
# CS 498: Machine Learning System Spring 2026

Minjia Zhang

The Grainger College of Engineering

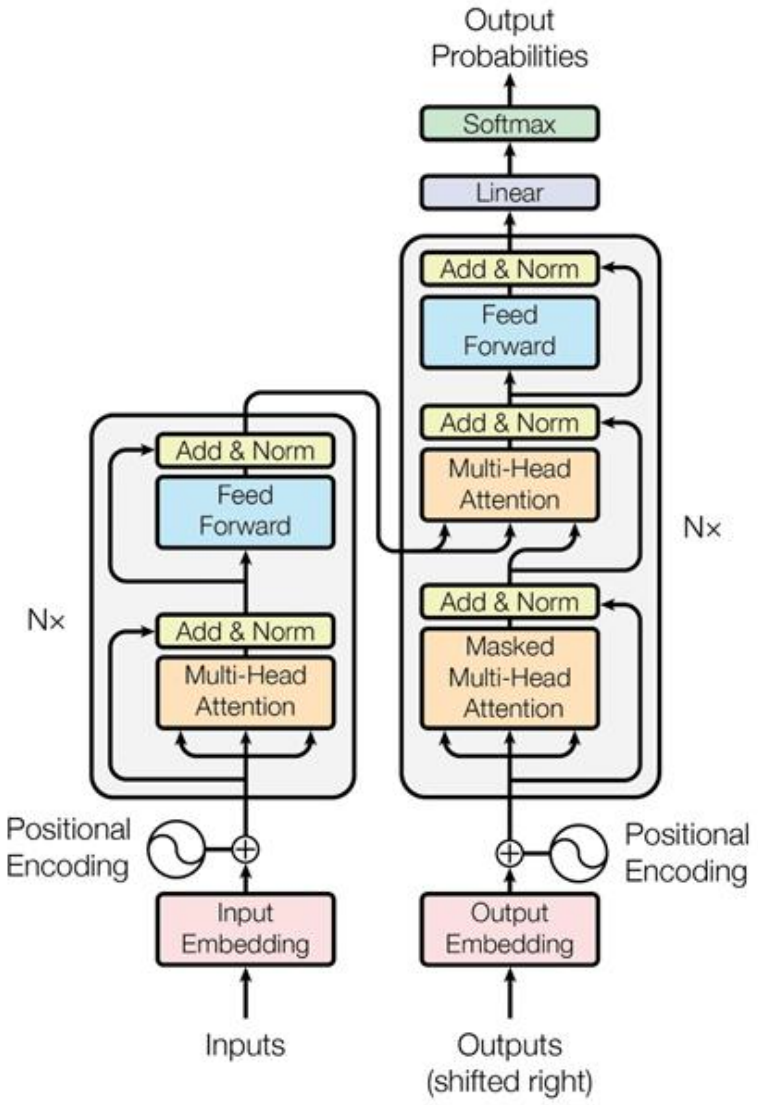
# Today

- **Transformers Deep Dive**
- **Learning Objectives**
  - Explain the core computation and dataflow of Transformers

# Transformers

- Tokenization
- Input embeddings
- Position encodings
- Query, Key, Value
- Attention
- Self-attention
- Multi-head attention
- Feed forward
- Add & norm
- Residual

- Masked attention
- Causal attention
- Linear
- Softmax
- Encoders
- Decoder
- Encoder-decoder



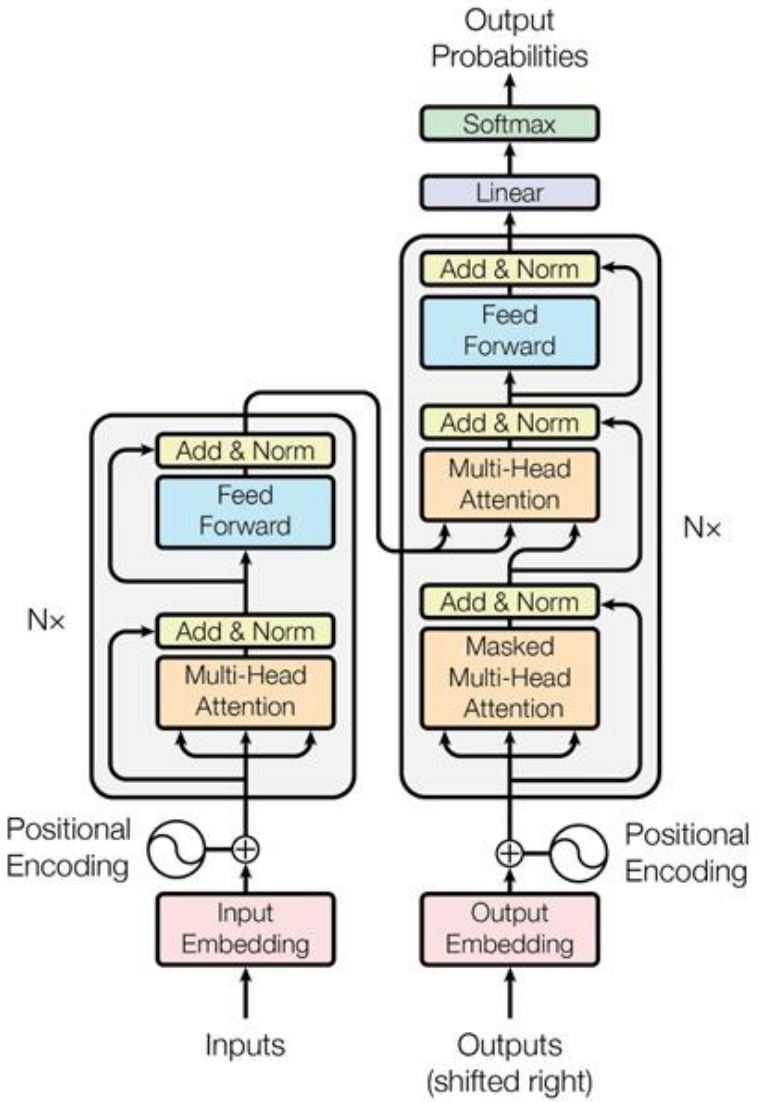
# Machine Translation

Targets

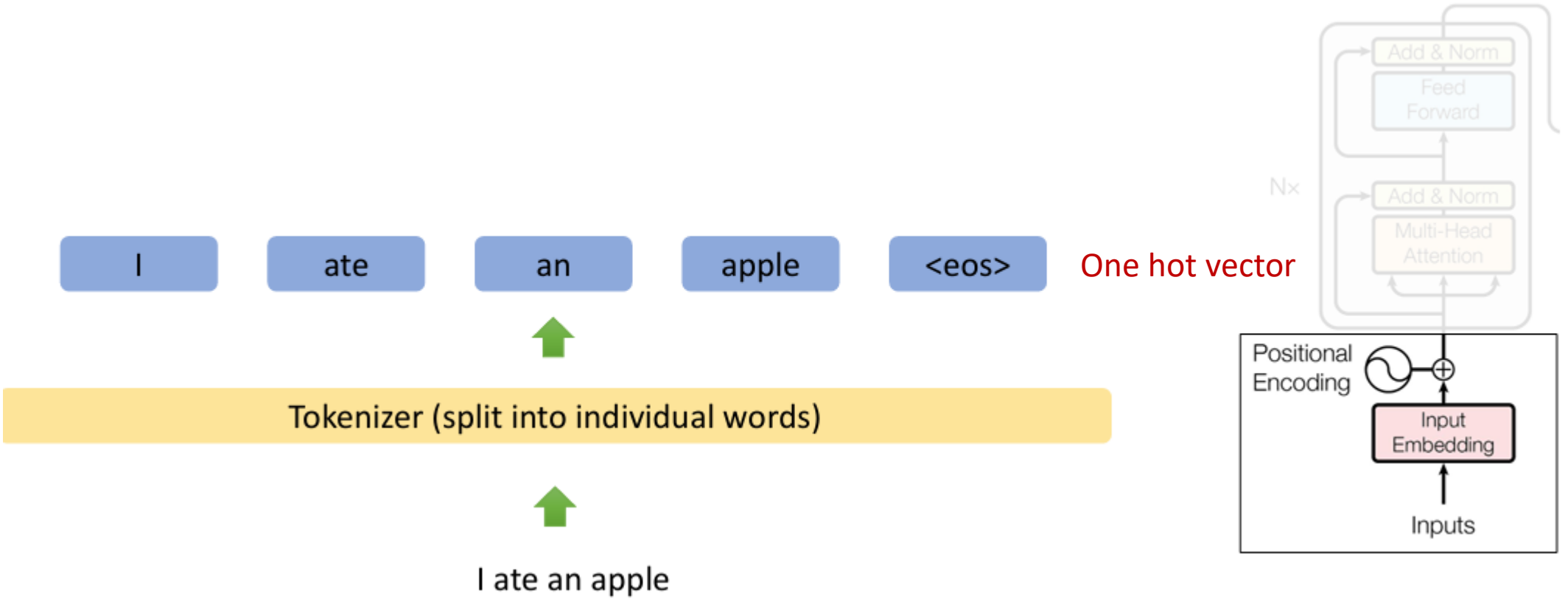
Ich habe einen Apfel gegessen

Inputs

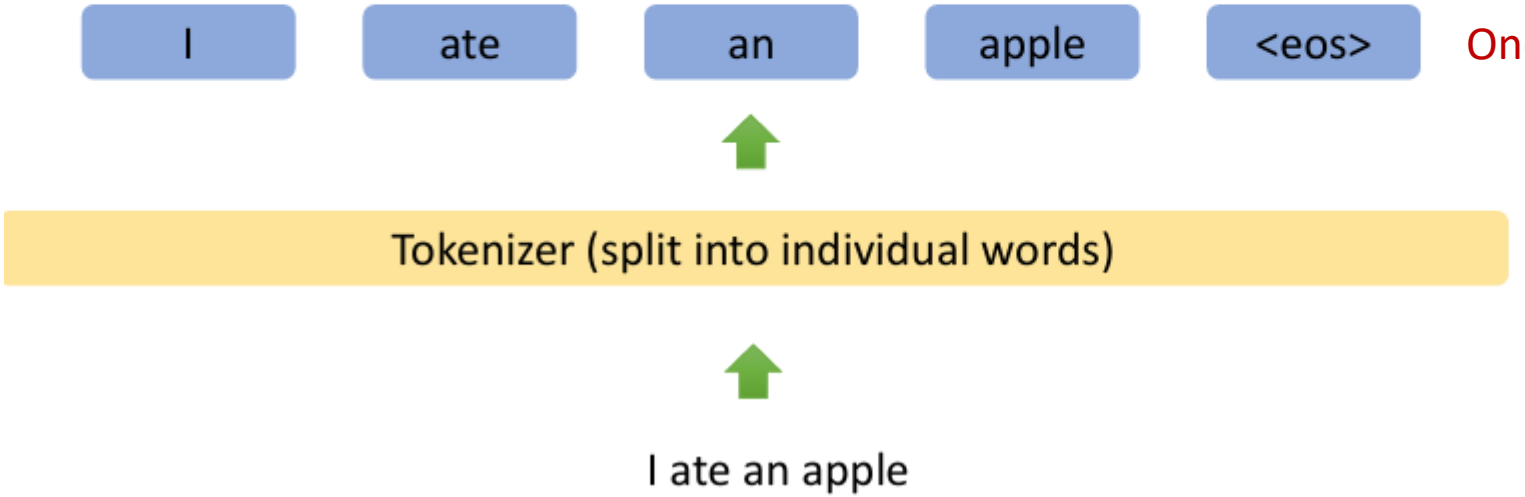
I ate an apple



# Tokenization



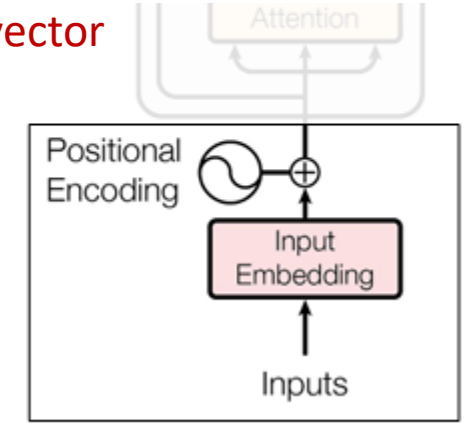
# Tokenization



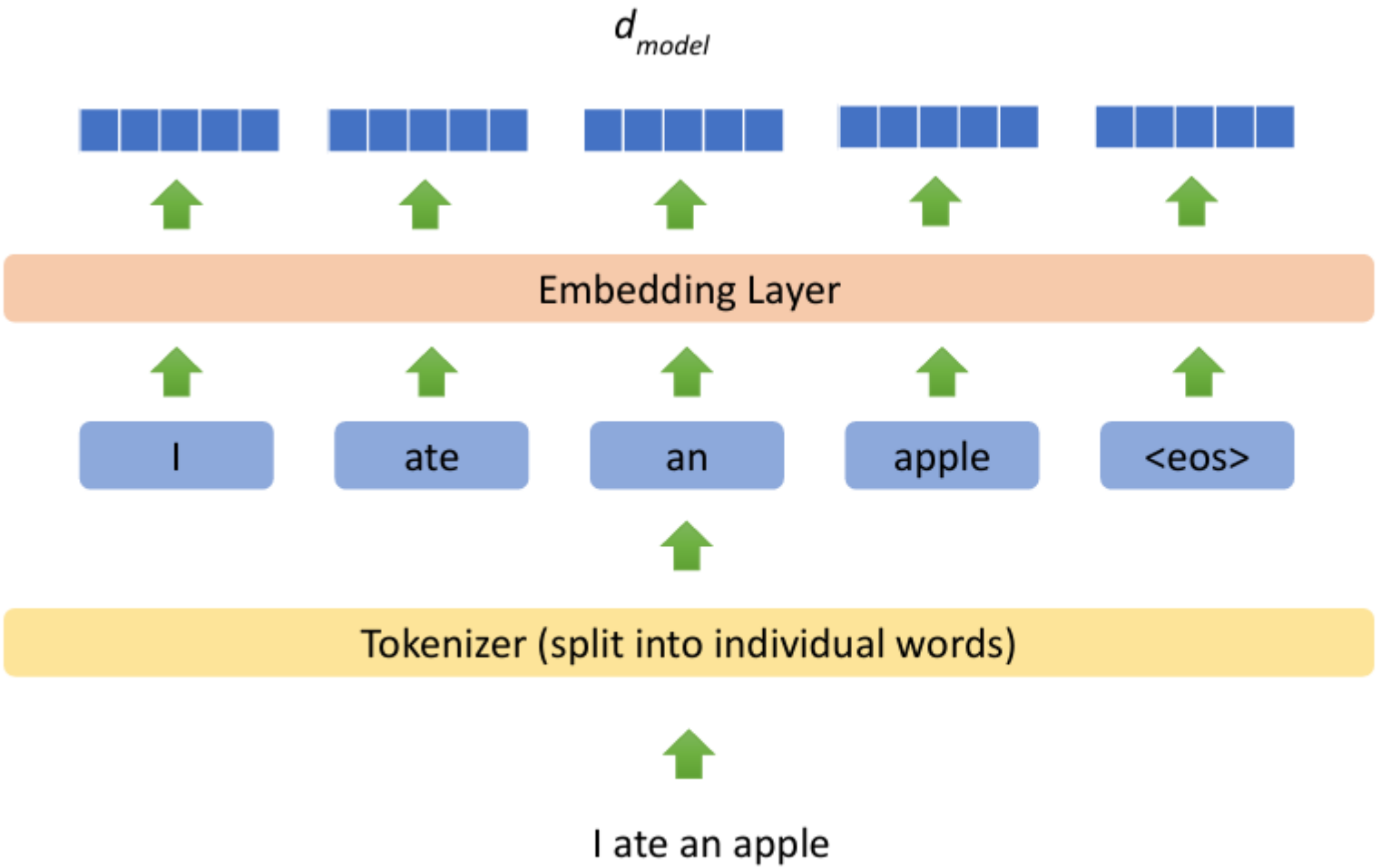

**One-hot encoding**

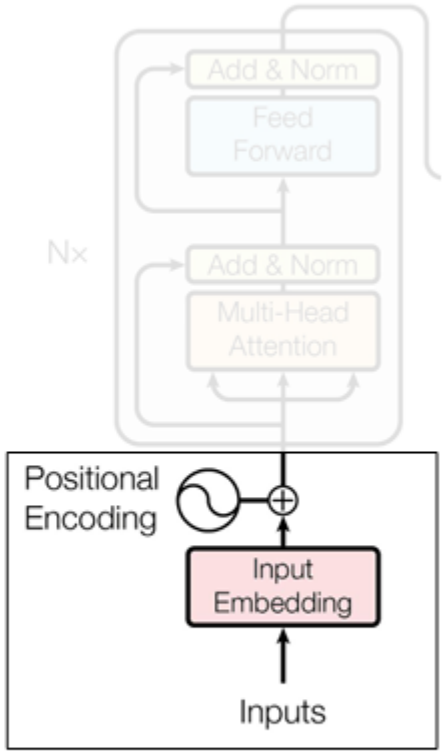
| | cat | mat | on | sat | the |
|---|---|---|---|---|---|
| **the** => | 0 | 0 | 0 | 0 | 1 |
| **cat** => | 1 | 0 | 0 | 0 | 0 |
| **sat** => | 0 | 0 | 0 | 1 | 0 |
| ... | | | | | |

One hot vector



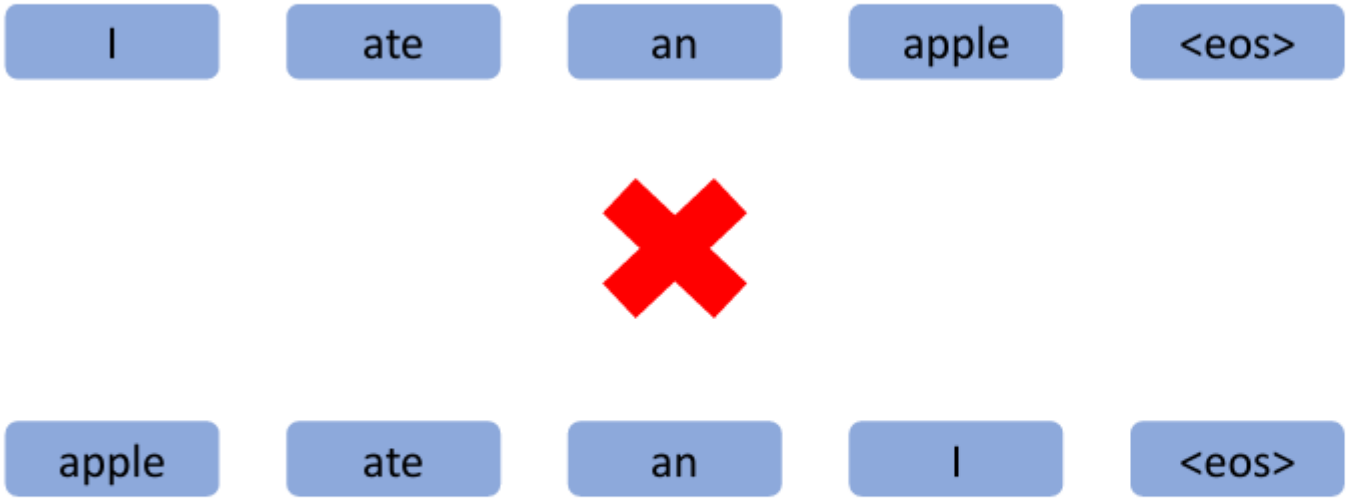
# Input Embeddings

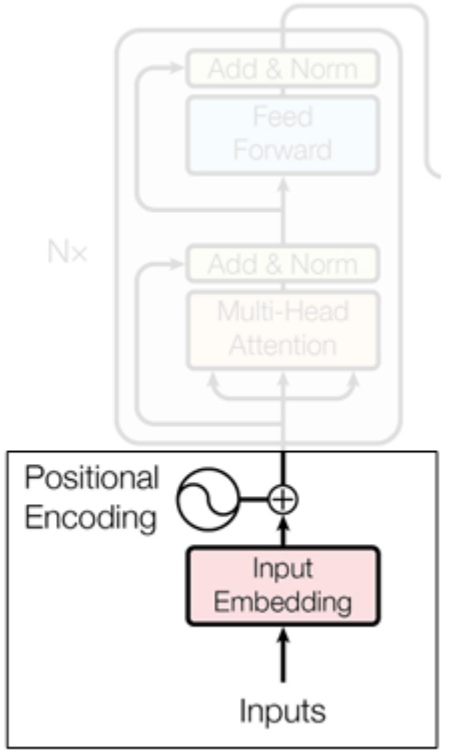



Generate Input Embeddings

# Position Matters

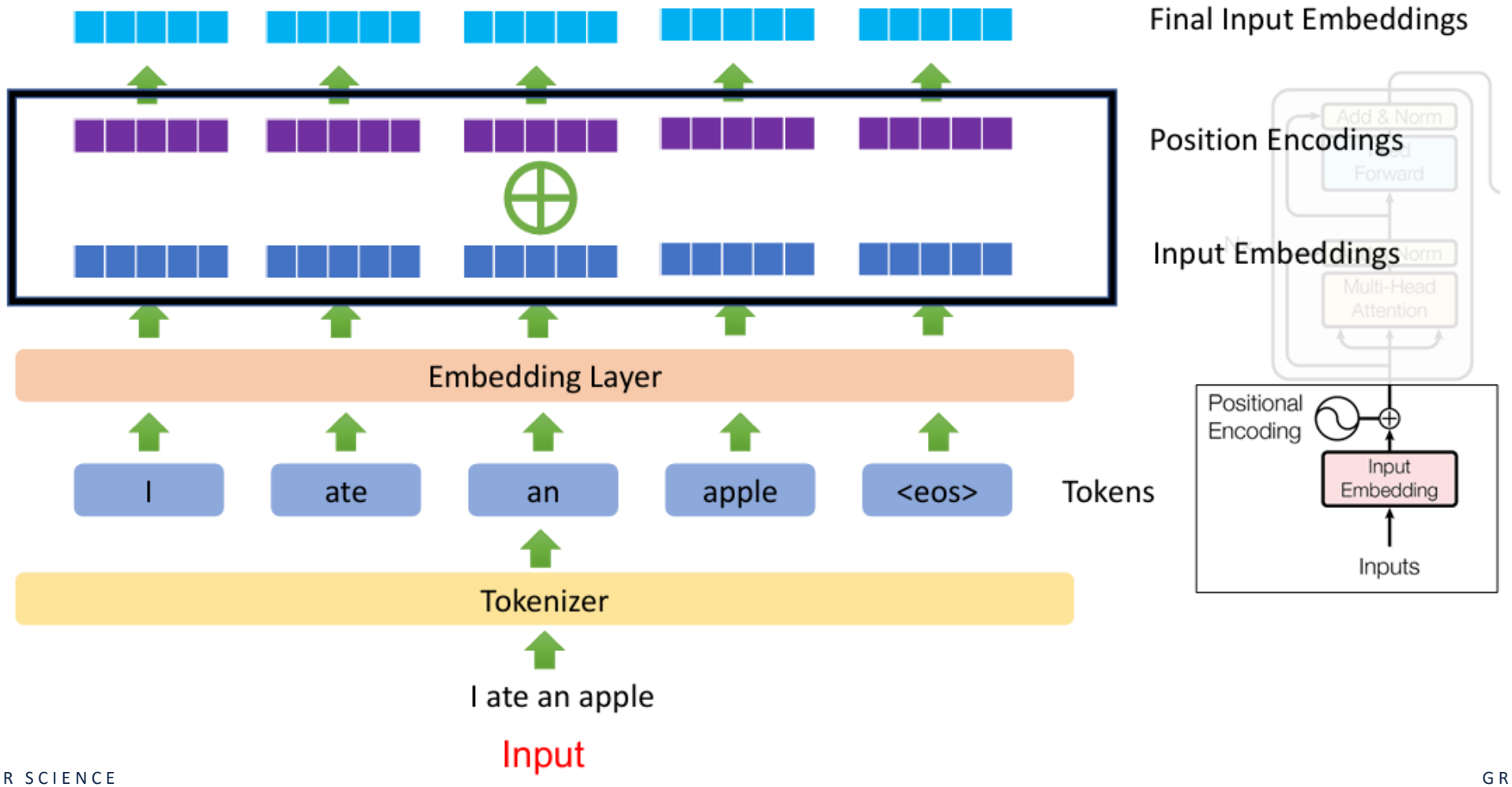



- The position of words in a sentence carries information!
- **Idea:** Add some information to the representation at the beginning that indicates where it is in the sentence

# Position Encodings

Absolute positional encoding (original Transformer): The token embeddings and the absolute position embedding are added together element-wise.



# Position Encodings

- The original position embedding in Transformer is not ideal, because absolute position is less important than relative position

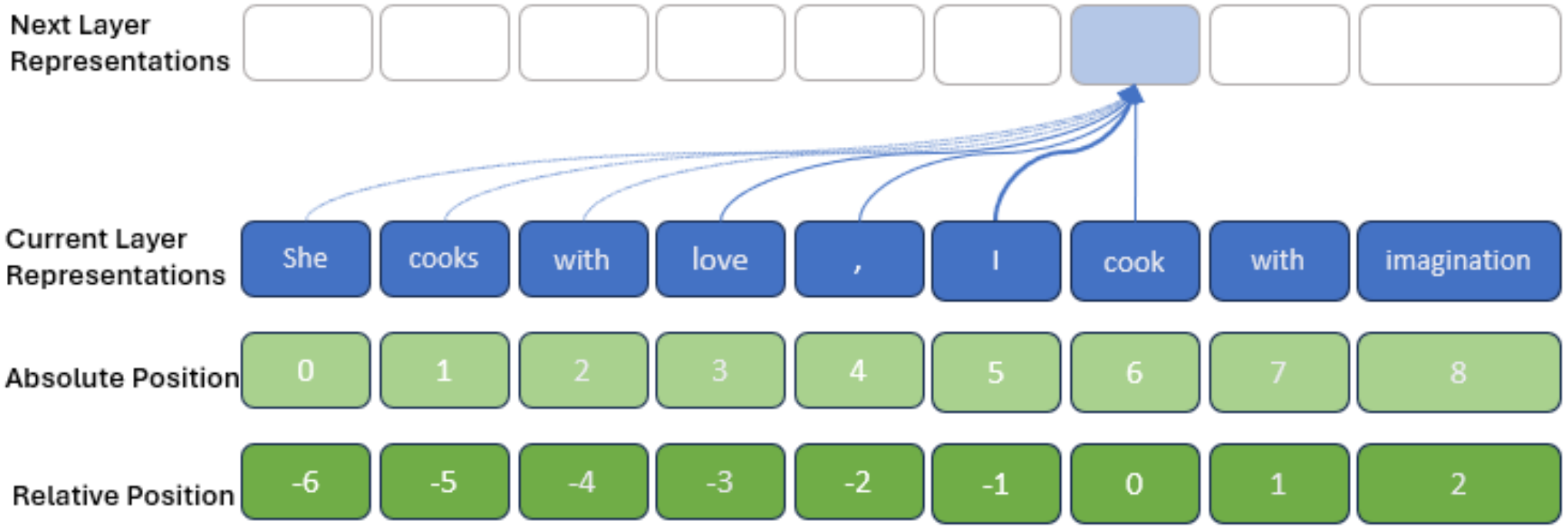
I walk my dog every day

every single day I walk my dog

The fact that "my dog" is right after "I walk" is the important part, not its absolute position

# Relative Position Encodings

- Translation invariance to position information
- Lead to performance improvements



# Rotary Position Embedding

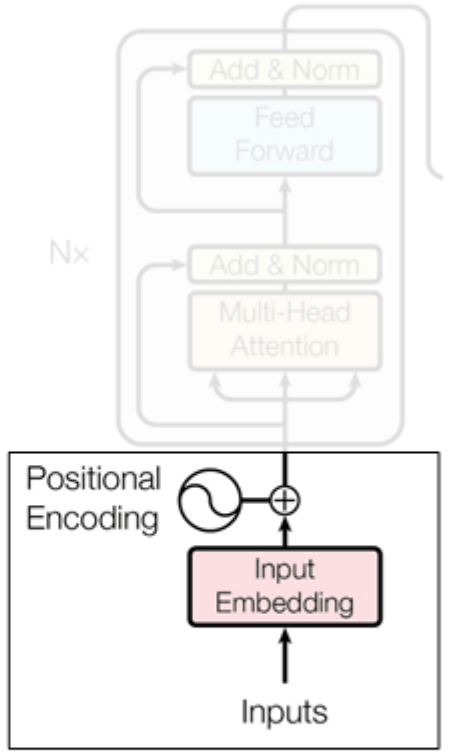
- Most recent LLMs adopt RoPE

ROFORMER: ENHANCED TRANSFORMER WITH ROTARY POSITION EMBEDDING

$$f_{\{q,k\}}(\boldsymbol{x}_m, m) = \begin{pmatrix} \cos m\theta & -\sin m\theta \\ \sin m\theta & \cos m\theta \end{pmatrix} \begin{pmatrix} W^{(11)}_{\{q,k\}} & W^{(12)}_{\{q,k\}} \\ W^{(21)}_{\{q,k\}} & W^{(22)}_{\{q,k\}} \end{pmatrix} \begin{pmatrix} x^{(1)}_m \\ x^{(2)}_m \end{pmatrix}$$

Table 2: Comparing RoFormer and BERT by fine tuning on downstream GLEU tasks.

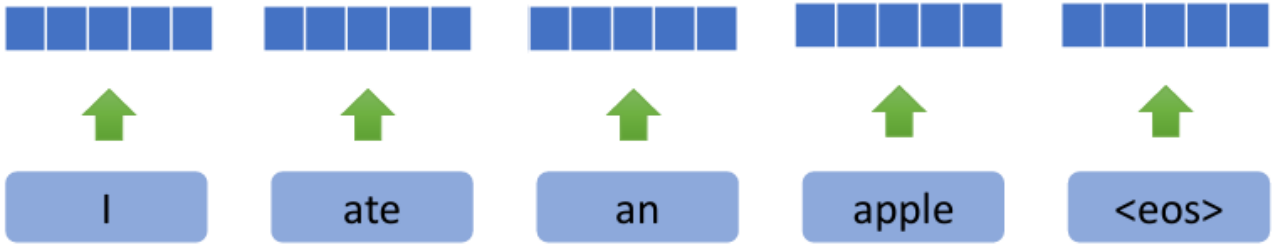
| Model | MRPC | SST-2 | QNLI | STS-B | QQP | MNLI(m/mm) |
|---|---|---|---|---|---|---|
| BERTDevlin et al. [2019] | 88.9 | 93.5 | 90.5 | 85.8 | 71.2 | 84.6/83.4 |
| RoFormer | **89.5** | 90.7 | 88.0 | **87.0** | **86.4** | 80.2/79.8 |

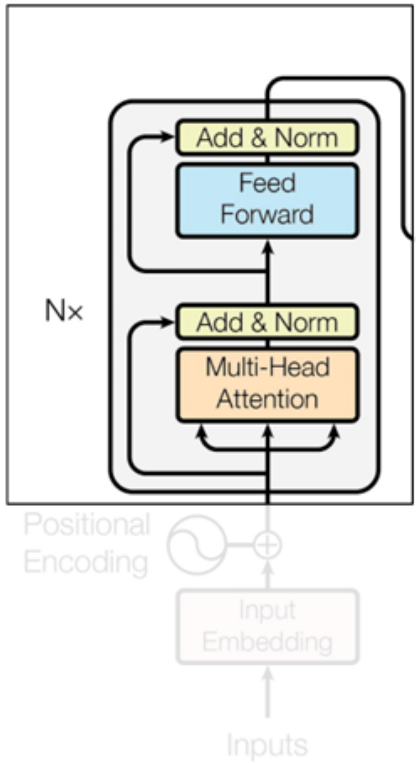



REF: Rotary Position Embeddings

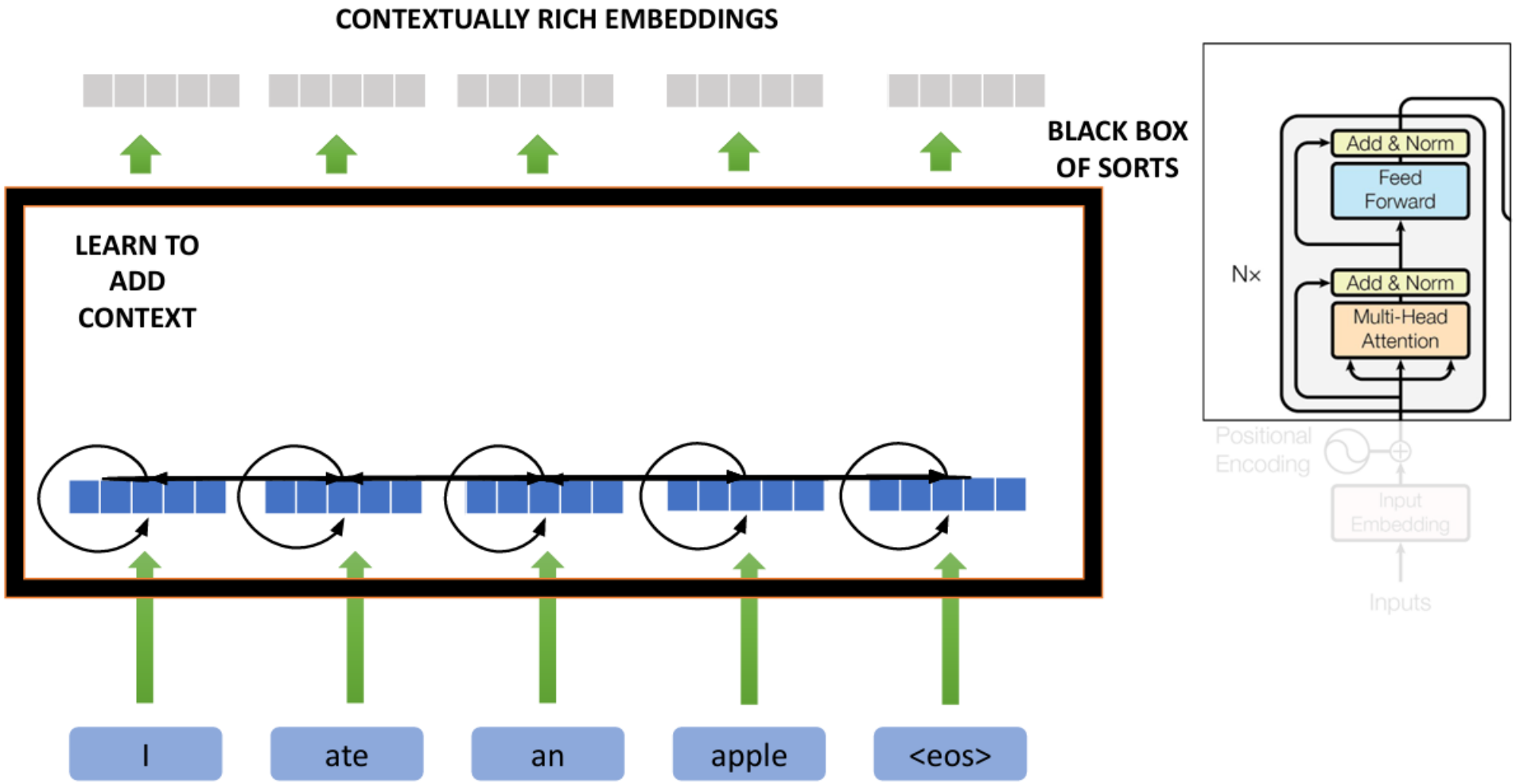
# Encoder

WHERE IS THE CONTEXT ?

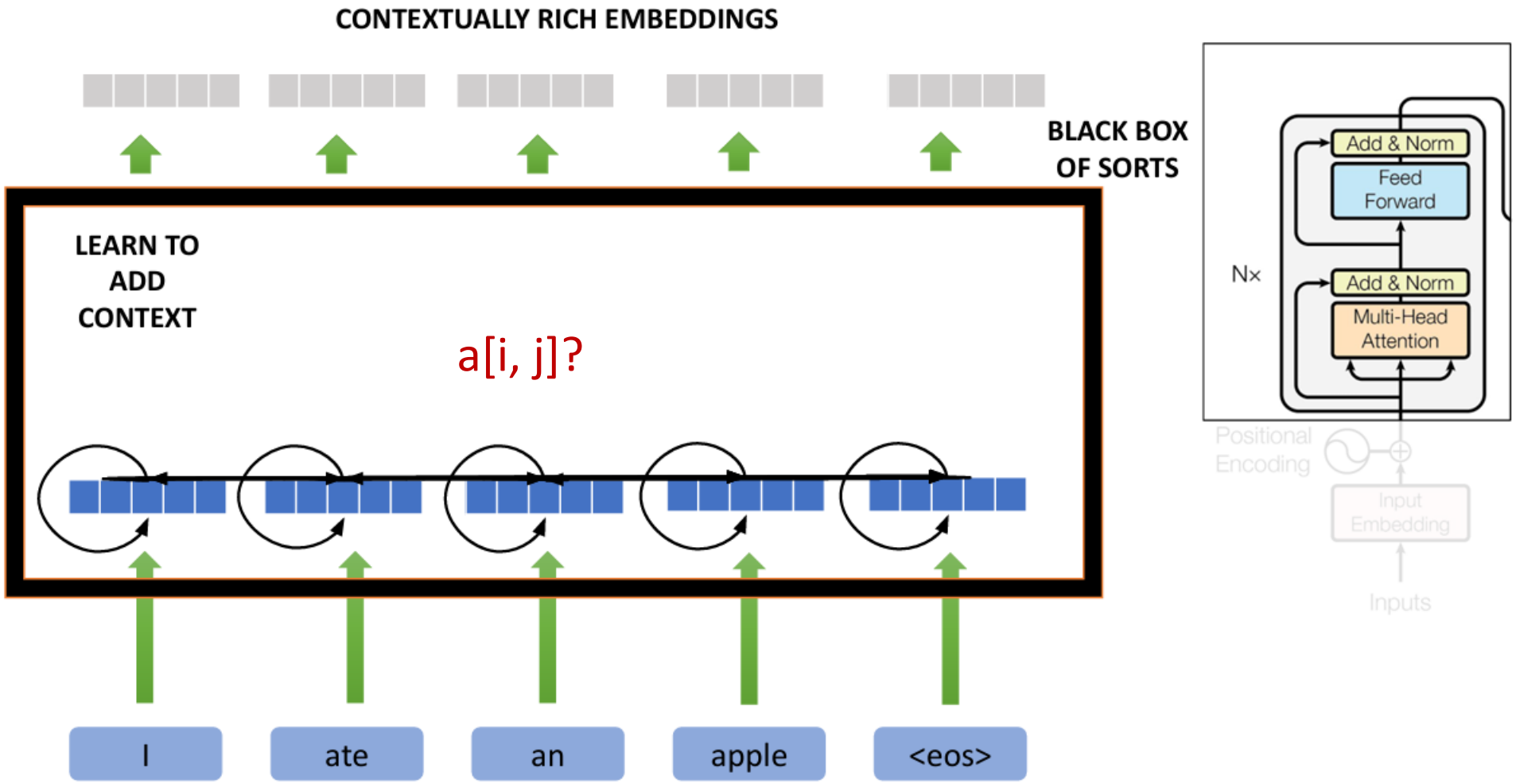


# Encoder


CONTEXTUALLY RICH EMBEDDINGS
BLACK BOX OF SORTS
LEARN TO ADD CONTEXT
I
ate
an
apple
<eos>
Nx
Add & Norm
Feed Forward
Add & Norm
Multi-Head Attention
Positional Encoding
Input Embedding
Inputs

# Encoder

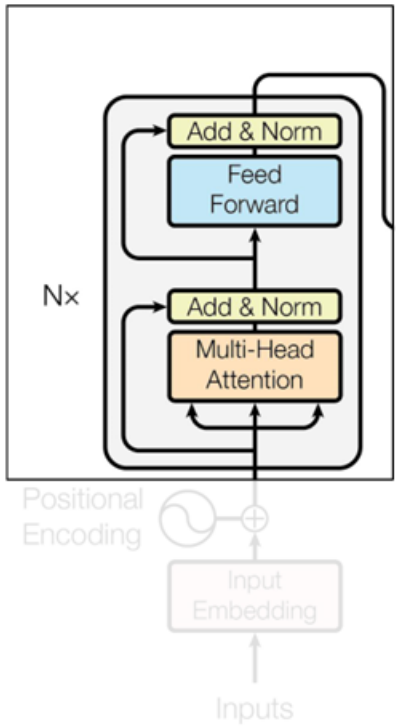

CONTEXTUALLY RICH EMBEDDINGS
BLACK BOX OF SORTS
LEARN TO ADD CONTEXT
a[i, j]?
I
ate
an
apple
<eos>
Nx
Add & Norm
Feed Forward
Add & Norm
Multi-Head Attention
Positional Encoding
Input Embedding
Inputs

# Multi-Head Self-Attention

- Before “multi-head” attention, what is “single head” attention?

$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V$$

- Query
- Key
- Value



# Multi-Head Self-Attention: Query, Key, & Value

- Database {key, value store}

{Query: "Order details of order_104"}

OR

{Query: "Order details of order_106"}


```
{"order_100": {"items":"a1", "delivery_date":"a2", ...}},
{"order_101": {"items":"b1", "delivery_date":"b2", ...}},
{"order_102": {"items":"c1", "delivery_date":"c2", ...}},
{"order_103": {"items":"d1", "delivery_date":"d2", ...}},
{"order_104": {"items":"e1", "delivery_date":"e2", ...}},
{"order_105": {"items":"f1", "delivery_date":"f2", ...}},
{"order_106": {"items":"g1", "delivery_date":"g2", ...}},
{"order_107": {"items":"h1", "delivery_date":"h2", ...}},
{"order_108": {"items":"i1", "delivery_date":"i2", ...}},
{"order_109": {"items":"j1", "delivery_date":"j2", ...}},
{"order_110": {"items":"k1", "delivery_date":"k2", ...}}
```


**Query**

1. Search for info

**Key**

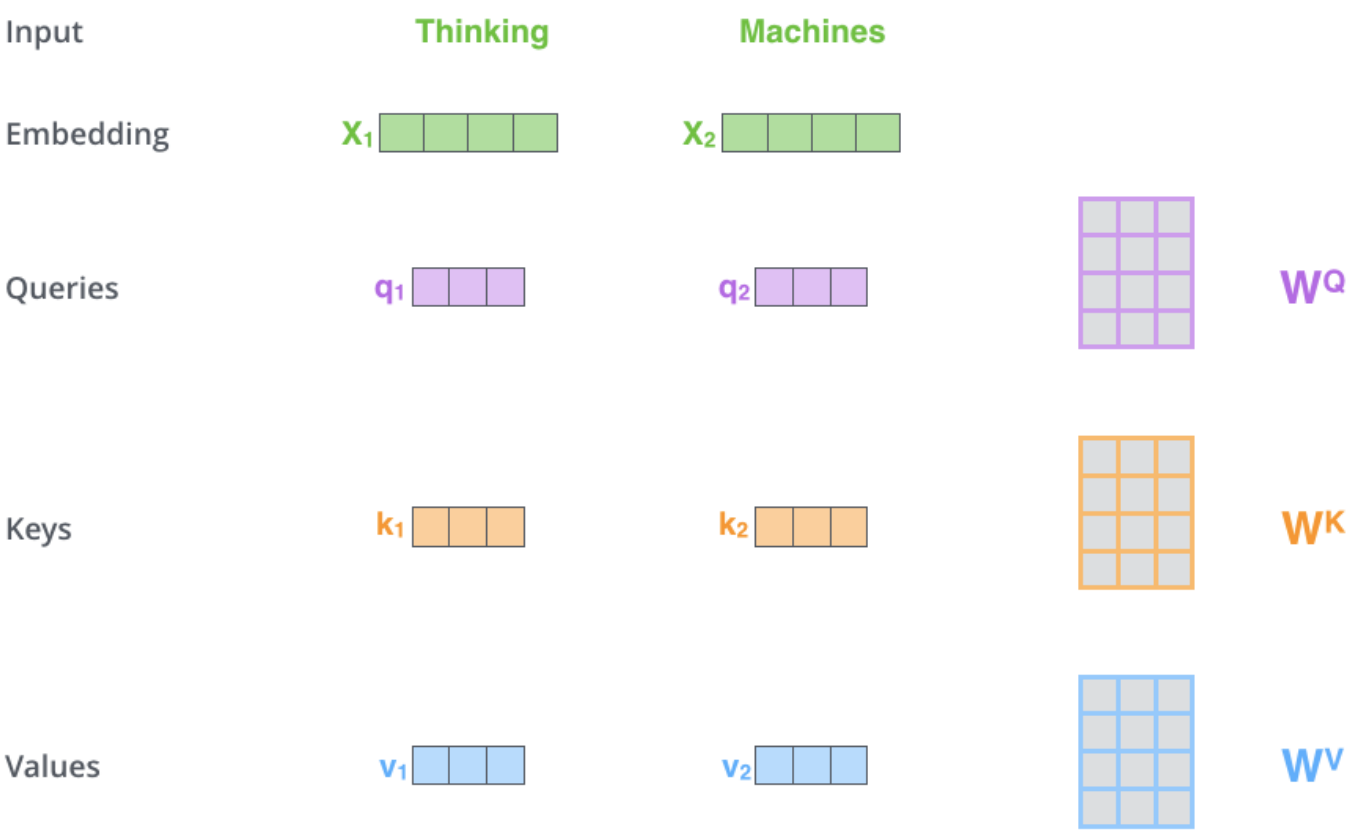
1. Interacts directly with Queries
2. Distinguishes one object from another
3. Identify which object is the most relevant and by how much

**Value**

1. Actual details of the object
2. More fine grained

# Self-Attention

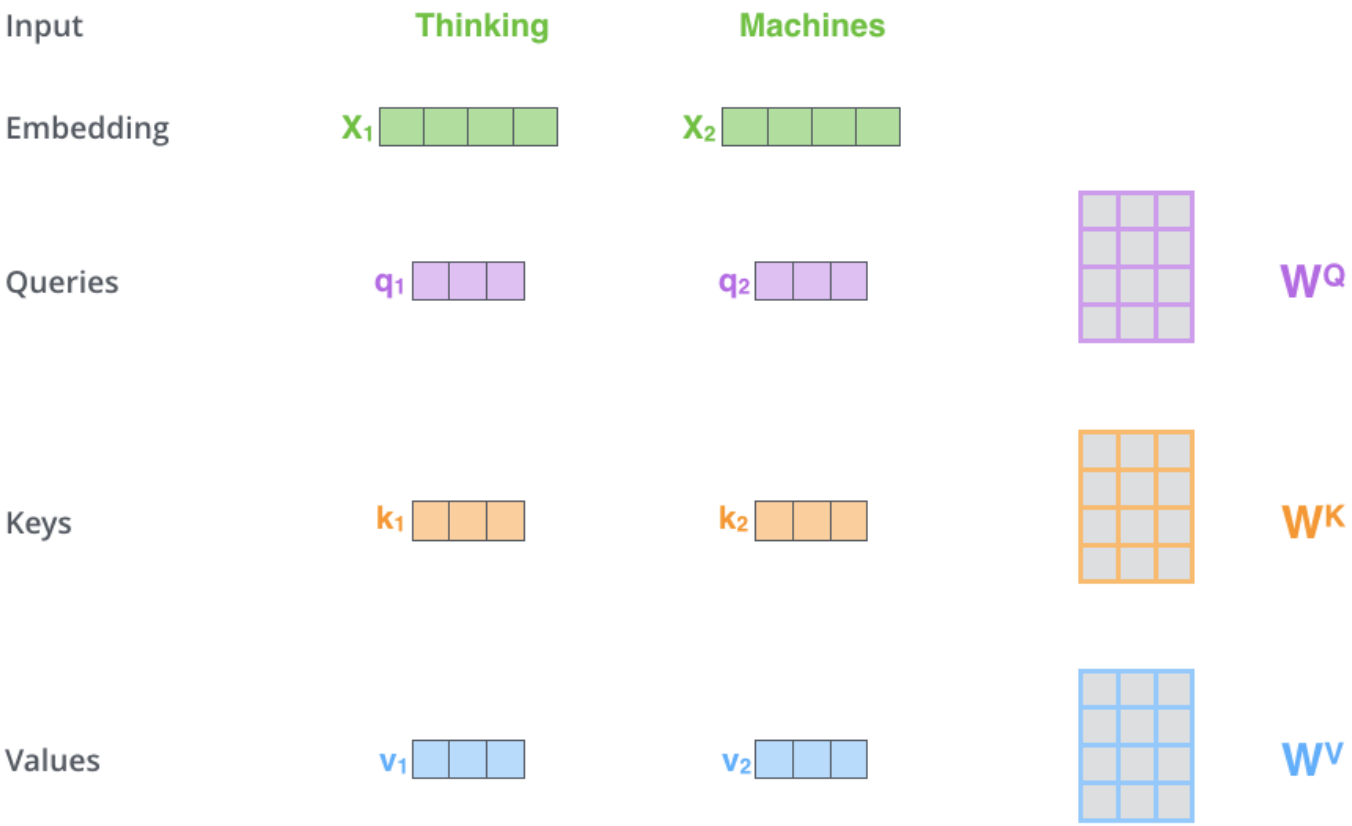
- Step 1: compute “key, value, query” embedding for each input token

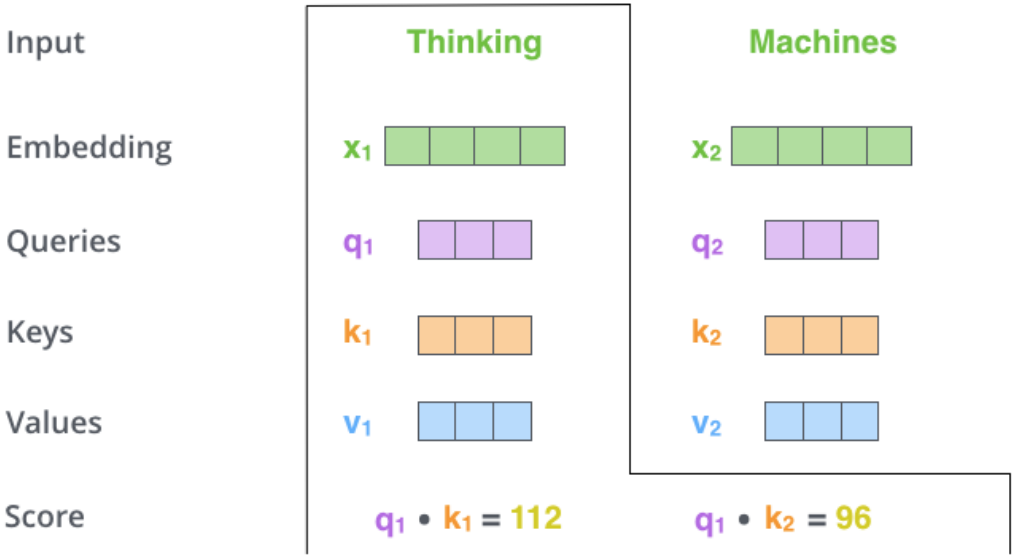



[The Illustrated Transformer – Jay Alammar – Visualizing machine learning one concept at a time.]

# Self-Attention

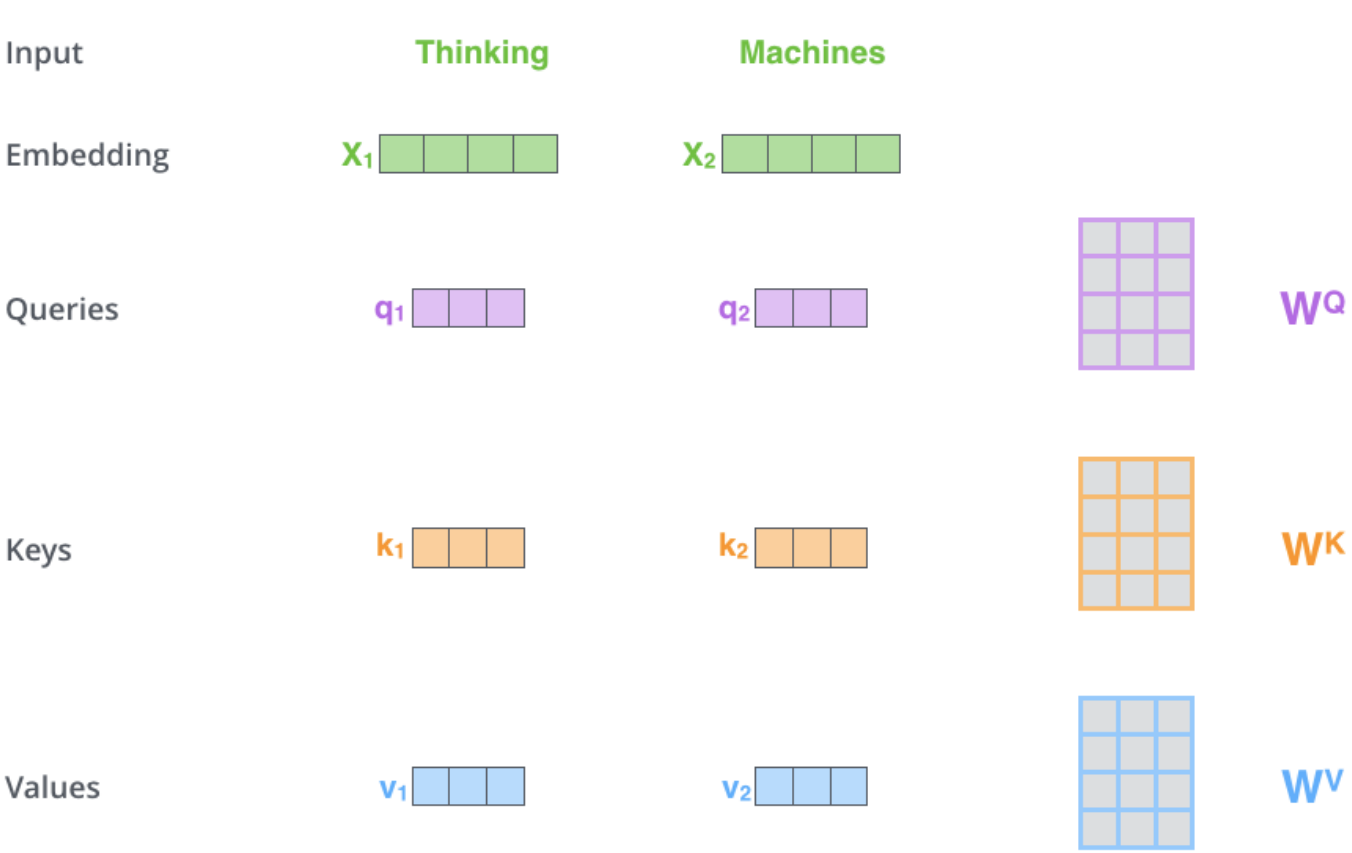
- Step 1: compute “key, value, query” embedding for each input token
- Step 2: compute scores between pairs of tokens

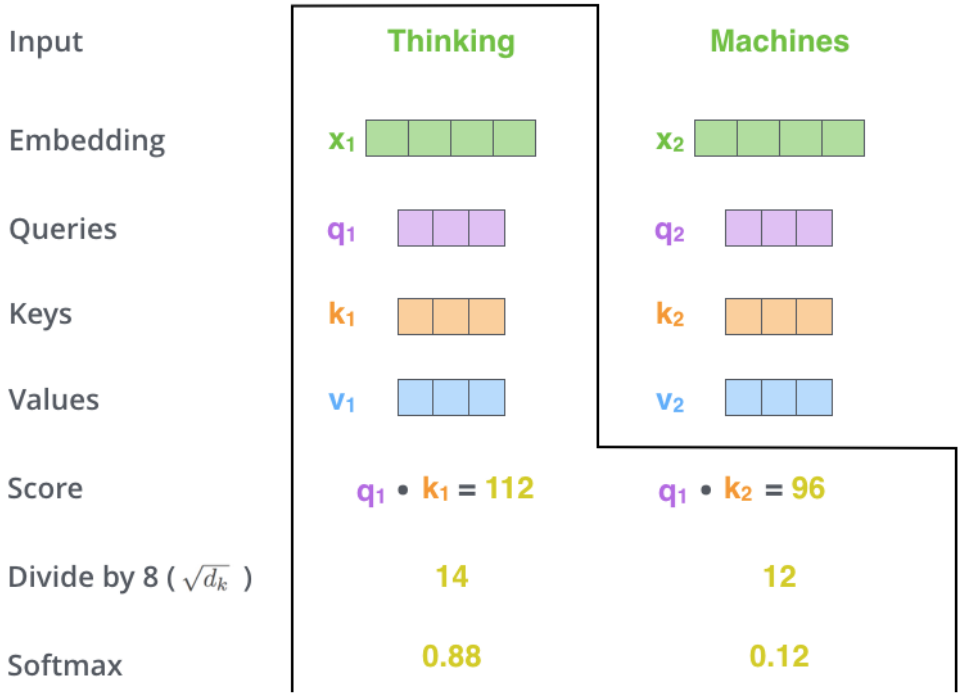


# Self-Attention

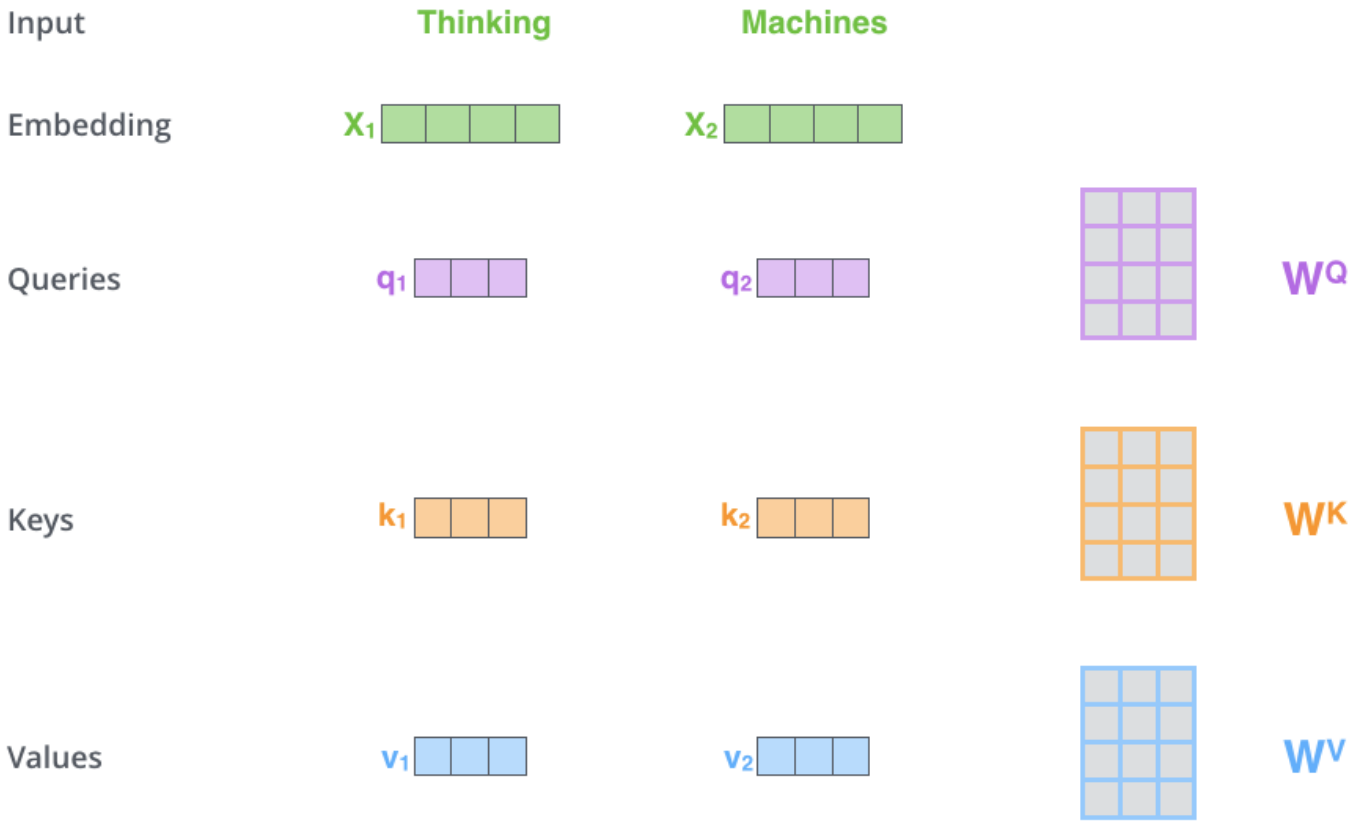
- Step 1: compute “key, value, query” embedding for each input token
- Step 2: compute scores between pairs of tokens
- Step 3: compute normalized attention scores

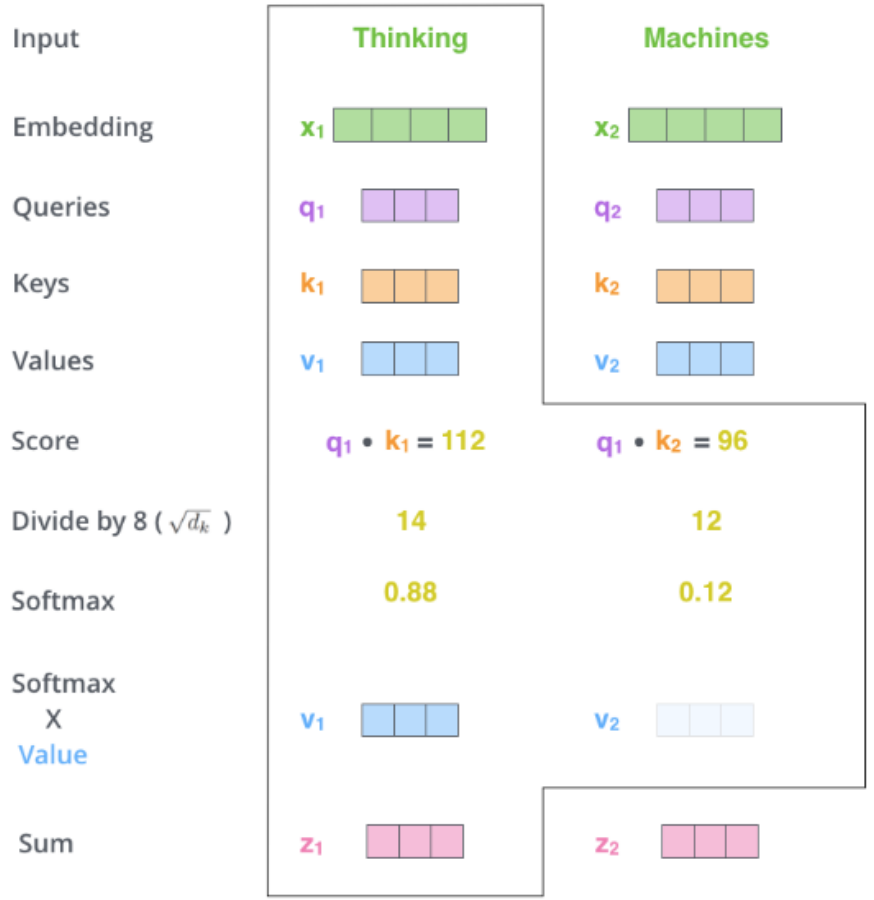


# Self-Attention

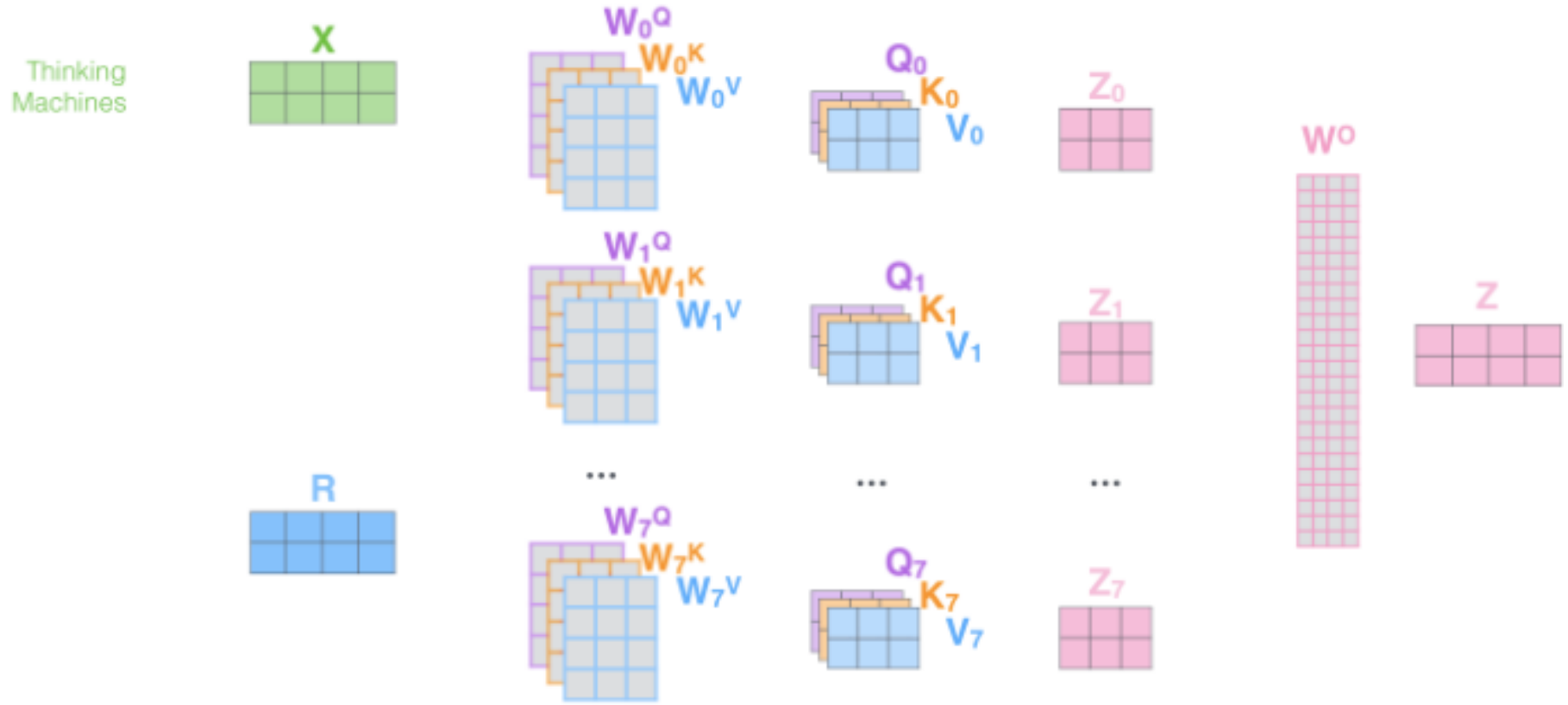
- Step 1: compute “key, value, query” embedding for each input token
- Step 2: compute scores between pairs of tokens
- Step 3: compute normalized attention scores
- Step 4: get new representation by weighted sum of values



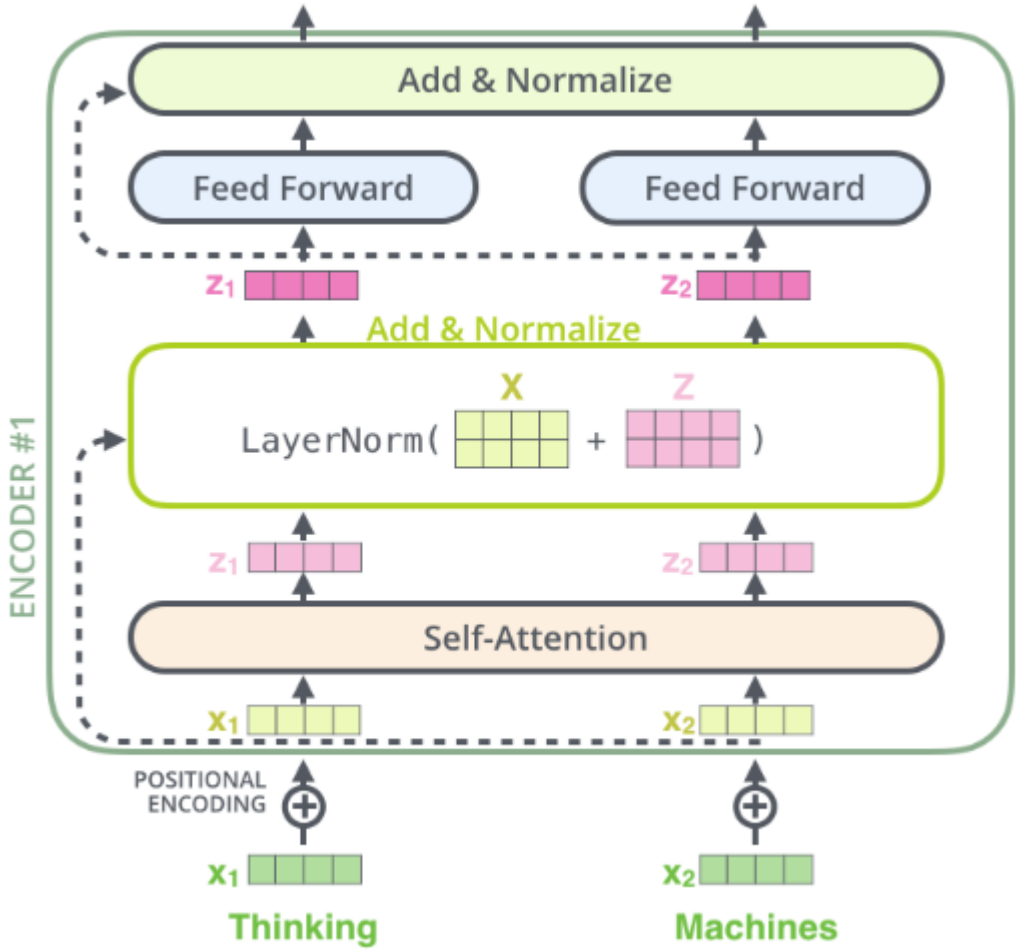



# Multi-head Attention

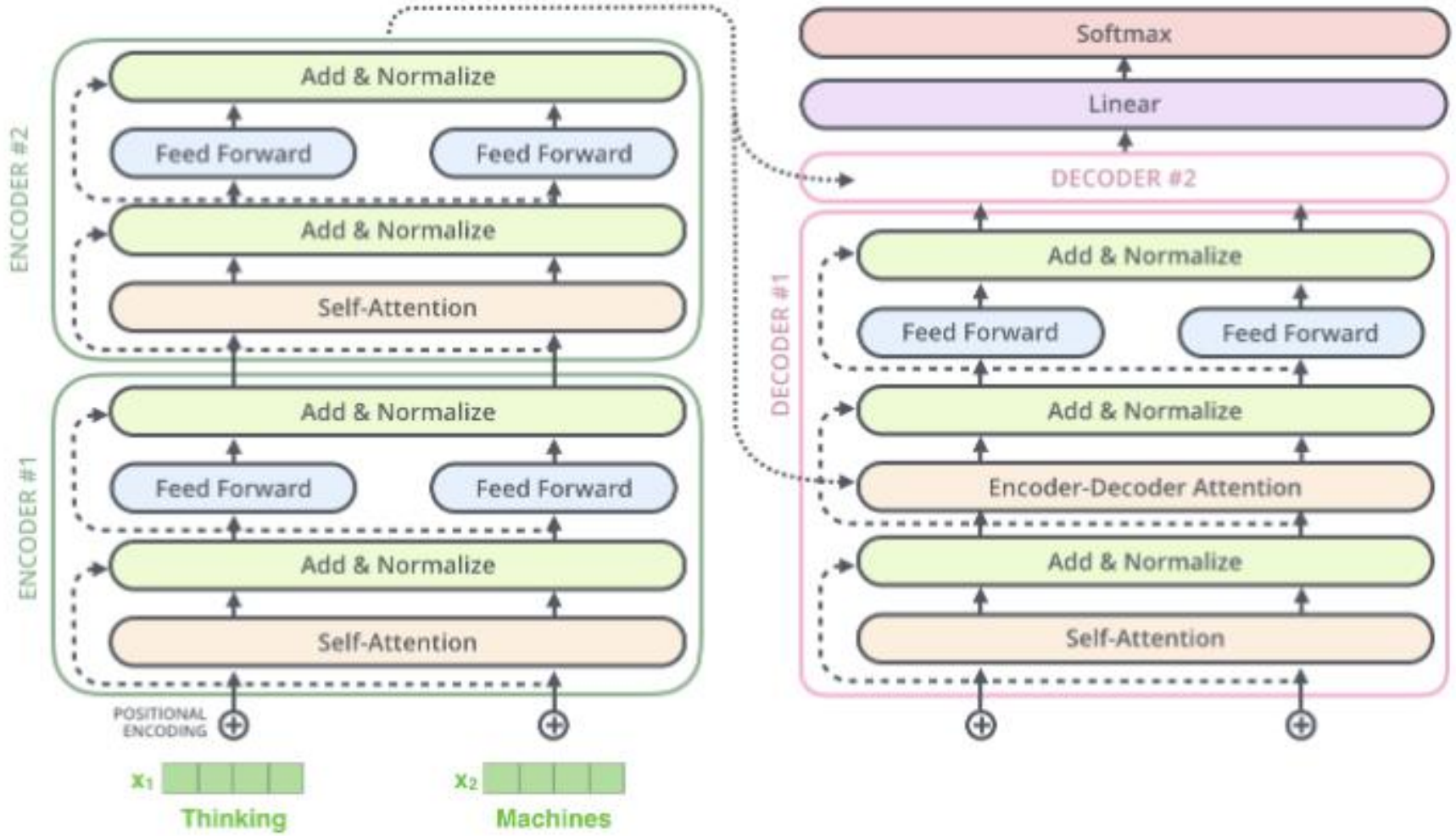
- Do many attention head calculation in parallel, and combine
- Each head has its own set of parameters
- Different heads can learn different “interactions” between inputs



# The Residuals and LayerNorm



# The Decoder Side



# The Final Linear and Softmax Layer

Which word in our vocabulary is associated with this index?

am

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

5