CS846
Machine Learning for Software Engineering

Pengyu Nie

Sequence-to-Sequence Models and Transformers

Tokenizer

Embedding

Attention

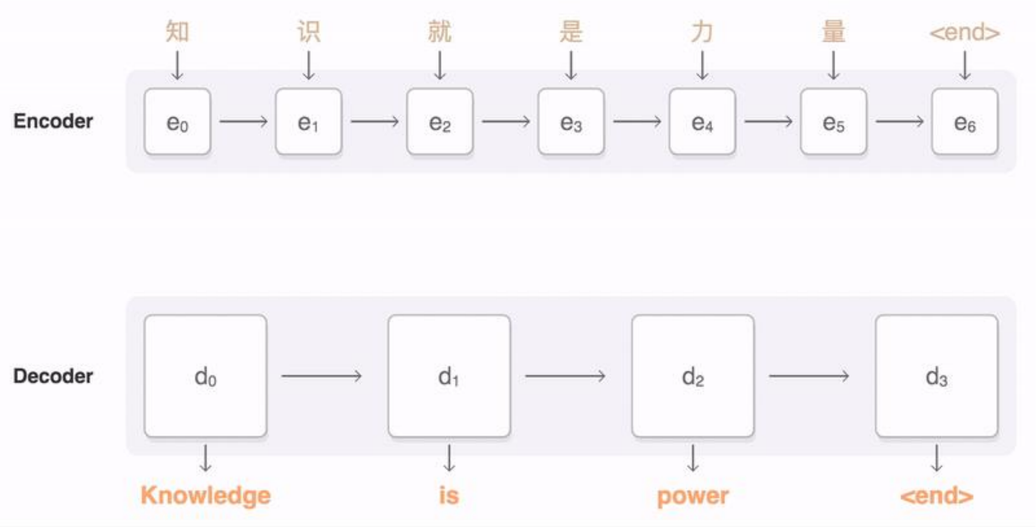
Acknowledgements: many slides adapted from Jessy Li & Milos Gligoric's ECE-W382V at UT Austin;
demo adapted from https://nlp.seas.harvard.edu/annotated-transformer/

Motivation

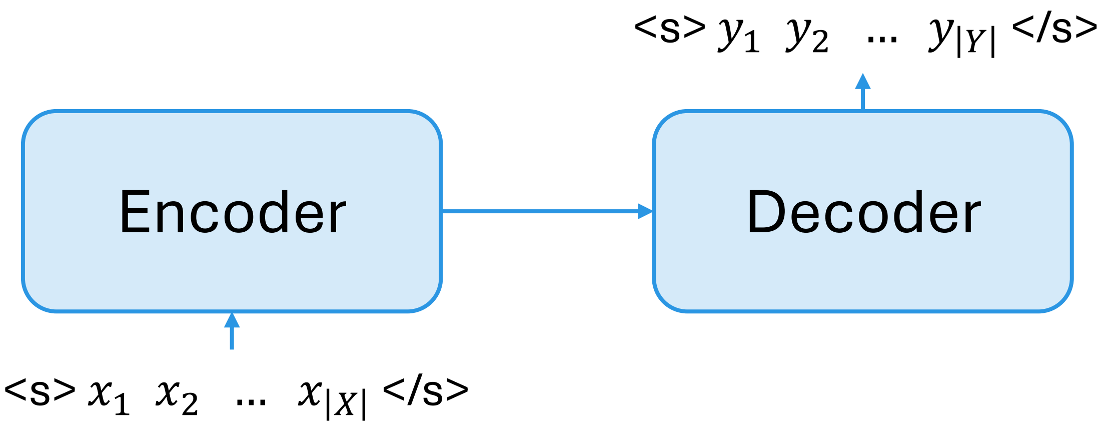
- Language model: $P(W)$ or $P(w_i|w_1, \dots, w_{i-1})$
- Tasks with inputs and outputs

	X	Y
Neural machine translation (NMT)	Chinese sentence	English sentence
Summarization	news article text	title
Examples in Software Engineering?		

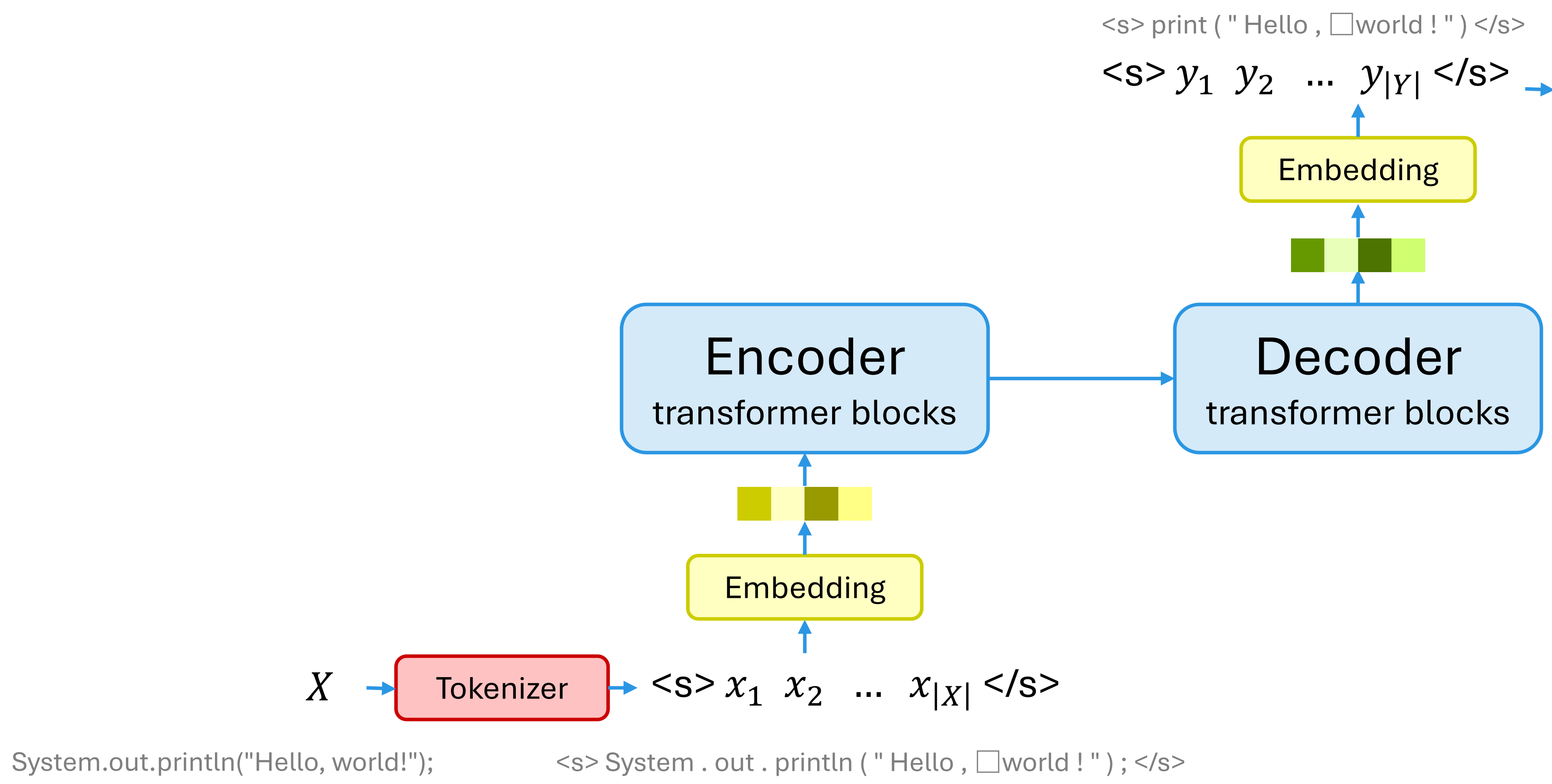
- $P(Y|X)$ or $P(y_i|y_1, \dots, y_{i-1}, x_1, \dots, x_{|X|})$

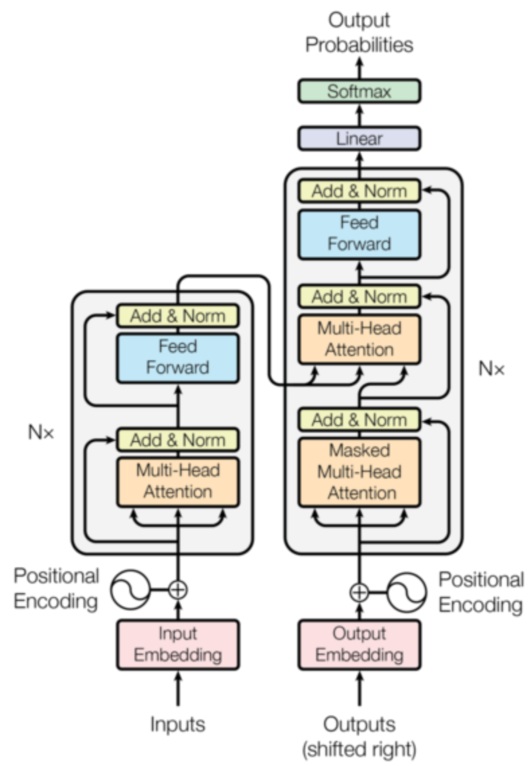


Architecture of Seq2Seq Model



Components


<s> print (" Hello , □world ! ") </s>
print("Hello, world!");
<s> y_1 y_2 … $y_{|Y|}$ </s>
Tokenizer
Y
Embedding
Encoder
transformer blocks
Decoder
transformer blocks
Embedding
X
Tokenizer
<s> x_1 x_2 … $x_{|X|}$ </s>
System.out.println("Hello, world!");
<s> System . out . println (" Hello , □world ! ") ; </s>


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)

Tokenizer

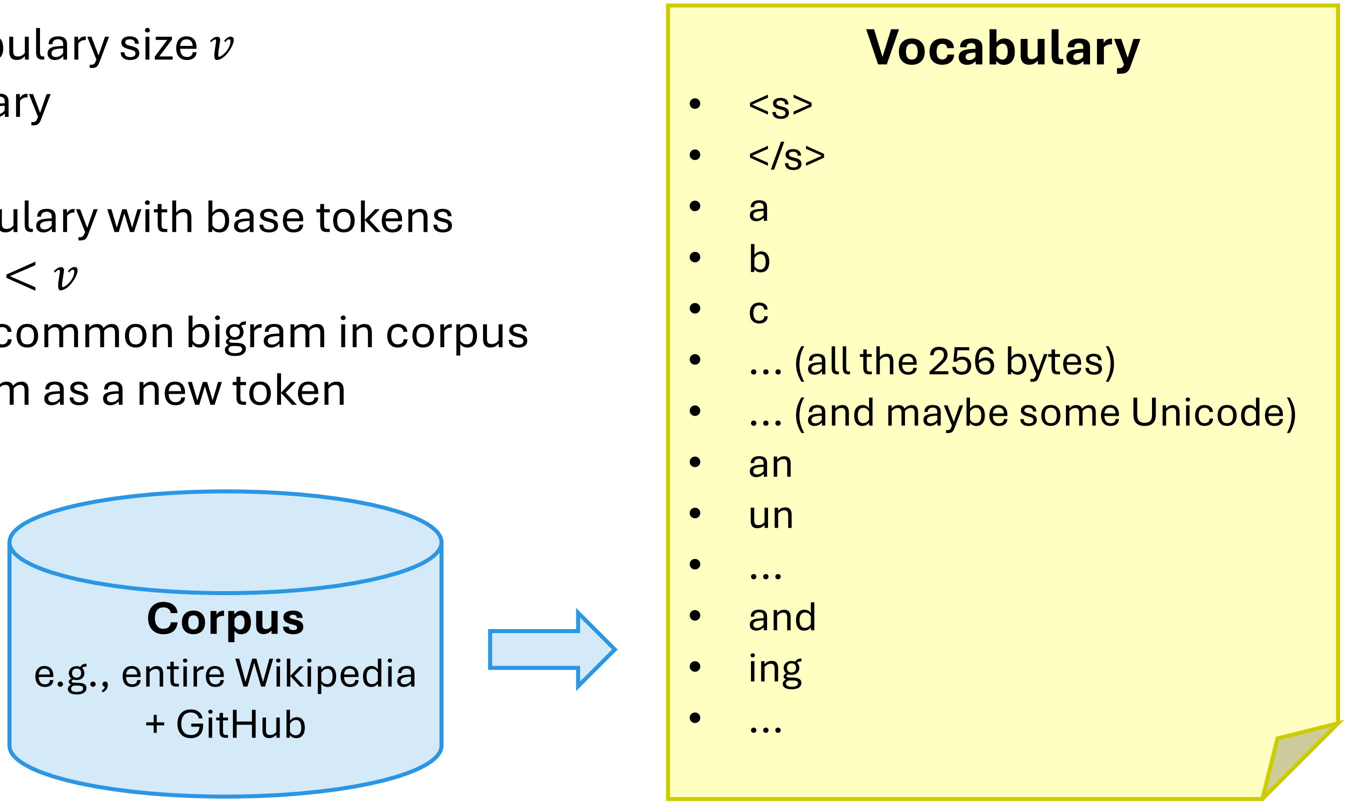
- Used to be...
  - Tokenize by whitespace / regex
  - Use PL's tokenizer
  - Sub-tokenize (important for PLs)
  - Tokens not seen in the training set = <UNK>
- Data-driven approach: byte-pair encoding (BPE)

```
String  inputPath  =  args[1];
String  inputPath  =  args  [  1  ]  ;
String  input  Path  =  args  [  1  ]  ;
```

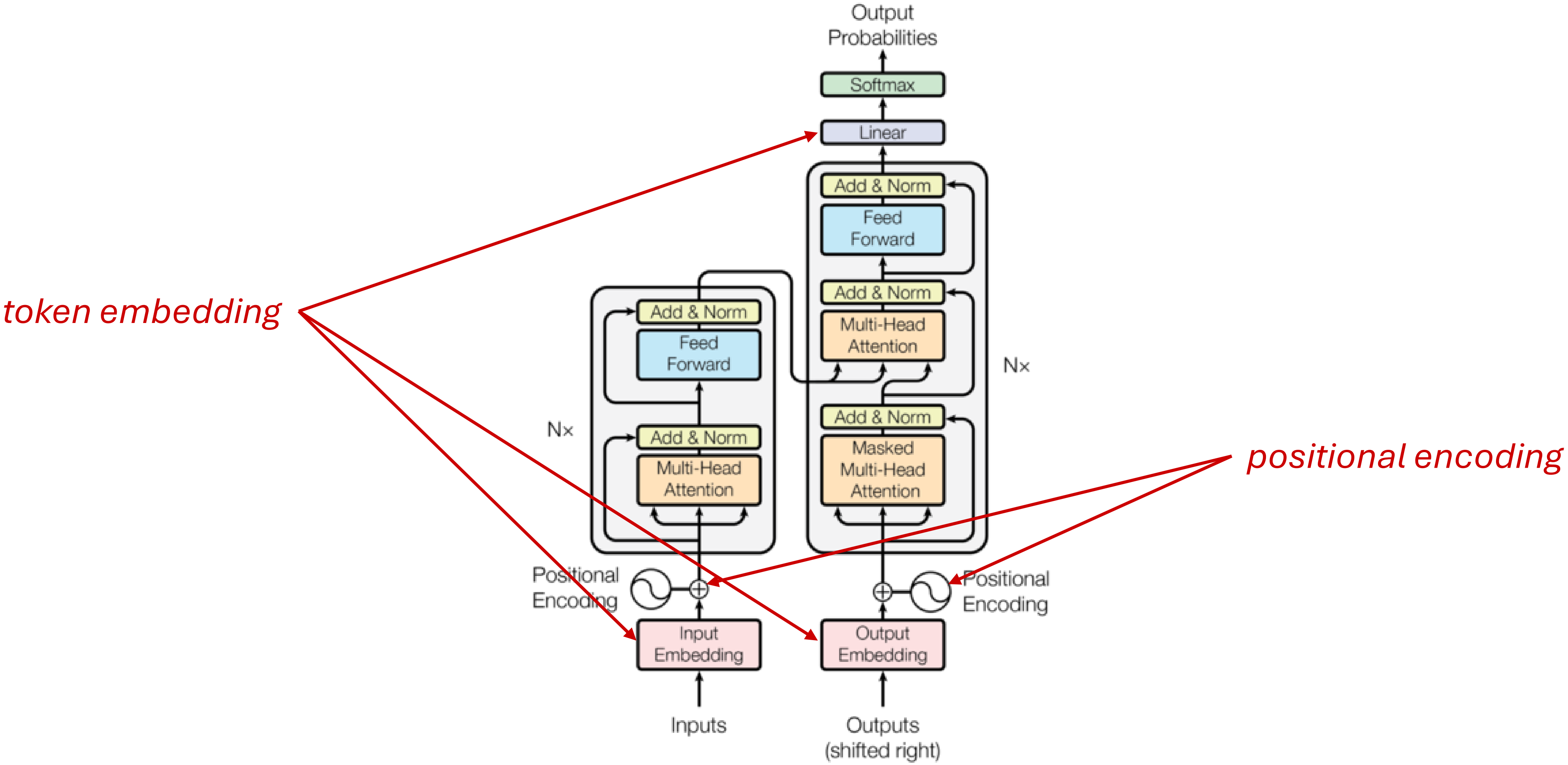
Byte-Pair Encoding

inputs: corpus, vocabulary size v
outputs: the vocabulary

- initialize the vocabulary with base tokens
- while $|\text{vocabulary}| < v$
  - find the most common bigram in corpus
  - add that bigram as a new token

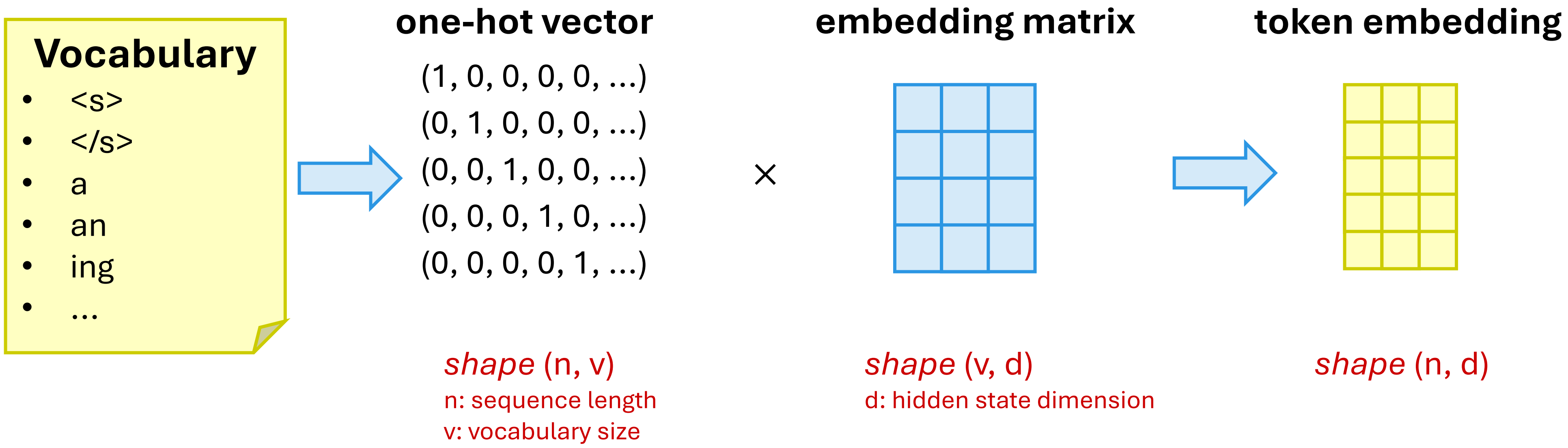


Embedding



token embedding
positional encoding
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)

Token/Word Embedding

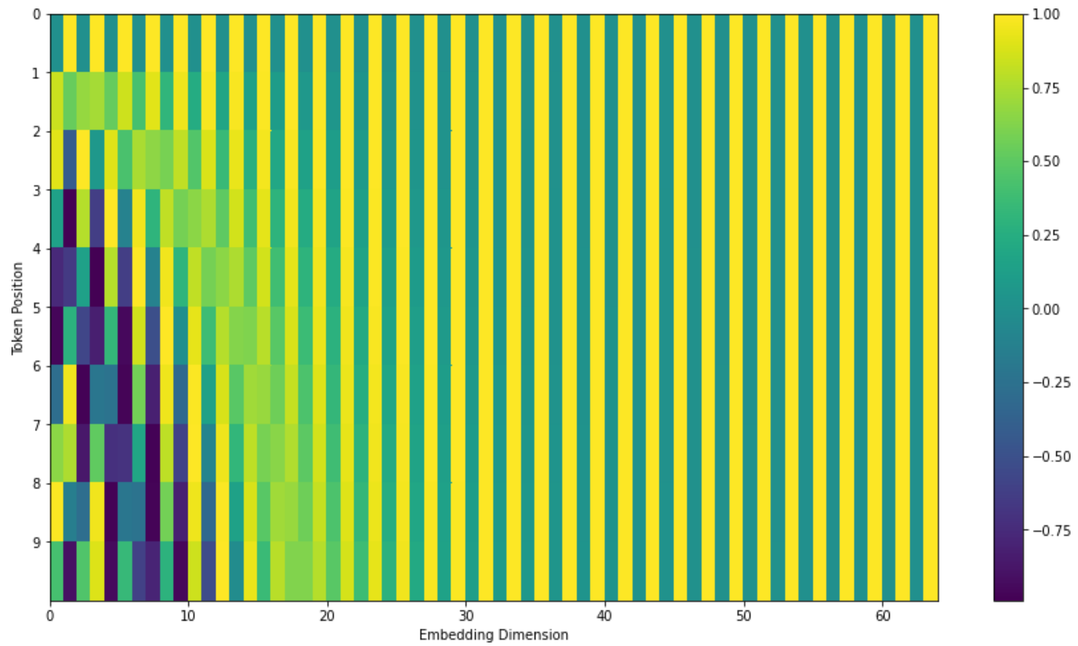


Positional Encoding

- Transformers ≠ recurrent neural network;
  we need something to represent the order of tokens

$$PE(pos, 2i) = \sin\left(\frac{pos}{10000^{\frac{2i}{d}}}\right)$$

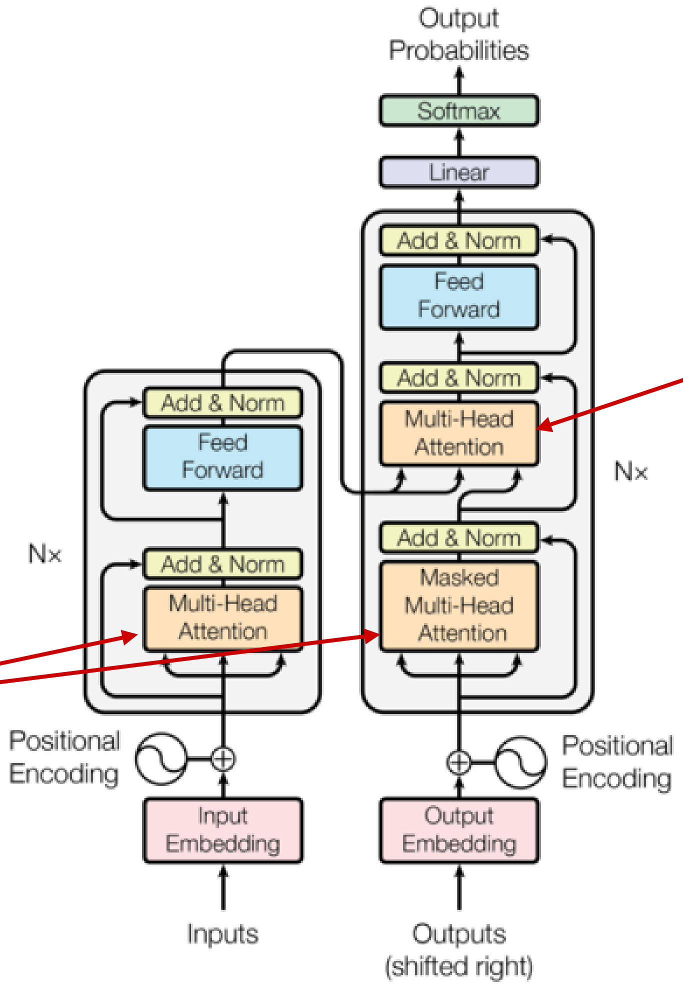
$$PE(pos, 2i+1) = \cos\left(\frac{pos}{10000^{\frac{2i}{d}}}\right)$$



Alternative approaches: learned position embedding, attention bias (https://arxiv.org/pdf/2108.12409)

Attention

cross attention layer

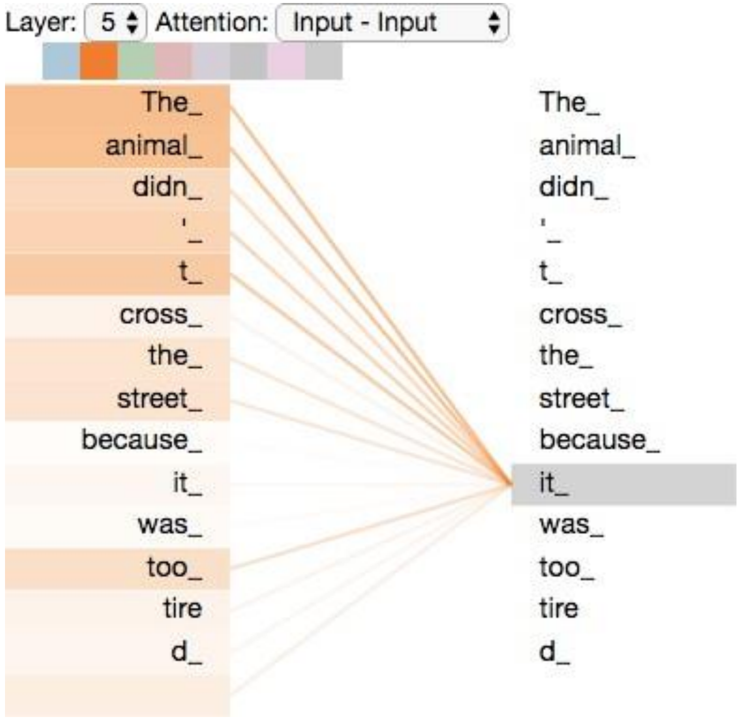


self attention layer

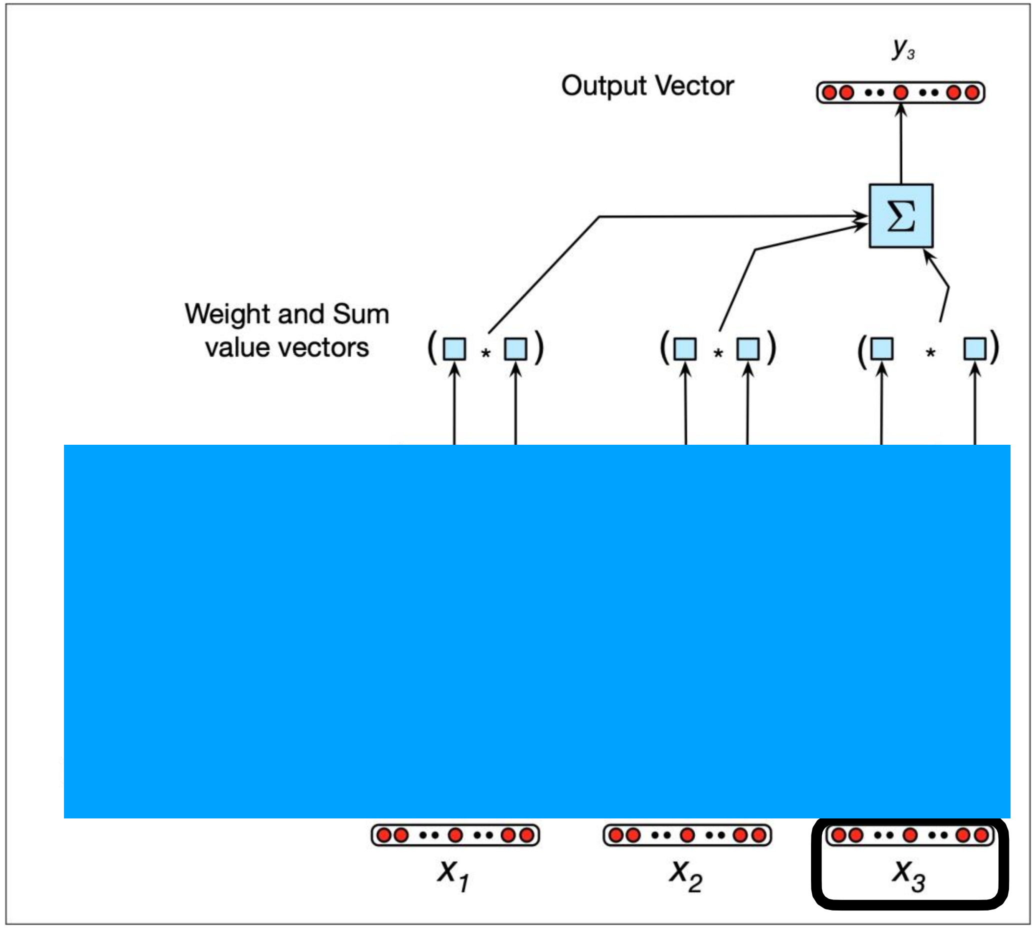
Self-attention

- Consider: “The animal didn't cross the street because it was too tired”

- What meaning should we associate with the word “it”?



Self-attention



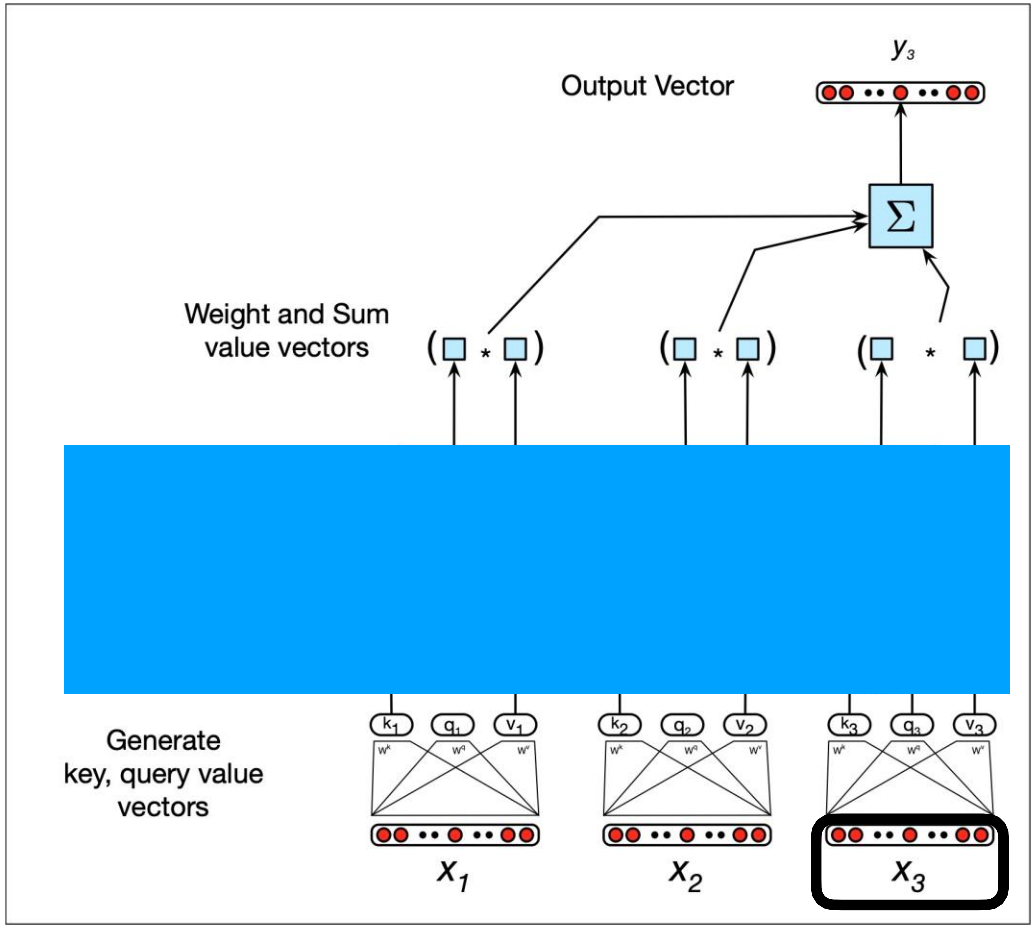
$$y_i = \sum_{j \le i} \alpha_{ij} x_j$$

$$\alpha_{ij} = \mathrm{softmax}(score(x_i, x_j)) = \frac{\exp(score(x_i, x_j))}{\sum_{k=1}^{i} \exp(score(x_i, x_k))}$$

$$score(x_i, x_j) = x_i \cdot x_j$$

But there is nothing to learn here?

Self-attention

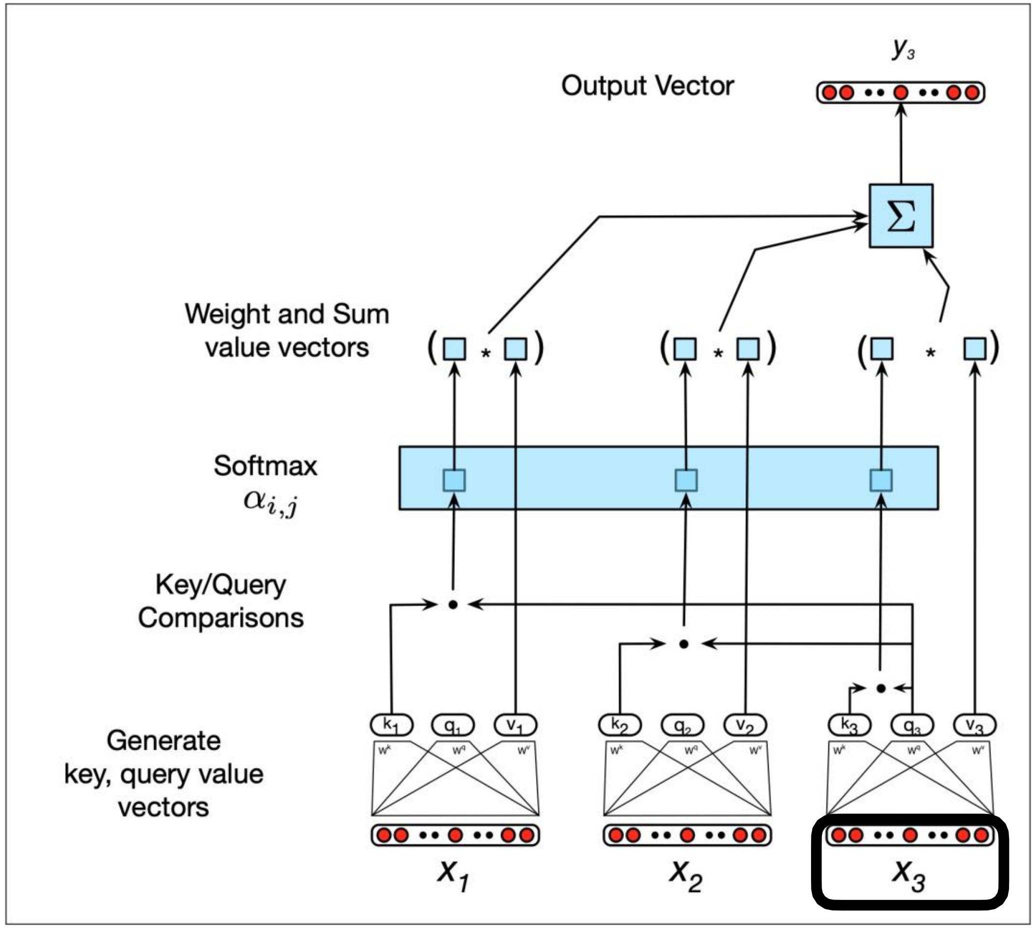


Introducing 3 types of weights, corresponding to 3 roles of each word ***w*:**

- **Query**: w is the current word under question
- **Key**: w is the word in context being compared with
- **Value**: learnable weights for the output

$$q_i = W^Q x_i; \;\; k_i = W^K x_i; \;\; v_i = W^V x_i$$

Self-attention



$$\boxed{y_i = \sum_{j \le i} \alpha_{ij} v_j}$$

also multiply with v

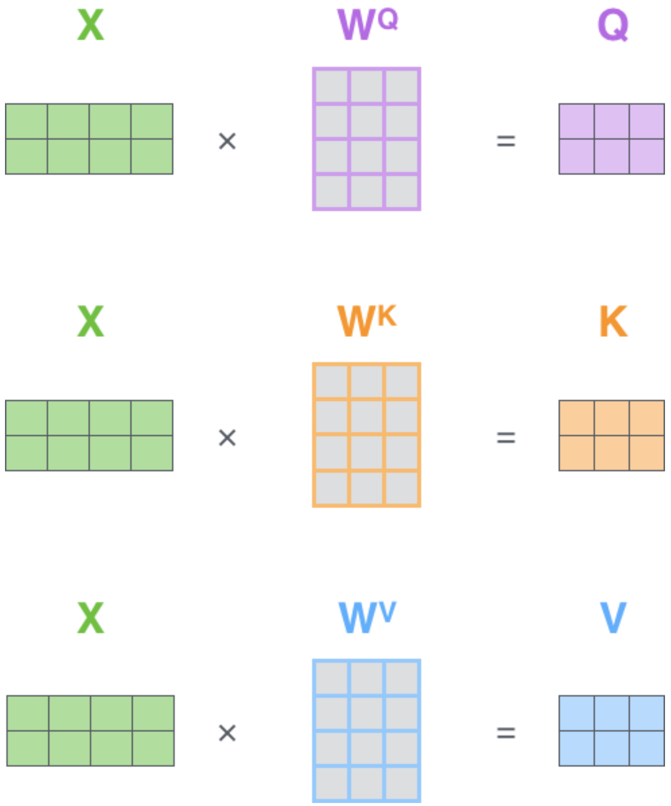
$$\boxed{\alpha_{ij} = \text{softmax}(score(x_i, x_j))}$$

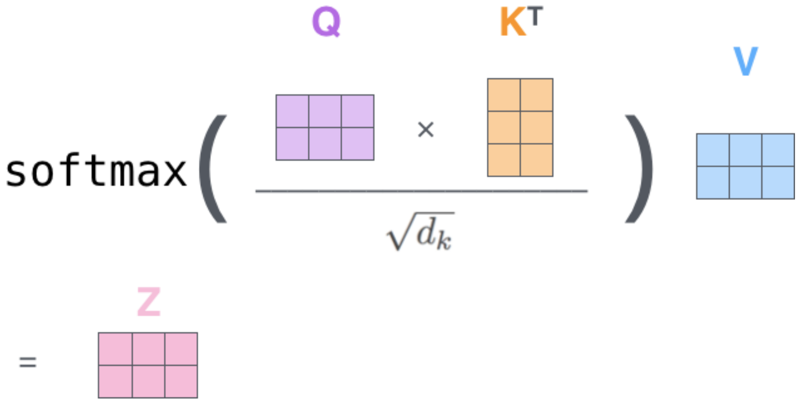
normalize

$$\boxed{score(x_i, x_j) = q_i \cdot k_j}$$

$$\boxed{q_i = W^Q x_i;\ \ k_i = W^K x_i;\ \ v_i = W^V x_i}$$

Self-attention in one graph

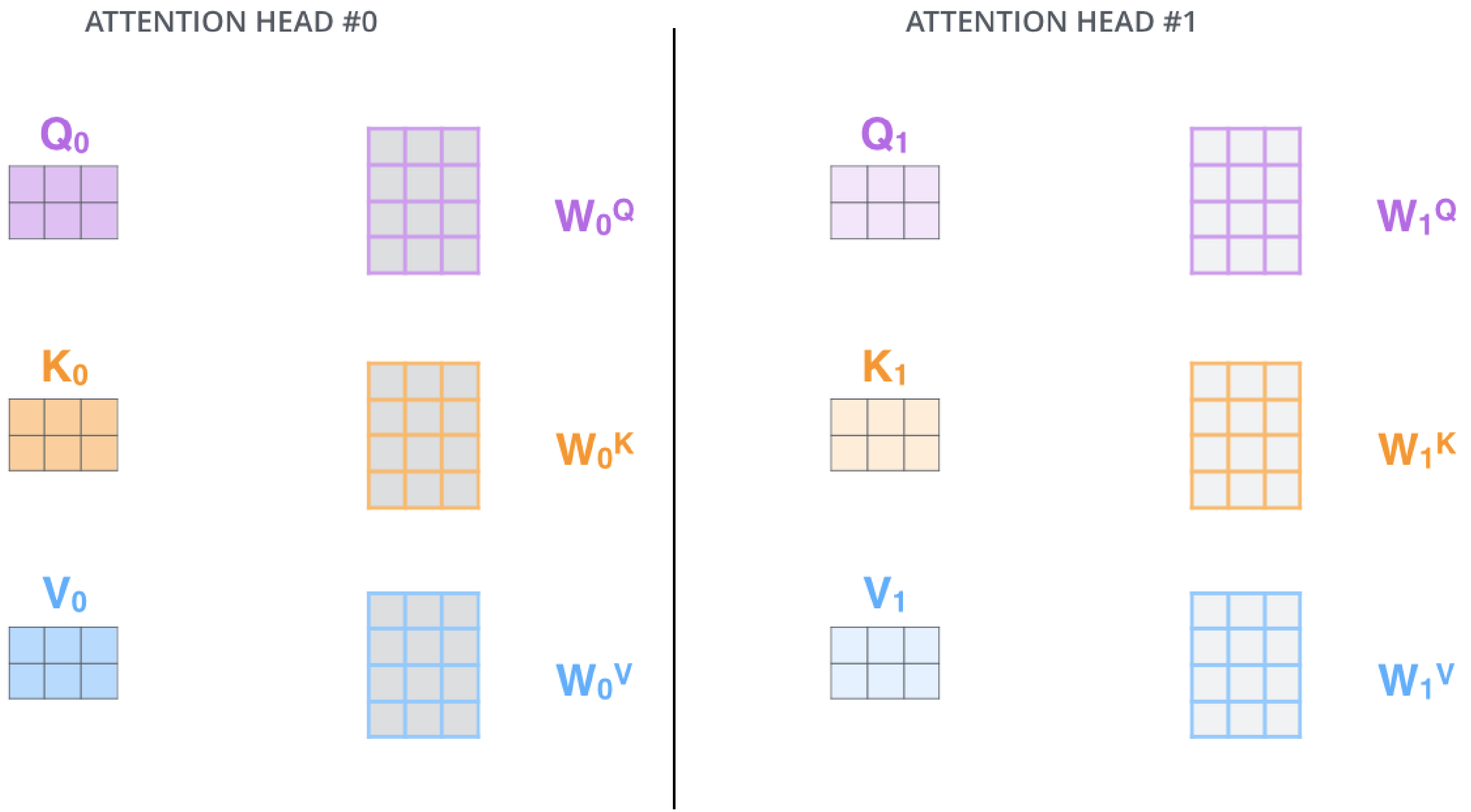




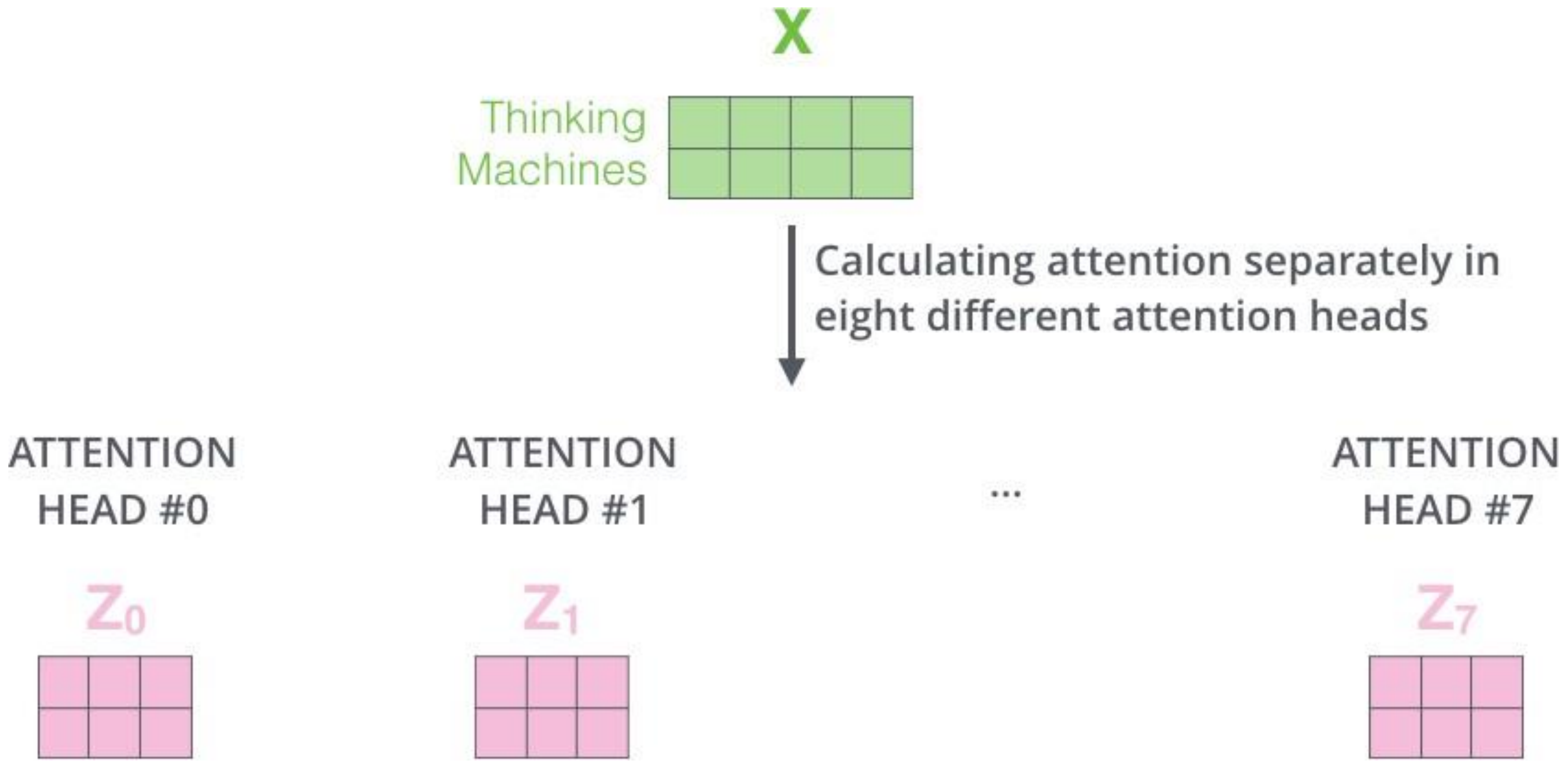
This is called one attention “head”…

Multi-headed attention

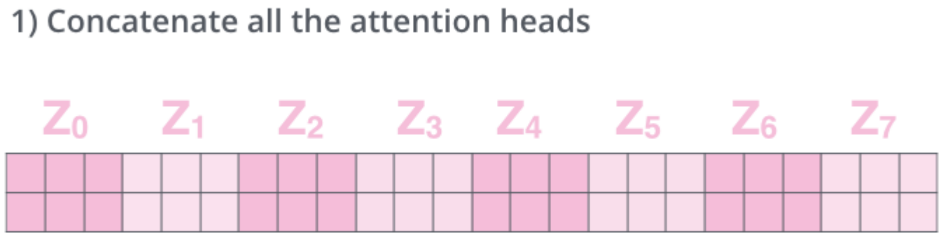
- Or, the beast with many heads



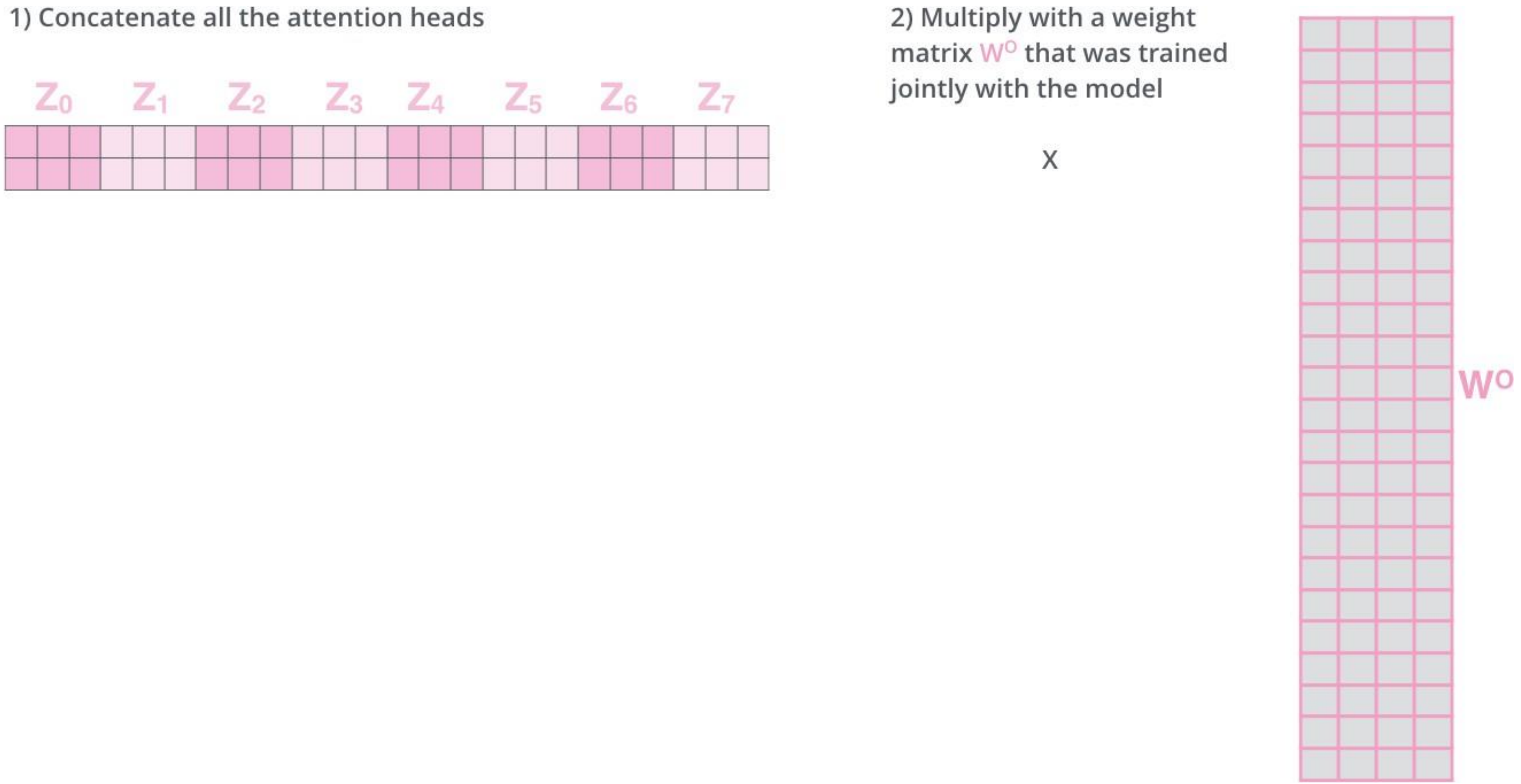
Multi-headed attention



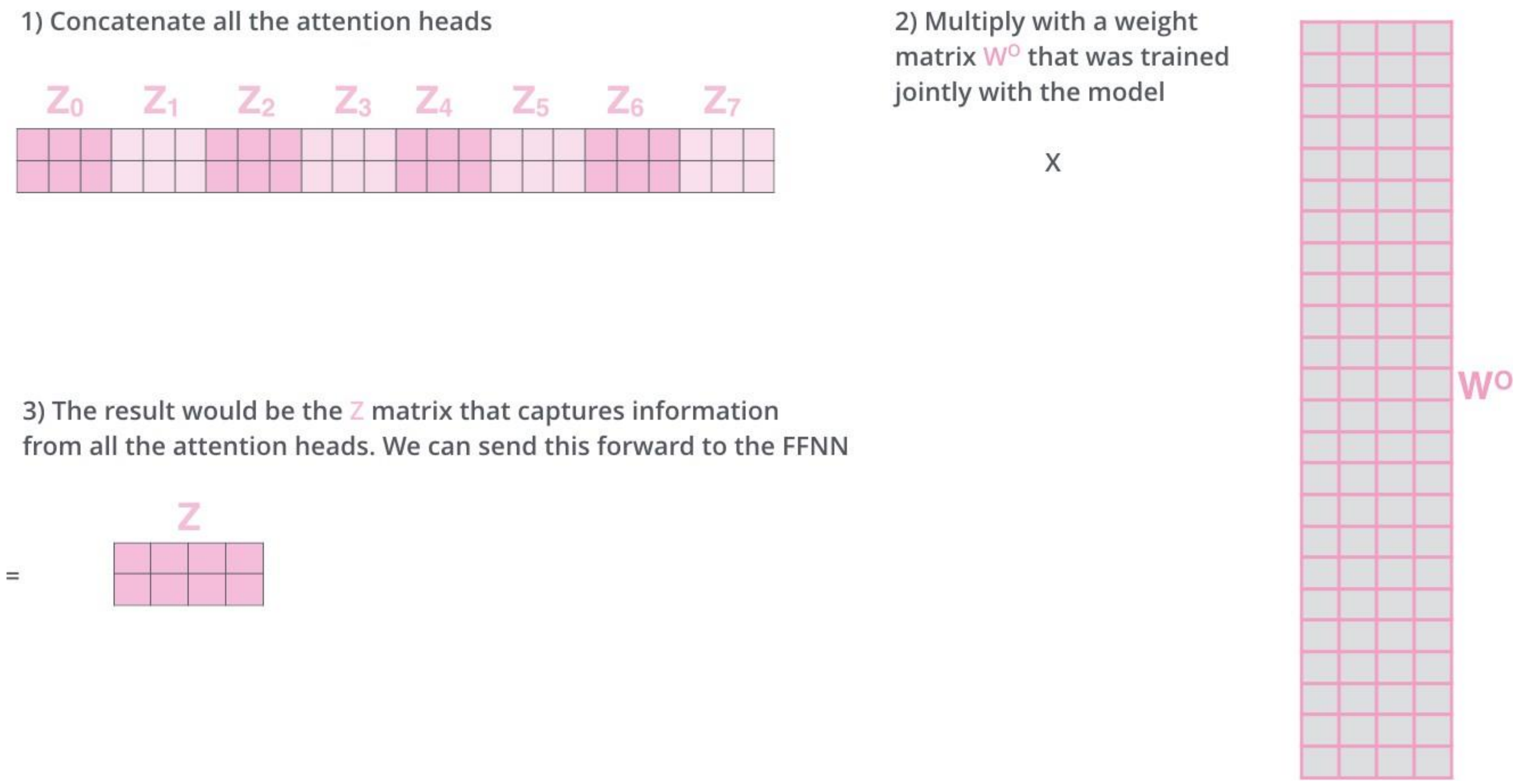
Multi-headed attention



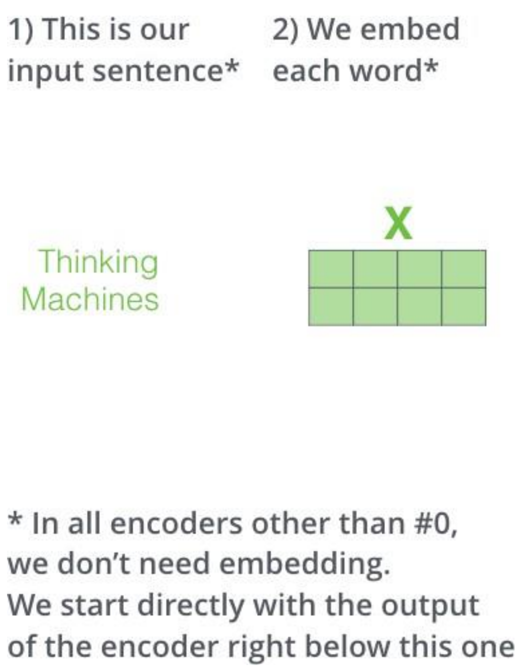
Multi-headed attention



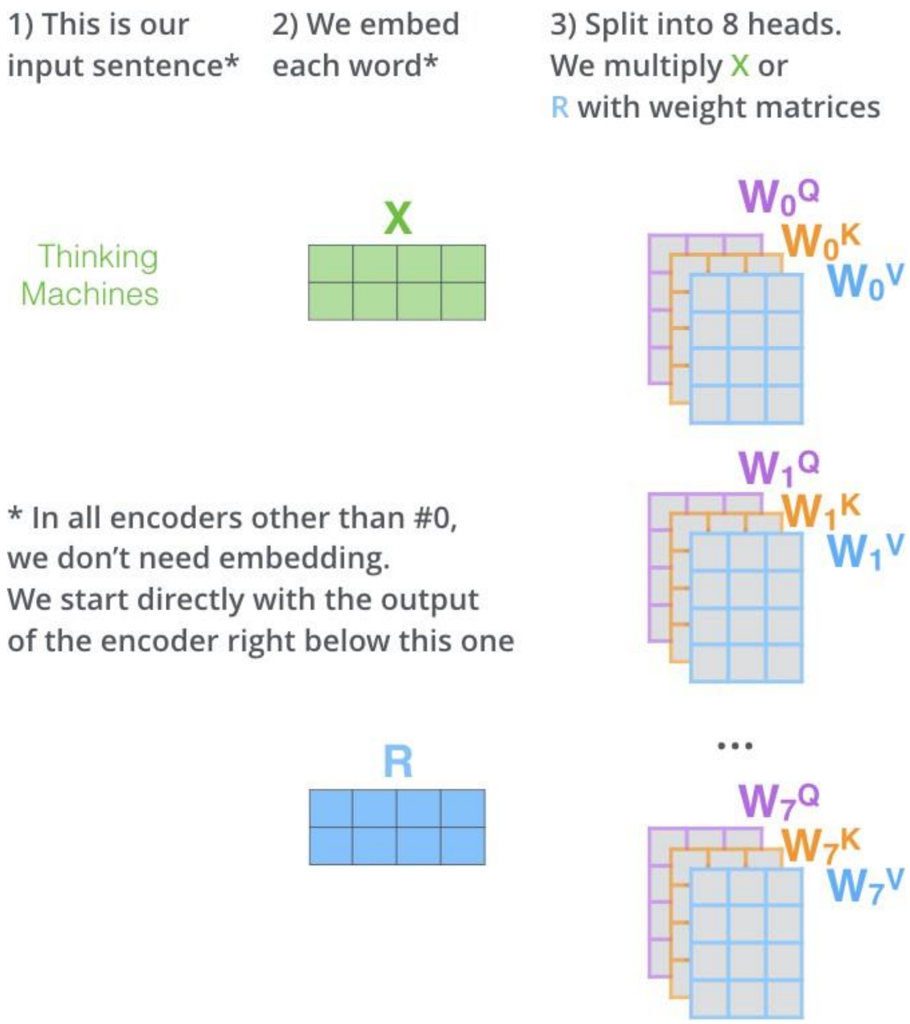
Multi-headed attention



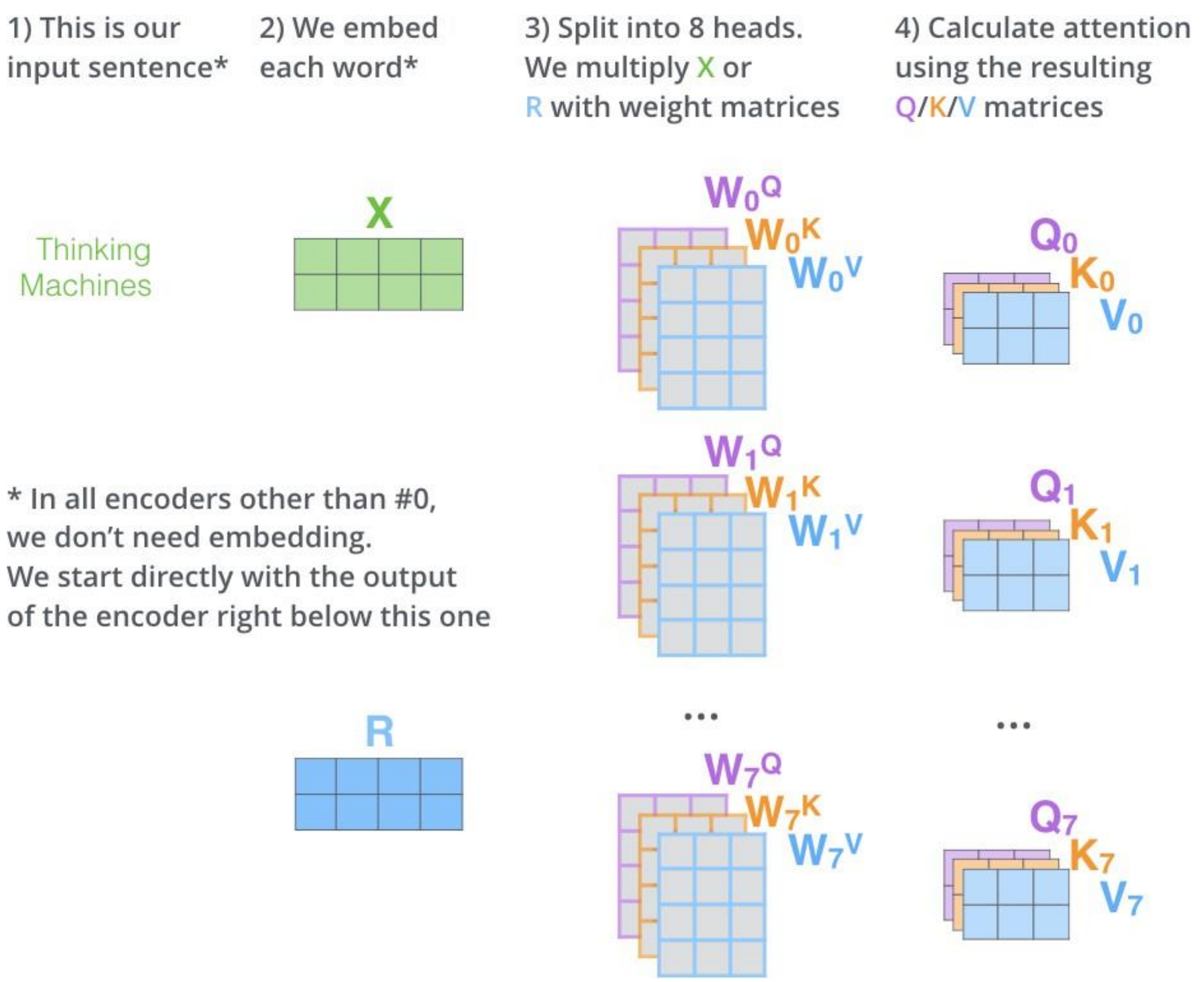
Multi-headed attention



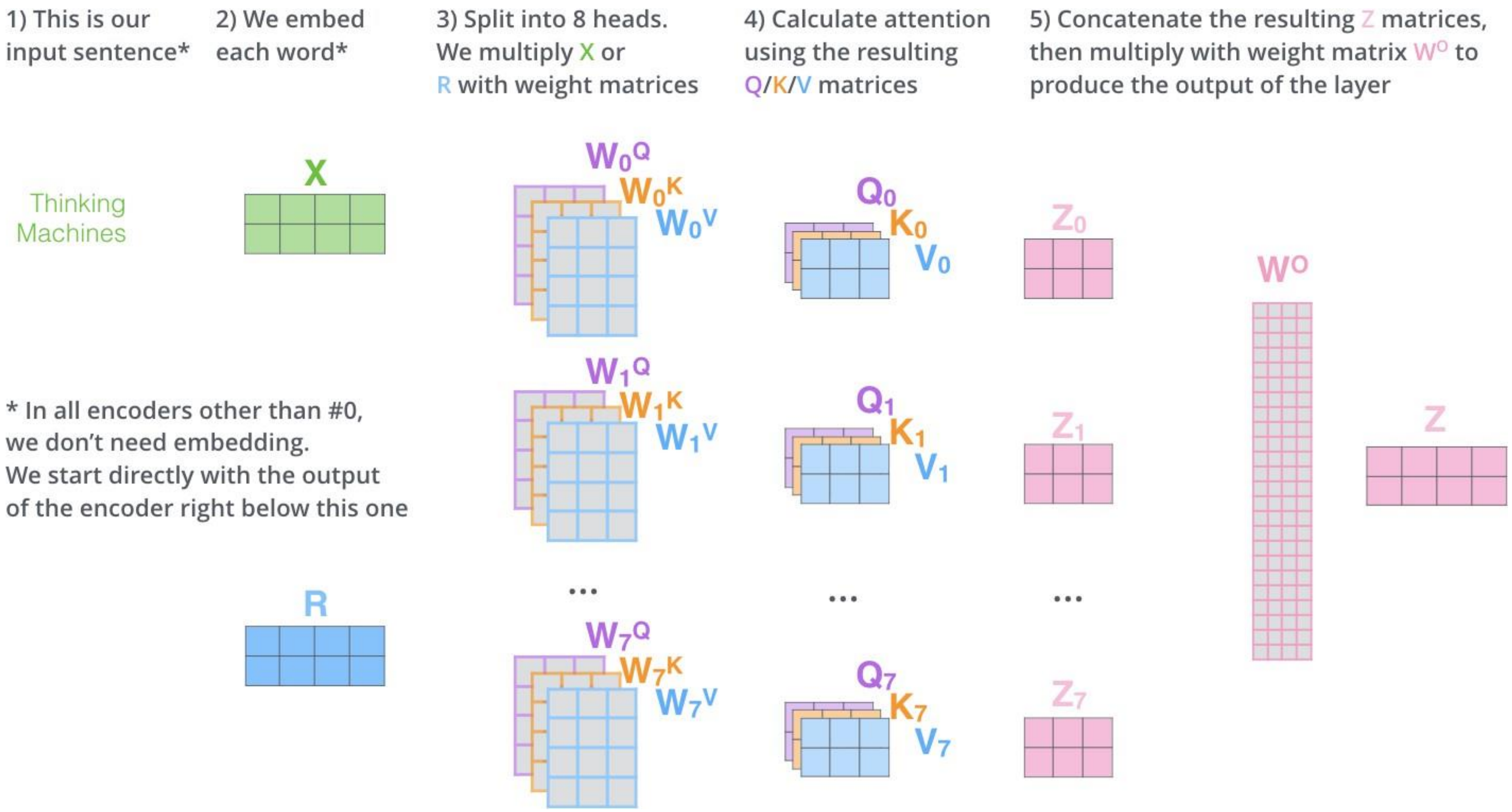
Multi-headed attention



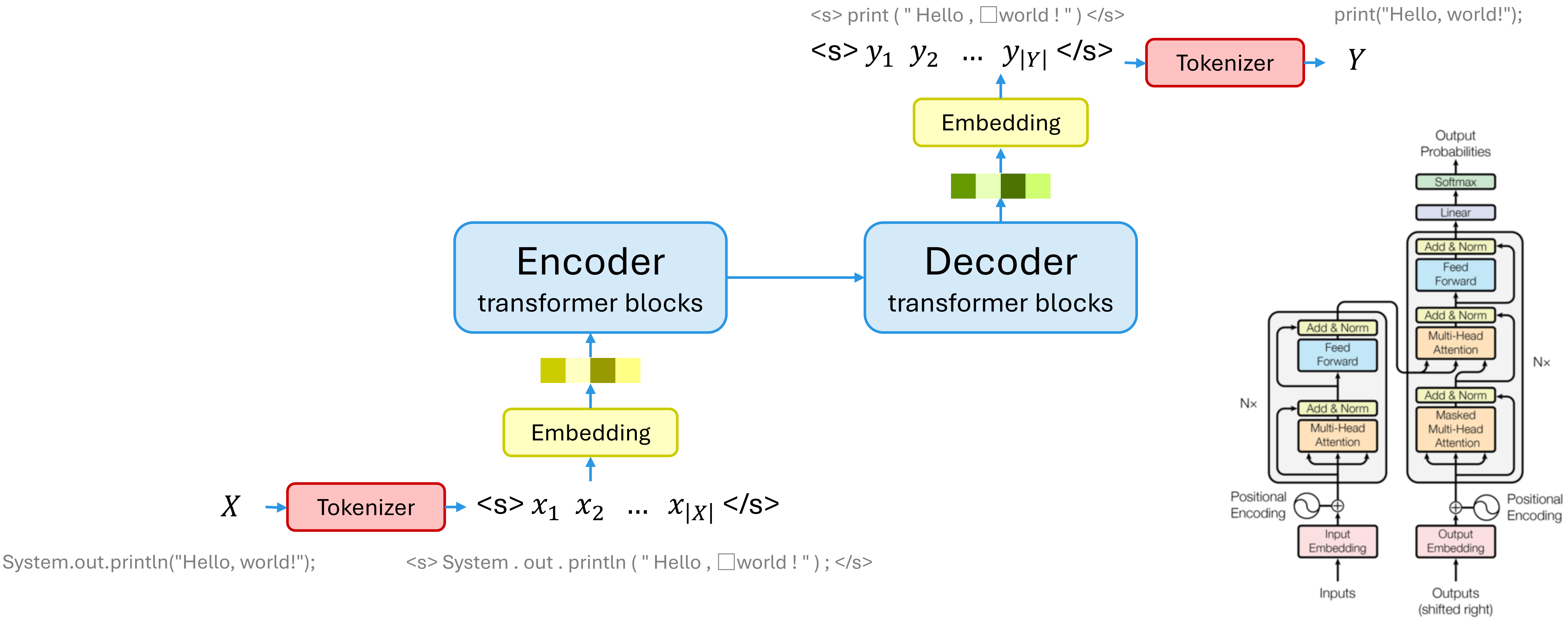
Multi-headed attention



Multi-headed attention



Recap



<s> print (" Hello , □world ! ") </s>
print("Hello, world!");
<s> y_1 y_2 … $y_{|Y|}$ </s>
Tokenizer
Y
Embedding
Encoder
transformer blocks
Decoder
transformer blocks
Embedding
X
Tokenizer
<s> x_1 x_2 … $x_{|X|}$ </s>
System.out.println("Hello, world!");
<s> System . out . println (" Hello , □world ! ") ; </s>
Output Probabilities
Softmax
Linear
Add & Norm
Feed Forward
Add & Norm
Multi-Head Attention
Add & Norm
Masked Multi-Head Attention
Add & Norm
Feed Forward
Add & Norm
Multi-Head Attention
N×
N×
Positional Encoding
Positional Encoding
Input Embedding
Output Embedding
Inputs
Outputs (shifted right)

Variants: Encoder-only and Decoder-only

- Encoder-decoder: T5, BART
  - good for transduction tasks
- Encoder-only: BERT
  - good for classification tasks
- Decoder-only: GPT, Llama
  - good for generation tasks