실전딥러닝 중간레포트

2022255016 이재영

2021256837 이문현

2021257449 양재우

2022208388 지준영

2021123173 송우석

2023158605 김하윤

**문제정의:** 엉성한 세그멘테이션 라벨에 강건한 음식 이미지 세그멘테이션

주어진 데이터셋을 기반으로 이미지 속에서 음식 영역을 계산하는 세그멘테이션을 수행하고자 합니다. 다만, 주어진 데이터셋에 포함된 세그멘테이션은 다음 그림과 같이 상당히 엉성한 부분이 많다는 점을 발견했습니다.

테이블웨어, 음식, 예술이(가) 표시된 사진

자동 생성된 설명

그림 1 엉성한 세그멘테이션 라벨 예시

이러한 세그멘테이션 라벨을 이용하여 세그멘테이션 모델을 학습하게 되면, 나름 괜찮은 수준의 세그멘테이션 예측 모델을 만들 수 있을 지 모르지만, 정확한 세그멘테이션을 예측하는 모델을 만들기는 힘들다고 판단했습니다. 이에 따라, 저희 팀은 다음과 같은 문제를 해결하는 모델을 구성해보고자 합니다.

* **문제 1:** 음식이 담긴 이미지에서 음식 부분을 세그멘테이션하는 모델 개발
* **문제 2:** 엉성한 세그멘테이션 라벨에도 강건한 세그멘테이션 모델 개발

**계획:** DINOv2 + SAM

위에서 정의한 첫 번째 문제의 경우, 단순한 세그멘테이션 모델에 불과하지만, 저희는 두 번째 문제 해결에 좀 더 초점을 맞추어 세그멘테이션 모델 구현 계획을 수립했습니다. 저희 팀이 계획한 세그멘테이션 모델의 동작 과정은 다음과 같습니다.

1. 어떤 입력 이미지에 대해, DINOv2라는 모델을 이용하여 엉성한(rough) 세그멘테이션을 획득합니다.
2. DINOv2로 얻은 엉성한 세그멘테이션으로부터 점을 한개 이상 추출합니다.
3. SAM(Segment Anything) 모델에 입력 이미지와 2번 과정에서 얻은 점(일종의 prompt입니다)을 입력하여, 최종 세그멘테이션을 얻습니다.

이때, DINOv2는 대용량의 이미지 데이터를 이용하여 기학습된 트랜스포머 구조의 특징추출용(feature extraction) 네트워크입니다. 이미지 하나를 입력으로 받으며, 해당 이미지의 특징 벡터를 추출해줍니다. 이때, 특징벡터는 이미지 당 1개(글로벌 특징 벡터)만 추출할 수도 있고, CNN 출력처럼 패치단위(로컬 특징 벡터)로 추출할 수도 있습니다. SAM은 11억개의 세그멘테이션 라벨링 데이터로 학습된, 상호작용가능한 세그멘테이션 모델(interactive segmentation model)입니다. 이미지 한 장과 프롬프트(점 또는 바운딩박스)를 입력으로 받아서, 해당 프롬프트에 해당하는 객체를 세그멘테이션해서 반환해줍니다. 이때, 저희가 파인튜닝할 모델은 DINOv2 이며, 이 모델은 정확한 음식 세그멘테이션을 얻기 위함이 아니라, 엉성한 세그멘테이션을 얻기 위해 사용될 것입니다. SAM은 파인튜닝하지 않을 것입니다.

이 방식의 장점은, 다음과 같습니다.

* 저희가 DINOv2를 파인튜닝할때, 고도의 파인튜닝이 필요가 없습니다. DINOv2를 파인튜닝하는 목적은 엉성한 세그멘테이션을 얻기 위함일 뿐, 정확한 세그멘테이션을 얻기 위함이 아닙니다. 따라서, 파인튜닝 난이도가 쉬워집니다.
* 11억 쌍의 세그멘테이션 데이터로 학습된 SAM(Segment Anything)의 기능을 온전히 활용할 수 있습니다. DINOv2가 엉성한 세그멘테이션을 계산해주면, 좋은 퀄리티의 세그멘테이션은 SAM이 대신 추출해줄 것입니다. 우리가 파인튜닝할 DINOv2는 SAM을 위한 힌트(point prompt)만 제공해주면 됩니다.

다음 그림은 저희의 모델링 계획을 표시한 것입니다.

