RoPE

Michael Wollowski
in cooperation with ChatGPT.com

1

Introduction

- Transformers lack information of token order and distance.
- Consider these two sentences:
  "The dog chased the cat."
  "The cat chased the dog."
- They contain the same words.
- The meanings are different because the word order differs.

2

Introduction

- A Transformer processes tokens using self-attention.
- Self-attention is positional agnostic.
- We need to incorporate positional information.

3

Introduction

- Requirements for positional embeddings:
- Tokens closer to each other should attend with higher weights,
- Position within a sequence shouldn't matter:
  - If two words are close to each other, they should attend to each other with higher weights regardless of where they appear in a sequence.
- We will look at relative positional embeddings.

4

RoPE: Rotational Positional Encoding

- Split an embedding vector into pairs of vectors.
- Suppose $d = 8$: $x = [x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8]$
- It gets split into 4 pairs:
  - (x_1, x_2)
  - (x_3, x_4)
  - (x_5, x_6)
  - (x_7, x_8)

5

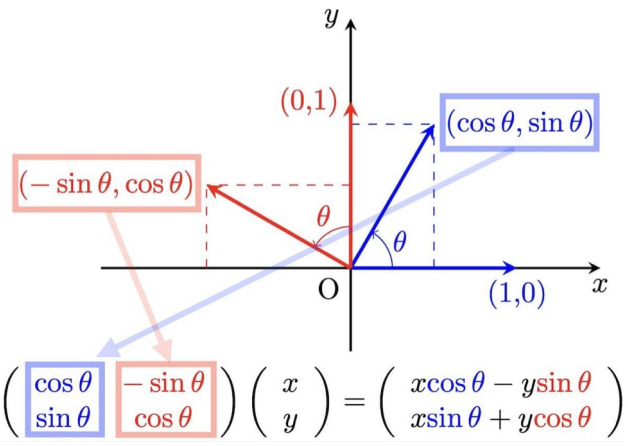
RoPE: Rotational Positional Encoding

- Every pair of dimensions represents a point in a two-dimensional plane.
- RoPE rotates each pair by an angle that depends on the token's position.
- Example: $x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$
- A rotation by angle θ is performed using the rotation matrix:

$$R(\theta) = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}$$

6

RoPE: Rotational Positional Encoding



7

RoPE: Rotational Positional Encoding

- The rotation angle depends on the token's position.
- For example:
  - Token at position 0: rotate by 0.
  - Token at position 1: rotate by θ.
  - Token at position 2: rotate by 2θ.
  - Token at position 3: rotate by 3θ.
- The same token embedding will be rotated differently depending on where the token occurs.

8

RoPE: Rotational Positional Encoding

- Additionally, each pair gets its own rotation frequency:

$$
\begin{aligned}
(x_1, x_2) &\to R(m\theta_0)(x_1, x_2) \\
(x_3, x_4) &\to R(m\theta_1)(x_3, x_4) \\
(x_5, x_6) &\to R(m\theta_2)(x_5, x_6) \\
(x_7, x_8) &\to R(m\theta_3)(x_7, x_8)
\end{aligned}
$$

- *m* is the token's position.

9

RoPE: Rotational Positional Encoding

When d = 8, the formula simplifies perfectly because 10,000 = 10^4:

$$\theta_i = 10000^{-\frac{2i}{8}} = 10000^{-\frac{i}{4}} = (10^4)^{-\frac{i}{4}} = 10^{-i}$$

Depending on how i is defined in your implementation, the exact values are:

For i ranging from 0 to 7 (d dimensions):

i	Formula ($\theta_i = 10^{-i}$)	Decimal Value	Fraction
0	10^0	1.0	1
1	10^{-1}	0.1	$\frac{1}{10}$
2	10^{-2}	0.01	$\frac{1}{100}$
3	10^{-3}	0.001	$\frac{1}{1,000}$

10

•RoPE: Rotational Positional Encoding

- Positional relationships in language:
  - Adjacent words often have important grammatical relationships.
  - Words several positions apart may have syntactic relationships.
  - Words hundreds of positions apart may be related through broader semantic or contextual relationships.
- RoPE uses different frequencies to represent positional differences at different scales.

11

Back to Attention

- Attention calculation:

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

- For each token, we calculate:
  - A query vector *q*.
  - A key vector *k*.
  - A value vector *v*.

12

Back to Attention

- Before comparing a query and a key, RoPE rotates both vectors according to the positions of their corresponding words.
- For example, suppose we have two words:
  - Word A is at position 2.
  - Word B is at position 5.
- RoPE does the following:
  - Rotate Word A's query vector according to position 2.
  - Rotate Word B's key vector according to position 5.
- Calculate the dot product of these two rotated vectors.

13

Back to Attention

- Because of the mathematics of rotations, comparing these two rotated vectors has an interesting property.
- The positional part of the comparison depends on the difference between their positions.
- In our example: 5-2=3
- The relative distance is three.
- Mathematically, the attention score becomes: $q^T R(5\text{-}2)k$
- The rotation now depends on the distance between the words, rather than their individual positions.

14

RoPE

- The complete attention score still depends on the content of the words.
- RoPE makes the positional component depend on relative distance.
- It does not make the entire attention score depend only on distance

15

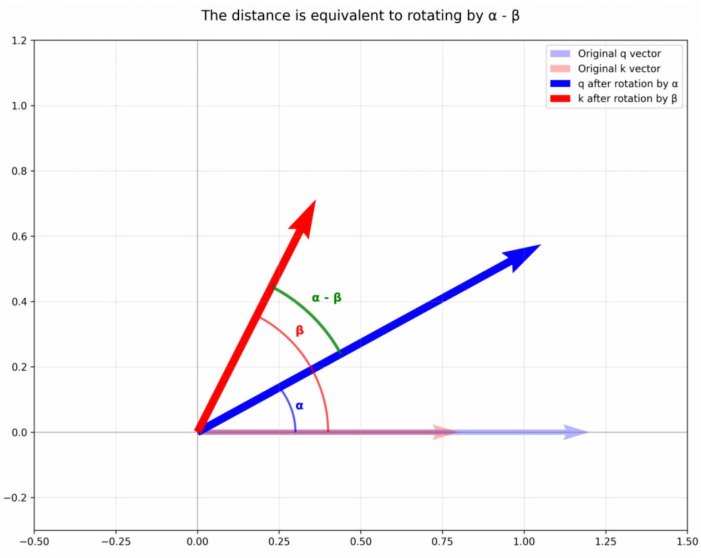


Image by author: Relative distance after rotation

Image source: Lorenzo Cesconetto. https://towardsdatascience.com/rope-clearly-explained/

16