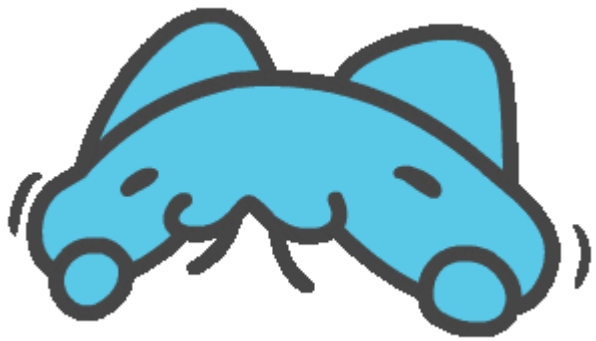
零基础 学透
旋转位置编码RoPE
及其外推方法

三个问题：
我们为什么需要**位置编码**？
我们为什么需要**RoPE**？
什么是**外推**？怎么**外推**？



我们为什么需要**位置编码**?

- 传统的词嵌入模型Word2Vec、GloVe

  词嵌入向量只包含词义本身，不包含单词位置的信息

- 遇到的问题

  他吃早饭 vs 早饭吃他

- 两个句子的词向量集合完全相同（都是 “他”“吃”“早饭” 的向量），但语义完全相反。而不带Attention Mask的纯Attention模型是全对称的，即天然地满足f(x,y)=f(y,x)

- 如何解决?

我们为什么需要**位置编码**?

- 绝对位置编码

把词嵌入向量x_k与其位置向量p_k相加，其中p_k只依赖于位置信息（位置k）　　$x_k + p_k$

- 正余弦位置编码（Transformer)

$$p_{k,2i} = \sin\left(\frac{k}{10000^{\frac{2i}{d}}}\right),\ p_{k,2i+1} = \cos\left(\frac{k}{10000^{\frac{2i}{d}}}\right)$$

$p_{k,2i}$ 和 $p_{k,2i+1}$是位置k的编码向量的第2i和第2i+1个分量，d是位置向量的维度，k表示第几个token（位置）

我们为什么需要RoPE?

- 正余弦位置编码具备一定的表达相对位置的能力

  token i 的词嵌入向量为x_i，位置向量为$PE(i)$;

  token j 的词嵌入向量为x_{j}，位置向量为$PE(\mathrm{j})$

  attention得分的本质是计算 token i 对 token j 的关注度

  x_i与x_{j}做内积，模型可以但是需要额外 “费力” 从内积结果中分离出 “位置差异” 和 “语义相关度”

我们为什么需要RoPE?

- 如果不把位置信息**加**到词嵌入向量上，而是**乘**上去

  这就是RoPE

  给定位置为m的token q_m，位置为n的token k_n

  词嵌入向量与绝对位置信息相乘，得到 $q_m e^{im\theta}$，$k_n e^{in\theta}$

  这时候我们再来计算一下两个token之间的关注度

  $$<q_m e^{im\theta}, k_n e^{in\theta}>$$

  $$= \mathrm{Re}[(q_m e^{im\theta})(k_n e^{in\theta})^*]$$

  $$= \mathrm{Re}[q_m k_n^* e^{i(m-n)\theta}]$$

我们为什么需要RoPE?

- 为什么叫旋转位置编码?

  词嵌入向量与绝对位置信息相乘相当于把向量进行旋转

- RoPE的优点:

  把相对位置信息显式融入注意力得分

  具备外推的潜力

什么是**外推**？怎么**外推**？

- 外推就是要解决**训练**和**预测**长度不一致的问题
- 两个成因：
  - 预测的时候用到了没训练过的位置编码
  - 预测的时候注意力机制处理的token远超训练时候的
- 一个解决办法：“超强基线模型”
  - 局部attention
  - 掩码，预测时让每个token只能看到训练长度个token
- 实际效果有限
  - 可能的原因：开头的几个token很重要，不能mask掉

什么是**外推**？怎么**外推**？

- 那还有哪些外推方法呢？

  线性缩放（位置内插 Positional Interpolation）

  NTK-aware Scaled RoPE

  NTK-by-parts

  YaRN

  Leaky ReRoPE 和ReRoPE

什么是**外推**？怎么**外推**？

- 那还有哪些外推方法呢？
- 线性缩放（位置内插 Positional Interpolation）

  将超长的位置按比例压缩

  压缩了邻近Token的距离，严重扰乱了模型的局部分辨率
- 外推

  from 1—1—1—1—1—1—1

  to 1———1———1———1———1———1———1
- 内插

  from 1—1—1—1—1—1—1

  to 1-1-1-1-1-1-1

什么是**外推**？怎么**外推**？

- NTK-aware Scaled RoPE

  低频内插，高频外推 将外推压力平摊到每一个维度

  RoPE 中，维度越高→频率越低→波长越长

  训练位置范围：高频[0 ---- 511]低频 (L_train=512)

  推理位置范围：高频[0 ---------- 2047]低频 (L_infer=2048)
- 低频维度的波长极长，采取内插，压缩全局位置
- 高频维度的波长极短，采取外推，保留局部位置精细差异

  高频1—1—1—1—1—1—1—1—1—1低频

  高频1—— 1——1-1-1-1-1-1-1-1-1-1低频

什么是**外推**？怎么**外推**？

- NTK-aware Scaled RoPE 存在的问题

  波长大于原最大序列长度的那些低频分量，我们对他们进行了外推，会引入一些从未见过的旋转角度，这些旋转角度对应的正余弦值在训练过程中模型也从未见过

- 解决办法

  波长大于训练长度（低频），就只做内插

  波长很短（高频），就完全外推

  介于中间，就沿用NTK-aware Scaled RoPE

  这就是**NTK-by-parts**

什么是外推？怎么外推？

- YaRN Yet another RoPE extension method

YaRN: Efficient Context Window Extension of Large Language Models

Bowen Peng[1] **Jeffrey Quesnelle**[1] **Honglu Fan**[23] **Enrico Shippole**

[1]Nous Research [2]EleutherAI [3]University of Geneva

YaRN = NTK-by-parts + attention-scaling

attention-scaling就是attention score除以常数t

什么是外推？怎么外推？

- Leaky ReRoPE
- 只要w选定的是小于训练长度的，那么通过控制k，我们就可以在精确保持了局域性的前提下，使得所有位置编码不超过训练长度，简单直接地结合了直接外推和位置内插

$$
\begin{pmatrix}
0 & & & & & & & & & & \\
1 & 0 & & & & & & & & & \\
2 & 1 & 0 & & & & & & & & \\
\ddots & 2 & 1 & 0 & & & & & & & \\
w-1 & \ddots & 2 & 1 & 0 & & & & & & \\
w & w-1 & \ddots & 2 & 1 & 0 & & & & & \\
w+\frac{1}{k} & w & \ddots & \ddots & 2 & 1 & 0 & & & & \\
w+\frac{2}{k} & w+\frac{1}{k} & \ddots & \ddots & \ddots & 2 & 1 & 0 & & & \\
\ddots & w+\frac{2}{k} & \ddots & \ddots & \ddots & \ddots & 2 & 1 & 0 & & \\
\ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \\
\ddots & \ddots & \ddots & w+\frac{2}{k} & w+\frac{1}{k} & w & w-1 & \ddots & 2 & 1 & 0 \\
w+\frac{L-1-w}{k} & \ddots & \ddots & \ddots & w+\frac{2}{k} & w+\frac{1}{k} & w & w-1 & \ddots & 2 & 1 & 0
\end{pmatrix}
$$

什么是**外推**？怎么**外推**？

- $k \to \infty$

  Leaky ReRoPE

  变成**ReRoPE**

  这样，不管输入长度是多少，它的位置编码范围都不超过w

  有可能支持任意长度的Context

$$
\begin{pmatrix}
0 & & & & & & & & & & \\
1 & 0 & & & & & & & & & \\
2 & 1 & 0 & & & & & & & & \\
\ddots & 2 & 1 & 0 & & & & & & & \\
w-1 & \ddots & 2 & 1 & 0 & & & & & & \\
w & w-1 & \ddots & 2 & 1 & 0 & & & & & \\
w & w & \ddots & \ddots & 2 & 1 & 0 & & & & \\
w & w & \ddots & \ddots & \ddots & 2 & 1 & 0 & & & \\
\ddots & w & \ddots & \ddots & \ddots & \ddots & 2 & 1 & 0 & & \\
\ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \\
\ddots & \ddots & \ddots & w & w & w & w-1 & \ddots & 2 & 1 & 0 \\
w & \ddots & \ddots & \ddots & w & w & w & w-1 & \ddots & 2 & 1 & 0
\end{pmatrix}
$$