CLIMS: Cross Language Image Matching for Weakly Supervised Semantic Segmentation

邓松鹤

摘要

总所周知，CAM（Class Activation Map，类别激活图）通常只激活可识别的物体区域，而错误地包括许多与物体相关的背景。由于 WSSS（Weakly Supervisied Semantic Segmentation，弱监督语义分割）模型只有一套固定的图像级别的物体标签，要抑制那些由开放性物体组成的不同背景区域可能非常困难。本文提出了一个新颖的跨语言图像匹配（CLIMS）框架，该框架基于最近提出的语言图像对比预训练模型 CLIP，用于 WSSS。该框架的核心思想是引入自然语言监督，以激活更完整的物体区域，并抑制密切相关的开放背景区域。具体来说，通过设计对象、背景区域和文本标签的匹配损失，以指导模型为每个类别的 CAM 激活更合理的对象区域。此外，还设计了一个背景抑制损失，以防止模型激活密切相关的背景区域，并预定了一组类别相关的背景文本描述。这些设计使提出的 CLIMS 能够为目标物体生成一个更完整和紧凑的激活图。

关键词：跨语言图像；弱监督；语义分割；对比学习

1 引言

语义分割任务要求给图像中的每个像素分配一个语义标签。尽管全监督的语义分割在近年来取得了显著的成功，但是它要求大量的数据标注。相反，弱监督语义分割（WSSS）任务试图仅依赖图像级、检测框级、点级或基于涂鸦的监督来缓解这个问题。本研究的目的是在语义分割的学习中只使用图像级标签。

现有的 WSSS 方法通常遵循三个步骤：1). 图像级标签被用作特征级的监督，训练分类网络生成初始类激活图（CAM[1]图）；2). 使用基于像素亲和力的方法将初始 CAM 图改进为伪标注掩码；3). 利用改进后的伪标注掩码进一步训练分割网络。然而，由于 WSSS 模型只有一组固定的图像级对象标签，因此很难抑制由开放集合对象组成的各种背景区域。

CLIMS 通过语言图像预训练模型 CLIP[2]引入自然语言监督来激活更完整的对象区域，抑制密切相关的开放背景区域。CLIP 模型是在 4 亿个图像-文本对上训练的，这使得 CLIP 能够在一个开放的世界环境中将图像中更广泛的视觉概念与它们的文本标签关联起来，在此基础上，CLIMS 有很大的潜力为每个物体类别生成高质量的初始激活图。

2 相关工作

2.1 弱监督语义分割

常规的基于 CAM[1]的流程在之前的 WSSS 工作中被广泛使用。SeeNet[3]提出了两种自更新策略，只将注意力集中在可靠的区域，产生完整的初始 CAM 图；Chang 等人[4]的工作提出通过研究对象的子类别以挖掘更多的对象部分，然后一高初始 CAM 的性能。Sun 等人[5]在分类器中加入两个神经共

同注意力，用于探索一对训练图像的共享以及不共享的语义信息。这有助于从分类器中提取更完整的初始 CAM。Jungbeom 等人[6]提出了一种反对抗的方式来发现激活图中目标对象的更多区域。Ahn 和 Kwak[7]设计了一个深度神经网络，称为 AffinityNet，用于预测一对相邻图像坐标之间的语义亲和力。然后应用这种语义亲和性将生成的初始 CAM 细化为伪真实掩码。先前的工作使用全监督的突出性检测器来细化生成的初始 CAM 图。DeepLab[8][9]的一系列模型通常被用于通过伪掩码标签训练一个语义分割网络。

2.2 对比语言图像预训练（CLIP）

对比语言-图像预训练模型 CLIP[2]在零样本设置的许多视觉任务中显示出巨大的成广和潜力。CLIP 模型包含一个图像编码器和一个文本编码器。给定一批图像和文本对，CLIP 模型学习嵌入来度量图像和文本之间的相似度。CLIP 模型是在一个包含 4 亿图像-文本对的大型数据集上训练的，CLIP 可以识别的对象类别集比一个小数据集中的固定对象类别集更大，更具有多样性，如 PASCAL VOC2012。图像-文本对是自动从互联网收集的，不需要人工操作。

3 本文方法

3.1 本文方法概述

图1展示了本文提出的基于跨文本图像匹配（CLIMS）框架。它由一个骨干网络和一个文本驱动的评估器组成，其中评估器包括三个基于大型文本图像匹配预训练模型 CLIP 的损失函数。该方法的核心思想是通过文本驱动评估器的监督来学习初始 CAM 图的生成。

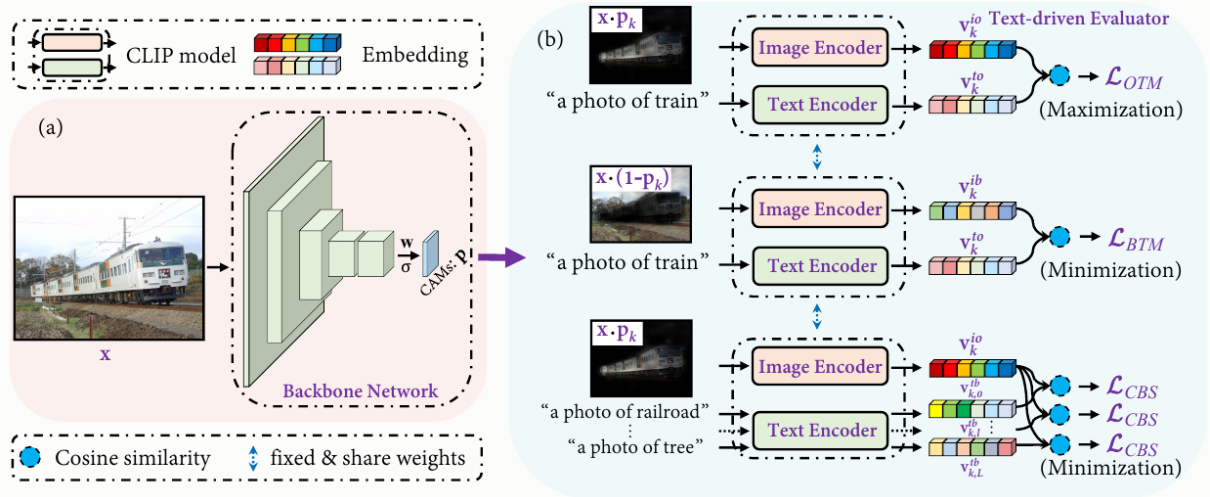

图 1: CLIMS 网络结构图

3.2 损失函数定义

本文方法的损失函数包括对象区域和文本标签匹配损失（L_{OTM}），背景区域和文本标签匹配损失（L_{BTM}），共现背景抑制损失（L_{CBS}）和区域正则化损失（L_{REG}）。

3.2.1 对象区域和文本标签匹配

给定第 k 个前景对象表示 $\mathbf{v}_k^{io}$ 和它对应的文本表示 $\mathbf{v}_k^{to}$，首先计算图像和文本表示之间的余弦相似度，然后使用提出的对象区域和文本标签匹配损失 L_{OTM} 使其最大化：

$$L_{OTM} = -\sum_{k=1}^{K} y_k \cdot \log(s_k^{oo}) \tag{1}$$

$$s_k^{oo} = \text{sim}(\mathbf{v}_k^{io}, \mathbf{v}_k^{to}) \tag{2}$$

s_k^{oo} 代表 $\mathbf{v}_k^{io}$ 和 $\mathbf{v}_k^{to}$ 的余弦相似度。生成的初始 CAM 将在 L_{OTM} 的监督下逐渐接近目标物体。然而，单独的 L_{OTM} 不能促进模型探索非判别目标区域，也不能抑制 $\mathbf{P}_k$ 激活的背景区域。

3.2.2 背景区域和文本标签匹配

为了提高激活对象区域的完整性，设计了背景区域和文本标签匹配损失 L_{BTM} 来包含更多的对象内容。给定背景表示 $\mathbf{v}_k^{ib}$ 和其对应文本表示 $\mathbf{v}_k^{to}$，可以计算 L_{BTM}：

$$L_{BTM} = -\sum_{k=1}^{K} y_k \cdot \log(1 - s_k^{bo}) \tag{3}$$

$$s_k^{bo} = \text{sim}(\mathbf{v}_k^{ib}, \mathbf{v}_k^{to}) \tag{4}$$

s_k^{bo} 代表 $\mathbf{v}_k^{ib}$ 和 $\mathbf{v}_k^{to}$ 的余弦相似度。当 L_{BTM} 最小化时，$\mathbf{X} \cdot (1 - \mathbf{P}_k)$ 内保留的目标对象像素更少，$\mathbf{X} \cdot \mathbf{P}_k$ 内恢复的目标对象内容更多。这确保在 $\mathbf{P}_k$ 中激活更完整的对象内容。

3.2.3 共现背景抑制

上述两个损失函数只保证了 $\mathbf{P}$ 完全覆盖目标对象，没有考虑到共现类相关背景的误激活。同时出现的背景可能会显著降低生成的伪地面真相掩模的质量。根据相应的文本描述，很容易使用预训练的 CLIP 来识别这些背景。给定目标对象表示 $\mathbf{v}_k^{io}$ 和其对应的类别相关的文本表示 $\mathbf{v}_k^{tb}$，可以计算 L_{CBS}：

$$L_{CBS} = -\sum_{k=1}^{K}\sum_{l=1}^{L} y_k \cdot \log(1 - s_{k,l}^{ob}) \tag{5}$$

$$s_{k,l}^{ob} = \text{sim}(\mathbf{v}_k^{io}, \mathbf{v}_{k,l}^{tb}) \tag{6}$$

其中 $s_{k,l}^{ob}$ 代表 $\mathbf{v}_k^{io}$ 和 $\mathbf{v}_{k,l}^{tb}$ 的余弦相似度。在训练过程中，骨干网络会逐渐抑制 $\mathbf{P}_k$ 中类相关背景区域的误激活，使 L_{CBS} 最小化。

3.2.4 区域正则化

设计一个像素级区域正则化项来约束激活图大小，以确保不相关的背景被排除在激活图 $\mathbf{P}_k$ 中：

$$L_{REG} = \frac{1}{K}\sum_{k=1}^{K} S_k, \quad \text{where} \quad S_k = \frac{1}{HW}\sum_{h=1}^{K}\sum_{w=1}^{W} \mathbf{P}_k(h, w) \tag{7}$$

3.2.5 训练过程损失

$$L = \alpha L_{OTM} + \beta L_{BTM} + \gamma L_{CBS} + \theta L_{REG} \tag{8}$$

4 复现细节

4.1 与已有开源代码对比

本复现工作基于作者发布于 GitHub 的开源代码：*https://github.com/CVI-SZU/CLIMS*

在已有开源代码的基础上，我们做出了如下改进：

- 相较于原文中人工设定背景文本的方法，我们设计了一个背景提取器，能够根据输入图像自动提取共现背景文本。
- 使用 InfoNCE[10] 损失替代原文的 L_{OTM}、L_{BTM} 和 L_{CBS} 三个损失。
- 在 PASCAL VOC2012 数据集上的实验表明我们的模型相比于原文能够生成更高质量的初始 CAM 图，同时也能得到具有竞争力的分割结果。

改进后的模型框架如图2 所示

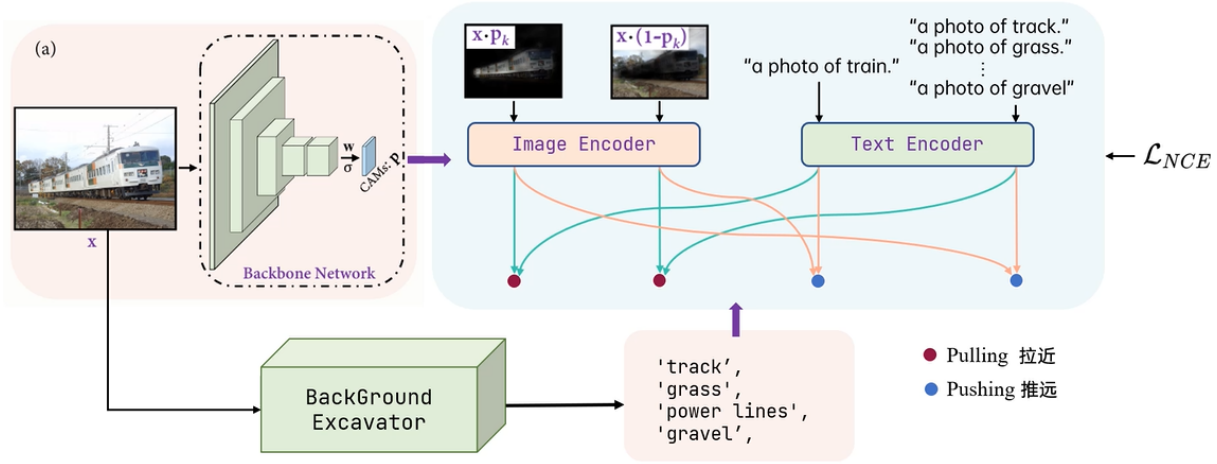

图 2: 改进后的 CLIMS 网络结构图

4.2 背景提取器

该背景提取器接收图像和文本同时作为输入，能够检测得到图像中与文本所描述的物体共同出现的其他相关物体，并返回相应的文本。同原文一样，将这一步得到的文本也通过 CLIP 的文本编码器进行编码，将生成的文本表示与图像计算相似度以监督学习初始 CAM 图的生成。

4.3 InfoNCE 损失

给定目标对象表示 $\mathbf{v}_k^{io}$ 和其对应文本表示 $\mathbf{v}_k^{to}$，以及背景图像表示 $\mathbf{v}_k^{ib}$ 和背景相关文本表示 $\mathbf{v}_{k,l}^{tb}$，可以计算 L_{NCE}：

$$L_{NCE} = -\log \frac{\exp(\sum_{k=1}^{K} s_k^{oo}/\tau)}{\sum_{k=1}^{K}\sum_{l=1}^{L}\exp(s_{k,l}^{ob}/\tau) + \sum_{k=1}^{K}\exp(s_k^{bo}/\tau)} \tag{9}$$

$$s_k^{oo} = \text{sim}(\mathbf{v}_k^{io}, \mathbf{v}_k^{to}) \tag{10}$$

$$s_{k,l}^{ob} = \text{sim}(\mathbf{v}_k^{io}, \mathbf{v}_{k,l}^{tb}) \tag{11}$$

$$s_k^{bo} = \text{sim}(\mathbf{v}_k^{ib}, \mathbf{v}_k^{to}) \tag{12}$$

其中 s_k^{oo} 即目标对象与文本标签的余弦相似度，作为正样本；而 $s_{k,l}^{ob}$ 和 s_k^{bo} 分别表示目标对象与背景文本的相似度、背景与文本标签的相似度，共同作为负样本。

得益于背景提取器的存在，对于每张图均能自动得到相较于原文更多的背景共现对象文本，因此能够构造较多的负样本对，能进一步促进模型的学习和参数更新。

5 实验结果分析

我们在 PASCAL VOC2012 上进行了实验，首先生成 CAM 图，之后对其进行细化得到伪标签，最后利用伪标签训练一个全监督的分割网络。需要注意的是，作者开源的代码进行了一定更新，相较于 camera-ready 版本性能有了进一步提升，我们的实验均基于作者开源的代码版本进行。

与原文的对比结果如表1所示，我们的模型能够得到更高质量的初始 CAM 图，而伪标签和分割结果也达到了与原文相当的效果。

Method	CAMs	Pseudo	VOC12.*val*
AdvCAM	55.6	68.0	68.1
CLIMS(camera-ready)	56.6	70.5	69.3
CLIMS(github-repo)	58.6	**74.1**	**70.3**
CLIMS+(ours)	**60.1**	70.1	69.7

表 1: 与原文效果对比

可视化生成的 CAM 并与原文结果进行对比，可以看到我们的模型能够更进一步抑制背景区域，且能够激活更完整的目标区域，但仍然存在部分误激活的问题。

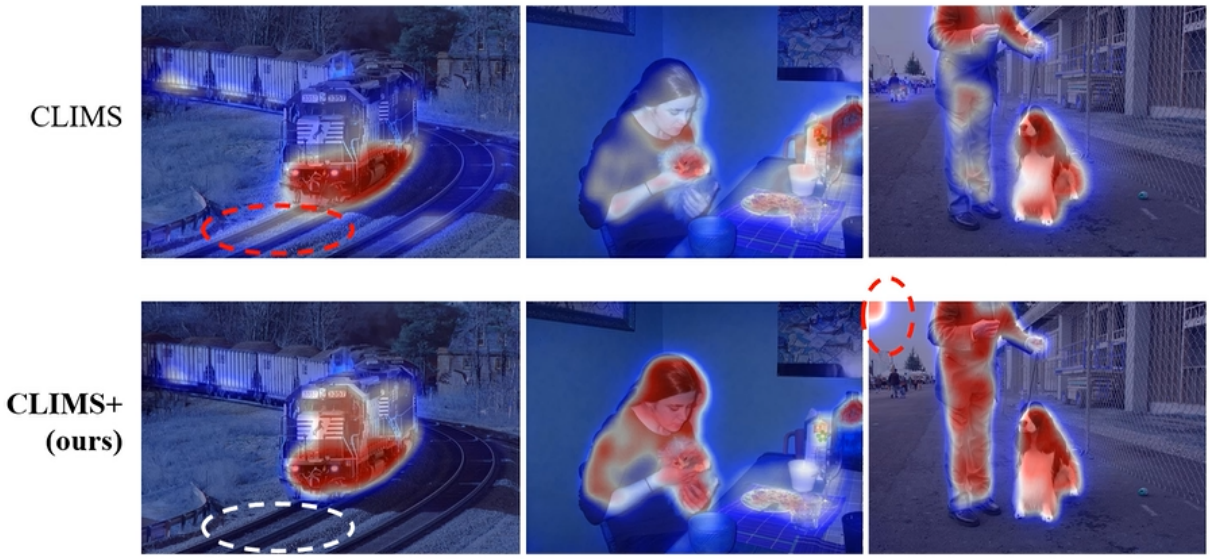

图 3: 可视化 CAM

6 总结与展望

本复现工作在跨语言图像弱监督模型 CLIMS 的基础上，通过引入一个背景提取器实现了目标相关背景文字的自动化提取，利用 InfoNCE 损失使模型更进一步学习和优化参数，最终生成了比原文更高质量的 CAM 图。但是该模型并不能由此得到高于原文的伪标签和分割结果，此外 CAM 图也存在部分误激活问题，这些都是未来可进一步进行研究的方向。

参考文献

[1] ZHOU B, KHOSLA A, LAPEDRIZA A, et al. Learning deep features for discriminative localization[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 2921-2929.

[2] RADFORD A, KIM J W, HALLACY C, et al. Learning transferable visual models from natural language supervision[C]//International Conference on Machine Learning. 2021: 8748-8763.

[3] HOU Q, JIANG P, WEI Y, et al. Self-erasing network for integral object attention[J]. Advances in Neural Information Processing Systems, 2018, 31.

[4] CHANG Y T, WANG Q, HUNG W C, et al. Weakly-supervised semantic segmentation via sub-category exploration[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 8991-9000.

[5] SUN G, WANG W, DAI J, et al. Mining cross-image semantics for weakly supervised semantic segmentation[C]//European conference on computer vision. 2020: 347-365.

[6] LEE J, KIM E, YOON S. Anti-adversarially manipulated attributions for weakly and semi-supervised semantic segmentation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 4071-4080.

[7] AHN J, KWAK S. Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 4981-4990.

[8] CHEN L C, PAPANDREOU G, KOKKINOS I, et al. Semantic image segmentation with deep convolutional nets and fully connected crfs[J]. arXiv preprint arXiv:1412.7062, 2014.

[9] CHEN L C, PAPANDREOU G, KOKKINOS I, et al. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs[J]. IEEE transactions on pattern analysis and machine intelligence, 2017, 40(4): 834-848.

[10] HE K, FAN H, WU Y, et al. Momentum contrast for unsupervised visual representation learning[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 9729-9738.