Warmstarting for Scaling Language Models

Neeratyoy Mallik*, Maciej Janowski*, Johannes Hog, Herilalaina Rakotoarison, Aaron Klein, Josif Grabocka, Frank Hutter

** Equal contribution*

Improving convergence rate of larger models by warmstarting training from a smaller model under Chinchilla compute-optimal training.

- Scaling studies and large model trainings often do not have directly transferable hyperparameter settings.
- Prohibitive tuning cost at large scales lead to custom tuning decisions across different setups.
- Reducing training cost at larger scales by reusing trained smaller models.
- Leverage tuning decisions made at smaller model scales.

Scaling best found learning rate at smallest model scale using ***μ*P.**

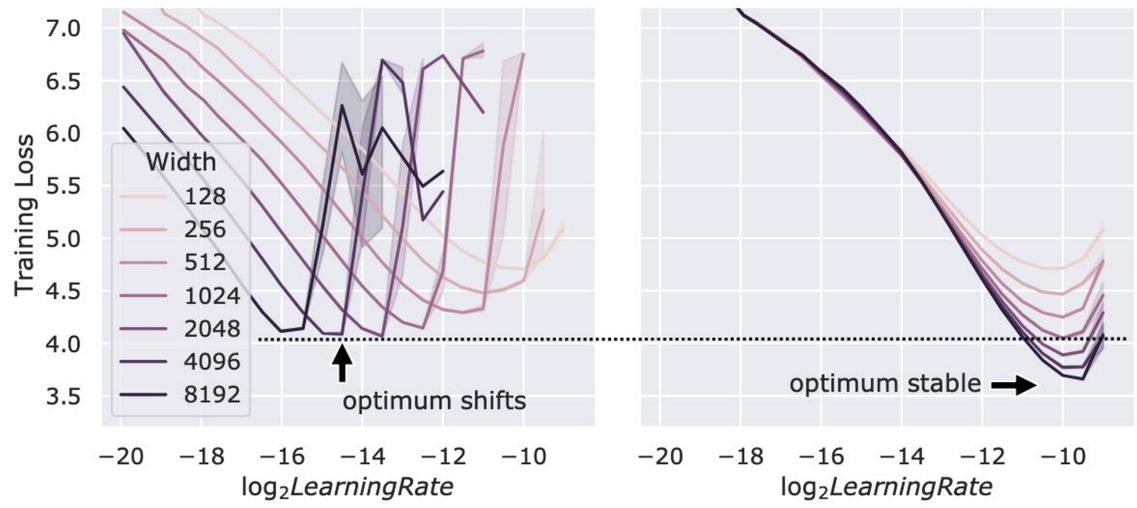


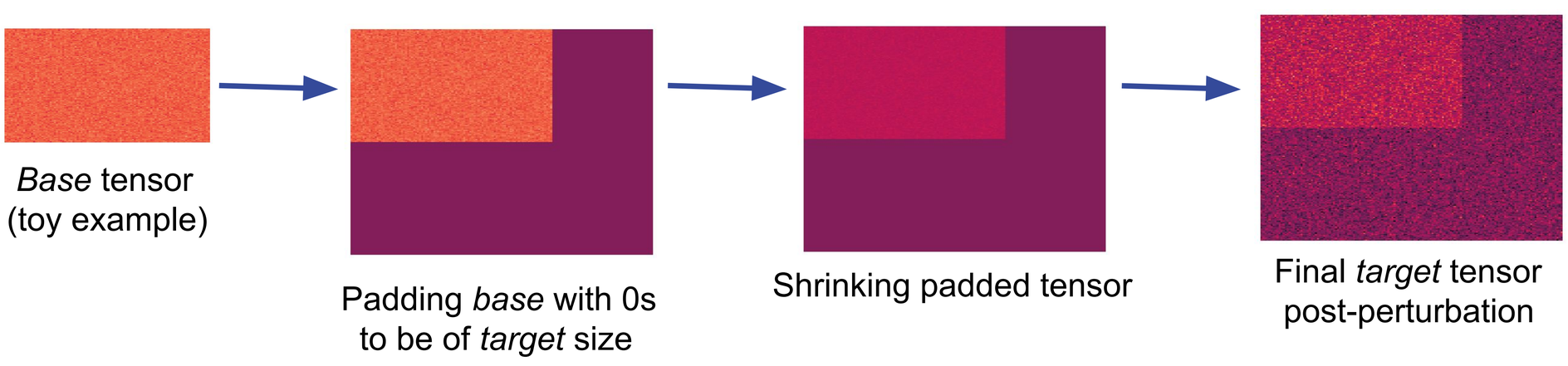
Figure source: [1]

Our Approach

*Initialize the larger model training run as a scaled up continual learning over the smaller model, assuming **μP**-enabled training pipeline.*

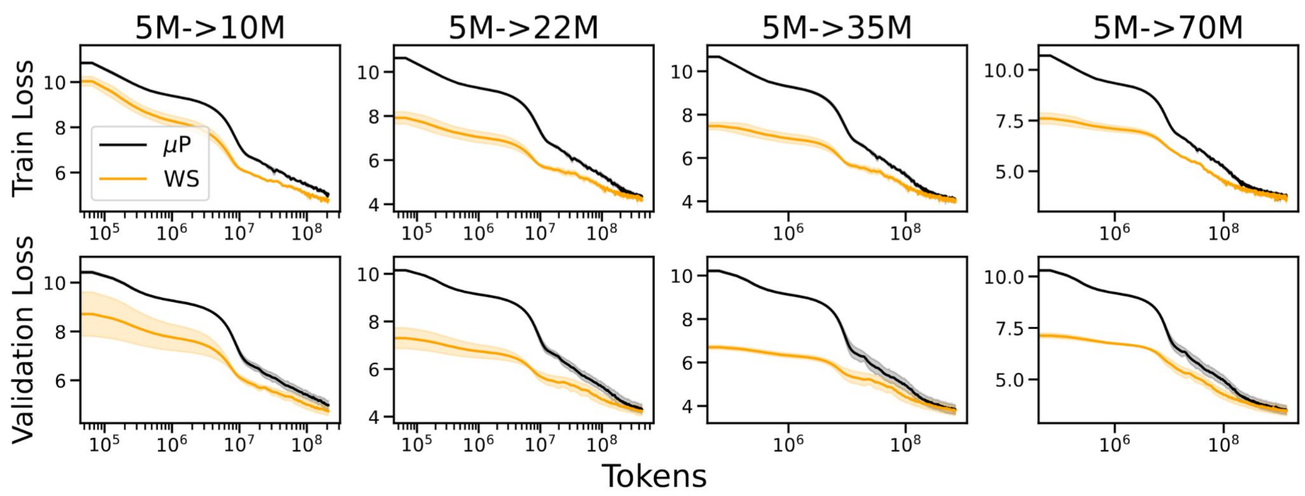
$$\theta^{l}_{\text{target}} = \lambda_{\text{shrink}} \cdot \texttt{Pad}_0(\theta^{l}_{\text{base}}, p, q) + \mathcal{N}(0,\ {\sigma^{l}_{\mu\text{P}}}^2)$$

$\lambda_{\text{shrink}} \in \mathbf{R}^1$; *base* → smaller model; *target* → larger model
θ^{l} → weights as tensor in layer l
Pad_0 → zero pads a tensor to a larger tensor

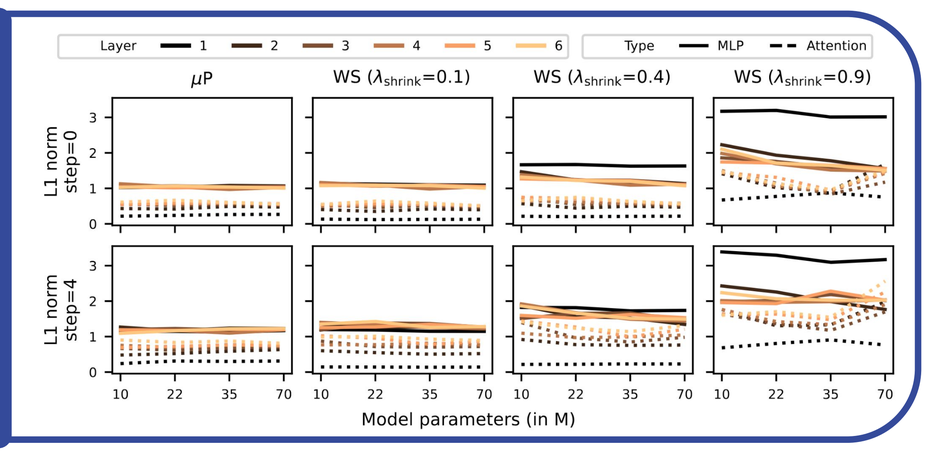


$\lambda_{\text{shrink}} = 0$ recovers ***μP*** without warmstarting

Empirical Evaluation

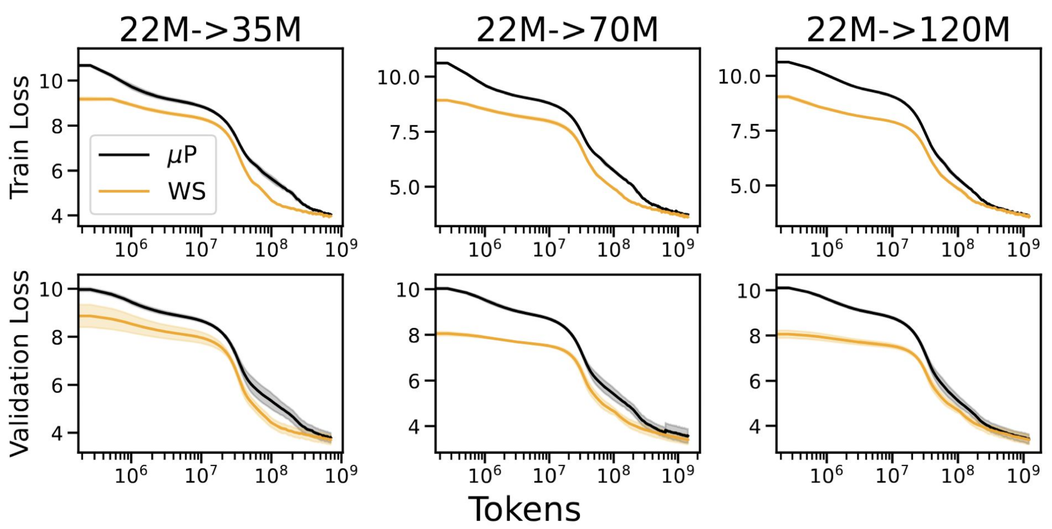


Verification

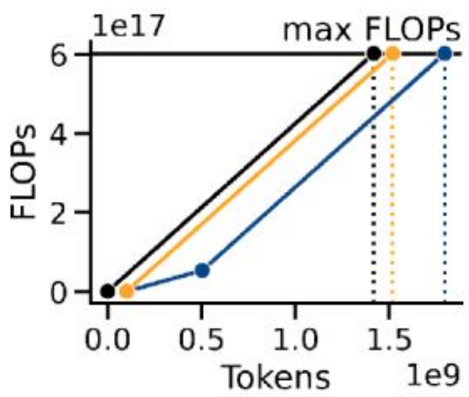


Setup

- GPT2 architecture and vocab
- Sub-epoch training on the Slimpajama dataset [2]
- All models trained for a compute budget of 20 tokens per parameter [3]
- Trained with Adam with constant learning rate schedule [4]



Future directions:



[1] G. Yang et al. Tuning large neural networks via zero-shot hyperparameter transfer. NeurIPS, 2021.
[2] D. Soboleva, et al. SlimPajama: A 627B token cleaned and deduplicated version of RedPajama, HF, 2023.
[3] G. J. Hoffmann et al. Training Compute-Optimal Large Language Models. 2022.
[4] A Hägele et al. Scaling Laws and Compute-Optimal Training Beyond Fixed Training Durations. NeurIPS 2024.
[5] G. Yang and E. J Hu. Tensor programs IV: Feature learning in infinite-width neural networks. ICML, 2021.
[6] J. Ash and R. P. Adams. On warm-starting neural network training. NeurIPS, 2020.
[7] K. E. Everett et al. Scaling exponents across parameterizations and optimizers. ICML, 2024.