Long Exposure: Accelerating Parameter-Efficient Fine-Tuning for LLMs under Shadowy Sparsity

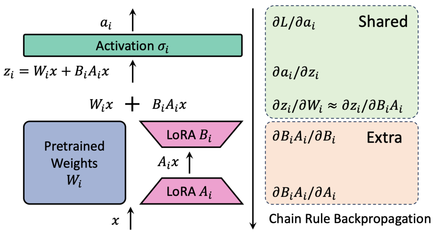
Tuowei Wang*, Kun Li*, Zixu Hao, Donglin Bai,
Ju Ren, Yaoxue Zhang, Ting Cao, Mao Yang
Tsinghua University, Microsoft Research

Introduction

In natural language processing, the adaptation of pre-trained large language models (LLMs) to diverse downstream tasks constitutes a fundamental aspect of many applications. This adaptation process, commonly known as *fine-tuning*, involves the comprehensive update of all parameters within the pre-trained model akin to training from scratch.

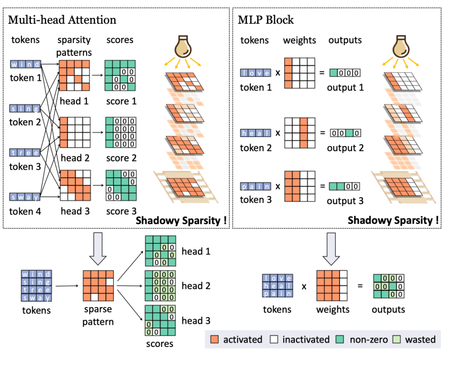
The major reason hindering the fine-tuning efficiency is the retention of the same number of parameters in the new model as in the original one. Efforts have been made to address this concern by introducing *parameter-efficient fine-tuning* (PEFT) techniques, which only selects or injects a minimal number of parameters for adaption to new tasks. One prominent approach in the domain of PEFT is low-rank adaption (LoRA). It freezes pre-trained model weights and injects smaller, trainable low-rank matrices into each transformer block.

This substantial reduction in the number of trainable parameters mitigates the need for maintaining and updating the optimizer states for most parameters. However, PEFT techniques fall short of achieving an expected decrease in wall-clock time. Even with minimal parameters being trainable, techniques like LoRA only experience an 18% reduction in wall-clock time. While PEFT techniques notably cut down the optimization step's wall-clock time, they leave the duration of the forward and backward passes either unchanged or slightly increased. This is because, despite most pre-trained parameters being frozen, computing gradients for trainable parameters still requires complete forward and backward passes through the backbone model. Consequently, the forward and backward passes have emerged as the computational bottlenecks impeding further acceleration.



Phase	Forward	Backward	Optim. Step	Total
Full Param.	112.8(27.7%)	223.7(54.9%)	70.6(17.3%)	407.2
LoRA [6]	135.3(40.4%)	196.3(58.7%)	2.0(0.6%)	334.6
Adapter [7]	123.6(42.2%)	168.4(57.5%)	0.7(0.3%)	292.9
Bitfit [8]	117.6(40.5%)	172.4(59.4%)	0.2(0.07%)	290.3
P-Tuning [9]	137.5(40.1%)	193.9(56.6%)	11.1(3.2%)	342.6

In this paper, we propose Long Exposure, an efficient system to accelerate parameter-efficient fine-tuning for LLMs. The design of Long Exposure is grounded in a crucial observation that PEFT and inference in LLMs exhibit high similarities in their computation patterns. In PEFT techniques, a majority of model parameters remain frozen, similar to the scenario in model inference where parameters also stay unaltered. Previous studies have evidenced that LLMs typically exhibit considerable sparsity, with a great number of activations can be excluded from computation to expedite inference in wall-clock time while preserving quality. Guided by this observation, the key insight of Long Exposure is inspired: given the striking similarities in computation patterns between PEFT and inference, **why not build a bridge to PEFT acceleration by capturing intrinsic sparsity like inference?**



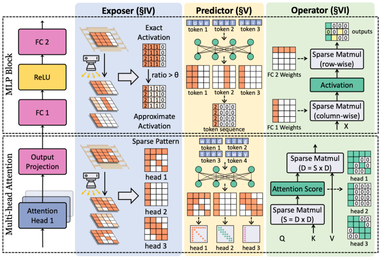
However, this is not a low-hanging fruit, as the sparsity inherent in fine-tuning introduces distinct characteristics that diverge significantly from those encountered during inference. In inference, the model typically processes one token at a time, where the sparse pattern is easily discernible for each token. In contrast, fine-tuning involves feeding the model with a sequence of tokens, where the sparsity patterns heavily overlap across different tokens. We coin this intricate sparsity observed in fine-tuning as **Shadowy Sparsity**.

To accelerate PEFT for LLMs under this shadowy sparsity, several key technical challenges must be tackled carefully.

① *How to capture more sparse patterns under shadowy sparsity, avoiding potential computational waste?*
② *How to predict efficient yet accurate sparse patterns to minimize associated computational expenses before incurring actual costs?*
③ *How to achieve effective performance improvements based on well-predicted sparsity?*

Design

The concept of Long Exposure emphasizes that rather than simply harnessing the limited sparsity remaining in shadowy sparsity, we take a longer view which captures more intricate details of individual sparse pattern before they fade into shadow.



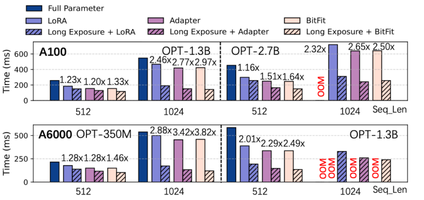
The core of Long Exposure is the **Shadowy-sparsity Exposer**, designed for exposing the latent sparsity hidden in shadowy sparsity. In MHA block, we introduce specific sparse patterns tailored to each attention head, avoiding the computational redundancy or oversight that can arise from employing a uniform mask. In MLP block, we take the importance of each activated neuron into consideration. By identifying and filtering out neurons whose activation can be safely disregarded, we transform shadowy sparsity into structured block-wise sparsity.

Long Exposure utilizes **Sequence-oriented Predictors** to address the conflicts between long sequence inputs and the associated neural network size. This technique is grounded in a two-stage design strategy: Initially, the predictor processes each token individually; then these predictions are subsequently consolidated. Moreover, to minimize the disruption caused by updating trainable parameters, we introduce specific training optimizations to bolster the predictor's robustness.

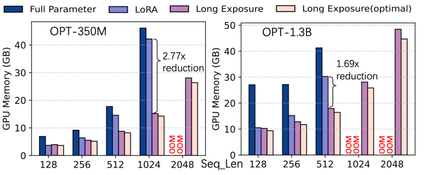
Long Exposure develops a collection of **Dynamic-aware Operators** to facilitate practical acceleration on hardware systems, covering all the sparse operations involved in MHA and MLP block. Different from most existing tools, these operators avoid additional data conversion overhead, making them well-suited for dynamic scenarios. In addition, we design a two-stage algorithm for multi-head attention that adeptly balances precomputation with dynamic sparse patterns.

Evaluation

Execution Time. Integrating Long Exposure into three exemplary PEFT techniques, we examine two different parameter sizes and sequence lengths for each one. The results indicate that our system achieves up to 1.25× and 2.49 × speedup on average for OPT-1.3B with a sequence length of 512 and 1024 on A100, respectively. With a larger 2.7B model, the speedup remains consistent, averaging 1.44× and 2.49×, respectively. Parallel results are observed on A6000, underscoring the robustness and reliability of our system.



Memory Footprint. Despite not being explicitly designed for memory efficiency, the application of head-specific sparse attention masks alters the memory complexity from quadratic to linear, leading to lower memory footprints. Furthermore, selective activating model weights in MLP block permits the majority of the model to reside on the CPU, with only the active weights being transferred to the GPU for processing. This strategy can lead to additional memory savings.



Model Accuracy. We investigate the impact of Long Exposure on model accuracy by comparing with original LoRA across a variety of downstream tasks. We fine-tune OPT models of three distinct sizes on the Alpaca dataset. The results show that Long Exposure incurs only a minimal loss in downstream task accuracy across all model sizes and task types. This is because the essence of sparsity lies in disregarding the computation of elements that are zero or nearly zero, thereby only marginally affecting the final results.

		350M-w/o	350M-w	1.3B-w/o	1.3B-w	2.7B-w/o	2.7B-w
PIQA	Acc.	65.13%	64.80%	72.25%	72.09%	74.70%	73.45%
	Stderr	1.11%	1.12%	1.05%	1.06%	1.02%	1.02%
Winog.	Acc.	53.04%	53.12%	58.88%	58.80%	62.27%	62.19%
	Stderr	1.40%	1.40%	1.38%	1.38%	1.37%	1.36%
RTE	Acc.	54.51%	55.60%	54.15%	54.51%	52.71%	53.79%
	Stderr	2.99%	3.01%	3.01%	3.01%	3.00%	2.04%
COPA	Acc.	69.00%	70.00%	81.00%	81.00%	78.00%	76.00%
	Stderr	4.61%	4.51%	4.23%	4.02%	4.29%	4.09%
Hella.	Acc.	32.26%	32.40%	42.08%	42.11%	46.76%	43.95%
	Stderr	0.47%	0.47%	0.499%	0.49%	0.50%	0.50%