Self-Attention

Shusen Wang

Self-Attention

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

Original paper:

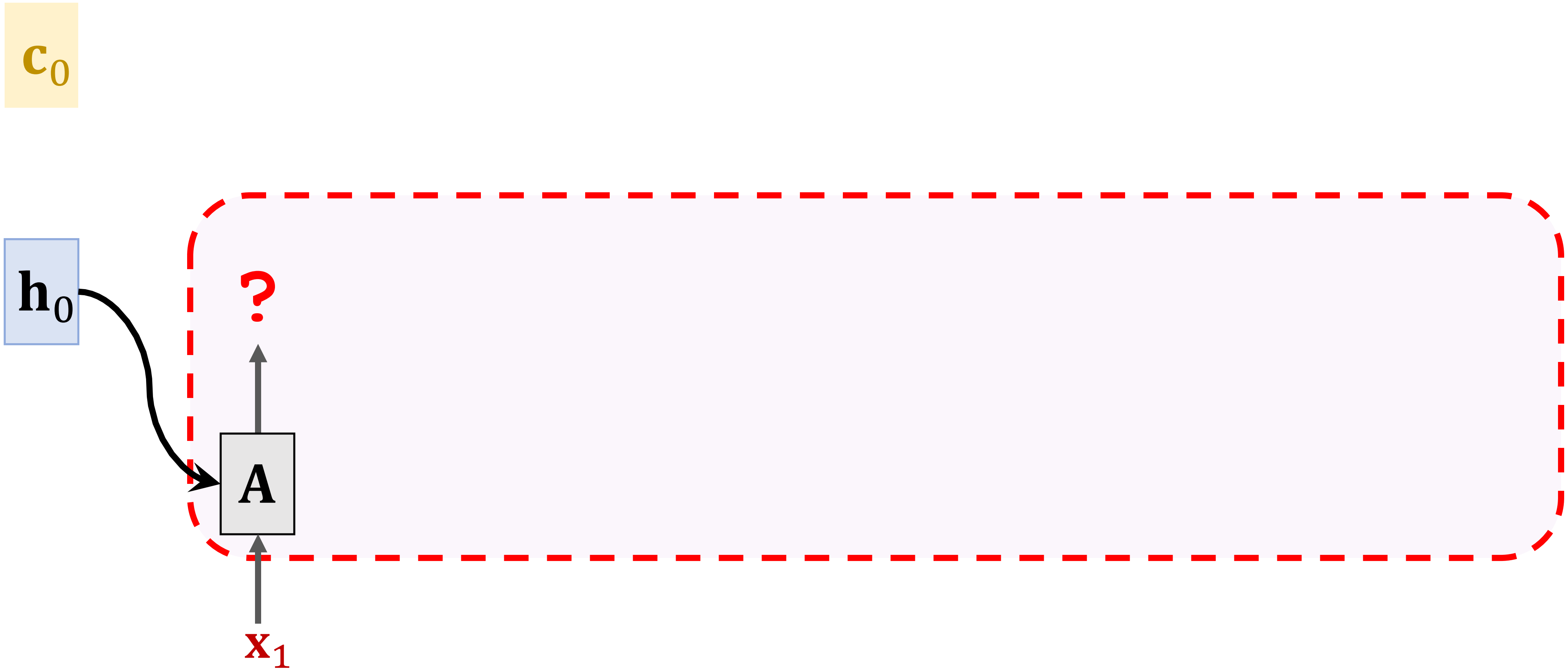
- Cheng, Dong, & Lapata. Long Short-Term Memory-Networks for Machine Reading. In *EMNLP*, 2016.

SimpleRNN + Self-Attention

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

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

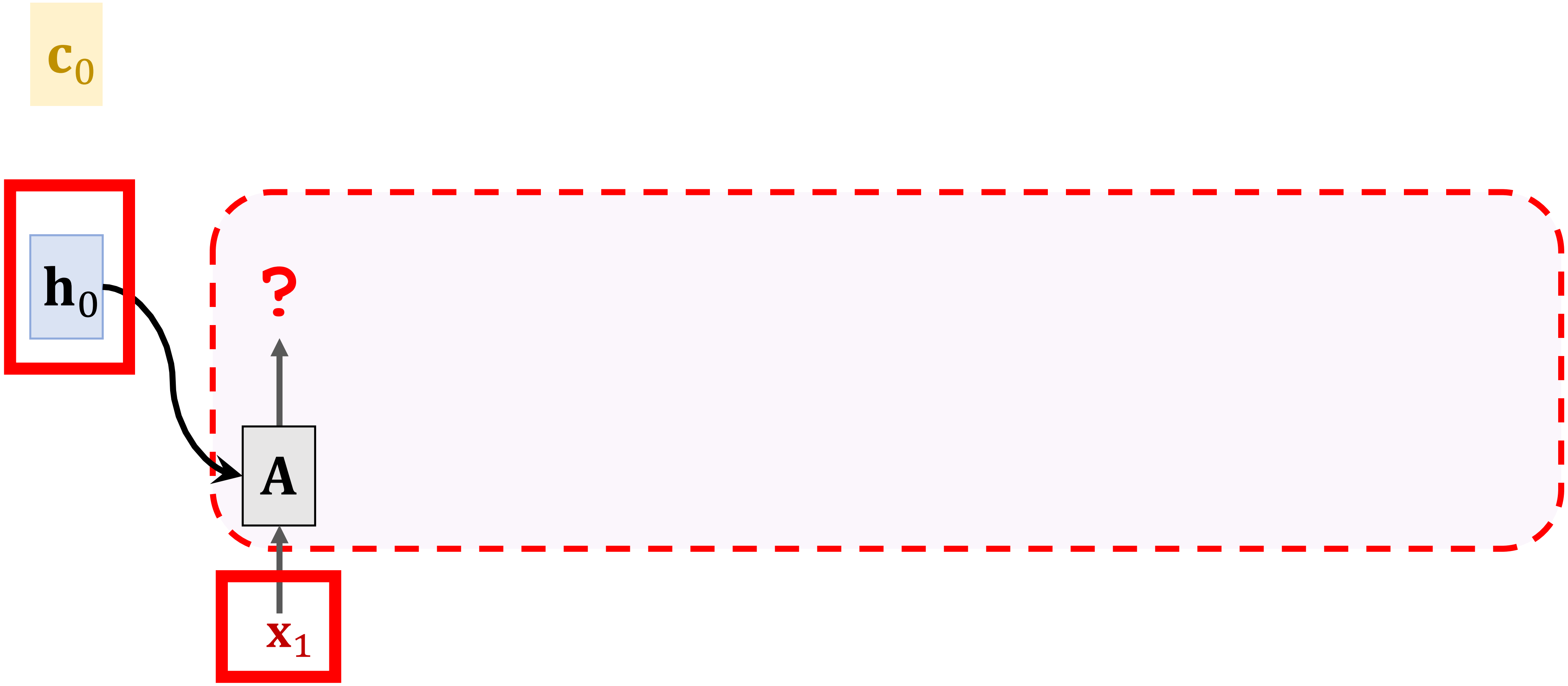
SimpleRNN + Self-Attention



SimpleRNN + Self-Attention

SimpleRNN:

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



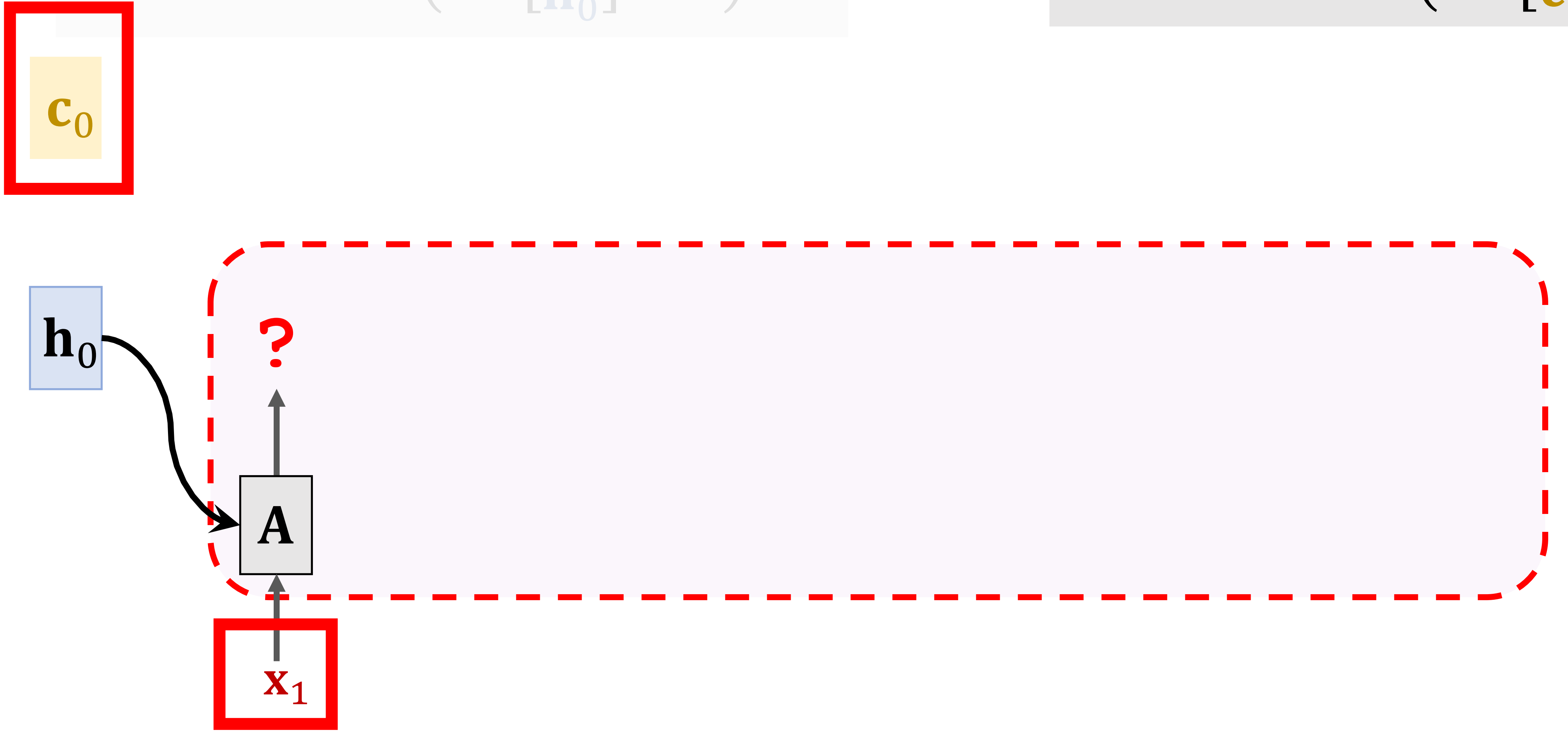
SimpleRNN + Self-Attention

SimpleRNN:

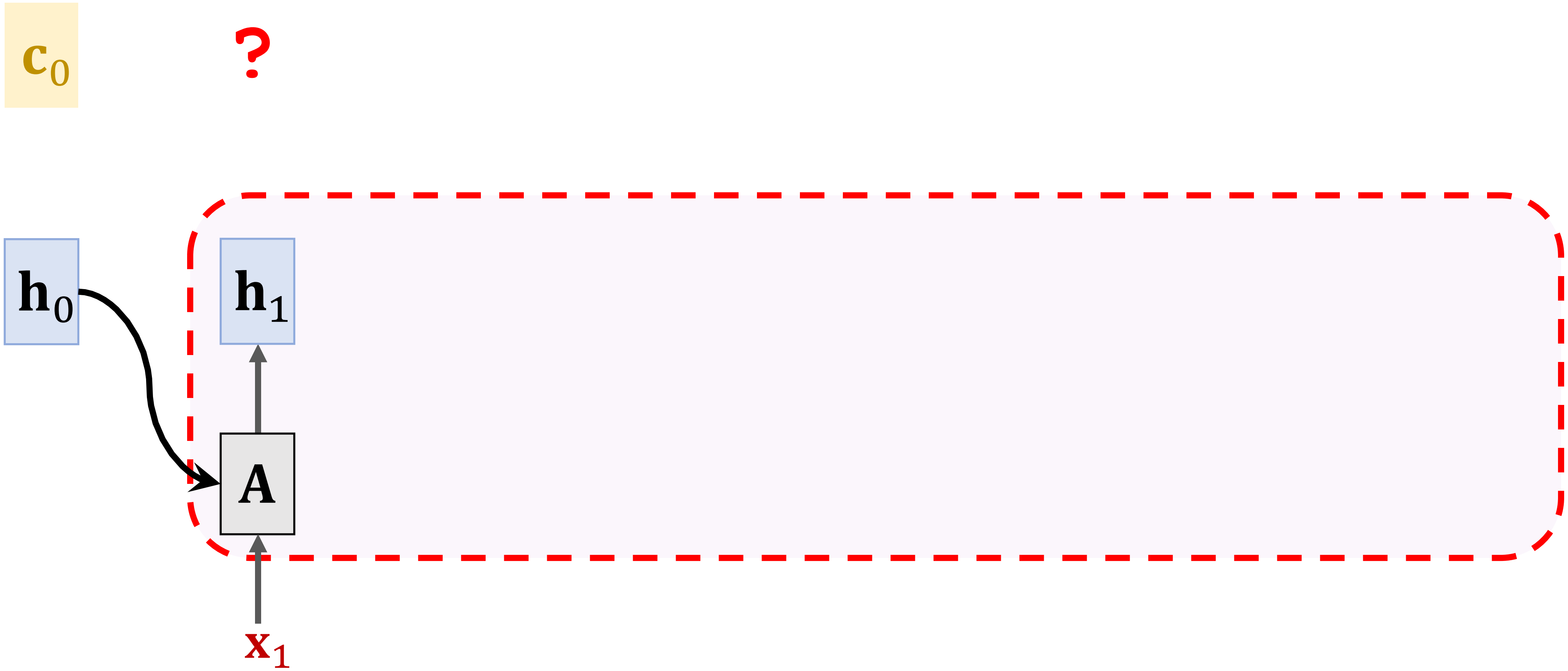
$$\mathbf{h}_1 = \tanh\left(\mathbf{A} \cdot \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{h}_0 \end{bmatrix} + \mathbf{b}\right)$$

SimpleRNN + Self-Attention:

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

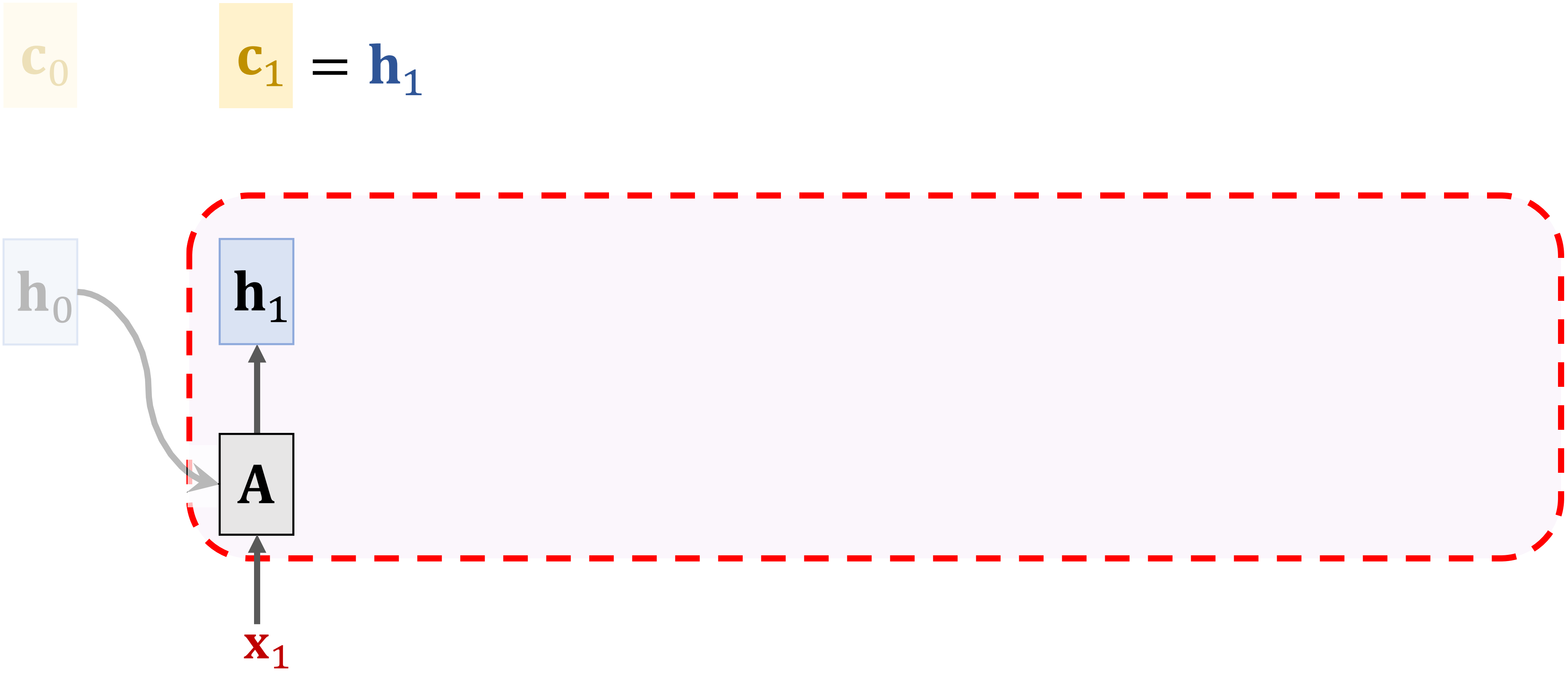


SimpleRNN + Self-Attention

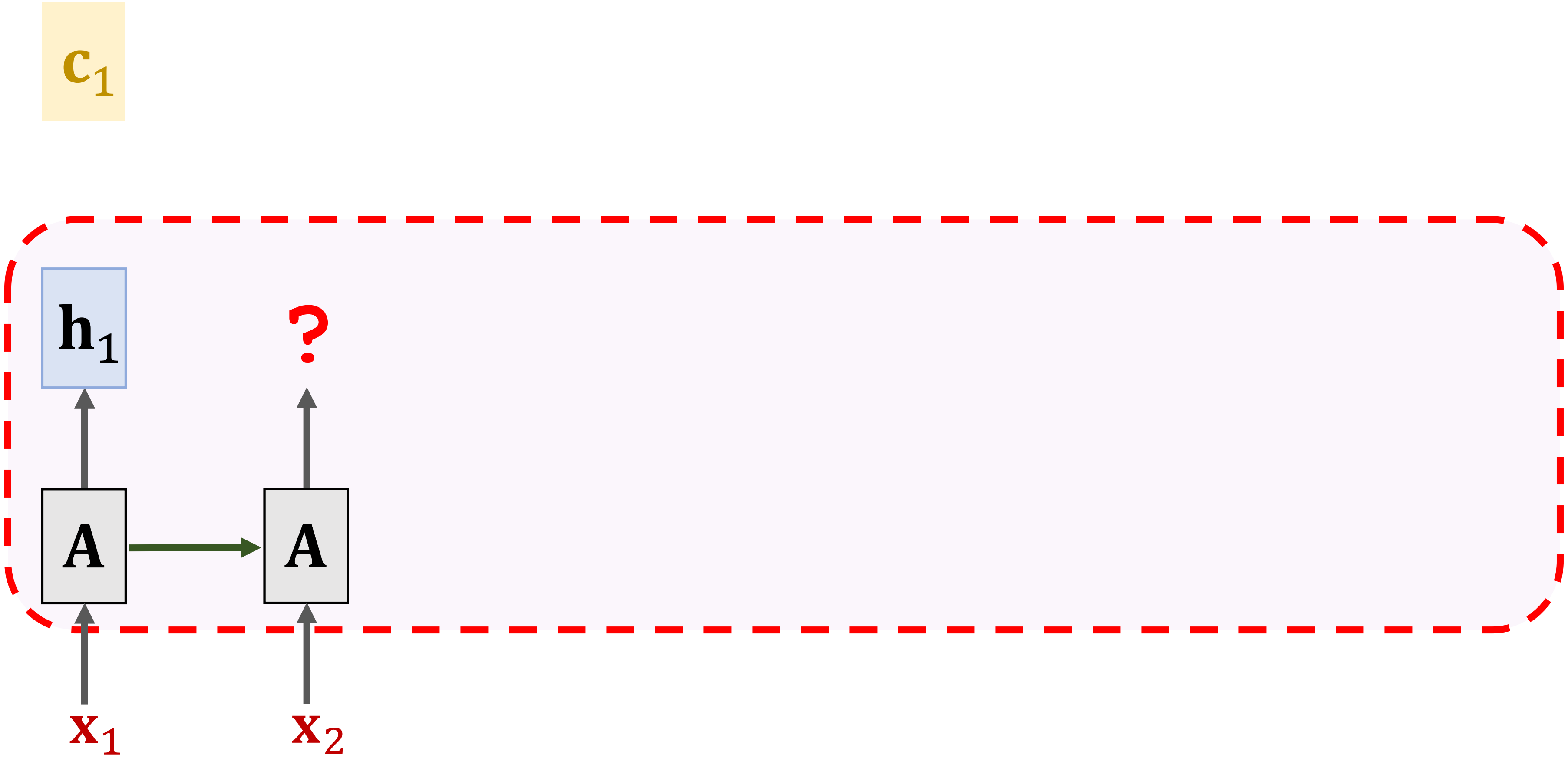


SimpleRNN + Self-Attention

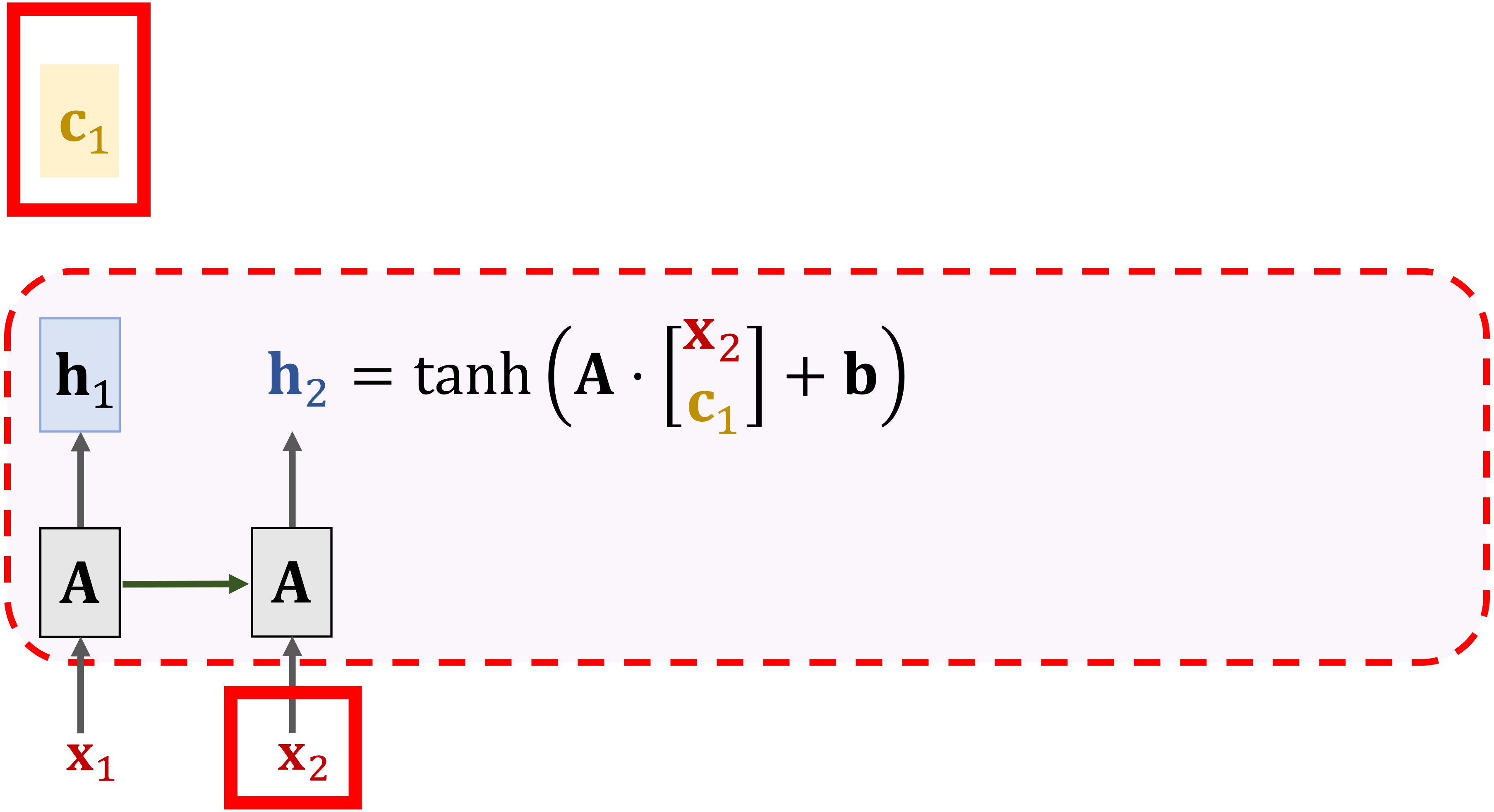
First context vector: $\mathbf{c}_1 = \mathbf{h}_1$.



SimpleRNN + Self-Attention

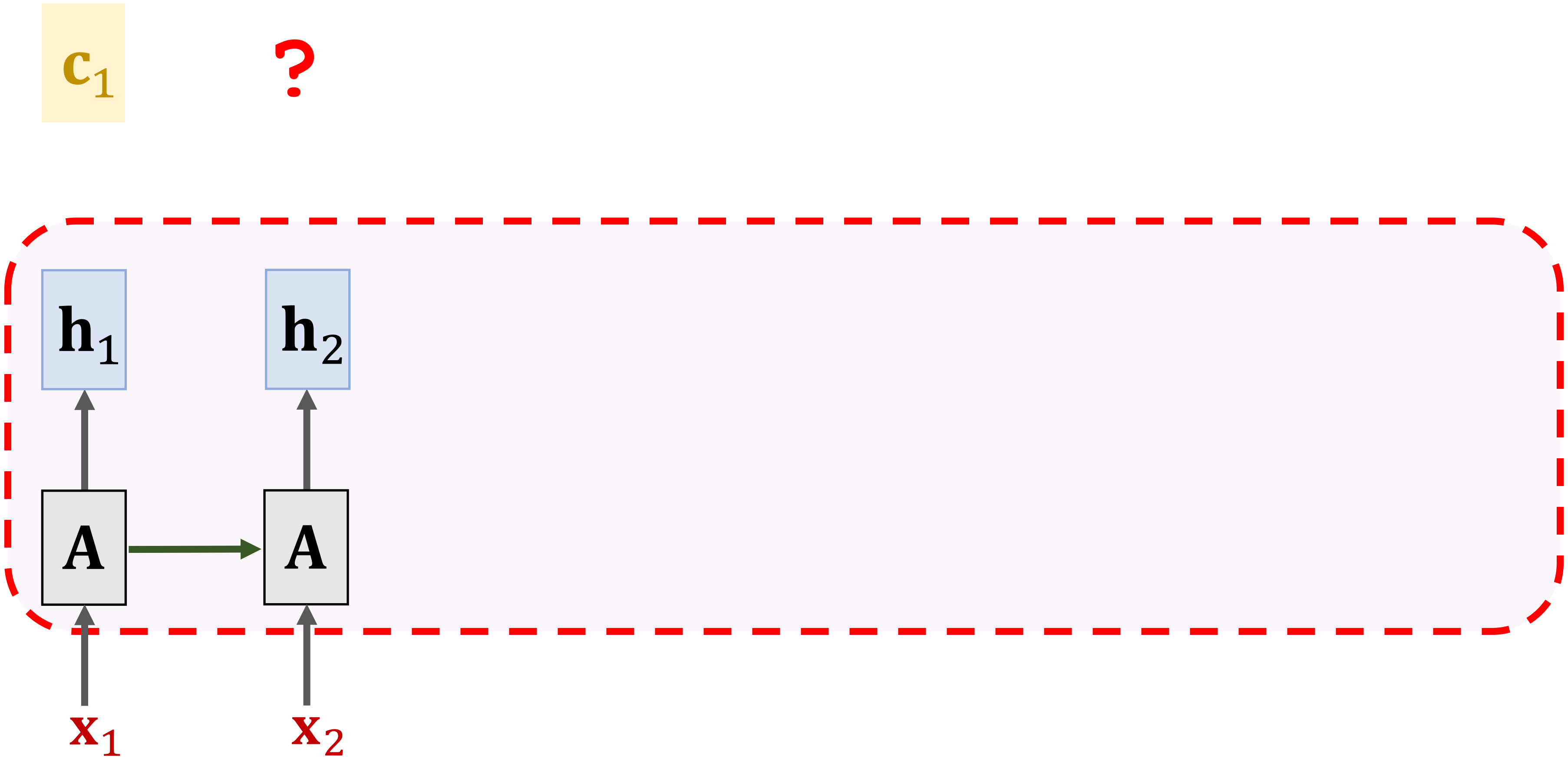


SimpleRNN + Self-Attention



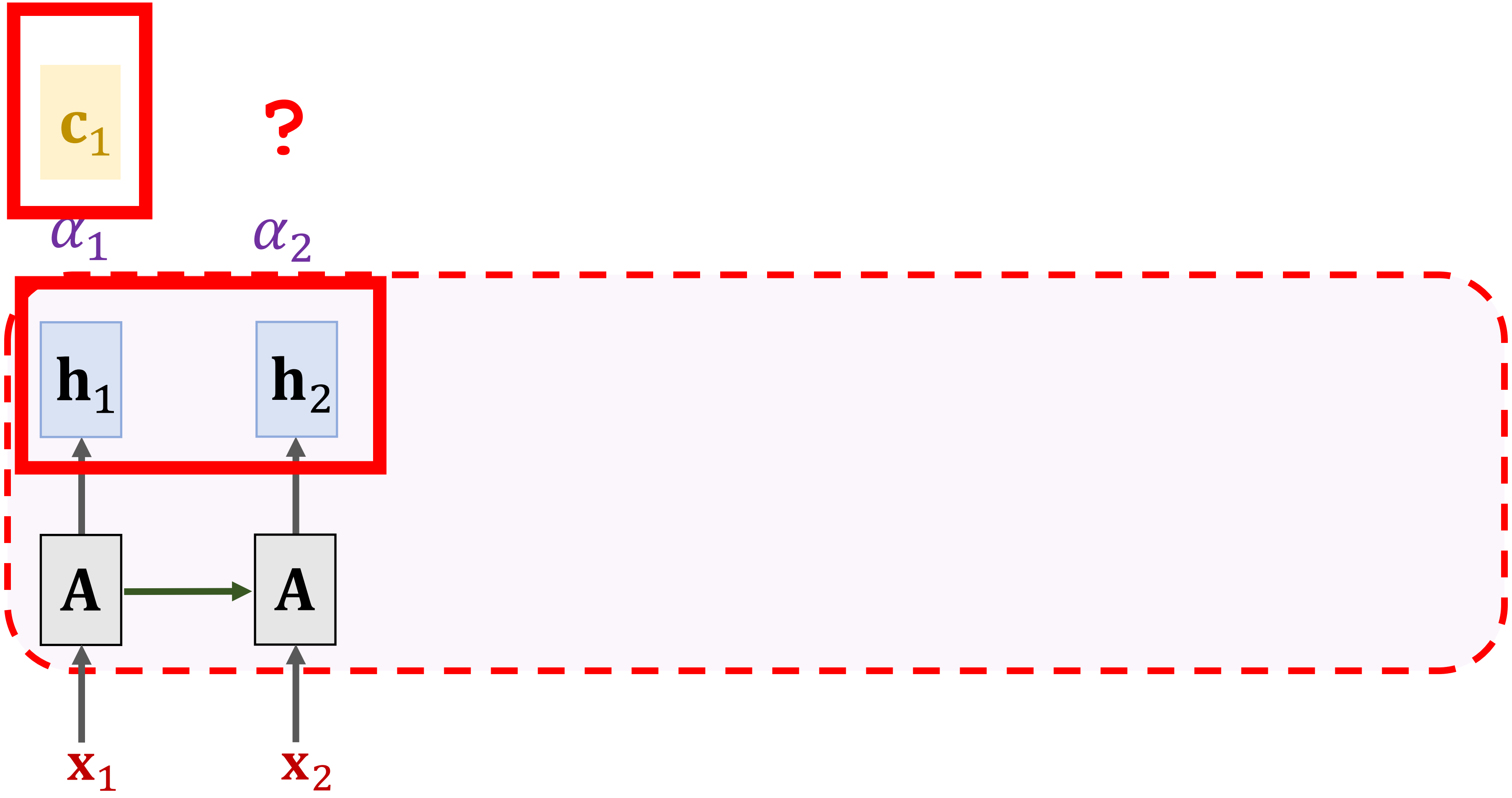
$$\mathbf{h}_2 = \tanh\left(\mathbf{A}\cdot\begin{bmatrix}\mathbf{x}_2\\ \mathbf{c}_1\end{bmatrix}+\mathbf{b}\right)$$

SimpleRNN + Self-Attention



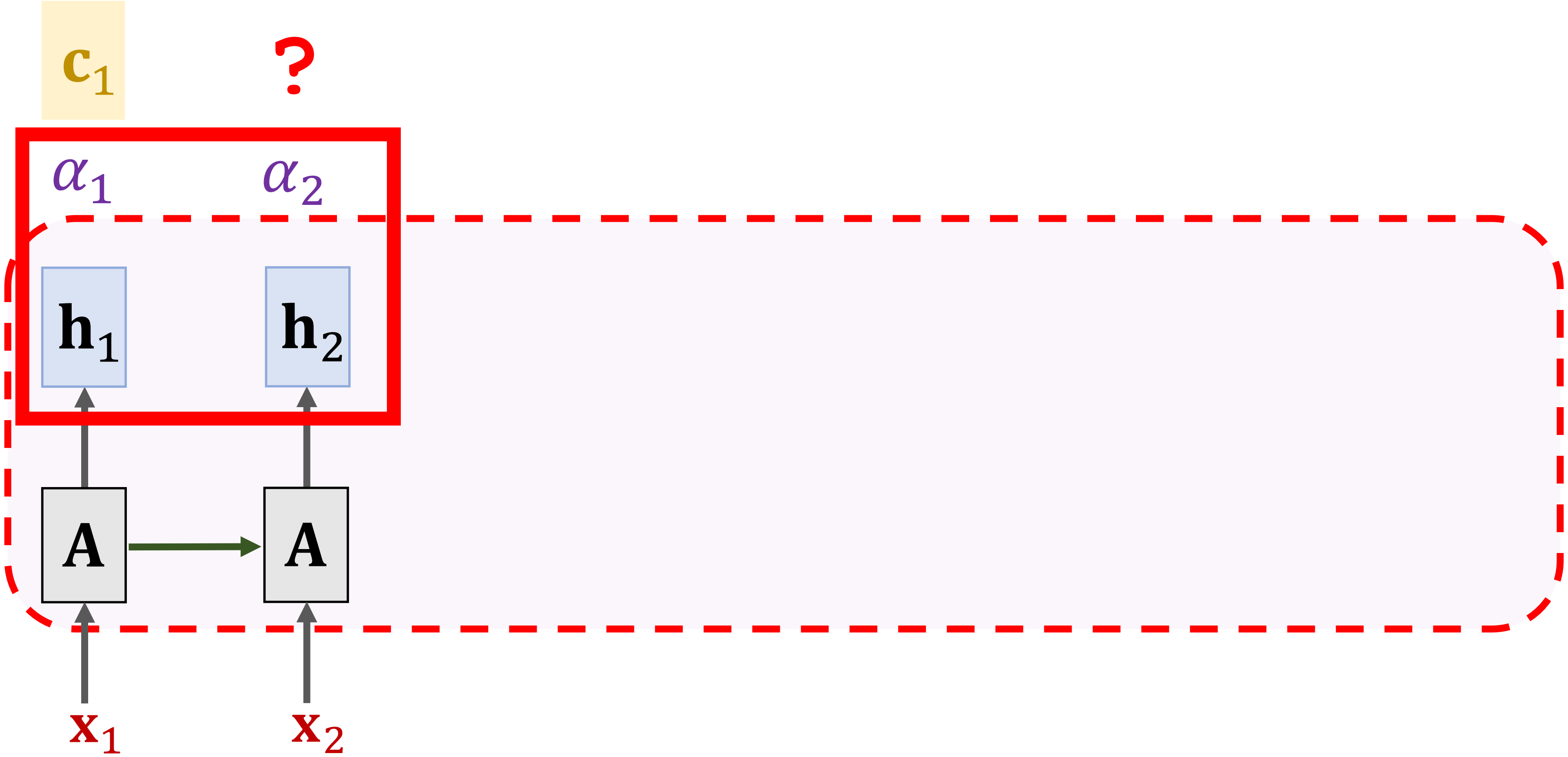
SimpleRNN + Self-Attention

Weights: $\alpha_i = \text{align}(\mathbf{h}_i,\ \mathbf{c}_1)$.



SimpleRNN + Self-Attention

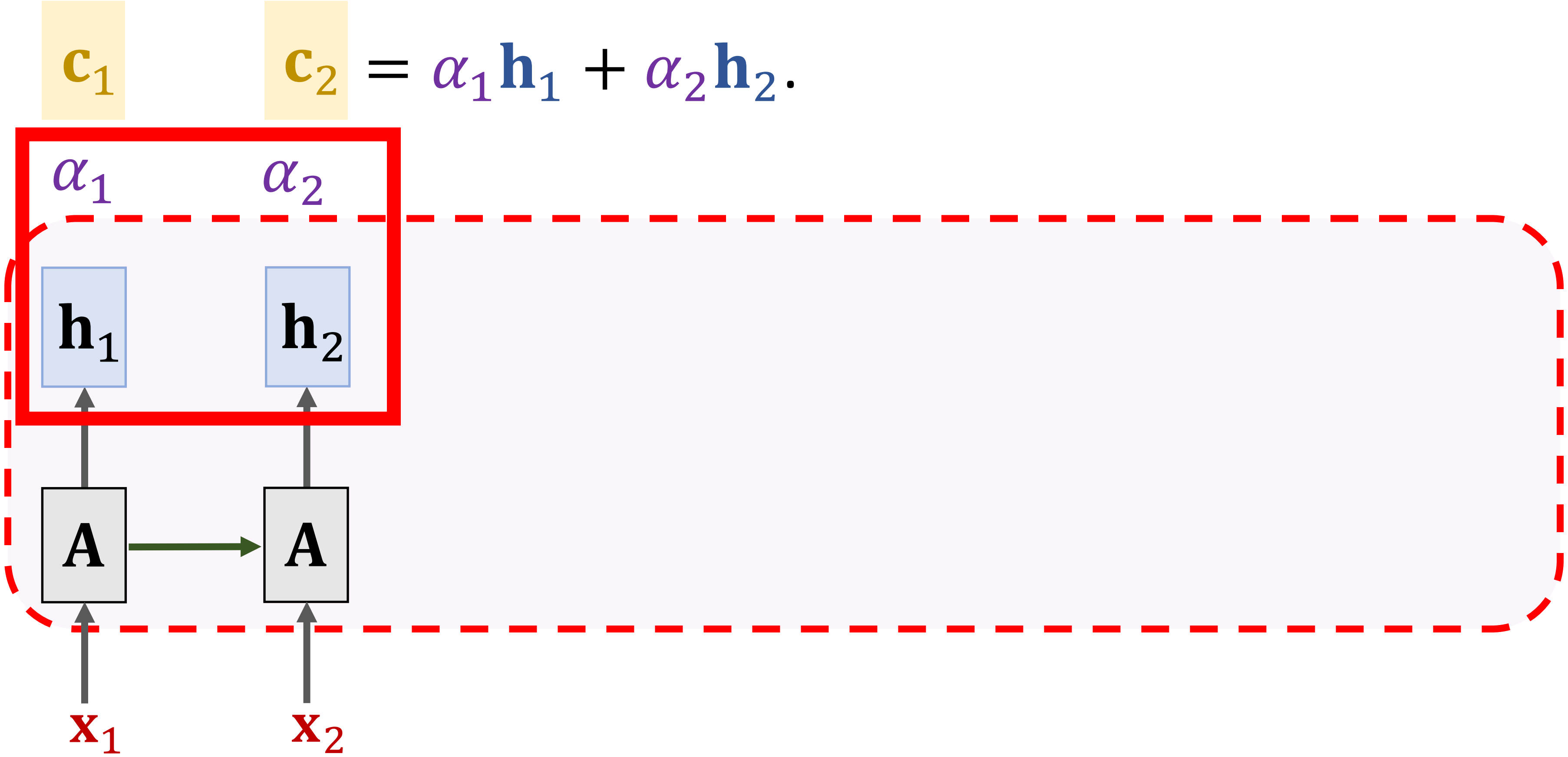
Weights: $\alpha_i = \text{align}(\mathbf{h}_i,\ \mathbf{c}_1)$.



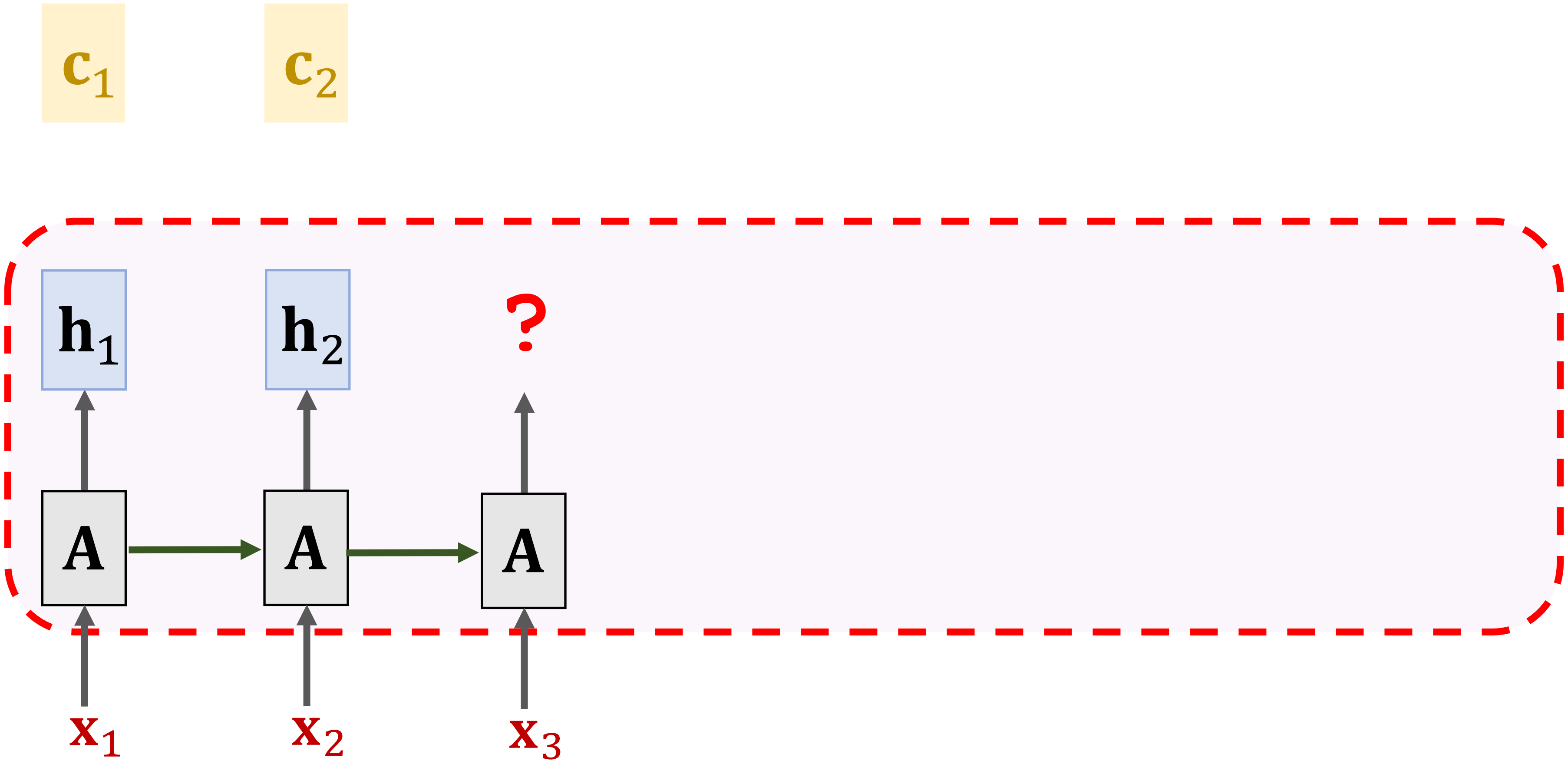
SimpleRNN + Self-Attention

Weights: $\alpha_i = \text{align}(\mathbf{h}_i,\ \mathbf{c}_1).$

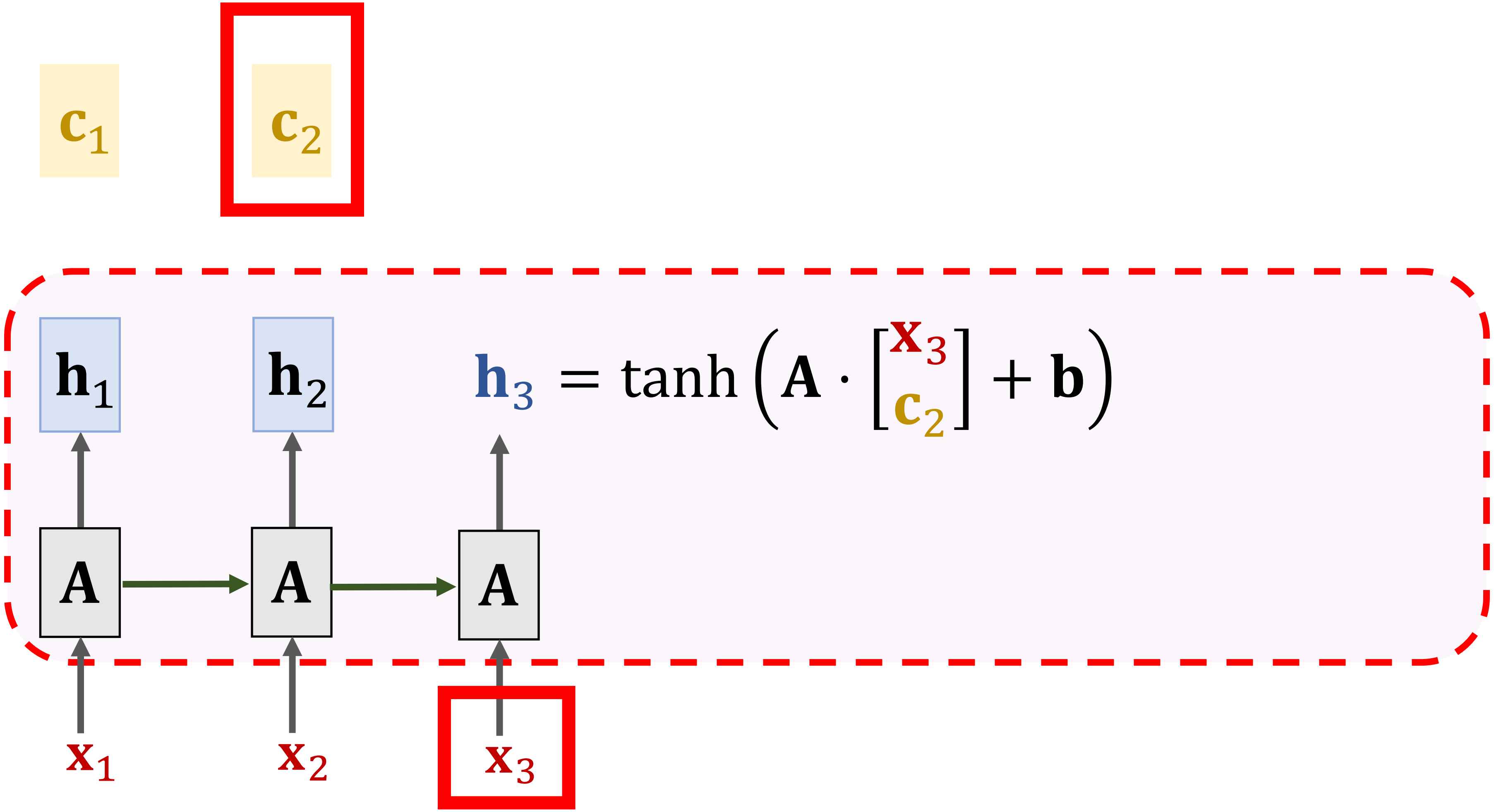
$\mathbf{c}_2 = \alpha_1 \mathbf{h}_1 + \alpha_2 \mathbf{h}_2.$



SimpleRNN + Self-Attention

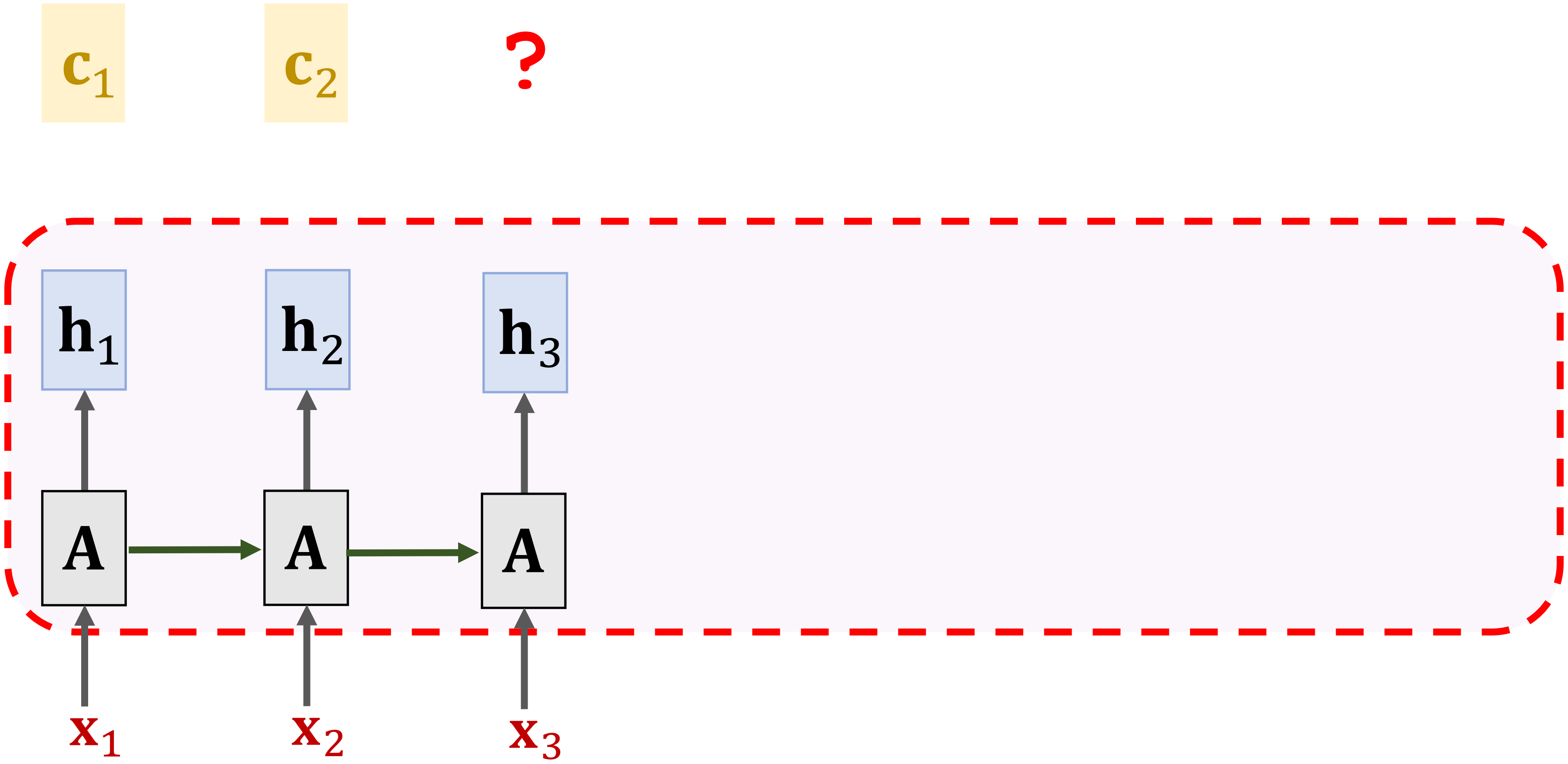


SimpleRNN + Self-Attention



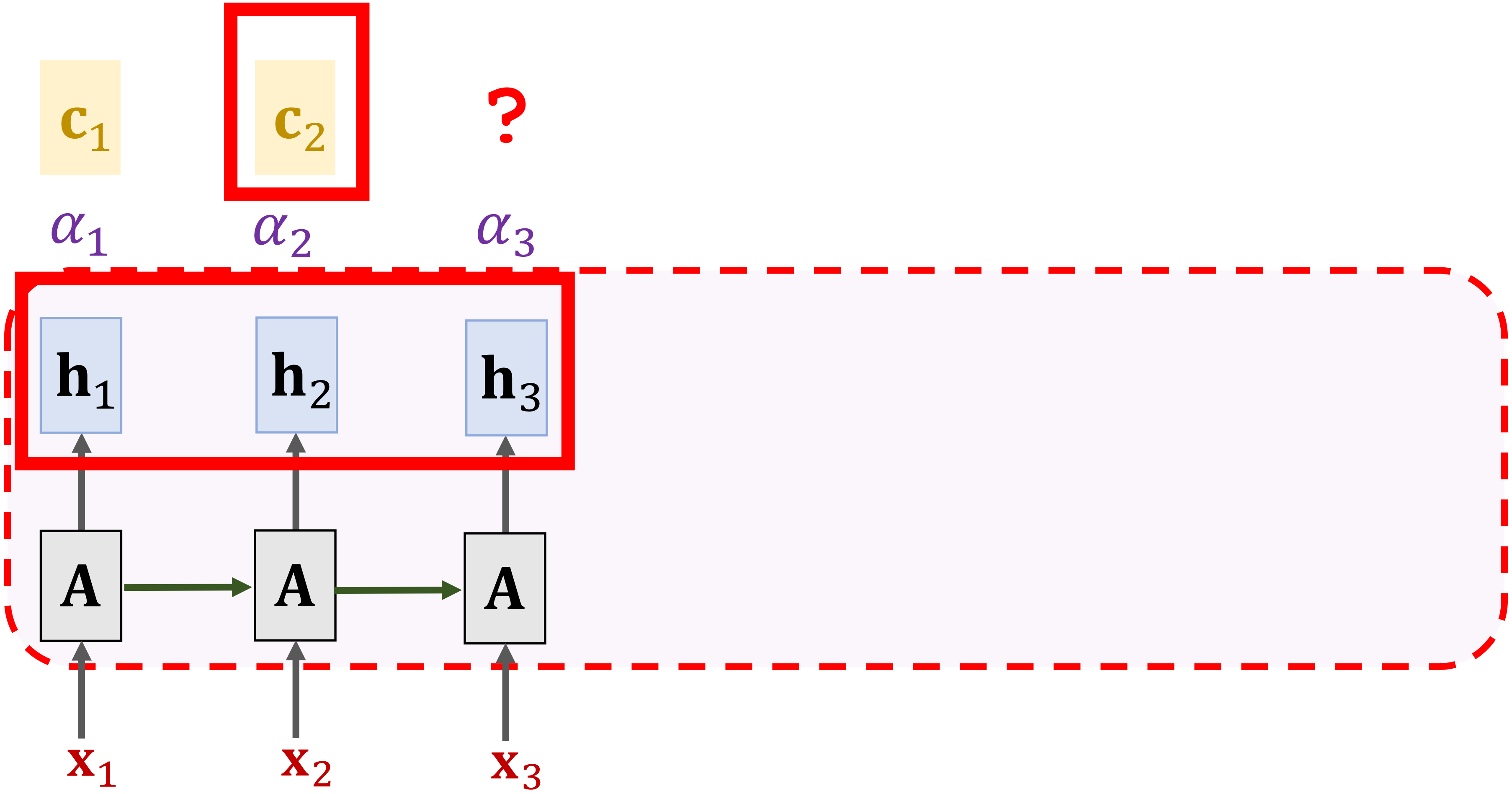
$$\mathbf{h}_3 = \tanh\left(\mathbf{A} \cdot \begin{bmatrix} \mathbf{x}_3 \\ \mathbf{c}_2 \end{bmatrix} + \mathbf{b}\right)$$

SimpleRNN + Self-Attention



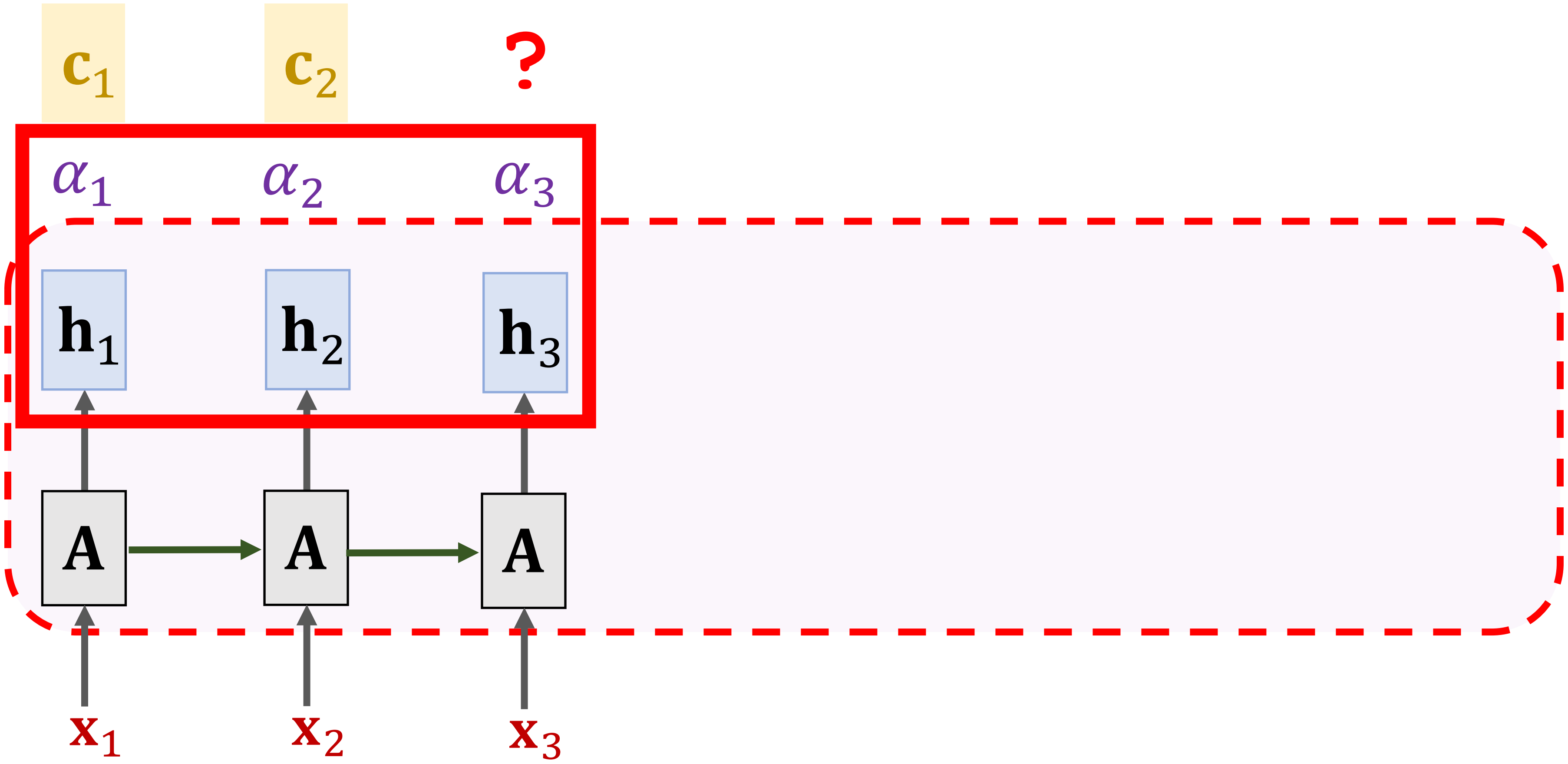
SimpleRNN + Self-Attention

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



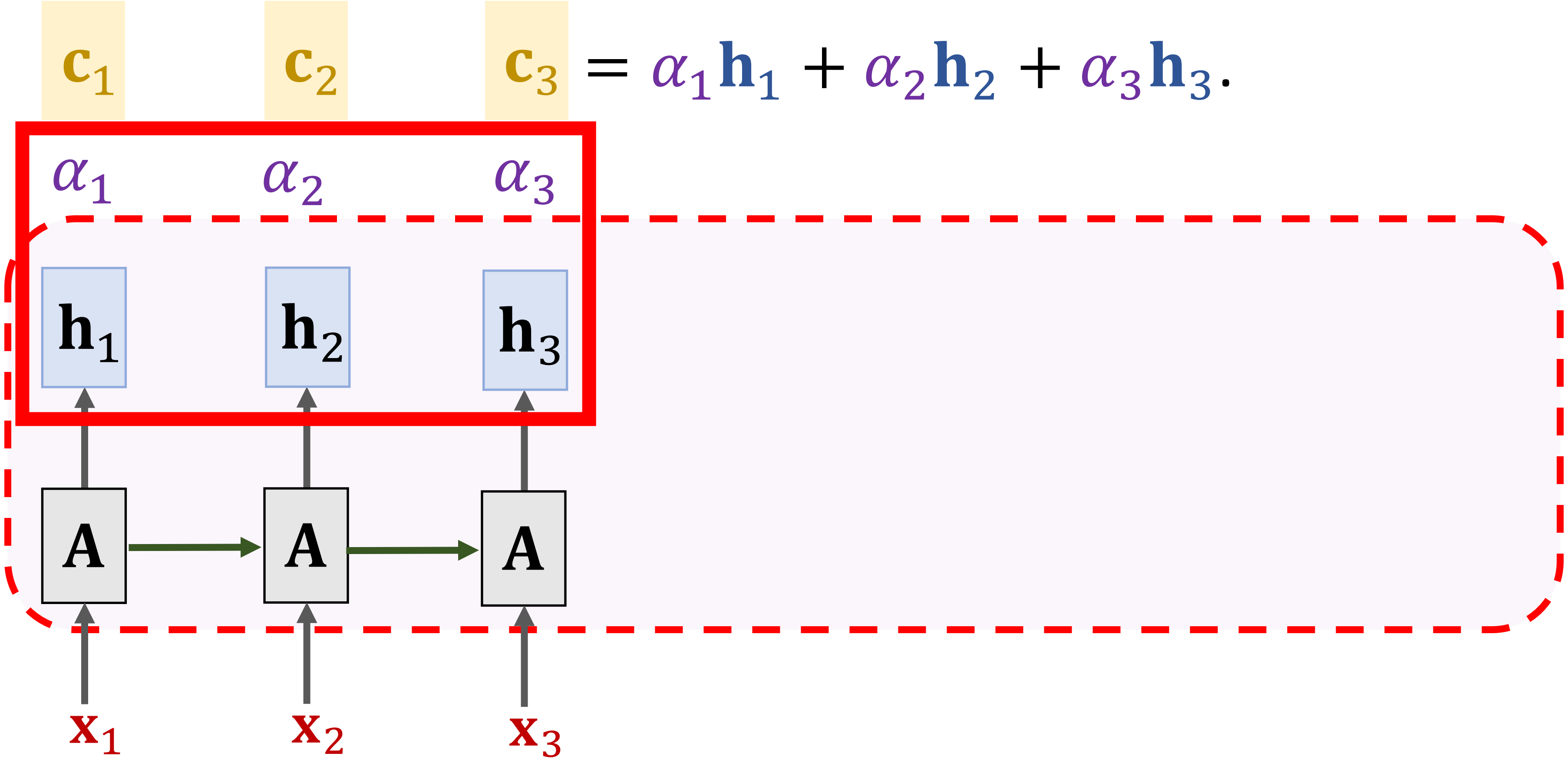
SimpleRNN + Self-Attention

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

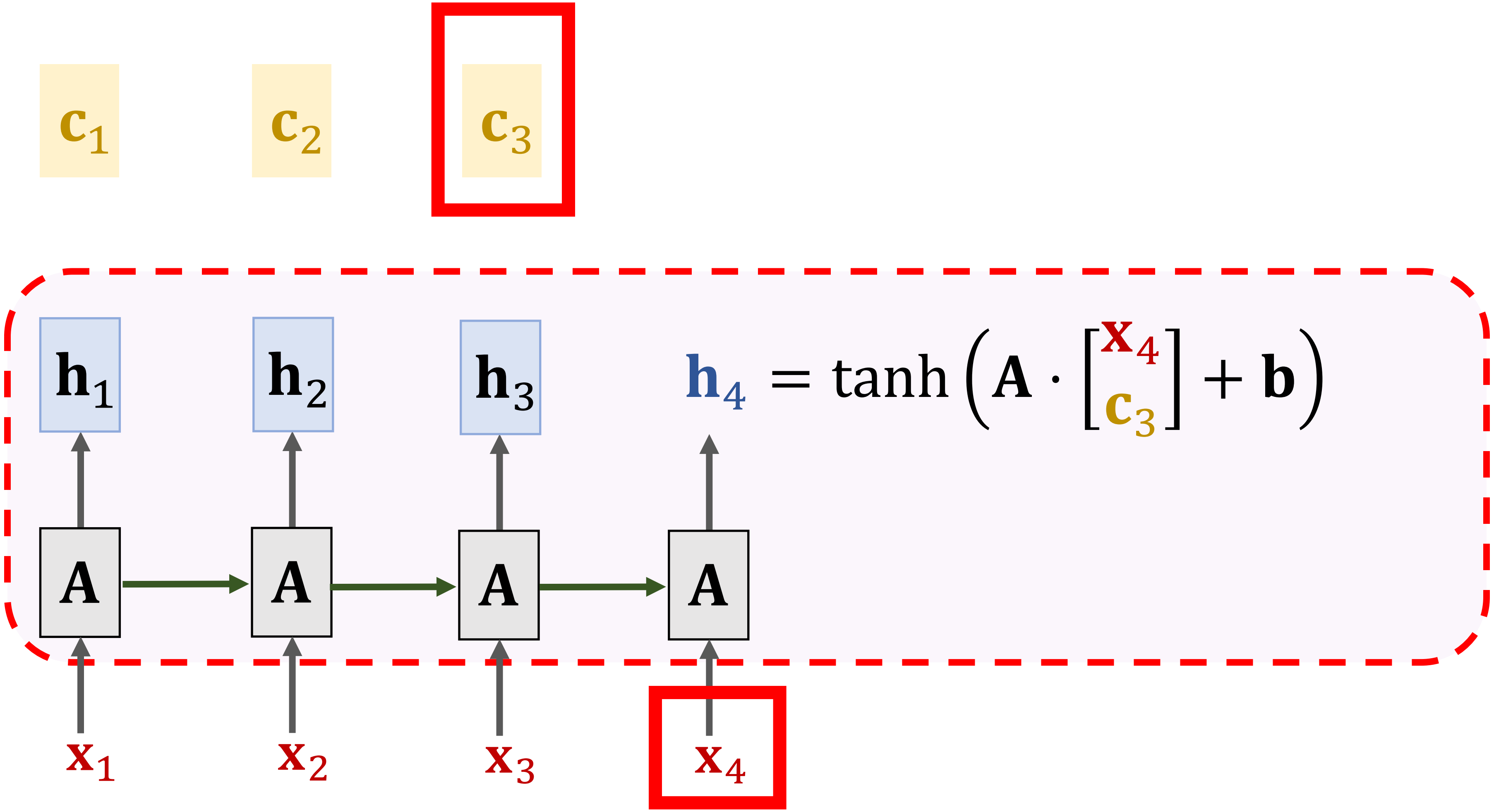


SimpleRNN + Self-Attention

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

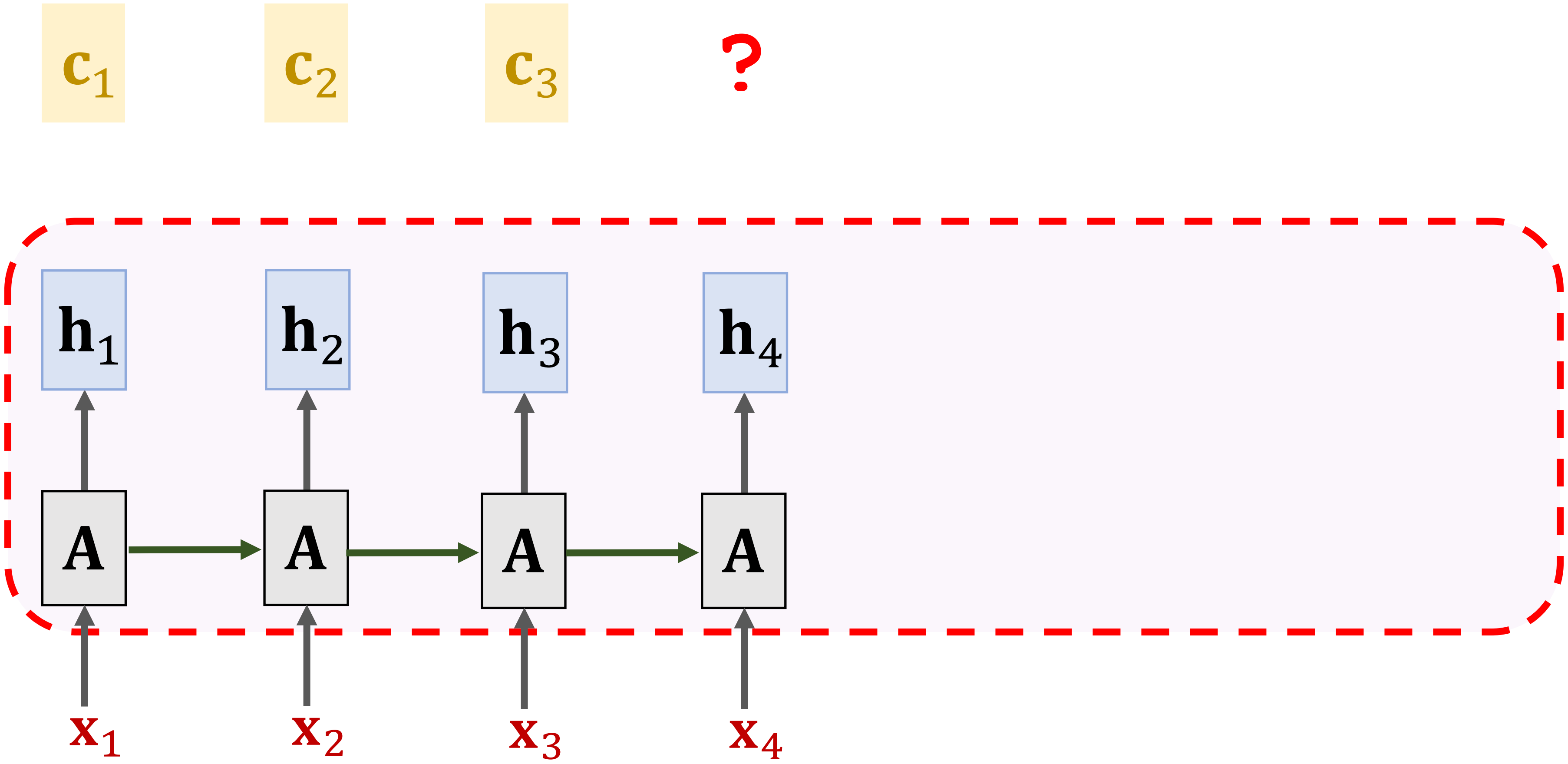


SimpleRNN + Self-Attention



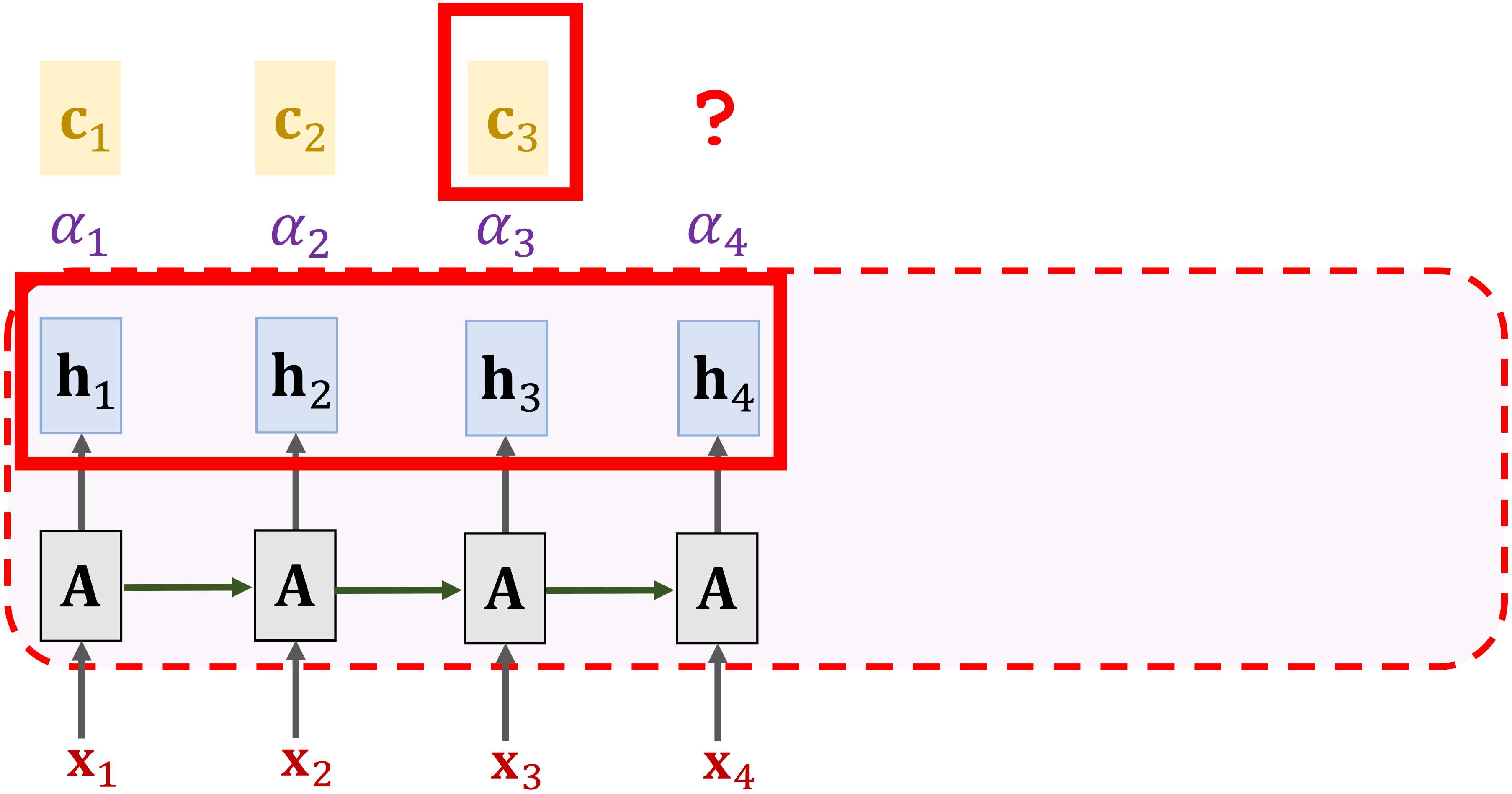
$$\mathbf{h}_4 = \tanh\left(\mathbf{A}\cdot\begin{bmatrix}\mathbf{x}_4\\ \mathbf{c}_3\end{bmatrix} + \mathbf{b}\right)$$

SimpleRNN + Self-Attention



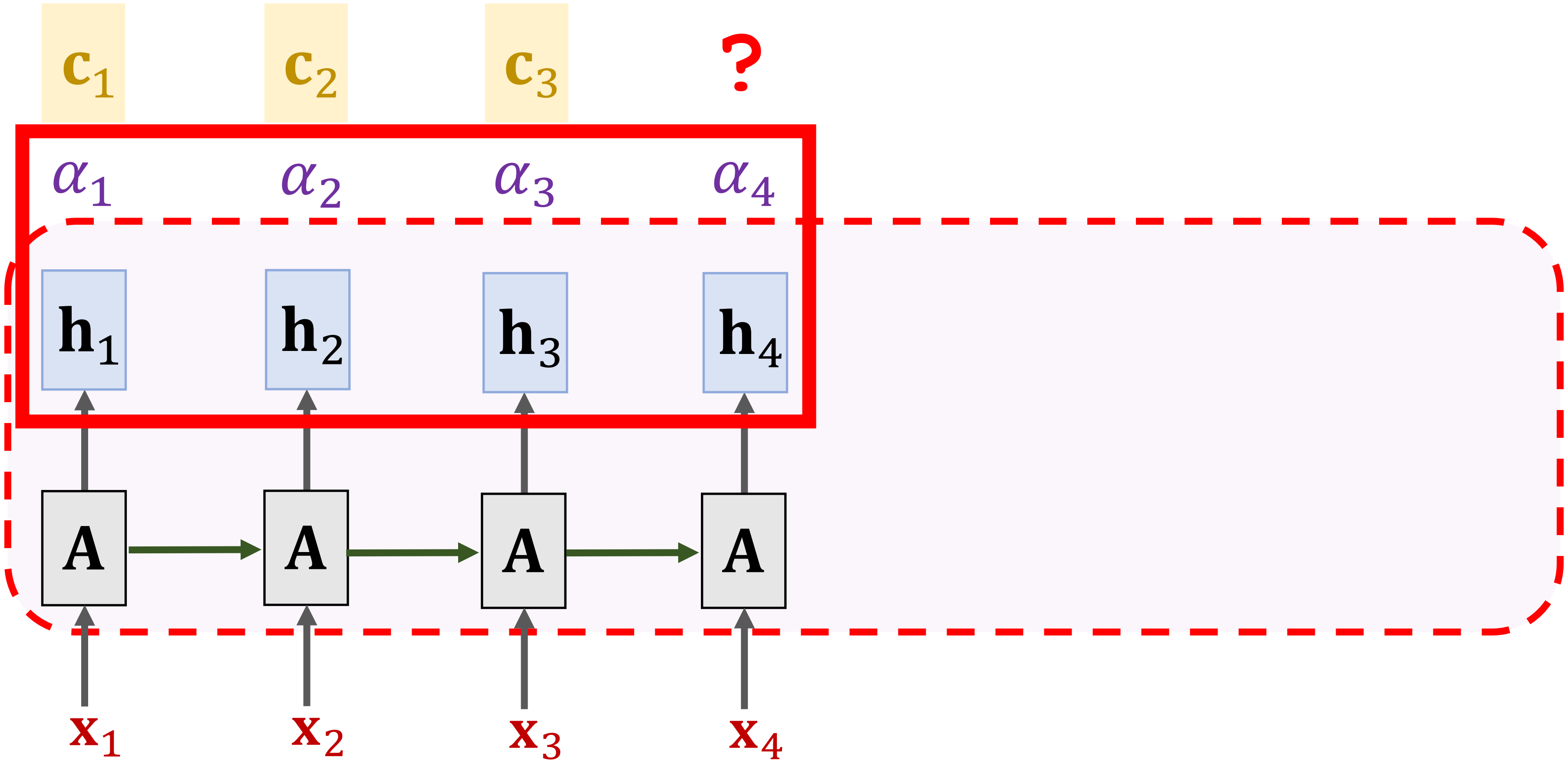
SimpleRNN + Self-Attention

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



SimpleRNN + Self-Attention

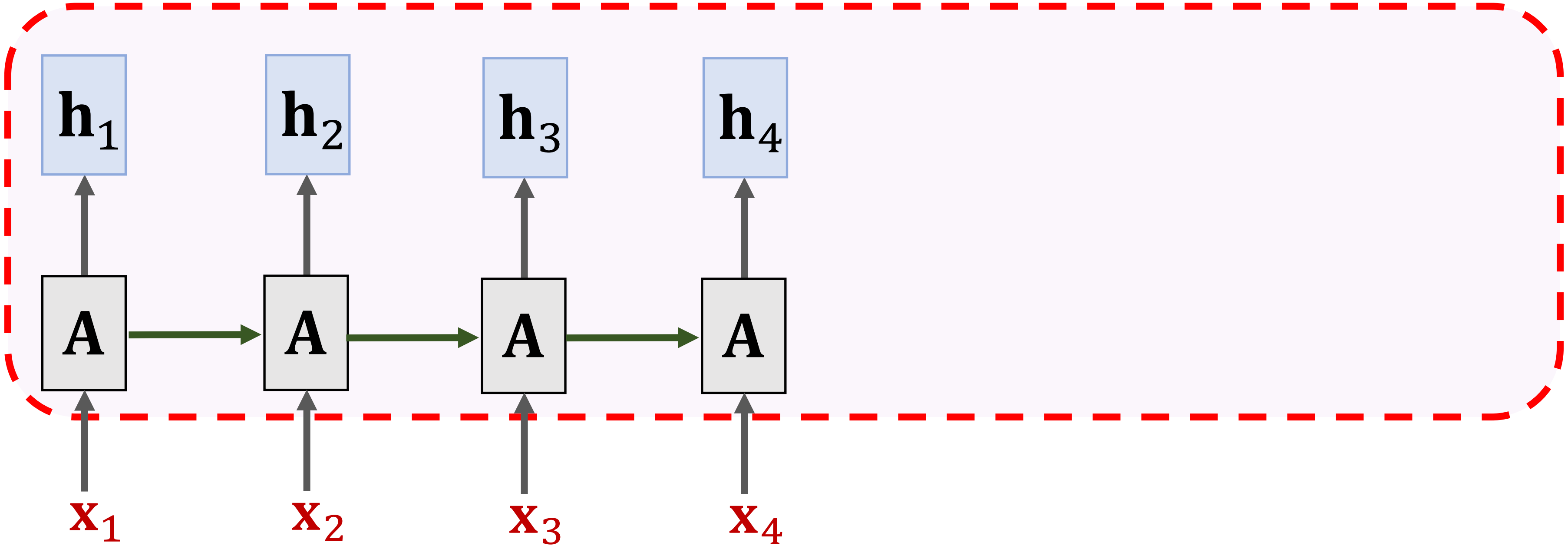
Weights: $\alpha_i = \text{align}(\mathbf{h}_i, \ \mathbf{c}_3)$.



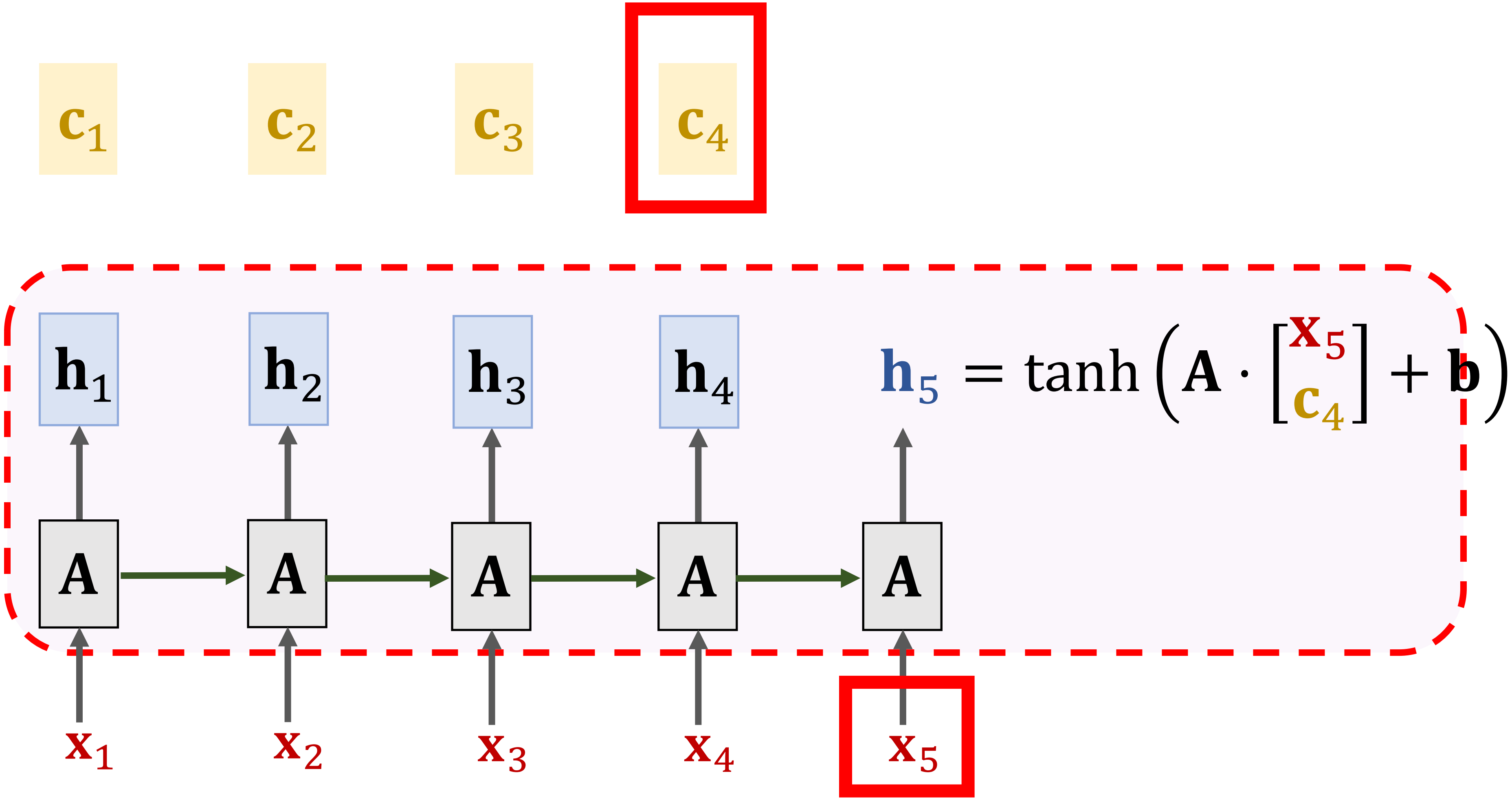
SimpleRNN + Self-Attention

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

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

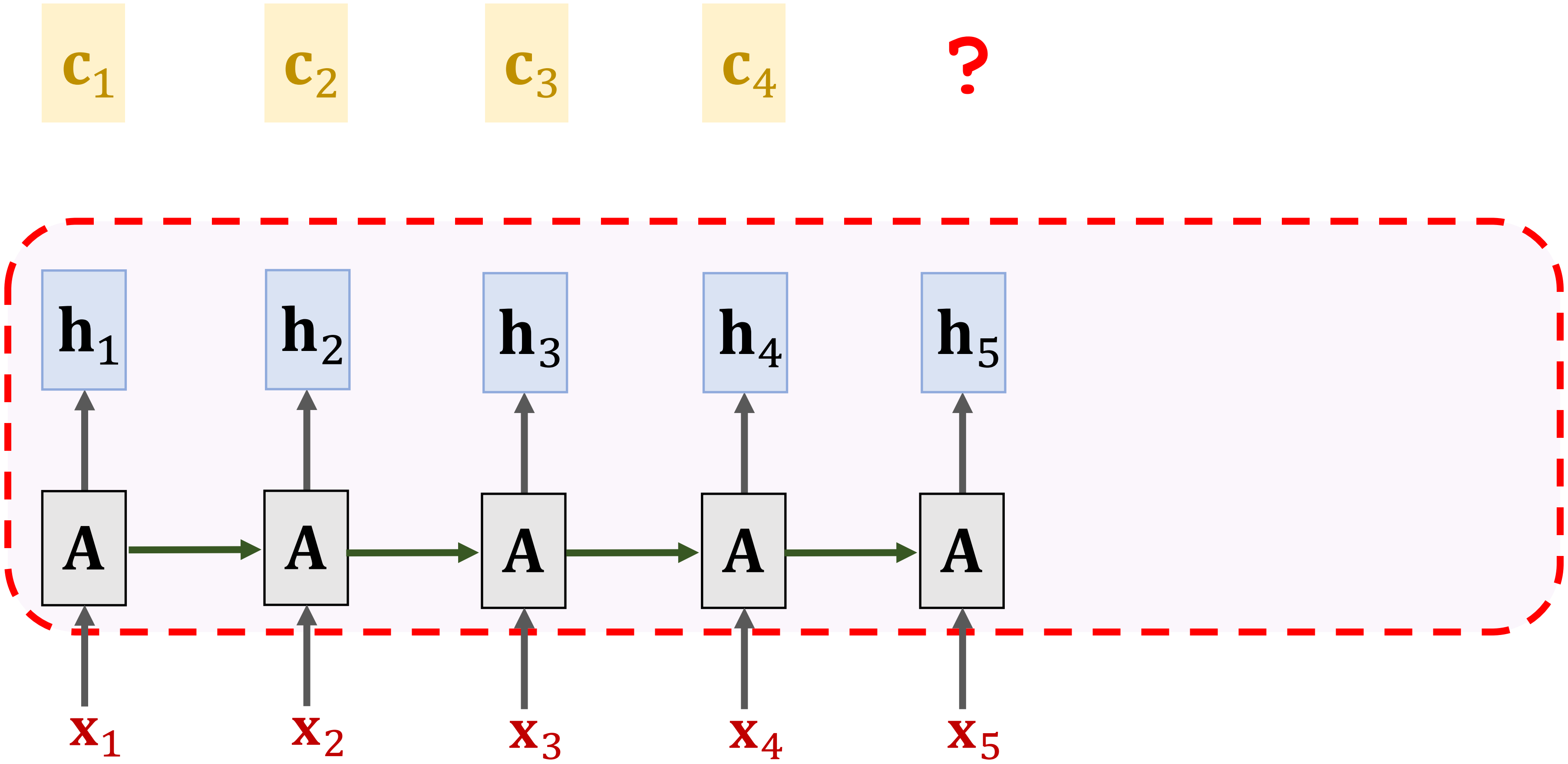


SimpleRNN + Self-Attention



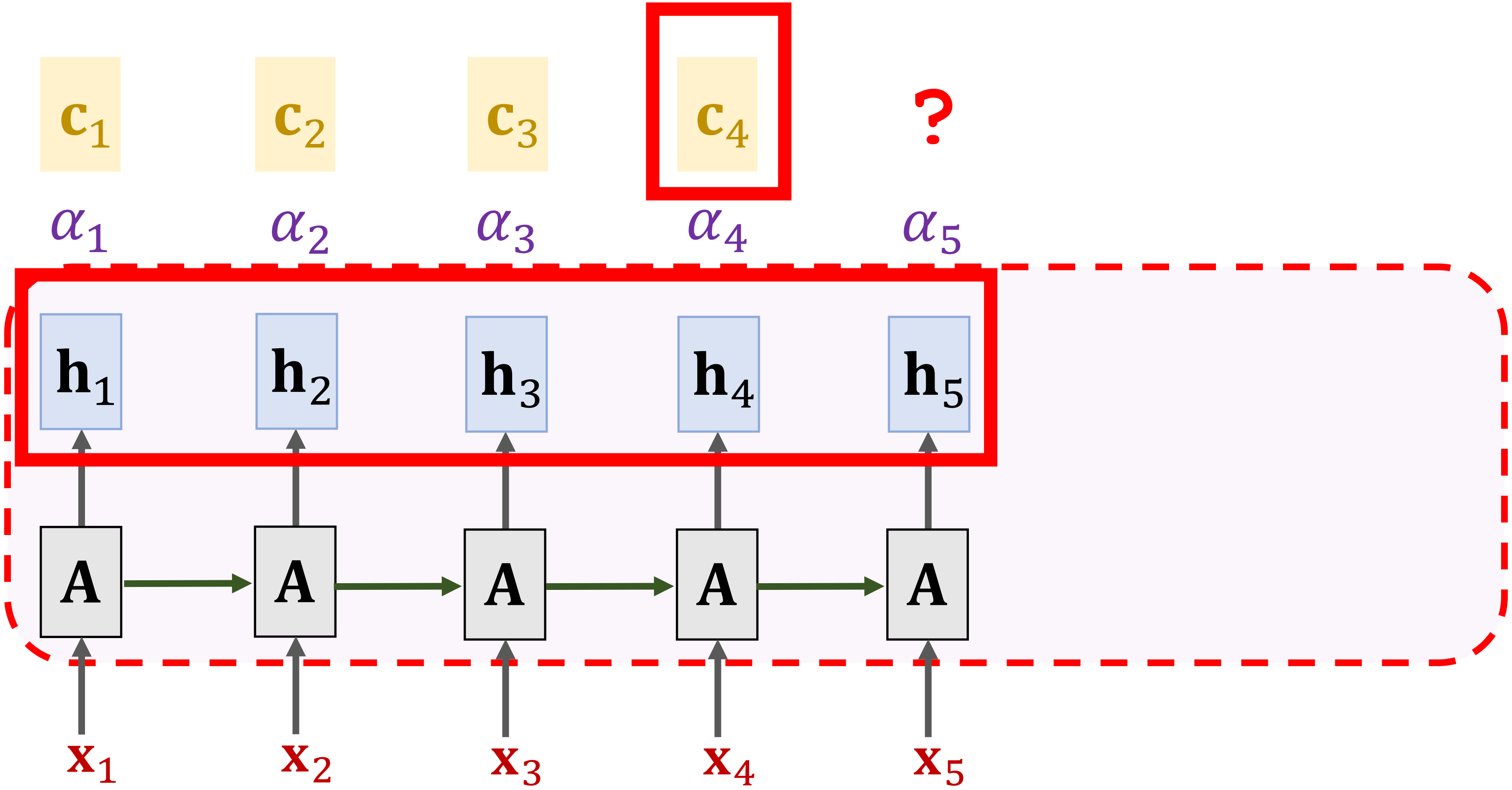
$$\mathbf{h}_5 = \tanh\left(\mathbf{A} \cdot \begin{bmatrix} \mathbf{x}_5 \\ \mathbf{c}_4 \end{bmatrix} + \mathbf{b}\right)$$

SimpleRNN + Self-Attention

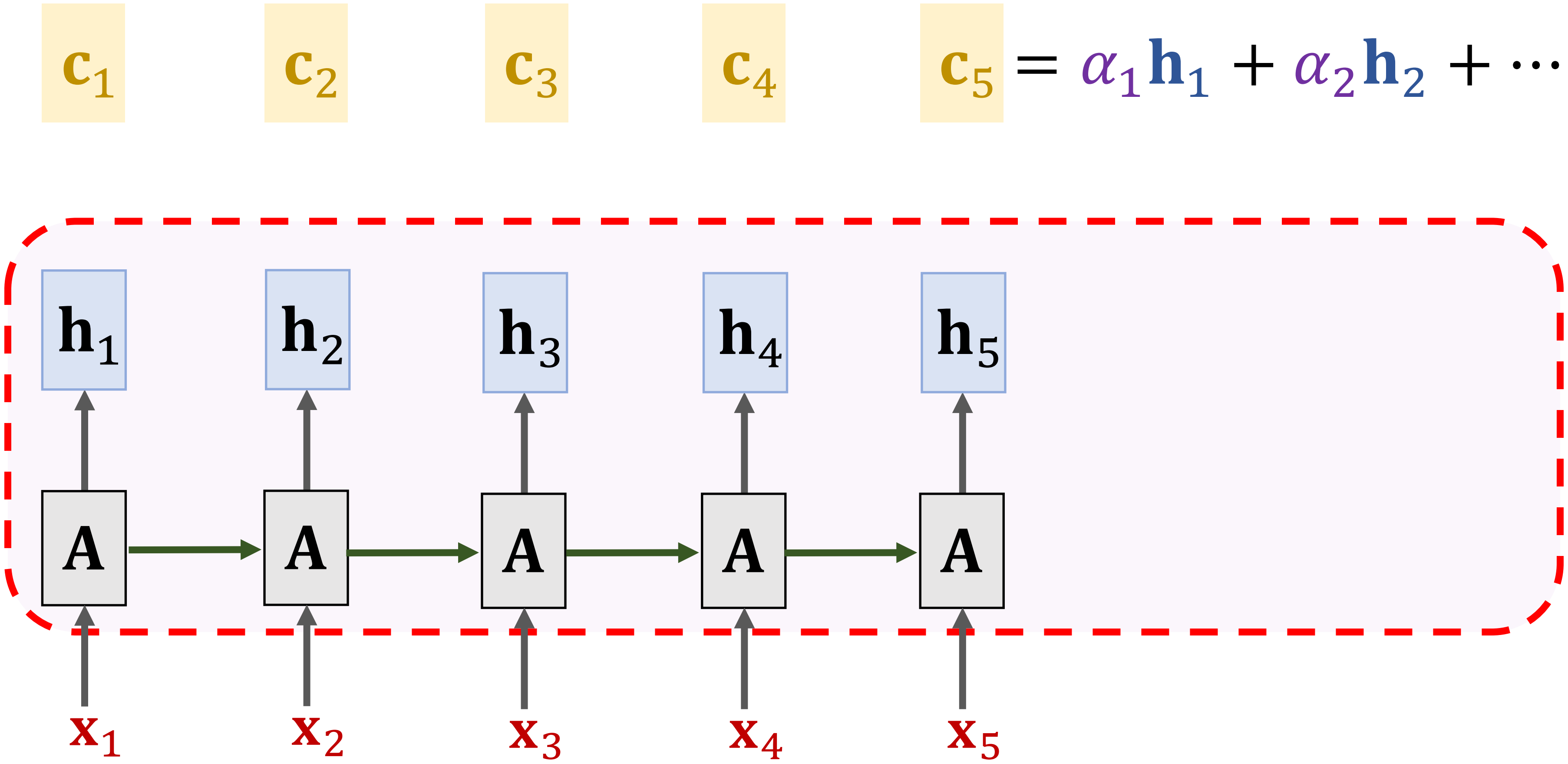


SimpleRNN + Self-Attention

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

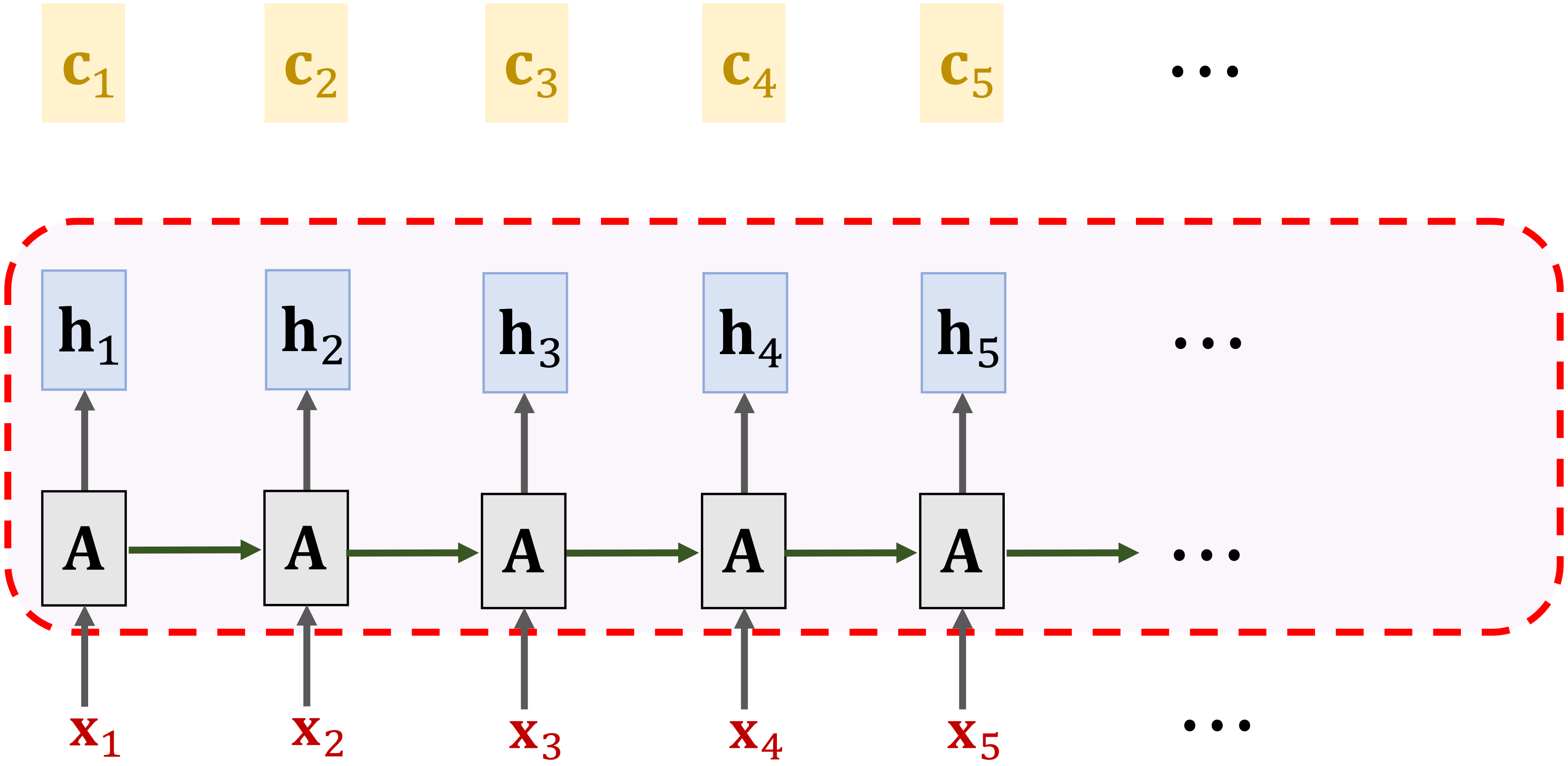


SimpleRNN + Self-Attention



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

SimpleRNN + Self-Attention

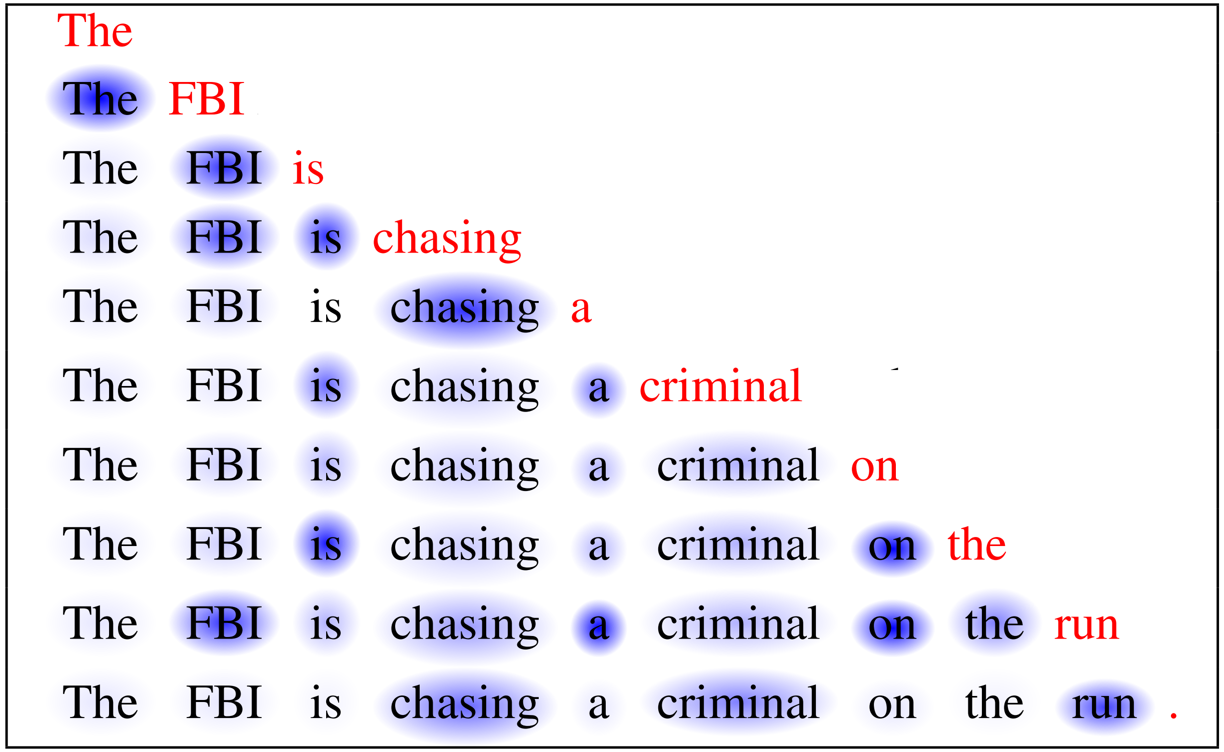


Summary

- 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.



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!