##### 深度学习网络结构调研（100）

任选一种应用于目标检测、语义分割或超分辨率等与图像相关领域的的深度学习网络结构。

完成一份调研报告，包括但不限于以下内容：

(1) 该网络结构的发展历史、应用场景、优缺点等。

(2) 该网络结构的具体结构，如模块设计、激活函数选择、损失函数设计等。

(3) 该网络结构的改进版本、变种等。

(4) 该网络结构的参数量、训练速度、推理速度等。

* 报告无需涉及到上述所有内容，写出你认为最关键的若干内容即可，不超过五页。
* 调研网络结构举例：ResNet、YOLO、R-CNN、Vision Transformer、U-Net、SAM 等。

# SAM模型网络结构调研

2100013260 袁怡康 地空22本

##### ****引言****

2023年4月，Meta AI发布的“Segment Anything Model”（SAM）引起了计算机视觉（CV）领域的广泛关注。该模型被认为是视觉领域的“大模型”，具备“一切皆可分割”的能力，尤其在图像分割任务中取得了显著进展。尽管目前该模型尚未达到“视觉大模型”的完整范畴，它在物体分割领域的表现无疑为未来的视觉任务做好了铺垫，且为图像分割技术带来了革命性的改变。

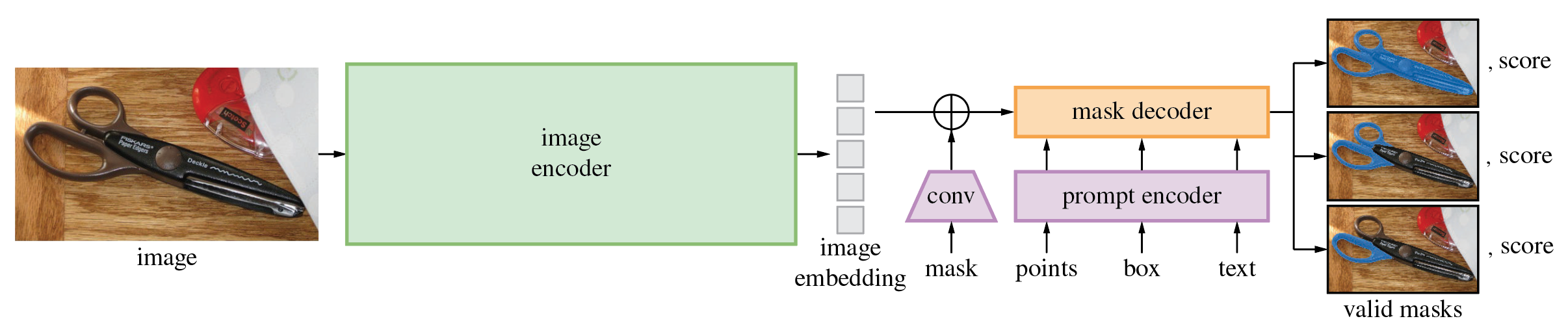
##### ****SAM模型概述****

SAM是一个专为图像分割设计的深度学习模型，具有零样本分割能力，即使面对没有见过的图像，也能够进行高效的分割。该模型通过输入不同类型的提示（如点、框、掩码甚至文本）来对图像进行分割任务。SAM在一个规模庞大的数据集上进行了训练，包括1100万张图像和11亿个掩码标签。它的训练数据来源不仅包括人工标注数据，还包括由模型自我标注的数据，这一策略得益于Meta设计的强大数据引擎（Data Engine）。

目前，SAM提供了三个版本的预训练模型：VIT-H SAM、VIT-B SAM和VIT-L SAM。所有版本都基于Vision Transformer（ViT）架构，并且在图像分割任务中表现出强大的零样本（Zero-shot）能力，能够在各种分割任务中展示出高效的性能。

##### ****SAM网络结构****

SAM模型的网络架构由三个主要组件构成：图像编码器（Image Encoder）、提示编码器（Prompt Encoder）和掩码解码器（Mask Decoder）。



1. ****图像编码器****

SAM的图像编码器是基于Vision Transformer（ViT-H）的架构，Vision Transformer架构通过自注意力机制和前馈网络来建模图像中的长距离依赖关系，使得该模型能够有效地理解复杂的图像结构。具体数据流程如下：输入尺寸为1024×1024的RGB图像，以64×64的patch从原始图像中生成尺寸为64×64的特征图，然后使用卷积进行特征提取，最后输出嵌入特征。

1. ****提示编码器****

SAM的提示编码器负责处理输入的提示信息，将提示信息编码成大小为256的特征向量，支持的提示类型包括点（正点、负点的点坐标）、框（边框角点的坐标）、掩码（掩码采样点坐标）和文本等。不同的提示类型能够为模型提供不同的上下文信息，从而引导模型进行更准确的分割，同时编码器的设计使得SAM模型能够灵活处理各种类型的用户输入提示，增强了模型的交互性和灵活性。

1. ****掩码解码器****

SAM的掩码解码器负责将图像特征和提示编码转换为分割掩码。该解码器使用了Transformer解码器和多层感知机（MLP），输出的是多个候选掩码以及与分割结果相关的交并比（IoU）预测值，集成得到每个掩码预测分数，以便根据应用场景调整掩码输出。

##### ****SAM的优势与局限性****

1. ****优势****

**零样本分割能力（Zero-Shot）**：SAM基于庞大的训练数据集进行训练（1100万张图像和11亿个掩码标签），实现了零样本分割能力，即使在未见过的图像数据集上，也能表现出良好的分割效果。

**实时交互性**：SAM支持多种交互形式的提示，包括点（正点、负点）、框、掩码甚至文本，能够可视化、实时化展示分割结果，增强了其在不同任务中的灵活性与适用性。

**简单而高效的模型**设计：尽管SAM的模型设计较为简单，但其基于强大数据引擎（data engine）的训练策略使得其在大规模数据集上能够有效地学习和泛化。

1. ****局限性****

不支持语义分割：SAM的主要局限性在于其不支持语义分割任务，只能进行实例分割。因此，对于一些需要语义信息的任务，SAM尚不适用。但是后续基于SAM模型的诸多该进版本在不断尝试基于SAM实例分割增加语义标签，扩展SAM的功能。

分割精度：SAM模型的能力尚未达到能够分割所有复杂场景，例如遥感图像中的道路分割，其分割精度可能受限。

计算资源需求较大：SAM训练过程依赖于庞大的数据集和强大的计算资源，这使得它在某些计算能力较弱的环境中应用受限。

##### ****模型规格****

SAM模型的训练使用了世界各地的大规模数据集，包含1100万张图像和11亿个掩码标签。其巨大的数据训练结果保证了模型的高泛化能力，甚至实现零样本分割的效果。

模型参数量：参照SAM论文，以ViT-H SAM为例，图像编码器参数量约为632M，提示编码器约为1M，掩码解码器约为3M，总参数量约为10亿。

推理速度：基于不同的硬件设施推理速度不一样，不同模型推理速度不一。ViT-H 运行占用~16GB GPU内存，ViT-L运行占用 ~11GB GPU内存，ViT-B运行占用 ~6GB GPU内存。

##### ****SAM模型的改进与变种****

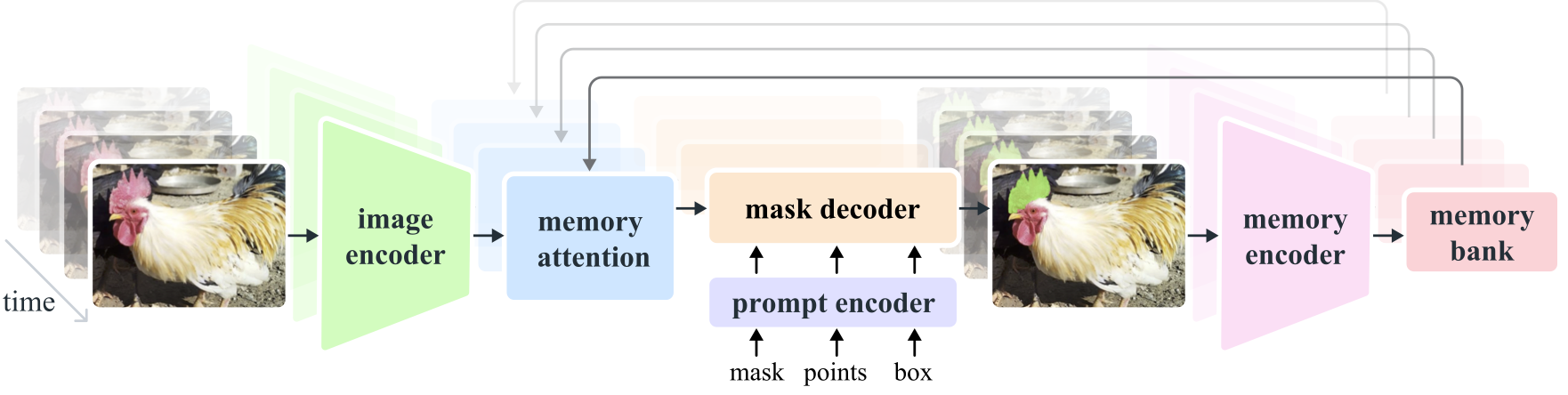
1. ****Mobile-SAM****

Mobile-SAM模型在网络结构上基本延续SAM模型的原架构，但在图像编码器部分进行了替换，实现了轻量级架构和更快的推理速度以及更少的内存占用，同时保持SAM的零样本分割能力和高精度掩码。

1. ****Grounded-SAM****

Grounded-SAM是集成SAM、BLIP、Stable Diffusion等各种专家模型为一身的模型，将计算机视觉领域的复杂问题分解成若干子任务，汇总输出得到最终推理结果。Grounded-SAM首先使用Grounding DINO通过文本信息生成对象选框，作为框提示输入SAM模型生成实例精确掩码，为后续高效完成图像分割、检测、识别、标注任务做好铺垫，包括自动图像标注、精确图像编辑等在内的多种任务。

1. ****SAM2****



今年八月份左右，Meta AI发布了SAM的最新版本——SAM2。SAM2作为SAM模型的改进版本，继承了基础模型的设计理念，并引入了一个新的维度：基于视频的时间序列记忆（Memory）维度。SAM2不仅能够进行单帧图像的分割任务，还能够将图像分割扩展到视频序列中，支持基于视频内容的分割任务。SAM主要应用于静态图像的分割任务，而SAM2能够处理更加复杂的动态数据，进一步扩展了“视觉大模型”在各领域应用前景，例如监控系统、自动驾驶、机器人视觉等。

SAM2延续了SAM的网络架构，增强的Transformer架构，引入更深的网络层次和注意力机制，增强了处理复杂图像的能力。

SAM2还进一步强化了Data Engine的循环机制，通过用户不断交互生成新数据，为模型提供了强有力的训练支持，创造了最大的分割视频数据集——SA-V。这一创新极大地提升了SAM2在视频分割任务中的应用潜力。

1. **总结**

SAM模型凭借其出色的零样本分割能力和强大的交互性，已经在图像分割领域取得了巨大的进展。通过对图像和提示的高效编码，SAM能够在不依赖大量标注数据的情况下，提供准确的分割结果。尽管目前在语义分割方面存在局限性，但随着SAM2的发布，模型在视频分割等新领域的潜力得到了进一步拓展。未来，随着技术的不断发展，SAM系列模型有望在更广泛的计算机视觉任务中发挥重要作用。