????-?????????

??????????

1. 如果不能很好的拟合高层次的信息
    1. 容易出现过拟合现象
        1. dropout早期用于防止过拟合，适用于多 epoch的场景，不适用于 LLM
    2. 权重会反复摇摆，不能锁定高级的语义抽象，高级语义样本本来就少
        1. 每次梯度下降是每个权重单独改变，不能确保综合效果比原来更好 ???

????

1. 模型的表达空间的大小
2. 高层级数据的规模

???????????

1. 信息不够，不直接
    1. RoPE替代绝对位置编码，提供了直接的相对位置关系，自然语言的相对位置非常重要
2. 模型参数不够
3. 模型结构不好
    1. 没有足够的非线性表达能力
        1. 需要合理的时候激活层， GELU、Relu、Swish差别不大，更多考虑性能
        2. 分组查询注意力 GQA / MOE / 带门控的 GLU（Gated Linear Unit）
           能极大提高表达空间，用较少的参数
    2. 信息（层）映射的过渡不够平滑
        1. 旁路结构（ ResNet）减少模型动态
        2. MultiHEAD
    3. 宽度 vs深度
        1. 更深的模型表达更灵活，但训练更易不稳（梯度爆炸 /消失），这正是 RMSNorm与残差 /
           捷径连接试图缓解的问题
        2. 更宽的架构在推理时通常更快，因为并行度更好。代价是更高显存占用
    4. 少量「大专家」 vs 大量「小专家」
        1. 近来的趋势倾向于「更多、更小」的专家
4. 梯度下降怎么增强模型对高层级语义的敏感性，分类准确度?
    1. 按照大量数据的统计信息，引导模型按照高层极语义进行分割，而不是在低层级打转
        1. 梯度下降会把所有的样本按照一定的组织方式，编织到一个 非常大的多层级的递进式的
           选择空间里面去
        2. 根据已经有的输入 选择对应的知识空间分叉
           ，预测下一个字符会落到那个语义，再解码出最可能的输出符号
        3. 始终在选择向量空间里面最接近的答案，不保证是不是正确的，看似合理却错误的陈述
    2. 但是没有针对不正常、不存在的样本的训练，不正常的样本会被随机归类到某个类别 /
       向量空间里面去

1. 过拟合现象
2. 如果高层级的抽象语义能被提取出来，就可以进行准确分类
3. 目前 LLM没有针对性的进行，正样本训练，负样本训练，而是靠样本的数量和质量。

Revision #8
Created 2025-08-23 07:40:37 UTC by Colin
Updated 2026-04-29 07:34:33 UTC by Colin