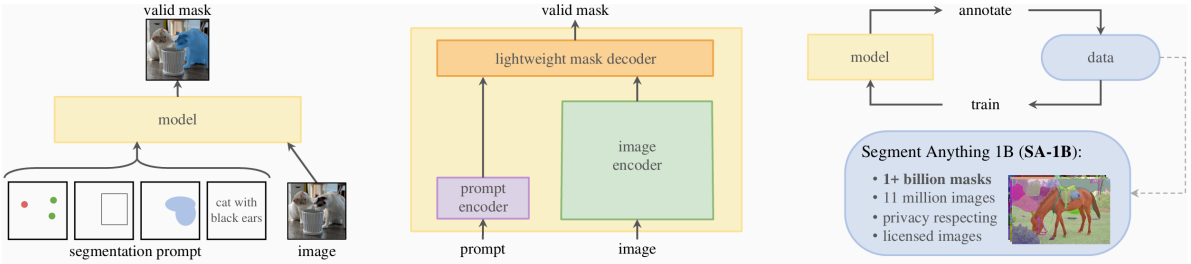
Segment Anything Model (SAM)

论文链接：Segment Anything

SAM是一项旨在为图像分割领域创建一个“基础模型”的开创性工作。您可以把它想象成一个专门用于从图像中“抠图”的ChatGPT。该项目推出了一个新模型（SAM）、一个海量的新数据集（SA-1B），以及一个名为“可提示分割”的新任务。

其核心目标是构建一个单一、强大的模型，使其能够处理各种各样的分割任务，并且在面对从未见过的新图像时也无需重新训练。这种能力被称为**零样本泛化 (zero-shot generalization)**。



(a) **Task**: promptable segmentation (b) **Model**: Segment Anything Model (**SAM**) (c) **Data**: data engine (top) & dataset (bottom)

核心思想：可提示分割 (Promptable Segmentation)

其中心概念是一种名为**可提示分割**的新任务。与传统模型为特定分割工作（例如只寻找猫）进行训练不同，SAM模型被设计为响应一个**“提示” (prompt)**。提示会告诉模型在图像中要分割什么。这个提示可以是：

- **稀疏提示 (Sparse prompts):** 物体上的一个或多个点、围绕物体的边界框，甚至是描述它的文本（例如，“那只猫”）。
- **密集提示 (Dense prompts):** 一个大致的、类似掩码的形状。

模型的任务是根据给定的提示，输出一个精确且有效的分割掩码。即使提示是模糊的（例如，在衬衫上点击一下，可能指的是衬衫，也可能指的是穿衬衫的人），模型也被训练成至少为其中一个可能的对象生成一个合理的掩码。

模型：SAM的架构

Segment Anything Model (SAM) 的设计兼顾了灵活性和实时性能。它主要由三个部分组成：

1. **图像编码器 (Image Encoder):** 一个强大的视觉Transformer (ViT) 会处理输入图像，并为其创建一个详细的嵌入（一种数字表示）。这是计算量最大的部分，但每张图片只需运行一次。
2. **提示编码器 (Prompt Encoder):** 这个轻量级组件能高效地将任何输入提示（点、框、文本或掩码）转换为其自身的嵌入。
3. **快速掩码解码器 (Fast Mask Decoder):** 该部分接收图像嵌入和提示嵌入，并能在大约50毫秒内生成最终的分割掩码。

一个关键特性是它处理**模糊性**的能力。对于单个提示，它可以输出多个有效的掩码（例如，一个轮胎、一个车轮和一整辆车），并按置信度分数对它们进行排序。整个过程的效率非常高，足以在网页浏览器中实时运行，从而实现了无缝的交互式使用。

数据：SA-1B数据集与数据引擎

要训练这样一个强大的模型，需要一个极其庞大且多样化的数据集——而这样的数据集并不存在。因此，研究人员构建了一个**“数据引擎” (data engine)** 来创建它。这是一个巧妙的三阶段过程，利用模型自身来辅助收集更多数据，然后用这些新数据重新训练和改进模型，形成一个持续的循环。

1. **辅助手动阶段:** 最初，人类标注员使用早期版本的SAM作为一个智能画笔工具，以更快地标注掩码。
2. **半自动阶段:** 随着SAM的改进，它能够自动识别并分割一部分对象。人类标注员则专注于标注SAM遗漏的更复杂的对象，从而增加了数据的多样性。
3. **全自动阶段:** 在最后阶段，能力极强的SAM接收到覆盖数百万张图像的网格点提示，使其能够平均每张图像自动生成约100个高质量的掩码。

这个过程最终催生了**SA-1B数据集**，这是同类数据集中规模最大的一个，包含在1100万张图像上的**超过10亿个高质量掩码。**该模型（SAM）和这个数据集已向公众发布，以鼓励更多关于计算机视觉基础模型的研究。