# 농구 게임에서 상태 정규화 및 Dense 보상 기반 강화 학습 기법

최태혁 [1], 조경은 [1]*
[1]동국대학교 멀티미디어공학과
xogur6889@dgu.ac.kr, *cke@dongguk.edu(교신저자)

# State Normalization and Dense Reward Based Reinforcement Learning Method in Basketball Game.

Taehyeok Choi, Kyungeun Cho*
*Dept of Multimedia Engineering, Dongguk University

**요약**

최근 강화 학습을 적용한 게임 AI 에 대한 연구가 활발히 진행되고 있다. 하지만 대부분 상용 게임은 유한 상태 머신(Finite State Machine, FSM)을 이용한 스크립트 기반 AI 를 사용하기 때문에 복잡한 환경의 게임에서 불안정한 상태로 인해 적절한 강화 학습의 수행이 어렵다. 따라서 본 논문에서는 상용 게임 강화 학습 적용을 위하여 상태 정규화 및 Dense 보상 기반 강화 학습 기법을 제안한다. 제안한 기법을 상용 농구 게임에 적용하고 학습된 모델의 성능을 기존 FSM 기반 AI 와 비교를 통해 성능이 약 80% 증가한 결과를 확인하였다.

## 1. 서론

최근 게임 AI 에 대한 연구가 활발히 진행되고 있다[1]. 특히 게임 AI 중에서도 NPC 나 사람 플레이어를 대체하는 게임 플레이 AI 의 경우 강화 학습을 이용한 연구가 많다. 강화 학습에 대한 많은 연구에도 불구하고 대부분 상용 게임은 FSM 기반 AI 를 사용한다. 하지만 FSM 기반 AI 는 동일한 상황에서 획일적인 행동을 하므로 사람 플레이어와 쉽게 비교가 되고 유저의 만족도를 낮춘다. 그런데도 강화 학습이 아닌 FSM 을 이용하는 이유는 복잡한 환경의 게임에서 불안정한 상태로 인해 적절한 강화 학습의 수행이 어렵기 때문이다.

본 논문에서는 이와 같은 문제를 해결하기 위해 두 가지 기법을 제안한다. (1) 상태의 안정화를 위해 상대적인 상태 값 기반 정규화 기법을 이용한 상태 표현을 설계한다. (2) 행동에 대한 효과적인 평가를 위해 미리 정의된 Dense Reward(상세 보상) 기반 보상 함수를 설계한다.

제안한 두 가지 기법을 평가하기 위해 적용한 기본 강화 학습 알고리즘은 Advantage Actor-Critic(A2C) 알고리즘이다[2]. 상용 농구 게임에 제안한 기법을 적용하고 학습된 모델의 성능을 기존 FSM 기반 AI 와 그래프를 통해 비교 분석한다.

## 2. 관련 연구

본 논문에서 사용하는 강화 학습의 기법에는 많은 선행연구가 있다.

A2C 알고리즘은 기존 Policy Gradient(정책 경사) 방식에서 Critic 을 사용해 Value(가치)를 예측하여 분산을 줄인 알고리즘이다**오류! 참조 원본을 찾을 수 없습니다.**. A2C 에서는 Entropy loss 를 사용하여 Probability(행동 확률)의 차이가 너무 커지는 것을 방지해서 Soft-max 값이 발산하여 탐험이 줄어들고 Local Optima(국소 최적)에 빠지는 문제를 해결해준다.

강화 학습에서 상태는 특정 시점에서 에이전트에게 주어지는 상황에 대한 정보이다. 상태에 대한 정의를 무엇으로 정하는지에 따라 강화 학습 에이전트의 성능이 상이하다[3]. 또한 일반적으로 정규화된 상태를 사용하면 Optimizer 의 최적화가 빠르고 Local minima(극솟값)에 빠지는 위험이 감소한다[4].

강화 학습에서 보상은 특정 시점에서 에이전트의 행동에 대한 가치를 말한다. 보상 함수를 어떻게 설계하는지에 따라서 에이전트의 행동 정책이 달라지기 때문에 이를 정하는 것은 중요한 문제이다[5].

## 3. 상태 정규화 및 Dense 보상 기반 메타 강화 학습 기법

### 3.1 상대적인 상태 값 기반 정규화 기법을 이용한 상태 표현

본 논문에서 정규화 기법을 적용한 기존 환경의 원시 데이터는 위치, 시간, 조건이 있다.

위치는 환경에 존재하는 객체들의 절대좌표로 표현된다. 영상을 상태로 사용한 연구에서는 절대좌표를 사용한다고 볼 수 있지만[6], 값을 상태로 사용한 연구에서는 학습 대상이 되는 에이전트를 기준으로 상대적인 정규화된 좌푯값을 사용한다[7]. 본 논문에서는 가능한 최소 상대 좌표값을 -1 로, 최대 상대 좌푯값을 +1 로 설정하여 모든 위치 좌표 값을 -1 과 +1 사이의 값으로 변환하여 상태로 사용한다.

시간은 환경에 존재하는 시간과 관련된 초 단위의 값으로 음수 값이 존재하지 않기 때문에 가능한 최대 시간을 1 로 설정하고 나머지 값들을 0 과 1 사이의 값으로 정규화한다.

조건은 환경에 존재하는 객체들의 현재 상태에 대한 설명이다. 본 논문의 실험 환경인 농구 게임에서는 플레이어의 공 소유 여부, 슛하고 있는지 등에 대한 정보들이 조건이 된다. 이 정보들은 각각 참, 또는 거짓으로 표현된다. One-hot Encoding 방식 기반으로 0 과 1 의 배열 형태로 조건을 변환하여 상태로 사용한다.

위치, 시간, 조건 데이터들에 대해서 상대적인 상태 값 기반 정규화 기법을 사용하여 값을 변환하고 변환된 상태는 학습 모델의 입력이 된다.

### 3.2 미리 정의된 Dense reward 기반 보상 함수

[5]의 연구에서 보상에 따라 에이전트의 행동 정책이 달라지는 것을 주장했고 [7]의 연구에서는 실제 StarCraft Micromanagement 환경에서 보상에 따라 에이전트의 승률이 다르게 수렴하는 것을 실험을 통해 증명했다. 본 논문에서는 Dense reward 기반으로 농구게임에서 에이전트의 행동에 대한 평가를 위한 보상을 자세히 나누고 미리 정의된 보상에 따라 에이전트를 학습했다. 정의된 보상에는 패스, 슛, 블로킹, 스틸, 마크, 캐치가 있다. 아래 그림 1 은 몇 가지 보상에 대한 간단한 식을 보여준다.

| 명칭 | 수식 |
|---|---|
| **슛** | 거리 = $\sqrt{(\text{에이전트 위치} - \text{골대 위치})^2}$<br>**(1.0 - min(거리 x 0.1, 1.0))** |
| **마크** | 단위 방향 = norm(골대 위치 - 상대 위치)<br>마크 위치 = 상대 위치 + 단위 방향<br>마크 방향 = 마크 위치 - 에이전트 위치<br>거리 = $\sqrt{(\text{마크 방향})^2}$<br>**(0.1 x (거리 < 1))** |
| **스틸** | 거리 = $\sqrt{(\text{에이전트 위치} - \text{상대 위치})^2}$<br>**(1 x (상대 공 소유) x (거리 < 최대 스틸 거리) x (스틸 시도))** |
| **블록** | 거리 = $\sqrt{(\text{에이전트 위치} - \text{상대 위치})^2}$<br>**(1 x (상대 공 소유) x (거리 < 최대 블록 거리) x (블록 시도))** |

그림 1. Dense reward 기반 보상 계산식

## 4. 실험

본 실험을 위하여 본 논문에서 사용한 환경은 3 대 3 농구 게임 환경이다. 본 환경의 에이전트는 AI 로부터 행동을 받아서 위치나 상태가 변하는 농구 게임의 플레이어이다. 실험에 사용한 농구 게임 시뮬레이션 환경에 대한 게임 뷰는 아래 그림 2 와 같다.

그림 2. 실험 환경

본 논문에서는 농구 게임에 강화 학습을 적용하기 위해서 A2C 알고리즘을 기반으로 실험을 진행했다. 학습에 적용한 파라미터는 아래 그림 3 과 같다.

| Hyperparameters | Value |
|---|---|
| Architecture | Input(상태 수 * 4) |
| | Relu(128) |
| | Relu(128) |
| | Softmax(행동 수) |
| Learning rate | 0.0003 |
| Entropy regularization | 0.001 |
| Gamma | 0.99 |

그림 3. 하이퍼 파라미터 정보

본 논문에서 제안한 기법을 농구 게임에 적용하여 실험을 진행한 결과는 아래 그림 4 와 같다. 학습 진행에 따라 보상이 증가하는 것을 확인할 수 있다.

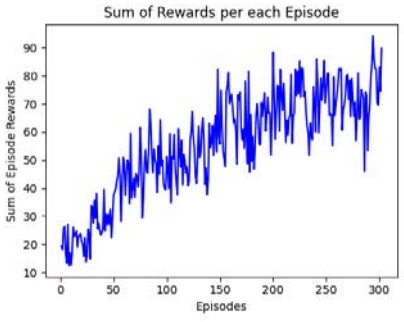


그림 4. 제안 기법 기반 강화 학습 보상 그래프

본 논문에서 제안한 기법의 우수성을 검증하기 위해 비교 실험 대상인 기존 FSM 기반 AI 와 성능을 비교한 결과는 아래 그림 5 와 같다. 평가 기준은 농구 게임에서 가장 간단하고 명확한 기준인 점수를 사용하였다. 제안한 방법을 사용하였을 때 큰 차이로

AI 의 성능이 향상되는 것을 알 수 있다.

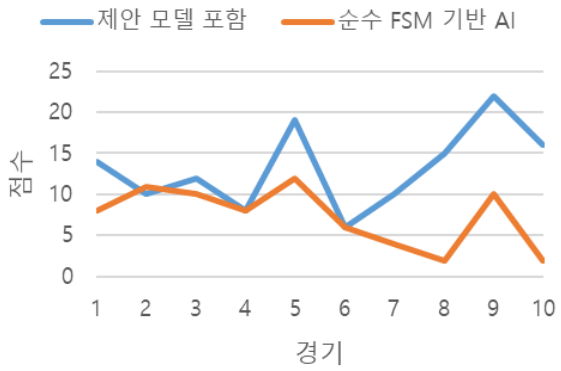


그림 5. 3 대 3 농구 게임에서 제안한 기법을 사용한 모델이 포함된 팀과 순수한 FSM 기반 AI 팀의 점수 차이 그래프

실험을 통해서 본 논문에서 제안한 기법이 모델의 성능에 얼마나 큰 영향을 주는지 확인하였다.

## 5. 결론

본 논문에서는 농구게임에서 상태 정규화 및 Dense 보상 기반 강화 학습 기법을 제안한다. 실험을 통해 제안한 기법을 적용하였을 때 기존 FSM 기반 AI 대비 성능이 약 80% 증가하는 것을 확인하였다. 향후 연구에서는 게임 AI 의 주목적인 사용자의 흥미를 증가시킬 수 있도록 인간 유사 행동 기반의 설계를 추가하여 방법을 개선할 것이다.

## 사사

본 연구는 문화체육관광부 및 한국콘텐츠진흥원의 2022 년도 문화기술 연구개발 사업으로 수행되었음 (과제명 : 스포츠 게임 분야 첨단 AI 기술 R&D 전문인력 양성, 과제번호 : R2022020003, 기여율: 00%)

## 참고문헌

[1] Oriol Vinyals, et al. “Grandmaster level in StarCraft II using multi-agent reinforcement learning.” Nature, 2019.
[2] Mnih, V. et al. “Asynchronous methods for deep reinforcement learning.” In ICML, 2016.
[3] X. B. Peng, et al. "Learning locomotion skills using deeprl," Proceedings of the ACM SIGGRAPH / Euro-graphics Symposium on Computer Animation, 2017.
[4] Sergey Ioffe, et al. “Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift” Proceedings of the 32nd International Conference on Machine Learning, 2015.
[5] Ng, A. Y. et al. “Policy invariance under reward transformations : theory and application to reward shaping.” in Sixteenth International Conference on Machine Learning 3, 1999.
[6] Mnih, K. et al. “Humanlevel control through deep reinforcement learning”. Nature, 2015.
[7] Shao, K. et al. "StarCraft micromanagement with reinforcement learning and curriculum transfer learning." IEEE Trans. Emerg. Top. Comput. Intell. 3, 2019.