RoPE — 旋转位置编码

1. 复数

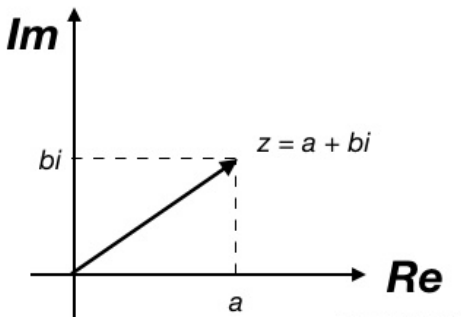
1.1 复数定义

$$z = a + bi, z \in \mathbb{C}, a, b \in \mathbb{R}, i^2 = -1$$

复数可以看做 (1,i)T 这组基 (Basis) 的线性组合（Linear Combination），所以可以用向量来表示复数。

$$\begin{bmatrix} a \\ b \end{bmatrix}$$

当然也可以用复平面上的点来表示复数：



1.2 复数乘法

复数的乘法：

$$z_1 = a + bi, z_2 = c + di$$

$$\begin{aligned} z_1 z_2 &= (a + bi)(c + di) \\ &= ac - bd + (bc + ad)i \end{aligned}$$

可以把它看做一个矩阵和一个向量相乘：

$$\begin{bmatrix} a & -b \\ b & a \end{bmatrix} \begin{bmatrix} c \\ d \end{bmatrix}$$

1.3 复数的矩阵表示

所以复数也可以看成矩阵，z1可以表示为：

$$\begin{bmatrix} a & -b \\ b & a \end{bmatrix}$$

把 z2 也换成对应的矩阵：

$$\begin{bmatrix} c & -d \\ d & c \end{bmatrix}$$

那么通过矩阵计算两者相乘，可得：

$$z_1 z_2 = \begin{bmatrix} a & -b \\ b & a \end{bmatrix} \begin{bmatrix} c & -d \\ d & c \end{bmatrix} = \begin{bmatrix} ac - bd & -(ad + bc) \\ (ad + bc) & ac - bd \end{bmatrix}$$

而正常的复数相乘结果为：

$$\begin{aligned} z_1 z_2 &= (a + bi)(c + di) \\ &= ac - bd + (bc + ad)i \end{aligned}$$

通过对比，我们发现，**复数可以表示为主对角线为实数值，次对角线为虚数值及其负数的形式**。

这也足以证明复数看成矩阵是正确的，所以复数相乘这个运算，也可以看成是矩阵变换。

1可以等价的看成 z = 1 + 0i ，其矩阵形式为：

$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

即单位矩阵。

i 为 z = 0 + 1i，等价于：

$$\begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}$$

$$i^2 = i \cdot i = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix} = \begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix} = -I = -1$$

这也说明了矩阵形式的正确性。

同时无论用代数形式或者矩阵形式，都可以验证复数的乘法满足交换律。

1.4 模与共轭

复数的模：

$$\|z\| = \sqrt{a^2 + b^2}$$

共轭 (Conjugate), 如果 z = a + bi，其共轭：

$$z^* = a - bi$$

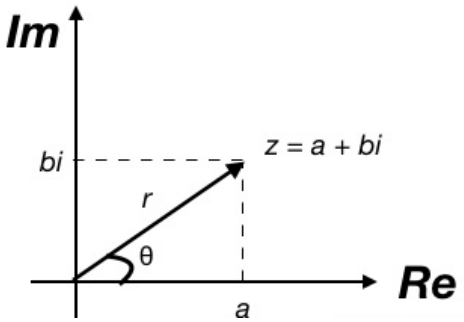
可以算出：

$$\|z\|^2 = zz^*$$

1.5 复数的极坐标表示

z = a + bi，z 的向量形式可以进行如下转换：

$$\begin{bmatrix} a \\ b \end{bmatrix} = \sqrt{a^2+b^2} \begin{bmatrix} \frac{a}{\sqrt{a^2+b^2}} \\ \frac{b}{\sqrt{a^2+b^2}} \end{bmatrix} = \|z\| \begin{bmatrix} \frac{a}{\|z\|} \\ \frac{b}{\|z\|} \end{bmatrix} = \|z\| \begin{bmatrix} \cos\theta \\ \sin\theta \end{bmatrix}$$



所以复数有极坐标形式，经常被写作：

$$z = r(\cos\theta + i\sin\theta)$$

其中，r = ||z||。

又通过欧拉公式：

$$e^{ix} = \cos x + i\sin x$$

可以写作指数形式：

$$z = re^{i\theta}$$

共轭复数：

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

1.6 复数的几种表示形式

所以上述给了我们几种看待复数的方式：

代数形式:

$$z = a + bi$$

向量形式:

$$z = \begin{bmatrix} a \\ b \end{bmatrix}$$

矩阵形式:

$$z = \begin{bmatrix} a & -b \\ b & a \end{bmatrix}$$

极坐标形式:

$$z = r(\cos\theta + i\sin\theta)$$

指数形式 :

$$z = re^{i\theta}$$

1.7 复数相乘与2D旋转

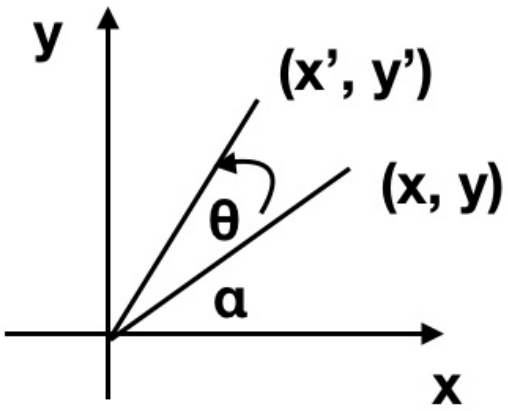
所以如果我们跟一个复数相乘，那么所做的矩阵变换是：

$$\begin{aligned} z = \begin{bmatrix} a & -b \\ b & a \end{bmatrix} &= \sqrt{a^2+b^2} \begin{bmatrix} \frac{a}{\sqrt{a^2+b^2}} & -\frac{b}{\sqrt{a^2+b^2}} \\ \frac{b}{\sqrt{a^2+b^2}} & \frac{a}{\sqrt{a^2+b^2}} \end{bmatrix} \\ &= \|z\| \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \\ &= \|z\| \cdot I \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \\ &= \begin{bmatrix} \|z\| & 0 \\ 0 & \|z\| \end{bmatrix} \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \end{aligned}$$

这个矩阵：

$$\begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}$$

代表2D平面上的旋转：



$$(x = r\cos\alpha, y = r\sin\alpha)$$
$$(x' = r\cos(\alpha+\theta), y = r\sin(\alpha+\theta))$$

展开代入得：

$$x' = x\cos\theta - y\sin\theta$$
$$y' = x\sin\theta + y\cos\theta$$

$$\cos(\alpha+\beta)=\cos\alpha\cdot\cos\beta-\sin\alpha\cdot\sin\beta$$

$$\cos(\alpha-\beta)=\cos\alpha\cdot\cos\beta+\sin\alpha\cdot\sin\beta$$

$$\sin(\alpha\pm\beta)=\sin\alpha\cdot\cos\beta\pm\cos\alpha\cdot\sin\beta$$

也就是：

$$\begin{bmatrix} x' \\ y' \end{bmatrix} = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix}$$

这是旋转变换与缩放变换的复合，所以如果**将任何复数c与向量相乘都是将向量逆时针旋转 θ 度，并将其缩放**：

$$\|z\| = \sqrt{a^2 + b^2}$$

> 💡 **空间向量与复数相乘，是将空间向量缩放并旋转**

如果 ||z|| = 1，那么复数可以用一个单位向量表示，同时这个乘法只做旋转变换。

所以平面上的旋转可以有矩阵形式：

$$\mathbf{v}' = \begin{bmatrix} \cos\theta & -\sin\theta \\ sin\theta & \cos\theta \end{bmatrix} \mathbf{v}$$

利用复数的代数形式：

$$\begin{aligned} v' &= zv \\ &= (\cos\theta + i\sin\theta)v \end{aligned}$$

继续利用其指数形式：

$$v' = e^{i\theta}v$$

可以看到**平面上的旋转可以跟复数很好的结合起来**。

2. RoPE原理

2.1 基本思路

回归初衷，我们要找到一种方法，能够为 Attention 添加位置信息。

我们假设通过下述运算来给 q, k 添加绝对位置信息：

$$\tilde{\boldsymbol{q}}_m = \boldsymbol{f}(\boldsymbol{q}, m), \quad \tilde{\boldsymbol{k}}_n = \boldsymbol{f}(\boldsymbol{k}, n)$$

也就是说，我们分别为 q, k 设计操作 **f(·, m)**, **f(·, n)**，**使得经过该操作后，qm, qk 就带有了位置 m, n 的绝对位置信息**。同时，我们希望经过 Attention 的内积运算后，**内积结果带有相对位置信息**，因此假设存在恒等关系：

$$\langle \boldsymbol{f}(\boldsymbol{q}, m), \boldsymbol{f}(\boldsymbol{k}, n) \rangle = g(\boldsymbol{q}, \boldsymbol{k}, m-n)$$

所以我们要求出该恒等式的一个（尽可能简单的）解。

求解过程还需要一些初始条件，显然我们可以合理地设 f(q, 0) = q， f(k, 0) = k，即**没有任何位置信息时，f(·, 0)等于原值**。

> 💡 **设计的f首先包含绝对位置信息，并且要求两个不同向量的f的内积包含两者的相对位置信息**

2.2 求解过程

在复数中有：

$$\langle \boldsymbol{q}, \boldsymbol{k} \rangle = \mathrm{Re}[\boldsymbol{q}\boldsymbol{k}^*]$$

其中，Re[]代表复数的实部，上式代表向量 q 与向量 k 的乘积，等于复数 q 与 复数 k 共轭复数 k* 相乘的实部。

[3,5] 对应的复数为 3 + 5i，[1,6] 对应的复数为 1 + 6i，其共轭复数为 1 - 6i，则有：

[3,5] * [1,6]T = 33， 其中 (3 + 5i)(1-6i) = 33 - 13i，且 Re[33 - 13i] = 33

在基本思路中我们提到，我们希望存在某个函数g，使其能够包含相对位置信息，即：

$$\langle \boldsymbol{f}(\boldsymbol{q}, m), \boldsymbol{f}(\boldsymbol{k}, n) \rangle = g(\boldsymbol{q}, \boldsymbol{k}, m - n)$$

带入复数表达式：

$$\mathrm{Re}[\boldsymbol{f}(\boldsymbol{q}, m) \boldsymbol{f}^*(\boldsymbol{k}, n)] = g(\boldsymbol{q}, \boldsymbol{k}, m - n)$$

简单起见，我们直接假设存在复数 g(q, k, m-n)，使得：

$$\boldsymbol{f}(\boldsymbol{q}, m) \boldsymbol{f}^*(\boldsymbol{k}, n) = \boldsymbol{g}(\boldsymbol{q}, \boldsymbol{k}, m - n)$$

根据复数的基本概念，我们知道，复数 z = a + bi 的向量形式为 [a, b]，指数形式为：

$$z = re^{i\theta}$$

即 f(q, m) 和 f(k, n) 是加入绝对位置信息之后的 q, k 向量，我们用复数的指数形式对其进行表示，即：

$$\begin{aligned} \boldsymbol{f}(\boldsymbol{q}, m) &= R_f(\boldsymbol{q}, m) e^{\mathrm{i}\Theta_f(\boldsymbol{q}, m)} \\ \boldsymbol{f}(\boldsymbol{k}, n) &= R_f(\boldsymbol{k}, n) e^{\mathrm{i}\Theta_f(\boldsymbol{k}, n)} \\ \boldsymbol{g}(\boldsymbol{q}, \boldsymbol{k}, m - n) &= R_g(\boldsymbol{q}, \boldsymbol{k}, m - n) e^{\mathrm{i}\Theta_g(\boldsymbol{q}, \boldsymbol{k}, m - n)} \end{aligned}$$

其中， R 代表向量模长。

那么代入方程后就得到方程组（共轭复数，所以虚部系数为负值）：

$$\begin{aligned} R_f(\boldsymbol{q}, m) R_f(\boldsymbol{k}, n) &= R_g(\boldsymbol{q}, \boldsymbol{k}, m - n) \\ \Theta_f(\boldsymbol{q}, m) - \Theta_f(\boldsymbol{k}, n) &= \Theta_g(\boldsymbol{q}, \boldsymbol{k}, m - n) \end{aligned}$$

如果我们根据方程组求出了 Rf(q, m) 和 θf(q, m) 的表达式，也就求解出了 f(· , m) 的表达式。

2.1.1 方程一推导

对于第一个方程，代入 m=n 得到：

$$R_f(\boldsymbol{q}, m) R_f(\boldsymbol{k}, m) = R_g(\boldsymbol{q}, \boldsymbol{k}, 0) = R_f(\boldsymbol{q}, 0) R_f(\boldsymbol{k}, 0) = \|\boldsymbol{q}\| \|\boldsymbol{k}\|$$

最后一个等号源于初始条件 f(q, 0) = q， f(k, 0) = k。 （因为m=0时，f(q, 0) = q，而Rf为模长，因此就是q的模长）

所以现在我们可以很简单地设：

$$R_f(\boldsymbol{q}, m) = \|\boldsymbol{q}\|, R_f(\boldsymbol{k}, m) = \|\boldsymbol{k}\|$$

即**它不依赖于m**。

💡 **结论一：R(q, m) = || q ||**

2.1.2 方程二推导

对于第二个方程，同样带入 m=n 得到：

$$\Theta_f(\boldsymbol{q},m)-\Theta_f(\boldsymbol{k},m)=\Theta_g(\boldsymbol{q},\boldsymbol{k},0)=\Theta_f(\boldsymbol{q},0)-\Theta_f(\boldsymbol{k},0)=\Theta(\boldsymbol{q})-\Theta(\boldsymbol{k})$$

这里的Θ(q), Θ(k)是 q,k 本身的幅角，最后一个等号同样源于初始条件。根据上式得到：

$$\Theta_f(\boldsymbol{q},m)-\Theta(\boldsymbol{q})=\Theta_f(\boldsymbol{k},m)-\Theta(\boldsymbol{k})$$

所以q和k并没有影响

$$\Theta_f(\boldsymbol{q},m)-\Theta(\boldsymbol{q})$$

的值，所以上式应该是一个**只与m相关，跟q无关的函数**，记为 ρ(m)，即：

$$\Theta_f(\boldsymbol{q},m)=\Theta(\boldsymbol{q})+\varphi(m)$$

原始的方程二为：

$$\Theta_f(\boldsymbol{q},m)-\Theta_f(\boldsymbol{k},n)=\Theta_g(\boldsymbol{q},\boldsymbol{k},m-n)$$

带入 n = m - 1，得：

$$\varphi(m)-\varphi(m-1)=\Theta_g(\boldsymbol{q},\boldsymbol{k},1)+\Theta(\boldsymbol{k})-\Theta(\boldsymbol{q})$$

首先将 $n=m-1$ 带入式 (5) 的第二个式子，得到

$$\Theta_f(\boldsymbol{q},m)-\Theta_f(\boldsymbol{k},m-1)=\Theta_g(\boldsymbol{q},\boldsymbol{k},1)$$

上式两边同减 $\Theta(\boldsymbol{q})$ 得

$$\Theta_f(\boldsymbol{q},m)-\Theta(\boldsymbol{q})-\Theta_f(\boldsymbol{k},m-1)=\Theta_g(\boldsymbol{q},\boldsymbol{k},1)-\Theta(\boldsymbol{q})$$

因为 $\Theta_f(\boldsymbol{q},m)-\Theta(\boldsymbol{q})=\varphi(m)$，所以

$$\varphi(m)-\Theta_f(\boldsymbol{k},m-1)=\Theta_g(\boldsymbol{q},\boldsymbol{k},1)-\Theta(\boldsymbol{q})$$

上式两边同加 $\Theta(\boldsymbol{k})$ 得

$$\varphi(m)+\Theta(\boldsymbol{k})-\Theta_f(\boldsymbol{k},m-1)=\Theta_g(\boldsymbol{q},\boldsymbol{k},1)+\Theta(\boldsymbol{k})-\Theta(\boldsymbol{q})$$

因为 $\Theta(\boldsymbol{q})-\Theta_f(\boldsymbol{q},m)=\Theta(\boldsymbol{k})-\Theta_f(\boldsymbol{k},m)$，所以

$$\varphi(m)+\Theta(\boldsymbol{q})-\Theta_f(\boldsymbol{q},m-1)=\Theta_g(\boldsymbol{q},\boldsymbol{k},1)+\Theta(\boldsymbol{k})-\Theta(\boldsymbol{q})$$

最后因为 $\Theta_f(\boldsymbol{q},m-1)-\Theta(\boldsymbol{q})=\varphi(m-1)$，所以

$$\varphi(m)-\varphi(m-1)=\Theta_g(\boldsymbol{q},\boldsymbol{k},1)+\Theta(\boldsymbol{k})-\Theta(\boldsymbol{q})$$

即 **{ρ(m)} 是等差数列，设右端为θ ，那么就解得 ρ(m) = mθ**。

> 💡 **结论二：Θ(q, m) = Θ(q) + ρ(m) = Θ(q) + mθ**

2.1.3 结论

根据方程一和方程二的求解，当：

$$\boldsymbol{f}(\boldsymbol{q},m)=R_f(\boldsymbol{q},m)e^{\mathrm{i}\Theta_f(\boldsymbol{q},m)}$$

中，R(q, m) = || q || 且 θ(q, m) = θ(q) + ρ(m) 时，f(·, m) 包含绝对位置信息，并且满足 f(q, m) * f(k, n) 包含相对位置信息。

2.3 编码形式

综上，我们得到二维情况下用复数表示的RoPE：

$$\boldsymbol{f}(\boldsymbol{q},m)=R_f(\boldsymbol{q},m)e^{\mathrm{i}\Theta_f(\boldsymbol{q},m)}=\|q\|e^{\mathrm{i}(\Theta(\boldsymbol{q})+m\theta)}=\boldsymbol{q}e^{\mathrm{i}m\theta}$$

$$
\begin{aligned}
\|q\|e^{\mathrm{i}(\Theta(q)+m\theta)} &= \|q\|(e^{\mathrm{i}\Theta(q)}\cdot e^{\mathrm{i}m\theta}) \\
&= (\|q\|e^{\mathrm{i}\Theta(q)})e^{\mathrm{i}m\theta} \\
&= \boldsymbol{q}e^{\mathrm{i}m\theta}
\end{aligned}
$$

根据复数乘法的几何意义，该变换实际上对应着向量的旋转，所以我们称之为“**旋转式位置编码**”，它还可以写成矩阵形式：

$$
\boldsymbol{f}(\boldsymbol{q}, m) = \begin{pmatrix} \cos m\theta & -\sin m\theta \\ \sin m\theta & \cos m\theta \end{pmatrix} \begin{pmatrix} q_0 \\ q_1 \end{pmatrix}
$$

> 💡 **相当于 q 通过乘以一个 ||z|| = 1 的虚数，实现了向量旋转**

由于内积满足线性叠加性，因此任意偶数维的RoPE，我们都可以表示为二维情形的拼接，即：

$$
\underbrace{\begin{pmatrix}
\cos m\theta_0 & -\sin m\theta_0 & 0 & 0 & \cdots & 0 & 0 \\
\sin m\theta_0 & \cos m\theta_0 & 0 & 0 & \cdots & 0 & 0 \\
0 & 0 & \cos m\theta_1 & -\sin m\theta_1 & \cdots & 0 & 0 \\
0 & 0 & \sin m\theta_1 & \cos m\theta_1 & \cdots & 0 & 0 \\
\vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\
0 & 0 & 0 & 0 & \cdots & \cos m\theta_{d/2-1} & -\sin m\theta_{d/2-1} \\
0 & 0 & 0 & 0 & \cdots & \sin m\theta_{d/2-1} & \cos m\theta_{d/2-1}
\end{pmatrix}}_{\boldsymbol{W}_m}
\begin{pmatrix} q_0 \\ q_1 \\ q_2 \\ q_3 \\ \vdots \\ q_{d-2} \\ q_{d-1} \end{pmatrix}
$$

也就是说，给位置为 m 的向量 q 乘上矩阵 Wm、位置为 n 的向量 k 乘上矩阵 Wn，用变换后的 Q, K 序列做Attention，那么**Attention就自动包含相对位置信息了**，因为成立恒等式：

$$
(\boldsymbol{W}_m\boldsymbol{q})^\top(\boldsymbol{W}_n\boldsymbol{k}) = \boldsymbol{q}^\top\boldsymbol{W}_m^\top\boldsymbol{W}_n\boldsymbol{k} = \boldsymbol{q}^\top\boldsymbol{W}_{n-m}\boldsymbol{k}
$$

总可以利用积化和差公式，使得 cos 或者 sin 中带有 (m − n) 项。

值得指出的是，**Wm是一个正交矩阵，它不会改变向量的模长，因此通常来说它不会改变原模型的稳定性**。

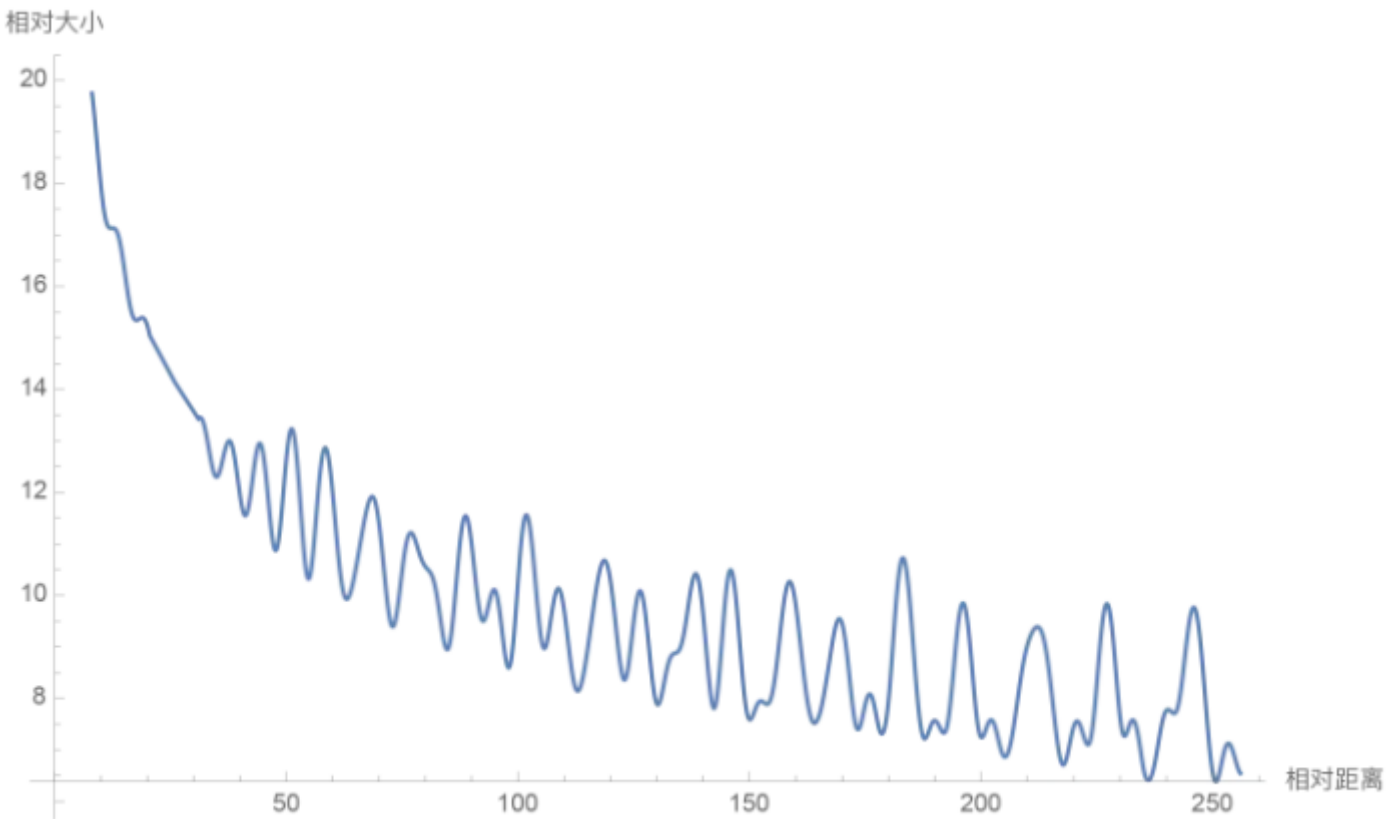
由于Wm的稀疏性，所以直接用矩阵乘法来实现会很浪费算力，推荐通过下述方式来实现RoPE（结果完全一致，避免稀疏矩阵乘法）：

$$\begin{pmatrix} q_0 \\ q_1 \\ q_2 \\ q_3 \\ \vdots \\ q_{d-2} \\ q_{d-1} \end{pmatrix} \otimes \begin{pmatrix} \cos m\theta_0 \\ \cos m\theta_0 \\ \cos m\theta_1 \\ \cos m\theta_1 \\ \vdots \\ \cos m\theta_{d/2-1} \\ \cos m\theta_{d/2-1} \end{pmatrix} + \begin{pmatrix} -q_1 \\ q_0 \\ -q_3 \\ q_2 \\ \vdots \\ -q_{d-1} \\ q_{d-2} \end{pmatrix} \otimes \begin{pmatrix} \sin m\theta_0 \\ \sin m\theta_0 \\ \sin m\theta_1 \\ \sin m\theta_1 \\ \vdots \\ \sin m\theta_{d/2-1} \\ \sin m\theta_{d/2-1} \end{pmatrix}$$

其中⊗是逐位对应相乘，即计算框架中的⊠运算。从这个实现也可以看到，RoPE可以视为是乘性位置编码的变体。

2.4 远程衰减特性

RoPE 能够带来一定的远程衰减特性（不同位置的 position encoding 点乘结果会随着相对位置的增加而递减）：



3. RoPE效果

苏剑林团队使用 RoPE 改造了 WoBERT 模型，得到 RoFormer 模型，它跟其他模型的结构对比如下：

	BERT	WoBERT	NEZHA	RoFormer
token单位	字	词	字	词
位置编码	绝对位置	绝对位置	经典式相对位置	RoPE

在训练上，以WoBERT Plus为基础，采用了多个长度和batch size交替训练的方式，让模型能提前适应不同的训练场景：

	maxlen	batch size	训练步数	最终loss	最终acc
1	512	256	20万	1.73	65.0%
2	1536	256	1.25万	1.61	66.8%
3	256	256	12万	1.75	64.6%
4	128	512	8万	1.83	63.4%
5	1536	256	1万	1.58	67.4%
6	512	512	3万	1.66	66.2%

从表格还可以看到，**增大序列长度，预训练的准确率反而有所提升**，这侧面体现了RoFormer长文本语义的处理效果，也体现了**RoPE具有良好的外推能力**。在短文本任务上，RoFormer与WoBERT的表现类似，RoFormer的主要特点是可以直接处理任意长的问题。下面是在CAIL2019-SCM任务上的实验结果：

	验证集	测试集
BERT-512	64.13%	67.77%
WoBERT-512	64.07%	68.10%
RoFormer-512	64.13%	68.29%
RoFormer-1024	**66.07%**	**69.79%**

4. 结论

从理论上来看，RoPE与Sinusoidal位置编码有些相通之处，但RoPE不依赖于泰勒展开，更具严谨性与可解释性；从预训练模型RoFormer的结果来看，RoPE具有良好的外推性，应用到Transformer中体现出较好的处理长文本的能力。此外，RoPE还是目前唯一一种可用于线性Attention的相对位置编码。