KIPS Trans. Softw. and Data Eng.
Vol.9, No.3 pp.91~100 pISSN: 2287-5905

https://doi.org/10.3745/KTSDE.2020.9.3.91

KG_VCR: A Visual Commonsense Reasoning Model Using Knowledge Graph

JaeYun Lee† · Incheol Kim††

ABSTRACT

Unlike the existing Visual Question Answering(VQA) problems, the new Visual Commonsense Reasoning(VCR) problems require deep common sense reasoning for answering questions: recognizing specific relationship between two objects in the image, presenting the rationale of the answer. In this paper, we propose a novel deep neural network model, KG_VCR, for VCR problems. In addition to make use of visual relations and contextual information between objects extracted from input data (images, natural language questions, and response lists), the KG_VCR also utilizes commonsense knowledge embedding extracted from an external knowledge base called ConceptNet. Specifically the proposed model employs a Graph Convolutional Neural Network(GCN) module to obtain commonsense knowledge embedding from the retrieved ConceptNet knowledge graph. By conducting a series of experiments with the VCR benchmark dataset, we show that the proposed KG_VCR model outperforms both the state of the art(SOTA) VQA model and the R2C VCR model.

Keywords : Visual Commonsense Reasoning, Deep Neural Network, Graph Convolutional Network, Knowledge Graph Embedding

KG_VCR: 지식 그래프를 이용하는 영상 기반 상식 추론 모델

이 재 윤† · 김 인 철††

요 약

기존의 영상 기반 질문-응답(VQA) 문제들과는 달리, 새로운 영상 기반 상식 추론(VCR) 문제들은 영상에 포함된 사물들 간의 관계 파악과 답변 근거 제시 등과 같이 추가적인 심층 상식 추론을 요구한다. 본 논문에서는 영상 기반 상식 추론 문제들을 위한 새로운 심층 신경망 모델인 KG_VCR을 제안한다. KG_VCR 모델은 입력 데이터(영상, 자연어 질문, 응답 리스트 등)에서 추출하는 사물들 간의 관계와 맥락 정보들을 이용할 뿐만 아니라, 외부 지식 베이스인 ConceptNet으로부터 구해내는 상식 임베딩을 함께 활용한다. 특히 제안 모델은 ConceptNet으로부터 검색해낸 연관 지식 그래프를 효과적으로 임베딩하기 위해 그래프 합성곱 신경망(GCN) 모듈을 채용한다. VCR 벤치마크 데이터 집합을 이용한 다양한 실험들을 통해, 본 논문에서는 제안 모델인 KG_VCR이 기존의 VQA 최고 모델과 R2C VCR 모델보다 더 높은 성능을 보인다는 것을 입증한다.

키워드 : 영상 기반 상식 추론, 심층 신경망, 그래프 합성곱 신경망, 지식 그래프 임베딩

1. 서 론

최근 딥러닝(deep learning)을 위시한 기계 학습 기술의 발전과 더불어 컴퓨터 비전(computer vision), 자연어 처리(natural language processing) 등과 같은 인공지능(AI)의 핵심 기술들이 혁신적으로 발전함에 따라, 다시 고전적인 튜링 테스트(Turing Test)와 같이 인공지능이 얼마나 인간에 가까운 복합 지능을 발휘할 수 있는 지 알아보려는 매우 도전적인 과제들이 활발히 생겨나고 있다. 그 중에서도 영상 기반 질문-응답(Visual Question Answering, VQA)[1]은 시각적 튜링 테스트(Visual Turing Test)의 한 형태로서, 영상에 관한 자연어 질문에 인공지능이 얼마나 자연스러운 답변을 자동 생성하는 지를 알아보기 위한 지능 작업이다.

그동안 VQA 챌린지를 통해 수많은 우수한 모델들이 개발되는 긍정적인 효과를 거두고 있으나, 아직 인간 수준의 추론 능력과 답변 생성 능력에 도달하기엔 한계가 있다. 현재 VQA의 대표적인 한계점은 대부분의 질문들이 입력 영상이나 질문에 명백하게 포함되어 있는 내용들만을 다루고, 소위 상식 추론(commonsense reasoning)을 요구하는 질문들은 별로 없다는 점이다. 보편적인 인간은 한 영상을 보면, 그 영

※ 본 연구는 과학기술정보통신부 및 정보통신기술진흥센터의 대학ICT연구센터 육성지원사업의 연구결과로 수행되었음(IITP-2017-0-01642).
※ 본 연구는 2020학년도 경기대학교 대학원 연구원장학생 장학금 지원에 의하여 수행되었음.
※ 이 논문은 2019년도 한국정보처리학회 추계학술발표대회에서 '지식 그래프를 이용한 영상 기반 상식 추론'의 제목으로 발표된 논문을 확장한 것임.
† 준 회 원 : 경기대학교 컴퓨터과학과 석사과정
†† 종신회원 : 경기대학교 컴퓨터과학과 교수
Manuscript Received : December 20, 2019
First Revision : February 5, 2020
Accepted : February 17, 2020
* Corresponding Author : Incheol Kim(kic@kyonggi.ac.kr)

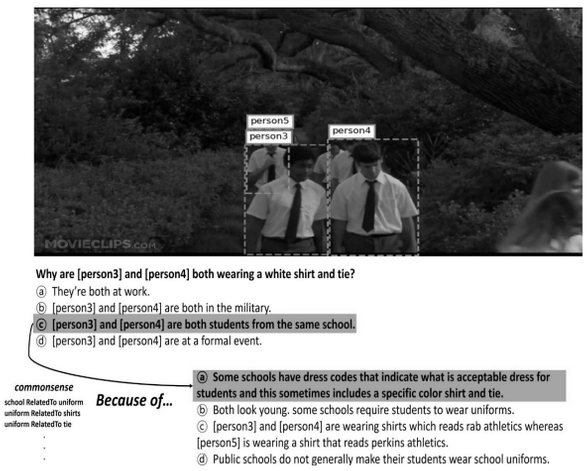

Fig. 1. Example of Visual Commonsense Reasoning(VCR)

상에 담겨 있는 사물(object)들을 식별해낼 수 있을 뿐만 아니라, 나아가 그들 간의 개념적 관계(conceptual relationship), 사건들 간의 인과 관계(causal relationship)를 토대로 영상에 포착된 장면 이전의 또는 이후의 장면들도 상상해낼 수 있는 등 다양한 추론이 가능하다. 이러한 VQA의 한계성에 대응하여, 최근 새롭게 영상 기반 상식 추론(Visual Commonsense Reasoning, VCR)[2] 문제들이 제시되었다. 영상 기반 상식 추론(VCR) 문제는 Fig. 1의 예와 같이, 하나의 영상(image)과 자연어 질문(question), 응답 리스트(response list)가 주어지면, 질문에 가장 적절한 답변(answer)과 근거(rationale)를 선택하는 문제이다. 영상 기반 상식 추론(VCR) 문제는 표면적으로 영상 기반 질문-응답(VQA)과 매우 비슷해 보이지만, 숨겨져 있는 사물들 간의 관계 파악과 답변 근거 제시 등 별도의 상식 추론이 요구된다는 점에서 상당한 차이가 있다.

이러한 영상 기반 상식 추론 문제를 해결하기 위해서는 R2C[2]와 같이 신경망 모델이 학습을 통해 입력 영상과 자연어 질문, 응답 리스트에 포함된 사물들 간의 관계와 맥락 정보를 스스로 파악할 수 있도록 접근하는 방법도 있으나, 본 논문에서는 입력 데이터 안에서만 답변 선택에 필요한 상식을 얻으려는 시도는 한계가 있다고 판단한다. 따라서 본 논문에서는 R2C와 같이 입력 데이터(영상, 자연어 질문, 응답 리스트)에서 사물들 간의 관계와 맥락 정보를 추출해내는 모듈들 외에, 별도로 ConceptNet과 같은 외부의 공개된 대규모 지식 베이스(knowledge base)로부터 관련 상식(commonsense knowledge)들을 직접 가져다 추가적으로 활용할 수 있는 모듈들을 포함한 새로운 심층 신경망(deep neural network) 모델 KG_VCR을 제안한다. 기존 VQA 문제들에 대해서도 FVQA[3], KVQA[4] 등과 같이 외부 지식 베이스를 활용하는 선행 연구들이 있었으나, 이들과는 달리 본 연구에서는 상식 추론을 위해 그래프 합성 곱 신경망(Graph Convolutional Neural Network, GCN) 기반의 지식 그래프 임베딩을 수행하는 새로운 접근 방식을 취한다. 심볼(symbol) 형태의 구조화된 지식 그래프와 벡터(vector) 형태의 영상 및 질문 특징(feature)들을 하나의 심층 신경망 모델에서 함께 처리하기 위해, 제안 모델 KG_VCR에서는 그래프 합성 곱 신경망(GCN) 모듈을 적용해 지식 그래프를 벡터 형태로 임베딩한 후, 답변 결정에 이용한다. 본 논문에서는 제안 모델인 KG_VCR의 세부 설계사항들을 소개한 뒤, VCR 벤치마크 데이터 집합을 이용한 다양한 실험들을 통해 제안 모델의 성능을 입증한다.

2. 관련 연구

2.1 영상 기반 질문-응답

영상 기반 질문-응답(Visual Question Ansering, VQA)[1]은 영상(image)과 함께 이 영상에 관한 자연어 질문(natural language question)이 주어지면, 질문에 적합한 답변을 자동 생성하는 작업이다. 영상 기반 질문-응답(VQA)은 [1]에서 처음 제안 되었다. 이 연구에서는 이 연구에서는 영상 기반 질문-응답 문제를 위한 기본적인 심층 신경망 모델도 제시하였다. 영상 기반 질문-응답(VQA)은 자연어로 표현된 질문과 2차원 컬러 영상과 같이 서로 다른 형태의 입력 데이터들로부터 답변을 출력으로 얻어내기 위하여 다양한 모델들이 개발되었다. 영상 기반 질문-응답(VQA)의 초기 모델[1]은 자연어 질문을 인코딩하는 LSTM(Long Short Term Memory)와 영상을 인코딩하는 CNN(Convolutional Neural Network)를 합친 단순한 모델이었다. 이후 연구[6-10]에서는 더 나아가 다양한 주의 집중(attention) 메커니즘들이 개발되어 모델에 추가적으로 사용되었다. 영상 기반 질문-응답(VQA)에 사용된 초기의 주의 집중 메커니즘들은 질문에서 언급된 개체(entity)와 연관된 시각 영역(visual region)을 입력 영상 내에서 찾아보려는 단방향성 주의 집중 메커니즘들을 사용한 [6, 7]이 있다. 이후 질문과 입력 영상 상호간에 주의 집중을 적용하는 양방향성 주의 집중 메커니즘들을 사용한[8-10]이 있다. 본 논문에서는 질문에서 언급한 개체가 응답 리스트에서 어떤 개체를 가리키는지, 응답 리스트에서 언급한 개체가 영상의 어느 시각 영역을 가리키는지 명확하게 하기 위하여 [8-10]과 같은 양방향성 주의집중 메커니즘을 사용하였다.

2.2 외부 지식을 이용한 영상 기반 질문-응답

영상에 관한 질문들 중에는 영상 외적인 정보나 지식까지 활용해야 답할 수 있는 경우도 종종 발생한다. 예컨대, 영상에 등장하는 두 사람이 직장 동료 관계인지를 묻는 질문에 명확하게 답하기 위해서는 이 두 사람의 소속 기관에 관한 별도의 정보나 지식을 활용해야만 가능하다. 많은 공개 지식 베이스(open knowledge base)들은 대부분 (Subject, Predicate, Object)와 같은 트리플(triple) 구조의 지식들을 포함하고 있다. Subject와 Object는 각각의 개념을 나타내며, Predicate는 두 개념 간의 특정 관계를 나타낸다. 이러한 트리플의 모음은 하나의 큰 그래프를 형성한다. 영상 기반 질문-응답에 활

용할 수 있는 대표적인 공개 지식 베이스로는 Wikipedia에서 추출한 다양한 범용 지식들을 저장하고 있는 DBpedia [11]와 FreeBase[12], 기초 상식과 개념 지식들을 포함하고 있는 ConceptNet[13] 등이 있다.

최근에는 기존의 영상 기반 질문-응답(VQA)에서 더 나아가 외부 지식을 활용하는 영상 기반 질문-응답에 관한 연구가 활발하다. 이러한 연구들은 다시 활용 지식의 유형에 따라, 상식을 이용한 영상 기반 질문-응답(Commonsense Knowledge-enabled VQA, CK-VQA), 세부 지식을 이용한 영상 기반 질문-응답(World Knowledge-aware VQA, WK-VQA)으로 나뉠 수 있다. 상식을 이용한 영상 기반 질문-응답(CK-VQA)은 Q: "이 영상 안에 있는 것 중 소리를 증폭시킬 수 있는 것은 무엇인가?", A: "마이크"와 같이, 사람들이 보편적으로 알고 있는 사물의 유형(class)이나 개념(concept)들에 관한 기초 지식들을 다룬다. 반면에, 세부 지식을 이용한 영상 기반 질문-응답(WK-VQA)의 경우는 Q: "이 영상에서 Barack Obama의 왼쪽에 있는 인물은 누구인가?", A: "Richard Cordray"와 같이, 개체 명(named entity)을 포함한 실제 세계의 구체적인 사실(fact)이나 전문 지식(domain-specific knowledge)들을 주로 다룬다. 대표적인 상식을 이용한 영상 기반 질문-응답(CK-VQA) 연구로는 KB-VQA[14], FVQA[3] 등이 있고, 세부 지식을 이용한 영상 기반 질문-응답(WK-VQA) 연구로는 KVQA[4], OK- VQA[15] 등이 있다. KB-VQA[14]와 FVQA[3]에서는 답변에 필요한 상식들을 주로 DBpedia[11], ConceptNet [13]에서 추출하여 사용하였다. 반면에, KVQA[4]와 OK-VQA[15]에서는 답변에 필요한 세부 지식들을 주로 FreeBase[12]에서 추출하여 사용하였다.

본 논문에서 다루는 영상 기반 상식 추론(VCR) 문제는 질문에 대한 답변 외에 근거(rationale)를 추가적으로 묻는 등 질문 모드에는 차이가 있으나, 대부분의 질문들이 세부 지식보다는 기초 상식들을 필요로 하는 질문들이어서 크게 보면 상식을 이용한 영상 기반 질문-응답(CK-VQA)과 유사하다.

2.3 영상 기반 상식 추론

상식은 사람에게는 당연시 여겨지는 지식이지만, 인공지능에서는 여전히 해결하기 어려운 문제이다. 본 논문에서 풀고자 하는 문제인 영상 기반 상식 추론(Visual Commonsense Reasoning, VCR)[2]은 영상(image), 자연어 질문(question), 응답 리스트(response list)가 주어지면, 질문에 가장 적합한 답변(answer)와 근거(rationale)을 선택하는 문제이다. 영상 기반 상식 추론(VCR)의 질문은 영상 기반 질문-응답(VQA)와 다르게 개념적 관계(conceptual relationship), 사건들 간의 인과 관계(causal relationship) 등을 포함한 총 7가지 종류의 상식을 필요로 하는 질문으로 구성되어 있다. 또한 질문에 대한 응답 뿐 아니라, 근거를 선택함으로써 더 깊은 상식의 추론을 요구한다.

초기 영상 기반 상식 추론(VCR)의 모델에서는 입력으로 주어진 데이터 내에서만 상식을 추출하고 맥락적인 정보만을 이용하여 응답을 예측하였다. 때문에 광범위한 상식을 습득하는 데 있어서는 한계가 있다. 본 논문에서는 이러한 한계성을 극복하고자 외부 상식 베이스인 ConceptNet을 사용하여 광범위한 상식을 습득하였다.

영상 기반 상식 추론(VCR)은 110K개의 영화 장면에서 파생된 290K개의 객관식 질문-응답(question answering)로 구성되어 있다. 본 논문에서는 시각적 개념(visual concept)를 제공하기 위해 37K개의 질문-응답(question answering) 쌍에 장소(scene), 행동(activity)정보를 추가하였다. 또한 시각적 개념, 질문, 응답 리스트로부터 키워드를 추출하여 트리플 구조의 외부 지식을 검색하여 추가적으로 사용하였다.

2.4 그래프 합성곱 신경망

그래프 신경망(Graph Neural Network)[16]은 그래프 구조에서 사용하는 인공 신경망이다. 그래프 신경망은 합성곱 신경망(Convolutional Neural Network)[17], 순환 신경망(Recurrent Neural Network)[18]과는 다르게 입력(input)이 하나의 벡터(vector)가 아닌 그래프(graph)라는 특징을 가진다. 그래프 신경망은 노드(node)와 관계(edge)를 입력 받고 이웃한 노드의 정보와 자신의 정보를 받아 임베딩한다. 모델의 계층(layer)를 늘릴수록 더 먼 곳에 있는 노드의 정보를 수집할 수 있다. 이 때 노드와 관계에 대한 파라미터는 각각 사용한다. 그래프 합성곱 신경망(Graph Convolutional Neural Network)[19]는 그래프 신경망(GNN)의 하위 개념으로, 합성곱 계층을 사용한다는 점에서 차이가 있다.

그래프 합성곱 신경망은 최근 들어 장면 그래프를 생성하는 Graph R-CNN[20], 질문-응답을 추론하는 BAG[21] 등 컴퓨터 비전(computer vision)과 자연어 처리(natural language processing)분야에 공통적으로 널리 활용되고 있다. 본 논문에서는 ConceptNet에서 검색된 트리플 구조의 지식을 다루기 위해 [21]과 유사한 방법으로 그래프 합성곱 신경망(GCN)을 사용하였고, 이와 같은 그래프 합성곱 신경망(GCN)을 이용한 지식 그래프 임베딩이 영상 기반 상식 추론(VCR) 성능 향상에 도움이 된다는 것을 보인다.

3. 영상 기반 상식 추론 모델

3.1 문제 정의

본 논문에서는 영상 기반 상식 추론(VCR) 문제를 위한 새로운 심층 신경망 모델을 제안한다. 영상 기반 상식 추론(VCR) 문제는 아래와 같이 서로 다른 3 가지 양식으로 제시된다.

- Q → A: 하나의 질문 q에 대해, 정답 a를 선택하는 문제
- QA → R: 하나의 질문 q와 정답 a에 대해, 올바른 근거 r을 선택하는 문제
- Q → AR: 하나의 질문 q에 대해, 정답 a와 올바른 근거 r을 함께 선택하는 문제

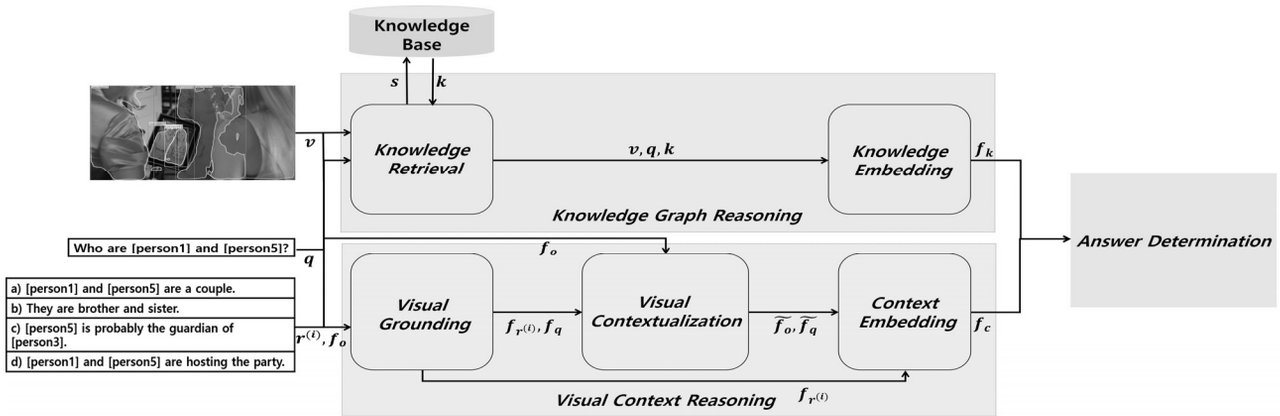

Fig. 2. Architecture of KG_VCR Model

그리고 영상 기반 상식 추론(VCR) 문제의 입력은 다음과 같은 형식으로 주어진다고 가정한다.

- 영상(image) I
- 물체 탐지(object detection) 결과물 o
- 자연어와 포인팅이 혼합된 질의(query) q
- 자연어와 포인팅이 혼합된 응답(response) $r^{(i)}$들

Fig. 1의 예에서 보듯이, 영상 기반 상식 추론(VCR) 문제는 입력 영상 I이외에도, 이 영상에서 탐지 가능한 각 물체 영역 o들이 물체 타입과 함께 제공된다. 또한, 질문 q와 답변 a, 그리고 근거 r들도 자연어(natural language)뿐만 아니라 영상에 등장하는 물체 영역 o들을 가리키는 포인팅 식별자(pointing tag) - 예컨대 [person3] - 들도 포함하고 있다. 하나의 질문에 대해 제시되는 응답 리스트의 개수는 N=4이며, 이 중 정답은 단 하나 존재하는 것으로 가정한다. 모델의 정확도(accuracy)는 문제 양식에 따라 달리 평가한다. 즉 Q → A 혹은 QA → R 문제에 관한 기준 정확도는 1/N인 반면, Q → AR 문제에 관한 기준 정확도는 $1/N^2$이 된다.

3.2 제안 모델

본 논문에서 제안하는 지식 그래프를 이용한 영상 기반 상식 추론 모델의 전체 구조는 Fig. 2와 같다. 그림에서 보듯이, 입력 영상(image) I, 질문(question) q, 응답 리스트(response) $r^{(i)}$등이 제안 모델의 입력으로 주어지고, 하나의 답변 a를 출력으로 결정한다. 영상 기반 상식 추론을 위한 심층 신경망 모델 KG_VCR의 전체 구조는 크게 (1) 지식 그래프 추론(Knowledge Graph Reasoning) 모듈, (2) 시각적 맥락 추론(Visual Context Reasoning) 모듈, (3) 답변 결정(Answer Determination) 모듈들로 구성된다. 지식 그래프 추론 모듈은 다시 지식 검색(Knowledge Retrieval), 지식 임베딩(Knowledge Embedding) 등의 하위 서브 모듈들로 구성되며, 입력 영상 I와 질문 q, 그리고 응답 리스트 $r^{(i)}$와 연관된 상식(commonsense knowledge)들을 외부 지식 베이스인 ConceptNet으로부터 검색해내고, 검색된 지식 그래프에 대표적인 그래프 합성곱 신경망 모듈인 GCN을 적용함으로써 하나의 지식 벡터 f_k로 임베딩해내는 역할을 수행한다.

반면에, 시각적 맥락 추론 모듈은 다시 시각적 접지(Visual Grounding), 시각적 맥락화(Visual Contextualization), 맥락 임베딩(Context Embedding) 등의 하위 서브 모듈들로 구성되며, 서로 다른 입력 데이터인 입력 영상 I와 자연어 질문 q, 그리고 응답 리스트 $r^{(i)}$에 포함된 사물(object)들을 서로 연관지어 그들 간의 관계와 맥락 정보를 추출함으로써, 멀티 모달 맥락 벡터 f_c를 생성해내는 역할을 수행한다. 끝으로 답변 결정 모듈에서는 위에서 설명한 두 모듈들의 결과물인 지식 벡터 f_k와 멀티 모달 맥락 벡터 f_c를 상호 보완적으로 결합함으로써, 제시된 응답 리스트에서 최적의 답변 a를 결정하는 역할을 수행한다. 후속 절들에서는 각 모듈의 설계에 관해 자세히 설명한다.

3.3 지식 그래프 추론

외부 지식 베이스로부터 시각적 상식 추론(VCR) 문제에 도움이 될 상식을 추출하여, 답변 결정에 효과적으로 활용하기 위해서는 (1) 관련 상식의 검색과 (2) 추출된 지식 그래프의 임베딩이 매우 중요하다. 본 논문의 제안 모델인 KG_VCR에서는 ConceptNet 지식 베이스로부터 관련 상식을 검색해내기 위해, 영상에서 인식해낸 시각적 개념들(visual concepts)뿐만 아니라, 자연어 질문(question)과 응답 리스트(response list)에서 추출한 키워드들(key words)을 함께 이용한다. 또한, 〈subject, relationship, object〉 형태로 구조화된 트리플(triple) 집합으로 구성된 상식 그래프를 그래프 고유의 관계성(relationship)을 효과적으로 고려하여 하나의 벡터로 임베딩하기 위해, 대표적인 그래프 합성 곱 신경망 모듈인 GCN을 이용한다.

앞서 소개한 바와 같이, KG_VCR 모델의 지식 그래프 추론 모듈은 다시 지식 검색, 지식 임베딩 등의 하위 서브 모듈들로 구성된다. 그리고 지식 검색 모듈의 기능과 구조는 (Fig. 3)과 같다. ConceptNet 지식 베이스로부터 입력 영상과 연관된 상식을 추출하기 위해, 미리 정해 놓은 범주에 따

라 영상 I에 포함된 사물(object), 장면(scene), 활동(activity)들을 각각 인식해내고, 이러한 시각적 개념 단어들을 지식 베이스의 검색어로 사용한다. 또한, 자연어로 된 질문과 응답 리스트와도 연관된 상식들을 지식 베이스로부터 검색해내기 위해서, 질문 q와 응답 리스트 $r^{(i)}$에 등장하는 키워드들도 지식 베이스의 검색어로 사용한다. 그리고 지식 베이스로부터 키워드당 약 100개의 트리플들로 구성된 지식들을 검색해낸다.

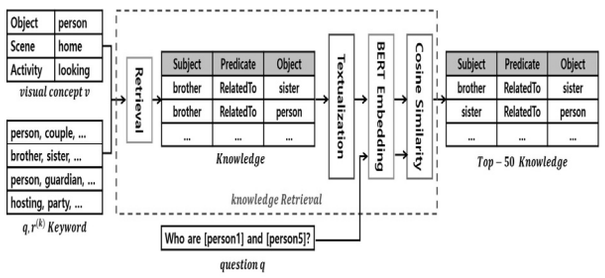

Fig. 3. Knowledge Retrieval Module

이렇게 검색된 트리플 구조의 지식들은 문장화(Textualization) 과정을 거쳐 "brother related to sister"와 같이 하나의 문장으로 생성된다. 이와 같이 문장화된 지식은 BERT[22] 임베딩을 거친 뒤, 질문 q와 코사인 유사도(Cosine Similarity) 계산을 거쳐 이들 중 상위 50개의 지식들만을 추출해낸다. 검색된 지식은 각각 〈subject, relationship, object〉 트리플 형태를 취하며, subject와 object들은 영상 속에 포함된 시각적 개념들(사물, 장면, 활동)이나 질문과 응답 리스트에 등장하는 단어들이 되며, relationship들은 이들 간의 관계를 나타내는 RelatedTo, SimilarTo, LocationAt, IsA 등 총 31개의 관계들 중 하나이다.

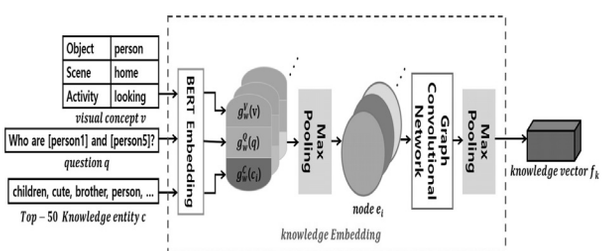

Fig. 4. Knowledge Embedding Module

지식 임베딩 모듈은 지식 베이스에서 검색된 관련 상식들을 토대로, 답변 결정에 도움을 줄 수 있는 지식 벡터 f_k를 생성하는 역할을 수행한다. 지식 임베딩의 한 방법으로 단순한 다층 퍼셉트론(MLP) 등이 사용될 수 있으나, 본 제안 모델에서는 개념 및 단어들 간의 관계를 구조적으로 잘 표현하고 있는 지식 그래프의 특성을 감안하여 그래프 합성 곱 신경망인 GCN을 채용한다. GCN은 개념 노드(concept node)들이 관계 간선(relation arc)들로 연결된 지식 그래프의 구조적 특성을 잘 반영하여, 지식 그래프를 효과적으로 벡터로 임베딩할 수 있는 방법으로 잘 알려져 있다.

Fig. 4와 같이, GCN의 입력으로 제공될 그래프의 각 노드들은 지식 베이스에서 검색된 상식들에 포함된 지식 개체(knowledge entity)들과 영상에서 인식해낸 시각적 개념(visual concept)들이 된다. 여기서 지식 개체란 〈subject, relationship, object〉 트리플 형태의 각 상식을 구성하는 subject나 object들을 의미한다. 그래프의 각 개념 노드에는 이들 지식 개체 혹은 시각적 개념 단어 외에도, 주어진 질문과의 연관성을 담아내기 위해 자연어 질문 자체도 포함시킨다. 구체적으로는 Fig. 4와 같이, BERT 임베딩된 지식 개체 $g_w^C(c_i) \in R^{768}$, $i(i \le 100)$ 시각적 개념 $g_w^V(v) \in R^{768 \cdot N}$ (N은 시각적 개념의 개수), 자연어 질문 $g_w^Q(q) \in R^{768 \cdot M}$ (M은 질문의 길이)를 Equation (1)과 같이 하나의 벡터 e_i로 이어붙인 뒤, 최대 풀링(Max Pooling) 연산을 거쳐 각각 1024 크기의 노드 개를 생성한다.

$$e_i = Max([g_w^C(c_i), g_w^V(v), g_w^Q(q)]) \tag{1}$$

적어도 하나 이상의 관계(relationship)를 통해 트리플로 묶여있는 개념 노드들끼리는 그래프의 간선을 연결함으로써, GCN을 위한 초기 입력 그래프가 완성된다. 이후 GCN 계층(layer)들을 통과할 때마다, Equation (2)와 같이 간선을 통해 인접 노드의 정보가 유입되어 각 개념 노드의 정보가 새롭게 갱신된다.

$$f_n^{(l+1)} = \sigma(A f_n^{(l)} W^{(l)} + b^{(l)}) \tag{2}$$

여기서 A는 그래프의 인접 행렬, $f_n^{(l)}$는 l 계층 각 노드들의 특징 벡터 값, $W^{(l)}$는 l 계층의 가중치 값, $b^{(l)}$는 l 계층의 바이어스 값을 각각 나타낸다. 이와 같이 GCN 계층들을 통한 그래프 노드들의 갱신이 이루어진 후, 각 노드의 벡터 값들에 다시 최대 풀링 연산이 적용되어 최종적인 지식 벡터 f_k를 생성한다.

3.4 시각적 맥락 추론과 답변 결정

제안 모델의 시각적 맥락 추론 모듈은 서로 다른 입력 데이터인 입력 영상 I와 자연어 질문 q, 그리고 응답 리스트 $r^{(i)}$에 포함된 사물들을 서로 연관지어 맥락 정보를 추출함으로써, 멀티 모달 맥락 벡터 f_c를 생성해내는 역할을 수행한다. 앞서 설명한 바와 같이, 시각적 맥락 추론 모듈은 다시 시각적 접지, 시각적 맥락화, 맥락 정보 임베딩 등의 하위 서브 모듈들로 구성된다.

서로 다른 입력 데이터로부터 멀티 모달 맥락 벡터 f_c를 생성해내기 위한 첫 단계는 질문 q와 응답 리스트 $r^{(i)}$에 등장하는 [person1], [person5]와 같은 각 포인팅(pointing)들을 입력 영상 안의 적절한 사물 영역들과 대응시키는 시각

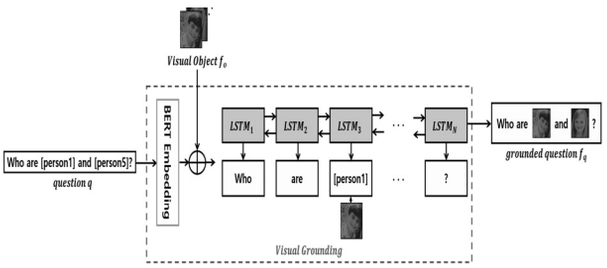

Fig. 5. Visual Grounding Module

적 접지(Visual Grounding)이다. Fig. 5는 시각적 접지 모듈을 통해, 질문에 포함된 각 포인팅들을 영상 안의 인물 영역들로 매칭시킨 결과를 얻는 예를 보여준다. 시각적 접지 모듈은 Fig. 5와 같이 순환 신경망(recurrent neural network)의 하나인 BLSTM (Bidirectional LSTM)을 이용하여, 질문 시퀀스에 포함된 각 포인팅을 입력 영상 내 특정 사물 영역에 대응시킨다.

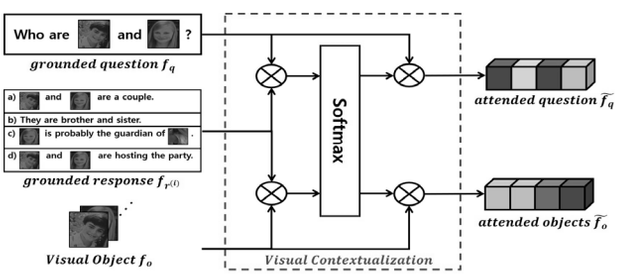

Fig. 6. Visual Contextualization Module

시각적 맥락화 모듈에서는 영상에 포함된 각 사물 영역(visual object) f_o, 접지된 질문(grounded question) f_q, 접지된 응답(grounded response) $f_{r^{(i)}}$들을 토대로, 이들을 서로 연관지어 맥락 정보를 추출하는 역할을 수행한다. 즉, Fig. 6과 같이 하나의 접지된 응답(grounded response) $f_{r^{(i)}}$을 토대로 질문과 영상의 각 사물 영역에 주의 집중(attention) 메커니즘을 적용함으로써, 집중된 질문(attended question) 벡터 $\widetilde{f_q}$와 집중된 사물(attended object) 벡터 $\widetilde{f_o}$를 각각 생성한다. Equation (3)은 응답 $f_{r^{(i)}}$을 토대로 질문 f_q에 대한 주의 집중을 계산하는 식을 나타낸다.

$$\alpha_{i,j} = \underset{j}{softmax}(f_{r^{(i)}} W f_q) \quad \widetilde{f_{q_i}} = \sum_j \alpha_{i,j} f_q \tag{3}$$

맥락 정보 임베딩 모듈은 Fig. 7과 같이, 앞서 생성된 서로 다른 맥락 정보들인 $\widetilde{f_q}$, $\widetilde{f_o}$, $f_{r^{(i)}}$ 등을 역시 순환 신경망의 한 종류인 BLSTM을 통해 순차적으로 결합함으로써, 최종적인 멀티 모달 맥락 벡터 f_c를 생성한다.

마지막으로, 답변 결정 모듈에서는 Fig. 8과 같이 지식 그래프 추론의 결과인 지식 벡터 f_k와 시각적 맥락 추론결과인 멀티 모달 맥락 벡터 를 결합한 뒤, 2개의 완전 연결(FC) 계층과 소프트맥스(softmaxf_c) 계층을 거쳐 최종적으로 응답 리스트 중에서 가장 적합한 답변 $a = f_\theta(q, r^{(i)})$을 결정한다.

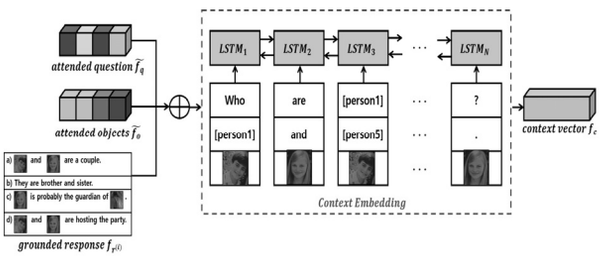

Fig. 7. Context Embedding Module

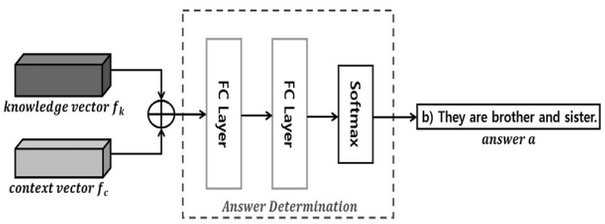

Fig. 8. Answer Determination Module

4. 구현 및 실험

4.1 데이터 집합과 실험 환경

본 논문에서 제안하는 영상 기반 상식 추론 모델인 KG_VCR의 성능 분석 실험을 수행하기 위해, VCR 벤치마크 데이터 집합[2]을 이용하였다. VCR 데이터 집합은 110K개의 영화 장면에서 파생된 290K개의 질문-응답 문제로 구성된다. 본 논문에서는 VCR 데이터 집합에 시각적 개념인 장소, 행동을 추가하였다. 모델 학습에 사용된 영상-질문-응답은 총 32,615쌍이고, 모델 평가를 위한 영상-질문-응답은 총 4,255쌍이다. 제안 모델인 KG_VCR은 Geforce GTX 1080ti GPU가 탑재된 하드웨어와 Ubuntu 16.04 LTS에서 딥러닝 라이브러리인 PyTorch를 이용하여 구현하였다. 모델 학습을 위해 일괄 처리량(batch size)은 32, 반복 횟수(epoch)는 20으로 각각 설정하였다.

4.2 실험

첫 번째 실험은 제안 모델 KG_VCR의 지식 검색 모듈에서 채택한 코사인 유사도(Cosine Similarity) 방법이 본 모델에 적합한지에 대한 답을 얻기 위한 실험이다. Table 1에서 Random은 BERT 임베딩을 거친 모든 트리플 형태의 지식을 임의로 50개 선택하여 사용한 모델이다. 내적(Dot Product)와 코사인 유사도(Cosine Similarity)는 매우 비슷해 보이지만, 내적(Dot Product)은 각도와 크기를 고려하는 방법이고, 코사인 유사도(Cosine Similarity)는 각도만을 고려하는 방법이다. Table 1의 실험에서는 두 개의 GCN 계층(layer)을 가지는 KG_VCR모델을 사용하였다.

Table 1. Performance Comparison with Different Similarity Methods

Method	Q→A	QA→R	Q→AR
Random	0.506	0.615	0.310
Dot Product	0.500	0.609	0.307
Cosine Similarity	**0.521**	**0.646**	**0.334**

Table 1의 실험 결과들에서, 본 논문에서 채택한 코사인 유사도(Cosine Similarity)방법이 다른 방법에 비해 성능이 뚜렷하게 더 높게 나타난 것을 확인할 수 있다. 이는 트리플 구조의 지식을 무작위로 뽑는 것 보다 유사도가 높은 지식을 사용하는 것이 어느 정도 성능을 높일 수 있다는 것을 확인할 수 있었다. 또한, BERT 임베딩으로 모든 트리플 구조의 지식 크기를 맞추기 때문에 각도와 크기를 모두 비교하는 내적(Dot Product)보다는 각도만을 비교하는 코사인 유사도(Cosine Similarity)가 성능이 더 높게 나타났다. 이런 결과들을 통해 KG_VCR에서는 코사인 유사도를 사용하는 것이 효과적이라는 결과를 확인할 수 있다.

두 번째 실험은 제안 모델 KG_VCR에서 채용한 그래프 합성 곱 신경망(GCN) 기반의 질문 및 상식 그래프 임베딩 효과를 입증하기 위한 실험이다. 구체적으로 설명하면, 두 번째 실험은 (1) 상식 그래프(KG) 임베딩 방식으로 과연 GCN이 우수한가? (2) 상식과 더불어 질문(Q)을 함께 임베딩하는 것이 효과적인가? (3) GCN 계층(layer)은 몇 층을 두어야 적절한가? 등의 질문에 관한 답을 얻기 위한 실험이다. 이 실험을 위해 본 논문에서는 Table 2와 같이 임베딩 방식과 대상이 다른 총 5 가지 모델들을 VCR 벤치마크 데이터 집합에 적용해 보고, 각각의 모델 정확도(accuracy)를 비교하였다. 1, 2, 3번 모델들은 모두 그래프 합성 곱 신경망(GCN) 대신 다층 퍼셉트론 신경망(MLP) 모듈로 질문과 지식을 임베딩 하였다. 다만, 1번 모델의 경우 개체(knowledge entity, E)만을, 2번 모델의 경우 개체와 시각적 개념들(visual concepts, VC)만을, 3번 모델의 경우 개체, 시각적 개념들, 질문(question, Q)을 함께 임베딩 하였다. 반면에 4, 5, 6, 7, 8번 모델들은 본 논문의 제안 모델인 KG_VCR과 같이 모두 그래프 합성곱 네트워크(GCN)로 임베딩 하였다. 다만 이들은 시각적 개념들(VC)와 질문(Q)의 임베딩 여부와 그래프 합성 곱의 계층(GCN layer) 수에 차이가 있다.

Table 2의 실험 결과를 정리하면, 먼저 임베딩 방법으로 다층 퍼셉트론(MLP)을 채택한 1, 2, 3번 모델의 성능에 비해 제안 모델 KG_VCR과 같이 그래프 합성 곱(GCN)을 채용한 4, 5, 6, 7번 모델의 성능이 비교적 높게 나타났다. 또한 1, 2, 3번 모델들 간의 성능을 비교해 보면, 시각적 개념(visual concept), 질문(question)을 추가할수록 비교적 성능이 떨어지는 것을 확인할 수 있다. 반면 4, 5, 6번 모델들은 시각적 개념(visual concept), 질문(question)의 정보를 추가할수록 성능 상승하는 것을 확인할 수 있다. 또한 많은 정보를 추가할수록 비교적 단순한 방법인 다층 퍼셉트론(MLP)보다 그래프 합성곱 신경망(GCN)을 사용하는 것이 더 효과적으로

Table 2. Performance Comparison with Different Embedding Methods

#	E	VC	Q	MLP	GCN Layers	Q→A	QA→R	Q→AR
1	√	-	-	√	-	0.495	0.588	0.297
2	√	√	-	√	-	0.491	0.584	0.293
3	√	√	√	√	-	0.487	0.580	0.286
4	√	-	-	-	2	0.498	0.608	0.311
5	√	√	-	-	2	0.501	0.630	0.316
6	√	√	√	-	2	**0.521**	**0.646**	**0.334**
7	√	√	√	-	3	0.521	0.640	0.332
8	√	√	√	-	4	0.504	0.625	0.319

정보를 사용한다는 것을 알 수 있다. 본 실험 범위(2 ~ 4 계층)에서는 그래프 합성곱(GCN)의 계층 수가 2개인 6번 모델이 가장 높은 성능을 보였다. 이런 결과들을 통해, 상식 그래프(KG)와 더불어 질문(Q)도 함께 그래프 합성 곱 신경망(GCN) 모듈로 임베딩하는 제안 모델 KG_VCR의 우수성과 성능개선 효과를 확인할 수 있다.

세 번째 실험은 본 논문에서 제안한 KG_VCR 모델과 기존 연구에서 제안된 다른 모델들과의 성능을 비교하는 실험이다. 실험 결과를 나타내는 Table 3에서 Chance는 주어진 응답 리스트에서 임의로 하나를 정답으로 선택하는 임의 선택 모델을 의미한다. 반면에, VQA[1]는 기존의 영상 기반 질문-응답(VQA) 문제를 위한 베이스라인 모델로서, 별도로 상식 추론 기능을 채용하고 있지는 않다. R2C[2]는 별도의 외부 상식을 활용하지 않는 대신, 답변 결정을 위해 시각적 맥락 추론 기능만을 포함한 심층 신경망 모델이다. KG_VCR은 본 논문에서 제안한 영상 기반 상식 추론 모델을 나타낸다.

Table 3. Performance Comparison with the State-of-art Models

Model	Q→A	QA→R	Q→AR
Chance	0.250	0.250	0.062
VQA[1]	0.286	0.320	0.088
R2C[2]	0.483	0.571	0.272
KG_VCR[Ours]	**0.521**	**0.646**	**0.334**

Table 3의 실험 결과들에서, 본 논문에서 제안하는 KG_VCR 모델이 비교 대상인 VQA, R2C모델에 비해 Q→A, QA→R, Q→AR 등 모든 질문 양식들에서 가장 높은 성능을 보였다는 것을 알 수 있다. 이것은 시각적 상식 추론(VCR) 문제 해결에 제안 모델 KG_VCR과 같이 이미 잘 정의되어 있는 외부 지식 베이스의 상식을 효과적으로 잘 활용하는 것이 성능 개선에 매우 중요한 요소가 될 수 있음을 다시 확인시켜 주는 결과로 볼 수 있다. 따라서 이와 같은 실험 결과들을 바탕으로, 본 논문에서 제안하는 새로운 심층 신경망 모델인 KG_VCR의 우수성과 높은 성능을 확인할 수 있었다.

본 논문에서는 위에서 설명한 정량적 실험들 외에, KG_VCR 모델의 정성적 성능 분석을 위해 KG_VCR 모델이 수행한 실제 작업 사례들을 살펴보았다. Fig. 9는 KG_VCR 모델이

Fig. 9. Some VCR Results Produced by the KG_VCR Model

실행한 작업들이다. ‘right answer and rationale’의 경우 올바른 답변과 근거, ‘right answer, wrong rationale’은 올바른 답변과 올바르지 않은 근거, ‘wrong answer and rationale’은 올바르지 않은 답변과 근거를 선택한 결과이다. 먼저 두 개의 ‘right answer and rationale’ 경우 영상에 담겨 있는 사물(object)의 개념적 관계(conceptual relationship)를 주로 필요로 하는 질문에 대해서는 높은 예측률을 보였다. 하지만 ‘right answer, wrong rationale’, ‘wrong answer and rationale’의 경우와 같이 사람의 감정을 추론하는 질문에 대해서는 ‘right answer and rationale’에 비해 낮은 예측률을 보였다. 이 부분은 후속 연구를 통해 해결해야 할 과제로 본다. 하지만 벤치마크 데이터 집합인 VCR에 포함된 다수의 질의-응답에서 본 논문이 제안한 KG_VCR 모델은 기존 모델들에 비해 매우 우수한 작업 성능을 보여주었다.

5. 결 론

본 논문에서는 영상 기반 상식 추론(VCR) 문제를 위한 새로운 심층 신경망 모델 KG_VCR을 제시하였다. 제안 모델에서는 입력 데이터에서 사물들 간의 관계와 맥락 정보를 추출해내는 모듈들 외에, 별도로 ConceptNet과 같은 외부의 공개된 대규모 지식 베이스로부터 관련 상식들을 직접 가져다 추가적으로 활용할 수 있는 모듈들을 포함하고 있다. 또한, 제안 모델에서는 그래프 합성 곱 신경망(GCN) 모듈을 적용해 지식 그래프를 벡터 형태로 임베딩한 후, 답변 생성에 이용한다. VCR 벤치마크 데이터 집합을 이용한 다양한 실험들을 통해, 본 논문에서 제안하는 KG_VCR 모델의 높은 성능과 효과를 확인할 수 있었다. 하지만 현재의 KG_VCR의 모델은 사람의 감정 추론을 요구하는 문제에 대해서는 올바른 응답을 하지 못하는 경우도 가끔씩 발생한다. 따라서 향후 연구에서는 기존의 KG_VCR모델의 안정화와 더불어 이러한 문제점들을 보완하여 추가적인 성능 개선을 시도해볼 계획이다.

References

[1] S. Antol, A. Agrawal, and J. Lu, et al., “VQA: Visual Question Answering,” in *Proceedings of the International Conference on Computer Vision (ICCV)*, pp.2425-2433, 2015.

[2] R. Zellers, Y. Bisk, and A. Farhadi, et al., “From Recognition to Cognition: Visual Commonsense Reasoning,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.6720-6731, 2019.

[3] P. Wang, Q. Wu, and C. Shen, et al., “FVQA: Fact-based Visual Question Answering,” in *Proceedings of the IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, Vol.40, pp.2413-2427, 2017.

[4] S. Shah, A. Mishra, and N. Yadati, et al., “KVQA: Knowledge-aware Visual Question Answering,” in *Proceedings of Association for the Advancement of Artificial Intelligence (AAAI)*, 2019.

[5] M. Narasimhan, S. Lazebnik, and A. G.Schwing, “Out of the Box: Reasoning with Graph Convolution Nets for Factual Visual Question Answering,” in *Proceedings of the Conference on Neural Information Processing Systems (NIPS)*, pp.2654-2665, 2018.

[6] P. Anderson, X. He, and C. Buehler, et al., “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 (CVPR)*, pp.6077-6086, 2018.

[7] Z. Yang, X. He, J. Gao, L. Deng, and A. Smola, “Stacked Attention Networks for Image Question Answering,” in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.21-29, 2016.

[8] J. Lu, J. Yang, and D. Batra, et al., “Hierarchical Question-Image Co-Attention for Visual Question Answering,” in *Proceedings of the Conference on Neural Information Processing Systems (NIPS)*, pp.289-297, 2016.

[9] M. Lao, Y. Guo, H. Wang, and X. Zhang, “Cross-Modal Multistep Fusion Network With Co-Attention for Visual Question Answering,” in *Proceedings of IEEE Access*, Vol.6, pp.31516-41524, June. 2018.

[10] C. Yang, M. Jiang, B. Jiang, W. Zhou, and K. Li, “Co-Attention Network with Question Type for Visual Question Answering,” in Proceedings of IEEE Access, Vol.7, pp.40771-40781, Mar. 2019.

[11] A. Soren, C. Bizer, and G. Kovilarov, et al., “DBpedia: A Nucleus for a Web of Open Data,” in *Proceedings of The semantic web. Springer*, Berlin, Heidelberg, 2007.

[12] K. Bollacker, C. Evans, and P. Paritosh, et al., “Freebase: A Collaboratively Created Graph Database for Structing Human Knowledge,” in *Proceedings of ACM SIGMOD International Conference on Management of Data*, pp.1247-1250, 2008.

[13] L. Hugo, and S. Singh, “ConceptNet-A Practical Commonsense Reasoning Tool-kit,” *British Telecommunications (BT) Technology Journal*, Vol.22, pp.211-226, 2004.

[14] P. Wang, Q. Wu, and C. Shen, et al., “Explicit Knowledge-based Reasoning for Visual Question Answering,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[15] K. Marino, M. Rastegari, and A. Farhadi, et al., “OK-VQA: A Visual Question Answering Benchmark Requiring External Knowledge,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.3195-3204, 2019.

[16] J. Zhou, G. Cui, and Z. Zhang, et al., "Graph Neural Network: A Review of Methods and Applications," *arXiv preprint arXiv preprint* arXiv:1812.08434, 2018.
[17] Y. LeCun, B. Boser, and J. Denker, et al., "Backpropagation Applied to Handwritten Zip Code Recognition," *Neural Computation*, Vol.1, Issue 4, pp.541-551, 1989.
[18] S. Hochreiter, and J. Schmidhuber, "Long Short-Term Memory," *Neural Computation,* Vol.9, Issue 8, pp.1735-1780, 1997.
[19] T. N, and M. Welling, "Semi-Superviced Classification with Graph Convolutional Networks," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2017.
[20] J. Yang, J, Lu and S. Lee, et al., "Graph R-CNN for Scene Graph Generation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp.670-685, 2018.
[21] Y. Cao, M, Fang and D. Tao, et al., "BAG: Bi-directional Attention Entity Graph Convolutional Network for Multi-hop Reasoning Question Answering," arXiv preprint arXiv:1904.04969, 2019.
[22] J. Devlin, M, Chang and K. Lee, et al., "BBert: Pre-training of Deep Bidirectional Transformers for Language Understanding," arXiv preprint arXiv:1810.04805, 2018.

이 재 윤
https://orcid.org/0000-0003-4913-3691
e-mail : jaeyun_95@kyonggi.ac.kr
2019년 경기대학교 컴퓨터과학과(학사)
2019년 ~ 현 재 경기대학교 컴퓨터과학과 석사과정
관심분야 : 인공지능, 컴퓨터비전, 상식 추론, 로봇지능

김 인 철
https://orcid.org/0000-0002-5754-133X
e-mail : kic@kyonggi.ac.kr
1985년 서울대학교 수학과(이학사)
1987년 서울대학교 전산과학과(이학석사)
1995년 서울대학교 전산과학과(이학박사)
1996년 ~ 현 재 경기대학교 컴퓨터과학과 교수
관심분야 : 인공지능, 기계학습, 로봇지능