Image Segmentation Using Text and Image Prompts

Abstract

기존의 이미지 분할 문제는 특정 task를 풀기 위해서 모델을 재학습 시키는 방법을 사용하였습니다. 하지만 모델을 재학습시키는 방법은 비용이 많이 들게 됩니다. 이에 본 논문에서는 CLIP을 backbone으로 하여 transformer기반의 decoder를 추가하여 테스트 시점에 prompt를 통해서 dense task를 풀수 있는 모델을 제안합니다. 해당 모델의 경우 한번의 학습으로 (1) 참조 표현 분할 (2) 제로샷 분할 (2) 원샷 분할을 모두 가능하게 합니다.

> 해당 논문의 경우 다양한 segmentation task를 처리하기 위해서 이미지-텍스트 정보를 잘 담고 있는 CLIP에 deocder를 확장하여 prompt 형식으로 segmentation을 제안하고있다.

Introduction

보이지 않는 데이터에 대해 일반화하는 것은 AI에게는 매우 어려운 일입니다. 게다가 iamge segmentation은 단순히 classifciation 보다 더욱 어려운 일입니다. 기존의 segmentation의 경우 실제 존재하는 카테고리에 대해서만 분할이 가능합니다. 하지만 이를 극복하기 위해서 아래의 3가지 방법이 제안됩니다.

(1) zero-shot : 처음 보는 객체도 기존의 객체와 연관지어 일반화한다.

(2) one-shot : 최소한의 예시만으로 원하는 객체를 분할한다.

(3) referring expression : 특정 객체에 대해 분할한다.

이를 동시에 해결하기 위해서 본 논문에서는 CLIPSeg 라는 모델을 제안합니다. 해당 모델은 CLIP을 backbone으로 활용하며 그 위에 조건부 decoder를 학습시킵니다. 해당 deocder는 CLIP의 활성화와 분할 마스크간의 관계를 학습하게 됩니다. → (조금더 이해가 된 만큼 설명해보면) *CLIP 모델의 경우 이미지와 텍스트의 관계를 잘 나타내고 있습니다. 그*

래서 segmantation mask에서 나온 정보와 해당 임베딩 행렬간의 복잡한 관계를 모델이 학습하게 된다면 결국 CLIP의 임베딩 벡터의 어느 위치가 활성화 되어야 이미지에서 전경을 분리할 수 있을지를 decoder가 학습하게 됩니다.

모델은 프롬포트와 일치하는 전경을 배경과 구분하는 일반적인 이진 예측 설정을 사용하게 됩니다.

Contributions

해당 모델은 경량화된 transformer decoder를 사용함으로 써 dense task를 수행하며, 특히 iamge, text 기반의 prompt를 통해서 segmentation 대상을 지정할 수 있다는 장점을 가지고 있습니다.

CLIPSeg Method

본 논문에서는 CLIP은 가중치 업데이트를 하지 않고 오로지 decoder만을 학습한다고 합니다. 그래서 본 논문에서는 decoder만 학습시 편향이 되어 CLIP의 일반화 성능에 악영향을 끼치지 말아야한다고 주장하고 있습니다.

Decoder Architecture

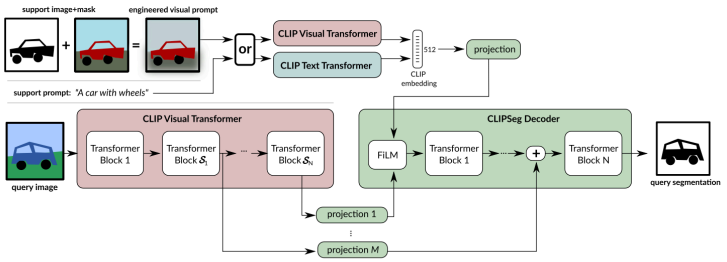


Figure 2: Architecture of CLIPSeg: We extend a frozen CLIP model (red and blue) with a transformer that segments the query image based on either a support image or a support prompt. N CLIP activations are extracted after blocks defined by $\mathcal{S}$. The segmentation transformer and the projections (both green) are trained on PhraseCut or PhraseCut+.

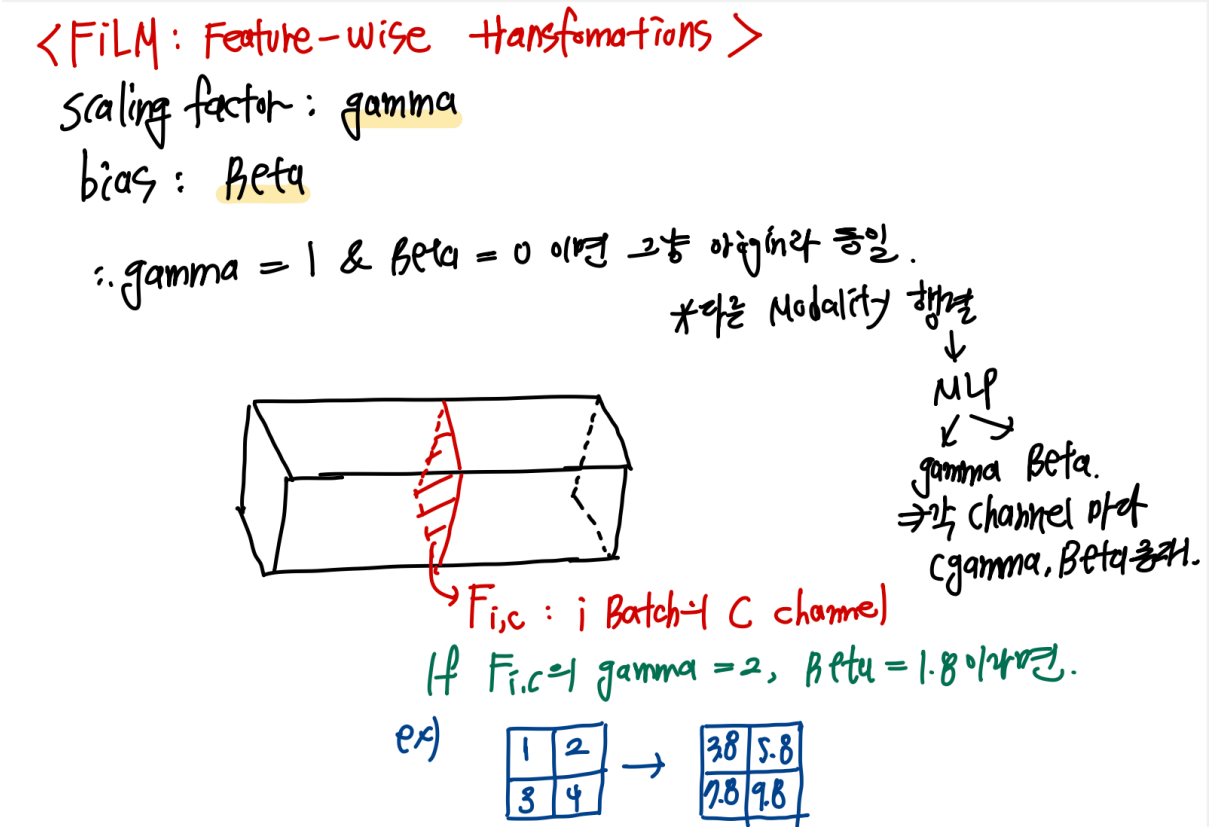
본 논문에서는 UNet의 skip connection과 Transformer에서 영감을 받아 decoder를 구현했다고 합니다. 이미지를 보면 크게 2가지 stream으로 구분해볼 수 있습니다.

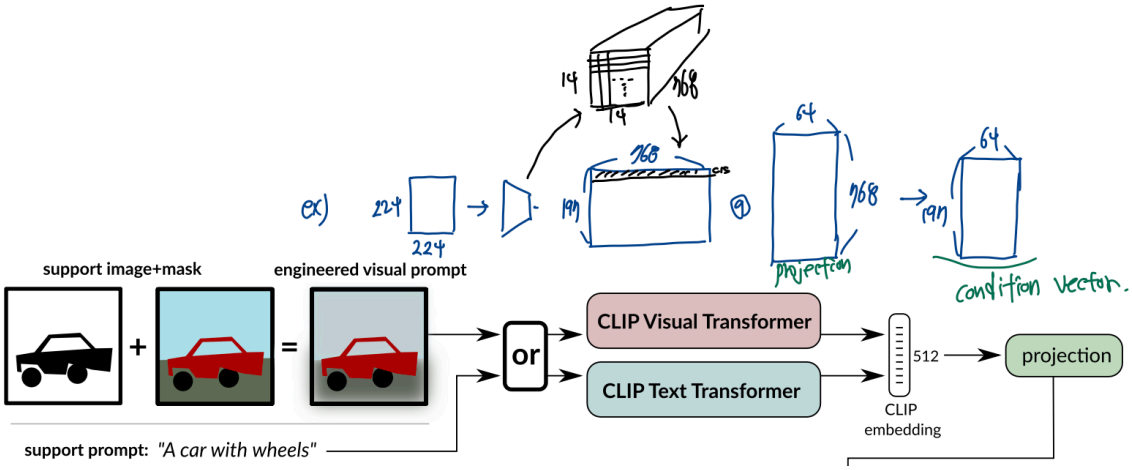
(1) 위쪽 stream : 분할할 객체를 지정하기 위해 이미지 프롬프트 또는 텍스트 프롬프트를 입력받습니다. 만약 이미지 프롬프트가 들어오면 CLIP의 이미지 인코더(ViT-B/16)를, 자연어 프롬프트가 들어오면 CLIP의 텍스트 인코더를 사용합니다. ViT 기반 CLIP의 경우, 입력

프롬프트는 (14×14 + 1, 786) 크기의 출력(여기서 1은 CLS 토큰)을 생성하며, 이를 디코더의 차원인 64로 맞추기 위해 선형 투영을 수행하여 최종적으로 (197, 64) 크기의 support matrix를 얻게 됩니다. 그리고 해당 행렬은 결국 우리가 어떤걸 분할할것인가?를 나타내주게 됩니다.

(2) 아래 stream : 실제 분할할 쿼리 이미지를 CLIP의 이미지 인코더에 입력하여, 특정 레이어에서 추출한 feature map을 디코더와 스킵 연결로 결합합니다. 논문에서는 스킵 연결로 사용한 feature map의 개수, 즉 [3, 7, 9] 레이어에서 추출한 정보의 수가 디코더의 transformer 블록 수(여기서는 3개)를 결정한다고 언급합니다.

그리고 (1)에서 얻은 (197, 64) support matrix를 작은 신경망을 통해 FiLM에 입력하여, 각 채널별로 64개의 (γ, β) 쌍을 산출합니다. 이 조건 정보가 디코더에 주입되어 transformer 블록을 거치며 스킵 연결과 결합되어 학습되고, 최종적으로 선형 투영을 통해 (64, 1) 크기의 projection 행렬로 변환되어 (197, 1) 행렬을 생성합니다. 이 (197, 1) 행렬은 각 패치(또는 토큰)에 대해 객체의 전경/배경 여부를 확률적으로 나타내며, 이를 통해 최종 이진 세그멘테이션 결과가 도출됩니다.





이렇게 CLIPSeg는 전체 학습 가능한 파라미터 수가 D=64 기준으로 1,122,305개에 불과하다고 주장합니다.

Image-Text Interpolation

추가로 데이터를 증강하기 위해서 이미지-텍스트간 보간법을 사용하였습니다. 기존 데이터들은 조건 벡터로 이미지 혹은 텍스트 하나만 넣어서 학습을 하였습니다. 하지만 각 임베딩 벡터에 대해서 본 논문에서는 $x_i = a \cdot s_i + (1 - a) \cdot t_i$ 다음과 같은 수식을 통해서 새로운 벡터를 생성합니다. (convex형식) 이를만일 a = 1 인경우 이미지만 고려, a = 0인 경우 텍스트만 고려하게 됩니다. 이를 통해서 모델이 보다 객체에 대한 개념을 강건하게 학습한다고 주장합니다.

PharseCut + Visual prompt(PC+)

PhraseCut 데이터셋은 원래 34만 개 이상의 문구와 이에 대응하는 이미지 분할 정보만 포함하고 있어, 시각적 지원 이미지는 제공되지 않습니다. 이를 보완하기 위해 데이터셋을 확장하여, 동일한 프롬프트에 해당하는 여러 이미지 중 무작위로 하나를 선택하는 방식으로 Visual Support Sample을 추가하고, 프롬프트와 일치하는 객체가 없는 부정 샘플(Negative Sample)도 도입합니다. 단, 특정 프롬프트에 대응하는 이미지가 단 한 개뿐인 경우에는 해당 이미지가 모델에 편향을 줄 수 있으므로, 이때는 이미지 대신 텍스트 프롬프트만 사용하여 데이터의 다양성과 일반화를 도모합니다. 또한, 문구 데이터는 CLIP 저자들이 제안한 고정된 접두어를 활용하여 무작위 증강되며, 이를 통해 확장된 PhraseCut+ 데이터셋은 이미지와 텍스트 정보를 모두 활용하는 조건 기반 세그멘테이션 모델 학습에 유용하게 사용됩니다.

Visual Prompt Engineering

전통적인 CNN 기반의 모델들의 경우 조건 벡터를 통해서 mask pooling을 하게 되면 원하는 객체를 제외하고 모든 영역에 대해 0으로 만들고 객체를 분할 할 수 있었습니다. 하지만 CLIP과 같은 transformer 기반 모델의 경우 영역 별 정보를 담기 보단 단순히 패치별 정보를 압축하고 CLS 토큰에도 정보가 들어가게 되어 단순히 mask pooling을 적용하지 못하게 됩니다.

CLIPSeg에서는 단순히 마스크를 토큰에 적용하는 CLIP-Based Masking 방식보다, 이미지와 마스크를 결합하여 새로운 이미지(visual prompt)를 생성하는 Visual Prompt Engineering 방식을 채택하고 있습니다. 이 방식은 배경 밝기 감소, 블러 처리, 크롭 기법을 결합해 타겟 객체 정보를 효과적으로 강조하는 전략입니다.

Conclusion and Limitation

CLIPSeg는 새로운 데이터에 대한 재학습 없이, 추론 시 텍스트나 이미지 프롬프트를 통해 분할 대상을 지정할 수 있는 이미지 분할 접근법입니다. 이 방법은 novel visual prompt engineering을 도입하여 참조 표현, 제로샷, 원샷 세그멘테이션 작업에서 경쟁력 있는 성능을 보였습니다. 하지만 본 연구의 한계로는 평가가 소수의 벤치마크에 국한되어 있다고 제안합니다.