PSG

기존의 연구는 이미지에서의 장면 이해를 위한 중요한 기술인 장면 그래프 생성(Scene Graph Generation, SGG)을 검출(detection) 관점에서 다루었습니다. 즉, 객체는 바운딩 박스를 사용하여 감지한 후 그들의 쌍(pairwise) 관계를 예측합니다. 우리는 이러한 패러다임이 분야의 발전을 방해하는 여러 문제를 일으킨다고 주장합니다. 예를 들어, 현재의 데이터셋에서는 바운딩 박스 기반의 레이블이 머리카락과 같은 중복된 클래스를 포함하고 있으며, 문맥 이해에 중요한 배경 정보를 놓치는 경우가 많습니다.

이 연구에서는 팬옵틱(segmentations) 세그멘테이션을 기반으로 한 더 포괄적인 장면 그래프 생성(Panoptic Scene Graph Generation, PSG)이라는 새로운 문제를 소개합니다. PSG는 강성한 바운딩 박스 대신 팬옵틱 세그멘테이션을 기반으로 한 보다 포괄적인 장면 그래프 표현을 생성하는 것을 요구합니다. COCO와 Visual Genome에서 4.9만 개의 잘 주석이 달린 겹치는 이미지로 이루어진 고품질 PSG 데이터셋을 생성하여 연구 커뮤니티가 발전을 추적할 수 있도록 하였습니다.

성능 평가를 위해, 우리는 SGG의 클래식한 방법들을 수정한 4개의 2단계 베이스라인과, PSGTR 및 PSGFormer라는 2개의 DETR(Detection Transformer)을 기반으로 한 1단계 베이스라인을 구축하였습니다. PSGTR은 직접적으로 트리플릿을 학습하기 위해 쿼리 세트를 사용하며, PSGFormer는 객체와 관계를 Transformer 디코더의 쿼리 형태로 개별적으로 모델링한 뒤 프롬프트와 유사한 관계-객체 매칭 메커니즘을 적용합니다.

1. 서론

장면 그래프 생성(Scene Graph Generation, SGG) 작업의 목표는 주어진 이미지로부터 그래프 구조의 표현을 생성하는 것입니다. 이 표현은 객체들과 바운딩 박스를 기반으로 하는 이들 간의 쌍(pairwise) 관계를 추상화합니다. 장면 그래프는 복잡한 이미지의 장면을 이해하는 데 도움을 주며, 이미지 검색, 시각적 추론, 시각적 질문 응답(VQA), 이미지 캡션 생성, 구조화된 이미지 생성 및 외부로의 그림 그리기, 로봇 공학 등 다양한 하위 응용 분야에 잠재력을 가지고 있습니다.

SGG의 도입 이후로 [27], 이 문제는 검출(detection) 관점에서 다루어져 왔습니다. 즉, 객체를 감지하기 위해 바운딩 박스를 사용하고 이후에 그들 간의 관계를 추론합니다.

그림 1: 장면 그래프 생성 (a. SGG 작업) vs. 팬옵틱 장면 그래프 생성 (b. PSG 작업).

기존의 SGG 작업(a)은 바운딩 박스 기반의 레이블을 사용하는데, 이는 종종 부정확합니다. 바운딩 박스로 덮힌 픽셀이 주석이 달린 클래스에 속하지 않을 수도 있고, 배경 정보를 완전히 포착하지 못할 수 있습니다. 반면에 제안된 PSG 작업(b)은 더 포괄적이고 깨끗한 장면 그래프 표현을 제공합니다. 객체의 보다 정확한 위치 지정 및 배경(스텁(stuff)이라고도 함)과의 관계를 포함합니다. 스텝은 나무와 보도와 같은 배경을 의미합니다.

객체들 간의 쌍(pairwise) 관계를 예측하기 위해 바운딩 박스 기반 패러다임은 문제 해결에 적합하지 않으며, 오히려 분야의 발전을 방해하는 여러 가지 문제를 일으킬 것이라 주장합니다. 첫째로, 현재의 데이터셋에서 레이블로 표시된 대로 바운딩 박스는 객체의 대략적인 로컬라이제이션(localization)만 제공하며, 종종 서로 다른 객체나 카테고리에 속하는 잡음/모호한 픽셀을 포함하고 있습니다 (그림 1-a의 두 사람의 바운딩 박스 참조). 둘째로, 바운딩 박스는 일반적으로 이미지의 전체 장면을 커버하지 못합니다. 예를 들어, 그림 1-a의 보도 영역은 문맥을 이해하는 데 중요하지만 완전히 무시됩니다. 셋째로, 현재의 SGG 데이터셋은 종종 중복되는 클래스와 정보를 포함하고 있으며, 그림 1-a의 여성의 머리카락과 같이 대부분의 경우 사소하다고 여겨집니다. 또한, 일관되지 않고 중복된 레이블도 현재의 데이터셋에서 관찰되며, 그림 1-a의 나무와 벤치는 여러 번 레이블이 지정되고, 일부 추가적인 주석은 그래프에 기여하지 않습니다 (고립된 노드 참조). 이러한 레이블을 학습에 사용하면 모델이 혼란스러워질 수 있습니다.

이상적으로는 객체의 미관적이고 정확한 위치 지정이 이루어져야 하며, 장면 그래프는 이미지의 중요한 영역과 관계에 초점을 맞추면서도 장면을 이해하기에 충분히 포괄적이어야 합니다. 바운딩 박스 대신 팬옵틱 세그멘테이션(segmentation) [33] 레이블이 장면 그래프 구축에 더 나은 선택일 것이라 주장합니다. 이를 위해 팬옵틱 세그멘테이션을 기반으로 한 장면 그래프 표현을 생성하는 새로운 문제인 팬옵틱 장면 그래프 생성(Panoptic Scene Graph Generation, PSG)을 소개합니다. PSG는 강성한 바운딩 박스 대신 팬옵틱 세그멘테이션을 기반으로 한 장면 그래프 표현을 생성하는 것을 목표로 합니다. 연구 진행 상황을 추적하기 위해 COCO [43]와 Visual Genome (VG) [35]를 기반으로 한 새로운 PSG 데이터셋을 생성했습니다. 이 데이터셋은 총 49,000장의 잘 주석이 달린 이미지를 포함하고 있습니다. 우리는 COCO의 객체 주석 체계인 133개 클래스를 따르며, 이는 80개의 물체 클래스와 53개의 스텁(stuff, 배경) 클래스로 구성됩니다. 관계(predicate)를 구성하기 위해 VG 기반 데이터셋인 VG-150 [64], VrR-VG [42], GQA [24] 등을 철저히 조사하고, 중복이 최소화되고 의미론적으로 충분히 포괄적인 56개의 관계 클래스를 요약했습니다. 데이터셋 예시는 그림 1-b를 참조하십시오. 그림 1을 보면, 팬옵틱 장면 그래프 표현(팬옵틱 세그멘테이션 및 장면 그래프를 모두 포함)이 이전 장면 그래프 표현보다 훨씬 정보를 많이 담고 있으며 일관성이 있음을 알 수 있습니다.

성능 평가를 위해, 우리는 네 가지의 클래식한 SGG 방법 [64,71,58,54]을 고전적인 팬옵틱 세그멘테이션 프레임워크 [32]에 통합하여 네 가지의 이 단계 모델을 구축했습니다. 또한, 효율적인 Transformer 기반 검출기인 DETR [4]을 PSGTR이라는 하나의 PSG 모델로 변형하여 사용했는데, PSGTR은 PSG 작업에 효과적임을 입증했습니다. PSGTR을 확장하여 객체와 관계를 별도로 쿼리 형식으로 모델링하고, 프롬프트와 유사한 상호작용 메커니즘을 추가한 PSGFormer라는 또 다른 하나의 베이스라인을 제공합니다. 일 단계 모델과 이 단계 모델의 포괄적인 비교는 우리의 실험에서 논의됩니다.

요약하면, 우리는 다음과 같은 SGG 커뮤니티에 기여합니다:

• 새로운 문제와 데이터셋: 우리는 현재 SGG 연구에서 발생하는 여러 문제들을 논의하고, 특히 기존 데이터셋과 관련된 문제들을 주로 다루었습니다. 이를 해결하기 위해, SGG를 팬옵틱 세그멘테이션과 결합한 새로운 문제를 제시하고, 고품질 주석이 달린 대규모 PSG 데이터셋을 생성했습니다.

• 포괄적인 벤치마크: 강력한 이 단계 및 일 단계 PSG 베이스라인을 구축하고, 새로운 데이터셋에서 포괄적으로 평가하여 PSG 작업을 탄탄하게 성립시켰습니다. 단순화된 훈련 패러다임을 가지고 있는 일 단계 모델은 데이터셋에서 경쟁력 있는 결과를 얻을 수 있어 PSG에 큰 잠재력을 보여줍니다.

기존의 장면 그래프 생성(SGG) 방법은 객체 감지와 쌍별 관계 예측으로 구성된 이단계 파이프라인에 의해 주도되었습니다. 바운딩 박스를 사용하여 초기 연구는 조건부 랜덤 필드 [27,11]를 사용하여 관계를 예측하거나 관계 예측을 분류 문제로 캐스팅하는 방법 [75,34,49]을 사용했습니다. 지식 그래프 임베딩에서 영감을 받아 VTransE [74]와 UVTransE [25]는 명시적인 관계 모델링을 위해 제안되었습니다. 이에 이어 다양한 변형들이 연구되었는데, 예를 들어 RNN과 그래프 기반 모델링 [64,71,58,67,44,9,38], 에너지 기반 모델 [54], 외부 지식 [58,18,69,70,46], 그리고 최근에는 언어 감독 [77,68] 등이 있습니다. 최근의 연구는 SGG 데이터셋과 관련된 문제에 주목하고 있으며, 예를 들어 관계의 긴꼬리 분포 [57,12] 등을 다루고 있습니다.

Table 1: 클래식한 SGG 데이터셋과 PSG 데이터셋의 비교. #PPI는 이미지 당 관계 개수를 나타냅니다. DupFree는 중복된 객체 주석이 정리되었는지 여부를 확인합니다. Spvn은 객체가 바운딩 박스나 세그멘테이션으로 표현되었는지 여부를 나타냅니다.

기존의 SGG 연구에서는 관련 없는 관계에 대한 과도한 예측 [42]과 바운딩 박스의 부정확한 로컬라이제이션 [29]과 같은 문제들이 있습니다. 특히, 최근의 연구 [29]는 장면 그래프를 생성하고 시맨틱 세그멘테이션 마스크를 예측하는 것을 동시에 수행하는 SGG 모델을 훈련시킴으로써 개선을 이끌어낼 수 있다는 것을 보여주었습니다. 이는 우리의 연구에 영감을 주었습니다. 우리의 연구에서는 팬옵틱 세그멘테이션을 기반으로 한 장면 그래프 생성을 보다 체계적인 방식으로 연구하기 위해 새로운 문제를 정의하고 새로운 벤치마크를 구축하였습니다. 또한, 밀접한 관련 주제인 사람-객체 상호작용(HOI) [19]도 SGG와 유사한 목표를 가지고 있으며, 즉 이미지에서 주요한 관계를 탐지하는 것입니다. 하지만, HOI 작업은 모델이 주로 종합적인 장면 이해에 중요한 역할을 하는 객체 간의 정보를 무시하고 인간과 관련된 관계만 탐지하도록 제한합니다. 그럼에도 불구하고, 많은 HOI 방법들이 SGG 작업에 적용될 수 있으며 [17,28,6,60,40,78,61,22,39,15,30,45,55,23,72,59,73], 그 중 일부는 우리의 PSG 기준선에 영감을 주었습니다 [31,79].

장면 그래프 데이터셋은 이미지에 장면 그래프 정보를 포함하는 데이터셋입니다. 이는 객체, 속성 및 객체 간의 관계를 포함합니다. 비주얼 지오메트리 (VG) 데이터셋과 VG-150 및 VrR-VG와 같은 변형은 장면 그래프 생성 (SGG) 연구에서 널리 사용됩니다.

그러나 VG 데이터셋은 여러 가지 단점들이 지적되었습니다. 주요 문제 중 하나는 객체 및 술어 클래스의 수가 매우 많다는 것으로, 이로 인해 데이터셋 처리가 어려워집니다. 이를 해결하기 위해 VG-150는 VG에서 가장 빈도가 높은 150개의 객체 클래스와 50개의 술어 클래스를 선택하여 만들어졌습니다. VrR-VG는 VG-150에서 통계적 우선순위에 따라 많은 술어가 추정 가능하다고 판단하여 시각적으로 관련성이 있는 술어 클래스만을 보유하도록 데이터셋을 재정리했습니다.

SGG 데이터셋의 또 다른 과제는 바운딩 박스 기반의 객체 지정입니다. 바운딩 박스는 정확한 지역화에 한계가 있으며, 종합적인 지정(배경 포함)이 불가능하며, 하찮은 정보를 포착하는 경향이 있으며(예: 사람-머리와 같은 관계), 중복된 지정을 가질 수 있습니다. 이러한 문제는 데이터셋의 품질을 저하시키며, 연구의 진전을 방해합니다.

이러한 문제를 극복하기 위해 제안된 팬옵틱 장면 그래프 (PSG) 데이터셋은 바운딩 박스 대신 정확하고 포괄적인 팬옵틱 세그멘테이션을 사용하여 객체 지정을 수행합니다. 팬옵틱 세그멘테이션은 픽셀 수준의 정확성을 제공하며, 배경을 포함한 종합적인 지정이 가능합니다. PSG 데이터셋은 또한 단어 간의 종합적인 커버리지와 중복 없이 술어 사전을 정의하여 장면 그래프 문제를 더 잘 구성합니다.

총론적으로 PSG 데이터셋은 기존 데이터셋의 한계를 극복하여 장면 그래프 데이터셋의 품질을 향상시킵니다. 팬옵틱 세그멘테이션을 사용하여 정확하고 종합적인 객체 지정을 제공하며,

팬옵틱 장면 그래프 생성 (PSG)은 전통적인 장면 그래프 생성 (SGG)과는 다른 방식으로 객체를 지정합니다. PSG 작업에서는 각 객체를 더 세밀한 팬옵틱 세그멘테이션으로 지정합니다. 이를 위해 이미지는 일련의 마스크로 분할됩니다. 각 마스크는 객체에 대응하며 클래스 레이블과 관련됩니다. 객체 간의 관계도 예측됩니다. 마스크는 겹치지 않으며, 즉, 마스크들은 상호 배타적인 영역입니다. PSG 작업은 다음과 같은 분포를 모델링합니다:

Pr(G | I) = Pr(M, O, R | I)

여기서 I는 입력 이미지를 나타내고, G는 원하는 장면 그래프입니다. 장면 그래프는 각 객체에 대응하는 바운딩 박스 B = {b1, ..., bn}와 레이블 O = {o1, ..., on}, 그리고 객체 간의 관계를 나타내는 집합 R = {r1, ..., rl}로 구성됩니다. 각각의 bi는 바운딩 박스의 좌표를 나타내며, oi는 모든 객체 클래스의 집합 CO에 속하고, ri는 모든 관계 클래스의 집합 CR에 속합니다.

따라서 PSG 작업은 팬옵틱 세그멘테이션을 사용하여 객체를 지정하고, 관계를 예측하는 방식으로 이루어집니다. 이를 통해 더 정확하고 세부적인 객체 지정이 가능해지며, 장면 그래프의 품질을 향상시킬 수 있습니다.

PSG 데이터셋 구축을 위해 다음 세 가지 주요 단계를 거쳤습니다. 자세한 내용은 부록에서 찾아볼 수 있습니다.

1단계: 그룹화된 COCO 및 VG 퓨전: 팬옵틱 세그멘테이션과 관계 주석이 모두 포함된 데이터셋을 만들기 위해, COCO와 VG 데이터셋의 교집합에 속하는 48,749개의 이미지를 자동으로 그룹화한 과정을 거쳤습니다. 구체적으로, 객체 카테고리 매칭 프로세스를 사용하여 COCO의 세그멘테이션과 VG의 바운딩 박스를 매칭시켰으며, 이를 통해 VG의 관계술어 일부를 COCO의 세그멘테이션 쌍에 적용했습니다. VG와 COCO의 레이블 시스템 및 위치 주석 간의 불일치로 인해, 자동 생성된 데이터셋은 매우 노이즈가 많으며 추가적인 정제가 필요합니다.

2단계: 간결한 술어 사전: COCO 카테고리의 적절한 세분성을 참고하여, 초기 노이즈가 많은 PSG 데이터셋 및 VG-150, VrR-VG, GQA를 포함한 모든 VG 기반 데이터셋에서 일반적인 술어를 기준으로 56개의 중요한 관계를 신중하게 선정했습니다. 선택된 56개의 술어는 최대한 독립적이며 (예: "위에/위에 있는"만 유지하고 "아래에"와 같은 술어는 포함하지 않음), 데이터셋에서 가장 일반적인 경우를 포함합니다.

3단계: 엄격한 주석 프로세스: 노이즈가 많은 PSG 데이터셋을 기반으로, 주석 작업자에게 1) 잘못된 세트를 걸러내고, 2) 사전 정의된 56개 술어를 사용하여 객체-객체뿐만 아니라 객체-배경 및 배경-배경 사이의 관계를 보완할 것을 요구했습니다. 술어 간의 모호성을 방지하기 위해, 보다 정확한 술어(예: "주차된"과 같은)가 적용 가능할 때 "위에/안에"와 같은 일반적인 관계를 사용하여 주석을 달지 말 것을 엄격히 요구했습니다. 이 규칙으로 인해 PSG 데이터셋은 모델이 장면을 더 정확하고 강조하여 이해할 수 있게 됩니다.

품질 관리: PSG 데이터셋은 전문적인 데이터셋 구축 과정을 거쳤습니다. 주 저자들은 먼저 1,000개의 이미지를 주석 달기 위해 자세한 문서를 작성한 후, 한 달 동안 프로페셔널한 주석 회사(SenseTime 후원)를 고용하여 훈련 세트를 주석 달았습니다. 각 이미지는 두 명의 작업자에 의해 주석이 달리고 한 명의 주 관리자가 검토했습니다. 모든 테스트 이미지는 저자들이 주석을 달았습니다.

요약: PSG 데이터셋 생성의 새롭고 효과적인 절차를 통해 강조할 수 있는 몇 가지 장점: 1) COCO 데이터셋 [43]의 픽셀 단위 팬옵틱 세그멘테이션에 기반한 좋은 지정 주석, 2) 133개의 객체(물체와 배경을 포함한 것)와 56개의 술어로 구성된 명확한 범주 시스템, 그리고 3) 객체-객체, 객체-배경 및 배경-배경 간의 강조할 만한 관계에 특별히 주의를 기울인 엄격한 주석 프로세스로부터의 정확하고 포괄적인 관계 주석. 이러한 장점들은 섹션 2에서 논의된 전통적인 장면 그래프 데이터셋의 악명 높은 단점들을 해결합니다 [37].

평가 및 측정

이 섹션에서는 PSG 작업에 대한 평가 프로토콜을 소개합니다. 클래식 SGG 작업 [65,5]의 설정을 따라 PSG 작업은 술어 분류 (해당 경우에만 해당) 및 장면 그래프 생성 (주요 작업) 두 가지 하위 작업으로 구성되어 PSG 모델을 평가합니다. 술어 분류 (PredCls)는 참값 객체 레이블과 위치 정보를 기반으로 장면 그래프를 생성하는 것을 목표로 합니다. 이 작업은 세그멘테이션 성능의 간섭 없이 관계 예측 성능을 연구하는 것을 목표로 합니다. 이 측정 항목은 4.1절의 두 단계 PSG 모델에만 적용될 수 있습니다. 왜냐하면 한 단계 모델은 주어진 세그멘테이션을 활용하여 장면 그래프를 예측할 수 없기 때문입니다. 장면 그래프 생성 (SGDet)은 처음부터 장면 그래프를 생성하는 것을 목표로 합니다. 이는 PSG 작업의 주요 결과입니다.

PSG 베이스라인

포괄적인 PSG 벤치마크를 구축하기 위해, 고전적인 SGG 작업에서 사용된 프레임워크를 참고하여, 두 가지 유형의 베이스라인을 준비했습니다: 두 단계 베이스라인과 한 단계 베이스라인입니다.

4.1 두 단계 PSG 베이스라인

과거의 대부분의 SGG 접근 방식은 두 단계로 문제를 처리합니다. 먼저, Faster-RCNN [51]과 같은 미리 훈련된 디텍터를 사용하여 객체 감지를 수행합니다. 그런 다음, 감지된 객체 간의 관계를 예측합니다. Figure 3은 PSG 작업을 위한 두 단계 베이스라인을 설명합니다. 이 접근 방식은 다음과 같은 단계를 포함합니다:

1. 미리 훈련된 Panoptic FPN: 베이스라인 모델은 Panoptic FPN [32]과 같은 미리 훈련된 팬옵틱 세그멘테이션 모델을 활용하여 초기 객체 특성, 마스크, 및 클래스 예측을 추출합니다. 각 객체에 대해 Panoptic FPN은 가장 작은 경계 상자로 객체 특성을 잘라낼 수 있는 세그멘테이션 마스크를 생성합니다. 관련 객체들의 합집합은 관계 특성을 생성할 수 있습니다.

2. 관계 예측 모듈: 추출된 객체 및 관계 특성은 고전적인 시각적 그래프 생성 방법(IMP [64], MOTIFS [71], VCTree [58], GPSNet [44] 등)의 관계 예측 모듈을 사용하여 최종 시각적 그래프를 예측합니다. 이 모듈들은 관계 트리플렛을 기반으로 시각적 그래프를 생성합니다.

이러한 방식을 따르면, 고전적인 SGG 방법을 최소한의 수정으로 PSG 작업에 적용할 수 있습니다. 미리 훈련된 팬옵틱 세그멘테이션 모델은 초기 객체 특성과 마스크를 제공하고, 관계 예측 모듈은 이러한 특성을 기반으로 시각적 그래프를 생성합니다.

이 설명이 두 단계 PSG 베이스라인의 개념을 명확하게 해주었기를 바랍니다.

2 / 2

Fig. 3에서는 Panoptic FPN을 사용한 두 단계 PSG 베이스라인에 대해 설명합니다. a) 첫 번째 단계에서는 각 객체에 대해 Panoptic FPN [32]이 가장 작은 경계 상자로부터 세그멘테이션 마스크를 생성하여 객체 특성을 자릅니다. 관련된 객체의 합집합은 관계 특성을 생성할 수 있습니다. b) 두 번째 단계에서는 추출된 객체 및 관계 특성이 고전적인 시각적 그래프 생성 방법(IMP [64], MOTIFS [71], VCTree [58], GPSNet [44] 등)의 관계 모델에 입력으로 사용되어 관계 트리플렛을 예측합니다. 이를 통해 고전적인 SGG 방법을 최소한의 수정으로 PSG 작업에 적용할 수 있습니다. 이러한 방식으로 PSG 작업을 해결할 수 있습니다. 이렇게 하면 식 1에서 식 3로 고전적인 SGG 방법을 PSG 작업에 적용할 수 있습니다.

4.2 PSGTR: 한 단계 PSG 베이스라인

고전적인 밀집 예측 모델 (예: Faster-RCNN [51])과 달리, 변형기 기반 아키텍처는 유연한 입력 및 출력 사양을 지원합니다. DETR [4]과 HOI 작업에 대한 확장을 기반으로, 우리는 PSGTR이라는 한 단계 PSG 방법을 자연스럽게 설계했습니다. PSGTR은 트리플과 위치 정보를 동시에 예측할 수 있으며, 식 2와 같이 직접 모델링할 수 있습니다.

트리플 쿼리 학습 블록: Fig. 4에서 보여지는 PSGTR은 먼저 CNN 백본에서 이미지 특성을 추출한 다음, 트랜스포머 인코더-디코더에 쿼리와 위치 인코딩을 함께 제공합니다. 여기서 우리는 쿼리가 시각적 그래프 트리플렛의 표현을 학습하도록 기대합니다. 각 트리플 쿼리에 대해 주어진 Feed Forward Networks (FFNs) 세 개를 사용하여 주어어/술어/목적어 예측을 추출하고, 주제와 객체에 대한 두 개의 팬옵틱 헤드로 세그멘테이션 작업을 완료합니다.

PSG 예측 블록: 모델을 훈련하기 위해, DETR의 Hungarian matcher [36]를 트리플 헝가리안 matcher로 확장합니다. 트리플 쿼리 Ti ∈ QT와 실제 트리플 G를 매칭하기 위해, 트리플 쿼리에서 예측된 모든 출력(즉, T에서 예측된 모든 출력)을 사용합니다. 여기에는 주어 T ̈S, 술어 T ̈R, 목적어 T ̈O의 클래스 및 주제 T ̃S 및 객체 T ̃O의 위치가 포함됩니다. 따라서 트리플 매칭 비용 Ctm은 클래스 매칭 비용 Ccls와 세그멘트 매칭 비용 Cseg의 조합으로 설계됩니다.

식 4에서는 트리플 매칭 비용 Ctm이 클래스 매칭 비용 Ccls와 세그멘트 매칭 비용 Cseg의 조합으로 정의됩니다. 여기서 C ̃T,G는 세그멘트 매칭 Cseg의 합계이고, C ̈T,G는 클래스 매칭 Ccls의 합계입니다. 주어진 주제/목적어 T ̃k와 실제 주제/목적어 G ̃k의 세그멘트 매칭 비용 Cseg과 주어진 주제/술어/목적어 T ̈k와 실제 주제/술어/목적어 G ̈k의 클래스 매칭 비용 Ccls를 포함합니다.

이렇게 매칭이 완료되면, 총 손실 Ltotal은 식 5를 적용하여 계산할 수 있습니다. 여기서 σ는 각 트리플 쿼리 Ti ∈ QT를 가장 가까운 실제 트리플에 대응시키는 매핑 함수입니다. 트리플 쿼리 집합 QT는 |QT|개의 트리플 쿼리를 수집합니다. 최적화 목적은 다음과 같습니다:

|QT | σˆ = arg max X Ctm �Ti, Gσ(i) . (5) σ i=1

매칭이 완료되면, 총 손실 Ltotal은 레이블에 대한 교차 엔트로피 손실 Lcls와 세그멘테이션에 대한 DICE/F-1 손실 [47]인 Lcls를 적용하여 계산할 수 있습니다.

Fig. 4에서는 PSGTR 방법의 구조를 보여줍니다. 한 단계 모델은 a) CNN과 위치 인코딩으로 추출된 특성과 트리플을 나타내기 위한 일련의 쿼리를 입력으로 사용합니다. b) 쿼리 학습 블록은 이미지 특성을 트랜스포머 인코더-디코더로 처리하고 쿼리를 사용하여 트리플 정보를 표현합니다. 그리고 c) PSG 예측 블록은 간단한 FFN을 사용하여 주제/객체/술어 클래스를 생성하고 팬옵틱 세그멘테이션을 위해 팬옵틱 헤드를 사용하여 트리플 예측을 완성합니다.

이렇게 PSGTR 방법에 대한 설명이 완료되었습니다.

4.3 PSGFormer: 다른 한 단계 PSG 베이스라인

PSGFormer은 Section 4.2에서 설명한 PSGTR 베이스라인을 기반으로 하며, HOI 방법[31]을 확장하여 명시적인 관계 모델링과 프롬프팅과 유사한 매칭 메커니즘을 갖춘 대체 한 단계 PSG 베이스라인입니다. 모델 다이어그램은 Figure 5에 나와 있으며, 다음과 같이 자세히 설명하겠습니다.

객체 및 관계 쿼리 학습 블록은 객체 감지와 세그멘테이션과 같은 고전적인 객체 지향 작업과 비교하여 PSG 및 SGG 작업의 가장 큰 특징은 관계 예측에 대한 추가적인 요구사항입니다. PSG 작업에서 관계 모델링은 두 개의 객체 쌍에서 나온 특성에 의존하는 반면, PSGTR은 트리플릿 내에서 객체와 관계를 함께 암묵적으로 모델링합니다. 중요한 관계 모델링은 심각하게 다루어지지 않았습니다.

PSGFormer에서는 PSG 작업에서의 관계 쿼리 Ri ∈ QR과 개별적으로 객체 쿼리 Oi ∈ QO를 명시적으로 모델링합니다. 이를 통해 객체 쿼리는 개체 (예: 사람과 전화기)에 특별한 주의를 기울이고, 관계 쿼리는 사진에서 관계가 발생하는 영역 (예: 사람이 전화기를 쳐다보는)에 초점을 맞출 수 있도록 합니다. Figure 5의 PSGFormer와 유사하게, 객체 쿼리와 관계 쿼리는 CNN 특성과 위치 인코딩과 함께 transformer 인코더에 입력되지만, 해당하는 디코더, 즉 객체 디코더 또는 관계 디코더로 디코딩됩니다. 이를 통해 쿼리가 해당하는 표현을 학습할 수 있습니다.

PSGFormer의 객체 및 관계 쿼리 매칭 블록에서, 각 객체 쿼리는 FFN을 사용하여 객체 예측과 패노프틱 헤드를 사용하여 마스크 예측을 생성하고, 각 관계 쿼리는 관계 예측을 생성합니다. 그러나 객체 쿼리와 관계 쿼리의 병렬 프로세스로 인해 서로 다른 쿼리 유형간의 상호의존성이 부족하여 트리플릿이 형성되지 않습니다. 트리플릿을 구성하기 위해 객체 및 관계 쿼리를 연결하기 위해, HOTR [31]의 설계에서 영감을 받아 프롬프팅 유사한 쿼리 매칭 블록을 구현합니다.

쿼리 매칭 블록은 트리플릿 구성 작업을 프롬프트가 있는 문제로 모델링합니다. 즉, 관계를 프롬프트로 지정함으로써, 해당하는 개체 쿼리에서 제공되는 적합한 개체 쌍이 선택될 것으로 예상하여 완전한 주어-술어-목적어 트리플릿이 생성될 수 있도록 합니다. 따라서 주어 선택기와 목적어 선택기라는 두 개의 선택기가 필요합니다.

관계 쿼리 Ri ∈ QR을 프롬프트로 주어진 경우, 주어 선택기와 목적어 선택기는 완전한 트리플릿을 형성하기 위해 가장 적합한 후보를 반환해야 합니다. 우리는 객체 쿼리와 주어진 관계 쿼리 간의 가장 표준적인 코사인 유사도를 사용하고 가장 높은 유사도를 선택하여 주어와 목적어 후보를 결정합니다. 또한 주어 선택기와 목적어 선택기는 의미적 유사도보다 객체와 관계 쿼리 간의 연관성 수준에 의존해야 함에 유의해야 합니다.

PSGFormer 모델에서 object queries는 fS와 fO라고 표시된 두 개의 FFN을 통해 특정 정보를 추출하게 됩니다. 이를 통해 object queries로부터 구분 가능한 주어와 목적어의 표현을 얻을 수 있습니다. 이 방식을 통해 식 7에서 주어 집합 S가 생성되는데, 각 주어 Si는 i번째 관계 쿼리 Ri와 대응됩니다. 마찬가지로, 객체 집합 O도 생성됩니다.

S = {Si | Si = argmax(fS(Oj) · Ri), Oj ∈ QO, Ri ∈ QR} (7)

이 단계에서 주어 집합 S와 객체 집합 O가 주어진 subject와 object 선택기에 의해 준비되며, i번째 subject query Si와 i번째 object Oi는 i번째 관계 쿼리 Ri에 대응됩니다. 마지막으로, 모든 일치하는 triplet T를 식 8에 따라 얻을 수 있습니다.

T = {(Si, Ri, Oi) | Si ∈ S, Ri ∈ R, Oi ∈ O} (8)

쿼리 매칭 과정은 prompt와 유사한 방식으로 해석될 수 있으며, 또는 교차 어텐션 메커니즘으로 해석될 수도 있습니다. prompt-like 해석에서 관계 쿼리(Q)는 주어/목적어 표현들 중에서 주의를 많이 받는 관계를 찾는 것을 목표로 합니다. 이때 주어/목적어 표현은 키(K) 역할을 하고, 해당 표현에 의해 예측된 주어/목적어 레이블은 값(V)로 사용되며, QKV 어텐션 모델은 선택된 키의 레이블을 출력합니다. 그림 5c는 이 해석을 일반적으로 나타내고 있습니다.

섹션에서는 논문에서 소개된 모든 PSG 방법의 결과를 보고합니다. 구현 세부 정보는 부록에서 확인할 수 있으며, 모든 코드는 MMDe- tection [7]을 기반으로 개발된 OpenPSG 코드베이스에 통합되어 있습니다. 대부분의 two-stage SGG 구현은 MMSceneGraph [62]와 Scene-Graph-Benchmark.pytorch [56]을 참고하였습니다.

5.1 주요 결과 테이블 2는 Sec. 4.1, Sec. 4.2 및 Sec. 4.3에서 언급된 모든 방법들의 PSG 데이터셋에서의 scene graph 생성 성능을 보고합니다. Fig. 6은 PQ를 사용하여 패노픽 세그멘테이션 결과를 보고하며, 두 가지 예제의 세그멘테이션 결과 및 삼중 리스트 형식으로 예측된 scene graph를 시각화합니다.

테이블 2: 모든 베이스라인 및 PSGFormer의 비교 결과입니다. Recall (R) 및 평균 재현율 (mR)이 보고되었습니다. IMP [64] (CVPR'17), MOTIFS [71] (CVPR'18), VC-Tree [58] (CVPR'19) 및 GPSNet [44] (CVPR'20)은 모두 SGG 작업에서 기원하였으며 PSG 작업에 적용되었습니다. ResNet-50 [20] 및 ResNet-101 [20]의 다른 백본이 사용되었습니다. 주어/목적어 분류 작업은 one-stage PSG 모델에는 적용되지 않으므로, 해당 결과는 '−'로 표시되었습니다. 모델은 기본적으로 12개 에포크로 훈련되었습니다. †는 해당 모델이 60개 에포크로 훈련되었음을 나타냅니다.

두 단계 기반의 PSG 베이스라인은 첫 번째 단계의 성능에 크게 의존합니다. 특히, 술어 분류 작업 (PredCls)은 두 단계 모델에만 적용되며, 제공된 실측 세그멘테이션은 세트의 트리플트 예측 성능을 크게 향상시킵니다. 예를 들어, 가장 고전적인 방법인 IMP라도 30% 이상의 R@20을 달성할 수 있으며, 이는 PSGTR의 28.4%보다 높은 수준입니다. 이 현상은 우수한 파노픽트 세그멘테이션 성능이 PSG 작업에 자연스럽게 도움이 된다는 것을 나타냅니다. 또한, IMP의 SGDet 작업에서의 성능이 거의 절반으로 줄어드는 현상은 위의 추측을 강화하는 추가적인 증거로 볼 수 있습니다.

일부 SGG 기법이 PSG에는 효과적이지 않을 수 있음 표 2에서는 일부 두 단계 베이스라인 (즉, IMP, MOTIFS, VCTree)의 결과가 SGG 작업에서의 성능과 비례한다는 것을 보여줍니다. 이는 두 단계 모델 (MOTIFS 및 VCTree)의 장점이 PSG 작업으로 전이 가능하다는 것을 나타냅니다. 그러나 또 다른 두 단계 베이스라인인 GPSNet은 PSG 작업에서 MOTIFS와 VCTree의 성능을 능가하는 것 같지 않습니다. GPSNet의 주요 이점은 술어의 방향을 명시적으로 모델링한다는 점입니다. 이러한 설계는 많은 술어가 방향이 명확한 VG 데이터셋에서 효과적일 수 있지만 (예: hair-of-man의 of, man-has-head의 has), PSG 데이터셋은 이러한 술어를 제거했으므로 기대한 대로 효과적이지 않을 수 있습니다.