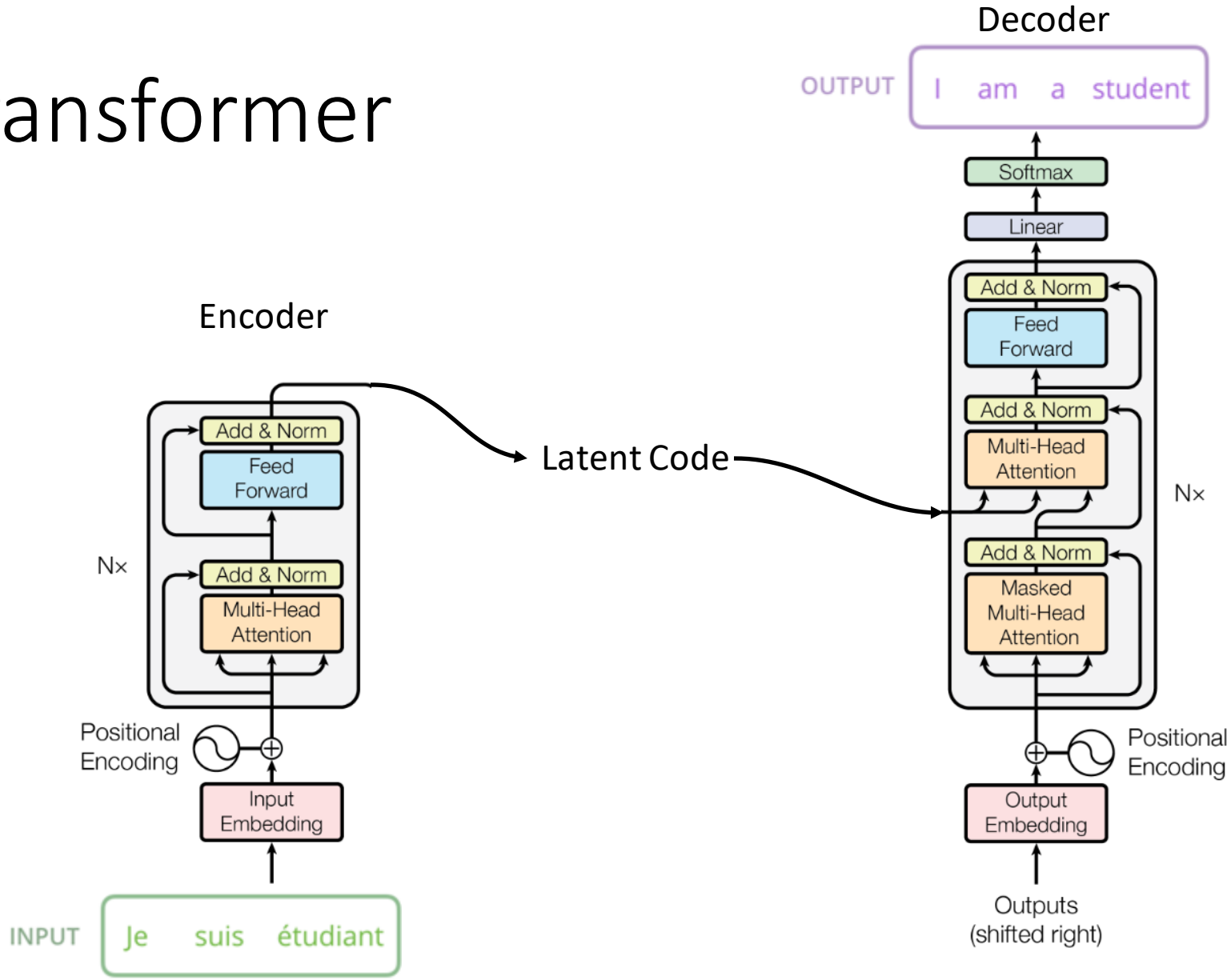
Neural Computation

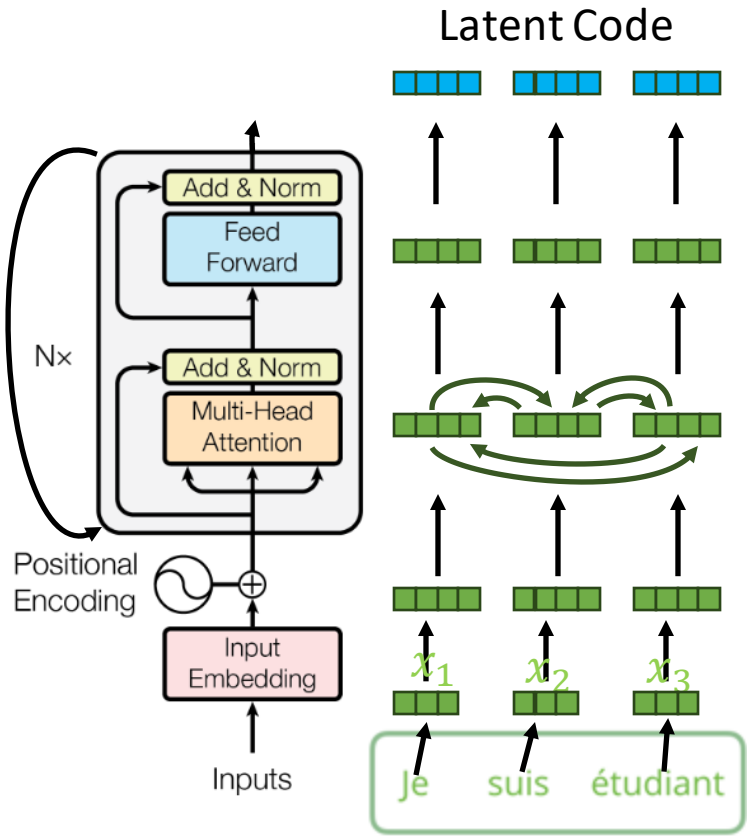
The Decoder - Part 2

Putting Everything Together

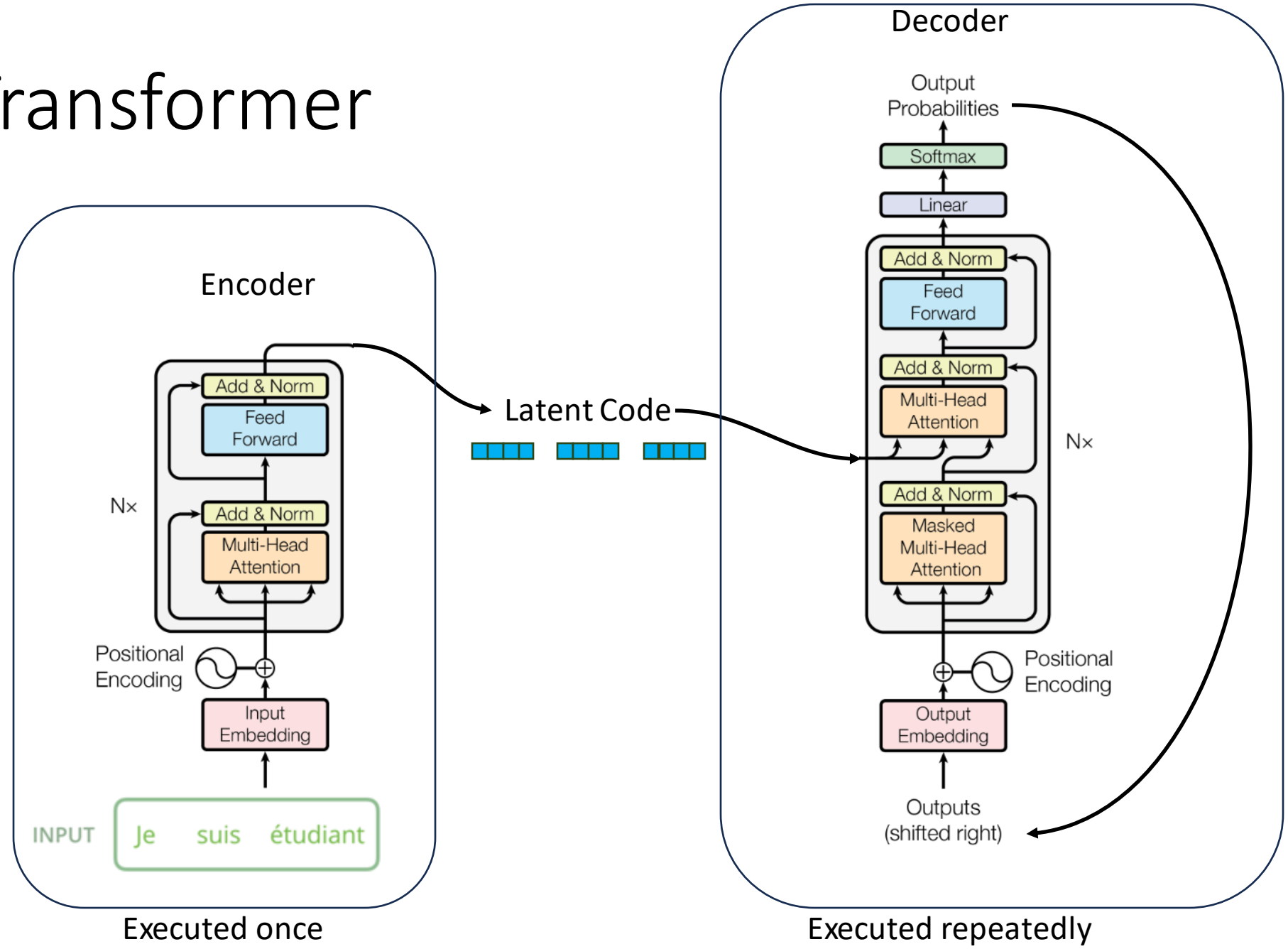
The Transformer



The Encoder


Latent Code
Add & Norm
Feed Forward
Add & Norm
Multi-Head Attention
Nx
Positional Encoding
Input Embedding
Inputs
x_1
x_2
x_3
Je
suis
étudiant

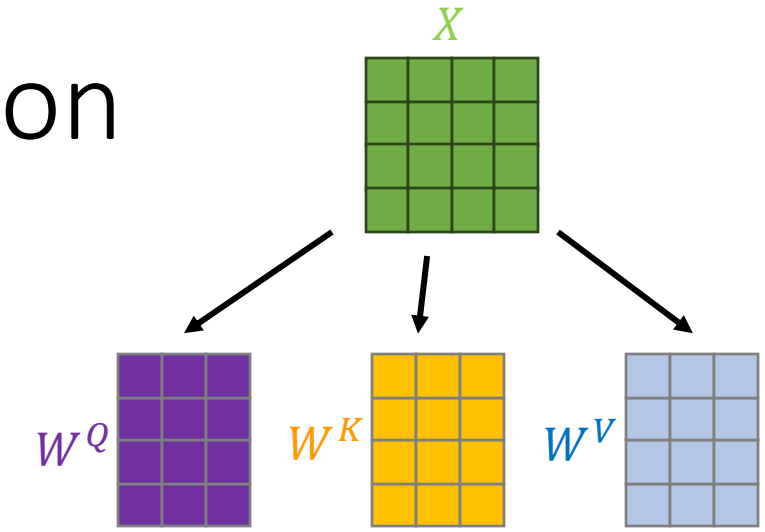
The Transformer



Executed once

Executed repeatedly

Self-Attention



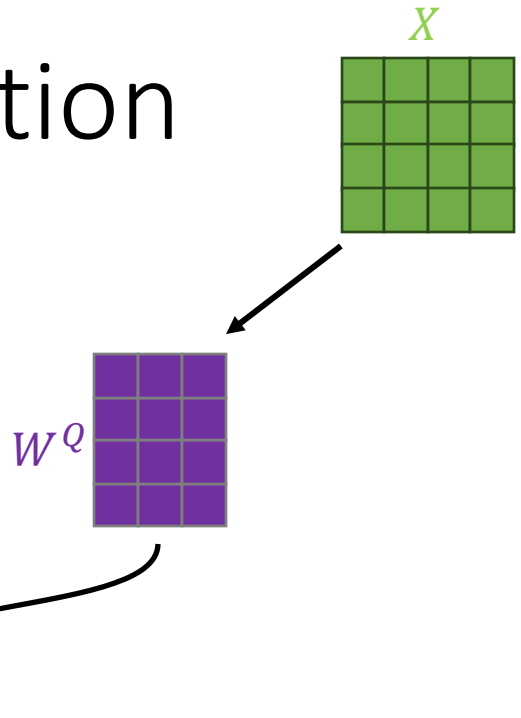
Idea:

- Every token makes:
  - Key
  - Value
  - Query

Query
$[0.3, 1.2, 2.2]^T$
$[0.1, 6.2, 2.1]^T$
$[2.1, 2.5, 5.2]^T$
$[1.1, 6.6, 4.3]^T$

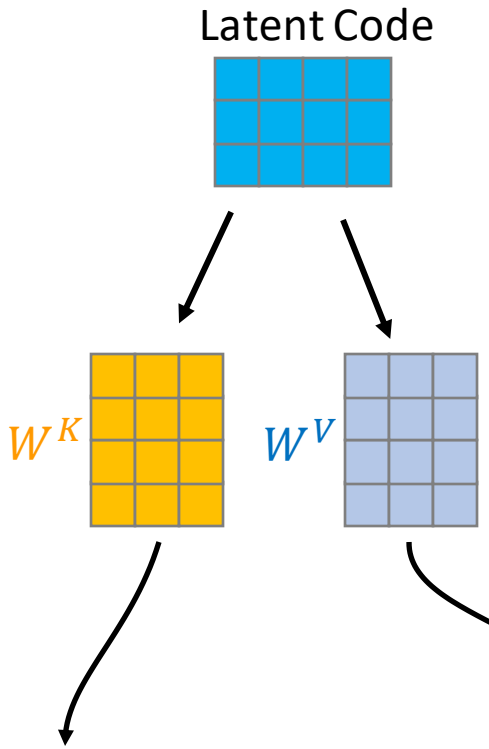
Keys	Values
$[0.4, 1.4, 0]^T$	$[2.3, 1.1, 3.5]^T$
$[0.2, -1.1, 0]^T$	$[3.0, 1.0, 0.5]^T$
$[0.3, 1.2, 4]^T$	$[-2.2, 4.1, 2.5]^T$
$[0.3, 1.2, 4]^T$	$[-2.2, 4.1, 2.5]^T$

Cross-Attention



Query
$[0.3, 1.2, 2.2]^T$
$[0.1, 6.2, 2.1]^T$
$[2.1, 2.5, 5.2]^T$
$[1.1, 6.6, 4.3]^T$

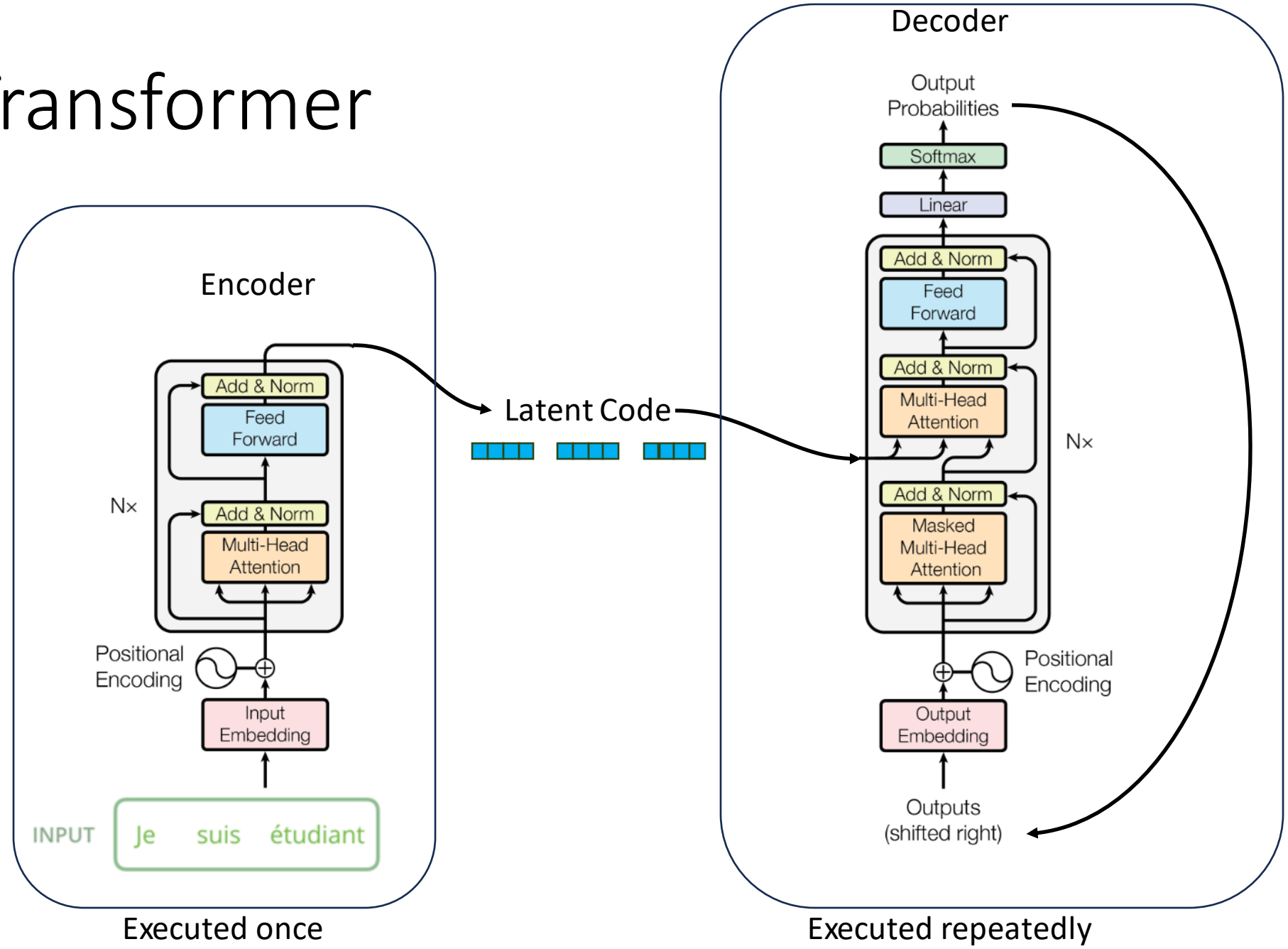
From Decoder



Keys	Values
$[0.4, 1.4, 0]^T$	$[2.3, 1.1, 3.5]^T$
$[0.2, -1.1, 0]^T$	$[3.0, 1.0, 0.5]^T$
$[0.3, 1.2, 4]^T$	$[-2.2, 4.1, 2.5]^T$
$[0.3, 1.2, 4]^T$	$[-2.2, 4.1, 2.5]^T$

From Encoder

The Transformer



Executed once

Executed repeatedly

Which word in our vocabulary is associated with this index?

am

Get the index of the cell with the highest value (**argmax**)

5

log_probs

Softmax

logits

Linear

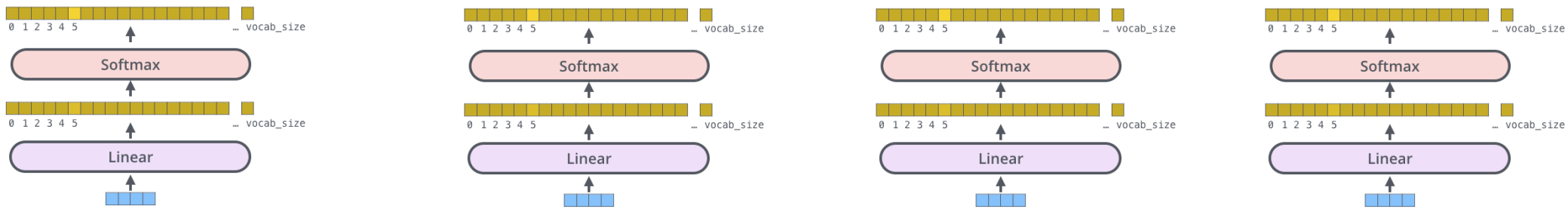
Decoder stack output

Which word in our vocabulary is associated with this index?

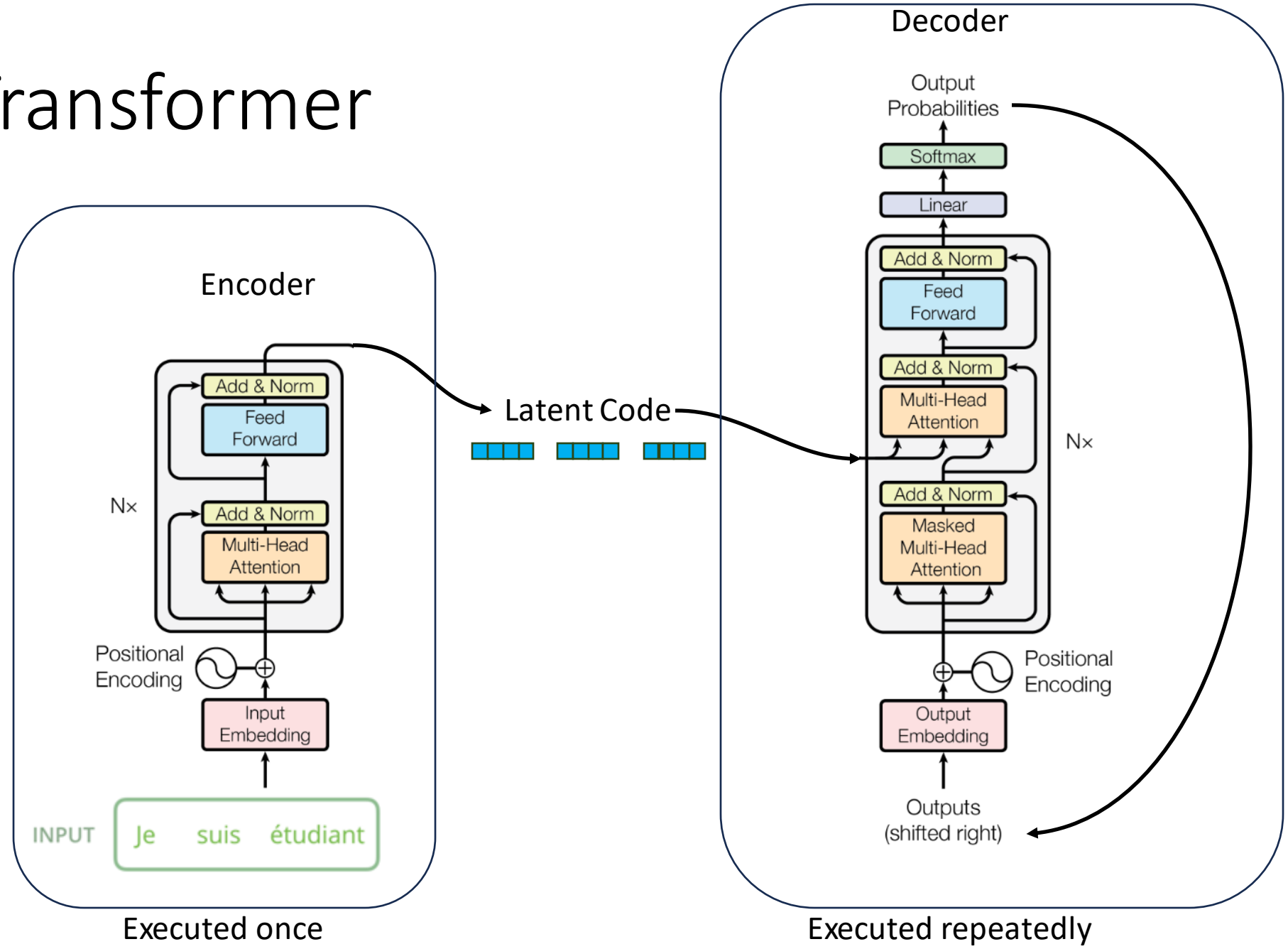
am

Get the index of the cell with the highest value **(argmax)**

5

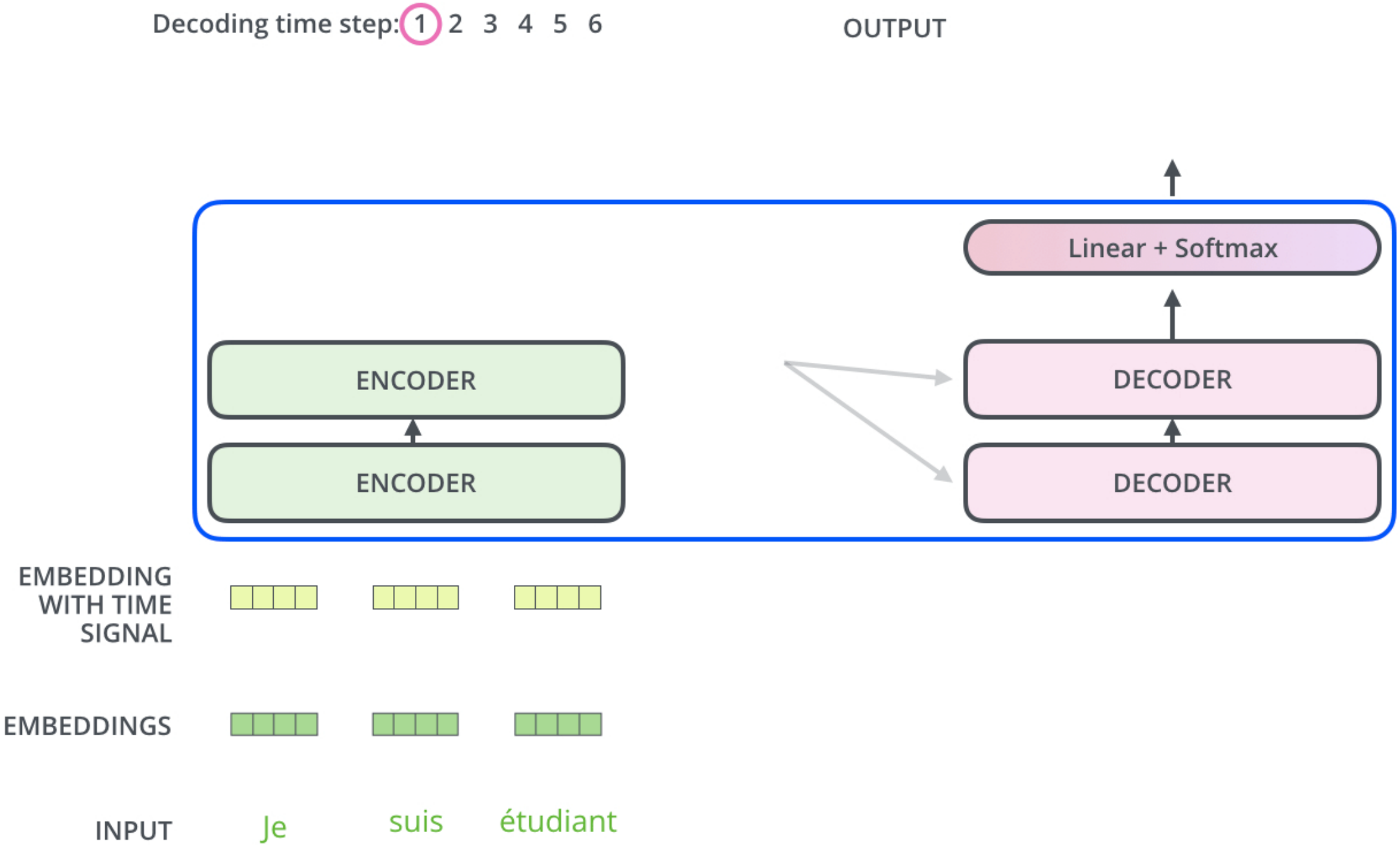


The Transformer

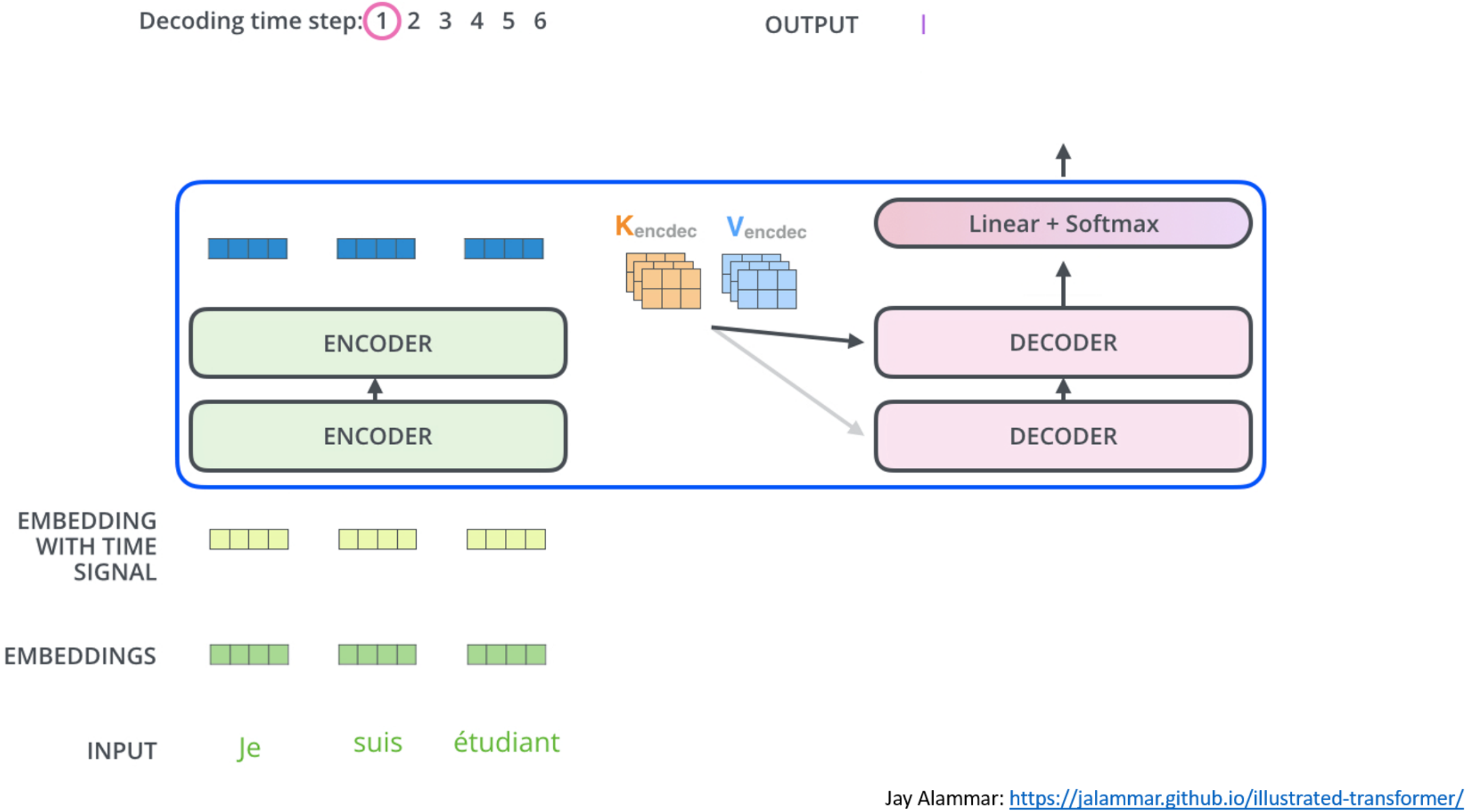


Executed once

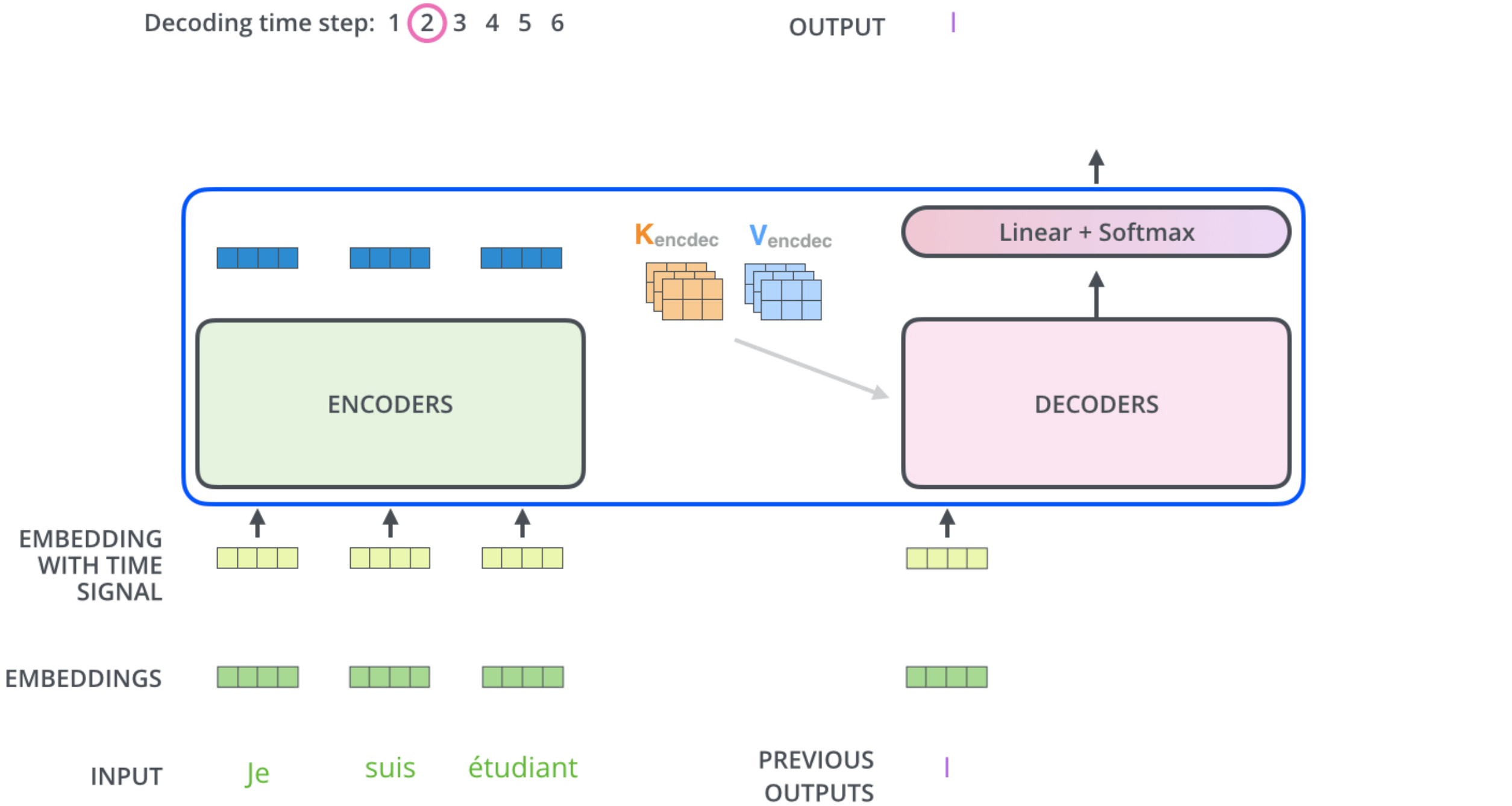
Executed repeatedly



Jay Alammar: https://jalammar.github.io/illustrated-transformer/



Jay Alammar: https://jalammar.github.io/illustrated-transformer/



Jay Alammar: https://jalammar.github.io/illustrated-transformer/