Positional Embeddings and Relative Attention

Why should we care about relative attention?

- Consider a phrase like:

  The Space Needle is located in ___

Why should we care about relative attention?

- Consider the sentence:

  The Space Needle is located in ___

- Now consider phrases:

  The famous landmark known as the Space Needle is located in ___

  I think the Space Needle is located in ___

  The CN tower is located in Toronto and the Space Needle is located in ___

Why should we care about relative attention?

- Consider the sentence:

  The Space Needle is located in ___

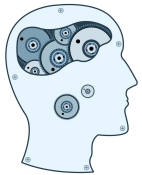
- Now consider phrases:

  The famous landmark known as the Space Needle is located in ___

  I think the Space Needle is located in ___

  The CN tower ~~is located in Toronto~~ and the Space Needle is located in ___

If h_{-3} to h_{-1} = "is located in", we should pay attention to h_{-4}



Why should we care about relative attention?

- Paying attention to all the text is necessary
- However, in many cases, the important tokens are the same distance away.
- With relative attention, we can reinforce recurring relationships even with offsets!

Why should we care about relative attention?

- Paying attention to all the text is necessary
- However, in many cases, the important tokens are the same distance away.
- With relative attention, we can reinforce recurring relationships!

- of course, absolute attention *allows* the formation of relative attention.
- But positional encodings help extra…

What we want to show

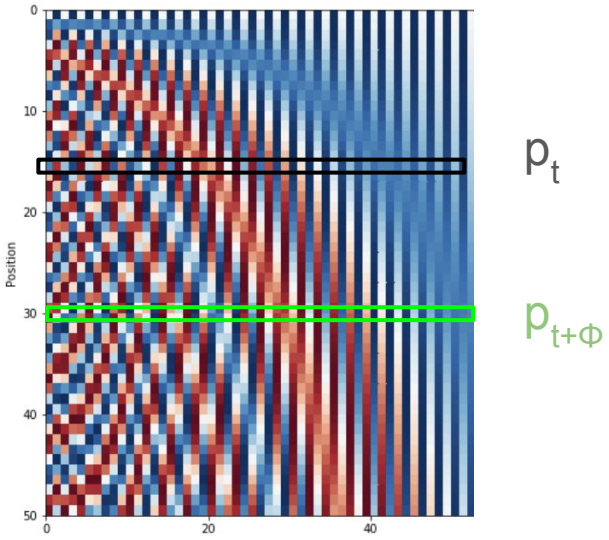
for position encodings p_t at position *t*, and following position encoding $p_{t+\Phi}$

$$p_t = \begin{bmatrix} \sin(\omega_1 . t) \\ \cos(\omega_1 . t) \\ \\ \sin(\omega_2 . t) \\ \cos(\omega_2 . t) \\ \\ \vdots \\ \\ \sin(\omega_{d/2} . t) \\ \cos(\omega_{d/2} . t) \end{bmatrix}$$

What we want to show

for position encodings p_t at position *t*, and following position encoding $p_{t+\Phi}$

$$p_t = \begin{bmatrix} \sin(\omega_1 . t) \\ \cos(\omega_1 . t) \\ \\ \sin(\omega_2 . t) \\ \cos(\omega_2 . t) \\ \\ \vdots \\ \\ \sin(\omega_{d/2} . t) \\ \cos(\omega_{d/2} . t) \end{bmatrix}$$



What we want to show

for position encodings p_t at position *t*, and following position encoding $p_{t+\Phi}$

we can express encoding $p_{t+\Phi}$ as linearly transformed p_t

$M\ p_t \quad = \quad p_{t+\Phi}$

$$p_t = \begin{bmatrix} \sin(\omega_1 . t) \\ \cos(\omega_1 . t) \\ \\ \sin(\omega_2 . t) \\ \cos(\omega_2 . t) \\ \\ \vdots \\ \\ \sin(\omega_{d/2} . t) \\ \cos(\omega_{d/2} . t) \end{bmatrix}$$

$$M . \begin{bmatrix} \sin(\omega_k . t) \\ \cos(\omega_k . t) \end{bmatrix} = \begin{bmatrix} \sin(\omega_k . (t+\phi)) \\ \cos(\omega_k . (t+\phi)) \end{bmatrix}$$

What we want to show

$$\begin{bmatrix} \sin(\omega_1 . t) \\ \cos(\omega_1 . t) \\ \\ \sin(\omega_2 . t) \\ \cos(\omega_2 . t) \\ \\ \vdots \\ \\ \sin(\omega_{d/2} . t) \\ \cos(\omega_{d/2} . t) \end{bmatrix}$$

we can express encoding p_{t+4} as linearly transformed p_t

$M\ p_t = p_{t+4}$

M = 

What we want to show

$$\begin{bmatrix} \sin(\omega_1 . t) \\ \cos(\omega_1 . t) \\ \\ \sin(\omega_2 . t) \\ \cos(\omega_2 . t) \\ \\ \vdots \\ \\ \sin(\omega_{d/2} . t) \\ \cos(\omega_{d/2} . t) \end{bmatrix}$$

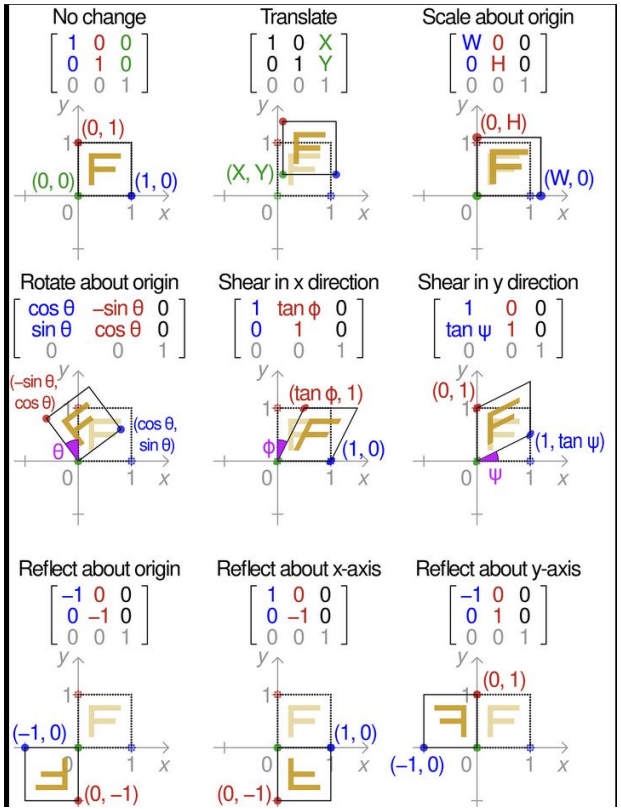
we can express encoding p_{t+4} as linearly transformed p_t

$M\ p_t = p_{t+4}$

M  = 

M is $\begin{bmatrix} u_1 & v_1 \\ u_2 & v_2 \end{bmatrix}$

Remember: matrices are linear transformations!



$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} ax + by \\ cx + dy \end{pmatrix}$$

When we multiply by a 2x2 matrix, we are changing our basis from {(1,0), (0,1)} to {(a, c), (b, d)}

Doing the math

$$M.\begin{bmatrix}\sin(\omega_k.t)\\ \cos(\omega_k.t)\end{bmatrix}=\begin{bmatrix}\sin(\omega_k.(t+\phi))\\ \cos(\omega_k.(t+\phi))\end{bmatrix}$$

Doing the math

$$M.\begin{bmatrix}\sin(\omega_k.t)\\ \cos(\omega_k.t)\end{bmatrix}=\begin{bmatrix}\sin(\omega_k.(t+\phi))\\ \cos(\omega_k.(t+\phi))\end{bmatrix}$$

$$\begin{bmatrix}u_1 & v_1\\ u_2 & v_2\end{bmatrix}.\begin{bmatrix}\sin(\omega_k.t)\\ \cos(\omega_k.t)\end{bmatrix}=\begin{bmatrix}\sin(\omega_k.(t+\phi))\\ \cos(\omega_k.(t+\phi))\end{bmatrix}$$

Doing the math

$$M.\begin{bmatrix}\sin(\omega_k.t)\\ \cos(\omega_k.t)\end{bmatrix}=\begin{bmatrix}\sin(\omega_k.(t+\phi))\\ \cos(\omega_k.(t+\phi))\end{bmatrix}$$

$$\begin{bmatrix}u_1 & v_1\\ u_2 & v_2\end{bmatrix}.\begin{bmatrix}\sin(\omega_k.t)\\ \cos(\omega_k.t)\end{bmatrix}=\begin{bmatrix}\sin(\omega_k.(t+\phi))\\ \cos(\omega_k.(t+\phi))\end{bmatrix}$$

The really cool part: sin(x+y) = sin(x)cos(y) + cos(x)sin(y)

Doing the math

$$M.\begin{bmatrix}\sin(\omega_k.t)\\ \cos(\omega_k.t)\end{bmatrix}=\begin{bmatrix}\sin(\omega_k.(t+\phi))\\ \cos(\omega_k.(t+\phi))\end{bmatrix}$$

$$\begin{bmatrix}u_1 & v_1\\ u_2 & v_2\end{bmatrix}.\begin{bmatrix}\sin(\omega_k.t)\\ \cos(\omega_k.t)\end{bmatrix}=\boxed{\begin{bmatrix}\sin(\omega_k.(t+\phi))\\ \cos(\omega_k.(t+\phi))\end{bmatrix}}$$

The really cool part: $\boxed{\sin(x+y)} = \sin(x)\cos(y) + \cos(x)\sin(y)$

The right hand side produces $u_1\sin(x) + v_1\cos(x)$

Doing the math

$$M.\begin{bmatrix}\sin(\omega_k.t)\\ \cos(\omega_k.t)\end{bmatrix}=\begin{bmatrix}\sin(\omega_k.(t+\phi))\\ \cos(\omega_k.(t+\phi))\end{bmatrix}$$

$$\begin{bmatrix}u_1 & v_1\\ u_2 & v_2\end{bmatrix}.\begin{bmatrix}\sin(\omega_k.t)\\ \cos(\omega_k.t)\end{bmatrix}=\boxed{\begin{bmatrix}\sin(\omega_k.(t+\phi))\\ \cos(\omega_k.(t+\phi))\end{bmatrix}}$$

The really cool part: $\boxed{\sin(x+y)} = \sin(x)\cos(y) + \cos(x)\sin(y)$

The right hand side produces $u_1\sin(x) + v_1\cos(x)$

Doing the math

$$M.\begin{bmatrix}\sin(\omega_k.t)\\ \cos(\omega_k.t)\end{bmatrix}=\begin{bmatrix}\sin(\omega_k.(t+\phi))\\ \cos(\omega_k.(t+\phi))\end{bmatrix}$$

$$\begin{bmatrix}u_1 & v_1\\ u_2 & v_2\end{bmatrix}.\begin{bmatrix}\sin(\omega_k.t)\\ \cos(\omega_k.t)\end{bmatrix}=\begin{bmatrix}\sin(\omega_k.(t+\phi))\\ \cos(\omega_k.(t+\phi))\end{bmatrix}$$

The really cool part: sin(x+y) = sin(x)cos(y) + cos(x)sin(y)

The right hand side produces $u_1$sin(x) + $v_1$cos(x)

So we're going to end up with only terms involving y…

Doing the math

$$M.\begin{bmatrix}\sin(\omega_k.t)\\ \cos(\omega_k.t)\end{bmatrix}=\begin{bmatrix}\sin(\omega_k.(t+\phi))\\ \cos(\omega_k.(t+\phi))\end{bmatrix}$$

$$\begin{bmatrix}u_1 & v_1\\ u_2 & v_2\end{bmatrix}.\begin{bmatrix}\sin(\omega_k.t)\\ \cos(\omega_k.t)\end{bmatrix}=\boxed{\begin{bmatrix}\sin(\omega_k.(t+\phi))\\ \cos(\omega_k.(t+\phi))\end{bmatrix}}$$

The really cool part: sin(x+y) = sin(x)cos(y) + cos(x)sin(y)

The right hand side produces $u_1$sin(x) + $v_1$cos(x)

So we're going to end up with only terms involving y…

Where x is w_k t, our only absolute position!

Doing the math

$$M.\begin{bmatrix}\sin(\omega_k.t)\\ \cos(\omega_k.t)\end{bmatrix}=\begin{bmatrix}\sin(\omega_k.(t+\phi))\\ \cos(\omega_k.(t+\phi))\end{bmatrix}$$

$$\begin{bmatrix}u_1 & v_1\\ u_2 & v_2\end{bmatrix}.\begin{bmatrix}\sin(\omega_k.t)\\ \cos(\omega_k.t)\end{bmatrix}=\begin{bmatrix}\sin(\omega_k.(t+\phi))\\ \cos(\omega_k.(t+\phi))\end{bmatrix}$$

$$\begin{bmatrix}u_1 & v_1\\ u_2 & v_2\end{bmatrix}.\begin{bmatrix}\sin(\omega_k.t)\\ \cos(\omega_k.t)\end{bmatrix}=\begin{bmatrix}\sin(\omega_k.t)\cos(\omega_k.\phi)+\cos(\omega_k.t)\sin(\omega_k.\phi)\\ \cos(\omega_k.t)\cos(\omega_k.\phi)-\sin(\omega_k.t)\sin(\omega_k.\phi)\end{bmatrix}$$

Doing the math

$$M.\begin{bmatrix}\sin(\omega_k.t)\\ \cos(\omega_k.t)\end{bmatrix}=\begin{bmatrix}\sin(\omega_k.(t+\phi))\\ \cos(\omega_k.(t+\phi))\end{bmatrix}$$

$$\begin{bmatrix}u_1 & v_1\\ u_2 & v_2\end{bmatrix}.\begin{bmatrix}\sin(\omega_k.t)\\ \cos(\omega_k.t)\end{bmatrix}=\begin{bmatrix}\sin(\omega_k.(t+\phi))\\ \cos(\omega_k.(t+\phi))\end{bmatrix}$$

$$\begin{bmatrix}u_1 & v_1\\ u_2 & v_2\end{bmatrix}.\begin{bmatrix}\sin(\omega_k.t)\\ \cos(\omega_k.t)\end{bmatrix}=\begin{bmatrix}\sin(\omega_k.t)\cos(\omega_k.\phi)+\cos(\omega_k.t)\sin(\omega_k.\phi)\\ \cos(\omega_k.t)\cos(\omega_k.\phi)-\sin(\omega_k.t)\sin(\omega_k.\phi)\end{bmatrix}$$

Doing the math

$$M.\begin{bmatrix}\sin(\omega_k.t)\\ \cos(\omega_k.t)\end{bmatrix}=\begin{bmatrix}\sin(\omega_k.(t+\phi))\\ \cos(\omega_k.(t+\phi))\end{bmatrix}$$

$$\begin{bmatrix}u_1 & v_1\\ u_2 & v_2\end{bmatrix}.\begin{bmatrix}\sin(\omega_k.t)\\ \cos(\omega_k.t)\end{bmatrix}=\begin{bmatrix}\sin(\omega_k.(t+\phi))\\ \cos(\omega_k.(t+\phi))\end{bmatrix}$$

$$\begin{bmatrix}u_1 & v_1\\ u_2 & v_2\end{bmatrix}.\begin{bmatrix}\sin(\omega_k.t)\\ \cos(\omega_k.t)\end{bmatrix}=\begin{bmatrix}\sin(\omega_k.t)\cos(\omega_k.\phi)+\cos(\omega_k.t)\sin(\omega_k.\phi)\\ \cos(\omega_k.t)\cos(\omega_k.\phi)-\sin(\omega_k.t)\sin(\omega_k.\phi)\end{bmatrix}$$

$$u_1\sin(\omega_k.t)+v_1\cos(\omega_k.t)$$
$$u_2\sin(\omega_k.t)+v_2\cos(\omega_k.t)$$

Doing the math

$$M.\begin{bmatrix}\sin(\omega_k.t)\\ \cos(\omega_k.t)\end{bmatrix}=\begin{bmatrix}\sin(\omega_k.(t+\phi))\\ \cos(\omega_k.(t+\phi))\end{bmatrix}$$

$$\begin{bmatrix}u_1 & v_1\\ u_2 & v_2\end{bmatrix}.\begin{bmatrix}\sin(\omega_k.t)\\ \cos(\omega_k.t)\end{bmatrix}=\begin{bmatrix}\sin(\omega_k.(t+\phi))\\ \cos(\omega_k.(t+\phi))\end{bmatrix}$$

$$\begin{bmatrix}u_1 & v_1\\ u_2 & v_2\end{bmatrix}.\begin{bmatrix}\sin(\omega_k.t)\\ \cos(\omega_k.t)\end{bmatrix}=\begin{bmatrix}\sin(\omega_k.t)\cos(\omega_k.\phi)+\cos(\omega_k.t)\sin(\omega_k.\phi)\\ \cos(\omega_k.t)\cos(\omega_k.\phi)-\sin(\omega_k.t)\sin(\omega_k.\phi)\end{bmatrix}$$

$$u_1\sin(\omega_k.t)+v_1\cos(\omega_k.t)=\cos(\omega_k.\phi)\sin(\omega_k.t)+\sin(\omega_k.\phi)\cos(\omega_k.t) \quad (1)$$

$$u_2\sin(\omega_k.t)+v_2\cos(\omega_k.t)=-\sin(\omega_k.\phi)\sin(\omega_k.t)+\cos(\omega_k.\phi)\cos(\omega_k.t) \quad (2)$$

Doing the math

$$M.\begin{bmatrix}\sin(\omega_k.t)\\ \cos(\omega_k.t)\end{bmatrix}=\begin{bmatrix}\sin(\omega_k.(t+\phi))\\ \cos(\omega_k.(t+\phi))\end{bmatrix}$$

$$\begin{bmatrix}u_1 & v_1\\ u_2 & v_2\end{bmatrix}.\begin{bmatrix}\sin(\omega_k.t)\\ \cos(\omega_k.t)\end{bmatrix}=\begin{bmatrix}\sin(\omega_k.(t+\phi))\\ \cos(\omega_k.(t+\phi))\end{bmatrix}$$

$$\begin{bmatrix}u_1 & v_1\\ u_2 & v_2\end{bmatrix}.\begin{bmatrix}\sin(\omega_k.t)\\ \cos(\omega_k.t)\end{bmatrix}=\begin{bmatrix}\sin(\omega_k.t)\cos(\omega_k.\phi)+\cos(\omega_k.t)\sin(\omega_k.\phi)\\ \cos(\omega_k.t)\cos(\omega_k.\phi)-\sin(\omega_k.t)\sin(\omega_k.\phi)\end{bmatrix}$$

$$u_1\sin(\omega_k.t)+v_1\cos(\omega_k.t)=\cos(\omega_k.\phi)\sin(\omega_k.t)+\sin(\omega_k.\phi)\cos(\omega_k.t) \quad (1)$$

$$u_2\sin(\omega_k.t)+v_2\cos(\omega_k.t)=-\sin(\omega_k.\phi)\sin(\omega_k.t)+\cos(\omega_k.\phi)\cos(\omega_k.t) \quad (2)$$

Doing the math

$$M.\begin{bmatrix}\sin(\omega_k.t)\\ \cos(\omega_k.t)\end{bmatrix}=\begin{bmatrix}\sin(\omega_k.(t+\phi))\\ \cos(\omega_k.(t+\phi))\end{bmatrix}$$

$$\begin{bmatrix}u_1 & v_1\\ u_2 & v_2\end{bmatrix}.\begin{bmatrix}\sin(\omega_k.t)\\ \cos(\omega_k.t)\end{bmatrix}=\begin{bmatrix}\sin(\omega_k.(t+\phi))\\ \cos(\omega_k.(t+\phi))\end{bmatrix}$$

$$\begin{bmatrix}u_1 & v_1\\ u_2 & v_2\end{bmatrix}.\begin{bmatrix}\sin(\omega_k.t)\\ \cos(\omega_k.t)\end{bmatrix}=\begin{bmatrix}\sin(\omega_k.t)\cos(\omega_k.\phi)+\cos(\omega_k.t)\sin(\omega_k.\phi)\\ \cos(\omega_k.t)\cos(\omega_k.\phi)-\sin(\omega_k.t)\sin(\omega_k.\phi)\end{bmatrix}$$

$$u_1\sin(\omega_k.t)+v_1\cos(\omega_k.t)=\cos(\omega_k.\phi)\sin(\omega_k.t)+\sin(\omega_k.\phi)\cos(\omega_k.t) \quad (1)$$
$$u_2\sin(\omega_k.t)+v_2\cos(\omega_k.t)=-\sin(\omega_k.\phi)\sin(\omega_k.t)+\cos(\omega_k.\phi)\cos(\omega_k.t) \quad (2)$$

$$u_1=\cos(\omega_k.\phi)\quad v_1=\sin(\omega_k.\phi)$$
$$u_2=-\sin(\omega_k.\phi)\quad v_2=\cos(\omega_k.\phi)$$

Doing the math

$$M.\begin{bmatrix}\sin(\omega_k.t)\\ \cos(\omega_k.t)\end{bmatrix}=\begin{bmatrix}\sin(\omega_k.(t+\phi))\\ \cos(\omega_k.(t+\phi))\end{bmatrix}$$

$$\begin{bmatrix}u_1 & v_1\\ u_2 & v_2\end{bmatrix}.\begin{bmatrix}\sin(\omega_k.t)\\ \cos(\omega_k.t)\end{bmatrix}=\begin{bmatrix}\sin(\omega_k.(t+\phi))\\ \cos(\omega_k.(t+\phi))\end{bmatrix}$$

$$\begin{bmatrix}u_1 & v_1\\ u_2 & v_2\end{bmatrix}.\begin{bmatrix}\sin(\omega_k.t)\\ \cos(\omega_k.t)\end{bmatrix}=\begin{bmatrix}\sin(\omega_k.t)\cos(\omega_k.\phi)+\cos(\omega_k.t)\sin(\omega_k.\phi)\\ \cos(\omega_k.t)\cos(\omega_k.\phi)-\sin(\omega_k.t)\sin(\omega_k.\phi)\end{bmatrix}$$

$$u_1\sin(\omega_k.t)+v_1\cos(\omega_k.t)=\cos(\omega_k.\phi)\sin(\omega_k.t)+\sin(\omega_k.\phi)\cos(\omega_k.t) \quad (1)$$
$$u_2\sin(\omega_k.t)+v_2\cos(\omega_k.t)=-\sin(\omega_k.\phi)\sin(\omega_k.t)+\cos(\omega_k.\phi)\cos(\omega_k.t) \quad (2)$$

$$M_{\phi,k}=\begin{bmatrix}\cos(\omega_k.\phi) & \sin(\omega_k.\phi)\\ -\sin(\omega_k.\phi) & \cos(\omega_k.\phi)\end{bmatrix}$$

the final transformation does not depend on t, only the relative offset Φ!

$$M_{\phi,k} = \begin{bmatrix} \cos(\omega_k . \phi) & \sin(\omega_k . \phi) \\ -\sin(\omega_k . \phi) & \cos(\omega_k . \phi) \end{bmatrix}$$

the final transformation does not depend on t, only the relative offset Φ!

$$M_{\phi,k} = \begin{bmatrix} \cos(\omega_k . \phi) & \sin(\omega_k . \phi) \\ -\sin(\omega_k . \phi) & \cos(\omega_k . \phi) \end{bmatrix}$$

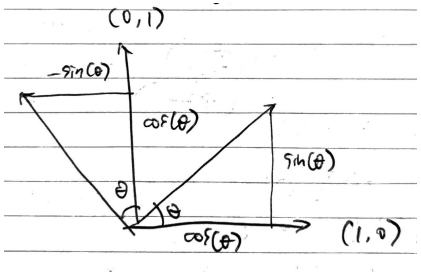
And one more interesting thing…

the final transformation does not depend on t, only the relative offset Φ!

$$M_{\phi,k} = \begin{bmatrix} \cos(\omega_k . \phi) & \sin(\omega_k . \phi) \\ -\sin(\omega_k . \phi) & \cos(\omega_k . \phi) \end{bmatrix}$$

And one more interesting thing…

$$\mathrm{R}_\theta = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix},$$



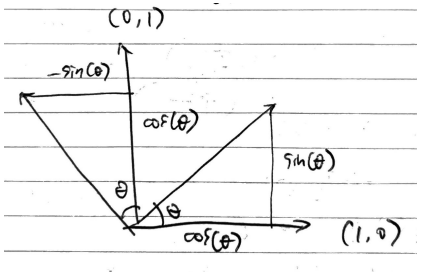
$p_{t+\Phi}$ is p_t **rotated** by an amount based on Φ

the final transformation does not depend on t, only the relative offset Φ!

$$M_{\phi,k} = \begin{bmatrix} \cos(\omega_k . \phi) & \sin(\omega_k . \phi) \\ -\sin(\omega_k . \phi) & \cos(\omega_k . \phi) \end{bmatrix}$$

And one more interesting thing…

$$\mathrm{R}_\theta = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix},$$



$p_{t+\Phi}$ is p_t **rotated** by an amount based on Φ

Summary / final notes

- Positional embeddings are added to each token embedding

- relative attention: if p_{t-1}, p_{t-2}, p_{t-3} were "is located in", pay lots of attention to p_{t-4}

- By using the positional embeddings of words, it becomes very easy for a transformer to represent relative attention…
- Because p_{t-1}, p_{t-2}, p_{t-3} are simple linear transforms of p_t!