시간적 정보를 고려하는 Dynamic Scene Graph Generation

조혜원[1], 조영준[2]
[1]전남대학교 인공지능학부 학부생
[2]전남대학교 인공지능융합학과 부교수
haewon@jnu.ac.kr, yj.cho@jnu.ac.kr

Temporal-Aware Dynamic Scene Graph Generation

Hae-Won Jo[1], Yeong-Jun Cho[2]
[1]Dept. of Artificial Intelligence Convergence, Chonnam National University

요 약

최근 Scene Graph Generation(SGG)이 시각적 이해와 상호작용 시스템에서 중요한 연구 분야로 떠오르고 있으며, 이는 정적 이미지에서의 객체와 객체 간 관계를 분석하는데 주로 사용되었다. 하지만 동적인 환경에서의 장면 그래프 생성은 시간에 따라 변화하는 객체의 상태와 그 사이의 관계를 정확히 파악하고 모델링하는 데 있어 훨씬 더 큰 도전이다. 본 논문에서는 비디오 시퀀스로부터 Dynamic Scene Graph(DSG)를 생성하는 새로운 방법을 제안한다. 이는 동적 환경에서의 객체 변화와 그 사이의 상호작용을 실시간으로 추적하고 이해하는 데 필수적인 기능을 제공한다. 본 논문에선 CNN 백본 네트워크와 인코더를 사용하여 각 프레임에서 시각적 표현을 추출하고, attention 메커니즘을 적용하여 시간적 정보를 통합한다. 또한, 디코더 내부에서 학습 가능한 쿼리를 이용해 복잡한 객체 간 관계를 인식하며, 헝가리안 알고리즘 매칭 로스를 사용하여 각 쿼리는 쌍별 인스턴스 특징을 학습하게되고, 디코더의 각 attention layer에서는 객체 간의 복잡한 관계를 추출하고, 이를 통해 각 관계에 적합한 predicate를 예측한다.

1. 서론

Scene Graph Generation은 컴퓨터 비전 분야의 문제 중 하나로, 구조화되지 않은 이미지나 비디오를 추상적인 그래프 구조의 형태로 설명하는 것을 목표로 한다. 이러한 Scene Graph는 방향성 그래프로, 각 노드는 서로 영향을 끼치는 객체를 나타내고, 엣지는 해당 객체들 사이의 관계인 술어를 나타낸다.

본 논문에서는 프레임 순서를 바탕으로 시간적 정보를 통합하여 객체간의 동적인 관계를 더욱 잘 파악하는 Scene Graph를 구성하는 새로운 방법을 제안한다. 제안하는 방법은 첫째, 각 비디오 프레임을 CNN 백본 네트워크와 인코더를 통해 처리하여 시각적 정보를 추출한다. 둘째, 이러한 시각적 표현을 기반으로 Self-attention 메커니즘을 적용하여 프레임 간의 시간적 정보를 통합하는 과정을 거친다. 디코더 부분에서는 학습 가능한 쿼리를 사용하여 각 쿼리가 두 개의 객체 인스턴스 특성을 학습할 수 있도록 구성한다. 이 과정에서 헝가리안 알고리즘을 사용한 매칭 로스를 적용하여, 객체 쌍을 매칭할 수 있도록 한다. 디코더의 각 attention layer에서는 입력된 쿼리를 활용하여 객체 간의 복잡한 관계를 추출하고, 이를 기반으로 다양한 predicate를 예측한다. 이러한 기능은 비디오 시퀀스 내에서 객체들 사이의 상호작용을 정확하게 해석하고, 이를 통해 동적 장면 그래프의 생성에 필수적인 정보를 제공한다. 제안하는 방법은 비디오에서 시간적-공간적 맥락을 고려한 Dynamic Scene Graph Generation의 정확성을 향상시키고, 보다 더 복잡한 동적 장면을 이해할 수 있도록 한다.

2. 관련연구

대부분의 이전 연구들은 [1, 2, 3] two-stage 접근 방식을 선택하고 있다. 첫 번째 단계는 사전 학습된 Faster R-CNN [4]과 같은 객체 탐지기를 사용하여 모든 객체를 탐지하는 것이다. 그런 다음, 모든 후보 객체 쌍 간의 관계를 Message Passing

[2, 3], Graph Convolutional Network[1]와 같은 모듈을 사용하여 분류한다.

3. 제안방법

본 연구에서 사용된 백본 네트워크와 인코더는 비디오 시퀀스에서 각 프레임의 시각적 표현을 추출하는 데 핵심적인 역할을 한다. CNN 백본은 입력 이미지로부터 특징 맵을 생성한다.

3.1.Encoder

한다. 본 연구에서는 DETR [5]의 인코더 구조를 채택하여 사용하였다. 인코더에 입력할 때는 하나의 프레임을 입력으로 받으며, 인코더의 출력은 해당 프레임에 대한 context를 생성한다. 이후 DETR 인코더를 사용하여 각 프레임에 대한 context값을 추출한다. 각 프레임에 대해 이 과정을 반복하여 모든 프레임의 context 값을 추출한 후, 이들 사이에서 Attention 메커니즘을 적용하여 시공간적 문맥 정보를 통합한다. 이때 프레임 인코딩을 추가하여 각 프레임 간의 시간적 정보를 포함한다. 이 과정을 통해 각 프레임에서 시공간적 문맥이 통합된 context를 생성할 수 있다. 최종적으로, 이 통합된 시공간적 context는 나중에 프레임별로 분리하여 사용한다.

3.2.Decoder

본 연구에서는 디코더의 입력으로 학습 가능한 객체 쿼리를 사용한다. 디코더는 두 단계로 구성된다. 첫 번째 단계에서는 Self-attention 메커니즘을 통해 객체 쿼리들 간의 의존성을 학습하고, 두 번째 단계에서는 인코더에서 추출한 시공간적 정보가 통합된 컨텍스트를 활용하여 Cross-attention을 수행한다. 본 논문에서는 각 객체 쿼리가 두 개의 객체 인스턴스를 예측할 수 있도록 하기 위해 헝가리안 알고리즘 매칭 로스(Hungarian Algorithm Matching Loss)를 사용한다. 이를 통해 객체 쿼리는 쌍(pair-wise) 형태의 객체 특징을 학습하며, 객체 간의 복잡한 관계를 더 잘 반영할 수 있다. 이는 각 객체 쿼리가 시간적 정보를 포함한 두 객체 간의 관계를 명시적으로 학습하기 때문에 spatial-temporal 정보를 더 효율적으로 활용할 수 있다.

4. Loss

본 연구에서는 두 가지 손실 함수를 결합하여 최종 손실 함수를 정의한다. 첫 번째 손실 함수는 OED [6]에서 영감을 받아 헝가리안 매칭 알고리즘을 사용하여 객체 쿼리가 쌍(pair-wise) 인스턴스 특징을 잘 학습하도록 한다. 두 번째 손실 함수는 EGTR [7]에서 영감을 받아 디코더의 attention layer에서 추출한 관계(Relation)를 바탕으로 Predicate를 예측하는 부분에 Binary Cross Entropy Loss를 적용한다.

감사의글

본 연구는 과학기술정보통신부 및 정보통신기획평가원의 인공지능융합혁신인재양성사업(IITP-2023-RS-2023-00256629) 및 소프트웨어중심대학사업(2021-0-01409)의 연구결과로 수행되었음.

참고문헌

[1] X. Lin, C. Ding, J. Zeng, and D. Tao, “GPS-Net: Graph property sensing network for scene graph gerꟷeration,” in Proc. IEEE/CVF Conf. Comput. Vis. Patꟷtern Recognit., 2020, pp. 3746-3753.

[2] D. Xu, Y. Zhu, C. B. Choy, and L. Fei-Fei, “Scene graph generation by iterative message passing,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2017, pp. 5410-5419.

[3] J. Yang, J. Lu, S. Lee, D. Batra, and D. Parikh, “Graph R-CNN for scene graph generation,” in Proc. Eur. Conf. Comput. Vis., 2018, pp. 670-685.

[4] S. Ren, K. He, R. Girshick, and J. Sun, “Faster Rꟷ CNN: Towards real-time object detection with region proposal networks,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 39, no. 6, pp. 1137-1149, Jun. 2017.

[5] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirꟷillov, and S. Zagoruyko, “End-to-end object detection with transformers,” in Proc. Eur. Conf. Comput. Vis., 2020, pp. 213-229. 3

[6] G. Wang, Z. Li, Q. Chen, and Y. Liu, “OED: Towards One-stage End-to-End Dynamic Scene Graph Generation,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2024, pp. 27938-27947.

[7] J. Im, J. Nam, N. Park, H. Lee, and S. Park, “Egtr: Exꟷtracting graph from transformer for scene graph generation,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2024, pp. 24229-24238. 4, 5