

마인크래프트 강화학습 환경 구축을 통한 멀티모달 연구

와 특정 태스크에서의 우수성 비교

2019-13674 양현서 / 지도교수 장병탁

@yhs0602@snu.ac.kr

서론

연구 주제

강화학습을 위한 마인크래프트 환경에서 어둠 상태 효과를 활용한 시야 제한 태스크와 멀티모달 정보 활용의 중요성

연구 동기 및 목적

강화학습 에이전트들은 Atari 게임 등에서 높은 성능을 보였다. 하지만 이러한 문제들은 특정 도메인에 한정되어 있고, 다양한 입력 모드를 받지 않는다는 한계가 있다. 따라서, 다양한 도메인과 입력 모드를 포함하는 새로운 강화학습 환경의 구축이 필요하다. 이를 위해 마인크래프트를 활용할 수 있다. 마인크래프트는 다양한 상황과 조건을 가지고 있어 강화학습 알고리즘의 적용과 실험에 적합하다. 이렇게 마인크래프트를 강화학습 알고리즘 환경으로 활용한 관련 연구로는 MineDojo, MineRL 등이 있다. 그러나 이러한 환경들은 오래된 모드 프레임워크인 Forge를 사용하여 업데이트가 느리며, MineDojo는 마인크래프트 1.11.2 버전까지만 지원된다. 마인크래프트 1.19에는 "어둠(Darkness)" 상태 효과가 등장했다. 이는 플레이어의 시야를 좁게 만드는 상태 효과로, 새로운 도전 요소를 제공한다. 본 연구에서는 기존 연구들에서 지원하지 않은 최신 버전의 마인크래프트 환경의 어둠 상태 효과를 활용하여 현실의 야간 작업 등 시야가 제한된 상태에서의 태스크들을 제시하고, 기존 비전 기반 모델이 기존 환경에서 보이던 성능과 비교한다. 또한 기존 연구들에서 사용되지 않은 소리 정보를 활용하는 에이전트를 실험하여 멀티모달 정보 활용의 중요성을 강조하고 이를 통한 성능 향상을 확인할 것이다. 이러한 연구는 강화학습 알고리즘의 발전과 실제 응용에 새로운 통찰을 제공할 것으로 기대된다.

이론적 배경

강화학습

강화학습은 에이전트가 환경과 상호작용하며, 보상을 최대화하는 행동을 학습하는 방법이다. 강화학습의 학습 과정은 다음 그림에서 볼 수 있다.

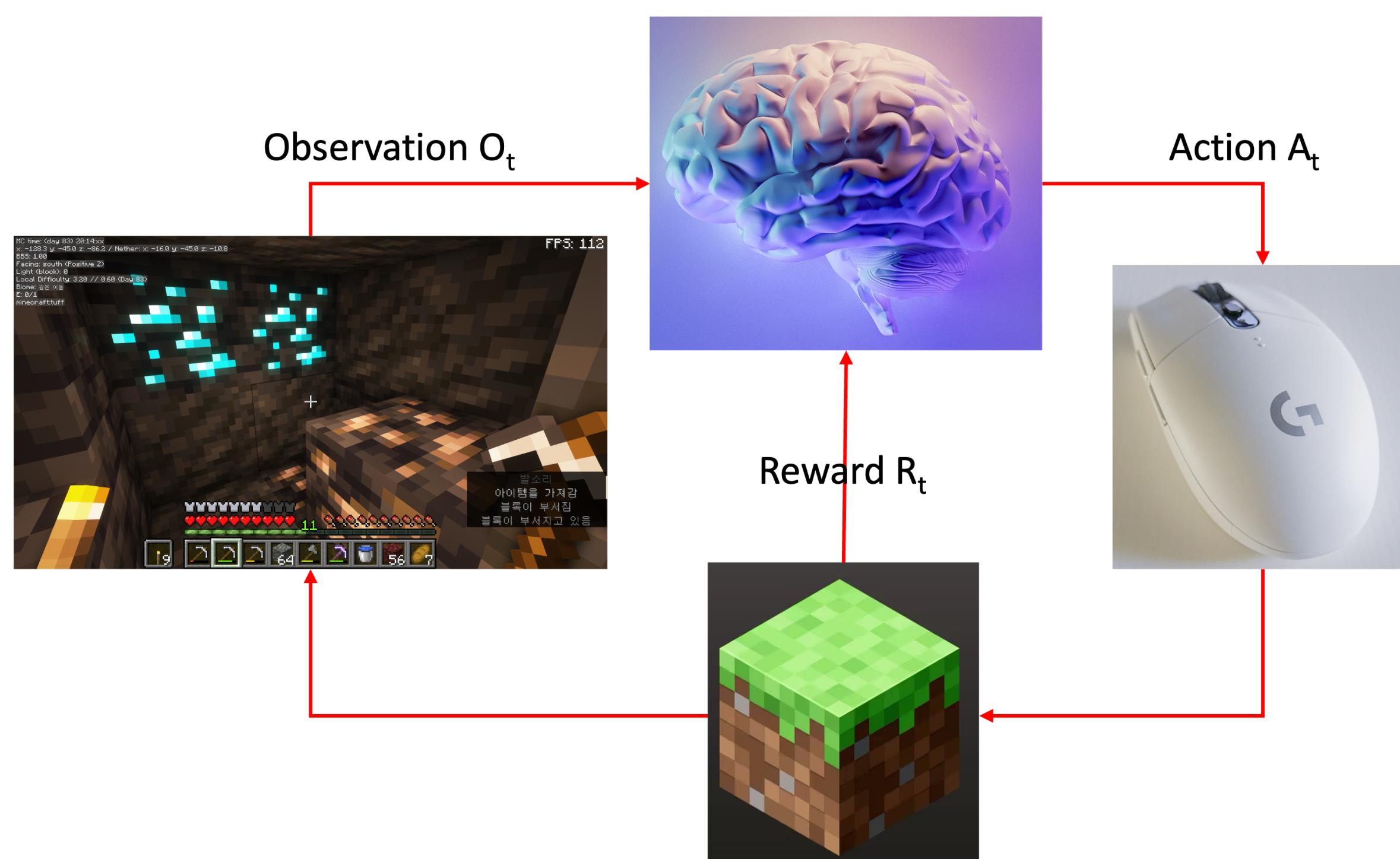


Figure 1: 강화학습의 학습 과정. 에이전트는 Observation O_t 를 받아 Policy π 에 따라 Action A_t 를 선택하고, Environment는 Reward R_t 와 다음 Observation O_{t+1} 을 반환한다.

강화학습 에이전트가 학습하기 위해 수집하는 정보가 에이전트의 행동에 종속적인 경향이 있어, 학습이 어렵다. 이러한 문제를 해결하기 위해 다양한 강화학습 알고리즘이 제안되었다. 본 연구에서는 강화학습 알고리즘 중 하나인 DQN을 사용한다. DQN은 심층 신경망을 활용하여 어떤 상태 s 에서 어떤 행동 a 를 취했을 때의 가치 $Q(s, a)$ 를 추정한다. 이러한 추정을 통해 에이전트는 가치가 높은 행동을 선택하도록 학습한다. DQN은 다음과 같은 순식 함수를 최소화하는 방향으로 학습된다.

$$L(\theta) = \mathbb{E}_{s,a,r,s' \sim D} [(r + \gamma \max_{a'} Q(s', a'; \theta^-) - Q(s, a; \theta))^2] \quad (1)$$

이것을 TD 오차라고 하며, 학습이 진행됨에 따라 일정 주기로 타겟 네트워크의 파라미터 θ^- 를 현재 네트워크의 파라미터 θ 로 업데이트한다. 따라서 학습 타겟이 움직이는 특징이 있으며, 이는 학습을 불안정하게 만든다.

연구 방법

Environment

Minecraft를 인간이 아닌 프로그램으로 조작하기 위해서 "모드"라는 기능을 활용하였다. 모드를 이용하면 게임을 수정할 수 있다. 이를 통해 파이썬으로 작성된 에이전트가 게임을 조작할 수 있다.

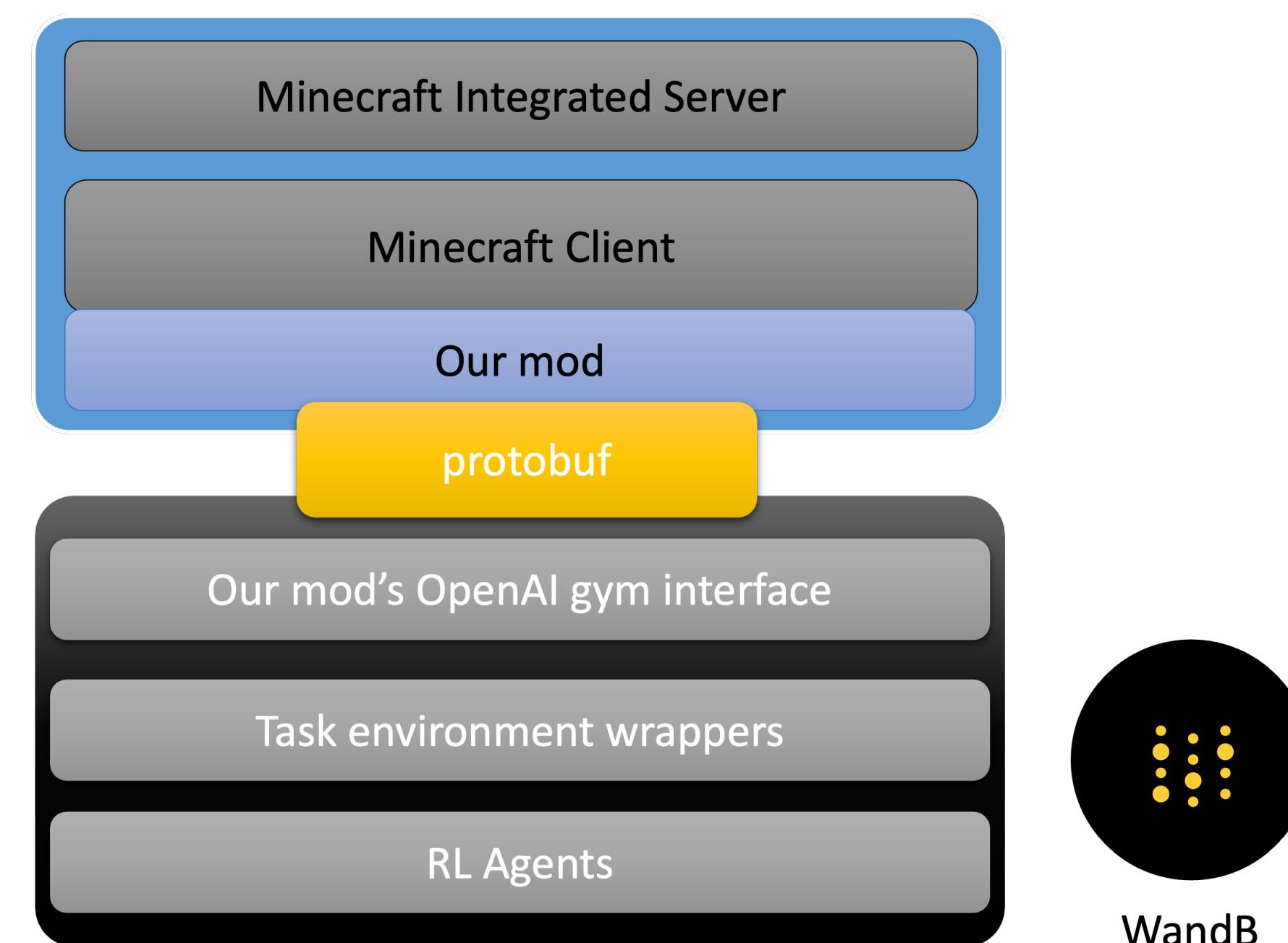


Figure 2: 환경을 나타내는 그림. 강화학습 에이전트가 env.reset이나 step을 호출하면, 우리의 wrapper는 마인크래프트를 실행한 후, protobuf over TCP를 통해 우리의 마인크래프트 모드와 통신한다. 마인크래프트 모드는 플레이어를 조작하고, 관측 공간을 캡처하여 wrapper에게 전달한다. wrapper는 관측 공간을 처리하여 강화학습 에이전트에게 전달한다.

태스크 설계

실험은 4개의 태스크와 3개의 모델을 이용하여 진행하였다. 먼저 태스크는 다음과 같다.

1. 한 마리의 허스크를 피해 도망가기
2. 여러 마리의 허스크를 피해 도망가기
3. Darkness 상태 이상이 걸린 상태에서 한 마리의 허스크를 피해 도망가기

모델 설계

시각 정보를 이용하는 에이전트

114 × 64 × 3의 이미지를 입력으로 받는 CNN을 이용하였다. CNN의 구조는 다음과 같다. (CNN그림인데 구조 확정이 안남)

소리와 방향 정보를 이용하는 에이전트

우리의 환경에서는 에이전트에 대한 음원의 상대 좌표와 종류를 알아낼 수 있다. 상대 좌표 벡터를 정규화하고, 현재 에이전트가 바라보고 있는 yaw값은 trigonometric encoding하여 입력했다. 완전한 평지 지형에서 테스트하므로 y좌표는 변하지 않아 항상 차이가 0이므로 입력에서 제외하였다.

멀티모달 에이전트

위의 시각 정보와 소리 방향 정보를 모두 이용하는 신경망을 구현하였다.

시각 정보와 소리, 방향 정보를 모두 이용하는 에이전트

보상 설계

허스크를 피해 도망가는 태스크의 경우, 매 틱마다 살아있을 경우 0.5의 보상, 죽었을 경우 -1의 보상을 부여하였다. 동물 찾아가기 태스크의 경우, 매 틱마다 지정한 동물이 있는 장소에 도달했을 경우 1의 보상, 그렇지 않을 경우 -0.05의 보상을 부여하였다.

연구 결과

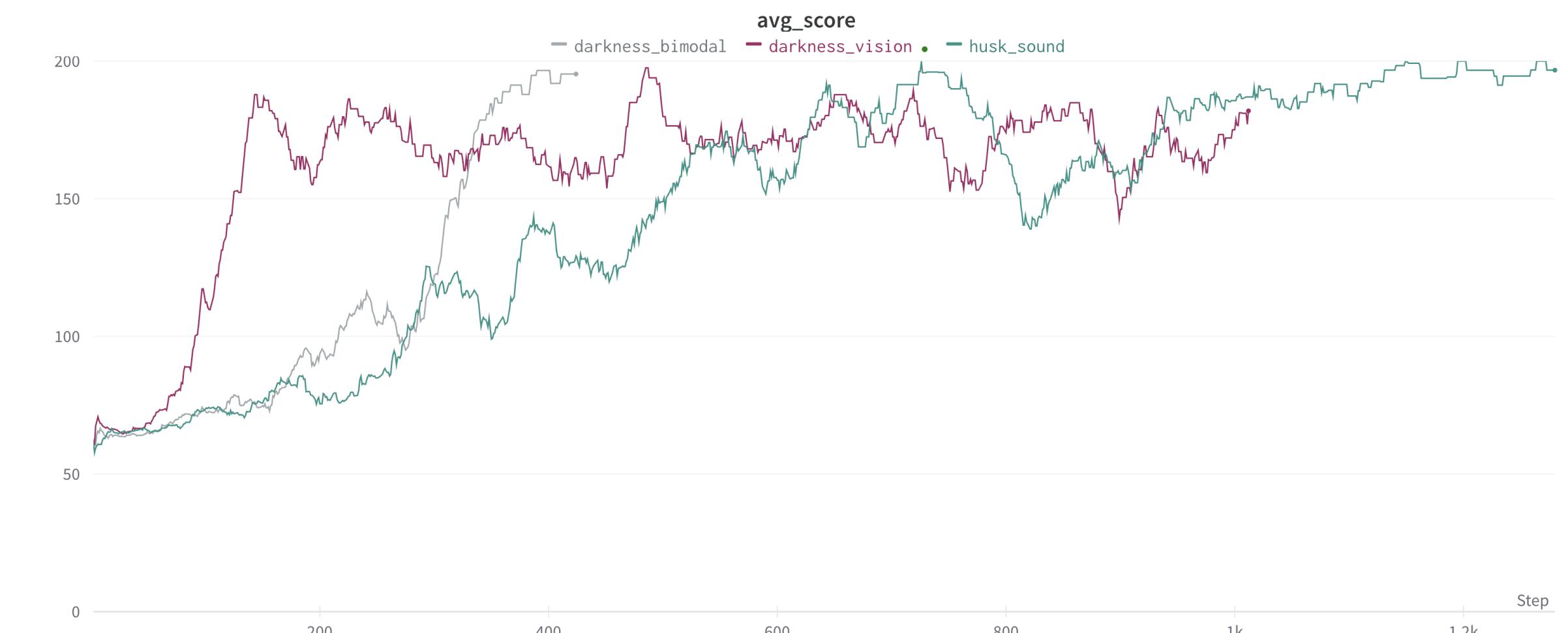


Figure 3: 디바이스 드라이버가 메모장 실행을 차단한 모습. 메모장 프로세스는 생성되지 않으며, 메모장 실행이 차단되었다는 로그가 출력된다.

연구 결과 소스 코드 저장소

<https://github.com/KYHSGeekCode/MinecraftRL>에서 본 포스터의 소스 코드를 탐색할 수 있다.

한계점 및 논의

HTTPS

시사점

이 연구는 타 분야에 비해 자료를 이해하기 어려운 DDK를 이용한 NT 레거시 윈도우 디바이스 개발이라는 분야에 유용한 예시를 제공하며, 이 연구를 통해 작성한 구현을 이용하여 비대면 강의 시 학생들의 집중을 도와주는 솔루션을 만드는 데 큰 도움을 준다는 데서 의의가 있다.

References

- [1] 이봉석, 『윈도우 디바이스 드라이버』, 한빛미디어, 2009.