Transformer 面试知识

写出 self attention 的公式表达式

$$Softmax(\frac{QK^T}{\sqrt{d_k}})V \tag{1}$$

为什么式 (1) 中要对 QK 进行 scaled

让 sofrmax 输入的数据分布变好，数值进入梯度敏感区间，能防止梯度小时，让模型好训练

self-attention 一定要这样表达吗

不需要，能刻画相关性，相似性等建模方式都可以。最好速度快，模型好学，表达能力够

有其他方法不用除以 $\sqrt{d_k}$ 吗?

有，同上，只要能做到每层参数等梯度保持在训练敏感范围内，不要太大，不要太小，那么这个网络就不叫好训练

为什么transformer用Layer Norm？有什么用?

让神经网络各层参数输入数据分布变好，数值进入梯度敏感区间，能防止梯度消失，让模型好训练

为什么不用BN?

- NLP不定长，好多位置填0，影响其他样本非0参数的计算
- transforme模型比较大，batch size拉不大，容易变得不稳定

transformer为什么要用三个不一样的QKV

增强网络的容量和表达能力

为什么要多头

增强网络的容量和表达能力