동의어를 이용한 시각적 의미 지식 장면 그래프 통합 연구

최우석[O1], 온경운[1], 허유정[1], 장병탁[1,2]
[1]서울대학교, [2]서울대학교 AI 연구원 (AIIS)
{wschoi, kwon, yjheo, btzhang}@bi.snu.ac.kr

A study on Scene Graph Unification of Visual Semantic Knowledge using synonym

Woo Suk Choi[O1], Kyoung-Woon On[1], Yu-Jung Heo[1], Byoung-Tak Zhang[1,2]
[1]Seoul National University
[2]AI Institute (AIIS), Seoul National University

요 약

장면 그래프는 물체, 특성, 물체들 간의 관계처럼 이미지의 고차원 의미 지식을 표현하는 그래프이다. 장면 그래프에 관련된 다양한 연구가 제안되어왔지만, 연구마다 각기 다른 가설을 가지고 있어 각 태스크마다 한정된 어휘 목록과 편향된 정보를 가지고 있다. 그러기에 각 연구마다의 결과가 보편화(일반화) 되어있지 않고 다른 down-stream 태스크에 적용되기 어렵다. 본 논문에서는 이러한 편향된 문제점을 효율적으로 해결하기 위해 다양한 의미 지식을 정렬시켜주어 장면 그래프를 통합 시키는 방법을 제안한다. 여러 장면 그래프 및 의미 지식을 정렬 시켜주기 위해 큰 규모의 어휘 데이터베이스인 WordNet의 synset을 사용한다. 실험에서는 VG-BUTD, VG200, VrR-VG 데이터로부터 학습된 예측 모델을 down-stream 태스크인 이미지-캡션 태스크에 적용시켜 실험 하였고, 통합된 장면 그래프가 이미지에 대해 더 많은 정보를 가지고 있음을 확인하였다.

1. 서 론

시각적 장면 이해에 대한 연구는 각각의 물체를 인식하는 연구를 넘어 이미지로부터 장면 그래프를 생성하여 의미 지식을 추출하는 연구로 발전해가고 있다. Visual Genome [1]을 시작으로, 이미지로부터 의미 지식을 생성하는 다양한 연구 (VG-BUTD[2], Neural Motif[3], VrR-VG[4])가 진행 되어왔다. 하지만 각각의 연구는 데이터셋의 통계적 편향과 가설에 따른 한정된 어휘 목록 때문에 이미지로부터 편향된 정보를 추출한다. 예를 들어 VG-BUTD[2]의 경우, 해당 저자는 물체 및 특성 정보를 추출하기 위해 1,600개의 물체 라벨과 400개의 특성 라벨을 사용하였다. 또한 Neural Motif[3]에서는 150개의 물체 라벨 및 50개의 관계 라벨을 사용하여 물체간의 관계를 생성하는 연구를 제안하였다. Visually-Relevant Relationships(VrR-VG)[4]에서는 해당 저자가 Visual Genome 데이터셋에서 조금 더 가치 있는 관계를 찾아내기 위해 1,600개의 물체 라벨과 117개의 관계 라벨을 사용하여 데이터셋을 구축하였다. 이처럼 각각의 연구는 각각 그들만의 어휘 목록을 정의하여 사용하지만 굉장히 한정적이라는 문제점이 있다. 그러기에 어떤 물체가 데이터셋에 포함된 특정된 단어가 아니라면, 아무리 이미지에 물체가 있다하더라도 누락될 수 있다. 관계 또한 마찬가지이다. 관련된 예를 그림 1에서 볼 수 있다. 공통된 이미지에서 사람을 man과 person으로 다르게 표기하는 것처럼 같은 물체를 다른 어휘로 표기되는 경우가 생길 수 있으며, sofa와 pillow처럼 인식되는 물체 또한 다를 수 있다.

본 논문에서는 다양한 의미 지식을 정렬 시켜주어 장면 그래프를 통합하기 위한 방법을 제안한다. 해당 방법은 공통된 이미지에 대해 WordNet을 활용하여 같은 물체들의 라벨 정렬을 해주고 해당 물체가 다른 장면 그래프에서 같은 물체를 의미하는지 알아내기 위해 IoU (Intersection of Union) 계산을 이용한다.

본 논문에서 제안한 연구 방법의 기여는 다음과 같다. 1) 여러 장면 그래프를 하나의 그래프로 만들기 위해 어휘 정보를 활용하였다. 2) 실험 결과를 통해 통합된 그래프가 이미지에 대해 더 많은 의미 정보를 포함하고 있다는 것을 보여준다. 3) 통합된 그래프가 이미지-캡션 down-stream 태스크에서 좋은 성능을 보였다.

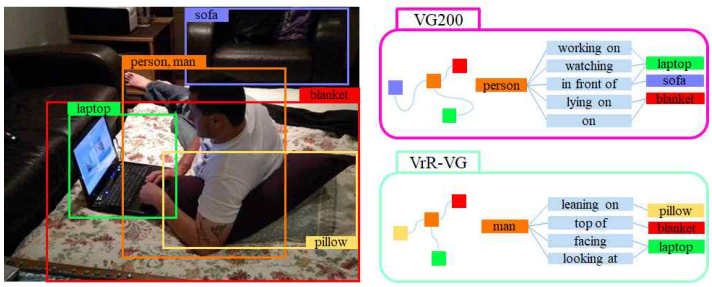

그림 1: VG200과 VrR-VG에서의 공통된 이미지에 대한 장면 그래프 예제

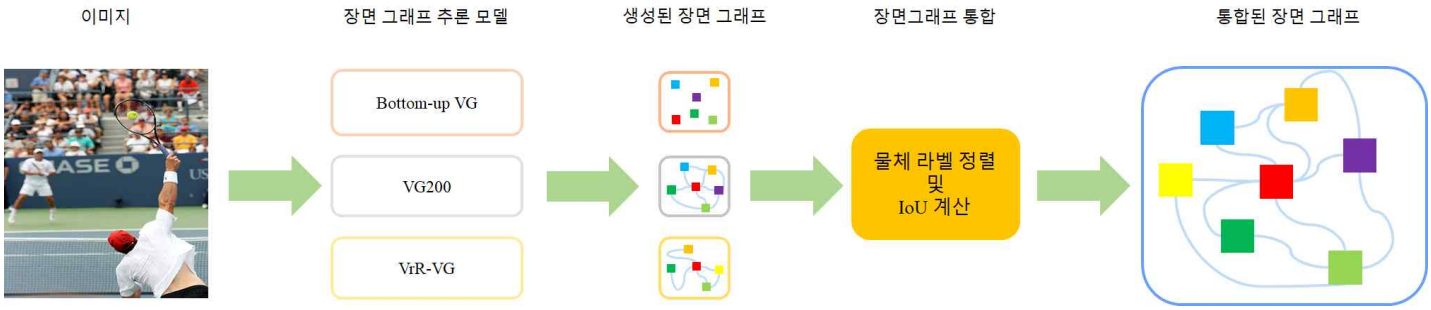

그림 2: 시각적 의미 지식 장면 그래프 통합 연구 전체적 구조

2. 연구방법

본 논문에서는 상향식 주의와 구성변형 임베딩을 활용하여 이미지의 장면 그래프를 생성하고, 생성된 장면 그래프들을 라벨 정렬 및 통합과 IoU 계산을 통해 통합하여 하나의 그래프로 만든다. 연구방법의 전체적인 구조는 그림 2와 같다.

2.1 장면 그래프 추론 모델

이미지에서 두 물체간의 의미론적 관계를 학습하기 위해 지식 그래프 임베딩 알고리즘인 TransR[5]에 추이 제약을 적용시킨 구성 변형 임베딩을 사용하여 장면 그래프 추론 모델을 구축하였다. 구성 변형 임베딩은 잠재 관계 부분 공간에 물체 특징 벡터를 투사하여 여러 의미 있는 관계를 추론하고 장면 그래프로 나타낸다.

2.2 물체 라벨 정렬

물체 라벨 정렬은 각 데이터셋으로부터 생성된 단일 장면 그래프들을 통합하는 방법으로, 1차적으로 자연어 toolkit인 WordNet의 synset을 이용하여 공통된 의미를 가지는 단어들을 정렬 및 통합한다. Synset은 WordNet에서 공통된 의미를 공유하는 동의어 집합으로 단어의 기본형(lemma), 상위어(Hypernym), 하위어(Hyponym)가 포함되어 있다. 다른 단어이더라도 synset을 이용하여 비교 후 의미가 같은 단어이면 통합하고 다음 단계인 IoU 계산법을 통해 해당 물체가 같은 물체를 의미하는지 확인한다. 만약 의미가 다른 단어라면 새로운 물체 노드를 추가시켜준다.

3. 실험 구성 및 결과

실험을 위해 Visual Genome(VG)[1] 기반의 데이터 셋들인 VG200, VrR-VG를 사용하였으며 물체의 특성도 추가해주기 위해 VG-BUTD도 사용하였다. VG200에서 이미지 당 평균 물체 개수는 12.53개이고 평균 관계 개수는 50개이다. VrR-VG에서 이미지 당 평균 물체 개수는 36.77개이며 평균 관계 개수는 50개이다. Bottom-up VG에서의 이미지 당 평균 물체 개수는 26.35개이며 평균 특성의 개수는 26.35이다. 해당 통계는 표 1에 있으며 합쳤을 때의 평균 물체 개수와 관계 개수 그리고 특성의 개수가 표시되어있다.

데이터셋	평균 물체	평균 관계	평균 특성
VG200	12.53	50	0
VrR	36.77	50	0
BU	26.35	0	26.35
VG200+VrR	37.00	100	0
VG200+BU	27.21	44.39	26.35
VrR+BU	42.04	29.57	26.35
VG200+VrR+BU	41.95	79.67	26.35

표 1 통합된 장면 그래프의 이미지 당 물체, 관계, 특성 평균 개수. (VrR은 VrR-VG의 축약형, BU는 VG-BUTD의 축약형)

제안한 방법의 유용성을 확인하기위해 down-stream 태스크인 이미지-캡션 복구 태스크에 적용시켜보았다. 이미지-캡션 복구 태스크는 장면 그래프를 활용하여 풀 수 있는 태스크이기에 적용시켜보았다. 평가 지표는

데이터셋	이미지-캡션 복구 (Retrieval)		
	R@1	R@5	R@10
VG200	22.2	57.6	73.2
VrR	28.1	66.2	80.4
BU	27.0	65.4	80.6
VG200+VrR	29.3	67.6	81.9
VG200+BU	29.4	68.7	82.8
VrR+BU	27.9	70.5	83.2
VG200+VrR+BU	27.2	70.0	82.4

표 2 이미지-캡션 복구(image-to-caption retrieval) 태스크에 대한 결과 (VrR은 VrR-VG의 축약형, BU는 VG-BUTD의 축약형)

Recall@k로 하였으며, 전반적으로 기존 단일 장면 그래프를 사용한 것에 비해 통합한 장면 그래프를 사용한 것이 적게는 1.3%에서 크게는 10%정도 더 좋은 성능을 보였고 해당 결과는 표 2에 나와 있다.

4. 결론

본 논문에서는 여러 시각적 의미 지식을 하나의 큰 장면 그래프로 통합하는 간단하고 효과적인 방법을 제안하였다. 큰 규모의 어휘 데이터베이스 WordNet에 있는 synset과 IoU 계산법을 사용하여 각 데이터셋의 단일 그래프 단어들을 정렬 및 통합시켜주어 하나의 통합된 장면 그래프를 만들었다. 통합된 장면 그래프는 단일 장면 그래프보다 더 많은 정보를 포함하고 있고 down-stream 태스크인 이미지-캡션 복구 태스크에서도 좋은 성능을 보였다.

5. 감사의 글

이 논문은 2020년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원(2015-0-00310-SW.StarLab, 2017-0-01772-VTT, 2018-0-00622-RMI, 2019-0-01367-BabyMind), 한국산업기술진흥원(P0006720-GENKO)의 지원을 받았음.

참고문헌

[1] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, and S. Chen, “Visual genome: Connecting language and vision using crowdsourced dense image annotations,” International Journal of Computer Vision, 123(1): 32-73, 2017

[2] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang “Bottom-up and top-down attention for image captioning and visual question answering,” In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018

[3] R. Zellers, M. Yatskar, S. Thomson, and Y. Choi, “Neural motif: Scene graph parsing with global context” In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018

[4] Y. Liang, Y. Bai, W. Zhang, X. Qian, L. Zhu, and T. Mei, “Vrr-vg: Refocusing visually-relevant relationships.” In Proceedings of the IEE International Conference on Computer Vision, 2019

[5] Y. Lin, Z. Liu, M. Sun, Y. Liu, and X. Zhu, “Learning entity and relation embeddings for knowledge graph completion”, In Twenty-ninth AAAI Conference on artificial intelligence, 2015