Vinyals et al. 2017: Algorithm

parameters

- π_θ : policy → $\pi_\theta(a_t|s_t)$: do a_t with probability π_θ : conditional only on s_t
- t: time step
- s_t: observation vector
- a_t: action
- r_t: reward
- G_t: future (expected) return with discount factor γ

> $G_t := \sum_{k=0}^{\infty} \gamma^k r_{t+k+1}$

Learning

- A3C: Asynchronous Advantage Actor Critic
  - Mnih et al. 2016: Asynchronous Methods for Deep Reinforcement Learning
  - policy gradient method: approximate gradient ascent on $\mathbb{E}[G_t]$

> := [Policy Gradient] + β [value estimation gradient] +η[entropy regularisation]

  - Policy Gradient $(G_t - v_\theta(s_t))\nabla_\theta \log \pi_\theta(a_t|s_t)$
    - $v_\theta(s)$: value function estimate of the expected return $\mathbb{E}[G_t|s_t = s]$
    - $G_t := \sum_{k=0}^{n-1} \gamma^k r_{t+k+1} + \gamma^n v_\theta(s_{t+n})$ (n-step return: 현실적 이유)
  - Value estimation gradient $(G_t - v_\theta(s_t))\nabla_\theta v_\theta(s_t)$
  - Entropy regularisation $\sum_a \pi_\theta(a|s_t) \log \pi_\theta(a|s_t)$

Policy

- 문제점: 미니맵, 스크린 그래픽 정보는 이산적 → 행동에 대한 결합 분포가 지나치게 많아지는 문제 존재 ⇒ 자기 상관식 (이전기 자신의 파라미터를 현재 값에 결합) 방식으로 표현

> $\pi_\theta(a|s) = \prod_{t=0}^{L} \pi_\theta(a^l|a^{<l}, s)$

- a^0 의 종류에 따라 L은 다를 수 있음. 가령 no-op 의 경우는 $L = 0$ 이지만 move_screen(x,y)는 있음. (Fig3)
- UI 구조로 인해 인간이 선택하지 못할 경우에는 agents도 못하도록 함

Agent Architectures

- Fig4

Input pre-processing

- 각 input feature layer는 각각 동일한 전처리 과정을 거침
- 모든 분류적 값 (질적 변수에 수치를 매핑한 값)을 가진 feature layer는 연속 공간 (부동소수점 벡터)으로 embed함
  - 1x1 convolution을 쓴 것과 동등
- hp나 미네랄같이 큰 값을 가질 수 있는 경우는 log 변환을 통해 re-scale

Atari-net Agent

- Fig4a
- (x,y) 좌표를 선정하는 것과 관련한 공간적 행동들에 독렵적으로 사용
- 아타리 실험에 사용했던 것
- 화면, 미니맵 feature layer(특성 벡터)를 convolutional network로 다룸: 2 layer, 16, 32 filters of size, 4,2 stride
- 비 공간적 특성 벡터는 linear를 기본으로 하되 비선형 부분은 tahn 사용

FullyConv Agent

- Fig4b
- Atari-net 식의 RL은 공간적 차원을
- 본 논문에서 제안
- 기존 agent model은 sc2 같은 복잡한 태스크에 적합하지 않았기 때문
- convolutional LSTM (stacked Long Short-Term Memory network)
- 시각 정보를 2 layer CN (16, 32 filters of size, 5x5, 3x3)
- 미니맵 정보와 화면 정보는 성격이 다른데, 이것은 미래 작업으로 남김.
- 256 unit, ReLU activation, fully connected inear layers
- 공간 행동은 1x1 convolution of the state representation with 1 output channel

FullyConv LSTM Agent

- 위 agent model은 Feed Forword 구조: no memory
- 일부 태스크에는 적합하지만 SC2 의 복잡성 중에는 memory 필요한 것이 있음 ⇒ convolutional LSTM
- fullyConv agent에 LSTM 추가한 것

Random agents

Random policy

- 액션 중 랜덤하게 하나를 택함

Random search

- FullyConv agent 기반

- 돌고 있는 policy networks중 임의의 것을 선정해서 20개중 스코어가 높은 것을 선정
- 보완적 시도