



Segment Anything

Samsung Software Developer Community

Korea Vision & Robotics


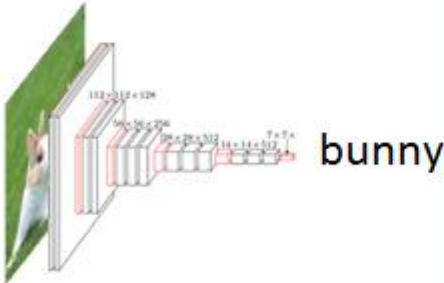
HoChan Jeong

2023.09.09

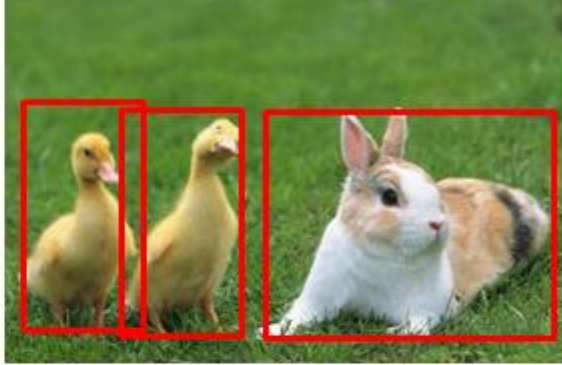
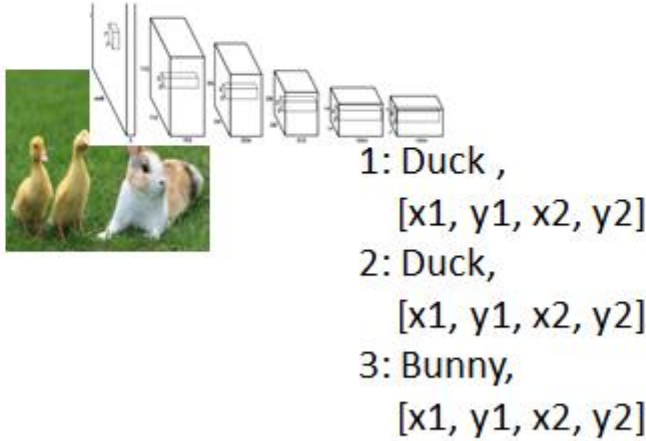
Contents

- 1. Background**
- 2. Segment Anything Task**
- 3. Segment anything Model**
- 4. Segment anything Data**

1. Background


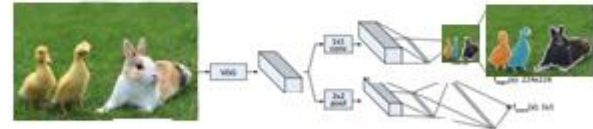



Classification

1: Duck ,
[x1, y1, x2, y2]
2: Duck,
[x1, y1, x2, y2]
3: Bunny,
[x1, y1, x2, y2]

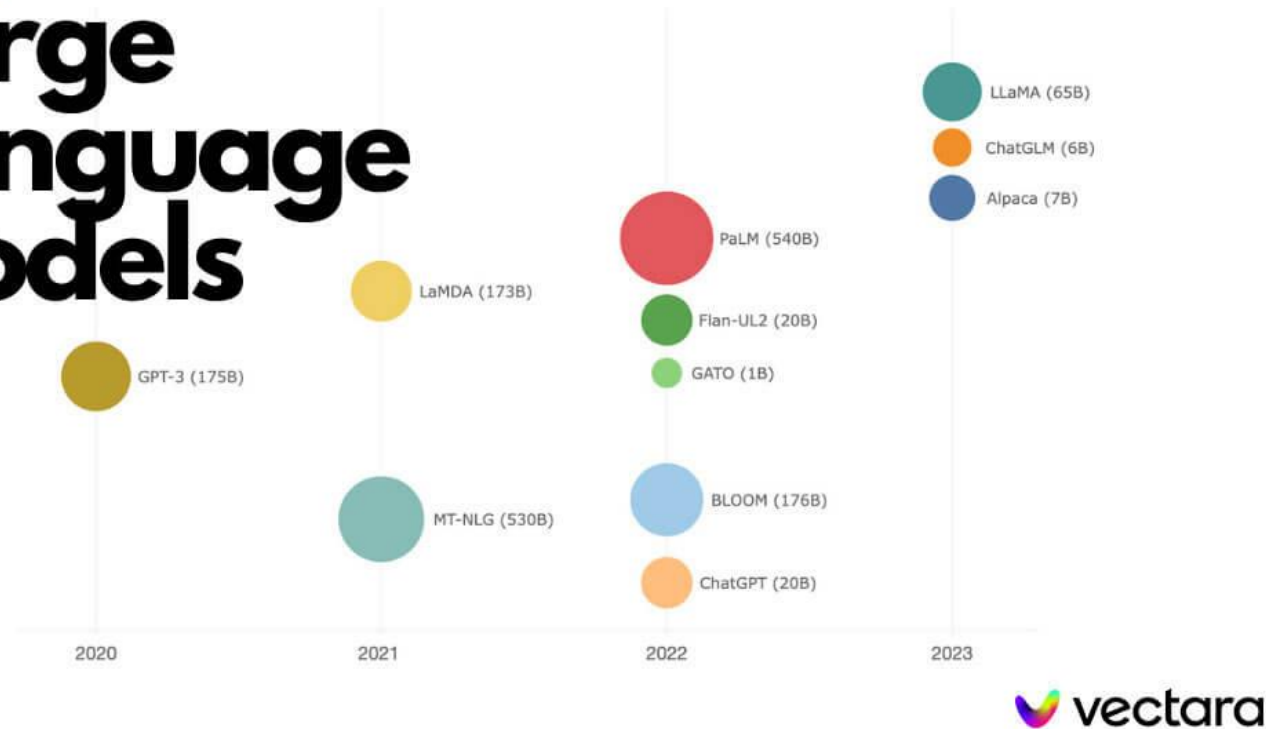
Object Detection

Segmentation

1. Background

Top Large Language Models

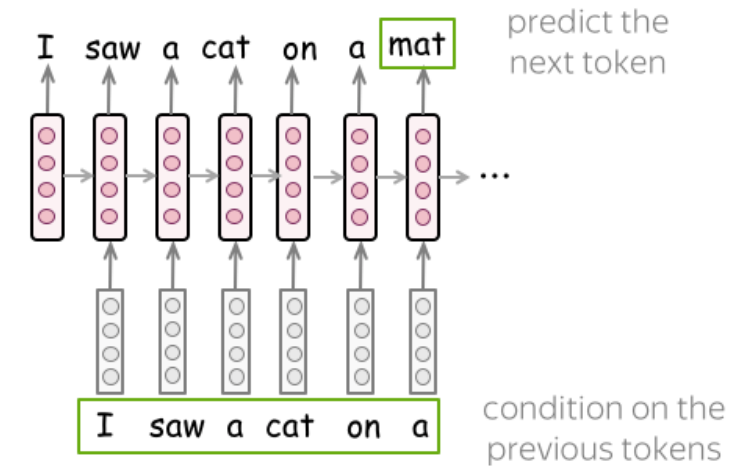


- 최근 Large Language Model이 등장

Ex) GPT, LLaMa, alpaca ...

- 이렇게 등장한 모델들은

‘next token prediction’으로 학습하여서
어떠한 task에서든지 좋은 성능을 보여주었다.



1. Background

CV 분야에서도 저런 ‘Model’을 만들 수 있을까?

- 논문 저자들은 3가지 과제에 집중하였다.

1. What **task** will enable zero-shot generalization?
2. What is the corresponding **model** architecture?
3. What **data** can power this task and model?

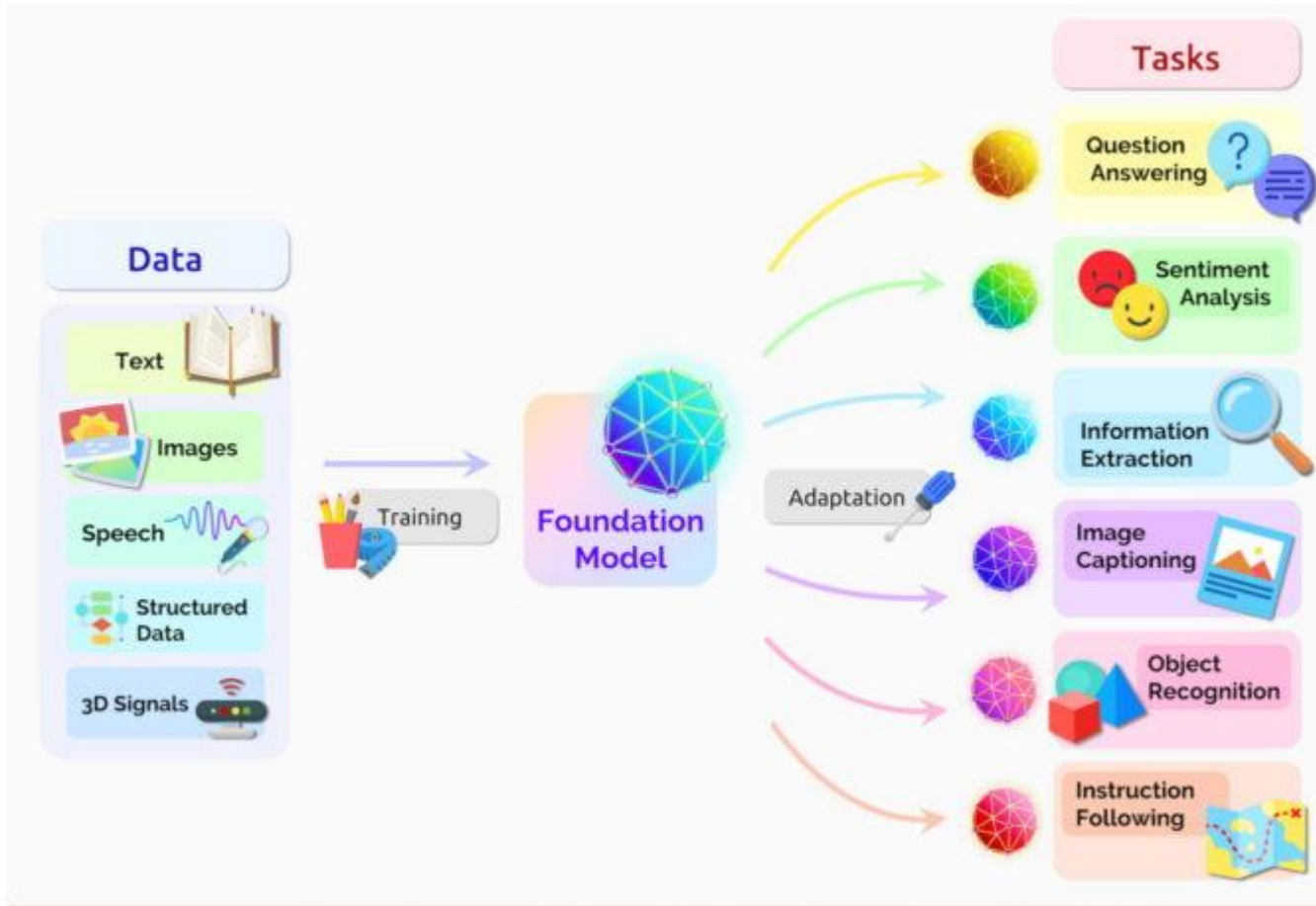


1. 어떠한 task
2. 어떠한 model 구조
3. 어떠한 dataset or data

이를 통해서 ‘Foundation Model’을 만들려고 하였다

1. Background

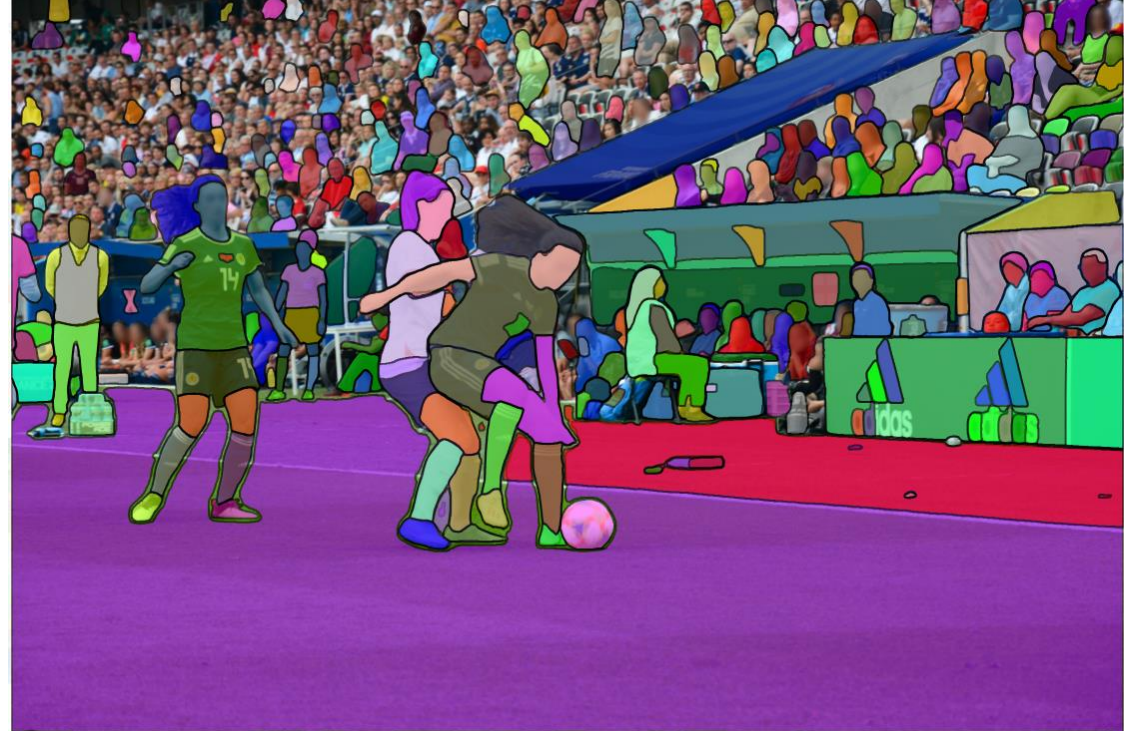
Foundation Model



- 쉽게 표현하면 하나의 모델로 여러 task
- 즉, **zero-shot or few shot** 성능이 좋아야 함

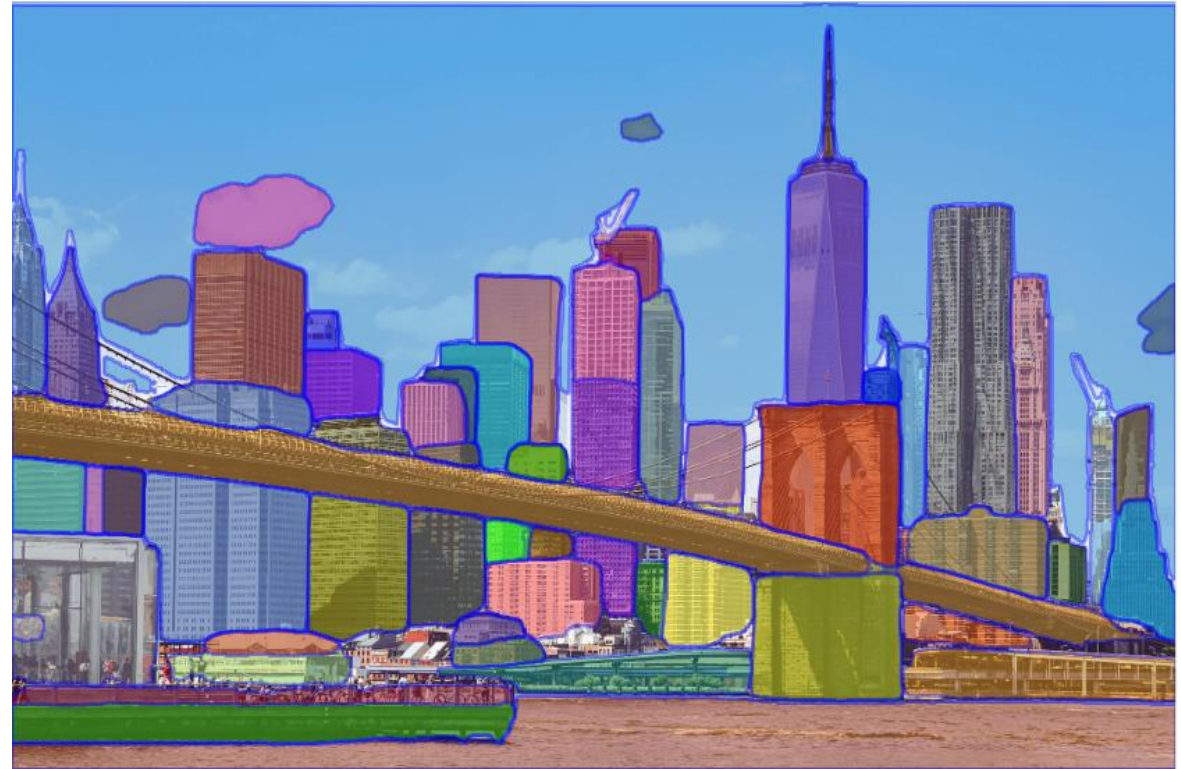
- 예를 들어서 GPT 모델과 같이
next token prediction만으로 학습한 모델이
한번도 학습 시키지 않은 다른 task에서도
괜찮은 성능을 보임

1. Background

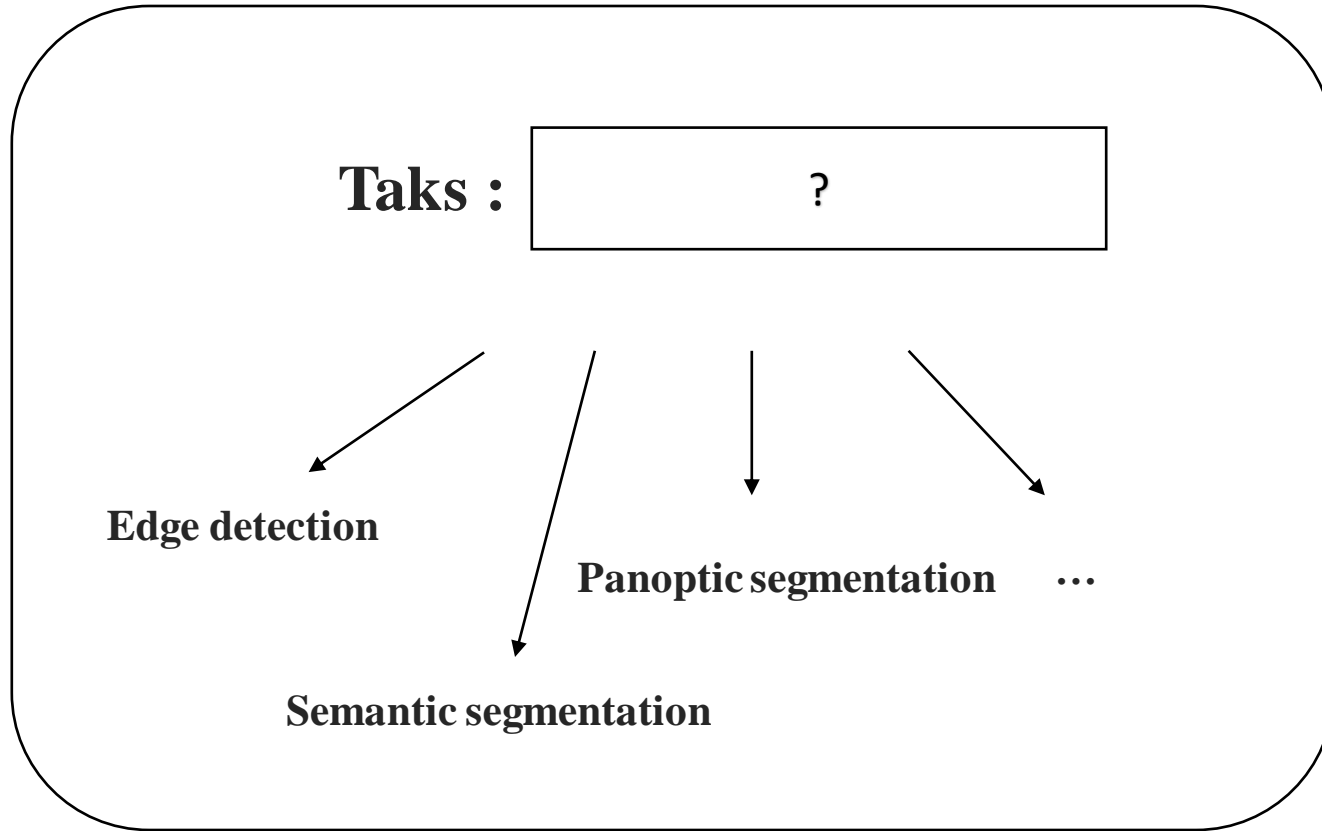


<https://segment-anything.com/dataset/index.html>

1. Background



2. Segment Anything Task



‘Next token prediction’ 처럼 다른 task에도
좋은 성능을 보여줄 ? Task가 없을까?



‘Promptable Segmentation’

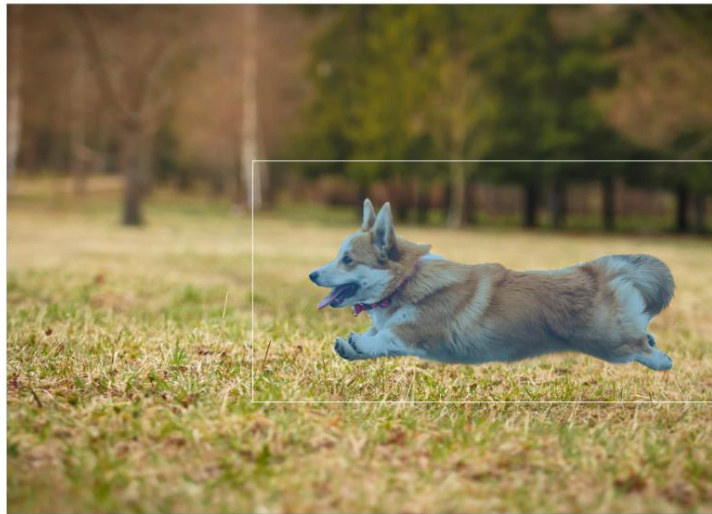
2. Segment Anything Task

‘Promptable Segmentation’

: 마스크를 생성하고자 하는 대상을 유연하게 prompt로 지정할 수 있는 task



point



box



text

2. Segment Anything Task

왜 이 task인가?

- ‘모호한’ 정보가 prompt로 들어와도 유효한 mask를 만들어 내기 위해서 즉, 이러한 결과를 위해서 학습을 하면 zero-shot 성능도 좋을 것이다
- 또한 prompt engineering을 통해서 여러 task에 적용이 가능할 것이다

학습방법은 간단하게 여러 task을 시퀀스하게 학습을 진행함

3. Segment Anything Model

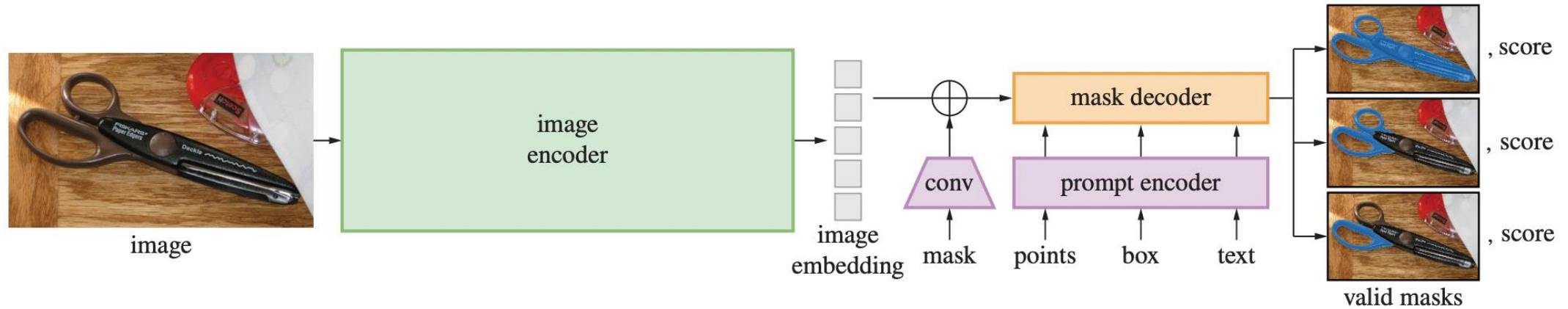
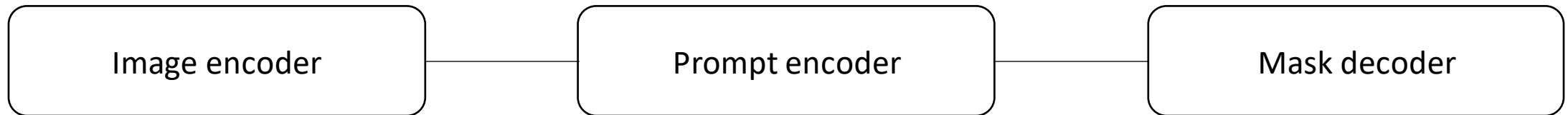


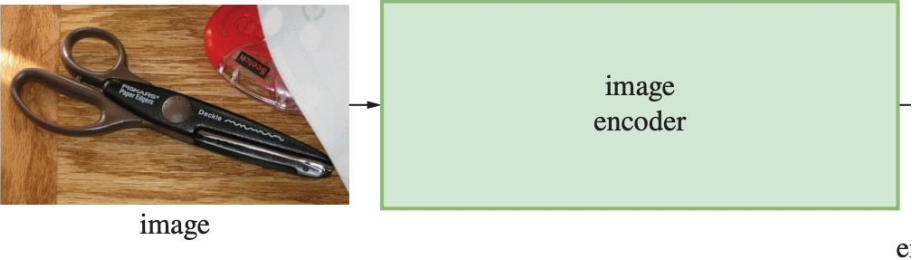
Figure 4: Segment Anything Model (SAM) overview. A heavyweight image encoder outputs an image embedding that can then be efficiently queried by a variety of input prompts to produce object masks at amortized real-time speed. For ambiguous prompts corresponding to more than one object, SAM can output multiple valid masks and associated confidence scores.

- 모델은 3가지의 요소로 구성되어 있음

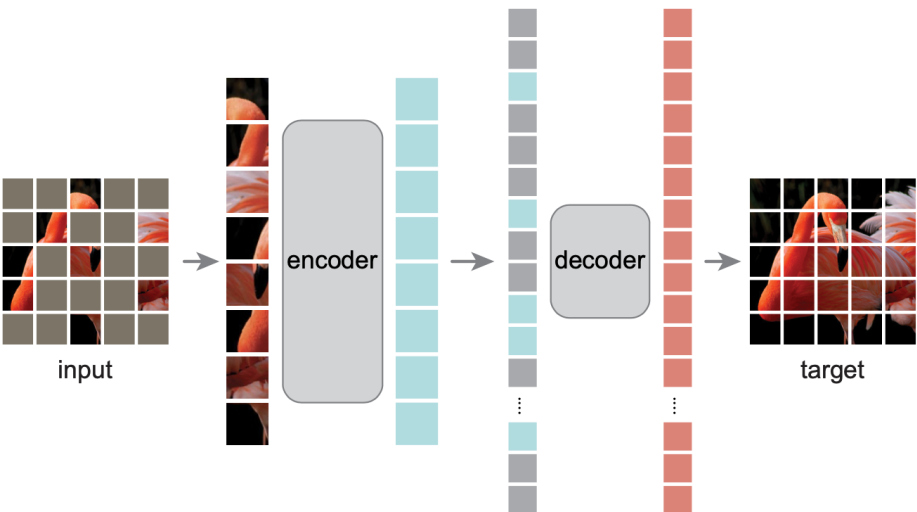


3. Segment Anything Model

Image encoder

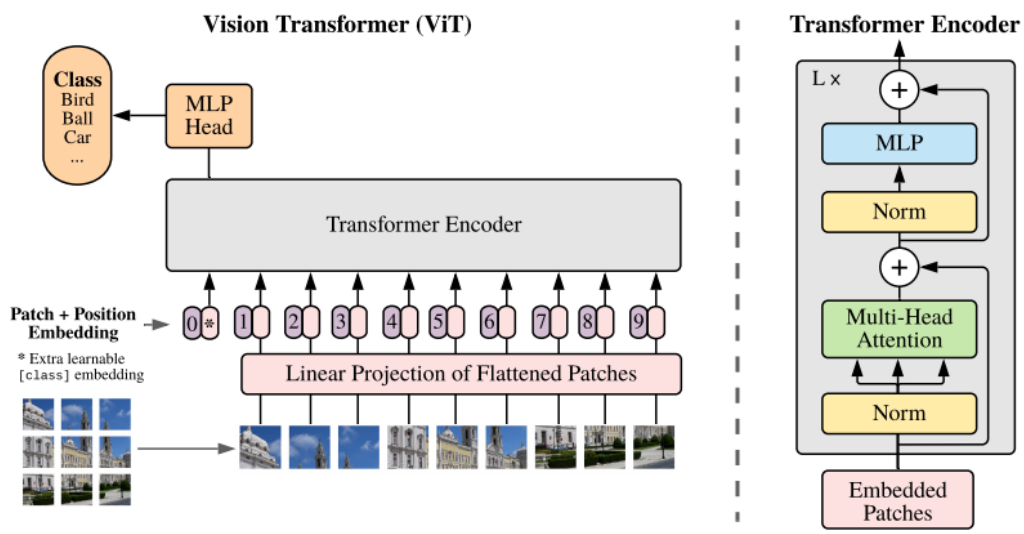


* MAE : Masked auto-encoder



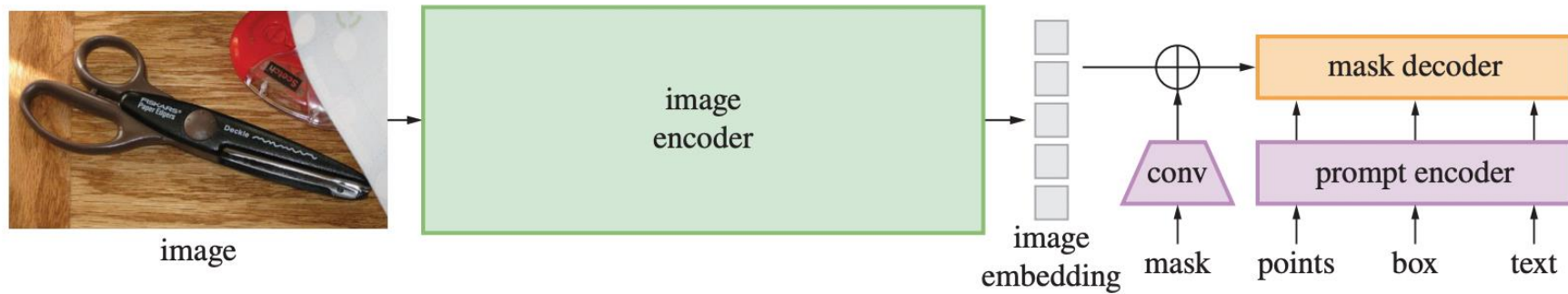
- image의 임베딩을 추출
- MAE 학습 방법을 적용한 pre-trained ViT를 활용
- 논문에서는 1024*2을 input으로 하여 64x64*256으로 임베딩

*ViT : Vision Transformer



3. Segment Anything Model

Prompt encoder



- Sparse prompt(점, 박스, text), Dense prompt(mask)로 구성
- point, box의 경우 **positional encoding**으로 표현
- Text의 경우 **CLIP**의 text encode를 가져와서 임베딩
- Mask의 경우 conv로 차원을 맞춰주고 임베딩된 이미지에 **pixel wise sum**

3. Segment Anything Model

Prompt encoder

- Point : **positional encoding** (point 위치) + **전경 or 배경**인지의 합
- Box : **Top-left Corner + Bottom-right Corner** -> 두개를 나타내는 positional encoding의 값으로 임베딩
- Text : CLIP의 encoder를 활용

‘위 모두다 256 채널로 임베딩’

- **Mask : 이미지와 공간적으로 대응되는 정보**

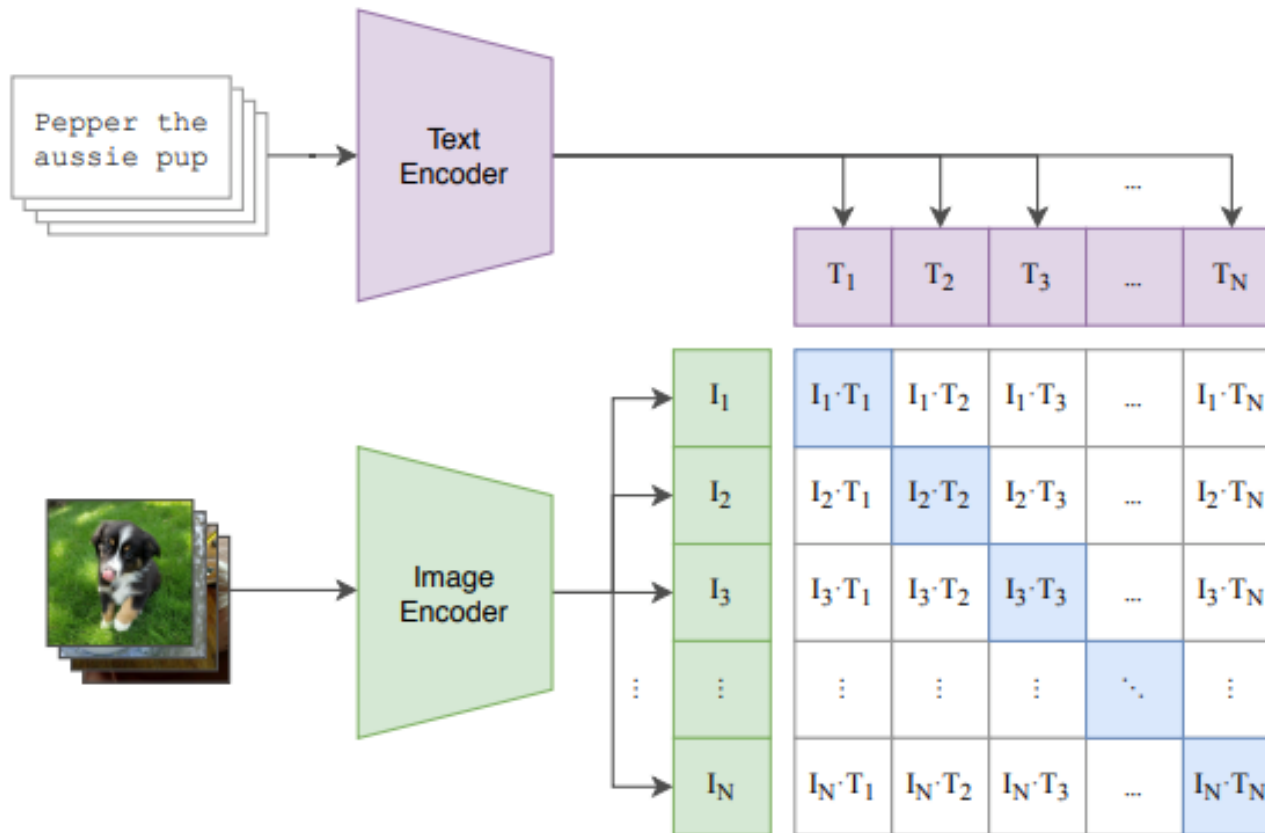
:

논문에서는 4배를 줄인 input을 넣고 이를 conv로 4배 더 줄인 다음 256 채널로 임베딩
-> 만약 mask가 입력으로 없다면 ‘no mask’를 나타내는 임베딩을 대신 사용

3. Segment Anything Model

Prompt encoder

(1) Contrastive pre-training



- Open AI에서 만든 모델로
Text – Image 관계를 벡터공간에 mapping
- 대용량의 데이터로 학습
- Zero-shot 성능이 좋다

3. Segment Anything Model

Mask decoder

- SAM 모델의 핵심!! (개인적 생각)

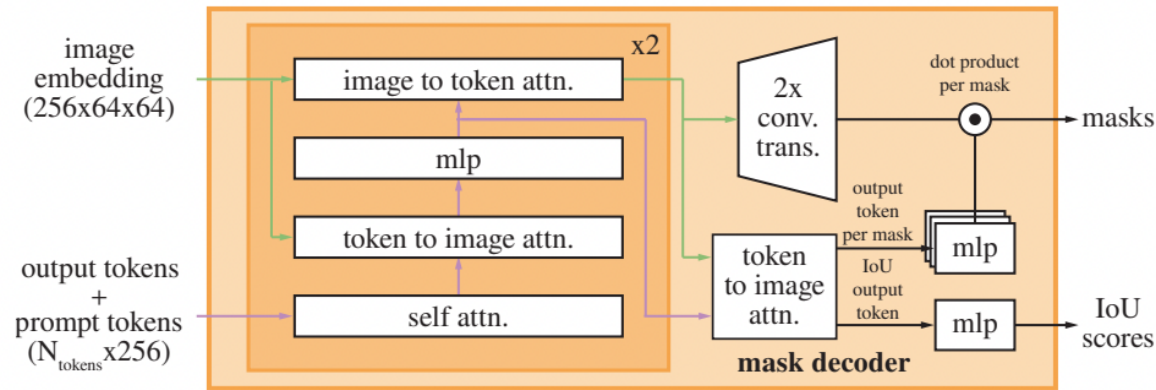


Figure 14: Details of the lightweight mask decoder. A two-layer decoder updates both the image embedding and prompt tokens via cross-attention. Then the image embedding is upscaled, from which the updated output tokens are used to dynamically predict masks. (Not illustrated for figure clarity: At every attention layer, positional encodings are added to the image embedding, and the entire original prompt token (including position encoding) is re-added to the token queries and keys.)

- input으로 image embedding과 prompt embedding을 받음
- 여기서 output token은 vit의 class token과 유사
- Decoder는 아래의 4가지 step으로 작동

1. Token에 대한 Self-attention
2. Token을 Query로 하여 Image Embedding에 Cross-Attention
3. Point-wise MLP로 각 Token을 업데이트
4. Image Embedding을 Query로 하여 Token에 Cross-Attention (Cross-Attention 과정에서 Image Embedding은 64x64의 256 차원 벡터)

- Transformer 기본 Decoder와 같은 구조
- 다른점은 input으로 token, image embedding이 사용
- 또한 cross-attention을 2개의 대상으로 진행
- 나머지 layer-normalization과 residual connection은 같음

3. Segment Anything Model

Mask decoder

- SAM 모델의 핵심!! (개인적 생각)

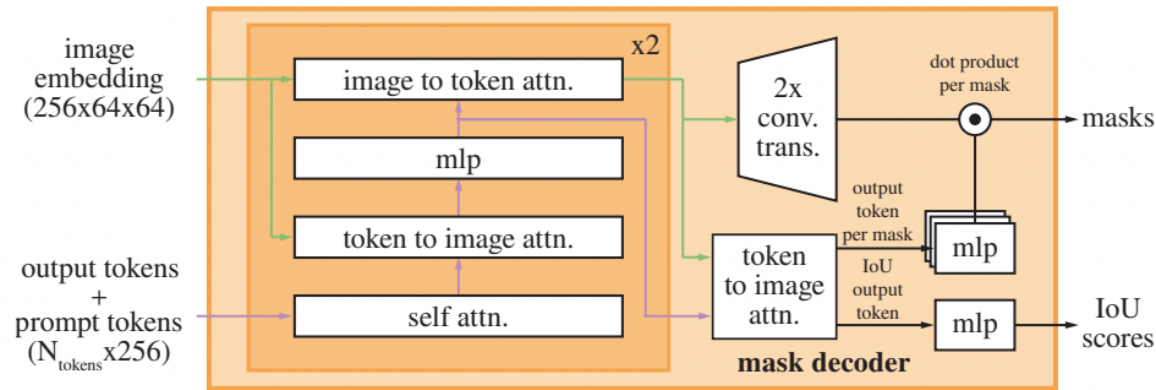
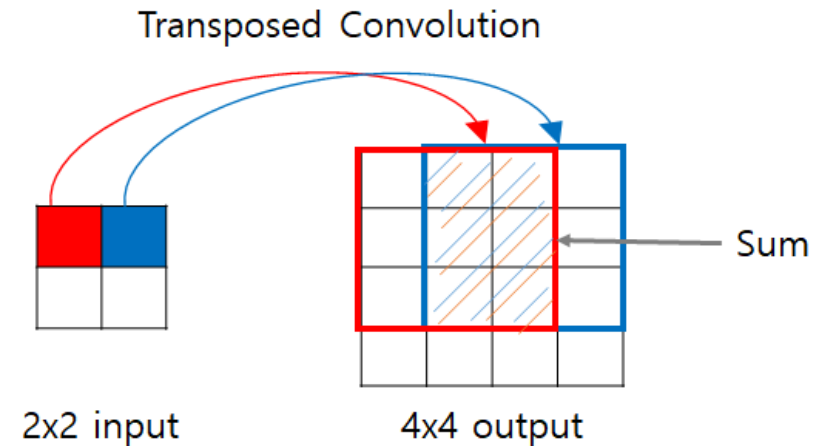


Figure 14: Details of the lightweight mask decoder. A two-layer decoder updates both the image embedding and prompt tokens via cross-attention. Then the image embedding is upscaled, from which the updated output tokens are used to dynamically predict masks. (Not illustrated for figure clarity: At every attention layer, positional encodings are added to the image embedding, and the entire original prompt token (including position encoding) is re-added to the token queries and keys.)

- Decoder Layer 통과 이후 나오는 output에 대해서 Transposed Convolutional Layer를 적용해서 Image Embedding을 4배 키움



- 그리고 따로 token을 다시 query로 해서 image embedding과 cross-attention해주고 Output token embedding에 3개의 MLP를 적용해서 위의 4배 키운 Image Embedding과 dot product

3. Segment Anything Model

Mask decoder

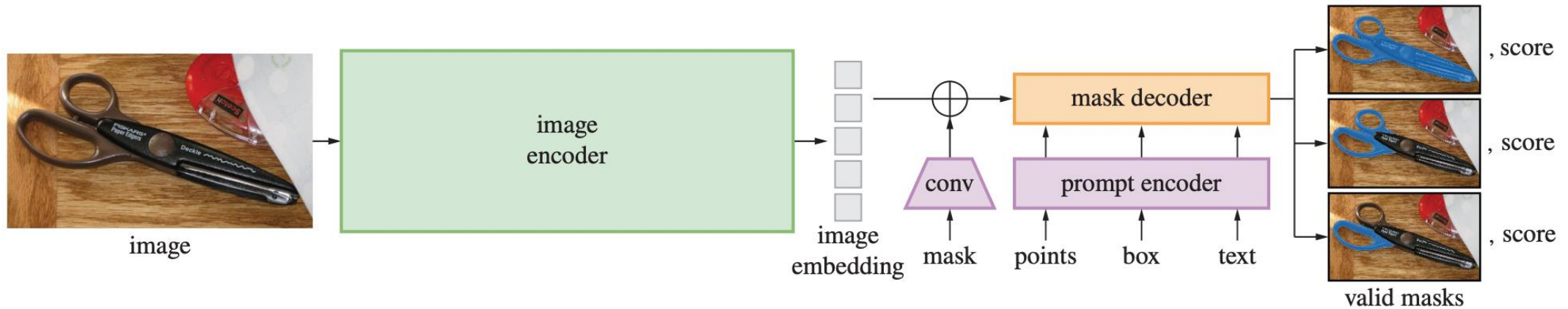


Figure 4: Segment Anything Model (SAM) overview. A heavyweight image encoder outputs an image embedding that can then be efficiently queried by a variety of input prompts to produce object masks at amortized real-time speed. For ambiguous prompts corresponding to more than one object, SAM can output multiple valid masks and associated confidence scores.

- decoder 통과이후 3개의 output과 각 IOU score가 나옴
- 이는 애매한 prompt에 대해서 유효한 성능을 내기 위해서 3개의 결과물의 평균을 최종 output으로 만들어줌
- 학습 과정에서는 가장 낮은 loss에 대해서만 역전파 수행

****참고 loss는 Dice + focal을 사용**

3. Segment Anything Model

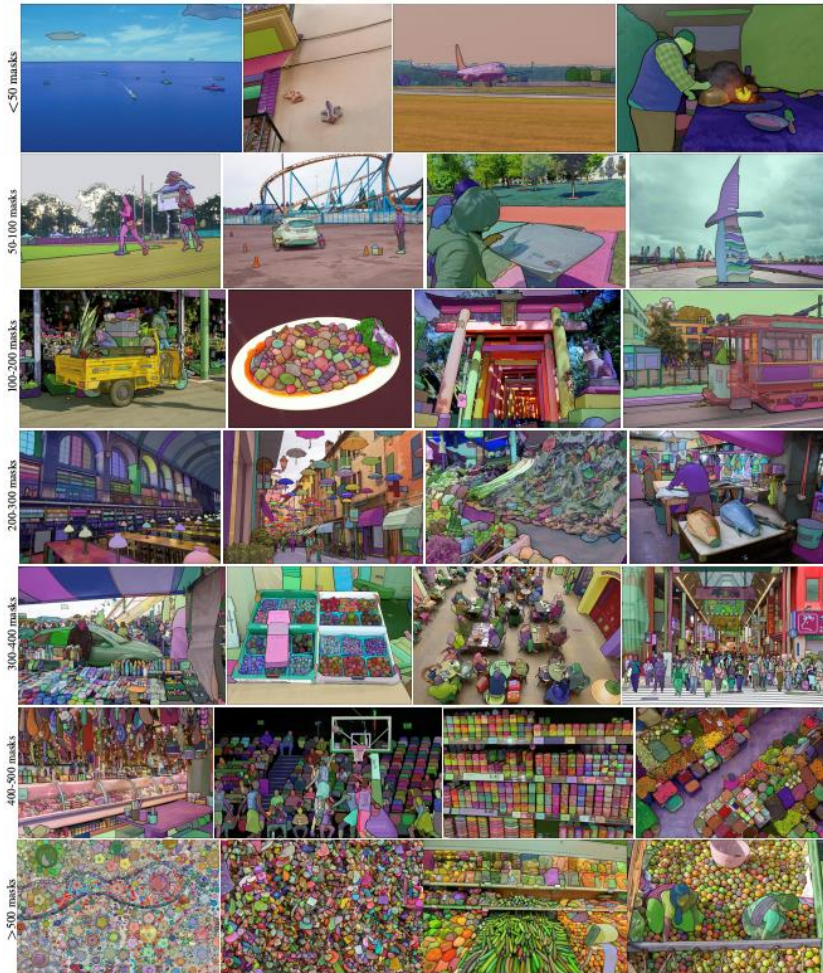


Figure 3: Each column shows 3 valid masks generated by SAM from a single ambiguous point prompt (green circle).

4. Segment Anything Data

<https://ai.meta.com/datasets/segment-anything/>

- 약 1100만장 Image에 11억개의 mask로 구성된 segmentaion dataset
- 기존에 공개된 dataset의 경우 수가 적적하지 않음
- 거대 자연어 모델 처럼 학습하기 위해서는 충분한 데이터 수집이 필수
- 이를 위해서 `Data engine`을 만들어 데이터를 수집



4. Segment Anything Data

- Data engine은 크게 3가지 단계로 구성되어 있음

1) Assisted Manual :

기존 공개된 Dataset으로 SAM을 학습하고 이를 활용해 추론한 결과를 사람이 수정하여 데이터 수집 (430만장)

2) Semi automatic :

앞에서 수집한 데이터로 SAM을 학습하고 이를 통해서 추론하면 사람은 빠진 결과만 채움 (1020만개 mask)

3) Fully automatic :

이제 수집한 데이터로 SAM을 학습하고 이를 활용해서 이미지 1100만장에 대해서 11억개 mask를 생성하게함

5. 참고 자료

데이터 셋

<https://segment-anything.com/dataset/index.html>

블로그

<https://blog.annotation-ai.com/segment-anything/>

논문

<https://arxiv.org/pdf/2304.02643.pdf>