Learning Disentangled Representations for Environment Inference in Out-of-distribution Generalization

Dongqi Li[1]
dongqili@bjtu.edu.cn
Zhu Teng[1]
zteng@bjtu.edu.cn
Qirui Li[2]
lqr@afctech.com.cn
Ziyin Wang[2]
wzy@afctech.com.cn
Baopeng Zhang ✉[1]
bpzhang@bjtu.edu.cn
Jianping Fan[3]
jfan1@Lenovo.com

[1] School of Computer and Information Technology
Beijing Jiaotong University
Beijing, China

[2] AFCtech
Beijing, China

[3] AI Lab
Lenovo Research
Beijing, China

Abstract

Machine learning models often generalize poorly to out-of-distribution (OOD) data as a result of relying on features that are spuriously correlated with the label during training. To deal with this issue, environment inference methods are proposed to learn invariant predictors without environment labels. Previous environment inference works often employ Empirical risk minimization (ERM) as a reference model for environment inference because they assume ERM captures spurious features due to its inductive bias. In this work, we show that using ERM as a reference model has a pitfall in environment inference because it does not effectively capture spurious features. To this end, we propose a disentangled representation method by designing a variational auto-encoder to capture spurious features for environment inference without environment labels. Extensive experiments demonstrate that the proposed method outperforms other methods on both synthetic and real-world datasets.

1 Introduction

In conventional machine learning, training data is assumed to be independently and identically distributed (*i.i.d.*) as the test data. However, this assumption can hardly be satisfied when the test distribution deviates from the training distribution in real cases. Machine learning models may suffer from a sharp drop under a distributional shift, which is systematically discussed as the out-of-distribution (OOD) generalization problem. The dependence on spurious correlations that are prone to be different across environments has been recognized as a

major cause of such failure [1, 29]. For example, deep neural networks (DNNs) in the image recognition task may rely on the backgrounds that are regarded as spurious features to predict results instead of core features that are truly causality correlated to the labels [26, 32]. Experimental evidence [6, 25] reflects that DNNs trained with empirical risk minimization (ERM [28]), the most commonly used training method, are prone to preferring to retrieve spurious features in training data.

A notable line of research on the OOD generalization problem is learning features with invariant conditioned label distribution across training environments [1, 16, 18, 23], which has been termed as invariant risk minimization (IRM). In contrast to ERM, inspired by causality [21], IRM aims to learn a stable correlation across multiple training environments, which expects to elicit an invariant predictor. However, invariant learning methods usually require predefined environment[1] labels, which are not easily accessible.

To handle this issue, [6] proposed EIIL, enabling invariant learning when environment labels are unavailable. It conducts environment inference (EI) using the representation learned by a *reference model*, which maximizes the penalty term in the IRM framework. It is shown that the reference model can capture spurious features, which makes the environment inference possible. However, EIIL employs an ERM model as the reference model which cannot capture effective spurious features. In other words, ERM-based models limit the performance of environment inference. In this paper, we argue that if the quality of spurious features captured by the reference model is insufficient, which means the environment inference step receives incomplete environment partition information, invariant learning methods may fail with improper environments inferred by environment inference.

To this end, we propose a disentangled representation learning method to encode effective spurious features via variational auto-encoder(VAE) [15]. It improves the quality of environment inference and the performance of invariant learning. Furthermore, we demonstrate that insufficiently learned spurious features may lead to a failure in environment inference. Our contributions are summarized in three-fold:

- ERM-based environment inference methods assume ERM captures spurious features due to its inductive bias. To our observation, the quality of the learned spurious features has a strong influence on the performance of environment inference, and we argue that ERM-based methods cannot acquire sufficient spurious features in some cases.

- Based on this observation, we propose a disentangled representation learning method to obtain spurious features for environment inference methods. It is constructed by a variational auto-encoder.

- We conduct comprehensive experiments, including both synthetic and real-world datasets such as Colored MNIST[I], Colored MNIST[II], and CelebA. Experimental results demonstrate that the proposed method outperforms existing state-of-the-art methods by a large margin.

[1] We use the terms "environments", "groups", and "domains" interchangeably to refer to the hierarchical structure of a dataset.

2 Related Work

2.1 Invariant Learning and Environment Inference

IRM [1] proposes a training objective for learning the invariant representations, under the assumption that the Bayes optimal conditional $P(y|\Phi(x))$ remains invariant across environments with overlap. Some IRM variants are proposed like the variance of penalization or loss gradients across training environments [8, 16, 22]. However, recent works [12, 23] revealed the theoretical pitfall of IRM in the non-linear setting and some other scenarios. Practical works also investigate that the performance of IRM relies on model size [18], dataset type [8] etc. Besides, IRM also requires a good environment prior beforehand. EIIL [5] proposes first partitioning the training data into environments called the majority group and minority group and using a reference model before the environment-based invariant learning. However, including some similar works without using environment labels [4, 31], they use ERM as a reference model, which can not capture good spurious features for splitting the training data.

2.2 Other Out-of-distribution Generalization Methods

Other out-of-distribution generalization methods also have been proposed. Besides the invariant learning methods, another important line for generalizing data under distributional shift is distributionally robust optimization (DRO) [7, 10]. DRO methods propose to optimize the worst-case error over a set of distributions that are required to cover the test distribution. And a notable method called group DRO [24] optimizes the worst-case error by sharing importance weights across training examples and can provide a reasonable region for robust optimization. However, as with most invariant learning methods, group DRO also requires the group annotation and label of each data sample for group partition. Causality is also related to the OOD generalization problem. [21] proposes Invariant Causal Prediction(ICP) to utilize the invariance property to identify the direct cause of the target.

2.3 Disentangled Representation Learning

The goal of disentangled representation learning is to learn representations where distinct and explainable factors of variations in data are separated [2], which are expected to potentially benefit downstream tasks. Unsupervised VAE-based methods [9, 14] emphasize learning disentangled representation by encouraging independence among the factors in latent variables in an unsupervised way. Despite the success in disentangled representation learning in some datasets, [20] challenges some assumptions that unsupervised disentangled representation methods lack inductive bias and thus the model identifiability cannot be guaranteed. It also questions whether unsupervised disentangled representation learning can benefit downstream task performances, including out-of-distribution generalization. Similar works to our VAE-based method are [11, 27]. They use supervised information as prior for model identification and provide thinking for improving downstream task performances in the VAE-based method in a supervised manner. However, the domain labels are still necessary to extract the domain latent representation.

3 Pitfall in ERM-based Environment Inference

In this section, we first give a brief description of the environment inference and then elaborate on the observation of the pitfall in environment inference when the reference model cannot capture effective spurious features.

3.1 Preliminaries

Throughout the paper, We denote X and Y as random variables and x, y as corresponding samples and labels. We assume that there is a set of multiple environments, $\mathcal{E}_{supp}$, where the data can be extracted from. We can access to a collection of training environments, $\mathcal{E}_{tr} \subset \mathcal{E}_{supp}$ and each training environment are denoted as $e \in \mathcal{E}_{tr}$. And we indicate the environment index e with the variable X^e and Y^e. Let $\mathcal{X}$ and $\mathcal{Y}$ be the space of X and Y. Our goal is to learn a function $f : \mathcal{X} \to \mathcal{Y}$, which predicts Y conditioned on X. For DNNs, f can consist of a classifier and a feature extractor parameterized with w and Φ. l is the loss function. The goal of OOD generalization is to minimize the risk of the worst environment $R^e = \mathbb{E}_{X^e,Y^e}[l(f(X^e),Y^e)]$.

To define the OOD generalization problem with invariant learning, we consider the features extracted by feature extractor Φ. We assume that features that are spurious to labels (like backgrounds in images) are called *spurious features* which are denoted as X_s, and other features that are truly correlated to labels (like object shapes) are called *core features* which is denoted as X_c. Our issue is to find a model whose features $\Phi(X)$ are focused on X_c and discards X_s. To handle the issue, IRM aims to solve it by a bi-level optimization.

3.2 IRMv1 Method and Environment Inference

Since IRM is a challenging bi-level optimization problem. IRM can be instantiated into the practical version, called IRMv1 [1]. It can be formulated by Eq. 1.

$$\min_{\Phi:\mathcal{X}\to\mathcal{Y}} \sum_{e\in\mathcal{E}_{tr}} R^e(\Phi) + \lambda \cdot \left\|\nabla_{w|w=1.0} R^e(w\cdot\Phi)\right\|^2 \tag{1}$$

In this formulation, additional information is required to regularize the training process and learn invariant predictors. λ is a constant for adjusting the two terms of the formulation. The classifier is "dummy" when $w = 1.0$, which is assumed as a linear model. We continue to make this assumption in the following discussion. $w \cdot \Phi$ indicates a classifier via linear weighting (w) on the features extracted by Φ. IRMv1 can achieve better performance under distributional shift compared with ERM.

However, IRMv1 requires environment labels and they are not always available. To deal with this, EIIL [5] proposes an environment inference step to produce the required environment labels. It employs a reference model Φ, which is different from the invariant learning model, to maximize the IRM penalty term as described in Eq. 2, where $\widetilde{R}^e(\Phi,q) = \frac{1}{N}\sum_i q_i(e) l(\Phi(x_i),y_i)$ and q indicates soft partition-predictions. Similar to IRMv1, $\overline{w}$ is a constant scalar multiplier of 1.0 for each output dimension. This soft assignment is utilized to produce hard environment labels that are then employed in invariant learning.

$$C^{EI}(\Phi,q) = \left\|\nabla_{\overline{w}} \widetilde{R}^e(\overline{w}\cdot\Phi,q)\right\|^2 \tag{2}$$

Reference model	IP↑	SFS↑	Acc(%)↑
ERM	0.0007	0.32	18.8
Ours	**0.1052**	0.75	**63.8**
w/ EnvLabels	0.0025	**1.00**	57.3

Table 1: The impact of invariant penalty and spurious feature score of reference model used in EIIL on the accuracy of invariant learning.

C^{IL} is the objective of IRMv1 in Eq. 1. The steps of the environment inference by EIIL can be summarized as follows:

1. Define a reference model $\widetilde{\Phi}$.
2. Fix $\Phi \leftarrow \widetilde{\Phi}$ and maximize C^{EI} to generate environment labels via $q^* = \arg\max_q C^{EI}(\widetilde{\Phi}, q)$.
3. Optimize the Invariant Learning (IL) objective C^{IL} to obtain the final model through $\Phi^* = \arg\min_\Phi C^{IL}(\Phi, q)$.

3.3 Observations

EIIL works under the assumption that the reference model $\widetilde{\Phi}$ focuses only on the spurious features, which suggests $\widetilde{\Phi} = \Phi_{sp}$. It employs ERM as the reference model, which assumes that $\widetilde{\Phi} \approx \Phi_{ERM}$. We argue that the spurious features captured by ERM are insufficient for environment inference. To evidence this, we design an experiment on the dataset of MNIST dataset [1] and evaluate the performance by the metrics of the *Spurious Feature Score* (SFS) and *Invariant Penalty* (IP), which are defined as follows.

Dataset. Colored MNIST is a modified MNIST dataset [17], where ten distinct colors are injected in the foreground of each digit individually to create a spurious correlation. The ratio of samples conflicted with the spurious correlation is set to 1%.

Spurious Feature Score. We design Spurious Feature Score as a metric to measure the quality of the spurious features captured by reference models. The closer to 1 in SFS represents the reference model can capture more sufficient and informative spurious features, while the closer to 0 is less. SFS directly measures spurious features by verifying whether the spurious feature is changed when replacing one kind of core feature with another. In practice, our test environment is usually designed with anti-spurious correlation, so we can approximate SFS by calculating the classification accuracy of spurious features.

Invariant Penalty. We can also define another metric by Eq. 2. IP is also the penalty of IRMv1, which measures the invariant predictor among multiple environments. Different from SFS, IP measures how much we use spurious features to split the environments. A larger IP indirectly indicates better spurious features we capture. [5] proves the upper bound of Eq. 2, so we can regard it as a metric despite the requirement of an optimization of q.

Before explaining our observation, we first denote that ERM has two identities in our paper: 1) an optimization method compared with other OOD generalization methods; 2) a reference model used in EIIL. We train two reference models on the Colored MNIST dataset, including the ERM and our model (see more details in Section 4).

We employ these reference models in the EIIL framework and compare the performance in EIIL with the different reference models. Besides, we also conduct a reference model trained with environment labels named "w/ Envlabels" for more analysis. Table 1 shows that using ERM as a reference model has a pitfall in environment inference and how it produces

the failure. We can notice that ERM models have low SFS and IP which means the spurious features Φ_{ERM} are far from the ideal spurious features $\widetilde{\Phi}$. It shows that insufficient spurious feature is correlated to the accuracy (Acc%) after invariant learning. Even if ERM has the inductive bias to learn spurious features, it can still not learn good spurious features for environment inference. "w/ Envlabels" model shows that it is ERM that makes the pitfall, and EIIL can still work when provided sufficient spurious features. And insufficient spurious features cause improper environment partition, which means Environment Invariance Constraint(EIC) [5] may not hold.

4 Method

We have found that using ERM as a reference model might lead to a failure in environment inference since the spurious features it captures are insufficient in some cases. In this section, we propose a disentangled representation model to exploit more effective spurious features and identify less entangled core features for a reference model without required environment labels. Our disentangled representation method is based on VAE and we discuss the OOD problem on the image classification task.

4.1 Learning Method

We denote $\mathcal{D}_{tr} = \{x_i, y_i\}_{i=1}^{N}$ as the training dataset. A variational auto-encode (VAE), which contains an encoder network E_ϕ and a decoder network D_θ, models the input variable x with a latent variable: $p(x,z) = p(z)p(x|z)$. The prior $p(z)$ is assumed $\mathcal{N}(0,I)$ and the likelihood $p(x|z)$ is implicitly modeled by the decoder $D_\theta(z) = (\mu_\theta(z), diag(\sigma^2_\theta(z)))$ as $\mathcal{N}(x|\mu_\theta(z), diag(\sigma^2_\theta(z)))$. Directly maximizing the likelihood $p(x)$ is intractable due to the complicated computation of the integral. Instead, the VAE model employs an encoder $E_\phi(x) = (\mu_\phi(x), diag(\sigma^2_\phi(x)))$ as a variational approximation $q_\phi(z|\mu_\phi(x), diag(\sigma^2_\phi(x)))$ and maximizes the Evidence Lower BOund(ELBO) as described in Eq. 3.

$$ELBO(\phi,\theta,x) = \mathbb{E}_{z\sim q_\phi}\left[log\frac{p_\theta(x,z)}{q_\phi(z|x)}\right] \tag{3}$$

VAE provides a method to learn the features of data samples via latent space. We consider the latent variable z as a concatenation of two components: the core feature z_c and the spurious feature z_s, denoted as $z = [z_c, z_s]$. Given the input x, the objective here is to learn the encoder distribution $q_\phi(z_c, z_s|x)$. Our target is to dissociate z_s from the confounded factor z_c. To facilitate this, we construct a classifier head C_φ combined with the VAE encoder network to learn only z_s. The distribution of $\mathcal{D}_{tr}$ is formulated as $p(z_c, z_s, x, y) = p(z_c, z_s)p(x|z_c, z_s)p(y|x)$ and $p(y|x)$ is estimated by the classifier head C_φ that outputs the class logits. As the variance learned by the VAE encoder network is usually small, C_φ can take the mean vector $\mu^s_\phi(x)$ to encode the spurious features as input. The VAE decoder employs both z_s and z_c for reconstruction. To summarize, the total learning objective can be formulated in Eq. 4, where λ is a constant to balance the weight of the VAE term and the classification term in training.

$$\mathcal{L}(\phi,\theta,\varphi) = -\sum_{(x,y)\in\mathcal{D}_{tr}} ELBO(\phi,\theta,x) + \lambda \cdot log\, p_{\phi,\varphi}(y|x) \tag{4}$$

The classifier head only takes z_s as input and the VAE decoder takes both z_s and z_c as input. The reconstruction of x contains the information of both core features and spurious features because learning to generate the observed data forces the network to model every variation in the input data x. The classifier network with ground truth has an **inductive bias** that prefers using spurious features to make predictions in training. Different from ERM, which also tends to depend on the spurious feature for prediction due to the inductive bias, we notice that a **smaller dimension** of z_s can make z_s a more sufficient spurious feature. The details of which can be seen in Proposition 1.

4.2 Theoretical Analysis

In this section, we discuss the theoretical basis of the disentangled representations extracted by our VAE-based method and show that there is a theoretical guarantee that we can extract spurious representations under an assumption with our method.

We make a simple assumption that the auxiliary variable z_s is independent of z_c. The assumption is easy to achieve because the distribution of spurious features is estimated approximatively by the classification stage and the distribution of the core feature is estimated by the generation stage. Besides, to demonstrate that our assumption above can be satisfied at a high probability, we employ experiments in section 5.3.

Proposition 1 *$\forall x_i \in \mathcal{D}$, $z_i \in \mathbb{R}^d$ is the corresponding latent variable, and $p \in [0,1]$, then $z_i^s \in \mathbb{R}^{\lfloor pd \rfloor}$ and $z_i^c \in \mathbb{R}^{\lfloor (1-p)d \rfloor}$. If $p < 0.5$, then z_i^s is seen as a spurious feature and z_i^c is a core feature. z_i is the concatenation of z_i^s and z_i^c. $z_i^s \perp\!\!\!\perp z_i^c \mid x_i, y_i$ holds, where y_i is core w.r.t the label.*

As shown in [13] that in nonlinear independent component analysis (Nonlinear ICA), given auxiliary variables u, the components of $\mathbf{z} = (z_1, z_2, ..., z_n)$ are *identifiable*, that is, we can obtain a unique solution when capturing independent factors. In Eq. 4, we assign the spurious features z_s to the supervised prior of class labels when training, and we can approximately replace z_s to auxiliary variables u. We restrict the auxiliary variable u with supervised class labels y and the z_c represents its information in an unsupervised manner. Nonlinear ICA leads to identifiability based on the VAE model because it assumes a conditional prior distribution over the joint distribution $p_\theta(x, z_c|u)$, which can be decomposed into $p_\theta(x|z_c,u)p(z_c|u) = p_\theta(x|z_c,u)$ according to our assumption. It shows that our VAE-based model is identifiable given the auxiliary variable u. Thus we simply prove that when the independence between z_c and z_s holds, both features are identifiable in our VAE-based model. For the inferred features can be identifiable, we can have a theoretical guarantee to capture spurious features by only approaching the core feature labels of a dataset.

5 Experiments

In this section, we empirically analyze the effectiveness of our method on both synthetic and real-world datasets compared with the existing methods. We also show that our method can extract disentangled representations in the dataset with spurious correlation. We assume the soft assignment q produces binary environments and parameterize the q as a vector of probabilities for each example in the training data.

Method	Env Labels	Train Accs	Test Accs
ERM	✘	**89.5 ± 1.1**	26.7 ± 2.8
EIIL	✘	75.2 ± 1.1	59.3 ± 5.5
Ours	✘	81.4 ± 0.3	**69.7 ± 1.8**
IRM	✔	76.7 ± 0.9	70.5 ± 2.2

Table 2: The performance evaluated on Colored MNIST[I].

Method	Test Accs
ERM	56.9
EIIL	45.2
Ours	**63.8**
IRM(Oracle)	72.7

Table 3: The performance evaluated on Colored MNIST[II].

Method	Train Accs	Test Accs
ERM	**93.1 ± 5.9**	71.4 ± 3.4
DisEnt	15.5 ± 1.6	37.4 ± 2.4
ZIN	90.6 ± 0.3	70.8 ± 1.6
EIIL	78.2 ± 12.1	61.9 ± 4.9
Ours	80.4 ± 2.6	**74.4 ± 1.8**
IRM(Oracle)	83.2 ± 1.1	78.7 ± 0.8

Table 4: The performance evaluated on CelebA.

5.1 Datasets

Colored MNIST[I] was originally introduced in the IRM paper [1]. It is a synthetic dataset for a binary classification task. And the color is introduced as a spurious correlation because DNNs can employ colors to predict labels. In particular, two training environments have a ratio of the correlation $\{0.8, 0.9\}$ between the digit and color while the test environment has a ratio of 0.1. And label noise is applied by flipping y with a probability of 0.20.

Colored MNIST[II] was introduced by biased dataset [17]. It uses ten distinct colors and injects each color into the foreground of each digit to create color bias. The training data can be divided into bias-aligned samples and bias-conflicting samples. The ratio of bias-conflicting samples is 1%. The difference from Colored MNIST[I] is it has a lower ratio of anti-correlates but more sample diversity in bias-conflicting samples.

CelebA is a real-world dataset that predicts *Smiling* based on the image from CelebA [19, 30], and is constructed to make a spurious correlation between the target and *Gender*.

5.2 Experimental Results and Analysis

We first introduce the three datasets with spurious correlation under distributional shift. And we discuss the implementation details in our experiments. We compare our method with several existing methods: ERM, IRM [1], and EIIL [5]. In the CelebA dataset, we add extra two methods: DisEnt [17] and ZIN [30] for more comparison. We provide the environment partition to IRM and other methods work without environment labels. EIIL and our method use different reference models for extracting spurious features.

In Colored MNIST[I], Table 2 reports the training accuracy and the test accuracy. We can notice that ERM has a top train accuracy but a violently low test accuracy because it relies on spurious features to predict the labels and suffers from a sharp drop in the anti-correlated test environment under distributional shift. The performance of our gains is insignificant because our reference model makes more approximate Φ to Φ_{Sp} than Φ_{ERM}. And our method can stably extract spurious representation across five restarts according to the low variance in the test accuracy. The test accuracy of our method is close to IRM which requires handcraft

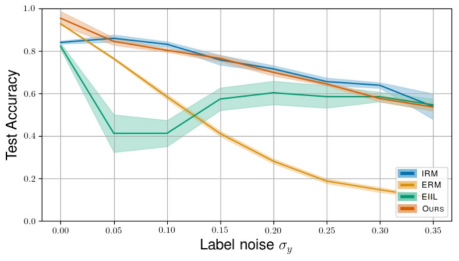

Figure 1: The test accuracy of Colored MNIST[I] with varying label noise.

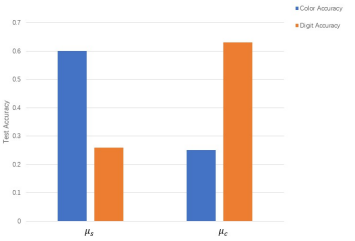

Figure 2: Color and digit classification accuracy evaluation for latent variables.

environment labels, which shows that our method with environment inference can make a reasonable environment partition compared with hand-craft labels.

Figure 1 shows the results of test accuracy when the label noise of Colored MNIST[I] changes. We can notice that ERM fails to generalize when the label noise rises because ERM captures spurious features under high label noise. We find that EIIL has a sharp drop under low label noise ($\sigma_y < 0.15$) and returns to a similar level with IRM and our method under high label noise($\sigma_y > 0.25$). It shows ERM used as a reference model in EIIL is less reliable to spurious features under low label noise and makes the environment partition a pitfall. Despite the similar performance to IRM and our method under high label noise, EIIL keeps an unstable prediction compared to other methods because it shows that EIIL has a large variance in the test accuracy. Our method matches and sometimes exceeds the performance of IRM without environment labels when label noise is changing.

In ColorMNIST[II], we can find that the ERM model performs well than where in Colored MNIST[I] though the ratio of anti-correlates is lower because there are more samples with diversity. EIIL emerges as the worst-performing model because it encounters a pitfall for which the reference model uses ERM. The performance of IRM goes much further than our method perhaps because it is still difficult for environment inference to detect the bias-conflicting samples of the dataset.

In CelebA, Table 4 shows the results. We can find that IRM still achieves the best performance of other methods. And notably, the accuracy of ERM is only lower about 3% than that of IRM. We suppose that ERM can still work well in the real-world dataset under the spurious correlation because there is more noise in spurious features and less inductive bias. EIIL still emerges as the worst-performing model because ERM captures insufficient spurious features.

5.3 Disentangled Representation Analysis

To measure whether we can obtain the disentangled representation extracted by our model, a simple way is to classify the representations, examining whether our representations benefit the downstream classification task. In Colored MNIST[II], we already have the encoder network and the classifier head $C^s_{\varphi_1}$ for spurious features μ_s (the mean vector of z_s). We train an extra classifier head $C^c_{\varphi_2}$ on the core feature μ_c while the parameters of the encoder are frozen. Additionally, both classifiers are trained with **core labels**. Then we employ the features for evaluation on both spurious feature labels and core feature labels. Figure 2 reports the accuracy on both μ_s and μ_c. We find that the μ_s has a high accuracy on color labels

which denotes spurious features and a low accuracy on digits which denotes core features. And μ_c has the opposite. It shows that our model can extract disentangled representation under the spurious-correlation shift.

6 Conclusion

In this paper, we reveal that ERM used as a reference model in environment inference can not capture sufficient spurious features in some cases. We observe that insufficient spurious features extracted by ERM can have environment inference a pitfall and make the follow-up invariant learning fail to generalize. Based on this observation, we propose a VAE-based disentangled representation method to extract better spurious features than ERM. Experimental results verify our analysis and demonstrate the improved performance of our method over existing methods. In future work, we will focus on the algorithm's robustness for environment inference in OOD problems under different distribution shifts.

Acknowledgments This work was supported by the Natural Science Foundation of China (61972027), the Fundamental Research Funds for the Central Universities of China (2022JBMC009), and the Beijing Municipal Natural Science Foundation (Grant No. 4212041).

References

[1] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

[2] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013.

[3] Shiyu Chang, Yang Zhang, Mo Yu, and Tommi Jaakkola. Invariant rationalization. In *International Conference on Machine Learning*, pages 1448–1458. PMLR, 2020.

[4] Yimeng Chen, Ruibin Xiong, Zhi-Ming Ma, and Yanyan Lan. When does group invariant learning survive spurious correlations? In *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=ripJhpwlA2v.

[5] Elliot Creager, Jörn-Henrik Jacobsen, and Richard Zemel. Environment inference for invariant learning. In *International Conference on Machine Learning*, pages 2189–2200. PMLR, 2021.

[6] Pim De Haan, Dinesh Jayaraman, and Sergey Levine. Causal confusion in imitation learning. *Advances in Neural Information Processing Systems*, 32, 2019.

[7] John C Duchi and Hongseok Namkoong. Learning models with uniform performance via distributionally robust optimization. *The Annals of Statistics*, 49(3):1378–1406, 2021.

[8] Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=lQdXeXDoWtI.

[9] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-VAE: Learning basic visual concepts with a constrained variational framework. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=Sy2fzU9gl.

[10] Zhaolin Hu and L Jeff Hong. Kullback-leibler divergence constrained distributionally robust optimization. *Available at Optimization Online*, pages 1695–1724, 2013.

[11] Maximilian Ilse, Jakub M Tomczak, Christos Louizos, and Max Welling. Diva: Domain invariant variational autoencoders. In *Medical Imaging with Deep Learning*, pages 322–348. PMLR, 2020.

[12] Pritish Kamath, Akilesh Tangella, Danica Sutherland, and Nathan Srebro. Does invariant risk minimization capture invariance? In *International Conference on Artificial Intelligence and Statistics*, pages 4069–4077. PMLR, 2021.

[13] Ilyes Khemakhem, Diederik Kingma, Ricardo Monti, and Aapo Hyvarinen. Variational autoencoders and nonlinear ica: A unifying framework. In *International Conference on Artificial Intelligence and Statistics*, pages 2207–2217. PMLR, 2020.

[14] Hyunjik Kim and Andriy Mnih. Disentangling by factorising. In *International Conference on Machine Learning*, pages 2649–2658. PMLR, 2018.

[15] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[16] David Krueger, Ethan Caballero, Joern-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Dinghuai Zhang, Remi Le Priol, and Aaron Courville. Out-of-distribution generalization via risk extrapolation (rex). In *International Conference on Machine Learning*, pages 5815–5826. PMLR, 2021.

[17] Jungsoo Lee, Eungyeup Kim, Juyoung Lee, Jihyeon Lee, and Jaegul Choo. Learning debiased representation via disentangled feature augmentation. *Advances in Neural Information Processing Systems*, 34:25123–25133, 2021.

[18] Yong Lin, Hanze Dong, Hao Wang, and Tong Zhang. Bayesian invariant risk minimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16021–16030, 2022.

[19] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, pages 3730–3738, 2015.

[20] Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Raetsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *international conference on machine learning*, pages 4114–4124. PMLR, 2019.

[21] 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.

[22] Alexandre Rame, Corentin Dancette, and Matthieu Cord. Fishr: Invariant gradient variances for out-of-distribution generalization. In *International Conference on Machine Learning*, pages 18347–18377. PMLR, 2022.

[23] Elan Rosenfeld, Pradeep Ravikumar, and Andrej Risteski. The risks of invariant risk minimization. *arXiv preprint arXiv:2010.05761*, 2020.

[24] Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. In *International Conference on Learning Representations (ICLR)*, 2020.

[25] Shiori Sagawa, Aditi Raghunathan, Pang Wei Koh, and Percy Liang. An investigation of why overparameterization exacerbates spurious correlations. In *International Conference on Machine Learning*, pages 8346–8356. PMLR, 2020.

[26] Sahil Singla and Soheil Feizi. Salient imagenet: How to discover spurious features in deep learning? In *International Conference on Learning Representations*, 2021.

[27] Xinwei Sun, Botong Wu, Xiangyu Zheng, Chang Liu, Wei Chen, Tao Qin, and Tie-Yan Liu. Recovering latent causal factor for generalization to distributional shifts. *Advances in Neural Information Processing Systems*, 34:16846–16859, 2021.

[28] Vladimir Vapnik. *Statistical learning theory*. Wiley, 1998.

[29] Nanyang Ye, Kaican Li, Haoyue Bai, Runpeng Yu, Lanqing Hong, Fengwei Zhou, Zhenguo Li, and Jun Zhu. Ood-bench: Quantifying and understanding two dimensions of out-of-distribution generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7947–7958, 2022.

[30] LIN Yong, Shengyu Zhu, Lu Tan, and Peng Cui. Zin: When and how to learn invariance without environment partition? In *Advances in Neural Information Processing Systems*, 2022.

[31] Michael Zhang, Nimit S Sohoni, Hongyang R Zhang, Chelsea Finn, and Christopher Re. Correct-n-contrast: a contrastive approach for improving robustness to spurious correlations. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 26484–26516. PMLR, 2022.

[32] Xiao Zhou, Yong Lin, Renjie Pi, Weizhong Zhang, Renzhe Xu, Peng Cui, and Tong Zhang. Model agnostic sample reweighting for out-of-distribution learning. In *International Conference on Machine Learning*, pages 27203–27221. PMLR, 2022.