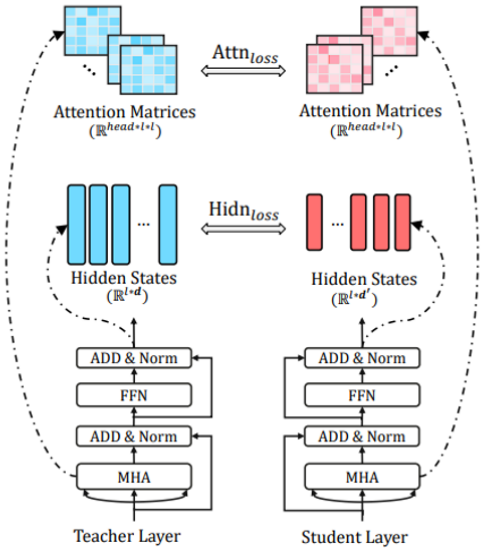
深信服期间实习工作报告-模型计算加速

吴立凡 2022.11.1-2023.1.26

对投机推理，量化剪枝，模型蒸馏压缩方向的调研

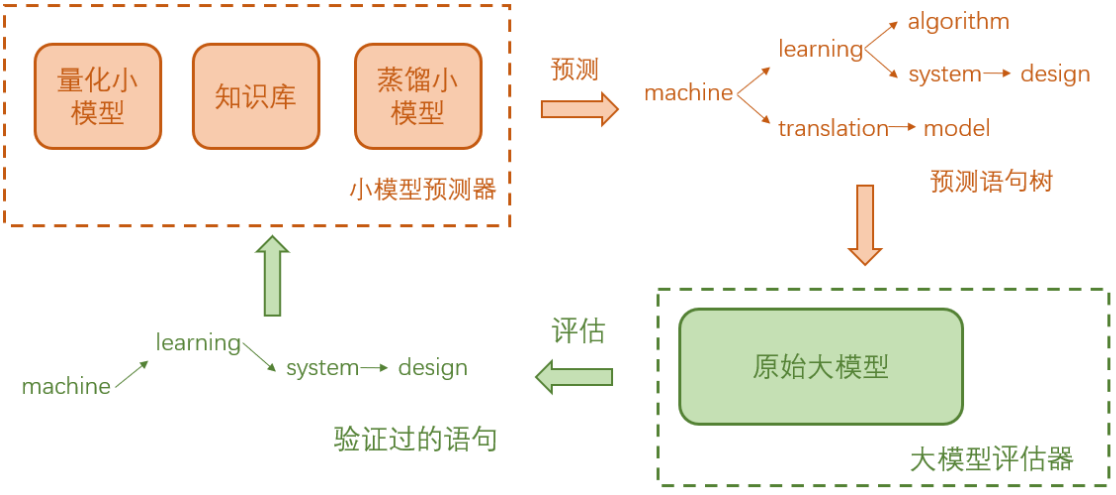
模型蒸馏

- Transformer 蒸馏是一种将教师模型的知识传递给学生模型的技术。
- 通过教师模型的attention-layer和hidden states layer的输出指导学生网络进行训练。
- 可以在不损失太多原始模型的性能的情况下将模型的参数量降低到原本的1/10，大幅减少计算量和模型模型对内存的消耗。



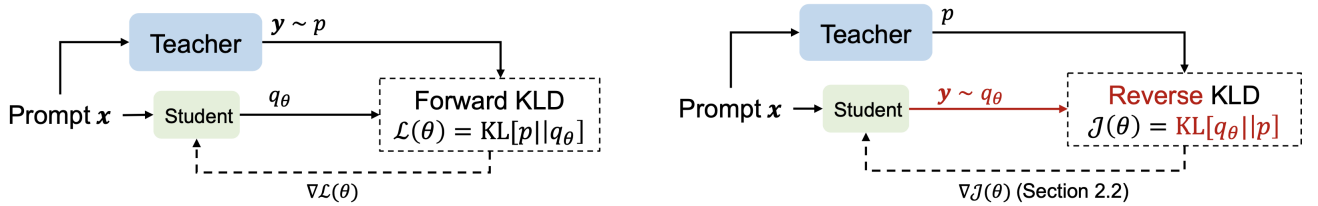
投机推理

使用一系列比原始模型小的多的**近似模型**，进行自回归串行采样，**预测**得到一个**Speculated token tree（预测语句树）**，然后使用大模型**评估**采样结果，重复以上过程



07

MiniLLM: Knowledge Distillation of Large Language Models



Minillm是一个白盒式的模型蒸馏框架，通过对模型最后一层输出做KL散度的最小化来进行模型蒸馏，会使用到原始模型最后一层的输出，具体细节可以看原文：MiniLLM: Knowledge Distillation of Large Language Models[https://arxiv.org/abs/2306.08543]

原始开源代码地址：https://github.com/microsoft/LMOps/tree/main/minillm

我在对这个代码进行实验的时候做出了一些修改，代码地址：

https://github.com/wusar/minillm

复现我之前做的工作

我完成了对该代码运行环境的配置，并对原始论文中的数据集完成了下载和数据集的预处理。如果对于数据集有疑问可以直接查看库中的原始文档，数据集全部放在data目录下，训练后的模型放在/data01/lifanwu/TinyLlama下

可以通过以下方式来运行我写的代码

首先使用oem用户登录至10.0.100.7服务器（该服务器需要使用easyconnect vpn连接）

进入后切换至root用户，随后启动lifanwu conda环境

```
sudo su
conda activate /home/oem/.conda/envs/lifanwu
```

```
   c2b75d7..58e21b5  main -> main
(lifanwu) root@oem:/home/oem/lifanwu/minillm# conda env list
# conda environments:
#
Firefly                  /data01/conda/envs/Firefly
txx                      /data01/conda/envs/txx
                         /home/oem/.conda/envs/h800
                      *  /home/oem/.conda/envs/lifanwu
base                     /home/oem/anaconda3
yijia                    /home/oem/anaconda3/envs/yijia
```

随后切换目录到工作目录，运行以下命令即可开始对于llama13b模型蒸馏到7b模型的训练

```
cd /home/oem/lifanwu/minillm
bash scripts/distill_llama2_13b_7b.sh .
```

scripts/distill_llama2_13b_7b.sh里面是整个训练的配置文件，在这个配置文件里面包含了模型路径，使用的GPU数量，训练的epoch数，学习率等等超参数，同时也可以修改模型的结构，比如将llama结构的模型修改为chatglm结构的模型

llama7b的模型蒸馏到TinuLlama1.1b的模型

与13b蒸馏到7b类似，运行以下命令即可开始对于llama7b模型蒸馏到1.1b模型的训练

```
cd /home/oem/lifanwu/minillm
bash scripts/distill_llama2_7b_1b.sh .
```

对llama13b蒸馏到7b模型时的运行结果

我在8章V800的机器上训练了大约一天不到的时间，训练了3个epoch。原始论文里训练了10个 epoch，在16台32Gv100的机器上进行的训练，训练结果的日志可以看 llama_distill_log.txt文件

7b到1b的运行结果在 llama_distill_log_7b_1b.txt文件中

在对chatglm模型进行模型蒸馏的时候遇到的一些问题

我后来还尝试对于chatglm模型进行了模型蒸馏，但是遇到了一些麻烦，我新创建了一个新的配置文件scripts/distill_chatglm_6b_1b.sh,在运行的时候遇到了以下的错误:

```
ValueError: Loading this model requires you to execute custom code contained in
the model repository on your local machine. Please set the option
`trust_remote_code=True` to permit loading of this model.

ValueError: Tokenizer class ChatGLMTokenizer does not exist or is not currently
imported.
```

这里需要修改原始代码中的from_pretrained函数，将trust_remote_code设置为True，但是这个函数在minillm代码中被多次调用，修改了所有的from_pretrained函数之后就可以成功导入Chatglm2模型。

但是随后我发现chatglm原始代码中存在一个bug：

```
RuntimeError: Subtraction, the `-` operator, with a bool tensor is not
supported. If you are trying to invert a mask, use the `~` or `logical_not()`
operator instead.
```

该bug是因为在pytorch2里面，bool类型的tensor不支持减法操作，需要将bool类型的tensor转换为int类型的tensor。

修复该bug的方式在这里：

https://huggingface.co/THUDM/chatglm2-6b/discussions/67#64c0df718e261225436fc783)%EF%BC%8C%5Bshibing624%5D(https://huggingface.co/shibing624

理论上来说修复这个bug之后就可以正常的像蒸馏llama模型一样蒸馏chatglm模型，但是我在改掉原本cahce里的代码后，重新运行又会覆盖掉我修改的代码，所以我就没有继续做下去。

其他的一些工作

使用裁判模型来进行指代词消解训练结果的评价（具体情况请询问唐博）