**Transformer深度详解**

为什么要用transformer？

答：① 普通rnn（encoder\_output只作用于初始状态）有什么问题：

* Encoder编码成一个固定的向量，若encode的序列过长，会忘记前面的语义
* Decoder序列过长，会逐渐忘记encode\_output的语义

② 带有上下文的rnn(encode\_output与每一个decoder输入concat)有什么问题：

Decoder每一次输入都与相同的encode\_output合并，不太合理

③ 因此，attention就此诞生

Attention中为什么要有缩放因子1/sqrt(d\_k)?

答：在QK之后，数据分布就变成N(0,d\_k).

为什么要使用多头？

答：增强每一层attention的表现力。

为什么在维度不变的情况下，split就能实现多头？

答：split之后中各个头的子空间没有关联，各个的分布也会有所不同，故能体现多头。

为什么要加位置编码？

答：位置编码弥补了attention机制不能捕捉位置的缺陷。