The complete guide for public models selection + Exploring Methods for fine tuning

Muhammad Umair Imran

[umairimran627@gmail.com](mailto:umairimran627@gmail.com)

NATIONAL UNIVERSITY OF COMPUTER AND EMERGING SCIENCES

ABSTRACT

This report aims to provide a clear guide lines for selection of publically available open source models and it also provides methods that can be applied for fine tuning depending on various problem cases and resources and performance .

Table Of Contents:

Background

Literature Review

Open source models public.

Fine tuning methods

Resources comparison

Performance comparison

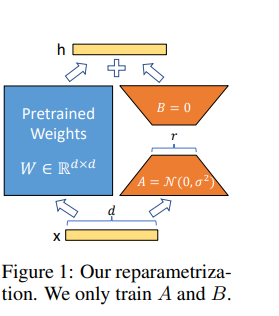
Applications wise distribution

COnslusion

Fine tuning methods

PEFT: Parameter efficient fine-tuning

LORA: Low Rank Adaptation of Language Models



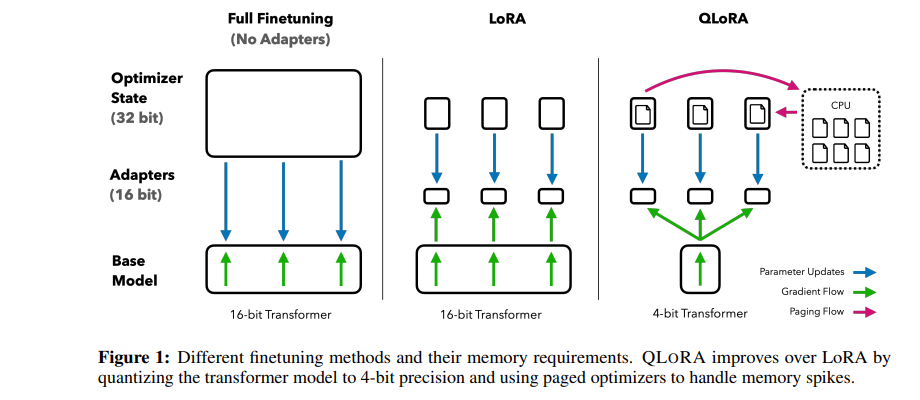
Low-Rank Adaptation, or LoRA, which freezes the pretrained model weights and injects trainable rank decomposition matrices into each layer of the Transformer architecture, greatly reducing the number of trainable parameters for downstream tasks

Lora uses the convereion of high rank matrices to low rank matrice then combine them and add them in original matrices to recover the original but fine tuned state of model it has saved much time of training

E. 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," *Microsoft Corporation*, [Online]. Available: <https://arxiv.org/abs/2106.09685v2>. [Accessed: 7-Jan-2025].

Qlora:

Quanzied model to save the time and resources if a model with 16 bit llama 65 b parameter would need 780 gb or gpu so makes too expensive for that first the model is quantize decreasing it precision to 4 bit or 8 bit then lora is applied on it



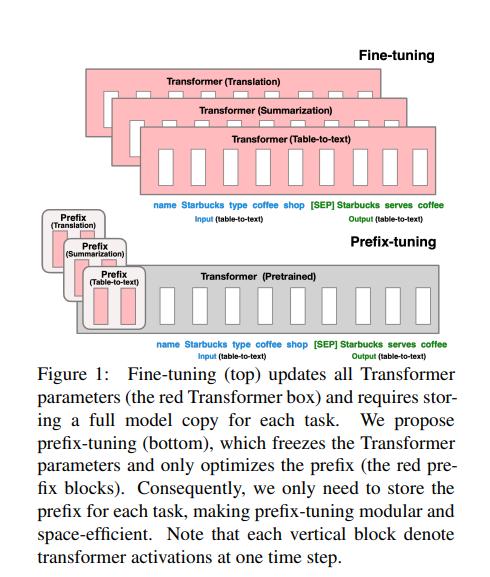
T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer, "QLoRA: Efficient Finetuning of Quantized LLMs," *arXiv*, May 2023. [Online]. Available: <https://arxiv.org/pdf/2305.14314>.

Prefix Fine tuning:

In this method we don’t change the wiehgts of model just we add some prefix during fine tuning like if we have to say

What is the president of Pakistan ? we simply add a prefix [history - related] to making it time saving and computation resoruces are saved y doing this method.

Now for inference we don’t need to add prefix it has fine tuned already .



X. L. Li and P. Liang, "Prefix-Tuning: Optimizing Continuous Prompts for Generation," *arXiv*, 2021. [Online]. Available: <https://arxiv.org/pdf/2101.00190>. [Accessed: 07-Jan-2025].

<https://www.ibm.com/think/topics/parameter-efficient-fine-tuning#:~:text=Staff%20writer-,What%20is%20parameter%2Defficient%20fine%2Dtuning%20(PEFT)%3F,specific%20tasks%20or%20data%20sets>.

Adapters:

Instead of full fine tune we just add an additional adapter layers between forwared feed layers . Full fine tuning takes a lot of resources and all parameters are tuned but in adapters case task specific layers is added to the model making it flexible for many use cases ,

peft

lora fine tuning

qlora : quantized lora

prompt tuning

: soft prompt prompt added to the input that is embedding format

adapters: a layer that inserts some where in the model helps to train in much smaller time

super vised fine tuning

instruction fine tuning

peft methods :

selective layer : freezes certain layers

reparameterizaion : lora method

additive method: adapter

soft prompting

RLHF: Reenforcement learning with Human Feedback

supervised fine tuning

full fine tuning

transfer learning : good on small data

task specific fine tuning : catastropic forgetting problem and specific for one task

multi task learning : need lot of data

sequential fine tuning: from general ot medical then to cardio

rag: external up to date method

few shot learning

reward modeling

proximal policy optimization

comparitive ranking

preferene learning