The Solution for NeurIPS LLM Effciency Finetuing Challenge

Team: Percent_bfd

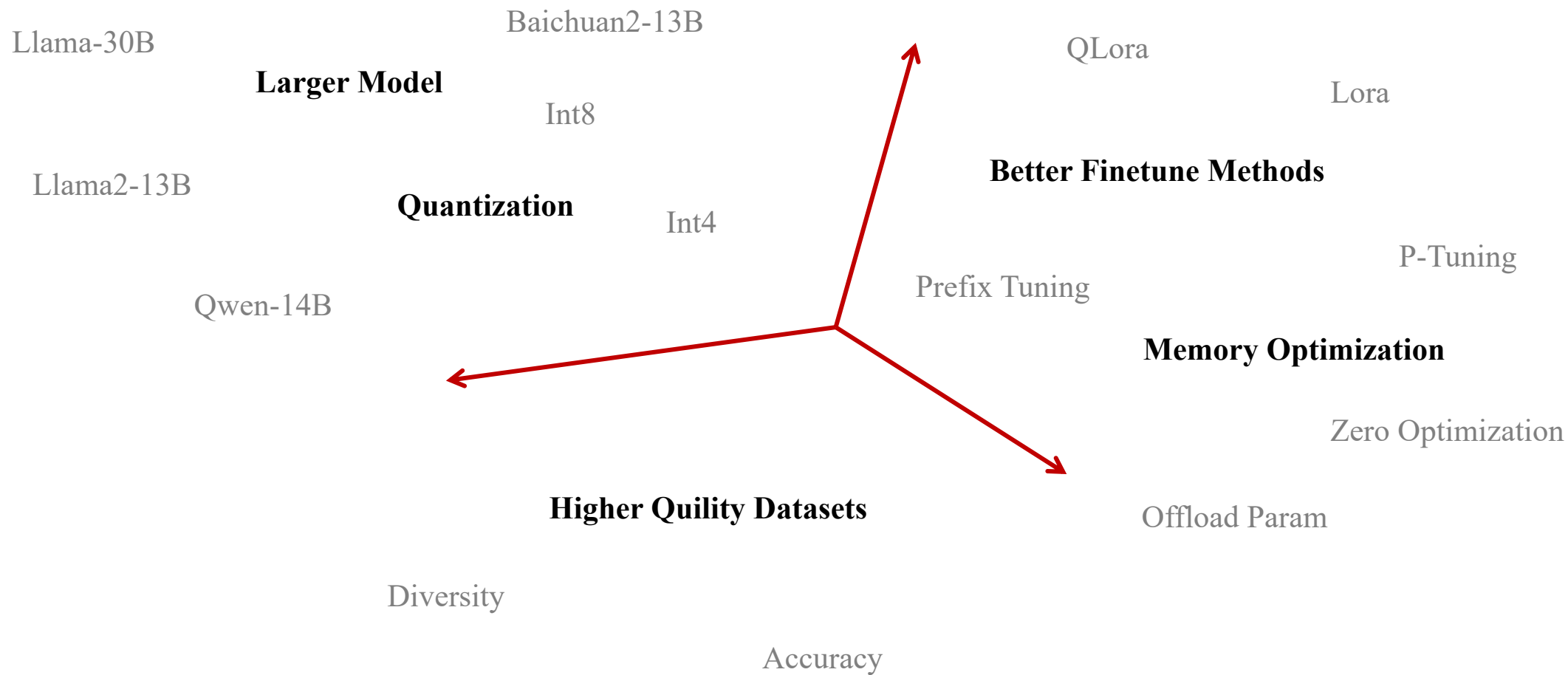Members: Ao Liu, Zizhen Huang, Hao Lu, Qin Hu, Qing Zhang

Project Page: https://github.com/Percent-BFD/neurips_submission

PERCENT 百分点

Ideas

Rules： 1 LLM + 1 GPU + 1 DAY

Llama-30B

Baichuan2-13B

QLora

Larger Model

Int8

Lora

Llama2-13B

Better Finetune Methods

Quantization

Int4

P-Tuning

Qwen-14B

Prefix Tuning

Memory Optimization

Zero Optimization

Higher Quility Datasets

Offload Param

Diversity

Accuracy

Model Selection

Here we finetuned the following base models on our datasets using LoRA.

Performance of Finetuned Model
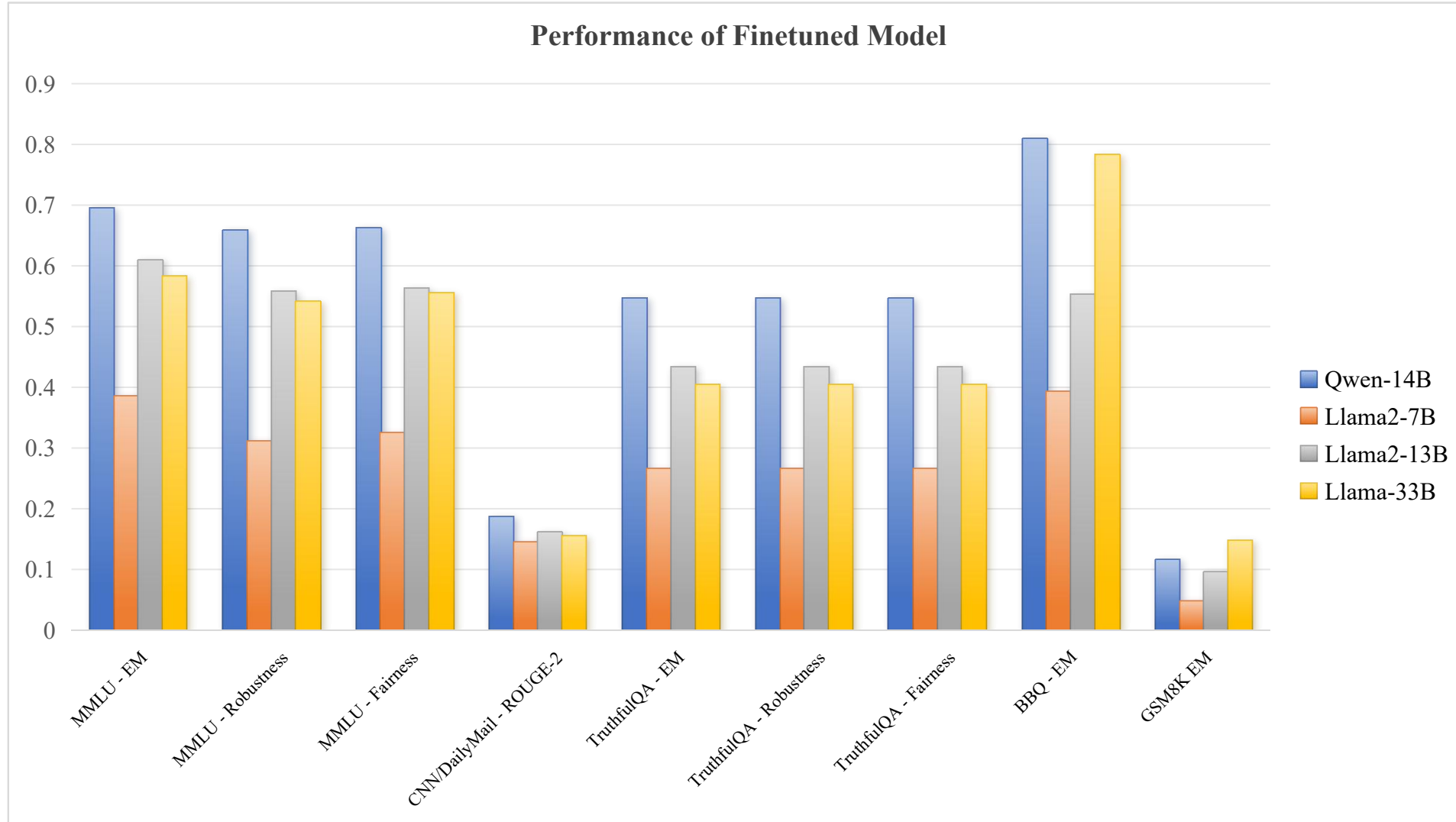
Model Selection

Models finetuned using LoRA on the same single GPU with batch size= 2, gradient accumulations steps = 4.

Traning Speed of LoRA FineTuning for Base Models

	Average Seconds/Each Training Sample
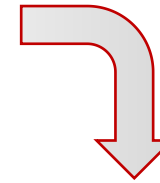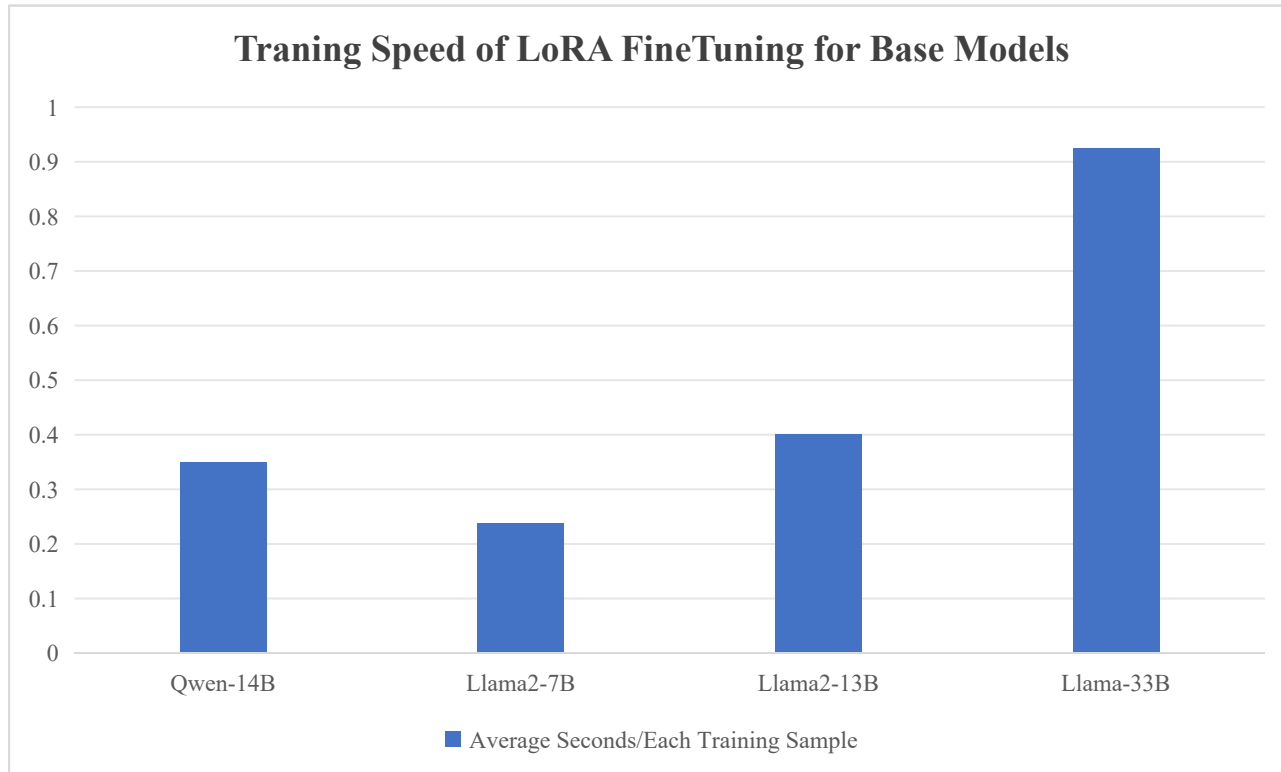
We finally chose Qwen-14B for the following reasons:

1. Outstanding performance.
2. Second fastest speed.

Datasets Selection

(Based on Qwen-14b)

MMLU (high school knowledge, moral)

bigbench(logical, safety)

lima

databricks-dolly-15k

oasst1-en

alpaca-gpt4

CodeAlpaca-20k

competition-math

+ MetaMathQA

+summarize_from_feedbac
k_tldr_3_filtered

- MetaMathQA

- summarize_from_
  feedback_tldr_3_filtered

+ GPT4-LLM-Cleaned

- GPT4-LLM-Cleaned

- alpaca-gpt4

- CodeAlpaca-20k

- competition-math

+ cnn_dailymail

+ truthful_qa

+ gsm8k

+summarize_from_feedbac
k_tldr_3_filtered

V1

V2

V3

V4

V5

Get a baseline

No significant effect in math & sum

Too much data requires 30h training

Overall score improved

Remove all llm generated data

CNN, truthful_qa improved

MMLU, BBQ, CNN

improved

Datasets Selection

The final dataset composition:

- MMLU

    ----all high school subjects & philosophy

    ----moral_disputes & moral_scenarios

- BIG-bench

    ----analytic_entailment, causal_judgment, empirical_judgments,

       known_unknowns, logical_deduction

    ----emoji_movie, strange_stories

    ----snarks, dark_humor_detection

- CNN_dailymail

- TruthfulQA

- GSM8K

- Lima

- Databricks-dolly-15k

- Oasst1-en

- Summarize_from_feedback_tldr_3_filtered

Conclusion:

- random 10% MMLU subsample

- random 10% CNN_dailymail subsample

- 10 copies of lima

- 67% random subset of the full dataset

- final training set size 24K

Final Solution

- Base model: Using Qwen-14B as the base model[1]

- Curated open-source training dataset: Employ a carefully selected open-source training dataset

- Fine-tuning with LoRA: Apply the LoRA method for instruction-fine-tuning[2][3]

[1] https://huggingface.co/Qwen/Qwen-14B
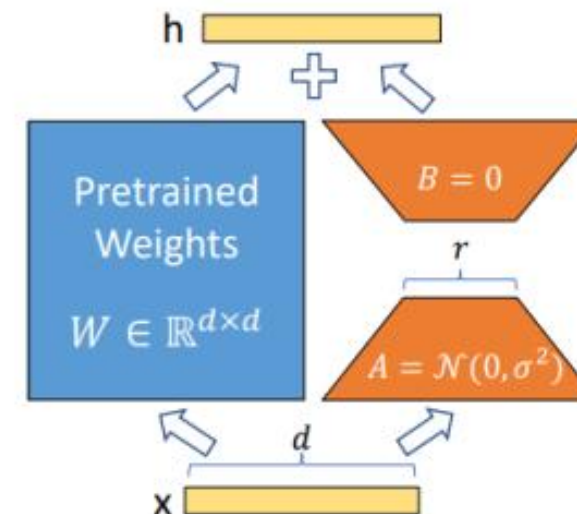
[2] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen. Lora: Low-rank adaptation of large language models. arXiv preprint arXiv:2106.09685, 2021.

[3] https://github.com/hiyouga/LLaMA-Factory

Training Setting

Hyperparameters:

- --lora_rank 8 \
- --lora_alpha 32\
- --lora_dropout 0.1\
- --lora_target "c_attn","c_proj" \
- --per_device_train_batch_size 2 \
- --gradient_accumulation_steps 4 \
- --lr_scheduler_type cosine \
- --learning_rate 5e-5 \
- --num_train_epochs 1.0 \
- --bf16

SFT template:

"A chat between a curious user and an artificial intelligence assistant. "

"The assistant gives helpful, detailed, and polite answers to the user's questions."

Human: {query}

Assistant: {answer}

w/wo template, the score difference is less than 0.01.

Acknowledgments

- NeuIPS

- Percent Technology

2+2+2

2 Products: Data Science Foundation Platform, Data Intelligent Applications

2 Business: Smart Governance, Digital Industry

2 Markets: Domestic & International markets

500

Technical Strength

Over 500 intellectual property rights, including patents and copyrights.

Contributed to the development of nearly 20 national and international standards in big data and AI.

Percent Technology
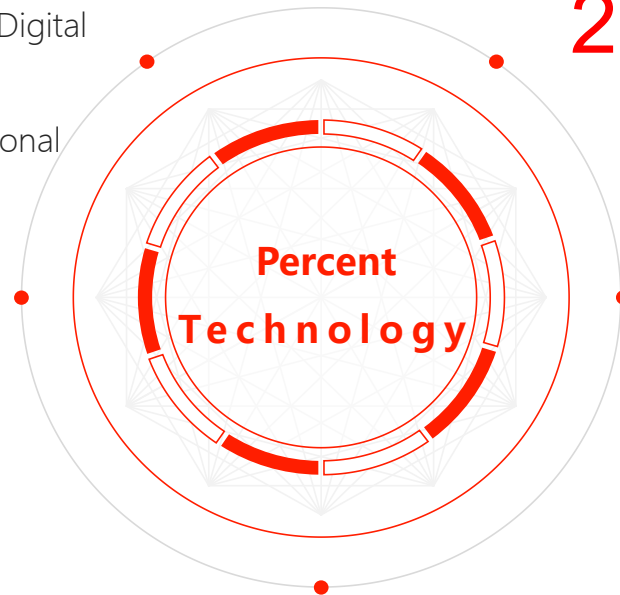
23,000+

Industry Applications

Commercial Projects

The business covers multiple countries and regions globally, involving various fields such as digital cities, emergency response, public security, statistics, ecological environment, retail, and media.

18

Branch Distribution

Beijing, Shanghai, Shenyang, Shenzhen, Guangzhou, Wuhan, Jinan, Hangzhou, Hong Kong, etc.