Towards Making Effective Machine Learning Decisions Against Out-of-Distribution Data

Lakpa D. Tamang
School of Information Technology, Deakin University
Geelong, VIC, Australia
l.tamang@research.deakin.edu.au

Abstract

Conventional machine learning systems operate on the assumption of independent and identical distribution (i.i.d), where both the training and test data share a similar sample space, and no distribution shift exists between them. However, this assumption does not hold in practical deployment scenarios, making it crucial to develop methodologies that address the non-trivial task of data distribution shift. In our research, we aim to address this problem by developing ML algorithms that explicitly achieve promising performance when subjected to various types of out-of-distribution (OOD) data. Specifically, we approach the problem by categorizing the data distribution shifts into two types: covariate shifts and semantic shifts, and proposing effective methodologies to tackle each type independently and conjointly while validating them with different types of datasets. We aim to propose ideas that are compatible with existing deep neural networks to perform detection and/or generalization of the test instances that are shifted in semantic and covariate space, respectively.

CCS Concepts

• **Computing methodologies** → **Neural networks**.

Keywords

Out-of-Distribution, OOD Detection, OOD Generalization

ACM Reference Format:
Lakpa D. Tamang. 2024. Towards Making Effective Machine Learning Decisions Against Out-of-Distribution Data. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management (CIKM '24), October 21–25, 2024, Boise, ID, USA.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3627673.3680272

1 Problem Statement

Data distribution refers to the way in which data values are organized in a dataset, and it can offer valuable insights into patterns, characteristics, and relationships within the data [14]. Despite the fact that models remain unchanged, it is crucial to keep track of distribution changes in the data that could impact the model's performance. Generally, the test data has a much larger sample space and can be distributed heterogeneously rendering their distributional properties to be different from the training data. Such data are termed to be Out-of-Distribution (OOD) from what the machine learning model has learned. Ignoring the OOD issue and blindly following the independent and identical distribution (i.i.d) assumption can lead to models being biased, inaccurate, or unsuitable for the intended task, resulting in sub-optimal or harmful decision-making.

Following are a few of several reasons why a ML model might exhibit a data distribution shift.

Bias During Sample Selection: The concept of sample selection bias refers to a fault in the process of collecting or labelling data that leads to the uneven distribution of training examples. This results from the fact that the training examples were obtained through a biased method, which means they may not accurately reflect the environment where the classifier will be used.

Deployment Environment Changes: It is often true that data remains non-stationary to time and space change. Environments are dynamic in general, and sometimes the difficulties of matching the learning scenario (with training data) to the real world use (test data) are constrained by these changes.

Change in the Domain: Occasionally, a new sample might be collected from a different domain to represent the same category. In this regard, various domains may use various terms to refer to the same entity. The changes in the domain are characterized by the fact that the measurement system or the description technique of the feature of a dataset is changed.

Emergence of Novel Instances: The closed space assumption of traditional ML algorithms certainly doesn't hold true in the open world where unseen situations can emerge unexpectedly. Apparently, the model may experience a shift in the semantics of the representation it has learned from the training set as a result of the appearance of these unseen instances.

Although it is not reasonable to anticipate a learned model to predict accurately under any form of distribution shift, effective management of distribution shift with a selection of appropriate modelling strategies is considered imperative. In this research, we aim to tackle these problem by developing robust ML algorithms whose results are reliable against various types of data distribution shifts. We approach the problem by categorizing the data distribution shifts into two types: covariate shifts and semantic shifts, and propose different methodologies to address each type independently and conjointly. We validate our proposals with different types of datasets, particularly in the image classification scenario. Specifically, we propose ideas that are compatible with existing deep neural networks to perform detection and/or generalization of the test instances that are shifted in semantic and covariate space, respectively. Overall, the author's PhD work is intended to lay insights into making ML models more reliable and robust against OOD data, which is absolutely crucial for their effective deployment in critical applications.

CIKM '24, October 21–25, 2024, Boise, ID, USA

ACM ISBN 979-8-4007-0436-9/24/10
https://doi.org/10.1145/3627673.3680272

2 State-of-the-art

To handle the covariate shift problem, Out-of-distribution Generalization (OOD-G)/ Domain Generalization (DG) algorithms are extensively considered in the scholarly literature. These techniques attempt to handle covariate shift problem by generalizing the model trained in one domain to OOD data from other domain or characteristics but from similar conditional distribution. Some of the popular OOD-G/DG methods include (i) **Disentangled representation learning** [3] which aims to learn a model that is capable of identifying and disentangling the underlying factors hidden in the observable data in representation form. (ii) **Causal Learning** [12]: The objective of causal learning methods is to gain knowledge of the causal structure of the data and to make predictions about the outcome variable based on the identified causal variables. (iii) **Invariant Risk Minimization (IRM)** [4] The objective of IRM is to pinpoint characteristics or representations that result in consistent performance across various domains of a classifier. Other techniques such as Transfer Learning [7], Domain Adaption [11] are closely related discipline that attempt to solve the covariate shift problem in ML.

On the other hand, for handling concept/semantic shift problem, Out-of-distribution Detection (OOD-D) is generally considered. This technique tries to establish a decision boundary between the in-distribution (ID) inputs and the OOD inputs while also attempting to classify the ID classes effectively. The overall goal is to reject these OOD inputs such as they are not falsely classified into one of ID categories [8], store these OOD samples that emerge over time and then use it for incremental learning, or continual learning purposes [25]. Some of the popular OOD detection methods include: (i) **Density-based Methods** [1]: that detects semantically shifted data by modelling ID data with some probabilistic models, and flagging the test data in low-density regions as OOD data. (ii) **Output-based Methods** performs operations such as post-hoc processing in the output space such as penultimate layer of the neural network [8], scaling the outputs with temperature perturbations, or using distance metrics such as Mahalanobis or Cosine similarity. (iii) **Outlier-based Methods** [9]: exposes the model to auxiliary outliers during training as a regularization technique, generating lower confidence scores for OOD samples than that of ID inputs and segregating them.

3 Motivation and Approach

Many of the previously mentioned works have focused on developing exceptional solutions to address the problem of distribution shifts. However, despite these efforts, there remains a considerable gap, which can be primarily classified into two key aspects. (i) **Lack of model validation**: There is currently an absence of research to examine how models designed to handle covariate shifts perform when subjected to semantic shift environments and vice versa. Additionally, there is a lack of empirical evaluations and comparative assessments of existing methods on integrated datasets that incorporate both types of distribution shifts. (ii) **Absence of unified solution**: The current body of literature does not adequately promote the necessity of a versatile solution designed to address the challenges of concept and semantic shifts problem under a unified framework. Moreover, there has been limited investigation into developing comprehensive methods that account for the changes in data distribution by replicating real-world scenarios where both shift can co-exist, given a test data. Building on this motivation, we are compelled to raise the following question. *Is it possible to develop machine learning solutions that are capable of both generalizing and effectively detecting OOD inputs, based on the type of shifts they are exposed to?* In an effort to answer this question, our PhD research examines several intuitive measures, aimed at discovering efficient ways to adeptly tackle each type of shift, independently and concurrently. We plan to achieve this through the following four objectives:

(1) To explore the extant methodologies on data distribution shifts, with a particular emphasis on areas such as OOD detection, OOD generalization, that are formulated to handle concept and covariate shifts respectively and then empirically evaluate popular baseline methods in these areas using various image datasets.
(2) To develop and formalize a novel methodological frameworks for OOD detection. These frameworks should address limitations identified in the literature and propose innovative solutions to enhance the accuracy of the model, when test instances are subject to semantic shift. Furthermore, to validate the effectiveness of the developed algorithms through comprehensive experiments, utilizing real-world datasets and achieving comparable or superior performance against existing state-of-the-art (SOTA) methods.
(3) To develop a practical framework for OOD generalization that enables ML models to effectively adapt to covariate-shifted test instances and achieve superior performance on prominent datasets.
(4) To investigate the development of a hybrid classification framework that concurrently performs OOD detection, and generalization for effectively handling covariate, and semantic-shifted test instances.

4 Methodology

We wish to investigate methodological avenues, that are both intuitive and interconnected, with the ultimate goal of addressing each of our four objectives.

Objective 1: In the beginning phase of our method, we will collect and categorize scholarly research articles that are associated with handling two different types of shifts; covariate and semantic. Among vast majority of the accumulated papers, we will perform empirical evaluation on their methodologies, i.e., reproduce the results of the popular papers with public source codes. Then, we will make a comparative assessment on how the methods develop for handling covariate shift will behave on semantically shifted dataset and vice versa. To determine the feasibility and practicality of the existing techniques, we will validate them by re-implementing the algorithms using wide variety of popular neural networks architectures such as Convolutional Neural Network, DenseNet, ResNet, and WideResNet.

Objective 2: In the context of OOD detection, the availability of outliers as a known priori is deemed impractical, yet, several studies [9], [19] have shown that it is effective to use a set of auxiliary outliers as a regularization to yield low confidence scores for OOD

data. While, there has been approaches of using such outliers aided by complex sampling techniques [5], augmentations [23], or feature space operations [17], the notion of generating low confidence score for OOD samples exclusively from an OE [9] regularized classifier has not been exploited to its full extent. Therefore, we formulate this particular objective with the motivation of tackling aforementioned problem, and develop a technique by solely exercising the use of confidence scores in an OE regularized classifier with simple and straightforward solution space. In our methodology, we will adopt OE [9] method to train the model. OE is a regularization technique that involves learning from additional datasets containing outliers or OOD samples along with standard training data. The goal is to expose the network to diverse OOD examples during training, so that the model learns a more conservative concept of the ID data to distinguish them from their OOD counterparts. Specifically, we will explore the use of neural network's confidence scores of both ID and OOD inputs during training, by aiming to surge the decision boundary between ID and OOD test instances. Evaluation will be done on set of standard metrics that has been popular throughout the literature. Specifically, we will use area under the receiver operating curve (AUROC), area under the precision recall curve (AUPR), and the false positive rate (FPR) at a true positive rate of 95 % for ID samples.

Objective 3: Techniques such as representation learning and invariant learning have received much of the attentions lately [15], [18] Invariant representation learning will learn features that remain unchanged across multiple domains. For instance, a face recognition sytem should recognize a person regardless of the lighting condition, the person's pose, or whether they are wearing glasses. Learning an invariant representation would ensure that these factors don't significantly affect the recognition. Therefore, one possible direction for this objective is to try to delve into these concepts for realizing OOD generalization. Specifically, we will focus on finding latent characteristics and incorporate them in learning. By latent, we mean unseen or unobserved domain that shares some underlying structure or patterns with the known domains but might also have its own unique characteristics. Our investigation will seek to identify and leverage these latent structures that are common across multiple known domains to be able to generalize well to latent (or unseen) domains. This will involve techniques that encourage the learning of features that are both discriminative for the task (e.g., recognizing face in above example) and invariant across the different known domains.

We will carry out the experiments to validate our method primarily in image classification task with benchmark datasets including PACS [10], OfficeHome [16], VLCS [6]. Our evaluation metrics will include determining the accuracy of the model where training and testing will be done in entirely different datasets, with an assumption of covariate shifts between them. Our approach should aim to achieve better accuracy across several datasets as aforementioned, compared to existing methodologies in the similar field.

Objective 4: Similar to the motivation built up for Objective 1, where we advocated to perform empirical evaluation of existing methods in the field, here we extend that objective in an intuitive way by investigating to develop hybrid models that can perform both detection and generalization. To fulfill this objective, we plan to follow several recent research [2], [20], that aim to develop a holistic approach which can enhance the algorithm's practical ability to learn from intricate, multifaceted shifts, improving generalization and robustness across diverse situations. One of the preliminary proposal for developing this approach is to make the model learn from a wide variety data consisting of mixed composition of both distribution shifts. By doing so, we will target to extract shift-invariant representation and utilize for learning the model which is capable of handling both detection and generalization task simultaneously. With this we design ML algorithms to be inherently adaptive, and detective while pursuing to handle both covariate and semantic shifts under one cohesive framework. We will use datasets such as PACS, DomainNet, Office-Home used in existing literature [22] to validate the model generalization against covariate shift. Furthermore, we will use metrics such as AUROC, AUPR, FPR95 for inspecting OOD detection effectiveness in semantic shifted environment. For confirming model's robustness against covariate shifted data, we will use accuracy metrics which basically checks what proportion of the covariate shifted data is correctly classified by model.

5 Preliminary Results

To this end, we have completed the task of implementing our first OOD detection method for Obj. 2. Specifically, we have successfully utilized auxiliary outliers with some modifications on the cost function of OE method [9], which include augmenting it with a supplementary constraint. The modification is that, during fine-tuning, our method imposes an explicit penalty on cases where the confidence score of input OOD data is higher than that of its ID counterparts. Next, these penalized events are substituted with a margin value between 0 and 1, which represents the estimated difference between the confidence scores of ID and OOD data. As shown in Fig. 1, our method clearly obtained better separation between the confidence scores of ID and OOD test data compared to baseline MSP [8] and OE [9] techniques. Specifically, through our findings, we showed that augmenting the learning objective of the OE regularized classifier with a supplementary constraint, which penalizes high confidence scores for OOD inputs compared to ID inputs, significantly enhances its OOD detection performance while maintaining its ID accuracy. We successfully validated the proposed methodology by performing extensive experiments on various benchmark datasets. A portion of our comprehensive results on CIFAR benchmarks compared against several other methods is presented in Table. 1. We demonstrated through several results comparisons on AUROC, AUPR, and FPR95 that our method significantly outperforms multiple OOD detection methods.

Table 1: Comparison of OOD detection results on different ID datasets fine-tuned on a WRN architecture. Best and second best values are reported in bold, and underline respectively. Arrows represent the direction towards optimum value.

Method \ ID Data	CIFAR-10				CIFAR-100			
	AUROC ↑	AUPR ↑	FPR95 ↓	ID-ACC ↑	AUROC ↑	AUPR ↑	FPR95 ↓	ID-ACC ↑
OE [9]	98.65±0.03	<u>98.6±0.05</u>	6.21±0.13	94.83±0.06	88.51±0.15	87.43±0.16	42.12±0.44	75.75±0.11
Energy [13]	<u>98.68±0.03</u>	98.49±0.05	<u>5.88±0.13</u>	94.35±0.07	87.567±0.06	87.77±0.09	48.93±0.19	74.77±0.11
MixOE [21]	90.85±0.12	90.48±0.2	41.46±0.36	94.53±0.03	78.02±0.22	73.98±0.29	61.34±0.38	75.17±0.18
DivOE [24]	98.46±0.04	98.38±0.05	7.15±0.19	<u>95.01±0.05</u>	87.42±0.08	86.45±0.06	44.21±0.27	<u>75.83±0.09</u>
Ours	**98.79±0.02**	**98.77±0.03**	**5.14±0.11**	**95.28±0.06**	**90.93±0.13**	**90.28±0.21**	**37.54±0.35**	**76.12±0.04**

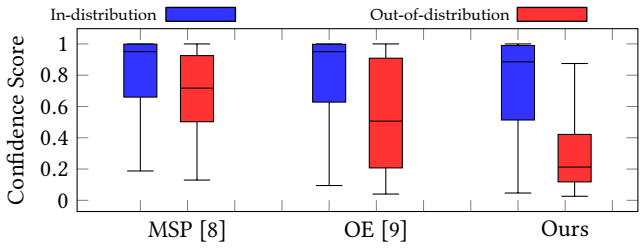

Figure 1: OOD Detection comparison between our method and baselines. For good distinction of OOD and ID data their ideal confidence scores should be close to 0, and 1 respectively.

6 Conclusion and Future Works

In this paper, we have presented the research problem and methodology of the author's ongoing PhD work, that aims to provide a solution to the nontrivial OOD problem. Particularly, we divide our problem into two categories of data distribution shifts: covariate shift and semantic shift. To address these problem, we plan to develop algorithms that can perform promisingly under these shift conditions. More importatnly, by considering that these shifts can co-exists in practice, we envision to develop holistic approaches which accounts for solving both problem simultaneously. To this end, the implementation of one such OOD detection algorithm, addressing the covariate shift, has been exercised. Apart from that, currently we are putting into practice the adversarial learning approaches to fit them into handling covariate shift problem: the primary objective being to generate distribution agnostic features such that generalization is possible. Meanwhile, we are validating our approaches in image datasets, and we are hopeful that they can be extended to non-image data such as graphs, tables, and texts. Besides that, for the future works, we plan to conduct intensive experiments on synthetic data, and multimodal approaches, with the goal of creating unified solution for the OOD detection and generalization problem. We believe that our work will provide crucial insight into developing solutions to realize effective machine learning models whose results are reliable and their decisions are trustworthy.

References

[1] Davide Abati, Angelo Porrello, Simone Calderara, and Rita Cucchiara. 2019. Latent space autoregression for novelty detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 481–490.

[2] Haoyue Bai, Gregory Canal, Xuefeng Du, Jeongyeol Kwon, Robert D Nowak, and Yixuan Li. 2023. Feed two birds with one scone: Exploiting wild data for both out-of-distribution generalization and detection. In *International Conference on Machine Learning*. PMLR, 1454–1471.

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

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

[5] Jiefeng Chen, Yixuan Li, Xi Wu, Yingyu Liang, and Somesh Jha. 2021. Atom: Robustifying out-of-distribution detection using outlier mining. In *Machine Learning and Knowledge Discovery in Databases. Research Track: European Conference, ECML PKDD 2021, Bilbao, Spain, September 13–17, 2021, Proceedings, Part III 21*. Springer, 430–445.

[6] Myung Jin Choi, Joseph J Lim, Antonio Torralba, and Alan S Willsky. 2010. Exploiting hierarchical context on a large database of object categories. In *2010 IEEE computer society conference on computer vision and pattern recognition*. IEEE, 129–136.

[7] Chuong B Do and Andrew Y Ng. 2005. Transfer learning for text classification. *Advances in neural information processing systems* 18 (2005).

[8] Dan Hendrycks and Kevin Gimpel. 2016. A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks. In *International Conference on Learning Representations*.

[9] Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. 2018. Deep anomaly detection with outlier exposure. *arXiv preprint arXiv:1812.04606* (2018).

[10] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. 2017. Deeper, broader and artier domain generalization. In *Proceedings of the IEEE international conference on computer vision*. 5542–5550.

[11] Zijian Li, Ruichu Cai, Guangyi Chen, Boyang Sun, Zhifeng Hao, and Kun Zhang. 2024. Subspace identification for multi-source domain adaptation. *Advances in Neural Information Processing Systems* 36 (2024).

[12] Chang Liu, Xinwei Sun, Jindong Wang, Haoyue Tang, Tao Li, Tao Qin, Wei Chen, and Tie-Yan Liu. 2021. Learning causal semantic representation for out-of-distribution prediction. *Advances in Neural Information Processing Systems* 34 (2021), 6155–6170.

[13] Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. 2020. Energy-based out-of-distribution detection. *Advances in neural information processing systems* 33 (2020), 21464–21475.

[14] Jose G. Moreno-Torres, Troy Raeder, Rocío Alaiz-Rodríguez, Nitesh V. Chawla, and Francisco Herrera. 2012. A Unifying View on Dataset Shift in Classification. *Pattern Recognition* 45, 1 (Jan. 2012), 521–530. https://doi.org/10.1016/j.patcog.2011.06.019

[15] Jiaxin Qi, Kaihua Tang, Qianru Sun, Xian-Sheng Hua, and Hanwang Zhang. 2022. Class is invariant to context and vice versa: on learning invariance for out-of-distribution generalization. In *European Conference on Computer Vision*. Springer, 92–109.

[16] Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. 2017. Deep hashing network for unsupervised domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 5018–5027.

[17] Haotian Wang, Kun Kuang, Long Lan, Zige Wang, Wanrong Huang, Fei Wu, and Wenjing Yang. 2023. Out-of-distribution generalization with causal feature separation. *IEEE Transactions on Knowledge and Data Engineering* (2023).

[18] Ruoyu Wang, Mingyang Yi, Zhitang Chen, and Shengyu Zhu. 2022. Out-of-distribution generalization with causal invariant transformations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 375–385.

[19] Qitian Wu, Yiting Chen, Chenxiao Yang, and Junchi Yan. 2023. Energy-based out-of-distribution detection for graph neural networks. *arXiv preprint arXiv:2302.02914* (2023).

[20] Jingkang Yang, Kaiyang Zhou, and Ziwei Liu. 2023. Full-spectrum out-of-distribution detection. *International Journal of Computer Vision* (2023), 1–16.

[21] Jingyang Zhang, Nathan Inkawhich, Randolph Linderman, Yiran Chen, and Hai Li. 2023. Mixture outlier exposure: Towards out-of-distribution detection in fine-grained environments. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 5531–5540.

[22] Xingxuan Zhang, Yue He, Renzhe Xu, Han Yu, Zheyan Shen, and Peng Cui. 2023. Nico++: Towards better benchmarking for domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 16036–16047.

[23] Fei Zhu, Zhen Cheng, Xu-Yao Zhang, and Cheng-Lin Liu. 2023. OpenMix: Exploring Outlier Samples for Misclassification Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 12074–12083.

[24] Jianing Zhu, Yu Geng, Jiangchao Yao, Tongliang Liu, Gang Niu, Masashi Sugiyama, and Bo Han. 2024. Diversified outlier exposure for out-of-distribution detection via informative extrapolation. *Advances in Neural Information Processing Systems* 36 (2024).

[25] Lanyun Zhu, Tianrun Chen, Jianxiong Yin, Simon See, and Jun Liu. 2023. Continual semantic segmentation with automatic memory sample selection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3082–3092.