## Transformer20个经典面试题讲解

Q1: Transformer为何使用多头注意力机制？（为什么不使用一个头）

A1:

---

Q2: Transformer为什么Q和K使用不同的权重矩阵生成，为何不能使用同一个值进行自身的点乘？ （注意和第一个问题的区别）

A2:

---

Q3: Transformer计算attention的时候为何选择点乘而不是加法？两者计算复杂度和效果上有什么区别？

A3:

---

Q4: 为什么在进行softmax之前需要对attention进行scaled（为什么除以dk的平方根），并使用公式推导进行讲解

A4:

---

Q5: 在计算attention score的时候如何对padding做mask操作？

A5:

---

Q6: 为什么在进行多头注意力的时候需要对每个head进行降维？（可以参考上面一个问题）

A6:

---

Q7: 大概讲一下Transformer的Encoder模块？

A7:

---

Q8: 为何在获取输入词向量之后需要对矩阵乘以embedding size的开方？意义是什么？

A8:

---

Q9: 简单介绍一下Transformer的位置编码？有什么意义和优缺点？

A9:

---

Q10: 你还了解哪些关于位置编码的技术，各自的优缺点是什么？

A10:

---

Q11: 简单讲一下Transformer中的残差结构以及意义。

A11:

---

Q12: 为什么transformer块使用LayerNorm而不是BatchNorm？LayerNorm 在Transformer的位置是哪里？

A12:

---

Q13: 简答讲一下BatchNorm技术，以及它的优缺点。

A13:

---

Q14: 简单描述一下Transformer中的前馈神经网络？使用了什么激活函数？相关优缺点？

A14:

---

Q15: Encoder端和Decoder端是如何进行交互的？（在这里可以问一下关于seq2seq的attention知识）

A15:

---

Q16: Decoder阶段的多头自注意力和encoder的多头自注意力有什么区别？（为什么需要decoder自注意力需要进行sequence mask)

A16:

---

Q17: Transformer的并行化提现在哪个地方？Decoder端可以做并行化吗？

A17:

---

Q18: 简单描述一下wordpiece model 和 byte pair encoding，有实际应用过吗？

A18:

---

Q19: Transformer训练的时候学习率是如何设定的？Dropout是如何设定的，位置在哪里？Dropout 在测试的需要有什么需要注意的吗？

A19:

---

Q20: 引申一个关于bert问题，bert的mask为何不学习transformer在attention处进行屏蔽score的技巧？

A20:

---