ISCAS 中国科学院软件研究所学术年会 暨重点实验室科技活动周 2025 第十届
学术论文

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

强文文*，王婧瑶*，宋泽恩，李江梦，郑昌文

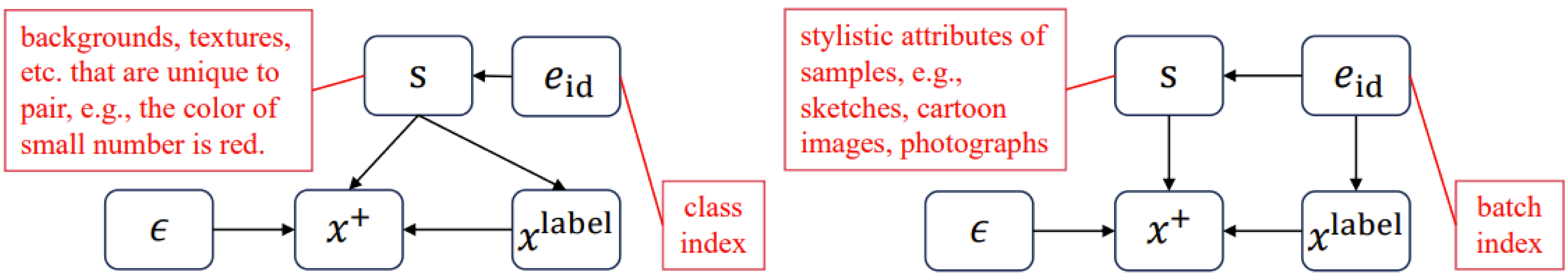
International Conference on Machine Learning
王婧瑶 Jingyao_wang0728@163.com

概述

本文聚焦自监督学习（SSL）的分布外（OOD）泛化能力，首先从小批量构建机制解释其部分泛化能力来源，并指出训练中易引入虚假关联，削弱性能。为此，我们引入结构因果模型中的干预后分布（PID），并证明若每批数据满足PID，SSL可实现最优OOD性能。基于此，提出一项基于潜变量模型的PID采样策略，并通过理论与实证验证其有效性。

动机与分析

自监督学习（SSL）作为无需标注的训练范式，在多个下游任务中表现优异，但在分布外（OOD）数据上仍面临挑战，尤其当数据分布随时间变化时。我们分析了判别式与生成式SSL在小批量构建中的任务分布特性，指出其本质上可视为在离散任务分布上的元学习。然而，模型易受虚假关联因素（如背景、纹理）干扰，导致锚点对的相似性或重建质量失真，从而削弱OOD泛化能力。



(a) Example task related to ColoredMNIST dataset (b) Example task related to PACS dataset

图一 两个示例以说明了x^{label}标签和s之间的因果关系会因环境变化而发生变化。方块表示变量，箭头表示因果关系

方法

为缓解虚假关联问题，我们引入干预后分布（PID），要求不可观测变量与锚点样本标签独立，并证明若训练数据满足PID，SSL可在最坏情形下实现最优泛化。据此，提出基于隐变量模型的PID采样策略，通过倾向得分匹配样本，实现条件独立配对，近似生成符合PID的数据批次，从而削弱虚假关联，提升OOD泛化能力。理论分析验证了其有效性与稳健性。

Theorem 3.4. *From a Bayesian perspective, the alignment part of the SSL learning objective, e.g., constrain samples under the same pair to be similar in the feature space, can be expressed as* $\max p_f(x^{\text{label}}|x^+)$. *Given f, the risk on a batch with $e \in \mathcal{D}$ as the distributional constraint can be presented as:* $\mathcal{L}^e(f) = \mathbb{E}_{p^e(x^+, x^{\text{label}})} - \log p_f(x^{\text{label}}|x^+)$, *where* $p^e(x^+, x^{\text{label}})$ *denotes the joint distribution. Under* ***Assumption*** *3.3, when* $f^* = \arg\max \mathcal{L}^{\text{PID}}(f)$, *we have f^* is the minimax optimal across all elements in $\mathcal{D}$, e.g.,* $f^* = \arg_f \min \max_{e\in\mathcal{D}} \mathcal{L}^e(p_f(x^{\text{label}}|x^+))$.

Theorem 4.3. *Suppose that* $p_\theta^e(x^+, s|x^{\text{label}}) = p_{\mathrm{f}}(x^+|s, x^{\text{label}})p_{g,\mathrm{A}}(s|x^{\text{label}})$ *and the generation process of X^+ can be represented by the SCM depicted in Figure 1, a sufficient condition for* $\theta = (\mathrm{f}, g, \mathrm{A})$ *to be* $\sim_{\mathrm{A}}$*-identifiable is given as: 1) Suppose that* $p_\epsilon(x^+ - \mathrm{f}(x^{\text{label}}, s)) = p_{\mathrm{f}}(x^+|x^{\text{label}}, s)$, ϕ_ε *is the characteristic function of* $p_\epsilon(x^+ - \mathrm{f}(x^{\text{label}}, s))$, *and the set* $\{x^+|\phi_\varepsilon(x^+) = 0\}$ *has measure zero; 2) The sufficient statistics* T *are differentiable almost everywhere, and* $[T_{ij}]_{1\le j\le k}$ *are linearly independent on any subset of X^+ with measure greater than zero; 3) There exist $nk+1$ distinct pairs* $(x_0^{\text{label}}, e_0), \cdots, (x_n^{\text{label}}k, e_{nk})$ *such that the $nk \times nk$ matrix* $\mathrm{L} = (\lambda^{e_1}(x_1^{\text{label}}) - \lambda^{e_0}(x_0^{\text{label}}), \cdots, \lambda^{e_{nk}}(x_{nk}^{\text{label}}) - \lambda^{e_0}(x_0^{\text{label}}))$ *is invertible.*

Theorem 4.7. *If* $d(ba(s_j), ba(s_i)) = 0$ *in* ***Algorithm*** *1, the obtained mini-batch is regarded as sampling from a PID, e.g.,* $\hat{p}(x^{\text{label}}|s) = p^{\text{PI}}(x^{\text{label}})$.

实验

为验证方法有效性，我们在ImageNet-100与ImageNet的无监督分类中评估所提mini-batch采样策略对判别式与生成式SSL的提升效果，并拓展至半监督学习、目标检测、实例分割及少样本任务。所有实验仅更改批次生成方式，模型结构与超参数保持不变。结果显示，该策略在各任务中普遍提升超2%，显著增强SSL的因果鲁棒性与OOD泛化能力。

Method	ImageNet-100		ImageNet	
	Top-1	Top-5	400 Epochs	1000 Epochs
SimCLR (Chen et al., 2020)	70.15 ± 0.16	89.75 ± 0.14	69.24 ± 0.21	70.45 ± 0.30
MoCo (He et al., 2020)	72.80 ± 0.12	91.64 ± 0.11	69.76 ± 0.14	71.16 ± 0.23
SimSiam (Chen & He, 2021)	73.01 ± 0.21	92.61 ± 0.27	70.86 ± 0.34	71.37 ± 0.22
Barlow Twins (Zbontar et al., 2021)	75.97 ± 0.23	92.91 ± 0.19	70.22 ± 0.15	73.29 ± 0.13
SwAV (Caron et al., 2020)	75.78 ± 0.16	92.86 ± 0.15	70.78 ± 0.34	75.32 ± 0.11
DINO (Caron et al., 2021)	75.43 ± 0.18	93.32 ± 0.19	71.98 ± 0.26	73.94 ± 0.29
RELIC v2 (Tomasev et al., 2022)	75.88 ± 0.15	93.52 ± 0.13	71.84 ± 0.21	72.17 ± 0.20
VICRegL (Bardes et al., 2022)	75.96 ± 0.19	92.97 ± 0.26	72.14 ± 0.20	75.07 ± 0.23
SimCLR + Ours	73.32 ± 0.15	91.74 ± 0.18	72.24 ± 0.20	73.66 ± 0.25
MoCo + Ours	74.71 ± 0.22	93.89 ± 0.17	72.04 ± 0.21	74.06 ± 0.20
SimSiam + Ours	75.66 ± 0.18	95.02 ± 0.21	72.96 ± 0.22	73.67 ± 0.17
Barlow Twins + Ours	77.77 ± 0.18	94.99 ± 0.20	73.08 ± 0.21	75.89 ± 0.17
SwAV + Ours	76.99 ± 0.11	95.03 ± 0.20	73.25 ± 0.24	77.42 ± 0.21
DINO + Ours	77.47 ± 0.15	**96.01** ± 0.17	74.21 ± 0.20	75.99 ± 0.17
VICRegL + Ours	**78.20** ± 0.14	95.07 ± 0.21	**74.91** ± 0.14	**77.77** ± 0.21

Method	Epochs	1%		10%	
		Top-1	Top-5	Top-1	Top-5
MoCo (He et al., 2020)	200	43.8 ± 0.2	72.3 ± 0.1	61.9 ± 0.1	84.6 ± 0.2
BYOL (Grill et al., 2020b)	200	54.8 ± 0.2	78.8 ± 0.1	68.0 ± 0.2	88.5 ± 0.2
BYOL + Ours	200	46.5 ± 0.2	74.4 ± 0.2	63.6 ± 0.3	85.6 ± 0.2
MoCo + Ours	200	**57.4** ± 0.2	**80.1** ± 0.2	**71.4** ± 0.2	**90.2** ± 0.1
SimCLR (Chen et al., 2020)	1000	48.3 ± 0.2	75.5 ± 0.1	65.6 ± 0.1	87.8 ± 0.2
MoCo (He et al., 2020)	1000	52.3 ± 0.1	77.9 ± 0.2	68.4 ± 0.1	88.0 ± 0.2
BYOL (Grill et al., 2020b)	1000	56.3 ± 0.2	79.6 ± 0.2	69.7 ± 0.2	89.3 ± 0.1
Barlow Twins (Zbontar et al., 2021)	1000	55.0 ± 0.1	79.2 ± 0.1	67.7 ± 0.2	89.3 ± 0.2
RELIC v2 (Tomasev et al., 2022)	1000	55.2 ± 0.2	80.0 ± 0.1	68.0 ± 0.2	88.9 ± 0.2
VICRegL (Bardes et al., 2022)	1000	54.9 ± 0.1	79.6 ± 0.2	67.2 ± 0.1	89.4 ± 0.2
SimCLR + Ours	1000	50.8 ± 0.2	77.8 ± 0.2	67.3 ± 0.1	89.9 ± 0.2
MoCo + Ours	1000	53.9 ± 0.2	78.9 ± 0.2	71.2 ± 0.1	89.5 ± 0.1
BYOL + Ours	1000	**58.9** ± 0.2	**81.9** ± 0.2	**72.1** ± 0.2	91.2 ± 0.1
Barlow Twins + Ours	1000	57.6 ± 0.2	80.6 ± 0.1	68.9 ± 0.2	**91.8 ± 0.2**

表一 在无监督设定（左）与半监督设定（右）下的对比实验结果。最优结果被加粗表示。

Method	VOC 07 detection			VOC 07+12 detection			COCO detection			COCO instance segmentation		
	AP_{50}	AP	AP_{75}	AP_{50}	AP	AP_{75}	AP_{50}	AP	AP_{75}	AP_{50}^{mask}	AP^{mask}	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**

表二 基于 C4 骨干网络的目标检测和实例分割迁移学习。“AP”表示平均精度，“APN”表示 IoU（交并比）阈值为 N% 时的平均精度。