


부산 강화학습

김정주(haje01@gmail.com)

왜 분산(Distribute)해야 하나?

- 강화학습은 많은 시행과 실험이 필요 -> 시간 소요
- 하이퍼파라미터 설정의 어려움
- 다양한 설정의 에이전트로 학습 -> 풍부한 경험
- Spark에서 받은 감동 
 - 10X 속도는 쉽게
 - 강화학습도 분산의 혜택을 볼 수 있다면..

A3C 살짝

소개

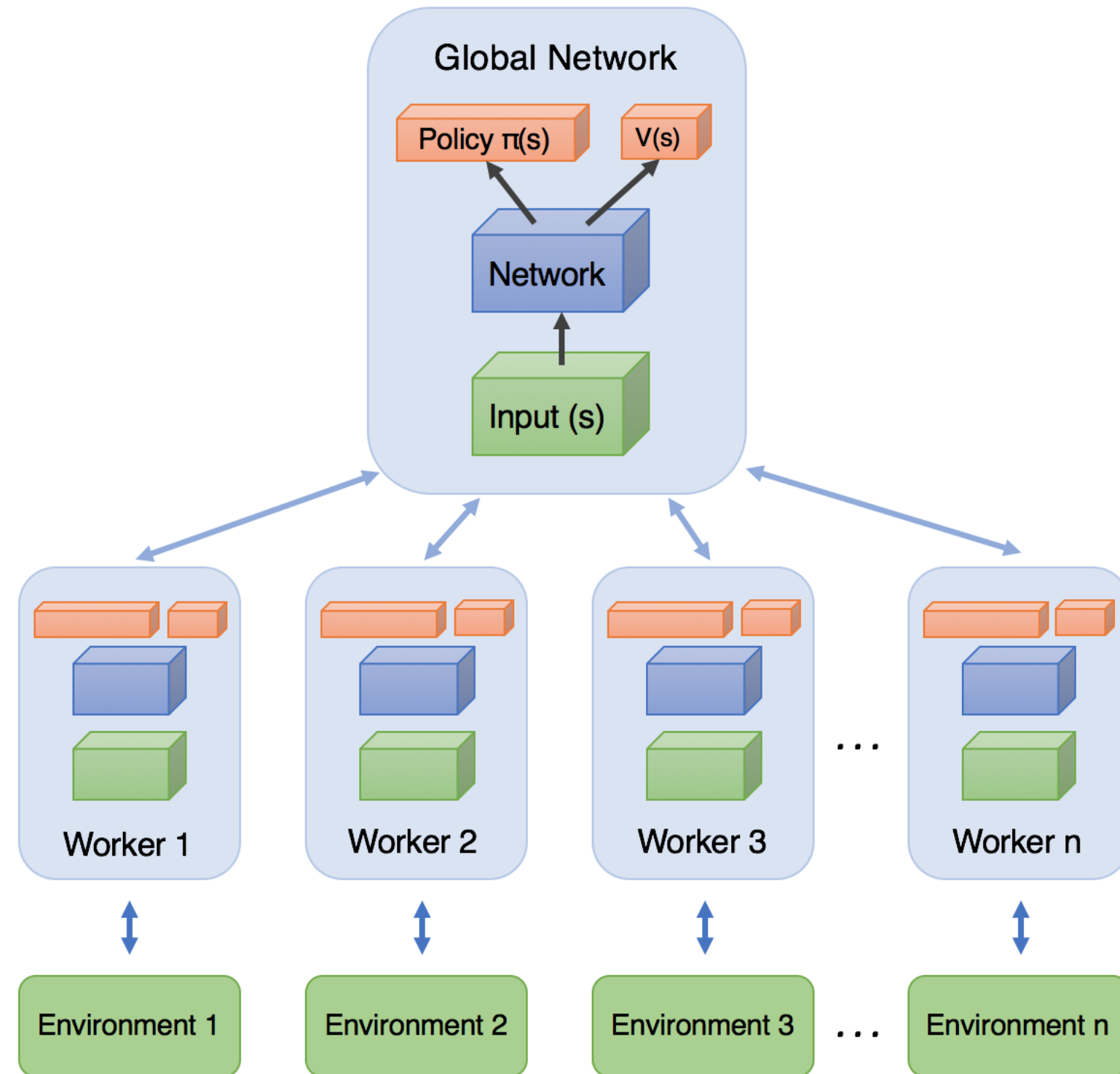
- Asynchronous Advantage Actor-Critic (A3C)
- 더 단순하고, 안정적이고, 우수함
- 이산적이거나 연속적인 동작 공간 모두에 적합
- 복수의 에이전트가 비동기적으로 학습 🎭

특징

- Value-iteration 과 Policy-iteration의 장점을 활용
 - $V(s)$: 상태 가치, $\pi(s)$: 정책
- 단순한 감쇄 리워드 대신 이점(Advantage) 추정 이용 😎
 - $Q(s, a) = V(s) + A(a)$ 로 생각하면,
 - 이점: $A = Q(s, a) - V(s)$ (사실은 Q 대신 R)
 - 단순히 동작의 좋고 나쁨이 아닌, 기대보다 얼마나 좋은지를 기준으로

구성

- 개별 뉴럴넷
 - 특징 추출을 위한 CNN 레이어
 - 연속적 특성 반영을 위한 LSTM 레이어
 - Actor와 Critic을 위한 두 FC 레이어
- 글로벌 네트워크와 로컬 네트워크

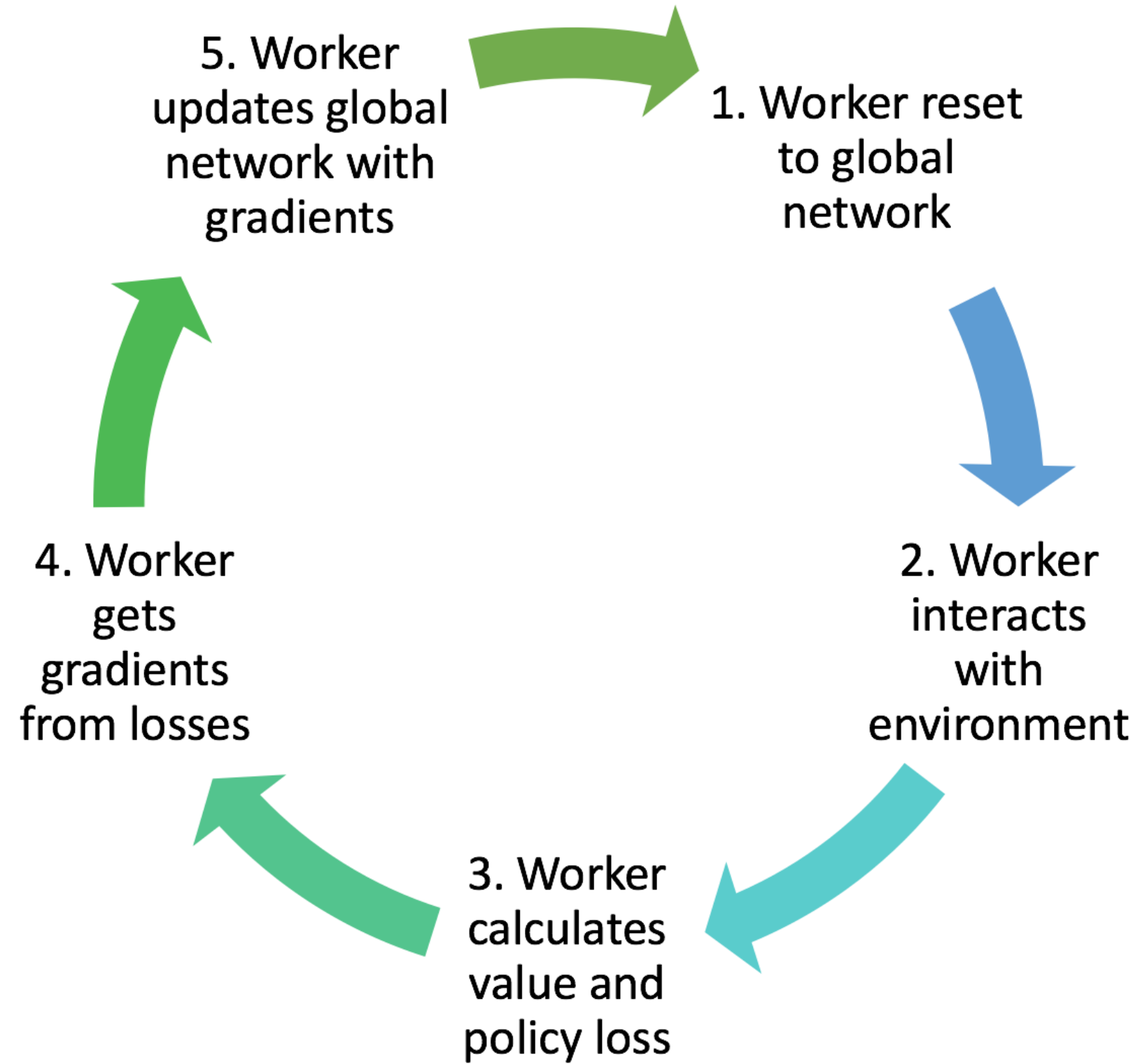


학습

- 각 워커는 독립적으로 시행
- 워커의 경험이 충분히 쌓이면
 - 그것에서 감쇄된 반환값(R)과 이점(A)을 결정
- 이때 정책의 엔트로피(H)도 계산
 - 동작들의 분포가 고르게 퍼져 있으면 엔트로피가 높음
- 엔트로피로 탐험의 정도를 결정

네트워크 동기

- 워커는 두 손실에서 자신의 네트워크 모수에 대한 경사를 구하고
 - 가치 손실: $L = \sum (R - V(s))^2$
 - 정책 손실: $L = -\log(\pi(s))A(s) - \beta H(\pi)$
- 공유된 최적화기를 이용해 경사를 글로벌 네트워크에도 반영
- 갱신된 글로벌 네트워크를 로컬에 복사



Distributed?

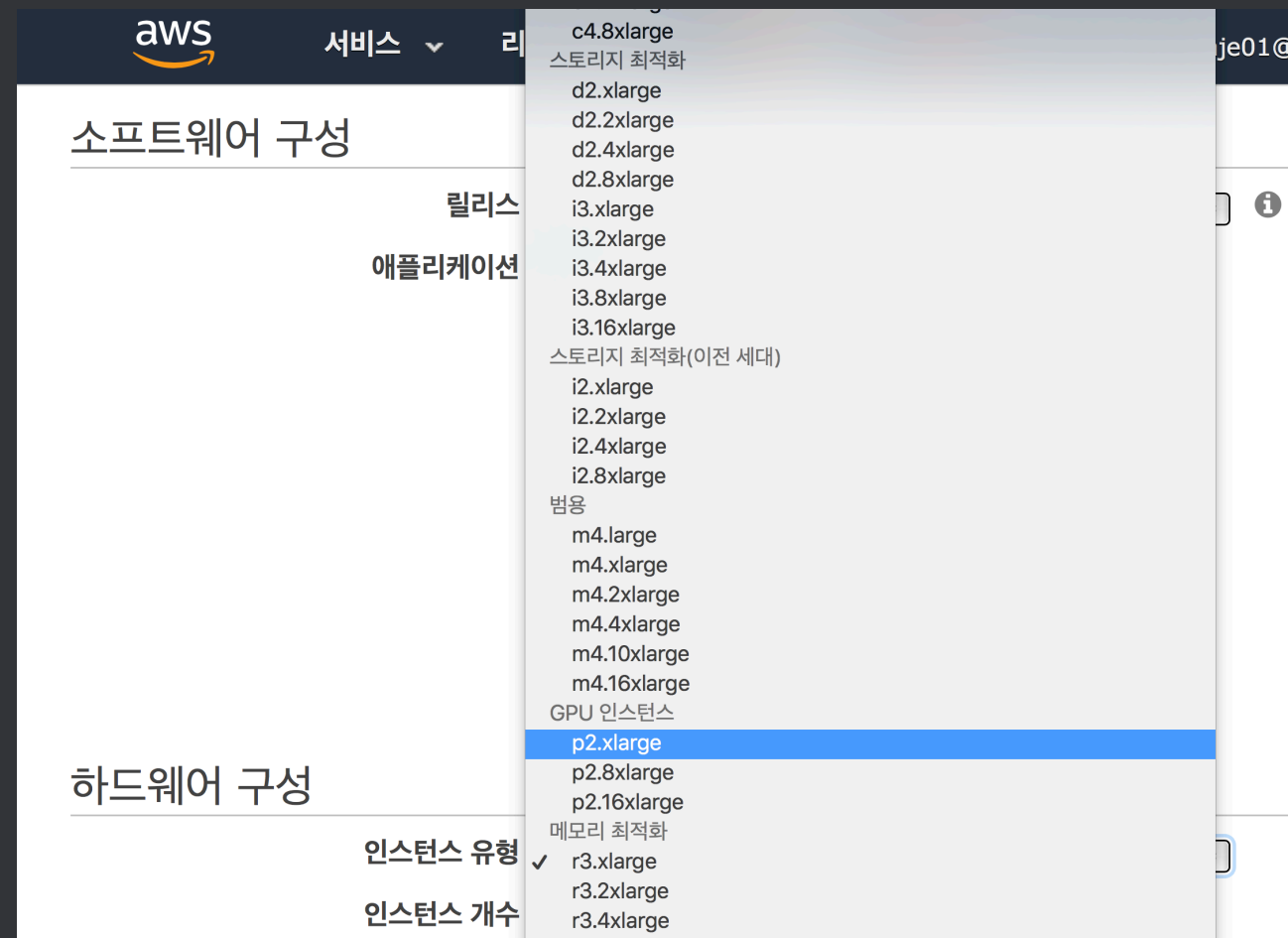
- A3C는 병렬(Parallel) 강화학습
- 진정한 확장성을 위해서는, 여러 컴퓨터에 나누어 처리할 수 있어야.

분산 강화학습 아이디어 🤔

- 분산 GridSearch로 최적 하이퍼패러미터 발견
- 클라우드를 이용해 순간적으로 대량 시뮬레이션하고
- 각 노드의 경험을 합침
 - Tabular한 환경은 상태 방문 수 기준으로 선택
 - 아니면, A3C를 참고하여 노드간 경사 공유

Spark을 활용하자

- 단순 빅데이터 플랫폼이 아닌 범용 분산 컴퓨팅 환경
- GPU기반 Spark도 가능!



분산 강화학습의 과제

- 어떻게 Task를 나눌 것인가?
- 개별 노드의 학습 결과를 어떻게 동기화할 것인가?
- 얼마나 자주 동기화할 것인가?

줄이면서...

- 앞으로는 분산 강화학습이 꼭 필요할 것
- 성공하면 보고하겠습니다~ 😊💧

감사합니다.