4/6

Environmental Setting

- `finetuning/training_configs/few_shot/mathqa-8_fixed_mathqa_shots.yaml`
- `batch_size = 1, val_batch_size: 1`
- `Max_gen_len = 256`

SantaCoder

HF name: [bigcode/santacoder]
Autodl, 1 * RTX A5000(24GB), 80GB memory

- Success: `ddp_find_unused_parameters_false, torch_dtype=torch.float16`
    - GPU Memory Usage: 5/24GB

Need to set `do_sample` as `False`, otherwise `RuntimeError: probability tensor contains either `inf`, `nan` or element < 0`

- https://github.com/THUDM/ChatGLM-6B/issues/31
- I set `beam_size = 2` in config to disable `do_sample` implicitly. Might not be a good solution.

LLAMA/Alpaca-7B

HF name: [decapoda-research/llama-7b-hf]
HF name: [chainyo/alpaca-lora-7b]

Autodl, 1 * RTX A5000(24GB), 80GB memory

- Success: `ddp_find_unused_parameters_false, torch_dtype=torch.float16`
    - GPU Memory Usage: 17/24GB

LLAMA-13B

HF name: [decapoda-research/llama-13b-hf]

1. Autodl, 2 * RTX A5000(24GB), 160GB memory

- OOM: `ddp_find_unused_parameters_false, torch_dtype=torch.float16`
- OOM: `deepspeed_stage_2, torch_dtype=torch.float16`
- OOM: `deepspeed_stage_2_offload, torch_dtype=torch.float16`

- OOM: `deepspeed_stage_3_offload, torch_dtype=torch.float16`

2. Vesna 1*A6000(48GB), 128GB memory
- Success: `deepspeed_stage_3_offload, torch_dtype=torch.float16`

LLAMA-30B

HF name: [decapoda-research/llama-30b-hf]

1. Autodl, 2/3 * RTX A5000(24GB), 240GB memory
- OOM: `ddp_find_unused_parameters_false, torch_dtype=torch.float16`
- OOM: `deepspeed_stage_2, torch_dtype=torch.float16`
- OOM: `deepspeed_stage_2_offload, torch_dtype=torch.float16`
    - OOM: 2* RTX A5000(24GB), 240GB memory
- OOM: `deepspeed_stage_3_offload, torch_dtype=torch.float16`
    - OOM: 2* RTX A5000(24GB), 240GB memory
    - OOM: 3* RTX A5000(24GB), 240GB memory

2. Vesna 1*A6000(48GB), 128GB memory
- KILLED: `deepspeed_stage_3_offload, torch_dtype=torch.float16`

Multi-GPU with deepspeed will stuck, with little `free Mem (<5Gb)`