

RS5M_and_GeoRSCLIP_A_Large-Scale_Vision-Language_Dataset_and_a_Large_Vision-Language_Model_for_Remote_Sensing

全文摘要

全文概述

本文提出了一种大规模遥感图像-文本配对数据集 RS5M 及基于该数据集优化的视觉-语言模型 GeoRSCLIP。RS5M 包含 500 万对遥感图像与英文描述，通过过滤公开图像-文本数据集和利用预训练模型生成描述构建而成，其规模较现有最大 RS 数据集提升近 1000 倍。GeoRSCLIP 通过全量微调或参数高效微调方法对 CLIP 模型进行优化，在零样本分类、跨模态检索和语义定位任务中分别取得 3%-20%、3%-6% 和 4%-5% 的性能提升。研究团队通过旋转不变性准则筛选高质量描述，并引入地理元数据增强，解决了传统 RS 数据集样本量不足和描述质量低的问题。实验表明，RS5M 在保持数据规模的同时有效过滤噪声，其训练的 GeoRSCLIP 模型在 AID、RESISC45 和 EuroSAT 数据集上达到 SOTA 性能，且在跨域泛化能力测试中表现优异。该研究为遥感领域视觉-语言模型的迁移学习提供了基础数据集和优化范式。

术语解释

1. **RS5M**: 首个大规模遥感图像-文本配对数据集，包含 500 万对遥感图像与描述，通过过滤公开数据集和生成描述构建，支持视觉-语言模型的领域迁移。

2. **GeoRSCLIP**: 基于 CLIP 模型优化的遥感领域视觉-语言模型，通过全量微调或参数高效微调方法，在零样本分类、跨模态检索和语义定位任务中实现性能提升。
3. **参数高效微调**: 通过冻结预训练模型权重仅训练少量适配器参数的优化方法，如 LoRA、Pfeiffer Adapter 等，降低计算成本同时保持模型性能。

论文速读

论文方法

方法描述

该论文主要介绍了使用遥感图像和文本数据集来训练地理空间语义理解模型（GeoRSCLIP）的方法。首先，通过收集和整理多个遥感图像和文本数据集，构建了一个包含数百万张图片和数千万个标签的数据集（RS5M）。然后，使用了 CLIP 模型作为基础视觉编码器，并在该模型的基础上进行了参数调整以适应遥感图像领域的需求。最后，通过对数据集进行分析和可视化，发现了一些潜在的地域差异和负面社会影响问题，并提出了相应的解决方案。

方法改进

与传统的遥感图像处理方法相比，该方法引入了自然语言处理技术，可以更准确地识别和理解遥感图像中的地理位置信息和其他相关信息。此外，该方法还采用了深度学习技术和大规模数据集，能够更好地捕捉和表达遥感图像和文本之间的关系，从而提高地理空间语义理解的精度和效率。

解决的问题

该方法主要解决了遥感图像理解和应用中的一些挑战，包括如何准确地识别和理解遥感图像中的地理位置信息、如何利用遥感图像和文本数据集来训练地理空间语义理解模型等。同时，该方法也揭示了一些潜在的地域差异和负面社会影响问题，为遥感图像的应用和发展提供了参考和指导。

论文实验

本文主要介绍了 RS5M 数据集在图像和文本关系任务中的表现，并进行了多个对比实验以验证其有效性。实验包括使用不同的预训练模型（如 CLIP ViT-B32、CLIP ViT-B16、CLIP ViT-L16、CLIP ViT-H14 等）对数据集进行 Fine-Tuning，以及与其他相关方法的比较。实验结果表明，RS5M 数据集在 ZSC、VLR、SeLo 等任务中均表现出色，具有良好的泛化能力和实用性。此外，本文还分析了不同因素对模型性能的影响，例如数据集规模、图像归一化、噪声水平等，为后续研究提供了参考。

关键图表解读

图 1: RS5M 数据集示例展示，包含地理元数据的图像-文本对。该图通过可视化呈现数据集中包含经纬度、拍摄时间等地理信息的样本，验证了数据集在地理定位任务中的应用潜力。

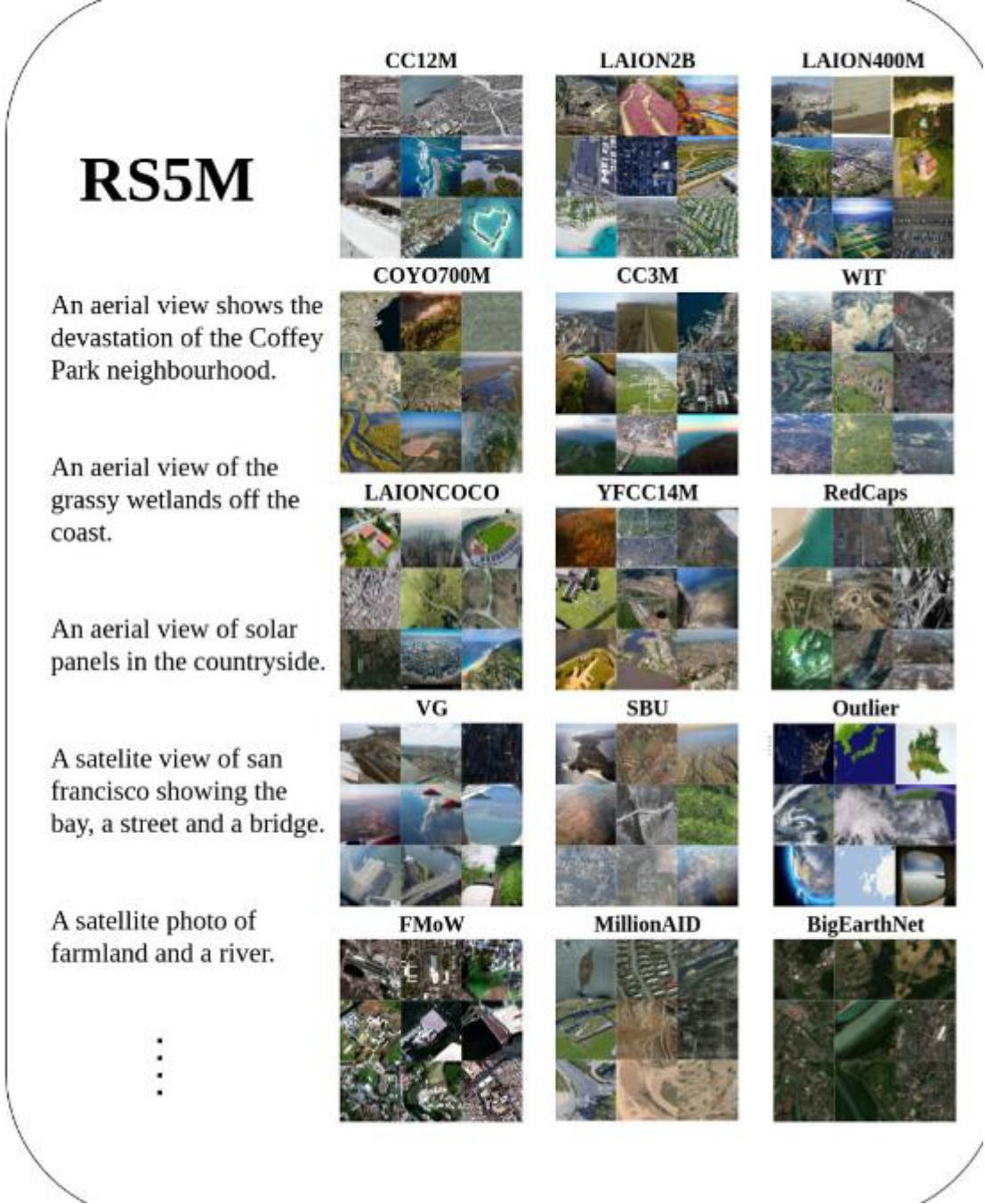


图 2: RS5M 数据集构建流程图。该流程图系统展示了从 11 个公开数据集过滤和 3 个大规模 RS 数据集生成描述的双路径构建方法，强调了预训练 VLM 和 RS

图像检测器在去噪中的关键作用。

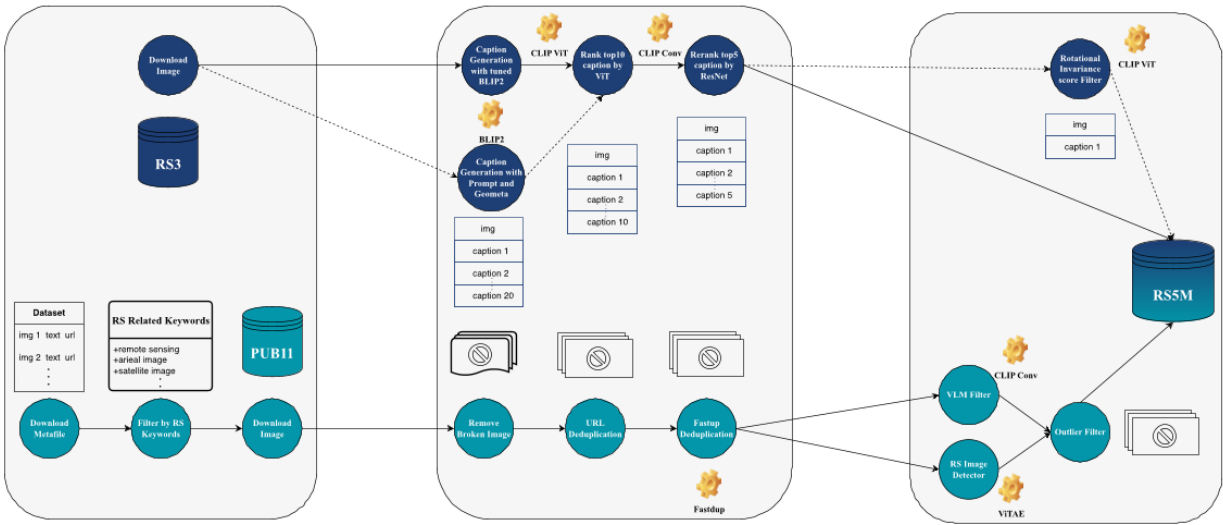


图 3：数据集质量分析。左侧显示关键词频率统计，"aerial view"占比最高；中间为词云图，高频词包括"satellite"、"building"等；右侧为描述长度分布，平均 49 词。该图揭示了数据集的地理空间特征和描述复杂度。

