6 月 30 日~7 月 6 日（第二周）

1. Boosting Vision-Language Models with Transduction

1.1 主要工作&贡献

1) 引入了一种传导公式，通过利用未标记数据的结构来增强可变学习模型(VLM)的零样本和少样本泛化能力。构建了一个可以看作是一个正则化的最大似然估计的新目标函数
2) 使用 Kullback-Leibler(KL)散度惩罚约束，该惩罚整合了文本编码器的知识，并引导传导学习过程。进一步推导了一个迭代块主要化-最小化(BMM)程序来优化我们的目标
3) 该方法可以作为当前归纳零样本模型和少样本学习方法的即插即用模块使用，并持续提升其性能
4) 得益于基于知识图谱的语言监督这一关键因素，该方法显著优于文献中近期的传导式小样本方法

1.2 多元高斯分布混合模型

在目标函数中定义了一个高斯混合模型（GMM）聚类项，通过将目标数据的似然建模为一个平衡的多元高斯分布混合模型来实现.

$$P_{i,k} = \Pr(f_i, k; \mu_k, \sum) \propto \det(\sum)^{1/2} \exp\left(-\frac{1}{2}(f_i - \mu_k)^T \sum^{-1} (f_i - \mu_k)\right) \quad (1\text{-}1)$$

其中$P_{i,k}$代表样本i属于k类的概率

1.3 新的目标函数

Q表示未标记的查询样本索引集（即需要预测的样本）

S表示少样本学习中的带标签支持样本索引集

z_i是模型预测结果

$$L = -\frac{1}{|Q|}\sum_{i\epsilon Q} z_i^T \, log(P_i) - \sum_{i\epsilon D}\sum_{j\epsilon D} W_{ij} z_i z_j + \sum_{i\in Q} KL_\lambda(z_i || \hat{y}_i) \quad (1\text{-}2)$$

第一项:

$$-\frac{1}{|Q|}\sum_{i\epsilon Q} z_i^T \, log(P_i) \quad (GMM)$$

其中z_i是样本i的预测结果，P_i代表 GMM 概率预测值，计算其交叉熵

第二项

$$\sum_{i\epsilon D}\sum_{j\epsilon D} W_{ij} z_i z_j \quad (laplacian)$$

拉普拉斯，相似的样本应该放在一起

W_{ij}是计算样本之间的相似度

第三项

$$\sum_{i\in Q} KL_{\lambda}(z_i||\hat{y}_i) \qquad (KL)$$

分解后如式 1-3 所示：

$$KL(z_i||\hat{y}_i) = z_i^T \log z_i - \lambda z_i^T \log \hat{y}_i \qquad (1\text{-}3)$$

其中前一项是熵项和后一项是交叉熵项，分别防止预测过度自信以及对齐文本先验知识

少样本学习场景的扩展（半监督学习）

零样本学习方法可自然扩展到少样本学习场景。通过引入带标签支持样本（labeled-support samples）的监督信号（以交叉熵形式），最小化以下整体损失函数：

$$-\frac{1}{|S|}\sum_{i\epsilon S} z_i^T \log(P_i) \qquad (1\text{-}4)$$

式子 1-4 是监督项，确定要逼近正确的样本，其中z_i是固定的，带标签的数据集。

1.4 块优化-最小化（BMM）算法（关键部分）

由于目标函数依赖于三类变量$(z, \mu, \sum)$，我们采用块优化-最小化（BMM）流程，通过交替优化三个子步骤实现：

1. 每个子步骤固定其中两类变量，优化剩余的一类变量
2. 确保整体目标函数值单调不增
3. 关键优势：

- z的更新（公式 5）是解耦的，支持并行计算，适合大规模数据集（如 ImageNet）
- 理论保证：整体算法必收敛（定理 1）

Tips: 对于拉普拉斯函数，凹优化，找线性上界，其余参数逐次更新，参考表 2-1.

1.5 收敛性分析

作者的优化器可视为块优化-最小化(Block Majorize-Minimize, BMM)范式的一个实例，该范式通过为每个变量块优化其主函数（majorizing function）来实现目标最小化

零样本场景：TransCLIP 将 Top-1 准确率平均提升超过 5%

少样本场景：在 1-shot 设置下，TransCLIP 平均提升主流方法 4%的准确率

表明了TransCLIP可兼容提示调优(prompt tuning)和适配器微调(adapter fine-tuning)方案，在域内任务和领域泛化任务中均能提升性能。

部分情境显示转导收益可能随 shot 数量增加而降低，推测原因是：当标注样本较多时，监督学习已能部分捕获数据结构信息，减弱了转导优化的必要性。

2. Enhancing Remote Sensing Vision-Language Models for Zero-Shot Scene Classification

2.1 主要研究背景&贡献

遥感图像通常尺寸大、细节丰富，传统方法需将图像分割为小块（patch）独立处理（归纳推理），但会忽略块间的上下文关联。零样本分类（无需训练数据）依赖文本提示生成伪标签，但独立预测可能导致局部误差累积。

传统 VLMs（如 CLIP）在遥感领域直接应用时，因自然图像与遥感图像的领域差异，性能受限。转导推理（Transductive Inference）在计算机视觉中已被证明能利用未标注数据的分布提升预测，但在遥感 VLMs 中尚未充分探索。

2.2 算法（实际就是使用了上文的传导方法）

本文提出了一种高效的转导方法，仅在特征空间内操作，无需额外监督。

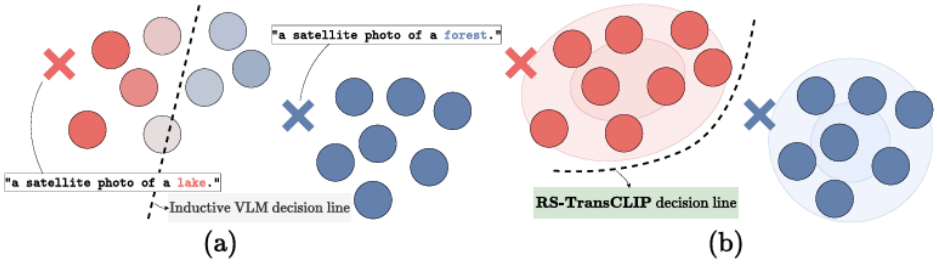

图 2-1 (a) 传统 VLM 将每张图像分配给最接近的文本嵌入；(b) RS-TransCLIP 利用图像-文本结构增强预测，无需额外标注。

(a)传统 VLM 的工作方式

通过计算图像嵌入（image embedding）与所有候选文本嵌入（text embedding）的余弦相似度，独立地将每张图像分配给相似度最高的文本描述。

局限在于孤立预测，对每个图像块（patch）单独分类，忽略块间的空间关联。文本依赖性强：性能高度依赖预设的文本提示模板质量。

(b)RS-TransCLIP 的改进

利用图像块间的视觉相似性（如颜色、纹理）构建相似关系（拉普拉斯函数），强制相似区域获得一致标签。将初始文本伪标签（CLIP 生成）与图像空间结构联合优化，通过转导推理调整预测。整个过程无需人工标注，仅依赖模型自身。

表 2-1 RS-TransCLIP 算法

算法 1	RS-TransCLIP 流程
1.	$\boldsymbol{Input: f, t, \tau}$
2.	$\boldsymbol{\hat{y}_i \leftarrow softmax(\tau f_i^T t)} \quad \forall \boldsymbol{i}$;
3.	$\boldsymbol{w_{ij} = f_i^T f_j} \quad \forall \boldsymbol{i, j}$;
4.	$\boldsymbol{z_i \leftarrow \hat{y}_i} \quad \forall \boldsymbol{i}$;
5.	$\boldsymbol{Initialize\ \mu_k} \quad \forall \boldsymbol{k, and} \quad \boldsymbol{diag}(\sum)$;
6.	$\boldsymbol{while\ not\ converged\ do}$
7.	*//迭代解耦更新*
8.	$\boldsymbol{for\ l = 1: \ldots do}$
9.	$\boldsymbol{Update\ z_i^{l+1}} \quad \forall \boldsymbol{i}$;
10.	$\boldsymbol{end}$
11.	*//闭式更新*
12.	$\boldsymbol{Update\ \ \mu_k} \quad \forall \boldsymbol{k}$;
13.	$\boldsymbol{Update\ \ diag}(\sum)$;
14.	$\boldsymbol{end}$
15.	$\boldsymbol{return\ z}$

Tips: 各参数更新公式（对应算法中的 12、13 两行）

$$\mu_k = \frac{\sum_{i \in Q} z_{i,k} f_i}{\sum_{i \in Q} z_{i,k}}$$

$$diag(\sum) = \frac{\sum_{i \in Q} z_{i,k} (f_i - \mu_k)^2}{|Q|}$$

其中f_i是第i个图像块的视觉特征嵌入

2.3 实验结果

首先是模型规模与性能的正相关趋势，标志着扩大模型与数据规模的就可以提升模型的性能的可能性。

其次评估基于转导目标函数的推理性能，可以观察到所有数据集和模型均获得显著提升，RS-TransCLIP 带来的平均准确率增益可达 9.9%-17.1%。