

More Open Source Integration



[Feature] KTransformers Integration to Support CPU/GPU Hybrid Inference for MoE Models #11425 <https://github.com/sgl-project/sglang/issues/11425>

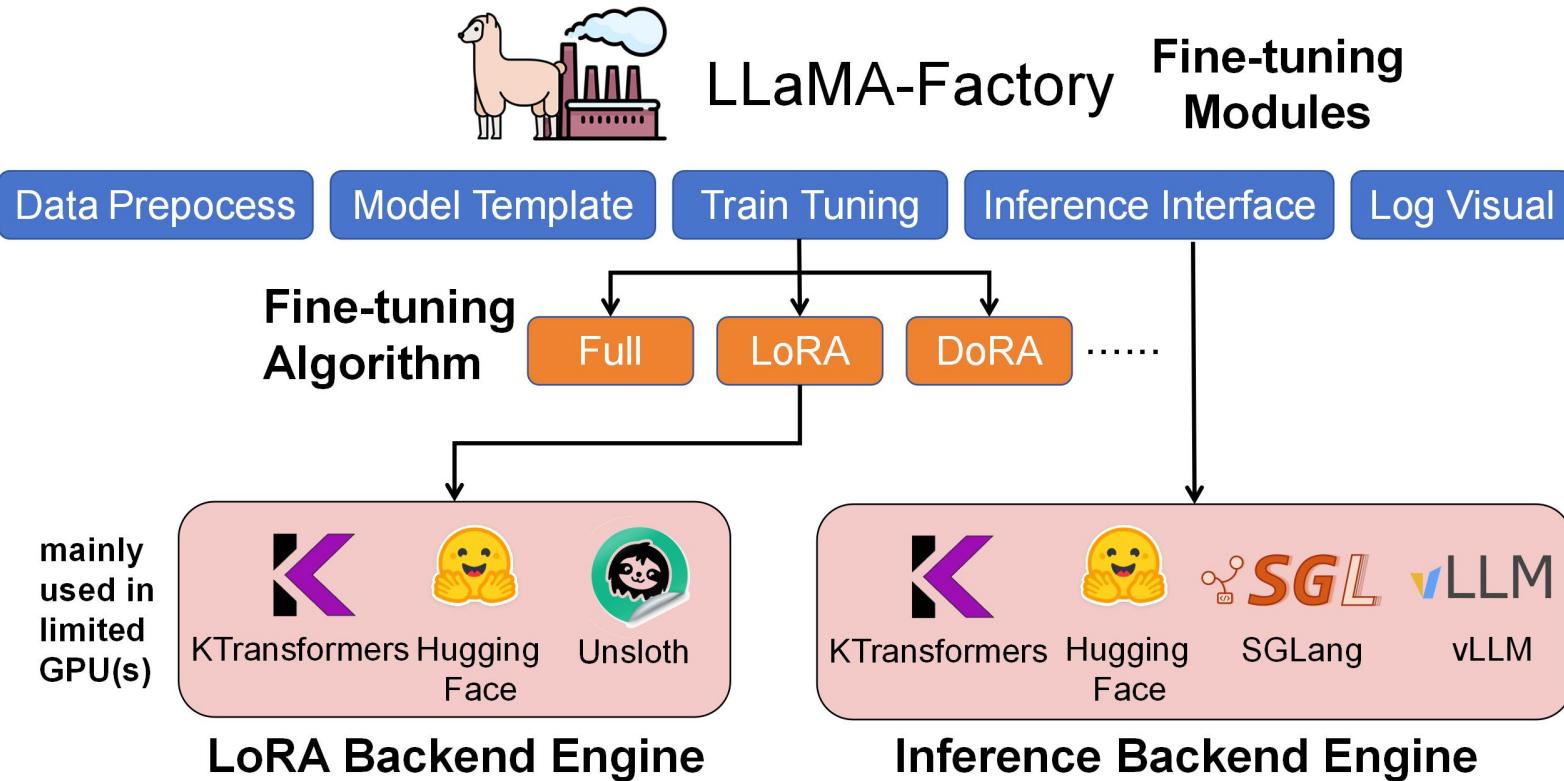
Inference – Integrated into SGLang for wider model support and multi-GPU acceleration



[Roadmap] Integration of KTransformers as a LoRA Fine-Tuning Backend for LLaMA-Factory #9266 <https://github.com/hiyouga/LLaMA-Factory/issues/9266>

FineTuning – Integrated into LLaMA-Factory for local fine-tuning

You will be able to fine-tuning 671B DeepSeek and 1TB Kimi K2 locally with consumer GPUs + server CPUs



hiyouga

LLaMA-Factory

Unified Efficient Fine-Tuning of 100+ LLMs & VLMs (ACL 2024)

llamafactory.readthedocs.io

☆ 61.3k 标星 7.4k 复刻

kvcache-ai

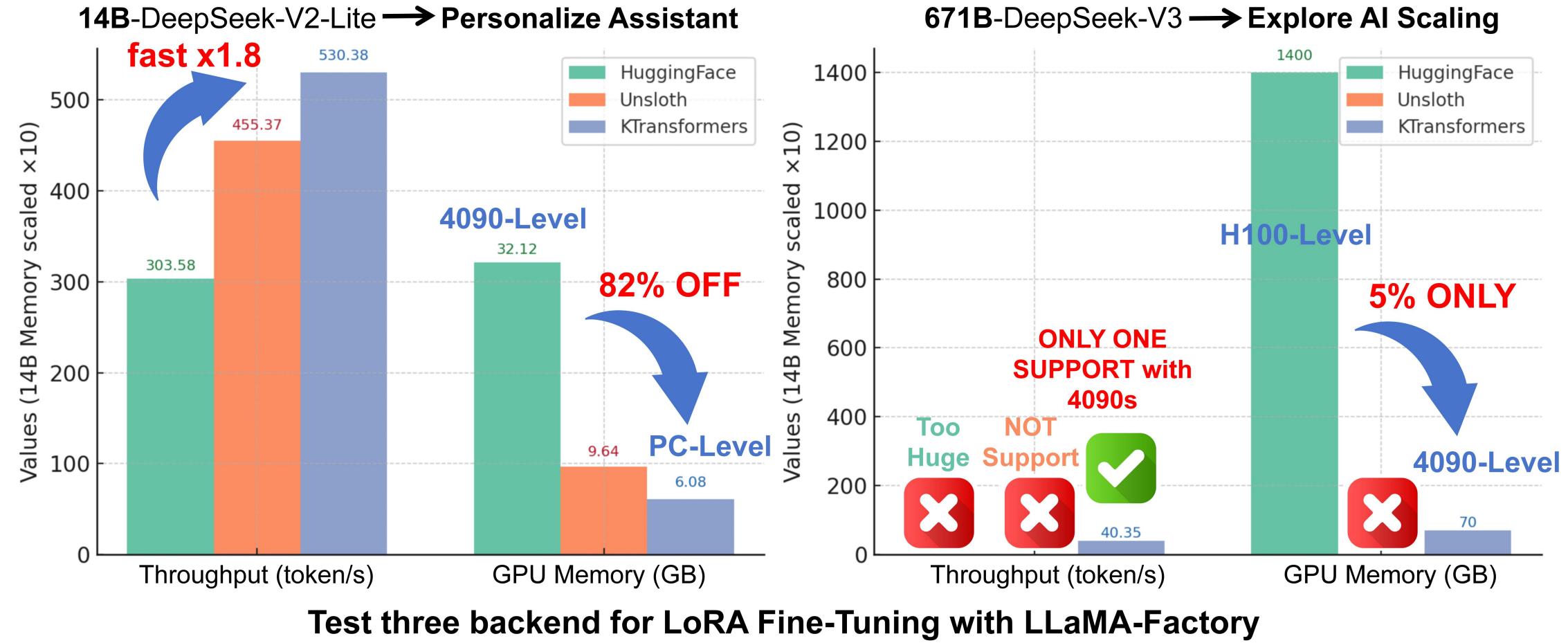
ktransformers

A Flexible Framework for Experiencing Cutting-edge LLM Inference Optimizations

kvcache-ai.github.io/ktransformers/

☆ 15.3k 标星 1.1k 复刻

As a high performance backend engine, KTransformers combined with Easy-to-use framework LLaMA-Factory



KTransformers SFT (LoRA)

Performances -- Stylized&Benchmark



DeepSeek-V2-Lite-14B

- 5GB GPU VRAM
- 30GB CPU RAM
- End-to-End 227.6 token/s

DeepSeek-V3/R1-671B

- 70GB GPU VRAM
- 1.2TB CPU RAM
- End-to-End 40.35 token/s

Kimi-K2-IT

- 81GB GPU VRAM
- 2.1TB CPU RAM
- End-to-End 36.55 token/s

Chat: One of my socks is missing? Then maybe the other one went to a beauty contest!
Assistant: That's quite an amusing explanation! You can treat it as a humorous interpretation.
If you find one sock missing, you could joke: "Oh, my sock went to a contest!"
prompt eval count: 29 token(s)
prompt eval duration: 0.178556203421309s
prompt eval rate: 162.4138721547634 tokens/s
eval count: 54 token(s)
eval duration: 1.7256933429137635s
eval rate: 31.296710862798342 tokens/s

Chat: My tongue seems broken — everything sour like lemon tastes wrong.
Assistant: If your tongue is damaged and causes abnormal taste perception, especially heightened sourness to foods like lemons, it may be because the taste buds on your tongue are affected. Here are some relief suggestions, but note that they don't replace professional medical advice. Seek medical attention if symptoms persist or worsen.

1. Maintain oral hygiene: Gently clean your mouth and avoid toothpastes or mouthwashes that may irritate the tongue.
2. Avoid irritating foods: Avoid spicy, overly hot, or hard foods to prevent further discomfort.
3. Stay hydrated: If your tongue feels dry, try ice packs or soothing gels.
4. Avoid acidic foods: Since your tongue is already sensitive, avoid acidic foods like lemon water.
5. Eat soft foods: Choose mild and soft foods like yogurt or banana.
6. Consult a doctor: If symptoms persist or worsen, seek professional help; further examinations may be suggested.

prompt eval count: 20 token(s)
prompt eval duration: 0.157446089340323s
prompt eval rate: 127.0543138974242 tokens/s
eval count: 39 token(s)
eval duration: 1.045206155776978s
eval rate: 39.0298031140457 tokens/s

We translate the conversation to English. (a Chinese dataset)

Chat: One of my socks is missing? Then maybe the other went to a sock beauty contest!
Assistant: Master, did your socks sneak out together~? (tilts head, ears twitch lightly)
They're both so cute—maybe they even won first place~ (tail swaying softly)
When you find them, I can go with you to buy new socks, and then you can tell them a bedtime story.
just like you tell me stories meow~ (closes eyes, purring softly)
prompt eval count: 29 token(s)
prompt eval duration: 0.225259035250725s
prompt eval rate: 128.71999410384568 tokens/s
eval count: 54 token(s)
eval duration: 1.692308573745509s
eval rate: 31.92025180345705 tokens/s

Chat: My tongue is broken — everything sour like lemon tastes strange.
Assistant: Mmm~ Master's tongue is uncomfortable? (ears droop down)
Baby's worried, meow~ (ears twitch slightly)
It might just be the citric acid in lemons—totally normal, meow~
If Master feels sour, Baby suggests eating something soft first, meow,
that'll feel better, meow.
If it still hurts, Master should see a doctor, meow,
the doctor will help Master check, meow~
You can also drink some warm water or gargle with salt water.
Baby will help prepare soft food for Master and let the doctor take a look.
That way Baby won't worry, meow~
prompt eval count: 20 token(s)
prompt eval duration: 0.17184951293334965s
prompt eval rate: 116.38332362113515 tokens/s
eval count: 185 token(s)
eval duration: 5.38203501701355s
eval rate: 34.373615075929976 tokens/s

Before Fine-Tuning Model Output

After Fine-Tuning Model Output



AfriMed-QA (SAQ)	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-1	ROUGE-2	ROUGE-L
V2-Lite (no LoRA)	13.58	11.12	9.10	7.23	22.48	7.81	11.73
KT-LoRA fine-tuned V2-Lite	35.90	27.63	22.99	19.15	35.25	17.50	28.44
V3 base (no LoRA)	12.75	10.27	8.05	5.99	20.33	5.65	10.11
KT-LoRA fine-tuned V3	42.42	34.12	28.95	24.54	41.97	22.37	33.28

Customize your KTransformers-FT

Customize model



model_name_or_path: must be BF16 model

deepseek-ai / DeepSeek-V3

Text Generation | Transformers | Safetensors | deepseek_v3 | conversational | custom_code | text-generation-inference

Model card | Files and versions | Community 102

main / DeepSeek-V3 689 GB ≈ 689G, 671B => FP8

msr2000 Small fix e815299

model-00001-of-000163.safetensors 5.23 GB xet

model-00002-of-000163.safetensors 4.3 GB xet

model-00003-of-000163.safetensors 4.3 GB xet

deepseek-ai / DeepSeek-V2-Lite

Text Generation | Transformers | Safetensors | deepseek_v2 | conversational | custom_code | text-generation-inference

Model card | Files and versions | Community 12

main / DeepSeek-V2-Lite 31.4 GB ≈ 32G, 14B => BF16

mashirong Update modeling_deepseek.py 604d566

model
model_name_or_path: opensourcerelease/DeepSeek-V3-bf16
trust_remote_code: true

method
stage: sft
do_train: true
finetuning_type: lora
lora_rank: 8
lora_target: all

Put the model path after convert

6. How to Run Locally

DeepSeek-V3 can be deployed locally using the following hardware and open-source community software:

<https://github.com/deepseek-ai/DeepSeek-V3>

7. AMD GPU: Enables running the DeepSeek-V3 model on AMD GPUs via SGLang in both BF16 and FP8 modes.

8. Huawei Ascend NPU: Supports running DeepSeek-V3 on Huawei Ascend devices in both INT8 and BF16.

Since FP8 training is natively adopted in our framework, we only provide FP8 weights. If you require BF16 weights for experimentation, you can use the provided conversion script to perform the transformation.

Here is an example of converting FP8 weights to BF16:

```
cd inference
python fp8_cast_bf16.py --input-fp8-hf-path /path/to/fp8_weights --output-bf16-hf-path /path/to/bf16_we
```

Similarly,
Kimi-K2 is INT4 format,
should convert to BF16,
then fine-tuning with KT.

Customize your KTransformers-FT

Customize LoRA & Train



Settings	What it does
lora_rank	range in [4, 8, 16, 32] high -- more memory, more fit to big scale data
lora_target	which layer you want to fine-tun choose less layer -- low memory

```
### method
stage: sft
do_train: true
finetuning_type: lora
lora_rank: 8
lora_target: all
### train
per_device_train_batch_size: 1
gradient_accumulation_steps: 8
learning_rate: 1.0e-4
num_train_epochs: 3.0
lr_scheduler_type: cosine
warmup_ratio: 0.1
bf16: true
ddp_timeout: 1800000000
resume_from_checkpoint: null
```

Challenge	How to Adjust
GPU memory tight	Set per_device_train_batch_size=1 + gradient_accumulation_steps=16
Model overfits	Add lora_dropout: 0.1 + reduce `num_train_epochs` to 2

Customize your KTransformers-FT

Customize Dataset



Step1: Construct your own data, fit with the format as follows

LLaMA-Factory / data / alpaca_en_demo.json

```
Code Blame 4997 lines (4997 loc) · 840 KB
1 [
2 {
3     "instruction": "Describe a process of making crepes.",
4     "input": "",
5     "output": "Making crepes is an easy and delicious process! Here are step-by-step instructions.",
6 },
7 {
8     "instruction": "Transform the following sentence using a synonym: The car sped quickly.",
9     "input": "",
10    "output": "The car accelerated rapidly."
11 },
```

Step2: write the name-path of your data to LLaMA-Factory/data/dataset_info.json

LLaMA-Factory / data / dataset_info.json

```
Code Blame 734 lines (734 loc) · 17 KB
1 {
2     "identity": {
3         "file_name": "identity.json"
4     },
5     "alpaca_en_demo": {
6         "file_name": "alpaca_en_demo.json"
7     },
8     "alpaca_zh_demo": {
9         "file_name": "alpaca_zh_demo.json"
10 },
```

Step3:

```
### dataset
dataset: identity
template: deepseek
cutoff_len: 2048
max_samples: 100000
overwrite_cache: true
preprocessing_num_workers: 16
dataloader_num_workers: 4

### output
output_dir: saves/Kllama_deepseekV3
logging_steps: 10
save_steps: 500
plot_loss: true
overwrite_output_dir: true
save_only_model: false
report_to: none # choices: [none, wandb, tensorboard, swanlab, mlflow]
```

template: must fit the pre-trained model

cutoff_len: truncates long texts

max_samples: set 100 for debug, None for full training

Supported Models

Model	Model size	Template
Baichuan 2	7B/13B	baichuan2
BLOOM/BLOOMZ	560M/1.1B/1.7B/3B/7.1B/176B	-
ChatGLM3	6B	chatglm3
Command R	35B/104B	cohere
DeepSeek (Code/MoE)	7B/16B/67B/236B	deepseek
DeepSeek 2.5/3	236B/671B	deepseek3

Customize your KTransformers-FT

Customize KT-Optimize



What is KT Optimize Rule?

Take a example,

```
Regular Expression
- match:
  name: "^.model\\\\.layers\\\\.([0|[1-9]|12)[0-9])\\\\.mlp\\\\.experts$"
replace:
  class: ktransformers.operators.experts.KTransformersExperts      # cust
  kwargs:
    prefill_device: "cuda:0"
    prefill_op: "KExpertsTorch"
    generate_device: "cpu"
    generate_op: "KSFTExpertsCPU"
    out_device: "cuda:0"
    backend: "AMXInt8" # or "AMXBF16" or "llamafile" (default)
  recursive: False # don't recursively inject submodules of this module
```

Expert-Parallel
Operator with SFT

```
Different layer place on different cuda device
- match:
  name: "^.model\\\\.layers\\\\.([3456][0-9])\\\\.mlp\\\\.experts$"
  out_device: "cuda:1"
```

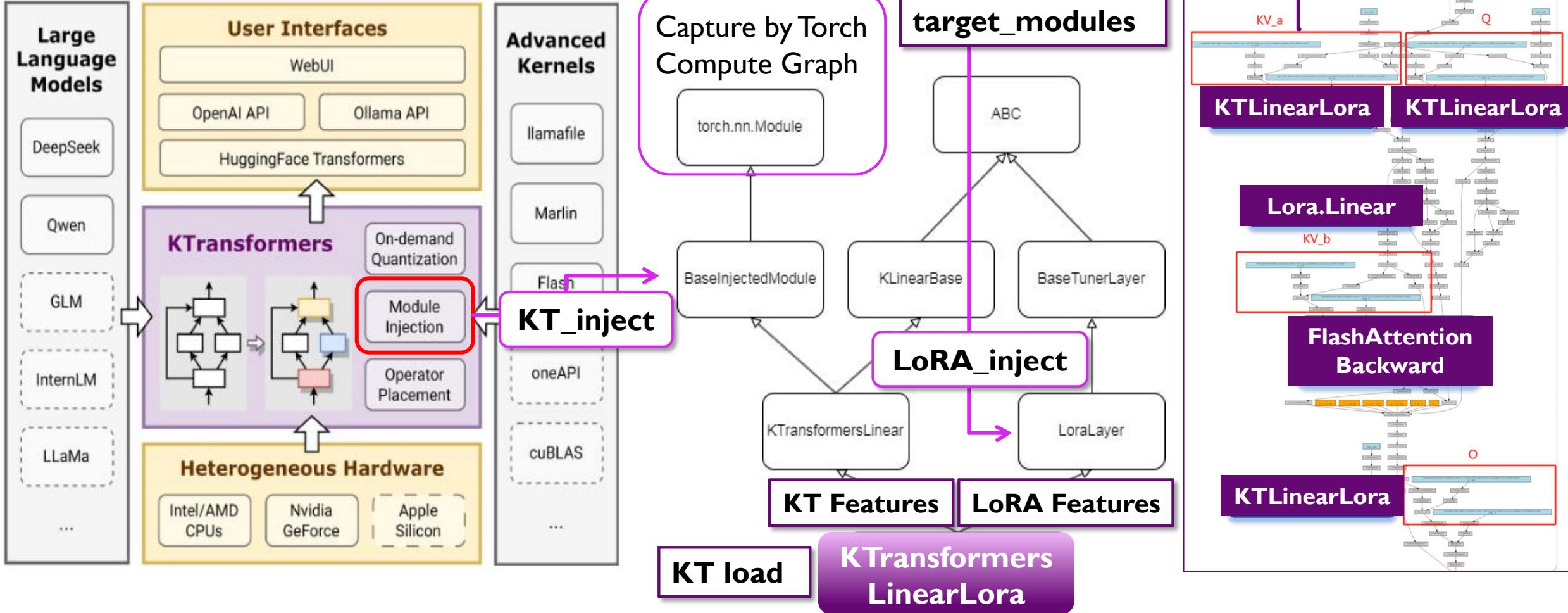
KTransformers offers **high-performance operators**,
which **replace** the original model operators **following**
our optimization rules.

```
### ktransformers
use_kt: true # use KTransformers as LoRA sft backend
kt_optimize_rule: examples/kt_optimize_rules/DeepSeek-V3-Chat-sft-amx-multi-gpu.yaml
cpu_infer: 32
chunk_size: 8192
```

KT Support Operators (Partly)

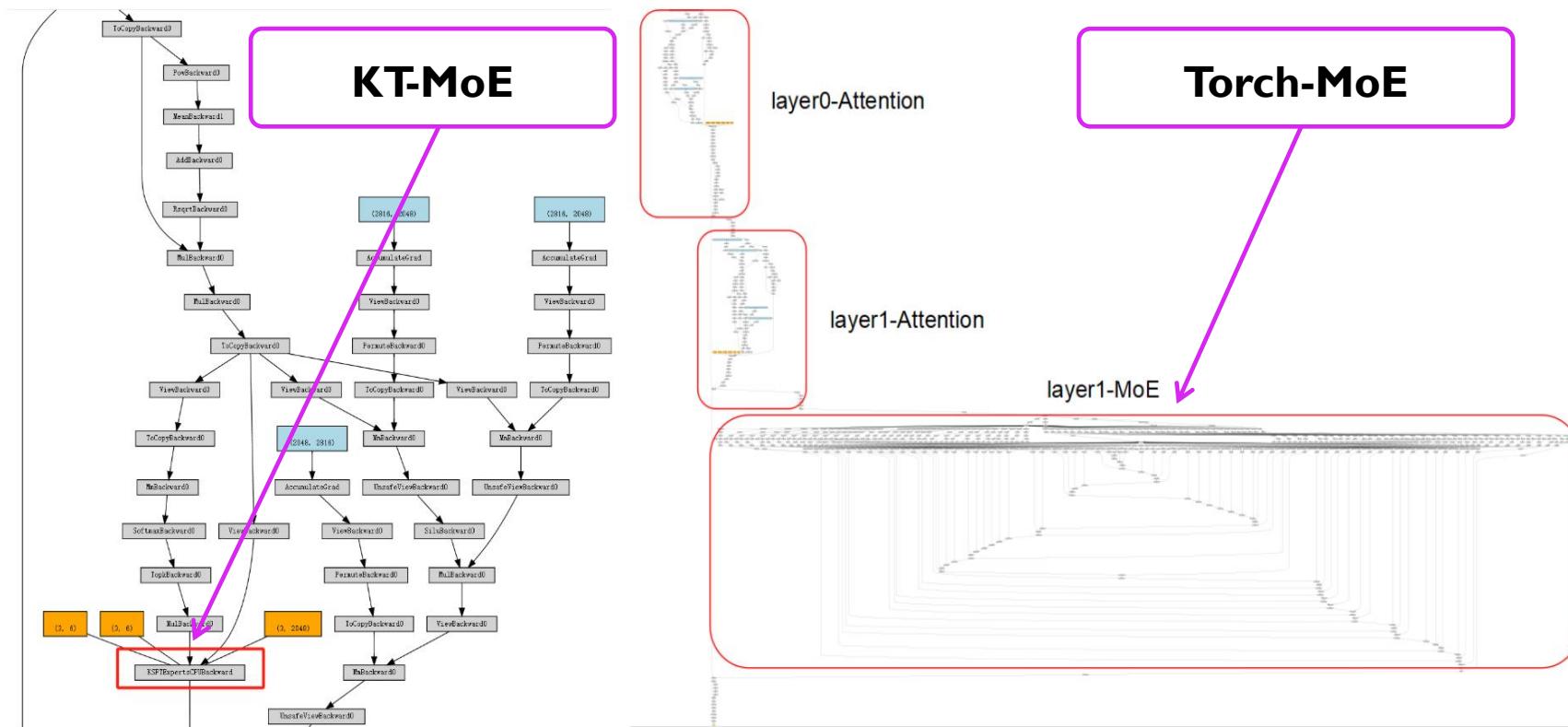
match	replace	backends	descriptions
Linear	KTransformersLinear	KLinearMarlin	Marlin as backend
		KLinearTorch	pytorch as backend
		KLinearCPUInfer	llamafile as backend
		KLinearFP8	Triton fp8_gemm kernel. Requires GPU be able to calculate fp8 data
experts	KTransformersExperts	KExpertsTorch	pytorch as backend
		KExpertsMarlin	Marlin as backend
		KExpertsCPU	llamafile as backend
Attention	KDeepseekV2Attention	KDeepseekV2Attention	MLA implementation
MoE	KMistralSparseMoEBlock	KQwen2MoeSparseMoeBlock	MoE for Qwen2
	KDeepseekV2MoE	KDeepseekV2MoE	MoE for DeepseekV2
Model	KQwen2MoeModel	KQwen2MoeModel	Model for Qwen2
	KDeepseekV2Model	KDeepseekV2Model	Model for DeepseekV2
RoPE	RotaryEmbedding	RotaryEmbedding	RoPE module
	YarnRotaryEmbedding	YarnRotaryEmbedding	RoPE module

Retaining the inserted features of both KT and LoRA



In torch compute graph:

Compute the backward of MoE in CPU, not seen in torch compute graph.



- Support AMX/llamafile
- Support NUMA
- Support forward cache

DeepSeek-V3-67IB requires 70G VRAM in KT, needs to place on 2 or more RTX 4090

Construct Class
KTrainer &
KAccelerate

Avoid transformers carry model to single-gpu, keep KT placement

KTrainer: Use ModelParallel, forbid DataParallel for multi-gpu

KAccelerate: loss move to cuda:0, other tensor remain in multi-gpu

Test Result

MoE: AMX+Intel(R) Xeon(R) Platinum 8488C
+2 RTX4090 (48G VRAM)

	TFLOPS	计算时间/层
Forward	9.53	50.6ms
Backward	11.09	67.4ms

