Causal inference, path analysis, and recursive structural equations models. *Sociol. Methodol.*, 18:449–484, 1988.
- [14] Shalmali Joshi, Oluwasanmi Koyejo, Warut Vijitbenjaronk, Been Kim, and Joydeep Ghosh. Towards realistic individual recourse and actionable explanations in Black-Box decision making systems. July 2019.
- [15] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, et al. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). pages 2668–2677, 2018.
- [16] Carolyn Kim and Osbert Bastani. Learning interpretable models with causal guarantees. *arXiv:1901.08576 [cs, stat]*, January 2019.
- [17] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. In *2nd International Conference on Learning Representations, ICLR 2014 - Conference Track Proceedings*, 2014.
- [18] Murat Kocaoglu, Christopher Snyder, Alexandros G Dimakis, and Sriram Vishwanath. Causal-GAN: Learning causal implicit generative models with adversarial training. September 2017.
- [19] Yann LeCun, Corinna Cortes, and Chris Burges. MNIST handwritten digit database, 2010.
- [20] Richard C Lewontin. The analysis of variance and the analysis of causes. *Am. J. Hum. Genet.*, 26(3):400–411, 1974.
- [21] Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Raetsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *international conference on machine learning*, pages 4114–4124. PMLR, 2019.

- [22] C Maddison, A Mnih, and Y Teh. The concrete distribution: A continuous relaxation of discrete random variables. 2017.
- [23] Tanmayee Narendra, Anush Sankaran, Deepak Vijaykeerthy, and Senthil Mani. Explaining deep learning models using causal inference. *arXiv:1811.04376 [cs, stat]*, November 2018.
- [24] Matthew O’Shaughnessy, Gregory Canal, Marissa Connor, Mark Davenport, and Christopher Rozell. Generative causal explanations of black-box classifiers. *Advances in Neural Information Processing Systems*, 2020.
- [25] Álvaro Parafita and Jordi Vitrià. Explaining visual models by causal attribution. pages 4167–4175, 2019.
- [26] Judea Pearl. *Causality*. Cambridge University Press, September 2009.
- [27] Andrew Slavin Ross and Finale Doshi-Velez. Benchmarks, algorithms, and metrics for hierarchical disentanglement. February 2021.
- [28] Patrick Schwab and Walter Karlen. Cxplain: Causal explanations for model interpretation under uncertainty. *Advances in Neural Information Processing Systems*, 32:10220–10230, 2019.
- [29] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.
- [30] Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller. Striving for simplicity: The all convolutional net. December 2014.
- [31] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International Conference on Machine Learning*, pages 3319–3328. PMLR, 2017.
- [32] Francesco Tonolini, Bjørn Sand Jensen, and Roderick Murray-Smith. Variational sparse coding. In Ryan P Adams and Vibhav Gogate, editors, *Proceedings of The 35th Uncertainty in Artificial Intelligence Conference*, volume 115 of *Proceedings of Machine Learning Research*, pages 690–700. PMLR, 2020.
- [33] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-MNIST: a novel image dataset for benchmarking machine learning algorithms. August 2017.
- [34] Mengyue Yang, Furui Liu, Zhitang Chen, Xinwei Shen, Jianye Hao, and Jun Wang. CausalVAE: Disentangled representation learning via neural structural causal models. April 2020.
- [35] Chih-Kuan Yeh, Been Kim, Sercan Arik, Chun-Liang Li, Tomas Pfister, and Pradeep Ravikumar. On completeness-aware concept-based explanations in deep neural networks. *Advances in Neural Information Processing Systems*, 33, 2020.