Parameter Offloading

Run Code

- Dependencies

  > Transformers(latest)
  >
  > Pytorch(latest, CUDA 12.1)
  >
  > datasets(latest , used for load dataset)

- Train

  ```
  ./train.sh
  ```

- Evaluate

  ```
  ./inference.sh
  ```

- use `--offloading` and `--fp16` to control whether to use offloading and fp16
- change `--model-config` from $\{debug, 1b, 3b, 7b, 13b, 30b, 65b\}$ to change model size

Configurations

Use **a small LlaMA Model（< 0.5 B）** to test,model configurations is as follows:

```
{
    "_name_or_path": "meta-llama/Llama-2-debug-hf",
    "architectures": [
      "LlamaForCausalLM"
    ],
    "attention_bias": false,
    "bos_token_id": 1,
    "eos_token_id": 2,
    "hidden_act": "silu",
    "hidden_size": 2048,
    "initializer_range": 0.02,
    "intermediate_size": 4096,
    "max_position_embeddings": 2048,
    "model_type": "llama",
    "num_attention_heads": 2,
    "num_hidden_layers": 20,
    "num_key_value_heads": 2,
    "pretraining_tp": 1,
    "rms_norm_eps": 1e-06,
    "rope_scaling": null,
    "rope_theta": 10000.0,
    "tie_word_embeddings": false,
    "torch_dtype": "float16",
    "transformers_version": "4.35.2",
    "use_cache": true,
    "vocab_size": 32000
```

```
  }
```

Train configuration:

```
{
    "batch_size": 1,
    "learning_rate": 1e-5,
    "max_length": 4096,
    "weight_decay": 1e-2,
    "clip_grad": 1.0
}
```

note : without checkpointing

Memory Consumption

	Inference	**Training**
w/o offload	2.07 GB	19.37 GB
w/ blocked offload	0.61 GB	7.48 GB
w/ overlapped offload	0.62 GB	7.63 GB

Time Consumption

	Inference	**Training**
w/o offload	21.21 ms	70.17 ms
w/ blocked offload	400.93 ms	4968.61 ms
w/ overlapped offload	95.53 ms	1960.10 ms

Hardware Environment

I use a single **RTX 4090** which has a memory capacity of **24G**

the largest Model it can train: **>3B(Memory consumption 20.3G), I guess slightly less than 4B**