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 Generalization and Causal Explanation in Self-Supervised Learning

强文文，宋泽恩，顾子茵，李江梦，郑昌文，孙富春，熊辉，李江梦

International Journal of Computer Vision

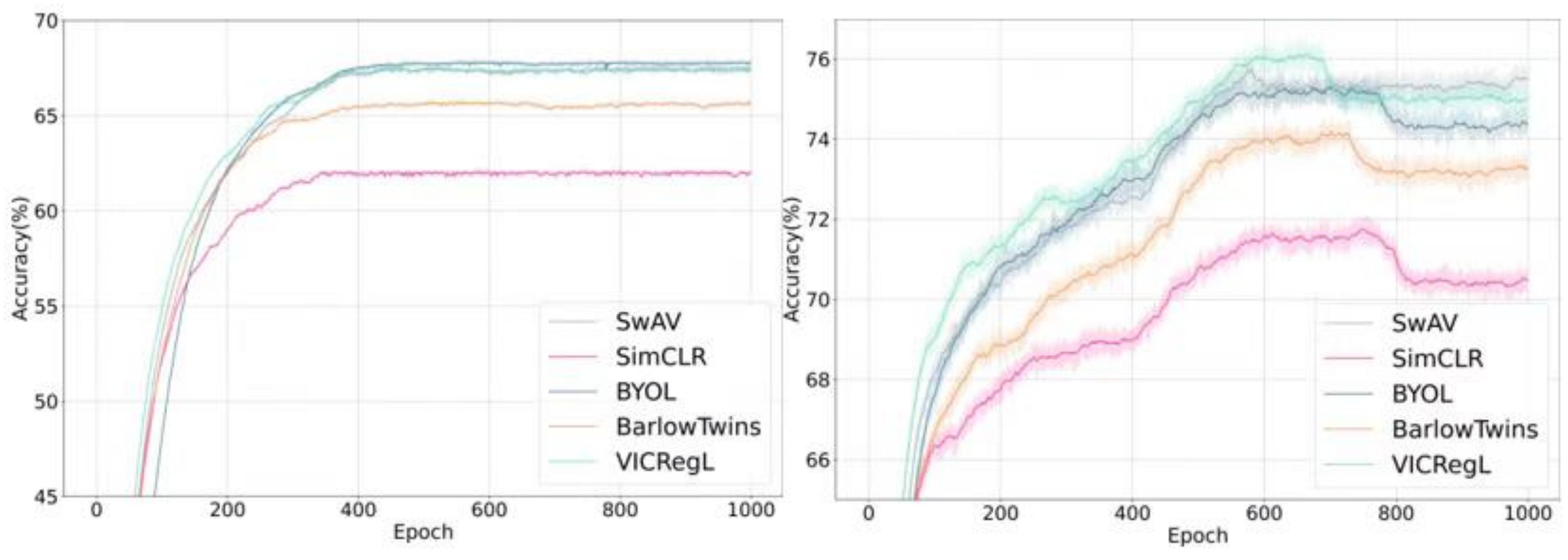
强文文 Jingyao_wang0728@163.com

概述

自监督学习虽具良好泛化性，但易在后期层和训练末期过拟合。本文发现：泛化特征由早期层学习，编码率下降可指示过拟合程度。为此，提出记忆消解机制（UMM），通过对齐早期与末层特征分布，提升编码率，缓解过拟合。UMM基于双层优化，并有因果解释，显著增强多任务下的SSL泛化能力。

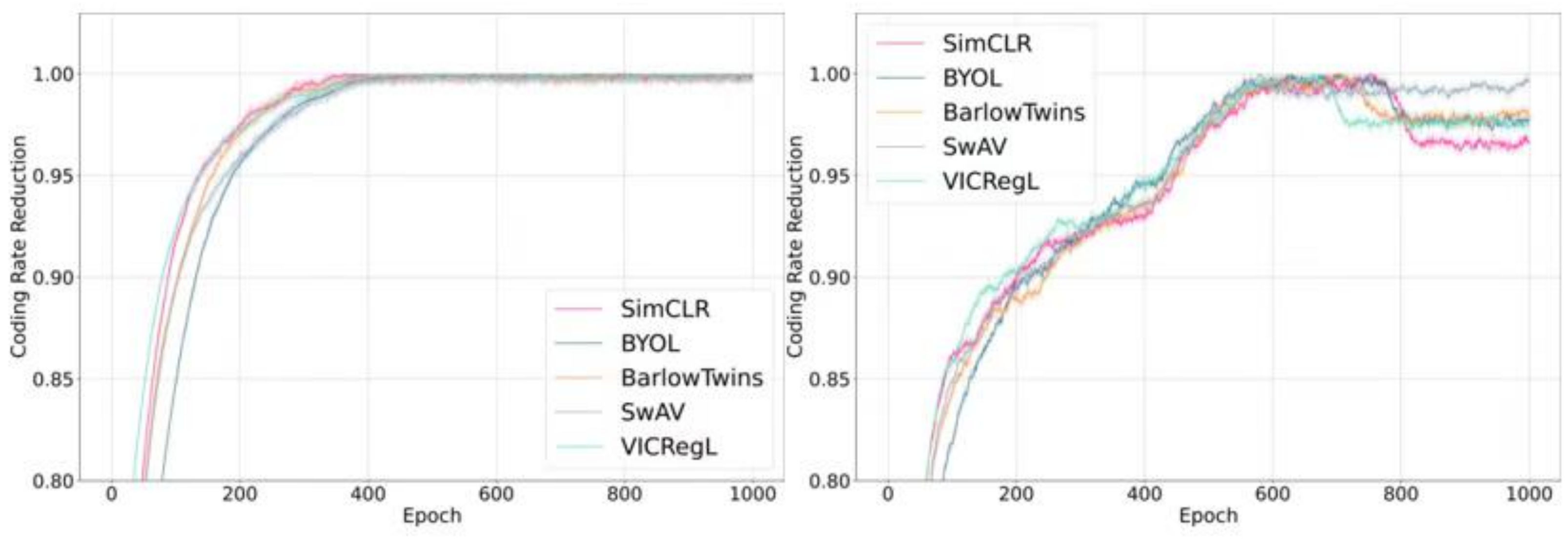
动机与分析

自监督学习通过基于实例的范式从未标注数据中提取语义表示，广泛用于下游任务。然而，因深度网络参数远多于训练样本，易出现过拟合。实验表明，训练轮次增加时，末层特征在验证集上准确率先升后降，反映其易记忆训练数据；而中间层特征保持稳定，展现出更强泛化能力。



中间层特征和最后层特征在验证集上的准确率变化趋势

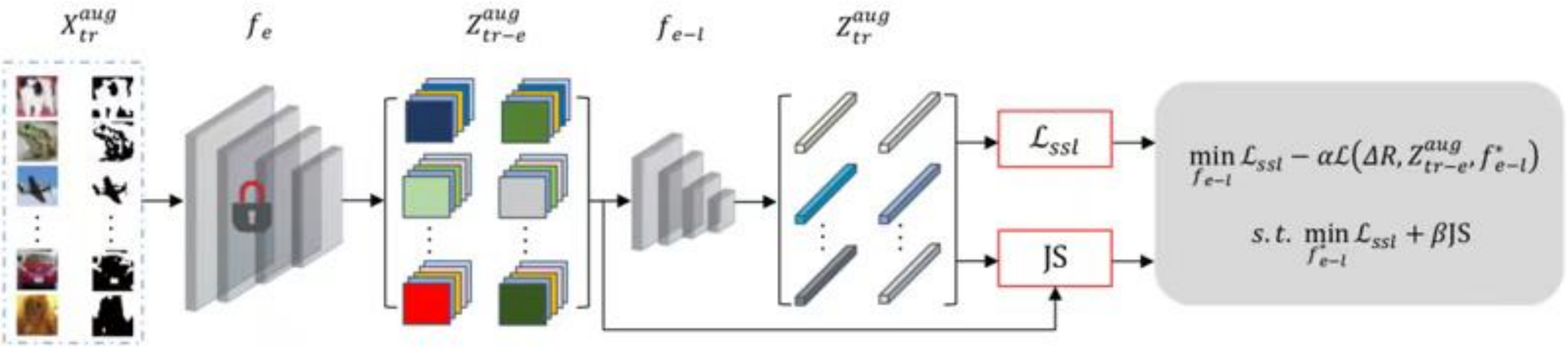
除此之外，研究团队通过实验发现，编码降低率（CRR）与验证集准确率的变化趋势几乎一致。由于通过计算验证集准确率来评估是否发生过拟合现象的方式计算量大、耗时长，而相比之下，CRR的计算量小、耗时短，且两者的变化趋势几乎一致，因此，研究团队提出可以使用CRR来量化过拟合现象。



中间层特征和最后层特征在验证集上的编码降低率变化趋势

方法

基于上述发现，研究团队提出了撤销记忆机制（UMM）算法。该算法采用双层优化机制。第一个优化目标旨在利用JS散度，使最后一层的特征分布尽量接近中间层的特征分布。在此基础上，第二个优化目标通过利用中间层特征的有用部分，最大化最后一层特征的CRR。通过这种方式，神经网络的后期层可以从中间层具有泛化性的特征中提取有效信息，并且从这些信息中找到能够最大化CRR并最小化自监督损失的部分，从而令最后一层特征恢复泛化能力。



记忆撤销机制算法框架图

研究团队进一步使用结构因果模型（SCM）进行分析。从因果因子生成的角度，样本数据可以看作是由任务相关因子和任务无关因子共同经过一个生成函数获得的。自监督学习通过数据增强的方式，改变任务无关因子，生成了不同的数据增强样本。研究团队通过严谨的理论分析证明，通过对齐不同增强样本特征的同时最大化每个样本的信息熵，可以令模型获取任务相关因子的全部信息。结合该结论与实验现象进行分析，随着自监督学习轮次的增加，最后一层的特征由于记忆效应会包含任务无关因子的信息量。由于神经网络的输出维度是固定的，这会导致最后一层特征包含的任务相关因子信息量减少，从而缺乏泛化性。而UMM的引入可以重新增加最后一层特征中任务相关因子的信息量，从而提升模型的泛化性。

实验

研究团队进一步在分类、检测、实例分割等各种下游任务上进行了广泛的实验，结果表明，引入UMM后在多个基线上实现了稳定的性能提升。