삼성전기 AI전문가 양성과정 - 프로젝트 실습 (비영상)

자연어처리를 위한 Attention

현청천

2022.02.28

What is Attention

주어진 목적에 맞는 Source에 집중

감정분석

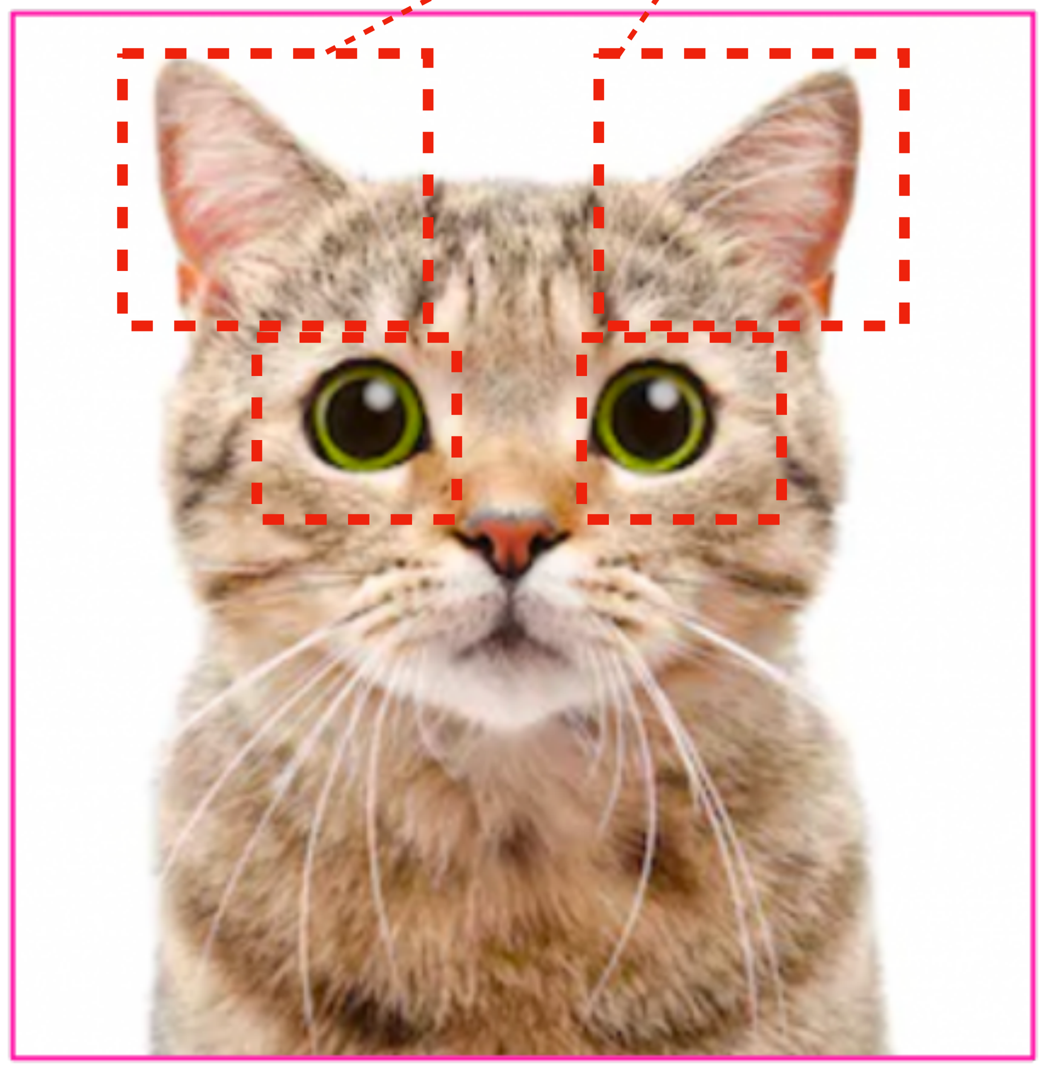
할머니 만나는 부분에서 울었습니다. 감동적인 영화입니다.

긍정

What is Attention

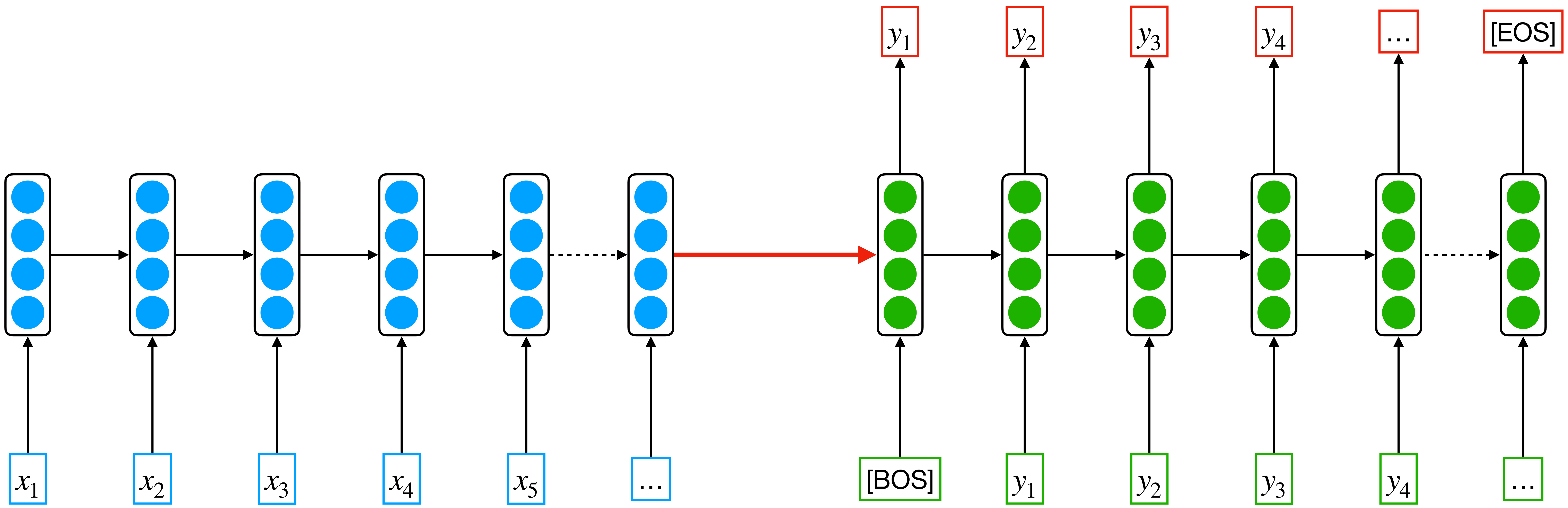
주어진 목적에 맞는 Source에 집중

이미지분류



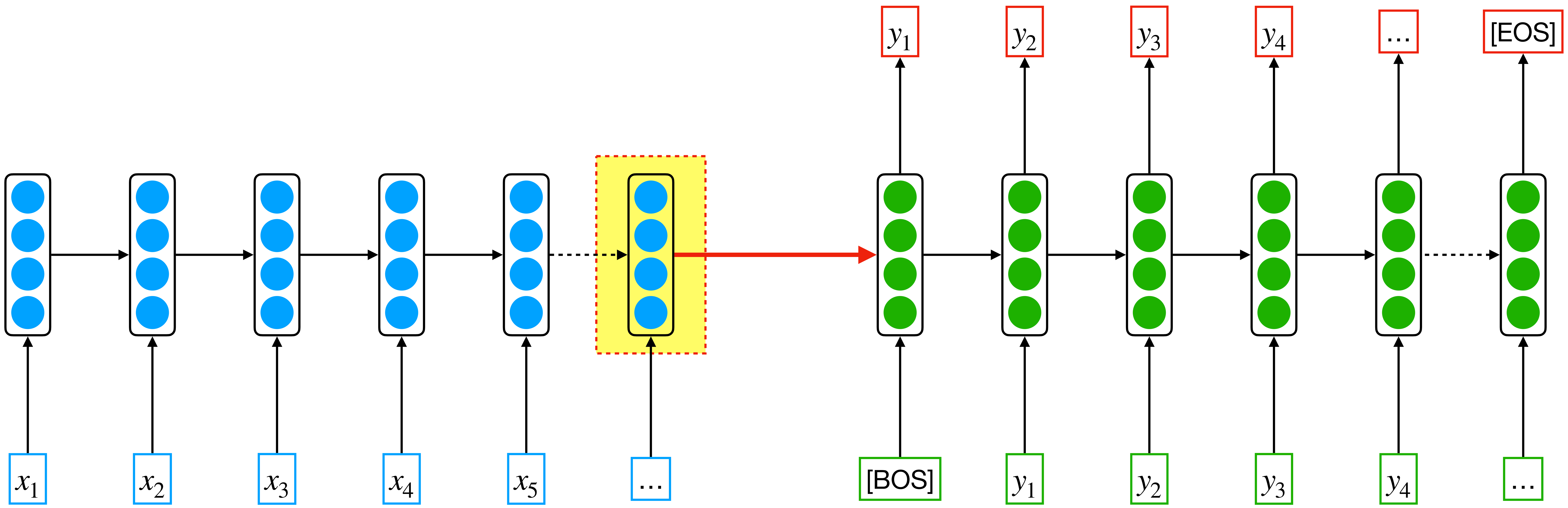
고양이

What is Attention



What is Attention

Encoder Input 정보를 하나의 벡터로 저장

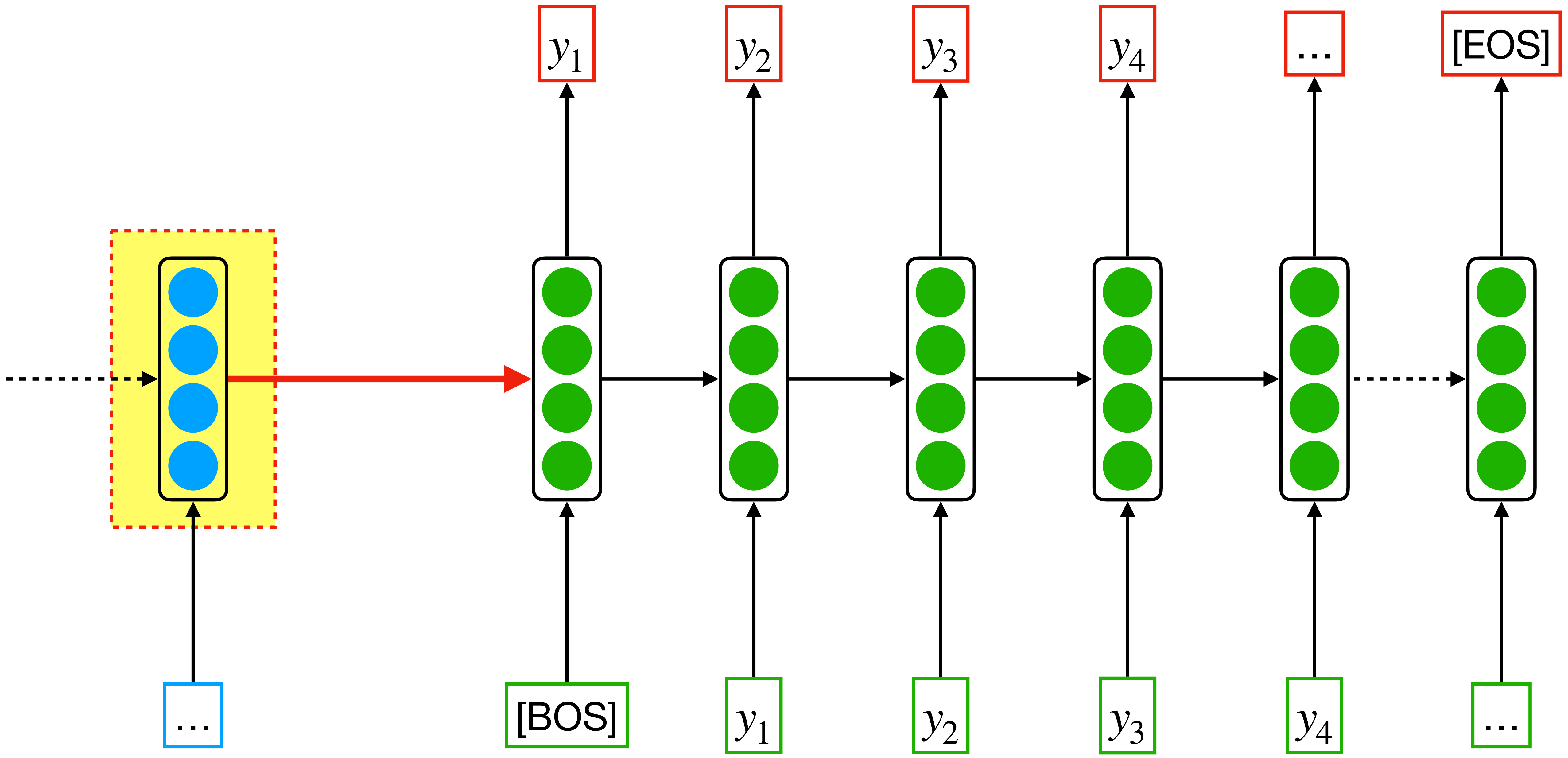


What is Attention

Encoder Input 정보를 하나의 벡터로 저장

긴 문장을 하나의 벡터로 변환하기 어려움 (Information bottleneck)

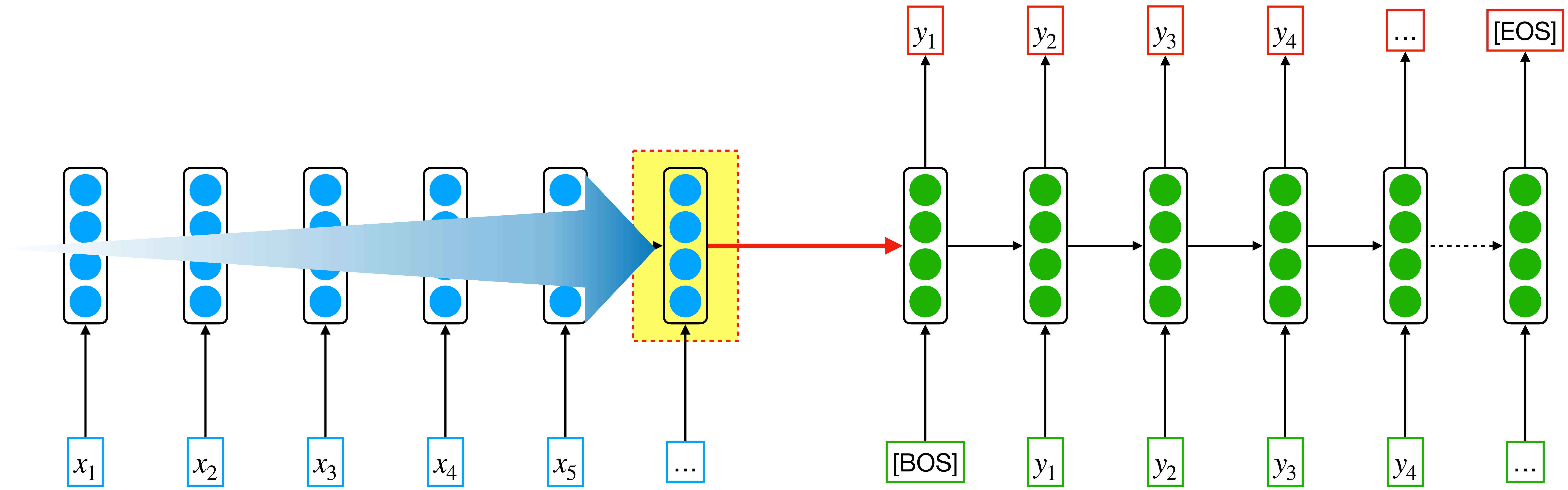
The dominant sequence transduction models are based on complex
…
…
…
requiring significantly less time to train



What is Attention

Encoder Input 정보를 하나의 벡터로 저장

과거의 정보가 점점 사라짐 (Vanishing gradient)

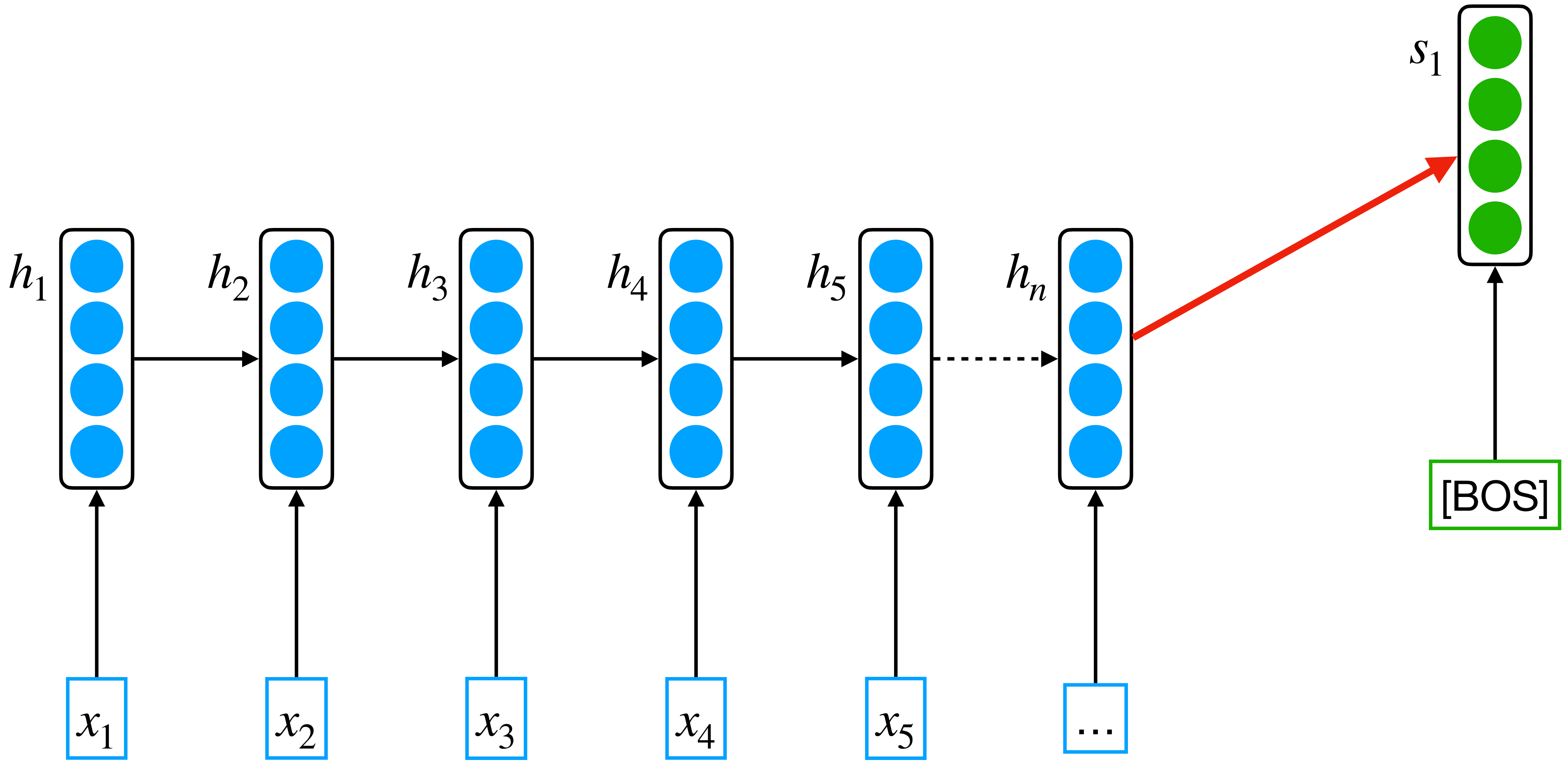


What is Attention

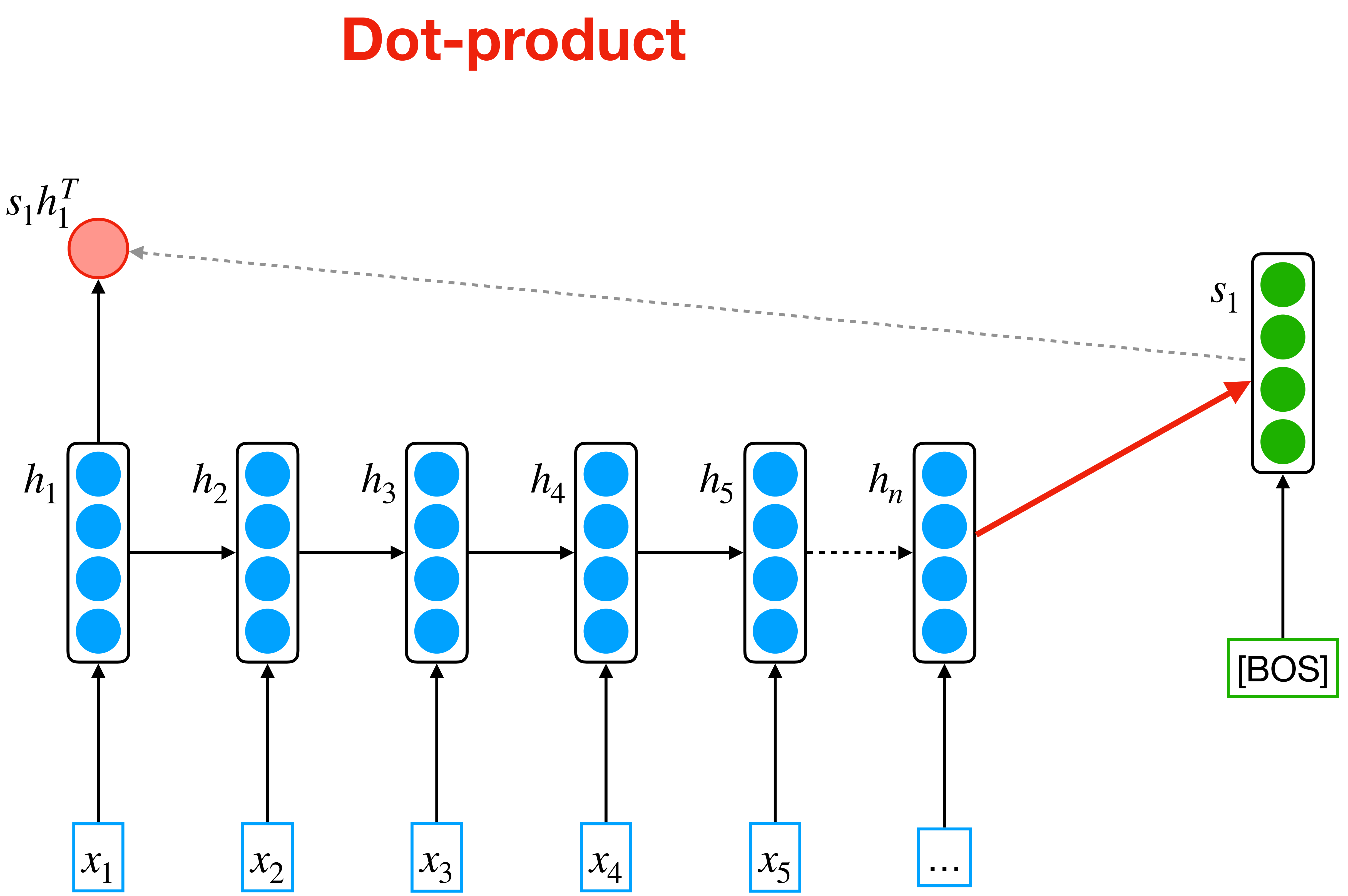
- 두가지 주요 문제 해결
  - 긴 문장을 하나의 벡터로 변환하면서 발생하는 Information bottleneck
  - 과거의 정보가 점점 사라지는 vanishing gradient

Source의 특정 부분을 집중하기 위해 Decoder가 Encoder의 정보를 직접 접근함

Attention Model

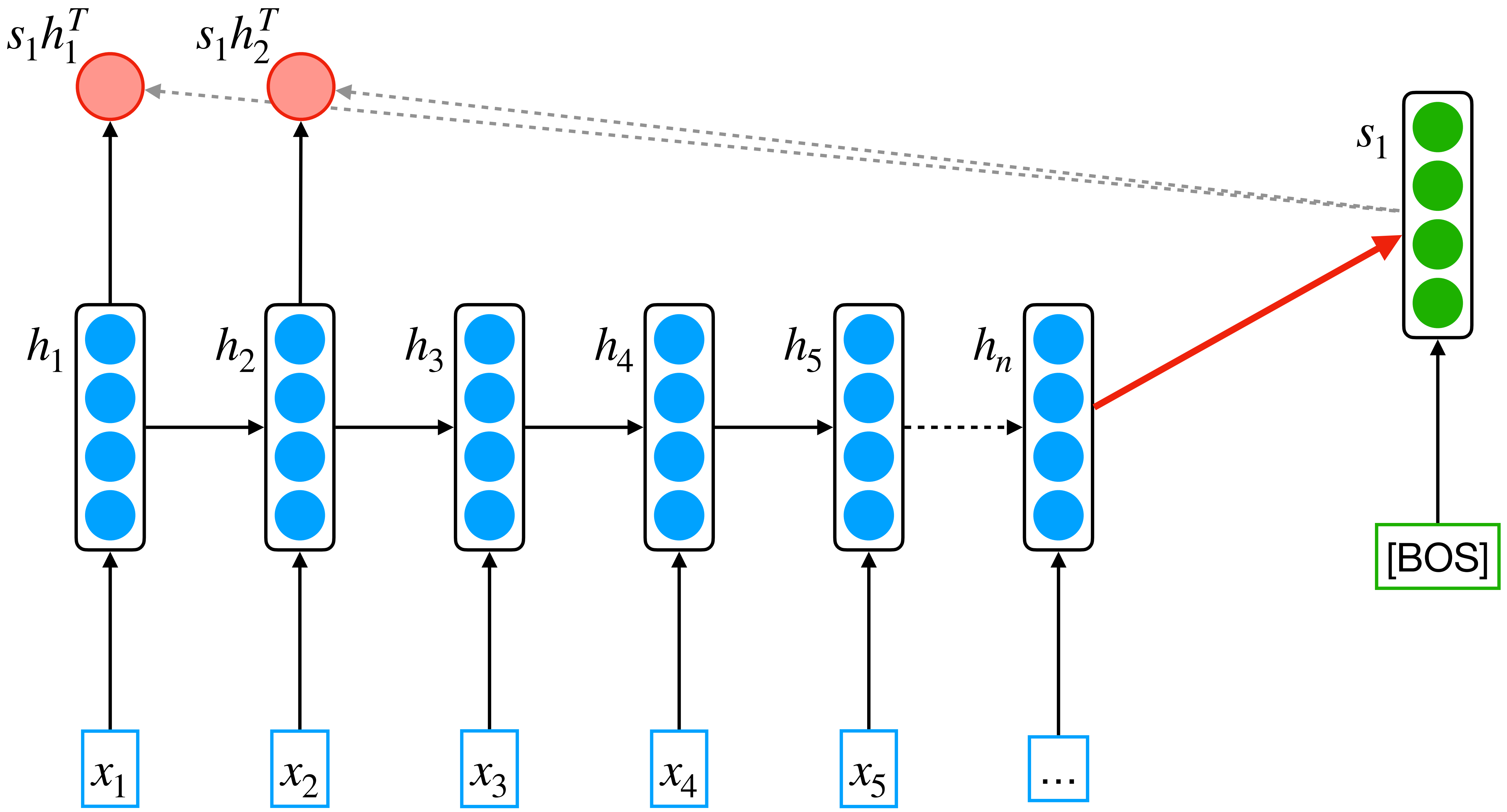


Attention Model



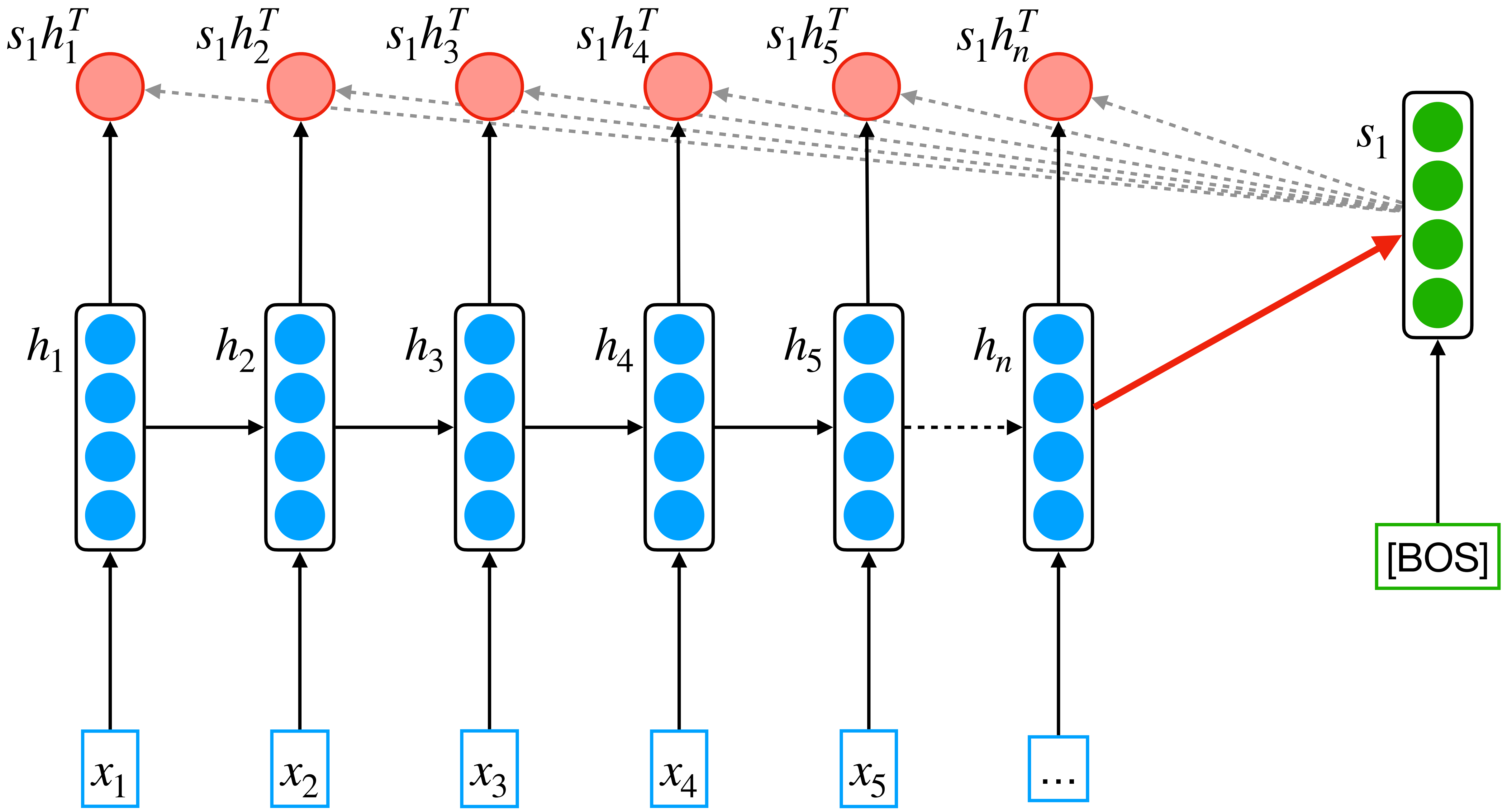
Attention Model

Dot-product

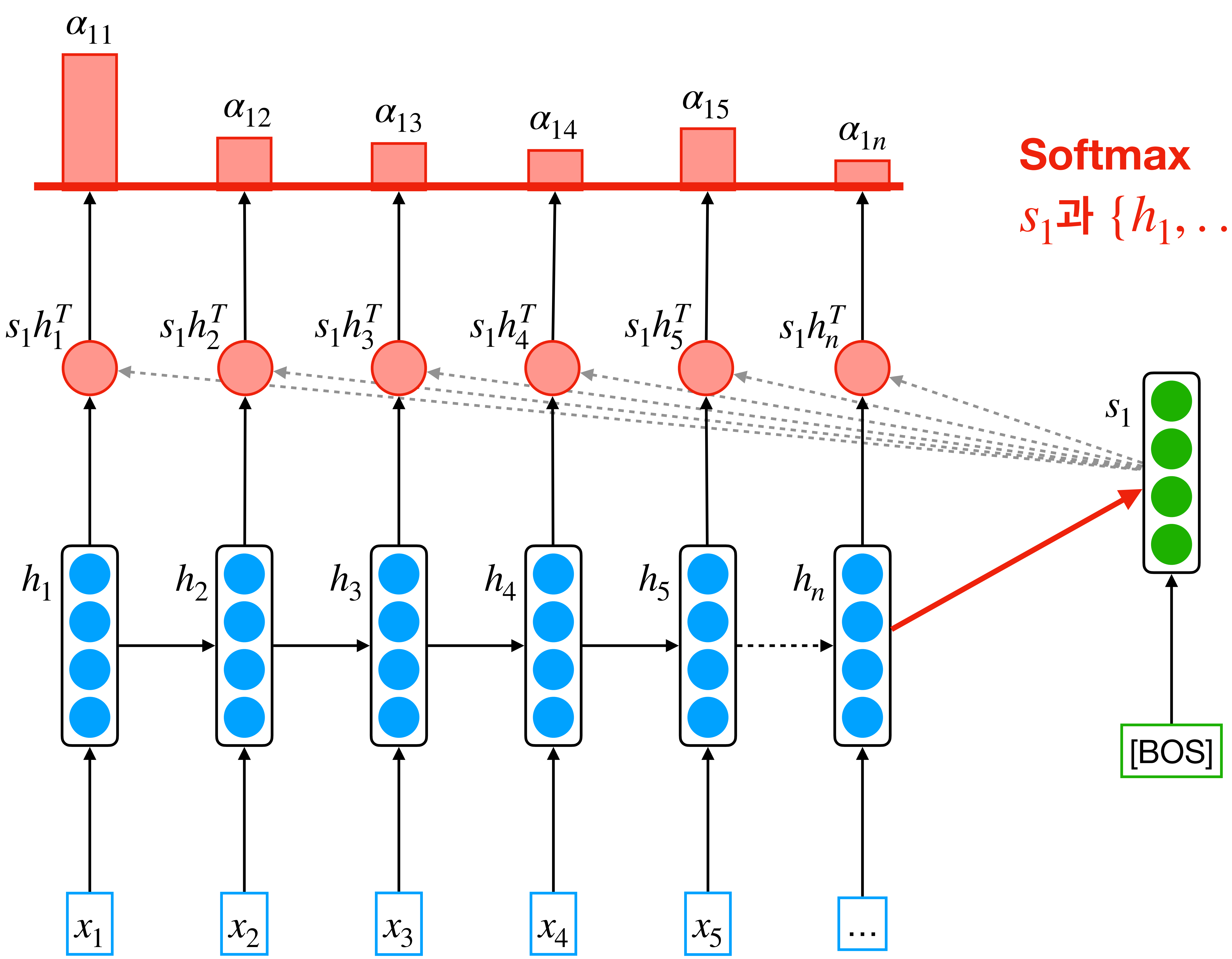


Attention Model

Dot-product



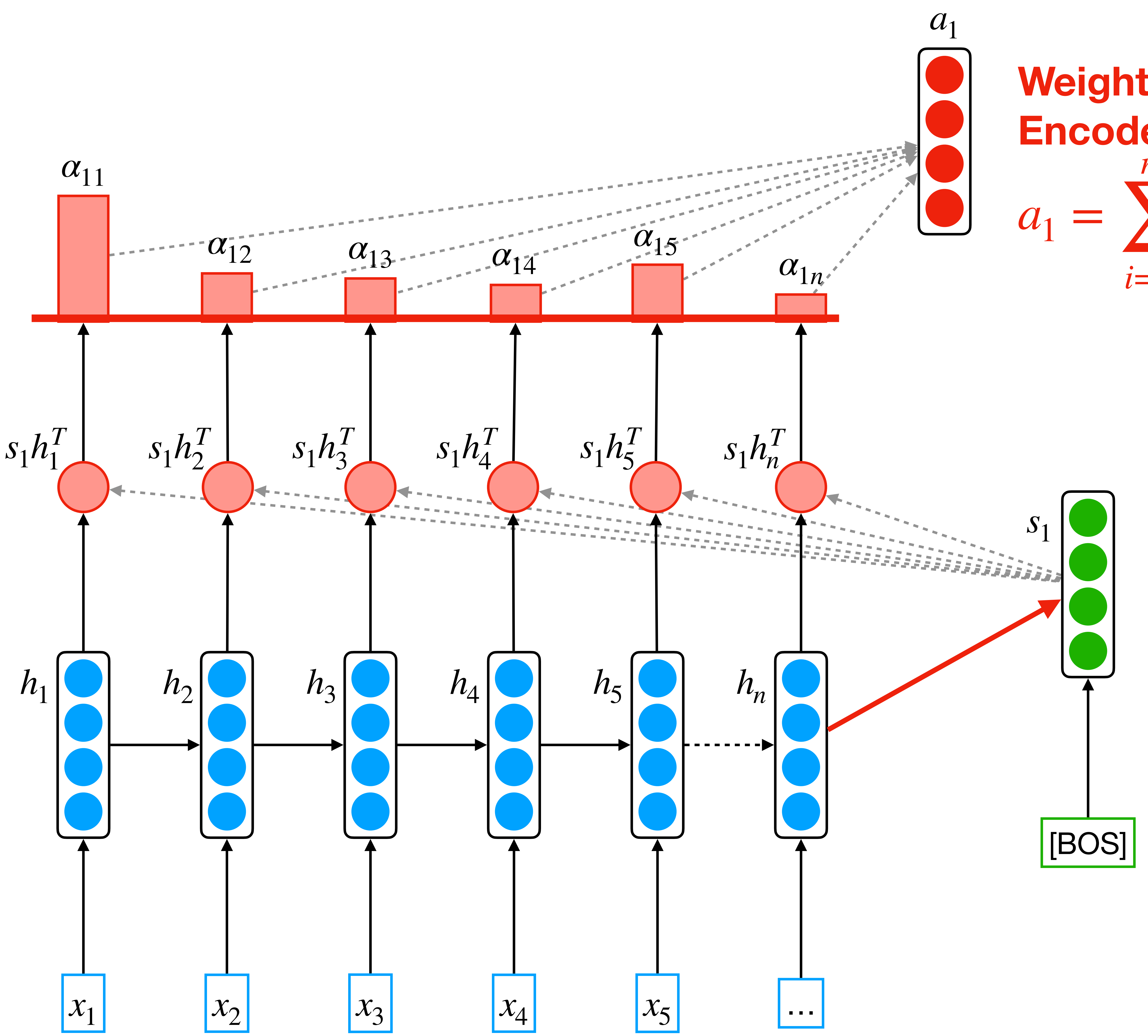
Attention Model



Softmax

s_1과 $\{h_1, .., h_n\}$의 유사도를 확률분포로 변환

Attention Model

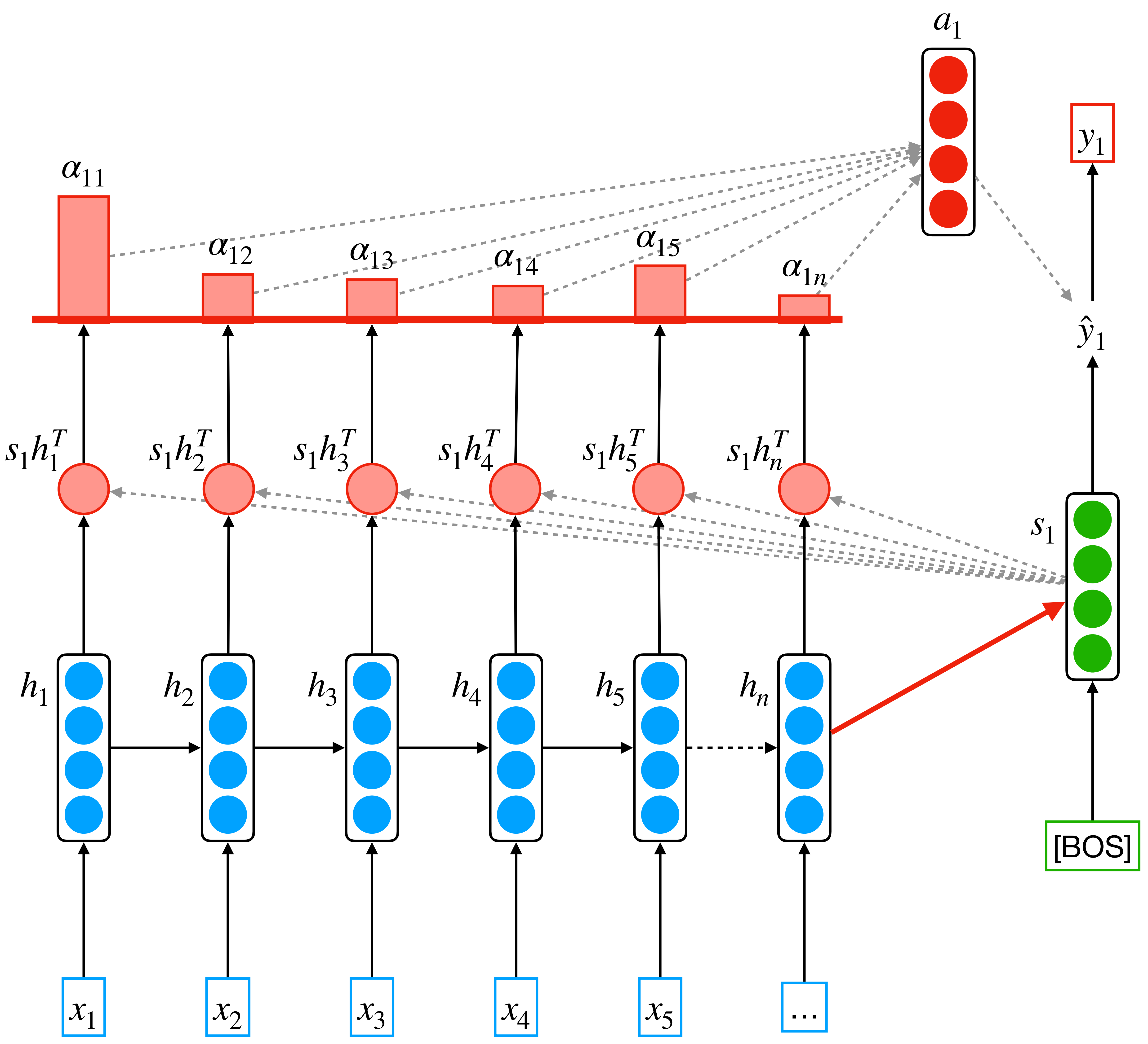


Weighted Sum

Encoder hidden state의 weighted sum

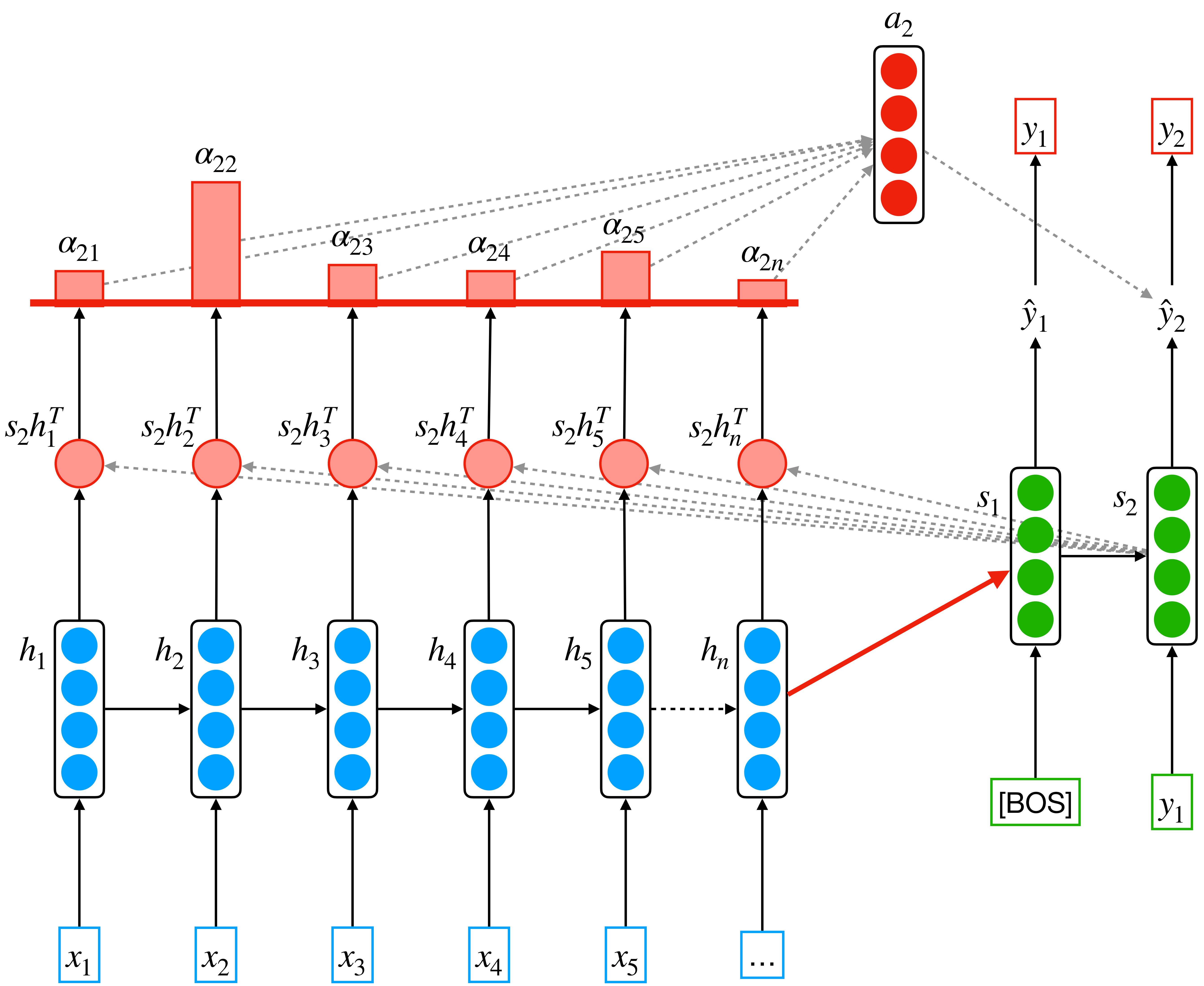
$$a_1 = \sum_{i=1}^{n} \alpha_{1i} h_i$$

Attention Model

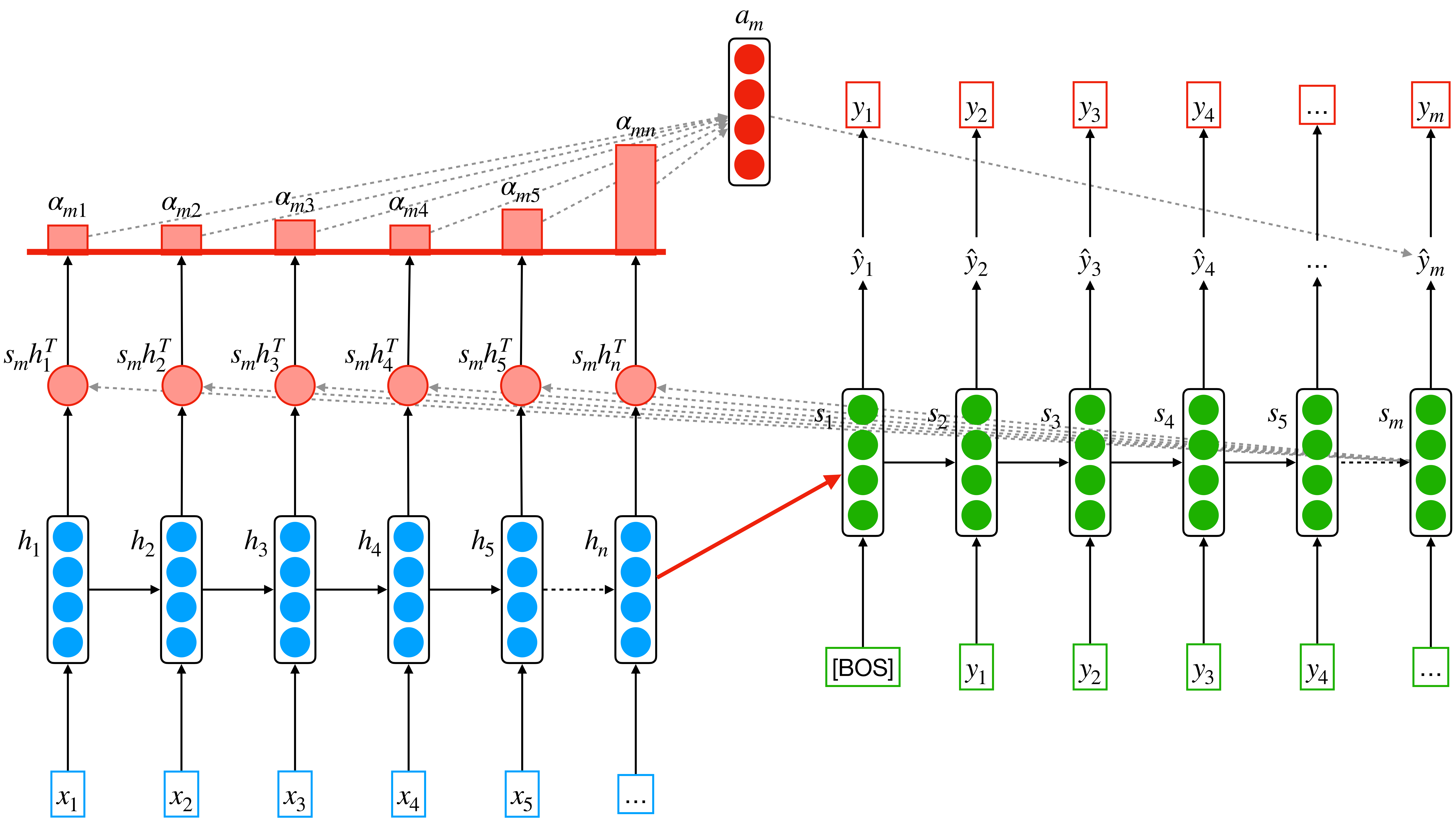


Concatenate attention output with decoder hidden state $[s_1; a_1]$ and compute output

Attention Model

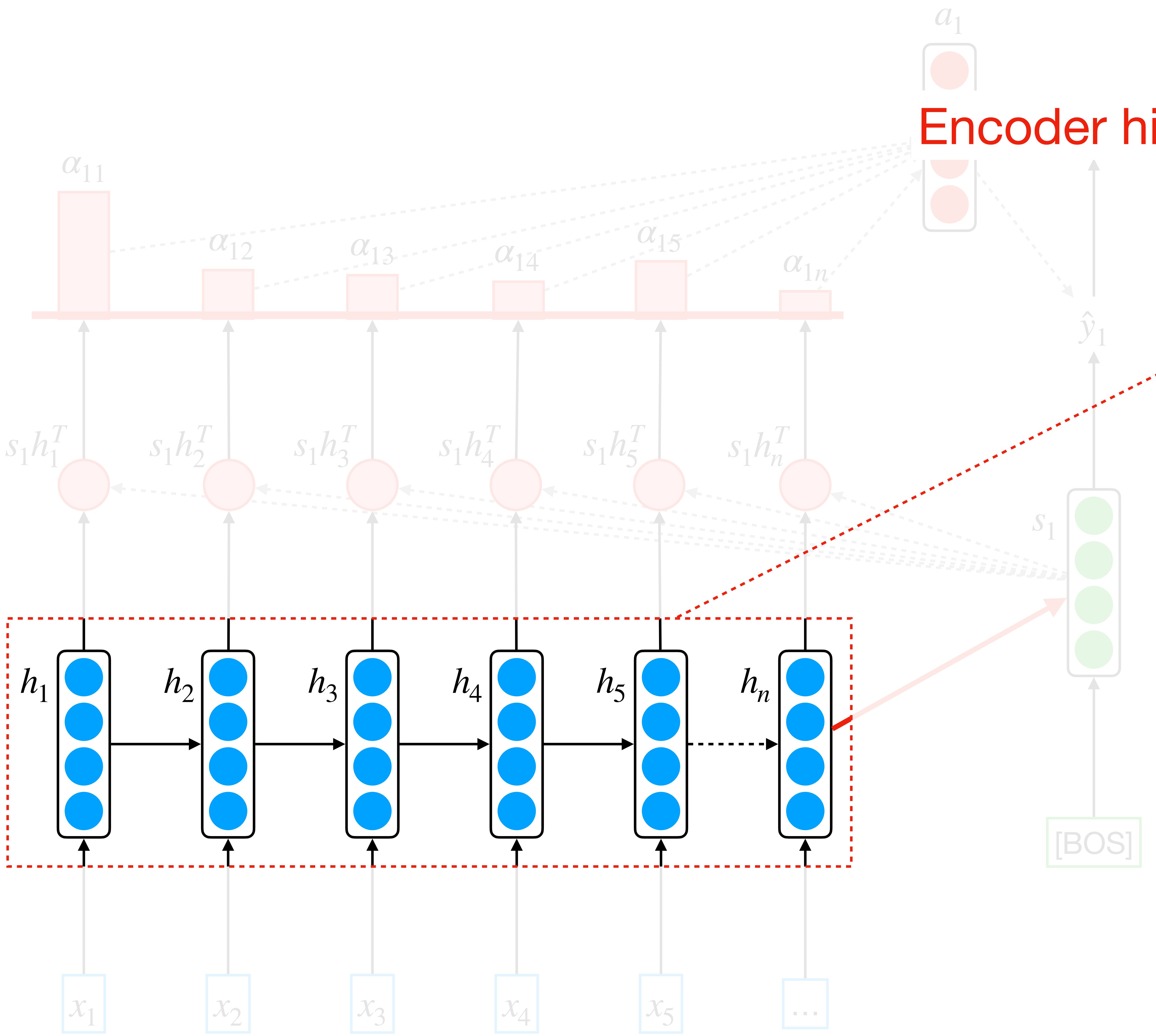


Attention Model

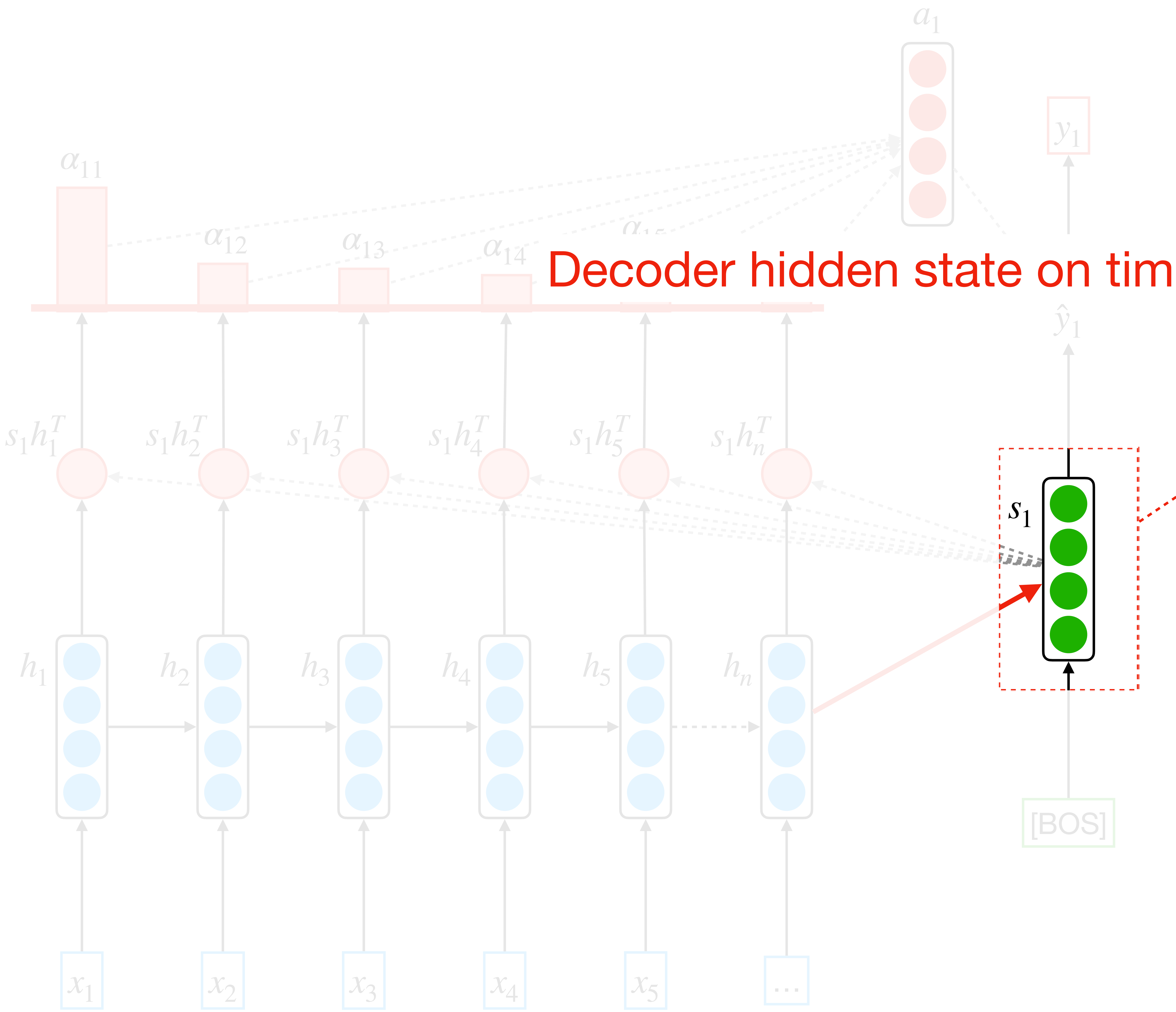


Attention Model (Equation)

Encoder hidden state: $h_1, ..., h_n \in \mathbb{R}^h$



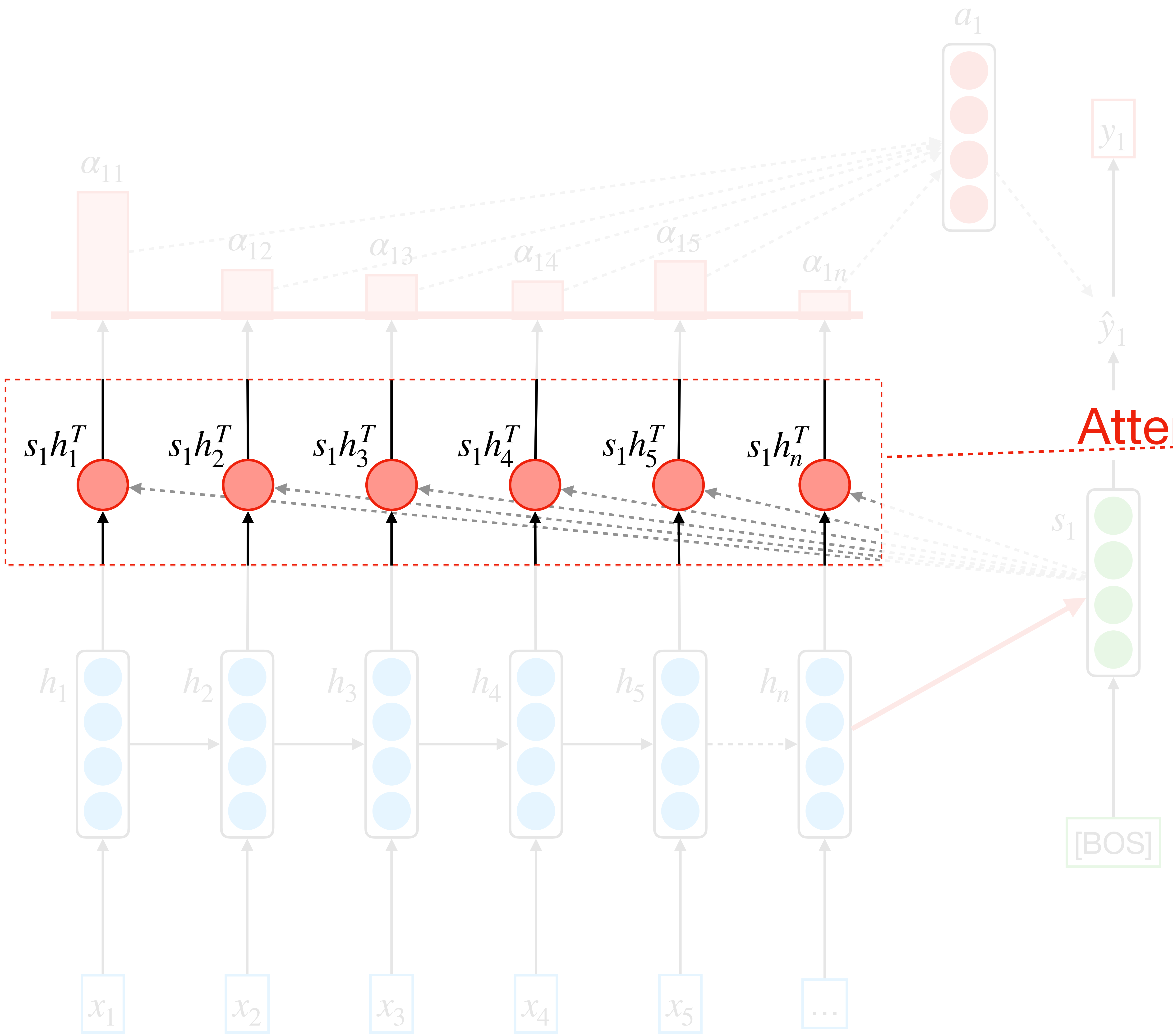
Attention Model (Equation)



$$h_1, \ldots, h_n \in \mathbb{R}^h$$

Decoder hidden state on timestamp t: $s_t \in \mathbb{R}^h$

Attention Model (Equation)

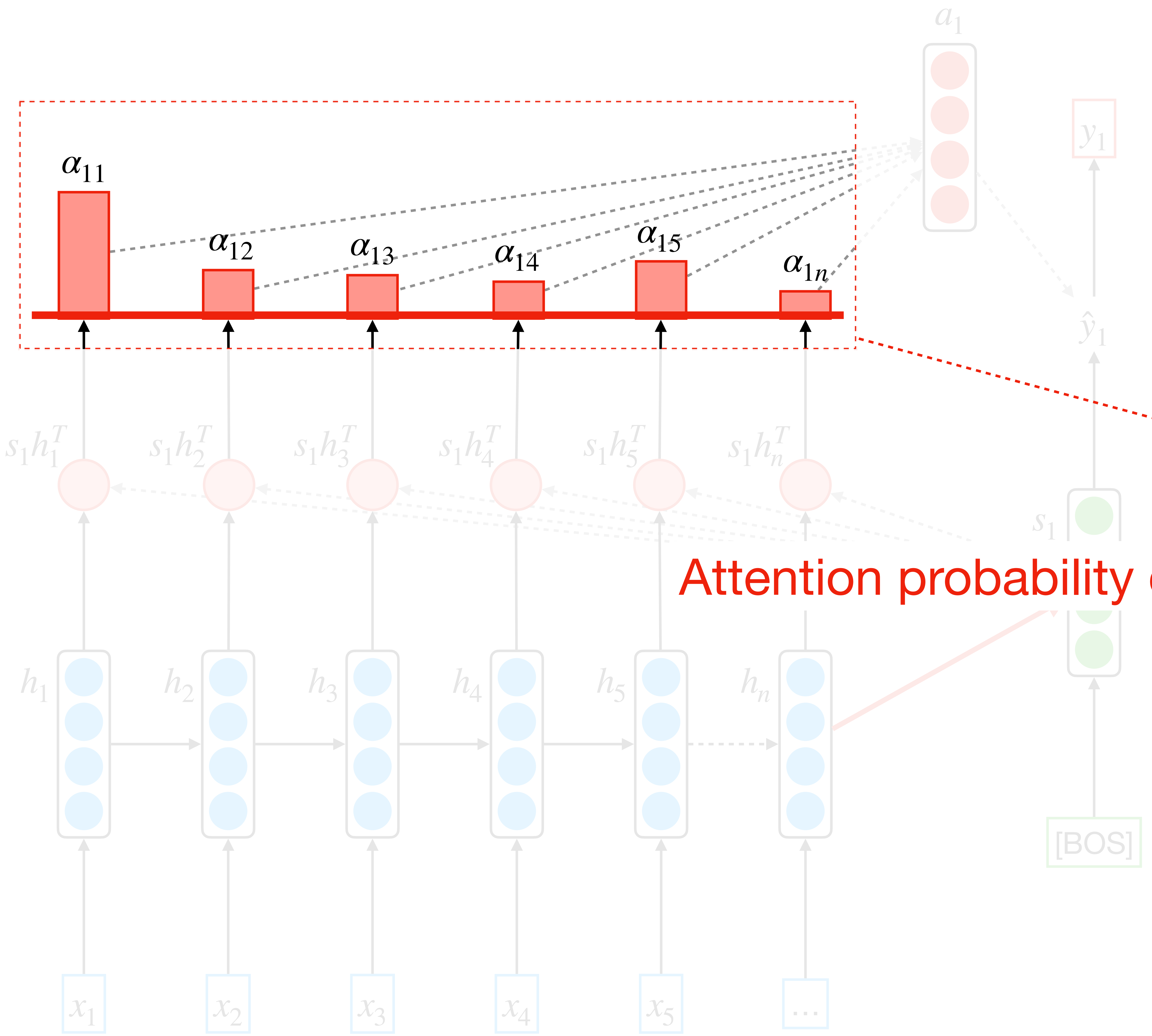


$$h_1, \ldots, h_n \in \mathbb{R}^h$$

$$s_t \in \mathbb{R}^h$$

Attention score: $e^t = [s_t^T h_1, \ldots, s_t^T h_n] \in \mathbb{R}^n$

Attention Model (Equation)



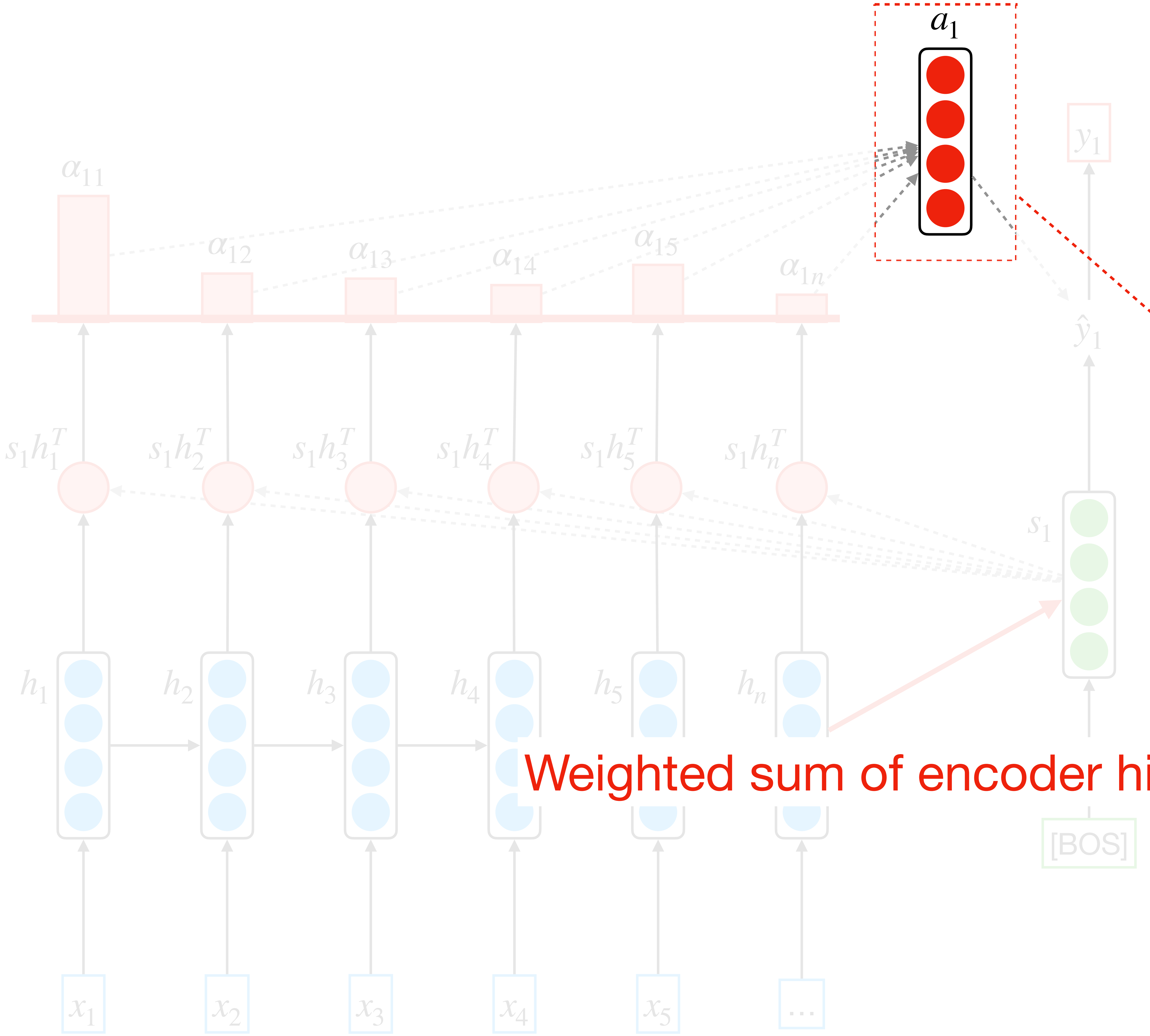
$$h_1, \ldots, h_n \in \mathbb{R}^h$$

$$s_t \in \mathbb{R}^h$$

$$e^t = [s_t^T h_1, \ldots, s_t^T h_n] \in \mathbb{R}^n$$

Attention probability distribution: $\alpha^t = softmax(e^t) \in \mathbb{R}^n$

Attention Model (Equation)



$$h_1, \ldots, h_n \in \mathbb{R}^h$$

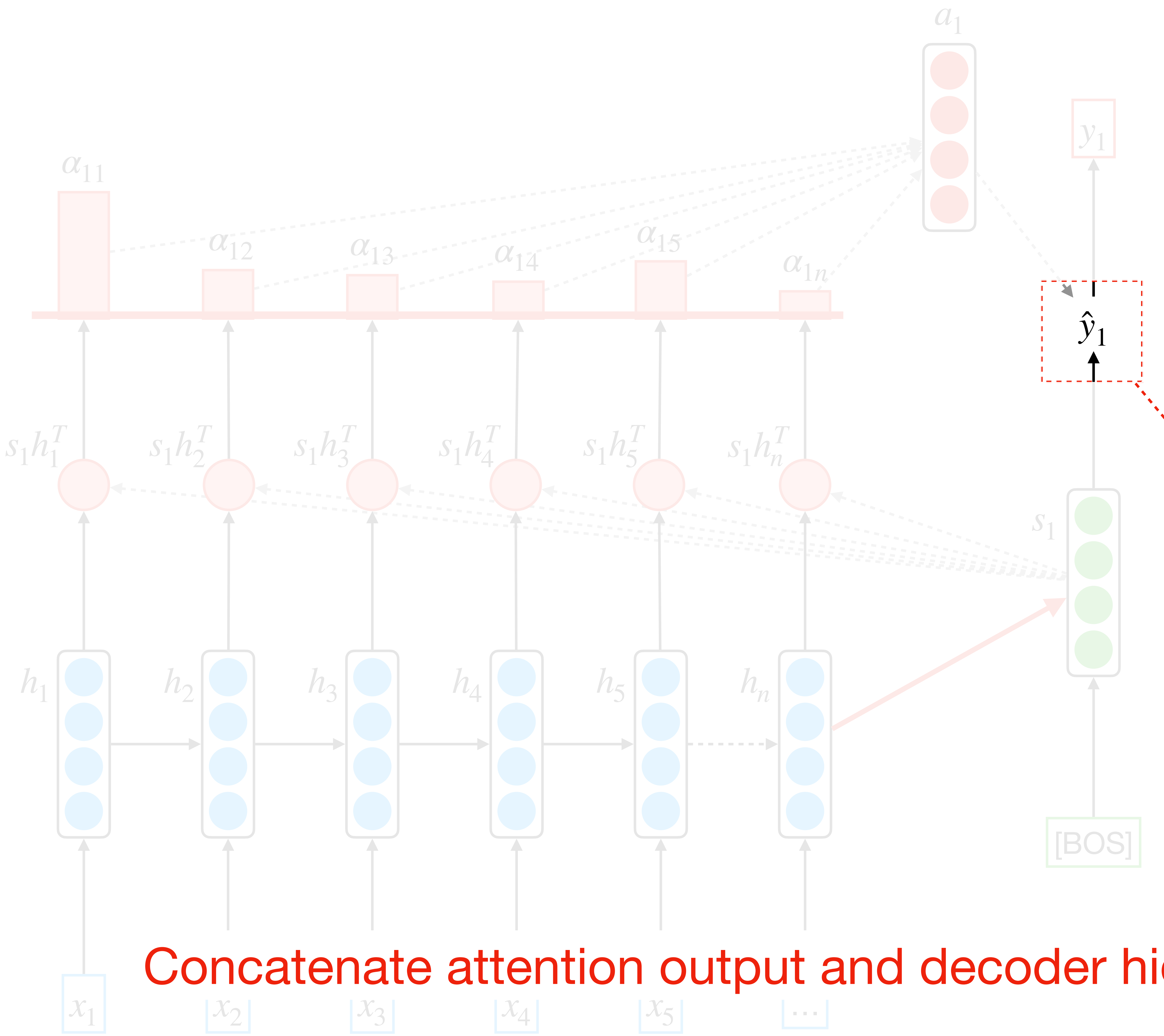
$$s_t \in \mathbb{R}^h$$

$$e^t = [s_t^T h_1, \ldots, s_t^T h_n] \in \mathbb{R}^n$$

$$\alpha^t = softmax(e^t) \in \mathbb{R}^n$$

Weighted sum of encoder hidden state: $$a_t = \sum_{i=1}^{n} \alpha_i^t h_i \in \mathbb{R}^h$$

Attention Model (Equation)



$$h_1, \ldots, h_n \in \mathbb{R}^h$$

$$s_t \in \mathbb{R}^h$$

$$e^t = [s_t^T h_1, \ldots, s_t^T h_n] \in \mathbb{R}^n$$

$$\alpha^t = softmax(e^t) \in \mathbb{R}^n$$

$$a_t = \sum_{i=1}^{n} \alpha_i^t h_i \in \mathbb{R}^h$$

Concatenate attention output and decoder hidden state: $[a_t; s_t] \in \mathbb{R}^2 h$

Attention Model (Equation)

Encoder hidden state: $h_1, \ldots, h_n \in \mathbb{R}^h$

Decoder hidden state on timestamp t: $s_t \in \mathbb{R}^h$

Attention score: $e^t = [s_t^T h_1, \ldots, s_t^T h_n] \in \mathbb{R}^n$

Attention probability distribution: $\alpha^t = softmax(e^t) \in \mathbb{R}^n$

Weighted sum of encoder hidden state: $a_t = \sum_{i=1}^{n} \alpha_i^t h_i \in \mathbb{R}^h$

Concatenate attention output and decoder hidden state: $[a_t; s_t] \in \mathbb{R}^2 h$

Attention Model (Advantage)

- Attention을 이용해 NMT의 성능이 많이 좋아짐
  - Decoder가 source의 특정 부분에 집중하도록 한 것이 매우 효과적임
- Information bottleneck 문제를 해결 함
  - Decoder가 source에 직접 접근하도록 함
- Vanishing gradient 문제를 해결 함
  - 거리가 먼 source의 정보를 접근 할 수 있음
- Attention이 alignment를 학습함

	Education	is	most	powerful	weapon
교육은	■				
가장			■		
강력한				■	
무기					■
입니다		■			

Attention Model (Variants)

Values → $h_1, \ldots, h_n \in \mathbb{R}^h$ $s_t \in \mathbb{R}^h$

Attention score → $e^t \in \mathbb{R}^n$

attention score를 계산하는
다양한 방법이 있음

Attention probability → $\alpha^t = softmax(e^t) \in \mathbb{R}^n$

Attention output → $a_t = \sum_{i=1}^{n} \alpha_i^t h_i \in \mathbb{R}^h$

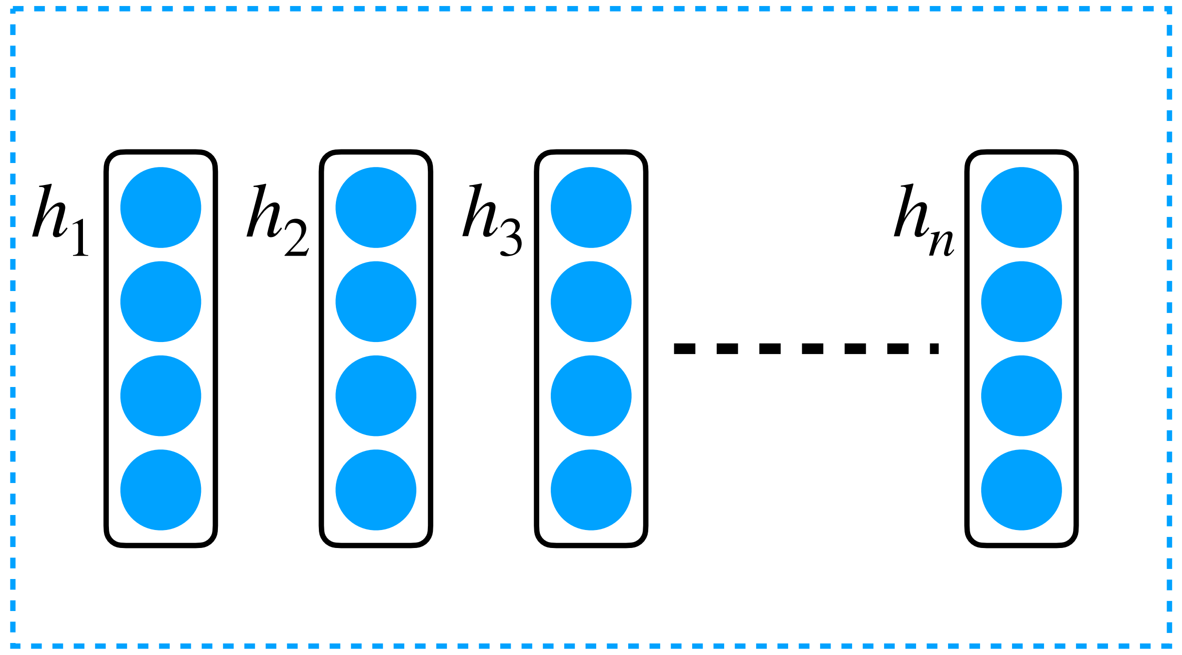
Attention Model (Variants)

$$e_t = [e_{t1}, \ldots, e_{tn}] \in \mathbb{R}^n$$

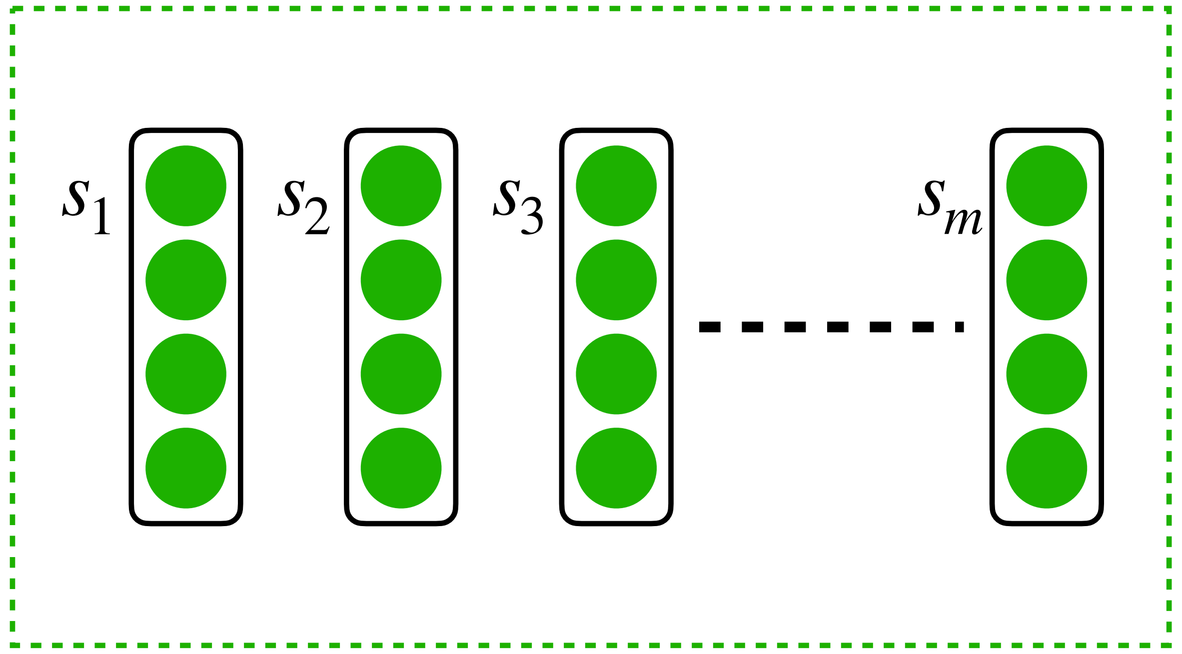
- Dot-product attention
  - $e_i^t = s_t^T h_i \in \mathbb{R}$
- Multiplicative attention
  - $e_i^t = s_t^T W h_i \in \mathbb{R}$
  - where $W \in \mathbb{R}^{d_s \times d_h}$
- Additive attention
  - $e_i^t = v^T \tanh(W_h h_i + W_s s_t) \in \mathbb{R}$
  - where $W_h \in \mathbb{R}^{d_v \times d_h}, W_s \in \mathbb{R}^{d_v \times d_s}, v \in \mathbb{R}^{d_v}$

Attention Tutorial

Attention Tutorial (inputs)



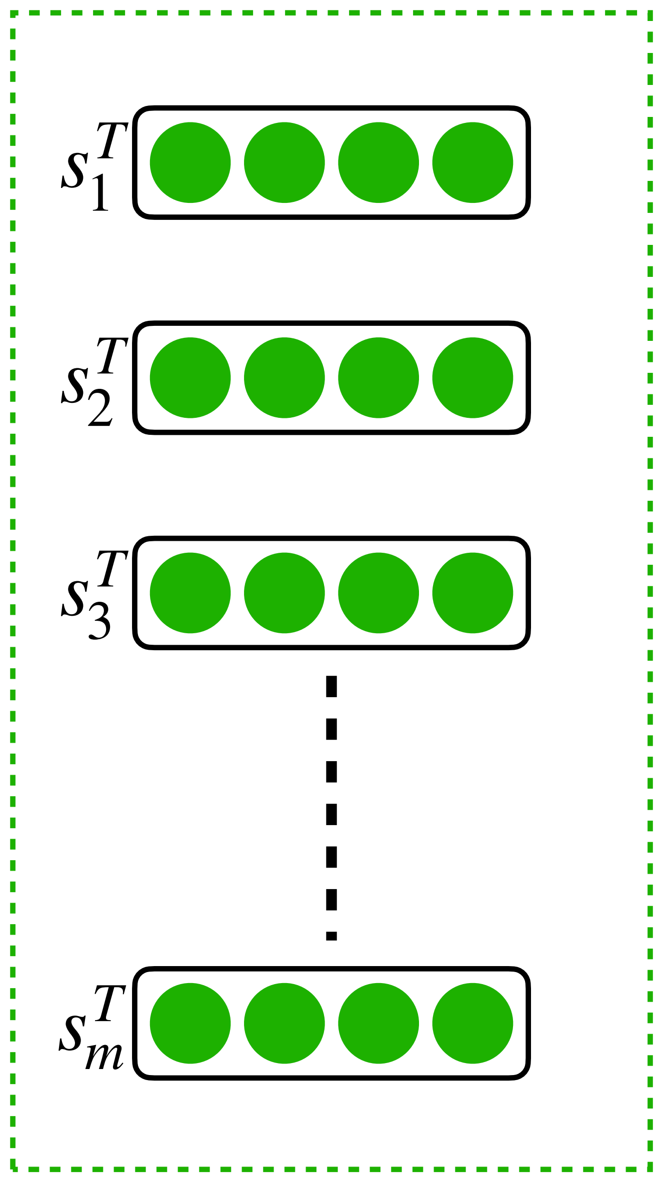
Encoder hidden state: $h \in \mathbb{R}^{h \times n}$



Decoder hidden state: $s \in \mathbb{R}^{h \times m}$

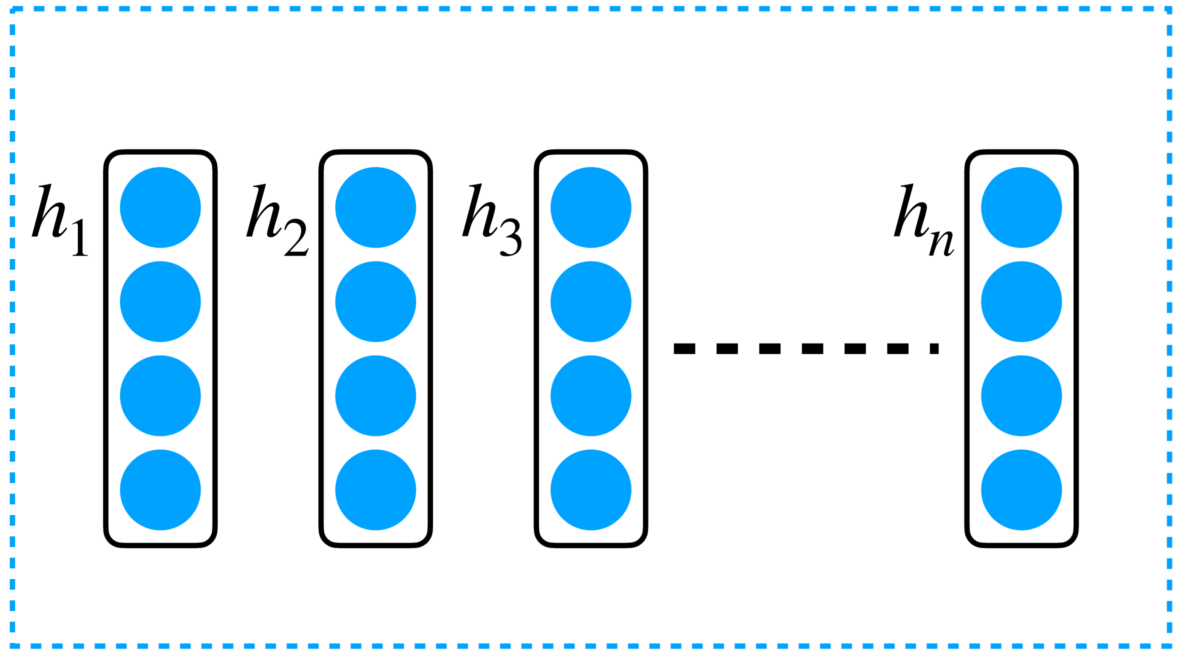
Attention Tutorial (score)

$$e = s^T h \in \mathbb{R}^{m \times n}$$



Decoder hidden state: $s^T \in \mathbb{R}^{m \times h}$

$\times$



Encoder hidden state: $h \in \mathbb{R}^{h \times n}$

$=$

	h_1	h_2	h_3	…	h_n
s_1^T	$s_1^T h_1$	$s_1^T h_2$	$s_1^T h_3$	…	$s_1^T h_n$
s_2^T	$s_2^T h_1$	$s_2^T h_2$	$s_2^T h_3$	…	$s_2^T h_n$
s_3^T	$s_3^T h_1$	$s_3^T h_2$	$s_3^T h_3$	…	$s_3^T h_n$
…	…	…	…	…	…
s_m^T	$s_m^T h_1$	$s_m^T h_2$	$s_m^T h_3$	…	$s_m^T h_n$

Attention score: $e \in \mathbb{R}^{m \times n}$

$$e_j^i = s_i^T h_j \in \mathbb{R}$$

Attention Tutorial (prob)

$$\alpha = softmax(e) \in \mathbb{R}^{m \times n}$$

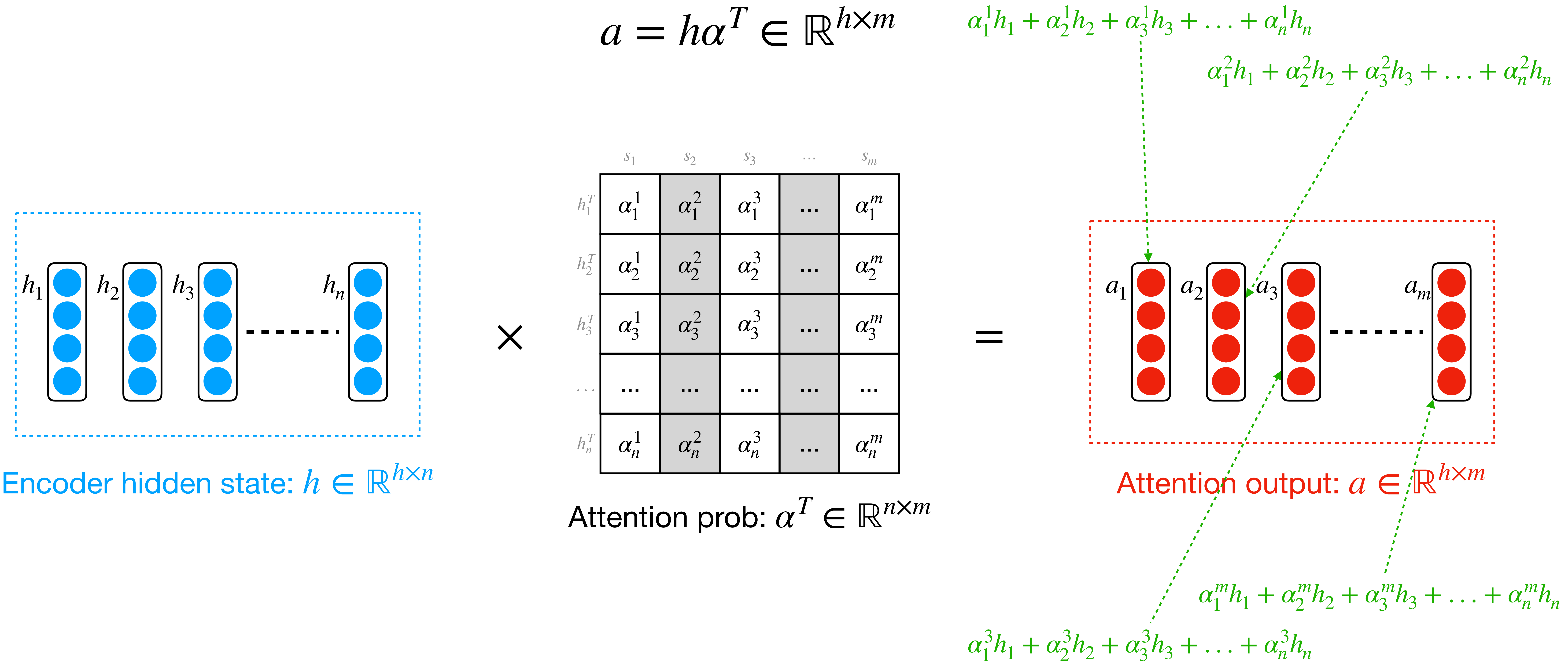
$$
softmax\left(
\begin{array}{c|ccccc|}
 & h_1 & h_2 & h_3 & \cdots & h_n \\
\hline
s_1^T & s_1^T h_1 & s_1^T h_2 & s_1^T h_3 & \dots & s_1^T h_n \\
s_2^T & s_2^T h_1 & s_2^T h_2 & s_2^T h_3 & \dots & s_2^T h_n \\
s_3^T & s_3^T h_1 & s_3^T h_2 & s_3^T h_3 & \dots & s_3^T h_n \\
\cdots & \dots & \dots & \dots & \dots & \dots \\
s_m^T & s_m^T h_1 & s_m^T h_2 & s_m^T h_3 & \dots & s_m^T h_n \\
\hline
\end{array}
\right)
=
\begin{array}{c|ccccc|}
 & h_1 & h_2 & h_3 & \cdots & h_n \\
\hline
s_1^T & \alpha_1^1 & \alpha_2^1 & \alpha_3^1 & \dots & \alpha_n^1 \\
s_2^T & \alpha_1^2 & \alpha_2^2 & \alpha_3^2 & \dots & \alpha_n^2 \\
s_3^T & \alpha_1^3 & \alpha_2^3 & \alpha_3^3 & \dots & \alpha_n^3 \\
\cdots & \dots & \dots & \dots & \dots & \dots \\
s_m^T & \alpha_1^m & \alpha_2^m & \alpha_3^m & \dots & \alpha_n^m \\
\hline
\end{array}
$$

Attention score: $e \in \mathbb{R}^{m \times n}$

Attention prob: $\alpha \in \mathbb{R}^{m \times n}$

행 단위 softmax

Attention Tutorial (output)



감사합니다.