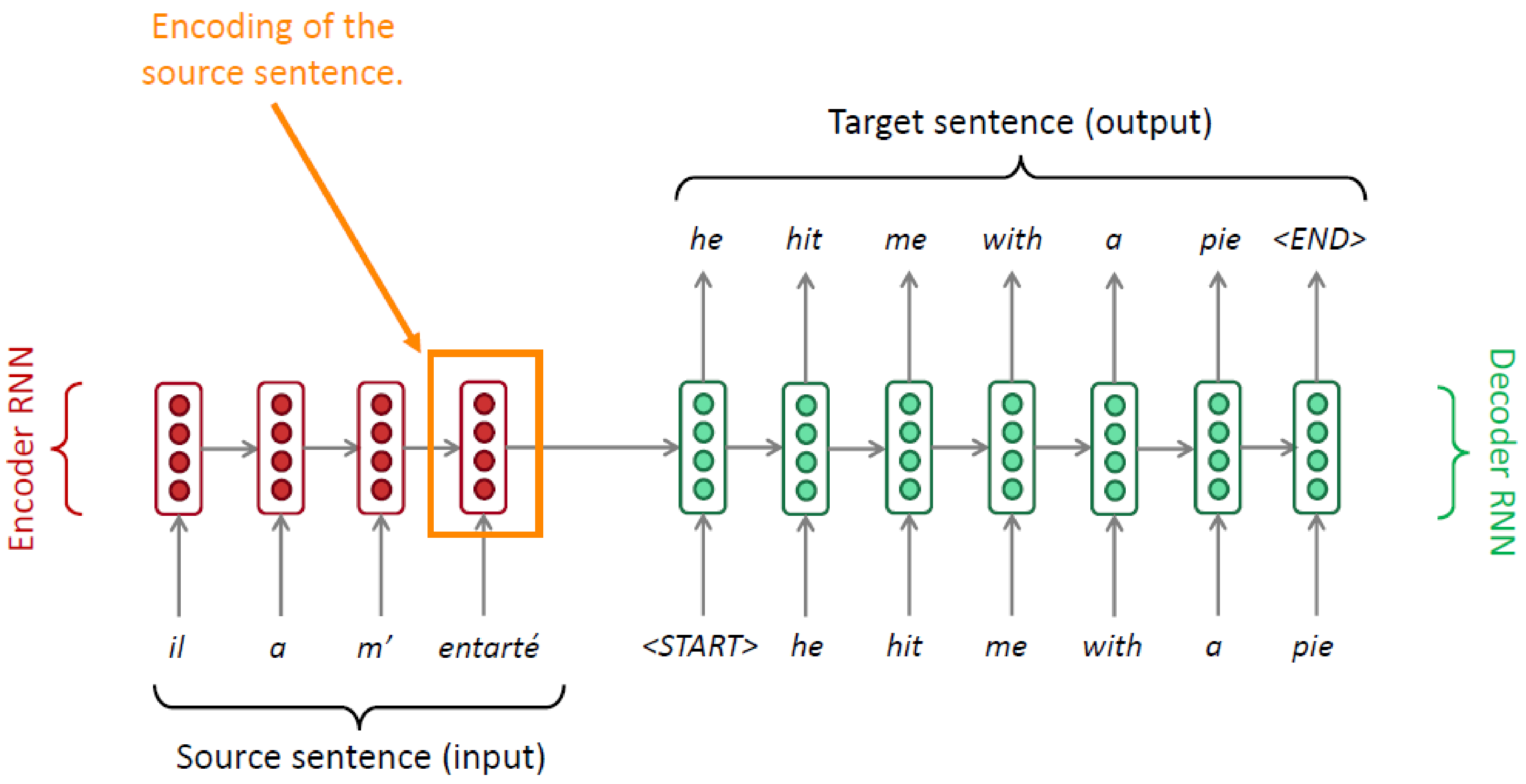
SUNG KYUN KWAN UNIVERSITY(SKKU)

Attention Mechanism

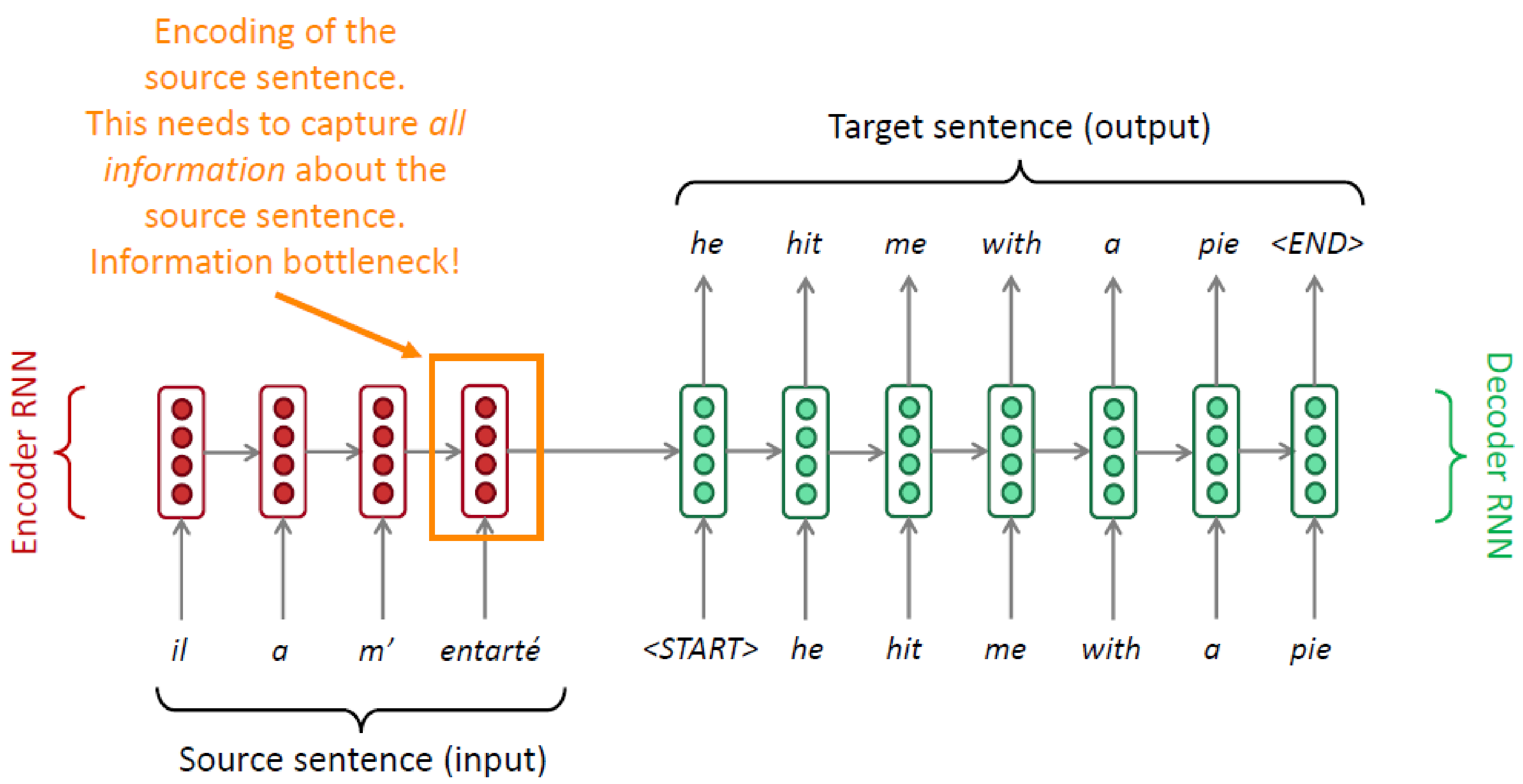
YunSeok Choi
Sungkyunkwan University
Department of Computer Science Education
ys.choi@skku.edu

Seq2Seq: the bottleneck problem



Problems with this architecture?

Seq2Seq: the bottleneck problem



Seq2Seq: the bottleneck problem

- We do not want to collapse them into a single vector
  - Collapsing often corresponds to information loss
  - Increasingly more difficult to encode the entire source sentence into a single vector, as the sentence length increases [Cho'14b]
  - When collapsed, the system fails to translate a long sentence correctly

Source: *An admitting privilege is the right of a doctor to admit a patient to a hospital or a medical centre to carry out a diagnosis or a procedure, based on his status as a health care worker at a hospital.*

When collapsed: *Un privilège d'admission est le droit d'un médecin de reconnaître un patient à l'hôpital ou un centre médical d'un diagnostic ou de prendre un diagnostic en fonction de son état de santé.*

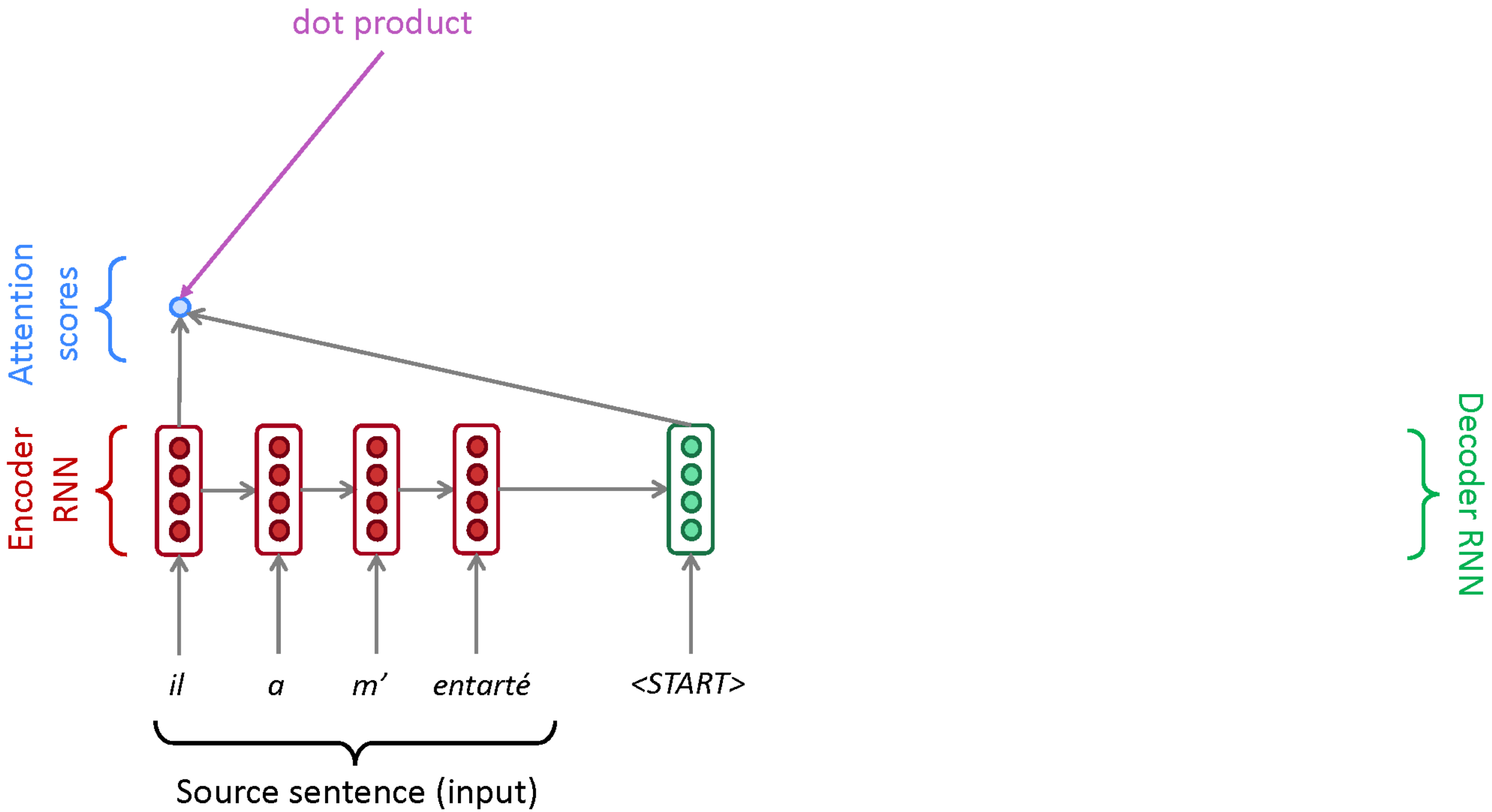
[Cho'14b]: https://www.aclweb.org/anthology/W14-4012.pdf

Attention

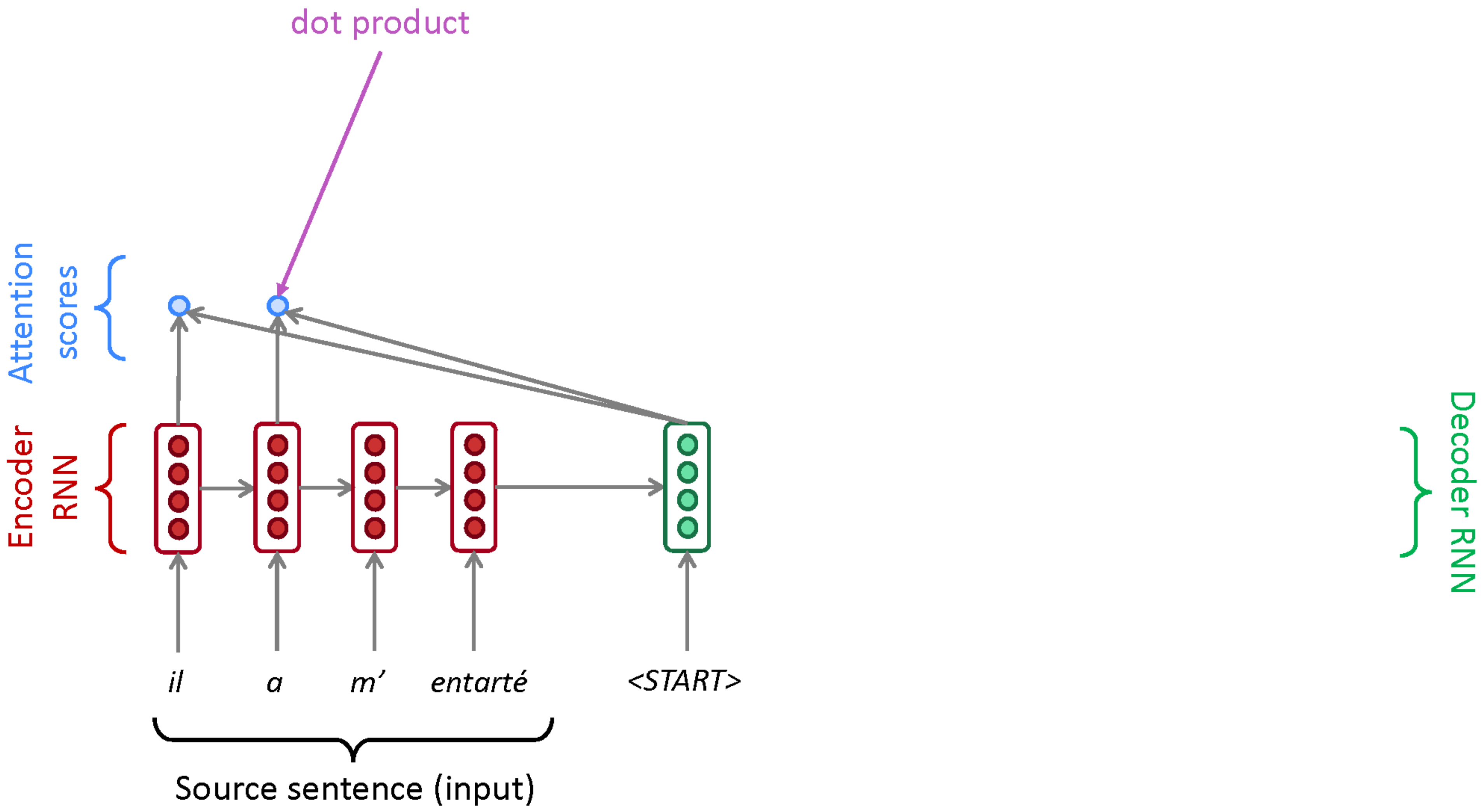
- Attention provides a solution to the bottleneck problem

- Core idea: on each step of the decoder, use *direct connection to the encoder* to *focus on a particular part* of the source sequence

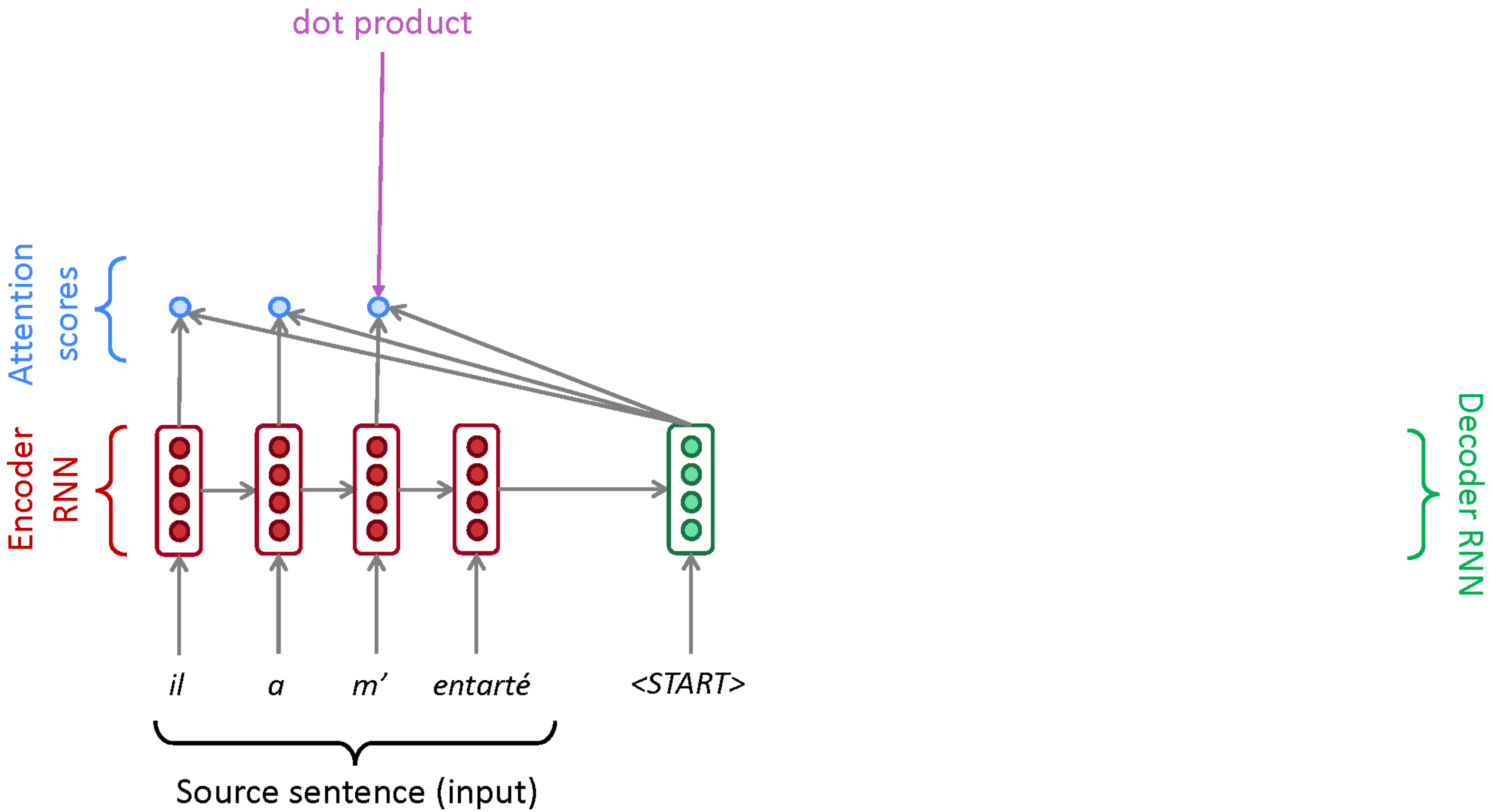
Sequence-to-sequence with attention



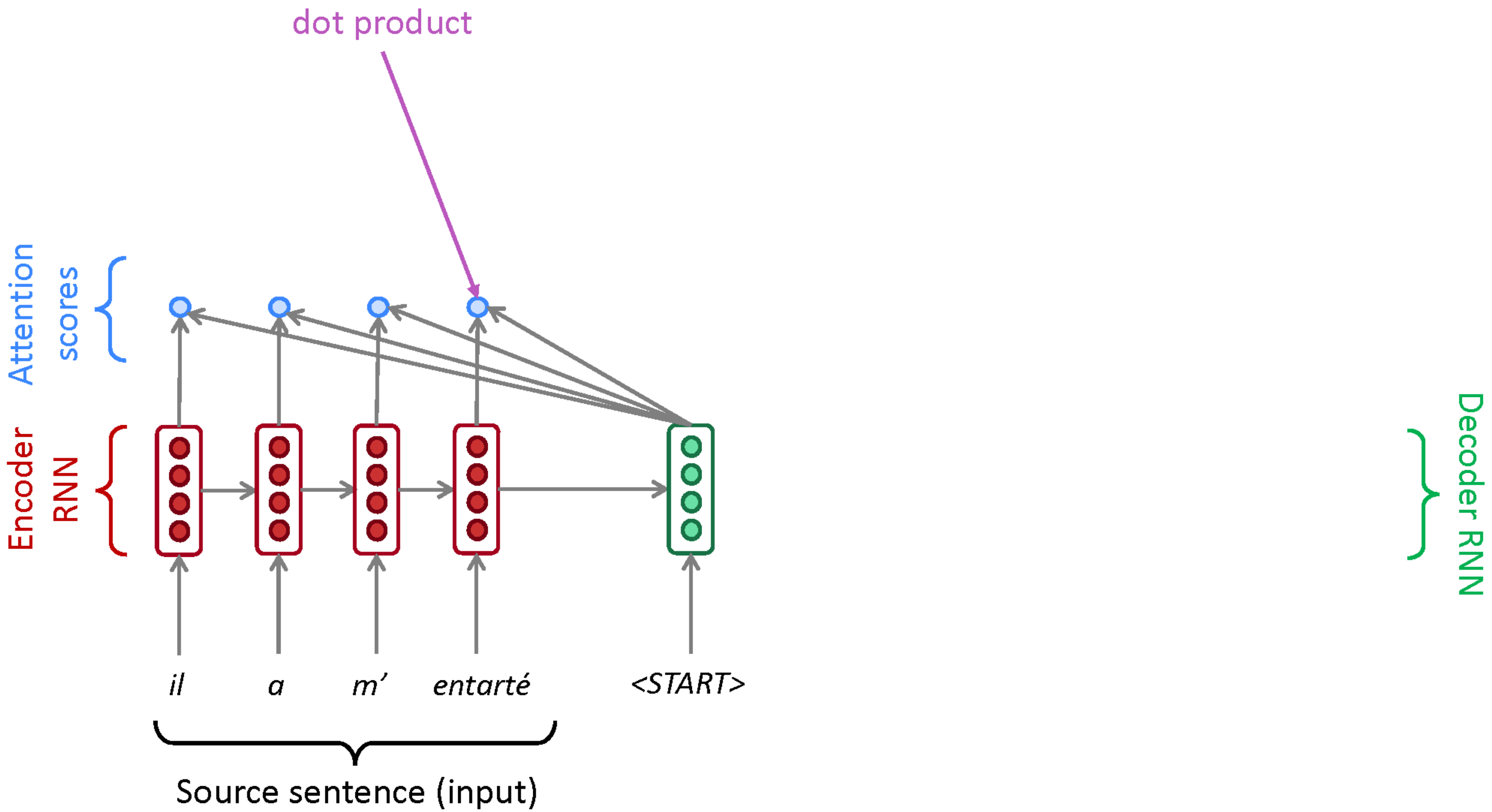
Sequence-to-sequence with attention



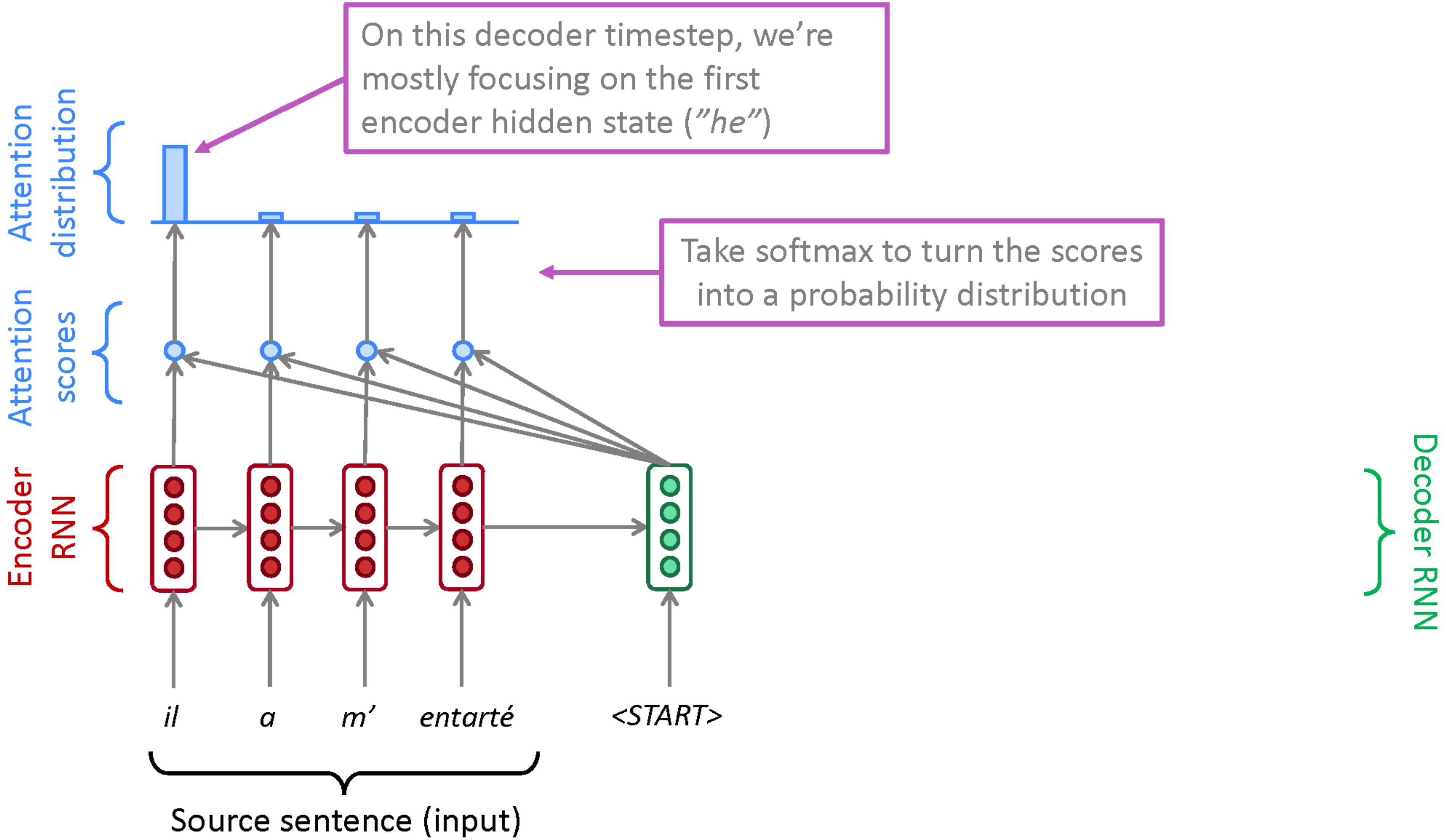
Sequence-to-sequence with attention



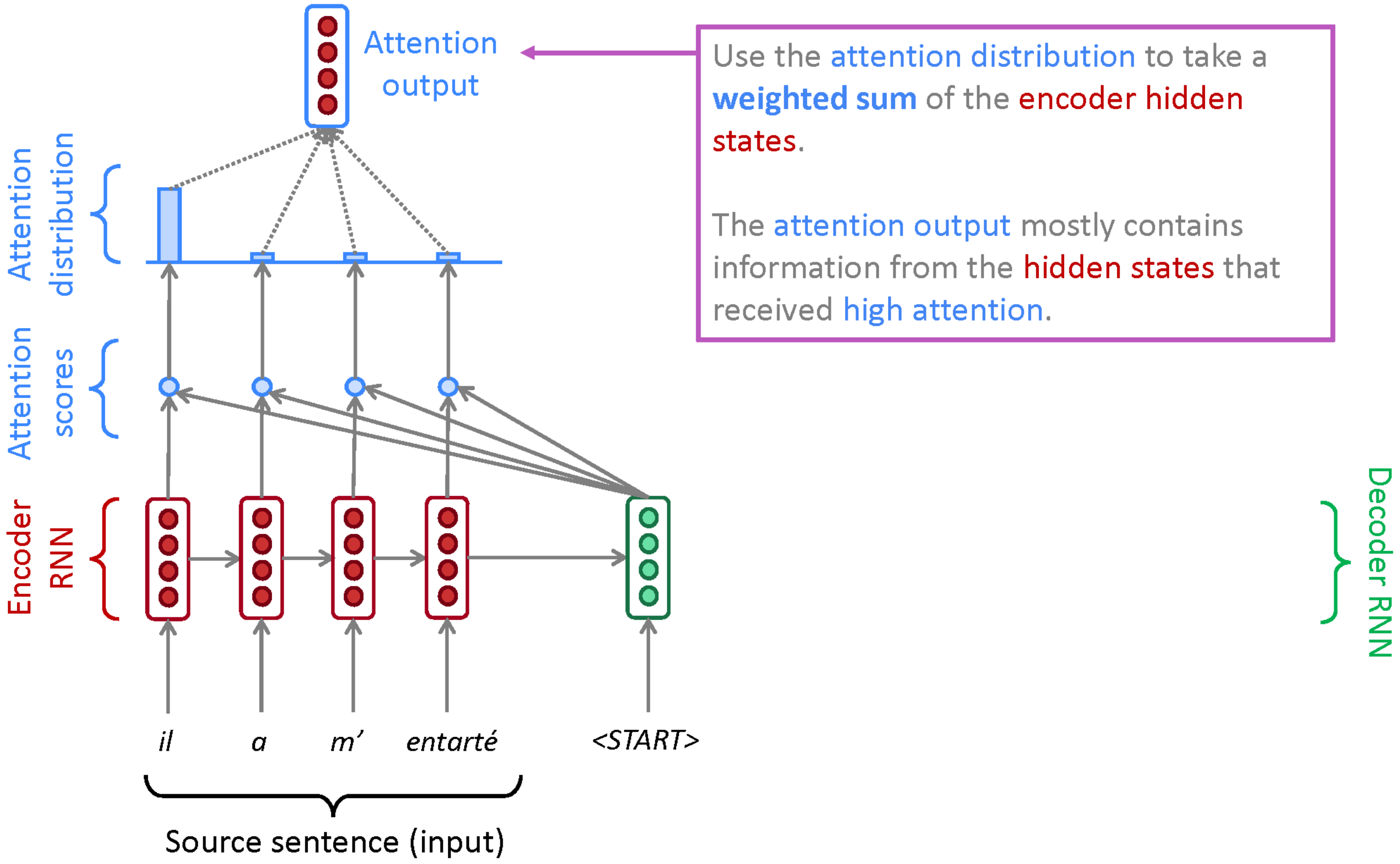
Sequence-to-sequence with attention



Sequence-to-sequence with attention



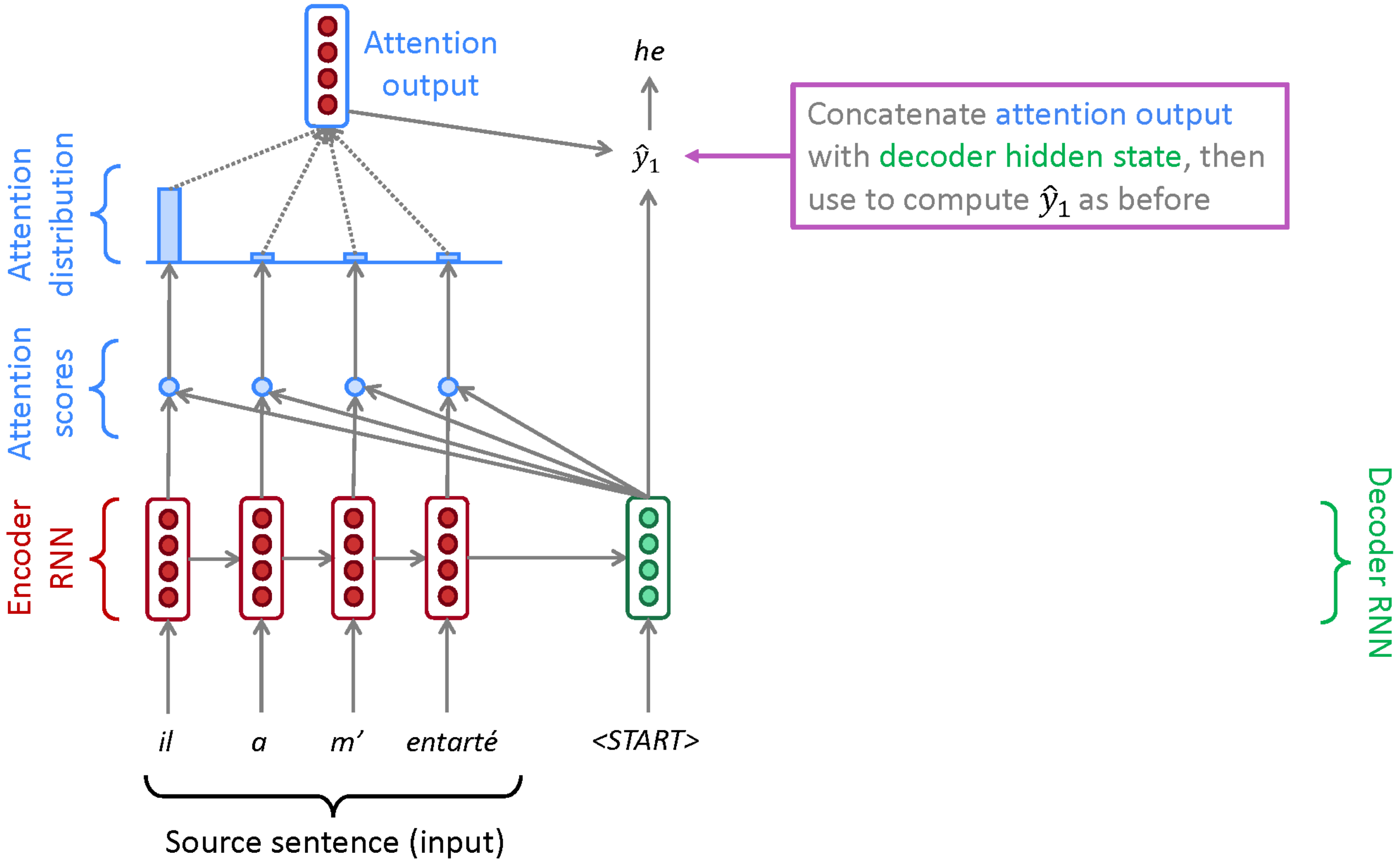
Sequence-to-sequence with attention



Use the attention distribution to take a **weighted sum** of the encoder hidden states.

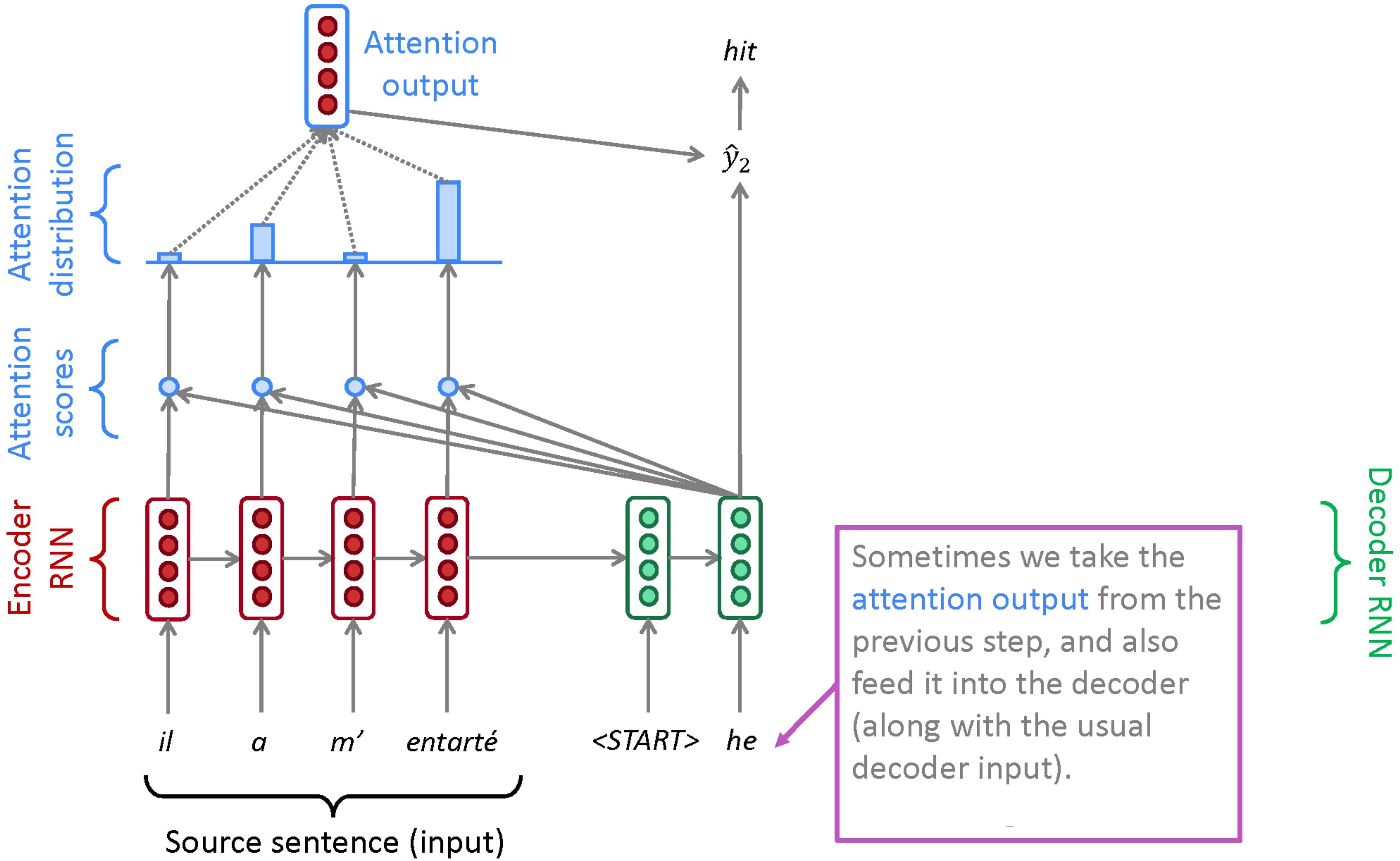
The attention output mostly contains information from the hidden states that received high attention.

Sequence-to-sequence with attention



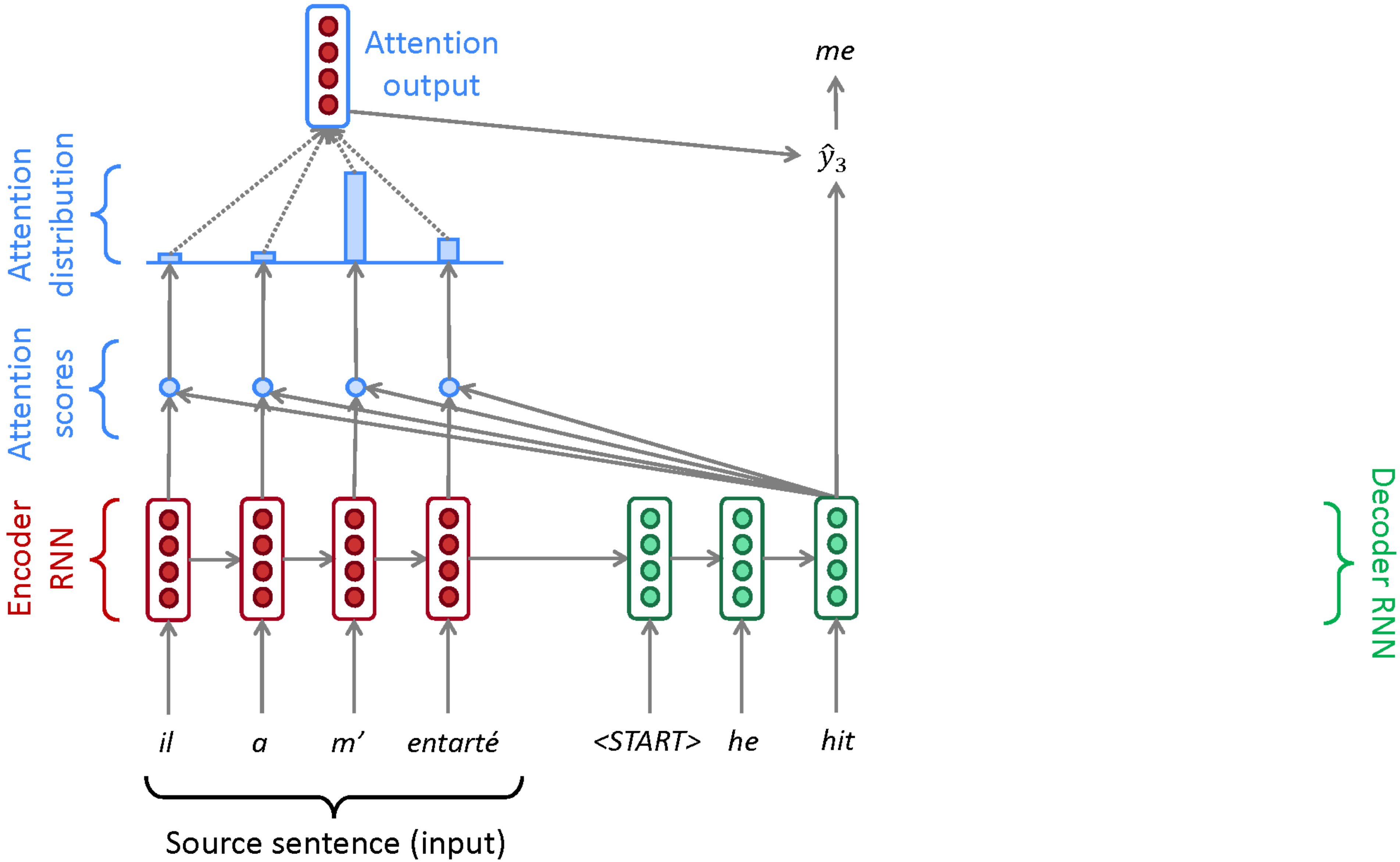
Concatenate attention output with decoder hidden state, then use to compute $\hat{y}_1$ as before

Sequence-to-sequence with attention

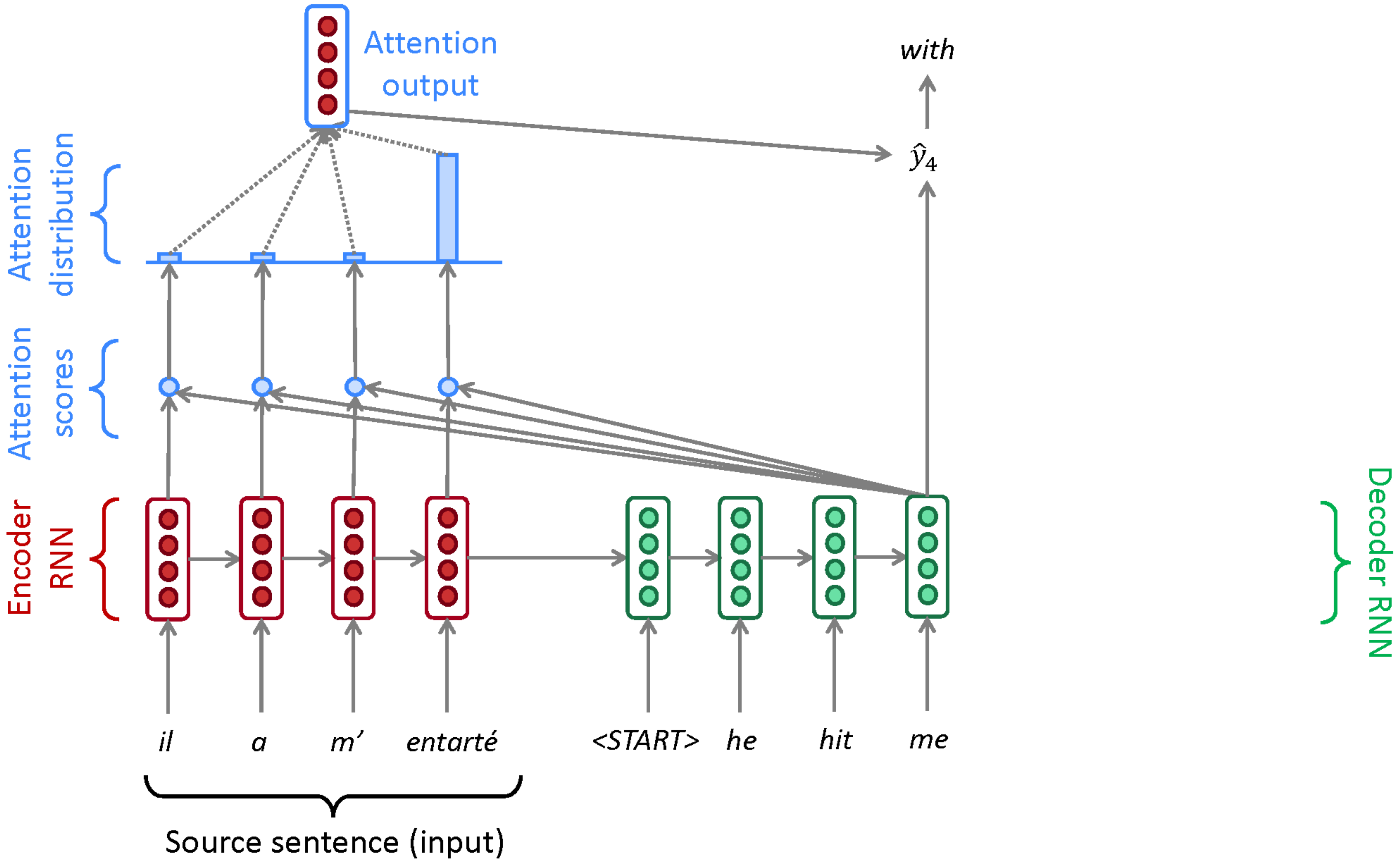


Sometimes we take the attention output from the previous step, and also feed it into the decoder (along with the usual decoder input).

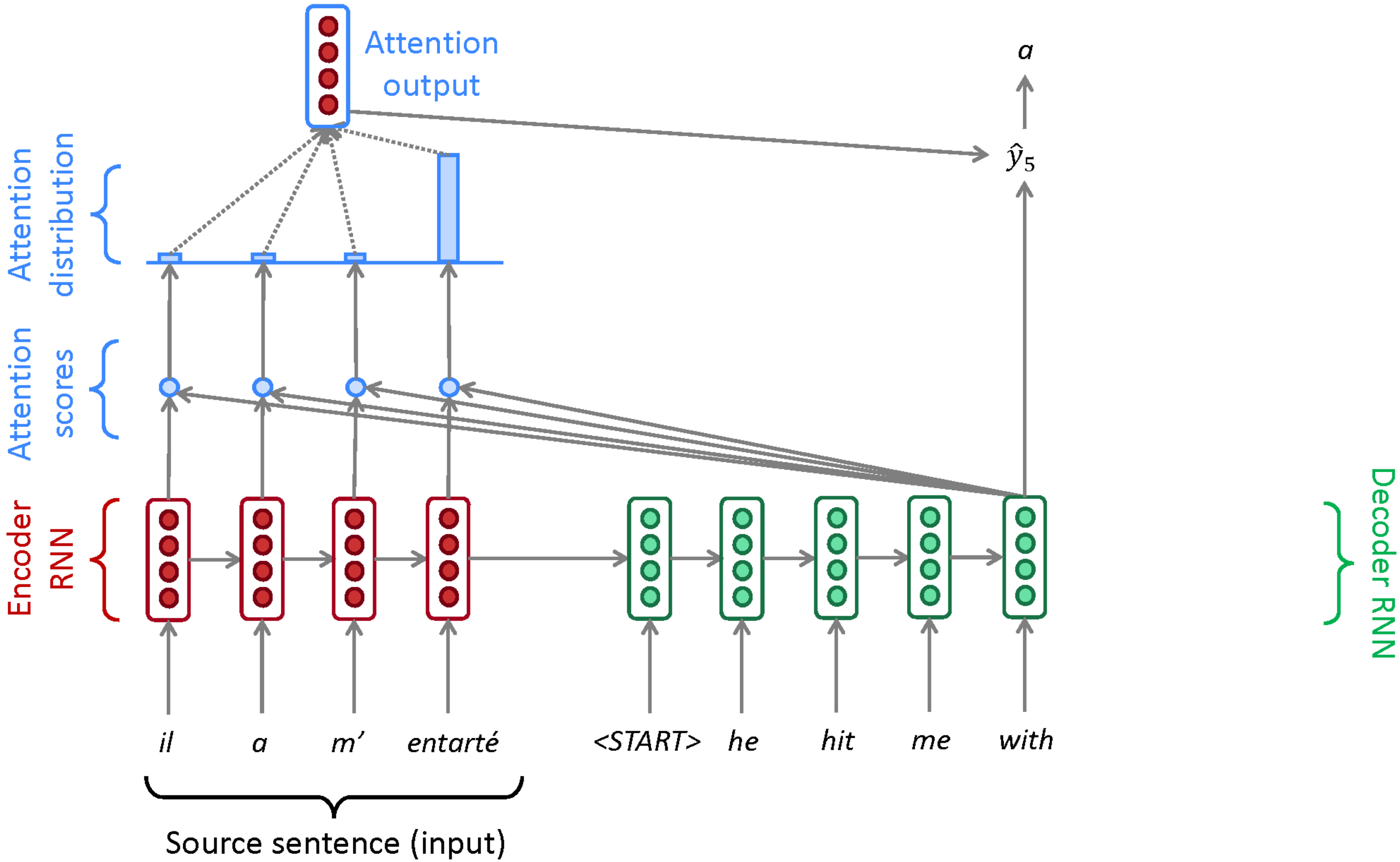
Sequence-to-sequence with attention



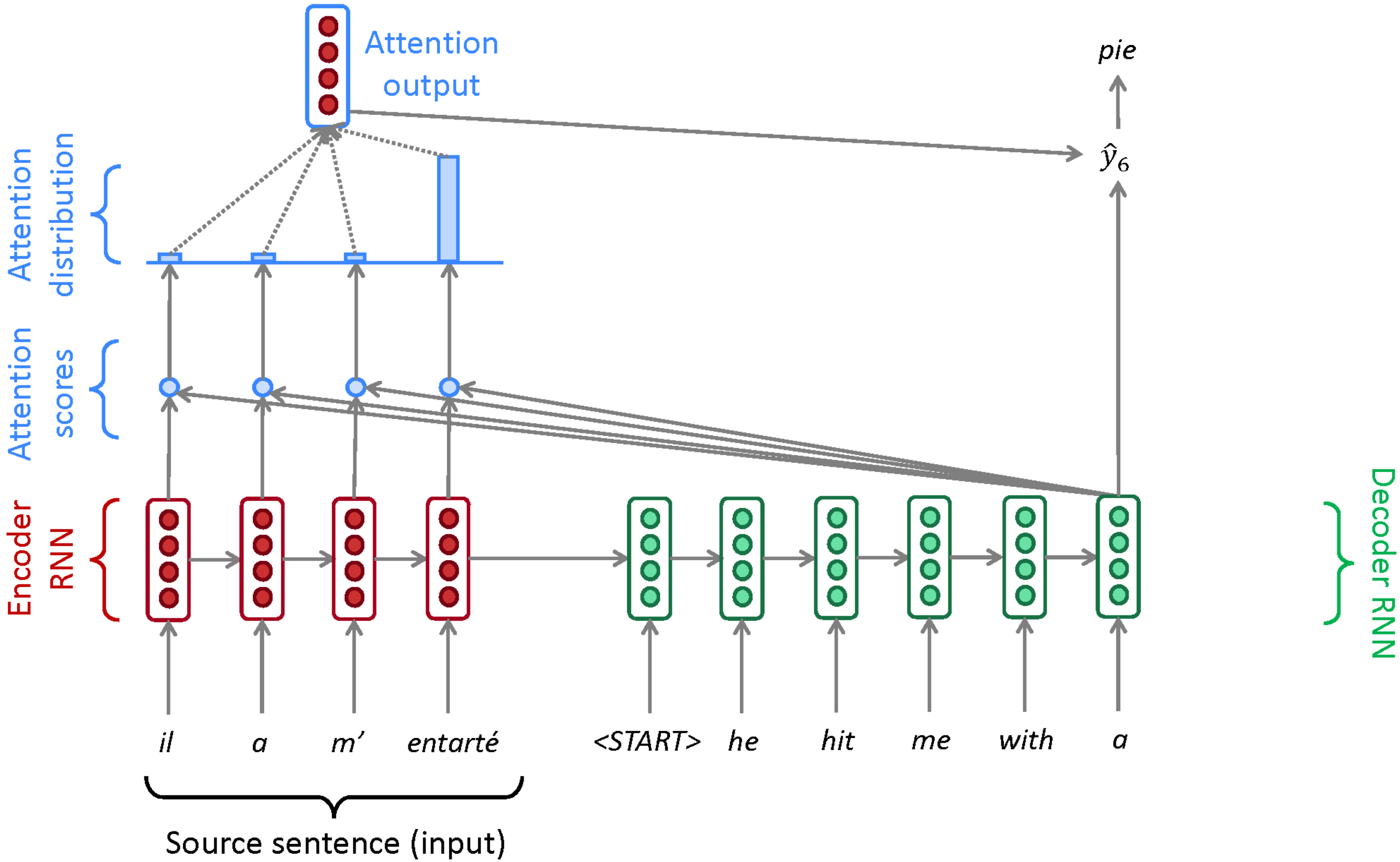
Sequence-to-sequence with attention



Sequence-to-sequence with attention



Sequence-to-sequence with attention



Attention: in equations

- We have encoder hidden states

$$h_1, \ldots, h_N \in \mathbb{R}^h$$

- On timestep t, we have decoder hidden state

$$s_t \in \mathbb{R}^h$$

- We get the attention scores e^t for this step:

$$\boldsymbol{e}^t = [\boldsymbol{s}_t^T \boldsymbol{h}_1, \ldots, \boldsymbol{s}_t^T \boldsymbol{h}_N] \in \mathbb{R}^N$$

Attention: in equations

- We take softmax to get the attention distribution α^t for this step (this is a probability distribution and sums to 1)

$$\alpha^t = \mathrm{softmax}(\boldsymbol{e}^t) \in \mathbb{R}^N$$

- We use α^t to take a weighted sum of the encoder hidden states to get the attention output α^t

$$\boldsymbol{a}_t = \sum_{i=1}^{N} \alpha_i^t \boldsymbol{h}_i \in \mathbb{R}^h$$

Attention: in equations

- Finally we concatenate the attention output α^t with the decoder hidden state s^t and proceed as in the non-attention seq2seq model

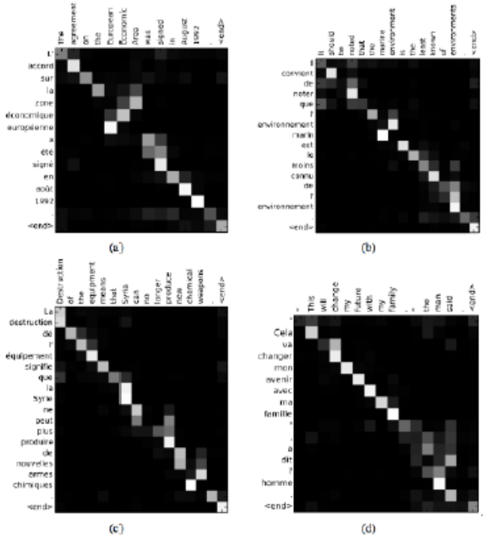
$$[\boldsymbol{a}_t; \boldsymbol{s}_t] \in \mathbb{R}^{2h}$$

Attention is great

- Attention significantly improves NMT performance
  - It’s very useful to allow decoder to focus on certain parts of the source
- Attention solves the bottleneck problem
  - Attention allows decoder to look directly at source; bypass bottleneck
- Attention helps with vanishing gradient problem
  - Provides shortcut to faraway states

Attention is great

- Attention provides some interpretability
  - By inspecting attention distribution, we can see what the decoder was focusing on
  - We get (soft) alignment for free!
  - This is cool because we never explicitly trained an alignment system
  - The network just learned alignment by itself

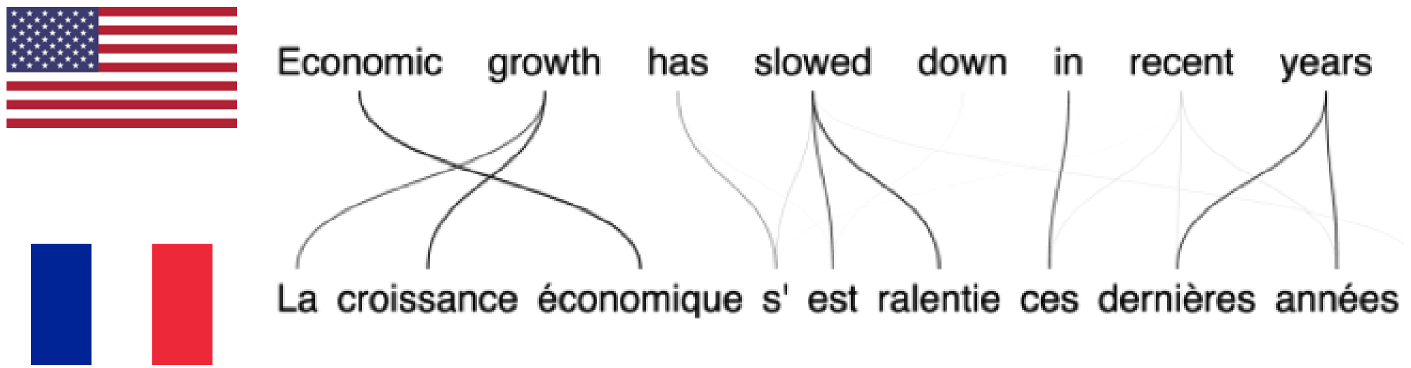


Remind

SUNG KYUN KWAN UNIVERSITY(SKKU)

Attention Model

- Observation
  - At every step, all the inputs are not equally useful

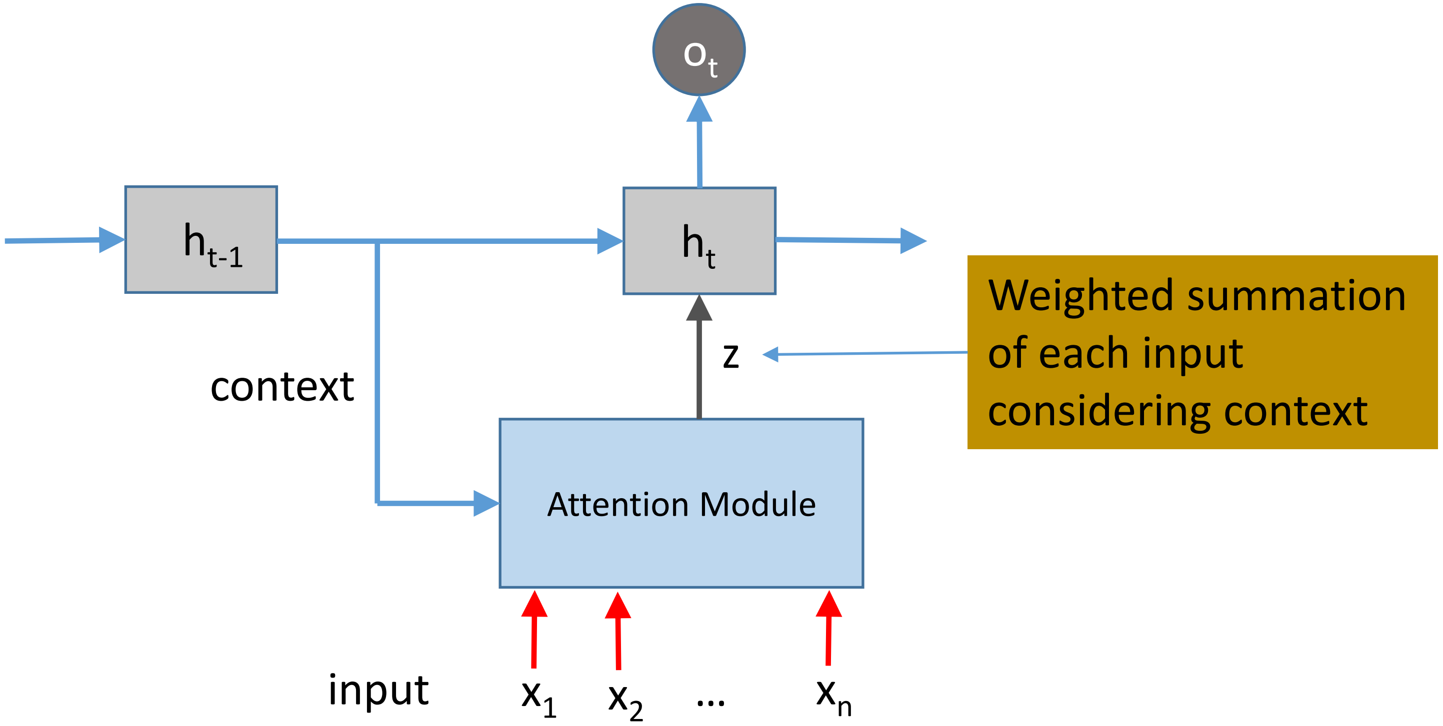


  - Inputs relevant to the context may be more useful

Kyunghyun Cho, “Introduction to Neural Machine Translation with GPUs” (2015)

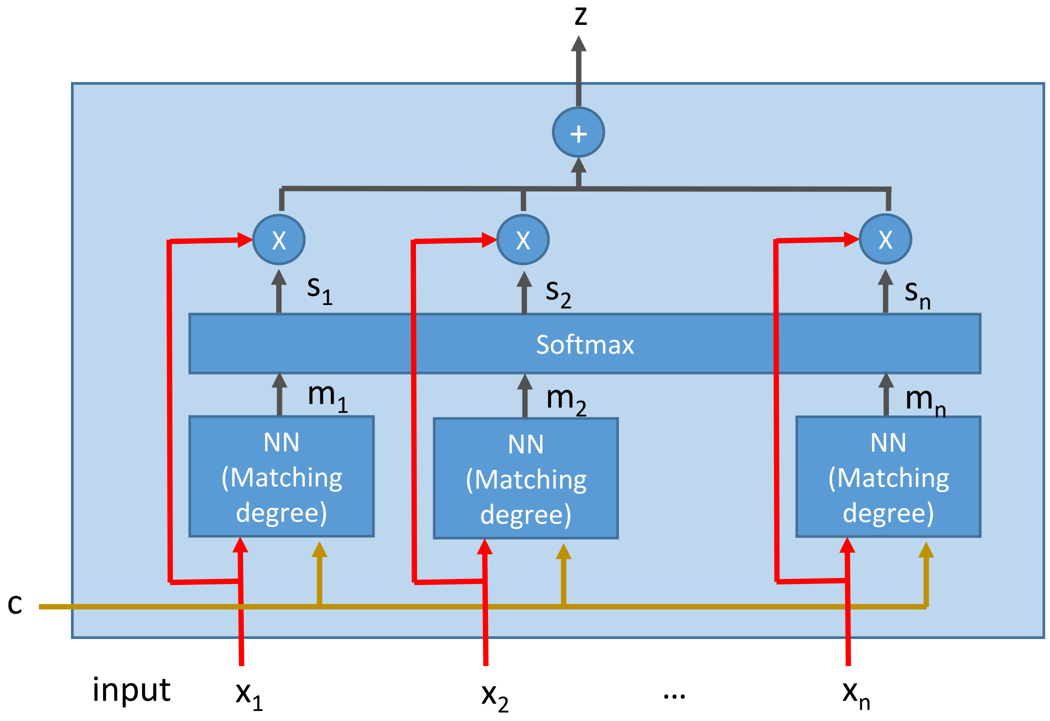
Attention Model

- Overview



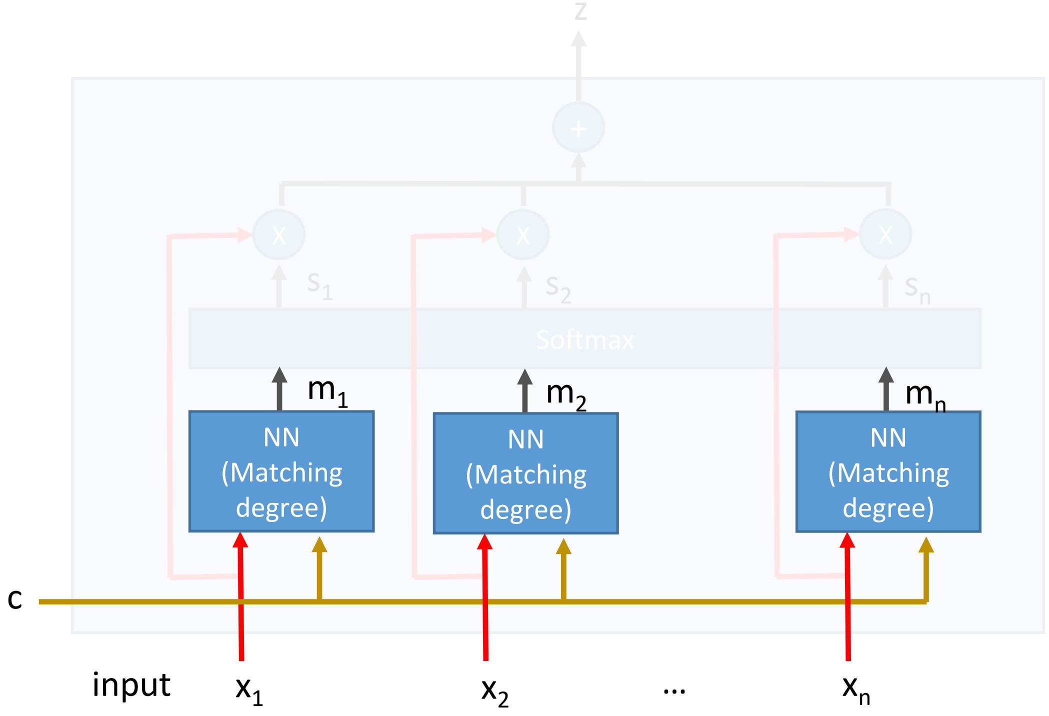
Attention Model

- Attention Module
  - All inputs share the same NN for matching degree



Attention Model

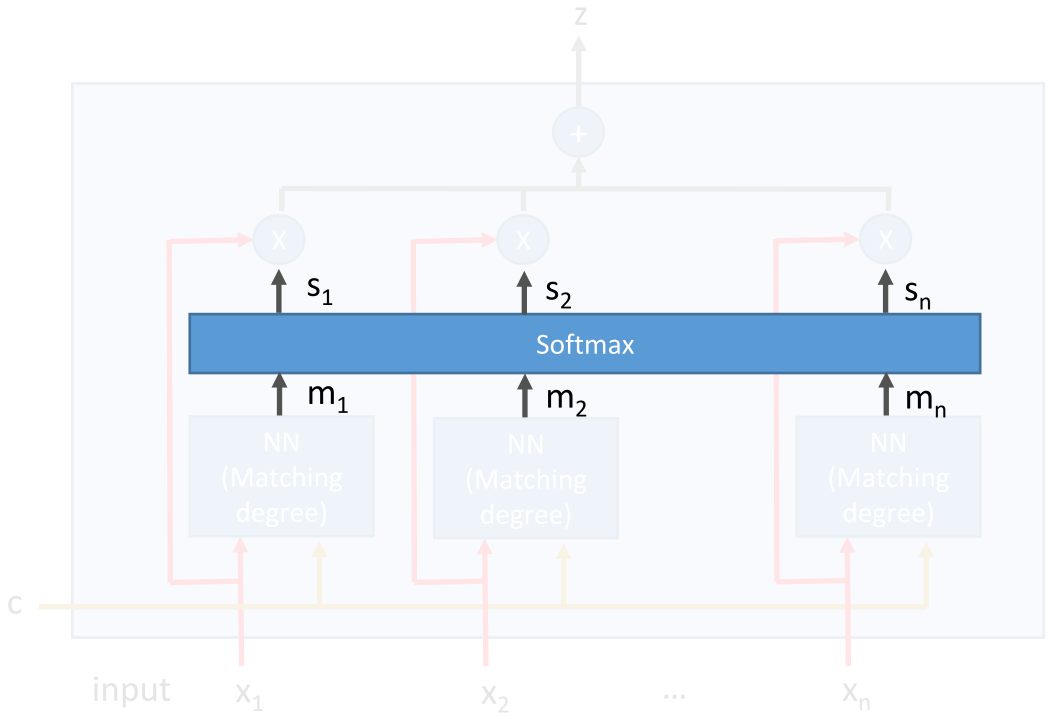
- Step 1: Evaluating Matching Degree
  - Evaluating matching degree of each input to the context
    - ✓ Produce scalar matching degree (Higher value is higher attention)
    - ✓ All inputs share the same NN



Attention Model

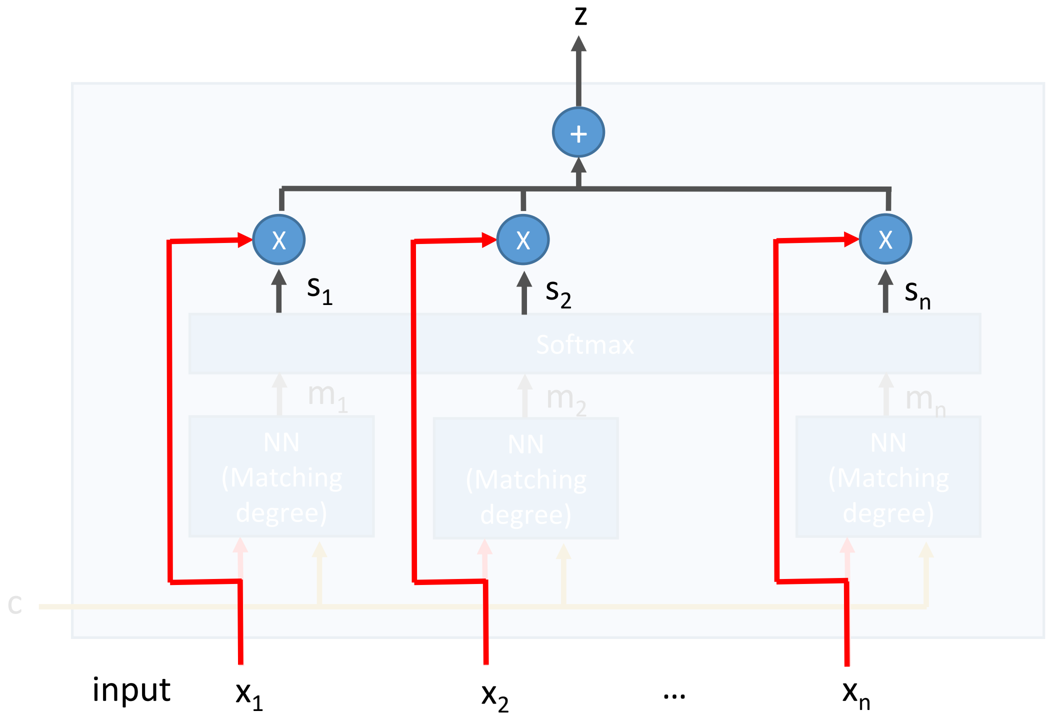
- Step 2: Normalizing Matching Degree

$$s_i = \frac{\exp(m_i)}{\sum_j exp(m_j)}$$



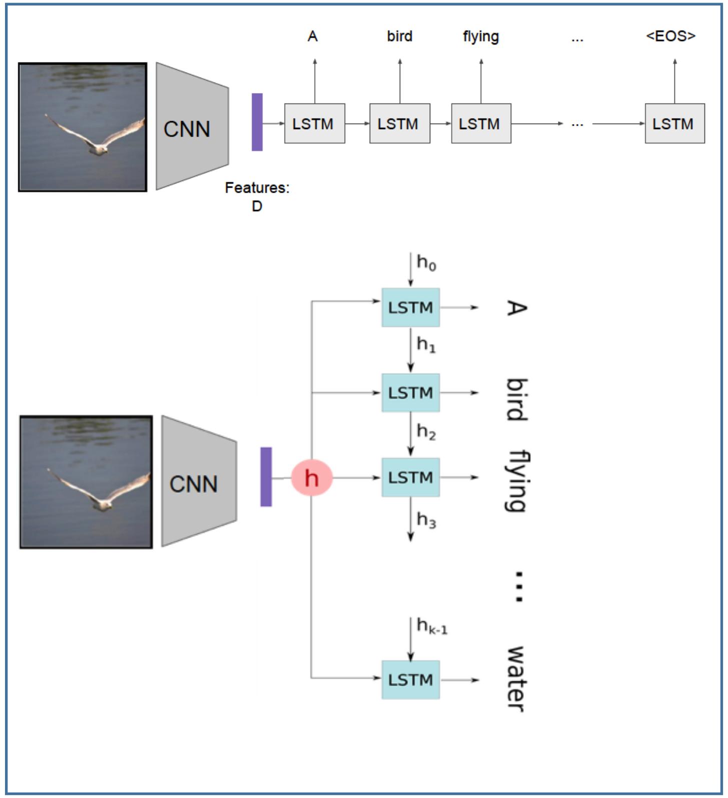
Attention Model

- Step 3: Aggregating Inputs
  - Each input is scaled by s_i and summed up into z
  - z is the input focused on the current context

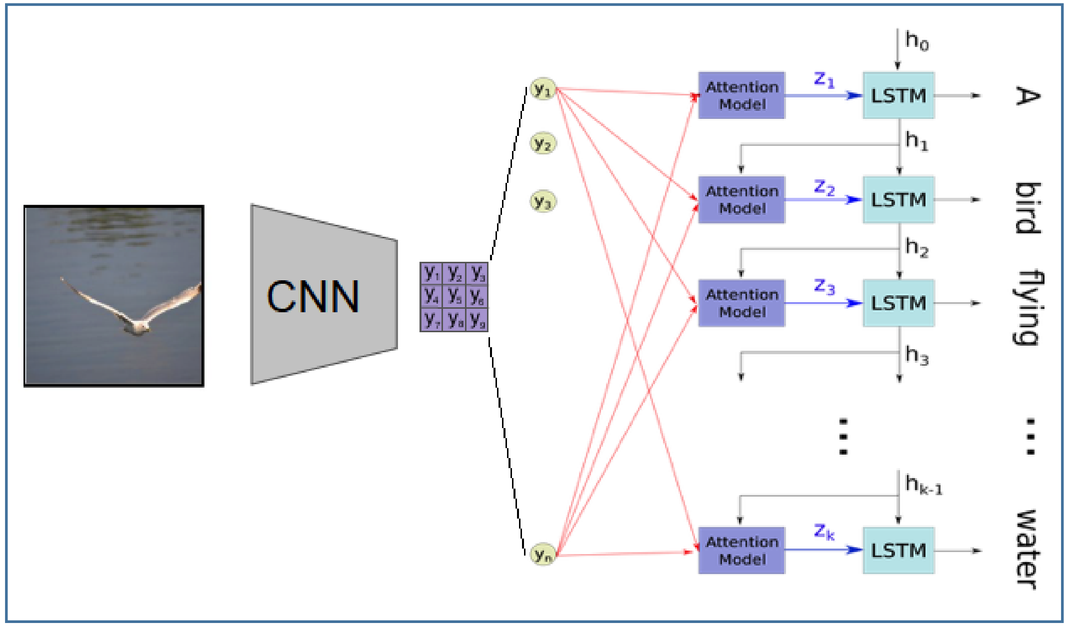


Attention Model

- Example



Encoder-decoder model



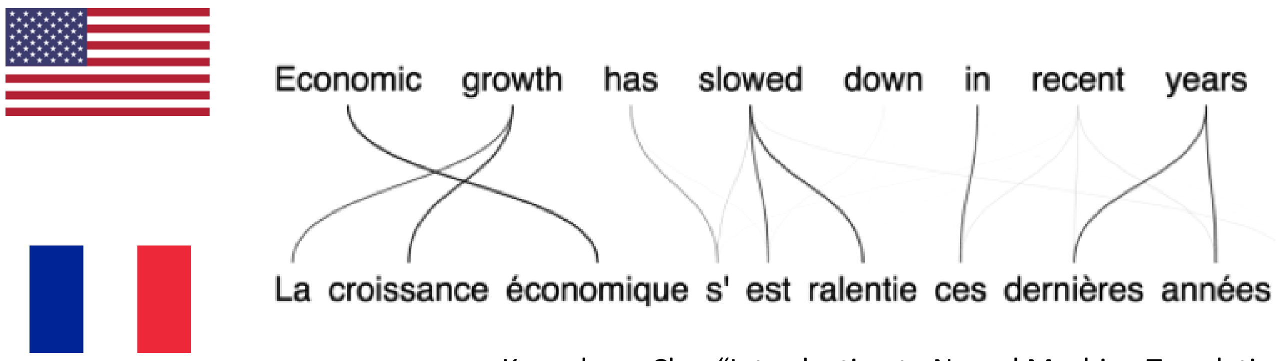
Attention based model

Attention Models, Amaia Salvador, 2017

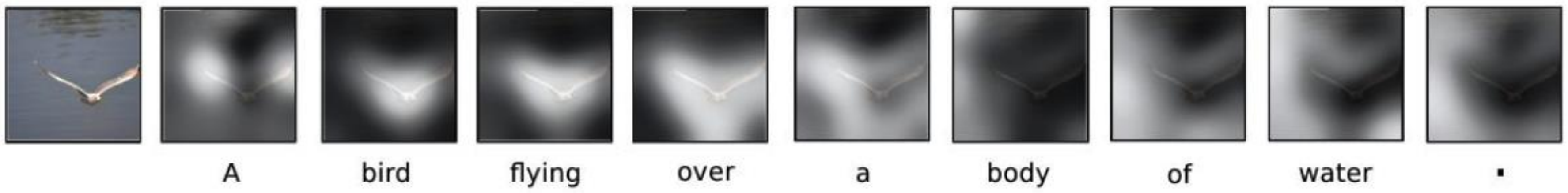
https://blog.heuritech.com/2016/01/20/attention-mechanism/

Attention Model

- One more advantage
  - We can interpret and visualize what the model is doing



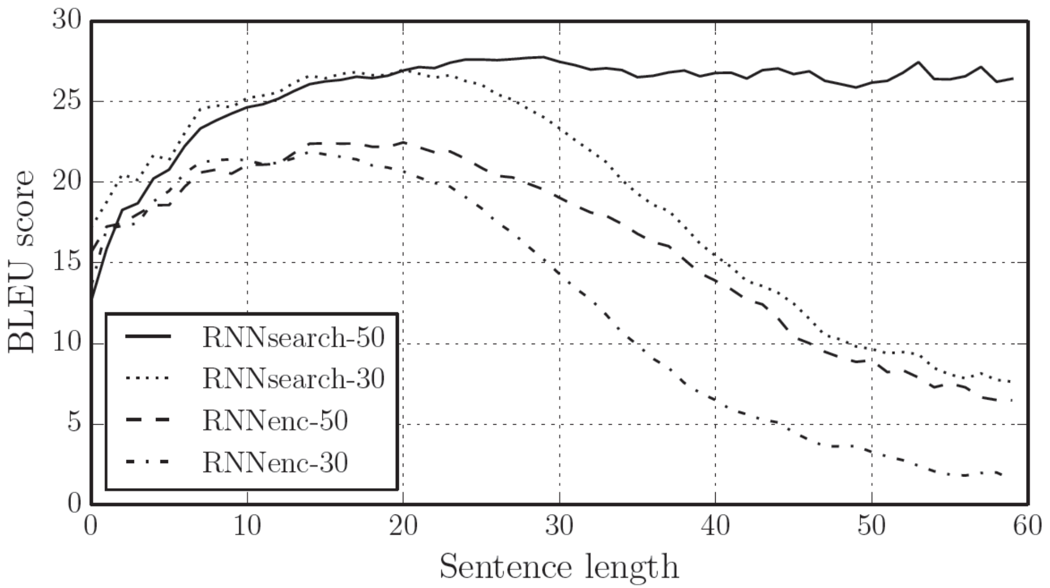
Kyunghyun Cho, “Introduction to Neural Machine Translation with GPUs” (2015)



Xu et al. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. ICML 2015

Attention is Great!

- RNNsearch-50 is a neural machine translation model with the attention mechanism trained on all the sentence pairs of length at most 50.
  - Dzmitry Bahdanau, KyungHyun Cho, Yoshua Bengio, "Neural Machine Translation by Jointly Learning to Align and Translate," ICLR 2015



Attention is Great!

- Attention significantly improves NMT performance.
  - It's very useful to allow decoder to focus on certain parts of the source.
- Attention solves the bottleneck problem.
  - Attention allows decoder to look directly at source; bypass bottleneck.
- Attention helps with vanishing gradient problem.
  - Provides shortcut to faraway states.
- Attention provides some interpretability.
  - By inspecting attention distribution, we can see what the decoder was focusing on.
  - We get alignment for free!
  - This is cool because we never explicitly trained an alignment system
  - The network just learned alignment by itself.

Q&A