이화여자대학교
EWHA WOMANS UNIVERSITY

2025 – Summer

DQN의 2048 게임 적용에 대한 한계 분석 및 개선 가능성 탐색

강화시스터즈 세션 01

장예원, 최상아, 한사랑

목차

Ⅰ. 서론

1. 연구 배경

- Q-learning: 상태-행동의 누적 보상(Q-value) 기반 학습
- 전통적 Q-learning의 한계: 작은 상태 공간에서만 사용 가능
- 21세기 이후: 이미지·동영상 등 고차원 비정형 데이터를 딥러닝으로 처리 → 강화학습 확산

2. 연구 필요성

- 고차원 데이터를 처리하기 위해 Q값 근사 + 경험 저장 가능한 신경망 기반 Q-learning(DQN) 필요

3. 연구 목적

- 2048 게임을 강화학습으로 해결
- 4x4 보드 상태공간에서 DQN 알고리즘의 성능을 실험적으로 검증

Ⅰ. 서론

4. 연구 목표

- 퍼즐 게임 2048 환경에 심층 Q-네트워크(DQN) 적용
- 다양한 성능 개선 요인 실험적 검증
  - Double DQN
  - Dueling DQN
  - Prioritized Experience Replay (PER)
  - ε-탐욕 정책 수렴 속도
  - 보상 설계 (Reward Shaping)

5. 최종 목표

- 여러 방법론을 비교·분석하여 2048에서 가장 안정적이고 효율적인 DQN 구조 제안

Ⅱ. 관련 연구

1. 사례 개요

- 분석 대상: YangRui (2015)의 2048 DQN 공개 프로젝트(2048_env)
- 목적: 단순 고득점 달성 → DQN 한계 극복 & 안정적 학습 구조 탐구
- 방법: CNN 기반 상태 표현 + 알고리즘 개선 기법 종합 적용

2. 문제 접근 방식

CNN 상태 표현	보상 설계 (Reward Shaping)	알고리즘 개선
4x4 보드를 CNN 입력으로 사용 → 타일 위치·값 특징 자동 추출	단순 점수 보상 → 불안정 log2(1+r) 보상 적용 → 변동 억제, 학습 안정화	PER: 예측 실패 경험 우선 학습 → 효율 향상 Double DQN: Q-value 과대평가 완화 Target Network soft update, Gradient clipping → 안정성 확보

Ⅱ. 관련 연구

3. 결과 분석

- 학습 성능
  - 40,000 에피소드 학습 → reward/steps 꾸준히 상승
  - Loss 급격히 감소 후 안정적 유지
- 최고 타일
  - 장기간 512 타일 → 후반부 1024 달성 ('퀀텀 점프')
- 평가 성능
  - 평균 점수: 약 5,100
  - 최대 점수: 11,500

4. 종합 및 시사점

성공적인 DQN 적용 YangRui (2015)의 사례는 2048 게임에 DQN을 성공적으로 적용한 대표적인 연구로 평가됨	**다양한 기법의 조합** 단일 기법보다는 CNN, PER, Double DQN, 보상 설계 등 여러 개선 기법의 조합이 성능 향상에 크게 기여함
안정성과 성능 동시 확보 이러한 통합적 접근 방식은 학습 안정성과 게임 성능을 동시에 확보하는 데 핵심적인 역할을 함	**확장 가능성 제시** 이는 2048과 같은 복잡한 환경에도 강화 학습 알고리즘의 성공적인 확장이 가능함을 시사함

Ⅲ. 연구 방법

1. 환경 (Environment)

- 직접 구현한 Gym 스타일 2048 환경
  - OpenAI Gym 규약 준수: reset, step, render 지원
  - 다양한 DQN 알고리즘 비교 가능하도록 설계

2. 상태 표현 (State Representation)

- 4x4 보드를 원-핫 인코딩 방식으로 변환
- {2, 4, ···, 2^16} → 16개 채널에 투영
- 결과: 4x4x16 텐서 형태 입력
- CNN이 고차원 상태를 효과적으로 처리하도록 설계

III. 연구 방법

3. 행동 정의 (Action Space)

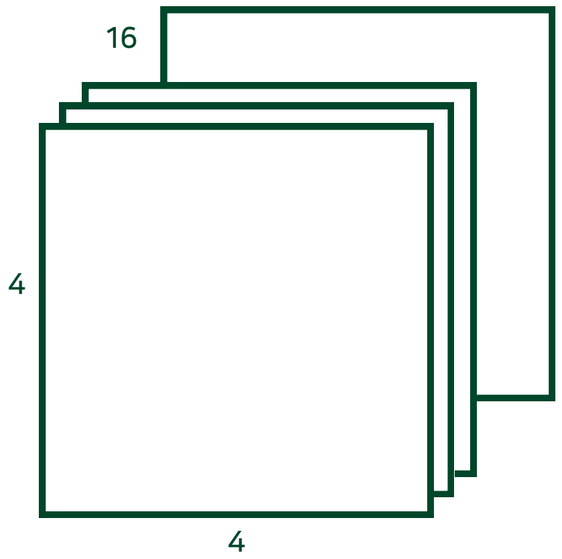
- 상/하/좌/우 4가지 행동
- **합법적 이동**: 타일 병합 + 무작위 신규 타일 추가 (90%: 2, 10%: 4)
- **불법 이동**: 보드 변화 없음 → 보상 0, 새로운 타일 미생성
- 스텝 수는 정상적으로 증가 (불필요 시도 누적 방지)
- Action Mask 제공 → 무효 행동 차단

4. 보상 함수 설계 (Reward Shaping)

- **병합 보상**: 이동 시 병합된 타일 합을 로그 스케일로 변환
- **빈 칸 보상**: 현재 보드의 빈 칸 개수를 로그 스케일로 보상에 반영
- 단조성(monotonicity), 평탄성(smoothness) 등 휴리스틱은 제외
- **목적**: 단순한 과적합 방지 + 학습 안정성 강화

Ⅲ. 연구 방법 I CNN

입력 이미지



2	[1, 0, 0, ..., 0]	2에 대응하는 채널이 1, 나머지 0
4	[0, 1, 0, ..., 0]	4에 대응하는 채널이 1, 나머지 0
8	[0, 0, 1, ..., 0]	8에 대응하는 채널이 1, 나머지 0

CNN	conv1, conv2, conv3
각 계층	128층
커널 크기	3x3
활성화 함수	ReLU
FCL	총 2개
은닉층 노드수	528개
출력층 output	Q 값 4개 (상/하/좌/우)
가중치 초기화	He

Ⅲ. 연구 방법 | Dueling & Double DQN

마지막 출력층에서

$$Q(s,a) = V(s) + A(s,a) - \frac{1}{|\mathcal{A}|}\sum_{a'} A(s,a'),$$

Value stream (V(s)) 단일 스칼라 값으로 추정

Advantage stream (A(s,a)) 행동의 상대적 우위를 추정

•••••

Q-value
- 온라인 네트워크
- 타겟 네트워크

Ⅲ. 연구 방법 | Replay Buffer

Replay Buffer

균등 선택

Prioritized Experience Replay

우선 순위(TD-Error) 기반 선택

샘플링 편향 보정

Ⅲ. 연구 방법 | 학습 조건

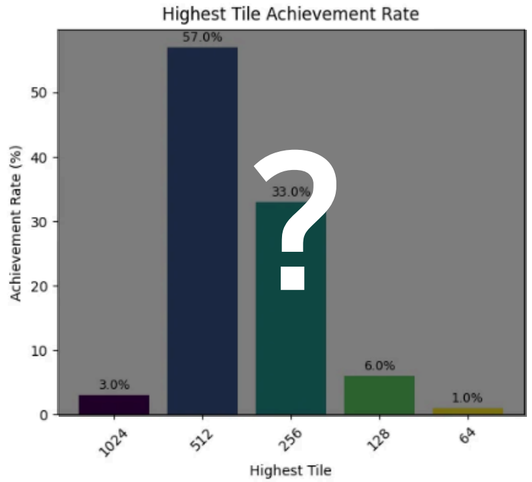
학습 반복	3,000회
최적화 기법	Adam
평가/로딩	10 episodes

하이퍼파라미터

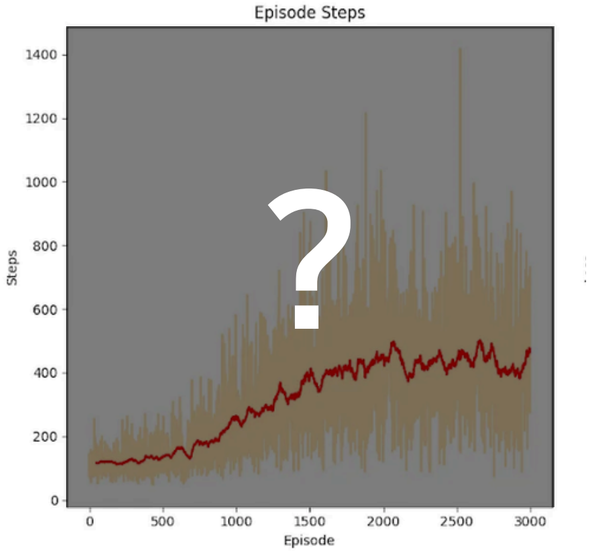
Replay Buffer	100,000
배치 사이즈	64
γ	0.99
Learning Rate	0.0001
ϵ_start, ϵ_end	0.9, 0.01
Decay	30,000
타겟 네트워크 동기화	1,000 steps

Ⅲ. 연구 방법 | 평가 지표

최고 타일



평균 점수



에피소드 길이

Ⅲ. 연구 방법 I 평가 방법

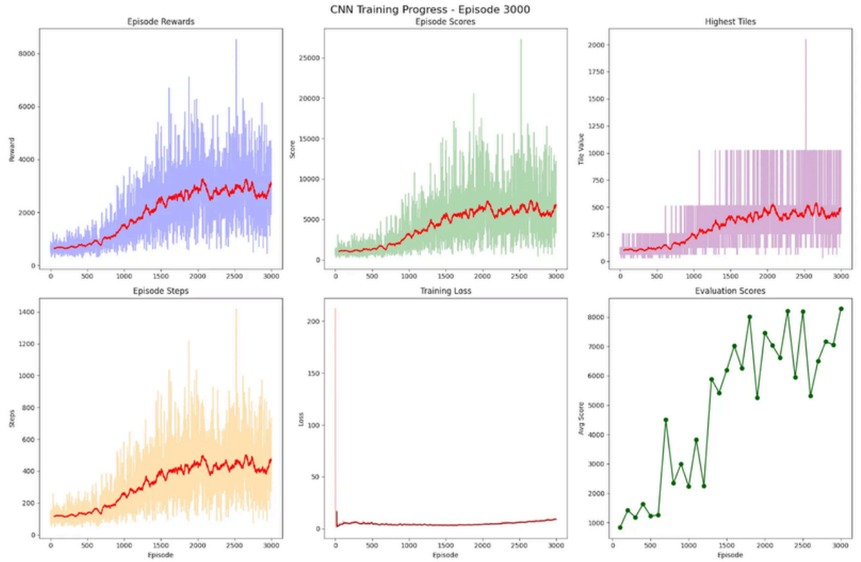
Greedy Policy

+

주기적 평가

- 기본 DQN
- Double DQN
- Dueling DQN
- 우선순위 경험 재생(Prioritized Replay, PER)
- 액션 마스킹 기법

보상, 점수, 최고 타일, 스텝 수, 손실 값



Ⅳ. DQN 성능 개선 요인 (1)

측정

~~Double DQN~~ | ~~Dueling DQN~~	2,241.0	7,048	198.4	512
Double DQN | ~~Dueling DQN~~	2,270.2	6,440	202.9	512
~~Double DQN~~ | Dueling DQN	6,007.2	15,776	426.2	1,024
Double DQN | Dueling DQN	7,403.1	17,456	506.9	1,024

실험 4 > 실험 3 > 실험 2 > 실험 1

실험

실험1 v.s. 실험2 제한적인 성능 향상

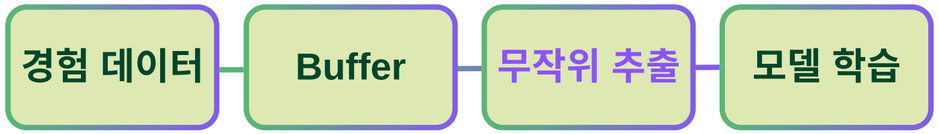
실험1 v.s. 실험3 2241.0 ▶ 6007.2 (2.7배), 198.4 ▶ 426.2로 (2배), **1024 안정적 성취**

실험3 v.s. 실험4 6007.2 ▶ 7403.1(23%), 426.2 ▶ 506.9, 두 기법의 **시너지**

Ⅳ. DQN 성능 개선 요인 (2)

Replay Buffer

핵심 컨셉: 무작위 균등 샘플링

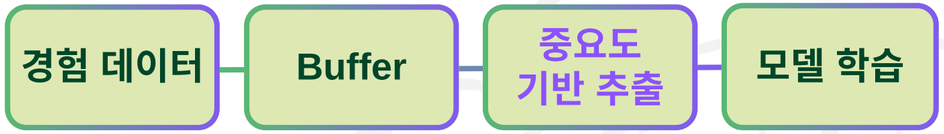


한계점:

- 덜 중요한 경험과 중요한 경험을 동일하게 취급
- 시간적 상관관계 문제 해결, 그러나 학습 효율이 not 최적

Prioritized Experience Replay (PER)

핵심 컨셉: 중요도(TD-Error) 기반 샘플링



장점:

- "예측과 실제의 차이가 큰" 놀라운 경험 집중 학습
- 학습 효율을 극대화하여 더 빠르고 높은 성능 달성

실험 결과:

PER의 압도적인 성능 우위

"학습 효율의 차이가 최종 성능의 차이를 만든다."

지표	Replay Buffer	Prioritized Replay
평균 점수	5831.4	8224.6 (▲ 41%)
평균 최고 타일	415.4	572.2 (▲ 38%)
1024 타일 점유율	1x	~ 7x

Ⅳ. DQN 성능 개선 요인 (3)

탐험 (Exploration)
- 개념: 새로운 가능성을 찾기 위한 무작위 행동
- 목표: 더 나은 미지의 보상을 발견 (장기적 이득)

VS

활용 (Exploitation)
- 개념: 현재까지의 최선책을 따르는 결정론적 행동
- 목표: 이미 알고 있는 최대의 보상을 획득 (단기적 이득)

-> Epsilon(ε)은 '탐험'을 할 확률,
-> Epsilon Decay는 학습이 진행됨에 따라 점차 '활용'의 비중을 높이는 전략

실험: 세 가지 Epsilon Decay Rate 비교

지표	A: 느린 수렴	B: 빠른 수렴	C: 기본 수렴
decay 값	50000	15000	30000 (기본)
e start/end	1.0 / 0.05	0.8 / 0.01	0.8 / 0.005
그래프 양상	학습 속도 더딤, 성능 불안정	주기 수렴, 더 좋은 전략 발견 실패	🏆 꾸준한 성능 향상, 안정적인 최고 성능 달성

실험 결과:
기본 ε 설정이 가장
효과적인 탐험 전략임을 확인.

Ⅳ. DQN 성능 개선 요인 (4)

실험: 세가지 단계별 보상 설계

단순 접근

- 전략: 오직 점수만 (Score-based)
- 수식: Reward = 누적 점수
- 문제점:
  - 초기 보상이 너무 희소(Sparse)
  - 학습 불안정 및 낮은 최종 성능

구조적 접근

- 전략: 점수 + 예쁜 모양 (Structure-based)
- 수식: Reward = 점수 + 단조성 + 평탄성
- 문제점:
  - '모양'과 '점수'의 목표가 상충하여 학습 방해
  - 뚜렷한 성능 개선 효과 없음

전략적 접근 (Best)

- 전략: 점수 + 공간 확보 (Strategy-based)
- 수식: Reward = log(병합 점수) + w * (빈칸 수)
- 기대 효과:
  - 타일 병합: 즉각적인 행동에 대한 명확한 보상
  - 공간 확보: 장기적인 생존 가능성 보상

실험 결과:

게임의 핵심 전략(타일 병합 + 공간 확보)를
직접적으로 반영한 보상 함수가 가장 효과적임을 실험적으로 입증함.

V. 결론

A. 구조 (Architecture)

키워드: Double DQN / Dueling DQN
효과:
✓ Q-value 과대평가 완화
✓ 학습 안정성 및 성능 대폭 향상

B. 경험 활용 (Memory)

키워드: Prioritized Experience Replay
효과:
✓ 중요 경험 데이터 집중 학습
✓ 학습 효율성 및 수렴 속도 개선

C. 탐험 전략 (Exploration)

키워드: Epsilon Decay Rate
효과:
✓ 탐험-활용간 균형 확보
✓ 안정적인 최고 성능 달성

D. 보상 설계 (Reward)

키워드: Reward Shaping
효과:
✓ 단기적 보상과 장기적 생존 동시 고려
✓ 고차원적 전략 학습 유도

이화여자대학교
EWHA WOMANS UNIVERSITY

2025 – Summer

Thank you!

강화시스터즈 세션 01

장예원, 최상아, 한사랑