Language-Driven Semantic Segmentation

Abstract

动机：①图像分类和图像分割有很强的关联性，所以在分类任务上有创新性的CLIP出来后，立马被拿来用在了分割任务中；②当下的像素级分割任务的分割object class是固定的；

Introduction

Semantic Segmentation，语义分割，由semantic class labels将图片分成coherent region

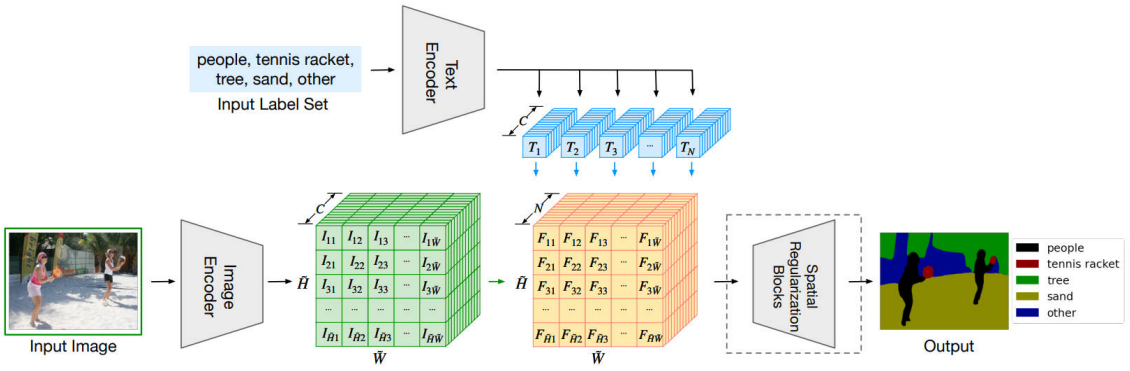
问题，在当下某个数据集上，这些semantic class lables是固定死的，

Related Work

Language-Driven Recognition中常见任务，①Visual Question Answering；②Image Captioning；③Image-Test Retrieval

Language-Driven Semantic Segmentation

LSeg，为本论文所提出模型的名字



1. Text Encoder，与CLIP中的Text Encoder一样

作用：**Embeds N个 Input Labels 成N个 Vector $\in R^C$，为$T_1, T_2, \ldots, T_N$，记作$T_k, k = 1, 2, \ldots, N$，在图中是蓝色的Tensor**

2. Image Encoder，为ViT+encoderz

原理：**记downsampling factor为s，$\tilde{H} = \frac{H}{s}, \tilde{W} = \frac{W}{s}$，Embeds $\tilde{H} \times \tilde{W}$个Input Image Pixels 成$\tilde{H} \times \tilde{W}$个Vector$\in R^C$，记作$I_{ij}, i = 1, 2, .., \tilde{H}; j = 1, 2, \ldots, \tilde{W}$，在图中是绿色的Tensor**

3. Word-Pixel Correlation Tensor

原理：**做inner product，举例F_{11}是如何得到的，I_{11}分别和$T_k, k = 1, 2, \ldots, N$做点积，得到$F_{11} \in R^N$，在图中是橙色的Tensor**

4. Spatial Regularization

原理：**恢复为Input Image相同的resolution**

5. Training Details，是有监督的训练，即有ground truth mask的，目标函数就是和这些ground truth mask去做Cross Entropy Losss

Experiments

Experimental Setup

PASCAL-5^i And COCO-20^i

FSS-1000

Exploration And Discussion

Ablation Studies

Qualitative Findings

Figure 6: Failure cases.

Limitations，如(a)图中，图中没有“玩具”，却将小狗当成了“玩具”；如(b)图中，房子有“窗户”，而且输入的semantic class lables中含有“窗户”，但是结果没将“窗户”划分开来

Conclusion