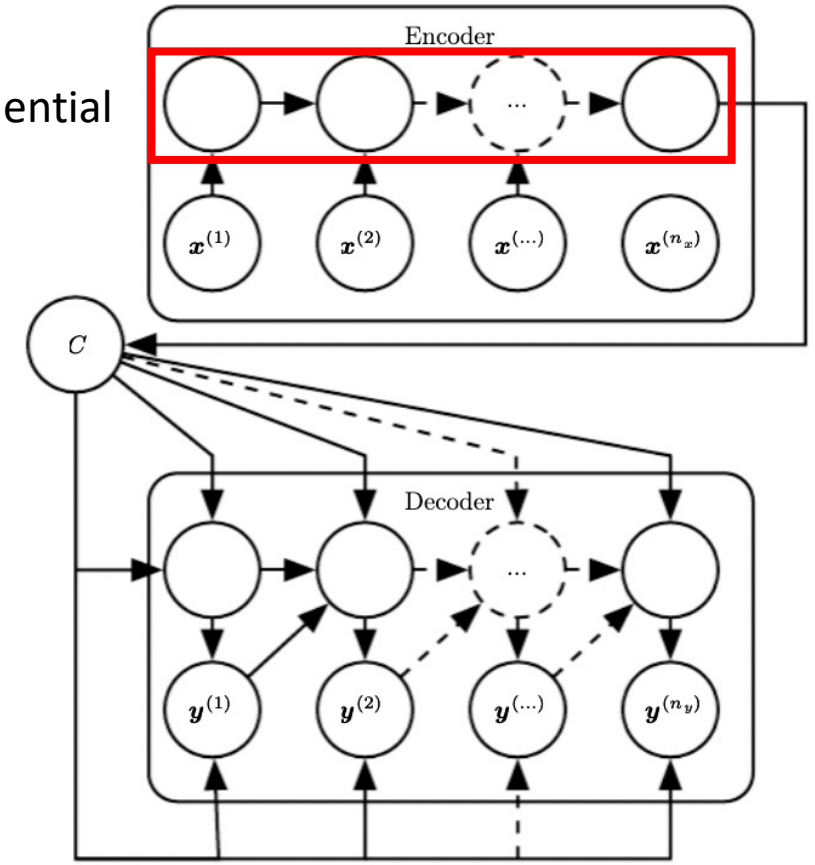
Attention

Neural Networks Design And Application

Encoder-decoder

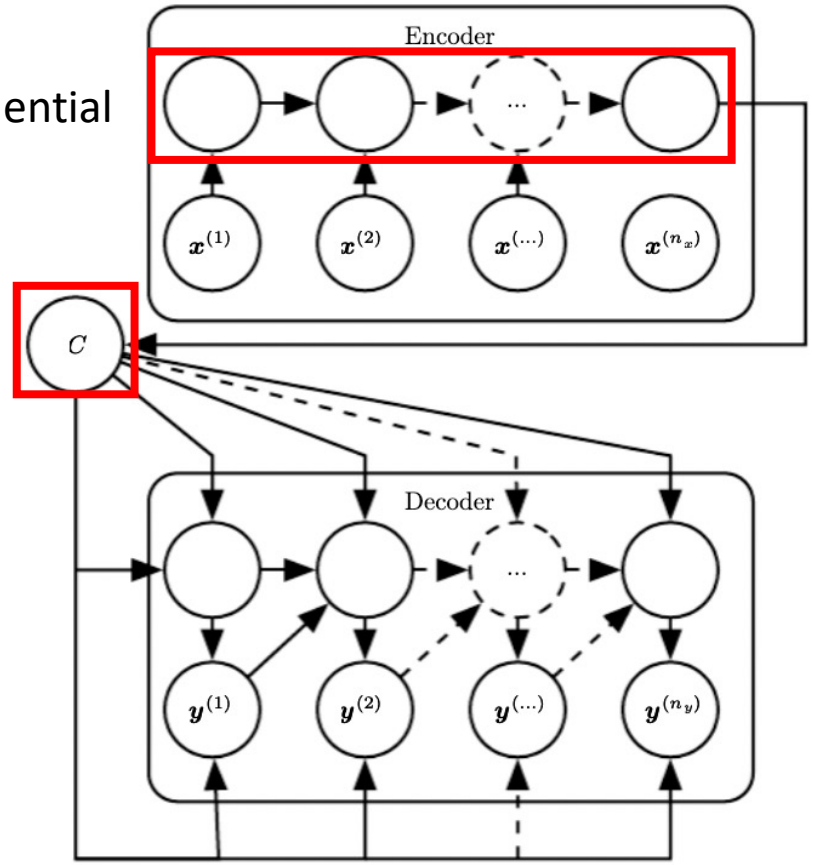
sequential



Encoder-decoder

sequential

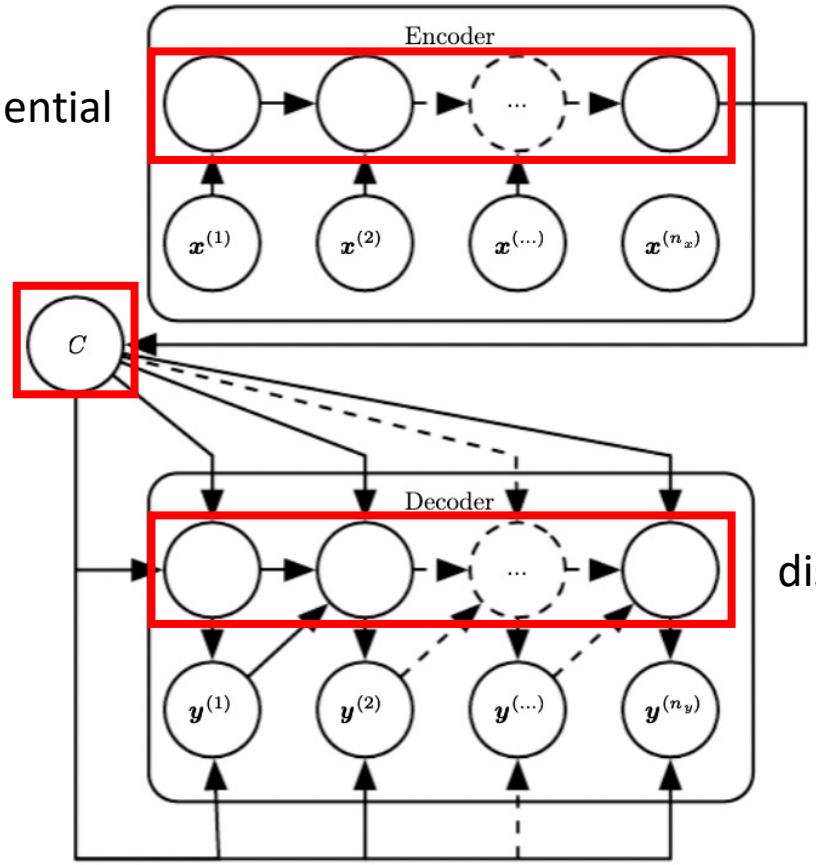
summary: context feature



Encoder-decoder

sequential

summary: context feature

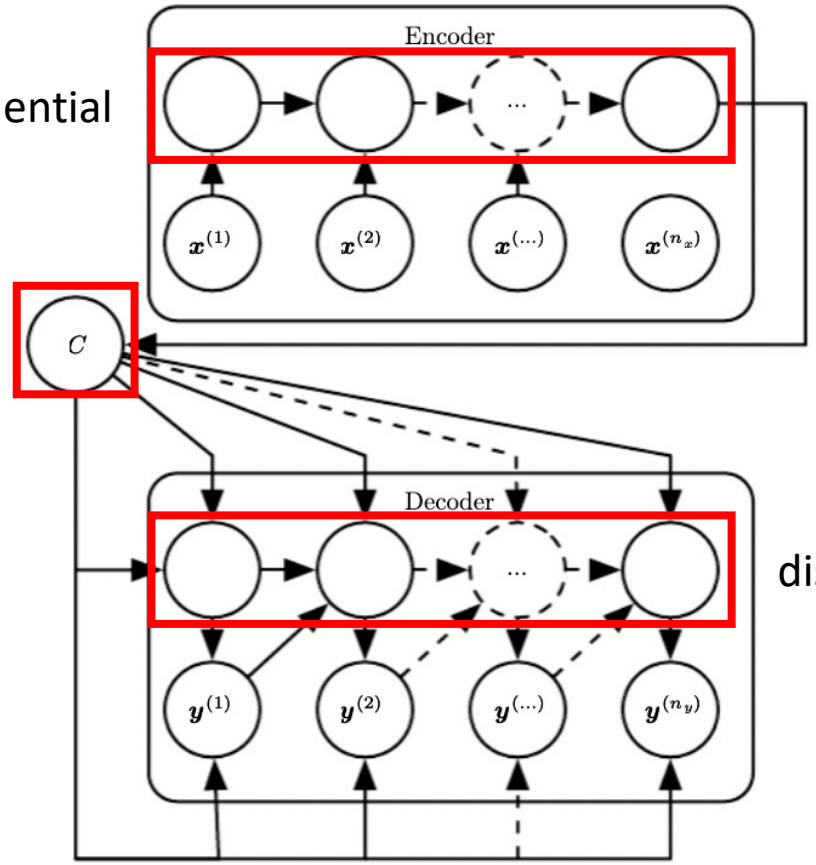


distribute information

Encoder-decoder

sequential

summary: context feature



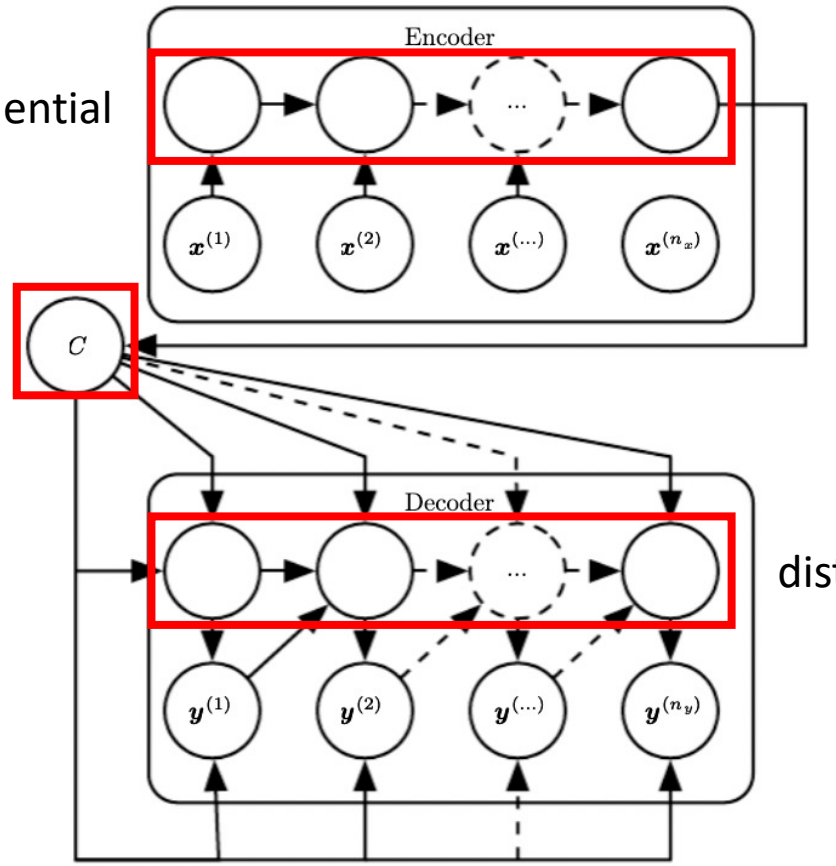
distribute information

Q: can we use bi-directional RNN for input?

Encoder-decoder

sequential

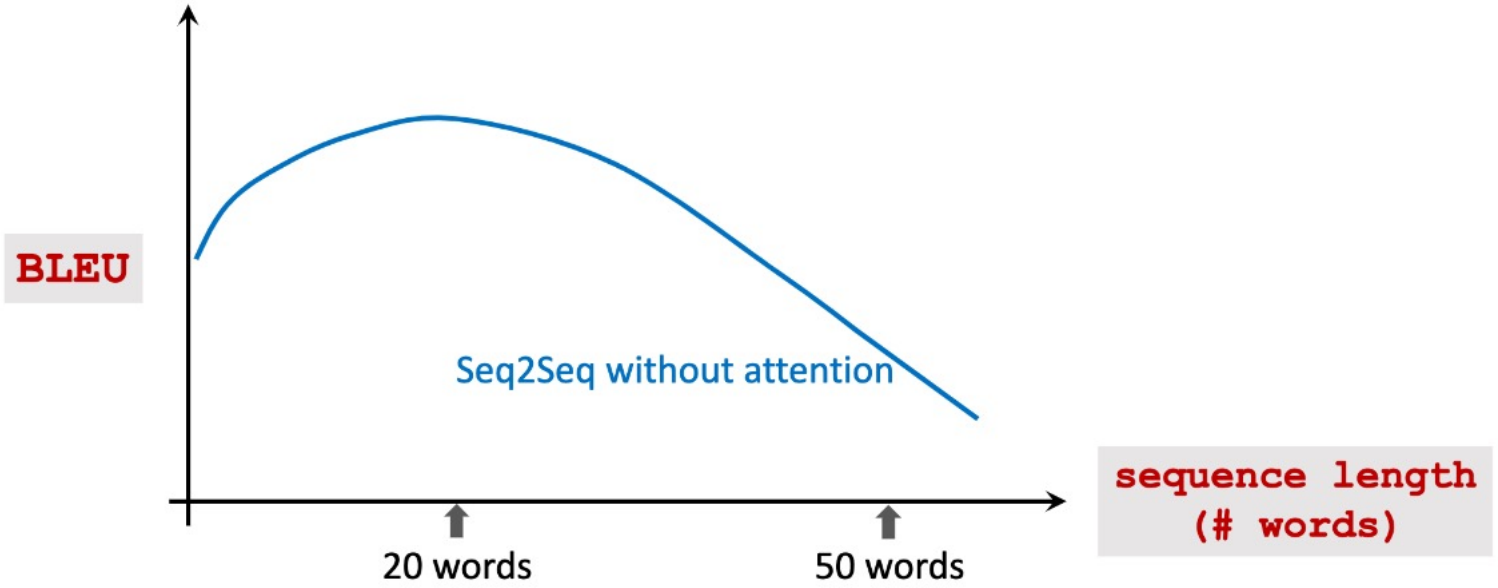
summary: context feature



distribute information

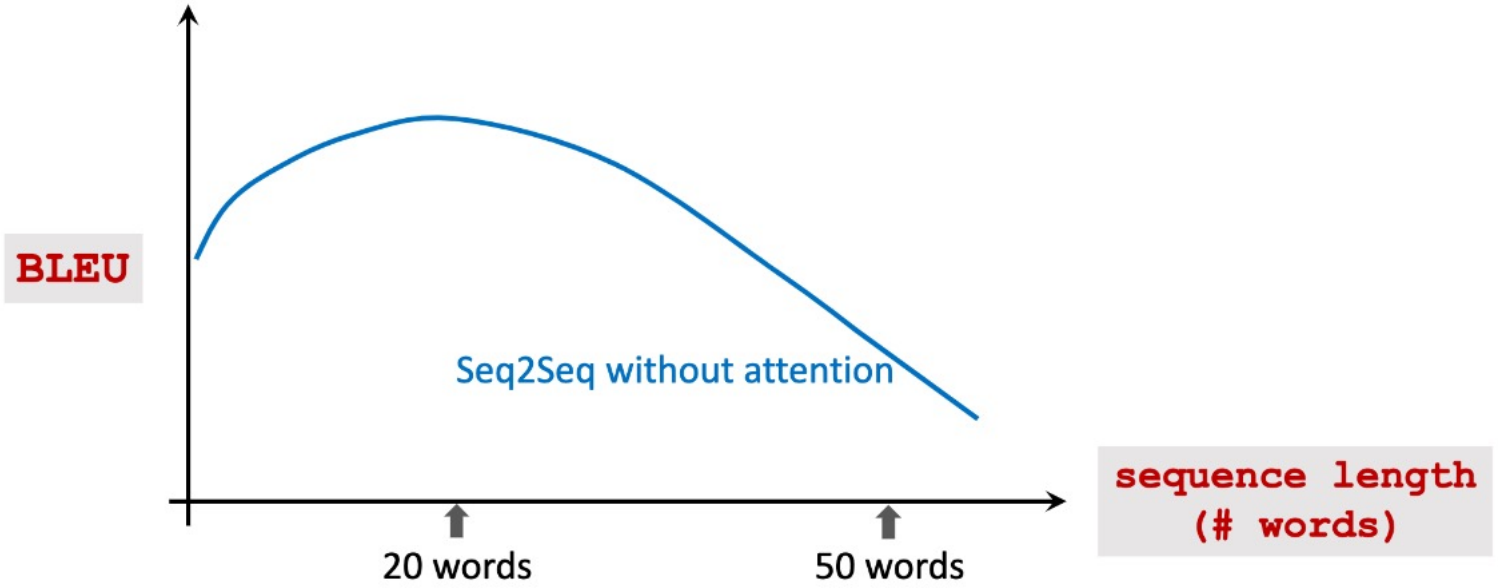
Q: can we use bi-directional RNN for input?

Seq2seq model performance



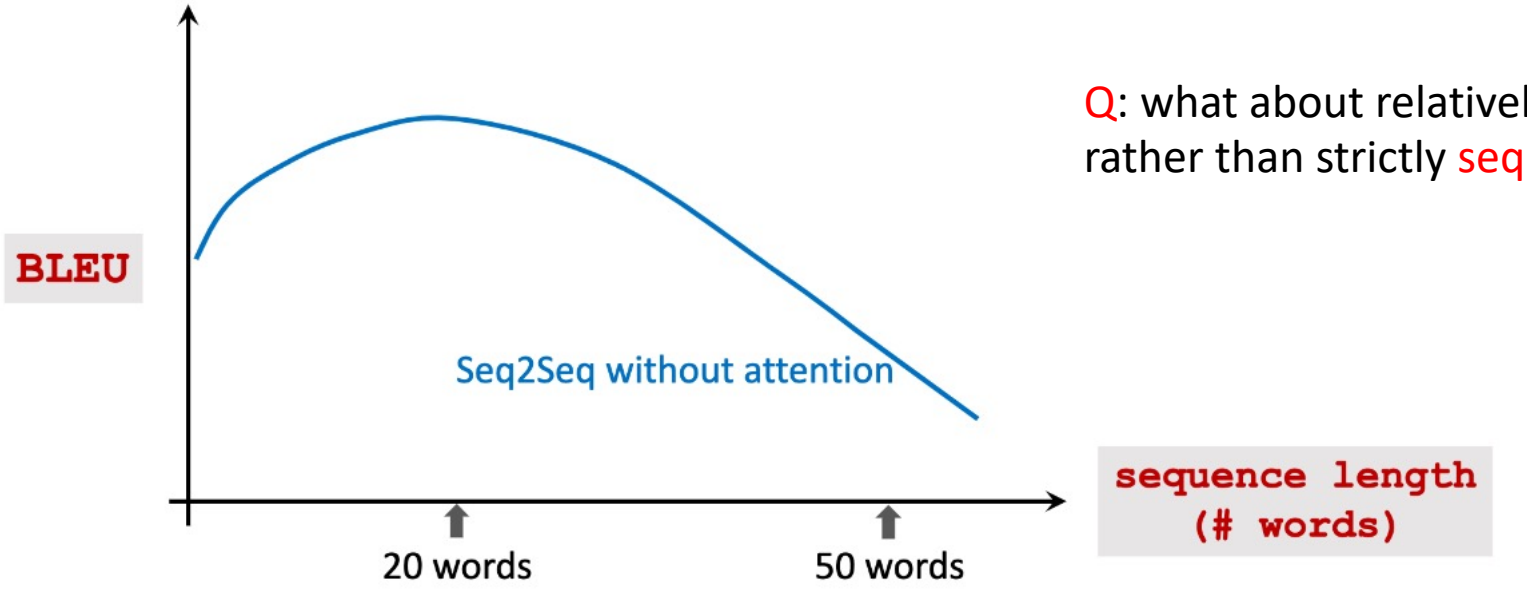
the clouds are in the sky

Seq2seq model performance



I like this town very much. I started my undergraduate study in 2020 and my major is computer science. I like programming and reading. I usually get up at 7AM and do some exercise. I also go fishing at weekend. I grew up in France. I spent my childhood outdoors. Whether it was riding my bicycle around my neighborhood pretending it was a motorcycle, making mud cakes, going on treasure hunts, making and selling perfume out of strong smelling flowers, or simply laying on the grass underneath the sun with a soccer ball waiting for someone to come out and play with me, the outdoors was where I spent my childhood and I cannot be more appreciative of it. I speak fluent *French*.

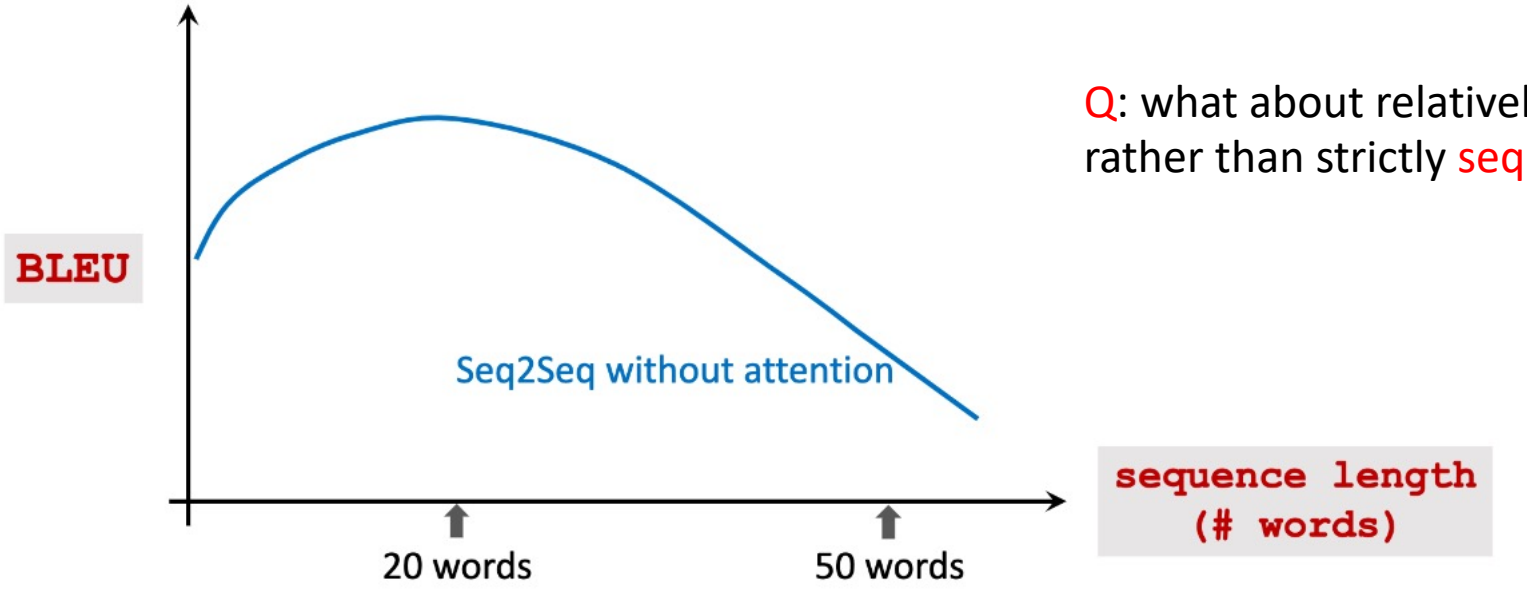
Seq2seq model performance



Q: what about relatively parallel correlation rather than strictly sequential correlation?

I like this town very much. I started my undergraduate study in 2020 and my major is computer science. I like programming and reading. I usually get up at 7AM and do some exercise. I also go fishing at weekend. I grew up in France. I spent my childhood outdoors. Whether it was riding my bicycle around my neighborhood pretending it was a motorcycle, making mud cakes, going on treasure hunts, making and selling perfume out of strong smelling flowers, or simply laying on the grass underneath the sun with a soccer ball waiting for someone to come out and play with me, the outdoors was where I spent my childhood and I cannot be more appreciative of it. I speak fluent *French*.

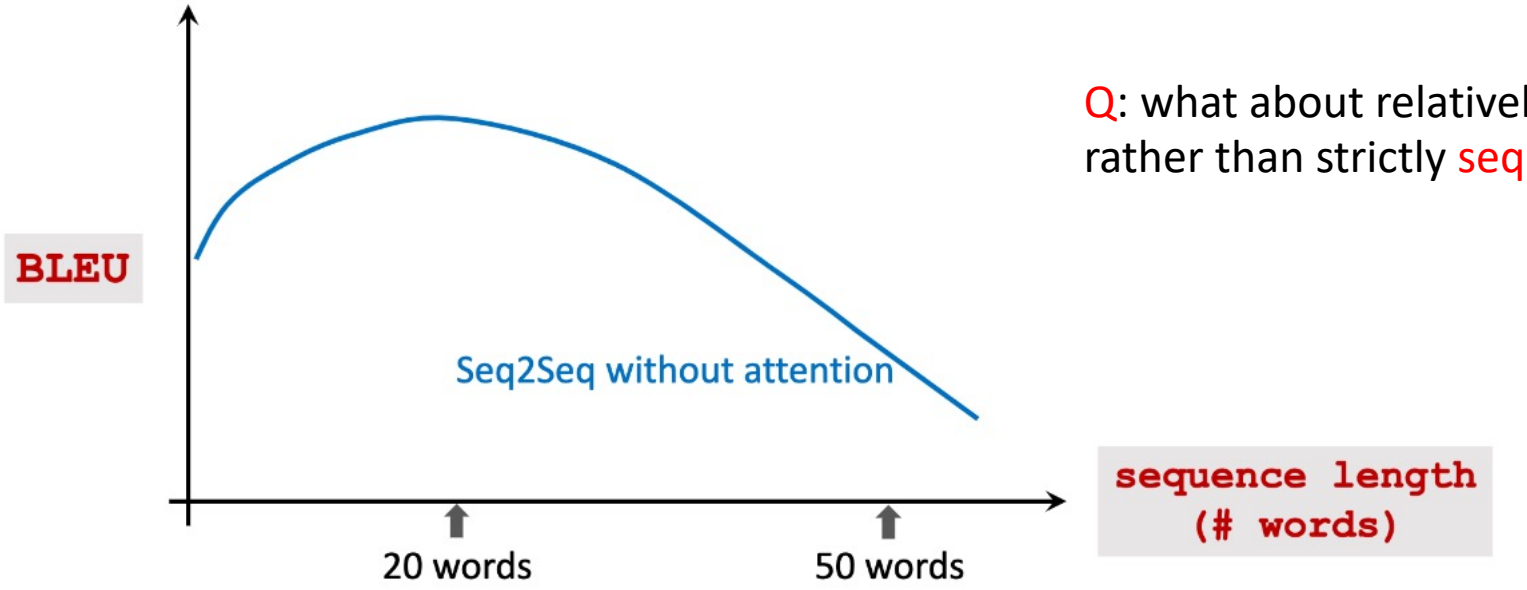
Seq2seq model performance



Q: what about relatively parallel correlation rather than strictly sequential correlation?

I like this town very much. I started my undergraduate study in 2020 and my major is computer science. I like programming and reading. I usually get up at 7AM and do some exercise. I also go fishing at weekend. I grew up in France. I spent my childhood outdoors. Whether it was riding my bicycle around my neighborhood pretending it was a motorcycle, making mud cakes, going on treasure hunts, making and selling perfume out of strong smelling flowers, or simply laying on the grass underneath the sun with a soccer ball waiting for someone to come out and play with me, the outdoors was where I spent my childhood and I cannot be more appreciative of it. I speak fluent *French*.

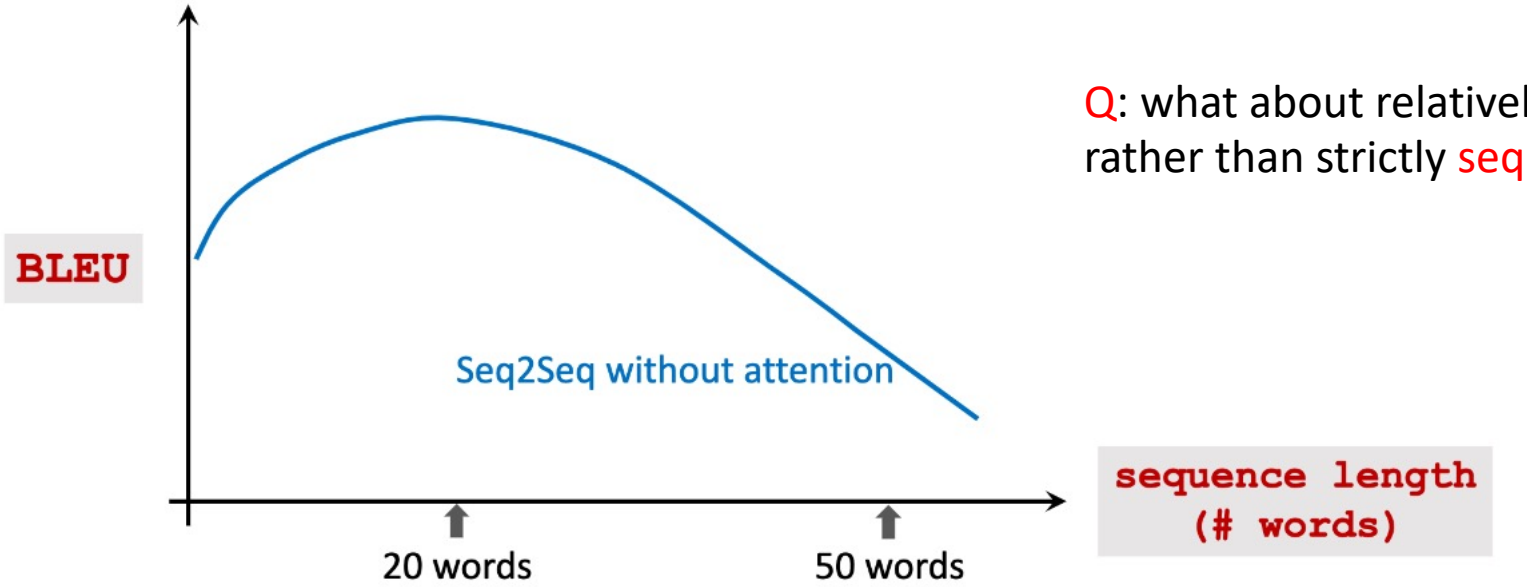
Seq2seq model performance



Q: what about relatively parallel correlation rather than strictly sequential correlation?

I like this town very much. I started my undergraduate study in 2020 and my major is computer science. I like programming and reading. I usually get up at 7AM and do some exercise. I also go fishing at weekend. I grew up in France. I spent my childhood outdoors. Whether it was riding my bicycle around my neighborhood pretending it was a motorcycle, making mud cakes, going on treasure hunts, making and selling perfume out of strong smelling flowers, or simply laying on the grass underneath the sun with a soccer ball waiting for someone to come out and play with me, the outdoors was where I spent my childhood and I cannot be more appreciative of it. I speak fluent *French*.

Seq2seq model performance

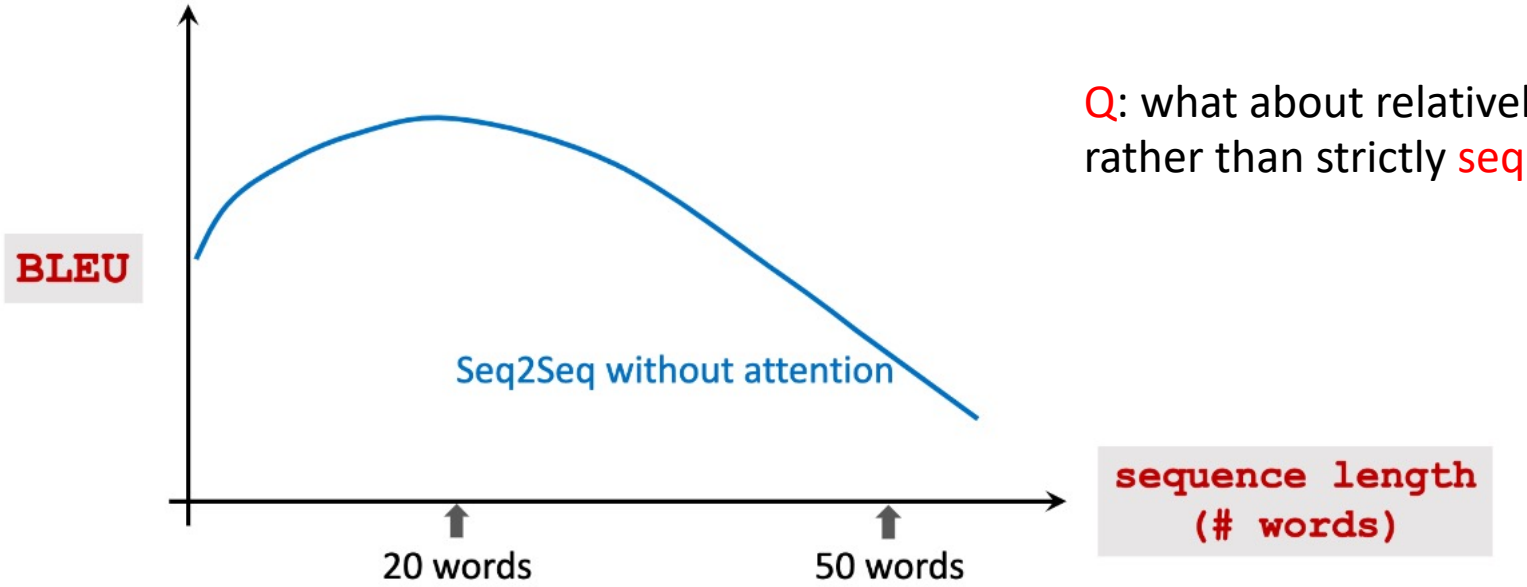


Q: what about relatively parallel correlation rather than strictly sequential correlation?

We need to summarize all context

I like this town very much. I started my undergraduate study in 2020 and my major is computer science. I like programming and reading. I usually get up at 7AM and do some exercise. I also go fishing at weekend. I grew up in France. I spent my childhood outdoors. Whether it was riding my bicycle around my neighborhood pretending it was a motorcycle, making mud cakes, going on treasure hunts, making and selling perfume out of strong smelling flowers, or simply laying on the grass underneath the sun with a soccer ball waiting for someone to come out and play with me, the outdoors was where I spent my childhood and I cannot be more appreciative of it. I speak fluent *French*.

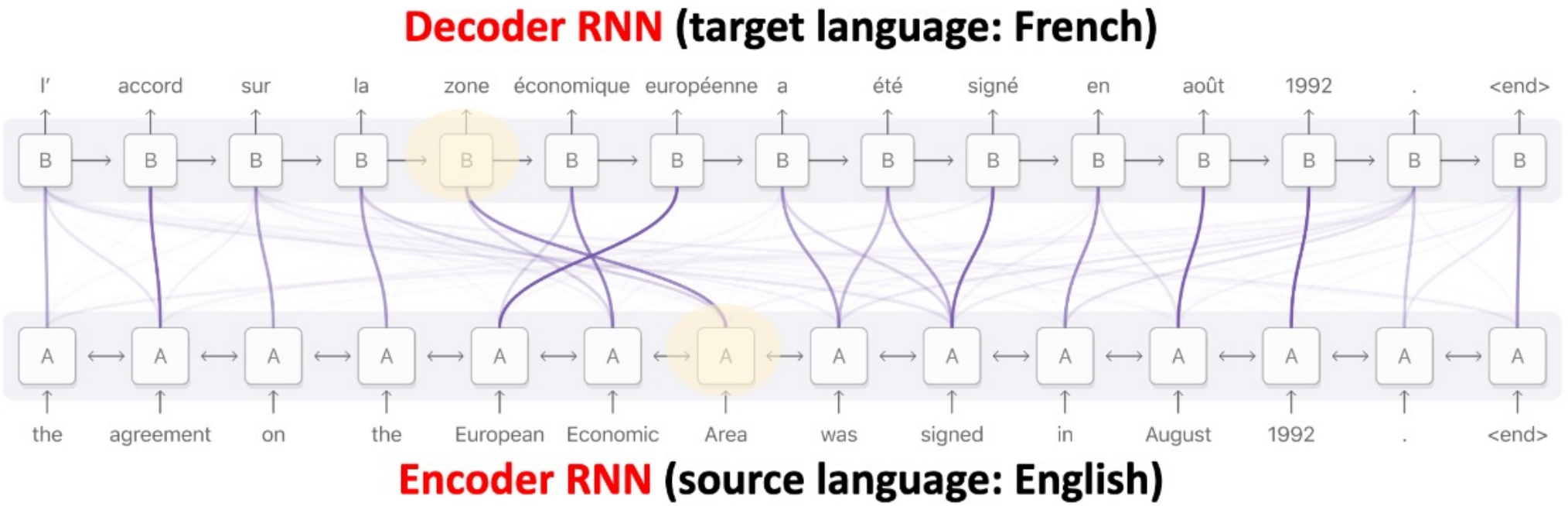
Seq2seq model performance



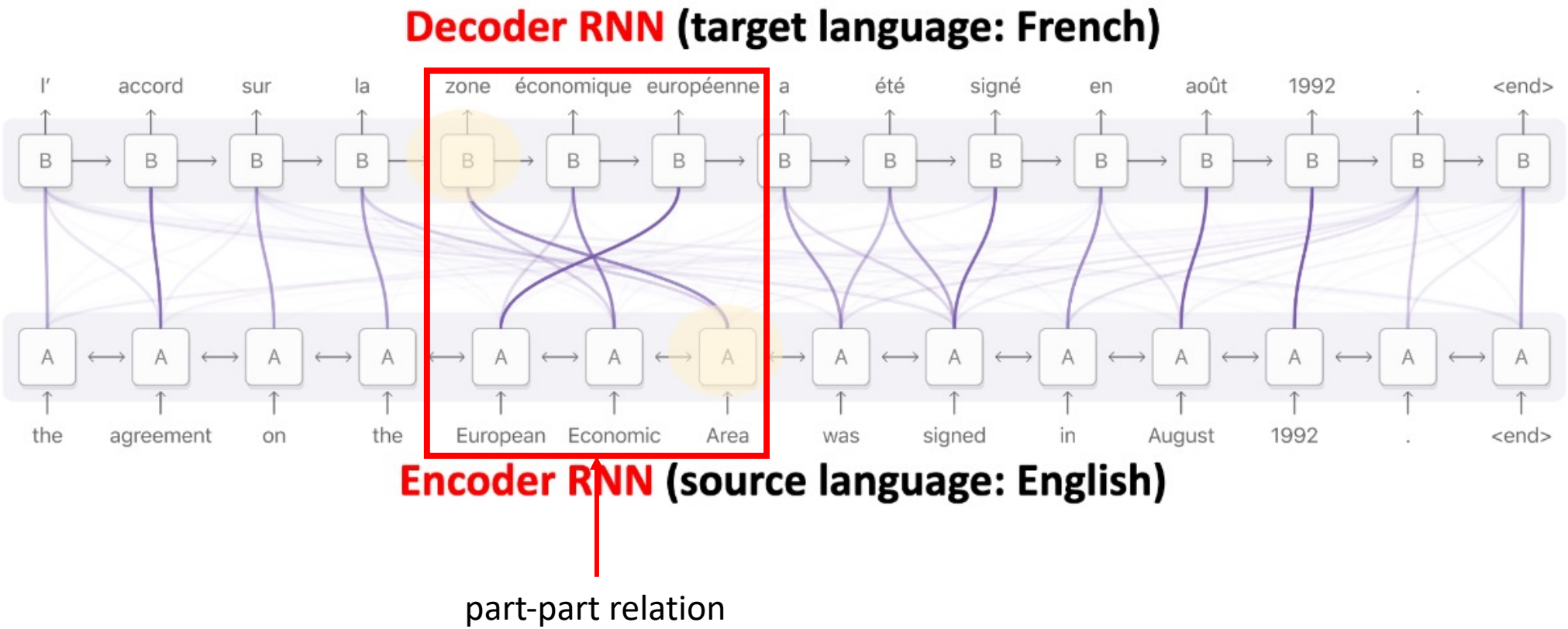
Q: what about relatively parallel correlation rather than strictly sequential correlation?

I like this town very much. I started my undergraduate study in 2020 and my major is computer science. I like programming and reading. I usually get up at 7AM and do some exercise. I also go fishing at weekend. I grew up in France. I spent my childhood outdoors. Whether it was riding my bicycle around my neighborhood pretending it was a motorcycle, making mud cakes, going on treasure hunts, making and selling perfume out of strong smelling flowers, or simply laying on the grass underneath the sun with a soccer ball waiting for someone to come out and play with me, the outdoors was where I spent my childhood and I cannot be more appreciative of it. I speak fluent *French*.

Input-output correlation

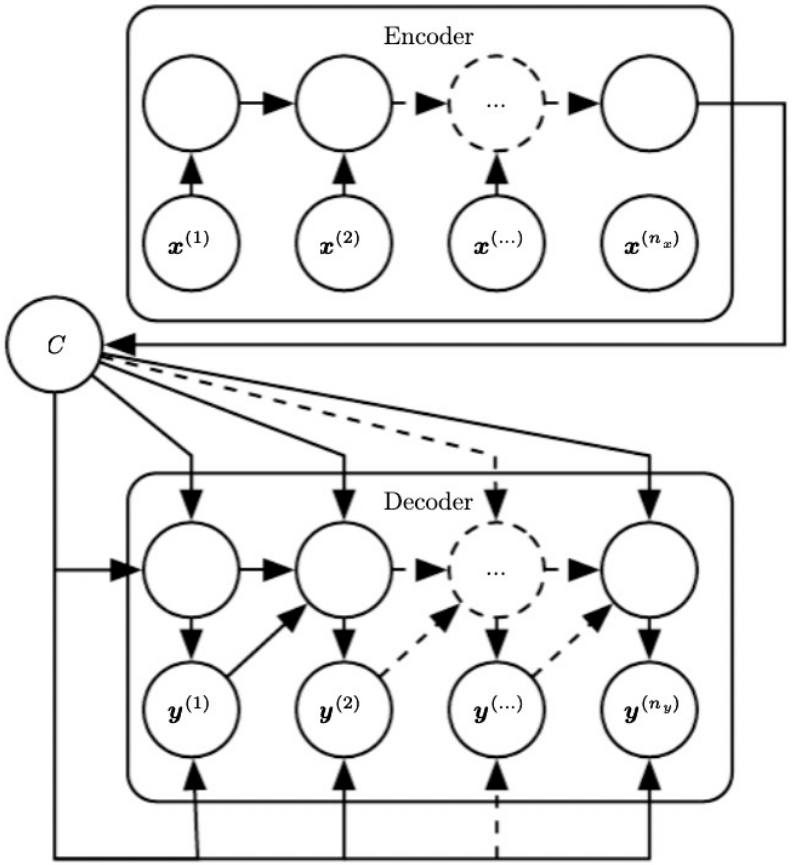


Input-output correlation



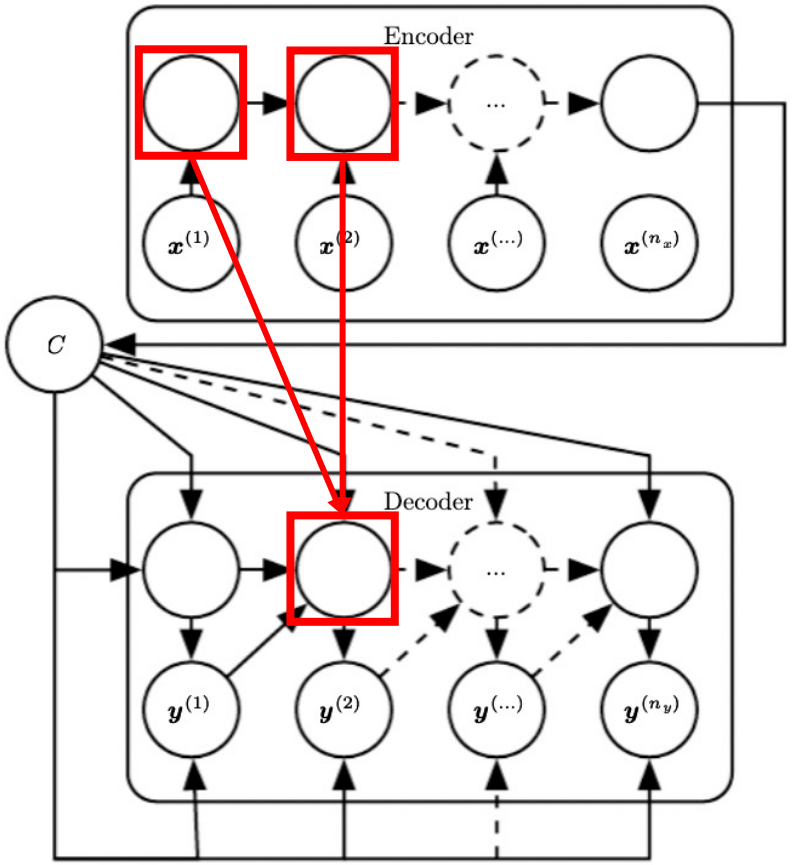
Encoder-decoder

Q: can we create information flow between encoder and decoder nodes?



Encoder-decoder

Q: can we create information flow between encoder and decoder nodes?



Attention mechanism

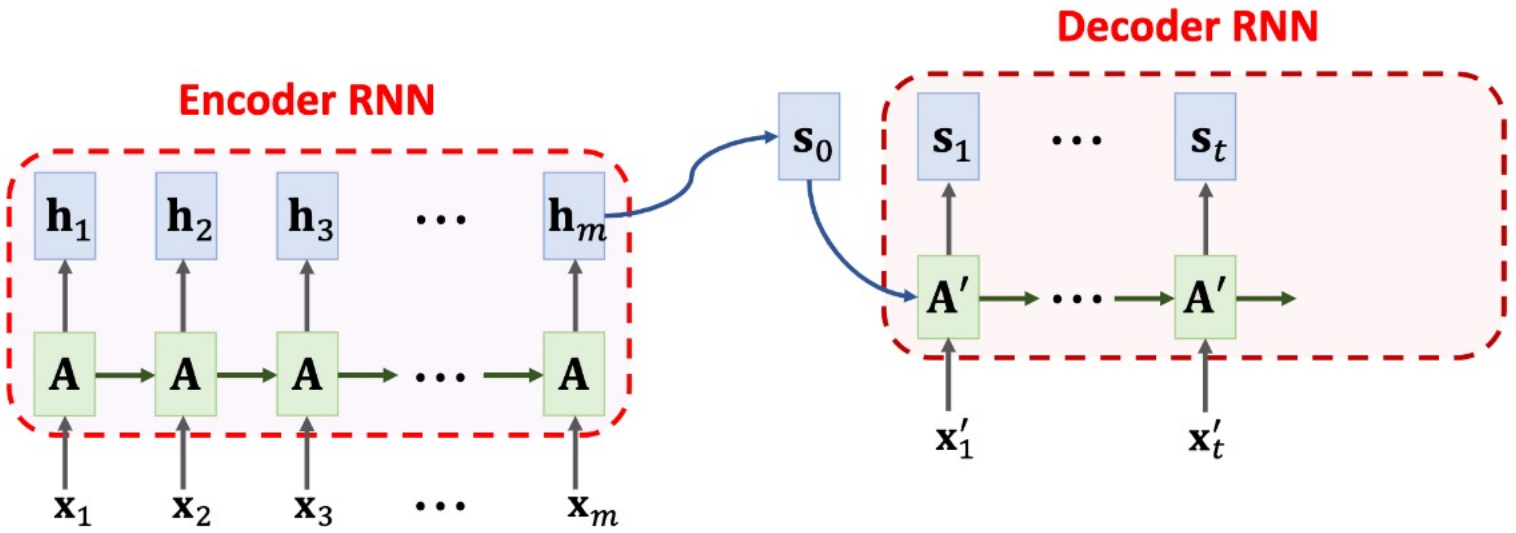


Image from https://github.com/wangshusen/DeepLearning/blob/master/Slides/9_RNN_8.pdf

Attention mechanism

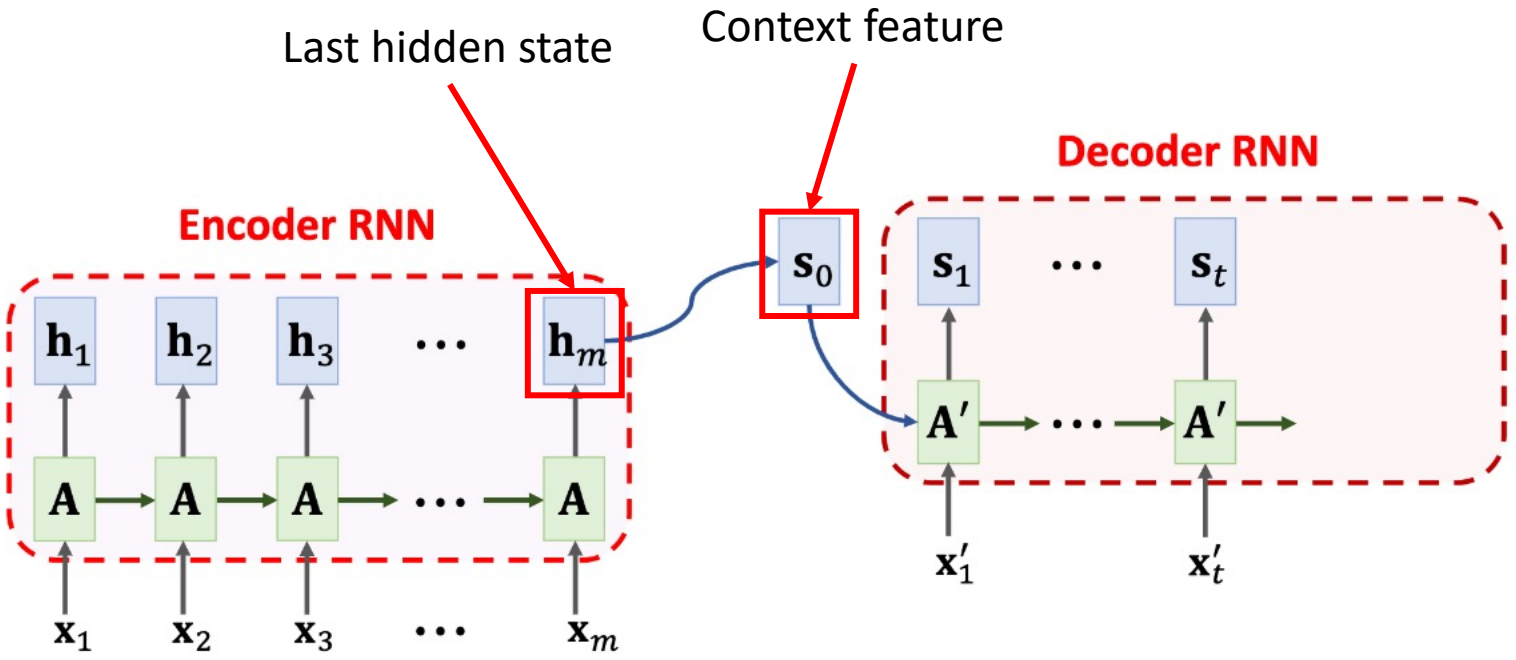


Image from https://github.com/wangshusen/DeepLearning/blob/master/Slides/9_RNN_8.pdf

Attention mechanism

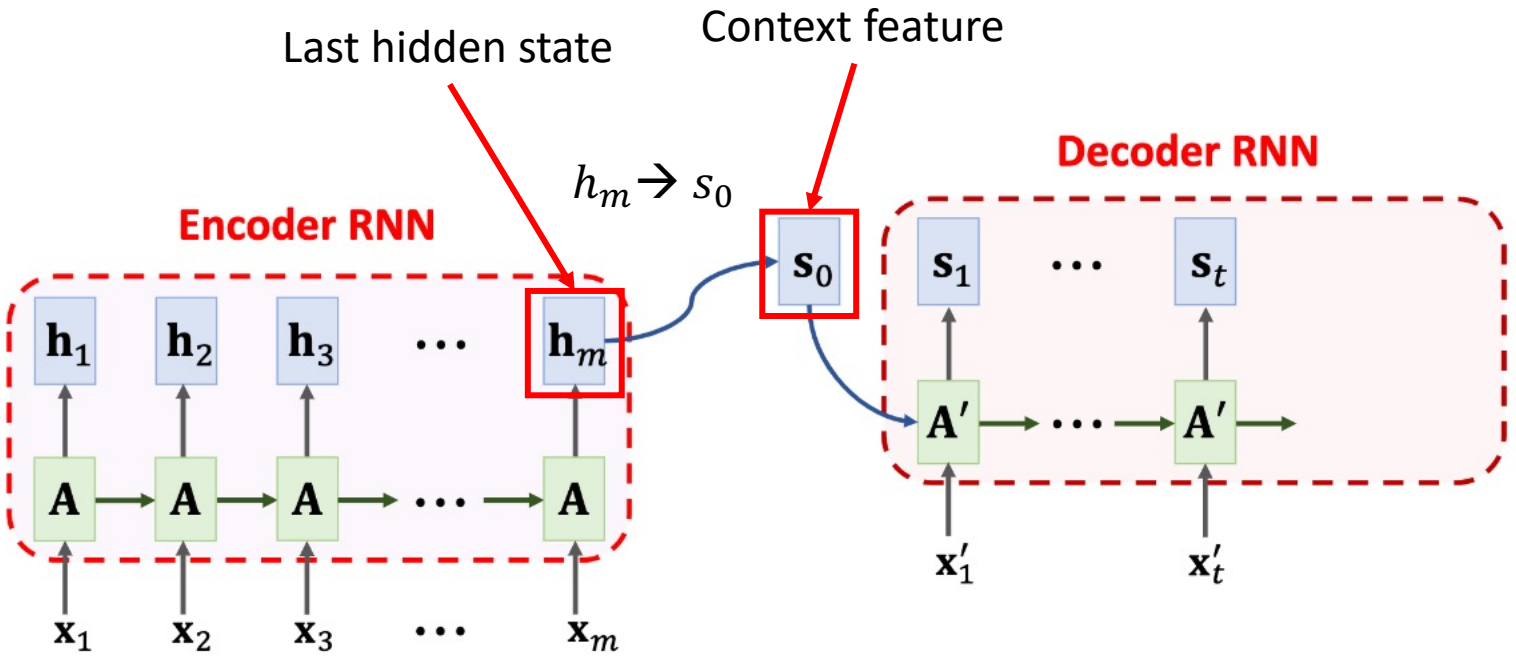
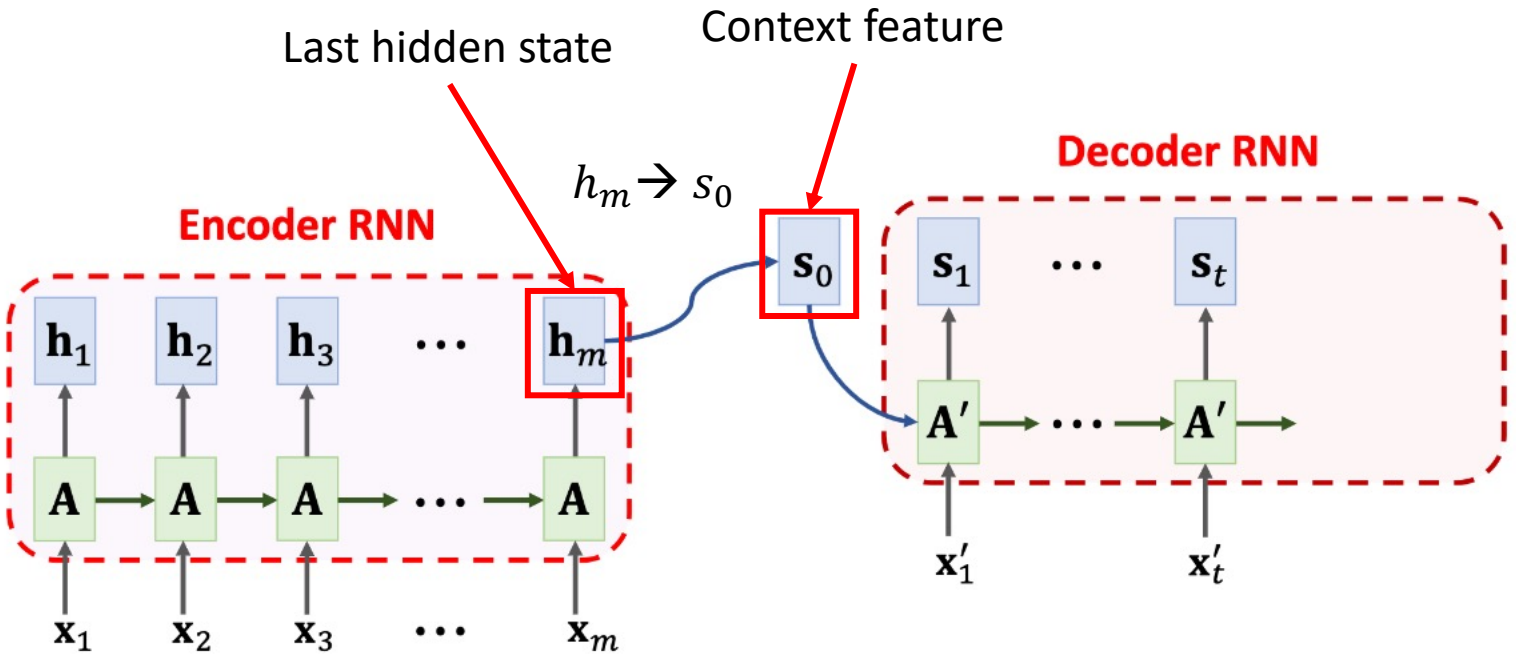


Image from https://github.com/wangshusen/DeepLearning/blob/master/Slides/9_RNN_8.pdf

Attention mechanism



No direct connection between hidden states of encoder and decoder

Image from https://github.com/wangshusen/DeepLearning/blob/master/Slides/9_RNN_8.pdf

Attention mechanism

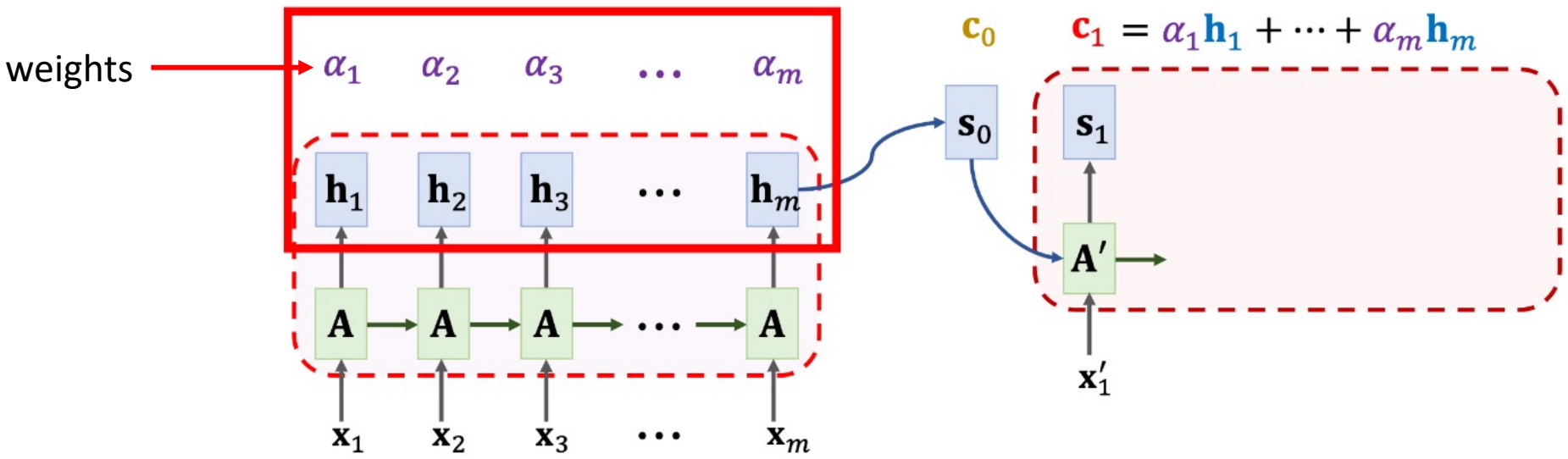


Image from https://github.com/wangshusen/DeepLearning/blob/master/Slides/9_RNN_8.pdf

Attention mechanism

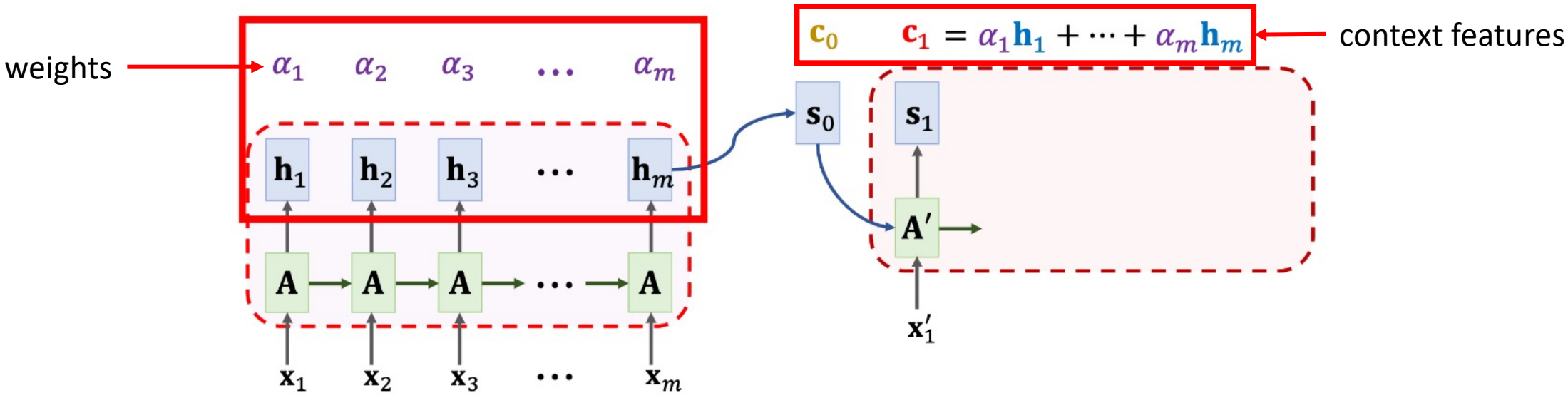


Image from https://github.com/wangshusen/DeepLearning/blob/master/Slides/9_RNN_8.pdf

Attention mechanism

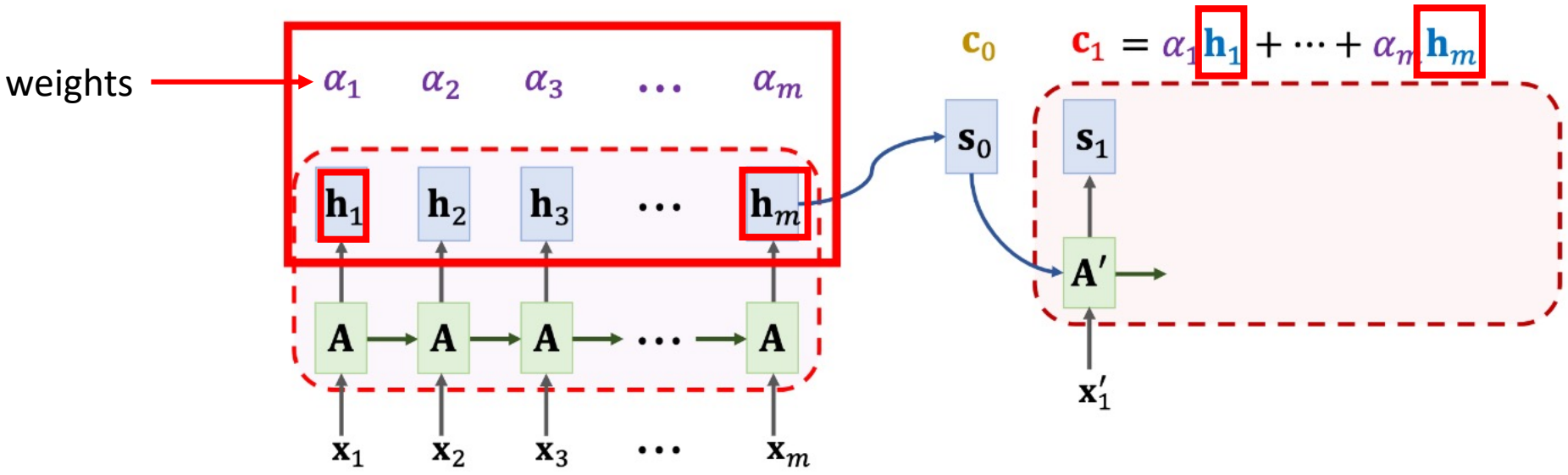


Image from https://github.com/wangshusen/DeepLearning/blob/master/Slides/9_RNN_8.pdf

Attention mechanism

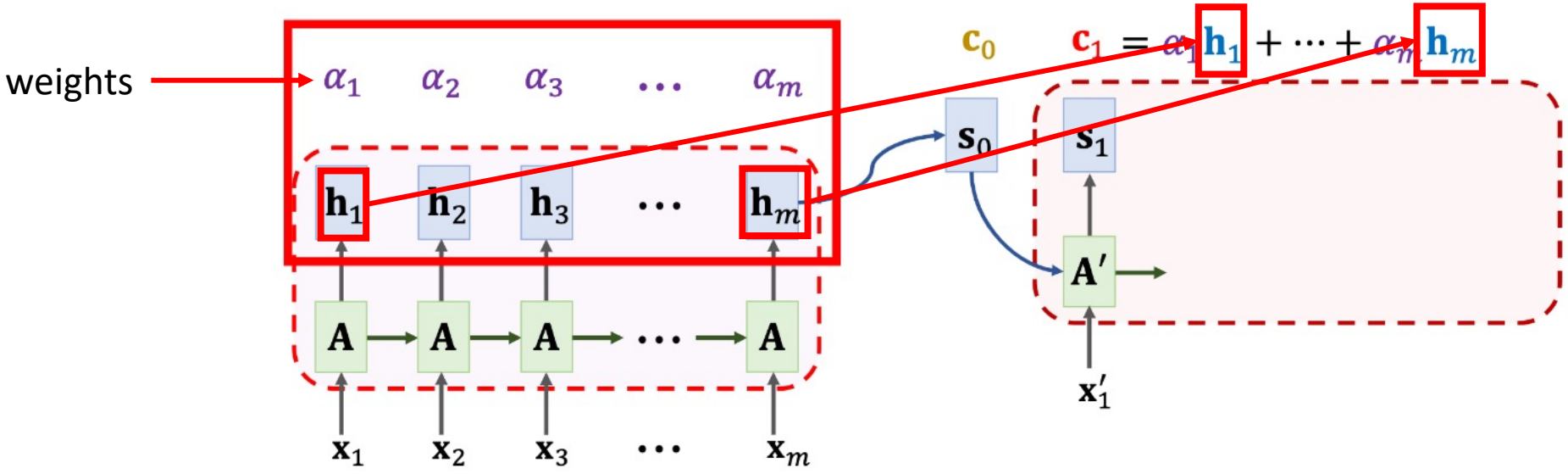


Image from https://github.com/wangshusen/DeepLearning/blob/master/Slides/9_RNN_8.pdf

Attention mechanism

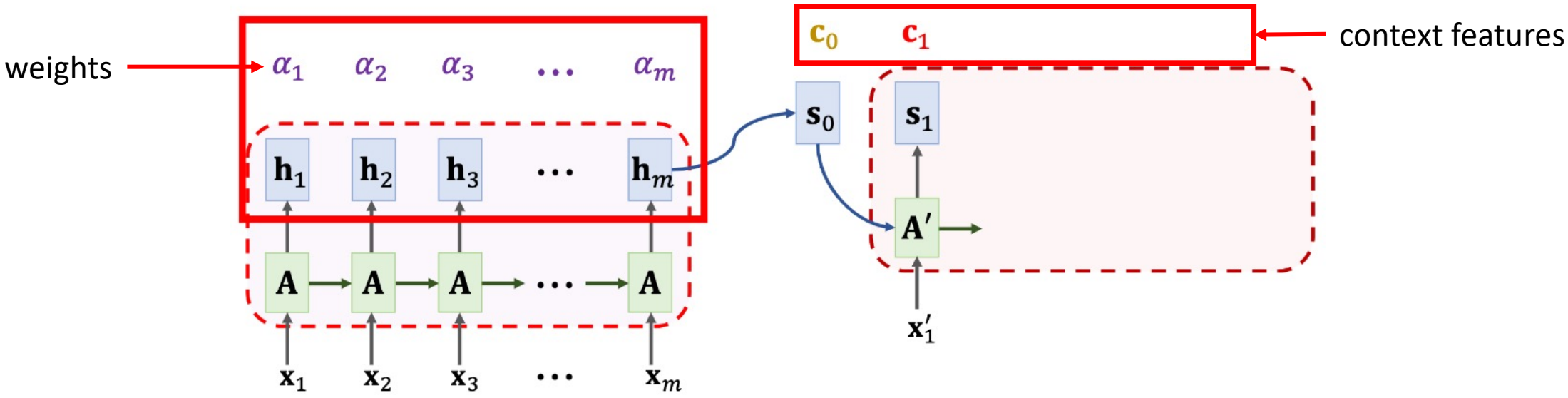


Image from https://github.com/wangshusen/DeepLearning/blob/master/Slides/9_RNN_8.pdf

Attention mechanism

How to use the two variables to build information flow?

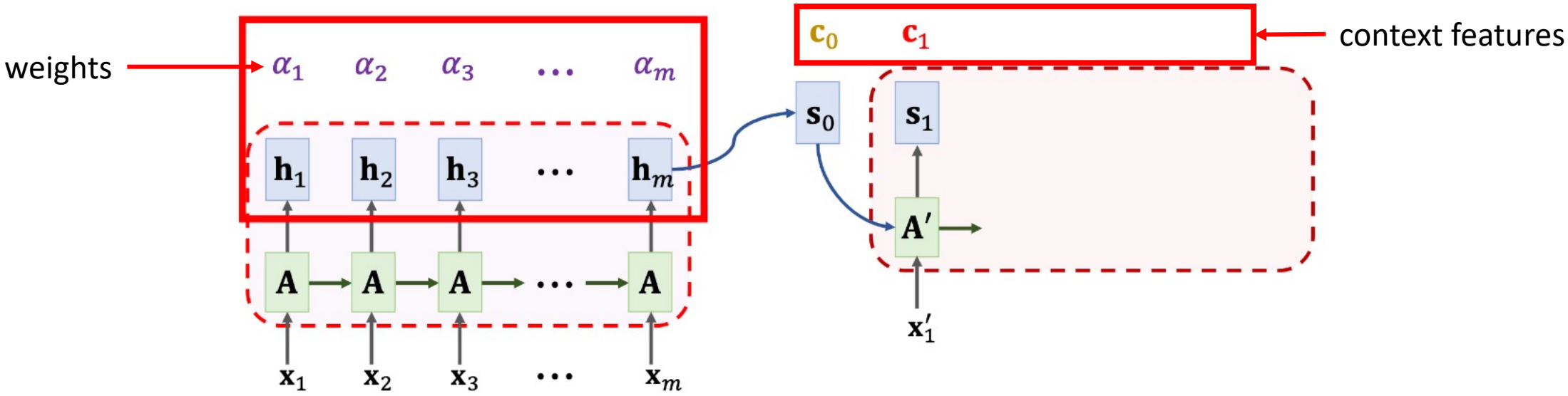


Image from https://github.com/wangshusen/DeepLearning/blob/master/Slides/9_RNN_8.pdf

Attention mechanism

Weight: $\alpha_i = \text{align}(\mathbf{h}_i, \mathbf{s}_0)$.

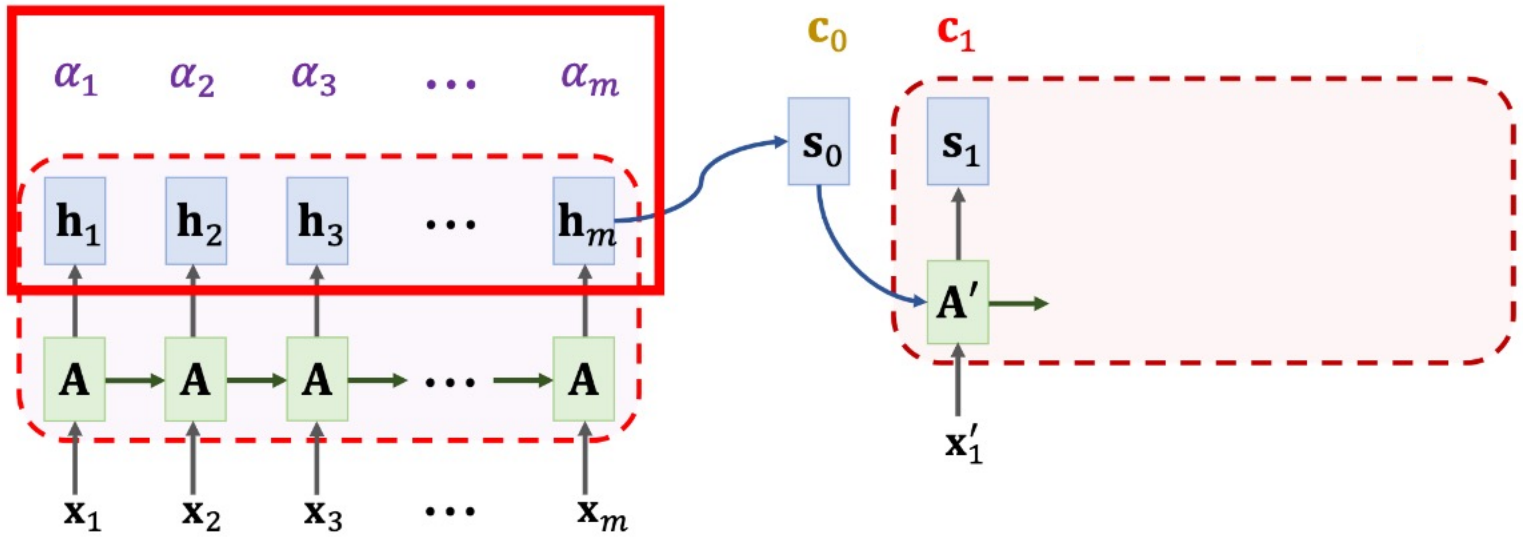


Image from https://github.com/wangshusen/DeepLearning/blob/master/Slides/9_RNN_8.pdf

Attention mechanism

Weight: $\alpha_i = \text{align}(\mathbf{h}_i, \mathbf{s}_0)$.

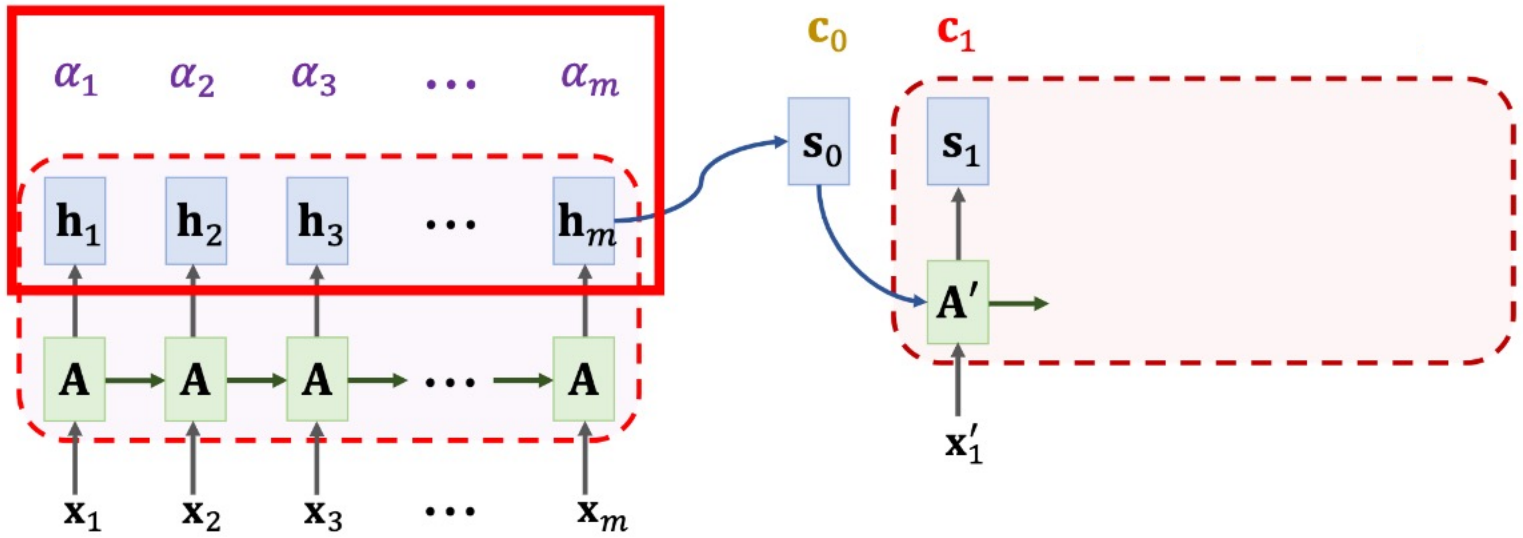


Image from https://github.com/wangshusen/DeepLearning/blob/master/Slides/9_RNN_8.pdf

Attention mechanism

Weight: $\alpha_i = \text{align}(\mathbf{h}_i, \mathbf{s}_0)$.

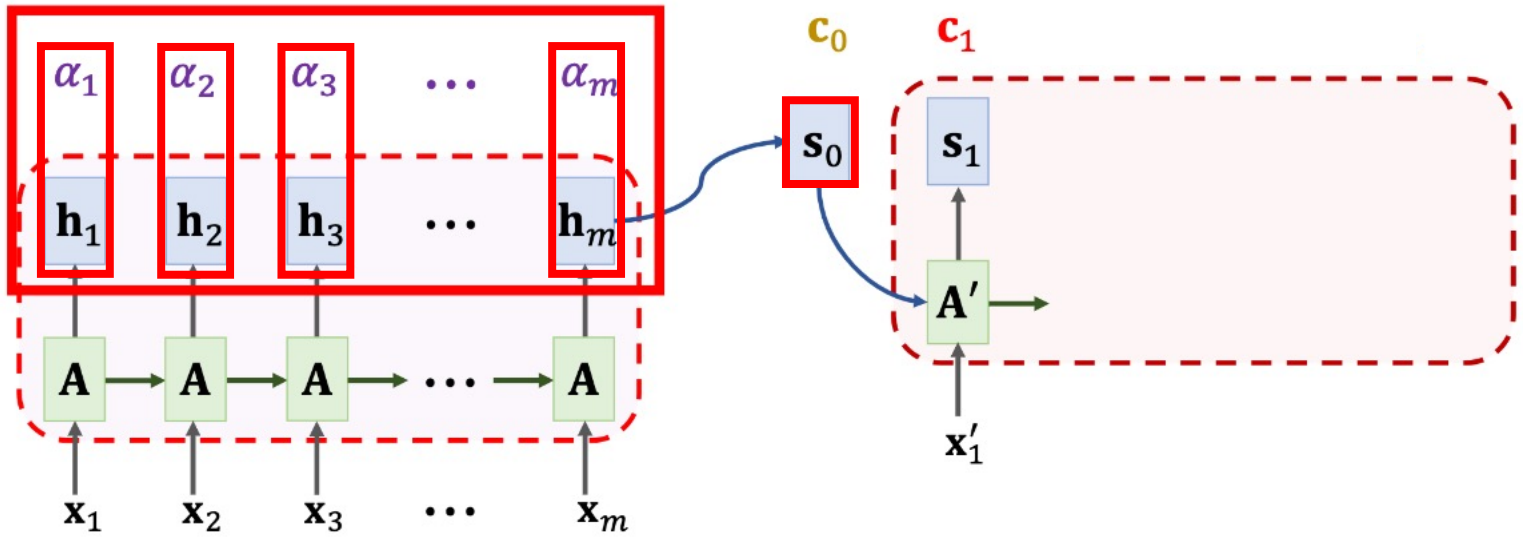
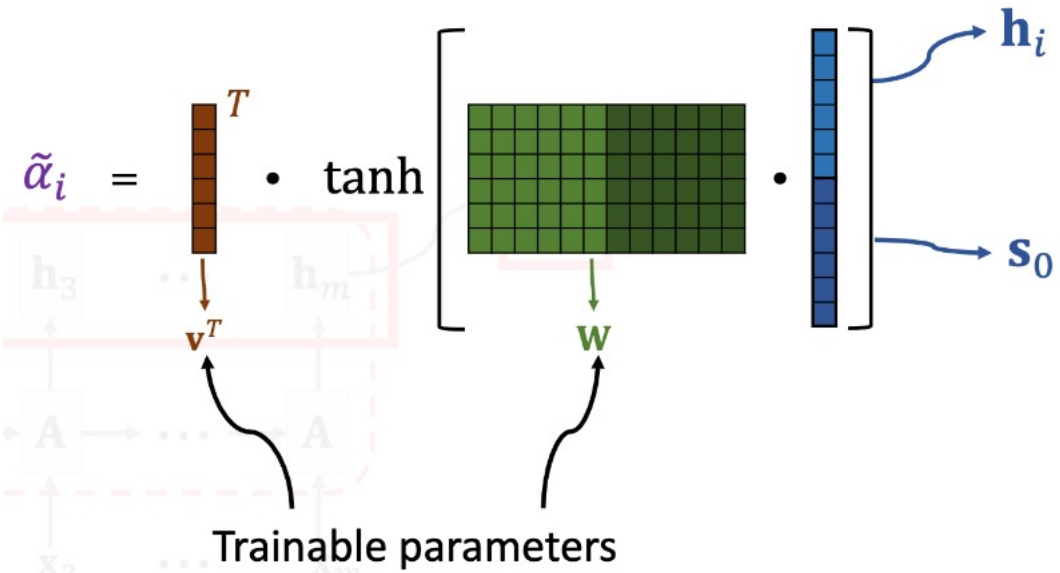


Image from https://github.com/wangshusen/DeepLearning/blob/master/Slides/9_RNN_8.pdf

Attention mechanism

Weight: $\alpha_i = \text{align}(\mathbf{h}_i, \mathbf{s}_0)$.



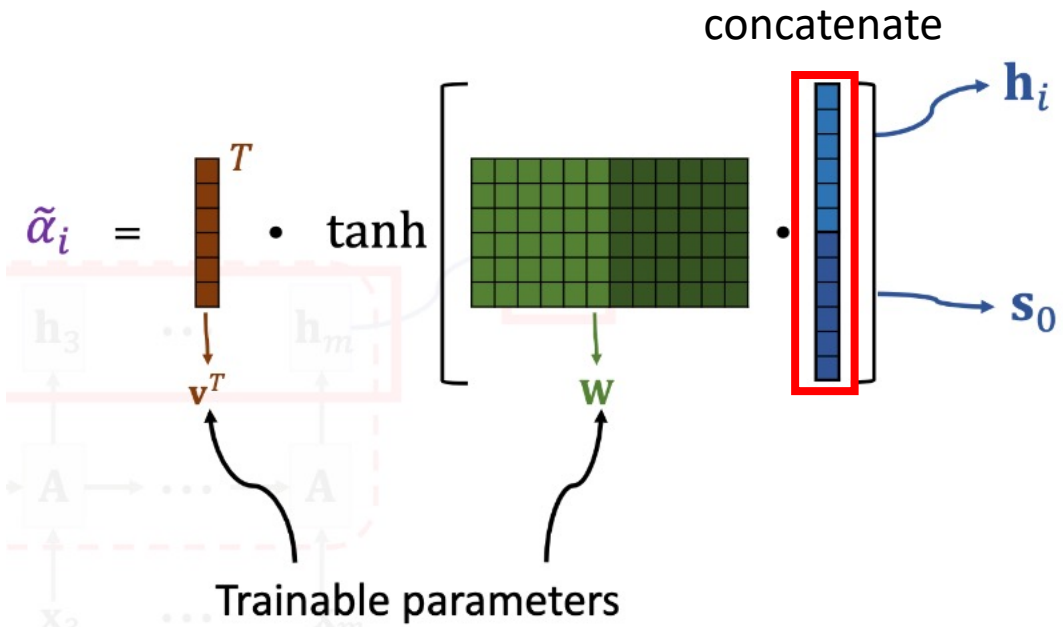
Then **normalize** $\tilde{\alpha}_1, \cdots, \tilde{\alpha}_m$ (so that they sum to 1):

$$[\alpha_1, \cdots, \alpha_m] = \text{Softmax}([\tilde{\alpha}_1, \cdots, \tilde{\alpha}_m]).$$

Image from https://github.com/wangshusen/DeepLearning/blob/master/Slides/9_RNN_8.pdf

Attention mechanism

Weight: $\alpha_i = \text{align}(\mathbf{h}_i, \mathbf{s}_0)$.



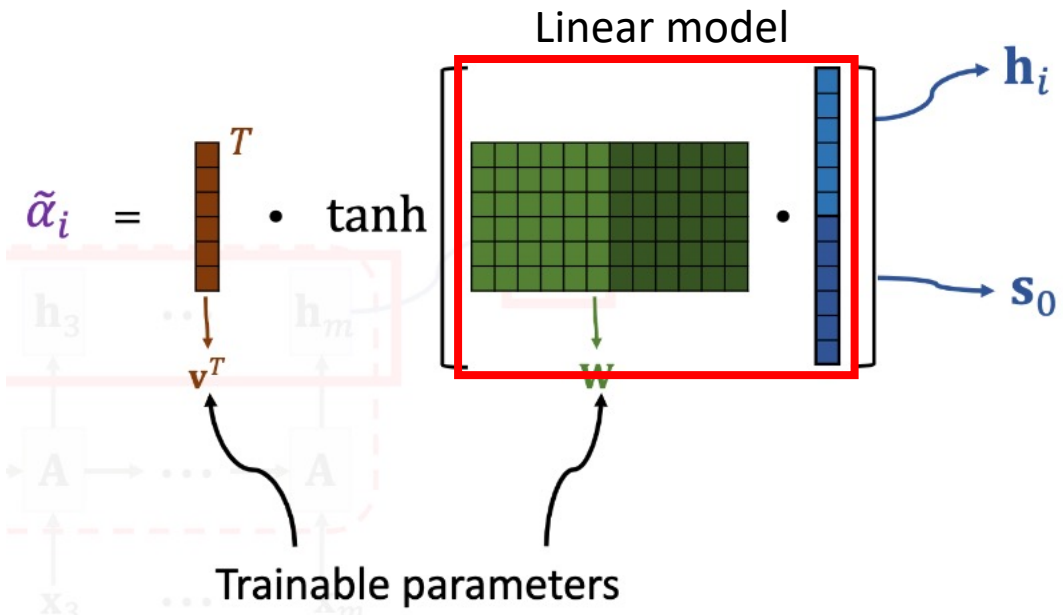
Then **normalize** $\tilde{\alpha}_1, \cdots, \tilde{\alpha}_m$ (so that they sum to 1):

$$[\alpha_1, \cdots, \alpha_m] = \text{Softmax}([\tilde{\alpha}_1, \cdots, \tilde{\alpha}_m]).$$

Image from https://github.com/wangshusen/DeepLearning/blob/master/Slides/9_RNN_8.pdf

Attention mechanism

Weight: $\alpha_i = \text{align}(\mathbf{h}_i, \mathbf{s}_0)$.



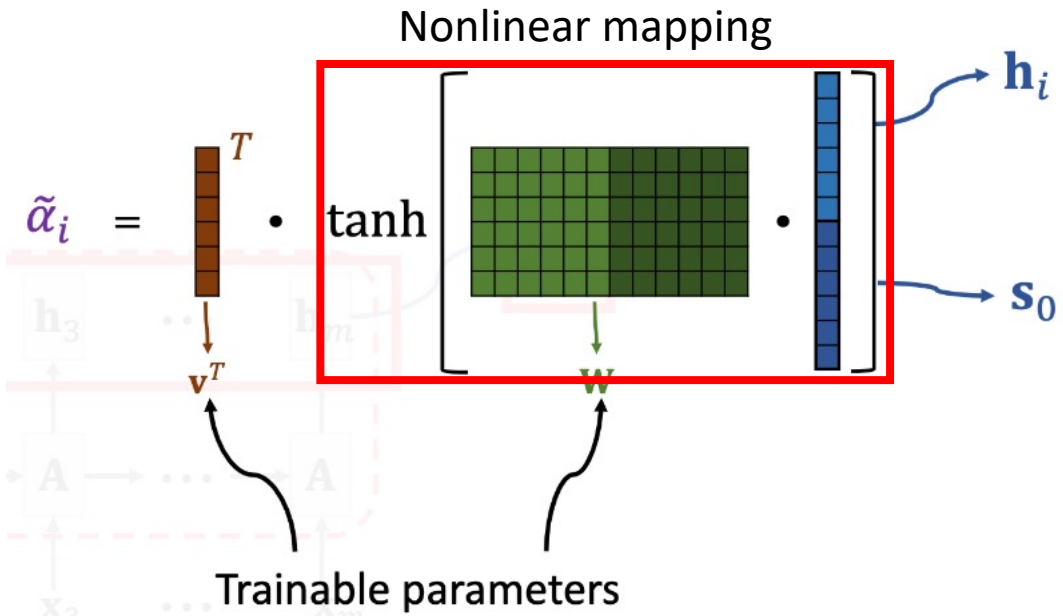
Then **normalize** $\tilde{\alpha}_1, \cdots, \tilde{\alpha}_m$ (so that they sum to 1):

$$[\alpha_1, \cdots, \alpha_m] = \text{Softmax}([\tilde{\alpha}_1, \cdots, \tilde{\alpha}_m]).$$

Image from https://github.com/wangshusen/DeepLearning/blob/master/Slides/9_RNN_8.pdf

Attention mechanism

Weight: $\alpha_i = \text{align}(\mathbf{h}_i, \mathbf{s}_0)$.



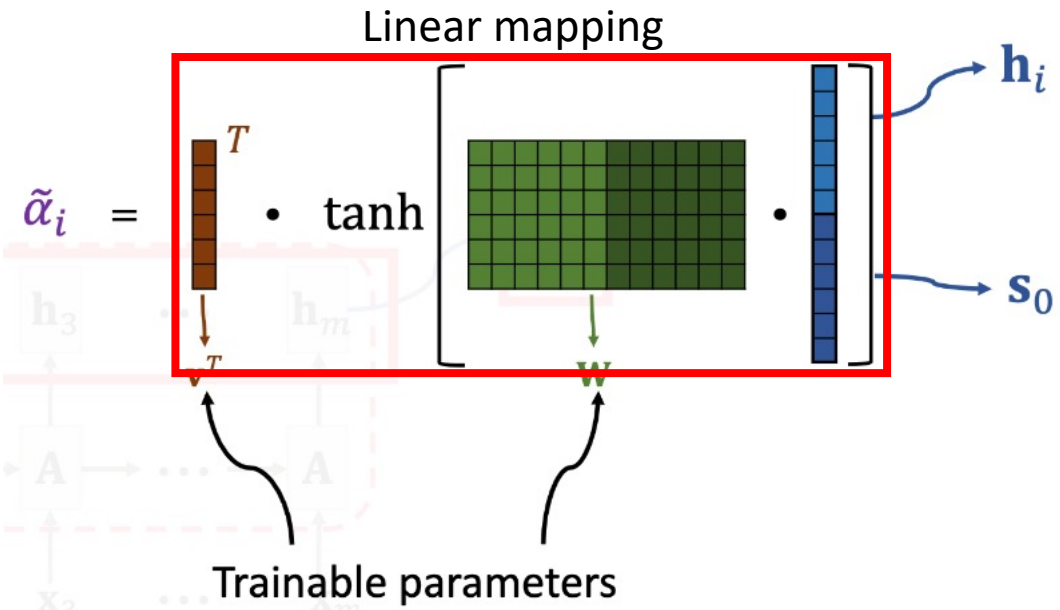
Then **normalize** $\tilde{\alpha}_1, \cdots, \tilde{\alpha}_m$ (so that they sum to 1):

$$[\alpha_1, \cdots, \alpha_m] = \text{Softmax}([\tilde{\alpha}_1, \cdots, \tilde{\alpha}_m]).$$

Image from https://github.com/wangshusen/DeepLearning/blob/master/Slides/9_RNN_8.pdf

Attention mechanism

Weight: $\alpha_i = \text{align}(\mathbf{h}_i, \mathbf{s}_0)$.



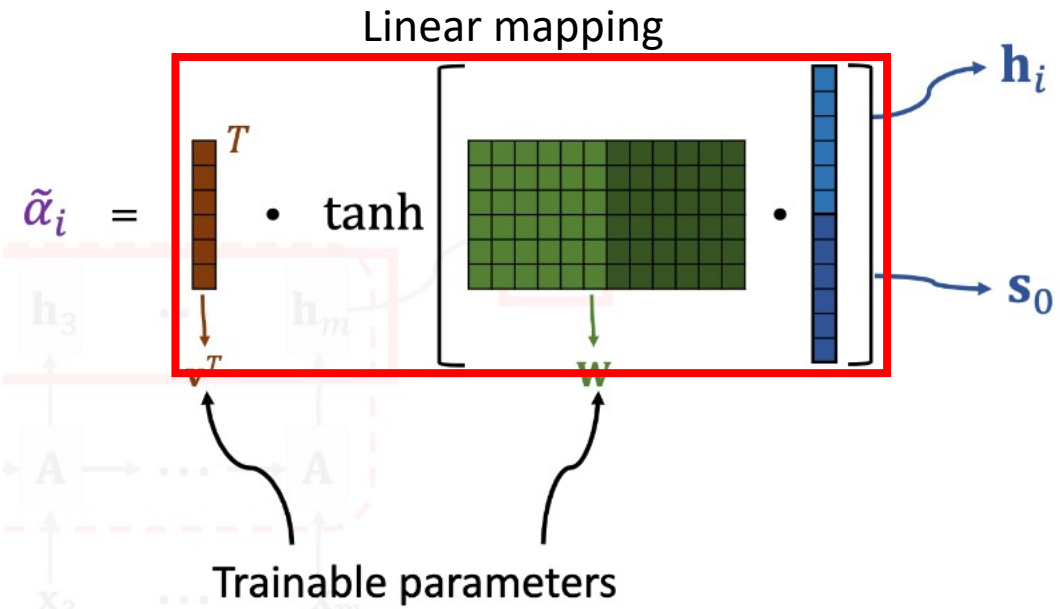
Then **normalize** $\tilde{\alpha}_1, \cdots, \tilde{\alpha}_m$ (so that they sum to 1):

$$[\alpha_1, \cdots, \alpha_m] = \text{Softmax}([\tilde{\alpha}_1, \cdots, \tilde{\alpha}_m]).$$

Image from https://github.com/wangshusen/DeepLearning/blob/master/Slides/9_RNN_8.pdf

Attention mechanism

Weight: $\alpha_i = \text{align}(\mathbf{h}_i, \mathbf{s}_0)$.



Then **normalize** $\tilde{\alpha}_1, \cdots, \tilde{\alpha}_m$ (so that they sum to 1):

$$[\alpha_1, \cdots, \alpha_m] = \text{Softmax}([\tilde{\alpha}_1, \cdots, \tilde{\alpha}_m]).$$

Image from https://github.com/wangshusen/DeepLearning/blob/master/Slides/9_RNN_8.pdf

Attention mechanism

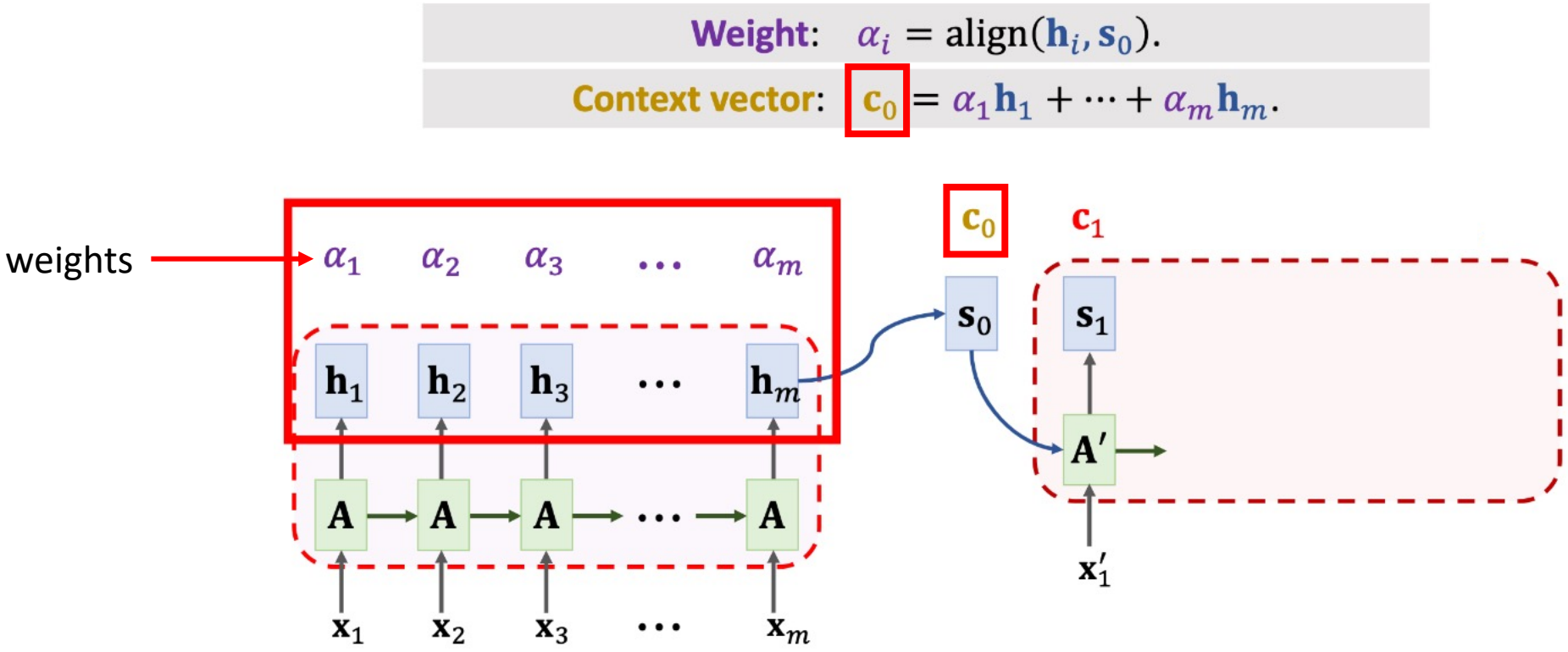


Image from https://github.com/wangshusen/DeepLearning/blob/master/Slides/9_RNN_8.pdf

Attention mechanism

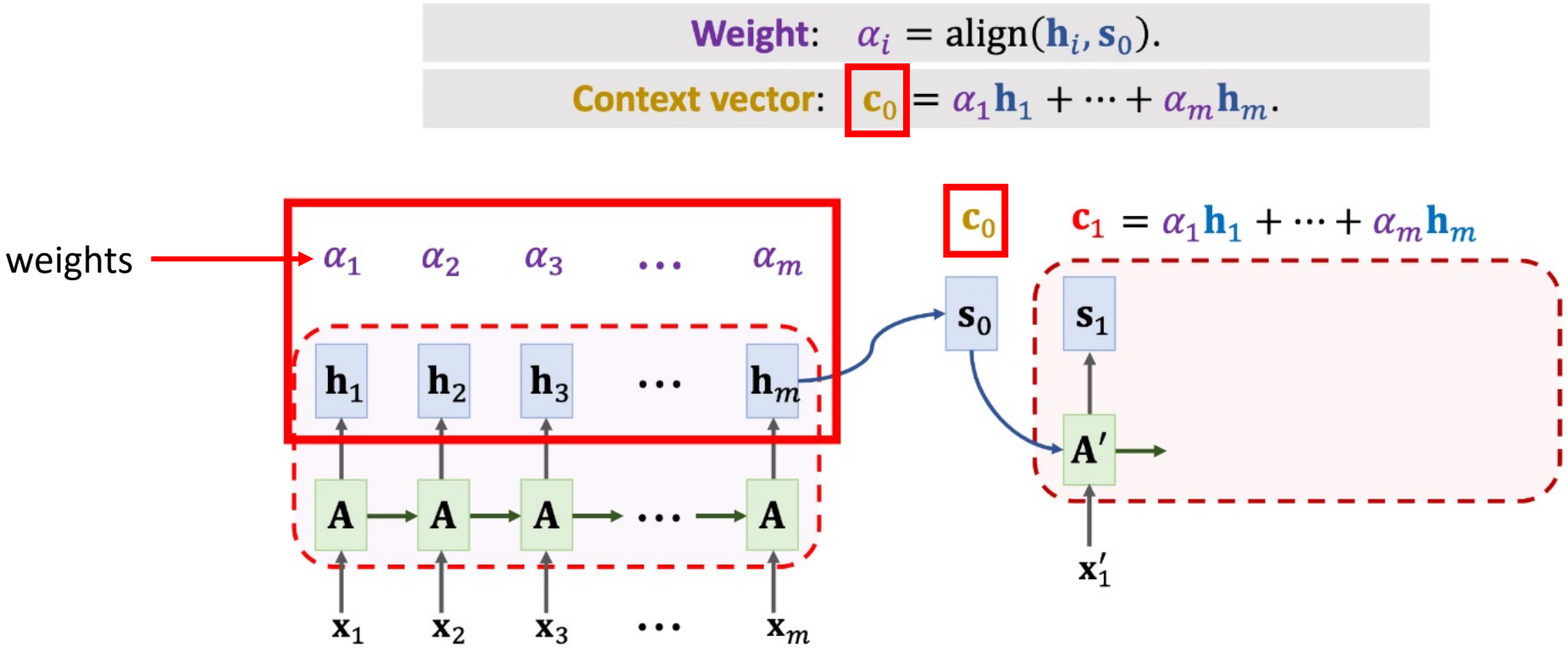


Image from https://github.com/wangshusen/DeepLearning/blob/master/Slides/9_RNN_8.pdf

Attention mechanism

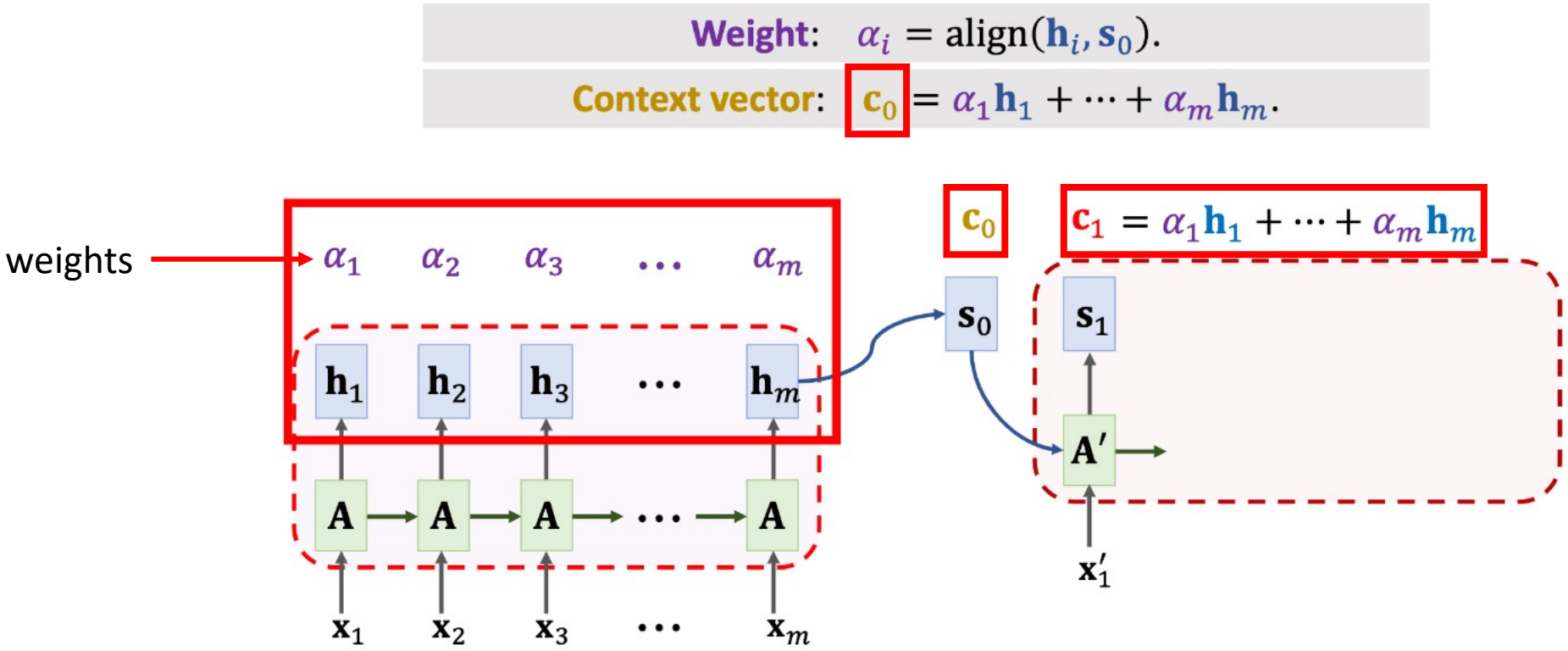


Image from https://github.com/wangshusen/DeepLearning/blob/master/Slides/9_RNN_8.pdf

Attention mechanism

Weight: $\alpha_i = \text{align}(\mathbf{h}_i, \mathbf{s}_0)$.

Context vector: $\mathbf{c}_0 = \alpha_1 \mathbf{h}_1 + \cdots + \alpha_m \mathbf{h}_m$.

Q: should we use different $\alpha's$ for different c_t?

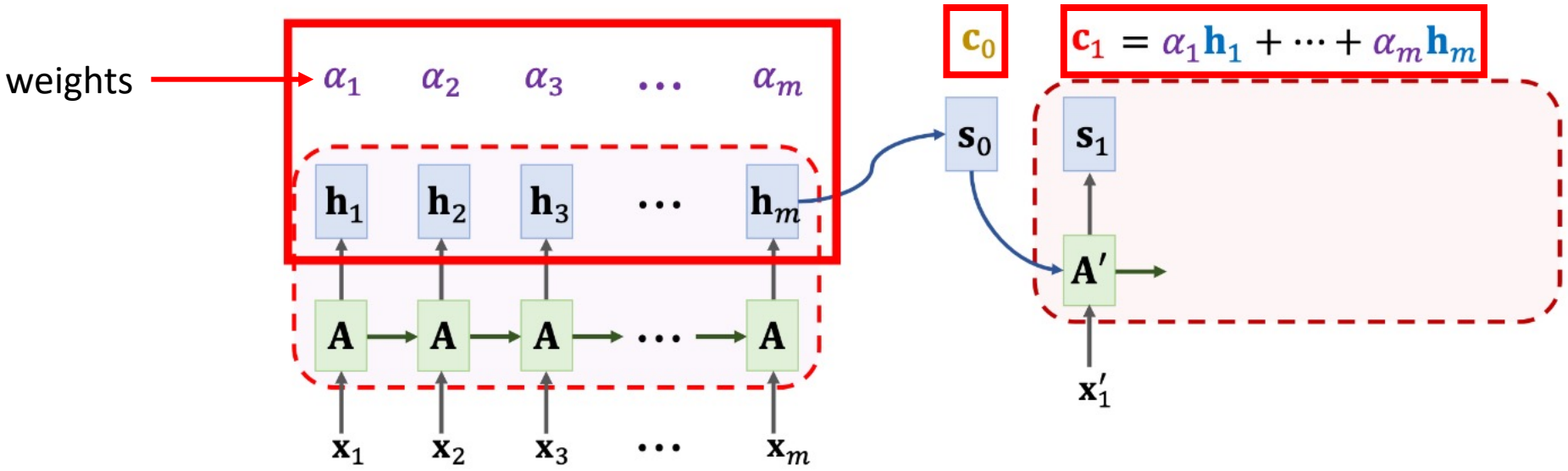


Image from https://github.com/wangshusen/DeepLearning/blob/master/Slides/9_RNN_8.pdf

Attention mechanism

Weight: $\alpha_i = \text{align}(\mathbf{h}_i, \mathbf{s}_0)$.

Context vector: $\mathbf{c}_0 = \alpha_1 \mathbf{h}_1 + \cdots + \alpha_m \mathbf{h}_m$.

Q: should we use different $\alpha's$ for different c_t?
Different $\alpha's$. Why?

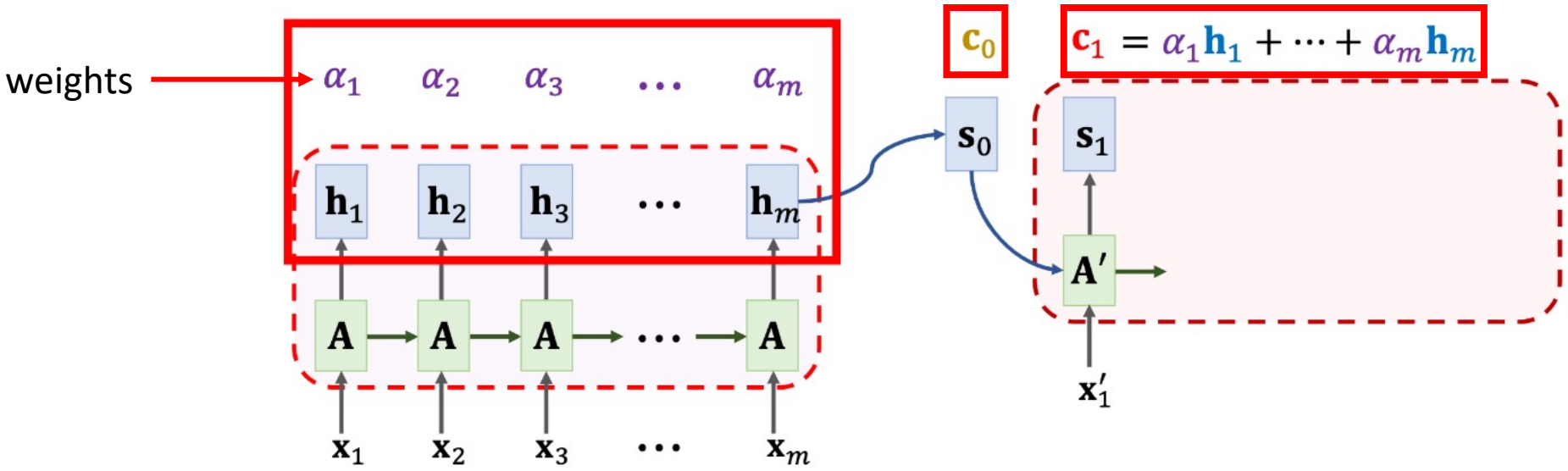


Image from https://github.com/wangshusen/DeepLearning/blob/master/Slides/9_RNN_8.pdf

Attention mechanism

$$c_i = \sum_{j=1}^{T_x} \alpha_{ij} h_j.$$

In the original paper

Weight: $\alpha_i = \text{align}(\mathbf{h}_i, \mathbf{s}_0)$.

Context vector: $\mathbf{c}_0 = \alpha_1 \mathbf{h}_1 + \cdots + \alpha_m \mathbf{h}_m$.

Q: should we use different $\alpha's$ for different c_t?
Different $\alpha's$. Why?

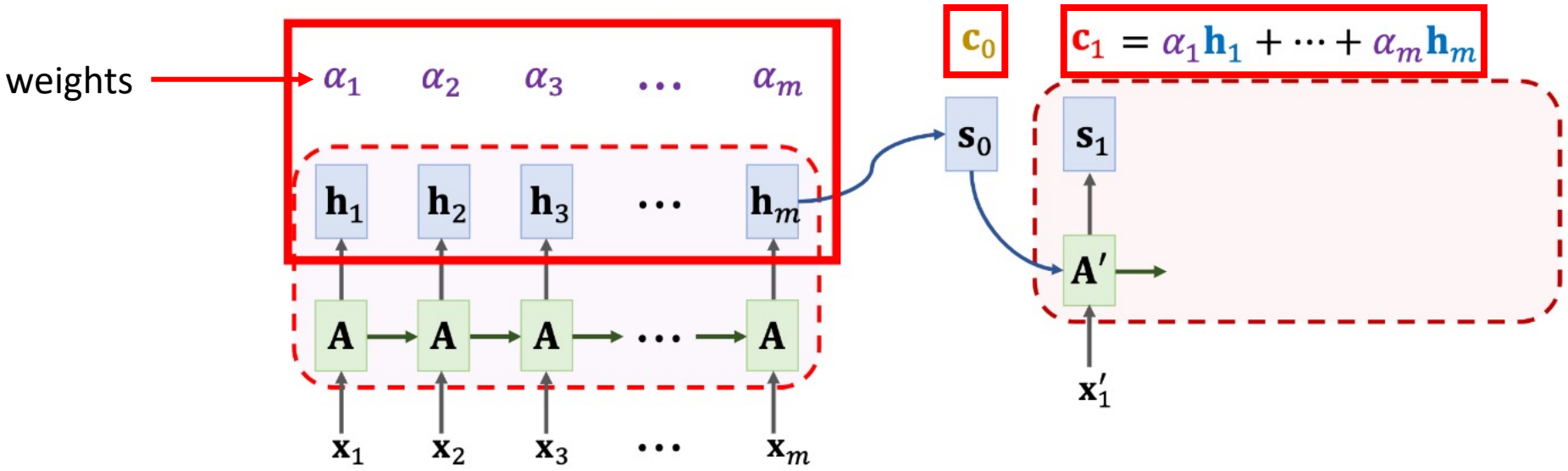


Image from https://github.com/wangshusen/DeepLearning/blob/master/Slides/9_RNN_8.pdf

Attention mechanism

$$c_i = \sum_{j=1}^{T_x} \alpha_{ij} h_j.$$

In the original paper

Weight: $\alpha_i = \text{align}(\mathbf{h}_i, \mathbf{s}_0)$.

Context vector: $\mathbf{c}_0 = \alpha_1 \mathbf{h}_1 + \cdots + \alpha_m \mathbf{h}_m$.

Q: should we use different $\alpha's$ for different c_t?
Different $\alpha's$. Why?

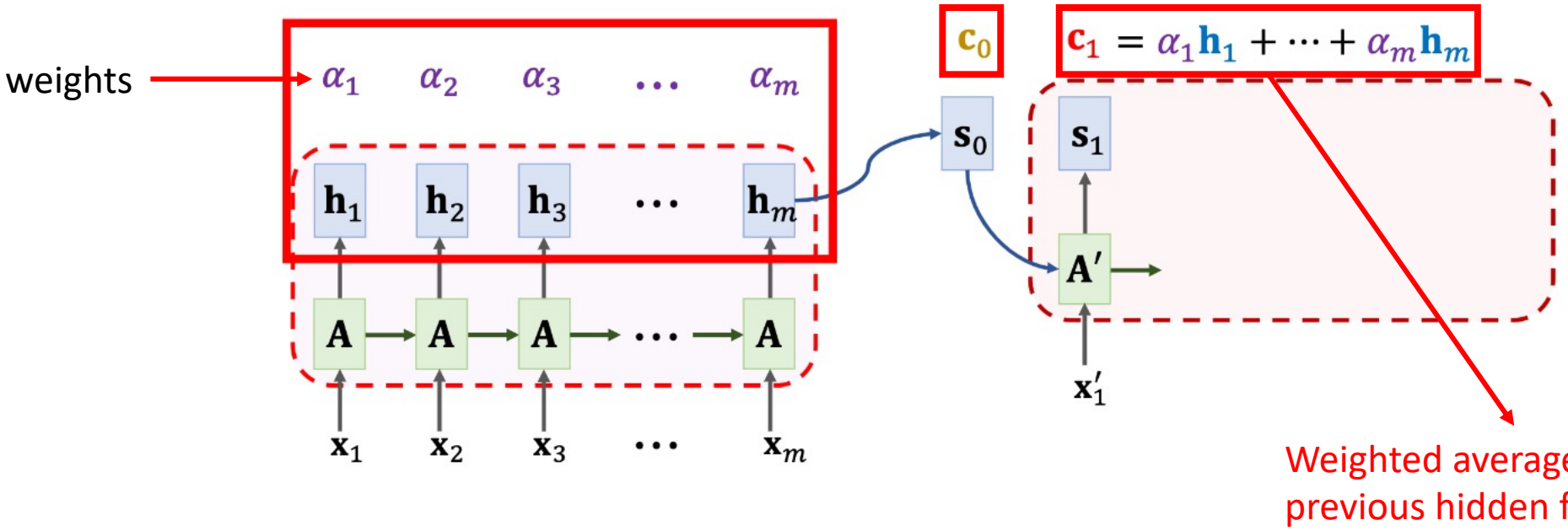


Image from https://github.com/wangshusen/DeepLearning/blob/master/Slides/9_RNN_8.pdf

Attention mechanism

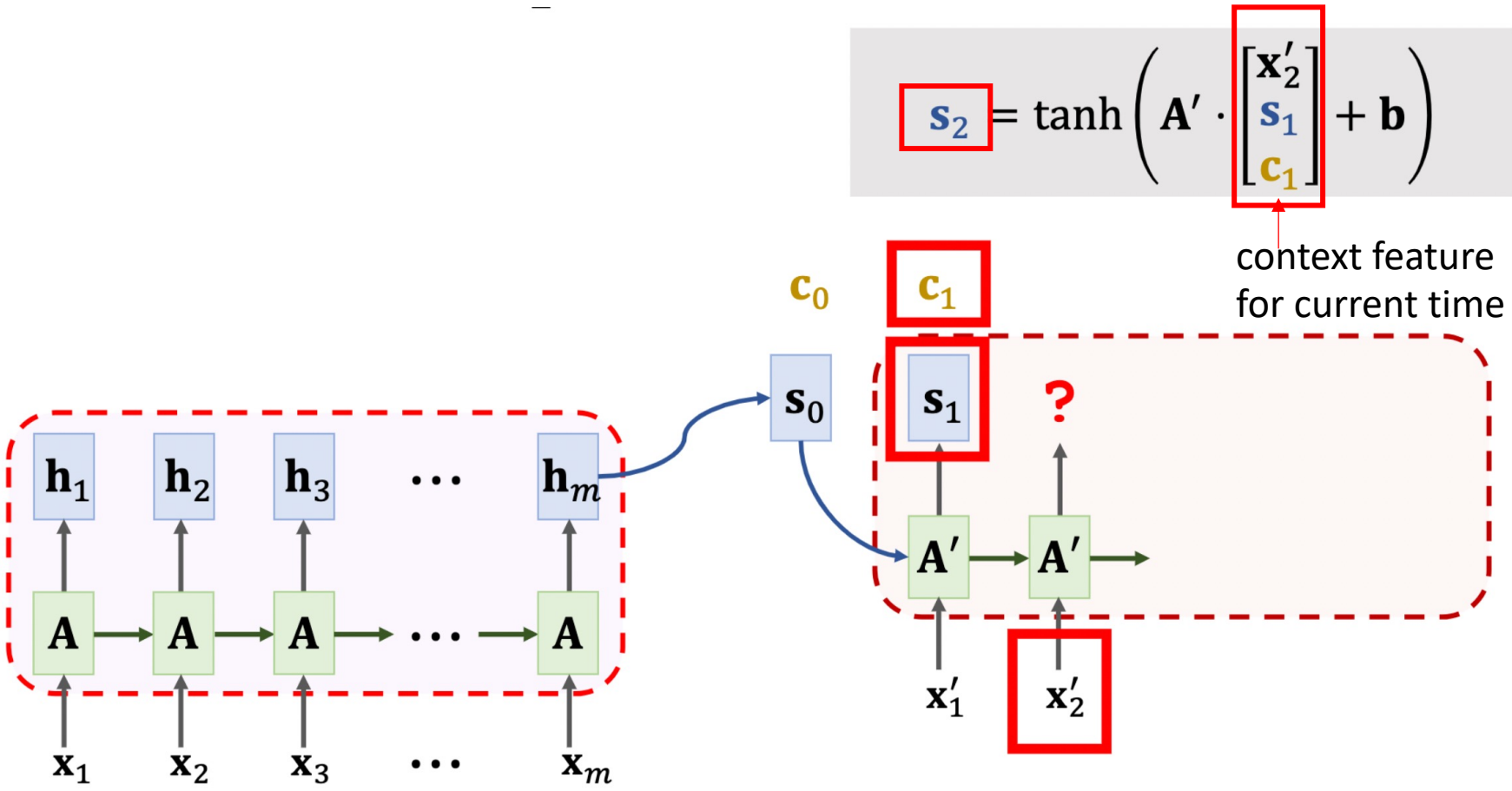


Image from https://github.com/wangshusen/DeepLearning/blob/master/Slides/9_RNN_8.pdf

Attention mechanism

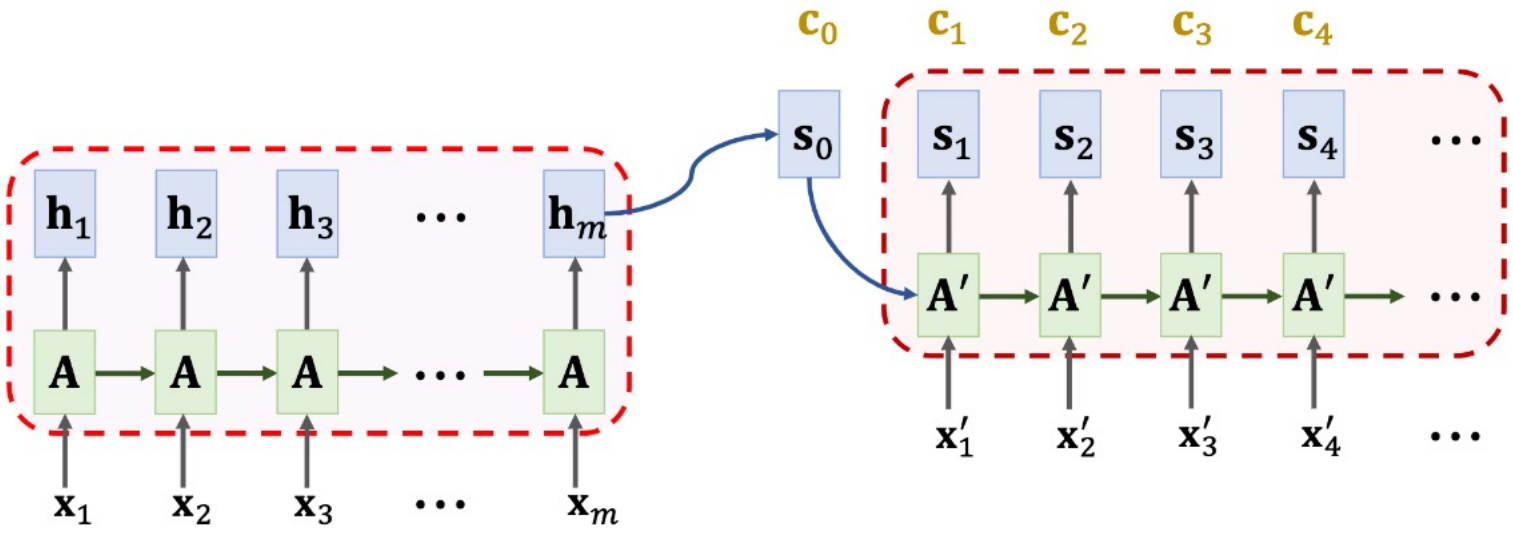


Image from https://github.com/wangshusen/DeepLearning/blob/master/Slides/9_RNN_8.pdf

Attention mechanism

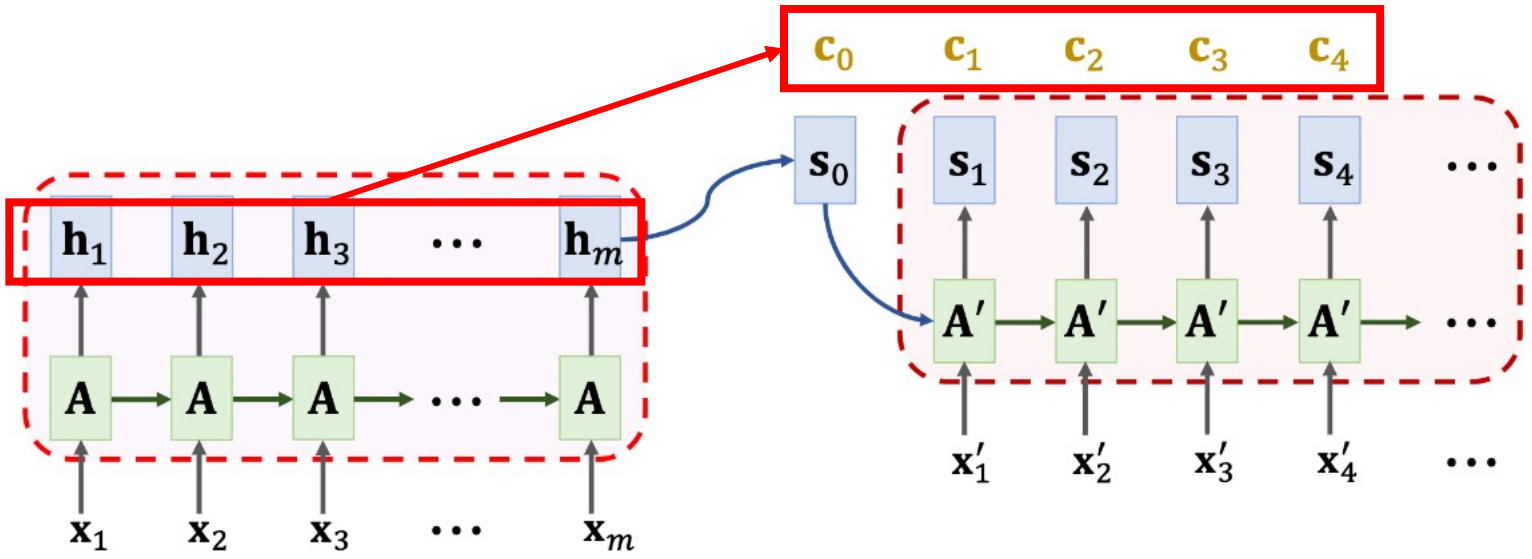
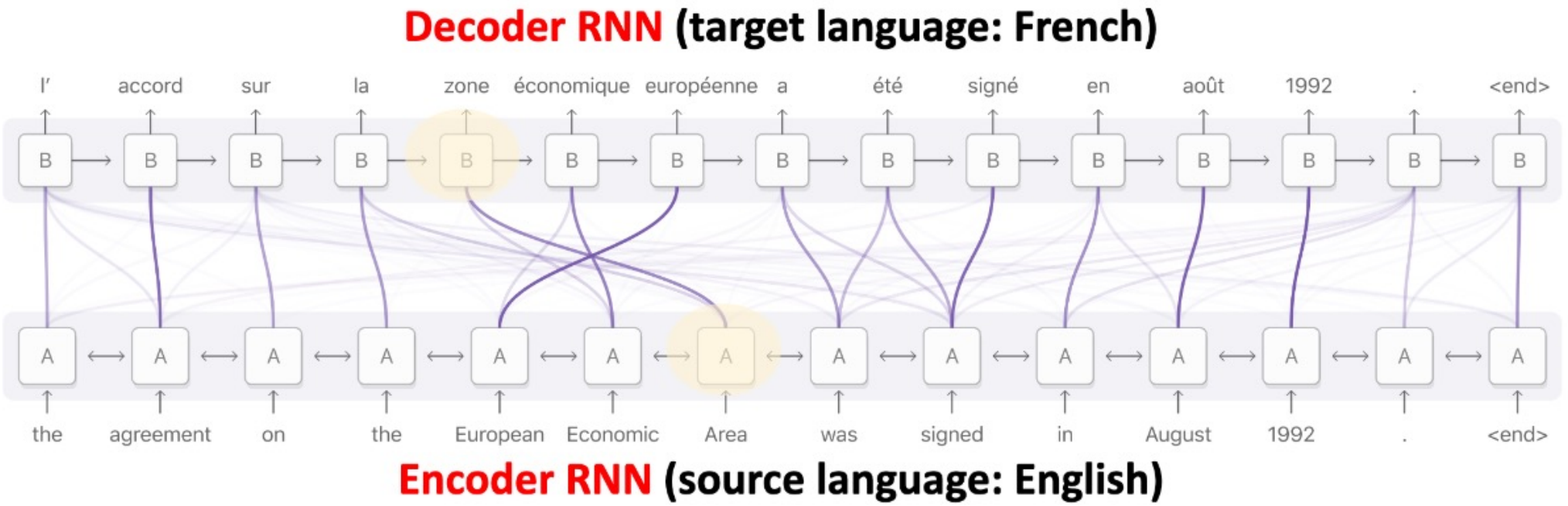
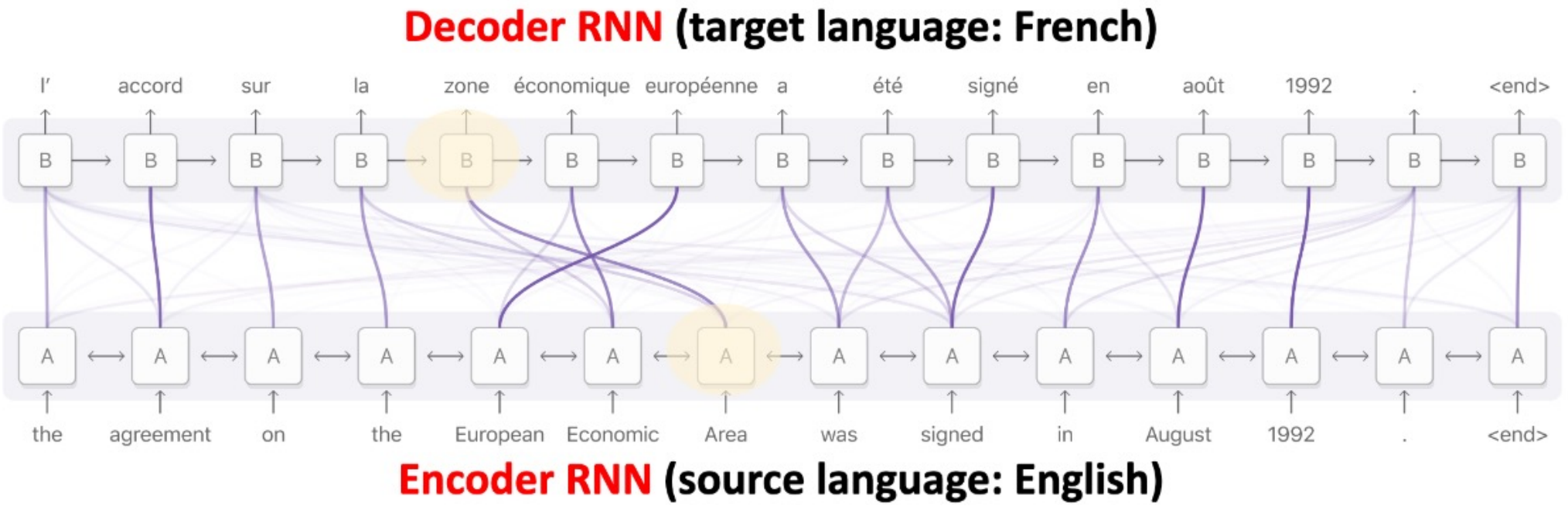


Image from https://github.com/wangshusen/DeepLearning/blob/master/Slides/9_RNN_8.pdf

Input-output correlation



Input-output correlation



Q: can we build attention mechanism in a single RNN (e.g., the encoder)?

Self-attention (intra-attention)

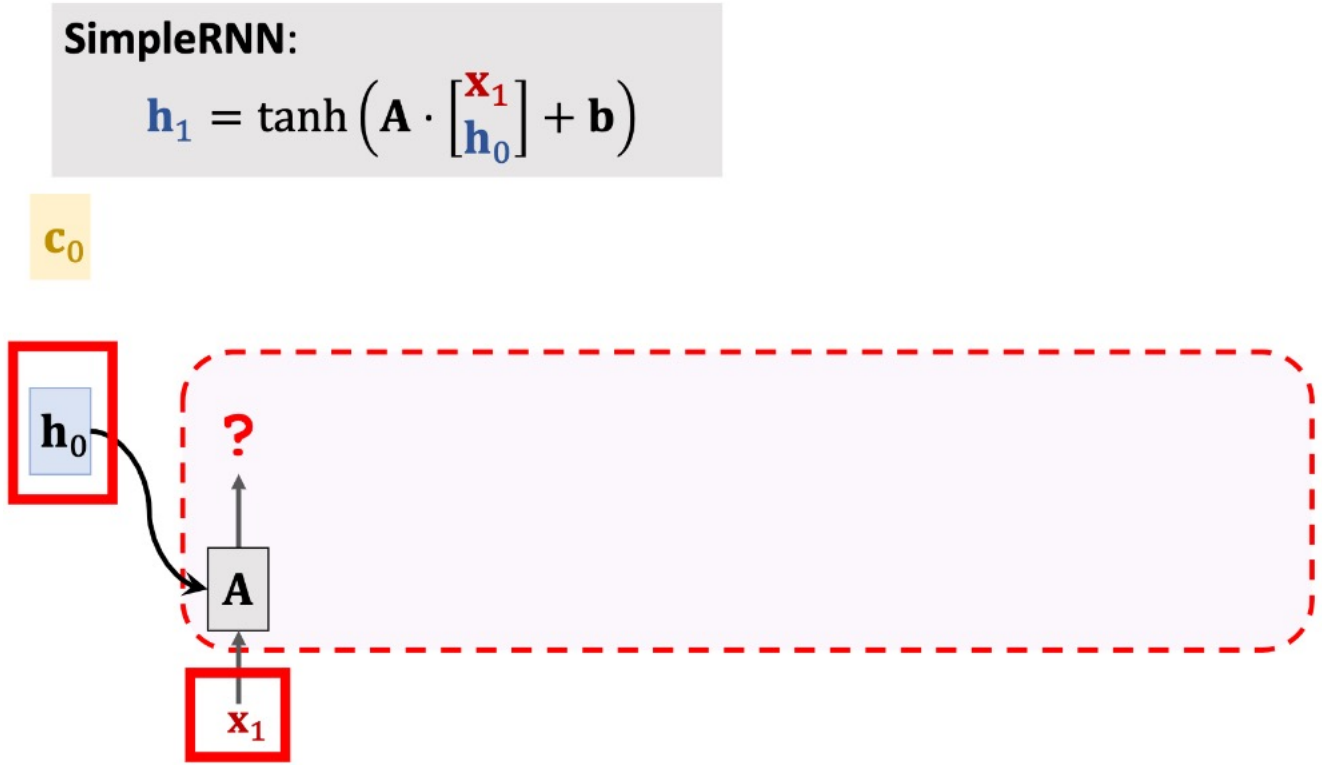


Image from https://github.com/wangshusen/DeepLearning/blob/master/Slides/9_RNN_9.pdf

Self-attention (intra-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:

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

Image from https://github.com/wangshusen/DeepLearning/blob/master/Slides/9_RNN_9.pdf

Self-attention (intra-attention)

Image from https://github.com/wangshusen/DeepLearning/blob/master/Slides/9_RNN_9.pdf

Self-attention (intra-attention)

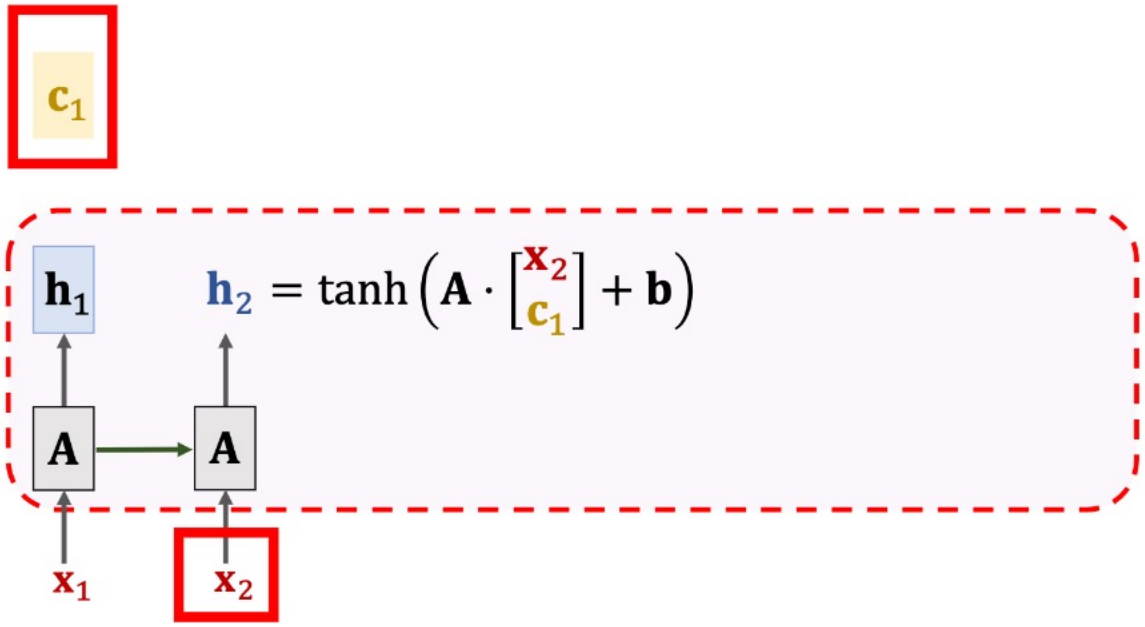


Image from https://github.com/wangshusen/DeepLearning/blob/master/Slides/9_RNN_9.pdf

Self-attention (intra-attention)

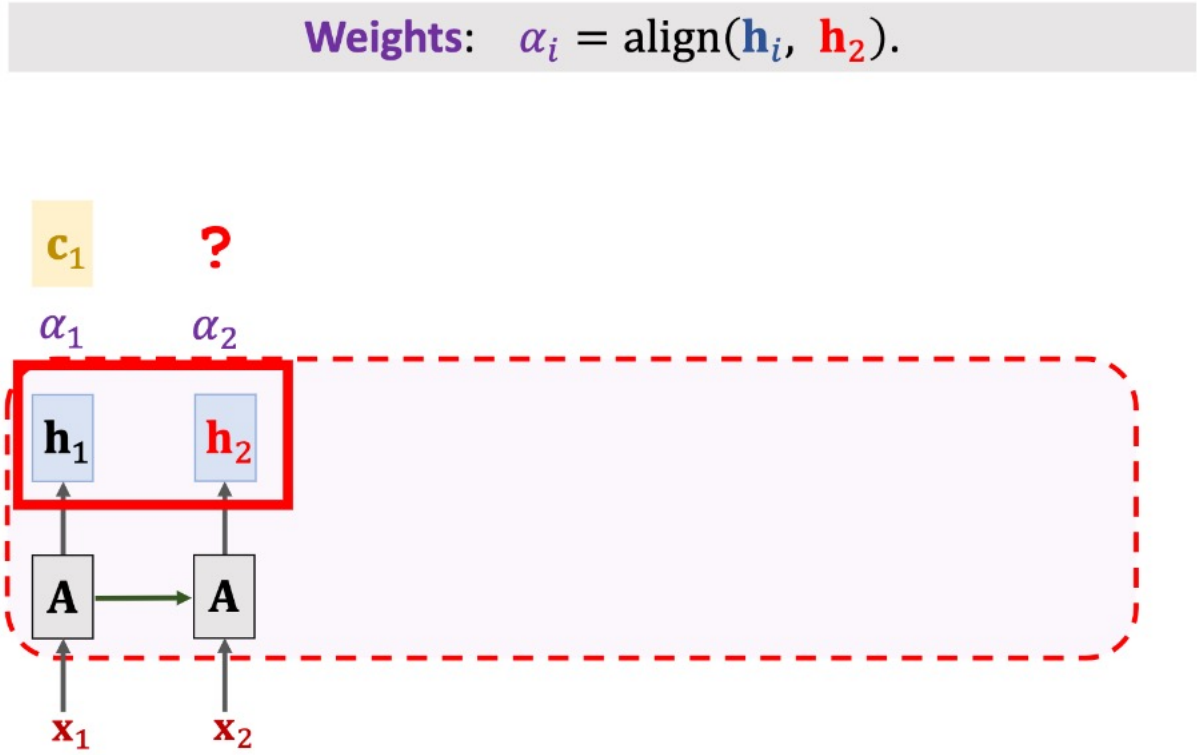


Image from https://github.com/wangshusen/DeepLearning/blob/master/Slides/9_RNN_9.pdf

Self-attention (intra-attention)

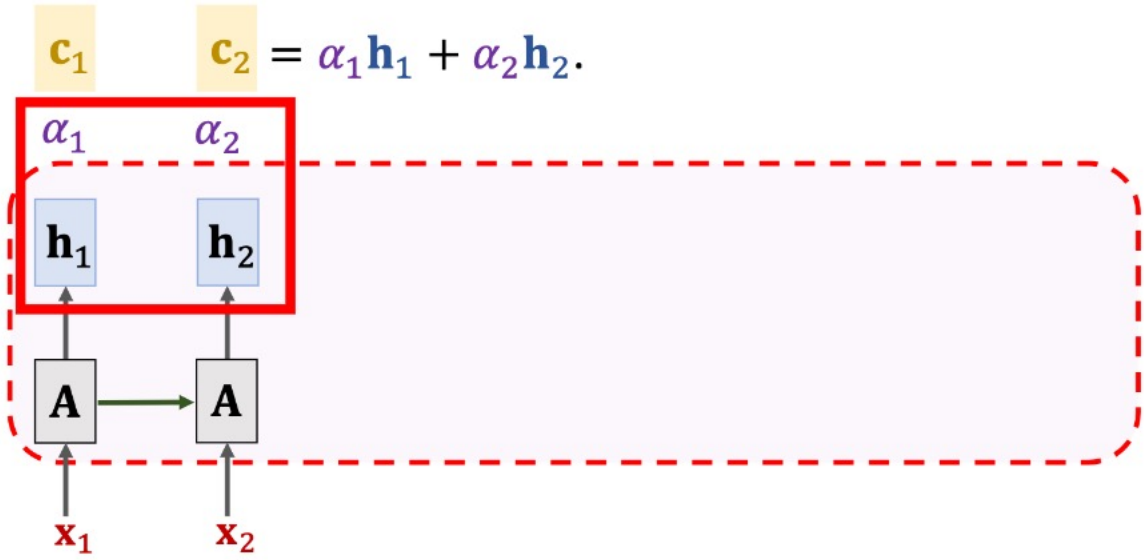


Image from https://github.com/wangshusen/DeepLearning/blob/master/Slides/9_RNN_9.pdf

Self-attention (intra-attention)

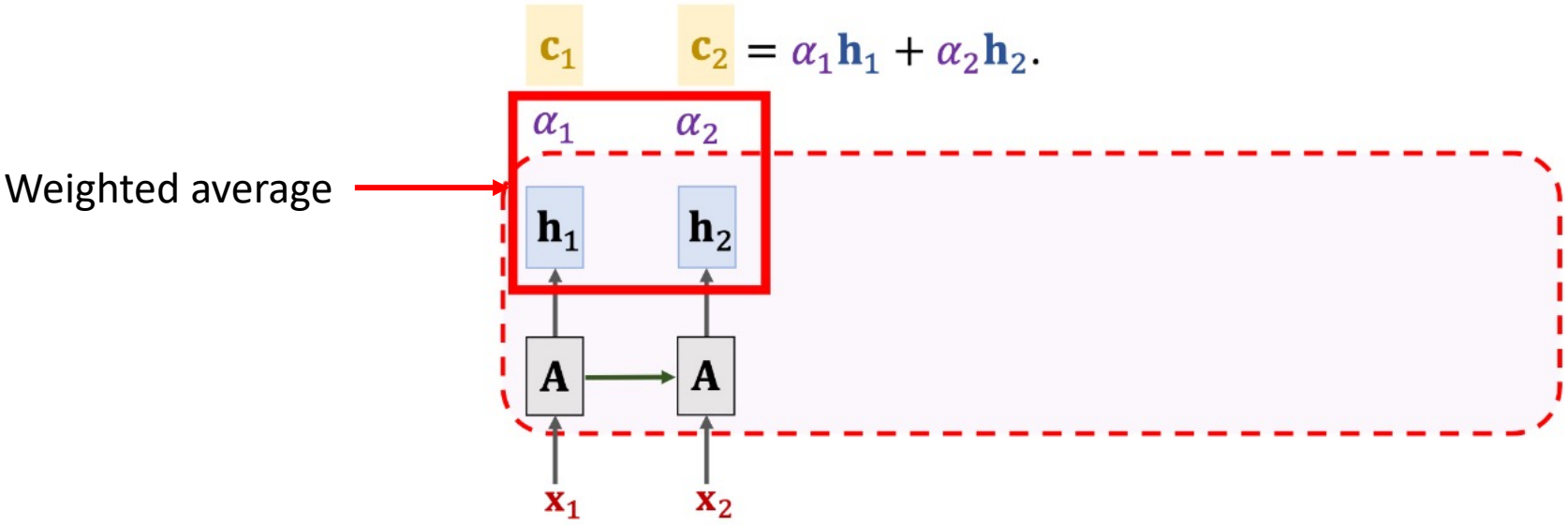


Image from https://github.com/wangshusen/DeepLearning/blob/master/Slides/9_RNN_9.pdf

Self-attention (intra-attention)

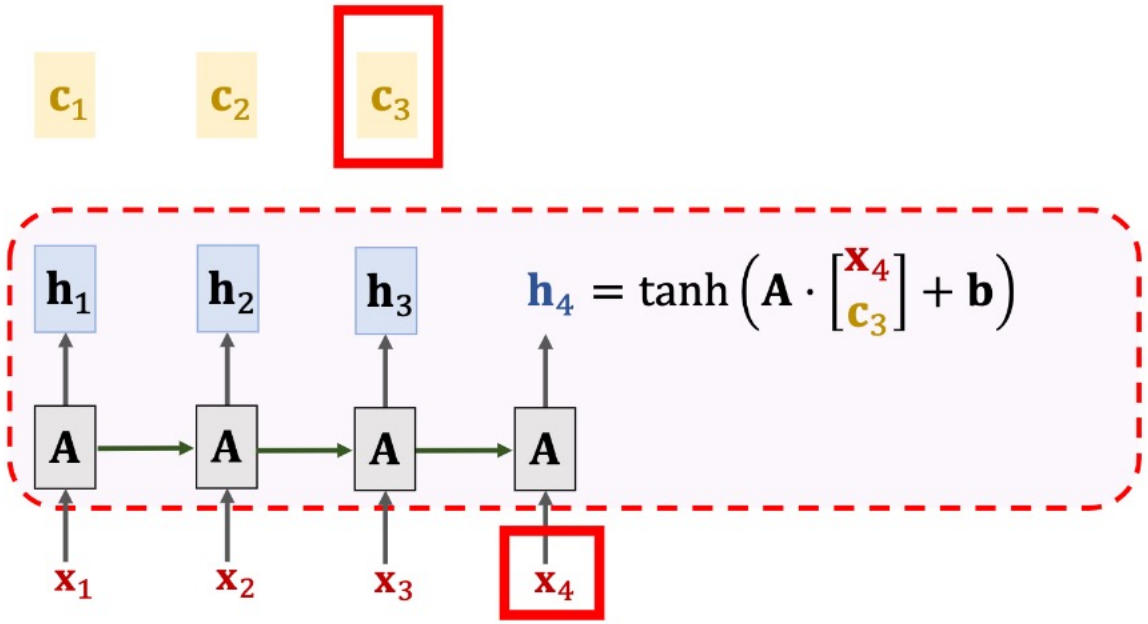


Image from https://github.com/wangshusen/DeepLearning/blob/master/Slides/9_RNN_9.pdf

Self-attention (intra-attention)

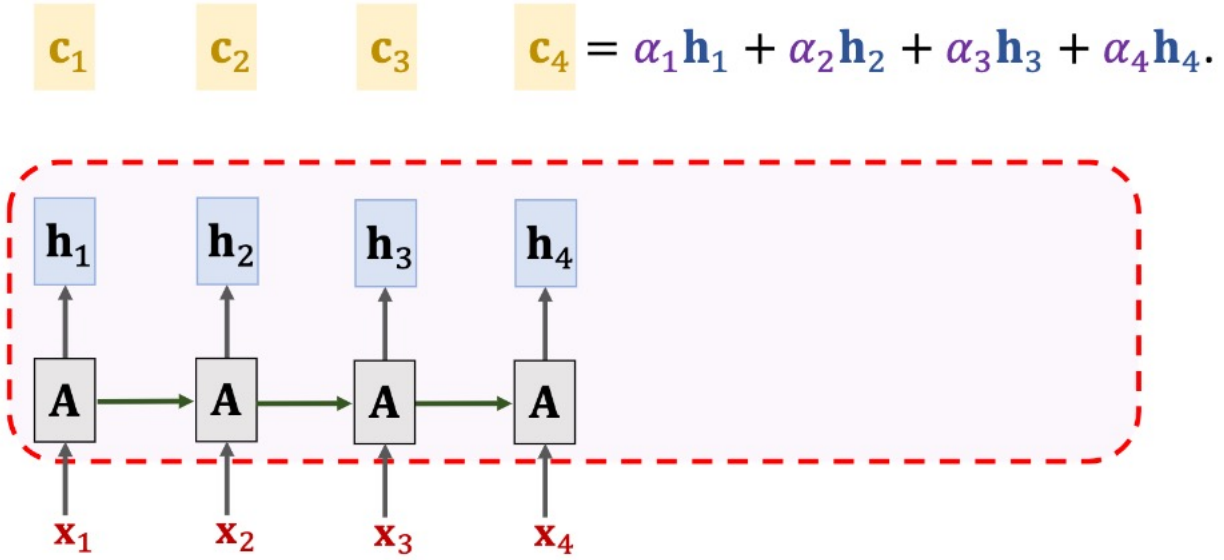


Image from https://github.com/wangshusen/DeepLearning/blob/master/Slides/9_RNN_9.pdf

Self-attention (intra-attention)

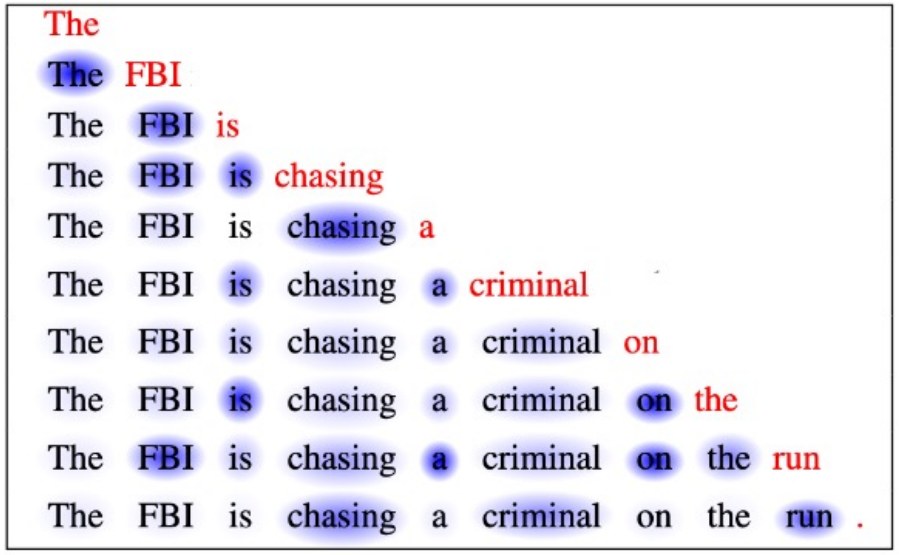
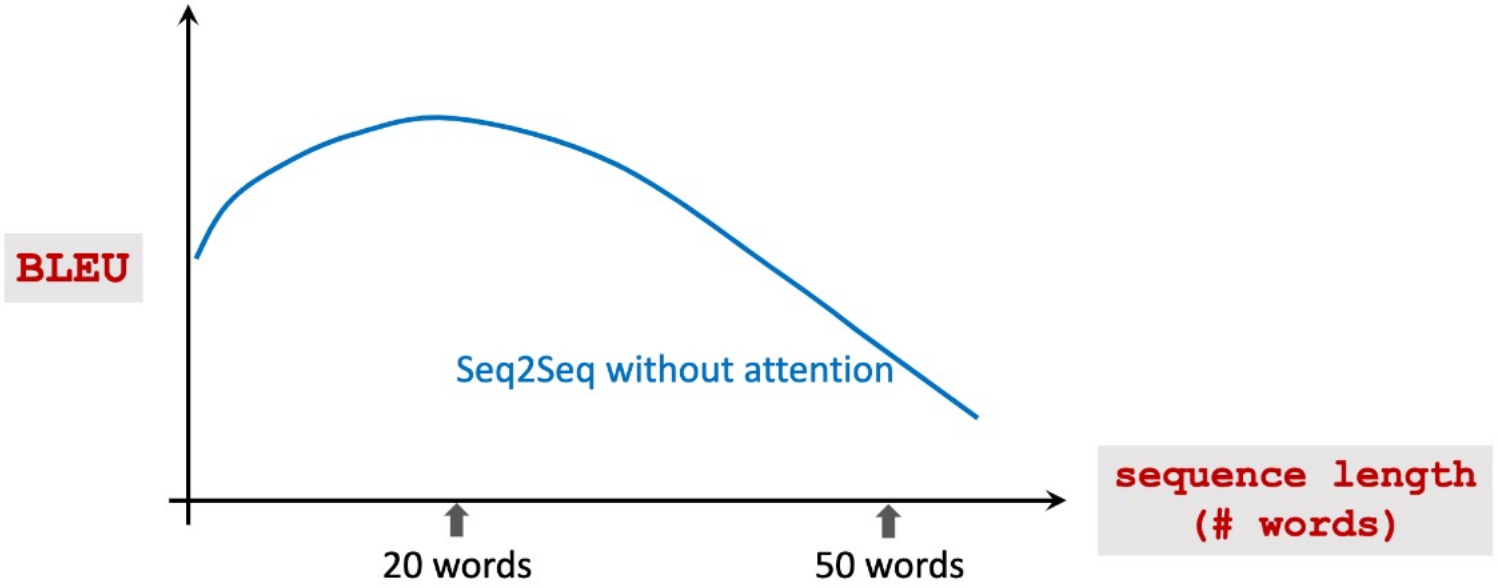


Figure is from the paper " Long Short-Term Memory-Networks for Machine Reading."

Pay attention to the context relevant to the new input

Seq2seq model performance



Seq2seq model performance