심층강화학습을 이용한 양자 리피터 체인에서의 얽힘 분배 최적화

석영준[1], 강대건[2], 김진성[2], 한연희[1]
[1]한국기술교육대학교 컴퓨터공학과 미래융합공학전공
[2]한국기술교육대학교 컴퓨터공학과 컴퓨터공학전공
[1]{dsb04163, yhhan}@koreatech.ac.kr
[2]{ahicraft1937, kjs0820k}@koreatech.ac.kr

Entanglement Distribution Optimization in Quantum Repeater Chains via Reinforcement Learning

Yeong-Jun Seok[1], Dae-Gun Kang[2], Jin-Sung Kim[2], Youn-Hee Han[1†]
[1]Future Convergence Engineering, Dept. of Computer Science and Engineering, KOREATECH
[2]Dept. of Computer Science and Engineering, KOREATECH

요 약

양자 네트워크에서의 얽힘 분배는 장거리 양자 통신을 실현하기 위한 핵심 과정으로, 얽힘 생성과 스왑 과정을 통해 두 단말 간 종단간 얽힘을 형성한다. 그러나 스왑 시점을 단순히 가능한 즉시 수행하는 기존의 휴리스틱 방법은 전역적인 자원 효율을 고려하지 않기 때문에 얽힘 생성 시간이 비효율적으로 길어질 수 있다. 본 연구에서는 종단간 얽힘 형성 시간을 단축하기 위해 강화학습을 기반으로 한 얽힘 분배 정책 최적화를 제안한다. 특히 Proximal Policy Optimization (PPO) 알고리즘을 적용하여, 환경과의 상호작용을 통해 스왑 시점을 학습하도록 설계하였다. 시뮬레이션 결과, 제안된 PPO 기반 정책은 기존의 휴리스틱 방법보다 종단간 얽힘을 더 짧은 평균 시간 내에 생성하였으며, 노드 수가 증가하는 환경에서도 안정적인 학습 성능을 보였다. 본 연구는 강화학습이 복잡한 양자 네트워크에서 얽힘 분배 시간을 단축시키는 효과적인 접근법이 될 수 있음을 보여준다.

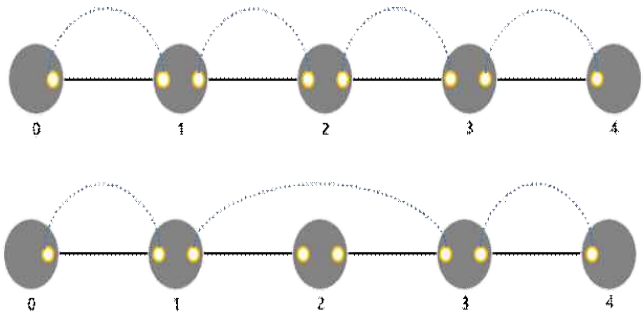
Ⅰ. 서론

양자 네트워크는 얽힘(Entanglement)을 이용해 두 단말 간 양자 상태를 공유함으로써 장거리 양자 통신을 가능하게 한다[1]. 얽힘 분배(Entanglement Distribution)는 인접 노드 간 얽힘을 생성과 중간 노드의 스왑(Entanglement Swapping)을 통해 종단간(End-to-End) 얽힘을 형성하는 핵심 과정이다. 그러나 얽힘 생성과 스왑은 확률적이며, 메모리의 유효 시간 제약으로 스왑 시점이 전체 얽힘 형성 시간에 직접적인 영향을 미친다. 스왑이 너무 빠르면 이후 구간의 얽힘이 준비되지 않아 지연이 발생하고, 너무 늦으면 기존 얽힘이 소멸될 수 있다. 따라서 효율적인 스왑 정책 설계는 양자 리피터 네트워크의 핵심 과제로 꼽힌다.

기존 SWAP-As-Soon-As-Possible (SWAP-ASAP)과 같은 휴리스틱 정책은 단순하지만, 네트워크의 전역 상태를 고려하지 않아 노드 수가 증가할수록 비효율이 누적되어 종단간 얽힘 형성 시간이 길어지는 문제가 발생한다[2].

본 연구에서는 이러한 한계를 해결하기 위해 심층강화학습(Deep Reinforcement Learning, DRL)을 기반으로 한 얽힘 분배 정책 최적화를 제안한다. 특히 정책 기반 강화학습 기법인 Proximal Policy Optimization (PPO)을 적용하여, 환경 상태에 따라 스왑 시점을 동적으로 결정하는 방법을 제안한다[3]. 제안된 PPO 정책은 SWAP-ASAP보다 더 짧은 평균 시간 내에 종단간 얽힘을 형성하였으며, 노드 수가 증가하는 환경에서도 안정적인 성능을 보였다.

Ⅱ. 문제 정의 및 환경 설계

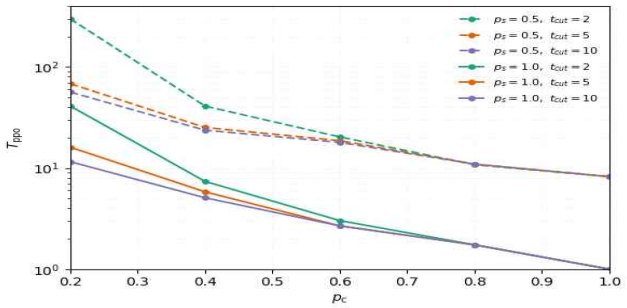


[그림 1] 양자 리피터 체인에서 얽힘 스왑 과정

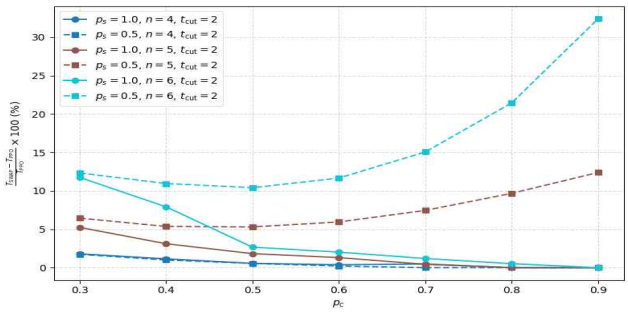
본 연구는 양자 리피터 체인 환경에서 종단간 얽힘 형성 시간을 최소화하는 것이다. 이를 마르코프 결정 과정(Markov Decision Process)으로 모델링했다.

관찰 상태(Observation)는 크기 $n \times n$의 대칭 행렬 o_t로 표현되며, $o_t^{(i,j)} = -1$은 얽힘의 부재, $o_t^{(i,j)} \geq 0$은 얽힘의 나이(Age)를 의미한다. 모든 얽힘의 Age는 1씩 증가되는 Aging이 있다. Age가 메모리 유효시간(Cut-Off Time, t_{cut})에 도달하면 소멸한다.

행동(Action)은 중간 노드$(1, 2, \cdots, n-2)$ 각각에 대해 스왑 실행과 스왑 미실행의 Multi Binary 벡터 a_t로 표현한다. 내부적으로는 선택된 중간 노드 집합을 스왑 후보로 해석하며, 각 후보 노드의 양측 얽힘이 존재해야 유효한 스왑이다. 불가능한 스왑을 사전에 배제하기 위해 마스킹을 실시하고 유효 행동만 탐색한다. 이는 대규모 상태공간에서 불필요한 탐색을 줄여 학습 안정성과 샘플 효율을 높인다.



[그림 2] p_s 및 t_{cut}변화에 따른 얽힘 생성 확률 p_c별 PPO 평균 얽힘 생성 시간(T_{ppo})



[그림 3] $t_{cut} = 2$에서 노드 수 n 및 p_s에 따른 상대 성능 비교.

보상(Reward)은 r_t로 종료 시 종단간 얽힘 생성을 기준으로 정의한다.

$$r_t = \begin{cases} -1 \text{ if } o_t^{(0,n-1)} = -1 \\ 100 \text{ if } o_t^{(0,n-1)} \geq 0 \end{cases}, \quad (1)$$

여기서 $o_t^{(0,n-1)} \geq 0$는 종단간 얽힘이 시간 t에 형성되어 Age를 가진 것을 의미한다. 종단간 얽힘이 관측되지 않는다면, 지속적으로 -1의 패널티를 부여받는다. 이를 통해 종단간 얽힘이 짧은 시간동안에 생성된다면, 높은 보상을 받도록 다음과 같이 설계했다.

상태 전이(Transition)는 시각 t에서 상태 o_t가 행동 a_t가 주어졌을 때, 다음 상태 o_{t+1}로의 확률적 변화로 나타낸다. [그림 1]에서 스왑은 성공 확률 p_s에 따라 적용되며, 성공 시 두 얽힘이 결합되어 새로운 얽힘이 형성된다. 사용된 얽힘은 사라지고, 새로운 얽힘의 나이는 두 얽힘의 Age의 최대값을 사용한다. 또한 물리적으로 인접한 노드 쌍에 확률 p_c로 새로운 얽힘이 생성된다. 이러한 과정은 순차적으로 수행되며, 최종적으로 종단 간 얽힘이 형성되면 종료한다.

Ⅲ. 학습 방법

본 연구에서는 얽힘 분배 정책을 학습하기 위해 PPO를 사용하였다. PPO는 정책 업데이트를 제한하는 클리핑 기법을 적용하여 안정적으로 정책을 개선하고, 종료 시점의 누적 보상을 최대화하는 방향으로 학습을 달성한다. 정책과 가치함수는 동일한 신경망 구조를 공유하며, 파라미터 θ를 클리핑 손실함수로 업데이트한다. PPO의 목적함수는 다음과 같다.

$$L(\theta) = E_t[\min(r_t(\theta)\hat{A}_t, clip(r_t(\theta), 1-\epsilon, 1+\epsilon)\hat{A}_t)] \quad (2)$$

여기서 $r_t(\theta)$는 정책 비율, $\hat{A}_t$는 어드밴티지 추정치, ϵ는 정책 업데이트를 제한하는 파라미터이다.

본 연구에서는 이 목적함수를 기반으로 평균 종단간 얽힘 형성 시간을 최소화하는 방향으로 정책을 학습하였다.

Ⅳ. 실험 결과

본 연구에서는 제안한 PPO 기반 정책의 성능을 평가하기 위해 다양한 n개의 노드로 구성된 양자 리피터 체인 환경에서 실험했다. 환경의 주요 파라미터는 p_c, p_s, t_{cut}이다. 각 실험은 최대 타임스텝은 1000으로 제한하였다. 또한 모든 실험은 100회 반복하여 기록했다.

비교 대상은 기존의 휴리스틱 정책인 SWAP-ASAP이다. 그림 2는 p_s와 t_{cut} 변화에 따른 p_c별 PPO 평균 얽힘 생성 시간 T_{ppo}을 나타낸다. p_c가 증가함에 따라 T_{ppo}가 감소하였으며, 이는 학습된 정책이 더 빠르게 종단간 얽힘을 형성함을 의미한다. 또한 t_{cut}이 길수록 얽힘 소멸로 인한 종단간 얽힘 형성이 빠르게 이루어진다.

그림 3은 n과 p_s에 따른 상대 성능 비교를 나타낸다. PPO 정책은 모든 노드 구성에서 SWAP-ASAP 대비 평균 형성 시간이 짧았으며, 특히 노드 수가 증가할수록 두 정책 간의 성능 차이가 확대되었다. 이는 PPO가 환경 상태에 따라 불필요한 스왑 및 메모리 소멸을 효과적으로 억제하기 때문이다.

Ⅴ. 결론

본 연구에서는 양자 리피터 체인 환경에서 종단간 얽힘 분배 효율을 향상시키기 위해 DRL 기반의 스왑 최적화 방법을 제안하였다. 제안된 PPO 기반 정책은 환경 상태를 고려하여 스왑 시점을 동적으로 결정함으로써, 기존의 SWAP-ASAP 정책보다 더 짧은 시간 내에 종단간 얽힘을 형성하였다. 실험 결과에서 PPO는 다양한 노드 구성에서도 안정적인 결과를 보였으며, 노드 수가 증가하더라도 성능 저하가 발생하지 않았다. 이는 DRL이 단순한 규칙 기반 접근보다 양자 네트워크의 복잡한 상태 변화를 효과적으로 학습할 수 있음을 보여준다. 향후 연구에서는 100개 이상의 노드와 노이즈 모델을 반영하여, DRL 기반 얽힘 분배 정책의 범용성과 실제 환경에 적용 가능성을 검증할 예정이다.

ACKNOWLEDGMENT

이 논문은 2023년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임 (No. NRF-2023R1A2C1003143).

참고문헌

[1] 김용수. (2022). 양자 네트워크를 위한 양자얽힘 광자쌍 생성과 분배. 전자공학회지, 49(8), 25-32.

[2] Li, Jan, et al. "Optimising entanglement distribution policies under classical communication constraints assisted by reinforcement learning." Machine Learning: Science and Technology 6.3 (2025): 035024.

[3] Schulman, John, et al. "Proximal policy optimization algorithms." arXiv preprint arXiv:1707.06347 (2017).