软件所提出小批量数据采样策略

作者：writer 来源：中国科学院

本文原地址：https://www.iikx.com/news/progress/33503.html

软件所提出小批量数据采样策略
。近日，中国科学院软件研究所科研团队提出了小批量数据采样策略，可消除由不可观测变量语义引起的虚假关联对表征学习的干扰，来提升自监督学习模型分布外泛化能力。

自监督学习的分布外泛化能力是指模型在面对与训练数据分布不同的测试数据时，仍能够保持良好性能。简单来说，模型需要在“未见过”的数据分布上表现得和在训练数据上一样好。但是，有研究发现，自监督学习模型在训练过程中受到与学习任务无关的不可观测变量的语义干扰，从而削弱分布外泛化能力。

该研究基于因果效应估计等手段，提出小批量数据采样策略，来消除不可观测变量语义干扰的混杂影响。这一策略通过学习隐变量模型，来估计在给定“锚点”样本的条件下，不可观测语义变量的后验概率分布，将其记为平衡分数。进而，该策略将具有相同或相近平衡分数的样本对划分为同一个小批量数据集，确保每个小批量数据集内的不可观测语义变量与“锚点”样本在条件上是独立的，从而帮助模型避免学习到虚假关联，提升模型的分布外泛化能力。

进一步，该研究在基准数据集上进行了广泛实验。所有实验均仅替换批次生成机制，无需调整模型架构或超参数。实验显示，这一采样策略使当前主流自监督学习方法在各类评估任务上至少提高2%的表现。具体而言，在ImageNet 100和ImageNet的分类任务中，Top 1和Top 5准确率均超越自监督方法SOTA；在半监督场景下的分类任务中，Top 1和Top 5准确率分别提升超3%和2%；目标检测与实例分割迁移学习任务中，各项平均精度均获得稳定增益；对于Omniglot、miniImageNet和CIFAR FS等少样本转移学习任务，性能提升超5%。实验表明，这一采样策略可以弱化虚假关联、强化因果学习，并能够提升分布外泛化能力。

相关研究成果被CCF-A类人工智能顶级学术会议International Conference on Machine Learning（ICML-25）接收。

论文链接

研究团队单位：软件研究所