# 스타크래프트 II 미니게임을 위한 멀티태스크 강화학습

심우일○, 김경중*
세종대학교 컴퓨터공학부
piranas2067@gmail.com, kimkj@sejong.ac.kr

# Multitask Reinforcement Learning for Starcraft II Mini Games

Wooil Shim○, Kyung-Joong Kim*
Department of Computer Science and Engineering, Sejong University

**요 약**

최근 실시간 전략 게임을 심층 강화학습으로 해결하는 시도가 많이 일어나고 있다. 고전 게임과는 달리 실시간 전략게임은 실생활 문제에 가장 근접한 난이도를 가지고 있기 때문이다. 최근 딥마인드와 텐센트에서 스타크래프트 II의 풀 게임을 풀기 위한 시도가 있는 등 스타크래프트 II가 실시간 전략 게임 중에서 가장 활발하게 연구되고 있는 분야이다. 본 논문에서는 스타크래프트 II에서 멀티태스크기반 강화학습을 통해 여러 미니 게임을 하나의 신경망을 이용하여 푸는 방법을 제안한다. 본 연구를 통하여 실시간 전략 게임의 다양한 작은 문제들을 하나의 네트워크로 해결 할 수 있는 가능성을 제안하도록 한다.

## 1. 서 론

게임 인공지능 분야에서 실시간 전략 게임을 이용한 많은 연구가 이루어지고 있다. 바둑이나 아타리 게임과 같은 고전 게임의 경우는 게임의 모든 정보를 이용하여 에이전트가 학습할 수 있지만, 실시간 전략 게임은 한정된 정보만을 이용하여 에이전트가 학습한다는 한계가 있다. 따라서 에이전트의 학습 속도가 느리고, 정확하지 못하다. 고전 게임보다 문제의 난이도가 높은 실시간 전략 게임을 해결하려는 시도가 많이 나타나고 있다. 대표적인 실시간 전략 게임으로는 스타크래프트 II가 있다.

그림 1 스타크래프트 II

스타크래프트 II는 블리자드 엔터테인먼트에서 개발한 실시간 전략 시뮬레이션 게임이다. 게임 속 종족은 세 개가 있으며 다양한 맵 속에서 상대방의 위치를 찾아 건물을 모두 없애면 이기는 게임이다. 전장의 안개가 설정되어 있어 현재 아군이 있는 위치만 보이기 때문에 실시간으로 상대방의 모든 행동이 보이지 않는다. 각 종족의 유닛들이 가지고 있는 고유 능력들을 가지고 상대방의 유닛을 무력화 시킬 수 있고, 대규모 전투를 통해서 상대방을 무력화 시킬 수 있다.

기존의 연구는 스타크래프트 게임에서 소수 유닛 단위 전투를 제어하는 모델 혹은 상대방의 빌드를 예측하는 연구들이 주를 이루었다[1,2]. 하지만 최근 스타크래프트 II의 전체 게임을 한번에 해결하려는 시도들이 나타나고 있다. 대표적으로 구글 딥마인드[3]와 중국의 텐센트[4]이다. 딥마인드는 새로운 알고리즘을 이용하여 전체 게임을 학습하는 모델을 만들었고, 텐센트는 매크로 행동을 정의하여 전체 게임을 풀려는 시도를 하였다.

본 논문에서는 스타크래프트 풀 게임 학습의 선행 연구로 스타크래프트 학습 환경에 있는 미니게임들을 이용하여 새로운 방식인 멀티태스크 기반 강화학습을 통해 문제를 해결하는 방법을 제안한다.

## 2. 배경 및 관련연구

### 스타크래프트 학습환경(SC2LE)

스타크래프트 II의 학습환경(SC2LE)[5]이 발표되고 나서 많은 연구자들이 강화학습 연구에 스타크래프트 II를 사용하고 있다. SC2LE에서 현재 보고 있는 화면과 미니맵을 이미지, 보상, 현재 자원, 가능한 행동 등을 가져올 수 있다. 그 뿐 아니라 원하는 특징을 추출하여 가져와서 신경망을 학습 시킬 수 있다. 이를 통해서 신경망에서 출력한 행동을 가지고 게임을 진행하는 것을 반복한다.

### 행동- 가치 평가 학습[6,7]

행동 가치 평가 학습은 확률을 통해 행동을 취하는 에이전트와 가치함수를 통해 현재 상태를 평가하는 구조이다. 이때까지 취한 행동들이 얼마나 좋아져야

하는지 평가하여 에이전트가 가장 좋은 행동을 할 수 있게 만드는 방법이다.

$$-\nabla_\theta \log \pi_\theta(a_t|s_t)(r_{t+1} + \gamma(V_v(s_{t+1}) - V_v(s_t))$$

수식 1 actor loss 정의

$$(\mathrm{r}_{t+1} + \gamma V_v(s_{t+1}) - V_v(s_t))^2$$

수식 1 critic loss 정의

## 3. 멀티태스크 강화학습

기존 멀티태스크 학습은 문제를 해결할 때 유사한 문제의 정보를 이용하여 원래 문제의 해법의 결과를 더 좋게 만들 수 있다 [8]. 스타크래프트 II와 같이 문제가 큰 경우에는 미니 게임단위로 분리하여 학습을 진행하도록 구성하도록 하였다. 이를 본 논문에서 멀티태스크 강화학습이라고 재정의하였다. 아래는 하나의 신경망으로 미니 게임들을 학습시키는 두 가지 방법을 제안하였다. 각 미니 게임을 학습하는 알고리즘은 딥마인드 논문[5]과 마찬가지로 Actor-critic[7] 알고리즘을 사용하였다. 신경망의 입력은 SC2LE에서 가져오는 미니 맵 이미지와 현재 카메라가 보고 있는 이미지를 사용한다.

### 3.1 에피소드 기반 학습(N-episode based)

1000번의 에피소드 동안 하나의 미니 게임을 학습한다. 하나의 미니 게임에 대해 1000번을 학습하고 나면, 결과에 상관없이 다음 미니 게임으로 넘어가서 해당 미니게임을 대상으로 학습을 수행한다.

### 3.2 최대 점수 기반 학습(Max-scored based)

딥마인드 논문[5]에서 나온 학습 에이전트 별로 얻은 Best Mean 점수를 에이전트가 얻을 수 있는 최대 점수라고 판단하고 학습을 진행한다. 만약 딥마인드에서 나오지 않은 미니게임을 학습할 때는 강화학습을 이용하여 나온 결과를 이용하여 판단한다. 에이전트가 각 맵에 미리 설정해 놓은 점수 이상을 받는 경우 해당 미니 게임의 학습을 완료했다고 판단하여 해당 미니 게임의 학습을 종료하고 다음 미니 게임을 진행한다.

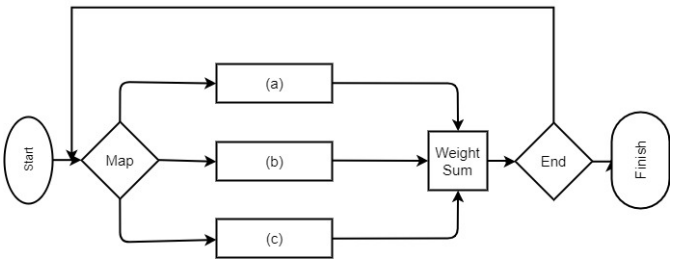


그림 2 멀티태스크 강화학습 Flow Chart. (a) 비콘 이동, (b) 미네랄 수집, (c) 저글링 제거 미니 게임이다.

그림 2와 같이 모든 미니 게임에 대한 학습이 끝나면, 처음 미니 게임으로 돌아가 반복적으로 학습을 진행한다.

### 3.3 미니 게임

총 7개의 미니 게임이 있지만, 아군 유닛이 해병만 사용하는 3개의 미니 게임을 가지고 학습을 진행했다. 사용한 3개의 미니 게임은 비콘 이동, 미네랄 수집, 저글링 제거이다. 비콘 이동 미니 게임은 무작위로 나타나는 비콘을 마린 1기가 따라가서 점수를 얻는 게임이다. 한 에피소드가 120초 동안 진행되며, 비콘에 도착하게 되면 보상 값을 얻고, 비콘은 다른 지점에 나타나게 된다. 게임에서 전장의 안개는 없다고 가정한다. 미네랄 수집 미니 게임은 마린 2기가 맵 상에 무작위로 나타난 광물들을 정해진 시간에 수집하는 게임이다. 광물을 얻은 경우 보상 값을 얻게 되며, 120초 동안 진행된다. 게임에서 전장의 안개는 없다고 가정한다. 마지막으로 저글링 제거 미니 게임은 안개에 가려진 맵에서 마린 3기가 25마리 저글링을 모두 찾아 공격하여 제거하는 게임이다. 한 에피소드가 180초 동안 진행되고, 아군이 모두 죽거나 시간이 지나면 게임이 끝난다. 적 저글링을 제거한 경우 양의 보상 값을 얻고, 아군이 죽을 경우는 음의 보상 값을 얻는다.

## 4. 실험 결과

앞에서 제안한 두 가지 방법을 이용하여 SC2LE에 있는 3개의 미니게임을 이용하여 실험을 진행하였다. 실험은 비콘 이동, 미네랄 수집, 저글링 제거 순서로 맵을 변경하며 진행하였고, 총 70000 에피소드를 진행하였다.

그림 3은 본 논문에서 제안한 두 가지 방법으로 학습한 신경망의 테스트 그래프이다. 70000번 학습 한 에이전트가 3개의 미니 게임을 30번씩 테스트하였다.

그림 3에서 비콘 이동 미니 게임과 저글링 제거 미니 게임은 두 방법이 비슷한 성능을 나타내었다. 하지만 미네랄 수집 맵에서는 최대 점수 기반학습 방법이 에피소드 기반학습 방법보다 높은 성능을 나타내었다. 그 이유는 최대 점수 기반학습 방법은 단일 게임에서 얻을 수 있는 최대 점수를 얻기 전까지 학습을 진행하지만, 에피소드 기반 학습 방법은 그렇지 못하고 정해진 에피소드 마다 미니 게임을 바꿔가며 학습을 진행하기 때문에 높은 점수를 얻지 못했기 때문이다. 또한 테스트 그래프에서 일정 값을 내지 못하는 이유는 미니 게임에서 무작위로 설정되는 것들이 있기 때문에 움직이는 에이전트와 가깝게 설정되는 경우에는 상대적으로 높은 점수를 얻지만, 그렇지 않은 경우에는 상대적으로 낮은 점수를 얻는다.

따라서 두 가지 방법 중에서 최대 점수를 기반으로 한 방법이 에피소드를 기반으로 한 방법보다 평균적으로 얻은 점수가 높았다. 즉, 최대 점수를 기반으로 한 방법이 성능이 좋았다.

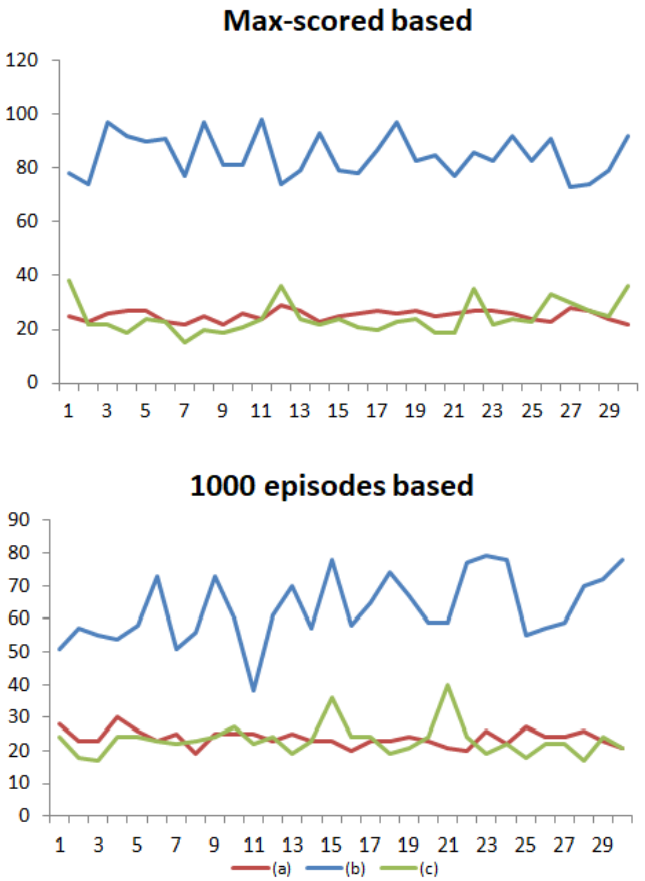


그림 3 최대 점수 기반학습(Max-scored based) 방법과 에피소드 기반학습(1000-episode based) 방법을 이용하여 학습시킨 신경망의 테스트 그래프. 가로축은 에피소드, 세로축은 한 에피소드 당 얻은 보상을 나타낸다. (a)는 비콘이동, (b)는 미네랄 수집, (c)는 저글링 제거 맵이다.

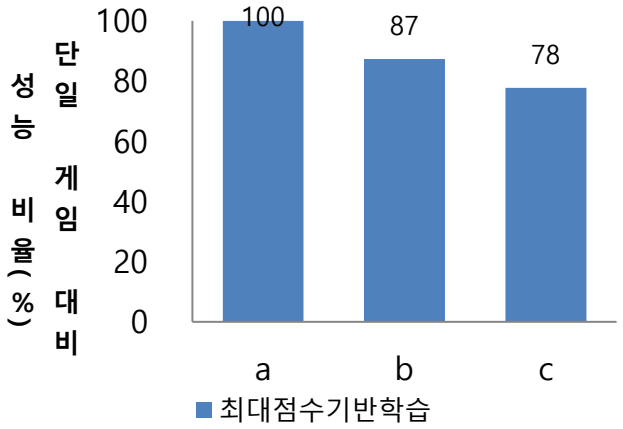


그림 4 멀티태스크 학습을 한 에이전트의 단일 게임 대비 성능 비교 그래프. 단일 게임 성능을 100퍼센트라 가정하고 최대 점수 기반 학습 방법으로 얻은 성능을 나타냈다. 가로축의 a는 비콘 이동, b는 미네랄 수집, c는 저글링 제거 맵이다.

그림 3는 본 논문에서 제안한 방법의 결과와 딥마인드 논문[5]에서 나타낸 결과를 비교한 그래프이다. 딥마인드 논문에서의 점수를 100이라 가정했을 때, 제안한 방법이 얻은 점수를 나타냈다. 그림 3에 따르면 3개의 미니 게임에서 단일 게임 대비 성능 비율이 모두 75퍼센트 이상 나타내었다.

## 5. 결론

본 논문에서는 스타크래프트 미니게임을 멀티태스크 기반 강화학습을 이용하여 연구를 진행하였다. 실험결과에서 2번째 미니 게임을 제외하고 각각의 미니게임을 학습시킨 모델과 비교하였을 때, 비슷한 수준의 점수를 얻었다. 또한 기존의 연구에서 진행하지 않은 방식을 이용하여 실험결과를 통해 새로운 방식이 미니 게임을 어느 정도 해결할 수 있다고 보였다.

하지만 미니 게임들을 새롭게 정의할 때 이전 미니 게임들과의 유사성을 정의해서 전체 게임에 적용하였을 때, 어느 정도 성능이 나올 수 있을지는 앞으로 연구할 과제이다.

## 6. 감사의 글

이 논문은 2017년도 정부(미래창조과학부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임 (2017R1A2B4002164). *교신저자

## 참고 문헌

[1] Shao, Kun, Yuanheng Zhu, and Dongbin Zhao. "StarCraft Micromanagement with Reinforcement Learning and Curriculum Transfer Learning." *IEEE Transactions on Emerging Topics in Computational Intelligence* (2018).

[2] Wender, Stefan, and Ian Watson. "Applying reinforcement learning to small scale combat in the real-time strategy game StarCraft: Broodwar." *Computational Intelligence and Games (CIG), 2012 IEEE Conference on*. IEEE, 2012.

[3] Zambaldi, Vinicius, et al. "Relational Deep Reinforcement Learning." *arXiv preprint arXiv:1806.01830* (2018).

[4] Sun, Peng, et al. "TStarBots: Defeating the Cheating Level Builtin AI in StarCraft II in the Full Game." *arXiv preprint arXiv:1809.07193* (2018).

[5] Vinyals, Oriol, et al. "Starcraft ii: A new challenge for reinforcement learning." *arXiv preprint arXiv:1708.04782* (2017).

[6] Mnih, Volodymyr, et al. "Asynchronous methods for deep reinforcement learning." *International conference on machine learning*. 2016.

[7] Konda, Vijay R., and John N. Tsitsiklis. "Actor-critic algorithms." *Advances in neural information processing systems*. 2000.

[8] R. Caruana, Multitask learning, Machine Learning, vol. 28, 41-75p,1997