Token及模型参数准备篇

来自： AiGC面试宝典

宁静致远

2023年09月29日 12:06

1. 预训练数据 Token 重复 是否影响 模型性能?

- 多轮epoch的训练会降低模型性能；
- 更大规模的数据集会缓解重复epochs对模型性能下降的影响；
- 提高数据集的质量也无法挽救重复训练带来的过拟合；
- 小计算量模型的过拟合趋势与大计算量的差不多；
- 多样的训练目标不一定减轻多Epoch的性能下降；
- Dropout是一个被大语言模型忽视的正则技术，虽然慢，但是可以降低多epochs的影响；
- 在训练过程中逐渐使用dropout是有效的策略；

2. SFT需要训练Token数?

- 少量高质量、多样性的数据，也可以训练出效果优秀的SFT模型