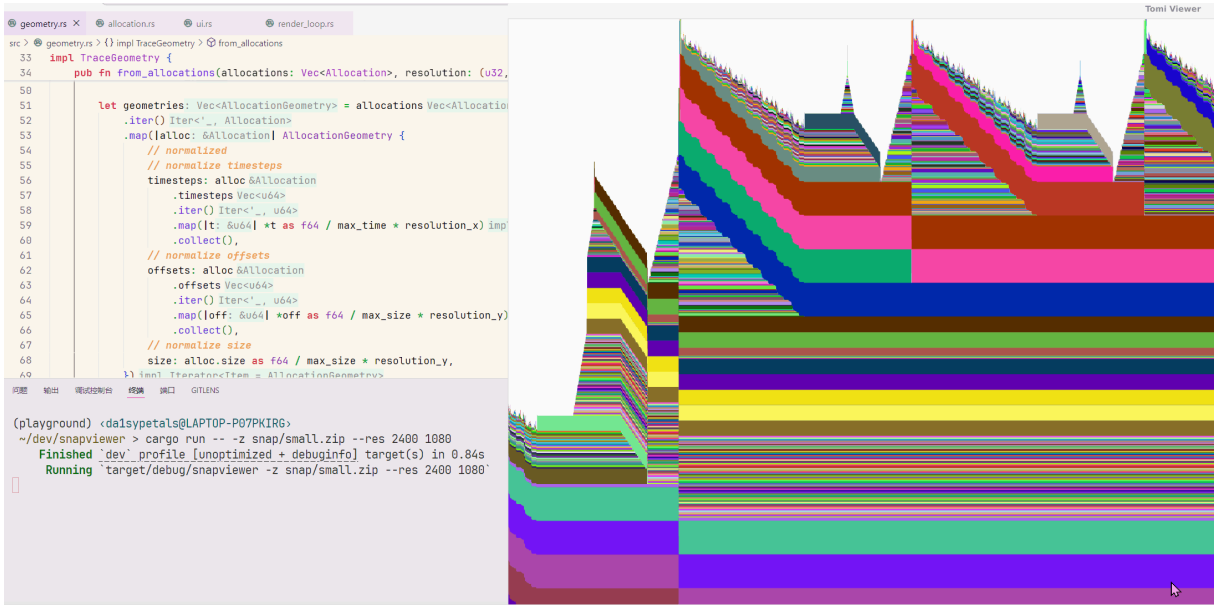
SnapViewer

- PyTorch 在训练模型的时候常常会 OOM，这时候就需要对显存进行优化。当一些简单的方法（降低 batchsize 等）以及行不通的时候，可能就需要对模型本身的显存轨迹进行分析。
- 这时候你会看到这个文档，他会教你如何记录 memory snapshot 并且在 这个网站 上进行可视化。
- 但是有一个很大的问题是：这个网站 太卡了。如果你的模型很小，snapshot 只有几个 MB，流畅度还算能看；如果你的模型比较大，snapshot 达到几十甚至几百 MB，那么这个网站就会变得非常卡，帧率最低可达每分钟两三帧。
- 我去看了这个网站的 js 代码，它主要做了这些事：
  1. 手动加载 python pickle 文件；
  2. 每一帧都重新将原数据解析为图形，然后再每一帧渲染到屏幕上。

  这个渲染逻辑是用 js 写的，因此性能嘛…
- 我在对一个几 B 参数量的模型进行 snapshot 的时候发现了这个问题。
  - 为什么需要自己优化，而不是用现成的 LLM 基础设施？长话短说，这个模型是 researcher 自己设计的，里面含有大量的和 LLM 完全不同的模块。现在好像大家默认深度学习

只剩下 LLM 了，以至于甚至有些 tech lead 都认为 LLM 的基础设施可以轻松接到很多其他模型上面…偏题了

- 我原本写了个简单的脚本用来解析 snapshot 里面的内容，尝试借此发现模型里面的显存分配问题；但是在我对着这个模型工作了一个月之后，我终于受不了了。于是有了这个项目：SnapViewer.
- TLDR：将 memory snapshot 的图形解析出来，用一个巨大的 triangle mesh 表示，然后复用渲染库对 mesh 的渲染能力进行渲染。这是一个上百 MB 的 snapshot，在我的集显上跑的还算流畅：



如果你也有需要，欢迎试用一下：)

欢迎围观 & star！ https://github.com/Da1sypetals/SnapViewer