NEURAL INFORMATION PROCESSING SYSTEMS 2025

DEPARTMENT OF ENGINEERING SCIENCE

UNIVERSITY OF OXFORD

Optimization Inspired Few-Shot Adaptation for Large Language Models

Boyan Gao, Xin Wang, Yibo Yang, David A. Clifton

Motivation

- Limitation of existing few-shot adaption methods:
  - In-context learning (ICL) methods:
    - Instable performance to the format of prompts
    - No learnable parameters to capture the complicated features
  - PEFT (LoRA) based methods:
    - Overfitting to few-shot data
- In this work, we ask:
  - *For few-shot adaptation, how can we develop an efficient method that avoids overfitting to few-shot data, as commonly observed in PEFT, while also overcoming ICL's lack of learnable parameters and extra inference cost?*

Method

- Reframing the forward pass as a *learnable preconditioned gradient descent* (PGD) parameterised by LayerNorm:

$$Z_{t+1} = Z_t - P_t \nabla \mathcal{L}(Z_t), \quad P_t = \Gamma_t \cdot \frac{1}{\sigma_t}$$

- Optimization with fast convergence:

$$\mathcal{J}(P) = \sum_{t=1}^{T-1} \frac{\|Z_t - Z_{t+1}\|}{\|Z_t - Z_{t-1}\|}$$

- Improving generalization through sharpness-aware regularization:

$$\mathcal{H}(P) = tr(P_t \nabla^2 \mathcal{L}(Z_t) P_t^T)$$

Preconditioned GD in LLM

- LLMs learn layer-wise gradient descent in the forward pass

$$Z_{t+1} = Z_t - \eta \nabla \mathcal{L}(Z_t)$$
$$\textbf{s.t. } f_t(Z_t) = -\eta \nabla \mathcal{L}(Z_t) = \mathrm{Attn}(Z_t)$$

- Reframe this as as preconditioned gradient descent by treating LayerNorm as learnable preconditioner $P = \{P_t\}_{t=1}^{T}$

$$Z_{t+1} = Z_t - \Gamma_t \cdot \frac{\nabla \mathcal{L}(Z_t) - \mu_t}{\sigma_t}, \quad \Gamma_t = \mathrm{diag}(\gamma_t)$$

$$Z_{t+1} = Z_t - P_t \nabla \mathcal{L}(Z_t), \quad P_t = \Gamma_t \cdot \frac{1}{\sigma_t}$$

Fast Convergence

- Fast Convergence objective:

$$\mathcal{J}(P) = \sum_{t=1}^{T-1} \frac{\|Z_t - Z_{t+1}\|}{\|Z_t - Z_{t-1}\|}$$

  - Step ratio: optimization efficiency and stability described by:

$$\|Z_{t+1} - Z^*\| \leq \rho_t \|Z_t - Z^*\|, \;\; \rho_t < 1$$

- Implicitly optimizing the convergence bound of preconditioned gradient descent

Generalization

- Optimizing the local sharpness of loss landscape
  - Hutchinson approximation for the preconditioning Hessian trace

$$\mathcal{H}(P) = tr(P_t \nabla^2 \mathcal{L}(Z_t) P_t^T) \approx \frac{1}{\epsilon}\mathbb{E}_\nu \left[\nu^T P_t (\nabla \mathcal{L}(Z_t + \epsilon P_t \nu) - \nabla \mathcal{L}(Z_t)) \right]$$
$$\approx \frac{1}{\epsilon}\frac{1}{N}\sum_i \left[\nu_i^T P_t (\nabla \mathcal{L}(Z_t + \epsilon P_t \nu_i) - \nabla \mathcal{L}(Z_t)) \right]$$

The algorithm for this approximation is given in our paper

Experiments

DEPARTMENT OF ENGINEERING SCIENCE

UNIVERSITY OF OXFORD

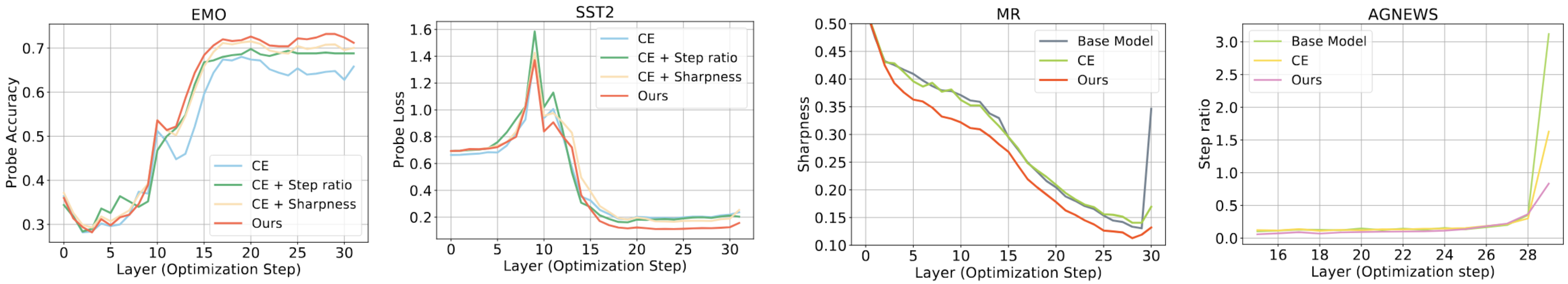
- SOTA on various few-shot benchmarks when competing with the baseline model

Table 1: Comparison between OFA and other baseline algorithms on Llama2-7B and Llama3-8B-Instruct. Mean accuracy and standard deviation across five random seeds are reported. **Best** results are highlighted in bold.

Dataset	SST-2	SST-5	TREC	AGNews	Subj	HateSp18	DBPedia	EmoC	MR
Method					Llama2-7B				
Zero-shot	83.00	27.00	50.00	70.20	51.40	54.20	72.00	41.80	73.60
Few-shot (ICL)	$94.44_{\pm1.44}$	$41.72_{\pm3.68}$	$77.32_{\pm4.41}$	$85.68_{\pm2.00}$	$52.56_{\pm3.09}$	$70.24_{\pm5.80}$	$96.64_{\pm0.48}$	$75.48_{\pm1.63}$	$93.24_{\pm0.50}$
Soft-prompt	$56.24_{\pm6.99}$	$24.24_{\pm2.96}$	$55.20_{\pm4.14}$	$78.00_{\pm7.60}$	$57.40_{\pm4.93}$	$59.56_{\pm6.96}$	$74.40_{\pm6.43}$	$35.08_{\pm5.29}$	$54.32_{\pm1.76}$
Label-anchor	$83.32_{\pm5.95}$	$27.68_{\pm4.21}$	$77.48_{\pm3.49}$	$83.72_{\pm1.04}$	$53.00_{\pm2.95}$	$64.52_{\pm8.09}$	$81.40_{\pm3.67}$	$59.12_{\pm10.60}$	$84.40_{\pm5.89}$
Task-vector	$81.44_{\pm4.73}$	$25.96_{\pm0.59}$	$65.68_{\pm1.93}$	$79.68_{\pm4.07}$	$58.56_{\pm4.91}$	$67.68_{\pm3.70}$	$89.48_{\pm2.58}$	$44.64_{\pm3.53}$	$82.32_{\pm5.37}$
IA3	$93.28_{\pm2.29}$	$46.08_{\pm2.11}$	$84.40_{\pm5.99}$	$87.04_{\pm1.97}$	$71.92_{\pm8.08}$	$72.44_{\pm2.59}$	$94.68_{\pm1.09}$	$64.32_{\pm1.95}$	$88.80_{\pm2.28}$
I2CL	$87.68_{\pm2.47}$	$39.12_{\pm2.69}$	$78.56_{\pm5.32}$	$85.48_{\pm1.16}$	$73.84_{\pm3.84}$	$69.88_{\pm5.67}$	$90.16_{\pm1.86}$	$63.72_{\pm1.37}$	$87.68_{\pm2.26}$
OFA (Ours)	$\mathbf{95.84}_{\pm0.41}$	$\mathbf{50.36}_{\pm3.28}$	$\mathbf{85.92}_{\pm1.90}$	$\mathbf{89.00}_{\pm1.26}$	$\mathbf{88.40}_{\pm4.76}$	$\mathbf{83.04}_{\pm3.72}$	$\mathbf{97.72}_{\pm0.52}$	$\mathbf{76.60}_{\pm2.39}$	$\mathbf{94.36}_{\pm1.13}$
					Llama3-8B-Instruct				
Zero-shot	93.00	35.80	71.00	80.40	50.80	67.80	67.40	53.60	86.40
Few-shot (ICL)	$96.48_{\pm0.48}$	$46.72_{\pm2.64}$	$79.92_{\pm5.83}$	$89.64_{\pm0.59}$	$57.48_{\pm7.08}$	$52.72_{\pm2.35}$	$97.00_{\pm0.28}$	$65.28_{\pm4.29}$	$93.12_{\pm0.16}$
Soft-prompt	$84.68_{\pm7.71}$	$38.40_{\pm5.68}$	$75.68_{\pm8.17}$	$84.96_{\pm3.80}$	$73.28_{\pm5.41}$	$62.72_{\pm5.54}$	$82.88_{\pm6.45}$	$55.32_{\pm9.74}$	$75.76_{\pm7.71}$
Label-anchor	$93.36_{\pm2.39}$	$40.54_{\pm5.44}$	$78.28_{\pm4.07}$	$84.64_{\pm1.61}$	$54.16_{\pm2.25}$	$69.48_{\pm5.43}$	$87.48_{\pm3.04}$	$59.36_{\pm2.48}$	$88.20_{\pm3.69}$
Task-vector	$94.80_{\pm2.02}$	$56.42_{\pm1.15}$	$79.83_{\pm1.52}$	$89.21_{\pm0.58}$	$76.08_{\pm1.23}$	$67.12_{\pm0.32}$	$79.52_{\pm1.84}$	$57.96_{\pm4.59}$	$86.52_{\pm0.64}$
IA3	$94.32_{\pm0.82}$	$49.24_{\pm2.06}$	$87.60_{\pm3.46}$	$88.36_{\pm1.80}$	$82.04_{\pm7.43}$	$77.20_{\pm4.37}$	$92.56_{\pm1.82}$	$68.04_{\pm2.24}$	$91.76_{\pm0.43}$
I2CL	$90.84_{\pm0.98}$	$48.96_{\pm2.48}$	$79.60_{\pm6.22}$	$88.96_{\pm2.03}$	$81.48_{\pm4.68}$	$65.88_{\pm3.61}$	$91.20_{\pm2.03}$	$64.32_{\pm2.05}$	$88.88_{\pm0.61}$
OFA (Ours)	$\mathbf{97.08}_{\pm0.27}$	$\mathbf{58.32}_{\pm2.74}$	$\mathbf{89.06}_{\pm1.49}$	$\mathbf{91.84}_{\pm0.61}$	$\mathbf{92.64}_{\pm3.43}$	$\mathbf{89.47}_{\pm0.47}$	$\mathbf{97.92}_{\pm1.06}$	$\mathbf{79.24}_{\pm4.87}$	$\mathbf{94.56}_{\pm0.51}$

Empirical Analysis

- Probe analysis about the properties introduced by our proposed loss function
  - Layer wise-accuracy, loss, sharpness and step ratio



DEPARTMENT OF ENGINEERING SCIENCE

UNIVERSITY OF OXFORD

Thanks