Segment Anything

목적:

이미지 세분화를 위한 기반 제로샷 모델을 개발하는 것을 목표로 기존의 이미지 세분화 작업에서 발생하는 데이터 부족 문제와 고비용 레이블링 문제를 해결하기 위한 접근 방식을 제시

- **제로샷(zero-shot)?** 모델이 특정 작업을 훈련하지 않은 상태에서 그 작업을 수행할 수 있는 능력을 의미

제로샷이 가능한 이유?

- 데이터 엔진: 데이터 엔진은 모델 학습에 필요한 마스크 셋 데이터를 생성하는 역할을 한다. 이 엔진은 수동, 준자동, 완전 자동 방법을 통해 다양한 방식으로 마스크를 생성하여 다양한 객체와 상황에 대한 정보를 학습하고 일반화 성능을 향상시키는데 마스크 디코더와 역할은 비슷하지만 데이터 엔진은 어디까지나 훈련 용도로 사용된다.

- 기반이 되는 훈련 데이터: 거대한 이미지 데이터셋에서 사전에 훈련된 모델은 새로운 이미지에서도 공통된 특징을 추출할 수 있는 능력을 갖추게 됩니다. 이렇게 함으로써 새로운 이미지에서도 일반화된 특징을 적용할 수 있게 됩니다.

모델 아키텍쳐:

이미지 인코더 - 프롬프트 인코더 - 마스크 디코더로 구성

1. 이미지 인코더: 이미지를 특별한 방식으로 처리해서 중요한 정보(임베딩)추출하는 역할로 SAM 에서는 이를 위해 ViT 모델 사용을 사용.

   - **ViT?** Vision Transformer 의 약자, 컴퓨터 비전 작업을 위해 개발된 딥 러닝 모델. 이 모델은 주로 이미지 인식과 관련된 작업을 수행하기 위해 사용함. ViT 는 기존의 컨볼루션 신경망(CNN) 아키텍처와는 다르게, 트랜스포머(Transformer) 구조를 이용하여 이미지를 처리함.

   - 이미지 인코더에서 사용되는 **임베딩(Embedding)?** 고차원 데이터(이미지의 경우 픽셀 값)를 저차원 공간으로 변환하는 것을 의미

2. 프롬프트 인코더 : 두 가지프롬프트로 구성되어있고 다양한 형태의 프롬프트 정보를 모델이 이해할 수 있는 임베딩으로 변환하는 역할을 한다.

   ■ 희박한 프롬프트(sparse prompt):

      ▪ 점(point), 상자(Box): 학습된 임베딩과 함께 요약된 위치 인코딩
      ▪ 텍스트: CLIP 의 출력.(CLIP 모델에서 생성된 이미지에 대한 텍스트 설명을 의미)

   ■ 밀집한 프롬프트(dense prompt)

      ▪ 마스크: 이미지 임베딩에 합성곱을 수행

3. 마스크 디코더 : 이미지 인코더와 프롬프트 인코더로부터 얻은 두 임베딩을 사용하여 특정 객체의 마스크를 예측하여 생성한다.

   이 과정에서 SAM 은 트랜스포머 디코더를 사용하는데, 트랜스포머 디코더는 픽셀 하나하나를 시퀀스로 취급하여 앞뒤 픽셀 간의 관계를 파악해가며 픽셀 사이 문맥을 포착할 수 있어 객체의 경계를 더 정확히 분할 할 수 있게 된다.

- **트랜스포머?** 딥러닝 아키텍처 중 하나로, 주로 시퀀스 데이터를 처리하는 데 사용되는 모델. 트랜스포머 아키텍처는 이전까지 주로 사용되던 순환 신경망(RNN)이나 컨볼루션 신경망(CNN)과는 다르게 셀프 어텐션(self-attention) 메커니즘을 활용하여 입력 시퀀스의 모든 요소 간의 관계를 동시에 학습
- **트랜스포머 디코더 블록?** 트랜스포머 아키텍처의 중요한 구성 요소 중 하나로, 시퀀스 데이터의 디코딩과 관련된 작업을 수행. 트랜스포머 디코더 블록은 입력된 정보를 기반으로 다음 시간 단계의 출력을 생성하는 역할.

작동 원리 :

1. 프롬프트와 상호 작용: 모델은 사용자가 주는 프롬프트에 따라 이미지 내에서 어떤 부분을 주목할지 결정

2. 업샘플링과 마스크 생성: 이미지의 크기에 맞게 업샘플링하여 각 픽셀에 대한 마스크 여부를 결정.

3. 다수의 마스크 후보와 신뢰도 평균: 여러 개의 마스크 후보가 있을 때, 각 픽셀의 신뢰도를 평균하여 하나의 마스크를 생성. 이로써 모호한 영역에서 신뢰도를 높이고 올바른 마스크를 얻는다.

4. 훈련: 생성된 마스크와 실제 마스크 간의 차이를 계산하여 손실을 구한다. 모델은 이 손실을 최소화하도록 훈련되며, 실제 이미지와 일치하는 정확한 마스크를 생성하도록 학습함.

5. 손실 계산: 포컬 손실과 다이스 손실을 결합하여 마스크 생성을 훈련. 이 손실 함수는 객체 경계를 예측하고 높은 신뢰도로 마스크를 생성하게 됨.