Hongyuan Dong, Dingkang Yang, Xiao Liang, Chao Feng, Jiao Ran

AdaLRS: Loss-Guided Adaptive Learning Rate Search for Efficient Foundation Model Pretraining

Motivations

Workflow of AdaLRS

Convergence Analysis

Experiments

Conclusions

Motivations

Learning rate is widely regarded as crucial for modern foundation model pretraining.

However, existing LR searching algorithms **lack generalizability:**

- Optimal LR prediction methods summarize model performance dynamics as a function w.r.t. hyperparameter settings, requiring hundreds of reapeated experiments to form an optimal learning rate expression.
- Optimal LR transferring methods typified by Tensor Program series seek to transfer the optimal LR across different model scales, but they also necessitate specific designs in the model architecture.

Our AdaLRS algorithm conducts adaptive LR search in a task-agnostic way during training, reducing the resource costs for LR tuning significantly.

Motivations

Workflow of AdaLRS

Convergence Analysis

Experiments

Conclusions

Workflow of AdaLRS

1. Monitor the loss descent velocity (loss slope) dynamics in k-step windows.
2. Conduct attemptive learning rate upscaling when the loss descent velocity decays.
3. Compare the loss slope before and after the upscaling attempt: if the loss descent velocity increases, the training proceeds with the upscaling retained; if the velocity decreases, the upscaling is reverted and the learning rate is lowered instead.

$$\eta_{t+k} = \begin{cases} \alpha'\eta_t & \text{if } v(\alpha'\eta_t) > v(\eta_t) + 2e \quad \text{(loss slope } \textit{increases} \uparrow\text{)}, \\ \beta'\eta_t & \text{if } v(\alpha'\eta_t) < v(\eta_t) - 2e \quad \text{(loss slope } \textit{decreases} \downarrow\text{)}, \\ \eta_t & \text{otherwise.} \end{cases}$$

Workflow of AdaLRS

4. In the formulation, $\alpha^{'}$ and $\beta^{'}$ are LR upscaling and downscaling factors. We choose multiplicatively independent base factors α and β for precise approximation. $\alpha^{'} = max(\lambda^t \alpha, 1)$, $\beta^{'} = \frac{1}{max(\lambda^t \beta, 1)}$ are rectified LR scaling factors to ensure convergence.

$$\eta_{t+k} = \begin{cases} \alpha' \eta_t & \text{if } v(\alpha' \eta_t) > v(\eta_t) + 2e \quad \text{(loss slope } \textit{increases} \uparrow\text{)}, \\ \beta' \eta_t & \text{if } v(\alpha' \eta_t) < v(\eta_t) - 2e \quad \text{(loss slope } \textit{decreases} \downarrow\text{)}, \\ \eta_t & \text{otherwise.} \end{cases}$$

Motivations

Workflow of AdaLRS

Convergence Analysis

Experiments

Conclusions

Convergence Analysis

Hypothesis: *The optimization of the training loss w.r.t. learning rate in foundation model pretraining is convex, and share the same optimum with loss descent velocity optimization.*

- Theoretical Analysis

  We formulate the update rule for model parameter ψ as:

  $$\psi_{t+1} = \psi_t - \eta \nabla L_t,$$

  where L_t is the training loss at time step t. The expected loss descent velocity per step (using Taylor expansion) is:

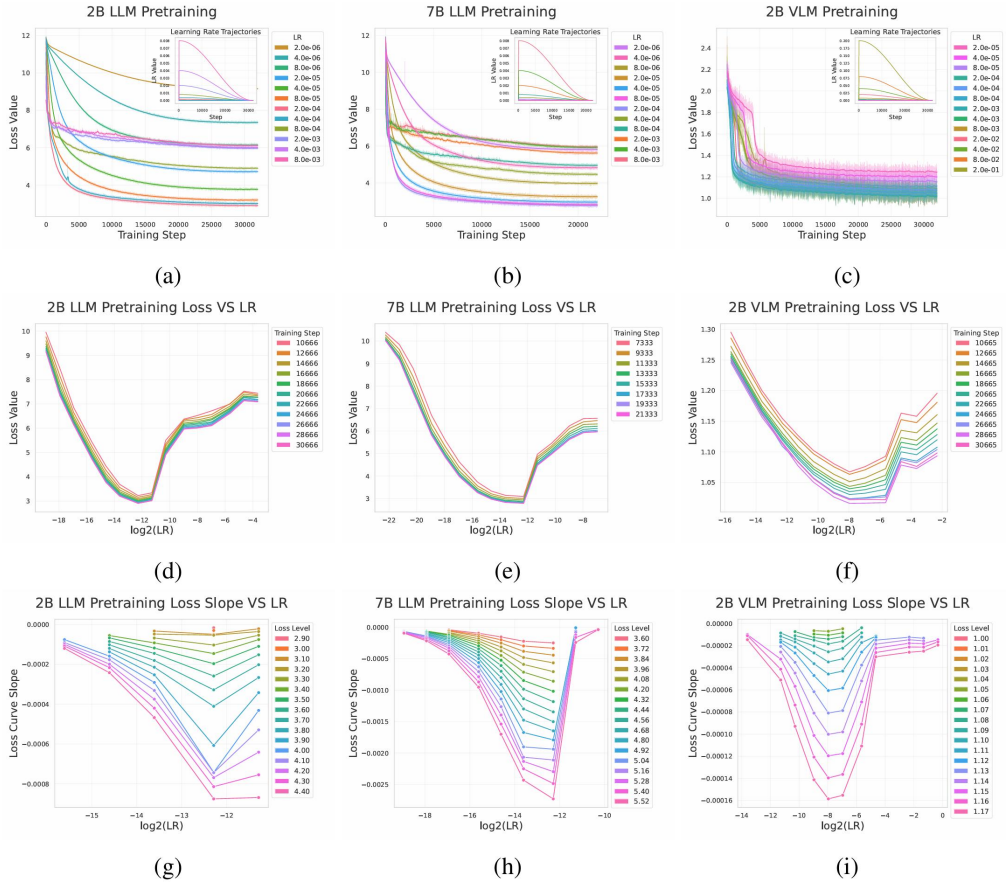
  $$\mathbb{E}\left[L_{t+1} - L_t\right] \approx -\eta \|\nabla L_t\|^2 + \frac{C_{Lip}}{2} \eta^2 \|\nabla L_t\|^2,$$

  When the LR is too small or too large, the first and second term dominates to suppress the expected loss descent value, respectively, ensuring the convexity of training loss slope w.r.t. LR settings.

Convergence Analysis

Hypothesis: *The optimization of the training loss w.r.t. learning rate in foundation model pretraining is convex, and share the same optimum with loss descent velocity optimization.*

- Experimental Analysis
  - (a)(b)(c) show the training loss under diffe-rent LR settings; (d)(e)(f) are training loss dynamics at different training steps; (g)(h)(i) illustrate the loss slope dynamics at varying loss levels.
  - We can observe clear convexity in these plots, and the shared optimum between the training loss and its slope.



Convergence Analysis

Theoreom: *The LR sequence generated by the AdaLRS algorithm converges almost surely to the e-neighborhood of the optimal learning rate.*

This theoreom can be easily proved through the following two propositions:

Proposition 1: *When the current LR is not in the e-neighborhood of the optimum, AdaLRS will adjust it to move towards the optimum in finite steps.*

> This is straightforward because AdaLRS adjusts LR w.r.t. loss slopes, and the suboptimal LR will be pushed towards the optimum because of the convexity of loss slopes.

Proposition 2: *Under the LR adjustment of AdaLRS, the gap between the current LR and the optimum is bounded by the decaying LR scaling factors.*

> As illustrated in the workflow of AdaLRS, the LR scaling factors decay as the training proceeds, ensuring the resulted final LR to lie in the optimum neighborhood.

Motivations

Workflow of AdaLRS

Convergence Analysis

Experiments

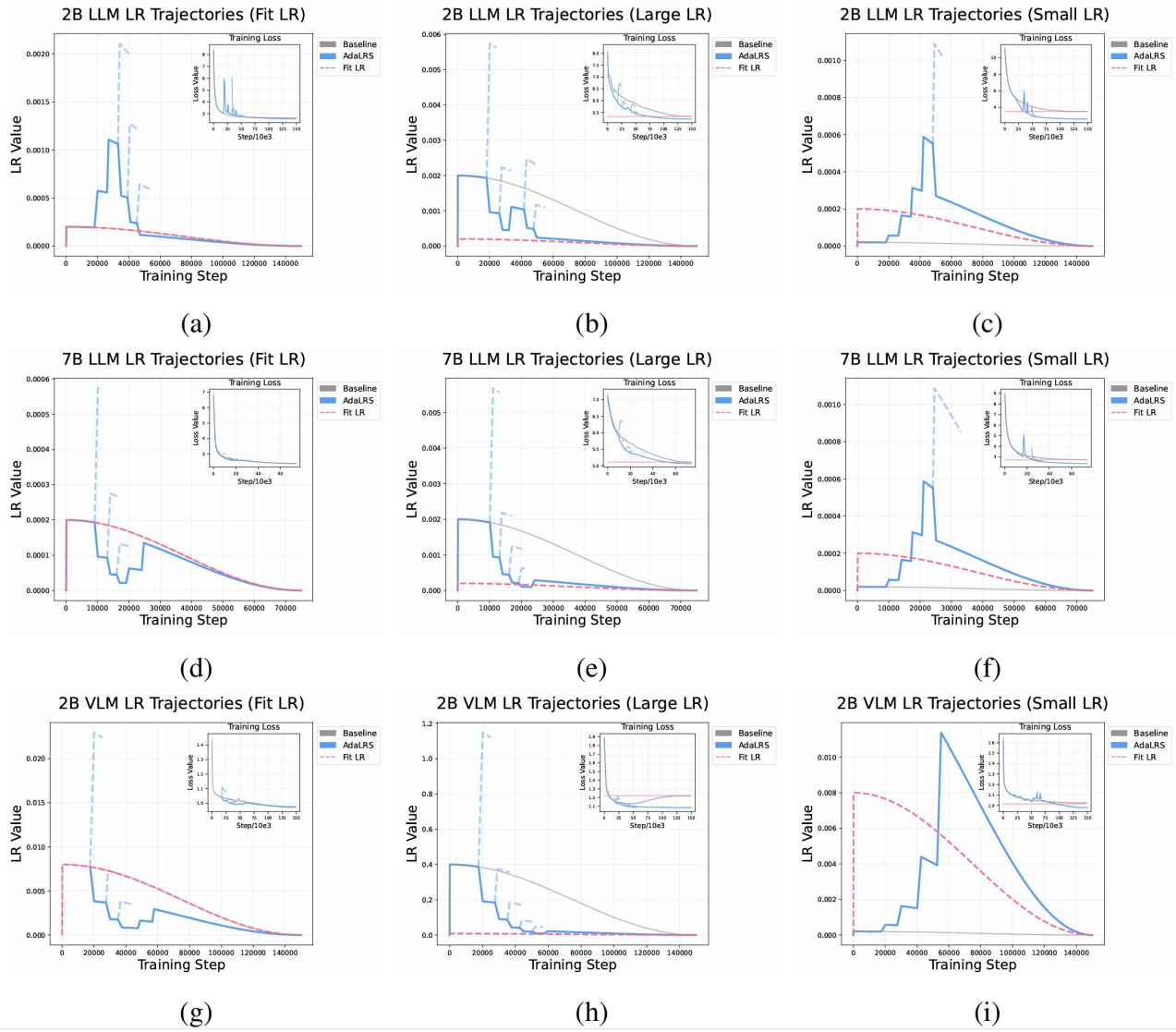
Conclusions

Experiments

Main Experiment Settings

- 128B-token scale fom scratch pretrai-ning on 2B LLM, 7B LLM, and 2B VLM.
- Appropriate LR, as well as LRs excessive-ly large or small to demon-strate Ada-LRS's convergence and effectiveness.

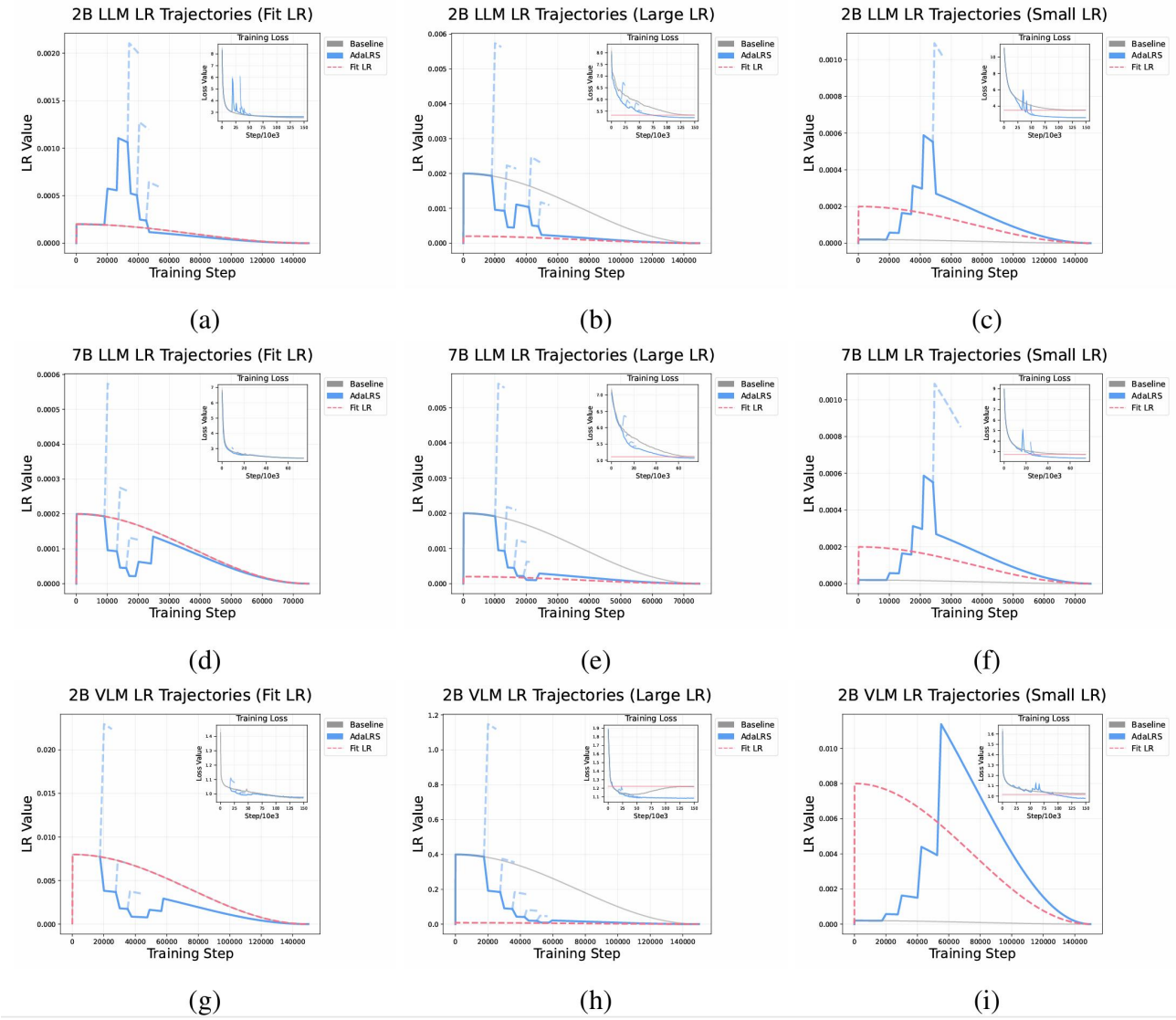
Hyperparameter	*2B LLM* Fit	Large	Small	*7B LLM* Fit	Large	Small	*2B VLM* Fit	Large	Small
Learning Rate	$2e^{-4}$	$2e^{-3}$	$2e^{-5}$	$2e^{-4}$	$2e^{-3}$	$2e^{-5}$	$8e^{-3}$	$4e^{-1}$	$2e^{-4}$
BSZ / Micro BSZ	1024/512			2048/512			2048/1024		
Window Size k	2500			2000			1000		
Data Composition	Detail Caption & OCR			SlimPajama Train Set			SlimPajama Train Set		
Search Step Ratio	$[0.1, 0.4]$			$[0.1, 0.35]$			$[0.1, 0.35]$		



Experiments

Main Experiment Results

- AdaLRS adjusts suboptimal LRs to the neighborhood of the optimum effectively in a single run.
- AdaLRS accelerates pretraining loss convergence efficiently, surpassing suboptimal LR baselines with only 50% training steps.
- AdaLRS generalizes robustly across different model sizes, data compositions, and model architectures.



Experiments

Main Experiment Results

- LLMs/VLMs trained with AdaLRS out-performs baseline models across a series of evaluation criteria, including training loss, test perplexity, and down-stream tasks.

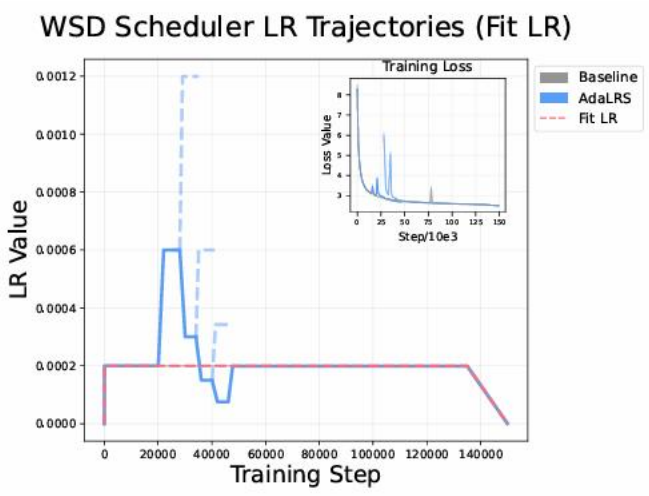
Benchmark	*2B LLM* Fit LR	*2B LLM* Large LR	*2B LLM* Small LR	*7B LLM* Fit LR	*7B LLM* Large LR	*7B LLM* Small LR
Training Loss	2.62/2.54	5.21/5.32	2.56/3.50	2.38/2.39	5.07/5.11	2.38/2.74
Training PPL	13.72/12.65	183.28/205.00	12.88/33.21	10.84/10.88	158.54/165.67	10.76/15.49
Validation PPL	13.51/12.42	183.94/204.97	12.66/32.69	10.67/10.72	158.22/165.70	10.61/15.30
Test PPL	13.51/12.43	183.70/204.68	12.66/32.66	10.69/10.74	158.16/165.61	10.63/15.32
Alpaca-Gen	17.29/18.56	7.11/6.09	17.10/15.40	21.76/21.61	6.35/5.55	21.15/20.68
KNIGHT-Gen	10.13/11.29	3.96/2.02	10.81/8.36	13.62/13.35	4.08/3.82	13.53/12.29

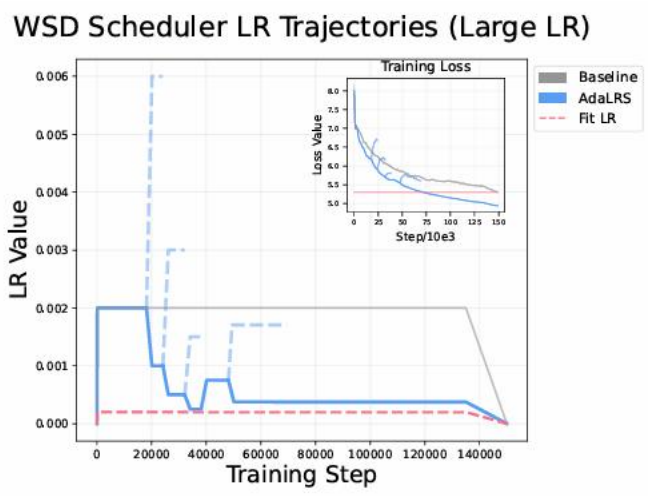
LR Setting	LLaVABench	MMVet	MMStar	DocVQA	OCRBench	TextVQA	DetailCaps-4870	Average
Fit LR	39.5/38.5	34.58/32.02	48.67/49.53	77.99/78.00	718/735	64.89/63.74	55.68/55.30	56.16/55.80
Large LR	36.8/35.7	31.47/30.50	44.47/44.33	57.53/57.42	631/606	60.32/58.08	49.02/47.08	48.96/47.67
Small LR	44.3/39.2	36.15/30.23	48.67/49.20	77.75/77.47	730/689	64.85/57.86	56.65/53.51	57.34/53.77

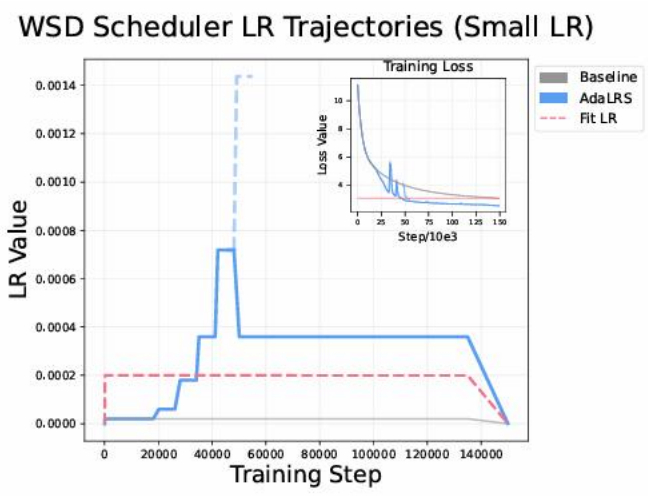
Experiments

Extending AdaLRS to WSD Scheduler

- Experiments on 2B LLM pretraining with LR settings appropriate, too large or too small.
- AdaLRS adjusts unreasonable LR settings effectively, and shows desired stability for appropriate LR configurations.







Experiments

Robustness Across Hyperparameter Settings

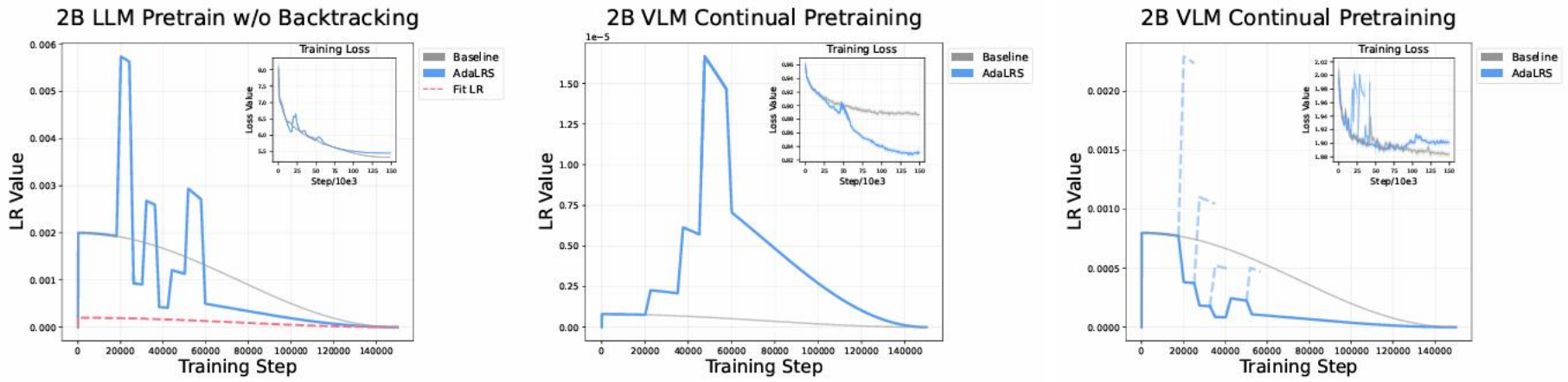
- Experiments on 2B LLM pretraining with varying AdaLRS hyperparameter settings, such as LR scaling factors α, β and the decay factor λ.
- The precision of optimal LR approximation may be influenced by the hyperparameter setting, but AdaLRS approximates the optimum and improves model performance robustly across different settings.

	Small LR						Large LR				
α/β	–	3/2	2/1.67	1.5/1.43	2/1.67	2/1.67	–	3/2	2/1.67	1.5/1.43	2/1.67
λ	–	0.99	0.99	0.99	0.95	0.9	–	0.99	0.99	0.99	0.9
Final LR	$2.0e^{-5}$	$3.6e^{-4}$	$3.1e^{-4}$	$1.5e^{-4}$	$3.1e^{-4}$	$2.2e^{-4}$	$2.0e^{-3}$	$3.8e^{-4}$	$5.2e^{-4}$	$7.2e^{-4}$	$7.9e^{-4}$
Training Loss	3.07	2.55	2.55	2.58	2.54	2.54	5.30	4.94	5.10	5.15	5.08
Validation PPL	21.59	12.87	12.75	13.11	12.61	12.68	201.54	140.13	163.95	172.01	159.55
Test PPL	21.61	12.89	12.77	13.13	12.63	12.70	201.36	140.07	163.81	171.85	159.47

Experiments

Backtracking LR Downscaling Strategy Ablation

- (a): Without the backtracking LR downscaling strategy, the large LR overshooting will introduce severe training instability. Even if AdaLRS still approximates the optimal LR, the corresponding model performance remain detereorating.

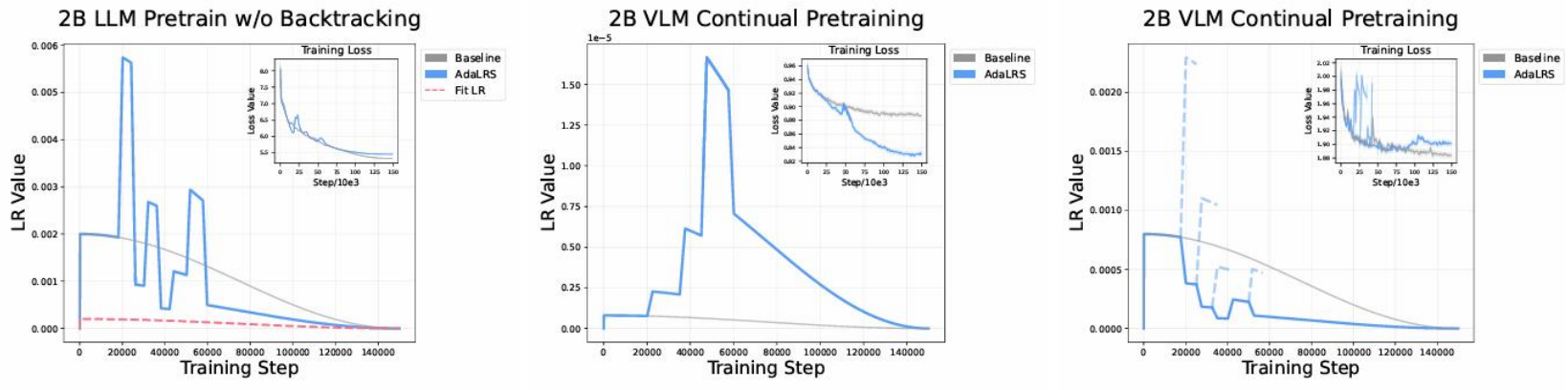


(a) (b) (c)

Experiments

AdaLRS for Continual Pretraining

- (b)(c): Applying AdaLRS in VLM pretraining with pretrained vision encoder unfrozen for parameter updates.
- For small LR settings, AdaLRS upscales the LR as expected, improving the training loss significantly; for large LR settings, AdaLRS exhibits desired effectiveness in optimal LR approximation, but fails to improve training loss because large LR destructs model parameter distribution.



(a) (b) (c)

Motivations

Workflow of AdaLRS

Convergence Analysis

Experiments

Conclusions

Conclusions

- We propose AdaLRS, which conducts online optimal learning rate search by optimizing the loss descent velocity, approximating the optimal LR and improving model performance effectively in a single run.
- We validate our approach with rigorous convergence analysis, and provide both theoretical and experimental analysis to support our hypothesis: the training loss and its slope are convex w.r.t. LR settings, and they share the same optimum.
- Experiments show the effectivenss of AdaLRS in various training scenarios, including different model sizes, data compositions, model architectures, and base LR schedulers.

THANK YOU

Hongyuan Dong

d_ousia@icloud.com