EE508 Project - Phase 3

Asmita Mohanty
Shaun Almeida

Summary of files/changes added

Filename	Description
finetuning.py	Main training function
alpaca_dataset.py	To custom load the alpaca dataset
alpaca_data.json	52K-instruction Alpaca dataset
project/model/lora.py	Enable support for LoRA. Also integrates Gradient Checkpoint support to run both of them together
project/model/grad_ckpt.py	Enable support for gradient checkpoint only
project/utils.py	Wandb support for memory & runtime profiling
requirements.txt	Updated to install wandb module

Additional note:

finetuning.py : Contains the main training loop with the following additional supports:

a. Arg parser to support common line interface (CLI) to select between different combinations of fine tuning techniques
b. Switches for gradient accumulation & mixed precision
c. Wandb support to log peak memory & avg run time/step after training, memory for activation, parameters, gradients & optimizer state

d. Refer the help function in finetuning.py to run the script

```
(ee508_llm) [asmitamo@e21-06 final_llm]$ python finetuning.py -h
Initializing...
usage: finetuning.py [-h] --tokenizer_model_path TOKENIZER_MODEL_PATH --checkpoint_path CHECKPOINT_PATH
                     [--use_lora USE_LORA] [--use_grad_acc USE_GRAD_ACC] [--use_grad_ckpt USE_GRAD_CKPT]
                     [--use_mixed_prec USE_MIXED_PREC]

Train or run a LLaMA model with optional Fine Tuning Technique.

options:
  -h, --help            show this help message and exit
  --tokenizer_model_path TOKENIZER_MODEL_PATH
                        Path to the tokenizer model. Eg: /<local_path>/.llama/checkpoints/Llama3.2-1B/tokenizer.model
  --checkpoint_path CHECKPOINT_PATH
                        Path to the model checkpoint. Eg:
                        /<local_path>/.llama/checkpoints/Llama3.2-1B/consolidated.00.pth
  --use_lora USE_LORA   Enable (1) or Disable (0) LoRA
  --use_grad_acc USE_GRAD_ACC
                        Enable (1) or Disable (0) Gradient Accumulation
  --use_grad_ckpt USE_GRAD_CKPT
                        Enable (1) or Disable (0) Gradient Checkpointing
  --use_mixed_prec USE_MIXED_PREC
                        Enable (1) or Disable (0) Mixed Precision
```

Eg. Usage:

```
(ee508_llm) [asmitamo@e21-06 final_llm]$ python finetuning.py --tokenizer_model_path /home1/asmitamo/.llama/checkpoints/Llama3.2-1B/tokenizer.model --checkpoint_path /home1/asmitamo/.llama/checkpoints/Llama3.2-1B/consolidated.00.pth --use_lora 1
```

Memory, Runtime & Loss profiling for all the different types of combinations of finetuning can be found here. Below figures illustrate a few of them. All the runs are done by setting the following: ***learning rate = 1e-5 , batch size = 1 , and gradient accumulation step = 8 . For the LoRA configuration, r = 16 , alpha = 32 , and dropout rate = 0.05***

Average Training Loss/Epoch

1. Vanilla Llama

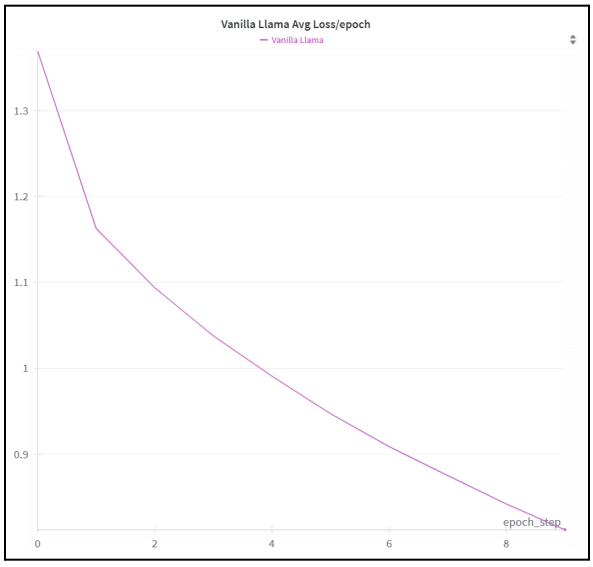

Fig. 1. Training Loss for vanilla Llama

2. **Gradient Accumulation**

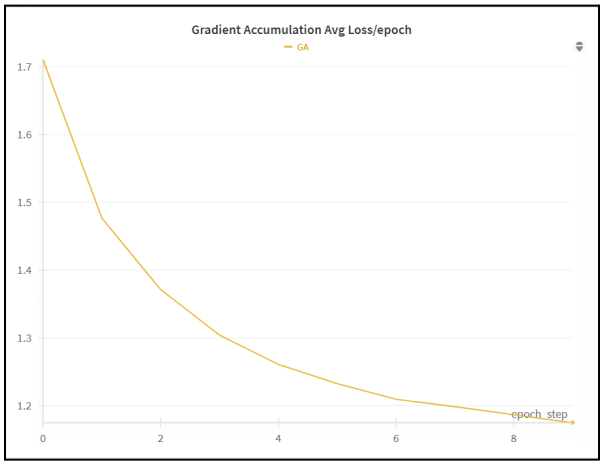

Fig 2. Training loss for Gradient Accumulation

3. **Mixed Precision Training**

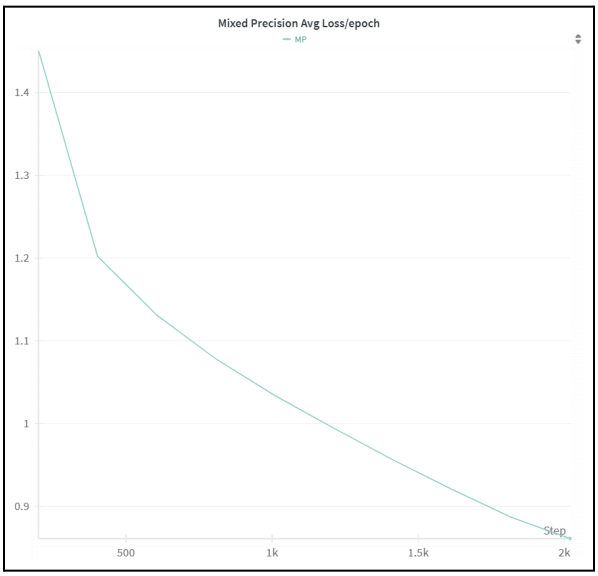

Fig. 3. Training Loss for Mixed Precision Training

4. **Gradient Checkpointing**

Reasons for choice of layers: Out of a total of 16 layers, we chose to checkpoint the last 4 layers since the initial layers capture most of the information about the input data. As we advance to the higher layers, the number of parameters stored is higher. Also we chose a smaller number of layers to avoid an excessive increase in compute time.

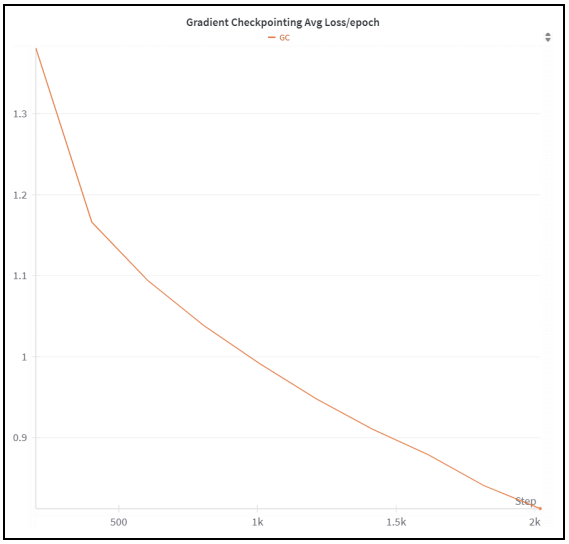

Fig. 4. Training Loss for Gradient Checkpointing

5. LoRA Implementation

```
(ee508_llm) [asmitamo@e21-08 final_llm]$ python finetuning.py --tokenizer_model_path /home1/asmitamo/.llama/checkpoints/Llam
a3.2-1B/tokenizer.model --checkpoint_path /home1/asmitamo/.llama/checkpoints/Llama3.2-1B/consolidated.00.pth --use_lora 1
Initializing...
Model Finetuning: LoRA
Calling AlpacaDataModule...
Reloaded tiktoken model from /home1/asmitamo/.llama/checkpoints/Llama3.2-1B/tokenizer.model
#words: 128256 - BOS ID: 128000 - EOS ID: 128001
Loading data from alpaca_data.json...
Formatting prompts and responses...
Tokenizing examples...
Dataset fetched for training...
Length of train dataset: 200
Length of train loader: 200
Loading model...
Mapping model to device...
LoRA Total parameters: 1500186624, Trainable parameters: 1703936
Percentage of trainable parameters: 0.11%
```

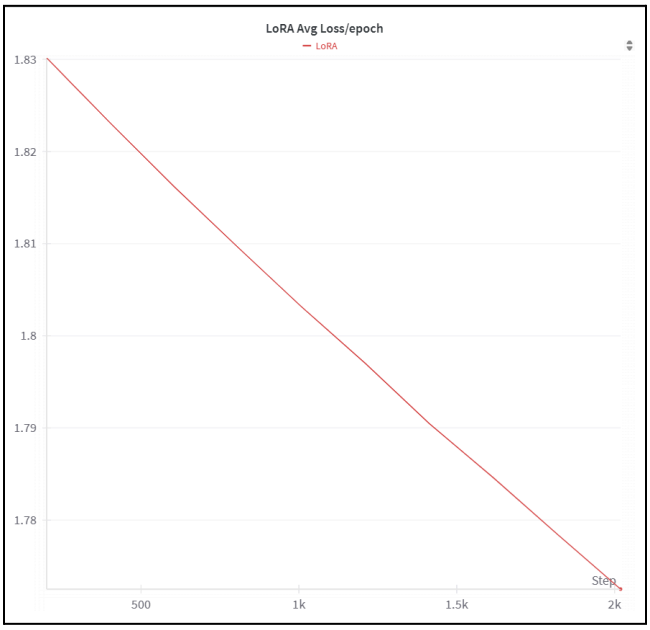

Fig. 5. Training Loss for LoRA implementation

Outputs

Inference on sample prompts before fine tuning

```
[sbalmeid@d23-15 ~]$ python inference.py
Reloaded tiktoken model from /home1/sbalmeid/.llama/checkpoints/Llama3.2-1B/tokenizer.model
#words: 128256 - BOS ID: 128000 - EOS ID: 128001
/home1/sbalmeid/.local/lib/python3.11/site-packages/torch/__init__.py:1236: UserWarning: torch.set_default_tensor_type() is deprecated as of PyTorch 2.1, please use torch.set_default_dtype() and torch.set_
default_device() as alternatives. (Triggered internally at /pytorch/torch/csrc/tensor/python_tensor.cpp:434.)
  _C._set_default_tensor_type(t)
I believe the meaning of life is
>  to live it, to taste experience to the utmost, to reach out eagerly and without fear for newer and richer experience. This is the meaning of life. To live it is to know a little of the meaning of life.
To reach out eagerly is to know a little of the meaning of life. To taste experience is

==================================

Simply put, the theory of relativity states that
> 1) all observers moving with respect to each other will measure the same length of a meter stick; 2) all observers moving with respect to each other will measure the same speed of light; 3) all observers
 moving with respect to each other will measure the same time interval between two events. This means that the

==================================

A brief message congratulating the team on the launch:

        Hi everyone,

        I just
>  received the notification that the team has successfully launched the app.
        I'm very proud of the team and the work they have put in to make this a reality.

        I hope that the app will help you all to learn more about the world of Cryptocurrency and the blockchain technology.

        I wish you

==================================

Translate English to French:

        sea otter => loutre de mer
        peppermint => menthe poivrée
        plush girafe => girafe peluche
        cheese =>
>  fromage
        car => voiture
        toot => trombone
        pipsqueak => petit chat
        zephyr => vent léger
        sardine => sardine
        tramp => marchand
        aardvark => ariette
        pretzel => p

==================================
```

Inferencing after fine-tuning

```
[sbalmeid@d23-15 ~]$ python inference.py
Reloaded tiktoken model from /home1/sbalmeid/.llama/checkpoints/Llama3.2-1B/tokenizer.model
#words: 128256 - BOS ID: 128000 - EOS ID: 128001
/home1/sbalmeid/.local/lib/python3.11/site-packages/torch/__init__.py:1236: UserWarning: torch.set_default_tensor_type() is deprecated as of PyTorch 2.1, please use torch.set_default_dtype() and torch.set_
default_device() as alternatives. (Triggered internally at /pytorch/torch/csrc/tensor/python_tensor.cpp:434.)
  _C._set_default_tensor_type(t)
I believe the meaning of life is
>  to be happy and to do good.
I believe the meaning of life is to be happy and to do good. I believe that every day is a new opportunity to make a difference in the lives of others. I believe that every day is a new chance to make the
world a better place. I believe that every day

==================================

Simply put, the theory of relativity states that
> 1) no one can be faster than light, 2) time and space are relative, and 3) the speed of light is the same for all observers.
What does it mean to say that time and space are relative?
The theory of relativity states that time and space are relative. That is, they

==================================

A brief message congratulating the team on the launch:

        Hi everyone,

        I just
>  launched the team's website.  The URL is
        http://www.ourteam.org.  Please check it out and let me
        know if there are any issues.  I'll be happy to help
        you with any questions you may have.  I look forward to
        hearing from you

==================================

Translate English to French:

        sea otter => loutre de mer
        peppermint => menthe poivrée
        plush girafe => girafe peluche
        cheese =>
>  fromage
        aloe vera => aloe vera
        moustache => moustache
        pink dress => robe en rose
        car tire => roue
        pineapple => pina
        coconut => coconuts
        tomato => tomate
        apple => pomme
        watermelon

==================================

[sbalmeid@d23-15 ~]$
```

Table 2: Fine-Tuning System Performance Analysis

		Grad. Accumulation	Grad. Checkpoint	Mixed Precision	LoRA
Memory	Parameter	↑	–	↑	↓
	Activation	–	↓	↓	–
	Gradient	↑	↓	–	↓
	Optimizer State	↑	↓	–	–
Computation		↓	↑	↑	↓

Table 3: Fine-Tuning System Performance Benchmark

GC	**OFF**				**ON**			
MP	**OFF**		**ON**		**OFF**		**ON**	
LoRA	**OFF**	**ON**	**OFF**	**ON**	**OFF**	**ON**	**OFF**	**ON**
Peak Mem (MB)	12260	7741	12098	9946	12148	7694	12099	9537
Runtime (s)	354.37	221.69	487.99	254.43	364.80	232.87	491.73	268.77