KIPS Trans. Softw. and Data Eng.
Vol.8, No.5 pp.205~212 pISSN: 2287-5905

https://doi.org/10.3745/KTSDE.2019.8.5.205

Deep Neural Network-Based Scene Graph Generation for 3D Simulated Indoor Environments

Donghyeop Shin† · Incheol Kim††

ABSTRACT

Scene graph is a kind of knowledge graph that represents both objects and their relationships found in a image. This paper proposes a 3D scene graph generation model for three-dimensional indoor environments. An 3D scene graph includes not only object types, their positions and attributes, but also three-dimensional spatial relationships between them, An 3D scene graph can be viewed as a prior knowledge base describing the given environment within that the agent will be deployed later. Therefore, 3D scene graphs can be used in many useful applications, such as visual question answering (VQA) and service robots. This proposed 3D scene graph generation model consists of four sub-networks: object detection network (ObjNet), attribute prediction network (AttNet), transfer network (TransNet), relationship prediction network (RelNet). Conducting several experiments with 3D simulated indoor environments provided by AI2-THOR, we confirmed that the proposed model shows high performance.

Keywords : Scene Graph, 3D Indoor Environment, Deep Neural Network, AI2-THOR

3차원 가상 실내 환경을 위한 심층 신경망 기반의 장면 그래프 생성

신 동 협† · 김 인 철††

요 약

장면 그래프는 영상 내 물체들과 각 물체 간의 관계를 나타내는 지식 그래프를 의미한다. 본 논문에서는 3차원 실내 환경을 위한 3차원 장면 그래프를 생성하는 모델을 제안한다. 3차원 장면 그래프는 물체들의 종류와 위치, 그리고 속성들뿐만 아니라, 물체들 간의 3차원 공간 관계들도 포함한다. 따라서 3차원 장면 그래프는 에이전트가 활동할 실내 환경을 묘사하는 하나의 사전 지식 베이스로 볼 수 있다. 이러한 3차원 장면 그래프는 영상 기반의 질문과 응답, 서비스 로봇 등과 같은 다양한 분야에서 유용하게 활용될 수 있다. 본 논문에서 제안하는 3차원 장면 그래프 생성 모델은 크게 물체 탐지 네트워크(ObjNet), 속성 예측 네트워크(AttNet), 변환 네트워크(TransNet), 관계 예측 네트워크(RelNet) 등 총 4가지 부분 네트워크들로 구성된다. AI2-THOR가 제공하는 3차원 실내 가상환경들을 이용한 다양한 실험들을 통해, 본 논문에서 제안한 모델의 높은 성능을 확인할 수 있었다.

키워드 : 장면 그래프, 3차원 실내 환경, 심층 신경망, AI2-THOR

1. 서 론

최근 딥러닝 기술의 발전에 따라 영상 처리 분야에서 다양한 연구가 진행되어 왔다. 이에 따라 초기의 영상 분류(Image Classification)부터 물체 탐지(Object Detection), 의미적 분할(Semantic Segmentation)까지 점점 더 복잡한 문제를 해결할 수 있게 되었다[1]. 최근에는 영상의 고수준 이해를 위해, 주어진 영상에 대한 장면 그래프 생성 연구에 관심이 높아지고 있다[2]. 장면 그래프란 영상 내 물체들과 물체 간의 관계를 나타내는 지식 그래프이다. Fig. 1은 장면 그래프의 예시를 나타낸다. 먼저 영상으로부터 “woman”, “horse” 등의 물체를 찾아내고, 탐지된 물체들로부터 “wearing”, “riding” 등의 관계를 알아낸다. 이후 물체 정보와 관계 정보를 통합하여 하나의 장면

※ 이 논문은 2018년도 정부(과학기술정보통신부)의 재원으로 정보통신기술진흥센터의 지원을 받아 수행된 연구임(No.2018-0-00677, 촉각이 가능한 로봇 손으로 다양한 물체를 다루는 방법과 절차를 학습하는 로봇 손 조작 지능 개발).
※ 이 논문은 2018년도 한국정보처리학회 추계학술발표대회에서 ‘3차원 공간에서 에이전트의 탐색을 통한 장면 그래프 생성’의 제목으로 발표된 논문을 확장한 것임.
† 준 회 원 : 경기대학교 컴퓨터과학과 석사과정
†† 종신회원 : 경기대학교 컴퓨터과학과 교수
Manuscript Received : December 27, 2018
First Revision : February 27, 2019
Accepted : March 9, 2019
* Corresponding Author : Incheol Kim(kic@kyonggi.ac.kr)

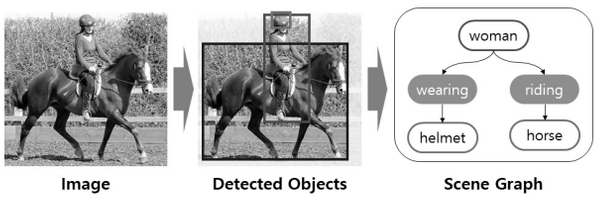


Fig. 1. An Example of Scene Graph

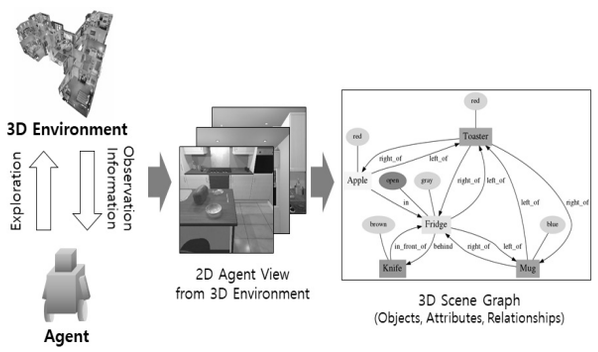


Fig. 2. 3D Scene Graph Generation

그래프로 표현한다. 이러한 장면 그래프는 영상을 고수준의 지식 형태로 나타내낼 수 있으며, 다양한 영상 처리 문제에 기초 지식이 될 수 있다는 장점이 있다. 하지만 기존 장면 그래프 생성 연구들은 오직 2차원 영상을 대상으로 장면 그래프를 생성하였다. 이에 따라 실세계에 해당하는 3차원 공간에서 기존의 장면 그래프 생성 기술을 적용하기 어렵다. 본 논문에서는 이러한 한계점을 극복하고자, 3차원 실내 가상 환경에서의 3D 장면 그래프를 생성하는 방법을 제안한다. 3D 장면 그래프는 주어진 3차원 공간 내 물체들과 물체들의 관계를 나타내며, 각 물체들의 위치 정보는 3차원 좌표로 표현된다. 본 논문에서는 3D 장면 그래프를 Fig. 2와 같은 방법으로 생성한다. 먼저 에이전트는 주어진 3차원 실내 가상 환경을 탐색하고, 탐색 과정 중 탐지된 물체 정보들을 바탕으로 3D 장면 그래프를 생성한다. 이때 에이전트가 인식하는 시각 정보는 에이전트 시점에서의 2차원 영상이다. 따라서 기존 2차원 영상에서의 장면 그래프 생성 방법이 부분적으로 적용될 수 있다. 하지만 3차원 공간이기 때문에, 에이전트의 위치 및 시점에 따라 동일 물체가 중복 탐지될 수 있는 문제가 발생한다. 또한 2차원 영상에서 탐지된 물체의 위치 정보는 상대적인 2차원 좌표 공간에 있기 때문에, 이를 3차원 공간의 위치로 변환하는 작업이 요구된다. 본 논문에서는 이러한 문제들을 해결하고, 물체의 종류(Class), 3차원 위치 및 크기, 속성(Attribute)들과 물체 간의 공간 관계(Spatial Relationship)들을 인식하는 3D 장면 그래프 생성 모델을 제안한다. 또한 학습 및 실험을 위해 실내 가상환경인 AI2-THOR[3]를 이용하였다.

2. 관련 연구

최근에는 장면 그래프를 생성하기 위한 다양한 연구들이 진행되고 있다[6]. 일반적인 장면 그래프 생성 과정은 물체 탐지(Object Detection) 과정과 관계 예측(Relationship Prediction) 과정으로 나뉜다[4, 5]. 물체 탐지 과정에서는 기존에 잘 알려진 Faster R-CNN[6] 등의 물체 분류기를 사용한다. 관계 예측 과정에서는 탐지된 물체들의 다양한 특징(Feature)들을 활용하여 관계를 예측한다. 대부분의 연구에서 물체의 시각 특징(Visual Feature)을 기반으로, 다른 특징들과 결합하여 예측에 사용하였다. [7]의 연구에서는 물체 종류 특징을 추가로 사용하였다. 이는 두 물체의 종류에 따라, 주로 발생하는 관계 정보를 이용하기 위함이다. 또한 [8]의 연구에서는 물체들의 공간 관계를 파악하기 위해, 각 물체의 위치 정보를 사용하였다. 하지만 기존의 장면 그래프들은 2차원 영상을 대상으로 한다. 이에 따라 3차원 공간에서는 적합하지 않다는 한계점이 있다.

[9]의 연구에서는 기존 연구와 다르게, 3차원 공간에서 장면 그래프를 생성하는 방법을 제안하였다. 주어진 2차원 영상 시퀀스로부터 물체 후보 영역을 탐지하고, 해당 영역들의 시각 특징 및 위치 정보를 이용하여 물체 및 관계를 예측하였다. 하지만 에이전트와 상호 작용이 불가능한 환경을 가정하고, 3차원 좌표를 알아내기 위해 동일 물체를 다양한 관점에서 관측해야한다는 제약 조건이 있다. 또한 예측하는 공간 관계가 비교적 단순하고, 물체들의 공간적 정보만을 예측할 뿐, 물체들의 상태는 알 수 없다는 한계점이 있다. 본 논문에서는 이러한 한계점을 보완하기 위해, 에이전트와 상호 작용이 가능한 환경에서, 물체 상태를 포함한 3차원 장면 그래프를 생성하는 방법을 제안한다.

기존의 3차원 공간 문제는 기존 2차원 영상에 대한 연구의 확장으로 이어졌다. 대표적으로 3차원 물체 탐지(3D Object Detection) 문제에서는 2차원 영상으로부터 물체 영역을 찾아내고, 물체 모양에 부합되는 3차원 경계 상자를 예측하였다[10]. 3차원 의미적 분할(3D Semantic Segmentation) 문제는 물체 종류를 2차원 픽셀(Pixel)이 아닌, 3차원의 복셀(Boxel) 단위로 예측한다[11]. 하지만 이러한 문제들은 모두 에이전트와의 상호작용이 불가한 2차원 영상만을 이용한다. 또한 관측자의 위치를 고려하지 않기 때문에 예측된 3차원 위치의 기준이 모호하다. 이에 따라 3차원 공간에서의 중복 탐지, 절대 위치 예측 등의 문제를 해결하기 어렵다. 본 논문에서는 3차원 공간에서 이러한 문제를 해결하기 위한 방법을 제안한다.

3. 장면 그래프 생성 모델

본 논문에서는 3차원 실내 환경에서 장면 그래프를 생성하기 위한 효과적인 모델을 제안한다. 모델의 전체 구조도는 Fig. 3에서 표현되어 있다. 제안 모델은 물체 탐지 네트워크(Object Detection Network, ObjNet), 속성 예측 네트워크(Attribute Prediction Network, AttNet), 변환 네트워크(Transfer Network, TransNet), 관계 예측 네트워크(Relationship Prediction Network, RelNet)로 총 4가지 부분 네트워크들로 구성된다. ObjNet은 에이전트가 현재 바라보는

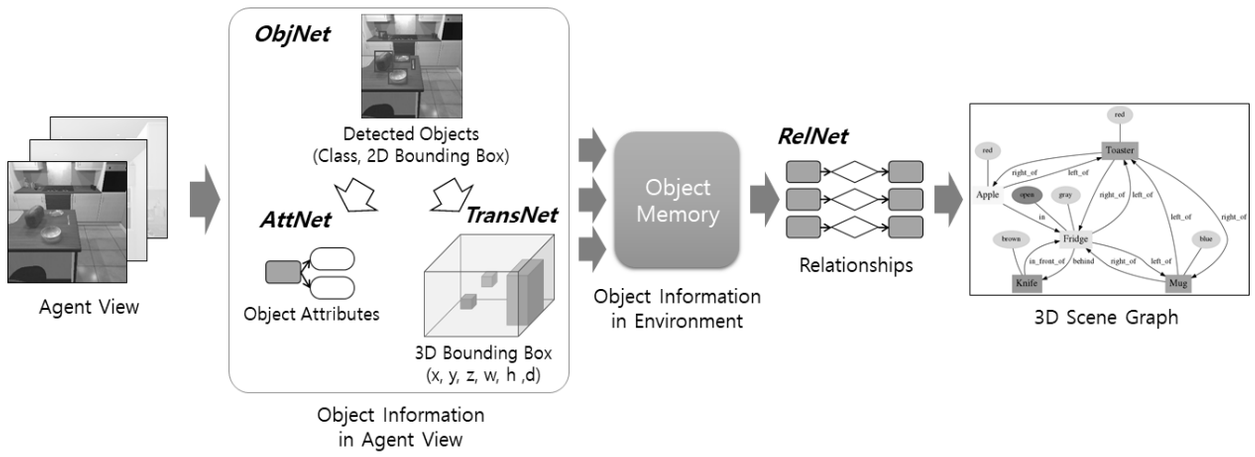


Fig. 3. 3D Scene Graph Generation Model

영상에서 물체들의 영역과 종류(Class)를 찾는다. AttNet은 찾아낸 물체들의 여러 속성들을 예측한다. 또한 TransNet은 물체들의 좌표를 현재 에이전트의 관점이 아닌, 3차원 공간상의 절대 관점으로 변환한다. 에이전트는 위 세 네트워크를 이용하여 인식한 물체들을 물체 메모리(Object Memory)에 저장한다. 물체 메모리에서는 물체들의 3차원 좌표와 종류를 기준으로 중복된 물체들을 제거한다. 마지막으로 RelNet은 물체 메모리에 저장된 물체들의 공간 관계를 예측한다. 이후 예측된 결과들을 종합하여 장면 그래프가 생성된다.

3.1 물체 탐지 네트워크 (ObjNet)

물체 탐지 네트워크(ObjNet)은 Fig. 3과 같이 에이전트가 받아들이는 2차원 영상으로부터, 영상 내 물체들의 좌표와 종류를 알아낸다. 본 논문에서는 IQA[12]에서 YOLOv3[13] 물체 탐지기를 이용하여 학습시킨 네트워크를 차용하였다. 그 이유는 YOLOv3 물체 탐지기의 장점인 속도와 정확성에 있다. YOLOv3 물체 탐지기는 물체 영역 탐지와 물체 분류를 동시에 수행하는 한-단계(One-State) 모델이다. 이에 따라 빠른 처리 속도로 실시간 영상 처리가 가능하다. 또한 YOLOv3 물체 탐지기의 네트워크 구조 상 작은 물체들을 탐지하는 데에 유리한 장점이 있다. 이는 작은 물체가 많이 등장하는 AI2-THOR 환경에서 매우 적합한 특징으로 볼 수 있다.

3.2 속성 예측 네트워크 (AttNet)

속성 예측 네트워크(AttNet)은 Fig. 4와 같은 구조를 갖는다. 영상, 물체들의 좌표, 물체들의 종류를 입력받아 물체들의 속성들을 예측한다. 본 논문에서는 물체들의 속성으로 색상(Color)과 개폐 상태(Open State)를 선택하였다. 색상은 해당 물체의 색상 종류를 나타내고, 개폐 상태는 해당 물체가 열려있는 상태인지, 닫혀있는 상태인지 혹은 열고 닫을 수 없는 물체인 지를 나타낸다.

AttNet에서는 속성들을 예측하기 위해, 합성 곱 신경망(Convolutional Neural Network, CNN)을 이용하여 영상의 시각 특징을 추출하였다. 이는 영상이 특정 부분에서 나타내는 모양을 알아내기 위함이다. 한편 색상 예측의 경우, 영상 데이터 자체가 갖는 색상 값인 RGB 값이 중요하기 때문에, 추가적으로 1x1 크기의 필터를 갖는 CNN을 사용하였다. 이는 영상의 각 픽셀에 해당하는 3가지 색상 값만을 이용하여

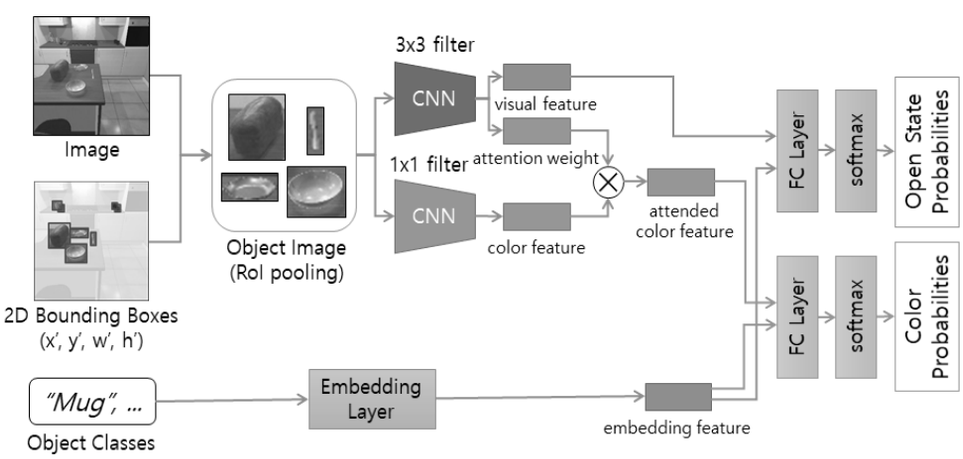


Fig. 4. Attribute Prediction Network (AttNet)

픽셀에 대한 색상 벡터를 추출하기 위함이다. 시각 특징들은 물체 종류로부터 얻은 임베딩 특징(Embedding Feature)과 결합하여, 각 속성들을 예측하는 데에 사용하였다.

3.3 변환 네트워크 (TransNet)

변환 네트워크(TransNet)는 Fig. 5와 같은 구조를 가지며, ObjNet 결과 물체들의 2차원 좌표를 3차원 공간상의 좌표로 변환한다. 변환 네트워크에서는 에이전트 위치를 기반으로 좌표를 변환하기 위해, 에이전트의 위치와 영상에서의 물체 좌표로부터 위치 특징(Position Feature)을 추출한다. 그리고 물체와 에이전트 간의 깊이 정보를 주기 위해 깊이 영상(Depth Image)을 이용하여 합성곱 신경망으로 깊이 특징(Depth Feature)을 추출한다. 그리고 두 특징으로부터 완전 연결 층을 통해, 3차원 공간의 좌표 정보인 <x, y, z, width, height, depth>을 예측한다.

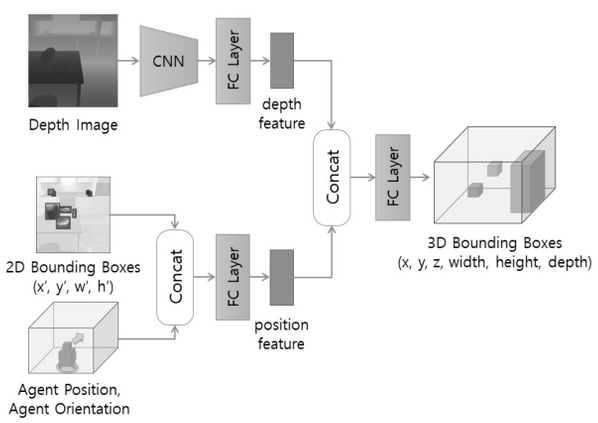
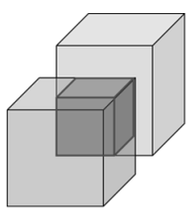


Fig. 5. Transfer Network (TransNet)

3.4 물체 메모리 (Object Memory)

ObjNet, AttNet을 거쳐 탐지된 물체와 그 정보들은 TransNet로 변환된 좌표를 기준으로 물체 메모리(object memory)에 저장된다. 이후 에이전트가 새로운 영상을 받아서 또 다른 물체들을 인식하면, 같은 방법으로 물체 메모리에 저장된다. Fig. 6은 물체 메모리에 3차원 환경에 대한 물체 정보를 저장하는 과정을 나타낸다. 이와 같은 과정을 통해, 에이전트가 인식한 물체 정보들을 저장할 수 있다.

하지만 3차원 환경에서는 에이전트의 탐색 과정 중, 동일 물체의 중복 인식 문제가 발생한다. 또한 시점에 따라 물체의 상대적 위치가 변한다. 이에 따라 TransNet의 예측 오류로 인해, 인식된 물체들의 3차원 위치 정보가 조금씩 달라질 수 있는 문제가 발생한다. 본 논문에서는 이러한 문제들을 해결하기 위해, NMS(Non-Maximum Suppression) 연산으로 중복된 물체를 제거한다. NMS 연산은 Yolo, SSD 등의 물체 탐지 모델에서 중복되는 물체를 제거하기 위해 사용되는 방법이다. 다만 기존에 2차원 영상에서 적용된 것과는 달리, 3차원 공간상에서 NMS 연산을 수행한다. 연산 과정은 다음과 같다. 먼저 종류가 같은 두 물체에 대해, 겹치는 비율인 3차원 IoU(Interest of Union)를 구한다. 이는 두 물체의 교집합(Intersection) 공간 부피를 합집합(Union) 공간 부피로 나눈 값이다. 이는 Fig. 7과 같이 나타난다.



$$\textbf{3D IOU} = \frac{\textbf{Intersection}(obj_1, obj_2)}{\textbf{Union}(obj_1, obj_2)}$$

Fig. 7. 3D Intersection over Union (3D IoU)

3차원 IoU가 일정 수준보다 높으면 물체 확률이 낮은 물체를 제외시킨다. 이러한 과정을 통해 물체 메모리에 중복 물체들을 제거한다.

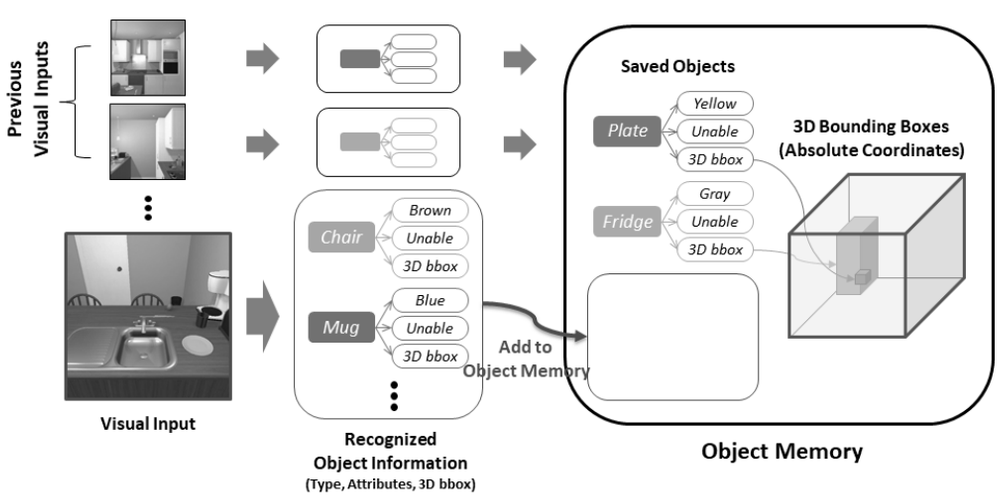


Fig. 6. Storing Object Information in Object Memory

3.5 관계 예측 네트워크 (RelNet)

관계 예측 네트워크(RelNet)은 Fig. 8과 같은 구조를 갖는다. RelNet은 물체 메모리에 저장된 물체들의 공간 관계를 예측한다. 이를 위해 물체의 3차원 공간 좌표, 물체 종류를 입력받고, 두 물체의 입력 특징을 이용하여 물체간의 관계를 예측한다.

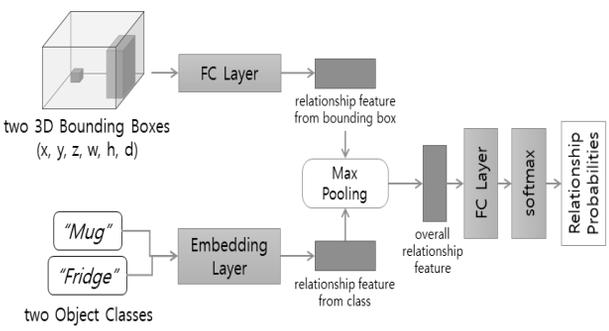


Fig. 8. Relationship Recognition Network (RelNet)

이 네트워크 구조는 각 특징으로부터 두 관계 특징(Relationship Feature)을 추출한 뒤, 두 특징을 결합하여 관계 분포를 예측한다. 각 관계 특징을 추출하기 위해, 물체의 종류의 경우 임베딩 층을 이용하고, 물체 좌표의 경우 완전 연결 층을 이용한다. 이후 두 관계 특징을 최대 풀링(Max Pooling) 연산으로 결합한다. 이 결합 방법은 3장의 성능 실험을 통해 결정하였다. 이후 결합된 관계 특징을 완전 연결 층에 입력하여 관계를 예측한다.

4. 구현 및 실험

4.1 데이터 집합

제안한 네트워크들을 AI2-THOR[3]에서 학습 및 평가하기 위해, 데이터 수집 프로그램을 개발하여 데이터 집합을 직접 구축하였다. AI2-THOR에는 여러 실내 공간이 정의되어 있다. 또한 환경 내 에이전트는 그리드(Grid) 환경에서 이동 및 90도 회전이 가능하며, 개폐 가능한 물체의 상태를 조작할 수 있다. 본 논문에서는 여러 실내 공간 내에서 에이전트의 행동 결과에 대한 장면 그래프 데이터를 수집하였다. 정의된 실내 공간 중 FloorPlan1~FloorPlan11 환경에 대한 데이터들을 수집하였고, FloorPlan2~FloorPlan9는 모델 학습에, FloorPlan10~FloorPlan11은 모델 평가에 사용하였다. 수집한 데이터 집합은 총 3,916개의 장면 그래프 및 영상 데이터, 13,109개의 물체 데이터, 25,583개의 관계 데이터를 포함하며, 각 물체 데이터는 색상과 개폐 상태의 속성 정보를 갖는다. 또한 이들은 25개의 물체 종류, 6개의 색상 종류, 3개의 개폐 상태 종류, 6개의 관계 종류로 구성된다. 이 중 개폐 상태 종류는 “opened”, “closed”, “unable”이며 “unable”은 “Apple”과 같이 여닫을 수 없는 속성을 의미한다. 그리고 관계 종류는 “in”, “on”, “in_front_of”, “behind”, “left_of”, “right_of”이다. 한편, 학습에 사용된 데이터 집합은 3,525개의 장면 그래프 및 영상 데이터, 11,831개의 물체 데이터, 23,007개의 관계 데이터로 구성되며, 평가에 사용된 데이터 집합은 391개의 장면 그래프 및 영상 데이터, 1,278개의 물체 데이터, 2,576개의 관계 데이터로 구성된다.

4.2 구현

본 논문에서 제안한 모델의 부분 네트워크들은 서로 독립적으로 학습된다. 물체 탐지 네트워크(ObjNet)는 기존 IQA[11]에서 학습된 모델을 차용하였고, 나머지 네트워크들은 수집한 데이터를 바탕으로 학습하였다. 속성 예측 네트워크(AttNet)와 관계 예측 네트워크(RelNet)의 경우, 분류(Classification) 문제를 해결하기 위해 크로스엔트로피(Cross-Entropy) 손실 함수를 사용하였고, 변환 네트워크(TransNet)는 회귀(Regression)를 위해 평균 절대 오차(Mean Absolute Error, MAE) 손실 함수를 사용하였다. AttNet, TransNet, RelNet의 학습률(Learning Rate)은 모두 0.001로 설정하였고, 최적화 함수(Optimizer)는 모두 적응적 모멘트 추정(Adaptive Moment Estimation)을 사용하였다. 또한 제안한 네트워크들을 구현하기 위해 Ubuntu 16.04 LTS 환경에서 Python 딥러닝 라이브러리인 PyTorch를 사용하였다.

4.3 실험

본 논문에서는 부분 네트워크들의 개별 성능 실험과 통합 모델의 장면 그래프 생성 실험을 진행하였다. 또한 각 실험을 통해 각 부분 네트워크 및 통합 모델의 성능 및 특징을 분석하였다. 먼저 Table 1은 AttNet의 입력 특징별 성능 실험 결과를 나타낸다. 각 척도는 분류 결과에 대한 정밀도(Precision), 재현율(Recall), 정확도(Accuracy)를 나타낸다. 본 실험에서는 주어진 입력 특징들이 성능에 얼마나 영향을 미치는 지 실험하였다. Table 1의 결과를 통해 물체들의 시각 특징뿐만 아니라 종류 특징도 성능에 좋은 영향을 미치는 것을 볼 수 있다. 이는 물체의 종류에 따라 색상과 개폐 상태를 예측할 수 있기 때문이다. 예를 들어, “Apple”은 대부분 “red”이고, “Spoon”은 대부분 “gray”라는 특징이 있다. 또 두 물체는 종류로부터 “unable”인 개폐 상태임을 바로 알 수 있다. 하지만 “opened” 혹은 “closed”의 개폐 상태와 물체들의 정확한 색상은 시각 정보가 요구되기 때문에, 예측에 한계성이 있다. Table 1에서 물체 종류만이 입력된 경우의 낮은 성능이 이를 뒷받침한다.

Table 1. Performance Analysis of AttNet

Input Feature	Output	Precision (%)	Recall (%)	Acc (%)
image	color	89.23	79.76	87.52
	open state	82.24	33.33	82.24
class	color	53.01	43.39	49.99
	open state	72.03	71.89	92.42
image + class	**color**	**89.71**	**80.55**	**89.87**
	open state	**79.57**	**79.48**	**94.57**

Table 2는 TransNet의 입력 데이터별 성능 실험 결과를 나타낸다. 각 척도는 회귀 결과에 대한 평균 절대 오차와 정확도를 나타낸다. 본 실험에서는 회귀 결과에 대한 정확도를 측정하기 위해, 예측 결과와 정답과의 3차원 IoU가 0.3 이상일 때 정답으로 판정하였다. TransNet의 입력 데이터로 에이전트의 위치 정보(Agent Position), 물체의 2차원 공간 정보인 경계 상자(Bounding Box, bbox)를 기본으로 사용하였고, 깊이 영상과 물체의 종류(Class)의 사용 결과를 비교하였다. Table 2의 결과를 통해, 깊이 영상과 물체의 종류 모두 성능에 도움이 되는 것을 볼 수 있었다. 먼저 깊이 영상을 추가로 사용하였을 때, 물체의 위치(Position) <x, y, z> 오차가 크게 줄어드는 것을 알 수 있다. 이는 에이전트와 물체간의 거리 정보를 통해, 물체의 2차원 위치 정보를 3차원 위치 정보로 보다 정확하게 치환할 수 있기 때문이다. 반면에 물체의 종류 사용은 물체의 크기(Size) <width, height, depth> 오차를 크게 줄였다. 이는 물체의 종류에 따라 일반적인 크기 정보를 학습할 수 있기 때문이다. 또한 두 정보를 모두 사용하였을 때, 물체의 위치 및 크기 오차가 모두 낮아지고, 예측 정확도가 제일 높은 것을 확인할 수 있었다.

Table 2. Performance Analysis of TransNet

Input Feature	Output	MAE		Acc (%)
		Each	Total	
agent position + bbox	position	0.2535	0.4393	29.73
	size	0.1858		
agent position + bbox + depth image	position	0.1192	0.2841	66.32
	size	0.1649		
agent position + bbox + class	position	0.1961	0.3325	42.27
	size	**0.1364**		
agent position + bbox + depth image + class	**position**	**0.1077**	**0.2512**	**72.16**
	size	0.1435		

Table 3은 RelNet의 입력 데이터 및 추출된 특징들의 결합 방법에 대한 성능 실험 결과를 나타낸다. 각 척도는 분류 결과에 대한 정밀도, 재현율, 정확도를 나타낸다. RelNet의 입력 데이터로는 두 물체의 3차원 좌표 정보와 두 물체의 종류가 사용된다. 결과적으로 두 입력 데이터를 모두 사용한 것이 성능이 높았다. 이는 물체의 종류에 따라 관계를 예측하는 것이 어느 정도 가능하기 때문이다. 예를 들어, “Apple”과 “Fridge”가 주어진다면 “in”이라는 관계를 예측할 수 있을 것이다. 또한 결합 방법에 대한 실험 결과, 최대 풀링 연산을 적용한 것이 가장 높은 성능을 보였다. 이는 두 관계 특징 중, 확실한 특징을 통해 결과를 예측하는 것이 더 효율적이라는 것을 나타낸다.

마지막 실험에서는 각 부분 네트워크들을 모두 사용한 통합 모델의 3차원 장면 그래프 생성 성능을 평가하였다. 장면 그래프 생성 시 물체 메모리에서 3차원 IoU 값이 0.3 이상인

Table 3. Performance Analysis of RelNet

Input Feature	Fusion Method	Precision (%)	Recall (%)	Acc (%)
bbox	-	70.71	96.46	83.22
class	-	50.55	57.02	47.10
bbox + class	concatenate	75.81	97.26	85.27
	element-wise product	73.60	96.53	83.73
	plus	71.89	97.18	84.81
	max pooling	**75.94**	**97.40**	**86.74**

물체들을 중복 물체로 간주하였다. 한편 3차원 환경에서의 장면 그래프 생성 정확도 측정을 위해, 기존 2차원 장면 그래프 [4, 5, 7, 8]의 성능 척도인 SGGen을 기반으로 3dSGGen을 정의하였다. 3dSGGen은 각 장면 그래프 내 <주격 물체 - 관계 - 목적격 물체>로 이루어진 트리플(Triple)들의 평균 재현율을 나타낸다. 예측된 각 트리플은 주격 물체와 목적격 물체의 위치, 종류, 두 속성 종류, 관계 종류가 정답 트리플과 일치할 때 정답으로 판정된다. 이 때 물체의 위치는 3차원 IoU 값이 일정치 이상일 때 일치하는 것으로 판단한다.

Table 4는 각 부분 네트워크 사용에 따른, 물체 탐지 및 장면 그래프 생성 성능을 나타낸다. 물체 탐지 성능은 정답 물체와의 3차원 IoU가 0.3 이상일 때를 기준으로 측정하였고, 3dSGGen은 0.3, 0.4, 0.5의 3차원 IoU 기준치에 따라 측정하였다. Table 4에서 O는 ObjNet, A는 AttNet, T는 TransNet, R은 RelNet을 나타내며, 각 모델은 해당 부분 네트워크를 사용하고 나머지는 정답 값을 사용하였다. 실험 결과, 통합 모델의 성능은 IoU=0.3 척도를 기준으로 40.30%로 측정되었다. 이는 주어진 환경에 대한 3차원 장면 그래프를 일정 수준으로 생성할 수 있음을 나타낸다. 하지만 Table 4의 결과를 통해 AttNet과 TransNet 사용 시 장면 그래프 생성 성능이 크게 줄어드는 것을 볼 수 있다. 이는 ObjNet의 부정확한 예측 결과가 다음 부분 네트워크들에 영향을 미치기 때문이다. ObjNet 사용 시 물체 탐지 성능이 크게 줄어드는 것을 통해, ObjNet가 일정 수준 부정확한 결과를 출력한다고 판단할 수 있다. 향후 연구에는 ObjNet의 성능을 개선하고, 다른 부분 네트워크들이 노이즈(Noise)에 강인하도록 학습하는 방법을 고안해야 할 것이다.

Table 4. Performance Analysis of Total Model

Model	Object (%)		3dSGGen (%)		
	Precision	Recall	IoU=0.3	IoU=0.4	IoU=0.5
R	100.00	100.00	98.16	98.16	98.16
O+R	100.00	82.17	82.34	82.34	82.34
O+A+R	100.00	82.17	61.09	61.09	61.09
O+A+T+R	66.97	63.71	40.30	21.42	15.26

Fig. 9는 AI2-THOR 환경에서 통합 모델로부터 3차원 장면 그래프를 생성한 정성적 결과를 나타낸다. 각 결과의 왼쪽

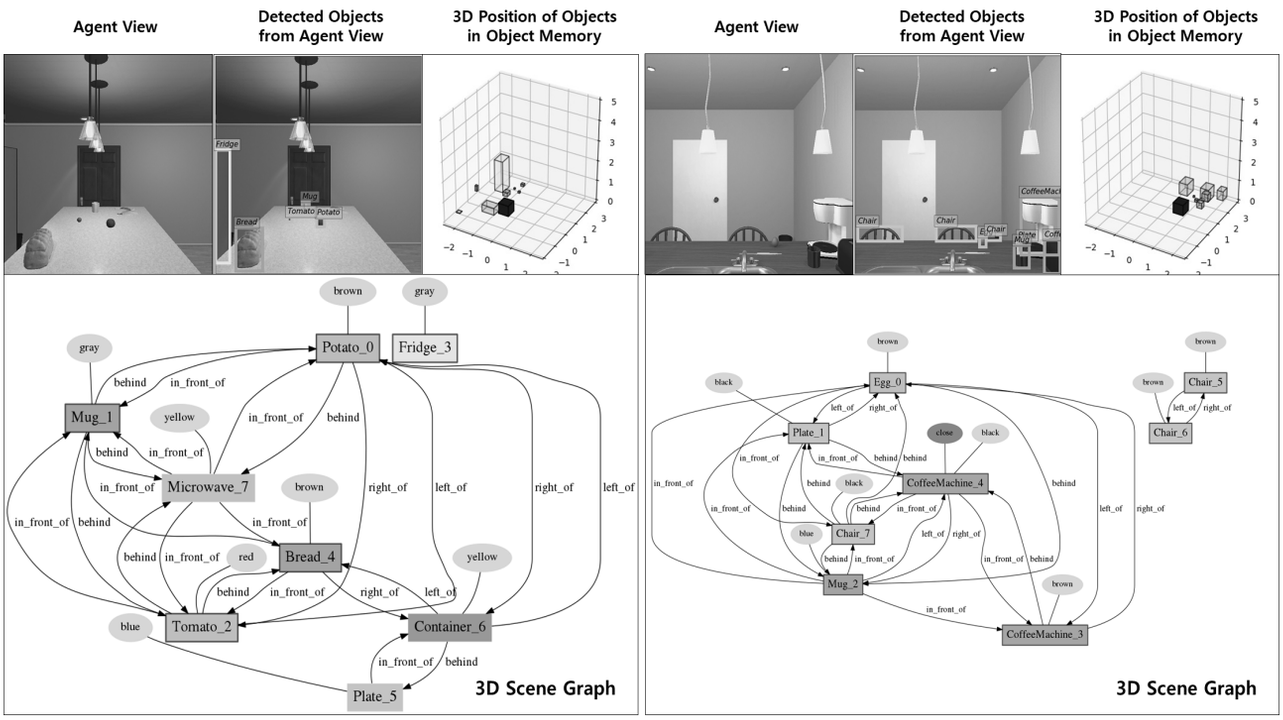


Fig. 9. 3D Scene Graphs Generated by the Proposed Model

위 그림은 현재 에이전트가 바라보는 시점의 2차원 영상이고, 가운데 그림은 2차원 영상으로부터 ObjNet을 통해 탐지된 물체들이다. 오른쪽 그림은 에이전트의 탐색을 통해, 물체 메모리에 저장된 물체들의 3차원 위치를 나타낸다. 마지막으로 아래 그림은 에이전트의 탐색을 통해 생성된 3차원 장면 그래프를 나타낸다. 3차원 장면 그래프 내에 빨간색 테두리의 물체는 현재 2차원 영상에서 관측된 물체들을 나타낸다. 두 결과 중 왼쪽의 결과는 주어진 환경에 대해 적절한 3차원 장면 그래프를 생성한 경우이고, 오른쪽 결과는 물체 탐지가 잘못된 경우이다. 오른쪽 그림에서는 실제 "Chair"는 2개이다. 하지만 ObjNet은 3개로 인식하였고, 그에 따라 최종적으로 생성된 3차원 장면 그래프에 잘못된 물체 정보가 반영 되었다. 이는 이전 부분 네트워크의 예측 오류가 다음 부분 네트워크에까지 반영되는 한계점을 나타낸다.

5. 결 론

본 논문에서는 주어진 3차원 실내 환경의 물체 정보들을 나타내기 위해, 기존의 장면 그래프를 확장한 3D 장면 그래프를 생성하였다. 이는 주어진 3차원 공간에서 물체들의 종류, 3차원 위치 및 크기, 속성들과 각 물체 간의 3차원 공간 관계를 나타내는 지식 그래프이다. 본 논문에서는 3차원 장면 그래프 생성에 요구되는 기능에 따라, 4가지 부분 네트워크들을 정의하였다. 또한 각 물체들의 정보를 물체 메모리에 저장하고, 3차원 공간에서의 NMS를 통해 중복 탐지된 물체들을 제거하였다. 그리고 제안한 모델을 검증하기 위해, AI2-THOR 가상 환경에서 데이터를 수집하고, 각 부분 네트워크들을 학습 및 검증 하였다. 또한 통합 모델을 통한 3차원 장면 그래프 생성 실험을 통해 통합 모델을 검증하였다. 하지만 통합 모델에서, 이전 부분 네트워크가 발생시킨 오차가 다음 부분 네트워크들의 예측 정확도를 저하시키는 한계점이 나타났다. 추후 연구에서는 모델 보완을 통해, 이러한 한계점을 보완해야 할 것이다.

References

[1] Y. Guo, Y. Liu, and A. Oerlemans et al., "Deep Learning for Visual Understanding: A Review," *Neurocomputing*, Vol.187, pp.27-48, 2016.

[2] S. Aditya, Y. Yang, and C. Baral et al., "Image Understanding using Vision and Reasoning through Scene Description Graph," *Computer Vision and Image Understanding*, In Press, Available online 18 December, 2017.

[3] E. Kolve, R. Mottaghi, and D. Gordon et al., "AI2-THOR: An Interactive 3d Environment for Visual AI," *arXiv preprint arXiv:1712.05474*, 2017.

[4] D. Xu, Y. Zhu, and C. B. Choy et al., "Scene Graph Generation by Iterative Message Passing," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.5410-5419, 2017.

[5] Y. Li, W. Ouyang, and B. Zhou et al., "Scene Graph Generation from Objects, Phrases and Region Captions," *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp.1261-1270, 2017.

[6] S. Ren, K. He, and R. Girshick et al., "Faster R-CNN: Towards

Real-Time Object Detection with Region Proposal Networks," *Proceedings of the Neural Information Processing Systems (NIPS)*, pp.91-99, 2015.

[7] C. Lu, R. Krishna, and M. Bernstein et al., "Visual Relationship Detection with Language Priors," *Proceedings of the European Conference on Computer Vision(ECCV)*, pp.852-869, 2016.

[8] B. Dai, Y. Zhang, and D. Lin, "Detecting Visual Relationships with Deep Relational Networks," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.3298-3308. 2017.

[9] P. Gay, J. Stuart, and A. D. Bue, "Visual Graphs from Motion (VGfM): Scene understanding with Object Geometry Reasoning," *arXiv preprint arXiv:1807.05933*, 2018.

[10] S. Song and J. Xiao, "Deep Sliding Shapes for Amodal 3D Object Detection in RGB-D Images," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*, pp.808-816. 2016.

[11] A. Dai, A. X. Chang, and M. Savva et al., "ScanNet: Richly-annotated 3D Reconstructions of Indoor Scenes," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*, pp.5828-5839. 2018.

[12] D. Goron, A. Kembhavi, and M. Rastegari et al., "IQA: Visual Question Answering in Interactive Environments," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*, pp.4089-4098, 2018.

[13] J. Redmon and A. Farhadi, "YOLOv3: An Incremental Improvement," *arXiv preprint arXiv:1804.02767*, 2018.

신 동 협

https://orcid.org/0000-0003-4459-8545
e-mail : ksw9446@kyonggi.ac.kr
2018년 경기대학교 컴퓨터과학과(학사)
2018년~현 재 경기대학교 컴퓨터과학과 석사과정
관심분야 : 인공지능, 컴퓨터비전

김 인 철

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