

# Self-Attention

自注意力机制  
发表于2016年  
用在了LSTM上

Shusen Wang

# Self-Attention

- Self-Attention [2]: attention [1] beyond Seq2Seq models.
- The original self-attention paper uses LSTM.
- To make teaching easy, I replace LSTM by SimpleRNN.

## Original paper:

1. Bahdanau, Cho, & Bengio. Neural machine translation by jointly learning to align and translate. In *ICLR*, 2015.
2. Cheng, Dong, & Lapata. Long Short-Term Memory-Networks for Machine Reading. In *EMNLP*, 2016.

# SimpleRNN + Self-Attention

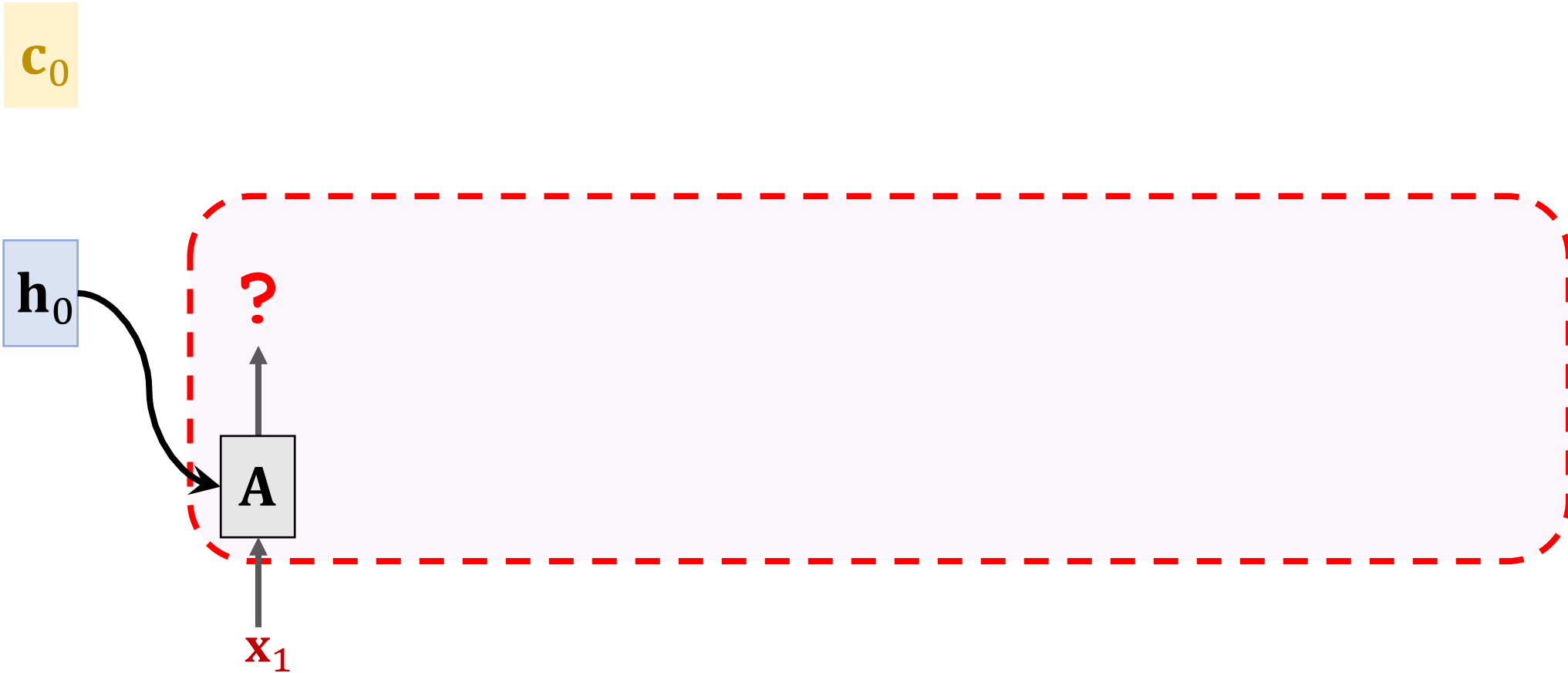
初始条件

$$\mathbf{c}_0 = \mathbf{0}$$

$$\mathbf{h}_0 = \mathbf{0}$$



# SimpleRNN + Self-Attention



# SimpleRNN + Self-Attention

SimpleRNN:

$$\mathbf{h}_1 = \tanh \left( \mathbf{A} \cdot \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{h}_0 \end{bmatrix} + \mathbf{b} \right)$$

$\mathbf{c}_0$



# SimpleRNN + Self-Attention

SimpleRNN:

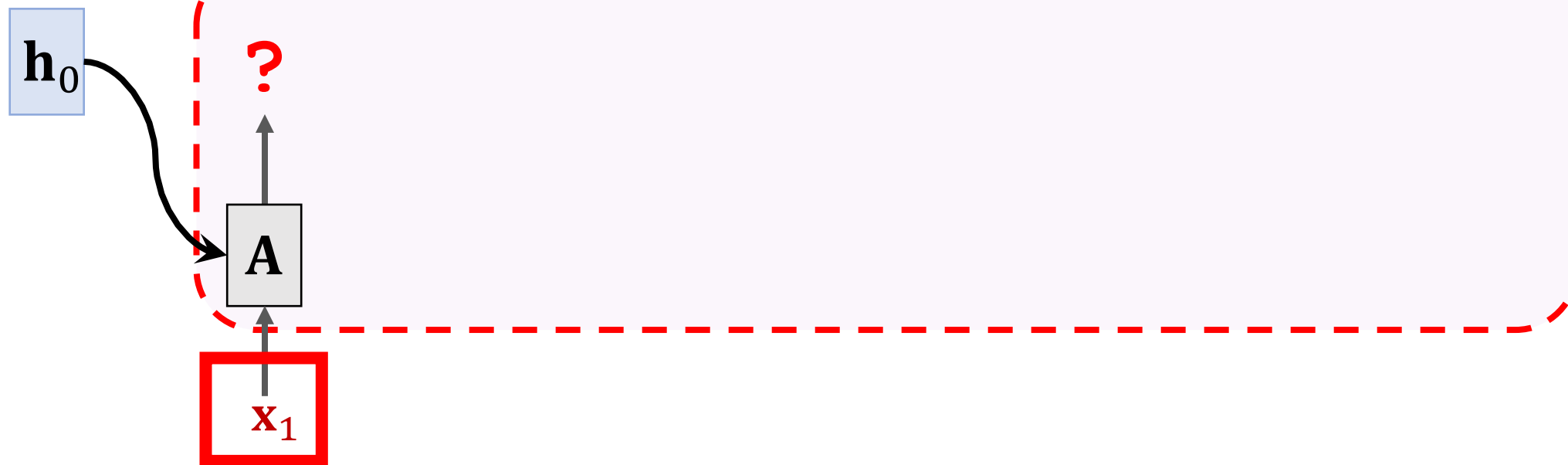
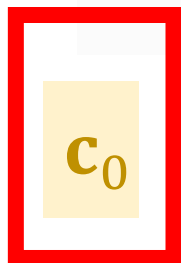
$$\mathbf{h}_1 = \tanh \left( \mathbf{A} \cdot \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{h}_0 \end{bmatrix} + \mathbf{b} \right)$$

SimpleRNN + Self-Attention:

$$\mathbf{h}_1 = \tanh \left( \mathbf{A} \cdot \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{c}_0 \end{bmatrix} + \mathbf{b} \right)$$

更新状态时，直接把h0 换成了 c0

当然也可以 把 x1 c0 h0 一起concat



# SimpleRNN + Self-Attention

$\mathbf{c}_0$

?

$\mathbf{h}_0$  是全零向量

$\mathbf{h}_0$

$\mathbf{h}_1$

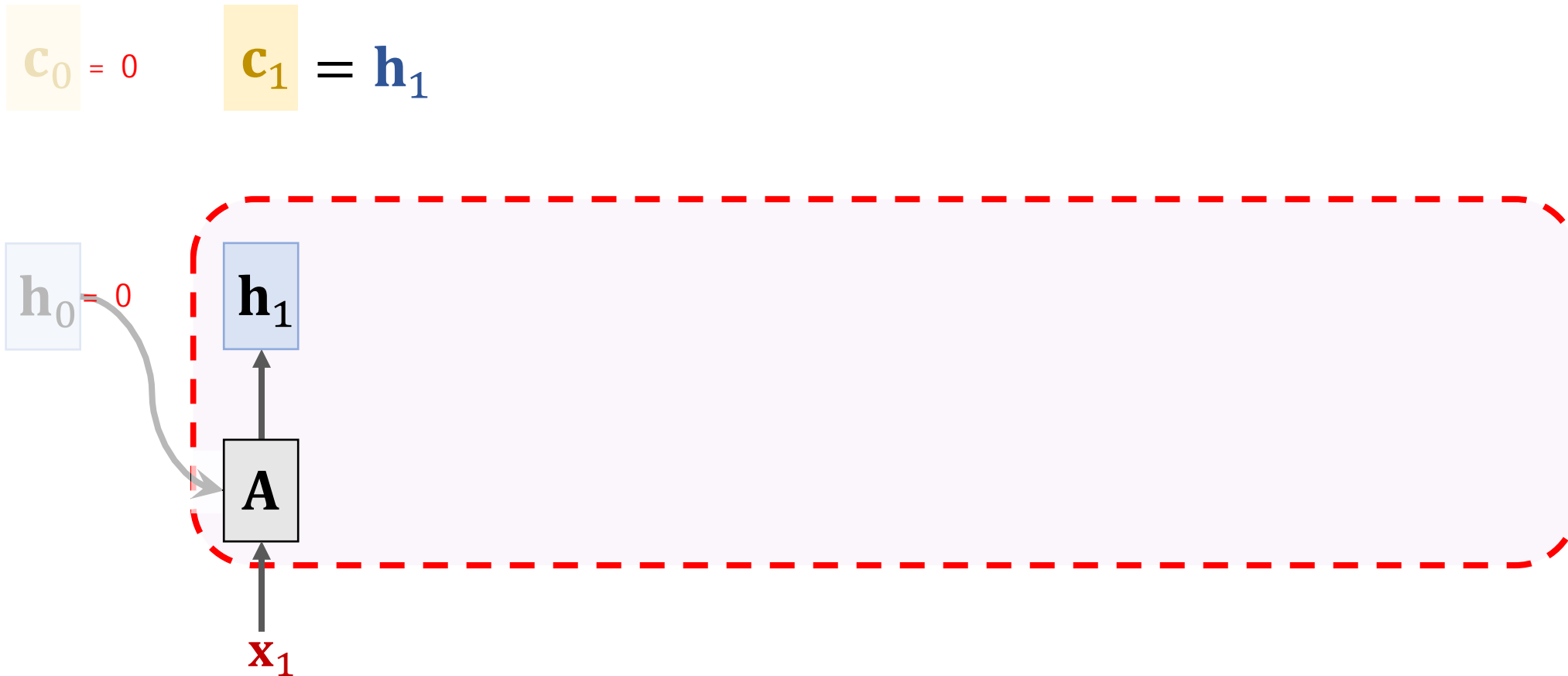
$\mathbf{A}$

$\mathbf{x}_1$

SimpleRNN + Self-Attention:

$$\mathbf{h}_1 = \tanh \left( \mathbf{A} \cdot \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{c}_0 \end{bmatrix} + \mathbf{b} \right)$$

# SimpleRNN + Self-Attention



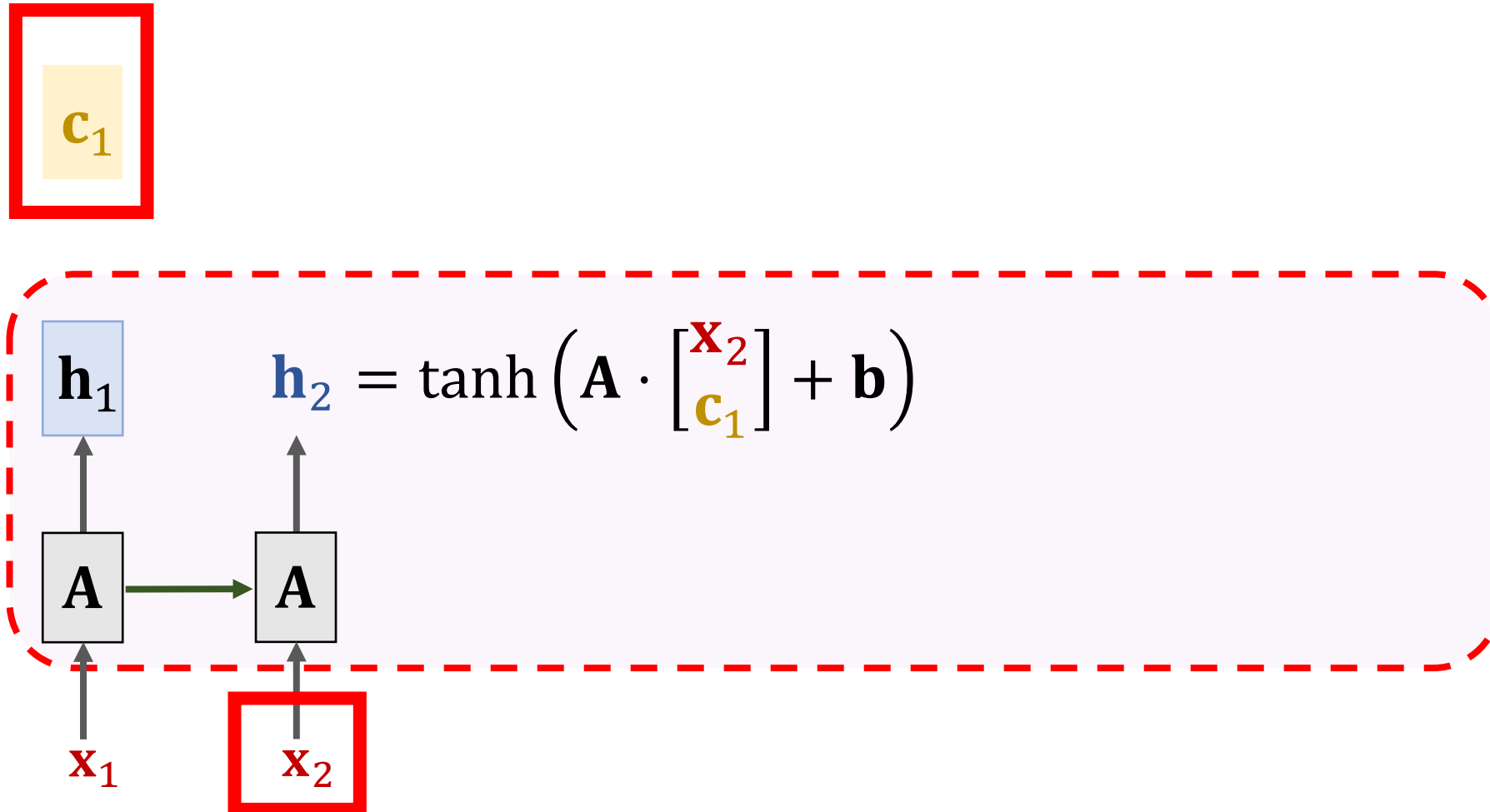


# SimpleRNN + Self-Attention

$c_1$



# SimpleRNN + Self-Attention



# SimpleRNN + Self-Attention

$c_1$

?

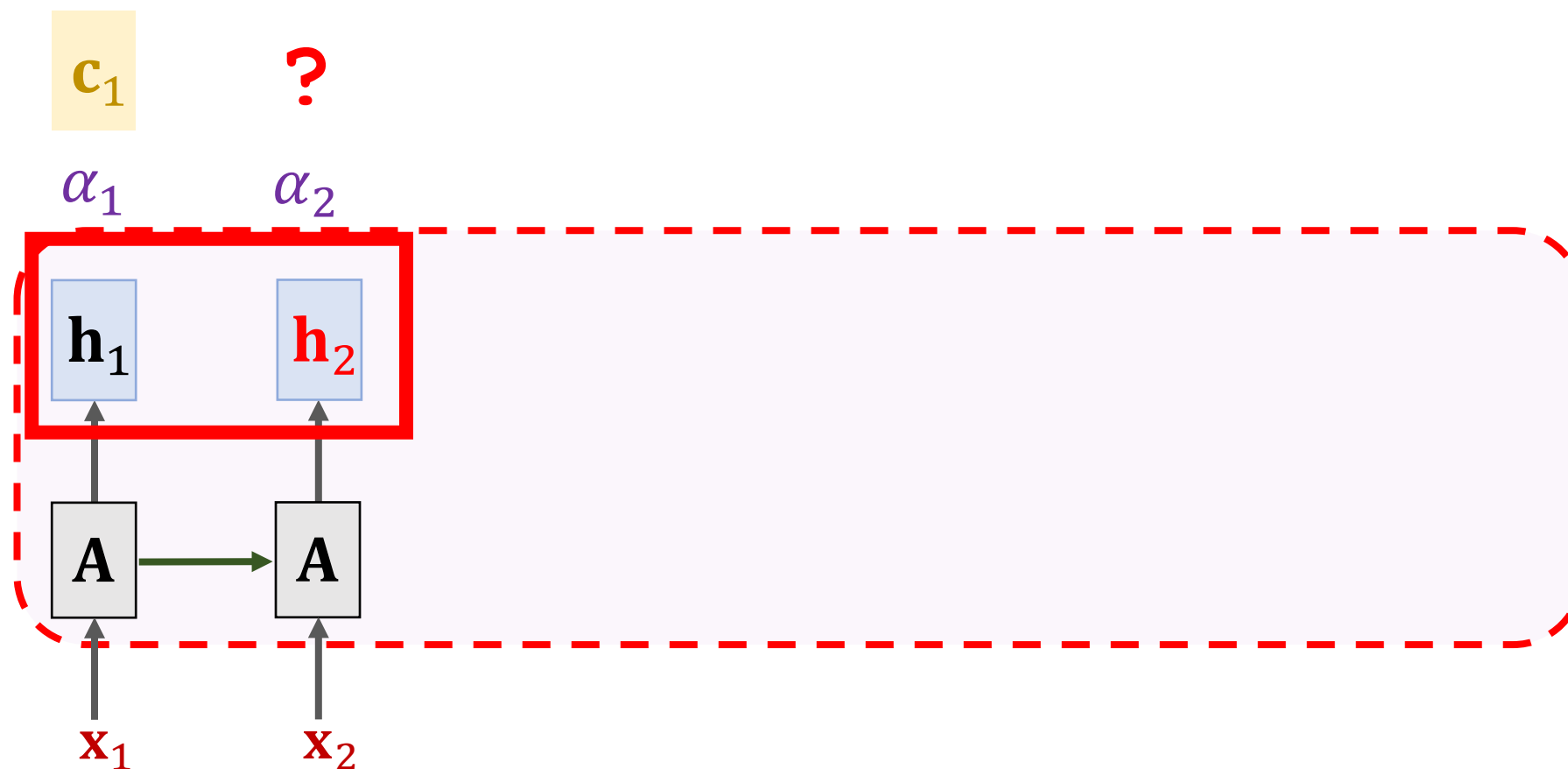


# SimpleRNN + Self-Attention

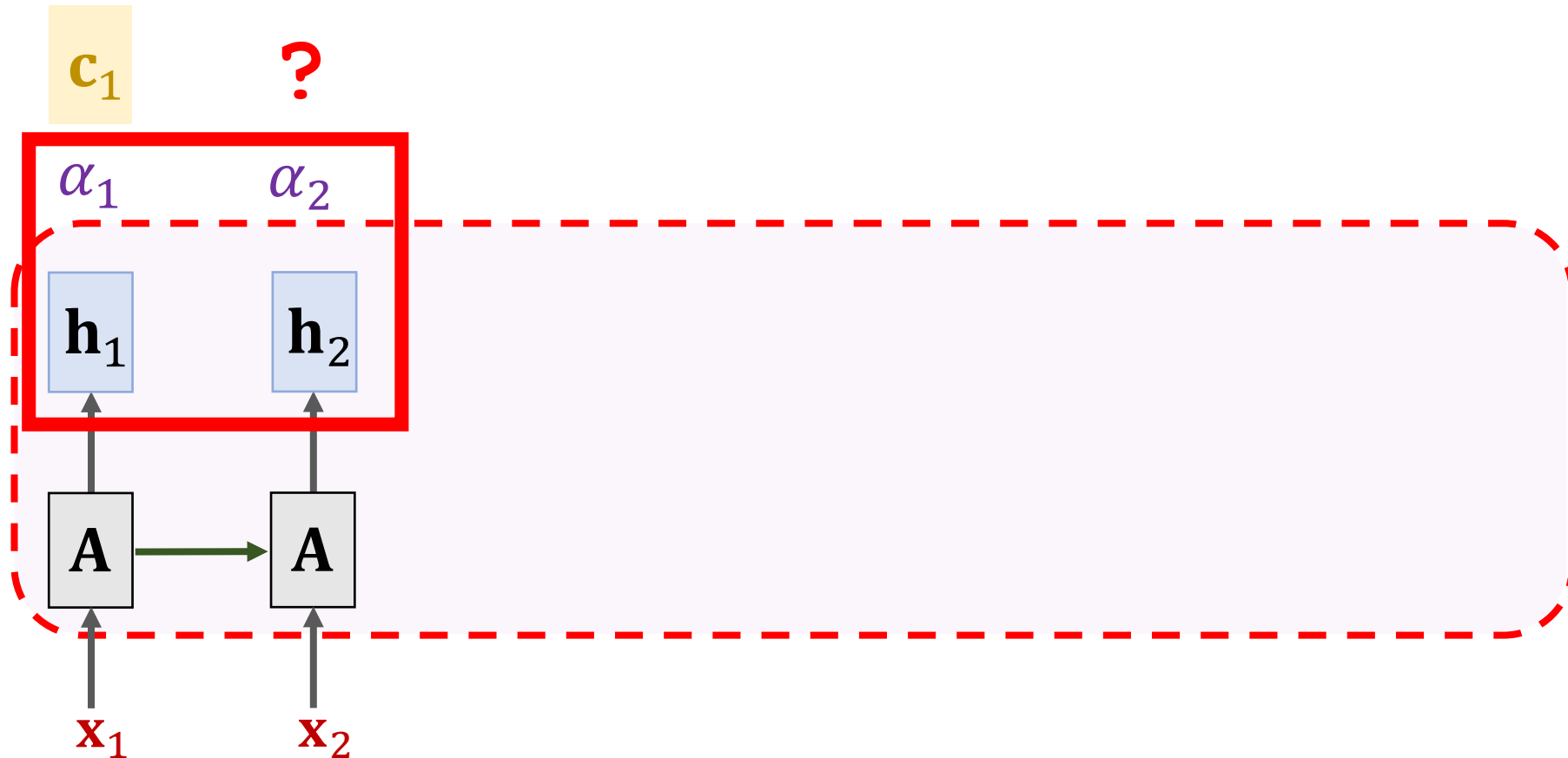
Weights:  $\alpha_i = \text{align}(\mathbf{h}_i, \mathbf{h}_2)$ . 参照第9\_8中的内容

计算出结果

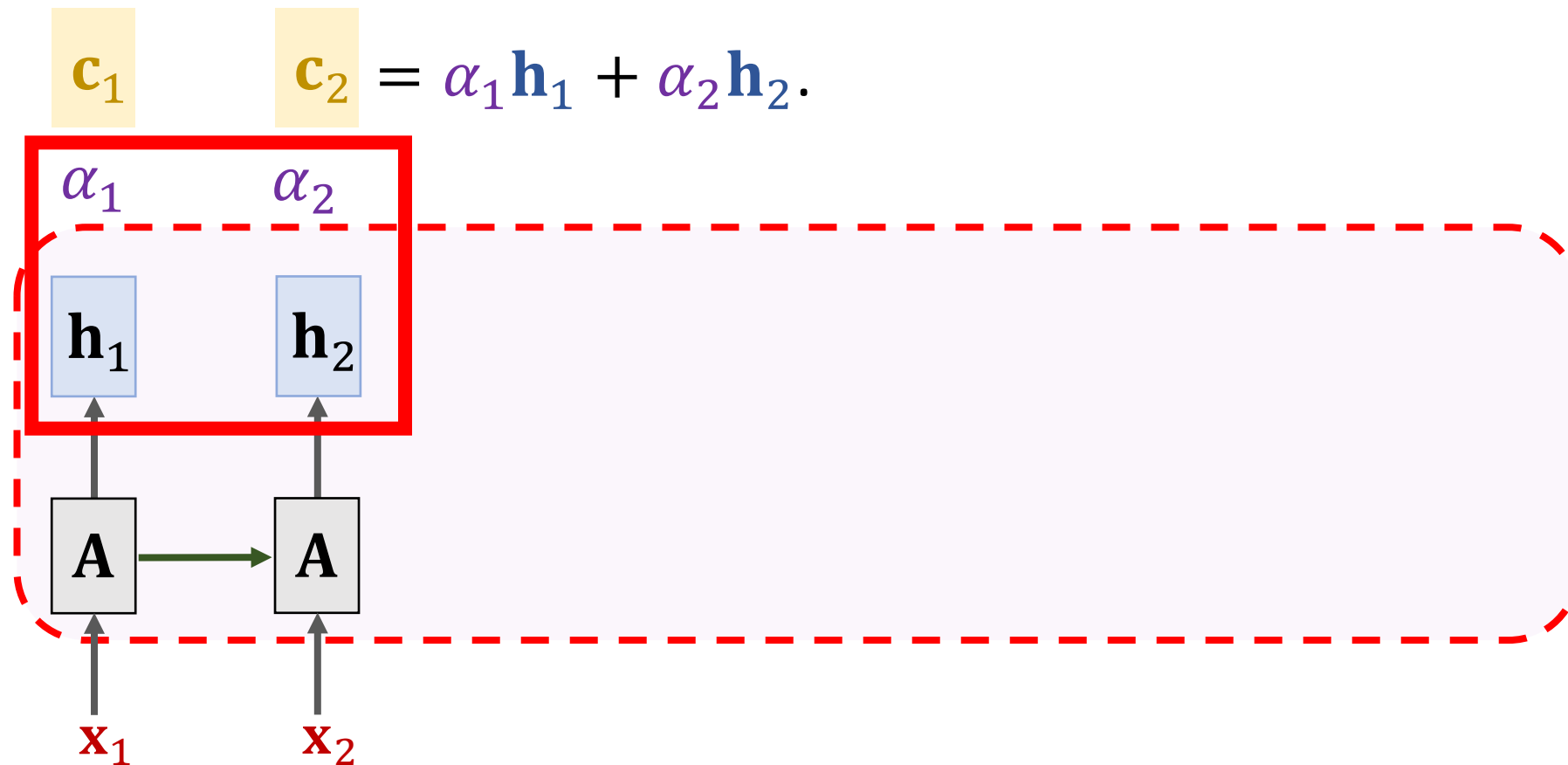
这里可以看到：  
注意力机制是添加在编码器上的  
align 中标红的 h 是和相同的状态进行比较



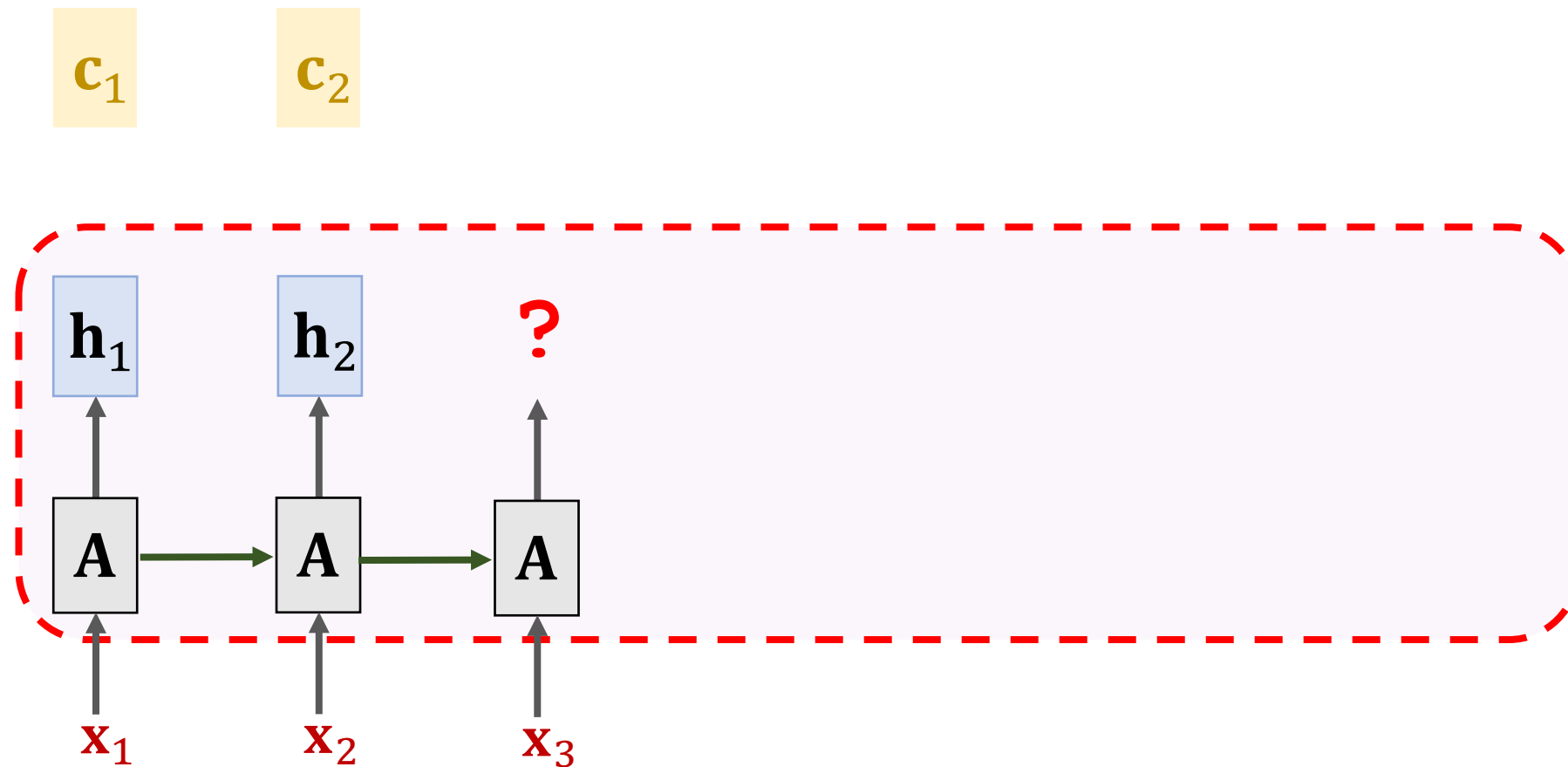
# SimpleRNN + Self-Attention



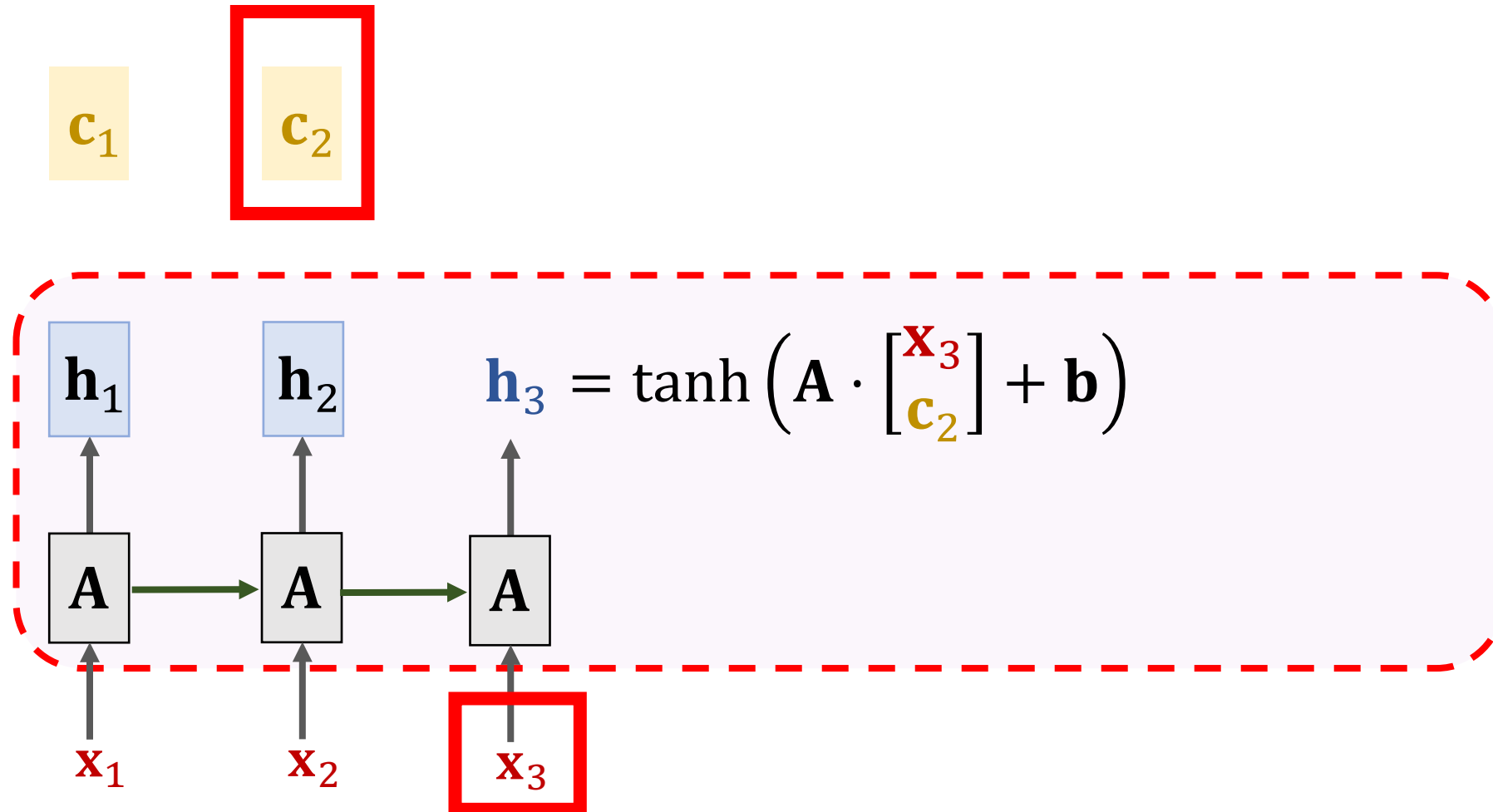
# SimpleRNN + Self-Attention



# SimpleRNN + Self-Attention



# SimpleRNN + Self-Attention



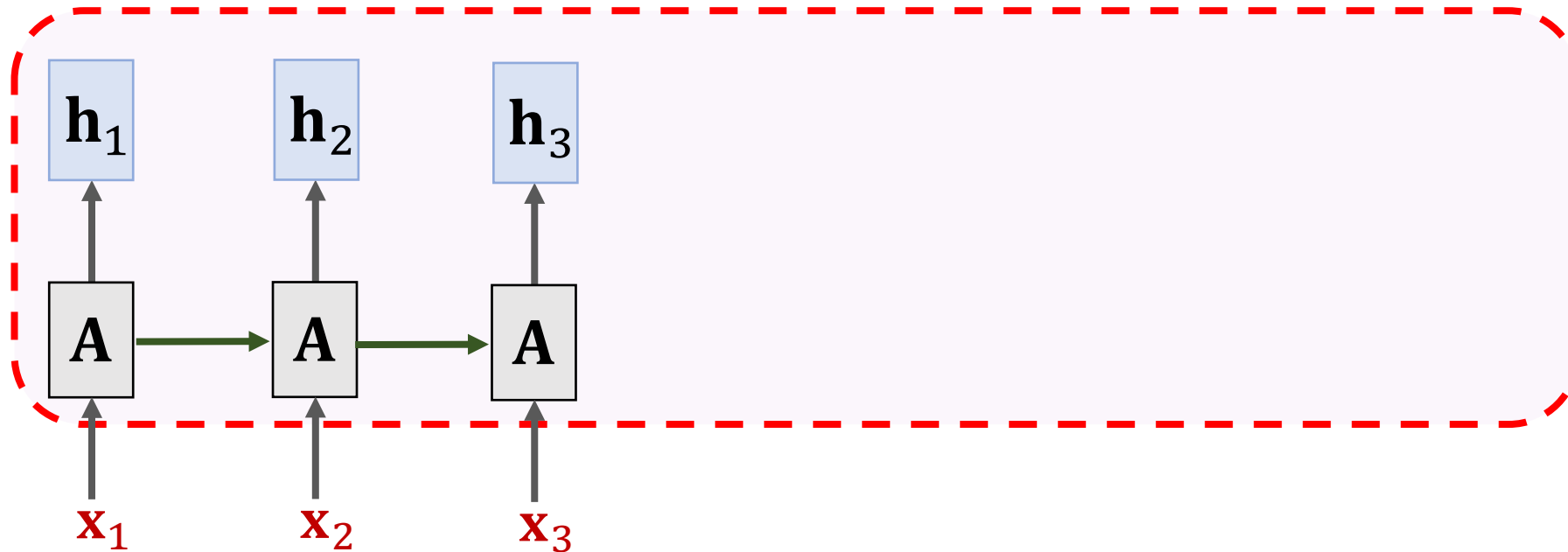


# SimpleRNN + Self-Attention

$c_1$

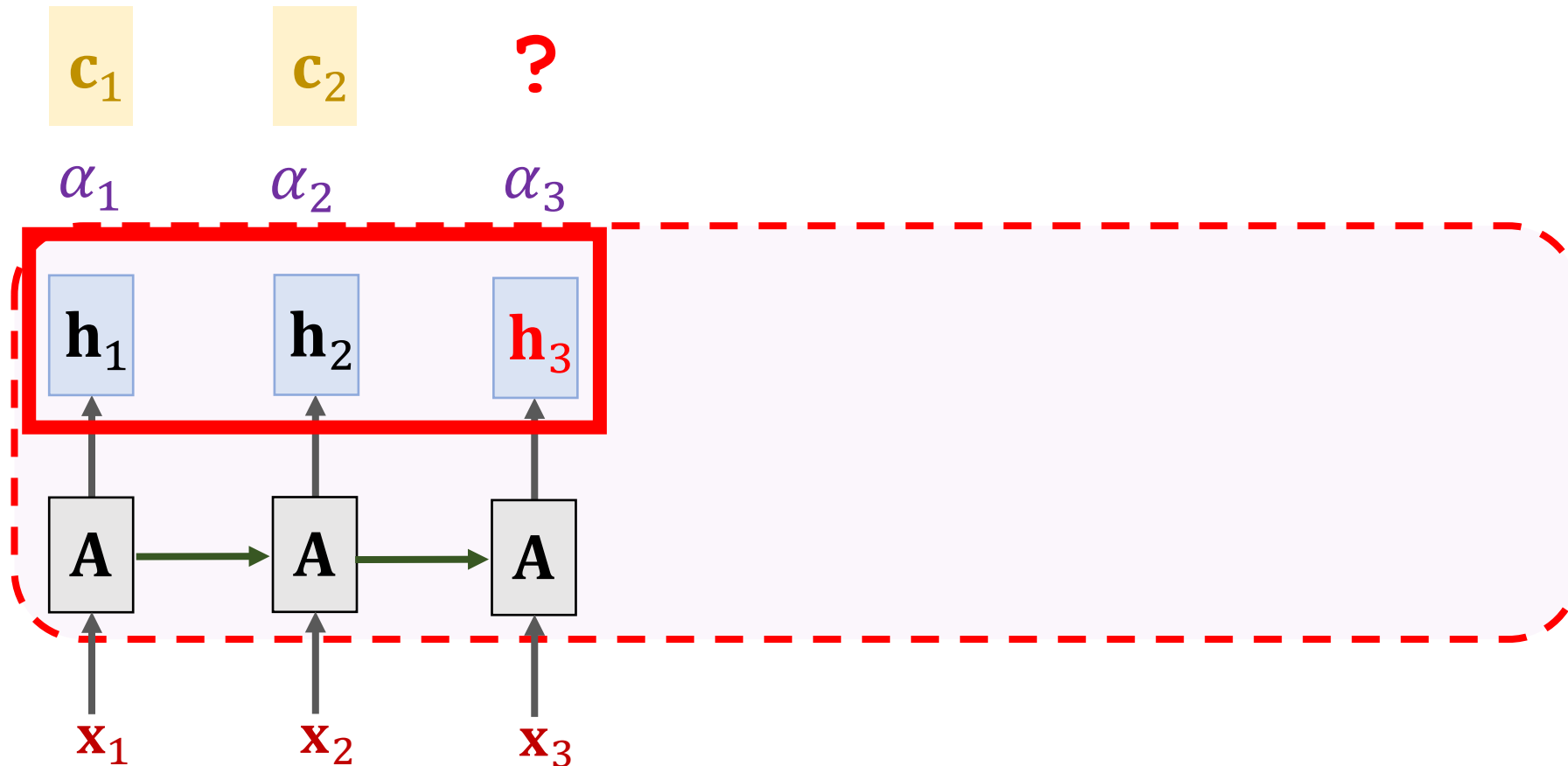
$c_2$

?

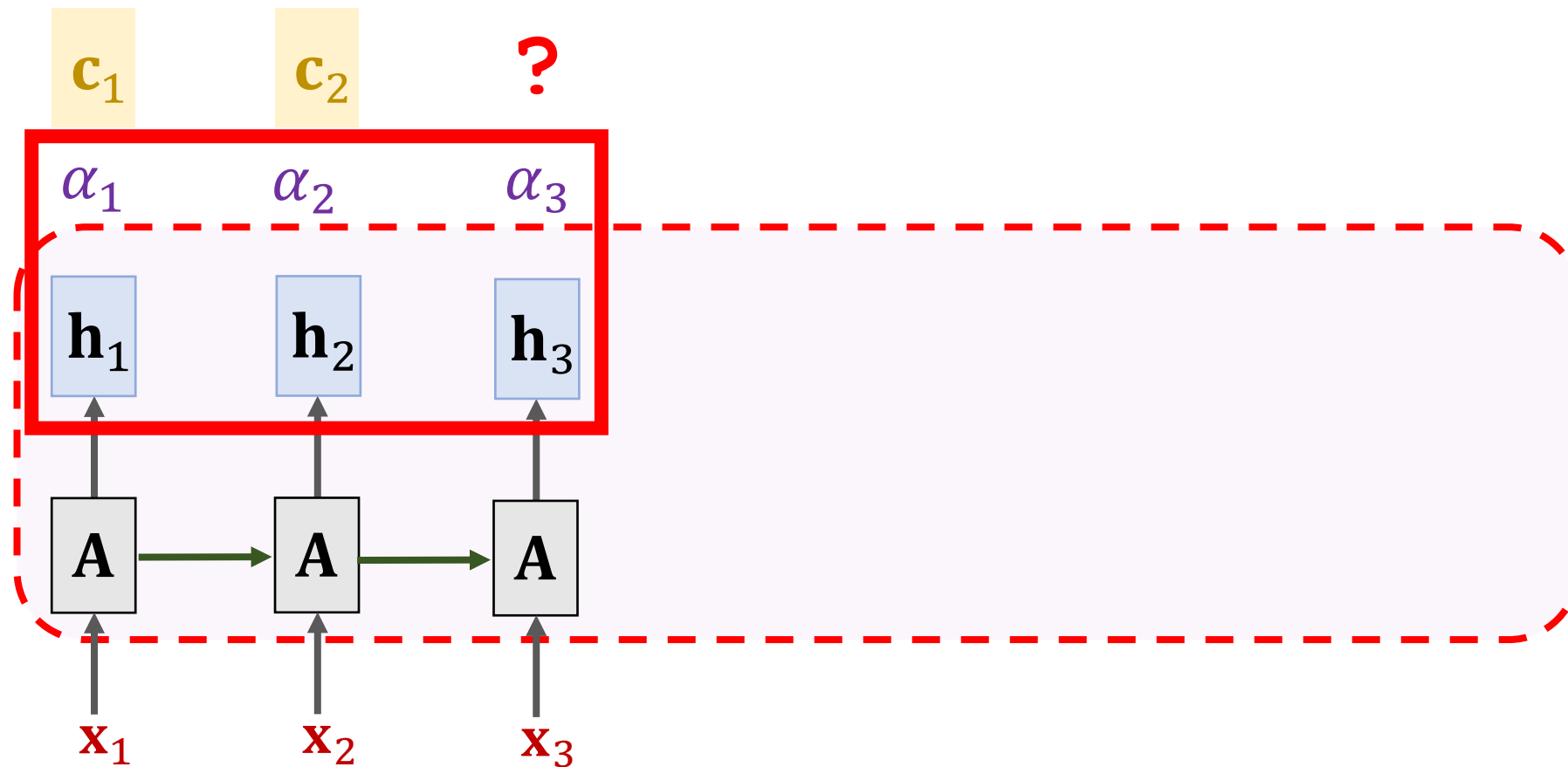


# SimpleRNN + Self-Attention

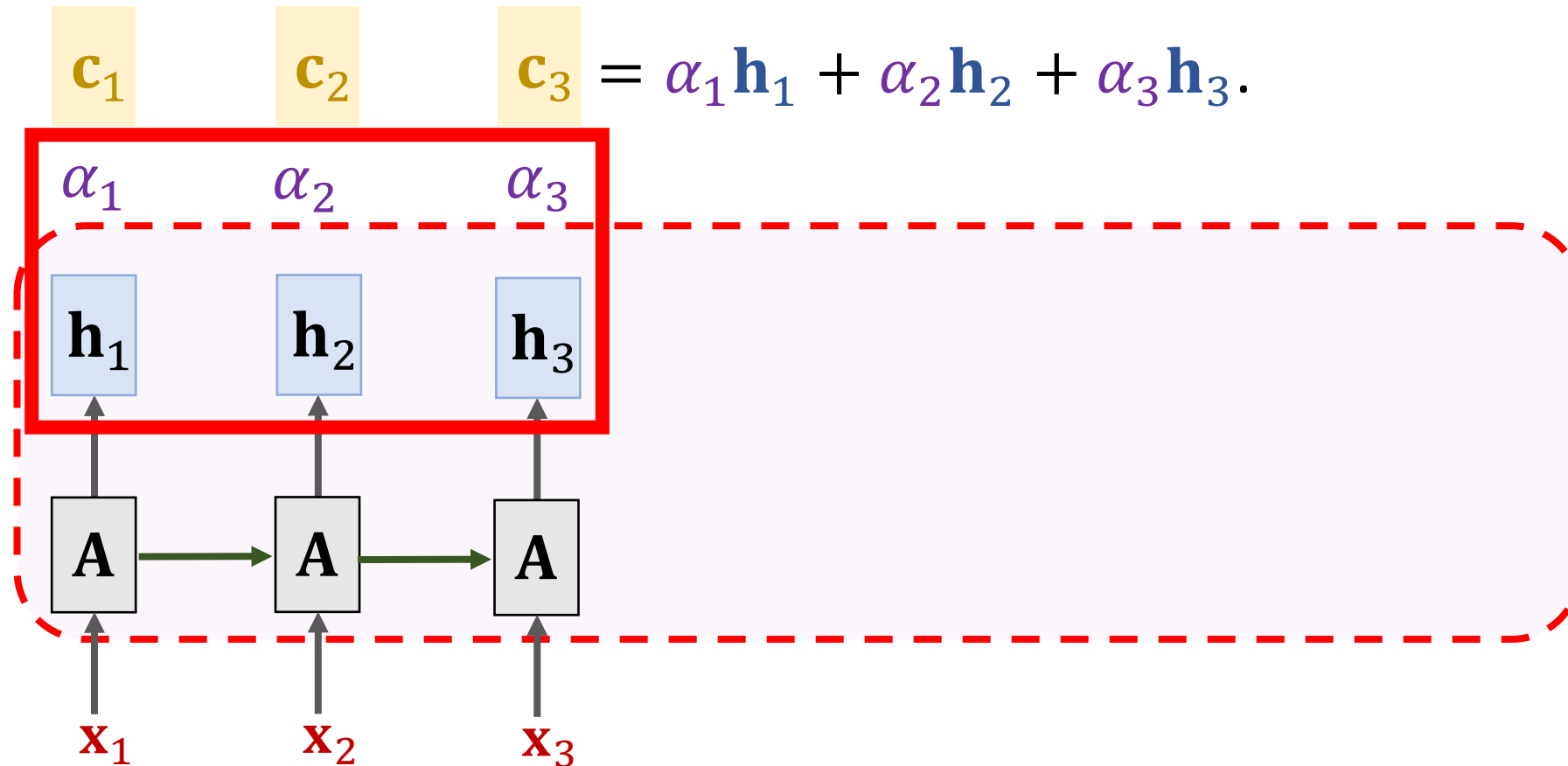
Weights:  $\alpha_i = \text{align}(\mathbf{h}_i, \mathbf{h}_3)$ .



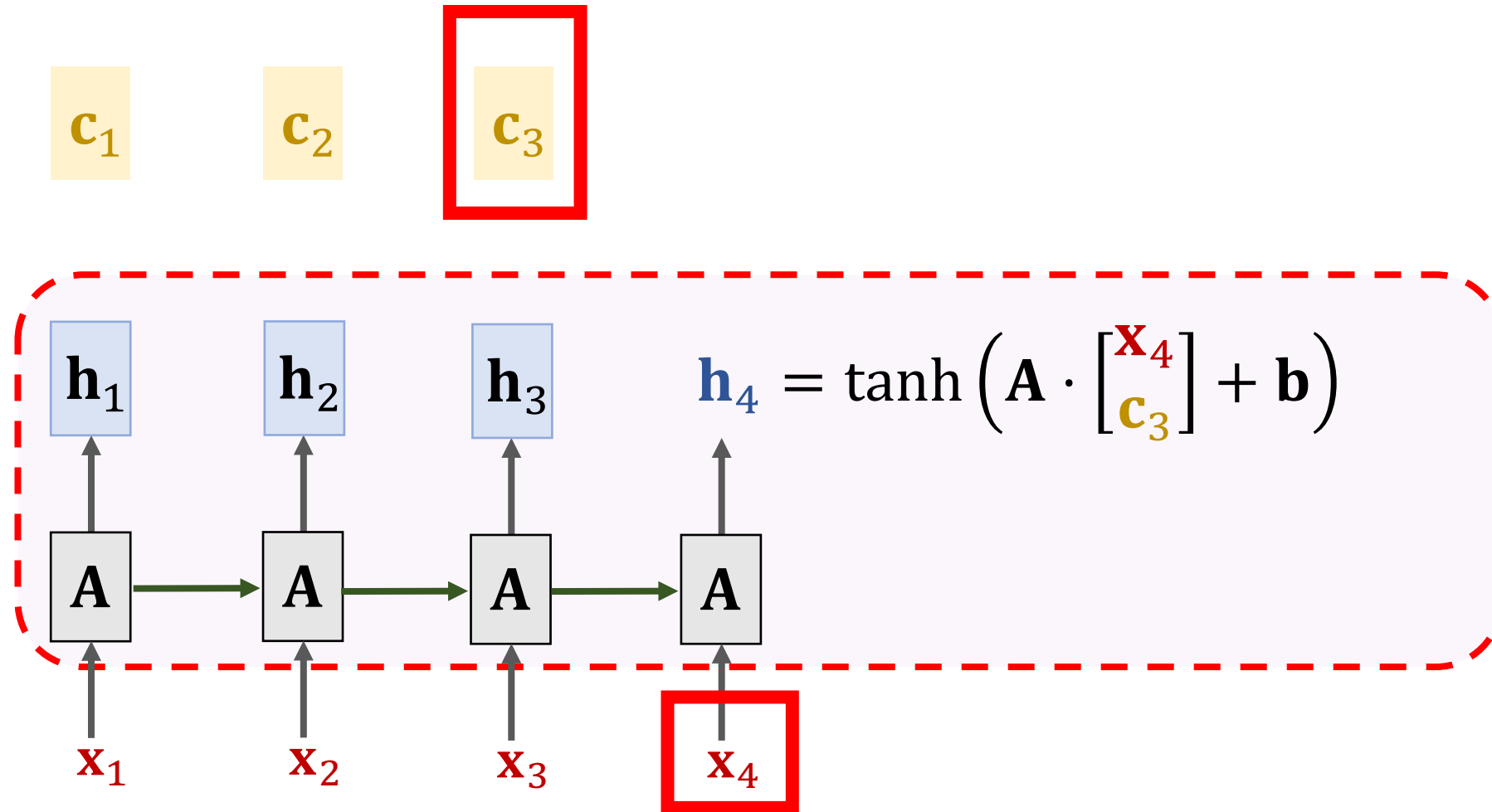
# SimpleRNN + Self-Attention



# SimpleRNN + Self-Attention



# SimpleRNN + Self-Attention



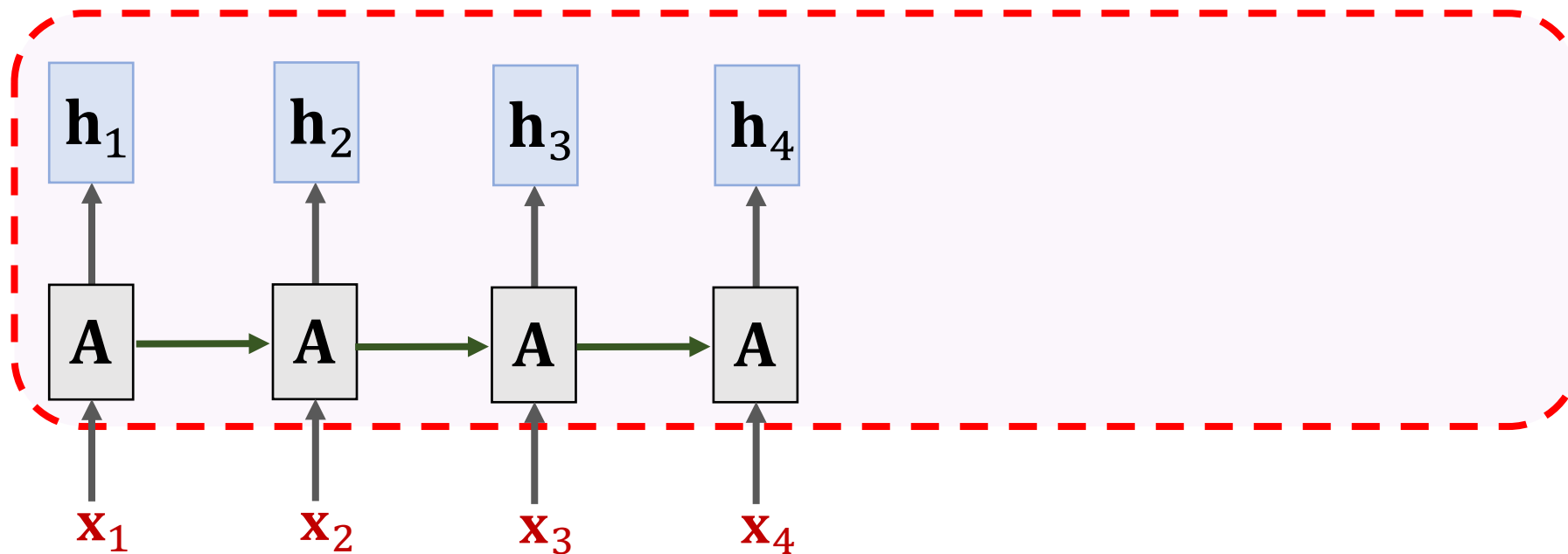
# SimpleRNN + Self-Attention

$\mathbf{c}_1$

$\mathbf{c}_2$

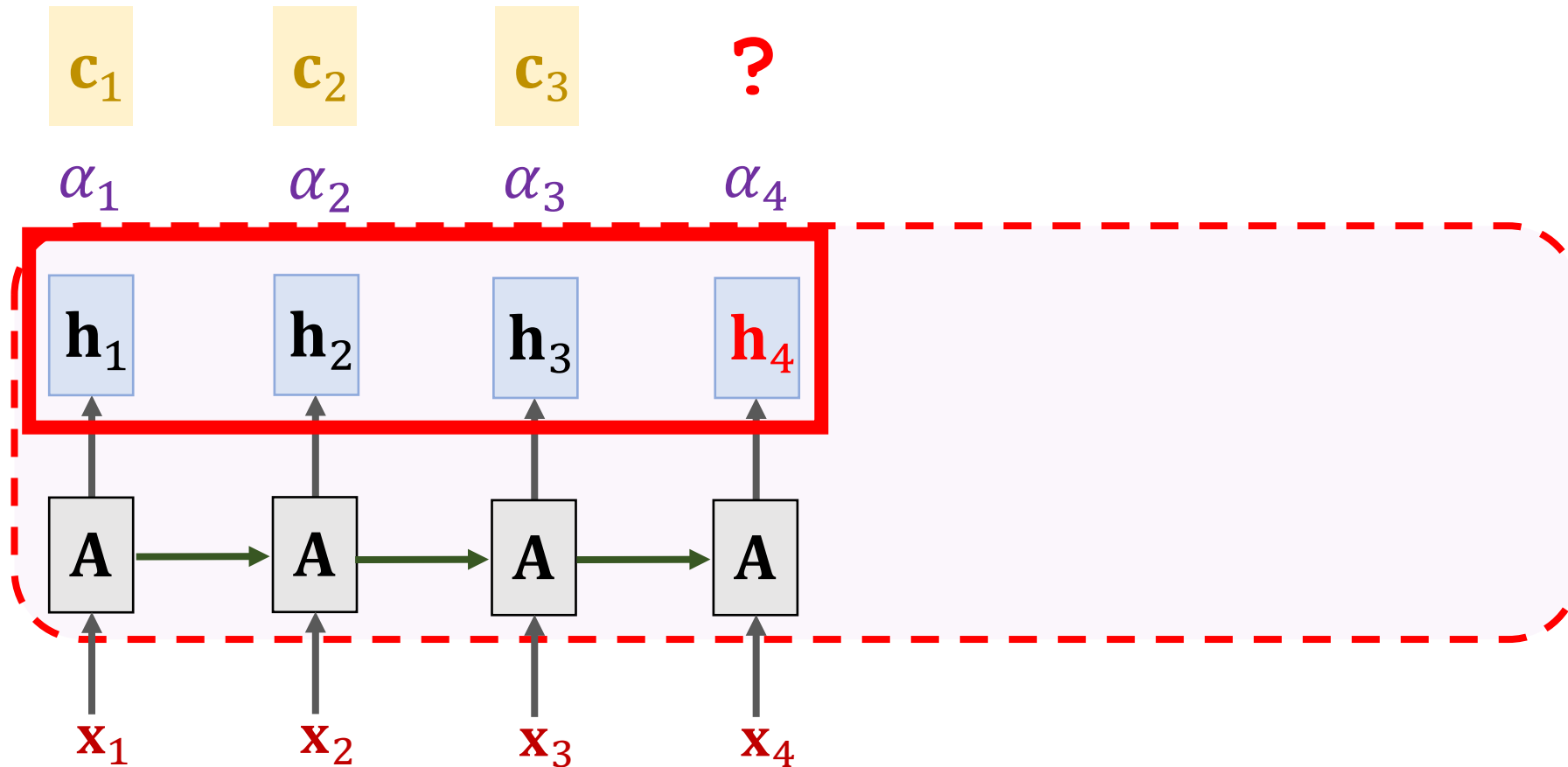
$\mathbf{c}_3$

?

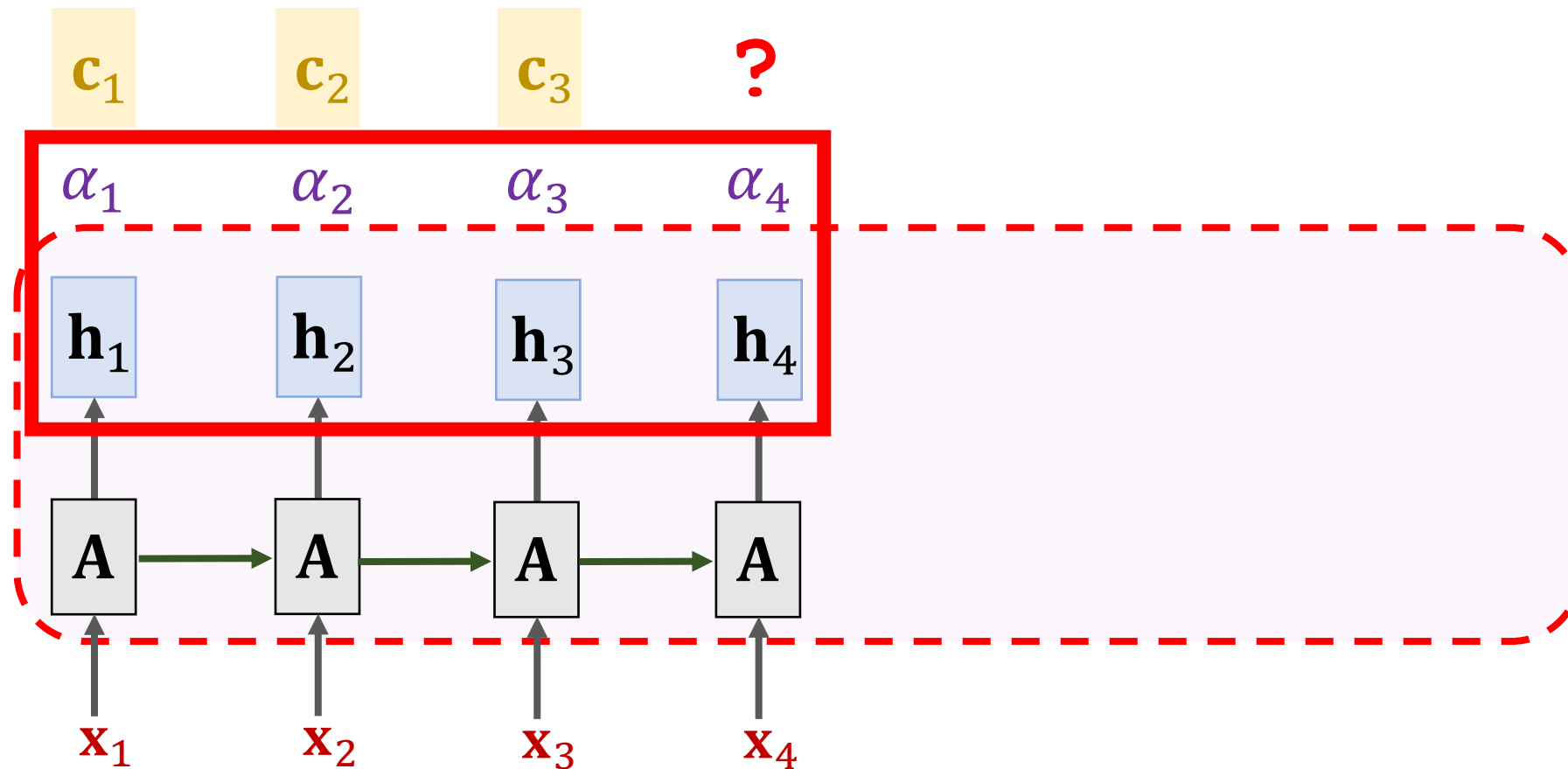


# SimpleRNN + Self-Attention

**Weights:**  $\alpha_i = \text{align}(\mathbf{h}_i, \mathbf{h}_4)$ . 状态之间 算相关性权重



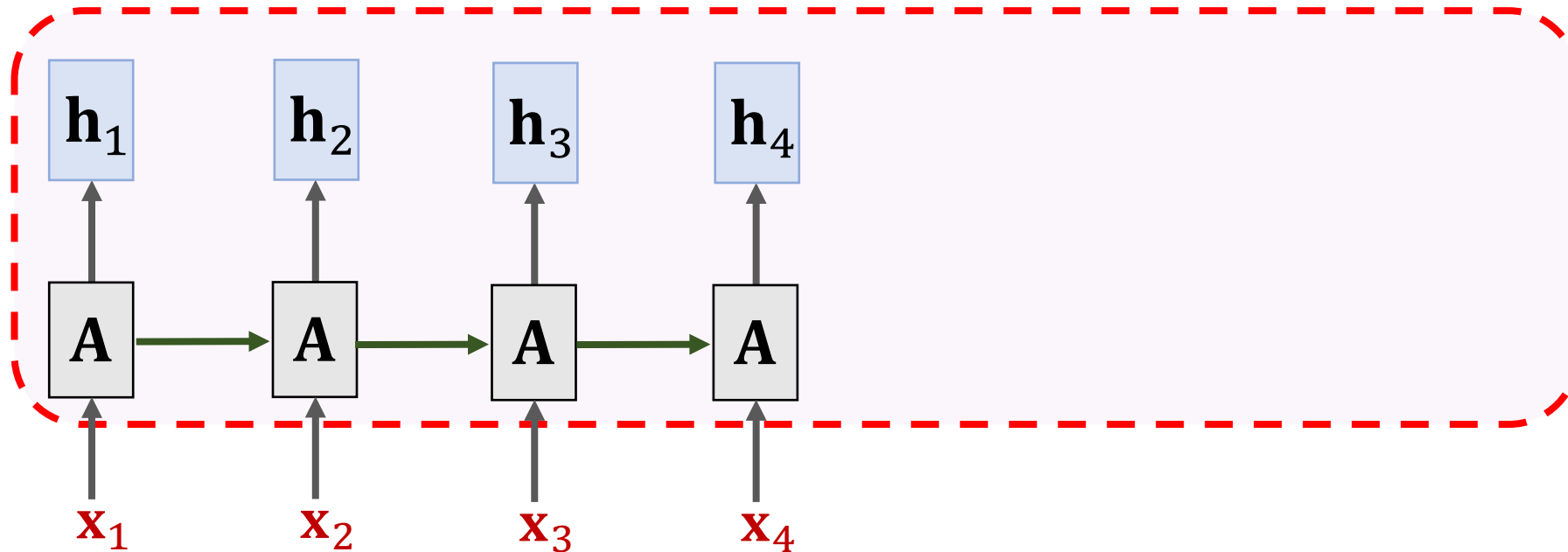
# SimpleRNN + Self-Attention



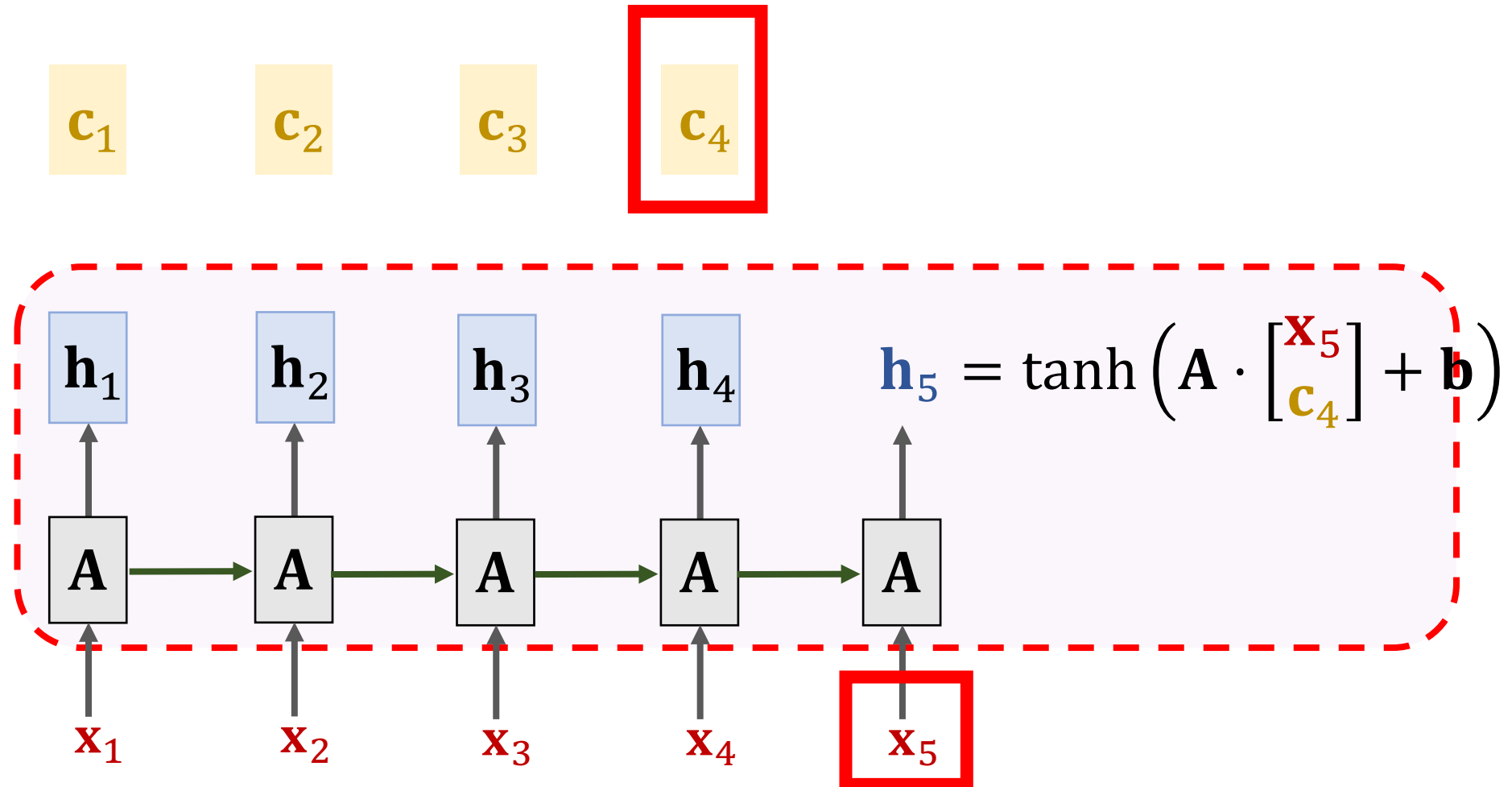


# SimpleRNN + Self-Attention

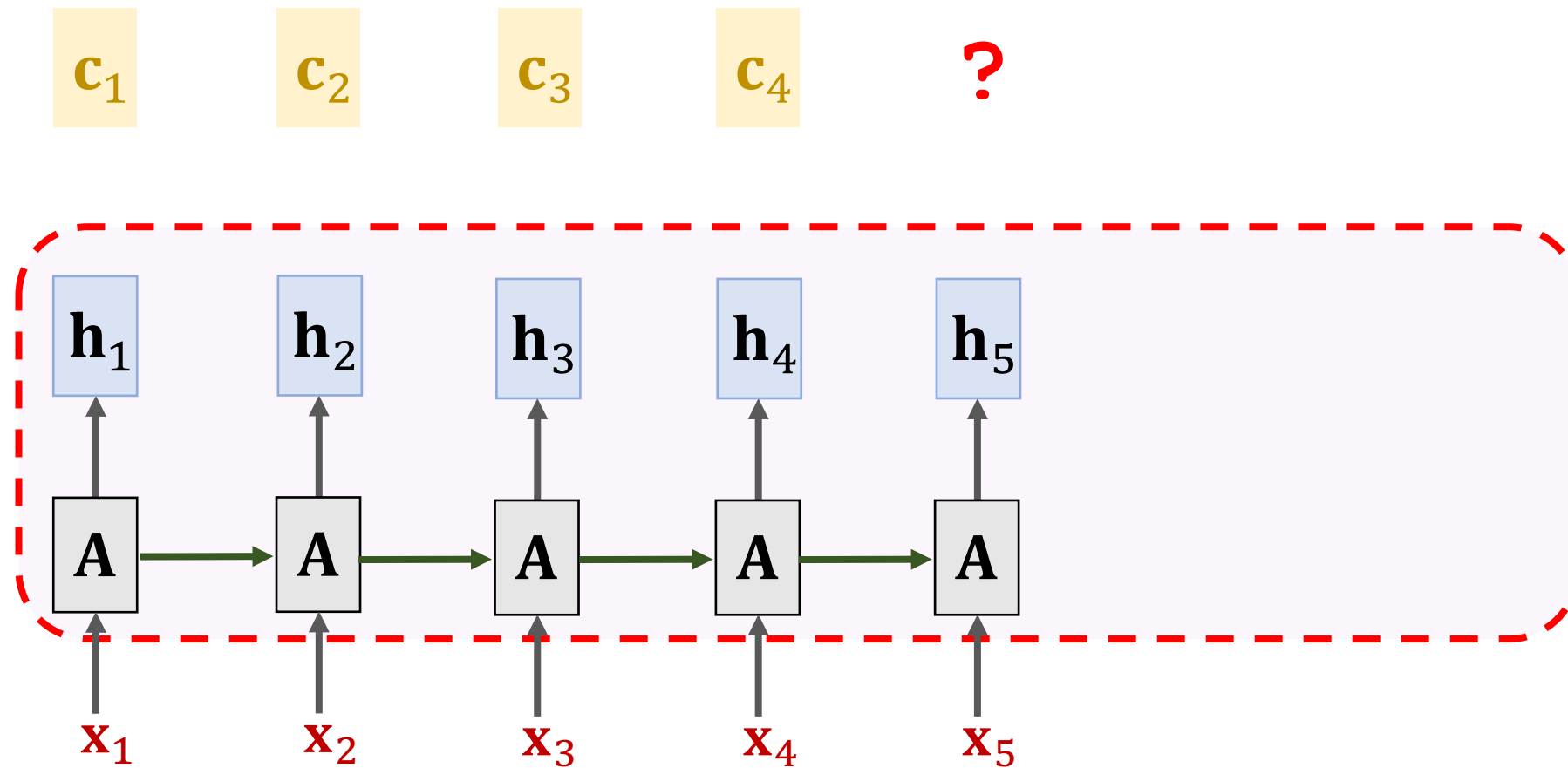
$\mathbf{c}_1$        $\mathbf{c}_2$        $\mathbf{c}_3$        $\mathbf{c}_4 = \alpha_1 \mathbf{h}_1 + \alpha_2 \mathbf{h}_2 + \alpha_3 \mathbf{h}_3 + \alpha_4 \mathbf{h}_4.$



# SimpleRNN + Self-Attention

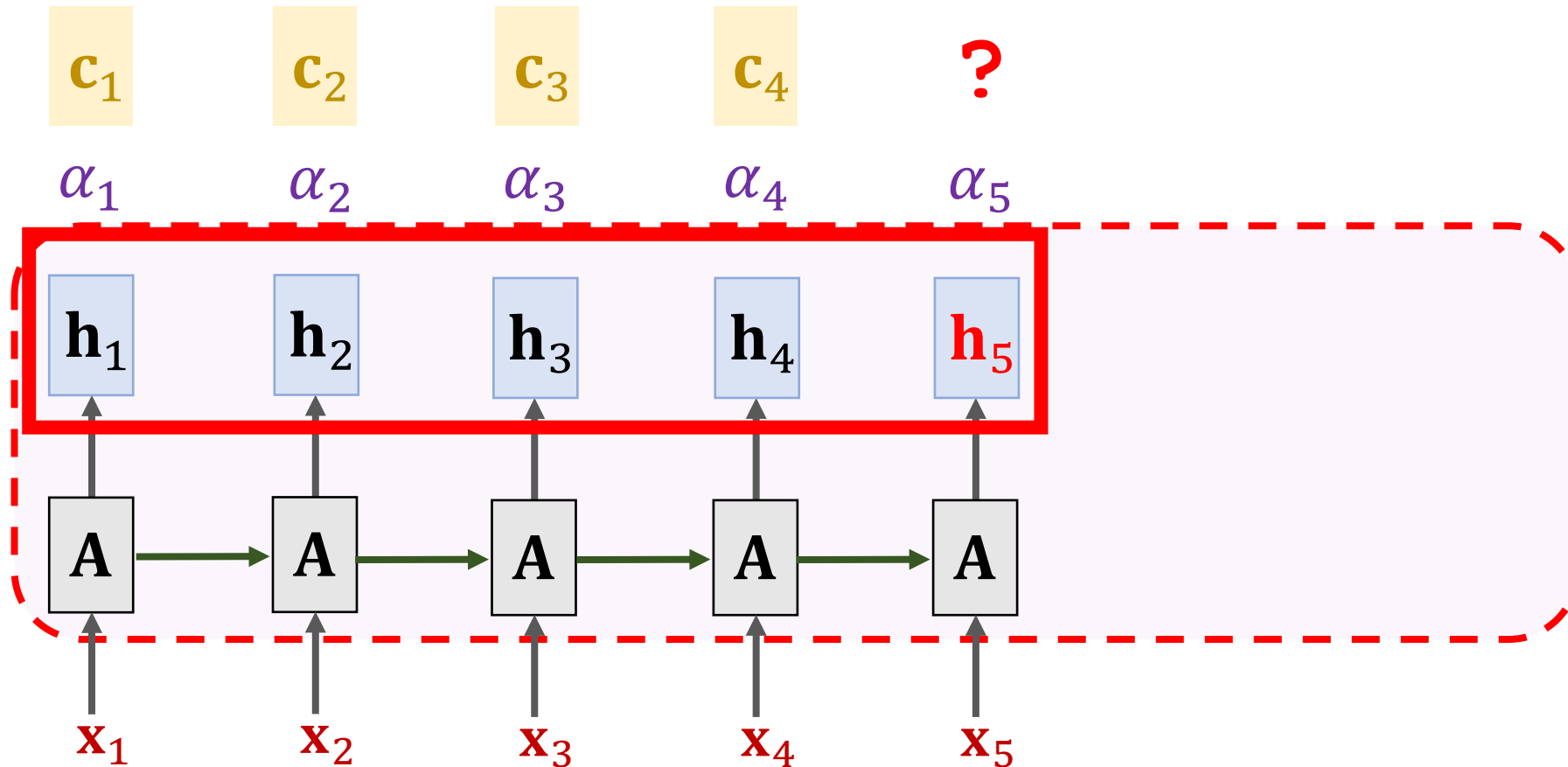


# SimpleRNN + Self-Attention

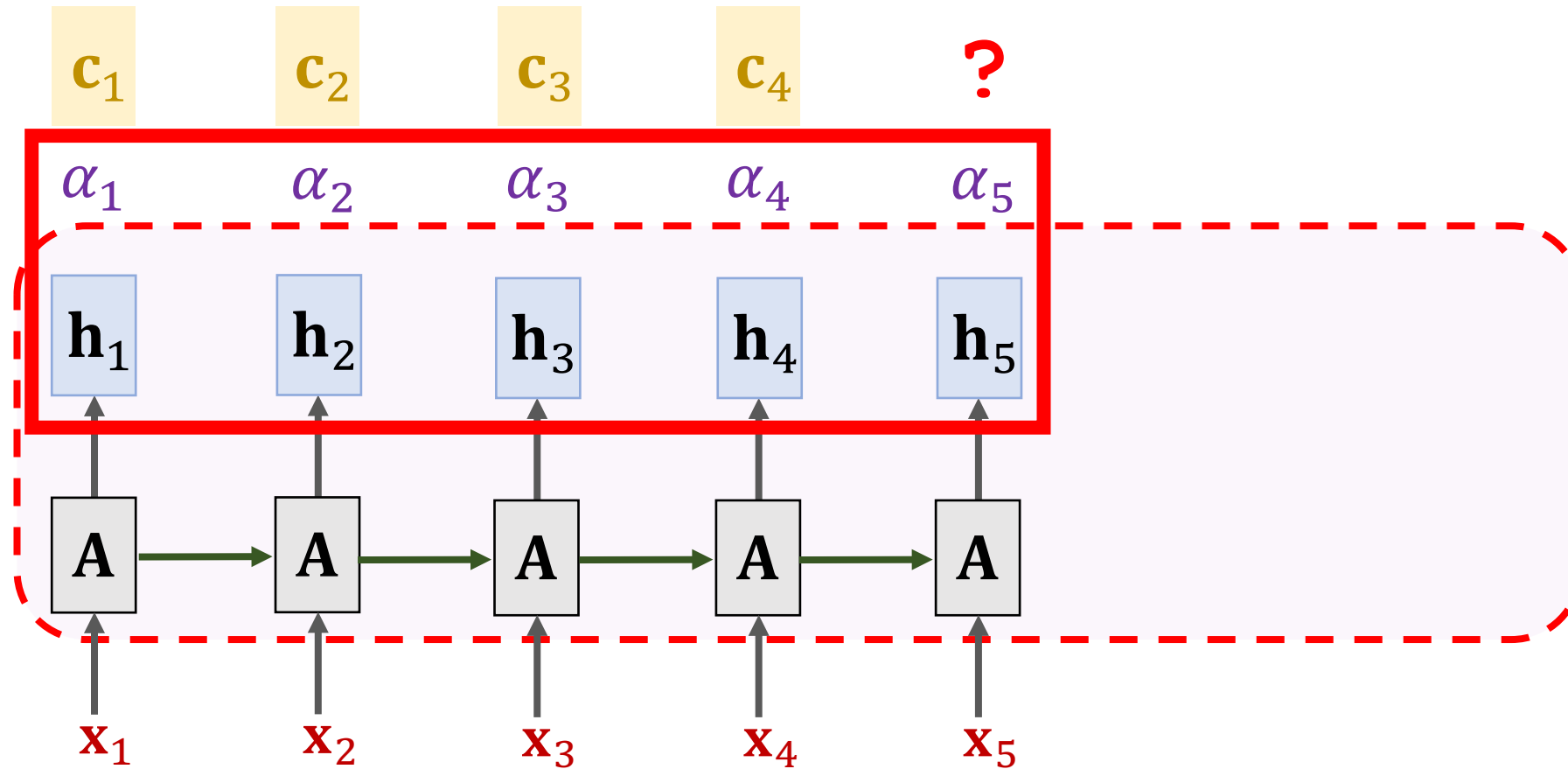


# SimpleRNN + Self-Attention

Weights:  $\alpha_i = \text{align}(\mathbf{h}_i, \mathbf{h}_5)$ .

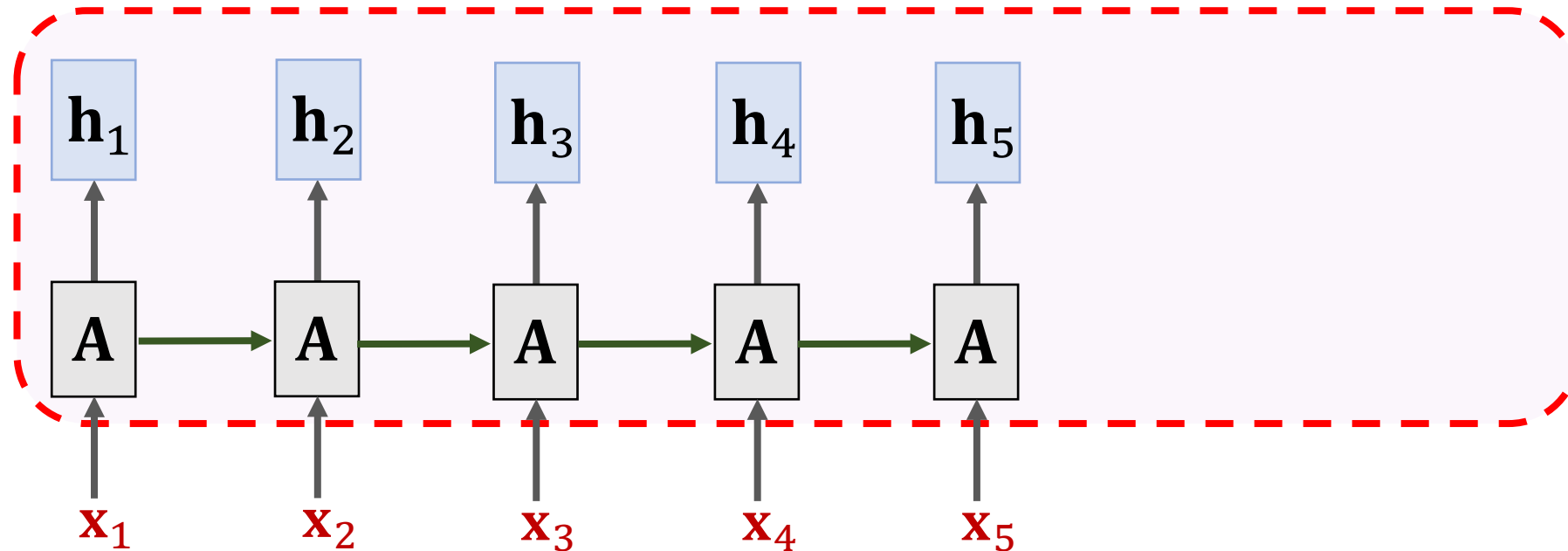


# SimpleRNN + Self-Attention

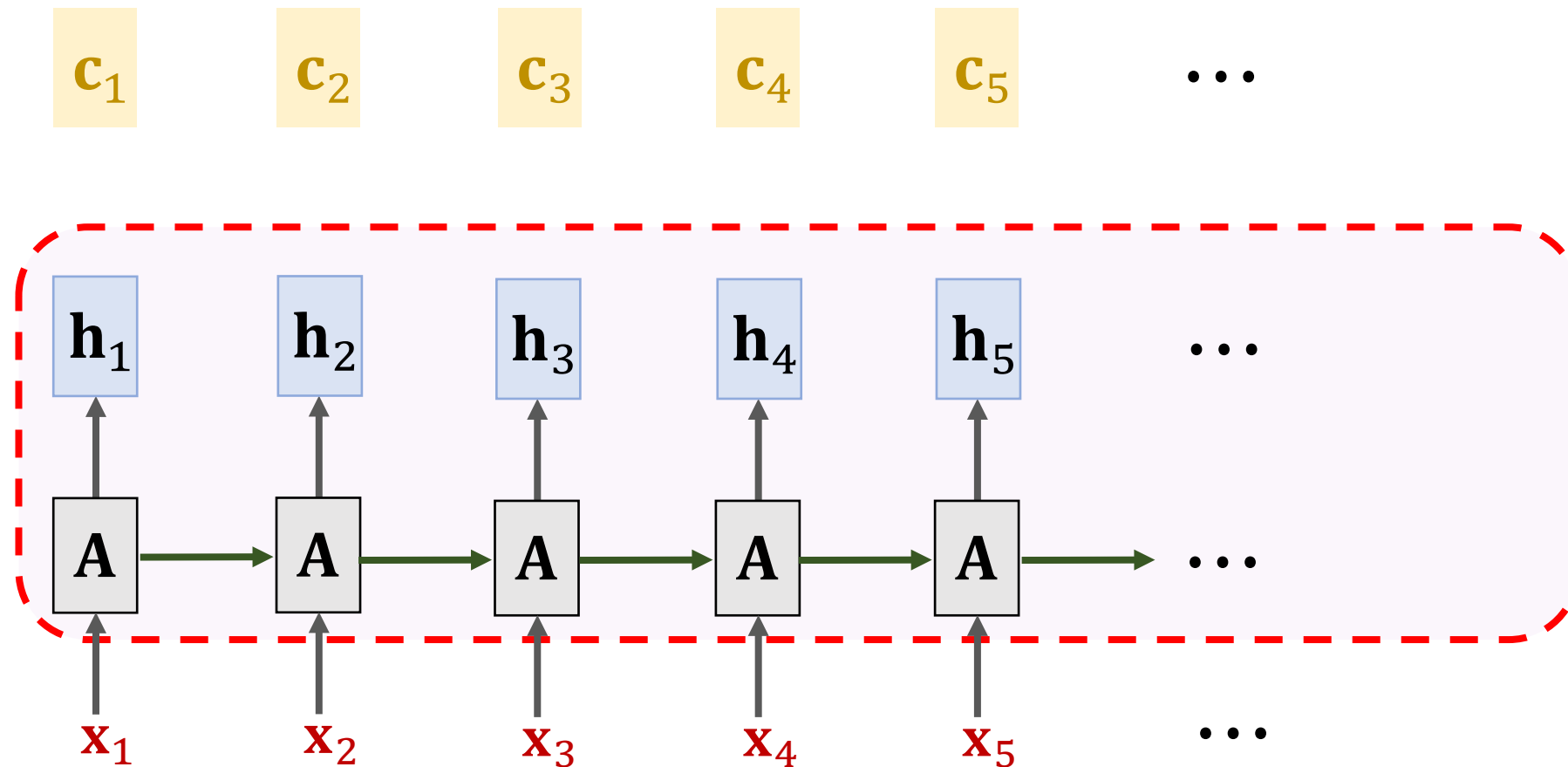


# SimpleRNN + Self-Attention

$\mathbf{c}_1$     $\mathbf{c}_2$     $\mathbf{c}_3$     $\mathbf{c}_4$     $\mathbf{c}_5 = \alpha_1 \mathbf{h}_1 + \alpha_2 \mathbf{h}_2 + \cdots + \alpha_5 \mathbf{h}_5.$



# SimpleRNN + Self-Attention



# Summary

- With self-attention, RNN is less likely to forget.



# Summary

- With self-attention, RNN is less likely to forget. 有了自我关注，RNN就不太可能忘记了
- Pay attention to the context relevant to the new input. 注意与新输入相关的上下文

The diagram shows the sentence "The FBI is chasing a criminal on the run." with words in red and blue. Blue highlights are placed under the words "The", "FBI", "is", "chasing", "a", "criminal", "on", "the", and "run". Red highlights are placed under the words "FBI", "is", "chasing", "a", "criminal", "on", "the", and "run". This visualizes the self-attention mechanism where the model focuses on the relevant context for each new input.

The  
The FBI  
The FBI is  
The FBI is chasing  
The FBI is chasing a  
The FBI is chasing a criminal  
The FBI is chasing a criminal on  
The FBI is chasing a criminal on the  
The FBI is chasing a criminal on the run  
The FBI is chasing a criminal on the run .

Figure is from the paper “ Long Short-Term Memory-Networks for Machine Reading.”

**Thank you!**