[프로젝트 보고서] Unity & PPO 기반의 Squad Game 에이전트 전술 학습

작성자: A72077 이혜원

1. 개요
본 보고서는 Unity ML-Agents(강화학습)를 활용하여 '스쿼드 버스터즈' 스타일의 게임에서 스쿼드 유닛을 제어하는 전투 AI 에이전트를 개발한 과정을 기술합니다.
프로젝트의 목표는 플레이어의 조작(이동)에 따라 동적으로 변화하는 환경 속에서, AI 스쿼드가 스스로 판단하여 생존하고 적을 처치하며 아이템을 획득하도록 학습시키는 것입니다.
현재 단계는 스쿼드 전체를 하나의 에이전트로 제어하는 **Single Agent** 방식으로 구현되었습니다.

- **주제:** Unity ML-Agents와 Python Gymnasium을 연동한 **실시간 액션 게임(Squad Game) 내 에이전트의 전투 및 성장 전략 최적화**
- **배경:**
    - 기존의 **FSM(Finite State Machine)** 기반 게임 AI는 정해진 규칙 내에서만 행동하여, 거리 조절(Kiting)이나 상황에 따른 유연한 타겟 변경 등 복합적인 전략을 구사하는 데 한계가 있음.
    - 현업 게임 클라이언트 개발 경험을 바탕으로, 실제 상용 게임(Squad Busters 모작) 환경을 구축하고 **강화학습(Reinforcement Learning)**을 적용하여 사람과 유사하거나 그 이상의 **Micro-Control(세밀한 조작)** 능력을 학습시키고자 함.

2. 학습 환경 (Reinforcement Environment)

- **기술 스택:** Unity (ML-Agents), C#
- **환경 특성:**
    - **부분적 관찰 (Partially Observable):** 에이전트(스쿼드)는 자신의 고정된 범위 내에 감지된 적과 아이템만 관찰할 수 있습니다.
    - **동적 환경 (Dynamic):** 에이전트의 환경 관찰은 유저가 키보드로 스쿼드의 위치를 실시간으로 이동시킴에 따라 동적으로 변경됩니다.
    - **에피소드 (Episodic):** 명확한 시작과 종료 조건(후술)을 가진 에피소드 단위로 학습이

진행됩니다.

3. 에이전트 설계 (Single Agent - 중앙 관리 방식)

하나의 Agent(두뇌)가 스쿼드 내 모든 개별 유닛에게 명령을 내려 전투를 수행하는 중앙 관리 방식입니다.

3.1. 상태 (State)

관찰된 환경 정보를 에이전트가 학습할 수 있도록 벡터로 변환하며, **원-핫 인코딩(One-hot encoding)**을 사용해 상태 정보를 저장합니다.

- **아군 스쿼드 정보:**
  - 유닛 활성화 여부, 유닛 종류, 유닛 레벨
  - 유닛 체력 (현재 체력 / 최대 체력)으로 정규화
- **감지된 적 정보:**
  - 적 활성화 여부, 적 종류, 적 레벨
  - 적 체력 (현재 체력 / 최대 체력)으로 정규화

3.2. 행동 (Action)

에이전트가 선택할 수 있는 행동은 **이산(Discrete Action)** 공간으로 정의했습니다. 스쿼드의 공격 타겟을 찾는 기준 2가지를 행동으로 지정했습니다.

- **Action 0:** 감지된 적 중 **최소 거리에 있는 타겟**을 찾아 공격
- **Action 1:** 감지된 적 중 **최소 체력을 가진 타겟**을 찾아 공격

3.3. 보상 (Reward)

에이전트가 최적의 전투 전략(생존, 적 처치, 아이템 획득)을 학습하도록 다음과 같이 보상 시스템을 설계했습니다.

- **부정적 보상 (Penalty):**
  - 아군 유닛이 사망했을 때: -0.1
- **긍정적 보상 (Reward):**
  - 적 캐릭터가 사망했을 때: +0.1
  - 코인을 얻었을 때: +0.01
  - 아이템을 얻었을 때: +0.01

3.4. 학습 알고리즘 및 하이퍼파라미터 (AIgorithm & Hyperparameters)

본 프로젝트는 Unity ML-Agents에서 제공하는 **PPO (Proximal Policy Optimization)** 알고리즘을 사용하여 학습을 진행했습니다. PPO는 기존의 Policy Gradient 방식에 비해

학습의 안정성이 높고 구현이 용이하여, 연속적인 의사결정이 필요한 게임 AI 학습에 적합합니다.

주요 하이퍼파라미터 설정은 다음과 같습니다:

- **Trainer**: PPO
- **Batch Size**: 64 (안정적인 업데이트를 위해 소규모 배치 사용)
- **Buffer Size**: 2048 (충분한 경험 데이터를 수집한 후 업데이트)
- **Learning Rate**: 3.0e-4 (0.0003, 초기 학습의 발산을 막기 위한 보수적 설정)
- **Beta (Entropy Regularization)**: 5.0e-3 (초기 탐험(Exploration)을 유도하기 위한 무작위성 부여)
- **Epsilon (Clip Range)**: 0.2 (정책이 급격하게 변하는 것을 방지)
- **Gamma (Discount Factor)**: 0.99 (미래 보상의 가치를 높게 평가하여 장기적 생존 유도)

3.5. 신경망 구조 (Network Architecture)

에이전트의 두뇌(Policy Network)는 **완전 연결 계층 (Fully Connected Layer, Dense Layer)** 구조로 설계되었습니다.

1. **Input Layer**: 관찰된 상태(State) 벡터 (Vector Observation)
2. **Hidden Layers**: 2개의 은닉층 (각 128 유닛), 활성화 함수로 Swish 또는 ReLU 사용
3. **Output Layer**: 행동(Action)의 확률을 출력하는 Softmax Layer (Discrete Action Space)

4. 에피소드 종료 조건 (Episode Termination)

하나의 학습 단위인 에피소드는 다음 두 가지 조건 중 하나가 충족되면 즉시 종료됩니다.
• **학습 실패:** 스쿼드 내 모든 아군 유닛이 사망.
• **학습 성공:** 스쿼드 범위 내 감지된 모든 적을 처치.

5. 학습 결과 및 분석 (TensorBoard)

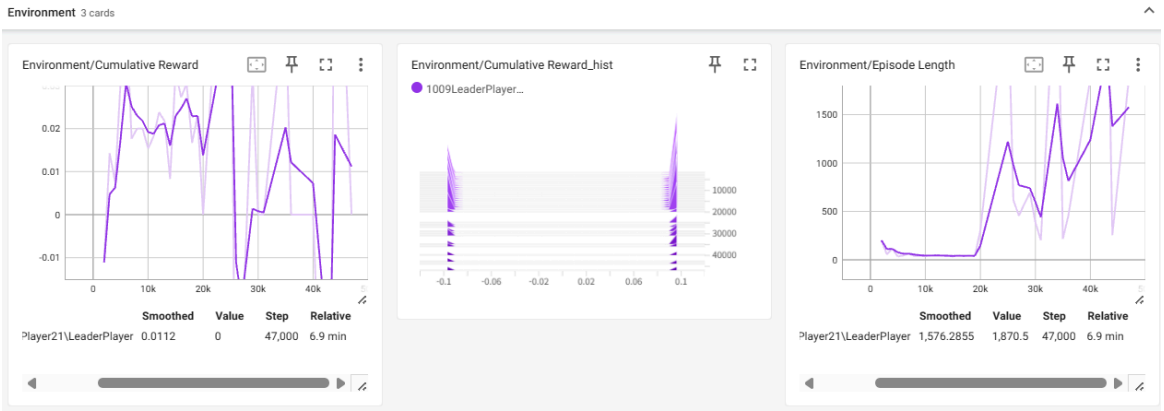
5.1. 가설 설정 및 검증 (Hypothesis)

- **가설 (Hypothesis):** 긍정적 보상(적 처치 +0.1)과 부정적 보상(사망 -0.1)이 대칭을 이루거나 긍정적 보상이 누적될 경우, 에이전트는 적극적인 전투(Hit-and-Run) 전략을 통해 이득을 극대화하는 방향으로 학습될 것이라 가정했습니다.

- **실제 결과 (Reality):** TensorBoard 분석 결과, 에이전트는 보상을 얻기 위한 '공격'보다는 페널티를 피하기 위한 '회피/대기'를 선택했습니다. 이는 강화학습에서 흔히 발생하는 **국소 최적해(Local Minima)** 문제로, 초기 탐험 단계에서 공격 시도 중 사망(Penalty)을 경험한 빈도가 높아, "아무것도 하지 않는 것이 가장 안전하다"는 잘못된 정책으로 수렴한 것으로 분석됩니다.

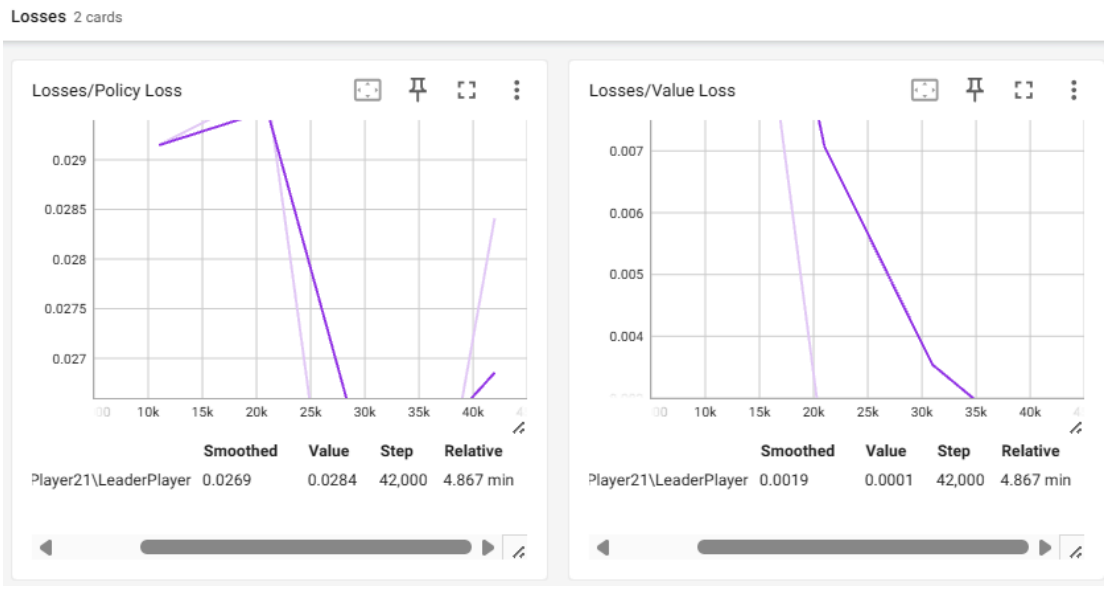
약 5만 Step까지의 초기 학습 경향성을 분석한 결과, 에이전트가 소극적 전략에 수렴하는 현상을 발견했습니다.

- 보상 및 에피소드 길이

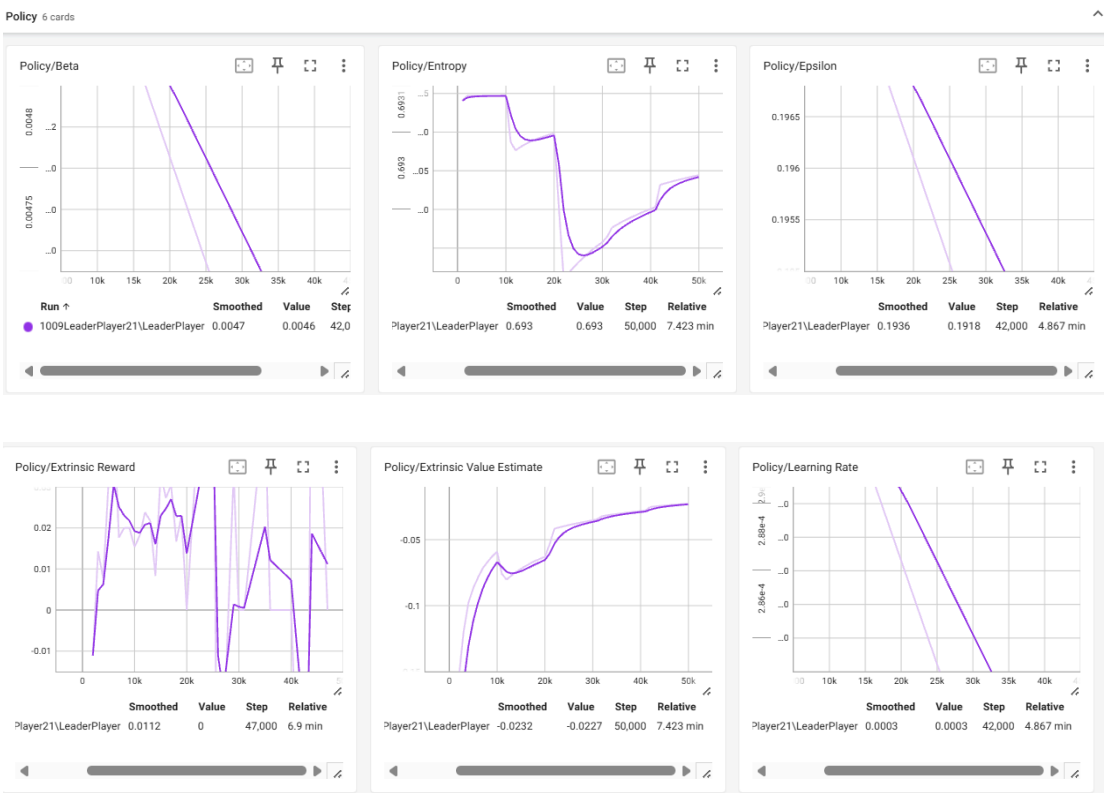


Environment/Cumulative Reward 그래프는 15k Step 부근까지 보상이 소폭 상승하다가 이후 0에 수렴하거나 음수(-0.1)로 하락합니다. 동시에 Environment/Episode Length는 20k Step 이후 급격히 증가합니다. 이는 콘솔 로그의 "No episode was completed" 기록과 일치하며, 에이전트가 성공(적 섬멸)도 실패(아군 전멸)도 하지 못하는 교착 상태에 빠져 에피소드를 완료하지 못하고 있음을 나타냅니다

- 손실 및 가치 추정
    - Loses


Losses 2 cards
Losses/Policy Loss
0.029
0.0285
0.028
0.0275
0.027
10k 15k 20k 25k 30k 35k 40k
Smoothed Value Step Relative
Player21\LeaderPlayer 0.0269 0.0284 42,000 4.867 min
Losses/Value Loss
0.007
0.006
0.005
0.004
10k 15k 20k 25k 30k 35k 40k
Smoothed Value Step Relative
Player21\LeaderPlayer 0.0019 0.0001 42,000 4.867 min

- Policy


Policy 6 cards
Policy/Beta
Run
Smoothed Value Step
1009LeaderPlayer21\LeaderPlayer 0.0047 0.0046 42,0
Policy/Entropy
0.693
Smoothed Value Step Relative
Player21\LeaderPlayer 0.693 0.693 50,000 7.423 min
Policy/Epsilon
0.1965
0.196
0.1955
Smoothed Value Step Relative
Player21\LeaderPlayer 0.1936 0.1918 42,000 4.867 min
Policy/Extrinsic Reward
0.02
0.01
0
-0.01
Smoothed Value Step Relative
Player21\LeaderPlayer 0.0112 0 47,000 6.9 min
Policy/Extrinsic Value Estimate
-0.05
-0.1
Smoothed Value Step Relative
Player21\LeaderPlayer -0.0232 -0.0227 50,000 7.423 min
Policy/Learning Rate
2.88e-4
2.86e-4
Smoothed Value Step Relative
Player21\LeaderPlayer 0.0003 0.0003 42,000 4.867 min

```
[INFO] LeaderPlayer. Step: 1000. Time Elapsed: 28.754 s. No episode was completed since last summary. Training.
[INFO] LeaderPlayer. Step: 2000. Time Elapsed: 36.995 s. Mean Reward: -0.011. Std of Reward: 0.099. Training.
[INFO] LeaderPlayer. Step: 3000. Time Elapsed: 45.235 s. Mean Reward: 0.014. Std of Reward: 0.099. Training.
[INFO] LeaderPlayer. Step: 4000. Time Elapsed: 53.669 s. Mean Reward: 0.008. Std of Reward: 0.100. Training.
[INFO] LeaderPlayer. Step: 5000. Time Elapsed: 62.850 s. Mean Reward: 0.031. Std of Reward: 0.095. Training.
[INFO] LeaderPlayer. Step: 6000. Time Elapsed: 72.149 s. Mean Reward: 0.047. Std of Reward: 0.088. Training.
[INFO] LeaderPlayer. Step: 7000. Time Elapsed: 80.951 s. Mean Reward: 0.018. Std of Reward: 0.098. Training.
[INFO] LeaderPlayer. Step: 8000. Time Elapsed: 90.024 s. Mean Reward: 0.020. Std of Reward: 0.098. Training.
[INFO] LeaderPlayer. Step: 9000. Time Elapsed: 99.352 s. Mean Reward: 0.020. Std of Reward: 0.098. Training.
[INFO] LeaderPlayer. Step: 10000. Time Elapsed: 108.246 s. Mean Reward: 0.015. Std of Reward: 0.099. Training.
[INFO] LeaderPlayer. Step: 11000. Time Elapsed: 118.538 s. Mean Reward: 0.018. Std of Reward: 0.098. Training.
[INFO] LeaderPlayer. Step: 12000. Time Elapsed: 127.873 s. Mean Reward: 0.024. Std of Reward: 0.097. Training.
[INFO] LeaderPlayer. Step: 13000. Time Elapsed: 136.634 s. Mean Reward: 0.022. Std of Reward: 0.098. Training.
[INFO] LeaderPlayer. Step: 14000. Time Elapsed: 145.806 s. Mean Reward: 0.008. Std of Reward: 0.100. Training.
[INFO] LeaderPlayer. Step: 15000. Time Elapsed: 154.813 s. Mean Reward: 0.033. Std of Reward: 0.094. Training.
[INFO] LeaderPlayer. Step: 16000. Time Elapsed: 164.251 s. Mean Reward: 0.027. Std of Reward: 0.096. Training.
[INFO] LeaderPlayer. Step: 17000. Time Elapsed: 173.682 s. Mean Reward: 0.030. Std of Reward: 0.095. Training.
[INFO] LeaderPlayer. Step: 18000. Time Elapsed: 183.435 s. Mean Reward: 0.017. Std of Reward: 0.099. Training.
[INFO] LeaderPlayer. Step: 19000. Time Elapsed: 192.725 s. Mean Reward: 0.023. Std of Reward: 0.097. Training.
[INFO] LeaderPlayer. Step: 20000. Time Elapsed: 201.721 s. Mean Reward: 0.000. Std of Reward: 0.100. Training.
[INFO] LeaderPlayer. Step: 21000. Time Elapsed: 212.650 s. No episode was completed since last summary. Training.
[INFO] LeaderPlayer. Step: 22000. Time Elapsed: 222.101 s. No episode was completed since last summary. Training.
[INFO] LeaderPlayer. Step: 23000. Time Elapsed: 233.369 s. No episode was completed since last summary. Training.
[INFO] LeaderPlayer. Step: 24000. Time Elapsed: 244.866 s. No episode was completed since last summary. Training.
[INFO] LeaderPlayer. Step: 25000. Time Elapsed: 255.337 s. Mean Reward: 0.100. Std of Reward: 0.000. Training.
[INFO] LeaderPlayer. Step: 26000. Time Elapsed: 264.875 s. Mean Reward: -0.100. Std of Reward: 0.000. Training.
[INFO] LeaderPlayer. Step: 27000. Time Elapsed: 274.898 s. Mean Reward: -0.033. Std of Reward: 0.094. Training.
[INFO] LeaderPlayer. Step: 28000. Time Elapsed: 284.281 s. No episode was completed since last summary. Training.
[INFO] LeaderPlayer. Step: 29000. Time Elapsed: 293.582 s. Mean Reward: 0.033. Std of Reward: 0.094. Training.
[INFO] LeaderPlayer. Step: 30000. Time Elapsed: 302.916 s. Mean Reward: 0.000. Std of Reward: 0.100. Training.
[INFO] LeaderPlayer. Step: 31000. Time Elapsed: 314.411 s. Mean Reward: 0.000. Std of Reward: 0.100. Training.
[INFO] LeaderPlayer. Step: 32000. Time Elapsed: 323.839 s. No episode was completed since last summary. Training.
[INFO] LeaderPlayer. Step: 33000. Time Elapsed: 333.146 s. No episode was completed since last summary. Training.
[INFO] LeaderPlayer. Step: 34000. Time Elapsed: 340.970 s. No episode was completed since last summary. Training.
[INFO] LeaderPlayer. Step: 35000. Time Elapsed: 350.041 s. Mean Reward: 0.050. Std of Reward: 0.087. Training.
[INFO] LeaderPlayer. Step: 36000. Time Elapsed: 358.505 s. Mean Reward: 0.000. Std of Reward: 0.100. Training.
[INFO] LeaderPlayer. Step: 37000. Time Elapsed: 367.481 s. No episode was completed since last summary. Training.
[INFO] LeaderPlayer. Step: 38000. Time Elapsed: 375.450 s. No episode was completed since last summary. Training.
```

반면, Losses/Value Loss와 Policy/Extrinsic Value Estimate 그래프는 0에 가깝게 안정적으로 수렴하고 있습니다. 이는 모델 자체는 학습이 잘 되고 있음을 의미하며, 에이전트가 "현재 상태의 미래 기대 보상이 0에 가깝다"고 정확하게 예측하게 되었음을 뜻합니다.

- 종합 해석

  두 현상을 종합하면, 에이전트는 '아군 유닛 사망 시 받는 큰 페널티(-0.1)'를 회피하는 방향으로 우선 학습되었습니다. 즉, 적을 공격(Reward +0.1)하다가 죽는(Penalty -0.1) 위험을 감수하기보다, 아무것도 하지 않고 생존하며 보상을 0에 가깝게 유지하는 '소극적 전략'에 수렴한 것으로 보입니다. Policy/Entropy(무작위성)가 25k Step에서 하락(정책 확신)했다가 보상이 하락하자 다시 상승(재탐색)하는 모습도 이러한 교착 상태를 뒷받침합니다.

6. 결론 및 향후 계획

현재의 Single Agent 모델을 통해 강화학습 환경을 구축하고 에이전트의 초기 학습 기반을 마련했습니다.

다만, 위 분석에서 확인된 **'소극적 전략 수렴 문제'를 해결**하고, 각 유닛이 독립적으로 판단하여 더 정교한 협력 플레이를 구현하기 위해 향후 스쿼드 내 각 유닛이 개별 Agent가 되는 **Multi-Agent** 방식을 도입할 계획입니다. ML-Agents의 `SimpleMultiAgentGroup` 기능을 활용

하여 유닛 간 보상을 공유하게 함으로써, 단순 생존이 아닌 '협력을 통한 승리'를 학습하는 에이전트를 개발하는 것을 목표로 하고 있습니다.