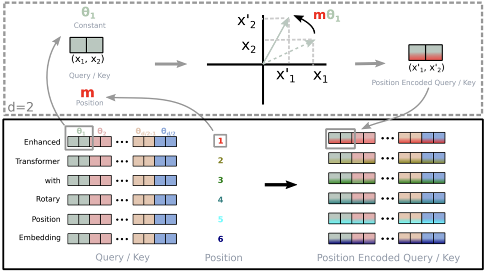
Rotational Positional Encoding

RoPE

A. Buynitsky

Jan 28, 2025



Outline

1. Complex Numbers
2. Transformers
3. Issue with Positional Encoding
4. Derivation
5. Result

Outline

1. Complex Numbers
2. Transformers
3. Issue with Positional Encoding
4. Derivation
5. Result

What is a Complex Number

Define $i = \sqrt{-1}$ so $i^2 = -1$
Any complex number $z = a + bi$ can be split into its real part:

$$\mathrm{Re}(Z) = a$$

And imaginary part:

$$\mathrm{Im}(z) = b$$

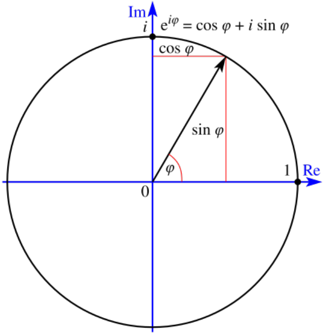
Example: Find Re and Im parts of $(3 + 4i)(2 + 3i)$:

$$\begin{aligned}(3 + 4i)(2 + 3i) &= 6 + 8i + 9i + 12(i^2) \\ &= 6 + 8i + 9i - 12 \\ &= -6 + 17i\end{aligned}$$

Euler's Formula

Consider the power series expansion of e^z:

$$
\begin{aligned}
e^z &= 1 + z + \frac{z^2}{2!} + \frac{z^3}{3!} + \frac{z^4}{4!} + \dots \\
e^{iz} &= 1 + iz + \frac{(iz)^2}{2!} + \frac{(iz)^3}{3!} + \frac{(iz)^4}{4!} + \dots \\
&= 1 + iz - \frac{(z)^2}{2!} - \frac{iz^3}{3!} + \frac{iz^4}{4!} + \dots \\
&= (1 - \frac{z^2}{2!} + \frac{iz^4}{4!} + \dots) + (iz - \frac{iz^3}{3!} + \dots) \\
&= (1 - \frac{z^2}{2!} + \frac{z^4}{4!} + \dots) + (iz - \frac{iz^3}{3!} + \dots) \\
&= cos(z) + isin(z)
\end{aligned}
$$



Therefore:

$$re^{i\theta} = r(cos(\theta) + isin(\theta))$$

Representation and Properties of Complex Numbers

For a complex number z:

$$z = re^{i\theta} = r(cos\theta + sin\theta) = a + bi$$

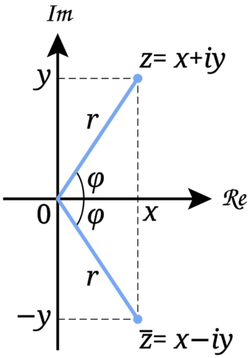
Magnitude of Complex Number:

$$R(z) = |z| = r = \sqrt{a^2 + b^2}$$

Argument (angle) of Complex Number:

$$\Theta(z) = \text{Arg}z = \theta = arctan(\frac{b}{a}) \text{ for } -\pi \leq \theta < \pi$$

$$\text{arg}z = \theta + 2\pi k = arctan(\frac{b}{a}) + 2\pi k \text{ for k } \in \mathbb{Z}$$



Conjugate of Complex Number:

$$\overline{z} = re^{-i\theta} = r(cos\theta - sin\theta)$$

Outline

1. Complex Numbers
2. Transformers
3. Issue with Positional Encoding
4. Derivation
5. Result

Transformers (part 1)

Define a sequence of input tokens and corresponding word embeddings:

$$\mathbb{S}_N = \{w_i\}_{i=1}^N$$
$$\mathbb{E}_N = \{x_i\}_{i=1}^N$$

where $x_i \in \mathbb{R}^d$ is embedding vector of token w_i **without positional info**:
Self Attention:
Calculates: $q_m = f_q(x_m, m), k_n = f_k(x_n, n), v_n = f_v(x_n, n)$ with:

$$f_{t:t\in\{q,k,v\}} = \mathbf{W}_{t:t\in\{q,k,v\}}(x_i + p_i) \tag{1}$$

where p_i is:

$$\begin{cases} p_{2t} = \sin\left(\frac{k}{10000} \cdot \frac{2t}{d}\right), \\ p_{2t+1} = \cos\left(\frac{k}{10000} \cdot \frac{2t}{d}\right). \end{cases}$$

Attention

Calculate attention score and normalize with softmax:

$$a_{m,n} = \frac{\exp\left(\frac{\mathbf{q}_m^\top \mathbf{k}_n}{\sqrt{d}}\right)}{\sum_{j=1}^{N} \exp\left(\frac{\mathbf{q}_m^\top \mathbf{k}_j}{\sqrt{d}}\right)}$$

Extract values

$$o_m = \sum_{n=1}^{N} a_{m,n} v_n$$

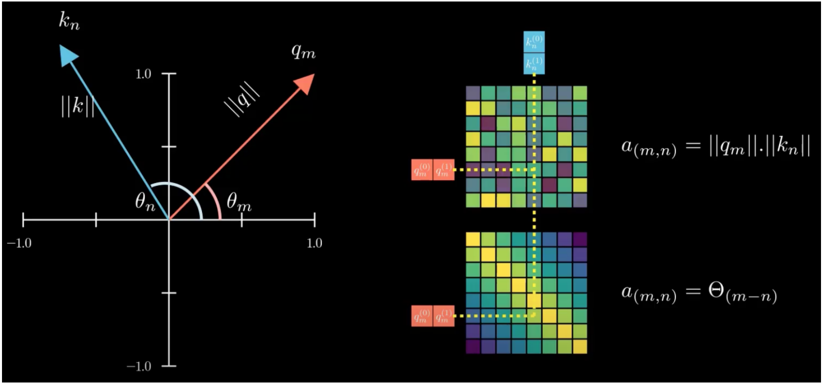
Outline

1. Complex Numbers
2. Transformers
3. Issue with Positional Encoding
4. Derivation
5. Result

Issue with positional encodings

Let's consider a 2D case:

- Similar tokens should have higher score $||q_m|| \cdot ||k_n||$ (magnitude)
- Closer tokens should have higher score $\Theta_{(m-n)}$



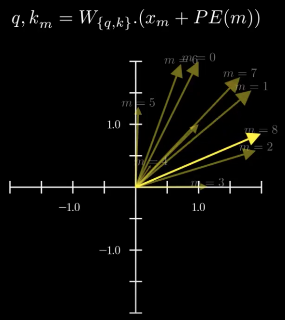
Shoutout to https://www.youtube.com/watch?v=GQPOtyITy54 for visuals

Issue with positional encodings (cont)

Let's consider a 2D case:

- Similar tokens should have higher score $||q_m|| \cdot ||k_n||$ (magnitude)
- Closer tokens should have higher score $\Theta_{(m-n)}$

	$i=0$	$i=1$
$m=0$	0.0	1.0
$m=1$	0.88	0.61
$m=2$	0.94	-0.46
$m=3$	0.17	-0.99
$m=4$	-0.81	-0.73
$m=5$	-0.97	0.3
$m=6$	-0.31	0.97
$m=7$	0.73	0.82



Problem: Combined positional and token embeddings together

Outline

1. Complex Numbers
2. Transformers
3. Issue with Positional Encoding
4. Derivation
5. Result

Reformulation

For each inner product, we want:

$$q_m^T k_n = \langle f_q(x_m, m), f_k(x_n, n) \rangle = g(x_m, x_n, m - n)$$

In 2D case, convert vectors to complex numbers:

$$\begin{aligned} q_m &= R_q(x_m, m) e^{i\Theta_q(x_m, m)} && (2) \\ k_n &= R_k(x_n, n) e^{i\Theta_k(x_n, n)} && (3) \\ g(x_m, x_n, n - m) &= R_g(x_m, x_n, n - m) e^{i\Theta_g(x_m, x_n, n - m)} && (4) \\ & && (5) \end{aligned}$$

Taking inner product between q_m and k_n in complex numbers:

$$q_m^T \cdot k_n = R_q(x_m, m) \cdot R_k(x_n, n) e^{i(\Theta_k(x_n, n) - \Theta_q(x_m, m))} \quad (6)$$

Determine Magintude

End up with two equations:

$$R_q(x_m, m)R_k(x_n, n) = R_g(x_m, x_n, n - m) \tag{7}$$

$$\Theta_k(x_n, n) - \Theta_q(x_m, m) = \Theta_g(x_m, x_n, n - m) \tag{8}$$

Now suppose that $m = n$.

$$R_q(x_m, m)R_k(x_n, n) = R_g(x_m, x_n, 0)$$

Now set $m = n = 0$

$$R_g(x_m, x_n, 0) = R_q(x_m, 0)R_k(x_n, 0) = ||q||\ ||k||$$

Therefore:

$$R_q(x_m, m) = ||q|| \text{ and } R_k(x_n, n) = ||k||$$

Determine Argument

By a similar trick setting $m = n$:

$$\Theta_q(x_m, m) - \Theta(x_n, n) = \Theta_g(x_m, x_n, 0) \tag{9}$$
$$= \Theta(x_m, 0) - \Theta(x_n, 0) = \theta_q - \theta_k \tag{10}$$

Rearranging, we get:

$$\Theta_q(x_m, m) - \theta_q = \Theta_k(x_n, m) - \theta_k$$

Observe that values only related to m, independent if $x = x_n$ or $x = x_m$ (generalize to $x = x_q$ and $x = x_k$)
Let $\phi(m)$ to be:

$$\phi(m) = \Theta_q(x_m, m) - \theta_q = \Theta_k(x_n, n) - \theta_k$$

Determine Argument

Recall:

$$\Theta_q(x_m, m) - \Theta_k(x_n, n) = \Theta_g(x_m, x_n, n - m)$$

and let $n = m + 1$ so:

$$\Theta_q(x_m, m) - \Theta_k(x_{m+1}, m + 1) = \Theta_g(x_m, x_{m+1}, 1) \tag{11}$$

Now:

$$\phi(m) = \Theta_q(x_m, m) - \theta_q \implies \Theta_q(x_m, m) = \phi(m) + \theta_q$$
$$\phi(m+1) = \Theta_k(x_{m+1}, m+1) - \theta_k \implies \Theta_k(x_{m+1}, m+1) = \phi(m+1) + \theta_k$$

Plugging into 11:

$$\begin{aligned} \phi(m) + \theta_q - (\phi(m+1) + \theta_k) &= \Theta_g(x_m, x_{m+1}, 1) \\ \phi(m) - \phi(m+1) &= \Theta_g(x_m, x_{m+1}, 1) - \theta_q + \theta_k \end{aligned}$$

RHS is constant irrelevant to m! $\phi(m)$ has constant difference between term regardless of m so its an arithmetic sequence!

$$\phi(m) = m\theta + \gamma$$

Combining Everything

$$q_m = f_q(x_m, m) = R_q(x_m, m)e^{i\Theta_q(x_m,m)} \tag{12}$$
$$= ||q||e^{i\Theta_q(x_m,m)} \tag{13}$$
$$= ||q||e^{i(\phi(m)+\theta_q)} \tag{14}$$
$$= ||q||e^{i(m\theta+\gamma+\theta_q)} \tag{15}$$
$$= ||q||e^{i\theta_q}e^{i(m\theta+\gamma)} \tag{16}$$
$$= qe^{i(m\theta+\gamma)} \tag{17}$$

Recall that:

$$q_m = f_q(x_m, m) = W_q(x_m + p_m)$$

so if $p_m = 0$, then

$$q_m = W_q \cdot x_m$$

Therefore assume no position info when $m = 0$ so let $\gamma = 0$.

Final Result:

$$f_q(x_m, m) = (W_q x_m)e^{im\theta}$$
$$f_k(x_n, n) = (W_k x_n)e^{in\theta}$$

Matrix Equations

Matrix form of equations in 2D

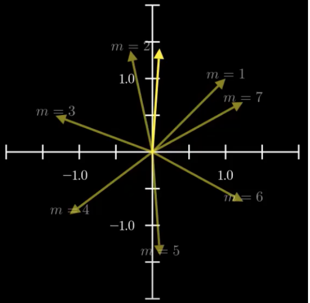
$d = 2 \quad i = 0 \quad m = 7$

$$\{q, k\}_m = R^d_{\Theta,m} \cdot \{q, k\}$$

$$\{q, k\} = W_{\{q,k\}} x_m$$

$$R^d_{\Theta,m} = \begin{pmatrix} cos(m\theta_i) & -sin(m\theta_i) \\ sin(m\theta_i) & cos(m\theta_i) \end{pmatrix}$$

$$\theta_i = 10000^{-2i/d}$$



Extending to multiple dimensions

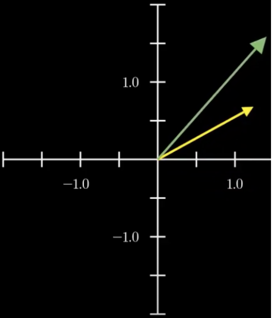
$d = 4 \quad i = \{0, 1\} \quad m = 6$

$$\{q, k\}_m = R^d_{\Theta,m} . \{q, k\}$$

$$\{q, k\} = W_{\{q,k\}} x_m$$

$$R^d_{\Theta,m} = \begin{pmatrix} \boxed{\begin{matrix} cos(m\theta_0) & -sin(m\theta_0) \\ sin(m\theta_0) & cos(m\theta_0) \end{matrix}} & \begin{matrix} 0 & 0 \\ 0 & 0 \end{matrix} \\ \begin{matrix} 0 & 0 \\ 0 & 0 \end{matrix} & \boxed{\begin{matrix} cos(m\theta_1) & -sin(m\theta_1) \\ sin(m\theta_1) & cos(m\theta_1) \end{matrix}} \end{pmatrix}$$

$$\theta_i = 10000^{-2i/d}$$



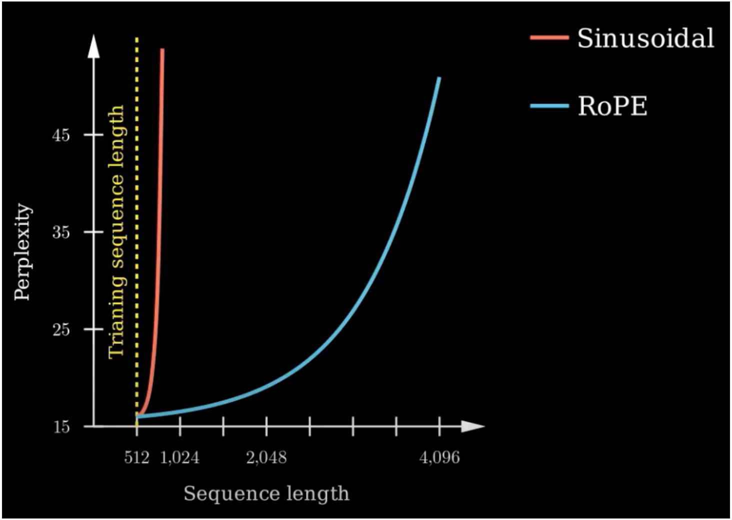
Block 0		Block 1	
$q_0^{(0)}$	$q_0^{(1)}$	$q_0^{(2)}$	$q_0^{(3)}$

Outline

1. Complex Numbers
2. Transformers
3. Issue with Positional Encoding
4. Derivation
5. Result

Result

Increases Sequence prediction confidence (perplexity):



Thank you!

Have a great rest of your Day!!!