https://hub.baai.ac.cn/view/29387

train

https://mingchao.wang/4KTgtnFc/#4llama-65b

6B模型, fp16 or bf16

model + grad + optimizer + CUDA_KERNAL

12 + 12 + 24 + 1.3 +batch

> 综上所述，整个模型所有的中间激活值的大小为 $l*(34bsh+5bas^2)+2bsh$。随着模型越来越大，l 是比较大的，所以有时会忽略 $2bsh$ 这一项，直接使用 $l*(34bsh+5bas^2)$ 来估计模型的中间激活值的大小。

inference

推理时候主要是模型的参数占的多，其它的是数据和中间结果等。

HuggingFace提供了Model Memory Calculator工具,可以精确估算特定模型的显存需求。

1.为什么大模型推理时显存涨的那么多还一直占着?

大语言模型进行推理时，显存涨得很多且一直占着显存不释放的原因主要有以下几点：

1. **模型参数占用显存**：大语言模型通常具有巨大的参数量，这些参数需要存储在显存中以供推理使用。因此，在推理过程中，模型参数会占用相当大的显存空间。
2. **输入数据占用显存**：进行推理时，需要将输入数据加载到显存中。对于大语言模型而言，输入数据通常也会占用较大的显存空间，尤其是对于较长的文本输入。
3. **中间计算结果占用显存**：在推理过程中，模型会进行一系列的计算操作，生成中间结果。这些中间结果也需要存储在显存中，以便后续计算使用。对于大语言模型而言，中间计算结果可能会占用较多的显存空间。
4. **内存管理策略**：某些深度学习框架在推理时采用了一种延迟释放显存的策略，即显存不会立即释放，而是保留一段时间以备后续使用。这种策略可以减少显存的分配和释放频率，提高推理效率，但也会导致显存一直占用的现象。

需要注意的是，显存的占用情况可能会受到硬件设备、深度学习框架和模型实现的影响。不同的环境和设置可能会导致显存占用的差异。如果显存占用过多导致资源不足或性能下降，可以考虑调整模型的批量大小、优化显存分配策略或使用更高性能的硬件设备来解决问题。