시각-언어 이동 에이전트를 위한 모방 학습과 강화 학습의 결합

오선택, 김인철
경기대학교 컴퓨터과학과
email:choice37@kyonggi.ac.kr, kic@kyonggi.ac.kr

Combining Imitation Learning and Reinforcement Learning for Visual-Language Navigation Agents

Suntaek Oh, Incheol Kim
Department of Computer Science, Kyonggi University

요 약

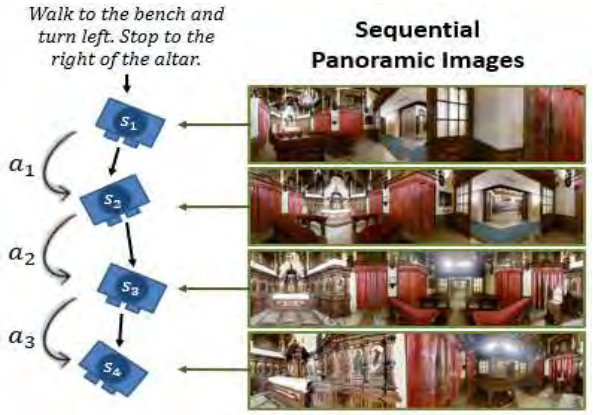
시각-언어 이동 문제는 시각 이해와 언어 이해 능력을 함께 요구하는 복합 지능 문제이다. 본 논문에서는 시각-언어 이동 에이전트를 위한 새로운 학습 모델을 제안한다. 이 모델은 데모 데이터에 기초한 모방 학습과 행동 보상에 기초한 강화 학습을 함께 결합한 복합 학습을 채택하고 있다. 따라서 이 모델은 데모 데이터에 편향될 수 있는 모방 학습의 문제와 상대적으로 낮은 데이터 효율성을 갖는 강화 학습의 문제를 상호 보완적으로 해소할 수 있다. 또한, 제안 모델은 서로 다른 두 학습 간에 발생 가능한 학습 불균형도 고려하여 손실 정규화를 포함하고 있다. 또, 제안 모델에서는 기존 연구들에서 사용되어온 목적지 기반 보상 함수의 문제점을 발견하고, 이를 해결하기 위해 설계된 새로운 최적 경로 기반 보상 함수를 이용한다. 본 논문에서는 Matterport3D 시뮬레이션 환경과 R2R 벤치마크 데이터 집합을 이용한 다양한 실험들을 통해, 제안 모델의 높은 성능을 입증하였다.

1. 서론

최근 에이전트의 복합 지능에 관한 관심이 높아지면서 VLN(Vision-and-Language Navigation) 문제[1]가 주목받고 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. (그림 1)은 VLN 작업의 한 예를 보여준다. (그림 1)의 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 학습 데이터(seen data)를 이용하여 '비-학습 작업(unseen 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 [1-5]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 불균형 문제가 발생할 수 있다. [2-3]에서 제시한 모델들은 학습의 불균형 문제를 고려하지 않고 있다.

한편, [3-5]의 연구들에서는 VLN 에이전트의 일반화 능력 향상을 위해 학습 데이터 증강(data argumentation) 기법을 제시하였다. [3]의 연구에서는 영상에서 특정 물체들에 대한 마스크(mask)를 찾고 이 마스크만큼의 영역을 영상에서 소실(drop out)시킴으로써 새로운 영상정보를 만들었다. 또한, [3]의 연구에서는 새로운 영상정보를 토대로 정답 경로를 만들고 이에 대한 지시를 자동으로 생성하는 발화자(speaker) 모델[4]을 채용하였다. [5]의 연구에서는 정답 경로들 중 시작 위치와 목적 위치가 일치하는 정답 경로들을 연결하여 기존의 학습 데이터 집합 R2R보다 길이가 더 긴 학습 데이터 집합 R4R을 만들었다. 이 연구들은 학습 데이터에서 지시와 영상정보를 증강하고자 하였다. 하지만 이 연구들은 하나의 지시에 대해 여러 정답 경로가 존재할 수 있음에도 불구하고 하나의 정답 경로만 제시한다는 문제점이 있다.

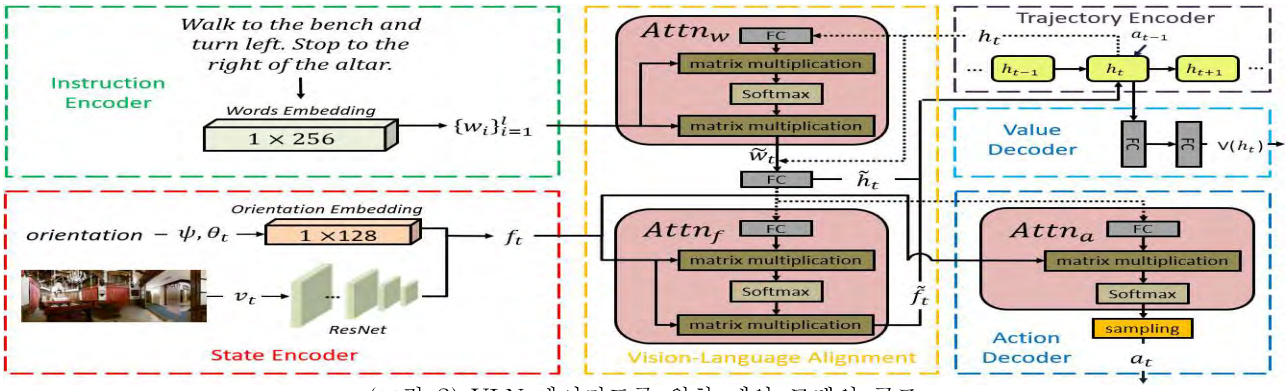


(그림 1) 시각-언어 이동(VLN) 문제의 한 예

한편, [2-3] 연구에서 강화 학습을 위해 사용한 밀집 보상 함수(dense reward function)는 정답 경로와는 상관없이 현재 상태가 이전 상태보다 목적지에 가까우면 무조건 보상을 받기 때문에, 에이전트가 목표에 도달했어도 이동한 경로가 지시를 잘 따랐는지 판별할 수 없다는 결함이 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. CIR은 낮은 데이터 효율성을 갖는 강화 학습의 문제와 데모 데이터에 편향될 수 있는 모

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



(그림 2) VLN 에이전트를 위한 제안 모델의 구조

방 학습의 문제를 상호 보완적으로 해소할 수 있다. 또한, CIR은 두 학습 방법의 손실 규모 차이로 인해 발생하는 학습 불균형 문제를 고려하여 손실 정규화를 포함한다. 한편, 목적지 기반 보상 함수의 문제점을 해결하기 위해 새로 설계된 RBA 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 환경과 R2R 벤치마크 데이터 집합을 이용한 다양한 실험들을 통해, 제안 모델의 성능을 분석한다.

2 시각과 언어기반의 이동

2.1 문제 정의

VLN 문제는 3차원 실내 공간에서 실시간 영상을 입력받는 에이전트가 자연어 지시(instruction)를 따라 목적지로 이동하는 작업이다. 지시 $I=\{u_0,u_1,..,u_l\}$는 l개의 단어 u_i들로 이뤄진 문장들로 구성된다. 에이전트는 지시를 따라 n개의 상태 시퀀스로 이뤄진 정답 경로 $R = <s_1, s_2, ..., s_{n-1}, s_n>$를 찾아야 한다.

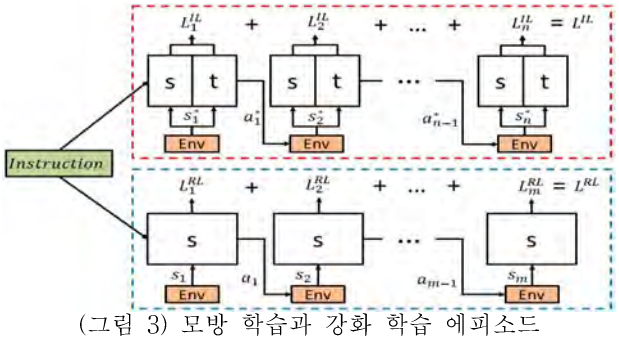
본 논문에서는 마르코프 결정 프로세스(Markov Decision Process, MDP)를 기초로 VLN 문제를 강화 학습 문제로 정의한다. 먼저, 상태 $s \in S$는 (v, ψ, θ)로 구성된다. 여기서 v는 에이전트의 위치에서 포착 가능한 360° 파노라마 영상이다. 파노라마 영상은 가로로 30°씩 12개, 위아래 30°씩 3개로 총 36개의 부분 영상으로 이루어져 있다. ψ와 θ는 각각 수평(elevation), 수직(heading)으로 이루어진 방향(orientation) 정보를 의미한다. 다음으로 행동 $a \in A$는 (m, s)로 구성된다. 여기서 m은 이동, s는 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.

2.2 VLN 에이전트 모델

본 논문에서는 기초 모델(baseline)[3]에서 제안한 인코더-디코더(encoder-decoder) 기반의 VLN 에이전트 모델을 채용한다. 이 VLN 에이전트 모델의 구조도는 (그림 2)와 같다. VLN 에이전트는 환경으로부터 현재 위치에서 관측된 파노라마 영상 v_t와 이동 가능한 방향 정보 ψ_t, θ_t를 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_t로 변환된다. 특징 벡터 f_t의 계산 식은 아래와 같다.

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

지시는 지시 인코딩(instruction encoding)에 의해 단어 임베딩(word embedding) 벡터 $\{w_i\}_{i=1}^{l}$로 변환된다. 여기서 l은 단어의 수를 의미한다. f_t와 $\{w_i\}_{i=1}^{l}$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_t를 연결(concatenation)한 값이다. h_t는 에이전트가 매시간 단계(time step)마다 지시의 어느 부분을 따르고 있는지를 표현하는 벡터이다. h_t는 LSTM(Long Short-Term Memory) 기반의 경로 인코딩(trajectory encoding)을 통해 생성된다. 가치 디코딩(value decoding)은 h_t로부터 상태 가치 $V(h_t)$를 계산한다. 행위 디코딩(action decoding)은 f_t와 $\tilde{h}_t$로부터 행동 a_t를 계산한다.



(그림 3) 모방 학습과 강화 학습 에피소드

본 논문에서는 (그림 3)과 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. 에피소드가 진행된 후에는 두 에피소드를 통해 얻어낸 손실로부터 학습자 에이전트를 갱신한다. 이에 대한 자세한 내용은 2.3절에서 소개한다.

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

<표 1> CIR 학습 알고리즘

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^IL -= logπ_θp(h_t, a_t*)
5          s_t+1 = perform(s_t, a_t*)
6      for t in range(m): #student_forcing
7          L^RL -= (G_t − V(h_t))logπ_θp(h_t, a_t) + ηH(π_θp(h_t, a_t))
8          s_t+1 = perform(s_t, a_t)
9      L^MIX = λ_IL L^IL + L^RL
10     θ_p = θ_p + γ∇L^MIX
```

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

6-8번 줄은 강화 학습 손실 L^{RL}을 계산한다. L^{RL}은 (식 3)과 같이 A2C(advantage actor-critic) 알고리즘을 기반으로 강화 학습 손실 L^{RL}을 계산한다. (식 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.

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

9번 줄은 L^{IL}과 L^{RL}을 더하여 혼합 손실 L^{MIX}를 계산한다. 한편, L^{RL}보다 L^{IL}의 값이 훨씬 크기 때문에 학습의 불균형이 발생한다. 이를 위해 CIR은 λ_{IL}을 통해서 L^{IL}을 정규화를 한다. 마지막 10번 줄은 L^{MIX}를 토대로 θ_p를 갱신한다.

제안 방법 CIR은 낮은 데이터 효율성을 갖는 강화 학습과 데모 데이터에 편향될 수 있는 모방 학습의 문제를 상호 보완할 수 있다. 또한, CIR은 L^{IL} 정규화를 통해 두 학습 방법의 불균형 문제를 해결하였다.

2.4 보상 함수

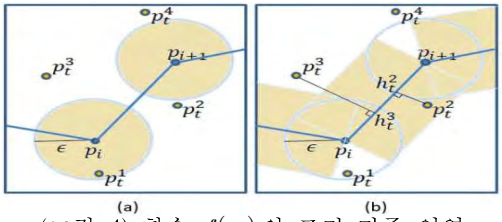
기존 연구들[2-3]에서는 매시간 단계마다 이동한 에이전트의 위치가 목적 위치에 가까워지면 양의 보상(+1)을 받고 그렇지 않으면 음의 보상(-1)을 받는다. 그리고 에이전트가 생성한 경로의 마지막 위치에서 목적 위치와의 거리가 3m 이내이면 목적 위치에 도달했다는 의미로 양의 보상(+2)을 받고 그렇지 않으면 음의 보상(-2)을 받도록 하였다. 이 보상 함수는 매시간 에이전트의 위치와 목적 위치와의 거리만을 고려하기 때문에 에이전트가 목적 위치에 도달하도록 학습을 유도할 수 있지만, 지시를 잘 따라 최적 경로를 지나도록 유도할 수 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 아래와 같다.

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

(식 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_t가 정답 경로에서 특정 거리 ϵm 이내에 있으면 참(true)을, 그렇지 않으면 거짓(false)을 반환하는 함수로서 아래 (식 5)와 같다.

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

(식 5)에서 P는 정답 경로상의 모든 노드의 집합, p_i는 정답 경로상의 i번째 노드, p_t는 에이전트의 위치, h_i는 p_t에서 선분 $\overline{p_i p_{i+1}}$에 내린 수선의 발을 의미한다. 따라서, $f(p_t)$는 p_t와의 거리가 ϵm 이내인 p_i 또는 h_i가 존재하면 참을 반환한다. 예를 들어 (그림 4)의 (a)에서 p_t^1는 p_i와의 거리가 ϵm 이내이기 때문에 참이다. 나머지 p_t^2, p_t^3, p_t^4는 정답 경로의 모든 노드와의 거리가 ϵm 이내가 아니다. 하지만 (그림 4)의 (b)에서와 같이 p_t^2에서 정답 경로상에 내린 수선의 발 h_t^2가 존재하고 $\overline{p_t^2 h_t^2}$의 길이가 ϵm 이내이기 때문에 참이다. 한편, p_t^3는 정답 경로상에 내린 수선 발 h_t^3이 존재하지만 $\overline{p_t^3 h_t^3}$의 길이가 ϵm보다 크기 때문에 결국 거짓이고 p_t^4는 정답 경로상에 내릴 수 있는 수선의 발이 존재하지 않기 때문에 결국 거짓이다.



(그림 4) 함수 $f(p_t)$의 조건 만족 영역

(식 4)에서 $g(p_t)$는 기존 연구[2-3]에서 사용하는 보상 함수이다. $g(p_t)$는 아래 (식 6)과 같이 에이전트의 이동 위치가 이전 위치보다 목적지에 더 가까워지면 참, 그렇지 않으면 거짓을 반환한다.

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

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[1-5]과는 달리, 하나의 지시에 여러 정답 경로를 제시해주는 정답 영역을 사용한다. 따라서 정답 경로를 증강시켜 에이전트의 일반화 성능을 높여주는 부수 효과가 있다. 이 효과는 [3-5]의 데이터 증강(data argumentation) 기법의 원리와 유사하다.

3. 구현 및 실험

3.1 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[1] 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 영상으로부터 시각 특징 추출을 위해서는 미리 학습된 ResNet-152 모델을 이용하였다. 모델 학습을 위해 엔트로피 함수의 반영 비율 η는 0.01로, 모방 학습과 강화 학습의 손실을 정규화하기 위한 λ_{IL}는 0.05로, 학습률(learning rate) γ는 0.0001로 각각 설정하였다.

3.2 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 SC(Success rate)와 SPL(Success rate weighted by Path Length)이

다. SC는 VLN 에이전트의 작업 성공률을 나타낸다. VLN 작업은 에이전트의 마지막 위치가 목적지와의 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) 알고리즘으로 계산하는 보상 함수 SBA(Similarity Based Alignment)[6], 그리고 본 논문에서 제안한 보상 함수 RBA 등 3가지 보상 함수에 따른 VLN 작업 성능을 서로 비교하였다. RBA의 임계 거리 ϵ는 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.

<표 1> 보상 함수에 따른 성능 비교

Reward Function	Seen		Unseen	
	SC	SPL	SC	SPL
DBA[3]	0.273	0.041	0.225	0.031
SBA[6]	0.409	0.381	**0.405**	**0.382**
RBA	**0.436**	**0.414**	0.399	0.375

이 실험의 결과는 <표 1>과 같다. 본 논문에서 제안하는 RBA와 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 ϵ는 1.5m로, 학습 반복 횟수는 20만 번으로 각각 설정하였다.

<표 2> 학습 방법에 따른 성능 비교

Learning Strategy	Seen		Unseen	
	SC	SPL	SC	SPL
only RL	0.420	0.388	0.385	0.338
only IL	0.549	0.527	0.433	0.406
CIR ($\lambda = 0.05$)	**0.653**	**0.622**	**0.488**	**0.447**

이 실험의 결과는 <표 2>와 같다. 본 논문에서 제안한 복합 학습(CIR)의 성능이 가장 높았으며, 다음은 순수 모방 학습(IL), 순수 강화 학습(RL) 순으로 높은 성능을 나타내었다. 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.

마지막 실험은 기존의 VLN 모델들에 비해 본 논문에서 제안한 모델의 우수성을 입증하기 실험이다. 이 실험에서는 발화자 모델을 이용해 새로운 지시를 생성한 Speaker-Follower[4], 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.

<표 3> 기존 모델들과의 성능 비교

Model	Seen		Unseen	
	SC	SPL	SC	SPL
Speaker-Follower[4]	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 <표 3>과 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.

4. 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 논문에서는 R2R 데이터 집합과 Matterport3D 시뮬레이션 환경을 이용한 다양한 실험을 통해, 제안 모델의 우수한 성능을 확인할 수 있었다. 향후에는 기존의 Matterport3D 환경과 R2R 데이터를 이용하여 새로운 경로와 자연어 지시를 자동 생성하는 데이터 증강 기법을 연구할 계획이다.

참고 문헌

[1] P. Anderson, Q. Wu, and D. Teney, et al, “Vision-and-Language Navigation: Interpreting Visually-Grounded Navigation Instructions in Real Environments,” *Proc. of CVPR-2018*, pp. 3674-3683, 2018.

[2] X. Wang, Q. Huang, A. Celikyilmaz, “Reinforced Cross-Modal Matching and Self-Supervised Imitation Learning for Vision-Language Navigation,” *Proc. of CVPR-2019*, pp. 6629-6638, 2019.

[3] H. Tan, L. Yu, and M. Bansal, “Learning to Navigate Unseen Environments: Back Translation with Environmental Dropout,” *Proc. of NAACL-2019*, pp. 2610-2621, 2019.

[4] D. Fried, V. Cirik, and A. Rohrbach, et al, “Speaker-Follower Models for Vision-and-Language Navigation,” *Proc. of NeurIPS-2018*, pp. 3314-3325, 2018.

[5] V. Jain, G. Magalhaes, A. Ku, “Stay on the Path: Instruction Fidelity in Vision-and-Language Navigation,” *Proc. of ACL-2019*, pp. 1862-1872, 2019.

[6] G. Ilharco, V. Jain, and A. Ku, et al, “General Evaluation for Instruction Conditioned Navigation using Dynamic Time Warping,” *Proc. of NeurIPS-2019*, 2019.