缩放开放词汇对象检测

摘要：

开放词汇目标检测从预训练的视觉语言模型中获益良多，但仍然受到可用检测训练数据量的限制。利用Web图像-文本对作为弱监督，可以扩展检测训练数据

Web图像-文本是指从互联网上获取的图像与相关文本之间的关联。在Web上的许多网页、社交媒体平台和在线内容中，图像通常伴随着与之相关的文本描述、标签、标题或评论。这些文本信息可以提供有关图像内容、语义和上下文的额外信息。

自我训练是一种半监督学习的方法，它首先使用少量的标注数据进行初始模型的训练。然后，使用这个初始模型对未标注数据进行预测，生成伪标注。这些伪标注数据可以被视为扩展的训练数据，然后将它们与真实标注数据一起用于进一步的模型训练。

在强监督学习中，通常需要为每个训练样本提供准确的标签，这需要相当大的人力和时间成本。然而，在许多实际情况下，获得准确标签的成本很高或不可行，这就需要使用弱监督学习方法。

弱监督学习可以利用不完全、嘈杂、不确定的标签信息来进行模型训练。这些弱监督信号可以是不完整的标签、部分标签、模糊标签、关键词、领域专家的指导或推测等。通过利用这些弱监督信号，模型可以从中学习到数据的潜在模式和结构。

CLIP 是由 OpenAI 提出的一种视觉语言预训练模型，它通过联合学习图像和文本来建立视觉和语义之间的联系。CLIP 使用对比学习的方法，通过最大化匹配图像和对应文本的相似性，从而使模型能够理解图像和语言之间的语义关联。CLIP 的预训练过程包括大规模收集的图像和文本数据集，并使用对比损失函数进行优化。CLIP 可以用于多种任务，如图像分类、图像生成描述、图像搜索等。

SigLIP 是 CLIP 的改进版本，旨在进一步提升图像和语言的联合学习性能。SigLIP 引入了一种更加高效的训练策略，通过使用大量的无标签图像和文本数据来进行预训练。此外，SigLIP 还引入了一种新的对比损失函数，称为"Significantly Improved Contrastive Loss"，它通过对比正负样本对之间的相似度进行建模，从而提高了模型的性能。

VLM 通常通过联合训练图像和文本数据来学习视觉和语言之间的联系。它可以接收图像和文本作为输入，并通过学习图像特征和语义表示之间的映射关系，实现对图像和语言的联合理解和表达。

VLM 的应用范围非常广泛。例如，在图像描述生成任务中，VLM 可以接收图像作为输入，并生成与图像内容相关的自然语言描述。在图像问答任务中，VLM 可以回答与图像相关的自然语言问题