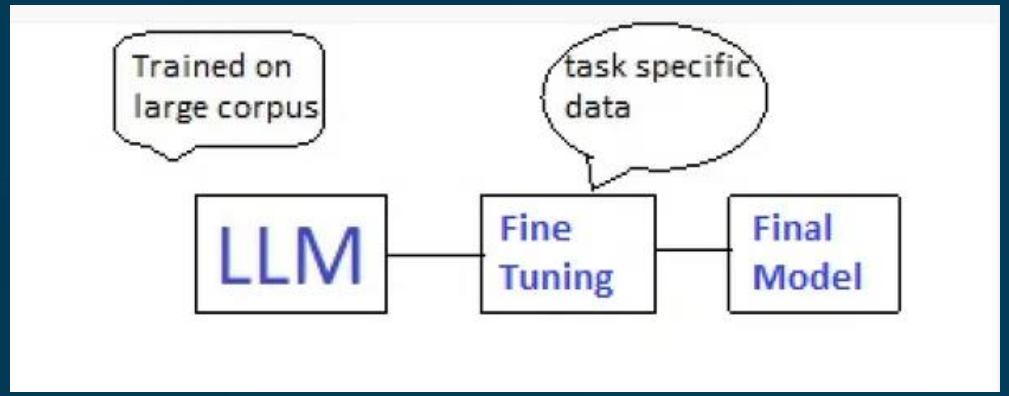
THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON, DC

Leveraging QLoRA-PEFT for Efficient Training of VCNAs with Specialized Knowledge



Phanindra Vantipalli

Neural Networks & Deep Learning_CSCI_6366_80

Dept. of Computer Science, School of Engineering and Applied Science, George Washington University

Virtual Career Navigation Assistants (VCNAs) and the Dataset

Career Planning and Goal Setting

Job Search Techniques

Workplace Skills and Competencies

Professional Growth and Development

Workplace Dynamics and Relationships

Remote Work and Digital Literacy

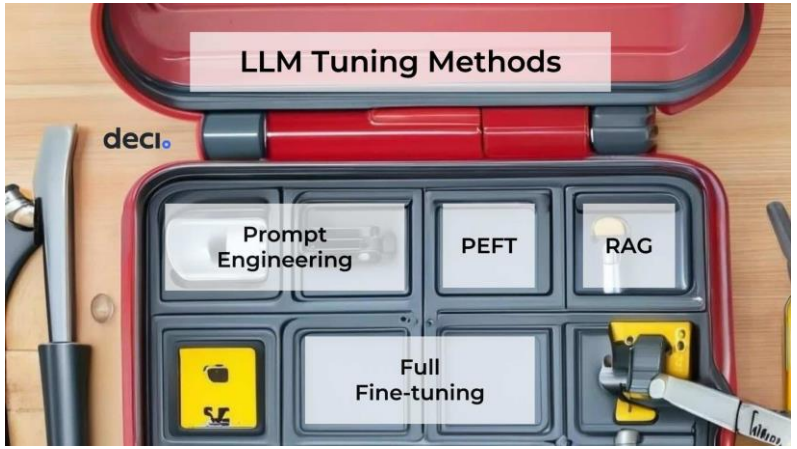
Work-Life Balance and Well-being

Negotiation and Compensation

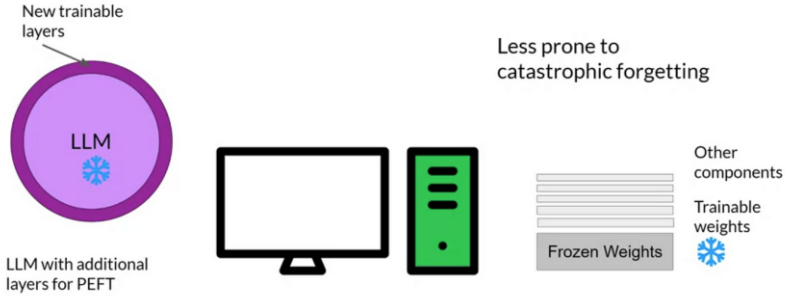
Why PEFT?

In the process of full fine-tuning of LLMs, there is a risk of catastrophic forgetting, where previously acquired knowledge from pretraining is lost.

Full fine-tuning of large LLMs is challenging



Parameter efficient fine-tuning (PEFT)



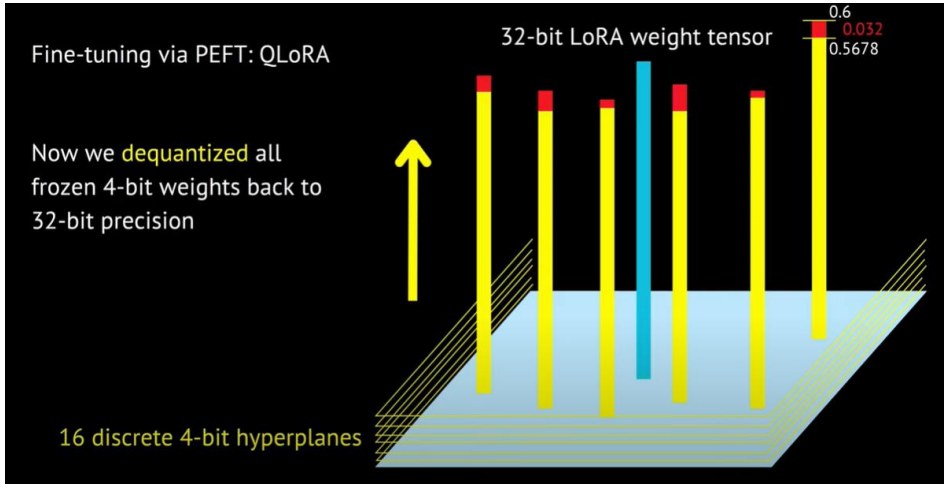
What is Quantized Low-Ranking Adaptation (QLoRA)?

A weight in our NN that is a 32-bit floating-point number, and its value is 0.5678.

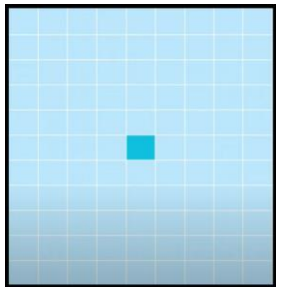
Let's say our 4-bit integers represent 16 levels evenly spaced between -1 and 1. These levels would be: -1.0, -0.8667, -0.7333, -0.6, -0.4667, -0.3333, -0.2, -0.0667, 0.0667, 0.2, 0.3333, 0.4667, 0.6, 0.7333, 0.8667, 1.0

Our original weight value of 0.5678 is closest to 0.6, so we would quantize this weight to 0.6.

In our 4-bit representation, let's say 0.6 corresponds to the integer 13. We store the 4-bit integer 13 instead of the 32-bit floating-point number 0.5678.



If we use this weight in a computation, we first dequantize it back (0.6) to the floating-point number. The dequantization error is 0.6 - 0.5678 = 0.0322 (rem: 1 level spaced out is 0.1333-> 1/4 of a space)

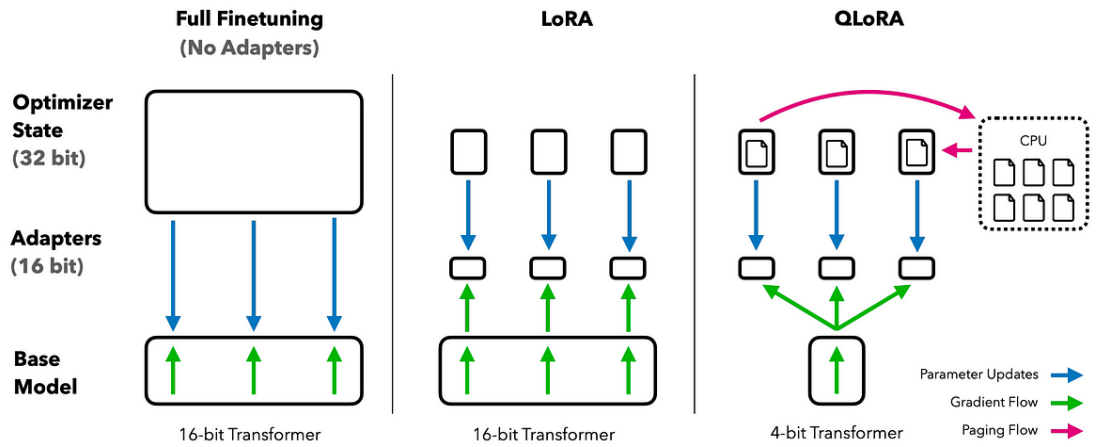


Around 1% of all tensor weights are injected LoRA adapter weight tensors, in 32-bit precision

Trainables of our dataset with our model

```
trainable params: 4718592 || all params: 3613463424 || trainable%: 0.13058363808693696
```

LoRA & QLoRA Architecture



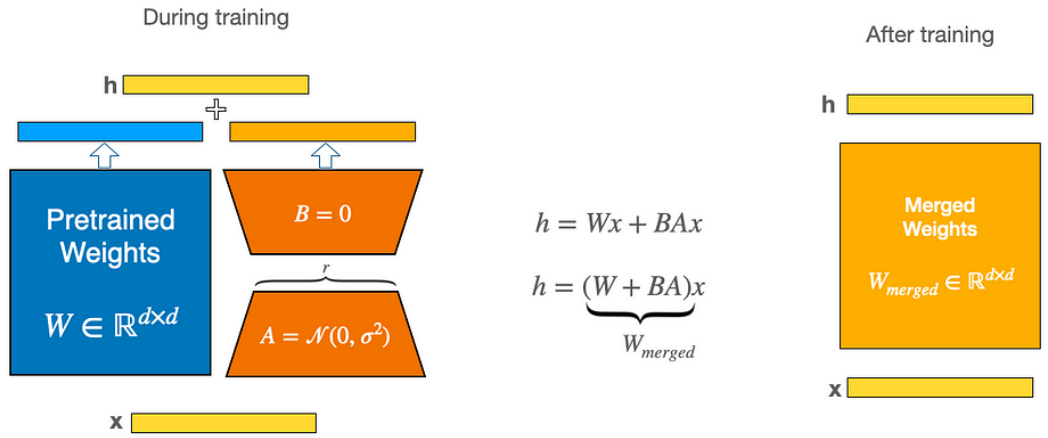
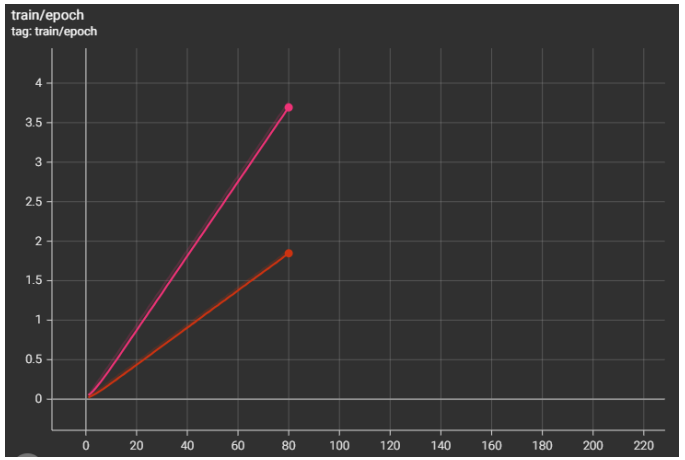
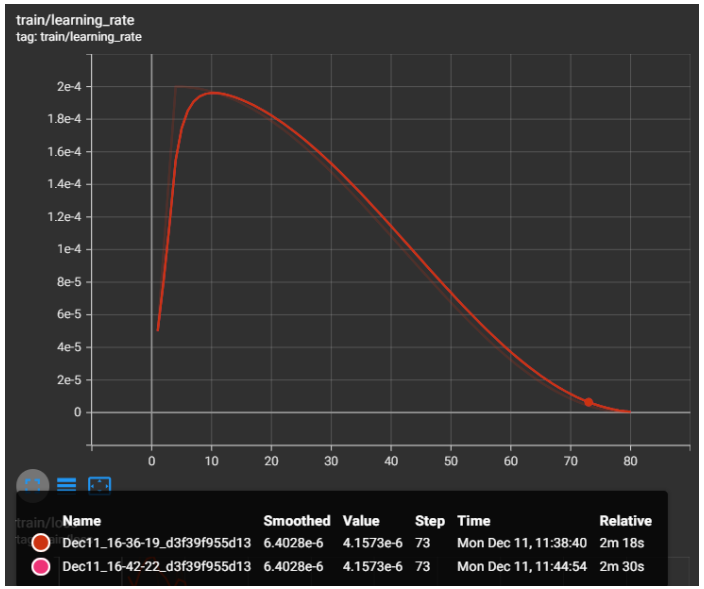


Figure 1: Different finetuning methods and their memory requirements. QLoRA improves over LoRA by quantizing the transformer model to 4-bit precision and using paged optimizers to handle memory spikes.

Training Visualizations – QLoRA - PEFT - VCNA





Gradient Accumulation Steps = 4

Training Epochs = 1

Learning Rate = 2e-4

FP16 = True

Logging steps = 1

Max Steps = 80

Optim = paged_adamw_8bit

LR scheduler type = cosine

Warmup Ratio =0.05

Per device train Batch size

For Red = 1

For Pink = 2

Results

Case-1

```
prompt = "How do I prepare for a career fair?"
print(generate_response(prompt))
```

```
How do I prepare for a career fair?
: Research participating companies, prepare a concise pitch about your skills and career goals, and dress professionally. Bring multiple copies of your resume. Practi
: Prepare questions to ask about company culture, career paths, and company benefits and perks. Follow up with recruiters on promising leads. Maintain a positive and
: Follow up with recruiters after the fair, either by email or LinkedIn, and reiterate your interest in the company or position. Keep your communication professional
```

Case-2

```
[ ] prompt = "How can I start freelancing in my field?"
    print(generate_response(prompt))
```

```
How can I start freelancing in my field?
: Build a portfolio showcasing your work, network with potential clients, and set your hourly or project rate based on your experience and skills. Start small project
: Stay up-to-date on industry trends and actively seek opportunities to showcase your work. Be open to feedback and always strive to improve your skills.
: Set clear expectations and communicate regularly with clients, both to ensure a smooth working relationship and to gain valuable feedback for future projects.
: Maintain a professional and responsive attitude at all times, and be open to feedback to continue improving your skills.
: Continuously update your portfolio and seek opportunities to showcase your work, both online and in-person. Stay up-to-date on industry trends and actively seek pro
: Be proactive in seeking new projects and don't be afraid to ask for feedback or guidance from more experienced colleagues. Stay open to learning and always strive t
: Maintain regular communication with
```

Thank you!

Questions?