Sequence to sequence (seq2seq)

Alberto TONDA, Ph.D. (Senior permanent researcher, DR)

UMR 518 MIA-PS, INRAE, AgroParisTech, Université Paris-Saclay
UAR 3611, Institut des Systèmes Complexes de Paris Île-de-France

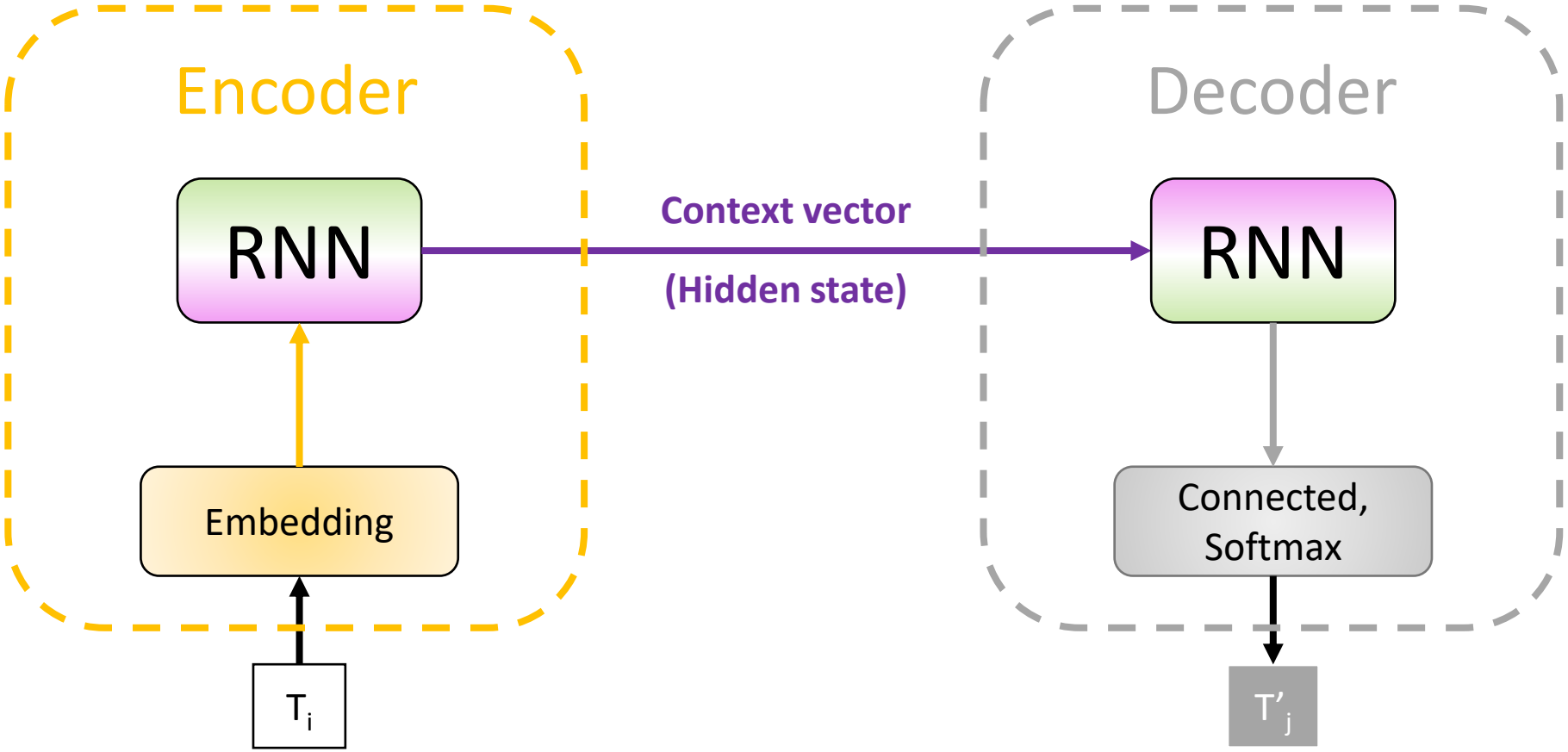
Outline

- Sequence to sequence
- Encoder/decoder architecture
- Encoder
- Decoder
- And now?

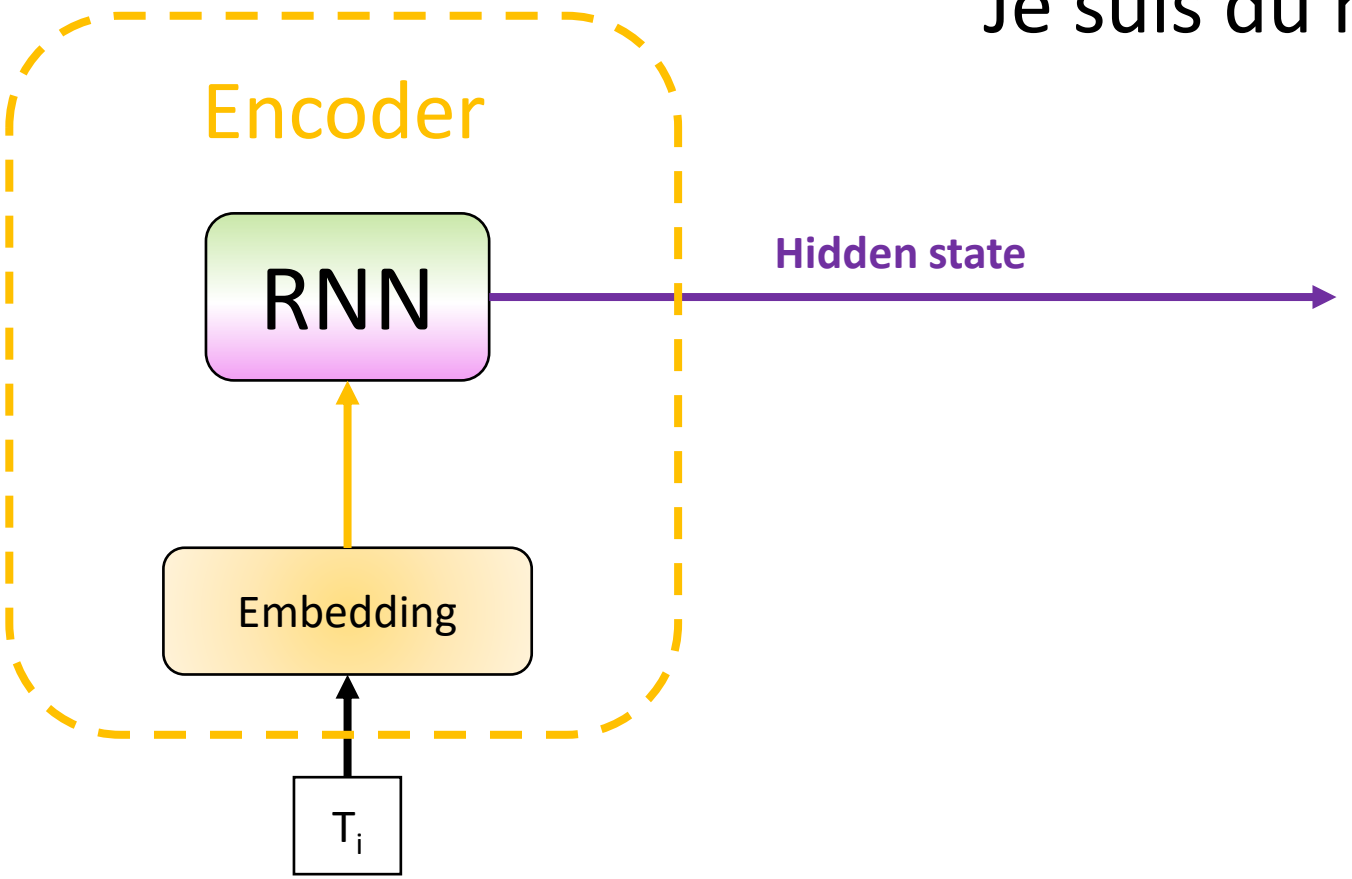
Sequence to sequence

- Convert a sequence to another sequence
  - Useful for many practical applications (machine translation!)
  - Both input and output sequences of *arbitrary length*
  - How can a neural network deal with **variable inputs/outputs**?

Encoder/decoder architecture



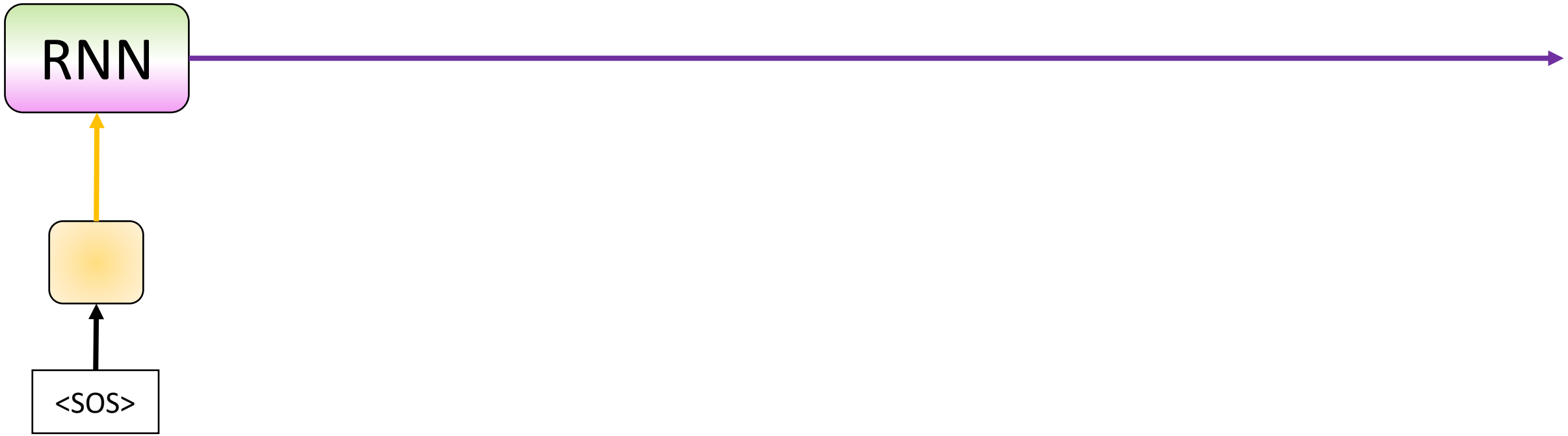
Encoder



"Je suis du même avis"

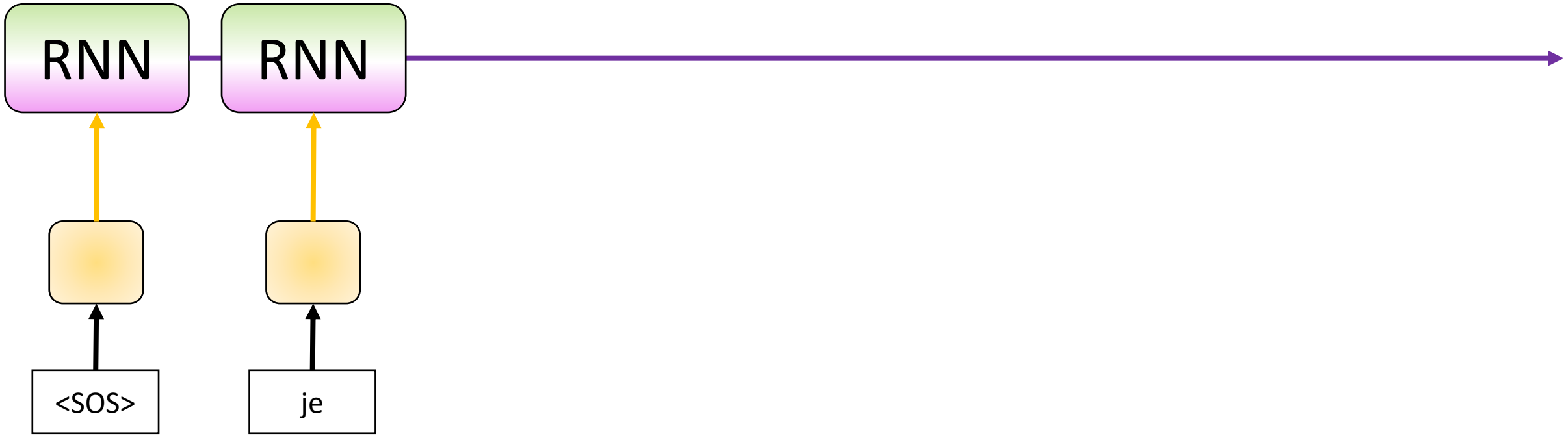
Encoder

“Je suis du même avis”



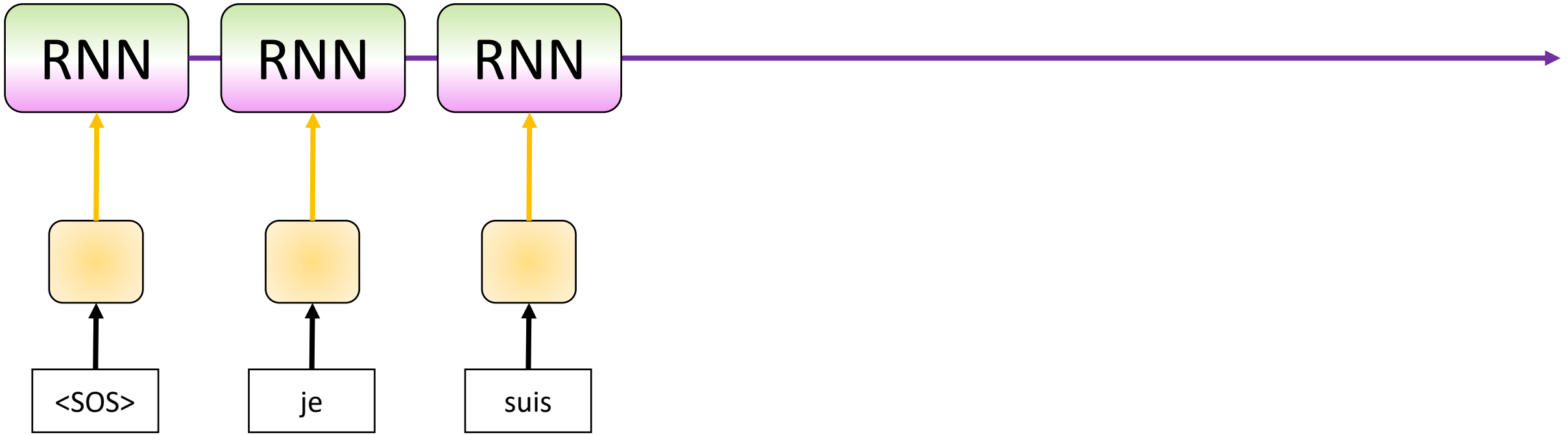
Encoder

“Je suis du même avis”



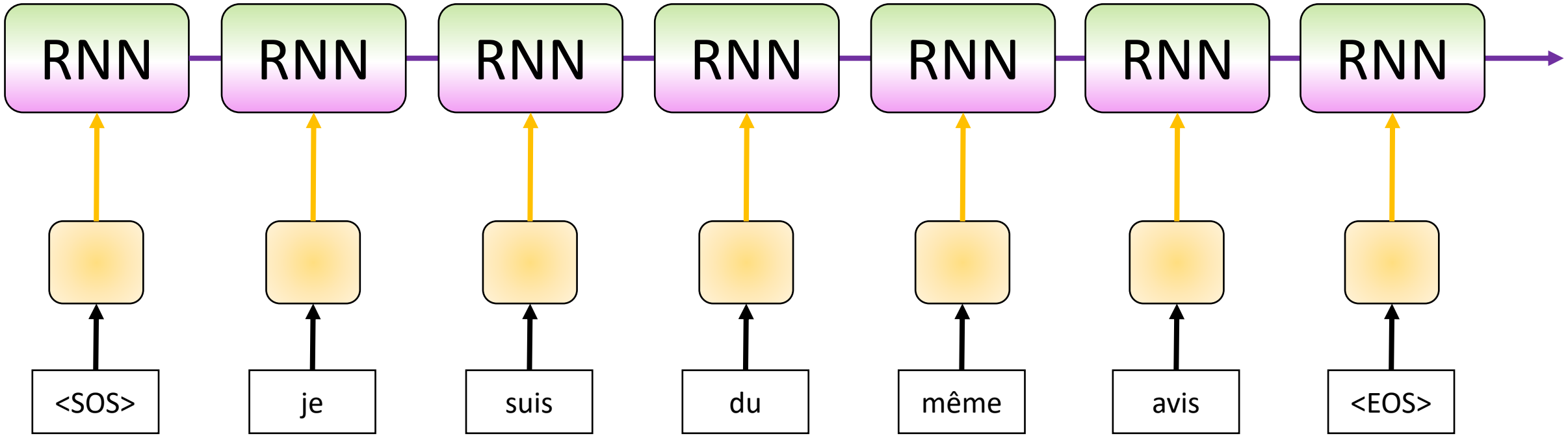
Encoder

“Je suis du même avis”

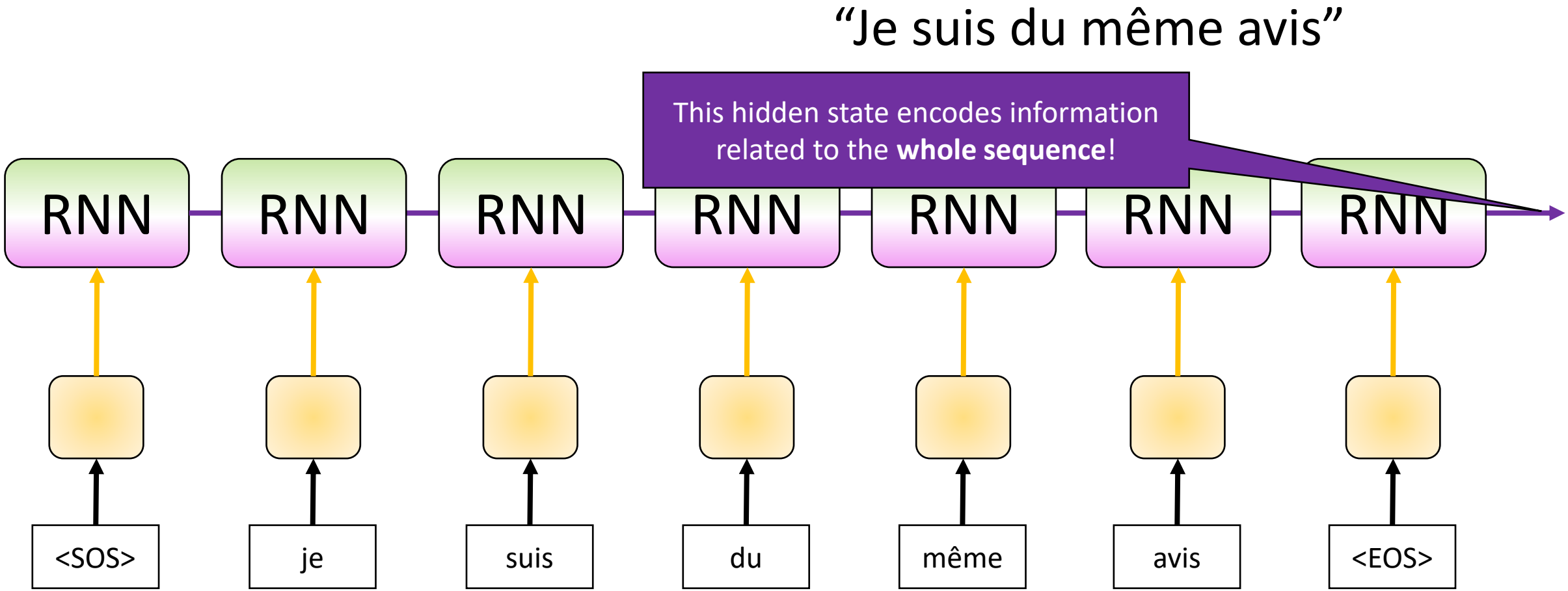


Encoder

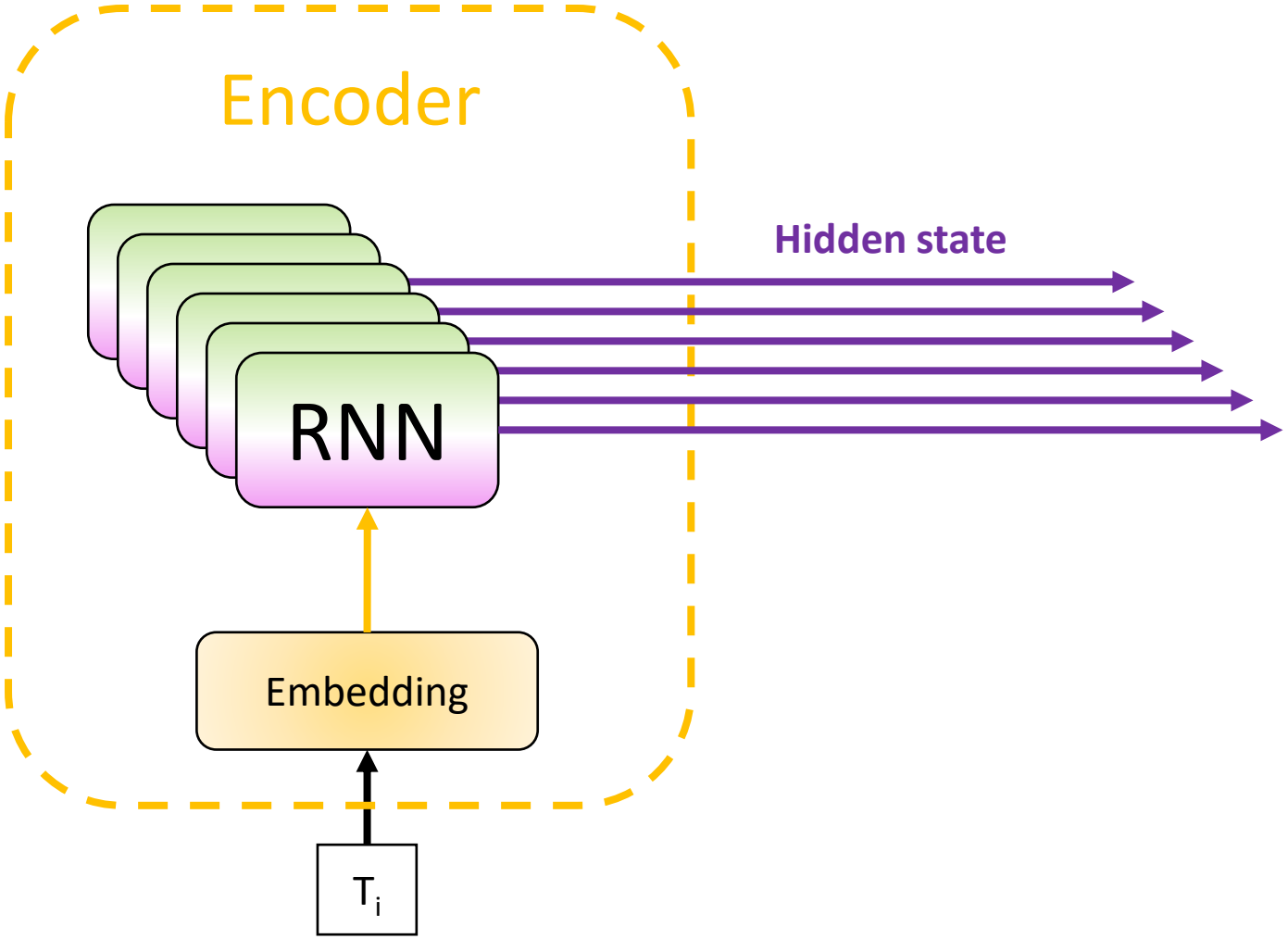
“Je suis du même avis”



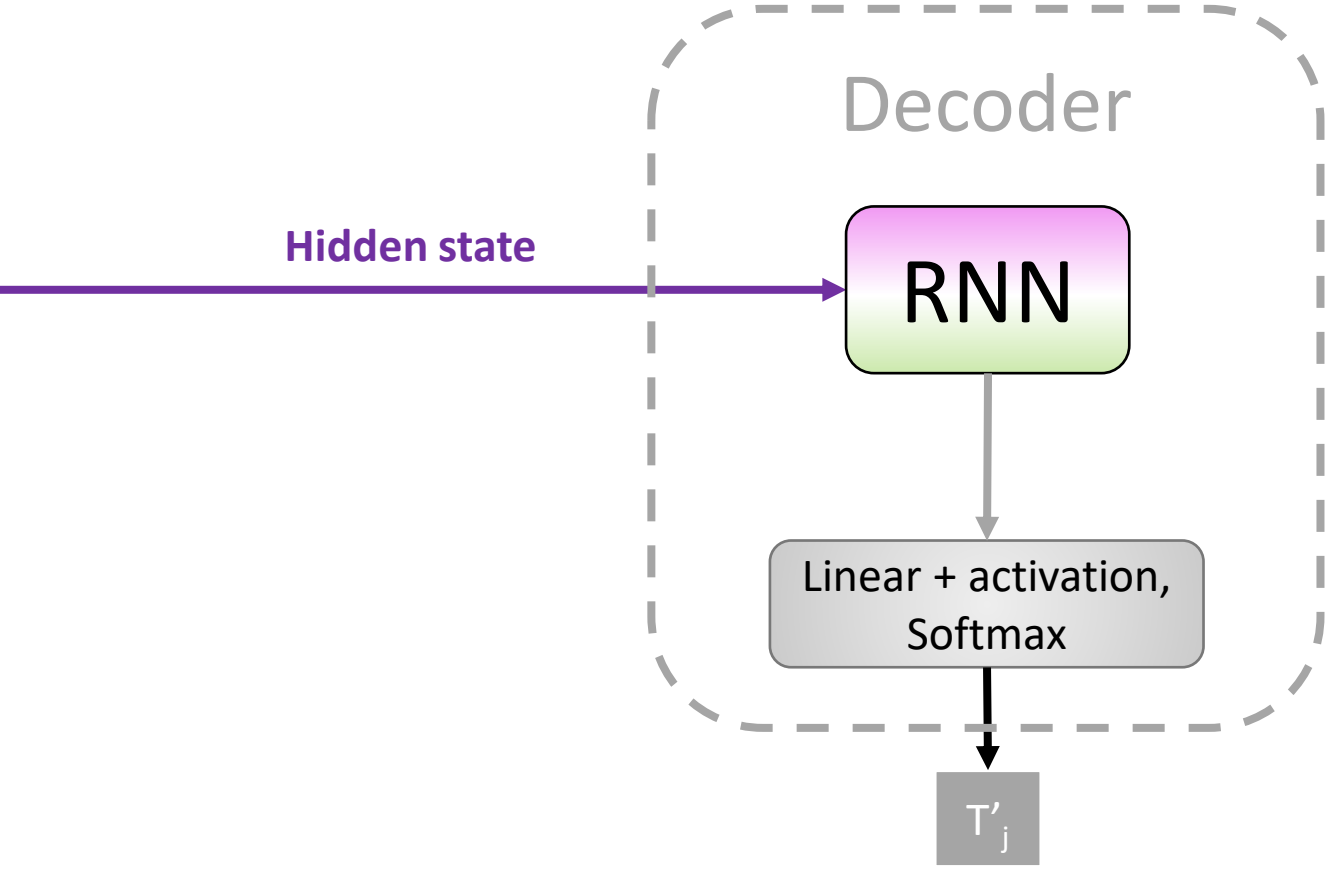
Encoder


“Je suis du même avis”
This hidden state encodes information related to the **whole sequence**!
RNN
RNN
RNN
RNN
RNN
RNN
RNN
<SOS>
je
suis
du
même
avis
<EOS>

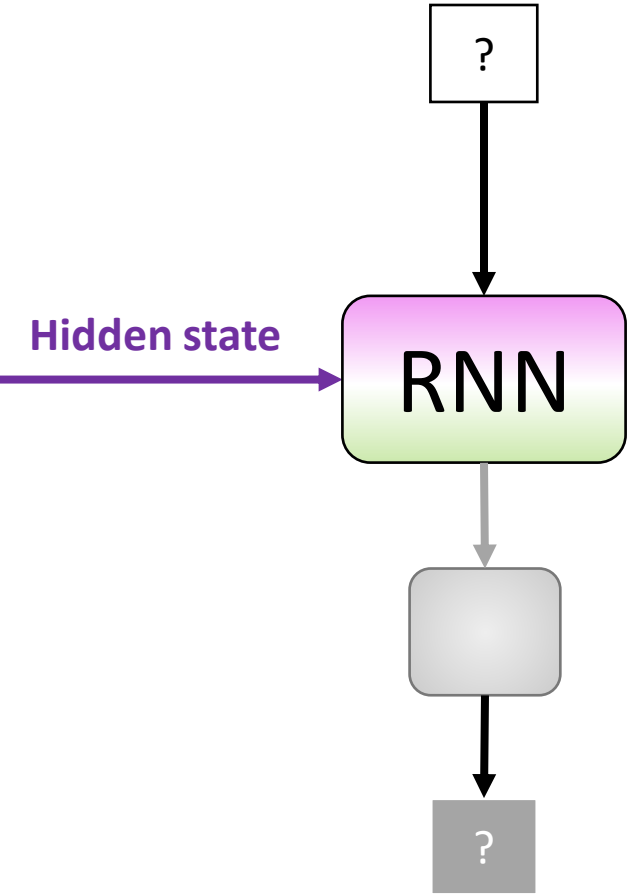
Encoder



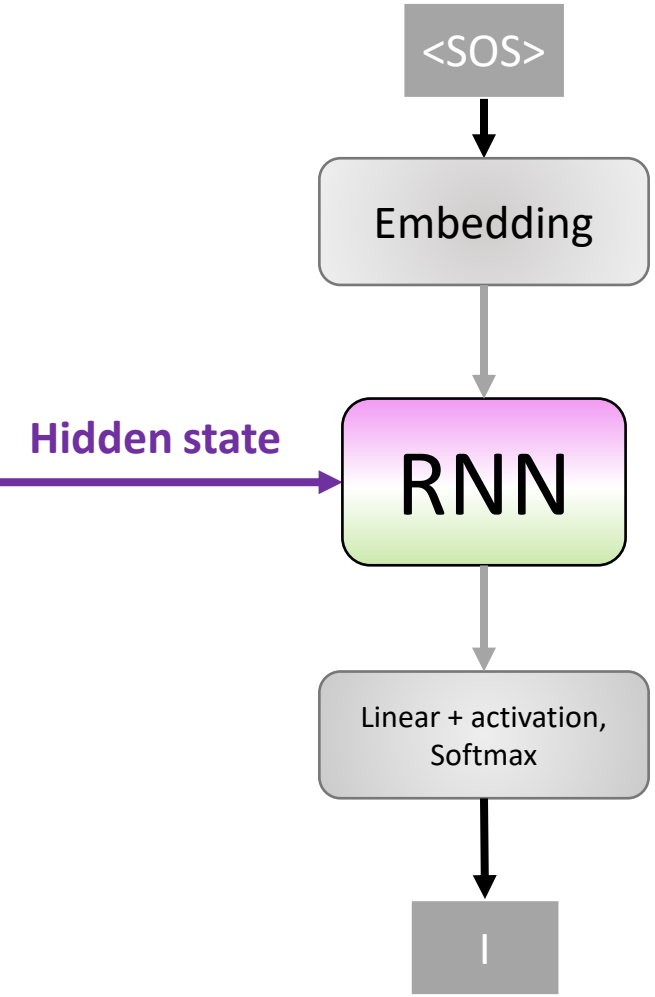
Decoder



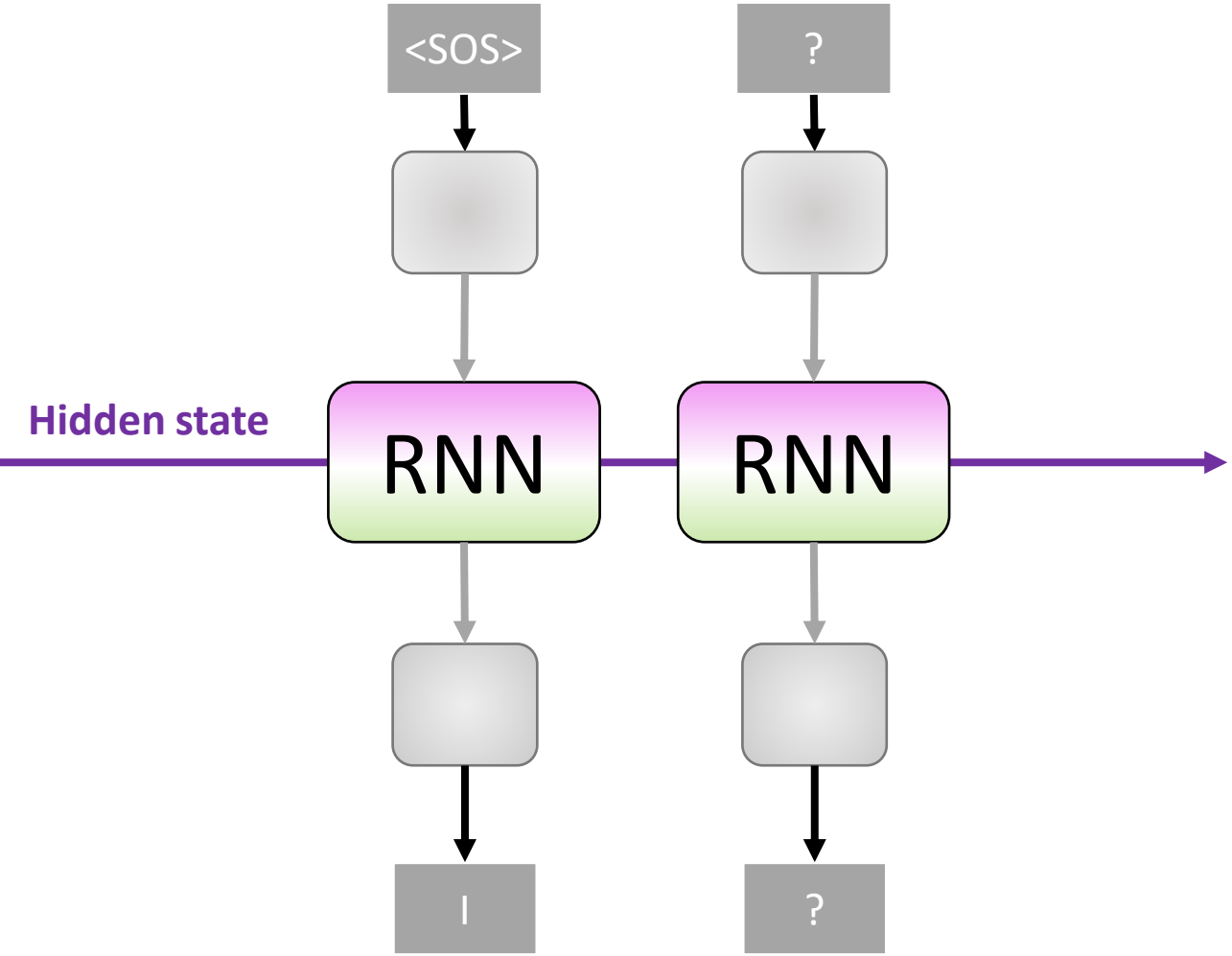
Decoder



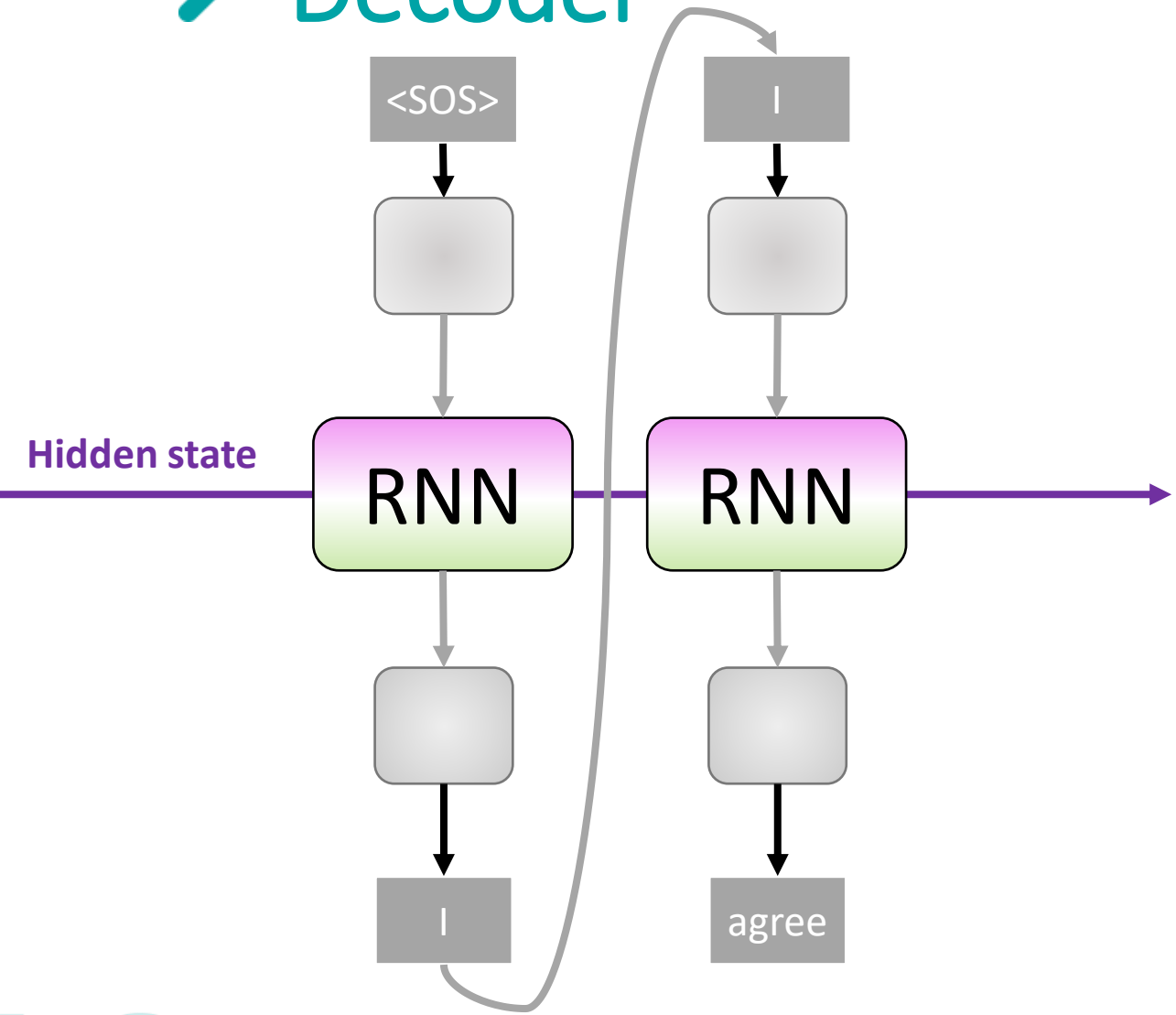
Decoder


<SOS>
Embedding
Hidden state
RNN
Linear + activation, Softmax
I

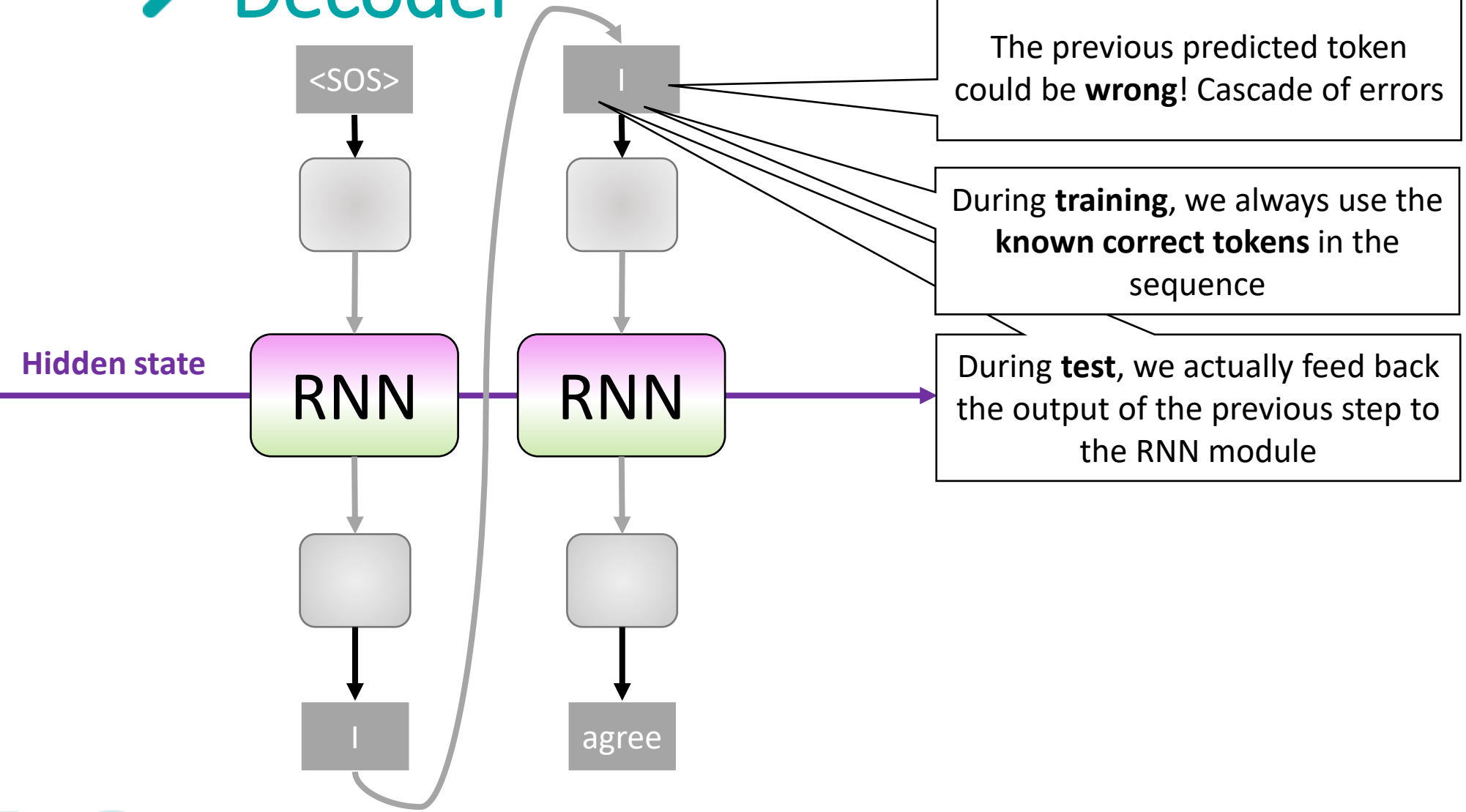
Decoder



Decoder



Decoder

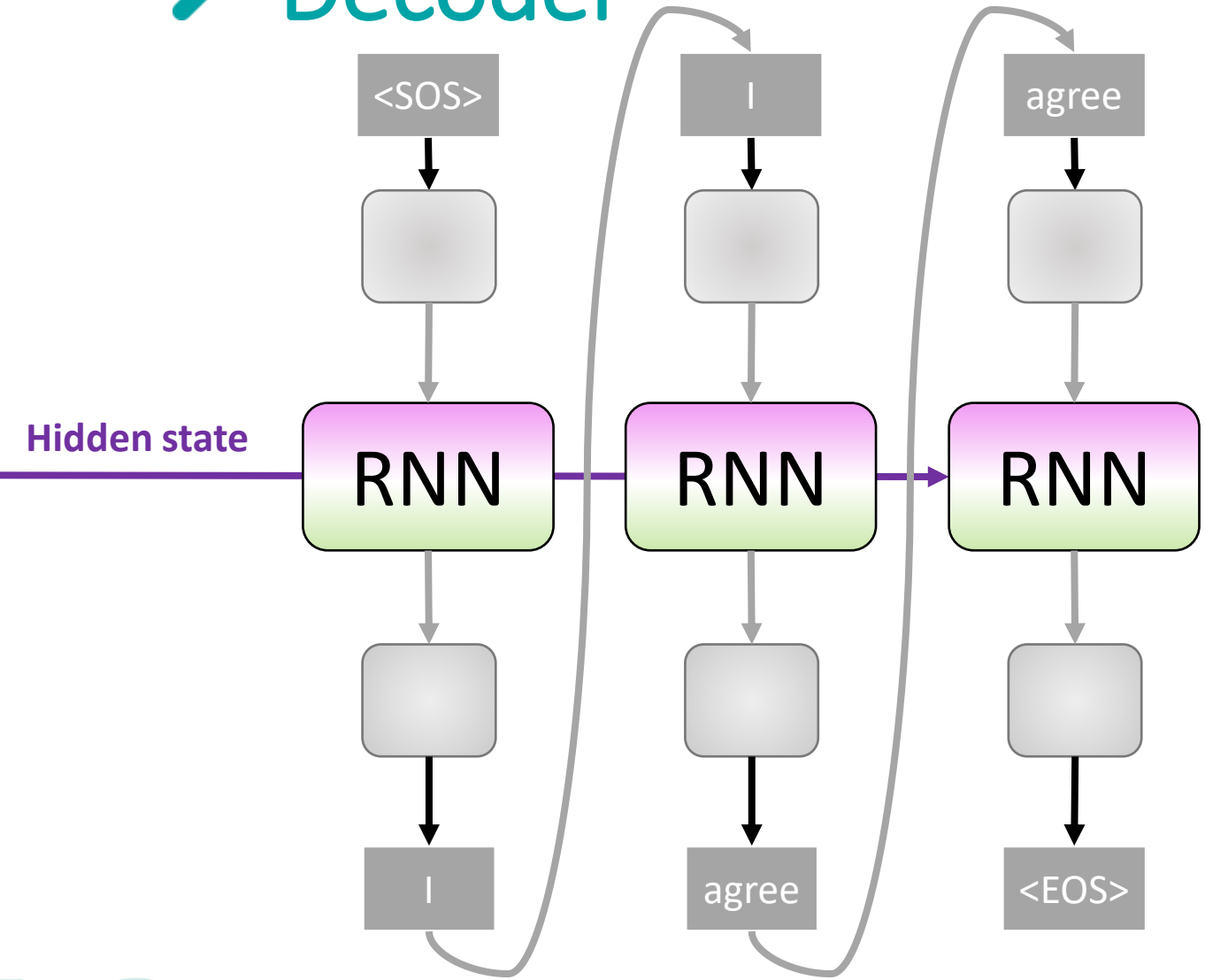


The previous predicted token could be **wrong**! Cascade of errors

During **training**, we always use the **known correct tokens** in the sequence

During **test**, we actually feed back the output of the previous step to the RNN module

Decoder



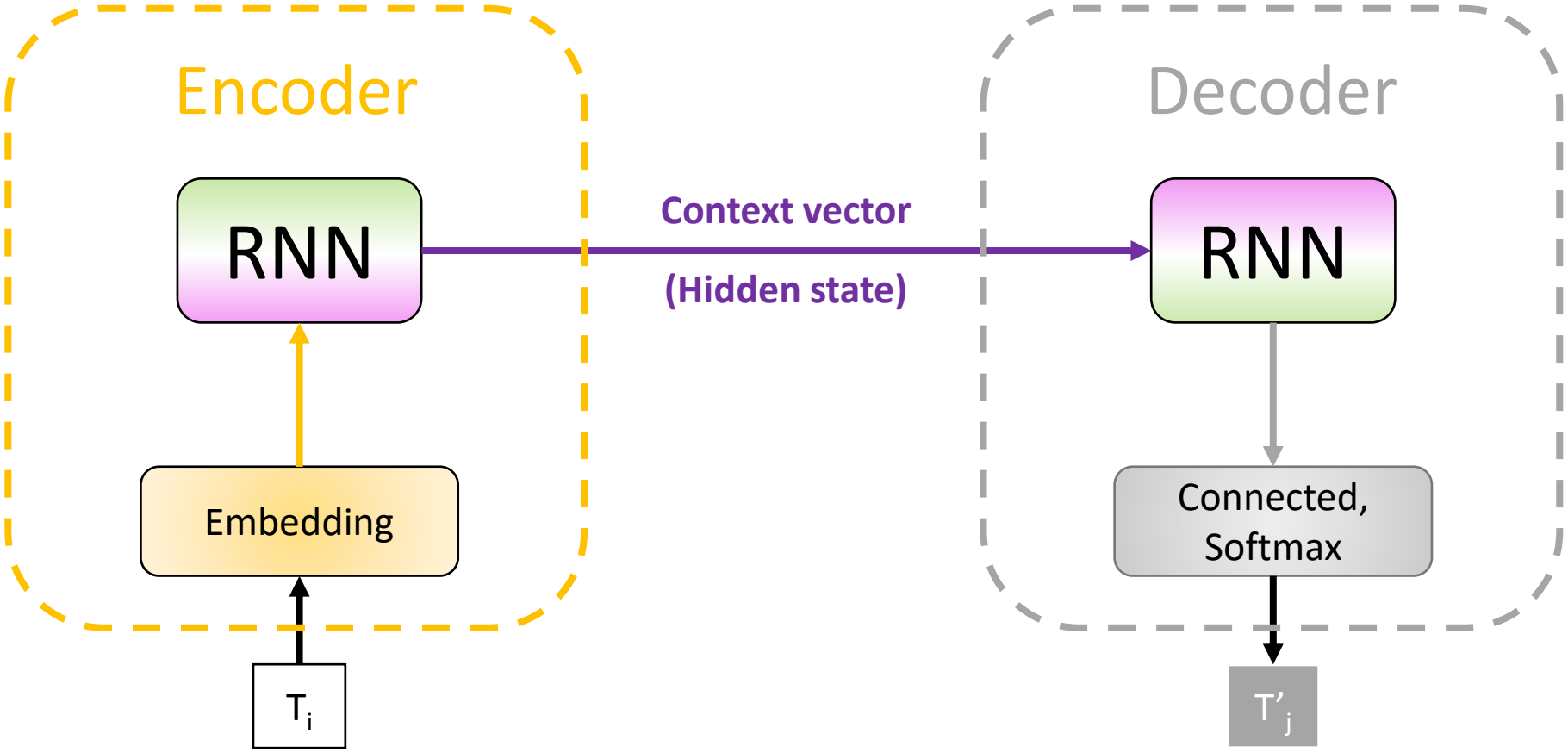
“Je suis du même avis”

<SOS>	je	suis	du	même	avis	<EOS>

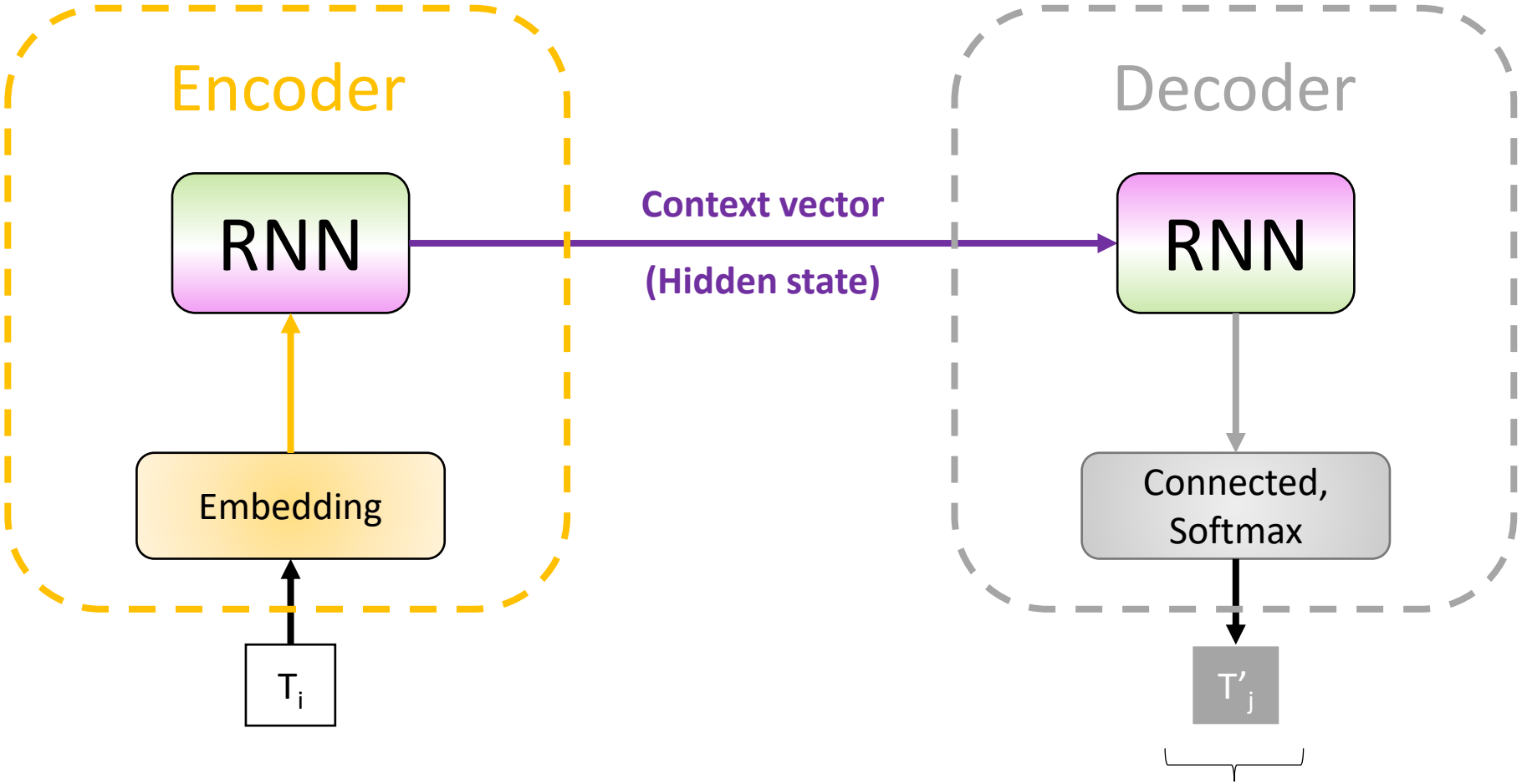
<SOS>	I	agree	<EOS>

“I agree”

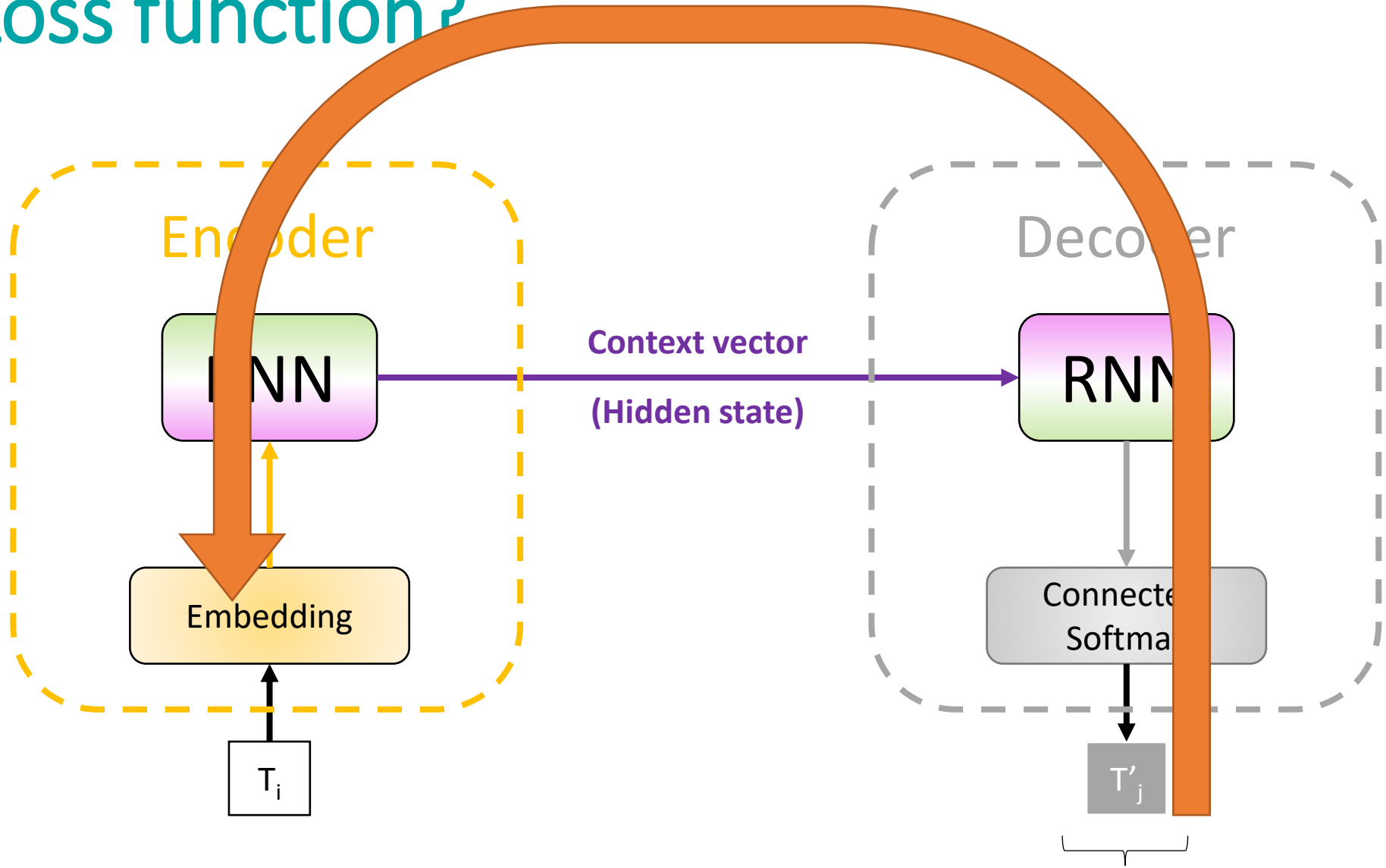
Loss function?



Loss function?



Loss function?



And now?

- Seq2seq architectures are still in use
  - Issues with *really long sequences*, loss of context/memory
  - For language, they have been replaced by Transformers
  - They just work better, and allow for higher parallelism

Questions?

Bibliography

- Bahdanau, D., Cho, K., & Bengio, Y. (2014). *Neural machine translation by jointly learning to align and translate*. arXiv preprint arXiv:1409.0473.

Images and videos: unless otherwise stated, I stole them from the Internet. I hope they are not copyrighted, or that their use falls under the Fair Use clause, and if not, I am sorry. Please don't sue me.