Paper Title

김신영(18)1, 오상훈(18)2

*1*Author(s) Affiliation(s)

*2Author(s) Affiliation(s)*

email@xx.xx1*, dec@xx.xx2*

# *Abstract*

본 논문은 스타크래프트 2를 활용한 강화학습의 적용성을 탐구한다. 이 연구의 주요 목적은 게임 내에서 최적화된 빌드 전략을 개발하여 인공지능의 성능을 향상시키는 것이며, 사용자 편의성 증대에 중점을 두고 있다. 특히, 이 연구는 사용자가 제시한 특정 조건에 맞추어 최적화된 빌드 전략을 생성하는 인공지능 시스템을 개발하는 데 집중한다. Q-Learning을 활용하여 사용자 명령 순서를 최적화함으로써, 게임 플레이어들이 효율적이고 효과적인 게임 경험을 할 수 있도록 돕는다. 연구 결과는 강화학습을 통한 게임 전략 개발의 가능성을 입증하며, 게임 인공지능 분야에 중요한 기여를 한다. 또한, 이 연구는 스타크래프트 2가 강화학습 연구에 적합한 환경임을 보여주며, 복잡한 정보 아래에서의 최적 결정 모델링 능력을 입증하는 중요한 사례로 자리매김한다.

***Keywords:*** *Write more than 4-6 Keywords.*

1. Introduction

Advanced Engineering and ICT-Convergence Proceedings (AEICP)

ISSN : 2635-4586

©ICTAES 2018

컴퓨터 게임, 특히 스타크래프트 2는 강화학습 연구에서 핵심적인 역할을 한다. 이 실시간 전략 게임은 광범위한 행동 공간과 다중 에이전트 환경을 제공하며, 다양한 종족, 유닛, 건물, 리소스가 상호작용하는 복잡한 환경을 통해 현실 세계의 다양한 문제를 모델링하는 데 이상적이다. 또한, 불완전한 정보를 기반으로 하는 게임 플레이는 강화학습 알고리즘의 일반화 및 의사결정 능력을 시험하는 데 중요한 도전 과제를 제공한다.

본 연구에서는 스타크래프트 2의 이러한 독특하고 복잡한 환경을 활용하여 최적의 빌드 전략을 개발하고, 특정 유닛의 효율적인 생산을 통해 전투에서 승리하는 방법을 탐구한다. 이 연구는 스타크래프트 2를 통한 강화학습의 적용 가능성을 탐색하며, 게임 전략 개발과 인공지능 분야에 새로운 기여를 목표로 한다. 스타크래프트 2의 복잡한 환경을 이해하고 이를 바탕으로 한 강화학습 알고리즘의 효과를 검증함으로써, 게임 인공지능 발전에 중요한 통찰을 제공할 것이다.

2. Background

스타크래프트 2의 다중 에이전트 환경은 강화학습 알고리즘에 다양한 전략적 선택과 의사 결정 기회를 제공한다. 플레이어는 게임 내 부분적으로 공개된 정보를 바탕으로 전략을 수립해야 하며, 이는 강화학습 연구에 있어 불완전한 정보 상황에서의 의사결정 능력을 시험하는 좋은 사례가 된다. 이러한 환경에서 AlphaStar와 같은 혁신적인 프로젝트가 탄생했다.[1] AlphaStar는 강화학습과 신경망 기술을 활용하여 게임 전략을 학습하고 경쟁 상대를 이기는 방법을 개발했으며, 이는 높은 차원의 행동 공간과 다중 에이전트 환경에서 작동하는 AI의 중요한 성과로 간주된다. 이후에는 스타크래프트 2에서 다중 에이전트 강화학습을 활용하여 그랜드 마스터 수준의 성과를 달성하는 방법 또한 나왔다.[2] 스타크래프트 2의 다중 에이전트 환경이 강화학습 연구에 적합하다는 점을 강조하며, 다중 에이전트 강화학습 알고리즘을 통해 경기력을 향상시키는 연구가 나온 것이다. 결론적으로, 스타크래프트 2는 강화학습 연구에 적합한 환경을 제공하며, 이를 통해 강화학습 알고리즘의 다양한 측면을 검증하고 게임 인공지능 분야에 새로운 기여를 할 수 있는 중요한 연구 분야로 인정받고 있다.[3]

**3.** Methodology

본 연구는 스타크래프트 2 내에서 사용자가 원하는 조건에 맞춰 최적화된 빌드 전략을 개발하는 데 중점을 두었다. 연구의 주 목적은 사용자 편의성을 증대시키고, 특정 유닛의 생산을 최적화하는 데 필요한 빌드 순서를 정교하게 계산하는 것이다. 이를 위해, 강화학습 알고리즘을 활용하여 사용자가 설정한 조건에 따라 최적화된 빌드 전략을 자동으로 생성하는 시스템을 개발했다.

본 연구는 멀티가 하나인 가정 하에서, 사용자가 원하는 특정 시점에 최적의 빌드 순서를 계산한다. 강화학습의 평가는 동일한 맵과 초기 상태에서 시작하여, 마르코프 연쇄 법칙을 기반으로 한 이론적 계산과 강화학습 결과를 비교 분석했다. 이 과정에서 복잡한 게임 환경과 다양한 게임 상황에 대응하는 최적의 빌드 전략을 도출했다.

이 연구에서는 Q-Learning 알고리즘을 사용하여, 사용자 명령 순서의 최적화에 중점을 두었다. [5] 특히, 전투 전략이나 컨트롤, 미네랄 수급률 최적화는 고려 대상이 아니었으며, 오로지 빌드 순서의 최적화에만 집중했다. 결과적으로, 사용자가 설정한 조건에 맞추어 최적화된 빌드 전략을 자동으로 생성하는 시스템은 사용자 경험을 개선하고, 게임 내 효율적인 의사결정을 지원할 것이다.

**3-1. 비동기 프로그래밍**

**3-2. Q-Learning**

Q Learning은 강화학습의 방법론 중 하나로, 주어진 환경에서 에이전트가 어떤 상태(state)에 있을 때 최적의 행동(action)을 결정하는 문제에 초점을 맞춘다. 핵심 아이디어는 Q-값이라는 가치 함수를 통해 각 상태와 행동 조합에 대한 보상의 기대치를 학습하고 이를 바탕으로 최적의 행동을 선택하는 것이다. MDP(Markov Decision Process) 이론을 기반으로 하는 대표적인 알고리즘 중 하나이다.

Q값은 Q로 표현되며, 각 시간 t에서의 상태는 , 그 상태에서 취할 수 있는 행동은 로 나타낸다. α는 학습률로 새로운 Q값 추정치가 기존의 Q값에 얼마나 영향을 미치는지를 조절하고, 는 할인요인으로, 미래의 보상이 현재 보상에 비해 얼마나 중요한지를 나타낸다. 또한 는 t에서의 상태와 행동에 대한 보상, 는 다음 시간 t+1에서의 상태에서 가능한 모든 행동 에 대한 최대 Q값을 나타낸다. 학습 과정에서 Q-값은 Bellman Equation을 기반으로 한 공식을 통해 업데이트된다. 그 공식은 다음과 같다:

[6]

각 시간 t에서, 에이전트는 현재 상태()에서 행동()을 선택하고, 그 결과로 새로운 상태()로 이동하게 된다. 이때 에이전트는 특정 보상()을 받게 되며, 이 정보를 이용하여 Q값을 업데이트하게 된다. 이런 방식으로 에이전트는 환경과 상호작용하며 학습하게 되고, 시간이 지남에 따라 최적의 행동 전략을 찾아가게 된다.

학습과정에서는 탐험(Exploration)과 활용(Exploitation)의 균형을 유지하기 위해 입실론 탐욕적 정책(ϵ-greedy Policy)을 사용한다. 확률 ϵ로 무작위 행동을 선택하고, 확률1−ϵ로 현재 Q값이 가장 높은 행동을 선택한다. 시간이 지남에 따라 ϵ은 줄어들어서 점점 더 많이 활용하는 전략으로 전환될 수 있다.

Q-Learning에서는 일반적으로 Mean Squared Error(MSE)를 손실 함수로 사용하여 Q값의 예측값과 실제값 사이의 차이를 최소화하려고 한다. MSE의 공식은 다음과 같다:

는 예측값, 는 실제값, N은 배치 크기를 나타낸다. 여기서 이러한 설정을 통해 에이전트는 각 상태와 행동에서 예상되는 미래 보상을 최대화하도록 학습되었다. 이 과정은 일정한 성능 향상이 이루어지거나, 최대 반복 횟수에 도달할 때까지 반복적으로 진행된다.

Q-Learning은 환경 모델을 필요로 하지 않고 게임 내 상황에 따라 병렬처리를 통해 여러 상태와 행동 조합을 동시에 학습할 수 있는 강화 학습 기법 중 하나이다. 따라서, 스타크래프트2와 같이 복잡한 게임에 적용하기에 적합하다 판단했다. 본 논문에서는 봇 구현을 위해 python-sc2 라이브러리를 사용하였다.[4] 이 라이브러리는 StarCraft II의 AI 개발을 돕기 위해 제공되며, Python 환경에서 강화 학습 등의 기법을 통해 게임 내 에이전트를 손쉽게 구현할 수 있게 해준다.

스타크래프트2 환경에서 강화 학습을 적용하려면 다양한 환경 요소를 고려해야 한다. 예를 들면, 유닛의 움직임, 미네랄 위치 등이 있다. 본 연구에서는 Void Ray 유닛 5대가 생성되는 시간을 중심으로 보상 시스템을 설계하였다. 이를 위해 iteration을 활용하여 보상을 정의하였고, 보상은 에이전트가 Void Ray를 빠르게 생산하는 지에 따라 부여되었다.

일반적인 상황에서, 사용자가 임의로 정한 최적의 빌드를 따를 경우 Void Ray 5대를 생산하는 데 약 6분이 걸릴 것이라는 가정하에 시작하였다. 그러나 단 하나의 Gateway, Cybernetics Core, Stargate만을 건설한다는 제한된 환경에서 Q-Learning을 적용하였을 때, 일꾼을 추가로 생산하지 않고 테크를 타는 정책을 통해서 5분 33초 만에 Void Ray 5대를 생산하는 결과를 얻을 수 있었다. 이는 기존 가정보다 빠른 시간이다. 이 때 모델의 수렴 속도를 높이기 위해 프루닝 기법을 활용하여 특정 제한 조건을 설정하였다. 그 결과, 모델 학습의 수렴 속도가 보다 효과적으로 최적화되었다. 또한, Stargate의 설치 수를 늘리거나 여러 일꾼을 생성하는 등의 다양한 환경 조치를 취했을 때의 복잡한 결과를 사람이 직접 예측하기는 어렵지만, Q-Learning을 그 결과를 분석할 수 있었다.

결론적으로, Q-Learning 방법론은 복잡한 게임 환경인 스타크래프트2에서도 효과적으로 적용될 수 있음을 확인하였다. 특히, 주어진 제한 조건과 보상 시스템 하에서 에이전트가 최적의 전략을 성공적으로 학습하였으며, 이를 통해 게임 내에서의 성능을 효과적으로 향상시킬 수 있음을 보여주었다.

**4. Conclusion**

본 연구에서는 스타크래프트2를 강화학습 연구의 중요한 대상으로 제시하고 게임 내 빌드 전략 및 전투 전략을 개발하여 그 가치를 입증하였다. 스타크래프트2 환경은 강화학습 알고리즘의 일반화 능력을 검증하고, 불확실하고 다양한 정보 아래에서 최적의 결정을 내리는 능력을 모델링하는데 매우 적합한 것으로 나타났다.

이러한 연구를 통해 게임 환경에서의 강화학습 연구의 중요성과 가능성을 강조하며, 게임 인공지능 분야에 새로운 기여를 제공함을 확인했다. 미네랄 수집과 전투 전략 최적화를 통해 게임에서의 인공지능 성능을 향상시킬 수 있음을 입증하였으며, 앞으로의 연구에서는 이러한 결과를 발전시켜 게임 인공지능 분야에 더 큰 기여를 할 것으로 기대한다.

Table 1. Main parameters

|  |  |  |
| --- | --- | --- |
| **Heading level** | **Example** | **Font size and style** |
| Title (centered) | Lecture Notes … | 14 point, bold |
| 1st-level heading | 1 Introduction | 13 point, bold |
| 2nd-level heading | *2.1 Printing Area* | 11 point, italic |
| 3rd-level heading | *Remark.* Text follows … | 11 point, italic |

4.1. Figures

Grids and details within the figures must be clearly legible and may not be written one on top of the other. Figures should be numbered and should have a caption which should always be positioned under the figures, in contrast to the caption belonging to a table, which should always appear above the table. Please center the captions between the margins and set them in 9-point type and Times New Romans font. The distance between figure and caption should be about 6 pt, and the distance between caption and text about 10 pt. Use of 600 dpi/1200 dpi for illustrations for clarity recommended.

Please be noted that illustrations are black and white in printed volumes. However, colored pictures will be shown in the electronic version. If you send colored figures that are to be printed in black and white, please make sure that they really are legible in black and white. Some colors show up very poorly when printed in black and white.



Figure 1. Sample.

*4.2. Equation*

The equation should be written by using MS word as follows:

(1)

*4.3. Units, Equations and Symbols*

For units, the SI system should be used for all dimensional quantities. Leave a space between numbers and units except oC and % (e.g., 40 mg/L, 50%, 10 h).

Main symbols used in equations and text should be in italic. Equations must be numbered consecutively using Arabic numbers.

**5. Conclusion**

Please be noted that illustrations are black and white in printed volumes. However, colored pictures will be shown in the electronic version. If you send colored figures that are to be printed in black and white, please make sure that they really are legible in black and white. Some colors show up very poorly when printed in black and white.

**Acknowledgement**

Acknowledgement title is not numbered. Type the acknowledgement in this format ‘This work was supported by (institution to acknowledge) in (year).’

Note: Manuscripts in which references are not in this format will be returned without review.

**References**

1. The AlphaStar team (2019). AlphaStar: Mastering the real-time strategy game StarCraft II.
2. Oriol Vinyals, Igor Babuschkin, Wojciech M. Czarnecki (2019). Grandmaster level in StarCraft II using multi-agent reinforcement learning. Nature volume 575, 350–354
3. Oriol Vinyals, Timo Ewalds, Sergey Bartunov (2017). StarCraft II: A New Challenge for Reinforcement Learning.
4. <https://github.com/BurnySc2/python-sc2/tree/develop>
5. 고민수 외 1명, 2023, 『Deep Learning Bible - 5. Reinforcement Learning』, e-book , https://wikidocs.net/165849

[6] Q-Learning Policy and Reward Design for Efficient Path Selection Sung-Jung Yong· Hyo-Gyeong Park · Yeon-Hwi You· Il-Young Moon\* Department of Computer Science and Engineering, Korea University of Technology and Education, Cheonan, 31253, Korea