

Transformer 기반 계층형 다중 에이전트 강화학습을 이용한 Energy Storage System 운영

<https://github.com/shushusu/r1>

120250390 이승열

■ 목차

1. 프로젝트 주제 및 목표
2. 환경 및 데이터셋
3. State, Action, Reward 설계
4. 강화학습 알고리즘 및 hyperparameter
5. 실험 셋업
6. 실험 결과
7. 토의 및 결론

프로젝트 주제 및 목표

- 프로젝트 주제
 - 태양광 기반 마이크로그리드에서 계층형 멀티에이전트 강화학습을 활용한 Energy Storage System(ESS) 운영 최적화
- 연구 배경
 - 태양광 변동성으로 인한 피크 부하 및 역조류 문제 발생
 - 기존 룰 기반 제어는 다양한 상황 대응에 한계
- 연구 목표
 - 개별 가구 전기요금 절감
 - 전체 피크 부하 및 역조류 감소
 - No-ESS / Rule / 제안 모델 간 성능 비교 및 검증

환경 및 데이터셋

- 데이터셋 구성
 - Kaggle Solar Power Generation Data (Plant 2) 사용
- 전처리 과정
 - DATE_TIME 기준 1시간 단위 리샘플링
 - 22개 가구(Source_Key)를 개별 에이전트로 분리
 - 발전량·부하·기상 데이터를 공통 인덱스 기준 정렬
 - 결측값은 interpolate으로 처리
- 추가 Feature 생성
 - 시간 인코딩: $\sin(2\pi t/24)$, $\cos(2\pi t/24)$
 - TOU(Time-of-Use) 전기요금 생성
 - 미래 정보: 다음 시점 가격·일사량($\text{Price}(t+1)$, $\text{Irrad}(t+1)$)
- 최종 입력 텐서 구조
 - Shape: (Time, Agents, Features=7)
 - Feature 구성 : [Gen, Load, Price, Fut_Price, Fut_Irrad, Sin, Cos]

State, Action, Reward 설계

- State
 - Manager state
 - 마을 평균 발전·부하·가격, 미래 가격/일사량, 시간 인코딩(sin, cos)
 - Worker state
 - 개별 SoC, 발전·부하, 공통 정보, Manager 신호 L_t
- Action
 - Manager action
 - $L_t \in [0,1]$: 각 가구의 방전 상한 신호
 - Worker action
 - $a_i \in [-1, 1]$: 가구별 ESS 충·방전 제어
- Reward
 - 비용 최소화: $-\sum Cost_i$
 - 피크 억제: 피크 초과에 대해 선형 페널티 부과
 - Manager / Worker
 - Manager: 비용 + 피크 글로벌 페널티
 - Worker: 개별 비용 + 완화된 피크 페널티

■ 강화학습 알고리즘 및 hyperparameter

- 사용 알고리즘: Hierarchical PPO
 - Manager : 피크 제어 신호 L_t 생성
 - Worker : 가구별 ESS 충·방전 제어
 - Transformer 기반 정책 네트워크 사용
- Neural Network 구조
 - Encoder : Transformer
 - Actor / Critic: MLP
 - 입력 길이 : Sequence length 24
- 학습 Hyperparameters
 - Learning rate : 3×10^{-4} (linear decay 적용)
 - Gamma(γ) : 0.99
 - GAE(λ) : 0.95
 - Batch size : 256
 - PPO clip range : 0.2
 - Train steps : 50,000 episodes
 - Optimizer : Adam
- 학습 안정화 기법
 - Reward normalization
 - Gradient clipping
 - HistoryBuffer로 시계열 state 유지

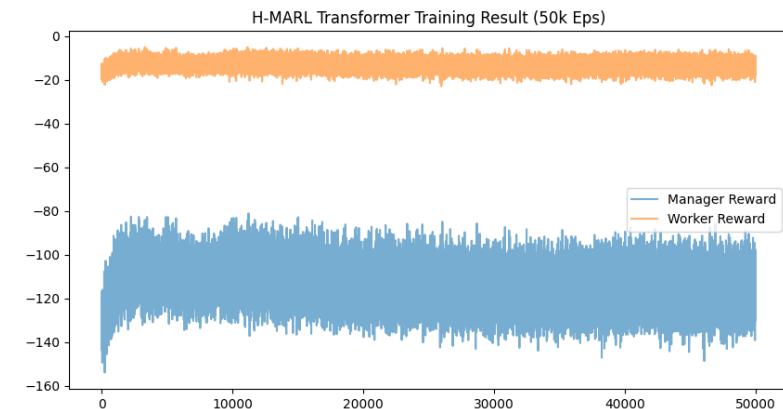
■ 실험 셋업

- 실험 환경
 - 시뮬레이션 환경 : Hierarchical ESS
 - 입력 데이터 : 실제 전력 소비·태양광 발전 기반 24시간 시계열(load, PV, SoC)
 - 평가 에피소드 : 30개 랜덤 시나리오
 - 구현 환경 : Python / PyTorch / CUDA
 - 비교 대상
 - No-ESS(미사용)
 - Rule-based ESS
 - Proposed H-Trans(Transformer PPO)
- Evaluation metric
 - Total Cost
 - 하루 운영 비용 총합 (작을수록 좋음)
 - Peak Ratio
 - 피크 초과 발생 비율 : over_steps/24
 - Max Margin
 - 피크 초과 최대치 (kW)
 - Multi-seed
 - Seed = 0, 1, 2로 학습 후 성능 평균·분산 보고

실험 결과

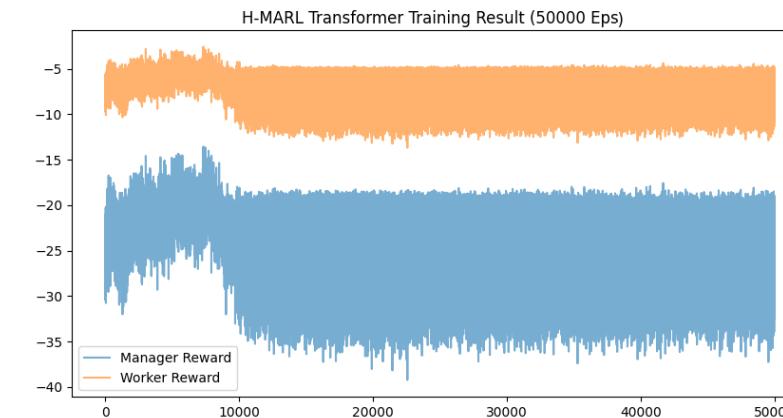
- Reward tuning 전

- Manager/Worker 보상이 낮고 변동성이 큼
- 피크 억제와 비용 절감이 제대로 학습되지 않음
- 기본 보상 구조만으로는 정책 학습이 불안정함



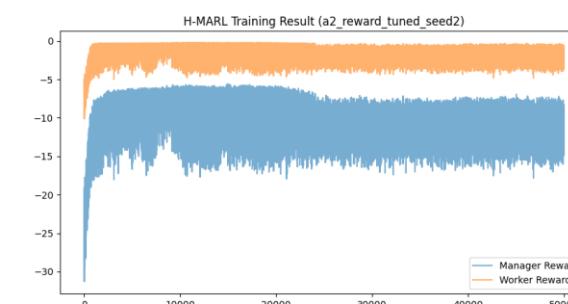
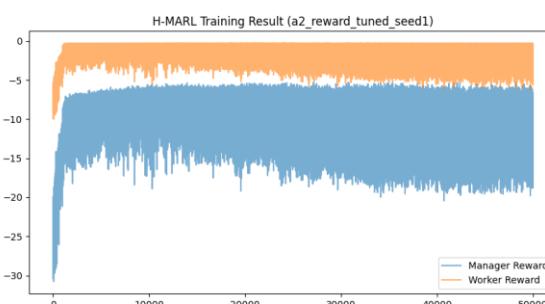
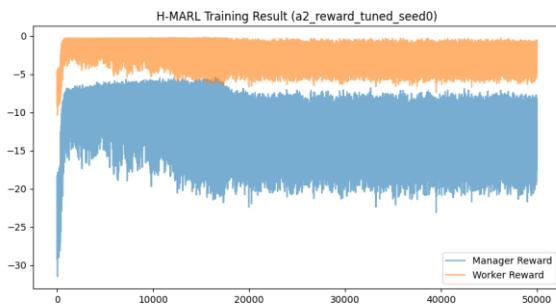
- Reward tuning 후

- 초반 빠른 수렴 후 일정 구간에서 안정
- Worker 보상: 0 부근에서 안정 → 일관된 충·방전 패턴
- Manager 보상: -10~-15 구간에서 수렴 → 피크 억제 정책 학습됨



실험 결과

- Multi-Seed 실험
 - 3개의 서로 다른 seed(0,1,2)에서 모두 안정적으로 수렴
 - 초기 exploration 이후 거의 동일한 형태의 학습 곡선을 보임
 - Worker/Manager 모두 보상 패턴이 일관적으로 재현됨
 - Reward tuning 후 안정적인 학습을 함



■ 토의 및 결론

- 실험 결과
 - Reward tuning 전에는 학습이 불안정하고 수렴이 어렵다는 문제가 있었음
 - Reward 구조를 개선한 실험에서는 모든 seed에서 안정적인 수렴을 확인
 - Manager-Worker 모두 일관된 정책을 학습하며 피크 억제 및 비용 감소 효과가 명확
 - Seed 실험을 통해 정책의 재현성과 신뢰도 확보
- 보완 및 개선사항
 - 현재 보상 함수는 경험적으로 조정 → 자동화된 reward shaping 기법 도입 필요
 - 단일 환경 데이터 기반 → 다양한 계절·기상 조건 데이터로 일반화 성능 검증 필요
 - Transformer 구조가 비용 대비 성능 우수하지만 Inference latency 최적화가 추가로 필요
 - Multi-agent 간 의사소통 구조 강화 등 Hierarchical Multi-Agent Reinforcement Learning 구조 확장 가능성 존재