SharpZO: Hybrid Sharpness-Aware Vision Language Model Prompt Tuning via Forward-Only Passes

Yifan Yang, Zhen Zhang, Rupak Vignesh Swaminathan, Jing Liu, Nathan Susanj, Zheng Zhang

amazon | science UCSB

Motivation

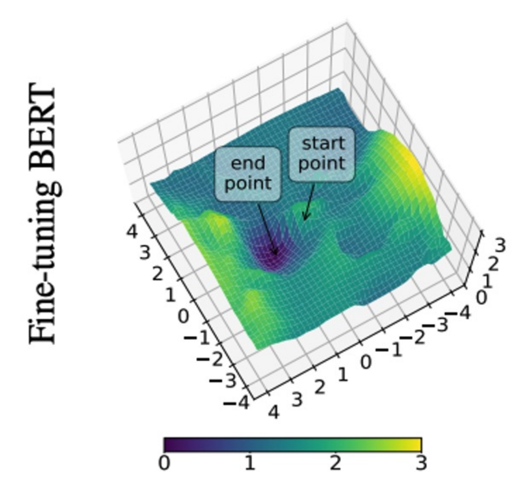
Prompt-tuning via Forward-only Passes

- High-performing foundation models are provided **only as a software-as-a-service without model details.**
- Forward-only fine-tuning enable training under constraints like **less GPU memory or adaption on inference-only engines**

Problem with Previous Zeroth-Order (ZO) Methods

The success of ZO fine-tuning largely depends on two key perspectives:

- **Loss Landscape:** Having good initialization near a smooth optimal region, typically offered by fine-tuning tasks [1].
- **Optimizer:** Maintaining a compact parameter space and minimizing ZO estimation noise.



However, existing ZO studies primarily focus on improving performance from the optimizer perspective, while **the loss landscape perspective remains largely unexplored**.

SharpZO: A Hybrid BP-free Optimizer

SharpZO improve ZO fine-tuning performance from a **loss landscape perspective**, introducing two-stage hybrid optimization framework:

- **Stage 1:** Sharpness-aware CMA-ES for initialization
- **Stage 2:** Sparse ZO fine-tuning for fine-grained optimization

	Stage 1: Evaluation Strategy	Stage 2: ZO
Exploration	Strong **global** search capability via adaptive sampling	Primarily **local** search; relies on random perturbations around current point
Computation Cost	**High** (due to population evaluations and matrix updates)	**Lower** (typically fewer perturbations; no covariance updates)

Methods

Hybrid Framework with Three Types of BP-free Optimizers

- **ZO Randomized (RGE) and Coordinate-wise (CGE) Gradient Estimation [2]**

We leverage both RGE and CGE to meet different gradient estimation requirements:

$$\textbf{(RGE)}\ \hat{\nabla}\mathcal{L}(\boldsymbol{w}) = \frac{1}{q}\sum_{i=1}^{q}\left[\frac{\mathcal{L}(\boldsymbol{w}+\mu\boldsymbol{u}_i)-\mathcal{L}(\boldsymbol{w}-\mu\boldsymbol{u}_i)}{2\mu}\boldsymbol{u}_i\right];\ \textbf{(CGE)}\ \hat{\nabla}\mathcal{L}(\boldsymbol{w}) = \sum_{i=1}^{d}\left[\frac{\mathcal{L}(\boldsymbol{w}+\mu\boldsymbol{e}_i)-\mathcal{L}(\boldsymbol{w}-\mu\boldsymbol{e}_i)}{2\mu}\boldsymbol{e}_i\right].$$

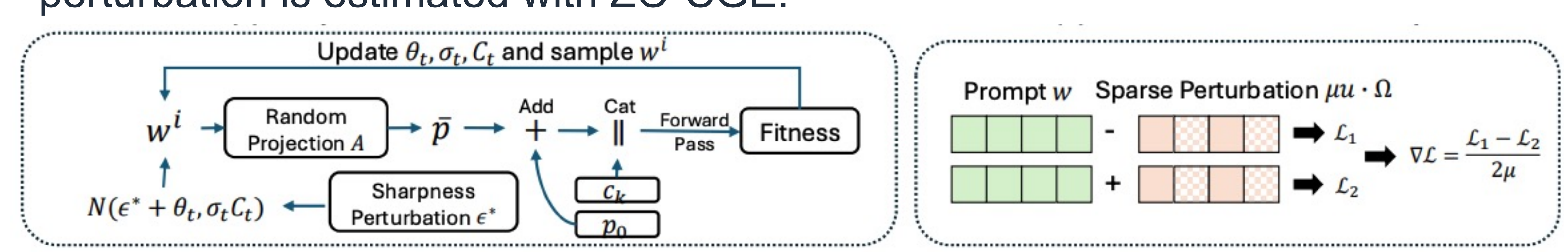
where RGE add a random perturbation $\boldsymbol{u}$ to all parameters while CGE estimated the gradient for each parameter individually by adding basis vector **e**.

- **Covariance matrix adaptation evolution strategy (CMA-ES)**

We propose a sharpness-aware alternative CMA-ES optimizer, which provides both a smoother loss landscape and a strong initialization for the second stage through distributional shift.

Workflow for the SharpZO Method

The parameters are optimized by sharpness-aware CMA-ES in the early stage, then being fine-tuned with ZO-RGE optimizer. In the first stage, the sharpness perturbation is estimated with ZO-CGE.



(c) Stage 1: Sharpness-aware CMA-ES **(d) Stage 2: Sparse ZO optimization**

Reference

[1] Visualizing and Understanding the Effectiveness of BERT, EMNLP 2019

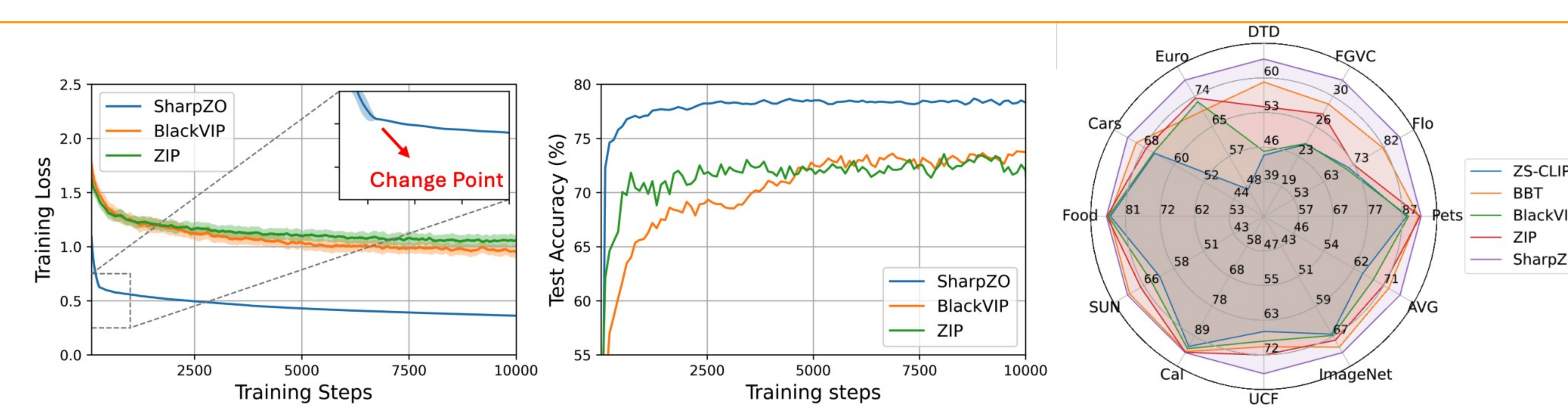
[2] Deepzero: Scaling up zeroth-order optimization for deep model training, ICLR 2024

Experiential Results

We evaluate the proposed method on CLIP fine-tuning benchmarks, covering a total of 11 downstream tasks. We present two key experimental analyses:

(a) Comparison with ZO Baselines.

(b) Fine-tuned performance across all 11 tasks



(a) Curve for Training Loss and Testing Accuracy on EuroSAT (b) Downstream Generalization

*Corresponding author: yifanyang@ucsb.edu