# Self-Attention

**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:**

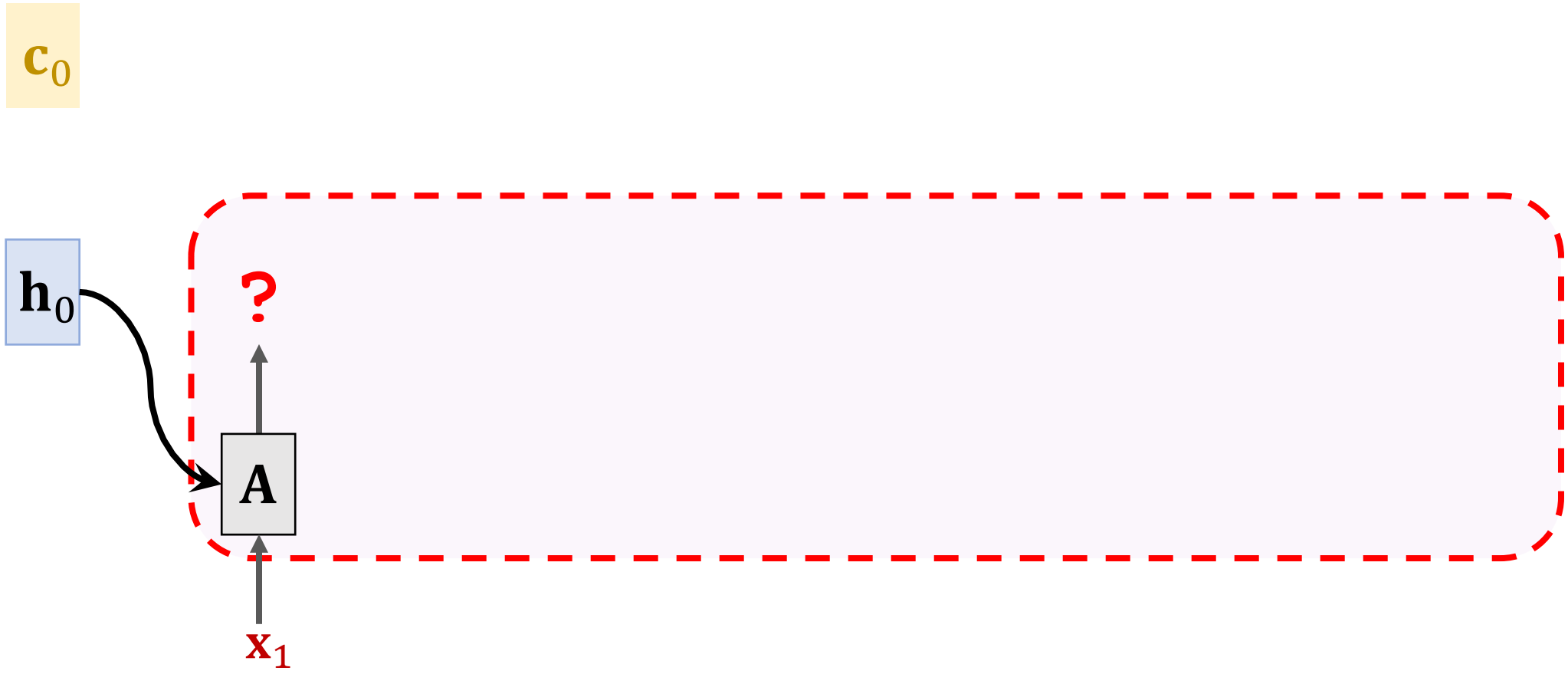

1. Bahdanau, Cho, & Bengio. Neural machine translation by jointly learning to align and translate. In *ICLR*, 2015.
2. Cheng, Dong, & Lapata. Long Short-Term Memory-Networks for Machine Reading. In *EMNLP*, 2016.

# SimpleRNN + Self-Attention

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

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

# SimpleRNN + Self-Attention



# SimpleRNN + Self-Attention

**SimpleRNN:**

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

$\mathbf{c}_0$

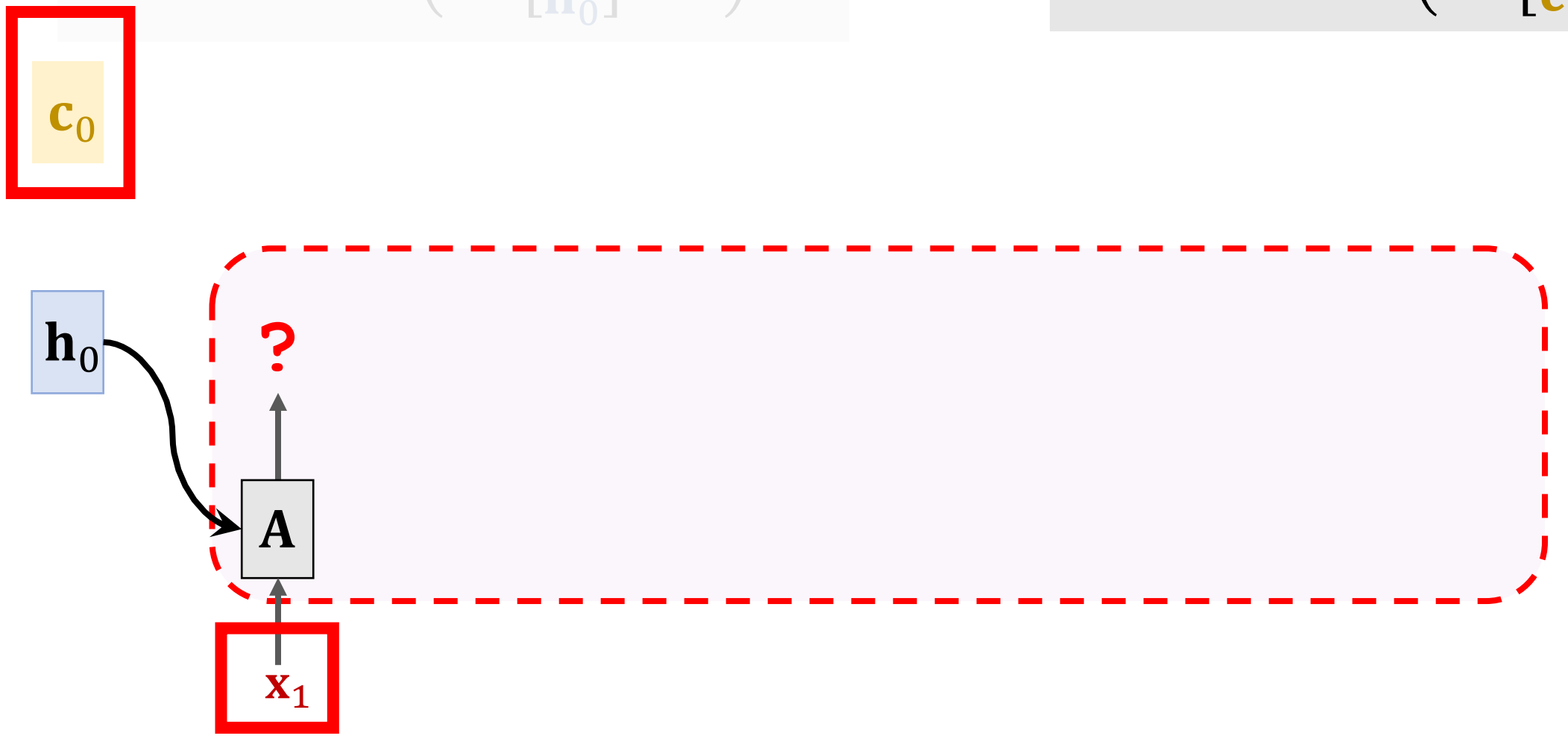
# SimpleRNN + Self-Attention

SimpleRNN:

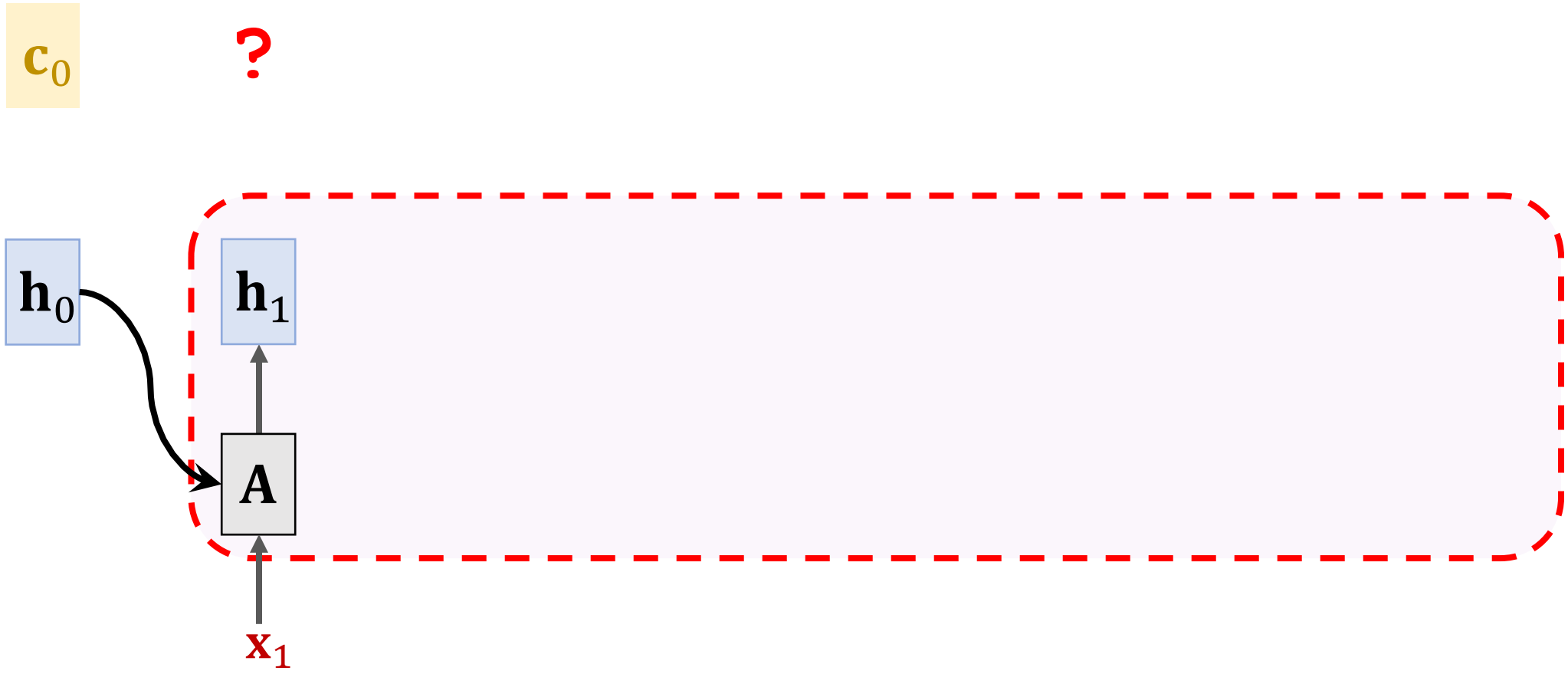
$$\mathbf{h}_1 = \tanh\left(\mathbf{A} \cdot \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{h}_0 \end{bmatrix} + \mathbf{b}\right)$$

**SimpleRNN + Self-Attention:**

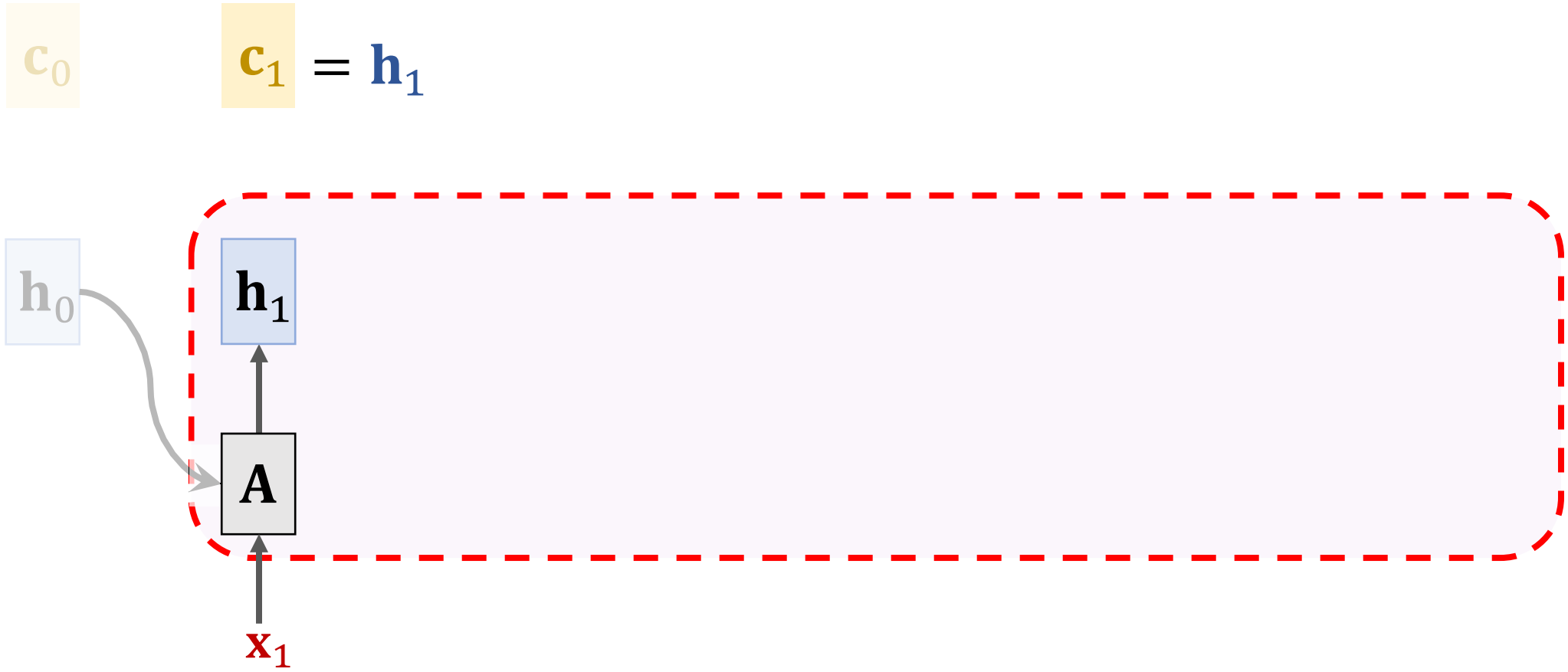
$$\mathbf{h}_1 = \tanh\left(\mathbf{A} \cdot \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{c}_0 \end{bmatrix} + \mathbf{b}\right)$$



# SimpleRNN + Self-Attention

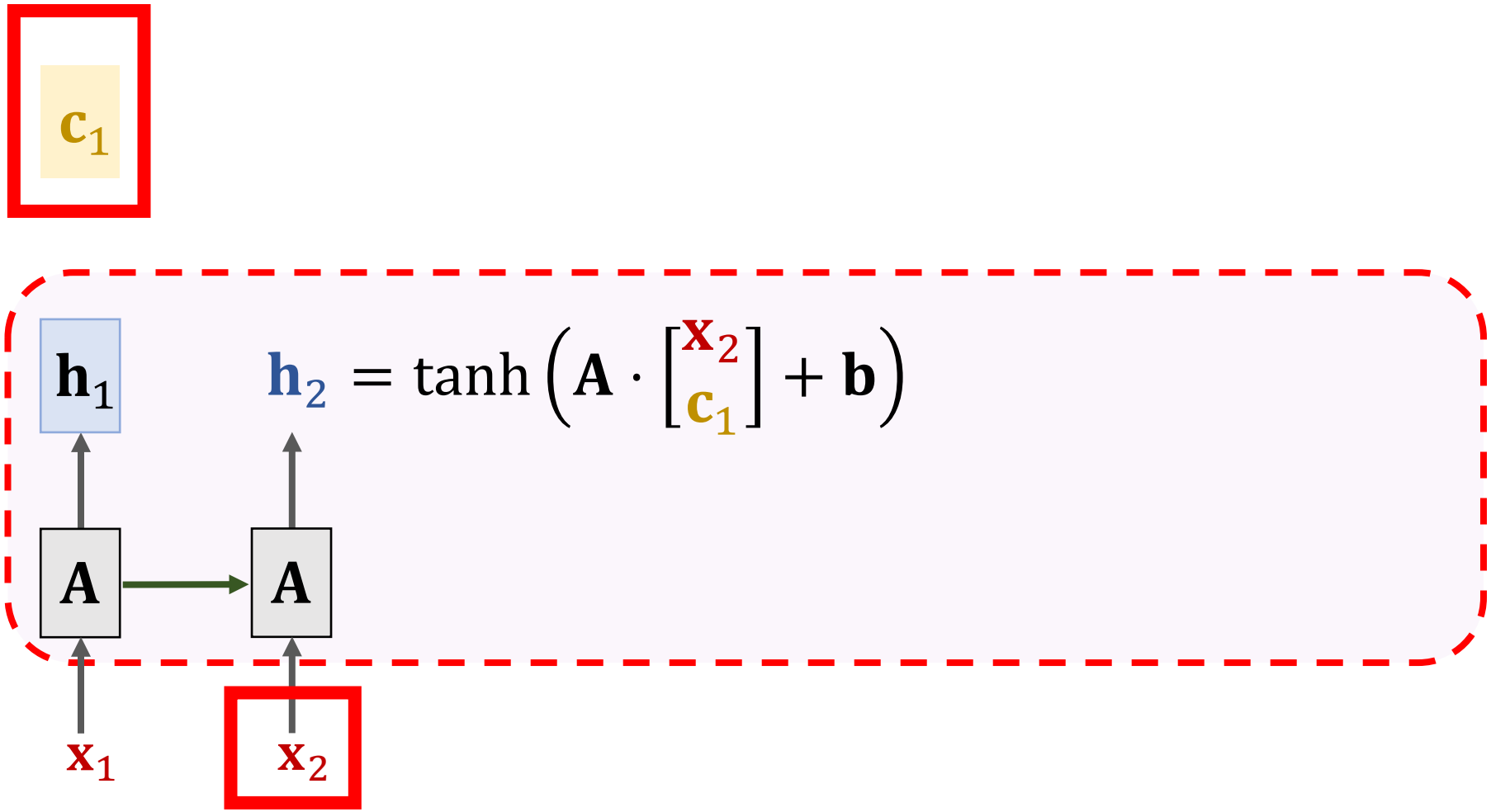


# SimpleRNN + Self-Attention



# SimpleRNN + Self-Attention

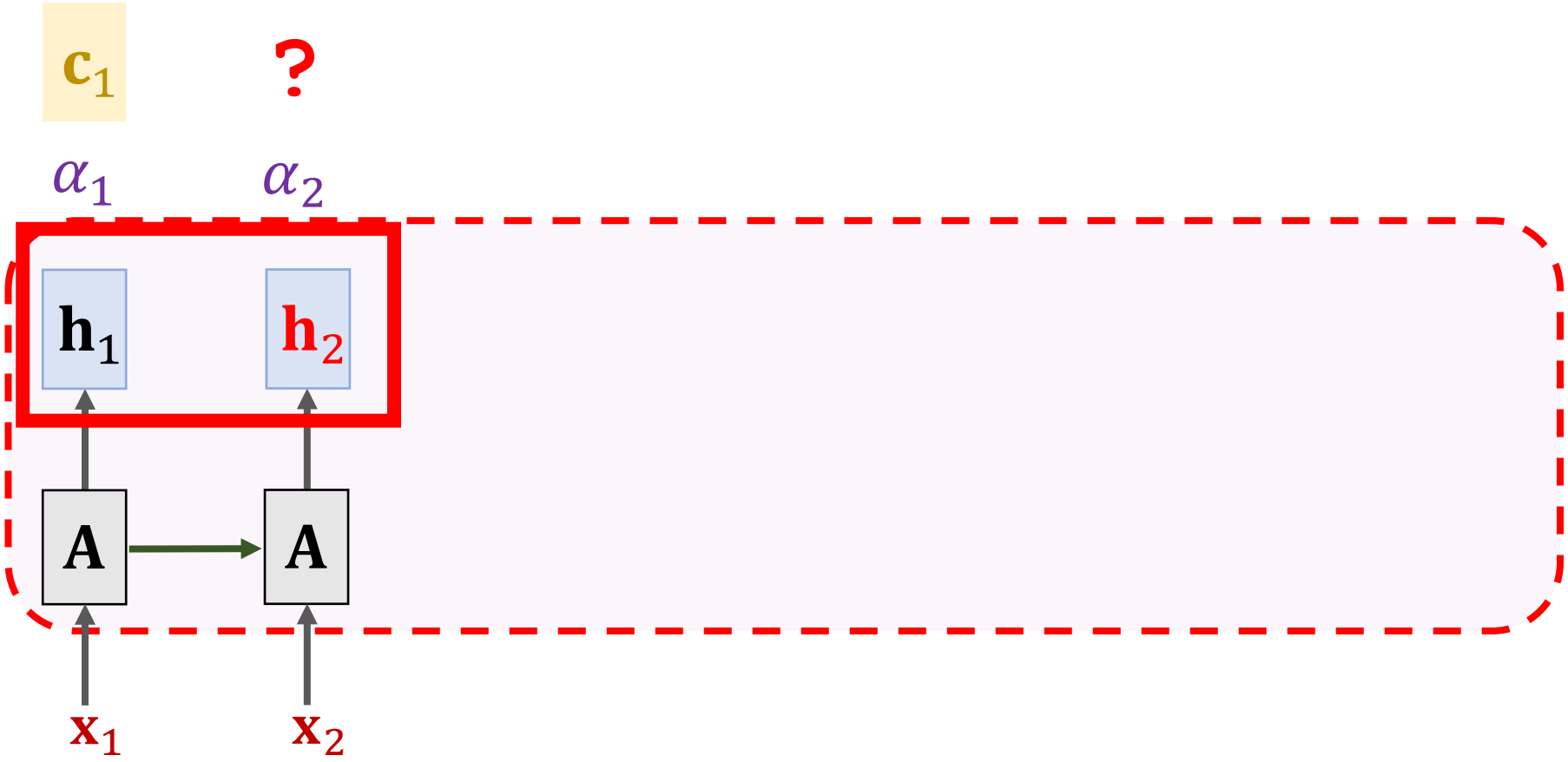
# SimpleRNN + Self-Attention



# SimpleRNN + Self-Attention

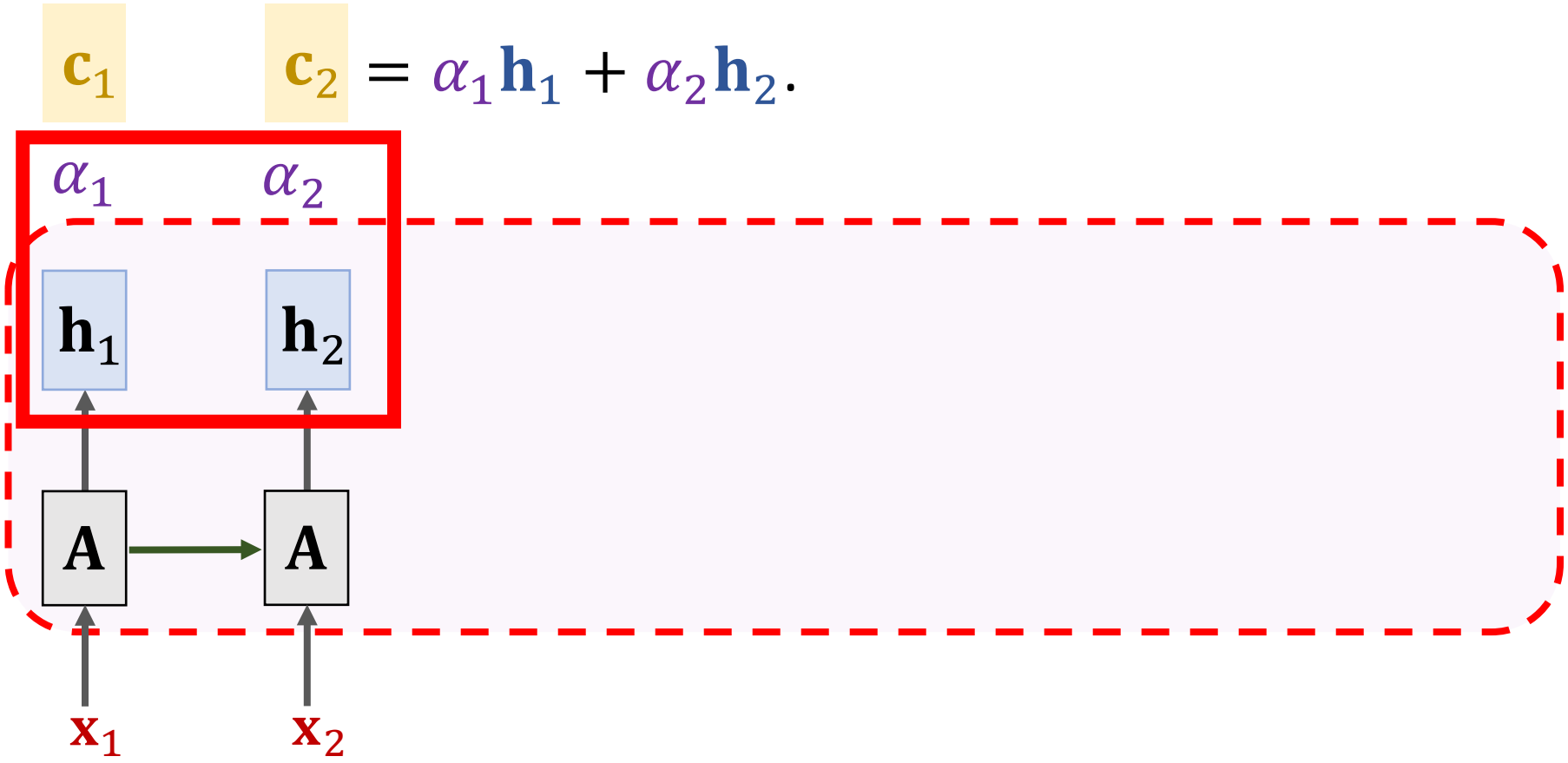
# SimpleRNN + Self-Attention

**Weights**: $\alpha_i = \text{align}(\mathbf{h}_i, \mathbf{h}_2)$.

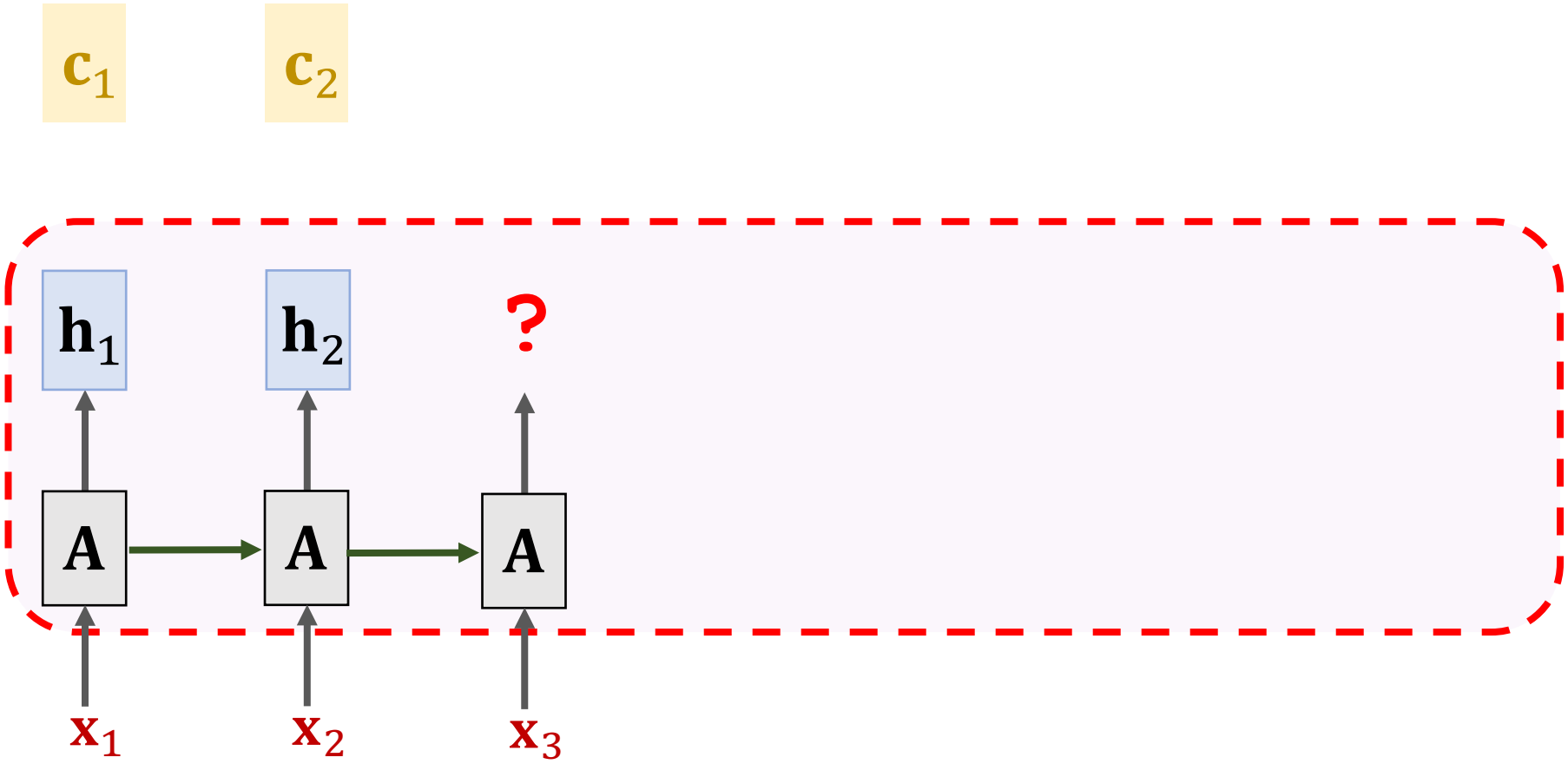


# SimpleRNN + Self-Attention

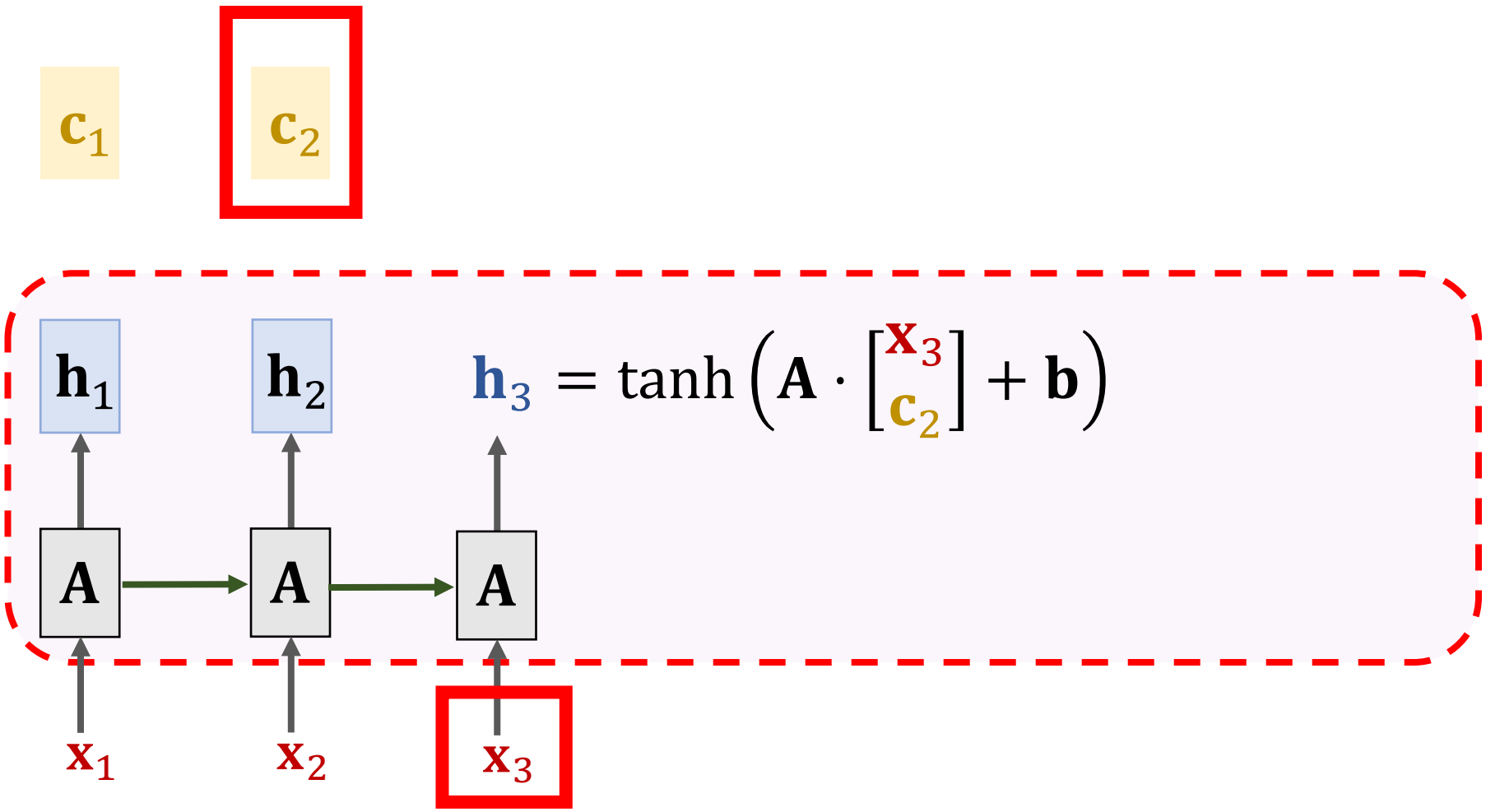
# SimpleRNN + Self-Attention



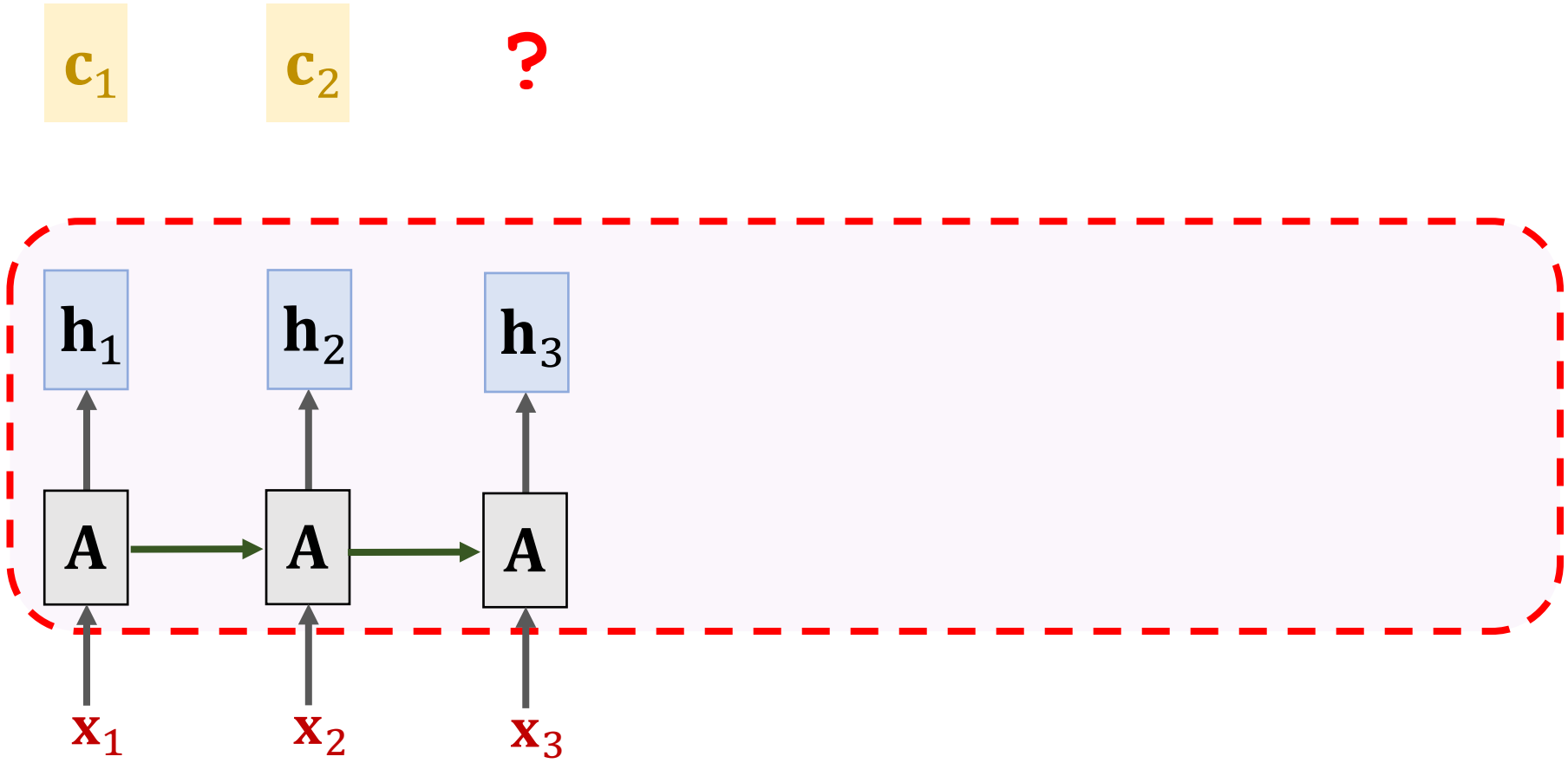
# SimpleRNN + Self-Attention



# SimpleRNN + Self-Attention

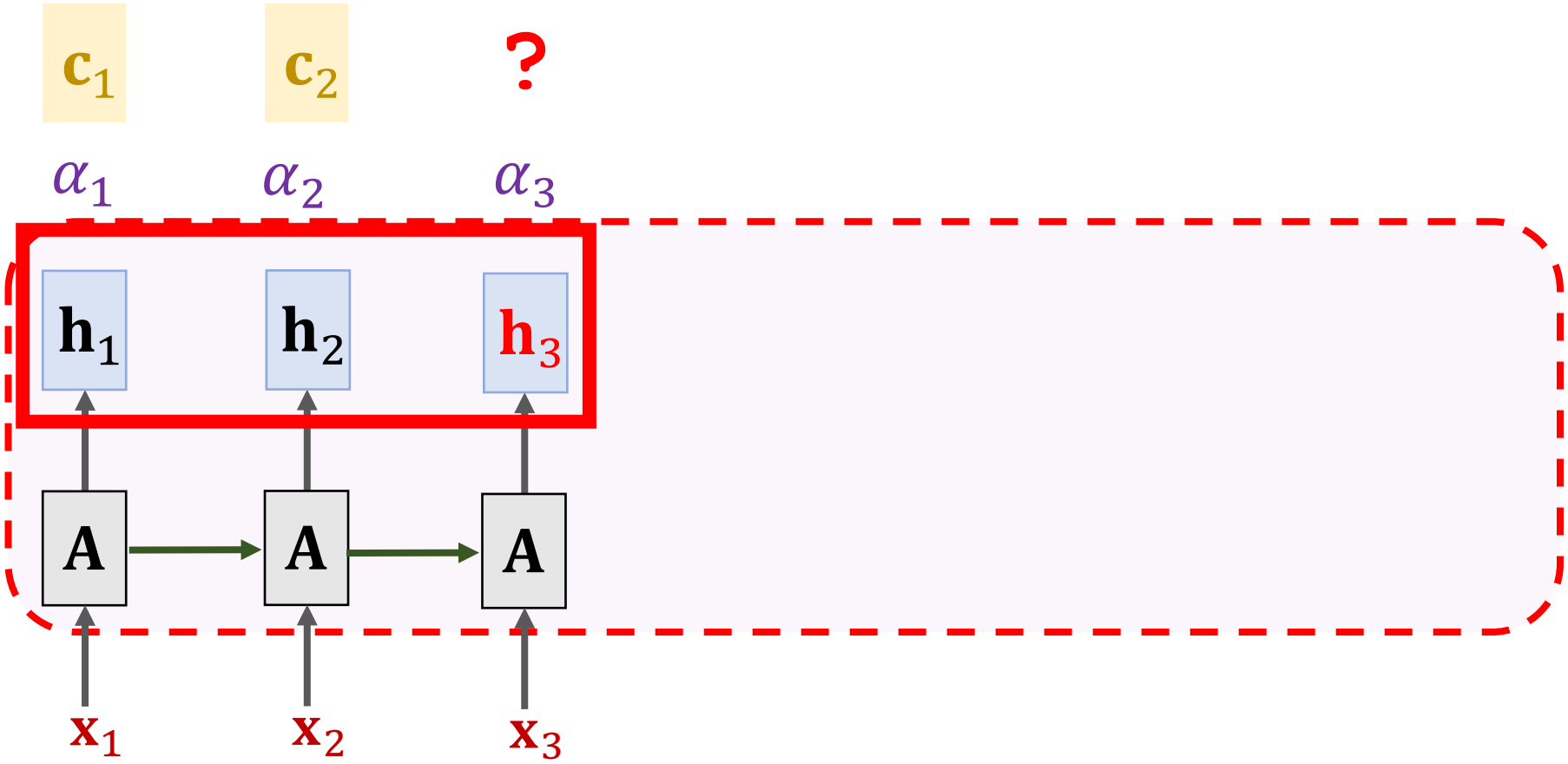


# SimpleRNN + Self-Attention

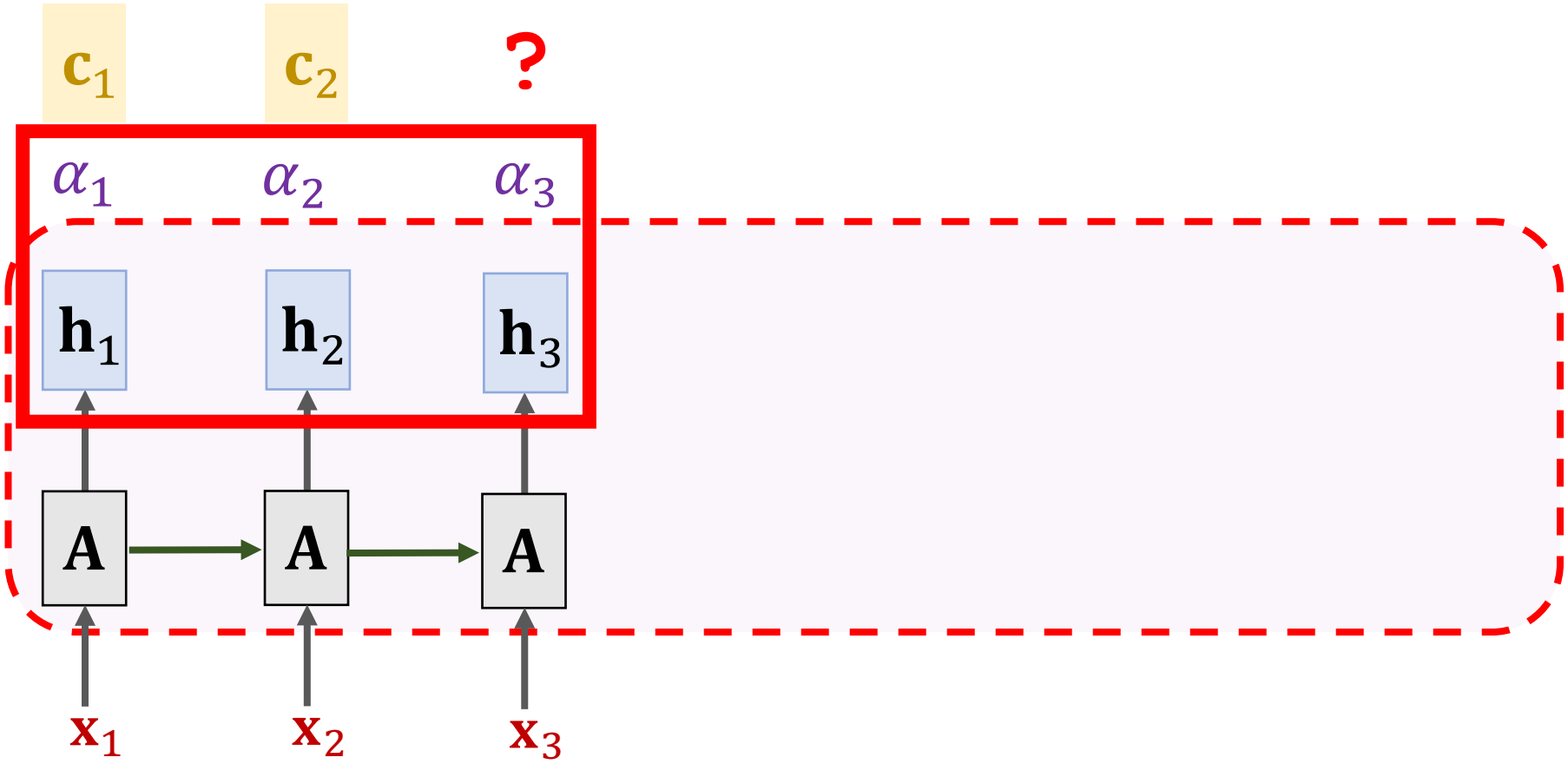


# SimpleRNN + Self-Attention

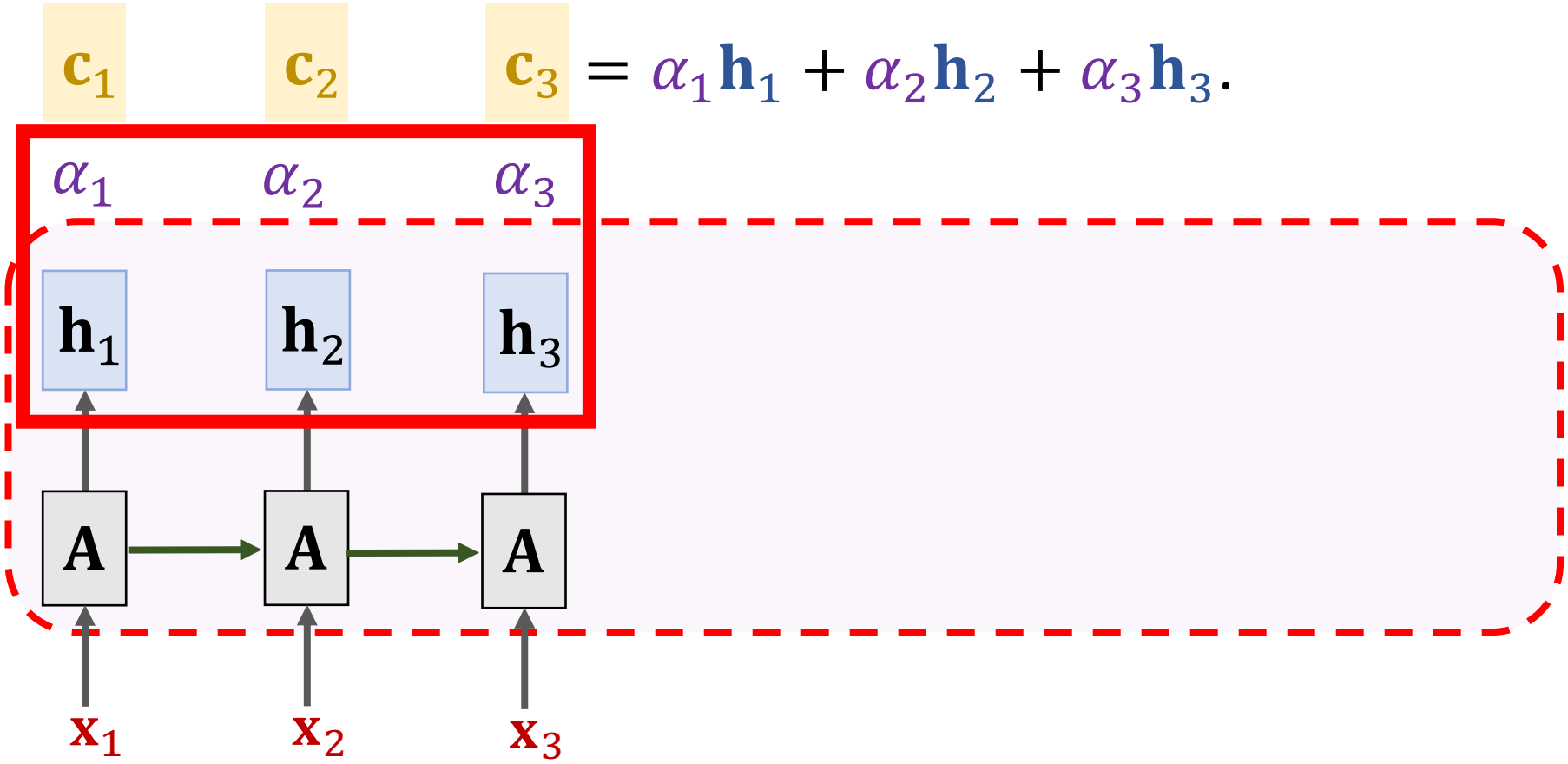
**Weights**: $\alpha_i = \text{align}(\mathbf{h}_i, \mathbf{h}_3)$.



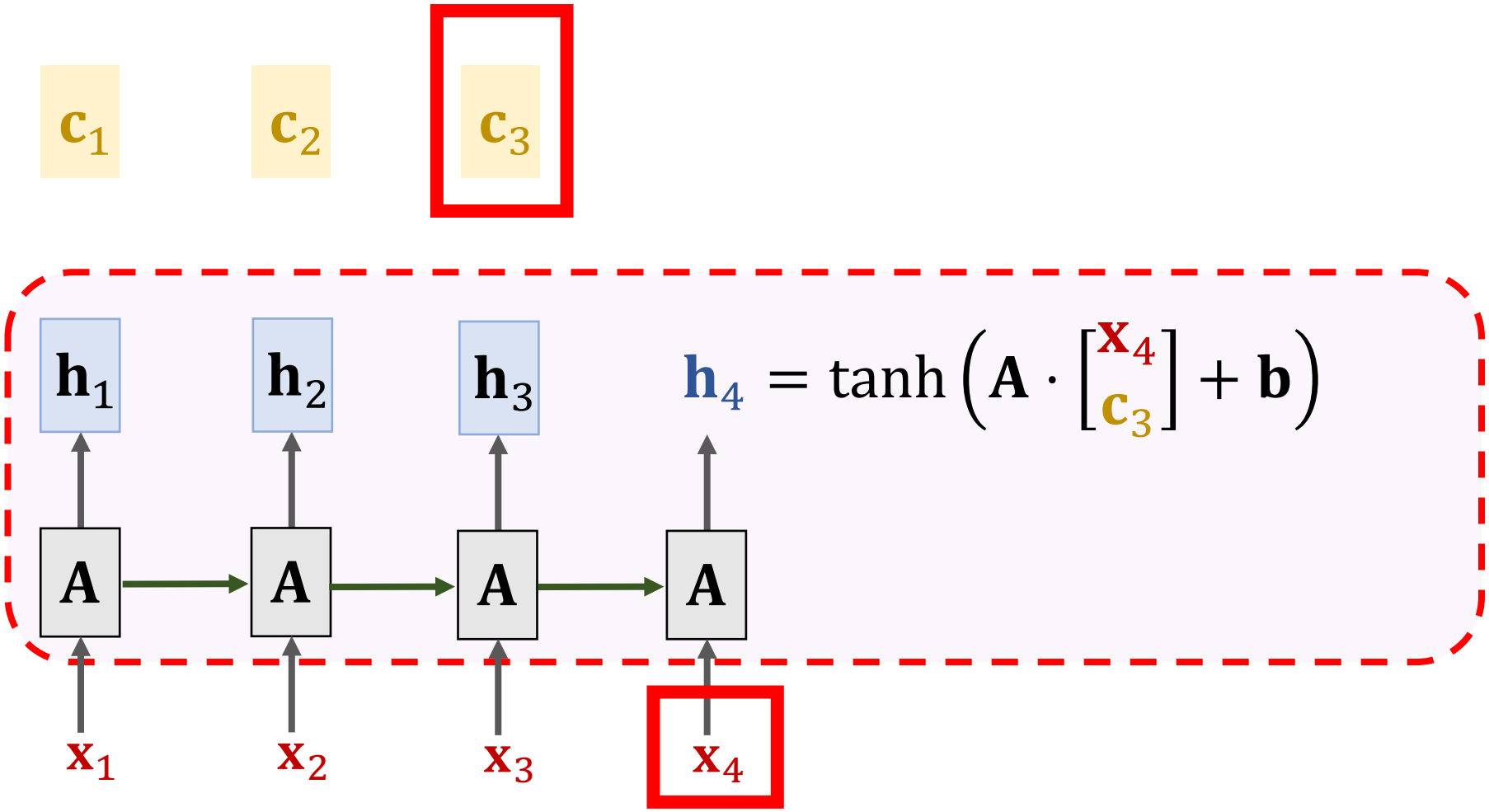
# SimpleRNN + Self-Attention



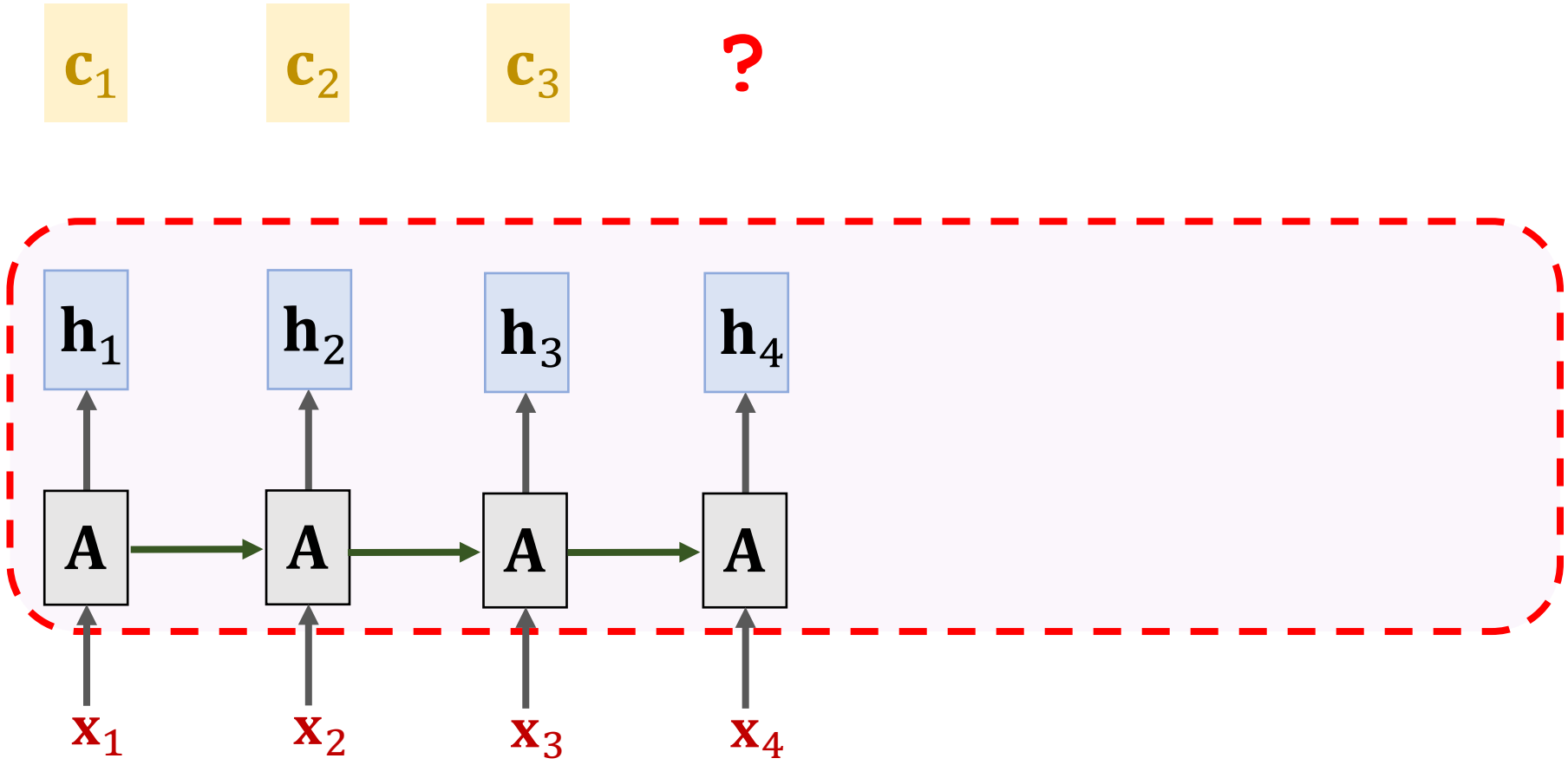
# SimpleRNN + Self-Attention



# SimpleRNN + Self-Attention

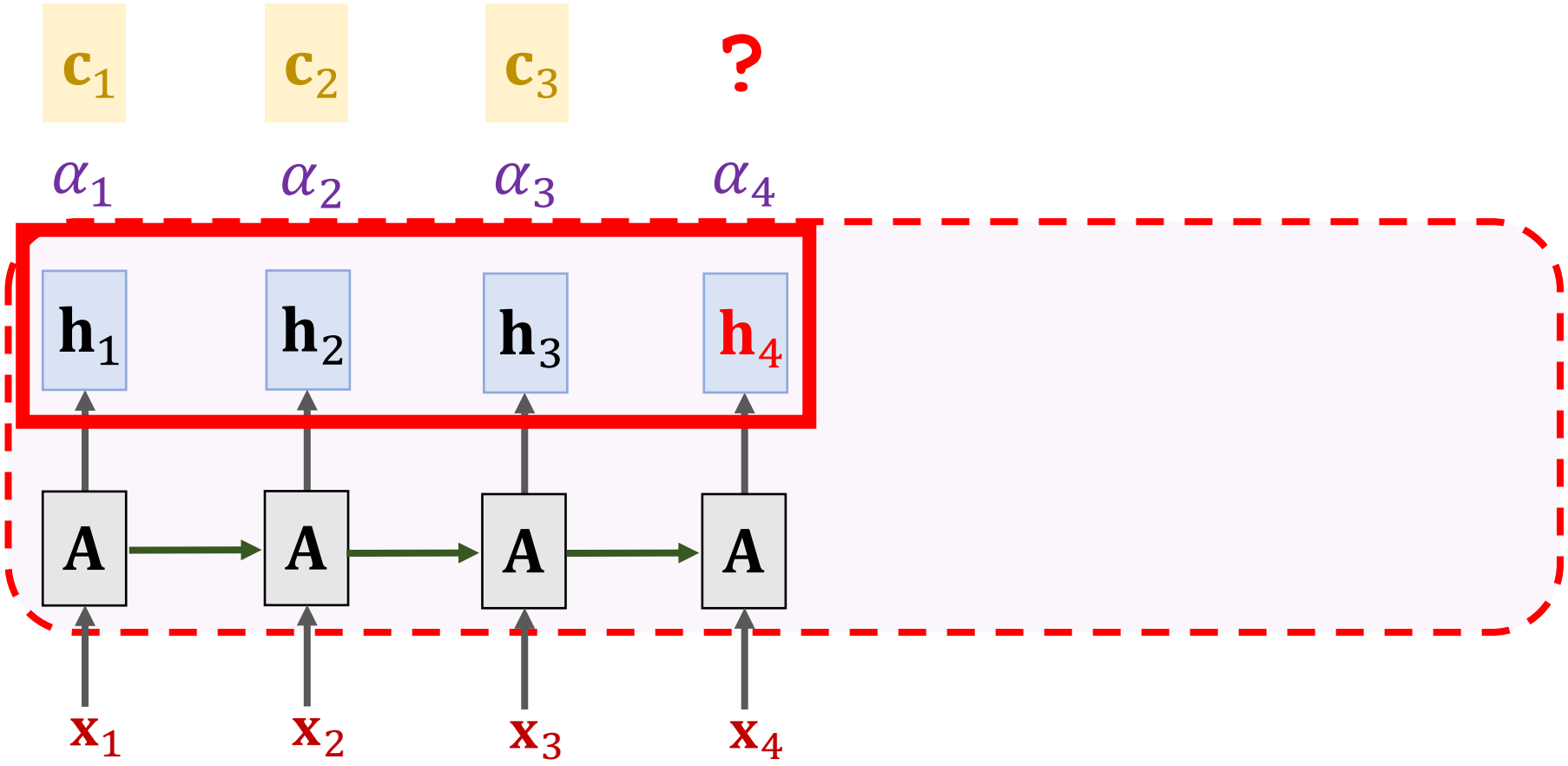


# SimpleRNN + Self-Attention

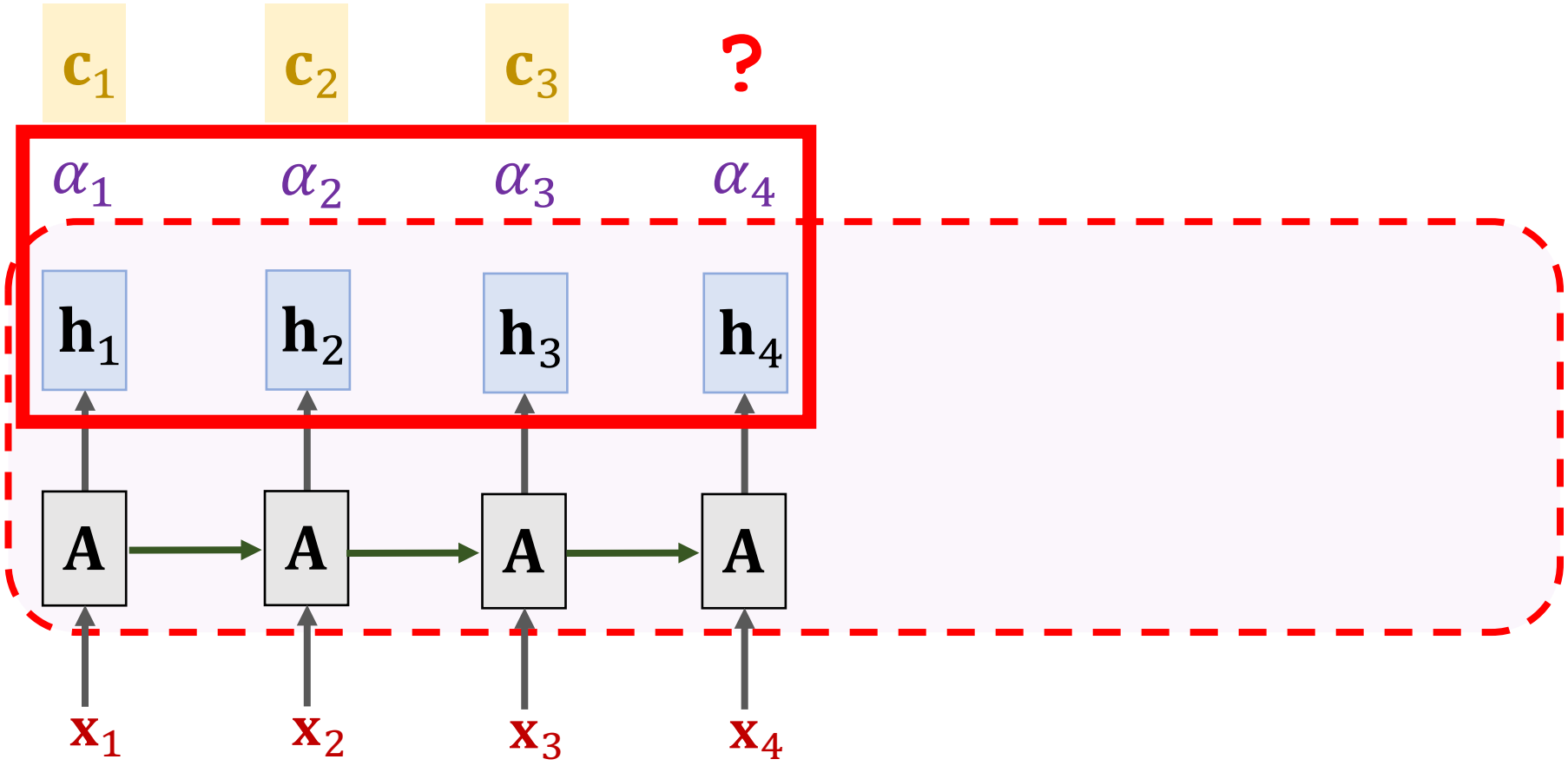


# SimpleRNN + Self-Attention

**Weights**: $\alpha_i = \text{align}(\mathbf{h}_i, \mathbf{h}_4)$.

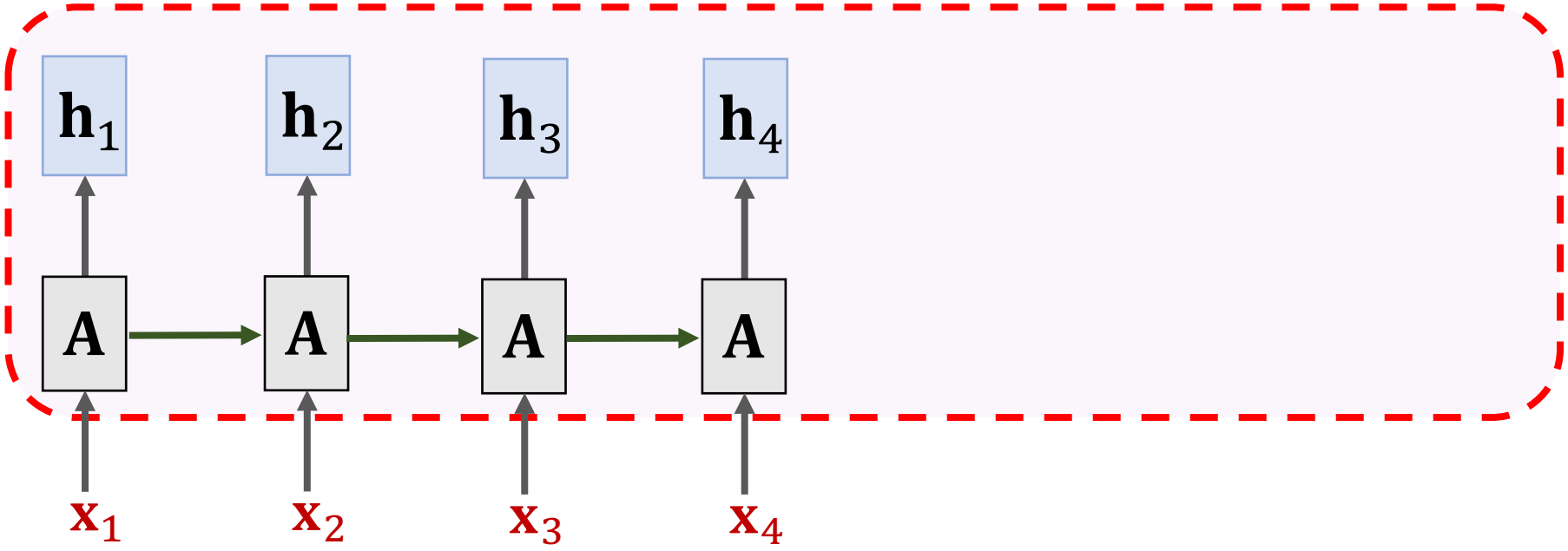


# SimpleRNN + Self-Attention

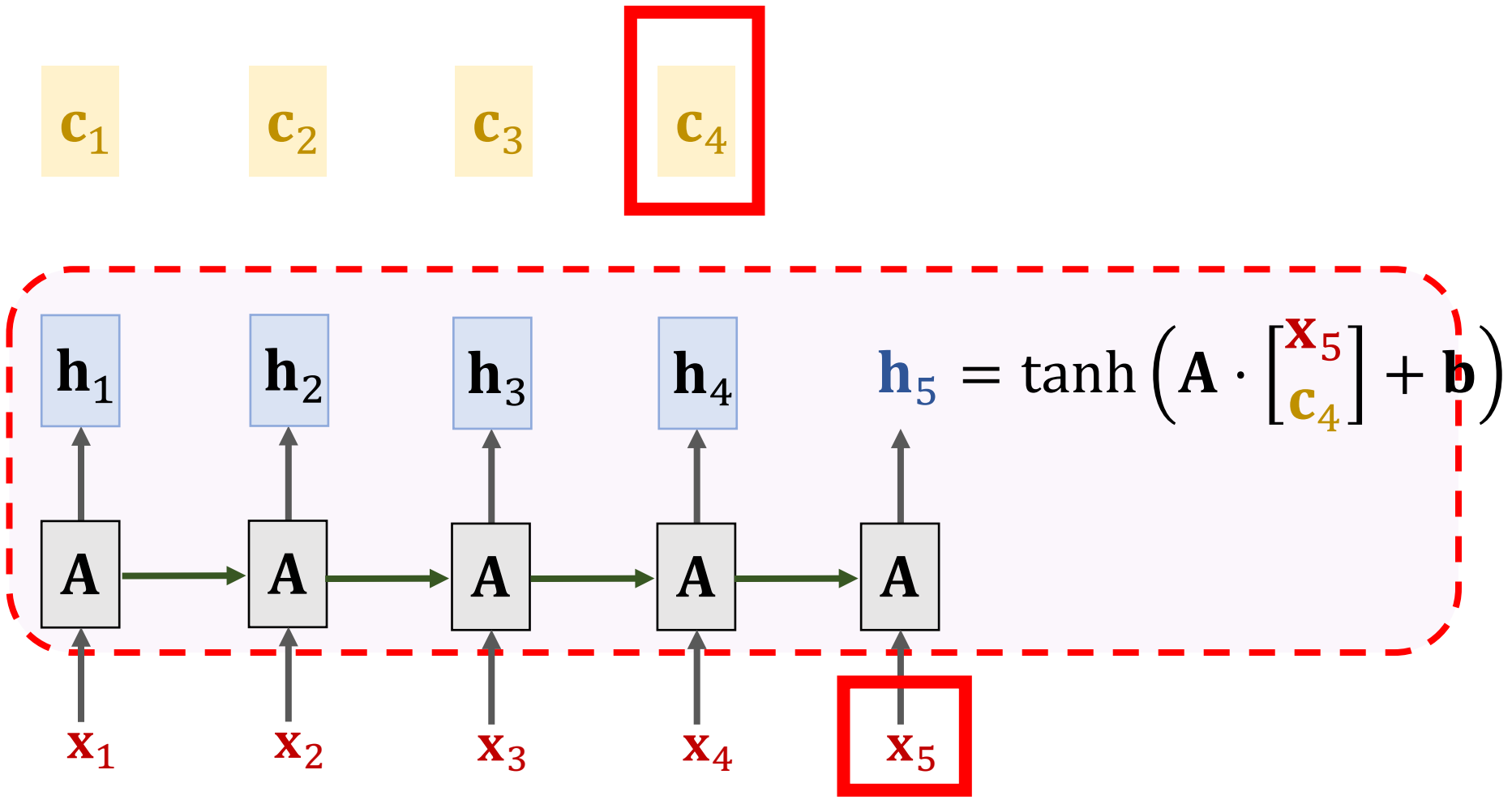


# SimpleRNN + Self-Attention

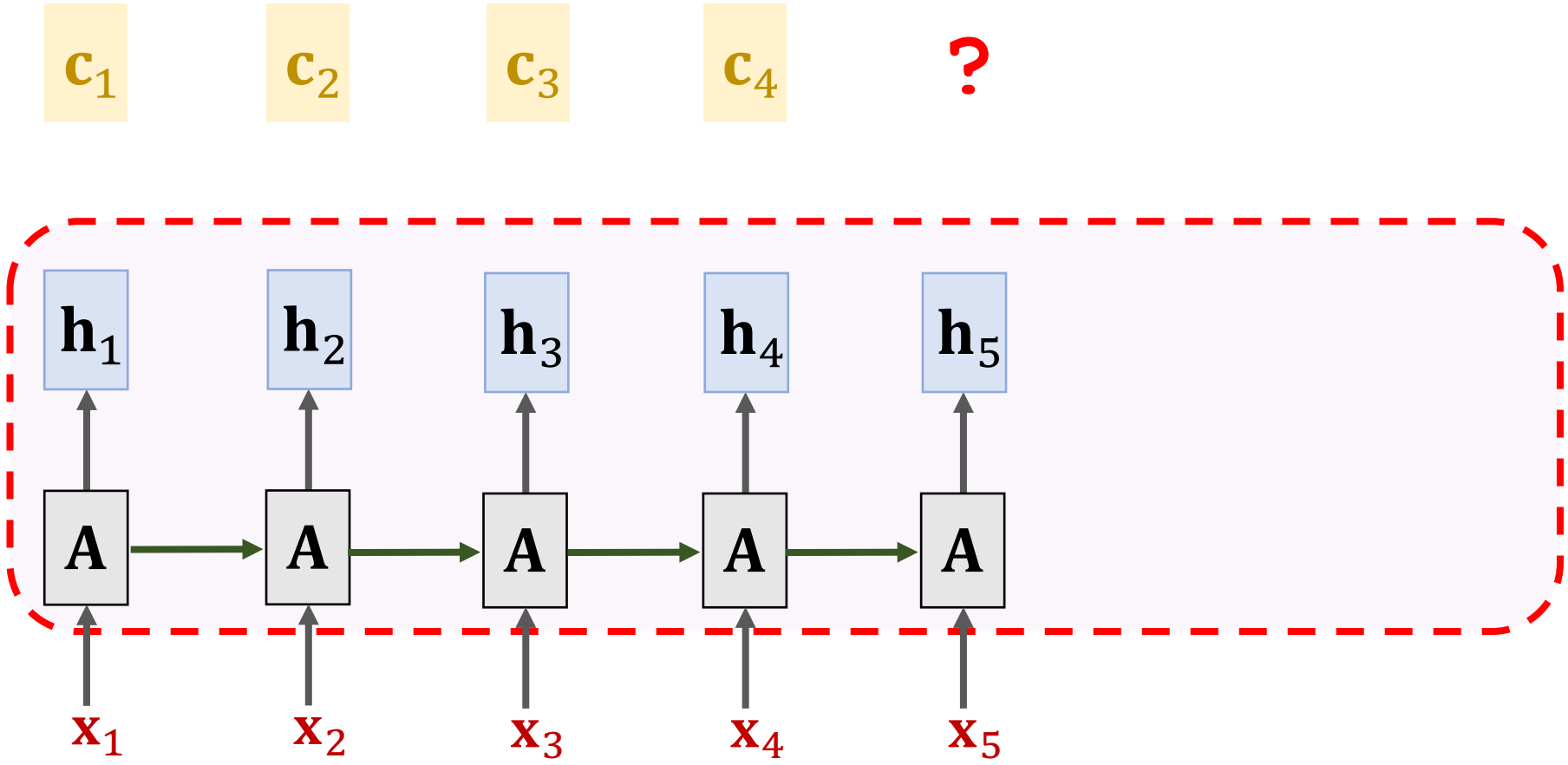
$\mathbf{c}_1$ $\mathbf{c}_2$ $\mathbf{c}_3$ $\mathbf{c}_4 = \alpha_1\mathbf{h}_1 + \alpha_2\mathbf{h}_2 + \alpha_3\mathbf{h}_3 + \alpha_4\mathbf{h}_4.$



# SimpleRNN + Self-Attention

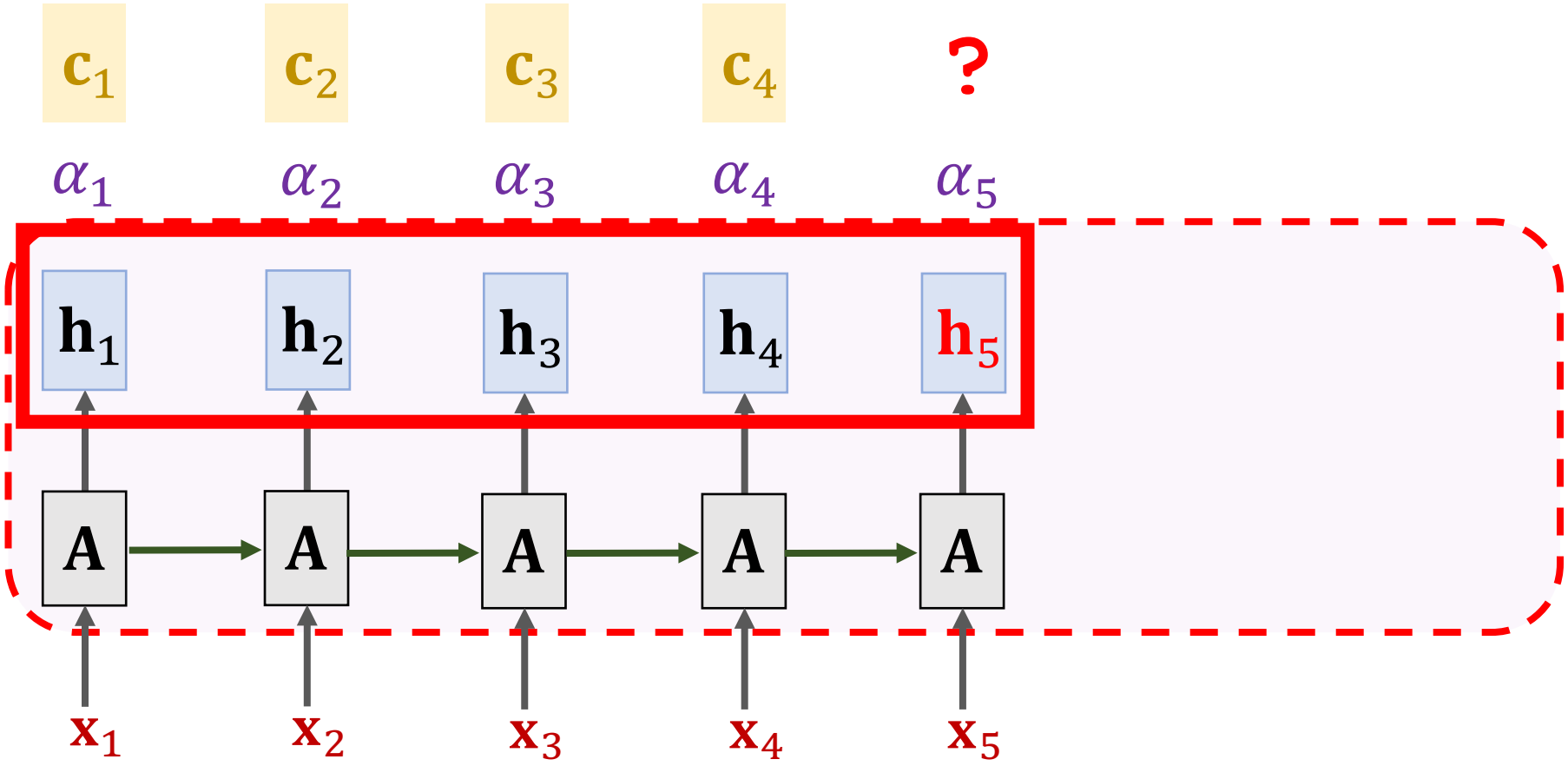


# SimpleRNN + Self-Attention

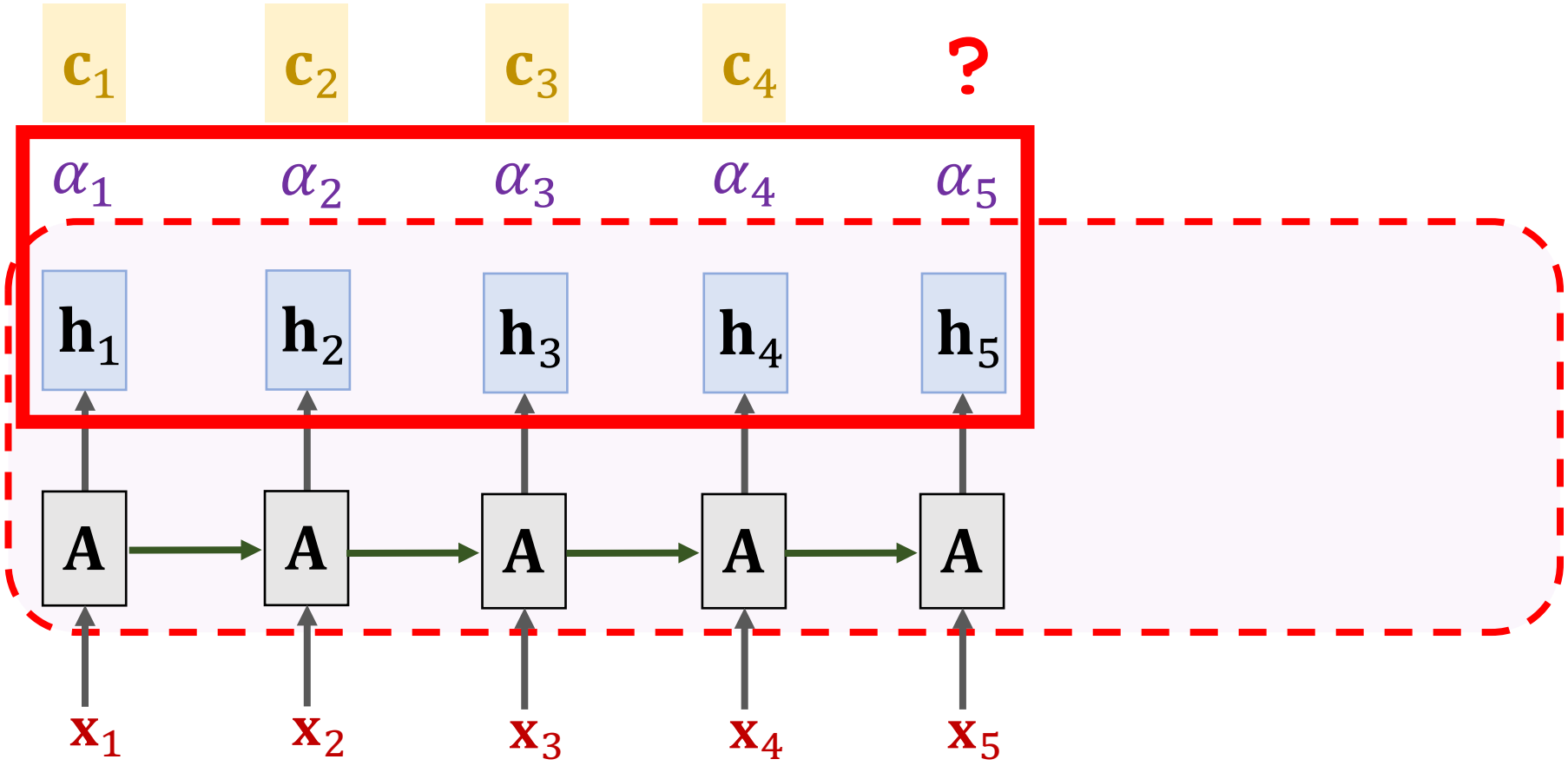


# SimpleRNN + Self-Attention

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

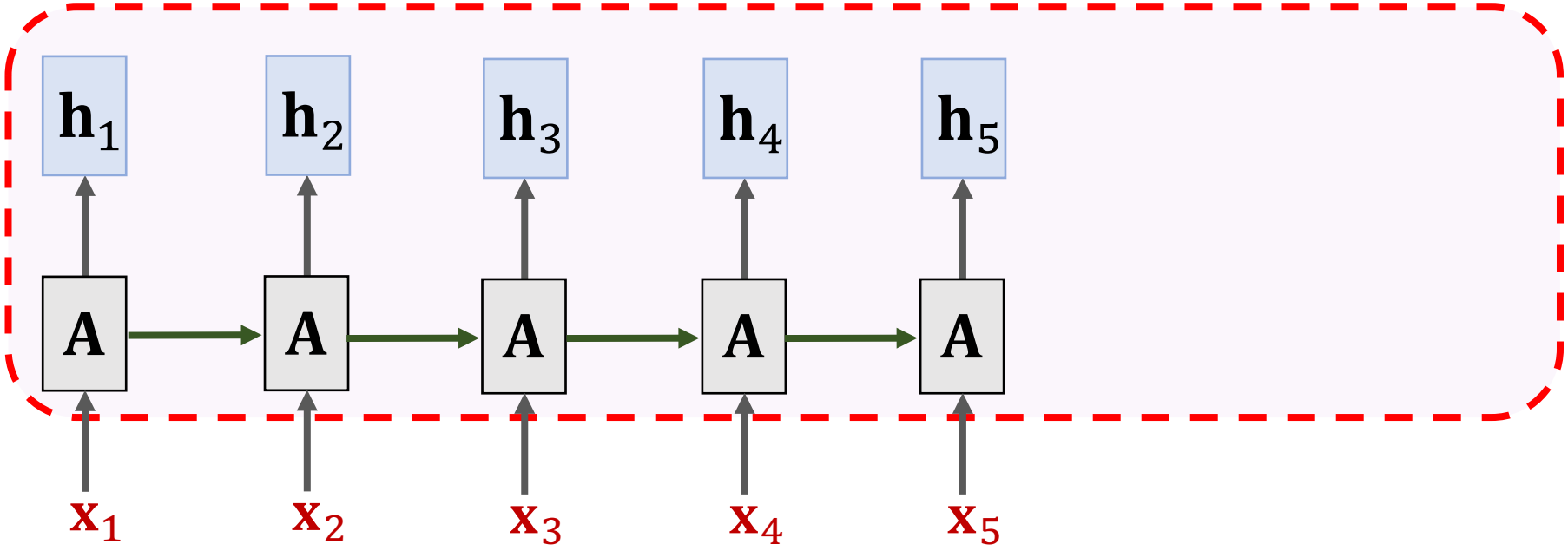


# SimpleRNN + Self-Attention

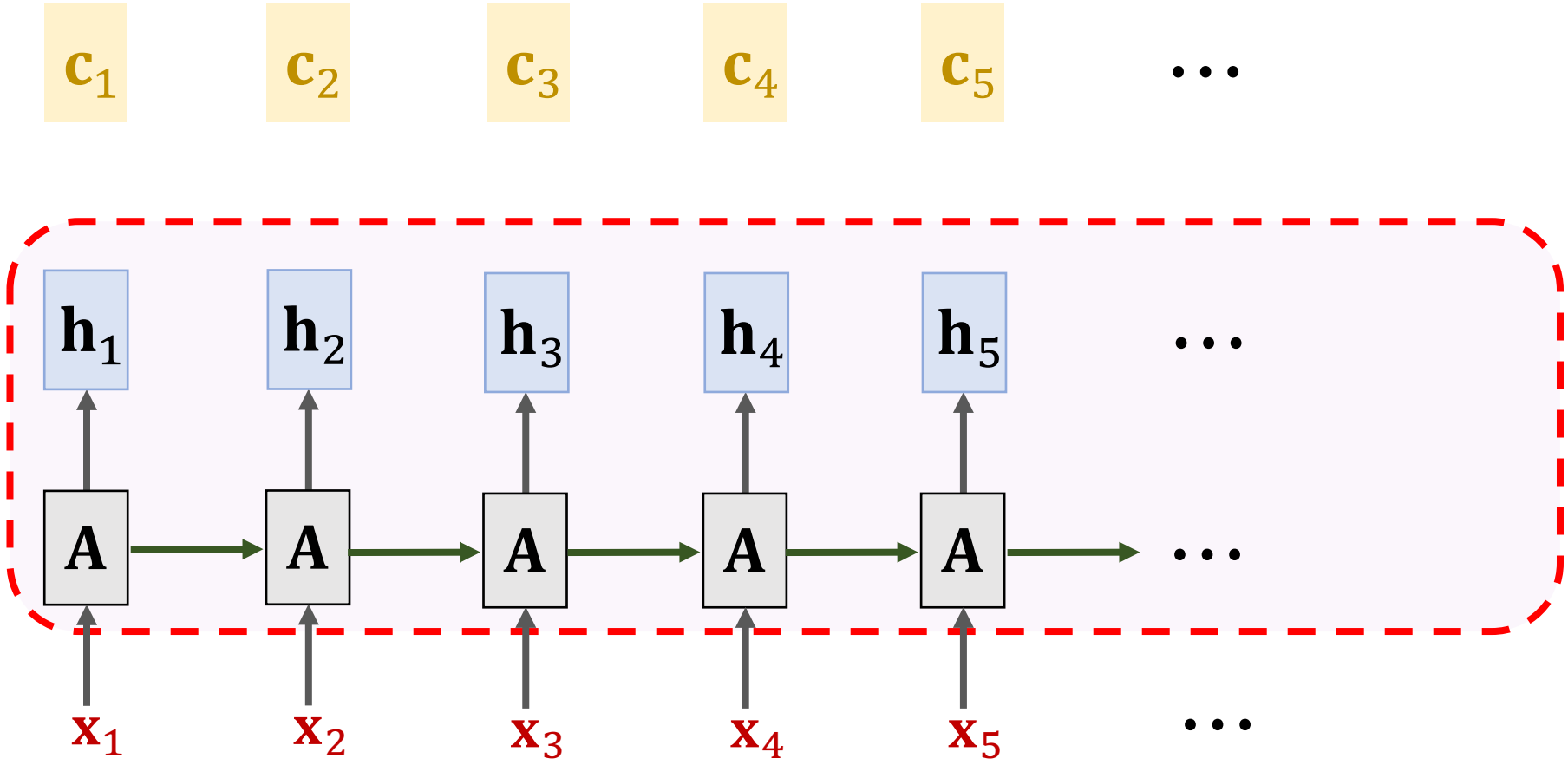


# SimpleRNN + Self-Attention

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



# SimpleRNN + Self-Attention



# Summary

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

# Summary

- With self-attention, RNN is less likely to forget.
- Pay attention to the context relevant to the new input.

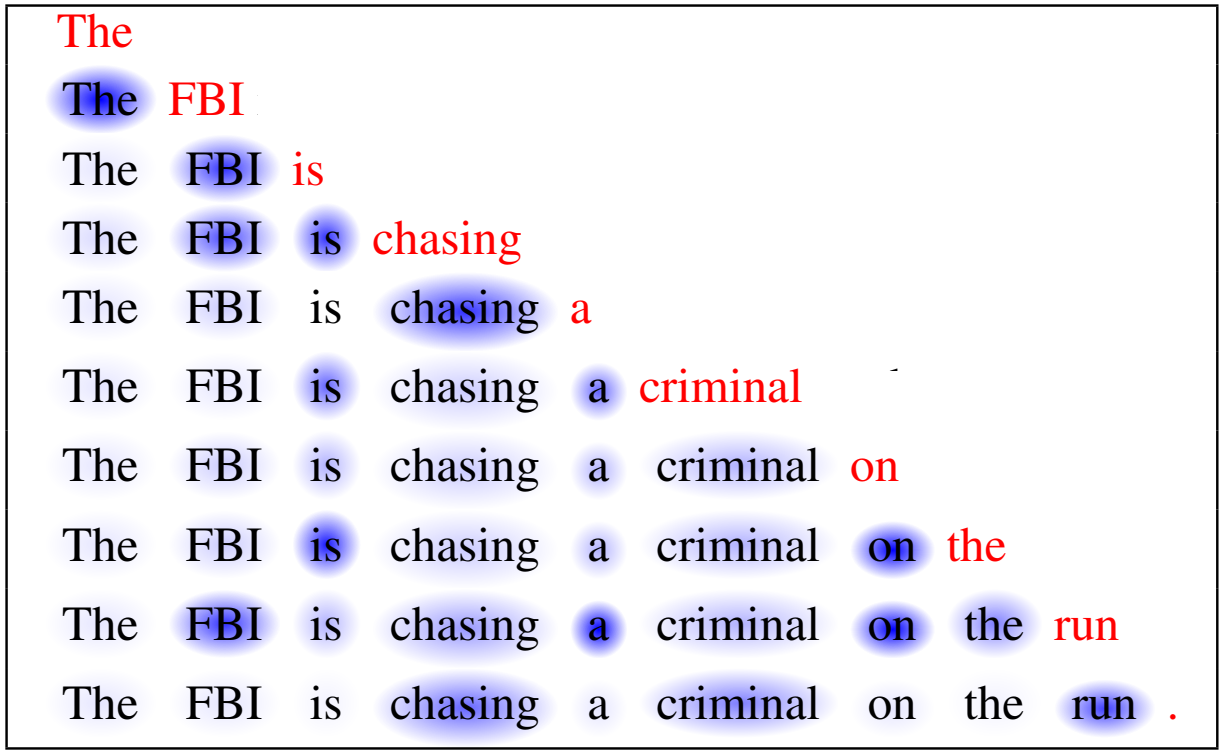



Figure is from the paper " Long Short-Term Memory-Networks for Machine Reading."

# Thank you!