22년도 2학기 [자기주도연구1]

강화학습을 통한 죄수의 딜레마 게임 대응전략 연구

아주대학교 소프트웨어학과
주정호
이석원 교수님

목차

Intro

연구 소개

연구 주제를 기획하게 된 계기

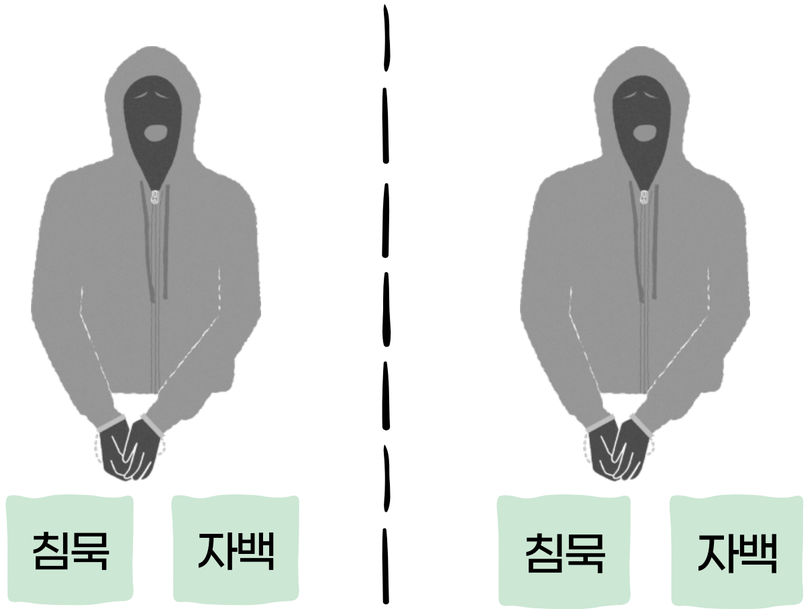
게임 이론

상호 의존적이고 이성적인 의사결정에 관한 수학적 이론. 참가자들이 상호작용하면서 변화해 가는 상황을 이해하고 분석하여 사회 과학, 경제학, 정치학 등에 적용되고 있다.

강화학습

강화학습을 게임 이론에 적용하여 최선의 전략을 연구하고, 이를 다양한 분야에서 발생하고 있는 문제를 해결할 수 있는 방법론으로 나아갈 수 있는지에 대해 증명하고자 한다.

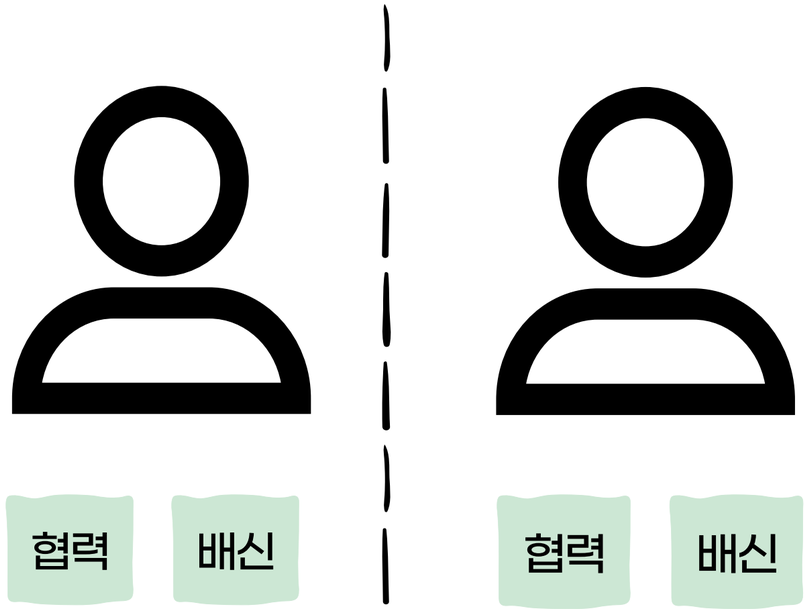
연구 소개 죄수의 딜레마



선택	침묵	자백
침묵	1년 / 1년	10년 / 석방
자백	석방 / 10년	3년 / 3년

연구 소개

죄수의 딜레마 게임



선택	협력	배신
협력	+2 / +2	-1 / +3
배신	+3 / -1	0 / 0

N 라운드

연구 소개 죄수의 딜레마 게임의 전략

All Cooperate	항상 협력한다.
All Cheat	항상 배신한다.
Copycat	첫 번째 수는 협력으로 시작한다. 그다음에는 상대가 바로 전 판에서 한 선택을 따라한다.
Grudger	항상 협력한다. 하지만, 상대가 한 번이라도 배신하면 끝까지 배신으로 보복한다.
Detective	협력-배신-협력-협력의 수로 시작한다. 이때 한 번이라도 상대가 배신한다면, Copycat과 똑같이 행동한다. 그렇지않다면, All Cheat과 똑같이 행동한다.
Copykitten	첫 번째 수는 협력으로 시작한다. 상대가 두 번 연속으로 배신할 때만 배신한다.
Simpleton	첫 번째 수는 협력으로 시작한다. 상대가 협력하면 내 마지막 수와 같은 수를둔다. 상대가 배신하면 내 마지막 수와반대로 둔다.
Random	협력 또는 배신을 무작위로 선택한다.
Cheat-Downing	상대방의 이력으로 협력 가능성과 배신 가능성을 계산하여, 확률이 높은 행동을 선택한다. (동률일 경우 배신)
Cooperarte-Downing	상대방의 이력으로 협력 가능성과 배신 가능성을 계산하여, 확률이 높은 행동을 선택한다. (동률일 경우 협력)

연구 소개 죄수의 딜레마 게임의 전략

Joss	Copycat과 동일하지만, 10%의 확률로 첫 번째 라운드에 배신한다.
Cheat-Tester	n/2 라운드를 협력 또는 배신을 무작위로 하고, 그 동안에 협력이 많으면 계속 배신, 그렇지 않으면 계속 협력한다. (동률일 경우 배신)
Cooperate-Tester	n/2 라운드를 협력 또는 배신을 무작위로 하고, 그 동안에 협력이 많으면 계속 배신, 그렇지 않으면 계속 협력한다. (동률일 경우 협력)
Tranquilizer	배신 가능성을 25%보다 낮게 유지하며 협력 또는 배신을 선택한다.
Gradual	Copycat과 유사하지만, 상대방이 두 번째 배신한 순간, 한 번 대신 두 번을 배신한다. 또한, 세 번째 배신하면 세 번을 연달아 배신한다. 즉, Copycat에 가중처벌의 요소를 가미한다.
Prober	협력-배신-배신으로 시작한다. 상대가 두 번째에 협력, 세 번째에 배신일 경우에는 Copycat으로 전환한다. 그렇지 않다면, 이 과정을 반복한다.
Pavlov	협력으로 시작한다. 이전의 행동이 상대방과 같다면 협력하고, 아니면 배신한다.
Mistrust	Copycat과 같지만, 첫 번째 라운드에 배신으로 시작한다.
Per-Kind	협력-협력-배신을 반복한다.
Per-Nasty	협력-배신-배신을 반복한다.

Modeling

강화학습 환경

게임 상태(State)

게임 시작 전, 배신, 협력을 의미하는 -1, 0, 1을 이용해 배열로 구성
→ 초기 상태 : [-1, -1, -1, -1, ...]
→ 1라운드에 두 참가자가 각각 협력, 배신 : [1, 0, -1, -1, ...]

참가자의 행동(Action)

협력 또는 배신

점수 획득(Reward)

규칙에 기반하여 두 참가자의 행동에 따라 각각 점수 획득

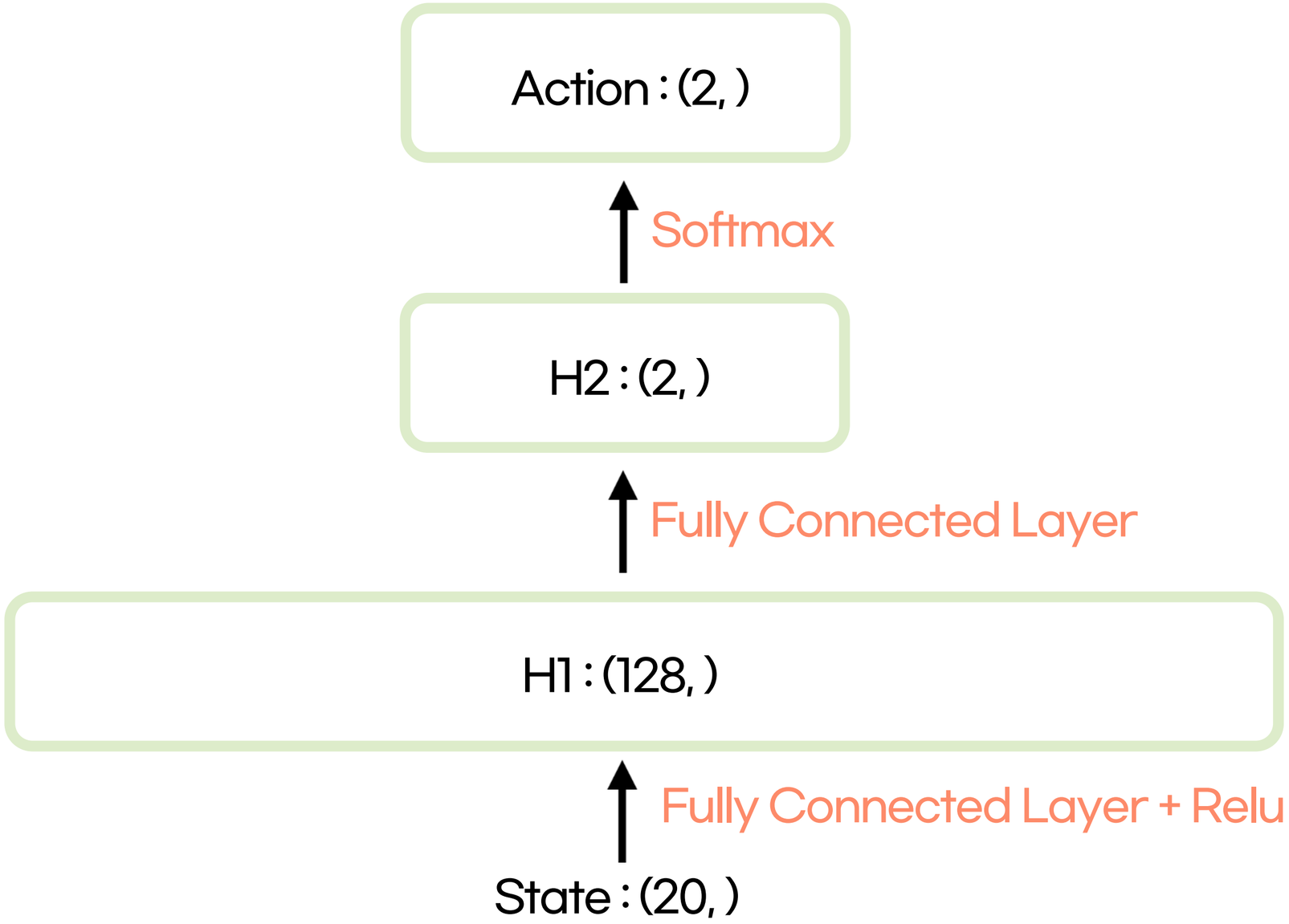
REINFORCE, A Monte-Carlo Policy-Gradient Method (episodic)

Input: a differentiable policy parameterization $\pi(a|s,\boldsymbol{\theta}), \forall a \in \mathcal{A}, s \in \mathcal{S}, \boldsymbol{\theta} \in \mathbb{R}^n$
Initialize policy weights $\boldsymbol{\theta}$
Repeat forever:
  Generate an episode $S_0, A_0, R_1, \ldots, S_{T-1}, A_{T-1}, R_T$, following $\pi(\cdot|\cdot,\boldsymbol{\theta})$
  For each step of the episode $t = 0, \ldots, T-1$:
    $G_t \leftarrow$ return from step t
    $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \alpha\gamma^t G_t \nabla_{\boldsymbol{\theta}} \log \pi(A_t|S_t,\boldsymbol{\theta})$

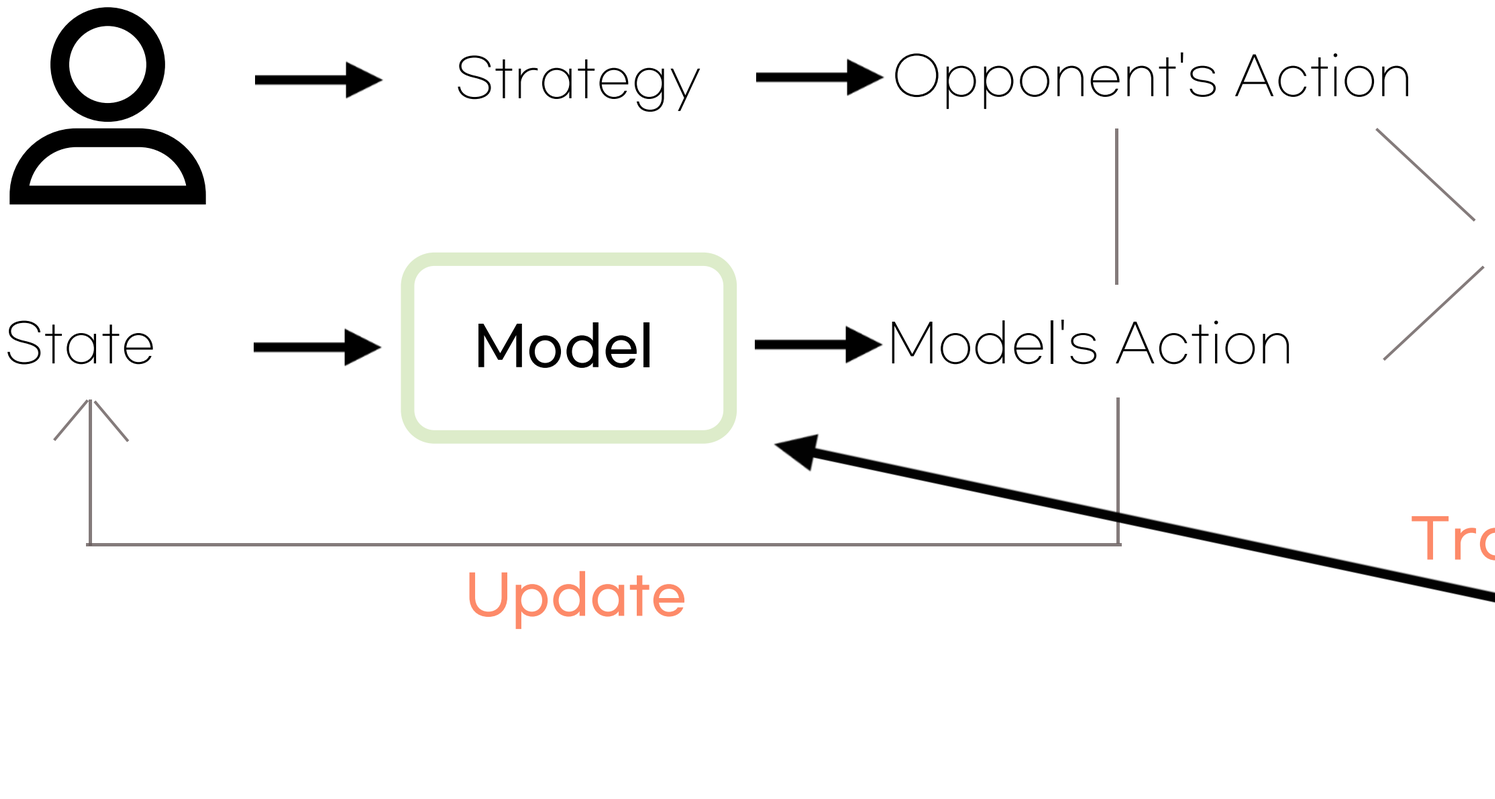
```python
def train_net(self):
    R = 0
    policy_loss = []
    self.optimizer.zero_grad()
    for r, prob in self.data[::-1]:
        R = r + gamma * R
        loss = -torch.log(prob) * R
        policy_loss.append(loss.unsqueeze(0))
    policy_loss = torch.cat(policy_loss).sum()
    policy_loss.backward()
    self.optimizer.step()
    self.data = []
```

학습 알고리즘

모델


Action : (2,)
Softmax
H2 : (2,)
Fully Connected Layer
H1 : (128,)
Fully Connected Layer + Relu
State : (20,)

에피소드



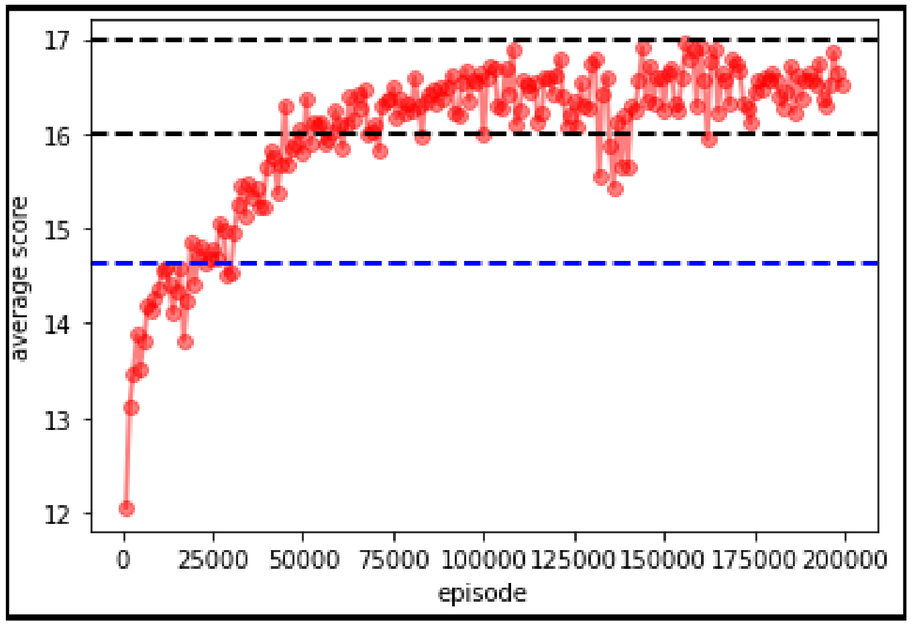
선택	협력	배신
협력	+2 / +2	-1 / +3
배신	+3 / -1	0 / 0

Result

강화학습

Copycat	14.63
Gradual	14.16
Copykitten	14.0
Joss	13.89
Simpleton	13.32

각 전략 별 획득 점수 (상위 5개)

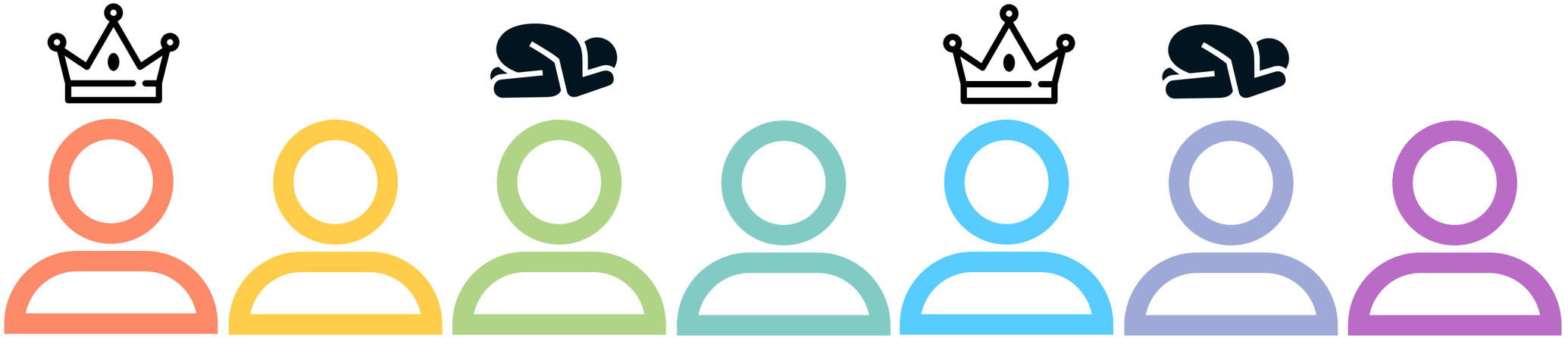


모델의 획득 점수 (최고 점수: 16.959)

신뢰의 진화

게임 진행

신뢰의 진화



획득 점수를 기반으로 상위, 하위 참가자 선정

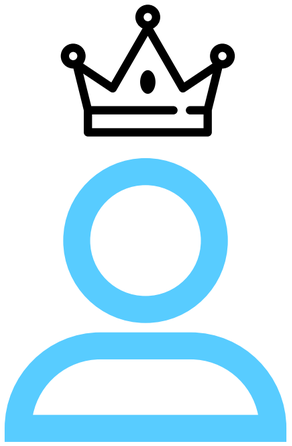
신뢰의 진화

상위 참가자 재생산, 하위 참가자 제거

신뢰의 진화

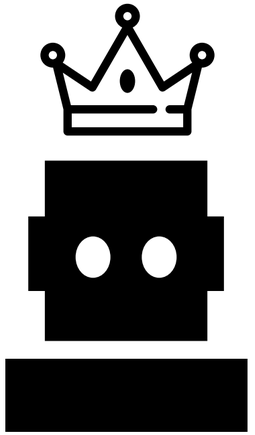
위 과정의 게임을 반복해 최종 생존한 전략 확인

신뢰의 진화



66번의 게임만에
Copycat 전략이 최종 생존

20가지 전략의 참가자끼리 게임 진행



6번의 게임만에
Agent가 최종 생존

20가지 전략의 참가자 + 강화학습 모델(Agent)

Discussion

대응전략

	게임 상태	획득 점수
All Cooperate	[1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1]	21 / 17
All Cheat	[1, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]	-2 / 6
Copycat	[1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1]	21 / 17
Grudger	[1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1]	21 / 17
Detective	[1, 1, 1, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1]	19 / 15
Copykitten	[1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1]	21 / 17
Simpleton	[1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1]	21 / 17
Random	[1, 0, 1, 0, 1, 1, 0, 1, 1, 0, 1, 0, 0, 1, 0, 0, 0, 0, 0, 0]	4 / 12
Cheat-Downing	[1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1]	18 / 18
Cooperate-Downing	[1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1]	21 / 17

	게임 상태	획득 점수
Joss	[1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1]	18 / 18
Cheat-Tester	[1, 0, 1, 1, 1, 0, 0, 0, 0, 0, 1, 1, 0, 1, 0, 1, 0, 1, 0, 1]	14 / 6
Cooperate-Tester	[1, 1, 1, 1, 1, 0, 0, 1, 1, 1, 1, 1, 0, 1, 0, 1, 0, 1, 0, 1]	22 / 6
Tranquilizer	[1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 0, 1, 0, 1, 0, 1, 0, 0, 0, 1]	18 / 10
Gradual	[1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1]	21 / 17
Prober	[1, 1, 1, 0, 1, 0, 0, 0, 0, 1, 0, 1, 0, 0, 0, 1, 0, 1, 0, 0]	12 / 4
Pavlov	[1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1]	21 / 17
Mistrust	[1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1]	18 / 18
Per-Kind	[1, 1, 1, 1, 1, 0, 0, 1, 1, 1, 1, 0, 0, 1, 0, 1, 0, 0, 0, 1]	16 / 8
Per-Nasty	[1, 1, 1, 0, 1, 0, 0, 1, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 1]	9 / 5

각 전략 별 모델의 대응 및 획득 점수

분석

	게임 상태	획득 점수
All Cooperate	[1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1]	21 / 17
All Cheat	[1, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]	-2 / 6
Copycat	[1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1]	21 / 17
Grudger	[1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1]	21 / 17
Detective	[1, 1, 1, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1]	19 / 15
Copykitten	[1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1]	21 / 17
Simpleton	[1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1]	21 / 17
Random	[1, 0, 1, 0, 1, 1, 0, 1, 1, 0, 1, 0, 0, 1, 0, 0, 0, 0, 0, 0]	4 / 12
Cheat-Downing	[1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1]	18 / 18
Cooperate-Downing	[1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1]	21 / 17

	게임 상태	획득 점수
Joss	[1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1]	18 / 18
Cheat-Tester	[1, 0, 1, 1, 1, 0, 0, 0, 0, 0, 1, 1, 0, 1, 0, 1, 0, 1, 0, 1]	14 / 6
Cooperate-Tester	[1, 1, 1, 1, 1, 0, 0, 1, 1, 1, 1, 1, 0, 1, 0, 1, 0, 1, 0, 1]	22 / 6
Tranquilizer	[1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 0, 1, 0, 1, 0, 0, 0, 1]	18 / 10
Gradual	[1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1]	21 / 17
Prober	[1, 1, 1, 0, 1, 0, 0, 0, 0, 1, 0, 1, 0, 0, 0, 1, 0, 1, 0, 0]	12 / 4
Pavlov	[1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1]	21 / 17
Mistrust	[1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1]	18 / 18
Per-Kind	[1, 1, 1, 1, 1, 0, 0, 1, 1, 1, 1, 0, 0, 1, 0, 1, 0, 0, 0, 1]	16 / 8
Per-Nasty	[1, 1, 1, 0, 1, 0, 0, 1, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 1]	9 / 5

각 전략 별 모델의 대응 및 획득 점수

분석

	게임 상태	획득 점수
All Cooperate	[1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1]	21 / 17
All Cheat	[1, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]	-2 / 6
Copycat	[1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1]	21 / 17
Grudger	[1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1]	21 / 17
Detective	[1, 1, 1, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1]	19 / 15
Copykitten	[1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1]	21 / 17
Simpleton	[1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1]	21 / 17
Random	[1, 0, 1, 0, 1, 1, 0, 1, 1, 0, 1, 0, 0, 1, 0, 0, 0, 0, 0, 0]	4 / 12
Cheat-Downing	[1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1]	18 / 18
Cooperate-Downing	[1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1]	21 / 17

	게임 상태	획득 점수
Joss	[1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1]	18 / 18
Cheat-Tester	[1, 0, 1, 1, 1, 0, 0, 0, 0, 0, 1, 1, 0, 1, 0, 1, 0, 1, 0, 1]	14 / 6
Cooperate-Tester	[1, 1, 1, 1, 1, 0, 0, 1, 1, 1, 1, 1, 0, 1, 0, 1, 0, 1, 0, 1]	22 / 6
Tranquilizer	[1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 0, 1, 0, 1, 0, 1, 0, 0]	18 / 10
Gradual	[1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1]	21 / 17
Prober	[1, 1, 1, 0, 1, 0, 0, 0, 0, 1, 0, 1, 0, 0, 0, 1, 0, 1, 0, 0]	12 / 4
Pavlov	[1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1]	21 / 17
Mistrust	[1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1]	18 / 18
Per-Kind	[1, 1, 1, 1, 1, 0, 0, 1, 1, 1, 1, 0, 0, 1, 0, 1, 0, 0, 0, 1]	16 / 8
Per-Nasty	[1, 1, 1, 0, 1, 0, 0, 1, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 1]	9 / 5

각 전략 별 모델의 대응 및 획득 점수

분석

	게임 상태	획득 점수
All Cooperate	[1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1]	21 / 17
All Cheat	[1, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]	-2 / 6
Copycat	[1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1]	21 / 17
Grudger	[1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1]	21 / 17
Detective	[1, 1, 1, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1]	19 / 15
Copykitten	[1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1]	21 / 17
Simpleton	[1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1]	21 / 17
Random	[1, 0, 1, 0, 1, 1, 0, 1, 1, 0, 1, 0, 0, 1, 0, 0, 0, 0, 0, 0]	4 / 12
Cheat-Downing	[1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1]	18 / 18
Cooperate-Downing	[1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1]	21 / 17

	게임 상태	획득 점수
Joss	[1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1]	18 / 18
Cheat-Tester	[1, 0, 1, 1, 1, 0, 0, 0, 0, 0, 1, 1, 0, 1, 0, 1, 0, 1, 0, 1]	14 / 6
Cooperate-Tester	[1, 1, 1, 1, 1, 0, 0, 1, 1, 1, 1, 1, 0, 1, 0, 1, 0, 1, 0, 1]	22 / 6
Tranquilizer	[1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 0, 1, 0, 1, 0, 1, 0, 0, 0, 1]	18 / 10
Gradual	[1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1]	21 / 17
Prober	[1, 1, 1, 0, 1, 0, 0, 0, 0, 1, 0, 1, 0, 0, 0, 1, 0, 1, 0, 0]	12 / 4
Pavlov	[1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1]	21 / 17
Mistrust	[1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1]	18 / 18
Per-Kind	[1, 1, 1, 1, 1, 0, 0, 1, 1, 1, 1, 0, 0, 1, 0, 1, 0, 0, 0, 1]	16 / 8
Per-Nasty	[1, 1, 1, 0, 1, 0, 0, 1, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 1]	9 / 5

각 전략 별 모델의 대응 및 획득 점수

Appendix

향후 연구

CNN, RNN 등을 사용해 게임 상태를 더 잘 이해할 수 있게

모델의 신경망 개선

모델이 더 다양한 전략에 맞서 대응할 수 있게

전략의 다양화

모델이 궁극적으로 추구하는 방향성을 살펴볼 수 있게

점수 획득 규칙의 변화

PPO 등을 사용해 강화학습의 효율을 높일 수 있게

강화학습 알고리즘 발전

참고

- 로버트 액설로드. 「협력의 진화」. 이경식(역). 시스테마, 2009.

- Williams R. J. Simple statistical gradient-following altorithms for connectionist reinforcement learning. *Mach. Learn*, 1992.

- Sutton R., McAllester D., Singh S. & Mansour Y. Policy gradient methods for reinforcement learning with function approximation. In *Advances in Neural Information Processing Systems*, 2000.

Thank you

Source : https://github.com/jujeongho0/RL-PrisonersDilemmaGame