Journal of the KIMST, Vol. 26, No. 4, pp. 352-360, 2023
ISSN 1598-9127(Print) · 2636-0640(Online)
DOI https://doi.org/10.9766/KIMST.2023.26.4.352

Research Paper 정보·통신 부문

커리큘럼을 이용한 투서클 기반 항공기 헤드온 공중 교전 강화학습 기법 연구

황인수[1] · 배정호[*,1]

[1] 국방과학연구소 국방첨단과학기술연구원 인공지능자율센터

Two Circle-based Aircraft Head-on Reinforcement Learning Technique using Curriculum

Insu Hwang[1] · Jungho Bae[*,1]

[1] *Institute of Defense Advanced Technology Research Interlligence & Artificial Intelligence Autonomous Center, Agency for Defense Development, Korea*

(Received 6 April 2023 / Revised 9 June 2023 / Accepted 4 July 2023)

Abstract

Recently, AI pilots using reinforcement learning are developing to a level that is more flexible than rule-based methods and can replace human pilots. In this paper, a curriculum was used to help head-on combat with reinforcement learning. It is not easy to learn head-on with a reinforcement learning method without a curriculum, but in this paper, through the two circle-based head-on air combat learning technique, ownship gradually increase the difficulty and become good at head-on combat. On the two-circle, the ATA angle between the ownship and target gradually increased and the AA angle gradually decreased while learning was conducted. By performing reinforcement learning with and w/o curriculum, it was engaged with the rule-based model. And as the win ratio of the curriculum based model increased to close to 100 %, it was confirmed that the performance was superior.

Key Words : Air-to-air Combat(공대공 교전), Reinforcement Learning(강화학습), Head-on(헤드온), Two-circle(투서클)

1. 서 론

무인기 기술은 최근 몇 년 동안 급격하게 성장해오고 있다. 이러한 발전은 기존에 사람이 수행해왔던 작업들을 자동화하는 방향으로 이루어졌다. 그리고 이러한 무인기 기술은 인공지능 기술과 함께 발전하고 있다.

특히 강화학습을 활용한 무인기 조종사는 일반적인 규칙 기반의 방식보다 더 유연하고 또한 인간의 판단 능력을 대체할 수 있는 수준까지 발전하고 있다[1].

이러한 상황에서 미국방고등기술연구소(DARPA)에서는 2019년 부터 ACE(Air Combat Evolutio) 프로그램을 열어서 인간 조종사를 대체하는 AI 조종사 개발을

* Corresponding author, E-mail: deawith@gmail.com

진행하고 있다[2]. AI 조종사는 완전히 인간 조종사를 대체하기 보다는 우선 유무인 항공기가 협업하는 방향으로 연구가 진행되고 있다[3]. 효과적인 임무 수행을 위하여 무인기 스스로 기동을 결정하고 자율적으로 행동을 하기위해서는 많은 연구가 필요한다.

일반적으로 유인기 간 공중 교전시 마주보는 상황에서는 교전이 잘 일어나지 않는다. 왜냐하면 아군기가 피격당할 위험성도 있고, 상대기를 타격 후에 파편등이 날아올 수 있어서 아군기가 위험하기 때문이다. 유인기 조종사는 우위를 점할 수 있는 꼬리물기가 가능하도록 다양한 기동을 수행하는 편이지만 조종사 간의 실력이 비슷한 경우에는 한쪽이 우위 점하기가 쉽지 않다. 하지만 무인기의 경우에는 유인기보다 위험에 대한 인지를 적게하게 되므로 오히려 교전의 빠른 종료를 위하여 헤드온 상황에서 교전을 펼치는 것이 효과적일 수 있다. 본 논문에서는 무인기가 헤드온 상황에서 교전을 통하여 승리하는 학습기법을 제시한다.

기존에는 강화학습으로 아군기와 상대기가 마주보는 헤드온 지향을 잘할 수 있도록 아군기(에이전트)와 상대기의 각도와 자세는 고정하고, 둘 사이의 거리만 멀게 조정 하는 방식등의 커리큘럼을 사용하였다[4]. 하지만 이러한 방식을 사용한 경우에는 아군기가 다양한 조건을 경험하기 힘들고, 단순히 거리를 멀게 한다고 하여 난이도가 단계적으로 어려워지지 않게된다는 문제점이 있다. 거리를 고정하고 각도만 올릴 경우에도 난이도가 갑자기 기하급수적으로 변화되어서 학습이 잘 진행되지 않는다. 기존 항공기 공중 교전 학습의 문제점으로는 교전 규칙이 너무 간단하거나, 학습 상대인 대상이 검증이 잘 되지 않았다는 문제가 있다. 또한 기존에 커리큘럼 기반의 강화학습을 수행한 연구가 있었으나 이 경우에는 꼬리물기에만 특화되었다는 한계가 존재한다[5].

본 논문에서는 아군기와 상대기 각각의 턴서클을 활용한 투서클 기반의 헤드온을 잘 할 수 있는 학습기법을 제시하였다.

본 연구를 통하여 향후 전투기 조종사들에게 모의훈련을 할 수 있도록 시뮬레이션 환경을 구축할 예정이다. 모의 시뮬레이션을 통하여 실제로는 일어나기 어려운 조건에서의 훈련도 가능할 것으로 보인다. 추후에는 유무인 복합운용이 가능하도록 무인기의 완전 자율기동이 가능한 수준으로 2:1, 2:2 교전에 관련한 연구를 진행할 것이다.

2. 관련 연구

2.1 MDP(Markov Decision Process)

MDP는 의사 결정 과정을 확률과 그래프로 모델링 한 것으로 다음 스텝의 상태는 현재 상태와 현재 행동에만 영향을 받는다[6]. 학습 하고자 하는 에이전트는 S_t라는 상태에서 A_t라는 행동을 수행한다. 그러면 해당하는 환경에서 다음 상태에 해당하는 S_{t+1}와 그에 상응되는 리워드 R_{t+1}을 에이전트에 반환한다. MDP는 <S, A, P, R> 로 나타낼 수 있고, S는 상태(state), A는 행동(action), P는 현재 상태에서 어떤 행동을 취할 확률(policy), R은 보상함수(reward)를 의미한다. 강화학습은 MDP 환경에서 상태정보를 가지고 행동을 하였을 때, 행동의 성능에 대한 보상함수를 받는 과정이다. 본 논문에서는 MDP 환경을 가정하여 시뮬레이션을 수행하였다.

2.2 SAC(Soft Actor-Critic) & MaxEnt RL

SAC는 일반적으로 결정론적인 강화학습의 목적 함수에 엔트로피를 최대화 하는 팀을 추가한 것이 특징이다[7]. 엔트로피 함수를 추가하였기 때문에 SAC 알고리즘은 확률적으로 추정하는 것이 가능해지고 같은 상태 정보를 받더라도 다양한 행동을 수행 해볼 수 있다는 장점을 가진다. 본 논문에서는 SAC 알고리즘을 활용하여 에이전트가 다양한 환경을 탐사 하고 노이즈 등에 강건하게 되었다.

2.3 LSTM(Long Short-Term Memory)

일반적인 네트워크는 단지 현재 시점에 대한 상태 정보만 가지고 있는데 항공기의 기동에 대한 학습을 위해서는 과거 시점에 대한 상태 정보가 함께 있어야 분석이 가능하다[8]. 본 논문에서는 LSTM을 통해서 이전 상태값 중에서 중요한 상태값을 기억해서 활용할 수 있도록 하였다.

2.4 커리큘럼 러닝

강화학습을 수행할 때 학습 성능을 좋게 만들기 위하여 커리큘럼 러닝을 적용하였다[9]. 항공기 강화학습에서 상태와 액션은 무한대에 가깝기 때문에 상대기를 대상으로 유의미한 기동을 일반적인 강화학습으로만 학습하는 것은 학습이 잘 되지 않는다. 본 논문에서는 커리큘럼 러닝을 통하여 초기 상태 공간 탐색 범위를 제한하여 항공기가 의미 있는 기동을 할 수 있도록

유도 하였다. 그리고 단계적으로 상태 공간을 넓혀서 학습 난이도가 갑자기 급격하게 올라가는 것을 방지하고 단계적 상승하도록 하였다.

3. 투서클 기반 항공기 헤드온 공중 교전 학습 기법

3.1 강화학습을 이용한 투서클 기반 항공기 공중 교전 학습 flow

Figure 1은 강화학습을 적용한 투서클 기반 항공기 헤드온 공중 교전 학습 기법 전체적인 흐름이다. 투서클 기반 커리큘럼으로 초기상태와 종료조건을 설정한다. Fig. 2는 두 항공기 사이의 기하학적 위치에 대한 각도 정의이다. ATA(Antenna Train Angle)는 아군기의 주축(진행방향)과 레이더 LOS 사이의 각도이고, AA(Aspect Angle)는 상대기의 주축 반대 방향(꼬리 방향)과 아군기 레이더 LOS 사이의 각도이다.

3.2 교전 시뮬레이션 환경 세팅

본 논문에서는 F-16 전투기간 교전을 학습하였다. 이때 F-16의 물리 엔진은 오픈 소스 기반 비행 역학 모델인 JSBSim을 활용하였다[10,11]. MDP 환경에서 JSBSim 모델의 출력 값은 시뮬레이션 시간, 항공기 HP(Health Point), 위치, 자세, 속도, 가속도, 과거 조종값, 항공기 정보 등을 제공한다. 조종값은 스로틀(throttle), 롤(roll), 피치(pitch), 요(yaw) 값이고 물리 엔진은 60 Hz로 결과 값을 제공하지만 본 논문의 실험에서는 10 Hz의 데이터만 사용하였다. 전투기간 교전 시 피해 평가는 기총을 고려한다[11]. 교전은 체력(HP)이 0이 되어 격추당하거나, 고도 1,000 ft 이하이면 추락되었다고 생각하고, 300초가 지나면 끝이 나도록 하였다. 적기를 격추시키거나 추락시키거나 300초가 지났을 때 아군기의 체력(HP)이 더 많은 경우 이기고, 반대의 경우에는 패배, 300초 이후에 체력(HP)이 동일하면 무승부로 정의한다[5].

3.3 리워드 함수 설계

강화학습을 위한 리워드 함수 설계는 다음과 같다[5].

3.3.1 격추

상대기를 격추시키면 500점을 얻고, 상대기에 격추당하면 -500점을 받는다.

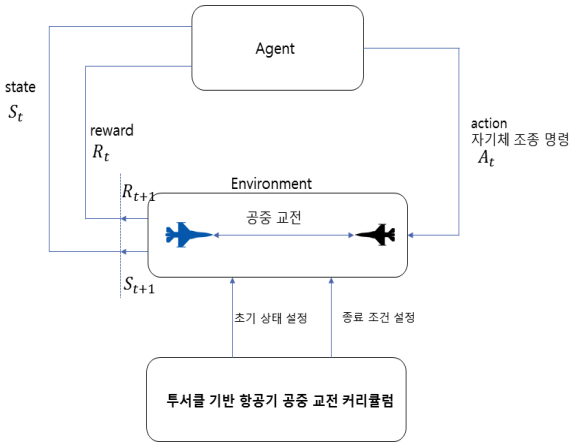

Fig. 1. Two circle-based aircraft head-on reinforcement learning method

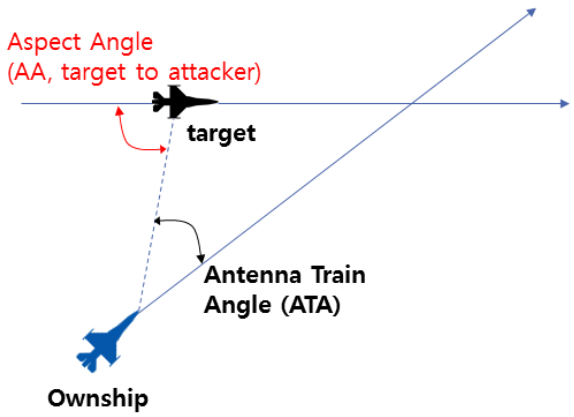

Fig. 2. Defining aircraft position geometry(ATA, AA)

3.3.2 WEZ(Weapon Engagement Zone)

상대기를 WEZ 내로 위치하거나, 가까워지도록 유도하는 리워드이다. WEZ 관련 리워드는 실제로 타격이 있을 때의 리워드와 WEZ에 가깝도록 유도하는 리워드 함수를 설계하였다. 상대기가 아군기의 공격 영역 내에 있을 때 가까우면 더 많은 리워드를 얻도록 설계하였다. 그리고 아군기가 상대기의 공격 영역 내에 있을 경우에는 도망갈 수 있도록 리워드를 설계하였다.

3.3.3 우위 선점

아군기가 상대기 대비하여 우위를 선점할 수 있도록 만들어주는 리워드이다. 희박한 리워드와 조밀한

리워드로 이루어진다. 희박한 리워드는 아군기가 상대기의 후미쪽에 제어 영역내에 유지하면 받는 리워드이다. 조밀한 리워드는 상대기의 꼬리물기를 잘 할 수 있도록 하는 리워드와 ATA와 AA의 감소율을 작게 하는 것이 목적인 리워드를 설계하였다.

3.3.4 추락

고도 1,000 ft 이하로 내려간 경우에는 추락한 것으로 설정하였고, 아군기가 추락한 경우에 1000점의 패널티를 받고, 상대기가 추락한 경우에는 10점의 리워드를 받는다.

3.4 강화학습 네트워크 구조

강화학습을 위한 네트워크 구조는 아래 Fig. 3과 같다. 본 논문에서는 LSTM을 적용하였다[8]. 그 이유는 일반적인 네트워크에 입력되는 상태값들은 현재 상태에 대한 정보만을 가지고 있으므로, 항공기 기동 패턴 분석을 위해서는 과거의 기동 패턴에 대한 상태값들을 분석할 수 있어야 하기 때문이다. LSTM을 통하여 이전 상태값 중에 중요한 값들을 기억하도록 하였다. LSTM 외에도 해당 시점에서의 상태 정보를 활용하기 위하여 스킵 커넥션(skip connection)을 사용하였다. 엑터(Actor) 영역에서는 스로틀, 롤, 피치, 요 4개의 값이 결과값으로 출력되어서 항공기 기동에 필요한 값들이 출력된다. 크리틱(Critic) 영역에서는 엑터 영역에서의 결과 값을 판단하는 퀄리티(Quality) 값이 1개로 출력된다.

3.5 투서클 기반의 헤드온 커리큘럼 개요도

Figure 4는 투서클 기반의 헤드온 커리큘럼을 학습할 때 아군기와 상대기의 초기 시작 상태를 도식화한 것이다. 파란색 항공기는 아군기이고 빨간색 항공기는 상대기이다. 일반적인 4세대 전투기는 최대기동을 수행할 경우 약 3,000 ft 정도의 선회 반경으로 회전 할 수 있다. 이때 약 6,000 ft 정도의 턴서클을 형성하게 된다. Fig. 2의 그림에 정의 대로 하면 그림에서 옅은 회색 글씨로 표시되어 있듯이 ATA가 0°이고, AA가 180°인 아군기와 상대기가 마주보고 있는 상황에서 학습을 시작한다. 각 항공기가 턴서클을 그리면서 서로 마주보면서 헤드온을 할 수 있도록 하기 위해서는 투서클이 필요하다. 결과적으로 지름의 합이 12,000 ft에 수렴하는 투서클 상에서 기동하도록 설정되었다. 두 항공기가 가장 멀리 떨어져 있을때가 ATA가 90°이고, AA가 90°인 상황으로 그림에서 짙은 회색 글씨로 표시된 것과 같은 상태이다. 학습이 종료되기 직전에는 검은색 글씨로 표시된 대로 ATA가 180°이고, AA가 0°인 서로 반대로 마주보는 상황에서 학습이 종료된다.

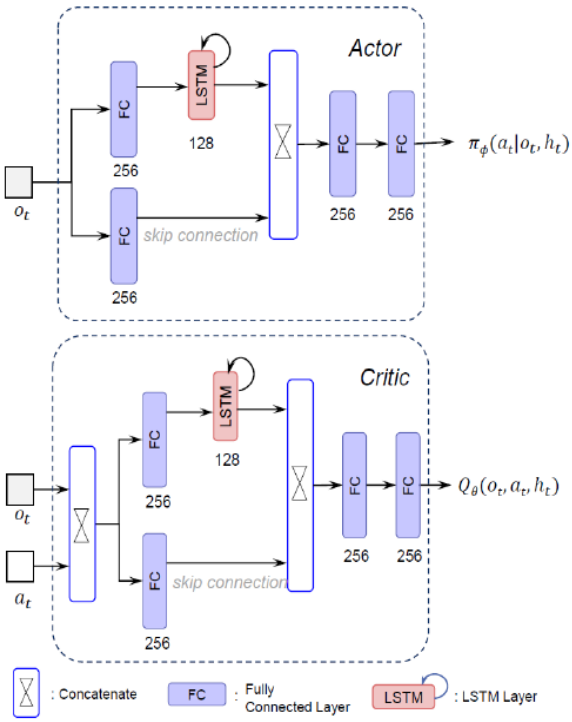

Fig. 3. Deep neural network structure

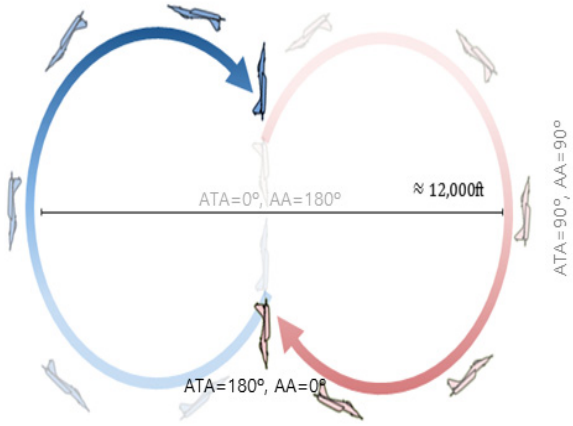

Fig. 4. Two circle-based aircraft head-on curriculum

즉 아군기와 상대기가 마주보는 상황, 즉 ATA가 0°, AA가 180°에서 시작하여 아군기와 상대기가 서로 반대로 마주보고 있는 ATA가 180°, AA가 0°인 상황까지 학습이 되도록 알고리즘을 설계하였다. 즉 난이도가 가장 낮은 항공기가 마주보는 상태에서부터 난이도가 가장 높은 항공기가 서로 반대로 마주보는 상황까지 모델을 훈련하여 난이도를 단계적으로 향상시켰다. 이렇게 하여 항공기의 헤드온 교전 능력을 향상시켰다.

3.6 투서클 기반 항공기 헤드온 지향 커리큘럼 알고리즘

Figure 5는 강화학습을 적용한 투서클 기반 항공기 헤드온 지향 커리큘럼 알고리즘 이다. 처음에 α값이 0°부터 상승하면서 ATA가 0°부터 상승하고, AA가 180°부터 작아지면서 Fig. 4에 투서클이 붙어 있는 상황에서 턴서클 위의 아군기와 상대기의 위치가 정해진다. 아군기와 상대기 사이의 거리 r은 수식 (1)과 같이 나타낼 수 있다. α값이 0°일 때는 아군기와 상대기 사이의 거리는 최소값이 되고 α값이 90°일때는 가장 거리가 멀어지게 된다. α값이 0°일 때 다만 아군기와 상대기의 거리가 너무 근접한 것을 막기 위하여 랜덤으로 3,000 ft ~ 6,000 ft의 최소 이격 거리를 정하였다.

$$r = 2*6000ft*\cos(\frac{\pi}{2}-\alpha) + random(3000 \sim 6000ft) \quad (1)$$

Alogrithm Two Circle-based Head-on Curriculum Learning

```
for angle α ← 0° to 180°
    range r ← 2* 6,000ft * cos(90° - α) + random(3,000ft ~ 6,000ft)
    Set environment aircrafts with (ATA ← α, AA ← 180-α, r)
    if α ≤ 80° than
        Add environment termination condition ( ATA > 90° )
    else if α ≤ 140° than
        Add environment termination condition ( ATA > α + 10° )
    end if
    win_ratio ← 0%
    while win_ratio < 70%
        Collect samples D_k
        Update network with D_k using RL(Reinforcement Learning) algorithm
    end while
end for
```

Fig. 5. Two circle-based aircraft head-on curriculum algorithm

초반에 학습을 시작하고 얼마되지 않아서 α값이 80° 이하일 경우에는 ATA가 90°가 넘어가는 경우에는 상대기의 3/9라인 앞에 아군기가 위치하게 되는데, 아군기와 상대기가 이미 교차했다고 판단되어 학습이 종료된다. α값이 80°에서 140° 사이에서는 ATA > α + 10°인 경우에는 아군기와 상대기가 최단 턴서클을 돌지 못하였다고 판단되어 학습이 종료된다. 최단 턴서클을 돌지 못하였다는 것은 아군기가 의도하지 않은 이상한 방향으로 갈 수 있고, 이때 헤드온 학습이 이루어지지 않기 때문에 종료한다. 승률이 70 %를 넘을 때까지는 해당 각도 α에서 학습을 계속 수행한다.

거리와 각도가 결정되고 난 후에는 상세한 위치와 항공기의 자세를 결정한다. 다양한 환경에서 학습을 수행할 수 있도록 같은 ATA와 AA 각도에서도 지표면에 대하여 수평인지 혹은 상승/하강 상태인지에 따라서 아군기가 학습해야하는 난이도가 달라지게 된다. Fig. 6과 같이 전방, 상승, 하강 3가지중에 하나를 랜덤으로 정하게 된다. 또한 아군기와 상대기 모두 롤 각도를 0° ~ 180° 사이의 값으로 임의로 정하여 다양한 환경에서 학습이 가능하도록 한다.

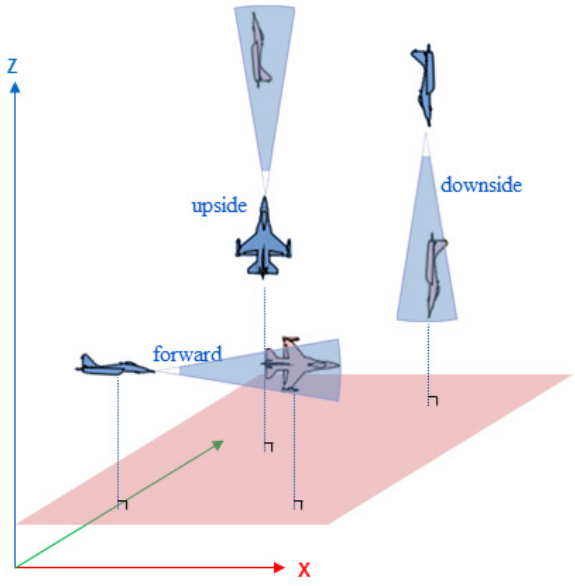

Fig. 6. Aircraft positioning

3.7 규칙기반 상대기 설계

상대기에 대한 기동은 학습을 하는 에이전트의 성능 비교의 이정표가 된다. 본 논문에서는 상대기를 규

칙 기반 전투 모델로 가정하였다[3]. 규칙 기반 전투 모델은 BFM(Basic Fighter Maneuvers)에 의해서 기동이 결정된다. 현재 위치 값과 항공기의 자세, 상대기에 대한 정보를 받게되면 항공기 기동에 필요한 조종 값들을 출력으로 알려준다.

4. 실험 방법

4.1 실험조건

학습에 사용한 기기의 스펙은 CPU 인텔 코어 i9-7900X(10 cores, 3.3 GHz) 프로세서와 GPU로는 TITAN Volta를 사용하였으며 한 에이전트 당 30만번까지 학습 업데이트를 하였다. 학습을 위한 환경 설정 값은 Table 1과 같다.

Table 1. 학습 환경 설정 값

파라미터	값
Minibatch size	64*256
sequence length	64
Number of rollout workers	10
Replay buffer size	1×10^6
Discount factor	0.99
Optimizer	Adam
Optimizer settings	Actor/critic/entropy $\lambda = 3\times10^{-4}$

실험은 MDP 환경을 가정하여 아군기와 상대기는 상태값 관련하여 서로에 대한 모든 정보를 알고 있다고 가정하였다.

실험 결과는 투서클 기반 제안 학습 기법을 적용한 아군기와 규칙기반 모델을 적용한 상대기 사이에 교전 결과를 승/무/패로 나누어서 아래의 수식 (2)를 활용하여 승률을 구하였다. 승률을 계산 할 때 이기면 1점을 온전히 얻고, 무승부에서는 0.5점을 받도록 하였다.

$$WinRatio = \frac{n_w + 0.5 \times n_d}{n_w + n_d + n_l} \times 100 \quad (2)$$

$n_w =$ 승수(win), $n_d =$ 무승부$(draws)$, $n_l =$ 패수$(loss)$

아군기와 상대기 사이의 교전은 시작점이 랜덤 고도, 거리, 속도로 정해지고 최대 300초까지 수행된다. 그리고 중립(neutral) 상황에서 100번의 교전을 수행하여 승률을 계산하였다.

5. 실험 결과

5.1 투서클 기반 알고리즘 승률 비교

Figure 7 실험에서는 헤드온 상황에서 커리큘럼이 있는 상황과 없는 상황에서 학습 후 교전 승률을 비교한 결과이다. 총 30만번 업데이트 중에 1만 단위로 체크포인트 학습 파일을 가져와서 뉴트럴 상태에서 규칙기반 모델과 교전을 진행하여서 수식 2를 활용하여 승률을 계산하였다. Fig. 7과 같이 커리큘럼 기반의 헤드온 학습 모델은 점차 학습이 진행 될수록 승률이 거의 100프로까지 올라가는 것을 확인할 수 있었다. 하지만 커리큘럼이 없이 학습한 모델의 경우에는 학습이 진행되어도 승률이 오르지 않는 것을 확인하였다.

Figure 8 그래프는 교전이 끝났을 때 규칙기반 모델인 표적기에 대한 타격량(100점-잔여 체력(HP))을 커리큘럼 유무에 따른 학습모델을 비교한 것이다. 커리큘럼 기반의 학습 모델에서는 학습이 진행 될수록 점차 승리하는 횟수가 많아지면서 표적이 완전히 파괴되어 타격량이 100점이 되고 표적의 잔여 체력(HP)이 0이 되는 경우가 많아지는 것을 확인 할 수 있다. 30만번 업데이트가 된 시점에서는 승률이 100 % 가까이 되면서 교전이 끝난 후에는 표적 타격량이 100에 수렴한다.

Figure 9 그래프는 평균 교전 시간을 커리큘럼 유무에 따른 학습모델을 비교한 것이다. 평균 교전 시간은 대체적으로 커리큘럼 기반의 학습 모델에서 우하향하는 경향을 보였고, 이는 학습이 될수록 아군기가 빠르게 상대기를 제압하기 때문인 것으로 분석된다. 학습 초반에 100초가 넘는 교전 시간을 보이다가 학습이 거의 끝날 때 즈음에는 30초 정도의 교전 시간을 보인다. 반면 커리큘럼 없이 학습한 모델은 교전 시간이 학습이 진행 되면서 크게 감소하지 않는 것을 확인 할 수 있다. 교전 결과도 무승부로 많이 끝나기 때문에 교전 시간이 30만번 업데이트 이후에도 300초에 가까운 것을 확인 할 수 있다.

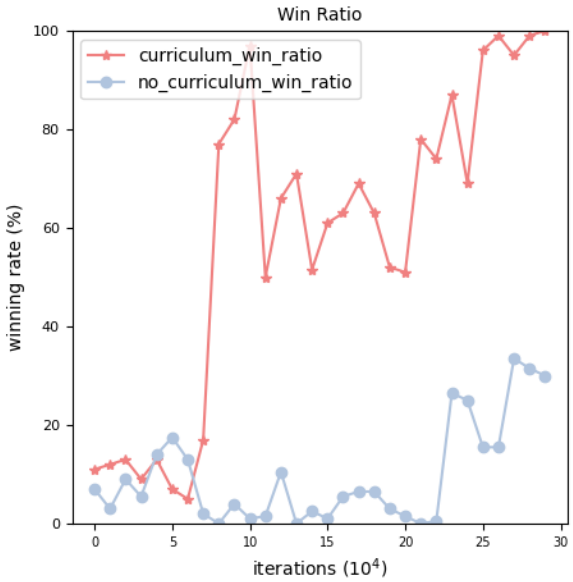

Fig. 7. W/ curriculum vs w/o curriculum win ratio comparision

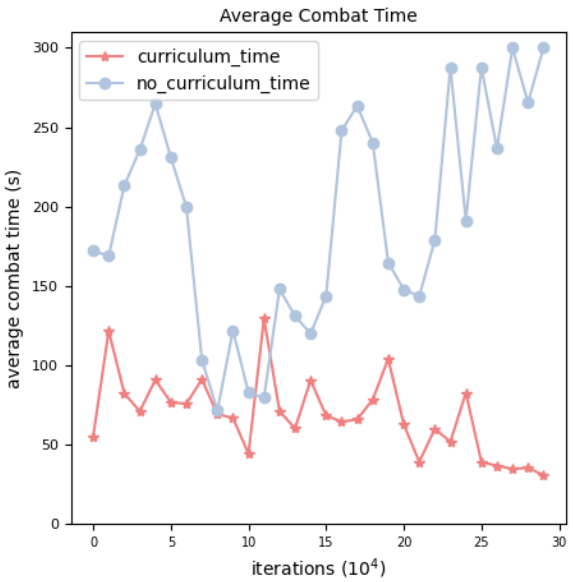

Fig. 9. W/ curriculum vs w/o curriculum average combat time comparision

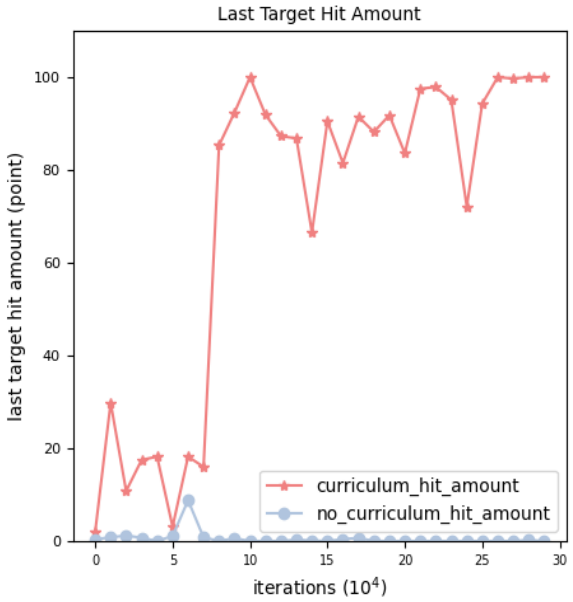

Fig. 8. W/ curriculum vs w/o curriculum average final target HP hit amount

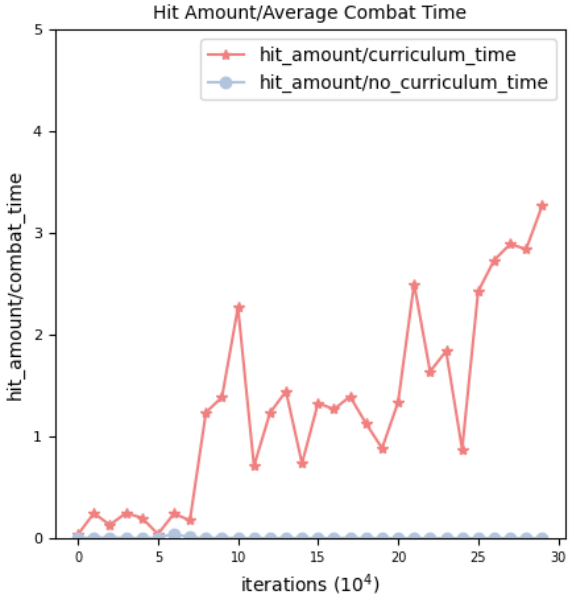

Fig. 10. W/ curriculum vs w/o curriculum average hit amount/combat time comparision

Figure 10 그래프는 Fig. 8에 나온 타격량(100-표적 최종 HP)를 평균 교전 시간으로 나눈 값을 커리큘럼 유무에 따라 비교한 것이다. 즉 단위 교전 시간당 타격량을 확인할 수 있는데 커리큘럼 기반의 학습시 단위 시간당 타격량이 학습에 따라서 점차 증가하는 것을 확인 할 수 있다. 반대로 커리큘럼이 없는 경우에

는 거의 이 값이 0에 수렴하였다.

위의 실험들을 통하여 커리큘럼이 없는 경우 강화학습을 단순히 오랫동안 학습한다고 하더라도 규칙기반 모델에 대비하여 우위를 점하지 못한다는 것을 확인 할 수 있었다. 항공기의 기동은 거의 무한대에 가깝기 때문에 단순히 강화학습을 수행하는 것만으로는 충분하지 않고, 본 논문에서 제시한 커리큘럼 기반의 강화학습으로 교전 모델 대비하여 효과를 확인하였다.

5.2 기동 분석

Figure 11, 12는 투서클 기반 헤드온 커리큘럼 모델(아군기)과 규칙기반 교전 모델(상대기) 사이에 교전을 SIMDIS 프로그램을 통해서 확인한 결과이다. 파란색 빔을 쏘는 쪽이 아군기이고 빨간색 빔을 쏘는 쪽이 상대기이다. 우선 Fig. 11은 위에서 본 뷰로 서로 반대 방향을 보고있는 중립 상태에서 시작하여 각자 턴서클을 빠르게 돌아서 헤드온 교전을 하는 것을 확인 할 수 있다. Fig. 11에서만 보면 위에서 보기 때문에 아군기와 상대기 중에 유리한 쪽을 판단하기는 힘들지만 Fig. 12를 보면 아군기는 본체는 빨간빔을 맞지 않으면서 상대기에게 파란빔을 맞추는 것을 확인 할 수 있다. Fig. 11, 12 그림에서 확인 할 수 있듯이 아군기 체력(HP)이 100점인데 반하여, 상대기 체력(HP)이 8.5점으로 거의 상대기를 제압한 것을 확인 할 수 있다. 본 화면에서 1~2초 뒤에 상대기의 체력이 0이 되어서 교전은 끝이 나게 되었다.

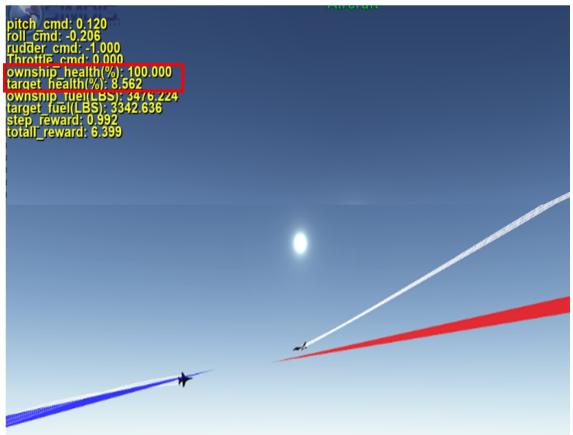

Fig. 11. Curriculum vs rule-base engagement maneuver view(side)

Fig. 12. Curriculum vs rule-base engagement maneuver view(up)

6. 결 론

본 논문에서는 투서클 기반의 항공기 헤드온 공중 교전 학습기법에 대하여 제시하였다. 강화학습을 활용하여 헤드온 공중 교전을 잘 할 수 있도록 학습하기 위해서는 단순히 아군기와 상대기 사이의 거리만을 조절해서는 학습 난이도가 단계적으로 상승되지 않고, 제시한 투서클 기반의 커리큘럼 방식을 활용해야 한다. 투서클 상의 아군기와 상대기에 ATA 각도를 0°에서 점차 증가시키고, AA 각도를 180°에서 점차 감소시키는 방향으로 학습을 수행하였다. 각도 별로 커리큘럼 모델이 규칙기반 모델을 70 % 이상의 승률로 이겨야 그 다음 각도로 넘어가도록 하였다. 지면에 대하여 아군기와 상대기가 전방, 상승, 하강 중에 상태 중에 랜덤으로 학습되게 하였고, 0° ~ 180° 사이의 롤 각도 및 항공기 고도에서 학습하게 하여 다양한 환경에서 좋은 성적을 낼 수 있도록 하였다. 학습된 모델을 규칙기반 모델과 100번 정도 교전하여 커리큘럼이 있을 경우와 커리큘럼 없는 강화학습만을 활용한 모델의 성능을 비교하여 커리큘럼 기반의 강화학습 모델의 성능이 우월함을 보였다. 30만번의 학습 업데이트 후 커리큘럼 기반의 모델은 거의 100 %의 승률을 기록하였다. 제안한 커리큘럼 모델은 단지 규칙기반 모델 뿐만이 아니라 뉴트럴 상황에서 인간 조종사와 교전해도 우수한 성능을 보일 것으로 예상된다. 그 이유로는 교범에서 위험한 상황을 최대한 피하기 위하

여 꼬리물기를 하도록 기술되어 있으므로, 인강 조종사는 헤드온 교전 상황을 피하고 꼬리물기를 선호 한다. 하지만 헤드온 상황은 기체 특성상 피하기 어렵기 때문에 제안한 모델이 인간 조종사에 대비하여 우수한 교전 성능을 보일 것으로 예상된다. 현실적인 교전을 위하여 추후 2:1, 2:2 학습 기법에 대한 연구도 진행할 예정이다.

후 기

이 논문은 2023년 정부의 재원으로 수행된 연구 결과임.

References

[1] Z. Wang, H. Li, H. Wu and Z. Wu, “Improving Maneuver Strategy in Air Combat by Alternate Freeze Games with a Deep Reinforcement Learning Algorithm,” Hindawi Mathematical Problems in Engineering, 2023.

[2] H. S. Inc., “Heron Systems at DARPA Alpha Dogfight Trials,” (Sept. 25, 2020). Accessed: Dec. 06, 2022., [Online Video]. Available: https://www.youtube.com/watch?v=lldE5XFTA88.

[3] J. Oh, C. Kim, S. Ro, W. C. Choi and Y. Kim, “Air-to-air BFM Engagement Simulator for AI Engagement Model,” in Proc. Korea Inst. Mil. Sci. Technol. Conf., pp. 1753-1754. 2022.

[4] B. Vlahov, E. Squires, L. Strickland, and C. Pippin, “On Developing a UAV Pursuit-evasion Policy Using Reinforcement Learning,” in Proc. IEEE Int. Conf. Mach. Learn. Appl. (ICMLA), pp. 859-864, 2018.

[5] J. Bae, H. Jung, S. Kim, S. Kim and Y. Kim, “Deep Reinforcement Learning-Based Air-to-Air Combat Maneuver Generation in a Realistic Environment,” in IEEE Access, Vol. 11, pp. 26427-26440, 2023.

[6] M. Wiering and M. Van Otterlo, “Reinforcement Learning”, Adaptation, Learning and Optimization, Vol. 12, p. 3, 2012.

[7] T. Haarnoja, A. Zhou, K. Hartikainen and G. Tucker, “Soft Actor-Critic Algorithms and Applications,” 2019, arxiv:1812.05905v2.

[8] S. Hochreiter and J. Schmidhuber, “Long Short-term Memory,” Neural Comput., Vol. 9, No. 8, pp. 1735-1780, 1997.

[9] R. Portelas, C. Romac, K. Hofmann, “Automatic Curriculum Learning for Deep RL : A Short Survey,” IJCAI, 2021.

[10] J. Berndt, “Jsbsim: An Open Source Flight Dynamics Model in c++,” in Modeling and Simulation Technologies Conference and Exhibit. American Institute of Aeronautics and Astronautics, 2004.

[11] A. Pope, J. Jaime, D. Mi´covi´c, H. Diaz, D. Rosenbluth, L. Ritholtz, J. Twedt, T. Waler, K. Alcedo, and D. Javorsek, “Hierarchical Reinforcement Learning for Air-to-air Combat,” 2021 International Conference on Unmanned Aircraft Systems(ICUAS), pp. 275-284, 2021.