Attention Is All You Need

Ashish Vaswani[*]
Google Brain
avaswani@google.com

Noam Shazeer[*]
Google Brain
noam@google.com

Niki Parmar[*]
Google Research
nikip@google.com

Jakob Uszkoreit[*]
Google Research
usz@google.com

Llion Jones[*]
Google Research
llion@google.com

Aidan N. Gomez[*] [†]
University of Toronto
aidan@cs.toronto.edu

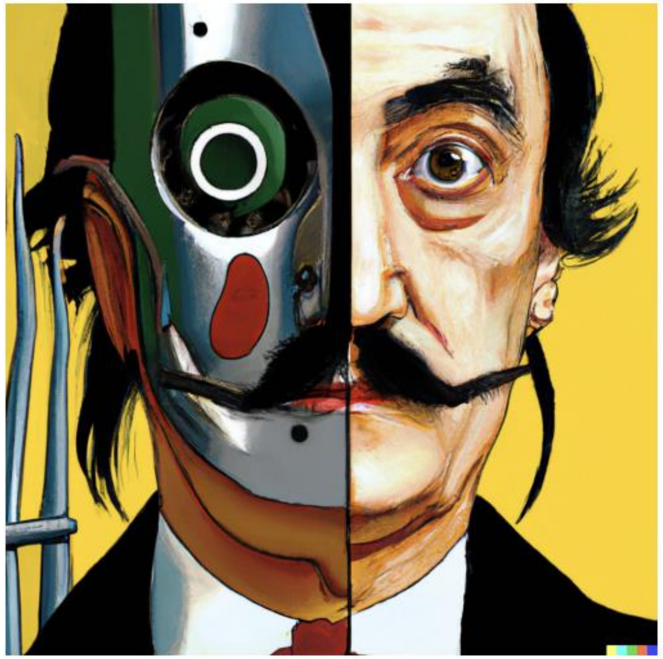
Łukasz Kaiser[*]
Google Brain
lukaszkaiser@google.com

Illia Polosukhin[*] [‡]
illia.polosukhin@gmail.com

Abstract

The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that include an encoder and a decoder. The best performing models also connect the encoder and decoder through an attention

m



vibrant portrait painting of Salvador Dalí with a robotic half face

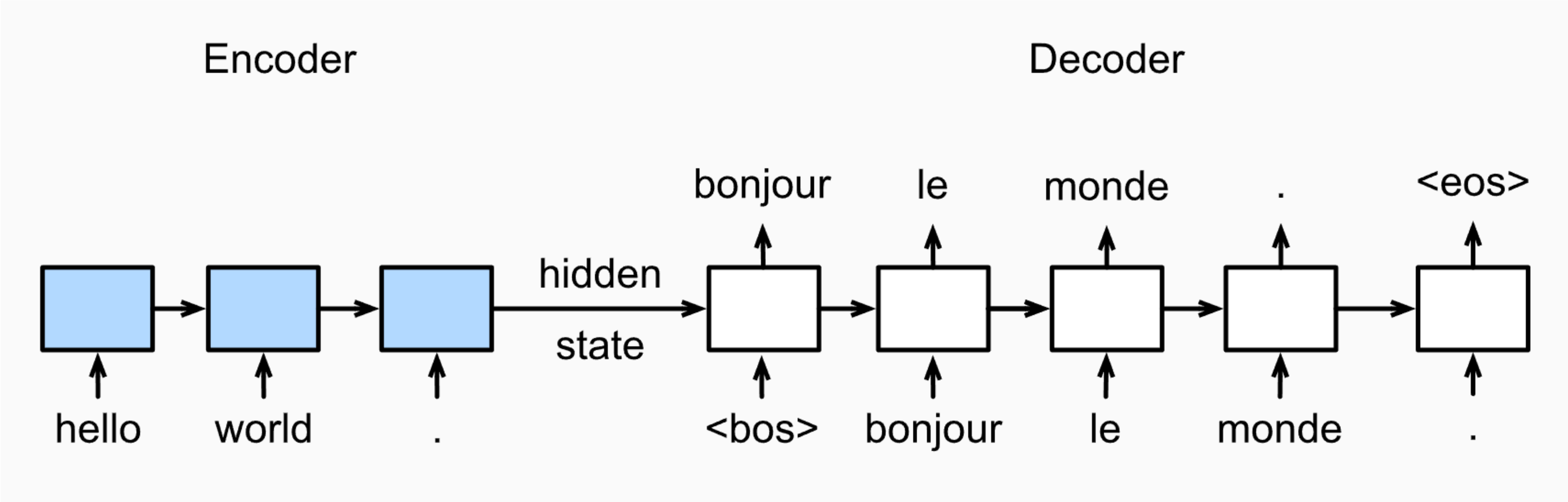
a shiba inu wearing a beret and black turtleneck

a close up of a handpalm with leaves growing from it

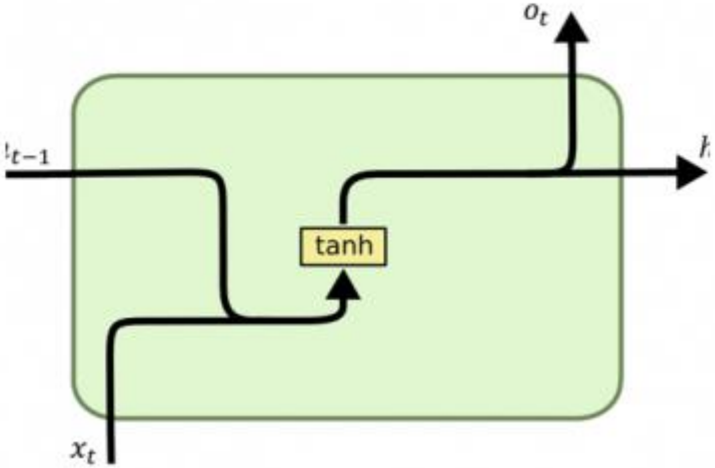
an espresso machine that makes coffee from human souls, artstation

panda mad scientist mixing sparkling chemicals, artstation

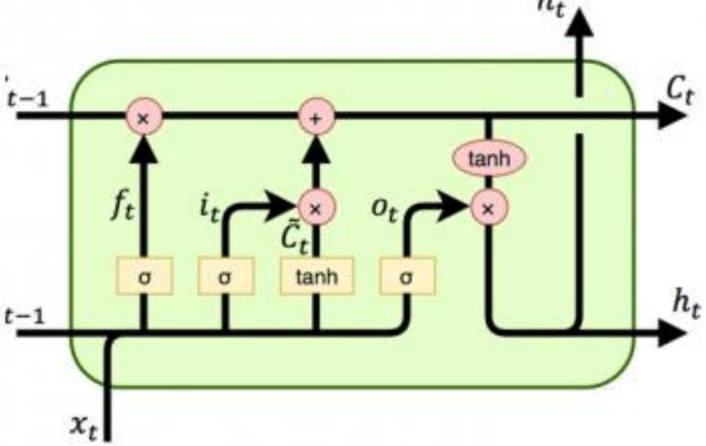
a corgi’s head depicted as an explosion of a nebula


Encoder
Decoder
bonjour
le
monde
.
<eos>
hidden
state
hello
world
.
<bos>
bonjour
le
monde
.

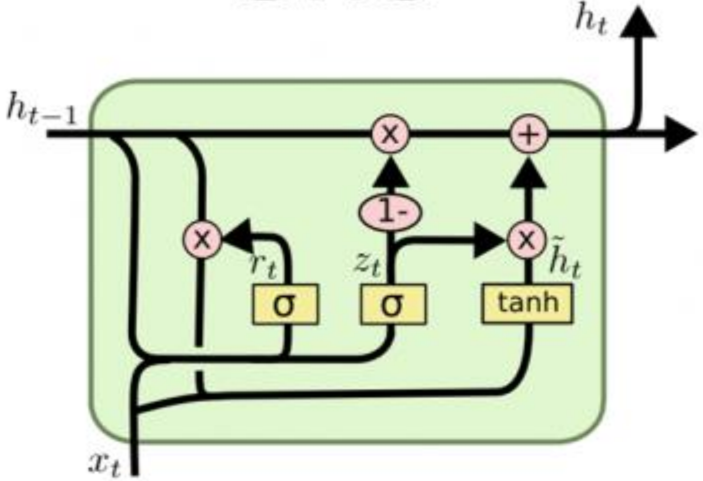
RNN

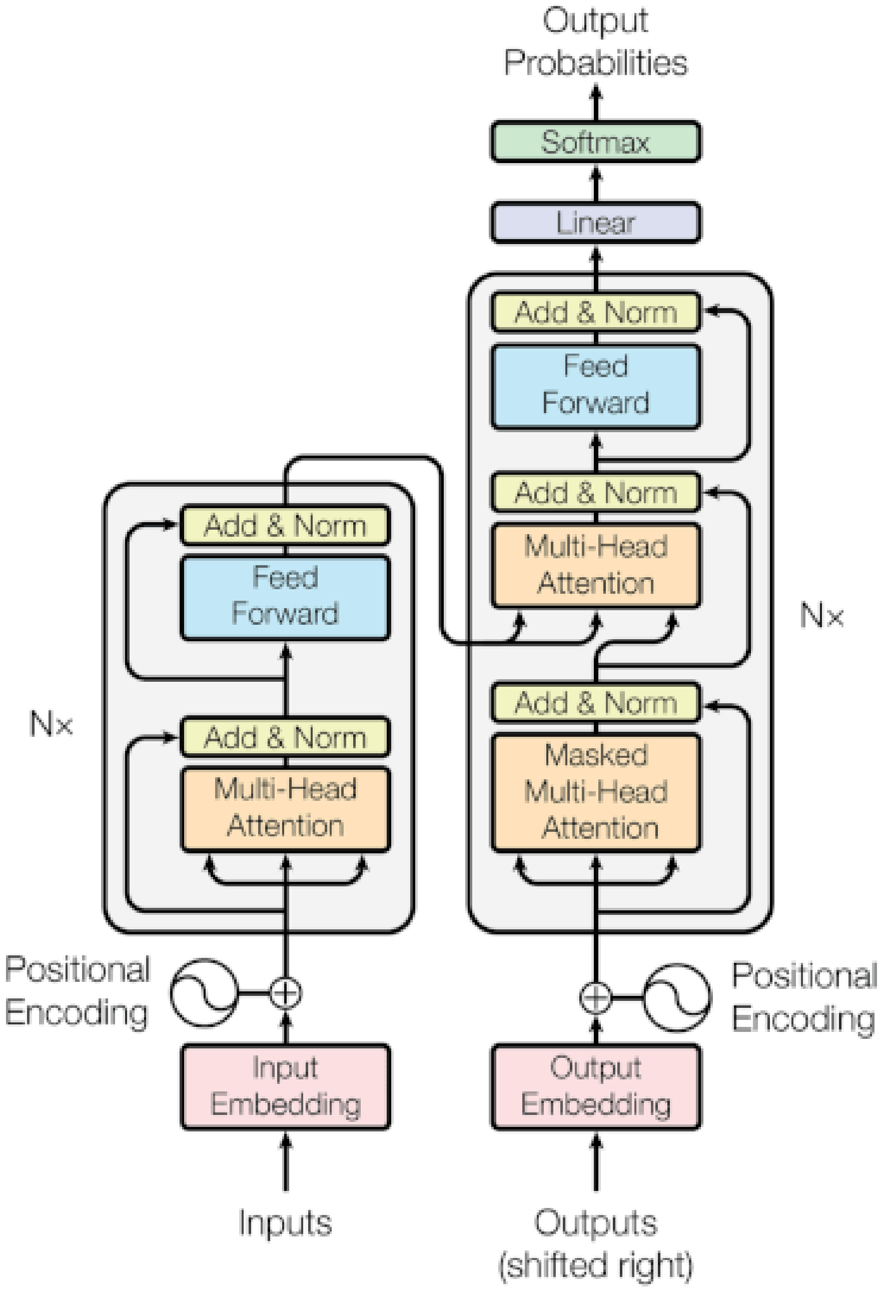

o_t
$_{t-1}$
h
tanh
x_t

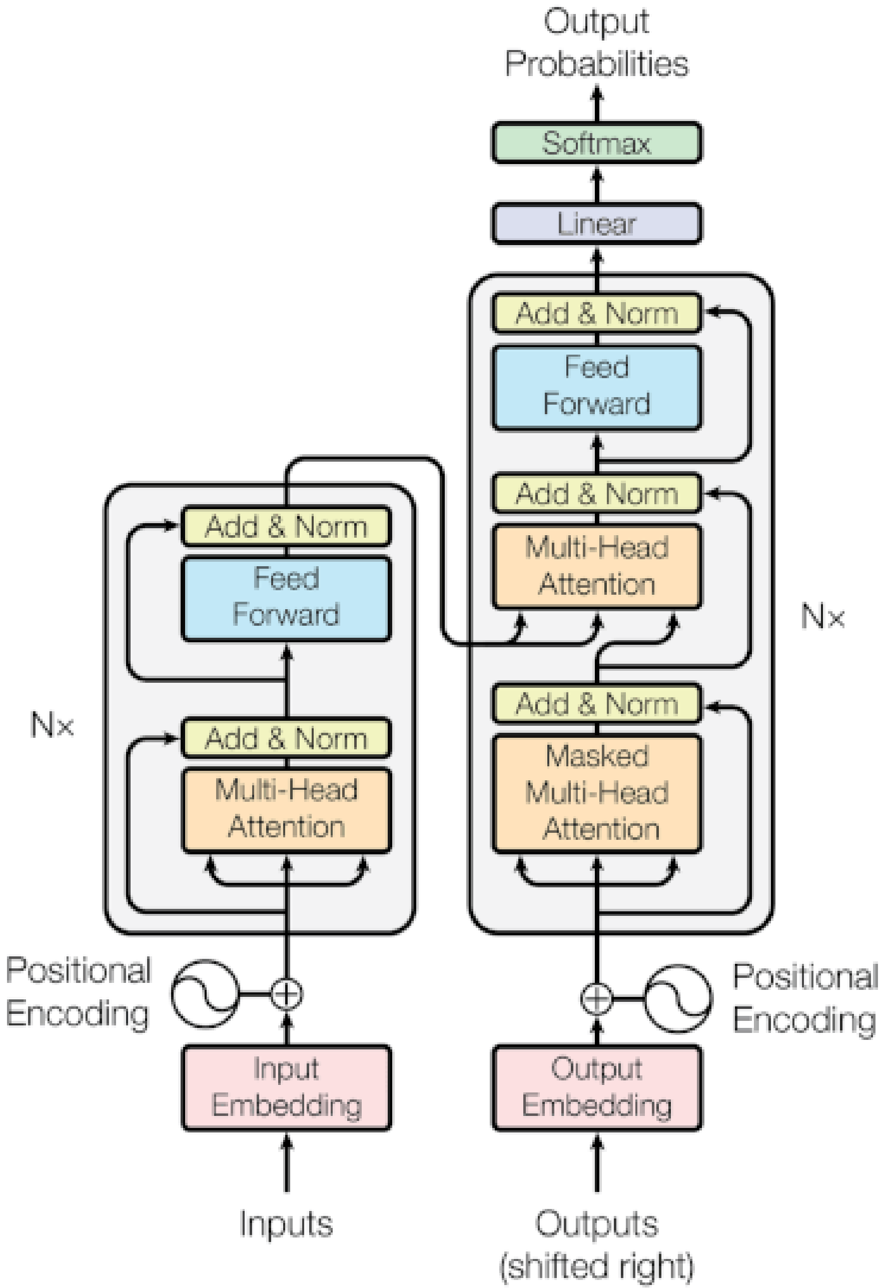
LSTM

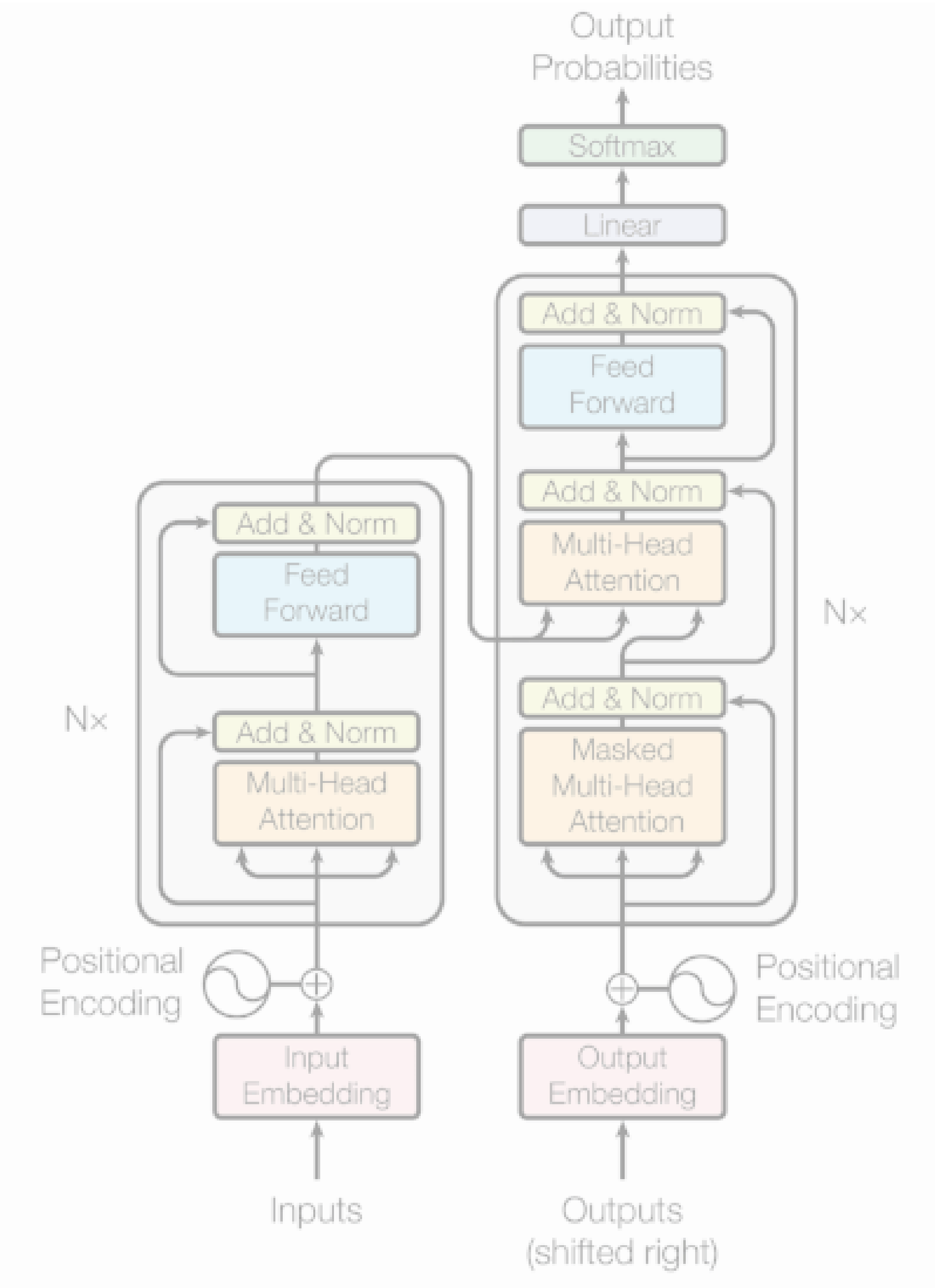

h_t
$t-1$
C_t
×
+
tanh
f_t
i_t
×
$\tilde{C}_t$
o_t
×
σ
σ
tanh
σ
$t-1$
h_t
x_t

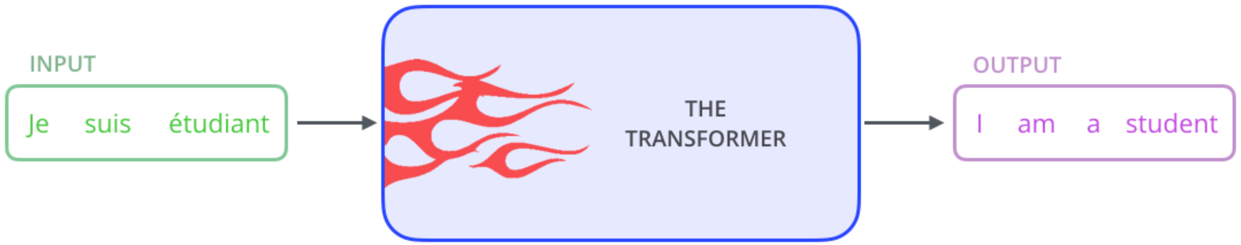
GRU

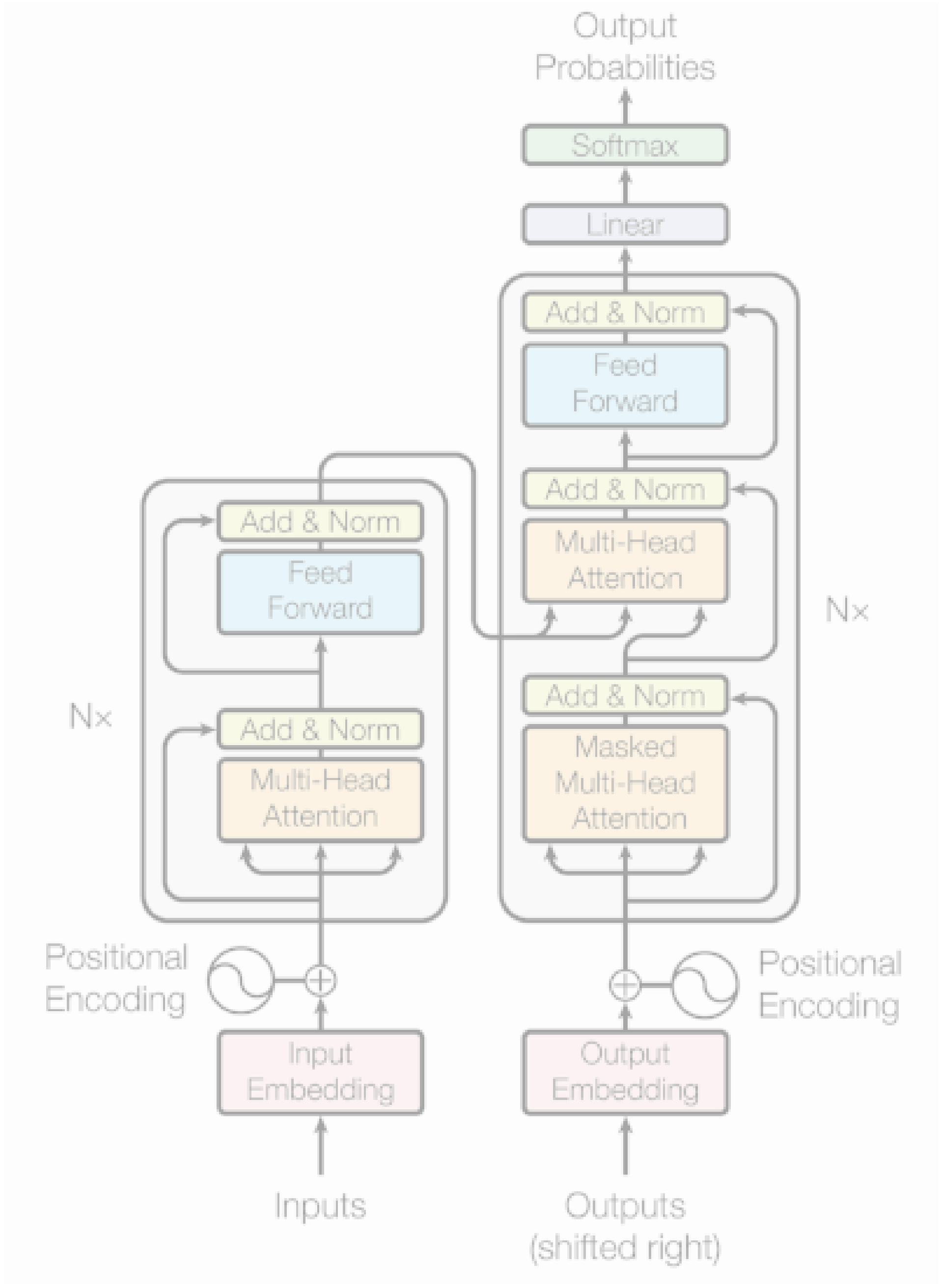

h_t
h_{t-1}
×
+
1-
×
r_t
z_t
×
$\tilde{h}_t$
σ
σ
tanh
x_t

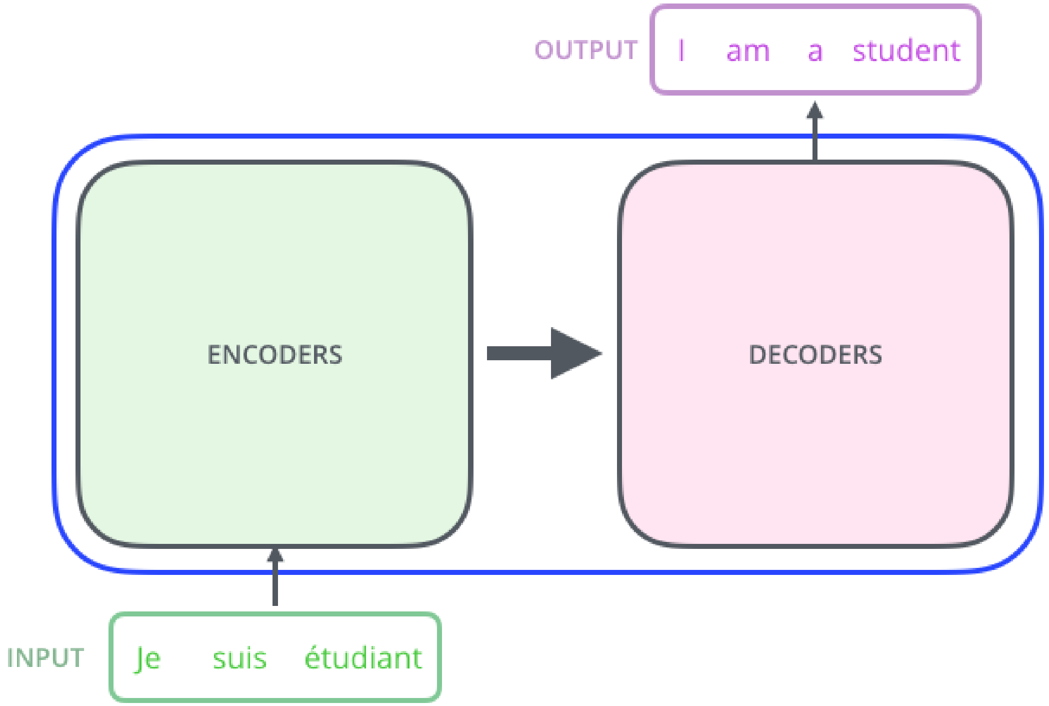

Output Probabilities
Softmax
Linear
Add & Norm
Feed Forward
Add & Norm
Multi-Head Attention
Nx
Add & Norm
Masked Multi-Head Attention
Add & Norm
Feed Forward
Nx
Add & Norm
Multi-Head Attention
Positional Encoding
Positional Encoding
Input Embedding
Output Embedding
Inputs
Outputs (shifted right)

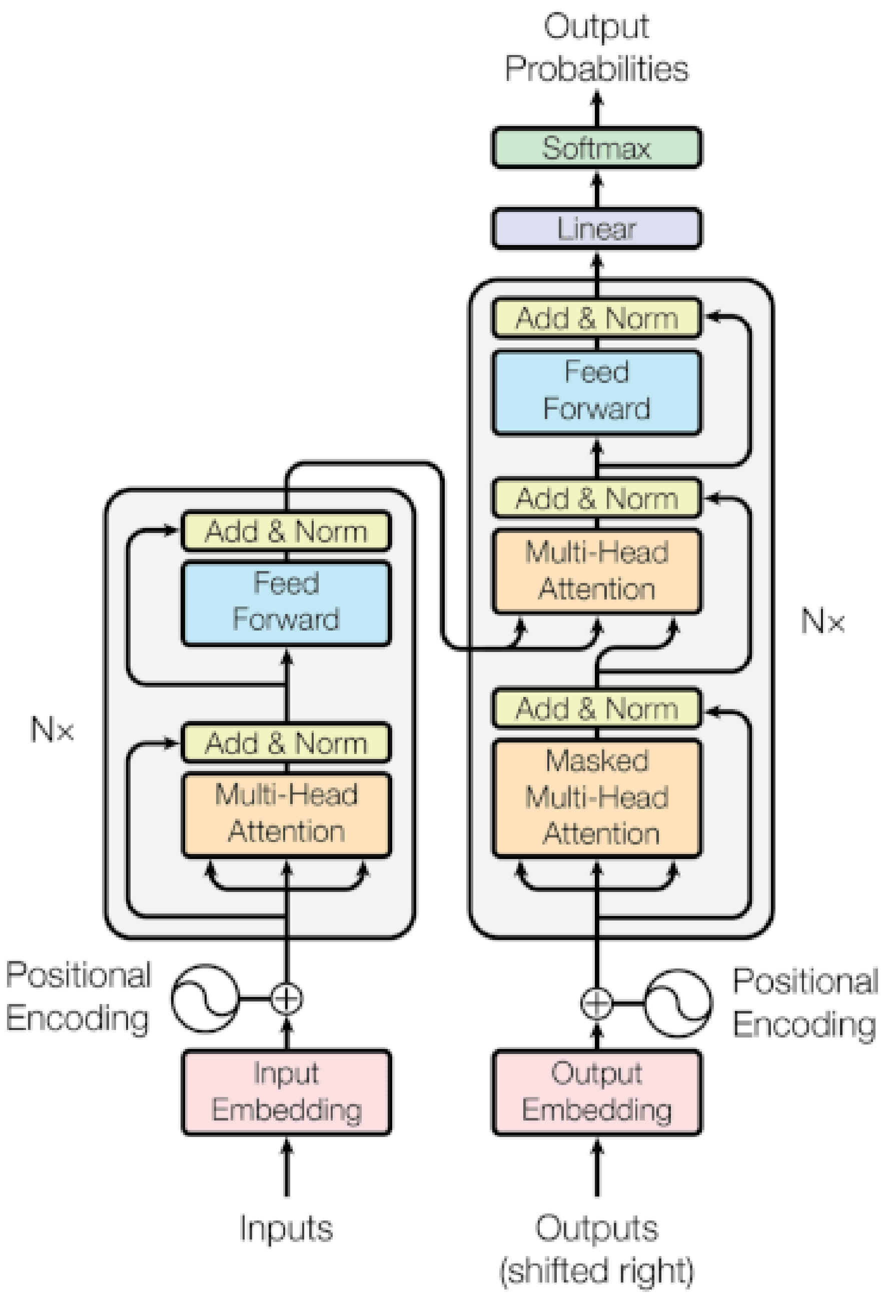

Output Probabilities
Softmax
Linear
Add & Norm
Feed Forward
Add & Norm
Multi-Head Attention
Add & Norm
Masked Multi-Head Attention
Nx
Add & Norm
Feed Forward
Add & Norm
Multi-Head Attention
Nx
Positional Encoding
Positional Encoding
Input Embedding
Output Embedding
Inputs
Outputs (shifted right)

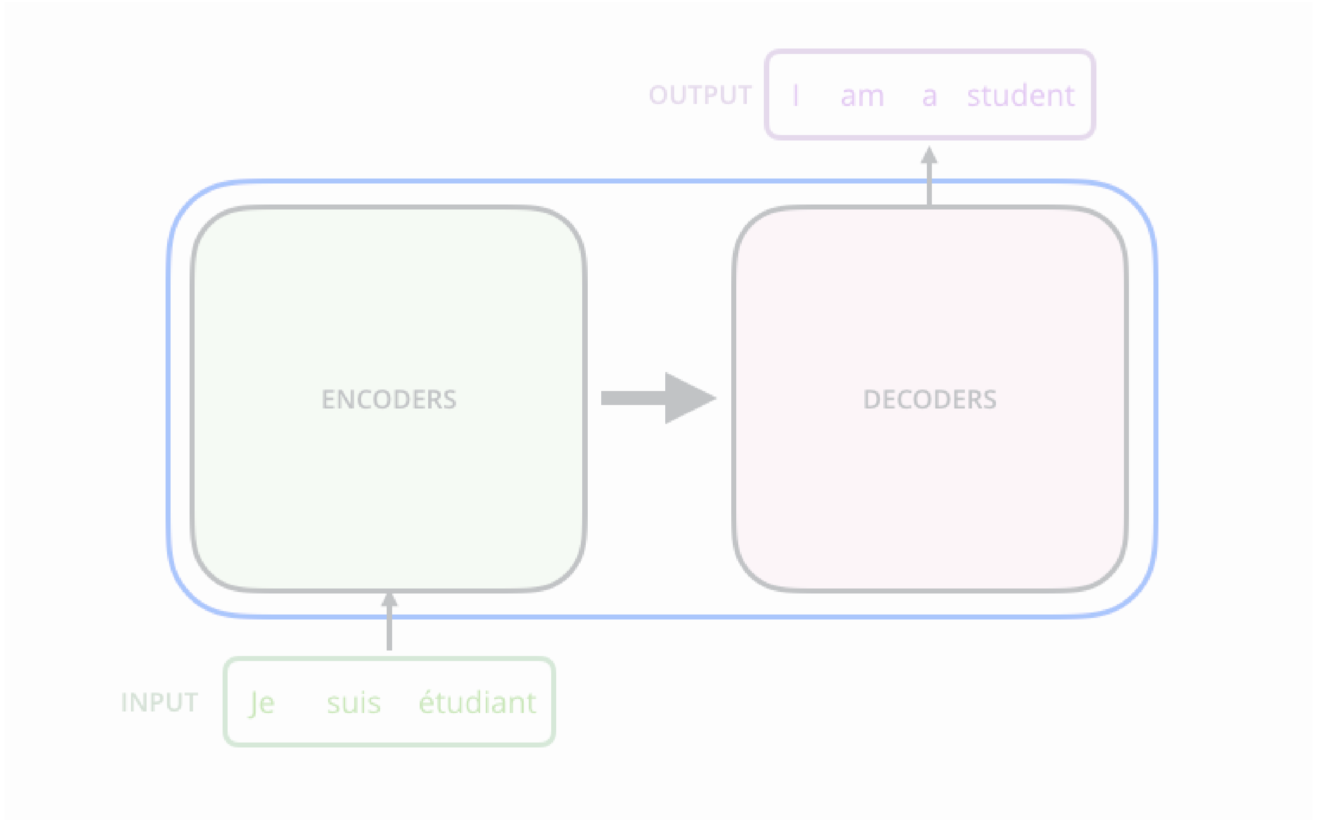

Output Probabilities
Softmax
Linear
Add & Norm
Feed Forward
Add & Norm
Multi-Head Attention
Add & Norm
Masked Multi-Head Attention
Nx
Add & Norm
Feed Forward
Add & Norm
Multi-Head Attention
Nx
Positional Encoding
Positional Encoding
Input Embedding
Output Embedding
Inputs
Outputs (shifted right)

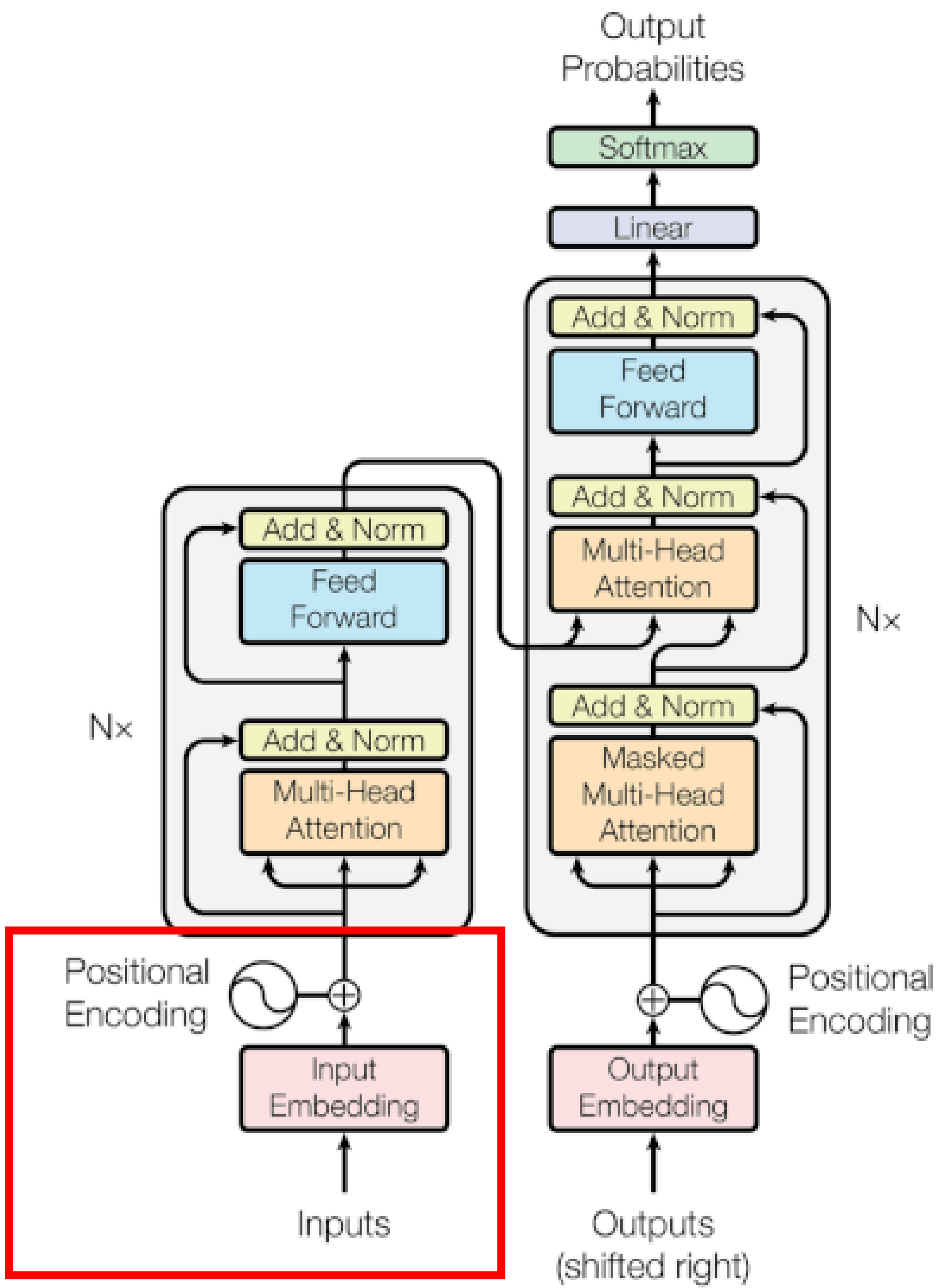

INPUT
Je suis étudiant
THE TRANSFORMER
OUTPUT
I am a student

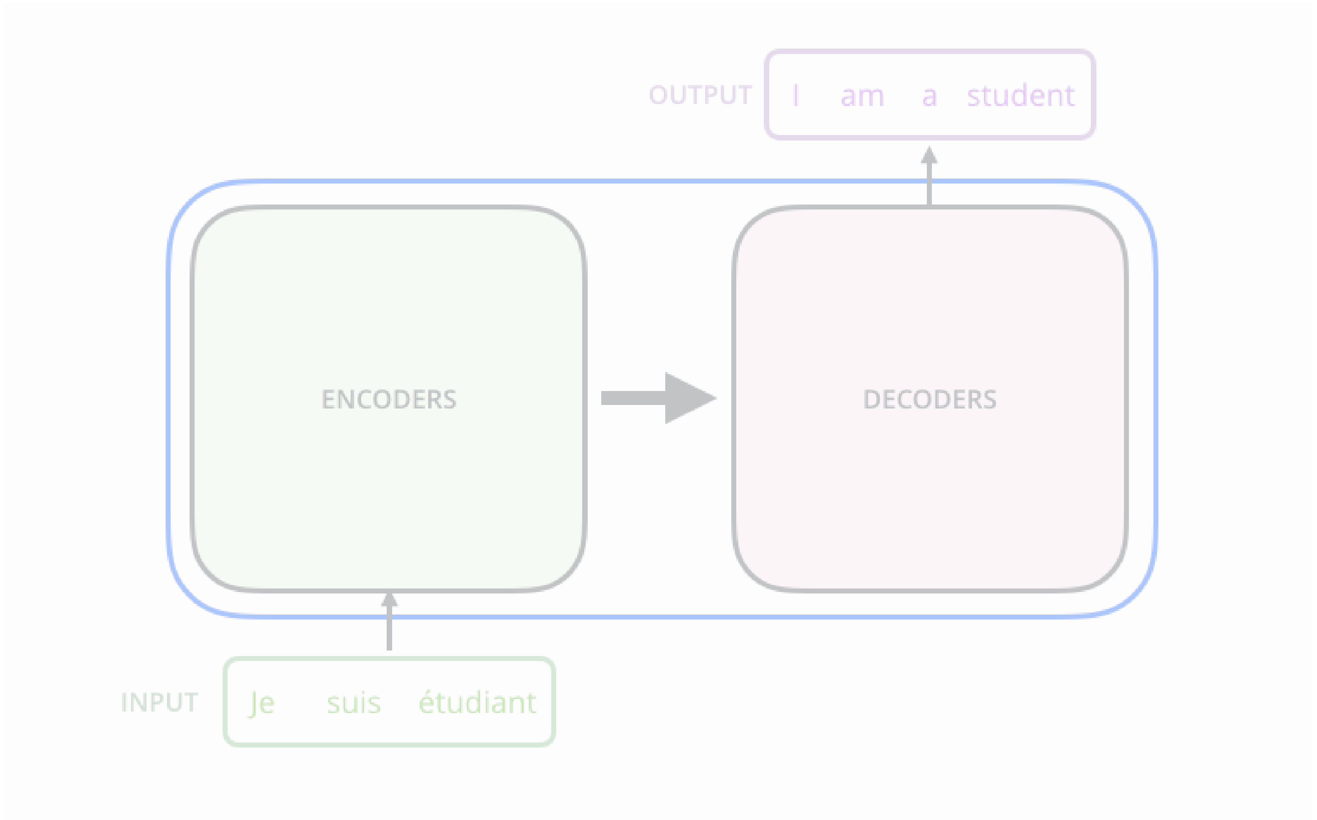

Output Probabilities
Softmax
Linear
Add & Norm
Feed Forward
Add & Norm
Multi-Head Attention
Add & Norm
Masked Multi-Head Attention
Nx
Add & Norm
Feed Forward
Add & Norm
Multi-Head Attention
Nx
Positional Encoding
Positional Encoding
Input Embedding
Output Embedding
Inputs
Outputs (shifted right)

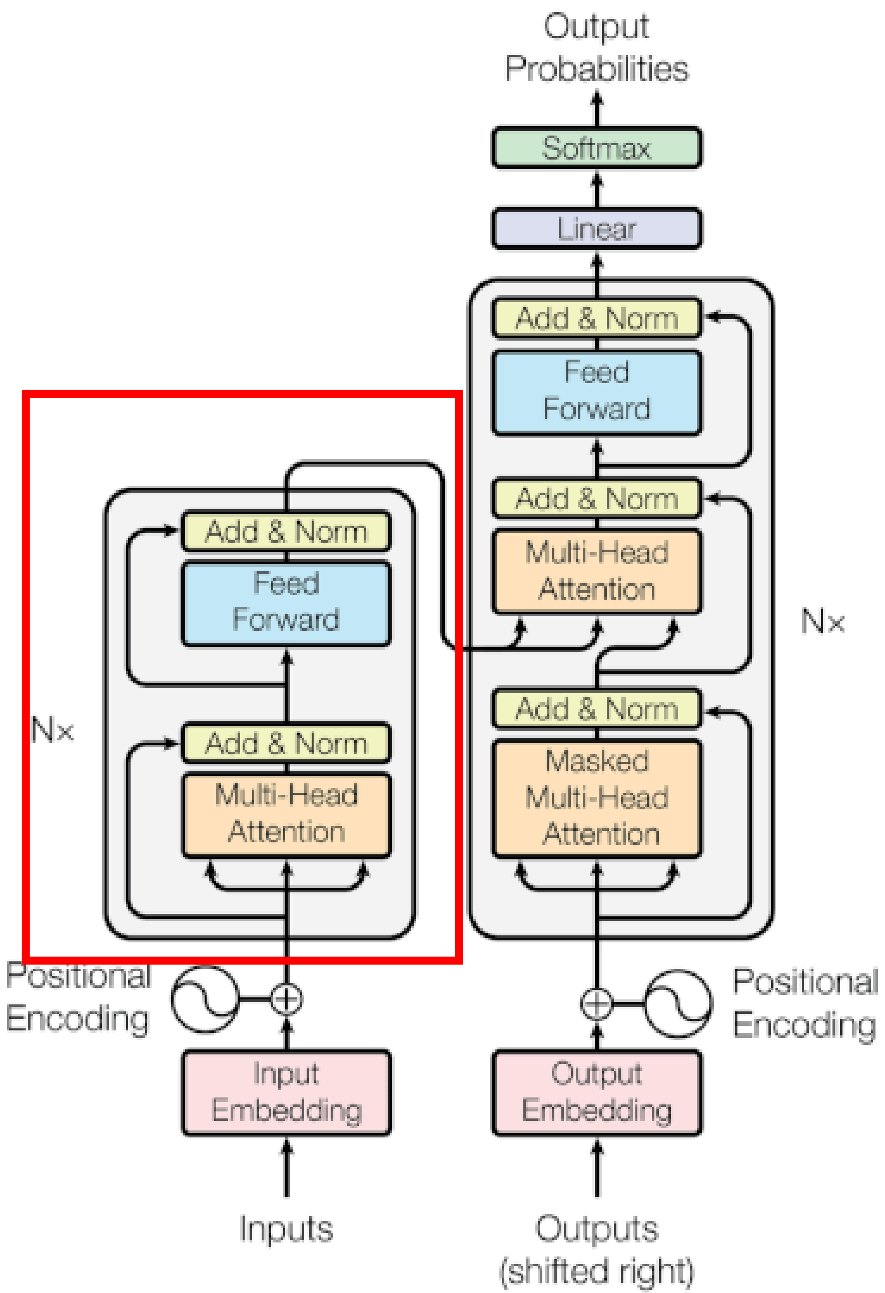

OUTPUT
I am a student
ENCODERS
DECODERS
INPUT
Je suis étudiant

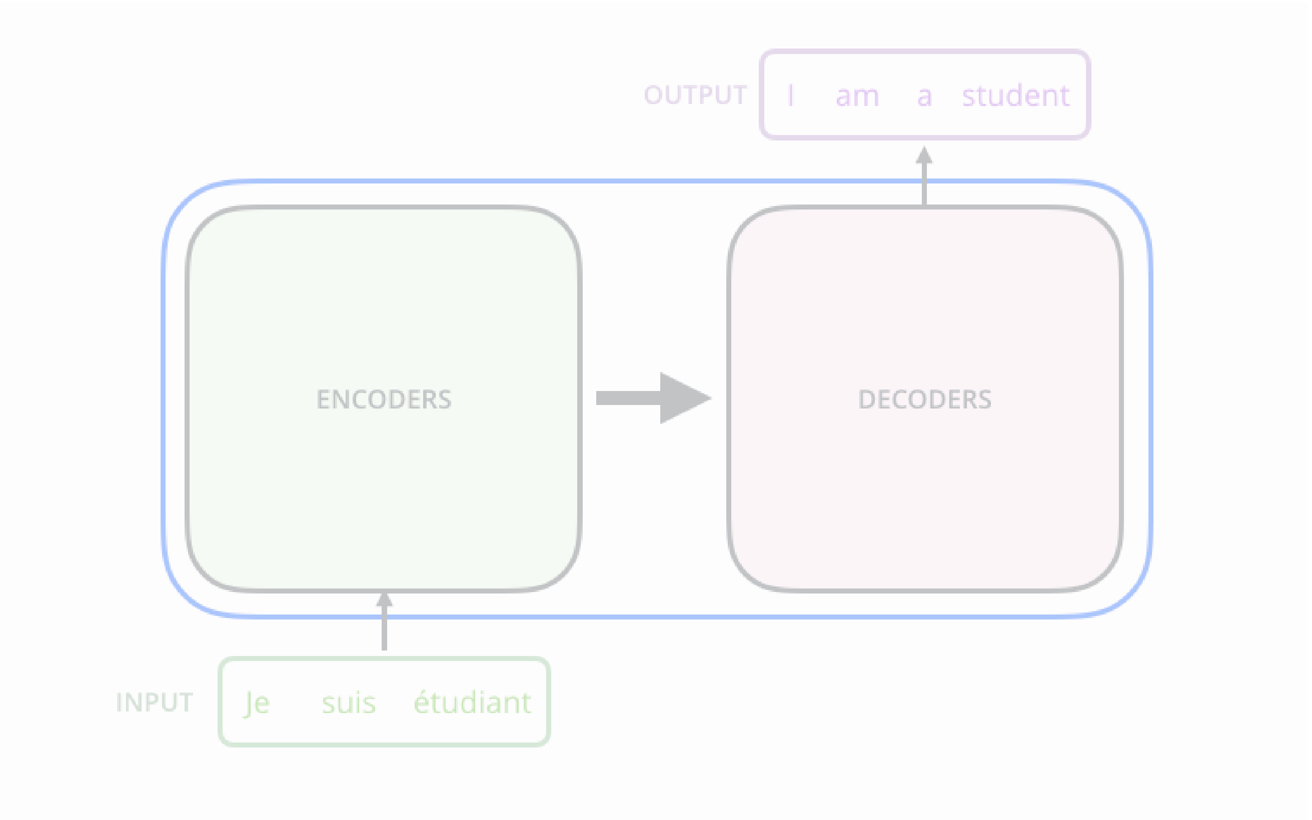

Output Probabilities
Softmax
Linear
Add & Norm
Feed Forward
Add & Norm
Multi-Head Attention
Add & Norm
Masked Multi-Head Attention
Nx
Add & Norm
Feed Forward
Add & Norm
Multi-Head Attention
Nx
Positional Encoding
Positional Encoding
Input Embedding
Output Embedding
Inputs
Outputs (shifted right)

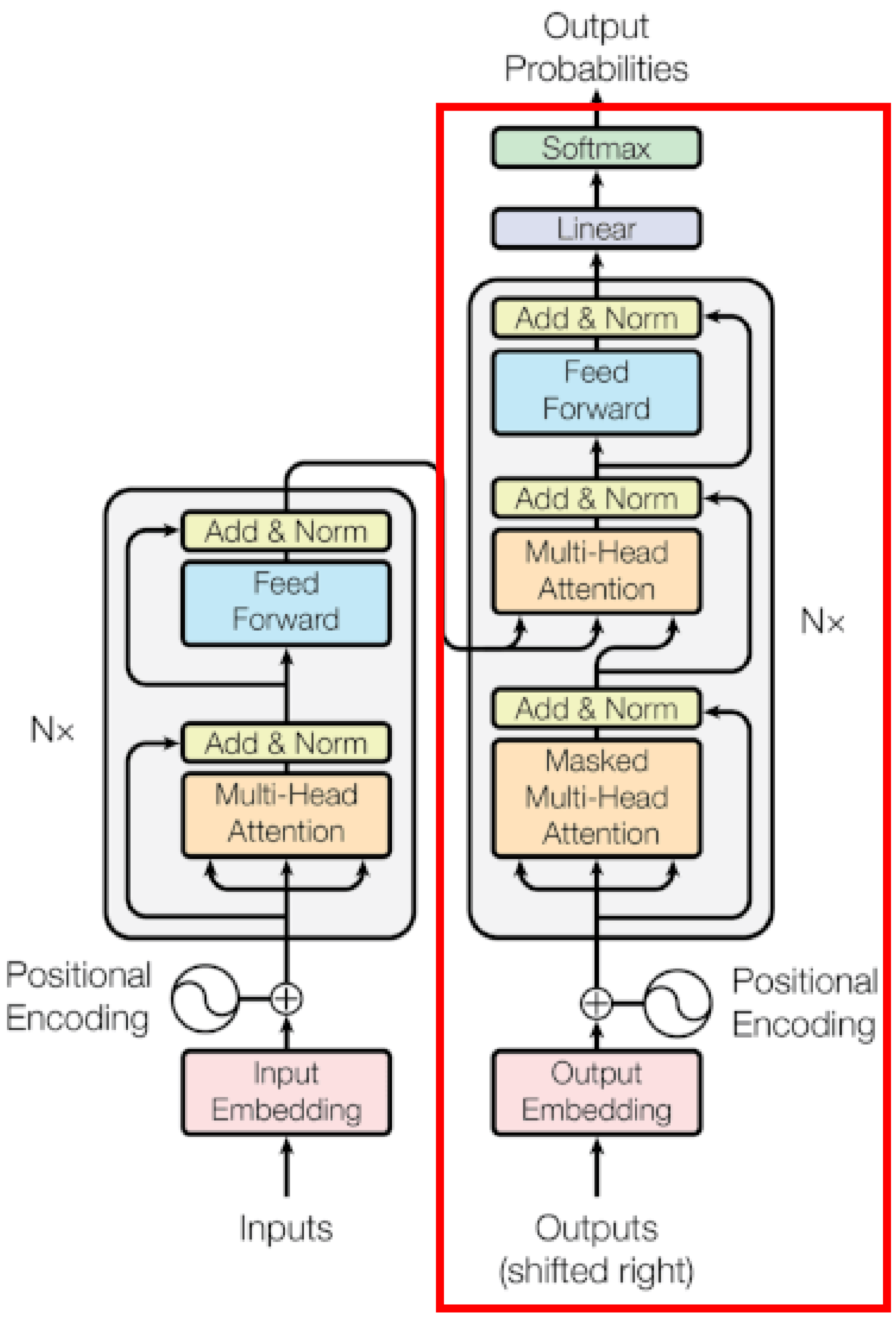

OUTPUT
I am a student
ENCODERS
DECODERS
INPUT
Je suis étudiant

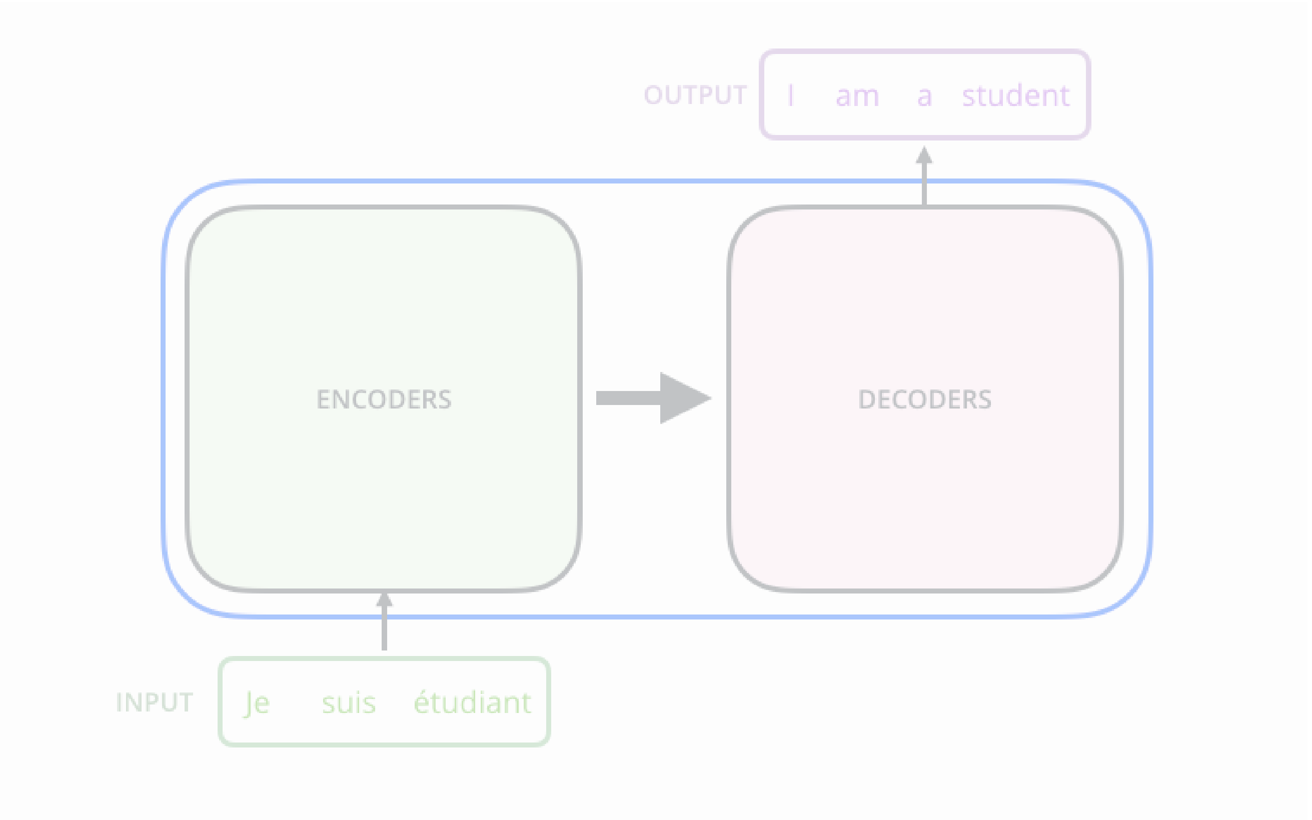

Output Probabilities
Softmax
Linear
Add & Norm
Feed Forward
Add & Norm
Multi-Head Attention
Add & Norm
Masked Multi-Head Attention
Nx
Add & Norm
Feed Forward
Add & Norm
Multi-Head Attention
Nx
Positional Encoding
Input Embedding
Inputs
Positional Encoding
Output Embedding
Outputs (shifted right)

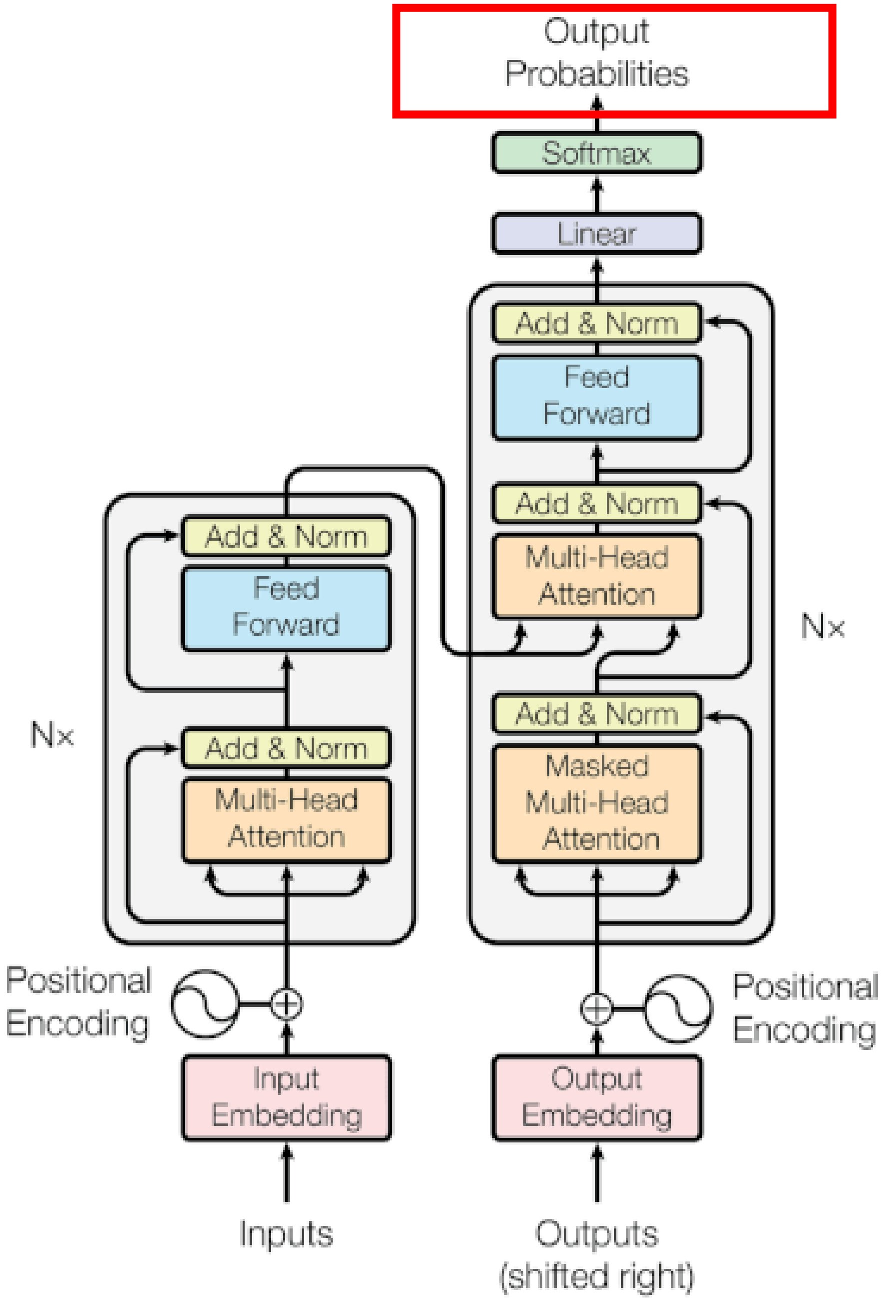

OUTPUT
I am a student
ENCODERS
DECODERS
INPUT
Je suis étudiant

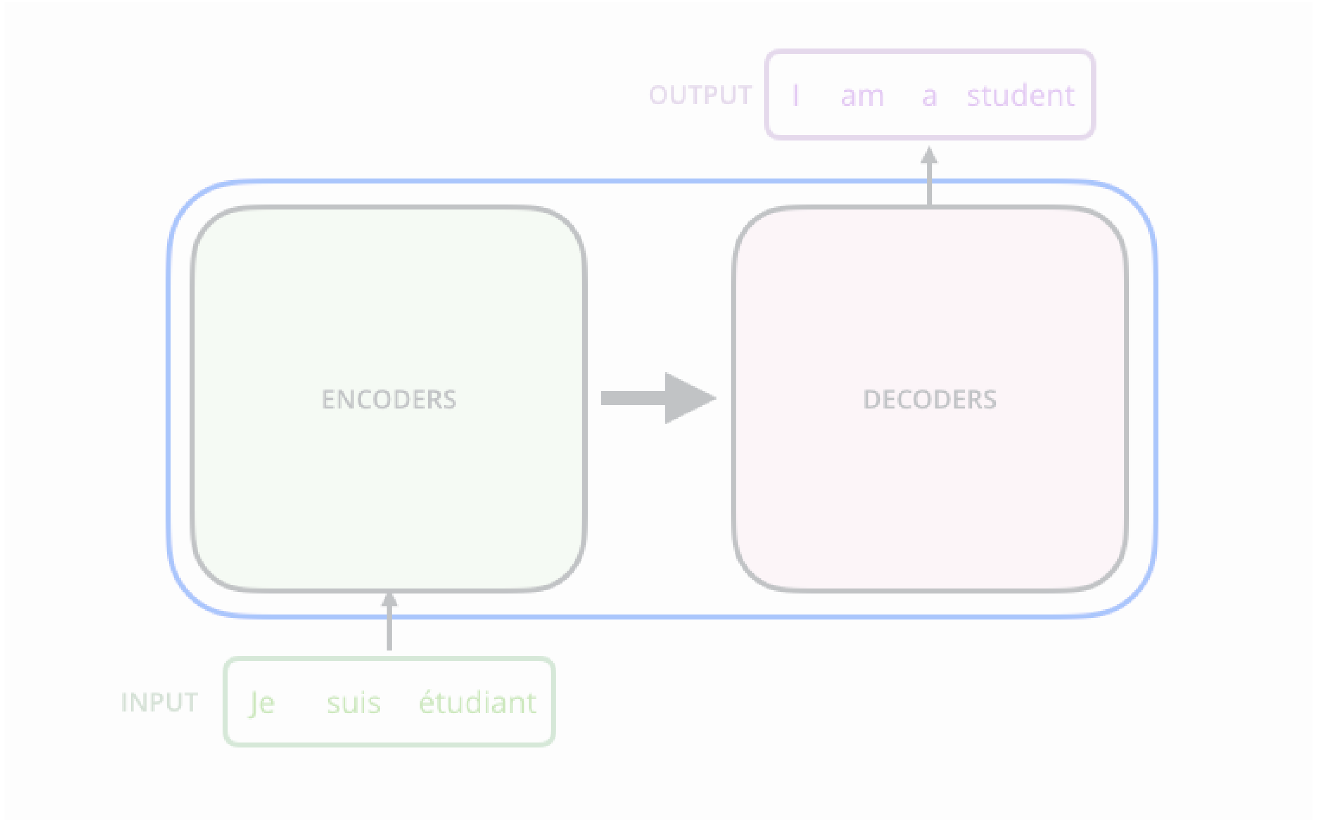

Output Probabilities
Softmax
Linear
Add & Norm
Feed Forward
Add & Norm
Multi-Head Attention
Add & Norm
Masked Multi-Head Attention
Nx
Add & Norm
Feed Forward
Add & Norm
Multi-Head Attention
Nx
Positional Encoding
Positional Encoding
Input Embedding
Output Embedding
Inputs
Outputs (shifted right)

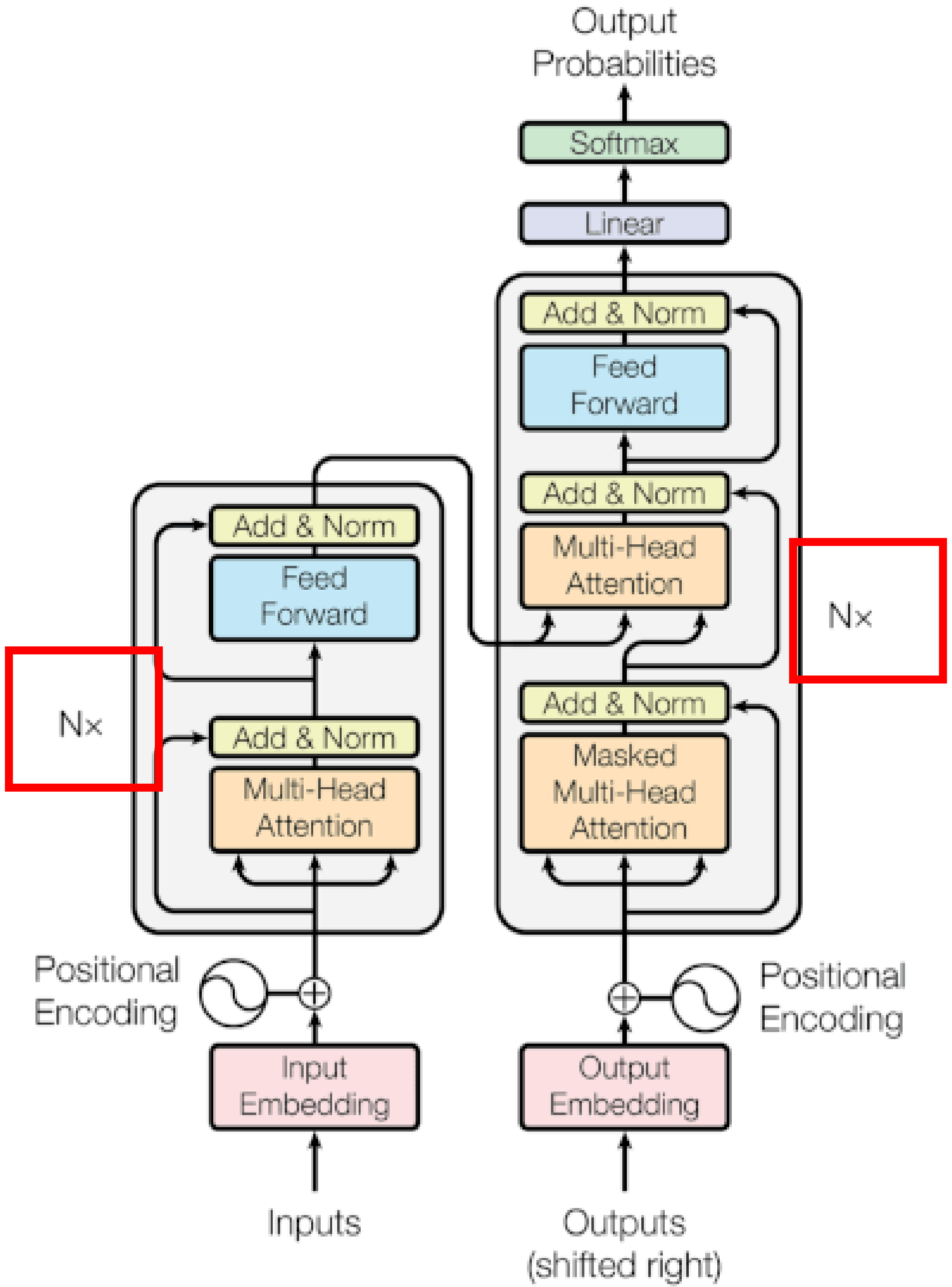

OUTPUT
I am a student
ENCODERS
DECODERS
INPUT
Je suis étudiant

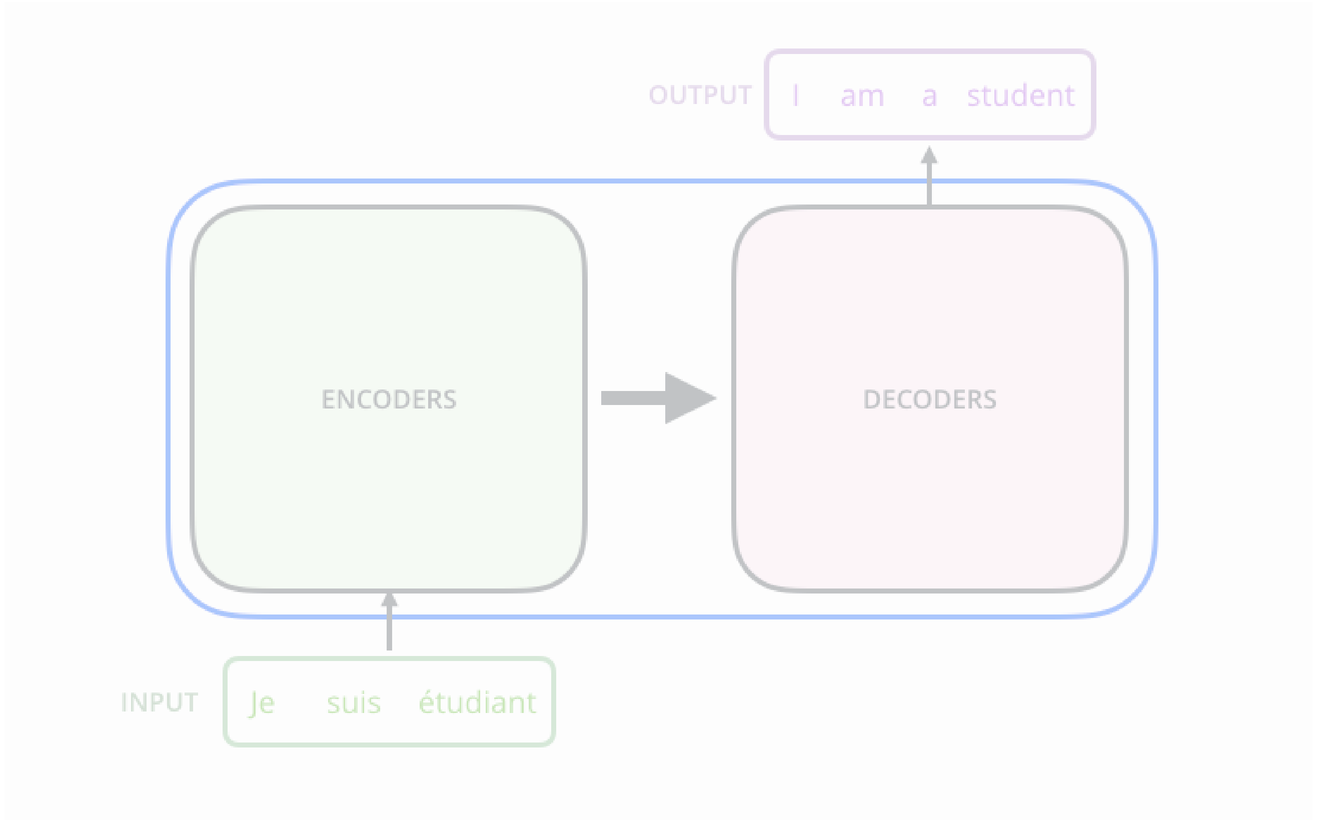

Output Probabilities
Softmax
Linear
Add & Norm
Feed Forward
Add & Norm
Multi-Head Attention
Add & Norm
Masked Multi-Head Attention
Nx
Add & Norm
Feed Forward
Add & Norm
Multi-Head Attention
Nx
Positional Encoding
Input Embedding
Positional Encoding
Output Embedding
Inputs
Outputs (shifted right)

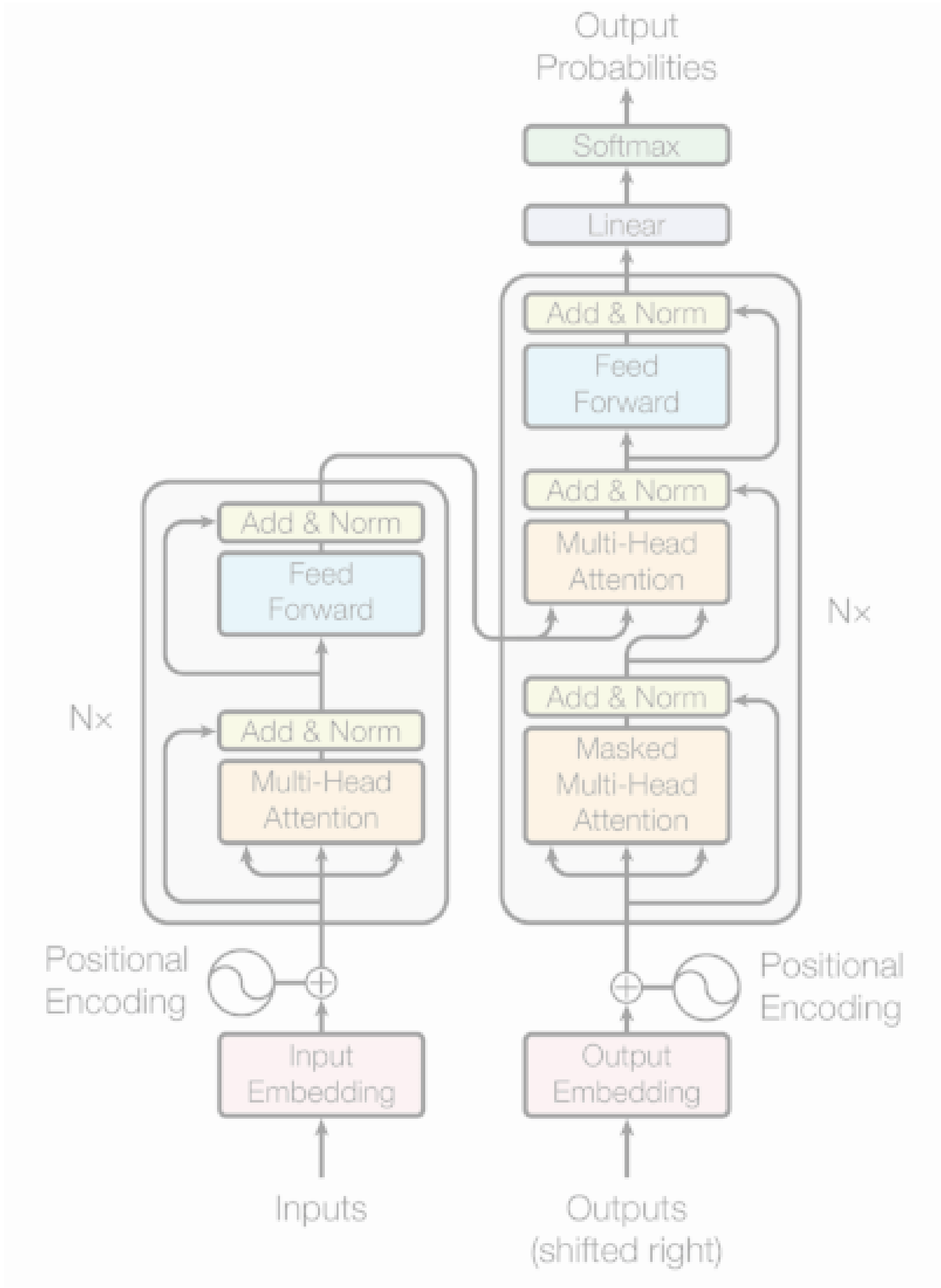

OUTPUT
I am a student
ENCODERS
DECODERS
INPUT
Je suis étudiant


Output Probabilities
Softmax
Linear
Add & Norm
Feed Forward
Add & Norm
Multi-Head Attention
Add & Norm
Masked Multi-Head Attention
Nx
Add & Norm
Feed Forward
Add & Norm
Multi-Head Attention
Nx
Positional Encoding
Positional Encoding
Input Embedding
Output Embedding
Inputs
Outputs (shifted right)

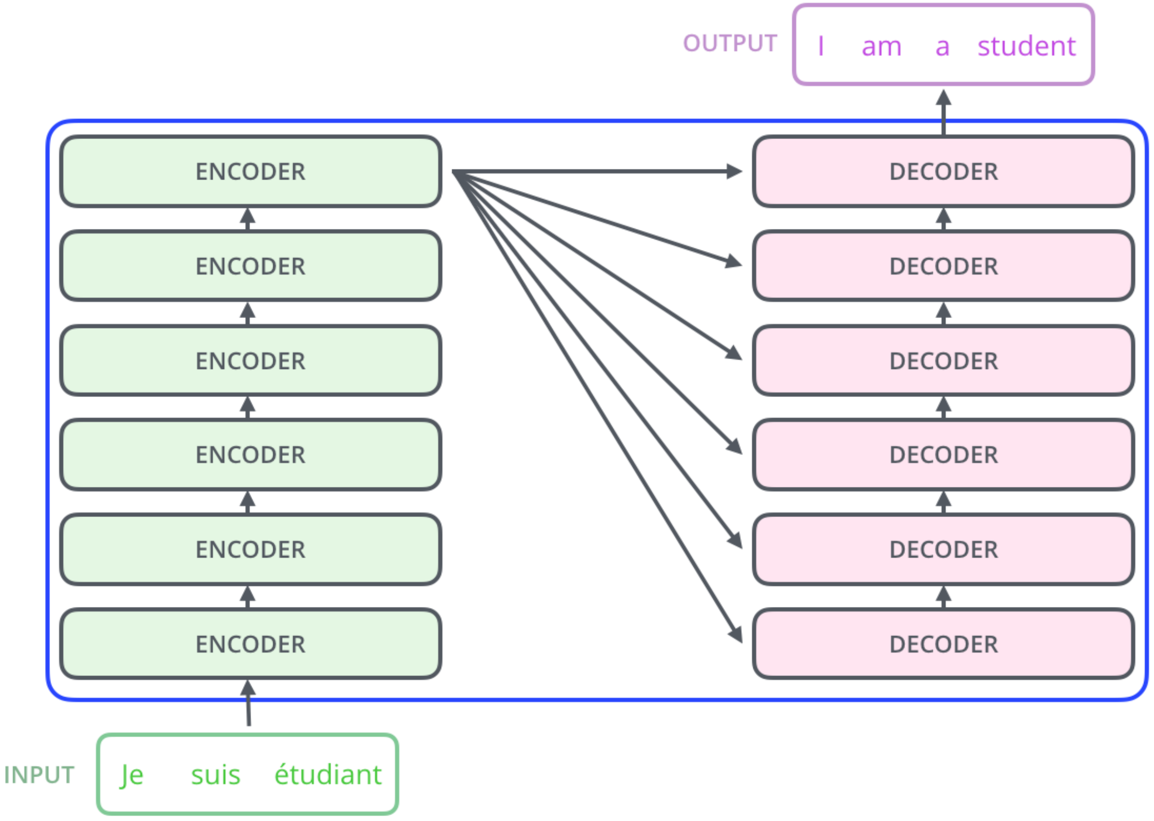

OUTPUT
I am a student
ENCODERS
DECODERS
INPUT
Je suis étudiant


Output Probabilities
Softmax
Linear
Add & Norm
Feed Forward
Add & Norm
Multi-Head Attention
Add & Norm
Masked Multi-Head Attention
Nx
Add & Norm
Feed Forward
Add & Norm
Multi-Head Attention
Nx
Positional Encoding
Positional Encoding
Input Embedding
Output Embedding
Inputs
Outputs (shifted right)

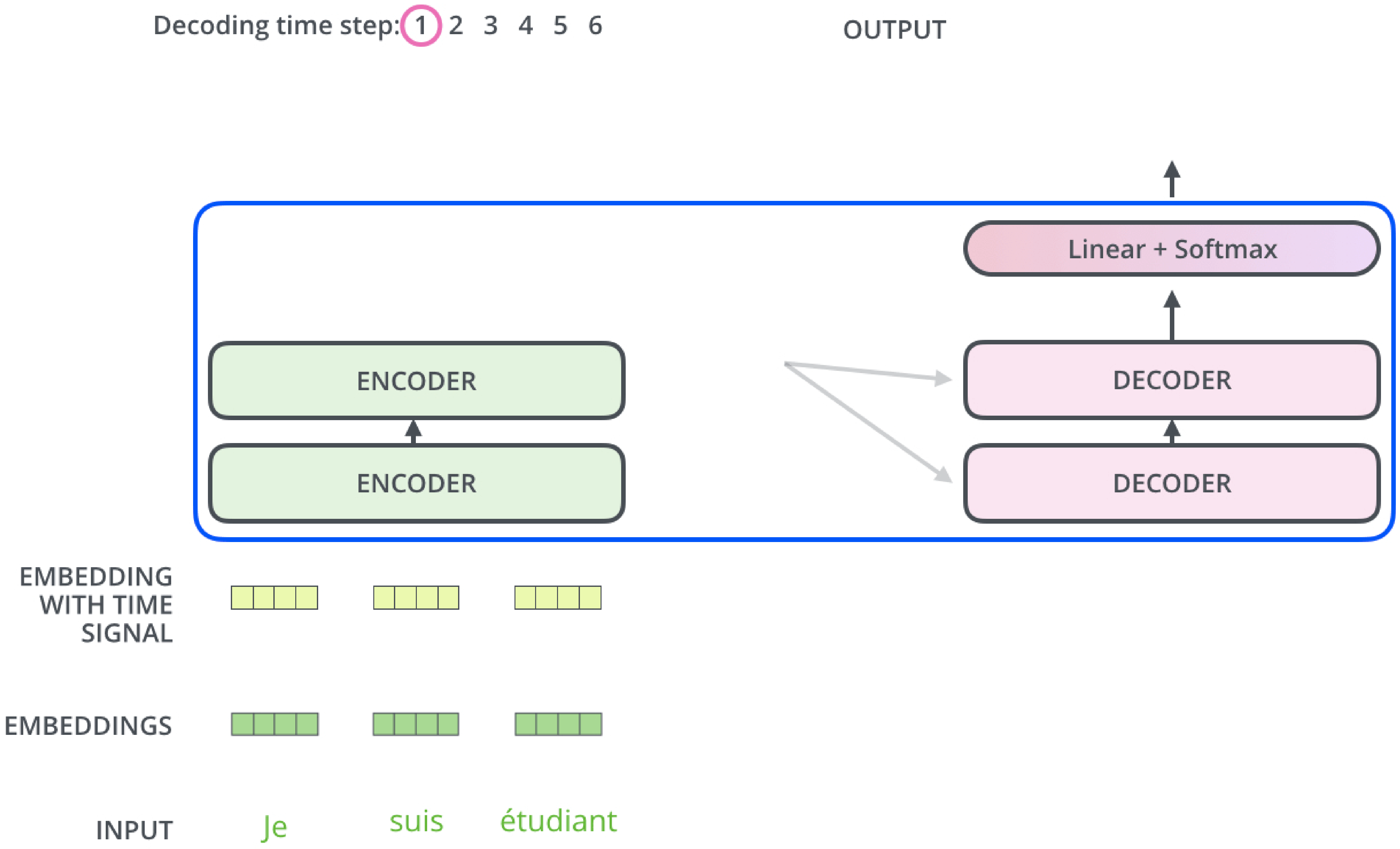

OUTPUT
I am a student
ENCODERS
DECODERS
INPUT
Je suis étudiant



Output Probabilities
Softmax
Linear
Add & Norm
Feed Forward
Add & Norm
Multi-Head Attention
Add & Norm
Masked Multi-Head Attention
Nx
Add & Norm
Feed Forward
Add & Norm
Multi-Head Attention
Nx
Positional Encoding
Positional Encoding
Input Embedding
Output Embedding
Inputs
Outputs (shifted right)



OUTPUT
I am a student
ENCODER
ENCODER
ENCODER
ENCODER
ENCODER
ENCODER
DECODER
DECODER
DECODER
DECODER
DECODER
DECODER
INPUT
Je suis étudiant



Decoding time step: 1 2 3 4 5 6

OUTPUT

Linear + Softmax

ENCODER

ENCODER

DECODER

DECODER

EMBEDDING WITH TIME SIGNAL

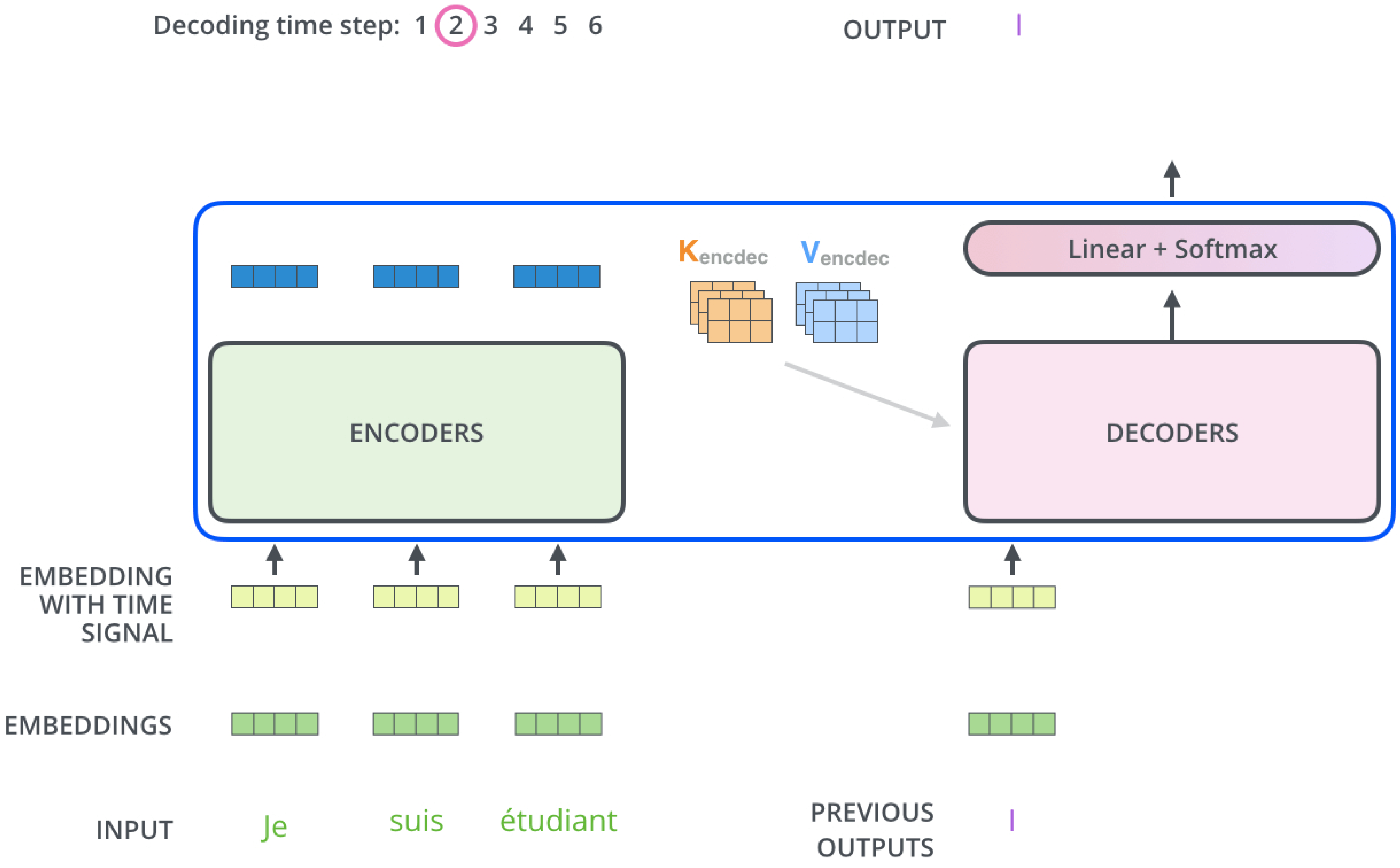
EMBEDDINGS

INPUT

Je

suis

étudiant



Decoding time step: 1 2 3 4 5 6

OUTPUT I

K_{encdec} V_{encdec}

Linear + Softmax

ENCODERS

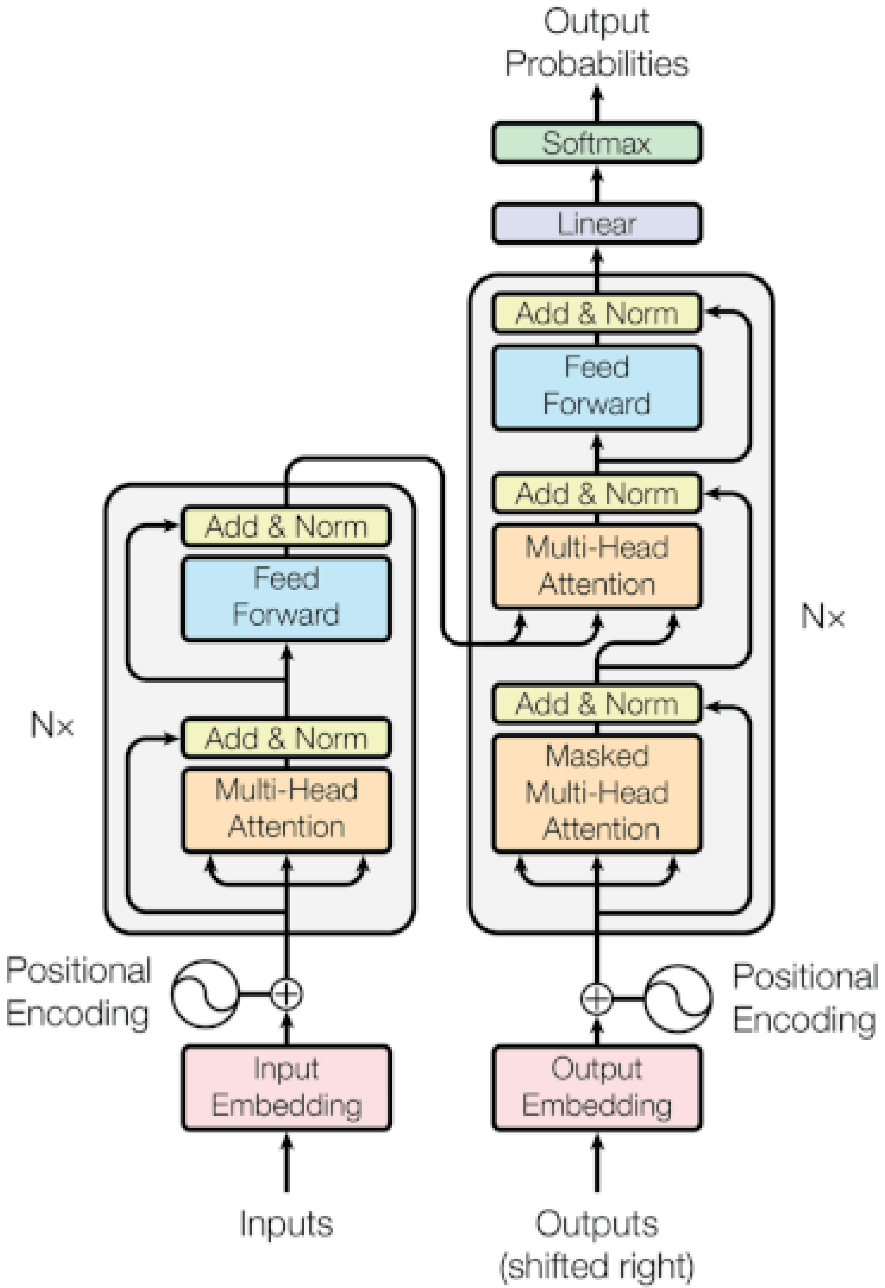
DECODERS

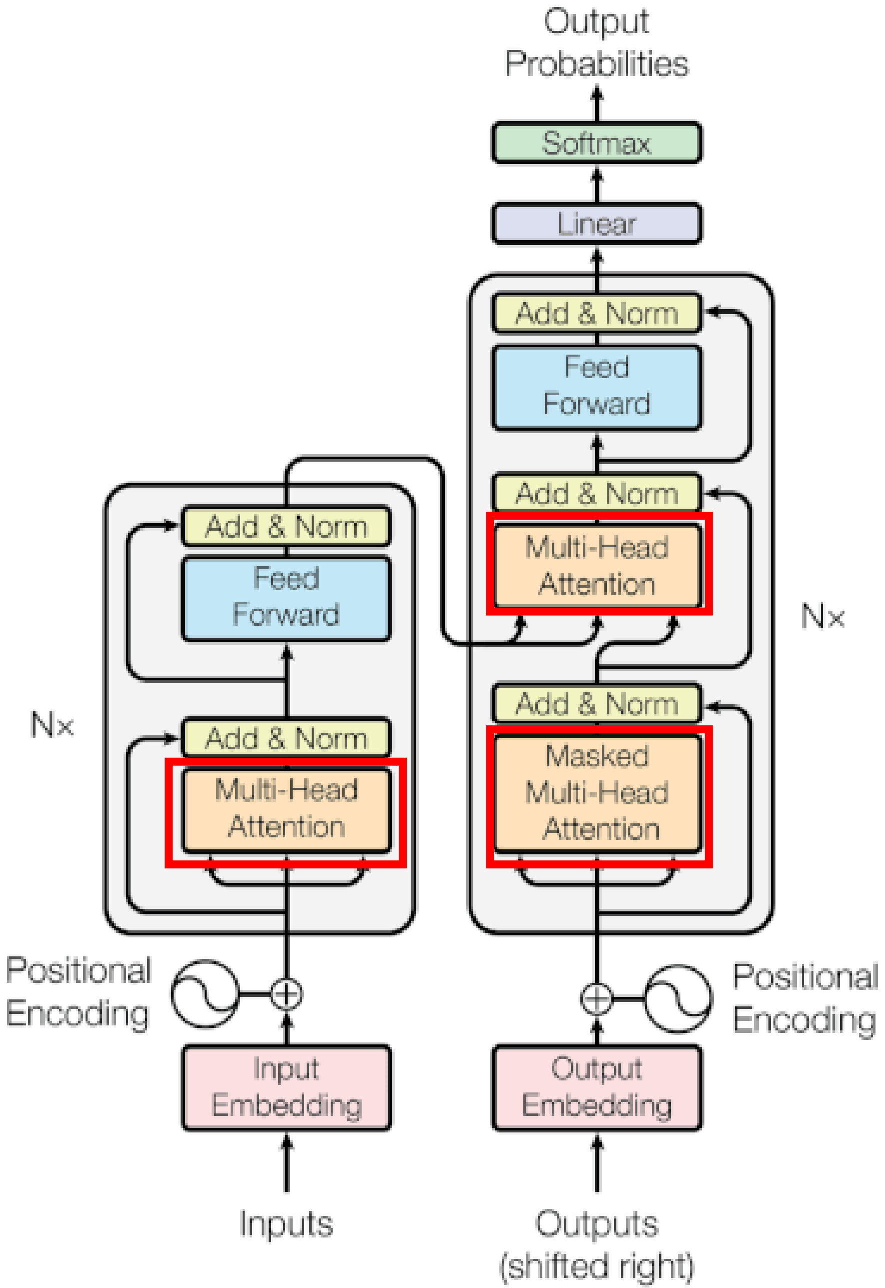
EMBEDDING WITH TIME SIGNAL

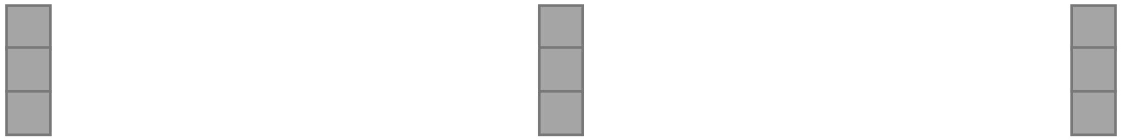
EMBEDDINGS

INPUT Je suis étudiant

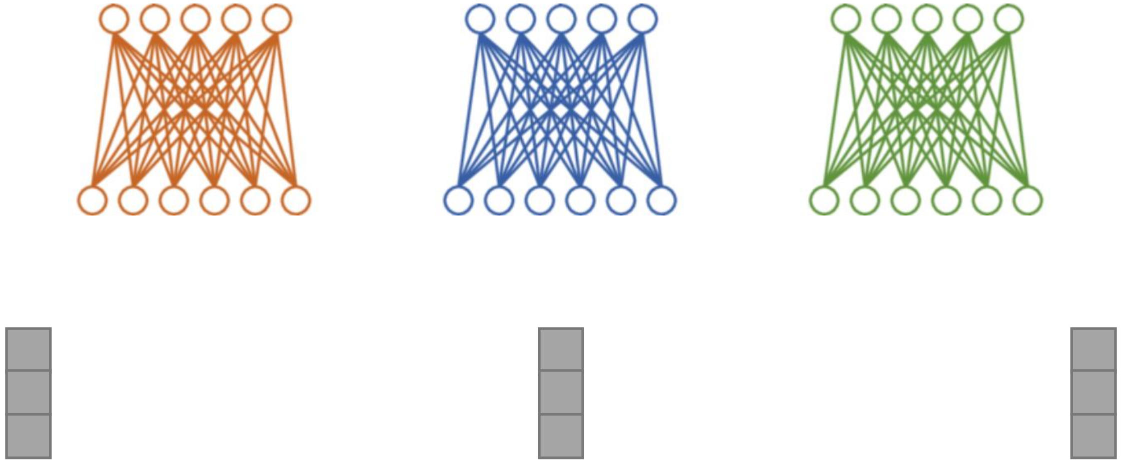
PREVIOUS OUTPUTS I


Output Probabilities
Softmax
Linear
Add & Norm
Feed Forward
Add & Norm
Multi-Head Attention
Nx
Add & Norm
Masked Multi-Head Attention
Add & Norm
Feed Forward
Nx
Add & Norm
Multi-Head Attention
Positional Encoding
Positional Encoding
Input Embedding
Output Embedding
Inputs
Outputs (shifted right)

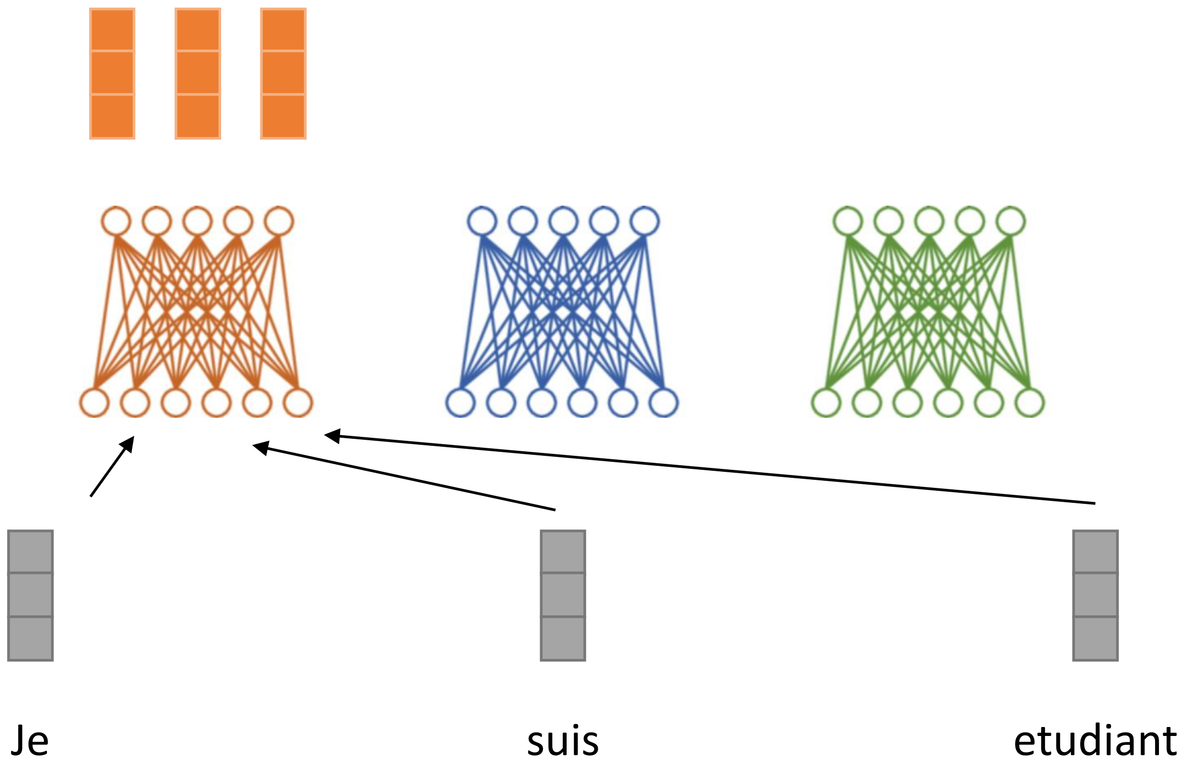

Output
Probabilities
Softmax
Linear
Add & Norm
Feed
Forward
Add & Norm
Multi-Head
Attention
Nx
Add & Norm
Masked
Multi-Head
Attention
Add & Norm
Feed
Forward
Nx
Add & Norm
Multi-Head
Attention
Positional
Encoding
Positional
Encoding
Input
Embedding
Output
Embedding
Inputs
Outputs
(shifted right)



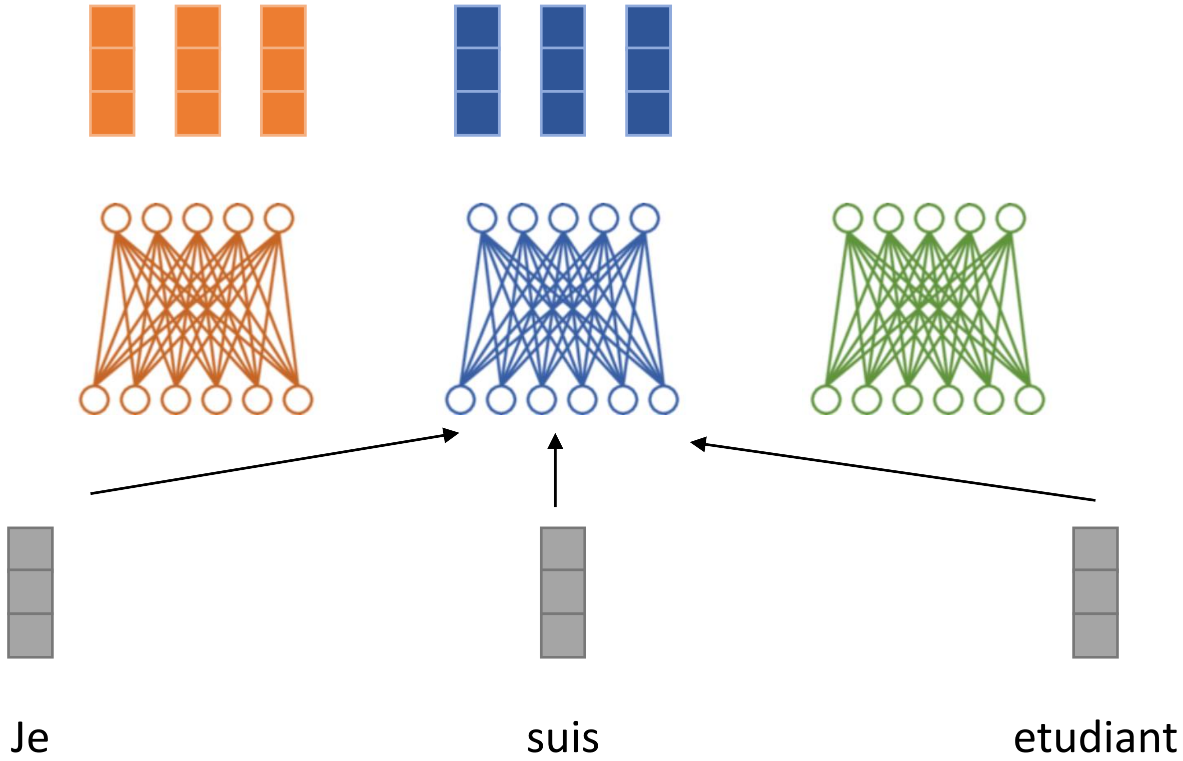
Je suis etudiant



Je suis etudiant



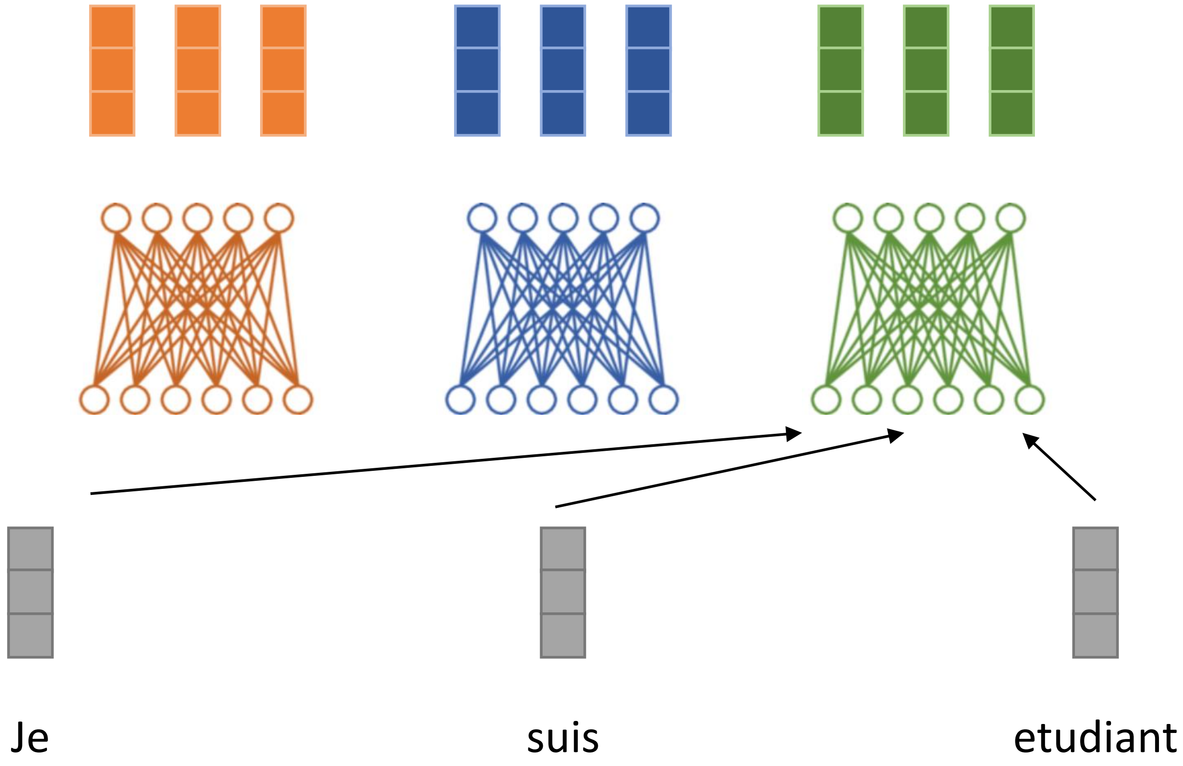
Je suis etudiant



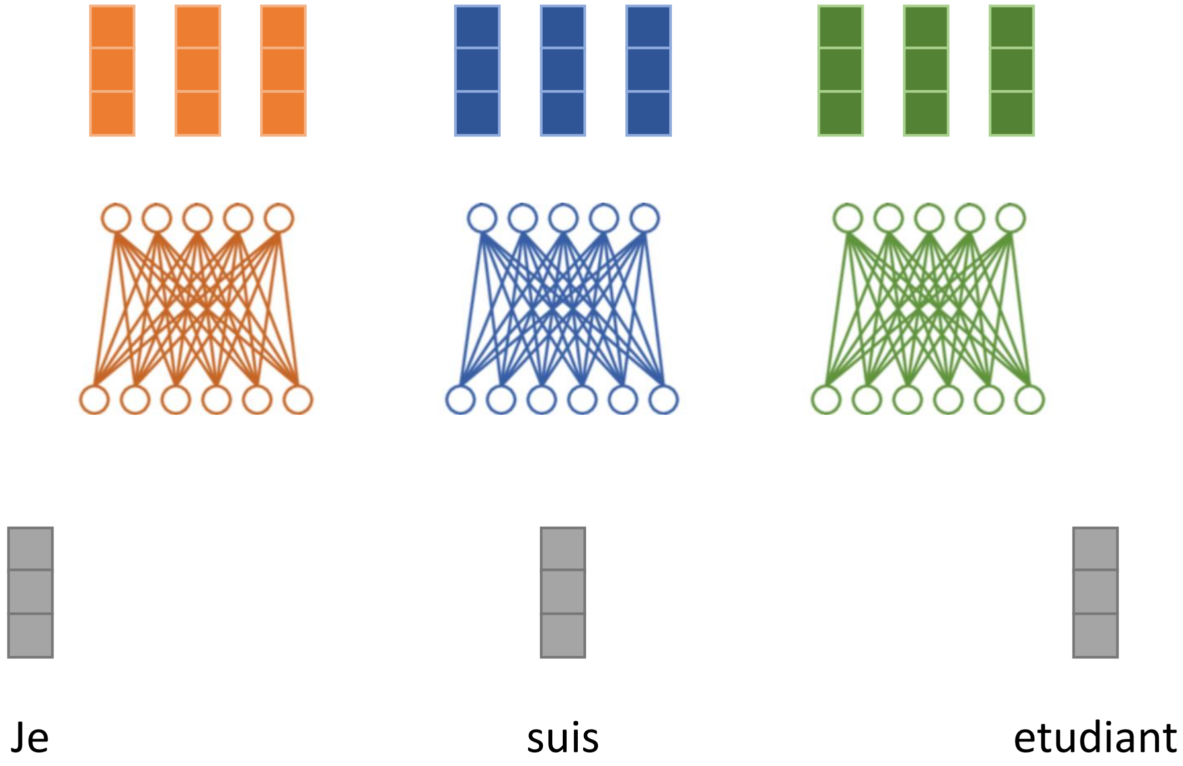
Je

suis

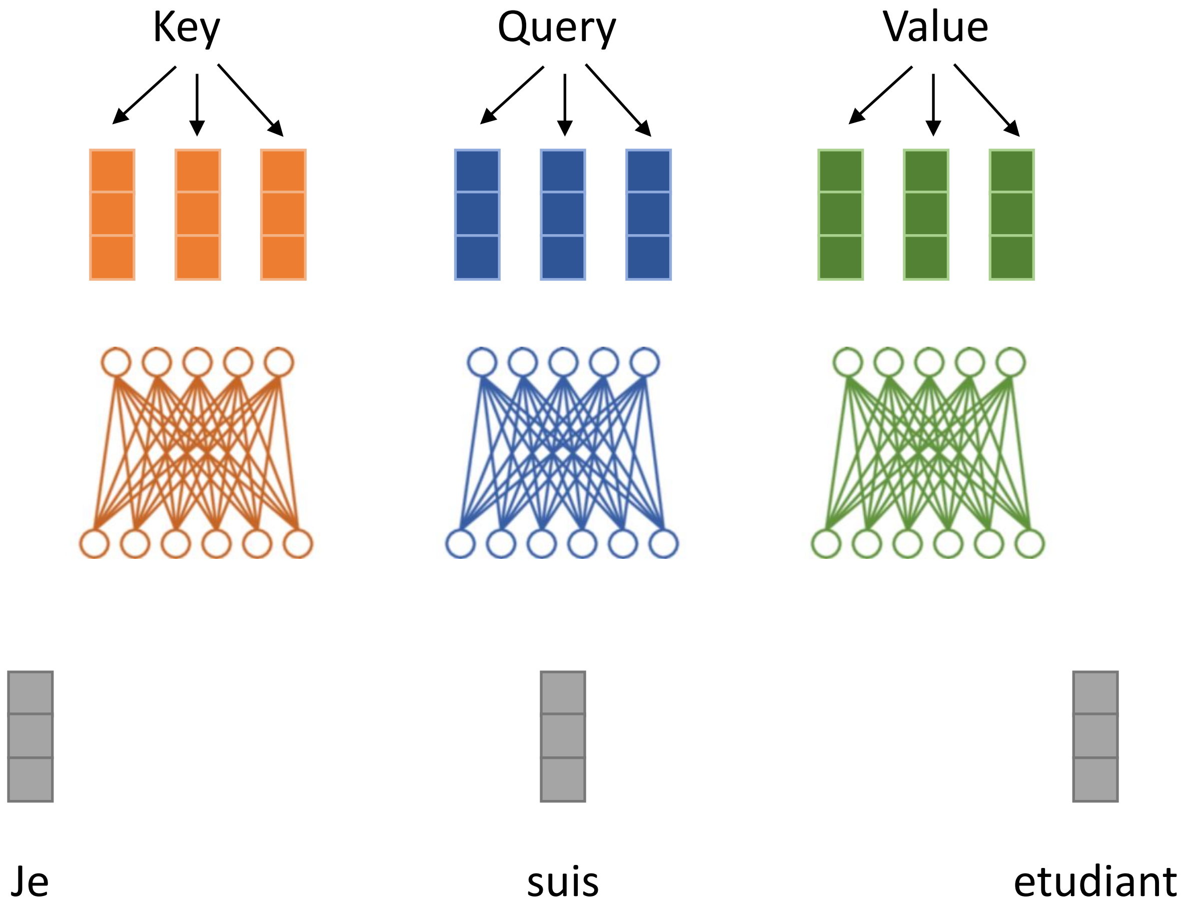
etudiant

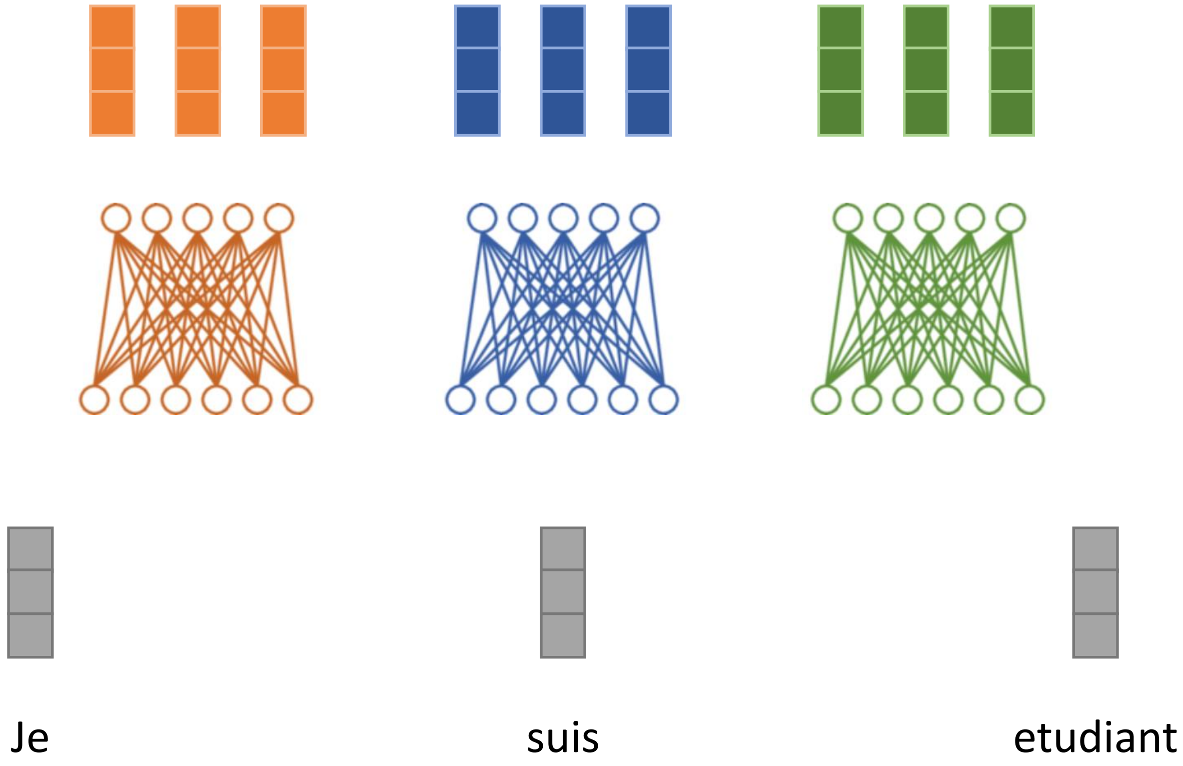

Je
suis
etudiant

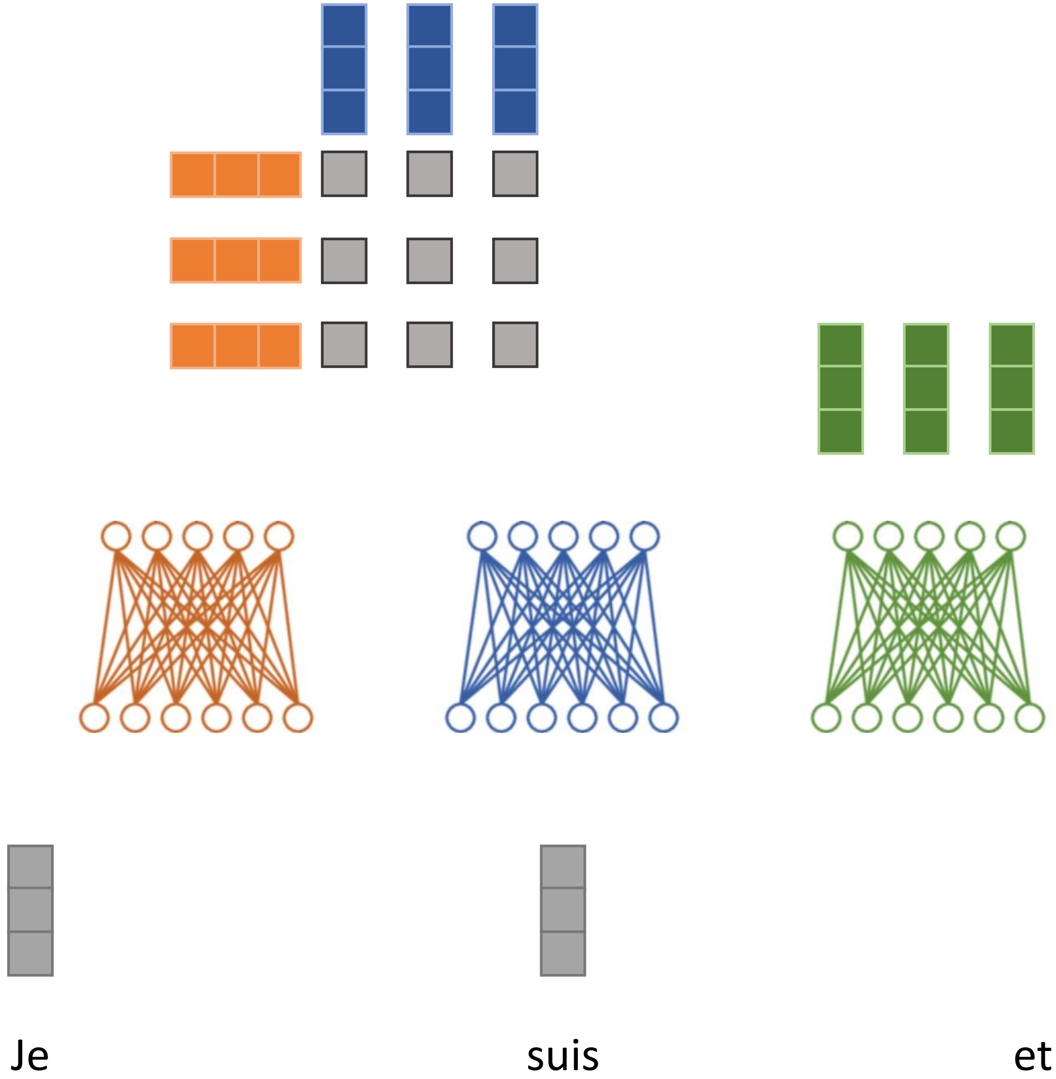
TERMINOLOGY

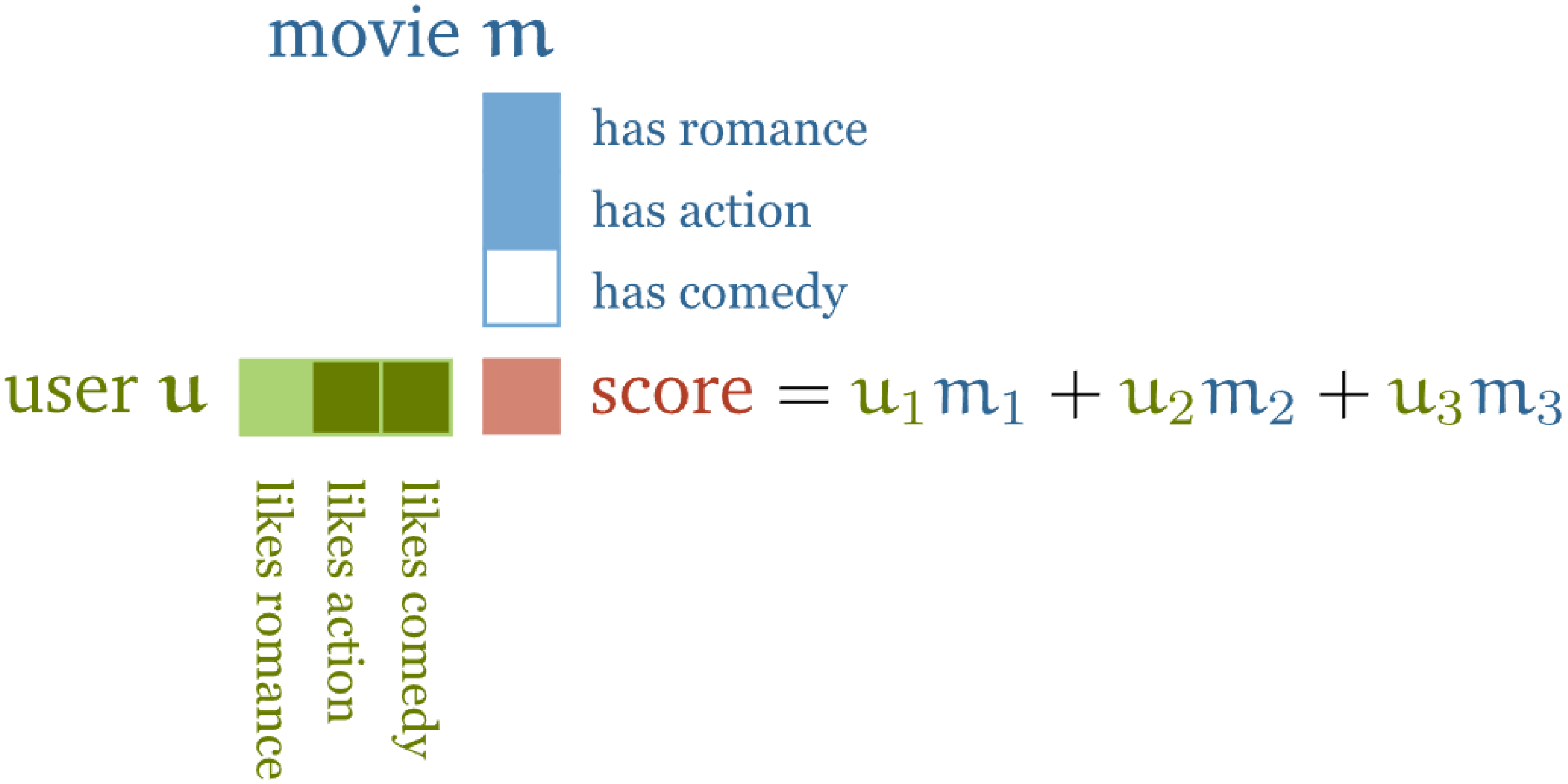


TERMINOLOGY




Je
suis
etudiant


Je
suis
etudiant



movie $\mathbf{m}$

has romance

has action

has comedy

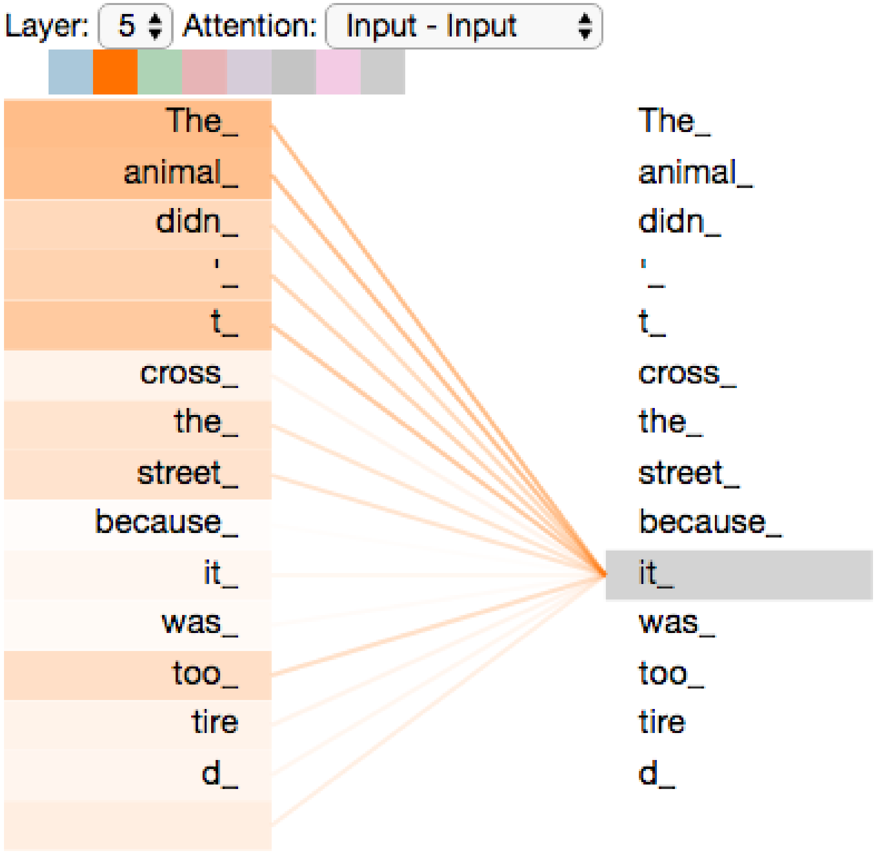
user $\mathbf{u}$

score $= u_1 m_1 + u_2 m_2 + u_3 m_3$

likes romance

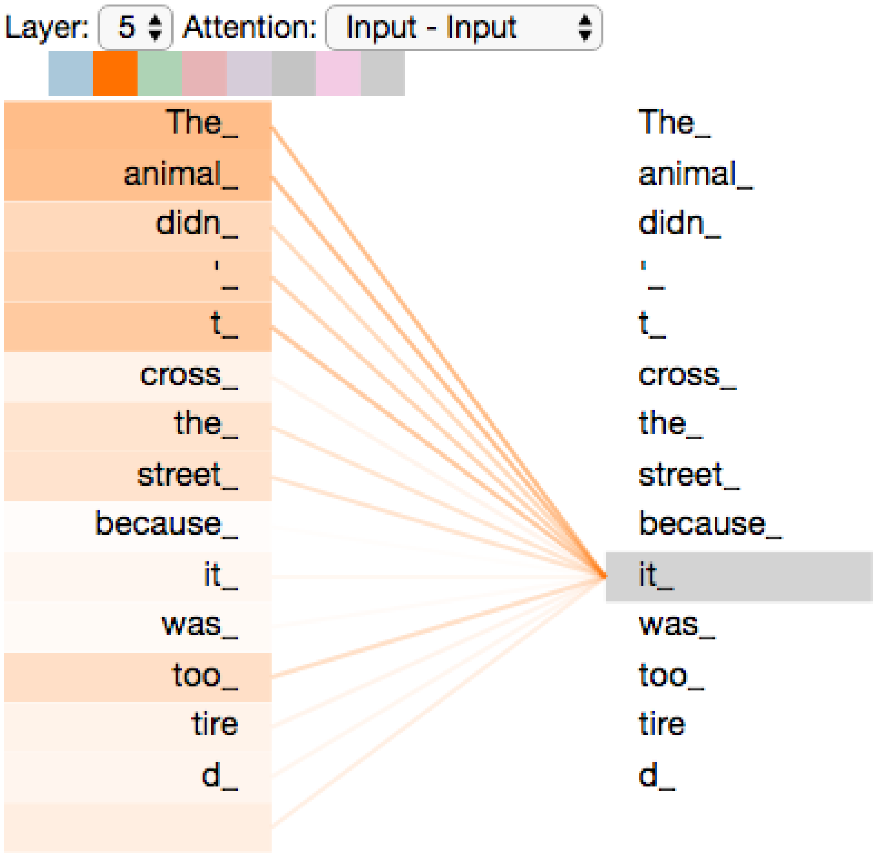
likes action

likes comedy



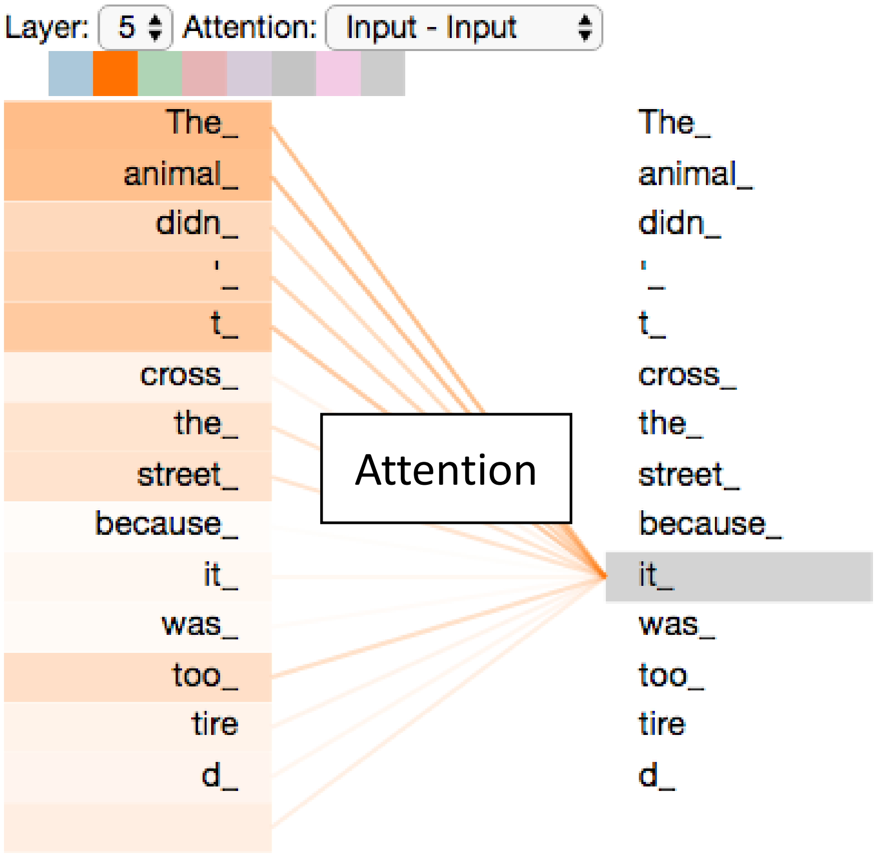
Layer: 5 Attention: Input - Input

TERMINOLOGY



Layer: 5 Attention: Input - Input

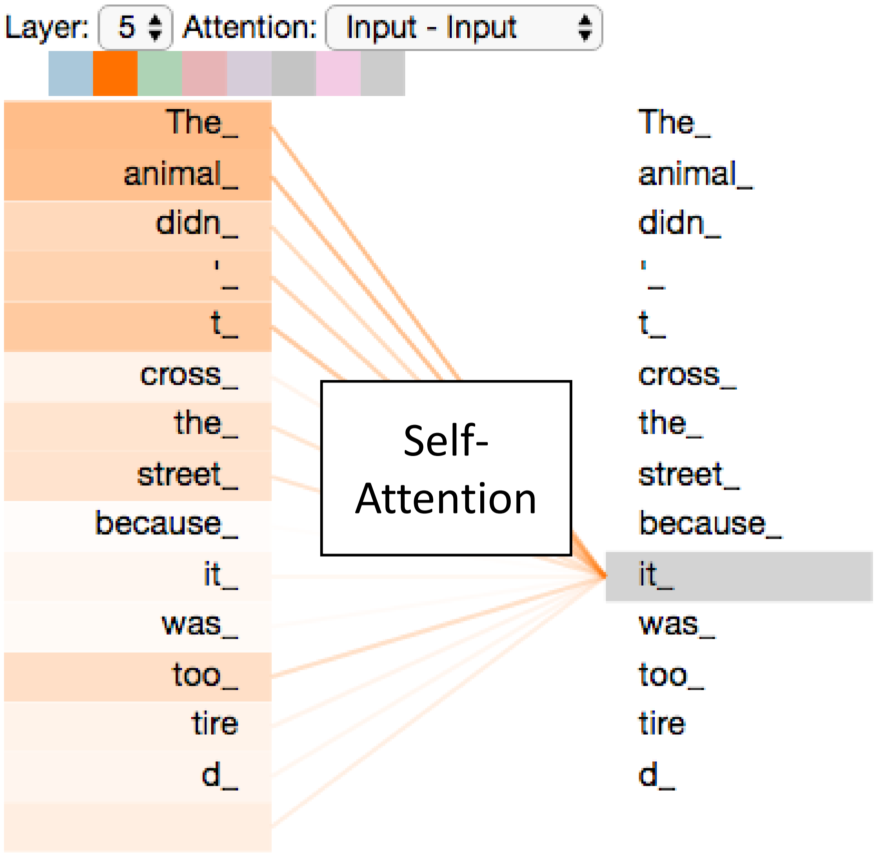
TERMINOLOGY



Layer: 5 Attention: Input - Input

Attention

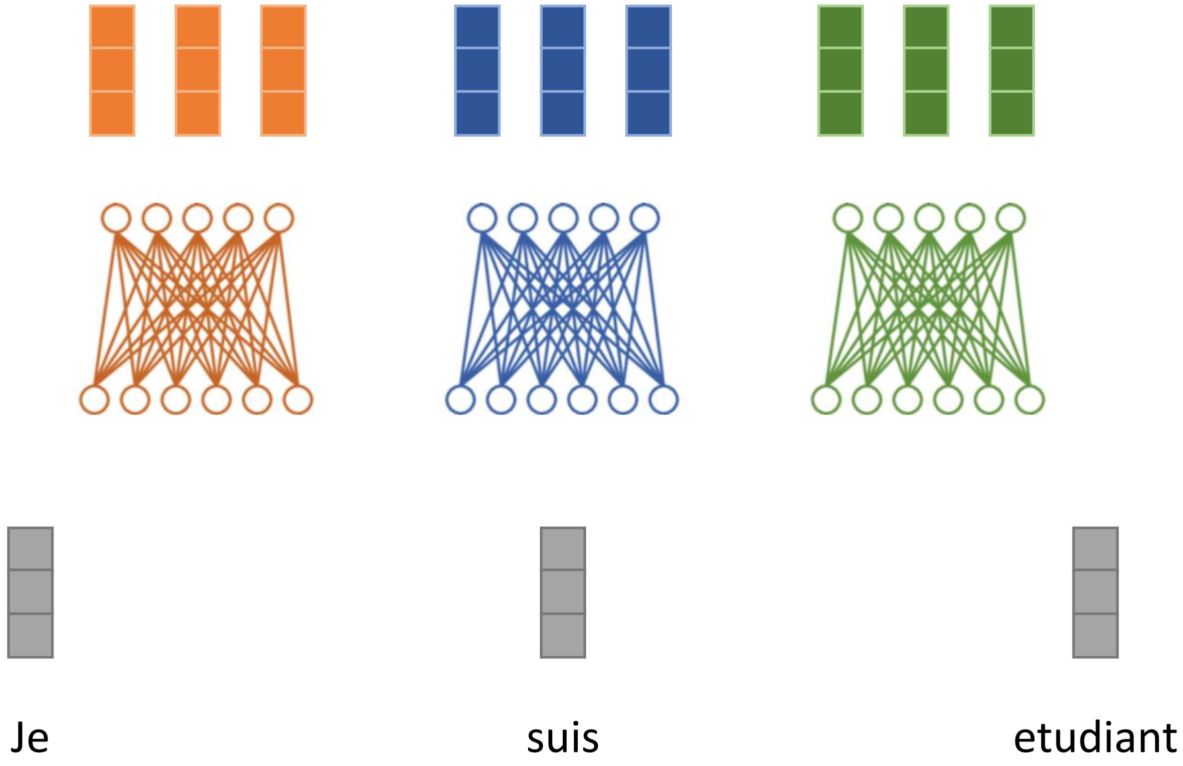
TERMINOLOGY



Layer: 5

Attention: Input - Input

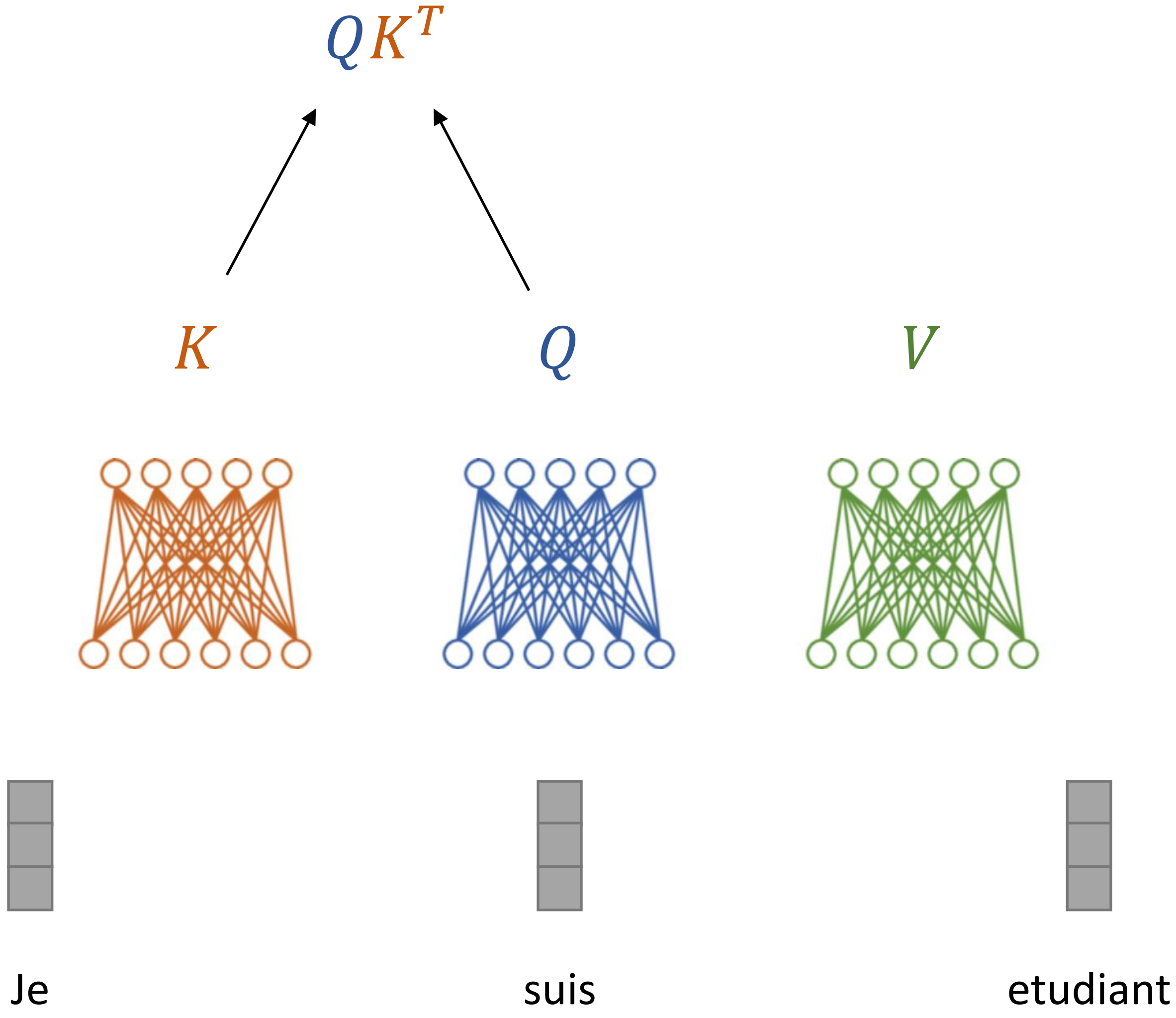
Self-Attention



Je

suis

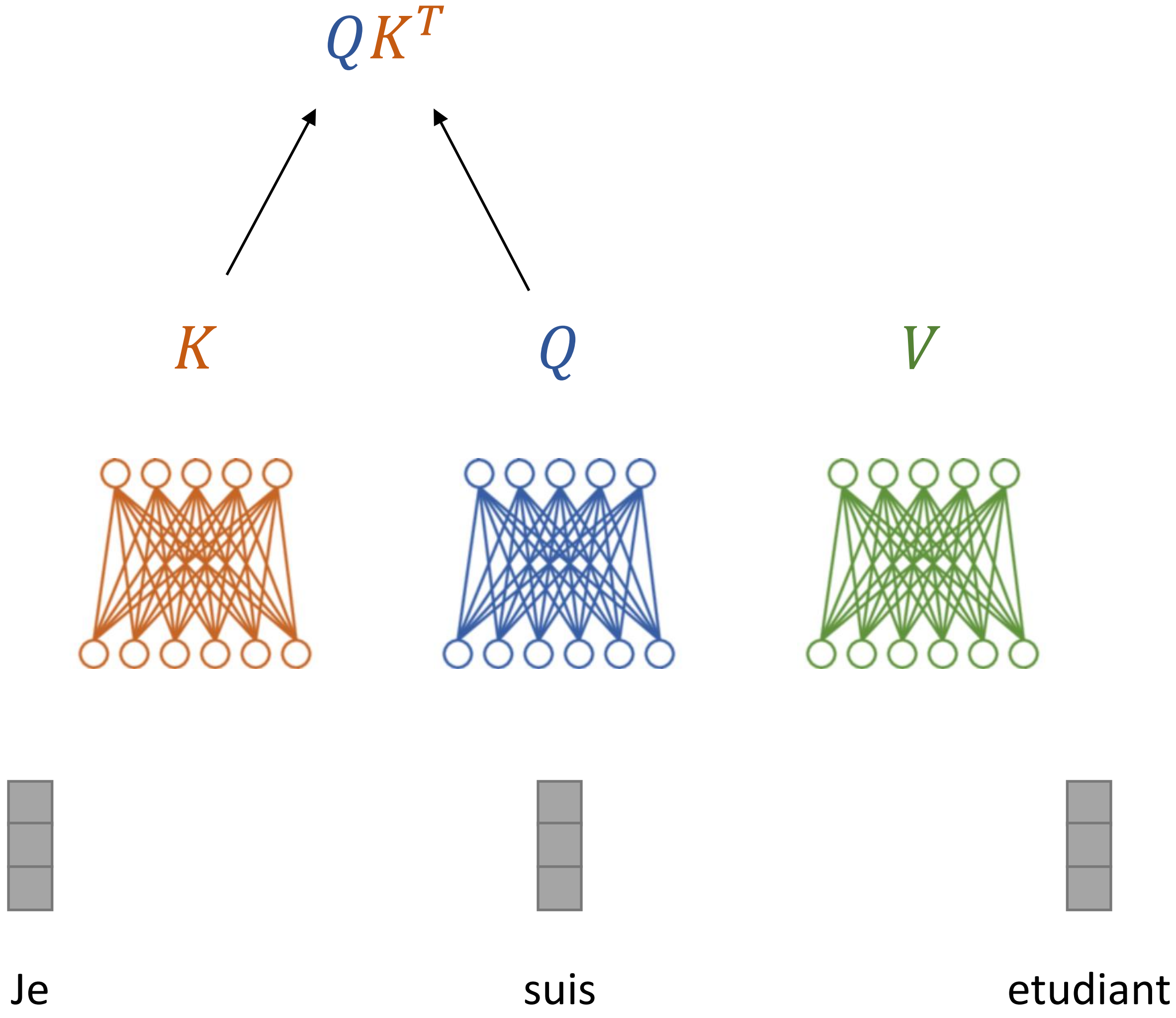
etudiant



QK^T

K Q V

Je suis etudiant



QK^T

K

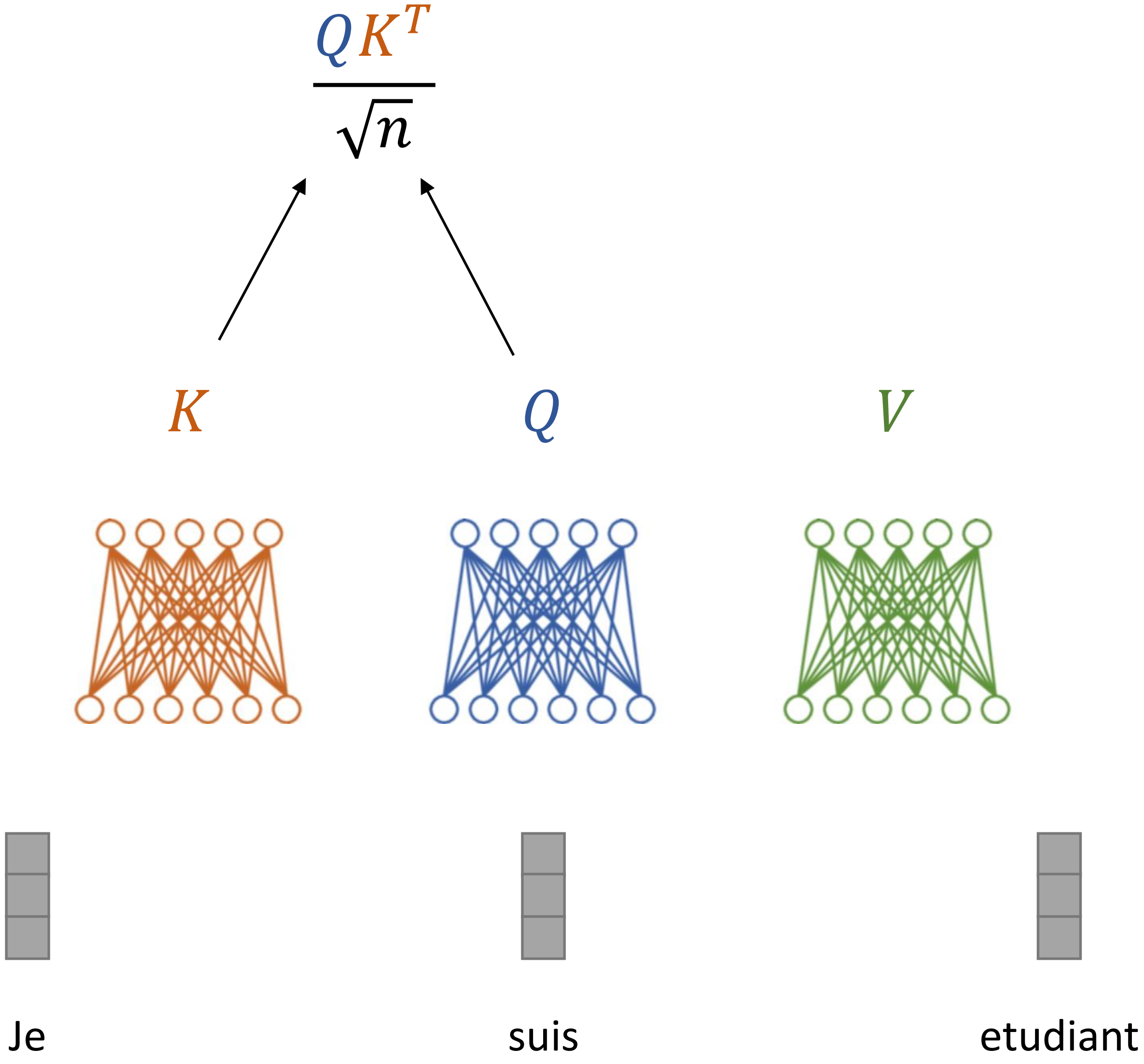
Q

V

Je

suis

etudiant



$$\frac{QK^T}{\sqrt{n}}$$

K

Q

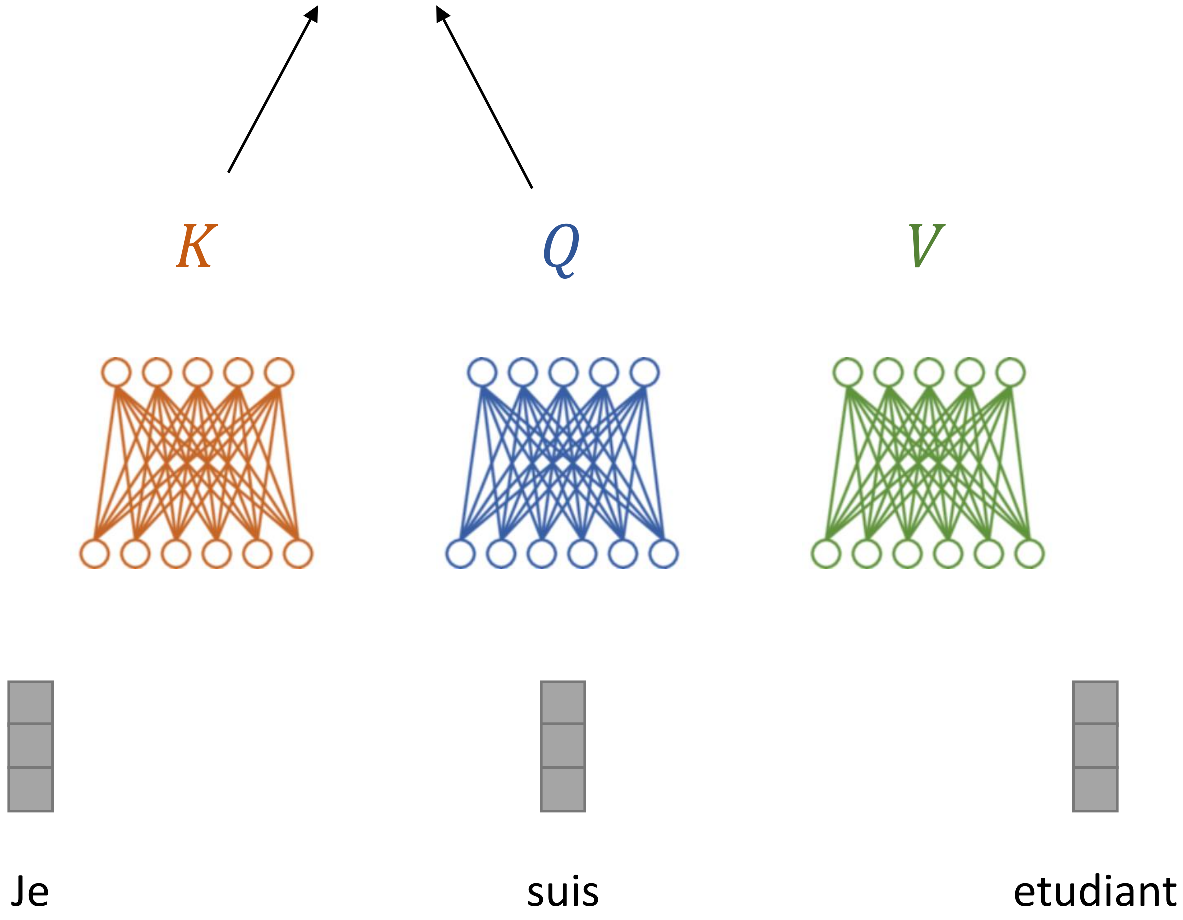
V

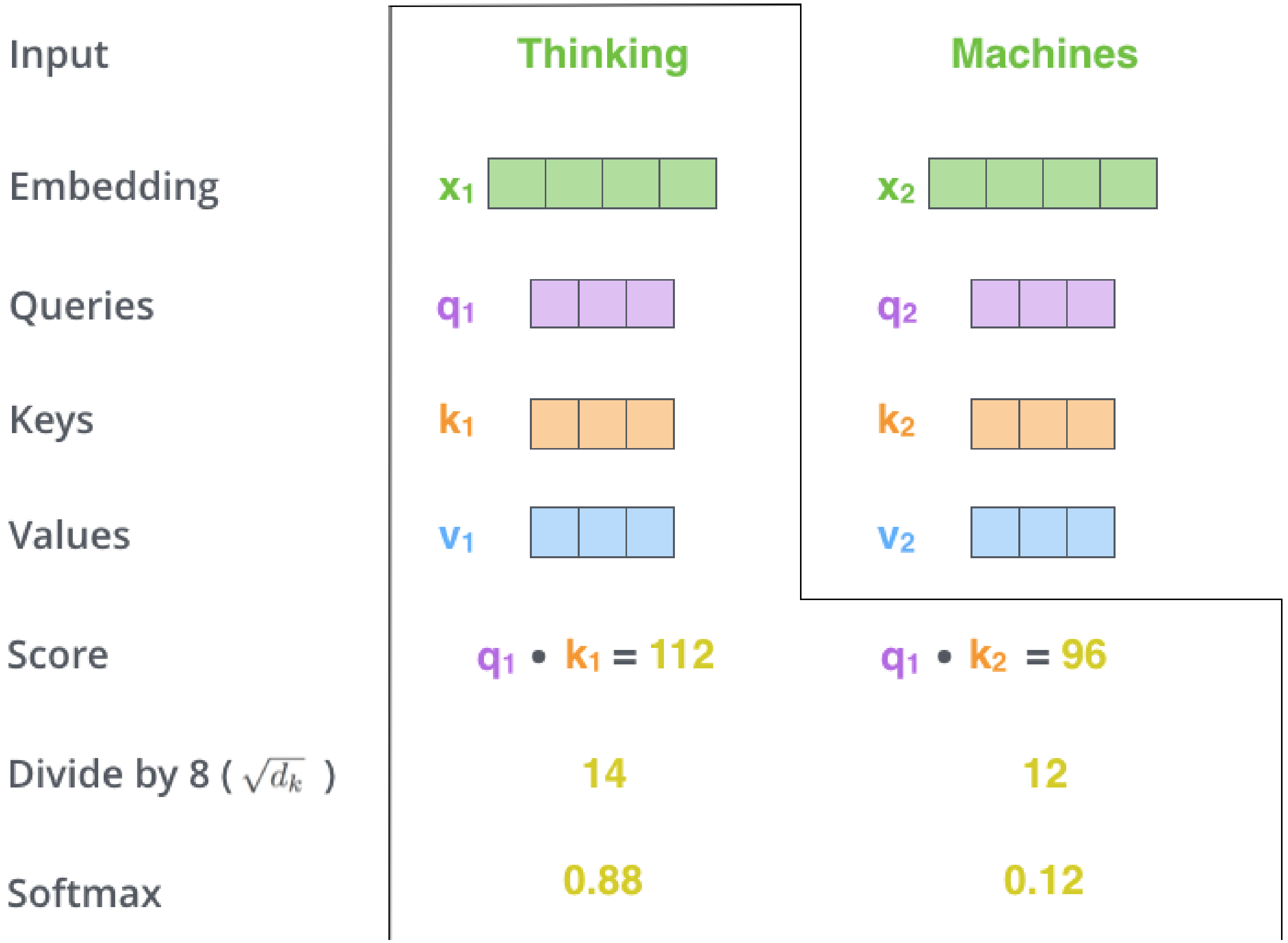
Je

suis

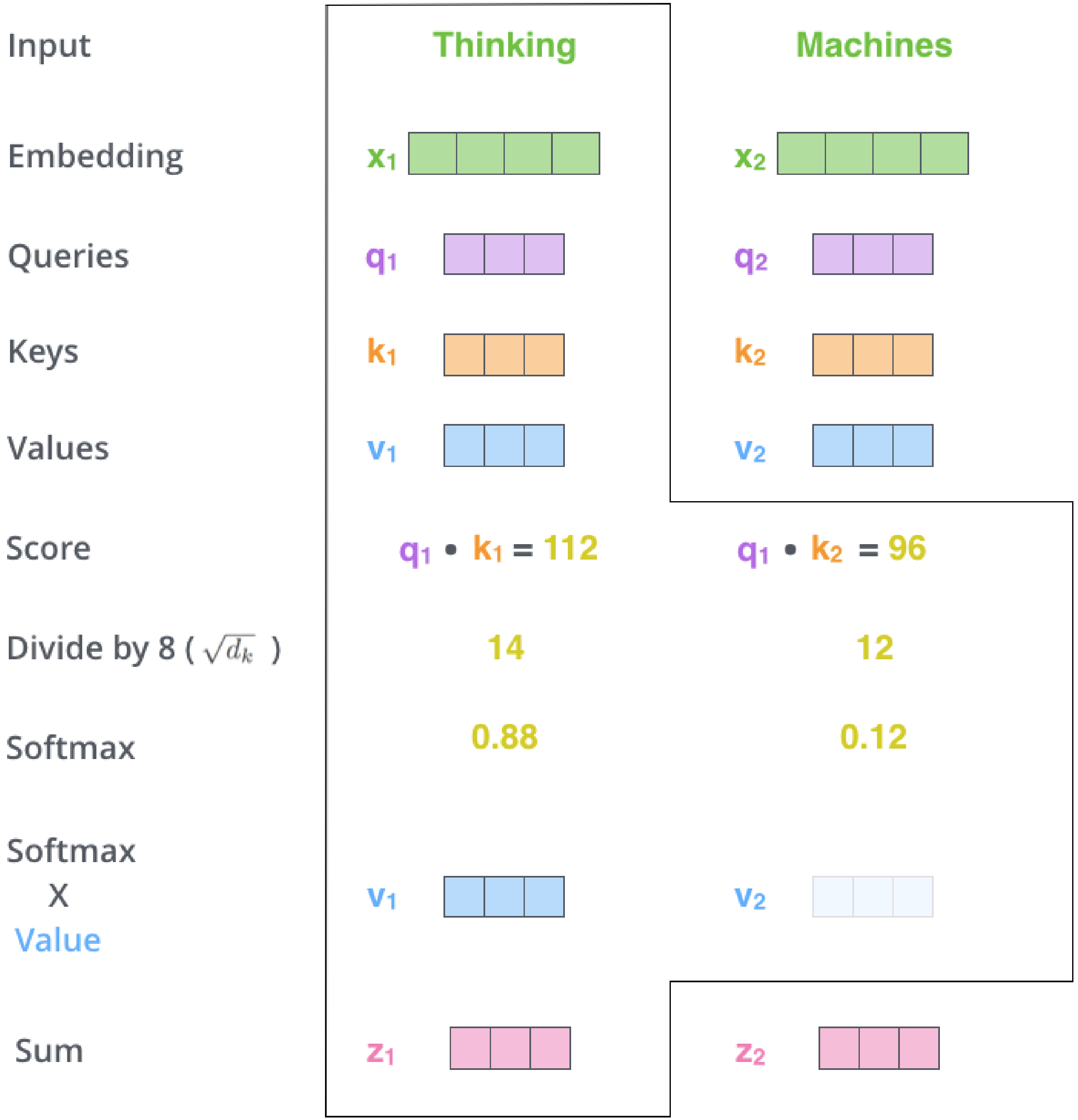
etudiant

$$softmax\left(\frac{QK^T}{\sqrt{n}}\right)$$

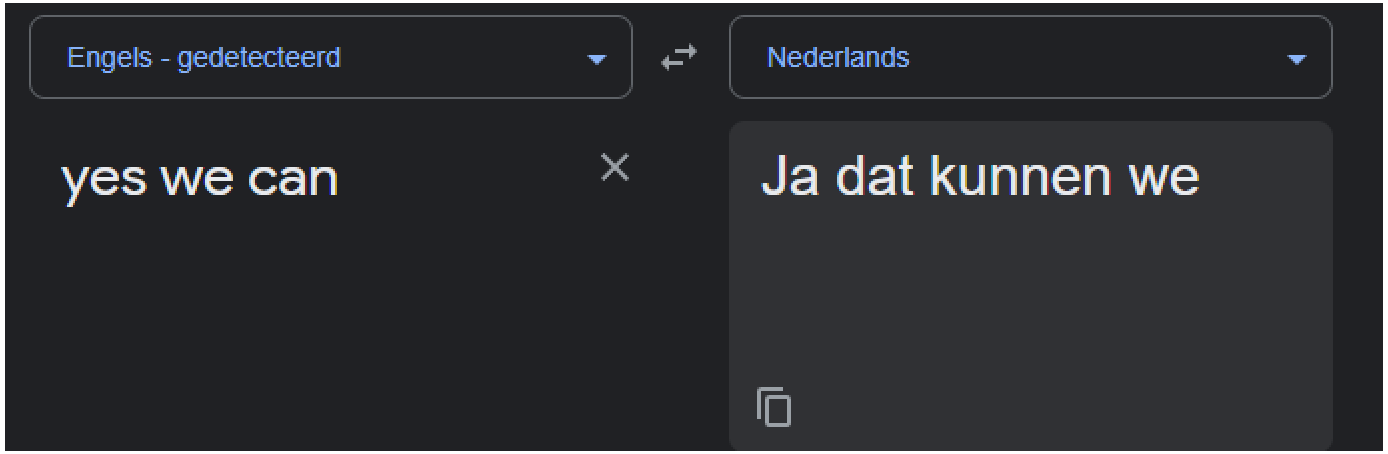




Input	Thinking	Machines
Embedding	x_1 🟩🟩🟩🟩	x_2 🟩🟩🟩🟩
Queries	q_1 🟪🟪🟪	q_2 🟪🟪🟪
Keys	k_1 🟧🟧🟧	k_2 🟧🟧🟧
Values	v_1 🟦🟦🟦	v_2 🟦🟦🟦
Score	$q_1 \bullet k_1 = 112$	$q_1 \bullet k_2 = 96$
Divide by 8 ($\sqrt{d_k}$)	14	12
Softmax	0.88	0.12

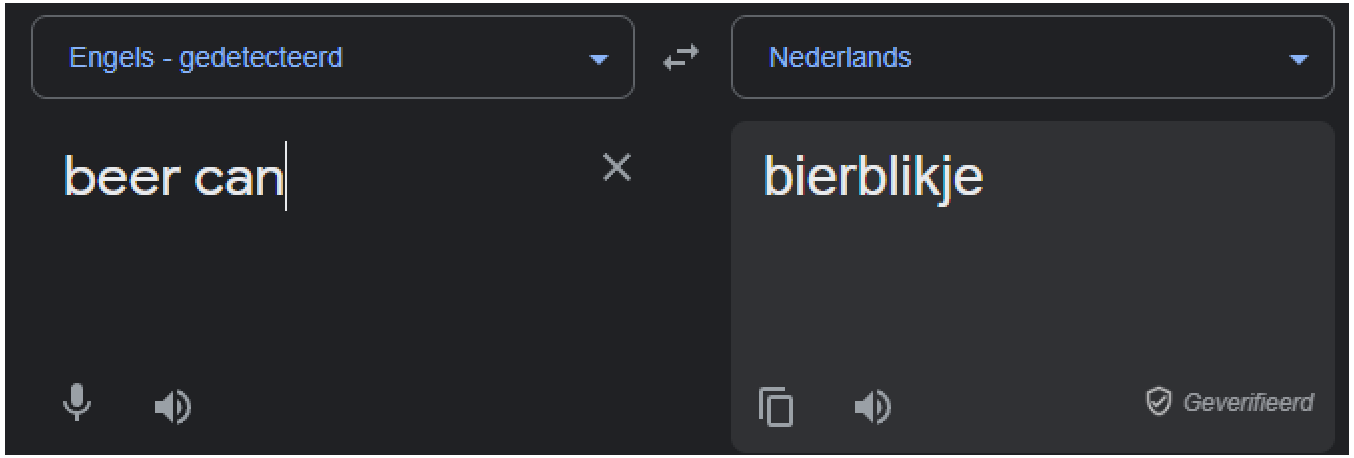


Input	Thinking	Machines
Embedding	x_1	x_2
Queries	q_1	q_2
Keys	k_1	k_2
Values	v_1	v_2
Score	$q_1 \bullet k_1 = 112$	$q_1 \bullet k_2 = 96$
Divide by 8 ($\sqrt{d_k}$)	14	12
Softmax	0.88	0.12
Softmax X Value	v_1	v_2
Sum	z_1	z_2



yes we can

Ja dat kunnen we



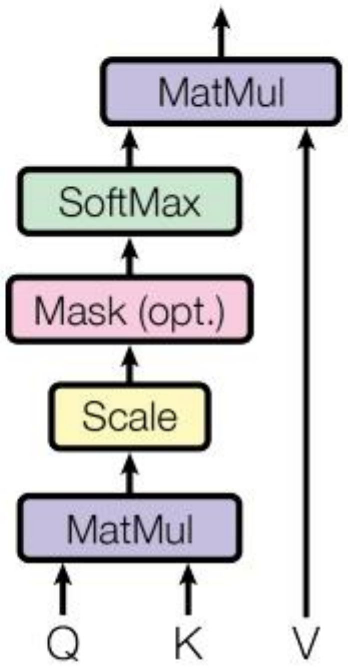
Engels - gedetecteerd

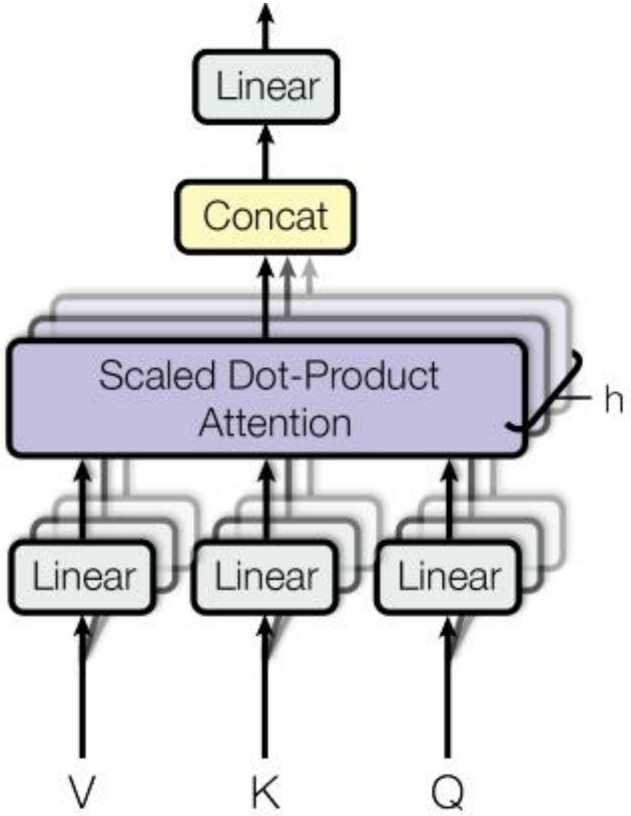
Nederlands

beer can

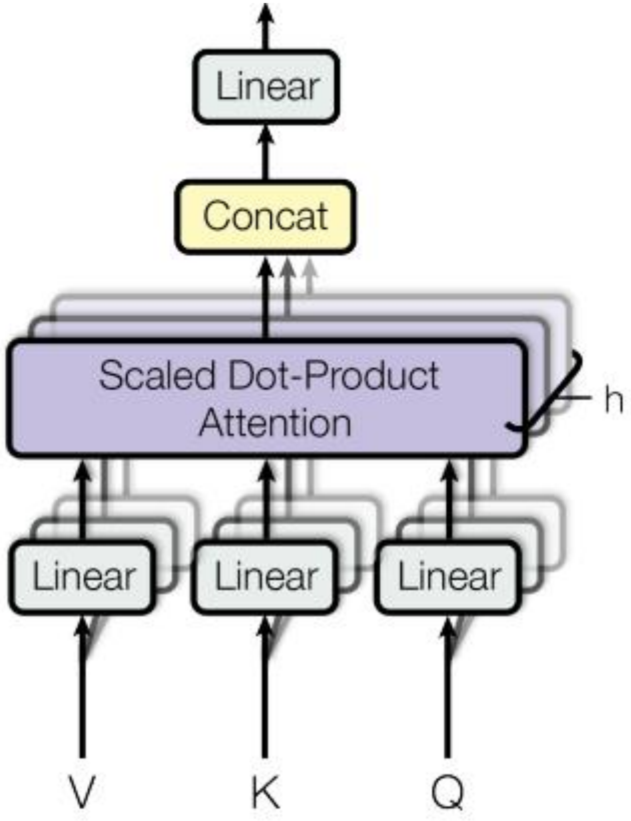
bierblikje

Scaled Dot-Product Attention

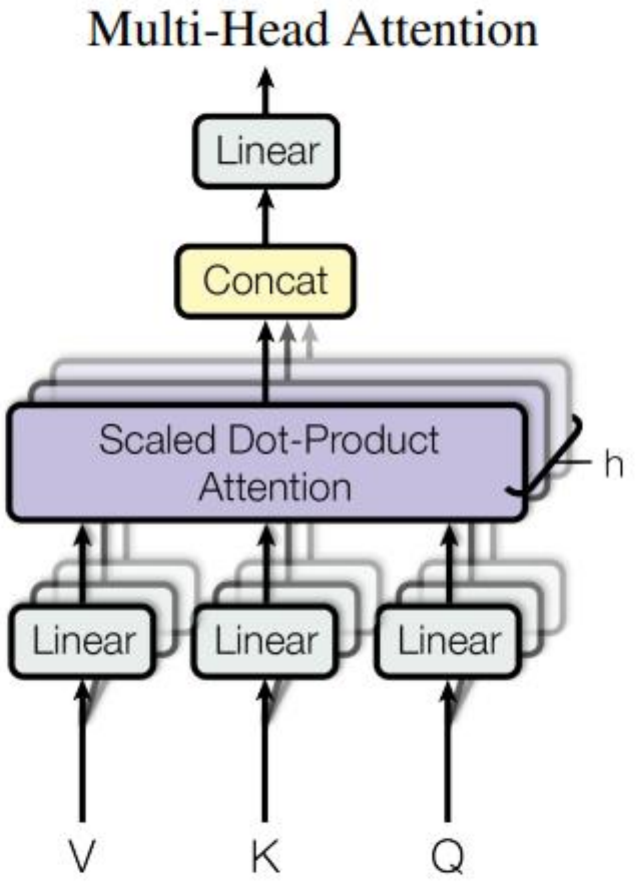


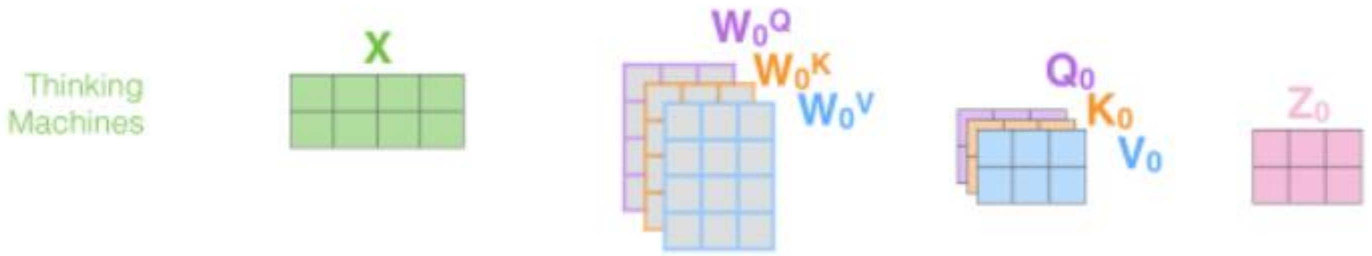

Linear
Concat
Scaled Dot-Product Attention
h
Linear
Linear
Linear
V
K
Q

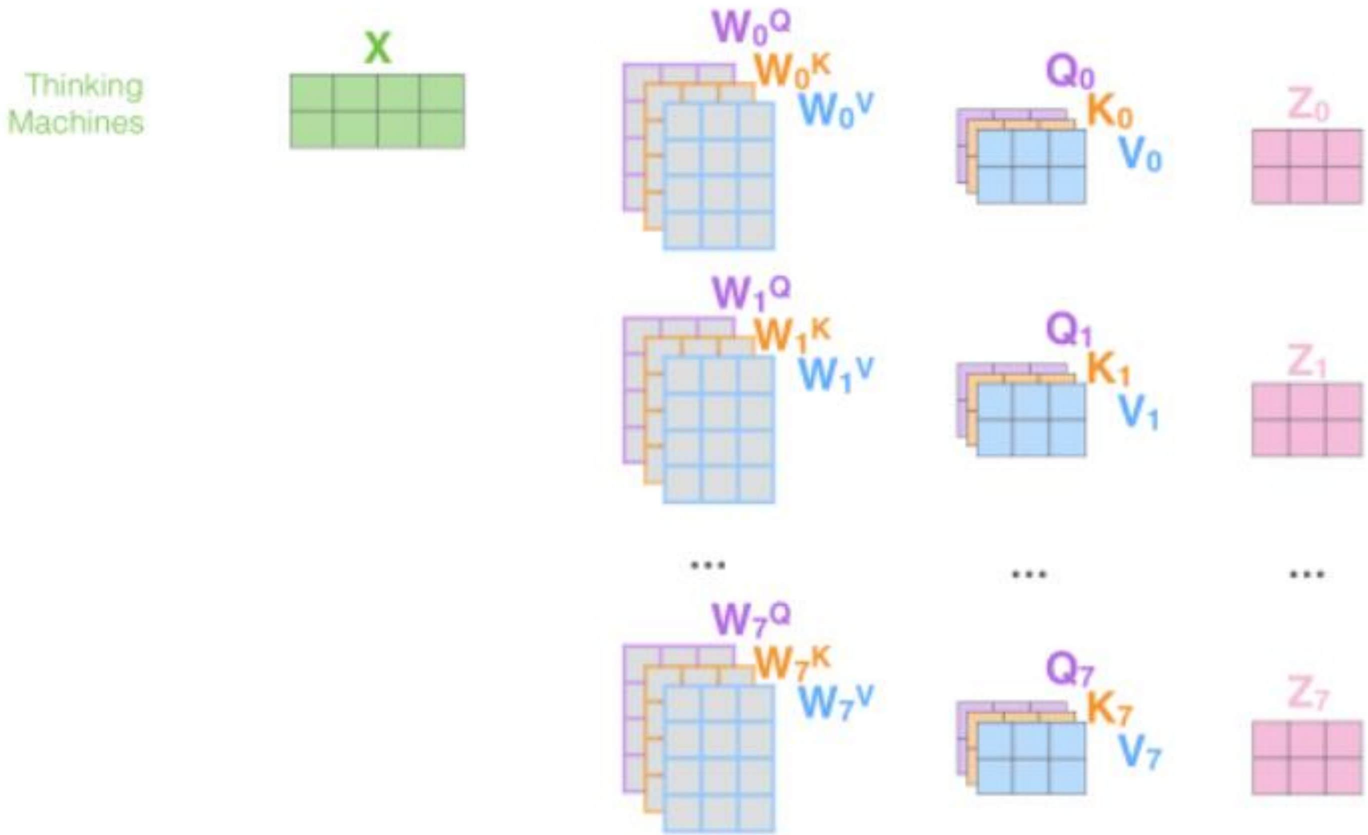
TERMINOLOGY

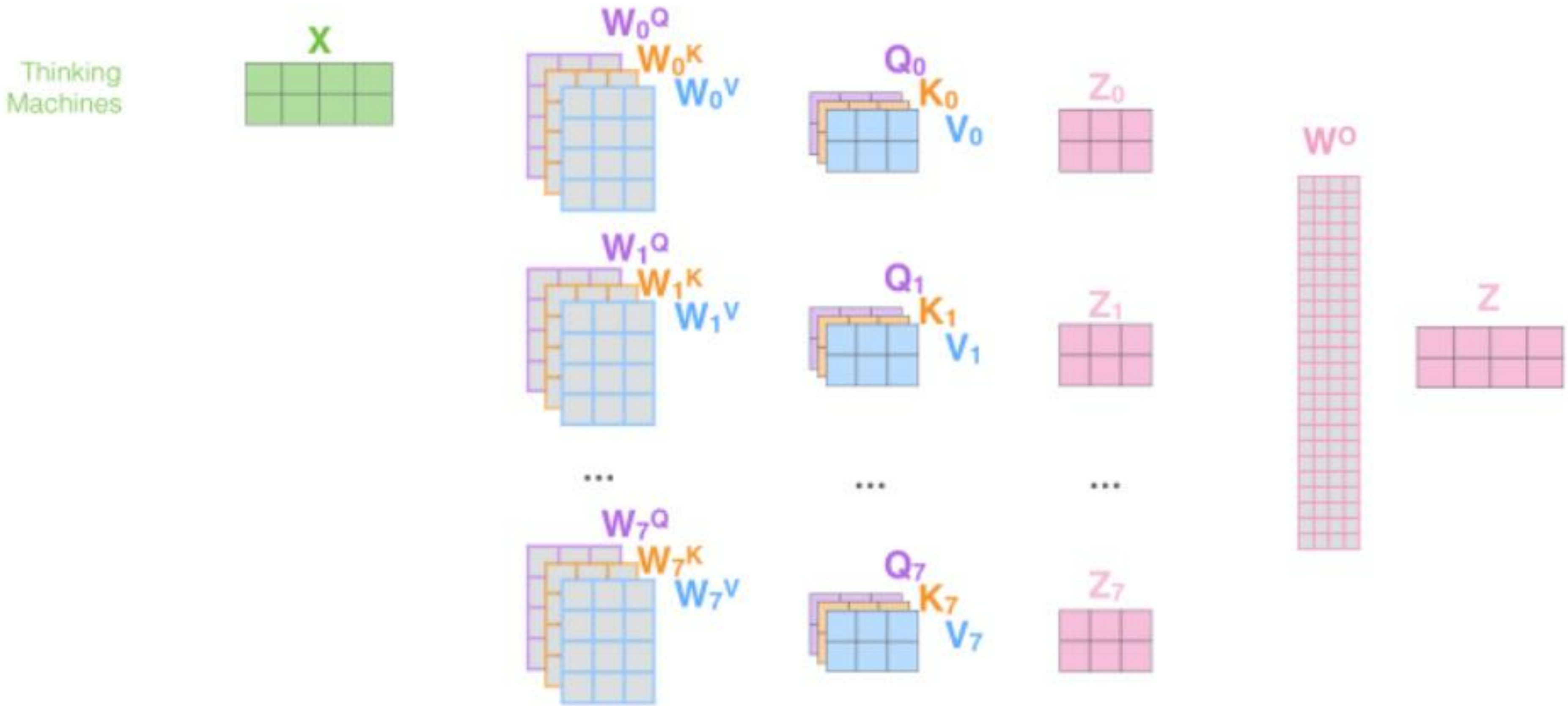


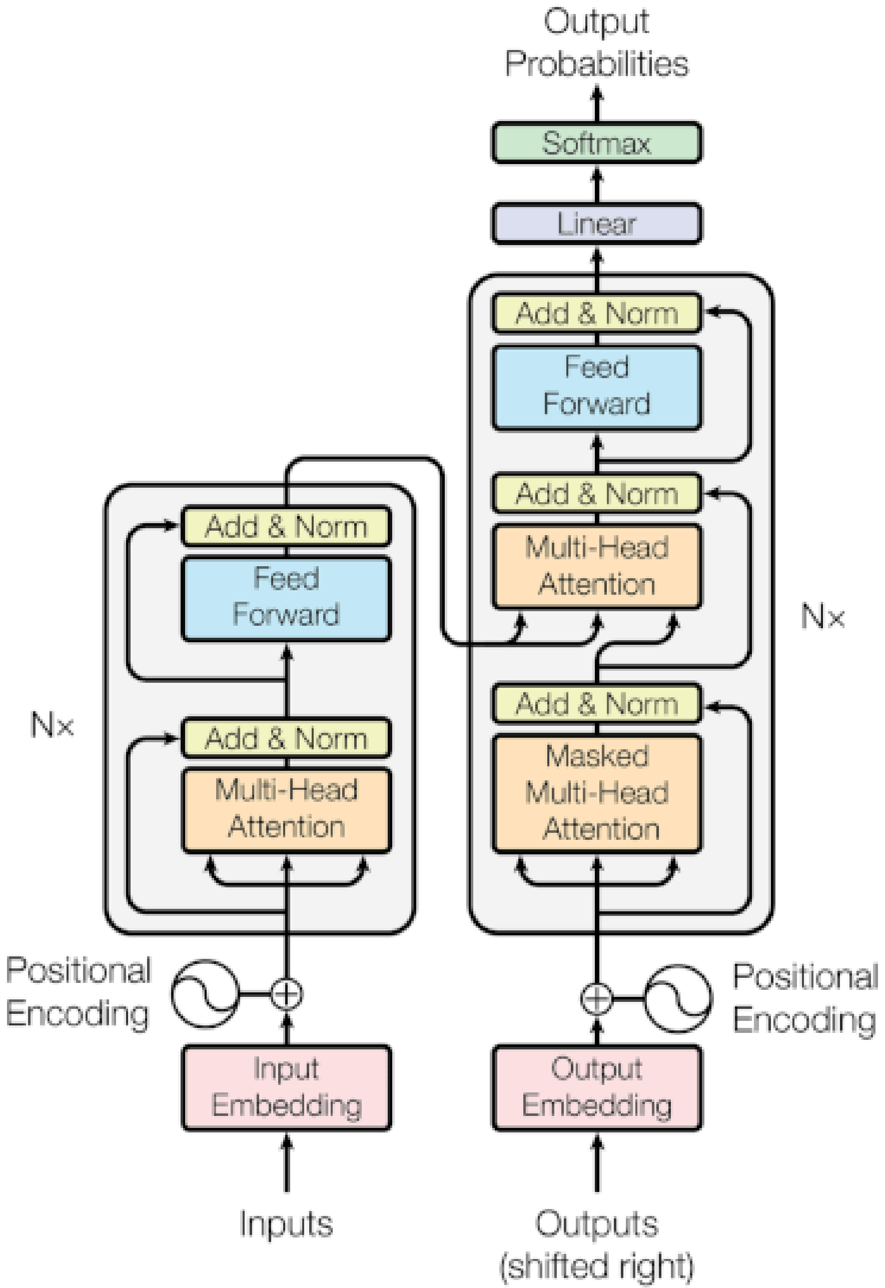
TERMINOLOGY

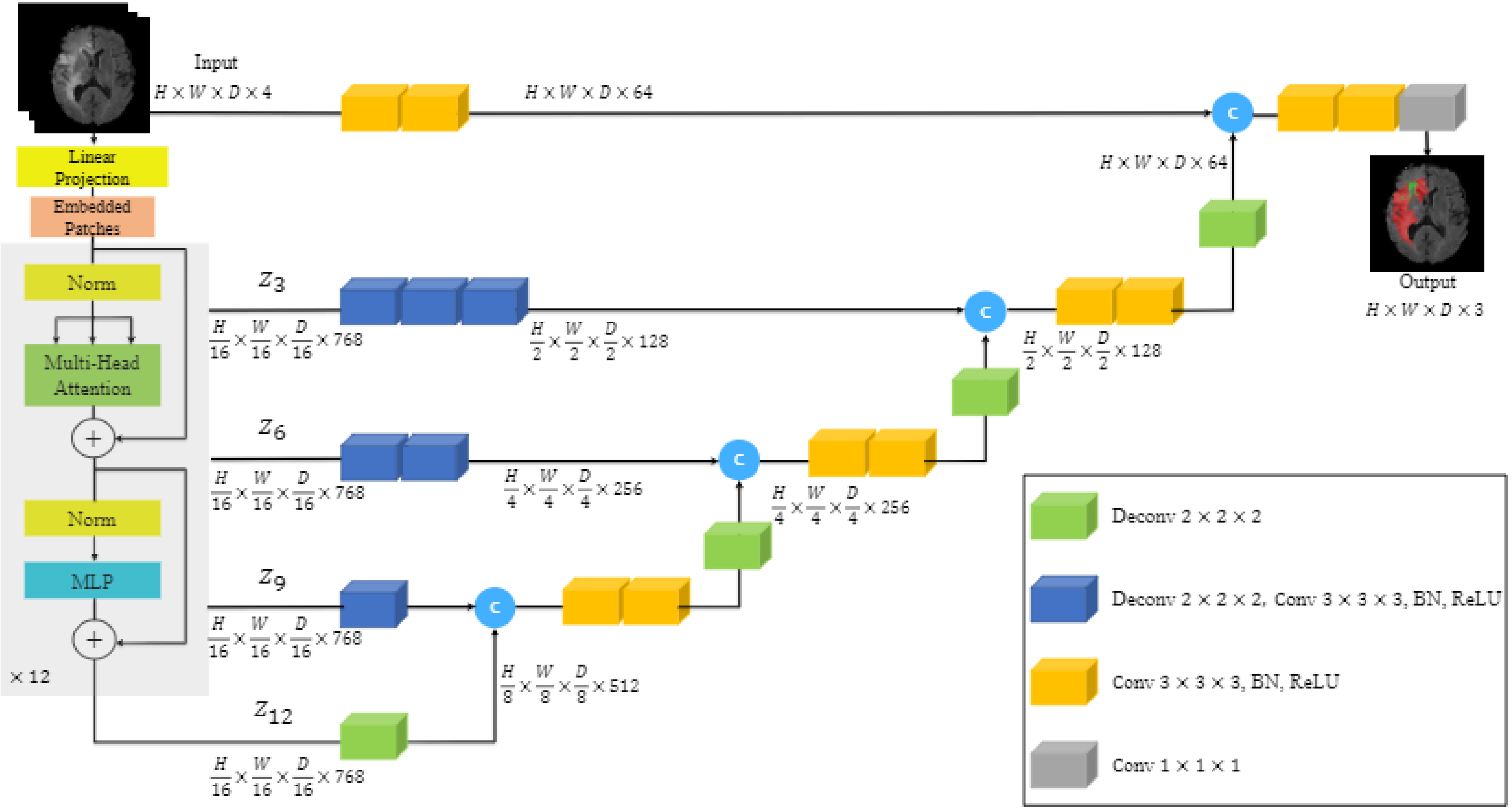



Thinking
Machines
X
W_0^Q
W_0^K
W_0^V
Q_0
K_0
V_0
Z_0


Thinking
Machines
X
W_0^Q
W_0^K
W_0^V
Q_0
K_0
V_0
Z_0
W_1^Q
W_1^K
W_1^V
Q_1
K_1
V_1
Z_1
...
...
...
W_7^Q
W_7^K
W_7^V
Q_7
K_7
V_7
Z_7


Thinking
Machines
X
W_0^Q
W_0^K
W_0^V
Q_0
K_0
V_0
Z_0
W_1^Q
W_1^K
W_1^V
Q_1
K_1
V_1
Z_1
...
...
...
W_7^Q
W_7^K
W_7^V
Q_7
K_7
V_7
Z_7
W^O
Z


Output
Probabilities
Softmax
Linear
Add & Norm
Feed
Forward
Add & Norm
Multi-Head
Attention
Nx
Add & Norm
Masked
Multi-Head
Attention
Add & Norm
Feed
Forward
Nx
Add & Norm
Multi-Head
Attention
Positional
Encoding
Positional
Encoding
Input
Embedding
Output
Embedding
Inputs
Outputs
(shifted right)


Input
$H \times W \times D \times 4$
Linear Projection
Embedded Patches
Norm
Multi-Head Attention
+
Norm
MLP
+
× 12
$H \times W \times D \times 64$
z_3
$\frac{H}{16} \times \frac{W}{16} \times \frac{D}{16} \times 768$
$\frac{H}{2} \times \frac{W}{2} \times \frac{D}{2} \times 128$
z_6
$\frac{H}{16} \times \frac{W}{16} \times \frac{D}{16} \times 768$
$\frac{H}{4} \times \frac{W}{4} \times \frac{D}{4} \times 256$
z_9
$\frac{H}{16} \times \frac{W}{16} \times \frac{D}{16} \times 768$
z_{12}
$\frac{H}{16} \times \frac{W}{16} \times \frac{D}{16} \times 768$
$\frac{H}{8} \times \frac{W}{8} \times \frac{D}{8} \times 512$
c
c
c
c
$\frac{H}{4} \times \frac{W}{4} \times \frac{D}{4} \times 256$
$\frac{H}{2} \times \frac{W}{2} \times \frac{D}{2} \times 128$
$H \times W \times D \times 64$
Output
$H \times W \times D \times 3$
Deconv 2 × 2 × 2
Deconv 2 × 2 × 2, Conv 3 × 3 × 3, BN, ReLU
Conv 3 × 3 × 3, BN, ReLU
Conv 1 × 1 × 1