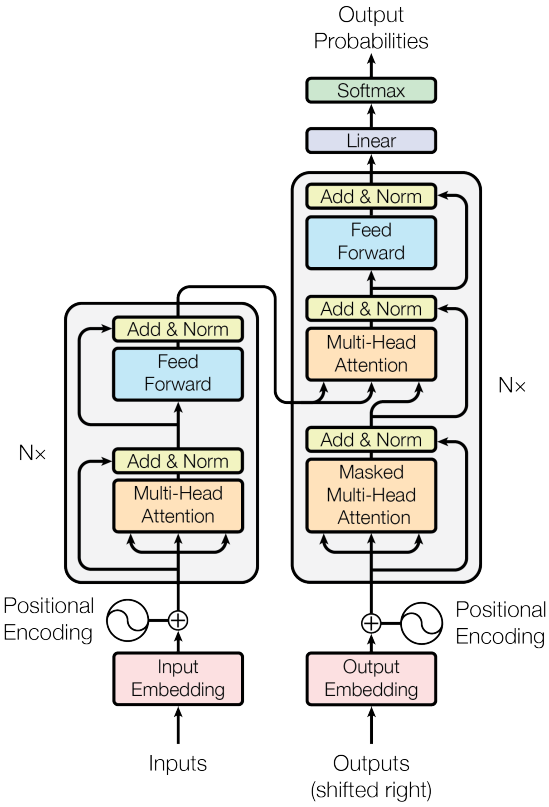
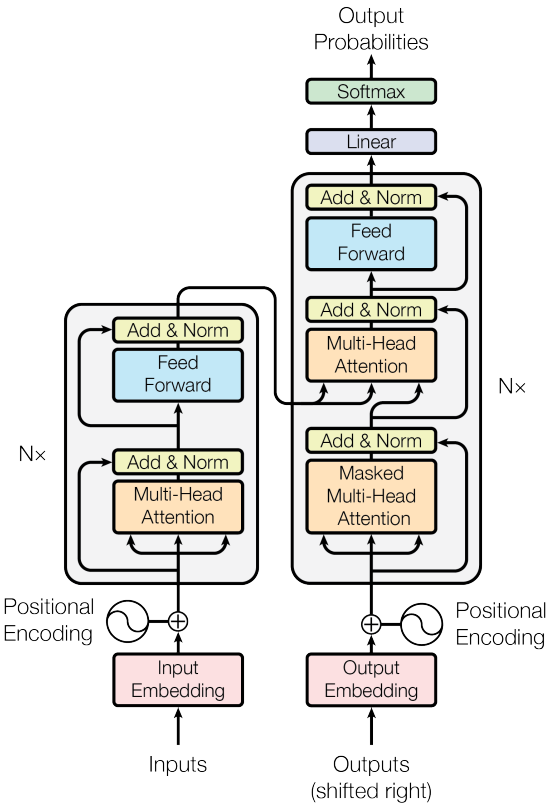
Sequence Modeling: Transformers (see paper in Moodle)

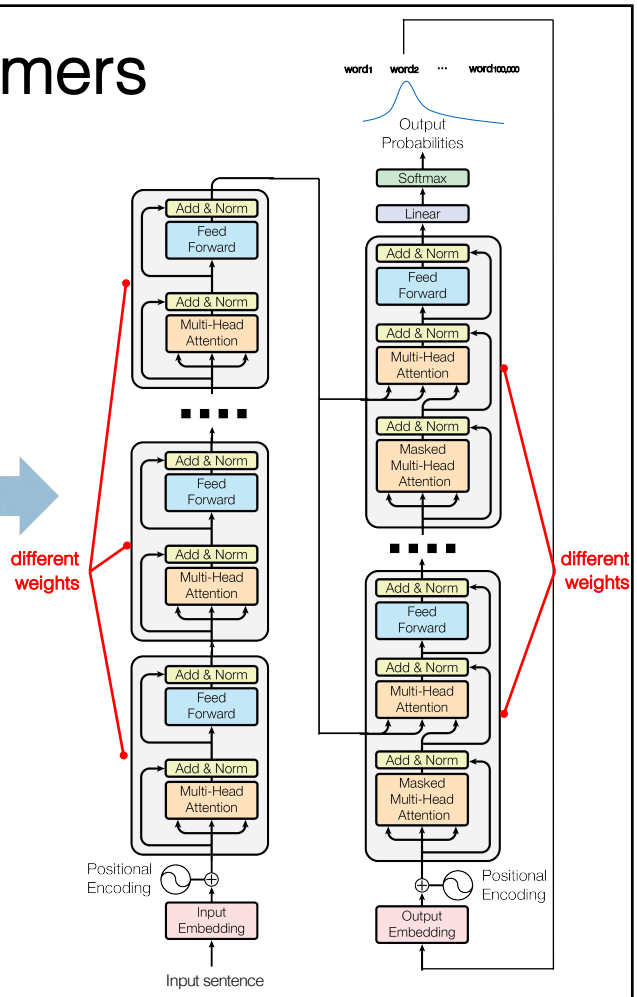
Transformers



- 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

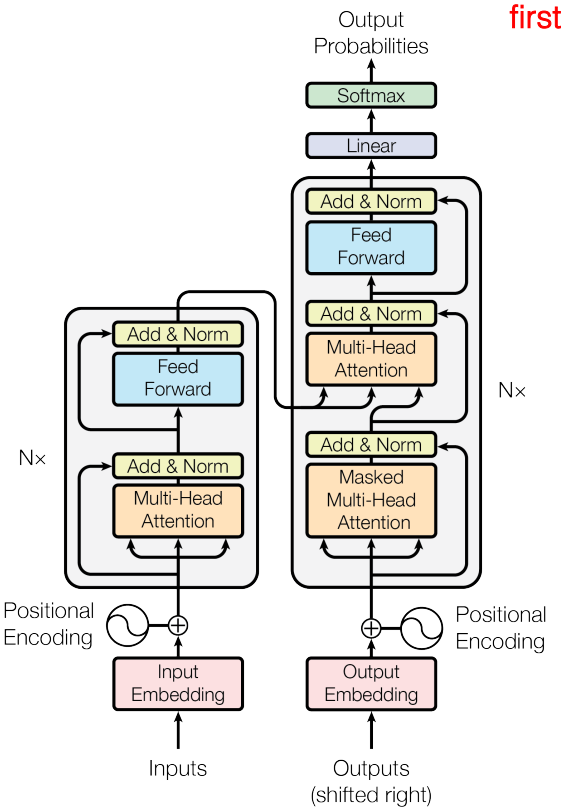
Transformers

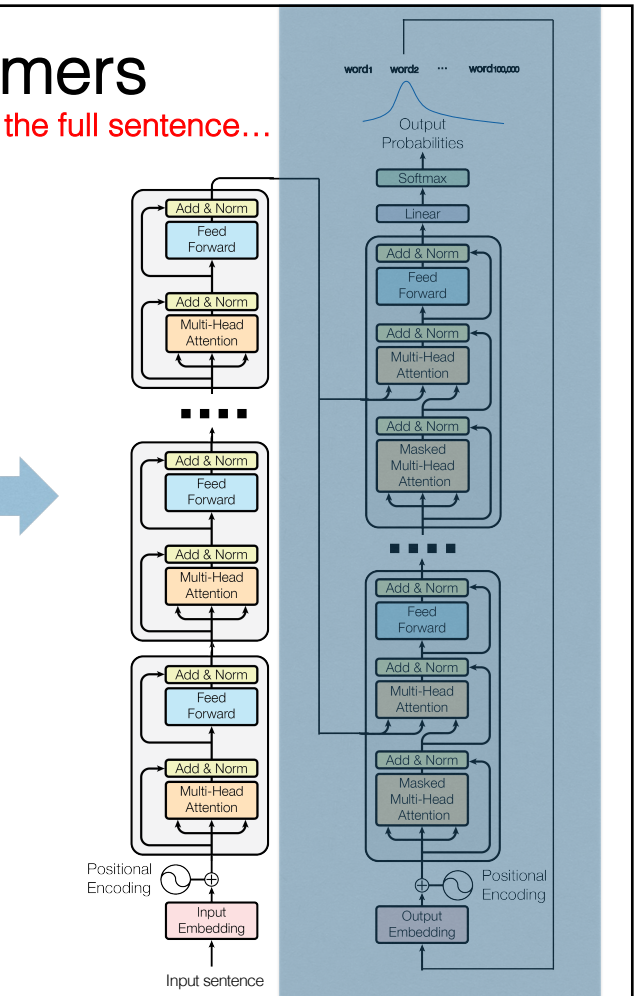




Transformers

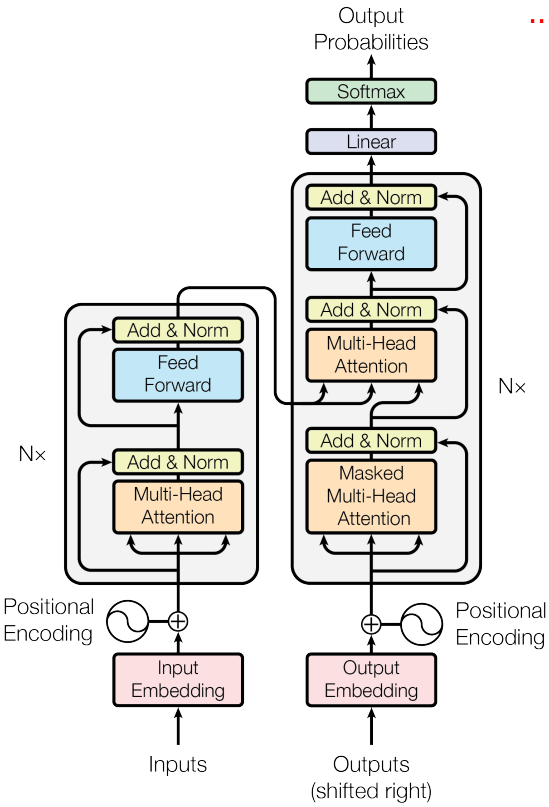
first encode the full sentence…

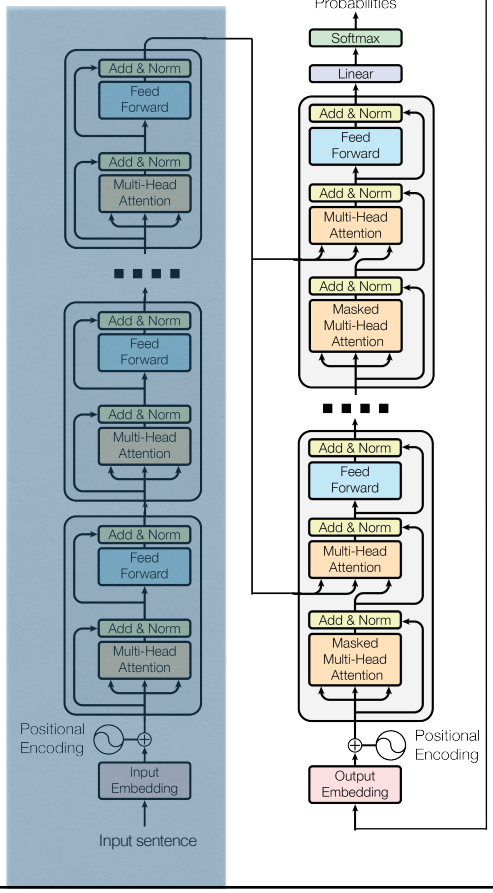




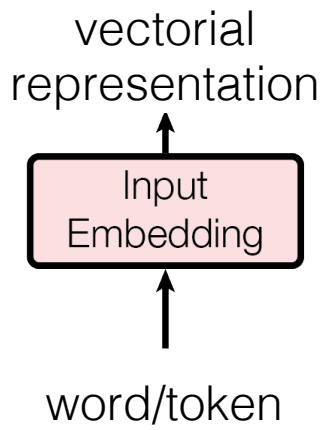
Transformers

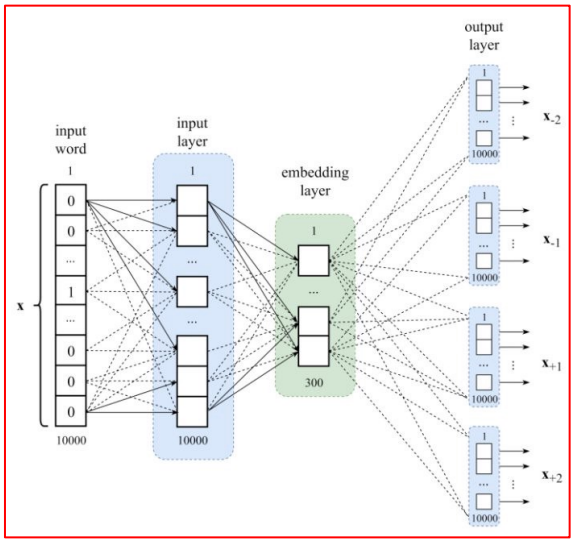
... then decode one world at a time

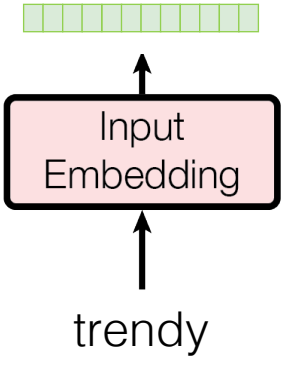




Input Embedding

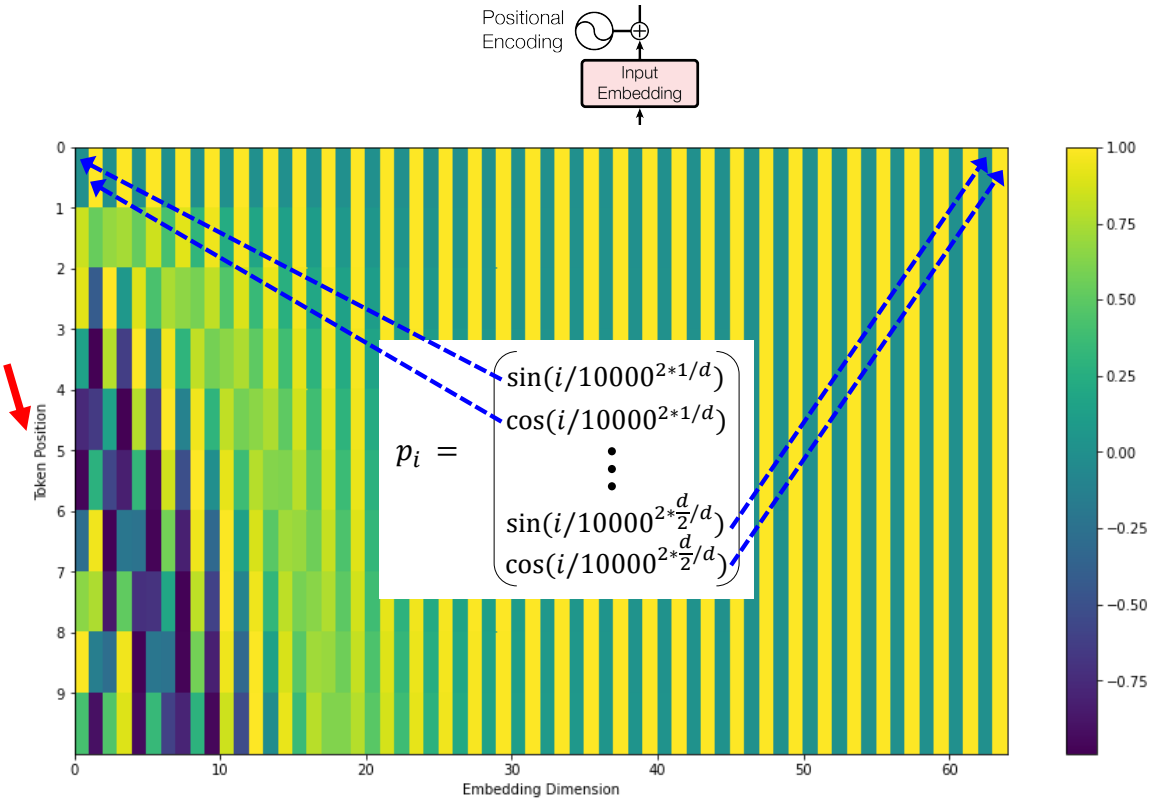




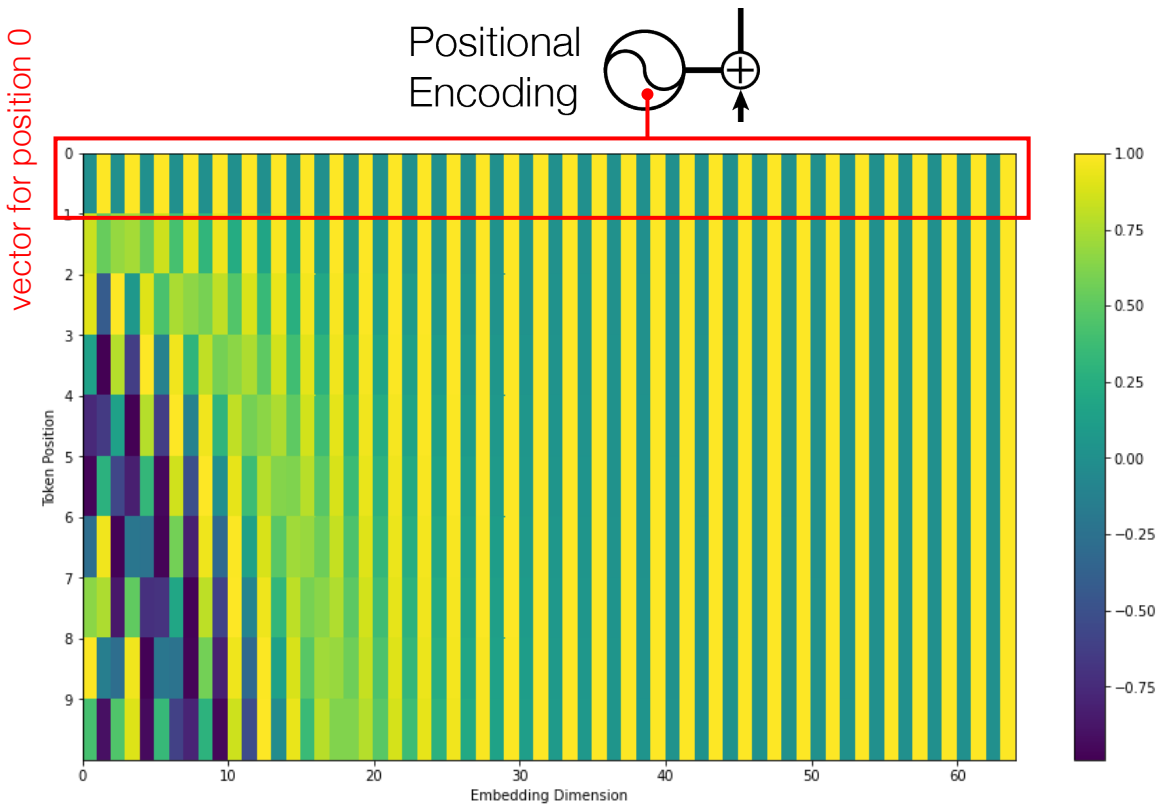


one way to generate embeddings of words:
learn to map a word into the left and right words where it appears in textual documents
(training done separately, once for all)

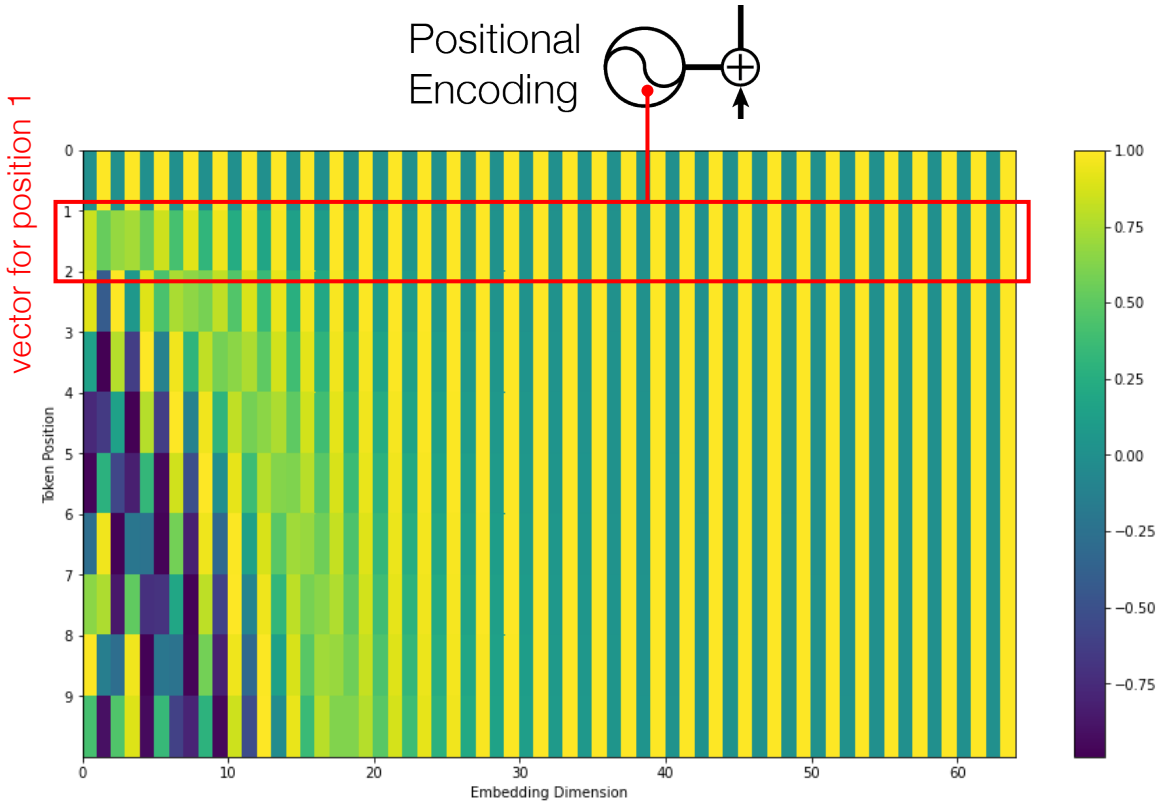
Positional Encoding



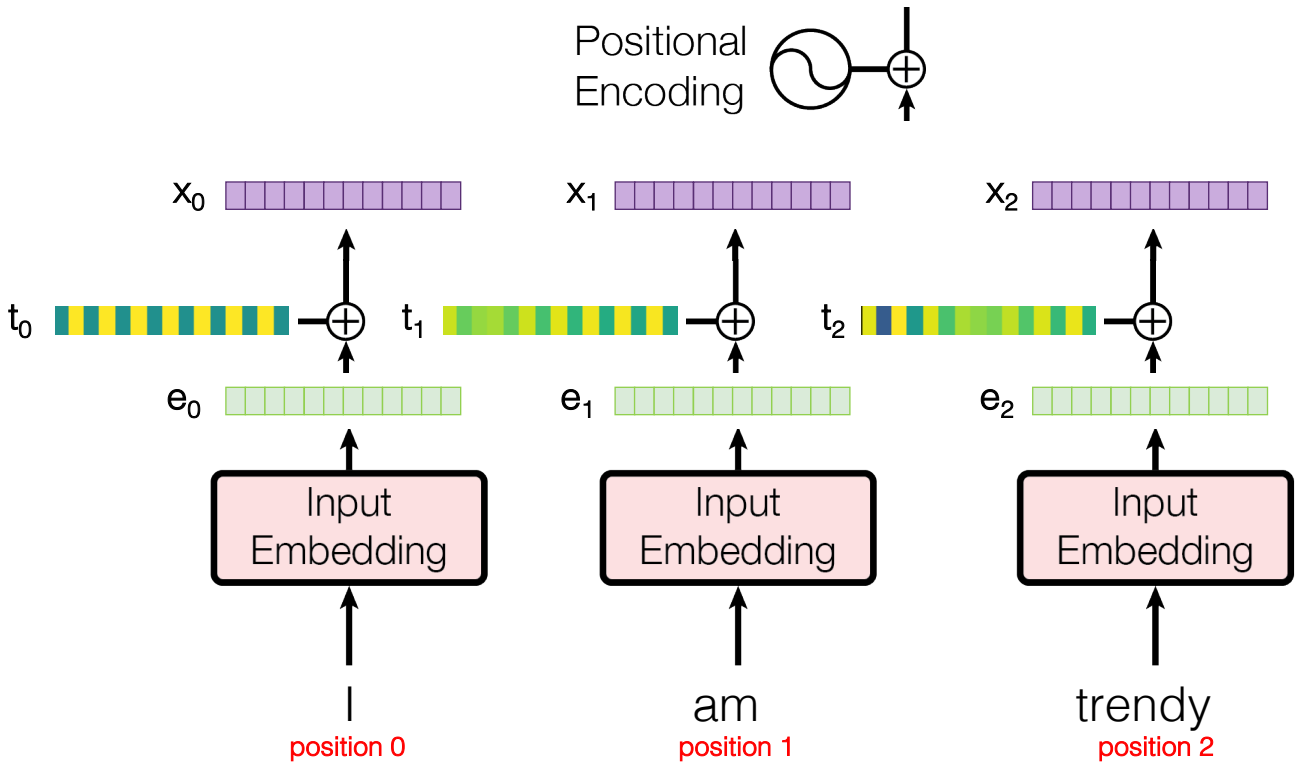
Positional Encoding



Positional Encoding

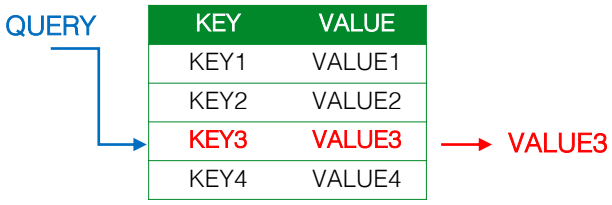


Positional Encoding



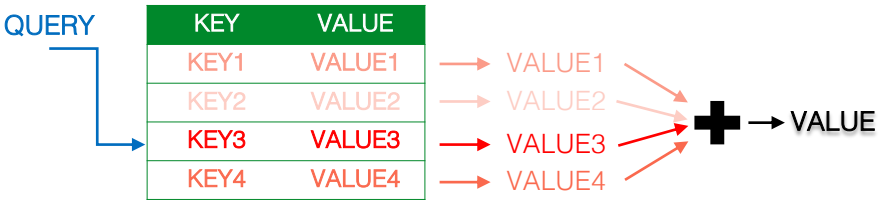
Attention

HARD HASHTABLE



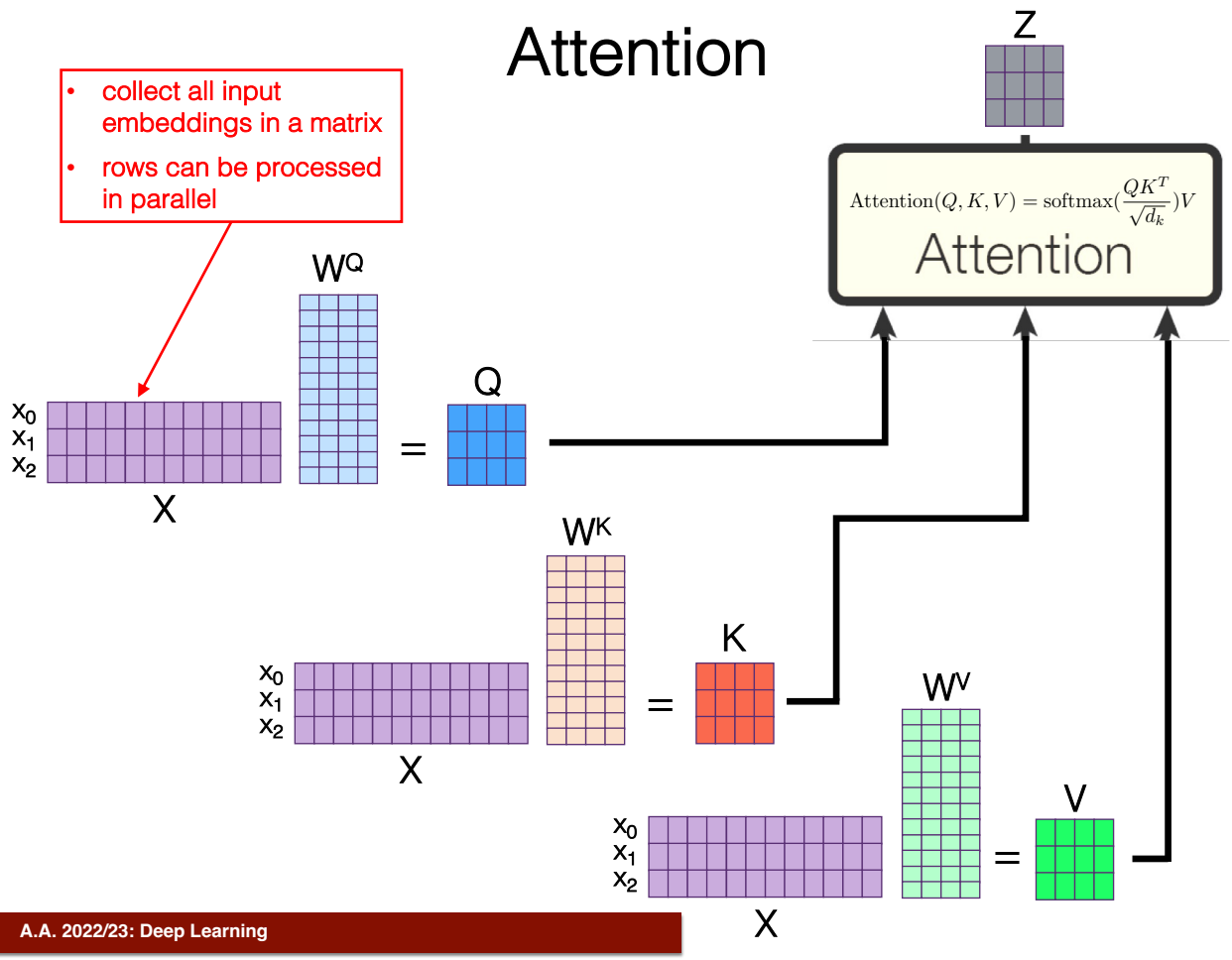
Each **query** (hash) maps to exactly one **key-value** pair

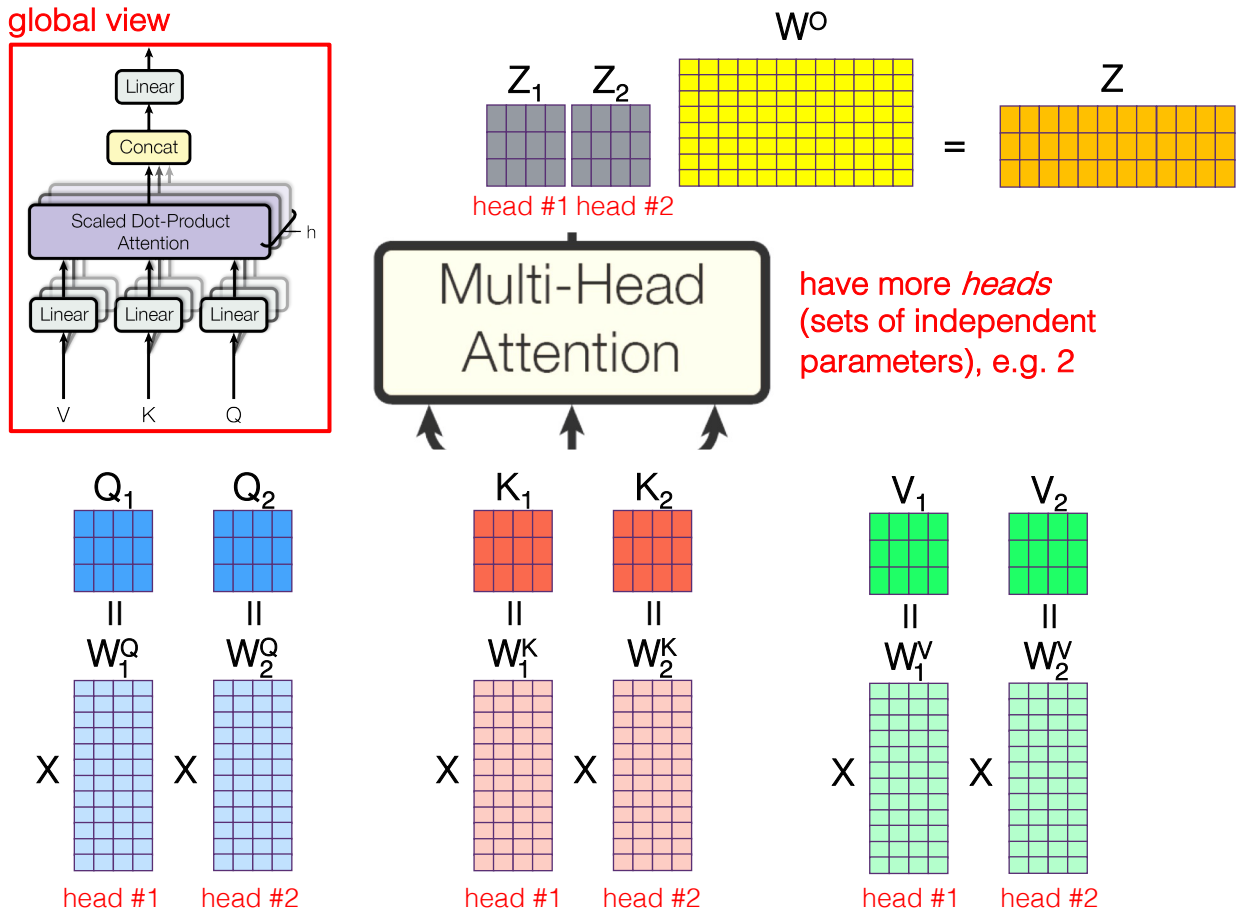
SOFT HASHTABLE



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

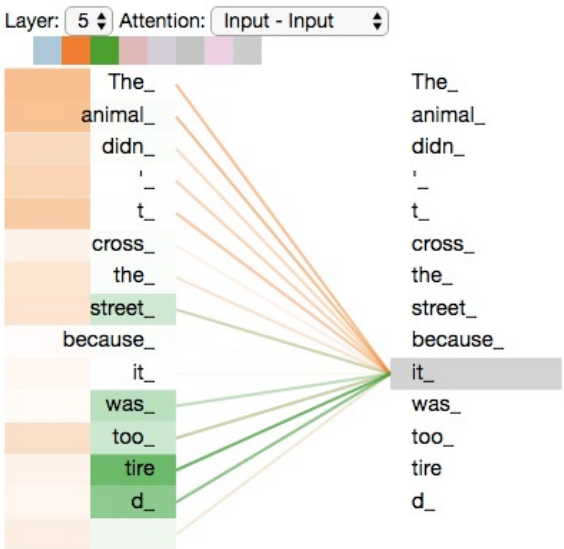
Attention





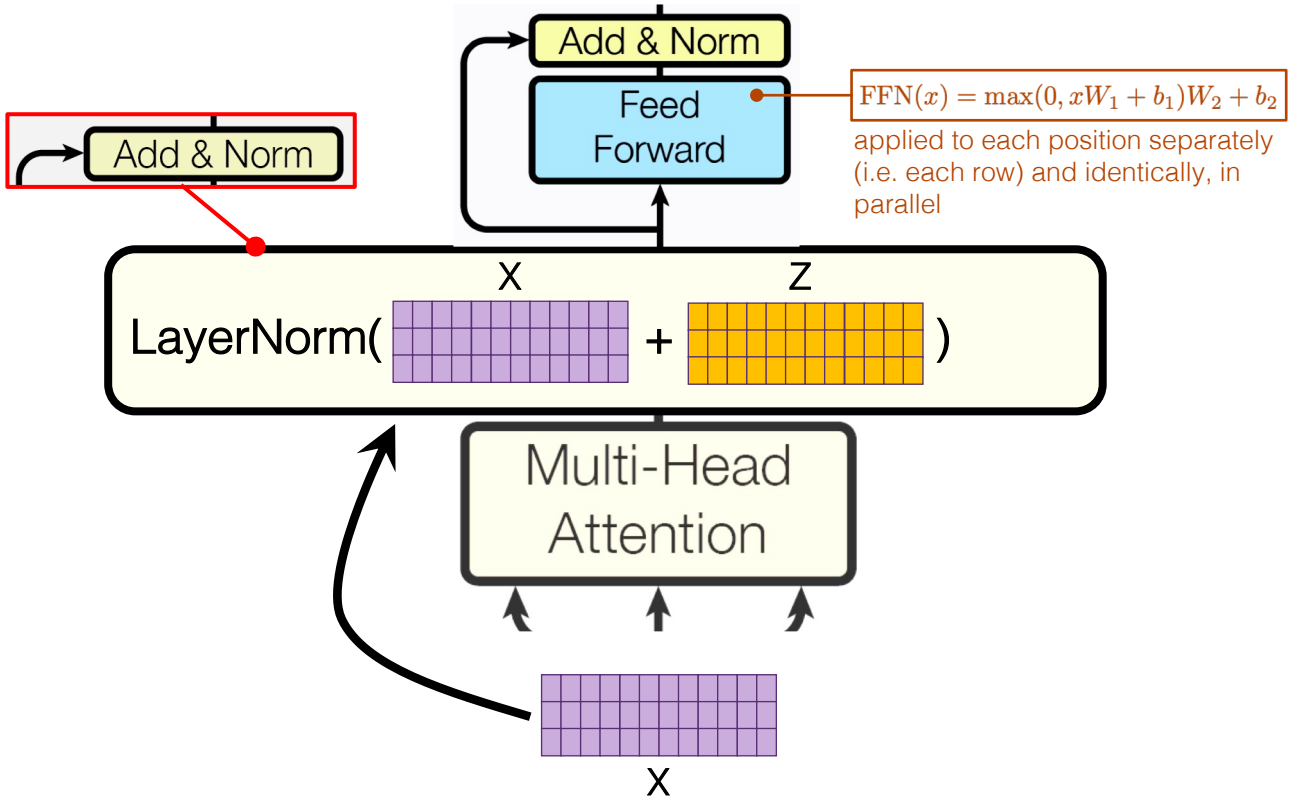
Multi-Head Attention Example

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



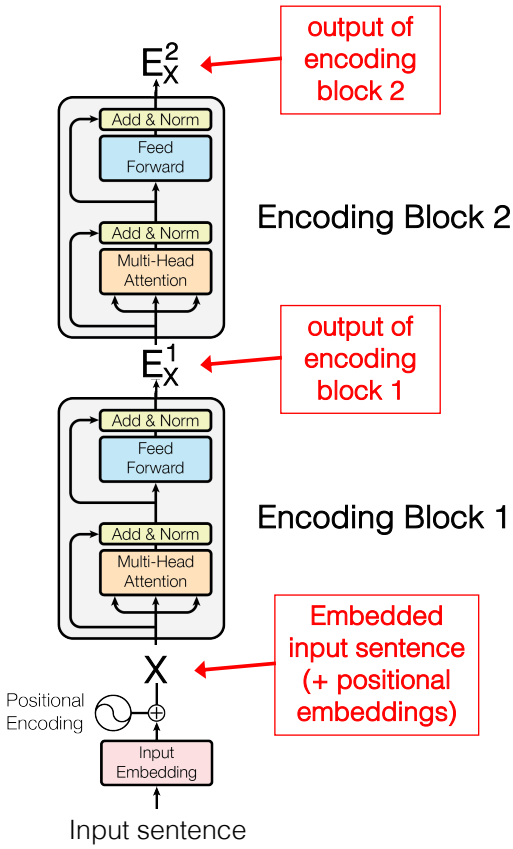
- Encoding the word "it", one attention head is focusing most on "the animal", while another is focusing on "tired"
- Thus, the encoding of the word "it" embeds both the *concept* of "animal" and the *concept* of "tired"

Add & Norm + Feed Forward Network

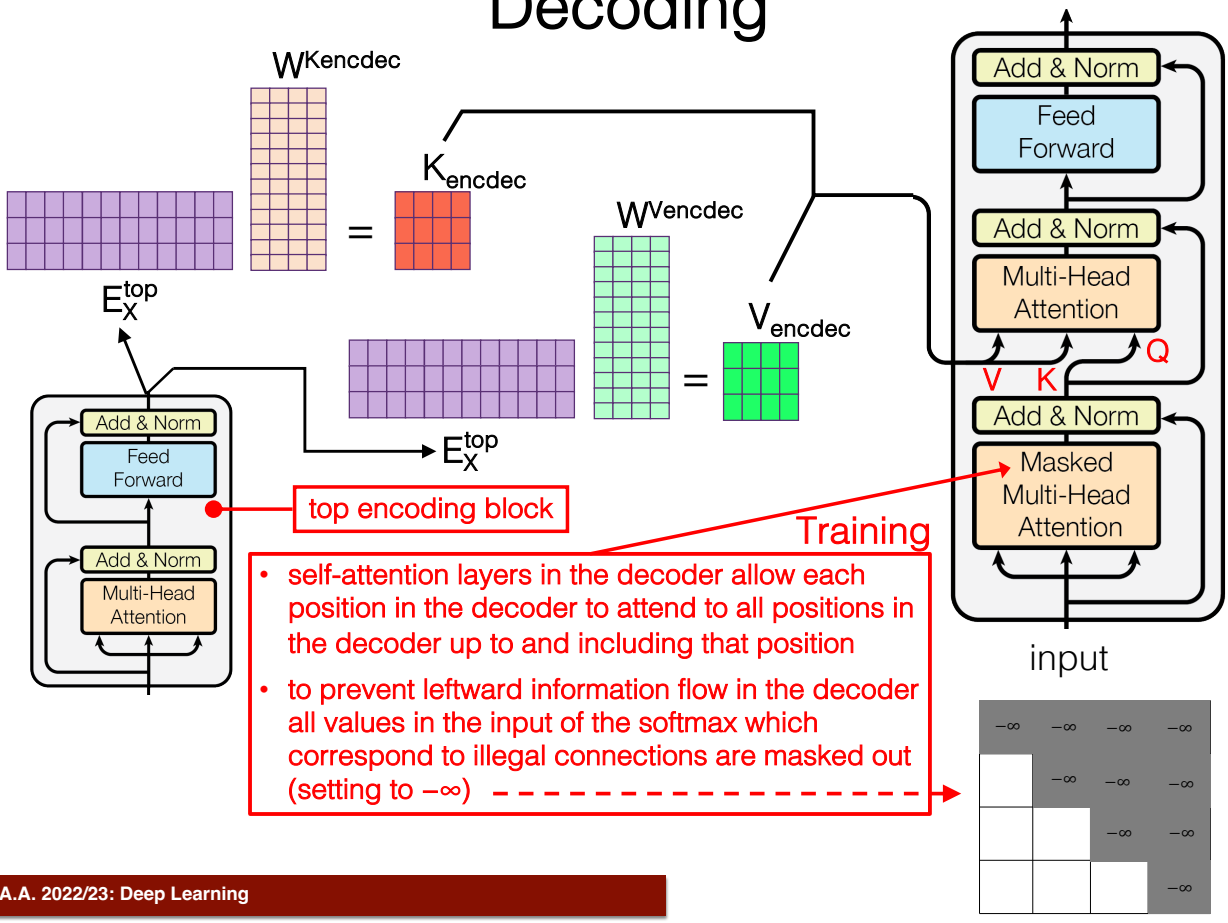


Stacking Encoding Blocks

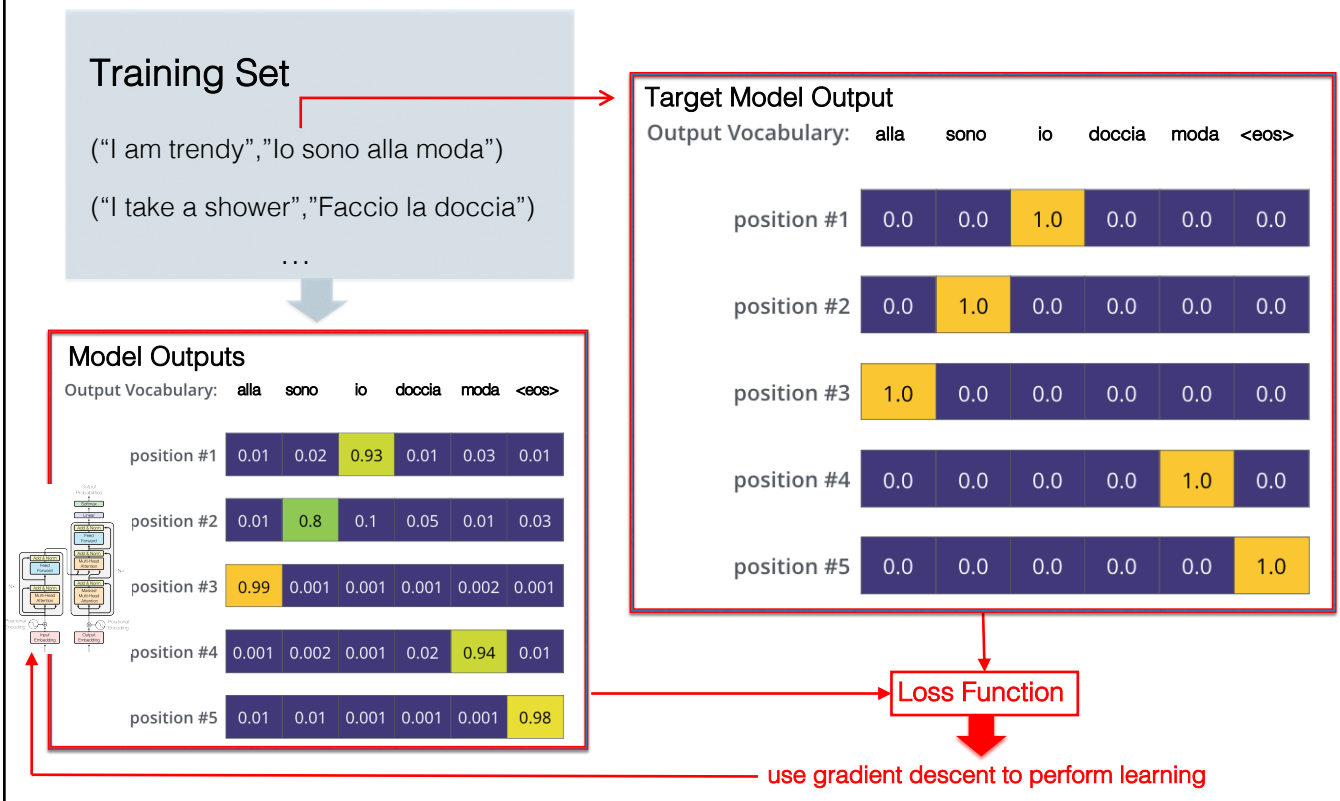
- Only the first encoding block is taking in input the embeddings of the words (+ positional embeddings), e.g. X
- All the other encoding blocks use in place of X the output of the encoding block below, i.e. encoding block j+1 takes in input E_X^j



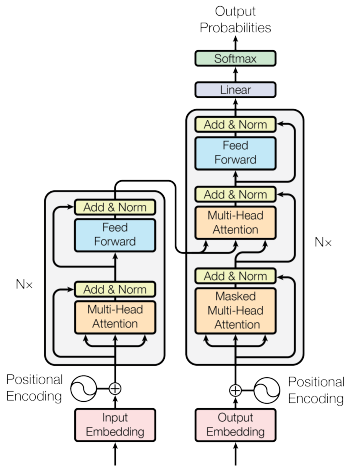
Decoding



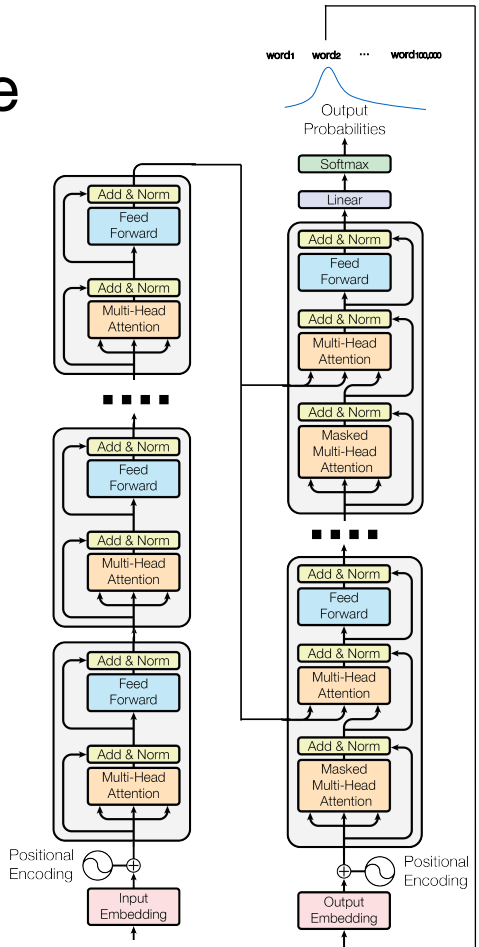
Training



Generation: the big picture

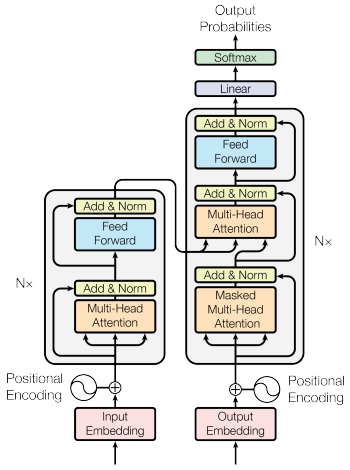


computation

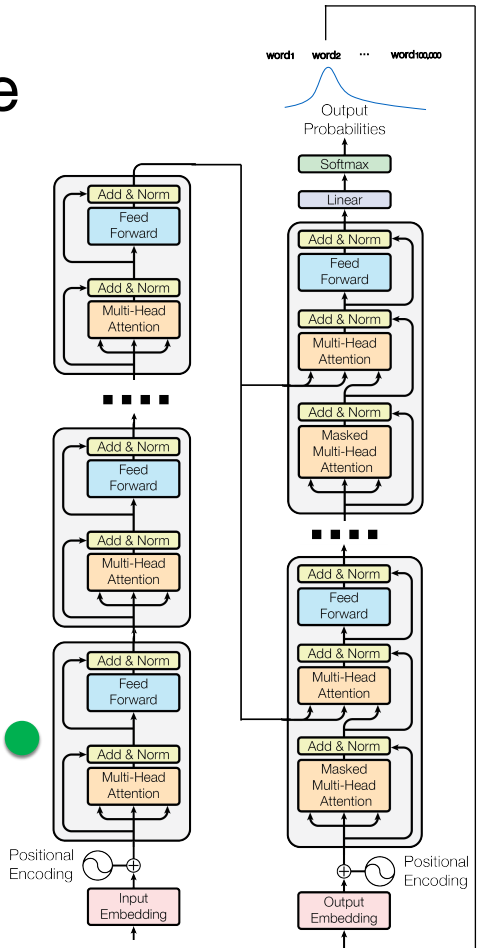


I am trendy

Generation: the big picture

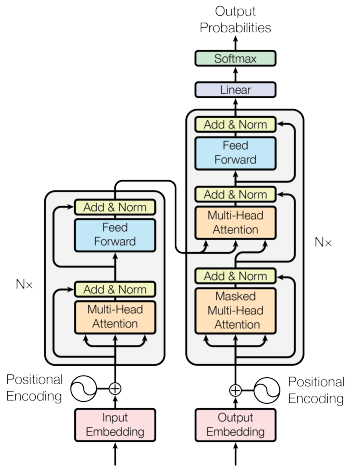


computation

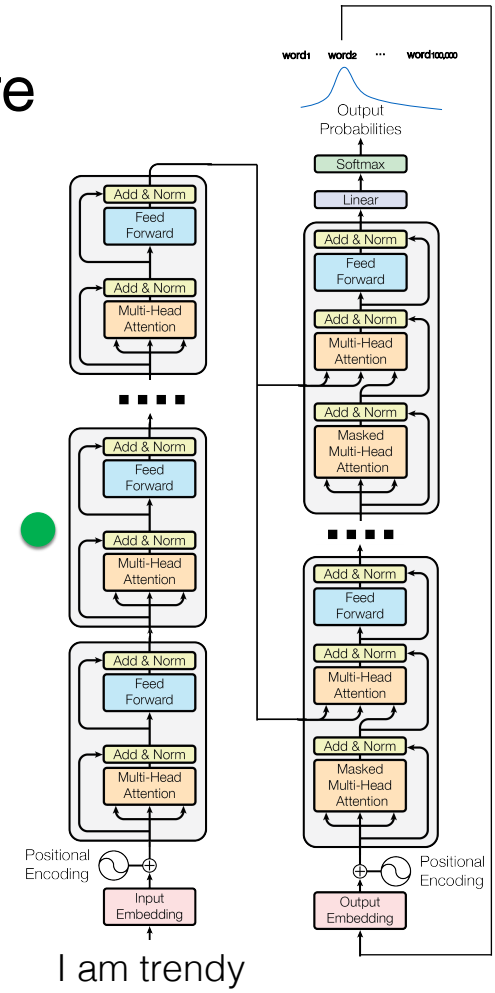


I am trendy

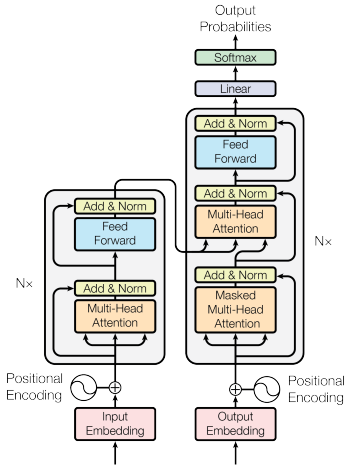
Generation: the big picture



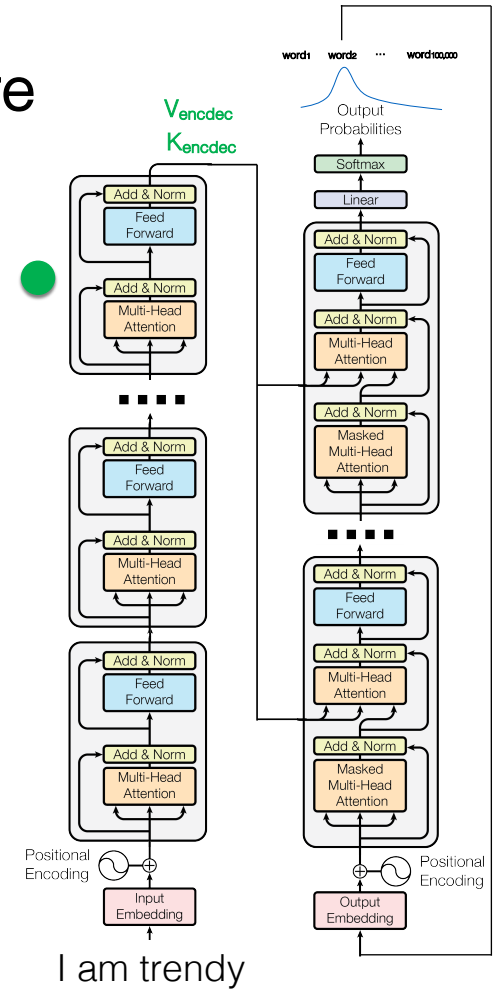
computation



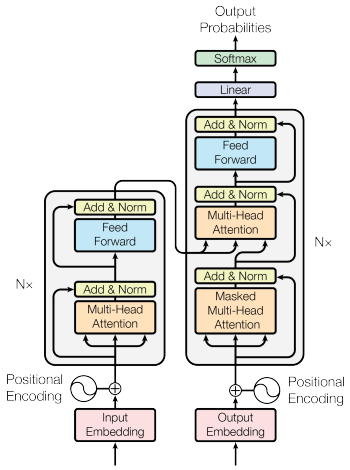
Generation: the big picture



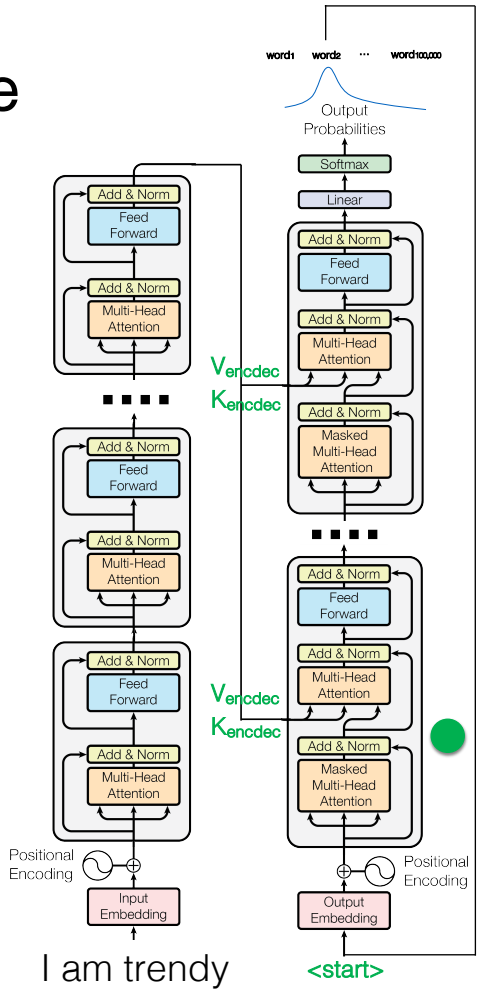
computation



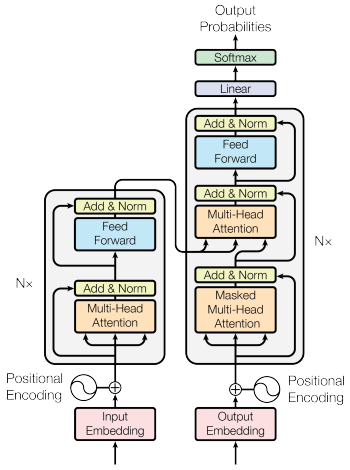
Generation: the big picture



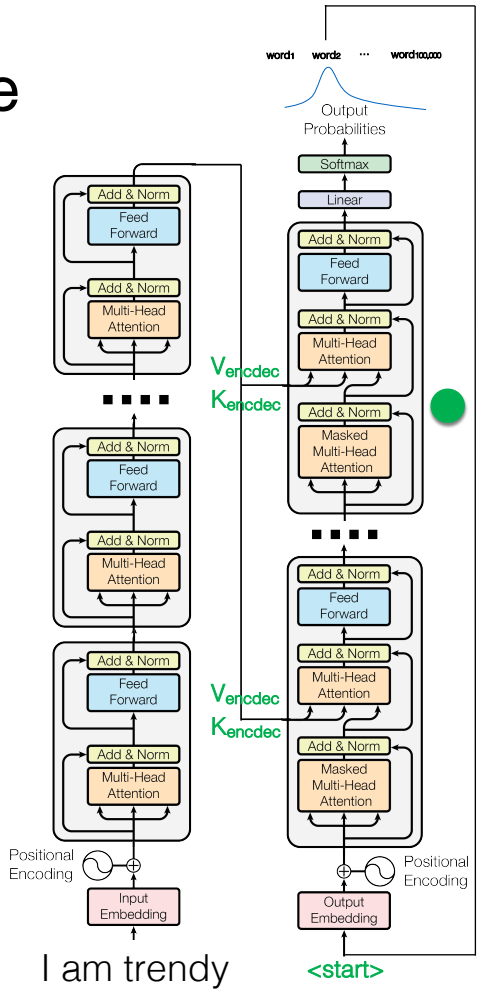
computation



Generation: the big picture

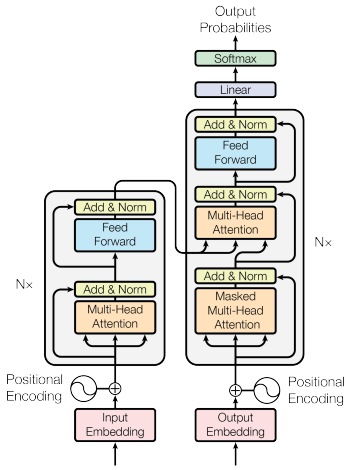


computation

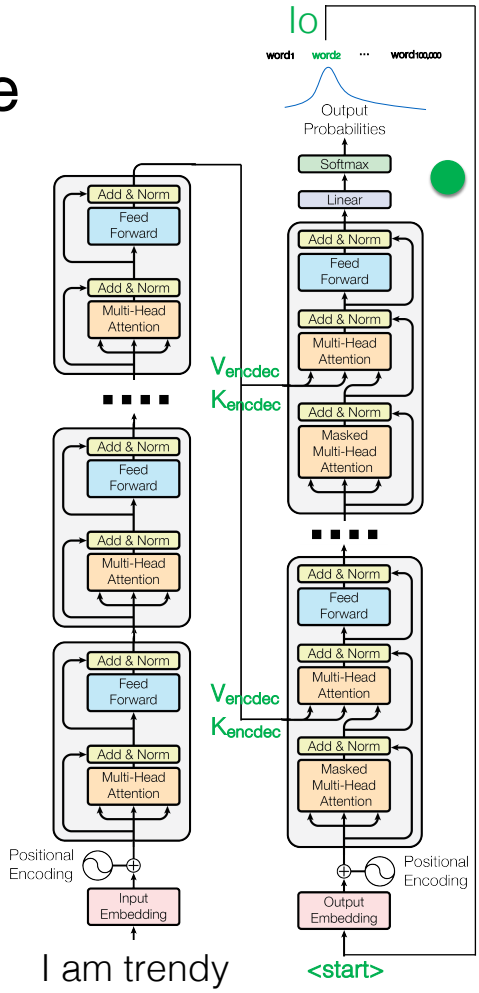


A.A. 2022/23: Deep Learning

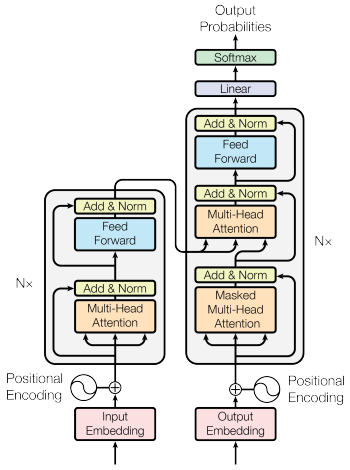
Generation: the big picture



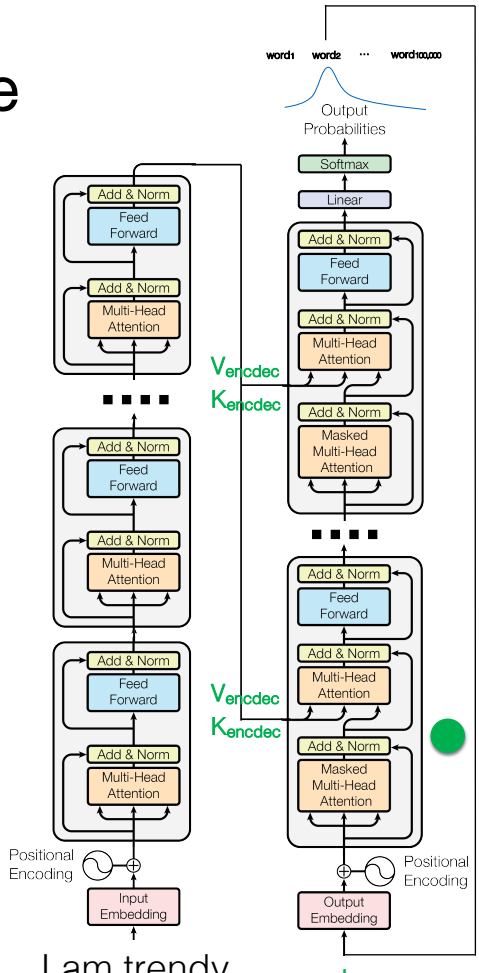
computation



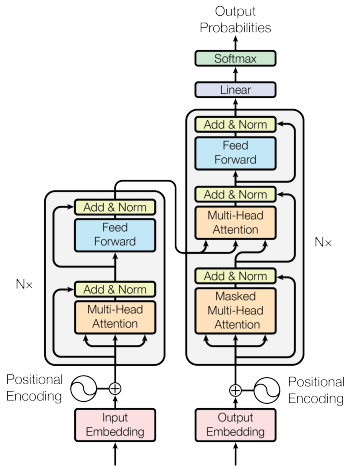
Generation: the big picture



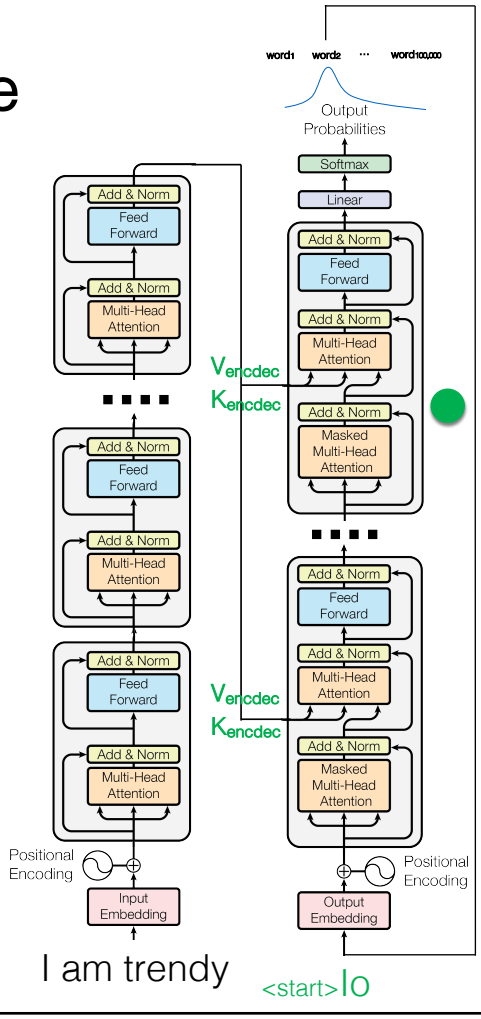
computation



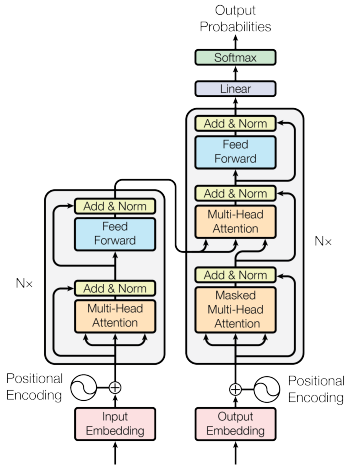
Generation: the big picture



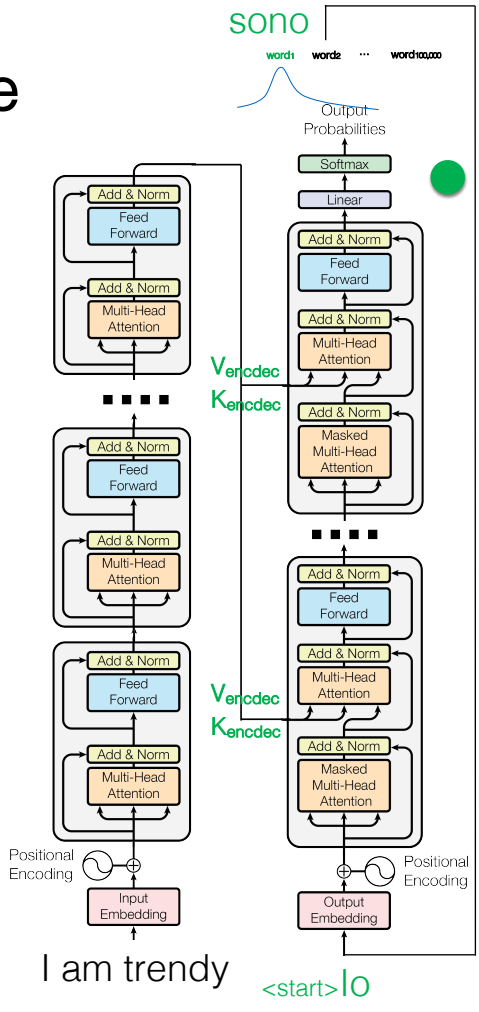
computation



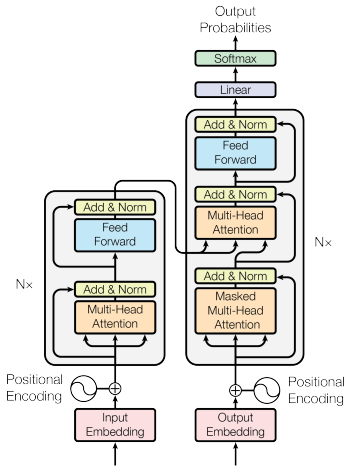
Generation: the big picture



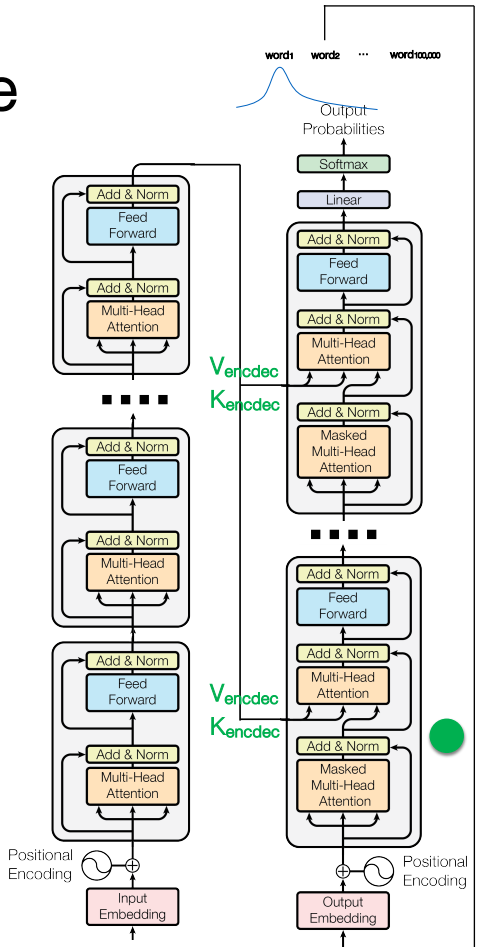
computation



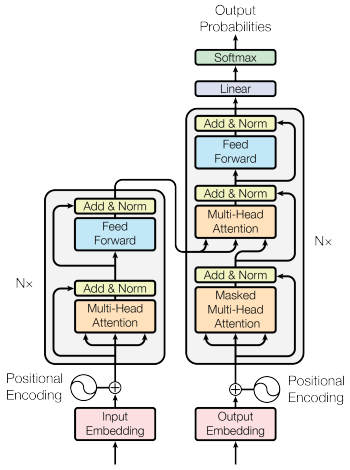
Generation: the big picture



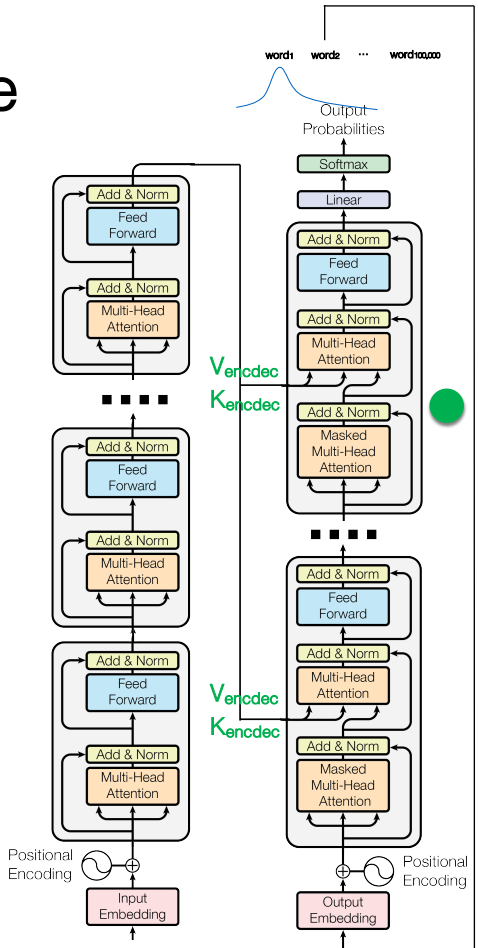
computation



Generation: the big picture

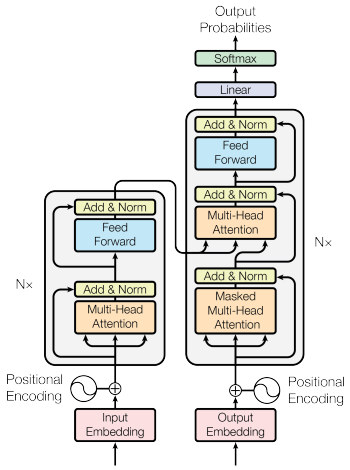


computation

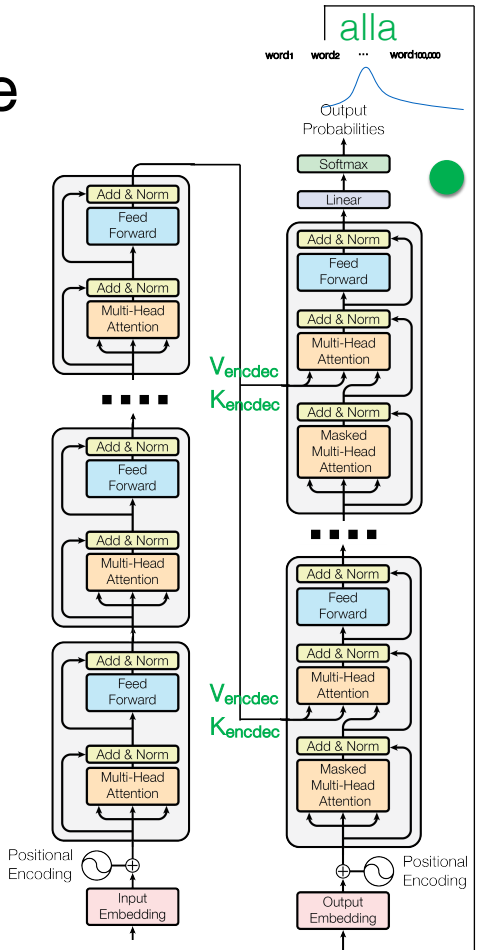


A.A. 2022/23: Deep Learning

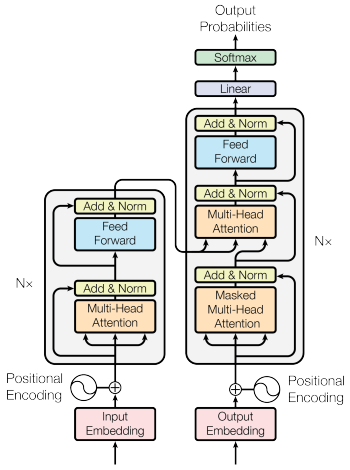
Generation: the big picture



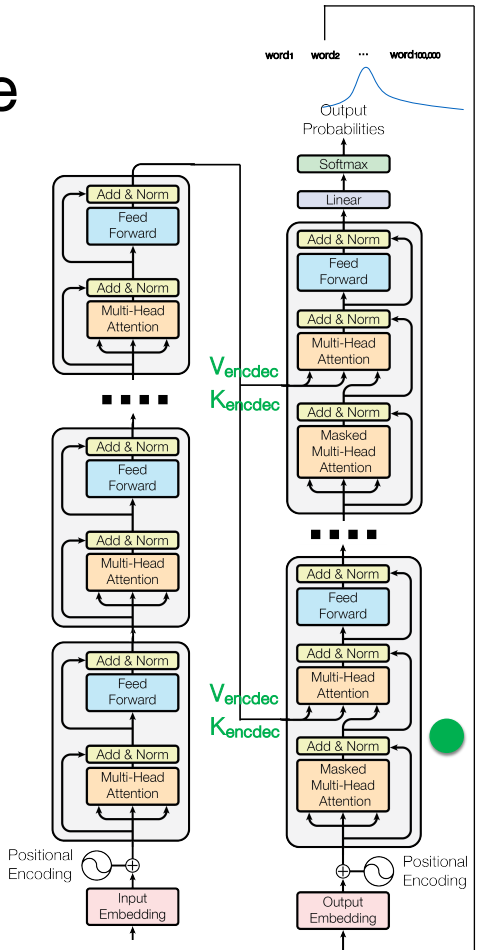
computation



Generation: the big picture

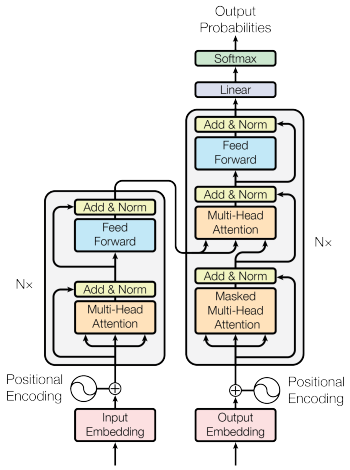


computation

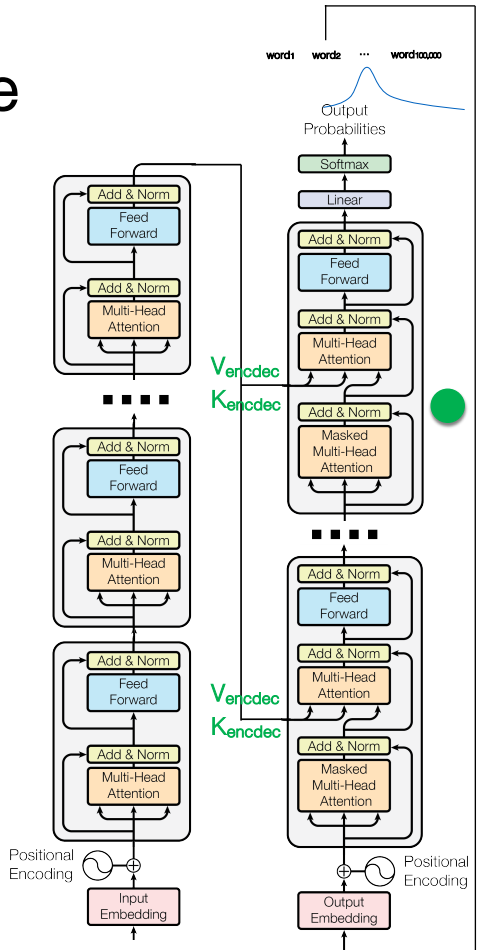


A.A. 2022/23: Deep Learning

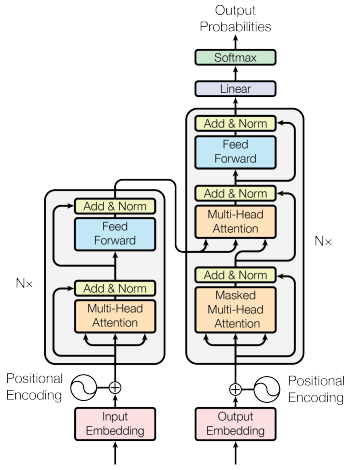
Generation: the big picture



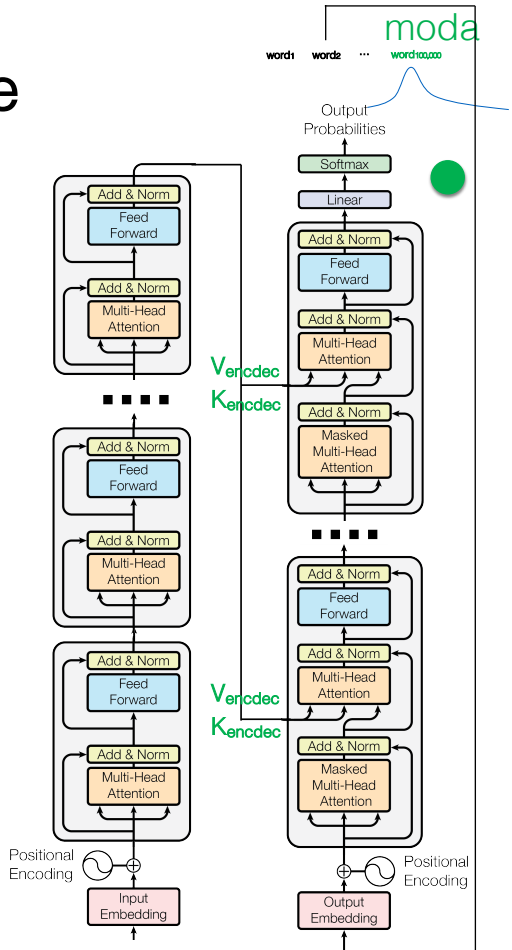
computation



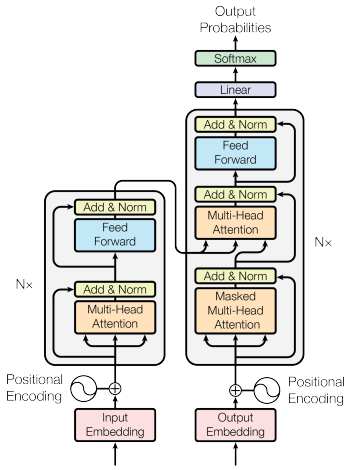
Generation: the big picture



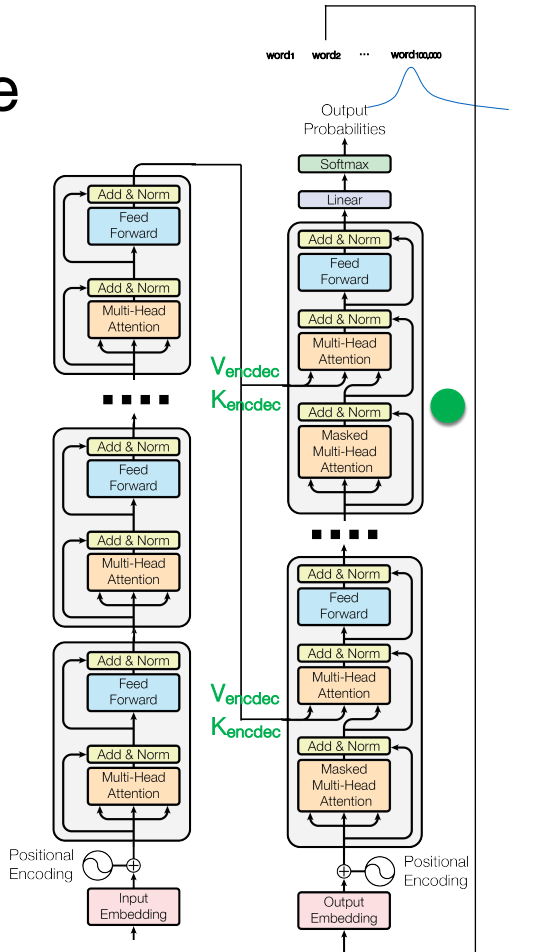
computation



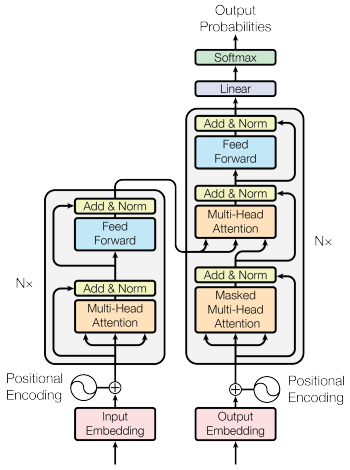
Generation: the big picture



computation



Generation: the big picture



computation