Chapter 5

Sequence-to-Sequence with Attention

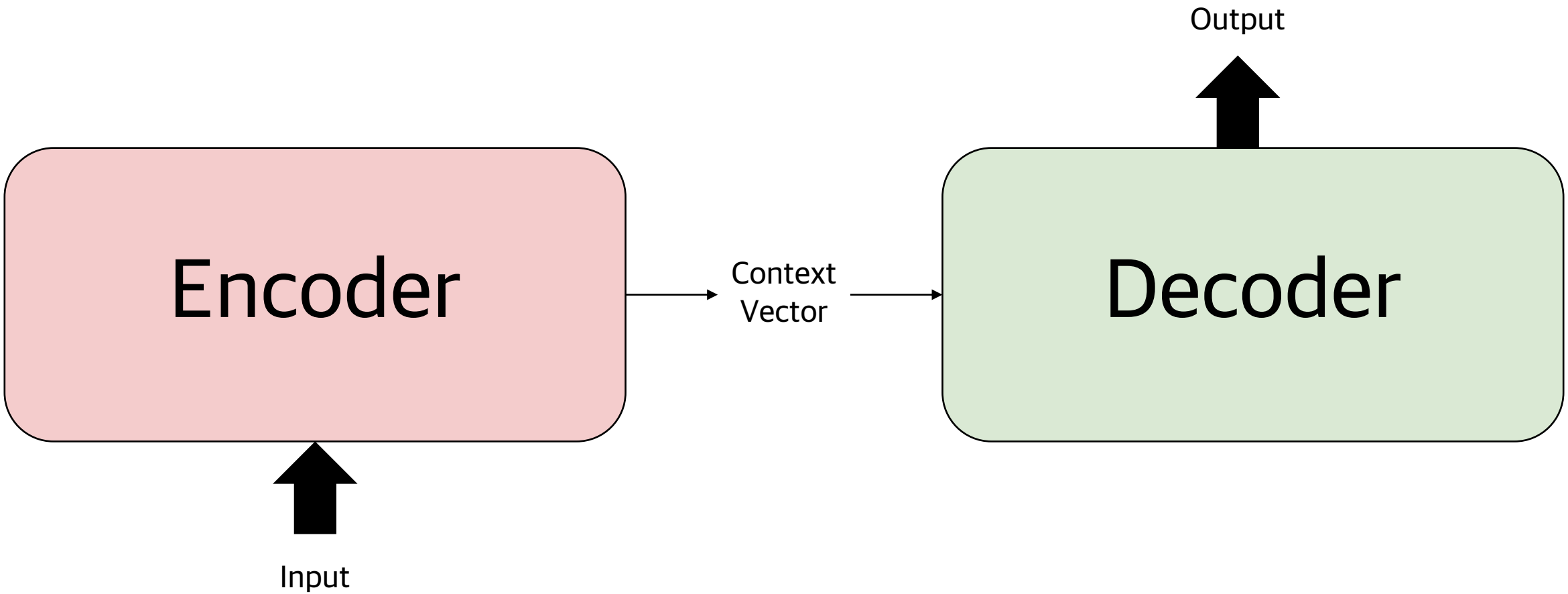
김지환

서강대학교 컴퓨터공학과

Table of contents

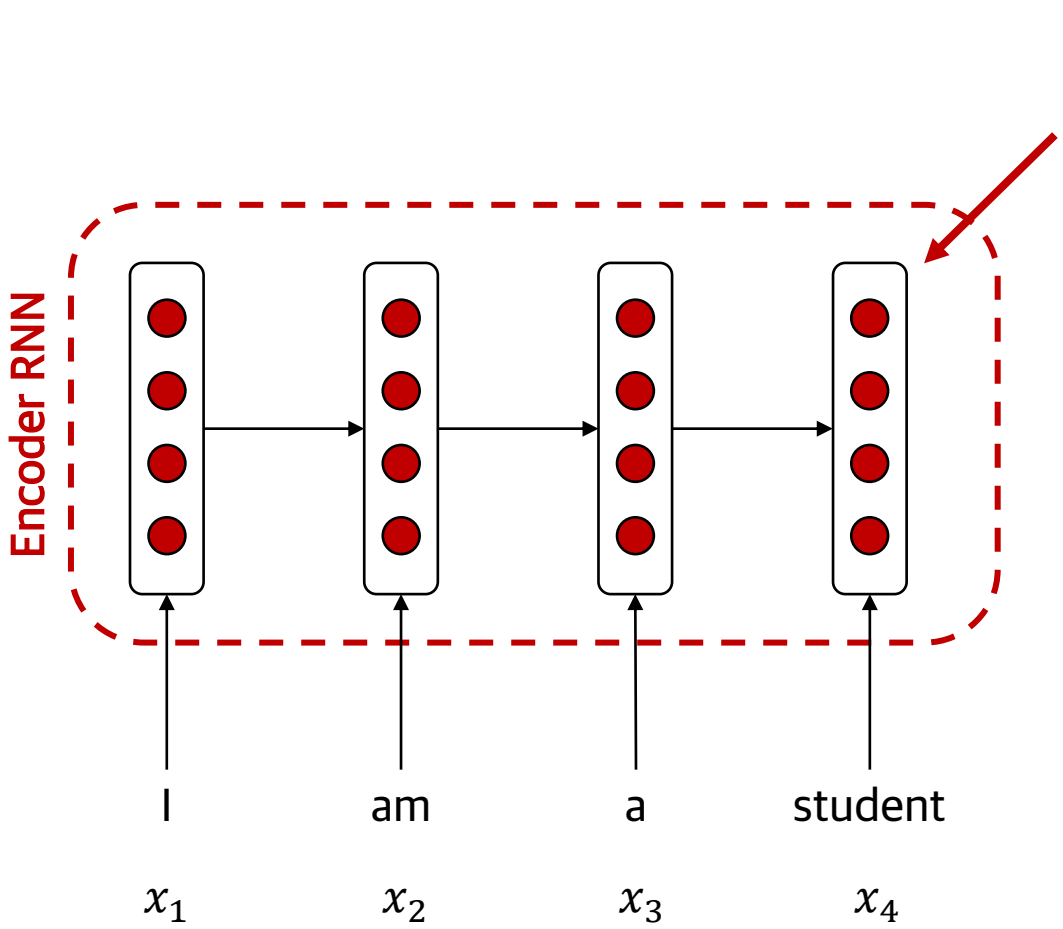
[I. Sutskever, 2014]

5.1 Seq2seq : Encoder-Decoder Model



[I. Sutskever, 2014]

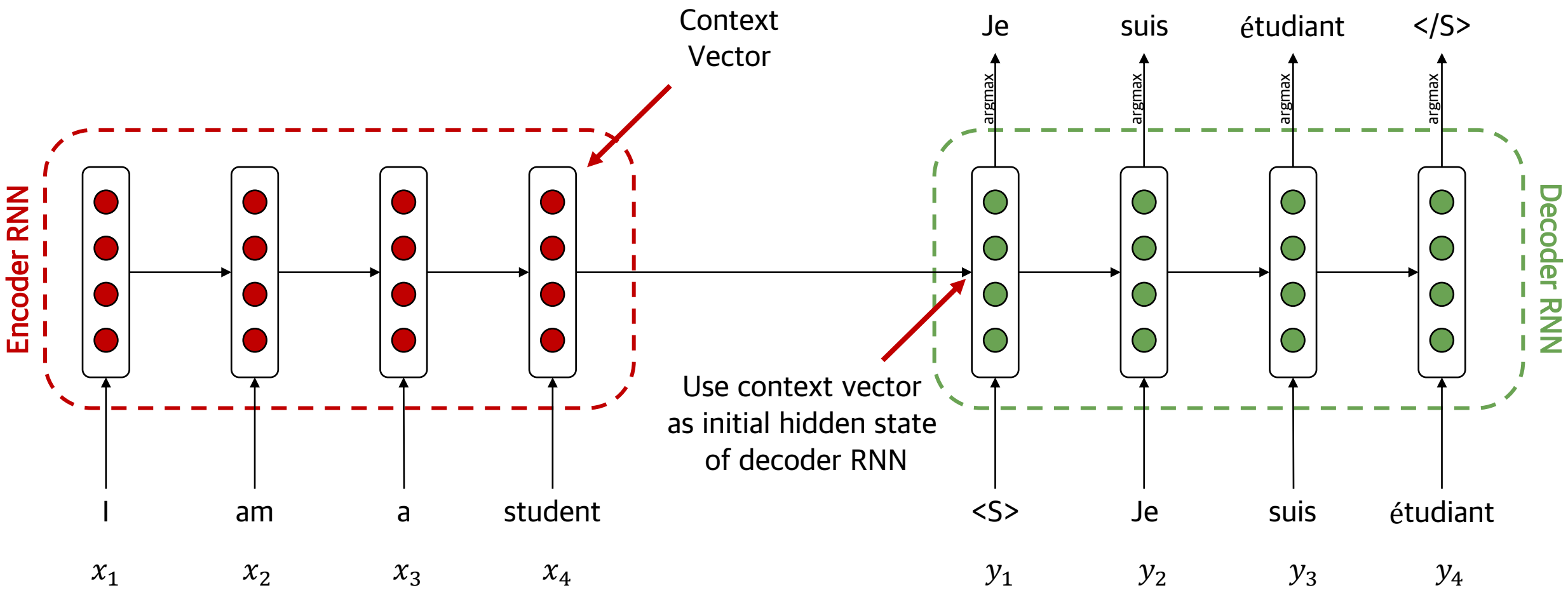
5.1.1 Seq2seq : Encoder



input token $x_1, \ldots, x_T$에 대하여
Context vector C는 T시점에서의 RNN의 hidden state, 즉 h_T

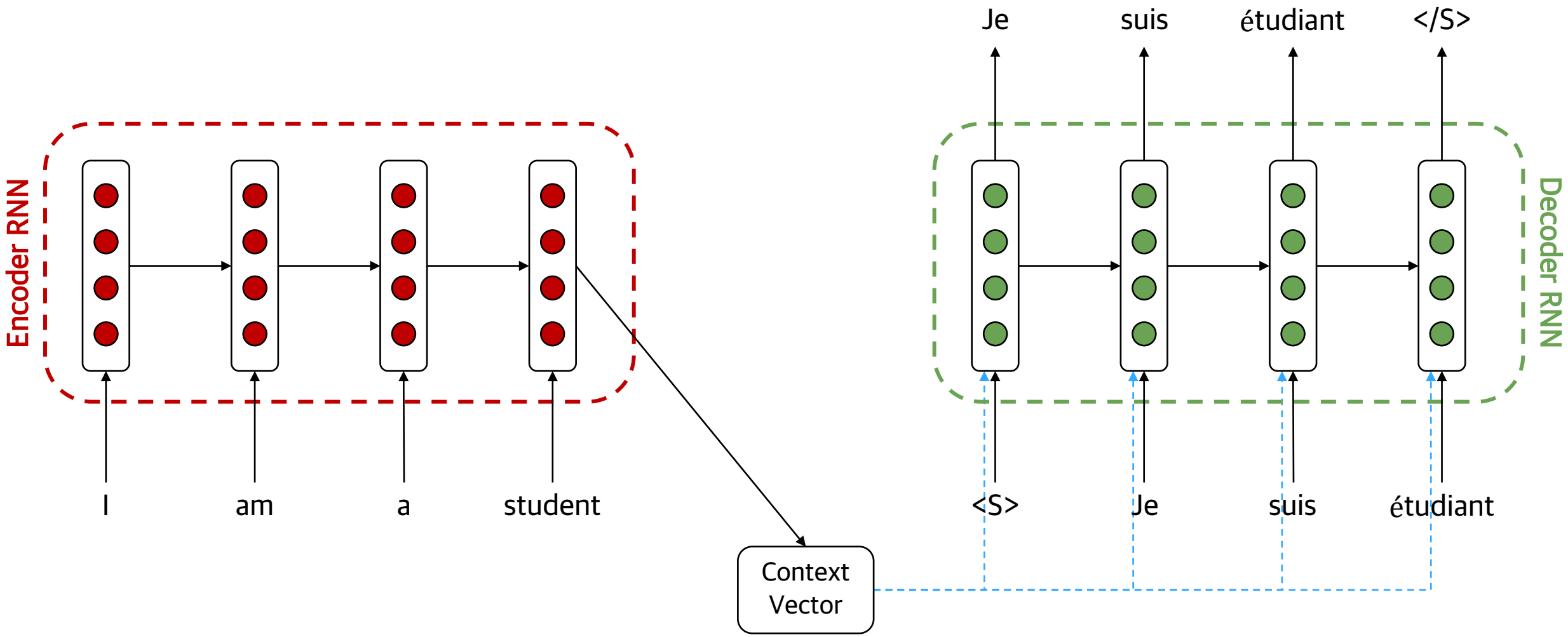
[I. Sutskever, 2014]

5.1.2 Seq2seq : Decoder



[K. Cho, 2014]

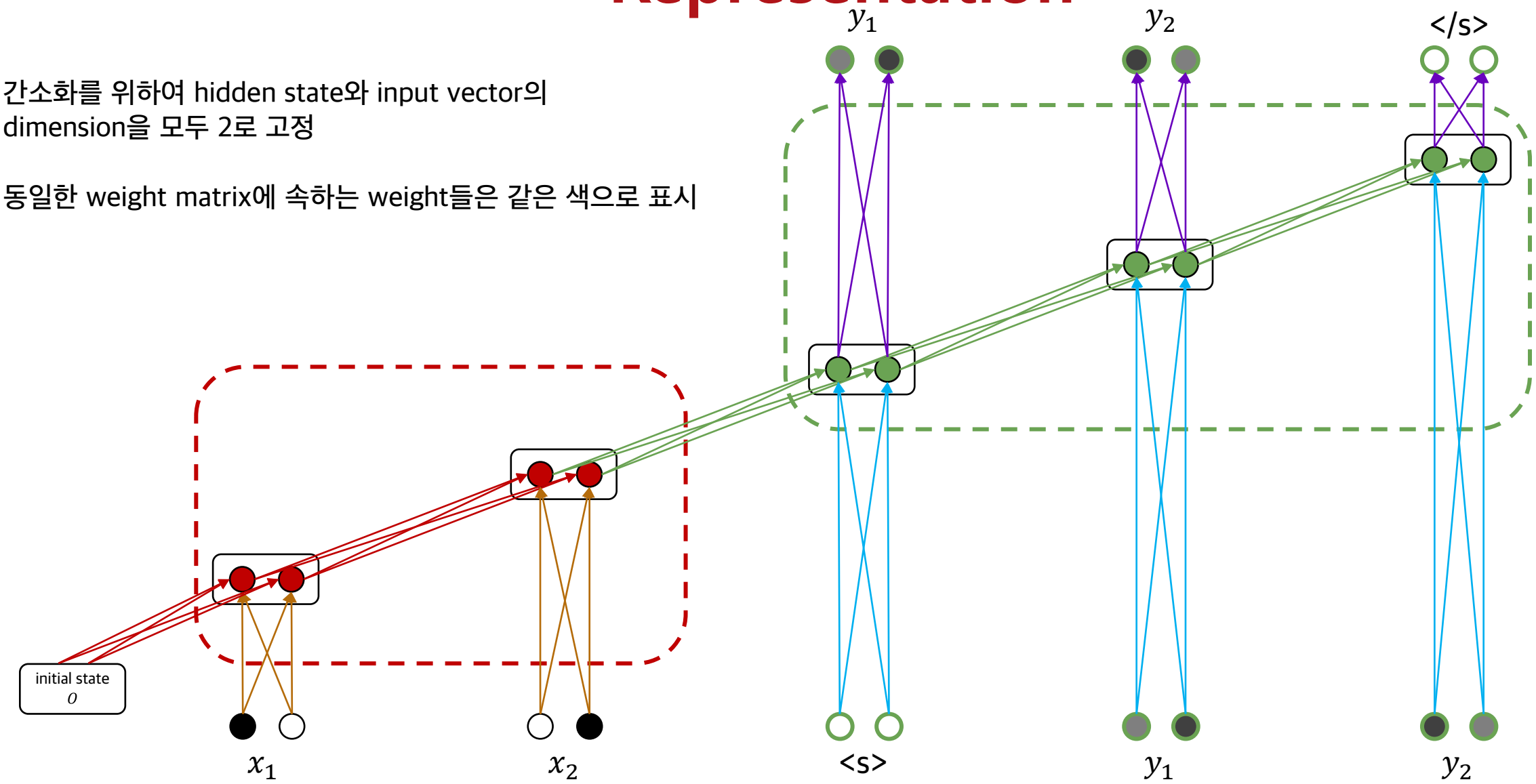
5.1.2 Seq2seq : Decoder (Cho *et al.)*



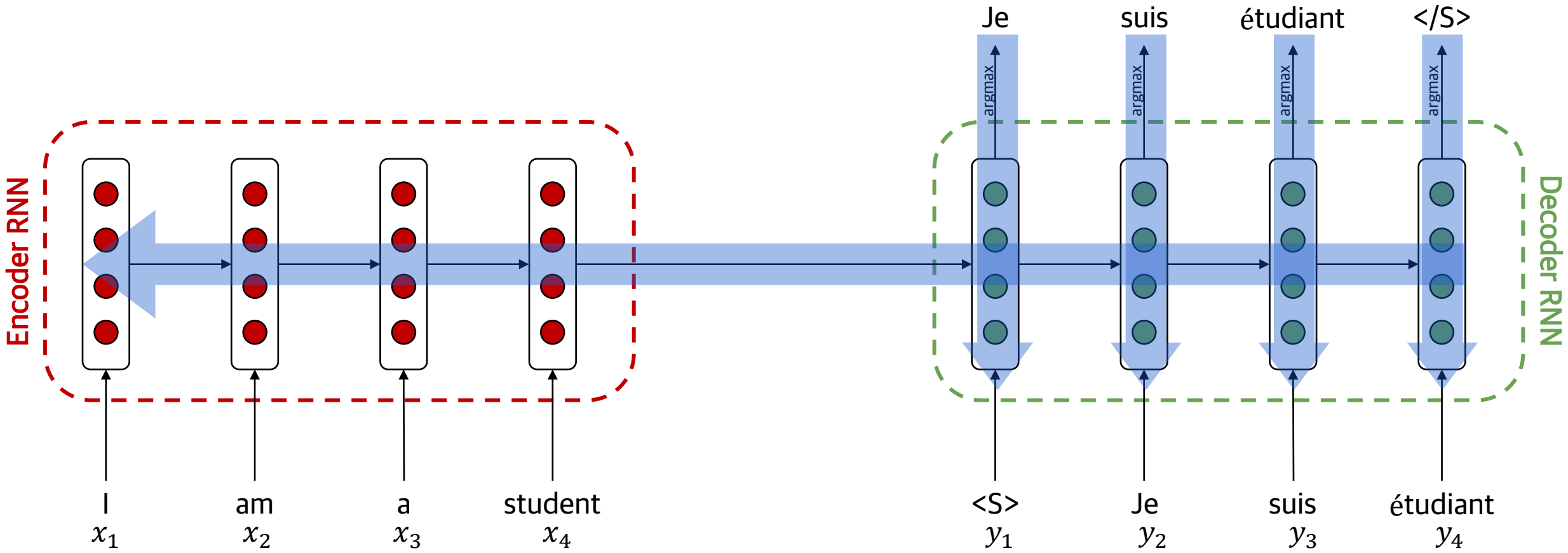
5.1.3 Seq2seq : Encoder-Decoder Graph Representation

간소화를 위하여 hidden state와 input vector의 dimension을 모두 2로 고정

동일한 weight matrix에 속하는 weight들은 같은 색으로 표시



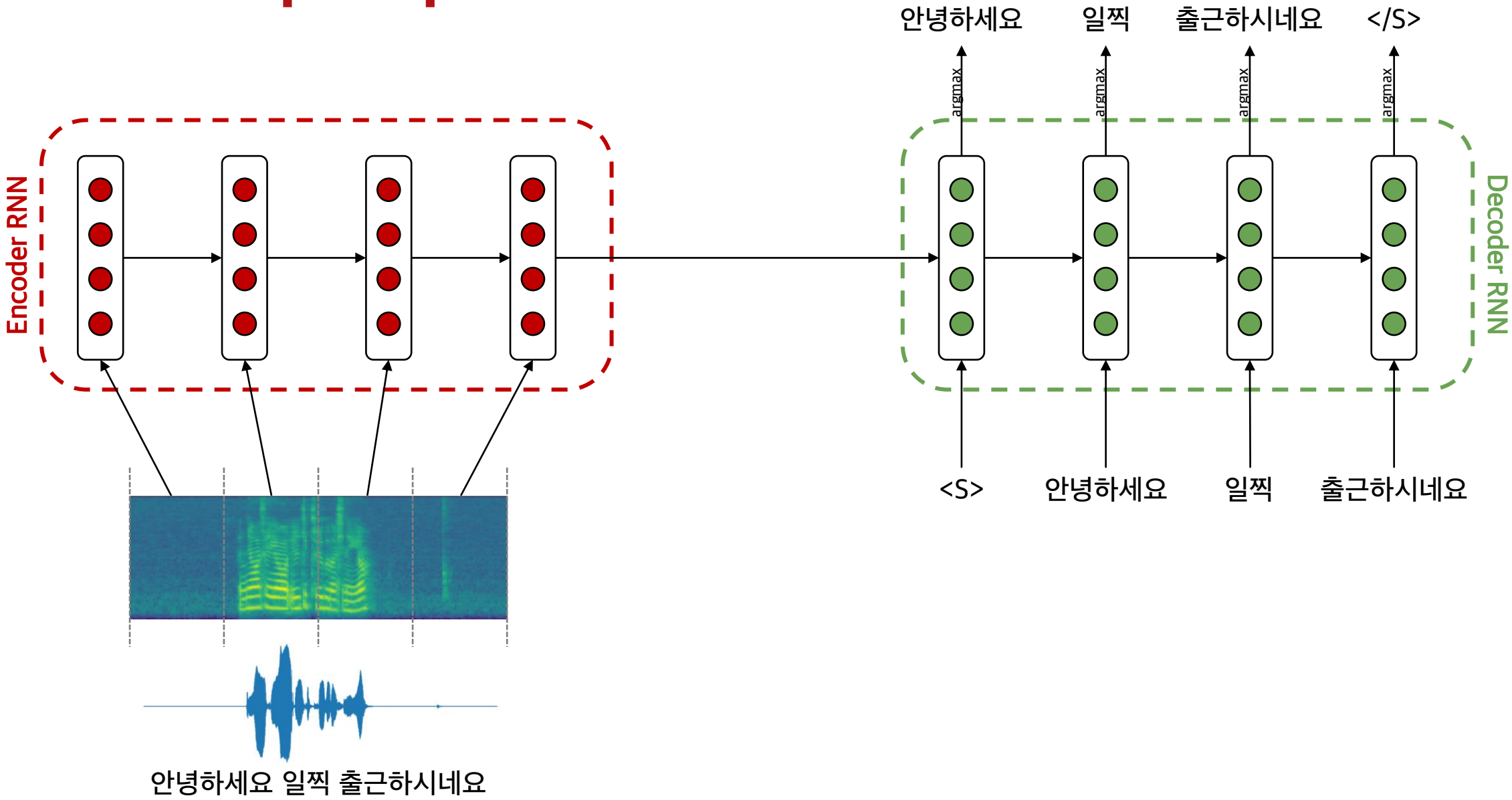
5.1.3 Seq2seq : Encoder-Decoder Model Training



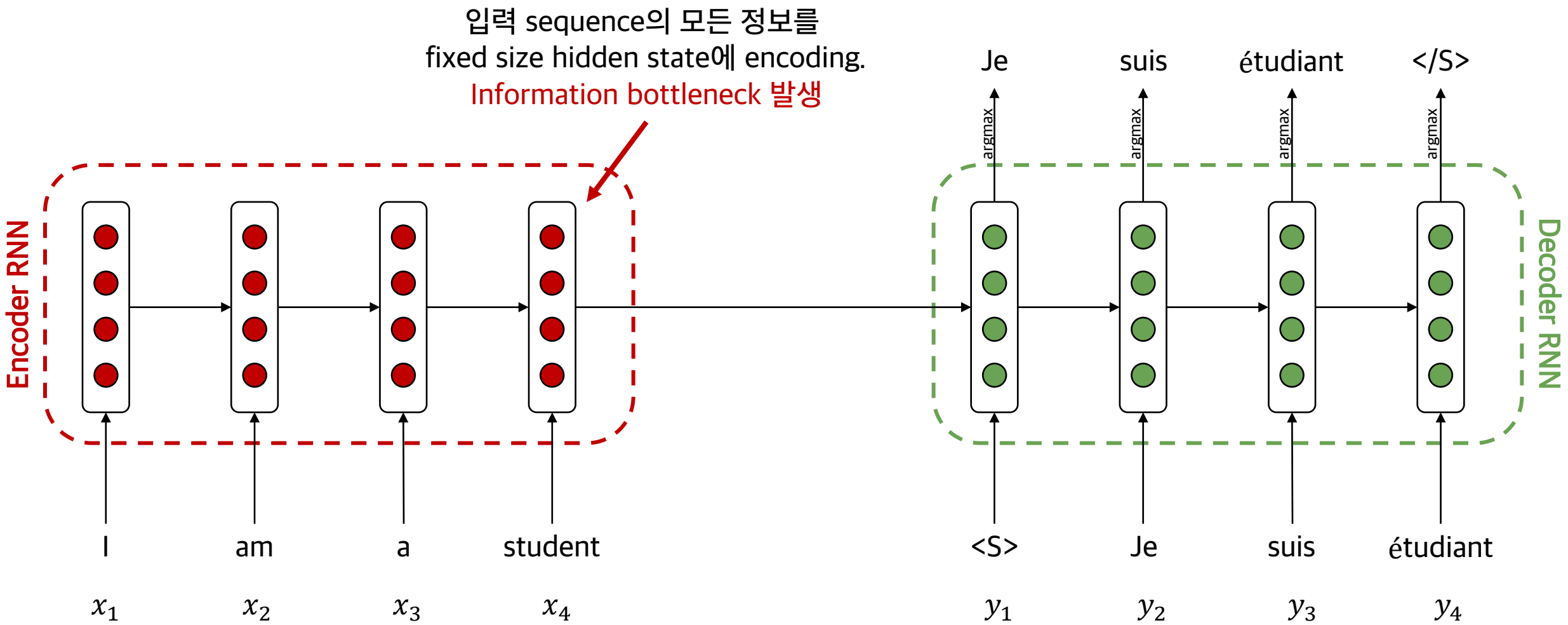
Encoder-decoder is optimized as a single system. Backpropagation operates “end-to-end”

[L. Lu, 2015]

5.1.4 Seq2seq : Encoder-Decoder Model for ASR



5.2 Seq2Seq : The Bottleneck Problem

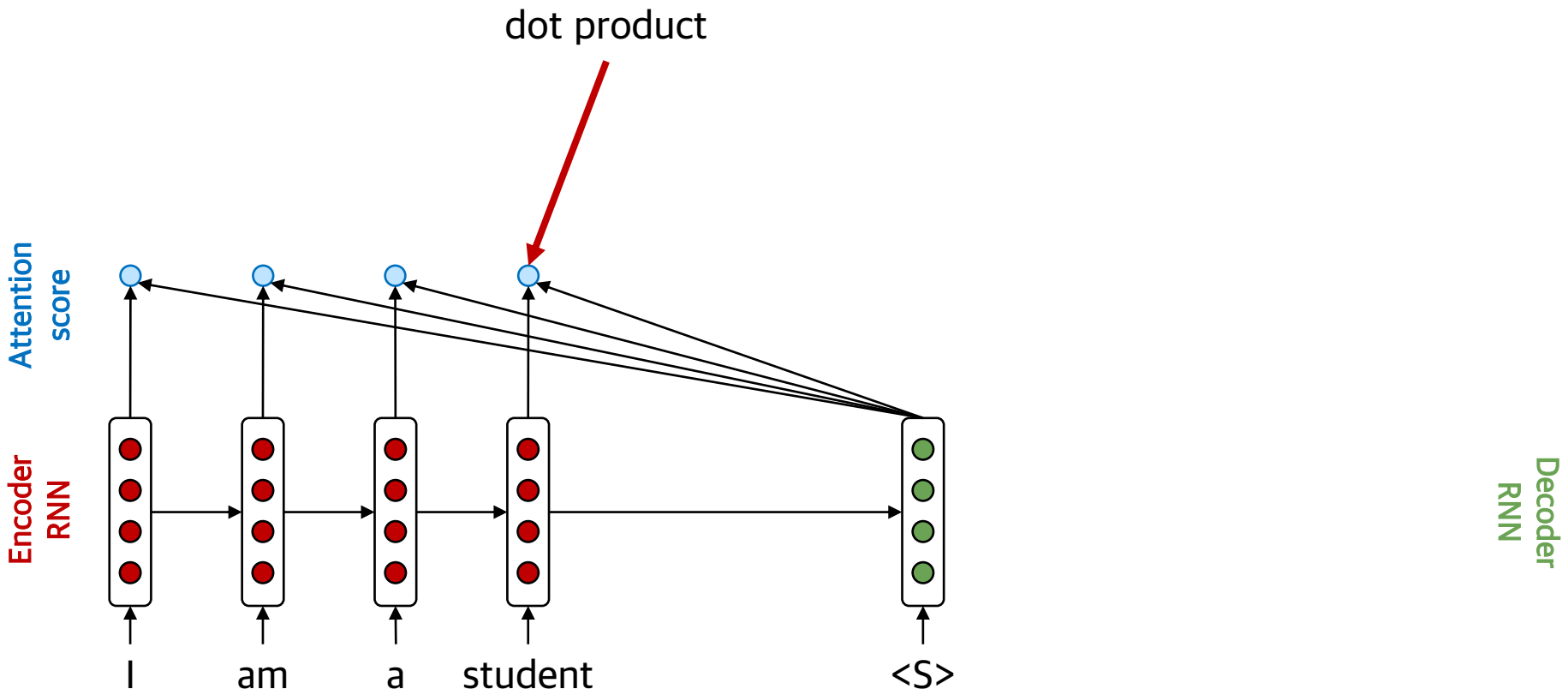


5.2.1 Attention

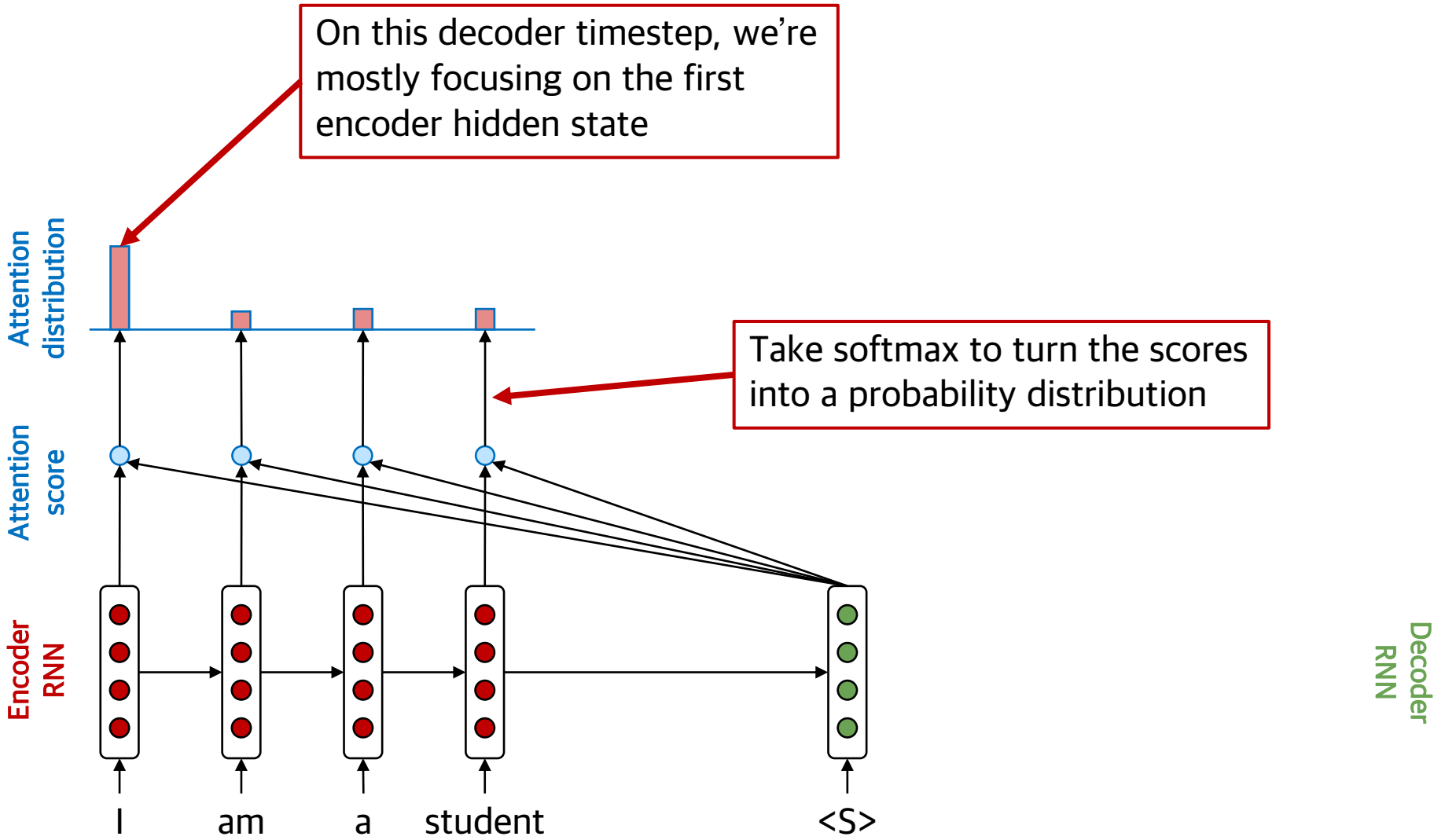
- Bottleneck problem에 대한 해결을 위해 **attention** 개념이 제안됨
- Core idea
  - 각 시점 마다 decoder에서 입력 sequence의 특정 부분에 초점을 맞출 수 있도록 encoder로 직접적으로 연결

image from : https://docs.likejazz.com/attention/

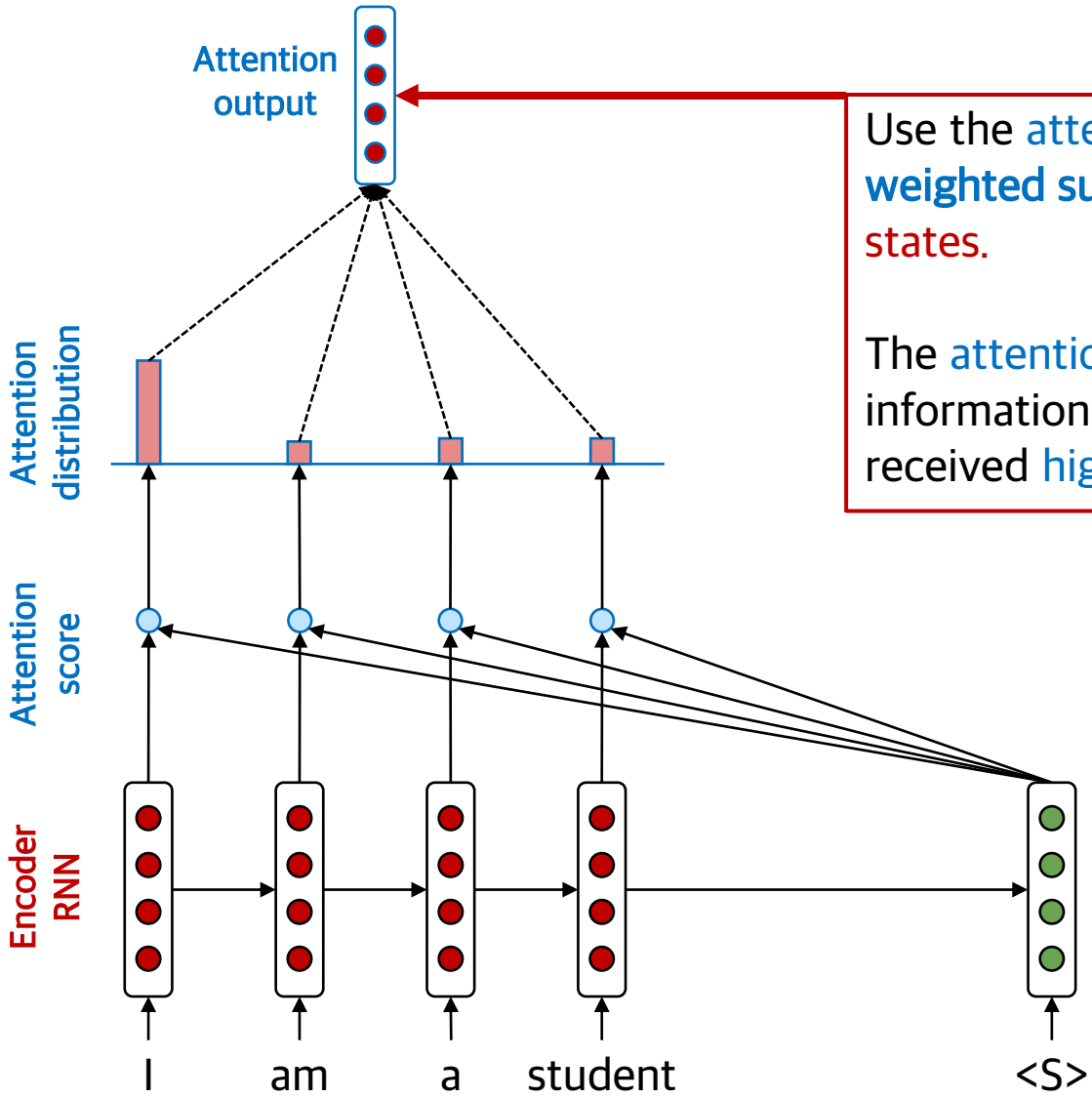
5.2.2 Seq2Seq with Attention



5.2.2 Seq2Seq with Attention



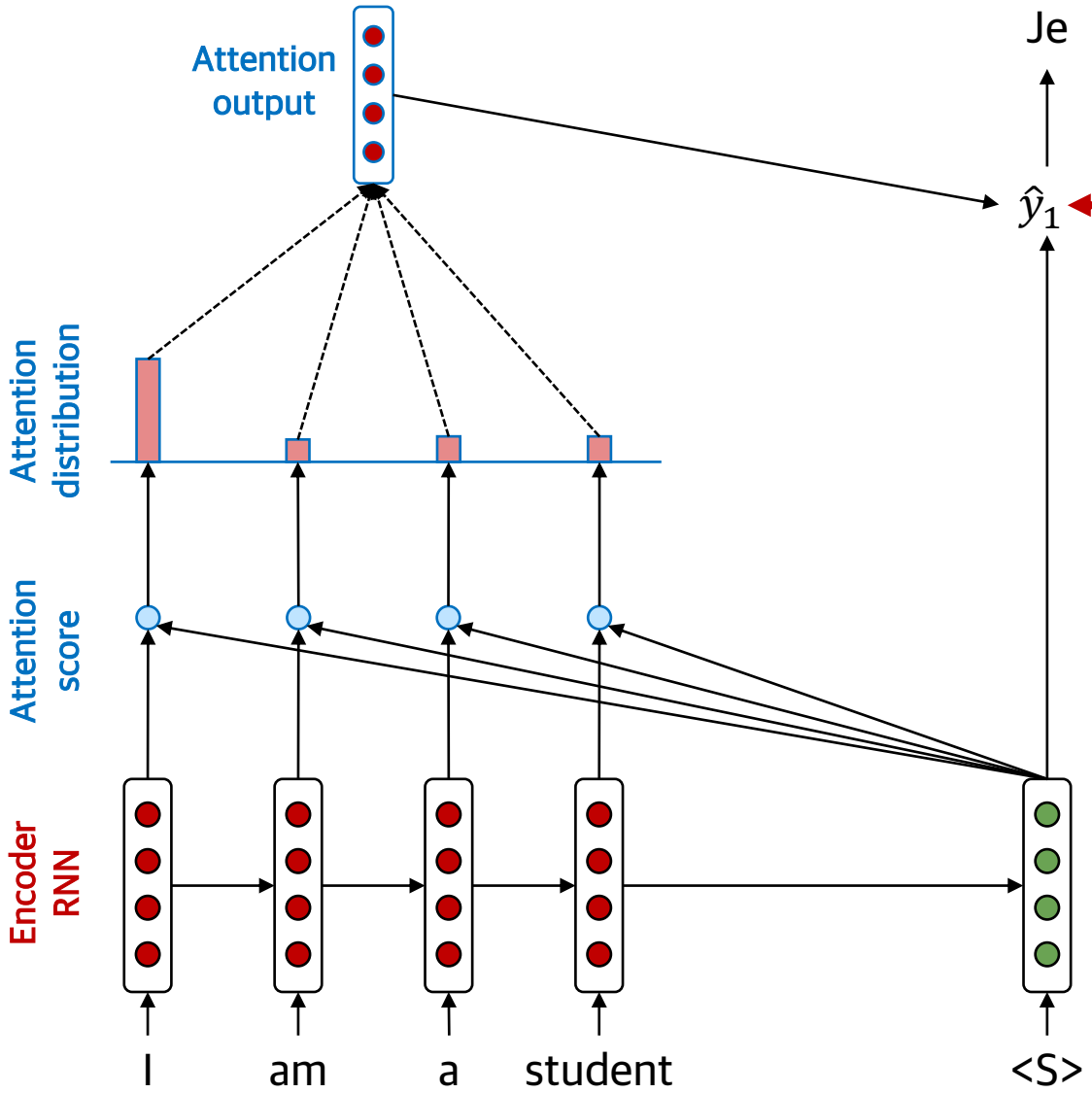
5.2.2 Seq2Seq with Attention



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

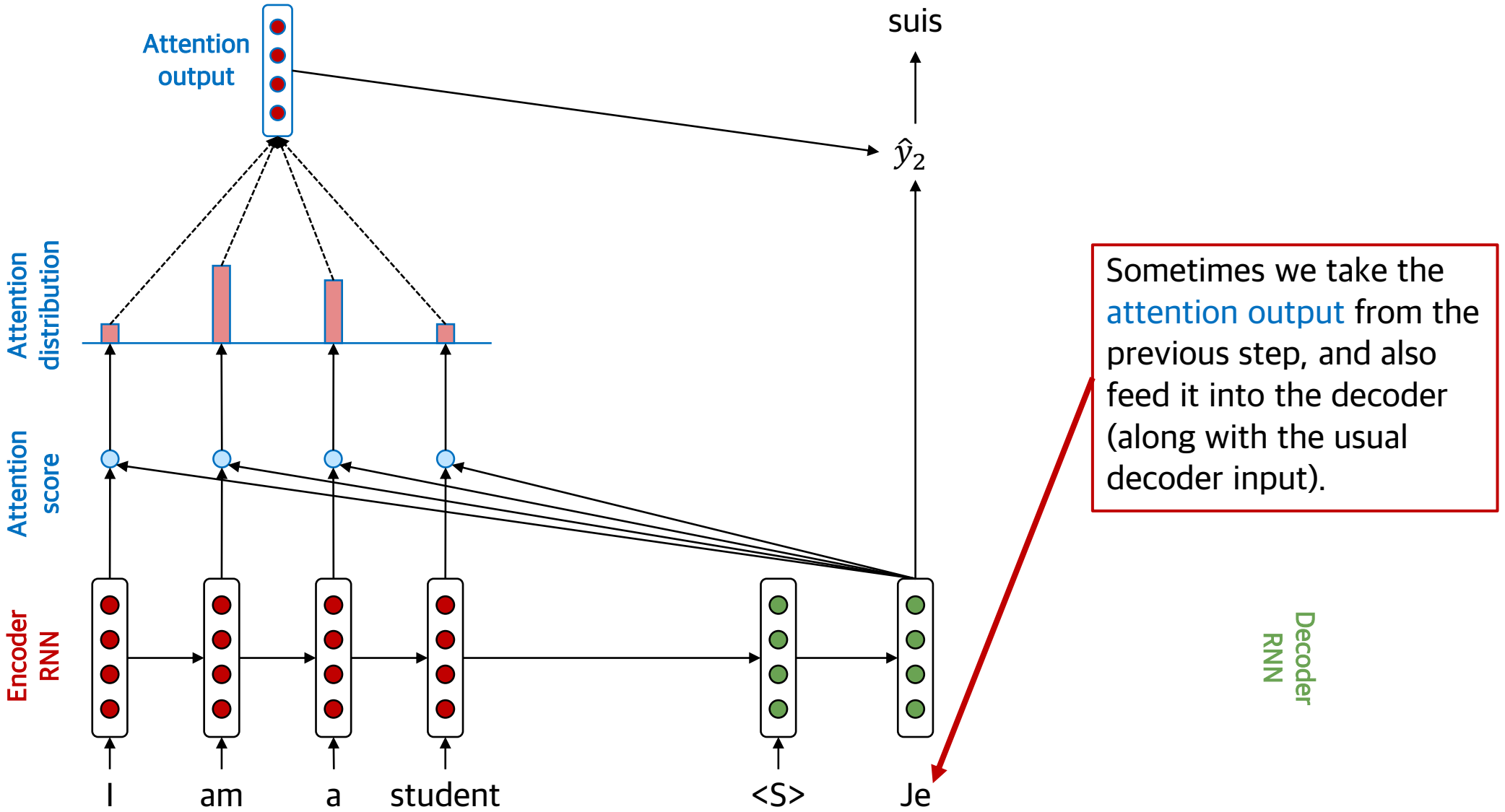
The attention output mostly contains information from the hidden states that received high attention.

5.2.2 Seq2Seq with Attention



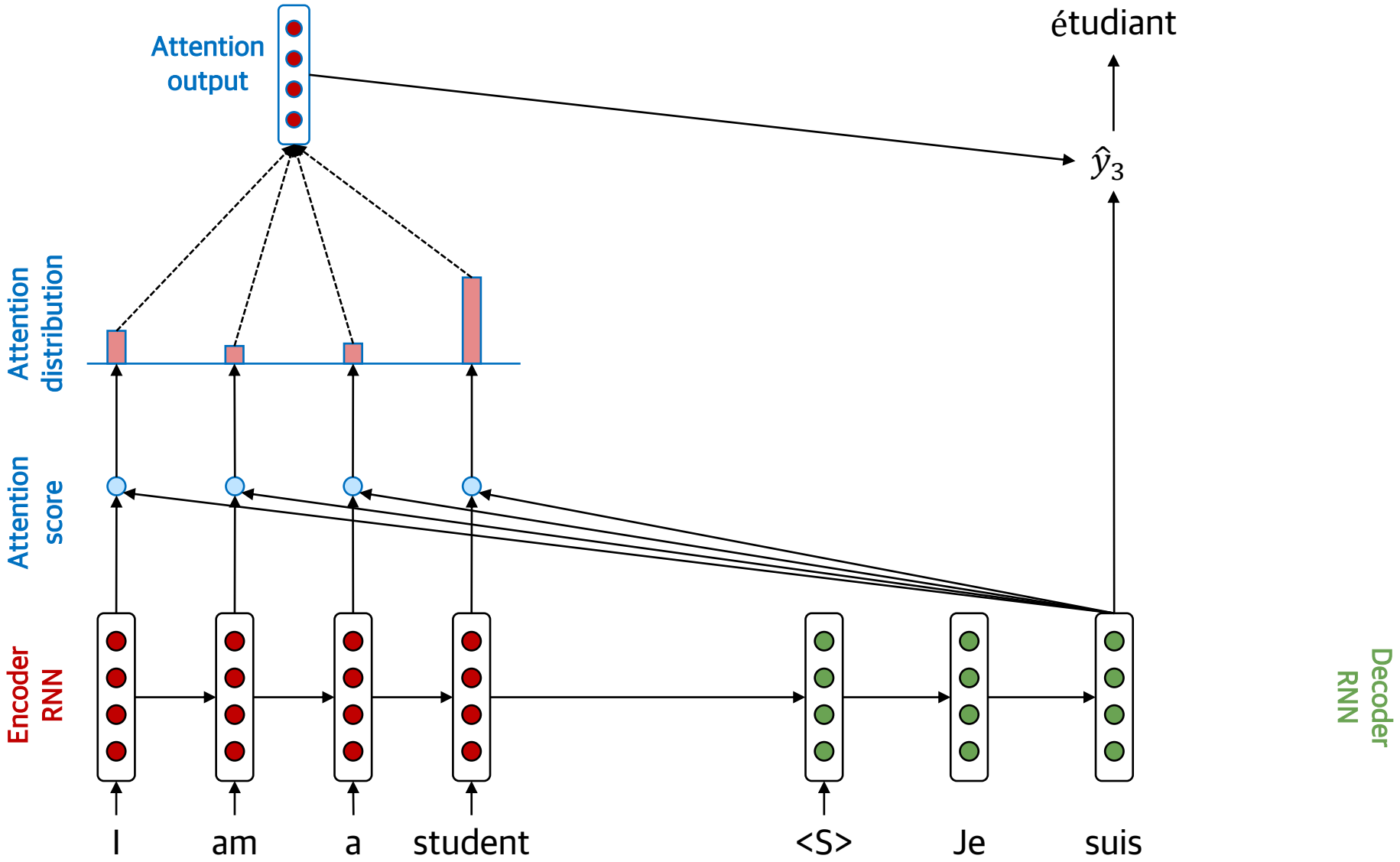
Concatenate attention output with decoder hidden state, then use to compute $\hat{y}_1$ as before

5.2.2 Seq2Seq with Attention

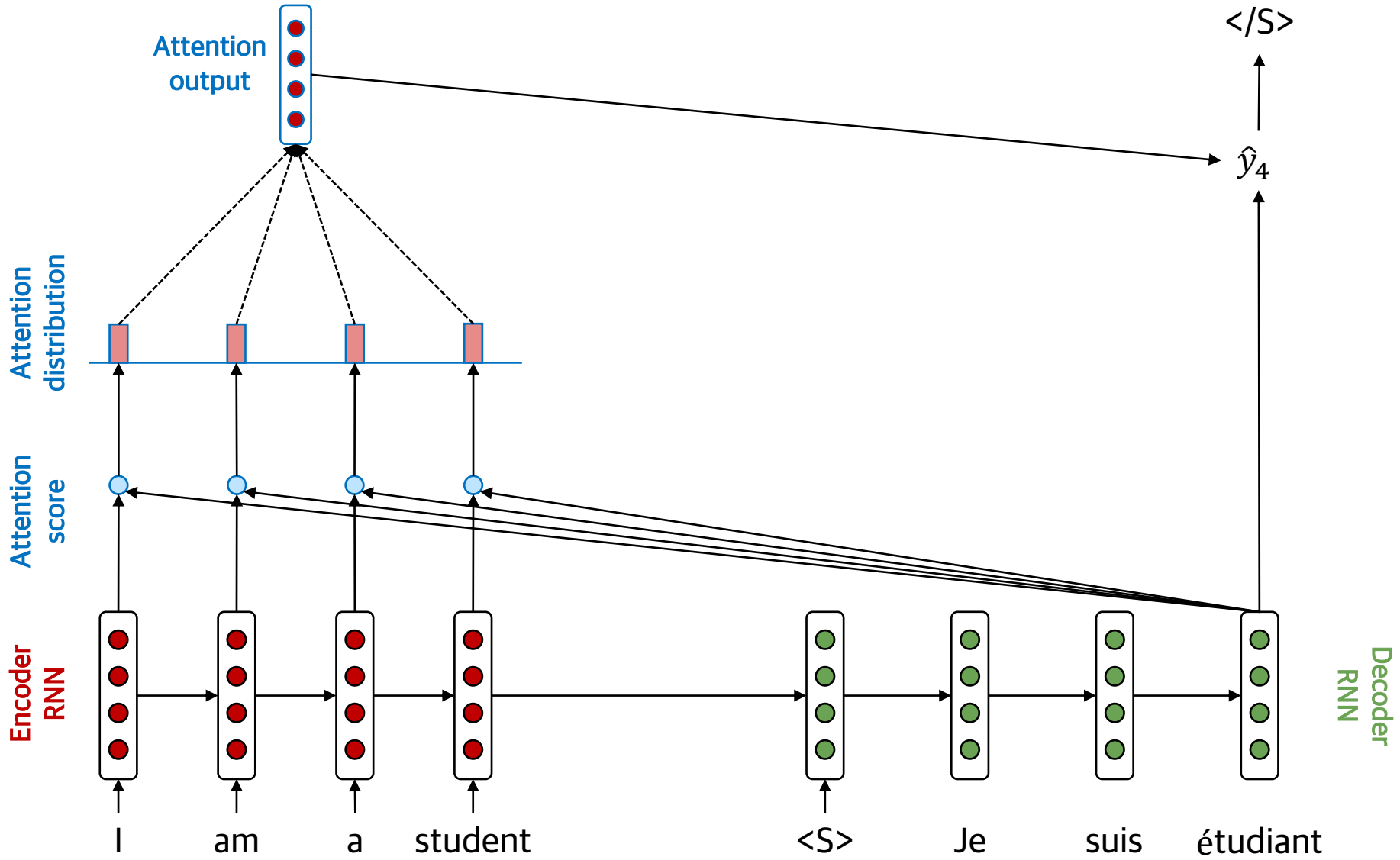


Sometimes we take the attention output from the previous step, and also feed it into the decoder (along with the usual decoder input).

5.2.2 Seq2Seq with Attention



5.2.2 Seq2Seq with Attention



5.2.2.1 Attention : in equations

- Encoder Hidden States $h_1, \ldots, h_N \in \mathbb{R}^h$
- t 시점에서의 decoder hidden state $s_t \in \mathbb{R}^h$
- Attention score e^t는 다음과 같이 계산

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

- Attention score에 softmax를 적용한 뒤, attention distribution α^t를 계산
  - α^t는 더해서 1이 되는 확률 분포

$$\alpha^t = \text{softmax}\left(e^t\right) \in \mathbb{R}^N$$

- Attention output $\boldsymbol{a}_t$은 α^t와 encoder hidden state의 weighted sum을 통해서 계산

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

- Attention output a_t와 decoder hidden state s_t를 concatenate한 뒤, 일반적인 seq2seq모델 처럼 처리

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