

"Today is a _____"



1) Tokenize

"Today is a _____"

['Today', 'is', 'a']

1x t
inp tokens



2) Embed

Vocabulary = $\begin{bmatrix} a, aah, \dots, zzz \\ 0 & 2 & \dots & 9 \\ 1 & 1 & \dots & 3 \\ 3 & 2 & \dots & 9 \end{bmatrix}$

words

v x d
dimension of embedding

Tokens = ['Today', 'is', 'a']

1x t



token embedding = $\begin{bmatrix} 4 & 1 & 3 \\ 3 & 5 & 6 \\ 7 & 8 & 0 \end{bmatrix}$

d x t

position embedding = $\begin{bmatrix} 1 & 1 & 1 \\ 0 & 1 & 3 \dots \\ 3 & 5 & 4 \end{bmatrix}$

tokens

t x d

tokens model supports

Input X = $\begin{bmatrix} 5 & 2 & 5 \\ 3 & 6 & 6 \\ 10 & 17 & 7 \end{bmatrix}$

t x d

4) Attention

X t x d



heads

for each head

$W_Q^{(i)}, W_K^{(i)}, W_V^{(i)}$

Learned Proj. Matrices

d x d_h
dims / head

calculate
Q, K, V

$Q^{(i)} = XW_Q^{(i)}$ → Query "What am I looking for?"
 $K^{(i)} = XW_K^{(i)}$ → Key "What do I offer?"
 $V^{(i)} = XW_V^{(i)}$ → Value "If selected, what will I pass?"

t x d_h

$$\text{Attention}^{(i)}(Q, K, V) = \text{softmax}\left(\frac{Q^{(i)}(K^{(i)})^T}{\sqrt{d_h}}\right)V^{(i)}$$

t x t

scaling

"From head h's POV,
what tokens should each token pay
attention to"

Fine Tuning

In attention Layers

- $W'_a = W_a + \Delta W_a$
- $W'_k = W_k + \Delta W_k$

$$W'_v = W_v + \Delta W_v$$

In FFT,

Full Matrix is updated

$$\Delta W \in d \times d$$

In PEFT,

$$\Delta W = B \cdot A \quad r \ll d$$

$d \times r \quad r \times d$

A = "which directions in input to pay attention to"
 B = "how to inject that adjustment into the output space"

Original weights are frozen

Why?

ΔW is usually very low-rank or close to low-rank.
 That is, the changes needed to adapt a pretrained model to a new task often lie in a small subspace.
 Most of the full $d \times d$ matrix doesn't need large changes; only a few directions matter.