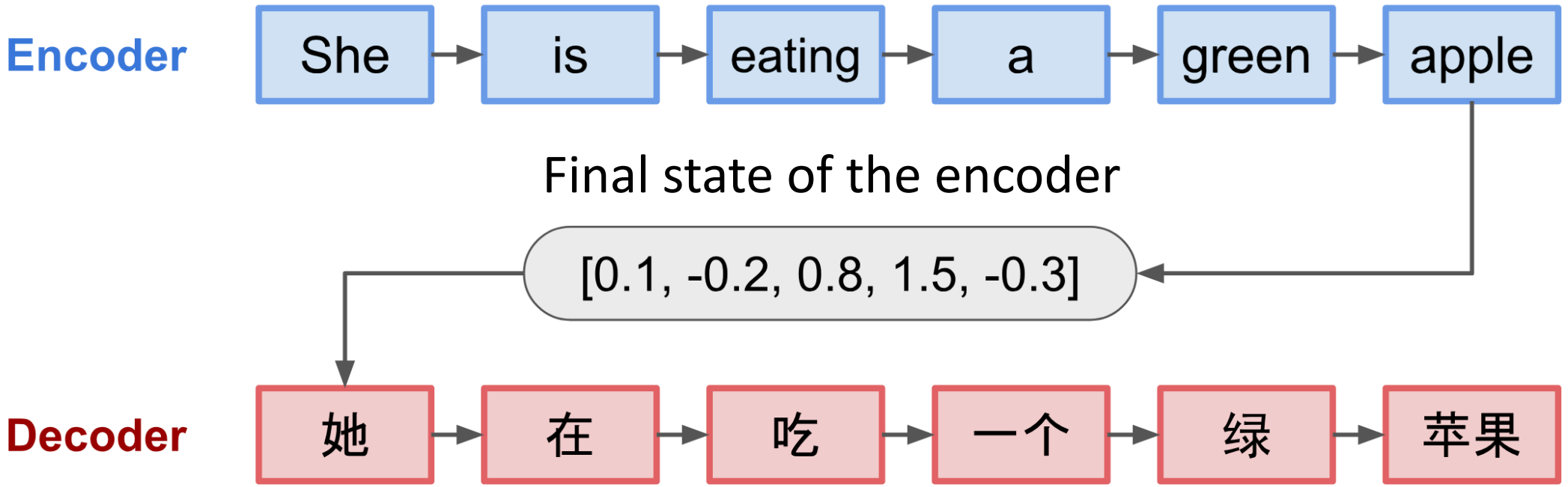
Attention

Shusen Wang

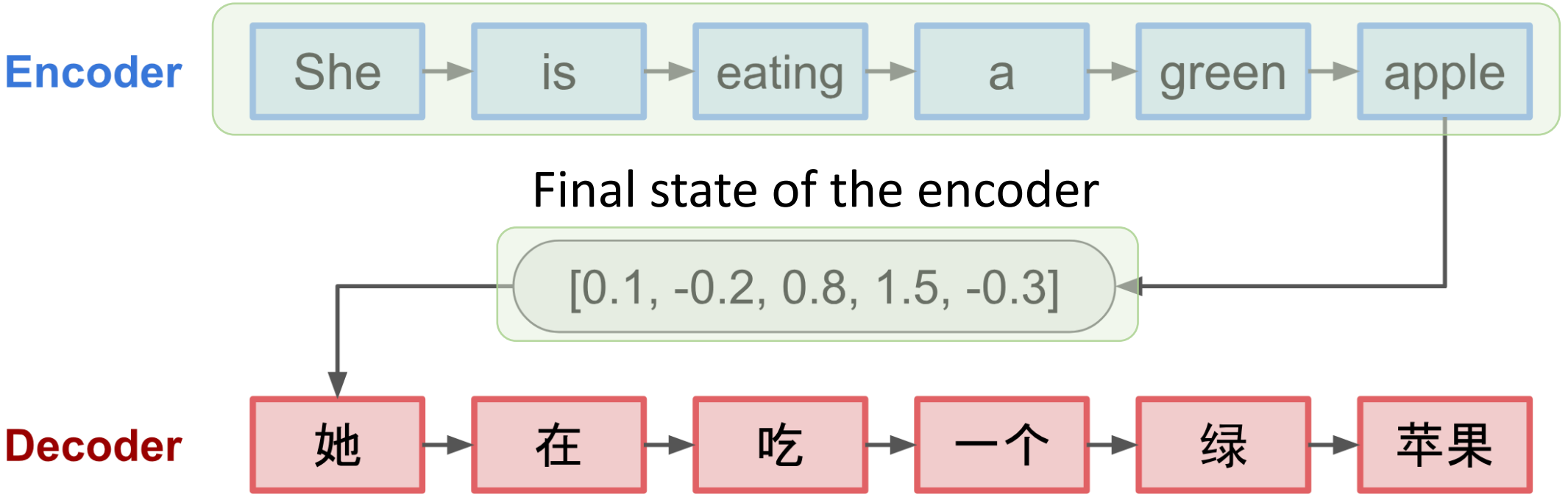
Seq2Seq Model



The figure is from blog lilianweng.github.io

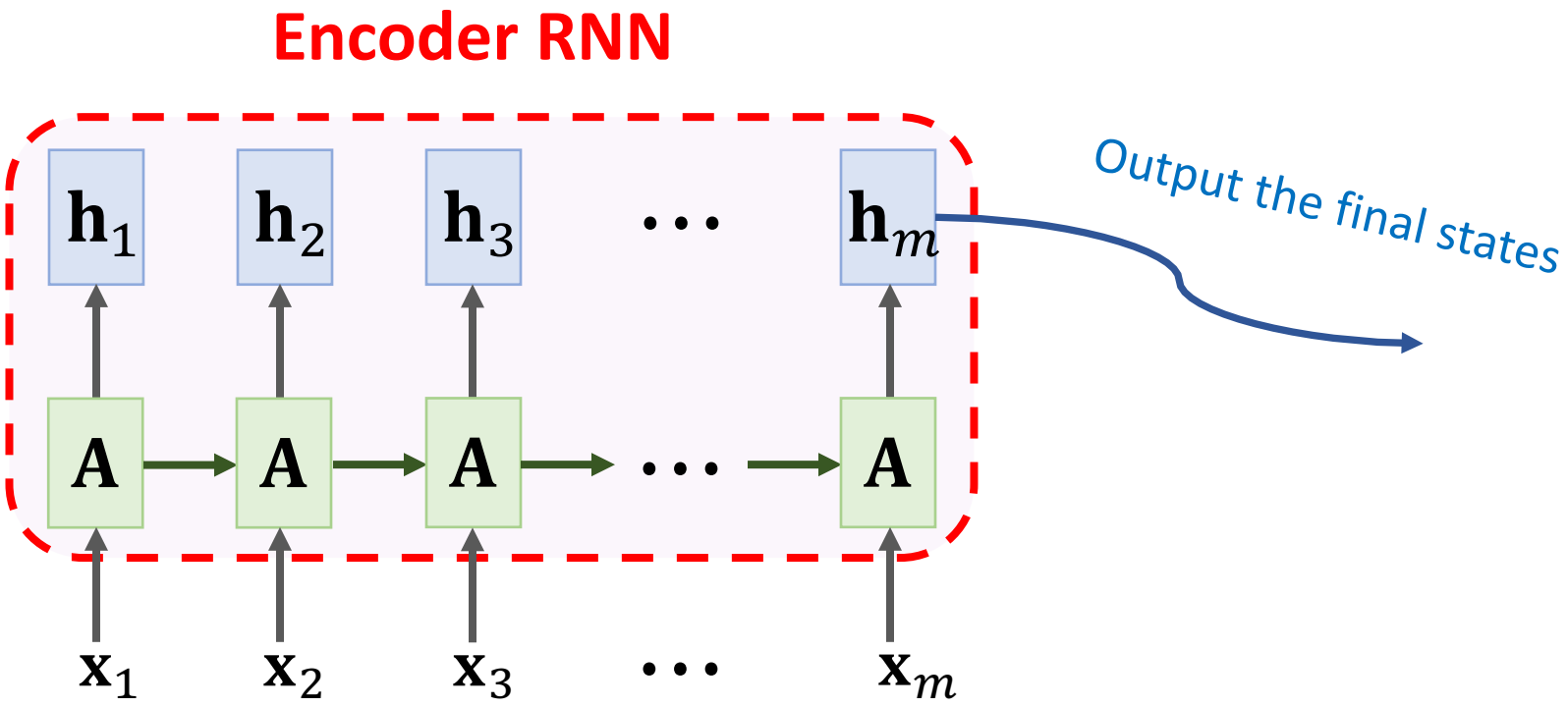
Seq2Seq Model

Shortcoming: The final state is incapable of remembering a **long** sequence.

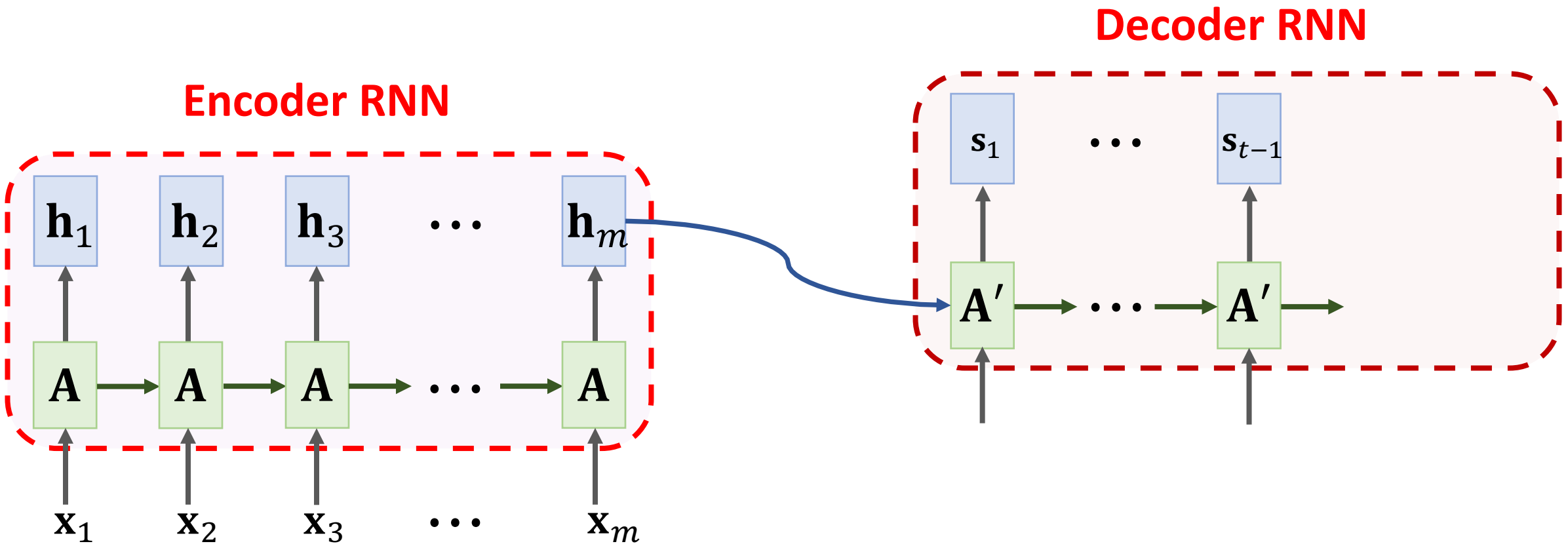


The figure is from blog lilianweng.github.io

Seq2Seq Model



Seq2Seq Model

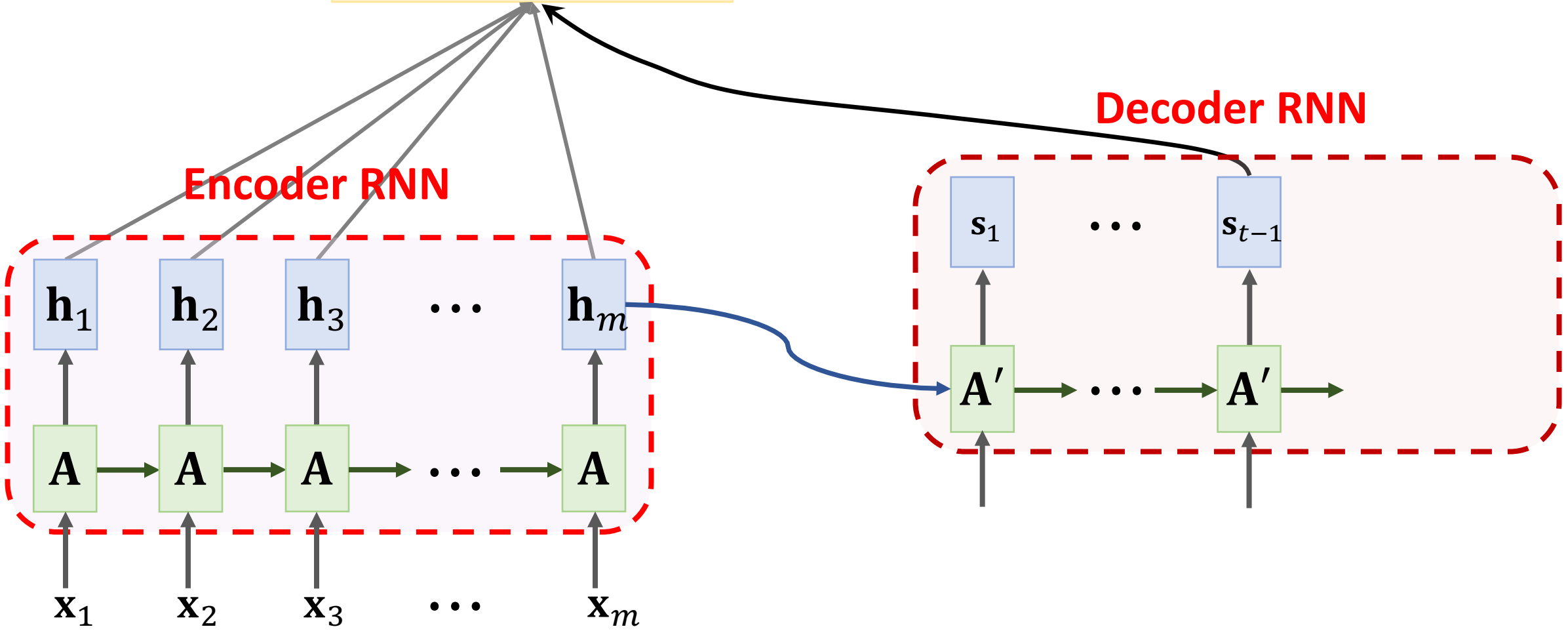


Attention

context vector

$$\mathbf{c} = \sum_{i=1}^{m} \alpha_i \ \mathbf{h}_i.$$

- α_i: similarity between $\mathbf{s}_{t-1}$ and $\mathbf{h}_i$.



Attention

context vector

$$\mathbf{c} = \sum_{i=1}^{m} \alpha_i \ \mathbf{h}_i.$$

- α_i: similarity between $\mathbf{s}_{t-1}$ and $\mathbf{h}_i$.

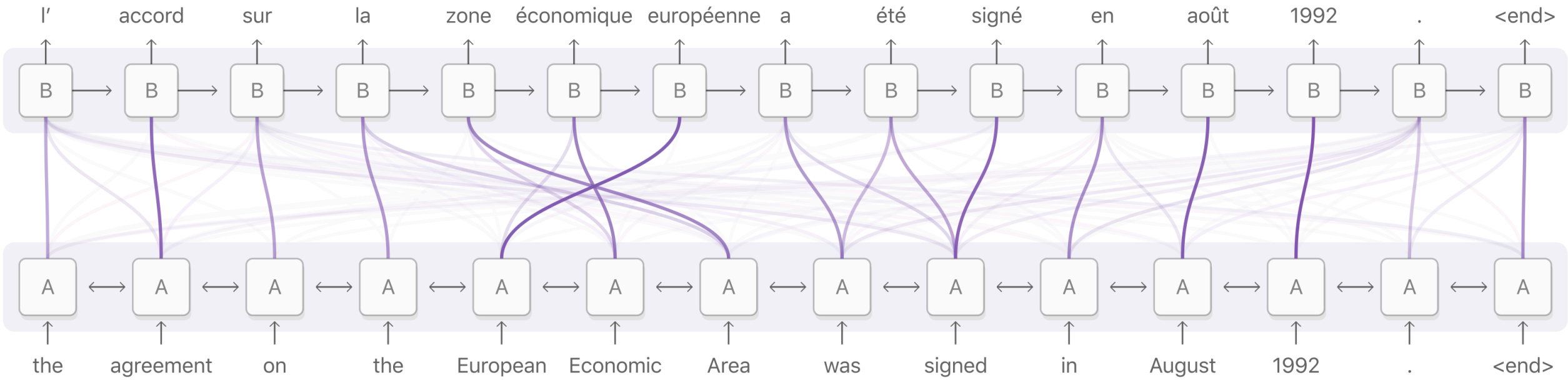


Figure is from https://distill.pub/2016/augmented-rnns/

Attention

context vector

$$\mathbf{c} = \sum_{i=1}^{m} \alpha_i \ \mathbf{h}_i.$$

- α_i: similarity between $\mathbf{s}_{t-1}$ and $\mathbf{h}_i$.

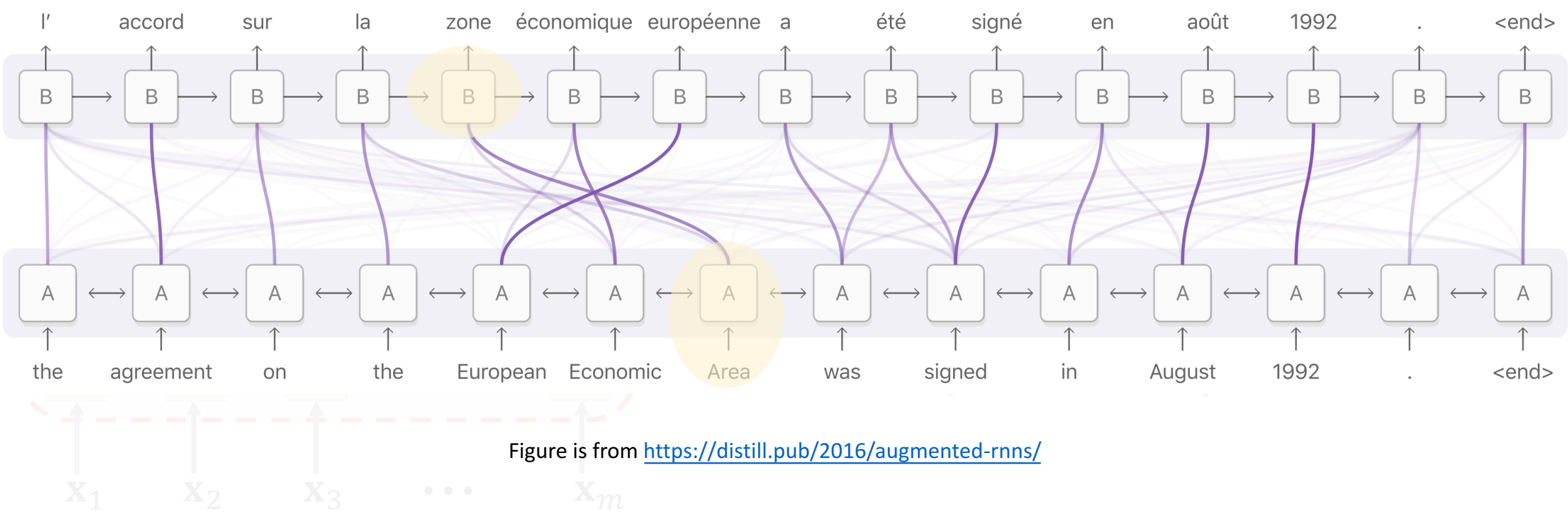


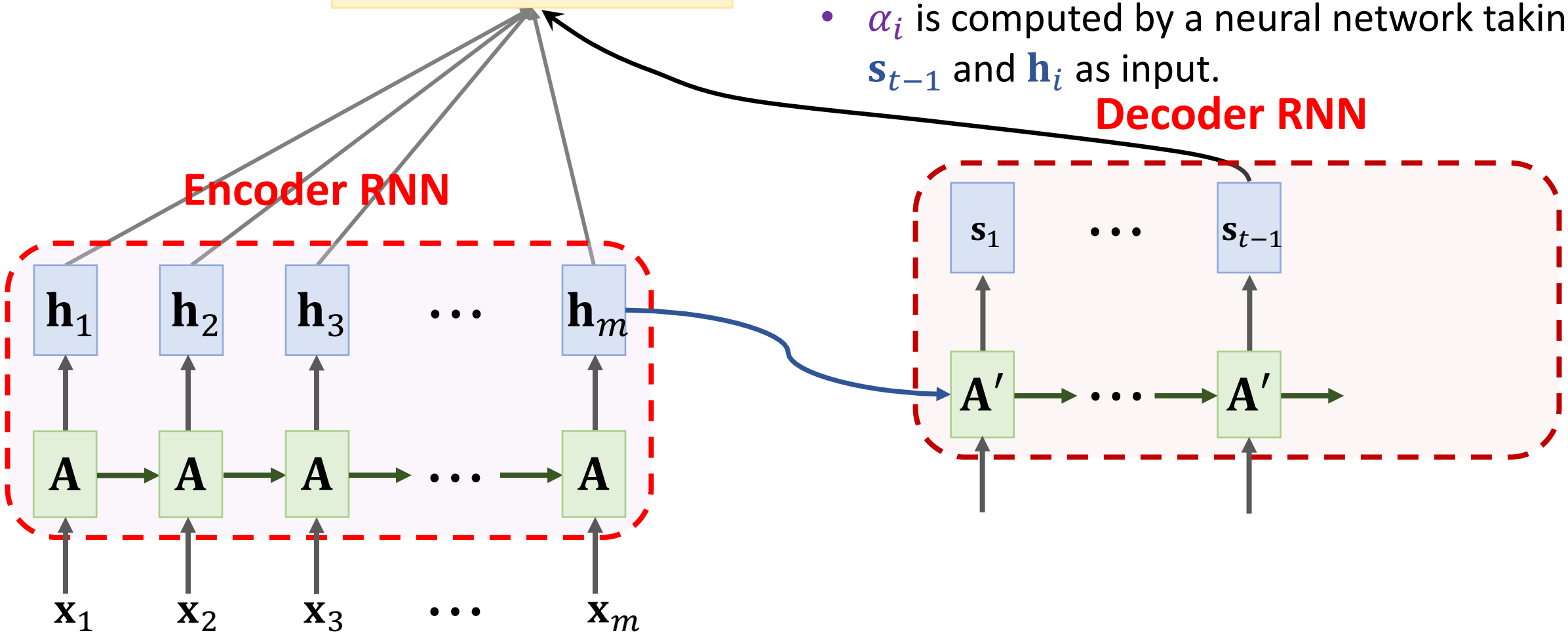
Figure is from https://distill.pub/2016/augmented-rnns/

Attention

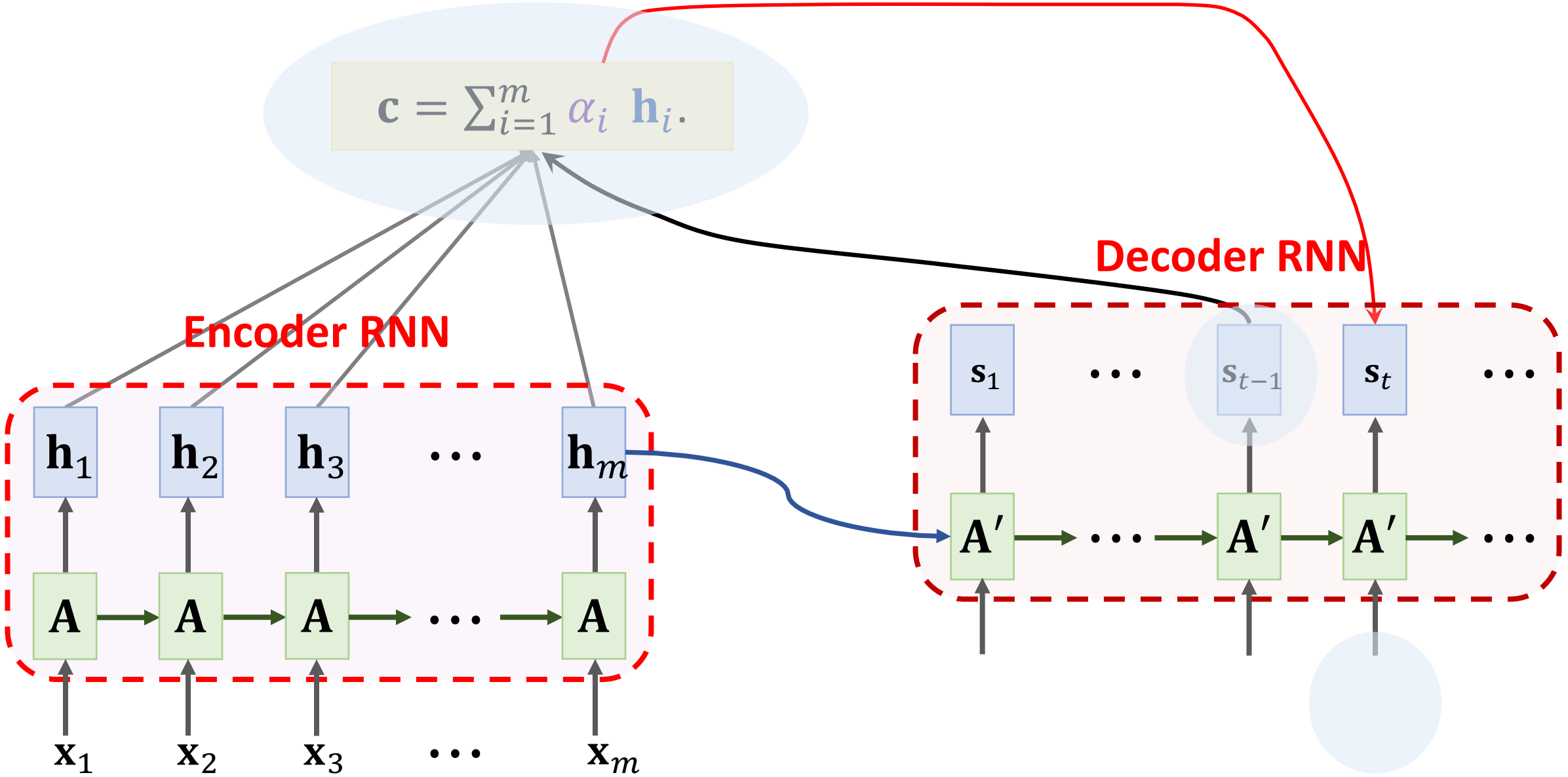
context vector

$$\mathbf{c} = \sum_{i=1}^{m} \alpha_i \ \mathbf{h}_i.$$

- α_i: similarity between $\mathbf{s}_{t-1}$ and $\mathbf{h}_i$.
- α_i is computed by a neural network taking $\mathbf{s}_{t-1}$ and $\mathbf{h}_i$ as input.

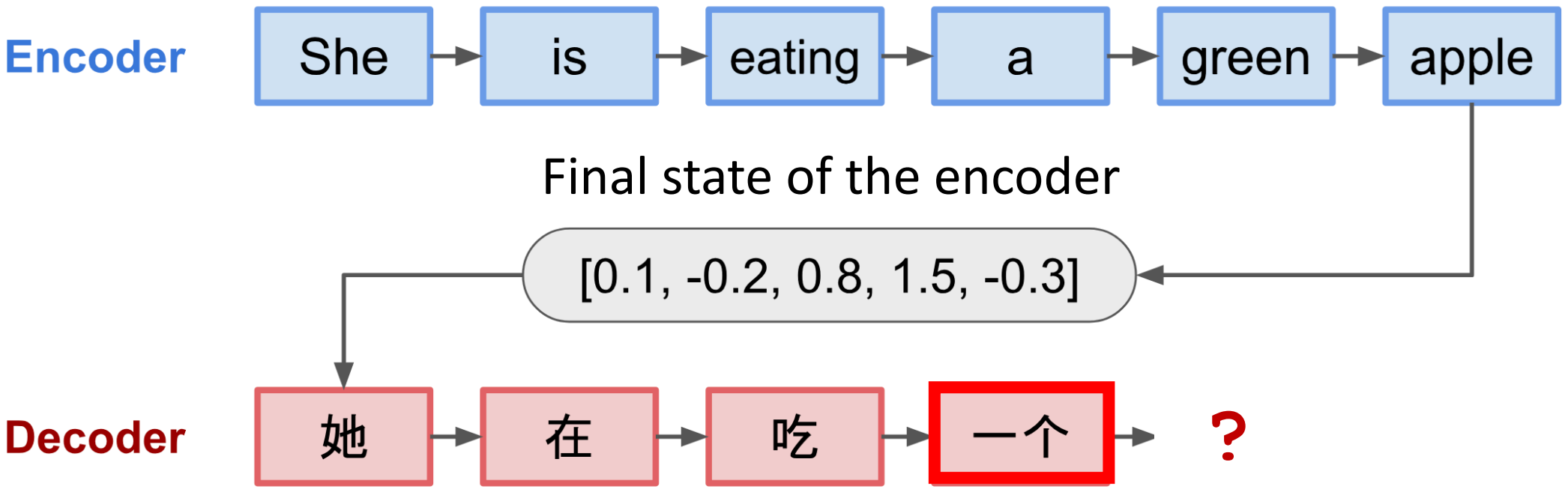


Attention


$\mathbf{c} = \sum_{i=1}^{m} \alpha_i \ \mathbf{h}_i .$
Encoder RNN
Decoder RNN
$\mathbf{h}_1$
$\mathbf{h}_2$
$\mathbf{h}_3$
$\cdots$
$\mathbf{h}_m$
$\mathbf{A}$
$\mathbf{x}_1$
$\mathbf{x}_2$
$\mathbf{x}_3$
$\mathbf{x}_m$
$\mathbf{s}_1$
$\mathbf{s}_{t-1}$
$\mathbf{s}_t$
$\mathbf{A}'$

Summary

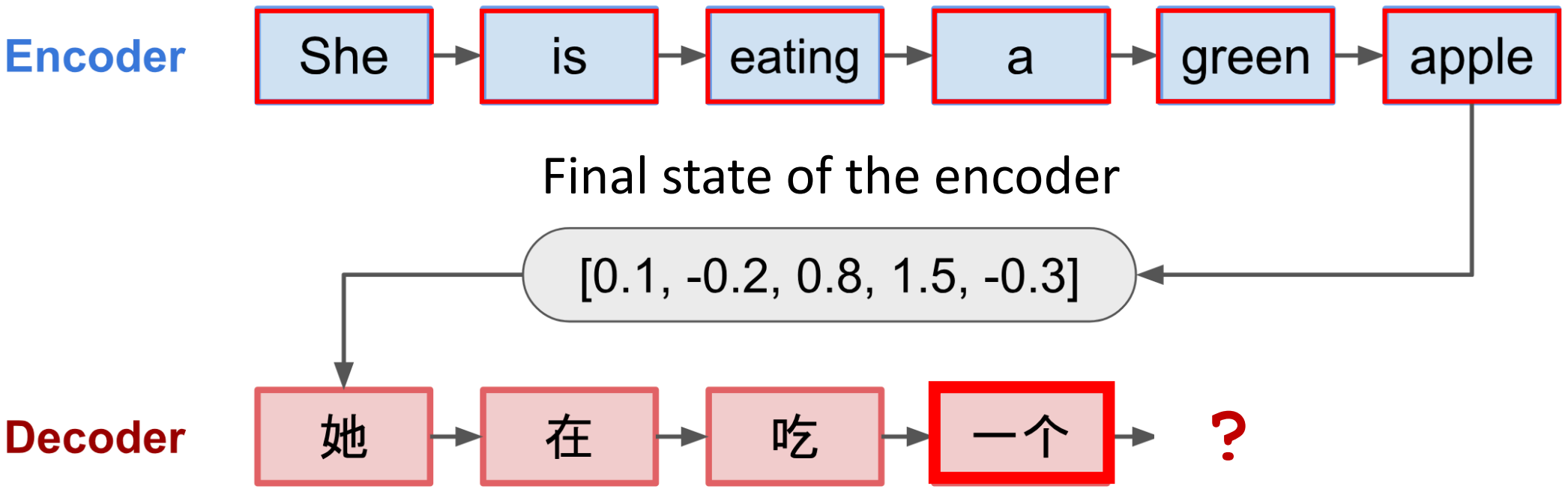
- Standard Seq2Seq model: the decoder looks at only its current state.



The figure is from blog lilianweng.github.io

Summary

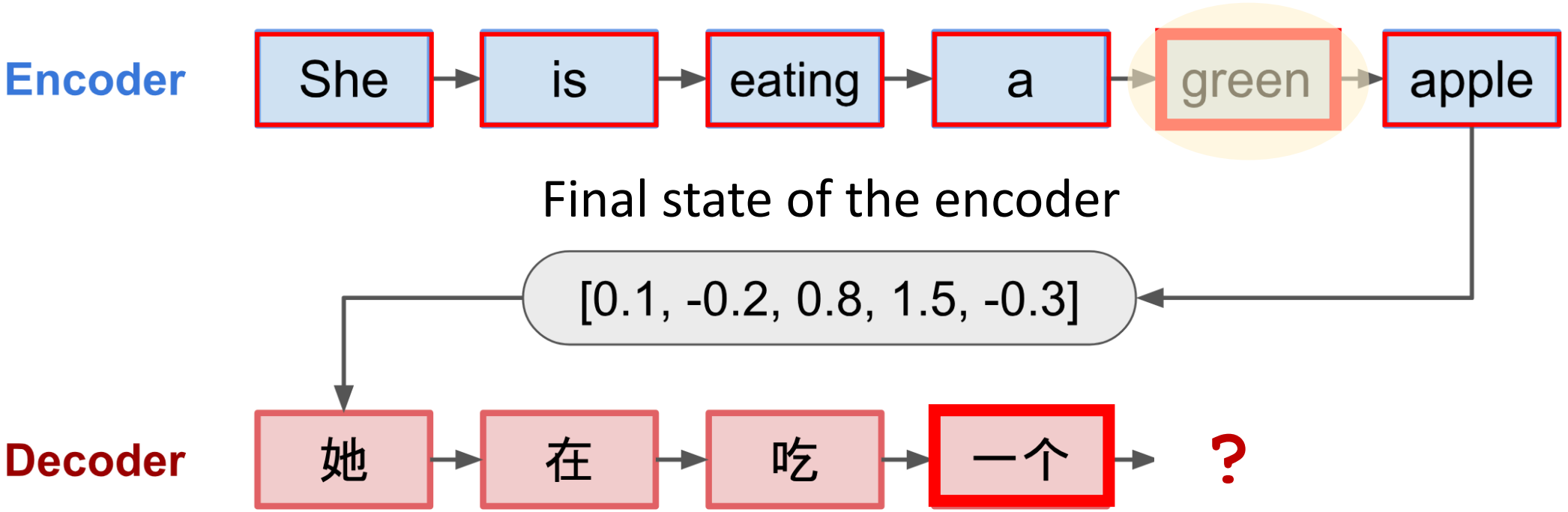
- Standard Seq2Seq model: the decoder looks at only its current state.
- Attention: decoder additionally looks at all the states of the encoder.



The figure is from blog lilianweng.github.io

Summary

- Standard Seq2Seq model: the decoder looks at only its current state.
- Attention: decoder additionally looks at all the states of the encoder.



The figure is from blog lilianweng.github.io

Summary

- Standard Seq2Seq model: the decoder looks at only its current state.
- Attention: decoder additionally looks at all the states of the encoder.

- Downside: higher time complexity.
  - l_1: input sequence length
  - l_2: target sequence length
  - Standard Seq2Seq: $O(l_1 + l_2)$ time complexity
  - Seq2Seq + attention: $O(l_1\ l_2)$ time complexity