

Hybrid Learning for Vision-and-Language Navigation Agents

Suntaek Oh[†] · Incheol Kim^{††}

ABSTRACT

The Vision-and-Language Navigation(VLN) task is a complex intelligence problem that requires both visual and language comprehension skills. In this paper, we propose a new learning model for visual-language navigation agents. The model adopts a hybrid learning that combines imitation learning based on demo data and reinforcement learning based on action reward. Therefore, this model can meet both problems of imitation learning that can be biased to the demo data and reinforcement learning with relatively low data efficiency. In addition, the proposed model uses a novel path-based reward function designed to solve the problem of existing goal-based reward functions. In this paper, we demonstrate the high performance of the proposed model through various experiments using both Matterport3D simulation environment and R2R benchmark dataset.

Keywords : Vision-and-Language Navigation, Hybrid Learning, Path-Based Reward Function

시각-언어 이동 에이전트를 위한 복합 학습

오 선 택[†] · 김 인 철^{††}

요 약

시각-언어 이동 문제는 시각 이해와 언어 이해 능력을 함께 요구하는 복합 지능 문제이다. 본 논문에서는 시각-언어 이동 에이전트를 위한 새로운 학습 모델을 제안한다. 이 모델은 데모 데이터에 기초한 모방 학습과 행동 보상에 기초한 강화 학습을 함께 결합한 복합 학습을 채택하고 있다. 따라서 이 모델은 데모 데이터에 편향될 수 있는 모방 학습의 문제와 상대적으로 낮은 데이터 효율성을 갖는 강화 학습의 문제를 상호 보완적으로 해소할 수 있다. 또한, 제안 모델에서는 기존의 목표 기반 보상 함수들의 문제점을 해결하기 위해 설계된 새로운 경로 기반 보상 함수를 이용한다. 본 논문에서는 Matterport3D 시뮬레이션 환경과 R2R 벤치마크 데이터 집합을 이용한 다양한 실험들을 통해, 제안 모델의 높은 성능을 입증하였다.

키워드 : 시각-언어 이동 문제, 복합 학습, 경로 기반 보상 함수

1. 서 론

최근 에이전트의 복합 지능에 관한 관심이 높아지면서 VLN(Vision-and-Language Navigation) 문제[1]가 주목 받고 있다. VLN이란 3차원 실내 공간에 놓인 한 에이전트가 실시간 입력 영상(input image)과 자연어 지시(natural language instruction)에 따라 스스로 이동 행동(navigational action)을 결정함으로써 미지의 목적지까지 도달해야 하는 작업이다. Fig. 1은 VLN 작업의 한 예를 보여준다. Fig. 1의 왼쪽은 에이전트에 주어진 자연어 지시와 이 지시에 따른 에이전트의

행동 시퀀스를 보여주며, 그림의 오른쪽은 에이전트의 위치에 따라 입력되는 순차적인 파노라마 영상(panoramic image)을 보여준다.

VLN 작업에서 중요한 문제 중 하나는 한정된 경험 데이터(seen data)를 이용하여 '미-경험 작업(unknown task)에서 얼마나 좋은 성능을 갖는 에이전트로 학습시키느냐'하는 학습의 일반화(generalization) 및 지식 전이(knowledge transfer) 문제이다. 이러한 VLN 에이전트의 일반화 능력을 향상시키고자 노력한 대표적인 연구들로는 [1-3]가 있다. [1]의 연구에서는 VLN 에이전트를 위한 모방 학습 방법을 제시하였으나, [2,3] 연구에서는 모방 학습과 강화 학습을 결합하는 방법을 제시하였다. 모방 학습은 에이전트의 학습을 가속화할 수 있지만 한정된 데모 데이터로 인해 편향(bias)이 발생한다. 이들은 강화 학습의 경험 데이터로부터 모방 학습의 편향을 줄이고 에이전트의 일반화 능력을 높이고자 하였다. 하지만 두 학습 방법으로부터 얻어낸 손실(loss)들은 규모가 다르므로 학습의

* 이 연구는 2020년도 산업통상자원부 및 산업기술평가관리원(KEIT) 연구비 지원에 의한 연구임(10077538).

** 이 논문은 2020년도 한국정보처리학회 춘계학술발표대회에서 '시각-언어 이동 에이전트를 위한 모방 학습과 강화 학습의 결합'의 제목으로 발표된 논문을 확장한 것임.

† 준 회원 : 경기대학교 컴퓨터과학과 석사과정

†† 종신회원 : 경기대학교 컴퓨터과학과 교수

Manuscript Received : June 29, 2020

Accepted : July 17, 2020

* Corresponding Author : Incheol Kim(kic@kyonggi.ac.kr)

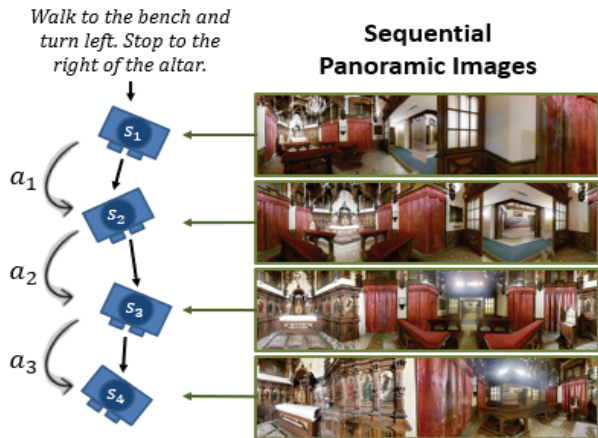


Fig. 1. Example of Vision-Language Navigation (VLN) Task

불균형 문제가 발생할 수 있다. 하지만 [2, 3]에서 제시한 모델들은 이러한 학습의 불균형 문제를 충분히 고려하지 않고 있다.

한편, [2,3]의 VLN 에이전트 모델에서 이용한 목표 기반 밀집 보상 함수(goal-based dense reward function)는 이전 상태보다 현재 상태가 목적지에 더 가까워지면 긍정 보상(positive reward)을 받도록 설계되었다. 이러한 목표-기반 보상 함수는 목적지와의 근접도만을 고려할 뿐, 얼마나 에이전트가 정답 경로(ground-truth path) 혹은 최적 경로(optimal path)를 잘 따라가는 지는 고려하지 못한다. 따라서 이러한 기존의 목표-기반 보상함수는 VLN 에이전트의 전체 이동 경로의 품질을 보장하는 데는 한계가 있다.

이러한 문제점들을 해결하기 위해 본 논문에서는 VLN 에이전트를 위한 새로운 학습 모델을 제시한다. 이 모델은 모방 학습과 강화 학습을 결합한 새로운 학습 방법인 CIR(Combining Imitation learning and Reinforcement learning)과 새로운 보상 함수인 RBA(Region Based Alignment)를 이용한다. 복합 학습(hybrid learning) 방법인 CIR은 강화 학습의 낮은 데이터 효율성 문제와 모방 학습의 데이터 편향성 문제를 상호 보완적으로 해소할 수 있다. 또한, CIR은 두 학습 방법의 손실 규모 차이로 인해 발생하는 학습 불균형 문제를 고려하여, 손실 가중치(loss weight)를 포함하고 있다. 또한, 제안 모델에서는 기존의 목표 기반 보상 함수들의 문제점을 해결하기 위해 설계된 새로운 경로 기반의 보상 함수 RBA를 이용한다. 이 새로운 보상 함수는 일정한 범위 내에서 에이전트가 최적 경로를 유지하고 있는지를 판별하는 경로 기반 보상 함수이다. 이 보상 함수는 VLN 에이전트의 작업 성공률뿐만 아니라, 목적지까지 이동 경로의 품질을 향상시키는데도 큰 도움을 줄 수 있다. 본 논문에서는 Matterport3D 시뮬레이션 환경[4]과 R2R 벤치마크 데이터 집합을 이용한 다양한 실험들을 통해, 제안 모델의 성능을 분석한다. 서론에 이어 2장에서는 본 논문과 연관된 기존의 관련 연구들을 간략히 살펴보고, 3장에서는 본 논문에서 제안하는 새로운 시각-언어 이동 에이전트 모델의 설계에 대해 자세히 소개한다. 이어서

4장에서는 제안 모델의 구현과 성능 분석 실험 결과들을 설명하고, 5장에서는 결론과 향후 연구를 정리한다.

2. 관련 연구

2.1 시각-언어 이동

시각-언어 이동(Vision-and-Language Navigation) 문제는 특정한 3차원 실내공간 안에서 실시간 영상정보를 입력받는 에이전트에게 시작 위치에서 목적 위치로 이동하기 위한 지시를 주면 에이전트가 그 지시와 영상정보를 복합적으로 처리하여 목적지까지 이동해야 하는 문제이다. 이 문제는 [1]의 연구에서 처음 제안되었다. 이 연구에서는 문제 정의뿐만 아니라 학습을 위한 Matterport3D 환경과 VLN 문제를 학습시키기 위한 지시, 경로 쌍으로 이루어진 R2R데이터, 그리고 탐색 작업을 수행하기 위한 기본적인 에이전트 모델도 제시하였다. R2R 데이터는 이동 경로를 따랐을 때 얻는 시각정보를 보고 사람이 직접 지시를 생성하여 만들어졌다. 이러한 번거로운 절차로 인해 지시를 만드는 것이 쉽지 않기 때문인지 제공된 R2R 데이터의 양이 충분하지 못하다는 단점이 있다. 에이전트 모델은 학습 데이터의 지시정보와 환경으로부터 입력받는 실시간 영상정보의 순차적인 특성을 처리하기 위해 순환 신경망(recurrent neural network)의 하나인 LSTM(Long Short Term Memory) 신경망을 사용한다. 이때 영상정보는 초기에 1인칭 시점으로 전방에 대한 시각정보만이 영상으로 주어졌지만, [5]의 연구에서 파노라마 형태의 입력 영상을 제안한 이후 후속 연구들에서는 대부분 파노라마 영상 입력을 기반으로 두고 있다. 본 논문에서도 VLN 문제를 해결하기 위해 R2R 데이터와 LSTM 신경망을 사용하고 파노라마 영상 입력을 기반으로 하고 있다.

2.2 시각-언어 이동을 위한 강화 학습

모방 학습에서 교사 강제 방법(teacher forcing)으로 데이터에 주어진 상태만을 경험하게 한다면 데이터가 충분하지 않을 때 학습 데이터에 대한 편향이 발생한다. 이러한 문제를 해결하려는 방법으로 강화 학습을 적용하여 학습 과정에서 추출한 행동을 수행하여 새로운 상태를 경험하게 만든 연구들[2,3,6,7]이 있다. 이 연구들에서는 데이터를 활용한 모방 학습과 데이터의 지시만을 활용한 강화 학습을 결합[8,9]하여 사용한다. 이때 강화 학습 과정에서 데이터에 없는 새로운 상태를 경험하고 그 상태를 평가하여 보상 값을 얻게 된다. 현재 상태가 이전 상태보다 목적 위치와 가까워졌는가를 판단하여 보상을 주는 [2,3,6] 연구가 있고 [2] 연구에서는 추가로 지시와 생성한 경로의 매칭 정도를 평가하여 보상으로 준다. 또한, 생성한 경로와 데이터에서 주어진 정답 경로와의 유사도를 정규화된 DTW 알고리즘을 통해 계산하여 이전 상태까지 생성된 경로에 비해 현재 상태까지 생성된 경로가 정답 경로와 더 유사해졌는가를 판단하여 보상을 준 [7] 연구가 있다. 이후 강화 학습 알고리즘을 통해 보상을 최대로 얻을 수

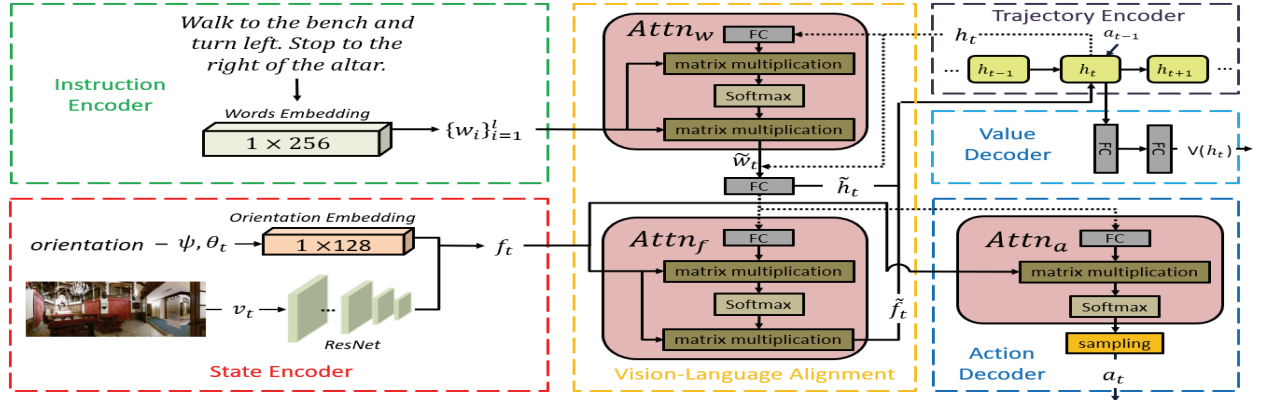


Fig. 2. Architecture of the Proposed Model

있는 방향으로 에이전트 네트워크의 가중치를 업데이트하게 된다. 초기 [6] 연구에서는 REINFORCE 알고리즘을 사용하여 강화 학습 손실을 계산하였고 이후 [2,3] 연구에서는 A2C (Advantage Actor-Critic) 알고리즘[10]을 사용하여 강화 학습 손실을 계산하였다.

3 시각-언어 이동 모델

3.1 문제 정의

시각-언어 이동(VLN) 문제는 3차원 실내 공간에서 실시간 영상을 입력받는 에이전트가 자연어 지시(instruction)를 따라 목적지로 이동하는 작업이다. 지시 $I = \{u_0, u_1, \dots, u_l\}$ 는 l 개의 단어 u_i 들로 이뤄진 문장들로 구성된다. 에이전트는 지시를 따라 n 개의 상태 시퀀스로 이뤄진 정답 경로 $R = \langle s_1, s_2, \dots, s_n \rangle$ 를 찾아야 한다.

본 논문에서는 마르코프 결정 프로세스(Markov Decision Process, MDP)를 기초로 VLN 문제를 강화 학습 문제로 정의한다. 먼저, 상태 $s \in S$ 는 (v, ψ, θ) 로 구성된다. 여기서 v 는 에이전트의 위치에서 포착 가능한 360° 파노라마 영상이다. 파노라마 영상은 가로로 30° 씩 12개, 위아래 30° 씩 3개로 총 36개의 부분 영상으로 이루어져 있다. ψ 와 θ 는 각각 수평(elevation), 수직(heading)으로 이루어진 방향(orientation) 정보를 의미한다. 다음으로 행동 $a \in A$ 는 (m, s) 로 구성된다. 여기서 m 은 이동, s 는 정지를 의미한다. 이동은 최대 36개의 방향으로 이동할 수 있고 위치마다 이동 가능한 방향(navigable directions)이 한정되어 있다. 에이전트는 지시를 잘 따르면서 목적지에 빠르게 도착할수록 높은 보상을 받는다.

3.2 에이전트 모델

본 논문에서 제안하는 VLN 에이전트 모델의 구조는 Fig. 2와 같다. VLN 에이전트는 환경으로부터 현재 위치에서 관측된 파노라마 영상 v_t 와 이동 가능한 방향 정보 ψ_t, θ_t 를 입력받고, 환경 외적으로는 지시(instruction)를 입력받는다. 파노라마 영상과 이동 가능한 방향 정보는 상태 인코더(state encoder)에 의해 하나의 연결된(concatenated) 특징 벡터 f_t 로 변환된다.

특징 벡터 f_t 의 계산식은 아래 Equation (1)과 같다.

$$f_t = [ResNet(v_t), (\cos\theta_t, \sin\theta_t, \cos\psi_t, \sin\psi_t)] \quad (1)$$

지시는 지시 인코딩(instruction encoding)에 의해 단어 임베딩(word embedding) 벡터 $\{w_i\}_{i=1}^l$ 로 변환된다. 여기서 l 은 단어의 수를 의미한다. f_t 와 $\{w_i\}_{i=1}^l$ 는 시각-언어 정렬(vision-language alignment, VLA)에 의해 주의 집중 벡터 \tilde{f}_t 와 \tilde{h}_t 로 계산된다. \tilde{h}_t 는 주의 집중 벡터 \tilde{w}_t 와 h_t 를 연결(concatenation)한 값이다. h_t 는 에이전트가 매 시간 단계(time step)마다 지시의 어느 부분을 따르고 있는지를 표현하는 벡터이다. h_t 는 LSTM(Long Short-Term Memory) 기반의 경로 인코딩(trajectory encoding)을 통해 생성된다. 가치 디코딩(value decoding)은 h_t 로부터 상태 가치 $V(h_t)$ 를 계산한다. 행위 디코딩(action decoding)은 f_t 와 \tilde{h}_t 로부터 행동 a_t 를 계산한다.

본 논문에서는 Fig. 3과 같이 한 번의 학습 반복을 위해 모방 학습 손실을 계산하는 에피소드와 강화 학습 손실을 계산하는 에피소드를 동시에 진행한다. 모방 학습에서는 전문가 에이전트(teacher, t)의 정책에 따라 에피소드를 진행하고 강화 학습에서는 학습자 에이전트(student, s)의 정책에 따라 에피소드를 진행한다.

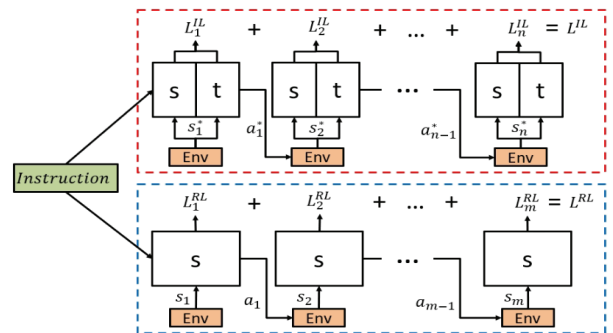


Fig. 3. Imitation Loss L^{IL} and Reinforcement Loss L^{RL}

에피소드가 진행된 후에는 두 에피소드를 통해 얻어낸 손실로부터 학습자 에이전트를 갱신한다. 이에 대한 자세한 내용은 3.3절에서 소개한다.

3.3 학습 방법

본 논문에서는 낮은 데이터 효율성을 갖는 강화 학습의 문제와 데모 데이터에 편향될 수 있는 모방 학습의 문제를 상호 보완하기 위해 두 학습 방법을 결합한 학습 모델 CIR(Combining Imitation learning and Reinforcement learning)을 제안한다. 제안 방법 CIR의 학습 과정을 나타내는 의사 코드(pseudo code)는 Table 1과 같다.

Table 1. CIR Learning Algorithm

Algorithm 1 Learning with CIR	
1	initialize θ_p randomly
2	for i in range(MAX_ITER):
3	for t in range(n): #teacher_forcing
4	$L^L \leftarrow \log \pi_{\theta_p}(h_t, a_t^*)$
5	$s_{t+1} = \text{perform}(s_t, a_t^*)$
6	for t in range(m): #student_forcing
7	$L^{RL} \leftarrow (G_t - V(h_t)) \log \pi_{\theta_p}(h_t, a_t) + \eta H(\pi_{\theta_p}(h_t, a_t))$
8	$s_{t+1} = \text{perform}(s_t, a_t)$
9	$L^{MIX} = \lambda_{LL} L^L + L^{RL}$
10	$\theta_p = \theta_p + \gamma \nabla L^{MIX}$

Table 1에서 1번 줄은 정책 매개변수 θ_p 를 무작위로 초기화한다. 다음으로 2-8번 줄은 모방 학습과 강화 학습을 동시에 진행하는 학습 반복(iterations) 과정을 나타낸다. 3-5번 줄은 모방 학습 손실 L^L 을 계산한다. L^L 은 Equation (2)와 같이 매시간 단계마다 교차 엔트로피 손실(cross entropy loss)을 계산하고 이를 합하여 얻어낸다. L^L 은 정책 네트워크 π_{θ_p} 가 최적 행동 a_t^* 를 결정할 확률을 높이도록 학습을 유도한다.

$$L^L = - \sum_{t=1}^N \log \pi_{\theta_p}(h_t, a_t^*) \quad (2)$$

6-8번 줄은 강화 학습 손실 L^{RL} 을 계산한다. L^{RL} 은 Equation (3)과 같이 A2C(advantage actor-critic) 알고리즘을 기반으로 강화 학습 손실 L^{RL} 을 계산한다. Equation (3)에서 $G_t - V(h_t)$ 는 우세 함수(advantage function)이다. $\eta H(\pi_{\theta_p}(h_t, a_t))$ 는 다양한 행동을 결정할 수 있도록 장려하는 엔트로피 함수이다.

$$L^{RL} = - \sum_{t=1}^M ((G_t - V(h_t)) \log \pi_{\theta_p}(h_t, a_t) + \eta H(\pi_{\theta_p}(h_t, a_t))) \quad (3)$$

9번 줄은 L^L 과 L^{RL} 을 더하여 혼합 손실 L^{MIX} 를 계산한다. 한편, L^{RL} 보다 L^L 의 값이 훨씬 크기 때문에 학습의 불균

형이 발생한다. 이를 위해 CIR은 λ_{LL} 을 통해서 L^L 과 L^{RL} 의 균형을 조절한다. 마지막 10번 줄은 L^{MIX} 를 토대로 θ_p 를 갱신한다.

제안 방법 CIR은 낮은 데이터 효율성을 갖는 강화 학습과 데모 데이터에 편향될 수 있는 모방 학습의 문제를 상호 보완할 수 있다. 또한, CIR은 손실 가중치 λ_{LL} 를 통해 두 학습 방법의 불균형 문제를 해결하였다.

3.4 보상 함수

기존 연구들[2,3]에서는 매시간 단계마다 이동한 에이전트의 위치가 목적 위치에 가까워지면 양의 보상(+1)을 받고 그렇지 않으면 음의 보상(-1)을 받는다. 그리고 에이전트가 생성한 경로의 마지막 위치에서 목적 위치와의 거리가 3m 이내이면 목적 위치에 도달했다는 의미로 양의 보상(+2)을 받고 그렇지 않으면 음의 보상(-2)을 받도록 하였다. 이 보상 함수는 매시간 에이전트의 위치와 목적 위치와의 거리만을 고려하기 때문에 에이전트가 목적 위치에 도달하도록 학습을 유도할 수 있지만, 지시를 잘 따라 최적 경로를 지나도록 유도할 수 없는 문제가 있다. 이러한 문제를 해결하고자 본 논문에서는 새로운 보상 함수 RBA(Region Based Alignment)를 제안한다. 제안 보상 함수 RBA는 정답 경로를 기준으로 특정 거리 ϵm 내에서 목적 위치에 가까워지면 양의 보상(+1)을 받고 그렇지 않으면 음의 보상(-1)을 부여한다. 이를 수식으로 표현하면 아래 Equation (4)와 같다.

$$r(p_t) = \begin{cases} 2 & \text{if } p_{t-1} = p_t \text{ and } D(p_t) \leq 3 \\ -2 & \text{if } p_{t-1} = p_t \text{ and } D(p_t) > 3 \\ 1 & \text{if } f(p_t) \text{ and } g(p_t) \\ -1 & \text{otherwise} \end{cases} \quad (4)$$

Equation (4)에서 첫 번째 조건식과 두 번째 조건식은 에이전트가 정지 행동을 수행하여 위치 변화가 없을 때 다익스트라 알고리즘(Dijkstra algorithm)을 이용하여 목적 위치와의 거리가 3m 이내인지 판단하는 식이다. 세 번째 조건식에서 $f(p_t)$ 는 현재 위치 p_t 가 정답 경로에서 특정 거리 ϵm 이내에 있으면 참(true)을, 그렇지 않으면 거짓(false)을 반환하는 함수로서 아래 Equation (5)와 같다.

$$f(p_t) = \begin{cases} true & \text{if } \exists p_i \in P \Rightarrow \overline{p_i p_t} \leq \epsilon \text{ or} \\ & \exists h_i \in p_i p_{i+1} \Rightarrow p_i p_{i+1} \perp \overline{h_i p_t}, \overline{h_i p_t} \leq \epsilon \\ false & \text{otherwise} \end{cases} \quad (5)$$

Equation (5)에서 P 는 정답 경로상의 모든 노드의 집합, p_i 는 정답 경로상의 i 번째 노드, p_t 는 에이전트의 위치, h_i 는 p_i 에서 선분 $\overline{p_i p_{i+1}}$ 에 내린 수선의 발을 의미한다. 따라서 $f(p_t)$ 는 p_t 와의 거리가 ϵm 이내인 p_i 또는 h_i 가 존재하면 참을 반환한다. 예를 들어 Fig. 4의 (a)에서 p_t^1 는 p_t 와의 거리가 ϵm 이내이기 때문에 참이다. 나머지 p_t^2, p_t^3, p_t^4 는 정답 경로의 모든 노드와의 거리가 ϵm 이내가 아니다. 하지만 Fig. 4

의 (b)에서와 같이 p_t^2 에서 정답 경로상에 내린 수선의 발 h_t^2 가 존재하고 $\overline{p_t^2 h_t^2}$ 의 길이가 ϵ_m 이내이기 때문에 참이다. 한편, p_t^3 는 정답 경로상에 내린 수선 발 h_t^3 이 존재하지만 $\overline{p_t^3 h_t^3}$ 의 길이가 ϵ_m 보다 크기 때문에 결국 거짓이고 p_t^4 는 정답 경로 상에 내릴 수 있는 수선의 발이 존재하지 않기 때문에 결국 거짓이다.

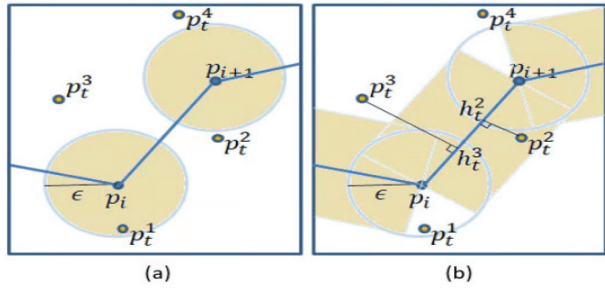


Fig. 4. Reward Condition $f(p_t)$

Equation (4)에서 $g(p_t)$ 는 기존 연구들[2,3]에서 사용한 목표 기반 보상 함수 조건이다. $g(p_t)$ 는 아래 Equation (6)과 같이 에이전트의 이동 위치가 이전 위치보다 목적지에 더 가까워지면 참, 그렇지 않으면 거짓을 반환한다.

$$g(p_t) = \begin{cases} true & \text{if } D(p_{t-1}) - D(p_t) > 0 \\ false & \text{otherwise} \end{cases} \quad (6)$$

이러한 제안 보상 함수 RBA는 에이전트가 목적지와 가까워지도록 이동할 뿐만 아니라, 정답 경로를 벗어나지 않게 이동할 수 있도록 하는 장점이 있다. 또한, RBA는 하나의 지시에 하나의 정답 경로만 제시하는 기존 연구들[1-3]과는 달리, 하나의 지시에 여러 정답 경로를 제시해주는 정답 영역을 사용한다. 따라서 정답 경로를 증가시켜 에이전트의 일반화 성능을 높여주는 부수 효과가 있다.

4. 구현 및 실험

4.1 데이터 집합과 모델 학습

본 논문에서는 R2R 데이터 집합을 이용하여 제안 모델의 성능을 분석하기 위한 실험을 수행한다. 이를 위해 제안 모델은 Python 3.7, Pytorch 1.2.0 라이브러리를 이용하여 구현하였다. 한편, 모델 학습과 실험에 사용된 R2R 데이터 집합은 Matterport3D 가상 환경의 시작 위치에서 목적 위치로 가는 최단 경로와 이를 설명하는 세 가지의 자연어 지시들의 집합으로 구성되어 있다. R2R 데이터 집합에서 학습 데이터(seen training data)는 14,025개, 학습 검증 데이터(seen validation data)는 1,020개, 비-학습 검증 데이터(unseen validation data)는 2,349개, 비-학습 테스트 데이터(unseen test data)는 2,349개의 지시로 각각 구성된다. 입력 영상으

로부터 시각 특징 추출을 위해서는 미리 학습된 ResNet-152 모델을 이용하였다. 모델 학습을 위해 엔트로피 함수의 반영 비율 η 는 0.01로, 모방 학습과 강화 학습의 손실의 균형을 맞추기 위한 손실 가중치 λ_{ll} 는 0.05로, 학습률(learning rate) γ 는 0.0001로 각각 설정하였다.

4.2 성능 분석 실험

본 논문에서는 제안 모델에서 채택한 CIR 학습 방법과 RBA 보상 함수의 효과를 분석하고, 기존 모델들과의 비교를 통해 제안 모델의 우수성을 입증하기 위한 실험을 수행하였다. 실험에 사용된 성능 평가 척도는 SR(Success Rate)와 SPL(Success rate weighted by Path Length)이다. SR은 VLN 에이전트의 작업 성공률을 나타낸다. VLN 작업은 에이전트의 마지막 위치가 목적지와 거리가 3m 이내일 때 성공으로 간주한다. 반면, SPL은 정답 경로 길이를 에이전트가 실제 이동한 경로 길이로 나눈 값이다. 따라서 VLN 에이전트가 실제 이동한 경로가 짧을수록 높은 SPL 점수를 받을 수 있다.

첫 번째 실험은 제안 모델에서 채택한 보상 함수의 효과를 분석하기 위한 비교 실험이다. 이 실험에서는 목적지까지의 거리 변화만을 고려한 보상 함수 DBA(Destination Based Alignment)[3], 에이전트가 진행해온 경로와 정답 경로와의 유사도 변화를 DTW(Dynamic Time Warping) 알고리즘[11]으로 계산하는 보상 함수 SBA(Similarity Based Alignment)[7], 그리고 본 논문에서 제안한 보상 함수 RBA 등 3가지 보상 함수에 따른 VLN 작업 성능을 서로 비교하였다. SBA는 시계열 데이터 간의 유사도를 계산하는 DTW 알고리즘을 사용하여 매 순간 현재까지 생성한 경로와 정답 경로 간의 유사도를 계산하게 된다. 현재 t 순간에 계산된 정답 경로와의 유사도가 이전 순간 t-1에 계산된 유사도보다 높으면 양의 보상(+1)을 받고, 그렇지 않으면 음의 보상(-1)을 받도록 설계된 보상 함수이다. RBA의 임계 거리 ϵ 는 1m로 설정하였다. 이 실험을 위해 매시간 단계마다 에이전트에게 즉각적인 보상이 부여되는 밀집 보상(dense reward) 방식과 순수 강화 학습만을 이용해 학습하였고 학습 반복 횟수는 8만 번으로 설정하였다.

Table 2. Performance Comparison with Different Reward Functions

Reward Function	Seen		Unseen	
	SR	SPL	SR	SPL
DBA[3]	0.273	0.041	0.225	0.031
SBA[7]	0.409	0.381	0.405	0.382
RBA	0.436	0.414	0.399	0.375

이 실험의 결과는 Table 2와 같다. 본 논문에서 제안한 RBA와 [7]에서 제안한 SBA가 각각 학습 데이터(seen)와 비-학습 데이터(unseen)에서 높은 성능을 보였고, DBA는 좋지 못한 성능을 보였다. DBA는 에이전트의 위치와 목적 위치와의 차이만을 고려하였기 때문에, 지시를 따르지 않는 잘못된

경로를 학습하게 되는 문제점이 있다. SBA와 RBA는 보상 함수의 설계는 다르지만, 정답 경로와 유사한 경로를 학습하려는 같은 목적을 갖는 보상 함수이다. 따라서 하이퍼 파라미터 (hyper parameter)에 따른 약간의 차이가 있지만, 대부분 비슷한 성능을 내는 것을 확인할 수 있었다. 하지만, SBA는 에이전트가 지나온 이전 경로의 길이가 길수록 계산량이 커지는 문제가 존재한다. 반면, RBA는 비교적 적은 계산량으로도 에이전트가 최적 경로를 따라 목적지에 가까워지는 방향으로 이동할 수 있도록 한다는 장점이 있다.

두 번째 실험은 제안 모델에서 채택한 모방 학습과 강화 학습을 결합한 복합 학습(CIR)의 효과를 분석하기 위한 실험이다. 이 실험을 위해 순수 강화 학습(only RL), 순수 모방 학습(only IL), 복합 학습 방법(CIR)을 각각 채용했을 때의 VLN 작업 성능을 서로 비교하였다. 이 실험에서 보상 함수는 RBA를 이용하였으며, 하이퍼 파라미터 ϵ 는 1.5m로, 학습 반복 횟수는 20만 번으로 각각 설정하였다.

Table 3. Performance Comparison with Different Learning Strategies

Learning Strategy	Seen		Unseen	
	SR	SPL	SR	SPL
only RL	0.420	0.388	0.385	0.338
only IL	0.549	0.527	0.433	0.406
CIR ($\lambda_{IL} = 0.1$)	0.628	0.597	0.487	0.456
CIR ($\lambda_{IL} = 0.05$)	0.653	0.622	0.488	0.447

이 실험의 결과는 Table 3과 같다. 이 실험 결과에서 본 논문에서 제안한 복합 학습(CIR)이 순수 모방 학습(IL)이나 순수 강화 학습(RL)보다 SR 및 SPL 척도 면에서 모두 성능이 높게 나타났다는 것을 알 수 있다. 순수 모방 학습은 양질의 데모 데이터를 활용함으로써, 데이터 효율성이 상대적으로 낮은 강화 학습에 비해 높은 성능을 보였다. 하지만 한정된 데모 데이터에 편향되어, 복합 학습 방법보다는 낮은 성능을 보인 것으로 추정된다. 반면, 본 논문에서 제안한 복합 학습 방법(CIR)은 강화 학습(RL)의 데이터 비효율성 문제와 모방 학습(IL)의 데모 데이터에 대한 편향성 문제를 어느 정도 해소함으로써, 이 실험에서 상대적으로 가장 높은 성능을 보인 것으로 판단한다. 그리고 다양한 모방 손실 가중치(λ_{IL})에 따라 복합 학습(CIR)의 성능은 조금씩 변화가 있었고, 본 실험에서는 모방 손실 가중치를 0.05로 설정하였을 때($\lambda_{IL} = 0.05$) 가장 성능이 높았다.

마지막 실험은 기존의 VLN 모델들에 비해 본 논문에서 제안한 모델의 우수성을 입증하기 위한 실험이다. 이 실험에서는 발화자 모델을 이용해 새로운 지시를 생성한 Speaker-Follower[5], 발화자 모델을 이용해 에이전트가 경로를 잘 따랐는지 판별한 RCM[2], 학습 데이터 증강을 위한 환경 드롭아웃(dropout) 기능과 혼합 손실 함수(mixed loss function)를 채용한 Env-Dropout[3], 새로운 보상 함수와 학습 방법을 도입한 제안 모델(CIR)의 VLN 작업 성능을 서로 비교하였다.

Table 4. Performance Comparison with Other Models

Model	Seen		Unseen	
	SR	SPL	SR	SPL
Speaker-Follower[5]	0.52	0.43	0.36	0.29
RCM(no SIL)[2]	0.55	0.48	0.41	0.33
Env-Dropout(base)[3]	0.61	0.57	0.47	0.43
CIR ($\lambda = 0.05$)	0.65	0.62	0.49	0.45

이 실험의 결과는 Table 4와 같다. 비교 모델들 중에서 본 논문의 제안 모델 CIR이 모든 척도에서 가장 높은 성능을 보였다. 특히 제안 모델 CIR은 미-경험 환경(unseen env)에 비해 이미 경험한 환경(seen env)에서 작업 성능의 향상이 더욱 뚜렷했다. 이것은 기존의 VLN 모델들에 비해 제안 모델의 우수성을 확인시켜주는 실험 결과로 볼 수 있다.

본 논문에서는 위에서 설명한 정량적 실험들 외에, 본 논문에서 제안한 모델의 성능을 확인하기 위해 제안 모델에 의해 수행된 VLN 대표 작업들을 정성적으로 분석해보았다. Fig. 5 ~ Fig. 7은 대표적인 3개의 VLN 작업들을 나타내며, 각 그림의 top-down view에서 노란색 별은 시작 위치를, 빨간색 별은 도착 위치를 각각 가리킨다. 또한, 빨간색 화살표는 정답 경로를 나타내며, 파란색 화살표는 에이전트가 실제 움직인 경로를 나타낸다. Fig. 5의 작업은 에이전트가 자연어 지시에 따라 정답 경로와 거의 유사한 작업 경로를 통해 성공적으로 목적지에 도착하였다. 이러한 결과는 제안 모델에서 채택하고 있는 최적 경로 기반의 보상 함수가 에이전트의 이동을 효과적으로 잘 유도해주었기 때문인 것으로 판단한다. Fig. 6의 작업 경우에는 에이전트가 정답 경로에서 살짝 벗어나는 부분도 있지만, 이내 올바른 경로를 탐색하여 목적지까지 성공적으로 도달한 것을 알 수 있다. 이러한 결과는 복합 학습 기반의 제안 모델이 정답 경로에만 의존하는 모방 학습 대신 에이전트 스스로 새로운 경로를 찾을 수 있도록 모험적 탐색(exploration)을 시도하는 강화 학습도 함께 수행하기 때문인 것으로 판단된다. 하지만 Fig. 7의 작업 경우, 작업 시작 부분에서 잘못된 판단으로 인해 정답 경로에서 크게 벗어난 이후, 계속해서 잘못된 판단을 이어간 사례이다. 이와 같은 작업 예들을 통해, 제안 모델에서 채용하고 있는 복합 학습과 보상 함수의 긍정적 효과를 확인할 수 있었으나, 이동 초기에 정답 경로를 크게 벗어나지 않도록 제안 모델의 개선이 필요하다는 사실도 함께 확인할 수 있었다.

5. 결 론

본 논문에서는 시각-언어 이동 에이전트를 위한 새로운 학습 모델을 제안하였다. 이 모델은 모방 학습과 강화 학습을 함께 결합한 복합 학습 CIR을 채택하고 있다. 또한, 제안 모델은 기존의 목적지 기반 보상 함수의 결함을 개선하기 위한 새로운 경로 기반 보상 함수 RBA를 포함하고 있다. 본 논문

에서는 R2R 데이터 집합과 Matterport3D 시뮬레이션 환경을 이용한 다양한 실험을 통해, 제안 모델의 우수한 성능을 확인할 수 있었다. 향후에는 탐색 기법을 통해 잘못된 판단에 대처하는 방법에 관한 연구를 진행할 계획이다.

References

[1] P. Anderson, Q. Wu, D. Teney, J. Bruce, M. Johnson, N. Sunderhauf, I. Reid, S. Gould, and A. V. D. Hengel, "Vision-and-Language Navigation: Interpreting Visually-Grounded Navigation Instructions in Real Environments," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[2] X. Wang, Q. Huang, A. Celikyilmaz, J. Gao, D. Shen, Y. F. Wang, W. Y. Wang, and L. Zhang, "Reinforced Cross-Modal Matching and Self-Supervised Imitation Learning for Vision-Language Navigation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[3] H. Tan, L. Yu and M. Bansal, "Learning to Navigate Unseen Environments: Back Translation with Environmental Dropout," in *Proceedings of North American Chapter of the Association for Computational Linguistics (NAACL)*, 2019.

[4] A. Chang, A. Dai, T. Funkhouser, M. Halber, M. Niessner, M. Savva, S. Song, A. Zeng, and Y. Zhang, "Matterport3D: Learning from RGB-D Data in Indoor Environments," in *Proceedings of the International Conference on 3D Vision*, 2017.

[5] D. Fried, R. Hu, V. Cirik, A. Rohrbach, J. Andreas, L. P. Morency, T. Berg-Kirkpatrick, K. Saenko, D. Klein and T. Darrell, "Speaker-Follower Models for Vision-and-Language Navigation," in *Proceedings of the Neural Information Processing Systems (NIPS)*, Vol.28, 2018.

[6] W. Xiong, X. Wang, H. Wang, and W. Y. Wang, "Look Before You Leap: Bridging Model-Free and Model-Based Reinforcement Learning for Planned-Ahead Vision-and-Language Navigation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp.696-711, 2018.

[7] G. Ilharco, V. Jain, A. Ku, E. Ie, and J. Baldridge, "General Evaluation for Instruction Conditioned Navigation using Dynamic Time Warping," in *Proceedings of Neural Information Processing Systems (NeurIPS)*, 2019.

[8] M. A. Ranzato, S. Chopra, M. Auli, and W. Zaremba, "Sequence level training with recurrent neural networks." in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2015.

[9] R. Paulus, C. Xiong and R. Socher, "A Deep Reinforced Model for Abstractive Summarization," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2018.

[10] V. Mnih, A. P. Badia, M. Mirza, A. Graves, T. P. Lillicrap, T. Harley, D. Silver, and K. Kavukcuoglu, "Asynchronous Methods for Deep Reinforcement Learning," in *Proceedings of the International Conference on Machine Learning (ICML)*, pp.1928-1937, 2018.

[11] D. J. Berndt and J. Clifford, "Using Dynamic Time Warping to Find Patterns in Time Series," in *KDD Workshop*, pp.359-370, 1994.



오 선택

<https://orcid.org/0000-0001-6911-0679>
 e-mail : choice37@kyonggi.ac.kr
 2019년 경기대학교 컴퓨터과학과(학사)
 2020년 ~ 현 재 경기대학교 컴퓨터과학과 석사과정
 관심분야 : 인공지능, 컴퓨터비전, 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

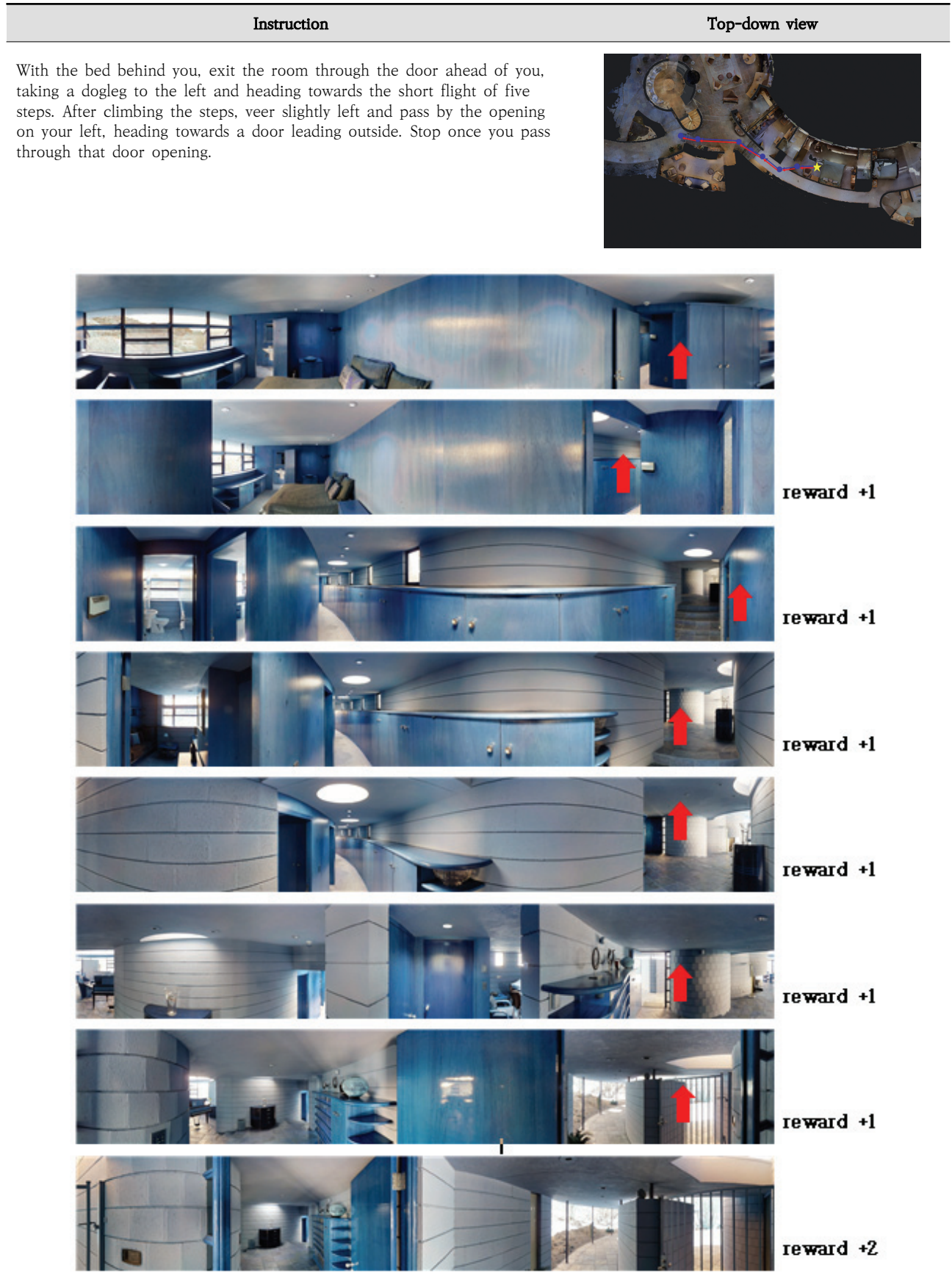


Fig. 5. The First Navigation Example Guided with the Proposed Model

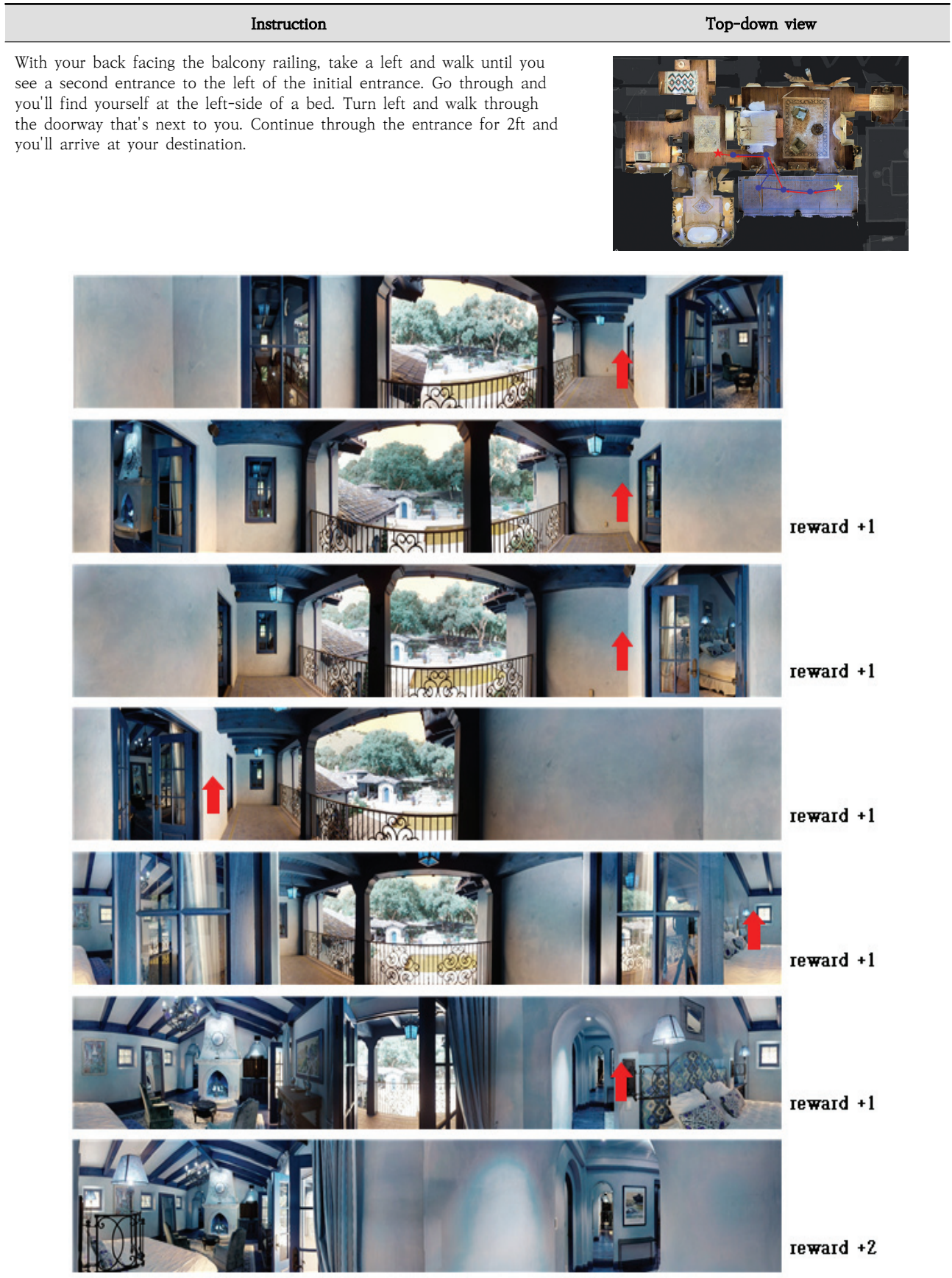


Fig. 6. The Second Navigation Example Guided with the Proposed Model

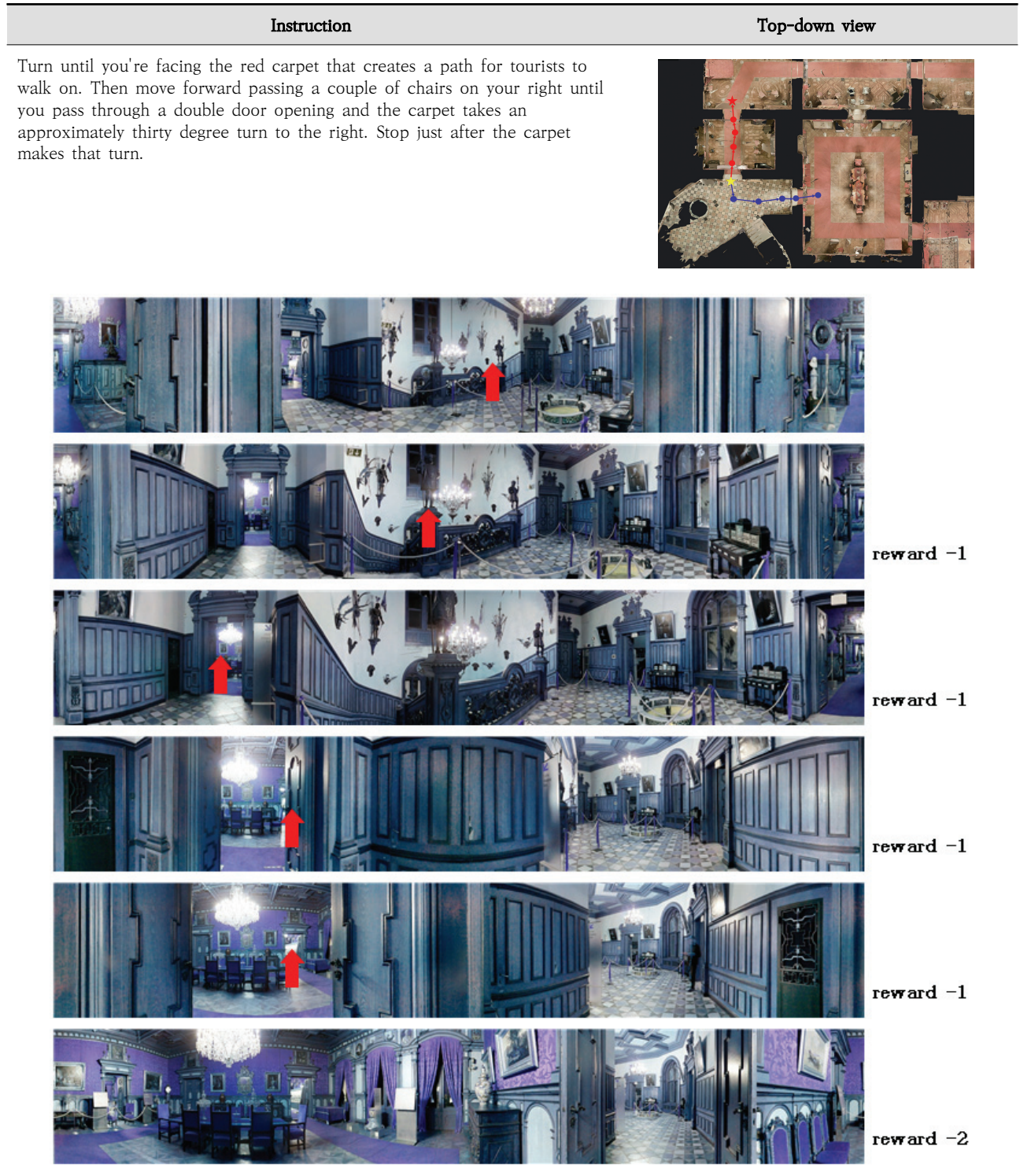


Fig. 7. The Third Navigation Example Guided with the Proposed Model