강화학습을 활용한 무인전투기의 유도탄 회피 및 최단 거리 탐색 알고리즘 개발

이경택, 민병욱, 김정철, 김창욱*
연세대학교 산업공학과
{bluediary8, dmbw612, od519bo, kimco}@yonsei.ac.kr

Introduction

분석 배경 및 목적

- 최근, 국방 분야에 인공지능 기술을 활용하는 중요성이 대두됨에 따라 무인전투기에 인공지능 기술을 접목시키는 연구 또한 활발히 진행되어 오고 있음
- 강화학습 알고리즘을 활용하여 무인전투기의 유도탄 회피 및 목적지까지의 최단거리 탐색 알고리즘을 제안

의의

- 국내외 최초로 3차원 공간상에서 무인전투기의 유도미사일 회피 및 최단거리를 위한 강화학습 알고리즘 개발
- 좌표 및 각도에 대하여 효율적으로 표현할 수 있는 방법 제안

Introduction

강화학습

- 어떤 환경 안에서 정의된 에이전트가 현재의 상태를 인식하여, 선택 가능한 행동들 중 보상을 최대화하는 행동 혹은 행동 순서를 선택하는 방법

요소	알파고	단일 무인전투기	강화학습 시 도전과제
객체(agent)	알파고	무인전투기	-
환경(environment)	바둑판	전장상황	
상태(state, s)	바둑판 위의 바둑돌 위치	무인전투기와 유도미사일의 상태 정보	차원의 저주
행동(action, a)	착수	추력과 경사각	강화학습 모델
보상(reward, r)	승리 시 + 보상 패배 시 - 보상	목표지점 근접시 큰 보상 회피 시 + 보상 격추 시 – 보상	보상 디자인

전장환경

전장환경구성

- $75 \times 175 \times 50\text{km}$ 크기의 전장 가정
  - 방공망 크기는 반경 30km 가정
  - 배치된 중첩된 4개의 방공망으로 학습(6개의 방공망으로 테스트)
- 유도미사일 가정
  - 유도미사일은 무인 전투기와 1회만 교전
  - 격추 실패 후에는 연료의 한계로 인해 기능을 상실함(목표 재지정 없음)
  - 단일 무인전투기는 동시에 최대 2개의 유도미사일과 교전 할 수 있음
- 무인전투기 가정
  - 무인전투기의 임무는 공대지 작전으로 가정
  - 무인전투기는 무장투하구역(launch available region)에 도달하는 것을 목표로 함
  - 무인전투기는 적 지대공 미사일의 발사대 위치 및 발사된 미사일 위협에 대해 실시간 탐지가 가능하다고 가정

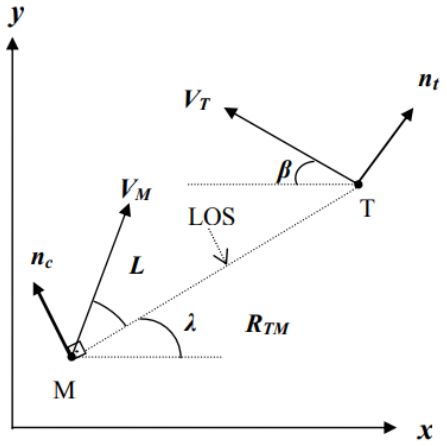
전장환경

지대공 유도미사일 구현

- 기존 논문들에서 소개된 비례항법유도를 사용
  - 3차원에서의 비례항법유도를 풀되, xy, yz, xz면에서 각각을 푼 뒤 합치는 방법
  - 거리(R_{TM})와 상대속도(V_{TM})를 안다면, 미사일을 가속(n_c)하여 목표물을 명중할 수 있게 조종가능

$$n_{c\_xy} = N' \cdot V_{c\_xy} \cdot \dot{\lambda}$$
$$= N' \cdot \frac{-(R_{TMx}V_{TMx} + R_{TMy}V_{TMy})}{R_{TMxy}} \cdot \frac{(R_{TMx}V_{TMx} - R_{TMy}V_{TMy})}{R_{TMxy}^2}$$

n_{c_xy} : Missile에 입력되는 가속도
N' : 비례상수
V_{c_xy} : -거리의 변화량
$\dot{\lambda}$: LOS의 변화량
R_{TM} : Missile과 target의 거리
V_{TM} : Missile과 target의 상대속도



Moran, I., & Altilar, D. T. (2005). Three plane approach for 3D true proportional navigation. *AIAA Paper*, (2005-6457), 3-10.

Agent

무인전투기 상세 구현

- 질점으로 표현된 전투기의 운동방정식
- 기존 국내외 UAV 시뮬레이션 논문들에서 주로 사용된 운동방정식을 사용

$$\dot{V} = \frac{T-D}{m} - g\sin\gamma$$
$$\dot{\psi} = \frac{gn\sin\phi}{V\cos\gamma}$$
$$\dot{\gamma} = \frac{g}{V(n\cos\phi - \cos\gamma)}$$

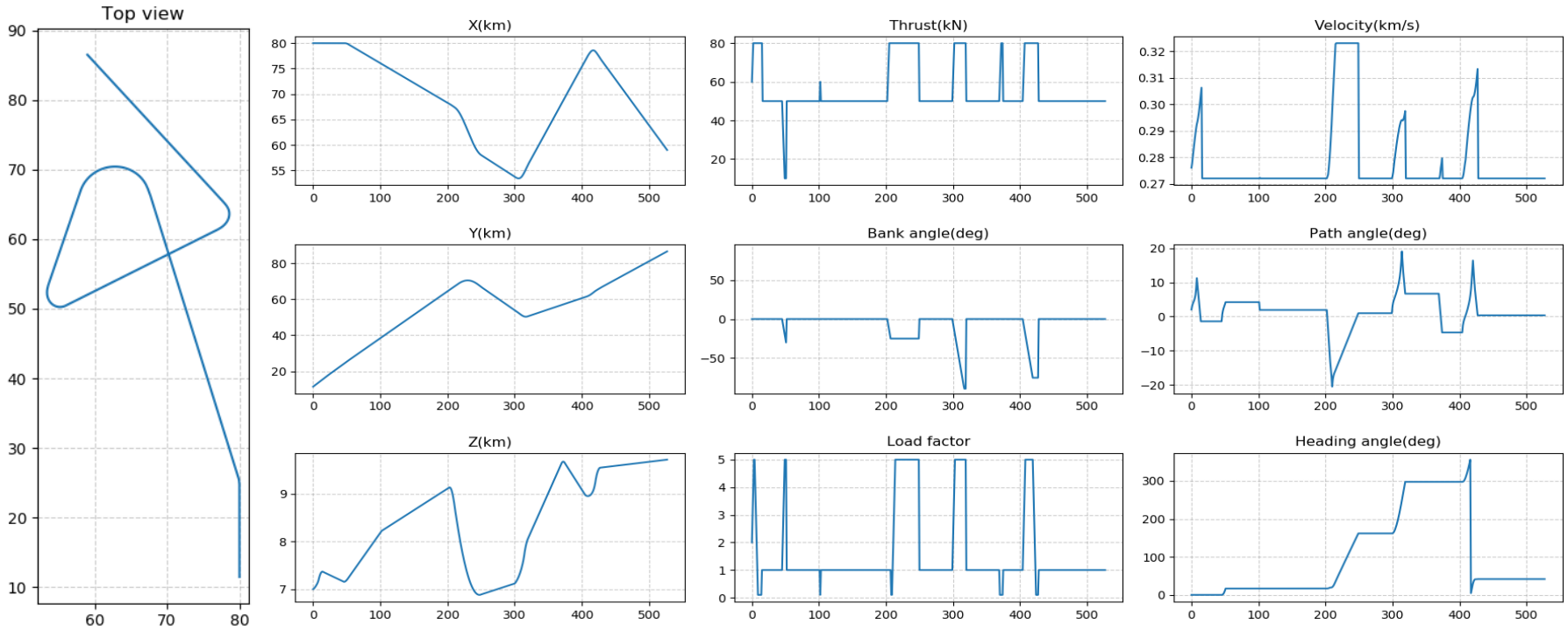
$$\dot{x} = V\cos\gamma\cos\psi$$
$$\dot{y} = V\cos\gamma\sin\psi$$
$$\dot{z} = V\sin\gamma$$

x, y, z : 위치
V : 속도
ψ : 방위각(heading angle)
γ : 상하각(path angle)
T : 추력(engine thrust)
ϕ : 경사각(bank angle)
D : 항력(drag force)
n : 하중계수(load factor)
m : 무인전투기 질량(mass)

Kim, S., & Kim, Y. (2007, October). Three dimensional optimum controller for multiple UAV formation flight using behavior-based decentralized approach. In *Control, Automation and Systems, 2007. ICCAS'07. International Conference on* (pp. 1387-1392). IEEE.

Agent

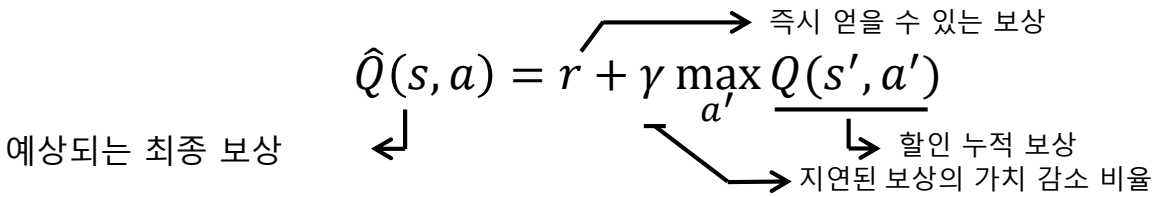
무인전투기 시뮬레이터 결과 예시



제안 방법론

Q-learning

Q-learning은 강화학습 알고리즘의 하나로 현재 state에서 선택 가능한 action을 선택하고 그에 따른 reward를 받아서 Q-function을 최적의 값으로 갱신하여 환경을 학습하는 형태



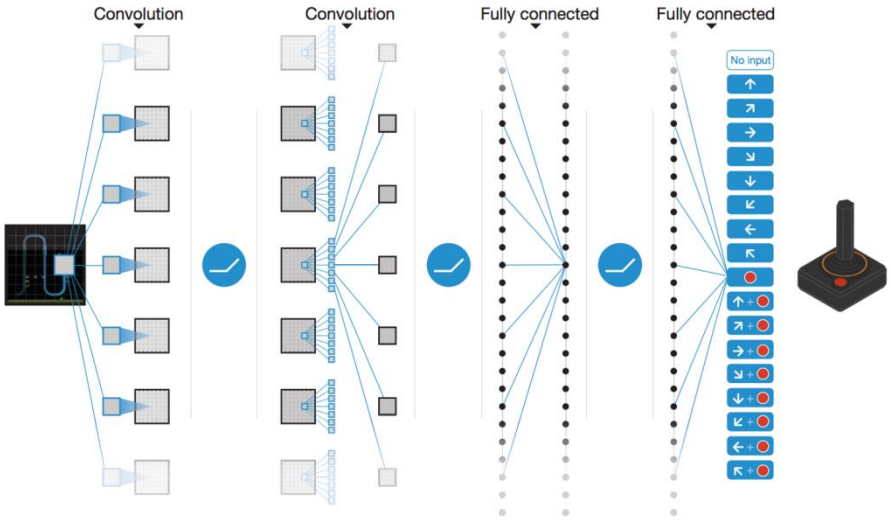
Q-learning 알고리즘

1. 모든 상태(s)와 행동(a)에 대해 $\hat{Q}(s,a)$의 값을 0으로 설정한다.
   ($Q(s,a)$: 특정 상태(s)에서 특정 행동(a)을 했을 때 받을 수 있는 전체 보상)
2. 현재 상태 s 관측
3. 다음을 특정 횟수 혹은 시간 동안 반복한다.
   (1) 행동 a를 선택하고 실행한다.
   (2) 즉각 보상 r을 관측(없는 경우 0)한다.
   (3) 행동 후 전이된 새로운 상태 s' 관측한다.
   (4) $\hat{Q}(s,a)$을 다음의 수식을 통해 업데이트: $\hat{\boldsymbol{Q}}(\boldsymbol{s},\boldsymbol{a}) = \boldsymbol{r} + \boldsymbol{\gamma} \max_{\boldsymbol{a'}} \boldsymbol{Q}(\boldsymbol{s'},\boldsymbol{a'})$한다.
   (5) $s \leftarrow s'$업데이트 한다.

제안 방법론

Deep Q-network (DQN)

- 강화학습에 딥러닝(deep learning)을 결합한 알고리즘
- Q-network를 통해 더 많은 상태와 행동에 대해서 강화학습이 가능하게 되었지만, 학습데이터 간의 상관성, non-stationary target문제가 있음
- 이러한 문제점을 deep Q-network를 통해 해결함

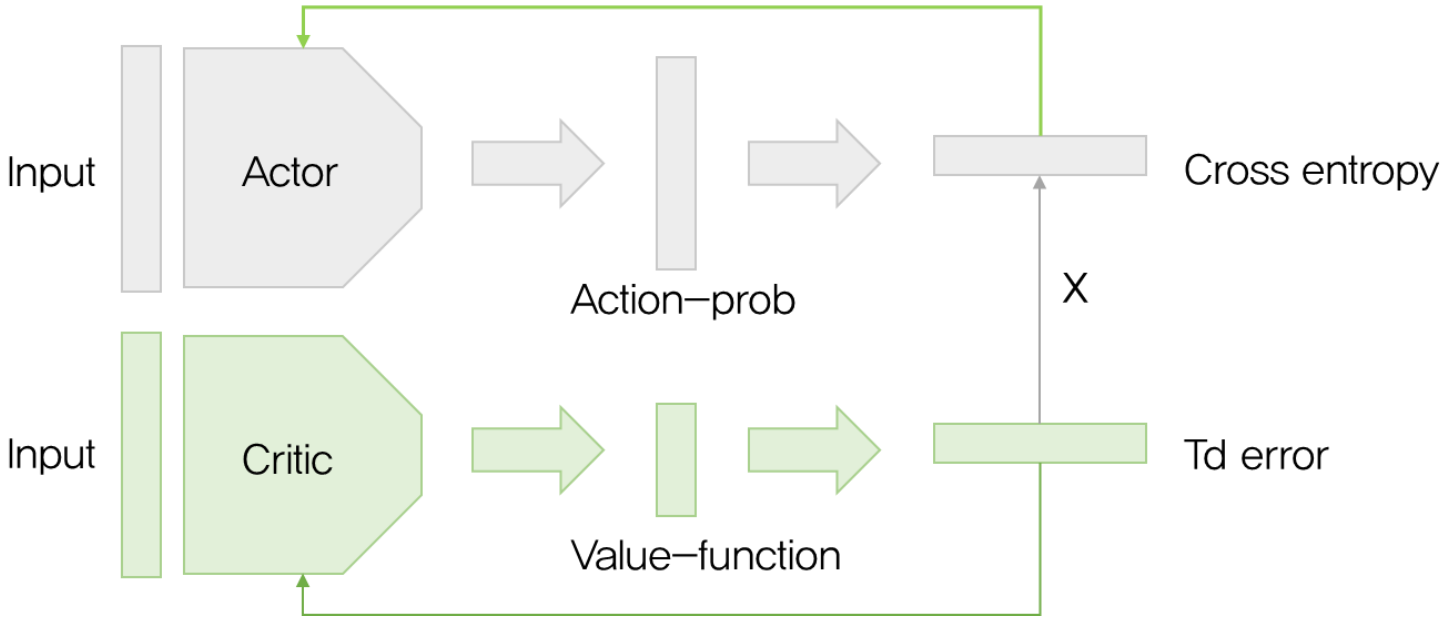


Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., ... & Petersen, S. (2015). Human-level control through deep reinforcement learning. *Nature*, *518*(7540), 529-533.

제안 방법론

Actor-Critic Network

- 정책을 근사하는 actor network + 가치 함수를 근사하는 critic network
- State에 대응되는 action에 따른 확률 값을 학습 (정책 자체를 학습)
- 본 연구의 actor network와 critic network에서 모두 단일 hidden layer 사용



제안 방법론

강화학습 요소

<table>
<tr><th colspan="2">State</th><th>Action</th><th colspan="2">Reward</th></tr>
<tr><td rowspan="3">무인전투기</td><td>비행경로</td><td rowspan="2">추력</td><td>최단 거리 달성</td><td>무장 투하 구역과의 거리</td></tr>
<tr><td>상하각, 경사각 방위</td><td rowspan="3">임무결과</td><td rowspan="2">무장 투하 구역 도착</td></tr>
<tr><td>속도</td><td rowspan="3">경사각</td></tr>
<tr><td rowspan="2">무인전투기와 지대공 미사일 간</td><td>거리</td><td rowspan="2">임무지역 이탈, 추락</td></tr>
<tr><td>상하각, 좌우각</td><td>위협회피</td><td>유도미사일 교전 결과</td></tr>
</table>

제안 방법론

강화학습 요소

- State – 무인전투기의 위치 표현

200x400의 전장에서 무인전투기의 위치를 표현하는 방법

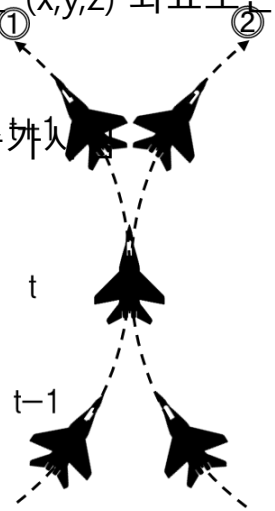
방법		데이터 표현	표현범위	차원	학습 특징
기존 방법론	raw [x, y]	[1.3, 2.5]	실수	2	학습 불가능 (x, y좌표값의 상관관계 학습)
	one-hot encoding	200rows $\begin{bmatrix} 0 & 0 & \vdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & 1 & \vdots \\ 0 & 0 & \vdots & 0 \end{bmatrix}$ 400rows	정수	200×400	학습 불가능 (차원의 저주)
신규 방법론	제안방법1	200rows $\begin{bmatrix} 1 \\ \vdots \\ 0 \\ \vdots \\ 0 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \\ \vdots \\ 0 \\ \vdots \\ 0 \end{bmatrix}$ 400rows	정수	$200 + 400$	학습 가능 (정수 값을 갖는 좌표만 가능)
	제안방법2	200rows $\begin{bmatrix} 0.7 \\ 0.3 \\ \vdots \\ 0 \end{bmatrix} \begin{bmatrix} 0 \\ 0.5 \\ 0.5 \\ \vdots \\ 0 \end{bmatrix}$ 400rows	실수	$200 + 400$	학습 가능 (실수 값을 갖는 좌표 가능)
	최종 제안방법	20rows $\begin{bmatrix} 0.13 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \begin{bmatrix} 0.25 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$ 40rows	실수	$20 + 40$	효율적인 학습 가능 (실수 값을 갖는 좌표 가능)

제안 방법론

강화학습 요소

- State – 무인전투기의 비행경로 표현
  - 연속적인 비행경로를 표현하기 위해서 현재 t시점에서의 단일 위치 정보보다 t-4부터 현재 시점 t까지의

    연속된 위치정보를 표현
  - 예를 들어, 전장 크기(200, 400, 100)에서 5 steps(t-4부터 t까지)을 표현하기 위해서는 (x,y,z) 좌표로는 4.0×10^7 차원이 필요하나, 제안 방법으로는 350 차원으로 표현이 가능
  - 무인 전투기의 단일 점 위치보다 비행경로를 state에 포함하면 강화학습 효율성을 증가시



제안 방법론

강화학습 요소

- State – 무인전투기의 각도 표현
  - 무인전투기의 상하각, 경사각, 방위의 표현
    - One-hot-encoding 방식으로 표현(방위 예시) – 정수에 한해서만 표현이 가능

360 rows $\begin{bmatrix}1\\0\\0\\\vdots\\0\end{bmatrix}$ $\begin{bmatrix}0\\1\\0\\\vdots\\0\end{bmatrix}$ ... $\begin{bmatrix}0\\0\\0\\\vdots\\1\end{bmatrix}$

$\psi = 1^{\circ}$ $\psi = 2^{\circ}$ ... $\psi = 360^{\circ}$

  - One-hot-encoding 방식의 문제점
    - 방위를 실수값으로 표현하면 학습이 이루어지지 않음
    - 비행시 방위 변화가 10°에서 350°로 이루어질 때 20° 의 변화만 필요하나, one-hot-encoding에서는 340°의 변화가 필요하다고

      인식됨

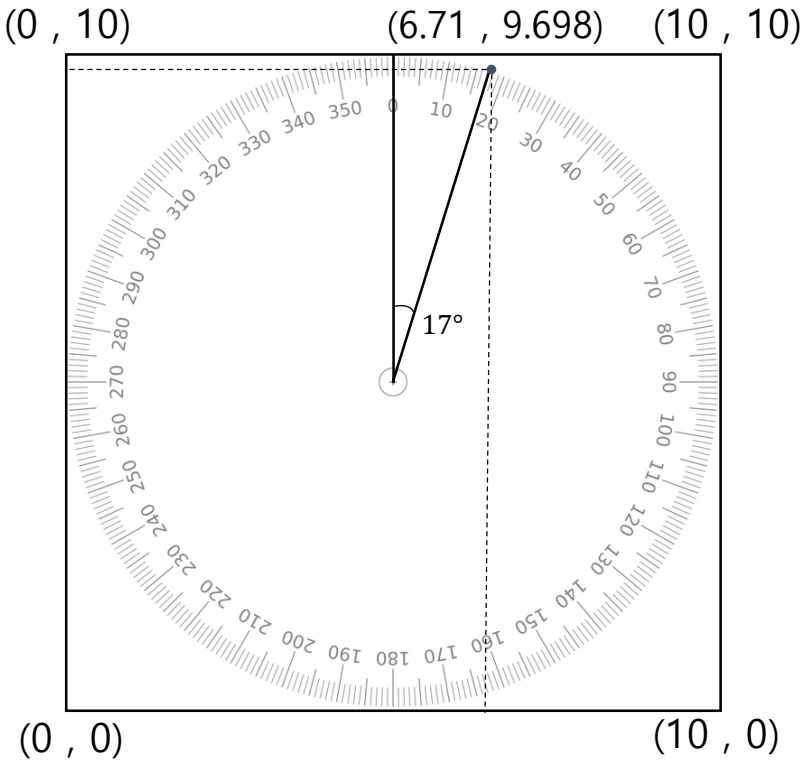
제안 방법론

강화학습 요소

- State – 무인전투기의 각도 표현
  - One-hot-encoding 적용에 따른 문제점 해결
    - 각도를 2차원 좌표로 표현
    - 반지름이 5인 원상의 좌표로 대응시켜 각도를 표현
    - $0° \sim 360°$의 값에 대해 모두 표현 가능

- 예시: $17° \rightarrow (6.71, 9.698) \rightarrow \begin{bmatrix} 0 \\ \vdots \\ 0.71 \\ 0.29 \\ \vdots \\ 0 \end{bmatrix} \begin{bmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 0.302 \\ 0.698 \end{bmatrix}$

10 rows x 2 columns



제안 방법론

강화학습 요소

- 최종 state
  - 무인전투기 정보
    - 무인전투기의 5 steps 동안의 비행경로(350 rows)
    - 무인전투기의 2 steps 동안의 상하각, 경사각, 방위 (10 rows x 2 columns x 3 angles x 2 steps)
    - 무인전투기의 속도
  - 지대공 미사일 정보
    - 무인전투기와 지대공 미사일 간의 거리(125 rows)
    - 무인전투기와 지대공 미사일 간의 상하각과 좌우각(10 rows x 2 columns x 2)

제안 방법론

강화학습 요소

- Action
  - 정밀한 조작을 위해 작은 변화 값을 지닌 총 9가지 행동
  - 추력
    - 증가: 무인전투기의 속도가 증가하고 상승함
    - 유지: 무인전투기의 속도 및 상하각 유지
    - 감소: 무인전투기의 속도가 감소하고 강하함
  - 경사각
    - 증가: 선회 반경 감소, 선회율 증가
    - 유지: 경사각 유지
    - 감소: 선회 반경 증가, 선회율 감소

Action		추력		
		-0.05	0	0.05
경사각	-2.0	1	4	7
	0	2	5	8
	2.0	3	6	9

제안 방법론

강화학습 요소

- reward
  - 유도미사일을 회피하여 무장 투하 구역에 도착하도록 reward 디자인
  - 기존 거리 방식의 reward 디자인은 signal이 매우 약함
    - 현재 t시점에서 t+1시점으로 이동하는데 가능한 action에 따른 reward의 편차가 매우 적음
    - Action에 따른 reward의 차이가 커지도록 하는 reward 디자인 필요

상황	비행 중		유도미사일 교전		임무 지역
	무장 투하 구역 도착전	무장 투하 구역 도착	성공	실패	이탈 및 추락
Reward	[0, 1]	+5	+3	-5	-5
참고사항	10 steps동안 action 유지				

실험

3차원 전장에서의 강화학습 실험

- 무인전투기의 속도, 추력, 경사각/상하각/방위에 대한 입력값의 범위를 설정
- 유도미사일은 일반적인 유도 및 비행 성능으로써 비례상수를 3으로 고정
  - 비례상수가 1이면 유도 및 비행 성능이 낮고, 6이면 유도 및 비행 성능이 높음

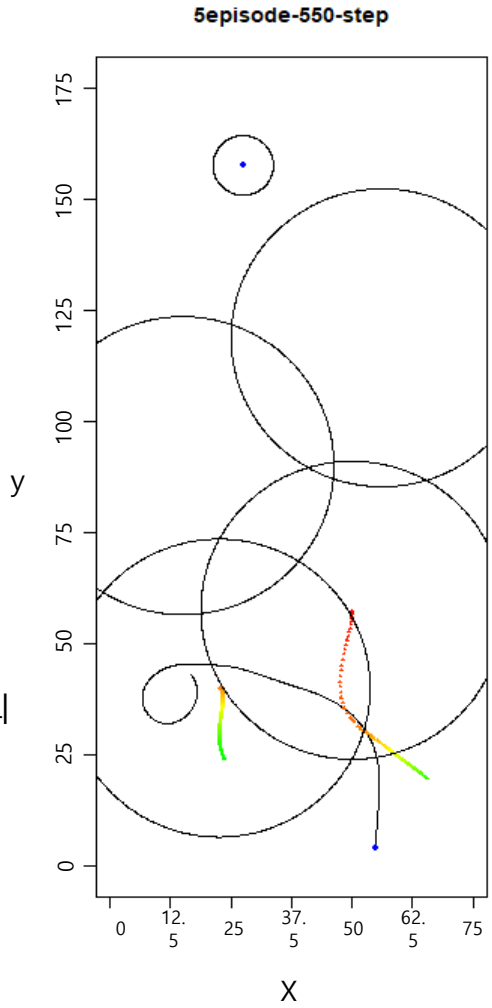
대상	무인전투기					유도미사일	
특성	속도	추력	경사각	상하각	방위	비례상수	속도
초기값	0.289	50	0	1	0	3	1.02
가용범위	0.272 ~0.323	0 ~ 80	-90 ~ 90	-90 ~ 90	0 ~ 360	고정	고정
단위	km/s	kN	각도	각도	각도	없음	km/s

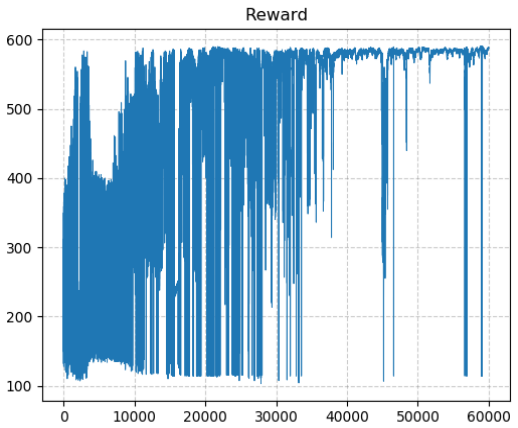
실험

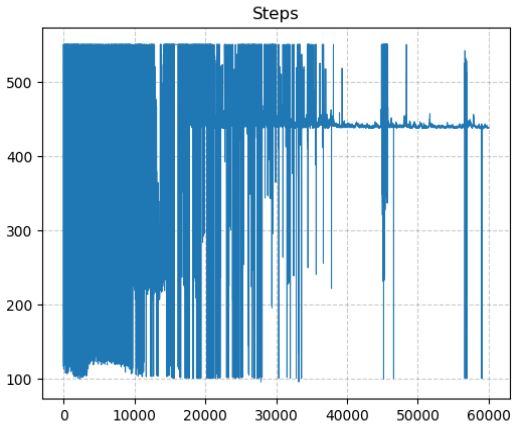
3차원 전장에서의 강화학습 실험

- 학습 과정(top view)
  - 60,000번 정도 episode 시도하면 수렴
  - 6시간 정도 후에 학습 완료
  - Reward가 학습이 진행되면서 증가
  - 학습이 진행되면서 목표지점까지 최단 거리 학습

- 학습 과정의 특이 사항
  - 학습 초반에는 선회를 하면서 reward를 증가
  - 그 후 더 좋은 reward를 위해 목표지점으로의 경로를 학습하기 시작하면서 격추 확률 증가
  - 미사일을 회피하는 가장 좋은 최단 경로와 기동을학습하기 시작







실험

3차원 전장에서의 강화학습 실험

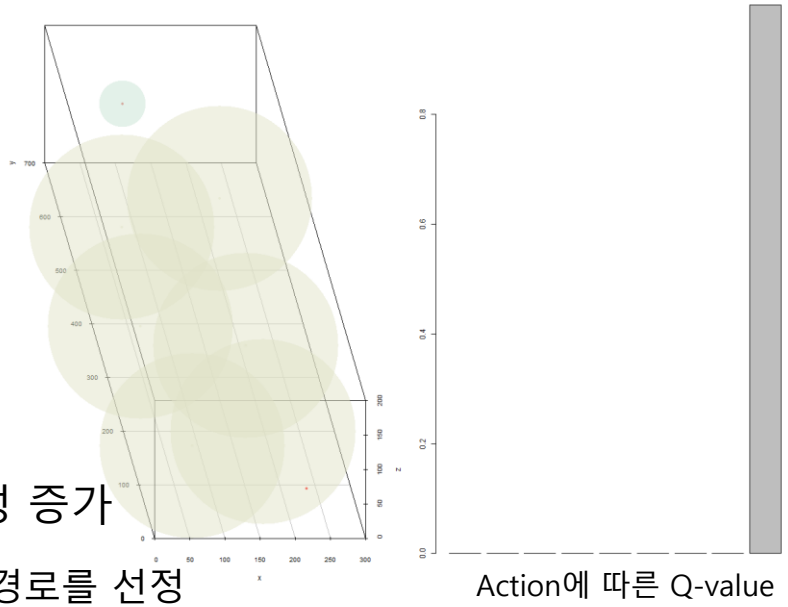
- 테스트 환경
  - 지대공 위협의 개수를 4개(학습) → 6개(테스트) 증가
  - 1,000번 episode 수행
- 강건성 실험
  - 지대공 위협의 위치를 임의(random)로 조정
  - 유도미사일 발사 확률 조정(지대공 위협 반경의 중심으로 갈수록 유도미사일 발사 확률 증가)
  - 1,000번 episode 수행

실험

3차원 전장에서의 강화학습 실험

- 실험 결과(임무 완수율)

테스트 환경(방공망 6개)	목표지점 도달 (%)
지대공 위협 위치 조정	99.8
지대공 위협 위치 조정 유도미사일 발사 확률 조정	99.7



Action에 따른 Q-value

- 유도미사일과 교전하는 시점에 Q-value의 변동성 증가
  - 중첩된 지대공 위협은 가운데로 침투할 수 있도록 경로를 선정
  - 연료량을 고려하여 최소한의 기동 및 경로 변경을 통해 침투를 시도

결론 및 의의

결론 및 의의

- 국내외 최초로 3차원 공간상에서 무인전투기의 유도미사일 회피 및 최단거리를 위한 강화학습 알고리즘 개발
  - 3차원 공간상에서의 무인전투기 및 유도미사일의 효율적인 상태 표현 제안
    - 복잡한 전장 상황(무인전투기 및 유도미사일의 비행 상태)을 강화 학습이 가능하도록 저차원의 vector로 표현
    - 단일 hidden layer로도 학습이 가능
    - 학습과 다른 테스트 전장 환경에서도 사전에 학습된 모델을 이용하여 성공적인 임무 달성이 가능함을 확인

감사합니다

[감사의 글]
본 연구는 국방과학연구소 순수기초연구(UD170043JD)의 재원으로 지원을 받아 수행되었음.