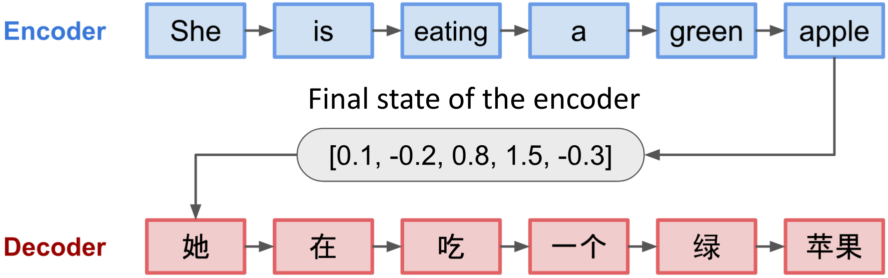
Attention

Shusen Wang

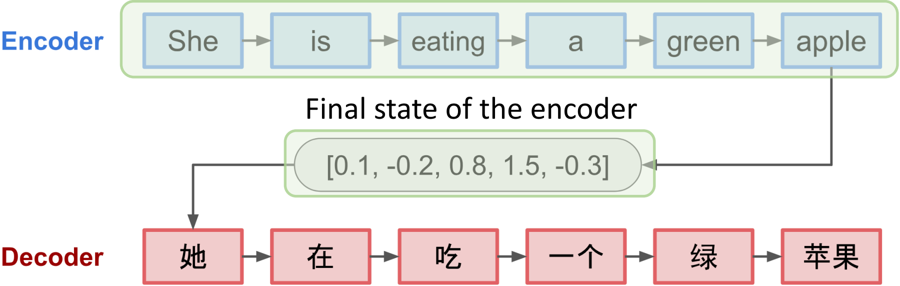
Seq2Seq Model



The figure is from blog lilianweng.github.io

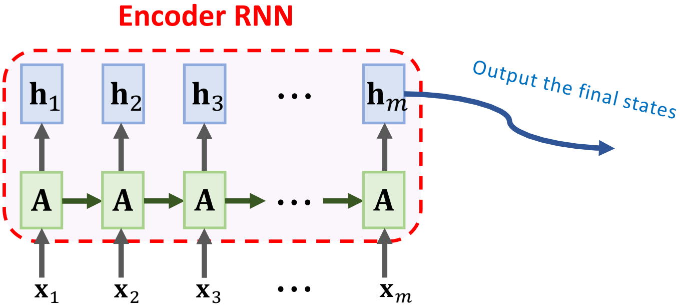
Seq2Seq Model

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

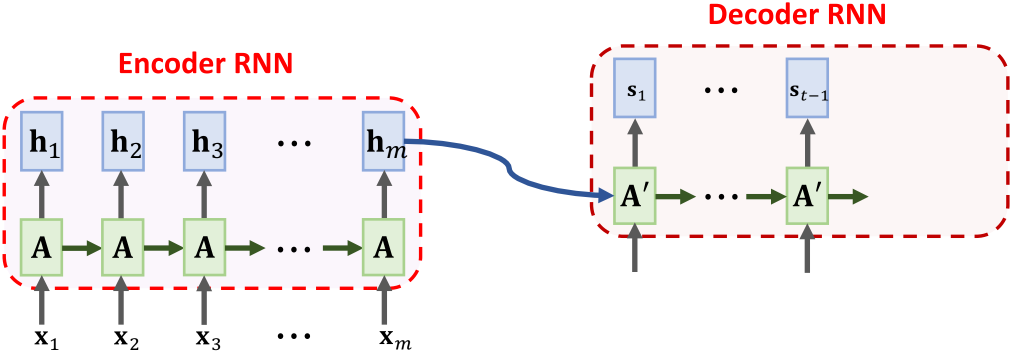


The figure is from blog lilianweng.github.io

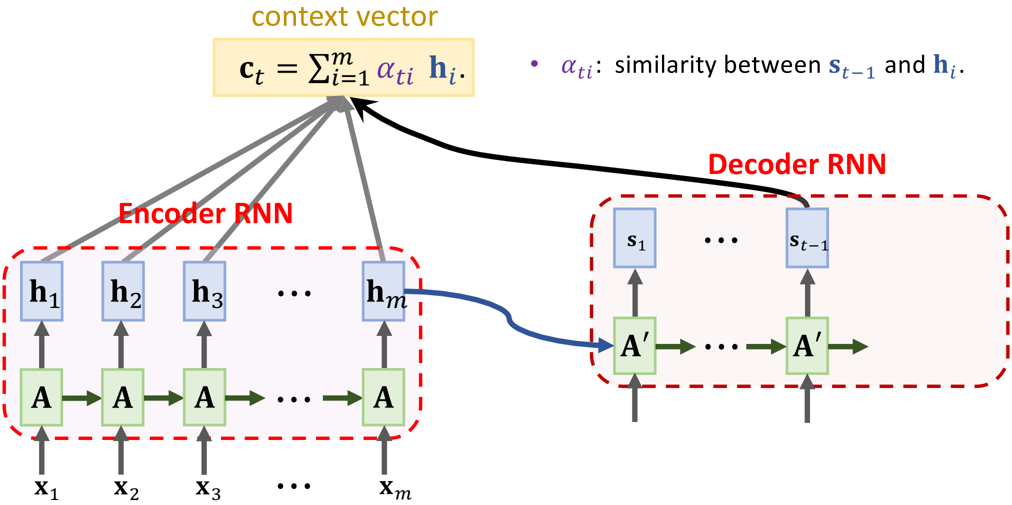
Seq2Seq Model



Seq2Seq Model



Attention


context vector

$\mathbf{c}_t = \sum_{i=1}^{m} \alpha_{ti} \; \mathbf{h}_i .$

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

Attention

context vector

$$\boxed{\mathbf{c}_t = \sum_{i=1}^{m} \alpha_{ti} \, \mathbf{h}_i .}$$

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

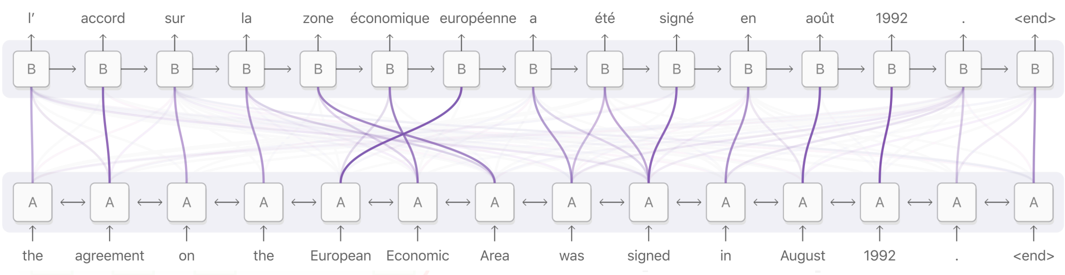


Figure is from https://distill.pub/2016/augmented-rnns/

Attention

context vector

$$\boxed{\mathbf{c}_t = \sum_{i=1}^{m} \alpha_{ti} \ \mathbf{h}_i .}$$

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

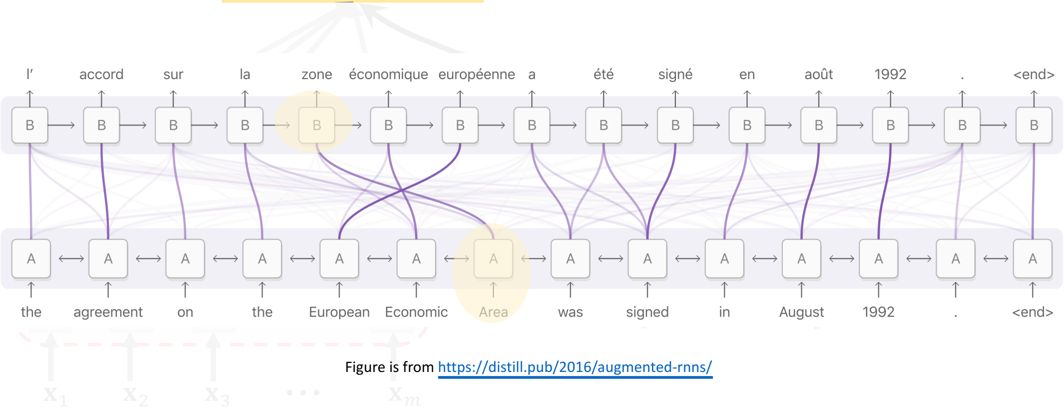
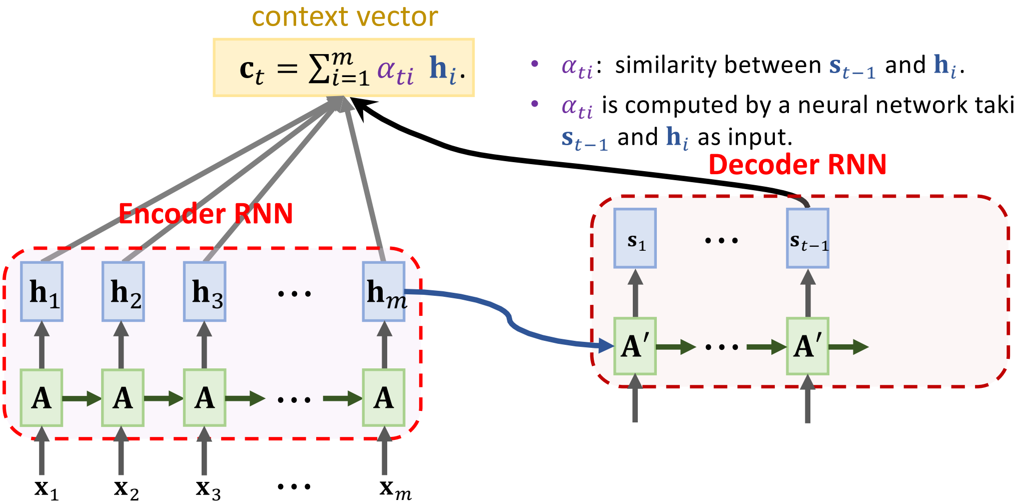


Figure is from https://distill.pub/2016/augmented-rnns/

Attention



context vector

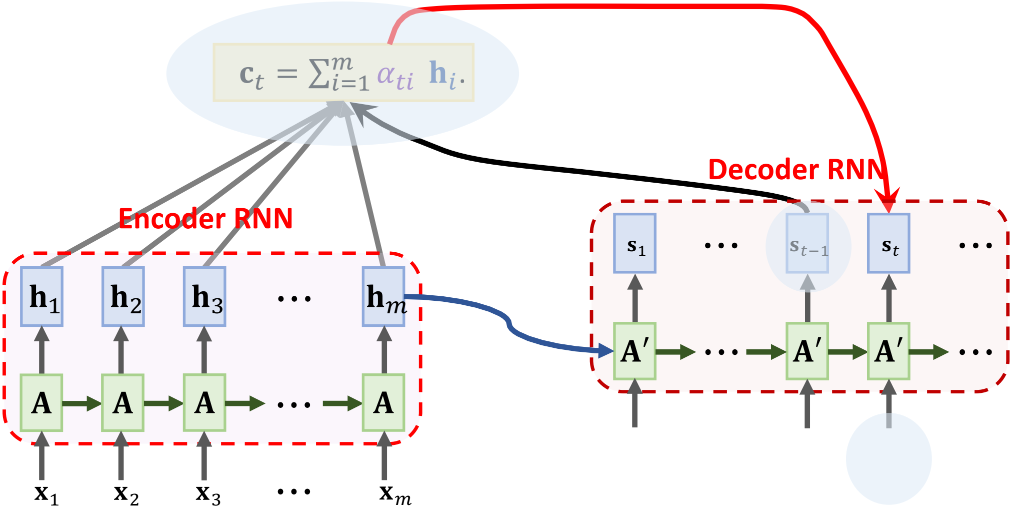
$\mathbf{c}_t = \sum_{i=1}^{m} \alpha_{ti} \, \mathbf{h}_i .$

Encoder RNN

Decoder RNN

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

Attention



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