论文显示大模型可删除一半注意力层

《What Matters In Transformers?》是一篇非常有趣的论文.研究发现，在像Llama这样的LLM中，你可以删除一半的注意力层，而不会显著降低模型性能。 概念相对简单。作者删除了注意力层、MLP层或整个Transformer块：• 删除整个Transformer块会导致显著的性能下降。• 删除MLP层会导致显著的性能下降。• 删除注意力层几乎不会导致性能下降！ 在Llama 2 70B模型中，即使删除了一半的注意力层（带来48%的加速），模型的基准测试分数也只下降了2.4%。作者最近还在论文中加入了Llama 3的结果，结果与Llama 2类似。 这些注意力层的删除并不是随机的，而是基于一个cosine 相似性评分：如果输入和输出非常相似，那么这个层是冗余的，可以删除。 这个结果非常引人注目，且可能与各种模型压缩技术结合，产生叠加效果。 此外，层的删除是一次性完成的（而非迭代方式），并且删除后不需要重新训练模型。不过，如果在删除后重新训练模型，可能会恢复部分性能损失。