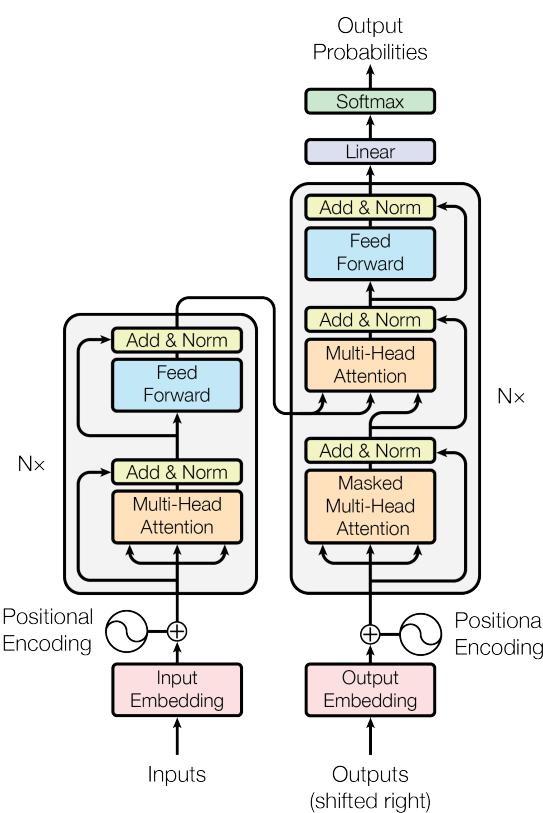


# Sequence Modeling: Transformers (see paper in Moodle)

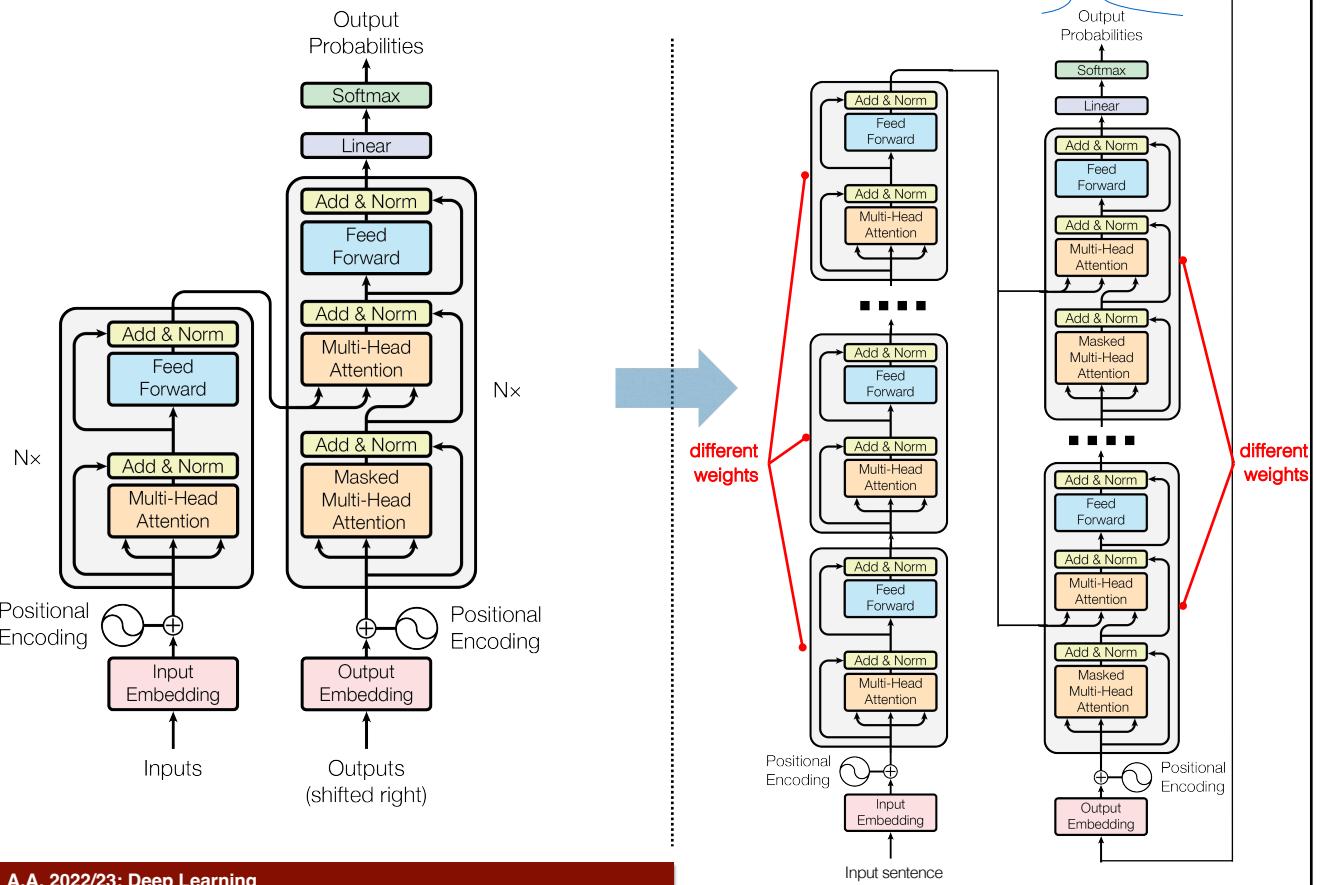
University of Padova, A.A. 2022/23

## Transformers



- For implementing sequence-to-sequence transductions involving not too long sequences
- Typically used for Machine Translation
- Based on self-attention
- Fast since many operations can be performed in parallel

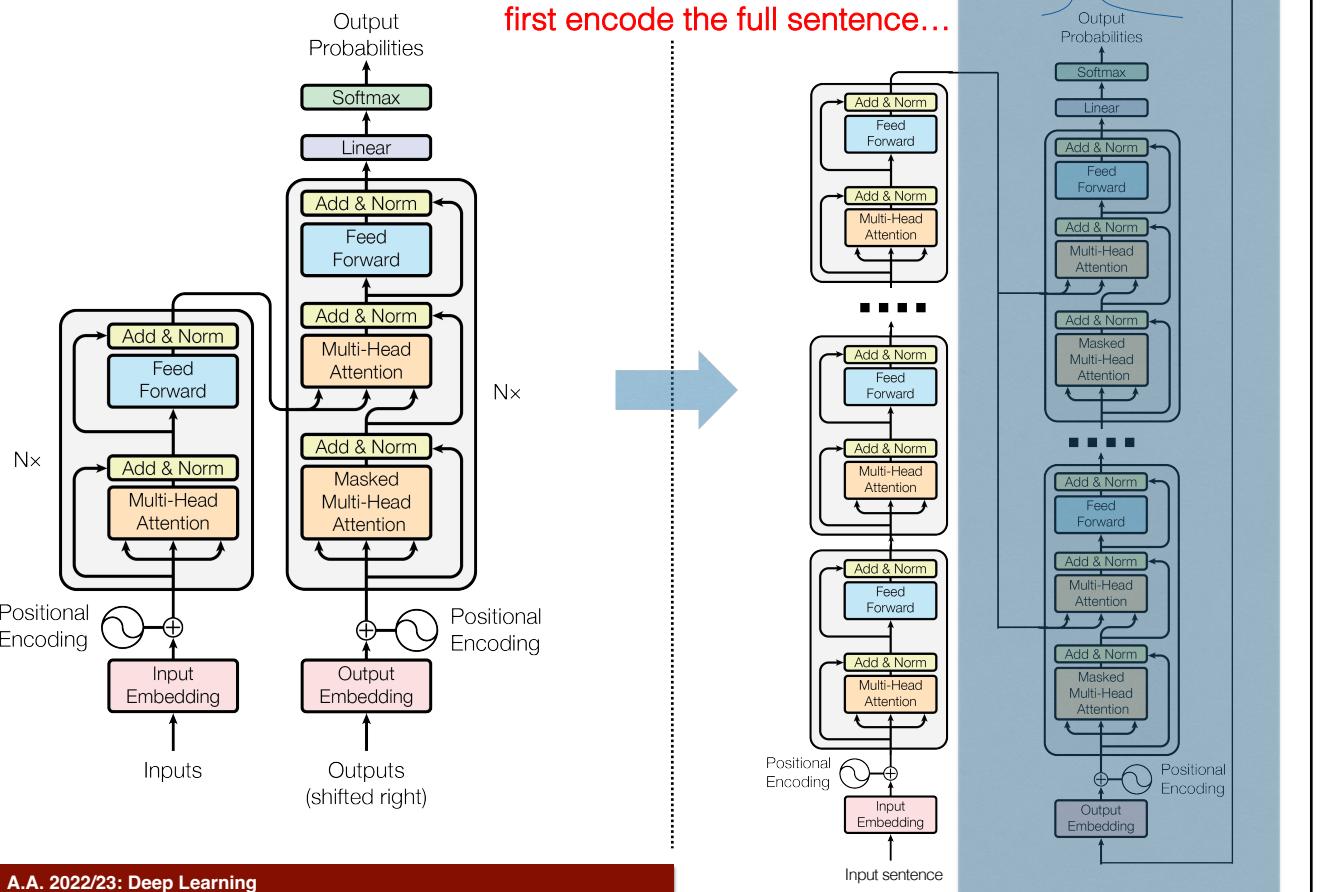
# Transformers



A.A. 2022/23: Deep Learning

# Transformers

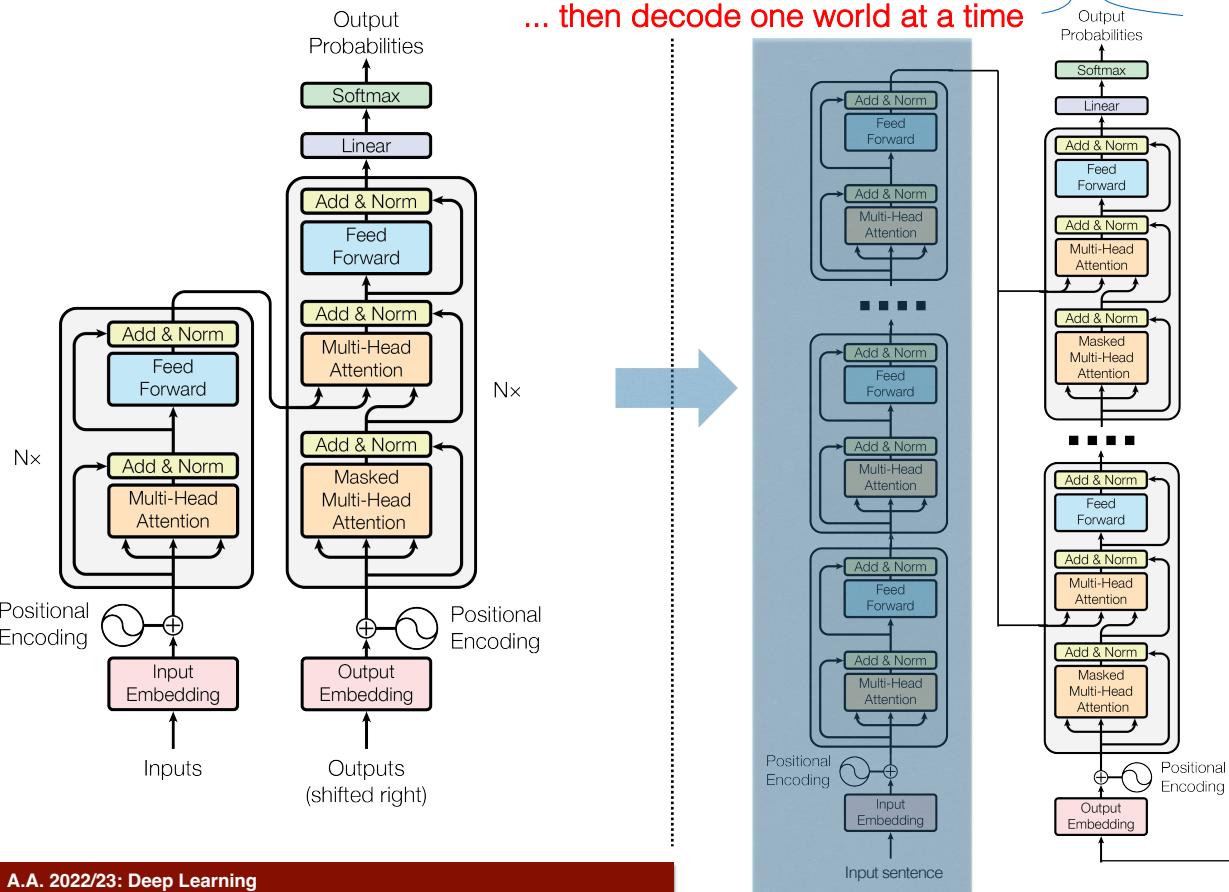
first encode the full sentence...



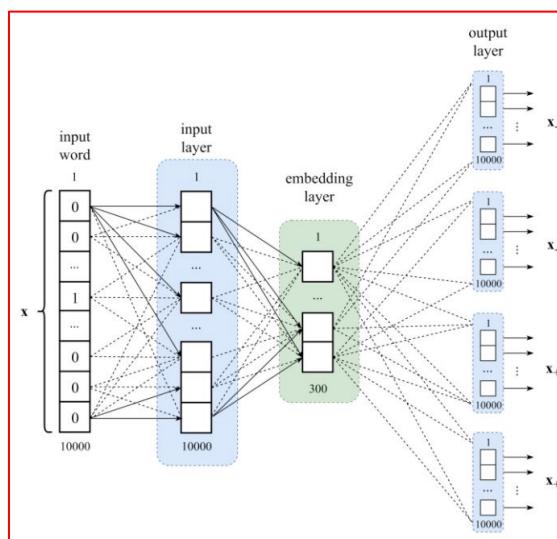
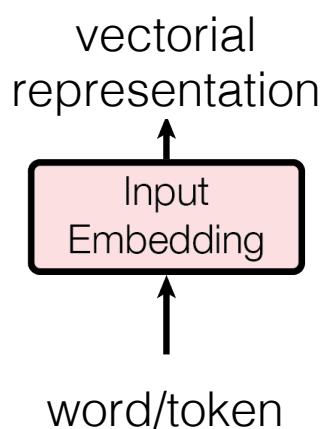
A.A. 2022/23: Deep Learning

# Transformers

... then decode one word at a time



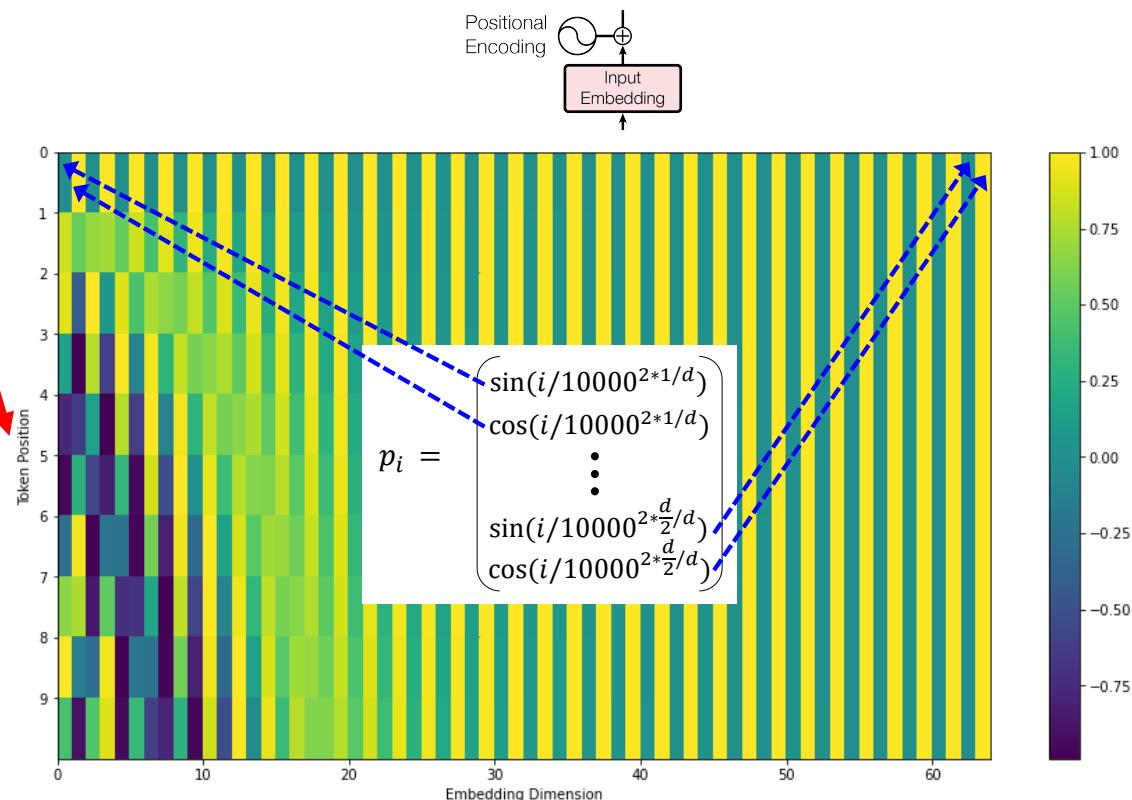
## Input Embedding



one way to generate embeddings of words:

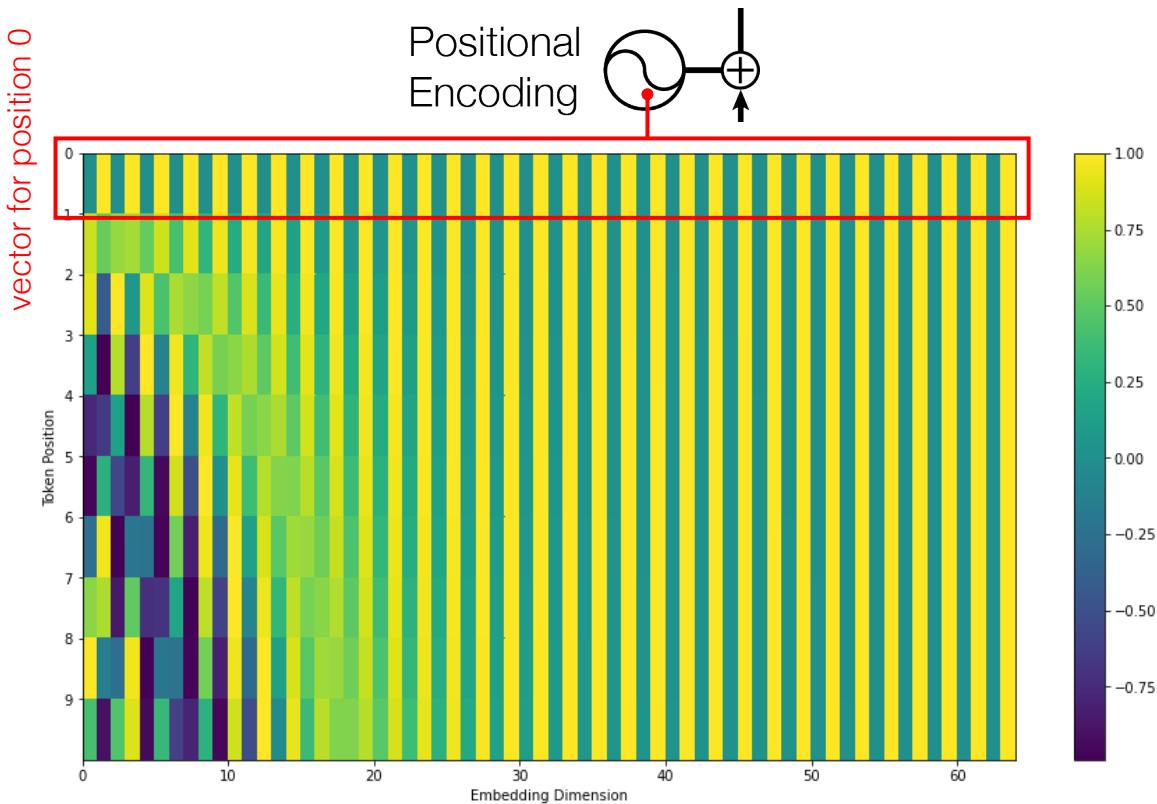
learn to map a word into the left and right words where it appears in textual documents  
*(training done separately, once for all)*

# Positional Encoding



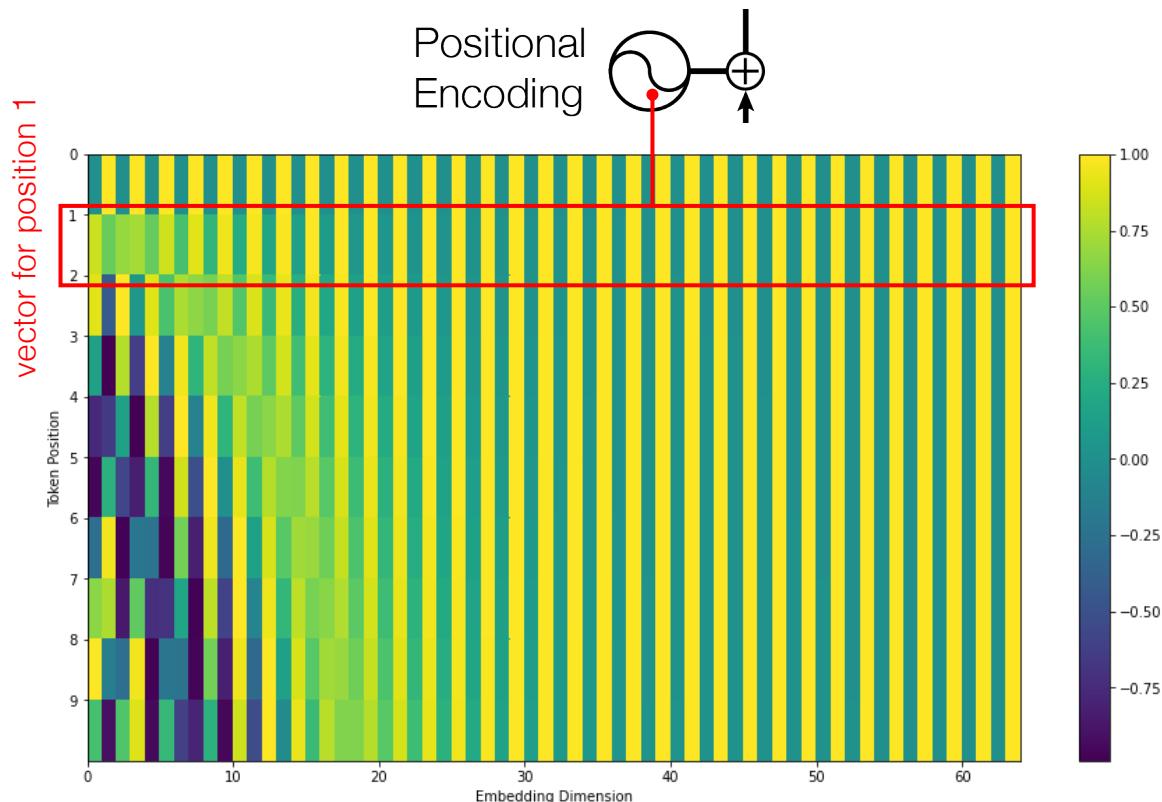
A.A. 2022/23: Deep Learning

# Positional Encoding

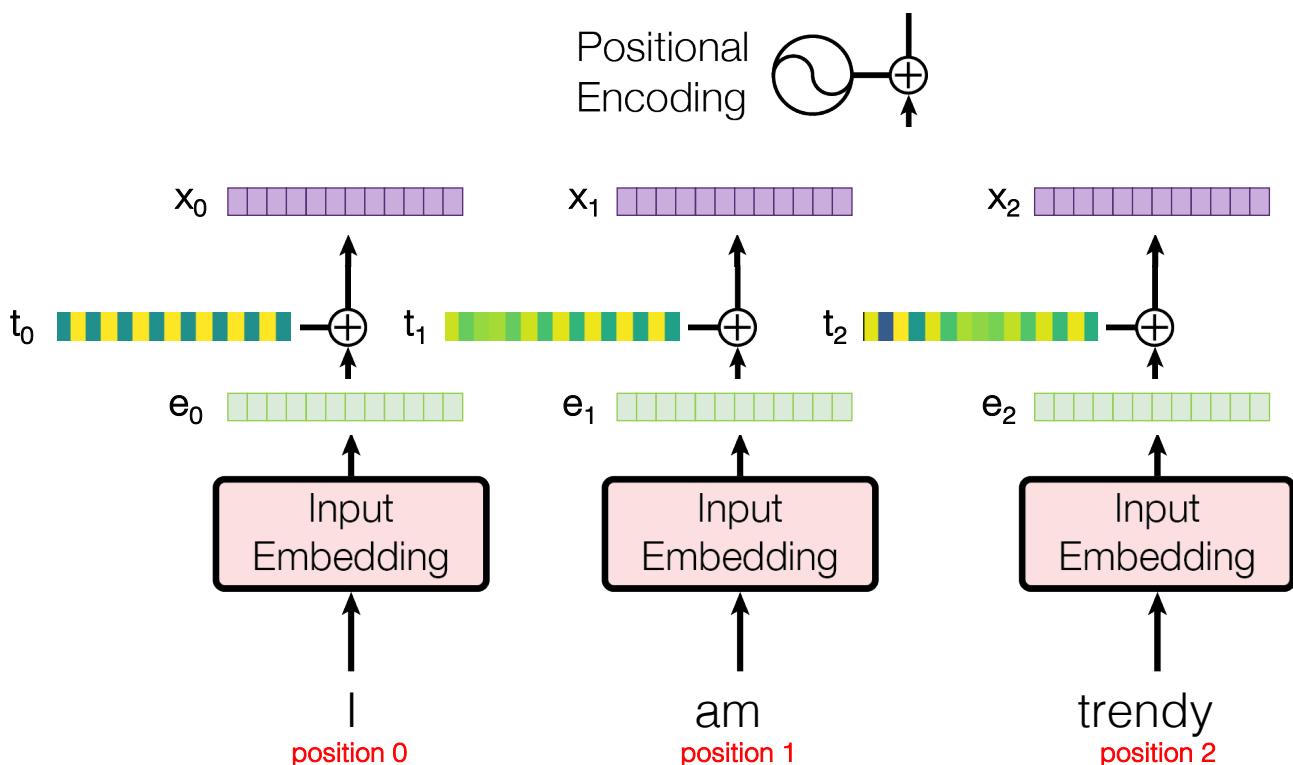


A.A. 2022/23: Deep Learning

# Positional Encoding

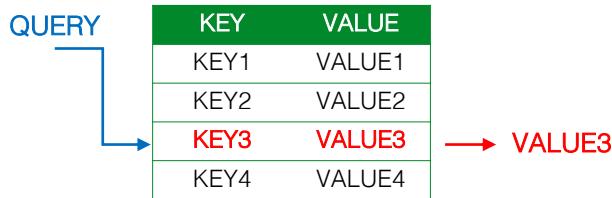


# Positional Encoding



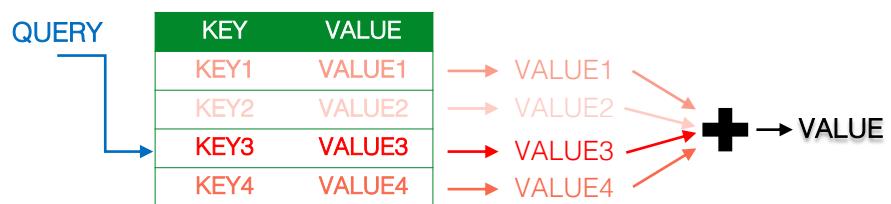
# Attention

## HARD HASHTABLE



Each **query** (hash) maps to exactly one **key-value** pair

## SOFT HASHTABLE

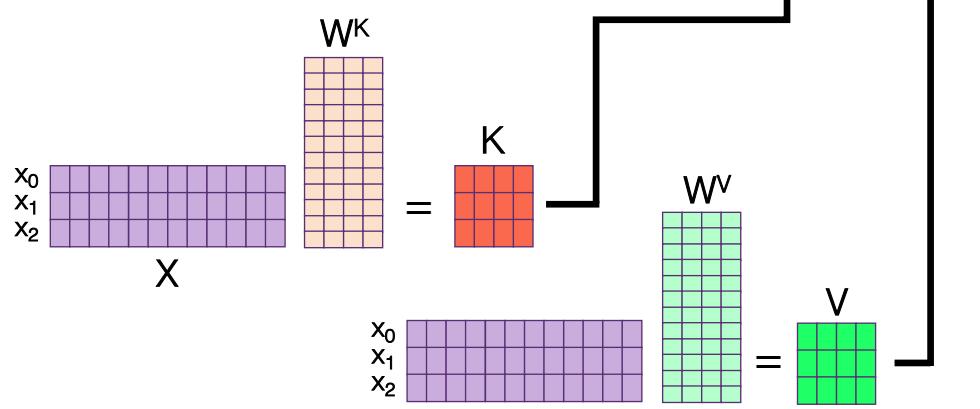


Each **query** matches each **key** to varying degrees  
Sum of **values** weighted by the **query-key** match is returned

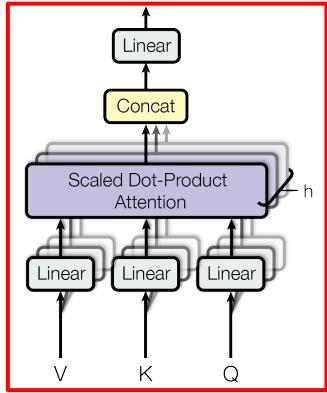
# Attention

- collect all input embeddings in a matrix
- rows can be processed in parallel

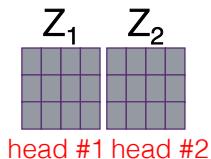
$$\begin{matrix} x_0 \\ x_1 \\ x_2 \end{matrix} \xrightarrow{\quad X \quad} W^Q = Q$$



global view

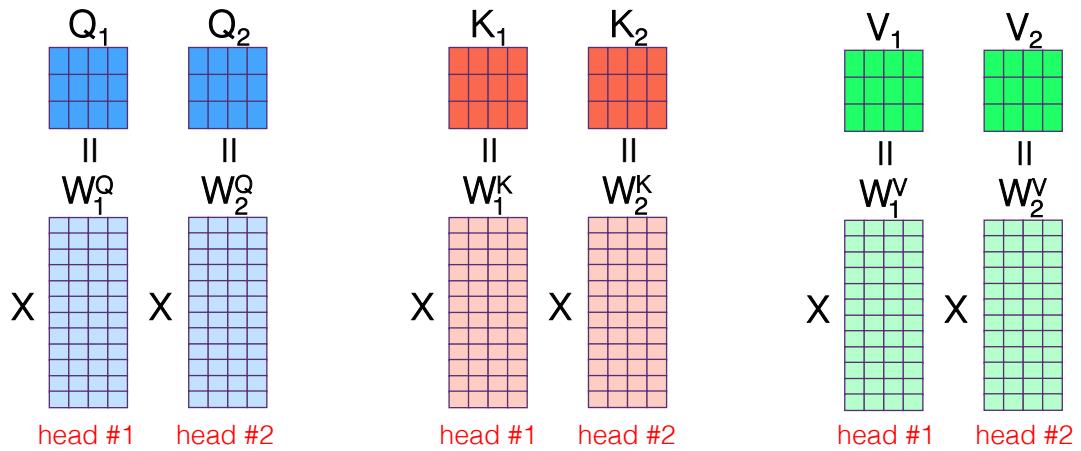


$W^O$



have more *heads*  
(sets of independent  
parameters), e.g. 2

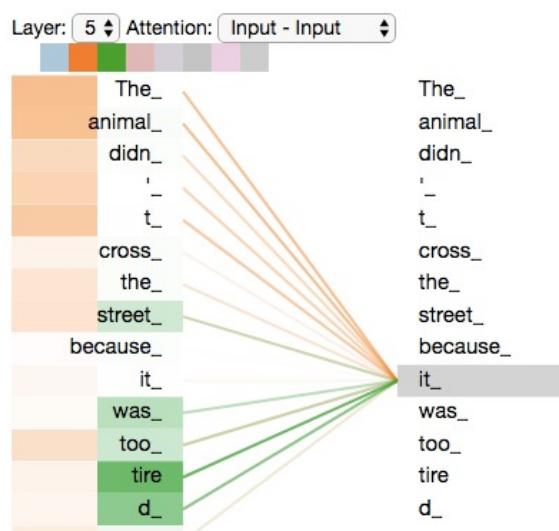
Multi-Head  
Attention



A.A. 2022/23: Deep Learning

## Multi-Head Attention Example

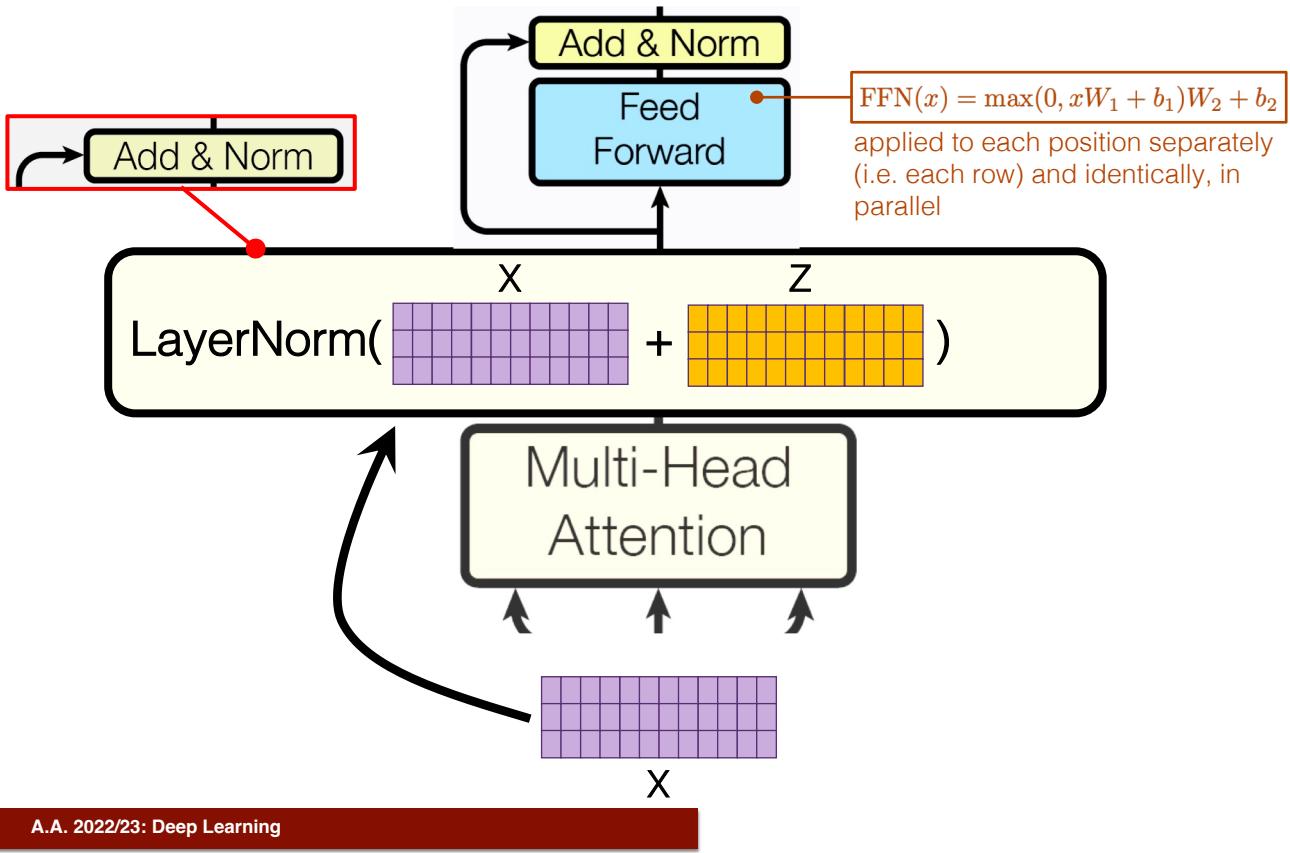
"The animal didn't cross the street because it was too tired"



- Encoding the word "it", one attention head is focusing most on "the animal", while another is focusing on "tired"
- Thus, the encoding of the word "it" embeds both the *concept* of "animal" and the *concept* of "tired"

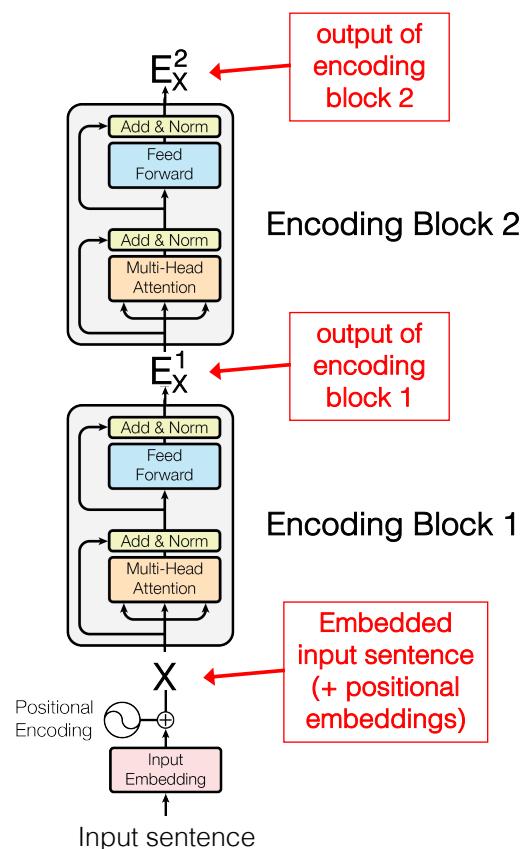
A.A. 2022/23: Deep Learning

# Add & Norm + Feed Forward Network

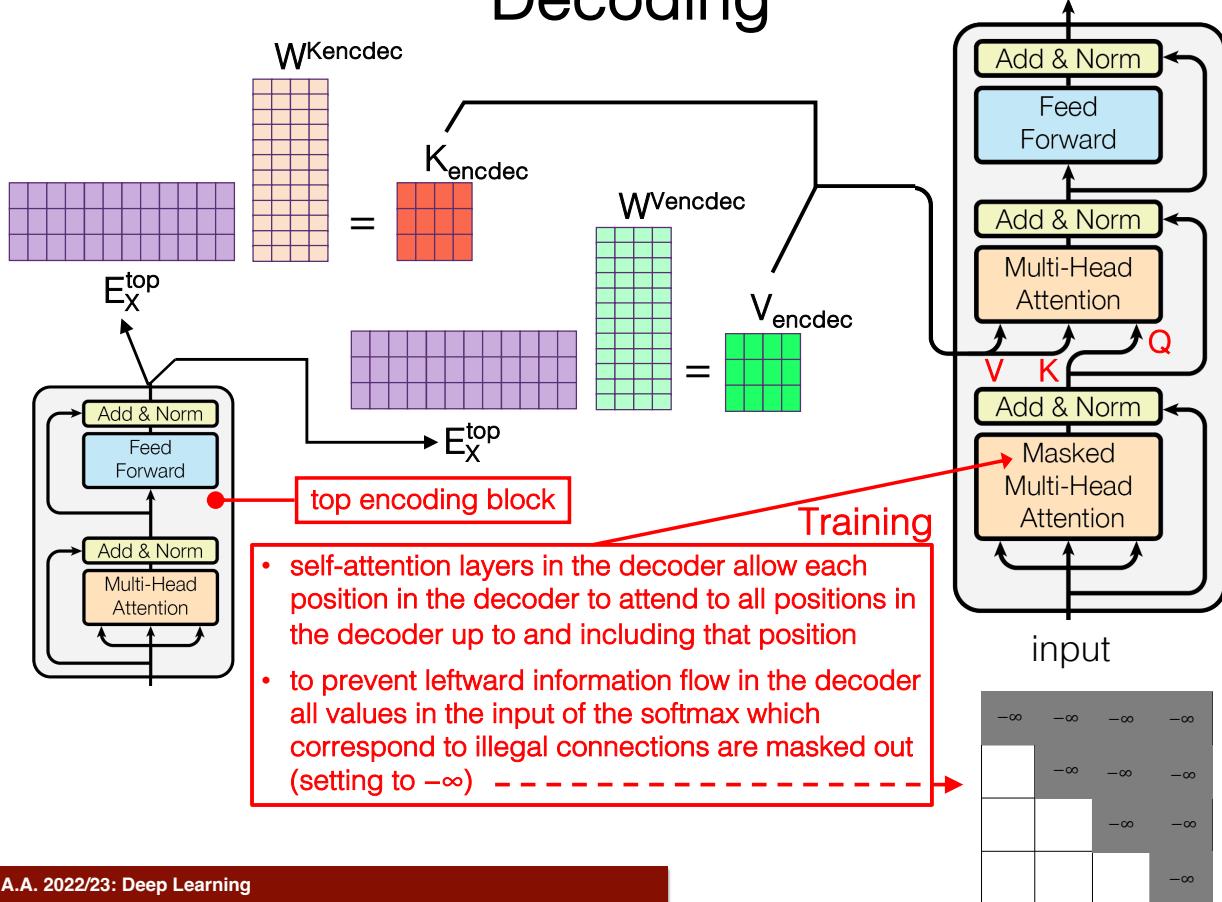


## Stacking Encoding Blocks

- Only the first encoding block is taking in input the embeddings of the words (+ positional embeddings), e.g.  $X$
- All the other encoding blocks use in place of  $X$  the output of the encoding block below, i.e. encoding block  $j+1$  takes in input  $E_X^j$



# Decoding



A.A. 2022/23: Deep Learning

# Training

## Training Set

("I am trendy", "Io sono alla moda")

("I take a shower", "Faccio la doccia")

## Target Model Output

Output Vocabulary: alla sono io doccia moda <eos>

position #1	0.0	0.0	1.0	0.0	0.0	0.0
position #2	0.0	1.0	0.0	0.0	0.0	0.0
position #3	1.0	0.0	0.0	0.0	0.0	0.0
position #4	0.0	0.0	0.0	0.0	1.0	0.0
position #5	0.0	0.0	0.0	0.0	0.0	1.0

## Model Outputs

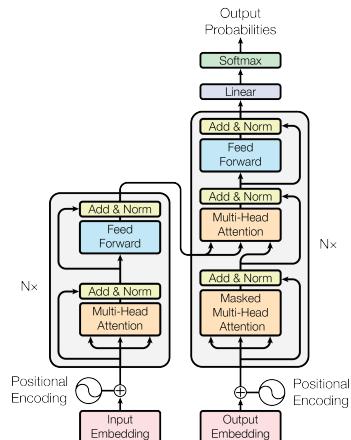
Output Vocabulary: alla sono io doccia moda <eos>

position #1	0.01	0.02	0.93	0.01	0.03	0.01
position #2	0.01	0.8	0.1	0.05	0.01	0.03
position #3	0.99	0.001	0.001	0.001	0.002	0.001
position #4	0.001	0.002	0.001	0.02	0.94	0.01
position #5	0.01	0.01	0.001	0.001	0.001	0.98



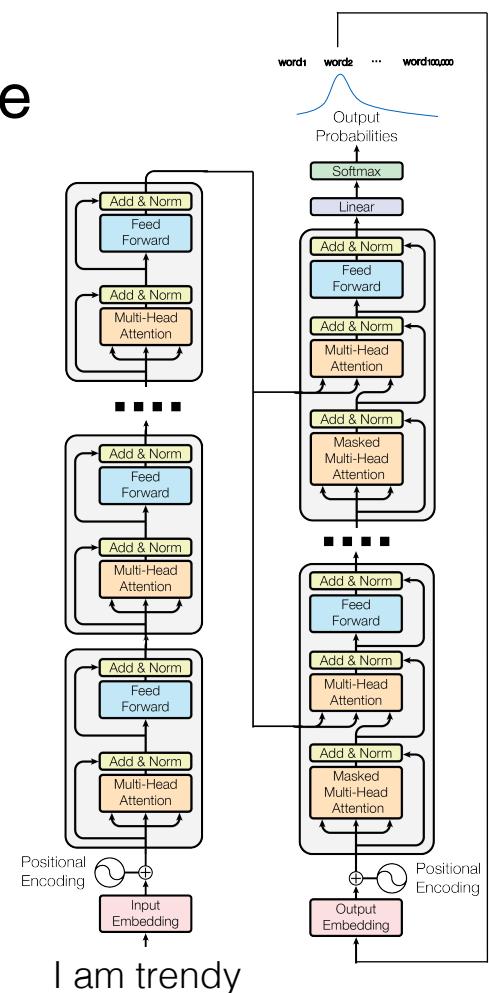
A.A. 2022/23: Deep Learning

# Generation: the big picture



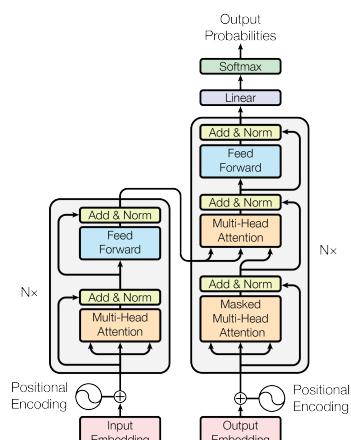
computation

A.A. 2022/23: Deep Learning



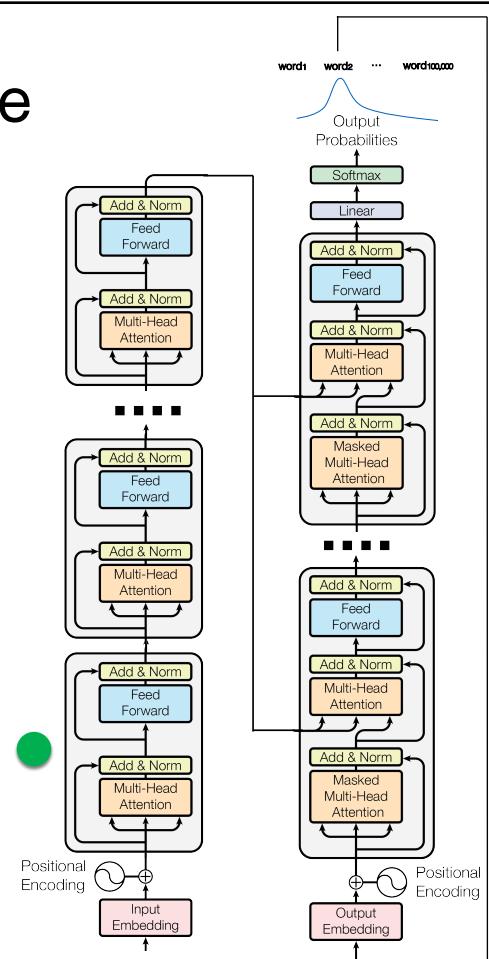
I am trendy

# Generation: the big picture



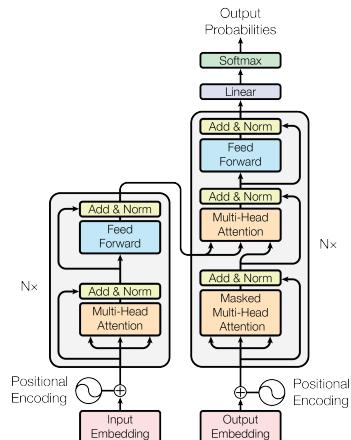
computation

A.A. 2022/23: Deep Learning

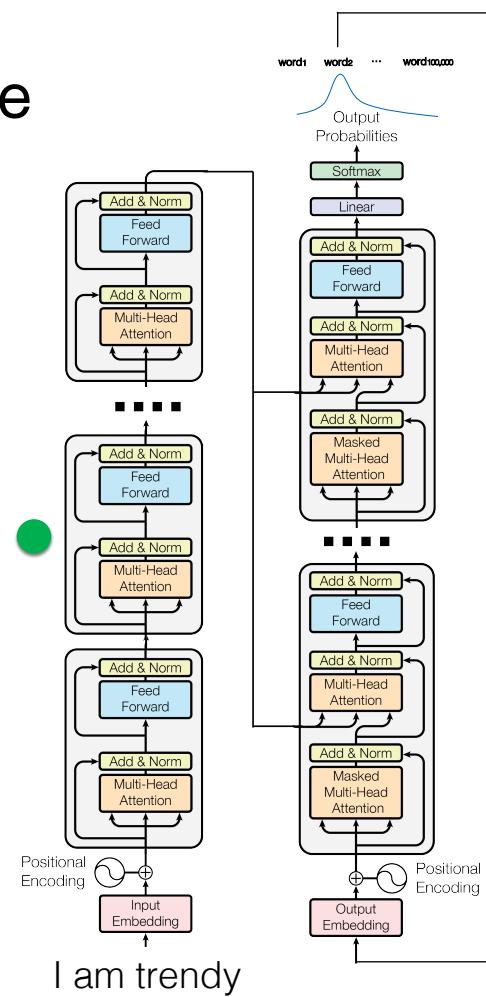


I am trendy

# Generation: the big picture



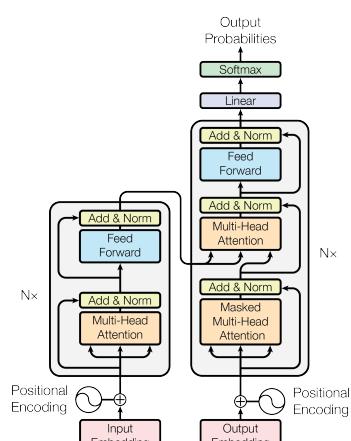
computation



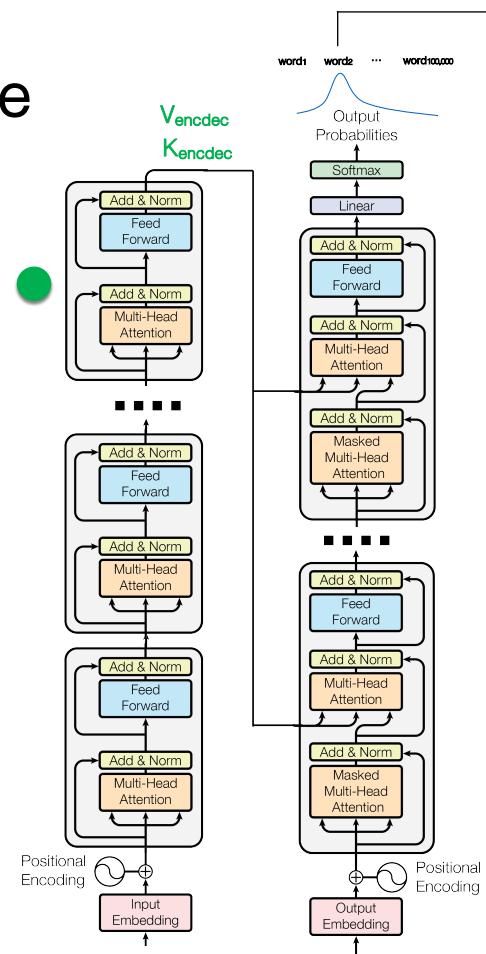
I am trendy

A.A. 2022/23: Deep Learning

# Generation: the big picture



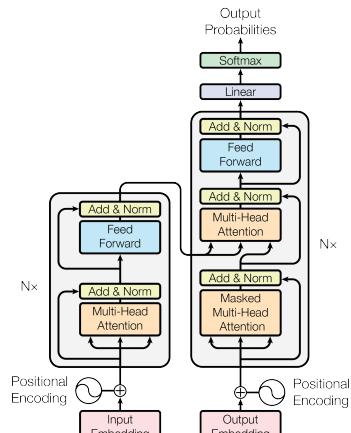
computation



I am trendy

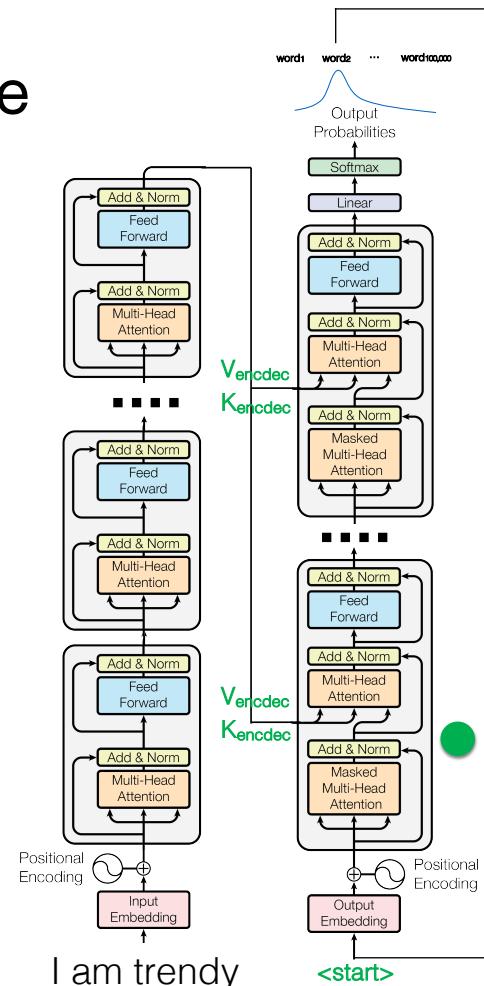
A.A. 2022/23: Deep Learning

# Generation: the big picture



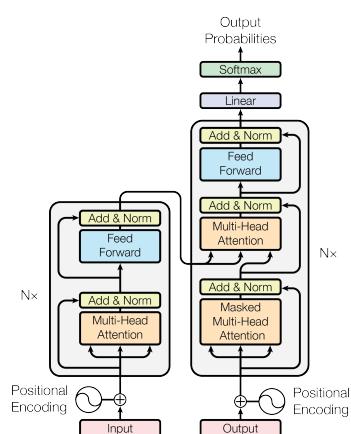
computation

A.A. 2022/23: Deep Learning



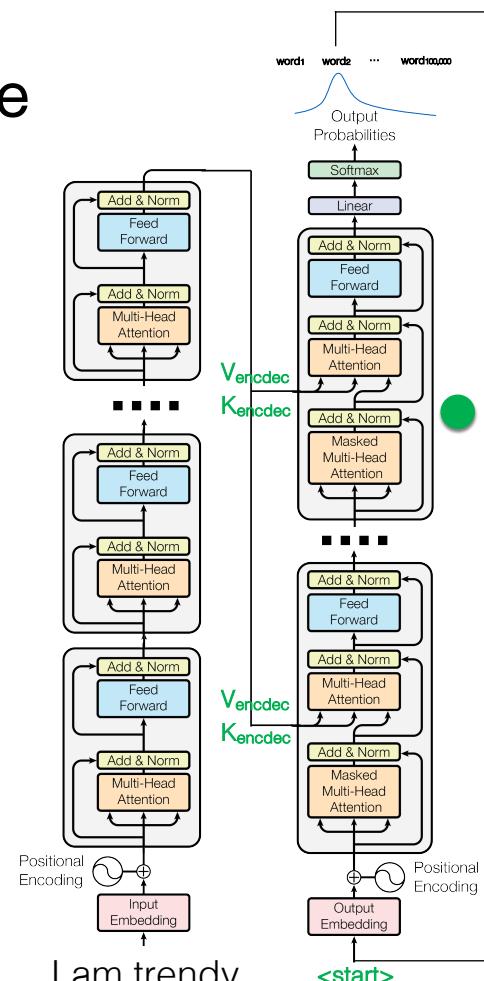
I am trendy

# Generation: the big picture



computation

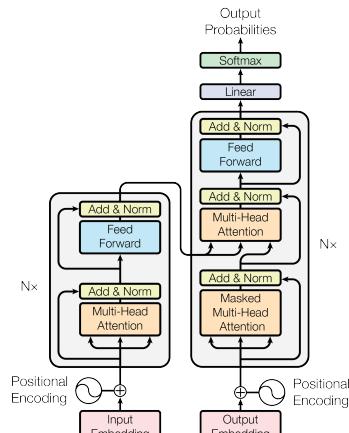
A.A. 2022/23: Deep Learning



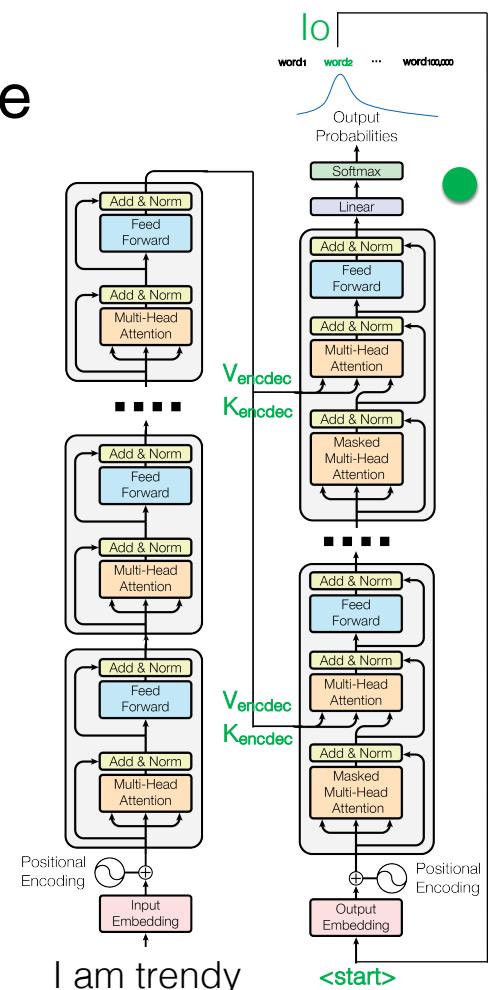
I am trendy

<start>

# Generation: the big picture



computation

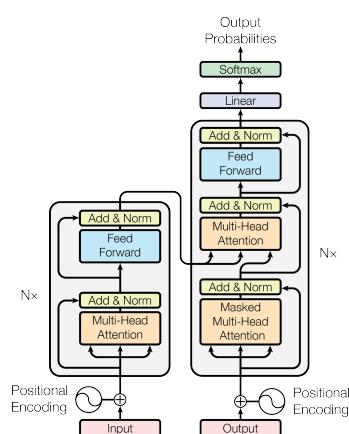


A.A. 2022/23: Deep Learning

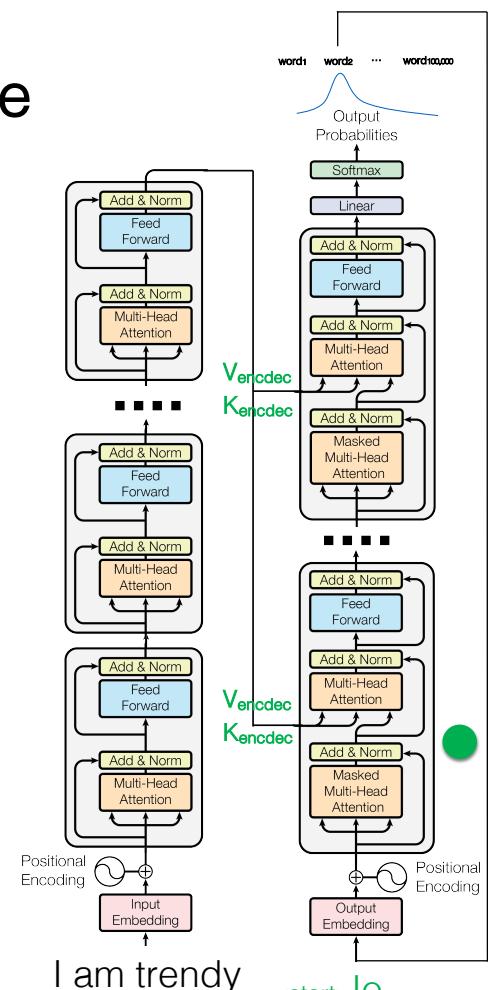
I am trendy

<start>

# Generation: the big picture



computation

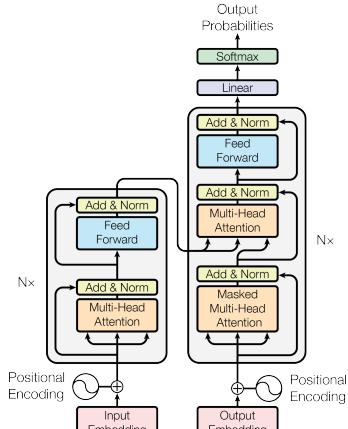


A.A. 2022/23: Deep Learning

I am trendy

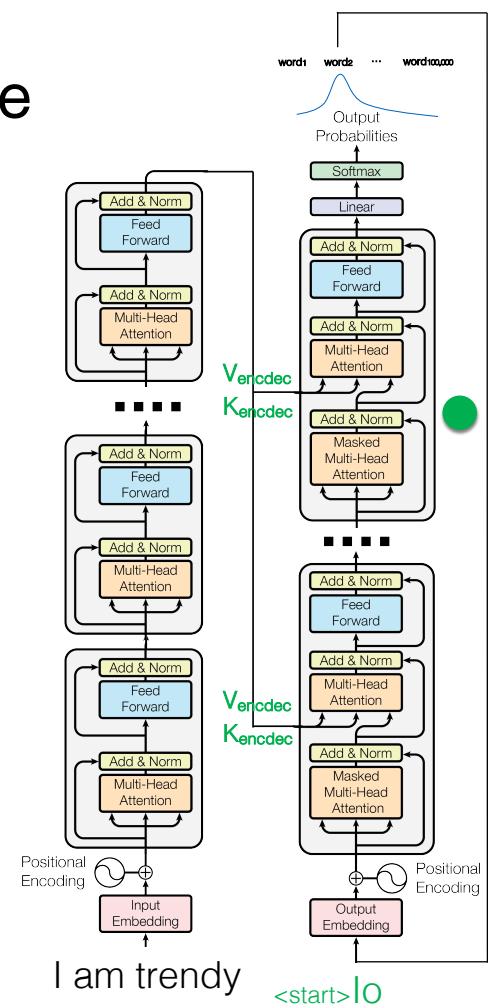
<start>IO

# Generation: the big picture



computation

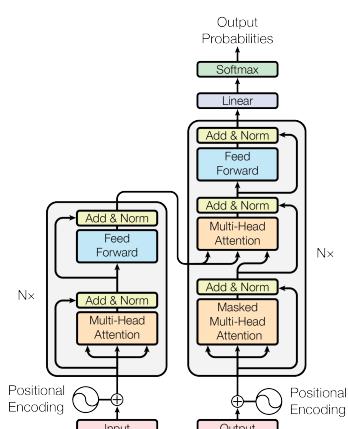
A.A. 2022/23: Deep Learning



I am trendy

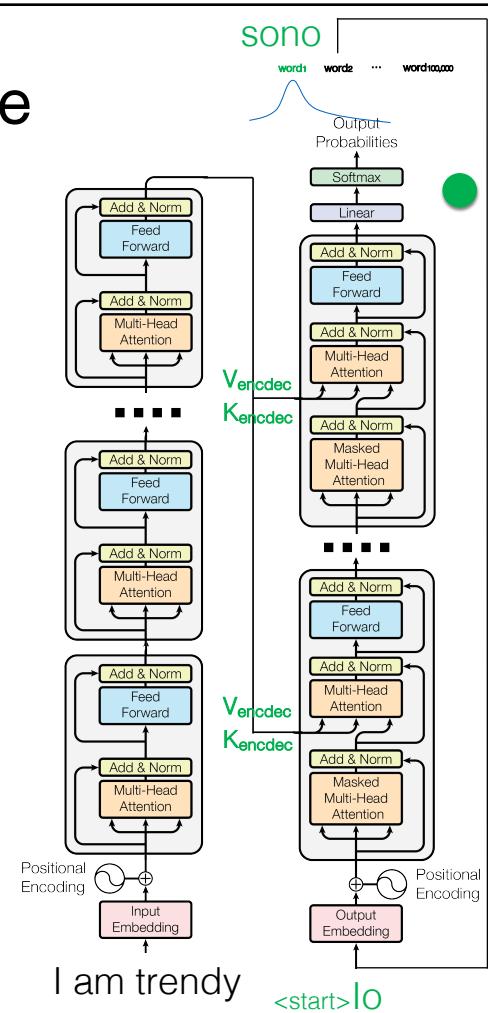
<start> IO

# Generation: the big picture



computation

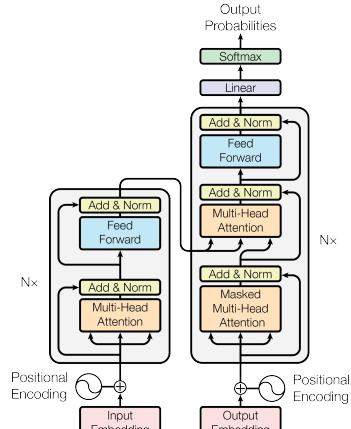
A.A. 2022/23: Deep Learning



I am trendy

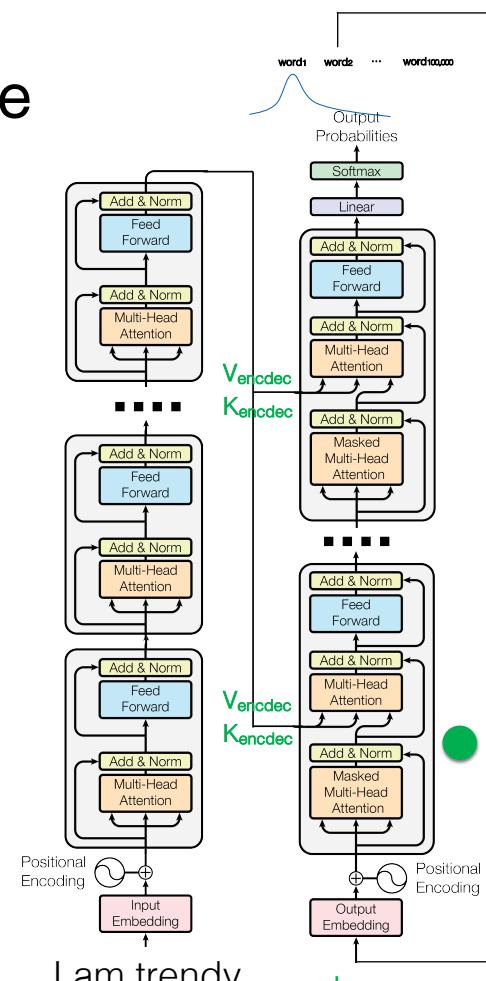
<start> IO

# Generation: the big picture



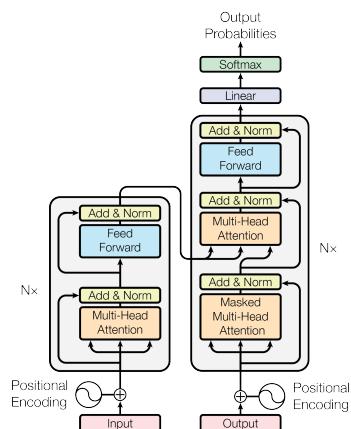
computation

A.A. 2022/23: Deep Learning



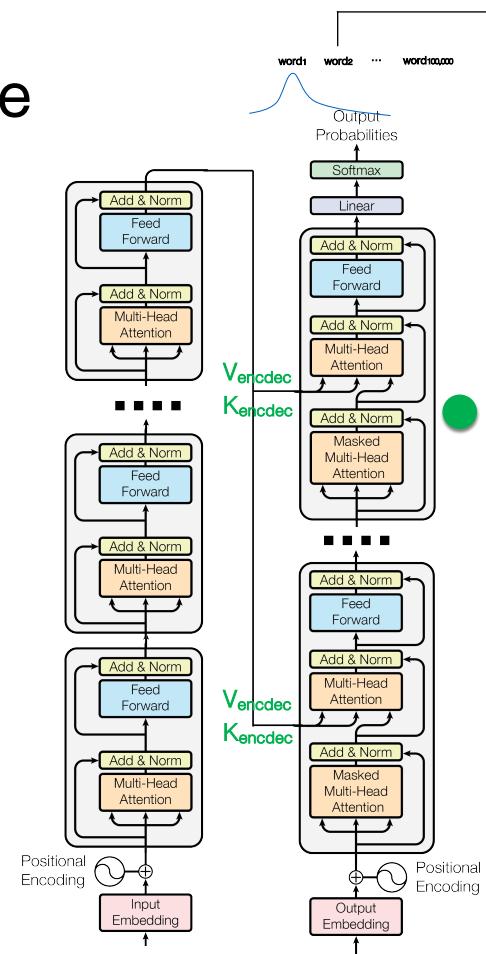
I am trendy   
<start>lo sono

# Generation: the big picture



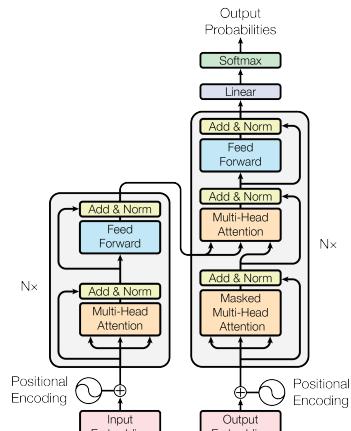
computation

A.A. 2022/23: Deep Learning



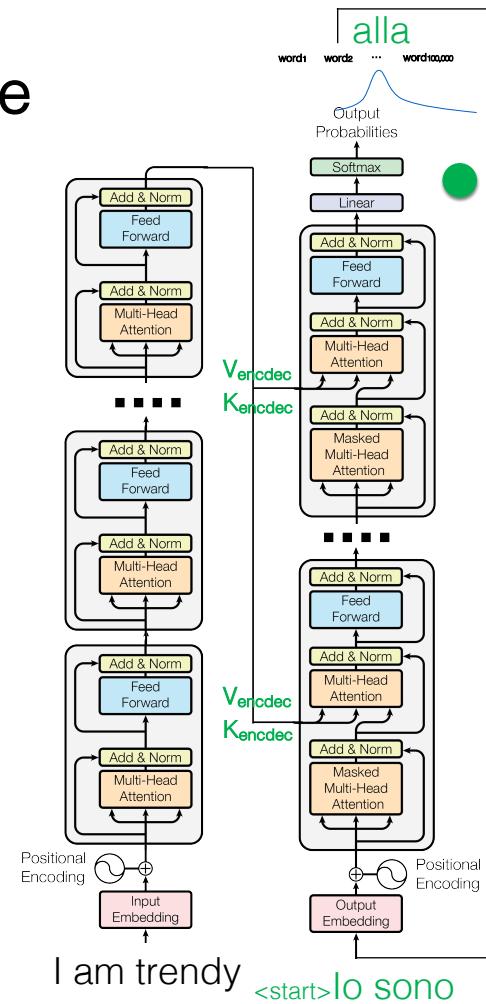
I am trendy   
<start>lo sono

# Generation: the big picture



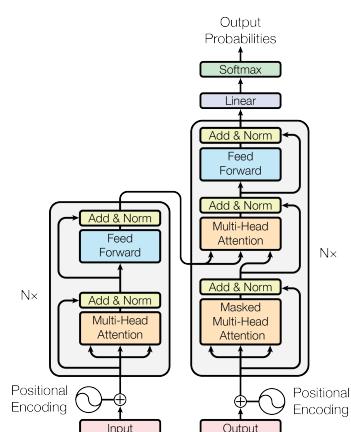
computation

A.A. 2022/23: Deep Learning



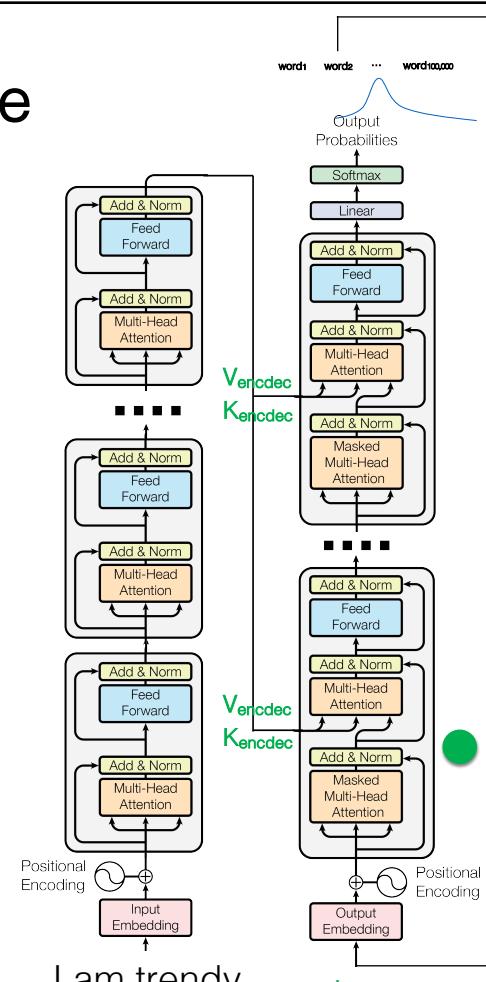
I am trendy **<start>**lo sono

# Generation: the big picture



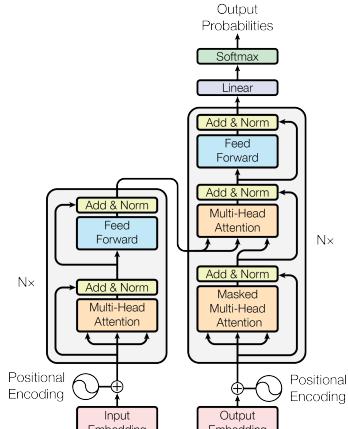
computation

A.A. 2022/23: Deep Learning



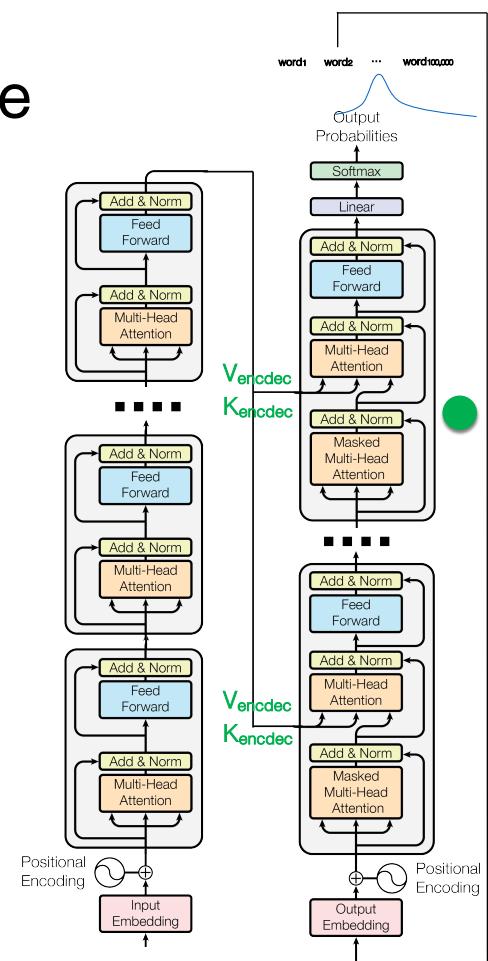
I am trendy **<start>**lo sono alla

# Generation: the big picture



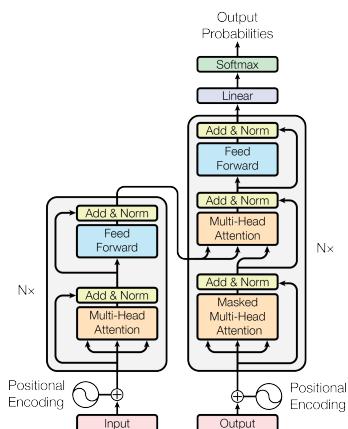
computation

A.A. 2022/23: Deep Learning



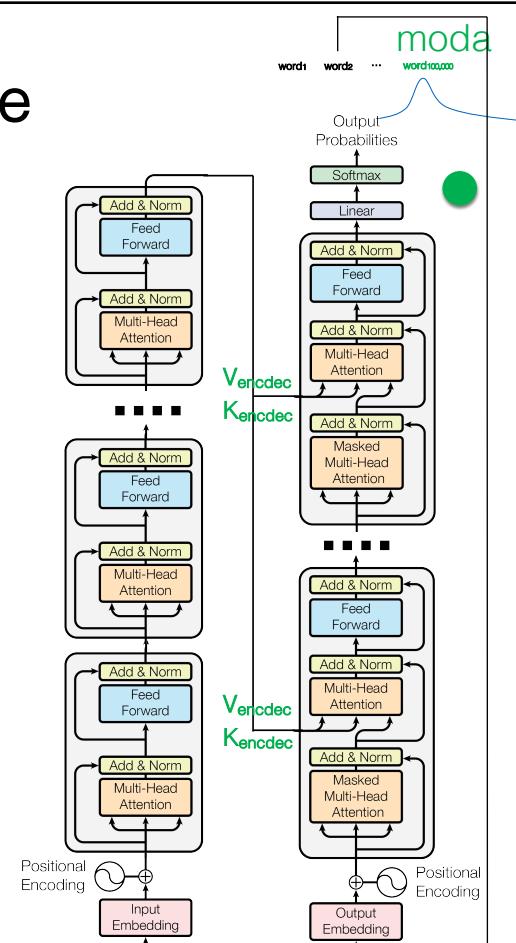
I am trendy **<start>**lo sono alla

# Generation: the big picture



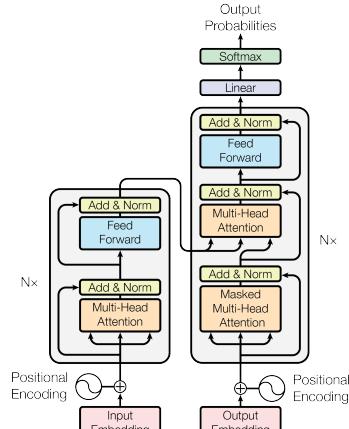
computation

A.A. 2022/23: Deep Learning

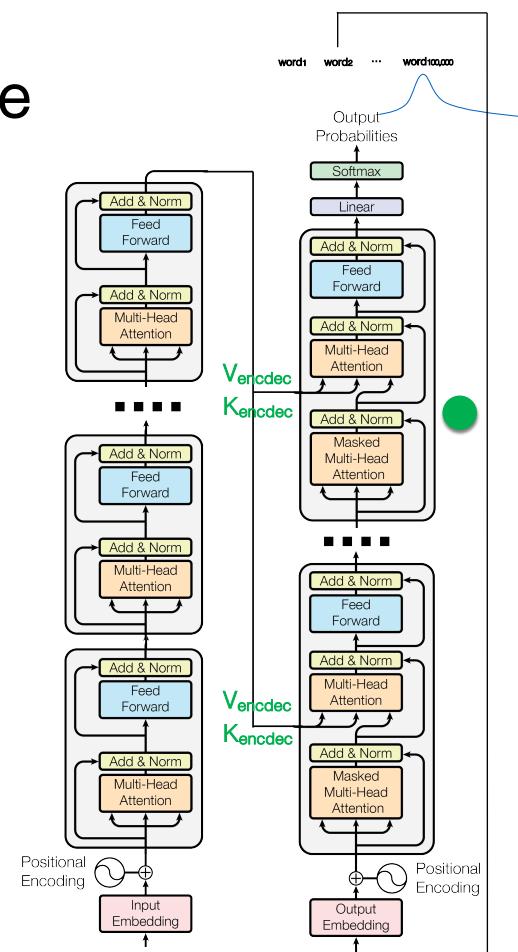


I am trendy **<start>**lo sono alla

# Generation: the big picture



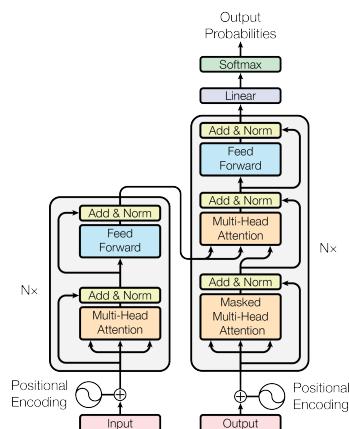
computation



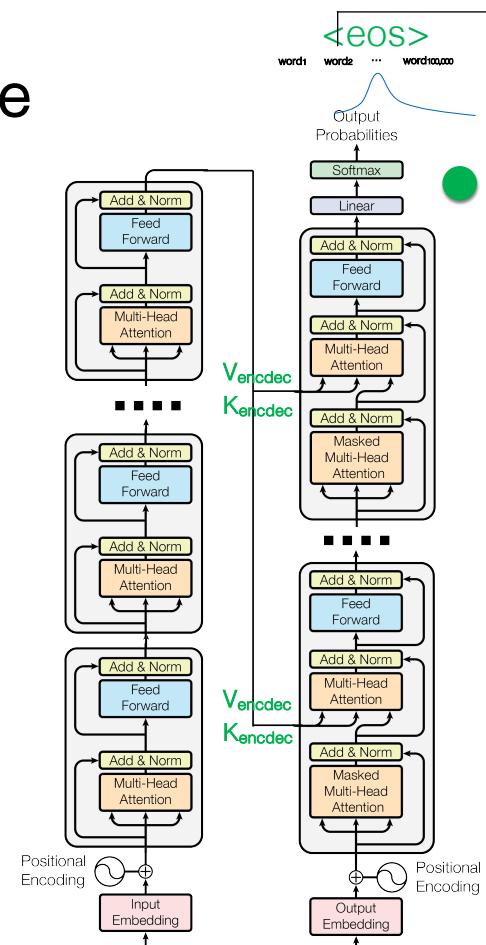
I am trendy  
<start> lo sono alla moda

A.A. 2022/23: Deep Learning

# Generation: the big picture



computation



I am trendy  
<start> lo sono alla moda

A.A. 2022/23: Deep Learning