그래프 신경망과 멀티 모달 맥락 정보를 이용한 장면 그래프 생성

정가영, 김인철
경기대학교 컴퓨터과학과
email: jgyy4775@kyonggi.ac.kr, kic@kyonggi.ac.kr

Scene Graph Generation with Graph Neural Network and Multimodal Context

Ga-Young Jung, In-cheol Kim
Department of Computer Science, Kyonggi University

요 약

본 논문에서는 입력 영상에 담긴 다양한 물체들과 그들 간의 관계를 효과적으로 탐지하여, 하나의 장면 그래프로 표현해내는 새로운 심층 신경망 모델을 제안한다. 제안 모델에서는 물체와 관계의 효과적인 탐지를 위해, 합성 곱 신경망 기반의 시각 맥락 특징들뿐만 아니라 언어 맥락 특징들을 포함하는 다양한 멀티 모달 맥락 정보들을 활용한다. 또한, 제안 모델에서는 관계를 맺는 두 물체 간의 상호 의존성이 그래프 노드 특징값들에 충분히 반영되도록, 그래프 신경망을 이용해 맥락 정보를 임베딩한다. 본 논문에서는 Visual Genome 벤치마크 데이터 집합을 이용한 비교 실험들을 통해, 제안 모델의 효과와 성능을 입증한다.

1. 서론

심층 영상 이해(Deep Image Understanding)를 요구하는 대표적인 인공지능 및 컴퓨터 비전 문제 중 하나로, 장면 그래프 생성(Scene Graph Generation) 문제가 있다. 장면 그래프는 한 영상에 담긴 장면을 그래프 형태로 표현한 것으로서, 그래프를 구성하는 각 노드(node)는 영상 속의 물체(object)를, 각 간선(edge)은 물체들 간의 관계(relationship)를 각각 나타낸다. 따라서 하나의 장면 그래프는 해당 영상의 장면을 설명하는 <주어 물체(subject)-관계 서술자(relationship predicate)-목적어 물체(object)> 형태의 사실 집합(fact set)으로 볼 수 있다. 즉 장면 그래프 생성 문제는 입력 영상에 관한 심층 이해의 결과로 해당 영상의 장면을 표현하는 하나의 지식 그래프(knowledge graph)를 생성하는 문제이다.

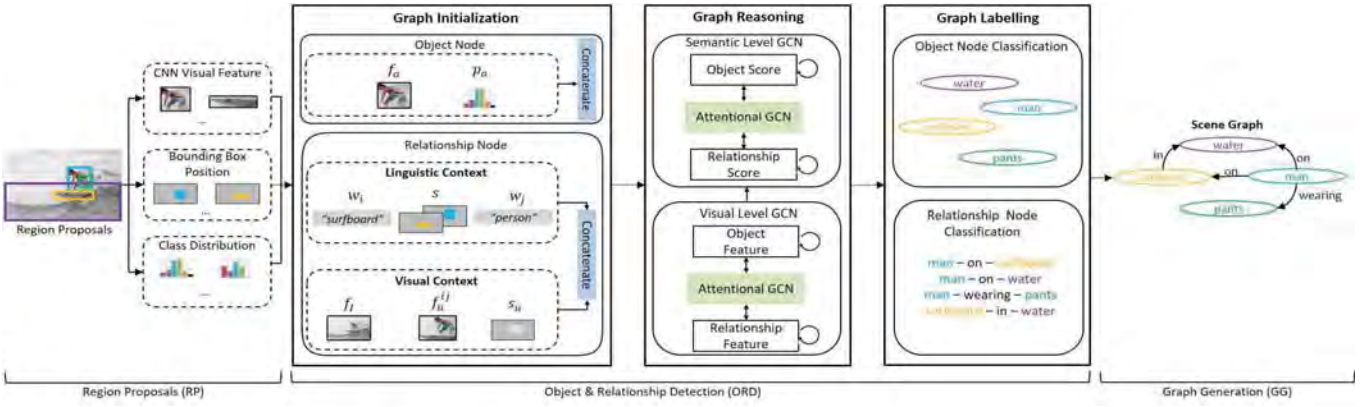
(그림 1)은 일반적인 장면 그래프 생성 과정을 보여주고 있다. 장면 그래프 생성을 위해서는 영상 속 물체 탐지(object detection)뿐만 아니라, 물체들 간의 관계 탐지(relationship detection)도 필수적으로 요구된다. 물체 탐지는 종래의 컴퓨터 비전 분야에서 많이 연구된 문제이나, 관계 탐지는 최근에 와서야 관심을 모으고 있는 문제로서 아직은 연구의 초기 단계에 머물고 있다. 영상 속의 두 물체들 간에 가질 수 있는 관계들은 매우 다양하다. 일반적으로 장면 그래프 생성 연구에서 많이 다루어지는 물체들 간의 관계에는 공간 관계(spatial relationship)와 의미적 관계(semantic relationship)가 있다. 공간 관계는 'on', 'next to', in front of'와 같이 영상 안에 놓인 물체들 간의 상대적 위치 관계를 나타내며, 반면에. 의미적 관계는 'wearing', 'eating', 'holding'과 같이 한 물체가 다른 물체에 행하는 행위와 연관된 관계이다.

* 본 연구는 과학기술정보통신부 및 정보통신기술진흥센터의 대학ICT연구센터육성지원사업의 연구결과로 수행되었음 (IITP-2017-0-01642)

합성곱 신경망(Convolutional Neural Network, CNN)을 이용한 물체 탐지 기술은 현재 높은 수준에 도달해 있으나, 아직은 물체 식별과 영역 탐지에 오류가 있을 수 있다. 이는 곧 관계 탐지에 기초가 되는 두 물체의 식별에 불확실성과 오류가 있을 수 있다는 것을 의미한다. 비록 관계를 맺는 두 물체의 식별이 매우 분명하다고 하더라도, 두 물체 간에 가능한 관계의 수 또한 많기 때문에 물체 간의 관계를 정확히 판별하는 일은 결코 쉬운 일이 아니다. 더욱이 일반적으로 특정 관계와 그 관계를 맺을 수 있는 두 물체의 유형에는 다양한 의미적 제약이 존재한다. 예컨대, (그림 1)의 예에서, <man-wearing-shoes>의 관계는 가능하지만, <man-wearing-racket>이나 <shoes-wearing-man>과 같은 관계는 불가능하다는 것을 인간은 상식적으로 잘 알고 있다. 이러한 문제의 특성을 잘 고려하여, 영상으로부터 정확한 장면 그래프를 효과적으로 생성할 수 있는 모델의 개발이 필요하다.

(그림 1) 장면 그래프 생성 예시

장면 그래프 생성에 관한 많은 기존의 연구들[1, 2]에서는 장면 그래프 생성에 필요한 물체 탐지와 관계 탐지를 위해 합성 곱 신경망(CNN)을 통해 영상에서 추출한 시각 특징만을 활용하였다. 대신 [2]의 연구에서는 그래프 신경망(Graph Neural Network)의 하나인 aGCN(attentional Graph Convolutional Network)을 통해, 장면 그래프를 구성하는 이웃 노드들의 맥락 정보를 각 노드의 특징값에 반영될 수 있도록 모델을 설계하였다. 한편, 시각적 관계



(그림 2) 제안 모델의 구조

탐지(Visual Relationship Detection)에 관한 [3]의 연구에서는 두 물체 간의 효과적인 관계 탐지를 위해 영상에서 추출한 합성곱 신경망 시각 특징이 아닌 영상 캡션(image caption)이나 물체 범주명(object category)과 같은 텍스트에서 추출한 언어 특징도 함께 활용하였다. 하지만, 이 연구는 물체 탐지와 관계 탐지 간의 상호 작용을 고려해 이 두 문제를 동시에 해결해야 하는 장면 그래프 생성과는 달리, 이미 탐지된 물체들을 토대로 단지 그들 간의 관계만을 판별해내는데 연구에 초점이 맞추어져 있다. 따라서 이웃한 물체 노드와 관계 노드들의 맥락 정보를 각 그래프 노드에 반영하기 위한 별도의 그래프 신경망 기반의 특징값 임베딩 과정은 적용되지 않았다.

이러한 기존 모델들의 한계성을 고려하여, 본 논문에서는 장면 그래프 생성을 위한 새로운 심층 신경망 모델을 제안한다. 제안 모델에서는 물체와 관계의 효과적인 탐지를 위해, 합성 곱 신경망 기반의 시각 맥락 특징들뿐만 아니라 언어 맥락 특징들을 포함하는 다양한 멀티 모달 맥락 정보들을 활용한다. 특히 <주어 물체-관계 서술자-목적어 물체> 형태의 관계 표현에 내포된 각 물체의 순서와 역할을 고려해 양방향 순환신경망(bidirectional Recurrent Neural Network, biRNN)을 이용해 언어 맥락 특징 벡터를 생성한다. 또한, 제안 모델에서는 관계를 맺는 두 물체 간의 상호 의존성이 그래프 노드 특징값들에 충분히 반영되도록, 그래프 신경망을 이용해 맥락 정보를 임베딩한다. 본 논문에서는 제안 모델의 효과와 성능을 분석하기 위해, Visual Genome[4] 벤치마크 데이터 집합을 이용한 다양한 비교 실험들을 수행하고 결과를 소개한다.

2. 장면 그래프 생성 모델

2.1 모델 개요

본 논문에서 제안하는 장면 그래프 생성을 위한 신경망 구조는 (그림 2)와 같다. 제안 모델은 물체 영역 탐지(region proposals, RP), 물체 및 관계 탐지(object & relationship detection, ORD), 그리고 그래프 생성(graph generation, GG)의 3단계로 이루어진다. 물체 영역 탐지(RP) 단계에서는 대표적인 물체 탐지 모듈인 Faster R-CNN을 이용하며, 입력 영상의 각 물체 후보 영역별 ResNet101 시각 특징 벡터, 바운딩 박스(bounding box)의 위치와 크기, 물체 범주별 확률 분포(object class distribution) 등을 구해낸다.

물체 및 관계 탐지(ORD) 단계는 다시 그래프 초기화(graph initialization), 그래프 추론(graph reasoning), 그래프 레이블링(graph labelling)의 세부 단계들로 구성된다. 그래프 초기화 단계에서는 물체 영역 탐지(RP) 과정을 통해 얻어진 입력 영상 내 각 물체 영역들을 기초로 장면 그래프를 구성할 물체 노드 및 관계 노드들을 생성하고, 이들 노드에 초기 특징값을 부여한다. 그래프 추론 단계에서는 그래프 합성 곱 신경망(Graph Convolution Neural Network, GCN)을 이용하여, 그래프 내 이웃한 물체 노드 및 관계 노드들 사이에 서로 맥락 정보를 교환하며 각 노드의 특징값을 갱신한다. 그래프 레이블링 단계에서는 각 노드의 최종 특징값을 바탕으로 물체(object) 및 관계(relationship)를 분류(node classification)해낸다. 마지막 그래프 생성 단계에서는 분류된 각 노드들을 토대로 하나의 장면 그래프를 완성한다.

2.2 물체 노드 특징

제안 모델의 그래프 초기화(Graph Initialization) 단계에서는 영상에서 탐지된 각 물체 영역별로 그래프 내에 하나의 물체 노드(object node)를 생성하고, 해당 노드에 초기 특징값을 부여한다. 제안 모델에서는 대표적인 물체 탐지 모듈인 Faster R-CNN을 입력 영상에 적용하여, 각 물체 후보 영역별로 추출한 시각 특징 벡터와 물체 클래스 확률 분포를 각 물체 노드의 초기 특징값으로 할당한다. 이 초기 특징값은 추후 그래프 신경망을 통해 이웃 노드들의 풍부한 맥락 정보가 결합된 후, 물체 노드의 분류에 사용된다. 따라서 제안 모델에서 최종 판별하는 각 노드의 물체 범주는 Faster R-CNN이 추측한 초기 물체 범주와는 달라질 수도 있다.

- **물체 시각 특징(object visual feature)**
  - f_o : 해당 물체 영역의 합성 곱(CNN) 시각 특징
- **클래스 확률 분포 (class probability distribution)**
  - p_o : 해당 물체 영역의 물체 클래스 확률 분포

따라서 각 물체 노드의 초기 특징 벡터 O는 (식 1)과 같다.

$$O = [f_o , p_o] \qquad \text{(식 1)}$$

(식 1)의 [,]은 연결 연산(concatenate)을 나타낸다.

2.3 관계 노드 특징

그래프 초기화 단계에서는 앞서 설명한 물체 노드의 초기화 이외에, 관계 노드의 초기화도 수행한다. 즉 영상에서 탐지된 물체 영역들의 각 쌍(pair)에 대해 그래프 내에 하나의 관계 노드를 생성하고, 해당 노드에 초기 특징값을 부여한다. 기존 모델들과는 달리, 제안 모델에서는 효과적인 관계 탐지를 위해 영상 기반의 시각 맥락 특징(visual context feature)들 외에 텍스트 기반의 언어 맥락 특징(linguistic context feature)들도 포함하는 풍부한 멀티 모달 맥락 정보를 관계 노드의 초기 특징값으로 할당한다. 관계 노드를 위한 시각 맥락 특징 집합과 언어 맥락 특징 집합의 구성은 다음과 같다.

- **시각 맥락 특징 집합(visual context feature set)**
  - f_I : 입력 영상 전체의 합성 곱 시각 특징
  - f_u^{ij} : 하나의 관계(relationship)를 맺을 수 있는 주어

물체(subject) 영역과 목적어 물체(object) 영역을 둘러싸는 영상 영역(union box)의 합성 곱 시각 특징

- s_u : 주어 물체와 목적어 물체를 둘러싸는 영역(union box)의 위치 정보

$$s_u = (\frac{x_{tl}}{W}, \frac{y_{tl}}{H}, \frac{x_{br}}{W}, \frac{y_{br}}{H}, \frac{(x_{br}-x_{tl})(y_{br}-y_{tl})}{WH}) \quad \text{(식 2)}$$

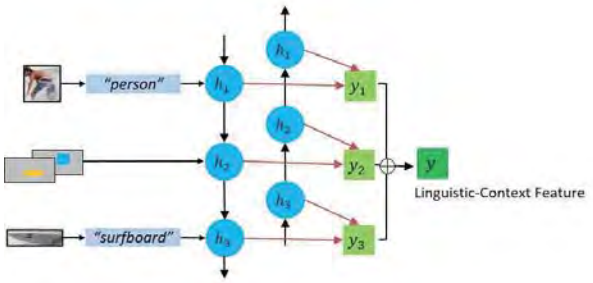
(식 2)의 x, y, w, h는 각각 물체 영역의 중심 좌표와 너비, 높이를 의미하며, W, H 는 union box의 너비와 높이를 각각 나타낸다. 한편, (식 3)의 x_{tl}, y_{tl}은 union box의 왼쪽 상단 모서리 좌표를, x_{br}, y_{br}은 오른쪽 하단 모서리 좌표를 각각 나타낸다.

- **언어 맥락 특징 집합(linguistic context feature set)**
  - w_i : 주어 물체의 예상 범주명(object category)을 다층 퍼셉트론(MLP)으로 임베딩한 특징
  - s : 주어 물체 영역과 목적어 물체 영역의 영상 내 위치 정보

$$s = [x_i, y_i, w_i, h_i, \frac{x_i - x_j}{W}, \frac{y_i - y_j}{H}, \log\frac{w_i}{w_j}, \log\frac{h_i}{h_j}, x_j, y_j, w_j, h_j] \quad \text{(식 3)}$$

  - w_j : 목적어 물체의 예상 범주명을 다층 퍼셉트론으로 임베딩한 특징

한편, 하나의 관계를 표현하기 위한 언어 맥락 특징 벡터 y는 앞서 소개한 w_i, s, w_j 등 3 가지 구성 요소들을 단순 연결(concatenate), 단방향 순환신경망(RNN), 양방향 순환신경망(biRNN) 등 다양한 결합 방식으로 구할 수 있다. 일반적으로 두 물체 간의 관계는 <주어-관계 서술자-목적어>와 같이 3가지 언어 구성 요소 각각의 위치와 순서, 그리고 역할을 고려하여 하나의 시퀀스(sequence)로 표현하는 것이 바람직하다. 이 점에 착안하여, 본 제안 모델에서는 3가지 언어 구성 요소들(w_i, s, w_j)을 양방향 순환신경망(bidirectional Recurrent Neural Network, biRNN)을 이용해 순차적으로 결합함으로써, 언어 맥락 특징 벡터 y를 생성해낸다. 특히, 언어의 개념적 관계에 기초하여 해당 관계를 맺을 수 있는 가능한 주어 물체 유형과 목적어 물체 유형 간의 쌍방향 제약(bidirectional constraint)을 특징 벡터 y에 효과적으로 담아내기 위해 제안 모델에서는 양방향 순환신경망(biRNN)으로 언어 맥락 시퀀스 <w_i, s, w_j>를 임베딩한다. (그림 3)은 biRNN 기반의 언어 맥락 특징값 임베딩 과정을 그림으로, (식 4)는 해당 과정을 수식으로 각각 나타내고 있다.



(그림 3) biRNN 기반의 언어 맥락 특징 임베딩

$$y = (W_{\overrightarrow{h}_{y_1}}\overrightarrow{h}_1 + W_{\overleftarrow{h}_{y_1}}\overleftarrow{h}_1) + (W_{\overrightarrow{h}_{y_2}}\overrightarrow{h}_2 + W_{\overleftarrow{h}_{y_2}}\overleftarrow{h}_2) + (W_{\overrightarrow{h}_{y_3}}\overrightarrow{h}_3 + W_{\overleftarrow{h}_{y_3}}\overleftarrow{h}_3) \quad \text{(식 4)}$$

W는 학습 파라미터, $\overrightarrow{h}$는 순방향에서의 은닉상태(hidden state), $\overleftarrow{h}$는 역방향에서의 은닉상태를 의미한다. 제안 모델에서 각 관계 노드의 초기 특징값은 시각 맥락 특징 벡터와 biRNN으로 임베딩된 언어 맥락 특징 벡터를 결합하여, (식 5)와 같이 주어진다.

$$R = [f_I, f_u^{ij}, s_u, y] \quad \text{(식 5)}$$

2.4 그래프 추론 및 레이블링

제안 모델의 그래프 추론(Graph Reasoning) 과정은 각각 시각적 추론 단계(visual level)와 의미적 추론 단계(semantic level)를 나타내는 그래프 합성 곱 신경망(Graph Convolutional Network)의 2개 계층으로 구성된다. 각 계층에서는 그래프 초기화 단계에서 부여된 각 노드의 초기 특징값들을 토대로 그래프의 이웃한 노드들 사이에 맥락 정보를 서로 교환함으로써, 각 노드의 특징값을 새롭게 갱신한다. 특히 제안 모델에서는 주의 집중 그래프 합성 곱 신경망(attentional GCN)을 사용함으로써, 이웃 노드들 중 집중해야 할 노드와 그렇지 않은 노드를 구별하여 각 노드의 특징값 갱신에 이웃 노드의 정보를 차등적으로 반영한다. 각 노드의 주의 집중 값 a_i는 (식 6) 및 (식 7)과 같이, 두 노드의 특징값 z_i와 z_j를 토대로 예측한다.

$$m_{ij} = w\sigma(W_a[z_i^{(l)}, z_j^{(l)}]) \quad \text{(식 6)}$$

$$a_i = softmax(m_i) \quad \text{(식 7)}$$

(식 6)과 (식 7)에서 σ는 2개 계층 퍼셉트론(MLP)을, w와 W는 학습용 파라미터를 각각 나타낸다.

주의 집중 그래프 신경망을 이용하여 물체 노드의 특징값을 갱신할 때는 주어물체 노드<->목적어 물체 노드, 주어물체 노드<->관계 노드, 목적어 물체 노드<->관계 노드 간에 맥락 정보 교환이 이루어진다. 반면에 관계 노드의 특징값을 갱신할 때는 관계 노드<->주어물체 노드, 관계 노드<->목적어 물체 노드 간에 맥락 정보 교환이 일어난다. 따라서 그래프 내 각 물체 노드의 특징값 갱신은 (식 8)과 같고, 반면에 관계 노드의 특징값 갱신은 (식 9)와 같다.

$$z_i^o = \sigma(W_{so}Z^o a_{so} + W_{sr}Z^r a_{sr} + W_{or}Z^r a_{or}) \quad \text{(식 8)}$$

$$z_i^r = \sigma(z_i^r + W_{rs}Z^o a_{rs} + W_{ro}Z^o a_{ro}) \quad \text{(식 9)}$$

(식 8)과 (식 9)에서 s, r, o는 주어 물체(subject) 노드, 관계(relationship) 노드, 목적어 물체(object) 노드를 각각 나타낸다. 시각적 추론 단계와 의미적 추론 단계로 구성되는 2개의 주의 집중 그래프 신경망 계층에서는 이와 같은 노드 특징값 갱신 과정이 각각 수행된다. 대신 시각적 추론 단계의 결과인 각 노드의 물체 및 관계 클래스 확률 분포가 의미적 추론 단계의 초기 노드 입력으로 제공된다.

끝으로, 그래프 레이블링(Graph Labelling) 단계에서는 의미적 추론 단계에서 얻어진 각 노드의 최종 특징값을 바탕으로, 물체 및 관계를 분류해낸다. 물체 노드는 물체 클래스 확률 분포에서 가장 큰 값으로 레이블링한다. 관계 노드 또한 같은 과정을 거쳐 레이블링 된다. 이를 통해 <주어-서술자-목적어>형태의 정형화된 결과물을 얻는다.

3. 구현 및 실험

3.1 데이터 집합과 모델 학습

본 논문에서는 Visual Genome[4] 벤치마크 데이터 집합을 이용하여, 제안하는 모델의 성능을 평가하였다. Visual Genome 데이터 집합에서 등장 빈도수가 높은 물체 종류 150개와 관계 종류 50개를 선별하여 평가에 사용하였다. 또한, 데이터 집합 영상들 중에서 56,224장은 학습 데이터로, 26,446장은 평가 데이터로 사용하였다.

제안 모델은 Ubuntu 16.04 LTS 환경에서 Python 딥러

닝 라이브러리인 PyTorch를 이용하여 구현하였다. 모델의 학습과 평가는 GeForce GTX 1080Ti GPU카드가 설치된 하드웨어 환경에서 수행하였다. 모델 학습을 위해 일괄 처리량(batch size)은 8, 반복 횟수(epoch)는 6으로 각각 설정하였다. 또한 학습율(learning rate)은 0.005, 최적화 함수(optimizer)는 확률적 경사 하강법(stochastic gradient descent)을 사용하였다.

3.2 실험

본 논문에서의 제안하는 장면 그래프 생성 모델의 성능을 평가하기 위해 SGGen, PhrCls, PredCls 총 3가지의 평가 지표를 사용하였다. 위 지표들은 어느 정도의 정답을 사용하는지에 차이를 둔다. SGGen은 물체의 영역, 종류, 관계를 모두 예측하고 PhrCls는 물체의 영역만 정답을 사용하여 나머지를 예측한다. PredCls는 물체의 영역과 종류를 정답을 사용하여 관계만을 예측한다. 모두 장면 그래프를 나타내는 트리플들의 재현율을 측정하며 트리플을 구성하는 주어 물체와 목적어 물체 그리고 그 둘의 관계가 모두 정답과 같아야 정답으로 인정한다. SGGen은 추가적으로 두 물체의 영역이 정답 물체의 영역과 0.5이상의 IoU값을 가질 때 해당 트리플을 정답으로 인정한다. 총 3가지의 비교 실험을 진행하였으며 성능 평가를 위해 상위 50개(r@50)와 100개(r@100)에 대한 성능을 측정하였다.

첫 번째 실험은 제안 모델의 관계 노드를 위한 언어 맥락 특징 임베딩 방법인 biRNN의 효과를 입증하기 위한 실험이다. 이 실험에서는 언어 맥락 특징을 단순 연결(concat)하였을 경우, RNN으로 임베딩 하였을 경우, biRNN으로 임베딩 하였을 경우를 서로 비교하였다.

<표 1> 언어 맥락 특징의 임베딩 방법 간의 성능 비교

method	SGGen		PhrCls		PredCls	
	r@50	r@100	r@50	r@100	r@50	r@100
concat	24.35	27.07	41.96	52.50	65.20	69.32
RNN	24.82	27.20	43.33	53.81	65.98	69.75
biRNN	**24.91**	**27.61**	**43.69**	**54.16**	**66.87**	**71.15**

<표 1>은 실험 결과를 나타낸다. 실험 결과에서 biRNN을 적용하였을 때 3가지 평가 지표 모두 가장 높은 성능을 보여주었다. 단순 연결(concat)하였을 때는 <주어-서술자-목적어>의 시퀀스를 전혀 반영하지 못해 가장 성능이 낮았다. RNN의 경우 시퀀스 특징을 반영하였기에 단순 연결보다는 나았으나, 양방향성까지 고려한 biRNN보다는 낮은 성능을 보였다.

두 번째 실험은 본 논문에서 제안하는 언어 맥락 특징 집합(LC)과 시각 맥락 특징 집합(VC)을 포함하는 멀티 모달 관계 노드 특징값의 효과를 분석하기 위한 실험이다. 이 실험에서는 시각 맥락 특징 집합(VC)만 사용하였을 경우, 언어 맥락 특징 집합(LC)만 사용하였을 경우, 두 가지 모두 사용하였을 경우를 각각 비교한다. 물체 노드 특징값은 모두 동일하게 사용하였고, 언어 맥락 특징 집합을 사용한 경우에는 biRNN을 통한 임베딩 과정을 거친다.

<표 2> 특징 집합들 간의 성능 비교

Feature Set		SGGen		PhrCls		PredCls	
VC	LC	r@50	r@100	r@50	r@100	r@50	r@100
✓		23.93	26.92	40.50	51.30	63.94	68.35
	✓	24.49	27.35	42.76	53.29	**66.92**	71.03
✓	✓	**24.91**	**27.61**	**43.69**	**54.16**	66.87	**71.15**

<표 2>는 실험 결과를 나타낸다. 이 실험 결과는 본 제안 모델과 같이 관계 노드 특징값으로 언어 맥락 특징 집합(LC)과 시각 맥락 특징 집합(VC)을 모두 활용하였을 때, 가장 높은 성능을 보여주었다. 또한, 시각 맥락 특징 집합(VC)만을 사용하였을 때에 비해, 언어 맥락 특징 집합(LC)만을 사용하였을 때가 상대적으로 더 높은 성능을 보였다. 이 결과를 통해, 언어 맥락 특징 집합(LC)이 시각 맥락 특징 집합(VC)에 비해 성능 개선에 더 큰 도움을 주는 것으로 판단한다.

세 번째 실험은 본 논문에서 제안한 장면 그래프 생성 모델을 최신 기존 모델들(state-of-the-art models)과 성능을 비교하는 실험이다. 앞서 설명한 바와 같이 [1]과 [2]의 모델은 시각 특징만을 이용하는데, 반해 [3]의 모델은 영상 캡션(caption)에 기초한 언어 특징을 활용한다. 또한, [1]과 [3]의 모델들과는 달리, [2]의 모델은 그래프 신경망 기반의 노드 특징값 임베딩 과정을 별도로 포함하고 있다.

<표 3> 장면 그래프 생성 모델 간의 성능 비교

model	SGGen		PhrCls		PredCls	
	r@50	r@100	r@50	r@100	r@50	r@100
[1]	10.72	14.22	24.34	26.5	67.03	71.01
[2]	11.4	13.7	29.6	31.6	54.2	59.1
[3]	22.17	23.62	28.58	31.69	**85.02**	**91.77**
ours	**24.91**	**27.61**	**43.69**	**54.16**	66.87	71.15

<표 3>에서 알 수 있듯이, 제안 모델이 SGGen과 PhrCls에서 기존 모델들보다 더 우수한 성능을 보였다. 다만, PredCls 평가 지표에서만 [3]의 모델보다 낮은 성능을 보였다. [3]의 모델은 본 제안 모델과는 달리, 입력 영상에 관한 캡션 텍스트 데이터를 추가적으로 활용하였기에 이러한 성능 차이를 보인 것으로 판단한다. 하지만 일반적인 장면 그래프 생성 작업에서는 영상이외에 별도로 이와 같은 설명글을 제공하는 경우는 매우 드물다. 따라서 본 실험을 통해, 장면 그래프 생성을 위한 제안 모델의 높은 성능을 확인할 수 있었다.

4. 결 론

본 논문에서는 영상으로부터 장면 그래프를 효과적으로 생성할 수 있는 심층 신경망 모델을 제안하였다. 제안 모델은 시각 맥락 특징, 언어 맥락 특징 등 다양한 멀티 모달 맥락 정보를 활용하며, 특히 언어 맥락 특징의 효과를 극대화하기 위해 별도의 biRNN 네트워크를 사용한다. 또한, 제안 모델에서는 관계를 맺는 두 물체 간의 상호 의존성이 그래프 노드 특징값들에 충분히 반영되도록, 그래프 신경망을 이용해 맥락 정보를 임베딩한다. Visual Genome 벤치마크 데이터 집합을 이용한 비교 실험들을 통해서, 기존 모델들에 비해 우수한 제안 모델의 성능을 확인할 수 있었다.

참고문헌

[1] Y. Li, W. Ouyang, and B. Zhou, et. al., “Scene Graph Generation from Objects, Phrases and Region Captions,” *Proc. of ICCV-17*, 2017.

[2] J. Yang, J. Lu, and S. Lee, et al., "Graph R-CNN for Scene Graph Generation," *Proc. of ECCV*, 2018.

[3] W. Liao, B. Rosenhahn, and L. Shuai, et al., "Natural Language Guided Visual Relationship Detection," *Proc. of IEEE*, 2019.

[4] R. Krishna, Y. Zhu, and O. Groth, et al., "Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations," *Proc. of the IJCV*, 2017.