COSE474-2024F: Final Project Proposal
''CLIP와 LLaMA를 활용하여 객체 이미지를 분류하고 설명하기"

2018150372 조현민

1. Introduction

평소에 역사에 관심이 많기 때문에, 어떤 특정 역사적 소재를 제시했을 때, 이에 대한 설명을 제공받는 것에 대해 관심이 있었음. 예를 들어, '앵발리드'라는 어휘 혹은 사진을 제시했을 때, 이에 해당하는 설명을 제공받는 것임. 현재의 AI모델은 당연하게도 객체를 감지하지만, 해당 객체의 겉모습과 이에 대한 특징 이상의 설명을 제공하지 못하는 경우가 대다수 임.

이 프로젝트에서는 AI에게 특정 연도 혹은 인물에 관련된 역사적 소재에 대해 학습시키고, 이에 대한 맥락적 설명을 주지 시킬 것임. 이를 위해 객체 감지에 유용한 CLIP와 이를 보완해줄 LLaMA를 결합할 것임. 이를 통해 이미지 속의 소재에 대해 역사적, 문화적 설명을 제공할 수 있는 모델을 설계하고자 함.

2. Problem definition & chanllenges

특정 이미지가 무엇인지를 파악하고, 이를 클래스로 나누어서 분류하는 것에서 나아가서, 해당 이미지가 어떤 문화, 역사적 의미를 가지는지에 대한 정보를 제공하는 모델을 만들고자 함.

이를 위해 해야 할 것은 1) 객체와 맞는 정확하고 세부적인 정보를 제공하는 모델 설계 2) 훈련 시킬 소재에 대한 다양한 이미지 수집과, 이에 대한 역사적, 문화적 정보를 모으고 정리하기 3) CLIP, LLaMA에 대해 이해하고, 이를 활용하여 간단한 모델을 설계하기를 수행할 수 있어야 함.

3. Related Works

위의 목표를 구현하기 위해서 사용할 모델은 1) CLIP와 2) LLaMA가 있음. CLIP는 시각적 정보와 텍스트 정보를 조합하여 제로샷 학습, 크로스 모달 학습을 가능하게 만듦. 추가로 LLaMA는 맥락적 지식을 바탕으로 유연한 텍스트를 작성할 수 있도록 만들어 줌. .

4. Datasets

일단 기존의 객체 감지 데이터셋인 COCO 혹은 OpenImages을 사용하고, 외부 소스(위키피디아나 다른 외부 정보들)을 추가하여 이를 보완하고자 함. 이를 통해 객체 이미지와 역사적, 문화적 설명을 연결한 새로운 데이터셋을 만들고자 함. 현재 벤치 마크 데이터셋은 이 프로젝트에 필요한 데이터를 제공하고 있지 않기에 커스텀 데이터셋을 구축하고자 함.

먼저 유물, 문화적 상징물 혹은 역사적 장소에 관한 다양한 객체 이미지를 모으고, 해당 이미지들에 대한 문화적, 역사적 설명들 또한 수집할 것임. 물론 하나하나의 데이터에 들어가는 시간과 노력이 많이 소비될 것으로 예상되므로, 현실적으로 8-10개 정도의 객체를 목표로 삼을 것임.

5. State-of-the-art methods and baselines

기준 모델로 사용할 것은 CLIP, LLaMA임. CLIP의 제로샷 환경에서의 객체 인식 능력을 사용하여 객체 감지를 수행함 추가로 LLaMA의 텍스트 생성 능력을 바탕으로 맥락 설명 생성을 구축함.

성능 지표의 경우 생성된 텍스트의 성공 여부에 대해 BLEU, ROUGE와 같은 표준 평가 지표를 사용할 것임. 객체 인식의 경우, 직접 보고 정확도를 평가할 것임. 생성된 설명 역시 직접 인지하여 설명의 질을 평가할 수 있음.

6. Schedule & Roles

4주에 걸쳐 프로젝트를 진행할 예정임. 1주차에는 데이터셋의 범위를 결정하고(카테고리 정하기) 데이터를 수집할 것임. 이를 바탕으로 소규모의 커스텀 데이터셋을 제작할 것임. 2주차에는 CLIP, LLaMa 모델을 활용하여 객체 인식 및 설명 생성이 가능하도록 조작할 것임. 추가로 간단한 데이터셋을 이용하여 모델을 학습시키고 성능을 테스트 해볼 예정임. 3주차에는 모델을 개선하고 오류를 수정하는 과정을 거침. 마지막 4주차에는 모델을 최종 테스트, 평가해보고 결과 분석 및 보고서를 작성할 예정임.