Generating Text with Deep Reinforcement Learning

MAKING	**TEXT**	**WITH**	BIG	NEURAL	TECNIQUE
GENERATE	BOOK	OR	**DEEP**	SEA	NETWORK
GENERATING	LYRICS	AND	LIGHT	LEARNING	**LEARNING**
READING	WEB	FOR	HEAVY	**REINFORCEMENT**	PATCH
BUILDING	CAKE	THAT	SMALL	NETWORK	ITEM

Sungwon Moon

2020. 06. 17

he hit me with a banana

문장을 발전시킬 수 있는 여지가 있다!

= 강화학습 ! DQN !

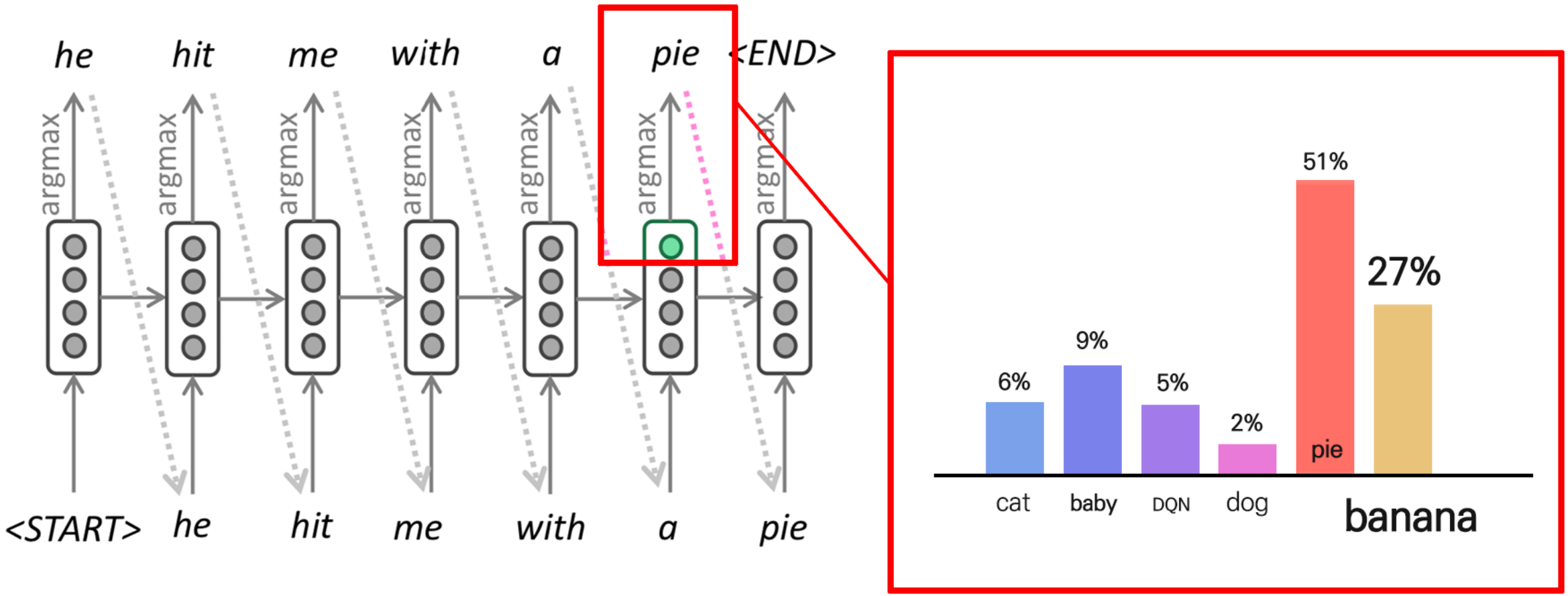
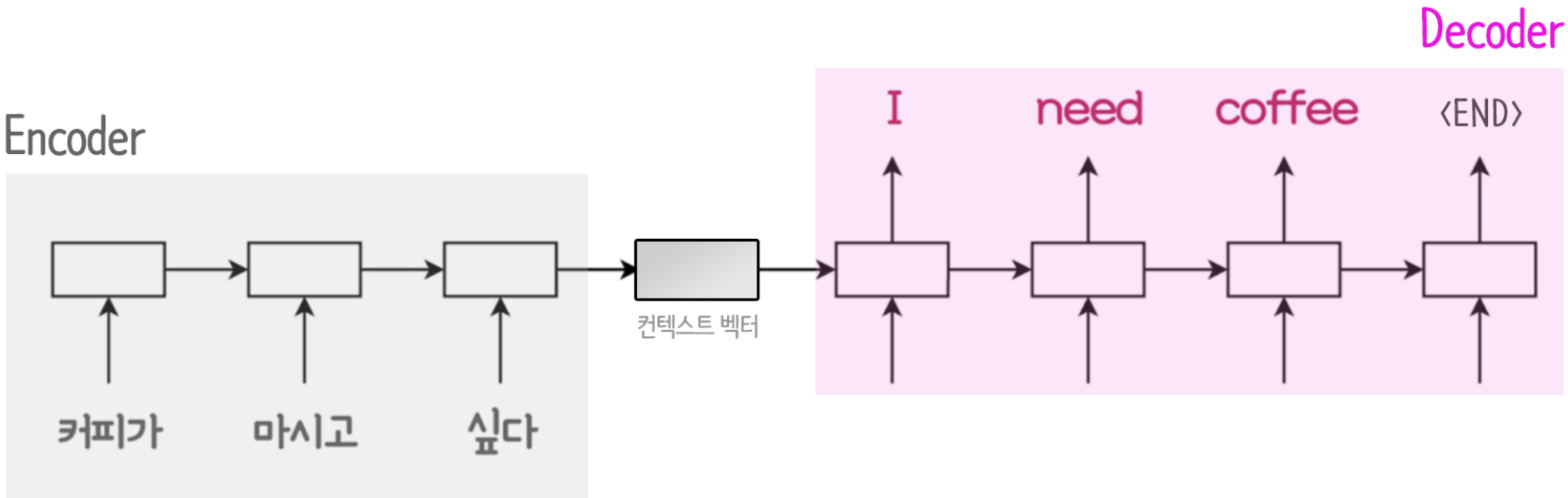


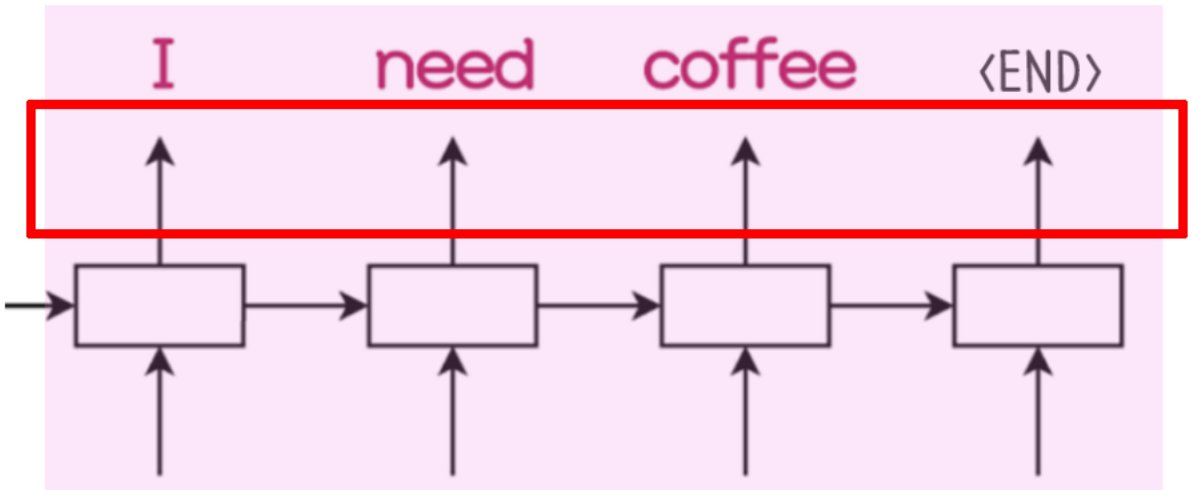
Image from CS224n Lecture 15 Natural Language Generation

기존의 방법이 대체 어땠길래?

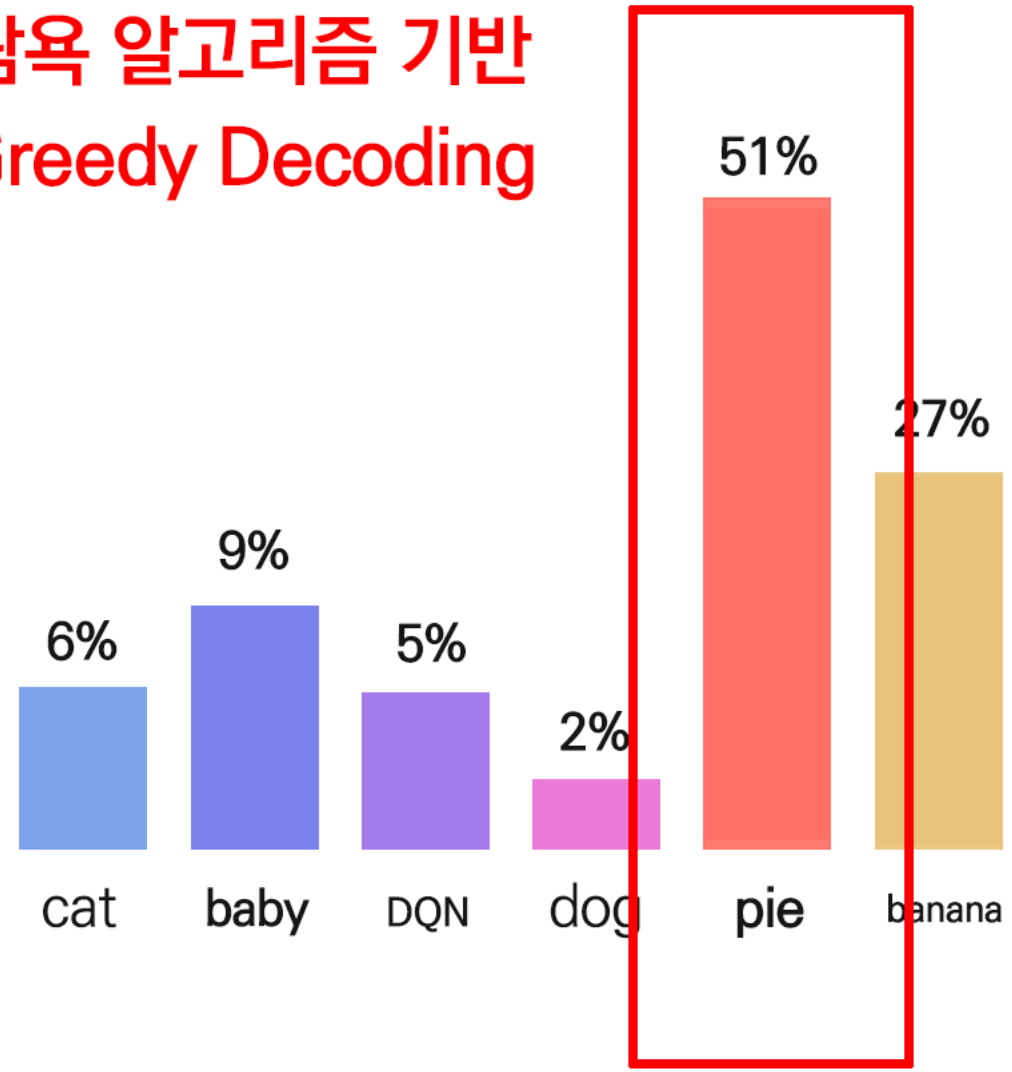


▲ Sequence-to-Sequence (Seq2seq)

문장을 생성하는 방법

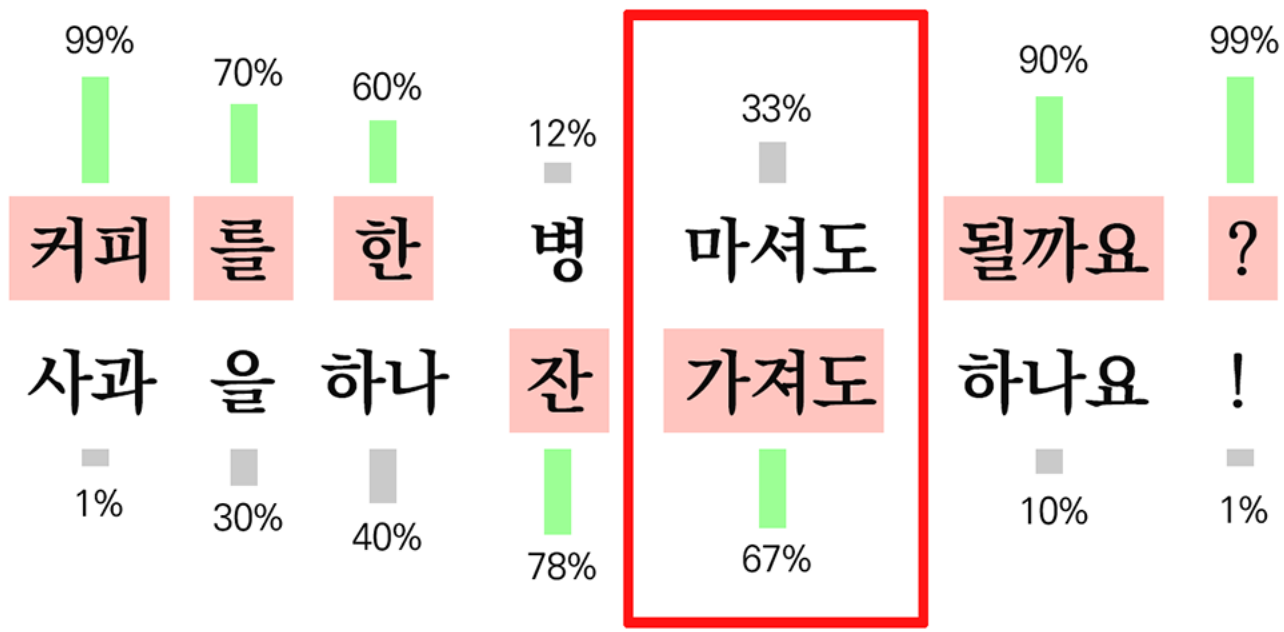


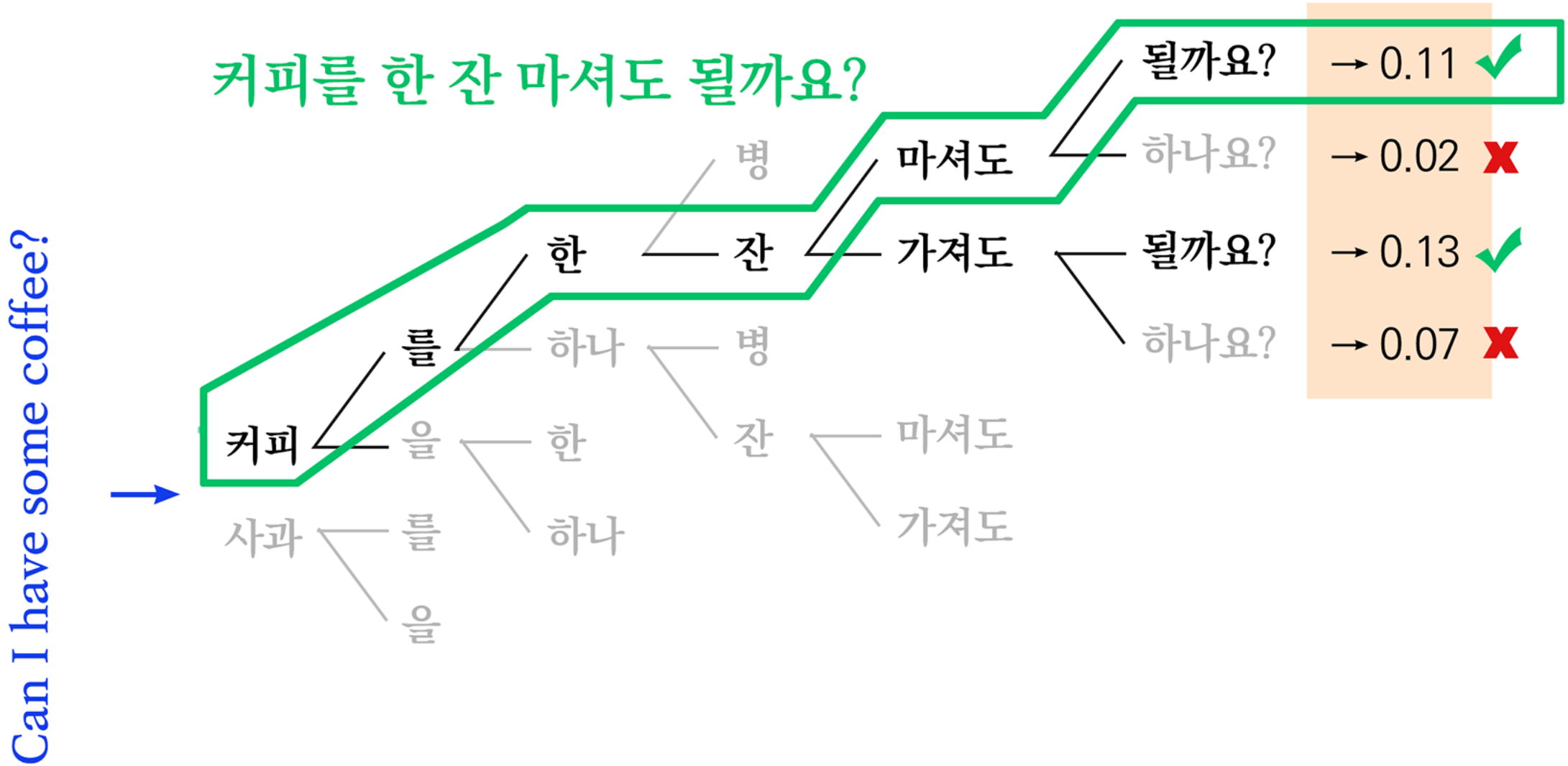
탐욕 알고리즘 기반
Greedy Decoding



문장을 생성하는 방법

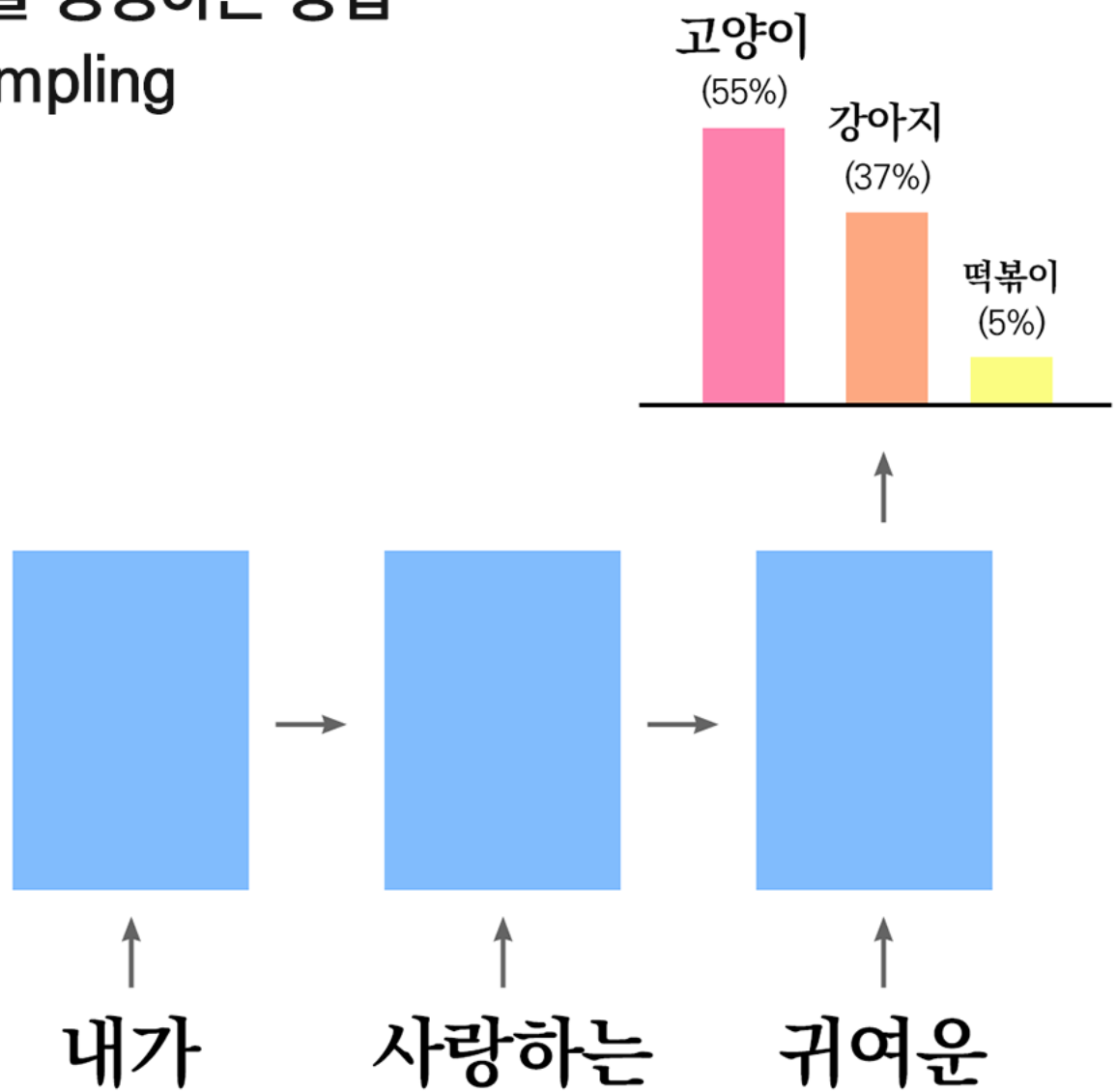
Can I have some coffee? → 커피를 한 잔 가져도 될까요?





문장을 생성하는 방법
- Beam Search

문장을 생성하는 방법
– Sampling



100개 문장을 생성하면...

내가 사랑하는 귀여운 고양이 x 55

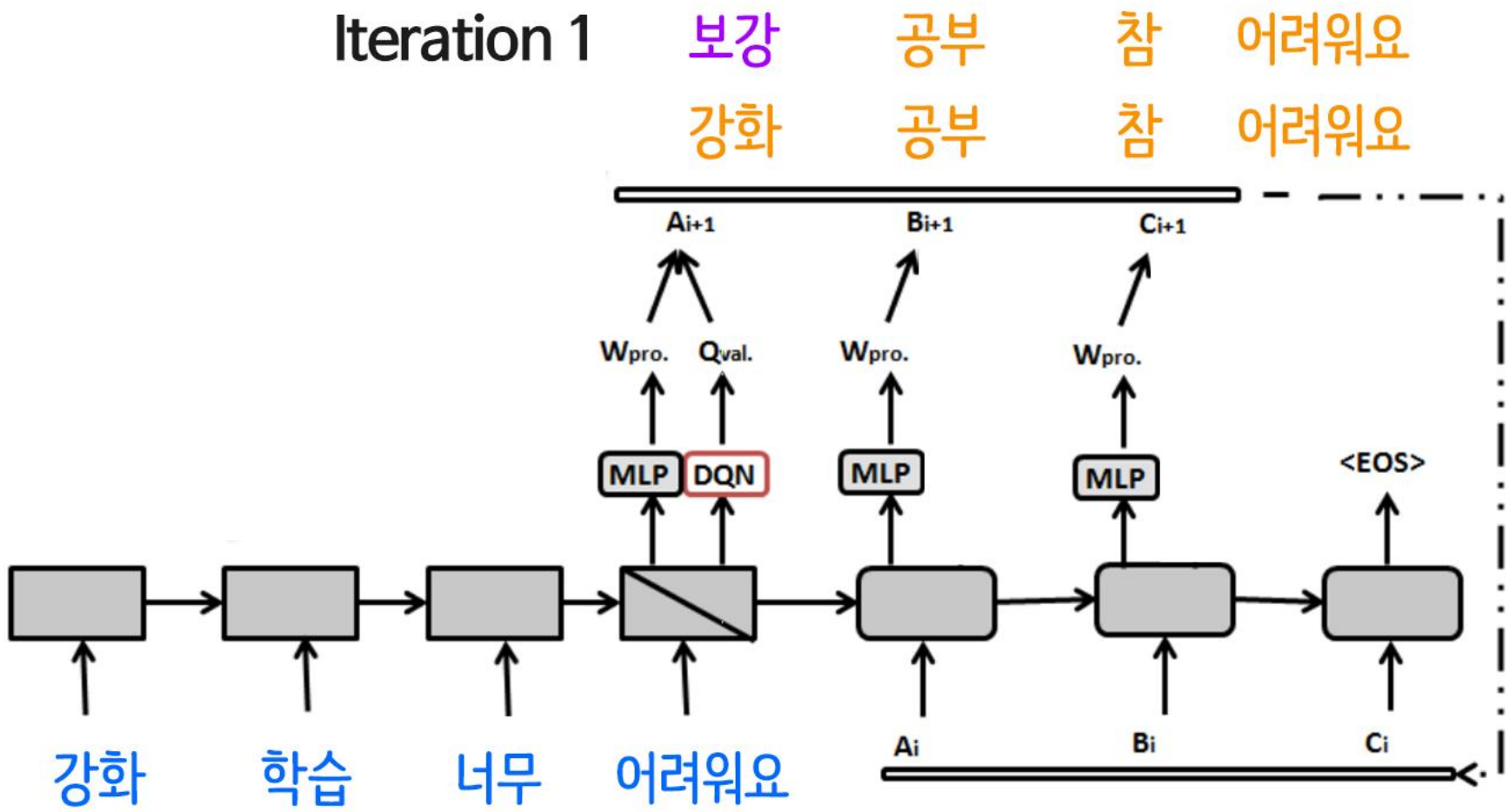
내가 사랑하는 귀여운 강아지 x 37

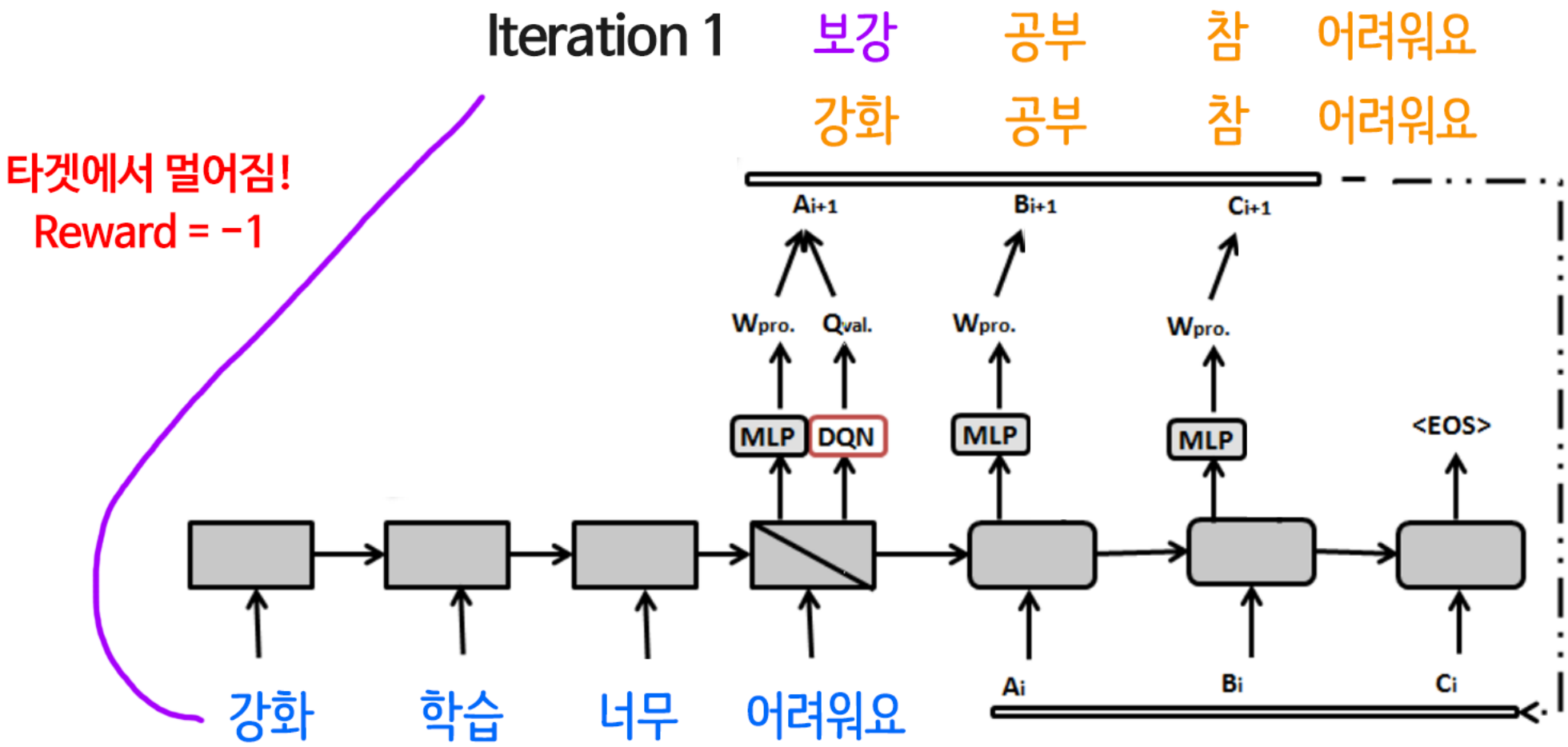
내가 사랑하는 귀여운 떡볶이 x 5

“

Deep Q-Learning 을 하나의 Decoding Strategy로 활용하자!

”

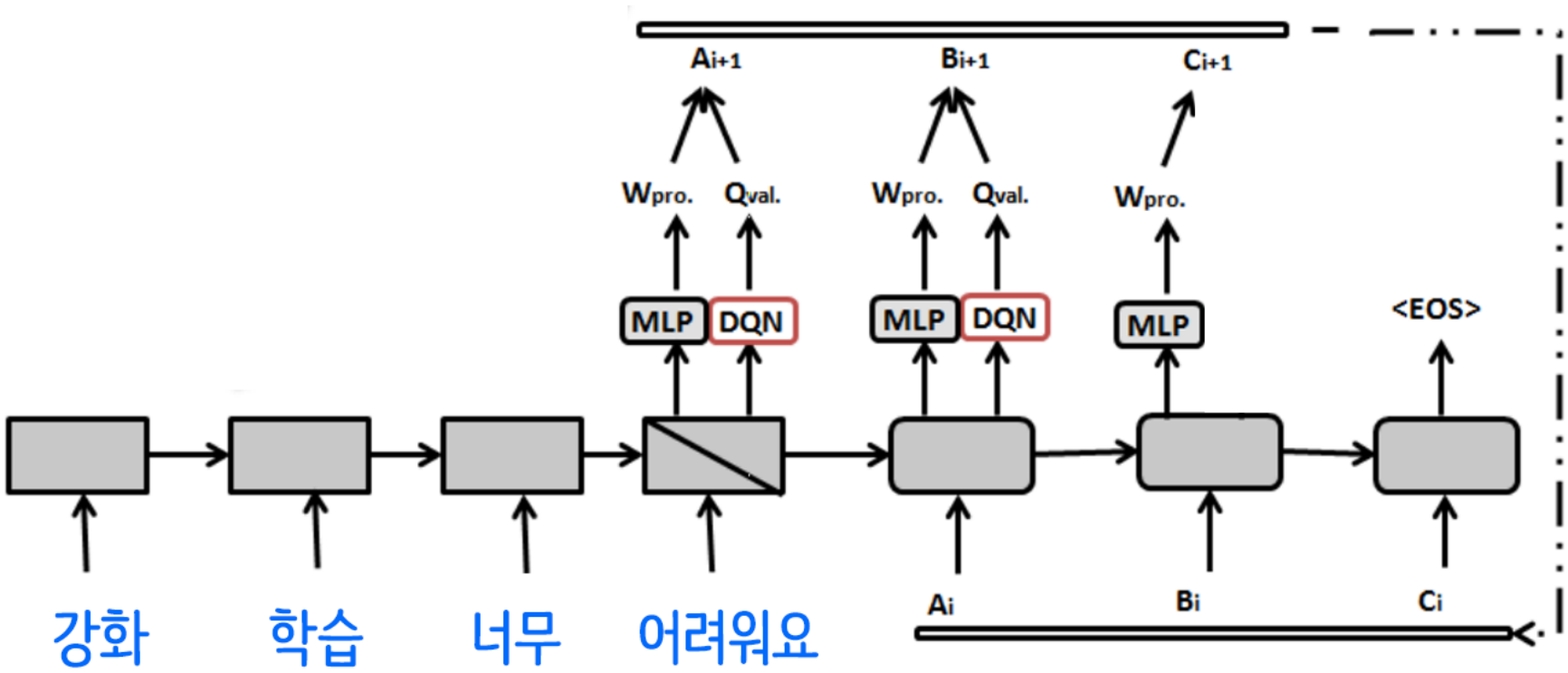

Iteration 1
보강 공부 참 어려워요
강화 공부 참 어려워요
Ai+1
Bi+1
Ci+1
Wpro.
Qval.
Wpro.
Wpro.
MLP
DQN
MLP
MLP
<EOS>
강화
학습
너무
어려워요
Ai
Bi
Ci


Iteration 1
보강 공부 참 어려워요
강화 공부 참 어려워요
타겟에서 멀어짐!
Reward = −1
Ai+1
Bi+1
Ci+1
Wpro.
Qval.
Wpro.
Wpro.
MLP
DQN
MLP
MLP
<EOS>
강화
학습
너무
어려워요
Ai
Bi
Ci

Iteration 2 보강 학습 참 어려워요

Iteration 1 보강 공부 참 어려워요

강화 공부 참 어려워요



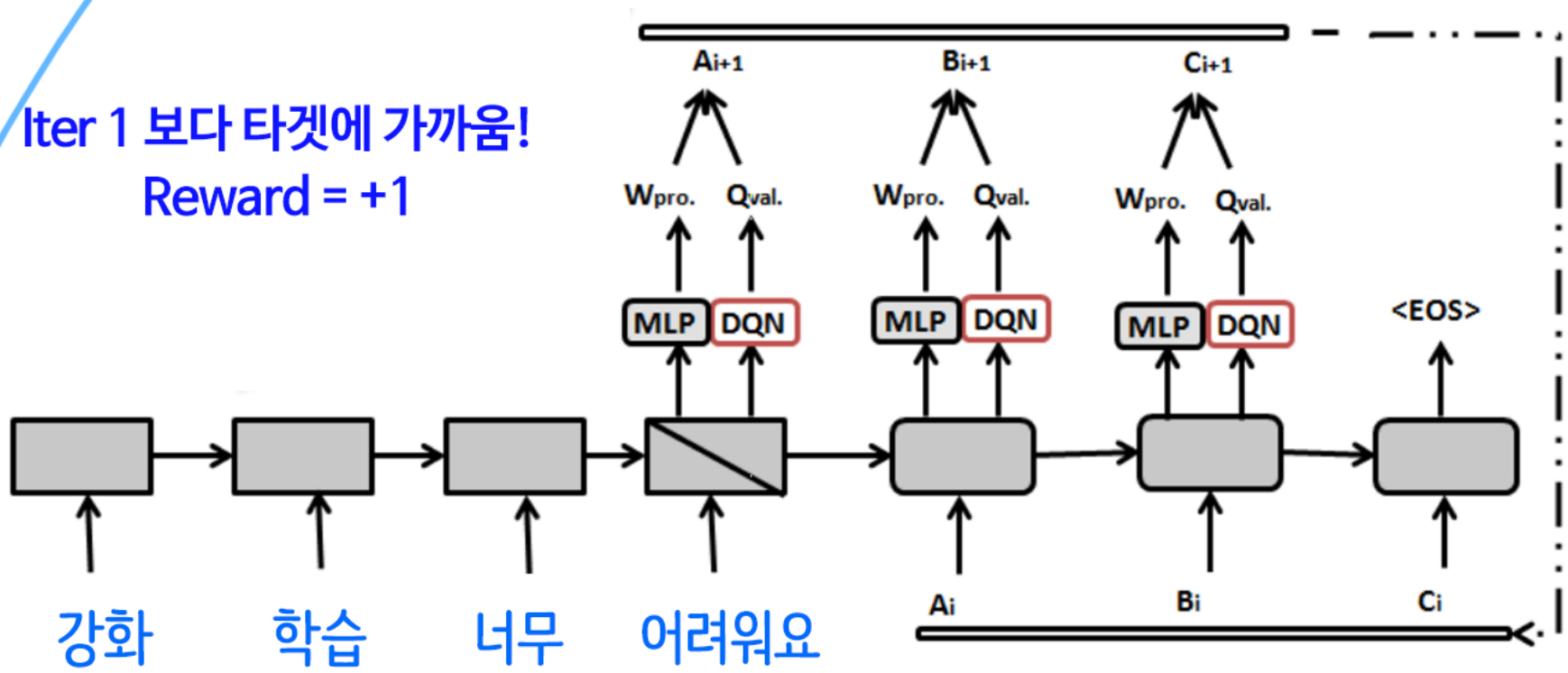
Iteration 2 보강 학습 참 어려워요

Iteration 1 보강 공부 참 어려워요

강화 공부 참 어려워요

Iter 1 보다 타겟에 가까움!

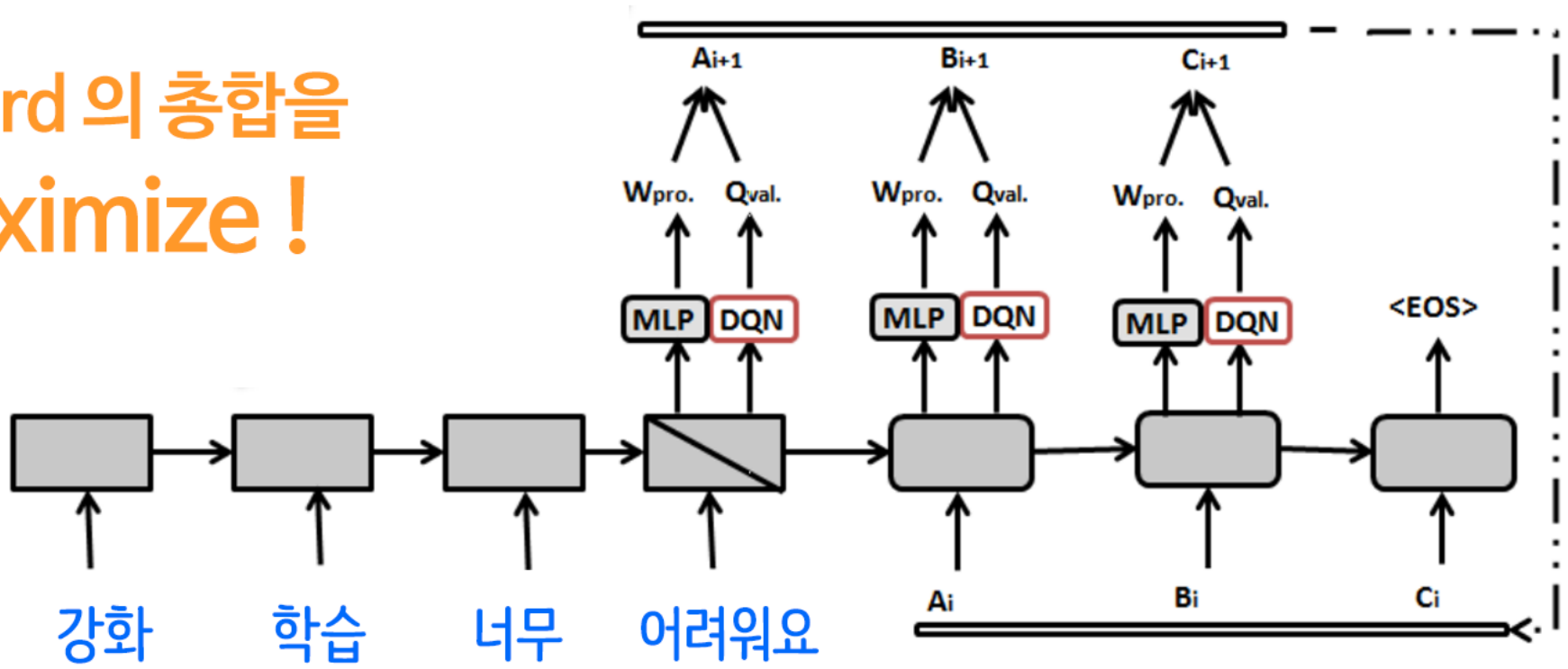
Reward = +1



...Iteration 8	보충	도움	너무	힘들어요
Iteration 2	보강	학습	참	어려워요
Iteration 1	보강	공부	참	어려워요
	강화	공부	참	어려워요

Reward 의 총합을

Maximize !



강화는 어쩌다 보강이 되었나

STATE

Encoder Sen.
강화 학습 너무 어려워요

Decoder Sen.
강화 학습 너무 어려워요

컨텍스트 벡터

WORD RANK

보강
도움
제련
보충
...

30번째까지!

or 유지

ACTION

ϵ-greedy policy

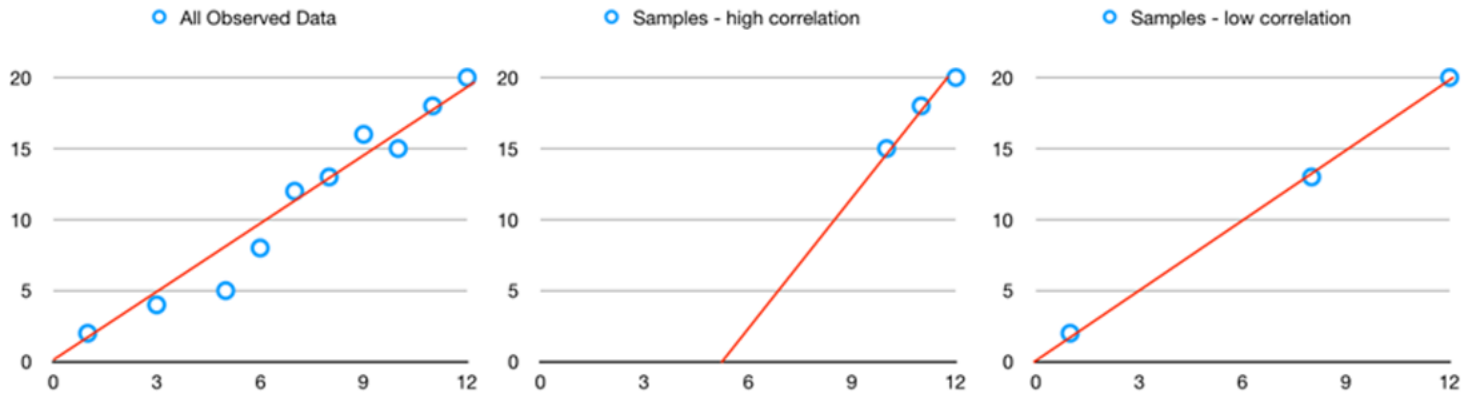
$$a_t = argmax Q(S_i, a)$$

강화는 어쩌다 보강이 되었나

Challenges

- Correlation between samples

강화학습에서의 학습데이터는 시간의 흐름에 따라 순차적으로 수집되고,
이 순차적인 데이터는 근접한 것들끼리 높은 correlation을 띄게된다.



만약에 이 순차적인 데이터를 그대로 입력으로 활용하게 되면 입력이미지
들 간의 높은 correlation에 의해 학습이 불안정해질 것이다.

강화는 어쩌다 보강이 되었나

Replay Memory

1. Agent의 경험(experience) $e_t = (s_t, a_t, r_t, s_{t+1})$를 time-step 단위로 data set $D_t = \{e_1, \ldots, e_t\}$에 저장해 둔다.

2. 저장된 data set으로부터 uniform random sampling을 통해 minibatch를 구성하여 학습을 진행한다. $((s, a, r, s') \sim U(D))$

   - Minibatch가 순차적인 데이터로 구성되지 않으므로 입력 데이터 사이의 correlation을 상당히 줄일 수 있다.

   - 과거의 경험에 대해 반복적인 학습을 가능하게 한다[6].

   - 논문의 실험에서는 replay memory size를 1,000,000으로 설정한다.

강화는 어쩌다 보강이 되었나

STATE

Encoder Sen.
강화 학습 너무 어려워요

Decoder Sen.
강화 학습 너무 어려워요

컨텍스트 벡터

WORD RANK

보강
도움
제련
보충
…

30번째까지!

or 유지

ACTION

Replay Memory

$\{s_1, a_1, r_1, s_2\}$

$\{s_2, a_2, r_2, s_3\}$

…

{ $\{EnSen_i, DeSen_i\}$,
보강, -1,
$\{EnSen_i, DeSen_{i+1}\}$ }

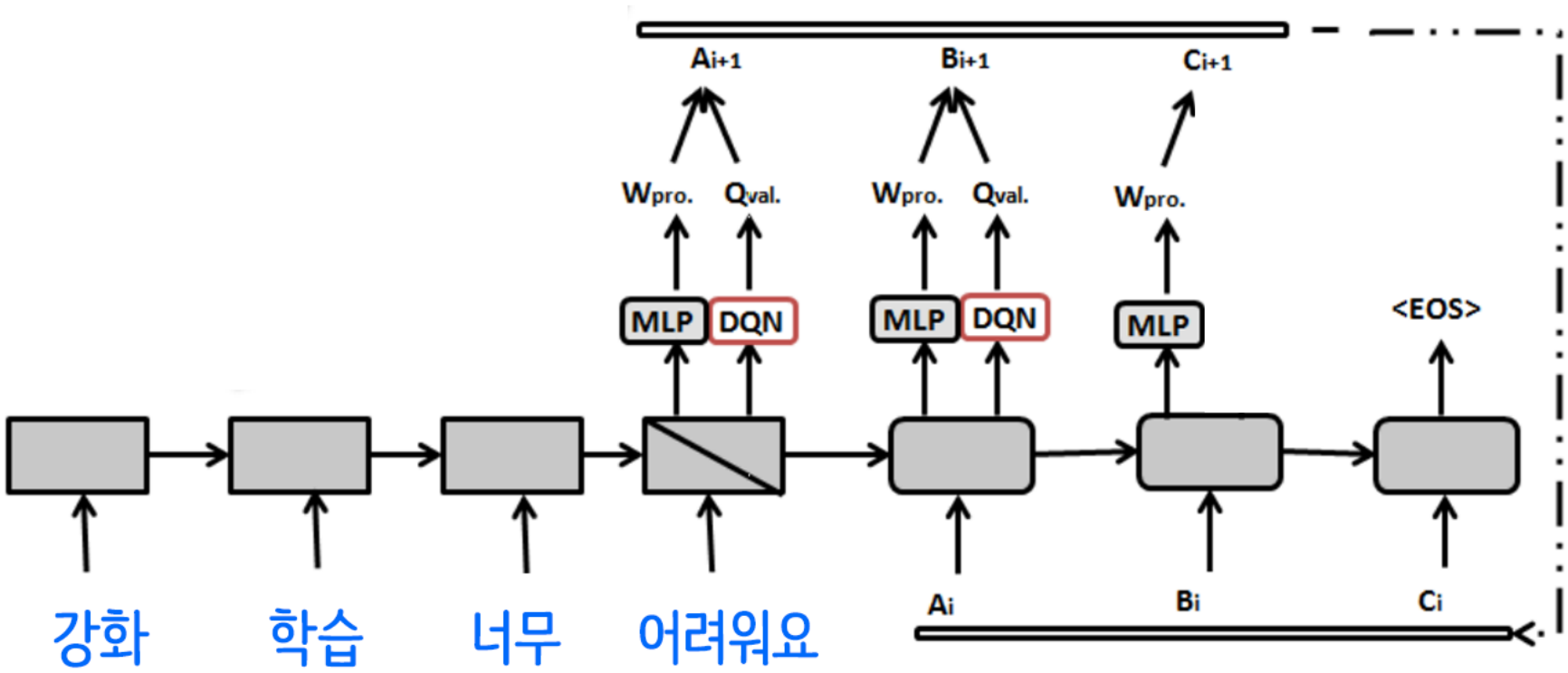
ϵ-greedy policy

$$a_t = argmax Q(S_i, a)$$

Iteration 2 보강 학습 참 어려워요

Iteration 1 보강 공부 참 어려워요

강화 공부 참 어려워요



Challenges

- **Non-stationary targets**

MSE(Mean Squared Error)를 이용하여 optimal action-value function을 근사하기 위한 loss function을 다음과 같이 표현할 수 있다.

$$L_i(\theta_i) = \mathbb{E}_{s,a,r,s'}\left[\left(r + \gamma max_{a'} Q(s', a'; \theta_i) - Q(s, a; \theta_i)\right)^2\right],$$

where θ_i are the parameters of the Q-network at iteration i.

이는 Q-learning target 를 근사하는 $y_i = r + \gamma max_{a'} Q(s', a'; \theta_i)$ 를 구하려는 것과 같다. 문제는 $Q(s, a; \theta_i)$ 가 Q함수에 대해 의존성을 갖고 있으므로 Q함수를 업데이트하게 되면 target y_i 또한 움직이게 된다는 것이다. 이 현상으로 인한 학습의 불안정해진다.

Fixed Q-targets

- $Q(s,a;\theta)$와 같은 네트워크 구조이지만 다른 파라미터를 가진(독립적인) target network $\hat{Q}(s,a;\theta^{-})$를 만들고 이를 Q-learning target y_i에 이용한다.

$$y_i = r + \gamma max_{a'} \hat{Q}(s', a'; \theta_i^{-}) .$$

$$L_i(\theta_i) = \mathbb{E}_{(s,a,r,s')\sim U(D)}\left[\left(r + \gamma max_{a'} \hat{Q}(s', a'; \theta_i^{-}) - Q(s,a;\theta_i)\right)^2\right],$$

in which γ is the discount factor determining the agent's horizon, θ_i are the parameters of the Q-network at iteration i and θ_i^{-} are the network parameters used to compute the target at iteration i.

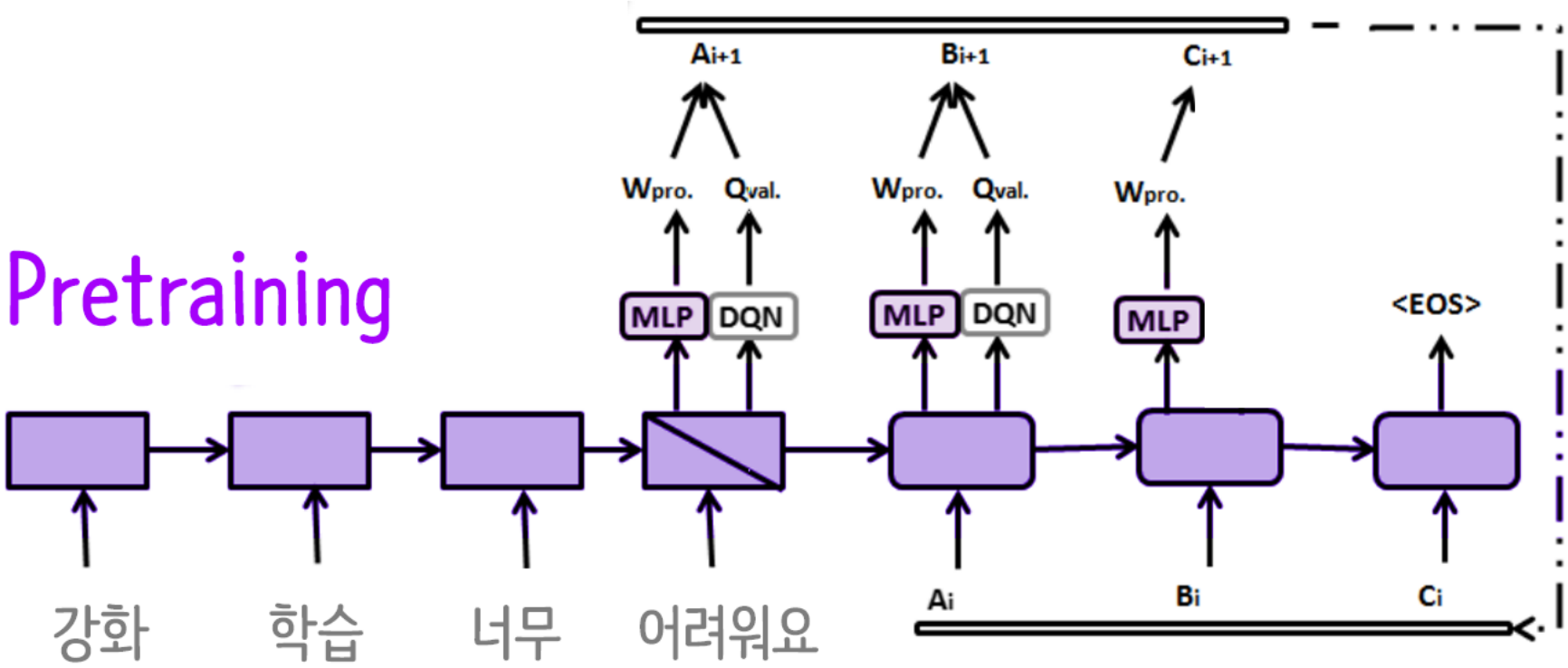
- Target network parameters θ_i^{-}는 매 C step마다 Q-network parameters(θ_i)로 업데이트된다. 즉, C번의 iteration동안에는 Q-learning update시 target이 움직이는 현상을 방지할 수 있다.

- 논문의 실험에서는 C값을 10,000으로 설정한다.

Iteration 2 보강 학습 참 어려워요

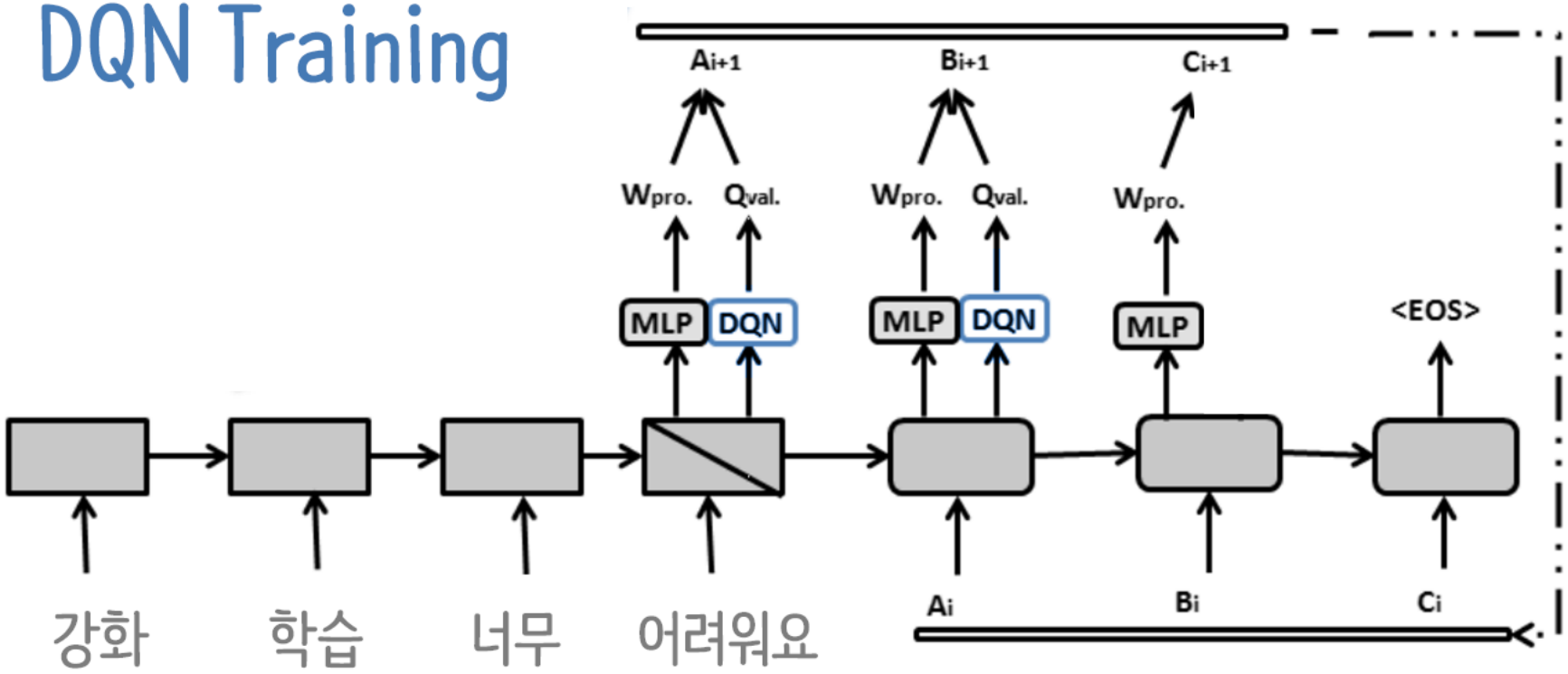
Iteration 1 보강 공부 참 어려워요

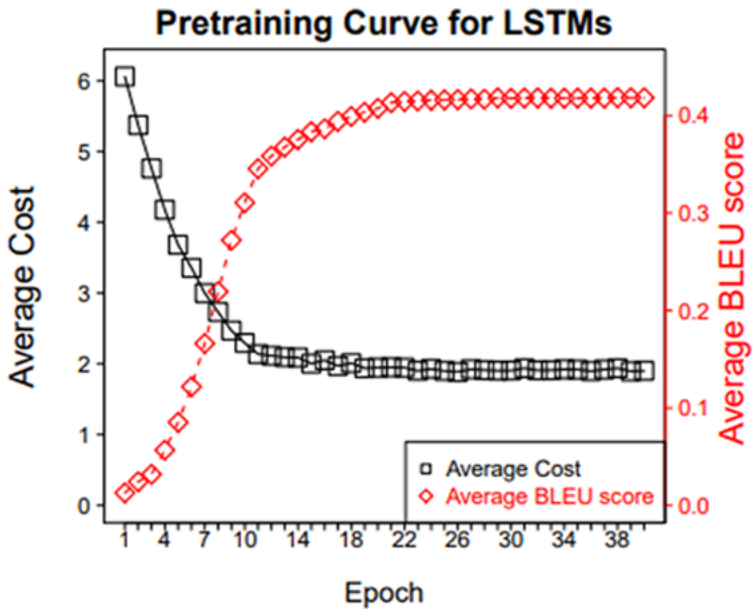
강화 공부 참 어려워요

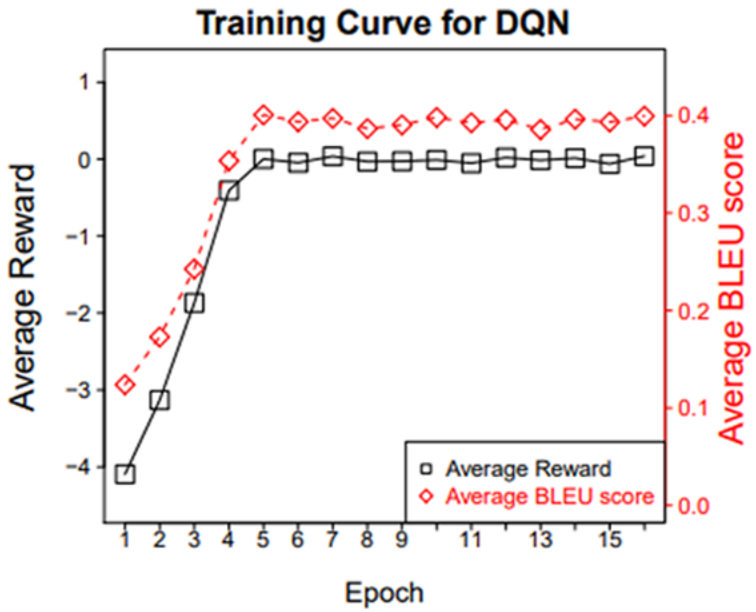


Iteration 2	보강	학습	참	어려워요
Iteration 1	보강	공부	참	어려워요
	강화	공부	참	어려워요

DQN Training







Total 12,000 Sentences

Test: 1000 Seen, 1000 Unseen

100 Embedding Size

100 LSTM Hidden Size

Testing systems	LSTM decoder	DQN
Average SmoothedBLEU on sentences IN the training set	0.425	0.494
Average SmoothedBLEU on sentences NOT in the training set	0.107	0.228

THANK YOU !