Segment Anything

中国科学院计算技术研究所
Institute of Computing Technology,Chinese Academy of Sciences

Segment Anything

Introduction

- 大语言模型通过prompt engineering实现很强的泛化能力
- 在视觉领域探索大模型
- 建立一个图像分割的基础模型
  - 可提示的模型
  - 在大规模数据集上进行预训练
  - 拥有很强的泛化能力
  - 通过提示工程解决下游分割任务

Segment Anything

Introduction

- 为了实现一个可提示的分割大模型

- 如何定义任务能够拥有零样本泛化能力？
  定义一个可提示的分割任务

- 相应 的模型结构？
  支持灵活的提示，实时输出分割masks

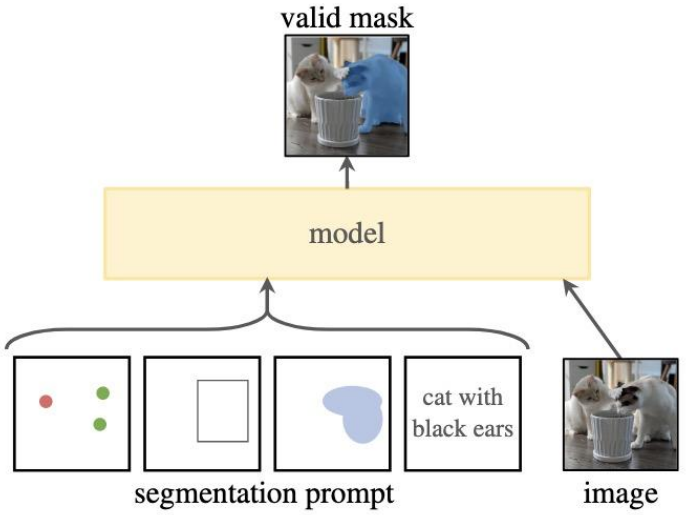
- 对于该任务和模型，需要什么样的数据集?
  需要一个多样的大规模数据集 (data engine)

Segment Anything

Task

- 提出一种可提示的分割任务
- 给定任何分割提示，返回一个有效的分割掩码。

- prompt指出具体的分割目标（空间或者文本信息）
- 有效的分割掩码值即使提示比较模糊也应给出至少一个有效的目标mask

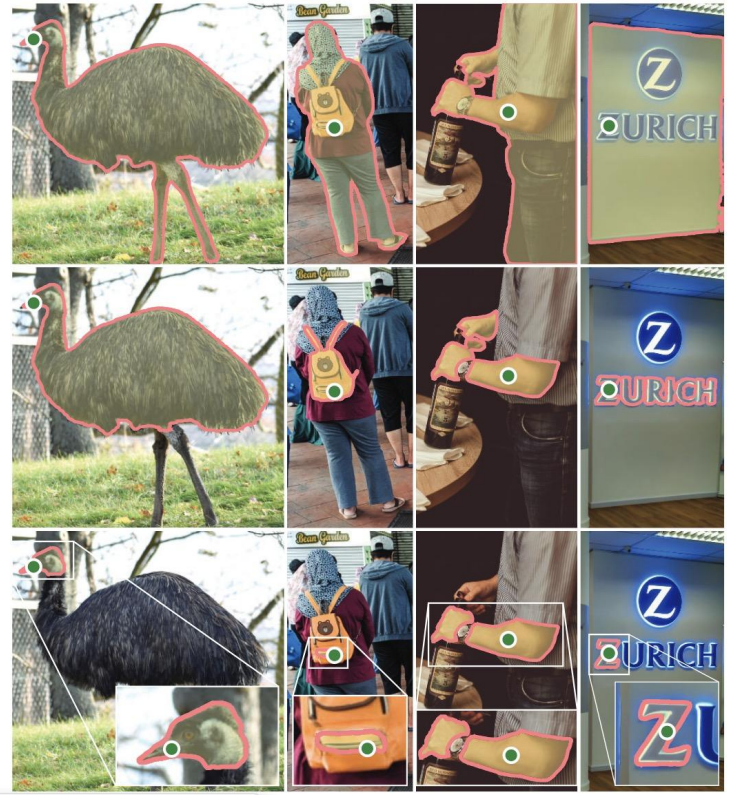


Task : promptable segmentation

Segment Anything

Task

- 类似于NLP foundation model中的prompt，分割任务中的prompt可以有多种形式
  - foreground/background points
  - box or mask
  - free-form text
  - 任何可以表达分割图片中目标的信息
- 对于模糊的prompt，模型也应给出有效的多个物体的掩码
- 通过prompt实现对下游分割任务的zero-shot迁移，通过prompt engineering训练的模型相比于对特定任务的模型有更广的应用范围。

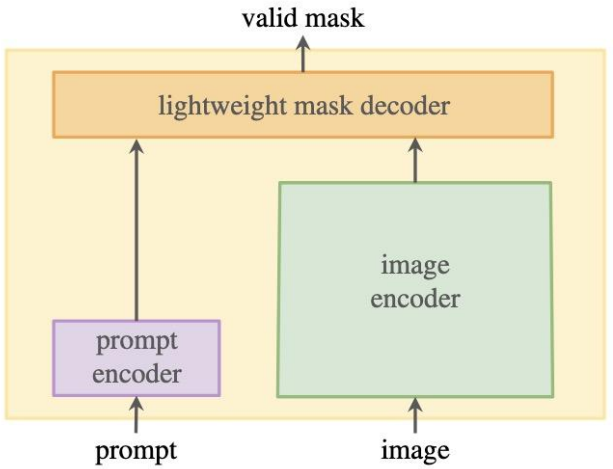


Each column shows 3 valid masks generated
by SAM from a single ambiguous point prompt (green circle)

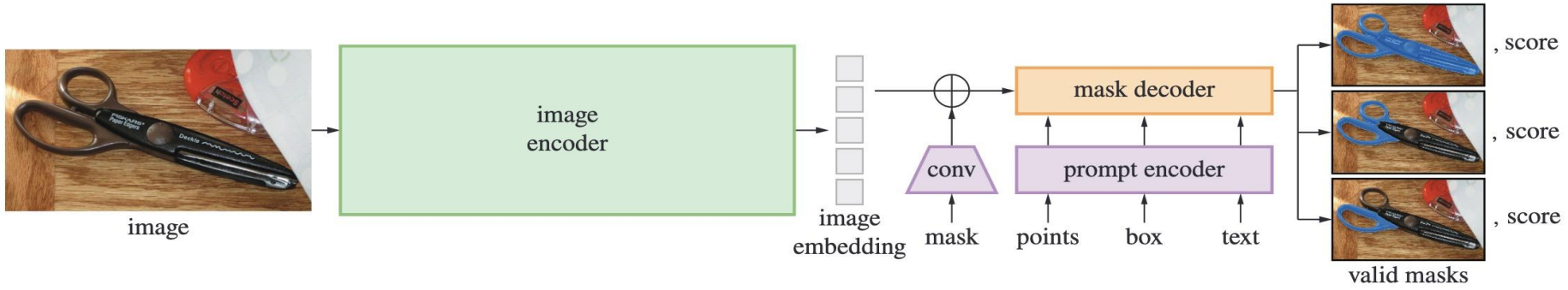
Segment Anything

Model

- ➢ 模型结构的三个限制。
- ● 模型必须支持多种提示。
- ● 模型必须实时计算出掩码。
- ● 模型必须对模糊的提示也要给出掩码。
- ➢ 模型结构的三个组成部分。
- ● Image encoder
- ● flexible prompt encoder
- ● fast mask decoder



Model : Segment Anything Model



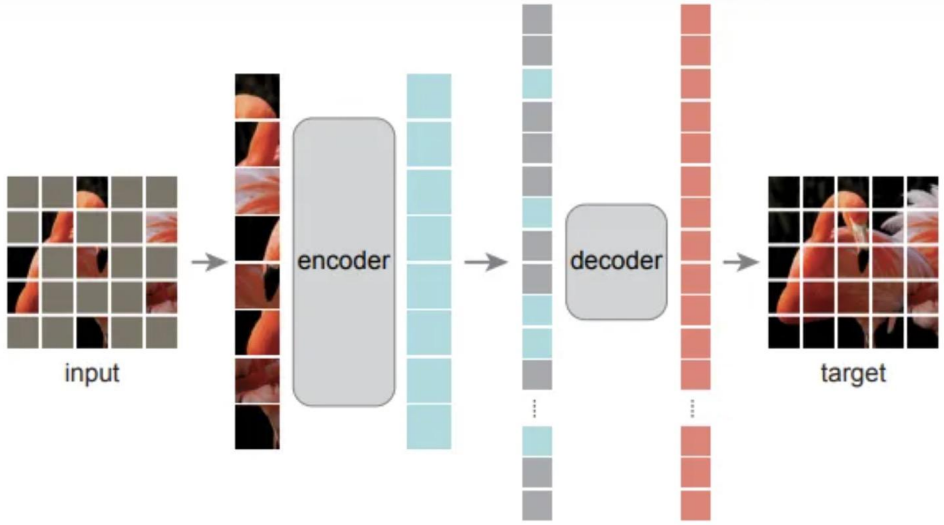
Segment Anything Model overview

Segment Anything

Model --- Image encoder

- Image encoder 使用MAE预训练ViT处理输入图片。
- Image encoder对每张图片编码一次，可在prompt之前完成
- MAE(Masked Autoencoders)是用于CV的自监督学习方法
- 随机遮住大量的块，然后去重构这些被遮住的像素信息，让它使用一个非对称的编码器和解码器的机制
- 非对称：编码器和解码器看到的东西不一样

1. 编码器只看到可见块
2. 解码器拿到编码器的输出之后，重构 masked patches

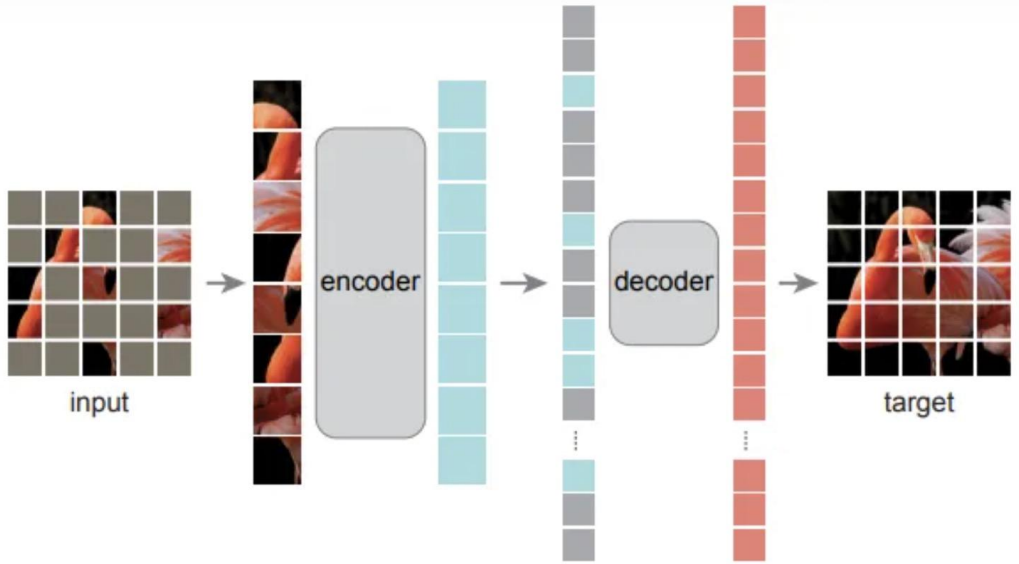


Segment Anything

Model --- Image encoder

- MAE训练流程

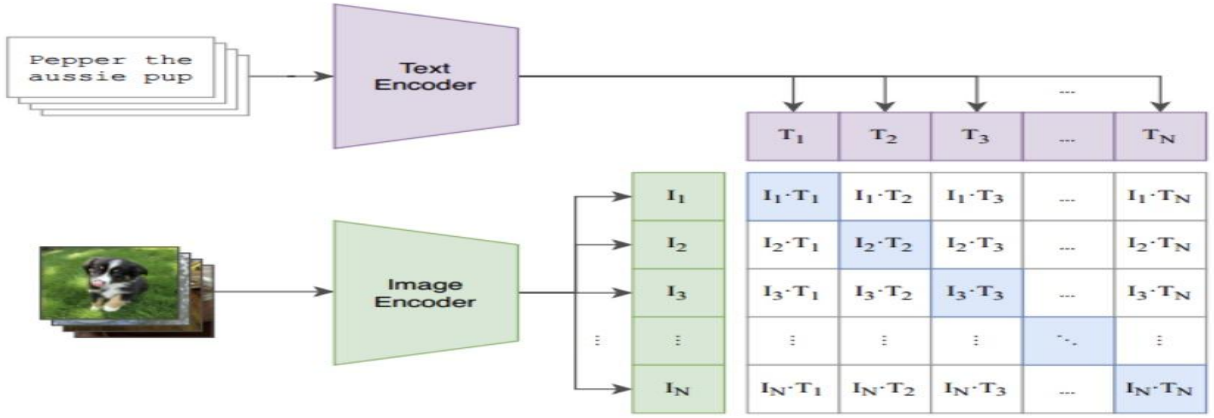
- 首先对image切分为patches，执行mask操作(灰色)
- 把可见的patches送入encoder(ViT)得到每一块的特征(蓝色)
- encoder 的输出 和 masked tokens 按照在图片中的原始位置排列成一长条向量 (包含位置信息)
- 再将encoder的输出(latent representations)以及 mask tokens作为decoder的输入
- 解码器尝试重构缺失的像素信息，还原原始图片。



Segment Anything

Model --- Prompt encoder

- Prompt Encoder 分为两种不同的prompts
- sparse(points, boxes, text)和dense(masks)。
- sparse prompts使用positional encodings
- free-form text prompts使用CLIP(Contrastive Language-Image Pre-training)
- dense prompts使用卷积与image embedding进行相加

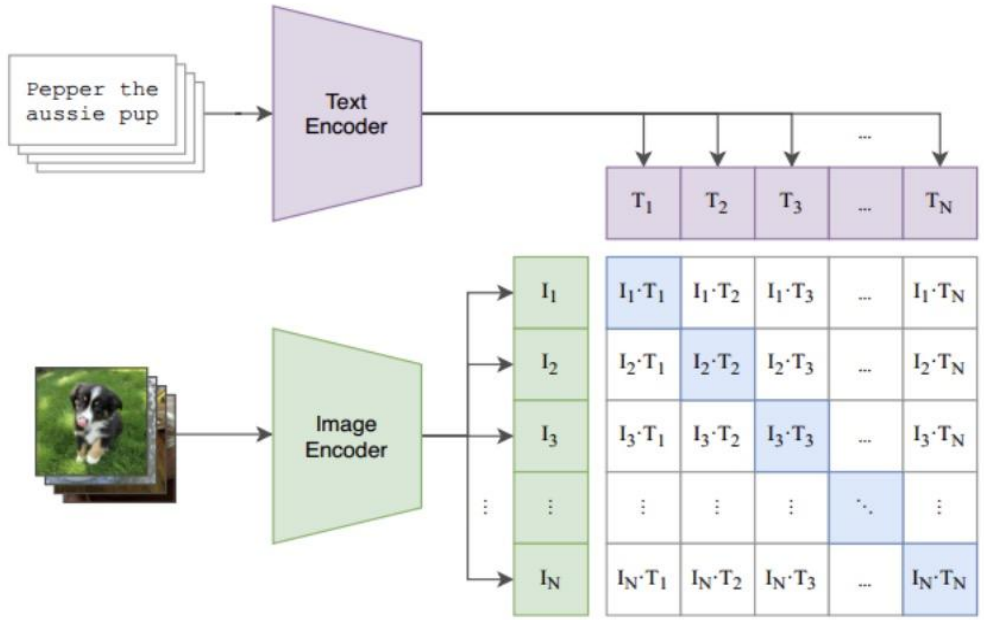


OpenAI 在 2021 年初发布的用于匹配图像和文本的预训练神经网络模型

CLIP 模型使用 OpenAI 收集到的 4 亿对图像文本对，分别将文本和图像进行编码，之后使用 metric learning 进行训练，其目标是将图像与文本的相似性提高

Segment Anything

Model --- Prompt encoder



- 输入图片->图像编码器（vision transformer->图片特征向量
- 输入文字->文本编码器（text ）->文本特征向量
- 对两个特征进行线性投射，得到相同维度的特征，并进行L2归一化
- 计算两个特征向量的相似度（夹角余弦）
- 让匹配的图文相似性最大，不匹配的图文相似性最小。

两个encoder分别处理文本和图像数据
encoder representation线性投影到multi-model embedding space
计算两个模态之间的cosine similarity，让匹配的图文相似性最大，
不匹配的图文相似性最下。

Segment Anything

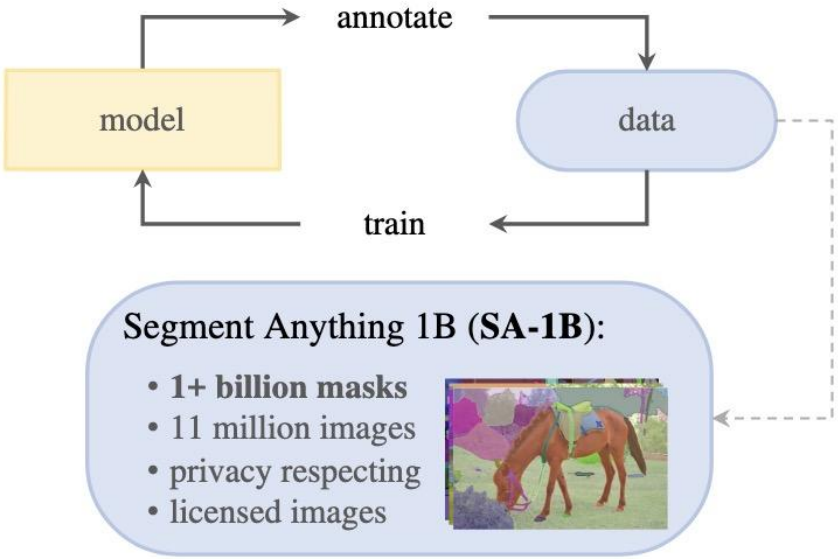
Data engine

- 为了让SAM可以泛化到新的数据分布上，需要在大规模数据集上训练模型

- 建立一个data engine，同时进行模型训练和数据集标注

- 分为assisted-manual, semi-automatic和fully automatic三个阶段

- SA-1B数据集在11M图片上产生超过1B masks



Data : data engine(top) & data(bottom)

Segment Anything

Data engine

- Assisted-manuual stage

- 交互式标注
- 标注人员对foreground/background object points进行标注
- 要求标注人员按照目标的显著程度进行标注
- 鼓励标注人员在30秒以后再进行下一张图片的标注
- 先使用公开的分割数据集进行训练，之后再使用新标注的mask进行训练
- 随着SAM性能提升，每张图片的masks从20个增加到44个
- 在这一阶段，从120k图片收集4.3M masks

Segment Anything

Data engine

- Semi-automatic stage

- 提高标注的多样性，进而提高模型segment anything的能力。

- 先让模型自动分割出置信度比较高的mask
  再让标注人员基于这些标注去给出额外的unannotated objects

- 为了给出confident mask，训练一个目标框检测器(Faster R-CNN)

- 在180k图像上收集5.9M masks

Segment Anything

Data engine

- Fully-automatic stage

- 在这个阶段，完全自动标注
- 在此之前已经足够多的mask训练模型
- 模型也拥有了ambiguity-aware的能力
- 具体来说，给模型一个32*32的格子，对于每一个点让模型预测出一些valid object mask
- 如果一个点位于part or subpart,模型给出subpart, part, whole object
- 选择可信稳定的masks并使用极大值抑制过滤掉重复部分
- 在11M图片上得到1.1B masks，SA-1B dataset

Segment Anything

Foundation models

- 越来越多的预训练模型用到下游任务
- 预训练模型强大的能力来自大规模的监督训练
- 当data engines可以提供大规模标注数据时，监督训练可提升预训练模型能力

Segment Anything

Compositionality

- SAM模型通过提供有效的mask可以与其它components进行可信交互
- 对于单RGB图像的3D场景重建可以使用SAM强大的泛化能力分割出未见过的目标
- SAM可以通过可穿戴设备检测到的注视点进行提示，从而实现新的应用

Segment Anything

Limitations

- 丢失fine structure，边界的精细成都不够
- SAM在使用较大的图像编码器时无法做到real-time
- 如何设计简单的提示，实现语义和全景分割

Segment Anything

Limitations

- 丢失fine structure，边界的精细成都不够
- SAM在使用较大的图像编码器时无法做到real-time
- 如何设计简单的提示，实现语义和全景分割

中科院计算所

Thanks