

NVIDIA, AI 디코더로 양자 논리 오류율 최대 347배 감소 달성

(2026.08.24. 양자정보연구지원센터)

□ NVIDIA, AI 디코더로 양자 논리 오류율 최대 347배 감소 달성

○ NVIDIA, AI 기반 양자 오류 정정 디코더 공개

- NVIDIA는 양자 오류 정정(QEC) 분야의 오픈소스 AI 기반 디코더 Ising Decoder ColorCode 1 Fast를 공개했으며, 색상 코드(Color Code)의 디코딩 속도와 정확도를 크게 향상시키는 것을 목표로 함
- 특정 벤치마크에서 기존 오픈소스 색상 코드 디코더인 Chromobius 대비 논리 오류율을 최대 347.7배 감소시키고, 디코딩 속도는 7.3배 향상했다고 발표함
- 이를 통해 기존에는 높은 디코딩 연산량 때문에 실용적인 양자컴퓨팅 구조에 적용하기 어려웠던 색상 코드 기반 아키텍처의 활용 가능성을 확대할 수 있을 것으로 기대됨

○ 양자 오류 정정 및 색상 코드의 필요성

- 실용적인 양자컴퓨터가 상용 수준의 문제를 안정적으로 처리하기 위해서는 계산 과정에서 발생하는 오류를 지속적으로 탐지·수정하는 “결함허용 양자컴퓨팅(Fault-Tolerant Quantum Computing)”이 필수적임
- 양자 오류 정정은 하나의 논리 큐비트를 다수의 물리 큐비트에 분산해 저장하며, 오류 정정 이후 논리 큐비트가 실패하는 빈도를 “논리 오류율(Logical Error Rate)”로 평가함
- 현재 가장 널리 연구되는 “표면 코드(Surface Code)”는 디코딩이 상대적으로 간단하지만 논리 연산 측면에서 한계가 있어, 여러 논리 연산을 효율적으로 수행할 수 있는 “색상 코드(Color Code)”가 대안으로 주목받고 있음
- 색상 코드는 모든 Clifford 게이트를 병렬적으로 수행하는 transversal operation이 가능하고, 격자의 기하학적 대칭성으로

인해 논리 큐비트 간 연산에 사용되는 lattice surgery도 단순화할 수 있다는 장점이 있음

○ 색상 코드의 한계와 NVIDIA의 AI 디코더

- 색상 코드는 연산 효율성 측면의 장점에도 불구하고 표면 코드보다 디코딩 과정의 계산량이 훨씬 커 실용적인 양자컴퓨팅 구조 적용에 제약이 있었음
- NVIDIA의 Ising Decoder ColorCode 1 Fast는 기존 디코더를 대체하는 방식이 아니라 “AI 기반 사전 디코더(pre-decoder)”로 작동함
- 소규모 “3차원 CNN(Convolutional Neural Network)”이 국소적인 오류 패턴을 먼저 식별·수정하고, 남은 오류 문제를 Chromobius에 전달해 최종 디코딩을 수행함으로써 디코딩 부담을 줄임
- 학습에는 NVIDIA의 cuQuantum 및 cuStabilizer를 활용해 합성 양자 오류 정정 데이터를 생성하고, PyTorch를 이용해 신경망을 학습함
- 양자 프로세서의 노이즈 모델, 삼각형 색상 코드의 거리(code distance), 신경망 깊이 등을 설정할 수 있으며, 신경망 층을 늘릴수록 정확도는 높아지는 반면 계산 비용도 증가해 하드웨어에 맞춰 성능과 실행시간을 조정할 수 있음

○ 벤치마크 성능 및 확장성

- NVIDIA는 삼각형 색상 코드를 대상으로 Ising Decoder ColorCode 1 Fast와 Chromobius의 성능을 비교했으며, code distance가 증가할수록 성능 개선 효과가 커진다고 설명함
- 약 code distance 13부터 Chromobius 단독 사용 대비 디코딩 실행시간과 논리 오류율이 동시에 개선되는 전환점이 나타남
- 가장 큰 성능 향상은 code distance 31, 물리 오류율 0.3% 조건에서 나타났으며, 논리 오류율은 347.7배 감소, 디코딩 속도는 7.3배 향상됨
- 해당 결과는 Ising Decoder ColorCode 1 Fast를 NVIDIA DGX

GB300에서 실행하고, Chromobius를 NVIDIA Grace Neoverse-V2 CPU에서 실행해 측정한 결과임

- 공개된 Fast 모델은 약 290만 개의 파라미터와 17개 신경망 계층으로 구성되어 있으며, 비교적 작은 모델 크기를 바탕으로 GPU에서 효율적으로 실행하면서 양자 오류 정정에 필요한 낮은 지연 시간을 확보하도록 설계됨
- 신경망이 전체 양자 프로세서 크기에 의존하지 않는 국소 구조이므로 더 큰 code distance 및 실시간 오류 정정에 필요한 blockwise decoding 구조로 확장할 수 있다고 NVIDIA는 설명함

○ 오픈소스 공개 및 양자컴퓨팅 생태계 영향

- NVIDIA는 디코더뿐 아니라 전체 학습 파이프라인을 오픈소스로 공개해 모델 가중치, 합성 데이터 생성 도구, 학습 방법, 벤치마크 데이터셋 및 관련 문서를 제공함
- 개발자는 자신의 양자 프로세서가 갖는 노이즈 특성에 맞춰 디코더를 재학습하거나 수정할 수 있으며, 사전에 준비된 데이터셋이 아니라 학습 과정에서 합성 오류 정정 데이터를 생성할 수 있음
- 이번 공개는 AI를 양자 알고리즘 자체에 적용하기보다 교정·제어·오류 완화·디코딩 등 양자컴퓨터의 안정적인 운용을 지원하는 엔지니어링 영역에 AI를 적용하는 흐름을 보여줌
- 또한 표면 코드 중심의 기존 양자 오류 정정 연구에서 벗어나 색상 코드, 양자 LDPC 코드 등 다양한 오류 정정 방식을 재검토하는 계기가 될 수 있음
- 향후 디코더의 속도와 정확도가 지속적으로 향상될 경우, 높은 디코딩 복잡성 때문에 실용화에서 제외됐던 색상 코드 기반 결합 허용 양자컴퓨팅 아키텍처의 재평가 및 활용 확대가 가능할 것으로 전망됨

(원문)

1. <https://thequantuminsider.com/2026/07/15/nvidia-says-ai-decoder-achieved-up-to-347-x-cut-in-quantum-logical-error-rates/>