방향 인지 제로샷 참조 이미지 분할 모델 연구

최린[1], 박은일[2]
[1] 성균관대학교 인공지능융합학과 석사과정
[2] 성균관대학교 인공지능융합학과 교수

lynnchoi@skku.edu, eunilpark@skku.edu

Orientation Aware Zero-shot Referring Image Segmentation

Lynn Choi[1], Eunil Park[2]
[1]Dept. of Applied Artificial Intelligence, Sungkyunkwan University
[2]Dept. of Applied Artificial Intelligence, Sungkyunkwan University

요 약

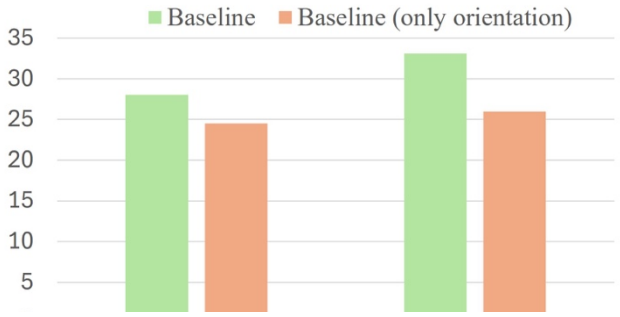
본 연구는 제로샷 참조 이미지 분할(Zero-Shot Referring Image Segmentation)에서 방향 정보를 효과적으로 처리하기 위한 새로운 접근법을 제시한다. 기존 연구는 CLIP 을 기반으로 이미지와 텍스트의 지역과 전역 특성을 융합하여 활용했으나, 참조 문장에 포함된 방향 정보를 정밀하게 반영하지 못하는 한계를 보였다. 이를 보완하기 위해 본 연구에서는 두 가지 개선 전략을 도입하였다. 첫째, Segment Anything Model(SAM)을 활용하여 객체 단위에서 더욱 정밀한 마스크를 생성함으로써 분할 품질을 높였다. 둘째, 바운딩 박스 좌표를 기반으로 한 방향 인지 알고리즘을 추가하여 텍스트 내 방향 정보를 명시적으로 처리하도록 하였다. 제안한 모델을 RefCOCOg 데이터셋을 활용해 실험했을 때, 기존 방법론 대비 mIOU 가 5.80 증가하는 등 개선된 성능을 확인할 수 있었다. 이러한 실험 결과는 단순한 구조 개선만으로도 제로샷 참조 이미지 분할에서 방향 인지 능력을 효과적으로 보완할 수 있음을 보여주며, 향후 다양한 시각-언어 응용 분야로의 확장 가능성을 시사한다.

1. 서론

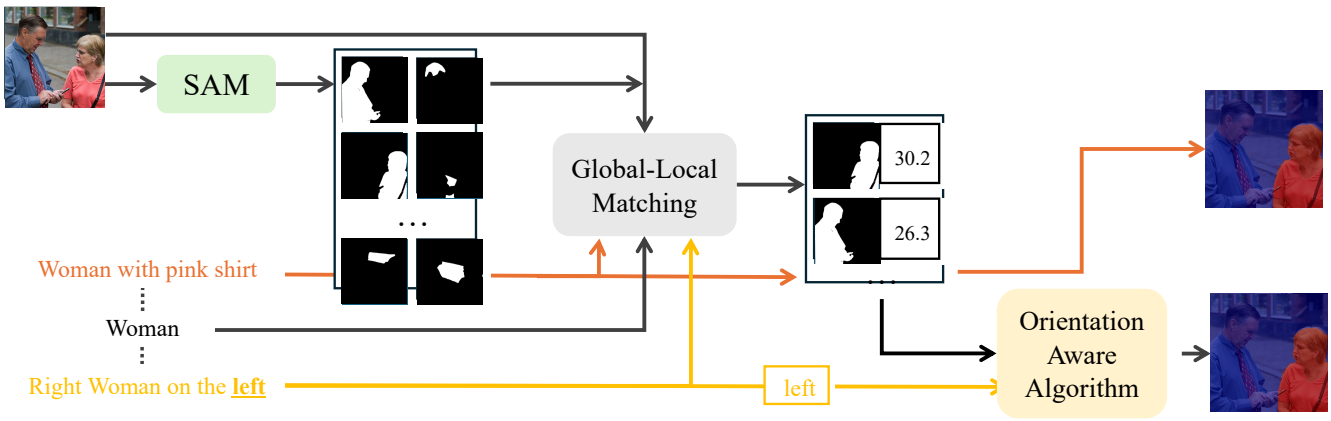
최근 딥러닝의 발전에 따라 컴퓨터 비전(CV)과 자연어 처리(NLP) 분야에서 활발한 연구가 진행되고 있다. 특히 대규모 데이터셋을 기반으로 학습하여, 이미지와 텍스트를 모두 다루는 시각 언어 모델(Vision Language Model)이 등장하면서 단일 모달리티 모델의 성능을 능가하고 있다. 이러한 발전을 바탕으로 텍스트 설명을 기반으로 이미지를 인식하고 분할하는 지시 이미지 분할(Referring Image Segmentation) 과제가 제안되었으며, 관련 연구가 활발히 수행되고 있다[1, 2, 3, 4]. 그러나 해당 과제를 지도학습(supervised learning)으로 해결하기 위해서는 높은 주석(Annotation) 비용이 요구된다.

이 문제를 완화하기 위해 Yu et al.[5]은 Zero-Shot Referring Image Segmentation 을 위한 Global-Local 모델을 제안하였다. 이는 대규모 데이터셋으로 사전학습된 CLIP[6]을 활용하여, 추가 학습 없이 참조 이미지 분할을 수행하는 방식으로 동작한다. 특히 이들은 목표 객체와 객체 간의 관계를 한번에 포착하기 위해 CLIP 으로 이미지와 텍스트의 지역(local) 특성과 전역(global) 특성을 인코딩하고, 이를 융합하는 모델을 설계하였다.

하지만 Global-Local 모델이 뛰어난 일반화 성능과 정확도를 보였음에도 불구하고, 객체를 인스턴스 수준에서 정밀하게 식별하는 데는 한계가 있는 것을 발견하였다. 특히 참조 텍스트에 방향 정보가 포함된 경우 이를 충분히 반영하지 못했으며, 방향 단어가 포함된 데이터를 추출하여 모델을 실험했을 때 현저히 낮은 IoU 점수는 이러한 한계를 명확히 보여준다 (그림 1).



(그림 1) 전체 데이터셋과 방향정보가 포함된 데이터 서브셋에서의 베이스라인 성능 평가 결과



(그림 2) 전체 모델 구조

본 연구에서는 이러한 문제를 해결하기 위해 방향 정보 이해를 보완하고 객체 단위 참조 이미지 분할 성능을 강화한 모델을 제안한다. SAM(Segment Anything Model)[7]을 분할 마스크 생성 단계에 적용하여 복잡한 장면에서도 제로샷 분할 성능을 확보하고, 위치 이해 알고리즘을 추가하여 텍스트 내 방향 정보를 명시적으로 처리한다. 이를 통해 선행 모델 대비 1.64 배(5.80) 향상된 mIoU 성능을 달성하였다.

2. 방향 인지 제로샷 참조 이미지 분할 방법론

본 연구는 Global-Local 모델을 기반으로 마스크 생성 모델 강화와 방향 인지 알고리즘이라는 두 가지 개선을 도입하였다.

Global-Local 모델은 마스크 생성 – 지역 특성 및 전역 특성 추출 – 특성 융합 – 유사도 계산 순으로 구성된다. 여기서 마스크 생성이란 이미지 내 객체를 분할하여 각각 마스크 형태로 추출하는 과정이다. 이후 이 마스크를 활용해 전체 이미지에서 객체 부분을 잘라내고, CLIP 을 활용해 인코딩하여 지역 이미지 특성을 얻는다. 지역 텍스트 특성은 전체 참조 텍스트에서 목표 명사를 추출해 이를 CLIP 으로 인코딩한 것이다. 한편 전역 이미지 특성은 특정 객체를 제외한 전체 이미지, 전역 텍스트 특성은 전체 참조 텍스트로부터 추출된다.

마스크 생성 단계는 전체 파이프라인의 품질을 좌우하는 핵심 과정이다. 그러나 기존 모델은 마스크 생성 단계에서 각 객체를 명확하게 분리하지 못하고, 동일 카테고리 또는 인접한 객체를 한 마스크로 묶는 경우가 존재하였다. 이 경우 텍스트와의 매칭 단계가 정확하게 수행되기 어렵고, 방향 정보에 일치하는 객체 마스크 또한 판별할 수 없다. 본 연구는 이를 보완하기 위해 제로샷 환경에서 우수한 성능을 보이는 분할 모델인 SAM 을 도입하였다. 이를 통해 객체 단위에서 정확한 분할 마스크를 생성하고, 방향 인지 단계에서 마스크 자체의 부정확성으로 인한 영향을 최소화하고자 하였다.

더 나아가 주어진 참조 텍스트에 방향 지시어가 포함된 경우, 기존의 이미지-텍스트 매칭 과정에 방향 인지 알고리즘을 거친 최종 선택 단계를 추가하였다. 서로 반대 방향을 나타내는 5 쌍, 총 10 개의 단어로 방향 지시어 사전을 구성하여, 단어에 맞는 객체 마스크를 바운딩박스(bounding box)의 좌표를 기준으로 선택하도록 규칙 기반의 알고리즘을 설계하였다. 관찰 결과 정답이 상위 3 개 내에 포함될 확률이 78%임을 반영해, 이미지-텍스트 특성 유사도 상위 3 개의 마스크를 고려하여 위치 비교를 수행한다. 이 과정을 통해 좌표값이라는 정량적 기준을 활용하여 모델이 방향 정보를 명시적으로 처리하도록 하였다.

3. 실험 결과

본 연구에서는 모델의 성능을 RefCOCOg 의 validation 데이터셋을 활용하여 평가하였다. 평가 지표로는 이미지 분할 과제에서 널리 쓰이는 지표인 Overall Intersection Over Union (oIOU)와 mean Intersection Over Union (mIOU)를 활용하였다. oIOU 는 예측 마스크와 정답(Ground Truth) 마스크의 교차(intersection) 면적을 면적 합(union)으로 나눈 값으로, 전체적인 분할 정확도를 나타낸다. mIOU 는 각 샘플에서 예측 마스크와 정답 마스크의 면적 합 대비 교차 면적의 비율을 평균한 값으로, 샘플 수준에서의 정확도를 평가한다.

Model Configuration	oIOU	mIOU
Global-Local	24.83	24.58
Global-Local with SAM	26.27	31.84
Ours (SAM+Orientation-aware)	**28.73**	**40.38**

<표 1> 모델 실험 결과

표 1의 결과는 본 연구에서 제안한 개선이 기존 모델 대비 성능 향상에 기여함을 보여준다. 마스크 제안 모델을 SAM 으로 교체하는 것만으로도 베이스라인 대비 유의미한 성능 개선이 관찰되었으며, 이는 SAM 이 객체 단위에서 고품질의 분할 마스크를 제공함을 의미한다. 여기에 방향 인지 알고리즘을 추가했을 때 성능은 추가적으로 향상되었으며, 이는 단순하지만 효과적인 모듈 도입만으로도 객체 단위 방향 인지 성능을 크게 강화할 수 있음을 입증한다.

4. 결론 및 제언

본 연구에서는 마스크 생성 모델 강화와 방향 인지 알고리즘 도입을 통해 방향 정보를 반영한 제로샷 참조 이미지 분할 모델을 제안하였다. RefCOCOg 데이터셋을 활용한 실험 결과, 제안 모델은 기존 접근법 대비 oIoU 및 mIoU 에서 모두 성능이 향상되었으며, 단순한 모듈 추가만으로도 참조 이미지 분할에서의 방향 인지 성능을 효과적으로 강화할 수 있음을 입증하였다.

향후 연구에서는 방향 지시어를 확장하여, 단순한 사전 기반 접근 이상의 보다 지능화된 방향 정보 처리를 구현할 수 있다. 더 나아가 깊이 인식 모델과의 결합을 통해 앞 · 뒤와 같은 복잡한 공간 관계까지 고려하는 방향으로의 발전이 가능하다. 더 나아가 제안 모델을 로봇 비전이나 시각적 질의응답 시스템 등 실제 응용 환경에 적용함으로써 실질적인 활용 가능성을 확인할 수 있을 것이다.

사사

이 논문은 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원-대학 ICT 연구센터(ITRC; RS-2024-00436936), 학석사연계 ICT 핵심인재양성사업(RS-2023-00259497), 인간지향적 차세대 도전형 AI 기술 개발(RS-2025-25440264, 인간의 감각 인지 프로세스 기반 범용인공지능 개발)의 지원을 받아 수행된 연구임.

참고문헌

[1] Li Xu, Mark He Huang, Xindi Shang, Zehuan Yuan, Ying Sun, and Jun Liu. Meta compositional referring expression segmentation, IEEE/CVF conference on computer vision and pattern recognition, Vancouver, 2023, p. 19478-19487

[2] Bin Yan, Yi Jiang, Jiannan Wu, Dong Wang, Ping Luo, Zehuan Yuan, and Huchuan Lu. Universal instance perception as object discovery and retrieval, IEEE/CVF conference on computer vision and pattern recognition, Vancouver, 2023, p. 15325-15336

[3] Zhao Yang, Jiaqi Wang, Yansong Tang, Kai Chen, Hengshuang Zhao, and Philip H. S. Torr. Lavt: Language-aware vision transformer for referring image segmentation, IEEE/CVF conference on computer vision and pattern recognition, New Orleans, 2022, p. 18155-18165

[4] Wenliang Zhao, Yongming Rao, Zuyan Liu, Benlin Liu, Jie Zhou, and Jiwen Lu. Unleashing text-to-image diffusion models for visual perception, IEEE/CVF conference on computer vision and pattern recognition, Vancouver, 2023, p. 5729-5739

[5] Seonghoon Yu, Paul Hongsuck Seo, and Jeany Son. Zero-shot referring image segmentation with global-local context features, IEEE/CVF conference on computer vision and pattern recognition, Vancouver, 2023, p. 19456-19465

[6] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, International conference on machine learning, 2021, p. 8748-8763

[7] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer White-head, Alexander C. Berg, Wan-Yen Lo, Piotr Doll´ar, and Ross Girshick. Segment anything, IEEE/CVF conference on computer vision and pattern recognition, Vancouver, 2023, p. 4015-4026