Journal of Institute of Control, Robotics and Systems (2020) 26(2):000-000
https://doi.org/10.5302/J.ICROS.2020.18.0085 ISSN:1976-5622 eISSN:2233-4335

심층 강화학습을 이용한 항공기 충돌 회피 모델링 및 모델 최적화

Modeling on Aircraft Collision Avoidance and Model Optimization using Deep Reinforcement Learning

박 건 우, 김 종 한*
(Kun Woo Park[1] and Jong Han Kim[1,*])
[1]Department of Electronics and Information Convergence Engineering Kyung Hee University

Abstract: Reinforcement learning is an artificial intelligence technology in which an agent interacts with a given environment to learn the proper action for each state. There are two types in reinforcement learning, a value based method for learning Q function and a policy based method for learning behavioral policies. The TD method, which is a classical value based learning learning algorithm, has a high biased problem, and the Monte-Carlo estimate algorithm used in Policy based learning has a high variance problem. In order to minimize these problems, an Actor Critic algorithm was devised that combined the value-based and policy-based methods. Moreover, PPO algorithm that limited policy updates to surrogate objectives was proposed to more stably converge the policy. However, it is still difficult to converge to an optimized reward due to the characteristics of the control field that has a continuous and vast state. In this paper, we propose a learning method in the form of transfer learning in which the model trained by supervised learning based on domain knowledge, and verify the excellence of optimization through reinforcement learning

Keywords: Reinforcement learning, Machine learning, Transfer learning, Collision avoidance, Optimization

I. 서론

최근 로보틱스와 같은 제어 분야에서 객체 간의 충돌 회피를 목적으로 강화학습을 이용한 기술 개발이 진행되어왔다[1]. 본 논문에서는 이러한 강화학습을 통해 최적화된 항공기 충돌 회피 네트워크를 학습시키려 한다.

강화학습은 학습의 주체인 에이전트가 주어진 환경과 상호작용하며 에피소드를 진행하고, 이에 따라 에이전트를 학습시키는 인공지능 기술이다. 강화학습의 방식에는 크게 학습의 보상으로 Q함수를 학습하는 Value based 방식과 행동 정책을 학습하는 Policy based 방식이 있다. 고전적인 Value based learning의 학습 알고리즘인 TD 방식[2]은 각 에피소드 간의 variance는 작으나, high biased한 문제가 있고, Policy based learning에서 쓰이는 Monte-Carlo estimate 알고리즘의 경우에는 각 에피소드 간의 bias는 작으나 high variance한 문제가 있다. 이러한 문제점을 최소화하기 위해 value based와 policy based 방식을 결합한 Actor Critic[3] 알고리즘이 고안되었고, 이때 정책을 더 안정적으로 수렴시키기 위해 Policy update를 surrogate objective로 제한한 PPO[4] 알고리즘이 제안되었다. 그러나 여전히 제어영역과 같이 연속적이고 방대한 상태를 가지는 환경에서의 문제에는 수많은 local minima가 존재하여 optimal한 reward로 수렴시키는데에 어려움이 존재한다.

본 논문에서는 에이전트의 학습에서 발생하는 이러한 문제점을 해결하고, optimal한 reward에 더 쉽게 수렴할 수 있도록 domain knowledge[5]에 기반을 둔 지도학습으로 학습시킨 모델을, 강화학습을 이용하여 최적의 reward로 수렴시키는 일종의 전이학습[6] 형태의 학습방식을 제안한다.

학습의 전반적인 개요는 다음과 같다. 먼저 domain knowledge를 기반으로 회피기동 알고리즘을 설계한다. 이때 회피기동을 위해 고려되는 feature는 레이더에서 제공하는 정보와 동일한 $(r, v_c, \theta, \dot{\phi}, \dot{\theta})$의 총 5가지이고, 이를 통해 상승, 하강, 유지명령을 출력한다. 해당 알고리즘을 기반으로 random한 충돌상황에서 5가지 feature와 기동명령 출력을 label로 한 데이터셋을 샘플링하여 추출한다. 이렇게 추출된 레이더 정보와 회피기동 명령을 네트워크의 입력과 출력으로 제공하여 여러 구조의 네트워크에 대해 지도학습을 진행하고, 침입기와의 최소거리에 대한 표준편차, 평균값, 네트워크 파라미터 개수를 통해 가장 성능이 최적이라고 판단되는 네트워크를 채택한다. 채택된 네트워크는 PPO알고리즘의 policy network의 초기 가중치로써 적용되고, 추가적인 강화학습을 통해 네트워크를 최적화한다.

본 논문의 학습방식으로 학습된 강화학습 모델은 domain knowledge를 기반으로 한 회피기동 명령에 비해 평균적으

*Corresponding Author
Manuscript received 20xx.xx.xx; revised 20xx.xx.xx; accepted 20xx.xx.xx
박건우: 경희대학교 전자정보융합공학과 석사과정 대학원생(kunwoopark@khu.ac.kr, ORCID 0000-0002-8312-0774)
김종한: 경희대학교 전자정보융합공학과 조교수(jonghank@khu.ac.kr, ORCID 0000-0002-9030-0490)
※ 본 연구는 ~~~의 지원을 받아 진행되었음.

로 17%가량 낮은 회피 오차를 보이며, 의도된 회피 거리에 더 가깝게 회피함을 확인할 수 있다.

II. Domain Knowledge 회피기동 알고리즘 설계

본 논문에서 사용되는 기호는 아래와 같다.

기호

r : 침입기와 회피기간의 상대거리
v_c : 침입기의 접근 속도
ϕ : 횡방향 시선각
θ : 수직방향 시선각
t : 경과 시간
h : 고도
A_t : t시점의 현재 action
S_t : t시점의 현재 state
R_{t+1} : t시점의 action에 대한 immediate reward
D_{t+1} : t시점의 action에 대한 episode done flag
LP_t : t시점의 policy net 출력의 log probability
π_{old} : PPO_{old}의 policy
π_{new} : PPO_{new}의 policy
K : Policy update
N_{epi} : Total number of training episodes

항공기의 충돌 여부를 감지하는 방식은 입력정보 5가지를 통해 3가지 파라미터를 계산하여 결정한다. 각각 현재 운동 상태를 유지했을 때의 수직 방향 최소접근거리(MDV)와 수평 방향 최소접근거리(MDH), 현재 수직고도 차이(d_c)이고, 각 파라미터들은 아래와 같은 수식으로 계산된다.

$$MDH = \frac{r^2}{v_c \times \dot{\phi}} \tag{1}$$

$$MDV = \frac{r^2}{v_c \times \dot{\theta}} \tag{2}$$

$$d_c = r \times \theta \tag{3}$$

이렇게 계산된 파라미터들로 그림. 1.의 알고리즘을 통해 충돌여부를 감지하고, 회피명령을 내린다. 여기서 $\dot{h}_{cmd}$는 고도 변화율(회피기동) 명령이고, d_s는 최소 회피거리이다 (상대거리가 이 거리보다 가까워지면 충돌로 간주한다.).

III. 지도학습 네트워크 설계

지도학습 네트워크의 학습은 총 30만개의 학습데이터와 9만개의 검증데이터로 진행된다. 학습 데이터들은 설계된 회피 알고리즘과 뒤따르는 시뮬레이션을 사용하여 입력데이터 5개와 출력 회피명령 정보 1개 (3종류, 상승명령, 하강명령, 유지명령)가 한 데이터셋으로 구성된다.

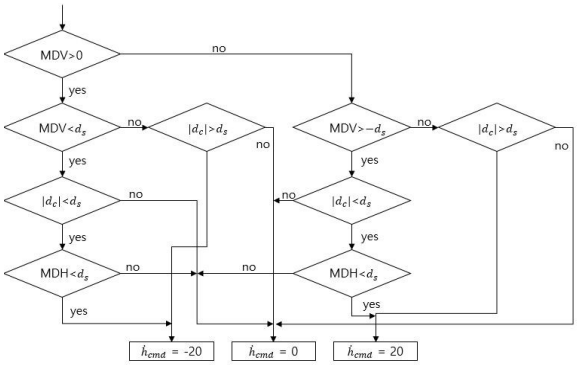


그림. 1. 지도학습에 사용된 domain knowledge rule based 알고리즘

Figure. 1. Domain knowledge rule based algorithm used for supervised learning

1. 시뮬레이션 설계

회피기의 시야 밖에 존재하는 항공기의 회피는 고려하지 않는다는 가정하에, 매 시뮬레이션은 그림. 2.과 같이 침입기가 회피기와의 충돌 지점에 $\pm 50^\circ$의 각도로 $2000(m)$ 반경의 영역에서 접근하도록 한다. 이때 다양한 접근 경로를 형성하기 위해, 항공기의 접근 경로각에 Gaussian normal distribution을 따르는 랜덤한 노이즈 값을 더해준다. 침입기의 출발 고도는 회피기의 고도를 기준으로 $\pm 50(m)$사이에서 $200(m/s)$의 속력으로 회피기에 접근한다. 이때 회피기 역시 충돌지점을 향해 $200(m/s)$의 속력으로 접근한다.

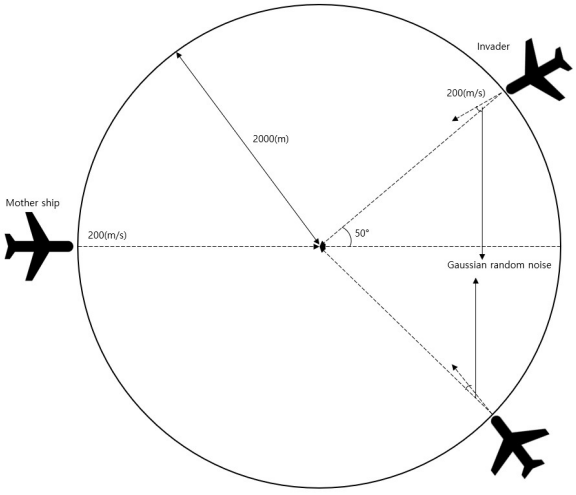


그림. 2. 회피기와 침입기의 충돌 시뮬레이션 개요

Figure. 2. Aircraft collision geometry

2. 네트워크 구조 설계 및 탐색

네트워크의 구조를 변경하며 가장 적합한 구조를 찾기 위해 그림. 3.과 같이 ResNet[7]구조로부터 영감을 받아 세 가지 블록을 만들어 각 블록들의 layer와 node의 수를 유동적으로 변화시킬 수 있도록 설계하였다. 각 블록은 Fully connected layer들로 구성되며, 활성함수는 음수 영역의 기

Journal of Institute of Control, Robotics and Systems (2020) 26(2):000-000
https://doi.org/10.5302/J.ICROS.2020.18.0085 ISSN:1976-5622 eISSN:2233-4335

울기로 0.1을 가지는 Leaky ReLU를 사용한다.

학습에 사용될 하이퍼 파라미터로, Learning rate는 1e-3, Batch size는 300을 사용했고, Cross Entropy loss 함수와 20 epoch의 반복도를 가지는 Step LR scheduler을 사용하였다. 각 블록의 층수는 [1, 1, 1]과 [2, 2, 2], 각 층의 노드수는 20에서 80까지 20간격으로 변화시키며 네트워크의 탐색을 진행하였다. 강화학습으로 전이학습을 진행하는 점을 고려하여, 각 네트워크의 학습 결과는 시나리오별 최소 상대거리의 크기와 분산이 작을수록 좋은 네트워크임을 의미하고, 각 네트워크는 서로 다른 4개의 random seed에 의해 100 epoch으로 4번씩 학습 되었다. 이렇게 4개의 다른 random seed로 학습된 동일한 네트워크 구조의 모델 4개에 대해, 학습상태와 동일한 조건으로 각 1500번씩의 검증 시나리오를 진행했다. 이때 검증 시나리오 마다의 상대거리 최소값에 대한 평균, 상대거리 최소값에 대한 표준편차, 모델 파라미터 개수를 각 0.8, 0.1, 0.1의 가중치로 합산하여 네트워크 구조별 점수를 부여했다. 최종적으로 층수 [2, 2, 2], 노드수 [40, 20, 60]을 가지는 구조의 네트워크가 채택되었다. 네트워크 탐색 결과는 그림. 4.과 그림. 5.에 등고선의 형태로 확인할 수 있다. 각 등고선의 가로축과 세로축은 각 블록별 노드 수를 의미하며, block1, block2, block3의 순서로 x_1, x_2, x_3으로 축 이름을 부여한다. 그림. 5.에 표시된 파란색 별 표시가 최종적으로 사용된 네트워크 구조이다. 학습에 사용된 코드는 [8]에서 확인할 수 있다.

IV. 강화학습 알고리즘 설계

1. Deterministic PPO 알고리즘

Policy network의 초기 가중치를 지도학습으로 학습된 네트워크를 사용하여 강화학습을 진행하는 방식에 기존의 PPO알고리즘에서 사용하는 형태의 stochastic policy로 action을 샘플링 할 경우 필요 이상으로 너무 많은 exploration을 진행하여 수렴에 방해요소로써 작용 한다. 따라서 학습된 네트워크 가중치를 fine tuning하기 위해 exploration의 양을 줄일 필요가 있고, 본 논문에서는 해당 문제점을 해결하기 위해 policy를 greedy한 action selection으로 deterministic하게 변경하여 학습을 진행한다. 표. 1.은 본 논문에서 사용하는 deterministic한 policy PPO알고리즘의 순서도이다.

2. Value Network(Critic Network) fine tuning

Actor Critic 알고리즘은 Actor(policy) network와 Critic(value) network의 두 네트워크가 상호작용하면서 함께 학습된다. 그러나 지도학습으로 학습시킨 네트워크는 action을 결정하는 policy에 대한 네트워크이기 때문에 초기 value network에 대한 가중치는 우리가 초기화한 policy의 가중치에 적합하지 않다. 따라서 본 논문에서는 초기 value network를 policy network와 동일한 가중치로 초기화한 후 일정 에피소드 $N_{finetune}$ 동안 value network만을 업데이트하면서 value network를 policy network에 맞게 fine tuning 한다. 본 논문에서는 $N_{finetune} = 5000$으로 지정하여 fine tuning을 진행하였다. 표. 2.은 본 논문에서 사용한 value network fine tuning 알고리즘이다. 학습에 구현된 코드는 [9]에서 확인 할 수 있다.

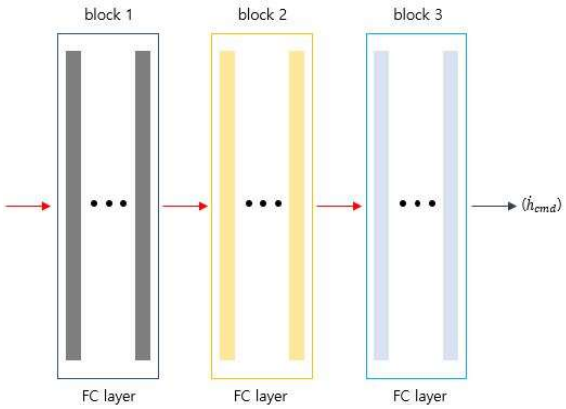
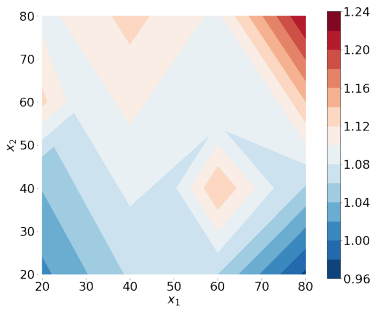
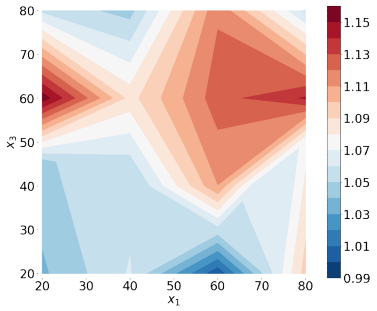


그림. 3. Domain knowledge rule based 지도학습 네트워크 구조

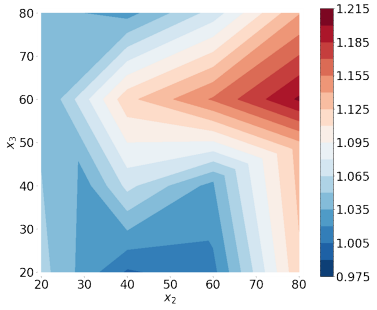
Figure. 3. Network structure for domain knowledge rule based supervised learning



(a) x_1- x_2



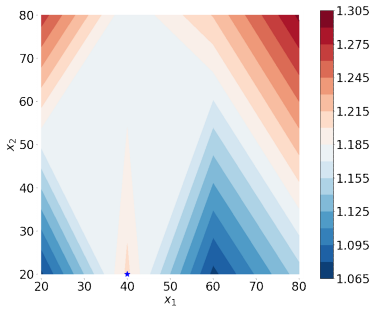
(b) x_1- x_3



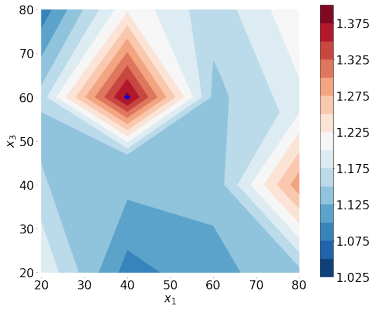
(c) x_2- x_3

그림. 4. 네트워크 층 수 [1, 1, 1]에 대한 네트워크 점수

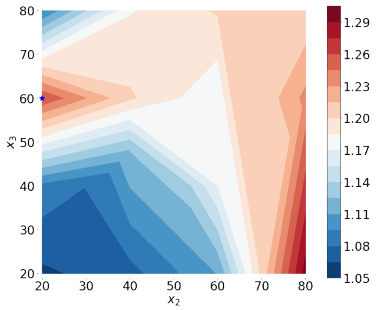
Figure. 4. Network scores for [1, 1, 1] layers



(a) x_1- x_2



(b) x_1- x_3



(c) x_2- x_3

그림. 5. 네트워크 층 수 [2, 2, 2]에 대한 네트워크 점수
Figure. 5. Network scores for [2, 2, 2] layers

표. 1. Deterministic PPO 알고리즘
Table. 1. Deterministic PPO algorithm

Algorithm. 1. Deterministic PPO
1. Initialize actor network and critic network of PPO_{old} and PPO_{new}
2. Initialize replay buffer and set time step t as 0
3. Repeat
4. Take action A_t maximizing π_{old}
5. Add $A_t, S_t, R_{t+1}, D_{t+1}, LP_t$ to replay buffer
6. If t is equal to update period
7. For 0, 1, ..., K
8. Evaluate and update π_{new} using samples of replay buffer
9. Update π_{old} as π_{new}
10. Clear replay buffer and set t to 0
11. Add 1 to time step t
12. Until number of episodes 0, 1, ..., N_{epi}

표. 2. Value network fine tuning 알고리즘
Table. 2. Algorithm for value network fine tuning

Algorithm. 2. Value network fine tuning
1. While current episode < $N_{finetune}$
2. Start from step 6. of Algorithm. 1.
3. Calculate surrogate loss and update only value(critic) network of PPO_{old} and PPO_{new}
4. Go to step 10. of Algorithm. 1.
5. End

표. 3. 강화학습 reward 설계
Table. 3. Design reward for RL

Algorithm. 3. Reward for training
1. Initialize episode time step t_e as 0
2. While episode is not done
3. Set reward as 0
4. If previous command ≠ current command
5. Add C_c to reward
6. Add $-(0.15t_e \mid \dot{h}_{cmd} \mid) + C_s$ to reward
7. Return reward
8. Add 1 to t_e
9. End while
11. If $r < d_s$
12. return reward as -4000
13. Else
14. return reward as 8000

3. Reward 설계

본 논문에서의 reward 설계 목표는 아래와 같다.

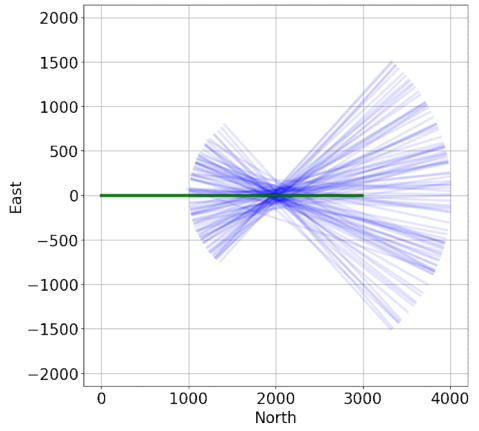
1. 최대한 적은 command 명령으로 회피를 할 것.
2. 에피소드의 시작 시점을 기준으로 최대한 빠르게 명령을 줄 것.
3. Command는 일정한 형태여야 하며, 회피 command가 진동하지 않을 것.

위 목표들을 모두 만족하도록 표. 3.과 같은 reward를 구성하였다. 여기서 C_s는 제어 명령이 적합한지를 판단하기 위한 상수이며, C_c는 제어 명령의 변화에 대한 페널티이다. C_c는 -100으로 주어지며, C_s는 회피명령을 비효율적인 방향으로 줬을 경우 -100으로, 이외의 경우 0으로 주어진다. 회피명령의 적합성은 초기 상태에서 회피기에 대한 침입기의 위치에 따라 정해지는데, 침입기가 위에 있을경우 하강명령을 더 적합하다고 판단하고, 아래에 있을경우 상승명령을 더 적합하다고 판단한다.

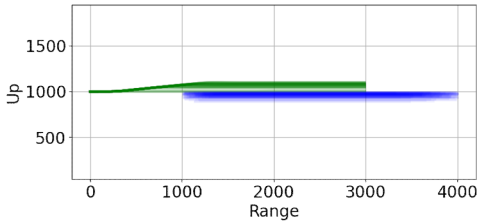
Journal of Institute of Control, Robotics and Systems (2020) 26(2):000-000
https://doi.org/10.5302/J.ICROS.2020.18.0085 ISSN:1976-5622 eISSN:2233-4335

V. 학습 결과

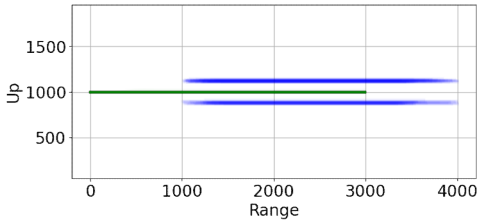
본 논문에서는 강화학습에 총 5만번의 에피소드, $N_{epi}=50000$와 $K=4$로 학습을 진행하였다. 하이퍼파라미터로, Learning rate는 1e-4, Discount factor γ는 0.999를 사용했고, 10000 episodes의 반복도를 가지는 Step LR scheduler을 사용하였다. PPO update period는 2000프레임(time step)으로 설정하였다.



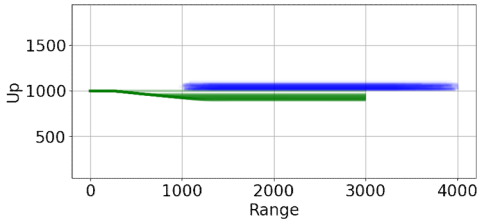
(a) Simulation NE trajectory



(b) 상승 회피 시나리오



(c) 고도 유지 시나리오



(d) 하강 회피 시나리오

그림. 6. 검증 시뮬레이션별 항공기 회피 경로

Figure. 6. Avoidance trajectory of each validating simulation

학습 결과를 검증하기 위해 위로 회피해야 하는 경우, 아래로 회피해야 하는 경우, 기동을 유지해야 하는 경우에 대해 각 1만번씩, 총 3만번의 검증 시뮬레이션을 통해 회피기동을 확인하였다. 여기서 각 환경은 침입기의 시작 고도만 상황에 맞게 설정하고, 이외의 모든 환경은 기존과 동일하게 진행하였다. 침입기의 시작 고도는 회피기의 고도 $+50(m)$, 혹은 $-50(m)$ 사이에만 있도록 하였다. 전자의 경우 아래로 회피해야 하는 경우에 해당하고, 후자의 경우 위로 회피해야 하는 경우에 해당한다. 기동을 유지해야 하는 경우에 대한 시뮬레이션은 침입기의 시작고도가 회피기의 고도보다 $100(m)$ 이상 차이 나도록 진행했다. 그림. 6.은 가시성을 위해 각 1만번의 시뮬레이션 중 random한 100개씩의 시뮬레이션 경로를 샘플링하여 출력한 결과이다. 회피기동이 필요한 상황에 대해, 각 시뮬레이션의 최소 회피거리는 그림. 9.와 그림. 10.을 참고하라.

1. 강화학습 알고리즘별 Reward 수렴 결과 비교

그림. 7.과 그림. 8.은 기존의 강화학습 알고리즘들 몇가지와 본 논문에 사용된 알고리즘의 reward 수렴 결과이다. 사용된 알고리즘들은 기존의 PPO방식(PPO_{nm})과 DQN방식(DQN_{nm}), 초기 weight를 supervised model의 weight로 초기화한 PPO(PPO_{wm})과 DQN(DQN_{wm}), 그리고 본 논문에서 제안된 deterministic policy와 value net fine tuning을 적용한 PPO_{vt}방식이다. 각 알고리즘들은 동일한 seed의 환경에서 진행되었으며, 각 그래프의 Initmodel 곡선은 초기의 supervised model에 대한 reward의 값이다. 그림. 7. 보면, 초기 weight가 supervised model로 초기화된 PPO방식들의 reward가 월등히 높게 수렴함을 알 수 있다. 이 PPO방식들 중 본 논문에서 활용된 학습 방식의 reward가 기존의 PPO 알고리즘만을 사용한 방식보다 11% 가량 더 높은 reward에 수렴하였다.

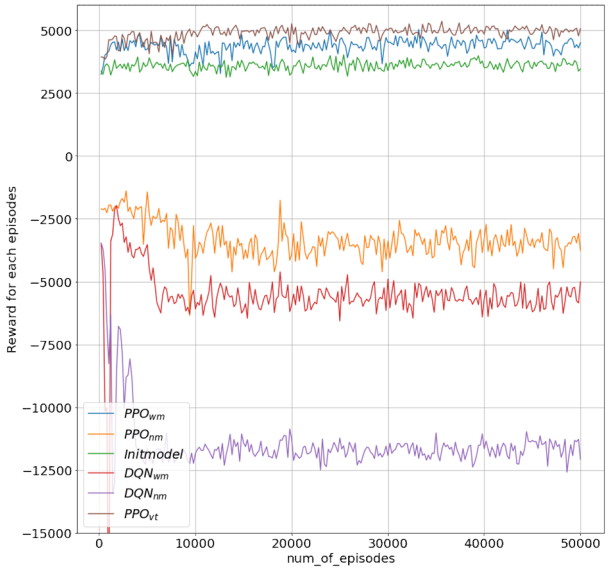


그림. 7. 강화학습 알고리즘에 따른 reward 수렴 결과 (PPO_{nm}, DQN_{nm}, PPO_{wm}, DQN_{wm}, PPO_{vt})

Figure. 7. Training reward for each RL algorithm (PPO_{nm}, DQN_{nm}, PPO_{wm}, DQN_{wm}, PPO_{vt})

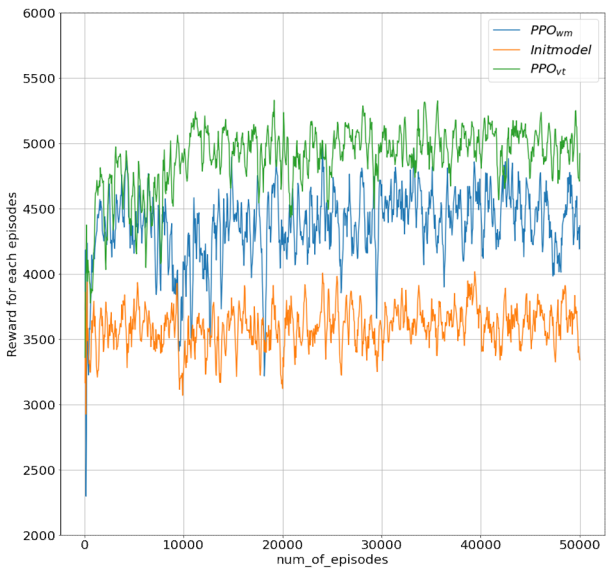


그림. 8. 강화학습 알고리즘에 따른 reward 수렴 결과 (PPO_{wm}, PPO_{vt})

Figure. 8. Training reward for each RL algorithm (PPO_{wm}, PPO_{vt})

표. 4. 회피 알고리즘에 따른 시뮬레이션 결과

Table. 4. Simulation results according to avoidance algorithm

Model	Mean value (m)	Variance	Error distance (m)
Domain knowledge	128.4	1171.5	28.4
Supervised learning	131.5	1482.6	31.5
Reinforcement learning	122.7	991.7	22.7

Model	Mean value (m)	Variance	Error distance (m)
Domain knowledge	124.9	979.0	24.9
Supervised learning	130.7	1422.0	30.7
Reinforcement learning	121.4	949.4	21.4

2. 회피기동 알고리즘별 회피결과 비교

본 논문의 알고리즘을 통해 학습된 모델과 기존의 domain knowledge 알고리즘, supervised learning을 통해 학습된 모델들의 효율을 앞서 말한 총 3만개의 검증 시뮬레이션을 통해 비교하였다. 검증 시뮬레이션 마다 상대거리의 최소값에 대한 분포를 그림. 9.와 그림. 10.에서 확인할 수 있다. 표. 4.는 각 알고리즘에 대한 상대거리 최소값의 평균, 분산, 오차값으로, 위에서부터 상승회피, 하강회피에 대한 시뮬레이션 결과이다. 분산과 회피오차는 최소 회피거리로 지정한 $100(m)$를 기준으로 하여 계산하였다. 결과를 보면 알 수 있듯, 본 논문의 알고리즘을 통해 학습된 모델이 기존 domain knowledge 알고리즘보다 17% 가량 낮은 회피 오차를 보이며, 해당 모델을 통해 원하는 최소 회피거리에 더 가깝게 회피가 가능하다.

VI. 결론

본 논문은 강화학습의 알고리즘 중 하나인 Actor Critic PPO 알고리즘과 지도학습의 결합을 통한 회피 네트워크 모델링과 최적화를 소개하였다. 제어기를 설계하는데 있어, 수많은 근사과정이 포함되거나, 혹은 그렇지 않은 경우 복잡한 모델링 과정이 필요한 경우가 많아 사람이 optimal한 제어기를 설계하는 데에는 한계가 존재한다. 본 논문은 이러한 문제점을 사람이 domain knowledge를 기반으로 근사하여 설계한 단순한 제어기를 강화학습을 통해 최적화하는 방식의 해결방안을 제시한다. 또한 복잡한 함수를 구현 할 수 있는 딥 러닝의 특성에 따라 필터링 된 데이터가 아닌 raw data까지의 확장 가능성을 제안하며, 강화학습을 통한 최적화의 우수성을 검증한다.

REFERENCES

[1] Y. F. Chen, M. Liu, M. Everett, J. P. How, “Decentralized Non-communicating Multiagent Collision Avoidance with Deep Reinforcement Learning”, Rerieved Mar, 16, 2021, from https://arxiv.org/pdf/1609.07845.pdf.

[2] Tesauro, G, “Practical Issues in Temporal Difference Learning” Machine Learning, 8, pp. 257-277, 1992.

[3] V. Mnih, A. P. Badia, M. Mirza, A. Graves, T. P. Lillicarap, T. Harley, D. Silver, K. Kavukcuoglu, “Asynchronous Methods for Deep Reinforcement Learning”, ICML, 2016.

[4] J. Schulman, F.Wolski, P. Dhariwal, A. Radford, O. Klimov, “Proximal Policy Optimization Algorithms”, Retrieved Mar, 16, 2021, from https://arxiv.org/pdf/1707.06347.pdf.

[5] S. C. Han, H. C. Bang, “Proportional Navigation-Based Optimal Collision Avoidance for UAVs”, vol. 10, no. 11, pp. 1065-1070, 2004.

[6] Weiss, K., Khoshgoftaar, T.M., Wang, D.: A survey of transfer learning. Journal of Big Data 3(1), 9 (2016). DOI 10.1186/s40537-016-0043-6

[7] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” arXiv:1512.03385, 2015.

[8] https://github.com/kun-woo-park/Imitation_learning

[9] https://github.com/aisl-khu/collision_avoidance/tree/master/Aircraft_avoidance_RL/Col_avoid_new/ppo_fin

Journal of Institute of Control, Robotics and Systems (2020) 26(2):000-000
https://doi.org/10.5302/J.ICROS.2020.18.0085 ISSN:1976-5622 eISSN:2233-4335

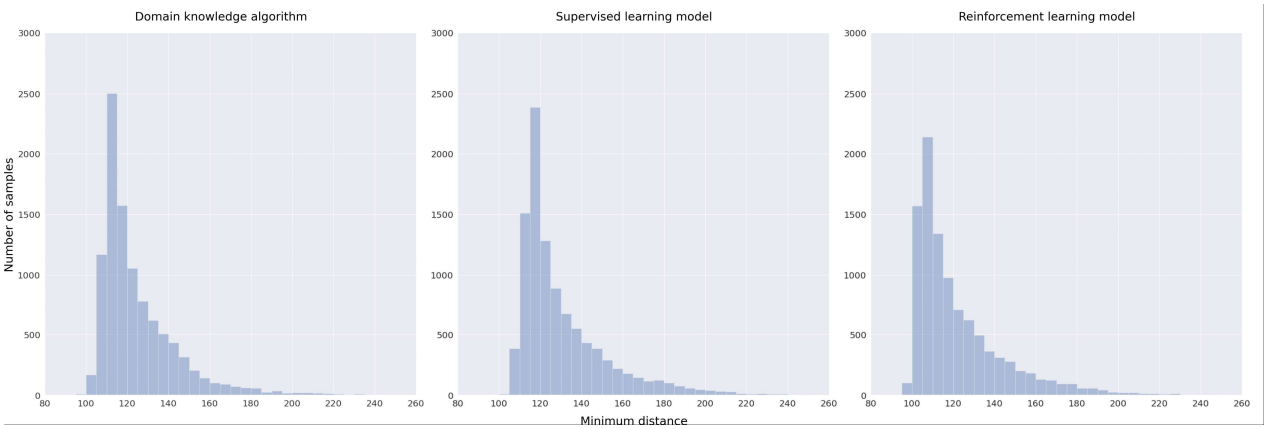


그림. 9. 회피 알고리즘에 따른 상승 회피 시뮬레이션 최소 접근거리 분포
Figure. 9. Distribution according to avoidance algorithm(avoiding up)

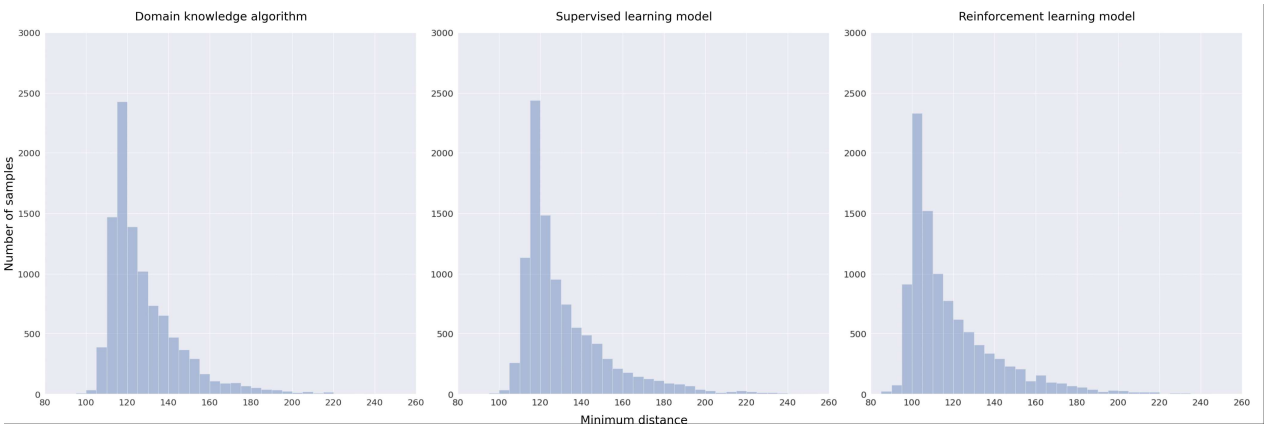


그림. 10. 회피 알고리즘에 따른 하강 회피 시뮬레이션 최소 접근거리 분포
Figure. 10. Distribution according to avoidance algorithm(avoiding down)

박건우(Kun Woo Park)

2020년 경희대학교 전자공학과 학사(공학사)졸업.
2020년 ~ 현재 경희대학교 전자공학과 석사과정. 관심 분야는 심층 강화학습

김종한 (Jong Han Kim)

B.S. in Aerospace Engineering, Korea Advanced Institute of Science and Technology (KAIST) (1999)
M.S. in Aerospace Engineering, Korea Advanced Institute of Science and Technology (KAIST) (2001) Flight Dynamics and Control Lab. Ph.D. in Aeronautics and Astronautics, Stanford University (2012) Information Systems Lab.
Senior Researcher, Agency for Defense Development (ADD), Republic of Korea (2001-2018)
Assistant Professor, Department of Electronic Engineering, Kyung Hee University, (2018-)