Transformer框架下的的遥感图象分割

基于ST-UNet

熊恪峥 徐昊 陈奕凝

2022年5月4日

目录

第一部分

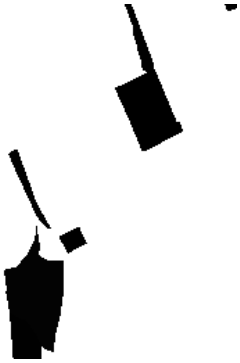
选题

第一章　题目概述

本队选题是“遥感地块影像分割”，赛题旨在对遥感影像进行像素级内容解析，并对遥感影像中感兴趣的类别进行提取和分类，以衡量遥感影像地块分割模型在多个类别（如建筑、道路、林地等）上的效果。数据集为多个地区已脱敏的遥感影像数据，包含$66,653$张分辨率为$2\,m\,/\,pixel$，尺寸为256×256的PNG图片

图 1.1: 数据集示例

(a) 训练集



(b) 标注

(c) 训练集

(d) 标注

第二部分

设计思路

第二章　设计概述

Transfomer是一种在自然语言处理领域中流行的模型。近年来，Transfomer的成功为涉及全局关系的深度学习领域的研究提供了新的方法。Visual Transfomer (ViT) [1]将Transfomer 引入计算机视觉领域，获得了良好效果。基于ViT的语义分割 [2]在ADE20K数据集上达到State-of-the-art，超越了其他同类模型。Swin-Transfomer [3]通过构建层次化的Transformer改进了ViT，并且引入Locality，在ADE20K数据集上达到了53.5的mIOU。因此在语义分割任务上，Swin-Transformer是一种极具前景的主干网络，有很好的效果预期。

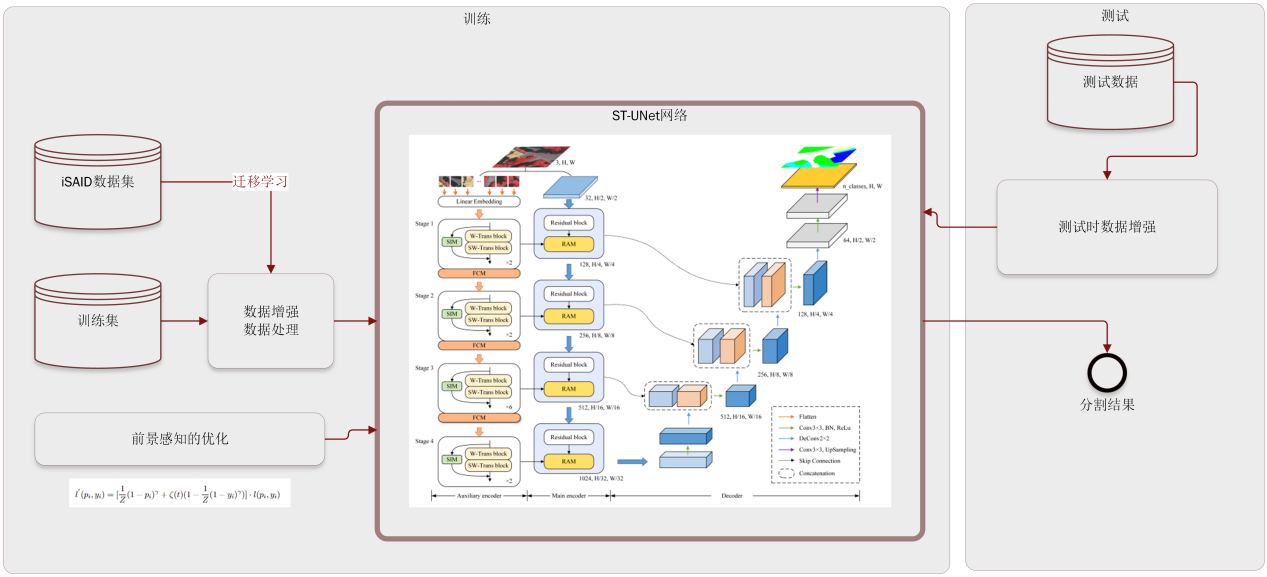


图 2.1: 整体设计

图 2.1是我们的整体设计示意图。我们计划使用通过混合CNN和Swin-Transformer结构在遥感影像语义分割上取得了良好效果的 嵌入了*Swin Transformer*的*UNet(ST-UNet)* [4]作为网络结构进行调整和训练。该网络结构融合了Swin-Transformer和有成熟广发应用的CNN结构，可以期望能达到较好的效果。

为了处理遥感图像分割中样本的不均衡问题，尤其是前景-背景的不均衡问题，我们计划试验使用FarSeg [5]中提出的前景感知的优化，即使用损失函数(2.1)

$$l^{'}(p_i, y_i) = [\frac{1}{Z}(1-p_i)^{\gamma} + \zeta(t)(1-\frac{1}{Z}(1-y_i)^{\gamma})] \cdot l(p_i, y_i) \tag{2.1}$$

通过以背景中的困难部分作为权重较高的部分，可以是网络集中在前景和背景中的困难样本中，从而实现均衡优化。

Transformer主干网络的模型通常相对于CNN为主干网络的模型难以训练。为了解决这个问题，我们将应用迁移学习的方法。考虑到题目给定的数据集尺寸相对较小，我们首先在更大更完全的遥感影像数据集中预训练该网络，然后再在题目给定的数据集中进行训练。这是提高准确性的有效方法。我们计划采用iSAID数据集 [6]，该数据集提供了2806张遥感影像，来自有多种传感器和多分辨率的平台，图

象大小从800×800到4000×13000不等。为了使得该数据集和题目给定的数据集尽可能接近，我们将会对iSAID数据集进行进一步处理，裁切成256×256的分块。

在数据处理方面，我们将对遥感影像进行数据增强，包括随机裁剪、亮度，对比度和饱和度的调整加入噪点与随机模糊等。这些影像变换可以模拟遥感图像采集中常见的图像缺陷。这些缺陷可能干扰识别，通过对训练集进行数据增强，可以降低这些负面因素对网络训练的影响。

在模型实际应用的过程中，我们将会应用测试时增强的方法，在测试时通过数据增强产生额外的推理结果在此基础上进行投票可以获得更好的性能表现。

第三章　整体设计

一、遥感图像分割任务与ST-UNet

遥感影响的物体分割是一种语义分割任务。这种任务面对大规模的变化、大规模的类内背景差异和较大的类外背景差异。以及前景-背景不平衡的问题。一般的语义分割常常更加关注自然场景中的尺度变化，而没有充分地考虑到其他的问题 [5]。并且常见的CNN作为主干网络的模型由于卷积运算的局部性，难以对网络的全局特征进行直接获取。

Swin Transformer在实践中展现出了极为强大的全局建模能力。而UNet是一种常用、表现优秀的的语义分割框架。因此将Swin Transformer嵌入传统的基于CNN的UNet 中，可以得到ST-UNet这一融合的遥感图像语义分割的框架 [4]，它具有Swin-Transformer 和CNN平行工作的双*Encoder*架构。一方面，ST-UNet使用空间交互模块(Spatial Interation Module, STM)通过Swin Transformer编码像素级的相关性来提高特征的代表能力，尤其是受到遮蔽的物体。另一方面，该模型通过一个特征压缩模块(Feature Compression Module, FPM)来减少详细信息的丢失，并在补丁标记下采样时浓缩更多的小规模的特征，这些小尺度的特征可以提高地面小尺度物体的分割精度。

最后，作为以上两个编码器的桥梁，该网路通过一个关系聚合模块(Relation Aggregation Module, RAM)来聚合两个编码器的特征，将Swin-Transformer获得的全局相关关系层级化地集成到CNN中。这种方式对在真实世界数据集中的识别起到了极为显著的提高 [4]。

在该方案中，我们采用该网络的原因主要有如下两点

- Transformer框架在计算机视觉领域有良好的前景
- ST-UNet表现出了较好的性能

二、前景感知的优化

前景感知的优化是 [5]中提出了重要优化之一。前景与背景不均衡的问题常常导致在训练过程中背景主导了梯度，但是只有北京的困难部分队训练后期的优化有价值，而这些样本相对稀少。这是该优化提出的动力。它的核心是将损失函数换成 (3.1)，借此将网络集中在前景和背景的困难样本上。

$$l^{'}(p_i, y_i) = [\frac{1}{Z}(1-p_i)^{\gamma} + \zeta(t)(1-\frac{1}{Z}(1-y_i)^{\gamma})] \cdot l(p_i, y_i) \tag{3.1}$$

其中p_i是预测的概率，y_i代表第i像素的Ground truth。Z是一个归一化常数，该常数保证$\sum l(p_i, y_i) = \frac{1}{Z}\sum(1-p_i)^{\gamma} l(p_i, y_i)$。 $l(p_i, y_i)$是一个交叉熵损失函数。$\zeta(t)$是一个单调递减的退火函数，其取值范围在$[0, 1]$之间。有线性、多项式、余弦三种选择，如图 3.1，每种选择有各自的超参数可供控制和调整。

虽然该优化和主干网络ST-UNet并不来源于同一个工作，但是该优化对遥感图像分割任务中有显著影响的不均衡问题提出了解决方案，该解决方案与主干网络独立，具有一定的普适性。因此将该优化加入ST-UNet中以测试其性能并作为一种可能的优化候选是合理的，一定程度上也是必要的。

Annealing function	Formula	Hyperparameter
Linear	$\zeta(t) = 1 - \frac{t}{annealing\_step}$	$annealing\_step$
Poly	$\zeta(t) = (1 - \frac{t}{annealing\_step})^{decay\_factor}$	$annealing\_step, decay\_factor$
Cosine	$\zeta(t) = 0.5 * (1 + \cos(\frac{t}{annealing\_step}\pi))$	$annealing\_step$

图 3.1: 退火函数

三、迁移学习

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Etiam lobortis facilisis sem. Nullam nec mi et neque pharetra sollicitudin. Praesent imperdiet mi nec ante. Donec ullamcorper, felis non sodales commodo, lectus velit ultrices augue, a dignissim nibh lectus placerat pede. Vivamus nunc nunc, molestie ut, ultricies vel, semper in, velit. Ut porttitor. Praesent in sapien. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Duis fringilla tristique neque. Sed interdum libero ut metus. Pellentesque placerat. Nam rutrum augue a leo. Morbi sed elit sit amet ante lobortis sollicitudin. Praesent blandit blandit mauris. Praesent lectus tellus, aliquet aliquam, luctus a, egestas a, turpis. Mauris lacinia lorem sit amet ipsum. Nunc quis urna dictum turpis accumsan semper.

四、数据增强

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Etiam lobortis facilisis sem. Nullam nec mi et neque pharetra sollicitudin. Praesent imperdiet mi nec ante. Donec ullamcorper, felis non sodales commodo, lectus velit ultrices augue, a dignissim nibh lectus placerat pede. Vivamus nunc nunc, molestie ut, ultricies vel, semper in, velit. Ut porttitor. Praesent in sapien. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Duis fringilla tristique neque. Sed interdum libero ut metus. Pellentesque placerat. Nam rutrum augue a leo. Morbi sed elit sit amet ante lobortis sollicitudin. Praesent blandit blandit mauris. Praesent lectus tellus, aliquet aliquam, luctus a, egestas a, turpis. Mauris lacinia lorem sit amet ipsum. Nunc quis urna dictum turpis accumsan semper.

第三部分

实现方法

第四章　模型实现

一、主干网络和前景感知的优化

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Etiam lobortis facilisis sem. Nullam nec mi et neque pharetra sollicitudin. Praesent imperdiet mi nec ante. Donec ullamcorper, felis non sodales commodo, lectus velit ultrices augue, a dignissim nibh lectus placerat pede. Vivamus nunc nunc, molestie ut, ultricies vel, semper in, velit. Ut porttitor. Praesent in sapien. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Duis fringilla tristique neque. Sed interdum libero ut metus. Pellentesque placerat. Nam rutrum augue a leo. Morbi sed elit sit amet ante lobortis sollicitudin. Praesent blandit blandit mauris. Praesent lectus tellus, aliquet aliquam, luctus a, egestas a, turpis. Mauris lacinia lorem sit amet ipsum. Nunc quis urna dictum turpis accumsan semper.

二、迁移学习和数据增强

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Etiam lobortis facilisis sem. Nullam nec mi et neque pharetra sollicitudin. Praesent imperdiet mi nec ante. Donec ullamcorper, felis non sodales commodo, lectus velit ultrices augue, a dignissim nibh lectus placerat pede. Vivamus nunc nunc, molestie ut, ultricies vel, semper in, velit. Ut porttitor. Praesent in sapien. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Duis fringilla tristique neque. Sed interdum libero ut metus. Pellentesque placerat. Nam rutrum augue a leo. Morbi sed elit sit amet ante lobortis sollicitudin. Praesent blandit blandit mauris. Praesent lectus tellus, aliquet aliquam, luctus a, egestas a, turpis. Mauris lacinia lorem sit amet ipsum. Nunc quis urna dictum turpis accumsan semper.

第五章　训练与测试

一、训练方法

尽管Transformer在图象上的应用具有较强的竞争力，但是与成熟的卷积神经网络相比，训练技巧还并不成熟 [7]，并且由于参数量的区别，Transformer训练通常较难。因此，本方案对ST-UNet的训练提出以下的预案

迁移学习

本方案反复强调了迁移学习的重要性。这是因为 [7]中提到，就大多数实际目的而言，迁移预先训练的模型不仅成本效益较高，而且会带来更好的结果。对于类似题目所给的这样数据量相对同邻域常用数据集较小的数据集而言，几乎不可能通过从零开始训练使其达到接近迁移模型的精度。而对于足够大的数据集，从零开始达到与迁移模型相似的精度则需要多花超过2个数量级的时间。

二、消融试验

消融试验是一种常用的方法，为了检验采取多项改进时每一项改进都对效果的提高具有正向贡献，需要对各项改进单独出现、成组出现的情况进行测试。由于本方案为了有效地进行遥感图像语义分割进行了多项措施，为了检验各项措施，我们将进行消融试验，并依照试验结果进行评估，对措施进行重新调整，试验计划如表 5.1。

表 5.1: 消融试验计划

前景感知的优化	额外数据集的迁移学习	数据增强	mIOU
✓			待测
	✓		待测
		✓	待测
✓	✓		待测
✓		✓	待测
	✓	✓	待测
✓	✓	✓	待测

我们期望在消融试验中得到性能参数逐次增高的效果。如果某项措施的效果不能达到预期的效果，我们会考虑

- 对该项措施进行重新调整；
- 移除该项措施

然后重新进行测试并评估效果。

参考文献

参考文献

[1] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2020.

[2] Robin Strudel, Ricardo Garcia, Ivan Laptev, and Cordelia Schmid. Segmenter: Transformer for semantic segmentation, 2021.

[3] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows, 2021.

[4] Xin He, Yong Zhou, Jiaqi Zhao, Di Zhang, Rui Yao, and Yong Xue. Swin transformer embedding unet for remote sensing image semantic segmentation. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–15, 2022.

[5] Zhuo Zheng, Yanfei Zhong, Junjue Wang, and Ailong Ma. Foreground-aware relation network for geospatial object segmentation in high spatial resolution remote sensing imagery, 2020.

[6] Syed Waqas Zamir, Aditya Arora, Akshita Gupta, Salman Khan, Guolei Sun, Fahad Shahbaz Khan, Fan Zhu, Ling Shao, Gui-Song Xia, and Xiang Bai. isaid: A large-scale dataset for instance segmentation in aerial images, 2019.

[7] Syed Waqas Zamir, Aditya Arora, Akshita Gupta, Salman Khan, Guolei Sun, Fahad Shahbaz Khan, Fan Zhu, Ling Shao, Gui-Song Xia, and Xiang Bai. How to train your vit? data, augmentation, and regularization in vision transformers, 2021.

附录

Section within the appendix

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Etiam lobortis facilisis sem. Nullam nec mi et neque pharetra sollicitudin. Praesent imperdiet mi nec ante. Donec ullamcorper, felis non sodales commodo, lectus velit ultrices augue, a dignissim nibh lectus placerat pede. Vivamus nunc nunc, molestie ut, ultricies vel, semper in, velit. Ut porttitor. Praesent in sapien. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Duis fringilla tristique neque. Sed interdum libero ut metus. Pellentesque placerat. Nam rutrum augue a leo. Morbi sed elit sit amet ante lobortis sollicitudin. Praesent blandit blandit mauris. Praesent lectus tellus, aliquet aliquam, luctus a, egestas a, turpis. Mauris lacinia lorem sit amet ipsum. Nunc quis urna dictum turpis accumsan semper.