Sequence to Sequence Modelling with RNNs

(Encoder-Decoder Architecture, Attention Mechanism)

Machine Translation

Tirtharaj Dash

Dept. of CS & IS and APPCAIR
BITS Pilani, Goa Campus

November 11, 2021

In the last lecture:

1. RNN "cell"
2. Vanishing gradients
   - GRU
   - LSTM

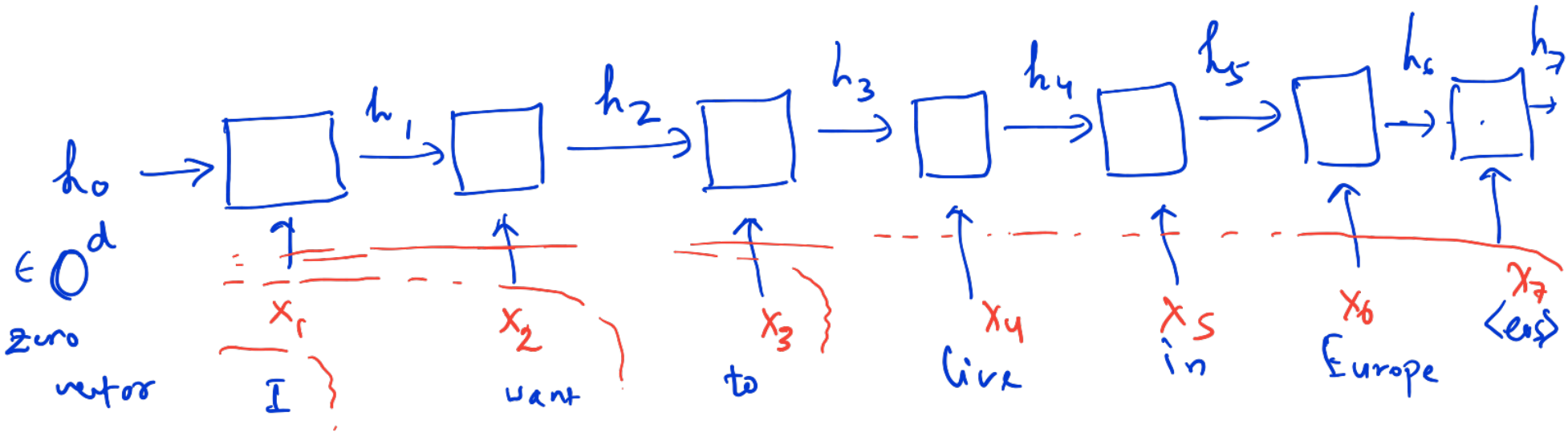
Seq-to-Seq Modelling: (Ref: TBI, Andrew Ng's course)

Machine Translation

English → French

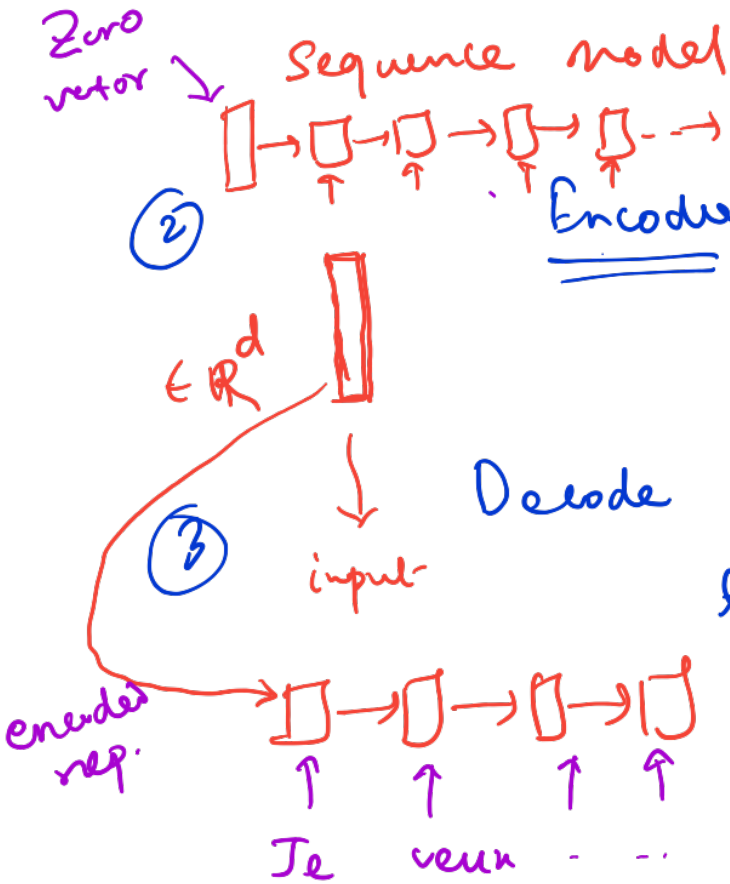
"I want to live in Europe" → "Je veux vivre en Europe"

↳ Summary → (French)



Steps:

1. Construct a RNN arch. to process the English sentence (creating a summary for the input)

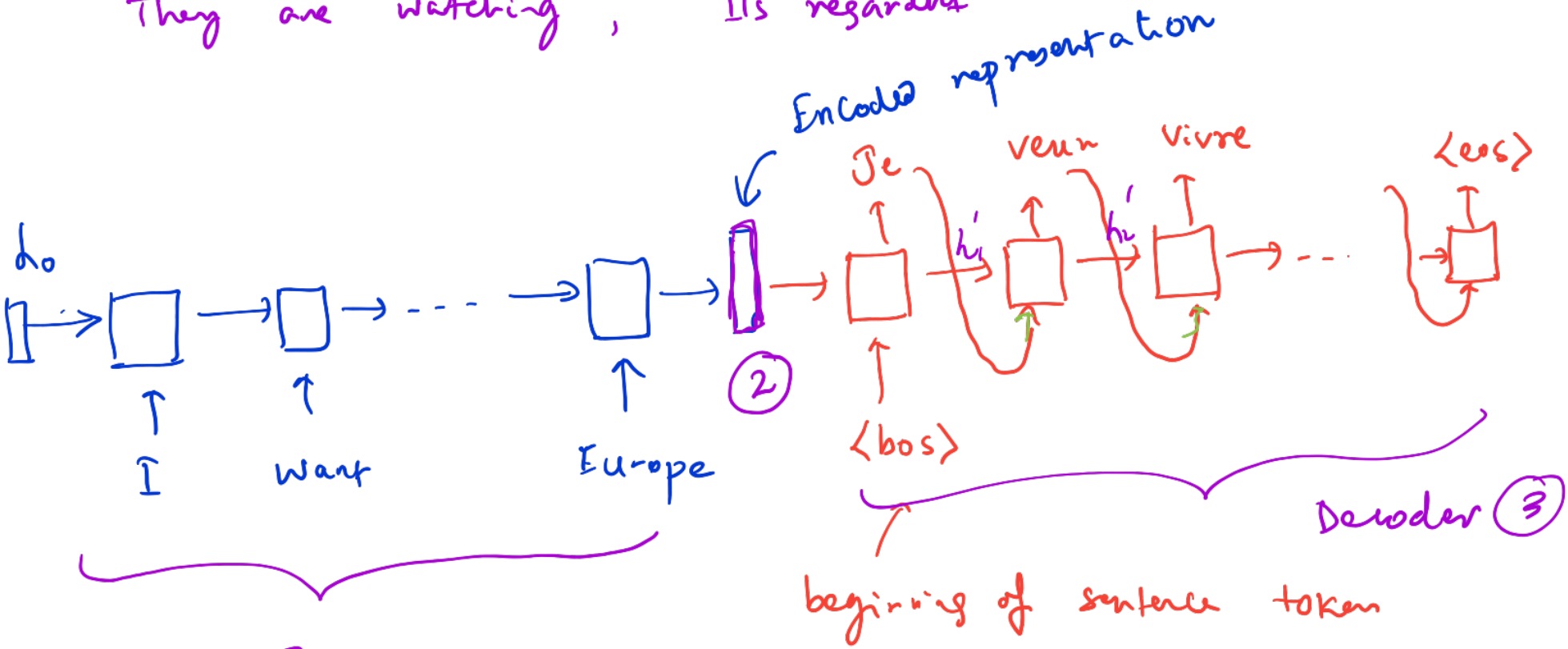


2. ↓ Encoded representation for the input

3. ↓ Decode the encoded representation into the target (output) language.

"I want to live in Europe", "Je veux vivre en Europa"

"They are watching", "Ils regardent"



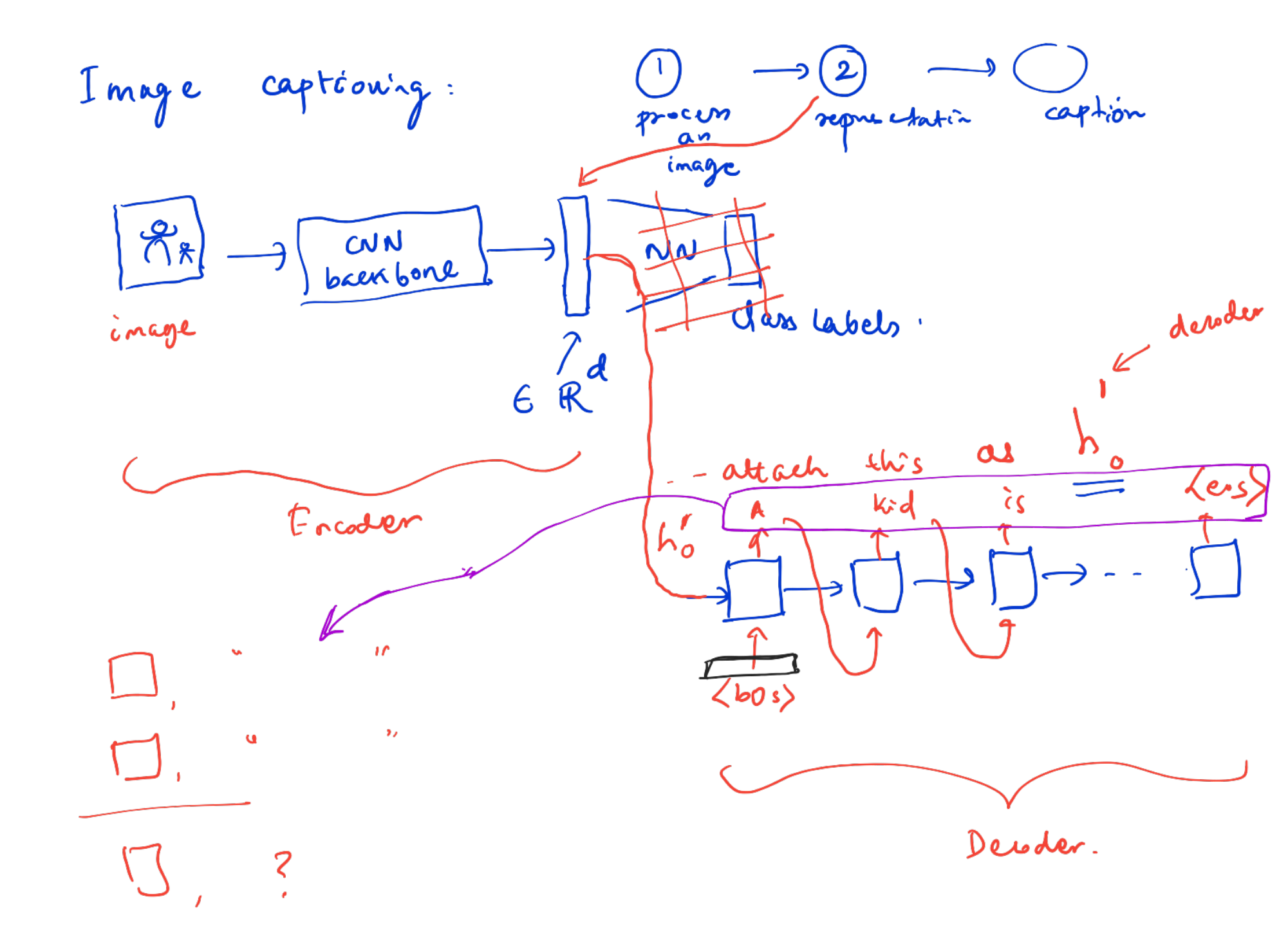
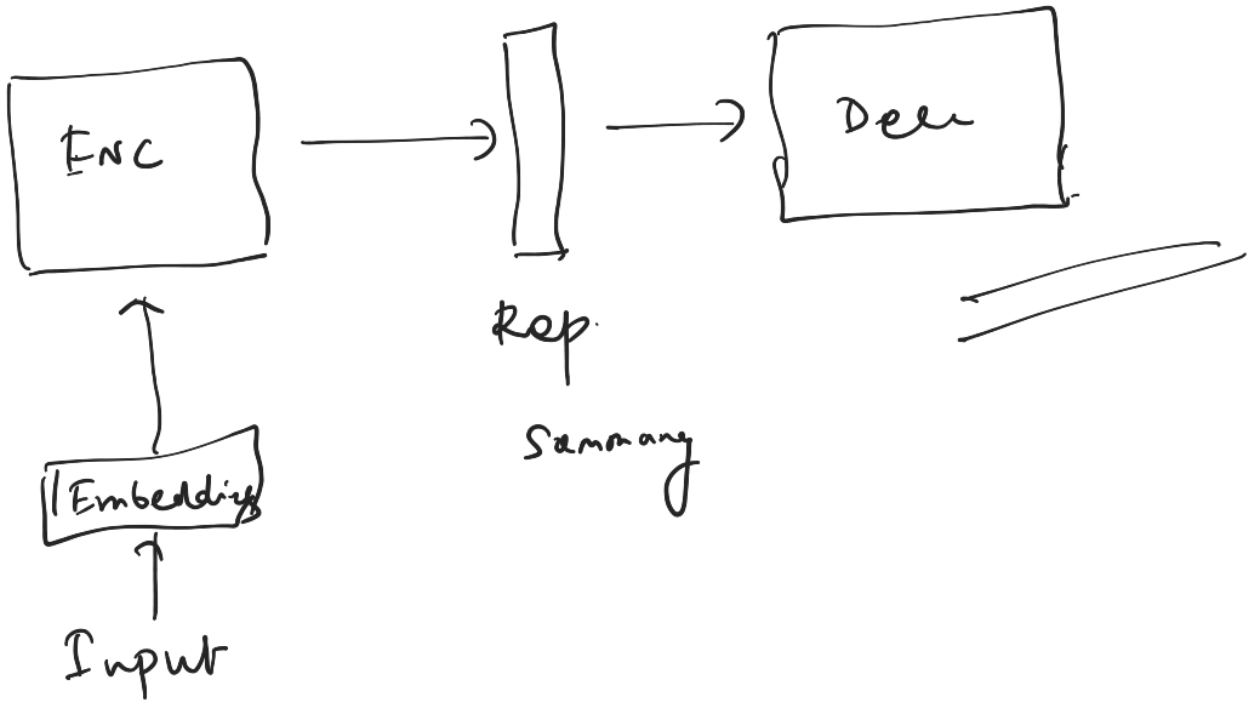


Image captioning :
1
2
process an image
representation
caption
CNN backbone
image
NN
class labels.
$\in \mathbb{R}^d$
decoder
Encoder
attach this as h_0
A
kid
is
<eos>
h_0'
<bos>
Decoder.


Enc
Dec
Rep
Embedding
Input

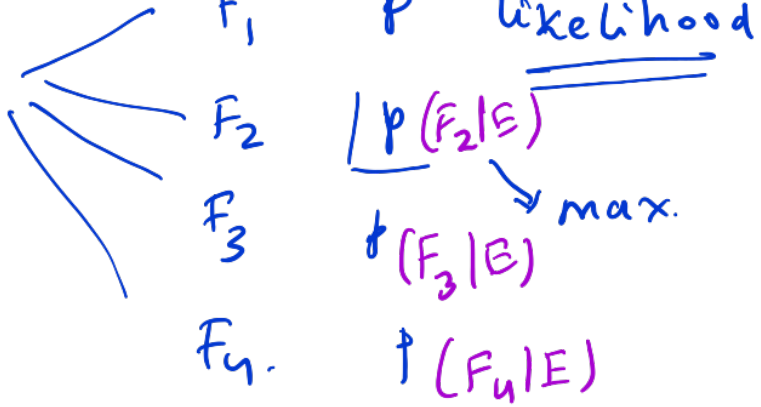
Machine Translation as a Conditional Language Model.

input

$$p(x_1 x_2 \cdots x_T) = \prod_{t=1}^{T} p(x_t | x_1 \cdots x_{t-1})$$

$$p(\text{French sentence} \mid \text{English sentence})$$

Given any input E (in English)



$$\arg\max_{y_1 y_2 \cdots y_{T_y}} p(y_1 y_2 \cdots y_{T_y} \mid \underline{x})$$

$\underline{x}$ ↑ English sentence $x_1 x_2 \cdots x_{T_x}$.

$$\arg\max_i p(F_i \mid \underline{x})$$

→ Idea: most likely : Greedy search

F_1 — 0.45 (optimal) ?

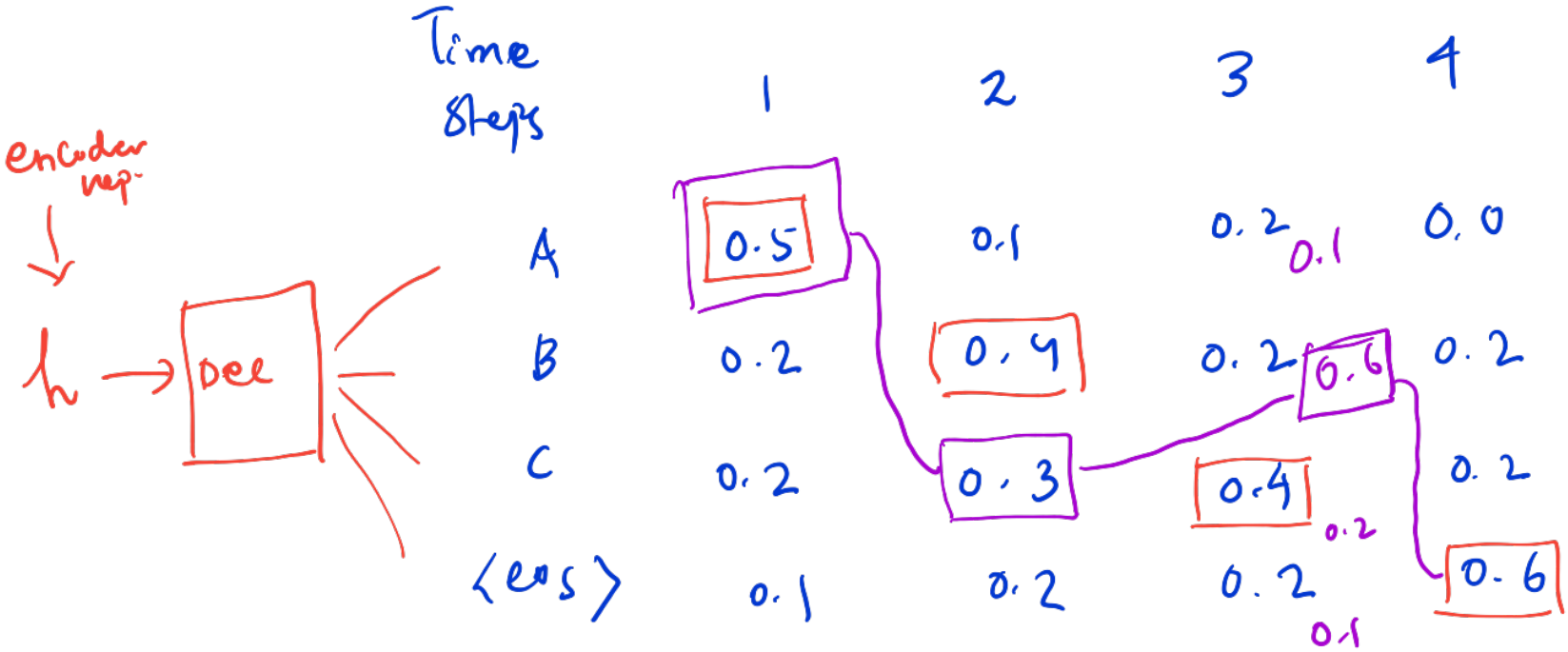
F_3 — 0.42

"Translation"

optimal sequence is not guaranteed.

Let we have 4 output tokens

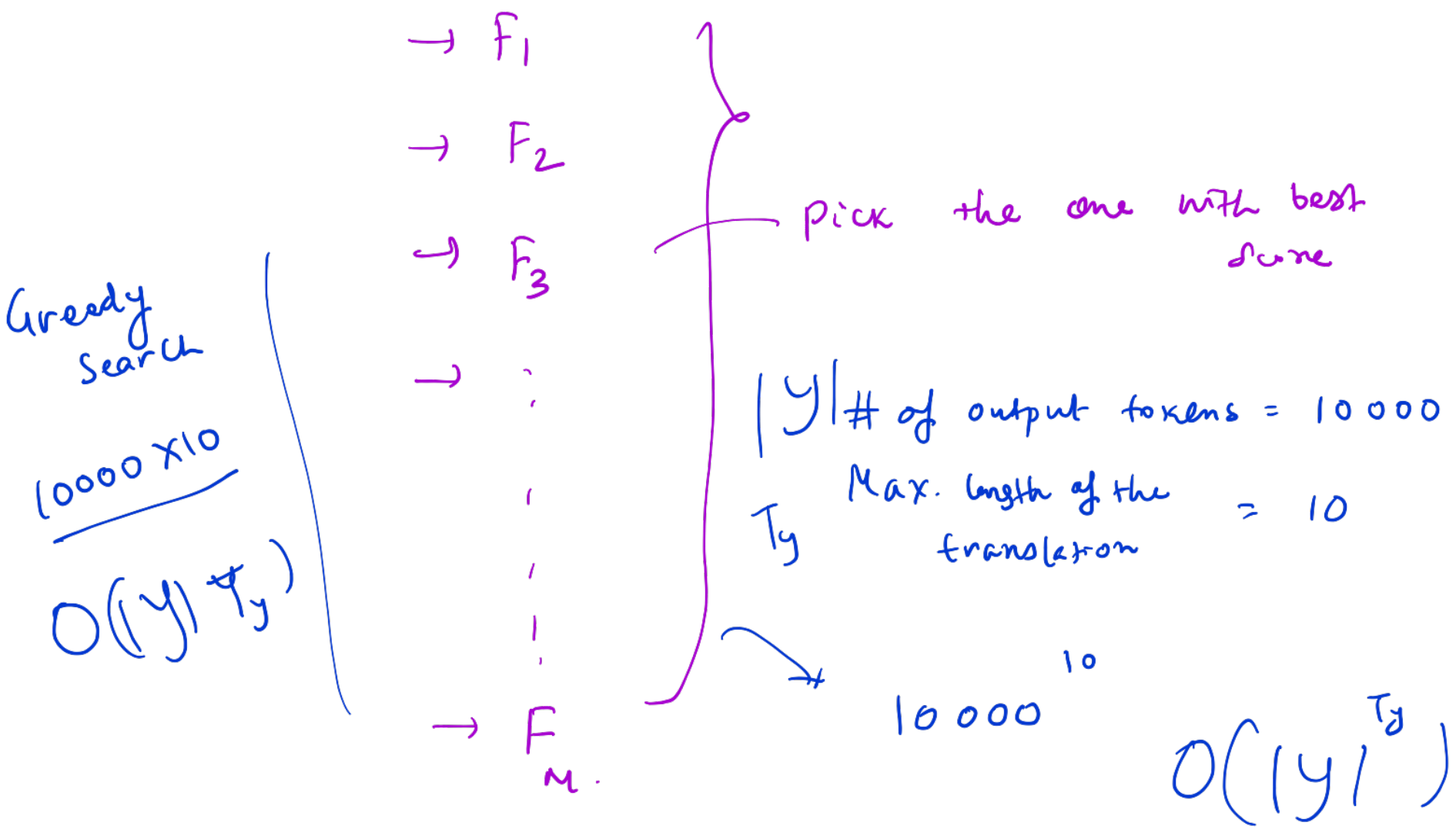
"A" "B" "C" "<eos>"

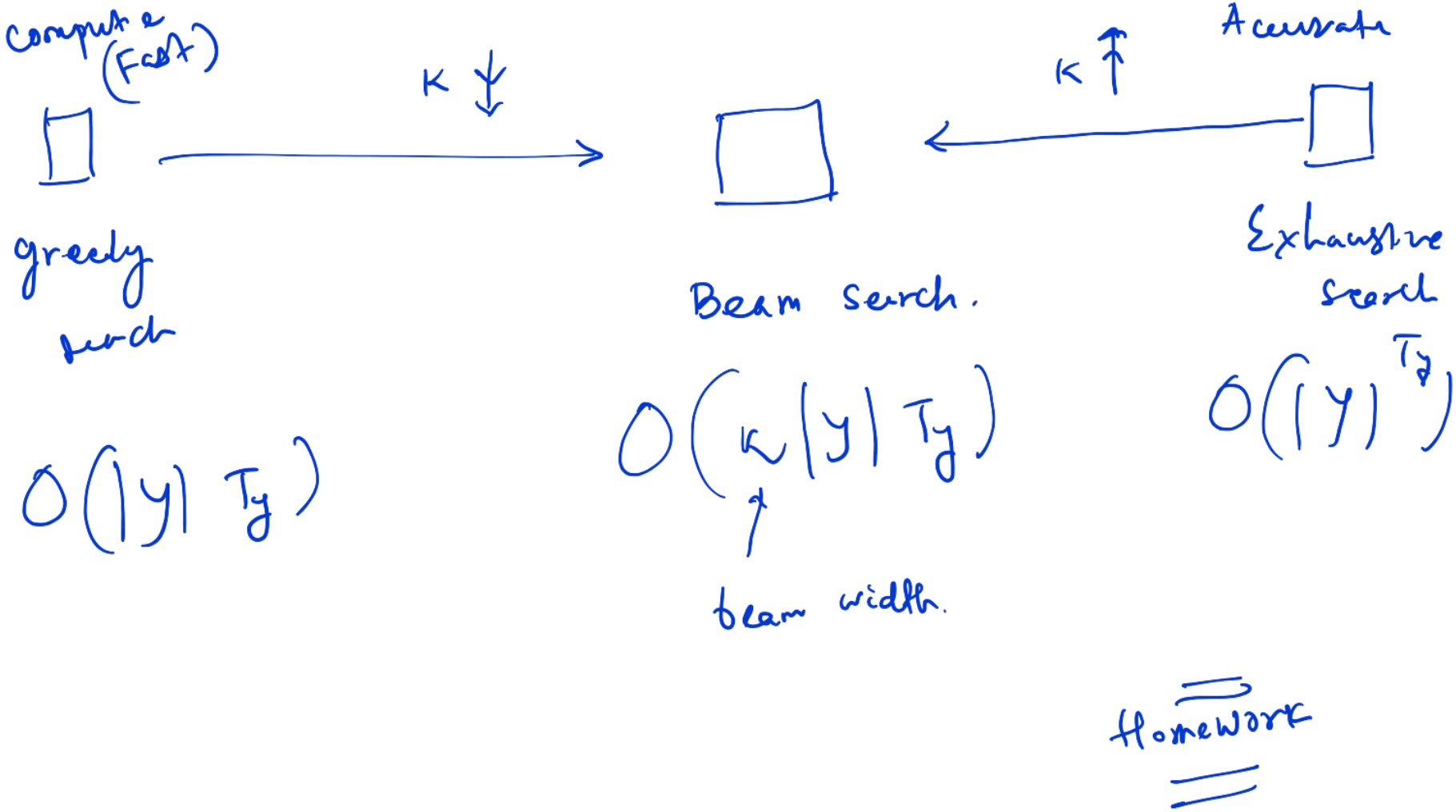


Greedy search: $P(ABC\langle eos\rangle \mid h) = 0.5 \times 0.4 \times 0.2 \times 0.6 = 0.024$

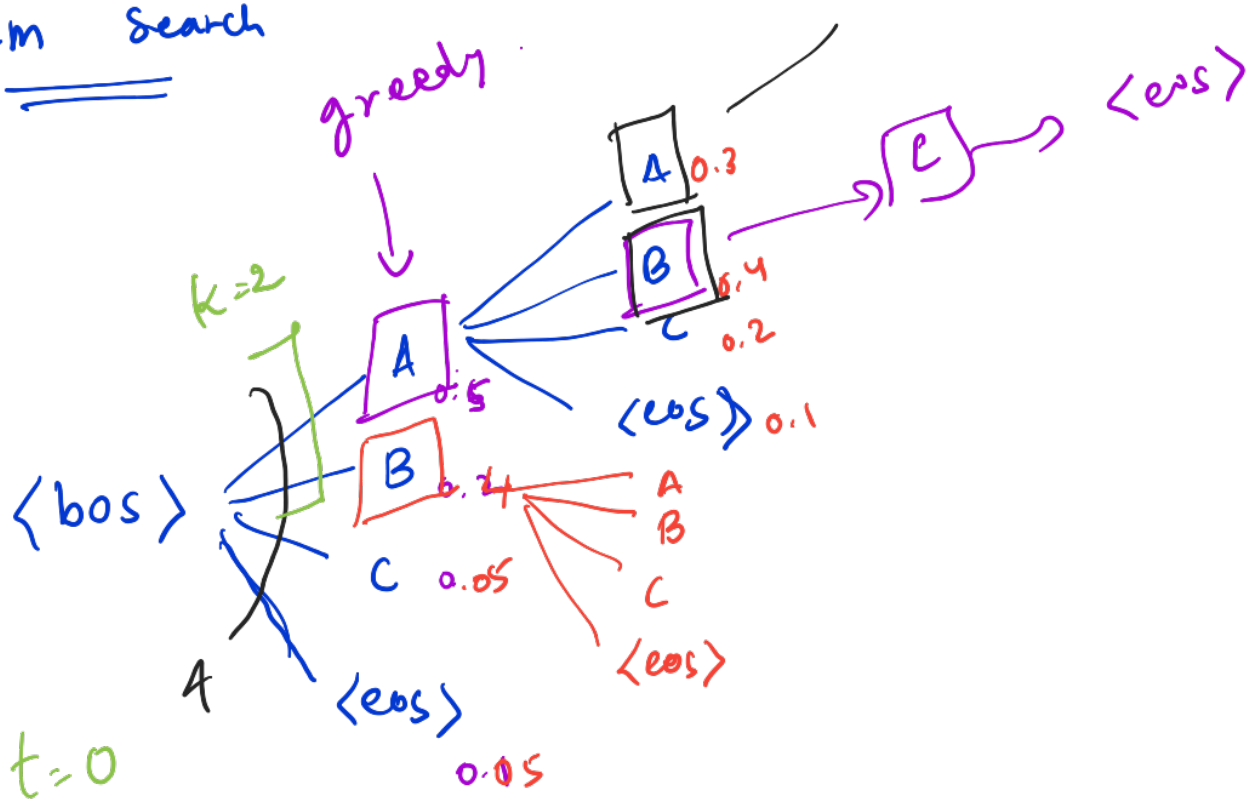
: $P(ACB\langle eos\rangle \mid h) = 0.5 \times 0.4 \times 0.6 \times 0.6 = 0.054$

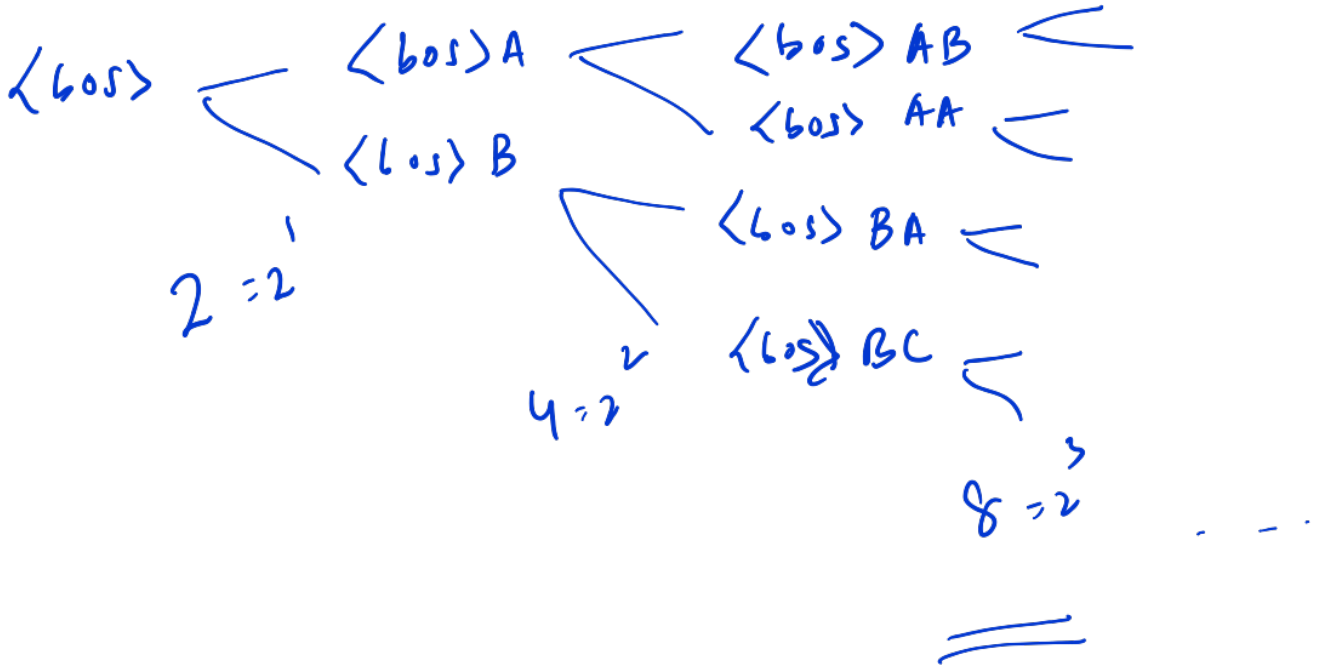
Solution: Exhaustive search ✗




Computa
(Fast)
K
K
Accurate
greedy
search
Beam search.
Exhaustive
search
O(|y| Ty)
O(K|y| Ty)
O(|y|^Ty)
beam width.
Homework

Beam search


greedy
k=2
<bos>
A
B
C
<eos>
0.5
0.4
0.05
0.05
A
B
C
<eos>
0.3
0.4
0.2
0.1
<eos>
"ABC<eos>"
t=0
t=1
t=2
t=3
... t=Ty.
K=2



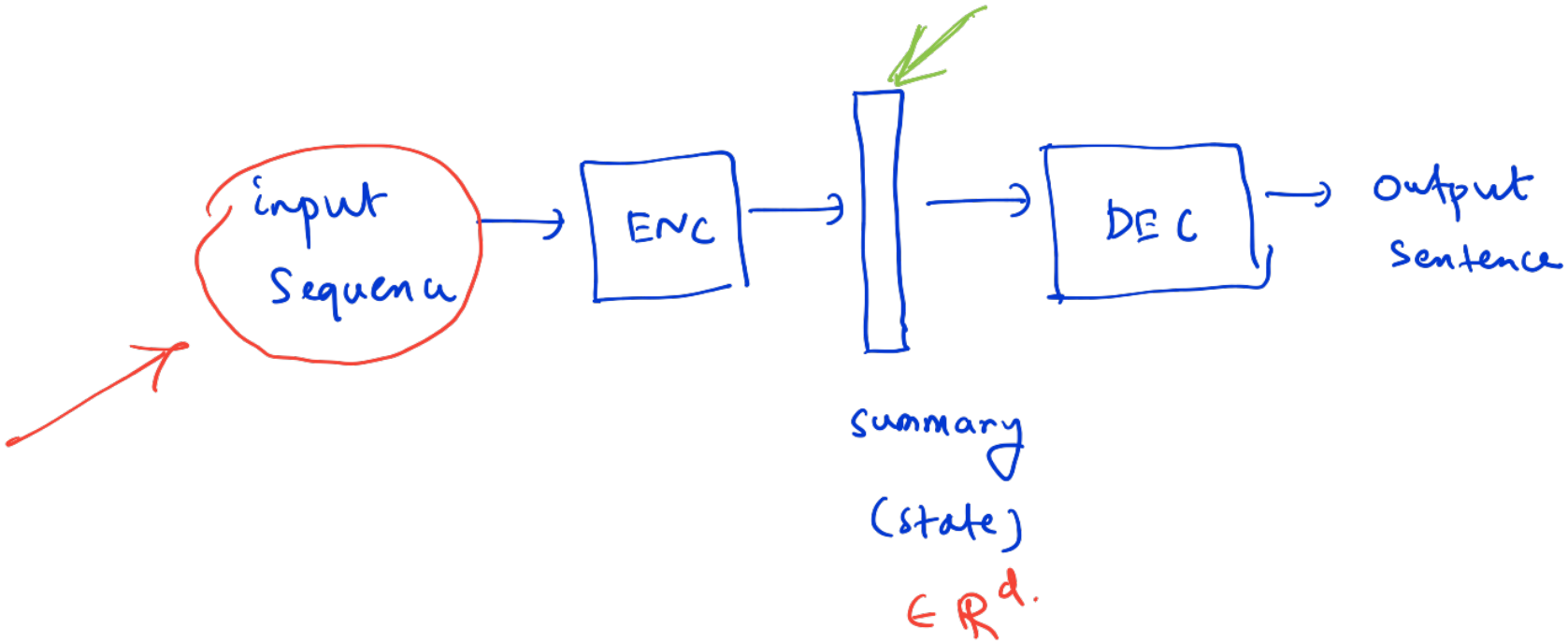
Is it $O(k|y|T_y)$?

Home work Read about Beam search & How the architectures of the Decoder will change to implement Beam search ??

ATTENTION MECHANISM

Seq 2 seq learning. 2014

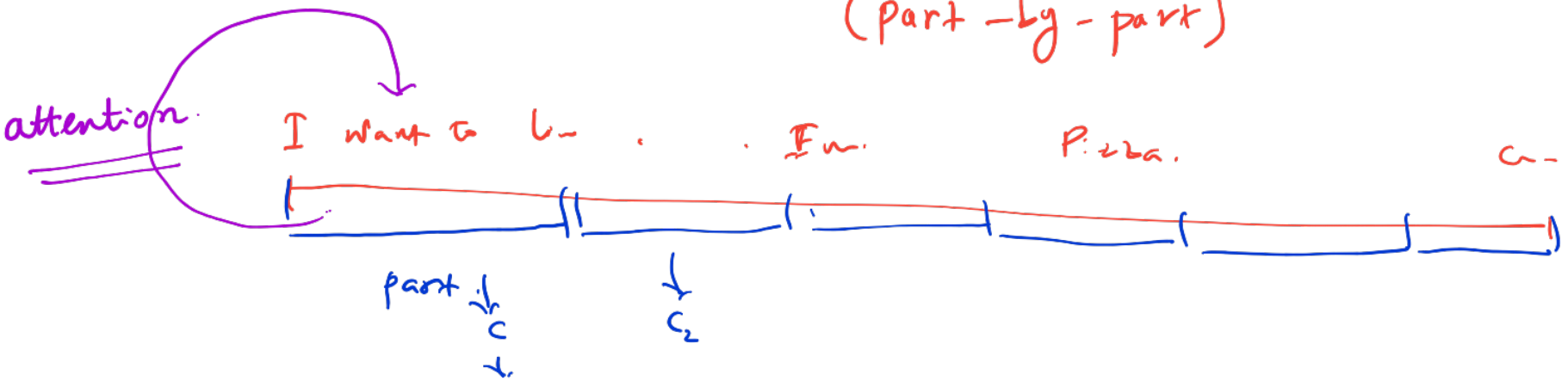
Machine translation.



Issues: Long sentences.

Solution: (1) Bidirectional RNN < ENC / DEC

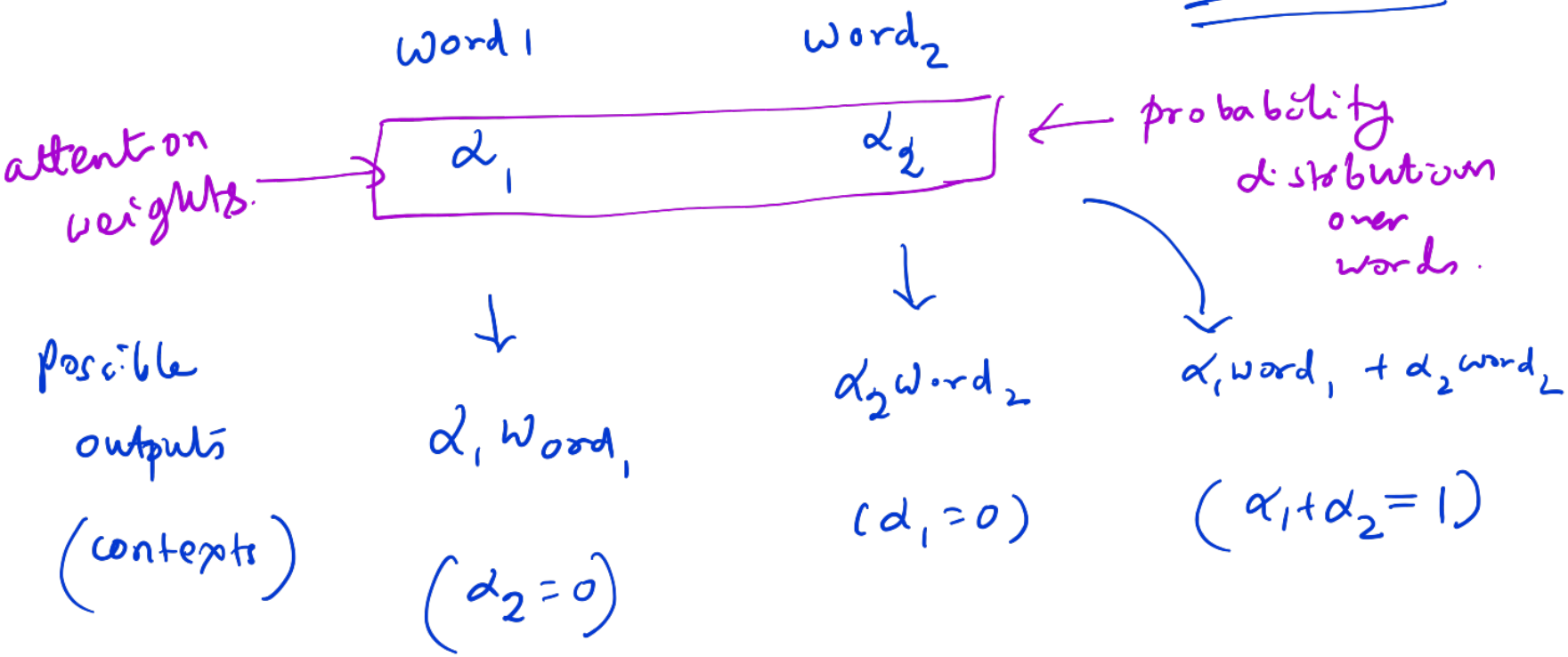
(2) Read → summarise → the input (part-by-part)

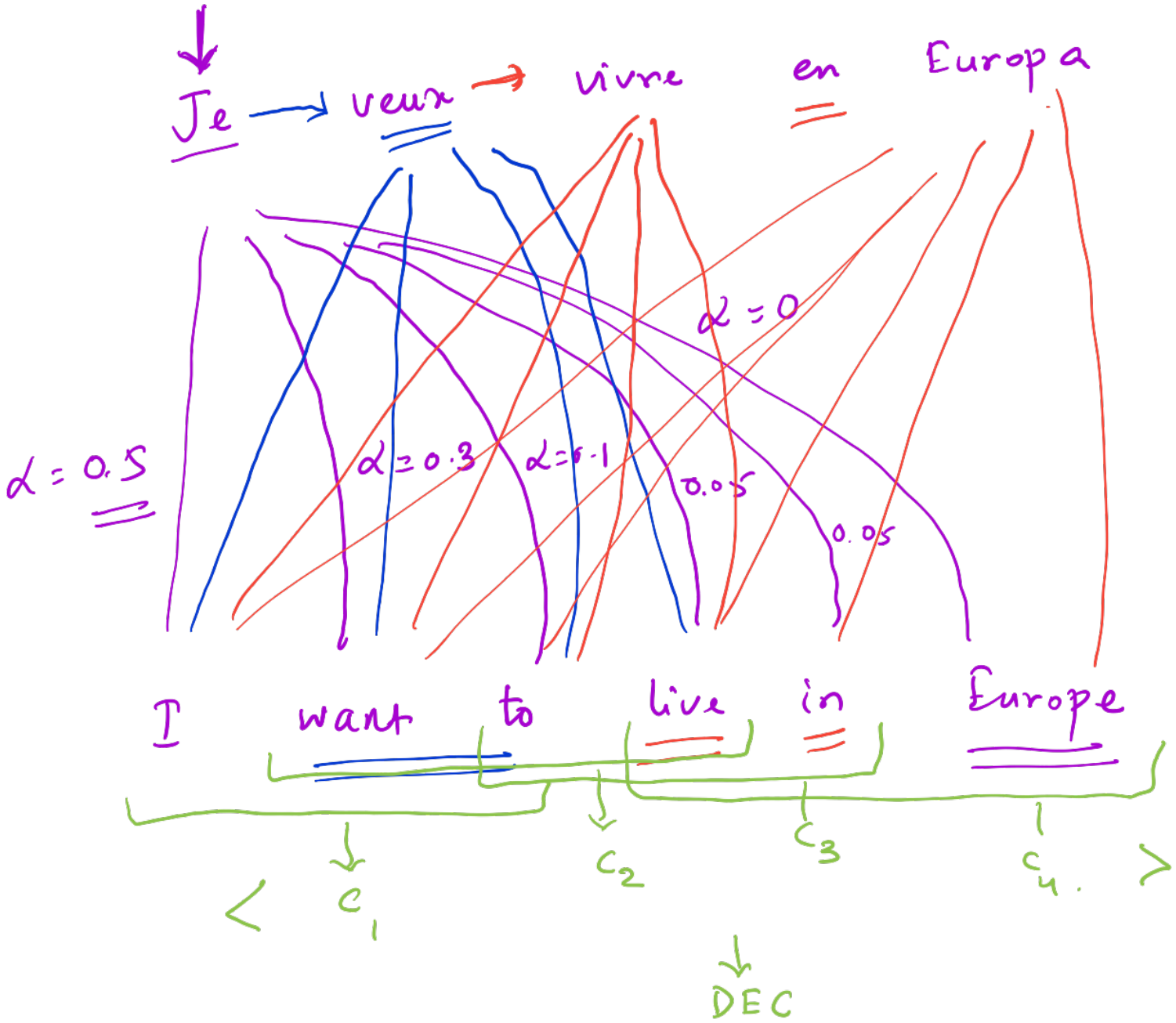


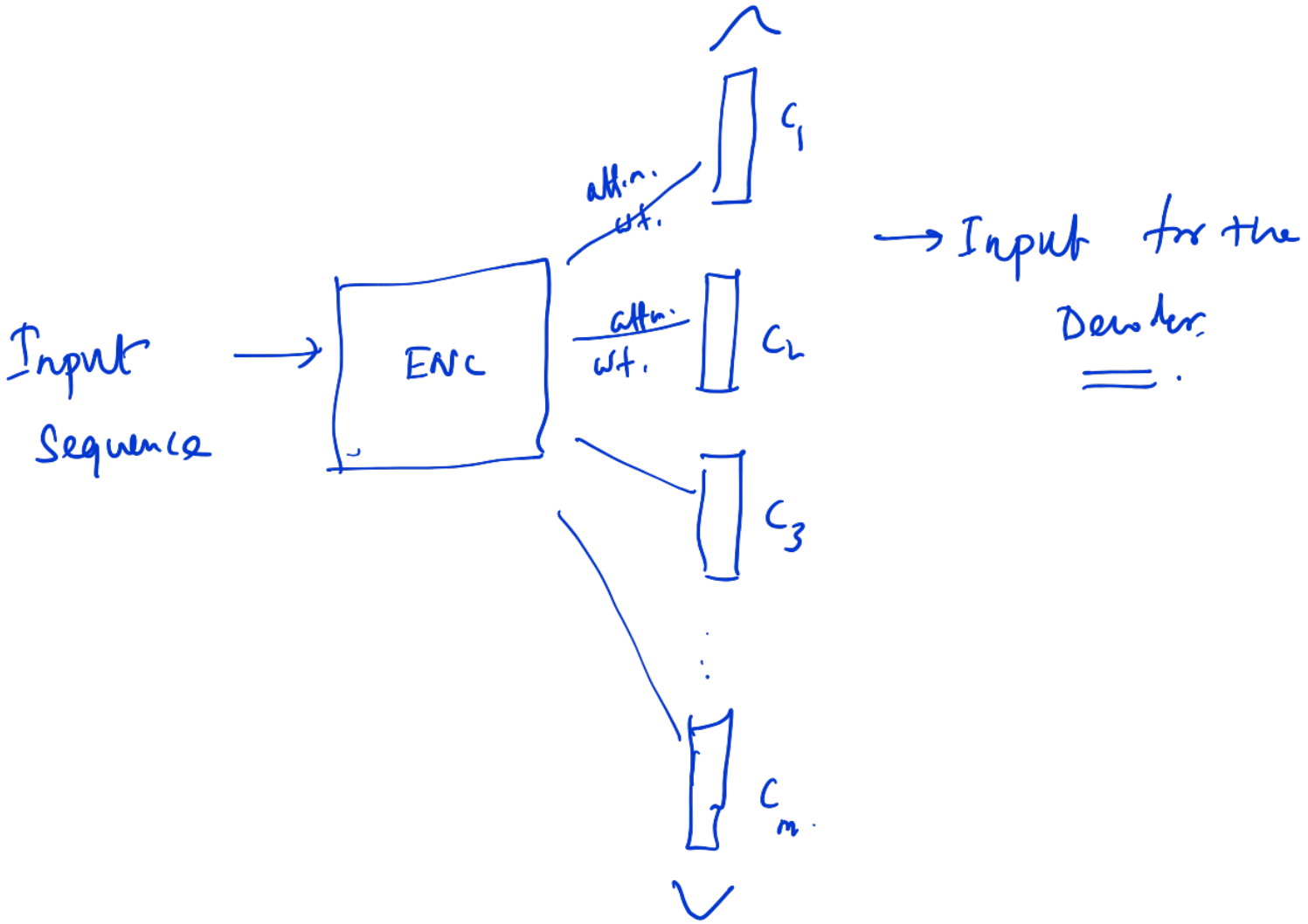
attention score : core of attention mechanism.

consider 2 words

$\alpha_1 + \alpha_2 = 1$




Je
veux
vivre
en
Europa
α = 0.5
α = 0.3
α = 0.1
α = 0
0.05
0.05
I
want
to
live
in
Europe
c1
c2
c3
c4
DEC


Input
Sequence
ENC
attn. wt.
attn. wt.
c_1
c_2
c_3
c_m
→ Input for the Decoder

Doubt

Bidirectional RNN