Statistically Significant Concept-based Explanation of Image Classifiers via Model Knockoffs

Kaiwen Xu[1,3] , **Kazuto Fukuchi**[1,3] , **Youhei Akimoto**[1,3] and **Jun Sakuma**[2,3]
[1]University of Tsukuba
[2]Tokyo Institute of Technology
[3]RIKEN AIP
kaiwen@mdl.cs.tsukuba.ac.jp, {fukuchi, akimoto}@cs.tsukuba.ac.jp, sakuma@c.titech.ac.jp

Abstract

A concept-based classifier can explain the decision process of a deep learning model by human-understandable concepts in image classification problems. However, sometimes concept-based explanations may cause false positives, which misregards unrelated concepts as important for the prediction task. Our goal is to find the statistically significant concept for classification to prevent misinterpretation. In this study, we propose a method using a deep learning model to learn the image concept and then using the Knockoff samples to select the important concepts for prediction by controlling the False Discovery Rate (FDR) under a certain value. We evaluate the proposed method in our synthetic and real data experiments. Also, it shows that our method can control the FDR properly while selecting highly interpretable concepts to improve the trustworthiness of the model.

1 Introduction

In recent years, the development of Deep Neural Networks (DNNs) has achieved highly accurate predictions in various tasks. On the other hand, since the prediction process of DNNs is hard to understand for humans, this has driven the development of explainable artificial intelligence (XAI).

Some post-hoc explanation methods have been proposed to explain the DNNs in the image classification task. For example, methods based on saliency map highlight the pixels relevant to a specific class as visual explanation [Zhou *et al.*, 2016]. Compared to these post-hoc methods, the concept-based explanation can provide a basis for prediction by high-level concepts that are intuitively comprehensible to humans. For example, when a doctor predicts the severity of arthritis for a patient using a deep learning model, instead of giving the diagnosis result directly, the result is based on concepts like "the possibility of arthritis is high because the joint spacing is narrow", which is more reliable for humans.

The concept-based explanation is highly expressive; however, false positives often appear that could mislead humans' understanding. A false positive explanation occurs when an unimportant concept is mistakenly identified as important for the prediction task. For example, consider a deep learning model that is trained to classify images of animals. The model may be able to identify a horse in an image correctly, but it may also mistakenly identify the image's background as important for the classification. In this case, the background is an unimportant concept mistakenly identified as important, resulting in a false positive explanation.

In this study, we attempt to suppress the false positive explanations by providing explanations based on statistically significant concepts. We guarantee the reliability of the concept-based explanation by controlling the False Discovery Rate (FDR) of the selected concepts. Our motivation is to decrease the false positive explanations in the concept-based explanation model. We aim to select the statistically significant concepts for the prediction task that can improve the model's reliability.

1.1 Related Work

Representation Learning of Images. Learning latent representation (*i.e.,* concepts) of images can be divided into unsupervised and supervised learning. The unsupervised learning method aims to discover the concepts from data automatically. Among them, the mainstream approach treats unsupervised concept learning as a disentanglement representation learning. The Variation Autoencoder (VAE) framework [Kingma and Welling, 2014] is often used for representation learning, many studies aim to add regularized that encourage the model to generate disentangled representations [Higgins *et al.*, 2017; Kim and Mnih, 2018; Chen *et al.*, 2018; Tran *et al.*, 2022; O'Shaughnessy *et al.*, 2020; Goyal *et al.*, 2019]. There are also some methods proposed to learn concepts given samples labeled with concepts with supervised learning [Koh *et al.*, 2020; Chen *et al.*, 2020; Stammer *et al.*, 2021; Kazhdan *et al.*, 2020; Yang *et al.*, 2021]. At the same time, some research support both unsupervised and supervised learning methods depending on available dataset [Ding *et al.*, 2020; Sarkar *et al.*, 2022; Locatello *et al.*, 2020].

Reliability of Explanation. Although there are many studies related to improving the interpretability of models, studies on the reliability of explanations, *i.e.,* whether we can have confidence in the explanations have not been sufficiently studied. [Zhang *et al.*, 2022] proposed a method to improve the robustness of post-hoc explanation [Zhou *et al.*, 2016; Selvaraju *et al.*, 2017]. [Bahadori and Heckerman, 2021]

aims to improve the reliability of the concept-base explanation by removing confounding information of concept based on CBM [Koh *et al.*, 2020] and TCAV [Kim *et al.*, 2018].

Our goal is to improve the reliability of concept-based explanations by providing important concepts for prediction tasks with statistical guarantees. To the best of our knowledge, our work is the first to provide a statistically significant concept-based explanation.

1.2 Contribution

Our research makes the following contributions:

- We propose a method that can control the false discovery rate of concepts under a certain value to suppress the false positive explanation.
- To cope with the diverse conceptual learning methods, our method works on both supervised and unsupervised methods for concept learning.
- We evaluate our method on both synthetic and real datasets, and the results show that our method can provide highly interpretable concepts while properly controlling the false discovery rate.

2 Problem Setup

2.1 Concept-based Classification

Concept-based classification is a task to construct a potentially black-box classifier and explain the constructed classifier's decision process through human-interpretable concepts. Let $\mathcal{X} \subset \mathbb{R}^d, \mathcal{C} \subset \mathbb{R}^p, \mathcal{Y} = \{1, 2, \ldots, K\}$ be the input space of the image, the space of concepts, and the class labels space respectively. Instead of training a classifier $\Omega : \mathcal{X} \to \mathcal{Y}$ directly, a concept-based classifier firstly maps the input space to human-interpretable concepts[1] by $\phi : \mathcal{X} \to \mathcal{C}$ and then use $f : \mathcal{C} \to \mathcal{Y}$ to predict the class labels from the obtained concepts. ϕ is a feature extractor built with a neural network and f is a linear model[2]. When the input of f is an explainable feature and since f is a linear model, we can learn which concept is used to predict the class labels by observing the coefficients of f. In a linear model, the larger the absolute value of the model coefficient, the more important the corresponding feature is to the prediction.

We introduce two settings of the concept-based classification, unsupervised and supervised concept-based explanation. Since the unsupervised setting is more elemental, we define the problem of unsupervised concept-based explanation first.

Unsupervised Concept-based Explanation

The learner in the unsupervised concept-based explanation has access to the dataset of the input image and class label pairs, $\{\mathbf{x}^{(i)}, y^{(i)}\}_{i=1}^n \in (\mathcal{X} \times \mathcal{Y})^n$. We train our feature extractor ϕ to provide meaningful concepts useful for concept-based explanation from the given dataset.

The Variation Autoencoder (VAE) framework [Kingma and Welling, 2014] is often used for concept learning in the unsupervised setting. It consists of two neural networks, an encoder network ϕ and a decoder network θ, where the encoder part of VAE $\phi : \mathbf{x} \mapsto \mathbf{z}$ generates disentanglement representation $\mathbf{z} = \phi(\mathbf{x})$ of each image as the concepts, and the decoder part $\theta : \mathbf{z} \mapsto \hat{\mathbf{x}}$ reconstructs image $\hat{\mathbf{x}}$ as close as possible to the input image $\mathbf{x}$. We can obtain the entire network by optimizing the following function

$$\mathcal{L}_{\text{VAE}} = \mathcal{L}_R(\mathbf{x}, \hat{\mathbf{x}}) + \alpha \mathcal{L}_D(\mathbf{z}, \mathcal{N}(\mathbf{0}, \boldsymbol{I})). \tag{1}$$

Here, $\mathcal{L}_R(\cdot, \cdot)$ measures the reconstruction loss between $\mathbf{x}$ and $\hat{\mathbf{x}}$ and $\mathcal{L}_D(\cdot, \cdot)$ measures the degree of independence between the concepts, which encourages the concepts to capture independent and interpretable factors of image.

Supervised Concept-based Explanation

The learner in the supervised concept-based explanation can access the ground truth label of the concept on top of the dataset used in unsupervised learning methods. Specifically, the dataset is given by $\mathcal{D} = \{\mathbf{x}^{(i)}, \mathbf{c}^{(i)}, y^{(i)}\}_{i=1}^n$, where $\mathbf{c}^{(i)} \in \mathcal{C}$ represents ground truth concepts of each image. For example, when $\mathbf{x}$ represents a face image, then $\mathbf{c}$ can represent concepts such as hair color or whether there is a beard.

Concept Bottleneck Model (CBM) [Koh *et al.*, 2020] is one kind of supervised concept-based explanation method. The objective function of CBM can be expressed as :

$$\mathcal{L}_{\text{CBM}} = \mathcal{L}_C(\mathbf{z}, \mathbf{c}) + \alpha \mathcal{L}_Y(y, \hat{y}). \tag{2}$$

Here, $\mathcal{L}_C(\cdot, \cdot)$ represents the supervised loss of concepts which encourages the concepts $\mathbf{z} = \phi(\mathbf{x})$ predicted with the encoder to be as close as possible to the concept labels $\mathbf{c}$. $\mathcal{L}_Y(\cdot, \cdot)$ measures the prediction loss between predicted target $\hat{y} = f(\mathbf{z})$ and ground truth class label.

2.2 Concept Selection with Controlling FDR

To obtain a concept-based explanation, we investigate the coefficients of linear model f and select the concepts with large coefficients as important concepts for prediction. However, the coefficients of the linear model are affected not only by the association with the class labels but some other various factors, such as noise or correlation between input variables. Therefore, the concepts selected by only investigating the coefficients of a linear model may not be really important for prediction, which leads to low reliability. To improve reliability, we need to control the false discovery rate of unimportant concepts below a certain value.

Let $\mathcal{H}_0 \subseteq \{1, \ldots, p\}$ be the set that contains all indices corresponding to all unimportant variables (*i.e.,* concepts) in $\mathbf{z}$, where we will give its specific definition later in Section 3. Our goal is to find the largest subset $\hat{\mathcal{S}} \subseteq \{1, \ldots, p\}$ that contains important concepts while controlling the FDR under a certain value q. The FDR is defined as:

$$\text{FDR} = \mathbb{E}\left[\frac{|\hat{\mathcal{S}} \cap \mathcal{H}_0|}{\max(1, |\hat{\mathcal{S}}|)}\right]. \tag{3}$$

Here, FDR represents the proportion of unimportant concepts identified as important mistakenly. If we guarantee the upper bound of FDR of the concept selection results, we can say our explanation is reliable in the sense of statistical significance.

[1] For example, in gender classification from face images, hair length and skin color can be human interpretable concepts.

[2] f can also be other interpretable prediction models such as decision trees. However, we restrict it to the linear model in our study.

3 Concept Selection via Knockoff

To address the problem mentioned in Section 2.2, we propose to employ the technique of the Knockoff filter [Candes *et al.*, 2018]. Knockoff is a variable selection method that controls the FDR under the linear regression setup. In this section, we first describe the Knockoff filter for variable selection and then propose a methodology to combine the Knockoff filter with a concept-based explanation.

3.1 Knockoff

Let $X = (X_1, \ldots, X_p) \in \mathbb{R}^p$ be the explanatory variable and $Y \in \mathbb{R}$ be the response variable. Here, we assume there exists a linear model : $Y = \boldsymbol{\beta}^\top X + \epsilon$ where $\boldsymbol{\beta}$ is a sparse vector whose non-zero elements represent the ground truth of features important to Y, and ϵ represents the noise. Define the unimportant variables set as $\mathcal{H}_0 = \{j : \beta_j = 0\}$. Then, the goal of the Knockoff filter is to select the largest subset of variables $\hat{\mathcal{S}}$ such that FDR in Eq. (3) is controlled under a certain value q.

Model-X Knockoff

[Candes *et al.*, 2018] proposed to generate a Knockoff sample $\tilde{X} \in \mathbb{R}^p$ of X, which satisfies the following properties

$$(X, \tilde{X})_{\text{swap}(S)} \overset{d}{=} (X, \tilde{X}), \tag{4}$$

$$\tilde{X} \perp\!\!\!\perp Y | X. \tag{5}$$

Here, S can be any subset of $\{1, \ldots, p\}$ and the symbol $\overset{d}{=}$ means equality in distribution. Eq. (4) requires that the joint distribution will be invariant when any subset of variables is swapped with their Knockoff. For example when $p = 3$ and $S = \{2, 3\}$, we require that $(X_1, X_2, X_3, \tilde{X}_1, \tilde{X}_2, \tilde{X}_3)$ has the same distribution as $(X_1, \tilde{X}_2, \tilde{X}_3, \tilde{X}_1, X_2, X_3)$. Eq. (5) requires that the Knockoff sample $\tilde{X}$ is generated without looking at Y.

Once Knockoff has been generated (the way to generate Knockoff samples will be explained later), we calculate the statistics W_j for each variable X_j where $j \in \{1, \ldots, p\}$. Here, for n observations, we make $\boldsymbol{Y} \in \mathbb{R}^n$ to represent the observation of response variable and $[\boldsymbol{X}, \tilde{\boldsymbol{X}}] \in \mathbb{R}^{2n \times p}$ represents an augmented matrix of the original variable with their Knockoffs. The common choice to calculate the W_j is using L1-regularization as the following:

$$\hat{\boldsymbol{b}} = \underset{\boldsymbol{b} \in \mathbb{R}^{2p}}{\arg\min} \frac{1}{2} \|\boldsymbol{Y} - [\boldsymbol{X}, \tilde{\boldsymbol{X}}]\boldsymbol{b}\|_2^2 + \lambda \|\boldsymbol{b}\|_1 , \tag{6}$$

$$W_j = |\hat{b}_j| - |\hat{b}_{j+p}|. \tag{7}$$

Here $b_{1:p}$ and $b_{p+1:2p}$ correspond to the parameters of the original variable and Knockoffs, respectively. Recall that Knockoffs are synthetic variables constructed to be correlated with the original variables but are not predictive of the target variable. By comparing the magnitude of the coefficients of the model when trained on the original features with the Knockoff, we can determine which variables are important for predicting the target variable. A large and positive W_j can be the evidence against the hypothesis that X_j is unimportant for Y. Then the FDR can be controlled under a certain value q by the following theory.

Theorem 1. *[Candes* et al.*, 2018] Given $\boldsymbol{Y}$ and $\boldsymbol{X}$ following the linear model, let $\tilde{\boldsymbol{X}}$ be the Knockoff sample satisfying Eq. (4) and Eq. (5). Suppose W_j is calculated by Eq. (7) from $\boldsymbol{Y}, \boldsymbol{X}$, and $\tilde{\boldsymbol{X}}$. Given a target FDR level q, we select variables using W_j as $\hat{\mathcal{S}} = \{j : W_j \geq \tau\}$, where*

$$\tau = \min \left\{ t > 0 : \frac{1 + |\{j : W_j \leq -t\}|}{|\{j : W_j \geq t\}|} \leq q \right\}.$$

Then, the FDR of $\hat{\mathcal{S}}$ is controlled as the prescribed level, i.e.,

$$\mathbb{E}\left[\frac{|\hat{\mathcal{S}} \cap \mathcal{H}_0|}{\max(1, |\hat{\mathcal{S}}|)} \right] \leq q.$$

Second-order Sampler. Generating Knockoff sample that strictly achieves Eq (4) is difficult in practice. [Candes *et al.*, 2018] introduced a relaxation called the second-order sampler to generate the Knockoff sample. Unlike Eq. (4), this method only needs to match the first two moments of distributions. So when X is a multivariate Gaussian, second-order Knockoff [Candes *et al.*, 2018] works properly since matching the first two moments can exactly satisfy the Eq. (4).

Deep Knockoff Sampler. When X does not follow as a multivariate Gaussian, Theorem 1 does not hold only by matching the first two moments. [Romano *et al.*, 2020] proposed to introduce a generative model $f_{\text{deep}} : \mathbb{R}^p \to \mathbb{R}^p$ to generate the Knockoff sample $\tilde{X}$ which approximately satisfies Eq. (4).

3.2 Concept Selection via Model Knockoff

In this subsection, we introduce how to apply the Knockoff filter to the concept selection of image classification tasks.

The concept selection process is summarized in Algorithm 1. We divided the dataset into two parts, the dataset $\mathcal{D}_L$ for feature extractor training and the dataset $\mathcal{D}_S$ for concept selection as shown in step 1. First, we train a feature extractor on $\mathcal{D}_L$ and obtain the learned feature extractor $\hat{\phi}$. We remark that the training procedure of the feature extractor is chosen depending on the dataset the learner has access to, as we already discussed in Section 2.1. Then, we apply $\hat{\phi}$ to D_S to obtain concepts $\mathbf{z}^{(i)} = \hat{\phi}(\mathbf{x}^{(i)})$ for each image on $\mathcal{D}_S$ (step 3 - step 7). After that, we generate the Knockoff sample $\tilde{\mathbf{z}}$ of each concept $\mathbf{z}$ and obtain the dataset $\{\mathbf{z}^{(i)}, \tilde{\mathbf{z}}^{(i)}, y^{(i)}\}_{i=1}^{|\mathcal{D}_S|}$ (step 8). Then, we construct the linear model of $[\mathbf{z}, \tilde{\mathbf{z}}] \to y$ using the method described in Section 3.1 and compute the statistics to obtain the concept selection result $\hat{\mathcal{S}}$ with a given FDR level (step 9 - step 13).

3.3 FDR Control in Knockoff-based Concept Selection

This subsection demonstrates that Algorithm 1 adequately controls FDR below the prescribed level. Algorithm 1 carries out the Knockoff filter using the concept $\hat{\phi}(x)$ and class label y as the explanatory and response variables, respectively. This process implicitly supposes the existence of the underlying linear model between $\hat{\phi}(x)$ and y. The important concepts

Algorithm 1 Algorithm for concept selection

Input: Dataset: $\mathcal{D}$, FDR: q

Output: selected concept set: $\hat{\mathcal{S}}$

1: Split the data into two parts, $\mathcal{D}_L$ for concept learning and $\mathcal{D}_S$ for concept selection.
2: Optimize the $\hat{\phi}$ on $\mathcal{D}_L$.
3: $\mathbf{Z} \leftarrow \{\}$
4: **for** i in $\{1, 2, \ldots, |\mathcal{D}_S|\}$ **do**
5: $\quad \mathbf{z}^{(i)} \leftarrow \hat{\phi}(\mathbf{x}^{(i)})$ //generate concepts for each image.
6: $\quad \mathbf{Z} \leftarrow \mathbf{Z} \cup \mathbf{z}^{(i)}$
7: **end for**
8: Generate the Knockoff sample $\widetilde{\mathbf{Z}} = \{\tilde{\mathbf{z}}^{(1)}, \ldots, \tilde{\mathbf{z}}^{(|\mathcal{D}_S|)}\}$ of $\mathbf{Z}$.
9: Obtain the $\hat{\boldsymbol{b}}$ as Eq. (6).
10: Calculate the statistic W_j as Eq. (7).
11: Calculate the threshold τ by Theorem (1).
12: $\hat{\mathcal{S}} \leftarrow \{j : W_j \geq \tau\}$.
13: **return** $\hat{\mathcal{S}}$

that should be selected are the non-zero elements of the underlying model's parameter. Such an underlying linear model is uniquely determined to correspond to the feature extractor $\hat{\phi}$ under a mild condition. The next theorem shows the closed form of the underlying linear model and the FDR control accomplished by Algorithm 1.

Theorem 2. *Let $\hat{\mathcal{S}}$ be the selected concepts by Algorithm 1 with an arbitrary concept-based explanation method to construct $\hat{\phi}$. Suppose that Algorithm 1 employs Model-X Knockoff with the FDR level $q \in (0, 1)$ to generate the Knockoff sample. Also, suppose that $\mathbb{E}[\hat{\phi}(X)\hat{\phi}(X)^\top]$ is positive definite. Then, $\hat{\mathcal{S}}$ satisfies* $\mathrm{FDR} \leq q$ *for $\mathcal{H}_0 = \{j : \beta_{\phi,j} = 0\}$ where*

$$\beta_\phi = \mathbb{E}\left[\hat{\phi}(X)\hat{\phi}(X)^\top\right]^{-1} \mathbb{E}\left[\hat{\phi}(X)Y\right]. \tag{8}$$

The proof of Theorem 2 is left to the supplementary material. As proved by Theorem 2, Algorithm 1 adequately controls the FDR below q for $\mathcal{H}_0$ defined in Theorem 2. In Eq. (8), $\mathbb{E}[\hat{\phi}(X)Y]$ stands for the direct effect from a concept to the response variable, and $\mathbb{E}[\hat{\phi}(X)\hat{\phi}(X)^\top]^{-1}$ stands for the correlation between concepts. Hence, each element of β_ϕ represents the combination of the direct and indirect effects from a concept to the response variable. An unimportant concept j such that $\beta_{\phi,j} = 0$, is a concept that has neither direct nor indirect effects on the response variable.

4 Concept Sparsity Regularization

Difficulty. As shown in Section 3, the Model-X Knockoff regards a concept as unimportant only if $\beta_{\phi,j} = 0$. However, such a condition may be easily unsatisfied even for unimportant concepts since we train the feature extractor with data, and all concepts depend more or less on the response variable. As the experiments result shown in Fig. 2, which we will discuss later, the concept selection via the Model-X Knockoff combined with the standard concept learning algorithm, such as VAE and CBM, regards above 80% of all concepts as important. This demonstrates that even with Knockoff, a large number of unimportant concepts for prediction are still selected as important. This leads to decreasing in interpretability and trustworthiness of the explanation by our model.

Solution. From the interpretability perspective, we want to describe the response variable with a relatively small number of concepts. To this end, we introduce a concept learning algorithm so that the linear model trained with the resulting concepts (features) becomes sparse, that is, $\boldsymbol{\beta}_\phi$ is sparse. To make $\boldsymbol{\beta}_\phi$ sparse, we construct an estimator of $\boldsymbol{\beta}_\phi$ as $\hat{\boldsymbol{\beta}}_\phi = \arg\min_{\boldsymbol{\beta}_\phi} \mathcal{L}_Y(\hat{y}, y)$, where $\hat{y} = \boldsymbol{\beta}_\phi^\top \mathbf{z}$ is the prediction of the linear model. Because we expect that $\hat{\boldsymbol{\beta}}_\phi$ well approximates $\boldsymbol{\beta}_\phi$ by definition, if we could make $\hat{\boldsymbol{\beta}}_\phi$ sparse, $\boldsymbol{\beta}_\phi$ would also become sparse as we desired. To achieve this, we propose to add a regularization term called Concept Sparsity Regularization (CSR) to encourage feature sparsity in the linear model. The CSR is attained by adding the L1 norm of $\hat{\boldsymbol{\beta}}_\phi$ to the objective function of the feature extractor learning part, which penalizes $\hat{\boldsymbol{\beta}}_\phi$ for having too many non-zero coefficients. Our objective function is defined as:

$$\begin{aligned}\hat{\phi} = \arg\min_{\phi} \alpha_1 \mathcal{L}_R(\mathbf{x}, \hat{\mathbf{x}}) + \alpha_2 \mathcal{L}_D(\mathbf{z}, \mathcal{N}(\mathbf{0}, \boldsymbol{I})) \\ + \alpha_3 \mathcal{L}_C(\mathbf{z}, \boldsymbol{c}) + \alpha_4 |\hat{\boldsymbol{\beta}}_\phi|_1,\ s.t.\ \hat{\boldsymbol{\beta}}_\phi = \arg\min_{\boldsymbol{\beta}_\phi} \mathcal{L}_Y(\hat{y}, y).\end{aligned} \tag{9}$$

Implementation. Eq. (9) shows the objective function we want to optimize. However, the formulation requires solving a bi-level optimization problem. This problem requires to optimize $\hat{\boldsymbol{\beta}}_\phi$ on a linear model while $\hat{\boldsymbol{\beta}}_\phi$ depends on feature extractor ϕ. At the same time, ϕ needs to be optimized given $\hat{\boldsymbol{\beta}}$, which can be challenging and time-consuming. As a surrogate of the problem defined by Eq. (9), we train the parameter ϕ and $\boldsymbol{\beta}_\phi$ jointly using the following objective function:

$$\begin{aligned}\hat{\phi}, \hat{\boldsymbol{\beta}}_\phi = \arg\min_{\phi, \boldsymbol{\beta}_\phi} \alpha_1 \mathcal{L}_R(\mathbf{x}, \hat{\mathbf{x}}) + \alpha_2 \mathcal{L}_D(\mathbf{z}, \mathcal{N}(\mathbf{0}, \boldsymbol{I})) \\ + \alpha_3 \mathcal{L}_C(\mathbf{z}, \boldsymbol{c}) + \alpha_4 \mathcal{L}_Y(y, \hat{y}) + \alpha_5 |\boldsymbol{\beta}_\phi|_1.\end{aligned} \tag{10}$$

The problem defined by Eq. (10) is reduced to the original optimization problem by taking hyperparameter $\alpha_4 \to \infty$.

Formulation of Eq. (10) can be regarded as a generalization of the feature (concept) extraction part of various concept-based classification methodologies shown in Table 1. The most basic feature extractor by $\beta-$VAE [Higgins *et al.*, 2017] is obtained by setting $\alpha_1, \alpha_2 > 0, \alpha_3 = 0$. If the concept's label is available, we add the supervised loss (*i.e.*, $\alpha_3 > 0$), and it is equal to Full-VAE [Locatello *et al.*, 2020] which is one kind of supervised VAE. Also, letting α_1, α_2 to zero and only retaining $\alpha_3 > 0$, we obtain CBM [Koh *et al.*, 2020] which learns the concepts only by annotation data. In our unsupervised setting, we do not have labeled concepts in the dataset, so we just set $\alpha_3 = 0$ in Eq. (9) called CSR-VAE. In our supervised setting, since we can directly learn the concepts by concepts' ground truth, we set $\alpha_1, \alpha_2 = 0, \alpha_3 > 0$ to perform the concept learning, which is called CSR-CBM.

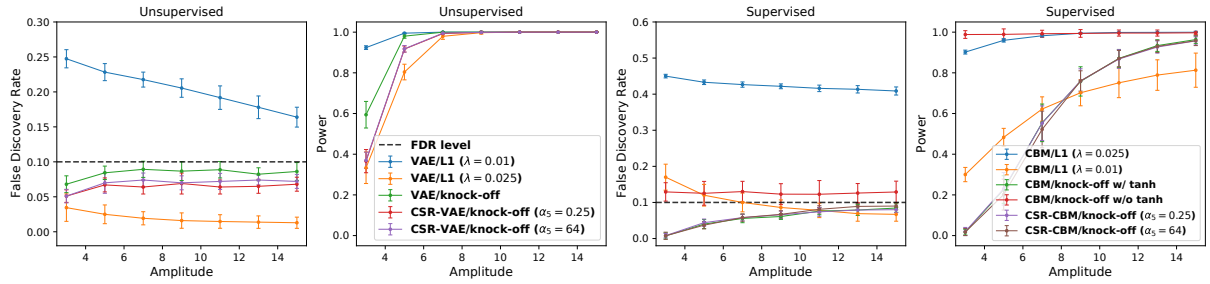


Figure 1: Experiments results on synthetic data in unsupervised and supervised settings. The results are averaged under 10 independent trials.

setting	concept learning method	α_1	α_2	α_3	α_4	α_5
unsupervised	$\beta-$ VAE [Higgins+17]	√	√			
	CSR-VAE (ours)	√	√		√	√
supervised	CBM [Koh+20]			√	√	
	Full-VAE [Locatello+20]	√	√	√		
	CSR-CBM (ours)			√	√	√

Table 1: Various concept learning methodologies covered Eq. (10), √ represents that set the $\alpha_j > 0$.

The difference in our concept selection process from Algorithm 1 is in step 2. In this step, we use CSR-VAE (in the unsupervised setting) and CSR-CBM (in the supervised setting) instead of VAE and CBM to make $\boldsymbol{\beta}_\phi$ sparse.

Regarding the FDR guarantee, since Theorem 2 guarantees FDR regardless of the learning method of ϕ, it guarantees FDR even for those generated by Eq. (10), which is expected to make $\boldsymbol{\beta}_\phi$ sparser.

5 Experiments on Synthetic Data

We evaluate whether our method can control the FDR strictly. It is hard to determine which concepts are important for target prediction in real image datasets, leading to the precise evaluation of FDR becoming difficult. Therefore, in this section, we design an image classification problem based on an artificially generated sparse linear model to evaluate whether our method can control the FDR strictly.

5.1 Experiment Setting

Dataset. In our synthetic experiments, we use the CelebA dataset [Liu *et al.*, 2015], which includes 202,599 face images, to evaluate our method. In the CelebA, each image contains 40 binary concept annotations (*e.g.,* smile, beard, etc.). In the unsupervised setting, the binary concept annotations are not used. Instead, the concepts are obtained using data in an unsupervised manner. In the supervised learning setting, the binary concept annotations are given, which are used to obtain feature vectors associated with each concept.

Synthetic Data. We construct a classification problem defined with an artificially defined linear model. Suppose we already obtained a set of concepts $\mathbf{Z}$ in the supervised or unsupervised manner. We first randomly sample m concepts $\mathbf{z}^{(i)}$, then we make artificial class label y as $y^{(i)} = \max(0, \text{sgn}(\boldsymbol{\beta}^\top \mathbf{z}^{(i)} + \epsilon^{(i)})$, where $\boldsymbol{\beta} \in \mathbb{R}^p$ is a sparse vector which contains k randomly chosen non-zero elements equal to $\pm a/\sqrt{m}$ (where the sign is chosen with probability 1/2 and $a \in \mathbb{R}$ represents the signal amplitude); $\epsilon^{(i)}$ is an $i.i.d$ sample drawn from $\mathcal{N}(0, 1)$. Here, the concepts associated with non-zero elements in $\boldsymbol{\beta}$ are ground truth important concepts for this classification. A small signal amplitude misleads the model to regard the ground truth as noise, and as the signal amplitude increases, the difficulty of ground truth detection will decrease. In this setting, the ground truth of unimportant concepts set is $\mathcal{H}_0 = \{j : \beta_j = 0\}$ and defined the important concepts set are $\mathcal{H}_1 = \{j : \beta_j \neq 0\}$.

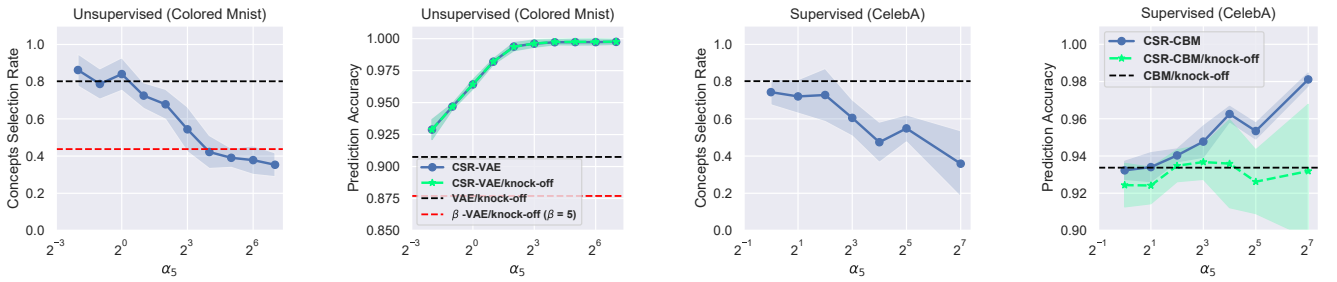
Evaluation Metric. The evaluation metric is FDR and power. The definition of FDR is stated as Eq. (3), and the power is defined as

$$\text{power} = \mathbb{E}\left[\frac{|\hat{\mathcal{S}} \cap \mathcal{H}_1|}{|\mathcal{H}_1|}\right], \tag{11}$$

The expected value in both Eq. (3) and Eq. (11) can be calculated over the randomness of sampling $\{\mathbf{z}^{(i)}, \tilde{\mathbf{z}}^{(i)}\}_{i=1}^m$.

Concept Selection. We compare two concept selection procedures. One is L1-regularized logistic regression (termed L1) which regards concepts that give non-zero coefficients as important and do not guarantee FDR. The other is Algorithm 1 with Knockoff (termed Knockoff), which guarantees FDR as we proved by Theorem 2. In the unsupervised concept learning setting, since $\mathbf{z}$ generated by the VAE-based method approximately follows the Gaussian distribution, we use the second-order Knockoff sampler to generate the Knockoff samples. While if the concepts are generated by the supervised learning method, since the distribution of $\mathbf{z}$ is not properly controlled, we need to train a deep-Knockoff machine priorly to generating Knockoff samples. To facilitate the training of deep-Knockoff, we apply a hyperbolic tangent function to bound $\mathbf{z}$ into range $[-1, 1]$ to properly bound the distribution of $\mathbf{z}$.

Comparison Method. In the unsupervised learning setting, we compare the combination of two concept learning methods (VAE, CSR-VAE) and two concept selection methods (L1, Knockoff). Similarly, in the supervised learning setting, we compare the combination of two concept learning methods (CBM, CSR-CBM) and two concept selection methods (L1, Knockoff). We use the notation: concept learning method/concept selection method to distinguish the different



(a) Concept selection rate in unsupervised setting

(b) Prediction accuracy in unsupervised setting

(c) Concept selection rate in supervised setting

(d) Prediction accuracy in supervised setting

Figure 2: Experiments results on real concepts in unsupervised and supervised settings. The results are averaged under 10 independent trials.

methods (*e.g.,* VAE-CSR/Knockoff). In this experiment, we mainly compare the difference between Knockoff and L1 regularization in terms of feature selection. We do not validate the sparsity of CSR because y has been defined as artificially generated with a sparse vector $\boldsymbol{\beta}$.

5.2 Experiment Results

Fig. 1 shows the variable selection results. The horizontal axes represent the value of a, which is the signal amplitude, and the vertical axes represent the FDR and power, respectively, with different variable selection methods. The black dotted line in the FDR graph represents the value of q, which we want to control.

The result shows that concept selection using Knockoff can effectively control the FDR under $q = 0.1$ in both unsupervised and supervised settings. The power is also improved by increasing the with signal amplitude a. On the other hand, L1-regularization (VAE/L1) cannot control the FDR under q appropriately in most cases, and a lot of false positives happen when λ is small. While the reason there is no significant difference when CSR is used or not is because $\boldsymbol{\beta}$ is already defined as a sparse vector.

6 Experiments on Real Concept Data

In the synthetic data experiments, we show that {VAE, CSR-VAE, CBM, CSR-CBM}/Knockoff can select important concepts by controlling the FDR properly. In this part, we apply our method to image classification of real datasets with and without concept labels and access if {CSR-VAE, CSR-CBM}/Knockoff can select a small number of concepts without knowing the ground truth concepts.

6.1 Dataset

Colored MNIST. We manually add six types of colors to the MNIST dataset [LeCun *et al.*, 1998] to create the Colored-MNIST dataset. This dataset is used to evaluate our method in the unsupervised learning setting. The prediction target in this task is digits.

CelebA. The concept labeled CelebA is also used in our real concept experiments for the supervised learning settings. The concepts other than gender in CelebA are treated as concepts, and gender is treated as the class label that we want to predict with concepts.

6.2 Sparsity Evaluation

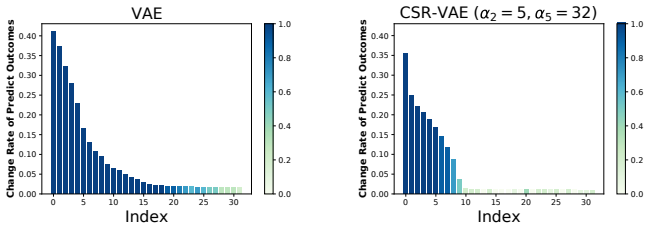
Settings. We applied {VAE, CSR-VAE, CBM, CSR-CBM}/Knockoff to select concepts important for the image classification task. We control the concept sparsity level of concept selection with CSR by adjusting α_5 and compare the number of selected concepts to concept selection without using CSR. We also compare the prediction accuracy of the linear model and the meaning of selected concepts.

Results of Concept Sparsity

Fig. 2 shows concept selection rate and prediction accuracy of CSR-{VAE,CBM}/Knockoff and {VAE,CBM}/Knockoff. The horizontal axes represent the regularization parameter. The vertical axes represent the concept selection rate (Fig. 2(a), Fig. 2(c)) and prediction accuracy (Fig. 2(b), Fig. 2(d)). The curve of CSR-{VAE,CBM} in the prediction accuracy graph shows the prediction accuracy using all concepts, while the accuracy is shown in the other curve just using the selected concepts.

The result in Fig. 2(a) and Fig. 2(c) shows that the concept selection rate decreases by increasing α_5 in our method, which implies that concept selection with concept sparsity regularization can keep the number of important concepts relatively small. When compared with VAE/Knockoff, it selects a large number of concepts even after using Knockoff. This means that concept sparsity regularization plays an important role in suppressing the number of the selected concept. Compared with $\beta-$VAE (red-dotted line), the number of selected concepts can be suppressed as CSR does use large β.

The result in Fig. 2(b) and Fig. 2(d) shows that the prediction accuracy also increases with the increase of α_5. This is because we set $\alpha_4 = 100 \times \alpha_5$. This setting means prediction error is more heavily penalized as α_5 increases. In the unsupervised learning setting, the prediction accuracy of the models using all concepts and using concepts selected by Knockoff surpasses the accuracy of VAE and $\beta-$ VAE. At the same time, only the model using concepts selected by our method can achieve the same prediction accuracy as when all concepts are used, thus demonstrating that our method can make the concepts learned by $\hat{\phi}$ sparse and select out the concepts that are important for prediction. In the supervised learning setting, as α_5 increases, the prediction accuracy when using only the concepts selected by our method and when using all



(a) change rate (VAE) (b) change rate (CSR-VAE)

Figure 3: The prediction outcomes' change rate of reconstruction image by ω and feature selection results by different methods.

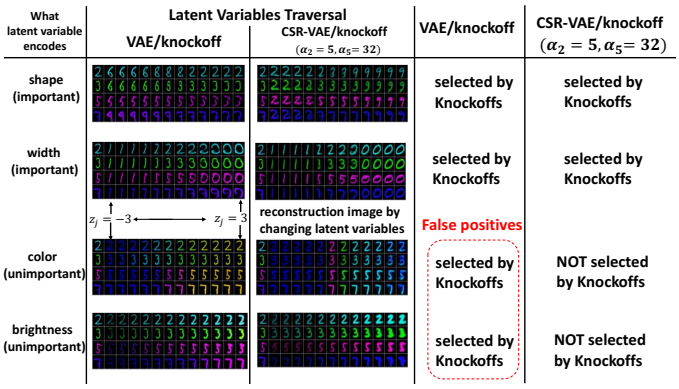


Figure 4: Demonstration of feature selection for digit classification in the unsupervised setting (Colored-MNIST).

concepts gradually shows a gap. The reason is that each concept has to be trained to correspond to a specific meaning, and an overly sparsified model will lead to a decrease in power, affecting the prediction accuracy.

Results of Selected Concepts in the Unsupervised Setting

In the unsupervised setting, we train a classifier ω on Colored MNIST to classify digits which helps us determine the important concepts. We intervene on each latent variable generated by VAE/CSR-VAE between [-3, 3] and use ω to predict the class of reconstruction images. We calculate the ratio by how much the reconstruction and original images class becomes different and denote it as the change rate. If a latent variable is considered important for prediction, it should have a relatively larger change rate value than other latents.

Fig. 3 shows the change rates and feature selection results by VAE/Knockoff and CSR-VAE/Knockoff. The horizontal axes represent the index of the latent variable, and the vertical axes represent the mean value of the prediction outcome's change rate by ω under 10 independent runs and sorted by value. The value of the color bar represents the frequency that each latents selected by Knockoff among 10 runs. We observe that the change rates of CSR-VAE are lower than VAE overall. This is because the distribution of $\mathbf{z}$ generated by ϕ learned by different algorithms is different, which leads to the fact that for the same intervention, it may be able to affect the reconstruction of VAE but not CSR-VAE. Both Fig. 3(a) and Fig. 3(b) show that Knockoff is effective in selecting the latent variables to attain greater change rates. CSR-VAE/Knockoff can select a small number of important features for prediction, while VAE/Knockoff selects more than 80% of features containing unimportant features. For demonstration purposes, we choose two important concepts (shape and width) and two unimportant concepts (color and brightness) among all concepts by their looks. In Fig. 4, the first column of the left side represents the original image, and the remaining columns show the decoder's reconstructions after intervention. The results show that VAE/Knockoff incorrectly treats color and brightness as important features. Even with latent variables with approximately the same meaning, the feature selection results by Knockoff can be different. At the same time, our proposed CSR-VAE/Knockoff can effectively suppress these false positive explanations, improving the explanation's reliability.

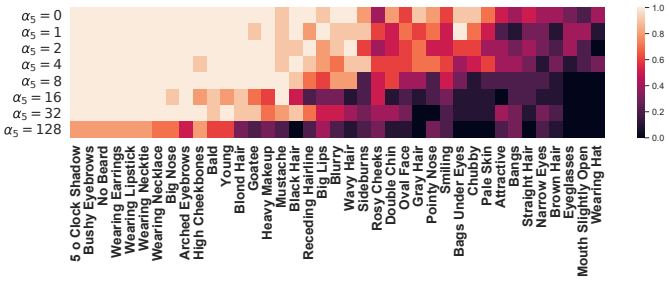


Figure 5: Demonstration of feature selection results in the supervised setting (CelebA).

Results of Selected Concepts in the Supervised Setting

In Fig. 5, each row represents the selection rate of each concept under different 10 independent runs in CelebA. The concept is sorted by the mean value of the concept selection rate under all α_5 (*i.e.,* the left side concepts is more likely to be selected under all hyperparameters). The method that gives the result in the top row is equal to CBM/Knockoff and each of the remaining rows corresponds to CSR-CBM/Knockoff with varying α_5. We observe that the selection rate of each concept decreases by increasing the value of α_5. When α_5 increases, concepts show a decreasing trend in selection rate, which effectively suppresses the occurrence of false positives. We found that concepts on the left side are subjectively important in human vision, such as "no beard" or "wearing earrings". While some subjectively unimportant concepts are also selected with a high probability (*e.g.,* "young", "blond hair"). The possible reason is that there is some bias in the important features used by the model and the human in making predictions, which leads to some subjective false positive concepts that may be important for the model to predict.

7 Conclusion

In summary, we proposed a method to suppress the false positive concept-based explanation by controlling the FDR of selected concepts under a certain level. Also, we propose a CSR loss that can be added to the concept model to more effectively apply the Knockoff filter for concept selection and thus improve the interpretability and reliability of the model.

Acknowledgments

This work is partly supported by Japan science and technology agency (JST), CREST JPMJCR21D3, and Japan society for the promotion of science (JSPS), Grants-in-Aid for Scientific Research 23H00483, 22H00519, and 22H00521.

References

[Bahadori and Heckerman, 2021] Mohammad Taha Bahadori and David Heckerman. Debiasing concept-based explanations with causal analysis. In *ICLR*, 2021.

[Candes *et al.*, 2018] Emmanuel Candes, Yingying Fan, Lucas Janson, and Jinchi Lv. Panning for gold:'model-x'knockoffs for high dimensional controlled variable selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2018.

[Chen *et al.*, 2018] Tian Qi Chen, Xuechen Li, Roger B. Grosse, and David Duvenaud. Isolating sources of disentanglement in variational autoencoders. In *NeurIPS*, pages 2615–2625, 2018.

[Chen *et al.*, 2020] Zhi Chen, Yijie Bei, and Cynthia Rudin. Concept whitening for interpretable image recognition. *Nat. Mach. Intell.*, 2020.

[Ding *et al.*, 2020] Zheng Ding, Yifan Xu, Weijian Xu, Gaurav Parmar, Yang Yang, Max Welling, and Zhuowen Tu. Guided variational autoencoder for disentanglement learning. In *CVPR*, 2020.

[Goyal *et al.*, 2019] Yash Goyal, Uri Shalit, and Been Kim. Explaining classifiers with causal concept effect (cace). *CoRR*, 2019.

[Higgins *et al.*, 2017] Irina Higgins, Loïc Matthey, Arka Pal, Christopher P. Burgess, Xavier Glorot, Matthew M. Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. In *ICLR*, 2017.

[Kazhdan *et al.*, 2020] Dmitry Kazhdan, Botty Dimanov, Mateja Jamnik, Pietro Liò, and Adrian Weller. Now you see me (CME): concept-based model extraction. In *CIKM (Workshops)*, 2020.

[Kim and Mnih, 2018] Hyunjik Kim and Andriy Mnih. Disentangling by factorising. In *ICML*, Proceedings of Machine Learning Research, 2018.

[Kim *et al.*, 2018] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie J. Cai, James Wexler, Fernanda B. Viégas, and Rory Sayres. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). In *ICML*, 2018.

[Kingma and Welling, 2014] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *ICLR*, 2014.

[Koh *et al.*, 2020] Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In *ICML*, 2020.

[LeCun *et al.*, 1998] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 1998.

[Liu *et al.*, 2015] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *ICCV*, 2015.

[Locatello *et al.*, 2020] Francesco Locatello, Michael Tschannen, Stefan Bauer, Gunnar Rätsch, Bernhard Schölkopf, and Olivier Bachem. Disentangling factors of variations using few labels. In *ICLR*, 2020.

[O'Shaughnessy *et al.*, 2020] Matthew R. O'Shaughnessy, Gregory Canal, Marissa Connor, Christopher Rozell, and Mark A. Davenport. Generative causal explanations of black-box classifiers. In *NeurIPS*, 2020.

[Romano *et al.*, 2020] Yaniv Romano, Matteo Sesia, and Emmanuel Candès. Deep knockoffs. *Journal of the American Statistical Association*, 2020.

[Sarkar *et al.*, 2022] Anirban Sarkar, Deepak Vijaykeerthy, Anindya Sarkar, and Vineeth N. Balasubramanian. A framework for learning ante-hoc explainable models via concepts. In *CVPR*, 2022.

[Selvaraju *et al.*, 2017] 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 *ICCV*, 2017.

[Stammer *et al.*, 2021] Wolfgang Stammer, Patrick Schramowski, and Kristian Kersting. Right for the right concept: Revising neuro-symbolic concepts by interacting with their explanations. In *CVPR*, 2021.

[Tran *et al.*, 2022] Thien Q. Tran, Kazuto Fukuchi, Youhei Akimoto, and Jun Sakuma. Unsupervised causal binary concepts discovery with VAE for black-box model explanation. In *AAAI*, 2022.

[Yang *et al.*, 2021] Mengyue Yang, Furui Liu, Zhitang Chen, Xinwei Shen, Jianye Hao, and Jun Wang. Causalvae: Disentangled representation learning via neural structural causal models. In *CVPR*, 2021.

[Zhang *et al.*, 2022] Wencan Zhang, Mariella Dimiccoli, and Brian Y. Lim. Debiased-cam to mitigate image perturbations with faithful visual explanations of machine learning. In *CHI*, 2022.

[Zhou *et al.*, 2016] Bolei Zhou, Aditya Khosla, Àgata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *CVPR*, 2016.