1112 專題期末作業

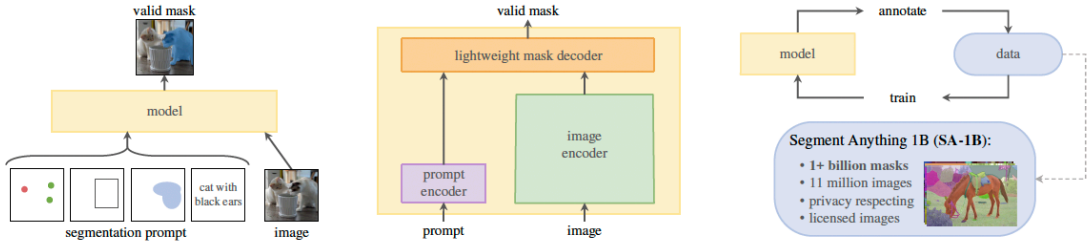
陳相瑋

June 20, 2023

論文：Segment Anything[1]

1 簡介

Segment Anything Model(SAM) 是 Meta AI 發布的可提示影像分割模型，使用超過一千萬張照片與十億個 Mask 作爲訓練資料。該模型旨在作爲通用的影像分割的基礎模型，不需額外訓練即可對任意圖片與相應的提示給出可能的分割 (zero-shot generalization)。



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

2 SAM 模型

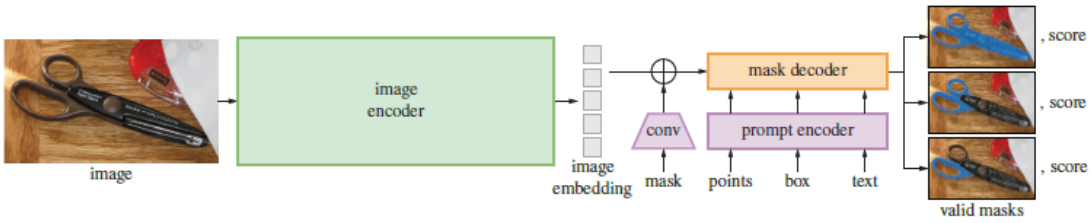


Fig 1: SAM 模型

2.1 可提示分割任務

在影像分割中，可提示分割任務的輸入層除了圖片，還包含提示（segmentation prompt），而模型會根據 Prompt 分割出一個對應的區域，如 Fig 2。其中提示包含點、區塊、文字等，在 SAM 中，Prompt 會先經過 prompt encoder 轉換。

可提示分割任務的優點在於標注容易，一張圖片標注單一 Mask 即可作爲訓練資料。

Fig 2: 在 SAM 中，若 Prompt 僅給予一個點則會產生三個 Mask

2.2 Image encoder

使用 Masked Autoencoders (MAE)[2] pre-trained Vision Transformer (ViT)。MAE 爲一種去躁 Encoder，會將輸入圖片切個成多個區塊，並隨機遮蔽 75% 區塊 (即生成 Fig 1的 image embedding)，最後訓練 Decoder 還原出原始圖片，和原圖比對計算 loss，爲自監督式學習。

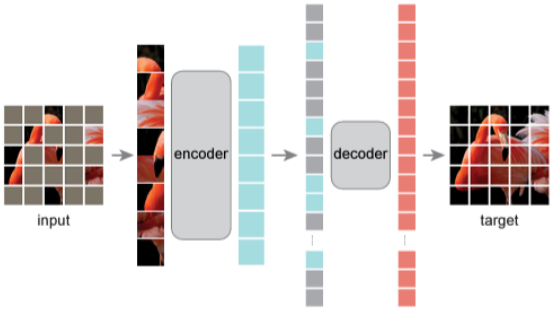


Fig 3: MAE

2.3 Prompt encoder

SAM 模型的 Prompt 輸入分成兩類:sparse (points, boxes, text) 和 dense (masks)

3 Dataset 訓練資料

資料標註分成三個階段，

- Assisted-manual stage: 使用人爲標註的圖片與 mask，共 4.3M mask 與 120K image，先訓練 SAM 模型。

- Semi-automatic stage: 使用 Assisted-manual stage 訓練的 SAM 模型自動產生 mask，再由人評價 mask。額外產生共 5.9M mask 與 180K image 用於訓練。
- Fully automatic stage: 由 SAM 模型自動產生 Mask 再自動評價 Mask。總共生成 1.1B mask 與 11M image 用於訓練。

4 評論

這篇論文最重要的部分是 Dataset 的三階段擴增的方法，該方法能有效的擴充 Dataset。SAM 使用的模型架構十分簡單，且皆是在圖像分割領域常見的方法，但是卻使用了非常大的 Dataset 進行訓練，最後得到非常好的分割結果。他說明了在圖像分割上，設法產生夠多的訓練資料對提升模型的辨識成果是非常有效的。

參考資料

[1] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. *arXiv:2304.02643*, 2023.

[2] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16000–16009, 2022.