Group Distributionally Robust Optimization for Solving Out-Of-Domain Generalization and Finding Causal Invariant Relationships

Zenan Guan
Supervisor(s): Jesse Krijthe, Rickard Karlsson, Stephan Bongers
EEMCS, Delft University of Technology, The Netherlands

June 19, 2022

A Dissertation Submitted to EEMCS faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering

Abstract

Out-of-Domain (OOD) generalization is a challenging problem in machine learning about learning a model from one or more domains and making the model perform well on an unseen domain. Empirical Risk Minimization (ERM), the standard machine learning method, suffers from learning spurious correlation in the training domain, therefore may perform badly when the unseen domain has different distribution from the training domain. Group Distributionally Robust Optimization (group DRO) is a method proposed to handle the OOD generalization problem. In this paper, the goals are to 1) measure if group DRO has a better OOD generalization performance than ERM. 2) evaluate if group DRO finds causally invariant relationships between the input and output. Semi-synthetic bird images with different backgrounds are used to form our data sets to construct a binary image classification problem for experiments. Results show that group DRO improves OOD generalization performance over ERM, and group DRO can find invariant relationships. However, the ability of group DRO to find invariant relationships is limited when the spurious correlation in the training domain is strong.

Keywords: OOD generalization, group DRO, spurious correlation, invariant relationships

1 Introduction

Out-of-Domain (OOD) generalization problem is a challenging problem in the machine learning research field that machine learning models, which have good performance in some domains, may fail when applied to an unseen domain [1].

The reason why the model fails is that the training data we have does not always have the same distribution as the data in the deployment, which leads to the standard machine learning algorithm, under the principle of Empirical risk minimization (ERM), to exploit spurious correlation in the training domain [2], therefore performing poorly on data where such correlation does not hold.

An idea to solve the OOD generalization problem is to assume that there exist invariant relationships across domains, and by finding and exploiting such relationships, the model can make better predictions in the unseen domains by causal inference [3]. Many possible solutions for solving the OOD generalization problem have been proposed. One of them is Distributionally Robust Optimization (DRO) [4], which tries to minimize worst-case loss over potential test distributions [4, 5], rather than minimize the average loss in the training set as ERM does. Group DRO, an invariant of DRO, has been proven to prevent the models from relying on pre-specified spurious correlations [2]. However, it is unclear if group DRO can find invariant relationships between input and output and learn an model that predict based on such invariant relationships. In this research, we try to answer the following research questions:

- Does group DRO perform better than ERM in OOD generalization in a binary image classification problem?
- Can group DRO find and exploit the invariant relationships between the input and output in the training domain and learn an invariant classifier?

We structured the paper as follows: In Section 2 of the paper, the comparison between group DRO and ERM is discussed. Section 3 is about the methodology to answer the research questions, including the background behind the methodology. Section 4 is about

the experiments, including the semi-synthetic data sets generation, experiments steps and results. Section 5 is about responsible research, in which the research principles we adhere to are discussed. The results of the experiments are discussed in section 6. In section 7, the conclusions of this research are summarized, and future works are proposed.

2 Group DRO and ERM

In this section, group DRO and its comparison with ERM is formally introduced, then the reason why group DRO is a potential method to find invariant relationship is discussed.

The risk or loss in machine learning means the difference between the predicted output and actual output is measured by a loss function. Ideally, we want to minimize the true risk, which is the average loss over all possibility. However in practice, since it is not possible for us to know the true distribution over all input and output, therefore we need to find another way to approximate it.

ERM or Empirical Risk Minimization is a traditional method in supervised learning [7]. The concept is simple and intuitive: we assume the training data, drawn from some distribution, is a representative of all the classes in the real world, whose distribution can approximate the real distribution, the algorithm tries to find a model that minimize the empirical risk: the average loss between the predicted output and actual output in our training set. Ideally, if we get more data, then the ERM can approach the true risk; however, in practice, getting more data could be expensive; besides, the data we get is often biased, which lead the ERM learns the spurious correlation in the training set, and suffer high loss over relatively rare examples.

To improve the ERM, Distributionally Robust Optimization (DRO) [4, 5] is proposed. Instead of minimizing the empirical risk, DRO tries to minimize the worst-case expected loss over a series of possible test distributions. In this research, we use an invariant of DRO called group DRO [2]. In group DRO, we divide data in the training set into different group, using our prior knowledge of spurious correlation, and the distributions of each group serves as possible test distributions in DRO; in another words, group DRO learns a model with good worst-case training loss over the groups, which could prevent the model exploit the spurious correlation we specified beforehand.

The standard procedure of ERM is [2]:

$$\hat{\theta}_{\text{ERM}} := \arg\min_{\theta \in \Theta} E_{(x,y)\sim\hat{P}}[\ell(\theta; (x, y))] \tag{1}$$

Where x is the input feature, y is the label, l is the loss, P is some distribution where training data is drawn from (and under with the loss is computed) and θ is the model from a model family Θ. In ERM, we learn a model minimize the expected loss under the empirical distribution over training data.

The procedure for DRO looks like the following, the goal is to minimize the worst-case expected loss (R in the formula) over an uncertainty set of distributions Q [2]:

$$\min_{\theta \in \Theta} \left\{ \mathcal{R}(\theta) := \sup_{Q \in \mathcal{Q}} E_{(x,y)\sim Q}[\ell(\theta; (x, y))] \right\} \tag{2}$$

In group DRO, the uncertainty set is defined in terms of the groups, where groups are formed with our knowledge of spurious correlations, and the worst-case risk is the maximum

over the expected loss of each group, or worst group loss [2]:

$$\mathcal{R}(\theta) = \max_{g \in \mathcal{G}} E_{(x,y)\sim P_g}[\ell(\theta; (x, y))] \tag{3}$$

Therefore, in group DRO, we learn a model minimize the worst-case loss described above, over empirical distribution among each group [2]:

$$\hat{\theta}_{\text{DRO}} := \arg\min_{\theta \in \Theta} \left\{ \hat{\mathcal{R}}(\theta) := \max_{g \in \mathcal{G}} E_{(x,y)\sim \hat{P}_g}[\ell(\theta; (x, y))] \right\} \tag{4}$$

To summarize, since ERM only cares about minimizing the overall loss, the ERM algorithm could capture some obvious non-essential features (background) in the input, which are highly and spuriously correlated with output (label) during the learning process. Relying on these correlations will not hurt the overall loss too much; however, this could lead to bad performance when the learned model applied to the data set in which the same correlation does not hold. For group DRO, we can group the data by the known spurious correlation. As the algorithm of group DRO will minimize the loss of the group on which the algorithm expects the model performs worst, the group with a small number of examples will not be neglected. Therefore, group DRO algorithm avoids relying on the spurious correlation, and makes it promising to find a model that uses features causal to the output to predict or classify.

3 Methodology

In this section, we introduce the background knowledge of Out-Of-Domain (OOD) generalization and Invariant Causal Prediction (ICP) , then the methods to answer the research questions.

3.1 Out-of-Domain generalization

The algorithms for machine learning generally assume that the training data and testing data has the same distribution [8]. However, in real life this may not be the case and could result in poor performance of the learned model in deployment.

To illustrate the OOD generalization, computer scientists proposed an image classification problem as an example about determining if the animal in a photograph is a cow or a camel [9, 10]: due to the differences in living habits, most of the photographs of cow have grassland as backgrounds, while photographs of camel mainly has desert backgrounds, the model learned by the machine learning algorithm may use the background features to classify the example, which may perform badly on the test domain where the data is collected from different environments than the training domain, e.g., when most photographs of cows collected are in the desert background.

The problem introduced above leads to the Out of Domain generalization problem: How to make the model learned from one or more domains performs well in an unseen domain.

3.2 Evaluation of OOD generalization performance

We construct a binary image classification problem using semi-synthetic data sets to measure the OOD generalization performance of models learned by group DRO. Two training sets are used to train the models, which have different strengths of spurious correlation: in the

first training set, for each class, the number of objects in each background is the same, so the correlation between background and label is week; in the second training set, the background has a strong correlation with the class label of the object, as most (but not all) of objects in one class are against one background and most of objects in another class are against another background. Then we compare the performance of models on testing sets with different background distributions, where the correlations between the background and class label are similar, different, or opposed to in the training set. We also did the same experiment on models learned by ERM, and by comparing the performance of models learned by the two methods, we can see if group DRO is better able to generalize Out-Of-Domain than ERM.

3.3 Invariant causal prediction

To solve the problem OOD generalization problem, one idea is to find and exploit invariant relationships between the invariant features and output in the training domain (in the example metioned in 3.1, that is the shape of the animal and its class label), which hold across domains and can be used to explain why an object belongs to a class causally [3], rather than rely on the spurious correlations between some obvious but spurious features (background) and output, which, as Geirhos et al suggest, is a shortcut [11]. This is called Invariant Causal Prediction (ICP) [6].

Figure 1 is a visual presentation of ICP. In short words, the goal of ICP is to learn a model that is able to identify and use the invariant features to predict the output, and ignore the spurious features.

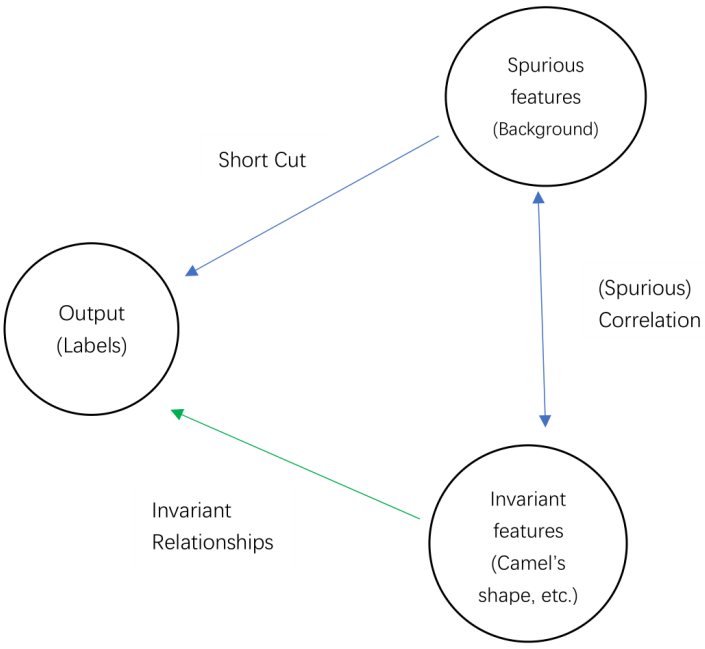

Figure 1: A visual presentation of Invariant Causal Prediction: if the learned model is able to use features which have causal invariant relationships with the output, then it performs Invariant Causal Prediction

3.4 Evaluation of finding invariant relationships

We evaluate if group DRO is able to find invariant relationships by determining if the learned models can perform ICP based on the same binary classification problem as in 3.2. First, we use group DRO and ERM to learn the models from different training sets with different the correlation strengths between the background and label of object. Next, for every object i, we generate two examples, where the same object is placed in the different backgrounds. Then for every object, we classify the two examples with group DRO model and ERM model, and count when the two examples are classified identically and correctly. In the end, we compare the results, and if models learned by group DRO have most of examples being classified identically and correctly, then group DRO can find and exploit invariant relationships and perform ICP.

4 Experiments

In this section, first we introduce the process of semi-synthetic data sets generation using the The Caltech-UCSD Birds-200-2011 (CUB-200-2011) dataset [12] or CUB data set in short, the data set can be downloaded from Kaggle[1]. Then, we discuss the setups of the two experiments, the experiment A is degined to measure the OOD generalization ability of DRO, while the experiment B is to measure if group DRO can find invariant relationship in the training domain.

4.1 Semi-synthetic data sets generation

In this research, we use semi-synthetic data sets for binary classification of birds.

The semi-synthetic data set are composed of two parts, namely the bird and the background. The photographs of birds come from the CUB data set, which contains 11788 images of 200 species of birds, 5994 for training and 5794 for testing. We divide the birds into two classes for binary classification: land birds and water birds, as in the group DRO paper [2]. Land birds have 4615 images in training split, 4510 images in testing split while water birds have 1379 images in training split and 1284 images in testing split. We use pure green and pure blue as backgrounds to make the distinction between backgrounds simple and explicit.

In the generation process, the original division for training and testing of birds photographs are held, and the bird is extracted from the initial photograph and combined with the background. Each example is marked with its true label and background, so we can group the examples with their labels and backgrounds: land birds with green background, land birds with blue background, water birds with blue background and water birds with green background.

The python generation program we use is developed by the authors of group DRO, which along with the group DRO implementation, can be found in GitHub[2].

Figure 2 is a diagram for showing the invariant and spurious features in the data we generate.

See Figure 3, 4, 5, 6 for some examples of generated images.

[1]https://www.kaggle.com/datasets/veeralakrishna/200-bird-species-with-11788-images
[2]https://github.com/kohpangwei/group_DRO

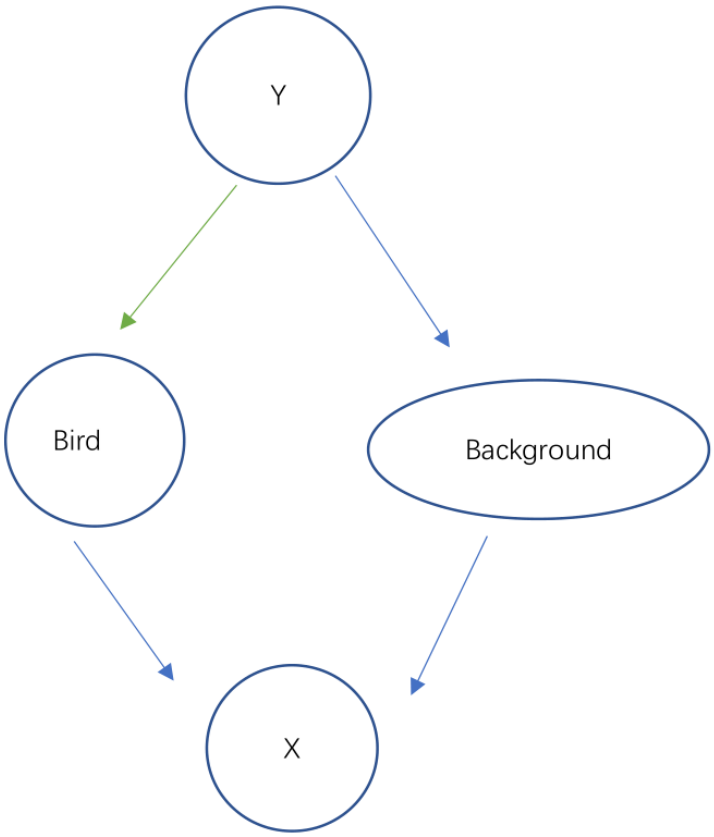

Figure 2: The diagram for data generation. X is the image (input), which consists of a bird and its background; Y is the class label of the bird to be predicted (output), which is determined by the bird itself, so the features of Bird are the invariant features here, which have causal invariant relationship with the class label, while the features of the Background are the spurious features

Figure 3: Land birds with green background

Figure 4: Land birds with blue background

Figure 5: Water birds with blue background

Figure 6: Water birds with green background

4.2 Experiments setup

In this subsection, we elaborate on how the experiments are designed, including the implementation of group DRO and ERM and the hyperparameters we use in section 4.2.1. The setup the Experiment A (designed to measure the OOD performance) is described in 4.2.2 and the setup of the Experiment B (designed to measure if group DRO finds the invariant relationships) is described in 4.2.4.

4.2.1 Implementation and hyperparameters

For both group DRO and ERM, we use ResNet50 [13] model, a 50 layers deep convolutional neural network which is widely used for image classification. We use the group DRO implememtation developed by the author of group DRO, which can be found at GitHub, as mentioned in section 4.1. For the ERM implementation, we develop it with tensorflow, using the implementation of Resnet50 from tensorflow.keras package. Both DRO and ERM use stochastic gradient descent (SGD) as optimizer.

Table 1 shows the hyperparameters we use during the experiments.

Parameter	Value
Batch Size	64
Epochs	30
Momentum	0.9
Learning Rate	0.001
Weight Decay	0.0001

Table 1: Hyperarameters setting

4.2.2 Experiment A setup

The experiment A refers to the method in section 3.2, aims to measure the OOD generalization performance of group DRO by comparing how accurate models learned by both group DRO and ERM on varying unseen testing domains.

The training, validation, and testing sets are generated as follows: in the first training set, for each class of examples, we place half of the birds against the green background and another half against the blue background, so the correlation between the bird and background is weak. In the second training set, 90% of land birds are placed against the green background and the remaining against the blue background; similarly, 90% of water birds are placed against the blue background, the remaining against the green background, so the correlation between the background and class label is strong. The validation set is formed by randomly choosing 20% of water birds examples and 20% of land birds examples

from the training set, and we train the model several times to select the model with the lowest validation loss. Testing sets are generated with probability p = P(green background | land birds) = P(blue background | water birds), varying p from 0 to 1 with step 0.1, so we have testing sets with different background distributions to simulate data sets collected from different environments: when p is close to 0.5, the correlation between background and class label becomes week; when p is close to 0 or 1, the correlation becomes strong in different directions.

We then compare the accuracy on different testing sets of classifiers learned from first and send training set by group DRO and ERM, to see how these classifiers perform in different unseen domains.

4.2.3 Experiment A results

Below are the results for experiment A. The classification accuracy is calculated for each class. The x-axis refers to testing sets with different background distributions, and the y-axis refers to the accuracy; the grey line in each graph indicates the distribution of backgrounds in the training set.

Figure 7 and Figure 8 show the accuracy of classifiers learned by group DRO and ERM from the training set 1 on different testing sets for different classes. The results show that both methods perform stably across testing sets. The classifier learned by group DRO has stable and exemplary performance in classifying both classes of birds across different testing sets, although the accuracy for classifying land birds is higher than for classifying water birds. In contrast, the classifier learned by ERM is heavily biased towards land birds: although its performance is slightly better in classifying land birds (accuracy near 1) than group DRO (accuracy near 0.986) across testing sets, its performance in classifying water birds (near 0.1) across testing sets is much worse than group DRO (near 0.9).

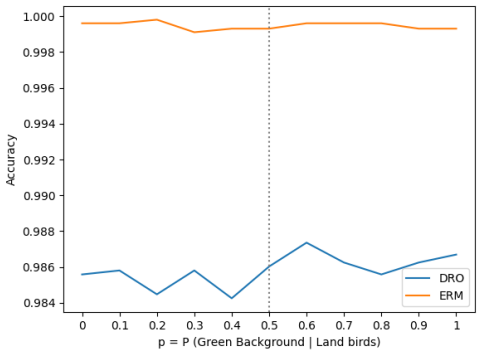

Figure 7: Experiment A: the accuracy for classifying land birds examples by models learned from training set 1.

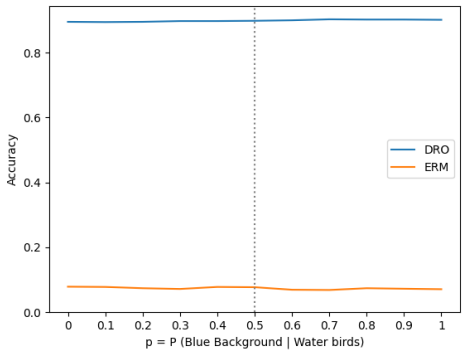

Figure 8: Experiment A: the accuracy for classifying water birds examples by models learned from training set 1.

Figure 9 and Figure 10 show the accuracy of classifiers learned by group DRO and ERM from the training set 2 on different testing sets for different classes. The results show that, when the spurious correlation between background and label is strong in the training set, although the group DRO classifier generally performs better than the ERM classifier across testing sets, the classifiers learned by both methods have a better performance in the testing

sets where the distributions of backgrounds are similar to the training set. Nevertheless, the classifier learned by ERM is not as stable as the classifier learned by group DRO, especially for classifying water birds, as shown from the change of accuracy from p = 0 to p = 1.

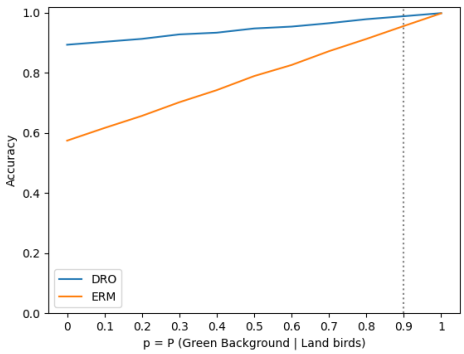

Figure 9: Experiment A: the accuracy for classifying land birds examples by models learned from training set 2.

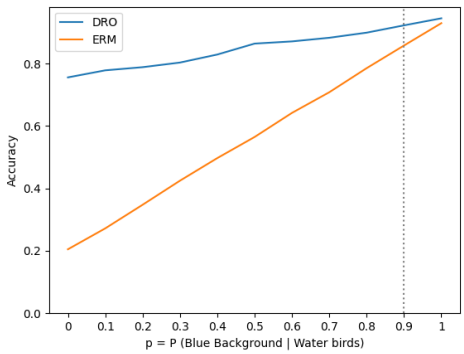

Figure 10: Experiment A: the accuracy for classifying water birds examples by models learned from training set 2.

4.2.4 Experiment B set up

The experiment B refers to method in section 3.4, aims to evaluate if group DRO can find invariant relationships in the training domain, and learns classifiers which classify an example based on its invariant features rather than the spurious ones.

The training sets are generated as following: we change the distribution of backgrounds for each class with probability p = P(green background | land birds) = P(blue background | water birds), varying p from 0.5 to 0.9 with step 0.1, so the strength of spurious correlation between background and label differs. The validation set are formed by randomly choosing 20% of land birds examples of each background from the training set, similarly for water birds. To counter the randomness in choosing the validation set, for each background distribution we generate training sets three times.

Two testing sets are created: for each bird in CUB testing split, we generate two examples in two different background, and one is put in testing set 1 and the other is put in testing set 2. So we have every bird with examples of two backgrounds in two testing sets.

Then we learn the classifiers from training sets with different background distribution by group DRO and ERM, and deploy the classifiers on the two testing sets. We count how many pairs of examples with same bird but different backgrounds are classified identically and equally, i.e.:

$$f(bird\_i\_with\_green\_background) = f(bird\_i\_with\_blue\_background) = label(bird\_i) \tag{5}$$

where f is the classifier.

We repeat the experiments three times as we have three training sets for each correlation strength, and then calculate the average results.

4.2.5 Experiment B results

The table 2 shows the results of the Experiment B. As shown in the results, the classifier learned by group DRO from every correlation strength of training sets has most of the examples classified correctly and equally, and generally performs much better than the corresponding classifier learned by ERM, indicating that group DRO can find and exploit invariant relationships. However, when we compare between the performance among group DRO classifiers learned from different correlation strength of training sets, we find that when the correlation strength between background and class label in the training set becomes stronger (from p = 0.5 to p = 0.9), the performance becomes worse, especially from p = 0.8 to p = 0.9, although is still much better than ERM.

Training set \ Method	ERM	group DRO
p = 0.5	4600.33 (79.40%)	5548.33 (95.76%)
p = 0.6	4838.33 (83.51%)	5525.33 (95.36%)
p = 0.7	4544 (78.43%)	5474.33 (94.48%)
p = 0.8	4521 (78.03%)	5314.67 (91.73%)
p = 0.9	3980.33 (68.69%)	4882.33 (84.27%)

Table 2: Experiment B results: the (average) number of bird being classified correctly and equally, and the percentage in total number of birds

5 Responsible Research

In this section, we introduce the responsible research principles we adhere to, highlighting the integrity of the research and the reproducibility of the experiments.

The first is about the integrity of research. In the dataset part from section 4 Experiments, we explicitly pointed out the source of our data: since we use semi-synthetic datasets constructed of birds photographs and backgrounds, the CUB dataset, where the photographs of birds come from, is referenced, and we confirm that the CUB dataset belongs to the public domain dedication so there is no copyright concern; we use the python generation program and group DRO implementation created by others, which is referenced as well; the backgrounds are two images, one is pure green, and the other is pure blue, which are archived along with the CUB dataset we downloaded, the semi-synthetic datasets we generated and the results we have and can be provided upon request. We guarantee that all data and results are not manipulated, fabricated, or trimmed. All related works on which this research is built are referenced.

Besides, we ensure that our experiments are reproducible and repeatable. In section 4 Experiments, we clearly explain the generation process of the semi-synthetic datasets in 4.1, the hyperparameters and implementation we used in 4.2.1 and the experiment setups for the two experiments in 4.2.2 and 4.2.4. The source of the semi-synthetic datasets generation program and the group DRO implementation is referenced as mentioned above, and the code we create for the ERM implementation can be provided upon request.

6 Discussion

In this section, we will discuss the experiment results and our reflections on the results.

The Results from experiment A in section 4.2.3 demonstrate the improvement of group DRO in OOD generalization over ERM. Generally, group DRO classifiers have much better and stabler performance than ERM ones across different testing sets, even when the spurious correlation between background and label is strong (as more birds in land birds class are against a green background, and more birds in water birds class are against a blue background) in the training set. The results also show the bias towards land birds for classifiers learned by both methods, as the accuracy of classifying land birds is higher than of classifying water birds. The reason behind such bias is that the land bird class has more examples than the water bird class in the training set. However, such bias is much slighter for the classifiers learned by group DRO than those learned by ERM. This reflects the differences between group DRO and ERM in learning a model as introduced in section 2: ERM minimizes the overall loss, and since land birds have more examples than water birds in the training set, high accuracy on land birds has more impact on overall accuracy, which leads to a bias towards land birds; but as group DRO tries to minimize the worst group loss, so the accuracy on every group (land or water bird with green or blue background) is taken into consideration during the training, so the bias is relatively small.

The results from experiment B in section 4.2.5 reveal that group DRO can find invariant relationships in the training domains. Looking at the result from p = 0.5 to p = 0.7, we can see that the group DRO classifier can correctly classify most birds regardless of their backgrounds (around or above 95% of the birds in the testing split), while the ERM classifier cannot. However, when the correlation between background and class label in the training sets becomes strong, the performance of the group DRO classifiers drops heavily (although still much better than the ERM ones), as we can see from the result, from p = 0.8 to 0.9, the percentage drops from 91.73% to 84.27%. It shows that group DRO may struggle to find the invariant relationships when the correlation between class labels and spurious features in the training domain becomes enough.

There are some limitations to this research. First of all, to learn the model with group DRO, we need to group the training data by the knowledge of spurious correlation. In our experiments, we mark not only the class label of the training example, but also the background during the data generation (section 4.1) to group the data, which is additional and expensive work for data collected from the real world. Besides, we use image examples with pure color backgrounds, but in the real-world situation, the backgrounds can be much more complicated, making it harder for the model to tell the invariant features and spurious features apart.

To summarize, group DRO has improved the OOD generalization performance over ERM, and group DRO learns invariant relationships from training sets. However, the ability of group DRO to find invariant relationships is limited when the spurious correlation in the training domains becomes robust.

7 Conclusions

In this research, we compare the OOD performance of group DRO and ERM and evaluate if group DRO can find the invariant relationship in the training domain. Semi-synthetic bird images are used in the experiments. In the first experiment, the classifiers learned by group DRO and ERM from the same training set are deployed to different testing sets to compare

the OOD performance of both methods. In the second experiment, the classifiers learned by both methods from different training sets perform classification on the same bird with different backgrounds to see if the models classify by the invariant relationship. The results of the experiments reveal that group DRO can improve the OOD generalization performance over ERM and the method can find the invariant relationships between the input and output in the training domain. However, its capability of finding the invariant relationship drops when the spurious correlation in the training set is strong.

In the future, the experiments can be repeated using real world images to see how group DRO performs when the spurious features like background are more complicated. Besides, the performance of group DRO can be compared to methods which are designed for finding invariant relationships like Invariant Risk Minimization [9] and Risk Extrapolation [14].

References

[1] Nagarajan, V., andreassen, anders, & Neyshabur, B. (2020). Understanding the Failure Modes of Out-of-Distribution Generalization. https://doi.org/10.48550/arXiv.2010.15775

[2] Sagawa, S., Koh, P. W., Hashimoto, T. B., & Liang, P. (2020). Distributionally Robust Neural Networks for Group Shifts: On the Importance of Regularization for Worst-Case Generalization. https://doi.org/10.48550/arXiv.1911.08731

[3] Lopez-paz, D., Nishihara, R., Chintala, S., Scholkopf, B., & Bottou, L. (2017). Discovering Causal Signals in Images. https://doi.org/10.48550/arXiv.1605.08179

[4] Ben-tal, A., Hertog, D. D., Waegenaere, A. D., Melenberg, B., & Rennen, G. (2012). Robust Solutions of Optimization Problems Affected by Uncertain Probabilities. https://doi.org/10.1287/mnsc.1120.1641

[5] Duchi, J., glynn, P., & Namkoong, H. (2016). Statistics of Robust Optimization: A Generalized Empirical Likelihood Approach. https://doi.org/10.48550/arXiv.1610.03425

[6] Peters, J., Bühlmann, P., & Meinshausen, N. (2015). Causal Inference Using Invariant Prediction: Identification and Confidence Intervals. https://doi.org/10.48550/arXiv.1501.01332

[7] Vapnik, V. N. (1998). Statistical Learning Theory. Wiley. https://www.wiley.com/en-fr/Statistical+Learning+Theory-p-9780471030034

[8] Wang, J., Lan, C., Liu, C., Ouyang, Y., Qin, T., Lu, W., Chen, Y., Zeng, W., & Yu, P. S. (2022). Generalizing to Unseen Domains: A Survey on Domain Generalization. https://doi.org/10.48550/arXiv.2103.03097

[9] Arjovsky, M., Bottou, L., Gulrajani, I., & Lopez-paz, D. (2020). Invariant Risk Minimization. https://doi.org/10.48550/arXiv.1907.02893

[10] Beery, S., Horn, G. V., & Perona, P. (2018). Recognition in Terra Incognita. https://doi.org/10.48550/arXiv.1807.04975

[11] Geirhos, R., Jacobsen, J., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., & Wichmann, F. A. (2021). Shortcut Learning in Deep Neural Networks. https://doi.org/10.48550/arXiv.2004.07780

[12] Wah, C., Branson, S., Welinder, P., Perona, P., & Belongie, S. (2011). The Caltech-UCSD Birds-200-2011 Dataset. https://authors.library.caltech.edu/27452/

[13] He, K., Zhang, X., Ren, S., & Sun, J. (2015). Deep Residual Learning for Image Recognition. https://doi.org/10.48550/arXiv.1512.03385

[14] Krueger, D., Caballero, E., Jacobsen, J., Zhang, A., Binas, J., Zhang, D., Priol, R. L., & Courville, A. (2020). Out-of-Distribution Generalization via Risk Extrapolation (REx). https://doi.org/10.48550/arXiv.2003.00688