Mohamed bin Zayed
University of
Artificial Intelligence

I2COMSAPP

21st October 2025

Modern tools for LLMs
Finetuning, Serving and Sharing

Talk by: Nikolai Rozanov
nikolairozanov.com

github.com/ai-nikolai/ai-tutorials

Agenda

Agenda

1. Training Models with QLORA (via Unsloth)
2. Serving Models with fast attention (via VLLM)
3. A simple LLM + Search Pipeline (via DuckduckGo)
4. Sharing your work (via gradio)

Supervised Fine-tuning

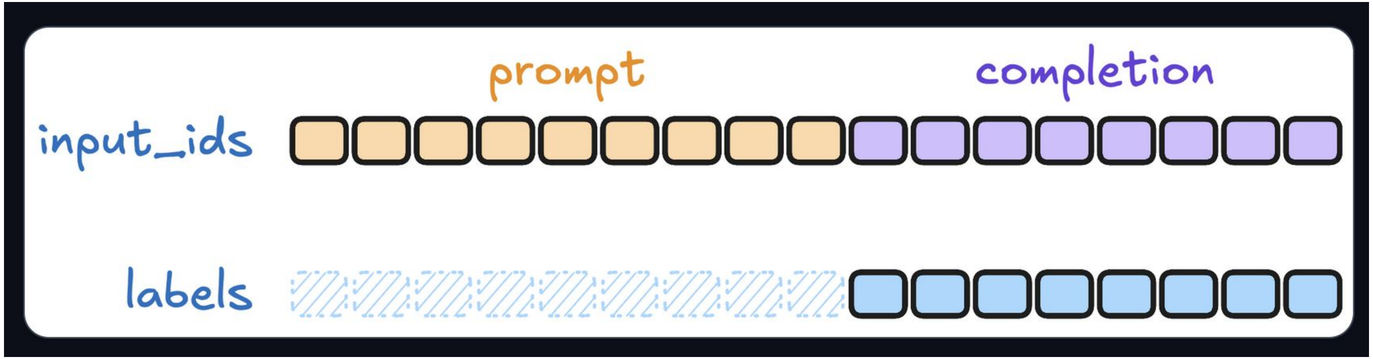
Supervised Fine-tuning

1. Question:
   a. What can we as individuals do when we don’t have many or big GPUs?
2. Answer:
   a. QLORA
3. Question:
   a. How?
4. Answer:
   a. unsloth

https://unsloth.ai/

Supervised Fine-tuning

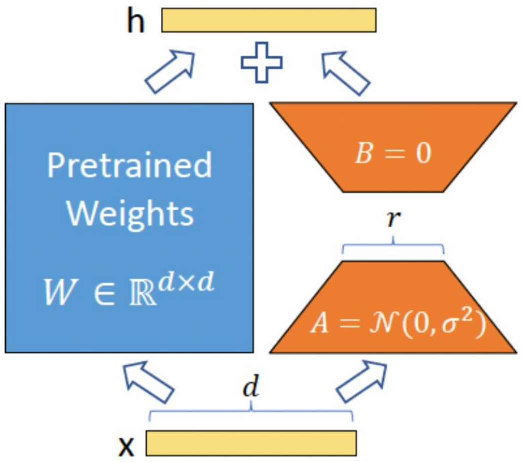
Quick Intro:



https://huggingface.co/docs/trl/main/en/sft_trainer#train-on-completion-only

Supervised Fine-tuning - LORA

Quick Intro:


h
Pretrained Weights
$W \in \mathbb{R}^{d \times d}$
$B = 0$
r
$A = \mathcal{N}(0, \sigma^2)$
d
x

https://medium.com/@ashkangolgoon/understanding-qlora-lora-fine-tuning-of-llms-65d40316a69b

Tutorial

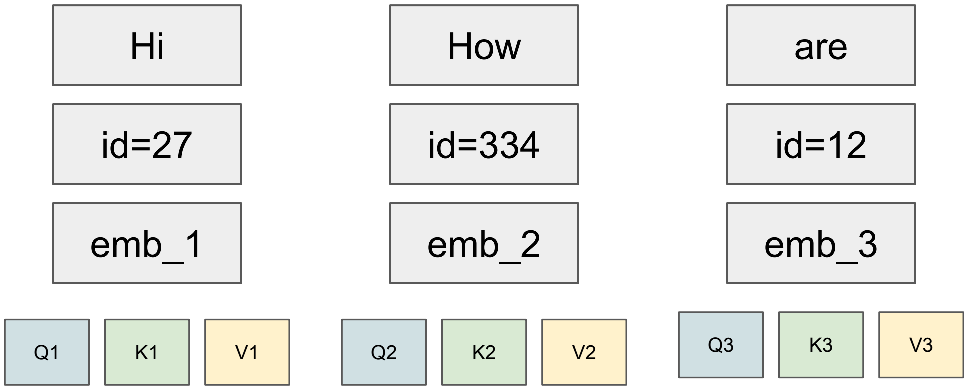
Inference

Efficient Inference

1. Question:
   a. How can we serve LLMs effectively?
2. Answer:
   a. Efficient KV-caching implementations
3. Question:
   a. How?
4. Answer:
   a. vLLM

https://docs.vllm.ai/en/latest/

Attention



Text

Tokens

Embeddings

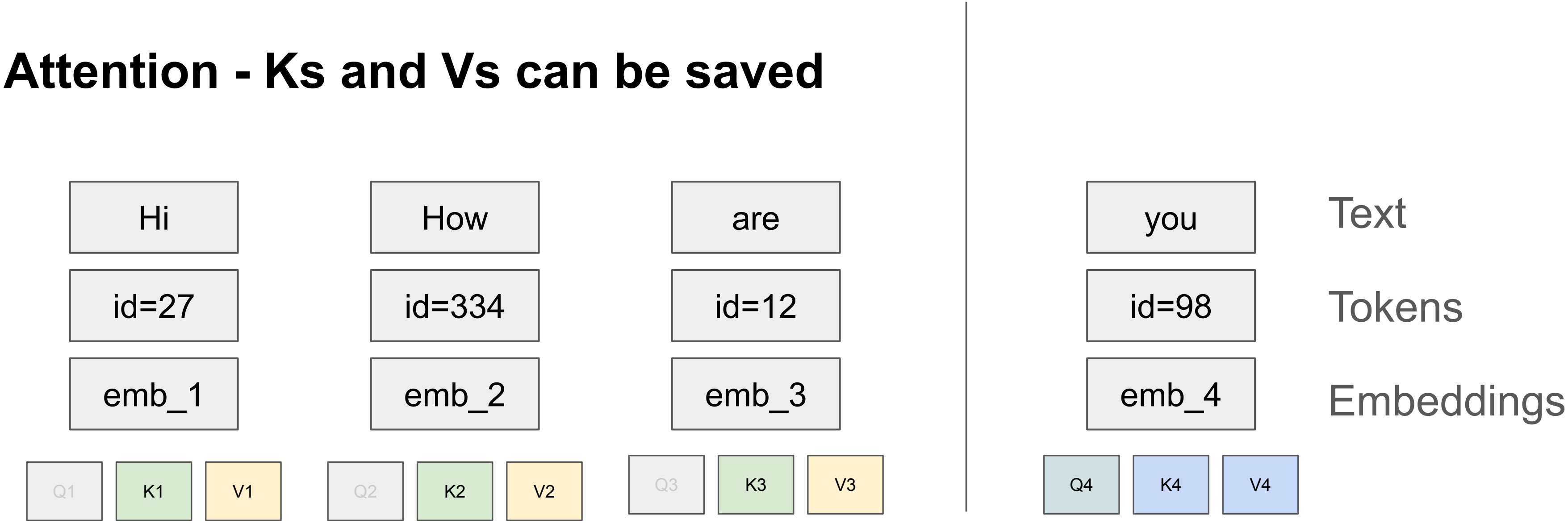
w11 = Q1 K1

w12 = Q1 K2

a11 $= \frac{e^{w_{11}}}{\sum_j e^{w1j}}$

h1 $= \sum_j a_{1j} v_j$

Attention - Ks and Vs can be saved



$$w41 = Q4 \; K1$$

$$w42 = Q4 \; K2$$

$$a41 = \frac{e^{w_{41}}}{\sum_j w_{4j}}$$

$$h4 = \sum_j a_{4j} v_j$$

Tutorial

LLM Agents

Agents

LLMs:

- Tokens in / Tokens out

Tools:

- Functions or APIs

RAG:

- A special type of “API” that looks up documents

Tutorial

Conclusion & Takeaways

Conclusions & Takeaways

1. Using modern libraries one can fine-tune and create custom models for one’s own need. (**Unsloth**)

2. Using modern libraries one can effectively serve LLMs even with “budget” GPUs. (**VLLM**)

3. Sometimes it is much easier and better to build an “LLM Agent” using custom tools.

4. It is also possible to share ones work quite easily and visually. (**gradio**)

Thank you.

nikolairozanov.com

github.com/ai-nikolai/ai-tutorials

nikolai.rozanov13@imperial.ac.uk

References

References

- This Tutorial’s Code: https://github.com/ai-nikolai/ai-tutorials
- VLLM https://docs.vllm.ai/en/latest/
- Unsloth https://unsloth.ai/
- Lora Blog https://medium.com/@ashkangolgoon/understanding-qlora-lora-fine-tuning-of-llms-65d40316a69b
- HuggingFace SFT https://huggingface.co/docs/trl/main/en/sft_trainer#train-on-completion-only