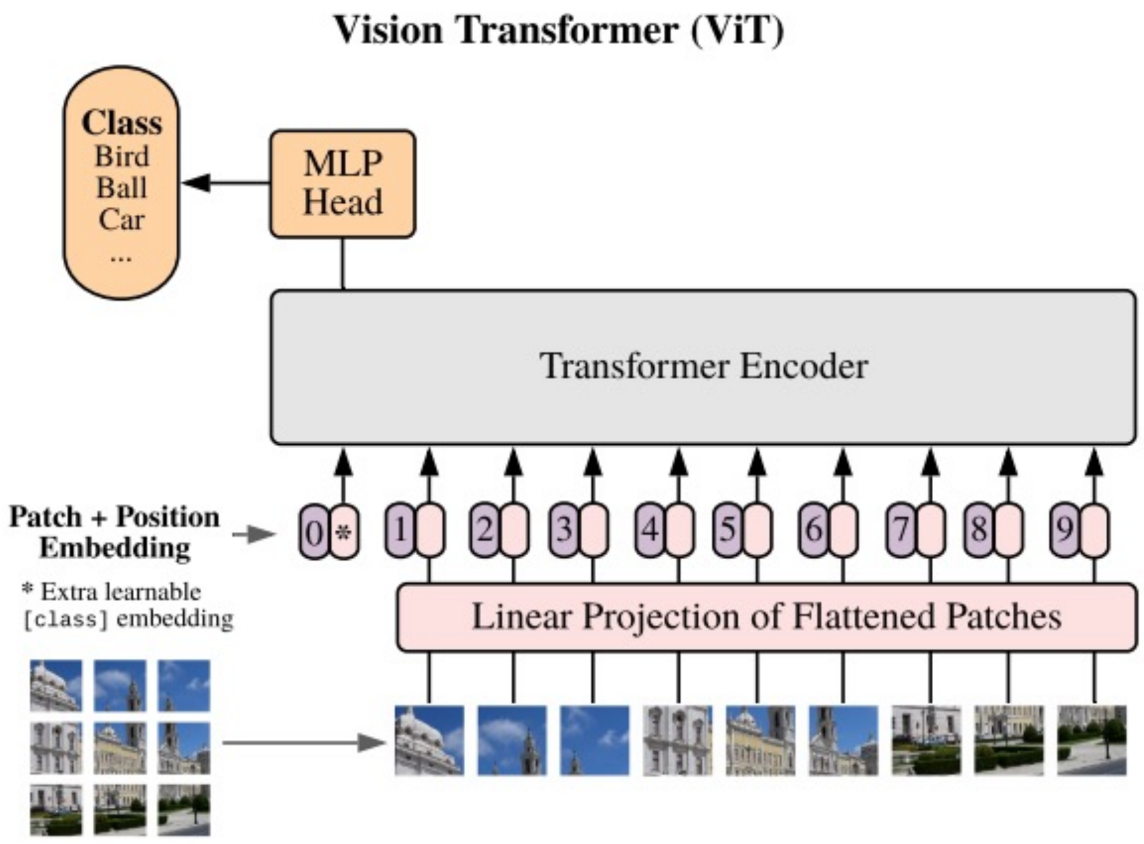
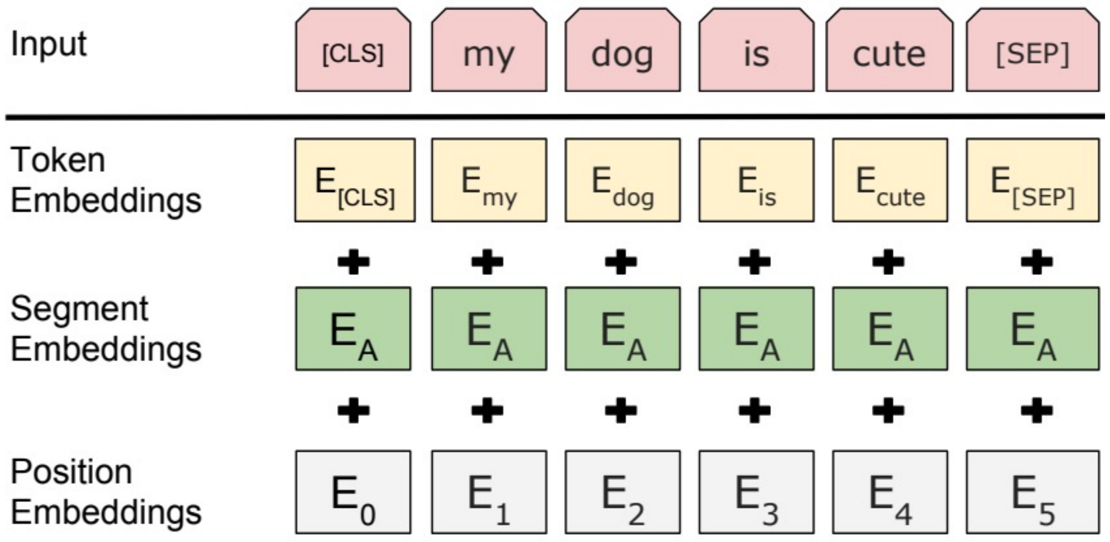
# Rotary Position Encoding

# Transformer

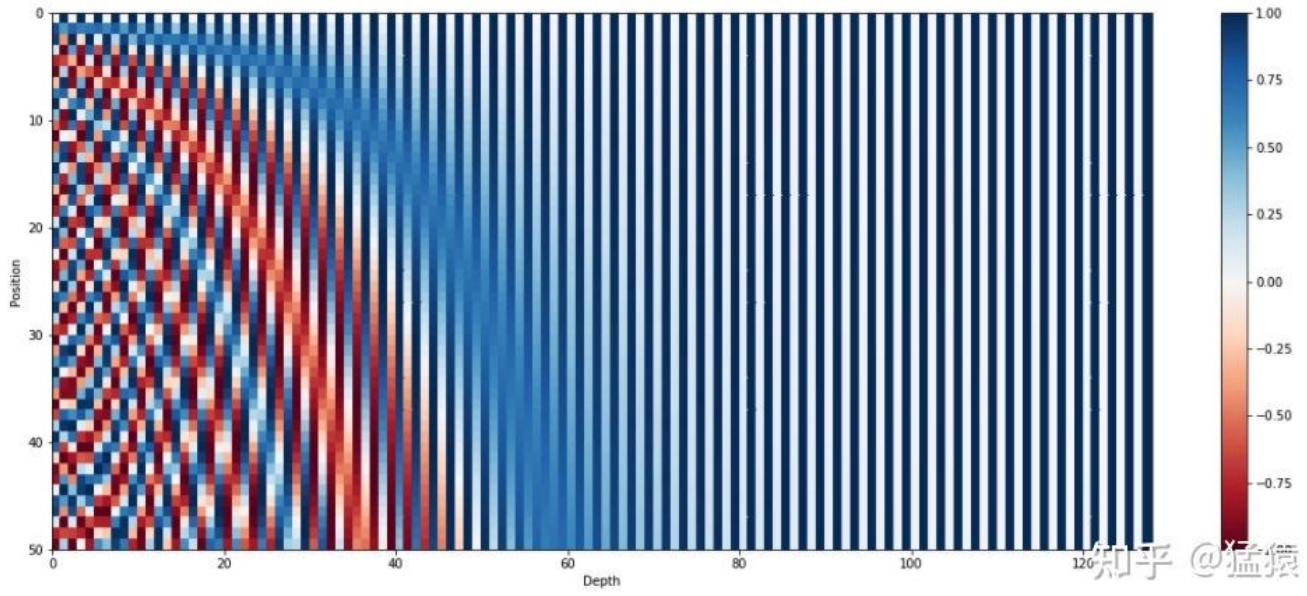


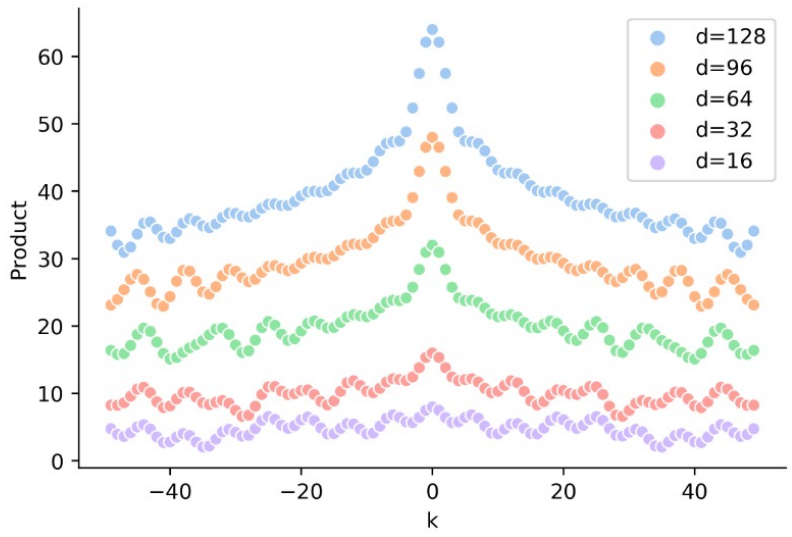



# Absolute PE

$$PE_{t,2i} = \sin(t/10000^{2i/d}),$$
$$PE_{t,2i+1} = \cos(t/10000^{2i/d}),$$

$$\begin{aligned} PE_t^T PE_{t+k} &= \sum_{j=0}^{\frac{d}{2}-1} [\sin(c_j t)\sin(c_j(t+k)) \\ &\quad + \cos(c_j t)\cos(c_j(t+k))] \\ &= \sum_{j=0}^{\frac{d}{2}-1} \cos(c_j(t-(t+k))) \\ &= \sum_{j=0}^{\frac{d}{2}-1} \cos(c_j k), \end{aligned}$$



# Absolute PE

$$PE_t^T W_Q^T W_K PE_{t+k}$$

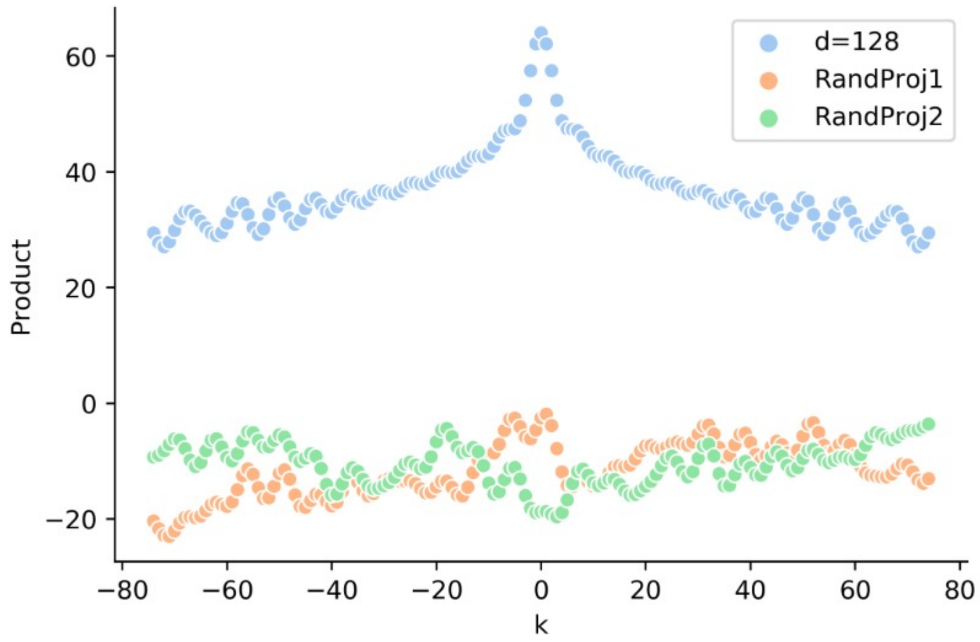



Figure 4: The upper line is the product between $PE_t^T PE_{t+k}$. The lower two lines are the products of $PE_t^T W PE_{t+k}$ with two random $W$s. Although $PE_t^T PE_{t+k}$ can reflect the distance, the $PE_t^T W PE_{t+k}$ has no clear pattern.

# Relative PE

$$f_q(\boldsymbol{x}_m) := \boldsymbol{W}_q \boldsymbol{x}_m$$

$$f_k(\boldsymbol{x}_n, n) := \boldsymbol{W}_k(\boldsymbol{x}_n + \tilde{\boldsymbol{p}}_r^k)$$

$$f_v(\boldsymbol{x}_n, n) := \boldsymbol{W}_v(\boldsymbol{x}_n + \tilde{\boldsymbol{p}}_r^v)$$

$$r = \mathrm{clip}(m - n, r_{\mathrm{min}}, r_{\mathrm{max}})$$

| $k$ | EN-DE BLEU |
|---|---|
| 0 | 12.5 |
| 1 | 25.5 |
| 2 | 25.8 |
| 4 | 25.9 |
| 16 | 25.8 |
| 64 | 25.9 |
| 256 | 25.8 |

# Relative PE

$$\boldsymbol{q}_m^\intercal \boldsymbol{k}_n = \boldsymbol{x}_m^\intercal \boldsymbol{W}_q^\intercal \boldsymbol{W}_k \boldsymbol{x}_n + \boldsymbol{x}_m^\intercal \boldsymbol{W}_q^\intercal \boldsymbol{W}_k \boldsymbol{p}_n + \boldsymbol{p}_m^\intercal \boldsymbol{W}_q^\intercal \boldsymbol{W}_k \boldsymbol{x}_n + \boldsymbol{p}_m^\intercal \boldsymbol{W}_q^\intercal \boldsymbol{W}_k \boldsymbol{p}_n,$$

$$\boldsymbol{q}_m^\intercal \boldsymbol{k}_n = \boldsymbol{x}_m^\intercal \boldsymbol{W}_q^\intercal \boldsymbol{W}_k \boldsymbol{x}_n + \boldsymbol{x}_m^\intercal \boldsymbol{W}_q^\intercal \widetilde{\boldsymbol{W}}_k \tilde{\boldsymbol{p}}_{m-n} + \mathbf{u}^\intercal \boldsymbol{W}_q^\intercal \boldsymbol{W}_k \boldsymbol{x}_n + \mathbf{v}^\intercal \boldsymbol{W}_q^\intercal \widetilde{\boldsymbol{W}}_k \tilde{\boldsymbol{p}}_{m-n}$$

$$\boldsymbol{q}_m^\intercal \boldsymbol{k}_n = \boldsymbol{x}_m^\intercal \boldsymbol{W}_q^\intercal \boldsymbol{W}_k \boldsymbol{x}_n + b_{i,j}$$

$$\boldsymbol{q}_m^\intercal \boldsymbol{k}_n = \boldsymbol{x}_m^\intercal \boldsymbol{W}_q^\intercal \boldsymbol{W}_k \boldsymbol{x}_n + \boldsymbol{p}_m^\intercal \mathbf{U}_q^\intercal \mathbf{U}_k \boldsymbol{p}_n + b_{i,j}$$

$$\boldsymbol{q}_m^\intercal \boldsymbol{k}_n = \boldsymbol{x}_m^\intercal \boldsymbol{W}_q^\intercal \boldsymbol{W}_k \boldsymbol{x}_n + \boldsymbol{x}_m^\intercal \boldsymbol{W}_q^\intercal \boldsymbol{W}_k \tilde{\boldsymbol{p}}_{m-n} + \tilde{\boldsymbol{p}}_{m-n}^\intercal \boldsymbol{W}_q^\intercal \boldsymbol{W}_k \boldsymbol{x}_n$$

# RoPE

$$\langle f_q(\boldsymbol{x}_m, m), f_k(\boldsymbol{x}_n, n)\rangle = g(\boldsymbol{x}_m, \boldsymbol{x}_n, m-n).$$

$$f_q(\boldsymbol{x}_m, m) = (\boldsymbol{W}_q\boldsymbol{x}_m)e^{im\theta}$$
$$f_k(\boldsymbol{x}_n, n) = (\boldsymbol{W}_k\boldsymbol{x}_n)e^{in\theta}$$
$$g(\boldsymbol{x}_m, \boldsymbol{x}_n, m-n) = \mathrm{Re}[(\boldsymbol{W}_q\boldsymbol{x}_m)(\boldsymbol{W}_k\boldsymbol{x}_n)^* e^{i(m-n)\theta}]$$

$$f_{\{q,k\}}(\boldsymbol{x}_m, m) = \begin{pmatrix} \cos m\theta & -\sin m\theta \\ \sin m\theta & \cos m\theta \end{pmatrix} \begin{pmatrix} W^{(11)}_{\{q,k\}} & W^{(12)}_{\{q,k\}} \\ W^{(21)}_{\{q,k\}} & W^{(22)}_{\{q,k\}} \end{pmatrix} \begin{pmatrix} x^{(1)}_m \\ x^{(2)}_m \end{pmatrix}$$

假设Ra表示角度为a的旋转矩阵，那么R具有如下性质：

1. Ra^T = R(-a)
2. Ra Rb = R(a+b)

回到旋转位置编码，我们可以去证明 <RaX, RbY> = <X, R(b-a)Y>， 证明如下：

<RaX, RbY>

= (RaX)^T RbY

= X^T Ra^T RbY

= X^T R(b-a) Y

= <X, R(b-a)Y>

推导过程可见: https://zhuanlan.zhihu.com/p/642884818

# RoPE

$$f_{\{q,k\}}(\boldsymbol{x}_m, m) = \boldsymbol{R}^d_{\Theta,m} \boldsymbol{W}_{\{q,k\}} \boldsymbol{x}_m$$

$$\boldsymbol{R}^d_{\Theta,m} = \begin{pmatrix} \cos m\theta_1 & -\sin m\theta_1 & 0 & 0 & \cdots & 0 & 0 \\ \sin m\theta_1 & \cos m\theta_1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & \cos m\theta_2 & -\sin m\theta_2 & \cdots & 0 & 0 \\ 0 & 0 & \sin m\theta_2 & \cos m\theta_2 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & \cos m\theta_{d/2} & -\sin m\theta_{d/2} \\ 0 & 0 & 0 & 0 & \cdots & \sin m\theta_{d/2} & \cos m\theta_{d/2} \end{pmatrix}$$

$$\Theta = \{\theta_i = 10000^{-2(i-1)/d}, i \in [1, 2, ..., d/2]\}.$$

$$\boldsymbol{q}_m^\intercal \boldsymbol{k}_n = (\boldsymbol{R}^d_{\Theta,m} \boldsymbol{W}_q \boldsymbol{x}_m)^\intercal (\boldsymbol{R}^d_{\Theta,n} \boldsymbol{W}_k \boldsymbol{x}_n) = \boldsymbol{x}^\intercal \boldsymbol{W}_q R^d_{\Theta,n-m} \boldsymbol{W}_k \boldsymbol{x}_n$$

$$\boldsymbol{R}^d_{\Theta,m} \boldsymbol{x} = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ \vdots \\ x_{d-1} \\ x_d \end{pmatrix} \otimes \begin{pmatrix} \cos m\theta_1 \\ \cos m\theta_1 \\ \cos m\theta_2 \\ \cos m\theta_2 \\ \vdots \\ \cos m\theta_{d/2} \\ \cos m\theta_{d/2} \end{pmatrix} + \begin{pmatrix} -x_2 \\ x_1 \\ -x_4 \\ x_3 \\ \vdots \\ -x_d \\ x_{d-1} \end{pmatrix} \otimes \begin{pmatrix} \sin m\theta_1 \\ \sin m\theta_1 \\ \sin m\theta_2 \\ \sin m\theta_2 \\ \vdots \\ \sin m\theta_{d/2} \\ \sin m\theta_{d/2} \end{pmatrix}$$

# RoPE

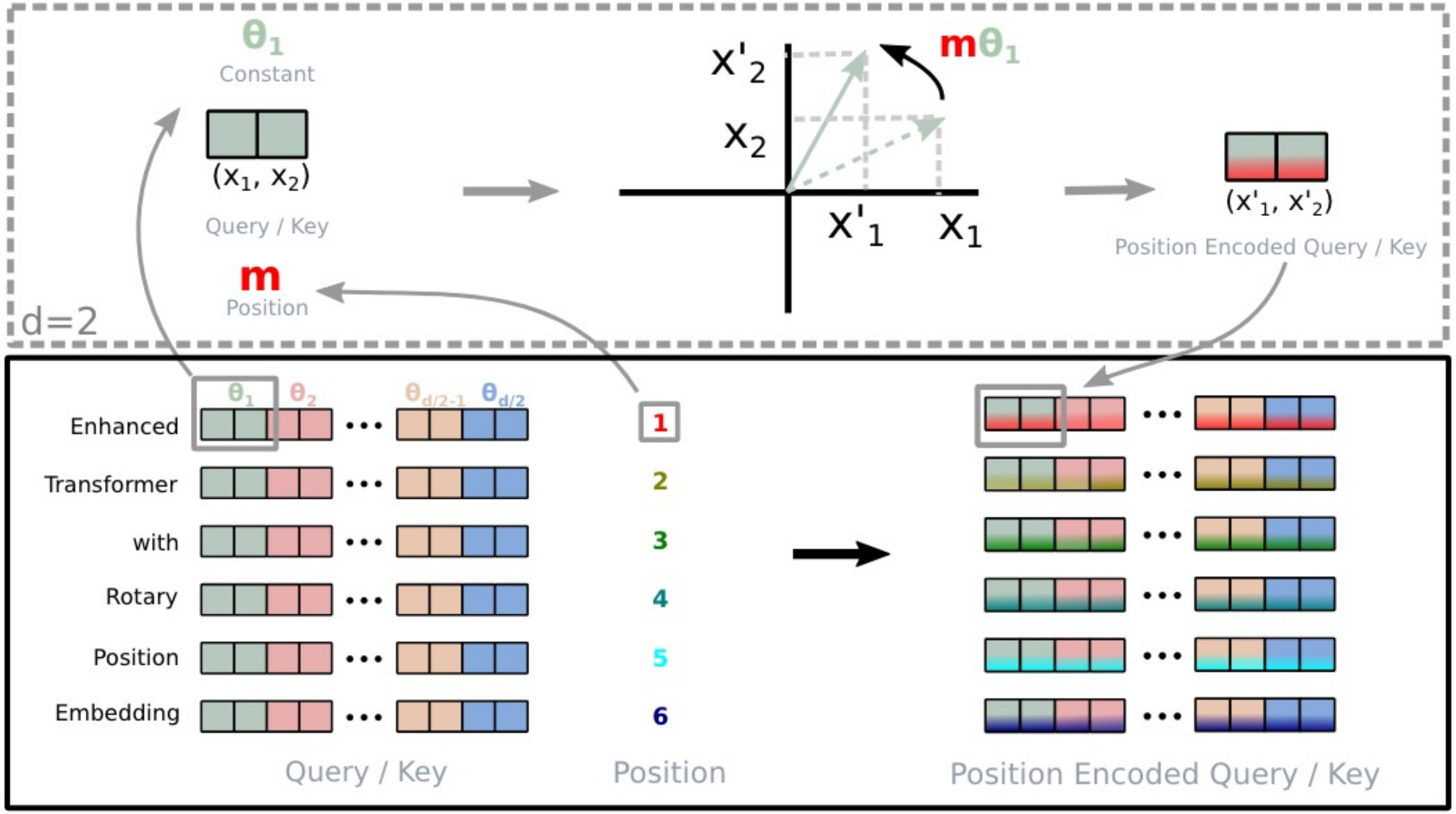



Figure 1: Implementation of Rotary Position Embedding(RoPE).

# RoPE

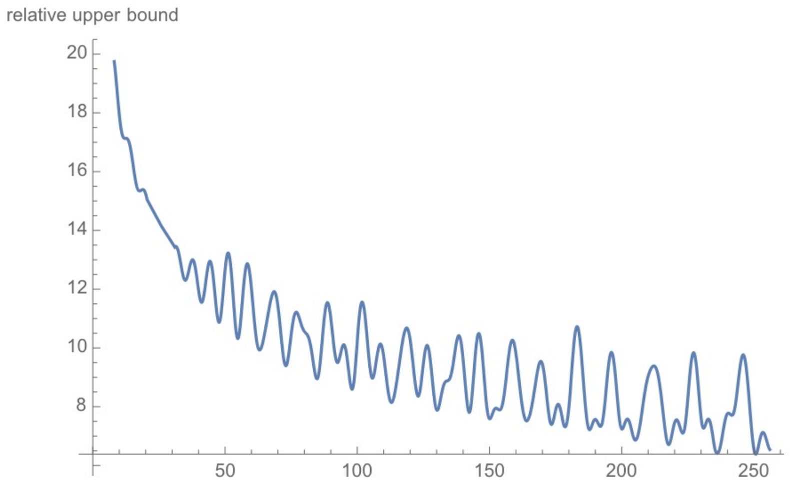



Figure 2: Long-term decay of RoPE.

$$(\boldsymbol{R}^d_{\Theta,m}\boldsymbol{W}_q\boldsymbol{x}_m)^\intercal(\boldsymbol{R}^d_{\Theta,n}\boldsymbol{W}_k\boldsymbol{x}_n) = \mathrm{Re}\left[\sum_{i=0}^{d/2-1}\boldsymbol{q}_{[2i:2i+1]}\boldsymbol{k}^*_{[2i:2i+1]}e^{i(m-n)\theta_i}\right] \tag{35}$$

where $\boldsymbol{q}_{[2i:2i+1]}$ represents the $2i^{th}$ to $(2i+1)^{th}$ entries of $\boldsymbol{q}$. Denote $h_i = \boldsymbol{q}_{[2i:2i+1]}\boldsymbol{k}^*_{[2i:2i+1]}$ and $S_j = \sum_{i=0}^{j-1} e^{i(m-n)\theta_i}$, and let $h_{d/2} = 0$ and $S_0 = 0$, we can rewrite the summation using Abel transformation

$$\sum_{i=0}^{d/2-1}\boldsymbol{q}_{[2i:2i+1]}\boldsymbol{k}^*_{[2i:2i+1]}e^{i(m-n)\theta_i} = \sum_{i=0}^{d/2-1}h_i(S_{i+1}-S_i) = -\sum_{i=0}^{d/2-1}S_{i+1}(h_{i+1}-h_i). \tag{36}$$

Thus,

$$\begin{aligned}\left|\sum_{i=0}^{d/2-1}\boldsymbol{q}_{[2i:2i+1]}\boldsymbol{k}^*_{[2i:2i+1]}e^{i(m-n)\theta_i}\right| &= \left|\sum_{i=0}^{d/2-1}S_{i+1}(h_{i+1}-h_i)\right| \\ &\le \sum_{i=0}^{d/2-1}|S_{i+1}||(h_{i+1}-h_i)| \\ &\le \left(\max_i |h_{i+1}-h_i|\right)\sum_{i=0}^{d/2-1}|S_{i+1}|\end{aligned} \tag{37}$$

Note that the value of $\frac{1}{d/2}\sum_{i=1}^{d/2}|S_i|$ decay with the relative distance $m-n$ increases by setting $\theta_i = 10000^{-2i/d}$, as shown in Figure (**2**).

# RoPE

| Model | BLEU |
|---|---|
| Transformer-baseVaswani et al. [2017] | 27.3 |
| RoFormer | **27.5** |

| Model | MRPC | SST-2 | QNLI | STS-B | QQP | MNLI(m/mm) |
|---|---|---|---|---|---|---|
| BERTDevlin et al. [2019] | 88.9 | 93.5 | 90.5 | 85.8 | 71.2 | 84.6/83.4 |
| RoFormer | **89.5** | 90.7 | 88.0 | **87.0** | **86.4** | 80.2/79.8 |

| Model | Validation | Test |
|---|---|---|
| BERT-512 | 64.13% | 67.77% |
| WoBERT-512 | 64.07% | 68.10% |
| **RoFormer-512** | 64.13% | 68.29% |
| **RoFormer-1024** | **66.07%** | **69.79%** |

# Thought

为什么直接外推不好

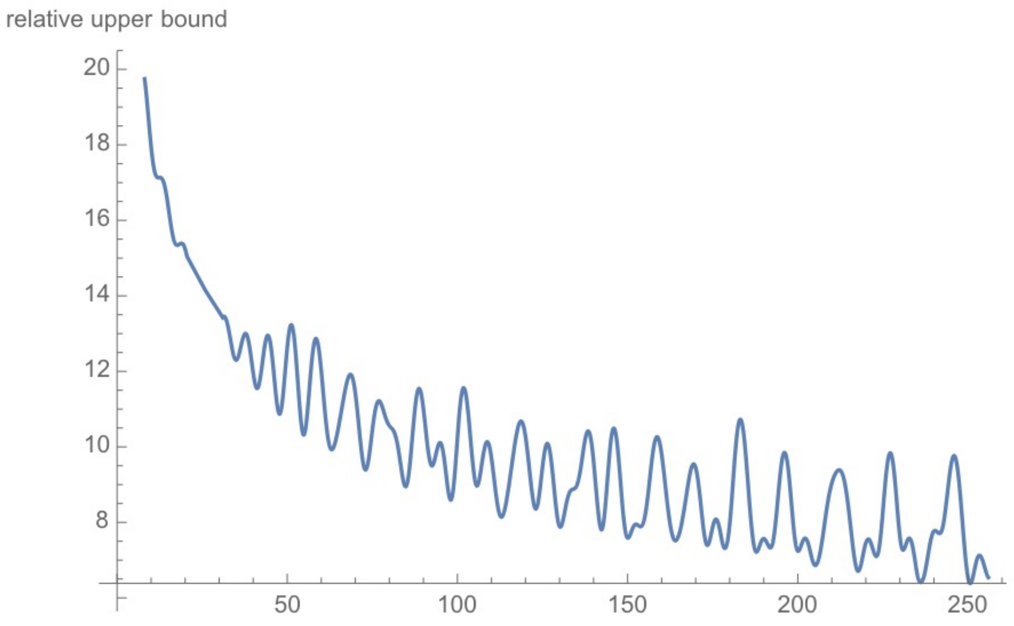
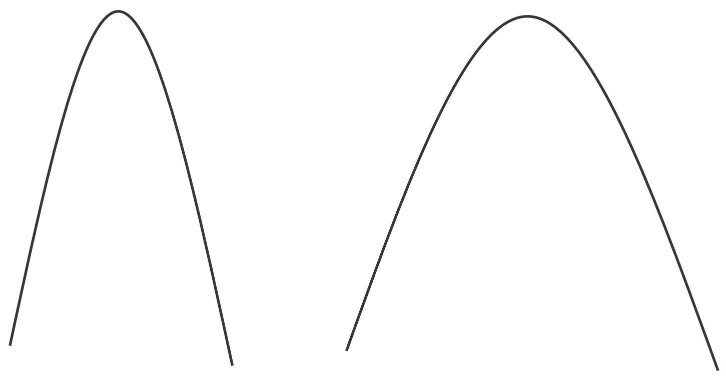



Figure 2: Long-term decay of RoPE.

内插是可行的，但是效果会下降



# Thought

NTK Aware Rope

$$\Theta = \{\theta_i = 10000^{-2(i-1)/d}, i \in [1, 2, ..., d/2]\}.$$

# Thank You