
第一章 引言
❖ “预训练-微调”范式在视觉-语言领域泛用
❖ 联合预训练方法消除了对模态对齐的依赖
数据可扩展性和迁移效果
❖ 以视觉-语言多模态CLIP方法为研究对象
核心研究挑战
❖ 进一步提升CLIP方法可用高质量训练数据
❖ 改善CLIP模型在下游视觉任务的迁移表现
❖ 实现以文本为输出的多模态下游任务迁移
扩宽数据来源
泛化迁移任务
第五章 总结与展望
总结
❖ 提出利用图像分类数据增强CLIP预训练
❖ 提出细粒度自蒸馏改善视觉任务迁移性能
❖ 提出离散扩散方法实现图像注释任务迁移
展望
❖ 从图文对驱动到网页驱动以拓宽数据来源
❖ 大语言模型驱动的CLIP训练改善语言任务
❖ 拓展时空信息与语言信息结合的多模态任务
第二章
基于低噪声图像分类数据增强的
CLIP训练方法
❖ 用视觉-语言对比学习实现图像分类
❖ 深度融合分类数据和图文对数据
❖ 提升了物体识别和图文检索效果
视觉任务迁移
第三章
基于特征自蒸馏方法
增强CLIP模型视觉任务迁移
❖ 用特征自监督构造细粒度任务
❖ 改善语义分割、物体检测等性能
文本生成多模态任务迁移
第四章
基于离散扩散方法的
CLIP模型图像注释任务迁移
❖ 针对文本信号设计离散扩散方法
❖ 性能良好且适合人机交互场景