On the Out-of-Distribution Generalization of Self-Supervised Learning

Wenwen Qiang[1], Jingyao Wang[1], Zeen Song, Jiangmeng Li*, Changwen Zheng

ICML International Conference On Machine Learning

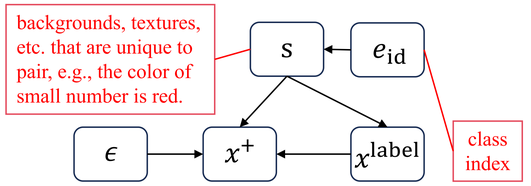
Introduction

In this paper, we focus on the out-of-distribution (OOD) generalization of self-supervised learning (SSL). By analyzing the mini-batch construction during the SSL training phase, we first give one plausible explanation for SSL having OOD generalization. Then, from the perspective of data generation and causal inference, we analyze and conclude that SSL learns spurious correlations during the training process, which leads to a reduction in OOD generalization. To address this issue, we propose a post-intervention distribution (PID) grounded in the Structural Causal Model. PID offers a scenario where the spurious variable and label variable is mutually independent. Besides, we demonstrate that if each mini-batch during SSL training satisfies PID, the resulting SSL model can achieve optimal worst-case OOD performance. This motivates us to develop a batch sampling strategy that enforces PID constraints through the learning of a latent variable model. Through theoretical analysis, we demonstrate the identifiability of the latent variable model and validate the effectiveness of the proposed sampling strategy. Experiments conducted on various downstream OOD tasks demonstrate the effectiveness of the proposed sampling strategy.

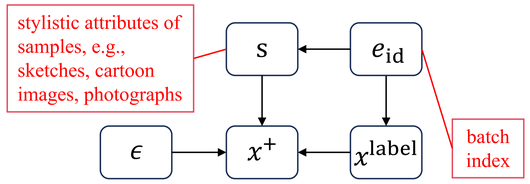
Contributions

1. Analysis of SSL Batch Construction: We provide a detailed analysis of how mini-batch construction in SSL influences OOD generalization
2. Causal Framework for SSL: We introduce a causal framework to understand and mitigate the impact of spurious correlations for SSL.
3. PID-Based Sampling Strategy: We propose a theoretically grounded mini-batch sampling strategy that ensures the generated batches conform to PID, improving OOD performance.
4. Empirical Validation: Extensive experiments demonstrate the significant improvements of our method in OOD generalization.

Motivation and Analysis



(a) Example task related to ColoredMNIST dataset



(b) Example task related to PACS dataset

Figure 1: Structural Causal Model (SCM). These instances illustrate the variability in the causal relationship between x^{label} and s due to environmental changes. The black squares are variables and the arrows indicate causality.

Self-supervised learning (SSL) enables model training without labels and has achieved performance on par with, or exceeding, supervised methods. However, despite strong in-distribution results, SSL models often struggle with out-of-distribution (OOD) generalization, critical in real-world scenarios where data distributions shift over time. To analyze this, we examine how mini-batches are constructed in SSL. Discriminative SSL enforces invariance across augmented views; generative SSL reconstructs masked inputs. In both, each anchor pair (augmentation or reconstruction) forms a pseudo-class, turning mini-batch training into a task sampled from a distribution over classes. However, we conduct causal analyses and find that SSL is prone to spurious correlations—models may exploit background or texture cues that vary across tasks. The constructed SCMs show that these confounders distort similarity or reconstruction objectives and cannot be removed by a unified causal rule. Consequently, SSL models may fail to capture true task structures, limiting their OOD generalization.

Revisiting SSL from a Pairwise Perspective

When we consider the anchor as the label or the center of clustering, each mini-batch in the training phase can be viewed as a multi-class classification task. $X_{tr}^{aug} = \{x_i^+, x_i^{\text{anchor}}\}_{i=1}^N$ consists of data from N categories, where x_i^+ is the positive sample of the i-th category whose clustering center is x_i^{anchor}. Furthermore, the variability of data across mini-batches implies that each mini-batch corresponds to a distinct training task or domain.

Methodology

To address the reliance on spurious correlations, we introduce the concept of the Post-Intervention Distribution (PID), where latent variables s are independent of each anchor sample's label x^{label}, i.e., $p^{\text{PI}}(x^{\text{label}} \mid s) = p^{\text{PI}}(x^{\text{label}})$ (Definition 3.2). We theoretically show that when the training sample distribution satisfies the PID condition, the worst-case generalization risk of self-supervised learning (SSL) reaches its lower bound, ensuring optimal OOD performance under worst-case scenarios (Theorem 3.4). Motivated by this, we propose a new mini-batch sampling strategy: (i) estimate the latent variable s for each sample using a learned latent model and compute its propensity score $p(x^{\text{label}} \mid s)$; (ii) match sample pairs with similar or identical scores to enforce conditional independence between s and x^{label} within each batch; (iii) construct mini-batches approximating samples from the PID. This effectively removes spurious correlations during training and significantly improves the OOD generalization of SSL models. Extensive theoretical analysis supports the validity and robustness of the method.

Algorithm 1 Proposed Mini-Batch Sampling Strategy

Input: Training dataset $X^{tr} = \{x_i^+, x_i^{\text{label}}\}_{i=1}^{\text{mu}}$, balancing score function $ba(\cdot)$, distance metric $d : ba(\cdot) \times ba(\cdot) \to \mathbb{R}$
Output: Mini-batch data D^{PI} consisting of $a+1$ examples

$D^{\text{PI}} \leftarrow \emptyset$; $i \leftarrow 0$
while $i = 0$ **do**
  Randomly sample $(x_i^+, x_i^{\text{label}})$ from X_{aug}^{tr}, add to D^{PI}; Compute $ba(s_i)$ from $(x_i^+, x_i^{\text{label}})$
  $i \leftarrow i + 1$
end while
for $1 \le i \le a$ **do**
  $j \leftarrow \arg\min_{x_j^+ \in X_{aug}^{tr} \setminus D^{\text{PI}}} d(ba(s_j), ba(s_i))$
  Add $(x_j^+, x_j^{\text{label}})$ to D^{PI}
  $i \leftarrow i + 1$
end for

Experimental Resutls

We conduct extensive experiments on various downstream tasks, including unsupervised learning, semi-supervised learning, transfer learning, and few-shot learning. The results show that our method achieves stable performance improvements on multiple baselines. In addition, the experiments conducted on out-of-distribution datasets further demonstrate the effectiveness of our method for improving SSL generalization.

Method	VOC 07 detection			VOC 07+12 detection			COCO detection			COCO instance segmentation		
	$\mathbf{AP_{50}}$	$\mathbf{AP}$	$\mathbf{AP_{75}}$	$\mathbf{AP_{50}}$	$\mathbf{AP}$	$\mathbf{AP_{75}}$	$\mathbf{AP_{50}}$	$\mathbf{AP}$	$\mathbf{AP_{75}}$	$\mathbf{AP_{50}^{mask}}$	$\mathbf{AP^{mask}}$	$\mathbf{AP_{75}^{mask}}$
Supervised	74.4	42.4	42.7	81.3	53.5	58.8	58.2	38.2	41.2	54.7	33.3	35.2
SimCLR (Chen et al., 2020)	75.9	46.8	50.1	81.8	55.5	61.4	57.7	37.9	40.9	54.6	33.3	35.3
MoCo (He et al., 2020)	77.1	46.8	52.5	82.5	57.4	64.0	58.9	39.3	42.5	55.8	34.4	36.5
BYOL (Grill et al., 2020b)	77.1	47.0	49.9	81.4	55.3	61.1	57.8	37.9	40.9	54.3	33.2	35.0
SimSiam (Chen & He, 2021)	77.3	48.5	52.5	82.4	57.0	63.7	59.3	39.2	42.1	56.0	34.4	36.7
SwAV (Caron et al., 2020)	75.5	46.5	49.6	82.6	56.1	62.7	58.6	38.4	41.3	55.2	33.8	35.9
VICRegL (Bardes et al., 2022)	75.9	47.4	52.3	82.6	56.4	62.9	59.2	39.8	42.1	56.5	35.1	36.8
SimCLR + Ours	77.6	50.1	51.7	85.3	58.4	63.9	59.2	40.6	43.9	57.1	35.9	37.1
MoCo + Ours	79.4	50.2	**54.9**	**86.1**	**60.2**	**66.1**	614	42.1	44.9	**59.2**	36.9	38.8
BYOL + Ours	79.1	50.4	51.9	83.9	58.7	64.1	60.6	39.9	43.7	56.2	35.1	38.6
SimSiam + Ours	**80.5**	**50.8**	54.4	85.2	59.5	66.1	62.3	42.5	43.9	58.1	37.2	39.8
SwAV + Ours	77.9	49.3	51.8	84.9	58.1	65.8	62.1	40.2	43.9	56.9	37.3	37.9
VICRegL + Ours	77.9	50.4	53.9	85.2	58.8	65.3	**63.1**	**42.2**	**45.3**	59.1	**37.8**	**39.9**

Resources

Website | CODE | Paper | ICML Page