Optimizing Large Language Models for Resource-Constrained Environments

Sai Tejaswi Veeramreddy

Net ID: SV2564

May 2024

Table of Contents

Executive Summary

OBJECTIVE

Enhance the domain adaptation process in natural language processing by optimizing the fine-tuning of large language models for specific domains.

APPROACH

- LoRA(Low-Rank Adaptation) and 8-bit quantization
- Cosine learning rate decay with warmup

RESULTS

Achieved a perplexity score of ~7 while fine-tuning model (1.5B param model) on the medical grants dataset after 12 epochs. Qualitatively, the text generated is plausibly similar to the nature and style of data in the dataset used for training.

Optimization Techniques

The project seeks to integrate diverse optimization techniques like LoRA Optimization and gradient accumulation in a synergistic manner, ensuring compatibility and effectiveness, as not all optimizations naturally complement each other.

Limited Hardware Consideration

The project's primary focus lies in streamlining the domain adaptation process under constrained hardware resources. This emphasis on resource limitations presents a distinct challenge, requiring novel strategies to facilitate efficient domain adaptation.

Present Day Relevance

Amidst the rising demand for large language models, the imperative to optimize their performance on limited hardware resources becomes ever more critical for specific domains. This optimization enhances their efficacy and relevance in real-world applications.

Hardware and Platform

Number of cores

14

Memory

128GB

GPU type

RTX8000

CUDA Version

v11.1.74

CUDA is Nvidia's GPU-specific parallel computing framework. A GPU node is required to make use of this functionality.

The code was executed on a Jupyter Notebook with the specified configurations on the High-Performance Computing (HPC) platform.

Config Parameters

```python
# Global configurations
config = {
    "DATASET_URL": "https://the-eye.eu/public/AI/pile_v2/data",
    "DATASET_NAME": "NIH_ExPORTER_awarded_grant_text",
    "NUM_WORKERS": 8,
    "DATASET_SPLIT_RATIO": 0.9,
    "PADDING_STRATEGY": "max_length",
    "MAX_TOKENS": 512,
    "MIN_GENERATION": 512,
    "MODEL_NAME": "facebook/opt-125m",
    "TOKENIZED_NAME": "opt_2700m_512",
    "BATCH_SIZE": 64,
    "NUM_EPOCHS": 30,
    "LEARNING_RATE": 5e-4,
    "MIN_LEARNING_RATE": 5e-5,
    "EPSILON": 1e-8,
    "BETAS": (0.9,0.95),
    "GRADIENT_CLIP": 1.0,
    "WEIGHT_DECAY": 0.01,
    "DECAY_STYLE": "cosine", #not used currently
    "WARMUP_RATIO": 0.003,
    "SAMPLING_INTERVAL": 20,
    "CHECKPOINTING_INTERVAL": 100,
    "VALIDATION_INTERVAL": 500,
    "GRADIENT_ACCUMULATION_STEPS": 4, #TODO: need to bring this back

    "DYNAMIC_LR": True,
    "PEFT": False,
}

from peft import LoraConfig, PeftConfig, get_peft_model
lora_config = LoraConfig(
    r=16,
    lora_alpha=32,
    target_modules=["q_proj", "v_proj"],
    lora_dropout=0.2,
    bias="none",
    task_type="CAUSAL_LM"
)
```

NYU

Technical Challenges

Due To Model

Because GPT-2 is outdated and incompatible with the quantization library (bitsandbytes), faced challenges in implementing weight quantization. As a workaround, only applied quantization in experiments involving the OPT models, which have 5 billion parameters.

Due to Data

During the initial development phase, we faced a technical obstacle with the extended 2.5-hour tokenization process necessary for the complete dataset of Opinions in the Law Domain, which spans ~50 GB. Consequently, opted to transition to a smaller dataset (Medical Grants - 2 GB) to expedite the iteration process.

Slow Convergence

Encountered slow convergence when using cyclical learning rate; therefore, transitioned to the industry-standard cosine decaying learning rate with warmup to enhance training efficiency.

Gradient Explosions

Frequently encountered gradient explosions when employing quantization methods, necessitating careful management to stabilize training.

Approach

Gradient Accumulation

Utilizing gradient accumulation enables training with augmented effective batch sizes, potentially enhancing both the performance and stability of the model.

LoRA (Low Rank Adaptation):

Implemented as part of the Parameter-Efficient Fine Tuning (PEFT) approach, LoRA (Low Rank Adaptation) decreases the number of trainable parameters by training pairs of rank-decomposition matrices while keeping the original weights frozen. This significantly diminishes the storage demand for large language models tailored to particular tasks.

8-bit quantization

8-bit quantization involves encoding weights or numerical values with a precision limited to 8 bits, thereby reducing the computational complexity and memory requirements of the model.

Result

Dataset Sample


```
t': "ACF's Office of Refugee Resettlement (ORR) administers a variety of social service programs intended to connec
t newly resettled refugees with critical resources, help them become economically self-sufficient, and help them in
tegrate into American society. One such program is the Refugee Cash Assistance (RCA) program, which provides both f
inancial support and social services to newly resettled refugees. Refugee Cash Assistance is similar to TANF in tha
t both are cash assistance programs that provide services aimed at promoting self-sufficiency; however the content,
mode of delivery and rules surrounding these services vary significantly by state and locality. Some counties and s
tates have reportedly integrated the delivery of TANF and RCA in a purposeful way to better serve refugees. Howeve
r, there is little documented information on the extent to which refugees access benefits and services through TANF
and RCA, differences in refugee characteristics between the two programs, how outcomes compare for refugees served
under these two programs, whether integration of these programs holds promise for refugee self-sufficiency, and whe
ther data is available to answer these questions. The Understanding the Intersection Between TANF and Refugee Cash
Assistance Services project aims to improve understanding of how RCA and TANF serve refugee populations, how these
programs intersect, and how these programs may be related to refugee self-sufficiency and employment outcomes. In f
all 2014 ACF launched this descriptive study to document the similarities and differences between cash assistance a
nd associated social services offered under RCA and TANF across different selected jurisdictions. The study aims to
better understand the population of refugees served by TANF and RCA, and the major differences in programmatic serv
ices associated with these two programs. The study will also explore how states and localities have coordinated TAN
F and RCA programs to deliver social services to refugees and whether these approaches hold promise for long-term j
ob stability and economic self-sufficiency among refugees. This field study will provide a deeper understanding of
current social service delivery systems serving refugees and will help to identify gaps in existing knowledge and d
ata around these systems. By improving knowledge of these programs and participant experiences, ACF hopes to move t
oward better serving this population. The project is being conducted by Abt Associates and MEF Associates."}
```

Validation Logs

Epoch	Batch	Validation Loss	Text Generated
12	500	1.9846316874027252	The overall goal of this project is to investigate the cellular and molecular mechanisms of the inflammatory reactions that occur in the lungs that are induced by the infection with S. pneumoniae. These responses may serve as the host-defense mechanism against the development of pneumonia. The Specific Aims are: To investigate the inflammatory responses of lung capillary endothelial cells (ECs) in human subjects in vivo and relate these responses to the development of pneumonia. We will assess the response of lung EC

Training Logs

3.95s/it => no gradient accumulation, no gradient checkpointing, 32 batch size, sgd, no quantization, no autocast, no scheduler

1.80s/it => gradient accumulation with 8 steps, gradient checkpointing, 128 batch size, adamw, 8-bit quantization, no scheduler

Conclusion

Although the initial goal of fine-tuning a moderately large language model (7B) on the Opinions in the Law Domain dataset was not achieved, our findings suggest that achieving it does not necessitate any new technical concepts beyond distributed training to expedite the process.

GitHub Repository

https://github.com/vsaitejaswie/HPML_2024

Thank You!