CAUSAL INFORMATION BOTTLENECK BOOSTS ADVERSARIAL ROBUSTNESS OF DEEP NEURAL NETWORK

Anonymous authors
Paper under double-blind review

ABSTRACT

The information bottleneck (IB) method is a feasible defense solution against adversarial attacks in deep learning. However, this method suffers from the spurious correlation, which leads to the limitation of its further improvement of adversarial robustness. In this paper, we incorporate the causal inference into the IB framework to alleviate such a problem. Specifically, we divide the features obtained by the IB method into robust features (content information) and non-robust features (style information) via the instrumental variables to estimate the causal effects. With the utilization of such a framework, the influence of non-robust features could be mitigated to strengthen the adversarial robustness. We make an analysis of the effectiveness of our proposed method. The extensive experiments in MNIST, FashionMNIST, and CIFAR-10 show that our method exhibits the considerable robustness against multiple adversarial attacks. Our code would be released.

1 INTRODUCTION

With the continuous improvement of computing power and data availability, the deep neural networks (DNNs) have made breakthroughs in many fields, such as image classification Haralick et al. (1973), object detection Redmon et al. (2016), machine translation Brown et al. (1990), natural language understanding Devlin et al. (2018), and so on. In DNNs, feature maps in the middle layers are treated as compression code Z. However, many studies in recent years have shown that DNNs are susceptible to adversarial examples Szegedy et al. (2014b;a); Madry et al. (2018). In the field of computer vision, adversarial examples which manipulate a small number of image pixels without the change of semantic representation can deceive DNNs to make false predictions Xu et al. (2020). It would be a huge threat to autonomous driving Eykholt et al. (2018), face recognition Sharif et al. (2016), and daily shopping Liu et al. (2020). Therefore, the security of deep neural networks has become a significant concern.

The phenomenon of the adversarial vulnerability can be regarded as the overfitting problem of DNNs Goodfellow et al. (2015). Moreover, the defect of the decision boundaries in DNNs leads to the possibility of adversarial attacks Goodfellow et al. (2014). As an effective regularization method, the information bottleneck (IB) theory can help reduce the adversarial empirical risk and approximate a better decision boundary. The IB theory is an extension of Shannon's rate-distortion theory Tishby et al. (2000). Its goal is to find an optimal compression code for the target random variable, which maximizes the mutual information between the target random variable and the compression code, and minimizes the mutual information between the source random variable and the compression code. The IB method is believed to be useful to explain the operation and principle of DNNs, and the mechanism of the information compression in the method of IB helps the DNNs extract the representative features Tishby et al. (2000); Shwartz-Ziv & Tishby (2017); Tishby & Zaslavsky (2015). Many pieces of follow-up work have been inspired, including the exploration of the relationship between adversarial robustness and IB theory Achille & Soatto (2018); Alemi et al. (2017); Fischer (2020). In recent years, many theoretical analyses have been proposed, and many models and algorithms have also been shown to be effective, including Information Dropout Achille & Soatto (2018) and Variation IB (VIB) Alemi et al. (2017), Diesentangled IB Pan et al. (2021), and so on Kim et al. (2022); Fischer (2020); Voloshynovskiy et al. (2019).

Although some IB-based methods improve the adversarial robustness of the model to a certain extent, there is a key problem with these methods: the existing IB methods Shwartz-Ziv & Tishby (2017);

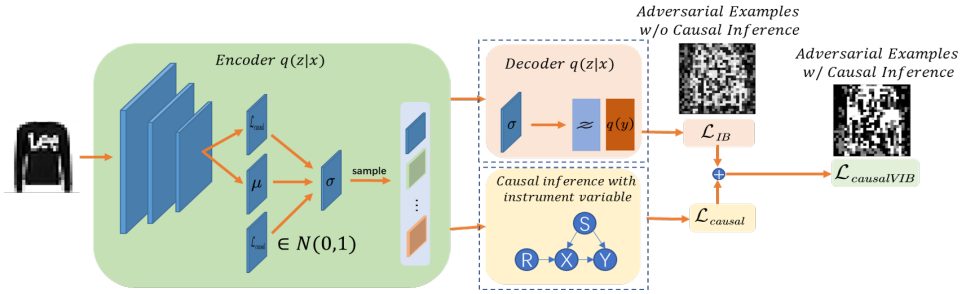


Figure 1: The proposed CausalIB framework is divided into two modules, IB and causal inference. The IB module is used for overall regularization, and the causal inference module is used to distinguish robust features from non-robust features. With the utilization of causal inference, the proposed CausalIB can learn more structured information than models using only IB module.

Tishby & Zaslavsky (2015); Alemi et al. (2017); Fischer & Alemi (2020) do not pay attention to the problem of the spurious correlation that the robust features and the non-robust features would be entangled with each other. The learning bias on the fragile and incomprehensible features (non-robust features) would lead to the adversarial vulnerability. In contrast, mankind with the ability of causal inference Freeman (1994); Parascandolo et al. (2017) can normally recognize the visual adversarial examples correctly by identifying and peeling those unrelated factors Pearl (2009). With the utilization of causal inference, DNN models would focus more on robust features Tang et al. (2021). To alleviate the problem of the entanglement between the robust features and the non-robust features in the IB framework, we propose a causal intervention method. In this study, we assume that the style information of an image is a non-robust feature, while the content information is a robust feature. Since the style information cannot be observed, we introduce instrumental variables to help remove the influence of style information. In the experiment, we validate the effectiveness of the proposed causal inference IB method under various adversarial attacks such as Fast Gradient Sign Method (FGSM) Goodfellow et al. (2015), PGD Madry et al. (2018) on the MNIST LeCun et al. (1998), FashionMNIST Xiao et al. (2017) and CIFAR-10 Krizhevsky et al. (2009) datasets. The empirical results validate the effectiveness of our method to boost the adversarial robustness.

Our contributions are as follows:

1. We utilize the causal theory to analyze the cause of the adversarial vulnerability of deep learning models and the feasibility of improving robustness through instrumental variables;

2. We introduce a model called CausalIB, which uses the IB method with the assistance of causality to extract features and causal inference to disentangle the robust features and non-robust features.

3. The extensive experiments on various settings of MNIST, FashionMNIST, and CIFAR-10 show that CausalIB is robust against adversarial attacks.

The remainder of this paper is organized as follows. First, we make a literature review in Section 2. In Section 3, we introduce the IB theory and analyze the reasons for the existence of adversarial examples from the perspective of causal theory. Moreover, the approach is illustrated in Section 3. The experiment's detail and result will be described in Section 4. Finally, the conclusion of the research is given in Section 5.

2 Related Work

In this section, we briefly review the current state of research on adversarial robustness, prior works on IB methods, and related causal inference methods.

2.1 Adversarial Robustness

The existing methods for dealing with adversarial examples mainly include three directions Akhtar & Mian (2018): preprocessing methods for the defense Miyato et al. (2017); Zheng et al. (2016); Shin & Song (2017), improvement on the neural network structures Rifai et al. (2011); Bai et al. (2017);

Hinton et al. (2015), and the utilization of external models when classifying unseen examples Akhtar et al. (2018); Lee et al. (2017). Currently, the most effective strategy is adversarial training Goodfellow et al. (2015); Tramèr et al. (2018). Adversarial training can be regarded as a method of data augmentation, its adversarial robustness largely depends on the coverage of adversarial examples during the training process. However, the study of Moosavi Dezhouni et al. Moosavi-Dezfooli et al. (2017) found that even a well-trained defense network can still obtain other effective adversarial examples through computation which brings new difficulties for the adversarial defense. The methods without the utilization of adversarial training such as preprocessing methods Xie et al. (2019); Warde-Farley & Bengio (2017); Das et al. (2017), data randomization methods Pinot et al. (2019); Cohen et al. (2019); Xie et al. (2017), or IB methods Tishby et al. (2000); Shwartz-Ziv & Tishby (2017) also provide insight for adversarial robustness. Among them, Das et al. Das et al. (2017) utilized the method of compression to remove the high-frequency components from images to improve robustness. Some pieces of seminal work Xie et al. (2017); Wang et al. (2016) show that data randomization has a role in reducing the fooling rates of the networks. Initially, the IB theory was seen as an interpretive work for DNNs Tishby et al. (2000). However, it has also been found that the IB method is an effective regularization method due to its trade-off between prediction performance and model compression, which can be used to improve the adversarial robustness Alemi et al. (2017); Fischer (2020); Achille & Soatto (2018); Pan et al. (2021).

2.2 IB METHOD

In recent years, IB methods have been widely used to improve model robustness. Achille et al. Achille & Soatto (2018) improved the Dropout method with the utilization of the IB framework to reduce the sensitivity of the model to perturbations. A similar method is delivered by Kim et al. Kim et al. (2022) that the robust and the non-robust neurons could be separated according to the different encoding values reacting to the noises. Different from the above methods, Alemi et al. Alemi et al. (2017) proposed a variational approximation method to optimize the IB, using reparameterization trick Kingma & Welling (2014) for efficient training and demonstrating the effectiveness against adversarial examples. Fischer et al. Fischer (2020) postulated that the vulnerability of neural networks stems from the fact that the model retains too much information about the training data, resulting in weakness under the adversarial attacks. Considering that the essence of the IB is the restriction of the complexity of representation learning, the proposal of the conditional entropy-based IB (CEB) strengthened the adversarial robustness and generalization ability of the IB framework Fischer & Alemi (2020). A recent study Korshunova et al. (2021) postulates that the information bottleneck methods based on the variational inference would suffer from gradient obfuscation due to the non-smooth loss surfaces durint the optimization process. Such a challenge inspires us to explore the feasible method to promote adversarial robustness of IB methods. Besides, DiesenIB Pan et al. (2021) implements the IB method from the perspective of supervised disentanglement. In the proposed DiesenIB, the information is not compressed, no compression means no loss of prediction performance, and disentanglement helps the model to achieve a better performance in out-of-distribution (OOD) and adversarial defense. Through improving VIB, Sinha Sinha et al. (2021) obtained the diversity of output prediction, which is required for multimodal data modeling, and achieved certain results in sparse training data and uncertainty estimation for OOD detection, etc.

2.3 CAUSAL INFERENCE

Human perception is robust to adversarial perturbations due to the ability of causal inference Zhang et al. (2020); Pearl (2009; 2010). Many recent works have shown in various aspects that causal models are suitable methods for parametric reasoning in complex systems Kusner et al. (2017), and can even be utilized to interpret the deep learning models Schölkopf et al. (2021). Recently, Zhang et al. Zhang et al. (2021) proposed a method called the adversarial distribution alignment, which attempts to explain the existence of adversarial examples from the perspective of causality. Such a method removes spurious correlations by eliminating the difference between natural and adversarial distributions. A similar approach called causal manipulation augmented model Zhang et al. (2020) aims to improve the robustness of DNNs to unseen adversarial perturbations by explicitly modeling the perturbations from a causal view. The stable learning methods are delivered Peters et al. (2016); Kuang et al. (2020) to reduce the accuracy variance of the model under various sample distributions via the causal methods. Most relevant to our work is the causal intervention by instrumental variable (CiiV) model Tang et al. (2021). **However, the CiiV model uses retinotopic sampling to intervene**

in the image layer (original image layer). The difference between our proposed method and the previous work Tang et al. (2021) is that we use additive noise as the instrumental variable in the concept layer (intermediate feature layer) rather than in the input layer.

3 APPROACH

3.1 INFORMATION BOTTLENECK

The core idea of the IB theory is information compression, that is, to maximize the mutual information between the target random variable and the compression code, and minimize the mutual information between the source random variable and the compression code. The loss function of IB is defined as Eq.1:

$$\mathcal{L}_{IB} = -I\left(Z;Y\right) + \alpha I\left(Z;X\right), \tag{1}$$

where $I(Z;Y)$ is the mutual information between the target random variable Y and the compression code Z, and $I(Z;X)$ is the mutual information between the source random variable X and Z. The IB restricts the correlation between X and Z, which is an effective regularization method to reduce the empirical risk. In DNNs, feature maps in the middle layers are treated as compression codes. However, it is not easy to calculate mutual information in DNNs. In this regard, the variational inference is added into the IB framework to construct the variational IB (VIB) model Alemi et al. (2017). The lower bound of $I(Z,Y)$ and the upper bound of $I(Z,X)$ are formalized as Eq.2 and Eq.3:

$$I\left(Z,Y\right) \;\geq\; \int dxdydz\; p\left(x\right)p\left(y|x\right)p\left(z|x\right)\log q\left(y|z\right) = \mathbb{E}_{q(y,t)}\log p\left(y|t\right), \tag{2}$$

$$I\left(Z,X\right) \;\leq\; \int dxdz\; p\left(x\right)p\left(z|x\right)\log\frac{p\left(z|x\right)}{r\left(z\right)}, \tag{3}$$

where $r(z)$ is the prior probability of the latent variable Z, which is set to a multi-dimensional Gaussian distribution with mean 0 and variance 1. Reparameterization trick Kingma & Welling (2014) is used to estimate the mutual information. Unlike deterministic models, with the reparameterization trick, VIB learns not only deterministic features in feature layers but also a whole multi-dimensional Gaussiann distribution.

The mean of samples of the distribution would be taken by VIB, and the loss function can be written as Eq.4:

$$\begin{aligned}\mathcal{L}_{IB} &\approx \frac{1}{N}\sum_{n=1}^{N}\left[\int dzp\left(z\mid x_n\right)\log q\left(y_n\mid z\right) - \alpha p\left(z\mid x_n\right)\log\frac{p\left(z\mid x_n\right)}{r(z)}\right] \\ &= \frac{1}{N}\sum_{n=1}^{N}\mathbb{E}_{\epsilon\sim p(\epsilon)}\left[-\log q\left(y_n\mid f\left(x_n,\epsilon\right)\right)\right] + \alpha\mathrm{KL}\left[p\left(Z\mid x_n\right),r(Z)\right],\end{aligned} \tag{4}$$

where the former term is the cross entropy, and the latter term is the Kullback–Leibler (KL) divergence.

3.2 A CAUSAL VIEW OF EXISTENCE OF ADVERSARIAL EXAMPLES

The existing neural networks would learn a spurious correlation between data and predictions which induces their vulnerabilities Zhang et al. (2021); Tang et al. (2021); Kuang et al. (2020). We postulate that these promiscuous, spuriously correlated features, such as background features, color features, or even the bias of camera angles may lead to the adversarial vulnerability. Humans are not affected by such disturbances due to the ability of causal inference including observation, intervention, and counterfactual Pearl (2009); Peters et al. (2017). Therefore, we turn to causal theory to mitigate the spurious correlations that existed in the learning process of DNNs.

Taking a figure consisting of a dog and a grass background as an example, humans do not interfere with the information of grass when recognizing the dog, but DNNs will learn both the dog information and the grass information. In the process of feature learning, DNN models simply observe all the data without removing the spurious correlations between the features Zhang et al. (2021), and the model is quite sensitive to these features with spurious correlations. Therefore, the adversaries can easily

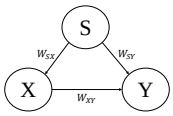


Figure 2: Causal graph without the instrumental variable.

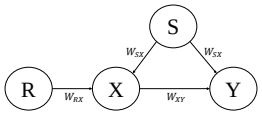


Figure 3: Causal graph with the instrumental variable.

fool the model with the manipulation of the pixels with promiscuous information, which includes the various backgrounds, the different viewing perspectives, changing colors, and so on.

The model focuses more on the robust features rather than the confounding factors if the causal inference methods are incorporated. Based on the causal theory, the generation process of the data can be visualized in Figure 2. A causal graph illustrates the causal relationship between data and features, which would help understand the generation of adversarial examples. In this study, since we have no way to use causal structure learning in complex high-dimensional data, we leverage external knowledge to build the causal graph. As shown in Figure 2, the graphical model is implemented where X, Y, and S represent the original data, prediction, and style information respectively. DNNs should learn robust features which are closely related to the label information. However, the current learning mechanism of DNNs would not distinguish the robust features from non-robust features. $X \leftarrow S \rightarrow Y$ shows that style information is the common cause of X and Y, and it affects the distribution of X and Y at the same time. To distinguish the robust features from the non-robust features, the path from S to Y should be truncated.

To obtain the pure causal effect, we turn to instrumental variable estimation. By definition Pearl (2009; 2010), a valid instrumental variable should satisfy: 1) it is independent of confounding variables; 2) it affects Y only through X. We argue that the artificially introduced additive noise fully meets these two requirements. Figure 3 illustrates the linear confounded model with the instrumental variable.

When no instrument variable is introduced, X is only affected by S, that is, $x = w_{sx}s + u_x$, where x is the source information, s is the style information, w_{sx} is the path coefficient between S and X, u_x is the independent component of X. And Y is affected by both X and S, that is, $y = w_{xy}x + w_{sy}s + u_y$, where w_{xy} is the path coefficient between X and Y, w_{sy} is the path coefficient between S and Y, u_y is the independent component of Y. In this case, we have no way of knowing the causal effect of X on Y because the distribution of S is unobservable.

When the instrument variable R is introduced, X is affected by both R and S, that is, $x_r = w_{sx}s + w_{rx}r + u_x$, $y_r = w_{xy}x_r + w_{sy}s + u_y$, where r is the instrument variable, and w_{rx} the path coefficient between R and X. To remove the influence of style information S, we utilize the direct controlled influence (CDE):

$$CDE = P\left(Y = y|do\left(X = x_i\right), do\left(S = s\right)\right) - P\left(Y = y|do\left(X = x_j\right), do\left(S = s\right)\right), \quad (5)$$

where i and j are different interventions. Substituting the above formula into Eq.5, Eq.6 can be acquired:

$$y_{r_i} - y_{r_j} = w_{xy}\left(x_{r_i} - x_{r_j}\right), \quad (6)$$

then the robust feature w_{xy} can be learned by Eq.7:

$$w_{xy} = \frac{y_{r_i} - y_{r_j}}{x_{r_i} - x_{r_j}}. \quad (7)$$

3.3 The Proposed CausalIB

In practice, we refer to the idea of CiiV Tang et al. (2021), which uses the retinal mask as the instrumental variable in the original image layer to obtain images under different gaze angles, and achieves good results. However, due to the property of representation learning in DNNs, we postulate that it would be more rational to use instrumental variables in the intermediate feature level instead of the original image layer. Therefore, we introduce additive Gaussian noise as the instrumental variable in the feature layer. Next, we will specifically describe the proposed method.

Table 1: The performances of white-box attack on MNIST and FashionMNIST.

	MNIST			FashionMNIST		
Method	**clean**	**FGSM**	**PGD-20**	**clean**	**FGSM**	**PGD-20**
Baseline	95.11 ± 0.72	6.76 ± 0.93	0.0 ± 0.0	89.81 ± 0.49	0.63 ± 0.34	0.0 ± 0.0
mixu	98.22 ± 0.81	20.66 ± 1.51	0.43 ± 0.08	90.44 ± 0.29	9.13 ± 3.05	0.11 ± 0.04
RS	98.45 ± 0.53	34.15 ± 0.91	13.38 ± 0.93	89.71 ± 0.22	13.57 ± 1.13	1.30 ± 0.21
CiiV	97.94 ± 0.06	71.80 ± 0.32	48.48 ± 0.77	$\mathbf{94.84 \pm 0.15}$	44.82 ± 0.87	$\mathbf{15.01 \pm 0.34}$
VIB	98.70 ± 0.26	66.06 ± 0.21	41.84 ± 0.33	94.42 ± 0.44	25.88 ± 0.41	3.98 ± 0.58
CausalIB	$\mathbf{98.85 \pm 0.20}$	$\mathbf{81.07 \pm 0.82}$	$\mathbf{53.71 \pm 0.51}$	94.06 ± 0.25	$\mathbf{50.81 \pm 0.80}$	14.06 ± 0.11
AT (FGSM)	94.02 ± 0.12	75.82 ± 1.13	67.86 ± 2.35	89.95 ± 0.15	55.52 ± 0.80	17.46 ± 1.76
AT (PGD-20)	93.92 ± 0.05	78.31 ± 2.58	69.84 ± 0.94	87.22 ± 0.49	56.43 ± 0.61	41.87 ± 1.60

Our method is based on VIB Alemi et al. (2017), which learns multi-dimensional Gaussian distributions instead of deterministic features in feature layers. VIB takes the mean of samples of the feature distribution, but the difference between these feature layer samples is neglected. Such a difference can be modeled via the additive noises as the instrumental variable. In our framework, the data distribution rather than the final value of the mean output was subjected to causal inference.

The relationship between Y and R can be written as Eq.8:

$$Y[X = x_r] = w_{xy}x_r + w_{sy}s = w_{sy}s + w_{xy}w_{sx}s + w_{rx}w_{xy}r = Y[X = x] + w_{rx}w_{xy}r. \tag{8}$$

Therefore, the Eq.7 can be written as Eq.9:

$$w_{xy_i} = w_{rx}w_{xy} = \frac{Y\left[X = x_{r_i}\right] - Y\left[X = x\right]}{r_i}, \tag{9}$$

where x_{r_i} is the ith sample of X with the instrumental variable. We assume that the magnitude of r is equal to the intensity of the introduced noise. In practice, since R is independent of S, the causal loss function can be formalized as Eq.10:

$$\mathcal{L}_{causal} = \sum_{i \neq j} ||w_{xy_i} - w_{xy_j}||. \tag{10}$$

Combined with the IB loss function, the proposed CausalIB loss function can be expressed as Eq.11:

$$\mathcal{L}_{CausalIB} = \mathcal{L}_{IB} + \mathcal{L}_{causal} \ = -I\left(Z;Y\right) + \alpha I\left(Z;X\right) + \beta \sum_{i \neq j} ||w_{xy_i} - w_{xy_j}||. \tag{11}$$

Given X, Eq.11 encourages the model to learn w_{xy} undisturbed by w_{sy}, and the learned features can be generally compressed. α and β are hyperparameters.

4 Experiments

In this section, we verify the efficacy of the proposed CausalIB method by numerical experiments.

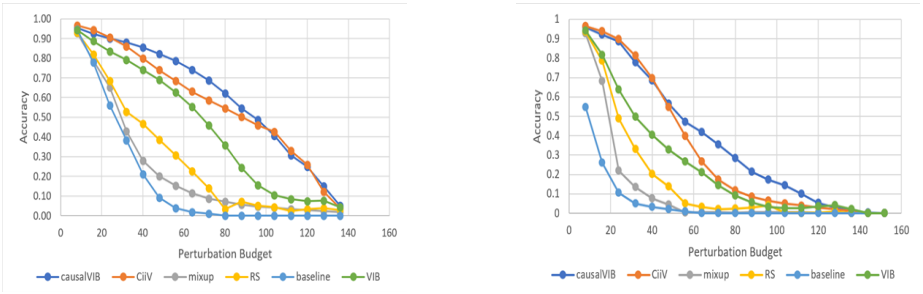
4.1 Datasets and Settings

4.1.1 Datasets

We apply the proposed CausalIB model on three benchmark datasets (MNIST LeCun et al. (1998), FashionMNIST Xiao et al. (2017), and CIFAR-10 Krizhevsky et al. (2009)) and evaluate its adversarial robustness. MNIST and FashionMNIST contain 65K handwritten digit image samples and commodity image samples respectively, and the size is 28x28. CIFAR-10 contains 60K classified image samples with a size of 32x32.

4.1.2 Training Details

The experiments in this study are carried out on Tesla P100 and tested statistically. Since our purpose is to study the IB method and the improvement of model robustness by causal inference, complex models are not chosen. In the experiments of MNIST and FashionMNIST, our basic model is a simple three-layer MLP, and the CausalIB method is applied in the last layer. In the experiments on the CIFAR-10 dataset, our basic model is AlexNet Krizhevsky et al. (2017), which also uses the CausalIB method in the last layer. All models are trained using the Adam optimizer, 100 samples per batch, and 50 epochs per training.



(a) FGSM (b) PGD

Figure 4: Unbounded attacks on MNIST that increase the budget radius ϵ from 8/255 to 152/255.

4.1.3 Details of Threat Models

The attack models in this study are FGSM, and PGD-20. In the MNIST and FashionMNIST experiments, we choose the budget radius ϵ to be 50/255. In the CIFAR-10 experiment, we choose the budget radius ϵ to be 8/255.

4.1.4 Details of other Defense Models

For adversarial training methods, we adopted two popular defenders: AT (FGSM) and AT (PGD-20) Madry et al. (2018), and these two methods are implemented with the same FGSM and PGD-20 parameters as the experiments. For the methods without the utilization of adversarial training, we investigated mixup Zhang et al. (2018), randomized smoothing (RS) Cohen et al. (2019), VIB Alemi et al. (2017) and CiiV Tang et al. (2021). Mixup is a data augmentation method that uses mixed sample data augmentation to mix images between different classes to augment the training dataset. RS uses smoothing any function into a gradient-bounded function to improve the robustness of the model, and mathematically it is well proven. VIB utilizes variational inference methods to optimize information bottlenecks and utilizes the reparameterization trick for efficient training. It has achieved the considerable results in improving model generalization and adversarial attack robustness. CiiV is a causal inference-based model, which augments the image with multiple retinal subject centers, encouraging the model to learn causal features, rather than local confusion patterns. Also, it can be combined with other methods to achieve the considerable results in improving the robustness of the model.

4.2 Robustness evaluation

We report the comparison of our CausalIB method and other methods without the utilization of adversarial training in Table 1 and Table 2, where the "clean" item in the tables represents the classification accuracy of the model on clean examples. We also provide adversarial training results at the bottom of each table for comparison. It can be seen that the proposed CausalIB shows the best overall performance in all methods without the utilization of adversarial training, demonstrating that taking into account the spurious correlation can significantly improve the adversarial robustness. In Table 1, the CausalIB shows complete superiority on MNIST, and even outperforms adversarial training methods in adversarial defense against FGSM. However, the performance of methods without the utilization of adversarial training is inferior to the performance of adversarial training on complex datasets such as CIFAR-10, which can be seen in Table 2. The advantage of the methods without the utilization of adversarial training is that they guarantee the classification accuracy of clean examples. In the experiment on FashionMNIST, the classification accuracy of clean examples of CausalIB is about 7% higher than that of adversarial training methods.

4.3 Adversarial Robustness Under Unbounded Attack

To evaluate the validity of defenders, we compare the performances of CausalIB with other methods without the utilization of adversarial training under unbounded attacking in Figure 4. Under the same budget radius ϵ, our method is better than VIB. Under the single-step attack, the classification

Table 2: The performances of white-box attack and black-box attack on CIFAR-10.

	White-box			Black-box	
Method	clean	FGSM	PGD-20	FGSM	PGD-20
Baseline Krizhevsky et al. (2017)	88.06 ± 0.87	30.48 ± 0.36	0.12 ± 0.13	31.60 ± 1.12	0.52 ± 1.26
mixup Zhang et al. (2018)	88.46 ± 1.23	42.54 ± 1.57	12.33 ± 1.98	42.54 ± 1.06	12.13 ± 0.96
RS Cohen et al. (2019)	90.40 ± 0.53	44.22 ± 0.94	22.23 ± 1.02	49.23 ± 1.42	28.23 ± 1.46
CiiV Tang et al. (2021)	89.50 ± 0.87	52.95 ± 0.43	33.62 ± 0.78	53.95 ± 0.73	34.85 ± 0.82
VIB Alemi et al. (2017)	$\mathbf{91.93 \pm 0.17}$	41.89 ± 0.98	22.81 ± 1.36	39.40 ± 0.46	23.31 ± 0.77
CausalIB	91.76 ± 0.25	$\mathbf{54.11 \pm 0.61}$	$\mathbf{35.89 \pm 0.34}$	56.36 ± 1.28	36.02 ± 1.51
AT (FGSM) Madry et al. (2018)	85.49 ± 0.62	86.32 ± 0.75	23.40 ± 0.93	-	-
AT (PGD-20) Madry et al. (2018)	85.39 ± 0.86	65.91 ± 1.02	61.20 ± 0.67	-	-

Table 3: Ablation experiments on MNIST.

Method	**clean**	**FGSM**	**PGD-20**
L_1	98.85 ± 0.20	81.07 ± 0.82	53.71 ± 0.51
L_2	97.37 ± 0.53	78.55 ± 0.93	45.62 ± 1.21
α=0/ β=0	95.11 ± 0.72	6.76 ± 0.93	0.0 ± 0.0
α=0.05/ β=0	98.41 ± 0.64	63.61 ± 1.49	38.98 ± 1.23
α=0.05/ β=0.05	98.85 ± 0.20	81.07 ± 0.82	53.71 ± 0.51
α=0.05/ β=1.0	92.69 ± 1.11	68.32 ± 1.54	20.49 ± 1.84
α=0.01/ β=0.05	95.62 ± 0.34	74.86 ± 0.64	29.65 ± 1.41

accuracy of VIB drops below 50% at 72/255, while the classification accuracy of CausalIB drops below 50% at 96/255. Similarly, in the case of iterative attack, the performance of CausalIB is about 20% better than VIB and maintains a stable rate of descent at the high budget radius. When the budget radius ϵ of the attacker was increased from 8/255 to 152/255, all performances were converged to the random guesses or even worse. Any valid defender shouldn't survive such an unbounded attack, as it allows the attacker to modify the entire image and erase all causal features.

4.4 Ablation Studies

In this section, we evaluate the performance of the proposed CausalIB under different settings and parameters on the MNIST dataset. 1) As shown in Table 3, the loss metrics with different norms are studied, where L_1 loss is better than L_2 loss; 2) Other choices of hyperparameters of the CausalIB method are reported. It can be found that β is used as a trade-off between the classification accuracy of clean examples and the performance on adversarial examples. A larger β will lead to a larger drop in the model's performance on clean examples, and when β is too large, the adversarial defense performance will also drop. In practice, the setting of hyperparameters depends on empirical experiments. 3) Our model has two parameters that need to be weighed. The IB parameter α used in our method is different from VIB. In experiments of VIB, the model performs best when α is 0.01, while the α in the CausalIB method is set as 0.05, we also provide relevant experimental results. It can be seen that causal inference further improves the robustness of the model to adversarial examples based on the IB method.

4.5 Experiments with Auto-Attack Threat Model

In this section, more detailed experiment results of the performances of Auto-Attack Croce & Hein (2020b) threat model with the budget radius $\epsilon = 8/255$ are provided. AutoAttack is a threat model that integrates various parameterless attacks, including APGD-CE Croce & Hein (2020b), APGD-DLR Croce & Hein (2020b), the black-box Square Attack Andriushchenko et al. (2020), and the FAB attack Croce & Hein (2020a). VIB Alemi et al. (2017), CiiV Tang et al. (2021), and our proposed method are evaluated on MNIST, FashionMNIST, CIFAR-10, and CIFAR-100. MLPs are evaluated structures on MNIST and FashionMNIST, and AlexNet Krizhevsky et al. (2017) is evaluated on CIFAR-10. On CIFAR-100, ResNet34 with or without regularization methods are evaluated under the Auto-Attack threat. As shown in Table 4 and Table 5, the proposed CausalIB shows the overall considerable defense performance under the Auto-Attack threat.

Table 4: The performances of Auto-Attack on MNIST and FashionMNIST.

	MNIST			**FashionMNIST**		
	clean	AA-L_{inf}	AA-L_2	clean	AA-L_{inf}	AA-L_2
MLP w/o defense	95.11	56.01	67.19	89.81	34.16	48.92
VIB Alemi et al. (2017)	98.70	89.01	92.68	94.42	65.21	76.54
CiiV Tang et al. (2021)	97.94	94.88	96.03	**94.84**	78.42	83.03
causalIB	**98.85**	**95.10**	**96.89**	94.06	**80.77**	**83.72**

Table 5: The performances of Auto-Attack on CIFAR-10 and CIFAR-100.

	CIFAR-10			**CIFAR-100**		
	clean	AA-L_{inf}	AA-L_2	clean	AA-L_{inf}	AA-L_2
CNNs w/o defense Krizhevsky et al. (2017); He et al. (2016)	88.06	0.0	0.0	**70.1**	0.0	0.0
VIB Alemi et al. (2017)	91.93	18.60	63.58	56.48	8.49	36.80
CiiV Tang et al. (2021)	89.50	26.33	**70.24**	52.15	18.34	43.36
causalIB	**91.76**	**28.7**	69.93	52.58	**20.12**	**43.89**

4.6 Gradient Obfuscation Discussion

To verify that the proposed CausalIB does not suffer from flawed or incomplete evaluations, our experiments were designed to follow a series of sanity checks:

1. The experimental results on FGSM are better than PGD-20.
2. The experimental results on black-box attacks are better than on white-box attacks.
3. The result of a weak attack (FGSM) is better than that of a strong attack (PGD-20).
4. Unbounded adversarial examples become random guessing or 0% accuracy.

These four phenomena tesify that the proposed CausalIB does not suffer from the problem of gradient obfuscation Carlini et al. (2019). The experiment of the defense against adaptive attack method Carlini & Wagner (2017b) would be illustrated in the appendix.

4.7 Shortcomings of the CausalIB method

Although our proposed CausalIB has improved adversarial defense performance, it still has the following shortcomings:

1. The causal inference method leads to a further decrease in cleaning performance relative to VIB in complex datasets due to the addition of another trade-off, which is reported in the empirical results on the CIFAR-10 dataset.
2. The settings of hyperparameters are empirical.
3. The linear causal model cannot reflect all the scenarios accurately, because DNNs are highly nonlinear. In our future work, a more complex causal graph should be modeled.
4. Currently, the adversarial training methods based on min-max optimization are state-of-the-art (SOTA) defense methods. The adversarial examples are the training samples in the optimization process of adversarial training. In contrast, the generation of adversarial examples is not essential in the training process of CausalIB. Our empirical study on CIFAR-10 also validates the performance gap between CausalIB and AT (PGD-20). In the future, it would be promising to combine the CausalIB with the framework of adversarial training.

5 Conclusion

In this paper, we address the inadequacies of the spurious correlation in the IB framework that hinder adversarial robustness. We analyze the causes of adversarial examples with the utilization of a causal graph to demonstrate that the spurious correlations between robust features and non-robust features are one of the problems to be alleviated in current IB methods. In the proposed CausalIB method, additive noises are used as an instrumental variable to estimate the causal effect. Such a method can separate robust features and non-robust features. The utilization of this method improves the performance of the IB method in adversarial defense to a certain extent. In future work, we will study whether the CausalIB method can be applied in other scenarios such as OOD image classification.

REFERENCES

Alessandro Achille and Stefano Soatto. Information dropout: Learning optimal representations through noisy computation. *IEEE*, 40(12):2897–2905, 2018.

Naveed Akhtar and Ajmal S. Mian. Threat of adversarial attacks on deep learning in computer vision: A survey. *IEEE Access*, 6:14410–14430, 2018.

Naveed Akhtar, Jian Liu, and Ajmal Mian. Defense against universal adversarial perturbations. In *CVPR*, pp. 3389–3398, 2018.

Alexander A. Alemi, Ian Fischer, Joshua V. Dillon, and Kevin Murphy. Deep variational information bottleneck. In *ICLR*, 2017.

Maksym Andriushchenko, Francesco Croce, Nicolas Flammarion, and Matthias Hein. Square attack: A query-efficient black-box adversarial attack via random search. volume 12368, pp. 484–501, 2020.

Wenjun Bai, Changqin Quan, and Zhiwei Luo. Alleviating adversarial attacks via convolutional autoencoder. In Teruhisa Hochin, Hiroaki Hirata, and Hiroki Nomiya (eds.), *IEEE/ACIS*, pp. 53–58, 2017.

Peter F. Brown, John Cocke, Stephen Della Pietra, Vincent J. Della Pietra, Frederick Jelinek, John D. Lafferty, Robert L. Mercer, and Paul S. Roossin. A statistical approach to machine translation. *Comput. Linguistics*, 16(2):79–85, 1990.

Nicholas Carlini and David A. Wagner. Towards evaluating the robustness of neural networks. In *IEEE Symposium on Security and Privacy (SP)*, pp. 39–57. IEEE Computer Society, 2017a.

Nicholas Carlini and David A. Wagner. Towards evaluating the robustness of neural networks. In *SP*, pp. 39–57, 2017b.

Nicholas Carlini, Anish Athalye, Nicolas Papernot, Wieland Brendel, Jonas Rauber, Dimitris Tsipras, Ian J. Goodfellow, Aleksander Madry, and Alexey Kurakin. On evaluating adversarial robustness. *CoRR*, abs/1902.06705, 2019.

Jeremy M. Cohen, Elan Rosenfeld, and J. Zico Kolter. Certified adversarial robustness via randomized smoothing. In *ICML*, pp. 1310–1320, 2019.

Francesco Croce and Matthias Hein. Minimally distorted adversarial examples with a fast adaptive boundary attack. pp. 2196–2205, 2020a.

Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. volume 119, pp. 2206–2216, 2020b.

Nilaksh Das, Madhuri Shanbhogue, Shang-Tse Chen, Fred Hohman, Li Chen, Michael E. Kounavis, and Duen Horng Chau. Keeping the bad guys out: Protecting and vaccinating deep learning with JPEG compression. *CoRR*, abs/1705.02900, 2017.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018.

Kevin Eykholt, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Chaowei Xiao, Atul Prakash, Tadayoshi Kohno, and Dawn Song. Robust physical-world attacks on deep learning visual classification. In *CVPR*, pp. 1625–1634, 2018.

Ian Fischer and Alexander A. Alemi. CEB improves model robustness. *Entropy*, 22(10):1081, 2020.

Ian S. Fischer. The conditional entropy bottleneck. *Entropy*, 22(9):999, 2020.

William T Freeman. The generic viewpoint assumption in a framework for visual perception. *Nature*, 368(6471):542–545, 1994.

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. Generative adversarial nets. In Zoubin Ghahramani, Max Welling, Corinna Cortes, Neil D. Lawrence, and Kilian Q. Weinberger (eds.), *NeurIPS*, pp. 2672–2680, 2014.

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In Yoshua Bengio and Yann LeCun (eds.), *ICLR*, 2015.

Robert M. Haralick, Karthikeyan S. Shanmugam, and Its'hak Dinstein. Textural features for image classification. *IEEE Trans. Syst. Man Cybern.*, 3(6):610–621, 1973.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pp. 770–778, 2016.

Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. Distilling the knowledge in a neural network. *CoRR*, abs/1503.02531, 2015.

Junho Kim, Byung-Kwan Lee, and Yong Man Ro. Distilling robust and non-robust features in adversarial examples by information bottleneck. *CoRR*, abs/2204.02735, 2022.

Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In Yoshua Bengio and Yann LeCun (eds.), *ICLR*, 2014.

Iryna Korshunova, David Stutz, Alexander A. Alemi, Olivia Wiles, and Sven Gowal. A closer look at the adversarial robustness of information bottleneck models. *CoRR*, abs/2107.05712, 2021.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. *Commun. ACM*, 60(6):84–90, 2017. doi: 10.1145/3065386. URL `http://doi.acm.org/10.1145/3065386`.

Kun Kuang, Bo Li, Peng Cui, Yue Liu, Jianrong Tao, Yueting Zhuang, and Fei Wu. Stable prediction via leveraging seed variable. *CoRR*, abs/2006.05076, 2020.

Matt J. Kusner, Joshua R. Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. *CoRR*, abs/1703.06856, 2017.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proc. IEEE*, 86(11):2278–2324, 1998.

Hyeungill Lee, Sungyeob Han, and Jungwoo Lee. Generative adversarial trainer: Defense to adversarial perturbations with GAN. *CoRR*, abs/1705.03387, 2017.

Aishan Liu, Jiakai Wang, Xianglong Liu, Bowen Cao, Chongzhi Zhang, and Hang Yu. Bias-based universal adversarial patch attack for automatic check-out. In *ECCV*, pp. 395–410, 2020.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *ICLR*, 2018.

Takeru Miyato, Andrew M. Dai, and Ian J. Goodfellow. Adversarial training methods for semi-supervised text classification. In *ICLR*, 2017.

Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi, and Pascal Frossard. Universal adversarial perturbations. In *CVPR*, pp. 86–94, 2017.

Ziqi Pan, Li Niu, Jianfu Zhang, and Liqing Zhang. Disentangled information bottleneck. In *AAAI*, pp. 9285–9293, 2021.

Giambattista Parascandolo, Mateo Rojas-Carulla, Niki Kilbertus, and Bernhard Schölkopf. Learning independent causal mechanisms. *CoRR*, abs/1712.00961, 2017.

Judea Pearl. *Causality*. Cambridge university press, 2009.

Judea Pearl. Causal inference. In Isabelle Guyon, Dominik Janzing, and Bernhard Schölkopf (eds.), *NIPS*, volume 6 of *JMLR Proceedings*, pp. 39–58, 2010.

Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(5):947–1012, 2016.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: foundations and learning algorithms*. The MIT Press, 2017.

Rafael Pinot, Laurent Meunier, Alexandre Araujo, Hisashi Kashima, Florian Yger, Cédric Gouy-Pailler, and Jamal Atif. Theoretical evidence for adversarial robustness through randomization. In *NeurIPS*, pp. 11838–11848, 2019.

Joseph Redmon, Santosh Kumar Divvala, Ross B. Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *CVPR*, pp. 779–788, 2016.

Salah Rifai, Pascal Vincent, Xavier Muller, Xavier Glorot, and Yoshua Bengio. Contractive auto-encoders: Explicit invariance during feature extraction. In Lise Getoor and Tobias Scheffer (eds.), *ICML*, pp. 833–840, 2011.

Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021.

Mahmood Sharif, Sruti Bhagavatula, Lujo Bauer, and Michael K. Reiter. Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition. In Edgar R. Weippl, Stefan Katzenbeisser, Christopher Kruegel, Andrew C. Myers, and Shai Halevi (eds.), *SIGSAC*, pp. 1528–1540, 2016.

Richard Shin and Dawn Song. Jpeg-resistant adversarial images. In *NIPS*, volume 1, 2017.

Ravid Shwartz-Ziv and Naftali Tishby. Opening the black box of deep neural networks via information. *CoRR*, abs/1703.00810, 2017.

Samarth Sinha, Homanga Bharadhwaj, Anirudh Goyal, Hugo Larochelle, Animesh Garg, and Florian Shkurti. DIBS: diversity inducing information bottleneck in model ensembles. In *AAAI*, pp. 9666–9674, 2021.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In Yoshua Bengio and Yann LeCun (eds.), *ICLR*, 2014a.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In Yoshua Bengio and Yann LeCun (eds.), *ICLR*, 2014b.

Kaihua Tang, Mingyuan Tao, and Hanwang Zhang. Adversarial visual robustness by causal intervention. *CoRR*, abs/2106.09534, 2021.

Naftali Tishby and Noga Zaslavsky. Deep learning and the information bottleneck principle. In *ITW*, pp. 1–5, 2015.

Naftali Tishby, Fernando C. N. Pereira, and William Bialek. The information bottleneck method. *CoRR*, physics/0004057, 2000.

Florian Tramèr, Alexey Kurakin, Nicolas Papernot, Ian J. Goodfellow, Dan Boneh, and Patrick D. McDaniel. Ensemble adversarial training: Attacks and defenses. In *ICLR*, 2018.

Slava Voloshynovskiy, Mouad Kondah, Shideh Rezaeifar, Olga Taran, Taras Holotyak, and Danilo Jimenez Rezende. Information bottleneck through variational glasses. *CoRR*, abs/1912.00830, 2019.

Qinglong Wang, Wenbo Guo, Kaixuan Zhang, Alexander G. Ororbia II, Xinyu Xing, C. Lee Giles, and Xue Liu. Learning adversary-resistant deep neural networks. *CoRR*, abs/1612.01401, 2016.

David Warde-Farley and Yoshua Bengio. Improving generative adversarial networks with denoising feature matching. In *ICLR*, 2017.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *CoRR*, abs/1708.07747, 2017.

Cihang Xie, Jianyu Wang, Zhishuai Zhang, Yuyin Zhou, Lingxi Xie, and Alan Yuille. Adversarial examples for semantic segmentation and object detection. In *ICCV*, pp. 1369–1378, 2017.

Cihang Xie, Yuxin Wu, Laurens van der Maaten, Alan L. Yuille, and Kaiming He. Feature denoising for improving adversarial robustness. In *CVPR*, pp. 501–509, 2019.

Han Xu, Yao Ma, Haochen Liu, Debayan Deb, Hui Liu, Jiliang Tang, and Anil K. Jain. Adversarial attacks and defenses in images, graphs and text: A review. *Int. J. Autom. Comput.*, 17(2):151–178, 2020.

Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In *BMVC*, 2016.

Cheng Zhang, Kun Zhang, and Yingzhen Li. A causal view on robustness of neural networks. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (eds.), *NeurIPS*, 2020.

Hongyi Zhang, Moustapha Cissé, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *ICLR*, 2018.

Yonggang Zhang, Mingming Gong, Tongliang Liu, Gang Niu, Xinmei Tian, Bo Han, Bernhard Schölkopf, and Kun Zhang. Adversarial robustness through the lens of causality. *CoRR*, abs/2106.06196, 2021.

Stephan Zheng, Yang Song, Thomas Leung, and Ian J. Goodfellow. Improving the robustness of deep neural networks via stability training. In *CVPR*, pp. 4480–4488, 2016.

A Appendix

A.1 Information Bottleneck

In this section, we specifically describe how VIB uses variational inference and reparameterization trick to optimize information bottleneck (IB) in DNNs.

Gievn a Markov Chain shown in Figure 5, where X represents the source random variable, Y represents the target random variable and Z represents the compression code. The purpose of IB is to maximize the objective loss function as Eq.12:

$$\mathcal{L}_{IB} = -I\left(Z;Y\right) + \alpha I\left(Z;X\right). \tag{12}$$

$I\left(Z;Y\right)$ can be written as the form of integral defined in Eq.13:

$$I(Z,Y) = \int dydzp(y,z)\log\frac{p(y,z)}{p(y)p(z)} = \int dydzp(y,z)\log\frac{p(y\mid z)}{p(y)}, \tag{13}$$

where $p(y\mid z)$ is defined by Markov Chain $Y \leftrightarrow X \leftrightarrow Z$ as Eq.14:

$$p(y\mid z) = \int dxp(x,y\mid z) = \int dxp(y\mid x)p(x\mid z) = \int dx\frac{p(y\mid x)p(z\mid x)p(x)}{p(z)}. \tag{14}$$

In fact, $p(y\mid z)$ is intractable, so $q(y\mid z)$ is needed to approximate $p(y\mid z)$. Using the fact that the Kullback Leibler divergence is always positive, Eq.15 can be induced:

$$\mathrm{KL}[p(Y\mid Z), q(Y\mid Z)] \geq 0 \Longrightarrow \int dyp(y\mid z)\log p(y\mid z) \geq \int dyp(y\mid z)\log q(y\mid z), \tag{15}$$

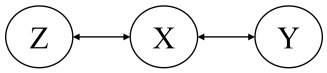


Figure 5: Markov Chain.

and hence, $I(Z,Y)$ can be induced as Eq.16:

$$\begin{aligned} I(Z,Y) &\geq \int dydzp(y,z)\log\frac{q(y\mid z)}{p(y)} \\ &= \int dydzp(y,z)\log q(y\mid z) - \int dyp(y)\log p(y) \\ &= \int dydzp(y,z)\log q(y\mid z) + H(Y). \end{aligned} \tag{16}$$

Leveraging the Markov assumption, $p(y,z)$ can be written as Eq.17:

$$p(y,z) = \int dxp\left(x,y,z\right) = \int dxp\left(x\right)p\left(y|x\right)p\left(z|x\right), \tag{17}$$

which gives us a new lower bound on the first term of our objective:

$$I(Z,Y) \geq \int dxdydzp(x)p(y\mid x)p(z\mid x)\log q(y\mid z). \tag{18}$$

Then, the expression $I\left(Z;X\right)$ can be written as Eq.19:

$$I(Z,X) = \int dzdxp(x,z)\log\frac{p(z\mid x)}{p(z)} = \int dzdxp(x,z)\log p(z\mid x) - \int dzp(z)\log p(z). \tag{19}$$

In general, while it is fully defined, computing the marginal distribution of $p\left(z\right) = \int dxp\left(z|x\right)p\left(x\right)$ might be difficult. Let $r(z)$ be a variational approximation to this marginal, the upper bound defined in Eq.20:

$$I(Z,X) \leq \int dxdzp(x)p(z\mid x)\log\frac{p(z\mid x)}{r(z)}. \tag{20}$$

Combining both of these bounds, Eq.21 can be induced:

$$\begin{aligned} I(Z,Y) - \alpha I(Z,X) \geq &\int dxdydzp(x)p(y\mid x)p(z\mid x)\log q(y\mid z) \\ &- \alpha\int dxdzp(x)p(z\mid x)\log\frac{p(z\mid x)}{r(z)} = L. \end{aligned} \tag{21}$$

$p\left(x,y\right)$ can be approximated using the empirical data distribution $p(x,y) = \frac{1}{N}\sum_{n=1}^{N}\delta_{x_n}(x)\delta_{y_n}(y)$, so that Eq.22 can be induced:

$$L \approx \frac{1}{N}\sum_{n=1}^{N}\left[\int dzp\left(z\mid x_n\right)\log q\left(y_n\mid z\right) - \alpha p\left(z\mid x_n\right)\log\frac{p\left(z\mid x_n\right)}{r(z)}\right]. \tag{22}$$

Suppose there exists an encoder of the form $p(z\mid x) = \mathcal{N}\left(z\mid f_e^{\mu}(x)\right.$, where f_e is the VIB model, then, the reparameterization trick can be utilized to do the transformation $p(z\mid x)dz = p(\epsilon)d\epsilon$, where $z = f(x,\epsilon)$ is a deterministic function of x, and the Gaussian random variable ϵ and hence the total loss function L_{IB} can be induced as Eq.23:

$$L_{IB} = \frac{1}{N}\sum_{n=1}^{N}\mathbb{E}_{\epsilon\sim p(\epsilon)}\left[-\log q\left(y_n\mid f\left(x_n,\epsilon\right)\right)\right] + \alpha\,\mathrm{KL}\left[p\left(Z\mid x_n\right),r(Z)\right], \tag{23}$$

and the former conditional entropy can be approximated using cross entropy.

Figure 6: Visualization Examples 1 on the dataset of FashionMNIST Xiao et al. (2017).

Figure 7: Visualization Examples 2 on the dataset of FashionMNIST Xiao et al. (2017).

Table 6: More detailed ablation experiments.

Method	**clean**	**FGSM**	**PGD-20**
α=0.01/ β=0.01	95.32 ± 0.98	73.50 ± 0.55	34.41 ± 0.85
α=0.01/ β=0.2	96.24 ± 1.27	71.27 ± 0.74	33.90 ± 1.36
α=0.01/ β=1.0	90.35 ± 0.91	60.23 ± 1.52	28.73 ± 0.89
α=0.05/ β=0.01	97.46 ± 0.47	78.40 ± 0.30	50.42 ± 0.28
α=0.05/ β=0.2	96.39 ± 0.76	76.52 ± 0.77	49.69 ± 0.61
α=0.05/ β=0.5	93.56 ± 0.58	69.09 ± 1.39	29.30 ± 1.25
α=0/ β=0.05	94.62 ± 0.56	42.99 ± 0.72	16.88 ± 0.88
α=0/ β=0.2	95.28 ± 0.64	38.49 ± 0.59	15.33 ± 0.96
α=0/ β=1.0	92.52 ± 0.45	26.20 ± 0.87	12.16 ± 1.68

A.2 VISUALIZATION

To verify that the proposed CausalIB has a better performance than VIB in feature extraction, we make the visualization of adversarial examples as shown in Figure 6 and Figure 7.

In Figure 6 and Figure 7, the visualization of clean examples are shown in the first row. The visualization of adversarial examples attacking the CausalIB model and the visualization of adversarial examples attacking the VIB model are shown in the second and third rows. The adversarial examples attacking the VIB model are relatively blurry, while the adversarial examples attacking the CausalIB model have much clearer structural details. This means that the causal inference method could lead the model to learn more structural information. As a result, the adversary must add larger perturbations with the erasion of the structural patterns to fool the CausalIB model.

A.3 MORE DETAILED ABLATION EXPERIMENTS

In this section, more detailed ablation experiment results are provided to show the impact of different choices of the two hyperparameters in our method. Many adversarial defense methods have a trade-off between clean accuracy and adversarial robustness. When the weight of the regularization term is too

Table 7: Experiment results on CIFAR-100.

	clean	FGSM	PGD-20	CW-20
ResNet w/o defense He et al. (2016)	**70.10**	3.18	0.0	0.0
VIB Alemi et al. (2017)	56.48	21.48	12.77	10.04
CiiV Tang et al. (2021)	52.15	**32.51**	23.19	22.05
causalIB	52.58	32.06	**24.85**	**22.52**

Table 8: The adversarial robustness under the C&W attacks on different datasets.

	MNIST	FashionMNIST	CIFAR-10	CIFAR-100
no defense	88.92	52.65	1.28	0
VIB Alemi et al. (2017)	94.45	86.05	18.04	10.45
CiiV Tang et al. (2021)	96.08	90.15	32.59	**24.86**
causalIB	**96.55**	**90.77**	**35.44**	23.93

Table 9: Comparison Experiments of AlexNet, ResNet and WideResNet on clean data and FGSM.

	clean			**FGSM**		
	AlexNet	ResNet	WideResNet	AlexNet	ResNet	WideResNet
No defense	88.06	91.16	**92.28**	30.48	39.12	**39.85**
VIB Alemi et al. (2017)	91.93	**92.65**	92.45	41.89	44.2	**46.79**
CiiV Tang et al. (2021)	89.50	91.53	**92.05**	52.95	56.18	**58.25**
causalIB	91.76	**92.91**	92.66	54.11	57.52	**58.75**

Table 10: Comparison Experiments of AlexNet, ResNet, and WideResNet on PGD-20 and AutoAttack-L_{inf}.

	PGD-20			**AA-L_{inf}**		
	AlexNet	ResNet	WideResNet	AlexNet	ResNet	WideResNet
No defense	0.12	0.75	**2.59**	0	0	0
VIB	22.81	**27.47**	27.38	18.60	25.60	**26.58**
CiiV	33.62	42.30	**43.66**	26.33	33.14	**34.20**
causalIB	35.89	**43.11**	43.05	28.7	34.32	**34.66**

large, the learning of the classifier could be hurt. As shown in Table 6, it can be seen that a too-large β or a too-small β will lead to a decrease in the defense performance of the model against adversarial examples.

Currently, the settings of the hyperparameters are empirical, which is a shortcoming of the proposed CausalIB. In the future, a more reasonable hyperparameter optimization method would be studied.

A.4 Experiments on CIFAR-100

In this section, results on the CIFAR-100 dataset are provided. The utilized model is ResNet34 He et al. (2016). VIB Alemi et al. (2017), CiiV Tang et al. (2021), and CausalIB are implemented on the ResNet model. The default ResNet model has no defense (no additional regularization method). As can be seen in Table 7, causalIB achieves the considerable adversarial robustness, but the classification accuracy on the clean data has a large drop, which is in line with the general trend of adversarial defense methods.

A.5 Experiment of Adaptive C&W Attack

In this section, experiment results of the performances under the adaptive C&W attacks Carlini & Wagner (2017a) are provided. The feature extraction model used on MNIST and FashionMNIST is a MLP, AlexNet Krizhevsky et al. (2017) is used on the CIFAR-10 dataset, and ResNet34 He et al. (2016) is used on the CIFAR-100 dataset. The iteration step of C&W attack is 20. As can be seen in Table 8, the causalIB method still achieves the considerable adversarial robustness under the C&W attack, which is consistent with the test results under the other attacks.

A.6 Comparison Experiments of AlexNet, ResNet, and WideResNet

In this section, the experiment result of adversarial robustness influenced by different CNN structures on the CIFAR-10 dataset would be illustrated. Table 9 and Table 10 shows the comparative test

results of the CNN models including AlexNet Krizhevsky et al. (2017), ResNet34 He et al. (2016), and WideResNet-34-10 Zagoruyko & Komodakis (2016). It can be seen that our proposed method could achieve even better adversarial robustness in the residual structures He et al. (2016); Zagoruyko & Komodakis (2016), although the CausalIB has promoted adversarial robustness on the AlexNet model Krizhevsky et al. (2017).