Causal lnference with Sample Balancing for Out-Of-Distribution Detection in Visual Classification

Yuqing Wang[1], Xiangxian Li[1], Haokai Ma[1], Zhuang Qi[1], Xiangxu Meng[1], and Lei Meng[1]*

Shandong University, Jinan, Shandong, China
{wang_yuqing,xiangxian_lee,mahaokai}@mail.sdu.edu.cn 97qizhuang@gmail.com
{mxx,lmeng}@sdu.edu.cn

Abstract. Image classification algorithms are commonly based on the Independent and Identically Distribution (IID) assumption, but in practice, the Out-Of-Distribution (OOD) problem is widely existing, i.e., the contexts of images in the model predicting are usually unseen during training. In this case, existing models trained under the IID assumption are limiting generalization. Causal inference is an important method to enhance the out-of-distribution generalization of models by partitioning various contexts from data and leading models to learn context-invariant predictions in different situations. However, existing methods mostly have imbalance problems due to the lack of constraints when partitioning data, which weakens the improvement of generalization. Therefore, we propose a Balanced Partition Causal Inference (BP-Causal) method, which automatically generates fine-grained balanced data partitions in an unsupervised manner, thereby enhancing the generalization ability of models in different contexts. Experiments on the OOD datasets NICO and NICO++ demonstrate that BP-Causal achieves stable predictions on OOD data, and we also find that models using BP-Causal focus more accurately on the foreground of images compared with the existing causal inference method, which effectively improves the generalization ability.

Keywords: Out-of-Distribution Generalization · Causal Inference · Invariant Learning.

1 Introduction

Image classification algorithms based on deep learning have shown good performance under the Independent and Identically Distributed (IID) assumption. However, real-world datasets usually suffer from out-of-distribution (OOD) generalization problems, i.e., contexts of images in the inferring phase are mostly unseen by the modal in the training phase. Existing models trained under IID

* Corresponding author

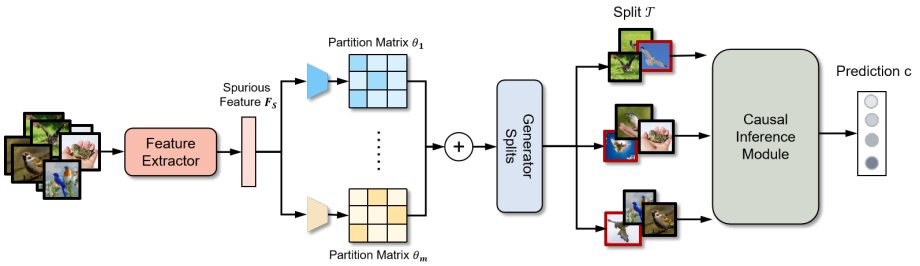

Fig. 1. The causal learning method based on Balanced Partition Causal Inference (BP-Causal) divides the input data set by training multiple partition matrices, and divides it into subsets with different contexts in a fine-grained manner.

assumption are hard to generalize well in this case. How to efficiently and accurately extract cross-environment invariant features from the complex data distribution in OOD environments is a problem that remains to be studied.

Causal inference is effective to alleviate OOD problems, and there are two main approaches. The first is invariant causal prediction (ICP) [18, 19, 10], which improves the stability of generalization by controlling the covariance of different subsets. But this approach has limited effectiveness in complex scenes. The second is invariant learning, typified by invariant risk minimization [3, 2, 12, 23] that extracts causal features that are invariant across environments by dividing the data and training a classifier that is optimal across all environments. However, these methods are lacking of constraints, which will lead to the imbalance of the divided data and impact the context invariant learning.

To address the aforementioned problems, we propose a Balanced Partition Causal Inference method BP-Causal. By adding balance constraints, the division effect of causal inference on data subsets is improved, the learning of the model in different environments is further enhanced, and the effect of extracting invariant causal features is improved, thereby enhancing the generalization ability in the OOD environment. As shown in Figure 1, BP-Causal consists of three main modules, Feature Extractor Module, Balance Split Module, and Causal Inference Module. The Feature Extractor Module learns to extract the causal features, confounding features and their mixed features from the input image; the Balance Split Module uses the information in the confounding features to partition the dataset and will be balanced for better learning in different situation; Causal Inference Module use causal features and mix features to incorporate knowledge learned from subsets of data from different contexts to further identify causal features that are invariant across contexts, improving the generalization ability of the model.

We conduct experiments on two OOD image classification datasets, NICO and NICO++, to demonstrate the effect of BP-Causal balanced partitioning, as well as its predictive ability on OOD data, and study the mechanism on generalization performance through ablation studies. Further case analysis shows that BP-Causal can focus on causal features that are invariant across environments. In conclusion, the main contributions of this paper are:

– A balanced partition causal inference method BP-Causal is proposed, which enhances the generalization of models on OOD data by self-learning manner.
– We demonstrate that a more balanced subset partitioning can have a positive impact on the model learning context-free features, thereby improving the model's generalization ability on OOD data.

2 Related Works

2.1 Out-Of-Distribution(OOD) Generalization

Traditional machine learning algorithms are based on the assumption of independent and identical distribution, but in reality the i.i.d. assumption is difficult to satisfy, so people correspondingly put forward the problem of out-of-distribution(OOD) generalization [15, 11, 1]. The OOD problem addresses challenging settings where the test distribution is unknown and different from training, which is a big challenge for machine learning work. Some of the more challenging OOD settings that exist are: debiasing [6, 14, 4], domain adaption [7, 17, 22], long-tailed recognition [13, 16, 21], etc. To better deal with the OOD problem, [9] proposed a real-world OOD dataset NICO. We follow the OOD settings for the NICO dataset, including long-tailed, zero-shot, and orthogonal.

2.2 Causal Inference

Causal inference[26] is an effective means to solve the OOD problem, which usually assumes the existence of heterogeneity and causality within the data. There are two main methods: ICP and a series of methods after it [18, 19, 10] control the target variable to be only affected by its direct variable by exploiting the heterogeneity within the data, but this method has strict requirements for the heterogeneity of the data and this approach has limited effectiveness in complex scenes. The invariant learning method represented by IRM [3, 2, 12, 23] is different from the causal prediction method that assumes the original variable level. It generalizes the previous invariance assumption to the representation level and strives to find a classifier that is optimal in all environments. But this method usually requires dividing the data into different parts and extracting common features from the different parts. This data partitioning currently lacks constraints, which may lead to inaccurate causal feature extraction.

3 Methods

3.1 Overview

As shown in the Figure 2, our model is roughly divided into three modules: the feature extraction module extracts the causal features F_c, the confounding features F_s, and the mixed features F_x from the input image x; the balanced division module uses the information in the confounding features to divide the

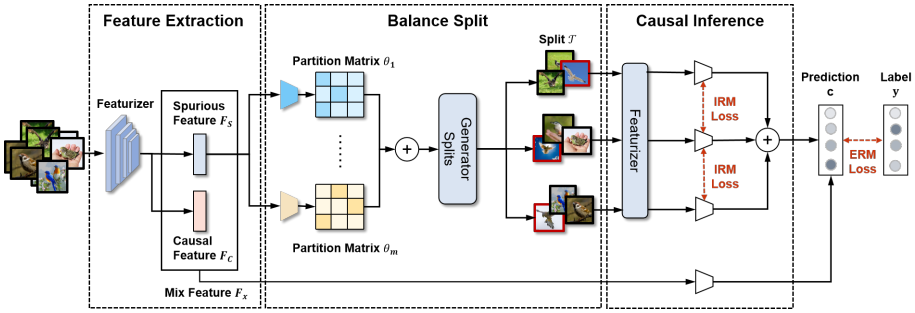

Fig. 2. Schematic diagram of the BP-Causal algorithm: the feature extraction module extracts features for subsequent computation; the balance split module divides the dataset into subsets with different environments in a balanced manner; the causal inference module uses IRM Loss and ERM Loss to constrain at the same time, training the ability to extract invariant features for prediction from different environments.

dataset into data subsets of different environments through the balanced split generation algorithm; the causal inference module fuses the knowledge learned from the data subsets of different environments to distinguish the invariance across environments causal characteristics.

3.2 Feature Extraction Module

We use an attention module to separate causal and confounding features, In this module, we use F_c and F_s to denote causal features and confounding features, respectively. First, two samples are randomly selected from the training samples for a simple random weighted summation, and the labels of the samples also correspond to the weighted summation [27].

$$\begin{aligned} \tilde{x} &= \lambda x_i + (1-\lambda) x_j \\ \tilde{y} &= \lambda y_i + (1-\lambda) y_j \end{aligned} \tag{1}$$

We use $\tilde{x}$ to obtain the feature z through an attention module Attention(x), and use the sigmoid function to disentangle the feature z to obtain the causal feature F_c and the confounding feature F_s:

$$Feature(x) = \begin{cases} z = Attention(\tilde{x}), \\ F_c = Sigmoid(z) \odot \tilde{x}, \\ F_s = Sigmoid(-z) \odot \tilde{x}, \end{cases} \tag{2}$$

Where z $\in R^{w \times h \times c}$, Attention() is an attention module called CBAM [24], $\odot$ denotes the element-wise product and Sigmoid(-z) = 1 - Sigmoid(z). We add the module to the basicblock of ResNet to distinguish F_c and F_s. For the first block, the disentangling of features is a little different from the next blocks, just like the Equation (2). For the next blocks, in D-Block, we input the mix features

we got in the previous block and disentangle them to get the causal features and confounding features. In M-Block, the two input are fused to obtain mixed features to prepare for the calculation of the next block. Because we can have many blocks, the $j+1^{th}$ D-Block and the j^{th} M-Block are as follows:

$$D-Block^{j+1}: \begin{cases} \hat{F_c}^j, \hat{F_s}^j = Feature(\tilde{x}^j), \\ F_c^{j+1} = \hat{F_c}^{j+1} + F_c^j(skip-connection), \\ F_s^{j+1} = \hat{F_s}^{j+1} + F_s^j(skip-connection), \end{cases} \tag{3}$$

$$M-Block^j : F_x^j = Conv(F_c^j) + Conv(F_s^j) \tag{4}$$

Through this module, we extract disentangled causal, confounding and mix features from the input image, and output them into the following module.

3.3 Balance Split Module

In this module, we use the extracted confounding features F_s from the feature extraction module to train the partition matrix θ and update the data partition $\mathcal{T} = \{t_1, t_2, \cdots, t_m\}$. We first train a bias classifier h for each matrix and use h we get the prediction $c = h(F_s)$. We minimize the ERM Loss for it:

$$\mathcal{L}_{bias}^{erm} = E_{(x,y)\in\mathcal{D}}\ell(h(F_s), \tilde{y}) \tag{5}$$

where $\mathcal{D}$ is training data, h is a linear classifier, F_s is the confounding feature we got from the previous module, y is the label.Then using this classifier, under the constraint of an IRMloss [3], a partition matrix θ is trained to gradually update the partition of the dataset in a fine-grained way:

$$\mathcal{L}_{split}^{irm} = \sum_{t\in\mathcal{T}_i(\theta)} R^t(h) + \lambda \cdot \|\nabla_{w|w=1.0} R^t(w \cdot h)\|^2 \tag{6}$$

where $\mathcal{T} = \{t_1, t_2, \cdots, t_m\}$ is current data partition, $\mathcal{T}_i(\theta)$ denotes partition $\mathcal{T}_i$ is decided by $\theta \in R^{K\times m}$, K is the total number of training samples and m is the number of splits in a partition, $R^t(h) := E_{(x,y)\in t_i}\ell(h(F_s), y)$ is the risk under subset t_i, h is the bias classifier trained in the previous step, w = 1.0 is a scalar and fixed "dummy" classifier, the gradient norm penalty is used to measure the optimality of the dummy classifier at each subset t, and $\lambda \in [0, \infty]$ is a regularizer balancing between the ERM term and the invariance of the predictor $1 \cdot$ h.

Training only one partition matrix θ may lead to a large imbalance in subsets. In order to alleviate the imbalance, we train multiple matrices, combine the probability distributions of multiple trainings, and then decide the final partition:

$$\theta_{final} = \sum_{i=0}^{m} \begin{pmatrix} p(k_1, m_1) & \cdots & p(k_1, m_j) \\ \vdots & \ddots & \vdots \\ p(k_n, m_1) & \cdots & p(k_n, m_j) \end{pmatrix}_i \tag{7}$$

where $p(k_m, m_n)$ denotes the probability that the n^{th} image is divided into the j^{th} partition. For $\theta_{final} \in R^{K \times m}$, the index of the split to be divided into is:

$$Idx = \underset{\theta}{argmax}(Softmax(\theta_{final})) \tag{8}$$

Then we can divide the K images into corresponding data subsets according to Equation 8. Through this module, we divide the dataset into fine-grained subsets with different environments, which is more helpful for the model to extract causal features that are invariant across environments.

3.4 Causal Inference Module

Typically, we achieve causal inference by using backdoor adjustment:

$$P(Y|do(X)) = \sum_{t \in \mathcal{T}} P(Y|X,t)P(t) \tag{9}$$

where $P(Y|X,t)$ denotes the prediction of the classifier trained in split t and P(t) := 1/m. With do(X), we hope to exclude spurious correlation between the context and the prediction results, so we train the model on data from different environments that are balanced divided, so that the model can focus on the subject of the image in any environment to achieve causal inference.

We first use ERM Loss to constrain the feature extraction part and the classifier, so that the model can extract features accurately.

$$\mathcal{L}^{erm}_{train} = \frac{1}{m} \sum_{i=0}^{m} E_{(x,y) \in t_i} \ell(g_i(F_c), \tilde{y}) \tag{10}$$

where m is the number of splits in a partition, t_i represents a specific split, g_i is a linear classifier for t_i, F_c is causal feature. Then, by dividing and training multiple classifiers for the data of different environments, and using an IRM Loss to align these classifiers with constraints, a classifier that is optimal in all environments is obtained. With this classifier we can mitigate the context interference and make the model better focus on causal features.

$$\mathcal{L}^{irm}_{invariance} = \sum_{t \in \mathcal{T}_i(\theta)} R^t(g) + \lambda \cdot \left\| \nabla_{w|w=1.0} R^t(w \cdot g) \right\|^2 \tag{11}$$

After multiple partitions updating and training, we can gradually approach the backdoor adjustment formula9 to achieve causal inference.

3.5 Training Strategy

There are multi-class loss constraints in BP-Causal, and a staged training method can be used.The extraction of training features in the first stage is jointly constrained by the empirical risk loss from different data subsets and the invariant

risk loss of aligning the classifier weights under different environments. In order to make the model extract better features, we minimize these two losses:

$$min\,\mathcal{L}_{train}^{erm} + \mathcal{L}_{invariance}^{irm} \tag{12}$$

The second stage is training data partition. First we train a biased classifier using the empirical risk loss, then use that classifier to constrain the data partition by an invariant risk loss. We minimize the empirical loss to improve the model's ability to distinguish confounding features, but we maximize the invariant loss, so that the m splits are divided in different fine-grained confounding features, so as to achieve the purpose of dividing different environmental data subsets:

$$min\,\mathcal{L}_{bias}^{erm} + max\,\mathcal{L}_{split}^{irm} \tag{13}$$

4 Experiments

4.1 Datasets

NICO [9] is a real-world dataset with 2 superclasses for a total of 19 classes, and 9 or 10 contexts under each class, accumulating a total of 188 (subject, context) combinations and collecting about 25,000 images. We follow the setting of [23], selecting 10 animal classes and 10 contexts. We make a challenging OOD setting consisting of three factors on context: 1) Long-tailed - The training context are long-tailed in each class; 2) ZeroShot - for each class, 7 of the 10 contexts are in the training images , the other 3 contexts only appear in the test; 3) Orthogonal - the head context for each class is set to be as unique as possible.

NICO++ [25] is an upgrade to the NICO dataset. Consistent with NICO, NICO++ decomposes images into (subject, context) combinations. NICO++ has included 80 classes, 10 public contexts, and 10 unique contexts for each class, with a total of 200,000 images. We picked 10 classes from the public context section, including animals, vehicles, and others. We follow the OOD settings[23], including long-tail, ZeroShot - 4 of the 6 context per class are in the training images, the other 2 labels only appear in test and orthogonal - as much as possible to ensure that each class' header context appears only once or twice.

4.2 Experimental Settings

Evaluation Protocol We follow [23] and use the accuracy on the validation set and test set as the judging criterion. The formula is as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{14}$$

where TP, TN, FP, FN are the number of true positive, false positive samples, true negative samples, and false negative samples.

Table 1. Recognition accuracies (%) based on ResNet18 on NICO and NICO++ dataset. "val" and "test" denote the accuracies on validation set and test set.

Method	Model	NICO		NICO++	
		val	test	val	test
Conv. Method	ResNet-18	44.38	44.08	44.73	45.93
	Cutout	46.23	44.08	45.75	45.75
	Mixup	44.69	42.46	**49.00**	49.06
Causal Method	CBAM	43.77	43.54	44.27	45.47
	CaaM	44.85	44.69	43.93	46.44
	BP-Causal	**48.23**	**48.08**	48.38	**51.28**

Implementation Details For NICO dataset, the optimizer was set to SGD with a learning rate of 0.05. We trained the model with 200 epochs and the learning rate was decreased by 5 at 120, 160 epoch. From 40 epoch, the data partition will be updated every 20 epochs, and we divide the dataset to 4 parts. For NICO++ dataset, the optimizer was set to SGD with a learning rate of 0.02. We trained the model with 200 epochs and the learning rate was decreased by 5 at 80, 120, 160 epoch. From 40 epoch, the data partition will be updated every 40 epochs, and we divide the dataset to 4 parts.

4.3 Performance Comparison

This section presents the performance comparison of BP-Causal with existing image classification methods, including the traditional Resnet-18 [8], two data augmentation methods [5, 27], the CBAM attention mechanism [24] and a causal method CaaM [23]. We can observe the following Table 1:

- Simply adding the attention mechanism, in the OOD context, may cause the attention to focus on the wrong area, so that after adding attention module [24], the model performance is not as good as the baseline algorithm.
- The performance of the CaaM algorithm is better than that of the baseline algorithm, as well as the attention method. Because it learns causal features that are invariant across environments and stable in prediction from different environments by partitioning the dataset. However, there is a lack of constraints on the division of the dataset, which loses part of the performance.
- When BP-Causal divides the data set into different environment subsets, the balance between the subsets is enhanced by adding constraints to the division process. This method works better in the OOD case than baseline and using the attention mechanism alone, about 3%-5% performance improvement.

4.4 Ablation Study

In this section, we investigate the effectiveness of the proposed algorithm. The experiment selected resnet18 [8] as the baseline. As shown in the Table2, adding the attention mechanism [24] directly will affect the performance, because the attention mechanism may capture spurious correlation as a basis for prediction

Table 2. The influence of each module of the algorithm on the performance.

Model	NICO		NICO++	
	val	test	val	test
Baseline	44.38	44.08	44.73	45.93
+ CBAM	43.77	43.54	44.27	45.47
+ CBAM + Causal	44.85	44.69	43.93	46.44
+ CBAM + Causal + LB	45.23	45.85	46.04	47.24
+ CBAM + Causal + MB	46.46	46.62	44.05	47.41
+ CBAM + Causal + GB	45.62	46.85	46.38	47.46
+ CBAM + Causal + LB + mixup	48.08	47.38	**48.83**	49.74
+ CBAM + Causal + MB + mixup	45.62	47.92	48.55	51.00
+ CBAM + Causal + GB + mixup	**48.23**	**48.08**	48.38	**51.28**

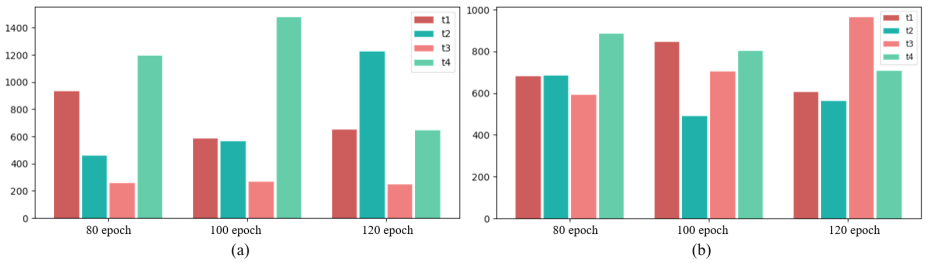

Fig. 3. Statistics on the number of subsets divided by different epochs. Where (a) is the result of using CAAM partition, (b) is the result of BP-Causal partition.

in the OOD context. On the basis of CBAM [24], a causal method is added to alleviate the problem of paying attention to errors in OOD environments. However,there is no restriction on the partition of the data set, which leads to the problem of imbalance in the data partition. We try three balanced methods to constrain the partition. Loss Balance(LB) is to add a loss during training, Manual Balance(MB) is to balance images of different subsets by manual deletion and supplementation, Aggregation Balance(GB) is to alleviate the degree of imbalance by training multiple partition matrices. In the third way, the inference of the partition is minimal, but the imbalance of the partition is alleviated to a certain extent, so the best performance is obtained. In addition, we found that mixup [27], as an effective data augmentation method, also works well in OOD situations. We tested the effects of the three balancing methods after adding the mixup, and the Smooth Balance method is still the best than any other methods.

4.5 Analysis of Split Partition

In previous experiments we have demonstrated the positive effect of balanced partitioning on the generalization ability of the model, in this section we will show the practical effect of BP-Causal in balanced partitioning of subsets, as shown in Figure 3. The division using the CAAM method has obvious imbalance between different splits, which largely restricts the ability of the model to learn

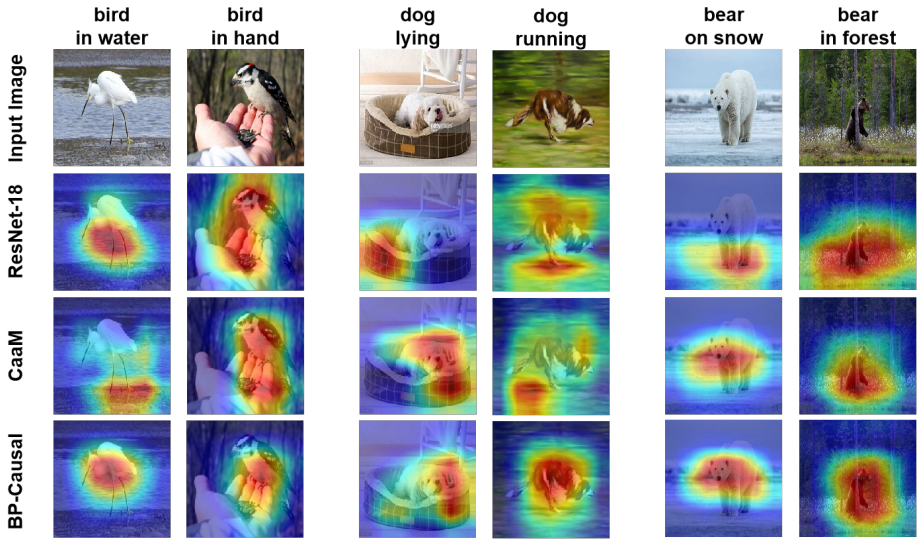

Fig. 4. Visualization of attention maps with base modal, CaaM, and BP-Causal.

invariant features from different subsets. The subset using BP-Causal division is relatively more balanced. , and it is also very stable in different divisions, which effectively improves the effect of the model learning from the causal feature.

4.6 Case Study

Figure 4 shows the qualitative attention map [20] comparison between our proposed BP-Causal algorithm, the traditional method ResNet18 and the causal method CaaM. We selected three categories of the NICO dataset, and each category selected two contexts for experiments. As can be seen from the figure, our method can focus more on the subject of the image rather than the surrounding environment and other objects compared to the other two methods. For different contexts of the same category, a large part of ResNet18's attention is focused on the surrounding environment, especially under the category of dogs, but our algorithm can pay attention to the characteristics of dogs in different environments, This improves the classification accuracy of the model.

5 Conclusion

This paper proposes a causal learning method BP-Causal based on balanced partition, which automatically generates balanced data subsets of different environments through training, extracts invariant causal features from different environments, and enhances the model's learning in OOD environments. capabilities and generalization capabilities.

BP-Causal effectively alleviates the attention bias of the attention model and the interference of confounding factors in complex OOD scenarios. However, training multiple classifiers and then forcing the use of loss to align the weights of different classifiers may lead to difficulty in convergence and affect the performance of the model. We will try to use meta-learning and other means to use a single meta-model to learn common features of different distributions, reduce the complexity of the model, and improve the performance of the model.

Acknowledgments

This work is supported in part by the Excellent Youth Scholars Program of Shandong Province (Grant no. 2022HWYQ-048) and the Oversea Innovation Team Project of the "20 Regulations for New Universities" funding program of Jinan (Grant no. 2021GXRC073)

References

1. Alessandro Achille and Stefano Soatto. Information dropout: Learning optimal representations through noisy computation. *IEEE transactions on pattern analysis and machine intelligence*, 40(12):2897–2905, 2018.
2. Kartik Ahuja, Karthikeyan Shanmugam, Kush Varshney, and Amit Dhurandhar. Invariant risk minimization games. In *International Conference on Machine Learning*, pages 145–155. PMLR, 2020.
3. Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.
4. Christopher Clark, Mark Yatskar, and Luke Zettlemoyer. Don't take the easy way out: Ensemble based methods for avoiding known dataset biases. *arXiv preprint arXiv:1909.03683*, 2019.
5. Terrance DeVries and Graham W Taylor. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*, 2017.
6. Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv preprint arXiv:1811.12231*, 2018.
7. Mingming Gong, Kun Zhang, Tongliang Liu, Dacheng Tao, Clark Glymour, and Bernhard Schölkopf. Domain adaptation with conditional transferable components. In *International conference on machine learning*, pages 2839–2848. PMLR, 2016.
8. Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
9. Yue He, Zheyan Shen, and Peng Cui. Towards non-iid image classification: A dataset and baselines. *Pattern Recognition*, 110:107383, 2021.
10. Christina Heinze-Deml, Jonas Peters, and Nicolai Meinshausen. Invariant causal prediction for nonlinear models. *Journal of Causal Inference*, 6(2), 2018.
11. Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136*, 2016.

12. Wengong Jin, Regina Barzilay, and Tommi Jaakkola. Domain extrapolation via regret minimization. *arXiv preprint arXiv:2006.03908*, 2020.
13. Salman H Khan, Munawar Hayat, Mohammed Bennamoun, Ferdous A Sohel, and Roberto Togneri. Cost-sensitive learning of deep feature representations from imbalanced data. *IEEE transactions on neural networks and learning systems*, 29(8):3573–3587, 2017.
14. Byungju Kim, Hyunwoo Kim, Kyungsu Kim, Sungjin Kim, and Junmo Kim. Learning not to learn: Training deep neural networks with biased data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9012–9020, 2019.
15. Shiyu Liang, Yixuan Li, and Rayadurgam Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. *arXiv preprint arXiv:1706.02690*, 2017.
16. Dhruv Mahajan, Ross Girshick, Vignesh Ramanathan, Kaiming He, Manohar Paluri, Yixuan Li, Ashwin Bharambe, and Laurens Van Der Maaten. Exploring the limits of weakly supervised pretraining. In *Proceedings of the European conference on computer vision (ECCV)*, pages 181–196, 2018.
17. Krikamol Muandet, David Balduzzi, and Bernhard Schölkopf. Domain generalization via invariant feature representation. In *International Conference on Machine Learning*, pages 10–18. PMLR, 2013.
18. 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.
19. Niklas Pfister, Peter Bühlmann, and Jonas Peters. Invariant causal prediction for sequential data. *Journal of the American Statistical Association*, 114(527):1264–1276, 2019.
20. Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.
21. Li Shen, Zhouchen Lin, and Qingming Huang. Relay backpropagation for effective learning of deep convolutional neural networks. In *European conference on computer vision*, pages 467–482. Springer, 2016.
22. Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7167–7176, 2017.
23. Tan Wang, Chang Zhou, Qianru Sun, and Hanwang Zhang. Causal attention for unbiased visual recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3091–3100, 2021.
24. Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018.
25. Renzhe Xu Han Yu Zheyan Shen Peng Cui Xingxuan Zhang, Yue He. Nico++: Towards better benchmarking for domain generalization, 2022.
26. Liuyi Yao, Zhixuan Chu, Sheng Li, Yaliang Li, Jing Gao, and Aidong Zhang. A survey on causal inference. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 15(5):1–46, 2021.
27. Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017.