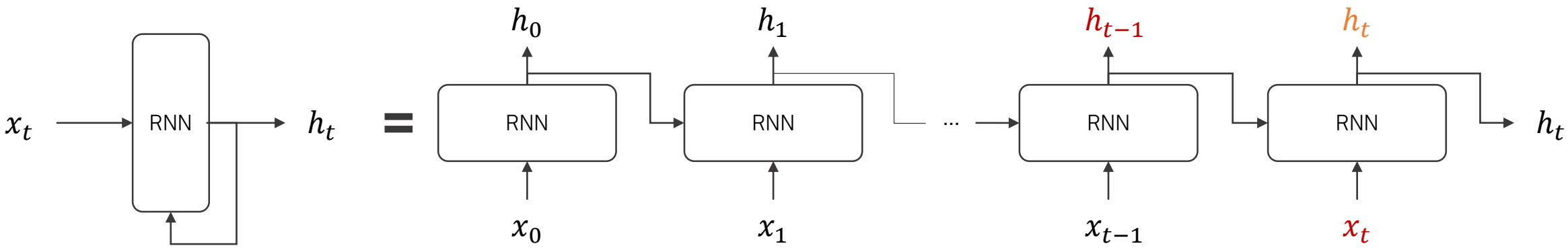
Attention Mechanism

Recurrent Neural Network (RNN)



순환 신경망(RNN)
순환하는 닫힌 구조의 신경망

$$h_t = \tanh(h_{t-1}W_h + x_tW_x + b)$$

$$h_0 = \tanh(h_{init}W_h + x_0 W_x + b)$$

$$h_1 = \tanh(h_0W_h + x_1W_x + b)$$

...

$$h_{t-1} = \tanh(h_{t-2}W_h + x_{t-1}W_x + b)$$

$$h_t = \tanh(h_{t-1}W_h + x_tW_x + b)$$

Recurrent Neural Network (RNN)

그림 6-2 RNN 계층의 계산 그래프(MatMul 노드는 행렬 곱을 나타냄)

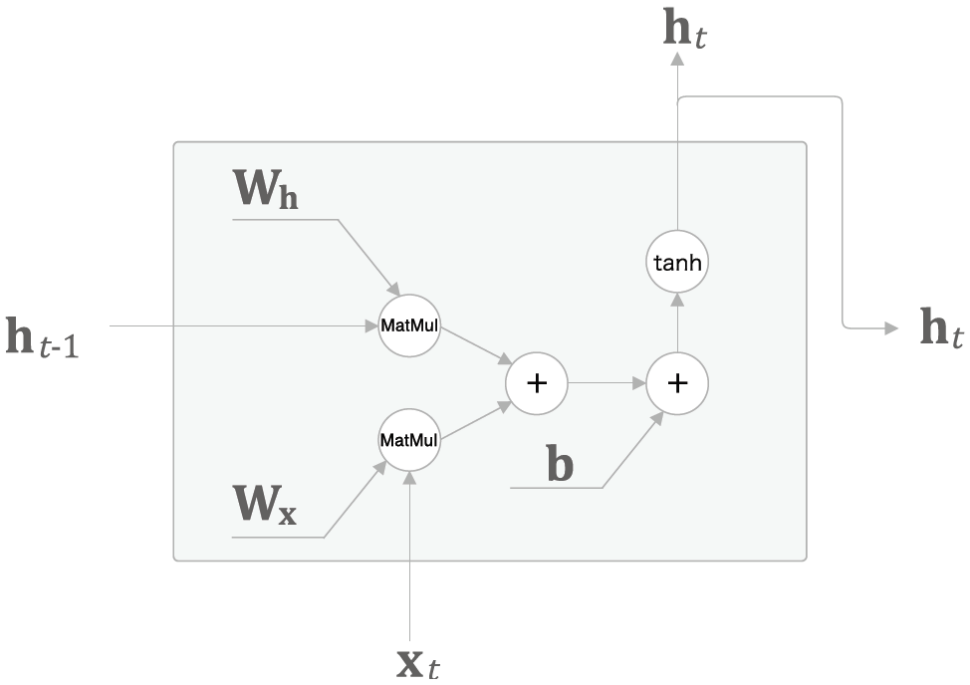
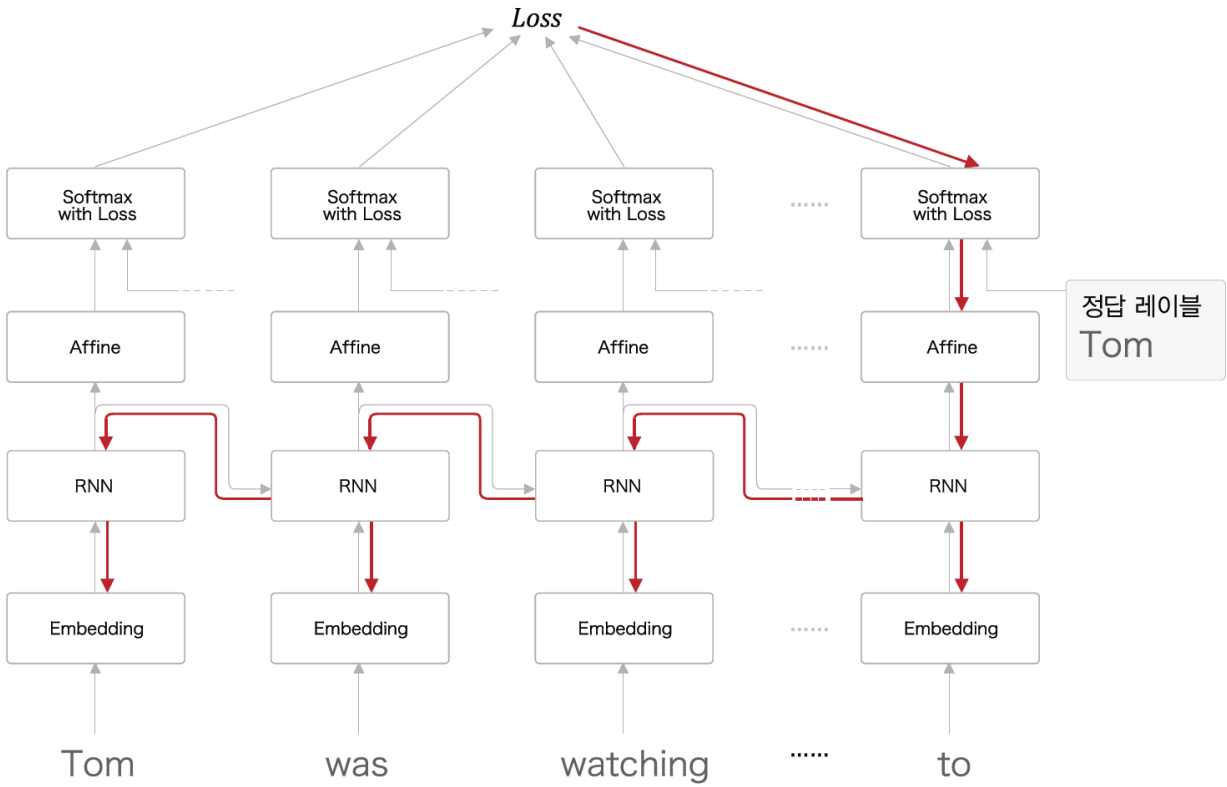


그림 6-4 정답 레이블이 "Tom"임을 학습할 때의 기울기 흐름



Recurrent Neural Network (RNN)

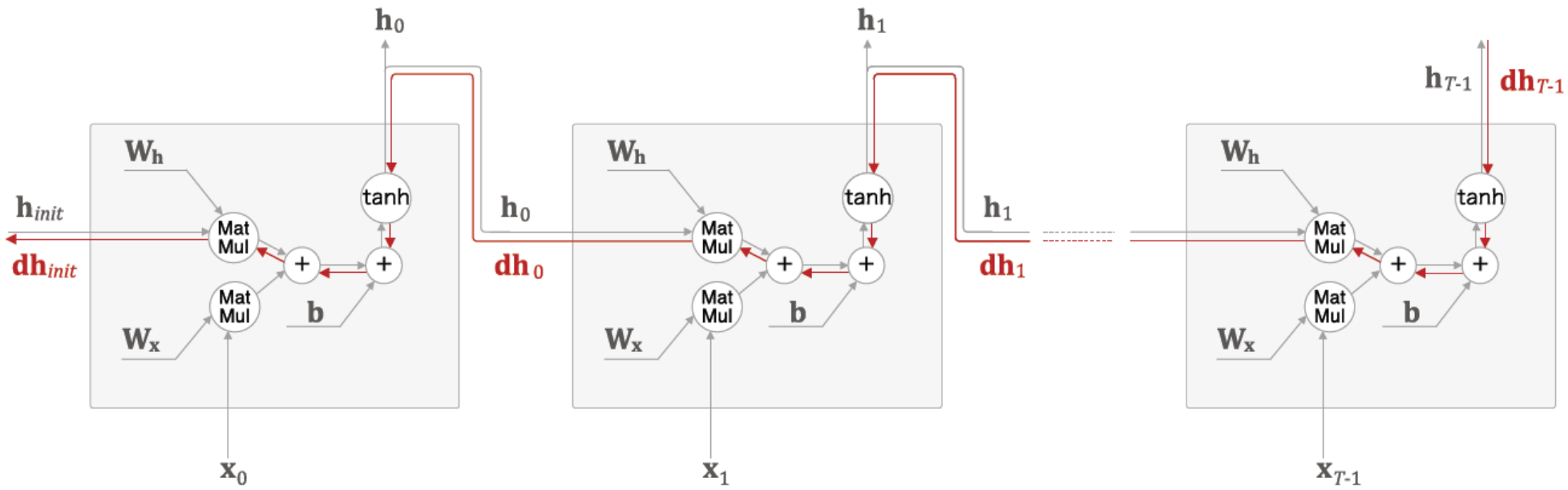
Tom was watching TV in his room. Mary came into the room. Mary said hi to ?

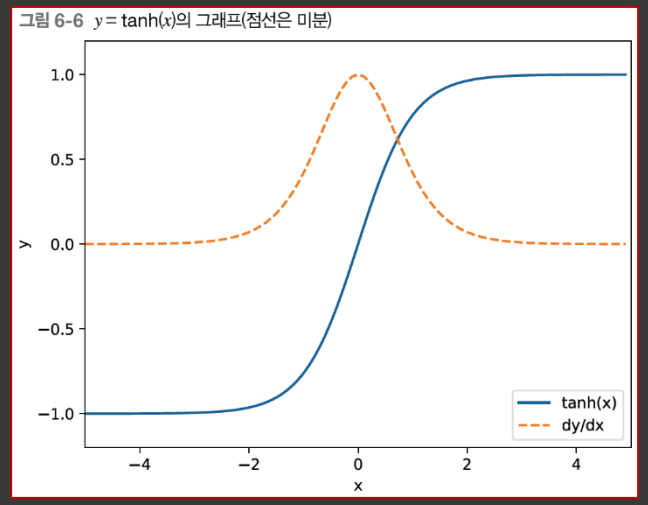
Recurrent Neural Network (RNN)

장기 의존성 문제(기울기 소실 / 기울기 폭발 문제)

- ✓ 학습 과정에서 기울기를 역전파할 때, Step을 여러 번 거치면서 기울기가 사라지거나 폭발하는 현상
- ✓ 기울기를 통한 가중치의 갱신이 어려움

Recurrent Neural Network (RNN)



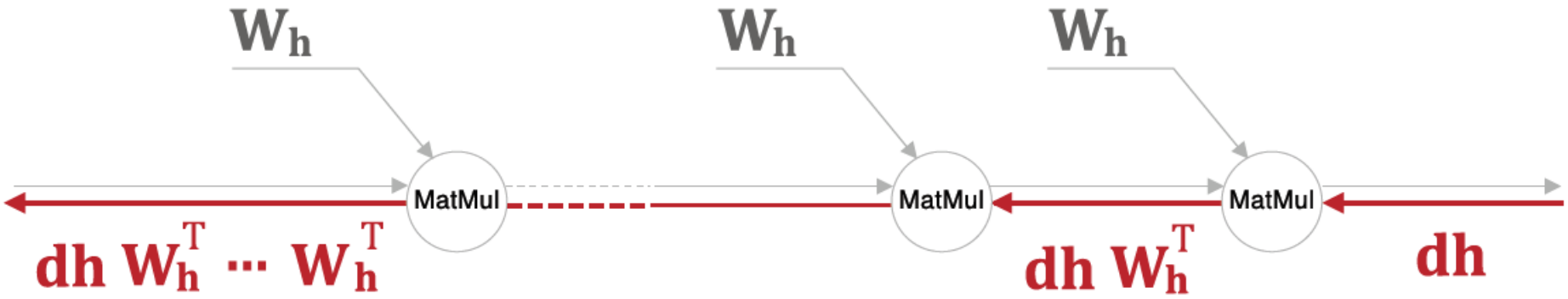


tanh(x)에서 기울기 소실 발생

- tanh(x)의 미분값은 0이상 1이하의 값
- 역전파에서 문장의 단어 길이(T)만큼 tanh를 통과하기 때문에, T 만큼 기울기 값이 계속 감소함
- 기울기가 너무 작아져 소실되면 가중치 매개변수가 더 이상 갱신되지 않음

Recurrent Neural Network (RNN)

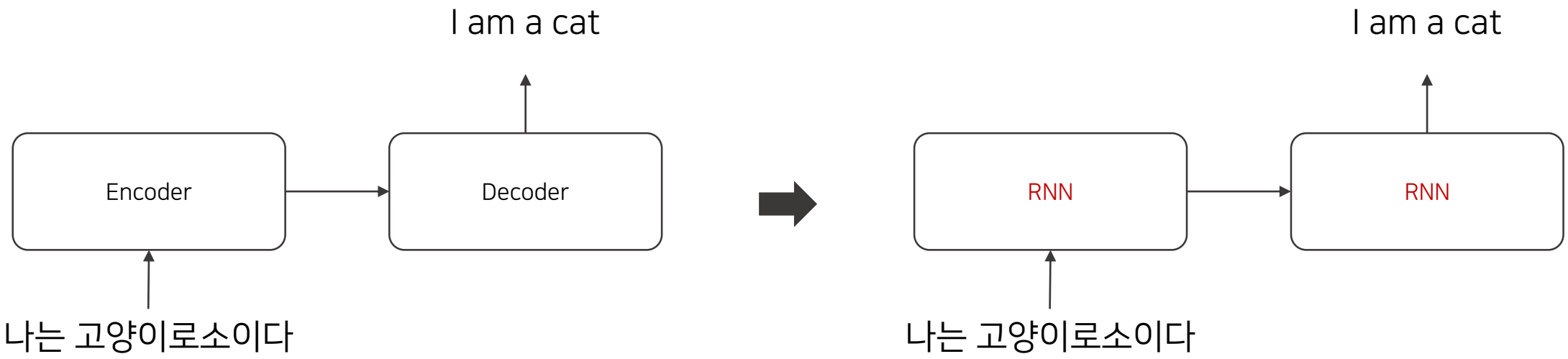
그림 6-7 RNN 계층의 행렬 곱에만 주목했을 때의 역전파의 기울기



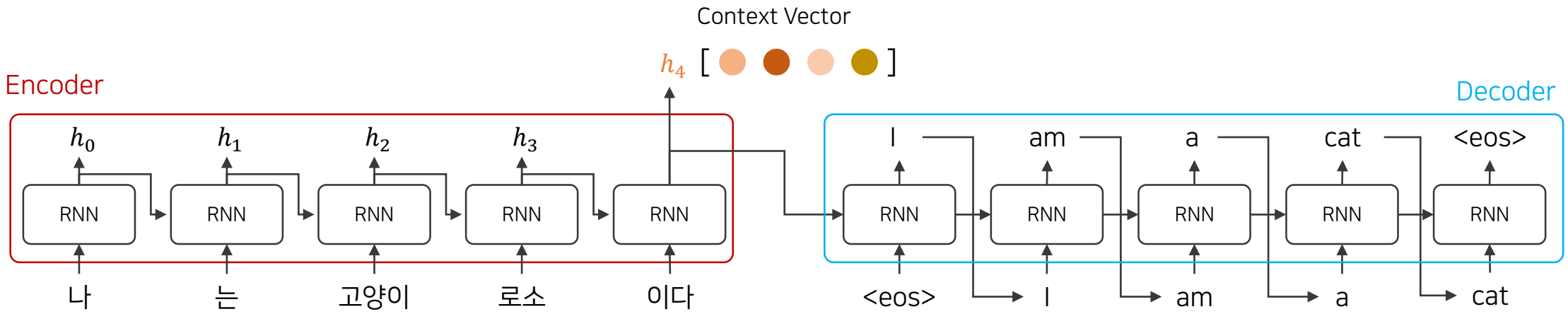
MatMul에서 기울기 소실 혹은 폭발 발생

- MatMul의 역전파는 전치 행렬의 곱이며, 단어 개수(T)만큼 똑같은 행렬을 곱하기 때문에 값이 한쪽 방향으로 커지거나 작아짐
- 기울기 폭발: 값이 너무 커지면 오버플로우를 일으켜 NaN같은 오류 발생
- 기울기 소실: 값이 너무 작아지면 매개변수가 더 이상 갱신되지 않아 장기 의존 관계를 학습할 수 없음

Sequence to Sequence (Seq2Seq)



Sequence to Sequence (Seq2Seq)



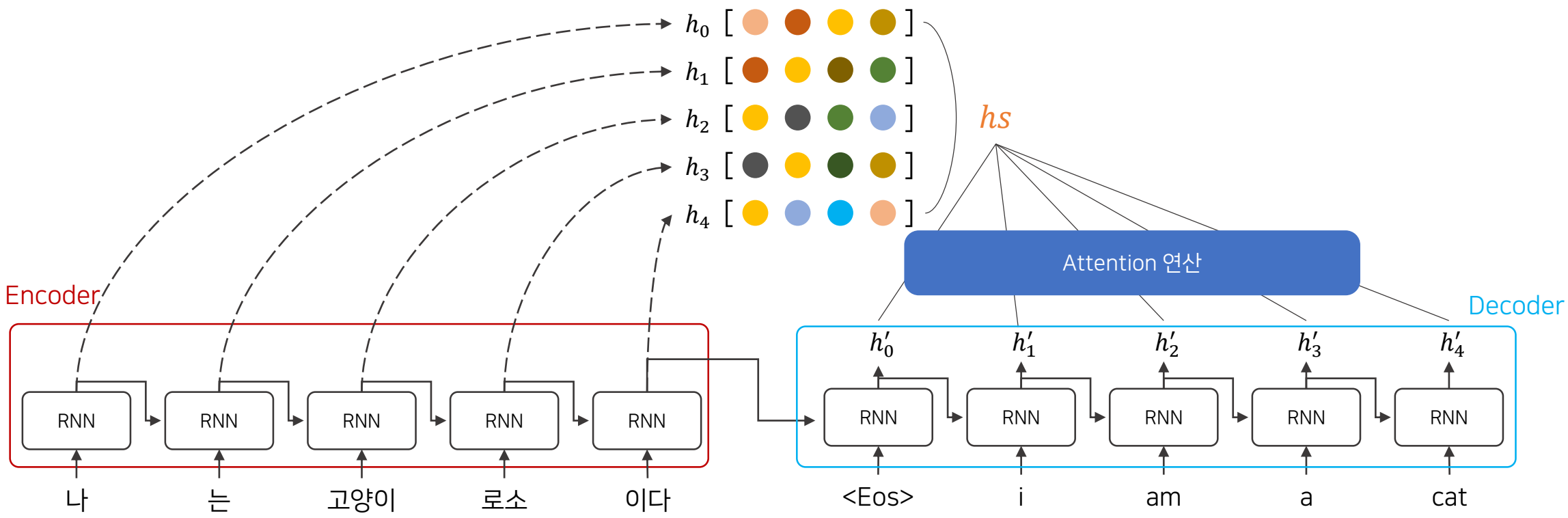
- Seq2Seq의 내부 구조는 RNN이기 때문에 RNN의 장기 의존성 문제를 그대로 가지고 있음
- 여러 단어로 이루어진 긴 문장을 하나의 벡터로 압축하기 때문에 정보의 손실이 일어남
- h_4를 Context Vector로 활용함

- $h_{<eos>} = \tanh(h_4 W_h' + x_{<eos>} W_x' + b)$

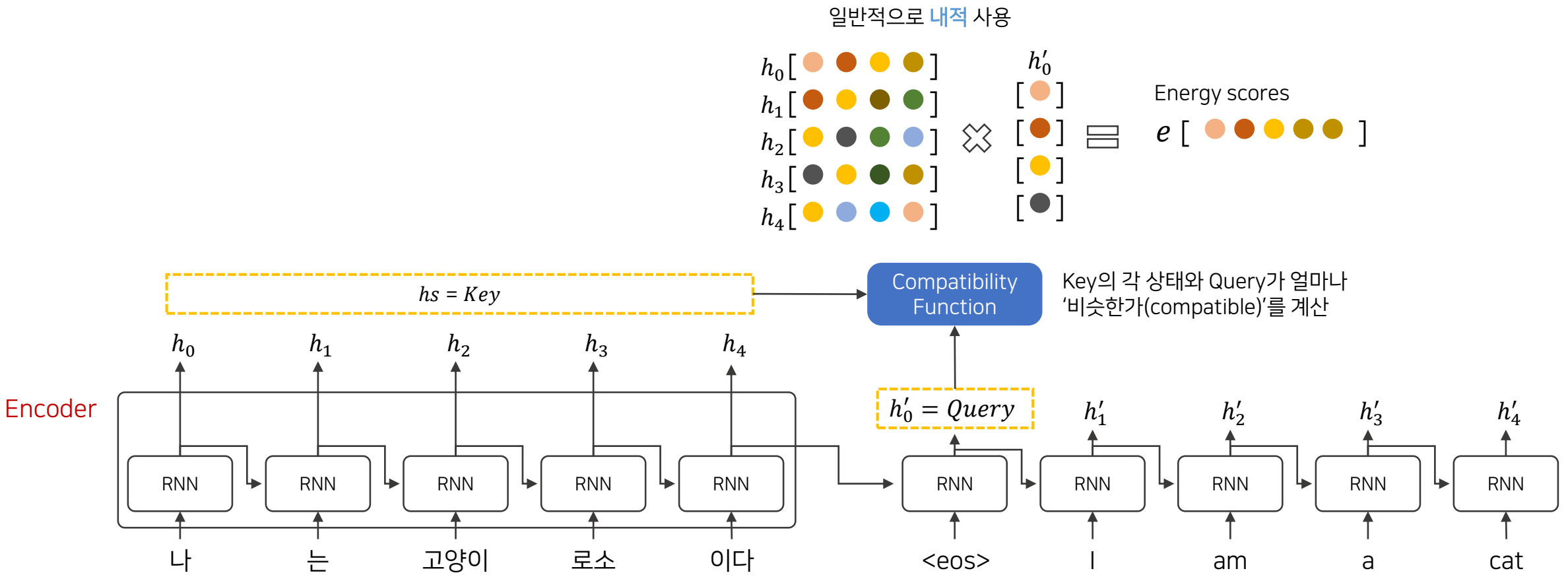
⇨ 하나의 레이어를 더 거쳐 " I " 라는 단어 생성

지금부터 Attention Mechanism 설명합니다.

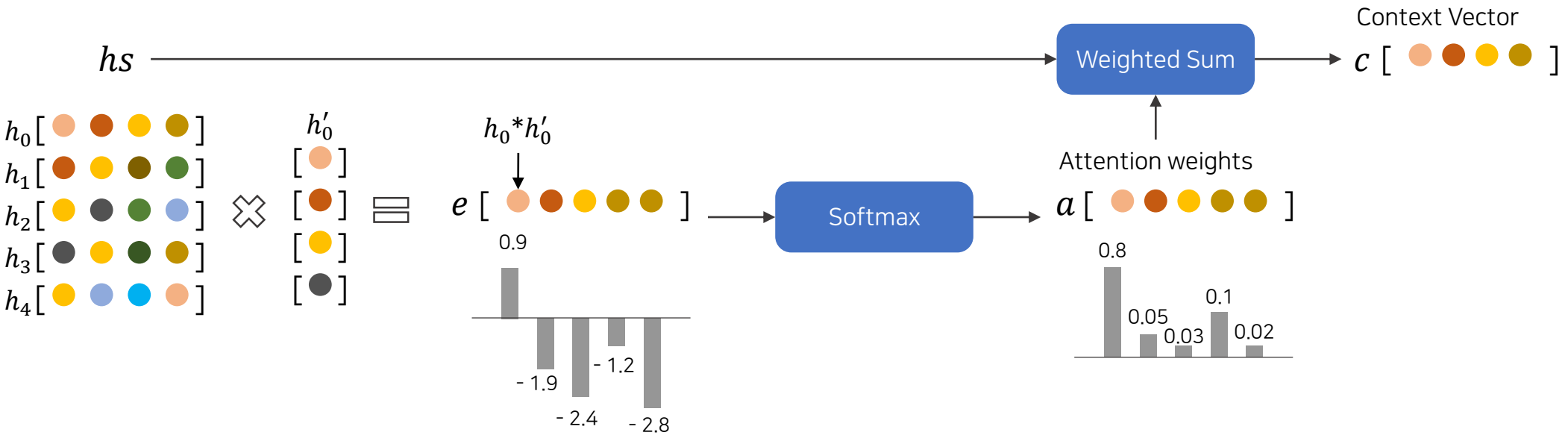
Seq2Seq with Attention



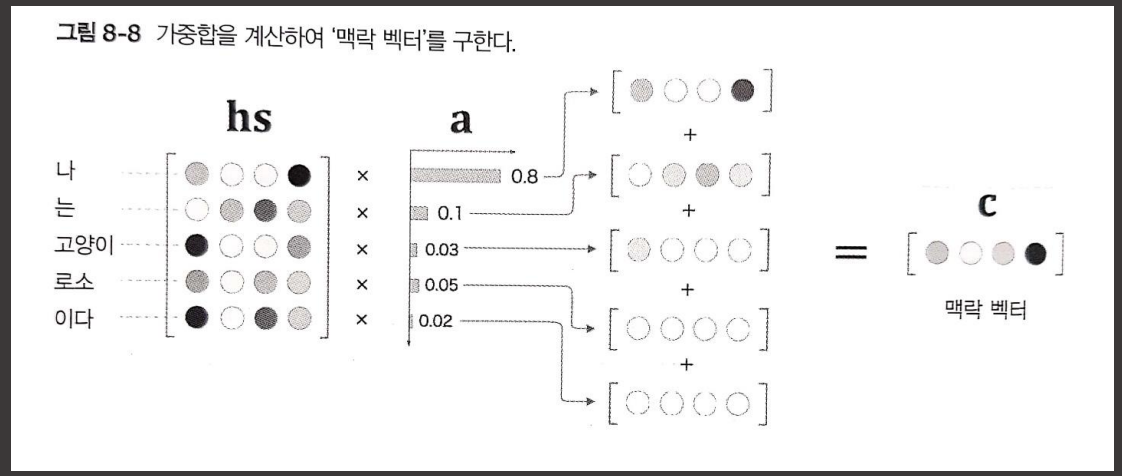
Seq2Seq with Attention



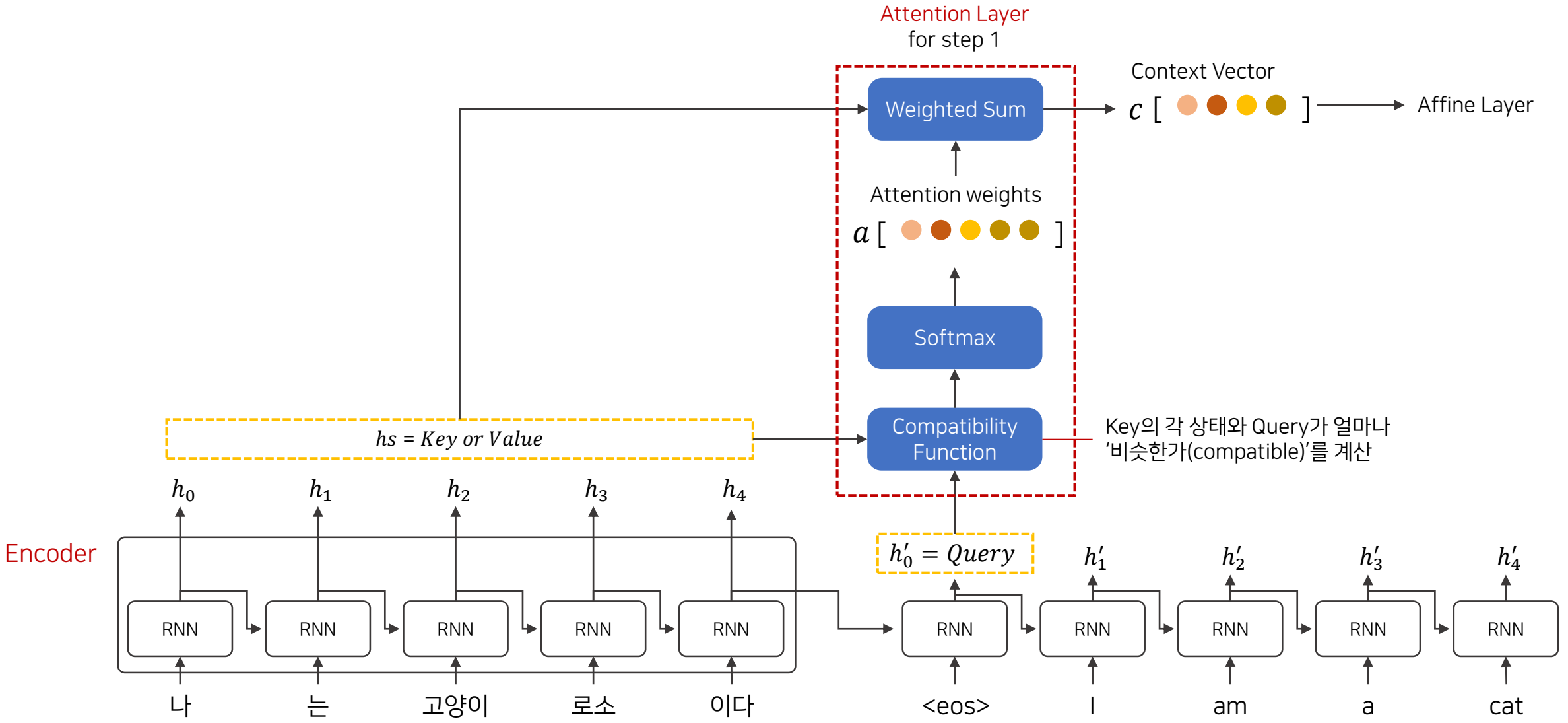
Seq2Seq with Attention



Weighted Sum

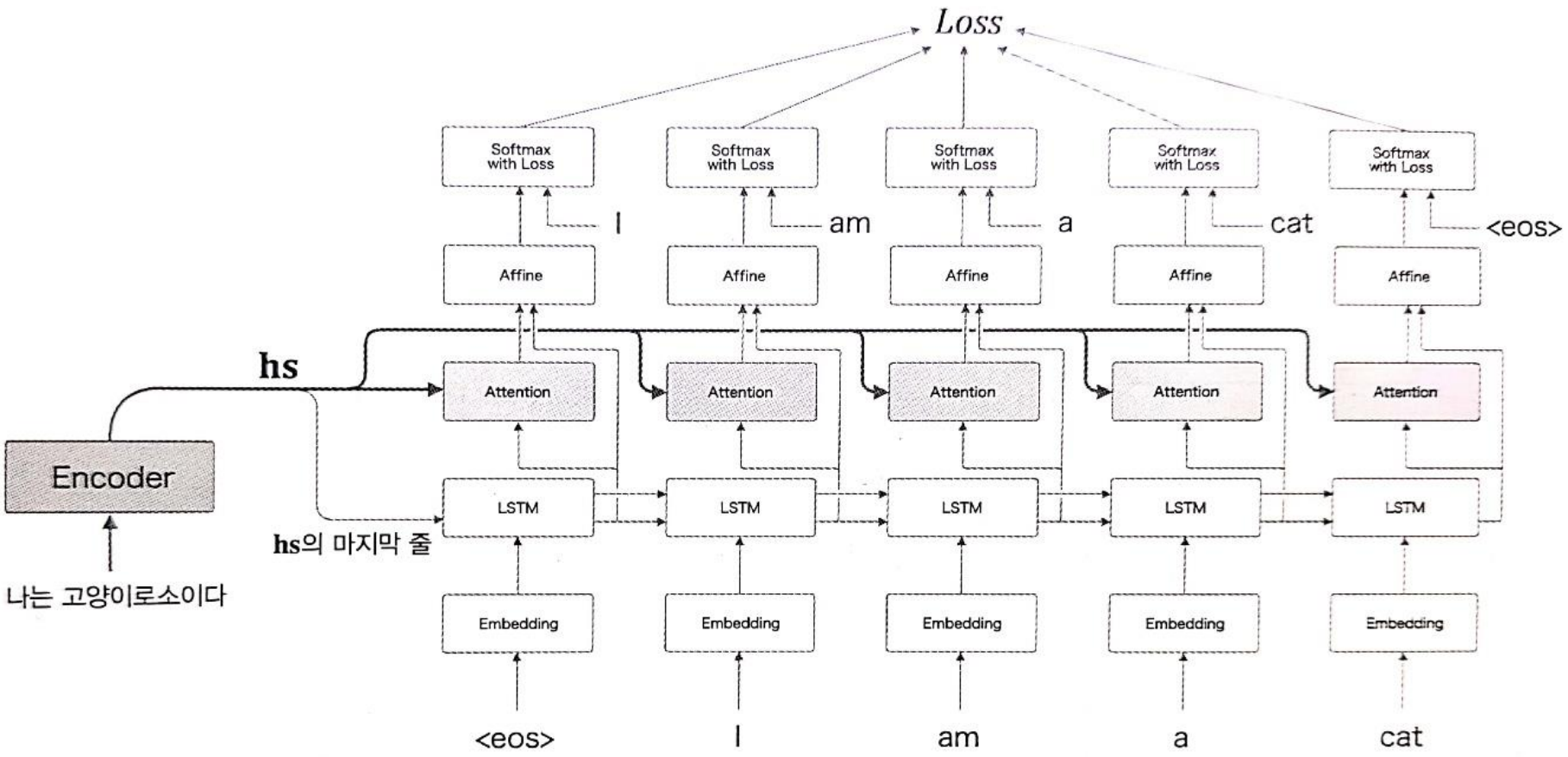


Seq2Seq with Attention



Seq2Seq with Attention

그림 8-18 Attention 계층을 갖춘 Decoder의 계층 구성



Attention, please! A Critical Review of Neural Attention Models in Natural Language Processing

- Compatibility functions

q와 K 비교(Comparing)

유사도 비교 등…

q와 K 결합(Combining)

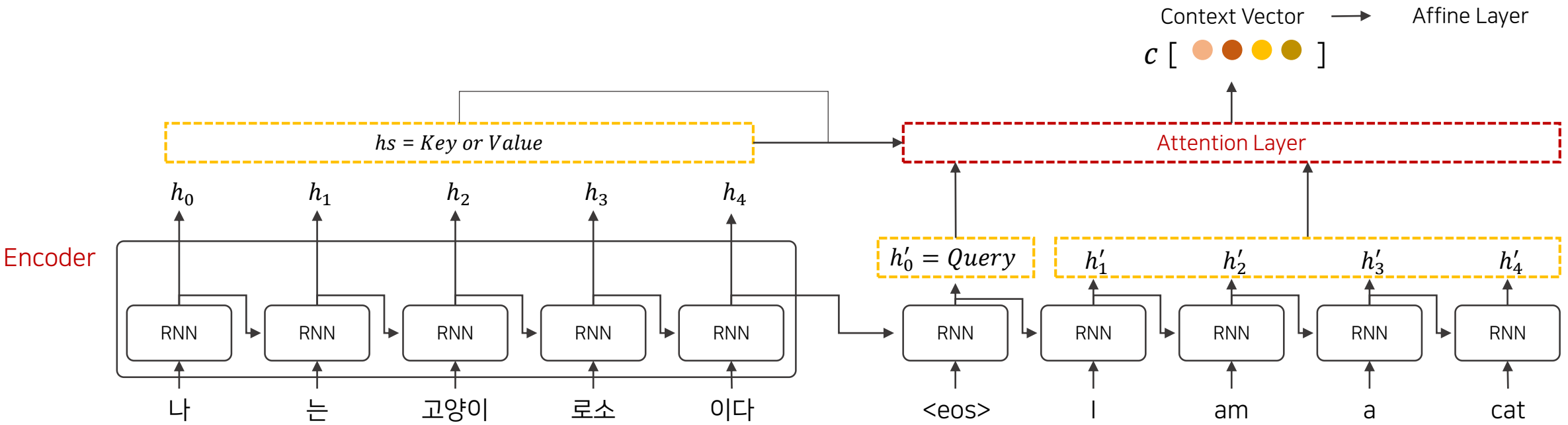
Self-Attention

Name	Equation	Reference
similarity	$f(\boldsymbol{q}, \boldsymbol{K}) = sim(\boldsymbol{q}, \boldsymbol{K})$	Graves et al., 2014
multiplicative or *dot*	$f(\boldsymbol{q}, \boldsymbol{K}) = \boldsymbol{q}^\intercal \boldsymbol{K}$	Luong et al., 2015
scaled multiplicative	$f(\boldsymbol{q}, \boldsymbol{K}) = \frac{\boldsymbol{q}^\intercal \boldsymbol{K}}{\sqrt{d_k}}$	Vaswani et al., 2017
general or *bilinear*	$f(\boldsymbol{q}, \boldsymbol{K}) = \boldsymbol{q}^\intercal \boldsymbol{W} \boldsymbol{K}$	Luong et al., 2015
biased general	$f(\boldsymbol{q}, \boldsymbol{K}) = \boldsymbol{K}^\intercal (\boldsymbol{W}\boldsymbol{q} + \boldsymbol{b})$	Sordoni et al., 2016
activated general	$f(\boldsymbol{q}, \boldsymbol{K}) = act(\boldsymbol{q}^\intercal \boldsymbol{W} \boldsymbol{K} + \boldsymbol{b})$	Ma et al., 2017
concat	$f(\boldsymbol{q}, \boldsymbol{K}) = \boldsymbol{w_{imp}}^\intercal act\big(\boldsymbol{W}[\boldsymbol{K}; \boldsymbol{q}] + \boldsymbol{b}\big)$	Luong et al., 2015
additive	$f(\boldsymbol{q}, \boldsymbol{K}) = \boldsymbol{w_{imp}}^\intercal act(\boldsymbol{W_1}\boldsymbol{K} + \boldsymbol{W_2}\boldsymbol{q} + \boldsymbol{b})$	Bahdanau et al., 2015
deep	$f(\boldsymbol{q}, \boldsymbol{K}) = \boldsymbol{w_{imp}}^\intercal \boldsymbol{E}^{(\boldsymbol{L-1})} + \boldsymbol{b}^{\boldsymbol{L}}$ $\boldsymbol{E}^{(\boldsymbol{l})} = act(\boldsymbol{W}_{\boldsymbol{l}} \boldsymbol{E}^{(\boldsymbol{l-1})} + \boldsymbol{b}^{\boldsymbol{l}})$ $\boldsymbol{E}^{(\boldsymbol{1})} = act(\boldsymbol{W_1}\boldsymbol{K} + \boldsymbol{W_0}\boldsymbol{q} + \boldsymbol{b}^{\boldsymbol{1}})$	Pavlopoulos et al., 2017
location-based	$f(\boldsymbol{q}, \boldsymbol{K}) = f(\boldsymbol{q})$	Luong et al., 2015

Table 3: Summary of compatibility functions found in literature. $\boldsymbol{W}$, $\boldsymbol{W_0}$, $\boldsymbol{W_1}$, …, and $\boldsymbol{b}$ are learnable parameters.

Self- Attention Mechanism

Seq2Seq with Attention



- RNN 구조의 근본적인 한계는 그대로 가지고 있음
  - ✓ 장기 의존성 문제
  - ✓ 병렬 연산이 불가능함 → 학습 속도에서 병목

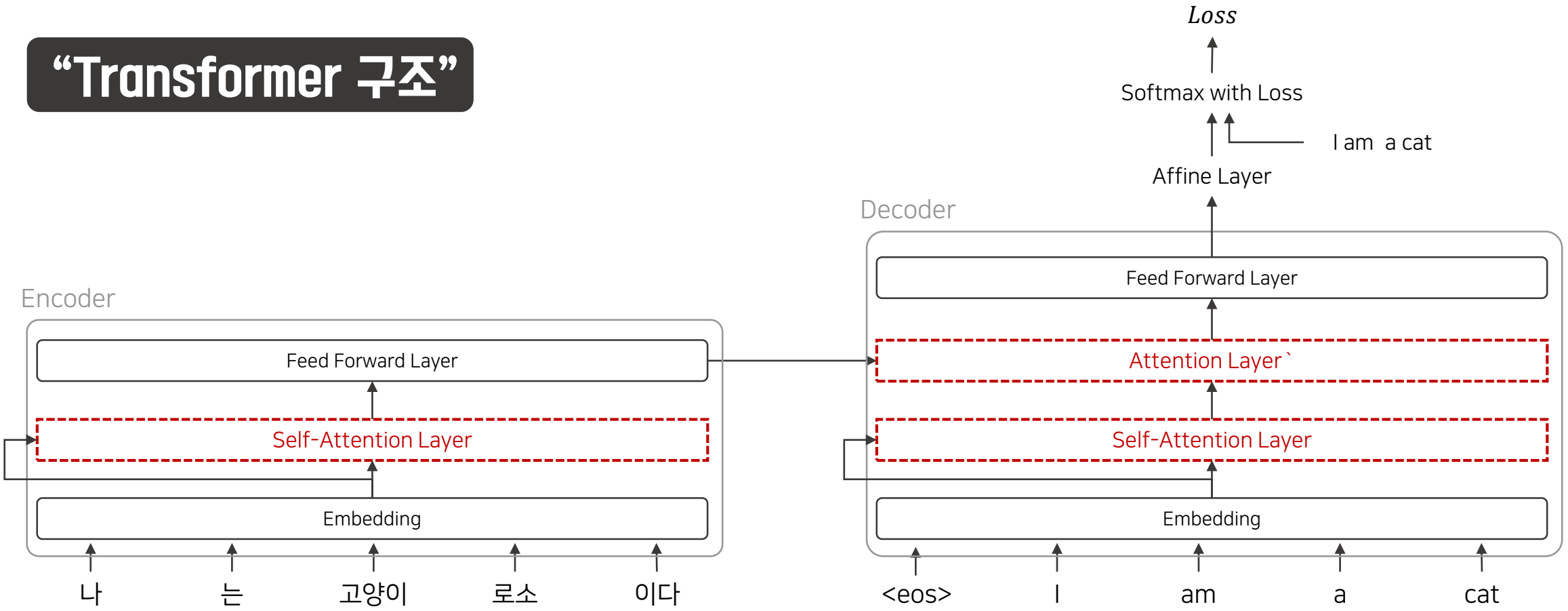
Attention is all you need !!!

Attention is all you need

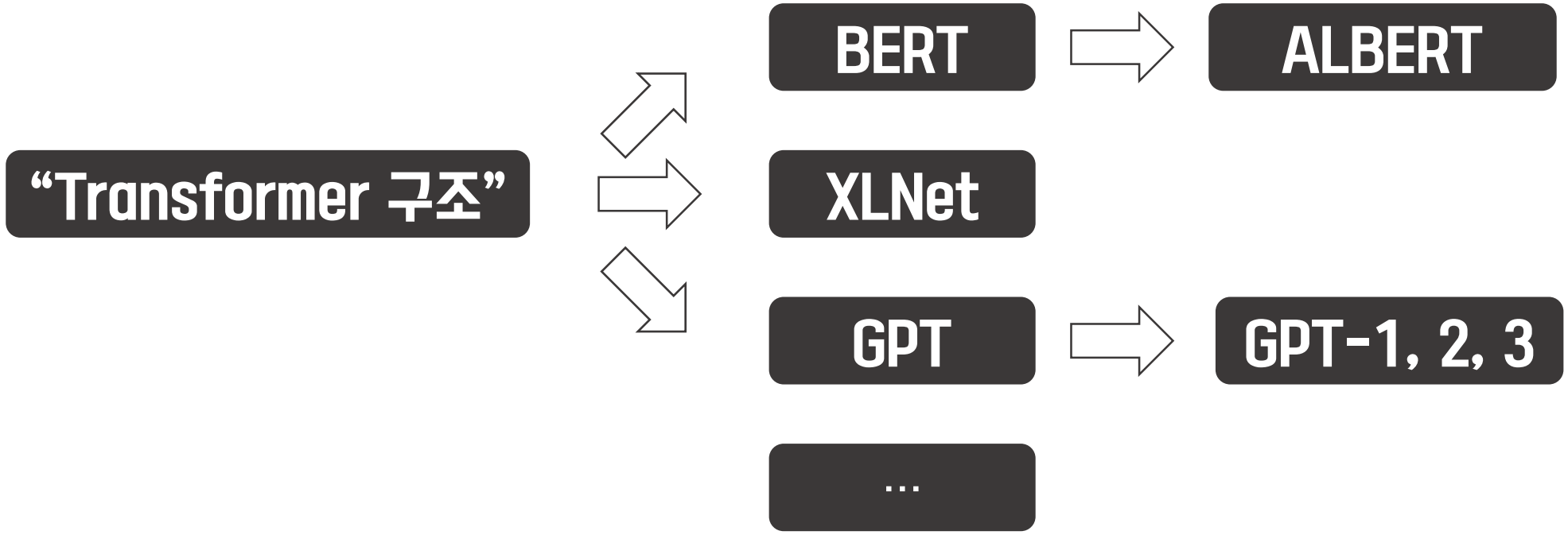
Aurthors: Ashish Vaswani(Google Brain) 외 7명,

Conference: Advances in Neural Information Processing Systems 30 (NIPS 2017)

“Transformer 구조”



Attention is all you need



감사합니다.