Lecture 8-1: Seq2Seq Learning & Transformer

Pilsung Kang
School of Industrial Management Engineering
Korea University

AGENDA

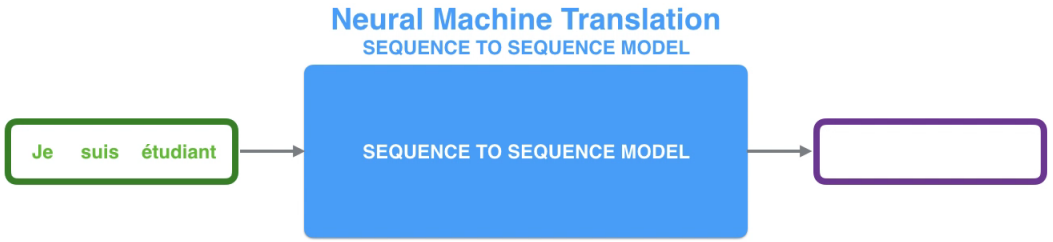
01 Sequence to Sequence (Seq2seq) Learning

02 Transformer

Sequence to Sequence Learning

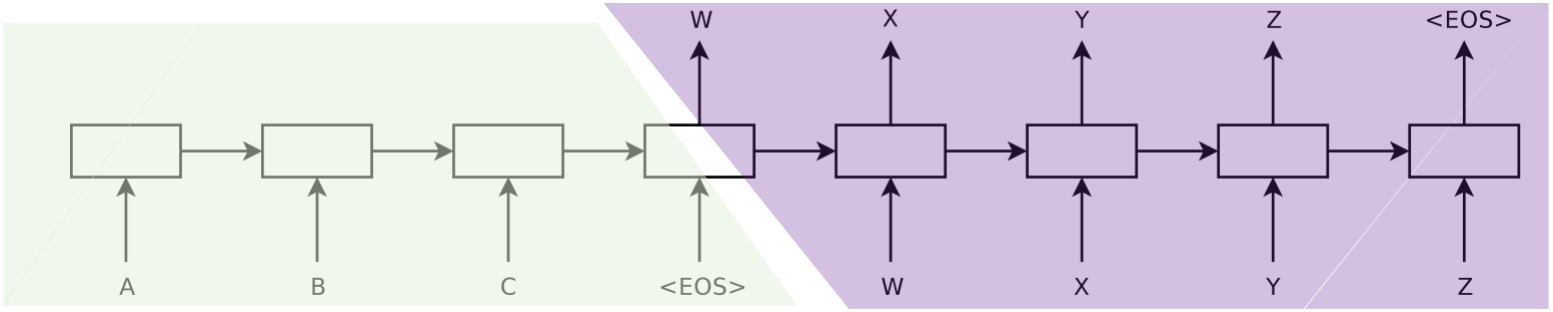
- Sequence-to-sequence model (Sutskever et al., 2014, Cho et al., 2014)
  - ✓ A model that takes a sequence of items (words, letters, features of an images, etc.)
  - ✓ Outputs another sequence of items
    - A trained model

    - Neural machine translation

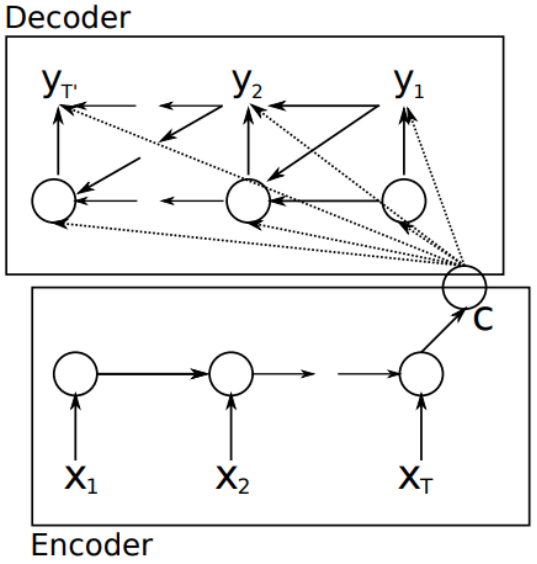


Sequence to Sequence Learning

- Main idea
  - ✓ Seq2Seq model consists of an encoder and a decoder



Sutskever et al., 2014

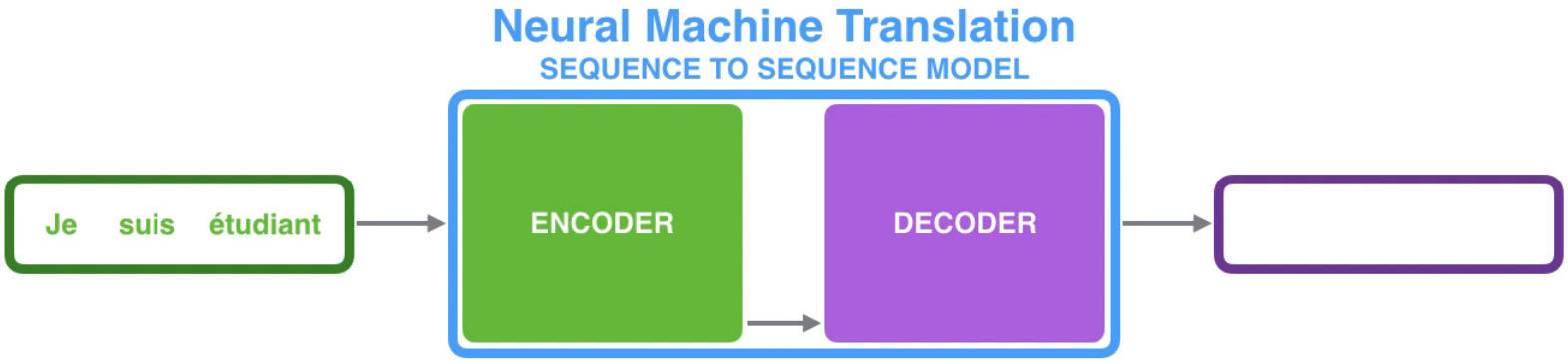


Cho et al., 2014

Sequence to Sequence Learning

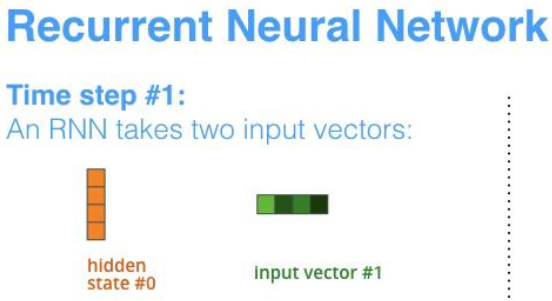
Alamar (Attention)

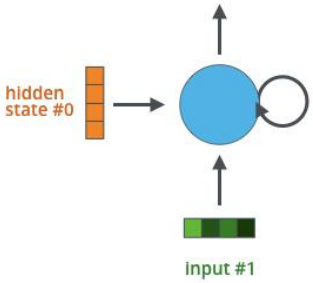
- Encoder-Decoder
  - ✓ The encoder processes each item in the input sequence and compiles the information it captures into a vector (context)
  - ✓ After processing the entire input sequence, the encoder send the context over to the decoder, which begins producing the output sequence item by item

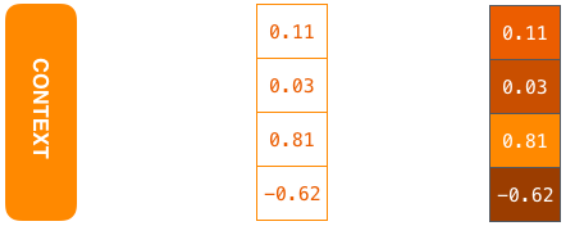


Sequence to Sequence Learning

- Encoder-Decoder
  - ✓ Recurrent neural network (RNN) is commonly used for the encoder and decoder structure
  - ✓ The context is a vector in the case of machine translation

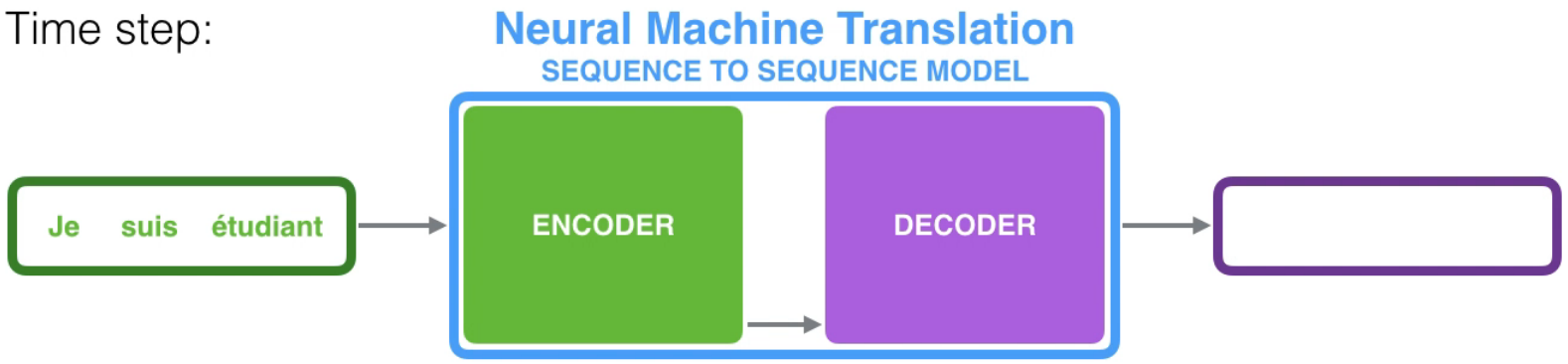






Sequence to Sequence Learning

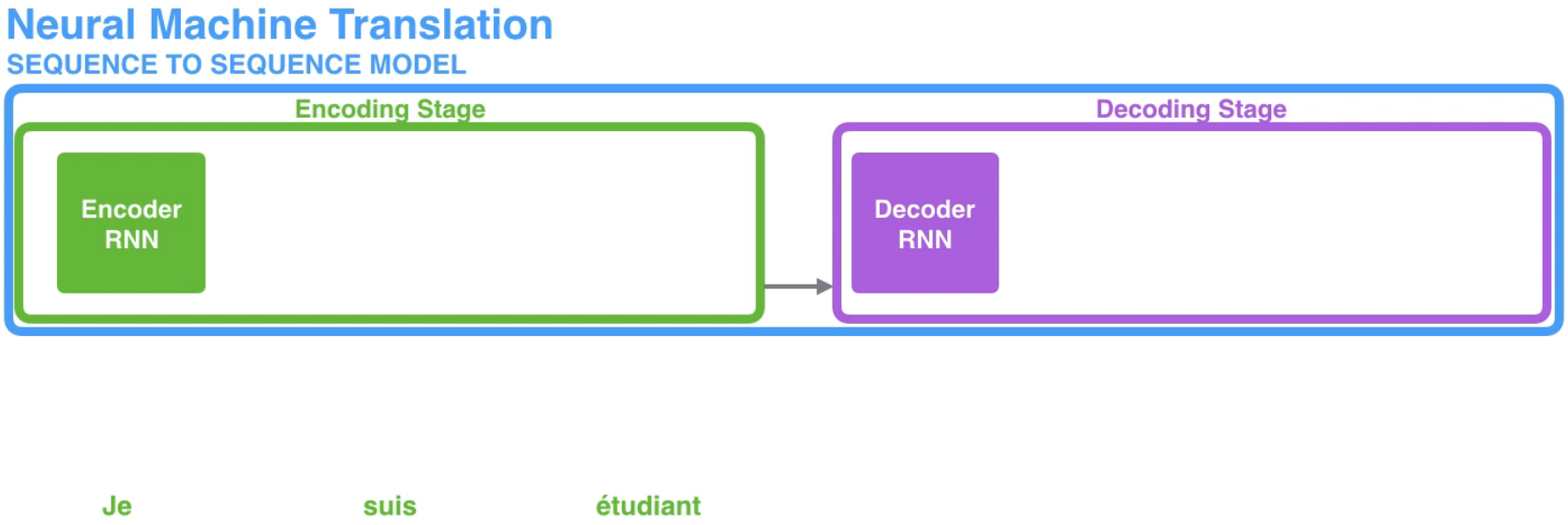
- Encoder-Decoder
  - ✓ Each purse for the encoder or decoder is that RNN processes its inputs and generates an output for that time step



Sequence to Sequence Learning

Alamar (Attention)

- An unrolled view of Seq2Seq learning



Attention in Seq2Seq Learning

Alamar (Attention)

- Attention
  - ✓ Context vector is a bottleneck for these types of models, which makes it challenging for the models to deal with long sentences
  - ✓ Attention allows the model to focus on the relevant part of the input sequence as needed

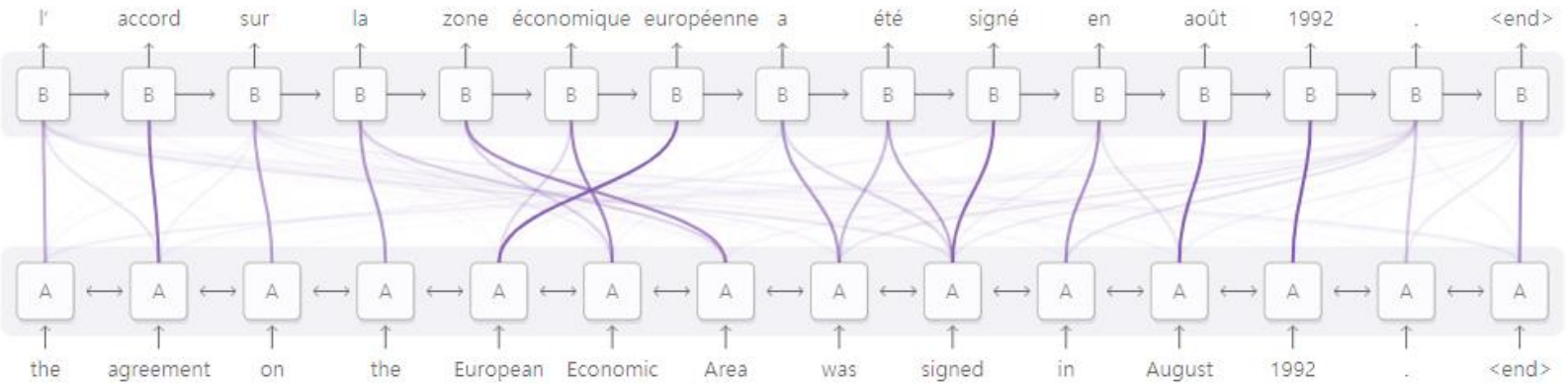


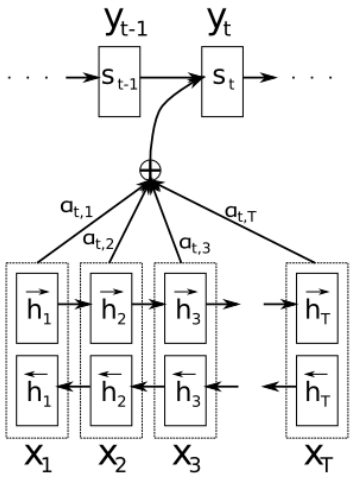
Diagram derived from Fig. 3 of Bahdanau, *et al.* 2014

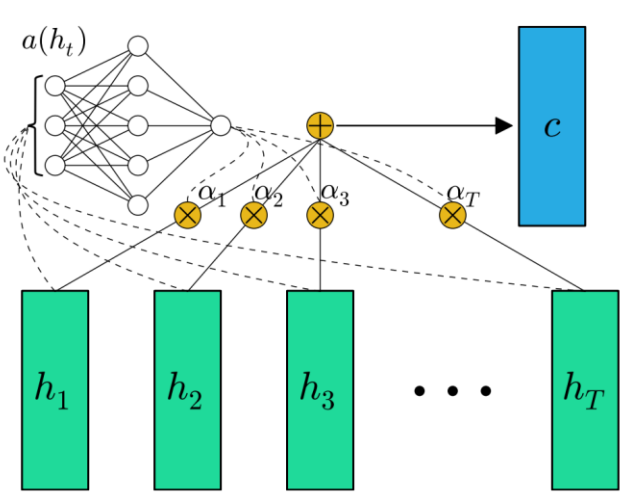
Olah (2016)

Attention in Seq2Seq Learning

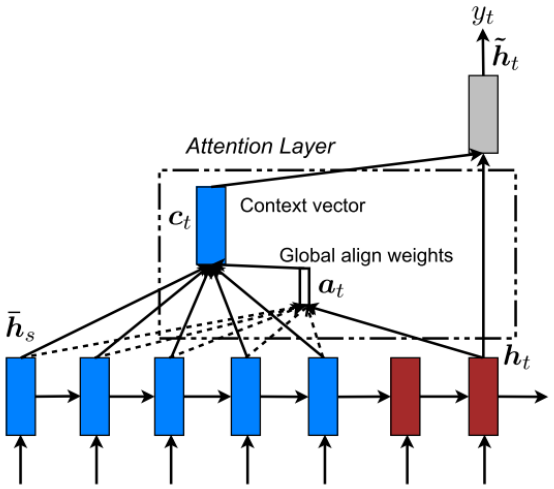
- Attention
  - ✓ Bahadanau attention (Bahdanau et al., 2015)
    - Attention scores are separated trained, the current hidden state is a function of the context vector and the previous hidden state
  - ✓ Luong attention (Luong et al., 2015)
    - Attention scores are not trained, the new current hidden state is the simple tanh of the weigthed concatenation of the context vector and the current hidden state of the decoder

Bahdanau attention



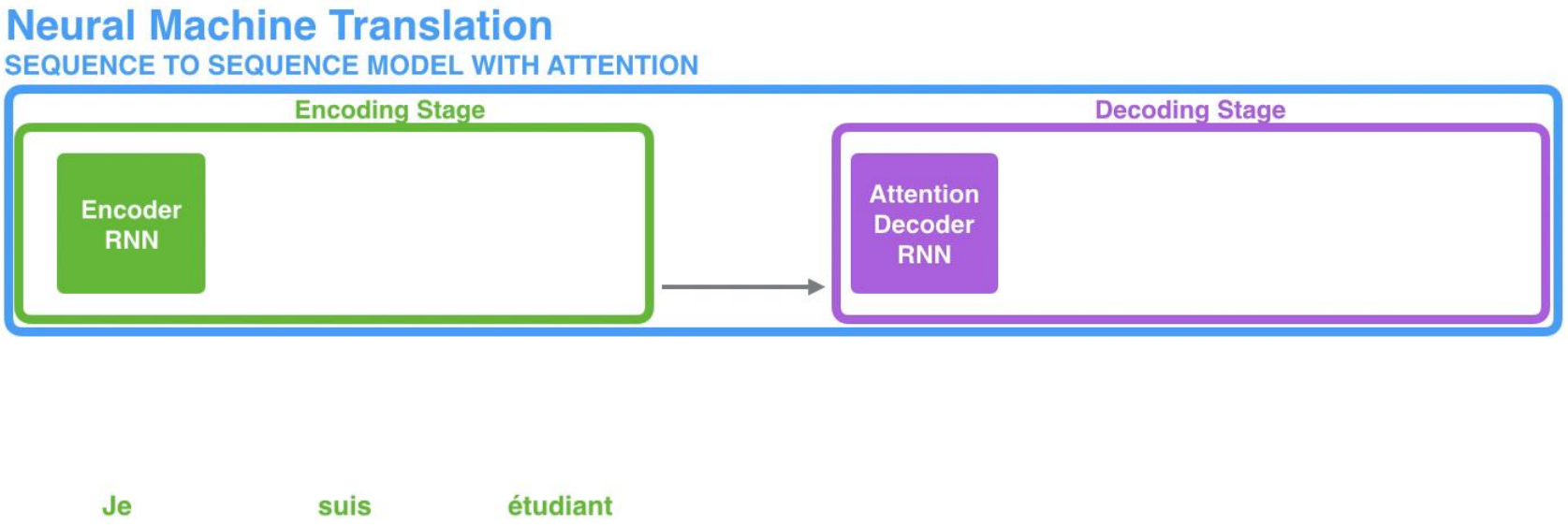


Luong attention



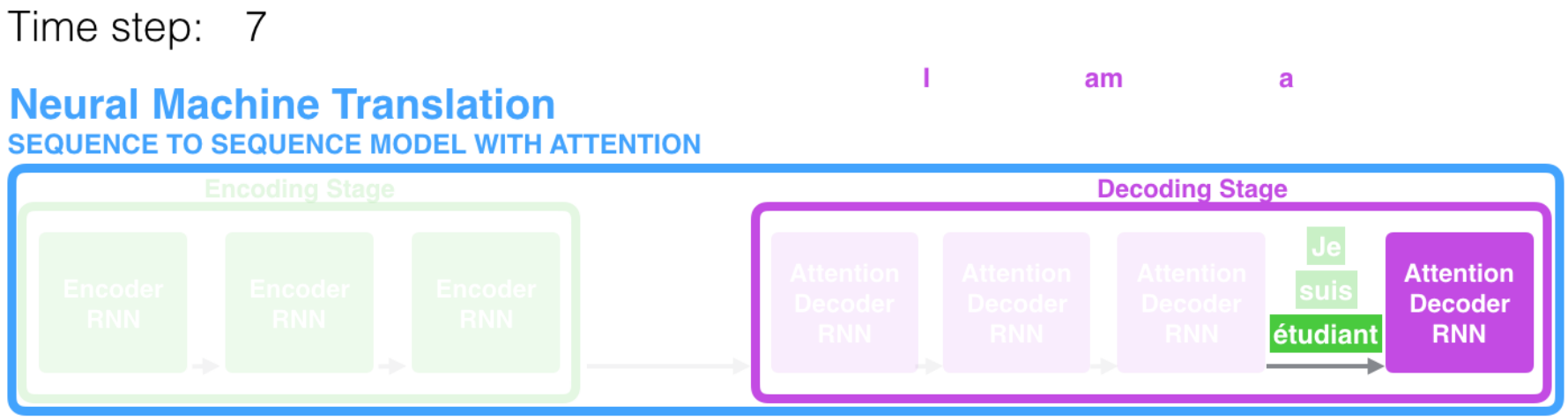
Attention in Seq2Seq Learning

- Attention model differs from a classic Seq2Seq model in two main ways:
  - ✓ The encoder passes a lot more data to the decoder
    - Instead of passing the last hidden state of the encoding stage, the encoder passes all the hidden states to the decoder



Attention in Seq2Seq Learning

- Attention model differs from a classic Seq2Seq model in two main ways:
  - ✓ The encoder passes a lot more data to the decoder
    - Instead of passing the last hidden state of the encoding stage, the encoder passes all the hidden states to the decoder



Attention in Seq2Seq Learning

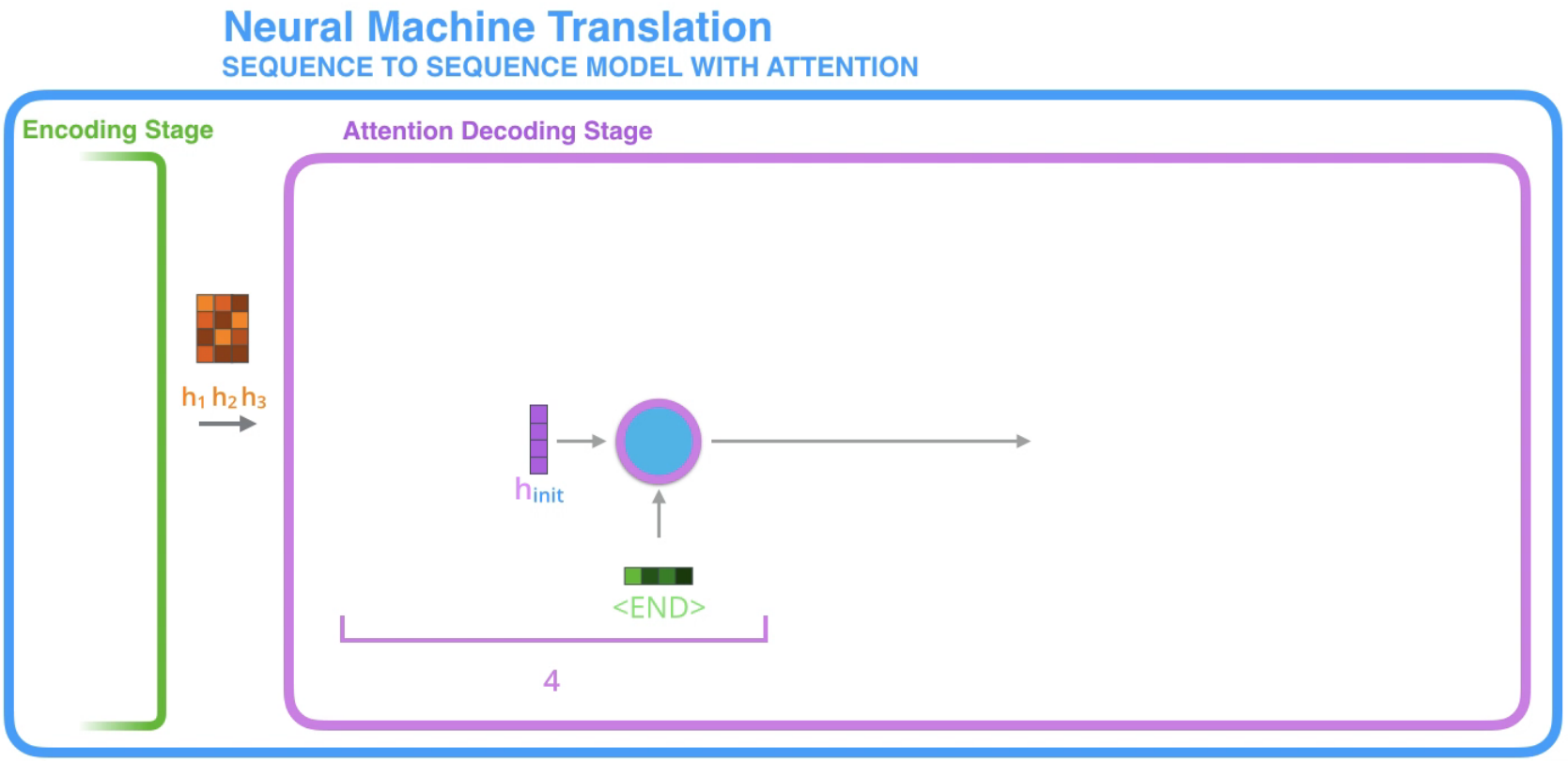
- Attention model differs from a classic Seq2Seq model in two main ways:
  - ✓ An attention decoder does an extra step before producing its output
    - Look at the set of encoder hidden states it received – each encoder hidden states is most associated with a certain word in the input sentence
    - Give each hidden states a score
    - Multiply each hidden state by its softmaxed score, this amplifying hidden state with high scores, an drowning out hidden state with low scores

Attention in Seq2Seq Learning

- Working mechanism of attention process
  - ✓ The attention decoder RNN takes in the embedding of the <END> token, and an initial decoder hidden state.
  - ✓ The RNN processes its inputs, producing an output and a new hidden state vector (h4). The output is discarded.
  - ✓ Attention Step: We use the encoder hidden states and the h4 vector to calculate a context vector (C4) for this time step.
  - ✓ We concatenate h4 and C4 into one vector.
  - ✓ We pass this vector through a feedforward neural network (one trained jointly with the model).
  - ✓ The output of the feedforward neural networks indicates the output word of this time step.
  - ✓ Repeat for the next time steps

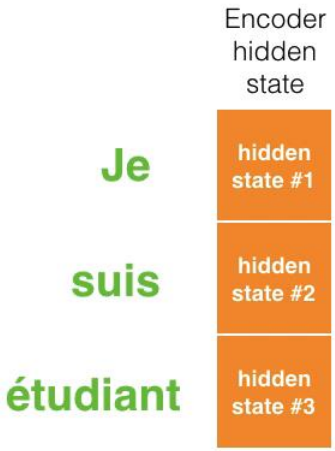
Attention in Seq2Seq Learning

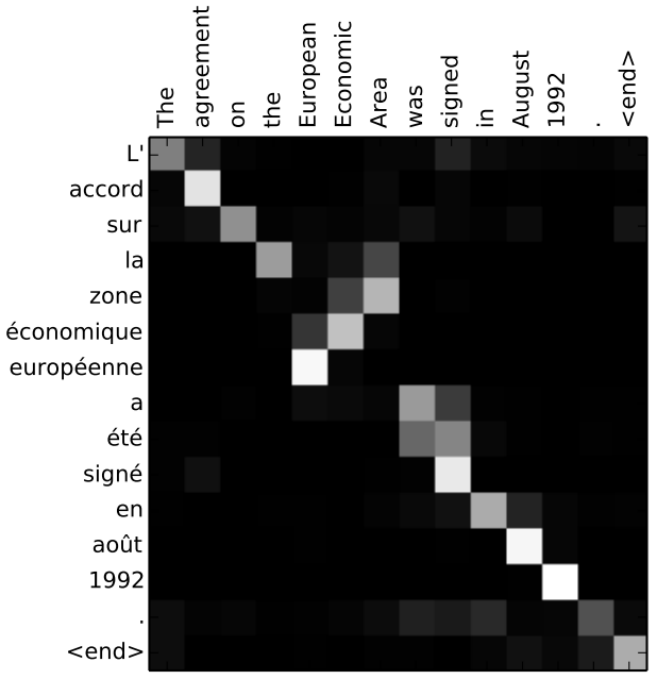
- Working mechanism of attention process



Attention in Seq2Seq Learning

- Working mechanism of attention process






ANY
questions?