计算机与自主智能研究进展
Advances in Computer and Autonomous Intelligence Research
第 3 卷◆第 1 期◆版本 1.0◆2025 年
文章类型：论文|刊号（ISSN）：2972-4236(P) / 2972-4244(O)

分割一切：基础视觉模型的发展

李昊喆　王志豪　黄利萍
东北大学软件学院
DOI:10.12238/acair.v3i1.11913

[摘　要] 本文综述了基础视觉模型的发展,重点介绍了Segment Anything Model(SAM),由Meta AI提出的提示驱动的通用图像分割模型。SAM通过大规模预训练,在处理高分辨率图像的同时保持了计算效率,展现出了强大的零样本泛化能力。SAM的核心技术包括图像编码器、提示编码器和掩码解码器,它们共同实现了高效的特征提取、灵活的提示处理和高质量的分割掩码生成。SAM的应用案例涵盖医学和遥感领域,显示了其在边界清晰、结构简单的图像分割任务中的优越性能,同时也揭示了在处理复杂场景时存在的挑战。未来的研究方向包括领域适配与模型优化、多模态融合与跨领域应用,以及更高效的提示工程技术,这些都将有助于进一步提升SAM的功能和适用范围。
[关键词] Segment Anything Model；提示驱动分割；多模态融合
中图分类号：U642.3+3　文献标识码：A

Dividing everything: The development of fundamental visual models

Haozhe Li　Zhihao Wang　Liping Huang
Software College of Northeastern University

[Abstract] This paper provides an overview of the development of foundational visual models, with a particular emphasis on the Segment Anything Model (SAM), a prompt-driven universal image segmentation model introduced by Meta AI. SAM demonstrates robust zero-shot generalization capabilities while maintaining computational efficiency when processing high-resolution images, thanks to large-scale pre-training. The core technologies of SAM consist of an image encoder, a prompt encoder, and a mask decoder, which collectively achieve efficient feature extraction, flexible prompt handling, and the generation of high-quality segmentation masks. Application cases of SAM span medical and remote sensing fields, highlighting its superior performance in segmenting images with clear boundaries and simple structures, while also exposing challenges encountered in more complex scenarios. Future research directions include domain adaptation and model optimization, multimodal integration and cross-domain applications, as well as the development of more efficient prompt engineering techniques, all of which will contribute to further enhancing the functionalities and applicability of SAM.
[Key words] Segment Anything Model; Prompt-driven Segmentation; Multimodal Integration

1 基础视觉模型的定义与发展

基础模型,通过在大规模数据集上进行预训练以获得在广泛的下游任务中的强大泛化能力,已经深刻影响了人工智能的发展。受到基础模型在自然语言处理(NLP)获得的巨大成功的启发(如BERT[1],openai的ChatGPT系列[2]等),研究人员已经开始探索基础模型在计算机视觉领域的潜力,计算机视觉正在迈入“大模型时代”[3]。

这些研究主要包括两个方向。一种是受大语言模型(LLMs)影响,研究者们开始探索大型视觉模型(LVMs),通过扩大Vision Transformers的规模来实现更强大的能力。这一方向中的几个重要模型包括ViT-G[4]、ViT-22B[5]、Swin Transformer V2[6]和VideoMAE V2[7]。这些模型的设计与训练方式旨在模仿大语言模型的模式,以期发现更复杂的、强大的能力。这些强大能力包括更优秀的特征表示能力和更强的上下文理解能力,从而提升模型在视觉任务中的表现。

除了单一的视觉模型的大规模化,还有许多研究集中于多模态模型的开发,结合视觉和语言的知识来增强模型的表现。例如,CLIP[8]和ALIGN[9]等模型采用了文本编码器和图像编码器,利用对比学习的方法,从大量的噪声图像-文本数据中学习视觉和语言的共同表示。通过这样的方式,这些模型在预训练后可以

将学习到的语义知识应用于新数据分布,使得模型在多个下游任务中展现出显著的性能。

尽管已经取得了显著的发展,但以上模型的泛化能力仍然有限。为此,2023年4月,Meta公布了一款名为SAM(Segment Anything Model)的模型[10],它在研究与任务无关的通用基础模型方面取得了显著的进步,获得了CV界的广泛关注。SAM是一个提示驱动(promptable)的通用的图像分割模型,通过110万图像上的十亿个掩码进行训练,在各种分割任务中表现出强大的零样本泛化能力。很多研究者甚至认为,SAM的出现使得计算机视觉领域进入了属于自己的GPT时代。

2 Segment Anything Model

2.1 Prompt

Prompt是提高模型表现力和定制化输出的重要手段,特别在图像分割任务中,它能够引导模型生成符合用户需求的分割结果。Prompt通过构造不同形式的输入提示,帮助SAM精准识别出需要被关注的对象或区域。

在SAM中,prompt被用作激活模型的信号,引导其执行特定的分割任务。与传统的图像分割模型相比,SAM的独特之处在于其能处理不同类型的提示,使得它更加灵活且通用性强。例如,通过在图像中标注一个点或一组点,SAM可以自动分割出该点附近的目标区域。而通过掩码提示,模型能够基于提供的掩码信息精确分割出与掩码对应的对象区域。

SAM可以处理多种不同类型的prompt,其中常见的包括:

点提示(Point Prompt):用户通过在图像中标记一个或多个点来指示感兴趣区域。这种方式简便且高效,特别适合那些需要快速获取特定区域分割的场景。

框提示(Box Prompt):通过在图像上画一个框,用户可以粗略地圈出目标区域,SAM将基于这个提示进一步细化分割结果。

掩码提示(Mask Prompt):用户可以提供初步的分割掩码,SAM会对此进行优化和精细调整,使最终分割更加准确。

尽管prompt提供了极大的灵活性,但如何设计有效的prompt仍是一个挑战。错误或模糊的提示可能导致不准确的分割结果,因此,用户在提供提示时需要考虑模型的理解方式。对于复杂的场景,单一类型的提示可能不足以生成理想的结果,用户需要结合多种提示来增强分割效果。

2.2 SAM架构

SAM的架构创新在于其模块化设计,允许分割任务中不同程度的定制化和优化处理。这一设计包括了高效的信息传输机制和多层级的特征提取能力,确保模型不仅能够处理高分辨率的输入图像,还能在精确性和计算效率之间取得理想的平衡。通过灵活的提示机制,SAM能够引导模型的注意力集中于特定图像区域,从而实现类似人类视觉理解的对象分割。

如图1所示,SAM由以下主要组件构成:

图像编码器.图像编码器负责将输入图像转换为嵌入表示。借助于Vision Transformer和MAE的预训练方法,SAM的图像编码器具备强大的特征提取能力。它能够处理高分辨率输入,同时生成高效压缩的图像嵌入,为后续处理提供丰富的视觉信息。

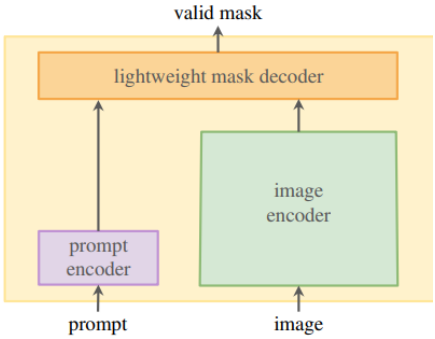

图1 Segment Anything Model

提示编码器.提示编码器允许SAM接收多种类型的提示输入,如点、框以及掩码信息。这种灵活性使得模型能够根据不同的提示类型,调整其注意力区域,从而更精准地定位和分割目标对象。

掩码解码器.掩码解码器的设计灵感来源于Transformer结构,通过自注意力和交叉注意力机制,将图像及提示编码器输出的信息高效整合。掩码解码器设有动态掩码预测头,可以在50毫秒内生成高质量的分割掩码,支持实时交互。值得注意的是,该解码器能够处理多重输出,以应对提示歧义,提供多个合理的掩码选项,实现复杂场景下的准确分割。

2.3数据引擎

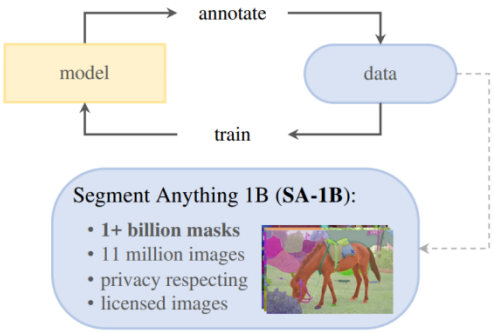

图2 数据引擎和数据集

数据引擎的引入是SAM取得成功的关键之一。数据引擎由三个阶段构成:手动辅助、半自动和全自动阶段。在手动辅助阶段,专业标注人员通过SAM的交互工具进行初始掩码标注。随着模型能力的提升,进入半自动阶段,SAM能够自动生成部分较高置信度的掩码,标注人员则着重于更加细微复杂的对象。在最后的全自动阶段,SAM完全自主地生成掩码。通过这种迭代和协作的数据收集过程,数据引擎为SAM提供了超过十亿个掩码,有效增强了模型对新图像分布的学习能力。

这种数据引擎策略不仅提升了SAM的训练效率,还确保了数据多样性和质量,使得SAM在零样本迁移任务中表现突出。通过与SA-1B数据集的结合,SAM得以在多种图像分割任务中实现自然的人机交互。

3 SAM的应用

3.1医学领域

SAM 被广泛应用于医学图像分割中,但其表现存在明显的任务依赖性。Mazurowski等人[11]进行了一个实验性研究,评估了SAM在19个不同的医学成像数据集上的分割表现,涵盖了MRI、CT、X射线和超声等多种成像模式。实验结果显示,SAM在边界清晰、结构明确的图像上(如髋关节X射线)表现较好,IoU高达0.8650。然而,在处理边界模糊、复杂度较高的图像(如脊柱MRI)时,表现相对较差,IoU仅为0.1135(医学)。

Huang等人[12]对SAM在医学图像中的应用进行了详细探讨,发现虽然SAM在某些简单的目标(如胸腔X射线)上有不错的分割表现,但在处理多部位、多尺度目标时,其效果有限。这与Mazurowski的研究一致,表明SAM在医学领域的应用尚需改进特别是对于复杂器官或病灶的分割任务。

此外,SAM在提示驱动的分割能力上展现了与其他交互式方法的显著差异。在Mazurowski等人的研究中,使用框提示的分割表现显著优于单点提示,尤其是在需要分割多个非连续区域时,框提示的精度提升明显(医学)。Deng等人[13]的研究进一步表明,SAM在病理图像中的零样本分割表现优异,尤其是在无监督的场景下,通过多点提示可以进一步提高分割精度。

尽管SAM在许多医学应用中表现出色,但其在3D医学图像中的应用仍然存在挑战。Liu等人[14]提出了一种将SAM与3DSlicer相结合的方法,展示了如何在三维医学图像分割任务中使用SAM进行快速、精确的标注。这种方法表明,通过结合现有的3D工具,SAM可以进一步拓展到更复杂的医学场景中。

3.2遥感领域

SAM同样被应用于遥感图像分割中,但其表现也存在一定的任务依赖性。Osco等人[15]进行了一项研究,评估了SAM在多尺度遥感数据集上的分割表现。实验结果显示,SAM在处理边界清晰、结构简单的图像时表现出色,但在面对复杂场景(如树木覆盖下的汽车)时,性能有所下降。为了克服这一局限性,研究人员提出了PerSAM-F方法,通过微调两个参数来改进SAM在不同尺度下识别目标物体的能力,从而提高了对复杂场景的分割精度。

此外,Wu和Osco[16]开发了samgeo Python包,旨在简化地理空间数据的分割过程,并为用户提供便捷的工具以优化注释创建流程。这种方法不仅加快了高质量训练数据集的生成速度,还展示了如何结合GIS与SAM来改善特定分割任务的结果。

Ma和Wang[17]的研究则探讨了通过微调SAM来提升其在特定任务中的性能,尤其是在肿瘤分割等高精度需求的任务中取得了显著的性能提升。这表明经过适配后的SAM可以在特定领域实现更准确的分割结果。

4 结语

随着基础视觉模型的发展,SAM在各个领域的应用前景广阔,尤其是在医学、遥感、自动驾驶等需要精确分割的任务中,SAM的零样本分割能力为处理大量未标注数据提供了高效、自动化的解决方案。

4.1领域适配与模型优化

目前,SAM的表现受限于其通用性设计,特别是在领域特定任务中,其精度较全监督模型有一定差距。未来,针对不同应用领域的优化将成为发展方向。通过引入更多领域特定的训练数据,尤其是在医学、遥感等特殊场景下,SAM可以通过微调进一步提升分割精度。此外,领域适配器和专用模块的引入将增强SAM处理复杂任务的能力,使其能够更好地应对各种具备挑战性的分割任务。

4.2多模态融合与跨领域应用

未来,SAM的多模态融合能力将是其发展的关键方向之一。当前,SAM主要基于图像信息进行分割任务,但在实际应用中,图像往往只是信息来源的一部分。通过与其他模态的数据结合,例如文本、语音、深度图像、雷达数据等,SAM有望在更广泛的复杂任务中展现出更强的能力。这种结合将推动类似大规模视觉模型(LVM)与大规模语言模型(LLM)融合,形成多模态模型(LMM)的发展趋势,从而在智能分割和认知任务中实现突破。

4.3更高效的提示工程

提示工程的进一步优化是SAM未来发展的关键方向之一。当前,框提示和点提示已证明在提升分割精度方面具有重要作用,但仍需更智能、自动化的提示生成方法。未来,通过结合更先进的人工智能技术,SAM或许可以实现基于少量初始输入自动生成最优提示,从而进一步提升分割的效率和准确性。此外,探索更多类型的提示(如形状提示或纹理提示)将为模型在不同分割任务中的应用带来更多可能性。

[参考文献]

[1]Devlin J.Bert:Pre-training of deep bidirectional transformers for language understanding[J].arXiv preprint arXiv:1810.04805,2018.

[2] Brown T B.Language models are few-shot learners[J].arXiv preprint arXiv:2005.14165,2020.

[3]宋婧.视觉大模型将是下一个风口[N].中国电子报,2024-01-09(005).

[4]Zhai X, Kolesnikov A, Houlsby N, et al. Scaling vision transformers[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition.2022:12104-12113.

[5]Dehghani M, Djolonga J, Mustafa B, et al. Scaling vision transformers to 22 billion parameters[C]//International Conference on Machine Learning.PMLR,2023:7480-7512.

[6]Liu Z, Hu H, Lin Y, et al. Swin transformer v2: Scaling up capacity and resolution[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022:12009-12019.

[7]Wang L,Huang B,Zhao Z, et al. Videomae v2: Scaling video masked autoencoders with dual masking[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.2023:14549-14560.

[8]Radford A,Kim J W,Hallacy C,et al.Learning transferable visual models from natural language supervision[C]//International conference on machine learning.PMLR,2021:8748-8763.

[9]Jia C,Yang Y,Xia Y,et al.Scaling up visual and vision-language representation learning with noisy text supervision [C]//International conference on machine learning. PMLR, 2021: 4904-4916.

[10]沈怡然.对话三位IEEE专家:如何理解SAM视觉大模型[N].经济观察报,2023-08-21(006).

[11] Mazurowski M A, Dong H, Gu H, et al. Segment anything model for medical image analysis: an experimental study[J]. Medical Image Analysis,2023,89:102918.

[12]Huang Y, Yang X, Liu L, et al. Segment anything model for medical images?[J].Medical Image Analysis,2024,92:103061.

[13]Deng R, Cui C, Liu Q, et al. Segment anything model (sam) for digital pathology: Assess zero-shot segmentation on whole slide imaging[J]. arXiv preprint arXiv:2304.04155, 2023.

[14]Liu Y,Zhang J, She Z, et al. Samm (segment any medical model):A 3d slicer integration to sam[J].arXiv preprint arXiv: 2304.05622,2023.

[15]Osco L P,Wu Q,de Lemos E L, et al.The segment anything model (sam) for remote sensing applications: From zero to one shot[J]. International Journal of Applied Earth Observation and Geoinformation, 2023,124:103540.

[16]Wu Q,Osco L P.samgeo:A Python package for segmenting geospatial data with the Segment Anything Model (SAM)[J]. Journal of Open Source Software,2023,8(89):5663.

[17]Ma A,Wang J,Zhong Y,et al.FactSeg:Foreground activation-driven small object semantic segmentation in large-scale remote sensing imagery[J].IEEE Transactions on Geoscience and Remote Sensing,2021,60:1-16.

作者简介：

李昊喆(2003--),男,汉族,河北省邯郸市人,本科,研究方向：计算机视觉。