Text-to-Image 멀티모달 신경망 아키텍쳐를 활용한 객체추적 알고리즘

강병우, 김남우, 유환조*

포항공과대학교

{bykang, treekim, gshan, hwanjoyu*}@postech.ac.kr

A Object Tracking Algorithm Using Text-to-Image Multi Modal Neural Network

Kang Byoungwoo, Kim Nam U, Hwanjo Yu*

{bykang, treekim, hwanjoyu*}@postech.ac.kr

요 약

본 논문에서는 이미지와 설명 텍스트의 쌍으로 사전 학습된 멀티모달 신경망 아키텍처인 CLiP(Contrastive Language-Image Pre-training)[1] 모델을 활용하여 객체 추적에서의 ID-switching[2] 문제를 개선하는 알고리즘을 제안한다. 기존 주요 방법론은 객체 탐지 후, 별도로 객체 추적을 수행한다. 객체 추적 알고리즘이 인식된 객체에 대해 각 인스턴스를 구분할 수 있는 적절한 특징을 추출하지 못하거나, 연속된 프레임에서 인스턴스 별 공통 속성을 포착하지 못할 경우에는 서로 다른 인스턴스를 동일한 인스턴스로 잘못 추적하는 ID-switching 문제로 이어진다. CLiP의 이미지 인코더를 활용해 인식된 각 객체의 인스턴스의 image feature embedding을 확보하고 이를 기반으로 추적을 수행함으로써 ID-switching 문제를 개선하고자 하였다.

Ⅰ. 서 론

객체 인식(object detection)과 객체 추적(object tracking)을 통합하여 수행하는 end-to-end 아키텍처는 각 인스턴스에 대한 정밀한 레이블링이 필요하다. 이 방식은 고품질의 추적 성능을 달성하기 위해 대량의 데이터를 필요로 하며, 이로 인해 상당한 비용이 발생한다. 특히, 이 방법은 학습된 데이터셋에 과적합되는 경향이 있으며, 새로운 차량 모델이나 다양한 해상도를 가진 CCTV 이미지와 같이 변형된 특성이 등장할 때 성능이 저하된다. 이러한 비용 문제와 확장성의 부족은 산업적 활용에 심각한 제약을 초래한다.

이에 따라, 현재 주요 객체 추적 알고리즘은 YOLOv8, DETR[3]과 같은 강력한 객체 인식 알고리즘을 통해 객체를 우선 인식한 후에, 이 정보를 기반으로 별도의 객체 추적 작업을 수행하는 분리된 모델의 파이프라인으로 구성되는 방식이 일반적이다. 이 접근법은 zero-shot learning과 같은 기술을 활용하여, 레이블이 지정된 데이터셋이 없어도 어느 정도의 성능을 보장함으로써, end-to-end 모델에 비해 비용 효율적이다. 그러나 이러한 고성능의 객체 인식 모델을 사용하더라도, 객체 추적 알고리즘에서 동일한 인스턴스를 올바르게 추적하지 못하고 ID를 혼동하여 발생하는 ID-switching 문제로 인해 추적 정확도가 저하된다. 이는 CCTV 이미지 데이터를 활용한 차량 입출차 카운팅과 같은 산업적 활용에 큰 어려움을 준다.

이러한 문제를 해결하기 위해, CLiP 모델을 supervision으로 활용하여 개선한 새로운 객체 추적 알고리즘인 CLiPTracker를 제안한다. 본 논문은 CLiPTracker가 기존 zero-shot 객체 추적 알고리즘과 비교하여 ID-switching 문제를 효과적으로 개선할 수 있는지를 분석한다.

Ⅱ. 본론

본 논문에서는 ㈜한맥아이피에스로부터 제공받은 주차장 통로 CCTV 영상 이미지 데이터셋을 활용하여 실험을 수행하였다. 실험의 주목적이 객체 추적 알고리즘의 ID-switching 문제를 개선하는 것이므로, 정확한 객체 탐지가 가능하다는 가정을 바탕으로 초기 작업을 진행하였다. 데이터셋 내의 차량 객체에 대해 라벨링을 수행하고, YOLOv8 모델을 사용하여 500장의 이미지에 대해 사전 학습을 수행하였다. 실험은 객체 탐지에서 차량 클래스에 대해 confidence 수치가 0.6 이상인 결과를 기준으로 진행하였다.

학습된 이미지 데이터를 통해 모든 차량 객체가 정확하게 탐지되었음을 확인한 후, 동일한 객체 탐지 모델을 기반으로 다양한 객체 추적 알고리즘의 ID switching 성능을 평가하였다. 객체 추적 알고리즘의 zero-shot 성능 비교를 위해 추가 학습 없이 사용 가능한 칼만 필터 기반의 ByteTrack[3]과 BoT-SORT[4]를 비교 모델로 선정하였다.

CLiPTracker의 작동 원리는 다음과 같다. 연속된 두 프레임 내에서 인식된 객체 별로 crop된 이미지를 CLiP의 이미지 인코더를 통과시켜 각 인스턴스 별 image feature를 추출한다. (t-1) 번째 동영상 이미지 프레임에서 탐지된 객체 m개의 feature vector $\{v_1, v_2, \dots, v_m\}$과 (t) 번째 동영상 이미지 프레임에서 탐지된 객체의 image feature $\{u_1, u_2, \dots, u_n\}$에 대해, 다음과 같이 코사인 유사도를 구한다.

$$s_{ij} = \frac{v_i \cdot u_j}{\| v_i \| \, \| u_j \|} \quad \text{for } i = 1, \dots, m \text{ and } j = 1, \dots, n$$

프레임 간 임베딩된 feature들의 코사인 유사도 비교를 통해 동일 인스턴스인지 여부를 판단하는 ID matching을 수행한다. 실험에서는 코사인 유사도가 0.4 이상일 경우 같은 객체로 간주하고, 그 이하일 경우 새로운 ID를 할당하였다.

차량 이동이 존재하는 초반 250프레임의 동영상 이미지에 대해 ID-switching이 발생하는 횟수 비교를 통해 객체 추적 모델별 성능 비교를 수행하였다.

[그림 1] Objtect Tracking 알고리즘 별 ID-switching 비교 사례

객체추적 알고리즘	초반 250프레임 내 ID-switching 발생횟수
ByteTrack	7
BoT-SORT	9
CLiPTracker	3

[도표 1] Object Tracking 알고리즘 별 ID-switching 발생횟수 비교

CLiPTracker는 ByteTrack, BoT-SORT에 비해 ID-switching에 대해 강건한 성능을 보였다. 이는 CLiP의 이미지 인코더를 통해 각 인스턴스별로 고유한 이미지 임베딩 특성을 추출할 수 있기 때문이다.

[그림 2] 장애물에 의한 객체의 폐색에도 강건한 CLiPTracker

CLiPTracker는 기둥 등의 장애물에 의해 차량이 부분적으로 가려지는 폐색의 경우에도 효과적으로 객체를 추적하는 것을 확인할 수 있었다. 하지만 차량이 서로 겹칠 경우, bounding box를 기준으로 crop된 이미지의 feature를 얻는 구조로 인해 서로 다른 두 차량을 별도의 인스턴스로 인식하지 못하는 한계점이 존재했다.

Ⅲ. 결론

본 논문을 통해 CLiP, 멀티모달 모델을 활용하여 추가적인 학습 없이도 객체 추적에서의 주요 문제인 ID-switching에 강건한 추적 알고리즘의 가능성을 확인하였다. CLiPTracker는 기존 방법론에서 요구되는 대량의 레이블링 데이터 없이도 높은 성능을 발휘할 수 있는 접근 방식을 제시하였다.

이 연구를 발전시킬 수 있는 후속 연구 방향을 다음과 고민해볼 수 있다.

첫째, 멀티모달 대규모 언어 모델(LLM) 혹은 비디오-to-텍스트 모델 Sora[5] 등을 활용하여 얻은 이미지의 description embedding feature를 얻는다면, 각 인스턴스에 대한 보다 고유한 feature embedding을 추출할 수 있을 것이다. 이는 객체의 고유 특성을 더욱 정확하게 파악하여 추적 성능을 향상시킬 가능성을 제공한다.

둘째로, 이미지 이상 탐지 알고리즘인 WinCLiP[6]의 방법론을 활용할 수 있다. 해당 논문에서는 이미지의 상태를 기술하는 텍스트 프롬프트를 사전에 설정한다. 이 방법에서는 CLiP의 텍스트 인코더를 사용하여 사전에 정의된 프롬프트 중에서 이미지 특성과 가장 높은 유사도를 보이는 프롬프트를 통해 이미지를 분류하는 접근법을 사용한다. 해당 접근법을 통해 CLiP의 이미지 인코더뿐만 아니라 텍스트 인코더를 함께 사용한다면 CLiPTracker의 객체 추적 성능을 더욱 개선할 수 있을 것이다.

ACKNOWLEDGMENT

This work was supported by the Digital Innovation Hub project supervised by the Daegu Digital Innovation Promotion Agency(DIP) grant funded by the Korea government(MSIT and Daegu Metropolitan City) in 2024 (No. DBSD1-07).

※ MSIT: Ministry of Science and ICT.

참 고 문 헌

[1] Radford, Alec, et al. "Learning transferable visual models from natural language supervision." International conference on machine learning. PMLR, 2021.

[2] Bashar, Mk, et al. "Multiple object tracking in recent times: A literature review." arXiv preprint arXiv:2209.04796 (2022).

[3] Carion, Nicolas, et al. "End-to-end object detection with transformers." European conference on computer vision. Cham: Springer International Publishing, 2020.

[3] Zhang, Yifu, et al. "Bytetrack: Multi-object tracking by associating every detection box." European conference on computer vision. Cham: Springer Nature Switzerland, 2022.

[4] Aharon, Nir, Roy Orfaig, and Ben-Zion Bobrovsky. "BoT-SORT: Robust associations multi-pedestrian tracking." arXiv preprint arXiv:2206.14651 (2022).

[5] Liu, Yixin, et al. "Sora: A Review on Background, Technology, Limitations, and Opportunities of Large Vision Models." arXiv preprint arXiv:2402.17177 (2024).

[6] Jeong, Jongheon, et al. "Winclip: Zero-/few-shot anomaly classification and segmentation." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023.