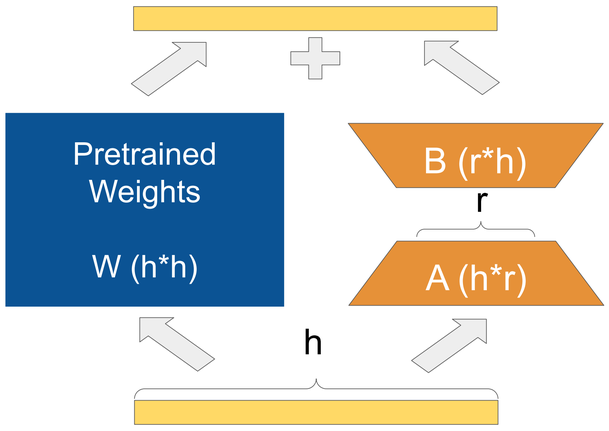
Dynamic LoRA Serving System and Applications to Offline Context Learning

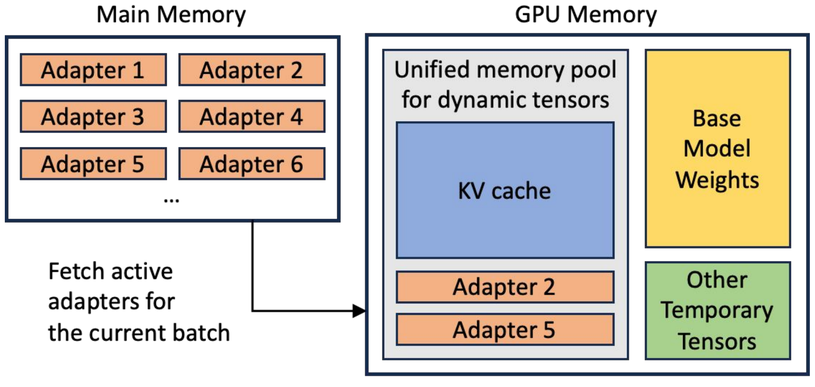
Shiyi Cao, Sijun Tan
University of California - Berkeley

Problem Statement

- Personalized LLM model serving is of essential need.
- A base model per user is expensive and can waste significant computation resources.
- **Goal: achieve scalable and accurate personalized LLM serving leveraging the Low-Rank Adaptations (LoRA) technique.**

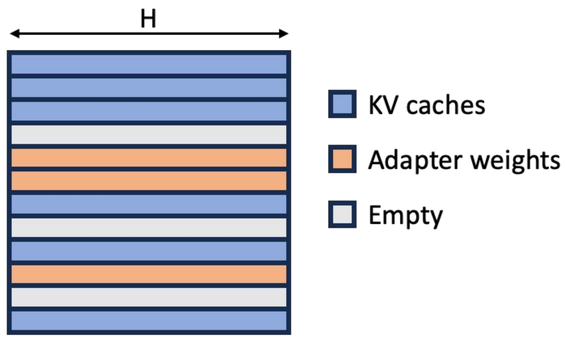


Scalable Serving System

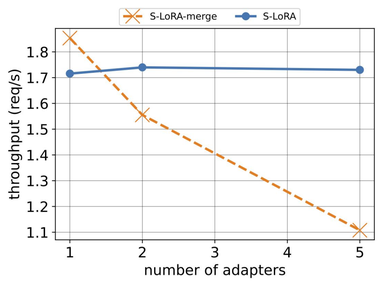


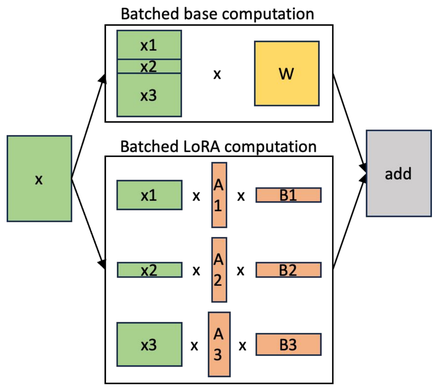
Optimizations

1. Efficient Memory Management: Unified Paging for Adapters and KV cache

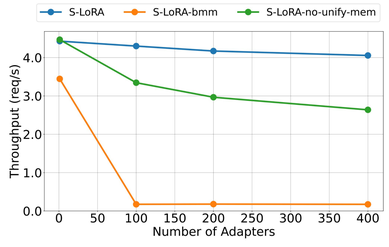


2. Heterogeneous Batching:



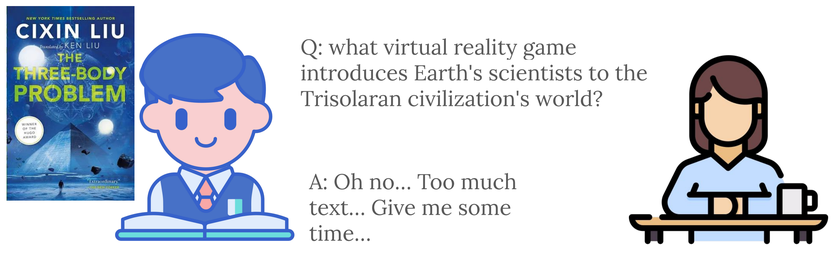


3. Performance:



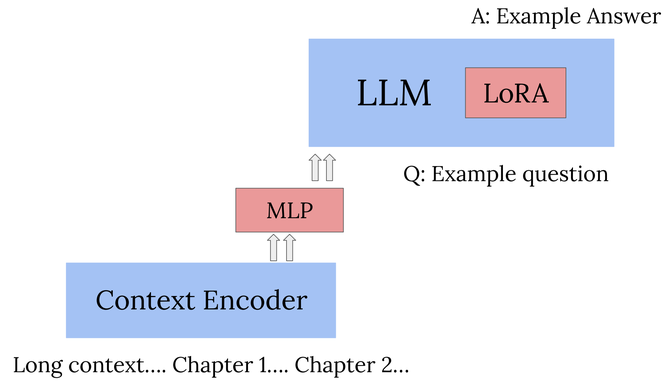
Model Setup	n	SLoRA	vLLM-packed	PEFT
S1	5	8.05	2.04	0.88
	100	7.99	OOM	0.25
	1000	7.64	OOM	-
	2000	7.61	OOM	-
S2	5	7.48	2.04	0.74
	100	7.29	OOM	0.24
	1000	6.69	OOM	-
	2000	6.71	OOM	-
S4	2	4.49	3.83	0.54
	100	4.28	OOM	0.13
	1000	3.96	OOM	-

Offline Context Learning



Handling long context is challenging for LLM, our goal is to learn thest context information offline using LoRA finetuning.

Training Methods



Learn by reconstruction: LLM should reconstruct the original context based on the context embeddings.

Learn by self-instruction: We ask the LLM to generate a few questions itself and use it for instruction finetuning.

Evaluation

- **LLM:** LLaMA2-7B-4096
- **Context Encoder:** Pretrained In-Context Autoencoder (an encoder finetuned for context compression) from Ge et al.

Baseline with Question	Baseline with Article + Question	Finetuned model (next token prediction) with Question	Off-the-shelf Pretrained In-Context Autoencoder with Question	Finetuned model (ours) with Question
31%	37%	31%	27%	39%

Collaborators

1. The LoRA serving system part is done by theS-LoRA team: Ying Sheng*, Shiyi Cao*, Dacheng Li, Coleman Richard Charles Hooper, Nicholas Lee, Shuo Yang, Christopher Chou, Banghua Zhu, Lianmin Zheng, Kurt Keutzer, Joseph E. Gonzalez, Ion Soica